Table des matières

CHAPITRE 1

Résolution numérique d'équations différentielles

Sommaire

1.1 Généralités

Les équations différentielles décrivent l'évolution de nombreux phénomènes dans des domaines variés. Une équation différentielle est une équation impliquant une ou plusieurs dérivées d'une fonction inconnue. Si toutes les dérivées sont prises par rapport à une seule variable, on parle d'équation différentielle ordinaire. Une équation mettant en jeu des dérivées partielles est appelée équation aux dérivées partielles.

1.1.1 Position du problème

Une équation différentielle (EDO) est une équation exprimée sous la forme d'une relation

$$F(y, y', y'', \ldots, y^{(n)}) = g(t)$$

* dont l'inconnue est une fonction $y : I \subset \mathbb{R} \to \mathbb{R}$ définie sur un intervalle I (à déterminer)
* dans laquelle cohabitent à la fois y et ses dérivées $y', y'', \ldots, y^{(p)}$ (p est appelé l'ordre de l'équation).

Si la fonction g, appelée *second membre* de l'équation, est nulle, on dit que l'équation en question est homogène. Dans ce cours, nous allons nous limiter aux équations différentielles du premier ordre, car une équation d'ordre $p > 1$ peut toujours se ramener à un système de p équations d'ordre 1.

Résoudre une équation c'est chercher toutes les valeurs de l'inconnue qui satisfont l'égalité. Dans les équations rencontrées jusqu'à présent, les inconnues étaient des nombres. Par exemple, résoudre l'équation $2x + 4 = 10$ signifie chercher toutes les valeurs de $x \int \mathbb{R}$ telles que $2x + 4 = 10$. Dans les équations différentielles, les inconnues sont des fonctions. Résoudre une équation différentielle, c'est chercher toutes les fonctions, définies sur un intervalle $I \subset \mathbb{R}$, qui satisfont l'équation (on dit aussi intégrer l'équation différentielle).

Exemple 1.1.1. Résoudre l'équation différentielle $y'(t) = -y(t)$ signifie chercher toutes les fonctions

$$\begin{aligned} y : I \subset \mathbb{R} &\to \mathbb{R}, \\ t &\mapsto y = f(t) \end{aligned}$$

telles que $f'(t) = -f(t)$ pour tout $t \in I$. On peut vérifier que $y(t) = 0$ pour tout $t \in \mathbb{R}$ est une solution de l'EDO mais aussi $y(t) = ce^{-ct}$ pour tout $t \in \mathbb{R}$ (où c est constante réelle quelconque).

1.1.2 Problème de Cauchy

Une EDO admet généralement une infinité de solutions. Pour en sélectionner quelques unes (parfois juste une), on doit imposer une condition supplémentaire qui correspond à la valeur prise par la solution en un point de l'intervalle d'intégration.

Définition 1.1.1. Une condition initiale (CI) est une relation du type $y(t_0) = y_0$ qui impose en t_0 la valeur y_0 de la fonction inconnue. En pratique, se donner une CI revient à se donner le point (t_0, y_0) par lequel doit passer le graphe de la fonction solution.

Définition 1.1.2. (Problème de Cauchy)
Soi $\varphi : \mathbb{R} \to \mathbb{R}$ une fonction donnée et y' la dérivée de y par rapport à t. On appelle *problème de Cauchy* le problème qui consiste à trouver une fonction $y : I \subset \mathbb{R} \to \mathbb{R}$ définie sur un intervalle I telle que

$$\begin{cases} y'(t) &= \varphi(t, y(t)), \qquad \forall t \in I, \\ \\ y(t_0) &= y_0, \end{cases} \tag{1.1}$$

avec t_0 un point de I et y_0 une valeur donnée.

Le but de ce chapitre est d'étudier des méthodes numériques pour approximer les solutions $y(x)$ d'un problème de Cauchy de type (1.1). Les méthodes numériques étudiées ici nous permettront de :

1. déterminer une famille de solutions $y = \psi(x, c)$,appelées *courbes intégrales*, dépendant d'un paramètre c.
2. choisir des membres de cette famille de courbes passant par (x_0, y_0). Ceci en déterminant une valeur particulière de c.

Du point de vue géométrique, l'équation (1.1) signifie que la pente de la courbe intégrale $y = \psi(x, c)$ passant par un point (x, y) quelconque est donnée par $f(x, y)$.

Définition 1.1.3. (Solution maximale)
On se donne une équation différentielle $y'(t) = \varphi(t, y(t))$ avec une condition initiale $y(t_0) = y_0$. Une solution maximale pour ce problème est une fonction $y = f(t)$, définie sur un intervalle I appelé intervalle de vie, telle que

* f est solution de l'équation différentielle et vérifie la condition initiale ;
* il n'existe pas de solution $\tilde{f}$ de la même équation, vérifiant la même condition initiale et définie sur un intervalle J contenant I et plus grand que I.

Définition 1.1.4. On dit que la fonction $f : [a, b] \times \mathbb{R} \to \mathbb{R}$ est Lipschitzienne en y dans $[a, b] \times \mathbb{R}$ s'il existe $A > 0$ tel que :

$$|f(x, y) - f(x, z)| \leq A\,|y - z|\,,\ \forall x \in [a, b],\ \forall y, z \in \mathbb{R}.$$

A est appelé *constante de Lipschitz.*

Théorème 1.1.1. Soit le problème de Cauchy (1.1). Si $f : [a, b] \times \mathbb{R} \to \mathbb{R}$ vérifie les hypothèses :

a. f est continue ;
b. f est lipschitzienne en y,

alors, le problème de Cauchy (1.1) admet une solution unique.

1.2 Schémas numériques classiques

Pour résoudre numérique le problème de Cauchy (1.1), on procède comme suit :

* On prend $x_i = a + ih, i = 0, \ldots, N$, avec $h = \dfrac{b-a}{N}$, une subdivision de l'intervalle $[a, b]$ en N sous-intervalles égaux.
* On cherche N nombres $y_1, \ldots, y_N$ où y_i est une valeur approchée de $y(x_i)$.
* Puis, on relie ces points par interpolation pour définir une fonction y_h sur $[a, b]$.

Si nous intégrons l'EDO $y'(t) = \varphi(t, y(t))$ entre t_n et t_{n+1}, nous obtenons

$$y(t_{n+1}) - y(t_n) = \int_{t_n}^{t_{n+1}} \varphi(t, y(t))dt. \tag{1.2}$$

Pour chaque nœud $t_n = t_0 + nh$ $(1 \leq n \leq N)$, on cherche la valeur inconnue u_n qui approche $y(t_n)$. L'ensemble des valeurs $\{u_0 = y_0, u_1, \ldots, u_N\}$ représente la solution numérique.

1.2.1 Schémas numériques à un pas

On peut construire différentes schémas selon la formule de quadrature utilisée pour approcher le membre de droite. Les schémas qu'on va construire permettent de calculer u_{n+1} à partir de u_n et il est donc possible de calculer successivement $u_1, u_2, \ldots$, en partant de u_0 par une formule de récurrence de la forme

$$\begin{cases} u_0 & = y_0, \\ u_{n+1} & = \varphi(u_n), \qquad \forall n \in \mathbb{N}. \end{cases}$$

— Si on utilise la formule de quadrature de rectangle à gauche, c'est-à-dire

$$\int_{t_n}^{t_{n+1}} \varphi(t, y(t))dt \approx h\varphi(t_n, y(t_n)),$$

on obtient le **schéma d'Euler progressif** défini par

$$\begin{cases} u_0 & = y(t_0) = y_0, \\ u_{n+1} & = u_n + h\varphi(t_n, y_n) \qquad n = 0, 1, \dots N-1. \end{cases}$$

Il s'agit d'un schéma explicite car il permet d'expliciter u_{n+1} en fonction de u_n.

— Si on utilise la formule de quadrature du rectangle à droite, c'est-à-dire

$$\int_{t_n}^{t_{n+1}} \varphi(t, y(t))dt \approx h\varphi(t_{n+1}, y(t_{n+1})),$$

on obtient le **schéma d'Euler rétrograde** défini par

$$\begin{cases} u_0 & = y(t_0) = y_0, \\ u_{n+1} & = u_n + h\varphi(t_{n+1}, y_{n+1}) \qquad n = 0, 1, \dots N-1. \end{cases}$$

Il s'agit d'un schéma implicite car il ne permet pas d'expliciter directement u_{n+1} en fonction de u_n lorsque la fonction f n'est pas triviale.

— Si on utilise la formule du trapèze, c'est-à-dire

$$\int_{t_n}^{t_{n+1}} \varphi(t, y(t))dt \approx \frac{h}{2}\left(\varphi(t_n, y(t_n)) + \varphi(t_{n+1}, y(t_{n+1}))\right),$$

on obtient le **schéma de du trapèze ou de Crank-Nicolson** défini par

$$\begin{cases} u_0 & = y(t_0) = y_0, \\ u_{n+1} - \frac{h}{2}\varphi(t_{n+1}, u_{n+1}) & = u_n + \frac{h}{2}\varphi\left(t_n, u_n\right) \qquad n = 0, 1, \dots N-1. \end{cases}$$

Il s'agit à nouveau d'un schéma implicite car il ne permet pas d'expliciter directement u_{n+1} en fonction de $_n$ lorsque la fonction f n'est pas triviale. En fait, ce schéma fait la moyenne des schémas d'Euler progressif et rétrograde.

Exemple 1.2.1. Soit à résoudre le problème de Cauchy défini par

$$\begin{cases} y' & = -y + x - 1 \\ y(x_0) & = 1,\ x_0 = 0. \end{cases}$$

La solution exacte de cette équation est donnée par $y(x) = x + e^{-x}$.

1. En prenant $N = 10$ et $h = 0.1$, donner une expression de l'équation aux différences associées à ce problème pour les différents schémas vus précédemment.
2. Compléter le tableau suivant afin de calculer les valeurs approchées de la courbe associée.

i	x_i	Solution exacte	Euler progressif	erreur	Euler retrograde	Erreur	CN	Erreur
0	0							
1	0.1							
2	0.2							
3	0.3							
4	0.4							
5	0.5							
6	0.6							
7	0.7							
8	0.8							
9	0.9							

3. Que constatez-vous quant à l'évolution de l'erreur ?

1.2.2 Schémas numériques de type prédicteur-correcteur

Au lieu d'utiliser les schémas classiques présentés dans la section précédente, on peut utiliser des schémas appelés multi-pas car ils ont la forme des schéma prédicteur-correcteur. Ils sont définis comme suit :

1. Si on utilise la formule de quadrature du point du milieu, c'est-à-dire

$$\int_{t_n}^{t_{n+1}} \varphi(t, y(t))dt \approx h\varphi\left(t_n + \frac{h}{2}, y\left(t_n + \frac{h}{2}\right)\right),$$

on obtient un nouveau schéma défini par

$$\begin{cases} u_0 & = y(t_0) = y_0, \\ u_{n+1} & = u_n + h\varphi\left(t_n + \frac{h}{2}, u_{n+1/2}\right) \qquad n = 0, 1, \dots N-1 \end{cases}$$

où $u_{n+1/2}$ est une approximation de $y(t_n + h/2)$. Nous pouvons utiliser une prédiction d'Euler progressive pour approcher le $u_{n+1/2}$ dans le terme $\varphi\left(t_n + \frac{h}{2}, u_{n+1/2}\right)$ par $\tilde{u}_{n+1/2} = u_n + (h/2)\varphi(t_n, u_n)$. Nous avons ainsi construit un nouveau schéma appelé **schéma d'Euler modifié** qui s'écrit :

$$\begin{cases} u_0 & = y(t_0) = y_0, \\ \tilde{u}_{n+1/2} & = u_n + (h/2)\varphi(t_n, u_n) \\ u_{n+1} & = u_n + h\varphi\left(t_n + \frac{h}{2}, \tilde{u}_{n+1/2}\right) \qquad n = 0, 1, \dots N-1. \end{cases}$$

2. Dans le schéma de Crank-Nicolson, pour éviter le calcul implicite de u_{n+1} dans le schéma, nous pouvons utiliser une prédiction d'Euler progressive et remplacer le u_{n+1} dans le terme $\varphi(t_{n+1}, u_{n+1})$ par $\tilde{u}_{n+1} = u_n + h\varphi(t_n, u_n)$. Nous avons construit ainsi un nouveau schéma appelé **schéma de Heun**. Plus précisément, la méthode de **Heun** s'écrit

$$\begin{cases} u_0 & = y(t_0) = y_0, \\ \tilde{u}_{n+1} & = u_n + h\varphi(t_n, u_n), \\ u_{n+1} & = u_n + \frac{h}{2}\left(\varphi(t_n, u_n) + \varphi(t_n, \tilde{u}_n)\right) \qquad n = 0, 1, \ldots N-1. \end{cases}$$

1.3 D'autres méthodes numériques

1.3.1 Méthode de Taylor

Une première façon d'améliorer la méthode d'Euler (au sens où l'erreur variera au moins comme h^2) consiste à utiliser un développement de Taylor jusqu'à l'ordre deux. Si $u_n = y(t_n)$ est donnée, on a

$$y(t_{n+1}) = y(t_n) + (t_{n+1}) - t_n)y'(t_n) + \frac{1}{2}(t_{n+1} - t_n)^2 y''(t_n) + \frac{1}{6}(t_{n+1} - t_n)^3 y^{(3)}(\xi_n),$$

où $\xi_n \in [t_n, t_{n+1}]$. Si $h = t_{n+1} - t_n$ est assez petit, on a donc l'égalité approchée

$$y(t_{n+1}) \simeq y(t_n) + hy'(t_n) + \frac{h^2}{2}y''(t_n).$$

Puisque $y'(t_n) = \varphi(t_n, y(t_n))$, alors on a :

$$y''(t_n) = \frac{\partial \varphi}{\partial t}(t_n, y(t_n)) + \frac{\partial \varphi}{\partial y}(t_n, y(t_n))\varphi(t_n, y(t_n)).$$

Ceci suggère donc la méthode aux différences définie par

$$u_{n+1} = u_n + h\varphi(t_n, u_n) + \frac{h^2}{2}\left[\frac{\partial \varphi}{\partial t}(t_n, u_n) + \frac{\partial \varphi}{\partial y}(t_n, y_n)\varphi(t_n, y_n)\right], \tag{1.3}$$

appelée **méthode de Taylor d'ordre 2**.

Remarques.

1. On peut montrer que pour chaque t_n fixé, l'erreur varie comme h^2.
2. Cette méthode possède le désavantage d'elle nécessite le calcul des dérivées partielles de $\varphi(t, y)$.
3. Cette méthode se généralise facilement, cependant les méthode de Taylor nécessitent le calcul des dérivées partielles d'ordre supérieur. Ce qui peut être très coûteux.

Exemple 1.3.1. soit l'équation différentielle définie par

$$\begin{cases} y' & = -y + x + 1, \\ y_0 & = y(x_0) = 1. \end{cases}$$

Trouver l'équation aux différences associée à cette équation, en utilisant la méthode de Taylor. Puis, en prenant $h = 0.1$ et $x_0 = 0$, calculer les trois premières valeurs de cette solution.

1.3.2 Méthodes de Runge-Kutta

On voudrait conserver l'avantage des méthodes d'ordre supérieur mais corriger les inconvénients dus au calcul des dérivées partielles de φ. Ces différentes méthodes sont basées sur la formule de Taylor à plusieurs variables :

$$\begin{aligned} f(x,y) &= f(x_0,y_0) + \left[(x-x_0)\frac{\partial\varphi}{\partial x} + (y-y_0)\frac{\partial\varphi}{\partial y}\right] \\ &+ \frac{1}{2!}\left[(x-x_0)^2\frac{\partial^2\varphi}{\partial x^2} + 2(x-x_0)(y-y_0)\frac{\partial^2\varphi}{\partial x\partial y} + (y-y_0)^2\frac{\partial^2\varphi}{\partial y^2}\right] \\ &+ \cdots + \frac{1}{(n+1)!}\sum_{j=0}^{n+1} \mathtt{C}_{n+1}^{j}\left[(x-x_0)^{n+1-j}(y-y_0)^j\frac{\partial^{n+1}\varphi}{\partial x^{n+1-j}\partial y^j}(\xi,\eta)\right], \end{aligned} \tag{1.4}$$

où $\xi \in [x, x_0]$ et $\eta \in [y, y_0]$.

Parmi toutes ces méthodes, la plus utilisée est celle d'ordre 4 mais les calculs menant à l'algorithme sont très laborieux. On illustre ici l'idée de base pour la méthode de Runge-Kutta d'ordre deux. On essaie de remplacer le terme

$$\varphi(t_i,y_i) + \frac{h}{2}\left[\frac{\partial\varphi}{\partial t}(t_i,y_i) + \frac{\partial\varphi}{\partial y}(t_i,y_i)\varphi(t_i,y_i)\right]$$

de l'équation aux différences obtenue pour la méthode de Taylor par une expression de type

$$a\varphi(t_i+\alpha, y_i+\beta),$$

en utilisant (1.4). On obtient alors,

$$a\varphi(t_i+\alpha,y_i+\beta) = a\varphi(t_i,y_i) + a\alpha\frac{\partial\varphi}{\partial t}(t_i,y_i) + a\beta\frac{\partial\varphi}{\partial y}(t_i,y_i) + aR_2(\xi_i,\eta_i),$$

avec $\xi_i \in [x_i, x_i+\alpha],\ \eta_i \in [y_i, y_i+\beta]$. En identifiant les cœfficients de φ, $\frac{\partial\varphi}{\partial t}$, $\frac{\partial\varphi}{\partial y}$, on voit que l'on doit choisir :

$$a = 1, \qquad a\alpha = \frac{h}{2}, \qquad a\beta = \frac{h}{2}\varphi(t_i,y_i),$$

c'est-à-dire $a = 1$, $\alpha = \frac{h}{2}$ et $\beta = \frac{h}{2}\varphi(t_i,y_i)$. En substituant dans l'équation aux différences de Taylor, on obtient donc la formule aux différences

$$y_{n+1} = y_n + h\varphi\left(t_n + \frac{h}{2}, y_n + \frac{h}{2}\varphi(t_n,y_n)\right). \tag{1.5}$$

D'où les algorithmes suivants :

Exemple 1.3.2. Soit l'équation différentielle définie par

$$\begin{cases} y'(t) &= y(t) + t^2 + 1, \qquad t \in [0,1] \\ y(0) &= 1. \end{cases}$$

poser $h = \dfrac{t_N - t_0}{N}$;
poser $i = 0$;
Tant que $(i < N)$ **faire**
    $k = h\varphi(t_i, y_i)$;
    $y_{i+1} = y_i + h\varphi(t_i + h/2, y_i + k/2)$;
    $t_{i+1} = t_i + h$;
    $i = i + 1$;
finTantQue

Algorithm 1: Méthode de Runge-Kutta d'ordre 2.

poser $h = \dfrac{t_N - t_0}{N}$;
poser $i = 0$;
Tant que $(i < N)$ **faire**
    $k_1 = h\varphi(t_i, y_i)$;
    $k_2 = h\varphi(t_i + h/2, y_i + k_1/2)$;
    $k_3 = h\varphi(t_i + h/2, y_i + k_2/2)$;
    $k_4 = h\varphi(t_i + h, y_i + k_3)$;
    $y_{i+1} = y_i + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4)$;
    $t_{i+1} = x_i + h$;
    $i = i + 1$;
finTantQue

Algorithm 2: Méthode de Runge-Kutta d'ordre 4.

1. Montrer que ce problème admet une et une seule solution dans $[0, 1]$, puis trouver cette solution exacte.
2. En prenant le nombre de points $N = 10$, utiliser la méthode de Runge-Kutta d'ordre deux pour approcher les solutions de ce problème.
3. Reprendre les questions précédentes, en prenant $\varphi(t, y) = -y(t) + t + 1$ et en utilisant la méthode de Runge-Kutta d'ordre 4.

1.4 Etude numérique de la convergence des schémas

1.4.1 Stabilité numérique d'un schéma

Lorsqu'on applique un schéma numérique pour résoudre un problème différentiel, deux questions naturelles se posent : que se passe-t-il lorsqu'on fait tendre le pas h vers 0 ? Que se passe-t-il lorsqu'on fixe le pas $h > 0$ mais on fait tendre T vers l'infini ? Dans les deux cas le nombre de nœuds tend vers l'infini mais dans le premier cas on s'intéresse à l'erreur en chaque point, dans le deuxième cas il s'agit du comportement asymptotique de la solution et de son approximation :

Zéro-stabilité. Soit T fixé et considérons la limite $h \to 0$ (ainsi $N \to +\infty$). On note $e_n^{(h)} \equiv y(t_n^{(h)}) - u_n^{(h)} = y(t_0 + nh) - u_n^{(h)}$ l'erreur au point $t_0 + nh$. Il s'agit d'estimer le comportement

de $e_n^{(h)}$ en tout point, c'est-à-dire $1 \leq n\mathbb{N}$. La méthode est zéro-stable si $e_n^{(h)} \to 0$ quand h tend vers 0 pour tout $n \in \mathbb{N}$.

Cette notion est très importante car le théorème de Lax-Richtmyer (ou théorème d'équivalence) affirme que une méthode consistante et zéro-stable est convergente.

A-stabilité. On considère un problème de Cauchy défini par (1.1) dont la solution exacte vérifie $y(t) \longrightarrow 0$ quand t tend vers l'infini. Soit h fixé et considérons la limite $T \to +\infty$ (ainsi $N \to +\infty$). On dit que la méthode est A-stable si $u_n^{(h)} \to 0$ quand n tend vers l'infini.

Définition 1.4.1. Soit $\beta > 0$ un nombre réel positif et considérons le problème de Cauchy

$$\begin{cases} y'(t) &= -\beta y(t), \qquad \forall t > 0, \\ y(0) &= y_0 \end{cases}$$

où $y_0 \neq 0$ est une valeur donnée. Sa solution est $y(t) = y_0 e^{-\beta t}$ et $\lim\limits_{t \to +\infty} y(t) = 0$. Soit $h > 0$ un pas de temps donnée, $t_n = nh$ pour tout $n \in \mathbb{N}$ et notons $u_n \approx y(t_n)$ une approximation de la solution y au temps t_n.
Si, sous d'éventuelles conditions sur h, on a

$$\lim_{n \to +\infty} u_n = 0,$$

alors, on dit que le schéma est A-stable.

Exemple 1.4.1. Etudions la stabilité des quelques schémas numériques vus précédemment.

* Le **schéma d'Euler progressif** devient

$$u_{n+1} = (1 - \beta h) u_n, \qquad n = 0, 1, 2 \ldots$$

et par la suie

$$u_n = (1 - \beta h)^n u_0, \qquad n = 0, 1, 2, \ldots$$

Par conséquent, $\lim\limits_{n \to +\infty} u_n = 0$ si, et seulement si

$$|1 - \beta h| < 1,$$

ce qui a pour effet de limiter h à

$$h \leq \frac{2}{\beta}.$$

Cette condition de A-stabilité limite le pas h d'avance en t lorsqu'on utilise le schéma d'Euler progressif. On dit donc que le schéma d'Euler progressif est **conditionnement stable**.

* Le **schéma d'Euler rétrograde** devient dans le cadre du même exemple :

$$(1 + \beta h) u_{n+1} = u_n, \qquad n = 0, 1, 2, \ldots$$

et par suite, on obtient

$$u_n = \frac{1}{(1+\beta h)^n}, \qquad n = 0, 1, 2, \ldots$$

Dans ce cas nous voyons que pour tout $h > 0$ nous avons $\lim_{n\to+\infty} u_n = 0$, le schéma d'Euler rétrograde est donc toujours stable, sans limitations sur h. On dit que le schéma d'Euler rétrograde est **inconditionnellement stable.**

* Etudier la A-stabilité des schémas de Crank-Nicolson et de Heun.

1.4.2 Consistance d'un schéma numérique

L'erreur de consistance au pas n est par définition l'erreur commise sur y_{n+1}, lorsqu'on prend pour les valeurs précédentes des y_k les valeurs exactes $z(t_k)$, ce qui donne la définition suivante où nous n'explicitons que pour les méthodes à un pas.

Définition 1.4.2.

1. L'erreur de consistance est la suite

$$e_n = z(t_{n+1}) - y_{n+1}(t_n, z(t_n), h_n) = z(t_{n+1}) - z(t_n) - h_n\Phi(t_n, z(t_n), h_n).$$

2. Une méthode numérique est dite consistante si

$$\lim_{h_{\max}\to 0} \sum_{n=0}^{N-1} |e_n| = 0.$$

Théorème 1.4.1. Une méthode numérique à un pas est consistance si, et seulement si, quel que soit (t, y), on a

$$\Phi(t, y, 0) = f(t, y).$$

Théorème 1.4.2. Une méthode numérique à un pas qui est stable et consistante est convergente.

1.5 Exercices

Exercice 1.5.1. On considère le problème différentiel défini par

$$\begin{cases} y''(t) &= f(t, y(t), y'(t)), \quad t \in [a, b] \\ \\ y(a) &= \alpha \quad \text{et} \quad y(b) = \beta \end{cases} \tag{1.6}$$

On divise $[a, b]$ en m intervalles de longueur h, dans (1.11) on remplace $y'(x_n)$ par $\dfrac{y_{n+1} - y_{n-1}}{2h}$ et y'' par $\dfrac{y_{n+1} - 2y_n + y_{n-1}}{h^2}$.

1. Montrer que le problème (1.11) se transforme en

$$\begin{cases} y_{n+1} &= \Phi(y_n, y_{n-1}) + h^2 f_n \\ \\ y_0 &= \alpha \quad \text{et} \quad y_m = \beta \end{cases} \tag{1.7}$$

avec $n = 1, \ldots, m-1$ et $f_n = f\left(t_n, y_n, \dfrac{y_{n+1} - y_{n-1}}{2h}\right)$. Donner Φ.

2. Écrire (1.12) sous forme matricielle.
3. On suppose maintenant que $f(t, y, y') = -\lambda y$; quel système linéaire obtient-on pour $y_0, \ldots, y_m$.

Exercice 1.5.2. Pour résoudre l'équation différentielle : $y' = f(t, y)$, où f est continue de $[a, b] \times \mathbb{R}$ dans $\mathbb{R}$, on propose la méthode à un pas suivante :

$$\begin{cases} y_{n+1} &= y_n + h\Phi(t_n, y_n, h) \\ \\ \Phi(t, y, h) &= \alpha f(t, h) + \beta f\left(t + \dfrac{h}{2}, y + \dfrac{h}{2} f(t, y)\right) + \gamma f(t + h, y + hf(t, y)) \end{cases} \tag{1.8}$$

(α, β, γ sont des réels de $[a, b]$).

1. Pour quelles valeurs du triplet (α, β, γ) retrouve-t-on la méthode d'Euler ? Même question pour la méthode RK2 ?
2. On suppose dans la suite de l'exercice que $f(t, y) \in \mathcal{C}^2([a, b] \times \mathbb{R})$ et L-lipschitzienne en y :
   (a) Pour quelles valeurs α, β, γ, la méthode proposée est stable ?
   (b) Quelle relation doit satisfaire (α, β, γ) pour que la méthode soit consistante ?
   (c) Que peut-on conclure quand à la convergence de la méthode ?

Exercice 1.5.3. On considère le problème de CAUCHY

$$\begin{cases} y'(t) &= -y(t), \\ y(0) &= 1, \end{cases} \tag{1.9}$$

sur l'intervalle $[0; 10]$.

1. Calculer la solution exacte du problème de CAUCHY.
2. Soit h le pas temporel. Écrire la méthode d'EULER explicite pour cette équation différentielle ordinaire (EDO).
3. En déduire une forme du type

$$u_{n+1} = g(h, n)$$

avec $g(h, n)$ à préciser (autrement dit, l'itérée en t_n ne dépend que de h et n et ne dépend pas de u_n).

4. Utiliser la formulation ainsi obtenue pour tracer les solutions

(a) exacte,

(b) obtenue avec la méthode d'Euler avec = $h = 2.5$,

(c) obtenue avec la méthode d'Euler avec = $h = 1.5$,

(d) obtenue avec la méthode d'Euler avec = $h = 0.5$,

5. Que peut-on en déduire sur la A-stabilité de la méthode ?
6. Montrer que la méthode d'Euler est convergente.
7. Pour $h = 0.5$, évaluer l'erreur absolue en $t = 2.0$ et interpréter géométriquement le résultat.

Exercice 1.5.4. Considérons l'équation différentielle

$$y'(t) = f(t, y), \quad a \leq t \leq b, \quad y(a) = \alpha. \tag{1.10}$$

1. Montrer que
$$y'(t) = \frac{-3y(t_i) + 4y(t_{i+1}) - y(t_{i+2})}{2h} + \frac{h^2}{3}y'''(\zeta_1),$$
tel que $t_i < \zeta_i < t_{i+2}$.
2. Montrer que les approximations y_i de la solution (1.10) vérifient
$$y_{i+2} = 4y_{i+1} - 3y_i - 2hf(t_i, y_i), \forall i = 0, 1, \ldots, N-2.$$
3. Utilise cette méthode pour résoudre l'équation
$$y' = 1 - y, \quad 0 \leq t \leq 1,$$
avec $h = 0.1$. On prendra les valeurs initiales $y_0 = 0$ et $y_1 = y(t_1) = 1 - e^{-0.1}$.
4. Refaire la question (3) avec $h = 0.01$ et $y_1 = 1 - e^{-0.01}$.
5. Que dire de la stabilité, de la consistance et de la convergence de cette méthode ?

Exercice 1.5.5. Soit $\beta > 0$ un nombre réel positif et considérons le problème de Cauchy

$$\begin{cases} y'(t) = -\beta y(t), & \forall t > 0, \\ y(0) = y_0, & \text{pour } t = 0. \end{cases} \tag{1.11}$$

où y_0 est une valeur donnée. Soit $h > 0$ un pas de temps donné, $t_n = nh$ pour $n \in \mathbb{N}$ et u_n une approximation de $y(t_n)$.

1. Écrire le schéma du trapèze (appelé aussi de Crank-Nicholson) permettant de calculer u_{n+1} à partir de u_n.
2. Sous quelle condition sur h le schéma du trapèze est-il A-stable ? Autrement dit, pour quelles valeurs de h la relation $\lim_{n\to+\infty} u_n = 0$ a-t-elle lieu ? Conclure quant à la stabilité du schéma de Crank-Nicholson.

3. À partir du schéma du trapèze, en déduire le schéma de Heun. Sous quelle condition sur h le schéma de Heun est-il A-stable ?

Exercice 1.5.6. L'évolution de la concentration de certaines réactions chimiques au cours du temps peut être décrite par l'équation différentielle

$$y'(t) = -\frac{1}{1+t^2}y(t).$$

1. Sachant qu'à l'instant $t = 0$ la concentration est $y(0) = 5$, calculer la solution exacte de cette équation différentielle.
2. Pour $h = 0.5$, établir une suite numérique correspondant aux schémas numériques d'Euler explicite, Euler implicite puis de Crank-Nicholson.
3. Déterminer la concentration à $t = 2$ pour chacun des schémas numériques précédents.
4. Comparer l'erreur absolue obtenue quant aux trois méthodes.

Exercice 1.5.7. On considère le problème de Cauchy défini par

$$\begin{cases} y'(t) = -\left(y(t)\right)^m + \cos(t), \quad \forall t > 0, \\ y(0) = 0, \end{cases} \tag{1.12}$$

où m est un entier impair.

1. Montrer que le problème (1.12) possède une solution unique locale.
2. Soit $h = 0$ un pas de temps donné, soit $tn = nh$ pour $n \in \mathbb{N}$ et u_n une approximation de $y(t_n)$. Écrire le schéma d'Euler rétrograde permettant de calculer u_{n+1} à partir de u_n.
3. À partir du schéma obtenu au point précédent, écrire un seul pas de la méthode de Newton pour calculer une nouvelle approximation de u_{n+1}. En déduire ainsi un nouveau schéma explicite.

Exercice 1.5.8. Soit $f : \mathcal{D} = [a, b] \times \mathbb{R} \to \mathbb{R}$ une fonction continue sur $\mathcal{D}$ et Lipschitzienne par rapport à la deuxième variable. On considère l'équation différentielle

$$\begin{cases} y' & = f(x, y), \qquad x \in [a, b], \\ y(a) & = Y_0 \end{cases}$$

dont la solution exacte est notée Y. Soit $x_0 = a, x_1, \ldots, x_N = b$, $N + 1$ points équidistants appartenant à $[a, b]$, et soit $h = x_{n+1} - x_n$,$n = 0, \ldots, N - 1$.

1. En utilisant la formule d'intégration numérique du point milieu, montrer que :

$$Y(x_{n+1}) - Y(x_n) = hf\left(x_{n+\frac{1}{2}}, Y(x_{n+\frac{1}{2}})\right) + E_1(f),$$

   où $x_{n+\frac{1}{2}} = \frac{1}{2}(x_{n+1} - x_n)$ et $E_1(f)$ désigne le terme d'erreur.
2. En utilisant la formule d'intégration numérique du rectangle à gauche, montrer que

$$Y(x_{n+\frac{1}{2}}) = Y(x_n) + \frac{h}{2}f(x_n, Y(x_n)) + E_2(f).$$

3. En déduire un schéma numérique à un pas :

$$y_{n+1} = y_n + h\Phi(x_n, y_n, h), \qquad \text{(S)}$$

pour une détermination d'une approximation y_n de $Y(x_n)$.

4. Montrer que le schéma (S) est stable, consistant et est d'ordre 2.

CHAPITRE 2

Valeurs et vecteurs propres d'une matrice

Sommaire

2.1 Rappels d'algèbre linéaire

Définition 2.1.1. Soit $\mathbf{A} \in \mathbb{C}^{n,n}$ une matrice d'ordre n à valeurs complexes et $x, y \in \mathbb{C}^n$ des vecteurs colonnes ayant n éléments.

1. On dit que $\lambda \in \mathbb{C}$ est une valeur propre de $\mathbf{A}$ s'il existe un vecteur non nul $x \in \mathbb{C}^n$ tel que $Ax = \lambda x$. Le vecteur x est appelé *vecteur propre* associé à la valeur propre λ.
2. L'ensemble des valeurs propres d'une matrice $\mathbf{A}$ est appelé le *spectre de* $\mathbf{A}$ et est noté $\sigma(\mathbf{A})$.
3. On appelle *matrice adjointe* $\mathbf{A}^*$ de $\mathbf{A}$ la matrice transconjuguée (transposée de la conjugué) définie par
$$\mathbf{A}^* = (a_{ij}^*) = (\overline{a}_{ji}).$$
4. On dit que les vecteurs x et y sont respectivement vecteur propre à droite et vecteur propre à gauche de $\mathbf{A}$ associés à la valeur propre λ, si
$$\mathbf{A}x = x; \qquad y^*\mathbf{A} = \lambda y^*,$$
avec y^* le vecteur adjoint de y.

La valeur propre λ correspondant au vecteur propre x peut être déterminée en calculant le quotient de Rayleigh
$$\lambda = \frac{x^*\mathbf{A}x}{x^*x}.$$

Le nombreλ est solution de l'équation caractéristique

$$P_{\mathbf{A}}(\lambda) = \det(\mathbf{A} - \lambda I) = 0,$$

où $P_{\mathbf{A}}(\lambda)$ est le polynôme caractéristique de $\mathbf{A}$. Ce polynôme étant de degré n par rapport à λ, on sait qu'il existe n valeurs propres (non nécessairement distinctes). On peut donc démontrer la propriété suivante :

$$\det(\mathbf{A}) = \prod_{i=1}^{n} \lambda_i, \qquad \text{tr}(\mathbf{A}) = \sum_{i=1}^{n} \lambda_i.$$

Puisque $\det(\mathbf{A}^\top - \lambda I) = \det((\mathbf{A} - \lambda I)^\top) = \det(\mathbf{A} - \lambda I)$, on déduit que $\sigma(\mathbf{A}) = \sigma(\mathbf{A}^\top)$ et d'une manière analogue, que $\sigma(\mathbf{A}^*) = \sigma(\overline{\mathbf{A}})$.

Définition 2.1.2. Le plus grand module des valeurs propres d'une matrice $\mathbf{A}$ est appelé *rayon spectral* de $\mathbf{A}$ noté $\rho(\mathbf{A})$ et défini par

$$\rho(\mathbf{A}) = \max_{\lambda \in \sigma(\mathbf{A})} |\lambda|.$$

En utilisant la caractérisation des valeurs propres d'une matrice comme racines du polynôme caractéristique, on voit en particulier que λ est une valeur propre de $\mathbf{A} \in \mathbb{C}^{n,n}$ si, et seulement si, $\overline{\lambda}$ est une valeur propre de $\mathbf{A}^*$. Une conséquence immédiate est que $\rho(\mathbf{A}) = \rho(\mathbf{A}^*)$. De plus, $\forall \mathbf{A} \in \mathbb{C}^{n,n}$ et $\forall \alpha \in \mathbb{C}$, on a

$$\rho(\alpha \mathbf{A}) = |\alpha| \rho(\mathbf{A}), \qquad \text{et} \qquad \rho(\mathbf{A}^k) = [\rho(\mathbf{A})]^k, \ \forall k \in \mathbb{N}.$$

Définition 2.1.3. Une matrice $\mathbf{A} \in \mathbb{C}^{n,n}$ est dite :

1. hermitienne (autoadjointe) si $\mathbf{A}^\top = \overline{\mathbf{A}}$ ou encore $\mathbf{A}^* = \mathbf{A}$.
2. normale si $\mathbf{A}^*\mathbf{A} = \mathbf{A}\mathbf{A}^*$.
3. unitaire si $\mathbf{A}^*\mathbf{A} = \mathbf{A}\mathbf{A}^* = I$.

2.2 Localisation géométrique des valeurs propres

Les valeurs propres d'une matrice $\mathbf{A}$ étant les racines du polynôme caractéristique $P_{\mathbf{A}}(\lambda)$, on ne peut donc les calculer qu'avec des méthodes itératives quand $n \geq 5$. Il est donc utile de connaître leur localisation dans le plan complexe pour accélérer la convergence.

On sait (on l'a vu en analyse numérique II) que $|\lambda| \leq \|\mathbf{A}\|, \forall \lambda \in \sigma(\mathbf{A})$, pour toute norme matricielle constante $\|\cdot\|$. Cette inégalité montre que toutes les valeurs propres de $\mathbf{A}$ sont contenues dans un disque de rayon $R_{\|\mathbf{A}\|} = \|\mathbf{A}\|$ centré à l'origine du plan complexe.

Théorème 2.2.1. Soit $\mathbf{A} \in \mathbb{C}^{n,n}$ et soient $H = \dfrac{\mathbf{A} + \mathbf{A}^*}{2}$ et $iS = \dfrac{\mathbf{A} - \mathbf{A}^*}{2}$ les parties hermitienne

et anti-hermitienne de $\mathbf{A}$. Pour tout $\lambda \in \sigma(\mathbf{A})$, on a

$$\begin{cases} \lambda_{\min}(H) \leq \mathrm{Re}(\lambda) \leq \lambda_{\max}(H), \\ \\ \lambda_{\min}(S) \leq \mathrm{Re}(\lambda) \leq \lambda_{\max}(S). \end{cases}$$

Démonstration. D'après la définition de H est S, on a $\mathbf{A} = H + iS$. Soit $u \in \mathbb{C}^n$ un vecteur propre associé à la valeur propre λ tel que $\|u\|_2 = 1$. Le quotient de Rayleigh s'écrit par

$$\lambda = u^* A u = u^* H u + i u^* S u.$$

Or, on remarque que H et S sont des matrices hermitiennes, tandis que iS est antihermitienne. Ainsi, les matrices H et S sont unitaires et semblables à une matrice réelle diagonale. Les valeurs propres sont donc réelles et on déduit que

$$\mathrm{Re}(\lambda) = u^* H u \text{ et } \mathrm{Im}(\lambda) = u^* S u.$$

D'où

$$\begin{cases} \lambda_{\min}(H) \leq \mathrm{Re}(\lambda) \leq \lambda_{\max}(H), \\ \\ \lambda_{\min}(S) \leq \mathrm{Re}(\lambda) \leq \lambda_{\max}(S). \end{cases}$$

□

Le résultat suivant donne une estimation à priori des valeurs propres de $\mathbf{A}$.

Définition 2.2.1. Soit $\mathbf{A} \in \mathbb{C}^{n,n}$. On appelle *disques de Gershgorin* associés à la matrice $\mathbf{A}$, les disques définis par

$$\mathcal{R}_i = \left\{ z \in \mathbb{C} \ : \ |z - a_{ii}| \leq \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}| \right\}.$$

Théorème 2.2.2. (Les disques de Gershgorin).

Soit $\mathbf{A} \in \mathbb{C}^{n,n}$ une matrice d'ordre n. Alors

$$\sigma(\mathbf{A}) \subseteq S_{\mathcal{R}} = \bigcup_{i=1}^{n} \mathcal{R}_i.$$

Démonstration. Décomposons $\mathbf{A}$ sous la forme $\mathbf{A} = D + E$ avec

$$d_{ij} = \begin{cases} a_{ij} & \text{si } i = j \\ 0 & \text{sinon} \end{cases} \qquad \text{et} \qquad e_{ij} = \begin{cases} 0 & \text{si } i = j, \\ a_{ij} & \text{sinon.} \end{cases}$$

Pour $\lambda \in \sigma(\mathbf{A})$ (avec $\lambda \neq a_{kk},\ 1 \leq k \leq n$), on introduit la matrice $B_\lambda = \mathbf{A} - \lambda I = (D - \lambda I) + E$. Comme B_λ est singulière, il existe un vecteur non nul $x \in \mathbb{C}^{n,n}$ tel que $B_\lambda x = 0$, c'est-à-dire

$$((D - \lambda I) + E)\, x = 0 \quad \Rightarrow \quad x = -(D - \lambda I)^{-1} E x.$$

En passant à la norme $\|\cdot\|_\infty$ on a ainsi :

$$\|x\|_\infty \leq \|(D-\lambda I)^{-1}E\|_\infty \|x\|_\infty.$$

Et donc, pour un certain k tel que $1 \leq k \leq n$,

$$1 \leq \|(D-\lambda I)^{-1}E\|_\infty = \sum_{j=1}^{n} \frac{|e_{ij}|}{|a_{kk}-\lambda|} = \sum_{\substack{j=1 \\ j\neq k}}^{n} \frac{|a_{kj}|}{|a_{kk}-\lambda|}.$$

On en déduit alors que $\lambda \in \mathcal{R}_k$. D'où le résultat.

□

Le théorème précédent assure que toutes les valeurs propres de $\mathbf{A}$ se trouvent dans la réunion des disques $\mathcal{R}_i$. De plus, les matrices $\mathbf{A}$ et $\mathbf{A}^\top$ ayant le même spectre, le théorème peut aussi s'écrire sous la forme

$$\sigma(\mathbf{A}) \subseteq S_{\mathcal{C}} = \bigcup_{j=1}^{n} \mathcal{C}_j \qquad \text{avec} \qquad \mathcal{C}_j = \left\{ z \in \mathbb{C} \ : \ |z-a_{jj}| \leq \sum_{\substack{i=1 \\ i\neq j}}^{n} |a_{ij}| \right\}.$$

Les disques $\mathcal{R}_i$ du plan complexe sont appelés les *disques lignes* tandis que les disques $\mathcal{C}_j$ sont appelés les *disques colonnes*.

Propriété 2.2.1. (Premier théorème de Gershgorin)
Pour une matrice $\mathbf{A} \in \mathbb{C}^{n,n}$ donnée, on a

$$\forall \lambda \in \sigma(\mathbf{A}), \qquad \lambda \in S_{\mathcal{R}} \cap S_{\mathcal{C}}.$$

Propriété 2.2.2. (Deuxième théorème de Gershgorin)
Soit $1 \leq m \leq n$ un entier et

$$S_1 = \bigcup_{i=1}^{m} \mathcal{R}_i, \qquad S_2 = \bigcup_{i=m+1}^{n} \mathcal{R}_i.$$

Si $S_1 \cap S_2 = \emptyset$, alors S_1 contient exactement m valeurs propres de $\mathbf{A}$, chacune étant comptée avec son ordre de multiplicité algébrique. Les autres valeurs propres sont dans S_2.

Définition 2.2.2. Une matrice $\mathbf{A} \in \mathbb{C}^{n,n}$ est dite réductible s'il existe une matrice de permutation P telle que

$$PAP^\top = \begin{bmatrix} B_{11} & B_{12} \\ \mathbf{0} & B_{22} \end{bmatrix},$$

où B_{11}, B_{12} et B_{22} sont des matrices carrées. La matrice $\mathbf{A}$ est dite irréductible dans le cas contraire.

Propriété 2.2.3. (Troisième théorème de Gershgorin)
Soit $\mathbf{A} \in \mathbb{C}^{n,n}$ une matrice irréductible. Une valeur propre $\lambda \in \sigma(\mathbf{A})$ ne peut pas appartenir au bord de $S_{\mathcal{R}}$ à moins qu'elle n'appartienne au bord de chaque disque $\mathcal{R}_i$ pour $i = 1, \ldots, n$.

Exemple 2.2.1. Considérons la matrice définie par

$$\mathbf{A} = \begin{pmatrix} 10 & 2 & 3 \\ -1 & 2 & -1 \\ 0 & 1 & 3 \end{pmatrix}$$

dont le spectre est $\sigma(\mathbf{A}) = \{9.687, 2.656 \pm 0.6933i\}$.

1. Vérifier que $|\lambda| \leq \|\mathbf{A}\|$ pour toute norme matricielle donnée.
2. Déterminer les disques lignes et colonnes de Gershgorin associés à la matrice $\mathbf{A}$.
3. Tracer ces disques sur le plan complexes puis en déduire un encadrement du rayon spectral de $\mathbf{A}$.
4. Reprendre les questions précédentes en considérant la matrice

$$\mathbf{A} = \begin{pmatrix} 4 & 1 & 1 \\ 0 & 2 & 1 \\ -2 & 0 & 9 \end{pmatrix}.$$

2.3 Méthodes de la puissance itérée

Ce paragraphe décrit un premier type de méthodes utilisées généralement pour le calcul d'une ou de quelques valeurs propres d'une matrice, ainsi que des vecteurs propres associés.

2.3.1 Itérations simples

Théorème 2.3.1. Soit $\mathbf{A} \in \mathbb{C}^{n,n}$. On suppose que la valeur propre de plus grand module est simple et vérifie $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_n|$ et que $\mathbf{A}$ possède n vecteurs propres $u_1, \cdots, u_n$ linéairement indépendants, autrement dit $\mathbf{A}$ est diagonalisable. Soit $|\cdot\|$ une norme vectorielle. On construit une suite $\mathbf{x}^{(k)} = (x_1^{(k)}, \ldots, x_n^{(k)})^\top$ d'éléments de $\mathbb{C}^n$ par récurrence :

$$\begin{cases} \mathbf{x}^{(0)} & \text{donné dans } \mathbb{C}^n, \\ \\ \mathbf{x}^{(k+1)} & = \dfrac{\mathbf{A}\mathbf{x}^{(k)}}{\|\mathbf{x}^{(k)}\|}, \qquad k \in \mathbb{N} \end{cases} \tag{2.1}$$

Si $\mathbf{x}^{(0)}$ n'est pas orthogonal à l'espace propre à gauche associé à λ_1, alors :

1. la suite $\left((\lambda_1/|\lambda_1|)^k \mathbf{x}^{(k)}\right)$ converge vers un vecteur propre associé à λ_1 quand $k \to +\infty$;
2. $\lim\limits_{k\to+\infty} (\mathbf{A}\mathbf{x}^{(k)})_j / x_j^{(k)} = \lambda_1$ pour au moins un $j \in \{1, \ldots, n\}$ où on a noté $\mathbf{x}^{(k)} = (x_1^{(k)}, \cdots, x_n^{(k)})^\top$.

On utilise le plus souvent la norme $\|\cdot\|_\infty$.

Démonstration. On procède en deux étapes :

1. Soit $\mathbf{A}$ une matrice diagonalisable dans $\mathbb{C}$. On note $\lambda_1, \ldots, \lambda_n$ les valeurs propres de $\mathbf{A}$ dans $\mathbb{C}$ et $\mathbf{u}_1, \ldots, \mathbf{u}_n$ ses vecteurs propres dans $\mathbb{C}^n$. On suppose que $|\lambda_1| > |\lambda_2| \geq \cdot \geq |\lambda_n|$.

Ceci implique que $\lambda_1 \in \mathbb{R}$: en effet, si λ_1 est valeur propre, alors $\overline{\lambda_1}$ l'est aussi et $|\lambda_1| = |\overline{\lambda_1}|$ ce qui contredit $|\lambda_1| > |\lambda_2|$.
Soit $\mathbf{x}^{(0)}$ non orthogonal à $\mathbf{v}_1$ défini comme suit :

$$\mathbf{x}^{(0)} = \sum_{i=1}^{n} \alpha_i \mathbf{u}_i,$$

où $\mathbf{v}_1$ est un vecteur propre à gauche de $\mathbf{A}$, $\mathbf{A}^\top \mathbf{v}_1 = \lambda_1 \mathbf{v}_1$. Alors $\alpha_1 \neq 0$ puisque

$$\langle \mathbf{x}^{(0)}, \mathbf{v}_1 \rangle = \sum_{i=1}^{n} \alpha_i \langle \mathbf{u}_i, \mathbf{v}_1 \rangle = \alpha_1 \langle \mathbf{u}_1, \mathbf{v}_1 \rangle,$$

où $\langle \mathbf{u}_1, \mathbf{v}_1 \rangle \neq 0$. Pour le reste, $\langle \mathbf{u}_i, \mathbf{v}_1 \rangle = 0$ pour $i = 2, \ldots, n$ puisqu'en effet,

$$\lambda_i \langle \mathbf{u}_i, \mathbf{v}_1 \rangle = \langle \mathbf{A}\mathbf{u}_i, \mathbf{v}_1 \rangle = \langle \mathbf{u}_1, \mathbf{A}^T \mathbf{v}_1 \rangle = \lambda 1 \langle \mathbf{u}_i, \mathbf{v}_1 \rangle.$$

D'où $\langle \mathbf{u}_i, \mathbf{v}_1 \rangle = 0$ si $\lambda_i \neq \lambda_1$ (sinon $\mathbf{v}_1 = \mathbf{0}$).

2. On calcule la suite $\mathbf{x}^{(k)}$ définie par $\mathbf{x}^{(k+1)} = \mathbf{A}x^{(k)}$. Autrement dit :

$$\mathbf{x}^{(k)} = \mathbf{A}^k x^{(0)} = \mathbf{A}^k \sum_{i=1}^{n} \alpha_i \mathbf{u}_i = \sum_{i=1}^{n} \alpha_i \mathbf{A}^k \mathbf{u}_i = \sum_{i=1}^{n} \alpha_i \lambda_i^k \mathbf{u}_i,$$

donc,

$$\frac{\mathbf{x}^{(k)}}{\lambda_1^k} = \alpha_1 \mathbf{u}_1 + \sum_{i=2}^{n} \alpha_i \left(\frac{\lambda_i}{\lambda_1} \right)^k \mathbf{u}_i.$$

Or, $\lim_{k \to \infty} \left(\frac{\lambda_i}{\lambda_1} \right)^k = 0$ car $|\lambda_i| < |\lambda_1|$ pour tout $i = 2, \ldots, n$. Finalement, on obtient

$$\lim_{k \to \infty} \frac{\mathbf{x}^{(k)}}{\lambda_1^k} = \alpha_1 \mathbf{u}_1,$$

qui est un vecteur propre associé à λ_1.

□

Remarques.

1. La suite $\mathbf{x}^{(k)}$ est normée à chaque étape pour éviter que les composantes du vecteur ne deviennent trop grandes ou trop petites.
2. Le facteur de convergence vers la première valeur propre est en $\mathcal{O}(|\lambda_2/\lambda_1|)$. La convergence est donc d'autant plus rapide que $|\lambda_1|$ et $|\lambda_2|$ sont distants l'un de l'autre.
3. Il y a encore convergence vers la première valeur propre et un vecteur propre correspondant lorsque la valeur propre $|\lambda_1|$ est multiple et vérifie $\lambda_1 = \lambda_2 = \cdots = \lambda_p$ et $|\lambda_p| > |\lambda_{p+1}|$.

Exemple 2.3.1. Soit $\mathbf{A}$ la matrice définie par $\mathbf{A} = \begin{pmatrix} -2 & -3 \\ 6 & 7 \end{pmatrix}$.

1. Calculer les valeurs propres de $\mathbf{A}$ et vecteurs propres associés.

2. En appliquant cinq itérations de la méthode de la puissance itérée, en partant de $\mathbf{x}^{(0)} = (1,1)^\top$, proposer une approximation de la valeur propre correspondant au rayon spectral de $\mathbf{A}$ ainsi que le vecteur propre associé.
3. Reprendre les questions précédentes en prenant $\mathbf{A} = \begin{pmatrix} -4 & -6 \\ -6 & -4 \end{pmatrix}$.

Pour calculer les autres valeurs propres suivantes, on utilise une méthode appelée la *méthode de déflation* caractérisée par le théorème suivante :

Théorème 2.3.2. (Méthode de déflation) Sous les hypothèses du théorème précédent et si $\mathbf{y}_1$ est un vecteur propre à gauche de $\mathbf{A}$ associé à la valeur propre λ_1, alors la matrice

$$\mathbf{B} = \mathbf{A} - \frac{\lambda_1}{\mathbf{y}_1^\top \mathbf{u}_1} \mathbf{u}_1 \mathbf{y}_1^\top,$$

a pour valeurs propres : $0, \lambda_2, \ldots, \lambda_n$ associées respectivement aux vecteurs propres $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n$.

Ces deux théorèmes permettent théoriquement de calculer toutes les valeurs propre d'une matrice $\mathbf{A}$ pour laquelle $|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n|$. Pratiquement, ce procédé est limité au calcul de quelques valeurs propres en raison de l'accumulation des erreurs numériques (erreurs d'arrondi).

2.3.2 Accélération de la convergence

La méthode de la puissance inverse définie ci-haut peut ne pas converger aussi rapidement pour le calcul d'une valeur propre. Lorsque la matrice $\mathbf{A}$ est symétrique, pour accélérer la convergence, on définit une suite des réels R_k appelée *méthode du quotient de Rayleigh*, comme suit

$$\begin{cases} \mathbf{x}^{(0)} & \text{donné dans } \mathbb{R}^n \\ \mathbf{x}^{(k+1)} & = \mathbf{A}\mathbf{x}^{(k)} \\ R_k & = \dfrac{\mathbf{x}^{(k)^\top} \mathbf{A}\mathbf{x}^{(k)}}{\mathbf{x}^{(k)^\top}\mathbf{x}^{(k)}} = \dfrac{\mathbf{x}^{(k)^\top}\mathbf{x}^{(k+1)}}{\|\mathbf{x}^{(k)}\|_2^2}. \end{cases}$$

Si $\mathbf{x}^{(0)}$ n'est pas orthogonal au sous-espace propre à gauche associé à $\mathbf{u}_1$ alors la suite (R_k) converge et $\lim\limits_{k\to\infty} R_k = \lambda_1$. La vitesse de convergence est $\mathcal{O}(|\lambda_2/\lambda_1|^2)$.

2.3.3 Méthode de la puissance itérée inverse

Si λ est une valeur propre de $\mathbf{A}$ et $\mathbf{A}$ est inversible, alors, de façon équivalente, λ^{-1} est une valeur propre de $\mathbf{A}^{-1}$ et les vecteurs propres associés sont les mêmes. On a, en effet :

$$\mathbf{A}\mathbf{y} = \lambda\mathbf{y} \iff \mathbf{y} = \lambda\mathbf{A}^{-1}\mathbf{y} \iff \mathbf{A}^{-1}\mathbf{y} = \lambda^{-1}\mathbf{y}.$$

Par ailleurs, définissons, pour q scalaire donné, la matrice

$$\mathbf{B} = \mathbf{A} - q\mathbf{I}.$$

Cette matrice admet comme valeurs propres $\lambda_i - q$ où λ_i sont les valeurs propres de $\mathbf{A}$. La matrice $\mathbf{B}^{-1}$ (si elle est définie, c'est-à-dire que si q n'est pas une valeur propre de $\mathbf{A}$) admet pour valeurs propres

$$\mu_i = \frac{1}{\lambda_i - q}.$$

En effet, si z est vecteur propre de $\mathbf{B}^{-1}$ associé à $(\lambda_i - q)^{-1}$, il est vecteur propre de $\mathbf{B}$, et donc de $\mathbf{A}$. Et il est évidemment associé à la valeur propre λ_i. Supposons que $\mathbf{A}$ est une matrice diagonalisable dans $\mathbb{R}$ et q est un réel. Soit λ_j une valeur propre de $\mathbf{A}$ qui vérifie

$$0 < |q - \lambda_j| < |q - \lambda_i|, \qquad \forall i \neq j,$$

c'est-à-dire que q n'est pas valeur propre de $\mathbf{A}$ et λ_j est la valeur propre la plus proche de q. Il résulte donc que la matrice $\mathbf{B}^{-1}$ définie par

$$\mathbf{B}^{-1} = (\mathbf{A} - q\mathbf{I})^{-1},$$

admet comme valeur propre dominante la valeur propre ν_1 définie par

$$\nu_1 = \frac{1}{\lambda_j - q}.$$

Donc, pour q donné, on peut s'inspirer de l'algorithme de la puissance itérée appliqué à $\mathbf{B}^{-1}$ pour calculer ν_1 et on retrouver λ_j en posant

$$\lambda_j = \frac{1}{\nu_1} + q.$$

Cette méthode appelée *méthode de la puissance itérée inverse* peut se structurer comme suit. Posons $\mathbf{B} = \mathbf{A} - q\mathbf{I}$. On donne $\mathbf{x}^{(0)}$ arbitraire et on définit la suite $\mathbf{x}^{(n+1)}$ par

$$\begin{cases} \mathbf{B}\mathbf{u}^{(k+1)} & = \mathbf{x}^{(k)}, \\ \mathbf{x}^{(k+1)} & = \dfrac{\mathbf{u}^{(k+1)}}{\|\mathbf{u}^{(k+1)}\|}, \quad k \geq 0. \end{cases}$$

La méthode de la puissance itérée inverse peut aussi être formulée comme suit :

$$\begin{cases} \mathbf{x}^{(0)} & \neq \mathbf{0} \text{ donné}, \\ \mathbf{x}^{(k+1)} & = \dfrac{\mathbf{A}^{-1}\mathbf{x}^{(k)}}{\|\mathbf{A}\mathbf{x}^{(k)}\|}, \quad k \geq 0. \end{cases}$$

Vu de cet angle, on peut, en pratique calculer $\mathbf{x}^{(k+1)}$ en effectuant une factorisation de $\mathbf{A}$ par la méthode de Cholesky(ou la méthode LU). Et on résout le système linéaire $\mathbf{L}\mathbf{L}^\top\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)}$ (ou $\mathbf{L}\mathbf{U}^\top\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)}$).

Exemple 2.3.2. Soit $\mathbf{A}$ la matrice définie par $\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$.

1. Calculer le polynôme caractéristique de $\mathbf{A}$

2. Expliciter les éléments propres de $\mathbf{A}$. En prenant $\mathbf{x}^{(0)} = (1,1)^\top$, appliquer trois itérations de la méthode de la puissance itérée pour approcher la plus grande valeur propre de $\mathbf{A}$ en module ainsi que le vecteur propre associé.
3. En prenant $\mathbf{x}^{(0)} = (1,1)^\top$, appliquer trois itérations de la méthode de la puissance itérée inverse pour approcher la plus petite valeur propre de $\mathbf{A}$ en module ainsi que le vecteur propre associé.
4. En déduire le rayon spectral de $\mathbf{A}$.

2.4 Méthode de Jacobi pour les matrices réelles symétriques

Définition 2.4.1. (Matrice orthogonale). Une matrice $\mathbf{U} \in \mathbb{C}^{n,p}$ est dite orthogonale si, et seulement si,

$$\mathbf{U}^*\mathbf{U} = \mathbf{I},$$

où $\mathbf{I}$ est l'identité dans $\mathbb{R}^{p,p}$. Les colonnes de $\mathbf{U}$ sont alors orthogonales.

Remarque. Si $\mathbf{U}$ est une matrice réelle, carrée ($p = n$) et orthogonale alors $\mathbf{U}$ est inversible est $\mathbf{U}^{-1} = \mathbf{U}^\top$.

Exemple 2.4.1.

1. Soit $\omega \in \mathbb{R}^n$ un vecteur tel que $\omega^\top\omega = 1$. Alors, la matrice $\mathbf{H}$ définie par $\mathbf{H} = \mathbf{I} - 2\omega\omega^\top$ est orthogonale est appelée la *transformation de Householder*. Remarquons qu'une transformation de Householder est aussi symétrique. Donc, $\mathbf{H}^{-1} = \mathbf{H}$. (prouvez-le ! ! !)
2. La matrice de rotation définie par

$$\mathbf{P} = \begin{pmatrix} \cos\theta & 0 & \sin\theta \\ 0 & 1 & 0 \\ -\sin\theta & 0 & \cos\theta \end{pmatrix}$$

est une matrice orthogonale (vérifiez-le ! ! !).

Lemme 2.4.1. Soit $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{n,n}$ une matrice symétrique et (i_0, j_0) un couple d'indices appartenant à l'ensemble $\{1, \ldots, n\}^2$ avec $i_0 \neq j_0$ et soient :

$$\varphi = \frac{1}{2}\arctan\left(\frac{2a_{i_0j_0}}{a_{j_0j_0} - a_{i_0i_0}}\right),$$

et $\mathbf{P} = (p_{ij}) \in \mathbb{R}^{n,n}$ définie par :

$$\begin{cases} p_{i_0i_0} &= p_{j_0j_0} = \cos\varphi \\ p_{i_0j_0} &= -p_{j_0i_0} = \sin\varphi \\ p_{ij} &= \delta_{ij} \qquad \text{sinon.} \end{cases} \tag{2.2}$$

Alors, la matrice $\mathbf{B} = \mathbf{P}^\top \mathbf{A}\mathbf{P}$ vérifie :

$$\begin{cases} b_{ii} = a_{ii}, & \text{si } i \neq i_0 \text{ et } j \neq j_0 \\ b_{i_0 j_0} = b_{j_0 i_0} & = 0 \\ b^2_{i_0 i_0} + b^2_{j_0 j_0} & = a^2_{i_0 i_0} + a^2_{j_0 j_0} + 2a_{i_0 j_0}. \end{cases}$$

On dit que la matrice $\mathbf{P}$ « réduit à zéro » l'élément $a_{i_0 j_0}$.

Théorème 2.4.1. Soit $\mathbf{A} \in \mathbb{R}^{n,n}$ une matrice réelle symétrique. On construit par récurrence une suite de matrices $(\mathbf{B}_k)$, $\mathbf{B}_k \in \mathbb{R}^{n,n}$ par

$$\begin{cases} \mathbf{B}_0 & = \mathbf{A} \\ \mathbf{B}_k & = \mathbf{P}_{k-1}^\top \mathbf{B}_{k-1} \mathbf{P}_{k-1}, \qquad \forall k \in \mathbb{N}^* \end{cases}$$

où $\mathbf{P}_{k-1}$ est la matrice orthogonale $\mathbf{P}$ donnée par (2.2) avec i_0 et j_0 tels que $i_0 \neq j_0$ et :

$$\left| b^{(k-1)}_{i_0 j_0} \right| = \max_{m,l : m \neq l} \left| b^{(k-1)}_{ml} \right|,$$

où $\mathbf{B}_k = (b^{(k)}_{ij})_{i,j=1,\ldots,n}$, alors :

— la suite $(\mathbf{B}_k)$ converge vers une matrice diagonale $\mathbf{\Lambda}$ quand k tend vers $+\infty$;

— les éléments diagonaux de $\mathbf{\Lambda}$ sont les valeurs propres de $\mathbf{A}$.

Si, de plus, on note $\mathbf{Q}_k = \mathbf{P}_0 \mathbf{P}_1 \cdots \mathbf{P}_k$, alors :

— la suite $(\mathbf{Q}_k)$ a une limite notée $\mathbf{X}$ quand k tend vers $+\infty$,

— $\mathbf{\Lambda} = \mathbf{X}^\top \mathbf{A}\mathbf{X}$;

— les colonnes de $\mathbf{X}$ sont des vecteurs propres de $\mathbf{A}$.

Démonstration. On note la matrice $\mathbf{B}_k = (b^{(k)}_{ij}, i, j = 1, \ldots, n$ et, d'après le lemme (2.4.1), on a : $b^{(k)}_{i_0 j_0} = b^{(k)}_{j_0 i_0} = 0$ et si $l \neq i_0$ et $l \neq j_0$, alors $b^{(k)}_{ll} = b^{(k-1)}_{ll}$. Montrons que, sous les hypothèses du théorème

$$\sum_{i \neq j} \left(b^{(k)}_{ij} \right)^2 \leq \left(1 - \frac{2}{n^2 - n} \right)^k \sum_{i \neq j} a^2_{ij}.$$

Pour tout $n \in \mathbb{N}^*$:

$$\operatorname{tr}\left(\mathbf{B}_k^\top \mathbf{B}_k \right) = \sum_{i,j=1} \left(b^{(k)}_{ij} \right)^2,$$

et

$$\sum_{i \neq j} \left(b^{(k)}_{ij} \right)^2 = \operatorname{tr}\left(\mathbf{B}_k^\top \mathbf{B}_k \right) - \sum_{l=1}^{n} \left(b^{(k)}_{ll} \right)^2 = \operatorname{tr}\left(\mathbf{B}_{k-1}^\top \mathbf{B}_{k-1} \right) - \sum_{l=1}^{n} \left(b^{(k)}_{ll} \right)^2,$$

car la matrice $\mathbf{B}_k$ est unitairement semblable à $\mathbf{B}_{k-1}$ (c'est-à-dire qu'il existe une matrice orthogonale $\mathbf{P}$ telle que $\mathbf{B}_k = \mathbf{P}^\top \mathbf{B}_{k-1} \mathbf{P}$). Donc $\mathbf{B}_k^\top \mathbf{B}_k$ est semblable) $\mathbf{B}_{k-1}^\top \mathbf{B}_{k-1}$ et les traces sont conservées. Par conséquent, d'après le lemme (2.4.1) :

$$\sum_{i \neq j} \left(b^{(k)}_{ij} \right)^2 = \operatorname{tr}\left(\mathbf{B}_{k-1}^\top \mathbf{B}_{k-1} \right) - \sum_{l=1}^{n} \left(b^{(k-1)}_{ll} \right)^2 - 2 \left(b^{(k-1)}_{i_0 j_0} \right)^2 = \sum_{i \neq j} \left(b^{(k-1)}_{ij} \right)^2 - 2 \left(b^{(k-1)}_{i_0 j_0} \right)^2.$$

Il en résulte du choix de i_0 et j_0 que :

$$\left(b_{i_0 j_0}^{(k-1)}\right)^2 \geq \left(\frac{1}{n^2-n}\right) \sum_{i \neq j} \left(b_{ij}^{(k-1)}\right)^2 .$$

D'où

$$\sum_{i \neq j} \left(b_{ij}^{(k)}\right)^2 \leq \left(1 - \frac{2}{n^2-n}\right) \sum_{i \neq j} \left(b_{ij}^{(k-1)}\right)^2 \leq \left(1 - \frac{2}{n^2-n}\right) \sum_{i \neq j} a_{ij}^2.$$

Il résulte alors que les termes non diagonaux de $\mathbf{B}_k$ tendent vers 0 lorsque k tend vers l'infini et que la suite $(\mathbf{B}_k)$ tend vers une matrice diagonale notée $\mathbf{\Lambda}$. La suite $(\mathbf{Q}_k)$ admet donc aussi une limite notée $\mathbf{X}$.

A la limite, nous avons $\mathbf{\Lambda} = \mathbf{X}^\top \mathbf{A}\mathbf{X}$, soit $\mathbf{A}\mathbf{X} = \mathbf{X}\mathbf{\Lambda}$. Si $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ sont les colonnes de $\mathbf{X}$ et $\lambda_1, \lambda_2, \ldots, \lambda_n$ les éléments diagonaux de $\mathbf{\Lambda}$, alors $\lambda_1 \mathbf{X}_1, \ldots, \lambda_n \mathbf{X}_n$ sont les colonnes de $\mathbf{X}\mathbf{\Lambda}$. Par conséquent, les $\lambda_1, \ldots, \lambda_n$ sont les valeurs propres de $\mathbf{A}$ associées aux vecteurs propres $\mathbf{X}_1, \ldots, \mathbf{X}_n$. □

2.4.1 Mise en œuvre de la méthode de Jacobi

Exemple 2.4.2. Soit $\mathbf{A}$ la matrice définie par

$$\mathbf{A} = \begin{pmatrix} 2 & -1 & 1 \\ -1 & 3 & -4 \\ 1 & -4 & 3 \end{pmatrix}.$$

Proposer une matrice semblable à $\mathbf{A}$ en appliquant une itération de la méthode de Jacobi.

2.5 Exercices

Exercice 2.5.1. Déterminer les cercles de Gershgorin for la matrice A définie par

$$A = \begin{pmatrix} 4 & 1 & 1 \\ 0 & 2 & 1 \\ -2 & 0 & 9 \end{pmatrix},$$

puis trouver un encadrement du rayon spectral de A.

Exercice 2.5.2.

1. Rappeler et démontrer le théorème de Gershgorin.
2. Localiser les valeurs propres des matrices suivantes

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & -1 \\ 2 & 7 & 0 \\ -1 & 0 & 5 \end{pmatrix} \qquad \mathbf{B} = \begin{pmatrix} 10 & 2 & 3 \\ -1 & 2 & -1 \\ 0 & 1 & 3 \end{pmatrix}.$$

Exercice 2.5.3.

1. Calculer les valeurs propres et les vecteurs propres de

$$\mathbf{A} = \begin{pmatrix} 10 & 0 \\ -9 & 1 \end{pmatrix}.$$

2. Que donne la méthode de la puissance pour la matrice $\mathbf{A}$ en partant de $\mathbf{x}_0 = (2,1)^\top$?
3. Calculer les valeurs propres et les vecteurs propres v_1 et v_2 de

$$\mathbf{A} = \begin{pmatrix} 1 & -3 \\ -3 & 1 \end{pmatrix}.$$

4. Exprimer $\mathbf{x}_0 = (1,0)^\top$ en fonction de v_1 et v_2. En déduire l'expression de $\mathbf{A}^k\mathbf{x}_0$, puis de $\mathbf{A}^k\mathbf{x}_0/\|\mathbf{A}^k\mathbf{x}_0\|$ et conclure.

Exercice 2.5.4.

1. Utiliser la méthode de la puissance itérée pour calculer une approximation de la valeur propre dominante de la matrice

$$A = \begin{pmatrix} -4 & 14 & 0 \\ -5 & 13 & 0 \\ -1 & 0 & 2 \end{pmatrix}$$

ensuite appliquer la méthode de déflation pour l'approximation des autres valeurs propres de la matrice A.

2. Appliquer la méthode de la puissance à la matrice suivante en utilisant la méthode d'accélération de la convergence

$$A = \begin{pmatrix} 4 & -1 & 1 \\ -1 & 3 & -2 \\ 1 & -2 & 3 \end{pmatrix}.$$

Exercice 2.5.5. Appliquer la méthode de la puissance inverse avec $\mathbf{x}^{(0)} = (1,1,1)^t$ à la matrice A suivante

$$A = \begin{pmatrix} -4 & 14 & 0 \\ -5 & 13 & 0 \\ -1 & 0 & 2 \end{pmatrix} \qquad \text{avec} \qquad q = \frac{\mathbf{x}^{(0)t} A \mathbf{x}^{(0)}}{\mathbf{x}^{(0)t}\mathbf{x}^{(0)}} = \frac{19}{3},$$

Exercice 2.5.6.

1. Trouver une matrice de rotation P avec la propriété que PA ait un zéro à la seconde ligne et première colonne, avec

$$A = \begin{pmatrix} 4 & 1 & -2 & 2 \\ 1 & 2 & 0 & 1 \\ -2 & 0 & 3 & -2 \\ 2 & 1 & -2 & -1 \end{pmatrix}$$

2. Appliquer deux itérations de la méthode de Jacobi à la matrice A pour estimer ses valeurs propres.

Exercice 2.5.7.

1. Trouver une matrice de rotation P avec la propriété que PA ait un zéro à la seconde ligne et première colonne, avec
$$A = \begin{pmatrix} 3 & 1 & 0 \\ 1 & 3 & 1 \\ 0 & 1 & 3 \end{pmatrix}.$$
2. Proposer une approximation des valeurs propres et des vecteurs propres de A en utilisant la méthode de Jacobi.

Exercice 2.5.8. Déterminer la matrice semblable à $\mathbf{A}$ donnée par la première itération de la méthode de Jacobi avec
$$\mathbf{A} = \begin{pmatrix} 2 & -1 & 3 & 1 \\ -1 & 3 & 2 & 0 \\ 3 & 2 & 1 & -1 \\ 1 & 0 & -1 & 2 \end{pmatrix}.$$

CHAPITRE 3

Introduction à l'optimisation

Sommaire

3.1 Introduction

Soit $\mathcal{V}$ un espace vectoriel normé ; muni de la norme $\|\cdot\|$. Dans ce chapitre, on s'intéresse au problème suivant :

$$\min_{x\in\mathcal{X}} f(x), \tag{3.1}$$

où $\mathcal{X} \subset \mathcal{V}$ et $f : \mathcal{X} \to \mathbb{R}$ est une fonction appelée *fonction coût* ou *critère*.

1. Si $\mathcal{X} = \mathcal{V}$, on dit que le problème (3.1) est un problème d'optimisation sans contraintes.
2. Si $\mathcal{X} \subsetneq \mathcal{V}$, on dit que le problème (3.1) est un problème d'optimisation sous contraintes.
3. Si $\dim(\mathcal{X}) < \infty$, on dit que le problème (3.1) est un problème d'optimisation en dimension finie.

Définition 3.1.1.

1. Un point $x^* \in \mathcal{X}$ est un minimiseur global du problème (3.1) si x^* est une solution de (3.1).
2. Un point x^* est un minimiseur local s'il existe un voisinage $\mathfrak{V} \subset \mathcal{X}$ de x^* tel que x^* soit solution du problème

$$\min_{x\in\mathfrak{V}} f(x).$$

Remarque. Le problème (3.1) peut aussi caractériser un problème de maximisation puisque

$$\max_{x\in\mathcal{X}} f(x) = -\min_{x\in\mathcal{X}} \left(-f(x)\right).$$

Proposition 3.1.1. Soit $\mathcal{X}$ une partie minorée non vide de $\mathbb{R}$. Alors les assertions suivantes sont équivalentes :

i. $m = \inf\{x, x \in \mathcal{X}\}$.

ii. $\forall \varepsilon > 0$, $\exists x \in \mathcal{X}$ tel que $m \leq x \leq m + \varepsilon$.

iii. m est un minimisant de $\mathcal{X}$ et il existe $(x_n)_{n\in\mathbb{N}} \subset \mathcal{X}^{\mathbb{N}}$ appelée *suite minimisante.*

3.2 Étude théorique des problèmes d'optimisation

On considère le problème défini comme suit

$$\begin{aligned} &\min_{x\in\mathbb{R}^n} \mathcal{J}(x) && \text{(Problème sans contrainte)} \\ &\min_{x\in\mathcal{C}} \mathcal{J}(x) && \text{(Problème sous contrainte)}. \end{aligned}$$

On envisage deux types de contraintes :

$$\begin{aligned} \text{Contraintes d'égalités :} \quad & \mathcal{C} = \{x \in \mathbb{R}^n : \varphi_i(x) = 0, \forall i \in I\} \\ \text{Contraintes d'inégalités :} \quad & \mathcal{C} = \{x \in \mathbb{R}^n : \varphi_i(x) \leq 0, \forall i \in I\} \end{aligned}$$

Les fonctions φ_i étant des fonctions continues de $\mathbb{R}^n$ vers $\mathbb{R}$.

On distingue les catégories suivantes des problèmes d'optimisation :

- Programmation linéaire lorsque $\mathcal{J}$ est linéaire et $\mathcal{V}$ est un polyèdre convexe défini par

$$\mathcal{V} = \{x \in \mathbb{R}^n : Bx \leq b\},$$

  où B est une matrice $m \times n$ et b un élément de $\mathbb{R}^{n\times 1}$.
- Programmation quadratique lorsque $\mathcal{J}$ est de la forme

$$\mathcal{J}(x) = \frac{1}{2}\langle Ax, x\rangle_{\mathbb{R}^n} + \langle b, x\rangle_{\mathbb{R}^n} + c,$$

  où A est une matrice symétrique, $b \in \mathbb{R}^{n,1}$ un vecteur colonne et c un réel et $\mathcal{V}$ un polyèdre convexe.

3.2.1 Existence et unicité de la solution

Théorème 3.2.1. Soit $\mathcal{J}$ une fonction continue sur un sous-ensemble $\mathcal{V}$ fermé de $\mathbb{R}^n$. On suppose que

— Ou bien $\mathcal{V}$ est borné.

— Ou bien $\mathcal{V}$ est non borné et $\lim\limits_{\|x\|\to+\infty} \mathcal{J}(x) = +\infty$ (on dit, dans ce cas que $\mathcal{J}$ est coercive).

Alors, $\mathcal{J}$ possède un minimum sur $\mathcal{V}$.

Définition 3.2.1.

1. Un ensemble $\mathcal{V}$ est dit convexe si, pour tous points x et y de $\mathcal{V}$, le segment $[x, y]$ est inclus dans $\mathcal{V}$, c'est-à-dire que $\forall t \in [0, 1]$, le point $tx + (1 - t)y \in \mathcal{V}$.
2. Une fonction $\mathcal{J}$ définie sur un ensemble convexe $\mathcal{V}$ est dite convexe si

$$\forall (x, y) \in \mathcal{V}^2, \forall t \in [0, 1], \quad \mathcal{J}(tx + (1 - t)y) \leq t\mathcal{J}(x) + (1 - t)\mathcal{J}(y).$$

Proposition 3.2.1. Soit $\mathcal{J}$ une fonction différentiable sur un convexe $\mathcal{V} \subset \mathbb{R}^n$ et à valeurs réelles. La fonction $\mathcal{J}$ est convexe si, et seulement si,

$$\forall (x, y) \in \mathcal{V}^2, \langle \nabla \mathcal{J}(x), y - x \rangle_{\mathbb{R}^n} \leq \mathcal{J}(y) - \mathcal{J}(x).$$

Proposition 3.2.2. Soit $\mathcal{J}$ une fonction convexe définie sur un ensemble convexe $\mathcal{V}$ de $\mathbb{R}^n$. Alors :

* Tout minimum local de $\mathcal{J}$ sur $\mathcal{V}$ est un minimum global.
* Si $\mathcal{J}$ est strictement convexe, il y a au plus un minimum global.

3.2.2 Conditions d'optimalités

On suppose, dans cette section, que $\mathcal{J} : \mathbb{R}^n \to \mathbb{R}$ est une fonction continue deux fois différentiable. On notera x^* un minimum local de $\mathcal{J}$.

Théorème 3.2.2. (Condition nécessaire d'optimalité)
Soit x^* un minimum du problème

$$\min_{x \in \mathbb{R}^n} \mathcal{J}(x).$$

Alors, x^* vérifie nécessairement :

- (Conditions au premier ordre) : Si $\mathcal{J}$ est différentiable en x^*, alors on a $\nabla \mathcal{J}(x^*) = 0$.
- (Conditions au deuxième ordre) : Si $\mathcal{J}$ est deux fois différentiable au point x^*, alors la forme quadratique $\mathcal{D}^2\mathcal{J}(x^*)$ est semi-définie positive, c'est-à-dire

$$\forall y \in \mathbb{R}^n, \quad \langle \mathcal{D}^2\mathcal{J}(x^*)y, y \rangle_{\mathbb{R}^n} \geq 0.$$

  où $\mathcal{D}^2\mathcal{J}(x^*)$ est la matrice hessienne, définie par les cœfficients $\dfrac{\partial^2 \mathcal{J}(x^*)}{\partial x_i \partial x_j}$.

Théorème 3.2.3. (Conditions suffisantes d'optimalité)
Soit $\mathcal{J}$ une fonction de classe $\mathcal{C}^1$ définie sur $\mathbb{R}^n$. On suppose que $\nabla \mathcal{J}(x^*) = 0$ et que $\mathcal{J}$ est deux fois différentiable en x^*. Alors, x^* est un minimum local de $\mathcal{J}$ si l'une des deux conditions suivantes est vérifiée :

1. $\mathcal{D}^2\mathcal{J}(x^*)$ est définie positive.
2. $\exists r > 0$ tel que $\mathcal{J}$ est deux fois différentiable sur $\mathcal{B}(x^*, r)$ et la forme quadratique $\mathcal{D}^2\mathcal{J}(x)$ est semi-définie positive pour tout $x \in \mathcal{B}(x^*, r)$.

Proposition 3.2.3. (CNS, cas convexe)
Soit $\mathcal{J}$ une fonction convexe de classe $\mathcal{C}^1$ définie positive sur $\mathbb{R}^n$ et x^* un point de $\mathbb{R}^n$. Alors, x^* est un minimum global de $\mathcal{J}$ si, et seulement si : $\nabla \mathcal{J}(x^*) = 0$.

3.3 Algorithmes pour l'optimisation sans contrainte

Dans cette partie, on considère essentiellement une classe d'algorithmes qui se formulent comme suit : avec $x^{(0)}$ une solution initiale donnée, on cherche à calculer :

$$x^{(k+1)} = x^{(k)} + \rho^{(k)} d^{(k)},$$

avec $d^{(k)}$ la direction de la descente et $\rho^{(k)}$ le pas de la méthode à la k-ème itération. En particulier, on s'arrange pour satisfaire :

$$\mathcal{J}(x^{(k+1)}) \leq \mathcal{J}(x^{(k)}).$$

De tels algorithmes sont souvent appelés méthodes de descente. Essentiellement, la différence entre ces algorithmes réside dans le choix de la direction de descente $d^(k)$. Cette direction étant choisie, nous sommes plus ou moins ramenés à un problème unidimensionnel pour la détermination de $\rho^(k)$. Pour ces raisons, commençons par analyser ce qui se passe dans le cas de la dimension un.

3.3.1 Méthode de la section dorée

Soit $\rho \longmapsto q(\rho)$ la fonction coût que l'on cherche à minimiser. On pourra prendre par exemple

$$q(\rho) = \mathcal{J}\left(x^{(k)} + \rho d^{(k)}\right)$$

afin d'appliquer les idées au de la méthode de descente. Supposons que l'on connaisse un intervalle $[a, b]$ contenant le minimum ρ^* de q et tel que q soit décroissante sur $[a, \rho^*]$ et croissante sur $]\rho^*, b]$. On appelle alors q une fonction unimodale.

On construit une suite décroissante d'intervalles $[a_i; b_i]$ qui contiennent tous le minimum ρ^*. Pour passer de $[a_i; b_i]$ à $[a_{i+1}; b_{i+1}]$, on procède de la manière suivante. On introduit deux nombres a' et b' de l'intervalle $[a_i; b_i]$ et tels que $a' < b'$. Puis, on calcule les valeurs $q(a')$ et $q(b')$. Trois possibilités se présentent alors à nous. Si $q(a') < q(b')$, alors, le minimum ρ^* se trouve nécessairement à gauche de b' . Ceci définit alors le nouvel intervalle en posant $a_{i+1} = a_i$ et $b_{i+1} = b'$. Considérons maintenant que l'inégalité : $q(a') > q(b')$ est satisfaite. Dans ce second cas, il est évident que le minimum se trouve cette fois à droite de a' . On pose alors : $a_{i+1} = a'$ et $b_{i+1} = b_i$. Enfin, le dernier cas consiste à avoir $q(a') = q(b')$. Alors, le minimum se trouve dans l'intervalle $[a'; b']$. On se restreint donc à $a_{i+1} = a'$ et $b_{i+1} = b'$.

La question suivante se pose : comment choisir a' et b' en pratique ? En général, on privilégie deux aspects :

i. on souhaite que le facteur de réduction τ, qui représente le ratio du nouvel intervalle par rapport au précédent, soit constant,

ii. on désire réutiliser le point qui n'a pas été choisi dans l'itération précédente afin de diminuer les coûts de calculs.

On peut montrer que la vérification simultanée de ces deux contraintes conduit à un choix unique des paramètres a' et b'. Plus précisément, supposons que q est unimodale. Alors, on

obtient l'algorithme (3) suivant dit de la section dorée, la méthode tirant son nom de la valeur du paramètre τ.

Données: Un intervalle $[a, b]$ et une fonction $\mathcal{J}$ à minimiser
Résultat: Minimum global x^* de $\mathcal{J}$.

poser $\tau = \dfrac{1+\sqrt{5}}{2}$;
poser $a_0 = a$;
poser $b_0 = b$;
pour $i = 0, \ldots, N_{\max}$ **faire**
    poser $a' = a_i + \frac{1}{\tau^2} b_i - a_i)$;
    poser $b' = a_i + \frac{1}{\tau}(b_i - a_i)$;
    si $(q(a') < q(b'))$ **alors**
        poser $a_{i+1} = a_i$;
        poser $b_{i+1} = b'$;
        sinon si $(q(a') > q(b'))$ **alors**
            poser $a_{i+1} = a'$;
            poser $b_{i+1} = b_i$;
            sinon si $(q(a') = q(b'))$ **alors**
                poser $a_{i+1} = a'$;
                poser $b_{i+1} = b'$;
            finsi
        finsi
    finsi
finPour

Algorithm 3: Algorithme de la méthode de la section dorée.

Ici, $N_{\max}$ est le nombre maximal d'itérations que l'on se fixe. A cette fin, on doit valider un critère d'arrêt de la forme : $|b_{i+1} - a_{i+1}| < \varepsilon$, où ε est l'erreur (ou tolérance) que l'on se permet sur la solution ρ^* du problème.

3.3.2 Méthode de gradient à pas fixe ou optimal

La méthode du gradient fait partie des classes de méthodes dites de descente. Quelle est l'idée cachée derrière ces méthodes ? Considérons un point de départ $x^{(0)}$ et cherchons à minimiser une fonction $\mathcal{J}$. Puisque l'on veut atteindre x^*, nous cherchons à avoir : $\mathcal{J}(x^{(1)}) < \mathcal{J}(x^{(0)})$. Une forme particulièrement simple est de chercher $x^{(1)}$ tel que le vecteur $x^{(1)} - x^{(0)}$ soit colinéaire à une direction de descente $d^{(0)} \neq 0$. Nous le noterons : $x^{(1)} - x^{(0)} = \rho^{(0)} d^{(1)}$, où $\rho^{(0)}$ est le pas de descente de la méthode. La méthode consiste au principe itératif suivant :

$$\begin{aligned} &\text{choisir } x^{(0)} \\ &\text{calculer } x^{(k+1)} \quad = \quad x^{(k)} + \rho^{(k)} \cdot d^{(k)}, \end{aligned}$$

avec $d^{(k)} \in \mathbb{R}^n$ et $\rho^{(k)} > 0$. Pour choisir $d^{(k)}$ et $\rho^{(k)}$, on procède comme suit. Le développement de Taylor de $\mathcal{J}$ au premier ordre au voisinage de $x^{(k+1)}$ donne

$$\begin{aligned}\mathcal{J}\left(x^{(k+1)}\right) &= \mathcal{J}\left(x^{(k)} + \rho^{(k)} \cdot d^{(k)}\right) \\ &= \mathcal{J}\left(x^{(k)}\right) + \rho^{(k)}\langle \nabla \mathcal{J}\left(x^{(k)}\right), d^{(k)}\rangle_{\mathbb{R}^n} + \rho^{(k)}\|d^{(k)}\|_{\mathbb{R}^n} E\left(x^{(k)}; \rho^{(k)} d^{(k)}\right),\end{aligned}$$

où

$$\lim_{\rho^{(k)} \cdot d^{(k)} \to 0} E\left(x^{(k)}, \rho^{(k)} \cdot d^{(k)}\right) = 0.$$

On peut alors prendre $d^{(k)} = -\nabla \mathcal{J}\left(x^{(k)}\right)$, puisque

$$\mathcal{J}\left(x^{(k+1)}\right) - \mathcal{J}\left(x^{(k)}\right) = -\rho^{(k)}\|\nabla \mathcal{J}\left(x^{(k)}\right)\|_{\mathbb{R}^n} + \mathcal{O}(\rho^k).$$

Si $\rho^{(k)}$ est suffisamment petit, on aura bien $\mathcal{J}\left(x^{(k+1)}\right) \leq \mathcal{J}\left(x^{(k)}\right)$. Le choix de $\rho^{(k)}$ se fait de la manière suivante :

* Soit $\rho^{(k)} = \rho$ fixé à priori. On parle d'une méthode de gradient à pas fixe ou constant.
* Soit $\rho^{(k)}$ est choisi comme le minimum de la fonction

$$q(\rho) = \mathcal{J}\left(x^{(k)} - \rho \nabla \mathcal{J}\left(x^{(k)}\right)\right),$$

et on parle de la méthode du gradient à pas optimal.

L'algorithme général de ces deux méthodes peut alors se structurer comme suit

Données: Un intervalle $[a, b]$ et une fonction $\mathcal{J}$ à minimiser
Résultat: Minimum global x^* de $\mathcal{J}$.

```
Poser k = 0;
Choisir x^(0) ∈ [a, b];
Tant que (‖x^(k+1) − x^(k)‖_{R^n} ≥ ε) et (k < k_max) faire
    Calculer d^(k) = −∇J(x^(k)) ;
    Calculer ρ^(k);
    Poser x^(k+1) = x^(k) + ρ^(k) · d^(k);
finTantQue
```

Algorithm 4: Méthodes de gradients à pas optimal.

Théorème 3.3.1. Soit $\mathcal{J} \in \mathcal{C}^1(\mathbb{R}^n) \to \mathbb{R}$, x^* un minimum de $\mathcal{J}$. Supposons que :

1. $\mathcal{J}$ est α-elliptique, c'est-à-dire :

$$\exists \alpha > 0,\ \forall (x, y) \in (\mathbb{R}^n)^2\ :\ \langle \nabla \mathcal{J}(x) - \nabla \mathcal{J}(y), x - y\rangle_{\mathbb{R}^n} \geq \alpha \|x - y\|^2_{\mathbb{R}^n}.$$

2. L'application $\nabla \mathcal{J}$ est lipschitzienne :

$$\exists M > 0;\ \forall (x, y) \in (\mathbb{R}^n)^2 : \|\nabla \mathcal{J}(x) - \nabla \mathcal{J}(y)\|_{\mathbb{R}^n} \leq M \|x - y\|_{\mathbb{R}^n}.$$

S'il existe deux réels a et b tels que $\rho^{(k)}$ satisfasse

$$0 < a < \rho^{(k)} < b < \frac{2\alpha}{M^2},\ \forall k \geq 0,$$

alors la méthode du gradient définie par

$$x^{(k+1)} = x^{(k)} - \rho^{(k)} \nabla \mathcal{J}\left(x^{(k)}\right),$$

converge pour tout choix de $x^{(0)}$ de façon géométrique, c'est-à-dire

$$\exists \beta \in]0,1[:\ \|x^{(k+1)} - x^*\|_{\mathbb{R}^n} \leq \beta^k \|x^{(0)} - x^*\|_{\mathbb{R}^n}.$$

Remarque. Même pour le gradient à pas optimal qui est en principe la meilleure de ces méthodes d'un point de vue de la rapidité de convergence, celle-ci peut être lente car altérée par un mauvais conditionnement de la matrice hessienne de $\mathcal{J}$. Par ailleurs, on peut considérer des critères de convergence sur le gradient de $\mathcal{J}$ en $x^{(k)}$: $\|\nabla \mathcal{J}\left(x^{(k)}\right)\| < \varepsilon_1$.

3.3.3 Méthode du gradient conjugué

Considérons une matrice A, définie positive, et $\mathcal{J}$ la fonctionnelle quadratique définie par

$$\begin{array}{rcl} \mathcal{J} & \mathbb{R}^n \to & \mathbb{R} \\ & x \longmapsto & \mathcal{J}(x) = \dfrac{1}{2}\langle Ax, x\rangle_{\mathbb{R}^n} - \langle b, x\rangle_{\mathbb{R}^n}. \end{array}$$

La fonction $\mathcal{J}$ est alors une fonctionnelle strictement convexe et de classe $\mathcal{C}^\infty$. On calcule

$$\nabla \mathcal{J}(x) = Ax - b \text{ et } \operatorname{Hess} \mathcal{J}(x) = A.$$

Par conséquent, le minimum (unique et global) de $\mathcal{J}$ est réalisé en x^* tel que $Ax^* = b$.

Définition 3.3.1. Deux vecteurs d_1 et d_2 sont dits conjugués pour la matrice A si

$$\langle Ad_2, d_1\rangle_{\mathbb{R}^n} = 0.$$

Ces deux vecteurs sont orthogonaux pour le produit scalaire associé à la matrice A, défini par

$$\langle x, y\rangle_A = \langle Ax, y\rangle_{\mathbb{R}^n}, \quad \forall (x,y) \in (\mathbb{R}^n)^2.$$

Si on connaît k directions conjuguées $d^{(0)}, \ldots, d^{(k-1)}$, en partant d'un point $x^{(0)} \in \mathbb{R}^n$, on calcule $x^{(k+1)}$ tel qu'il satisfasse :

$$\mathcal{J}\left(x^{(k+1)}\right) = \mathcal{J}\left(x^{(k)} + \rho^{(k)} \cdot d^{(k)}\right) = \min_{\rho \in \mathbb{R}} \mathcal{J}\left(x^{(k)} + \rho d^{(k)}\right).$$

En utilisant en utilisant la condition de minimum au premier ordre, la valeur de $\rho^{(k)}$ est donnée

par

$$\rho^{(k)} = -\frac{\langle Ax^{(k)} - b, d^{(k)}\rangle_{\mathbb{R}^n}}{\|d^{(k)}\|_A^2} = -\frac{\langle x^{(k)}, d^{(k)}\rangle_A}{\|d^{(k)}\|_A^2} + \frac{\langle b, d^{(k)}\rangle_{\mathbb{R}^n}}{\|d^{(k)}\|_A^2},$$

en posant $\|d^{(k)}\|_A^2 = \langle d^{(k)}, d^{(k)}\rangle_A^{\frac{1}{2}}$. Or, $x^{(k)} - x^{(0)}$ s'exprime selon les vecteurs $d^{(0)}, \ldots, d^{(k-1)}$, puisque

$$\begin{aligned} x^{(k)} &= x^{(k-1)} + \rho^{(k-1)} \cdot d^{(k-1)} \\ &= x^{(k-2)} + \rho^{(k-2)} \cdot d^{(k-2)} + \rho^{(k-1)} \cdot d^{(k-1)} \\ &= x^{(0)} + \sum_{l=0}^{k-1} \rho^{(l)} \cdot d^{(l)}. \end{aligned}$$

Ainsi, on simplifie :

$$\rho^{(k)} = -\frac{\langle x^{(0)}, d^{(k)}\rangle_A}{\|d^{(k)}\|_A^2} + \frac{\langle b, d^{(k)}\rangle_{\mathbb{R}^n}}{\|d^{(k)}\|_A^2} = -\frac{\langle r^{(0)}, d^{(k)}\rangle_A}{\|d^{(k)}\|_A^2},$$

où le vecteur $r^{(0)}$ est appelé résidu et défini à l'instant initial par

$$r^{(0)} = Ax^{(0)} - b.$$

Le succès de l'algorithme du gradient conjugué est intimement lié à la proposition importante suivante.

Proposition 3.3.1. Le point $x^{(k)}$ est le minimum de $\mathcal{J}$ sur le sous-espace affine passant par $x^{(0)}$ engendré par les vecteurs $\{d^{(0)}, \ldots, d^{(k-1)}\}$.

La connaissance de n directions $\{d^{(0)}, \ldots, d^{(n-1)}\}$ détermine explicitement le minimum du problème. En fait, l'algorithme du gradient conjugué consiste à construire simultanément ces directions conjuguées par le procédé de Gram-Schmidt. Plus précisément, l'algorithme (5) sui-

vant décrit décrit la structure de l'algorithme du gradient conjugué.

Données: Un intervalle $[a, b]$ et une fonction $\mathcal{J}$ à minimiser
Résultat: Minimum global x^* de $\mathcal{J}$.

Poser $k = 0$;
Choisir $x^{(0)} \in [a, b]$;
Choisir $\varepsilon > 0$;
Choisir $\varepsilon_1 > 0$;
Poser $r^{(0)} = \nabla\mathcal{J}\left(x^{(0)}\right)$;
Tant que $\left(\|x^{(k+1)} - x^{(k)}\|_{\mathbb{R}^n} \geq \varepsilon\right)$ *et* $(k \leq k_{\max})$ **faire**
  si $(\|r^{(k)}_{\mathbb{R}^n} < \varepsilon_1)$ **alors**
    Arrêter l'algorithme;
    sinon si $(k = 0)$ **alors**
      Poser $d^{(k)} = r^{(k)}$;
      sinon si $(k \neq 0)$ **alors**
        Calculer $\alpha^{(k)} = -\dfrac{\langle r^{(k)}, d^{(k-1)}\rangle_A}{\|d^{(k-1)}|^2_A}$;
        Poser $d^{(k)} = r^{(k)} + \alpha^{(k)} d^{(k-1)}$;
      finsi
    finsi
    Calculer $\rho^{(k)} = -\dfrac{\langle r^{(k)}, d^{(k)}\rangle_A}{\|d^{(k)}\|^2_A}$;
    Poser $x^{(k+1)} = x^{(k)} + d\rho^{(k)} d^{(k)}$;
    Calculer $r^{(k+1)} = Ax^{(k+1)} - b$;
    Poser $k = k + 1$;
  finsi
finTantQue

Algorithm 5: Méthode du gradient conjugué.

3.3.4 Méthode de Newton

La méthode de Newton n'est pas a proprement parlé une méthode d'optimisation. C'est une méthode de recherche de zéros d'une fonction $f : \mathbb{R}^n \to \mathbb{R}^n$ selon : $F(x) = 0$. L'idée de cette méthode ici est de résoudre l'équation $\nabla\mathcal{J}(x) = 0$, condition nécessaire de premier ordre pour la détection d'extrema d'une fonction. L'équation $\nabla\mathcal{J}(x) = 0$ est donnée par un système $n \times n$ d'équations non linéaires. La méthode s'écrit formellement

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \left[D^2\mathcal{J}(\mathbf{x}^{(k)})\right]^{-1} \nabla\mathcal{J}(\mathbf{x}^{(k)}).$$

Si nous cherchons à résoudre l'équation $f(x) = 0$, tout en supposant que f est de classe $\mathcal{C}^1$, alors l'algorithme de Newton est donné par l'équation

$$x^{(k+1)} = x^{(k)} - \frac{f\left(x^{(k)}\right)}{f'\left(x^{(k)}\right)}.$$

Une interprétation géométrique simple de cette méthode est donnée à partir de la tangente au point $x^{(k)}$. La convergence de la méthode doit être précisée ainsi que la définition de la suite $(f'(x^{(k)}))_k$.

Généralisons maintenant cette méthode au cas d'un champ scalaire $\mathcal{J}$ donné de $\mathbb{R}^n$ dans $\mathbb{R}$. Soit F une fonction de classe $\mathcal{C}^1$. On suppose que l'équation $F(x) = 0$ possède au moins une solution notée x^* et que la matrice $DF(\mathbf{x}^*)$ est une matrice inversible. La continuité de DF permet en fait d'assurer l'inversibilité de $DF(\mathbf{x}^{(k)})$ pour tout point $\mathbf{x}^{(k)}$ se trouvant dans un voisinage de $\mathbf{x}^*$ et permet de définir l'itéré $\mathbf{x}^{(k+1)}$. L'extension de la seconde étape est réalisée par la résolution du système linéaire

$$\left[DF(\mathbf{x}^{(k)}\right]\delta^{(k)} = F(\mathbf{x}^{(k)}),$$

puisqu'on pose $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \delta^{(k)}$. En résumé, l'algorithme de Newton peut s'écrire sous la forme

```
poser k = 0;
choisir x^(0) dans un voisinage de x x*;
choisir ε > 0;
Tant que (||x^(k+1) − x^(k)|| ≥ ε) et (k ≤ k_max) faire
    résoudre le système linéaire [DF(x^(k))]δ^(k) = F(x^(k));
    poser x^(k+1) = x^(k) − δ^(k);
    poser k = k + 1;
finTantQue
```

Théorème 3.3.2. Soit $F : \mathbb{R}^n \to \mathbb{R}^n$ une fonction de classe $\mathcal{C}^1$ et $\mathbf{x}^*$ un zéro de F. On suppose que ce zéro est isolé et que $DF(x^*)$ est inversible (DF est la dérivée premier de F). Alors, il existe une boule $\mathcal{B}(\mathbf{x}^*)$ telle que, pour tout point $x^{(0)} \in \mathcal{B}(\mathbf{x}^*)$, la suite $(\mathbf{x}^{(k)})_k$ définie par la méthode de Newton est entièrement contenue dans $\mathcal{B}$ et converge vers $\mathbf{x}^*$, seul zéro de F dans $\mathcal{B}$. De plus, la convergence est géométrique : il existe $\beta \in]0;1[$ tel que

$$\forall k \geq 0, \qquad \left\|\mathbf{x}^{(k+1)} - \mathbf{x}^*\right\| \leq \beta^k \left\|\mathbf{x}^{(0)} - \mathbf{x}^*\right\|.$$

Par conséquent, si nous choisissons $\mathbf{x}^{(0)}$ suffisamment près de $\mathbf{x}^*$, la méthode de Newton converge.

3.4 Exercices

Exercice 3.4.1. On considère la fonction

$$f(x, y) = \frac{1}{2}x^2 + \frac{1}{2}y^2.$$

En partant du point initial $(x_0, y_0) = (1, 1)$ et en appliquant la méthode du gradient à pas optimal, calculer (x_1, y_1),(x_2, y_2),(x_3, y_3).

Exercice 3.4.2. Soit $a \in \mathbb{R}$. On définit $f_a : (x, y) \mapsto x^2 + y^2 + axy - 2c - 2y$

1. Pour quelles valeurs de a, la fonction f_a est-elle convexe ? Et strictement convexe ?
2. Discuter en fonction des valeurs du paramètre a de l'existence de solutions au problème d'optimisation défini par
$$\inf_{(x,y)\in\mathbb{R}^2} f_a(x,y).$$
3. Lorsque $a \in]-2,2[$, résoudre le problème précédent.

Exercice 3.4.3. Soit a une forme bilinéaire symétrique de $\mathbb{R}^n \times \mathbb{R}^n$ dans $\mathbb{R}$.

1. Montrer que l'on peut trouver une matrice symétrique A d'ordre n telle que :
$$\forall u, v \in \mathbb{R}^n \qquad a(u,v) = \langle Au, v\rangle_{\mathbb{R}^n}.$$
2. Calculer le gradient et la dérivée seconde (le hessien) de la fonctionnelle J définie sur $\mathbb{R}^n$ par :
$$J(v) = \frac{1}{2}\langle Av, v\rangle - \langle b, v\rangle,$$
où $b \in \mathbb{R}^n$ est fixé.
3. A quelle condition sur A, la fonction J est-elle convexe ? strictement convexe ?

Exercice 3.4.4. En partant du point $(x_0, y_0) = (-\frac{1}{2}, -\frac{1}{4})$, appliquer $n = 5$ itérations de la méthode du gradient conjugué puis la méthode de Newton, pour déterminer le minimum de la fonction f définie de $\mathbb{R}^2$ par :
$$f(x,y) = e^{x+y} + x^2 + 2y^2.$$

Exercice 3.4.5. Soient $A \in \mathcal{M}_n(\mathbb{R}), b \in \mathbb{R}^n$, et f la fonction de $\mathbb{R}^n$ dans $\mathbb{R}$ définie par
$$f(x) = \frac{1}{2}\langle Ax, x\rangle_{\mathbb{R}^n} - \langle b, x\rangle_{\mathbb{R}^n}.$$

1. Montrer que $f \in \mathcal{C}^\infty(\mathbb{R}^n, \mathbb{R})$ et calculer le gradient et la matrice hessienne de f en tout point.
2. Montrer que si A est symétrique définie positive alors il existe un unique $x^* \in \mathbb{R}^n$ qui minimise f, et que ce x^* est l'unique solution du système linéaire $Ax = b$.

Exercice 3.4.6. Soit $f \in \mathcal{C}^1(\mathbb{R}^n, \mathbb{R}), (n \geq 1)$. On suppose que f est α-elliptique et lipschitzienne.

1. Montrer que
$$f(y) - f(x) \geq \langle \nabla f(x), y - x\rangle_{\mathbb{R}^n} + \frac{\alpha}{2}\|y - x\|^2, \ \forall x, y \in \mathbb{R}^n.$$
2. Montrer que f est strictement convexe et que $f(x) \to \infty$ quand $|x| \to \infty$.
3. En déduire qu'il existe un et un seul $x^* \in \mathbb{R}^n$ tel que $f(x^*) \leq f(x)$ pour tout $x \in \mathbb{R}^n$.
4. Soient $\rho \in \left]0, \frac{2\alpha}{M^2}\right[$ et $x_0 \in \mathbb{R}^n$. Montrer que la suite $(x_n)_{n\in\mathbb{N}}$ définie par
$$x_{n+1} = x_n + \rho\nabla f(x_n), \ n \in \mathbb{N},$$
converge vers x^*.

Exercice 3.4.7. On veut résoudre le système suivant par la méthode de gradient à pas optimal :

$$\begin{cases} \frac{1}{2}x = 0 \\ \\ \frac{c}{2}y = 0 \end{cases} \quad \text{où } \ c \geq 1.$$

1. Ecrire le système sous la forme $Ax = b$ et calculer les valeurs propres de A.
2. Soit r le résidu : $b - Ax$. Calculer r et le paramètre α correspondant à la minimisation sur $\mathbb{R}$ de la fonction qui à α associe $J(x_k + \alpha r_k)$.
3. Soit P_k le point de coordonnées x_k et y_k. Exprimer x_{k+1} et y_{k+1} en fonction de x_k et y_k.
4. Soit $t_k = \dfrac{y_k}{x_k}$ la pente de la droite (OP_k) avec $O(0,0)$. Exprimer t_{k+2} en fonction de t_k. Interpréter géométriquement le résultat puis conclure.
5. Soit $t \in \{t_k, t_{k+1})$ (k donné). On appelle τ le facteur moyen de réduction de l'erreur :

$$\tau^2 = \frac{y_{k+2}}{y_k} = \frac{x_{k+2}}{x_k}.$$

   Montrer que :

$$\tau^2 = \left[\frac{c-1}{c+1}\right] \frac{1}{1 + \dfrac{c}{(c+1)^2}\left(ct - \dfrac{1}{ct}\right)^2}.$$

6. Pour quelle valeur de t, τ est-il maximum ?

Exercice 3.4.8. Soit $J : \mathbb{R}^n \to \mathbb{R}$, une fonctionnelle que l'on suppose de classe $\mathcal{C}^1$.

1. Montrer que si J est α-elliptique, alors

$$J(v) - J(u) \geq \langle \nabla J(u), v - u \rangle_{\mathbb{R}^n} + \frac{\alpha}{2} \|v - u\|^2_{\mathbb{R}^n}, \ \forall (u, v) \in (\mathbb{R}^n)^2.$$

   En déduire que J est strictement convexe et coercive.

2. On suppose que J est deux fois dérivable en tous points de $\mathbb{R}^n$. Montrer que J est α-elliptique si, et seulement si,

$$\left\langle D^2 J(u) w, w \right\rangle_{\mathbb{R}^n} \geq \alpha \|w\|^2_{\mathbb{R}^n}, \ \forall (u, w) \in (\mathbb{R}^n)^2.$$

Exercice 3.4.9. On veut minimiser la fonctionnelle $J(x) = \frac{1}{2}\langle Ax, x \rangle - \langle b, x \rangle_{\mathbb{R}^n}$, où $b \in \mathbb{R}^n$ et A est une matrice réelle symétrique définie positive. Cela revient à résoudre le système $Ax = b$. Pour cela, nous allons employer la méthode du gradient à pas constant. Soit x^* la solution de ce système. On propose l'algorithme suivant : on se donne un vecteur initial $x^{(0)}$, et on calcule la suite d'itérés par l'algorithme

$$\begin{cases} r^{(k)} & = b - Ax^{(k)} \\ x^{(k+1)} & = x^{(k)} + \alpha r^{(k)} \end{cases}$$

où α est une constante réelle donnée.

1. Soit $e^{(k)} = x^{(k)} - x^*$, pour $k \geq 0$. Donner une relation entre $e^{(k)}$ et $e^{(0)}$.

2. Montrer que l'algorithme converge si et seulement si $0 < \alpha < \frac{2}{\lambda_n}$ où λ_n est la plus grande valeur propre de A.
3. Donner le meilleur choix pour α en fonction des valeurs propres de A. Que concluez-vous ?

Exercice 3.4.10. On considère la fonction f définie sur $\mathbb{R}^n$ par

$$f(x, y) = x^4 + y^4 - 2(x - y)^2.$$

1. Montrer qu'il existe $(\alpha, \beta) \in \mathbb{R}^2_+$ (et les déterminer) tels que

   $$f(x, y) \geq \alpha \|(x, y)\|^2 + \beta,$$

   pour tous $(x, y) \in \mathbb{R}^2$, où la notation $\| \cdot \|$ désigne la norme euclidienne de $\mathbb{R}^2$.
2. En déduire que le problème

   $$\inf_{(x,y)\in\mathbb{R}^2} f(x, y) \qquad (\mathcal{P})$$

   admet au moins une solution.
3. La fonction f est-elle convexe sur $\mathbb{R}^2$?
4. Déterminer les points critiques de f, et préciser leur nature (minimum local, maximum local, point-selle, ...).
5. Résoudre le problème $(\mathcal{P})$.