Mathématiques

Mazen Saad
mazen.saad@ec-nantes.fr

Quelques notes sur les cours :

- 16h de cours, 44h de TD, 12h de TP,
- 7 chapitres,
- 2 DS (novembre & janvier): note EVI coef. 1/3 par DS,
- compte-rendus de TP (note EVC) coef 1/3.

Les chapitres de cette année :

(1) Introduction à l'analyse numérique (méthodes directes de résolution de systèmes linéaires),
(2) Méthodes itératives de résolution de systèmes linéaires d'approximation de valeurs propres,
(3) Optimisation sans contrainte,
(4) Optimisation avec contraintes,
(5) Probabilités,
(6) Statistiques,
(7) Interpolation et approximation, intégration numérique.

Table des matières

I. Introduction à l'analyse numérique, méthodes directes de résolution de systèmes linéaires

I.1. Arithmétiques flottantes

Comment les nombres sont représentés dans une machine (un ordinateur) ? Un nombre est représenté par un nombre fini de caractères. Des nombres comme $\sqrt{2}$ doivent être tronqués.

Un nombre réel est représenté par un un *nombre flottant*, où un nombre flottant est de la forme suivante :

$$\pm a \cdot 10^q$$

où q est un entier tel que $-M \leq q \leq M$, et a est la *mantisse* de la forme :

$$0.d_1 d_2 \dots d_t$$

avec $d_1 \neq 0$ et t le nombre de chiffres de la mantisse.

Un nombre peut être représenté de plusieurs manières.

Simple précision. On utilise 4 octets (32 bits), on a $M = 38$ et $t = 7$. La précision est de 7 chiffres décimaux.

Double précision. On utilise 8 octets (64 bits), on a $M = 308$ et $t = 15$. La précision est de 15 chiffres décimaux. Ceci est le type le plus utilisé (type `double` en C).

Les conséquences de cette précision sont :

(1) Il existe un plus petit nombre flottant machine (appelé *zéro machine*). Il est accessible avec `eps` en Matlab. Ce nombre est donc de la forme $0,100\dots00 \cdot 10^{-M}$ (en valeur absolue).

(2) Il existe un plus grand nombre : $0,\underbrace{99\dots9}_{t \text{ fois}} \cdot 10^{+M}$.

(3) Tous les réels ne sont pas représentés :

$$\sqrt{2} \approx 0,1414234 \cdot 10^1 \quad \text{et} \quad \pi \approx 0,31415925 \cdot 10^1.$$

(4) Toute opération élémentaire ($+, \times, \div$) est en général entachée d'une erreur.

Exemple. Avec une machine avec $t = 2$, on pose $a = 0,63 \cdot 10^1$ et $b = 0,82 \cdot 10^{-4}$. On fait la somme et l'ordinateur réduit au même exposant (le plus grand) :

$$\begin{aligned} a : &\quad 0,6300000 \cdot 10^1 \\ b : &\quad 0,0000082 \cdot 10^1 \\ a+b : &\quad 0,6300082 \cdot 10^1 \end{aligned}$$

On a $\mathrm{fl}(a+b) = 0,63 \cdot 10^1$, donc $\mathrm{fl}(a+b) = \mathrm{fl}(a)$ et $b \neq 0$. Par contre, $a + b > a$ car $b > 0$.

Si on note $z = a + b$, alors

$$\frac{1}{\mathrm{fl}(z-a)} = \frac{1}{0} = +\infty \quad \text{au lieu de} \quad \frac{1}{b}.$$

Le réel $\sqrt{2}$ est calculé 10^-7 près en simple précision. Si on l'utilise 10^7 fois, alors l'erreur est d'environ 1.

I.2. Calcul matriciel (document sur le serveur pédagogique)

I.3. Systèmes linéaires creux

I.3.a. Équation de la chaleur

On note $u(x)$ la distribution de température d'une plaque chauffée aux deux extrémités (Figure 1).

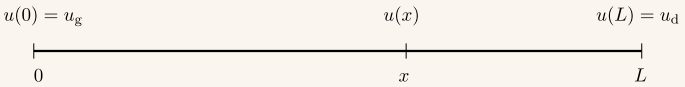

Figure 1: Plaque chauffée

On ajoute les conditions du système ci-dessous :

$$\begin{cases} -u''(x) = f(x) \quad 0 < x < L \\ u(0) = u_{\mathrm{g}} \\ u(L) = u_{\mathrm{d}} \end{cases}$$

où f est le *terme source*. Si f est intégrable, alors u est exacte. Sinon, on approxime la solution u.

Analyse numérique. On approxime la solution par une méthode d'approximation comme la *méthode des différences finies*. On décompose l'intervalle $[0, L]$ en $(N+1)$ intervalles de pas h, où

$$h = \frac{L}{N+1}$$

, comme montré dans la Figure 2.

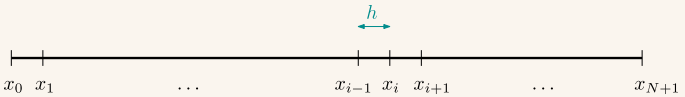

Figure 2: Approximation de la plaque chauffée

Ainsi, on a $x_i = i \cdot h$, pour $i = 0, N+1$, et donc $x_{i+1} = x_i + h$.

On écrit le développement de Taylor :

$$-u''(x_i) = \frac{-u(x_{i-1}) + 2u(x_i) - u(x_{i+1})}{h^2} + \mathcal{O}(h^2).$$

On note $u_i \approx u(x_i)$ une approximation de la solution exacte au point x_i. La méthode des différences finies s'écrit comme le système :

$$(S_L) : \begin{cases} \frac{-u_{u+1}+2u_i-u_{i+1}}{h^2} = f(x_i), \quad i = 1, N \\ u(0) = u_0 = u_{\mathrm{g}} \\ u(L) = u_{N+1} = u_{\mathrm{d}} \end{cases}$$

Le système (S_L) est un système linéaire. On cherche $\boldsymbol{U}_h = (U_1, U_2, \ldots, U_N)^\top$ solution de (S_L). Le système linéaire (S_L) est équivalent à résoudre $\boldsymbol{A}_h \cdot \boldsymbol{U}_h = \boldsymbol{b}_h$ avec

$$\boldsymbol{A}_h = \frac{1}{h^2}\begin{pmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \ddots & \vdots \\ 0 & -1 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & -1 \\ 0 & \cdots & 0 & -1 & 2 \end{pmatrix} \in \mathscr{M}_{N,N}(\mathbb{R}) \quad \text{et} \quad \boldsymbol{b}_h = \begin{pmatrix} f(x_1) + \frac{u_{\mathrm{g}}}{h^2} \\ \vdots \\ f(x_i) \\ \vdots \\ f(x_N) + \frac{u_{\mathrm{d}}}{h^2} \end{pmatrix}.$$

Remarques.

(1) $\boldsymbol{A}_h$ est une matrice creuse : le nombre de coefficients non nuls est $3(N-2)+4 \approx 3N$, à comparer avec N^2 coefficient pour une matrice pleine.

(2) $\boldsymbol{A}_h^{-1}$ est une matrice pleine.

I.3.b. Méthodes directes pour résoudre $\boldsymbol{Ax} = \boldsymbol{b}$

I.3.b.i. Méthode de Cramer

On pose $x_i = \det \boldsymbol{A}_i / \det \boldsymbol{A}$ où $\boldsymbol{A}_i$ est la matrice de $\boldsymbol{A}$ en replaçant la i ème colonne par $\boldsymbol{b}$. La *complexité de l'algorithme* est le nombre d'opérations nécessaires pour calculer la solution. On a besoin de $(N+1) \times N!$ opérations pour la méthode de Cramer (le $N!$ vient du calcul du déterminant). Pour $N = 25$, et la vitesse de calcul d'un ordinateur est $1\ \mathrm{G/s}$, c'est à dire 10^9 opérations chaque seconde. Pour résoudre un système linéaire, il faut $26 \times 25! \approx 10$ milliards d'années (deux fois l'âge de la Terre).

I.3.b.ii. Élimination de Gauss

Elle consiste à trouver un système équivalent à $\boldsymbol{Ax} = \boldsymbol{b}$ mais le système est triangulaire supérieur :

$$\boldsymbol{A} \text{ matrice quelconque } \rightsquigarrow \tilde{\boldsymbol{A}} \text{ triangulaire supérieure}$$
$$\boldsymbol{Ax} = \boldsymbol{b} \iff \tilde{\boldsymbol{A}}\boldsymbol{x} = \tilde{\boldsymbol{b}}$$

I.3.b.iii. Méthode de Choleski

La méthode de Choleski consiste à décomposer la matrice $\boldsymbol{A}$ sous la forme $\boldsymbol{A} = \boldsymbol{L} \cdot \boldsymbol{U}$ où $\boldsymbol{L}$ est une matrice triangulaire supérieure avec $\ell_{i,i} = 1$ et $\boldsymbol{U}$ est une matrice triangulaire supérieure. Le calcul de ces

deux matrices est possible via un algorithme. Une fois ces matrices calculées, trouver la solution est facile à résoudre :

$$\boldsymbol{Ax} = \boldsymbol{b} \iff \boldsymbol{A}\underbrace{\boldsymbol{Ux}}_{\boldsymbol{y}} = \boldsymbol{b} \iff \begin{cases} \boldsymbol{Ly} = \boldsymbol{b} \\ \boldsymbol{Ux} = \boldsymbol{y} \end{cases}$$

I.4. Graphe associé à une matrice et inversement

Soit $\boldsymbol{A} = \left(a_{i,j}\right)_{i,j}$ une matrice carrée d'ordre N. À chaque colonne, on fait correspondre un sommet S_i pour $i = 1, N$. Un arc relie deux sommets s_i et s_j si $a_{i,j} \neq 0$. Un graphe est formé de l'ensemble des sommets et des arcs.

Exemple. On considère la matrice

$$\boldsymbol{A} = \begin{pmatrix} 4 & 3 & 0 & 0 \\ 0 & 2 & 1 & 2 \\ 0 & 1 & 0 & 3 \\ 6 & 5 & 0 & 0 \end{pmatrix}$$

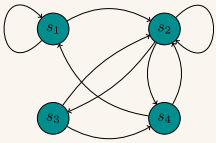

Figure 3: Graphe associé à la matrice $\boldsymbol{A}$

/!\ Le graphe donne la « structure creuse » de la matrice $\boldsymbol{A}$.

Un *chemin* allant de S_i à S_j est une suite d'arcs, si elle existe, telle que la suite

$$\left(s_i, s_{i_1}\right), \left(s_{i_1}, s_{i_2}\right), \ldots, \left(s_{i_p}, s_j\right)$$

soient des arcs du graphe.

Un graphe est *fortement connexe* s'il existe au moins un chemin de tout sommet s_i à tout sommet s_j.

Exemple. Avec $\boldsymbol{A} = \begin{pmatrix} 3 & 2 & 5 \\ 4 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$, le graphe n'est pas fortement connexe : il n'existe pas de chemin de s_3 à s_1, comme le montre la Figure 4.

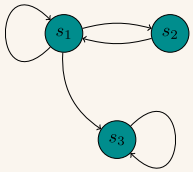

Figure 4: Graphe associé à la matrice $\boldsymbol{A}$

I.5. Les matrices irréductibles

Soit $\boldsymbol{A} \in \mathscr{M}_{N,N}(\mathbb{R})$ de la forme

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{1,2} \\ \boldsymbol{0} & \boldsymbol{A}_{2,2} \end{pmatrix} \quad \text{et} \quad \boldsymbol{x} = \begin{pmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \end{pmatrix}$$

Ainsi, il est équivalent :

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \iff \begin{cases} \boldsymbol{A}_{1,1}\boldsymbol{x}_1 = \boldsymbol{b}_1 - \boldsymbol{A}_{1,2}\boldsymbol{x}_2 & \rightsquigarrow \text{on obtient } \boldsymbol{x}_1 \\ \boldsymbol{A}_{1,2}\boldsymbol{x}_2 = \boldsymbol{b}_2 & \rightsquigarrow \text{on obtient } \boldsymbol{x}_2 \end{cases}$$

Le système est réduit à la résolution de deux systèmes de plus petite taille.

Définition. Une matrice $\boldsymbol{A}$ d'ordre N est réductible si, et seulement si, il existe $\boldsymbol{P}$ une matrice de permutation telle que :

$$\boldsymbol{B} = \boldsymbol{P}^\top \boldsymbol{A}\boldsymbol{P} = \begin{pmatrix} \boldsymbol{B}_{1,1} & \boldsymbol{B}_{1,2} \\ \boldsymbol{0} & \boldsymbol{B}_{2,2} \end{pmatrix},$$

si, et seulement si, il existe une permutation $\sigma : \{1, \dots, N\} \mapsto \{I, J\}$ où $b_{i,j} = 0$ pour tout $i \in I$, et $j \in J$.

Exemple. Avec la matrice $\boldsymbol{B}$ ci-dessous, on applique la permutation σ définie ci-dessous pour trouver la matrice $\widetilde{\boldsymbol{B}} = \boldsymbol{B}_\sigma = \left(b_{\sigma(i),\sigma(j)}\right)_{i,j}$.

$$\boldsymbol{B} = \begin{pmatrix} 30 & 0 & 20 & 0 \\ 11 & 10 & 6 & 5 \\ 13 & 0 & 11 & 0 \\ 4 & 2 & 3 & -1 \end{pmatrix} \quad \sigma : \begin{pmatrix} 1 \mapsto 4 \\ 2 \mapsto 2 \\ 3 \mapsto 3 \\ 4 \mapsto 1 \end{pmatrix} \quad \widetilde{\boldsymbol{B}} = \begin{pmatrix} \begin{pmatrix} -1 & 0 \\ 5 & 6 \end{pmatrix} & \begin{pmatrix} 3 & 4 \\ 10 & 11 \end{pmatrix} \\ \mathbf{0} & \begin{pmatrix} 11 & 13 \\ 20 & 30 \end{pmatrix} \end{pmatrix}$$

La matrice $\boldsymbol{B}$ est donc réductible.

Proposition (admis) $\boldsymbol{A}$ est irréductible si, et seulement si, le graphe de $\boldsymbol{A}$ est fortement connexe.

I.6. Localisation des valeurs propres

Théorème 1 (Gerschgörin-Hadamond) Soit λ une valeur propre de $\boldsymbol{A}$, alors

$$\lambda \in \bigcup_{k=1}^{N} \bar{D}_k \quad \text{avec} \quad \bar{D}_k = \left\{ z \in \mathbb{C}^2 \mid |z - a_{k,k}| \leq \sum_{\substack{j=1 \\ j \neq k}}^{N} |a_{k,j}| \right\}$$

Les valeurs propres de $\boldsymbol{A}$ appartiennent à l'union de N disques $\bar{D}_k$ de Gerschgörin.

Preuve. Soit $(\lambda, \boldsymbol{u})$ un élément propre de $\boldsymbol{A}$ tel que $\max_{i=1,N}|u_i| = |u_k| = 1$.
On a $\boldsymbol{Au} = \lambda \boldsymbol{u}$ donc $(\boldsymbol{Au})_k = \lambda u_k$.
Ainsi, $\sum_{i=0}^{N} a_{k,j} u_k = \lambda u_k$, ce qui est équivalent à, $(\lambda - a_{k,k}) u_k = \sum_{k \neq j} a_{k,j} u_k$.
Donc $| \lambda - a_{k,k} | \cdot |u_k| \leq \sum_{j \neq k} |a_{k,j}| \cdot |u_j|$.
Et donc $| \lambda - a_{k,k}| \leq \sum_{j \neq k} |a_{k,j}|$, *i.e.*, $\lambda \in \bar{D}_k$.

Exemple.

$$\boldsymbol{A} = \begin{pmatrix} 1+i & i & 2 \\ -3 & 2+i & 2 \\ 1 & i & 6 \end{pmatrix}$$

La matrice $\boldsymbol{A}$ a pour disques de Gerschgörin :

- $\bar{D}_1 = \{\lambda \mid |\lambda - (1+i)| \leq |i| + 2 = 3\}$,
- $\bar{D}_2 = \{\lambda \mid |\lambda - (2+i)| \leq |-3| + 1 = 4\}$,
- $\bar{D}_3 = \{\lambda \mid |\lambda - 6| \leq 1 + |i| = 2\}$.

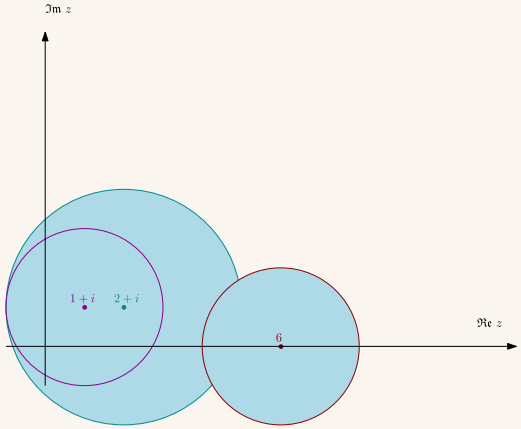

Figure 5: Disques de Gerschgörin pour la matrice $\boldsymbol{A}$

Dans la Figure 5, les valeurs propres se situent dans la zone bleue, correspondant à l'union des disques de Gerschgörin.

Si 0 n'est dans aucun disque, alors on sait que la matrice est inversible. On peut ainsi localiser les valeurs propres d'une matrice : on exclut tous les points en dehors des disques.

Théorème 2 (Gerschgörin-Hadamond) Soit λ une valeur propre de $\boldsymbol{A}$ et $\boldsymbol{A}$ une matrice irréductible[1]. Si une valeur propre λ est située sur la frontière de la réunion des disques alors tous les cercles de Gerschgörin passent par λ :

$$\lambda \in \partial\left(\bigcup_{k=1}^{N} \bar{D}_k\right) \Longrightarrow \lambda \in \bigcap_{k=1}^{N} \partial\bar{D}_k$$

On utilise souvent cela pour montrer que $\lambda = 0$ n'est pas valeur propre.

[1] *i.e.* le graphe est fortement connexe

II. Méthodes itératives de résolution de systèmes linéaires

II.1. Convergence des méthodes itératives

Le principe est d'engendrer une suite de vecteurs $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$, que l'on nomme les *itérés*, convergent vers la solution $\boldsymbol{x}$ du système linéaire $\boldsymbol{Ax} = \boldsymbol{b}$. La méthode itérative consiste à la procédure suivante.

> Soit $\boldsymbol{x}_0 \in \mathbb{R}^N$ (ou potentiellement $\mathbb{C}^N$). On engendre une suite $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ définies par :
>
> $$(\mathbf{1}) : \quad \boldsymbol{x}_{k+1} = \boldsymbol{B}\boldsymbol{x}_k + \boldsymbol{c},$$
>
> où $\boldsymbol{B} \in \mathscr{M}_{N\times N}$ est la *matrice d'itération* et $\boldsymbol{c} \in \mathbb{R}^N$.

Définition. Une méthode itérative de la forme $(\mathbf{1})$ est convergente si, et seulement si,

$$\forall x_0 \in \mathbb{R}^N, \quad \boldsymbol{x}_k \underset{k\to+\infty}{\longrightarrow} \boldsymbol{x} \quad \text{et} \quad \boldsymbol{Ax} = \boldsymbol{b}$$

La méthode est *consistante* si, et seulement si, si $\boldsymbol{x}_k \longrightarrow \boldsymbol{x}$, alors $\boldsymbol{x} = \boldsymbol{B}\boldsymbol{x} + \boldsymbol{c} \iff \boldsymbol{Ax} = \boldsymbol{b}$.

Définition. L'erreur d'approximation à la k-ième iétération est

$$\begin{aligned}\boldsymbol{e}_k = \boldsymbol{x}_k - \boldsymbol{x} &= \boldsymbol{B}\boldsymbol{x}_{k-1} + \boldsymbol{c} - \boldsymbol{B}\boldsymbol{x} + \boldsymbol{c} \\ &= \boldsymbol{B}(\boldsymbol{x}_{k-1} - \boldsymbol{x}) \\ &= \boldsymbol{B}\boldsymbol{e}_{k-1} \\ &= \boldsymbol{B}^2\boldsymbol{e}_{k-2}\end{aligned}$$

Donc $\boldsymbol{e}_k = \boldsymbol{B}^k\boldsymbol{e}_0$, où $\boldsymbol{e}_0 = \boldsymbol{x}_0 - \boldsymbol{x}$.

Définition. Une méthode itérative est convergente si, et seulement si,

$$\begin{aligned}\forall \boldsymbol{x}_0, \quad \boldsymbol{e}_k \underset{k\to+\infty}{\longrightarrow} \boldsymbol{0} &\iff \boldsymbol{B}^k \underset{k\to+\infty}{\longrightarrow} \boldsymbol{0} \\ &\iff \left\|\boldsymbol{B}^k\right\| \underset{k\to+\infty}{\longrightarrow} 0 \\ &\iff \boldsymbol{B}^k x \underset{k\to+\infty}{\longrightarrow} \boldsymbol{0}, \forall \boldsymbol{x} \in \mathbb{R}^N.\end{aligned}$$

Théorème (Convergence des méthodes itératives).

$$\lim_{k\to+\infty} \boldsymbol{B}^k = \boldsymbol{0} \iff \rho(\boldsymbol{B}) < 1.$$

où $\rho(\boldsymbol{B}) = \max|\lambda(\boldsymbol{B})|$.

Preuve.

$\Longrightarrow$. Si $\rho(\boldsymbol{B}) > 1$, il existe λ une valeur propre de $\boldsymbol{B}$ telle que $|\lambda| \geq 1$. Soit $\boldsymbol{x} \neq \boldsymbol{0}$ un vecteur propre associé à la valeur propre λ. Alors, $\boldsymbol{B}^k\boldsymbol{x} = \lambda^k\boldsymbol{x}$. Et, $|\lambda^k| \longrightarrow 1$ ou vers $+\infty$. On en déduit donc que $\boldsymbol{B}^k\boldsymbol{x} \longrightarrow \boldsymbol{0}$.

$\Longleftarrow$. On suppose $\rho(\boldsymbol{B}) < 1$, *i.e.* $|\lambda| < 1$ pour $\lambda \in \operatorname{Sp} \boldsymbol{B}$.

- Si $\boldsymbol{B}$ est diagonalisable, alors on pose $\boldsymbol{B} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^{-1}$ où $\boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$. Ainsi, $\boldsymbol{B}^k = \boldsymbol{P}\boldsymbol{\Lambda}^k\boldsymbol{P}^{-1}$. Et, on sait que $\boldsymbol{\Lambda}^k = \operatorname{diag}\left(\lambda_1^k, \ldots, \lambda_n^k\right)$. Comme $|\lambda_i| < 1$, on sait que $\lambda_i^k \longrightarrow 0$ et donc $\boldsymbol{B}^k \longrightarrow \boldsymbol{0}$.
- Cas général : toute matrice est semblable à une matrice de Jordan.

$$\widetilde{\boldsymbol{B}} = \boldsymbol{S}\boldsymbol{B}\boldsymbol{S}^{-1} = \operatorname{diag}(\boldsymbol{J}_1, \ldots, \boldsymbol{J}_r)$$

où $\boldsymbol{J}_k = \lambda_k\boldsymbol{I} + \boldsymbol{E}$, où $\boldsymbol{E} = \begin{pmatrix} \boldsymbol{0} & \boldsymbol{I} \\ 0 & \boldsymbol{0} \end{pmatrix}$ (matrice de surdiagonale de 1). Donc $\boldsymbol{J}_k^n \longrightarrow \boldsymbol{0}$ car $\boldsymbol{E}^k = 0$ à partir d'un certain rang, car la matrice est nilpotente. Ainsi, comme $\widetilde{\boldsymbol{B}}^k = \boldsymbol{S}\boldsymbol{B}^k\boldsymbol{S}^{-1}$, on a l'équivalence $\widetilde{\boldsymbol{B}}^k \longrightarrow \boldsymbol{0} \iff \boldsymbol{B}^k \longrightarrow \boldsymbol{0}$.

Corrolaire. Si $\|\boldsymbol{B}\| < 1$, alors $\rho(\boldsymbol{B}) < 1$ et la méthode converge donc.

(Ceci vient de $\rho(\boldsymbol{B}) \leq \|\boldsymbol{B}\| < 1$, montré en TD.)

II.2. Méthodes itératives classiques

II.2.a. Méthode de Jacobi

Soit $\boldsymbol{A}$ une matrice d'ordre n telle que $a_{i,i} \neq 0$, pour tout i. On décompose $\boldsymbol{A}$ sous la forme $\boldsymbol{A} = \boldsymbol{D} - \boldsymbol{E} - \boldsymbol{F}$, où

- $\boldsymbol{D}$ est la diagonale,
- $-\boldsymbol{E}$ est la partie inférieure stricte,[2]
- $-\boldsymbol{F}$ est la partie supérieure stricte,

Rermarque. On a l'équivalence suivante :

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \iff (\boldsymbol{D} - \boldsymbol{E} - \boldsymbol{F})\boldsymbol{x} = \boldsymbol{b} \iff \boldsymbol{D}\boldsymbol{x} = (\boldsymbol{E} + \boldsymbol{F})\boldsymbol{x} + \boldsymbol{b}.$$

Le principe de la *méthode de Jacobi* est :

$$\begin{cases} \boldsymbol{x}_0 \in \mathbb{R}^N \\ \boldsymbol{D}\boldsymbol{x}_{k+1} = (\boldsymbol{E} + \boldsymbol{F})\boldsymbol{x}_k + \boldsymbol{b} \end{cases} \iff \begin{cases} \boldsymbol{x}_0 \in \mathbb{R}^N \\ \boldsymbol{x}_{k+1} = \underbrace{\boldsymbol{D}^{-1}(\boldsymbol{E} + \boldsymbol{F})}_{\boldsymbol{J}}\boldsymbol{x}_k = \boldsymbol{D}^{-1}\boldsymbol{b}. \end{cases}$$

La matrice $\boldsymbol{J} = \boldsymbol{D}^{-1}(\boldsymbol{E} + \boldsymbol{F})$ est la matrice d'itération de Jacobi. On peut ré-écrire $\boldsymbol{J}$ comme $\boldsymbol{D}^{-1}(\boldsymbol{D} - \boldsymbol{A}) = \boldsymbol{I} - \boldsymbol{D}^{-1}\boldsymbol{A}$. Ainsi, la méthode de Jacobi converge si, et seulement si, $\rho(\boldsymbol{J}) < 1$.

Algorithme. Pour calculer les itérés de la méthode de Jacobi :

$$\boldsymbol{D}\boldsymbol{x}_{k+1}(\boldsymbol{E} + \boldsymbol{F})\boldsymbol{x}_k + \boldsymbol{b} \iff a_{i,i}(\boldsymbol{x}_{k+1})_i = -\sum_{j \neq i} a_{i,j}(\boldsymbol{x}_k)_j + b_i$$

Ainsi, $(\boldsymbol{x}_{k+1})_i = \left(b_i - \sum_{j \neq i} (\boldsymbol{x}_k)_j\right)$ pour tout i.

II.2.b. Méthode de Gauss-Seidel

Remarque. On décompose $\boldsymbol{A} = \boldsymbol{D} - \boldsymbol{E} - \boldsymbol{F}$. Il y a l'équivalence suivante :

$$(\boldsymbol{D} - \boldsymbol{E} - \boldsymbol{F})\boldsymbol{x} = \boldsymbol{b} \iff (\boldsymbol{D} - \boldsymbol{E})\boldsymbol{x} = \boldsymbol{F}\boldsymbol{x} + \boldsymbol{b}.$$

[2] On veut dire ici sans la diagonale.

Le principe de la méthode de Gauss-Seidel est : on pose $\boldsymbol{x}_0 \in \mathbb{R}^N$, et on itère avec $(\boldsymbol{D}-\boldsymbol{E})\boldsymbol{x}_{k+1} = \boldsymbol{F}\boldsymbol{x}_k + \boldsymbol{b}$. Sous forme matricielle, cela devient : $\boldsymbol{x}_{k+1} = \underbrace{(\boldsymbol{D}-\boldsymbol{E})^{-1}\boldsymbol{F}}_{\boldsymbol{G}}\boldsymbol{x}_k + (\boldsymbol{D}-\boldsymbol{E})^{-1}$.

La méthode de Gauss-Seidel converge si, et seulement si, $\rho(\boldsymbol{G}) < 1$. Pour calculer $\boldsymbol{x}_{k+1}$, il suffit d'appliquer l'algorithme de descente sur $\boldsymbol{D}-\boldsymbol{E}$. Pour $i = 1, N$, on a

$$a_{i,i}(\boldsymbol{x}_{k+1})_i = -\sum_{j=1}^{i-1} a_{i,j}(\boldsymbol{x}_{k+1})_j - \sum_{j=i+1}^{N} a_{i,j}(\boldsymbol{x}_k)_j + \boldsymbol{b}_i.$$

II.2.c. Méthode de relaxation

Soit $\omega \neq 0$. On décompose $\boldsymbol{A}$ en $\boldsymbol{A} = \left(\frac{\boldsymbol{D}}{\omega} - \boldsymbol{E}\right) + \left(\boldsymbol{D} - \frac{\boldsymbol{D}}{\omega} - \boldsymbol{F}\right)$.

Le principe de la méthode de relaxation s'écrit :

$$\left(\frac{\boldsymbol{D}}{\omega} - \boldsymbol{E}\right)\boldsymbol{x}_{k+1} = \left(\frac{1-\omega}{\omega}\boldsymbol{D} + \boldsymbol{F}\right)\boldsymbol{x}_k + \boldsymbol{b}$$
$$\iff (\boldsymbol{D} - \omega\boldsymbol{E})\boldsymbol{x}_{k+1} = ((1-\omega)\boldsymbol{D} + \omega\boldsymbol{F})\boldsymbol{x}_k + \omega\boldsymbol{b}.$$

Ainsi, la méthode de relaxation converge si, et seulement is, $\rho(\boldsymbol{L}_\omega) < 1$, où $\boldsymbol{L}_\omega = (\boldsymbol{D} - \omega\boldsymbol{E})^{-1}((1-\omega)\boldsymbol{D} + \omega\boldsymbol{F})$.

II.2.d. Méthode de Richardson

On écrit $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_k) = \underbrace{(\boldsymbol{I} - \alpha\boldsymbol{A})}_{R}\boldsymbol{x}_k + \alpha\boldsymbol{b} = \boldsymbol{R}\boldsymbol{x}_k + \boldsymbol{b}$.

C'est une méthode très simple, car il n'y a pas de système linéaire à résoudre. La recherche d'une valeur de α pour que la méthode soit convergente sera faite en TD.

II.2.e. Méthode du gradient à pas optimal

On écrit $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_k) = \overbrace{(\boldsymbol{I} - \alpha_k\boldsymbol{A})}^{\boldsymbol{G}_k}\boldsymbol{x}_k + \alpha_k\boldsymbol{b}$. On choisit α_k de façon optimale.

La matrice d'itération $\boldsymbol{G}_k = \boldsymbol{I} - \alpha_k \boldsymbol{A}$ *dépend* de k. On ne peut donc pas utiliser le théorème précédent pour montrer que la méthode converge (*c.f.* TD).

C'est la méthode la plus efficace dans le cas général.

Définition. Une matrice est dit
- *à diagonale strictement dominante* si, et seulement si

$$\forall i, \quad |a_{i,i}| > \sum_{j \neq i} |a_{i,j}|$$

- *à diagonale fortement dominante* si, et seulement si

$$\forall i, \quad |a_{i,i}| \geq \sum_{j \neq i} |a_{i,j}|$$

et il existe au moins un i_0 tel que $|a_{i_0,i_0}| > \sum_{j \neq i_0} |a_{i_0,j}|$.

Théorème.
(1) Soit $\boldsymbol{A}$ une matrice réelle symétrique définie positive (ou hermitienne définie positive dans $\mathbb{C}$). Alors, la méthode de relaxation converge si $0 < \omega < 2$.
(2) Soit $\boldsymbol{A}$ une matrice à diagonale strictement dominante (ou $\boldsymbol{A}$ une matrice à diagonale fortement dominante ***et*** irréductible). Alors,
   (a) la méthode de Jacobi converge.
   (b) la méthode de Gauss-Seidel converge.
   (c) la méthode de relaxation converge pour $0 < \omega \leq 1$.

Preuve. faite en TD.

Remarque. On peut montrer que $\rho(\boldsymbol{G}) < \rho(\boldsymbol{J})$. C'est à dire que, la méthode de Gauss-Seidel converge plus vite que celle de Jacobi. On pourra le vérifier en TP.

II.3. Calcul de valeurs et de vecteurs propres

II.3.a. Méthode de puissance itérées pour calculer la plus grande des valeurs propres en module

On pose une matrice exemple : $\boldsymbol{A} = \begin{pmatrix} 10 & 0 \\ -0 & 1 \end{pmatrix}$. Ses valeurs propres sont 10 et 1. Deux vecteurs propres associés sont $\begin{pmatrix} 1 & -1 \end{pmatrix}^\top$ et $\begin{pmatrix} 0 & 1 \end{pmatrix}^\top$. Soit $\boldsymbol{x}_0 = \begin{pmatrix} 1 & 1 \end{pmatrix}^\top$. On calcule

(1) $\boldsymbol{x}_1 = \boldsymbol{A}\boldsymbol{x}_0 = \begin{pmatrix} 10 & -8 \end{pmatrix}^\top$,
(2) $\boldsymbol{x}_2 = \boldsymbol{A}\boldsymbol{x}_1 = \begin{pmatrix} 100 & -98 \end{pmatrix}^\top$,
(3) $\boldsymbol{x}_3 = \boldsymbol{A}\boldsymbol{x}_2 = \begin{pmatrix} 1000 & -998 \end{pmatrix}^\top$,
(4) $\dots$

On voit que $\boldsymbol{x}_{k+1} \approx 10\boldsymbol{x}_k$ et $\boldsymbol{x}_{k+1} \parallel \begin{pmatrix} 1 & -1 \end{pmatrix}^\top$.

Algorithme. Pour le calcul des puissances itérées : soit $\boldsymbol{q}_0 \in \mathbb{C}^N$ tel que $\|\boldsymbol{q}_0\| = 1$; alors, pour $k = 1, 2, \dots$, on pose $\boldsymbol{x}_k = \boldsymbol{A}\boldsymbol{x}_{k-1}$ et $\boldsymbol{q}_k = \boldsymbol{x}_k / \|\boldsymbol{x}_k\|$.

Théorème. Soit $\boldsymbol{A}$ une matrice d'ordre N, diagonalisable dont la valeur propre du plus grand module λ_1 est unique et vérifie :

$$|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \cdots \geq |\lambda_N|.$$

Soit $q_0 \in \mathbb{C}^N$ choisi *convenablement*. Alors,

(1) $\lim_{k\to+\infty} (\lambda_i / |\lambda_1|)^k \boldsymbol{q}_k = \boldsymbol{q}$ est un vecteur propre de norme 1 associé à λ_1.
(2) $\lim_{k\to+\infty} \|\boldsymbol{A}\boldsymbol{q}_k\| = \lim_{k\to+\infty} \|\boldsymbol{x}_k\| = |\lambda_1|$
(3) $\lim_{k\to+\infty} (\boldsymbol{x}_{k+1})_j / (\boldsymbol{q}_k)_i = \lambda_1$ pour $1 \leq j \leq 1$ (?) tel que $(\boldsymbol{q}_k)_j \neq 0$.

Preuve. On suppose $\boldsymbol{A}$ diagonalisable, et $(\boldsymbol{u}_i)_i$ une base de vecteurs propres. On pose $\boldsymbol{q}_0 = \alpha_1 \boldsymbol{u}_1 + \sum_{i=2}^N \alpha_i \boldsymbol{u}_i$ en supposant $\alpha_k \neq 0$.

(1) On pose $\boldsymbol{q}_1 = \boldsymbol{x}_1 / \|\boldsymbol{x}_1\| = \boldsymbol{A}\boldsymbol{q}_0 / \left\|\boldsymbol{A}\boldsymbol{q}_0\right\|$. Par récurrence on montre que $\boldsymbol{q}_k = (\boldsymbol{A}^k \boldsymbol{q}_0) / \left\|\boldsymbol{A}^k \boldsymbol{q}_0\right\|$. Ainsi,

$$\begin{aligned} \boldsymbol{A}^k \boldsymbol{q}_0 &= \boldsymbol{A}^k \left(\alpha_1 \boldsymbol{u}_1 + \sum_{i\geq 2} \alpha_i \boldsymbol{u}_i \right) \\ &= \alpha_1 \boldsymbol{A}^k (\boldsymbol{u}_1 + \boldsymbol{e}_k) \end{aligned}$$

où $\boldsymbol{e}_k = \sum_{i\geq 2}(\alpha_i/\alpha_1)(\lambda_i/\lambda_1)^k \boldsymbol{u}_i$.

(2) Exercice

(3) Exercice

II.3.b. Méthode de la puissance inverse pour calculer la plus petite des valeurs propres en module

Les valeurs propres de $\boldsymbol{A}^{-1}$ sont $\mu_i = 1/\lambda_i$. On suppose que l'on peut ordonner les valeurs propres comme $|\lambda_1| \geq \cdots \geq |\lambda_{N-1}| > |\lambda_N|$. Ainsi, par passage à l'inverse, $1/\,|\lambda_1| \leq \cdots \leq 1/\,|\lambda_{N-1}| < 1/\,|\lambda_N|$. Autrement dit, $|\mu_1| \leq \cdots < |\mu_N|$.

Le réel $|\mu_N| = 1/\,|\lambda_N|$ est la plus grande des valeurs propres de $\boldsymbol{A}^{-1}$. On applique l'algorithme des puissances itérées sur $\boldsymbol{A}^{-1}$.

III. Optimisation sans contrainte

III.1. Introduction

L'optimisation consiste à traiter le ou les elements d'un ensemble « admissible » de données, pour trouver la meilleure solution, nommée solution optimale, qui minimise un certain critère.

Un *problème d'optimisation* s'écrit sous la forme suivante.

> Soit E un espace vectoriel normé (*evn*) de dimension finie n (ou non[3]). Soit K un sous-ensemble (en général convexe et fermé) de E, et soit $J : E \to \mathbb{R}$ une fonction.

On nomme E l'espace des états admissibles, K l'ensemble des constantes, et J la fonction de coût, ou fonction objectif.

Dans ce cours, on supposera $K = E = \mathbb{R}^N$. On dit que c'est un problème sans contraintes.

> ***Exemple.* (Problème du sac à dos)**
> On souhaite maximiser l'utilité du sac à dos sous la contrainte du poids. On liste n objets : une bouteille d'eau, une loupe, un briquet, ...
> On suppose que le sac à dos ait une contrainte maximale de poids P, par exemple 15 kg. On note
> - $p_1, p_2, \dots, p_n$ les poids respectifs des objets ;
> - $u_1, u_2, \dots, u_n$ les utilités respectives des objets ;
> - x_i valant 1 si on a l'objet i dans le sac à dos, et 0 sinon.
>
> Le problème d'optimisation est
> - $J(\boldsymbol{x}) = x_1u_1 + x_2u_2 + \dots + x_nu_n$,
> - $K = \{\boldsymbol{x} \in \mathbb{R}^n \mid x_1p_1 + x_2p_2 + \dots + x_np_n \leq P\}$.
>
> On cherche $x^\star$ tel que $J(\boldsymbol{x}^\star) = \max_{\boldsymbol{x}\in K} J(\boldsymbol{x})$.

[3] Espace fonctionnel $\mathscr{C}^0(\mathbb{R})$

Exemple. **(Détection d'un objet)**
Un objet est situé sur une droite et sa position est inconnue. On dispose de deux radars donnant les distances *entachées d'erreurs*.

Les radars mesurent approximativement les distances $a - s_1 \approx d_1$ et $s_2 - a \approx d_2$. Pour résoudre ce problème, on cherche à minimiser $J(a) = \min_{x \in \mathbb{R}} J(x)$, où

$$J(x) = |x - s_1 d_1|^2 + |s_2 - x - d_2|^2.$$

On dit que c'est un *problème quadratique*.

III.2. Dérivabilité.

Soient E et F deux espaces vectoriels normés, et $f : E \to F$.

Définition.
On dit que f est dérivable en $\boldsymbol{x}$ s'il existe $\ell_{\boldsymbol{x}} \in \mathscr{L}(E, F)$ telle que :

$$f(\boldsymbol{x} + \boldsymbol{h}) = f(\boldsymbol{x}) + \ell_{\boldsymbol{x}}(\boldsymbol{h}) + \|\boldsymbol{h}\|\varepsilon(\boldsymbol{h})$$

où $\varepsilon(\boldsymbol{h}) \to \mathbf{0}$ quand $\|\boldsymbol{h}\| \to 0$. On note alors $f'(\boldsymbol{x}) = \ell_{\boldsymbol{x}}$.

Exemple.
Dans le cas où $f : \mathbb{R}^N \to \mathbb{R}$ est définie par $f(\boldsymbol{x}) = \langle \boldsymbol{x} \mid \boldsymbol{x} \rangle = \sum_{i=1}^{n} x_i^2$. Alors $f'(\boldsymbol{x})\boldsymbol{h} = 2\langle \boldsymbol{x} \mid \boldsymbol{h} \rangle$.

Exemple.
Dans le cas où $E = \mathbb{R}^N$, on note $\frac{\partial f}{\partial x_i}$

Définition.
Pour $f : \mathbb{R}^N \to \mathbb{R}$ deux fois dérivable, on a

$$f''(\boldsymbol{x}) = \nabla^2 f(\boldsymbol{x})$$

est une matrice symétrique de $\mathscr{M}_{N,N}(\mathbb{R})$. On la note $\mathbf{H}_f$. À compléter

Formule de Taylor Lagrange.
Il existe $\theta \in]0,1[$ tel que

$$f(\boldsymbol{x}+\boldsymbol{h}) = f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}) \mid \boldsymbol{h} \rangle + \frac{1}{2}\langle \nabla^2 f(\boldsymbol{x}+\theta\boldsymbol{h})\boldsymbol{h} \mid \boldsymbol{h} \rangle.$$

Formule de Taylor Young.

$$f(\boldsymbol{x}+\boldsymbol{h}) = f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}) \mid \boldsymbol{h} \rangle + \frac{1}{2}\langle \nabla^2 f(\boldsymbol{x})\boldsymbol{h} \mid \boldsymbol{h} \rangle + \|\boldsymbol{h}\|^2 \varepsilon(\boldsymbol{h}),$$

où $\varepsilon(\boldsymbol{h}) \to \mathbf{0}$ quand $\|\boldsymbol{h}\| \to 0$.

Formule de Taylor avec reste intégral.

$$f(\boldsymbol{x}+\boldsymbol{h}) = f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}) \mid \boldsymbol{h} \rangle + \int_0^1 (1-t)\langle \nabla^2 f(\boldsymbol{x}+t\boldsymbol{h}) \mid \boldsymbol{h} \rangle \, \mathrm{d}t.$$

Formule de Taylor à l'ordre 1.

$$f(\boldsymbol{x}+\boldsymbol{h}) = f(\boldsymbol{x}) + \int_0^1 \langle \nabla f(\boldsymbol{x}+t\boldsymbol{h}) \mid \boldsymbol{h} \rangle \, \mathrm{d}t.$$

III.3. Existence des minimums sur $\mathbb{R}^N$

On note V un espace vectoriel normé de dimension finie, dans ce cas $V = \mathbb{R}^N$. On suppose $J : V \to \mathbb{R}$ continue. On cherche $\min_{\boldsymbol{v}\in V} J(\boldsymbol{v})$.

Remarque.
On cherche toujours le minimum. En effet, pour trouver le maximum, on peut se ramener à la recherche d'un minimum par :

$$\max_{\boldsymbol{v}\in V} J(\boldsymbol{v}) = -\min_{\boldsymbol{v}\in V}(-J(\boldsymbol{v})).$$

Définition (minimum local et global).

- On dit que $\boldsymbol{x}$ est un point de minimum *global* de J si $J(\boldsymbol{y}) \geq J(\boldsymbol{x})$, quel que soit $\boldsymbol{y} \in V$.
- On dit que $\boldsymbol{x}$ est un point de minimum *local* de J si $J(\boldsymbol{y}) \geq J(\boldsymbol{x})$, quel que soit $\boldsymbol{y}$ voisin de $\boldsymbol{x}$. C'est à dire :

$$\forall \boldsymbol{y} \in V, \quad [\exists \varepsilon > 0, \|\boldsymbol{y} - \boldsymbol{x}\| \leq \varepsilon] \Longrightarrow J(\boldsymbol{y}) \geq J(\boldsymbol{x}).$$

- On dit que $\boldsymbol{x}$ est un point de *selle* de J si
  - $\boldsymbol{x}$ est un minimum local dans une direction,
  - $\boldsymbol{x}$ est un maximum local dans une autre direction.

Définition (coercivité).
Soit $J : \mathbb{R}^n \to \mathbb{R}$ continue. On dit que J est *coercive* sur $\mathbb{R}^N$ si, $J(\boldsymbol{x}) \to +\infty$ quand $\|\boldsymbol{x}\| \to +\infty$. On dira que J est *infinie* à l'infini.

La définition précédente en précise pas la « direction » de la limite. Par exemple, dans $\mathbb{R}$, une fonction est coercive si, et seulement si, sa limite en $+\infty$ *et* en $-\infty$ est $+\infty$.

Exemples.
- La fonction J définie par $J(\boldsymbol{x}) = x_1^3 + x_2^2$ est non coercive. En effet, on a $J(x_1, 0) \to -\infty$ quand $x_1 \to -\infty$.
- La fonction J définie par $J(\boldsymbol{x}) = x_1^2 + x_2^2$ est coercive.

Théorème. (Existence d'un minimum)
Soit J une fonction continue et coercive sur $\mathbb{R}^n$. Alors J admet au moins un point de minimum global.

La preuve est sur le serveur pédagogique. C'est le seul résultat d'existence.

Remarque.
Si J n'est pas coercive, on ne peut pas conclure. En effet, voici les deux cas :
- J n'est pas coercive et admet un minimum,

- J n'est pas coercive et n'admet pas de minimum.

III.4. Points stationnaires, et caractérisation d'un minimum

Définition (point stationnaire).
Soit J dérivable en $\boldsymbol{x}^\star$. On dit que $\boldsymbol{x}^\star$ est un *point stationnaire* de J si $\nabla J(\boldsymbol{x}^\star) = \boldsymbol{0}$.

Théorème. (Caractérisation d'un minimum)[4]
Soit J dérivable en $\boldsymbol{x}^\star$, où $\boldsymbol{x}^\star$ est un point d'extremum local de J, alors $\nabla J(\boldsymbol{x}^\star) = \boldsymbol{0}$. Un *extremum local* est un minimum ou un maximum local.

Preuve.
Soit $\varphi : t \mapsto J(\boldsymbol{x}^\star + t\boldsymbol{y})$ est définie au voisinage de $\boldsymbol{x}^\star$. On a $\varphi'(t) = \langle \nabla J(\boldsymbol{x}^\star + t\boldsymbol{y}) \mid \boldsymbol{y} \rangle$. Ainsi, en $t = 0$, on a donc

$$\varphi'(0) = \langle \nabla J(\boldsymbol{x}^\star) \mid \boldsymbol{y} \rangle.$$

- Si $\boldsymbol{x}^\star$ est un minimum local, on a donc $J(\boldsymbol{x}^\star + t\boldsymbol{y}) \geq J(\boldsymbol{x}^\star)$ et donc $J(\boldsymbol{x}^\star + t\boldsymbol{y}) - J(\boldsymbol{x}^\star) \geq 0$. À finir ...
- De même si $\boldsymbol{x}^\star$ est un maximum local.

III.4.a. Étude de la nature d'un point stationnaire.

Soit $\boldsymbol{x}^\star$ un point stationnaire. On suppose que J est de classe $\mathscr{C}^2$. On écrit le théorème de Taylor Young à l'ordre 2. Soit $\boldsymbol{h} \in \mathbb{R}^n$. On a :

$$J(\boldsymbol{x}^\star + \boldsymbol{h}) = J(\boldsymbol{x}^\star) + \underbrace{\langle \nabla J(\boldsymbol{x}^\star) \mid \boldsymbol{h} \rangle}_{=0} + \frac{1}{2}\langle \nabla^2 J(\boldsymbol{x}^\star)\boldsymbol{h} \mid \boldsymbol{h} \rangle + \|\boldsymbol{h}\|^2 \varepsilon(\boldsymbol{h}).$$

On a $\nabla^2 J(\boldsymbol{x}^\star)$ qui est symétrique et réelle. D'après le quotient de Rayleigh, on a

$$\lambda^\star_{\min} \|\boldsymbol{h}\|^2 \leq \langle \nabla^2 J(\boldsymbol{x}^\star)\boldsymbol{h} \mid \boldsymbol{h} \rangle \leq \lambda^\star_{\max} \|\boldsymbol{h}\|^2,$$

[4] On nomme aussi ce théorème « l'équation d'Euler ».

où $\lambda^\star_{\min}$ et $\lambda^\star_{\max}$ sont la plus petite et la plus grande valeur propre des valeurs propres de $\nabla^2 J(\boldsymbol{x}^\star)$. On en déduit que

$$(\mathbf{E}) : \|\boldsymbol{h}\|^2 \left(\frac{1}{2}\lambda^\star_{\min} + \varepsilon(\boldsymbol{h})\right) \leq J(\boldsymbol{x}^\star + \boldsymbol{h}) - J(\boldsymbol{x}^\star) \leq \|\boldsymbol{h}\|^2 \left(\frac{1}{2}\lambda^\star_{\max} + \varepsilon(\boldsymbol{h})\right).$$

Conclusion.

(1) Si $\lambda^\star_{\min} > 0$, alors toutes les valeurs propres sont positives, et pour $\boldsymbol{h}$ petit, on a $\frac{1}{2}\lambda^\star_{\min} + \varepsilon(\boldsymbol{h}) > 0$ et donc $J(\boldsymbol{x}^\star + \boldsymbol{h}) - J(\boldsymbol{x}^\star) \geq 0$. C'est donc un minimum local.

(2) Si $\lambda^\star_{\max} < 0$, alors toutes les valeurs propres sont négatives, et pour $\boldsymbol{h}$ petit, on a $\frac{1}{2}\lambda^\star_{\max} + \varepsilon(\boldsymbol{h}) < 0$ et donc $J(\boldsymbol{x}^\star + \boldsymbol{h}) - J(\boldsymbol{x}^\star) \leq 0$. C'est donc un maximum local.

(3) Si les valeurs propres sont de signes différents, et non nulles, que l'on notera λ_+ et λ_-. Alors, pour $\lambda_+ > 0$, soit $\boldsymbol{u}$ un vecteur propre associé à λ_+. Alors, on pose $\boldsymbol{h} = t\boldsymbol{u}$ pour t petit. Ainsi, $J(\boldsymbol{x}^\star + t\boldsymbol{u}) - J(\boldsymbol{x}^\star) = \frac{t^2}{2}\lambda_+\|\boldsymbol{u}\|^2 + t^2\|\boldsymbol{u}\|^2\varepsilon(t\boldsymbol{u}) > 0$ pour t petit. Ainsi $\boldsymbol{x}^\star$ est un minimum local.

(4) Si $\lambda = 0$, alors $J(\boldsymbol{x}^\star + t\boldsymbol{u}) - J(\boldsymbol{x}^\star) = t^2\|\boldsymbol{u}\|^2\varepsilon(t\boldsymbol{u})$, qui n'a pas de signe. On ne peut donc pas conclure.

On a donc montré le théorème suivant.

Théorème.
Soit J de classe $\mathscr{C}^2$ sur $\mathbb{R}^n$, et $\boldsymbol{x}^\star$ un point stationnaire de J. On a alors les situations suivantes.

Signe des valeurs propres de $\nabla^2 J(\boldsymbol{x}^\star)$	Nature du point stationnaire
$\forall i \in [\![1, n]\!], \lambda_i > 0$	$\boldsymbol{x}^\star$ est un minimum local
$\forall i \in [\![1, n]\!], \lambda_i < 0$	$\boldsymbol{x}^\star$ est un maximum local
$\forall i \in [\![1, n]\!], \lambda_i \neq 0$ de signes différents	$\boldsymbol{x}^\star$ est un point de selle
$\exists i \in [\![1, n]\!], \lambda_i = 0$	$\boldsymbol{x}^\star$ est un point dégénéré

III.5. Convexité

Les définitions de la convexité ci-dessous sont équivalentes.

Définition.
Soit J continue sur $\mathbb{R}^N$
(1) On dit que J est *convexe* si

$$\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n, \forall t \in]0,1[, J(t\boldsymbol{x} + (1-t)\boldsymbol{y}) \leq tJ(\boldsymbol{x}) + (1-t)J(\boldsymbol{y}).$$

(1) On dit que J est *convexe* si l'inégalité ci-dessus est stricte :

$$\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n, \forall t \in]0,1[, J(t\boldsymbol{x} + (1-t)\boldsymbol{y}) < tJ(\boldsymbol{x}) + (1-t)J(\boldsymbol{y}).$$

Autrement dit,
(1) J est convexe si elle est au dessous de toutes ses cordes.
(2) J est strictement convexe si elle est strictement au dessous de toutes ses cordes.

Définition.
Soit J de classe $\mathscr{C}^1$ sur $\mathbb{R}^N$. On dit que J est *convexe* si

$$\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n, \quad \langle \nabla J(\boldsymbol{y}) - \nabla J(\boldsymbol{x}) \mid \boldsymbol{y} - \boldsymbol{x} \rangle \geq 0$$

i.e.

$$\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n, \quad J(\boldsymbol{y}) \geq J(\boldsymbol{x}) + \langle \nabla J(\boldsymbol{x}) \mid \boldsymbol{y} - \boldsymbol{x} \rangle$$

Définition.
Soit J de classe $\mathscr{C}^1$ sur $\mathbb{R}^N$. On dit que J est α-*convexe* s'il existe $\alpha > 0$ tel que :

$$J(\boldsymbol{y}) \geq J(\boldsymbol{x}) + \langle \nabla J(\boldsymbol{x}) \mid \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{\alpha}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2$$

ou

$$\langle \nabla J(\boldsymbol{y}) - \nabla J(\boldsymbol{x}) \mid \boldsymbol{y} - \boldsymbol{x} \rangle \geq \alpha \|\boldsymbol{x} - \boldsymbol{y}\|^2.$$

Définition.
Soit J de classe $\mathscr{C}^2$ sur $\mathbb{R}^N$. On dit que :

(1) J est convexe si $\langle \nabla^2 J(\boldsymbol{x})\boldsymbol{y} \mid \boldsymbol{y}\rangle \geq 0$, quels que soient $\boldsymbol{x}$ et $\boldsymbol{y}$, *i.e.* que toutes les valeurs propres se $\nabla^2 J(\boldsymbol{x}$ sont positives ou nulles ;
(2) J est strictement convexe si $\langle \nabla^2 J(\boldsymbol{x})\boldsymbol{y} \mid \boldsymbol{y}\rangle > 0$, quels que soient $\boldsymbol{x}$ et $\boldsymbol{y}$, *i.e.* que toutes les valeurs propres se $\nabla^2 J(\boldsymbol{x})$ sont strictement positives ;
(3) J est α-convexe si $\langle \nabla^2 J(\boldsymbol{x})\boldsymbol{y} \mid \boldsymbol{y}\rangle \geq \alpha\|\boldsymbol{y}\|$, quels que soient $\boldsymbol{x}$ et $\boldsymbol{y}$, *i.e.* que toutes les valeurs propres se $\nabla^2 J(\boldsymbol{x})$ sont supérieurs ou égales à α.

Attention, α de dépend pas de $\boldsymbol{x}$.

Proprositions.
Soit J une fonction α-convexe sur $\mathbb{R}^n$. Alors, J est strictement convexe et coercive.

Théorème.

(1) Soit J une fonction convexe. Alors, si $\boldsymbol{x}^\star$ est un minimum de J, alors $\boldsymbol{x}^\star$ est un minimum global.

« Si J est convexe, alors tout minimum local est global. »

(2) Soit J une fonction strictement convexe. Alors, J a au plus un minimum global.

« Si J est strictement convexe, alors tout minimum local est unique et global. »

(3) Soit J une fonction strictement convexe et coercive. Alors, J a un, et un seul, minimum global.

(4) Soit J une fonction α-convexe. Alors, J a un, et un seul, minimum global.

Attention, on n'a pas l'implication J convexe $\not\Longrightarrow$ $\boldsymbol{x}^\star$ existe. Par exemple, $\exp : x \mapsto e^x$ est convexe mais n'a pas de minimum global.

IV. Optimisation non linéaire avec contraintes

La forme générale du problème est :

$$\inf_{\boldsymbol{v} \in K} J(\boldsymbol{v})$$

où $K \subset V$ est l'ensemble des contraintes, $J : K \to \mathbb{R}$ et $V = \mathbb{R}^n$.[5]

IV.1. Existence d'un minimum

Théorème (Existence dans $K \subset \mathbb{R}^n$).
Soient $K \subset \mathbb{R}^n$, où K est fermé et non vide, et $J : K \to \mathbb{R}$ continue.
On suppose, de plus,
(1) ou bien K est borné,
(2) ou bien J est coercive sur K (*i.e.* $J(\boldsymbol{x}) \to +\infty$ quand $\|\boldsymbol{x}\| \to +\infty$).

Alors, il existe au moins un minimum sur K.

Preuve.

(1) Soit K fermé et bornée, donc K est compact. Comme J est continue, alors J atteint ses bornes sur K. Il y a donc au moins un minimum et un maximum de J sur K.

(2) On suppose K fermé et J coercive. Soit $(\boldsymbol{u}_n)_{n \in \mathbb{N}}$ une suite minimisante de J sur K. Alors, $\boldsymbol{u}_n \in K$ et $J(\boldsymbol{u}_n) \to \inf_K J$. La suite $(\boldsymbol{u}_n)$ est bornée car sinon, il existe une sous suite $\left(\boldsymbol{u}_{\varphi(n)}\right)$ telle que $J(\boldsymbol{u}_{\varphi(n)} \to +\infty$, ce qui contredit que $J(\boldsymbol{u}_n) \to \inf_K J$. D'où, la suite $(\boldsymbol{u}_n)$ est bornée et dans un fermé. On peut en extraire une sous-suite qui converge vers $\boldsymbol{u}$ dans K (car K fermé). Par continuité de J, on a $J(\boldsymbol{u}_n) \to J(\boldsymbol{u}) = \inf_K J$ quand n tend vers $+\infty$.

□

IV.2. Unicité du minimum

On s'intéresse à la convexité de l'ensemble K.

[5] On peut choisir un espace vectoriel mini d'un produit scalaire et complet.

Définition (K convexe).
On dit que K est convexe dans $V = \mathbb{R}^n$ si, pour tout couple de vecteurs $(\boldsymbol{x}, \boldsymbol{y}) \in K^2$, alors $t\boldsymbol{x} + (1-t)\boldsymbol{y} = \boldsymbol{z} \in K$ pour $t \in [0,1]$.

Théorème 2 (minimum local → minimum global).
Soient K convexe sur V, $J : K \to \mathbb{R}$ continue et convexe sur K. Alors, un minimum local de J sur K est un minimum global de J sur K.

Preuve.
Soit $\boldsymbol{x}$ un minimum local de J sur K, *i.e.* $J(\boldsymbol{z}) \geq J(\boldsymbol{x})$ pour $\boldsymbol{z} \in K$ et $\boldsymbol{z}$ proche de $\boldsymbol{x}$ (*i.e.* $\exists \eta > 0, \|\boldsymbol{z} - \boldsymbol{x}\| \leq \eta$). Soit $\boldsymbol{y} \in K$, alors

$$\boldsymbol{z} = \boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x}) = t\boldsymbol{y} + (1-t)\boldsymbol{x} \in K,$$

pour $t \in [0,1]$, car K est convexe. Pour t petit, on considère que $\boldsymbol{z}$ est un voisin de $\boldsymbol{x}$, et donc

$$J(\boldsymbol{x}) \leq J(\boldsymbol{z}) = J(t\boldsymbol{y} + (1-t)\boldsymbol{x}) \leq tJ(\boldsymbol{y}) + (1-t)J(\boldsymbol{x}),$$

car J est convexe. Ainsi,

$$tJ(\boldsymbol{y}) + (1-t)J(\boldsymbol{x}) - J(\boldsymbol{x}) = t(J(\boldsymbol{y}) - J(\boldsymbol{x})) \geq 0$$

et on a $t > 0$, car $\boldsymbol{y} \neq \boldsymbol{x}$. Ceci est vrai quel que soit $\boldsymbol{y}$:

$$\forall \boldsymbol{y} \in K, J(\boldsymbol{y}) \geq J(\boldsymbol{x}).$$

On en déduit que $\boldsymbol{x}$ est un minimum global. □

Attention. Ce théorème ne montre ni l'unicité, ni l'existence.

Théorème 3 (unicité).
Soient K convexe sur V, $J : K \to \mathbb{R}$ continue et strictement convexe sur K. Alors, J admet au plus un minimum sur K. C'est à dire, un minimum local est unique.

Preuve.
On suppose J strictement convexe. Supposons que J admet deux mi-

nima globaux $\boldsymbol{x}_1$ et $\boldsymbol{x}_2$. Alors, $J(\boldsymbol{x}_1) = J(\boldsymbol{x}_2) = m$. Si $\boldsymbol{x}_1 \neq \boldsymbol{x}_2$, alors $\boldsymbol{z} = t\boldsymbol{x}_1 + (1-t)\boldsymbol{x}_2 \in K$, pour $t \in [0,1]$. Ainsi,

$$\begin{aligned} J(\boldsymbol{x}_1) = m \leq J(\boldsymbol{z}) &= J(t\boldsymbol{x}_1 + (1-t)\boldsymbol{x}_2) \\ &< tJ(\boldsymbol{x}_1) + (1-t)J(\boldsymbol{x}_2) \\ &= tm + (1-t)m \\ &= m \end{aligned}$$

Ceci démontre $m < m$, ce qui est *absurde*.

On en déduit $\boldsymbol{x}_1 = \boldsymbol{x}_2$. □

Proposition (Existence *et* unicité)**.**
Soit K un convexe, fermé, non vide de $V = \mathbb{R}^n$. Soit $J : K \to \mathbb{R}$ continue. Si de plus,
(1) ou bien J est strictement convexe et coercive sur K,
(2) ou bien J est α-convexe sur K.

Alors J admet un unique minimum global sur K.

Preuve.
On regroupe les théorèmes précédents. □

IV.3. Caractérisation du minimum sur un convexe

Exemple. On considère $f : x \mapsto x^2$ sur $\mathbb{R}$, avec $K = [1,2]$. On a $f'(x) = 2x$ et donc $f''(x) = 2 > 0$, d'où f est α-convexe. Ainsi $x^\star = 0$ est le minimum de f sur $\mathbb{R}$. Sur $K = [1,2]$, on a $\min_K f$ est atteint en $x^\star = 1$. On peut remarquer que $f'(1) = 2 \neq 0$. L'égalité d'Euler $\nabla f(x^\star) = 0$ n'est pas valide sur K.

Théorème (condition d'optimalité sur un convexe)**.**
Soient K convexe non vide de V, et $J : V \to \mathbb{R}$ continue (ou sur un ouvert contenant K), et dérivable en $\boldsymbol{x}^\star \in K$.
(1) *Condition nécessaire.* Si $\boldsymbol{x}^\star$ est un minimum local de J sur K alors $\boldsymbol{x}^\star$ vérifie l'*inégalité d'Euler* :

$$\forall \boldsymbol{x} \in K, \langle \nabla J(\boldsymbol{x}^\star) \mid \boldsymbol{x} - \boldsymbol{x}^\star \rangle \geq 0.$$

(2) *Condition nécessaire et suffisante.* Si, de plus, K est convexe sur K, alors $\boldsymbol{x}^\star$ est un minimum si, et seulement si, l'inégalité d'Euler est vérifiée. Dans ce cas, $\boldsymbol{x}^\star$ est un minimum global.

Preuve.

(1) Soit $\boldsymbol{x}^\star \in K$. On a $J(\boldsymbol{z}) \geq J(\boldsymbol{x}^\star)$, quel que soit $\boldsymbol{z}$ voisin de $\boldsymbol{x}^\star$. Soit $\boldsymbol{x} \in K$, alors $\boldsymbol{z} = \boldsymbol{x}^\star + t(\boldsymbol{x} - \boldsymbol{x}^\star)$, pour $0 < t < 1$. La fonction J est dérivable en $\boldsymbol{x}^\star$. On applique la formule de Taylor-Young :

$$\begin{aligned} J(\boldsymbol{z}) &= J(\boldsymbol{x}^\star) + \langle \nabla J(\boldsymbol{x}^\star) \mid \boldsymbol{z} - \boldsymbol{x}^\star \rangle + \|\boldsymbol{z} - \boldsymbol{x}^\star\| \cdot \varepsilon(\boldsymbol{z} - \boldsymbol{x}^\star) \\ &= J(\boldsymbol{x}^\star) + t \cdot \langle \nabla J(\boldsymbol{x}^\star) \mid \boldsymbol{x} - \boldsymbol{x}^\star \rangle + \|\boldsymbol{z} - \boldsymbol{x}^\star\| \cdot \varepsilon(t(\boldsymbol{x} - \boldsymbol{x}^\star)) \end{aligned}$$

Alors,

$$\begin{aligned} \langle \nabla J(\boldsymbol{x}^\star) \mid \boldsymbol{x} - \boldsymbol{x}^\star \rangle &= \frac{1}{t}\overbrace{(J(\boldsymbol{z}) - J(\boldsymbol{x}^\star))}^{\geq 0} - \|\boldsymbol{x} - \boldsymbol{x}^\star\| \cdot \varepsilon(t(\boldsymbol{x} - \boldsymbol{x}^\star)) \\ &\geq -\|\boldsymbol{x} - \boldsymbol{x}^\star\| \cdot \varepsilon(t(\boldsymbol{x} - \boldsymbol{x}^\star)) \end{aligned}$$

On en déduit que

$$\forall \boldsymbol{x} \in K, \langle \nabla J(\boldsymbol{x}^\star) \mid \boldsymbol{x} - \boldsymbol{x}^\star \rangle \geq 0.$$

(2) Supposons J convexe. Ainsi, $J(\boldsymbol{x}) \geq J(\boldsymbol{x}^\star) + \overbrace{\langle \nabla J(\boldsymbol{x}^\star) \mid \boldsymbol{x} - \boldsymbol{x}^\star \rangle}^{\geq 0}$. Alors, si $\langle \nabla J(\boldsymbol{x}^\star) \mid \boldsymbol{x} - \boldsymbol{x}^\star \rangle \geq 0$, alors $J(\boldsymbol{x}) \geq J(\boldsymbol{x}^\star)$ quel que soit $\boldsymbol{x} \in K$. D'où, $\boldsymbol{x}^\star$ est un minimum global de J sur K.

□

Remarques.

(1) *Minimum de $f(x) = x^2$ sur $[1, 2]$.* L'inégalité d'Euler s'écrit $f'(x^\star) \cdot (x - x^\star) \geq 0$, quel que soit $x \in [0, 1]$ et pour $x^\star \in [1, 2]$. Cette inégalité est vérifiée si, et seulement si $2x^*(x - x^*) \geq 0$, d'où $x^\star = 1$.

(2) Si $K = V = \mathbb{R}^n$, alors $\nabla J(\boldsymbol{x}^\star) = \boldsymbol{0}$.

(3) Si $\boldsymbol{x}^\star \in \overset{\circ}{K}$, l'intérieur de K, alors $\langle \nabla J(\boldsymbol{x}^\star) \mid \boldsymbol{x} - \boldsymbol{x}^\star \rangle = 0$.

(4) Si $\boldsymbol{x}^\star \in \partial K$, le bord de K, alors

$$\langle \nabla J(\boldsymbol{x}^\star) \mid \boldsymbol{x} - \boldsymbol{x}^\star \rangle = \|\nabla \dots\|.$$

IV.4. Multiplicateurs de Lagrange

IV.4.a. Cas des contraintes d'égalité

On pose $V = \mathbb{R}^n$ et $K = \{\boldsymbol{x} \in \mathbb{R}^n \mid \forall i \in \{1, \dots, p\}, h_i(\boldsymbol{x}) = 0\}$, il y a donc p contraintes. Pour chaque contrainte, on associe un multiplicateur de Lagrange λ_i, on définit le lagrangien par

$$L(\boldsymbol{x}, \boldsymbol{\lambda}) = J(\boldsymbol{x}) + \sum_{i=1}^{n} \lambda_i h_i(\boldsymbol{x}).$$

Théorème KKT (Contrainte d'égalité)**.**
On suppose que J et les h_i sont $\mathscr{C}^1$ sur $\mathbb{R}^n$. On supposera également que, pour $\boldsymbol{x}^\star \in K$, les $(\nabla h_i(\boldsymbol{x}^\star))$ sont linéairement indépendants.

(1) Si $\boldsymbol{x}^\star$ est un minimum global de J sur K alors il existe un vecteur de multiplicateurs $\boldsymbol{\lambda}^\star = (\lambda_i^\star) \in \mathbb{R}^n$ tel que

$$(\mathbf{KKT}) : \begin{cases} \nabla J(\boldsymbol{x}^\star) + \sum_{i=1}^{n} \lambda_i^\star \cdot \nabla h_{i(\boldsymbol{x}^\star)} = 0 \\ \forall i \in \{1, \dots, p\}, \quad h_i(\boldsymbol{x}^\star) = 0 \end{cases}$$

Il faut résoudre un système de $n + p$ équations.

(2) Si J est convexe sur K et K convexe. Les conditions $(\mathbf{KKT})$ sont suffisantes.

Preuve.
On fait la démonstration dans le cas linéaire :

$$h_i(\boldsymbol{x}) = \sum_{j=1}^{n} a_{i,j} x_j - b_i = (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})_i \iff h(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}.$$

où $\boldsymbol{A} = (a_{i,j}) \in \mathscr{M}_{p,n}$ et $i \in \{1, \dots, p\}$, avec $p \leq n$. Ainsi,

$$\nabla h_i(\boldsymbol{x}) = \frac{\partial h_i}{\partial x_j} = \begin{pmatrix} a_{i,1} \\ \vdots \\ a_{i,n} \end{pmatrix}$$

On a $K = \{\boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}\}$.

...

On a $\boldsymbol{x}^\star$ qui vérifie $\langle \nabla J(\boldsymbol{x}^\star) \mid \boldsymbol{x} - \boldsymbol{x}^\star \rangle \geq 0$, que l que soit $\boldsymbol{x} \in K$. On a $\boldsymbol{A}\boldsymbol{x}^\star = \boldsymbol{b}$ et $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ d'où $\boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}^\star) = 0$. On a $\boldsymbol{x} - \boldsymbol{x}^\star \in K_0$ quel que soit $\boldsymbol{x} \in K$. Alors, $\langle \nabla J(\boldsymbol{x}^\star) \mid \boldsymbol{y} \rangle \geq 0$, quel que soit $\boldsymbol{y} \in K_0$. Or, $-\boldsymbol{y} \in K_0$ et donc $\langle \nabla J(\boldsymbol{x}^\star) \mid \boldsymbol{y} \rangle \leq 0$. On en déduit $\langle \nabla J(\boldsymbol{x}^\star) \mid \boldsymbol{y} \rangle = 0$ quel que soit $\boldsymbol{y} \in K_0 = \ker \boldsymbol{A}$. Ainsi, $\nabla J(\boldsymbol{x}^\star) \perp \ker \boldsymbol{A}$. D'où, $\nabla J(\boldsymbol{x}^\star) \in (\ker \boldsymbol{A})^\perp$. Or, $(\ker \boldsymbol{A})^\perp = \operatorname{im} \boldsymbol{A}^\top$. Et donc $\nabla J(\boldsymbol{x}^\star) \in \operatorname{im} \boldsymbol{A}^\top$ d'où, $\boldsymbol{A}^\top \in \mathscr{M}_{p,n}$. Il existe $\tilde{\boldsymbol{\lambda}} \in \mathbb{R}^p$ tel que $\nabla J(\boldsymbol{x}^\star) = \boldsymbol{A}^\top \tilde{\boldsymbol{\lambda}}$. Soit encore $\nabla J(\boldsymbol{x}^\star) + \boldsymbol{A}^\top \boldsymbol{\lambda}^\star = \boldsymbol{0}$ et $\boldsymbol{\lambda}^* = \tilde{\boldsymbol{\lambda}}$. On a

$$\left(\boldsymbol{A}^\top \boldsymbol{\lambda}^\star\right)_j = \sum_{i=1}^{p} a_{j,i}^\top \lambda_i^\star = \sum_{i=1}^{p} a_{i,j}^\top \lambda_i^\star = \sum_{i=1}^{n} \frac{\partial h_i}{\partial x_j} \lambda_i^\star$$

Enfin $\boldsymbol{A}^\top \boldsymbol{\lambda}^\star = \sum_{i=1}^{p} \lambda_i^\star \nabla h_i$. On a un système non linéaire à résoudre : trouver $\boldsymbol{\lambda}^\star \in \mathbb{R}^p$ et $\boldsymbol{x}^\star \in K$ solution du système $(\mathbf{KKT})$. □

Exemple. On calcule $\min_{x+y=1} x^2 + y^2 - 4$. On pose $h(x, y) = x + y + 1 = 0$; On a un seul multiplicateur de Lagrange : on cherche $\boldsymbol{\lambda}^\star$ et $\boldsymbol{x}^\star = (x^\star, y^\star)$ tel que

$$\begin{cases} \nabla J(\boldsymbol{x}^\star) + \boldsymbol{\lambda}^\star \nabla h(\boldsymbol{x}^\star) = 0 \\ h(\boldsymbol{x}^\star) = 0 \end{cases} \Longrightarrow x = y = -\frac{1}{2}$$

et $\lambda = 1$.

IV.4.b. Multiplicateurs de Lagrange – cas général

On considère des contraintes d'égalité et d'inégalité :

$$K = \left\{\boldsymbol{x} \in \mathbb{R}^n \mid \forall i \in \{1, \dots, p\}, h_i(\boldsymbol{x}) = 0 \text{ et } \forall j \in \{1, \dots, q\}, g_j(\boldsymbol{x}) \leq 0\right\}.$$

On suppose

- J, les h_i et les g_j de classe $\mathscr{C}^1$;

- pour $\boldsymbol{x}^\star \in K$, $(\nabla h_i(\boldsymbol{x}^\star))$ sont linéairement indépendants ;
- *constantes qualifiées*, il existe $\boldsymbol{d} \neq \boldsymbol{0}$ un vecteur de $\mathbb{R}^n$ tel que l'on ait $\langle \nabla h_i(\boldsymbol{x}^\star) \mid \boldsymbol{d} \rangle = 0$ et $\langle \nabla g_i(\boldsymbol{x}^\star) \mid \boldsymbol{d} \rangle < 0$.

Le multiplicateur de Lagrange est donc

$$L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = J(\boldsymbol{x}) + \sum_{i=1}^{p} \lambda_i h_i(\boldsymbol{x}) + \sum_{j=1}^{q} \mu_j g_j(\boldsymbol{x}).$$

Le nom **KKT** vient des noms Karuch, Kuhn et Tucker.

Théorème KKT.

(1) Si $\boldsymbol{x}^\star$ est un minimum local de J sur K, alors il existe $\boldsymbol{\lambda}^\star$ et $\boldsymbol{\mu}^\star$ qui vérifie :
   (a) $L(\boldsymbol{x}^\star, \boldsymbol{\lambda}, \boldsymbol{\mu}) = 0$,
   (b) $h_i(\boldsymbol{x}^\star) = 0$,
   (c) $\mu_j^* g_j(\boldsymbol{x}^\star) = 0$,
   (d) $\mu_j^\star \geq 0$,
   (e) $g_j(\boldsymbol{x}^\star) \geq 0$.

   On a $n + p + p$ équations à résoudre.

(2) Si, de plus, J est convexe, et les g_j sont convexes. Alors les conditions sur des conditions nécessaires et suffisantes, et $\boldsymbol{x}^\star$ est un minimum global.

Remarque. $\mu_j^\star g_j(\boldsymbol{x}^\star) = 0$, alors $g_j(\boldsymbol{x}^\star) = 0$ alors la contrainte est active ; et si $g_j(\boldsymbol{x}^\star) < 0$ alors elle est inactive car $\mu_j^\star = 0$.

V. Probabilités

Ce chapitre sera vu en deux CM :

(1) Mesure de probabilités, lois usuelles, variables aléatoires, densité de probabilité
(2) Vecteur aléatoire, convergence d'une suite aléatoire (convergence presque sûre, convergence en loi)

Exemple (Ω fini : dé)**.**
Soit $\Omega = \{1,2,3,4,5,6\}$. On considère les événements $A_1 = \{1,3,5\}$ « lancé de dé impair », $A_2 = \Omega$ « lancé de dé », $A_3 = \{1,3\}$ « lancé de dé impair, inférieur à 5 ».

On dit que $M = \wp(\Omega)$ est l'ensemble des événements de Ω. On peut définir

$$P : M \to [0,1]$$
$$A \mapsto P(A) = \sum_{a \in A} P(a) = \frac{\text{card } A}{6}$$

On note $\text{card } A = \#A$ le nombre d'éléments de A.

Définition (Mesure de probabilité uniforme discrète sur Ω fini)**.**
On pose $\Omega = \{\omega_0, \dots, \omega_n\}$ et $M = \wp(\Omega)$. Pour tout i, on pose $P(\omega_i) = 1/n$, équiprobabilité. Ainsi,

$$\forall A \in M, \quad P(A) = \frac{\text{card } A}{n}.$$

Exemple (Tirage d'un nombre dans $\Omega = [0,1]$, infini)**.**
On pose $M = \{[0,1], \emptyset, [0,1/4], (1/4,1]\}$. On définit $P : M \to [0,1]$ par $P(\Omega) = 1$, $P([0,1/4]) = \alpha$ et $P((1/4,1]) = 1 - \alpha$, pour $\alpha \in [0,1]$.

Définition.
On dit que :

- Ω est l'événement certain,
- $\emptyset$ est l'événement impossible,
- $A \neq \Omega$ et $P(A) = 1$ est un événement presque sûr,
- $A \neq \emptyset$ et $P(A) = 0$ est un événement négligeable.

Définition (Espace mesurable)**.**
Soient Ω un ensemble, et M un ensemble de parties de Ω (*i.e.* $M \subseteq \wp(\Omega)$). On dit que M est une *tribu* si
- $\Omega \in M$,
- si $A \in M$, alors $A^c \in M$ (le complémentaire de A dans Ω),
- si $\{A_i\}_{i\in\mathbb{N}}$ est une famille d'éléments de M, alors $\bigcup_{i\in\mathbb{N}} A_i \in M$.

On dira que (Ω, M) est un *espace mesurable* ou *espace prbabilisable.*
De plus, pour $A \in M$, on dit que A est un *événement mesurable.*

(1) L'ensemble des parties $M = \wp(\Omega)$ est une tribu.
(2) L'ensemble $\{\Omega, \emptyset, [0, 1/4], (1/4, 1]\}$ est une tribu engendrée pour $[0, 1/4]$.
(3) Pour $\Omega = \{0, 1, 2\}$, les ensembles $C_1 = \{\emptyset, \Omega, \{0\}, \{1, 2\}\}$, et $C_2 = \{\emptyset, \Omega, \{1\}, \{0, 2\}\}$ sont tous deux des tribus, mais $C_1 \cup C_2$ n'en n'est pas une.
En effet, $A = \{0, 1\} \in C_1 \cup C_2$, mais $A^c = \{2\} \notin C_1 \cup C_2$.

Définition (Tribu borélienne, Borel 1871–1956)**.**
(1) On appelle la *tribu borélienne* de $\mathbb{R}$, notée $\mathscr{B}(\mathbb{R})$, la tribu engendré par les ouverts de $\mathbb{R}$ (ou les fermés). La tribu $\mathscr{B}(\mathbb{R})$ est engendrée par tous les ouverts de la forme $(-\infty, x]$, pour $x \in \mathbb{R}$.
(2) La tribu $\mathscr{B}(\mathbb{R}^n)$ ens engendrée par tous les ouverts de la forme $\prod_{i=1}^n (-\infty, x_i)$, pour $(x_1, \ldots, x_n) \in \mathbb{R}^n$.

Définition (Espace probabilisé, Kolmogorov 1930)**.**
Soit (Ω, M) un espace mesurable. On appelle *probabilité* toute application $P : M \to [0, 1]$ telle que :

- $P(\Omega) = 1$,
- si $\{A_n\}_{(n\in N\}}$ est une famille finie ou dénombrable d'événements de M disjoints deux à deux, alors $P\left(\bigcup_{n\in N} A_n\right) = \sum_{n\in N} P(A_n)$.

Le triplet (Ω, M, P) s'appelle *espace probabilisé.*

(1) L'équiprobabilité est $P(\{\omega\}) = 1/(\text{card } \Omega)$, pour $\omega \in \Omega$.
(2) La loi de poisson $\mathscr{P}(\lambda)$ est définie sur $\mathbb{N}$ par $P(n) = \mathrm{e}^{-\lambda}\lambda^n/n!$.
(3) Pour $\Omega = [0,1]$, et $M = \mathscr{B}(\Omega)$, on définit P sur M par $P(A) = |A|$, pour tout intervalle A de Ω.

Proposition/Définition.
(1) $P(\emptyset) = 0, P(\Omega) = 1, P(A^c) = 1 - P(A)$
(2) $P(A\cup B) = P(A) + P(B) - P(A\cap B), P(A \mid B) = P(A\cap B)/P(B)$
(3) Deux événements A et B sont indépendants si $P(A\cap B) = P(A)\cdot P(B)$.
(4) Des événements $(A_i)_{i\in I}$ sont mutuellement indépendants si

$$\forall p \le |I|, \forall (i_1,\dots,i_p) \in I^p, \quad P\left(A_{i_1}\cap\cdots\cap A_{i_n}\right) = P\left(A_{i_1}\right)\cdots P\left(A_{i_n}\right)$$

V.1. Variables aléatoires réelles

On définit une fonction $X : \Omega \to \mathbb{R}, \omega \mapsto X(\omega)$, où ω est une éventualité sur Ω. C'est une fonction du hazard.

Définition.
Soit (Ω, M, P) un espace probabilisé.

On dit que X est une variable aléatoire réelle si $\forall x \in \mathbb{R}, P(X \in (-\infty, x])$ est défini.

Ceci est équivalent à $\forall B \in \mathscr{B}(\mathbb{R}), P(X^{-1}(B))$ est bien défini, où

$$X^{-1}(B) = \{\omega \in \Omega \mid X(\omega) \in B\}.$$

On note $P(X \in (-\infty, x]) = P(X \le x) = P(\omega \in \Omega \mid X(\omega) \le x)$.

Définition (Fonction de répartition)**.**
La fonction de répartition de X est définie par $F_X : \mathbb{R} \to \mathbb{R}$ telle que $F_X(x) = P(X \leq x)$, quel que soit $x \in \mathbb{R}$.

Exemple 1 (Cas discret)**.**
On pose $\Omega = \{1, 2, 3, 4, 5, 6\}$, et $M = \wp(\Omega)$. La probabilité P est discrète uniforme : $P(i) = 1/6$. On pose $X : \Omega \to \mathbb{R}, X(i) = i$, la valeur du dé. Ainsi, la fonction de répartition de X est donnée par :

$$F_X(x) = P(X \leq x) = \begin{cases} 0 \text{ si } x < 1 \\ \frac{1}{6} \text{ si } 1 \leq x < 2 \\ \frac{1}{2} \text{ si } 2 \leq x < 3 \\ \frac{1}{2} \text{ si } 3 \leq x < 4 \\ \frac{2}{3} \text{ si } 4 \leq x < 5 \\ \frac{5}{6} \text{ si } 5 \leq x < 6 \\ 1 \text{ si } x \geq 6 \end{cases}$$

Exemple 2 (Loi uniforme sur $[0, 1]$)**.**
On pose $\Omega = [0, 1]$, et $M = \mathscr{B}(\Omega)$. On pose P la probabilité continue uniforme, et $P(A) = |A|$. Soit $X : \Omega \to \mathbb{R}$ une variable aléatoire. Alors,

$$\forall x \in \mathbb{R}, F_X(x) = P(X \leq x) = \begin{cases} 0 \text{ si } x < 0 \\ x \text{ si } 0 \leq x \leq 1 \\ 1 \text{ si } x > 1 \end{cases}$$

Dans les deux exemples, on remarque que la fonction de répartition est une fonction *croissante*.

Proposition.

$$P(X \in (a, b]) = F_X(b) - F_X(a).$$

Définition. On dit que deux variables X et Y suivent la même loi si leurs fonction de répartition sont égales $F_X = F_Y$, *i.e.*

$$\forall x \in \mathbb{R}, P(X \leq x) = P(Y \leq x).$$

Défintition (Densité de probabilité)**.**
Soit $f : \mathbb{R} \to \mathbb{R}^+$ une fonction positive et continue (par morceaux) telle que $\int_{-\infty}^{+\infty} f(x)\ \mathrm{d}x = 1$. Alors f est une densité de probabilité.

On dit que X est une variable aléatoire à densité s'il existe f_X une fonction de probabilité et $F_X(x) = \int_{-\infty}^{x} f_X(t)\ \mathrm{d}t = P(X \leq x)$.

Exemples.
(1) Loi uniforme $X \sim \mathcal{U}(a, b)$ de densité $f_X(x) = \mathbb{1}_{[a,b]}(x)/(b-a)$.
(2) Loi normale $X \sim \mathcal{N}(m, \sigma^2)$ de densité

$$F_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot \mathrm{e}^{-\dfrac{(x-m)^2}{2}\sigma^2}.$$

(3) Loi exponentielle $X \sim \mathcal{E}(\lambda)$ de densité $f_X(x) = \lambda \mathrm{e}^{-\lambda x} \mathbb{1}_{\{X \geq 0\}}(x)$

Définition. Soit X une variable aléatoire à valeurs dans l'ensemble $\{x_1, x_2, \ldots, x_n, \ldots\}$. On a $X = \sum_{i=1}^{\infty} x_i \mathbb{1}_{\{X = x_i\}}$.