Classification of Matrix

A

- 2×2: **det, tr**
  - $\det > 0,\ tr > 0 \to$ pos. def.
  - $\det \geq 0,\ tr \geq 0 \to$ pos semi-def.
  - $\det > 0,\ tr < 0 \to$ neg. def.
  - $\det \geq 0,\ tr \leq 0 \to$ neg. semi-def.
- $n\times n$, $n \geq 3$: **Diagonal Dominance**

  每行其余元素绝对值的和小于diagonal的

  - strict $d_i > 0 \to$ pos. def.
  - strict $d_i < 0 \to$ neg. def.
  - Otherwise $\to$ **Principal Minor**

    $\left(D_1(A) = a_{11},\ D_2(A) = \det\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \cdots \right)$

    - $D_k(A) > 0\ \forall k. \Leftrightarrow$ pos. def.
    - $(-1)^k D_k(A) < 0\ \forall k \Leftrightarrow$ neg. def.
    - otherwise $\to$ **Try some vectors**
      - $x^T A x > 0,\ y^T A y < 0 \to$ indefinite
      - Otherwise $\to$ **Eigenvalues**
        - pos. semi-def. $\Leftrightarrow \lambda_i \geq 0$
        - pos. def. $\Leftrightarrow \lambda_i > 0$
        - neg. def. $\Leftrightarrow \lambda_i < 0$
        - neg. semi-def. $\Leftrightarrow \lambda_i \leq 0$

Optimality Condition

Local

local optimum $\Rightarrow$ ($\frac{\partial f}{\partial x_i}$ all exists)

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

Stationary Point (twice-diff.)

- $\nabla^2 f(x^*) \succ 0 \to$ strict local min
- $\nabla^2 f(x^*) \prec 0 \to$ strict local max
- $\nabla^2 f(x^*)$ indef. $\to$ saddle
- Otherwise $\to$ check by definition: local min / local max / saddle

  or show $\nabla f(x)$ is + on one side

  − on the other side

  - $- \Rightarrow +$ local min
  - $+ \Rightarrow -$ local maxi

② complete square!

① find path $x^* + t\vec{a},\ t\in\mathbb{R}$

on which x^* is minima/maxima of f

Global Coercive + domain closed $\longrightarrow$ $\exists$ global min

Special cases $\begin{cases} f \text{ convex, domain } C \text{ convex} \\ \forall x,\ \nabla^2 f(x) \succcurlyeq 0 \ (\cdot \nabla^2 f(x) \preccurlyeq 0) \\ \qquad \text{Stationary} \Rightarrow \text{global optimality} \\ \text{Weierstrass': } f \text{ continuous on non-empty compact } C \\ \qquad \Rightarrow \exists \text{ global min \& global max} \end{cases}$

Otherwise, check definition $\forall x,\ f(x) \geq f(x^*)$

e.g. via completing square

Convex Set

Def. $\lambda \underset{\sim}{x} + (1-\lambda)\underset{\sim}{y} \in C \quad \forall \underset{\sim}{x}, \underset{\sim}{y} \in C, \ \lambda \in [0,1]$

- $\{C_i\}_{i\in I}$ set of convex sets $\Rightarrow \bigcap_{i\in I} C_i$ convex
- $\mu_i \in \mathbb{R}$, C_i convex sets $\Rightarrow \mu_1 C_1 + \cdots + \mu_k C_k$ convex
- C_i convex $\Rightarrow C_1 \times C_2 \times \cdots \times C_m$ convex
- M convex, $A \in \mathbb{R}^{m\times n} \Rightarrow A(M) := \{Ax : x \in M\}$ convex
  $A^{-1}(M) := \{y : Ay \in M\}$ convex

Examples
- line $L = \{\underset{\sim}{z} + t\underset{\sim}{d} : t \in \mathbb{R}\}$
- hyperplane $H = \{x : \underset{\sim}{a}^T \underset{\sim}{x} = b\}$ $\underset{\sim}{a} \neq 0$
- half space $H^- = \{x : a^T x \leq b\}$
- open ball closed ba ,
- ellipsoid $E = \{x : \underset{\sim}{x}^T Q \underset{\sim}{x} + 2\underset{\sim}{b}^T x + c \leq 0\}$
  where $Q \succeq 0$ $\underset{\sim}{b} \in \mathbb{R}^n$,

Convex Hull
$$\mathrm{conv}(S) = \left\{ \sum_{i=1}^{k} \lambda_i x_i : x_i \in S, \underset{\sim}{\lambda} \in \Delta_k \right\}$$

Caratheodory Theorem If $x \in \mathrm{conv}(S)$ $(S \subseteq \mathbb{R}^n)$

$\exists\, x_1, \cdots, x_{n+1} \in S$ s.t. $x \in \mathrm{conv}(\{x_1, \cdots, x_{n+1}\})$

Rough Proof: If $x = \sum_{i=1}^{k} \lambda_i x_i$ where $k \geq n+2$,

$\{x_2 - x_1, x_3 - x_1, \cdots, x_k - x_1\}$ is linearly dependent

$\exists\, \mu_2, \cdots, \mu_k$ not all zero s.t. $\sum_{i=2}^{k} \mu_i (x_i - x_1) = 0$

$$\Rightarrow \sum_{i=1}^{k} \mu_i x_i = 0 \text{ where } \mu_1 := -\sum_{i=2}^{k} \mu_i$$

so $\sum \mu_i = 0$, $\exists\, i$ s.t. $\mu_i < 0$

Note $x = \sum_{i=1}^{k} (\lambda_i + \alpha \mu_i) x_i$ is another convex representation $\forall \alpha > 0$

$\mu_i < 0$ for some i, so by picking appropriate α, can make $\lambda_i + \alpha\mu_i \geq 0\ \forall i$, $\lambda_i + \alpha\mu_i = 0$ for some i

i.e. it is a convex representation in less than k vectors

Repeat until only $n+1$ vectors left. $\square$

Follow this proof to reduce convex representation to $n+1$ vectors

<u>Extreme Point</u> $\nexists\, x_1, x_2\ (x_1 \neq x_2)$, $\lambda \in (0,1)$ s.t. $x = \lambda x_1 + (1-\lambda) x_2$

<u>Krein-Milman Theorem</u>

$$S \begin{cases} \text{compact} \\ \text{convex} \end{cases} \Rightarrow S = \operatorname{conv}(\underbrace{\operatorname{ext}}_{\text{set of extreme points}}(S))$$

Convex Functions

Domain	Equivalent Characterisation
convex	$(\nabla f(y) - \nabla f(x))^T (y-x) \geq 0$ (monotonic) $f(x) + \nabla f(x)^T (y-x) \leq f(y)$ (gradient) (1st-order characterization of convex f)
open + convex	$\nabla^2 f(x) \succcurlyeq 0 \quad \forall x$ (second order)

Combination

(linear transform on output) convex, $\alpha \geq 0 \Rightarrow \alpha f$ convex

f_i convex $\Rightarrow \sum_i f_i$ convex

(linear transform on input) f convex $\Rightarrow g(y) := f(Ay+b)$ convex

(Compound)
f convex on C
g non-decreasing on $f(C)$ and convex $\Rightarrow h(x) := g(f(x))$ convex

(full max) f_i convex $\Rightarrow \max_i f_i(x)$ convex

(partial min) $f(x,y)$ convex $\Rightarrow g(x) := \min_y f(x,y)$ convex

Properties

1. local Lipschitz continuous
2. all directional derivative exists (for interior points)
3. No maximum in int(C)
4. At least one maximiser is an extreme point

Examples (may be used in exams)

Name	Function	Domain
Affine	$x \mapsto a^T x + b$	$\mathbb{R}^n$
Norm	$x \mapsto \|x\|$	$\mathbb{R}^n$
quad-over-lin	$x \mapsto \frac{x_1^2}{x_2}$	$\mathbb{R} \times \mathbb{R}_{++}$
log-sum-exp	$x \mapsto \log(\sum_{i=1}^{n} e^{x_i})$	$\mathbb{R}^n$
sum of smallest k values	$x \mapsto \sum_{i=1}^{k} x_{[i]}$	$\mathbb{R}^n$
distance to convex set C	$x \mapsto \inf_{y \in C} \|y - x\|$	$\mathbb{R}^n$
	$x \mapsto \sqrt{x^T Q x + 1}$	$\mathbb{R}^n$

Convex Optimisation

Convex problem: minimisation of a convex function over closed & convex set.

$g_i(x) \le 0$ (convex) $h_j(x) = 0$ (affine)

Stationarity $\nabla f(x^*)^T (x - x^*) \geq 0 \quad \forall x \in C$

$\left.\begin{array}{l} f \text{ cont. diff.} \\ \text{domain nonempty, closed} \\ \text{convex} \end{array}\right\}$ local min $\Rightarrow$ stationarity

stationarity $\Leftrightarrow x^* = P_C(x^* - s\nabla f(x^*)) \ \forall s > 0$

$\left.\begin{array}{l} f \text{ cont. diff., convex} \\ \text{domain non-empty, closed,} \\ \text{convex} \end{array}\right\}$ global min (optimality) $\Leftrightarrow$ stationarity

Projection: $P_C(x) = \operatorname{argmin}\{\|y - x\|^2 : y \in C\}$

- first proj. Thm.: $C \begin{cases} \text{non-empty} \\ \text{closed} \\ \text{convex} \end{cases} \Rightarrow P_C(x)$ exists and is unique
- second proj. Thm $z = P_C(x) \Leftrightarrow (x - z)^T (y - z) \le 0$

Gradient Projection

$$x^{k+1} = P_C(\underbrace{x^k - t^k \nabla f(x^k)}_{\text{Gradient descent}})$$

Existence & Uniqueness of solutions

f continuous, coercive, S non-empty, closed $\}$ $\exists$ global minimum on S

f convex, C convex, non-empty, compact $\}$ all extreme points are maximisers

f (strict) convex, C convex $\}$ local min must be (strict) global min

f strictly convex, C convex $\}$ at most one minimiser

KKT

$\min f(x)$
s.t. $g_i(x) \le 0$
$h_j(x) = 0$

KKT:
$L := f + \lambda_i g_i + \mu_j h_j$
(1) $\nabla L(x) = 0$ (3) $\lambda_i \geq 0$
(2) $\lambda_i g_i(x) = 0$

f (objective)	g_i (inequality)	h_j (equality)	condition	result
cont. diff	affine	affine		optimal $\Rightarrow$ KKT
convex	affine	affine		optimal $\Leftarrow$ KKT
convex	convex	affine		optimal $\Leftarrow$ KKT
convex	convex	affine	Slater	optimal $\Leftrightarrow$ KKT

Slater: $\exists x$ s.t. $g_i(x) < 0$, $h_j(x) = 0$

Gradient Descent

$\min_{x \in \mathbb{R}^n} \{f(x)\}$ Iterative Algorithm $x^{k+1} = x^k + t^k d^k$

$d^k = -\nabla f(x^k)$

choice of t^k $\begin{cases} \text{constant} \\ \text{exact line search} : t^k \in \operatorname{argmin}_{t \geq 0} f(x^k + t^k d^k). \end{cases}$

Zig-Zag: $\{x^k\}_{k\geq 0}$ generated by exact line search

$(x^{k+2} - x^{k+1})^T (x^{k+1} - x^k) = 0$

<u>Lipschitz Gradient</u> $\|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\|$ $C_L^{1,1}(\mathbb{R}^n)$ class of functions with Lipschitz gradient

$f \in C_L^{1,1}(\mathbb{R}^n) \Leftrightarrow \|\nabla^2 f(x)\| \leq L \quad \forall x \in \mathbb{R}^n$

<u>Condition Number</u> $\kappa(A) = \dfrac{\lambda_{max}(A)}{\lambda_{min}(A)}$

Scaled GD

Improves conditioning of Hessian $\nabla^2 f(x)$

Obtained from change of variable $x = Sy$ Let $g(y) := f(Sy)$

$x^{k+1} = x^k - t^k \underbrace{SS^T}_{=:D} \nabla f(x^k)$

$\nabla^2 g(y) = D^{\frac{1}{2}} \nabla^2 f(x) D^{\frac{1}{2}}$ Choices of D: $\begin{cases} D^k = (\nabla^2 f(x^k))^{-1} \\ \text{diagonal scaling} \\ (D^k)_{ii} = \left(\dfrac{\partial^2 f(x^k)}{\partial x_i^2}\right)^{-1} \end{cases}$

Gauss-Newton Method.

$\min_x \{g(x)\}$ $g(x) := \sum_{i=1}^m (f_i(x) - c_i)^2 = \|F(x)\|^2$

where $F(x) = (f_1(x) - c_1, \cdots, f_m(x) - c_m)^T$

The algorithm: $x^{k+1} = \operatorname{arg\,min}_{x \in \mathbb{R}^n} \left(\sum_{i=1}^m [\underbrace{f_i(x^k) + \nabla f_i(x^k)(x - x^k)}_{\text{linear Approximation}} - c_i]^2 \right)$

$= (J(x^k)^T J(x^k))^{-1} J(x^k)^T b^k$

$= x^k - \frac{1}{2}(J^T J)^{-1} \nabla g(x^k)$

where $J(x^k) := \begin{pmatrix} - \nabla f_1(x^k)^T - \\ \vdots \\ - \nabla f_m(x^k)^T - \end{pmatrix}$

$b^k := J(x^k) x^k - F(x^k)$

Algebraic Techniques

!! C-S $|a \cdot b| \leq \|a\| \|b\|$ $\min_{\|x\| \leq 1} a^T x = -\|a\|$

$\nabla(\|x\|) = \frac{x}{\|x\|}$ $\max_{\cdots} \cdots = \|a\|$

Transform: $\min_{x,y} f(x,y) = \min_{u,v} g(u,v)$

where $g(u,v) := f(x(u,v), y(u,v))$

Be careful with possible additional constraint to make transformation $(x,y) \mapsto (u,v)$ well-defined

Trace $Tr(A) := \sum_{i=1}^{n} a_{ii}$

- Tr is linear
- $Tr(A^T) = Tr(A)$
- $Tr(AB) = Tr(BA)$
- Traces of similar matrices are the same $Tr(P^{-1}AP) = Tr(APP^{-1}) = Tr(A)$
- $Tr(ba^T) = a^T b$

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

- $\det(cA) = c^n \det(A)$
- $\det(A^T) = \det(A)$
- $\det(AB) = \det(A)\det(B)$
- $\det(A^{-1}) = \frac{1}{\det(A)}$

If λ_i are Eigenvalues of matrix A

$\sum_{i=1}^{n} \lambda_i = Tr(A)$, $\prod_{i=1}^{n} \lambda_i = \det(A)$

Least Square

$\min_x \|Sx - b\|^2$ where $S \in \mathbb{R}^{m \times n}$, rank$(S) = n$ $(m > n)$

Normal Equation $(S^T S) x = S^T b$

Solution: $x_{LS} = (S^T S)^{-1} S^T b$

Least Square with Quadratic Regularisation

$\min_x \|Sx - b\|^2 + \lambda \|Dx\|^2 \quad \lambda > 0$

solution: $x_{RLS} = (S^T S + \lambda D^T D)^{-1} S^T b$

requires $\mathrm{Null}(S) \cap \mathrm{Null}(D) = \{0\}$

Separation thm.

Orthogonal projections