Mathématiques pour l’intelligence artificielle

M2 informatique parcours IAAA, Aix-Marseille Université

Thomas Schatz

Mercredi 6 septembre 2023

Plan du cours (prévisionnel)

Introduction

1. Notions de bases sur les preuves

2. Algèbre linéaire

3. Optimisation

4. Probabilités

5. Statistique

Discussion

Algèbre linéaire

1. Espaces vectoriels et fonctions linéaires
2. Matrices
3. Angles et orthogonalité
4. Structure des applications linéaires et matrices

Angles et orthogonalité

Définition: dans un espace vectoriel réel (i.e. K = **R**), un *produit scalaire* est une forme bilinéaire symétrique définie positive

Forme : $(\mid) : \left| \begin{array}{ccc} E \times E & \to & \mathbf{R} \\ (v, w) & \mapsto & (v \mid w) \end{array} \right.$

Bilinéaire : linéaire en v pour w fixé et en w pour v fixé

Symétrique : $\text{Pour tout } v, w \in E^2, \ (v \mid w) = (w \mid v)$

Définie positive : $\text{Pour tout } v \in E \setminus \{0_E\}, (v \mid v) > 0$

Angles et orthogonalité

Norme associée au produit scalaire : $\|v\| = \sqrt{(v \mid v)}$

Exemple: produit scalaire et norme Euclidienne sur $\mathbf{R}^n$

Vecteurs orthogonaux: si et seulement si $(v \mid v) = 0$

Famille orthogonale:

$$(v_i)_{i \in I}, \text{ telle que pour tout } (i, j) \in I^2,\ i \neq j,\ (v_i \mid v_j) = 0$$

Famille orthonormale: famille orthogonale telle que

$$\text{pour tout } i \in I,\ \|v_i\| = 1$$

Angles et orthogonalité

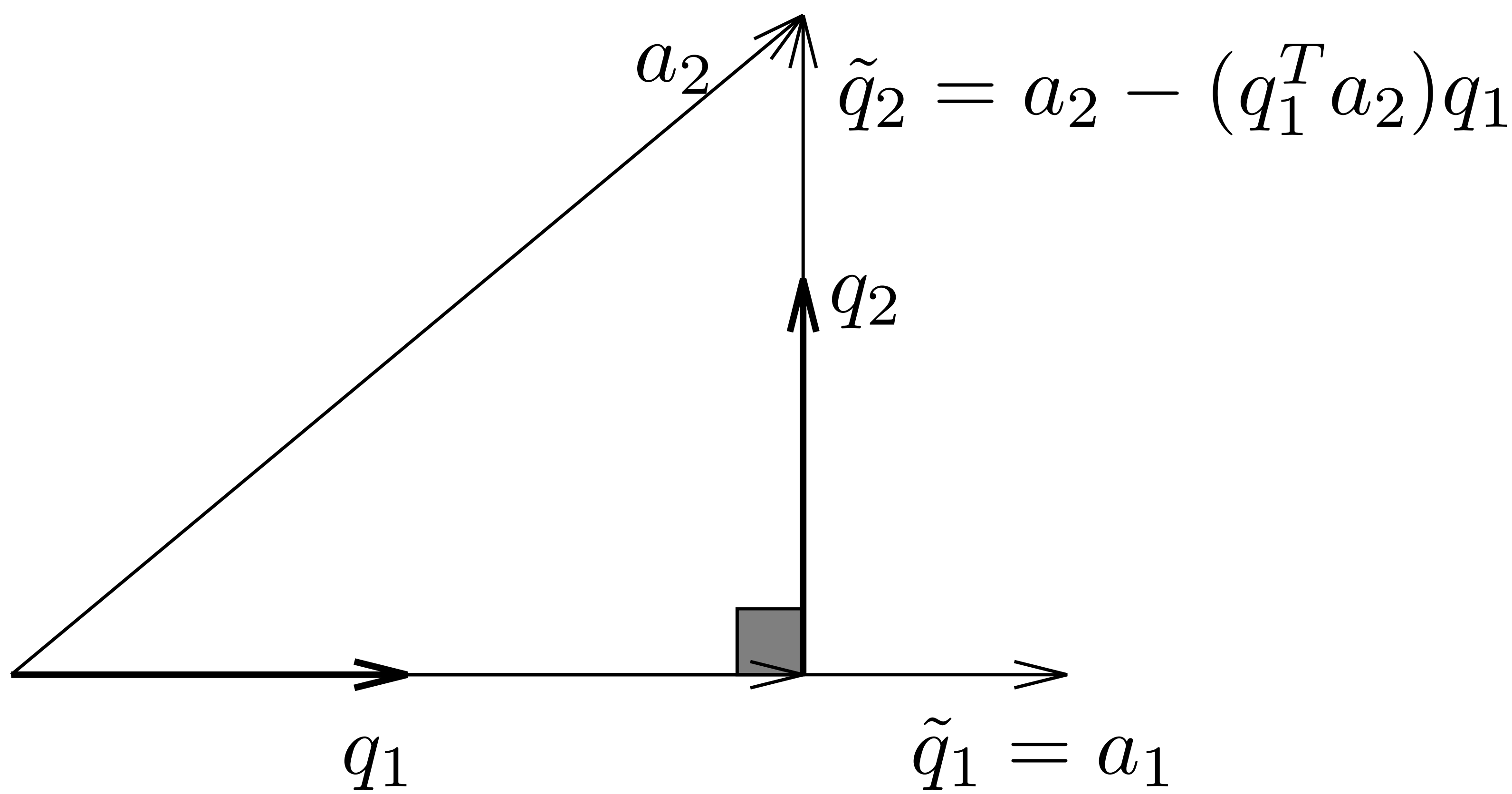
Théorème : une famille de vecteurs orthogonaux est libre

Comment construire des familles orthonormales?

Gram-Schmidt procedure : étant donné k vecteurs indépendants $(a_1, a_2, \ldots, a_k) \in E^k$, trouver des vecteurs orthonormaux $(q_1, q_2, \ldots, a_k)$, tels que pour tout $r \leq k$, $(q_1, q_2, \ldots, q_r)$ soit une base orthonormale de $\mathrm{Vect}(a_1, a_2, \ldots, a_r)$

- step 1a. $\tilde{q}_1 := a_1$
- step 1b. $q_1 := \tilde{q}_1 / \|\tilde{q}_1\|$ (normalize)
- step 2a. $\tilde{q}_2 := a_2 - (q_1^T a_2) q_1$ (remove q_1 component from a_2)
- step 2b. $q_2 := \tilde{q}_2 / \|\tilde{q}_2\|$ (normalize)
- step 3a. $\tilde{q}_3 := a_3 - (q_1^T a_3) q_1 - (q_2^T a_3) q_2$ (remove q_1, q_2 components)
- step 3b. $q_3 := \tilde{q}_3 / \|\tilde{q}_3\|$ (normalize)
- etc.

Orthogonalité



Gram-Schmidt et factorisation QR

$$\underbrace{\begin{bmatrix} a_1 & a_2 & \cdots & a_k \end{bmatrix}}_{A} = \underbrace{\begin{bmatrix} q_1 & q_2 & \cdots & q_k \end{bmatrix}}_{Q} \underbrace{\begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1k} \\ 0 & r_{22} & \cdots & r_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & r_{kk} \end{bmatrix}}_{R}$$

k : rang de la matrice A

$$r_{ij} = q_i^T a_j$$ Calculé durant Gram-Schmidt

Une matrice est dite semi-orthogonale si ces colonnes sont orthonormales.

Si Q est une matrice semi-orthogonale de taille $m \times k$, alors $Q^T Q = I_k$

Une matrice semi-orthogonale carrée est dite orthogonale.

Angles et orthogonalité

Intérêt ?

Soit E un espace vectoriel de dimension finie n muni d'un produit scalaire $(\,.\, | \,.\,)$ et $V = (v_1, \ldots, v_n)$ une base orthonormale de E. Alors, pour tout $v \in E$,

$$v = \sum_{i=1}^{n} (v \mid v_i) v_i.$$

Algèbre linéaire

1. Espaces vectoriels et fonctions linéaires
2. Matrices
3. Angles et orthogonalité
4. Structure des applications linéaires et matrices

Décomposition en valeurs singulières

Etant donné une matrice rectangulaire quelconque

(à coefficients réels)

$$A = \begin{pmatrix} a_{1,1} & \cdots & a_{1,n} \\ \vdots & \ddots & \vdots \\ a_{m,1} & \cdots & a_{m,n} \end{pmatrix}$$

Quelle transformation linéaire représente-t-elle ?

Par exemple : $M = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}$

Décomposition en valeurs singulières

$$\boldsymbol{AV} = \boldsymbol{U\Sigma} \qquad A\begin{bmatrix} \\ \boldsymbol{v}_1 \, .. \, \boldsymbol{v}_r \, .. \, \boldsymbol{v}_n \\ \\ \end{bmatrix} = \begin{bmatrix} \\ \boldsymbol{u}_1 \, .. \, \boldsymbol{u}_r \, .. \, \boldsymbol{u}_m \\ \\ \end{bmatrix} \left[\begin{array}{ccc|c} \boldsymbol{\sigma_1} & & & \\ & \ddots & & \mathbf{0} \\ & & \boldsymbol{\sigma_r} & \\ \hline & \mathbf{0} & & \mathbf{0} \end{array}\right]$$

Flashback: interpretation des opérations matricielles

$$c_{ij} = \tilde{a}_i^T b_j = \langle \tilde{a}_i, b_j \rangle$$

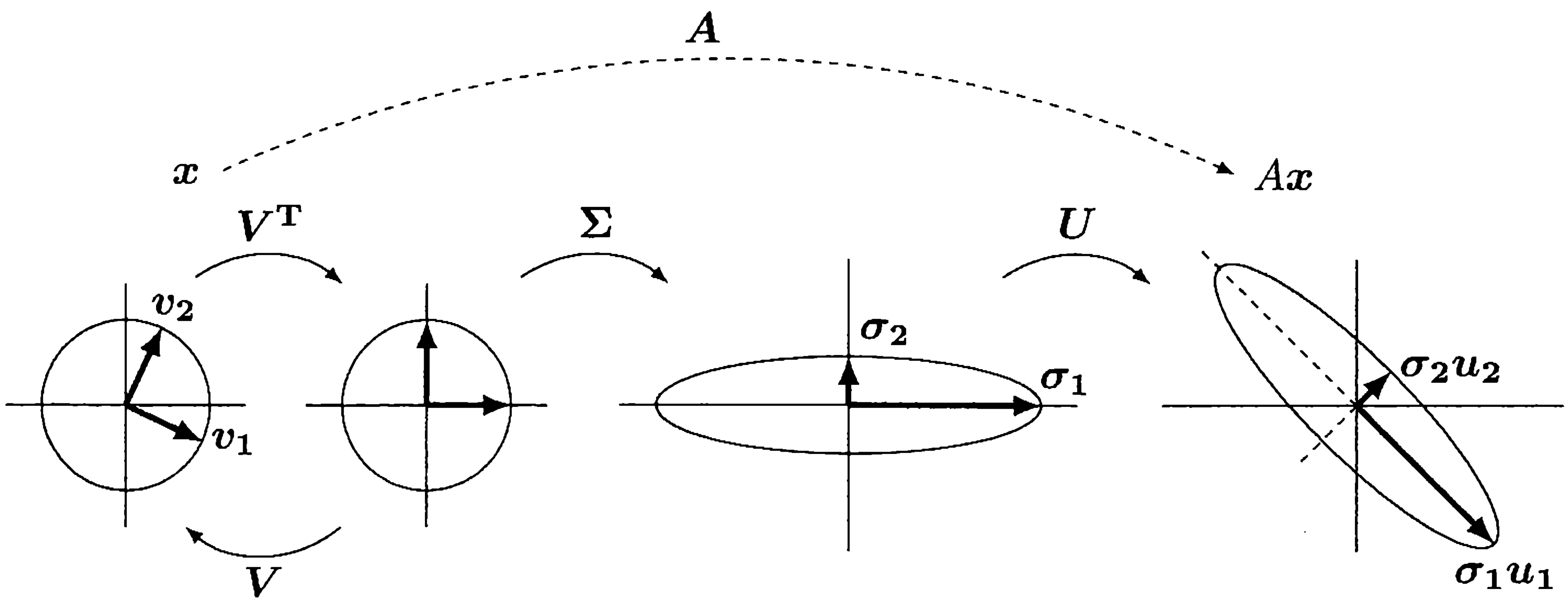
$$C = \left[\begin{array}{c} c_1 \cdots c_p \end{array} \right] = AB = \left[\begin{array}{c} Ab_1 \cdots Ab_p \end{array} \right]$$

$$C = \left[\begin{array}{c} \tilde{c}_1^T \\ \vdots \\ \tilde{c}_m^T \end{array} \right] = AB = \left[\begin{array}{c} \tilde{a}_1^T B \\ \vdots \\ \tilde{a}_m^T B \end{array} \right]$$

$$C = \sum_i a_i \tilde{b}_i^T$$

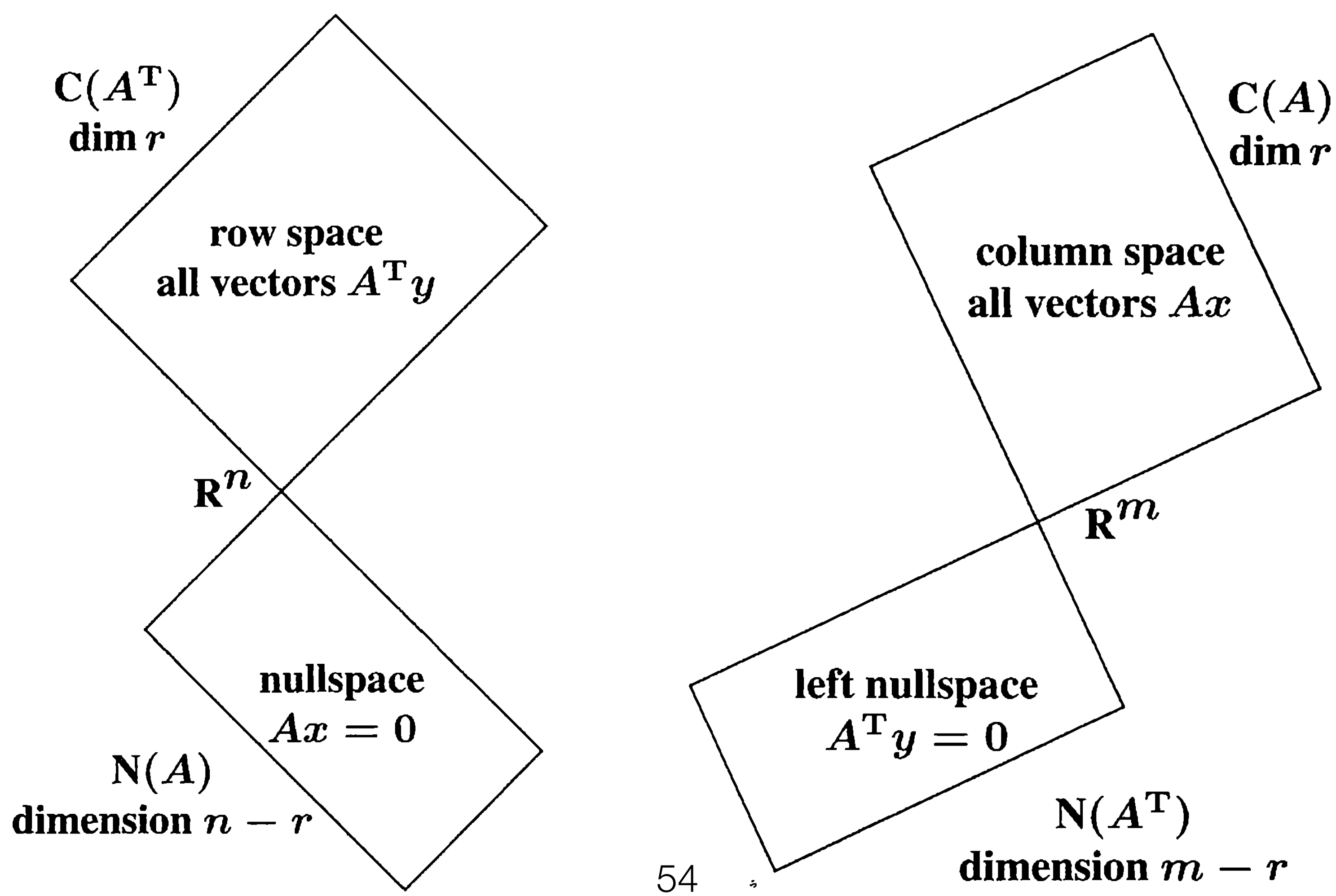
Décomposition en valeurs singulières

$$\boldsymbol{AV = U\Sigma} \qquad A\begin{bmatrix} \\ \boldsymbol{v}_1 \,..\, \boldsymbol{v}_r \,..\, \boldsymbol{v}_n \\ \\ \end{bmatrix} = \begin{bmatrix} \\ \boldsymbol{u}_1 \,..\, \boldsymbol{u}_r \,..\, \boldsymbol{u}_m \\ \\ \end{bmatrix} \left[\begin{array}{ccc|c} \boldsymbol{\sigma_1} & & & \\ & \ddots & & \mathbf{0} \\ & & \boldsymbol{\sigma_r} & \\ \hline & \mathbf{0} & & \mathbf{0} \end{array}\right]$$



Espaces associés à une matrice

$$\boxed{AV = U\Sigma \qquad A\begin{bmatrix} \\ v_1 \;..\; v_r \;..\; v_n \\ \\ \end{bmatrix} = \begin{bmatrix} \\ u_1 \;..\; u_r \;..\; u_m \\ \\ \end{bmatrix}\left[\begin{array}{ccc|c} \sigma_1 & & & \\ & \ddots & & \mathbf{0} \\ & & \sigma_r & \\ \hline & \mathbf{0} & & \mathbf{0} \end{array}\right]}$$



Espaces associés à une matrice

$$\boxed{\boldsymbol{AV = U\Sigma} \qquad A\begin{bmatrix} \\ \boldsymbol{v}_1 \,..\, \boldsymbol{v}_r \,..\, \boldsymbol{v}_n \\ \\ \end{bmatrix} = \begin{bmatrix} \\ \boldsymbol{u}_1 \,..\, \boldsymbol{u}_r \,..\, \boldsymbol{u}_m \\ \\ \end{bmatrix} \left[\begin{array}{ccc|c} \boldsymbol{\sigma_1} & & & \\ & \ddots & & \mathbf{0} \\ & & \boldsymbol{\sigma_r} & \\ \hline & \mathbf{0} & & \mathbf{0} \end{array}\right]}$$

Par exemple : $M = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}$

$$U \approx \begin{pmatrix} -.21 & -.89 & .41 \\ -.52 & -.25 & -.82 \\ -.83 & .39 & .41 \end{pmatrix} \qquad V \approx \begin{pmatrix} -.48 & -.57 & -.66 \\ .78 & .08 & -.62 \\ .41 & -.82 & .41 \end{pmatrix}$$

$$\begin{pmatrix} \sigma_1 & \sigma_2 & \sigma_3 \end{pmatrix} \approx \begin{pmatrix} 16.85 & 1.07 & 0 \end{pmatrix}$$

Théorème spectral

Si S est une matrice symétrique, réelle de taille $m \times m$, alors il existe une matrice orthogonale réelle Q de taille $m \times m$ et une matrice diagonale réelle Λ de taille $m \times m$ telles que $S = Q\Lambda Q^T$.

Positivité

Une matrice symétrique réelle est dite définie positive, noté $S \succ 0$ ssi pour toute matrice colonne u, $u^T S u > 0$.

Une matrice symétrique réelle est dite semi-définie positive, $S \succeq 0$, ssi pour toute matrice colonne u, $u^T S u \geq 0$.

Relation d'ordre sur les matrices symétriques réelles:
$S_1 \prec S_2$ ssi $S_2 - S_1 \succ 0$
et
$S_1 \preceq S_2$ ssi $S_2 - S_1 \succeq 0$

Diagonalisation

Soit A une matrice à coefficients réels de taille $m \times m$. $\lambda \in \mathbf{C}$ est une valeur propre de A ssi il existe $v \in \mathbf{C}^m$, non-nul, tel que $Av = \lambda v$, i.e. l'image de v par A est dans la même direction que v. Un tel v est appelé un vecteur propre de A associé à la valeur propre λ.

Diagonalisation

suppose $v_1, \ldots, v_n$ is a *linearly independent* set of eigenvectors of $A \in \mathbf{R}^{n \times n}$:

$$Av_i = \lambda_i v_i, \quad i = 1, \ldots, n$$

express as

$$A \left[\begin{array}{ccc} v_1 & \cdots & v_n \end{array} \right] = \left[\begin{array}{ccc} v_1 & \cdots & v_n \end{array} \right] \left[\begin{array}{ccc} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{array} \right]$$

define $T = \left[\begin{array}{ccc} v_1 & \cdots & v_n \end{array} \right]$ and $\Lambda = \mathbf{diag}(\lambda_1, \ldots, \lambda_n)$, so

$$AT = T\Lambda$$

and finally

$$T^{-1}AT = \Lambda$$

Forme canonique de Jordan

any matrix $A \in \mathbf{R}^{n\times n}$ can be put in *Jordan canonical form* by a similarity transformation, $i.e.$

$$T^{-1}AT = J = \begin{bmatrix} J_1 & & \\ & \ddots & \\ & & J_q \end{bmatrix}$$

where

$$J_i = \begin{bmatrix} \lambda_i & 1 & & \\ & \lambda_i & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_i \end{bmatrix} \in \mathbf{C}^{n_i\times n_i}$$

is called a *Jordan block* of size n_i with eigenvalue λ_i (so $n = \sum_{i=1}^q n_i$)

Déterminant et trace

A matrice carrée $n \times n$

$$\det(A) = \sum_{\sigma \in S_n} \left(\operatorname{sgn}(\sigma) \prod_{i=1}^{n} a_{i,\sigma_i} \right),$$

$\operatorname{sgn}(\sigma)$ est la parité du nombre d'élement dans une décomposition de σ en une séquence de transpositions (échange de deux éléments).

$$\det(AB) = \det(A)\det(B)$$

$$\operatorname{Tr}(A) = \sum_{i=1}^{n} a_{i,i} \quad \operatorname{Tr}(\mathrm{AB}) = \operatorname{Tr}(\mathrm{BA})$$

$$\operatorname{Tr}(A_1 A_2 \ldots A_k) = \operatorname{Tr}(A_2 A_3 \ldots A_k A_1) = \cdots = \operatorname{Tr}(A_k A_1 A_2 \ldots A_{k-1})$$

Polynôme caractéristique et théorème de Cayley-Hamilton

Cayley-Hamilton theorem: for any $A \in \mathbf{R}^{n\times n}$ we have $\mathcal{X}(A) = 0$, where $\mathcal{X}(s) = \det(sI - A)$

Corollaire : pour tout entier naturel p, $A^p \in \mathrm{Vect}(I, A, A^2, \ldots, A^{n-1})$

Algèbre linéaire

1. Espaces vectoriels et fonctions linéaires
2. Matrices
3. Angles et orthogonalité
4. Structure des applications linéaires et matrices

Plan du cours (prévisionnel)

Introduction

1. Notions de bases sur les preuves

2. Algèbre linéaire

3. Optimisation

4. Probabilités

5. Statistique

Discussion

Optimisation sans contraintes

$$f : \mathbf{R}^n \mapsto \mathbf{R}$$

$$\min_x f(x) \ ?$$

x^* est un minimum global de f ssi pour tout $x \in \mathbf{R}^n$, $f(x^*) \leq f(x)$

x^* est un minimum local de f ssi il existe un ensemble ouvert $U \subset \mathbf{R}^n$ contenant x^* tel que pour tout $x \in U$, $f(x^*) \leq f(x)$

Analyse

Concept central $\lim_{x \to a} f(x) = l$?

Notions de base de topologie

Dans $\mathbf{R}^n$, La boule fermée $B_f(x, \epsilon)$, centrée sur x et de rayon ϵ associée à la norme Euclidienne est l'ensemble des points y de $\mathbf{R}^n$ tels que $\|x - y\|_2 \leq \epsilon$

Dans $\mathbf{R}^n$, La boule ouverte $B_o(x, \epsilon)$, centrée sur x et de rayon ϵ associée à la norme Euclidienne est l'ensemble des points y de $\mathbf{R}^n$ tels que $\|x - y\|_2 < \epsilon$

U est un ouvert de $\mathbf{R}^n$ ssi $U \subset \mathbf{R}^n$ et pour tout x dans U, il existe $\epsilon > 0$, tel que $B_f(x, \epsilon) \subset U$.

U est un fermé de $\mathbf{R}^n$ ssi son complément dans $\mathbf{R}^n$ est un ouvert de $\mathbf{R}^n$

Jacobienne, gradient

$$f : U \subset \mathbf{R}^n \to \mathbf{R}^m,\ \text{de classe C}^1$$

Dérivée partielle $f : x \mapsto (f_1(x), f_2(x), \ldots, f_m(x))$

$$\frac{\partial f_i}{\partial x_j} : (a_1, \ldots, a_n) \mapsto \lim_{h \to 0} \frac{f_i(a_1, \ldots, a_{j-1}, a_j + h, a_{j+1}, \ldots a_n) - f_i(a_1, \ldots, a_j, \ldots, a_n)}{h}$$

Matrice Jacobienne (transposée du gradient si m=1)

$$\mathbf{J} = \begin{bmatrix} \frac{\partial \mathbf{f}}{\partial x_1} & \cdots & \frac{\partial \mathbf{f}}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \nabla^{\mathrm{T}} f_1 \\ \vdots \\ \nabla^{\mathrm{T}} f_m \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}$$

“Chain rule”

$$J_{f \circ g}(\mathbf{a}) = J_f(g(\mathbf{a})) J_g(\mathbf{a}),$$

Conditions d'optimalité

Conditions nécessaires, premier ordre

Si x^* est un minimum local et f est de classe $\mathcal{C}^1$ sur un ouvert $U \subset \mathbf{R}^n$ contenant x, alors $\nabla f(x^*) = 0$

Analyse

Hessienne

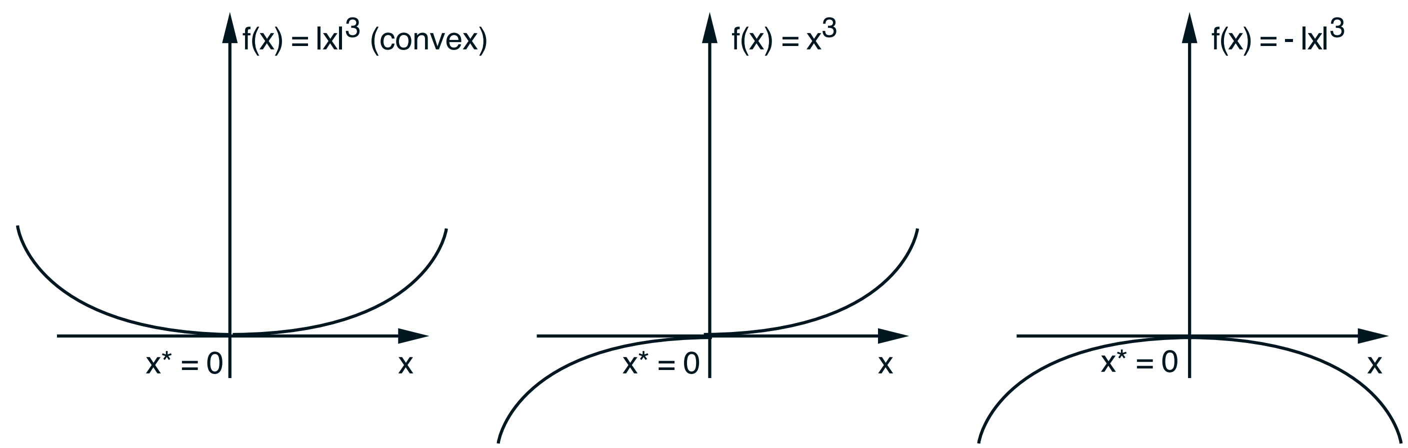
$$f : U \subset \mathbf{R}^n \mapsto \mathbf{R}$$

$$\nabla^2 f = \boldsymbol{H}(f) = \begin{bmatrix} \dfrac{\partial^2 f}{\partial {x_1}^2} & \dfrac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \dfrac{\partial^2 f}{\partial x_1 \partial x_n} \\ \dfrac{\partial^2 f}{\partial x_2 \partial x_1} & \dfrac{\partial^2 f}{\partial {x_2}^2} & \cdots & \dfrac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial^2 f}{\partial x_n \partial x_1} & \dfrac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \dfrac{\partial^2 f}{\partial {x_n}^2} \end{bmatrix}$$

Conditions d'optimalité

Conditions nécessaires, premier ordre

Si x^* est un minimum local et f est de classe $\mathcal{C}^1$ sur un ouvert $U \subset \mathbf{R}^n$ contenant x, alors $\nabla f(x^*) = 0$



Conditions nécessaires, second ordre

Si x^* est un minimum local et f est de classe $\mathcal{C}^2$ sur un ouvert $U \subset \mathbf{R}^n$ contenant x, alors $\nabla f(x^*) = 0$ et $\nabla^2 f(x^*)$ est semi-définie positive

Conditions suffisantes, second ordre

Si f est de classe $\mathcal{C}^2$ sur un ouvert $U \subset \mathbf{R}^n$ contenant x, si $\nabla f(x^*) = 0$ et si $\nabla^2 f(x^*)$ est définie positive, alors x^* est un minimum local.

Existence de minimum

Théorème de Weierstrass

Si f est définie et continue sur un sous-ensemble fermé et borné de $\mathbf{R}^n$, alors f admet un minimum global.

Théorème

Si f est définie et continue sur $\mathbf{R}^n$ et *coercive* (i.e. $f(x) \to +\infty$ when $\|x\| \to +\infty$, alors f admet un minimum global sur tout sous-ensemble fermé de $\mathbf{R}^n$.

Unicité du minimum ?

Prochaine séance : convexité