线性优化

Linear and Convex Optimization

与凸优化

秦一航 Y. Qin

5200309101SS

Introduction to Optimization

Def. $\min_{x \in X} f(x)$ i.e. minimize $f(x)$ subject to $x \in X$

where $f(x): \mathbb{R}^n \to \mathbb{R}$ is objective function (目标函数)

x is the optimization / decision variable

X is the feasible / constrain set

We say "x is *feasible*" if $x \in X$.

x is *infeasible* if $x \notin X$.

We say the problem is *unconstrained* if $X = \mathbb{R}^n$.

and *constrained* if $X \neq \mathbb{R}^n$.

X is often specified by *constrained functions*, i.e.

$$\min f(x)$$
$$\text{s.t. } g_i(x) = 0 \quad i = 1, 2, \dots, m$$
$$h_i(x) \le 0 \quad i = 1, 2, \dots, k$$

here, $X = \{ x \in \mathbb{R}^n \mid g_i(x) = 0,\ i = 1, \dots, m;\ h_i(x) \le 0,\ i = 1, \dots k \}$

Note. Under strict definition, only "$y = kx$" is referred to as *linear function* while "$y = kx + b\ (b \neq 0)$" is named *affine function*

In this notebook, we use "*linear function*" to mean both of them.

e.g. Data Fitting: $\min_{k,b>0} \sum \epsilon_i^2$ where ϵ_i is the "measurement error" (误差).

Def. A linear model predicts a response:

$$\hat{y} = f(x) = w^T x + b = \sum_{i=1}^{n} w_i x_i + b \quad \sim \text{Target: } y$$

($\hat{y}$: Prediction)

$$\text{Target: } \min_{w \in \mathbb{R}^n,\, b \in \mathbb{R}} \sum_{i=1}^{m} (f(x_i) - y_i)^2 = \min_{w \in \mathbb{R}^n,\, b \in \mathbb{R}} \sum_{i=1}^{m} (x_i^T w + b - y_i)^2$$

$$=: \min_{w \in \mathbb{R}^n,\, b \in \mathbb{R}} \| Xw + b\mathbf{1} - y \|^2$$

where $X = (x_1 \ \cdots \ x_m)^T = \begin{pmatrix} x_1^T \\ \vdots \\ x_m^T \end{pmatrix}$, $\mathbf{1} = (1, \cdots, 1)^T$

$$\|\alpha\| = \sqrt{\alpha^T \alpha} = \sqrt{\sum_{i=1}^{n} \alpha_i^2}$$

Norm 范数

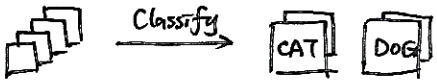
In Practice, we use $\tilde{x}_i = \begin{pmatrix} x_i \\ 1 \end{pmatrix}$ $\tilde{X} = \begin{pmatrix} x_1^T & 1 \\ \vdots & \vdots \\ x_m^T & 1 \end{pmatrix}$ $\tilde{w} = \begin{pmatrix} w \\ b \end{pmatrix}$

The target is then $\min_{\tilde{w} \in \mathbb{R}^{n+1}} \| \tilde{X} \tilde{w} - y \|^2$

Def. Linear Program 线性规划

$$\min \quad f(X) \qquad \text{s.t.} \quad \begin{aligned} g_i(x) &= 0 \qquad i = 1, 2, \dots, m \\ h_i(x) &\geq 0 \qquad i = 1, 2, \dots, k \\ x_{ij} &\geq 0 \qquad i, j \in \{\ \} \times \{\ \} \end{aligned}$$

Def. Binary Classification 分类器



x $\quad f(x)$ $\quad \{+1, -1\}$

Use x (vector) to represent images.

$\{-1, +1\}$ represents {dog, cat}.

$$\begin{cases} f(x_i) > 0 & \text{iff. } y_i = +1 \\ f(x_i) < 0 & \text{iff. } y_i = -1 \end{cases} \iff f(x_i)\, y_i > 0$$

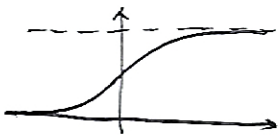
↑ Classifier

To predict an image: result: $\operatorname{sgn} f(x)$

How to find "f"?

1) Logistic Regression

$$\sigma(z) := \frac{1}{1 + e^{-z}}$$



* $\sigma(z)$ can be used to mean any "S-curve" functions in fact.

$$p(y|x) = \sigma(yf(x)) = \sigma(y(w^T x + b))$$

Target: $\max_{w,b} \mathcal{L}(w,b) = \max_{w,b} \prod p(y_i|x_i)$ Likelihood

product: unstable

also difficult when you apply gradient. (求导: $(uv)' = u'v + v'u$)

$$\iff \min_{w,b} NLL(w,b) = -\log \mathcal{L}(w,b) = \sum_{i=1}^{m} \log\left(1 + e^{-y_i(x_i^T w + b)}\right)$$

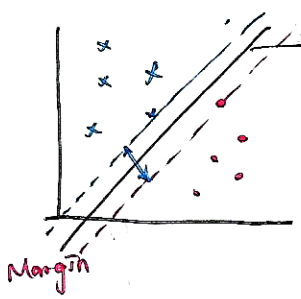
Negative Log Likelihood

↑ more terms

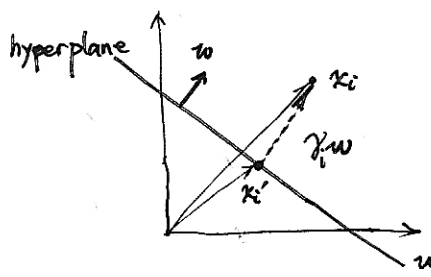
↑ a convex func. in fact !!

log: turn the product into a sum. simplify the question.

2) Support Vector Machine



margin ↑ ⟶ more robust to noise



$x_i = x_i' + \gamma_i w$ (γ_i: "Distance")

$w^T x_i' + b = 0$

$\Rightarrow w^T(x_i - \gamma_i w) + b = 0$

Thus, $\gamma_i = \dfrac{w^T x_i + b}{w^T w}$ ← the "distance" of x_i to the hyperplane $w^T x + b = 0$

the "margin" $\gamma_0 = \min_{1 \le i \le m} \|\gamma_i w\|$ ⟶ the target: $\max_{w,b} \gamma_0 = \max_{w,b} \min_{1 \le i \le m} \|\gamma_i w\|$

$= \max_{w,b} \min_{1 \le i \le m} \left| \dfrac{w^T x_i + b}{w^T w} \right| \dfrac{\|w\|^2}{\|w\|}$

$= \max_{w,b} \min_{1 \le i \le m} \dfrac{|w^T x_i + b|}{\|w\|}$

s.t. $y_i (w^T x_i + b) \ge 0$ "Hard Margin"

No disclassification Allowed

Reform

$w^T x + b = 0$ Let $\tilde{w} = \alpha w$, $\tilde{b} = \alpha b$.

$\tilde{w}$ and $\tilde{b}$ determines a hyperplane.

$x \in P \Leftrightarrow w^T x + b = 0 \Leftrightarrow \tilde{w}^T x + \tilde{b} = 0$ (hyperplane)

Choosing α properly, we can assume $\min_{1 \le i \le m} y_i (\tilde{w}^T x_i + \tilde{b}) = 1$.

$\Rightarrow |\tilde{w}^T x_i + \tilde{b}|$ (since $y_i = \mathrm{sgn}(\tilde{w}^T x_i + \tilde{b})$)

$= \mathrm{sgn}(\tilde{w}^T x_i + \tilde{b})$

$\max_{w,b} \gamma_0 = \max_{w,b} \dfrac{1}{\|w\|} \leftrightarrow \min_{w,b} \dfrac{1}{2}\|w\|^2$ (since $\min_{w,b} \|w\| \Leftrightarrow \min_{w,b} \frac{1}{2}\|w\|^2$)

Thus, Target: $\min_{w,b} \dfrac{1}{2}\|w\|^2$, s.t. $y_i (w^T x_i + b) \ge 1$ □

Soft Margin SVM

Introduce slack variables $\xi = (\xi_1, \cdots \xi_n)^T$

$\min_{w,b} \dfrac{1}{2}\|w\|_2^2 + c \sum_i \xi_i$ ($c > 0$: a hyper parameter)

s.t. $y_i (w^T x_i + b) \ge 1 - \xi_i$, $i = 1, 2, \cdots, n$

$\xi \ge 0$ (i.e. $\xi_1 \ge 0, \cdots \xi_n \ge 0$)

Def. Global Optima: $x^* \in X$ is a global minimum of f over X if

$f(x^*) \le f(x)$. $\forall x \in X$

optimal value of $\min_{x \in X} f(x)$ ↑ x^* is an optimal solution of $\min_{x \in X} f(x)$.

Extremum: Maximum and minimum are called extremum.

(might not exist) e.g. $\inf f(x)$ → unbounded below ($-\infty$) / not achievable

Def. (Linear Algebra Reprise)

Euclidean Inner Product: $\langle x, y \rangle = x^T y = \sum_{i=1}^{n} x_i y_i$

Euclidean Norm (2-norm, ℓ_2-norm): $\|x\|_2 = \sqrt{x^T x} = \sqrt{\sum_{i=1}^{n} x_i^2}$.

2范数

A norm on $\mathbb{R}^n$ is a function $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$ satisfying

范数

1) $\|x\| \geq 0 \quad \forall x \in \mathbb{R}^n$ 2) $\|x\| = 0$ iff. $x = 0$

3) $\|ax\| = |a|\|x\|$, $\forall a \in \mathbb{R}$, $x \in \mathbb{R}^n$ (positive homogeneity)

4) $\|x+y\| \leq \|x\| + \|y\|$. $\forall x, y \in \mathbb{R}^n$ (Triangle Inequality)

p-norm: $\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}$, $p \geq 1$ → * Minkowski Inequality Used to prove Property 4).

1-norm: $\|x\|_1 = \sum_i |x_i|$ (abs)

2-norm: $\|x\|_2 = \sqrt{x^T x}$ (By default, $\|x\| = \|x\|_2$)

Euclidean Norm

∞-norm (Max): $\|x\|_\infty = \max_i |x_i| = \lim_{p \to \infty} \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}$

Def. Open ball $B(x_0, r) = \{x : \|x - x_0\| < r\}$

Closed ball $\bar{B}(x_0, r) = \{x : \|x - x_0\| \leq r\}$

A set is open if $\forall x \in S$ exists $\varepsilon > 0$ s.t. $B(x, \varepsilon) \subset S$

is closed if its complement S^c is open.

Convergence: a sequence $\{x_n\}$ converges to x: $x_n \to x$

收敛

$\lim_{n \to \infty} x_n = x$ (if $\lim_{n \to \infty} \|x_n - x\| = 0$)

Thm. if S is closed, $\forall$ sequence $\{x_n\} \subset S$. $x_n \to x \in S$.

Def. Compactness

A set is bounded if exists $M \leq \infty$. s.t. $\|x\| < M$, $\forall x \in S$

A set is compact if it is closed and bounded.

Def. Continuity

A function $f : X \subset \mathbb{R}^n \to \mathbb{R}$ is continuous at $x \in X$ if

$\forall \varepsilon > 0$, $\exists \delta > 0$ s.t.

$y \in X \cap B(x, \delta) \Rightarrow |f(y) - f(x)| < \varepsilon$

i.e. $\forall \{x_n\} \subset X,\ x_n \to x \Rightarrow f(x_n) \to f(x)$

We say "f is continuous on X" if it is continuous at every $x \in X$.

Thm. If $f: \mathbb{R}^n \to \mathbb{R}$ is continuous on $\mathbb{R}^n$. then

$\forall\, c \in \mathbb{R}$. 1) $\{x : f(x) < c\}$ is open

2) $\{x : f(x) \le c\}$ is closed.

Thm. (Extreme Value Thm.)

If f is continuous on a compact set X. then

exists $x_1, x_2 \in X$ (not necessary unique) s.t. $f(x_1) \le f(x) \le f(x_2)$ $(\forall x \in X)$

sufficient conditions but not necessary

e.g. $f(x) = x^2$ on $(-1, 2]$. ~~global min exists~~. global minimum exists. (when $x=0$)

Corollary. If f is continuous on $\mathbb{R}^n$ and $f(x) \to \infty$ as $\|x\| \to \infty$.

then global minimum exists, i.e. exists x^* s.t. $f(x^*) \le f(x)$. $\forall x$.

Proof. Since $\|x\| \to \infty$. $f(x) \to \infty$. exists $M > 0$ s.t. when $\|x\| > M$. $f(x) > f(0)$

The closed ball $[-M, M] =: \bar{B}(0, M) = \{x \mid \|x\| \le M\}$ is compact.

By Extreme Value Thm., exists $x^* \in \bar{B}(0, M)$. s.t. $f(x^*) \le f(x)$. $\forall x \in \bar{B}(0, M)$

On the other side, $f(x^*) \le f(0) < f(x)$. $\forall x \in \mathbb{R}^n \setminus \bar{B}(0, M)$.

Thus, $\forall x$. $f(x^*) \le f(x)$. i.e. global min exists. □

Def.

A function is called coercive if $f(x) \to \infty$ as $\|x\| \to \infty$

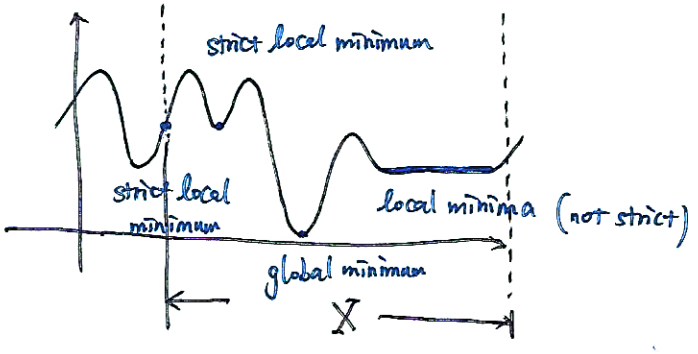
Def. $x^* \in X$ is a *local minimum* of f on X if exists $\varepsilon > 0$ s.t.

$f(x^*) \le f(x)$. $\forall x \in X \cap \bar{B}(x, \varepsilon)$

open.

x^* is a *strict local minimum* if strict inequality holds for $x \neq x^*$.

(none in $X \cap B(x, \varepsilon) \setminus \{x\}$ s.t. $f(x) = f(x^*)$)



global min is always a local min.

not vice versa.

(→ Convex. will reprise)

Def. x is an interior point of $X \subset \mathbb{R}^n$ if exists $\varepsilon > 0$ s.t. $B(x, \varepsilon) \subset X$.

The interior of X. (denoted by $\operatorname{int} X$) is the set of all interior points of X.

Def. A function $f: X \subset \mathbb{R}^n \to \mathbb{R}^m$ is differentiable at $x_0 \in \operatorname{int} X$ if exists a matrix $A \in \mathbb{R}^{m \times n}$ s.t.

$$\lim_{X \ni x \to x_0} \frac{\| f(x) - f(x_0) - A(x - x_0) \|}{\| x - x_0 \|} = 0$$

A: derivative We write $f'(x_0) = \mathcal{D} f(x_0) = A$

First-order Approximation: $f(x) = f(x_0) + A(x - x_0) + o(\|x - x_0\|)$

$f(x) \in \mathbb{R}^m$, $f(x_0) \in \mathbb{R}^m$, $A \in \mathbb{R}^{m \times n}$, $x - x_0 \in \mathbb{R}^n$

The derivative is given by Jacobian matrix of $f = (f_1, \cdots f_m)$

$$[f'(x_0)]_{ij} = \frac{\partial f_i(x_0)}{\partial x_j} \qquad i \in \{1, 2, \cdots, m\} \quad j \in \{1, 2, \cdots n\}$$

Proof. $f_i(x) - f_i(x_0) \approx \sum_k \frac{\partial f_i}{\partial x_k} \Delta x_k$.

$$f(x) - f(x_0) = \begin{pmatrix} f_1(x) - f_1(x_0) \\ \vdots \\ f_n(x) - f_n(x_0) \end{pmatrix} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix} \begin{pmatrix} \Delta x_1 \\ \vdots \\ \Delta x_n \end{pmatrix}$$

e.g. $\vec{f}(\vec{x}) = A\vec{x} + \vec{b}$. $\vec{f}'(x) = A$ (Can be proved).

For symmetric A, $\vec{f}(\vec{x}) = \vec{x}^T A \vec{x} = \sum_i \sum_j A_{ij} x_i x_j$ $\vec{f}'(\vec{x}) = 2\vec{x}^T A$

Proof. $f(\vec{x}) = \sum_i \sum_j A_{ij} x_i x_j$

$$\frac{\partial f(\vec{x})}{\partial x_k} = 2A_{kk} x_k + \sum_{i \neq k} (A_{ki} + A_{ik}) x_i = 2\left(A_{kk} x_k + \sum_{i \neq k} A_{ik} x_i\right)$$

(Symmetric!) $= 2 \sum_i A_{ki} x_i$

Thus, $f'(\vec{x}) = (x_1 \cdots x_n) \begin{pmatrix} A_{11} & \cdots & A_{1n} \\ A_{21} & \ddots & \vdots \\ \vdots & & \\ A_{n1} & \cdots & A_{nn} \end{pmatrix} = x^T A$ □

Note. For general A (not necessarily symmetric),

$x^T A x = x^T \tilde{A} x$. where $\tilde{A} = \frac{1}{2}(A + A^T)$.

Thus, $f'(x) = 2x^T \tilde{A} = x^T (A + A^T)$.

---

* To speed up note-taking, we discard "$\vec{x}$", "x" in the following chapters. But we'll note that it's a vec (might)

Prof. Since x^TAx is a number, we have $(x^TAx)^T = x^TAx$.

i.e. $x^TA^Tx = x^TAx$

Thus, $x^TAx = \frac{1}{2}(x^TAx + x^TA^Tx) = x^T\tilde{A}x$ ($\tilde{A}$: symmetric!) □

Def. For a function $f: \mathbb{R}^n \to \mathbb{R}$, the gradient of f at x (denoted by $\nabla f(x)$) is the transpose of derivative.

i.e. $\nabla f(x) = [f'(x)]^T$. $\quad [\nabla f(x)]_i = \frac{\partial f(x)}{\partial x_i}$

$\nabla f(x)$: a column vector s.t. $f'(x)\Delta x = \langle \nabla f(x), \Delta x \rangle = [\nabla f(x)]^T \Delta x$.

(inner product)

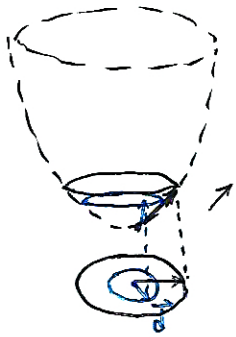
First-order Approximation: $f(x) = f(x_0) + \nabla f(x)^T \Delta x + o(\|\Delta x\|)$

$\nabla f(x)$ is the direction of fastest rate of increase of f at x.

$f(x+\vec{d}) - f(x) \doteq \nabla f(x)^T \vec{d} \le \|\nabla f(x)\| \cdot \|d\|$ (Let $\|d\| = 1$) □

where equality holds in the last step iff. $\vec{d} = \alpha \nabla f(x)$ for $\alpha > 0$. (denoted by $\vec{d} \propto \nabla f(x)$)

Directive ▶ Gradient: $\lim_{t \to 0} \frac{f(x+t\vec{d}) - f(x)}{t} = \vec{d}^T \nabla f(x)$



↗ $\nabla f(x)$: fastest of ~~test~~ increase

Thm. CHAIN RULE

If $f: X \subset \mathbb{R}^n \to \mathbb{R}^m$ is differentiable at $x_0 \in X$

$g: Y \subset \mathbb{R}^m \to \mathbb{R}^p$ is differentiable at $y_0 = f(x_0)$.

then the composition of f and g, denoted by $h(x) = g(f(x))$ is differentiable at x_0.

and $h'(x_0) = g'(y_0) f'(x_0) = g'(f(x_0)) f'(x_0)$

Proof. $\mathbb{R}^n \xrightarrow{f} \mathbb{R}^m \xrightarrow{g} \mathbb{R}^p$ (h: $\mathbb{R}^n \to \mathbb{R}^p$)

$x_0 \quad y_0 \quad z_0$

$\Delta x \to \Delta y = f'(x_0)\Delta x \to \Delta z = g'(y_0)\Delta y = g'(f(x_0)) f'(x_0)\Delta x$

$\Delta z = h'(x_0)\Delta x$

Thus, $h'(x_0) = g'(f(x_0)) f'(x_0)$ □

e.g. $g(x) = f(Ax+b)$ $\quad \nabla g(x) = A^T \nabla f(Ax+b)$

Thm. If x^* is a local minimum, then the gradient at x^* must vanish.

i.e. $\nabla f(x^*) = \left(\frac{\partial f(x^*)}{\partial x_1}, \cdots \frac{\partial f(x^*)}{\partial x_n} \right)^T = 0.$

Proof. Let $\vec{d} \in \mathbb{R}^n$. Define $g(t) = f(x^* + td)$

- Since x^* is a local minimum, $g(t) \geq g(0)$.
- For $t > 0$, $\frac{g(t) - g(0)}{t} \geq 0 \rightarrow g'(0) = \lim_{t \downarrow 0} \frac{g(t) - g(0)}{t} \geq 0$

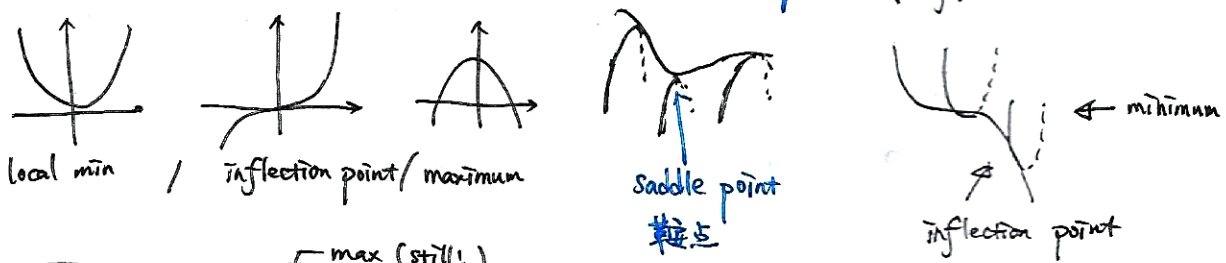
By Chain Rule, $g'(0) = \sum_{i=1}^{n} d_i \frac{\partial f(x^*)}{\partial x_i} = d^T \nabla f(x^*) \geq 0$

Setting $d = -\nabla f(x^*) \Rightarrow -\| \nabla f(x^*) \| \geq 0$ Thus, $\| \nabla f(x^*) \| = 0$.

Therefore, $\nabla f(x^*) = 0$ □

First-order Necessary Condition.

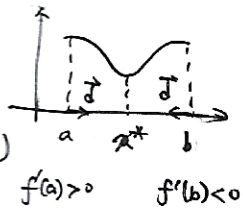
Def. A point x^* with $\nabla f(x^*) = 0$ is called a **stationary point** of f.



We can see that stationary points might be sufficient for convex functions. ← not proved here

For all $x \in X$ except for x^* ($\min/\max$ x-value) $\rightarrow \vec{d}^T f'(x) \geq 0$ holds

(for all feasible directions $\vec{d}$)



Second-order Conditions

Def. The second-order partial derivatives of $f: X \subset \mathbb{R}^n \rightarrow \mathbb{R}$ at $x_0 \in \text{int } X$

are $\frac{\partial^2 f(x_0)}{\partial x_i \partial x_j}$

Thm. If $\frac{\partial^2 f(x)}{\partial x_i \partial x_j}$ and $\frac{\partial^2 f(x)}{\partial x_j \partial x_i}$ exist in a neighborhood of x_0 and are continuous at x_0.

we have $\frac{\partial^2 f(x_0)}{\partial x_i \partial x_j} = \frac{\partial^2 f(x_0)}{\partial x_j \partial x_i}$

Hessian Matrix:

$$[\nabla^2 f(x)]_{ij} = \frac{\partial f(x)}{\partial x_i \partial x_j}$$

$$\begin{pmatrix} \frac{\partial^2 f}{\partial^2 x_1} & \cdots & \frac{\partial^2 f}{\partial x_1 x_n} \\ \vdots & & \vdots \\ \frac{\partial^2 f}{\partial x_1 \partial x_n} & \cdots & \frac{\partial^2 f}{\partial^2 x_n} \end{pmatrix}$$

e.g. $f(x) = x^T A x \longrightarrow \nabla f(x) = 2x^T A \longrightarrow \nabla^2 f(x) = 2A$

Proof. $\nabla(x^T A)$. $\frac{\partial f}{\partial x_i} = 2\sum_k A_{ki} x_k$ $\frac{\partial f}{\partial x_i x_j} = 2\sum_k \frac{\partial x_i}{\partial x_j} A_{ki} = 2A_{jk}$

CHAIN RULE. $g(x) = f(Ax+b) \longrightarrow \nabla^2 g(x) = A^T \nabla^2 f(Ax+b) A$

Thm. Second-order Taylor Expansion

Taylor Expansion: $g(a+t) = g(a) + g'(a)t + \frac{1}{2}g''(a)t^2 + \frac{1}{3!}g'''(a)t^3 + \cdots$

~~2nd order T.E. $g(a+t)$~~

$\vdots$

$\frac{1}{2} \cdot t \cdot g''(a) \cdot t$

$\vdots$

2^{nd}-order T.E. $f(x+d) = f(x) + \nabla f(x)^T d + \frac{1}{2} d^T \nabla^2 f(x) d + o(\|d\|^2)$

$= f(x) + \sum_i \frac{\partial f(x)}{\partial x_i} d_i + \frac{1}{2}\sum_{i,j} \frac{\partial^2 f(x)}{\partial x_i \partial x_j} d_i d_j + o(\|d\|^2)$

Denotation: $f(x+d) =: f(x+t\vec{d})$ where $t = \|d\|$. at $x = 0$.

then $g(\|d\|) = g(0) + g'(0)\|d\| + \frac{1}{2}g''(0)\|d\|^2 + o(\|d\|^2)$

Lagrange Remains: $g(a+t) = g(a) + g'(a)t + \frac{1}{2}g''(a+\xi)t^2$. $\xi \in (0, t)$.

Def. A matrix A is *positive semidefinite* (PSD) (半正定), denoted by $A \succeq 0$. if.

1) A is symmetric, i.e. $A = A^T$

2) $x^T A x \geq 0$ $(\forall x \in \mathbb{R}^n)$

positive definite: if 1) ... 2)' $x^T A x > 0$ $(\forall x \in \mathbb{R}^n$ and $x \neq 0)$ denoted by $A \succ 0$. (正定)

negative (semi-) definite: if $-A$ is positive (semi-) definite.

indefinite: $\exists x_1, x_2 \in \mathbb{R}^n$ s.t. $x_1^T A x_1 > 0 > x_2^T A x_2$.

[Trick (Reprise)] A: not symmetric $\leftarrow$ $\tilde{A} = \frac{1}{2}(A + A^T)$. symmetric.

[Trick] $A = BB^T$. must be *positive semidefinite*. (where B is a vector)

1) $A^T = (BB^T)^T = (B^T)^T B^T = BB^T = A$.

2) $x^T A x = x^T \underbrace{BB^T}_{\text{a vector, let it be } y} x = y^T y = \langle y, y \rangle = \|y\|^2 \geq 0$.

When is A *definite*? $\Rightarrow$ 2)' $\|y\|^2 > 0$ $(\forall x \in \mathbb{R}^n$ and $x \neq 0)$

$y = B^T x = (b_1 \cdots b_n)\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ $\begin{cases} x \neq 0 \Rightarrow y \neq 0 \\ y = 0 \Rightarrow x = 0 \end{cases}$ 唯一解 (零解) Thus, B must be *linearly independent*.

Def. Vector x is called *eigenvector* of a matrix A with associated *eigenvalue* λ if

$$Ax = \lambda x$$

(No change in direction, no rotation. only change the "Scale")

We can find all λ s by solving $\det(\lambda I - A) = 0$ where $I = \begin{pmatrix} 1 & & \\ & \ddots & \\ & & 1 \end{pmatrix}$

Thm. Let A be a symmetric matrix.

1) $A \succeq 0$ iff all $\lambda \geq 0$ 2) $A \succ 0$. iff all $\lambda > 0$,

Def. Principal Submatrix 主子矩阵

a $k\times k$ p.s. of matrix A consists of k rows and k columns with exactly same indices

$I = \{ i_1 \leq i_2 \leq \cdots \leq i_k \}$, i.e. $A_I = \begin{pmatrix} a_{i_1 i_1} & a_{i_1 i_2} & \cdots & a_{i_1 i_k} \\ \vdots & & \ddots & \vdots \\ a_{i_k i_1} & a_{i_k i_2} & \cdots & a_{i_k i_k} \end{pmatrix}$

e.g. $\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}$. a 2×2 p.s. : $A_{\{1,3\}} = \begin{pmatrix} 1 & 3 \\ 7 & 9 \end{pmatrix}$.

A *principal minor of order k* of A is $\det A_I$ for some I with $|I| = k$.
k阶主子式

If $I = \{1, 2, \cdots k\}$. $D_k(A) \triangleq \det A_I$ is called *the leading principal minor of order k*. k阶顺序主子式

Thm. (Sylvester) A = symmetric matrix

1) $A \succ 0$ iff $D_k(A) > 0$ for $k = 1, 2, \cdots, n$

2) $A \succeq 0$ iff $\det A_I \geq 0$ for all $I \subset \{1, 2, \dots n\}$

* We'll use $\succ$ (longer than $>$) to denote $\succ$ and $\succeq$ (longer than $\geq$) to denote $\succeq$ in this notebook (since it writes faster)

Def. a symmetric matrix $A \in \mathbb{R}^{n\times n}$ has the following *eigendecomposition*

$$A = Q \Lambda Q^T = \sum_{i=1}^{n} \lambda_i v_i v_i^T$$

where $\Lambda = \text{diag}(\lambda_1, \cdots \lambda_n)$ $Q = (v_1, \cdots v_n)$ is an *orthogonal matrix*

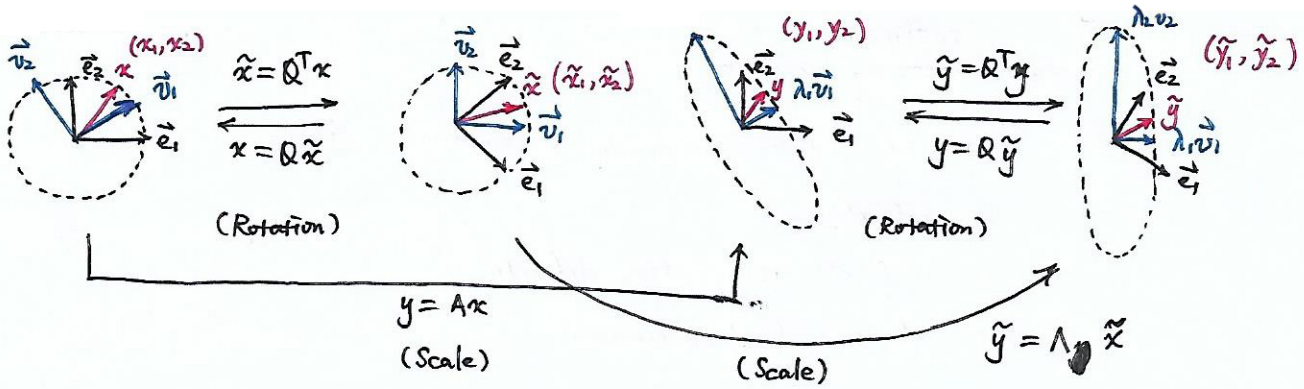
i.e. $Q^T Q = QQ^T = I = \text{diag}(1, \cdots 1)$. and $A v_i = \lambda_i v_i$.
(必正交) in fact $(v_1, \cdots v_n)$ is a basis of $\mathbb{R}^{n\times n}$.

Proof. $\underset{n\times n}{A}\underset{n\times n}{Q} = A(v_1, \cdots v_n) = \underset{n\times n}{(\lambda_1 v_1, \cdots \lambda_n v_n)} = \underset{n\times n}{(v_1, \cdots v_n)} \underset{n\times n}{\begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}} = Q\Lambda$.

~~$A = Q\Lambda Q^T$~~ Meanwhile, $Q^T Q = QQ^T = I$.

$\Rightarrow A = Q\Lambda Q^T$. □

$$= (v_1 \cdots v_n) \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix} \begin{pmatrix} v_1^T \\ \vdots \\ v_n^T \end{pmatrix} = \sum_i \lambda_i v_i v_i^T$$



$y = Ax = Q\Lambda Q^T x \Rightarrow Q^T y = \Lambda Q^T x$

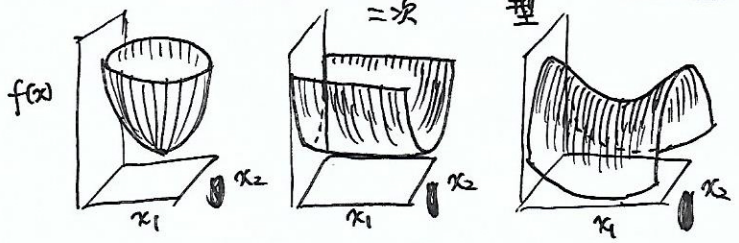
$\tilde{y} = \Lambda \tilde{x}$

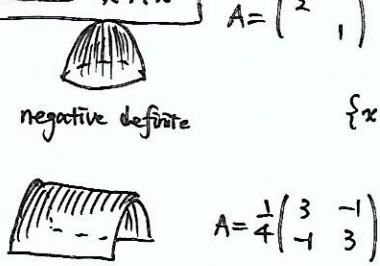
$\tilde{x} = Q^T x \Leftrightarrow x = Q\tilde{x}$ (since $QQ^T = I$)

$\tilde{x} = \tilde{x}_1 v_1 + \tilde{x}_2 v_2 + \cdots + \tilde{x}_n v_n$

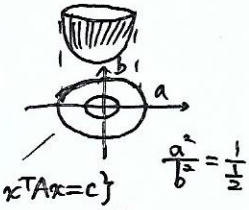
Def. Geometry of Quadratic Forms

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

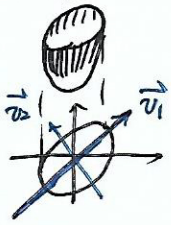




$A = \begin{pmatrix} \frac{1}{2} & \\ & 1 \end{pmatrix}$



$A = \frac{1}{4}\begin{pmatrix} 3 & -1 \\ -1 & 3 \end{pmatrix}$



Def. Bounds on Quadratic Forms

A: symmetric matrix

$$\lambda_{min} \|x\|_2^2 \leq x^T A x \leq \lambda_{max} \|x\|_2^2$$

Proof. 1) $A = \text{diag}(\lambda_1, \cdots \lambda_n)$. Obvious. $\lambda_{min}\|x\|_2^2 \leq \sum_i \lambda_i x_i^2 \leq \lambda_{max}\|x\|_2^2$

2) $A = Q\Lambda Q^T$ (since A is symmetric) ← eigendecomposition

$\lambda_{min}\|\tilde{x}\|_2^2 \leq x^T A x = x^T Q \Lambda Q^T x = \tilde{x}^T \Lambda \tilde{x}$ (Let $\tilde{x} = Q^T x$) $\leq \lambda_{max}\|\tilde{x}\|_2^2$

$\|\tilde{x}\|_2^2 = \tilde{x}^T\tilde{x} = (Q^T x)^T Q^T x = x^T \underbrace{QQ^T}_{I} x = x^T x = \|x\|_2^2$.

Qed. □

- Second-order necessary conditions

Thm. $f: \mathbb{R}^n \to \mathbb{R}$. twice continuously differentiable x^*: local minimum of f.

then its Hessian matrix $\nabla^2 f(x^*)$ is positive semidefinite. i.e. $d^T \nabla^2 f(x^*) d \geq 0$ $(\forall d \in \mathbb{R}^n)$

Proof: Taylor Expansion. $f(x^* + td) = f(x^*) + \frac{t^2}{2} d^T \nabla^2 f(x) d + o(t^2\|d\|^2) \geq f(x^*)$

$\Rightarrow \frac{1}{2} d^T \nabla^2 f(x^*) d + \frac{o(t^2\|d\|^2)}{t^2} \geq 0$. $t \to 0$: $d^T \nabla^2 f(x^*) d \geq 0$

$\Rightarrow g''(\text{local minimum}) \geq 0$.

- Second-order ~~necessary~~ sufficient condition

Thm. Suppose f is twice continuously differentiable, if

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

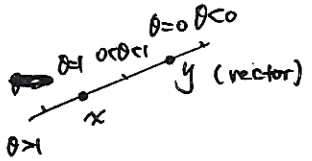
2) $\nabla^2 f(x^*)$ is positive definite.

then x^* is a local minimum.

Convex Functions and Convex Sets

- Convex Sets

Def. Lines, line segments and Rays



$z = y + \theta(x-y) = \theta x + (1-\theta)y. \quad \theta \in \mathbb{R}$

ray: endpoint y. $y \to x$ direction. $z = \theta x + (1-\theta)y. \quad \theta \geq 0$

segment: $z = \theta x + (1-\theta)y. \quad \theta \in [0,1]$

Sometimes use $\bar{\theta} := 1-\theta$.

Def. Convex Sets

A set $C \subset \mathbb{R}^n$ is • convex if the line segment between any two points $x, y \in C$ lies entirely in C, i.e.

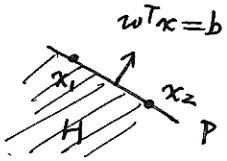
$$\forall x \in C,\ y \in C,\ \theta \in [0,1] \Rightarrow \theta x + \bar{\theta} y \in C$$

Here, $\theta x + \bar{\theta} y$ is called a convex combination.

In a more symmetric form: $\theta_1 x + \theta_2 y$ where $\theta_1 \geq 0,\ \theta_2 \geq 0,\ \theta_1 + \theta_2 = 1$.

e.g. $\emptyset$, $\mathbb{R}^n$, singletons (单个点) (points), lines, line segments and rays.

Hyperplane: $P = \{x \in \mathbb{R}^n : w^T x = b\}$



Proof. $x_1, x_2 \in P$. $\theta \in [0,1] \to \theta x_1 + \bar{\theta} x_2 \in P$.

To prove

i.e. to prove $w^T(\theta x_1 + \bar{\theta} x_2) \in P$

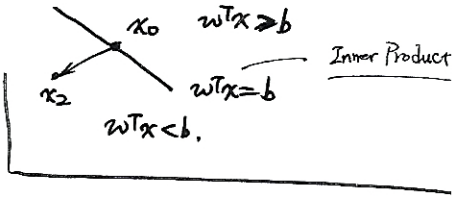
$= \theta \cdot (w^T x_1) + \bar{\theta} \cdot (w^T x_2)$

$= \theta b + \bar{\theta} b = b.$

Qed. □

Halfspace. $H = \{x \in \mathbb{R}^n : w^T x \leq b\}$.

$\{f(x) \leq 0\}$ — sublevelset



Proposition. The intersection of two arbitrary convex sets is convex.

$\{C_i\}_{i \in I}$ is a set of convex sets.

$C = \bigcap_{i \in I} C_i$ is also convex.

Easy to prove.

e.g. Affine spaces: $\{x \in \mathbb{R}^n : Ax = b\}$ is convex. $\longrightarrow$ Multi-hyperplanes' intersection. $(a_i^T x = b_i)$

$A \in \mathbb{R}^{m \times n}$

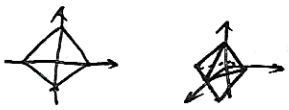
[eg.] polyhedron

Polyhedra: $\{x \in \mathbb{R}^n : Ax \leq b\}$ is convex. $\longrightarrow$ Multi-halfspaces' intersection.

Def. The inequality "$\leq$", "$\geq$" of vectors is interpreted component-wise.

Trick. $Ax=b \Leftrightarrow \begin{cases} Ax \geq b \\ Ax \leq b \end{cases} \Leftrightarrow \begin{pmatrix} A \\ -A \end{pmatrix} x \leq \begin{pmatrix} b \\ -b \end{pmatrix}$

(An affine space is in fact a polyhedron.)



1-norm balls $B_1^{(2)} = \{x \in \mathbb{R}^2 : |x_1|+|x_2| \leq a\}$ $B_1^{(3)} = \{x \in \mathbb{R}^3 : |x_1|+|x_2|+|x_3| \leq a\}$

four faces — four constrains; eight faces

Thm. A closed ball $\bar{B}(x_0, r) = \{x \in \mathbb{R}^n : \|x - x_0\| \leq r\}$ is convex.

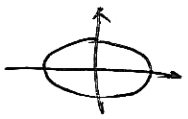
"Norm Ball"

Proof. $\forall x_1, x_2 \in \bar{B}(x_0, r)$ and $\forall \theta \in [0,1]$

$$\|(\theta x_1 + \bar{\theta} x_2) - x_0\| = \|\theta(x_1 - x_0) + \bar{\theta}(x_2 - x_0)\| \leq \theta\|x_1 - x_0\| + \bar{\theta}\|x_2 - x_0\| \leq \theta r + \bar{\theta} r = r.$$

True for any norm $\|\cdot\|$, also true for open ball

e.g. An ellipsoid $\left\{x \in \mathbb{R}^2 : \frac{x_1^2}{\lambda_1^2} + \frac{x_2^2}{\lambda_2} \leq 1\right\}$ is convex.



Proof. Transformation $\begin{cases} x_1 = \lambda_1 x_1^* \\ x_2 = \lambda_2 x_2^* \end{cases}$ $\mathcal{E} \longrightarrow$ ball (shifted ball) proved. Qed.

Formal Proof. Transformation: $u_1 = \frac{x_1}{\lambda_1}$, $u_2 = \frac{x_2}{\lambda_2}$. $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} \lambda_1 & \\ & \lambda_2 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \Lambda u$

[redacted], $\bar{B} = \{u : \|u\|_2 \leq 1\}$ is convex.

Now we prove $\mathcal{E} = \{\Lambda u : u \in \bar{B}\}$. is convex.

$\forall x_1, x_2 \in \mathcal{E}$. $\theta \in [0,1]$. $\exists\, u_1, u_2 \in \bar{B}$ s.t. $x_1 = \lambda_1 u_1 = \Lambda u_1$, $x_2 = \lambda_2 u_2 = \Lambda u_2$

Thus, $\theta x_1 + \bar{\theta} x_2 = \theta \lambda_1 u_1 + \bar{\theta} \lambda u_2 = \Lambda(\theta u_1 + \bar{\theta} u_2) \in \mathcal{E}$

($\theta u_1 + \bar{\theta} u_2 \in \bar{B}$)

e.g.'

$\mathcal{E}' = \{x_0 + Au : \|u\|_2 \leq 1\}$. $A \in \mathbb{R}^{n \times n}$, $A \succ 0$.

$A = Q \Lambda Q^T$. Let $\tilde{u} = Q^T u$. $\Rightarrow$ $\mathcal{E}' = \{x_0 + Q\Lambda\tilde{u} : \|\tilde{u}\|_2 \leq 1\}$

(shifted and rotated ball)

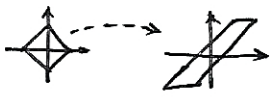
$\theta x_1 + \bar{\theta} x_2 = x_0 + A(\theta u_1 + \bar{\theta} u_2)$, with $\theta u_1 + \bar{\theta} u_2 \in \bar{B}$

Rewrite: $\mathcal{E}' = \{x : (x - x_0)^T P^{-1} (x - x_0) \leq 1\}$ where $P = A^2$, $P^{-1} = A^{-1}A^{-1}$

Proof. $x = x_0 + Au \Rightarrow u = A^{-1}(x - x_0)$. $u^T u \leq 1$ (ball) $\Rightarrow (x - x_0)^T A^{-1} A^{-1} (x - x_0) \leq 1$

$P = A^2$

Thm. The image of a convex set under an affine transformation is also convex.



Proof. Let $C \subset \mathbb{R}^n$ (convex set). $f(x) = Ax+b$ an affine transformation : $\mathbb{R}^n \to \mathbb{R}^m$.

$\forall\ y_1, y_2 \in f(C) := \{f(x) : x \in C\}$. $\theta \in [0,1]$. By definition. $\exists\ x_1, x_2 \in C$ s.t. $y_1 = f(x_1)$, $y_2 = f(x_2)$

Since y is affine, $\theta y_1 + \bar{\theta} y_2 = \theta(Ax_1+b) + \bar{\theta}(Ax_2+b) = A(\theta x_1 + \bar{\theta} x_2) + b$.

Since C is convex, $z := \theta x_1 + \bar{\theta} x_2 \in C$. Thus, $\theta y_1 + \bar{\theta} y_2 \in f(C)$ $\longrightarrow$ Qed. □

Proposition. The inverse image of a convex set under an affine transformation is convex.

逆 映射 f^{-1}

Proof.

Similarly.

e.g. The set $S_+^n = \{ A \in \mathbb{R}^{n\times n} : A \succeq 0 \}$. ↑ symmetric of course. $\longrightarrow$ $\theta A + \bar{\theta} B \in S_+^n$. $\theta \in [0,1]$.

{ Property I. easy to prove <symmetry>

Property II. $x^T(\theta A + \bar{\theta} B)x \geq 0$ $(\forall x)$

Def. A convex combination of $x_1, \cdots x_m \in \mathbb{R}^n$ is a point of the form

$$\sum_{i=1}^m \theta_i x_i = \theta_1 x_1 + \cdots + \theta_m x_m \quad \text{where } \theta_i \geq 0\ (\forall i) \text{ and } \sum_{i=1}^m \theta_i = 1.$$

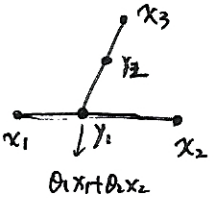
Thm. If C is convex, $x_1, \cdots x_m \in C$ then any convex combination $\sum_{i=1}^m \theta_i x_i \in C$.

Proof. By Induction. when $n=2$. Obvious

when $n=k$ the proposition holds

$n=k+1$: $\sum_{i=1}^{k+1} \theta_i x_i = \bar{\theta}_{k+1} \left(\sum_{i=1}^{k} \frac{\theta_i}{\bar{\theta}_{k+1}} x_i \right) + \theta_{k+1} x_{k+1}$

$\sum_{i=1}^{k} \frac{\theta_i}{\bar{\theta}_{k+1}} x_i \in C$



$\frac{\theta_1 x_1 + \theta_2 x_2}{\theta_1 + \theta_2 =: \bar{\theta}_2}$

(TBC)

Qed. □

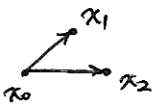
Def. The *convex hull* of a set $S \subset \mathbb{R}^n$ (denoted $\operatorname{conv} S$) is the smallest convex set containing S.

Thm.

$$\operatorname{conv} S = \left\{ \sum_{i=1}^m \theta_i x_i : m \in \mathbb{N};\ x_i \in S,\ \theta_i \geq 0,\ i = 1, \cdots, m;\ \sum_{i=1}^m \theta_i = 1 \right\}.$$

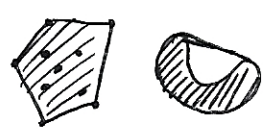
Def. Affinely independent points

$(m+1)$ points $x_0, \cdots x_m \in \mathbb{R}^n$ are *affinely independent* if $x_1 - x_0, \cdots x_m - x_0$ are linearly independent.

e.g. $\mathbb{R}^2$: 

(affinely independent).

e.g.



A :

$$\begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_0 & x_1 & \cdots & x_m \end{pmatrix} \begin{pmatrix} c_0 \\ \vdots \\ c_m \end{pmatrix} = 0.$$

↑ $\operatorname{rank} A = m+1$.

Proposition. $x_0, \cdots x_m$ are affinely independent iff. $\left\{ \begin{array}{l} \sum_{i=0}^m c_i x_i = 0 \\ \sum_{i=0}^m c_i = 0 \end{array} \right. \Rightarrow c_i = 0\ (\forall i \in \{1, 2, \cdots, m\})$

Proof is easy (definition.)

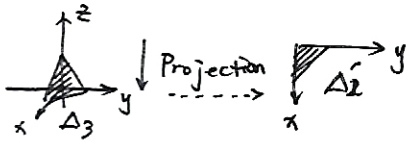
Def. An m-dimensional *simplex* $\Delta = \text{conv}\{x_0, \dots x_m\}$, where $x_0, \dots x_m$ are affinely independent. 单纯形 (also called m-simplex)

$\mathbb{R}^n$ only have m-simplexes with $m \le n$.

The probability n-simplex: $\Delta_n = \{\vec{\theta} \in \mathbb{R}^{n+1} : \vec{\theta} \ge \vec{0},\ 1^T\vec{\theta} = 1\}$ i.e. determined by std vectors $\vec{e}_1, \vec{e}_2, \dots \vec{e}_{n+1}$.

The unit n-simplex: $\Delta_n' = \{\vec{\theta} \in \mathbb{R}^n : \vec{\theta} \ge \vec{0},\ 1^T\vec{\theta} \le 1\}$

In fact: Δ_k' is the projection of Δ_{k+1}!

e.g. 

Relationship between Δ_m and general simplexes:

$$\vec{\theta} = \sum_{i=1}^{m} \theta_i \vec{e}_i \ \mapsto\ \vec{x} = \sum_{i=0}^{m} \theta_i \vec{x}_i = X\vec{\theta} \quad \text{where } X = (x_0, \dots x_m)_{n\times(m+1)},\ \vec{\theta} = (\theta_0, \dots \theta_m)^T \in \Delta_{m+1}$$

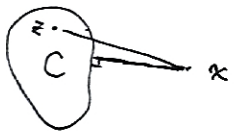
$$\vec{x} = \vec{x}_0 + \sum_{i=1}^{m} \theta_i(\vec{x}_i - \vec{x}_0) = \vec{x}_0 + B\vec{\theta}'. \text{ where } \vec{\theta}' \in \Delta_{m-1}'$$

(since $1^T\theta = 1$) $\quad B = (x_1 - x_0, \dots x_m - x_0)$

- **Projection onto Convex Sets**

Def. The Distance between a point x and C (a set)(nonempty) is

$$\text{dist}(x, C) := \inf_{z\in C} \|x - z\|$$



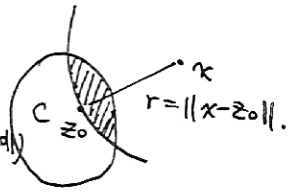
Thm. If $C \subset \mathbb{R}^n$ is nonempty, closed and convex, then for any x, exists a unique $\hat{x} \in C$ s.t. $\|x - \hat{x}\| = \text{dist}(x, C)$

$\hat{x}$ is called *the projection* of x onto C, denoted by $P_C(x)$.

Specially, $P_C(x) = x$ iff. $x \in C$. (obvious by the definition)

Proof. First show existence.

- Let $z_0 \in C$. $\text{dist}(x, C) \le \|z_0 - x\|$.
- Let $K = \{z \in C : \|x - z\| \le \|x - z_0\|\} = C \cap \bar{B}(x, \|x - z_0\|)$
- $\|x - z\|$ is continuous.



($f(z) = \|x - z\|$. $|f(z) - f(y)| = |\,\|x - z\| - \|x - y\|\,| \le \|z - y\|$). ($\to 0$, $\to 0$)

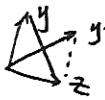
Also, K is compact $\Rightarrow \exists \hat{x} \in K$. s.t. $\text{dist}(x, C) = \|x - \hat{x}\|$.

Now show uniqueness. (by contradiction)

If exists $\hat{x}_1, \hat{x}_2$ s.t. $\|x - \hat{x}_1\| = \|x - \hat{x}_2\| = \text{dist}(x, C) =: d$.

$\hat{x}_c = \frac{\hat{x}_1 + \hat{x}_2}{2} \in C$. $\quad \|x - \hat{x}_c\| \le \|x - \hat{x}_1\|$

Let $y = x - \hat{x}_1$. $z = x - \hat{x}_2$. We have $\|y + z\|^2 + \|y - z\|^2 = 2\|y\|^2 + 2\|z\|^2$



(Since $\|y\pm z\|^2 = \langle y\pm z, y\pm z\rangle = \|y\|^2+\|z\|^2 \pm 2\langle y, z\rangle$)

Thus, we have, $\|2x-\hat{x}_1-\hat{x}_2\|^2+\|\hat{x}_1-\hat{x}_2\|^2 = 2\|x-\hat{x}_1\|^2+2\|x-\hat{x}_2\|^2 = 4d^2$

i.e. $0\le\|\hat{x}_1-\hat{x}_2\|^2 = 4d^2 - \|2(x-x_c)\|^2 = 4d^2-4\|x-x_c\|^2 \le 0$

$\therefore \|\hat{x}_1-\hat{x}_2\| = 0 \Rightarrow \hat{x}_1 = \hat{x}_2$

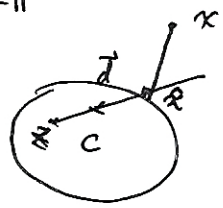
[Proposition]. Let $C\subset\mathbb{R}^n$ be nonempty, closed and convex. Given $\hat{x}\in C$.

$\hat{x} = P_C(x)$. iff. $\langle x-\hat{x}, z-\hat{x}\rangle \le 0$. $\forall x\in C$.

Proof. "$\Rightarrow$". Assume $\hat{x} = P_C(x) = \underset{z\in C}{\operatorname{argmin}}\ \|x-z\| = \underset{z\in C}{\operatorname{argmin}}\ \|x-z\|^2$

Let $f(z)=\|x-z\|^2$. $\forall z\in C$, $\vec{d} = z-\hat{x}$

$g(t) = \|x-(\hat{x}+td)\|^2$. $\nabla f^T d \ge 0$.



$f(z) = (x-z)^T(x-z)$. $\langle f'(\hat{x}), d\rangle = \langle \hat{x}-x, z-\hat{x}\rangle \ge 0$, i.e. $\langle x-\hat{x}, z-\hat{x}\rangle \le 0$.

"$\Leftarrow$". Assume $\langle x-\hat{x}, z-\hat{x}\rangle \le 0$ for any $z\in C$.

Then $\|x-z\|^2 = \|x-\hat{x}\|^2+\|z-\hat{x}\|^2-2\langle x-\hat{x}, z-\hat{x}\rangle \ge \|x-\hat{x}\|^2$. $\forall x\in C$.

By definition, $\hat{x} = P_C(x)$.

□

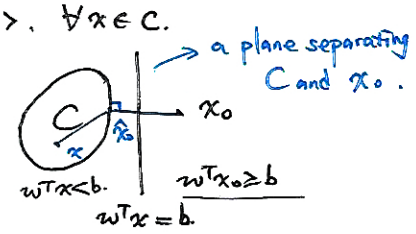
[Corollary]. Let $C\subset\mathbb{R}^n$ be nonempty, closed and convex. For $x_0\notin C$.

exists a $w\in\mathbb{R}^n\setminus\{0\}$. s.t. $\langle w, x\rangle < \langle w, x_0\rangle$. $\forall x\in C$.

Proof. Let $w = x_0 - P_C(x_0)$. Let $\hat{x}_0 = P_C(x_0)$.

Since $x_0\notin C$. $P_C(x_0)\in C$. $w\neq 0$.

By proposition above. $\langle \underbrace{x_0-\hat{x}_0}_{w}, x-\hat{x}_0\rangle \le 0$.

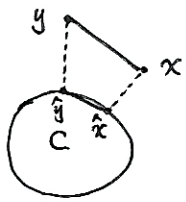


$\langle w, x\rangle \le \langle w, \hat{x}_0\rangle = \langle w, x_0\rangle - \langle w, w\rangle < \langle w, x_0\rangle$

$(w\neq 0.\ \langle w, w\rangle > 0)$.

□

[Corollary]. The projection operation is nonexpansive, i.e. $\|P_C(x)-P_C(y)\| \le \|x-y\|$.



Proof. Let $\hat{y} = P_C(y)$. $\hat{x} = P_C(x)$.

$\|x-y\|^2 = \|\hat{x}-\hat{y}\|^2 + \underbrace{\|x-y-(\hat{x}-\hat{y})\|^2}_{\ge 0} + 2\langle x-y-(\hat{x}-\hat{y}), \hat{x}-\hat{y}\rangle$

$\ge \|\hat{x}-\hat{y}\|^2 + 2\langle x-y, \hat{x}-\hat{y}\rangle - 2\langle \hat{x}-\hat{y}, \hat{x}-\hat{y}\rangle$

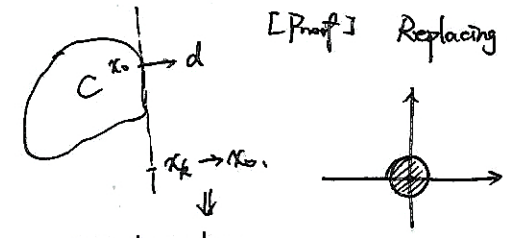
$\ge \|\hat{x}-\hat{y}\|^2$

□

Def. Supporting Hyperplane.

Proof. Lemma. If C is convex, $\operatorname{int} C$ and $\bar{C}$ are convex.

Lemma. If $C \subset \mathbb{R}^n$ is convex and $x_0 \in \partial C$. exists a direction d s.t. $x_0 + td \notin \bar{C}$ for any $t > 0$.



Supporting hyperplane

[Proof] Replacing C with $C - x_0$. Assume $x_0 = 0$. ($0 \in \bar{C}$. $0 \notin C$.) Prove the Lemma by contradiction.

1) Let $e_1, \cdots e_n$ be std basis vectors. $e_0 = -\sum_{i=1}^{n} e_i$

2) $\exists\, t > 0$ s.t. $d_i := t e_i \in \bar{C}$, $(\forall i)$

3) $(d_0\ d_1\ d_2)\begin{pmatrix}\theta_0\\ \theta_1\\ \theta_2\end{pmatrix} = x \in \Delta(d_0, d_1, d_2)$

At least $0 \in \Delta(d_0, d_1, d_2)$

4) Add a little "perturbance".

$$\begin{pmatrix} 1 & 1 & 1 \\ \underbrace{d_0+u_0}_{a_0} & \underbrace{d_1+u_1}_{a_1} & \underbrace{d_2+u_2}_{a_2} \end{pmatrix}\begin{pmatrix}\theta_0\\ \theta_1\\ \theta_2\end{pmatrix} = \begin{pmatrix}1\\ x\end{pmatrix}$$

$\det \begin{pmatrix} 1 & 1 & 1 \\ a_0 & a_1 & a_2 \end{pmatrix} \neq 0$

$$\Rightarrow \text{if } \begin{pmatrix}\theta_0\\ \theta_1\\ \theta_2\end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 \\ a_0 & a_1 & a_2 \end{pmatrix}^{-1}\begin{pmatrix}1\\ x\end{pmatrix} \geq 0. \text{ Then } x \in \Delta(a_0, a_1, a_2).$$

$\searrow \theta_i \geq 0\ (\forall i)$

Considering $\begin{pmatrix} 1 & 1 & 1 \\ d_0 & d_1 & d_2 \end{pmatrix}^{-1}\begin{pmatrix}1\\ 0\end{pmatrix} > 0$. when $x \to 0$. $u_i \to 0$. $\begin{pmatrix}\theta_0\\ \theta_1\\ \theta_2\end{pmatrix} > 0$. i.e. $x \in \Delta(a_0, a_1, a_2)$.

Thus, $x_0 \in \operatorname{int} C$. Contradiction. (Since $x_0 \in \partial C$)

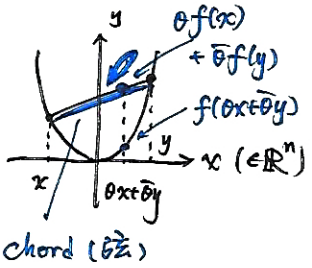
- Convex Functions

Def. A function $f: S \subset \mathbb{R}^n \to \mathbb{R}$ is convex if

1) its domain $\mathrm{dom} f = S$ is a convex set.

2) $\forall x, y \in S$. $\forall \theta \in [0,1]$ Jensen's Inequality holds,

i.e. $f(\theta x + \bar{\theta} y) \le \theta f(x) + \bar{\theta} f(y)$



Note. Condition 1 guarantees $\theta x + \bar{\theta} y \in \mathrm{dom} f$. Actually only need to check 2) for $x \neq y$ and $\theta \in (0,1)$.

Def. f is strictly convex if 1)

2)′ for any $x \neq y \in S$ and any $\theta \in (0,1)$,

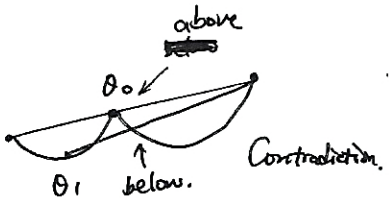
$$f(\theta x + \bar{\theta} y) < \theta f(x) + \bar{\theta} f(y).$$

[Proposition] Let f be convex.

If $f(\theta x + \bar{\theta} y) = \theta f(x) + \bar{\theta} f(y)$ for some $\theta_0 \in (0,1)$.

then $\forall \theta \in [0,1]$. $f(\theta x + \bar{\theta} y) = \theta f(x) + \bar{\theta} f(y)$

i.e. $g(\theta) = f(\theta x + \bar{\theta} y)$ is an affine function.



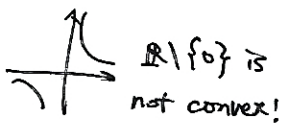
Def. f is (strictly) concave if $-f$ is (strictly) convex.

An affine function is both convex and concave, and is neither strictly convex nor strictly concave.

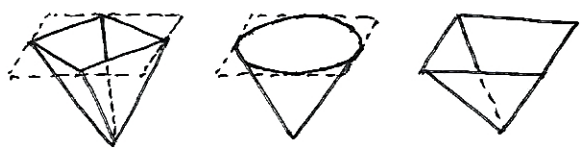
e.g. $f(x) = e^{ax}$ is convex, and strictly convex when $a \neq 0$.

$f(x) = \log_a x$ is $\begin{cases} \text{concave} & (a>0) \\ \text{convex} & (a<0). \end{cases}$

$f(x) = x^a$ is $\begin{cases} \text{convex on } \mathbb{R}^+ & \text{for } a \ge 1 \text{ or } a \le 0 \\ \text{concave on } \mathbb{R}^+ & \text{for } 0 \le a \le 1 \end{cases}$



$f(x) = \|x\|$ is convex.



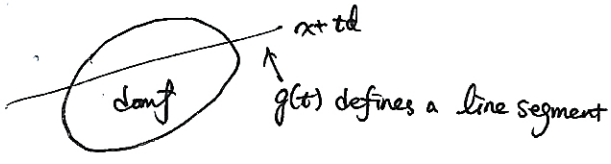
not strictly convex:

Let $x = 0$. $y \neq 0$

$\|\theta x + \bar{\theta} y\| = \|\bar{\theta} y\| = \bar{\theta}\|y\| = \theta\|x\| + \bar{\theta}\|y\|$ □

[Proposition] f is (strictly) convex iff. $\forall x \in \mathrm{dom} f$, any direction d. $(\neq 0)$

$g(t) = f(x + td)$ is (strictly) convex on $\mathrm{dom} g := \{t : x + td \in \mathrm{dom} f\}$



Proof: ⇐. $x, y \to$ find line segment $\to g(t^*)$. convex.

⇒. Obvious

Proof. $\Rightarrow$: $\forall t_1, t_2 \in \text{dom}\, g$, $\theta \in [0,1]$. $\bar{t} := \theta t_1 + \bar{\theta} t_2 \in \text{dom}\, g$

We have $x_1 = x + t_1 d \in \text{dom}\, f$. $x_2 = x + t_2 d \in \text{dom}\, f$. $\bar{x} = x + \bar{t} d \in \text{dom}\, f$

$(\because \bar{x} = x + (\theta t_1 + \bar{\theta} t_2) d = \theta \underbrace{(x + t_1 d)}_{\in \text{dom}\, f} + \bar{\theta} \underbrace{(x + t_2 d)}_{\in \text{dom}\, f})$

$$g(\bar{t}) = f(x + \bar{t} d) = f(\theta x_1 + \bar{\theta} x_2) \leq \theta f(x_1) + \bar{\theta} f(x_2) = \theta g(t_1) + \bar{\theta} g(t_2). \quad \square$$

$\Leftarrow$: Likewise.

Def. Extended-Value Extension

Given convex $f: S \subset \mathbb{R}^n \to \mathbb{R}^n$. its extended-value extension

$$\tilde{f}: \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}. \quad \tilde{f}(x) = \begin{cases} f(x). & x \in S \\ \infty & x \notin S. \end{cases}$$

with extended arithmetic and ordering: $\begin{cases} a + \infty = \infty & \text{for } a > -\infty \\ a \cdot \infty = \infty & \text{for } a > 0. \\ 0 \cdot \infty = 0 \end{cases}$

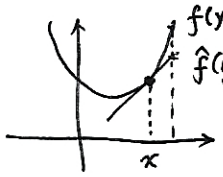
- Effective domain: $\text{dom}\, \tilde{f} = \text{dom}\, f = S$
- f is convex iff. $\tilde{f}$ is convex.

* We can also extend a concave function using $-\infty$.

- First-order condition for convexity

Thm. A differentiable f with an open convex domain $\text{dom}\, f$ is convex

iff. $f(y) \geq f(x) + \nabla f(x)^T (y - x)$. $\forall x, y \in \text{dom}\, f$



* Actually Taylor approximation underestimates a convex function.

e.g. $f(x) = e^x$.

→ To prove $\forall x \forall y$ $e^y \geq e^x + e^x (y - x)$ → $e^{y-x} \geq 1 + y - x$. i.e. $e^z \geq 1 + z$

$g(z) = e^z - z - 1 \geq g(0) = 0$ ($z \to \infty, -\infty$, $g(z) \to +\infty, +\infty$. coercive. global min exists).

$g'(z) = e^z - 1 = 0 \Rightarrow z = 0$

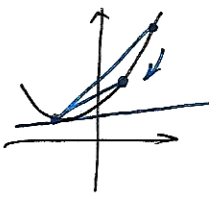
Proof. $\Rightarrow$. Assume f is convex. $d \triangleq y - x$. $\bar{t} := 1 - t$.

$$f(x + td) = f(x + t(y - x)) = f(ty + \bar{t}x) \leq t f(y) + \underbrace{\bar{t} f(x)}_{(1-t) f(x)} \quad t \in (0,1)$$

i.e. $f(x + td) - f(x) \leq t f(y) - t f(x)$

$$\Rightarrow \lim_{t \to 0} \frac{f(x + td) - f(x)}{t} \leq f(y) - f(x)$$

$\nabla f(x)^T d$. i.e. $\nabla f(x)^T (y - x) \leq f(y) - f(x)$.



$\frac{f(x + td) - f(x)}{t \|d\|}$ is actually the slope of the secant line through x and $x + td$.

$\Leftarrow$. Assume the first-order condition holds.

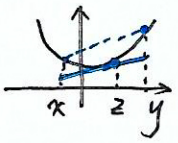
$z := \theta x + \bar{\theta} y$. then $f(y) \geq f(z) + \nabla f(z)^T (y - z)$ ①

$f(x) \geq f(z) + \nabla f(z)^T (x - z)$ ②

$\bar{\theta} \cdot ① + \theta \cdot ②$ yields $\theta f(x) + \bar{\theta} f(y) \geq f(z) + \nabla f(z)^T (\bar{\theta} y + \theta x) - \nabla f(z)^T z$

$= f(z) = f(\theta x + \bar{\theta} y)$

i.e. f is convex.



□

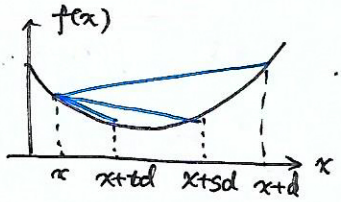
Thm. ..... $\mathrm{dom} f$ is strictly convex

iff. $f(y) > f(x) + \nabla f(x)^T (y - x)$. $\forall x \neq y \in \mathrm{dom} f$.

Proof. $\Leftarrow$: Likewise.

$\Rightarrow$: $d := y - x$.

$\forall 0 < t < s < 1$. $f(x + td) < \frac{t}{s} f(x + sd) + (1 - \frac{t}{s}) f(x)$ (strictly convex.)



$\Rightarrow \frac{f(x+td) - f(x)}{t} < \frac{f(x+sd) - f(x)}{s} < f(x+d) - f(x)$

$t \to 0$. yield $\nabla f(x)^T d \leq \frac{f(x+sd) - f(x)}{s} < f(y) - f(x)$.

□

Why we need "s"? To maintain "<".

[Corollary]. f: differentiable. $I \to \mathbb{R}$ defined on an open interval $I \subset \mathbb{R}$ is (strictly) convex iff. f' is (strictly) increasing on I.

Proof. $\Leftarrow$: $\forall x, y \in I$. $x < y$. By Mean Value Thm. can find $c \in (x, y)$ s.t. $f(y) - f(x) = f'(c)(y - x) \geq f'(x)(y - x)$ ≥ 0

$\therefore f$ convex (By First-order Condition)

$\Rightarrow$: $f(y) \geq f(x) + f'(x)(y - x)$. $f(x) \geq f(y) + f'(y)(x - y)$.

$f'(x) \leq \frac{f(y) - f(x)}{y - x} \leq f'(y)$.

□

Optimality of stationary points

[Corollary.] $\nabla f(x^*) = 0$ is a global minimum if f is convex.

unique global minimum — strictly convex

Proof. Obvious. $f(y) \geq f(x^*) + \underbrace{f'(x^*)(y - x)}_{=0} = f(x^*)$

global minimum

□

Concave → global maximum.

- Second-order Condition for Convexity

Thm. A twice (continuously) differentiable f with an open convex domain $\mathrm{dom} f$ is convex iff. $\nabla^2 f(x) \succeq 0$. at every $x \in \mathrm{dom} f$.

Proof. $\Rightarrow$: Assume f is convex,

$g(t) = f(x+td)$. is convex $\Rightarrow$ $g'(t)$. $\uparrow$ $\Rightarrow$ $g''(t) \geq 0$ for every $t \in$ domain,

i.e. $d^T \nabla^2 f(x) d = g''(0) \geq 0$. for any d $\Rightarrow$ $\nabla^2 f(x) \succeq 0$.

$\Leftarrow$ Also use $g(t)$. Likewise. □

* **Thm.** f is strictly convex if $\nabla^2 f(x) \succ 0$. Proof is Likewise.

e.g. $f(x) = x^4$ is strictly convex, while $f''(x) = 0$ @ $x = 0$.

* For Quadratic $f(x) = x^T Q x + b^T x + c$, it is necessary. can be proved.

e.g. $f(x) = \log\left(\sum_{i=1}^n e^{x_i}\right)$. $\longrightarrow$ Softmax. Smoothly approximates $\max_{1 \leq i \leq n} x_i$.

$$\log(e^{x_1} + e^{x_2}) = \log\left(e^{x_1}(1 + e^{x_2 - x_1})\right) = x_1 + \log(1 + e^{\Delta x}) \approx \begin{cases} 0 & x_1 >> x_2 \\ \Delta x. & x_1 << x_2 \end{cases}$$

$$= \begin{cases} x_1 & x_1 >> x_2 \\ x_2. & x_1 << x_2 \end{cases}$$

Show that it is convex.

Proof. $g(t) := f(\vec{x} + t\vec{d})$ ← univariant function

$$g(t) = \log\left(\sum_{i=1}^n e^{x_i + t d_i}\right) =: \log s(t)$$

$$g'(t) = \frac{s'(t)}{s(t)}. \qquad g''(t) = \frac{s''(t)s(t) - s'(t)^2}{s(t)^2}. \qquad g''(0) = d^T \nabla^2 f(x) d^T.$$

@ Suffices to show $g''(0) \geq 0$. for every $x, d \in \mathbb{R}^n$.

$$g''(0) = \frac{\sum_{i=1}^n e^{x_i} d_i^2}{\sum_{j=1}^n e^{x_j}} - \left(\frac{\sum_{i=1}^n e^{x_i} d_i}{\sum_{j=1}^n e^{x_j}}\right)^2 = \sum_{i=1}^n \frac{e^{x_i}}{\sum_{j=1}^n e^{x_j}} d_i^2 - \left(\sum_{i=1}^n \frac{e^{x_i}}{\sum_{j=1}^n e^{x_j}} d_i\right)^2$$

$$=: \sum_{i=1}^n p_i d_i^2 - \left(\sum_{i=1}^n p_i d_i\right)^2 \geq 0. \qquad \sum_{i=1}^n p_i = 1.$$

↑

$$f(\theta x + \bar{\theta} y) \leq \theta f(x) + \bar{\theta} f(y)$$

Generalization of Jensen's Inequality 下下页 → $f\left(\sum_{i=1}^n \theta_i x_i\right) \leq \sum_{i=1}^n \theta_i f(x_i)$. $\theta_i \geq 0$. $\sum_{i=1}^n \theta_i = 1$.

$h(x) := x^2$. Then $\sum_{i=1}^n h(d_i) p_i - h\left(\sum_{i=1}^n p_i d_i\right) \geq 0$. □

- Convexity-preserving Operations & Properties of Convex Functions

Thm. If x^* is a local minimum of a convex function f over S, then x^* is also a global minimum of f over S.

Proof. By Contradiction.

Suppose $\exists x_0 \in S$ s.t. $f(x_0) < f(x^*)$. $\forall \theta \in (0,1)$ $f(\theta x_0 + \bar{\theta} x^*) = \theta f(x_0) + \bar{\theta} f(x^*) < f(x^*)$.

$\|(\theta x_0 + \bar{\theta} x^*) - x^*\| = \|\theta(x_0 - x^*)\| = \theta \|x_0 - x^*\| \to 0$ when $\theta \to 0$.

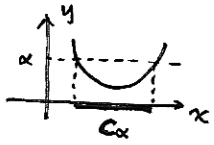
On the other side, since x^* is a local minimum. $\exists \delta > 0$. s.t. $\|x - x^*\| < \delta$. $f(x) \geq f(x^*)$.

Contradiction.

[Proposition]. strictly convex f $\longrightarrow$ unique global minimum. (By contradiction.)

Def. The α-sublevel set of a function f is: $C_\alpha = \{x \in \text{dom} f : f(x) \leq \alpha\}$.

Thm. Sublevel sets of a convex function are convex.



Similarly, the superlevel set $\{x \in \text{dom} f : f(x) \geq \alpha\}$ of a concave f is convex.

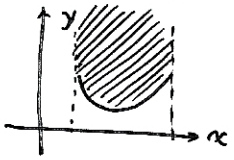
* Not-convex functions can have convex sublevel sets.

e.g. $f(x) = \sqrt{|x|}$. $f(x) = -e^x$

Def. The epigraph of $f: S \subset \mathbb{R}^n \to \mathbb{R}$ is: $\text{epi} f = \{(x,y) \in \mathbb{R}^{n+1} : x \in S,\ y \geq f(x)\}$.

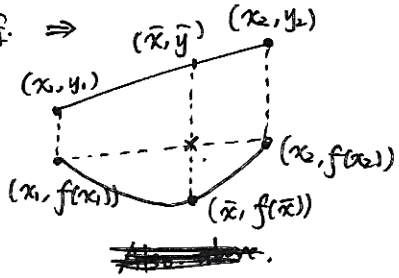
epi- : above

f and its extended-value extension $\tilde{f}$ have the same epigraph.



Thm. $f: S \subset \mathbb{R}^n \to \mathbb{R}$ is a convex function iff. epi$f$ is a convex set.

Proof. $\Rightarrow$



$(\bar{x}, \bar{y})$ is also above $(\bar{x}, f(\bar{x}))$.

$\Leftarrow$: Consider $(x_1, f(x_1))$. $(x_2, f(x_2))$

epi$f$ is convex: $\bar{x} \in \text{epi} f$.

Then $\theta f(x_1) + \bar{\theta} f(x_2) = \bar{y} \geq f(\bar{x})$.

□

* The projection of a convex set C ($\{x : (x,y) \in C \text{ for some } y\}$) is also convex.

Proof Likewise.

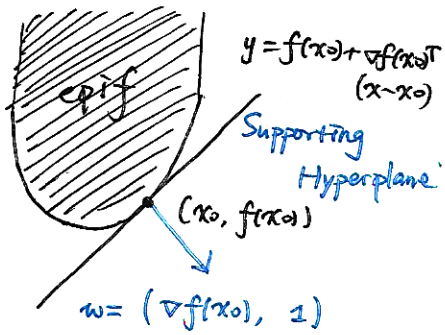
$(x,y) \in \text{epi} f$. $y \geq f(x) \geq f(x_0) + \nabla f(x_0)^T (x - x_0)$

i.e. $\nabla f(x_0)^T x - y \leq \nabla f(x_0)^T x_0 - f(x_0)$

i.e. $\begin{pmatrix} \nabla f(x_0) \\ -1 \end{pmatrix}^T \begin{pmatrix} x \\ y \end{pmatrix} \leq \begin{pmatrix} \nabla f(x_0) \\ -1 \end{pmatrix}^T \begin{pmatrix} x_0 \\ f(x_0) \end{pmatrix}$

$=: w$. A Hyper plane

$\forall (x,y) \in \text{epi} f$



Thm. Jensen's Inequality

For convex function f, $\forall x_i \in \mathrm{dom} f$. $\theta_i \geq 0$ s.t. $\sum_{i=1}^n \theta_i = 1$.

$$f\left(\sum_{i=1}^n \theta_i x_i\right) \leq \sum_{i=1}^n \theta_i f(x_i).$$

e.g. $f(x)=x^2 \Rightarrow \frac{1}{n}\sum_{i=1}^n x_i \leq \sqrt{\frac{1}{n}\sum_{i=1}^n x_i^2}$. e.g. $f(x)=\log x$ (concave) : $\frac{1}{n}\sum_{i=1}^n x_i \geq \sqrt[n]{\prod_{i=1}^n x_i}$

⇒ **Hölder's Inequality**

Let $p, q \in (1,+\infty)$ be conjugate exponents, i.e. $\frac{1}{p}+\frac{1}{q}=1$. $\sum_{i=1}^n |x_i y_i| \leq \|x\|_p \|y\|_q$.

$(x=(x_1,\cdots x_n)^T,\ y=(y_1,\cdots y_n)^T)$

$(\|x\|_p = \left(\sum_{i=1}^n |x_i|^p\right)^{1/p})$

Proof. $x^{\frac{1}{p}} y^{\frac{1}{q}} \leq \frac{1}{p}x+\frac{1}{q}y$. $(x,y\geq 0) \Leftrightarrow \frac{1}{p}\log x + \frac{1}{q}\log y \leq \log\left(\frac{1}{p}x+\frac{1}{q}y\right)$.

(since $\log x$ is concave.)

Thus, we have $\sum_{i=1}^n \left(\frac{|x_i|^p}{\sum_{j=1}^n |x_j|^p}\right)^{\frac{1}{p}} \left(\frac{|y_i|^q}{\sum_{j=1}^n |y_j|^q}\right)^{\frac{1}{q}} \leq \frac{1}{p}\frac{\sum_{i=1}^n |x_i|^p}{\sum_{j=1}^n |x_j|^p} + \frac{1}{q}\frac{\sum_{i=1}^n |y_i|^q}{\sum_{j=1}^n |y_j|^q}$

(p-norm, q-norm; the two fractions on the right equal 1)

i.e.

$$\frac{\sum_{i=1}^n |x_i||y_i|}{\|x\|_p \|y\|_q} \leq 1 \Rightarrow \sum_{i=1}^n |x_i y_i| \leq \|x\|_p \|y\|_q. \quad \square$$

⇒ **Minkowski's Inequality**

For $1<p<+\infty$. $\|x+y\|_p \leq \|x\|_p + \|y\|_p$.

Proof. $\|x+y\|_p^p = \sum_{i=1}^n |x_i+y_i|^{p-1}|x_i+y_i| \leq \sum_{i=1}^n |x_i+y_i|^{p-1}|x_i| + \sum_{i=1}^n |x_i+y_i|^{p-1}|y_i|$

Convexity-persering Operations:

$\sum_{i=1}^n c_i f_i(x)$ $(c_i \geq 0)$, $f(Ax+b)$ $(g(x))$, $h(f_1(x),\cdots f_n(x))$ $(h(x))$, $\sup_{i\in I} f_i(x)$, $\inf_{y\in C} f(x,y)$ are convex if f_i are convex.

e.g. $f(x_1,x_2) = (x_1-2x_2)^4 + 2e^{x_1-3x_2+5}$.

Let $g(y) = y_1^4 + 2e^{y_2}$. $y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} x_1-2x_2 \\ x_1-3x_2+5 \end{pmatrix} = \begin{pmatrix} 1 & -2 \\ 1 & -3 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} 0 \\ 5 \end{pmatrix}$.

f_1-f_2? $0-f$; x^4-x^2. f_1/f_2? $\frac{f}{-1}$; $\frac{1}{1+e^{+x}}$.

$f_1 \cdot f_2$? $-1\cdot f$; $x\cdot x^2$

$f(x) = h(g(x))$. $h: \mathbb{R} \to \mathbb{R}$. $g: \mathbb{R}^n \to \mathbb{R}$ Scalar Composition.

- f is convex if h is convex and increasing, g is convex
  if h is convex and ↘, g is concave
- f is concave if h is concave and ↗, g is concave
  if h is concave and ↘, g is convex

$h^* = h(-x)$ or $g^* = -g(+x)$ Transformed into Case 01.

Proof. (CASE 02). Consider $n=1$. $f''(x) = h''(g(x))\,[g'(x)]^2 + h'(g(x))\,g''(x) \geq 0$.

$h''(g(x)) \geq 0$ (convex); $[g'(x)]^2 \geq 0$; $h'(g(x)) \geq 0$ (increasing); $g''(x) \geq 0$

Formal Proof. $\forall x, y \in \mathbb{R}^n$. $\theta \in [0,1]$. Since g is convex. $g(\theta x + \bar{\theta} y) \leq \theta g(x) + \bar{\theta} g(y)$.

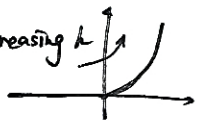
Since h is monotonic, (increasing) $h(g(\theta x + \bar{\theta} y)) \leq h(\theta g(x) + \bar{\theta} g(y)) \leq \theta h(g(x)) + \bar{\theta} h(g(y))$

$\because h$ is convex

Qed. □

e.g. To prove $f(x) = \|x\|^2$ is convex.

→ $g(x) = \|x\|$. $h(x) = \begin{cases} 0, & x < 0 \\ x^2, & x \geq 0 \end{cases}$ (Need to be monotonic)



Note. If ~~g~~ $\mathrm{dom} g$, $\mathrm{dom} h$ are not $\mathbb{R}/\mathbb{R}^n$. use extended-value extensions $\tilde{g}$, $\tilde{h}$ instead.

Vector Composition

Def. $h: \mathbb{R}^m \to \mathbb{R}$ is increasing if. $x \geq y$ (componentwise) $\Rightarrow h(x) \geq h(y)$.

decreasing ≤

Pointwise Maximum

If f_i is convex. $f := \max_{1 \leq i \leq m} f_i(x)$ is convex. $\mathrm{dom} f = \bigcap_{i=1}^m \mathrm{dom} f_i$

$f_i(x) < \infty \to \exists\, x \in \mathrm{dom} f_i$. $(\forall i)$

Pointwise Supreme: $\sup_{i \in I} f_i(x)$ I could be "$\mathbb{R}$"

e.g. $\sup_{x \in \mathbb{R}^n} \left(f(x) + \lambda g(x)\right) =: h(\lambda)$ (x: variable, λ: index)

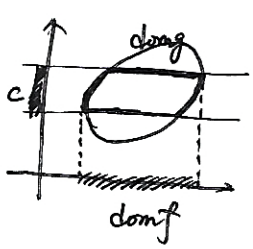
→ We prove $\mathrm{epi} f = \bigcap_i \mathrm{epi} f_i$ Intersection → Convex set. Prove. → f is Convex function.

Partial Minimization

here y is regarded as "i" (Index)

$f: \mathbb{R}^n \times \mathbb{R}^m \to (-\infty, \infty]$. is convex. $\emptyset = C \subset \mathbb{R}^m$ is convex.

then $f(x) = \inf_{y \in C} g(x,y)$ is convex, provided $f(x) > -\infty$ $(\forall x)$.



$\mathrm{dom} f := \{x \in \mathbb{R}^n : f(x) = \inf_{y \in C} g(x,y) < +\infty\} = \{x : \exists\, y \in C.\ \text{s.t. } g(x,y) < \infty\}$

$= \{x : \exists\, y \in C \text{ s.t. } (x,y) \in \mathrm{dom} g\} = \{x \in \mathbb{R}^n : (x,y) \in \mathrm{dom} g \cap \mathbb{R}^n \times C\}$

Proof. $\forall x_1, x_2 \in \text{dom} f$. $\theta \in [0,1]$

By definition,

$$f(\theta x_1 + \bar{\theta} x_2) = \inf_{y \in C} g(\theta x_1 + \bar{\theta} x_2, y) \leq g(\theta x_1 + \bar{\theta} x_2, \theta y_1 + \bar{\theta} y_2)$$

since $\theta y_1 + \bar{\theta} y_2 \in C$.

$$\leq \theta g(x_1, y_1) + \bar{\theta} g(x_2, y_2) < \theta(f(x_1) + \varepsilon) + \bar{\theta}(f(x_2) + \varepsilon)$$

convex

$g(x_i, y_i) < f(x_i) + \varepsilon$.

$$= \theta f(x_1) + \bar{\theta} f(x_2) + \varepsilon$$

(can find ε)

Convex Optimizations

Std Form for Optimization Prob.s

$$\min_{x \in \mathbb{R}^n} f(x) \quad \text{s.t.} \quad \begin{cases} h_i(x)=0 & (i=1,2,\cdots,k) \\ g_i(x)\le 0 & (i=1,2,\cdots,m) \end{cases} \quad (P)$$

Solution Set $X=\{x\in D: h_i(x)=0, g_j(x)\le 0\}$ ($i=\cdots$, $j=\cdots$)

$D = \mathrm{dom} f \cap \bigcap_{i=1}^{k} \mathrm{dom}\, h_i \cap \bigcap_{i=1}^{m} \mathrm{dom}\, g_i$

Def. Optimal Value (Generalized Definition)

The *optimal value* of Problem (P) is $f^* = \inf_{x\in X} f(x)$. allow f^* to take $\pm\infty$.

* $f^*=\infty$ if (P) is infeasible, i.e. $X=\emptyset$. $\sup\emptyset=-\infty$. $\inf\emptyset=+\infty$.
* $f^*=-\infty$: unbounded below. $\exists\{x_i\}\in X$ s.t. $f(x_i)\to -\infty$ as $i\to+\infty$
* x^* is an *optimal point* or *solves* (P). if $x^*\in X$ and $f(x^*)=f^*$.
  Note that f^* is not always attainable.
* x_0 is an *ε-suboptimal* if $x_0\in X$ and $f(x_0)\le f^*+\varepsilon$.
* locally optimal. omitted

Def. Convex Optimization Problem

(P) is a *convex optimization problem* if. (1) f, g_i are convex functions
(2) h_i are affine functions, i.e. $h_i(x)=a_i^T x-b_i$

* $\max_x f(x)$ where f is concave is also a convex optimization problem.
  ↳ $\min_x -f(x)$, $-f$ is convex.
  ↳ We transform the prob. into std form.

$D = \mathrm{dom} f \cap \bigcap_{i=1}^{m} \mathrm{dom}\, g_i$ (Since $\mathrm{dom}\, h_j = \mathbb{R}^n$)

$$X = \{x\in D: g_i(x)\le 0,\ 1\le i\le m;\ h_i(x)=0,\ 1\le i\le k\} = D \cap \bigcap_{i=1}^{m}\{x: g_i(x)\le 0\} \cap \bigcap_{i=1}^{k}\{x: h_i(x)=0\}$$

Both D and X are convex.

$$X_{opt} := \{x^*\in X: f(x^*)\le f(x),\ \forall x\in X\}.$$

sublevel set; hyperplane; polyhedron

- ▷ Any local minimum is a global minimum.
- ▷ If f is strictly convex. $|X_{opt}|\le 1$.

First-order Optimality Condition

Thm. a convex prob. with differentiable f. $x^*\in X$ is optimal iff. $\nabla f(x^*)^T(x-x^*)\ge 0$. $\forall x\in X$.

Proof. $\Leftarrow$: $f(x) \ge f(x^*)+\nabla f(x^*)^T(x-x^*) \ge f(x^*)$. $\forall x\in X$. ✓

(f is convex)

$\Rightarrow$: Assume x^* is optimal. $\alpha \in [0,1]$. $x^* + \alpha(x - x^*) = \alpha x + \bar{\alpha} x^* \in X$

$g(\alpha) := f(x^* + \alpha(x - x^*))$ $\quad d := x - x^*$ is a direction.

$g(\alpha) - g(0) \geq 0 \rightarrow \lim_{\alpha \to 0} \frac{g(\alpha) - g(0)}{\alpha} \geq 0$. i.e. $\nabla f(x^*)^T (x - x^*) \geq 0$

Corollary. $x^* \in \text{int } X$. then $\nabla f(x^*) = 0$

[Proof.] By Thm. $\nabla f(x^*)^T d \geq 0$. $\quad \nabla f(x^*)^T (-d) \geq 0$

$\Downarrow$

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

(Can choose $d = k \nabla f(x^*)$ $\quad \|k \nabla f(x^*)\| < \delta$.

$\Leftrightarrow \nabla f(x^*) \neq 0$ iff. $x^* \in \partial X$.

[Proof.] $w = \nabla f(x^*)^T$. $\quad w^T(x - x^*) \leq 0 \Rightarrow w^T x \leq w^T x^*$.

Thus. $w^T x = w^T x^*$ is a supporting hyperplane. of X @ x^*

Linear Program

Def. A *linear program* is an optimization problem of $\min_x c^T x$ s.t. $Bx \leq d$, $Ax = b$.

Std form: $\min_x c^T x$ s.t. $Ax = b$, $x \geq 0$.

Inequality form: $\min_x c^T x$ s.t. $Ax \leq b$

[Conversion] $Bx \leq d \xrightarrow{①}$ $\begin{cases} Bx + s = d. \\ s \geq 0. \end{cases}$ ② $x_i = x_i^+ - x_i^-$, $x_i^+, x_i^- \in \mathbb{R}_+$

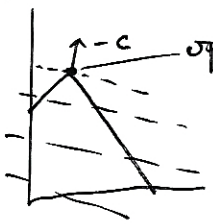
$x^+ = \begin{cases} x. & \text{if } x \geq 0 \\ 0. & \text{if } x < 0 \end{cases}$ $\quad x^- = \begin{cases} 0. & \text{if } x \geq 0 \\ -x. & \text{if } x < 0 \end{cases}$

$\Rightarrow$ std form.

[Conversion] Method 1). $Ax = b \Rightarrow \begin{pmatrix} A \\ -A \end{pmatrix} x \leq \begin{pmatrix} b \\ -b \end{pmatrix}$

Method 2). Solve one $x_i = f(x_1, \dots x_{i-1}, x_{i+1}, \dots x_m)$ $\quad$ fewer variables

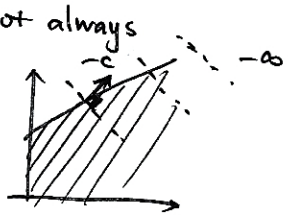
$\searrow \dots \rightarrow$ Inequality form.



optimal solution exists? not always

optimal solution unique?

not always (parallel)

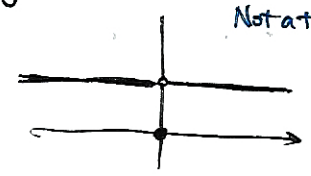


e.g. $y \in \mathbb{R}^n$. $X \in \mathbb{R}^{n \times p}$. rank $X = n < p$. Underdetermined Sys.: $Xw = y$. has infinitely many solutions

$\min_w \|w\|_0$ s.t. $Xw = y$. where l_0 "norm" $\|w\|_0 = \sum_{j=1}^p \mathbb{I}\{w_j \neq 0\}$

Not a true norm

not convex prob.! Why?



$\begin{cases} 1. & w_j \neq 0 \\ 0. & w_j = 0. \end{cases}$

The ℓ_1 approximation, called **basis pursuit** is convex.

$$\min_w \|w\|_1 = \sum_{j=1}^{p} |w_j| \quad \text{s.t.} \quad Xw = y. \qquad \text{(not an LP)}$$

Transform ① $t_j = |w_j|$: $\min_{t,w} \sum_{j=1}^{p} t_j$ s.t. $Xw = y$. $t_j = |w_j|$ (not affine!)

② $\mathbf{1}^T t = \min_{t,w} \sum_{j=1}^{p} t_j$ s.t. $Xw = y$, $t_j \geq |w_j|$.

Why equivalent? $\min_{t,w} f(t,w) = \min_{t} \min_{w} f(t,w) = \cdots\cdots$

$\Rightarrow \min_{t,w} \mathbf{1}^T t$ s.t. $Xw = y$. $-t \leq w \leq t$.

---

e.g. Piecewise Linear Minimization $\min_x f(x) = \max_{1 \leq i \leq m} a_i^T x + b$

$\Rightarrow \min_{x,t} t$ s.t. $t \geq a_i^T x + b$ $(1 \leq i \leq m)$

Quadratic Program

Def. $\min_x \frac{1}{2} x^T Q x + c^T x$ s.t. $Bx \leq d$. $Ax = b$ $(Q \succeq 0)$: QP

(affine)

convex when Q is positively semidefinite

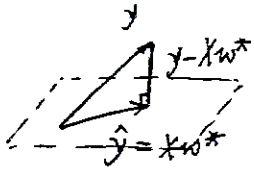
reduce to LP when $Q = 0$.

Def. $\min_x \frac{1}{2} x^T Q x + c^T x$ s.t. $\frac{1}{2} x^T Q_i x + c_i^T x + d_i \leq 0$. $Ax = b$: Quadratically Constrained QP (QCQP)

$(1 \leq i \leq m)$

convex iff. Q, Q_i are positively semidefinite. reduce to QP when all $Q_i = 0$.

e.g. Linear Least Squares $\min_w \|y - Xw\|_2^2$.



$f(w) = w^T X^T X w - 2 y^T X w + y^T y$.

$\nabla f(w) = 2 X^T X w - 2 X^T y = 0$ i.e. $X^T(Xw - y) = 0$

Let $Z = y - Xw$. $X^T Z = 0 \Rightarrow x_1, x_2, \cdots x_i \perp Z$

(perpendicular)

i.e. $y - Xw^*$ is perpendicular to the column space of X.

CASE 01. $\operatorname{rank} X = p$ (full column rank). $X^T X \succ 0$. unique solution $w^* = (X^T X)^{-1} X^T y$.

CASE 02. $\operatorname{rank} X = r < p$. $X = (X_1, X_2)$ $\hat{y} = Xw = (X_1, X_2)\begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = X_1 w_1 + X_2 w_2$, $w_2 = 0$.

↓

There is a solution with the last $(p-r)$ components being 0.

Thus, if w_1^* solves $\min_{w_1 \in \mathbb{R}^r} \|y - X_1 w_1\|$, $\begin{pmatrix} w_1^* \\ 0 \end{pmatrix}$ solves $\min_{w_1 \in \mathbb{R}^r} \|y - X_1 w_1\|$. $w_1^* = (X_1^T X_1)^{-1} X_1^T y$.

<not unique> < $\hat{y}$ is unique>

General Unconstrained QP

$$\min_x f(x) = \frac{1}{2}x^TQx + b^Tx + c \quad (Q \succeq 0) \quad \rightarrow \quad \nabla f(x) = Qx + b = 0,$$

① $Q \succ 0$. $x^* = Q^{-1}b$

② $\det Q = 0$. ● $b \in$ column space of Q. ~ infinitely many solution

③ $\det Q = 0$. $b \notin$ column space of Q. ~ no solution $f^* = -\infty$.

$Qx = -b$. $-b \in$ column space of Q $\Rightarrow b \in$ ——

e.g. $Q = \text{diag}(\lambda_1, \lambda_2, 0)$ $b_3 \neq 0$ $\rightarrow$ $f(x) = \frac{1}{2}\lambda_1 x_1^2 + \frac{1}{2}\lambda_2 x_2^2 + b_1x_1 + b_2x_2 + b_3x_3$ ← unconstrained

$x_3 \rightarrow -\infty$. $f(x) \rightarrow -\infty$.

Q is not diagonal. → use orthogonal matrix!

$Q = U\Lambda U^T$. $\Rightarrow y = U^Tx$ ($\downarrow$ $x = Uy$)

$\frac{1}{2}x^TQx = \frac{1}{2}x^TU\Lambda U^Tx = \frac{1}{2}y^T\Lambda y$

Thus, "$f(x)$" → "$g(y)$" $= \frac{1}{2}y^T\Lambda y + \underbrace{b^TU}_{\tilde{b}^T}y + c$.

$\det Q = 0 \iff \det \Lambda = 0$.

($(\det U)^2 \det \Lambda$; $\det U = 1$ (Since $U^TU = E$))

$Qx + b = 0 \iff \Lambda x + \tilde{b} = 0$

$U^TU\Lambda \underbrace{U^TU}_{E} y + U^Tb$

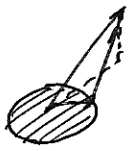
For ③. $\exists i_0$. $\lambda_{i_0} = 0$. $\tilde{b}_{i_0} \neq 0$ $\rightarrow$ $f(x) = (\Sigma \text{ rest}) + \tilde{b}_{i_0}y_{i_0}$ ($= g(y)$) $\rightarrow -\infty/+\infty$. $\rightarrow -\infty$

e.g. Lasso (Least Absolute Shrinkage and Selection Optimization)

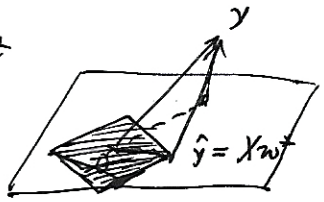
Given $y \in \mathbb{R}^n$. $X \in \mathbb{R}^{n\times p}$. $t > 0$.

$\min_w \|y - Xw\|_2^2$ s.t. $\|w\|_1 \leq t$

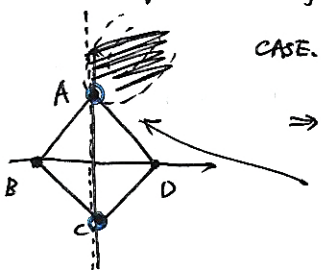
Ridge Regression



$\min_w \|y - Xw\|_2^2$ s.t. $\|w\|_2^2 \leq t$



unique solution if $n \geq p$ and $X^TX \succ 0$.



CASE 1. Constrain → "(x_0)"

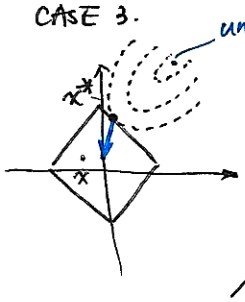
⇒ On boundary.

A/B/C/D/...

CASE 2. The global optimal of objective function is inside the constrained space

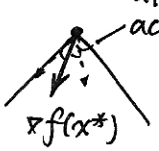
⇒ 约束无用.

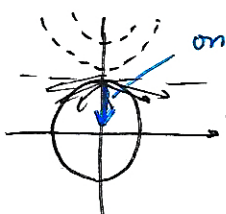
CASE 3.



$\langle \nabla f(x^*), x - x^* \rangle \geq 0$.

all: acute angle! (锐角!)





Ridge: More difficult to satisfy

Lasso: → easier to satisfy

Geometric Program

Def. A monomial : $f(x) = \gamma x_1^{a_1} x_2^{a_2} \cdots x_n^{a_n}$ $(\gamma > 0,\ x \in \mathbb{R}^n_{++},\ a_i \in \mathbb{R})$

A posynomial 正项式 $f(x) = \sum_{k=1}^{p} \gamma_k x_1^{a_{k1}} x_2^{a_{k2}} \cdots x_n^{a_{kn}}$

Geometric Program : $\min_x f(x)$ s.t. $g_i(x) \le 1,\ \eta_j x_1^{c_{j1}} x_2^{c_{j2}} \cdots x_n^{c_{jn}} = 1.$

$\Leftrightarrow$ $\min \log(\sum e^{\square})$ s.t. $\log(\sum e^{\square}) \le 1.$. affine.

"$x_k \to e^{u_k}$"

Gradient Descent

Consider $\min_x f(x)$ • where f is convex and differentiable on $\mathbb{R}^n$.

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

Descent Method

choose $x_0 \in \mathbb{R}^n$

repeat

descent direction $d_k \in \mathbb{R}^n$. step size $t_k > 0$

$x_{k+1} \leftarrow x_k + t_k d_k$ s.t. $f(x_{k+1}) < f(x_k)$

until (stopping criterion is satisfied.)

1. descent direction.

if $g(t) := f(x_k + t d_k) < f(x_k) = g(0)$ for all small enough t. we call d_k a ~.

- d_k is a descent direction $\Rightarrow g'(0) = d_k^T \nabla f(x_k) \le 0$
- $g'(0) = d_k^T \nabla f(x_k) < 0 \Rightarrow d_k$ is a descent direction.

Choose $d_k = -\nabla f(x_k)$. (fastest rate of descent) $d_k^T \nabla f(x_k) = -\|\nabla f(x_k)\|_2^2 < 0$. unless $\nabla f(x_k) = 0$.

for convex: optimum!

i.e. $x_{k+1} \leftarrow x_k - t_k \nabla f(x_k)$.

$x \leftarrow x_0 \in \mathbb{R}^n$

while $\|\nabla f(x)\| < \delta$. do

$x \leftarrow x - t \nabla f(x)$

end while

return x

practical then $\nabla f(x) = 0$

also: $\|f(x_{new}) - f(x_{old})\| \le \delta$

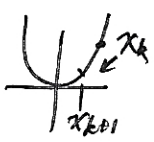
$\frac{\|f(x_{new}) - f(x_{old})\|}{\|f(x_{old})\|} < \delta$, etc.

How to choose a decent "t" (step)?

e.g. $f(x) = \frac{1}{2} a x^2$ $(a > 0)$ $\Rightarrow$ $x_{k+1} = x_k - t f'(x_k) = (1 - at) x_k$

$|1 - at| < 1$

$(1 - at)^k x_k \to 0$



Note: f satisfies $|f'(x)-f'(y)| = a|x-y| \rightarrow$ Lipschitz continuity (f').

$f''(x) = a$ ← Lipschitz constant

Def. Lipschitz continuity: $\|f(x)-f(y)\| \le L\|x-y\| \quad \forall x \forall y$ "L-Lipschitz"

↑ norm usually 2-norm.

⇓ implies

Uniform continuity 一致连续 $\forall x.\ \exists\, \delta(\varepsilon).$ s.t. $\|f(x)-f(y)\|<\varepsilon.\ \forall y$ s.t. $\|x-y\|<\delta.$

e.g. $f(x) = a^T x$ is $\|a\|$-Lipschitz. $|a^Tx - a^Ty| = |a^T(x-y)| \le \|a\|\|x-y\|$

Cauchy-Schwarz Inequality

$f(x) = \frac{1}{2}|a|x^2$ is $|a|$-Lipschitz.

Ex. $f(x) = Qx$ with $Q \succeq 0$. is $\lambda_{max}(Q)$-Lipschitz.

Proof. $f(x)-f(y) = Q(x-y) =: Qd. \quad (d := x-y)$

$\|f(x)-f(y)\|^2 = (Qd)^T(Qd) = d^T Q^T Q d \le \lambda_{max}(Q^TQ)\|d\|^2$

$Q \succeq 0 \rightarrow Q^T = Q$ (symmetric). $Qx = \lambda x \rightarrow Q^2x = \lambda^2 x$

$= (\lambda_{max}(Q)\|d\|)^2$ ← Lipschitz constant

For general cases, $f(x) = Qx$ is $\sqrt{\lambda_{max}(Q^TQ)}$-Lipschitz.

Def. L-smoothness: ∇f is L-Lipschitz. ← upper bound the rate of change of ∇f.

e.g. for convex (twice cont. diff.) $f: \mathbb{R}^n \to \mathbb{R}^1$. L-smooth iff. $\nabla^2 f(x) \preccurlyeq LE$.

↑ diag(1, ..., 1)

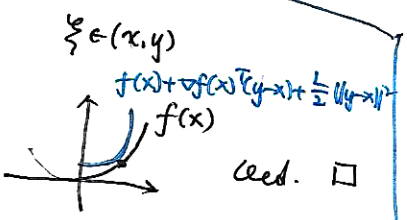
f is L-smooth $\Rightarrow f(y) \le f(x) + \nabla f(x)^T(y-x) + \frac{L}{2}\|y-x\|^2.$

Proof. $f(y) = f(x) + \nabla f(x)^T(y-x) + \frac{1}{2}(y-x)^T\nabla^2 f(\xi)(y-x)$

$\le f(x) + \nabla f(x)^T(y-x) + \frac{1}{2}\lambda_{max}(\nabla^2 f(\xi))\|y-x\|^2$

$\le f(x) + \nabla f(x)^T(y-x) + \frac{L}{2}\|y-x\|^2.$

$\xi \in (x, y)$



Qed. □

Another proof. Prove 1D case.

$g(t) := f(x + t(y-x))$. $g(1) = g(0) + g'(0) + \frac{1}{2}Lg.$

$|g'(t) - g'(s)| = |\nabla f(x+ty-x)^T(y-x) - \nabla f(x+s(y-x))^T(y-x)|$

$\le L\|y-x\|\|y-x\||t-s| \rightarrow Lg = L\|y-x\|^2.$

Back to Gradient Descent: L-smooth f.

$\{x_k\}$: $f(x_{k+1}) \le f(x_k) - t\left(1 - \frac{Lt}{2}\right)\|\nabla f(x_k)\|^2$

$(x_{k+1} = x_k - t\nabla f(x_k)$ ↗ $\Rightarrow f(x_k) - f(x_{k+1}) \ge t\left(1-\frac{Lt}{2}\right)\|\nabla f(x_k)\|^2$

As long as $0 < t \le \frac{1}{L}$. $f(x_k) - f(x_{k+1}) \ge \frac{t}{2}\|\nabla f(x_k)\|^2 \ge 0$

↓

always descent! → which is good

f: convex & L-smooth. x^*: a minimum of f. for step size $t \in (0, \frac{1}{L}]$

$\{x_k\}$ produced by gradient descent satisfies

$$f(x_k) - f(x^*) \leq \frac{\|x_0 - x^*\|^2}{2tk}.$$

Note.

- $f(x_k) \downarrow f^*$ as $k \to +\infty$.
- $\lim_{k\to+\infty} x_k =: \hat{x}_k$. $\hat{x}_k$ is a optimal solution.
- Problem: $f(x_k) - f(x^*) \leq \frac{c}{k}$. difference $< \varepsilon \Rightarrow k > \frac{1}{\varepsilon}$. k="10P" !! (iterations)
  ($\varepsilon \downarrow 10^{-P}$)
- Faster convergence?

→ Extend to: $t \in (0, \frac{2}{L})$. $f(x_k) - f(x_{k+1}) \geq t\|\nabla f(x_k)\|^2 \geq 0$.

$f(x_k) - f(x^*) < \frac{\|x_0 - x^*\|^2}{tk}$. not very helpful.

- $\|x_0 - x^*\|$ smaller. might help.

Proof. $\|x_{k+1} - x^*\|^2 = \|(x_k - x^*) - t\nabla f(x^*)\|^2 = \|x_k - x^*\|^2 + t^2\|\nabla f(x^*)\|^2 - 2t\nabla f(x^*)^T(x_k - x^*)$

$f(x_k) - f(x_{k+1}) \geq \frac{t}{2}\|\nabla f(x_k)\|^2$.

First-order condition of Convexity: $f(x^*) - f(x_k) \geq \nabla f(x_k)^T(x^* - x_k)$

Thus, $\|x_{k+1} - x^*\|^2 \leq \|x_k - x^*\|^2 + 2t[f(x_k) - f(x_{k+1})] + 2t[f(x^*) - f(x_k)]$

$= \|x_k - x^*\|^2 + 2t[f(x^*) - f(x_{k+1})]$

$$\therefore \sum_{k=0}^{N-1} f(x_{k+1}) - f(x^*) \leq \sum_{k=0}^{N-1} \frac{1}{2t}\left(\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2\right)$$

i.e. $(N\cdot)\left(f(x_N) - f(x^*)\right) \leq \frac{1}{2t}\left(\|x_0 - x^*\|^2 - \|x_N - x^*\|^2\right) \leq \frac{1}{2t}\|x_0 - x^*\|^2$

i.e. $f(x_N) - f(x^*) \leq \frac{1}{2Nt}\|x_0 - x^*\|^2$.

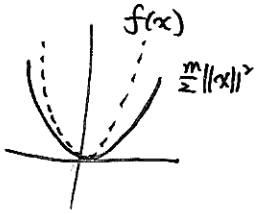
Qed. □

Faster Convergence? — Strong Convexity

Def. Strong Convexity.

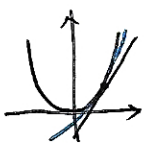
$f(x)$ is strongly convex with $m>0$ (m-strongly convex) if

$\tilde{f}(x) = f(x) - \frac{m}{2}\|x\|^2$ is convex.



e.g. x^4 is not strongly convex (for any m)

→ Def. First-order Condition for Strong Convexity



f is m-strongly convex iff. $f(y) \geq f(x) + \nabla f(x)^T(y - x) + \frac{m}{2}\|x - y\|^2 \quad \forall x \forall y$

strong convexity ⇒ strict convexity ⇒ convexity

* m-strong convexity and L-smoothness together imply $\frac{m}{2}\|x-y\|^2 \leq f(y) - f(x) - \nabla f(x)^T(y-x) \leq \frac{L}{2}\|x-y\|^2$

First-order condition for strong convexity.

Proof. $\Rightarrow$: $\tilde{f}(x) = f(x) - \frac{m}{2}\|x\|^2$ convex. $\Rightarrow$ $f(x) = \tilde{f}(x) + \frac{m}{2}\|x\|^2$

$$\tilde{f}(y) \geq \tilde{f}(x) + \langle \nabla \tilde{f}(x), y-x \rangle$$

$$\|y\|^2 = \|y-x\|^2 + \|x\|^2 + 2\langle x, y-x \rangle$$

Thus, $f(y) = \tilde{f}(y) + \frac{m}{2}\|y\|^2 \geq \tilde{f}(x) + \frac{m}{2}\|x\|^2 + \frac{m}{2}\|y-x\|^2 + \langle \nabla \tilde{f}(x) + mx, y-x \rangle$

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

$\Leftarrow$: Obvious.

□

Second-order condition for Strong Convexity

Thm. twice continuously differentiable f. f is m-strongly convex iff. $\nabla^2 f(x) \succeq mE$. $\forall x$.

iff. $\lambda_{min}(\nabla^2 f) \geq m$, $\forall x$

Proof. $\tilde{f}(x) = f(x) - \frac{m}{2}\|x\|^2$ is convex iff $\nabla^2 \tilde{f} = \nabla^2 f - mE \succeq 0$.

□

$\rightarrow$ $\lambda_{min} \leq \lambda \leq \lambda_{max}$

- λ_{min}: restrict ↑ Strong Convexity
- λ: eigen-value
- λ_{max}: restrict ↑ L-smoothness

Convergence Restricted by Smoothness and Strong Convexity

Thm. f: m-strongly convex, L-smooth. x^*: minimum of f

Step size $t \in (0, \frac{1}{L})$. $\{x_k\}$ produced by G.D. satisfies:

- $f(x_k) - f(x^*) \leq \frac{L(1-mt)^k}{2}\|x_0 - x^*\|^2$
- $\|x_k - x^*\|^2 \leq (1-mt)^k \|x_0 - x^*\|^2$.

Note: $f(x_k) \rightarrow f(x^*)$ exponentially fast.

$f(x_k) - f(x^*) \leq \varepsilon$. $\rightarrow$ $O(\log \frac{1}{\varepsilon})$. faster convergence.

$\nabla f(x^*) = 0$ $\Rightarrow$ $\frac{m}{2}\|x_k - x^*\|^2 \leq f(x_k) - f(x^*) \leq \frac{L}{2}\|x_k - x^*\|^2$.

(First-order condition)

Proof. $\|x_{k+1} - x^*\|^2 = \|x_k - t\nabla f(x_k) - x^*\|^2 = \|x_k - x^*\|^2 + t^2\|\nabla f(x_k)\|^2 + \underbrace{2t\nabla f(x_k)^T(x^* - x_k)}_{\leq 0}$

By L-smoothness: $t^2\|\nabla f(x_k)\|^2 \leq 2t\left(f(x_k) - f(x_{k+1})\right)$.

By m-strong convexity: $\nabla f(x_k)^T(x^* - x_k) \leq f(x^*) - f(x_k) - \frac{m}{2}\|x_k - x^*\|^2$

$\Rightarrow \|x_{k+1} - x^*\|^2 \leq (1-mt)\|x_k - x^*\|^2 + 2t[f(x^*) - f(x_{k+1})]$

$f(x^*) < f(x_{k+1})$ $\Rightarrow$ $\|x_{k+1} - x^*\|^2 \leq (1-mt)\|x_k - x^*\|^2$

Thus, $\|x_k - x^*\|^2 \leq (1-mt)^k \|x_0 - x^*\|^2$.

Note. L-smoothness $\Rightarrow f(x_k) - f(x^*) \le \frac{\|x_0 - x^*\|^2}{2tk}$. $t\uparrow$. faster.

converge $\rightarrow$ $0 < t \le \frac{1}{L}$. $\rightarrow$ $t_{max} = \frac{1}{L}$. fastest descanting rate.

Def. Condition number.

$Q \in \mathbb{R}^{n\times n}$. $Q \succ 0$. $\kappa(Q) = \frac{\lambda_{max}(Q)}{\lambda_{min}(Q)}$. (Kappa)

$\kappa(Q) = 1$.

well-conditioned

$\kappa(Q) = 10$.

ill-conditioned

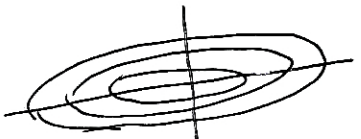
$m \le \lambda \le L$. $f(x_k)$ dominates the slower one in $(1-mt)^{2k}$ and $(1-Lt)^{2k}$.

Convergence Analysis.

$\frac{L}{2}(1-\frac{m}{L})^k \|x_0 - x^*\|^2 < \varepsilon \Rightarrow (1-\frac{m}{L})^k = \frac{\varepsilon}{c}$. $1 - \frac{m}{L} \approx e^{-\frac{m}{L}}$ (if $\frac{m}{L}$ is small)

Thus, $k \approx \frac{L}{m} \log\frac{c}{\varepsilon} \ge \kappa(Q) \log\frac{c}{\varepsilon}$.

$\kappa(Q) \le \frac{L}{m}$



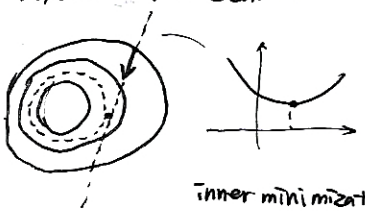
slower

faster $\leftarrow$ use a constant t. $\rightarrow$ not ideal.

Step Size. $x_{k+1} \leftarrow x_k - t_k \nabla f(x_k)$

- constant step size : t
- exact line search: $t_k = \underset{s}{\operatorname{argmin}}\ f(x_k - s\nabla f(x_k))$
- backtracking line search (Armijo's rule) : t_k satisfies $f(x_k) - f(x_k - t_k \nabla f(x_k)) \ge \alpha t_k \|\nabla f(x_k)\|_2^2$.

Exact Line Search



Impractical

used only if the inner minimization is easy.

(t_k)

Note. Can prove that

$0 = h'(t_k) = \nabla\left(f(x_k - t_k \nabla f(x_k))\right)^T \nabla f(x_k)$

$= \nabla f(x_{k+1})^T \nabla f(x_k)$

相邻两次 x_k 下降方向必垂直.

Thm. exact line search. $f(x_k) - f(x^*) \le \left(1 - \frac{m}{L}\right)^k [f(x_0) - f(x^*)]$.

Proof. L-smoothness: $f(x_k - t\nabla f(x_k)) - f(x_k) \le -t\left(1 - \frac{Lt}{2}\right)\|\nabla f(x_k)\|^2$

$\downarrow \min_t$ $\qquad$ $\downarrow \min_t$

$f(x_{k+1}) - f(x_k) \le \frac{-1}{2L}\|\nabla f(x_k)\|^2$

m-strong convexity: $f(x) \ge f(x_k) + \nabla f(x_k)^T(x - x_k) + \frac{m}{2}\|x - x_k\|^2$.

$\min_x \downarrow$ $\qquad$ $\downarrow \min_x$

$f(x^*) \ge f(x_k) - \frac{1}{2m}\|\nabla f(x_k)\|^2$.

Thus, $f(x_{k+1}) - f(x^*) \le \left(1 - \frac{m}{L}\right)\left[f(x_k) - f(x^*)\right]$. □