# CALCUL SCIENTIFIQUE

## Polycopié de cours
## Version 2.0 (2024)

Antoine RALLU

UNIVERSITÉ DE LYON
ENTPE
L'école de l'aménagement durable des territoires
LTDS Tribologie et Dynamique des Systèmes
LABEX CELYA
UNIVERSITÉ DE LYON

# Avant-propos

La première version de ce polycopié de cours était le résultat d'une concaténation de mes notes cours de Calcul Scientifique. Cette "numérisation avancée" avait été très fortement accélérée lors du printemps 2020, période pendant laquelle l'enseignement (en particulier du cours de Calcul Scientifique) avait été totalement chamboulé par la crise du COVID : les étudiants de première année n'avaient suivi en présentiel que le premier chapitre sur les EDO, le reste (et donc la grande majorité du cours) ayant été enseigné en visioconférence, ce qui a été perturbant à la fois pour les étudiants et pour les enseignants. Par conséquent, nous avons dû nous adapter aux conditions du confinement pour assurer au maximum la continuité pédagogique.

A la rentrée 2023 a eu lieu une réforme des enseignements à l'ENTPE. Concernant les cours de maths, ce qui a principalement changé est que la réduction de 38h à 30 h pour le cours d'Analyse a entraîné une modification du programme, à savoir : (i) les distributions ne sont plus enseignées en tronc commun (mais un cours électif est ouvert au second semestre), (ii) Nicolas Vigneaud (professeur principal du cours d'Analyse) et moi avons décidé de transférer le début du chapitre deux de Calcul scientifique dédié à l'analyse de problèmes elliptiques dans le cours d'Analyse. C'est la raison pour laquelle cette nouvelle version du polycopié de Calcul Scientifique a vu le jour.

Le contenu qui n'est plus enseigné à partir de cette année est toujours disponible dans le polycopié. Par rapport à la version précédente, les modifications majeurs sont les suivantes :

1. Ajout d'un chapitre dédié à la résolution d'EDP non-stationnaires, rédigé autour du problème de la châleur unidimensionnel. L'intérêt est à la fois d'étendre la méthodologie vue pour les problèmes elliptiques, et de faire le lien avec le chapitre sur les EDO.
2. Ajout de la notion de convergence (consistance et stabilité) des schémas aux différences finies en elliptique.
3. Ajout d'un chapitre final consacré à la concaténation de tous les rappels et compléments d'algèbre linéaire précédemment distillés tout le long du précédent polycopié. J'y ai également rajouté des rappels d'analyse de fonctions réelles, ainsi qu'un formulaire (trigonométrie, dérivation, primitivation, développements limités).

L'idée est de se recentrer sur les questions de schémas numérique, et de laisser aux éudiants la lecture des parties liées aux théorèmes d'existence et d'unicité des solutions des différents problèmes.

Le contenu du cours est dans la continuité de celui proposé par mon prédécesseur Jean-Marc Malasoma, qui m'a appris à découvrir l'Histoire passionnante des mathématiques, des babylonniens jusqu'à aujourd'hui. Tant que faire se peut, des parenthèses historiques seront donc apportées au document, qui je l'espère permettront aux lecteurs de prendre du recul sur les notions et de voir les mathématiques sous un autre angle que celui purement scolaire.

Je tiens à saluer et remercier l'ensemble de l'équipe enseignante de l'année scolaire 2023-2024 : Simon Charlemagne, Anne-Sophie Colas, Emmanuel Gourdon, Maxime Morell, Nicolas Vigneaud et Aniss Ziad. Un merci particulier à Nicolas, qui apporte toujours un regard mathématique rigoureux à mes propositions.

Cette version sera améliorée et enrichie au fil des années. Pour toutes les remarques, n'hésitez pas à me contacter à l'adresse antoine.rallu@entpe.fr.

Vaulx-en-Velin, le 9 février 2026.
Antoine RALLU

# Introduction

## 0.1 Contexte

A quoi fait référence le "calcul scientifique" ? Ce terme, qui sonne comme un pléonasme, prend tout son sens dès lors qu'on comprend son interdisciplinarité basée à la fois sur les mathématiques (analyse, algèbre linéaire, analyse numérique,...) et l'informatique. Lorsqu'un ingénieur ou un chercheur doit réaliser un dimensionnement, une estimation, une optimisation, un contrôle, (...) il peut effectuer une simulation numérique des phénomènes physiques (au sens large) en jeu. Il doit en premier lieu modéliser (en fonction de la précision attendue du résultat) le problème par une approche phénoménologique qui, on le verra, débouche généralement sur l'un et/ou l'autre de ces deux types de résolution :

- un système différentiel (traduisant les variations des variables principales du problème) couplé à des conditions aux limites et/ou initiales du problème : propagation d'ondes, diffusion thermique, évolution démographique, cinétique chimique,...
- un système algébrique ou transcendant à résoudre : modes de vibration d'un bâtiment en mécanique, équation d'état de Dieterici en thermodynamique, détermination du tirant d'eau en hydrostatique,...

A noter que les ces problèmes peuvent être linéaires ou non-linéaires, c'est-à-dire que ceux-ci dépendent de façon linéaire ou non-linéaire des variables à déterminer. Ces non-linéarités peuvent être géométriques (ex : flambement d'une structure), matérielles (ex : loi de comportement élasto-plastique) ou phénoménologiques (ex : frottements). Même s'il existe des exceptions remarquables, la plupart des problèmes non linéaires ne possèdent pas de solution analytique, d'où la nécessité de mettre en œuvre des méthodes d'approximations numériques.

Cependant, avant de foncer tête baissée dans la résolution du problème, il est nécessaire de réaliser l'étude théorique du problème afin de s'assurer qu'il admet au moins une solution, voire d'une unique solution. Vient ensuite l'analyse numérique de la méthode de résolution proposée, qui dépend des propriétés théoriques du problème : c'est à cette étape qu'on peut notamment estimer la stabilité, la vitesse de convergence de la méthode et une estimation de l'écart à la solution théorique à laquelle on peut s'attendre.

On peut alors passer à la résolution informatique du problème avec l'outil adapté à nos besoins, voir paragraphe 0.3. Cette étape est non sans difficulté, car même si aujourd'hui les ordinateurs ont des capacités de calcul (mémoire vive, processeur,...) très confortables, les simulations numériques peuvent vite être très lourdes d'où la nécessité de simplifier au maximum la modélisation en amont (par exemple en prenant en compte les symétries du système) et d'adapter la méthode de résolution au problème. Il peut être tout de même nécessaire de recourir à des techniques avancées d'architecture de systèmes informatiques, comme le parallélisme. Mais ceci est en dehors du cadre de ce cours.

Vous l'aurez compris, le "calcul scientifique" ne consiste donc pas uniquement à effectuer un calcul, mais à prendre en compte l'ensemble des étapes évoquées ci-dessus.

## 0.2 Le calcul scientifique dans le cursus de l'ENTPE

Ce cours du second semestre de 1A prend pour bases :
- les notions d'analyse et d'algèbre linéaire vues en classes préparatoire
- le cours d'analyse du premier semestre
- le cours d'informatique et le tutorat de Maxima

Il sert à résoudre les problèmes des cours de mécanique au sens large :

**1A :** Mécanique des milieux continus, Résistance des matériaux, Mécanique des sols, Acoustique, Energétique

**2A :** Calcul des structures, démarche expérimentale en GC, Dynamique des structures, Mécanique des sols appliquée

**3A :** Analyse Limite et Calcul à la Rupture, Comportement des matériaux

Mais pas que des problèmes de mécanique :

**Sciences humaines :** Economie

**Sciences et vie de la Terre :** (2A) Modélisation en hydrogéologie ; (3A) Hydraulique ? Hydrologie ?, Master mécanique des fluides

**Transports :** Modélisation des transports et de leurs usages

Il se voit naturellement complété par les cours suivants :
- MRO (Méthodes de recherche opérationnelle)
- Méthodes numériques, Eléments finis appliqués aux structures et aux ouvrages géotechniques, Travaux souterrains en sols et roches, Dynamique non linéaire, Mécanique des milieux poreux, Génie parasismique

Il s'appliquera notamment dans les différents projets du cursus ingénieur : projet de modélisation, PIC, projet pont, ...et peut-être le PAST ?

## 0.3 Choix de l'outil de résolution

Ce paragraphe n'a pas pour objectif de présenter une liste exhaustive de langages/logiciels de programmation, mais d'en présenter quelques uns utiles au cours.

### 0.3.1 Langages de programmation

Aujourd'hui de nombreuses méthodes sont déjà pré-implémentées ou disponibles dans la communauté des utilisateurs dans les langages informatiques

**open-source** — l'historique et performant Fortran
- le très populaire et efficace Python
- le très prometteur et unificateur Julia

**payant** : la plus répandue plateforme de calcul numérique Matlab

En dehors d'une habitude historique à l'utilisation de Matlab, il n'existe donc plus de raison valable de se tourner vers cette solution payante de nos jours !

A noter l'existence de logiciels de calcul formel, permettant en autres du calcul analytique : sans constest, les deux plus performants (mais payants) sont :
- Mathematica, qui possède notamment avec une version (restreinte) gratuite en ligne wolfram alpha
- Maple

Une alternative open-source est le logiciel Maxima, dont un cours/tutoriel de 4h vous est proposé en 1A.

### 0.3.2 Logiciels de résolution d'EDP

La résolution numérique d'EDP n'est pas au cœur du cours de Calcul scientifique, mais est son prolongement naturel. Dans ce cours seule la résolution d'EDP elliptiques (par exemple l'équation de la châleur en régime permanent) par la méthode des différences finies est abordée dans le chapitre 2. Il est tout à fait possible d'implémenter nous-même cette méthode (tout comme les méthodes des éléments finis, des éléments de frontière, ...) dans un des langages vus au paragraphe précédent. A ce titre citons entre autres les bibliothèques Python dédiées aux éléments finis FENICS ou aux élément de frontière bempp. Cependant, dans des cas compliqués il peut être utile d'utiliser des codes dédiés à chacune de ces méthodes. Voir par exemple :

**Eléments finis :** FreeFem++ (Gratuit) ou COMSOL (Simulations multiphysiques)

**Différences finies :** FLAC3D (dédié à la géotechnique)

# Table des matières

# Chapitre 1

# Schémas numériques pour l'approximation des solutions des équations différentielles ordinaires

*Ce premier chapitre est consacré à l'étude et la résolution d'équations différentielles ordinaires (EDO). La résolution d'EDO linaires à coefficients constants/variables a été traitée en classes préparatoires (voir par exemple [Gourdon, 2008]) et n'est pas abordée ici.*

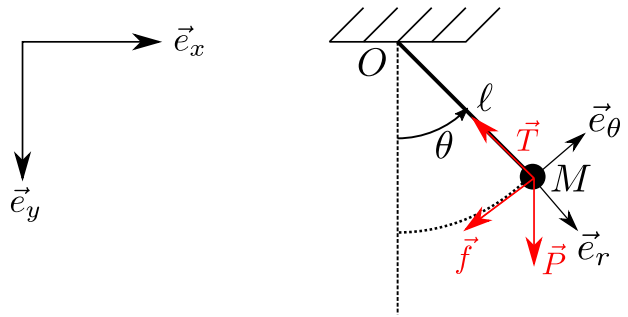
*Le chapitre commence par des exemples de modélisations de phénomènes multidisciplinaires menant à l'écriture de problèmes différentiels, i.e d'équations différentielles ordinaires accompagnées de conditions initiales. De plus, dans la grande majorité des cas, ces EDO sont non linéaires, et hormis quelques rares exceptions les EDO non linéaires n'admettent pas de solution analytique. Arrivent alors (au moins) trois questions : existe-t-il au moins une solution au problème différentiel ? Avec un peu de chance, une unique solution ? Et . . .comment en obtenir une approximation ?*

**Sommaire**

## 1.1 Exemples introductifs

### 1.1.1 Pendule pesant

Considérons le problème du pendule simple encastré au point $O$, sous les hypothèses : (i) fil de longueur $\ell$ rigide et sans masse, (ii) masse $(m)$ infiniment rigide et ponctuelle (au point $M$) et (iii) le pendule oscille dans le plan fixe $(\vec{e}_x, \vec{e}_y)$, voir figure 1.1 : où $\vec{P}$,$\vec{T}$ et $\vec{f}$ représentent respectivement le poids, la tension du fil




$$\vec{f} = -c\,\dot{\theta}\,\vec{e}_\theta$$
$$\vec{P} = mg\ (\cos(\theta)\vec{e}_r - \sin(\theta)\vec{e}_\theta)$$
$$\vec{T} = -\|T\|\,\vec{e}_r$$

FIGURE 1.1 – Pendule pesant

et le frottement de l'air, supposé orthoradial, opposé au sens du mouvement et proportionnel à $\dot{\theta} = \frac{d\theta}{dt}$. Le repère polaire tournant $(\vec{e}_r, \vec{e}_\theta)$ s'exprime dans le repère fixe :

$$\begin{cases} \vec{e}_r = \sin(\theta)\vec{e}_x + \cos(\theta)\vec{e}_y \\ \vec{e}_\theta = \cos(\theta)\vec{e}_x - \sin(\theta)\vec{e}_y \end{cases} \Rightarrow \begin{cases} \dfrac{d\vec{e}_r}{dt} = \dot{\theta}\vec{e}_\theta \\ \dfrac{d\vec{e}_\theta}{dt} = -\dot{\theta}\vec{e}_r \end{cases}$$

Le principe fondamental de la dynamique appliqué à la masse $m$ s'écrit :

$$m\,\frac{d^2\overrightarrow{OM}}{dt^2} = \vec{P} + \vec{T} + \vec{f}$$

Rappelons que dans le cas général où $\overrightarrow{OM} = r(t)\vec{e}_r$, ses dérivées successives par rapport au temps s'écrivent :

$$\frac{d\overrightarrow{OM}}{dt} = \dot{r}\,\vec{e}_r + r\dot{\theta}\vec{e}_\theta \qquad ; \qquad \frac{d^2\overrightarrow{OM}}{dt^2} = \left(\ddot{r} - r\,\dot{\theta}^2\right)\,\vec{e}_r + \left(r\,\ddot{\theta} + 2\,\dot{r}\,\dot{\theta}\right)\,\vec{e}_\theta$$

Ainsi, dans le cas $r(t) = \ell$ constant, et en projetant le PFD sur le repère polaire :

$$\begin{cases} -m\,\ell\,\dot{\theta}^2 = -\|T\| + m\,g\,\cos(\theta) \\ m\,\ell\,\ddot{\theta} = -c\,\dot{\theta} - m\,g\,\sin(\theta) \end{cases}$$

En posant $\omega_0 = \sqrt{\frac{g}{\ell}}$ et $\xi = \frac{c}{2m\sqrt{g\ell}}$ représentant respectivement la pulsation propre l'amortissement spécifique du système, l'équation pilotant le mouvement de la masse s'écrit :

$$\ddot{\theta} + 2\,\xi\,\omega_0\,\dot{\theta} + \omega_0^2\,\sin(\theta) = 0 \tag{1.1}$$

L'EDO (1.1) est non linéaire à cause du terme inertiel en $\sin(\theta)$. Dans le cas où il n'y a pas de frottement et pour de petits déplacement, i.e $\theta \ll 1$, l'approximation $\sin(\theta) \approx \theta$ est valide et l'EDO non linéaire (1.1) devient linéaire, on retrouve l'équation classique de l'oscillateur $\ddot{\theta} + \omega_0^2\,\theta = 0$.

### 1.1.2 Bruxellateur irréversible

Prigogine, Lefever et Nicolis ont proposé en 1967 un modèle à deux variables $X$ et $Y$ qui présente des oscillations périodiques asymptotiquement stables. Soit le modèle constitué par le schéma réactionnel suivant :

$$\begin{cases} A & \overset{k_1}{\to} X \\ B + X & \overset{k_2}{\to} Y + D \\ 2X + Y & \overset{k_3}{\rightleftharpoons} 3X \qquad \text{(autocatalyse)} \\ X & \overset{k_4}{\to} E \end{cases} \tag{1.2}$$

où les lettres $A, B, D, E$ désignent des espèces dont la concentration n'évolue pas au cours du temps, ce qui traduit que la réaction a lieu dans un réacteur ouvert. Pour alléger les notations, la concentration d'une espèce sera notée de la même manière que l'espèce elle-même.

Rappelons que dans le cas général d'une réaction $\alpha\, A + \beta\, B \overset{k}{\to} \gamma\, C$, les expressions des différentes vitesses sont :

- vitesse d'apparition de l'espèce $C$ : $\dot{C} = \gamma\, k\, A^\alpha\, B^\beta$
- vitesse de disparition de $A$ : $\dot{A} = -\alpha\, k\, A^\alpha\, B^\beta$
- vitesse de la réaction : $v = \frac{\dot{C}}{\gamma} = -\frac{\dot{A}}{\alpha} = -\frac{\dot{B}}{\beta}$

Ainsi, les deux équations cinétiques du modèle (1.2) représentant l'évolution des concentrations de $X$ et $Y$ en fonction du temps sont :

$$\begin{cases} \dot{X} = k_1\, A - (k_4 + k_2\, B)\; X + k_3\, X^2\, Y \\ \dot{Y} = k_2\, B\, X - k_3\, X^2\, Y \end{cases} \tag{1.3}$$

Le système (1.3) est un système de deux EDO non linéaires (à cause de l'autocatalyse) couplées. Ce modèle permet de démontrer que certaines réactions chimiques peuvent être à l'origine d'une auto-réorganisation du système. Son étude met en évidence les notions d'ondes chimiques et de cycle limite pour le système.

### 1.1.3 Démographie

L'exemple qui suit provient intégralement du cours [Popier et Winterberger, 2006]. Il a pour avantage de sortir des domaines classiques de la physique pour aborder la biologie (dynamique des populations). Les modèles démographiques les plus simples concernent une population isolée. Pour tout instant $t$, on note $N(t)$ l'effectif d'une population à cet instant, on suppose que cet effectif est assez important pour supposer $N(t)$ réel (et non pas entier). L'évolution de cette population est décrite par une équation différentielle :

$$\frac{dN(t)}{dt} = naissances - décès + migrations$$

S'il n'y a pas de migrations, et si les naissances et les décès sont proportionnelles à la taille de la population, on obtient une équation linéaire :

$$\frac{dN(t)}{dt} = a\, N(t) - b\, N(t)$$

Dans cette approche (modèle de Malthus (1766-1834)), l'effectif de la population croît ou décroît exponentiellement vite, ce qui n'est pas réaliste, surtout à long terme. On peut alors ajouter une correction lorsque la population grossit : il existe une taille idéale, dite capacité biotique. En dessous de cette capacité, la population augmente, au dessus elle diminue. Un modèle possible, dû à Verhulst (1804-1849) et dit logistique, est le suivant :

$$\frac{dN(t)}{dt} = r\, N(t) \left(1 - \frac{N(t)}{K}\right)$$

où $K$ est la capacité biotique et les solutions de cette équation sont :

$$N(t) = \frac{N(0)\, K\, e^{rt}}{K + N(0)\, (e^{rt} - 1)}$$

Cette équation admet deux points d'équilibre 0 et $K$. En effet si $N(0) = 0$, l'effectif de la population reste nulle au cours du temps. Si $N(0) = K$, la taille de la population reste stable, égale à $K$. De plus, si $N(0) > 0$, alors $\lim_{t\to\infty} N(t) = K$. Ceci signifie que $K$ est un point d'équilibre stable, tandis que 0 est instable : si $N(0)$ est proche de $K$, la solution reste proche de $K$ ; si $N(0)$ est proche de 0, elle ne reste pas près de zéro.

On peut aussi faire entrer en jeu l'interaction avec d'autres populations. On va se limiter à deux populations avec une interaction proie-prédateur. Le premier modèle est du à Volterra (1860-1940) et est appelé modèle de Lotka-Volterra, car il a été introduit presque simultanément par Lotka comme représentation d'un système chimique exhibant un caractère oscillant.
Ce modèle est basé sur les hypothèses suivantes :

- sans prédateurs, la population des proies croît exponentiellement vite (dynamique de Malthus) ;
- sans proies, le taux de décès parmi les prédateurs est proportionnel à la taille de la population ;
- le taux de disparition des proies est proportionnel au nombre de rencontres entre une proie et un prédateur, supposé lui-même proportionnel au produit des deux effectifs ;
- la taux de croissance des prédateurs est aussi proportionnel au nombre de rencontres entre proie et prédateur.

Si $N$ est l'effectif des proies et $P$ celui des prédateurs, on obtient le système d'équations :

$$\begin{cases} \dfrac{dN(t)}{dt} = N(t)\ (a - b\,P(t)) \\ \dfrac{dP(t)}{dt} = P(t)\ (c\,N(t) - d) \end{cases} \quad ; \qquad (a,b,c,d) \in \left(\mathbb{R}_+^\star\right)^4 \tag{1.4}$$

Le système dynamique (1.4) a des propriétés très particulières. Il a deux points d'équilibre $(0\,;\ 0)$ et $(d = c; a = b)$, et dans le plan de phase, i.e. dans le plan dont les coordonnées sont $N$ et $P$, les trajectoires d'un mobile données par le système, sont fermées, donc $N$ et $P$ sont périodiques. Ce modèle a lui aussi ces limites. En effet en l'absence de prédateurs, il n'est pas réaliste de penser que la population des proies va croître indéfiniment.

À travers ces exemples, nous avons vu différents cas : équations linéaires, non linéaires, systèmes d'équations, sans savoir pourquoi ces équations avaient des solutions, ni, le cas échéant, si elles admettaient une unique solution.

## 1.2 Résultats d'existence et d'unicité

### 1.2.1 Contexte

**Proposition 1.1.** *Soient $U$ un ouvert de $\mathbb{R}\times\mathbb{R}^m$, $m \in \mathbb{N}^\star$ et $f : U \to \mathbb{R}^m$ une application continue. On considère l'équation différentielle ordinaire*

$$(E) \qquad y' = f(t,y),\ (t,y) \in U,\ t \in \mathbb{R},\ y \in \mathbb{R}^m$$

**Remarque 1.1.**
Toute EDO peut s'écrire sous la forme (E), quel que soit son ordre. En effet, une EDO scalaire d'ordre $p$ peut s'écrire comme un système de $p$ EDO du premier ordre de manière équivalente. Si on reprend l'exemple du pendule (1.1) :

$$\ddot{\theta} + 2\,\xi\,\omega_0\,\dot{\theta} + \omega_0^2\,\sin(\theta) = 0 \Leftrightarrow \begin{cases} \dot{\theta} = \Omega \\ \dot{\Omega} = -2\,\xi\,\omega_0\,\Omega + \omega_0^2\,\sin(\theta) \end{cases}$$

### 1.2.2 Rappels - Définitions

**Problème de Cauchy**

**Définition 1.1.**
Une solution de $(E)$ sur un intervalle $I \subset \mathbb{R}$ est une fonction dérivable $y : I \to \mathbb{R}^m$ telle que :
i) $\forall t \in I$, $(t, y(t)) \in U$
ii) $\forall t \in I$, $y'(t) = f(t, y(t))$

**Définition 1.2.**
Soient $f : U \to \mathbb{R}^m$ et $(t_0, y_0) \in U$. Le problème de Cauchy consiste à trouver une solution $y : I \subset \mathbb{R} \to \mathbb{R}^m$ de $(E)$ sur un intervalle $I$ comprenant $t_0$ et telle que $y(t_0) = y_0$.

**Solution maximale**

**Définition 1.3.**
Soient $y : I \to \mathbb{R}^m$ et $\tilde{y} : \tilde{I} \to \mathbb{R}^m$ des solutions de $(E)$. On dit que $\tilde{y}$ est un prolongement de $y$ si $I \subset \tilde{I}$ et $\tilde{y}|_I = y$.

**Définition 1.4.**
On dit que $y : I \to \mathbb{R}^m$ solution de $(E)$ est maximale si $y$ n'admet pas de prolongement $\tilde{y} : \tilde{I} \to \mathbb{R}^m$, avec $I \subset \tilde{I}$.

**Théorème 1.1.**
*Toute solution $y$ se prolonge en une solution maximale $\tilde{y}$ (pas nécessairement unique).*

**Solution globale**

On suppose que l'ouvert $U$ est de la forme $I \times U'$ où $J \subset \mathbb{R}$ et $U'$ ouvert de $\mathbb{R}^m$.

**Définition 1.5.**
Une solution globale est une solution définie sur l'intervalle $I$ tout entier.

**Remarque 1.2.**
Toute solution globale est maximale, mais la réciproque est fausse.

**Exemple 1.1.** Soit l'EDO $(E) \quad y' = y^2$ sur $U = \mathbb{R} \times \mathbb{R}$.
— solution globale $\forall t \in \mathbb{R}$, $y(t) = 0$
— si $y$ ne s'annule pas, (E) peut se réécrire $\frac{y'}{y^2} = 1$ d'où par intégration (avec $C \in \mathbb{R}$)

$$-\frac{1}{y(t)} = t - C \Rightarrow y(t) = -\frac{1}{t - C}$$

Ainsi il existe deux solutions définies sur $]-\infty, C[$ et $]C, +\infty[$, qui sont maximales mais pas globales.

**Régularité des solutions**

**Théorème 1.2.**
*Si $f : U \to \mathbb{R}^m$ est de classe $\mathcal{C}^k$, toute solution de $(E)$ est de classe $\mathcal{C}^{k+1}$.*

### 1.2.3 Théorèmes d'existence

**Proposition 1.2.** *Une fonction $y : I \to \mathbb{R}^m$, $I \subset \mathbb{R}$ est une solution du problème de Cauchy de données initiales $(t_0, y_0)$ si et seulement si :*

*i) $y$ est continue et $\forall t \in I$, $(t, y(t)) \in U$*

*ii) et $\forall t \in I$,*

$$y(t) = y_0 + \int_{t_0}^{t} f(s, y(s))\, ds \tag{1.5}$$

Pour résoudre $(E)$, on va plutôt chercher à résoudre (1.5). On va d'abord montrer qu'une solution passant par $(t_0, y_0) \in U$ ne peut pas s'éloigner trop vite de $y_0$. On note $\|.\|$ une norme (quelconque) sur $\mathbb{R}^m$ et $\mathcal{B}(x, r)$ (respectivement $\bar{\mathcal{B}}(x, r)$) la boule ouverte (respectivement fermée) de centre $x$ et de rayon $r$. Comme $U$ est ouvert, il existe un cylindre

$$C_0 = [t_0 - T_0, t_0 + T_0] \times \bar{\mathcal{B}}(y_0, r_0) \subset U$$

$C_0$ est fermé et borné dans $\mathbb{R}^{m+1}$, donc compact. Par conséquent $f$ est bornée sur $C_0$, c'est-à-dire :

$$\exists M > 0 / \; M = \sup_{(t,y) \in C_0} \|f(t, y)\| < +\infty$$

Enfin on désigne par $C = [t_0 - T, t_0 + T] \times \bar{\mathcal{B}}(y_0, r_0) \subset C_0$ un cylindre de demi-longueur $T \leq T_0$.

**Définition 1.6.**
On dit que $C$ est un cylindre de sécurité pour l'équation $(E)$ si toute solution $y : I \to \mathbb{R}^m$ du problème de Cauchy $y(t_0) = y_0$ avec $I \subset [t_0 - T; t_0 + T]$, reste contenue dans $\bar{\mathcal{B}}(y_0, r_0)$.

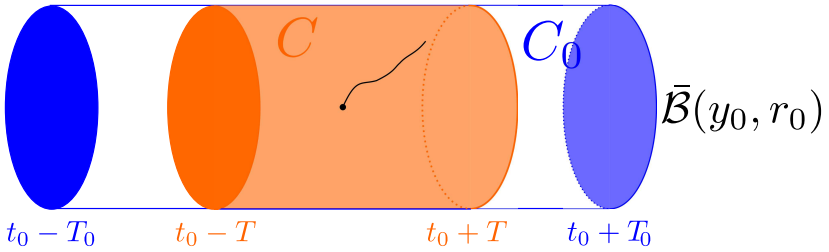



FIGURE 1.2 – Cylindres de sécurité $C_0$ et $C$, et une solution.

**Proposition 1.3.** *Pour que $C$ soit un cylindre de sécurité, il suffit de prendre :*

$$T \leq \min\left(T_0, \frac{r_0}{M}\right)$$

**Théorème 1.3 (de Cauchy-Peano-Arzela).**
*Soit $C$ un cylindre de sécurité pour $(E)$. Alors il existe une solution $y : [t_0 - T, t_0 + T] \to \bar{\mathcal{B}}(y_0, r_0)$ de (E) avec condition initiale $(t_0, y_0)$.*

**Remarque 1.3.**
Le théorème 1.3 prouve l'existence d'une solution locale de (E) sur $[t_0 - T, t_0 + T]$. Son prolongement en une solution maximale est assuré par le théorème 1.1.

**Théorème 1.4 (Condition suffisante d'existence de solution globale).**
*Soit $f : U \to \mathbb{R}^m$ une application continue sur un ouvert de la forme $U = I \times \mathbb{R}^m$, où $I \subset \mathbb{R}$ ouvert. On suppose qu'il existe une fonction continue $k : I \to \mathbb{R}^+$ telle que pour tout $t \in I$, l'application $y \mapsto f(t, y)$ est lipschitzienne de rapport $k(t)$ sur $\mathbb{R}^m$. Alors une solution maximale de $(E)$ est globale, i.e définie sur $I$ tout entier.*

### 1.2.4 Théorème d'existence et d'unicité de Cauchy-Lipschitz

On suppose de plus que $f$ est localement lipschitzienne pour son deuxième argument, i.e pour tout point $(t_0, y_0) \in U$ il existe une cylindre $C = [t_0 - T, t_0 + T] \times \bar{\mathcal{B}}(y_0, r_0) \subset U$ et une constante $k = k(t_0, y_0)$ tels que $f$ est $k$-lipschitzienne sur $C$ pour son deuxième argument :

$$\forall (t, y_1), (t, y_2) \in C \times C,\ \|f(t, y_1) - f(t, y_2)\| \leq k\|y_1 - y_2\|$$

**Théorème 1.5 (de Cauchy-Lipschitz).**
*Si $f : U \to \mathbb{R}^m$ est localement lipschitzienne pour son deuxième argument, alors pour tout cylindre de sécurité $C$ le problème de Cauchy avec donnée initiale $(t_0, y_0)$ admet une unique solution $y : [t_0 - T, t_0 + T] \to \bar{\mathcal{B}}(y_0, r_0)$.*

**Preuve.** *Notons $\tilde{\mathcal{C}}^0 = \mathcal{C}\left([t_0 - T, t_0 + T], \bar{\mathcal{B}}(y_0, r_0)\right)$ l'ensemble des fonctions continues de $[t_0 - T, t_0 + T]$ dans $\bar{\mathcal{B}}(y_0, r_0)$ muni de la distance $d$ de la convergence uniforme. A toute fonction $y \in \tilde{\mathcal{C}}^0$, associons la fonction $\Phi(y)$ telle que*

$$\forall t \in [t_0 - T, t_0 + T],\ \Phi(y)(t) = y_0 + \int_{t_0}^{t} f(s, y(s))\, ds$$

*D'après la proposition 1.2, $y$ est solution de $(E)$ si et seulement si $y$ est un point fixe de $\Phi$. On va donc appliquer le théorème du point fixe de Picard :*

- *l'espace $\left(\tilde{\mathcal{C}}^0, d\right)$ est un espace métrique complet (admis) ;*
- *$\Phi : \tilde{\mathcal{C}}^0 \to \tilde{\mathcal{C}}^0$. En effet,*

$$\|\Phi(y)(t) - y_0\| = \| \int_{t_0}^{t} f(s, y(s))\, ds\| \leq M\,|t - t_0| \leq MT \leq r_0$$

  *donc $\Phi(y) \in \tilde{\mathcal{C}}^0$*
- *Afin de montrer que $\Phi$ est strictement contractante, procédons en deux étapes : montrer que (i) $\Phi^p = \underbrace{\Phi \circ .. \circ \Phi}_{p \text{ fois}}$ puis (ii) $\Phi$ sont strictement contractantes.*

  *(i) Soient $(y, z) \in \tilde{\mathcal{C}}^0 \times \tilde{\mathcal{C}}^0$ et $y_p = \Phi^p(y)$, $z_p = \Phi^p(z)$. On a alors :*

$$\|y_1 - z_1\| = \| \int_{t_0}^{t} (f(s, y(s)) - f(s, z(s)))\ ds\| \leq k\,|t - t_0|\, d(y, z)$$

  *De même,*

$$\|y_2 - z_2\| \leq \int_{t_0}^{t} k\,\|y_1 - z_1\|\, dr \leq k^2\, \frac{|t - t_0|^2}{2}\, d(y, z)$$

  *Ainsi par récurrence,*

$$\|y_p - z_p\| \leq k^p\, \frac{|t - t_0|^p}{p!}\, d(y, z)$$

*En particulier,*

$$d\left(\Phi^p(y), \Phi^p(z)\right) \leq k^p \, \frac{|t-t_0|^p}{p!} \, d(y,z)$$

*Ainsi $\Phi^p$ est lispchitzienne sur $\tilde{\mathcal{C}}^0$ de rapport $k^p \, \frac{|t-t_0|^p}{p!}$ et pour p suffisamment grand, $k^p \, \frac{|t-t_0|^p}{p!} < 1$ i.e $\Phi^p : \tilde{\mathcal{C}}^0 \to \tilde{\mathcal{C}}^0$ est strictement contractante . Par application du théorème de Picard, $\Phi^p$ admet donc un unique point fixe.*

*(ii) Par conséquent, il existe $q \geq p$ tel que $\Phi^q$ est strictement contractante et admet un unique point fixe $y$ : $\Phi^q(y) = y$. Ainsi, $\Phi \circ \Phi^q(y) = \Phi(y)$, or $\Phi \circ \Phi^q(y) = \Phi^q \circ \Phi(y)$. Donc $\Phi(y)$ est point fixe de $\Phi^q$, et par unicité du point fixe de $\Phi^q$, $y = \Phi(y)$. Au final, comme les points fixe de $\Phi$ sont les points fixes de $\Phi^q$, $y$ est l'unique point fixe de $\Phi$.*

**Théorème 1.6.**
*Soient $y_1 : I \to \mathbb{R}^m$ et $y_2 : I \to \mathbb{R}^m$ deux solutions de (E), où $f$ est localement lipschizienne pour son deuxième argument. Si $y_1$ et $y_2$ coïncident en un point de $I$, alors $y_1 = y_2$ sur $I$.*

## 1.3 Schémas numériques

### 1.3.1 Principe

Soit une subdivision $t_0 < t_1 < \ldots < t_n$ de pas $h_i = t_{i+1} - t_i$. D'après la proposition 1.2, on sait que les solutions du problème de Cauchy vérifient :

$$y(t_{i+1}) = y(t_i) + \int_{t_i}^{t_{i+1}} f(s, y(s)) \, ds$$

La résolution numérique de ce problème consiste à approximer $y(t_{i+1}) \approx y_{i+1}$, $y(t_i) \approx y_i$ et le schéma de résolution définit l'approximation de l'intégrale $\int_{t_i}^{t_{i+1}} f(s, y(s)) \, ds$.

Il existe différents types de schémas :

**Définition 1.7.**
Le schéma de résolution est dit explicite si $y_{i+1}$ peut être calculé directement à partir des valeurs précédentes. Dans le cas contraire, le schéma est dit implicite.

**Définition 1.8.**
Le schéma de résolution est dit à un pas si la méthode numérique permettant de calculer $y_{i+1}$ ne dépend que de $y_i$ et $t_i$. Dans le cas contraire il est dit à pas multiple.

**Proposition 1.4.** *L'écriture générale d'une méthode explicite à un pas est la suivante :*

$$y_{i+1} = y_i + h_i \, \Phi(t_i, y_i, h_i) \tag{1.6}$$

*où $\Phi$ donne la définition du schéma.*

**Exemple 1.2 (**de schémas explicites à un pas**).**

$$\begin{cases} \text{Schéma d'Euler explicite (EE)} & \Phi(t,y,h) = f(t,y) \\ \text{Schéma d'Euler modifié} & \Phi(t,y,h) = f\left(t + \dfrac{h}{2}, y + \dfrac{h}{2} f(t,y)\right) \\ \text{Schéma de Heun} & \Phi(t,y,h) = \dfrac{1}{2}\left[f(t,y) + f\left(t+h, y + h\, f(t,y)\right)\right] \end{cases}$$

**Exemple 1.3 (**de schémas implicites à un pas**).**

$$\begin{cases} \text{Schéma d'Euler implicite (EI)} & y_{i+1} = y_i + h_i\, f(t_i + h_i, y_{i+1}) \\ \text{Schéma de Cranck-Nicholson (trapèze)} & y_{i+1} = y_i + \dfrac{h_i}{2}\,[f(t_i, y_i) + f(t_i + h_i, y_{i+1})] \end{cases}$$

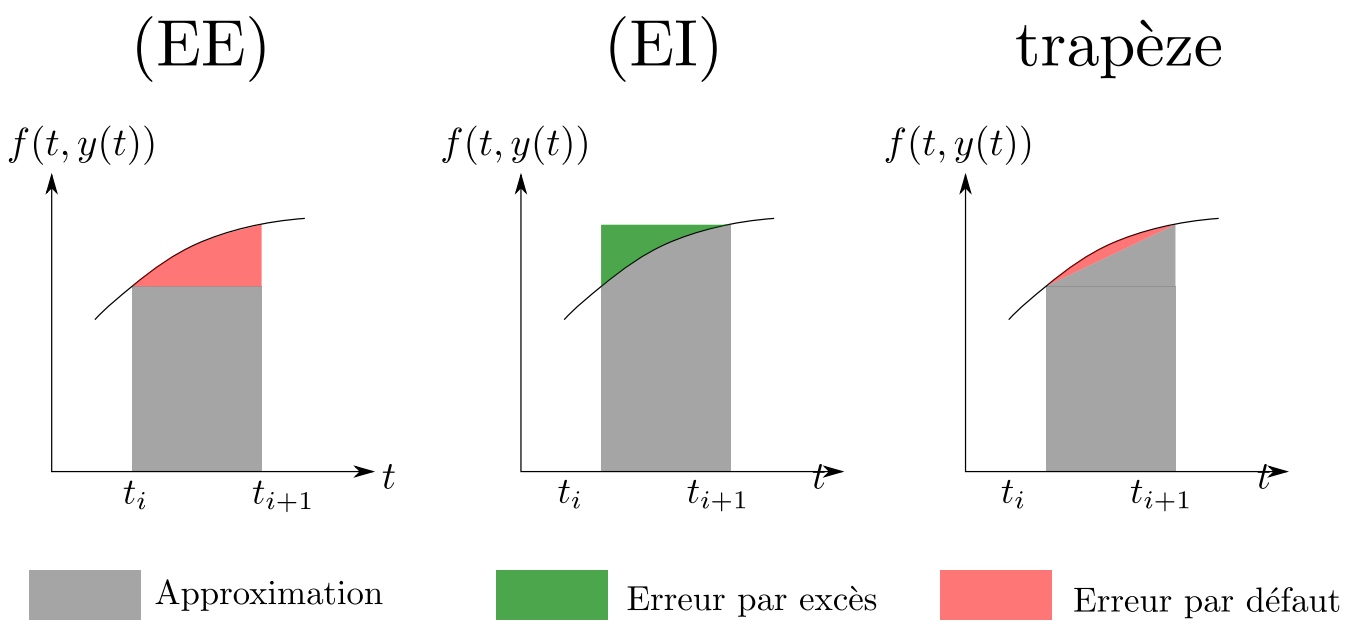



FIGURE 1.3 – Approximations de l'intégrale $\Phi$ selon les schémas (EE), (EI) et du trapèze.

Sur la Figure 1.3 sont représentées (en gris) les approximations numériques de l'intégrale

$$\int_{t_i}^{t_{i+1}} f(s, y(s))\, ds \approx h_i\, \Phi(t_i, y_i, h_i)$$

selon les schémas d'Euler explicite (EE), d'Euler implicite (EI) et de Cranck-Nicholson (trapèze). Dans l'exemple, la fonction $t \mapsto f(t, y(y))$ est croissante entre $t_i$ et $t_{i+1} = t_i + h_i$ donc le schéma (EI) mène à une erreur par excès (en vert) alors que les deux autres engendrent des erreurs par défaut (en rouge). Les aires correspondant à ces erreurs pour les schémas (EE) et (EI) sont du même ordre de grandeur, alors que celle pour la méthode du trapèze est beaucoup plus petite. Ce contraste sera quantifié au paragraphe 1.3.2, où on définira l'ordre de convergence d'un schéma numérique explicite.

**Exercice 1.1.** Soit le problème de Cauchy $(E)$ $\quad y' = y^2 + 1$, $y(0) = 0$ pour $t \in I = [0, 1]$.

1. Montrer que $(E)$ admet une unique solution locale puis résoudre analytiquement.
2. Résoudre $(E)$ grâce au schéma d'Euler Implicite. On prendra comme pas $h = 0.1$ et on fera 3 itérations (donner les résultats $10^{-5}$ près).
3. Comparer les deux approches.

### 1.3.2 Etudes de convergence des schémas explicites à un pas

**Consistance**

**Définition 1.9.**
La quantité $\varepsilon_i = y(t_{i+1}) - y(t_i) - h_i\, \Phi(t_i, y_i, h_i)$ est l'erreur de consistance à l'instant $t_i$

**Définition 1.10.**

Un schéma à un pas est dit consistant avec $(E)$ si $y$ étant une solution de cette équation, on a

$$\lim_{h\to 0}\sum_{i=0}^{n-1}\|\varepsilon_i\| = 0 \qquad \text{où } h = \max_i h_i$$

**Théorème 1.7.**

*Une condition nécessaire et suffisante pour qu'un schéma défini par $\Phi$ soit consistant avec (E) est que*

$$\forall (t,y)\in C,\ \Phi(t,y,0) = f(t,y)$$

**Stabilité**

La notion de stabilité d'un schéma numérique assure que la solution d'un problème où les données sont faiblement perturbées (perturbations désignées dans la définition suivante par $\varphi_i$) reste bornée (donc contrôlée) quand $h$ tend vers 0.

**Définition 1.11.**

Un schéma est dit stable s'il existe une constante $M > 0$ indépendante de $h$ telle que pour toutes suites $\{y_i, z_i, \varepsilon_i\}_{0\le i\le n}$ vérifiant :

$$y_{i+1} = y_i + h_i\,\Phi(t_i, y_i, h_i) \qquad ; \qquad z_{i+1} = z_i + h_i\,\Phi(t_i, z_i, h_i) + \varphi_i$$

Alors :

$$\|y_i - z_i\| \le M\|y_0 - z_0\| + \sum_{i<n}\|\varphi_i\|$$

**Théorème 1.8.**

*Une condition suffisante pour qu'un schéma soit stable est que $\Phi$ soit lipschitzienne par rapport à son deuxième argument, i.e*

$$\exists M > 0/\ \forall (t,y)\in C, \forall (t,z)\in C, \forall h\in [0,H[, \|\Phi(t,y,h) - \Phi(t,z,h)\| \le M\|y-z\|$$

*avec $M$ indépendant de $h$.*

**Convergence**

**Définition 1.12.**

Un schéma est dit convergent si, $y$ étant solution de (E), on a $\lim_{h\to 0}\|y_i - y(t_i)\| = 0$.

**Théorème 1.9 (de Lax-Richtmyer).**

*Un schéma consistant et stable est convergent.*

**Ordre de convergence**

**Définition 1.13.**
Un schéma convergent est dit d'ordre au moins $p$ si, $y$ étant solution de (E), on a

$$\frac{y(t+h)-y(t)}{h} - \Phi(t,y,h) = O(h^p) = o(h^{p-1})$$

**Théorème 1.10.**
*Un schéma convergent est dit d'ordre au moins $p$ si et seulement si $\Phi$ est tel que :*

$$\partial_3^\ell \Phi(t,y,0) = \frac{1}{\ell+1} f^{[\ell]}(t,y), \quad 0 \leq \ell \leq p-1$$

*Il est strictement d'ordre $p$ si en plus $\partial_3^p \Phi(t,y,0) \neq \frac{1}{p+1} f^{[p]}(t,y)$.*

**Rappel : règle de la chaîne**
Soit $g : \mathbb{R}^n \to \mathbb{R}$ une fonction à $n$ arguments $\{v_i\}_{1\leq i\leq n}$. La dérivée "droite" de $g$ par rapport à $t$ s'écrit :

$$\frac{d}{dt} g(v_1,\ldots,v_n) = \sum_{i=1}^{n} \partial_i g(v_1,\ldots,v_n) \times \frac{dv_i}{dt}$$

**Notations :**

$$\begin{cases} f^{[0]}(t,y) = y'(t) = f(t,y) \\ f^{[1]}(t,y) = y''(t) = \partial_1 f(t,y) + f(t,y)\,\partial_2 f(t,y) \\ f^{[2]}(t,y) = y'''(t) = \partial_{11}^2 f(t,y) + 2\,f(t,y)\,\partial_{12}^2 f(t,y) + f(t,y)^2 \partial_{22}^2 f(t,y) + (\partial_1 f(t,y) + f(t,y)\partial_2 f(t,y))\;\partial_2 f(t,y) \\ f^{[p]}(t,y) = y^{(p+1)}(t) = \partial_1 f^{[p-1]}(t,y) + f(t,y)\,\partial_2 f^{[p-1]}(t,y) \end{cases}$$

**Preuve.** *Soit le développement de Taylor à l'ordre $p$ de la fonction $y \in \mathcal{C}^p(\mathbb{R})$ :*

$$\begin{aligned} y(t+h) &= y(t) + h\,y'(t) + \ldots + \frac{h^p}{p!} y^{(p)}(t) + o(h^p) \\ &= y(t) + h\,f^{[0]}(t) + \ldots + \frac{h^p}{p!} f^{[p-1]}(t) + o(h^p) \end{aligned}$$

*Soit de plus le développement de Taylor à l'ordre $p-1$ de la fonction $h \mapsto \Phi(t,y,h)$, supposée au moins de classe $\mathcal{C}^{p-1}(\mathbb{R})$ :*

$$\Phi(t,y,h) = \Phi(t,y,0) + h\,\partial_3\Phi(t,y,0) + \ldots + \frac{h^{p-1}}{(p-1)!} \partial_3^{p-1}\Phi(t,y,0) + o(h^{p-1})$$

*Par conséquent on a :*

$$\begin{aligned} \frac{y(t+h)-y(t)}{h} - \Phi(t,y,h) = {} & (f(t,y) - \Phi(t,y,0)) + h\left(\frac{1}{2} f^{[1]}(t,y) - \partial_3\Phi(t,y,0)\right) + \ldots \\ & + \frac{h^{p-1}}{(p-1)!}\left(\frac{1}{p} f^{[p-1]}(t,y) - \partial_3^{p-1}\Phi(t,y,0)\right) + o(h^{p-1}) \end{aligned}$$

**Proposition 1.5.** *Un schéma convergent est d'ordre au moins 1 (car consistant).*

**Exemple 1.4 (**Schéma d'Euler explicite**).** Soient $y$ solution de (E), $I = [t_0, T] \subset \mathbb{R}$ et $f : I \times \mathbb{R} \to \mathbb{R}$ une fonction $k$-lipschitzienne par rapport à son deuxième argument. Montrer que le schéma d'Euler explicite est convergent. De quel ordre est ce schéma ?

1) Consistance : $\forall t \in I, \forall (y,z) \in \mathbb{R}^2, \Phi(t,y,0) = f(t,y)$
2) Stabilité : $H > 0, \forall t \in I, \forall (y,z) \in \mathbb{R}^2, \forall h \in [0, H[, \, |\Phi(t,y,h) - \Phi(t,z,h)| \leq k\,|y-z|$
3) Théorème de Lax-Richtmyer : schéma consistant et stable donc convergeant d'ordre au moins 1.
4) Ordre du schéma : $\partial_3 \Phi(t,y,h) = 0 \neq \frac{1}{2} f^{[1]}(t,y)$. Le schéma est strictement d'ordre 1.

**Exercice 1.2 (**Schéma de Heun**).** Soient $y$ solution de (E), $I = [t_0, T] \subset \mathbb{R}$ et $f : I \times \mathbb{R} \to \mathbb{R}$ une fonction $k$-lipschitzienne par rapport à son deuxième argument. Montrer que le schéma de Heun est convergent. De quel ordre est ce schéma ?

## 1.4 Généralisation des schémas explicites à un pas.

### 1.4.1 Méthode des approximations successives de Picard

La méthode de Picard se base sur la solution (1.5) du problème de Cauchy $(E)$ avec condition initiale $y(t_0) = y_0$ :

$$\forall t \in I, \quad y(t) = y_0 + \int_{t_0}^{t} f(s, y(s))\, ds$$

L'idée est de construire une suite $\tilde{y}_i$ d'approximations de la solution exacte en substituant, à chaque itération, la fonction inconnue $y(t)$ sous le signe intégrale par l'itérée précédente, tel que l'algorithme s'écrive :

$$\begin{cases} \tilde{y}_0(t) = y_0 \\ \tilde{y}_i(t) = y_0 + \displaystyle\int_{t_0}^{t} f(s, \tilde{y}_{i-1}(s))\, ds \end{cases}$$

Historiquement, en 1890 Emile Picard a introduit cette méthode afin de démontrer le théorème de Cauchy-Lipschitz dans son *Mémoire sur la théorie des équations aux dérivées partielles et la méthode des approximations successives*. La preuve décrite dans le théorème (1.5) reprend tous ces éléments et permet de nous convaincre de la convergence de la suite $\tilde{y}_i$.

Cette méthode n'est pas très utilisée en pratique puisqu'il est nécessaire de calculer analytiquement un grand nombre d'intégrales, dont il faut être capable de déterminer effectivement les primitives appropriées.

**Exercice 1.3 (**Méthode de Picard 1**).** Utiliser la méthode de Picard pour trouver la solution approchée de l'EDO

$$y' = t^2 + y^2$$

qui satisfait la condition initiale $y(0) = -1$. Calculer deux approximations successives.

**Exercice 1.4 (**Méthode de Picard 2**).** Appliquer la méthode des approximations de Picard pour résoudre le problème de Cauchy suivant :

$$\begin{cases} y' = t + y \qquad t \geq 0 \\ y(0) = 0 \end{cases}$$

En déduire la solution exacte.

### 1.4.2 Méthode de Taylor

Un moyen simple de construire une méthode d'ordre $p$ est de développer $y$ en série de Taylor à l'ordre $p$, de telle sorte que :

$$\frac{y(t+h)-y(t)}{h} = y'(t)+\frac{h}{2}\,y''(t)+\ldots+\frac{h^{p-1}}{p!}y^{(p)}(t)+o(h^{p-1}) = f(t,y)+\frac{h}{2}\,f^{[1]}(t,y)+\ldots+\frac{h^{p-1}}{p!}f^{[p-1]}(t,y)+o(h^{p-1})$$

Si on choisit $\Phi(t,y,h) = f(t,y) + \frac{h}{2}\,f^{[1]}(t,y) + \ldots + \frac{h^{p-1}}{p!}f^{[p-1]}(t,y)$ alors la méthode est clairement d'ordre au moins $p$ : c'est la méthode du développement de Taylor.

**Exercice 1.5 (**Méthode de Taylor**).** Soit le problème de Cauchy :

$$(E)\begin{cases} y' = y & t \in [0,1] \\ y(0) = 1 \end{cases}$$

Déterminer l'algorithme de Taylor d'ordre $p$ correspondant à $(E)$.

### 1.4.3 Schémas explicites de Runge et Kutta

#### Première idée

Essayons de construire une méthode d'ordre deux. Le problème de la méthode de Taylor est qu'à chaque pas il faut évaluer $f$ et $\{f^{[i]}\}_{0\leq i\leq p}$, ce qui peut être compliqué. L'idée est alors d'introduire une méthode d'ordre 2 ne nécessitant que 2 itérations de $f$. Cherchons alors $\Phi$ de la forme :

$$\begin{cases} \Phi(t,y,h) = a_1\,k_1(t,y,h) + a_2\,k_2(t,y,h) \\ k_1(t,y,h) = f(t,y) \\ k_2(t,y,h) = f(t+\alpha\,h, y+\beta\,h\,k_1(t,y,h)) \end{cases}$$

où les coefficients $(a_1,a_2,\alpha,\beta) \in \mathbb{R}^4$ sont à déterminer afin d'obtenir une méthode d'ordre 2. Premièrement calculons $\partial_3\Phi$ :

$$\partial_3\Phi(t,y,h) = a_2\,[\alpha\,\partial_1 f(t+\alpha\,h, y+\beta\,h\,k_1(t,y,h)) + \beta\,f(t,y)\,\partial_2 f(t+\alpha\,h, y+\beta\,h\,k_1(t,y,h))]$$

Par conséquent la méthode est d'ordre au moins 2 si :

$$\begin{cases} \Phi(t,y,0) = f(t,y) = (a_1+a_2)\;f(t,y) \\ \partial_3\Phi(t,y,0) = a_2\,[\alpha\,\partial_1 f(t,y) + \beta\,f(t,y)\,\partial_2 f(t,y)] = \dfrac{1}{2}\,[\partial_1 f(t,y) + f(t,y)\,\partial_2 f(t,y)] \end{cases}$$

i.e

$$\begin{cases} a_1 + a_2 = 1 \\ a_2\,\alpha = \dfrac{1}{2} \\ a_2\,\beta = \dfrac{1}{2} \end{cases}$$

On obtient ainsi une famille de méthodes d'ordre au moins deux, appelé $\theta$-schéma en posant $a_2 = \theta$ :

$$\begin{cases} \Phi(t,y,h) = (1-\theta)\;k_1(t,y,h) + \theta\,k_2(t,y,h) \\ k_1(t,y,h) = f(t,y) \\ k_2(t,y,h) = f(t+\dfrac{h}{2\theta}, y+\dfrac{h}{2\theta}\,k_1(t,y,h)) \end{cases}$$

— $\theta = \frac{1}{2}$ : schéma de Heun
— $\theta = 1$ : schéma d'Euler modifié

**Définition**

**Définition 1.14.**

On généralise ce qui précède pour définir la méthode de Runge-Kutta explicite d'ordre $q$ :

$$\begin{cases} \Phi(t,y,h) = \sum_{i=1}^{q} a_i\, k_i(t,y,h) \\ \forall 1 \leq i \leq q, \quad k_i(t,y,h) = f\left(t + \alpha_i\, h, y + h \sum_{j=1}^{i-1} \beta_{ji}\, k_j(t,y,h)\right) \end{cases}$$

où $\alpha_1 = 0$ est fixé et tous les coefficients $a_i, \alpha_i, \beta_{ji}$ sont à déterminer.

**Définition 1.15.**

Un tableau de Butcher rassemble les coefficients d'une méthode de Runge-Kutta :

| | |
|---|---|
| $\alpha_i$ | $\beta_{ji}$ |
| | $a_i$ |

**RK1**

Soit le schéma RK1 : $\begin{cases} \Phi(t,y,h) = a_1\, k_1(t,y,h) \\ k_1(t,y,h) = f(t,y) \end{cases}$

La consistance du schéma impose : $\Phi(t,y,0) = a_1 f(t,y) = f(t,y) \Rightarrow a_1 = 1$ Par conséquent, le schéma RK1 correspond au schéma d'Euler-explicite.

**RK2**

Vu plus haut, $\theta$-schéma.

**RK3**

Un des RK3 les plus utilisés est :

$$\begin{cases} \Phi(t,y,h) = \frac{1}{6}\left(k_1(t,y,h) + 4\, k_2(t,y,h) + k_3(t,y,h)\right) \\ k_1(t,y,h) = f(t,y) \\ k_2(t,y,h) = f(t + \frac{h}{2}, y + \frac{h}{2}\, k_1(t,y,h)) \\ k_3(t,y,h) = f(t+h, y - h\, k_1(t,y,h) + 2\, h\, k_2(t,y,h)) \end{cases}$$

| | | | |
|---|---|---|---|
| 0 | 0 | 0 | 0 |
| 1/2 | 1/2 | 0 | 0 |
| 1 | -1 | 2 | 0 |
| | 1/6 | 2/3 | 1/6 |

**RK4**

Un des RK4 les plus utilisés est :

$$\begin{cases} \Phi(t,y,h) = \dfrac{1}{6}\left(k_1(t,y,h) + 2\,k_2(t,y,h) + 2\,k_3(t,y,h) + k_4(t,y,h)\right) \\ k_1(t,y,h) = f(t,y) \\ k_2(t,y,h) = f(t+\dfrac{h}{2}, y+\dfrac{h}{2}\,k_1(t,y,h)) \\ k_3(t,y,h) = f(t+\dfrac{h}{2}, y+\dfrac{h}{2}\,k_2(t,y,h)) \\ k_4(t,y,h) = f(t+h, y+h\,k_3(t,y,h)) \end{cases}$$

| | | | | |
|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 |
| 1/2 | 1/2 | 0 | 0 | 0 |
| 1/2 | 0 | 1/2 | 0 | 0 |
| 1 | 0 | 0 | 1 | 0 |
| | 1/6 | 1/3 | 1/3 | 1/6 |

Le gros avantage de ce choix de coefficients est que chaque fonction $k_i$ ne dépend que d'une seule fonction $k$.

## 1.5 Application : pendule pesant

Prenons l'exemple d'un pendule pesant (1.1), sans prendre en compte les frottements de l'air, positionné initialement avec un angle de 45 degrés par rapport à la verticale et lâché sans vitesse. Si on pose le vecteur $\underline{X} = \begin{pmatrix} \theta \\ \Omega \end{pmatrix}$, le problème (E) s'écrit :

$$\begin{cases} \theta''(t) + \omega_0^2 \sin(\theta(t)) = 0 \\ \theta(0) = \dfrac{\pi}{4} \\ \theta'(0) = 0 \end{cases} \Leftrightarrow \begin{cases} \underline{X}'(t) = \underline{F}(t, \underline{X}(t)) = \begin{pmatrix} \Omega(t) \\ -\omega_0^2 \sin(\theta(t)) \end{pmatrix} \\ \theta(0) = \dfrac{\pi}{4} \\ \Omega(0) = 0 \end{cases}$$

Pour l'exercice écrivons les fonctions $\underline{k_i}$ du schéma RK4 :

$$\underline{k}_1(t,\underline{X},h) = \underline{F}(t,\underline{X}) = \begin{pmatrix} \Omega \\ -\omega_0^2 \sin(\theta) \end{pmatrix}$$

$$\underline{k}_2(t,\underline{X},h) = \underline{F}\left(t+\frac{h}{2}, \underline{X}+\frac{h}{2}\underline{k}_1(t,\underline{X},h)\right) = \begin{pmatrix} \Omega + \frac{h}{2}\left(-\omega_0^2 \sin(\theta)\right) \\ -\omega_0^2 \sin\left(\theta + \frac{h}{2}\Omega\right) \end{pmatrix}$$

$$\underline{k}_3(t,\underline{X},h) = \underline{F}\left(t+\frac{h}{2}, \underline{X}+\frac{h}{2}\underline{k}_2(t,\underline{X},h)\right) = \begin{pmatrix} \Omega + \frac{h}{2}\left(-\omega_0^2 \sin\left(\theta + \frac{h}{2}\Omega\right)\right) \\ -\omega_0^2 \sin\left(\theta + \frac{h}{2}\left(\Omega + \frac{h}{2}\left(-\omega_0^2 \sin(\theta)\right)\right)\right) \end{pmatrix}$$

$$\underline{k}_4(t,\underline{X},h) = \underline{F}(t+h, \underline{X}+h\underline{k}_3(t,\underline{X},h)) = \begin{pmatrix} \Omega + h\left(-\omega_0^2 \sin\left(\theta + \frac{h}{2}\left(\Omega + \frac{h}{2}\left(-\omega_0^2 \sin(\theta)\right)\right)\right)\right) \\ -\omega_0^2 \sin\left(\theta + h\left(\Omega + \frac{h}{2}\left(-\omega_0^2 \sin\left(\theta + \frac{h}{2}\Omega\right)\right)\right)\right) \end{pmatrix}$$

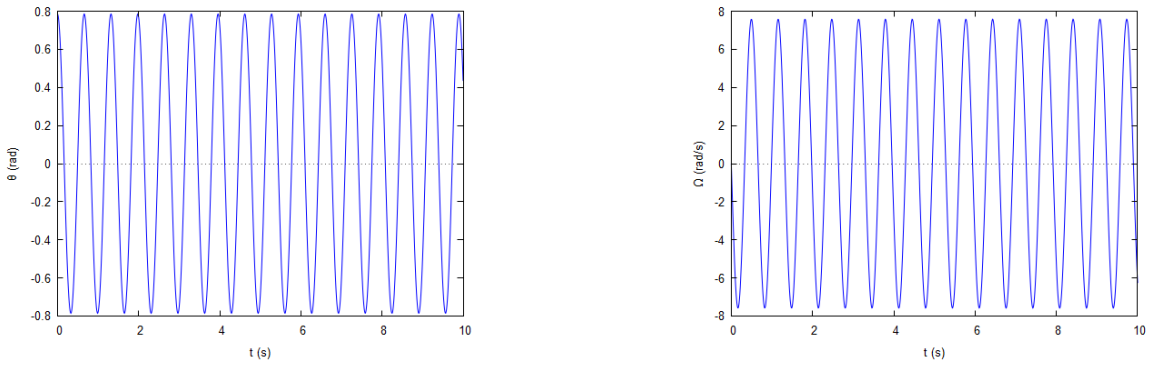



FIGURE 1.4 – Exemple : résolution par RK4 (h=0.01s) de l'équation du pendule d'une longueur de 10 cm sans frottements.

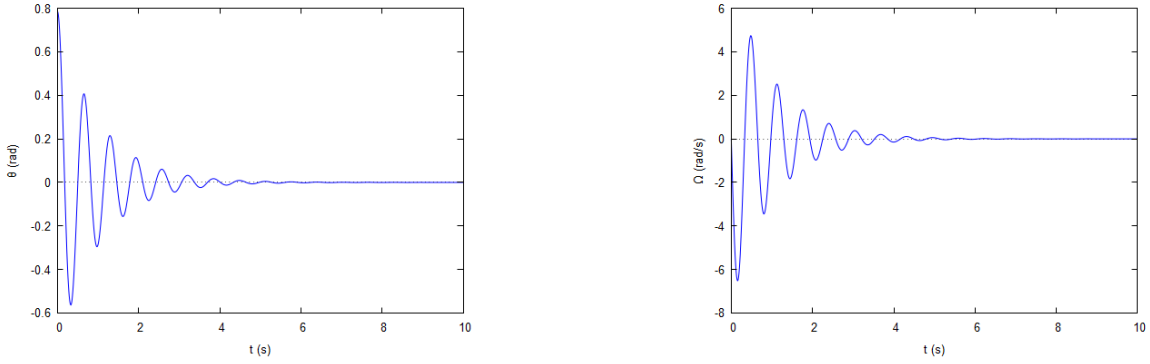



FIGURE 1.5 – Exemple : résolution par RK4 (h=0.01s) de l'équation du pendule d'une longueur de 10 cm avec frottements, $\xi = 0.1$

## 1.6 Exercices supplémentaires

### 1.6.1 Exercices d'application

**Exercice 1.6.** Résoudre l'équation suivante par la méthode d'Euler explicite :

$$\begin{cases} \theta_0 = \dfrac{\pi}{6}\,,\ \dot{\theta}_0 = -1 \\ \ddot{\theta} + 2\dot{\theta} + 10\sin\theta = 0 \end{cases}$$

On prendra comme pas $h = 0.1$ et on fera 2 itérations (donner les résultats à $10^{-3}$ près).
Bonus : faire varier le pas et observer l'influence sur les résultats.

**Exercice 1.7.** Appliquer le schéma de Runge-Kutta d'ordre 2 pour résoudre :

$$\begin{cases} x_0 = 1 \\ \dot{x} - 4x = \cos t \end{cases}$$

On calculera $x(0,025)$ et $x(0,05)$ à $10^{-5}$ près.

### 1.6.2 TD

**Exercice 1.8 (**TD 1**).** On s'intéresse à l'équation aux dérivées ordinaires suivante :

$$\begin{cases} y(0) = 0 \\ \dot{y} = y^2 + 2y + 2 \end{cases}$$

1. Résoudre analytiquement l'équation et donner la solution exacte.
2. On veut maintenant résoudre cette EDO numériquement. On utilise pour ce faire le schéma d'Euler implicite :
3. Ecrire le schéma correspondant à cette équation.
4. Exprimer $y_{i+1}$ en fonction de $y_i$ ; on précisera les valeurs de $h$ pour lesquelles c'est possible.
5. Calculer $y_i$ pour $i \in [1, 5]$ pour un pas constant $h = t_{i+1} - t_i = 0.05$. Comparer avec les valeurs exactes.
6. Peut-on approcher $y(0, 5)$ avec cette méthode ? Qu'en est-il si on choisit un pas de $h = 0.1$ ?

**Exercice 1.9 (**TD 2**).** On veut étudier les propriétés du schéma d'Euler sous ses différentes formes.

1. Montrer que le schéma d'Euler implicite est convergent d'ordre 1 :

$$y_{n+1} = y_n + h\, f(y_{n+1}, t_{n+1})$$

2. Montrer que le schéma d'Euler modifié est convergent d'ordre 2 :

$$\begin{cases} y_{n+\frac{1}{2}} = y_n + \dfrac{h}{2}\, f(y_n, t_n) \\ y_{n+1} = y_n + h\, f(y_{n+\frac{1}{2}}, t_{n+\frac{1}{2}}) \qquad \text{où } t_{n+\frac{1}{2}} = t_n + \dfrac{h}{2} \end{cases}$$

3. Soit $k \in \mathbb{R}^+$. On considère l'équation différentielle :

$$\begin{cases} y(0) = 1 \\ y' + k\, y^2 = 0 \end{cases}$$

   (i) Résoudre cette équation différentielle.

(ii) Écrire le schéma d'Euler explicite et étudier le comportement de la solution lorsque $n$ tend vers l'infini.

(iii) On considère le schéma :

$$\frac{1}{h}\,(y_{n+1}-y_n)+\frac{k}{4}\,((y_{n+1}-y_n))^2=0$$

Calculer la solution $y_{n+1}$ en fonction de $y_n$, $k$ et $h$ et en déduire une condition sur $h$ pour que le comportement de la solution numérique soit correct lorsque $n\to\infty$. Montrer que ce schéma est consistant d'ordre 2.

### 1.6.3 Annales

**Exercice 1.10 (**Annale 2019**).** Soient $[t_0,T]$ in intervalle de $\mathbb{R}$, $f:[t_0,T]\times\mathbb{R}\to\mathbb{R}$ une fonction de classe $\mathcal{C}^2$, $L$-lipschitzienne par rapport à son deuxième argument, $y_0\in\mathbb{R}$ et le problème de Cauchy suivant :

$$\begin{cases} y'(t)=f(t,y(t)) & t\in[t_0,T] \\ y(t_0)=y_0 \end{cases}$$

On considère, pour intégrer ce problème de Cauchy, le schéma de Runge-Kutta défini par le tableau suivant, où $(\alpha,\beta,\gamma)$ sont trois réels strictement positifs :

| | | | |
|---|---|---|---|
| 0 | 0 | 0 | 0 |
| 1/2 | $-1/2$ | 0 | 0 |
| 1 | -1 | 2 | 0 |
| | $\alpha$ | $-\beta$ | $\gamma$ |

1. Ecrire la fonction $\Phi$ associé à ce schéma.
2. Montrer par le calcul que ce schéma est stable pour tout $(\alpha,\beta,\gamma)$ et exprimer la constante de stabilité.
3. Sous quelles conditions ce schéma est-il au moins d'ordre 2 ?

**Exercice 1.11 (**Annale 2020**).** Soit $(\mathcal{E})$ le problème de Cauchy suivant, où $\dot\varphi=\frac{d\varphi}{dt}$ et $(\omega_0,\xi)$ deux réels positifs :

$$(\mathcal{E})\begin{cases} \ddot\varphi+2\,\xi\,\omega_0\,\dot\varphi+\omega_0^2\,\sin(\varphi)=0 \qquad (E) \\ \varphi(0)=\dfrac{\pi}{4} \\ \dot\varphi(0)=0 \end{cases}$$

1. Réécrire $(E)$ sous la forme $\dot{\underline{X}}=\underline{F}(\underline{X})$.
2. Le problème $(\mathcal{E})$ admet-il une unique solution ? [On pourra utiliser la $\|.\|_\infty$]
3. On cherche une solution approchée grâce au $\theta$-schéma (RK2) avec $\theta=1/4$ dans un intervalle $I=[t_0,T]$ et pour des pas de temps $h\in[0,H]$ ($0<t_0<T$ et $H>0$). Dans le cas général :
   — Ecrire le tableau de Butcher correspondant ;
   — Démontrer que ce schéma est convergent d'ordre au moins deux ;
   — Décrire la démarche (sans la mettre en œuvre) de résolution numérique.
4. Ecrire explicitement (en faisant intervenir en particulier $\underline{X}$, $h$, $\xi$, $\omega_0$) l'expression de $\underline{\Phi}(\underline{X},h)$ en appliquant le schéma précédent à $(E)$. Vérifier la consistance du schéma avec $(E)$.
5. Sur la Figure 1.7 sont représentées quatre résolutions de $(\mathcal{E})$ selon le schéma proposé pour $t\in[0,5]s$. Selon vous quel paramètre a changé ? Quelle(s) solution(s) est/sont acceptable(s) ? Justifier votre réponse.

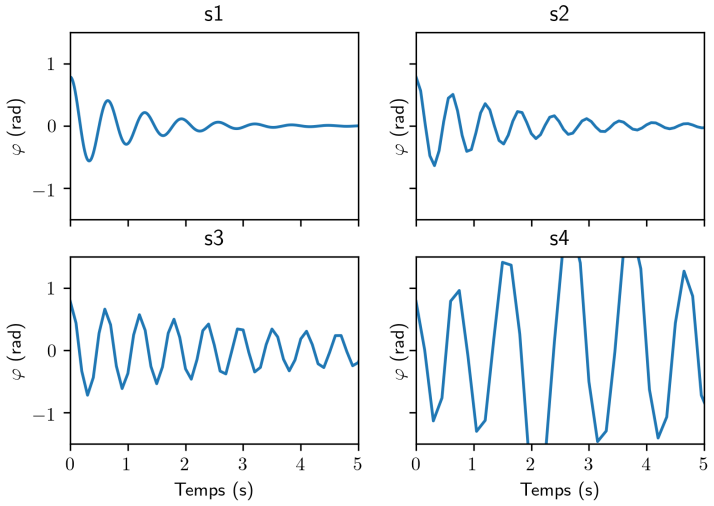



FIGURE 1.6 – Quatre résolutions de $(\mathcal{E})$ avec le même schéma pour $t \in [0, 5]s$.

**Exercice 1.12.** Soient $[0, H]$ un intervalle de $\mathbb{R}^+$, $[t_0, T]$ un intervalle de $\mathbb{R}$, $(E, \|.\|)$ un espace de Banach et $f : [t_0, T] \times E \mapsto E$ une fonction suffisamment régulière. On suppose qu'il existe $L > 0$ tel que pour tout $t$ dans $[t_0, T]$ et pour tout $(x_1, x_2)$ dans $E^2$ la fonction $f$ vérifie la condition de Lipschitz :

$$\|f(t, x_1) - f(t, x_2)\| \leq L\|x_1 - x_2\|$$

On considère le schéma à un pas :

$$\begin{cases} y_0 \in E \\ y_{n+1} = y_n + h_n F(t_n, y_n, h_n) \qquad n \in [\![0, N-1]\!] \end{cases}$$

où la fonction $F$ est définie sur $[t_0, T] \times E \times [0, H]$ par la formule :

$$\begin{aligned} F(t, y, h) =& \frac{1}{6} f(t, y) + \frac{2}{3} f\left(t + \frac{1}{2}h, y + \frac{1}{2}hf(t, y)\right) \\ &+ \frac{1}{6} f\left(t + h, y - hf(t, y) + 2hf\left(t + \frac{1}{2}h, y + \frac{1}{2}hf(t, y)\right)\right) \end{aligned}$$

1. Démontrer que ce schéma est convergent.
2. Démontrer que ce schéma est au moins d'ordre 2.
3. En utilisant la fonction $f$ définie sur $\mathbb{R}^2$ par $f(t, y) = t + y$, montrer que $F(t, y, h)$ est un polynôme en $h$ dont on déterminera le degré et en déduire que le schéma n'est pas, en général, d'ordre 4.
4. Montrer que ce schéma est un schéma de Runge-Kutta en déterminant son tableau de Butcher. Quel est son rang ou nombre d'étages ?

**Exercice 1.13 (**Annale 2016**).** Soient $N \in \mathbb{N}^\star$, $(x_0, X, u_0, H) \in (\mathbb{R}^\star)^4$, $I = [x_0, X]$ un intervalle compact d'intérieur non vide et $(E, \|.\|)$ un espace de Banach réel. On considère une fonction $f : I \times E \mapsto E$ de classe $\mathcal{C}^3$ et $L-$lipschitzienne par rapport à son second argument uniformément par rapport au premier :

$$\forall x \in I, \forall (y, \tilde{y}) \in E^2, \qquad \|f(x, y) - f(x, \tilde{y})\| \leq L\|y - \tilde{y}\|$$

1. Démontrer que pout tout $(x, y, h)$ dans $I \times E \times [0, H]$, où $H$ vérifie $0 < H < \frac{3}{L}$, l'équation d'inconne $s$ à trois paramètres $x, y, h$ :

$$f(x + \frac{2}{3}h, y + \frac{1}{3}hf(x, y) + \frac{1}{3}hs) = s$$

admet une unique racine dans $E$ que l'on notera $k(x, y, h)$.

L'existence pour tout $(x, y, h)$ dans $I \times E \times [0, H]$ de cette unique racine permet de définir la fonction $k : I \times E \times [0, H] \mapsto E$. On admettra sans démonstration que $k$ admet une dérivée partielle $\partial_3 k$ par rapport à son troisième argument et que les deux fonctions $k$ et $\partial_3 k$ sont continues sur $I \times E \times [0, H]$.

On considère le problème de Cauchy

$$\begin{cases} u'(x) = f(x, u(x)) & \forall x \in I \\ u(x_0) = u_0 \end{cases}$$

une subdivision de $I$ à pas variabes $h_n = x_{n+1} - x_n$ :

$$x_0 < x_1 < \ldots < x_n < \ldots < x_{N-1} < x_N = X$$

et on note $y_n$ l'approximation de $u(x_n)$ calculée par le schéma numérique à un pas défini par :

$$y_{n+1} = y_n + \frac{1}{4} h_n f(x_n, y_n) + \frac{3}{4} h_n k(x_n, y_n, h_n) \quad \text{avec} \quad 0 \leq n \leq N - 1$$

2. S'agit-il d'un schéma explicite ou bien implicite ? Pourquoi ?
3. Etablir l'expression du schéma sous sa forme canonique :

$$y_{n+1} = y_n + h_n \Phi(x_n, y_n, h_n)$$

où pour tout $(x, y, h) \in I \times E \times [0, H]$, $\Phi(x, y, h)$ sera explicité en fonction de $x, y, h, f(x, y)$ et $k(x, y, h)$.
4. Démontrer que la fonction $k$ est lipschitzienne par rapport à son deuxième argument.
5. Démontrer que le schéma est convergent.
6. Démontrer que le schéma est d'ordre au moins 2.
7. On considère la fonction $f : \mathbb{R}^2 \mapsto \mathbb{R}$ définie par $f(x, y) = x + y$. Démontrer que $f$ est 1-lipschitzienne par rapport à son seconde argument. Déterminer $k(x, y, h)$ et $\Phi(x, y, h)$ pour tout $(x, y, h) \in I \times E \times [0, H]$ avec $H < 3$. Pour cette fonction particulière, le schéma est-il d'ordre 3 ?
8. Montrer que ce schéma est un schéma de Runge-Kutta en déterminant son tableau de Butcher. Quel est son rang (nombre d'étages) ?

# Chapitre 2

# Equations aux dérivées partielles elliptiques : analyse et résolution par différences finies.

*Cette partie est dédiée à l'étude d'un problème aux dérivées partielles de type elliptique, puis à sa résolution par un schéma aux différences finies. Pour montrer que ce problème aux limites est bien posé au sens de Hadamard, c'est-à-dire qu'il admet une unique solution dépendant continûment de ses données, nous nous plaçons dans un espace de Sobolev et nous suivons l'approche variationnelle.*

**Sommaire**



## 2.1 EDP linéaires du second ordre

### 2.1.1 Définition et exemples

**Définition 2.1.**

Soit une fonction $u : \Omega \subset \mathbb{R}^N \to \mathbb{R}$ satisfaisant l'équation différentielle :

$$\underline{\underline{A}}(\underline{x}) : \underline{\underline{H}}_u(\underline{x}) + \underline{B}(\underline{x}).\underline{\nabla} u(\underline{x}) + c(\underline{x})u(\underline{x}) = f(\underline{x}) \qquad \forall \underline{x} \in \Omega \tag{2.1}$$

$\left(\underline{\underline{H}}_u(\underline{x}), \underline{\nabla} u(\underline{x})\right)$ représentent respectivement la matrice hessienne de $u$ (i.e $(\underline{\underline{H}}_u(\underline{x}))_{ij} = \partial^2_{ij} u(\underline{x})$) et le gradient de $u$ (i.e $(\underline{\nabla} u(\underline{x}))_i = \partial_i u(\underline{x})$). Les coefficients des termes d'ordre deux sont représentés par la

matrice $\underline{\underline{A}} \in \mathcal{M}_N(\mathbb{R})$ symétrique (donc diagonalisable), ceux d'ordre un par le vecteur $\underline{B} \in \mathbb{R}^N$ et ceux d'ordre zéro par $c$. L'équation différentielle (2.1) est une EDP linéaire d'ordre au plus deux portée par la variable $u$ et soumise au terme source $x \mapsto f(x)$.

**Exemple 2.1.** Quelques exemples en coordonnées cartésiennes en 3D, donc $\underline{x} = (t, x_1, x_2, x_3)$ :

**Equation de Poisson :** $\Delta u(\underline{x}) = f(x) \qquad \forall \underline{x} \in \mathbb{R}^3$. Dans ce cas il n'y a pas de dépendance temporelle ($\underline{x} = (x_1, x_2, x_3)$) ni dérivées spatiales croisées, donc on a :

$$\underline{\underline{A}}(\underline{x}) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad ; \quad \underline{B}(\underline{x}) = \underline{0} \quad ; \quad c(\underline{x}) = 0$$

**Equation de propagation d'ondes** à célérité constante $\mathcal{C}$ :

$$\Delta u(\underline{x}) - \frac{1}{\mathcal{C}^2}\partial_1^2 u(\underline{x}) = f(\underline{x}) \qquad \forall t \geq 0, \forall \underline{x} \in \mathbb{R}^3$$

Dans ce cas il n'y a pas de dérivées croisées, donc

$$\underline{\underline{A}}(\underline{x}) = \begin{bmatrix} -\frac{1}{\mathcal{C}^2} & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad ; \quad \underline{B}(\underline{x}) = \underline{0} \quad ; \quad c(\underline{x}) = 0$$

**Equation de diffusion** avec $\underline{\underline{D}} \in \mathcal{M}_3(\mathbb{R})$ définie positive symétrique représentant les coefficients de diffusion variables en espace :

$$\operatorname{div}\left(\underline{\underline{D}}(\underline{x}).\underline{\nabla} u(\underline{x})\right) - \partial_1 u(\underline{x}) = f(\underline{x}) \qquad \forall t \geq 0, \forall \underline{x} \in \mathbb{R}^3$$

Le développement du premier terme s'écrit, en notations d'Einstein (pour $2 \leq i, j \leq 4$) :

$$\left(D_{ij} u_{,j}\right)_{,i} = D_{ij} u_{,ij} + (D_{ij})_{,i} u_{,j}$$

soit en écriture tensorielle :

$$\operatorname{div}\left(\underline{\underline{D}}(\underline{x}).\underline{\nabla} u(\underline{x})\right) = \underline{\underline{D}}(\underline{x}) : \underline{\underline{H}}_u(\underline{x}) + \underline{\operatorname{div}}(\underline{\underline{D}}(\underline{x})).\underline{\nabla} u(\underline{x})$$

On a alors :

$$\underline{\underline{A}}(\underline{x}) = \left[\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \hline 0 & & & \\ 0 & & \underline{\underline{D}}(\underline{x}) & \\ 0 & & & \end{array}\right] \quad ; \quad \underline{B}(\underline{x}) = \begin{pmatrix} -1 \\ \hline \\ \underline{\operatorname{div}}\left(\underline{\underline{D}}\,(\underline{x})\right) \\ \\ \end{pmatrix} \quad ; \quad c(\underline{x}) = 0$$

### 2.1.2 Classification

Comme $\underline{\underline{A}}$ est diagonalisable, les EDP linéaires elliptiques du 2nd ordre peuvent être rangées selon trois grandes classes : les EDP elliptiques, paraboliques et hyperboliques.

**Remarque 2.1.**
Ces appellations proviennent directement des trois classes de coniques (ellipses, paraboles et hyperboles) que l'on retrouve en calculant les courbes vérifiant ${}^t\underline{x}.\underline{\underline{A}}.\underline{x} = \textit{constante}$ avec $\underline{\underline{A}}$ est constante.

**Définition 2.2.**

Soit $\underline{\underline{A}}$ la matrice des coefficients des termes du second ordre de l'EDP (2.1).

- Si $\underline{\underline{A}}$ n'admet que des valeurs propres non nulles et toutes de même signe, alors (2.1) est **elliptique** ;
- Si $\underline{\underline{A}}$ n'admet que des valeurs propres non nulles et toutes de même signe sauf une de signe opposé, alors (2.1) est **hyperbolique** ;
- Si $\underline{\underline{A}}$ admet $N-1$ valeurs propres non nulles de même signe et une valeur propre nulle, et si de plus le noyau de $\underline{\underline{A}}$ vérifie

$$\underline{\ker}(\underline{\underline{A}}(\underline{x})).\underline{B}(\underline{x}) \neq 0 \tag{2.2}$$

alors (2.1) est **parabolique**.

**Remarque 2.2.**

Une EDP elliptique (resp. hyperbolique resp. parabolique) sur tout $\Omega$ traduit généralement des phénomènes stationnaires (resp. de propagation d'onde resp. de diffusion). Si la condition (2.2) n'est pas vérifiée ce n'est pas une EDP parabolique mais un couplage entre une EDP et une équation algébrique.

**Exercice 2.1.** Déterminer les classes d'EDP des exemples 2.1.

## 2.2 Etude des problèmes aux limites

### 2.2.1 Rappels utiles

Cette partie rappelle quelques notions du cours d'Analyse qui sont fondamentales pour aborder la résolution de problèmes aux limites.

#### Intégrabilité au sens de Lebesgue

Soit $\Omega$ un ouvert de $\mathbb{R}^N$, muni de la mesure de Lebesgue. On définit l'espace de fonction $L^2(\Omega)$ comme l'espace des fonctions de carré intégrable (au sens de Lebesgue), i.e [1] :

$$L^2(\Omega) = \{f \text{ mesurable dans } \Omega \text{ et } \int_\Omega |f(x)|^2 \, d\Omega < \infty\}$$

De plus, $L^2(\Omega)$ muni du produit scalaire

$$\langle f, g \rangle_{L^2(\Omega)} = \int_\Omega f(x) g(x) \, d\Omega$$

est un espace de Hilbert. La norme induite par ce produit scalaire est la norme classique pythagoricienne :

$$\|f\|_{L^2(\Omega)} = \sqrt{\int_\Omega |f(x)|^2 \, d\Omega}$$

Enfin, rappelons qu'une propriété est vérifiée *presque partout*, notée $\mu pp$, si l'espace sur lequel cette propriété n'est pas vérifiée est de mesure de Lebesgue nulle. Par exemple :

$$\int_\Omega f(x) \, d\Omega = \int_\Omega g(x) \, d\Omega \Rightarrow f(x) = g(x) \, \mu pp \, x \in \Omega$$

1. Dans la pratique, bien souvent la mesurabilité d'une fonction dans $\Omega$ est réglée par sa continuité dans $\Omega$ (Attention la réciproque est fausse, par exemple du Dirac). Sinon il faut repartir de la définition d'une fonction mesurable.

**Dérivation faible - lien avec les distributions**

L'objectif de ce paragraphe est d'introduire la notion de dérivation faible, qui est une généralisation de la dérivation "classique" et un cas particulier de la dérivation au sens des distributions. On note $\mathcal{D}(\Omega) = \mathcal{C}_c^\infty(\Omega)$ l'espaces des fonctions $\mathcal{C}^\infty(\Omega)$ à support compact dans $\Omega$.

**Définition 2.3.**

Soit $v \in L^2(\Omega)$. On dit que $v$ est dérivable au sens faible dans $L^2(\Omega)$ s'il existe des fonction $\{w_i \in L^2(\Omega)\}_{i\in[\![1;N]\!]}$ telles que :

$$\forall \phi \in \mathcal{D}(\Omega),\ \int_\Omega v(x)\partial_{x_i}\phi(x)\, d\Omega = -\int_\Omega w_i(x)\phi(x)\, d\Omega$$

où les fonctions $w_i = \partial_{x_i} v$ sont les dérivées partielles faibles de $v$.

Dans la théorie des distributions, $\mathcal{D}(\Omega)$ est l'espace des fonctions tests, et son dual $\mathcal{D}'(\Omega)$ l'espace des distributions, c'est-à-dire l'espace des formes linéaires "continues" (Attention pas de norme définie mais seulement la convergence) sur $\Omega$. La dérivation au sens faible est bien un cas particulier de la dérivation au sens des distributions, dans laquelle si $T \in \mathcal{D}'(\Omega)$, on définit $\partial_{x_i} T \in \mathcal{D}'(\Omega)$ tel que :

$$\forall \phi \in \mathcal{D}(\Omega),\ \langle \partial_{x_i} T, \phi\rangle = -\langle T, \partial_{x_i}\phi\rangle$$

Il est évident mais utile de rappeler que si une fonction est dérivable au sens "classique", alors ses dérivées au sens des distributions et au sens "classique" correspondent.

**Conséquence.**

Soit $v \in L^2(\Omega)$ dérivable au sens faible, telle que toutes ses dérivées partielles faibles sont nulles. Alors, pour chaque composante connexe de $\Omega$, $\exists C > 0 / \, v(x) = C \ \mu pp$ dans cette composante connexe. Ceci généralise le résultat très connu en dérivation classique $v'(x) = 0 \Rightarrow \exists C > 0 / \, v(x) = C$.

Enfin, les espaces $L^p(\Omega)$ et $H^m(\Omega)$ (espaces de Sobolev) sont des sous-espaces de $\mathcal{D}'(\Omega)$, de telle sorte que toutes les notions de convergences dans ces espaces impliquent la convergence au sens des distributions (Attention la réciproque est fausse). Enfin, comme nous le verrons, les égalités dans les formulations variationnelles impliquent des égalités au sens des distributions.

**Espaces de Sobolev $H^1(\Omega)$, $H_0^1(\Omega)$**

**Définition 2.4.**

L'espace de Sobolev $H^1(\Omega)$ est défini tel que :

$$H^1(\Omega) = \{v \in\ L^2(\Omega) / \forall i \in [\![1; N]\!] \partial_{x_i} v \in\ L^2(\Omega)\}$$

où $\partial_{x_i} v$ est la dérivée partielle de $v$ au sens faible. Muni du produit scalaire

$$\langle f, g\rangle_{H^1(\Omega)} = \int_\Omega \left(f(x)g(x) + \nabla f(x).\nabla g(x)\right)\, d\Omega$$

$\left(H^1(\Omega), \langle ., .\rangle_{H^1(\Omega)}\right)$ est un espace de Hilbert. La norme induite par ce produit scalaire est :

$$\|f\|_{H^1(\Omega)} = \sqrt{\int_\Omega \left(|f(x)|^2 + |\nabla f(x)|^2\right)\, d\Omega}$$

**Remarque 2.3.**

si $N \geq 2$, les fonctions de $H^1(\Omega)$ ne sont en général ni continues ni bornées.

**Théorème 2.1** (de trace)**.**
*Soit $\Omega$ un ouvert borné régulier de classe $\mathcal{C}^1$. On définit l'application trace $\gamma_0$ telle que :*

$$\begin{aligned} H^1(\Omega) \cap \mathcal{C}(\bar{\Omega}) &\to L^2(\partial\Omega) \cap \mathcal{C}(\partial\bar{\Omega}) \\ v &\to \gamma_0(v) = v|_{\partial\Omega} \end{aligned}$$

*Cette application se prolonge par continuité en une application linéaire continue de $H^1(\Omega)$ dans $L^2(\partial\Omega)$, notée encore $\gamma_0$. En particulier :*

$$\exists C > 0 / \forall v \in H^1(\Omega), \quad \|v\|_{L^2(\partial\Omega)} \leq C\|v\|_{H^1(\Omega)}$$

**Remarque 2.4.**
Ce théorème de trace est fondamental, car il nous permet de définir la valeur au bord (ou trace) de $\Omega$ pour une fonction $H^1(\Omega)$. En effet, en lien avec la remarque 2.3, la valeur ponctuelle d'une fonction de $H^1(\Omega)$ n'est définie que presque partout, et le bord $\partial\Omega$ est un ensemble de mesure nulle. Ce résultat est d'autant plus fort qu'il n'est pas vrai pour une fonction de $L^2(\Omega)$.

**Définition 2.5.**
L'espace de Sobolev $H_0^1(\Omega)$ est défini comme l'adhérence de $\mathcal{D}(\Omega)$ dans $H^1(\Omega)$. Il correspond au sous-espace de $H^1(\Omega)$ constitué des fonctions qui s'annulent le bord $\partial\Omega$ (bien défini grâce au thérorème de trace 2.1). Une conséquence importante est que l'espace $\mathcal{D}(\Omega)$ est dense dans $H_0^1(\Omega)$. $H_0^1(\Omega)$ étant un sous-espace fermé de $H^1(\Omega)$ Hilbert, l'espace $H_0^1(\Omega)$ muni du produit scalaire $\langle ., . \rangle_{H^1(\Omega)}$ est aussi un Hilbert.

### 2.2.2 Différents problèmes

**Définition 2.6.**
On appelle problème aux limites la donnée (i) d'une équation aux dérivées partielles dans $\Omega$ et (ii) d'une ou plusieurs conditions aux limites sur $\partial\Omega$.

Le type de problème aux limites dépend donc de la nature des conditions sur la frontière $\partial\Omega$, voir figure 2.1. Les plus classiques sont :

— **Condition aux limites de Dirichlet :** les valeurs du champ sont imposées sur $\partial\Omega = \partial\Omega_d$

$$\begin{cases} -\Delta u = f & \text{dans} \quad \Omega \\ u = g_d & \text{sur} \quad \partial\Omega_d \end{cases}$$

— **Condition aux limites de Neumann :** les valeurs des gradients du champ sont imposées sur $\partial\Omega = \partial\Omega_n$.

$$\begin{cases} -\Delta u = f & \text{dans} \quad \Omega \\ \dfrac{\partial u}{\partial n} = g_n & \text{sur} \quad \partial\Omega_n \end{cases}$$

— **Condition aux limites de Robin (ou de Fourier ) :** C'est une généralisation des deux cas précédents : on impose une combinaison linéaire entre les valeurs du champ et des gradients du champ sur $\partial\Omega$

$$\begin{cases} -\Delta u = f & \text{dans} \quad \Omega \\ \dfrac{\partial u}{\partial n} + \alpha u = g_n & \text{sur} \quad \partial\Omega \end{cases}$$

— **Condition aux limites mêlée :** $\partial\Omega = \partial\Omega_d \bigcup \partial\Omega_n$

$$\begin{cases} -\Delta u = f & \text{dans} \quad \Omega \\ u = g_d & \text{sur} \quad \partial\Omega_d \\ \dfrac{\partial u}{\partial n} + \alpha u = g_n & \text{sur} \quad \partial\Omega_n \end{cases}$$

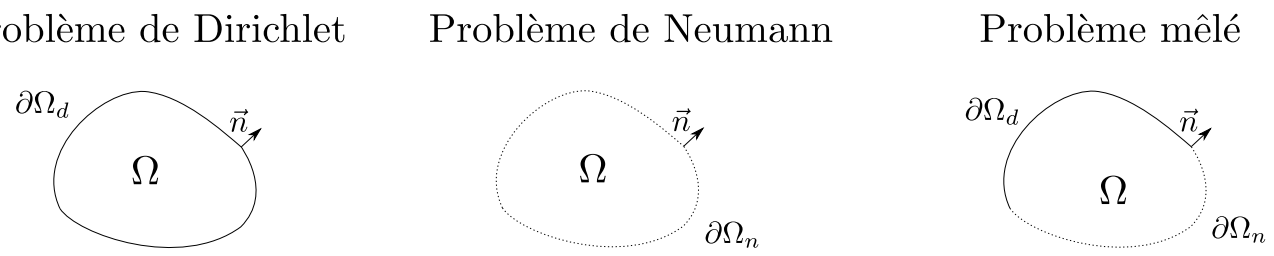



FIGURE 2.1 – Problèmes aux limites

Notons que l'étude des problèmes aux limites pour l'opérateur Laplacien peut être généralisé à des opérateurs elliptiques du second ordre à coefficients tensoriels variables. Ce type de problème est par exemple rencontré lors de l'étude du comportement statique d'un milieu élastique hétérogène, piloté par une EDP telle que :

$$\underline{\text{div}} \left( \underline{\underline{\underline{a}}}(\underline{x}) : \underline{\underline{\epsilon}}(\underline{u}(x)) \right) = \underline{0}$$

où $\underline{\underline{\underline{a}}}$ représente le tenseur élastique, et $\underline{u}, \underline{\underline{\epsilon}}(\underline{u})$ le champs de déplacement et le tenseur de déformation.

### 2.2.3 Conditions aux limites de Dirichlet

Soient $\Omega$ un ouvert borné de $\mathbb{R}^N$ de frontière $\partial\Omega$, $f \in L^2(\Omega)$ un second membre et $\vec{n}$ la normale sortante à $\partial\Omega$ telle que $\|\vec{n}\| = 1$. Nous étudions par exemple le problème aux limites de Dirichlet suivant :

$$\begin{cases} -\Delta u = f & \text{dans} \quad \Omega \\ u = 0 & \text{sur} \quad \partial\Omega \end{cases} \tag{2.3}$$

La méthode de l'approche variationnelle consiste alors à (i) trouver la formulation variationnelle équivalente à (2.3), (ii) montrer qu'il en existe une unique solution et (iii) démontrer l'équivalence entre le problème variationnel et (2.3).

**Remarque 2.5.**
La condition aux limites du problème (2.3) est homogène, i.e $u = 0$ sur $\partial\Omega$. Soit par exemple $g_d$ la trace sur $\partial\Omega$ d'une fonction $H^1(\Omega)$. Le problème s'écrit alors :

$$\begin{cases} -\Delta u = f & \text{dans} \quad \Omega \\ u = g_d & \text{sur} \quad \partial\Omega \end{cases} \tag{2.4}$$

En posant $\tilde{u} = u + g_d$ et $\tilde{f} = f + \Delta g_d$ on retombe sur le problème homogène (2.3).

#### Formulation variationnelle

Comme vu en cours d'Analyse, il faut trouver une forme bilinéaire $a(.,.)$, une forme linéaire $\ell(.)$ et un espace de Hilbert $V$ tels que le problème (2.3) soit équivalent à :

$$\text{Trouver } u \in V \text{ tel que } \forall v \in V,\ a(u,v) = \ell(v) \tag{2.5}$$

Pour ce faire, on multiplie l'EDP (2.3) par une fonction cinématiquement admissible $v$ (on verra par la suite dans quel espace il appartient) puis on intègre par parties. On obtient ainsi, à l'aide de la formule de Green :

$$\int_\Omega f(x)\, v(x)\, d\Omega = -\int_\Omega \Delta u(x)\, v(x)\, d\Omega = \int_\Omega \nabla u(x).\nabla v(x)\, d\Omega - \int_{\partial\Omega} v(x)\nabla u(x).n\, d\partial\Omega \tag{2.6}$$

le champ $v$ étant cinématiquement admissible, il satisfait la condition aux limites de Dirichlet sur $\partial\Omega$ i.e $u = v = 0$ sur $\partial\Omega$. Ainsi (2.6) se simplifie en :

$$\int_\Omega f(x)\, v(x)\, d\Omega = \int_\Omega \nabla u(x).\nabla v(x)\, d\Omega \tag{2.7}$$

Pour que le membre de gauche soit bien défini, il faut que $(f, v) \in L^2(\Omega)\times L^2(\Omega)$ (pour $f$ c'était une hypothèse du problème), et pour que le membre de droite soit bien défini, il faut que $(\nabla u, \nabla v) \in L^2(\Omega) \times L^2(\Omega)$. Par conséquent, l'espace $V$ cinématiquement admissible est l'espace de Sobolev $H_0^1(\Omega)$ défini par :

$$V = H_0^1(\Omega) = \Big\{v \in L^2(\Omega), \nabla v \in L^2(\Omega) \text{ et } v = 0 \text{ sur } \partial\Omega\Big\}$$

$V$ est bien un espace de Hilbert (espace vectoriel complet) muni du produit scalaire :

$$\langle u, v\rangle_V = \int_\Omega \left(u(x)v(x) + \nabla u(x)\nabla v(x)\right)\, d\Omega$$

Finalement, le problème variationnel s'écrit :

$$\text{Trouver } u \in H_0^1(\Omega) \text{ tel que } \forall v \in H_0^1(\Omega),\ \int_\Omega \nabla u(x).\nabla v(x)\, d\Omega = \int_\Omega f(x)\, v(x)\, d\Omega \tag{2.8}$$

et par identification avec la forme faible générale (2.5) :

$$a(u, v) = \int_\Omega \nabla u(x).\nabla v(x)\, d\Omega \qquad ; \qquad \ell(v) = \int_\Omega f(x)v(x)\, d\Omega$$

**Existence et unicité de la solution**

Le théorème de Lax-Milgram spécifie que si $a(.,.)$ est une forme bilinéaire continue coercive sur V et si $\ell(.)$ est une forme linéaire continue sur V, alors (2.7) admet une unique solution qui dépend continûment de la forme linéaire $\ell$.
Clairement la forme $a(.,.)$ est bilinéaire. Pour montrer sa continuité sur $V$, appliquons l'inégalité de Cauchy-Schwartz et utilisons la définition du produit scalaire dans $V$, tel que

$$\forall v \in V,\ \|v\|_V^2 = \langle v, v\rangle_V = \|v\|_{L^2(\Omega)}^2 + \|\nabla v\|_{L^2(\Omega)}^2 \geq \|\nabla v\|_{L^2(\Omega)}^2$$

La norme étant positive on obtient la relation $\|\nabla v\|_{L^2(\Omega)} \leq \|v\|_V$, $\forall v \in V$ Ainsi,

$$\forall (u, v) \in V \times V,\ |a(u, v)| \leq \|\nabla u\|_{L^2(\Omega)}\|\nabla v\|_{L^2(\Omega)} \leq \|u\|_V\|v\|_V$$

La coercivité de $a(.,)$ est assurée par l'inégalité de Poincaré (applicable içi car $\Omega$ ouvert borné de $\mathbb{R}^N$, voir les détails dans le paragraphe 2.2.5) : $\exists M > 0/\forall v \in V,\ \|v\|_{L^2(\Omega)}^2 \leq M\|\nabla v\|_{L^2(\Omega)}^2$. Ainsi $\forall v \in V,\ \|v\|_V^2 = \|v\|_{L^2(\Omega)}^2 + \|\nabla v\|_{L^2(\Omega)}^2 \leq (M + 1)\|\nabla v\|_{L^2(\Omega)}^2$, et donc :

$$\forall v \in V,\ a(v, v) = \|\nabla v\|_{L^2(\Omega)}^2 \geq \frac{1}{M + 1}\, \|v\|_V^2$$

La forme $\ell(.)$ est linéaire, et sa continuité par rapport à $V$ est démontrable grâce à l'inégalité de Cauchy-Schwartz :

$$\forall v \in V, |\ell(v)| \leq \|f\|_{L^2(\Omega)}\, \|v\|_{L^2(\Omega)} \leq \|f\|_{L^2(\Omega)}\, \|v\|_V$$

En conclusion, comme $V$ est un espace de Hilbert et que toutes les hypothèses du théorème de Lax-Milgram sont respectées, la formulation variationnelle (2.8) admet une unique solution $u^\star$.

Si de plus la forme $a(.,.)$ est symétrique, alors l'unique solution $u^\star$ du problème variationnelle est caractérisé par :

$$J(u^\star) = \min_{v \in V} J(v)$$

c'est-à-dire que $u^\star$ minimise l'énergie, définie par la fonctionnelle $J$ définie par :

$$\forall v \in V,\ J(v) = \frac{1}{2} a(v,v) - \ell(v)$$

Ce résultat fournit alors un outil permettant de calculer $u^\star$ de façon itérative.

**Retour au problème initial**

On vient de montrer qu'il existe une unique solution $u^\star \in V$ au problème variationnel (2.8). Mais $u^\star$ est-elle aussi la solution de (2.3) ?
Ce problème peut s'avérer complexe si les conditions de régularité de $u$ et de $\Omega$ ne sont pas assez fortes. Nous ne regardons que le cas le plus simple. Supposons que $u^\star \in H^2(\Omega) = \{u \in L^2(\Omega), \nabla u \in L^2(\Omega) \text{ et } \nabla\nabla u \in L^2(\Omega)\}$. Par intégrations par parties et en utilisant la formule de Green,

$$\forall v \in V,\ \int_\Omega \nabla u^\star . \nabla v \, d\Omega = -\int_\Omega \Delta u^\star v \, d\Omega \Rightarrow \int_\Omega (\Delta u^\star + f) v \, d\Omega = 0$$

Ainsi $-\Delta u^\star = f$ dans $L^2(\Omega)$ et

$$-\Delta u^\star = f \text{ dans } \Omega \, \mu pp$$

Enfin, si on suppose que $\Omega$ est un ouvert borné régulier de classe $\mathcal{C}^1$, le théorème de la trace assure que toute fonction de V a une trace sur $\partial\Omega$ nulle dans $L^2(\Omega)$. Par conséquent,

$$u^\star = 0 \text{ sur } \partial\Omega \, \mu pp$$

### 2.2.4 Conditions aux limites de Neumann

Soient $\Omega$ un ouvert borné connexe régulier de classe $\mathcal{C}^1$ de $\mathbb{R}^N$ de frontière $\partial\Omega$, $f \in L^2(\Omega)$ et $g \in L^2(\partial\Omega)$ et $\vec{n}$ la normale sortante à $\partial\Omega$ telle que $\|\vec{n}\| = 1$. Considérons le problème aux limites de Neumann suivant :

$$\begin{cases} -\Delta u = f & \text{dans} \quad \Omega \\ \nabla u.n = g_n & \text{sur} \quad \partial\Omega \end{cases} \tag{2.9}$$

La principale difficulté rencontrée est que pour qu'il n'existe une solution au problème (2.9) que si les fonctions $f$ et $g$ vérifient une condition de compatibilité. Pour cela intégrons (2.9) sur $\Omega$ :

$$\int_\Omega f(x) \, d\Omega = -\int_\Omega \Delta u(x) \, d\Omega = -\int_{\partial\Omega} \nabla u(x).n \, d\partial\Omega = -\int_{\partial\Omega} g(x) \, d\partial\Omega$$

La condition de compatibilité s'écrit alors :

$$\int_\Omega f(x) \, d\Omega + \int_{\partial\Omega} g(x) \, d\partial\Omega = 0 \tag{2.10}$$

La condition (2.10) peut s'interpréter comme une relation d'équilibre entre une force volumique $f$ et $g$ un flux entrant sur la surface. C'est une condition nécessaire et suffisante d'existence d'un état d'équilibre du système à une constante près. En effet, si $u$ est solution, alors $u + \alpha$ est aussi solution, $\alpha \in \mathbb{R}$. Ceci peut s'interpréter comme l'absence d'origine de référence sur l'échelle qui mesure les valeurs de $u$ (par exemple la température).

**Formulation variationnelle**

Afin d'obtenir la formulation variationnelle, multiplions (2.9) par un champ $v$ cinématiquement admissible, que l'on démontrera a posteriori :

$$\int_\Omega \nabla u(x).\nabla v(x)\, d\Omega = \int_\Omega f(x)v(x)\, d\Omega + \int_{\partial\Omega} g(x)v(x)\, d\partial\Omega \tag{2.11}$$

On pourrait penser que choisir l'espace $H^1(\Omega)$ comme Hilbert serait un choix raisonnable. Cependant il serait impossible de démontrer la coercivité de la forme bilinéaire, notamment à cause de l'existence d'une solution à une constante près. Pour ce faire, introduisons l'espace V des fonctions de $H^1(\Omega)$ à valeur moyenne nulle, i.e :

$$V = \{v \in H^1(\Omega),\ \int_\Omega v(x)d\Omega = 0\}$$

La formulation variationnelle s'écrit donc :

$$\text{Trouver } u \in V \text{ tel que } \forall v \in V,\ \int_\Omega \nabla u(x).\nabla v(x)\, d\Omega = \int_\Omega f(x)v(x)\, d\Omega + \int_{\partial\Omega} g(x)v(x)\, d\partial\Omega \tag{2.12}$$

et par identification avec la forme faible générale (2.5) :

$$a(u,v) = \int_\Omega \nabla u(x).\nabla v(x)\, d\Omega \qquad ; \qquad \ell(v) = \int_\Omega f(x)v(x)\, d\Omega + \int_{\partial\Omega} g(x)v(x)\, d\partial\Omega$$

**Existence et unicité de la solution**

La seule difficulté par rapport aux conditions aux limites Dirichlet est de démontrer la coercivité de la forme bilinéaire. En effet, l'inégalité de Poincaré ne s'applique plus, et doit être généralisée par l'inégalité de Poincaré-Wirtinger (voir les détails au paragraphe 2.2.5), qui stipule que si $\Omega$ un ouvert borné connexe, il existe une constante $M > 0$ telle que :

$$\forall v \in H^1(\Omega), \|v - m(v)\|_{L^2(\Omega)} \leq M\|\nabla v\|_{L^2(\Omega)} \text{ où } m(v) = \frac{\int_\Omega v(x)\, d\Omega}{\int_\Omega d\Omega}$$

L'espace $V$ décrivant les fonctions de $H^1(\Omega)$ étant de moyenne nulle, on retombe sur la démonstration de la coercivité dans le cas des conditions aux limites de Dirichlet.

**Retour au problème initial**

A partir de la formulation variationnelle (2.12), utilisons la formule de Green, applicable pour $u \in H^2(\Omega)$ et $v \in H^1(\Omega)$, afin de remonter au laplacien :

$$\forall v \in V,\ \int_\Omega (\Delta u + f)\, v\, d\Omega = \int_{\partial\Omega} (\nabla u.n - g_n)\, v\, d\partial\Omega \tag{2.13}$$

Cependant, pour toute fonction $w \in H^1(\Omega)$, la fonction $v = w - m(w) \in V$. En injectant cette fonction $v$ dans (2.13), on obtient :

$$\begin{aligned}\forall w \in H^1(\Omega),\ \int_\Omega (\Delta u + f)\, w\, d\Omega - \int_{\partial\Omega} (\nabla u.n - g_n)\, w\, d\partial\Omega = m(w) \Big( \Big[ \int_\Omega \Delta u\, d\Omega - \int_{\partial\Omega} \nabla u.n d\partial\Omega \Big] \\ + \Big[ \int_\Omega f\, d\Omega + \int_{\partial\Omega} g_n\, d\partial\Omega \Big] \Big)\end{aligned} \tag{2.14}$$

Dans le terme de droite, le premier crochet s'annule par intégration directe, et le deuxième grâce à la condition de compatibilité (2.10). Ainsi (2.14) se réduit en :

$$\forall w \in H^1(\Omega),\ \int_\Omega (\Delta u + f)\, w\, d\Omega - \int_{\partial\Omega} (\nabla u.n - g_n)\, w\, d\partial\Omega \tag{2.15}$$

Finalement, $u^\star \in V$ est l'unique solution du problème variationnel (2.12). De plus , $u^\star \in H^2(\Omega)$ est solution de est bien solution de (2.9), dans le sens où :

$$\begin{cases} -\Delta u^\star = f & \text{dans} \quad \Omega\, \mu pp \\ \nabla u^\star.n = g & \text{sur} \quad \partial\Omega\, \mu pp \end{cases}$$

### 2.2.5 Démonstration des inégalités de Poincaré

**Inégalité de Poincaré dans $H_0^1(\Omega)$**

**Théorème 2.2** (Inégalité de Poincaré)**.**
*Soit $\Omega \in \mathbb{R}^N$ un ouvert borné dans au moins une direction de l'espace. Il existe alors une constante $M > 0$ telle que, pour toute fonction $v \in H_0^1(\Omega)$ :*

$$\|v\|_{L^2(\Omega)} \leq M \|\nabla v\|_{L^2(\Omega)}$$

**Preuve.** *Dans un premier temps, montrons que cette inégalité est vrai pour des fonctions de $\mathcal{C}_c^\infty(\Omega)$. Sans perte de généralité, considérons que pour $x \in \Omega$, sa première composante $x_1$ est bornée, c'est-à-dire que $-\infty < a \leq x_1 \leq b < \infty$. Soit $v$ une fonction de $\mathcal{C}_c^\infty(\Omega)$ qui s'annule sur $\partial\Omega$. Ainsi,*

$$\forall x \in \Omega,\ v(x) = \int_a^{x_1} \partial_1 v(t, x_2, \ldots, x_N) dt$$

*Grâce à l'inégalité de Cauchy-Schwarz,*

$$|v(x)|^2 \leq \int_a^{x_1} |\partial_1 v(t, x_2, \ldots, x_N)|^2 \, dt \int_a^{x_1} 1 \, dt \leq (b-a) \int_a^b |\partial_1 v(t, x_2, \ldots, x_N)|^2 \, dt$$

*Par intégration sur $\Omega$, on obtient l'inégalité recherchée :*

$$\int_\Omega |v(x)|^2 \, d\Omega \leq (b-a) \int_\Omega \int_a^b |\partial_1 v(t, x_2, \ldots, x_N)|^2 \, dt \, d\Omega \leq (b-a)^2 \int_\Omega |\partial_1 v(x)|^2 \, d\Omega \leq (b-a)^2 \int_\Omega |\nabla v(x)|^2 \, d\Omega$$

*Par exemple si on considère $(v_n)_{n\in\mathbb{N}}$ une suite de $\in \mathcal{C}_c^\infty(\Omega)$, l'inégalité précédente est vérifiée :*

$$\exists M > 0, \quad \int_\Omega |v_n(x)|^2 \, d\Omega \leq M \int_\Omega |\nabla v_n(x)|^2 \, d\Omega$$

*Ensuite, par densité de $\mathcal{C}_c^\infty(\Omega)$ dans $H_0^1(\Omega)$ :*

$$\lim_{n\to\infty} \|v - v_n\|_{H_0^1(\Omega)}^2 = \lim_{n\to\infty} \int_\Omega \left( |v - v_n|^2 + |\nabla v - \nabla v_n|^2 \right) d\Omega = 0$$

*Par conséquent*

$$\lim_{n\to\infty} \int_\Omega |v_n|^2 \, d\Omega = \int_\Omega |v|^2 \, d\Omega \,;\ \lim_{n\to\infty} \int_\Omega |\nabla v_n|^2 \, d\Omega = \int_\Omega |\nabla v|^2 \, d\Omega$$

*Finalement, pour $v \in H_0^1(\Omega)$ :*

$$\exists M > 0, \quad \|v(x)\|_{L^2(\Omega)}^2 \leq M \|\nabla v(x)\|_{L^2(\Omega)}^2$$

**Inégalité de Poincaré dans $H^1(\Omega)$**

Plaçons nous à présent dans $H^1(\Omega)$. La démonstration précédente de fonctionne plus, car la condition de nullité au bord $\partial\Omega$ n'existe plus. Par conséquent, les fonctions constantes non nulles annulent le terme de droite dans l'inégalité mais pas celui de gauche. On va donc démontrer une généralisation de l'inégalité de Poincaré, appelée inégalité de Poincaré-Wirtinger :

**Théorème 2.3** (Inégalité de Poincaré-Wirtinger)**.**

$$\exists M > 0,\ \forall v \in H^1(\Omega),\ \|v - m(v)\|_{L^2(\Omega)} \leq C \|\nabla v\|_{L^2(\Omega)} \ \text{où}\ m(v) = \frac{\int_\Omega v(x) \, d\Omega}{\int_\Omega d\Omega}$$

Avant de la démontrer, il nous faut énoncer le théorème de Rellich qui nous servira dans la démonstration :

**Théorème 2.4** (Théorème de Rellich)**.**
*Si $\Omega$ ouvert borné régulier de classe $\mathcal{C}^1$, alors de toute suite bornée de $H^1(\Omega)$ on peut extraire une sous-suite convergente dans $L^2(\Omega)$.*

**Preuve.** *La démonstration de l'inégalité de Poincaré-Wirtinger se fait par contradiction. Supposons cette inégalité fausse. Ainsi,*

$$\forall n \geq 1,\ \exists u_n \in H^1(\Omega), \|u_n - m(u_n)\|_{L^2(\Omega)} > n \|\nabla u_n\|_{L^2(\Omega)}$$

*Posons $v_n = (u_n - m(u_n)) / \|u_n - m(u_n)\|_{L^2(\Omega)}$. Par conséquent,*

$$1 = \|v_n\|_{L^2(\Omega)} > n \|\nabla v_n\|_{L^2(\Omega)}$$

*Ainsi la suite $(v_n)_{n\geq 1}$ est bornée. De plus $\Omega$ étant un ouvert borné régulier, le théorème de Rellich nous permet d'extraire une sous-suite $\tilde{v}_n \in L^2(\Omega)$ convergente vers $v \in L^2(\Omega)$. Premièrement, comme $(\tilde{v}_n)_{n\geq 1}$ est convergente dans $L^2(\Omega)$, c'est une suite de Cauchy de $L^2(\Omega)$. Deuxièmement, $\lim_{n\to\infty} \nabla \tilde{v}_n = 0$ donc $(\nabla \tilde{v}_n)_{n\geq 1}$ est convergente dans $L^2(\Omega)$ et c'est aussi une suite de Cauchy de $L^2(\Omega)$. Par conséquent, $(\tilde{v}_n)_{n\geq 1}$ est une suite de Cauchy dans $H^1(\Omega)$, Hilbert donc complet. Ainsi, toute suite de Cauchy est convergente et $(\tilde{v}_n)_{n\geq 1}$ converge vers $v \in H^1(\Omega)$. De plus,*

$$\begin{cases} \|\nabla v\|_{L^2(\Omega)} = \lim\limits_{n\to\infty} \|\nabla \tilde{v}_n\|_{L^2(\Omega)} \leq \dfrac{1}{n} = 0 \\ m(v) = \lim\limits_{n\to\infty} m(\tilde{v}_n) = 0 \\ \|v\|_{L^2(\Omega)} = \lim\limits_{n\to\infty} \|\nabla \tilde{v}_n\|_{L^2(\Omega)} = 1 \end{cases}$$

*Puisque $\nabla v = 0$, $m(v) = 0$ et $\Omega$ est connexe, $v$ est une constante de moyenne nulle donc $v = 0$. Or $\|v_n\|_{L^2(\Omega)} = 1$ ce qui est absurde.*

## 2.3 Schéma aux différences finies

### 2.3.1 En dimension 1

#### Problème de Dirichlet

Considérons le problème aux limites 1D suivant, qui pourrait modéliser une poutre sur deux appuis simples soumise à une densité linéique de forces $f(x)$, dont on chercherait à décrire le déplacement transversal $u(x)$ :

$$\begin{cases} -u''(x) = \mathcal{L}(u)(x) = f(x) \quad \text{dans } \Omega = ]0,1[ \\ u(0) = u(1) = 0 \end{cases} \tag{2.16}$$

où $\mathcal{L}$ désigne l'opérateur différentiel elliptique linéaire du laplacien en 1D. L'approximation de (2.16) est obtenue grâce aux développement de Taylor suivants :

$$\forall x \in \Omega,\ \forall h \text{ tel que } [x-h, x+h] \in \bar{\Omega} = [0,1], \quad u(x \pm h) = u(x) \pm h\, u'(x) + \frac{h^2}{2} u''(x) + o(h^2) \tag{2.17}$$

Le schéma aux différences finies du problème (2.16) correspond donc à la discrétisation du Laplacien : on remplace une dérivée, i.e la limite d'un taux de variation, par un taux de variation dont le dénominateur est non nul. Concrètement, $h$ est le **pas de discrétisation** (non obligatoirement constant) qui découpe

$\bar{\Omega} = \Omega \bigcup \partial\Omega$ en $N+2$ points. Les fonctions continues $x \mapsto f(x)$ et $x \mapsto u(x)$ sont prises aux valeurs discrètes $x_n = n\,h$, de telle sorte que :

$$\forall n \in [\![0; N+1]\!], \qquad f(x_n) = f_n \quad \text{et} \quad u(x_n) = u_n$$

Par conséquent :

**Théorème 2.5.**
*Le schéma aux différences finie du problème* (2.16) *s'écrit :*

$$\frac{-u_{n-1} + 2\,u_n - u_{n+1}}{h^2} = f_n \quad \forall n \in [\![1; N]\!], \text{ avec } u_0 = u_{N+1} = 0 \tag{2.18}$$

*Ce schéma faisant intervenir $u_{n-1}$, $u_n$ et $u_{n+1}$, il est dit "à trois points".*

Le problème aux différences finies (2.18) peut s'écrire sous forme matricielle, de telle sorte que :

$$\underline{\underline{A}} \in \mathcal{M}_N(\mathbb{R}),\ \underline{f} \in \mathbb{R}^N, \text{on cherche } \underline{u} \in \mathbb{R}^N \text{ tel que } \quad \underline{\underline{A}}.\underline{u} = \underline{f} \text{ où } \underline{\underline{A}} = \frac{1}{h^2}\begin{bmatrix} 2 & -1 & 0 & \dots & 0 \\ -1 & 2 & -1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & -1 & 2 & -1 \\ 0 & \dots & 0 & -1 & 2 \end{bmatrix} \tag{2.19}$$

**Proposition 2.1.** *La matrice $\underline{\underline{A}} \in \mathcal{M}_N(\mathbb{R})$ possède les propriétés suivantes :*

1. *tridiagonale*
2. *symétrique définie positive (à valeurs réelles), donc diagonalisable*
3. *monotone (voir définition 7.10)*
4. $\left\|\left|\underline{\underline{A}}^{-1}\right|\right\|_\infty \leq \frac{1}{8}$

**Preuve.** *1. évident*

2. *Démontrons que $\underline{\underline{A}}$ est définie-positive :*

$$\forall \underline{v} \in \mathbb{R}^N,\ {}^t\underline{v}.\underline{\underline{A}}.\underline{v} = \sum_{i,j=1}^{N} v_i A_{ij} v_j = v_1(2v_1 - v_2) + \sum_{i=2}^{N-1} v_i\left(-v_{i-1} + 2v_i - v_{i+1}\right) + v_N(-v_{N-1} + 2v_N)$$
$$= 2\left(\sum_{i=1}^{N} v_i^2 - \sum_{i=1}^{N-1} v_i v_{i+1}\right) = v_1^2 + \sum_{i=1}^{N-1}\left(v_i - v_{i+1}\right)^2 + v_N^2 \geq 0$$

   *Si de plus ${}^t\underline{v}.\underline{\underline{A}}.\underline{v} = 0$, alors $v_1 = v_N = 0$ et par conséquent $\underline{v} = \underline{0}$. Par conséquent, $\underline{\underline{A}}$ est une matrice à coefficients réels symétrique (évident) définie positive, donc diagonalisable.*

3. *La preuve de la monotonie de $\underline{\underline{A}}$ est laissée à titre d'exercice en utilisant la proposition 7.4.*

4. *Remarquons que $\sum_{j=1}^{N}\left(\underline{\underline{A}}^{-1}\right)_{ij} = \sum_{j=1}^{N}\left(\underline{\underline{A}}^{-1}\right)_{ij}\delta_j$ avec $\delta_j = 1\,\forall 1 \leq j \leq N$. Cette situation correspond à une force $f \equiv 1$. Ainsi*

$$\underline{\underline{A}}.\underline{u} = \underline{\delta} \Rightarrow \underline{u} = \underline{\underline{A}}^{-1}.\underline{\delta}$$

*Donc* $\forall i \in [\![1;N]\!]$, $\sum_{j=1}^{N}\left(\underline{\underline{A}}^{-1}\right)_{ij} = u(x_i)$. *et* $\sum_{j=1}^{N}\left(\underline{\underline{A}}^{-1}\right)_{ij} \leq \sup_{x\in[0,1]}|u(x)|$ *La solution* $u$ *à ce problème où* $f \equiv 1$ *est directement calculable, i.e* $\forall x \in [0,1]$, $u(x) = \frac{x}{2}(1-x)$ *et* $\sup_{x\in[0,1]}|u(x)| = \frac{1}{8}$. *Donc*

$$\forall i \in [\![1;N]\!],\ \sum_{j=1}^{N}\left(\underline{\underline{A}}^{-1}\right)_{ij} \leq \frac{1}{8}$$

*et par conséquent* $\left|\left|\left|\underline{\underline{A}}^{-1}\right|\right|\right|_\infty \leq \frac{1}{8}$.

**Consistance et précision**

**Définition 2.7.**
On appelle $\mathcal{L}_h$ l'opérateur aux différences finies associé à $\mathcal{L}$.

**Exemple 2.2.** Dans notre cas,

$$\mathcal{L}_h(u)(x) = \frac{-u(x-h)+2u(x)-u(x+h)}{h^2}$$

**Définition 2.8.**
L'erreur de troncature (ou de consistance) du schéma aux différences finies est défini, pour $u$ solution de (2.16) par :

$$\varepsilon_h(x) = \left(\mathcal{L}_h(u) - \mathcal{L}(u)\right)(x)$$

Le schéma est dit consistant avec (2.16) si $\lim_{h\to 0}\varepsilon_h(x) = 0$. De plus, il est précis d'ordre $p \in \mathbb{N}^\star$ si

$$\varepsilon_h(x) = O(h^p)$$

**Exemple 2.3.** Poussons plus loin le développement de Taylor (2.17), en supposant $u \in \mathcal{C}^4(\Omega)$ :

$$u(x \pm h) = u(x) \pm h\,u'(x) + \frac{h^2}{2}u''(x) \pm \frac{h^3}{6}u^{(3)}(x) + \frac{h^4}{24}u^{(4)}(x) + o(h^4)$$

On a donc

$$\mathcal{L}_h(u)(x) = -u''(x) - \frac{h^2}{12}u^{(4)}(x) + o(h^2)$$

L'erreur de consistance s'écrit finalement :

$$\varepsilon_h(x) = -\frac{h^2}{12}u^{(4)}(x) + o(h^2) = O(h^2)$$

Comme $\lim_{h\to 0}\varepsilon_h(x) = 0$, le schéma aux différences finies (2.18) est donc bien consistant avec (2.16). Le développement de Taylor précédent montre qu'il est précis d'ordre 2.

**Définition 2.9.**
Soit la norme discrète du maximum, telle que $\|\underline{u}\|_\infty = \max_{1\leq n\leq N}|u_n|$. On notera par extension $\|u\|_{h,\infty} = \max_{1\leq n\leq N}|u(x_n)|$.

**Proposition 2.2.** *Si* $u \in \mathcal{C}^4(\Omega)$, *alors* $\|\varepsilon_h\|_{h,\infty} \leq \frac{h^2}{12}\|u^{(4)}\|_{h,\infty}$.

**Théorème 2.6** (Principe du maximum continu)**.**
*Soit le problème* (2.16) *(ce serait également vrai en dimension supérieure). Si* $f \geq 0\,\mu.p.px \in \Omega$*, alors* $u \geq 0\,\mu.p.px \in \Omega$*.*

**Remarque 2.6.**
Considérant l'équation de la chaleur en régime stationnaire avec des conditions de Dirichlet homogènes, le principe du maximum traduit le fait que si on chauffe le domaine $\Omega$ avec une source $f \geq 0$, alors la température $u$ dans le domaine sera toujours plus élevée que celle aux bords $u \geq 0$.

**Théorème 2.7** (Principe du maximum discret)**.**
*Si* $\forall n \in [\![1; N]\!]$, $f_n \geq 0$*, alors* $\forall n \in [\![0; N+1]\!]$, $u_n \geq 0$*.*

**Proposition 2.3.** *Le schéma aux différences finies* (2.18) *respecte le principe du maximum discret.*

**Preuve.** *Soit* $j \in [\![0; N+1]\!] / u_j = \min_{n \in [\![0;N+1]\!]} u_n$ *(s'il y a plusieurs minima on prend le plus petit* $j$*).*

- *Si* $j = 0$*, alors* $\forall n \in [\![1; N+1]\!]$, $u_n \geq u_j = 0$
- *Si* $j \geq 1$*, alors* $u_0 = 0 > u_j$ *et comme* $u_{N+1} = 0$*, alors* $j \neq N+1$ *et donc* $j \leq N$*. Ecrivons le laplacien discrétisé en* $1 \leq j \leq N$*, où par hypothèse* $h^2 f_j \geq 0$ *:*

$$-u_{j-1} + 2u_j - u_{j+1} = h^2 f_j \geq 0$$

*ce qui est impossible pour* $u_j$ *minimum :* $u_j - \frac{u_{j-1}+u_{j+1}}{2} < 0$

*Au final, seul le cas* $j = 0$ *est possible, et le schéma* (2.18) *respecte bien le principe du maximum discret.*

Un autre démonstration se base sur la monotonie de $\underline{\underline{A}}$ :

**Preuve.** *Comme* $\underline{\underline{A}}$ *est monotone, alors on a la propriété :*

$$\forall \underline{X} \in \mathbb{R}^N,\ \underline{\underline{A}}.\underline{X} \geq 0 \Rightarrow \underline{X} \geq 0$$

*Pour* $\underline{X} = \underline{u}$*,* $\underline{\underline{A}}.\underline{u} = \underline{f} \geq 0$ *par hypothèse. Donc le principe du maximum est vérifié*

**Stabilité** Notons $u_h$ la solution du problème discrétisé, vérifiant donc $\mathcal{L}_h(u_h) = f$ dans $\Omega_h$ l'espace discrétisé, et $u_h(0) = u_h(1) = 0$. Par conséquent, dans $\Omega_h$ on a $\mathcal{L}_h(u_h) - \mathcal{L}(u) = f - f = 0$. Or on a défini $\varepsilon_h(u) = (\mathcal{L}_h(u) - \mathcal{L}(u))$ l'erreur de consistance. Alors on peut redéfinir cette erreur de la façon suivante :

$$0 = \mathcal{L}_h(u_h) - \mathcal{L}(u) = \mathcal{L}_h(u_h) - \mathcal{L}_h(u) + \mathcal{L}_h(u) - \mathcal{L}(u) = \mathcal{L}_h(u_h - u) + \varepsilon_h(u) \Rightarrow \varepsilon_h(u) = \mathcal{L}_h(u - u_h)$$

On a vu dans la Proposition 2.2 que $\|\varepsilon_h\|_{h,\infty} \leq \frac{h^2}{12}\|u^{(4)}\|_{h,\infty}$. La notion de stabilité peut se poser de la façon suivante : $u - u_h$ est petit si $\varepsilon_h|_{h \ll 1}$ est petit ?

**Définition 2.10** (Stabilité selon une norme)**.**
On dit que le schéma aux différences finies est stable selon la norme $\|.\|$ si pour toute source $f$, la solution discrétisée est majorée par $\|f\|$ :

$$\exists C > 0 / \ \|u_h\| \leq C\|f\|$$

La notion de stabilité dépend donc de la norme, et un schéma peut être stable selon une norme et pas selon une autre.

**Proposition 2.4.** *Le schéma aux différences finies* (2.18) *est stable selon la norme discrète du maximum (on dit abusivement stable $L^\infty$).*

**Preuve.**

$$\|u_h\|_{h,\infty} = \|\underline{u}\|_\infty = \|\underline{\underline{A}}^{-1}.\underline{f}\|_\infty \leq \left\|\left|\underline{\underline{A}}^{-1}\right|\right\|_\infty \|\underline{f}\|_\infty \leq \frac{1}{8}\|\underline{f}\|_{h,\infty}$$

**Convergence du schéma.**

**Définition 2.11.**
Un schéma aux différences finies est convergent d'ordre $p \in \mathbb{N}^\star$ (ordre de l'erreur de consistance) selon la norme $\|.\|$ si

$$\exists C > 0 / \|u - u_h\| \leq Ch^p$$

**Théorème 2.8.**
*Le schéma aux différences finies* (2.18) *convergent $L^\infty$ d'ordre 2 :*

$$\|u - u_h\|_{h,\infty} \leq \frac{h^2}{96}\|u^{(4)}\|_{h,\infty}$$

**Preuve.** *Notons le vecteur erreur $\underline{e} \in \mathbb{R}^N$, tel que $e_n = u_n - u(x_n)$. On a alors*

$$\forall n \in [\![1; N]\!],\ \left(\underline{\underline{A}}.\underline{e}\right)_n = \varepsilon_h(x_n) \Rightarrow \underline{\varepsilon_h} = \underline{\underline{A}}.\underline{e}$$

*$\underline{\underline{A}}$ étant inversible, on peut estimer la norme de l'erreur $\underline{e}$ :*

$$\|\underline{e}\|_\infty = \|\underline{\underline{A}}^{-1}.\underline{\varepsilon_h}\|_\infty \leq \left\|\left|\underline{\underline{A}}^{-1}\right|\right\|_\infty \|\underline{\varepsilon_h}\|_\infty \leq \frac{1}{8}\|\underline{\varepsilon_h}\|_\infty \leq \frac{h^2}{96}\|u^{(4)}\|_{h,\infty}$$

**Exercice 2.2 (**Problèmes 1D**).** Soit $\Omega = ]0, \ell[$ un ouvert de $\mathbb{R}$. On considère sur $\Omega$ les problèmes aux limites suivants :

$$(\mathcal{P})\begin{cases} -u''(x) = f(x) & \text{dans } \Omega \\ u(0) = u_0 \\ u(\ell) = u_\ell \end{cases} \quad ; \quad (\mathcal{P}')\begin{cases} -u''(x) = f(x) & \text{dans } \Omega \\ u'(0) = \theta_0 \\ u(\ell) = u_\ell \end{cases}$$

avec $f$ un terme source tel que $f(x) = F_0 x(x - \ell) + F_1$. Ecrire le système linéaire associé à $(\mathcal{P})$ et $(\mathcal{P}')$, avec un pas de discrétisation $h$ constant.

### 2.3.2 En dimension 2

Considérons à présent le problème 2D suivant, qui pourrait modéliser le déplacement $u$ d'une membrane élastique soumise à une force verticale de densité surfacique $f$ :

$$\begin{cases} -\Delta u(x,y) = f(x,y) & \text{dans } \Omega = ]0,1[\times]0,1[ \\ u = 0 & \text{sur } \partial\Omega \end{cases} \tag{2.20}$$

La procédure est la même que pour le cas 1D. On approxime l'ensemble $\bar{\Omega} = \Omega \cup \partial\Omega$ par un ensemble fini résultant d'un maillage du domaine, avec des pas de discrétisation $h_x = \frac{1}{N_x+1}$ et $h_y = \frac{1}{N_y+1}$ qui découpent $\Omega$ en $N_x \times N_y$ points. Les fonctions continues $(x,y) \mapsto u(x,y)$ et $(x,y) \mapsto f(x,y)$ sont discrétisées aux points $x_i = i\,h_x$, $y_j = j\,h_y$.

**Théorème 2.9.**
*Le schéma aux différences finie du problème* (2.20) *s'écrit :*

$$\forall 1 \le i \le N_x,\ 1 \le j \le N_y, \begin{cases} \dfrac{-u_{i-1,j} + 2\,u_{i,j} - u_{i+1,j}}{h_x^2} + \dfrac{-u_{i,j-1} + 2\,u_{i,j} - u_{i,j+1}}{h_y^2} = f_{i,j} \\ u_{0,j} = u_{N_x+1,j} = u_{i,0} = u_{i,N_y+1} = 0 \end{cases} \tag{2.21}$$

*Ce schéma est consistant avec l'EDP et est précis à l'ordre 2 selon chaque pas de discrétisation :* $\varepsilon_{h_x,h_y} = O(h_x^2 + h_y^2)$.

**Remarque 2.7.**
Ce schéma est dit "à cinq points". Il est courant d'identifier chacun des quatre points voisin par le point cardinal correspondant : nord $(N)$, sud $(S)$, est $(E)$ et ouest $(O)$. Le point central est alors le point $(P)$ (voir Figure 2.4)

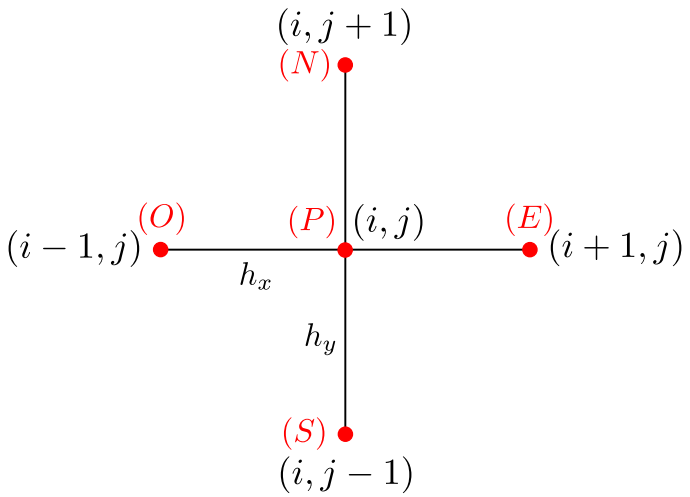



FIGURE 2.2 – Schéma à cinq points.

**Preuve.** *La preuve se fait par développements de Taylor. Soient* $(x,y) \in \Omega$ *et* $(h_x,h_y)$ *tels que* $[x-h_x, x+h_x] \times [y-h_y, y+h_y] \in \bar{\Omega}$. *Alors,* $\exists(\theta_x,\theta_y) \in ]-1,1[\times]-1,1[$ *tels que :*

$$\begin{cases} -\partial_1^2 u(x,y) = \dfrac{-u(x-h_x,y) + 2u(x,y) - u(x+h_x,y)}{h_x^2} + \dfrac{h_x^2}{12}\partial_1^4 u(x+\theta_x\,h_x,y) \\ -\partial_2^2 u(x,y) = \dfrac{-u(x,y-h_y) + 2u(x,y) - u(x,y+h_y)}{h_y^2} + \dfrac{h_y^2}{12}\partial_2^4 u(x,y+\theta_y\,h_y) \end{cases}$$

*Ainsi le développement de Taylor du laplacien est :*

$$\begin{aligned}-\Delta u(x,y) =& \frac{-u(x-h_x,y)+2u(x,y)-u(x+h_x,y)}{h_x^2} + \frac{-u(x,y-h_y)+2u(x,y)-u(x,y+h_y)}{h_y^2} \\ &+ \frac{h_x^2}{12}\partial_1^4 u(x+\theta_x\, h_x, y) + \frac{h_y^2}{12}\partial_2^4 u(x, y+\theta_y\, h_y) = f(x,y)\end{aligned}$$

**Remarque 2.8.**
Une telle discrétisation ne prend en compte que des maillages "rectangulaires" de longueur $(\ell_x, \ell_y)$. La relation entre le pas de discrétisation $h$ (lié au nombre $n$ de points sur le segment unité, tel que $n+1 = 1/h$) et le nombre de points $N$ dans le volume intérieur $\Omega$ dans une des deux directions est : $d = \{x, y\}\,, \quad N_d = \ell_d\,(n_d+1) - 1$. Il y a donc $N_x \times N_y = (\ell_x\,(n_x+1)-1) \times (\ell_y\,(n_y+1)-1)$ points de discrétisation dans $\Omega$ et $2\,(N_x+N_y)+4$ points de discrétisation sur $\partial\Omega$.

Prenons l'exemple qu'un maillage pour lequel $h_x = h_y = h = \frac{1}{N+1}$, par conséquent dans notre cas on donc l'égalité $N = n$. Ainsi, le schéma à cinq point s'écrit :

$$\begin{cases} \dfrac{-u_{i-1,j} - u_{i,j-1} + 4\,u_{i,j} - u_{i+1,j} - u_{i,j+1}}{h^2} = f_{i,j} & \forall 1 \le i,j \le N \\ u_{0,j} = u_{N+1,j} = u_{i,0} = u_{i,N+1} = 0 & \forall 0 \le i,j \le N+1 \end{cases} \tag{2.22}$$

Le schéma (2.22) peut s'écrire sous forme matricielle, telle que :

$$\underline{\underline{B}} \in \mathcal{M}_{N^2}(\mathbb{R}),\ \underline{F} \in \mathbb{R}^{N^2}, \text{on cherche } \underline{U} \in \mathbb{R}^{N^2} \text{ tel que } \quad \underline{\underline{B}}.\underline{U} = \underline{F}$$

avec

$$\underline{\underline{B}} = \frac{1}{h^2}\begin{bmatrix} \underline{\underline{T}}_N & -\underline{\underline{I}}_N & \underline{\underline{0}}_N & \cdots & \underline{\underline{0}}_N \\ -\underline{\underline{I}}_N & \underline{\underline{T}}_N & -\underline{\underline{I}}_N & \ddots & \vdots \\ \underline{\underline{0}}_N & \ddots & \ddots & \ddots & \underline{\underline{0}}_N \\ \vdots & \ddots & -\underline{\underline{I}}_N & \underline{\underline{T}}_N & -\underline{\underline{I}}_N \\ \underline{\underline{0}}_N & \cdots & \underline{\underline{0}}_N & -\underline{\underline{I}}_N & \underline{\underline{T}}_N \end{bmatrix}$$

où $\underline{\underline{I}}_N$ et $\underline{\underline{0}}_N$ sont respectivement les matrices identité et nulle de dimension $N$, et $\underline{\underline{T}}_N$ une matrice tridiagonale, telle que :

$$\underline{\underline{T}}_N = 2\underline{\underline{I}}_N + h^2\underline{\underline{A}} = \begin{bmatrix} 4 & -1 & 0 & \dots & 0 \\ -1 & 4 & -1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & -1 & 4 & -1 \\ 0 & \dots & 0 & -1 & 4 \end{bmatrix}$$

**Théorème 2.10.**
*La matrice $\underline{\underline{B}}$ est tridiagonale par blocs, symétrique, définie-positive et monotone.*

**Preuve.** *A titre d'exercice.*

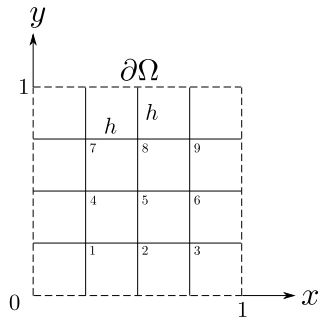
On pourrait également montrer le théorème suivant (plus difficile que dans le cas 1D) :

**Théorème 2.11.**
*Si $u \in \mathcal{C}^4(\Omega)$, une estimation de l'erreur est donnée par :*

$$\|\underline{e}\|_\infty \leq \alpha\, h^2\,(M_3 + h\, M_4) \quad \text{où} \quad \begin{cases} \alpha > 0 \text{ constante indépendante de } u \text{ et de } h \\ I = \{3,4\}, \quad M_I = \max\left( \sup\limits_{(x,u)\in\bar{\Omega}} \left|\partial_1^I u(x,u)\right|, \sup\limits_{(x,u)\in\bar{\Omega}} \left|\partial_2^I u(x,u)\right| \right) \end{cases}$$

**Exercice 2.3 (**2D : problème de Dirichlet**).** Soit $\Omega = ]0,1[\times]0,1[$ un ouvert de $\mathbb{R}^2$ de frontière $\partial\Omega = \bar{\Omega}/\Omega$. On considère sur $\Omega$ le problème aux limites suivant : Ecrire le système linéaire associé à $(\mathcal{P})$ si on choisit le

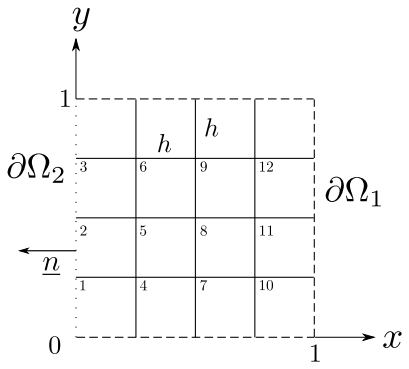
$$(\mathcal{P}) \begin{cases} -\Delta u = f & \text{dans } \Omega \text{ où } f \in L^2(\Omega) \\ u = 0 & \text{sur } \partial\Omega \end{cases}$$




même maillage horizontal et vertical $h = 1/4$.

**Exercice 2.4 (**2D : problème mêlé**).** Soit $\Omega = ]0,1[\times]0,1[$ un ouvert de $\mathbb{R}^2$ de frontière $\partial\Omega = \partial\Omega_1 \bigcup \partial\Omega_2$ où $\partial\Omega_2 = \{x = 0, 0 < y < 1\}$ de normale sortante $\underline{n} = (-1, 0)$. On considère sur $\Omega$ le problème aux limites suivant ($\alpha \in \mathbb{R}$) :

$$(\mathcal{P}) \begin{cases} -\Delta u = f & \text{dans } \Omega \text{ où } f \in L^2(\Omega) \\ u = 0 & \text{sur } \partial\Omega_1 \\ \underline{\nabla} u.\underline{n} + \alpha\, u = 0 & \text{sur } \partial\Omega_2 \end{cases}$$




Ecrire le système linéaire associé à $(\mathcal{P})$ si on choisit le même maillage horizontal et vertical $h = 1/4$.

## 2.4 Exercices supplémentaires

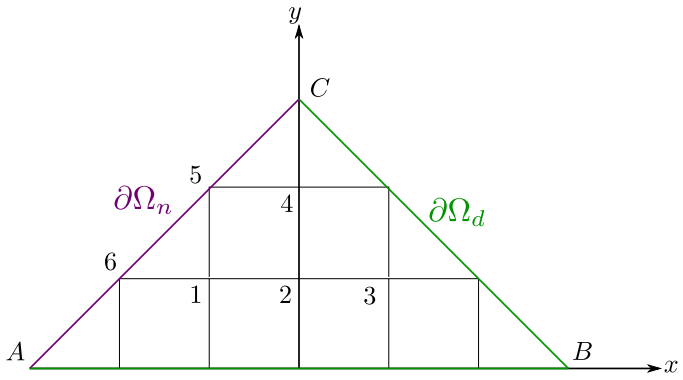
### 2.4.1 Exercices avancés

**Exercice 2.5.** Soient $B$ la boule unité ouverte de $\mathbb{R}^2$ et la fonction $u(x) = |\ln(|x|)|^\alpha$, $\forall x \in B$. Montrer que $u \in H^1(B)$ pour $0 < \alpha < 1/2$ mais qu'elle n'est pas bornée au voisinage de l'origine.

**Exercice 2.6.** Soit $\underline{\underline{A}} \in \mathcal{M}_N(\mathbb{R})$, telle que : $\underline{\underline{A}} = \begin{bmatrix} 2 & -1 & 0 & \dots & 0 \\ -1 & 2 & -1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & -1 & 2 & -1 \\ 0 & \dots & 0 & -1 & 2 \end{bmatrix}$. Montrer que $\underline{\underline{A}}$ est monotone, i.e qu'elle est inversible et d'inverse positive.

### 2.4.2 TD

**Exercice 2.7 (TD).** Dans le repère orthonormé $\mathcal{R} = (0, x, y)$, on appelle $\Omega$ l'ensemble délimité par le triangle $ABC$ avec $A(-1, 0)$, $B(1, 0)$ et $C(0, 1)$.




On pose :

$$\partial\Omega_d = [AB] \cup [BC] \qquad \text{et} \qquad \partial\Omega_n = ]AC[$$

et on considère le problème mélé suivant :

$$(\mathcal{P}) \begin{cases} -\Delta u = f & \text{dans } \Omega \\ u = g_d & \text{sur } \partial\Omega_d \\ \dfrac{\partial u}{\partial \nu} = g_n & \text{sur } \partial\Omega_n \end{cases}$$

On applique enfin sur $\Omega$ un maillage de pas horizontal $h = \frac{1}{3}$ dans les deux directions.

1. Ecrire le système linéaire associé à ce système.
2. Vérifier que $u(x, y) = (1 - x - y)\, y$ est solution du problème lorsque $g_d = 0$. Évaluer alors $f$ et $g_n$.
3. Résoudre le problème linéaire par la méthode de Gauss avec les $f$ et $g_n$ calculés précédemment.
4. Comparer les résultats exacts et approchés.

### 2.4.3 Annales

**Exercice 2.8 (Rattrapage 2019).** Soient $\mathcal{R} = (0, x, y)$ un repère orthonormé et les points A=(0,0) B=(2,0) C=(2,1) D=(1,1) E=(1,2) et F=(0,2). On pose $\Omega$ l'ouvert formé par l'intérieur du polygone $ABCDEFA$ de frontière $\partial\Omega$, voir figure 2.5.

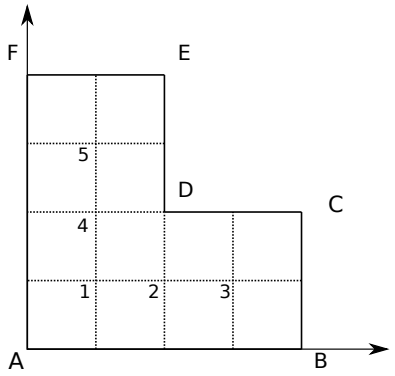



FIGURE 2.3 – Discrétisation du domaine d'étude avec un pas $h = 1/2$.

On souhaite résoudre le problème de Dirichlet 2D $(\mathcal{P})$ suivant :

$$(\mathcal{P}) \begin{cases} -\Delta u(x,y) + \alpha\, u(x,y) = f(x,y) & \forall x \in \Omega \\ u|_{\partial\Omega} = 0 \end{cases}$$

où $(x,y) \mapsto f(x,y) = 1 + x^2 + y^2$ et $\alpha \in \mathbb{R}_+^\star$

**Partie 1 : Formulation variationnelle**

1. On pose $V = H_0^1(\Omega) = \{v \in L^2(\Omega), \nabla v \in L^2(\Omega), v_{|\partial\Omega} = 0\}$ muni du produit scalaire

   $$\forall (u,v) \in V \times V,\ \langle u, v\rangle_V = \langle u, v\rangle_{L^2(\Omega)} + \langle \nabla u, \nabla v\rangle_{L^2(\Omega)}$$

   Etablir la formulation variationnelle $(\mathcal{P}_v)$ associée à $(\mathcal{P})$ sous la forme :

   $$\text{"Trouver } u \in V, \forall v \in V,\ a(u,v) = \ell(v)\text{"}$$

2. Soit $\mathcal{G}(v,w) = \displaystyle\int_\Omega (\nabla v \nabla w + \alpha v w)\, d\Omega$. Montrer rigoureusement qu'il existe deux réels $M > 0$ et $N > 0$ tel que

   $$\forall (v,w) \in V \times V,\ |\mathcal{G}(v,w)| \leq M \|v\|_V \|w\|_V$$

   et

   $$\forall w \in V,\ \mathcal{G}(w,w) \geq N \|w\|_V^2$$

3. Soit $\mathcal{H}(w) = \displaystyle\int_\Omega f w d\Omega$. Montrer rigoureusement que $f \in L^2(\Omega)$ puis qu'il existe un réel $\gamma > 0$ tel que

   $$\forall w \in V,\ |\mathcal{H}(w)| \leq \gamma \|w\|_V$$

4. En déduire l'existence et l'unicité de la solution au problème $(\mathcal{P}_v)$

**Partie 2 : Discrétisation par différences finies**

On souhaite résoudre le problème $(\mathcal{P})$ par la méthode des différences finies en discrétisant $\bar{\Omega} = \Omega \bigcup \partial\Omega$ avec un pas constant $h_x = h_y = h$.

1. Etablir le schéma à cinq points classique de discrétisation du Laplacien pour les points dans $\Omega$.
2. Dans un premier temps posons $h = 1/2$. On décide arbitrairement de numéroter les points ligne par ligne de gauche à droite en commençant par celle du bas (voir figure 2.5). Ecrire le système linéaire sous forme matricielle résultant de la méthode des différences finies. Pour alléger les notations on gardera les expressions formelles, i.e on ne cherchera pas à calculer les différentes valeurs numériques.
3. Montrer que ce système peut être résolu par la méthode du gradient conjugué. Combien d'itérations au maximum sont nécessaires à la résolution ? Rappeler le principe de la méthode.

**Partie 3 : Raffinement de la discrétisation**

On décide maintenant d'affiner le maillage en discrétisant $\Omega$ avec un pas $h_x = h_y = h = 1/3$.

1. En suivant les règles de numérotation de la partie 2, dessiner le maillage du modèle.
2. Ecrire le système linéaire sous forme matricielle résultant de cette discrétisation, en mettant en avant les blocs dans la matrice.

**Exercice 2.9 (**Annale 2019**).** On souhaite résoudre le problème d'élasticité 1D $(\mathcal{P})$ suivant par la méthode des différences finies :

$$(\mathcal{P}) \begin{cases} -\dfrac{d}{dx}\left(k(x)\,\dfrac{d}{dx}\left(u(x)\right)\right) = 1 \quad \forall x \in \Omega = ]0,1[ \\ u(0) = u(1) = 0 \end{cases}$$

où $\forall x \in \Omega$, $k(x) = 1 + x^2$.

1. Existence et unicité de la solution au problème $(\mathcal{P})$.
   (a) Montrer que la formulation variationnelle $(\mathcal{P}_v)$ associée au problème différentiel $(\mathcal{P})$ s'écrit :

   $$(\mathcal{P}_v) \text{ "Trouver } u \in V \text{ (à déterminer) tel que } \forall v \in V,\ a(u,v) = \ell(v)\text{" où } \begin{cases} a(u,v) = \displaystyle\int_\Omega k(x)\,u'(x)\,v'(x)\,dx \\ \ell(v) = \displaystyle\int_\Omega v(x)\,dx \end{cases}$$

   (b) Montrer qu'il existe une unique solution au problème $(\mathcal{P}_v)$. On rappellera l'inégalité de Poincaré dans $H_0^1(\Omega)$ :

   $$\exists M > 0 \text{ tel que, } \forall v \in H_0^1(\Omega), \quad \|v\|_{L^2(\Omega)} \leq M \|v'\|_{L^2(\Omega)}$$

   Sous quelle(s) condition(s) une solution de $(\mathcal{P}_v)$ est une solution de $(\mathcal{P})$ ?
2. Afin de résoudre $(\mathcal{P})$ par la méthode des différences finies, considérons le schéma suivant :

   $$-\frac{1}{h}\left[k\left(\left(i+\frac{1}{2}\right)h\right)\frac{u_{i+1}-u_i}{h} - k\left(\left(i-\frac{1}{2}\right)h\right)\frac{u_i - u_{i-1}}{h}\right] = 1$$

   (a) Interpréter qualitativement ce schéma au regard du problème $(\mathcal{P}_v)$.
   (b) En prenant un pas $h = 0.2$, écrire le système linéaire à résoudre.
   (c) Sans calcul supplémentaire, quelle(s) méthode(s) de résolution préconisez-vous ?

**Exercice 2.10 (**Annale 2020**).** Soit $(\mathcal{P})$ le problème aux limites dans $\Omega = ]0, L[$, où $L > 0$ et $(a,b) \in \mathbb{R}^2$ :

$$(\mathcal{P}) \begin{cases} -u''(x) = 1 \quad \forall x \in \Omega \\ u(0) = a \\ u(L) = b \end{cases}$$

1. Déterminer le changement de variable tel que le problème $(\mathcal{P})$ piloté par $u$ devienne le problème $(\tilde{\mathcal{P}})$ piloté par $\tilde{u}$ :

   $$(\tilde{\mathcal{P}}) \begin{cases} -\tilde{u}''(x) = 1 \quad \forall x \in \Omega \\ \tilde{u}(0) = \tilde{u}(L) = 0 \end{cases}$$

2. Ecrire la formulation variationnelle $(\mathcal{P}_v)$ associée à $(\tilde{\mathcal{P}})$. Montrer que $(\mathcal{P}_v)$ admet une unique solution.
3. Résoudre analytiquement $(\tilde{\mathcal{P}})$ et en déduire la solution de $(\mathcal{P})$.
4. On veut résoudre $(\tilde{\mathcal{P}})$ par la méthode des différences finies avec un pas $h = L/6$. Ecrire le système linéaire $\underline{\underline{A}}.\underline{u} = \underline{f}$ associé.
5. Décrire la structure de $\underline{\underline{A}}$. Quelle méthode de résolution préconisez-vous ? [Ne pas résoudre] Justifier votre réponse.

# Chapitre 3

# Résolution par la méthode des différences finies de problèmes aux limites non stationnaires.

*Ce chapitre traite de la résolution de problèmes aux limites dépendant du temps. L'exemple qui sert de fil rouge est celui de l'équation de la chaleur unidimensionnel avec des conditions de Dirichlet homogènes, mais la méthodologie serait la même pour toute EDP parabolique ou elliptique. La première partie permet d'établir des résolutions analytiques qui serviront de référence. Une première discrétisation (uniquement en espace) mène à un problème de Cauchy, que l'on peut résoudre grâce aux méthodes du chapitre 1. Ensuite une discrétisation totale en espace et en temps est proposée. On verra en particulier que certains schémas ne seront effectivement utilisables que sous une condition (appelée CFL) reliant le pas de temps et le pas d'espace.*

**Sommaire**

## 3.1 Résolution analytique de l'équation de la chaleur unidimensionnelle

### 3.1.1 Position du problème

Nous allons étudier le problème de de diffusion (équation de la chaleur) unidimensionnel en temps $t > 0$ dans un domaine ouvert $\Omega \in \mathbb{R}$ :

$$\rho(t,x)c(t,x)\partial_1 u(t,x) = \partial_2 \left(K(t,x)\partial_2(u(t,x))\right) + S(t,x)$$

où $\rho, c, K$ représentent respectivement la masse spécifique, la chaleur spécifique et la conductivité thermique du matériau, et $S$ est une densité de source de chaleur. Le problème aux limites comporte de plus une condition initiale :

$$u(t=0,x) = u^0(x) \qquad \forall x \in \Omega$$

et des conditions aux limites, qui peuvent être pour $(t,x) \in \mathbb{R}_\star^+ \times \partial\Omega$ du type :

**Dirichlet** $u(t,x)$ donné : par exemple un corps plongé dans de la glace fondue

**Neumann** $\partial_2 u(t,x)$ donné. Par exemple $\partial_2 u(t,x) = 0$ correspond à une isolation

**Fourier** $\partial_2 u(t,x) + bu(x,t)$ donné. Par exemple échange thermique avec l'extérieur (loi de convection) : $K\partial_2 u(t,x) = \alpha(u_{\text{ext}} - u)$

**Périodique** $u(t,x+X) = u(t,x)$ si $X$ est la longueur du domaine

Pour simplifier, on considèrera les coefficients mécaniques constants en espace et en temps, ce qui mènera à l'EDP :

$$\partial_1 u(t,x) - \kappa\partial_2^2 u(t,x) = s(t,x) \qquad \forall (t,x) \in \mathbb{R}_\star^+ \times \Omega \tag{3.1}$$

où $\kappa = \frac{K}{\rho c}$ et $s(x,t) = \frac{S(x,t)}{\rho c}$. A titre d'exemple, nous allons résoudre analytiquement deux problèmes.

### 3.1.2 Résolution dans $\Omega = \mathbb{R}$

Soit le problème de Cauchy :

$$\begin{cases} \partial_1 u(t,x) - \kappa\partial_2^2 u(t,x) = 0 & \forall (t,x) \in \mathbb{R}_\star^+ \times \mathbb{R} \\ u(0,x) = u^0(x) & \forall x \in \Omega \end{cases} \tag{3.2}$$

On suppose les hypothèses suivantes : les fonctions $x \mapsto u^0(x)$, $x \mapsto u(t,x)$, $x \mapsto \partial_1 u(t,x)$, $x \mapsto \partial_2 u(t,x)$ et $x \mapsto \partial_2^2 u(t,x)$ sont $L^1(\mathbb{R})$. La transformée de Fourier de $u$ selon sa variable spatiale s'écrit alors :

$$\forall k \in \mathbb{R}, \forall t > 0,\ U(t,k) = \mathcal{F}_2(u)(t,k) = \int_{\mathbb{R}} u(t,x)e^{-I2\pi kx}\,dx$$

Grâce au théorème de dérivation d'une fonction définie par une intégrale, on montre facilement que $\forall k \in \mathbb{R}$, la fonction $t \mapsto \mathcal{F}_2(u)(k,t)$ est dérivable et que $\partial_1 U(t,k) = \mathcal{F}_2(\partial_1 u)(t,k)$. Par conséquent, l'application de la transformée de Fourier à l'EDP (3.2) mène à l'EDO en temps :

$$\forall (t,k) \in \mathbb{R}_\star^+ \times \mathbb{R},\ \partial_1 U(t,k) + \kappa 4\pi^2 k^2 U(t,k) = 0$$

La résolution de cette EDO conduit à :

$$\forall (t,k) \in \mathbb{R}_\star^+ \times \mathbb{R},\ U(t,k) = U_0(k)e^{-\kappa(2\pi k)^2 t}$$

où $U_0 = \mathcal{F}_2(u^0)$ est la transformée de Fourier de la condition initiale. La transformée de Fourier d'une gaussienne ($a > 0$) étant

$$\forall k \in \mathbb{R},\ \mathcal{F}(x \mapsto e^{-ax^2})(k) = \sqrt{\frac{\pi}{a}}e^{-\frac{\pi^2 k^2}{a}}$$

donc

$$e^{-\kappa(2\pi k)^2 t} = \mathcal{F}_2((t,x) \mapsto \frac{1}{2\sqrt{\kappa\pi t}}e^{-\frac{x^2}{4\kappa t}})(k,t) = \mathcal{F}_2(G)(k,t)$$

on trouve au final la solution de (3.2) comme un produit de convolution entre $G$ le noyau de l'opérateur de diffusion et $u^0$ la condition initiale :

$$\forall (t,x) \in \mathbb{R}^+ \times \mathbb{R}, \quad u(t,x) = \frac{1}{2\sqrt{\kappa \pi t}} \int_{\mathbb{R}} u^0(x-\xi) e^{-\frac{\xi^2}{4\kappa t}} \, d\xi = \left(u^0 \star G\right)(t,x) \tag{3.3}$$

**Remarque 3.1.**

— Si $u^0$ est continue et bornée, alors $u \in \mathcal{C}^\infty(\mathbb{R} \times \mathbb{R}^+)$ (effet régularisant)
— Pour tout $t > 0$, $u$ dépend de toutes les valeurs de $u^0$ : il y a vitesse de propagation infinie.

**Exemple 3.1.** Si $u^0(x) = e^{-x^2}$, alors (3.3) devient :

$$\forall (t,x) \in \mathbb{R}^+ \times \mathbb{R}, \quad u(t,x) = \frac{1}{\sqrt{1+4\kappa t}} e^{-\frac{x^2}{1+4\kappa t}} \tag{3.4}$$

Une représentation graphique de l'évolution de $u$ au cours du temps est proposée sur la Figure 3.1 pour $\kappa = 1$.

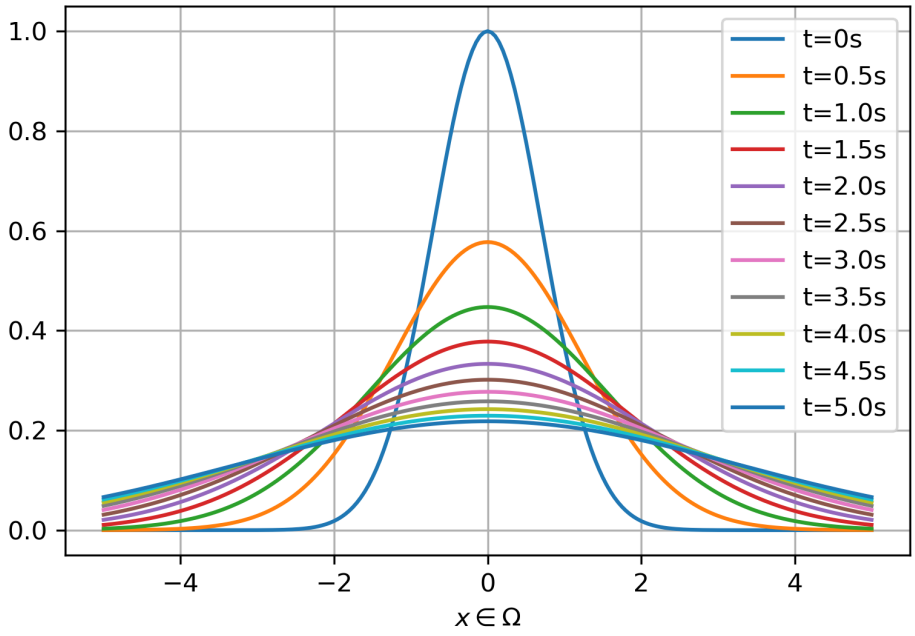



FIGURE 3.1 – Solution (3.4) exacte de (3.2) pour $\kappa = 1$

### 3.1.3 Résolution dans $\Omega = ]0, L[$

Soit le problème aux limites modélisant par exemple la diffusion de chaleur dans un mur d'épaisseur $L$

$$\begin{cases} \partial_1 u(t,x) - \kappa \partial_2^2 u(t,x) = 0 & \forall (t,x) \in \mathbb{R}_\star^+ \times \Omega \\ u(0,x) = u^0(x) & \forall x \in \Omega \\ u(t,x) = 0 & \forall t > 0, \forall x \in \partial\Omega \end{cases} \tag{3.5}$$

On va résoudre ce problème par la méthode de séparation de variables. On cherche la solution non nulle de (3.5) sous la forme :

$$u(t,x) = \phi(t)\psi(x) \neq 0$$

De plus, les conditions aux limites deviennent, pour tout $t > 0$ :

$$\begin{cases} u(t,0) = \phi(t)\psi(0) = 0 \Rightarrow \psi(0) = 0 \\ u(t,L) = \phi(t)\psi(L) = 0 \Rightarrow \psi(L) = 0 \end{cases} \tag{3.6}$$

En injectant cette forme dans l'EDP, on obtient :

$$\kappa \frac{\psi''(x)}{\psi(x)} = \frac{\phi'(t)}{\phi(t)} = -\lambda \in \mathbb{R}$$

ce qui mène à deux EDO :

$$\psi''(x) + \frac{\lambda}{\kappa}\psi(x) = 0 \tag{3.7a}$$

$$\phi'(t) + \lambda\phi(t) = 0 \tag{3.7b}$$

Ecrivons alors l'énergie liée à (3.7a) :

$$\frac{\lambda}{\kappa}\int_\Omega \psi(x)^2\,dx = -\int_\Omega \psi''(x)\psi(x)\,dx = -[\psi'(x)\psi(x)]_0^L + \int_\Omega (\psi'(x))^2\,dx$$

En appliquant les conditions aux limites (3.6) on obtient :

$$\frac{\lambda}{\kappa}\|\psi\|^2_{L^2(\Omega)} = \|\psi'\|^2_{L^2(\Omega)}$$

ce qui impose que $\lambda \geq 0$.

— Si $\lambda = 0$, alors $\psi(x) = a_0 x + b_0$ et en appliquant (3.6) on trouve $\psi(x) = 0\,\forall x \in \bar{\Omega}$

— Si $\lambda > 0$, alors $\psi(x) = a\cos(\sqrt{\frac{\lambda}{\kappa}}x) + b\sin(\sqrt{\frac{\lambda}{\kappa}}x)$. En appliquant (3.6) :

$$\psi(0) = a \quad ; \quad \psi(L) = b\sin(\sqrt{\frac{\lambda}{\kappa}}L) = 0$$

On trouve ainsi les valeurs et vecteurs propres du problème :

$$\forall q \in \mathbb{N}^\star,\ \lambda_q = \kappa\left(\frac{q\pi}{L}\right)^2 \quad ; \quad \psi_q(x) = \sqrt{\frac{2}{L}}\sin(q\pi\frac{x}{L}) \text{ tels que } \langle\psi_p,\psi_q\rangle = \int_0^L \psi_p(x)\psi_q(x)\,dx = \delta_{pq}$$

Résolvons ensuite (3.7a) pour les valeurs propres : $\phi_q(t) = U_q e^{-\lambda_q t}$, de telle sorte que les solutions élémentaires vérifiant les conditions aux limites s'écrivent :

$$\forall q \in \mathbb{N}^\star,\ u_q(t,x) = U_q\sqrt{\frac{2}{L}}\sin(q\pi\frac{x}{L})e^{-\lambda_q t}$$

La solution générale de l'EDP linéaire s'écrit comme la somme des solutions élémentaires :

$$\forall (t,x) \in \mathbb{R}_\star^+ \times [0,L],\ u(t,x) = \sum_{q\geq 1} U_q\sqrt{\frac{2}{L}}\sin(q\pi\frac{x}{L})e^{-\lambda_q t}$$

Pour déterminer les constantes $\{U_q\}_{q\geq 1}$ on projète la condition initiale sur la base des vecteurs propres :

$$\forall p \geq 1,\ \langle u^0,\psi_p\rangle = \sum_{q\geq 1} U_q\langle\psi_q,\psi_p\rangle = U_p$$

Au final la solution s'écrit :

$$\forall (t,x) \in \mathbb{R}^+ \times [0,L],\ u(t,x) = \sqrt{\frac{2}{L}}\sum_{q\geq 1}\langle u^0,\psi_q\rangle\sin(q\pi\frac{x}{L})e^{-\kappa\left(\frac{q\pi}{L}\right)^2 t} \tag{3.8}$$

**Remarque 3.2.**
Bien qu'analytique, l'expression (3.8) n'est pas fermée, i.e que pour en faire une application il est nécessaire de tronquer la base modale $q \in [\![1; Q]\!]$ (choix du nombre d'harmoniques $Q$ à prendre en compte), ce qui amène une erreur de troncature à minimiser.

**Exemple 3.2.** Soit le problème (3.5) dans $\Omega = ]-\frac{L}{2}, \frac{L}{2}[$ avec $u^0(x) = e^{-x^2}$ et $\kappa = 1$. Sa solution par séparation de variables est donnée par (3.8) en translatant de $\frac{L}{2}$. Sur la Figure 3.2 sont représentées la solution exacte issue du problème de Cauchy (vrai tant qu'on reste loin des bords) (3.4), ainsi que les solutions pour un nombre $Q$ croissant de modes pris en compte dans la série de Fourier.

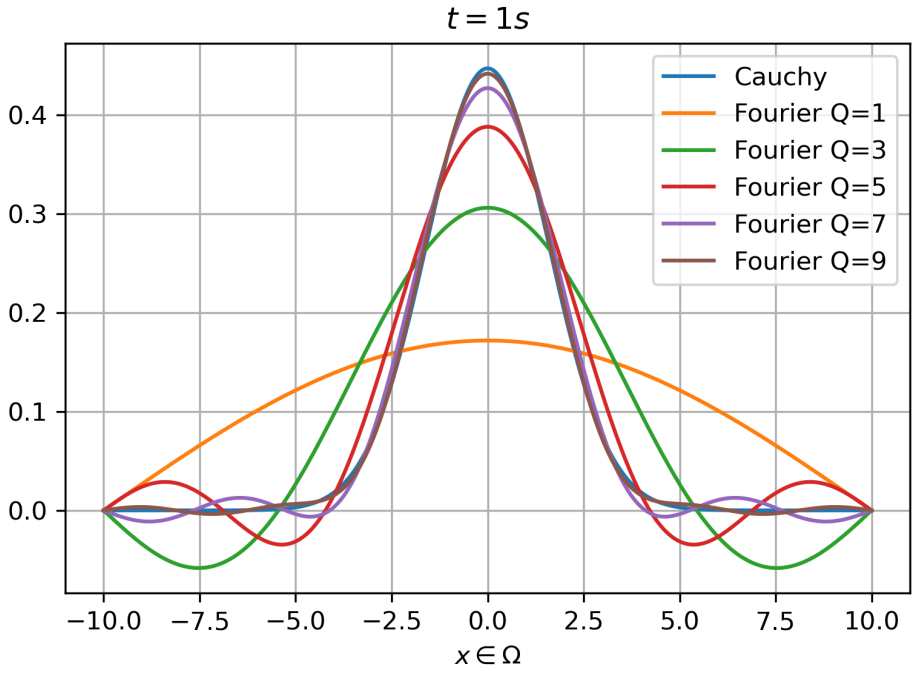



FIGURE 3.2 – Solutions de (3.5) pour $t = 1s$

Sur la Figure 3.3 sont représentées la solution issue du problème de Cauchy (qui n'est plus vraie car elle dépasse les bords !) (3.4), et la solution de variable ($Q = 10$ harmoniques), qui elle respecte les conditions aux limites.

### 3.1.4 Semi-discrétisation du problème

Dans un premier temps, nous proposons une discrétisation uniquement en espace (ou semi-discrétisation) de (3.5). La méthode des différences finies est appliquée en discrétisant de façon uniforme $\Omega$ avec un pas $\Delta x = \frac{L}{N+1}$. Ainsi

$$\forall j \in [\![0; N+1]\!],\ x_j = j\Delta x \qquad ; \qquad u(t, x_j) = u_j(t)$$

Par développement de Taylor on retrouve le schéma centré en espace pour l'approximation du laplacien

$$-\partial_2^2 u(t, x_j) \approx \frac{-u_{j-1}(t) + 2u_j(t) - u_{j+1}(t)}{\Delta x^2} \tag{3.9}$$

et la semi-discrétisation de (3.5) s'écrit alors :

$$\begin{cases} \underline{u}'(t) = -\underline{\underline{A}}.\underline{u}(t) = \underline{F}(t, \underline{u}(t)) \\ \underline{u}(0) = \underline{u}^0 \end{cases} \tag{3.10}$$

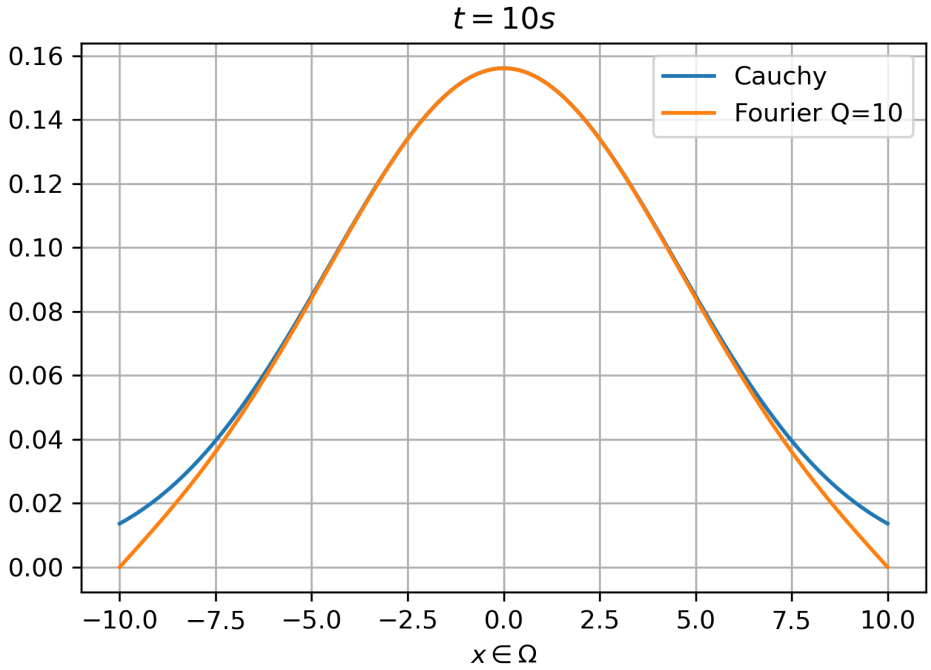



FIGURE 3.3 – Solutions de (3.5) pour $t = 10s$

Avec

$$\underline{u}(t) = \begin{pmatrix} u_1(t) \\ u_2(t) \\ \vdots \\ u_N(t) \end{pmatrix} ; \underline{u}^0 = \begin{pmatrix} u^0(x_1) \\ u^0(x_2) \\ \vdots \\ u^0(x_N) \end{pmatrix} ; \underline{\underline{A}} = \frac{\kappa}{\Delta x^2} \begin{bmatrix} 2 & -1 & 0 & \dots & \dots & \dots & \dots \\ -1 & 2 & -1 & 0 & \dots & \dots & \dots \\ 0 & -1 & 2 & -1 & 0 & \dots & \dots \\ \dots & \ddots & \ddots & \ddots & \ddots & \ddots & \dots \\ \dots & \dots & 0 & -1 & 2 & -1 & 0 \\ \dots & \dots & \dots & 0 & -1 & 2 & -1 \\ \dots & \dots & \dots & \dots & 0 & -1 & 2 \end{bmatrix}$$

Le système différentiel (3.1) (système de $N$ EDO d'ordre 1) est alors un problème de Cauchy que l'on peut résoudre numériquement avec l'un des schémas numériques vus au chapitre 1.

**Remarque 3.3.**

La matrice $\underline{\underline{A}} \in \mathcal{M}_N(\mathbb{R})$ étant symétrique, elle est diagonalisable. Si on note $\{\lambda_i, \underline{X}_i\}_{1 \leq i \leq N}$ les modes propres de $\underline{\underline{A}}$, tels que $\underline{\underline{A}}.\underline{X}_i = \lambda_i \underline{X}_i$ et $\langle \underline{X}_i, \underline{X}_j \rangle = \delta_{ij}$, alors une solution de (3.1) est

$$\underline{u}(t) = \sum_{i=1}^{N} \tilde{u}_i e^{-\lambda_i t} \underline{X}_i$$

Pour prendre en compte la condition initiale et ainsi déterminer les coefficients $(\tilde{u}_i)_{1 \leq i \leq N}$, on projète $\underline{u}^0$ sur la base modale, de telle sorte qu'au final :

$$\forall t \geq 0, \ \underline{u}(t) = \sum_{i=1}^{N} \langle \underline{u}^0, \underline{X}_i \rangle e^{-\lambda_i t} \underline{X}_i$$

**Exercice 3.1.** Résoudre numériquement le problème par la méthode de votre choix.

## 3.2 Discrétisation totale du problème

### 3.2.1 Différents schémas aux différences finies

Dans cette partie nous discrétisons le problème (3.5) en temps (pas $\Delta t$) et en espace (pas $\Delta x = \frac{L}{N+1}$). Ainsi :

$$\forall j \in [\![0; N+1]\!], \forall n \geq 0 \, (t_n, x_j) = (n\Delta t, j\Delta x) \qquad ; \qquad u(t_n, x_j) = u_j^n$$

Suivant la discrétisation en espace (3.9), le schéma centré en espace s'écrit :

$$-\partial_2^2 u(t_n, x_j) \approx \frac{-u_{j-1}^n + 2u_j^n - u_{j+1}^n}{\Delta x^2}$$

**Définition 3.1.**
Un schéma aux différences finies est dit à $N$ niveaux (ou $N-1$ pas de temps) si l'expression de $u_{n+1}$ dépend de $N-1$ termes précédents.

**Remarque 3.4.**
On note $\chi = \frac{\kappa \Delta t}{\Delta x^2}$ le coefficient CFL (Courant, Friedrichs et Lewy), qui relie le pas de temps spatial $\Delta x$ et le pas de temps temporel $\Delta t$. On verra pas la suite que la stabilité des schémas numériques peut être conditionnée par la valeur de ce coefficient. A noter que l'expression de ce coefficient dépend de l'EDP étudiée.

Discrétisons maintenant la dérivée temporelle. Commençons par écrire le développement de Taylor selon un pas $\Delta t$ :

$$u(t \pm \Delta t, x) = u(t, x) \pm \Delta t \partial_1 u(t, x) + \frac{\Delta t^2}{2} \partial_1^2 u(t, x) + o(\Delta t^2)$$

Il y a plusieurs choix de discrétisation de $\partial_1 u$ :

1. Schéma centré $\partial_1 u(t_n, x_j) \approx \frac{u_j^{n+1} - u_j^{n-1}}{2\Delta t}$, qui mène au schéma explicite à deux pas de temps (ou trois niveaux) dit **schéma de Richardson** (ou saute-mouton) :

$$u_j^{n+1} = u_j^{n-1} - 2\chi \left( -u_{j-1}^n + 2u_j^n - u_{j+1}^n \right) \tag{3.11}$$

   ou écrit vectoriellement :

$$\underline{u}^{n+1} = -2\chi \underline{\underline{C}}^{1} . \underline{u}^n + \underline{u}^{n-1} \quad \text{où } \underline{\underline{C}}^{1} = \begin{bmatrix} 2 & -1 & 0 & \dots & \dots & \dots & \dots \\ -1 & 2 & -1 & 0 & \dots & \dots & \dots \\ 0 & -1 & 2 & -1 & 0 & \dots & \dots \\ \dots & \ddots & \ddots & \ddots & \ddots & \ddots & \dots \\ \dots & \dots & 0 & -1 & 2 & -1 & 0 \\ \dots & \dots & \dots & 0 & -1 & 2 & -1 \\ \dots & \dots & \dots & \dots & 0 & -1 & 2 \end{bmatrix}$$

   Remarque : Il faut déterminer $(u_j^1)_j$ à l'aide d'un autre schéma.

2. Schéma décentré amont $\partial_1 u(t_n, x_j) \approx \frac{u_j^n - u_j^{n-1}}{\Delta t}$ mène au schéma implicite à deux niveaux dit **schéma d'Euler rétrograde** :

$$-\chi u_{j-1}^n + (1 + 2\chi) u_j^n - \chi u_{j+1}^n = u_j^{n-1} \tag{3.12}$$

   A chaque pas de temps il faut résoudre le système linéaire suivant :

$$\underline{\underline{C}}^{2} . \underline{u}^n = \underline{u}^{n-1} \quad \text{où } \underline{\underline{C}}^{2} = \begin{bmatrix} 1+2\chi & -\chi & 0 & \dots & \dots & \dots & \dots \\ -\chi & 1+2\chi & -\chi & 0 & \dots & \dots & \dots \\ 0 & -\chi & 1+2\chi & -\chi & 0 & \dots & \dots \\ \dots & \ddots & \ddots & \ddots & \ddots & \ddots & \dots \\ \dots & \dots & 0 & -\chi & 1+2\chi & -\chi & 0 \\ \dots & \dots & \dots & 0 & -\chi & 1+2\chi & -\chi \\ \dots & \dots & \dots & \dots & 0 & -\chi & 1+2\chi \end{bmatrix}$$

**Remarque 3.5.**

— A chaque pas de temps c'est toujours la même matrice à inverser.
— $\underline{\underline{\overset{2}{C}}}$ est inversible car symétrique définie positive :

$$\forall \underline{X} \in \mathbb{R}^N \neq \underline{0},\ {}^t\underline{X}.\underline{\underline{\overset{2}{C}}}.\underline{X} = \sum_{j=1}^{N} \frac{X_j^2 + X_{j+1}^2}{2} + \chi\left(X_{j+1} - X_j\right)^2 > 0$$

On peut aussi directement conclure que $\underline{\underline{\overset{2}{C}}}$ est inversible car à diagonale strictement dominante, voir le théorème 7.2.

3. Schéma décentré aval $\partial_1 u(t_n, x_j) \approx \frac{u_j^{n+1} - u_j^n}{\Delta t}$ mène au schéma explicite à deux niveaux dit **schéma d'Euler progressif** :

$$u_j^{n+1} = \chi u_{j-1}^n + (1 - 2\chi)\, u_j^n + \chi u_{j+1}^n \tag{3.13}$$

ou écrit vectoriellement :

$$\underline{u}^{n+1} = \underline{\underline{\overset{3}{C}}}.\underline{u}^n \quad \text{où } \underline{\underline{\overset{3}{C}}} = \begin{bmatrix} 1-2\chi & \chi & 0 & \dots & \dots & \dots & \dots \\ \chi & 1-2\chi & \chi & 0 & \dots & \dots & \dots \\ 0 & \chi & 1-2\chi & \chi & 0 & \dots & \dots \\ \dots & \ddots & \ddots & \ddots & \ddots & \ddots & \dots \\ \dots & \dots & 0 & \chi & 1-2\chi & \chi & 0 \\ \dots & \dots & \dots & 0 & \chi & 1-2\chi & \chi \\ \dots & \dots & \dots & \dots & 0 & \chi & 1-2\chi \end{bmatrix}$$

Il est également possible de créer un $\theta$-schéma par combinaison convexe ($0 \leq \theta \leq 1$) des schémas (3.12) (en passant de $n$ à $n+1$) et (3.13) :

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + \theta\kappa \frac{-u_{j-1}^{n+1} + 2u_j^{n+1} - u_{j+1}^{n+1}}{\Delta x^2} + (1-\theta)\kappa \frac{-u_{j-1}^n + 2u_j^n - u_{j+1}^n}{\Delta x^2} = 0 \tag{3.14}$$

Les cas suivants sont à remarquer :
— $\theta = 0$ correspond au schéma (3.13)
— $\theta = 1$ correspond au schéma (3.12)
— $\theta = \frac{1}{2}$ correspond au schéma de **Cranck-Nicolson**

Se pose alors la question du choix du schéma pour résoudre l'EDP. A l'instar de ce que nous avons fait dans le chapitre un, nous allons étudier les conditions de convergence en étudiant leur consistance et leur stabilité.

### 3.2.2 Consistance et précision

Notons $\mathcal{L}(u) = f$ l'EDP à résoudre, où $\mathcal{L} = \partial_1 - \kappa\partial_2^2$ est l'opérateur différentiel de l'équation de la chaleur, et $\mathcal{L}_{\Delta t, \Delta x}$ l'opérateur aux différences finies associé au problème approché, tel que par exemple pour le schéma d'Euler progressif :

$$\mathcal{L}_{\Delta t, \Delta x}(u)(t,x) = \frac{u(t+\Delta t, x) - u(t,x)}{\Delta t} - \kappa \frac{u(t, x-\Delta x) - 2u(t,x) + u(t, x+\Delta x)}{\Delta x^2}$$

**Définition 3.2.**
L'erreur de troncature (ou de consistance) du schéma aux différences finies est défini par

$$\varepsilon_{\Delta t, \Delta x}(t,x) = \left(\mathcal{L}_{\Delta t, \Delta x}(u) - \mathcal{L}(u)\right)(t,x)$$

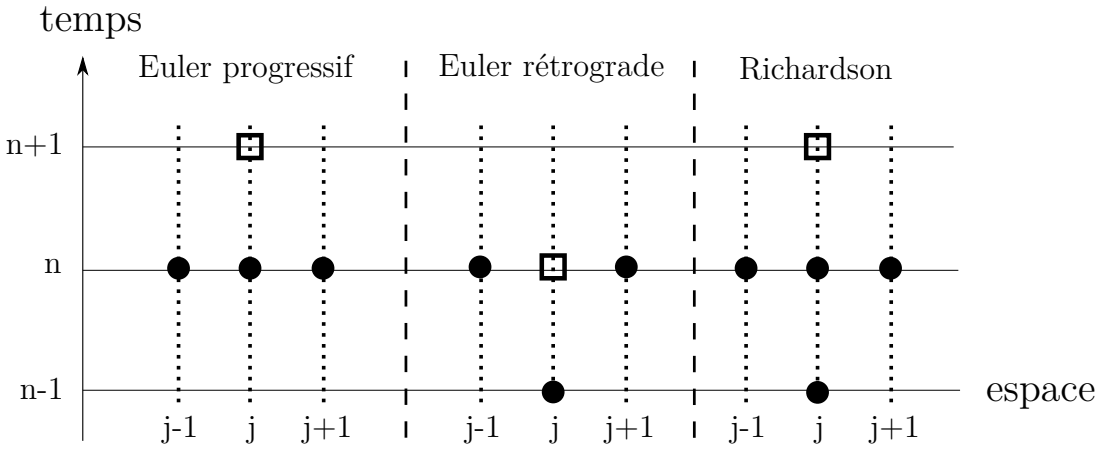



FIGURE 3.4 – Stencils de trois schémas. Les carrés représentent les points calculés, et les ronds les points nécessaires au calcul.

**Remarque 3.6.**

Si le terme source $f$ est nul, alors l'erreur de troncature se réduit à

$$\varepsilon_{\Delta t,\Delta x}(t,x) = \mathcal{L}_{\Delta t,\Delta x}(u)(t,x)$$

**Définition 3.3.**

Le schéma aux différences finies est dit consistant avec l'EDP si, pour toute solution $u$ de cette équation, l'erreur de troncature tend vers 0 uniformément par rapport à $(t,x)$ lorsque les pas de discrétisation tendent vers 0 indépendemment :

$$\lim_{\Delta t,\Delta x\to 0} \varepsilon_{\Delta t,\Delta x}(t,x) = 0$$

De plus, il est précis d'ordre $(p,q) \in \mathbb{N}^\star$ en espace et en temps si $\varepsilon_{\Delta t,\Delta x}(t,x) = O\left(\Delta x^p + \Delta t^q\right)$.

**Remarque 3.7.**

Afin de faciliter ce calcul, il est fortement recommandé d'utiliser l'expression de l'EDP pour ne transformer les dérivées en temps en dérivées en espace : $\partial_1 u(t,x) = \kappa \partial_2^2(t,x)$.

**Exemple 3.3 (**Consistance et précision du schéma Euler progressif**).** Commençons par le développement de Taylor en temps :

$$u(t+\Delta t,x) = u(t,x) + \Delta t \partial_1 u(t,x) + \frac{\Delta t^2}{2}\partial_1^2 u(t,x) + o(\Delta t^2)$$

par conséquent, et en utilisant l'expression de l'EDP dont $u$ est solution :

$$\frac{u(t+\Delta t,x)-u(t,x)}{\Delta t} = \partial_1 u(t,x) + \frac{\Delta t}{2}\partial_1^2 u(t,x) + o(\Delta t) = \kappa \partial_2^2 u(t,x) + \frac{\Delta t}{2}\kappa^2 \partial_2^4 u(t,x) + o(\Delta t)$$

De même en espace :

$$u(t,x\pm\Delta x) = u(t,x) \pm \Delta x \partial_2 u(t,x) \pm + \frac{\Delta x^2}{2}\partial_2^2 u(t,x) \pm \frac{\Delta x^3}{6}\partial_2^3 u(t,x) + \frac{\Delta x^4}{24}\partial_2^4 u(t,x) \pm \frac{\Delta x^5}{120}\partial_2^5 u(t,x) + o(\Delta x^5)$$

le schéma en espace donne donc :

$$\frac{u(t,x-\Delta x) - 2u(t,x) + u(t,x+\Delta x)}{\Delta x^2} = \partial_2^2 u(t,x) + \frac{\Delta x^2}{12}\partial_2^4 u(t,x) + o(\Delta x^3)$$

Au final, l'approximation de l'EDP par le schéma Euler progressif est :

$$\frac{u(t+\Delta t,x)-u(t,x)}{\Delta t}-\kappa\frac{u(t,x-\Delta x)-2u(t,x)+u(t,x+\Delta x)}{\Delta x^2}=\frac{\kappa}{2}\left(\kappa\Delta t-\frac{\Delta x^2}{6}\right)\partial_2^4u(t,x)+O\left(\Delta t^2+\Delta x^4\right)$$
$$=O\left(\Delta t+\Delta x^2\right)$$

Le schéma Euler progressif est donc
- consistant avec l'EDP
- précis à l'ordre 1 en temps et 2 en espace (et si $\chi=\frac{\kappa\Delta t}{\Delta x^2}=\frac{1}{6}$, alors il est précis à l'ordre 2 en temps et 4 en espace).

**Exercice 3.2.** Montrer que le schéma Euler rétrograde est précis à l'ordre 1 et temps et 2 en espace.

**Exercice 3.3.** Montrer que le $\theta$-schéma (3.14) est précis à l'ordre 1 en temps et 2 en espace si $\theta\neq 1/2$, et d'ordre 2 en temps et 2 en espace si $\theta=1/2$.

### 3.2.3 Stabilité

On définit les normes vectorielles classiques sur $\mathbb{R}^N$, mais pondérées par $\Delta x$ :

$$\begin{cases}\|\underline{u}^n\|_{p,\Delta x}=\left(\sum_{j=1}^N\Delta x|u_j^n|^p\right)^{\frac{1}{p}} & \text{pour } p\geq 1\\ \|\underline{u}^n\|_\infty=\max\limits_{1\leq j\leq N}|u_j^n|\end{cases}$$

La norme $\|.\|_{p,\Delta x}$ est une norme admissible, i.e $\lim\limits_{\Delta x\to 0}\|\underline{u}\|_{p,\Delta x}=\|u\|_p$ où $u\in L^p(\Omega)$. On pourra alors parler de normes $L^p(\Omega)$, car elles correspondent aux normes fonctionnelles pour des fonctions constantes par morceaux sur chaque intervalle de discrétisation spatiale.

**Définition 3.4.**
Un schéma aux différences finies est dit stable pour la norme $\|.\|$ si il existe une constante $K>0$ indépendante des pas de discrétisation telle que :

$$\forall\underline{u}\in\mathbb{R}^N,\forall n\geq 0,\ \|\underline{u}^n\|\leq K\|\underline{u}^0\| \tag{3.15}$$

**Remarque 3.8.**
Bien que toutes les normes soient équivalentes dans $\mathbb{R}^N$, la stabilité selon une norme n'implique pas la stabilité selon une autre norme.

**Proposition 3.1.** *Cas des schémas linéaires à deux niveaux. Dans ce cas, le schéma s'écrit*

$$\underline{u}^{n+1}=\underline{\underline{C}}.\underline{u}^n=(\underline{\underline{C}})^n.\underline{u}^0$$

*où $\underline{\underline{C}}$ est appelée matrice d'itération. Ainsi, la condition de stabilité* (3.15) *s'écrit :*

$$\|(\underline{\underline{C}})^n\,\underline{u}^0\|\leq K\|\underline{u}^0\|\Rightarrow\left|\left|\left|(\underline{\underline{C}})^n\right|\right|\right|\leq K$$

*où $|||.|||$ est la norme subordonnée à la norme vectorielle $\|.\|$.*

**Exemple 3.4.** Par exemple, pour le schéma d'Euler progressif, $\underline{\underline{C}} = \overset{3}{\underline{\underline{C}}}$ et pour le schéma d'Euler rétrograde, $\underline{\underline{C}} = (\overset{2}{\underline{\underline{C}}})^{-1}$.

**Proposition 3.2.** *Cas des schémas linéaires multiniveaux. Pour les schémas multiniveaux, $\underline{u}^{n+1}$ dépend de $\underline{u}^n$ mais également des termes précédents. Prenons l'exemple d'un schéma à trois niveaux. En posant*

$$\underline{U}^n = \begin{pmatrix} \underline{u}^n \\ \underline{u}^{n-1} \end{pmatrix}$$

*alors il existe deux matrices $(\underline{\underline{C}}_1, \underline{\underline{C}}_2) \in \mathcal{M}_N(\mathbb{R})$ telles que le schéma s'écrit*

$$\underline{U}^{n+1} = \underline{\underline{C}}.\underline{U}^n \qquad où \qquad \underline{\underline{C}} = \begin{bmatrix} \underline{\underline{C}}_1 & \underline{\underline{C}}_2 \\ \underline{\underline{I}} & \underline{\underline{0}} \end{bmatrix}$$

.

*Ainsi, la condition de stabilité s'écrit :*

$$\|(\underline{\underline{C}})^n \, \underline{u}^0\| \leq K\|\underline{u}^0\| \Rightarrow \left|\left|\left|(\underline{\underline{C}})^n\right|\right|\right| \leq K$$

*où $|||.|||$ est la norme subordonnée à la norme vectorielle $\|.\|$, voir définition 7.13.*

**Exemple 3.5.** Le schéma de Richardson (3.11) est un schéma à trois niveaux, tel que $\underline{\underline{C}}_1 = -2\chi\overset{1}{\underline{\underline{C}}}$; $\underline{\underline{C}}_2 = \underline{\underline{I}}$.

**Stabilité $L^\infty$**

**Définition 3.5 (**Principe du maximum discret**).**
Un schéma aux différences finies vérifie le principe du maximum discret si :

$$\forall \underline{u}^0 \in \mathbb{R}^{N+2} \, \forall n \geq 0, \forall j \in [\![1; N]\!], \min(0, \min_{0 \leq j \leq N+1} u_j^0) \leq u_j^n \leq \max(0, \max_{0 \leq j \leq N+1} u_j^0)$$

**Remarque 3.9.**
Les 0 présents dans le min et le max correspondent à la valeur des conditions aux limites de Dirichlet du problème traité (3.5). En effet rien n'oblige $\underline{u}^0$ à respecter cette condition.

**Théorème 3.1.**
*Si un schéma vérifie le principe du maximum discret, alors il est stable $L^\infty$.*

**Proposition 3.3.** *Le schéma d'Euler progressif* (3.13) *est stable $L^\infty$ ssi la condition CFL $\chi \leq \frac{1}{2}$ est vérifiée.*

**Preuve.** *Deux cas peuvent apparaître pour le schéma* (3.13) *:*

**Si** $1 - 2\chi \geq 0 \Leftrightarrow \chi \leq \frac{1}{2}$ *, alors par récurrence*

$$\forall \underline{u}^0 / m \leq u_j^0 \leq M \Rightarrow \forall n \geq 0, \, m \leq u_j^n \leq M$$

*Ce qui prouve que les oscillations de $\underline{u}^n$ restent bornées.*

**Si** $1-2\chi<0 \Leftrightarrow \chi>\frac{1}{2}$ *, alors en fonction des conditions initiales le schéma peut être stable ou instable :*
- *si* $\underline{u}^0=\underline{0}$*, le schéma est stable*
- *si* $u_j^0=(-1)^j$ *uniformément borné, alors par récurrence on montre que :*

$$u_j^n=(-1)^j(1-4\chi)^n$$

*donc* $\lim_{n\to\infty}|u_j|^n=+\infty$ *le schéma est instable.*

**Remarque 3.10.**
En pratique, un schéma instable est inutilisable car, même si pour certaines conditions initiale il peut êtte stable, avec les erreurs numériques d'arrondis qui existent le schéma diverge.

**Proposition 3.4.** *Le schéma d'Euler rétrograde* (3.12) *est inconditionnellement stable* $L^\infty$*.*

**Preuve.** *Soient* $m\leq 0\leq M$ *tels que* $\forall j\in[\![1;N]\!]$, $m\leq u_j^0\leq M$*. Avec les conditions de Dirichlet nulles, on montre par récurrence que :*

$$\forall j\in[\![0;N+1]\!],\ m\leq u_j^n\leq M$$

*Soit* $M'=\max_{1\leq j\leq N}u_j^{n+1}$*. Montrons que* $M'\leq M$ *:*

**Si** $M'=0$ *, vrai car* $M\geq 0$*.*

**Si** $M'>0$ *, en notant* $J$ *l'indice tel que* $u_J^{n+1}=M'$*, alors par définition du schéma* (3.12)

$$(1+2\chi)u_J^{n+1}=u_J^n+2\chi\frac{u_{J-1}^{n+1}+u_{J+1}^{n+1}}{2}\leq u_J^n+2\chi u_J^{n+1}\Rightarrow M'=u_J^{n+1}\leq u_J^n\leq M$$

*En notant* $m'=\min_{1\leq j\leq N}u_j^{n+1}$*, montrons que* $m'\geq m$ *:*

**Si** $m'=0$ *, vrai car* $m\leq 0$*.*

**Si** $m'<0$ *, en notant* $J$ *l'indice tel que* $u_J^{n+1}=m'$*, alors par définition du schéma* (3.12)

$$(1+2\chi)u_J^{n+1}=u_J^n+2\chi\frac{u_{J-1}^{n+1}+u_{J+1}^{n+1}}{2}\geq u_J^n+2\chi u_J^{n+1}\Rightarrow m'=u_J^{n+1}\geq u_J^n\geq m$$

**Exercice 3.4.** Montrer que le $\theta$-schéma est stable $L^\infty$ ssi $\chi\leq\frac{1}{2(1-\theta)}$ (avec des conditions aux limites de Dirichlet).

### Stabilité $L^2$

Certains schémas ne respectent pas le principe du maximum discret et sont donc instables $L^\infty$. Dans ce cas il peut être intéressant d'étudier leur stabilité $L^2$. Pour simplifier l'étude, considérons le problème de la chaleur avec des conditions aux limites périodiques :

$$\begin{cases}\partial_1 u(t,x)-\kappa\partial_2^2u(t,x)=0 & \forall(t,x)\in\mathbb{R}_\star^+\times\Omega=]0,L[\\ u(0,x)=u^0(x) & \forall x\in\Omega\\ u(t,x+L)=u(t,x) & \forall t>0,\forall x\in\Omega\end{cases}\tag{3.16}$$

**Remarque 3.11.**
Lorsque les conditions aux limites du problème ne sont pas des conditions périodiques, il est tout de même possible d'étudier la stabilité $L^2$ d'un schéma en ignorant les conditions aux bords et en considérant uniquement l'EDP.

**Principe de l'étude de stabilité $L^2$** Pour tout $\underline{u}^n = (u_j^n)_{0\leq j\leq N}$, on peut définir $u$ la fonction constante par morceaux par :

$$\forall x \in \bar{\Omega} = [0, L],\ u^n(x) = u_j^n \qquad \text{si } x_{j-\frac{1}{2}} \leq x \leq x_{j+\frac{1}{2}} \text{ ,où } \begin{cases} x_{j+\frac{1}{2}} = (j+\frac{1}{2})\Delta x & \forall j \in [\![0; N]\!] \\ x_{-\frac{1}{2}} = 0 \\ x_{N+1+\frac{1}{2}} = L \end{cases}$$

Pour tout $n \geq 0$, $u^n \in L^2_{\#}(\bar{\Omega}) = \{u^n \in L^2(\bar{\Omega})/u^n$ périodique sur $\Omega : u^n(x+L) = u^n(x) \forall x \in \bar{\Omega}\}$. Il est donc possible d'en faire sa décomposition en série de Fourier :

$$u^n(x) = \sum_{k\in\mathbb{Z}} \hat{u}^n(k) e^{I2\pi kx} \qquad \text{où } \hat{u}^n(k) = \int_\Omega u^n(x) e^{-I2\pi kx}\, dx \tag{3.17}$$

**Proposition 3.5.** *Soit $v^n(x) = u^n(x + \Delta x)$. Alors $\hat{v}^n(k) = \hat{u}^n(k) e^{I2\pi k\Delta x}$.*

La théorème de Parseval donne alors :

$$\int_\Omega |u^n(x)|^2\, dx = \sum_{k\in\mathbb{Z}} |\hat{u}^n(k)|^2 \tag{3.18}$$

**Définition 3.6.**

Pour les schémas à deux niveaux, on appelle facteur d'amplification $A(k)$ le scalaire tel que

$$\hat{u}^n(k) = A(k)\hat{u}^{n-1}(k) = A(k)^n \hat{u}^0(k)$$

Pour les schémas multiniveaux, $\underline{\underline{A}}(k)$ est la matrice d'amplification telle que

$$\underline{U}^n(k) = \underline{\underline{A}}(k).\underline{U}^{n-1}(k)$$

**Proposition 3.6.** *On appelle condition (nécessaire) de stabilité de Von Neumann l'inégalité*

$$\forall k \in \mathbb{Z},\ \rho(\underline{\underline{A}}(k)) \leq 1$$

*(qui devient $|A(k)| \leq 1$ pour les schémas à deux niveaux).*

**Remarque 3.12.**

Si $\underline{\underline{A}}(k)$ est une matrice normale, alors $\left|\left|\left|\underline{\underline{A}}(k)\right|\right|\right|_2 = \rho(\underline{\underline{A}}(k))$ et la condition de Von Neumann devient une condition nécessaire et suffisante de stabilité $L^2$.

**Exemples d'étude de stabilité $L^2$** Etudions la stabilité $L^2$ des schémas vus précédemment.

**Proposition 3.7.** *Le schéma Euler progressif* (3.13) *est stable $L^2$ ssi $\chi \leq \frac{1}{2}$.*

**Preuve.** *Le schéma peut s'écrire*

$$\frac{u^{n+1}(x)-u^n(x)}{\Delta t}-\kappa\frac{u^n(x-\Delta x)-2u(x)+u^n(x+\Delta x)}{\Delta x^2}=0$$

*Par décomposition en séries de Fourier, on trouve le facteur d'amplification*

$$A(k)=\Big(1-4\chi\sin^2(\pi k\Delta x)\Big)$$

*Par conséquent, $\hat{u}^n(k)$ est bornée pour $k\in\mathbb{Z}$ ssi $|A(k)|\leq 1$, ce qui est vérifié sous la condition CFL $\chi\leq 1/2$. Enfin, grâce à la formule de Parseval* (3.18) *:*

$$\|u^n\|_{L^2}^2=\int_0^L|u^n(x)|^2\,dx=\sum_{k\in\mathbb{Z}}|\hat{u}^n(k)|^2\leq\sum_{k\in\mathbb{Z}}|\hat{u}^0(k)|^2=\int_0^L|u^0(x)|^2\,dx=\|u^0\|_{L^2}^2$$

*Ce qui prouve la stabilité $L^2$.*

**Proposition 3.8.** *Le schéma Euler rétrograde* (3.12) *est inconditionnellement stable $L^2$.*

**Preuve.** *De la même manière que pour le schéma précédent, on obtient le facteur d'amplification*

$$A(k)=\frac{1}{1+4\chi\sin^2(\pi k\Delta x)}$$

*Par conséquent, $\hat{u}^n(k)$ est bornée quelle que soit la valeur $\chi$. La formule de Parseval* (3.18) *prouve ensuite la stabilité $L^2$.*

**Exercice 3.5.** Etudier la stabilité $L^2$ du $\theta$-schéma (3.14).

**Proposition 3.9.** *Le schéma de Richardson* (3.11) *est inconditionnellement instable en norme $L^2$.*

**Preuve.** *La matrice d'amplification du schéma de Richardson s'écrit :*

$$\underline{\underline{A}}(k)=\begin{bmatrix}-8\chi\sin^2(\pi k\Delta x) & 1\\ 1 & 0\end{bmatrix}$$

*$\underline{\underline{A}}(k)$ étant symétrique à valeurs réelles, elle est normale et donc la condition de Von Neumann est nécessaire et suffisante. Calculons alors ses valeurs propres. Le polynome caractéristique s'écrit :*

$$\det(\underline{\underline{A}}(k)-\lambda\underline{\underline{I}})=0=\lambda^2+\lambda 8\chi\sin^2(\pi k\Delta x)-1$$

*dont le discriminant strictement positif implique qu'il admet deux racines réelles dont le produit vaut -1. Par conséquent $\rho(\underline{\underline{A}}(k))>1$, ce qui prouve que le schéma de Richardson est inconditionnellement instable en norme $L^2$.*

### 3.2.4 Convergence du schéma

**Définition 3.7.**

On appelle erreur de convergence à l'instant $t_n$ le vecteur $\underline{e}^n$ tel que :

$$\forall j \in [\![0; N+1]\!],\ e_j^n = u_j^n - u(t_n, x_j)$$

Le théorème de Lax assure qu'un schéma stable et consistant avec l'EDP converge :

**Théorème 3.2** (de Lax)**.**

*Un schéma consistant et stable selon la norme $\|.\|$ est convergent, i.e*

$$\forall T > 0,\ \lim_{\Delta t, \Delta x \to 0} \sup_{t_n \leq T} \|\underline{e}^n\| = 0$$

*De plus, si l'erreur de consistance est $p$ en espace et $q$ en temps, alors :*

$$\forall T > 0, \exists C > 0 / \sup_{t_n \leq T} \|\underline{e}^n\| \leq C\left(\Delta x^p + \Delta t^q\right)$$

**Exemple 3.6.** Il est possible de définir la convergence des schémas présentés suivants :

- Le schéma d'Euler progressif (3.13) est convergent $L^2$ et $L^\infty$ sous la condition CFL $\chi \leq \frac{1}{2}$, et est précis d'ordre 1 en temps et 2 en espace.
- Le schéma d'Euler rétrograde (3.12) est inconditionnellement convergent $L^2$ et $L^\infty$ , et est précis d'ordre 1 en temps et 2 en espace.
- Le $\theta$-schéma (3.14) ($\theta \neq 1/2$) est convergent $L^2$ sous la condition CFL $\chi \leq \frac{1}{2(1-2\theta)}$, et $L^\infty$ sous la condition CFL $\chi \leq \frac{1}{2(1-\theta)}$. Ce schéma est précis d'ordre 1 en temps et 2 en espace.
- Le schéma de Cranck-Nicolson (3.14) ($\theta = 1/2$) est inconditionnellement convergent $L^2$, mais stable $L^\infty$ si $\chi \leq 1$. Le schéma de Cranck-Nicholson est précis d'ordre 2 en temps et 2 en espace.
- Le schéma de Richardson (3.11) est inconditionnellement non convergent $L^2$

A titre illustratif, sur la Figure 3.5 sont représentés la résolution de (3.5) pour deux conditions CFL : $\chi = 0.4 \leq \frac{1}{2}$, pour lequel le schéma est stable, et $\chi = 0.55 > \frac{1}{2}$, pour lequel il est instable.

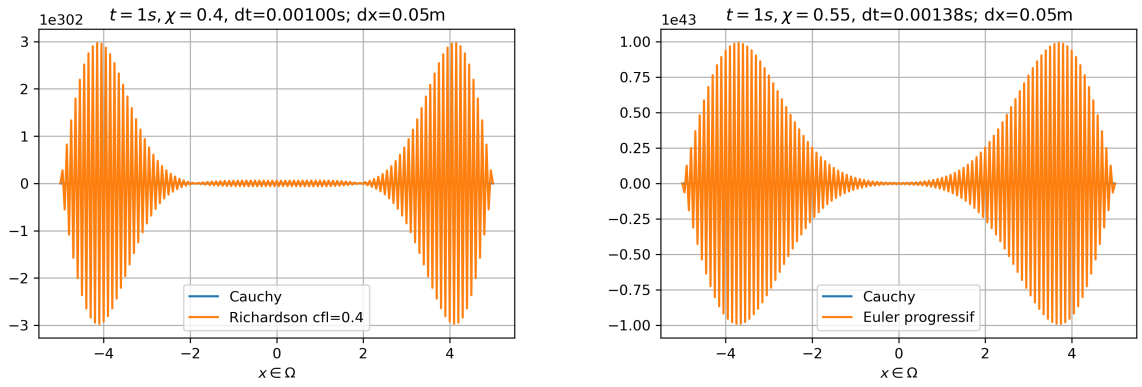



FIGURE 3.5 – Solutions de (3.5) par le schéma Euler progressif pour deux valeurs de CFL

# Chapitre 4

# Méthodes directes pour la résolution des systèmes linéaires

*Les chapitres 2 et 3 ont montré que le processus de discrétisation de l'opérateur différentiel par la méthode des différences finies peut mener à la résolution de systèmes linéaires, i.e à l'inversion d'une matrice qui est très souvent de grande taille, mais aussi en grande partie creuse.*

*En premier lieu la notion fondamentale de conditionnement de matrice est introduite, permettant de quantifier la sensibilité de la solution du système linéaire aux perturbations de la matrice ou du second membre. Ensuite les principales méthodes directes de résolution de systèmes linéaires, toutes basées sur le pivot de Gauss, sont passées en revue et leur coût de calcul respectif est estimé. Enfin, la méthode des moindres carrées est présentée, permettant la résolution de systèmes surdéterminés.*

**Sommaire**

# Introduction

**Définition 4.1.**
Un algorithme de calcul de résolution de système linéaire est dit direct s'il permet d'obtenir théoriquement la solution exacte d'un problème en un nombre fini d'opérations.

Dans ce chapitre on souhaite résoudre par des méthode directes les systèmes linéaires du type :

$$\text{Soient } \underline{\underline{A}} \in \mathcal{M}_{mn}(\mathbb{R}), \underline{b} \in \mathbb{R}^m. \text{Trouver } \underline{x} \in \mathbb{R}^n \text{ tel que } \underline{\underline{A}}.\underline{x} = \underline{b} \tag{4.1}$$

La résolution de ce problème est très fréquente, par exemple dans le cas de la résolution d'un problème aux limites par un schéma aux différences finies, ou par éléments finis.

Les points clés sont :
- étudier 3 méthodes directes : Pivot de Gauss, Factorisation LU et Cholesky (et bannir la méthode de Cramer pour $n > 2$ !)
- pour chacune, estimer le coût de calcul

## 4.1 Conditionnement d'une matrice

Dans un procédé de modélisation d'un problème, la résolution d'un système linéaire est l'étape finale résultant d'approximations numériques successives ; diverses erreurs peuvent ainsi provenir des incertitudes sur la matrice $\underline{\underline{A}}$ ou sur le second membre $\underline{b}$, mais aussi d'erreurs sur les arrondis. Ainsi au lieu de résoudre le système $\underline{\underline{A}}.\underline{x} = \underline{b}$ on résout

$$\left(\underline{\underline{A}} + \underline{\underline{\Delta A}}\right).\underline{\tilde{x}} = (\underline{b} + \underline{\Delta b})$$

On va donc chercher à majorer la différence $\underline{x} - \underline{\tilde{x}}$ en fonction des majorations de $\underline{\underline{\Delta A}}$ et $\underline{\Delta b}$.

**Exemple 4.1.** Soit le système linéaire formé par $\underline{\underline{A}} = \begin{bmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{bmatrix}$ et $\underline{b} = \begin{pmatrix} 32 \\ 23 \\ 33 \\ 31 \end{pmatrix}$. L'unique solution est $\underline{x} = {}^t\begin{pmatrix} 1 & 1 & 1 & 1 \end{pmatrix}$.
Considérons les perturbations suivantes :

$$\underline{\underline{\Delta A}} = \begin{bmatrix} 0 & 0 & 0.1 & 0.2 \\ 0.08 & 0.04 & 0 & 0 \\ 0 & -0.02 & -0.11 & 0 \\ -0.01 & -0.01 & 0 & -0.02 \end{bmatrix} \quad ; \quad \underline{\Delta b} = \begin{pmatrix} 0.01 \\ -0.01 \\ 0.01 \\ -0.01 \end{pmatrix}$$

La résolution du système $\left(\underline{\underline{A}} + \underline{\underline{\Delta A}}\right).\underline{y} = \underline{b}$ donne $\underline{y} = {}^t\begin{pmatrix} -81 & 137 & -34 & 22 \end{pmatrix}$ et celle du système $\underline{\underline{A}}.\underline{z} = (\underline{b} + \underline{\Delta b})$ donne $\underline{z} = {}^t\begin{pmatrix} 1.82 & -0.36 & 1.35 & 0.79 \end{pmatrix}$ On voit donc que d'infimes perturbations provoquent des grands changements dans la solution de ce système, qui sera considéré comme mal conditionné.

### Définition et propriétés

**Définition 4.2.**
Soit $\|.\|$ une norme matricielle. Le conditionnement d'une matrice inversible $\underline{\underline{A}} \in \mathcal{M}_n(\mathbb{K})$ associé à cette norme est défini par :

$$\text{cond}(\underline{\underline{A}}) = \|\underline{\underline{A}}\| \, \|\underline{\underline{A}}^{-1}\| = K(\underline{\underline{A}})$$

Afin de spécifier la norme utilisée, on note $\text{cond}_p$ le conditionnement relatif à la norme $\|.\|_p$

**Proposition 4.1.** *Soit* $\underline{\underline{A}} \in \mathcal{M}_n(\mathbb{K})$ *matrice inversible :*
— $\text{cond}(\underline{\underline{A}}) \geq 1$
— $\text{cond}(\underline{\underline{A}}) = \text{cond}(\underline{\underline{A}}^{-1})$
— $\forall \alpha \in \mathbb{R}^\star \ \text{cond}(\alpha \underline{\underline{A}}) = \text{cond}(\underline{\underline{A}})$

**Théorème 4.1 (**Caractérisation du conditionnement pour la norme 2**).**
*Soit* $\underline{\underline{A}} \in \mathcal{M}_n(\mathbb{K})$ *matrice inversible.*
— *Si* $\underline{\underline{A}}^{\text{H}}.\underline{\underline{A}}$ *est diagonalisable de valeurs propres* $0 < |\mu_1|^2 \leq |\mu_2|^2 \leq \ldots \leq |\mu_n|^2$*, alors*

$$\text{cond}_2(\underline{\underline{A}}) = \frac{|\mu_n|}{|\mu_1|}$$

— *Si de plus* $\underline{\underline{A}}$ *est symétrique, alors en notant* $0 < |\lambda_1| \leq |\lambda_2| \leq \ldots \leq |\lambda_n|$ *les valeurs propres de* $\underline{\underline{A}}$*, alors son conditionnement s'écrit :*

$$\text{cond}_2(\underline{\underline{A}}) = \frac{|\lambda_n|}{|\lambda_1|}$$

*Le gros avantage est qu'il n'est pas nécessaire de calculer* $\underline{\underline{A}}^{-1}$*.*

**Proposition 4.2 (**Propriétés du conditionnement pour la norme 2**).** *Soient* $\underline{\underline{A}} \in \mathcal{M}_n(\mathbb{K})$ *une matrice inversible et* $\underline{\underline{Q}} \in \mathcal{M}_n(\mathbb{K})$ *une matrice unitaire. On a alors les propriétés suivantes :*
— $\text{cond}_2(\underline{\underline{A}}) = 1$ *si et seulement si* $\exists \alpha \in \mathbb{R}^\star / \ \underline{\underline{A}} = \alpha \underline{\underline{Q}}$ *;*
— *Invariance par transformation unitaire :* $\text{cond}_2(\underline{\underline{A}}) = \text{cond}_2(\underline{\underline{A}}.\underline{\underline{Q}}) = \text{cond}_2(\underline{\underline{Q}}.\underline{\underline{A}}) = \text{cond}_2(\underline{\underline{Q}}^{\text{H}}.\underline{\underline{A}}.\underline{\underline{Q}})$.

**Exercice 4.1.** Reprenons la matrice de l'exemple 4.1 : $\underline{\underline{A}} = \begin{bmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{bmatrix}$ En vous aidant d'un logiciel, calculer le conditionnement de $\underline{\underline{A}}$ par rapport aux normes 1, 2, $\infty$ et de Frobenius.

**Estimation d'erreurs dans un système linéaire**

**Théorème 4.2.**
*Supposons que* $\underline{\underline{A}}.\underline{x} = \underline{b}$ *et* $\underline{\underline{A}}.\underline{\tilde{x}} = \underline{\tilde{b}}$*. Notons* $\underline{\Delta x} = \underline{\tilde{x}} - \underline{x}$ *et* $\underline{\Delta b} = \underline{\tilde{b}} - \underline{b}$*. On a alors la relation :*

$$\frac{\|\underline{\Delta x}\|}{\|\underline{x}\|} \leq \text{cond}(\underline{\underline{A}}) \, \frac{\|\underline{\Delta b}\|}{\|\underline{b}\|}$$

*Si au contraire, on modifie* $\underline{\underline{A}}$ *en* $\underline{\underline{\tilde{A}}} = \underline{\underline{A}} + \underline{\underline{\Delta A}}$ *avec* $\underline{b} = \underline{\tilde{b}}$*, donc* $\left(\underline{\underline{A}} + \underline{\underline{\Delta A}}\right).\underline{\tilde{x}} = \underline{b}$ *:*

$$\frac{\|\underline{\Delta x}\|}{\|\underline{\tilde{x}}\|} \leq \text{cond}(\underline{\underline{A}}) \frac{\|\underline{\underline{\Delta A}}\|}{\|\underline{\underline{A}}\|}$$
$$\frac{\|\underline{\Delta x}\|}{\|\underline{x}\|} \leq \text{cond}(\underline{\underline{A}}) \frac{\|\underline{\underline{\Delta A}}\|}{\|\underline{\underline{A}}\|} \left(1 + O(\|\underline{\underline{\Delta A}}\|)\right)$$

**Preuve.** *Par soustraction des systèmes* $\underline{\underline{A}}.\underline{x} = \underline{b}$ *et* $\underline{\underline{A}}.\underline{\tilde{x}} = \underline{\tilde{b}}$ *on obtient :*

$$\underline{\Delta b} = \underline{\underline{A}}.\underline{\Delta x} \Rightarrow \underline{\Delta x} = \underline{\underline{A}}^{-1}.\underline{\Delta b} \Rightarrow \|\underline{\Delta x}\| \leq \|\underline{\underline{A}}^{-1}\| \, \|\underline{\Delta b}\|$$

*De plus,*

$$\underline{\underline{A}}.\underline{x} = \underline{b} \Rightarrow \|\underline{b}\| \leq \|\underline{\underline{A}}\| \, \|\underline{x}\| \text{ ou } \frac{1}{\|\underline{x}\|} \leq \frac{\|\underline{\underline{A}}\|\}}{\|\underline{b}\|}$$

*Par conséquent on obtient :*

$$\frac{\|\underline{\Delta x}\|}{\|\underline{x}\|} \leq \text{cond}(\underline{\underline{A}}) \, \frac{\|\underline{\Delta b}\|}{\|\underline{b}\|}$$

*Pour la deuxième partie, on part de l'égalité :*

$$\underline{b} = \underline{\tilde{b}} \Rightarrow \underline{0} = \underline{\underline{A}}.\underline{x} - \underline{\underline{\tilde{A}}}.\underline{\tilde{x}} = \underline{\underline{A}}.\left(\underline{x} - \underline{\tilde{x}}\right) + \left(\underline{\underline{A}} - \underline{\underline{\tilde{A}}}\right).\underline{\tilde{x}}$$

*Par conséquent,*

$$\underline{x} - \underline{\tilde{x}} = \underline{\underline{A}}^{-1}.\left(\underline{\underline{\tilde{A}}} - \underline{\underline{A}}\right).\underline{\tilde{x}}$$

*Ce qui entraîne :*

$$\|\underline{x} - \underline{\tilde{x}}\| = \|\underline{\Delta x}\| \leq \|\underline{\underline{A}}^{-1}\| \, \|\underline{\underline{\Delta A}}\| \, \|\underline{\tilde{x}}\| = \text{cond}(\underline{\underline{A}}) \frac{\|\underline{\underline{\Delta A}}\|}{\|\underline{\underline{A}}\|}$$

*De la même manière,*

$$\underline{0} = \left(\underline{\underline{A}} - \underline{\underline{\tilde{A}}}\right).\underline{x} + \underline{\underline{\tilde{A}}}.\left(\underline{x} - \underline{\tilde{x}}\right) \Rightarrow \underline{\tilde{x}} - \underline{x} = \underline{\underline{\tilde{A}}}^{-1}.\left(\underline{\underline{\tilde{A}}} - \underline{\underline{A}}\right).\underline{x}$$

*et donc :*

$$\frac{\|\underline{\Delta x}\|}{\|\underline{x}\|} \leq \frac{\|\left(\underline{\underline{A}} + \underline{\underline{\Delta A}}\right)^{-1}\|}{\|\underline{\underline{A}}^{-1}\|} \, \frac{\|\underline{\underline{\Delta A}}\|}{\|\underline{\underline{A}}\|} \, \text{cond}(\underline{\underline{A}})$$

*Et le terme* $\frac{\|\left(\underline{\underline{A}}+\underline{\underline{\Delta A}}\right)^{-1}\|}{\|\underline{\underline{A}}^{-1}\|} \xrightarrow[\|\underline{\underline{\Delta A}}\| \to 0]{} 1$

## 4.2 Méthode du pivot de Gauss

Soit le système "plein" $\underline{\underline{A}}.\underline{x} = \underline{b}$, où $\underline{\underline{A}} \in \mathcal{M}_n(\mathbb{R})$ telle que $\det(\underline{\underline{A}}) \neq 0$. La méthode de résolution dite du pivot de Gauss comporte deux phases : une première phase **d'élimination**, qui permet de transformer un système "plein" en un système triangulaire supérieur de type $\underline{\underline{T}}.\underline{x} = \underline{c}$. Une seconde phase dite de **remontée du système triangulaire**.

### 4.2.1 Phase d'élimination

Le principe de ce processus itératif (avec $n-1$ itérations ) réside en une succession de combinaisons linéaires des lignes. Notons $\underline{\underline{A}}^{(i)}$ et $\underline{b}^{(i)}$ les résultats de ces combinaisons linéaires à l'itération $i$.

**Itération 0 :** Initialisation de la méthode. $\underline{\underline{A}}^{(0)} = \underline{\underline{A}} = \begin{bmatrix} a_{11} & a_{12} & \dots & \dots & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & \dots & \dots & a_{2n} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & a_{ij} & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & \dots & \dots & a_{nn} \end{bmatrix}$ et $\underline{b}^{(0)} = \underline{b}$

**Itération 1 :** On suppose que $a_{11} \neq 0$, que l'on choisit comme pivot $\Pi_1 = a_{11}^{(0)} = a_{11}$ afin d'éliminer $x_1$ des équations aux lignes supérieures à 2. On applique ainsi la combinaison linéaire, en notant $L_i$ la ligne $i$ :

$$\forall i \in [\![2; n]\!], \quad L_i \leftarrow L_i - \frac{a_{i1}}{\Pi_1} \, L_1$$

Soit

$$\forall (i,j) \in [\![2;n]\!] \times [\![2;n]\!], \quad \ell_{i1} = \frac{a_{i1}}{\Pi_1}$$

$$\begin{cases} a_{ij}^{(1)} = a_{ij}^{(0)} - \ell_{i1}\, a_{1j}^{(0)} \\ b_i^{(1)} = b_i^{(0)} - \ell_{i1}\, b_1^{(0)} \end{cases} \Rightarrow \underline{\underline{A}}^{(1)} = \begin{bmatrix} a_{11} & a_{12} & \dots & \dots & \dots & a_{1n} \\ 0 & a_{22}^{(1)} & \dots & \dots & \dots & a_{2n}^{(1)} \\ \vdots & \dots & \dots & \dots & \dots & \dots \\ \vdots & \dots & \dots & a_{ij}^{(1)} & \dots & \dots \\ \vdots & \dots & \dots & \dots & \dots & \dots \\ 0 & a_{n2}^{(1)} & \dots & \dots & \dots & a_{nn}^{(1)} \end{bmatrix}$$

**Itération $k$ :** De la même manière, on choisit le pivot non nul $\Pi_k = a_{kk}^{(k-1)}$ afin d'éliminer $x_k$ des équations aux lignes supérieures à $k+1$, grâce à la combinaison linéaire :

$$\forall i \in [\![k+1;n]\!], \quad L_i \leftarrow L_i - \frac{a_{ik}^{(k-1)}}{\Pi_k} L_k$$

Soit

$$\forall (i,j) \in [\![k+1;n]\!] \times [\![k+1;n]\!], \quad \ell_{ik} = \frac{a_{ik}^{(k-1)}}{\Pi_k}$$

$$\begin{cases} a_{ij}^{(k)} = a_{ij}^{(k-1)} - \ell_{ik}\, a_{kj}^{(k-1)} \\ b_i^{(k)} = b_i^{(k-1)} - \ell_{ik}\, b_k^{(k-1)} \end{cases} \Rightarrow \underline{\underline{A}}^{(k)} = \begin{bmatrix} a_{11} & a_{12} & \dots & \dots & \dots & a_{1n} \\ 0 & a_{22}^{(1)} & \dots & \dots & \dots & a_{2n}^{(1)} \\ \vdots & \ddots & \ddots & \dots & \dots & \dots \\ 0 & \dots & 0 & a_{k+1,k+1}^{(k)} & \dots & a_{k+1,n}^{(k)} \\ \vdots & \dots & \vdots & \vdots & \dots & \dots \\ 0 & \dots & 0 & a_{n,k+1}^{(k)} & \dots & a_{nn}^{(k)} \end{bmatrix}$$

**Itération $n-1$ :** Finalement après $n-1$ itérations le système initial est devenu :

$$\underline{\underline{A}}^{(n-1)}.\underline{x} = \begin{bmatrix} a_{11} & a_{12} & \dots & \dots & \dots & a_{1n} \\ 0 & a_{22}^{(1)} & \dots & \dots & \dots & a_{2n}^{(1)} \\ \vdots & \ddots & \ddots & \dots & \dots & \dots \\ 0 & \dots & 0 & a_{k+1,k+1}^{(k)} & \dots & a_{k+1,n}^{(k)} \\ \vdots & \dots & \vdots & \ddots & \ddots & \dots \\ 0 & \dots & 0 & \dots & 0 & a_{nn}^{(n-1)} \end{bmatrix} .\underline{x} = \begin{pmatrix} b_1 \\ b_2^{(1)} \\ \dots \\ b_{k+1}^{(k)} \\ \dots \\ b_n^{(n-1)} \end{pmatrix}$$

Notons $\underline{\underline{T}} = \underline{\underline{A}}^{(n-1)}$ et $\underline{c} = \underline{b}^{(n-1)}$ respectivement la matrice triangulaire supérieure et le second membre ainsi obtenus.

### 4.2.2 Phase de remontée d'un système triangulaire

Le système triangulaire à résoudre est donc le suivant :

$$\begin{cases} t_{11}x_1 + t_{12}x_2 + \ldots + t_{1n}x_n = c_1 \\ \qquad\quad t_{22}x_2 + \ldots + t_{2n}x_n = c_2 \\ \qquad\qquad\qquad\qquad \ddots \quad \vdots \\ \qquad\qquad\qquad\qquad t_{nn}x_n = c_n \end{cases}$$

Par remontée du système, la solution est donnée par :

$$\begin{cases} x_n = \dfrac{c_n}{t_{nn}} \\ x_i = \dfrac{1}{t_{ii}} \left( c_i - \displaystyle\sum_{k=i+1}^{n} t_{ik} x_k \right) \end{cases} \quad \forall i \in [\![1; n-1]\!]$$

### 4.2.3 Algorithme

**Phase d'élimination :**

$$\left[\begin{array}{l} k = 1 \dots n-1 \\ \quad \left[\begin{array}{l} i = k+1 \dots n \\ \quad \ell = \dfrac{a_{ik}}{a_{kk}} \\ \quad \quad \left[ j = k+1 \dots n \; \middle| \; a_{ij} = a_{ij} - \ell\, a_{kj} \right] \\ \quad b_i = b_i - \ell\, b_k \end{array}\right. \end{array}\right.$$

**Coût de calcul :**

divisions : $\displaystyle\sum_{k=1}^{n-1} \sum_{i=k+1}^{n} 1 = \sum_{k=1}^{n-1} (n-k) = \frac{n(n-1)}{2}$

additions : $\displaystyle\sum_{k=1}^{n-1} \sum_{i=k+1}^{n} \left( 1 + \sum_{j=k+1}^{n} 1 \right) = \sum_{k=1}^{n-1} (n-k)(n-k+1) = \frac{n(n^2-1)}{3}$

multiplications : $\displaystyle\sum_{k=1}^{n-1} \sum_{i=k+1}^{n} \left( 1 + \sum_{j=k+1}^{n} 1 \right) = \sum_{k=1}^{n-1} (n-k)(n-k+1) = \frac{n(n^2-1)}{3}$

total : $\dfrac{n}{6}\left(4n^2 + 3n - 7\right) \approx \dfrac{2}{3} n^3$ si $n \gg 1$

**Phase de remontée :**

$$\left[\begin{array}{l} x_n = \dfrac{c_n}{t_{nn}} \\ \quad \left[\begin{array}{l} i = n-1 \dots 1 \\ \quad s = c_i \\ \quad \quad \left[ k = i+1 \dots n \; \middle| \; s = s - t_{ik} x_k \right] \\ \quad x_i = \dfrac{s}{t_{ii}} \end{array}\right. \end{array}\right.$$

**Coût de calcul :**

divisions : $1 + \displaystyle\sum_{i=1}^{n-1} 1 = n$

additions : $\displaystyle\sum_{i=1}^{n-1} (n-i) = \frac{n(n-1)}{2}$

multiplications : $\displaystyle\sum_{i=1}^{n-1} (n-i) = \frac{n(n-1)}{2}$

total : $n^2$

**Coût total :**

$$\text{TotalDivisions} = \frac{n(n+1)}{2} \; ; \; \text{TotalAdditions} = \text{TotalMultiplications} = \frac{2n^3 + 3n^2 - 5n}{6}$$

Soit :

$$\text{CoûtTotal} = \frac{4n^3 + 9n^2 - 7n}{6} \approx \frac{2\,n^3}{3}$$

**Exercice 4.2.** Calcul le nombre d'opérations nécessaire pour résoudre un système symétrique par pivot de Gauss.

### 4.2.4 Choix du pivot

Si à l'étape $k$, le coefficient $a_{kk}^{(k)}$ est nul, il faut alors permuter la ligne avec une autre ligne dont le premier coefficient est non nul.

Il faut aussi éviter les pivots trop petits.

**Exemple 4.2.** Soit le système, $0 < \varepsilon \ll 1$ :

$$\begin{cases} \varepsilon x_1 + x_2 = 1 \\ x_1 + x_2 = 2 \end{cases} \tag{4.2}$$

La formule de Cramer donne la solution exacte :

$$\begin{cases} x_1 = \dfrac{1}{1-\varepsilon} \\ x_2 = \dfrac{1-2\varepsilon}{1-\varepsilon} \end{cases} \tag{4.3}$$

Supposons que la précision des calculs flottants (notée $fl$) en machine conduit à :

$$fl(1 - \frac{1}{\varepsilon}) = fl(2 - \frac{1}{\varepsilon}) = -\frac{1}{\varepsilon}\,;\; fl(1-\varepsilon) = fl(1-2\varepsilon) = 1 \tag{4.4}$$

Appliquons alors la méthode du pivot en effectuant la combinaison linéaire : $L_2 \leftarrow L_2 - \frac{1}{\varepsilon} L_1$ Alors le système (4.2) devient :

$$\begin{cases} \varepsilon x_1 + x_2 = 1 \\ fl(1 - \dfrac{1}{\varepsilon}) x_2 = fl(2 - \dfrac{1}{\varepsilon}) \end{cases}$$

En appliquant les approximations (4.4), la deuxième fournit $x_2 = 1$. Puis en réinjectant cette valeur de $x_2$ dans la première ligne, on obtient :

$$\varepsilon x_1 + 1 = 1 \Rightarrow x_1 = 0$$

La valeur ainsi obtenue pour $x_1$ est très mauvaise, $x_1 + x_2 = 1 \neq 2$ ! Maintenant recommençons en permuttant les deux lignes, afin que le pivot ne soit plus $\varepsilon$ :

$$\begin{cases} x_1 + x_2 = 2 \\ fl(1-\varepsilon) x_2 = fl(1-2\varepsilon) \end{cases}$$

La remontée du système fournit le résultat $x_1 = x_2 = 1$, correspondant à approximation au premier ordre de la solution exacte (4.3).

**Remarque 4.1.**
Une stratégie de calcul peut être d'amener, par permutations de lignes et de colonnes, en position pivot le coefficient le plus grand en valeur absolue.

**Exercice 4.3.** Soir $\underline{\underline{A}}.\underline{x} = \underline{b}$ le système linéaire défini par (on pourra prendre par exemple $\varepsilon = 1.0 \times 10^{-4}$) :

$$\underline{\underline{A}} = \begin{bmatrix} \varepsilon & 1.0 \\ 1.0 & 1.0 \end{bmatrix} \quad ; \quad \underline{b} = \begin{pmatrix} 1.0 \\ 2.0 \end{pmatrix}$$

Résoudre ce système pour $\varepsilon \ll 1$ par pivot de Gauss sans et avec permutation, puis calculer le conditionnement (selon la norme $\|.\|_1$) des systèmes triangulaires.

### 4.2.5 Calcul du déterminant par la méthode du pivot

Les opérations de combinaisons linéaires des lignes effectuées sur la matrice $\underline{\underline{A}}$ conservent tous les déterminants partiels de $\underline{\underline{A}}$. A la fin de la phase d'élimination une matrice triangulaire supérieure est obtenue, dont les coefficients diagonaux sont les pivots $\{\Pi_k\}_{1 \leq k \leq n}$, où par extension $\Pi_n = a_{nn}^{(n-1)}$. Ainsi,

$$\Delta_A = \det(\underline{\underline{A}}) = \prod_{k=1}^{n} \Pi_k \tag{4.5}$$

Le coût de calcul du déterminant est donc : $\frac{n}{6}\left(4n^2 + 3n - 7\right)$ opérations pour la phase d'élimination et $n$ multiplications pour le produit des pivots, soit une estimation du nombre d'opérations de $O(\frac{2}{3}n^3)$ si $n \gg 1$. Rappelons que le coût de calcul d'un déterminant par la formule de Cramer est $O(n\,n!)$ : si $n = 20$, $\text{coût}_{\text{Cramer}} = 4.9\,10^{19} \gg \text{coût}_{\text{Gauss}} = 5.5\,10^3$, soit 570 ans $\gg 2\mu s$ avec un processeur de 2.7 GHz...

**Exercice 4.4.** Calculer le déterminant de la matrice $\underline{\underline{A}} = \begin{bmatrix} 2 & 4 & 4 \\ 1 & 3 & 1 \\ 1 & 5 & 6 \end{bmatrix}$ par pivot de Gauss.

**Remarque 4.2.**
Attention, si des permutations des lignes sont effectuées il faut mémoriser leur nombre $N_p$ afin de rétablir le signe du déterminant. La formule (4.5) devient donc :

$$\Delta_A = \det(\underline{\underline{A}}) = (-1)^{N_p} \prod_{k=1}^{n} \Pi_k$$

### 4.2.6 Systèmes creux

Une matrice est dite "creuse" si le nombre de coefficients $a_{ij}$ non nuls est petit devant $n^2$. A l'étape $k$, $a_{ij}$ est modifié ssi $a_{ik}$ et $a_{kj}$ sont non nuls. Pour une matrice creuse dont les zéros sont "bien placés" il y a une énorme économie de calculs.

**Exemple 4.3 (**Matrice bande**).** Soit une matrice bande, i.e

$$\forall (i,j) \in [\![1;n]\!]^2 / \, |i-j| > m \Rightarrow a_{ij} = 0$$

La demi-largeur de bande $m$ est un entier petit devant $n$. La méthode de Gauss nécessite beaucoup moins d'opérations que dans le cas général : à chaque étape $k \in [\![1; n-m]\!]$, les seuls éléments $a_{ij}$ à modifier sont localisés dans une matrice carrée d'ordre $m$.
**Coût de calcul :**

$$\text{Elimination} \begin{cases} \text{Multiplications/Additions :} & \displaystyle\sum_{k=1}^{n-m} \sum_{i=k+1}^{k+m} \left(1 + \sum_{j=k+1}^{k+m} 1\right) + \sum_{k=n-m+1}^{n-1} \sum_{i=k+1}^{n} \left(1 + \sum_{j=k+1}^{n} 1\right) \\ & = mn\,(m+1) - \dfrac{m}{3}\left(1 + 3m + 2m^2\right) \\ \text{Divisions :} & \displaystyle\sum_{k=1}^{n-m} \sum_{i=k+1}^{k+m} 1 + \sum_{k=n-m+1}^{n-1} \sum_{i=k+1}^{n} 1 = mn - \frac{m(m+1)}{2} \end{cases}$$

$$\text{Remontée} \begin{cases} \text{Multiplications/Additions :} & \displaystyle\sum_{k=1}^{n-m} \sum_{i=k+1}^{k+m} 1 + \sum_{k=n-m+1}^{n-1} \sum_{i=k+1}^{n} 1 = mn - \frac{m(m+1)}{2} \\ \text{Divisions :} & \displaystyle 1 + \sum_{k=1}^{n-1} 1 = n \end{cases}$$

TotalDivisions $= (m+1)\dfrac{2n-m}{2}$ ; TotalAdditions = TotalMultiplications $= (n-m)m(m+2) + \dfrac{m}{6}(m-1)(2m+5)$

Soit :

$$\text{CoûtTotal} \approx m^2 n \text{ si } m \ll n$$

**Exercice 4.5.** Calculer le nombre d'opérations nécessaires à la résolution d'un système tridiagonal par pivot de Gauss.

## 4.3 Factorisation LU d'une matrice

On cherche à décomposer $\underline{\underline{A}} \in \mathcal{M}_n(\mathbb{R})$ tel que $\underline{\underline{A}} = \underline{\underline{L}}.\underline{\underline{U}}$ avec :
$\underline{\underline{L}}$ : matrice triangulaire inférieure ("Lower") dont les coefficients diagonaux valent 1.
$\underline{\underline{U}}$ : matrice triangulaire supérieure ("Upper").

### 4.3.1 LU comme pivot de Gauss matriciel

Chaque itération de la méthode de la méthode du pivot de Gauss (section 4.2) peut être écrite matriciellement, de telle sorte qu'à l'itération $k$, le passage de $\underline{\underline{A}}^{(k-1)}$ à $\underline{\underline{A}}^{(k)}$ s'écrit :

$$\begin{cases} \underline{\underline{A}}^{(k)} = \underline{\underline{G}}^{(k)}.\underline{\underline{A}}^{(k-1)} \\ \underline{b}^{(k)} = \underline{\underline{G}}^{(k)}.\underline{b}^{(k-1)} \end{cases} \quad \text{où} \quad \underline{\underline{G}}^{(k)} = \begin{bmatrix} 1 & 0 & \dots & \dots & \dots & \dots & 0 \\ 0 & 1 & 0 & \dots & \dots & \dots & \dots \\ \vdots & \ddots & \ddots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & 1 & \dots & \dots & \dots \\ \vdots & \dots & \vdots & -\ell_{k+1,k} & \ddots & \dots & \dots \\ \vdots & \dots & \vdots & \vdots & \dots & \ddots & 0 \\ 0 & \dots & 0 & -\ell_{n,k} & \dots & \dots & 1 \end{bmatrix}$$

Les seuls coefficients non nuls de $\underline{\underline{G}}^{(k)}$ sont la diagonale de 1 et la colonne des $-\ell_{ik} = -\frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}}$ en dessous de l'élément diagonale $k$. Le système triangulaire $\underline{\underline{T}}.\underline{x} = \underline{c}$ (voir section 4.2.1) est par conséquent obtenu en multipliant les matrices de passage $\underline{\underline{G}}^{(k)}$ successives, de telle sorte que :

$$\underline{\underline{T}} = \underline{\underline{G}}.\underline{\underline{A}} \qquad ; \qquad \underline{c} = \underline{\underline{G}}.\underline{b}$$

où

$$\underline{\underline{G}} = \prod_{k=1}^{n-1} \underline{\underline{G}}^{(n-k)} = \underline{\underline{G}}^{(n-1)}.\underline{\underline{G}}^{(n-2)} \dots \underline{\underline{G}}^{(1)}$$

**Remarque 4.3.**
L'ordre du produit de matrice est très important.

Au final, on obtient la décomposition $\underline{\underline{A}} = \underline{\underline{G}}^{-1}.\underline{\underline{T}}$, où $\underline{\underline{T}}$ est une matrice triangulaire supérieure et $\underline{\underline{G}}^{-1}$ une matrice triangulaire inférieure de diagonale 1. On peut de plus spécifier $\underline{\underline{G}}^{-1}$ :

$$\underline{\underline{G}}^{-1} = \prod_{k=1}^{n-1} \left(\underline{\underline{G}}^{(k)}\right)^{-1} = \left(\underline{\underline{G}}^{(1)}\right)^{-1} . \left(\underline{\underline{G}}^{(2)}\right)^{-1} \dots \left(\underline{\underline{G}}^{(n-1)}\right)^{-1}$$

où l'inverse de chaque matrice $\underline{\underline{G}}^{(k)}$ a pour expression (noter le changement de signe des coefficients $\ell_{i,j}$) :

$$\left(\underline{\underline{G}}^{(k)}\right)^{-1} = \begin{bmatrix} 1 & 0 & \dots & \dots & \dots & \dots & 0 \\ 0 & 1 & 0 & \dots & \dots & \dots & \dots \\ \vdots & \ddots & \ddots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & 1 & \dots & \dots & \dots \\ \vdots & \dots & \vdots & +\ell_{k+1,k} & \ddots & \dots & \dots \\ \vdots & \dots & \vdots & \vdots & \dots & \ddots & 0 \\ 0 & \dots & 0 & +\ell_{n,k} & \dots & \dots & 1 \end{bmatrix}$$

dont le produit matriciel fournit l'expression finale de $\underline{\underline{L}} = \underline{\underline{G}}^{-1}$ en fonction des coefficients des pivots de Gauss successifs :

$$\underline{\underline{L}} = \underline{\underline{G}}^{-1} = \begin{bmatrix} 1 & 0 & \dots & \dots & \dots & \dots & 0 \\ \ell_{2,1} & 1 & 0 & \dots & \dots & \dots & \dots \\ \vdots & \ddots & \ddots & \dots & \dots & \dots & \dots \\ \ell_{k,1} & \dots & \ell_{i,j} & 1 & \dots & \dots & \dots \\ \vdots & \dots & \vdots & +\ell_{k+1,k} & \ddots & \dots & \dots \\ \vdots & \dots & \vdots & \vdots & \dots & \ddots & 0 \\ \ell_{n,1} & \dots & \ell_{n,j} & +\ell_{n,k} & \dots & \dots & 1 \end{bmatrix}$$

### 4.3.2 Théorème d'existence et d'unicité

**Définition 4.3.**
Soit $\underline{\underline{A}} \in \mathcal{M}_n(\mathbb{R})$. Une sous-matrice principale $\underline{\underline{A}}_i$, $i < n$ de $\underline{\underline{A}}$ est une matrice carrée de la forme $\underline{\underline{A}}_{(1:i;1:i)}$.
Le déterminant de $\underline{\underline{A}}_i$ est appelé le mineur de $\underline{\underline{A}}$, noté $D_i$.

**Théorème 4.3.**
*$\underline{\underline{A}}$ inversible est décomposable en LU si et seulement si toutes ses sous-matrices principales sont inversibles (i.e $\forall 1 \leq i < n$, $D_i \neq 0$). Si elles le sont la décomposition est **unique**.*

**Exemple 4.4.** La matrice $\underline{\underline{A}} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ est-elle décomposable en LU ? Non, pourtant elle est inversible.

### 4.3.3 Algorithme

$$
\begin{array}{l}
\left[ u_{11} = a_{11} \right. \\
\left[ \begin{array}{l} j = 2 \dots n \\ \left[ \begin{array}{l} u_{1j} = a_{1j} \\ \ell_{j1} = \dfrac{a_{j1}}{a_{11}} \end{array} \right. \end{array} \right. \\
\left[ \begin{array}{l} i = 2 \dots n-1 \\ u_{ii} = a_{ii} - \displaystyle\sum_{k=1}^{i-1} \ell_{ik}\, u_{ki} \\ \left[ \begin{array}{l} j = i+1 \dots n \\ \left[ \begin{array}{l} u_{ij} = a_{ij} - \displaystyle\sum_{k=1}^{i-1} \ell_{ik}\, u_{kj} \\ \ell_{ji} = \dfrac{1}{u_{ii}} \left( a_{ji} - \displaystyle\sum_{k=1}^{i-1} \ell_{jk}\, u_{ki} \right) \end{array} \right. \end{array} \right. \end{array} \right. \\
\left[ u_{nn} = a_{nn} - \displaystyle\sum_{k=1}^{n-1} \ell_{nk}\, u_{kn} \right.
\end{array}
$$

**Coût de calcul :**

divisions : $\displaystyle\sum_{j=2}^{n} 1 + \sum_{i=2}^{n-1} \sum_{j=i+1}^{n} 1 = \frac{n(n-1)}{2}$

additions : $\displaystyle\sum_{k=1}^{n-1} 1 + \sum_{i=2}^{n-1} \left[ \sum_{k=1}^{i-1} 1 + \sum_{j=i+1}^{n} 2 \sum_{k=1}^{i-1} 1 \right] = \frac{n(n-1)(2n-1)}{6}$

multiplications : $\displaystyle\sum_{k=1}^{n-1} 1 + \sum_{i=2}^{n-1} \left[ \sum_{k=1}^{i-1} 1 + \sum_{j=i+1}^{n} 2 \sum_{k=1}^{i-1} 1 \right] = \frac{n(n-1)(2n-1)}{6}$

total : $\frac{n}{6}\left(4n^2 - 3n - 1\right) \approx \frac{2}{3}n^3$ si $n \gg 1$

### 4.3.4 Résolution de système par décomposition LU

Une fois la factorisation LU effectuée, résoudre le système $\underline{\underline{A}}.\underline{x} = \underline{b}$ revient à résoudre deux systèmes triangulaires :

$$
\begin{cases} \underline{\underline{L}}.\underline{y} = \underline{b} \\ \underline{\underline{U}}.\underline{x} = \underline{y} \end{cases}
$$

Le coût de calcul est le suivant :

1. Décomposition LU :
   — $\frac{n(n-1)}{2}$ divisions
   — $\frac{n(n-1)(2n-1)}{6}$ additions
   — $\frac{n(n-1)(2n-1)}{6}$ multiplications
2. Remontée des deux systèmes linéaires :
   — $1 \times n$ divisions (pas de division pour le système en $\underline{\underline{L}}$ car diagonale de 1)

— $2 \times \frac{n(n-1)}{2}$ additions
— $2 \times \frac{n(n-1)}{2}$ multiplications

L'intérêt de la méthode apparaît lors de la résolution d'une suite de système où seul le second membre change (par exemple calcul de structure avec différents cas de charges). Si on note $p$ le nombre de seconds membre, le coût total est :

— $\frac{n(n-1)}{2} + pn$ divisions
— $\frac{n(n-1)(2n-1)}{6} + pn(n-1)$ additions
— $\frac{n(n-1)(2n-1)}{6} + pn(n-1)$ multiplications

Soit pour $n \gg 1$, $\frac{n^3}{3} + pn^2 \ll \frac{2pn^3}{3}$ si on avait résolu $p$ pivots de Gauss.

**Exercice 4.6.** Résoudre le système $\underline{\underline{A}}.\underline{x} = \underline{b}$ suivant en décomposant $\underline{\underline{A}}$ en LU

$$\underline{\underline{A}} = \begin{bmatrix} 2 & 0 & 3 & 2 \\ -6 & -1 & -7 & -10 \\ 2 & -3 & 10 & -12 \\ 4 & 3 & 2 & 6 \end{bmatrix} \quad ; \quad \underline{b} = \begin{pmatrix} 1 \\ -6 \\ -9 \\ -3 \end{pmatrix}$$

### 4.3.5 Déterminant d'une matrice par décomposition LU

Une fois la décomposition LU d'une matrice $\underline{\underline{A}}$ effectuée, calculer le déterminant de la matrice (a priori pleine) $\underline{\underline{A}}$ revient à calculer le déterminant de la matrice triangulaire $\underline{\underline{U}}$. En effet,

$$\det(\underline{\underline{A}}) = \det(\underline{\underline{L}}) \times \det(\underline{\underline{U}}) \text{ où } \det(\underline{\underline{L}}) = 1$$

Calculer le déterminant de $\underline{\underline{A}}$ équivaut à multiplier les termes diagonaux de $\underline{\underline{U}}$, voir paragraphe 4.2.5

### 4.3.6 Inversion de matrice par décomposition LU

Disposant de la décomposition LU d'une matrice $\underline{\underline{A}}$, il suffit d'inverser $\underline{\underline{L}}$ et $\underline{\underline{U}}$ pour obtenir $\underline{\underline{A}}^{-1} = \underline{\underline{U}}^{-1}.\underline{\underline{L}}^{-1}$. Pour $n \gg 1$, on peut montrer que le coût total est de l'ordre de $n^3$

## 4.4 Factorisations de Crout et de Cholesky

### 4.4.1 Crout et Cholesky comme un cas particulier de LU

Si $\underline{\underline{A}} \in \mathcal{M}_n(\mathbb{R})$ est décomposable en LU et symétrique, alors :

$$\underline{\underline{A}} = \underline{\underline{L}}.\underline{\underline{U}} = \underline{\underline{L}}.\underline{\underline{\Delta}}.\underline{\underline{V}}$$

où $\underline{\underline{\Delta}}$ matrice diagonales des $u_{ii}$ et $\underline{\underline{V}}$ matrice triangulaire supérieure à diagonale unité. Par symétrie de $\underline{\underline{A}}$ on a :

$${}^t\underline{\underline{A}} = {}^t\underline{\underline{V}}.\left(\underline{\underline{\Delta}}.{}^t\underline{\underline{L}}\right) = \underline{\underline{A}} = \underline{\underline{L}}.\underline{\underline{U}}$$

L'unicité de la décomposition en LU implique que $\underline{\underline{L}} = {}^t\underline{\underline{V}}$ et $\underline{\underline{U}} = \underline{\underline{\Delta}}.{}^t\underline{\underline{L}} = \underline{\underline{\Delta}}.\underline{\underline{V}}$. Par conséquent,

$$\underline{\underline{A}} = {}^t\underline{\underline{V}}.\underline{\underline{\Delta}}.\underline{\underline{V}}$$

Cette décomposition est appelée factorisation de Crout.

De plus, si les éléments diagonaux de $\underline{\underline{\Delta}}$ sont strictement positifs, alors on peut écrire $\Delta_{ii} = \delta_{ii}^2$ et former la matrice diagonale $\underline{\underline{\delta}}$. Finalement, on obtient la factorisation :

$$\underline{\underline{A}} = {}^t\underline{\underline{R}}.\underline{\underline{R}} \qquad \text{où} \qquad \underline{\underline{R}} = \underline{\underline{\delta}}.\underline{\underline{V}}$$

et en imposant $r_{ii} = \delta_{ii} > 0$, la décomposition est unique.

### 4.4.2 Condition nécessaire et suffisante de factorisation

**Théorème 4.4** (Factorisation de Crout)**.**
*Si $\underline{\underline{A}} \in \mathcal{M}_n(\mathbb{R})$ est factorisable en $LU$ et si de plus $\underline{\underline{A}}$ est symétrique, alors $\underline{\underline{A}}$ est factorisable sous la forme :*

$$\underline{\underline{A}} = \underline{\underline{L}}.\underline{\underline{\Delta}}.{}^t\underline{\underline{L}}$$

*où $\underline{\underline{\Delta}}$ est la matrice diagonales des $u_{ii}$.*

**Théorème 4.5** (Factorisation de Cholesky)**.**
*Une condition nécessaire et suffisante pour qu'une matrice $\underline{\underline{A}} \in \mathcal{M}_n(\mathbb{R})$ se factorise en ${}^t\underline{\underline{R}}.\underline{\underline{R}}$, $\underline{\underline{R}} = \underline{\underline{\delta}}.{}^t\underline{\underline{L}}$ triangulaire supérieure à éléments diagonaux strictement positifs est que $\underline{\underline{A}}$ soit symétrique définie positive.*

**Preuve.** *1) Condition nécessaire :*
*Soit $\underline{x} \neq \underline{0}$. Alors ${}^t\underline{x}.\underline{\underline{A}}.\underline{x} = {}^t\left(\underline{\underline{R}}.\underline{x}\right).\left(\underline{\underline{R}}.\underline{x}\right) = \|\underline{\underline{R}}.\underline{x}\|^2 > 0$ car $\underline{\underline{R}}$ étant à diagonale strictement positive, $\|\underline{\underline{R}}.\underline{x}\|$ ne peut s'annuler.*

*2) Condition suffisante :*
*$\underline{\underline{A}}$ est symétrique définie positive ssi $\forall k \in [\![1;n]\!]$, $D_k > 0$. Ainsi $\Delta_{ii} > 0$, ce qui prouve l'unicité de la décomposition.*

**Remarque 4.4.**
Pour montrer que $\underline{\underline{A}}$ est symétrique définie positive on pourra s'appuyer sur les éléments de la section 7.1.5.

### 4.4.3 Algorithme

$$\left[\begin{array}{l} r_{11} = \sqrt{a_{11}} \\ \left[\begin{array}{l} j = 2 \dots n \\ r_{1j} = \dfrac{a_{j1}}{r_{11}} \end{array}\right. \\ \left[\begin{array}{l} i = 2 \dots n-1 \\ r_{ii} = \sqrt{a_{ii} - \displaystyle\sum_{k=1}^{i-1} r_{ki}^2} \\ \left[\begin{array}{l} j = i+1 \dots n \\ r_{ij} = \dfrac{1}{r_{ii}} \left( a_{ij} - \displaystyle\sum_{k=1}^{i-1} r_{ki}\, r_{kj} \right) \end{array}\right. \end{array}\right. \\ \left[ r_{nn} = \sqrt{a_{nn} - \displaystyle\sum_{k=1}^{n-1} r_{kn}^2} \right. \end{array}\right.$$

**Coût de calcul :**

divisions :
$$\sum_{j=2}^{n} 1 + \sum_{i=2}^{n-1} \sum_{j=i+1}^{n} 1 = \frac{n(n-1)}{2}$$

additions :
$$\sum_{k=1}^{n-1} 1 + \sum_{i=2}^{n-1} \left[ \sum_{k=1}^{i-1} 1 + \sum_{j=i+1}^{n} \sum_{k=1}^{i-1} 1 \right] = \frac{n(n^2-1)}{6}$$

multiplications :
$$\sum_{k=1}^{n-1} 1 + \sum_{i=2}^{n-1} \left[ \sum_{k=1}^{i-1} 1 + \sum_{j=i+1}^{n} \sum_{k=1}^{i-1} 1 \right] = \frac{n(n^2-1)}{6}$$

racines :
$$1 + \sum_{k=2}^{n-1} 1 + 1 = n$$

total :
$$\frac{n}{6}\left(2n^2 + 3n + 1\right) \approx \frac{n^3}{3} \text{ si } n \gg 1$$

### 4.4.4 Résolution de système par factorisation de Cholesky

Une fois la factorisation effectuée, résoudre le système $\underline{\underline{A}}.\underline{x} = \underline{b}$ revient à résoudre deux systèmes triangulaires :

$$\begin{cases} {}^t\underline{\underline{R}}.\underline{y} = \underline{b} \\ \underline{\underline{R}}.\underline{x} = \underline{y} \end{cases}$$

Le coût de calcul est le suivant :

1. Décomposition Cholesky :
   — $\frac{n(n-1)}{2}$ divisions
   — $\frac{n(n^2-1)}{6}$ additions
   — $\frac{n(n^2-1)}{6}$ multiplications
   — $n$ racines
2. Remontée des deux systèmes linéaires :
   — $2 \times n$ divisions
   — $2 \times \frac{n(n-1)}{2}$ additions
   — $2 \times \frac{n(n-1)}{2}$ multiplications

Le coût de calcul est proche de celui de Gauss pour une matrice symétrique (avec $n$ racines carrées et $n$ division en plus). L'avantage est l'introduction de racines carrées qui améliorent le comportement numérique (atténuation de la propagation des arrondis).

### 4.4.5 Exercices

**Exercice 4.7.** Résoudre par la méthode de Cholesky le système suivant :

$$\underline{\underline{A}}.\underline{x} = \underline{b} \text{ où } \underline{\underline{A}} = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix} \text{ et } \underline{b} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

**Exercice 4.8.** Soit le système défini par :

$$\underline{\underline{A}}.\underline{x} = \underline{b} \quad \text{où } \underline{\underline{A}} = \begin{bmatrix} 4 & 8 & 2 & 4 \\ 8 & 17 & 4 & 10 \\ 2 & 4 & 10 & 5 \\ 4 & 10 & 5 & 10 \end{bmatrix} ; \underline{b} = \begin{pmatrix} 2 \\ 6 \\ -2 \\ 7 \end{pmatrix}$$

Choisir la méthode qui semble la plus adaptée et résoudre le système.

## 4.5 Résolution de systèmes linéaires sur-déterminés par la méthode des moindres carrés

### 4.5.1 Systèmes linéaires sur-déterminé

Soient $(n, m) \in \mathbb{N}^\star \times \mathbb{N}^\star$ tels que $m > n$, $\underline{\underline{A}} \in \mathcal{M}_{mn}(\mathbb{R})$ et $\underline{b} \in \mathbb{R}^m$ connus. On cherche $\underline{x} \in \mathbb{R}^n$ tel que $\underline{\underline{A}}.\underline{x} = \underline{b}$ Ce système sur-déterminé n'admet pas de solution exacte dans le cas général. Le principe est de trouver une solution approchée $\tilde{\underline{x}}$ tel que la norme de l'erreur de reconstruction de $\underline{b}$ sous la forme $\underline{\underline{A}}.\underline{x}$ soit la plus faible possible.

Plaçons nous dans l'espace $\mathbb{R}^n$, qui est un espace de Hilbert muni du produit scalaire classique

$$\forall (\underline{x}, \underline{y}) \in \mathbb{R}^n \times \mathbb{R}^n, \langle \underline{x}, \underline{y} \rangle = {}^t \underline{x}.\underline{y}$$

dont la norme induite ($\|.\|_2$) a l'avantage de pouvoir être interprétée physiquement comme une énergie. Le problème s'écrit donc :

$$\frac{1}{2}\|\underline{\underline{A}}.\tilde{\underline{x}} - \underline{b}\|_2^2 = \min_{\underline{x} \in \mathbb{R}^n} \frac{1}{2} \|\underline{\underline{A}}.\underline{x} - \underline{b}\|_2^2 = \min_{\underline{x} \in \mathbb{R}^n} \varepsilon(\underline{x}) \tag{4.6}$$

Le théorème de projection orthogonale (voir cours d'Analyse) sur $\mathbb{R}^n$, sous-espace vectoriel fermé de $\mathbb{R}^m$, assure l'existence et l'unicité d'un tel $\tilde{\underline{x}}$. Ce minimum est atteint lorsque les dérivées partielles de $\varepsilon(\underline{x})$ par rapport à chaque composante $x_\ell$, $1 \leq \ell \leq n$ s'annulent. Ainsi, en utilisant les notations d'Einstein :

$$2\varepsilon(\underline{x}) = (A_{ij}\, x_j - b_i)\,(A_{ik}\, x_k - b_i) = -2\, b_i\, A_{ij}\, x_j + A_{ij}\, A_{ik}\, x_j\, x_k + b_i^2$$

et :

$$\forall \ell \in [\![1;n]\!], \quad \partial_{x_\ell}\varepsilon = \frac{1}{2}\left[-2\,b_i\,A_{ij}\,\delta_{j\ell} + A_{ij}\,A_{ik}\,(\delta_{j\ell}\,x_k + \delta_{k\ell}\,x_j)\right] = 0 = -b_i\,A_{i\ell} + A_{ij}\,A_{i\ell}\,x_j$$

La solution approchée du problème sur-déterminé est donc solution exacte du système carré suivant :

$$\left({}^t\underline{\underline{A}}.\underline{\underline{A}}\right).\underline{x} = {}^t\,\underline{\underline{A}}.\underline{b}$$

Si la matrice $\underline{\underline{A}} \in \mathcal{M}_{mn}(\mathbb{R})$ est de rang $n$, alors ${}^t\underline{\underline{A}}.\underline{\underline{A}}$ est inversible :

**Preuve.** *Il suffit de vérifier que* $\forall \underline{x} \in \mathbb{R}^n$, ${}^t\underline{\underline{A}}.\underline{\underline{A}}.\underline{x} = \underline{0} \Rightarrow \underline{x} = \underline{0}$. *En effet dans ce cas* $\|\underline{\underline{A}}.\underline{x}\|_2 = {}^t\,\underline{x}.{}^t\underline{\underline{A}}.\underline{\underline{A}}.\underline{x} = 0 \Rightarrow \underline{\underline{A}}.\underline{x} = \underline{0}$. *Ainsi* $\underline{x} \in \ker(\underline{\underline{A}})$. *Or d'après le théorème du rang,* $n = \dim(\ker(\underline{\underline{A}})) + \mathrm{rang}(\underline{\underline{A}}) = \dim(\ker(\underline{\underline{A}})) + n$ *donc* $\ker(\underline{\underline{A}}) = \{0\}$ *et* $\underline{x} = \underline{0}$

Dans ce cas, $\left({}^t\underline{\underline{A}}.\underline{\underline{A}}\right)^{-1}.{}^t\underline{\underline{A}}$ est appelée pseudo-inverse et la solution au problème (4.6) s'écrit :

$$\underline{\tilde{x}} = \left({}^t\underline{\underline{A}}.\underline{\underline{A}}\right)^{-1}.{}^t\underline{\underline{A}}.\underline{b} \tag{4.7}$$

### 4.5.2 Exemple 1 : Régression linéaire

En science expérimentale, dans de nombreux cas on cherche à approcher un nuage de points par une fonction représentant un modèle (lien cours Méthode Statistiques pour l'Ingénieur 2A). Supposons que nous disposons de $n$ couples $(x_i,\, y_i)$ issues de $n$ observations indépendantes. De plus, on suppose que nous cherchons une fonction de régression linéaire, de telle sorte que pour chaque point $1 \leq i \leq n$ :

$$y_i = a\,x_i + b + \varepsilon_i$$

où les $\varepsilon_i$ sont des réalisations indépendantes d'une variable $\varepsilon$ d'espérance nulle et de variance constante pour tous les $x_i$. Trouver ces coefficients $a$ et $b$ est le problème de minimisation suivant :

$$\min_{(a,b)\in\mathbb{R}^2} \frac{1}{2}\sum_{i=1}^{n}\left(y_i - (a\,x_i + b)\right)^2 \tag{4.8}$$

Pour retrouver le formalisme du paragraphe précédent, posons les notations suivantes :

$$\underline{\underline{M}} = \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix} \quad ; \quad \underline{C} = \begin{pmatrix} a \\ b \end{pmatrix} \quad ; \quad \underline{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad \text{tels que} \quad (4.8) \Leftrightarrow \frac{1}{2}\|\underline{\underline{M}}.\underline{\tilde{C}} - \underline{Y}\|_2^2 = \min_{\underline{C}\in\mathbb{R}^2}\frac{1}{2}\,\|\underline{\underline{M}}.\underline{C} - \underline{Y}\|_2^2$$

La solution est alors donnée par (4.7), dont l'expression des coefficients est (où $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$) :

$$\tilde{a} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\mathrm{Cov}(x,y)}{\mathrm{Var}(x)} \qquad ; \qquad \tilde{b} = \bar{y} - \tilde{a}\bar{x}$$

Par conséquent, la fonction de régression linéaire a pour expression :

$$y(x) = \bar{y} + \frac{\mathrm{Cov}(x,y)}{\mathrm{Var}(x)}\,(x - \bar{x})$$

En utilisant le formaliste des moindres carrés, il faut résoudre le système :

$$\underline{\tilde{C}} = \left({}^t\underline{\underline{M}}.\underline{\underline{M}}\right)^{-1}.{}^t\underline{\underline{M}}.\underline{Y}$$

**Exercice 4.9.** Le fichier data.txt fourni contient des données issues d'observations expérimentales structurées en deux colonnes, la première correspondant aux abscisses et la seconde aux ordonnées. Déterminer, par la méthode des moindres carrés, les courbes de régression linéaire, quadratique et cubique et comparer les résidus.

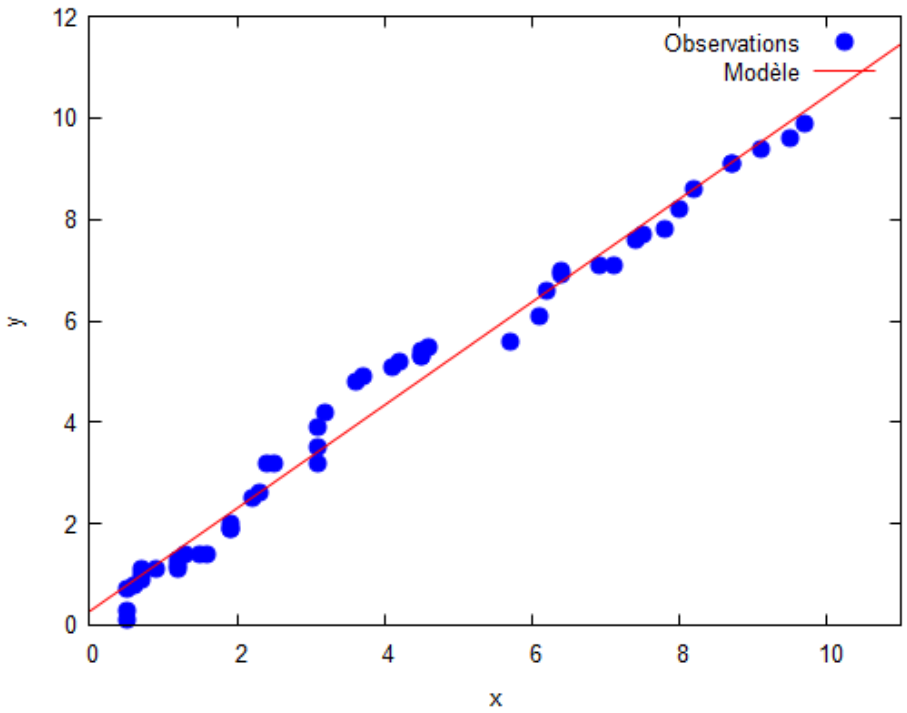



FIGURE 4.1 – Exemple de régression linéaire

### 4.5.3 Exemple 2 : calcul de convergence en tunnel

Le creusement d'un ouvrage souterrain dans le terrain environnant induit une modification du champ de contraintes initiales qui est associé à un champ de déplacement, dont l'intensité dépend des caractéristiques mécaniques du terrain et des contraintes initiales. On parle de convergence du terrain lorsque ce dernier exerce des efforts de poussée sur l'ouvrage souterrain. Dans ce cas, des mesures de convergence sont effectuées durant la vie de l'ouvrage afin de s'assurer de sa non dégradation.

Afin de comprendre les mécanismes de convergence d'une section donnée, des plots de convergence sont scellés sur la voûte du tunnel, et les distances (appelées *cordes*) entre ces plots sont mesurées. Sauf si un topographe a déterminé les emplacements initiaux des plots, aucune information n'est disponible sur les coordonnées initiales des plots. L'objectif que nous nous fixons est de calculer les coordonnées (relatives) de ces plots dans le plan formé par la section du tunnel.

Si la section est instrumentée par $N$ plots, $2N$ coordonnées sont à déterminer. Comme les coordonnées sont relatives, nous fixons un plot comme origine et l'ordonnée d'un autre plot afin de former l'axe des abscisses : il ne reste donc plus que $2N-3$ coordonnées à déterminer. Soit $M$ le nombre de mesures de cordes entre plots. Au maximum, $M_{\max} = \frac{N(N-1)}{2} \geq 2N-3$ mesures peuvent être effectuées. Sauf qu'il est fréquent que toutes les cordes ne soient pas mesurées, par exemple s'il y a un obstacle entre plots, ou si un appareil est défectueux. Dès lors que le nombre de mesures est strictement supérieur au nombre d'inconnues, i.e $M > 2N-3$, le système est dit surdéterminé et n'admet pas de solution exacte. La méthode des moindres carrés consiste à trouver une solution approchée qui minimise l'erreur.

Le premier travail consiste à initialiser le problème, i.e à calculer une solution au système "carré" équivalent (i.e en considérant $M = 2N-3$ mesures) afin de calculer un premier jeu de coordonnées, notées avec un $^\star$. On peut alors calculer une distance entre deux plots $I$ et $J$, notée $d^\star_{IJ}$. Soit $m_{IJ}$ la mesure de la corde entre les plots $I$ et $J$. Par définition, en notant $\underline{X}_{IJ} = (x_I, y_I, x_J, y_J)$ le vecteur dont les composantes sont

les coordonnées des points $I$ et $J$ :

$$m_{IJ} = \sqrt{(x_I - x_J)^2 + (y_I - y_J)^2} = f(\underline{X}_{IJ}) \quad ; \quad d^\star_{IJ} = f(\underline{\tilde{x}}_{IJ}) \tag{4.9}$$

La fonction $f$ représente clairement la distance "pythagoricienne", i.e la norme $\|.\|_2$ entre les coordonnées des points $I$ et $J$. Ainsi $f$ donne la forme selon laquelle nous voulons ajuster les coordonnées des plots sous la contrainte des mesures de cordes. Le développement en séries de Taylor au premier ordre de $f$ - équation (4.9)- autour de $\underline{\tilde{x}}_{IJ}$ s'écrit :

$$m_{IJ} = d^\star_{IJ} +^T \underline{\nabla} f(\underline{\tilde{x}}_{IJ}).\,(\underline{X}_{IJ} - \underline{\tilde{x}}_{IJ}) + r_{IJ} =^T \underline{\nabla} f(\underline{\tilde{x}}_{IJ}).\underline{X}_{IJ} + r_{IJ} \quad \text{avec} \quad \begin{cases} \underline{\nabla} f(\underline{\tilde{x}}_{IJ}) = \dfrac{1}{d^\star_{IJ}} \begin{pmatrix} x^\star_I - x^\star_J \\ y^\star_I - y^\star_J \\ -(x^\star_I - x^\star_J) \\ -(y^\star_I - y^\star_J) \end{pmatrix} \\ d^\star_{IJ} =^T \underline{\nabla} f(\underline{\tilde{x}}_{IJ}).\underline{\tilde{x}}_{IJ} \end{cases}$$

où $r_{IJ}$ est le résidu du développement en séries. Ce développement au premier ordre a permis de linéariser la fonction $f$, non linéaire. Ainsi, en concaténant les $M$ mesures dans le vecteur $\underline{y} \in \mathbb{R}^M$, le problème devient :

$$\text{"Trouver } \underline{X} \in \mathbb{R}^{2N-3} \text{ tel que : } \underline{y} = \underline{\underline{J}}^\star.\underline{X} + \underline{r}\text{"}$$

où $J^\star$ est la jacobienne de $f$ appliquée aux coordonnées $\underline{\tilde{x}}$ et $\underline{r}$ le vecteur des résidus $r_{IJ}$. La méthode des moindres carrés appliquée à ce problème consiste à minimiser la norme du vecteur résidu, et fournit l'approximation suivante :

$$\underline{\tilde{X}} = \left({}^t\underline{\underline{J}}^\star.\underline{\underline{J}}^\star\right)^{-1}.{}^t\underline{\underline{J}}^\star.\underline{y}$$

**Exemple 4.5.** Soit un profil comportant $N = 5$ plots (cas très courant), dont les mesures de cordes sont fournies dans le tableau suivant :

| Corde | 1-2 | 1-3 | 1-4 | 1-5 | 2-3 | 2-4 | 2-5 | 3-4 | 3-5 | 4-5 |
|---|---|---|---|---|---|---|---|---|---|---|
| Longueur (mm) | 3561 | 8755 | 11251 | 11550 | 5987 | 9326 | 10551 | 4058 | 6605 | 3210 |

On notera $\underline{y}$ le vecteur "données" regroupant ces valeurs de cordes, tel que :

$$\underline{y} = \begin{bmatrix} 3561 \\ 8755 \\ 11251 \\ 11550 \\ 5987 \\ 9326 \\ 10551 \\ 4058 \\ 6605 \\ 3210 \end{bmatrix}$$

**Etape 1 : Déterminer un premier jeu de coordonnées**

Fixons arbitrairement le point 1 comme origine du repère, et l'abscisse passant par le point 5. En considérant sept mesures judicieusement, i.e considérant les cinq cordes liées aux points 1 et les trois restantes liées au point 5 on peut initialiser le calcul des coordonnées :

$$\begin{cases} x^\star_1 = y^\star_1 = y^\star_5 = 0 \\ x^\star_5 = m_{15} \\ \forall i \in [\![2;4]\!], x^\star_i = \dfrac{m^2_{1i} - m^2_{i5} + m^2_{15}}{2\,m_{15}} \quad ; \quad y^\star_i = \sqrt{m^2_{1i} - (x^\star_i)^2} \end{cases}$$

Numériquement,

$$\underline{X}^\star = \begin{pmatrix} x_2^\star \\ y_2^\star \\ x_3^\star \\ y_3^\star \\ x_4^\star \\ y_4^\star \\ x_5^\star \end{pmatrix} = \begin{pmatrix} 1504.75 \\ 3227.45 \\ 7204.61 \\ 4974.30 \\ 10808.81 \\ 3123.26 \\ 11550 \end{pmatrix} \quad ; \quad \underline{d}^\star = \begin{pmatrix} d_{12}^\star \\ d_{13}^\star \\ d_{14}^\star \\ d_{15}^\star \\ d_{23}^\star \\ d_{24}^\star \\ d_{25}^\star \\ d_{34}^\star \\ d_{35}^\star \\ d_{45}^\star \end{pmatrix} = \begin{pmatrix} 3561.0 \\ 8755.0 \\ 11251.0 \\ 11550.0 \\ 5961.54 \\ 9304.64 \\ 10551.0 \\ 4051.74 \\ 6605.0 \\ 3210.0 \end{pmatrix} = \begin{pmatrix} m_{12} \\ m_{13} \\ m_{14} \\ m_{15} \\ m_{23} \\ m_{24} \\ m_{25} \\ m_{34} \\ m_{35} \\ m_{45} \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ -25.46 \\ -21.36 \\ 0 \\ -6.26 \\ 0 \\ 0 \end{pmatrix}$$

**Remarque 4.5.**

Dans certains cas moins favorables peuvent apparaître des choix de racines carrées, toutes deux positives. Un bon moyen d'automatiser le choix d'une de ces racines est de respecter la concavité de la voûte du tunnel, i.e que tous les points doivent vérifier la négativité de la "dérivée seconde", que l'on peut exprimer sous la forme de taux d'accroissements successifs (différences divisées) :

$$\frac{\frac{y_{i-1}-y_i}{x_{i-1}-x_i} - \frac{y_i-y_{i+1}}{x_i-x_{i+1}}}{x_{i-1}-x_{i+1}} < 0$$

**Etape 2 : Calcul de la jacobienne**

Les expressions des composantes de la jacobienne sont les suivantes :

$$\underline{\underline{J}}^\star = \begin{bmatrix} -\frac{x_1^\star-x_2^\star}{d_{12}^\star} & -\frac{y_1^\star-y_2^\star}{d_{12}^\star} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -\frac{x_1^\star-x_3^\star}{d_{13}^\star} & -\frac{y_1^\star-y_3^\star}{d_{13}^\star} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -\frac{x_1^\star-x_4^\star}{d_{14}^\star} & -\frac{y_1^\star-y_4^\star}{d_{14}^\star} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -\frac{x_1^\star-x_5^\star}{d_{15}^\star} \\ \frac{x_2^\star-x_3^\star}{d_{23}^\star} & \frac{y_2^\star-y_3^\star}{d_{23}^\star} & -\frac{x_2^\star-x_3^\star}{d_{23}^\star} & -\frac{y_2^\star-y_3^\star}{d_{23}^\star} & 0 & 0 & 0 \\ \frac{x_2^\star-x_4^\star}{d_{24}^\star} & \frac{y_2^\star-y_4^\star}{d_{24}^\star} & 0 & 0 & -\frac{x_2^\star-x_4^\star}{d_{24}^\star} & -\frac{y_2^\star-y_4^\star}{d_{24}^\star} & 0 \\ \frac{x_2^\star-x_5^\star}{d_{25}^\star} & \frac{y_2^\star-y_5^\star}{d_{25}^\star} & 0 & 0 & 0 & 0 & -\frac{x_2^\star-x_5^\star}{d_{25}^\star} \\ 0 & 0 & \frac{x_3^\star-x_4^\star}{d_{34}^\star} & \frac{y_3^\star-y_4^\star}{d_{34}^\star} & -\frac{x_3^\star-x_4^\star}{d_{34}^\star} & -\frac{y_3^\star-y_4^\star}{d_{34}^\star} & 0 \\ 0 & 0 & \frac{x_3^\star-x_5^\star}{d_{35}^\star} & \frac{y_3^\star-y_5^\star}{d_{35}^\star} & 0 & 0 & -\frac{x_3^\star-x_5^\star}{d_{35}^\star} \\ 0 & 0 & 0 & 0 & \frac{x_4^\star-x_5^\star}{d_{45}^\star} & \frac{y_4^\star-y_5^\star}{d_{45}^\star} & -\frac{x_4^\star-x_5^\star}{d_{45}^\star} \end{bmatrix}$$

$$\underline{\underline{J}}^\star = \begin{bmatrix} 0.42 & 0.91 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.82 & 0.57 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.96 & 0.28 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 1.0 \\ -0.96 & -0.29 & 0.96 & 0.29 & 0.0 & 0.0 & 0.0 \\ -1.0 & 0.011 & 0.0 & 0.0 & 1.0 & -0.011 & 0.0 \\ -0.95 & 0.31 & 0.0 & 0.0 & 0.0 & 0.0 & 0.95 \\ 0.0 & 0.0 & -0.89 & 0.46 & 0.89 & -0.46 & 0.0 \\ 0.0 & 0.0 & -0.66 & 0.75 & 0.0 & 0.0 & 0.66 \\ 0.0 & 0.0 & 0.0 & 0.0 & -0.23 & 0.97 & 0.23 \end{bmatrix}$$

et

$$ {}^t\underline{\underline{J^\star}}.\underline{\underline{J^\star}} = \begin{bmatrix} 3.0 & 0.36 & -0.91 & -0.28 & -1.0 & 0.011 & -0.91 \\ 0.36 & 1.0 & -0.28 & -0.086 & 0.011 & -1.310^{-4} & 0.29 \\ -0.91 & -0.28 & 2.8 & -0.15 & -0.79 & 0.41 & -0.43 \\ -0.28 & -0.086 & -0.15 & 1.2 & 0.41 & -0.21 & 0.5 \\ -1.0 & 0.011 & -0.79 & 0.41 & 2.8 & -0.38 & -0.053 \\ 0.011 & -1.310^{-4} & 0.41 & -0.21 & -0.38 & 1.2 & 0.22 \\ -0.91 & 0.29 & -0.43 & 0.5 & -0.053 & 0.22 & 2.4 \end{bmatrix} $$

**Etape 3 : Résolution du système linéaire**

Soit le système carré à résoudre :

$$ {}^t\underline{\underline{J^\star}}.\underline{\underline{J^\star}}.\underline{\tilde{X}} = {}^t\underline{\underline{J^\star}}.\underline{y} \tag{4.10} $$

Le conditionnement de la matrice ${}^t\underline{\underline{J^\star}}.\underline{\underline{J^\star}}$ est de l'ordre de 18.0, le système (4.10) est donc bien conditionné. Nous proposons de résoudre ce système par la méthode LU puis par la méthode de Cholesky.

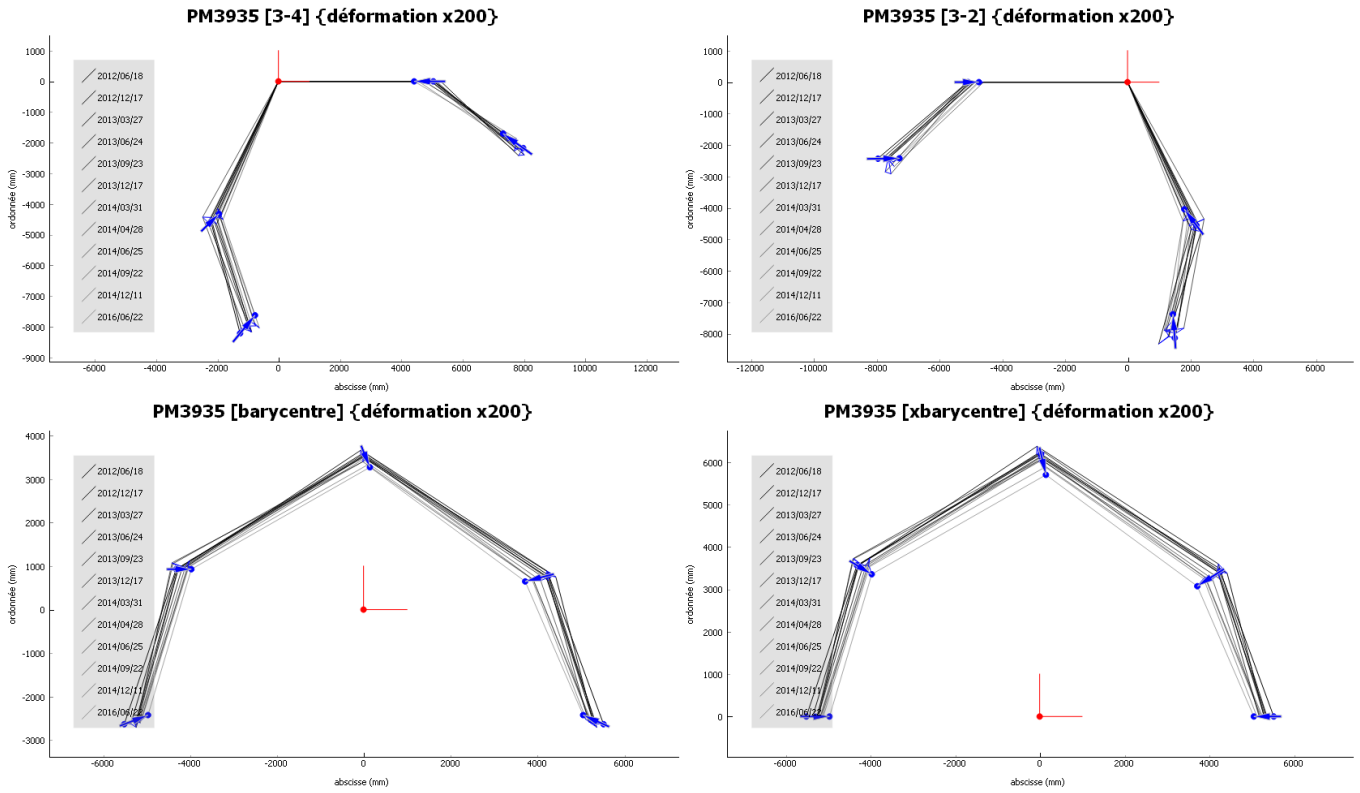



FIGURE 4.2 – Exemple de suivi de convergences

## 4.6 Exercice de synthèse

Soit le système linéaire $(S)\ \underline{\underline{A}}.\underline{x} = \underline{b}$ défini par

$$\underline{\underline{A}} = \begin{bmatrix} 3 & -1 & 0 & 0 \\ -1 & 3 & -1 & 0 \\ 0 & -1 & 3 & -1 \\ 0 & 0 & -1 & 3 \end{bmatrix} \quad ; \quad \underline{b} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

Q1/ Montrer que $\underline{\underline{A}}$ est inversible.

Q2/ Résoudre le système $(S)$ par pivot de Gauss.

Q3/ Montrer que $\underline{\underline{A}}$ est décomposable en LU puis calculer les matrices $\underline{\underline{L}}$ et $\underline{\underline{U}}$. En déduire le déterminant de $\underline{\underline{A}}$, puis résoudre $(S)$.

Q4/ Montrer que la factorisation de Crout de $\underline{\underline{A}}$ existe. Déduire de la décomposition LU de $\underline{\underline{A}}$ la factorisation de Crout de $\underline{\underline{A}}$.

Q5/ Montrer que la factorisation de Choleski de $\underline{\underline{A}}$ existe. Déduire de la décomposition Crout de $\underline{\underline{A}}$ la factorisation de Choleski de $\underline{\underline{A}}$. Résoudre $(S)$ par cette méthode.

# Chapitre 5

# Méthodes itératives pour la résolution des systèmes linéaires

*Lorsque le coût de calcul de la résolution exacte d'un système linéaire devient trop important, il peut être intéressant de recourir à des méthodes approchées qui convergent globalement par itérations successives vers la solution théorique du système. Deux grandes familles de méthodes itératives de résolution de sytèmes linéaires sont présentées : les méthodes d'éclatement, puis les méthodes de gradient. La dernière partie est consacrée au préconditionnement de matrice pour la méthode de gradient conjugué, permettant d'améliorer un mauvais conditionnement de la matrice du système linéaire.*

**Sommaire**

## 5.1 Méthodes d'éclatement classiques

### 5.1.1 Principe des méthodes itératives de résolution de systèmes linéaires

Soit le problème $(\mathcal{P})$ : "Trouver $\underline{x} \in \mathbb{K}^n$ tel que $\underline{\underline{A}}.\underline{x} = \underline{b}$, où $\underline{\underline{A}} \in \mathcal{M}_n(\mathbb{K})$, $\det(\underline{\underline{A}}) \neq 0$, $\underline{b} \in \mathbb{K}^n$". Ce problème admet une unique solution notée $\underline{x}^\star \in \mathbb{K}^n$.

**Définition 5.1.**
Une méthode itérative pour résoudre un tel problème consiste à créer une suite $\left(\underline{x}^{(k)}\right)_{k\in\mathbb{N}^\star}$ telle que :

$$\begin{cases} \underline{x}^{(0)} \in \mathbb{K}^n \\ \underline{x}^{(k+1)} = \underline{f}\left(\underline{x}^{(k)}\right) = \underline{\underline{C}}.\underline{x}^{(k)} + \underline{d} \end{cases} \quad \text{où } \underline{\underline{C}} \in \mathcal{M}_n(\mathbb{K}),\ \underline{d} \in \mathbb{K}^n.$$

et $\lim\limits_{k\to\infty} \underline{x}^{(k)} = \underline{x}^\star$. La fonction $\underline{f}$ et la matrice $\underline{\underline{C}}$ sont respectivement appelées fonction et matrice d'itération.

**Définition 5.2.**
Le principe des méthodes d'éclatement est de définir deux matrices $\left(\underline{\underline{M}}, \underline{\underline{N}}\right) \in \mathcal{M}_n(\mathbb{K}) \times \mathcal{M}_n(\mathbb{K})$ telles que $\underline{\underline{A}} = \underline{\underline{M}} - \underline{\underline{N}}$ où $\det(\underline{\underline{M}}) \neq 0$. Ainsi :

$$\underline{\underline{A}}.\underline{x} = \underline{b} \Leftrightarrow \underline{\underline{M}}.\underline{x} = \underline{\underline{N}}.\underline{x} + \underline{b} \Leftrightarrow \underline{x} = \underline{\underline{M}}^{-1}.\underline{\underline{N}}.\underline{x} + \underline{\underline{M}}^{-1}.\underline{b} = \underline{\underline{C}}.\underline{x} + \underline{d} = \underline{f}(\underline{x})$$

Chercher la solution du système linéaire est équivalent à chercher les points fixes de la fonction d'itération. De la définition de l'éclatement de $\underline{\underline{A}}$ va dépendre la définition de la méthode itérative de résolution. De plus, on introduit les matrices :

- $\underline{\underline{D}} \in \mathcal{M}_n(\mathbb{K})$ matrice diagonale ;
- $\underline{\underline{E}} \in \mathcal{M}_n(\mathbb{K})$ matrice triangulaire inférieure ;
- $\underline{\underline{F}} \in \mathcal{M}_n(\mathbb{K})$ matrice triangulaire supérieure

telles que : $\underline{\underline{A}} = \underline{\underline{D}} - \underline{\underline{E}} - \underline{\underline{F}} = \begin{bmatrix} \ddots & & -\underline{\underline{F}} \\ & \underline{\underline{D}} & \\ -\underline{\underline{E}} & & \ddots \end{bmatrix}$

### 5.1.2 Méthode de Jacobi

**Définition 5.3.**
Soit $\underline{\underline{A}} \in \mathcal{M}_n(\mathbb{K})$ de coefficients $(a_{ij})_{1\leq i,j\leq n}$ tels que $\forall i \in [\![1;n]\!]$, $a_{ii} \neq 0$. La méthode de Jacobi consiste à définir l'éclatement de $\underline{\underline{A}}$ tel que $\underline{\underline{M}} = \underline{\underline{D}}$ et $\underline{\underline{N}} = \underline{\underline{E}} + \underline{\underline{F}}$. Ainsi la méthode itérative est la suivante :

$$\underline{f}_J(\underline{x}) = \underline{\underline{C}}_J.\underline{x} + \underline{d}_J \quad \text{où} \quad \begin{cases} \underline{\underline{C}}_J = \underline{\underline{M}}^{-1}.\underline{\underline{N}} = \underline{\underline{D}}^{-1}.\left(\underline{\underline{E}} + \underline{\underline{F}}\right) = \underline{\underline{I}} - \underline{\underline{D}}^{-1}.\underline{\underline{A}} \\ \underline{d}_J = \underline{\underline{M}}^{-1}.\underline{b} = \underline{\underline{D}}^{-1}.\underline{b} \end{cases}$$

La matrice d'itération $\underline{\underline{C}}_J$, notée aussi $\mathcal{J}(\underline{\underline{A}})$ est appelée matrice de Jacobi associée à la matrice $\underline{\underline{A}}$. La suite des itérés s'écrit donc :

$$\forall \underline{x}^{(0)} \in \mathbb{K}^n,\ \underline{x}^{(k+1)} = \underline{f}_J(\underline{x}^{(k)}) = \mathcal{J}(\underline{\underline{A}}).\underline{x}^{(k)} + \underline{d}_J$$

soit terme à terme [1] :

$$\begin{cases} \forall \underline{x}^{(0)} \in \mathbb{K}^n \\ \forall i \in [\![1;n]\!],\ x_i^{(k+1)} = \dfrac{1}{a_{ii}}\left(b_i - \displaystyle\sum_{j\neq i} a_{ij}\, x_j^{(k)}\right) \end{cases}$$

### 5.1.3 Méthode de Gauss-Seidel

**Définition 5.4.**

Soit $\underline{\underline{A}} \in \mathcal{M}_n(\mathbb{K})$ de coefficients $(a_{ij})_{1\leq i,j\leq n}$ tels que $\forall i \in [\![1;n]\!]$, $a_{ii} \neq 0$. La méthode de Gauss-Seidel consiste à définir l'éclatement de $\underline{\underline{A}}$ tel que $\underline{\underline{M}} = \underline{\underline{D}} - \underline{\underline{E}}$ et $\underline{\underline{N}} = \underline{\underline{F}}$. Ainsi la méthode itérative est la suivante :

$$\underline{f}_{GS}(\underline{x}) = \underline{\underline{C}}_{GS}.\underline{x} + \underline{d}_{GS} \quad \text{où} \quad \begin{cases} \underline{\underline{C}}_{GS} = \underline{\underline{M}}^{-1}.\underline{\underline{N}} = \left(\underline{\underline{D}} - \underline{\underline{E}}\right)^{-1}.\underline{\underline{F}} \\ \underline{d}_{GS} = \underline{\underline{M}}^{-1}.\underline{b} = \left(\underline{\underline{D}} - \underline{\underline{E}}\right)^{-1}.\underline{b} \end{cases}$$

La matrice d'itération $\underline{\underline{C}}_{GS}$, notée aussi $\mathcal{L}_1(\underline{\underline{A}})$ est appelée matrice de Gauss-Seidel associée à $\underline{\underline{A}}$. La suite des itérés s'écrit donc :

$$\forall \underline{x}^{(0)} \in \mathbb{K}^n,\ \underline{x}^{(k+1)} = \underline{f}_{GS}(\underline{x}^{(k)}) = \mathcal{L}_1(\underline{\underline{A}}).\underline{x}^{(k)} + \underline{d}_{GS} = \left(\underline{\underline{D}} - \underline{\underline{E}}\right)^{-1}.\left(\underline{\underline{F}}.\underline{x}^{(k)} + \underline{b}\right)$$

Ainsi

$$\forall \underline{x}^{(0)} \in \mathbb{K}^n,\ \underline{x}^{(k+1)} = \underline{\underline{D}}^{-1}\left(\underline{b} + \underline{\underline{E}}.\underline{x}^{(k+1)} + \underline{\underline{F}}.\underline{x}^{(k)}\right)$$

soit terme à terme :

$$\begin{cases} \forall \underline{x}^{(0)} \in \mathbb{K}^n \\ \forall i \in [\![1;n]\!],\ x_i^{(k+1)} = \dfrac{1}{a_{ii}}\left(b_i - \displaystyle\sum_{j=1}^{i-1} a_{ij}\, x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}\, x_j^{(k)}\right) \end{cases}$$

**Remarque 5.1.**

Le calcul des composantes doit nécessairement être effectué dans l'ordre croissant des indices, sinon le calcul de $x_i^{(k+1)}$ est impossible.

### 5.1.4 Méthode de la relaxation ou SOR (Successive Over Relaxation)

**Définition 5.5.**

Soit $\underline{\underline{A}} \in \mathcal{M}_n(\mathbb{K})$ de coefficients $(a_{ij})_{1\leq i,j\leq n}$ tels que $\forall i \in [\![1;n]\!]$, $a_{ii} \neq 0$. La méthode SOR consiste à définir l'éclatement de $\underline{\underline{A}}$ tel que :

$$\underline{\underline{M}} = \frac{1}{\omega}\left(\underline{\underline{D}} - \omega\,\underline{\underline{E}}\right) \qquad ; \qquad \underline{\underline{N}} = \frac{1}{\omega}\left((1-\omega)\,\underline{\underline{D}} + \omega\,\underline{\underline{F}}\right)$$

Ainsi la méthode itérative est la suivante :

$$\underline{f}_{SOR}(\underline{x}) = \underline{\underline{C}}_{SOR}.\underline{x} + \underline{d}_{SOR} \quad \text{où} \quad \begin{cases} \underline{\underline{C}}_{SOR} = \underline{\underline{M}}^{-1}.\underline{\underline{N}} = \left(\underline{\underline{D}} - \omega\underline{\underline{E}}\right)^{-1}.\left((1-\omega)\,\underline{\underline{D}} + \omega\,\underline{\underline{F}}\right) \\ \underline{d}_{SOR} = \underline{\underline{M}}^{-1}.\underline{b} = \omega\left(\underline{\underline{D}} - \omega\underline{\underline{E}}\right)^{-1}.\underline{b} \end{cases}$$

La matrice d'itération $\underline{\underline{C}}_{GS}$, notée aussi $\mathcal{L}_\omega(\underline{\underline{A}})$ est appelée matrice SOR associée à $\underline{\underline{A}}$. La suite des itérés s'écrit donc :

$$\forall \underline{x}^{(0)} \in \mathbb{K}^n,\ \underline{x}^{(k+1)} = \underline{f}_{SOR}(\underline{x}^{(k)}) = \mathcal{L}_\omega(\underline{\underline{A}}).\underline{x}^{(k)} + \underline{d}_{SOR} = \left(\underline{\underline{D}} - \omega\underline{\underline{E}}\right)^{-1}.\left[(1-\omega)\,\underline{\underline{D}}.\underline{x}^{(k)} + \omega\,\underline{\underline{F}}.\underline{x}^{(k)} + \omega\,\underline{b}\right]$$

Ainsi

$$\forall \underline{x}^{(0)} \in \mathbb{K}^n,\ \underline{x}^{(k+1)} = (1-\omega)\ \underline{x}^{(k)} + \omega\,\underline{x}_{\mathrm{GS}}^{(k+1)}$$

La méthode SOR est par conséquent une généralisation de la méthode de Gauss-Seidel (avec $\omega = 1$).

### 5.1.5 Etude de convergence

#### Théorèmes généraux

Soit $\underline{\varepsilon}^{(k)} = \underline{x}^{(k)} - \underline{x}^\star$ l'erreur à l'itération $k$, où $\underline{x}^\star$ est la solution du système linéaire, tel que $\underline{\underline{A}}.\underline{x}^\star = \underline{b}$. La suite des itérés $\underline{x}^{(k)}$ converge vers $\underline{x}^\star$ ssi $\underline{\varepsilon}^{(k)}$ tend vers 0. De plus,

$$\underline{\underline{A}}.\underline{x}^\star = \underline{b} \Leftrightarrow \underline{x}^\star = \underline{\underline{C}}.\underline{x}^\star + \underline{d}$$

Ainsi,

$$\forall k \in \mathbb{N},\ \underline{\varepsilon}^{(k+1)} = \underline{x}^{(k+1)} - \underline{x}^\star = \underline{\underline{C}}.\left(\underline{x}^{(k)} - \underline{x}^\star\right) = \underline{\underline{C}}.\underline{\varepsilon}^{(k)}$$

Par récurrence, l'erreur à l'itération $k$ peut donc s'exprimer comme :

$$\underline{\varepsilon}^{(k)} = \underline{\underline{C}}^k.\underline{\varepsilon}^{(0)}$$

**Théorème 5.1.**
*Soit une norme matricielle sous-multiplicative notée $\|.\|$. Si $\|\underline{\underline{C}}\| < 1$ alors pour tout $\underline{x}^{(0)} \in \mathbb{K}^n$ la suite des itérés par la fonction d'itération définie par $\underline{x} \mapsto f(\underline{x}) = \underline{\underline{C}}.\underline{x} + \underline{d}$ converge vers la solution $\underline{x}^\star$ du système linéaire. La convergence est donc globale.*

**Preuve.** *La norme matricielle $\|.\|$ étant sous-multiplicative, il existe donc une norme matricielle associée, que l'on note aussi $\|.\|$. Ainsi,*

$$\|\underline{\varepsilon}^{(k)}\| = \|\underline{\underline{C}}^k.\underline{\varepsilon}^{(0)}\| \leq \|\underline{\underline{C}}^k\|\,\|\underline{\varepsilon}^{(0)}\| \leq \|\underline{\underline{C}}\|^k\,\|\underline{\varepsilon}^{(0)}\|$$

*Comme par hypothèse $\|\underline{\underline{C}}\| < 1$ alors $\lim_{k\to\infty} \|\underline{\underline{C}}\|^k = \underline{0}$ et donc :*

$$\forall \underline{x}^{(0)},\ \lim_{k\to\infty} \|\underline{\varepsilon}^{(k)}\| = \underline{0} \Rightarrow \lim_{k\to\infty} \underline{x}^{(k)} = \underline{x}^\star$$

**Théorème 5.2.**
*Une condition nécessaire et suffisante pour que la méthode itérative définie par la fonction $f(\underline{x}) = \underline{\underline{C}}.\underline{x} + \underline{d}$ converge est que $\rho(\underline{\underline{C}}) < 1$, où $\rho(\underline{\underline{C}})$ est le rayon spectral de la matrice $\underline{\underline{C}}$.*

**Remarque 5.2.**
Le théorème 5.1 est intéressant en pratique, car il est plus souvent facile de vérifier la condition de convergence portant sur la norme d'une matrice que de vérifier la condition de convergence du théorème 5.2 sur le rayon de convergence de cette même matrice. Cependant on verra dans la section suivante que si on note $\|.\|$ une norme matricielle quelconque, on a la relation suivante :

$$\forall \underline{\underline{A}} \in \mathcal{M}_n(\mathbb{K}),\ \rho(\underline{\underline{A}}) \leq \|\underline{\underline{A}}\|$$

Cette relation est vraie particulièrement pour la matrice d'itération $\underline{\underline{C}}$. Il existe des cas pour lesquels on a

$$\rho(\underline{\underline{C}}) < 1 < \|\underline{\underline{C}}\|$$

Dans ces cas, le théorème 5.1 est inutilisable, alors que le théorème 5.2 prouve que la méthode converge globalement.

**Remarque 5.3.**
On rappelle que $\underline{\underline{M}}^{\mathrm{H}}$ désigne la matrice transconjuguée de $\underline{\underline{M}}$, i.e $\underline{\underline{M}}^{\mathrm{H}} = {}^t\,\overline{\underline{\underline{M}}}$ , voir Définition 7.5.

**Théorème 5.3.**
*Soit $\underline{\underline{A}} \in \mathcal{M}_n(\mathbb{K})$ une matrice hermitienne définie positive telle que $\underline{\underline{A}} = \underline{\underline{M}} - \underline{\underline{N}}$ . Si $\underline{\underline{M}}^{\mathrm{H}} + \underline{\underline{N}}$, qui est hermitienne, est aussi définie positive alors la suite des itérés $\underline{x}^{(k+1)} = \underline{\underline{C}}.\underline{x}^{(k)} + \underline{d}$ est convergente.*

**Preuve.** *La matrice $\underline{\underline{A}} \in \mathcal{M}_n(\mathbb{K})$ étant hermitienne définie positive on peut définir un produit scalaire $\langle .,. \rangle_A$ sur $\mathbb{K}^n$ à partir du produit scalaire canonique $(.,.)$, tel que :*

$$\forall (\underline{x}, \underline{y}) \in \mathbb{K}^n \times \mathbb{K}^n, \quad \langle \underline{x}, \underline{y} \rangle_A = \left( \underline{\underline{A}}.\underline{x}, \underline{y} \right) = \left( \underline{\underline{A}}.\underline{x} \right)^{\mathrm{H}} .\underline{y} = {}^t\, \overline{\underline{\underline{A}}.\underline{x}}.\underline{y}$$

*On peut également montrer les égalités suivantes (preuve à titre d'exercice) :*

$$\forall (\underline{x}, \underline{y}) \in \mathbb{K}^n \times \mathbb{K}^n, \forall \underline{\underline{B}} \in \mathcal{M}_n(\mathbb{K}), \quad (\underline{\underline{B}}.\underline{x}, \underline{y}) = (\underline{x}, \underline{\underline{B}}^{\mathrm{H}}.\underline{y}) \; ; \; (\underline{x}, \underline{\underline{B}}.\underline{y}) = (\underline{\underline{B}}^{\mathrm{H}}.\underline{x}, \underline{y})$$

*On en déduit ainsi que ce produit scalaire est bien hermitien car $\underline{\underline{A}} = \underline{\underline{A}}^{\mathrm{H}}$, i.e $A_{ij} = \overline{A}_{ji}$, et donc*

$$\forall (\underline{x}, \underline{y}) \in \mathbb{K}^n \times \mathbb{K}^n, \; \langle \underline{x}, \underline{y} \rangle_A = (\underline{\underline{A}}.\underline{x}, \underline{y}) = (\underline{x}, \underline{\underline{A}}.\underline{y}) = \overline{(\underline{\underline{A}}.\underline{y}, \underline{x})} = \overline{\langle \underline{y}, \underline{x} \rangle_A}$$

*De ce produit scalaire découle la norme vectorielle induite $\|.\|_A$, telle que :*

$$\forall \underline{x} \in \mathbb{K}^n, \quad \|\underline{x}\|_A = \sqrt{\langle \underline{x}, \underline{x} \rangle_A}$$

*et aussi une norme matricielle subordonnée à la norme vectorielle $\|.\|_A$ (notée de la même manière). Montrons que :*

$$\|\underline{\underline{M}}^{-1}.\underline{\underline{N}}\|_A = \|\underline{\underline{C}}\|_A < 1$$

*Premièrement, l'ensemble des $\underline{x} \in \mathbb{K}^n$ tels que $\|\underline{x}\|_A = 1$ est un compact (fermé et borné) de $\mathbb{K}^n$ (de dimension finie). De plus la fonction $f : \begin{cases} \mathbb{K}^n \to \mathbb{R} \\ \underline{x} \mapsto \|\underline{x} - \underline{\underline{M}}^{-1}.\underline{\underline{A}}.\underline{x}\|_A \end{cases}$ est continue. Cette fonction est donc bornée et atteint ses bornes sur ce compact, ainsi :*

$$\|\underline{\underline{M}}^{-1}.\underline{\underline{N}}\|_A = \|\underline{\underline{M}}^{-1}.\left( \underline{\underline{M}} - \underline{\underline{A}} \right)\|_A = \|\underline{\underline{I}} - \underline{\underline{M}}^{-1}.\underline{\underline{A}}\|_A = \sup_{\|\underline{x}\|_A = 1} \|\underline{x} - \underline{\underline{M}}^{-1}.\underline{\underline{A}}.\underline{x}\|_A$$

*Montrons alors que $\sup_{\|\underline{x}\|_A = 1} \|\underline{x} - \underline{\underline{M}}^{-1}.\underline{\underline{A}}.\underline{x}\|_A < 1$.*

*La matrice $\underline{\underline{M}}^{\mathrm{H}} + \underline{\underline{N}}$ est hermitienne car :*

$$\underline{\underline{M}}^{\mathrm{H}} + \underline{\underline{N}} = \underline{\underline{M}}^{\mathrm{H}} + \underline{\underline{M}} - \underline{\underline{A}} = \underline{\underline{M}} + \left( \underline{\underline{M}} - \underline{\underline{A}} \right)^{\mathrm{H}} = \underline{\underline{M}} + \underline{\underline{N}}^{\mathrm{H}} = \left( \underline{\underline{M}}^{\mathrm{H}} + \underline{\underline{N}} \right)^{\mathrm{H}}$$

*De plus, $\underline{\underline{A}}$ étant définie positive elle est inversible et $\underline{\underline{M}}$ l'est par définition. Ainsi pour tout $\underline{x} \in \mathbb{K}^n$ tel que $\|\underline{x}\|_A = 1$ on a $\underline{x} \neq \underline{0}$ et par conséquent le vecteur $\underline{\underline{M}}^{-1}.\underline{\underline{A}}\,\underline{x} \neq \underline{0}$. Par conséquent, $\underline{\underline{M}}^{\mathrm{H}} + \underline{\underline{N}}$ étant définie positive on a en particulier :*

$$\left( \left( \underline{\underline{M}}^{\mathrm{H}} + \underline{\underline{N}} \right) .\underline{\underline{M}}^{-1}.\underline{\underline{A}}.\underline{x}, \underline{\underline{M}}^{-1}.\underline{\underline{A}}.\underline{x} \right) > 0$$

*De plus,*

$$\|\underline{x} - \underline{\underline{M}}^{-1}.\underline{\underline{A}}.\underline{x}\|_A^2 = \left( \underline{\underline{A}}.\underline{x}, \underline{x} \right) - \left( \underline{\underline{A}}.\underline{\underline{M}}^{-1}.\underline{\underline{A}}.\underline{x}, \underline{x} \right) - \left( \underline{\underline{A}}.\underline{x}, \underline{\underline{M}}^{-1}.\underline{\underline{A}}.\underline{x} \right) + \left( \underline{\underline{A}}.\underline{\underline{M}}^{-1}.\underline{\underline{A}}.\underline{x}, \underline{\underline{M}}^{-1}.\underline{\underline{A}}.\underline{x} \right)$$

*Pour les trois premiers termes :*

1. $\left( \underline{\underline{A}}.\underline{x}, \underline{x} \right) = \|\underline{x}\|_A^2 = 1$
2. $\left( \underline{\underline{A}}.\underline{\underline{M}}^{-1}.\underline{\underline{A}}.\underline{x}, \underline{x} \right) = \left( \underline{\underline{M}}^{-1}.\underline{\underline{A}}.\underline{x}, \underline{\underline{A}}.\underline{x} \right) = \left( \underline{\underline{M}}^{-1}.\underline{\underline{A}}.\underline{x}, \underline{\underline{M}}.\underline{\underline{M}}^{-1}.\underline{\underline{A}}.\underline{x} \right) = \left( \underline{\underline{M}}^{\mathrm{H}}.\underline{\underline{M}}^{-1}.\underline{\underline{A}}.\underline{x}, \underline{\underline{M}}^{-1}.\underline{\underline{A}}.\underline{x} \right)$
3. $\left( \underline{\underline{A}}.\underline{x}, \underline{\underline{M}}^{-1}.\underline{\underline{A}}.\underline{x} \right) = \left( \underline{\underline{M}}.\underline{\underline{M}}^{-1}.\underline{\underline{A}}.\underline{x}, \underline{\underline{M}}^{-1}.\underline{\underline{A}}.\underline{x} \right)$

*Ainsi :*

$$\|\underline{x} - \underline{\underline{M}}^{-1}.\underline{\underline{A}}.\underline{x}\|_A^2 = 1 - \left( \left( \underline{\underline{M}}^{\mathrm{H}} + \underline{\underline{M}} - \underline{\underline{A}} \right) .\underline{\underline{M}}^{-1}.\underline{\underline{A}}.\underline{x}, \underline{\underline{M}}^{-1}.\underline{\underline{A}}.\underline{x} \right) = 1 - \left( \left( \underline{\underline{M}}^{\mathrm{H}} + \underline{\underline{N}} \right) .\underline{\underline{M}}^{-1}.\underline{\underline{A}}.\underline{x}, \underline{\underline{M}}^{-1}.\underline{\underline{A}}.\underline{x} \right) < 1$$

*Au final $\|\underline{\underline{C}}\|_A < 1$ et on peut conclure grâce au théorème 5.1 que la méthode est convergente globalement.*

**Théorèmes spécifiques à certaines méthodes**

**Théorème 5.4 (**Condition nécessaire de convergence de la méthode SOR**).**
*Si la méthode SOR converge, alors* $|\omega - 1| < 1$.

**Preuve.**

$$\det(\mathcal{L}_\omega(\underline{\underline{A}})) = \det\left(\left(\underline{\underline{D}} - \omega\underline{\underline{E}}\right)^{-1}.\left((1-\omega)\ \underline{\underline{D}} + \omega\underline{\underline{F}}\right)\right) = \frac{1}{\det(\underline{\underline{D}})}\det\left((1-\omega)^n\ \det(\underline{\underline{D}})\right) = (1-\omega)^n$$

*Soient* $\lambda_i$ *les* $n$ *valeurs propres de la matrice d'itération* $\mathcal{L}_\omega(\underline{\underline{A}})$. *Alors :*

$$\left[\rho(\mathcal{L}_\omega(\underline{\underline{A}}))\right]^n \geq \prod_{i=1}^{n}|\lambda_i| = \left|\prod_{i=1}^{n}\lambda_i\right| = \left|\det(\mathcal{L}_\omega(\underline{\underline{A}}))\right| = |1-\omega|^n$$

*Par conséquent* $|1-\omega| \leq \rho(\mathcal{L}_\omega(\underline{\underline{A}}))$. *De plus si la méthode converge, alors* $\rho(\mathcal{L}_\omega(\underline{\underline{A}})) < 1$, *donc au final* $|1-\omega| \leq \rho(\mathcal{L}_\omega(\underline{\underline{A}})) < 1$

**Théorème 5.5.**
*Si* $\underline{\underline{A}} \in \mathcal{M}_n(\mathbb{K})$ *est hermitienne et définie positive, alors la méthode SOR est convergente si et seulement si* $|\omega - 1| < 1$.

**Preuve.** *Si* $\underline{\underline{A}} = \underline{\underline{D}} - \underline{\underline{E}} - \underline{\underline{F}}$ *est hermitienne, alors* $\underline{\underline{D}}^{\mathrm{H}} = \underline{\underline{D}}$ *et* $\underline{\underline{E}}^{\mathrm{H}} = \underline{\underline{F}}$. *Ainsi on a, par définition des matrices* $\underline{\underline{M}}$ *et* $\underline{\underline{N}}$ *pour la méthode SOR :*

$$\begin{aligned}\underline{\underline{M}}^{\mathrm{H}} + \underline{\underline{N}} &= \frac{1}{\omega}\left(\underline{\underline{D}} - \omega\underline{\underline{E}}\right)^{\mathrm{H}} + \frac{1-\omega}{\omega}\underline{\underline{D}} + \underline{\underline{F}} \\ &= \frac{1}{\omega}\left(\underline{\underline{D}} - \omega\underline{\underline{F}}\right) + \frac{1-\omega}{\omega}\underline{\underline{D}} + \underline{\underline{F}} \\ &= \frac{2-\omega}{\omega}\underline{\underline{D}}\end{aligned}$$

*De plus, comme* $\underline{\underline{A}}$ *est définie positive alors* $\underline{\underline{D}}$ *l'est aussi. Ainsi* $\underline{\underline{M}}^{\mathrm{H}} + \underline{\underline{N}}$ *est définie positive ssi* $|\omega - 1| < 1$ *(voir théorème 5.3).*

**Théorème 5.6.**
*Si* $\underline{\underline{A}} \in \mathcal{M}_n(\mathbb{R})$ *symétrique définie positive et* $2\,\underline{\underline{D}} - \underline{\underline{A}}$ *définie positive alors :*

1. *la méthode de Jacobi converge*
2. *la méthode SOR converge ssi* $|\omega - 1| < 1$

**Preuve.** $\underline{\underline{A}} = \underline{\underline{M}} - \underline{\underline{N}} = \underline{\underline{D}} - (\underline{\underline{E}} + \underline{\underline{F}})$ *et avec* $\underline{\underline{M}} = \underline{\underline{D}}$, $\underline{\underline{N}} = \underline{\underline{D}} - \underline{\underline{A}}$ *on a :* ${}^t\underline{\underline{M}} + \underline{\underline{N}} = 2\,\underline{\underline{D}} - \underline{\underline{A}}$. *On conclut avec le théorème 5.3 dans le cas réel. Pour SOR, voir le théorème 5.5 qui demande des hypothèses plus faibles que celui-ci (juste* $\underline{\underline{A}} \in \mathcal{M}_n(\mathbb{R})$ *symétrique définie positive).*

**Théorème 5.7.**
*Si* $\underline{\underline{A}} \in \mathcal{M}_n(\mathbb{K})$ *est tridiagonale par blocs et si* $\forall i, \det(\underline{\underline{A}}_{ii}) \neq 0$ *et si les valeurs propres de la matrice*

*d'itération $\mathcal{J}$ sont réelles, alors les méthodes de Jacobi et SOR convergent ou divergent simultanément : $\rho(\mathcal{L}_\omega) = \rho(\mathcal{J})^2$. En cas de convergence il existe $\omega^\star$ valeur optimale de $\omega$ telle que :*

$$\omega^\star = \frac{2}{1 + \sqrt{1 - \rho(\mathcal{J})^2}}$$

**Théorème 5.8.**
*Si $\underline{\underline{A}} \in \mathcal{M}_n(\mathbb{K})$ est tridiagonale par blocs et si $\forall i$, $\det(\underline{\underline{A}}_{ii}) \neq 0$ alors les méthodes de Jacobi et de Gauss-Seidel convergent ou divergent simultanément : $\rho(\mathcal{L}_1) = \rho(\mathcal{J})^2$*

**Théorème 5.9.**
*Si $\underline{\underline{A}}$ est une matrice à diagonale strictement dominante (voir définition 7.2) alors, pour le système $\underline{\underline{A}}.\underline{x} = \underline{b}$ les méthodes de Jacobi et de Gauss-Seidel convergent globalement.*

**Exercice 5.1.** Soit le système linéaire défini par les matrices :

$$\underline{\underline{A}} = \begin{pmatrix} 3 & 1 \\ -2 & -4 \end{pmatrix} \quad ; \quad \underline{b} = \begin{pmatrix} 1 \\ 6 \end{pmatrix}$$

Montrer que les méthodes de Jacobi et de Gauss-Seidel convergent puis résoudre selon ces deux méthodes en partant de $\underline{x}^{(0)} = (0, 0)$. Que se passe-t-il si on part de $\underline{x}^{(0)} = (1, -2)$

**Exercice 5.2.** Soient les matrices :

$$\underline{\underline{A}} = \begin{pmatrix} 1 & \frac{3}{4} & \frac{3}{4} \\ \frac{3}{4} & 1 & \frac{3}{4} \\ \frac{3}{4} & \frac{3}{4} & 1 \end{pmatrix} ; \underline{\underline{B}} = \begin{pmatrix} 1 & 2 & -2 \\ 1 & 1 & 1 \\ 2 & 2 & 1 \end{pmatrix} ; \underline{\underline{C}} = \begin{pmatrix} 2 & -1 & 1 \\ 2 & 2 & 2 \\ -1 & -1 & 2 \end{pmatrix}$$

Montrer que la méthode de Jacobi diverge pour les matrices $\underline{\underline{A}}$ et $\underline{\underline{C}}$ et converge pour $\underline{\underline{B}}$. Montrer également que la méthode de Gauss-Seidel diverge pour $\underline{\underline{B}}$ mais converge pour $\underline{\underline{C}}$.

**Exercice 5.3.** Soit le problème aux limites $(\mathcal{P})$ dans $\mathbb{R}$ défini par :

$$(\mathcal{P}) \begin{cases} -u''(x) = f(x) & \text{dans } \Omega = ]0, 1[ \\ u(x) = 0 & \text{sur } \partial\Omega = \{0\} \cup \{1\} \end{cases}$$

où $f \in L^2(\Omega)$.

1. Appliquer la méthode des différences finies à $(\mathcal{P})$ en prenant pour pas $h = \frac{1}{n+1}$ et écrire le problème linéaire $\underline{\underline{A}}.\underline{x} = \underline{b}$ correspondant ;
2. Décomposer $\underline{\underline{A}}$ en LU ;
3. Donner $\mathcal{J}(\underline{\underline{A}})$, la matrice de Jacobi de A. Montrer que ses vecteurs propres s'écrivent sous la forme $\{\sin(i\theta)\}_{1 \leq i \leq n}$ et en déduire les valeurs propres correspondantes. La méthode de Jacobi converge-t-elle ? Aurait-on pu procéder autrement pour déterminer sa convergence ?
4. Que peut-on dire de la convergence des méthodes de Gauss-Seidel et SOR ? Préciser, si elle existe, la valeur optimale de $\omega$.
5. On pose $n = 3$ et $f(x) = 4(3x - 1)$. Après avoir montré que $f \in L^2(\Omega)$, résoudre $(\mathcal{P})$ par la méthode LU puis par la méthode de Jacobi (5 itérations) en prenant $x^{(0)} = (0, 0, 0)$. Comparer les résultats.

## 5.2 Méthodes de gradient

### 5.2.1 Principe des méthodes de gradient

Considérons le problème :

"$(\mathcal{P})$ : Trouver $\underline{x} \in \mathbb{R}^n$ tel que $\underline{\underline{A}}.\underline{x} = \underline{b}$, où $\underline{b} \in \mathbb{R}^n$ et $\underline{\underline{A}} \in \mathcal{M}_n(\mathbb{R})$ matrice symétrique définie positive".

Autrement dit $\underline{\underline{A}}$ possède les caractéristiques suivantes : ${}^t\underline{\underline{A}} = \underline{\underline{A}}$ ; $\forall \underline{x} \in \mathbb{R}^n$, $\underline{x} \neq \underline{0}$, ${}^t\underline{x}.\underline{\underline{A}}.\underline{x} > 0$. $\underline{x}^\star$ solution de $(\mathcal{P})$ est l'unique solution du problème de minimisation $(\mathcal{P}_\mathcal{J})$ de la fonctionnelle quadratique $\mathcal{J}$ définie par, où $c \in \mathbb{R}$ :

"$(\mathcal{P}_\mathcal{J})$ : Trouver $\underline{x} \in \mathbb{R}^n$ tel que $\mathcal{J}(\underline{x}^\star) = \min_{\underline{x} \in \mathbb{R}^n} \{\mathcal{J}(\underline{x})\}$ où $\mathcal{J}(\underline{x}) = \frac{1}{2}\,{}^t\underline{x}.\underline{\underline{A}}.\underline{x} - {}^t\underline{x}.\underline{b} + c$"

En effet, grâce à la symétrie de $\underline{\underline{A}}$, on montre que $\forall \underline{x} \in \mathbb{R}^n$, $\mathcal{J}(\underline{x}) - \mathcal{J}(\underline{x}^\star) = \frac{1}{2}\,{}^t(\underline{x} - \underline{x}^\star).\underline{\underline{A}}.(\underline{x} - \underline{x}^\star)$. Ainsi, puisque $\underline{\underline{A}}$ est définie positive, $\forall \underline{x} \in \mathbb{R}^n$, $\mathcal{J}(\underline{x}^\star) < \mathcal{J}(\underline{x})$.

Le principe de résolution est alors décomposé ainsi :

1. Les méthodes de résolution itératives sont du type suivant :
$$\begin{cases} \underline{x}^{(0)} \in \mathbb{R}^n, \\ \underline{x}^{(k+1)} = \underline{x}^{(k)} + \alpha^{(k)}\,\underline{d}^{(k)} \end{cases}$$
Avec $\alpha^{(k)}$ déterminé par $\mathcal{J}(\underline{x}^{(k+1)}) = \min_{\alpha \in \mathbb{R}} \left\{\mathcal{J}(\underline{x}^{(k)} + \alpha\,\underline{d}^{(k)})\right\}$ et $\underline{d}^{(k)}$ représente la direction de la descente, dont le choix caractérise la méthode.
**Définition 5.6.**
2. ‖ On appelle résidu à l'étape $k$ le vecteur $\underline{r}^{(k)}$ tel que $\underline{r}^{(k)} = \underline{\underline{A}}.\underline{x}^{(k)} - \underline{b}$
Ainsi, $\underline{r}^{(k+1)} = \underline{r}^{(k)} + \alpha^{(k)}\underline{\underline{A}}.\underline{d}^{(k)}$
3. Calcul de $\alpha^{(k)}$ : par définition, $\partial_\alpha \mathcal{J}(\underline{x}^{(k)} + \alpha\,\underline{d}^{(k)}) = 0$. On montre que :
$$\alpha^{(k)} = -\frac{{}^t\underline{d}^{(k)}.\underline{r}^{(k)}}{{}^t\underline{d}^{(k)}.\underline{\underline{A}}.\underline{d}^{(k)}} \quad ; \quad {}^t\underline{d}^{(k)}.\underline{r}^{(k+1)} = 0 \tag{5.1}$$
Un tel $\alpha^{(k)}$ est bien défini grâce aux propriétés de $\underline{\underline{A}}$ et si $\underline{d}^{(k)} = 0$, c'est qu'on a atteint le résultat attendu puisque le résidu est nul.
4. On montre que $\mathcal{J}(\underline{x}^{(k+1)}) \leq \mathcal{J}(\underline{x}^{(k)})$, ce qui en terme d'énergie revient à "descendre à travers les niveaux d'énergie", d'où le nom de la méthode de la descente. En effet,
$$\mathcal{J}(\underline{x}^{(k+1)}) - \mathcal{J}(\underline{x}^{(k)}) = -\frac{1}{2}\frac{\left({}^t\underline{d}^{(k)}.\underline{r}^{(k)}\right)^2}{{}^t\underline{d}^{(k)}.\underline{\underline{A}}.\underline{d}^{(k)}} \leq 0 \tag{5.2}$$

**Exercice 5.4.** Démontrer l'expression de $\alpha^{(k)}$ (5.1) ainsi que l'inégalité (5.2).

### 5.2.2 Méthodes classiques de gradient

#### Méthode de la plus profonde descente

Cette méthode correspond au choix de la direction de descente : $\underline{d}^{(k)} = -\underline{r}^{(k)}$, autrement dit la méthode itérative correspondante s'écrit :
$$\begin{cases} \underline{x}^{(0)} \in \mathbb{R}^n, \\ \underline{x}^{(k+1)} = \underline{x}^{(k)} - \dfrac{{}^t\underline{r}^{(k)}.\underline{r}^{(k)}}{{}^t\underline{r}^{(k)}.\underline{\underline{A}}.\underline{r}^{(k)}}\,\underline{r}^{(k)} \end{cases}$$

### Méthode du gradient conjugué

Cette méthode est beaucoup plus performante que la méthode de la descente. Le principe est le suivant :

- on initialise avec un pas de plus profonde descente : $\underline{d}^{(0)} = -\underline{r}^{(0)}$
- on choisit la direction suivante de descente telle qu'elle soit conjuguée de la précédente par rapport au produit scalaire défini par $\underline{\underline{A}}$ : ${}^t\underline{d}^{(k)}.\underline{\underline{A}}.\underline{d}^{(k+1)} = 0$
- on impose à $\underline{d}^{(k+1)}$ la forme suivante : $\underline{d}^{(k+1)} = -\underline{r}^{(k+1)} + \beta^{(k)}\underline{d}^{(k)}$
- on calcule $\beta^{(k)} = \frac{{}^t\underline{d}^{(k)}.\underline{\underline{A}}.\underline{r}^{(k+1)}}{{}^t\underline{d}^{(k)}.\underline{\underline{A}}.\underline{d}^{(k)}}$. Si $\underline{d}^{(k)} \neq \underline{0}$, il n'y a pas de problème de définition de $\beta^{(k)}$. Si $\underline{d}^{(k)} = \underline{0}$, c'est que $\underline{r}^{(k)} = \underline{0}$ et donc $\underline{\underline{A}}.\underline{x}^{(k)} = \underline{b}$ et donc $\underline{x}^{(k)} = \underline{x}^\star$ solution du système linéaire.

**Proposition 5.1.** *On peut alors démontrer la propriété d'orthogonalité (i) et les expressions condensées des paramètres $\beta$ (ii) et $\alpha$ (iii) suivantes :*

$$(1)\ {}^t\underline{r}^{(k)}.\underline{r}^{(k+1)} = 0 \quad ; \quad (2)\ \beta^{(k)} = \frac{{}^t\underline{r}^{(k+1)}.\underline{r}^{(k+1)}}{{}^t\underline{r}^{(k)}.\underline{r}^{(k)}} \quad ; \quad (3)\ \alpha^{(k)} = \frac{{}^t\underline{r}^{(k)}.\underline{r}^{(k)}}{{}^t\underline{d}^{(k)}.\underline{\underline{A}}.\underline{d}^{(k)}}$$

**Exercice 5.5.** Démontrer les égalités de la proposition 5.1.

**Proposition 5.2** (Algorithme GC)**.** *Le principe de calcul est donc le suivant :*

1. *Les données d'entrée sont $\underline{x}^{(0)}$ et l'erreur souhaitée $\varepsilon$.*
2. *Initialisation : $\underline{r}^{(0)} = \underline{\underline{A}}.\underline{x}^{(0)} - \underline{b}$ ; $\underline{d}^{(0)} = -\underline{r}^{(0)}$*
3. *Pour $k \geq 0$, tant que $\|\underline{r}^{(k)}\| > \varepsilon\|\underline{b}\|$ (résidu normalisé par $\|.\|$ norme sur $\mathbb{R}^n$) :*

$$\begin{cases} \alpha^{(k)} = \dfrac{{}^t\underline{r}^{(k)}.\underline{r}^{(k)}}{{}^t\underline{d}^{(k)}.\underline{\underline{A}}.\underline{d}^{(k)}} \\ \underline{x}^{(k+1)} = \underline{x}^{(k)} + \alpha^{(k)}\underline{d}^{(k)} \\ \underline{r}^{(k+1)} = \underline{r}^{(k)} + \alpha^{(k)}\underline{\underline{A}}.\underline{d}^{(k)} \\ \beta^{(k)} = \dfrac{{}^t\underline{r}^{(k+1)}.\underline{r}^{(k+1)}}{{}^t\underline{r}^{(k)}.\underline{r}^{(k)}} \\ \underline{d}^{(k+1)} = -\underline{r}^{(k+1)} + \beta^{(k)}\underline{d}^{(k)} \end{cases}$$

**Théorème 5.10.**
*$\forall \underline{x}^{(0)} \in \mathbb{R}^n$, il existe un plus petit entier $\ell$ tel que :*

$$\begin{cases} \underline{d}^{(\ell)} = \underline{0} \\ \underline{x}^{(\ell)} = \underline{x}^\star \\ \forall 0 \leq j < i \leq \ell,\ {}^t\underline{r}^{(i)}.\underline{d}^{(j)} = 0\,;\ {}^t\underline{d}^{(i)}.\underline{\underline{A}}.\underline{d}^{(j)} = 0\,;\ {}^t\underline{r}^{(i)}.\underline{r}^{(j)} = 0 \\ \mathcal{J}(\underline{x}^{(k+1)}) = \min\limits_{(\chi^{(j)})_{1\leq j\leq n}\in\mathbb{R}^n} \left\{ \mathcal{J}\left(\underline{x}^{(0)} + \sum\limits_{j=1}^{n} \chi^{(j)}\underline{d}^{(j)}\right)\right\} \end{cases}$$

Ainsi la méthode du gradient conjugué cest une **méthode directe** qui converge en au plus $n$ itérations. Mais en pratique, lorsque $n$ est grand on s'arrête en général bien avant l'entier $\ell$.

**Théorème 5.11.**
*Si $\underline{\underline{A}} \in \mathcal{M}_n(\mathbb{R})$ symétrique définie positive admet $\ell \leq n$ valeurs propres distinctes, alors l'algorithme du gradient conjugué converge vers $\underline{x}^\star$ en au plus $\ell$ itérations.*

**Théorème 5.12.**
*Soient $\underline{x}^\star$ la solution du système linéaire et $\underline{\varepsilon}^{(k)}$ l'erreur à l'étape $k$ : $\underline{\varepsilon}^{(k)} = \underline{x}^{(k)} - \underline{x}^\star$. Nous avons le résultat suivant :*

$$\|\underline{\varepsilon}^{(k)}\|_A < 2\frac{\left(\sqrt{\kappa(\underline{\underline{A}})} - 1\right)^k}{\left(\sqrt{\kappa(\underline{\underline{A}})} + 1\right)^k} \|\underline{\varepsilon}^{(0)}\|_A$$

*où $\|\underline{\varepsilon}^{(k)}\|_A^2 = {}^t\,\underline{\varepsilon}^{(k)}.\underline{\underline{A}}.\underline{\varepsilon}^{(k)}$ et $\kappa(\underline{\underline{A}}) = \text{cond}_2(\underline{\underline{A}})$ est le conditionnement de la matrice $\underline{\underline{A}} \in \mathcal{M}_n(\mathbb{R})$ symétrique définie positive. Ce résultat permet d'estimer le nombre d'itérations nécessaires à l'obtention d'une erreur donnée.*

**Remarque 5.4.**
Si la matrice est mal conditionnée, alors la méthode du gradient conjugué convergera en plus que $n$ itérations. Nous ne sommes pas en mesure de justifier cela rigoureusement car il nous faudrait faire la démonstration du théorème 5.10. Mais si on raisonne "avec les mains", cette démonstration met en évidence qu'à chaque itération les éléments de descente participent à la construction d'une base, complète au bout de $n$ itérations. Le mauvais conditionnement de la matrice entraîne que les éléments ne sont pas exactement orthogonaux entre eux et donc la convergence du schéma est très fortement ralentie. D'où la nécessité d'améliorer le conditionnement du système, notamment par des techniques de préconditionnement.

## 5.3 Préconditionnement de matrice

### 5.3.1 Principe

Si la matrice $\underline{\underline{A}}$ du système linéaire est mal conditionnée, une technique de préconditionnement permet d'améliorer la résolution du système. Le principe est d'introduire une matrice $\underline{\underline{P}} \in \mathcal{M}_n(\mathbb{R})$ inversible telle que

$$\underline{\underline{P}}^{-1}.\underline{\underline{A}}.\underline{x} = \underline{\underline{P}}^{-1}.\underline{b} \Leftrightarrow \underline{\underline{A}}.\underline{x} = \underline{b}$$

La matrice $\underline{\underline{P}}$ est choisie telle que le conditionnement de $\underline{\underline{P}}^{-1}.\underline{\underline{A}}$ soit plus petit que celui de $\underline{\underline{A}}$. Idéalement on voudrait que $\underline{\underline{P}}^{-1} = \underline{\underline{A}}^{-1}$ mais ce choix de préconditionneur n'est pas réalisable en pratique. Par conséquent, le choix du préconditionneur $\underline{\underline{P}}$ se fait selon les critères suivants

1. $\underline{\underline{P}}$ doit "ressembler" le plus possible à $\underline{\underline{A}}$, i.e :
   — $Sp(\underline{\underline{P}}^{-1}.\underline{\underline{A}}) \simeq Sp(\underline{\underline{I}})$
   — $\|\underline{\underline{P}}^{-1}.\underline{\underline{A}} - \underline{\underline{I}}\| < \varepsilon$ où $\varepsilon > 0$ le plus petit possible
2. La structure de $\underline{\underline{P}}$ doit respecter la structure de $\underline{\underline{A}}$, par exemple la symétrie ;
3. Le stockage des composantes de $\underline{\underline{P}}$ ne soit pas alourdir le stockage global du système.

Un famille de préconditionneurs très importante est celle générée par les méthodes itératives d'éclatement.

### 5.3.2 Préconditionnement par les méthodes d'éclatement

Rappelons le principe : on décompose $\underline{\underline{A}} = \underline{\underline{D}} - \underline{\underline{E}} - \underline{\underline{F}} = \underline{\underline{M}} - \underline{\underline{N}}$ avec $\underline{\underline{M}}$ inversible tels que :

$$\begin{aligned}
\underline{\underline{A}}.\underline{x} = \underline{b} &\Leftrightarrow \underline{\underline{M}}.\underline{x} = \underline{\underline{N}}.\underline{x} + \underline{b} \\
&\Leftrightarrow \underline{x} = \underline{\underline{M}}^{-1}.\underline{\underline{N}}.\underline{x} + \underline{\underline{M}}^{-1}.\underline{b} = \underline{\underline{C}}.\underline{x} + \underline{d} \\
&\Leftrightarrow \left(\underline{\underline{I}} - \underline{\underline{M}}^{-1}.\underline{\underline{N}}\right).\underline{x} = \underline{d} \qquad \text{or } \underline{\underline{N}} = \underline{\underline{M}} - \underline{\underline{A}} \\
&\Leftrightarrow \underline{\underline{M}}^{-1}.\underline{\underline{A}}.\underline{x} = \underline{\underline{M}}^{-1}.\underline{b}
\end{aligned}$$

Par conséquent on voit que la matrice inversible $\underline{\underline{M}}$ est un préconditionneur du système linéaire $\underline{\underline{A}}.\underline{x} = \underline{b}$. On va pouvoir ainsi décliner les différentes méthodes d'éclatement vues précédemment. Ces préconditionneurs ont l'avantage d'utiliser les composantes de $\underline{\underline{A}}$, déjà stockées en mémoire.

**Jacobi** Soit le préconditionneur $\underline{\underline{P}} = \underline{\underline{D}}$. Les avantages de ce préconditionnement est qu'il est simple à réaliser, qu'il ne nécessite pas beaucoup de mémoire de stockage et que si $\underline{\underline{A}}$ est symétrique alors $\underline{\underline{P}}$ conserve la symétrie.

**SOR** $\underline{\underline{P}} = \frac{1}{\omega}\left(\underline{\underline{D}} - \omega\underline{\underline{E}}\right)$. Lorsque $\omega = 1$, on obtient le préconditionnement de Gauss-Seidel. En général, si la matrice $\underline{\underline{A}}$ est symétrique alors $\underline{\underline{P}}$ ne l'est pas. Dans ce cas, on utilisera plutôt le préconditionnement SSOR.

**SSOR (Symetric Successive Over Relaxation)** $\underline{\underline{P}} = \frac{1}{\omega(2-\omega)}\left(\underline{\underline{D}} - \omega\underline{\underline{E}}\right).\underline{\underline{D}}^{-1}.\left(\underline{\underline{D}} - \omega\underline{\underline{F}}\right)$ où $\underline{\underline{E}} = {}^t\underline{\underline{F}}$ puisque $\underline{\underline{A}} = {}^t\underline{\underline{A}}$. Ce type de préconditionnement, assuré si $\omega \in ]0,2[$, est particulièrement adapté pour la méthode du gradient conjugué préconditionné. Si $\omega = 1$, c'est la méthode SGS (Symetric Gauss Seidel).

### 5.3.3 Méthode du gradient conjugué préconditionné

**Proposition 5.3** (Algorithme PGC). *L'algorithme de calcul étant très proche de celui du gradient conjugué, les différences sont notées en rouge :*

1. *Les données d'entrée sont $\underline{x}^{(0)}$, le préconditionneur $\underline{\underline{P}}$ et l'erreur souhaitée $\varepsilon$.*
2. *Initialisation :* $\underline{r}^{(0)} = \underline{\underline{A}}.\underline{x}^{(0)} - \underline{b} \quad ; \quad \underline{z}^{(0)} = \underline{\underline{P}}^{-1}.\underline{r}^{(0)} \quad ; \quad \underline{d}^{(0)} = -\underline{z}^{(0)}$
3. *Pour $k \geq 0$, tant que $\|\underline{r}^{(k)}\| > \varepsilon\|\underline{b}\|$ (résidu normalisé par $\|.\|$ norme sur $\mathbb{R}^n$) :*

$$\begin{cases}
\alpha^{(k)} = \dfrac{{}^t\underline{z}^{(k)}.\underline{r}^{(k)}}{{}^t\underline{d}^{(k)}.\underline{\underline{A}}.\underline{d}^{(k)}} \\
\underline{x}^{(k+1)} = \underline{x}^{(k)} + \alpha^{(k)}\underline{d}^{(k)} \\
\underline{r}^{(k+1)} = \underline{r}^{(k)} + \alpha^{(k)}\underline{\underline{A}}.\underline{d}^{(k)} \\
\underline{z}^{(k+1)} = \underline{\underline{P}}^{-1}.\underline{r}^{(k+1)} \\
\beta^{(k)} = \dfrac{{}^t\underline{z}^{(k+1)}.\underline{r}^{(k+1)}}{{}^t\underline{z}^{(k)}.\underline{r}^{(k)}} \\
\underline{d}^{(k+1)} = -\underline{z}^{(k+1)} + \beta^{(k)}\underline{d}^{(k)}
\end{cases}$$

**Remarque 5.5.**
L'algorithme proposé n'est pas l'algorithme du gradient conjugué appliqué à $\underline{\underline{P}}^{-1}.\underline{\underline{A}}$.

**Remarque 5.6.**
Le produit $\underline{\underline{P}}^{-1}.\underline{\underline{A}}$ n'est pas nécessairement symétrique. Ce n'est pas un problème car l'algorithme modifié avec le préconditionnement ne prend en compte que la symétrie de $\underline{\underline{A}}$. En effet, à partir des expressions suivantes issues de la méthode générale de gradient :

$$\underline{r}^{(k+1)} = -\underline{r}^{(k)} + \alpha^{(k)}\underline{\underline{A}}.\underline{d}^{(k)} \quad ; \quad {}^t\underline{d}^{(k)}.\underline{r}^{(k+1)} = 0 \quad ; \quad \alpha^{(k)} = -\frac{{}^t\underline{d}^{(k)}.\underline{r}^{(k)}}{{}^t\underline{d}^{(k)}.\underline{\underline{A}}.\underline{d}^{(k)}}$$

, de la relation de conjugaison

$${}^t\underline{d}^{(k)}.\underline{\underline{A}}.\underline{d}^{(k+1)} = 0$$

et de l'expression de la descente prenant en compte le préconditionneur

$$\underline{d}^{(k+1)} = -\underline{z}^{(k+1)} + \beta^{(k)}\underline{d}^{(k)}$$

on montre que (voir démonstration pour le gradient conjugué sans préconditionnement) :

$${}^t\underline{z}^{(k)}.\underline{r}^{(k+1)} = 0 \quad ; \quad \beta^{(k)} = \frac{{}^t\underline{z}^{(k+1)}.\underline{r}^{(k+1)}}{{}^t\underline{z}^{(k)}.\underline{r}^{(k)}} \quad ; \quad \alpha^{(k)} = \frac{{}^t\underline{z}^{(k)}.\underline{r}^{(k)}}{{}^t\underline{d}^{(k)}.\underline{\underline{A}}.\underline{d}^{(k)}}$$

**Exercice 5.6.** Reprenons la matrice de l'exemple initial : $\underline{\underline{A}} = \begin{bmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{bmatrix}$, $\underline{b} = \begin{pmatrix} 32 \\ 23 \\ 33 \\ 31 \end{pmatrix}$. Résoudre ce système par la méthode du gradient conjugué

1. sans préconditionnement,
2. avec le préconditionnement de Jacobi
3. avec le préconditionnement SSOR avec $\omega = 1$.

## 5.4 Exercices sur les méthodes d'éclatement

### 5.4.1 Exercices avancés

**Exercice 5.7.** On considère le système linéaire

$$\underline{\underline{A}}.\underline{x} = \underline{b}$$

où $\underline{b}$ est un vecteur de $\mathbb{R}^3$ et $\underline{\underline{A}} \in \mathcal{M}_3(\mathbb{R})$ une matrice fonction du paramètre $a \in \mathbb{R}$ telle que :

$$A = \begin{bmatrix} 1 & a & a \\ a & 1 & a \\ a & a & 1 \end{bmatrix}$$

À quelle condition sur $a$ peut-on utiliser :

1. la méthode de Cholesky ?
2. la méthode de Jacobi ?

**Exercice 5.8.** Soit le problème aux limites $(\mathcal{P})$ dans $\mathbb{R}$ défini par :

$$(\mathcal{P}) \begin{cases} -u''(x) = f(x) & \text{dans } \Omega = ]0,1[ \\ u(x) = 0 & \text{sur } \partial\Omega = \{0\} \cup \{1\} \end{cases}$$

où $f \in L^2(\Omega)$.

1. Appliquer la méthode des différences finies à $(\mathcal{P})$ en prenant pour pas $h = \frac{1}{n+1}$ et écrire le problème linéaire $\underline{\underline{A}}.\underline{x} = \underline{b}$ correspondant ;
2. Décomposer $\underline{\underline{A}}$ en LU ;
3. Donner $\mathcal{J}(\underline{\underline{A}})$, la matrice de Jacobi de A. Montrer que ses vecteurs propres s'écrivent sous la forme $\{\sin(i\theta)\}_{1\leq i\leq n}$ et en déduire les valeurs propres correspondantes. La méthode de Jacobi converge-t-elle ? Aurait-on pu procéder autrement pour déterminer sa convergence ?
4. Que peut-on dire de la convergence des méthodes de Gauss-Seidel et SOR ? Préciser, si elle existe, la valeur optimale de $\omega$.
5. On pose $n = 3$ et $f(x) = 4(3x - 1)$. Après avoir montré que $f \in L^2(\Omega)$, résoudre $(\mathcal{P})$ par la méthode LU puis par la méthode de Jacobi (5 itérations) en prenant $x^{(0)} = (0, 0, 0)$. Comparer les résultats.

### 5.4.2 TD

**Exercice 5.9.** Soient $\underline{\underline{A}} \in \mathcal{M}_n(\mathbb{R})$ une matrice symétrique définie positive et $\underline{\underline{B}} \in \mathcal{M}_{n,m}(\mathbb{R})$ $(m \neq n)$ vérifiant $\ker \underline{\underline{B}} = \{0\}$.
Étant donné un vecteur $\underline{b} \in \mathbb{R}^n$, on cherche $\underline{x} \in \mathbb{R}^n$ et $\underline{y} \in \mathbb{R}^m$ tels que :

$$(\mathcal{S}) \begin{cases} \underline{\underline{A}}.\underline{x} + \underline{\underline{B}}.\underline{y} & = \underline{b} \\ {}^t\underline{\underline{B}}.\underline{x} & = \underline{0} \end{cases}$$

1. Montrer que $\underline{\underline{A}}^{-1}$ et ${}^t\underline{\underline{B}}.\underline{\underline{A}}^{-1}.\underline{\underline{B}}$ sont symétriques définies positives. En déduire que $(\mathcal{S})$ à une solution unique.

   On veut calculer l'unique solution de $(\mathcal{S})$ par la méthode itérative suivante :

$$\begin{cases} \underline{\underline{A}}.\underline{x}^{(k+1)} & = \underline{b} - \underline{\underline{B}}.\underline{y}^{(k)} \\ \underline{y}^{(k+1)} & = \underline{y}^{(k)} + r\,{}^t\underline{\underline{B}}.\underline{x}^{(k+1)} \end{cases}$$

   avec $r \in \mathbb{R}^{+*}$ et $\underline{y}^{(0)} \in \mathbb{R}^m$ donnés.

2. Exprimer $\underline{y}^{(k+1)}$ en fonction de $\underline{y}^{(k)}$.
3. Montrer que cette méthode converge si et seulement si :

$$0 < r < \frac{2}{R}$$

où $R = \rho\left({}^t\underline{\underline{B}}.\underline{\underline{A}}^{-1}.\underline{\underline{B}}\right)$.

4. Montrer que toute valeur propre de ${}^t\underline{\underline{B}}.\underline{\underline{A}}^{-1}.\underline{\underline{B}}$ est aussi valeur propre de $\underline{\underline{B}}\,{}^t\underline{\underline{B}}.\underline{\underline{A}}^{-1}$ ; en déduire que :

$$R < \frac{\beta}{\alpha}$$

où $\alpha$ est la plus petite valeur propre de $\underline{\underline{A}}$ et $\beta$ la plus grande valeur propre de ${}^t\underline{\underline{B}}.\underline{\underline{B}}$.

On pourra pour ce faire s'appuyer sur la propriété :

$$\forall \underline{\underline{M}} \in \mathcal{M}_n(\mathbb{R}) \qquad \|\underline{\underline{M}}\|_2 = \sqrt{\rho\left({}^t\underline{\underline{M}}.\underline{\underline{M}}\right)} \geq \rho\left(\underline{\underline{M}}\right)$$

### 5.4.3 Annale

**Exercice 5.10 (**Annale 2009**).** Soit $n$ un entier naturel strictement positif et $a$ un réel. On considère la matrice $\underline{\underline{A}} \in \mathcal{M}_n(\mathbb{R})$ est tridiagonale telle que :

$$A_{ij} = \begin{cases} a & \text{si } i = j \\ -1 & \text{si } |j-i| = 1 \\ 0 & \text{sinon} \end{cases}$$

On veut résoudre le système $\underline{\underline{A}}.\underline{x} = \underline{b}$ par la méthode de Jacobi. On rappelle que cette méthode est une méthode itérative qui consiste à décomposer $\underline{\underline{A}}$ en $\underline{\underline{A}} = \underline{\underline{D}} - \underline{\underline{N}}$ où $\underline{\underline{D}}$ est la partie diagonale de $\underline{\underline{A}}$. À quelle condition sur $a$ peut-on appliquer cette méthode ? Cette condition étant remplie, en déduire la matrice d'itération $\mathcal{J}(\underline{\underline{A}})$ correspondant à ce problème.

## 5.5 Exercices sur les méthodes de gradient

### 5.5.1 Applications numériques

**Exercice 5.11.** On cherche à minimiser la fonctionnelle quadratique :

$$\mathcal{J}(x_1, x_2) = \frac{3}{2}x_1^2 - x_1x_2 + x_2^2 - x_1 - 3x_2 + \sqrt{\pi}$$

Donner le nombre d'itération nécessaires à la convergence de l'algorithme du gradient conjugué et résoudre à partir de $\underline{x}^{(0)} = (0, 0)$ et pour critère d'arrêt $10^{-4}$.

**Exercice 5.12 (**Minimisation d'une fonction**).** Trouver par la méthode de la plus profonde descente et du gradient conjugué le minimum de la fonction :

$$J(x, y) = x^2 + \frac{1}{2}y^2 + xy - y - 2$$

On prendra pour valeur initiale $\underline{x}^{(0)} = (0, 0)$ et pour critère d'arrêt $10^{-4}$.

**Exercice 5.13.** Résoudre par la méthode du gradient conjugué et de la plus profonde descente le système linéaire défini par :

$$\underline{\underline{A}} = \begin{bmatrix} 4 & -2 & 0 & \dots & 0 \\ -2 & 4 & -2 & & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & & -2 & 4 & -2 \\ 0 & \dots & 0 & -2 & 4 \end{bmatrix} \quad ; \quad \underline{b} = \begin{pmatrix} 1 \\ \vdots \\ \vdots \\ \vdots \\ 1 \end{pmatrix}$$

On testera plusieurs valeurs de $n$.

**Exercice 5.14.** Reprenons la matrice de l'exemple initial : $\underline{\underline{A}} = \begin{bmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{bmatrix}$, $\underline{b} = \begin{pmatrix} 32 \\ 23 \\ 33 \\ 31 \end{pmatrix}$. Résoudre ce système par la méthode du gradient conjugué

1. sans préconditionnement,
2. avec le préconditionnement de Jacobi
3. avec le préconditionnement SSOR avec $\omega = 1$.

### 5.5.2 TD

**Exercice 5.15.** On veut résoudre par la méthode de plus profonde descente le système linéaire $\underline{\underline{A}}.\underline{x} = \underline{b}$, où $\underline{\underline{A}}$ est une matrice symétrique définie positive. Dans un premier temps, on va montrer que cette méthode est convergente.

1. En vous appuyant sur l'inégalité de Kantorovitch :

$$1 \leq \frac{(\underline{\underline{A}}.\underline{x}, \underline{x})(\underline{\underline{A}}^{-1}.\underline{x}, \underline{x})}{\|\underline{x}\|^4} \leq \frac{\left(K(\underline{\underline{A}})^{\frac{1}{2}} + K(\underline{\underline{A}})^{\frac{-1}{2}}\right)^2}{4}$$

   montrer que :

$$E(\underline{x}^{(k)}) \leq \left(\frac{K(\underline{\underline{A}}) - 1}{K(\underline{\underline{A}}) + 1}\right)^{2k} E(\underline{x}^{(0)})$$

2. Montrer maintenant que pour tout vecteur $\underline{x} \in \mathbb{R}^n$ :

$$(\underline{\underline{A}}.\underline{x}, \underline{x}) \geq \lambda_{min} \|\underline{x}\|^2$$

   où $\lambda_{min}$ désigne la plus petite valeur propre de $\underline{\underline{A}}$.
   On pourra pour ce faire exprimer les différents vecteurs dans la base orthonormale de $\mathbb{R}^n$ formée par les vecteurs propres de $\underline{\underline{A}}$.

3. En vous appuyant sur les questions (1) et (2), montrer finalement que l'erreur $\underline{\varepsilon}^{(k)} = \underline{x}^{(k)} - \underline{x}^{\star}$ tend vers $\underline{0}$ quand $k$ tend vers l'infini.
   Que peut-on en déduire pour cette méthode de gradient ?

   On va maintenant utiliser cette méthode pour résoudre le système suivant :

$$\begin{cases} 3x_1 - 2x_2 & = 1 \\ -2x_1 + 4x_2 & = 2 \end{cases}$$

4. Vérifier que la méthode de plus profonde descente est bien adaptée à la résolution de ce système.
5. Résoudre en prenant $\underline{x}^{(0)} = \left(0, \frac{1}{2}\right)$. On s'arrêtera à la 2$^{\text{e}}$ itération.

### 5.5.3 Annale

**Exercice 5.16 (**Rattrapage 2020**).** On cherche à résoudre le système linéaire suivant :

$$(S): \qquad \underline{\underline{A}}.\underline{X} = \underline{b} \quad \text{où} \quad \underline{\underline{A}} = \begin{bmatrix} 2 & -1 & -1 \\ -1 & 2 & 1 \\ -1 & 1 & 2 \end{bmatrix} \quad \text{et} \quad \underline{b} = \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix}$$

**0 - Préliminaire**

1. Montrer que le spectre (i.e l'ensemble des valeurs propres) de $\underline{\underline{A}}$ vaut $Sp(\underline{\underline{A}}) = \{1, 4\}$ .
2. **Bonus :** En déduire le conditionnement spectral de $\underline{\underline{A}}$. Quelle conclusion en tirez-vous ?

**1 - Résolution par méthodes directes**

On supposera dans un premier temps que ce système admet une unique solution.

**Pivot de Gauss**

1. Résoudre $(S)$ par pivot de Gauss en spécifiant toutes les étapes.
2. En déduire le déterminant de $\underline{\underline{A}}$, puis prouver que le système $(S)$ admet un unique solution.

**Décomposition LU**

1. La matrice $\underline{\underline{A}}$ est-elle décomposable en LU ? Rappeler la structure des matrices $\underline{\underline{L}}$ et $\underline{\underline{U}}$.
2. Déduire des étapes de résolution de $(S)$ par pivot de Gauss la décomposition LU de $\underline{\underline{A}}$.
3. Résoudre $(S)$ par décomposition LU.

**Résolution directe**

1. Déduire de la décomposition LU de $\underline{\underline{A}}$ la matrice inverse de $\underline{\underline{A}}$.
2. Résoudre $(S)$ par résolution directe.

**2 - Résolution par méthodes itératives**

**Méthodes d'éclatement**

1. Rappeler, dans le cas général, le principe de résolution d'un système linéaire par une méthode d'éclatement. Sous quelle(s) conditions(s) converge-t-elle ?
2. Calculer les matrices d'itération de Jacobi et de Gauss-Seidel. Pour le système $(S)$, ces méthodes convergent-elles ?

**Méthode du gradient conjugué**

1. Montrer que la méthode du gradient conjugué converge pour le système $(S)$. En combien d'itérations au maximum cette méthode converge-t-elle en théorie ? **Bonus :** A votre avis, en pratique cette estimation est-elle vérifiée ? [On pourra s'appuyer, pour toute cette question, sur le paragraphe 5.19]
2. Résoudre $(S)$ par la méthode du gradient conjugué, dont les étapes sont rappelées ci-après, en partant de $\underline{x}^{(0)} = \underline{0}$.

# Chapitre 6

# Méthodes de résolution itérative de systèmes

*Ce dernier chapitre est dédié aux méthodes itératives de résolution de fonctions (a priori) non linéaires, qu'elles soient algébriques ou transcendantes. Tout d'abord, l'émergence de ces problèmes de recherche de racines est illustré sur plusieurs exemples historiques. La notion de convergence globale ou locale est ensuite introduite, découlant sur les notions de bassin d'attraction et d'attractivité de points fixes. Différentes méthodes à un pas seront étudiées, et leur ordre de convergence déterminé.*

**Sommaire**

## 6.1 Introduction

Les exemples historiques des approximations babyloniennes et de Héron d'Alexandrie sont tirés intégralement du polycopié de calcul scientifique (2013) de Jean-Marc Malasoma. Ces textes sont présentés à titre de culture générale.

### 6.1.1 Approximations Babyloniennes

Les recherches archéologiques ont mis au jour un très grand nombre de tablettes d'argile babyloniennes dont certaines contiennent des informations mathématiques. L'étude de ces documents, en écriture cunéiforme, nous apprend que les Babyloniens utilisaient un système de numération sexagésimale de position.
L'utilisation de la base 60 a été héritée des Sumériens qui l'utilisaient déjà au IIIe millénaire av. J.-C. Cette base nécessite l'utilisation de 60 chiffres (si l'on tient compte du zéro qui ne sera inventé par les Babyloniens que vers la fin du dernier millénaire av. J.-C.). Pour écrire les chiffres de 1 à 59, les Babyloniens utilisaient un système décimal et uniquement deux symboles, le clou pour l'unité et le chevron pour la dizaine.
Ainsi, le nombre 59 était représenté par cinq chevrons suivis de neuf clous. Mais le nombre 60 était de nouveau représenté par un clou ! Trois chevrons pourront ainsi correspondre au nombre 30 ou bien au nombre $\frac{30}{60} = \frac{1}{2}$ ou encore plus généralement à $30 \times 60^n$ pour tout entier relatif $n$, etc.
Les calculs étaient effectués en virgule flottante ce qui constitue un principe très pratique pour effectuer les multiplications. Nos ordinateurs utilisent ce même principe, mais en base 2, avec exposant et mantisse. Toutefois, les Babyloniens n'écrivaient pas les exposants et les scribes devaient donc garder en tête le bon ordre de grandeur c'est-à-dire la bonne puissance 60. Certaines tablettes ont révélé que, près de 1000 ans avant Pythagore, les Babyloniens avaient connaissance de la formule qui lie le carré de l'hypoténuse d'un triangle rectangle à la somme des carrés des deux autres côtes de ce triangle. De plus, ils utilisaient, pour mener certains calculs, des approximations de diverses racines carrées, mettant en jeu des précisions souvent remarquables.
La tablette VAT (Vorderasiatische Abteilung Tontafeln) 6598, conservée au Staatliche Museen de Berlin, datée du début du IIe millénaire av. J.-C. (-2000 à -1700), comporte une liste de problèmes mathématiques. Dans le sixième problème de cette tablette, le scribe donne la longueur de la diagonale d'une porte de largeur 10 et de hauteur 40, c'est-à-dire le nombre $\sqrt{40^2 + 10^2} = 10\sqrt{17}$. Il note sa réponse sous la forme de quatre chevrons suivis d'un clou et d'un chevron lui-même suivi de cinq clous soit en notation sexagésimale 41 15, c'est-à-dire en utilisant la bonne puissance de soixante !

$$41 + \frac{15}{60} = \frac{165}{4}$$

Aucune indication n'est donnée sur la méthode de calcul utilisée pour parvenir à ce résultat. Pour nous, il est évident qu'il ne s'agit que d'une approximation et pas de n'importe laquelle. En effet, un développement limité au premier ordre permet d'expliquer le résultat donné par le scribe :

$$\sqrt{40^2 + 10^2} = 40\sqrt{1 + \left(\frac{1}{4}\right)^2} \simeq 40\left[1 + \frac{1}{2}\left(\frac{1}{4}\right)^2\right] = \frac{165}{4}$$

Cette remarque ne prouve évidemment pas que le scribe connaissait le début du développement limité que nous avons utilisé, ni même qu'il était conscient de donner une approximation de la longueur de la diagonale cherchée. Cependant la coïncidence, si coïncidence il y a, interpelle.

### 6.1.2 Méthode de Héron d'Alexandrie

L'exemple connu, le plus ancien, de description d'une méthode itérative pour la résolution d'une équation est celui rapporté par Héron d'Alexandrie.
Cela ne signifie évidemment pas que ce mathématicien et physicien grec, dont les historiens supposent qu'il a vécu dans la seconde moitié du Ier siècle de notre ère, soit l'inventeur du procédé qu'il décrit. D'ailleurs, certains spécialistes pensent même que le procédé est d'origine babylonienne !

Comme 720 n'est pas le carré d'un nombre rationnel et que le carré parfait immédiatement supérieur est $(27)^2 = 729$, il divise 720 par 27 il ajoute ensuite 27 et il en prend la moitié ce qui donne un rationnel qui élevé au carré dépasse 720 de seulement $\frac{1}{36}$. Il poursuit son raisonnement en ajoutant qu'en remplaçant 729 par le carré du nombre trouvé et en procédant de la même façon on trouverait une différence entre les carrés beaucoup plus petite que $\frac{1}{36}$.
Analysons, dans un cadre plus général, le procédé de construction de cette méthode itérative. Si $a$ est un réel strictement positif et $x$ est une approximation par excès de $\sqrt{a}$ alors

$$0 < \sqrt{a} < x$$

et par conséquent, en multipliant les membres de cette inégalité par le nombre positif $\frac{\sqrt{a}}{x}$, nous obtenons

$$0 < \frac{a}{x} < \sqrt{a} < x \tag{6.1}$$

Ce qui montre que le rapport $\frac{a}{x}$ est une approximation par défaut de $\sqrt{a}$. De plus, d'une part pour tout $a > 0$ et pour tout réel $x$ on a

$$x^2 - 2\sqrt{a}x + a = \left(x - \sqrt{a}\right)^2 > 0$$

et par conséquent pour tout $x > 0$

$$\sqrt{a} < \frac{1}{2}\left(x + \frac{a}{x}\right) \tag{6.2}$$

et d'autre part, les inégalités (6.1) impliquent

$$x + \frac{a}{x} < 2x$$

et donc

$$\frac{1}{2}\left(x + \frac{a}{x}\right) < x \tag{6.3}$$

En regroupant (6.1), (6.2) et (6.3), nous obtenons finalement

$$0 < \frac{a}{x} < \sqrt{a} < \frac{1}{2}\left(x + \frac{a}{x}\right) < x \tag{6.4}$$

Nous venons ainsi de montrer que si $x$ est une approximation par excès de $\sqrt{a}$ alors $\frac{a}{x}$ en est une approximation par défaut et que la moyenne arithmétique de ces deux approximations est alors une meilleure approximation par excès. Nous obtenons ainsi à chaque itération une meilleure approximation par excès de la racine carrée cherchée.
Puisque (6.2) est vraie pour tout $x > 0$, elle le reste si $x$ est une approximation positive par défaut de $\sqrt{a}$. De sorte que si l'on commence le procédé de Héron d'Alexandrie en utilisant une approximation par défaut de $\sqrt{a}$, l'itération suivante se fera à partir d'une approximation par excès.
Remarquons enfin que l'idée, sous-jacente à cette méthode, consiste à remplacer la résolution de l'équation polynômiale $x^2 - a = 0$ par celle de l'équation non polynômiale mais trivialement équivalente

$$x = \frac{1}{2}\left(x + \frac{a}{x}\right)$$

c'est-à-dire par la recherche des points fixes de la fonction $H_a$ définie par la formule

$$H_a(x) = \frac{1}{2}\left(x + \frac{a}{x}\right)$$

### 6.1.3 Méthode d'Al-Kashi

Depuis l'antiquité beaucoup de tables trigonométriques ont été établies avec une précision croissante pour les besoins de l'astronomie. Le dernier ouvrage de l'astronome et mathématicien persan al Kashi, intitulé *Epitre de la Corde et du Sinus* a été achevé après sa mort par Qadi Zada al Rumi dans les années 1400.
Bien que perdu, cet ouvrage est toutefois connu à travers un commentaire des tables trigonométriques d'Ulugh-Beg. On y trouve le calcul de $\sin(1^\circ)$ avec une précision remarquable pour l'époque.
Le $\sin(3^\circ)$ peut être obtenu géométriquement et exprimé à l'aide de racines carrées. Diverses expressions de ce résultat ont été publiées, les unes plus compliquées que les autres, mais elles sont toutes bien évidemment égales entre elles. La plus compacte et aussi certainement la plus élégante est celle qui a été publiée en 1770 par le mathématicien alsacien Lambert dans sa table (XIX) des expressions des sinus des arcs croissant par trois degrés à partir de trois degrés.
On peut tous les calculer au moyen de $\sin(18^\circ)$, $\sin(30^\circ)$ et $\sin(45^\circ)$. Par exemple $\sin(48^\circ) = \sin(18^\circ + 30^\circ)$ et $\sin(3^\circ) = \sin(48^\circ - 45^\circ)$ Cette expression est la suivante

$$\sin(3^\circ) = \frac{1}{8}\left[-\sqrt{\frac{1}{2}} - \sqrt{\frac{3}{2}} + \sqrt{\frac{5}{2}} + \sqrt{\frac{15}{2}} - \sqrt{5+\sqrt{5}}\left(\sqrt{3}-1\right)\right]$$

On peut en obtenir une approximation, aussi précise que nécessaire, en utilisant par exemple la méthode de Héron d'Alexandrie pour le calcul des diverses racines carrées qui figurent dans cette expression. La relation reliant le sinus d'un angle au sinus de l'angle triple

$$\sin(3\theta) = 3\sin(\theta) - 4\sin(\theta)^3$$

était connue d'al Kashi qui pouvait donc accéder au calcul de $\sin(1^\circ)$ en résolvant l'équation algébrique du troisième degré

$$f_a(x) = a - 3x + 4x^3 = 0$$

dans laquelle l'inconnue est $x = \sin(1^\circ)$ et le paramètre $a = \sin(3^\circ)$.
Toutefois, la méthode directe de résolution des équations algébriques de degré trois ne sera découverte qu'à la Renaissance et al Kashi a utilisé une méthode itérative pour résoudre l'équation de point fixe équivalente

$$x = K_a(x) = \frac{1}{3}\left(a + 4x^3\right)$$

En partant d'une valeur initiale approchée pour $x$ et en itérant le procédé un nombre suffisant de fois, al Kashi obtient le nombre suivant, écrit en sexagésimal

$$0\,1\,2\,49\,43\,11\,14\,44\,16\,26\,17$$

c'est-à-dire

$$\begin{aligned} x &= \frac{0}{60^0} + \frac{1}{60^1} + \frac{2}{60^2} + \frac{49}{60^3} + \frac{43}{60^4} + \frac{11}{60^5} + \frac{14}{60^6} + \frac{44}{60^7} + \frac{16}{60^8} + \frac{26}{60^9} + \frac{17}{60^{10}} \\ &\simeq 0.017452406437283510 \end{aligned}$$

et on peut montrer que

$$|\sin(1^\circ) - 0.017452406437283510| \leq 2 \times 10^{-18}$$

Si on utilise l'approximation grossière (mais naturelle) $x_0 = 0$ en effectuant les calculs avec 18 décimales dans le seul but de comparer les résultats des itérations avec celui d'al Kashi on trouve alors que 6 itérations suffisent, mais bien entendu c'est une autre histoire de les effectuer sans machine comme a du le faire al Kashi !

$$\begin{aligned} x_1 &= 0.017445318747647944 \\ x_2 &= 0.017452397805531902 \\ x_3 &= 0.017452406426767058 \\ x_4 &= 0.017452406437270700 \\ x_5 &= 0.017452406437283497 \\ x_6 &= 0.017452406437283512 \\ x_7 &= 0.017452406437283512 \end{aligned}$$

**Exercice 6.1.** Déterminer géométriquement la valeur de $\sin(3°)$. [On pourra penser à $\sin(3°) = \sin(18° - 15°)$ puis déterminer $\sin(18°)$ géométriquement en traçant un triangle isocèle dont la base vaut $x$ à déterminer, les deux côté égaux valent 1 et les deux angles égaux à la base 72°.]

## 6.2 Typologie des méthodes itératives

### 6.2.1 Principe

Soit $E$ un espace vectoriel sur un corps $\mathbb{K}$ et $D$ une partie non vide de $E$. L'objectif est de résoudre l'équation $\underline{f}(\underline{x}) = \underline{0}$ où $\underline{f} : D \to E$ une application continue, linéaire ou non linéaire. Contrairement aux méthodes directes de résolution, les méthodes de résolution itératives se basent sur le principe de construction d'une suite d'éléments de $D$, en définissant une application continue $\underline{F} : D \to E$ telle que :

$$\exists \underline{x}^\star \in D / \underline{F}(\underline{x}^\star) = \underline{x}^\star \Leftrightarrow \underline{f}(\underline{x}^\star) = \underline{0}$$

En fonction de la classe de la méthode et du nombre de points d'évaluation, on construit par récurrence la suite $(\underline{x}^{(k)})_{k\in\mathbb{N}}$ des itérés successifs, tels que $\underline{x}^{(k+1)} = \underline{F}(\underline{x}^{(k)}, \underline{x}^{(k-1)}, \ldots)$ à partir des conditions initiales $\underline{x}^{(0)}, \underline{x}^{(1)}, \ldots$ Par conséquent $\underline{F}$ est appelée fonction d'itération. Si tous les itérés existent et que $\lim\limits_{k\to\infty} \underline{x}^{(k)} = \underline{x}^\star$, la continuité de $\underline{F}$ implique $\underline{F}(\underline{x}^\star) = \underline{x}^\star$.

**Remarque 6.1.**
Attention la réciproque n'est pas vraie : tout point fixe de $\underline{F}$ n'est pas nécessairement limite d'une suite itérative construite avec $\underline{F}$ en partant d'un $\underline{x}^{(0)} \neq \underline{x}^\star$.

On définit deux classes de méthodes itératives en fonction de l'information nécessaire au calcul de chaque itération : les méthodes avec ou sans mémoire.

### 6.2.2 Méthode itérative sans mémoire

#### Méthode à un point sans mémoire

**Définition 6.1.**
Nous appelons méthode à un point sans mémoire toute méthode itérative pour laquelle le calcul de $\underline{x}^{(k+1)}$ s'obtient uniquement grâce à une évaluation de $\underline{f}$ et éventuellement de ses dérivées. En choisissant un point $\underline{x}^{(0)} \in D$, on construit par récurrence la suite $(\underline{x}^{(k)})_{k\in\mathbb{N}}$ des itérés successifs, tels que :

$$\begin{cases} \underline{x}^{(0)} \in D \\ \underline{x}^{(k+1)} = \underline{F}(\underline{x}^{(k)}) \end{cases}$$

**Exemple 6.1.** Plusieurs exemples de méthodes à un point sans mémoire :

— Soient $\underline{\Omega}$ la fonction suffisament régulière telle que $\underline{\Omega}(\underline{0}) = \underline{0}$ et $\underline{F}$ définie telle que :

$$\underline{F}(\underline{x}) = \underline{x} + \underline{\Omega}(\underline{f}(\underline{x})) \Rightarrow \underline{F}(\underline{x}^\star) = \underline{x}^\star$$

Ainsi $\underline{F}$ est la fonction d'itération de la méthode à un point sans mémoire telle que :

$$\begin{cases} \underline{x}^{(0)} \in D \\ \underline{x}^{(k+1)} = \underline{F}(\underline{x}^{(k)}) = \underline{x}^{(k)} + \underline{\Omega}(\underline{f}(\underline{x}^{(k)})) \end{cases}$$

— Al-Kashi : on veut trouver les racines de $f_a(x) = 4x^3 - 3x + a$ grâce à la fonction d'itération

$$K_a(x) = \frac{1}{3}\left(a + 4x^3\right) = x + \frac{1}{3} f_a(x) \Rightarrow K_a(x^\star) = x^\star$$

.

— Dans le cas vectoriel notons $\underline{\underline{A}} \in \mathcal{M}_n(\mathbb{R})$, $\underline{b} \in \mathbb{R}^n$ tels que $\underline{f}(\underline{x}) = \underline{\underline{A}}.\underline{x} - \underline{b}$ (ce qui revient à résoudre le système linéaire $\underline{\underline{A}}.\underline{x} = \underline{b}$). Soit $\underline{\underline{M}} \in \mathcal{M}_n(\mathbb{R})$ inversible, alors on pose

$$\underline{F}(\underline{x}) = \underline{x} - \underline{\underline{M}}^{-1}.\underline{f}(\underline{x}) \Rightarrow \underline{F}(\underline{x}^\star) = \underline{x}^\star$$

.

— C'est le cas de la plupart des méthodes vues dans ce cours : Héron, Newton-Raphson, Halley, Chebychev, ...

**Méthode multi-points sans mémoire**

**Définition 6.2.**

Nous appelons méthode multi-points (ou à $r$ points) sans mémoire toute méthode itérative qui nécessite l'évaluation de $\underline{f}$ et éventuellement de ses dérivées en $r$ points (notés $\{\underline{x}^{(k)}\} \bigcup_{m\in[\![1;r-1]\!]}\{\underline{y}_m^{(k)}\}$) :

$$\begin{cases} \underline{x}^{(0)} \in D \\ \underline{x}^{(k+1)} = \underline{F}(\underline{x}^{(k)}, \underline{y}_1^{(k)}, \underline{y}_2^{(k)}, \ldots, \underline{y}_{r-1}^{(k)}) \end{cases}$$

**Exemple 6.2.** La méthode de Steffensen est définie par :

$$S_f(x) = x - \frac{f(x)^2}{f(x + f(x)) - f(x)}$$

C'est une méthode à deux points ($x$ et $y = x + f(x)$) sans mémoire ($x^{(k+1)}$ ne dépend que de $x^{(k)}$).

### 6.2.3 Méthode itérative à mémoire

**Définition 6.3.**

Nous appelons méthode itérative à mémoire une méthode itérative pour laquelle la détermination de $\underline{x}^{(k+1)}$ s'obtient non seulement en fonction de $\underline{x}^{(k)}$ mais aussi en réutilisant un nombre fini ($m$) de quantités numériques calculées lors des itérations antérieures. Par exemple dans le cas d'une méthode à un point :

$$\begin{cases} \underline{x}^{(0)}, \underline{x}^{(1)}, \ldots, \underline{x}^{(m-1)} \in D^m \\ \underline{x}^{(k+1)} = \underline{F}(\underline{x}^{(k)}, \underline{x}^{(k-1)}, \underline{x}^{(k-2)} \ldots, \underline{x}^{(k-m+1)}) \end{cases}$$

Ces méthodes peuvent être à un point ou à $r$ points.

**Exemple 6.3.** Méthode de la sécante : exemple de méthode à un point ($x^{(k)}$) avec un point de mémoire ($x^{(k-1)}$)

$$\begin{cases} x^{(0)}, x^{(1)} \in D^2 / f(x^{(0)}) \neq f(x^{(1)}) \\ x^{(k+1)} = x^{(k)} - f(x^{(k)}) \dfrac{x^{(k)} - x^{(k-1)}}{f(x^{(k)}) - f(x^{(k-1)})} \end{cases}$$

**Exemple 6.4.** Méthode de Traub : exemple de méthode à deux points ($x^{(k)}$, $x^{(k)} + \gamma^{(k)} f(x^{(k)})$) avec un point de mémoire ($x^{(k-1)}$)

$$\begin{cases} x^{(0)}, x^{(1)} \in D^2 / f(x^{(0)}) \neq f(x^{(1)}) \\ \gamma^{(k)} = -\dfrac{x^{(k)} - x^{(k-1)}}{f(x^{(k)}) - f(x^{(k-1)})} \\ x^{(k+1)} = x^{(k)} - \dfrac{\gamma^{(k)} f(x^{(k)})^2}{f(x^{(k)} + \gamma^{(k)} f(x^{(k)})) - f(x^{(k)})} \end{cases}$$

## 6.3 Etude de convergence dans $\mathbb{R}^n$

### 6.3.1 Bassin d'attraction

**Définition 6.4 (**Méthode à un pas**).**
On appelle point fixe attractif de $\underline{F}$ tout vecteur $\underline{x}^\star \in D$ s'il existe $U \subset D$ ouvert tel que :

(i) $\underline{x}^\star \in U$ est un point fixe de $\underline{F}$

(ii) $F(U) \subset U$

(iii) $\forall \underline{x}^{(0)} \in U, \quad \lim_{k\to\infty} \underline{x}^{(k)} = \underline{x}^\star$ où $\underline{x}^{(k+1)} = \underline{F}(\underline{x}^{(k)})$

Dans le cas contraire le point fixe est dit répulsif.

**Définition 6.5.**
On appelle bassin d'attraction de $\underline{x}^\star$ associé à la fonction d'itération $\underline{F}$, noté $\mathcal{B}_F(\underline{x}^\star)$, le plus grand ouvert (i.e la réunion des ouverts) vérifiant la définition précédente. Si $\mathcal{B}_F(\underline{x}^\star)$ est un sous ensemble propre de l'intérieur de $D$, noté $\overset{\circ}{D}$, alors la méthode itérative converge localement. Si $\mathcal{B}_F(\underline{x}^\star)$ coïncide avec $\overset{\circ}{D}$, alors la méthode itérative est converge globalement.

**Définition 6.6.**
On appelle sous-ensemble propre d'un ensemble $E$ tout sous-ensemble de $E$ distinct de $E$.

**Remarque 6.2.**
Pour une méthode à $r$ pas la définition précédente s'étend facilement.

**Exemple 6.5 (**Héron**).** On cherche les racines de $f_a(x) = x^2 - a$ où $a \in \mathbb{R}_+^\star$ grâce à la fonction d'itération $H_a(x) = \frac{1}{2}\left(x + \frac{a}{x}\right) = x - \frac{f_a(x)}{2x}$ définie sur $D_{H_a} = \mathbb{R}^\star$. Comme $H_a$ est une fonction impaire, nous n'étudierons dans un premier temps cette fonction que pour des réels positifs. Déterminons tout d'abord les tableaux de variation de $f_a$ et de $H_a$ sur $\mathbb{R}_+$ :

$$\forall x \in \mathbb{R},\ f_a'(x) = 2x \quad ; \quad \forall x \in \mathbb{R}^\star,\ H_a'(x) = \frac{x^2 - a}{2x}$$

Remarquons tout d'abord grâce au tableau de variation de $H_a$ que :

$$\forall x > 0 / x \neq \sqrt{a},\ H_a(x) > \sqrt{a},\ \text{i.e} \begin{cases} H_a(]0, \sqrt{a}[) & = ]\sqrt{a}, +\infty[ \\ H_a(]\sqrt{a}, +\infty[) & = ]\sqrt{a}, +\infty[ \end{cases}$$

Ensuite, du tableau de variation de $f_a$ que, $\forall x > 0$ :

$$H_a(x) - x = -\frac{f_a(x)}{2x} > 0 \Leftrightarrow f_a(x) < 0 \quad \forall 0 < x < \sqrt{a}$$
$$H_a(x) - x = -\frac{f_a(x)}{2x} < 0 \Leftrightarrow f_a(x) > 0 \quad \forall x > \sqrt{a}$$

Afin de déterminer le bassin d'attraction de cette méthode pour $x^\star = +\sqrt{a}$, énumérons tous les cas de figure (voir illustration 6.2) :

1. si $x^{(0)} > \sqrt{a}$, on a l'encadrement $\sqrt{a} < H_a(x^{(0)}) = x^{(1)} < x^{(0)}$ et $x^{(1)} \in ]\sqrt{a}, +\infty[$, la suite des itérés est décroissante et minorée par $\sqrt{a}$ donc convergente. Comme $x \mapsto H_a(x)$ est continue et que $\forall x > 0, H_a(x) \geq \sqrt{a}$, cette suite converge vers $\sqrt{a}$.
2. si $x^{(0)} = \sqrt{a}$ la suite des itérés est stationnaire et $\forall k, x^{(k)} = \sqrt{a}$
3. si $x^{(0)} < \sqrt{a}$, $H_a(x^{(0)}) = x^{(1)} \in ]\sqrt{a}, +\infty[$ donc dès la première itération on retombe sur le premier cas : en effet $H_a(]0, \sqrt{a}[) \subset ]\sqrt{a}, +\infty[$.

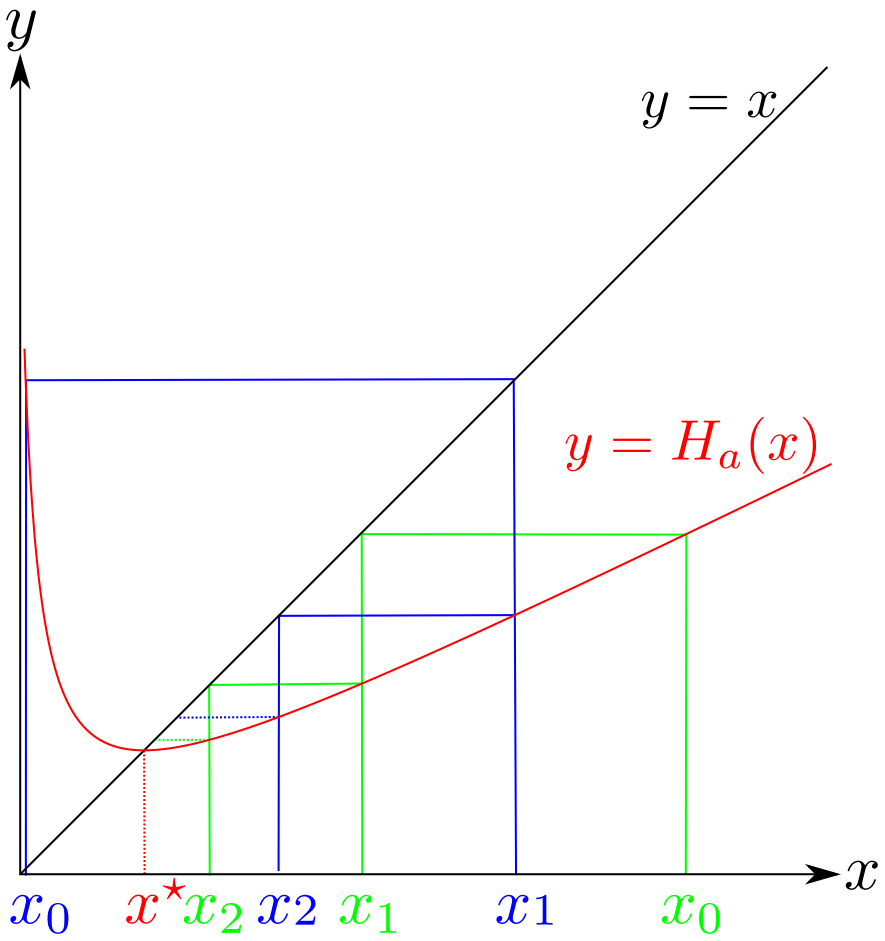



FIGURE 6.1 – Convergence des suites itératives avec $x^{(0)} > \sqrt{a}$ et $x^{(0)} < \sqrt{a}$

Au final, quel que soit $x^{(0)} \in \mathbb{R}_+^\star$, la suite des itérés est convergente et converge vers $\sqrt{a}$ donc $\mathcal{B}_{H_a}(\sqrt{a}) = \mathbb{R}_+^\star$ sous espace propre de $D_{H_a} = \mathbb{R}^\star$. Par conséquent la méthode est localement convergente. De même, par imparité de la fonction $H_a$, $\mathcal{B}_{H_a}(-\sqrt{a}) = \mathbb{R}_-^\star$. Au final on obtient $D_{H_a} = \mathcal{B}_{H_a}(-\sqrt{a}) \bigcup \mathcal{B}_{H_a}(+\sqrt{a})$

**Exercice 6.2.** Déterminer les bassins d'attraction pour la fonction $K_a$ de la méthode d'Al-Kashi.

### 6.3.2 Théorème d'Ostrowski scalaire

**Théorème 6.1.**
*Soient $D \subset \mathbb{R}$ un ouvert non vide, $F : D \to \mathbb{R}$ une fonction et $x^\star \in D$ un point fixe de $F$. Si $F$ est dérivable en $x^\star$ et si de plus $|F'(x^\star)| < 1$ alors il existe $r > 0$ et $I = ]x^\star - r; x^\star + r[ \subset D$ sur lequel $F(I) \subset I$. Par conséquent la suite définie par $\begin{cases} x^{(0)} \in I \\ x^{(k+1)} = F(x^{(k)}) \end{cases}$ reste dans $I$ et converge vers $x^\star$. Le point $x^\star$ est donc un point attractif de la méthode itérative associée à $F$.*

**Preuve.** *Soit $F$ dérivable en $x^\star \in D$. Il existe donc un voisinage de $x^\star$, noté $\mathcal{V}(x^\star)$, dans $D$ tel que :*

$$\lim_{\substack{x \to x^\star \\ x \in \mathcal{V}(x^\star)}} \frac{F(x) - F(x^\star)}{x - x^\star} = F'(x^\star)$$

| $x$ | $0$ | $\sqrt{a}$ | $+\infty$ |
|---|---|---|---|
| $f_a'(x)$ | | $+$ | |
| $f_a$ | $-a$ | $0$ | $+\infty$ |

| $x$ | $0$ | $\sqrt{a}$ | $+\infty$ |
|---|---|---|---|
| $H_a'(x)$ | | $-$ $0$ $+$ | |
| $H_a$ | $+\infty$ | $\sqrt{a}$ | $+\infty$ |

FIGURE 6.2 – Héron : tableaux de variations des fonctions $f_a$ et $H_a$

*Par conséquent* $\forall \varepsilon > 0$, $\exists r > 0$ *tels que* $I = ]x^\star - r; x^\star + r[$ *et*

$$\begin{aligned} \forall x \in I, \left| \frac{F(x) - F(x^\star)}{x - x^\star} - F'(x^\star) \right| &\leq \varepsilon \\ \Leftrightarrow \left| F(x) - F(x^\star) - F'(x^\star)\,(x - x^\star) \right| &\leq \varepsilon \left| x - x^\star \right| \end{aligned}$$

*Ainsi :*

$$\begin{aligned} |F(x) - x^\star| = |F(x) - F(x^\star)| &= \left| F(x) - F(x^\star) - F'(x^\star)\,(x - x^\star) + F'(x^\star)\,(x - x^\star) \right| \\ &\leq \left| F(x) - F(x^\star) - F'(x^\star)\,(x - x^\star) \right| + \left| F'(x^\star) \right| \left| x - x^\star \right| \\ &\leq \left( \varepsilon + \left| F'(x^\star) \right| \right) \left| x - x^\star \right| = C \left| x - x^\star \right| \end{aligned}$$

*Or par hypothèse* $|F'(x^\star)| < 1$ *donc on peut choisir* $\varepsilon > 0$ *tel que* $0 < C < 1$. *Finalement, par récurrence on obtient* $\forall k \in \mathbb{N}^\star$, $\left| x^{(k)} - x^\star \right| \leq C^k \left| x^{(0)} - x^\star \right|$ *et* $\lim\limits_{k \to \infty} x^{(k)} = x^\star$.

**Remarque 6.3.**
Le théorème d'Ostrowski ne dit pas que si $|F'(x^\star)| \geq 1$ alors $x^\star$ est un point fixe répulsif. Dans ce cas il faut déterminer le bassin d'attraction "à la main" comme vu précédemment.

**Définition 6.7.**
Soient $U \subset \mathbb{R}$ ouvert non vide et $F : U \to \mathbb{R}$ une fonction. $x^\star \in U$ point fixe de $F$ est dit super-attractif si $F$ est dérivable en $x^\star$ et $F'(x^\star) = 0$.

**Exercice 6.3.** Déterminer par le théorème d'Ostrowski si les points fixes des méthodes de Héron et d'Al-Kashi sont attractifs.

### 6.3.3 Théorème d'Ostrowski vectoriel (Admis)

**Définition 6.8 (**Différentielle au sens de Fréchet**).**
Soient $(E, \|.\|_E)$ et $(F, \|.\|_F)$ deux espaces vectoriels normé sur $\mathbb{K}$, $U \subset E$ un ouvert non vide, $f : U \to F$ une application et $x$ un point de $U$. On dit que $f$ est différentiable au point $x$ s'il existe une application linéaire continue $df_x \in \mathcal{L}(E, F)$ telle que :

$$\lim_{\|h\|_E \to 0} \frac{\|f(x+h) - f(x) - df_x(h)\|_F}{\|h\|_E} = 0$$

L'application linéaire et continue $df_x$ est appelée différentielle de $f$ en $x$.

**Remarque 6.4.**
On en déduit que pour $h \in E$ tel que $h + x \in U$ le développement limité :

$$f(x+h) = f(x) + df_x(h) + o(\|h\|_E)$$

**Théorème 6.2.**
*Soient $D \subset \mathbb{R}^d$ un ouvert non vide, $\underline{F} : D \to \mathbb{R}^d$ une application et $\underline{x}^\star \in D$ un point fixe de $\underline{F}$. Si $\underline{F}$ est différentiable en $\underline{x}^\star$ et si de plus le rayon spectral de la jacobienne de $\underline{F}$ en $\underline{x}^\star$, noté $\rho(\underline{J}_{\underline{F}}(\underline{x}^\star))$, vérifie $\rho(\underline{J}_{\underline{F}}(\underline{x}^\star)) < 1$ alors il existe $r > 0$ tel que la boule ouverte $B(\underline{x}^\star, r) \subset D$ sur lequel $F(B(\underline{x}^\star, r)) \subset B(\underline{x}^\star, r)$. Par conséquent la suite définie par $\begin{cases} \underline{x}^{(0)} \in B(\underline{x}^\star, r) \\ \underline{x}^{(k+1)} = \underline{F}(\underline{x}^{(k)}) \end{cases}$ reste dans $B(\underline{x}^\star, r)$ et converge vers $\underline{x}^\star$. Le point $\underline{x}^\star$ est donc un point attractif de la méthode itérative associée à $\underline{F}$.*

### 6.3.4 Théorème de point fixe

**Théorème 6.3 (**du point fixe, Picard**).**
*Soient $(E, d)$ un espace métrique complet et $D \subset E$. Si $\underline{F} : D \to D$ est strictement contractante de coefficient $0 < \ell < 1$, alors :*
*(i) $\underline{F}$ admet un unique point fixe : $\exists! \underline{x}^\star \in D / \underline{F}(\underline{x}^\star) = \underline{x}^\star$*
*(ii) Pour tout $\underline{x}^{(0)} \in D$, la suite $\begin{cases} \underline{x}^{(0)} \in D \\ \underline{x}^{(k+1)} = \underline{F}(\underline{x}^{(k)}) \end{cases}$ converge vers $\underline{x}^\star$*
*On a de plus :*
*(iii) $\forall k \in \mathbb{N}$, $d(\underline{x}^{(k+1)}, \underline{x}^\star) \leq \ell d(\underline{x}^{(k)}, \underline{x}^\star)$*
*(iv) $\forall k \in \mathbb{N}$, $d(\underline{x}^{(k+1)}, \underline{x}^\star) \leq \frac{1}{1-\ell} d(\underline{x}^{(k)}, \underline{x}^{(k+1)}) \leq \frac{\ell^k}{1-\ell} d(\underline{x}^{(0)}, \underline{x}^{(1)})$*

**Preuve.** *Commençons par démontrer le point (i) en deux étapes, existence et unicité :*

***Point (i-1) Existence :***
*L'idée est de montrer que la suite $\begin{cases} \underline{x}^{(0)} \in D \\ \underline{x}^{(k+1)} = \underline{F}(\underline{x}^{(k)}) \end{cases}$ est une suite de Cauchy : en effet, comme $E$ est complet, toute suite de Cauchy converge. Grâce à l'inégalité triangulaire et à la propriété de contraction de $F$, on a :*

$$\begin{aligned} \forall (p, q) \in \mathbb{N},\ d(\underline{x}^{(p)}, \underline{x}^{(p+q)}) &\leq d(\underline{x}^{(p)}, \underline{x}^{(p+1)}) + d(\underline{x}^{(p+1)}, \underline{x}^{(p+2)}) + \ldots + d(\underline{x}^{(p+q-1)}, \underline{x}^{(p+q)}) \\ &\leq d(\underline{x}^{(p)}, \underline{x}^{(p+1)}) \left[1 + \ell + \ldots + \ell^{q-1}\right] = d(\underline{x}^{(p)}, \underline{x}^{(p+1)}) \frac{1 - \ell^q}{1 - \ell} \end{aligned}$$

*Or, par récurrence, la distance entre deux termes consécutifs est majorée par :*

$$\forall k \in \mathbb{N},\ d(\underline{x}^{(k)}, \underline{x}^{(k+1)}) = d(\underline{F}(\underline{x}^{(k-1)}), \underline{F}(\underline{x}^{(k)})) \leq \ell d(\underline{x}^{(k-1)}, \underline{x}^{(k)}) \leq \ldots \leq \ell^k d(\underline{x}^{(0)}, \underline{x}^{(1)})$$

*Ainsi :*

$$\forall (p, q) \in \mathbb{N},\ d(\underline{x}^{(p)}, \underline{x}^{(p+q)}) \leq \frac{1 - \ell^q}{1 - \ell} d(\underline{x}^{(p)}, \underline{x}^{(p+1)}) \leq \ell^p \frac{1 - \ell^q}{1 - \ell} d(\underline{x}^{(0)}, \underline{x}^{(1)}) \tag{6.5}$$

*Au final, comme $\ell < 1$ :*

$$\lim_{(p,q) \to +\infty} d(\underline{x}^{(p)}, \underline{x}^{(p+q)}) = 0$$

*La suite $(\underline{x}^{(k)})_k$ est donc bien une suite de Cauchy dans $E$ complet, donc $(\underline{x}^{(k)})_k$ converge vers $\underline{x}^\star \in E$.*

***Point (i-2)Unicité :***
*Par l'absurde, s'il existait $\underline{x}^\star, \tilde{\underline{x}}$ deux points fixes distincts ($d(\underline{x}^\star, \tilde{\underline{x}}) > 0$) de $F$ alors :*

$$0 \leq d(\underline{x}^\star, \tilde{\underline{x}}) = d(\underline{F}(\underline{x}^\star), \underline{F}(\tilde{\underline{x}})) \leq \ell d(\underline{x}^\star, \tilde{\underline{x}}) \Rightarrow \ell \geq 1$$

*contradictoire avec la propriété de contraction stricte de $\underline{F}$.*

***Point (ii) :*** *l'application $\underline{F}$ étant continue, par passage à la limite $\underline{x}^\star$ est bien point fixe de $\underline{F}$ :*

$$\lim_{k\to+\infty} \underline{x}^{(k)} = \underline{x}^\star \Rightarrow \underline{x}^{(k+1)} = \underline{F}(\underline{x}^{(k)}) \xrightarrow[k\to+\infty]{} \underline{x}^\star = \underline{F}(\underline{x}^\star)$$

***Point (iii) :*** *comme $\underline{x}^\star$ point fixe de $\underline{F}$ on a :*

$$\forall k \in \mathbb{N},\ d(\underline{x}^{(k+1)}, \underline{x}^\star) = d(\underline{F}(\underline{x}^{(k)}), \underline{F}(\underline{x}^\star)) \leq \ell d(\underline{x}^{(k)}, \underline{x}^\star)$$

***Point (iv) :*** *$F$ contraction stricte de coefficient $\ell < 1 \Rightarrow \ell^q \xrightarrow[q\to+\infty]{} 0$, donc en ne faisant tendre que $q$ vers l'infini dans l'équation* (6.5)*, on obtient :*

$$\forall p \in \mathbb{N},\ d(\underline{x}^{(p)}, \underline{x}^\star) \leq \frac{1}{1-\ell} d(\underline{x}^{(p)}, \underline{x}^{(p+1)}) \leq \frac{\ell^p}{1-\ell} d(\underline{x}^{(0)}, \underline{x}^{(1)})$$

**Remarque 6.5.**

Le point (iv) du théorème (6.3) permet d'estimer le nombre d'itérations nécessaires pour atteindre le point fixe à une erreur $\mu$ (donné) près : en effet, la grandeur $d(\underline{x}^{(k+1)}, \underline{x}^\star) = \varepsilon^{(k)}$ correspond à la distance entre l'itération $k$ et la solution théorique, i.e l'erreur à l'itération $k$. A une valeur $\mu$ donnée ; et connaissant $\ell$ le coefficient de Lipschitz de $\underline{F}$ et la distance entre les deux premières itérations $d(\underline{x}^{(0)}, \underline{x}^{(1)})$, le nombre d'itérations $k^\star$ permettant d'estimer $\underline{x}^\star$ à $\mu$ près est au plus :

$$\mu \leq \frac{\ell^k}{1-\ell} d(\underline{x}^{(0)}, \underline{x}^{(1)}) \Rightarrow k \geq k^\star = \frac{\ln\left(\frac{\mu(1-\ell)}{d(\underline{x}^{(0)}, \underline{x}^{(1)})}\right)}{\ln(\ell)} \tag{6.6}$$

**Exercice 6.4.** Soient $(a, b, c, d) \in \mathbb{R}^4$ tels que $\gamma = \max(|a| + |b|, |c| + |d|) < 1$ et le système :

$$(S) \begin{cases} a\sin(x) + b\cos(y) - x = 0 \\ c\cos(x) + d\sin(y) - y = 0 \end{cases}$$

Montrer que $(S)$ admet une unique solution $\underline{x}^\star \in \mathbb{R}^2$. Proposer une estimation par excès du nombre d'itérations nécessaires pour obtenir une approximation de $\underline{x}^\star$ à $10^{-4}$ près en partant de $\underline{X}^{(0)} = (0, 0)$ et avec $(a, b, c, d) = (0.1, 0.5, 0.3, -0.4)$. Cette estimation est-elle satisfaisante ?

## 6.4 Ordre de convergence

### 6.4.1 Illustration numérique

Afin d'introduire la notion d'ordre de convergence d'une méthode itérative, prenons l'exemple de deux méthodes permettant d'évaluer la racine carrée d'un réel positif $a$ : la méthode de Héron et la méthode de Lambert (voir exercice 6.7). On montrera dans cet exercice que, tout comme pour la méthode de Héron, la méthode de Lambert est localement convergente pour $\sqrt{a}$ (point fixe de la méthode) et que son bassin d'attraction est $\mathbb{R}_+^\star$. Parmi ces deux méthodes, laquelle choisir pour obtenir une approximation de $\sqrt{a}$ ? Quels critères sont à prendre en compte ?

Par exemple prenons $a = 2$ et en partant de $x^{(0)} = 2$, cherchons combien d'itérations sont nécessaires pour obtenir une approximation de $\sqrt{2}$ avec 190 décimales exactes. Sur la figure 6.4.1 est représenté, pour chaque itération, le nombre de décimales exactes pour les méthodes de Héron (en bleu) et de Lambert (en rouge). Cette vitesse de convergence représente l'ordre de convergence de la méthode. On verra en effet que la méthode de Héron est d'ordre 2 alors que la méthode de Lambert est d'ordre 3. Ainsi, selon ce critère de vitesse de convergence, clairement la méthode de Lambert semble préférable à la méthode de Héron.

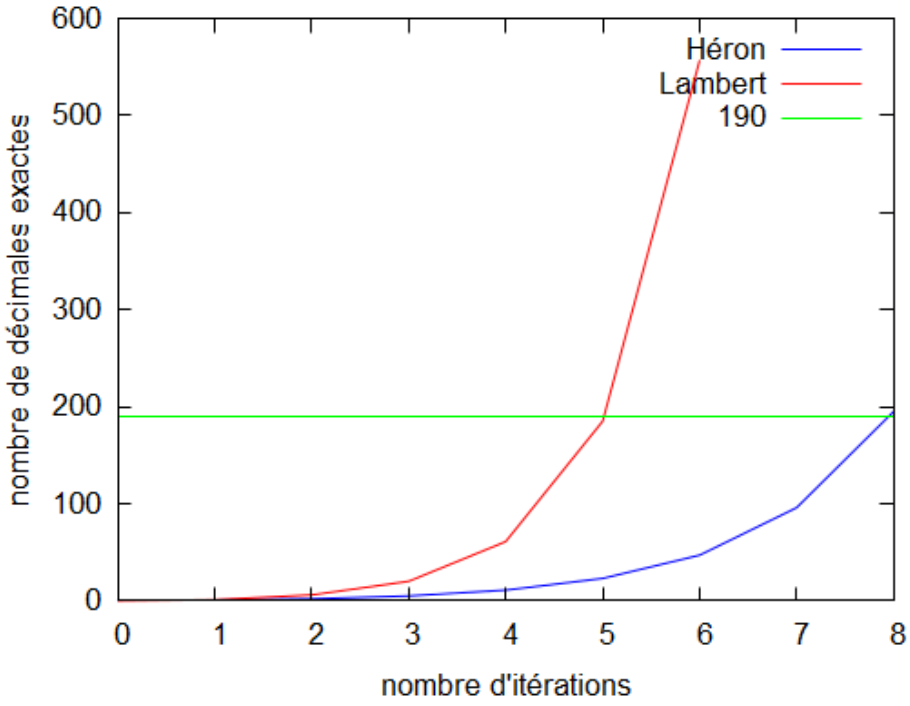



FIGURE 6.3 – Ordre de convergence : évolution du nombre de décimales exactes en fonction du nombre d'itérations pour les méthodes de Héron et de Lambert.

### 6.4.2 Ordre de convergence d'une suite convergente

**Définitions - propositions**

**Proposition 6.1.** *Soient $(E,d)$ espace métrique et $\{x^{(k)}\}_{k\in\mathbb{N}}$ suite de $E$ qui converge vers $x^\star \in E$. S'il existe $N \in \mathbb{N}^\star$ et $(a,b,\tau) \in (\mathbb{R}_+^\star)^3$ tels que :*

$$\forall k \geq N,\ x^{(k)} \neq x^\star \text{ et } a\, d(x^{(k)},x^\star)^\tau \leq d(x^{(k+1)},x^\star) \leq b\, d(x^{(k)},x^\star)^\tau$$

*alors $\tau \geq 1$ et si $\tau = 1$ alors $0 < a < 1$.*

**Preuve.** *Notons $\varepsilon^{(k)} = d(x^{(k)},x^\star)$ l'erreur au rang $k$. Par hypothèse $\lim\limits_{k\to\infty} \varepsilon^{(k)} = 0$ et $a\,\varepsilon^{(k)^\tau} \leq \varepsilon^{(k+1)}$.*

- *Montrons par récurrence que $(H)\ \forall k \geq N,\ \varepsilon^{(k)} \geq C\, D^{\tau^{k-N}}$ où $C = a^{\frac{1}{1-\tau}} > 0$ et $D = \varepsilon^{(N)}\, a^{\frac{-1}{1-\tau}} > 0$*
  - *$\varepsilon^{(N)} = CD$ donc $(H)$ est vérifiée pour $k = N$*
  - *Supposons que $(H)$ est vérifiée pour $k \geq N$ :*

$$\varepsilon^{(k+1)} \geq a\,\varepsilon^{(k)^\tau} \geq a\, C^\tau\, D^{\tau^{k+1-N}}$$

  *or $a\, C^\tau = a\, a^{\frac{\tau}{1-\tau}} = C$ donc $\varepsilon^{(k+1)} \geq C\, D^{\tau^{k+1-N}}$.*
- *Si $0 < \tau < 1$ alors $\lim\limits_{k\to+\infty} \tau^{k-N} = 0$. Or on a montré que*

$$\varepsilon^{(k)} \geq C\, D^{\tau^{k-N}} \qquad \text{avec} \begin{cases} \lim\limits_{k\to\infty} \varepsilon^{(k)} = 0 \\ \lim\limits_{k\to\infty} C\, D^{\tau^{k-N}} = C > 0 \end{cases}$$

*D'où la contradiction. Ainsi $\tau \geq 1$.*

— *Supposons que* $\tau = 1$ *et* $a \geq 1$ *tels que* $\forall k \in \mathbb{N}$, $a\,\varepsilon^{(k)} \leq \varepsilon^{(k+1)}$. *Montrons par récurrence que* $(\tilde{H})$ $\forall k \geq N$, $\varepsilon^{(k)} \geq \varepsilon^{(N)}\,a^{k-N}$ *:*
— $\varepsilon^{(N)} \geq \varepsilon^{(N)}$ *donc* $(\tilde{H})$ *est vérifiée pour* $k = N$
— $\varepsilon^{(k+1)} \geq a\,\varepsilon^{(k)} \geq a^{k+1-N}\,\varepsilon^{(N)}$.
*Pat conséquent :*
— *Si* $a = 1$, $\varepsilon^{(k)} \geq \varepsilon^{(N)} \neq 0$ *or* $\lim\limits_{k\to\infty} \varepsilon^{(k)} = 0$ *d'où la contradiction*
— *Si* $a > 1$, $\varepsilon^{(k)} \geq \varepsilon^{(N)}\,a^{k-N} \xrightarrow[k\to\infty]{} +\infty$ *d'où la contradiction*
*Ainsi si* $\tau = 1$ *alors nécessairement* $0 < a < 1$.

**Définition 6.9.**
Soient $(E, d)$ espace métrique et $\{x^{(k)}\}_{k\in\mathbb{N}}$ suite de $E$ qui converge vers $x^\star \in E$. On dit que la suite est d'ordre $\tau \in [1, +\infty[$ s'il existe $(a, b, \tau) \in (\mathbb{R}_+^\star)^3$ tels que :

$$0 < a < b \quad \text{et} \quad \exists N \in \mathbb{N}^\star / \forall k \geq N, a\,d(x^{(k)}, x^\star)^\tau \leq d(x^{(k+1)}, x^\star) \leq b\,d(x^{(k)}, x^\star)^\tau$$

et si $\tau = 1$ alors $0 < a < 1$.

**Remarque 6.6.**
Si $\tau = 1$ la convergence est dite linéaire, si $\tau = 2$ elle est dite quadratique, mais l'ordre de convergence n'est pas nécessairement entier, comme le montre l'exemple suivant.

**Exemple 6.6.** $\forall \tau \geq 1$ il existe des suites convergentes d'ordre $\tau$.
— Considérons dans un premier temps la suite $\{x^{(k)}\}_{k\in\mathbb{N}}$ telle que $\forall C \in ]0, 1[$, $x^{(k)} = C^k \xrightarrow[k\to\infty]{} 0$. De plus, $\frac{C}{2} < C$ donc $\frac{C}{2}x^{(k)} \leq x^{(k+1)} \leq C\,x^{(k)}$ la suite est d'ordre 1.
— Soit maintenant $\forall C \in ]0, 1[, \forall \tau > 1$, $x^{(k)} = C^{\tau^k} \xrightarrow[k\to\infty]{} 0$. On a clairement que $\frac{1}{2}\,x^{(k)^\tau} < x^{(k)^\tau} = x^{(k+1)}$ donc

$$\frac{1}{2}x^{(k)^\tau} \leq x^{(k+1)} \leq 2x^{(k)^\tau}$$

La suite $\{x^{(k)}\}_{k\in\mathbb{N}}$ est d'ordre $\tau$.

**Remarque 6.7.**
Une suite peut être convergente sans avoir d'ordre de convergence.

**Exemple 6.7.** Soit la suite $\{x^{(k)}\}_{k\in\mathbb{N}}$ telle que $x^{(2q)} = \frac{1}{(1+q)^2}$ ; $x^{(2q+1)} = \frac{1}{1+q}$. Clairement $\forall k \in \mathbb{N}$, $x^{(k)} > 0$ et $\lim\limits_{k\to\infty} x^{(k)} = 0$.

Supposons qu'il existe $b > 0$, $\tau \geq 1$ et $N \in \mathbb{N}^\star$ tels que $\forall k \in \mathbb{N}$, $x^{(k+1)} \leq b\,x^{(k)^\tau}$. Alors :

$$\frac{1}{1+q} \leq b\left(\frac{1}{(1+q)^2}\right)^\tau \Leftrightarrow (1+q)^{2\tau-1} \leq b$$

Or $2\tau - 1 > 0$ donc pour $k$ suffisament grand $(1+q)^{2\tau-1}$ ne peut être borné. Cette suite n'a donc pas d'ordre de convergence.

**Proposition 6.2.** *Soient* $(E, d)$ *espace métrique et* $\{x^{(k)}\}_{k\in\mathbb{N}}$ *suite convergente de* $E$. *Si la suite est d'ordre* $\tau \geq 1$ *alors son ordre est unique.*

**Preuve.** *Soit $x^\star \in E$ et supposons deux ordres différents $1 \leq \tau_1 \leq \tau_2$. Par conséquent $\exists (N_1, N_2) \in \mathbb{N}^\star \times \mathbb{N}^\star$ et $0 < a_1 < b_1$ et $0 < a_2 < b_2$ tels que*

$$\begin{cases} \forall k \geq N_1, & a_1 {\varepsilon^{(k)}}^{\tau_1} \leq \varepsilon^{(k+1)} \leq b_1 {\varepsilon^{(k)}}^{\tau_1} \\ \forall k \geq N_2, & a_2 {\varepsilon^{(k)}}^{\tau_2} \leq \varepsilon^{(k+1)} \leq b_2 {\varepsilon^{(k)}}^{\tau_2} \end{cases}$$

*Soit $N = \max(N_1, N_2)$, alors*

$$\forall k \geq N,\ a_1 {\varepsilon^{(k)}}^{\tau_1} \leq \varepsilon^{(k+1)} \leq b_2 {\varepsilon^{(k)}}^{\tau_2}$$

*Si on divise l'inégalité par $b_2\, {\varepsilon^{(k)}}^{\tau_1} > 0$ on obtient :*

$$0 < \frac{a_1}{b_2} \leq {\varepsilon^{(k)}}^{\tau_2 - \tau_1} \xrightarrow[k\to\infty]{} 0$$

*Ce qui est impossible.*

**Condition suffisante pour qu'une suite convergente soit d'ordre $\tau \geq 1$**

**Théorème 6.4.**
*Soient $(E, d)$ espace métrique et $\{x^{(k)}\}_{k\in\mathbb{N}}$ suite de $E$ qui converge vers $x^\star \in E$ tels que*

$$\exists \tau \in [1, +\infty[,\ \exists N \in \mathbb{N}^\star / \forall k \geq N, x^{(k)} \neq x^\star \text{ et } \lim_{k\to\infty} \frac{\varepsilon^{(k+1)}}{{\varepsilon^{(k)}}^\tau} = C > 0$$

*alors la suite est d'ordre $\tau$. Si de plus $\tau = 1$ alors on a nécessairement $0 < C \leq 1$.*

**Preuve.** *Soit $\mu \in \mathbb{R}$ tel que $0 < \mu < C$. L'existence de la limite $\lim_{k\to\infty} \frac{\varepsilon^{(k+1)}}{{\varepsilon^{(k)}}^\tau} = C > 0$ prouve que :*

$$\exists M > N / \forall k \geq M,\ \left| \frac{\varepsilon^{(k+1)}}{{\varepsilon^{(k)}}^\tau} - C \right| \leq \mu$$

*De plus $\forall k \geq M > N$, $\varepsilon^{(k)} > 0$ d'où :*

$$(C - \mu)\, {\varepsilon^{(k)}}^\tau \leq \varepsilon^{(k+1)} \leq (C + \mu)\, {\varepsilon^{(k)}}^\tau$$

*Ainsi la suite est d'ordre $\tau$. De plus on a montré que si $\tau = 1$ alors $\forall \mu \in ]0, C[$, $C - \mu < 1$ donc $C < 1 + \mu$ et $C \leq 1$ : en effet par l'absurde, si $C > 1$, alors en choisissant $0 < \mu = C - 1$ on obtient : $C < 1 + \mu \Rightarrow C < C$ ce qui est absurde.*

**Remarque 6.8.**
Une suite convergente peut avoir un ordre de convergence $\tau \geq 1$ sans que $\lim_{k\to\infty} \frac{\varepsilon^{(k+1)}}{{\varepsilon^{(k)}}^\tau}$ n'existe.

**Exemple 6.8.** Soient $(a, b, x^{(0)}, \tau) \in \mathbb{R}^4$ tels que $0 < a < b < 1$, $0 < x^{(0)} < 1$ et $\tau \geq 1$. On définit la suite $\{y^{(k)}\}_{k\in\mathbb{N}}$ telle que $y^{(k)} = \frac{1+\sin(k)}{2} \in [0, 1]$ qui n'a pas de limite. De plus, soit la suite $\{x^{(k)}\}_{k\in\mathbb{N}}$ telle que

$$x^{(k+1)} = \left(a + (b - a) y^{(k)}\right) {x^{(k)}}^\tau \leq b {x^{(k)}}^\tau$$

Par récurrence on a

$$\forall k \in \mathbb{N}^\star,\ 0 \leq x^{(k)} \leq b^{\sum_{i=0}^{k-1} \tau^i}\, {x^{(0)}}^{\tau^k}$$

or $b < 1$ donc $x^{(k)} \xrightarrow[k\to\infty]{} x^\star = 0$. Par ailleurs on a aussi

$$a {\varepsilon^{(k)}}^\tau = a {x^{(k)}}^\tau \leq \varepsilon^{(k+1)} = x^{(k+1)} \leq b\, {\varepsilon^{(k)}}^\tau = b {x^{(k)}}^\tau$$

donc la suite est d'ordre $\tau$. Cependant la quantité

$$\left|\frac{\varepsilon^{(k+1)}}{{\varepsilon^{(k)}}^{\tau}}\right| = \frac{x^{(k+1)}}{{x^{(k)}}^{\tau}} = a + (b-a)y^{(k)}$$

n'a pas de limite.

### 6.4.3 Ordre de convergence d'une méthode itérative scalaire localement convergente

**Théorème 6.5.**
*Soient $U \subset \mathbb{R}$ ouvert non vide et $F : U \to \mathbb{R}$ une fonction. Si $F$ admet en un point $x^\star \in U$ le développement limité d'ordre $p \geq 1$ suivant :*

$$F(x^\star + h) = x^\star + A\,h^p + o(h^p) \qquad \text{avec } A \neq 0 \text{ et } |A| < 1 \text{ si } p = 1$$

— *Si $p = 1$, alors $x^\star$ est un point fixe attractif de $F$ (th d'Ostrowski $|F'(x^\star)| < 1$)*
— *Si $p \geq 2$, alors $x^\star$ est un point fixe super-attractif de $F$ ($F'(x^\star) = 0$)*
*Il existe de plus un ouvert $V \subset U$ contenant $x^\star$ tel que toutes les suites*

$$\begin{cases} x^{(0)} \in V \\ x^{(k+1)} = F(x^{(k)}) \end{cases}$$

*sont dans $V$ et sont toutes convergentes d'ordre $p$.*

**Preuve.** *Soit le développement en séries de Taylor, en posant $h = x^{(k)} - x^\star = \varepsilon^{(k)}$ :*

$$F(x^\star + h) = F(x^\star) + hF'(x^\star) + \frac{h^2}{2}F''(x^\star) + \ldots + \frac{h^p}{p!}F^{(p)}(x^\star) + o(h^p)$$

— *Si $p = 1$, $A = F'(x^\star)$ or $|A| < 1$*
— *Si $p \geq 2$, $F'(x^\star) = 0$*
*De plus*

$$\lim_{k\to\infty} \frac{|\varepsilon^{(k+1)}|}{|\varepsilon^{(k)}|^p} = \lim_{k\to\infty} \frac{|F(x^{(k)}) - F(x^\star)|}{|\varepsilon^{(k)}|^p} = \lim_{k\to\infty} \left|\frac{F(x^\star + h) - F(x^\star)}{h^p}\right| = \lim_{k\to\infty} |A + o(1)| = |A| > 0$$

*La suite est donc convergente d'ordre $p$.*

**Remarque 6.9.**

— L'ordre commun de toutes ces suites définies dans un voisinage $V$ de $x^\star$ est nécessairement l'entier $p$ ;
— Si $p \geq 2$ et si on n'a pas d'autre hypothèse sur la régularité de $F$ on ne peut rien affirmer sur l'existence des dérivées d'ordre supérieur de $F$ en $x^\star$ ;
— En pratique si $F$ est suffisament dérivable on calcule les dérivées successives en $x^\star$ jusqu'à la première dérivée non nulle, d'où la proposition suivante.

**Théorème 6.6.**
*Soit $F \in \mathcal{C}^p(V)$ telle que $\forall i \in [\![1; p-1]\!]$, $F^{(i)}(x^\star) = 0$ et $F^{(p)}(x^\star) \neq 0$ alors la méthode itérative associée à $F$ converge et son ordre de convergence est $p$. On a alors*

$$\lim_{k\to\infty} \left|\frac{x^{(k+1)} - x^\star}{(x^{(k)} - x^\star)^p}\right| = \left|\frac{F^{(p)}(x^\star)}{p!}\right|$$

**Preuve.** *Il suffit d'écrire le développement en série de Taylor de $F$ à l'ordre $p$ en prenant en compte les dérivées nulles :*

$$x^{(k+1)} = F(x^{(k)}) = F\left(x^\star + \left(x^{(k)} - x^\star\right)\right) = F(x^\star) + \frac{F^{(p)}(x^\star)}{p!} + o\left(\left(x^{(k)} - x^\star\right)^p\right)$$

*D'où*

$$\lim_{k\to\infty} \left| \frac{x^{(k+1)} - x^\star}{(x^{(k)} - x^\star)^p} \right| = \left| \frac{F^{(p)}(x^\star)}{p!} \right|$$

**Exercice 6.5 (**Echauffement aux développements limités **).** Soient $(A, B, \alpha, \beta, \gamma)$ cinq réels non nuls tels que $D_1$ et $D_2$ sont définis par les grandeurs :

$$D_1 = Ah + Bh^2 + o(h^2) \quad ; \quad D_2 = \alpha h^2 + \beta h^3 + \gamma h^4 + o(h^4)$$

Calculer les développements limités suivants :

$$dl_1 = D_1 \times D_2 \,;\, dl_2 = D_1^2 \,;\, dl_3 = D_2^2 \,;\, dl_4 = \frac{1}{D_2} \,;\, dl_5 = \frac{D_1}{D_2} \,;\, dl_6 = \frac{D_2}{D_1}$$

**Exercice 6.6.** Déterminer l'ordre de convergence de la méthode de Héron définie par $H_a(x) = \frac{1}{2}\left(x + \frac{a}{x}\right)$.

**Exercice 6.7.** On définit la méthode de Lambert par $L_a(x) = \frac{3a+x^2}{a+3x^2}x$ où $a \in \mathbb{R}_+^\star$, permettant d'extraire les racines carrées de $a$. Déterminer :

- les points fixes de $L_a$
- les bassins d'attraction des points fixes attractifs.
- l'ordre de convergence de la méthode pour ces derniers.

### 6.4.4 Ordre de convergence d'une méthode itérative vectorielle localement convergente

**Théorème 6.7.**

*Soient $(E, \|.\|)$ un espace normé réel et $U$ une partie non vide de $E$. Si une application $\underline{F} : E \to E$ admet en un point $\underline{x}^\star \in U$ le développement limité d'ordre $p \geq 1$ suivant :*

$$\underline{F}(\underline{x}^\star + \underline{h}) = \underline{x}^\star + \underline{\underline{M}}\,(\underline{h}, \ldots, \underline{h}) + o(\|\underline{h}\|^p)$$

*où $\underline{\underline{M}}$ est une application d-linéaire, non nulle, symétrique et continue de $E^d$ dans $E$, de rayon spectral $\rho(\underline{\underline{M}}) < 1$ :*

- *si $p = 1$, le point fixe $\underline{x}^\star$ est un point attractif de la méthode itérative définie par $\underline{F}$ et cette méthode est d'ordre 1*
- *si $p \geq 2$, le point fixe $\underline{x}^\star$ est un point super-attractif de la méthode itérative définie par $\underline{F}$ et cette méthode est d'ordre $p$*

**Exemple 6.9 (**Méthode de Schultz**).** Soient $d \in \mathbb{N}^\star$ et $\underline{\underline{A}} \in \mathcal{M}_d(\mathbb{C})$ inversible. On considère l'application :

$$\begin{aligned} \underline{\underline{F}}_A :& \mathcal{M}_d(\mathbb{C}) \to \mathcal{M}_d(\mathbb{C}) \\ & \underline{\underline{X}} \mapsto \underline{\underline{X}}.\left(2.\underline{\underline{I}}_d - \underline{\underline{A}}.\underline{\underline{X}}\right) = \left(2.\underline{\underline{I}}_d - \underline{\underline{X}}.\underline{\underline{A}}\right).\underline{\underline{X}} \end{aligned}$$

Les points fixes sont tels que :

$$\underline{\underline{F}}_A(\underline{\underline{X}}^\star) = \underline{\underline{X}}^\star = 2.\underline{\underline{X}}^\star - \underline{\underline{X}}^\star.\underline{\underline{A}}.\underline{\underline{X}}^\star \Leftrightarrow \underline{\underline{X}}^\star.\left(\underline{\underline{I}}_d - \underline{\underline{A}}.\underline{\underline{X}}^\star\right) = \underline{\underline{0}}$$

$\underline{\underline{F}}_A$ a donc au moins deux points fixes : $\underline{\underline{X}}^\star = \{\underline{\underline{0}}, \underline{\underline{A}}^{-1}\}$. Donnons l'expression du développement limité de $\underline{\underline{F}}_A$ autour de ces deux points :

$$\begin{aligned}\underline{\underline{F}}_A(\underline{\underline{A}}^{-1} + \underline{\underline{H}}) &= 2\underline{\underline{A}}^{-1} + 2\underline{\underline{H}} - \left(\underline{\underline{A}}^{-1} + \underline{\underline{H}}\right)\left(\underline{\underline{I}}_d - \underline{\underline{A}}.\underline{\underline{H}}\right) \\ &= 2\underline{\underline{A}}^{-1} + 2\underline{\underline{H}} - \underline{\underline{A}}^{-1} - 2\underline{\underline{H}} - \underline{\underline{H}}.\underline{\underline{A}}.\underline{\underline{H}} \\ &= \underline{\underline{F}}_A(\underline{\underline{A}}^{-1}) + 0 - \underline{\underline{H}}.\underline{\underline{A}}.\underline{\underline{H}}\end{aligned}$$

On en déduit que la différentielle de $\underline{\underline{F}}_A$ en $\underline{\underline{A}}^{-1}$ est nulle et donc que la méthode est convergente d'ordre 2 pour le point supre-attractif $\underline{\underline{A}}^{-1}$. Pour le point fixe nul :

$$\underline{\underline{F}}_A(\underline{\underline{0}} + \underline{\underline{H}}) = \underline{\underline{F}}_A(\underline{\underline{0}}) + 2\underline{\underline{H}} - \underline{\underline{H}}.\underline{\underline{A}}.\underline{\underline{H}}$$

On en déduit que la différentielle de $\underline{\underline{F}}_A$ en $\underline{\underline{0}}$ est l'application linéaire $\underline{\underline{H}} \mapsto 2\underline{\underline{H}}$ dont le rayon spectral est 2. Le théorème d'Ostrowski ne s'applique pas dans ce cas, mais on pourrait montrer que ce point fixe n'est pas attractif.

### 6.4.5 Indice d'efficacité d'une méthode itérative

L'ordre de convergence d'une méthode itérative fournit une estimation de la vitesse de convergence, mais cette information ne permet pas de quantifier le temps nécessaire pour qu'une méthode donne l'approximation du calcul d'une racine. En effet, en général plus la méthode a un ordre de convergence élevé et plus il est nécessaire d'évaluer un grand nombre de fonctions (voir exemple 6.11). Pour illustrer ce propos reprenons l'exemple de la section 6.4.1 :

**Exemple 6.10.** On a vu que pour obtenir une approximation de $\sqrt{2}$ avec 190 décimales exactes, la méthode de Héron nécessitait huit itérations contre six pour la méthode de Lambert. Maintenant, regardons le nombre de d'opérations nécessaires pour obtenir cette convergence :

**Héron :** A chaque itération, il y a trois opérations élémentaires (une addition et deux divisions). Pour obtenir 190 décimales exactes, huit itérations sont nécessaires donc il faut $8 \times 3 = 24$ opérations ;

**Lambert :** A chaque itération, il y a six opérations élémentaires et pour obtenir 190 décimales exactes, six itérations sont nécessaires donc il faut $6 \times 6 = 36$ opérations.

On voit que dans ce cas la conclusion est inversée, la méthode de Héron semble préférable par rapport à la méthode de Lambert.

Il est alors nécessaire d'introduire un indice d'efficacité permettant de prendre en compte ces deux paramètres :

**Définition 6.10.**

Soit une méthode itérative $F$ d'ordre $\tau \geq 1$ qui à chaque itération nécessite $\sigma \geq 1$ évaluations de fonctions. L'indice d'efficacité de cette méthode est alors défini par :

$$\mathcal{E}_F(\tau, \sigma) = \tau^{1/\sigma}$$

Le nombre $\sigma$ d'évaluation de fonctions prend en compte les évaluations de la même fonction en différents points (pour les méthodes multi-points) et les évaluations de fonctions différentes, y compris des dérivées d'une même fonction.

**Exemple 6.11.** Reprenons pour exemple les méthodes de Héron et de Lambert, méthodes permettant d'extraire les racines carrées d'un réel $a > 0$, i.e de résoudre l'équation :

$$f_a(x) = x^2 - a$$

Ces deux méthodes sont localement convergentes pour $\sqrt{a}$ qui est un point fixe super-attractif et leur bassin d'attraction est identique ($\mathbb{R}_+^\star$). :

— Méthode de Héron : cette méthode est convergente d'ordre 2. Afin d'estimer son indice d'efficacité, réécrivons la fonction $H_a$ :

$$H_a(x) = x - \frac{f_a(x)}{f_a'(x)}$$

Chaque itération nécessite donc l'évaluation de $f_a$ et de $f_a'$, soit $\sigma = 2$. Ainsi

$$\mathcal{E}_{H_a} = 2^{1/2}$$

— Méthode de Lambert : cette méthode est convergente d'ordre 3. On réécrit la fonction $L_a$ telle que :

$$L_a(x) = x - \frac{f_a(x)\, f_a'(x)}{g_a(x)} \text{ où } g_a(x) = a + 3x^2$$

Chaque itération nécessite donc l'évaluation de $f_a$, de $f_a'$ et de $g_a$, soit $\sigma = 3$. Ainsi

$$\mathcal{E}_{L_a} = 3^{1/3}$$

Au final, d'après ce critère il est préférable d'utiliser la méthode de Lambert car $\mathcal{E}_{L_a} > \mathcal{E}_{H_a}$.

**Exercice 6.8 (**Méthode de Steffensen**).** Soient $U$ un ouvert non vide de $\mathbb{R}$, $f : U \to \mathbb{R}$ une fonction de classe $\mathcal{C}^2(U)$ et $x^\star \in U$ une racine simple de $f$, i.e $f(x^\star) = 0$ et $f'(x^\star) \neq 0$. On définit la méthode de Steffensen par

$$S_f(x) = x - \frac{f(x)^2}{f(x + f(x)) - f(x)}$$

1. Montrer qu'il existe un voisinage de $x^\star$, sur lequel $S_f$ est bien définie et dérivable et en déduire que $x^\star$ est super-attractif.
2. Déterminer l'ordre de convergence de la méthode.
3. En déduire l'indice d'efficacité de cette méthode.

## 6.5 Méthodes à un pas sans mémoire en dimension 1

L'idée générale est de remplacer au voisinage du point $(x^{(k)}, f(x^{(k)}))$ le graphe de $f$ par une courbe osculatrice $y_n$, i.e une fonction qui va épouser le graphe de $f$ "de mieux en mieux" au point $x^{(k)}$ :
— intersection (figure 6.7) : $f(x^{(k)}) = y_k(x^{(k)})$
— tangence (figure 6.8) : $f'(x^{(k)}) = {y_k}'(x^{(k)})$
— concavité (figure 6.9) : $f''(x^{(k)}) = {y_k}''(x^{(k)})$
Le point $x^{(k+1)}$ est alors obtenu par l'intersection entre $y_k$ et l'axe des abscisses. Autrement dit $y_k(x^{(k+1)}) = 0$.

### 6.5.1 Approximation par une forme affine

Géométriquement la première idée est de remplacer au voisinage du point $(x^{(k)}, f(x^{(k)}))$ le graphe de $f$ par une droite d'équation :

$$y_k(x) = f(x^{(k)}) + p_k \left(x - x^{(k)}\right)$$

où $p_k$ représente la pente de la droite. Cette droite vérifie au moins une condition d'osculation : l'intersection $y_k(x^{(k)}) = f(x^{(k)})$.

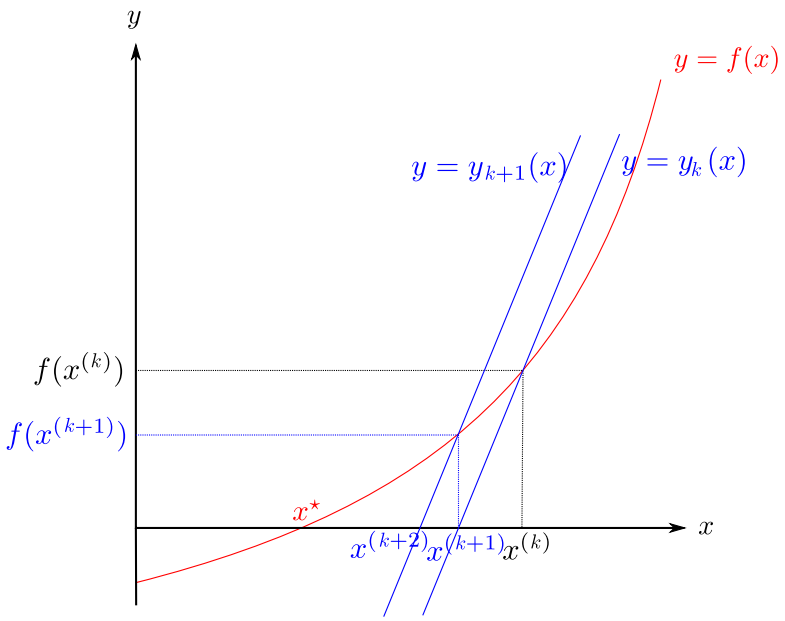



FIGURE 6.4 – Illustration de la méthode de la parallèle

### Méthode de la parallèle

Si toutes les droites sont définies avec la même pente $p_k = 1/\alpha$ où $\alpha \neq 0$, la relation de récurrence entre $x^{(k+1)}$ et $x^{(k)}$, illustrée sur la figure 6.7, est définie par :

$$y_k(x^{(k+1)}) = 0 = f(x^{(k)}) + \frac{1}{\alpha}\left(x^{(k+1)} - x^{(k)}\right) \Leftrightarrow x^{(k+1)} = x^{(k)} - \alpha f(x^{(k)})$$

**Proposition 6.3.** *Soient $U$ un ouvert non vide de $\mathbb{R}$, $f : U \to \mathbb{R}$ une fonction $\mathcal{C}^1(U)$, $x^\star \in U$ une racine de $f$ et $\alpha \in \mathbb{R}^\star$. On note $P_f$ la fonction d'itération de la méthode de la parallèle associée à la fonction $f$, définie par :*

$$P_f(x) = x - \alpha f(x)$$

*Cette méthode est convergente d'ordre au moins 1 si $|1 - \alpha f'(x^\star)| < 1$. Si $\alpha = 1/f'(x^\star)$, elle est au moins d'ordre deux.*

**Preuve.** *On applique directement le théorème d'Ostrowski. $x^\star \in U$ point fixe attractif de la méthode de la parallèle si*

$$|P_f'(x^\star)| < 1 \Leftrightarrow |1 - \alpha f'(x^\star)| < 1$$

*On voit au passage que si $x^\star$ est racine double, i.e $f'(x^\star) = 0$, alors la méthode ne converge pas. Si de plus $\alpha = 1/f'(x^\star)$, alors $P_f'(x^\star) = 0$ le point fixe est super-attractif et la méthode est d'ordre au moins deux.*

### Méthode de Newton-Raphson

Cette méthode consiste à imposer, en plus de l'intersection, une seconde condition d'osculation : la tangence de $y_k$ à $f$ en $x^{(k)}$, i.e :

$$\begin{cases} f(x^{(k)}) = y_k(x^{(k)}) \\ f'(x^{(k)}) = {y_k}'(x^{(k)}) = p_k \end{cases}$$

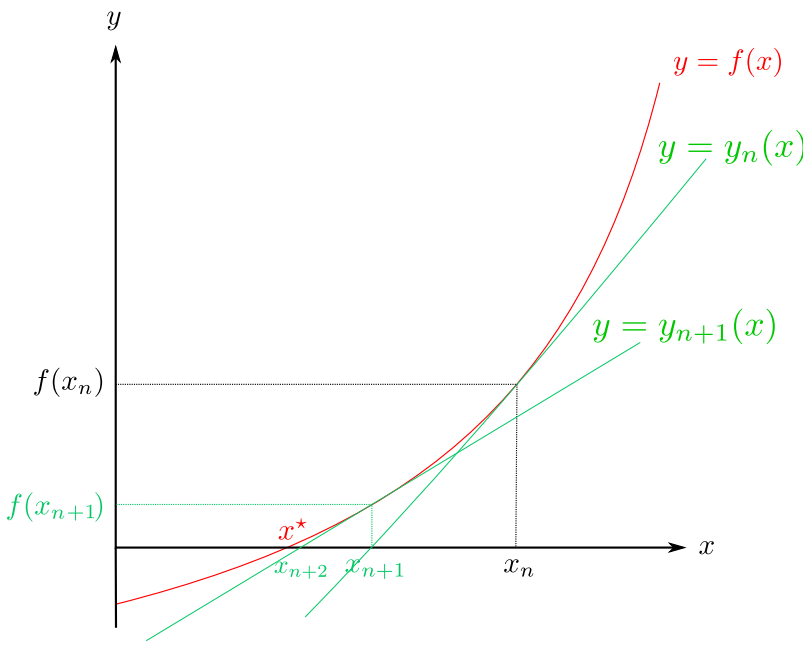



FIGURE 6.5 – Illustration de la méthode de Newton-Raphson

Ainsi la relation de récurrence, illustrée sur la figure 6.8, est définie par :

$$y_k(x^{(k+1)}) = 0 = f(x^{(k)}) + f'(x^{(k)})\left(x^{(k+1)} - x^{(k)}\right) \Leftrightarrow x^{(k+1)} = x^{(k)} - \frac{f(x^{(k)})}{f'(x^{(k)})}$$

**Proposition 6.4.** *Soient $U$ un ouvert non vide de $\mathbb{R}$, $f : U \to \mathbb{R}$ une fonction $\mathcal{C}^2(U)$, $x^\star \in U$ une racine simple de $f$, i.e $f(x^\star) = 0$ et $f'(x^\star) \neq 0$. On note $N_f$ la fonction d'itération de la méthode de la Newton-Raphson associée à la fonction $f$, définie par :*

$$N_f(x) = x - \frac{f(x)}{f'(x)}$$

*Il existe alors un ouvert $V \subset U$ contenant $x^\star$ tel que la fonction $f'$ ne s'annule pas. Si $f''(x^\star) \neq 0$, cette méthode est localement convergente d'ordre 2, et son indice d'efficacité est $\mathcal{E}_{N_f} = \sqrt{2}$.*

**Preuve.** *$f$ étant $\mathcal{C}^2(U)$, on peut écrire les développements limités de $f$ et de $f'$ autour de $x^\star$, tels que*

$$\begin{cases} f(x^\star + h) = f(x^\star) + hf'(x^\star) + \dfrac{h^2}{2}f''(x^\star) + o(h^2) = hf'(x^\star)\left[1 + \dfrac{h}{2}\dfrac{f''(x^\star)}{f'(x^\star)} + o(h)\right] \\ f'(x^\star + h) = f'(x^\star) + hf''(x^\star) + o(h) = f'(x^\star)\left[1 + h\dfrac{f''(x^\star)}{f'(x^\star)} + o(h)\right] \end{cases}$$

*Ainsi le développement limité de l'inverse de $f'(x^\star + h)$ vaut :*

$$\frac{1}{f'(x^\star + h)} = \frac{1}{f'(x^\star)}\left[1 - h\frac{f''(x^\star)}{f'(x^\star)} + o(h)\right]$$

*donc son produit par $f(x^\star + h)$*

$$\frac{f(x^\star + h)}{f'(x^\star + h)} = h\left[1 - h\frac{f''(x^\star)}{f'(x^\star)} + o(h)\right] \times \left[1 + \frac{h}{2}\frac{f''(x^\star)}{f'(x^\star)} + o(h)\right] = h\left[1 - \frac{h}{2}\frac{f''(x^\star)}{f'(x^\star)} + o(h)\right]$$

*Au final, le développement limité de $N_f$ a pour expression :*

$$N_f(x^\star + h) = x^\star + \frac{h^2}{2}\frac{f''(x^\star)}{f'(x^\star)} + o(h^2)$$

*On aurait également pu penser déterminer l'ordre de convergence en calculant les dérivées successives de $N_f$ en $x^\star$ :*

$$N_f'(x) = \frac{f(x)\,f''(x)}{f'(x)^2} \Rightarrow N_f'(x^\star) = 0$$

*Ceci montre bien que $x^\star$ point fixe super-attractif de $N_f$. Cependant si $f$ est uniquement $\mathcal{C}^2(U)$, on ne peut pas calculer $N_f''$ qui ferait intervenir $f^{(3)}$. Cependant, il existe bien un terme d'ordre deux en $h$ dans le développement limité de $N_f$. Ainsi on voit que la méthode du développement limité est plus puissante que celle des dérivées successives car elle nécessite moins de contraintes sur la régularité de $f$.*

*A chaque itération, il faut évaluer $f(x)$ et $f'(x)$ donc $\sigma = 2$ et la méthode est d'ordre 2 donc $\mathcal{E}_{N_f} = \sqrt{2}$.*

**Proposition 6.5.** *Si $x^\star \in U$ racine multiple d'ordre $r > 1$ (i.e $f(x^\star) = f'(x^\star) = \ldots = f^{(r-1)}(x^\star) = 0$ et $f^{(r)}(x^\star) \neq 0$), alors la suite itérative formée sur la méthode de Newton-Raphson converge linéairement, et plus $r$ est grand et plus la convergence est lente.*

**Preuve.** *On peut factoriser $f$, de telle sorte que :*

$$f(x) = (x - x^\star)^r\ g(x)\ \text{où}\ g \in \mathcal{C}^2(U)\ \text{et}\ g(x^\star) \neq 0$$

*Ainsi les deux premières dérivées de $f$ sont :*

$$\begin{cases} f'(x) = r\,(x - x^\star)^{r-1}\ g(x) + (x - x^\star)^r\ g'(x) \\ f''(x) = r\,(r-1)\,(x - x^\star)^{r-2}\ g(x) + 2r\,(x - x^\star)^{r-1}\ g'(x) + (x - x^\star)^r\ g''(x) \end{cases}$$

*La dérivée de $N_f$ vaut donc :*

$$\begin{aligned} N_f'(x) = \frac{f(x)\,f''(x)}{f'(x)^2} &= \frac{(x - x^\star)^r\ g(x)\ (x - x^\star)^{r-2}\left[r\,(r-1)\ g(x) + 2r\,(x - x^\star)\ g'(x) + (x - x^\star)^2\ g''(x)\right]}{(x - x^\star)^{2r-2}\left[r\,g(x) + (x - x^\star)\ g'(x)\right]^2} \\ &= \frac{g(x)\ \left[r\,(r-1)\ g(x) + 2r\,(x - x^\star)\ g'(x) + (x - x^\star)^2\ g''(x)\right]}{\left[r\,g(x) + (x - x^\star)\ g'(x)\right]^2} \end{aligned}$$

*Ainsi $N_f'(x^\star) = 1 - \frac{1}{r}$.*

### 6.5.2 Approximation par une forme quadratique

#### Méthode de Halley

Soit l'hyperbole d'équation

$$y_k(x) + a_k x y_k(x) + b_k x + c_k = 0$$

Cette hyperbole et la fonction $f$ vérifient trois conditions d'osculation en $x^{(k)}$ : l'intersection, la tangence et la concavité :

$$\begin{cases} f(x^{(k)}) + a_k x^{(k)} f(x^{(k)}) + b_k x^{(k)} + c_k = 0 \\ f'(x^{(k)}) + a_k x^{(k)} f'(x^{(k)}) + a_k f(x^{(k)}) + b_k = 0 \\ f''(x^{(k)}) + a_k x^{(k)} f''(x^{(k)}) + 2a_k f'(x^{(k)}) = 0 \end{cases} \Leftrightarrow \begin{bmatrix} x^{(k)} f(x^{(k)}) & x^{(k)} & 1 \\ x^{(k)} f'(x^{(k)}) + f(x^{(k)}) & 1 & 0 \\ x^{(k)} f''(x^{(k)}) + 2f'(x^{(k)}) & 0 & 0 \end{bmatrix} \cdot \begin{pmatrix} a_k \\ b_k \\ c_k \end{pmatrix} = - \begin{pmatrix} f(x^{(k)}) \\ f'(x^{(k)}) \\ f''(x^{(k)}) \end{pmatrix}$$

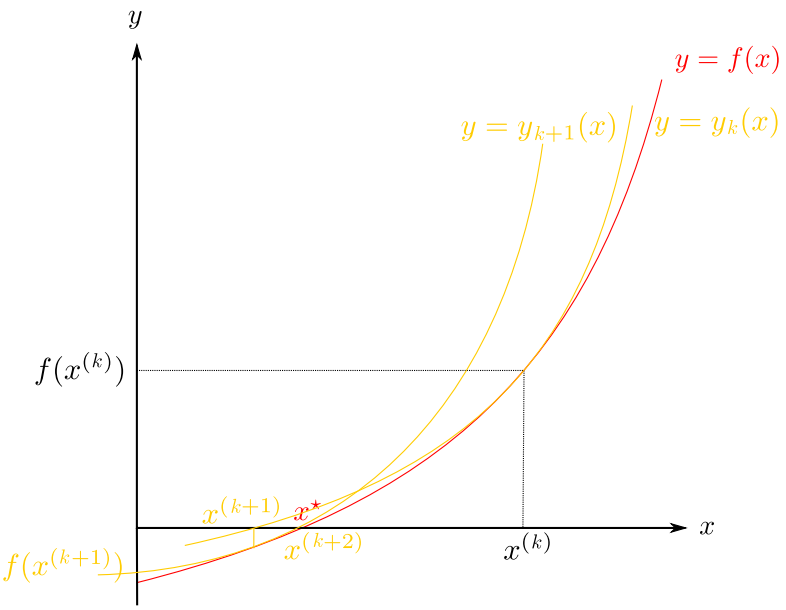



FIGURE 6.6 – Illustration du principe d'une méthode quadratique (Halley, Chebychev)

La résolution de ce système linéaire triangulaire fournit les expressions de $a_k, b_k, c_k$ :

$$\begin{cases} a_k = -\dfrac{f''(x^{(k)})}{x^{(k)} f''(x^{(k)}) + 2f'(x^{(k)})} \\ b_k = \dfrac{f(x^{(k)})f''(x^{(k)}) - 2f'(x^{(k)})^2}{x^{(k)} f''(x^{(k)}) + 2f'(x^{(k)})} \\ c_k = \dfrac{x^{(k)}\left[2f'(x^{(k)})^2 - f(x^{(k)})f''(x^{(k)})\right] - 2f(x^{(k)})f'(x^{(k)})}{x^{(k)} f''(x^{(k)}) + 2f'(x^{(k)})} \end{cases}$$

On obtient ainsi l'expression de l'hyperbole osculatrice :

$$y_k(x) - f(x^{(k)}) - \left(x - x^{(k)}\right) f'(x^{(k)}) - \frac{1}{2}\left(x - x^{(k)}\right)\left(y_k(x) - f(x^{(k)})\right)\frac{f''(x^{(k)})}{f'(x^{(k)})} = 0$$

Le point $x^{(k+1)}$ est définit l'intersection entre $y_k$ et l'axe des abscisses : $y_k(x^{(k+1)}) = 0$ d'où la relation de récurrence :

$$x^{(k+1)} = x^{(k)} - \frac{2f(x^{(k)})f'(x^{(k)})}{2f'(x^{(k)})^2 - f(x^{(k)})f''(x^{(k)})}$$

**Proposition 6.6.** *Soient $U$ un ouvert non vide de $\mathbb{R}$, $f : U \to \mathbb{R}$ une fonction $\mathcal{C}^3(U)$, $x^\star \in U$ une racine simple de $f$. On note $H_f$ la fonction d'itération de la méthode de la Halley associée à la fonction $f$, définie par :*

$$H_f(x) = x - \frac{2f(x)f'(x)}{2f'(x)^2 - f(x)f''(x)}$$

*Par conséquent*

- *il existe alors un ouvert $V \subset U$ contenant $x^\star$ tel que la fonction $2f'^2 - ff''$ est continue et ne s'annule pas.*
- *le point $x^\star$ est super-attractif*

— *si de plus* $\frac{3}{2}\left(\frac{f''(x^\star)}{f'(x^\star)}\right)^2 - \frac{f'''(x^\star)}{f'(x^\star)} \neq 0$, *cette méthode est localement convergente d'ordre 3, et son indice d'efficacité est* $\mathcal{E}_{H_f} = \sqrt[3]{3}$.

**Preuve.** *La fonction* $f$ *étant* $\mathcal{C}^3(U)$, *on peut faire les développements limités suivants :*

$$\begin{cases} f(x^\star+h) = hf'(x^\star) + \dfrac{h^2}{2}f''(x^\star) + \dfrac{h^3}{6}f'''(x^\star) + o(h^3) = hf'(x^\star)\left[1 + \dfrac{h}{2}\dfrac{f''(x^\star)}{f'(x^\star)} + \dfrac{h^2}{6}\dfrac{f'''(x^\star)}{f'(x^\star)} + o(h^3)\right] \\ f'(x^\star+h) = f'(x^\star) + hf''(x^\star) + \dfrac{h^2}{2}f'''(x^\star) + o(h^3) = f'(x^\star)\left[1 + h\dfrac{f''(x^\star)}{f'(x^\star)} + \dfrac{h^2}{2}\dfrac{f'''(x^\star)}{f'(x^\star)} + o(h^2)\right] \\ f''(x^\star+h) = f''(x^\star) + hf'''(x^\star) + o(h) = f'(x^\star)\left[\dfrac{f''(x^\star)}{f'(x^\star)} + h\dfrac{f'''(x^\star)}{f'(x^\star)} + o(h)\right] \end{cases}$$

*Afin d'alléger les expressions, utilisons les notations suivantes :*

$$c_2 = \frac{1}{2}\frac{f''(x^\star)}{f'(x^\star)} \qquad ; \qquad c_3 = \frac{1}{6}\frac{f'''(x^\star)}{f'(x^\star)}$$

*de telle sorte que le système précédent s'écrit :*

$$\begin{cases} f(x^\star+h) = hf'(x^\star)\left[1 + hc_2 + h^2c_3 + o(h^2)\right] \\ f'(x^\star+h) = f'(x^\star)\left[1 + 2hc_2 + 3h^2c_3 + o(h^2)\right] \\ f''(x^\star+h) = f'(x^\star)\left[2c_2 + 6hc_3 + o(h)\right] \end{cases}$$

*On peut alors calculer les développements limités suivants :*

$$\begin{aligned} f(x^\star+h)f'(x^\star+h) &= hf'(x^\star)^2\left[1 + 3hc_2 + h^2\left(4c_3 + 2c_2^2\right) + o(h^2)\right] \\ f(x^\star+h)f''(x^\star+h) &= hf'(x^\star)^2\left[2c_2 + h\left(6c_3 + 2c_2^2\right) + o(h)\right] \\ f'(x^\star+h)^2 &= f'(x^\star)^2\left[1 + 4hc_2 + h^2\left(6c_3 + 4c_2^2\right) + o(h^2)\right] \end{aligned}$$

*Dans un premier temps :*

$$2f'(x^\star+h)^2 - f(x^\star+h)f''(x^\star+h) = 2f'(x^\star)^2\left[1 + 3hc_2 + 3h^2\left(c_3 + c_2^2\right) + o(h^2)\right]$$

*soit en inversant ce développement limité grâce à la formule* $\frac{1}{1+u} = 1 - u + u^2 + o(u^2)$ *:*

$$\begin{aligned} \frac{1}{2f'(x^\star+h)^2 - f(x^\star+h)f''(x^\star+h)} &= \frac{1}{2f'(x^\star)^2}\left[1 - \left\{3hc_2 + 3h^2\left(c_3 + c_2^2\right)\right\} + \left\{3hc_2\right\}^2 + o(h^2)\right] \\ &= \frac{1}{2f'(x^\star)^2}\left[1 - 3hc_2 + 3h^2\left(-c_3 + 2c_2^2\right) + o(h^2)\right] \end{aligned}$$

*Par conséquent :*

$$\frac{2f(x^\star+h)f'(x^\star+h)}{2f'(x^\star+h)^2 - f(x^\star+h)f''(x^\star+h)} = h\left[1 + 0 \times h + h^2\left(c_3 - c_2^2\right) + o(h^2)\right]$$

*Au final, le développement limité de la fonction d'itération de la méthode de Halley vaut :*

$$H_f(x^\star+h) = x^\star + \frac{h^3}{6}\left(6c_2^2 - 6c_3\right) + o(h^3) = x^\star + \frac{h^3}{6}\left(\frac{3}{2}\left(\frac{f''(x^\star)}{f'(x^\star)}\right)^2 - \frac{f'''(x^\star)}{f'(x^\star)}\right) + o(h^3)$$

*Les conclusions sont donc les suivantes :*

— *le terme en h étant nul et* $f \in \mathcal{C}^3(U)$*, le terme* $H'_f(x^\star) = 0$ *le point* $x^\star$ *est donc super-attractif ;*
— *Si* $\frac{3}{2}\left(\frac{f''(x^\star)}{f'(x^\star)}\right)^2 - \frac{f'''(x^\star)}{f'(x^\star)} \neq 0$*, alors la méthode est d'ordre 3 ; De plus à chaque itération il faut évaluer* $f(x^{(k)}), f'(x^{(k)}), f''(x^{(k)})$ *donc* $\sigma_{H_f} = 3$ *et* $\mathcal{E}_{H_f} = \sqrt[3]{3}$*.*
— $f \in \mathcal{C}^3(U)$*, on ne peut pas associer :*
  — $H''_f(x^\star)$ *à 0 (sauf si* $f \in \mathcal{C}^4(U)$*)*
  — $H'''_f(x^\star)$ *à* $\frac{3}{2}\left(\frac{f''(x^\star)}{f'(x^\star)}\right)^2 - \frac{f'''(x^\star)}{f'(x^\star)}$ *(sauf si* $f \in \mathcal{C}^5(U)$*)*

**Exercice 6.9.** Déterminer l'ensemble des fonctions vérifiant l'EDO $(E)$ $\frac{3}{2}\left(\frac{f''(x)}{f'(x)}\right)^2 - \frac{f'''(x)}{f'(x)} = 0$ et montrer que chacune des ces solutions possède une unique racine. En déduire que si $f$ appartient à cet ensemble, la méthode de Halley converge globalement exactement en une seule itération.

### Méthode de Chebychev

Soit la parabole d'équation

$$y_k(x) + a_k {y_k}^2(x) + b_k x + c_k = 0$$

Cette parabole et la fonction $f$ vérifient trois conditions d'osculation en $x^{(k)}$ : l'intersection, la tangence et la concavité :

$$\begin{cases} f(x^{(k)}) + a_k f(x^{(k)})^2 + b_k x^{(k)} + c_k = 0 \\ f'(x^{(k)}) + 2a_k f(x^{(k)}) f'(x^{(k)}) + b_k = 0 \\ f''(x^{(k)}) + 2a_k \left\{ f'(x^{(k)})^2 + f(x^{(k)}) f''(x^{(k)}) \right\} = 0 \end{cases}$$

$$\Leftrightarrow \begin{bmatrix} f(x^{(k)})^2 & x^{(k)} & 1 \\ 2f(x^{(k)}) f'(x^{(k)}) & 1 & 0 \\ 2\left\{ f'(x^{(k)})^2 + f(x^{(k)}) f''(x^{(k)}) \right\} & 0 & 0 \end{bmatrix} . \begin{pmatrix} a_k \\ b_k \\ c_k \end{pmatrix} = - \begin{pmatrix} f(x^{(k)}) \\ f'(x^{(k)}) \\ f''(x^{(k)}) \end{pmatrix}$$

La résolution de ce système linéaire triangulaire fournit les expressions de $a_k, b_k, c_k$ :

$$\begin{cases} a_k = -\dfrac{f''(x^{(k)})}{2\left\{ f'(x^{(k)})^2 + f(x^{(k)}) f''(x^{(k)}) \right\}} \\ b_k = -\dfrac{f'(x^{(k)})^3}{f'(x^{(k)})^2 + f(x^{(k)}) f''(x^{(k)})} \\ c_k = \dfrac{-2f(x^{(k)}) f'(x^{(k)})^2 - f(x^{(k)})^2 f''(x^{(k)}) + 2f'(x^{(k)})^3 x^{(k)}}{2\left\{ f'(x^{(k)})^2 + f(x^{(k)}) f''(x^{(k)}) \right\}} \end{cases}$$

On obtient ainsi l'expression de la parabole osculatrice :

$$y_k(x) - f(x^{(k)}) - \left(x - x^{(k)}\right) f'(x^{(k)}) - \frac{1}{2}\left(y_k(x) - f(x^{(k)})\right)^2 \frac{f''(x^{(k)})}{f'(x^{(k)})} = 0$$

Le point $x^{(k+1)}$ est définit l'intersection entre $y_k$ et l'axe des abscisses : $y_k(x^{(k+1)}) = 0$ d'où la relation de récurrence :

$$x^{(k+1)} = x^{(k)} - \frac{2f(x^{(k)}) f'(x^{(k)})^2 + f(x^{(k)})^2 f''(x^{(k)})}{2f'(x^{(k)})^3}$$

**Proposition 6.7.** *Soient* $U$ *un ouvert non vide de* $\mathbb{R}$*,* $f : U \to \mathbb{R}$ *une fonction* $\mathcal{C}^3(U)$*,* $x^\star \in U$ *une racine simple de* $f$*. On note* $C_f$ *la fonction d'itération de la méthode de la Chebychev associée à la fonction* $f$*, définie par :*

$$C_f(x) = x - \frac{2f(x) f'(x)^2 + f(x)^2 f''(x)}{2f'(x)^3}$$

*Par conséquent*
- *il existe alors un ouvert $V \subset U$ contenant $x^\star$ tel que la fonction $f'$ ne s'annule pas.*
- *le point $x^\star$ est super-attractif*
- *si de plus $3\left(\frac{f''(x^\star)}{f'(x^\star)}\right)^2 - \frac{f'''(x^\star)}{f'(x^\star)} \neq 0$, cette méthode est localement convergente d'ordre 3, et son indice d'efficacité est $\mathcal{E}_{C_f} = \sqrt[3]{3}$.*

**Preuve.** *Comme pour la méthode de Halley, posons les coefficients $c_2$ et $c_3$ suivants afin d'alléger les notations :*

$$c_2 = \frac{1}{2}\frac{f''(x^\star)}{f'(x^\star)} \qquad ; \qquad c_3 = \frac{1}{6}\frac{f'''(x^\star)}{f'(x^\star)}$$

*La fonction $f$ étant $\mathcal{C}^3(U)$, on peut donc écrire les développements limités suivants :*

$$\begin{cases} f(x^\star+h) = hf'(x^\star)\left[1 + hc_2 + h^2c_3 + o(h^2)\right] \\ f'(x^\star+h) = f'(x^\star)\left[1 + 2hc_2 + 3h^2c_3 + o(h^2)\right] \\ f''(x^\star+h) = f'(x^\star)\left[2c_2 + 6hc_3 + o(h)\right] \end{cases}$$

*On en déduit tout d'abord,*

$$f(x^\star+h)^2 = h^2f'(x^\star)^2\left[1 + 2c_2h + o(h)\right]$$
$$f(x^\star+h)^2f''(x^\star+h) = h^2f'(x^\star)^3\left[2c_2 + h\left(4c_2^2 + 6c_3\right) + o(h)\right]$$

*On voit ici qu'on aurait pu pousser le développement limité de $f(x^\star+h)^2$ un ordre plus loin, mais dans le but de le multiplier avec $f''(x^\star+h)$ cela n'aurait servi à rien car le développement limité de $f''(x^\star+h)$ est $o(h)$.*
*Ensuite, on peut calculer le développement limité restant au numérateur :*

$$f'(x^\star+h)^2 = f'(x^\star)^2\left[1 + 4c_2h + h^2\left(6c_3 + 4c_2^2\right) + o(h^2)\right]$$
$$f(x^\star+h)f'(x^\star+h)^2 = hf'(x^\star)^3\left[1 + 5c_2h + h^2\left(8c_2^2 + 7c_3\right) + o(h^2)\right]$$

*Ce qui pemet d'exprimer le développement limité du numérateur :*

$$2f(x^\star+h)f'(x^\star+h)^2 + f(x^\star+h)^2f''(x^\star+h) = 2hf'(x^\star)^3\left[1 + 6hc_2 + h^2\left(10c_3 + 10c_2^2\right) + o(h^2)\right]$$

*Maintenant calculons le développement limité du dénominateur :*

$$f'(x^\star+h)^3 = f'(x^\star)^3\left[1 + 6c_2h + h^2\left(9c_3 + 12c_2^2\right) + o(h^2)\right]$$
$$\frac{1}{f'(x^\star+h)^3} = \frac{1}{f'(x^\star)^3}\left[1 - 6c_2h + h^2\left(-9c_3 + 24c_2^2\right) + o(h^2)\right]$$

*Ainsi,*

$$\frac{2f(x^\star+h)f'(x^\star+h)^2 + f(x^\star+h)^2f''(x^\star+h)}{2f'(x^\star+h)^3} = h\left[1 + h^2\left(c_3 - 2c_2^2\right) + o(h^2)\right]$$

*Au final, le développement limité de la fonction de Chebychev est :*

$$C_f(x^\star+h) = x^\star + \frac{h^3}{6}\left(12c_2^2 - 6c_3\right) + o(h^3) = x^\star + \frac{h^3}{6}\left(3\left(\frac{f''(x^\star)}{f'(x^\star)}\right)^2 - \frac{f'''(x^\star)}{f'(x^\star)}\right) + o(h^3)$$

*Les conclusions sont donc les suivantes :*
- *le terme en $h$ étant nul et $f \in \mathcal{C}^3(U)$, le terme $C_f'(x^\star) = 0$ le point $x^\star$ est donc super-attractif;*

— *Si* $3\left(\frac{f''(x^\star)}{f'(x^\star)}\right)^2 - \frac{f'''(x^\star)}{f'(x^\star)} \neq 0$*, alors la méthode est d'ordre 3 ; De plus à chaque itération il faut évaluer* $f(x^{(k)}), f'(x^{(k)}), f''(x^{(k)})$ *donc* $\sigma_{C_f} = 3$ *et* $\mathcal{E}_{C_f} = \sqrt[3]{3}$.
— $f \in \mathcal{C}^3(U)$*, on ne peut pas associer :*
  — $C''_f(x^\star)$ *à 0 (sauf si* $f \in \mathcal{C}^4(U)$*)*
  — $C'''_f(x^\star)$ *à* $3\frac{f''(x^\star)}{f'(x^\star)} - \frac{f'''(x^\star)}{f'(x^\star)}$ *(sauf si* $f \in \mathcal{C}^5(U)$*)*

**Exercice 6.10.** Déterminer l'ensemble des fonctions vérifiant l'EDO $(E) \quad 3\left(\frac{f''(x)}{f'(x)}\right)^2 - \frac{f'''(x)}{f'(x)} = 0$ et montrer que chacune des ces solutions possède une unique racine. En déduire que si $f$ appartient à cet ensemble, la méthode de Chebychev converge globalement exactement en une seule itération.

**Exercice 6.11.** On note $C_{NR}$, $C_H$ et $C_C$ les coefficients suivants :

$$C_{NR} = f''(x^\star) \quad ; \quad C_H = \frac{3}{2}\left(\frac{f''(x^\star)}{f'(x^\star)}\right)^2 - \frac{f'''(x^\star)}{f'(x^\star)} \quad ; \quad C_C = 3\left(\frac{f''(x^\star)}{f'(x^\star)}\right)^2 - \frac{f'''(x^\star)}{f'(x^\star)}$$

Comparer les indices d'efficacité des méthodes de Newton-Raphson, de Halley et de Chebychev dans les conditions suivantes :

1. $\{C_{NR} \neq 0\} \cap (\{C_H \neq 0\} \cup \{C_C \neq 0\})$
2. $\{C_{NR} = 0\} \cap (\{C_H \neq 0\} \cup \{C_C \neq 0\})$

## 6.6 Méthodes à un pas sans mémoire en dimension $n$

Soit le problème $(\mathcal{P}_n)$ "Trouver $\underline{x}^\star \in \mathbb{R}^n$ tel que $\underline{f}(\underline{x}^\star) = \underline{0}$ où $\underline{f} : \mathbb{R}^n \to \mathbb{R}^n$.

**Remarque 6.10.**
Un cas particulier est la résolution de systèmes linéaires

**Remarque 6.11.**
Il existe deux types de méthodes afin de résoudre $(\mathcal{P}_n)$ :
— Ramener la solution de $\underline{f}(\underline{x}^\star) = \underline{0}$ à la résolution (exacte ou approchée) d'une suite de systèmes linéaires. ex : méthode de Newton
— Ramener la solution de $\underline{f}(\underline{x}^\star) = \underline{0}$ à la résolution (plus ou moins approchée) d'une suite d'équations (linéaires ou non) à une seule inconnue. ex : Méthodes de type Gauss-Seidel (et s'appliquant aux problèmes linéaires)

### 6.6.1 Méthodes de la parallèle et de Newton

#### Méthode de la parallèle

Une généralisation naturelle du cas 1D est de considérer la fonction d'itération suivante :

$$\underline{F}(\underline{x}) = \underline{x} - \underline{\underline{H}}.\underline{f}(\underline{x})$$

où $\underline{\underline{H}}$ matrice fixe qui joue le rôle de $\alpha \in \mathbb{R}^\star$ en 1D. Pour que $\underline{x}^\star \in \mathbb{R}^n$ point fixe de $\underline{F}$ soit aussi racine de $\underline{f}$, i.e $\underline{F}(\underline{x}^\star) = \underline{x}^\star \Leftrightarrow \underline{\underline{H}}.\underline{f}(\underline{x}^\star) = \underline{0} \Leftrightarrow \underline{f}(\underline{x}^\star) = \underline{0}$, alors il faut que $\underline{\underline{H}}$ soit inversible.

La convergence locale de $\underline{x}^\star$ est assurée si $\rho\left(\underline{\underline{I}}_n - \underline{\underline{H}}.\underline{\underline{\mathcal{J}}}_{\underline{f}}(\underline{x}^\star)\right) < 1$ où $\underline{\underline{\mathcal{J}}}_{\underline{f}}(\underline{x}^\star)$ est la matrice jacobienne de $\underline{f}$ évaluée en $\underline{x}^\star$ et $\underline{\underline{I}}_n$ la matrice identité. La convergence est alors linéaire.

Si de plus on a $\underline{\underline{H}} = \alpha \underline{\underline{I}}_n$ on a alors $\underline{F}(\underline{x}) = \underline{x} - \alpha.\underline{f}(\underline{x})$.

**Méthode de Newton**

Comme dans le cas unidimensionnel, on cherche la droite osculatrice

$$\underline{y}_k(\underline{x}) = \underline{f}(\underline{x}^{(k)}) + \underline{\underline{\mathcal{J}}}_{\underline{f}}(\underline{x}^{(k)})\left(\underline{x} - \underline{x}^{(k)}\right)$$

telle que $\underline{y}_k(x^{(k+1)}) = \underline{0}$, i.e en supposant que $\underline{\underline{\mathcal{J}}}_{\underline{f}}(\underline{x}^{(k)})$ est inversible on a la suite itérative suivante :

$$\underline{x}^{(k+1)} = \underline{F}(\underline{x}^{(k)}) = \underline{x}^{(k)} - \underline{\underline{\mathcal{J}}}_{\underline{f}}^{-1}.\underline{f}(\underline{x}^{(k)})$$

En pratique, il faut résoudre une suite de systèmes linéaires de matrice $\underline{\underline{\mathcal{J}}}_{\underline{f}}(\underline{x}^{(k)})$, d'inconnue $\underline{x}^{(k+1)} - \underline{x}^{(k)}$ et de second membre $-\underline{f}(\underline{x}^{(k)})$ :

$$\underline{\underline{\mathcal{J}}}_{\underline{f}}(\underline{x}^{(k)}).\left(\underline{x}^{(k+1)} - \underline{x}^{(k)}\right) = -\underline{f}(\underline{x}^{(k)})$$

**Remarque 6.12.**
Si $n \gg 1$, chaque pas est coûteux car même si $\underline{\underline{\mathcal{J}}}_{\underline{f}}(\underline{x}^{(k)})$ est creuse il faut calculer les dérivées partielles puis résoudre le système.

**Proposition 6.8.** *$\underline{\underline{\mathcal{J}}}_{\underline{F}}(\underline{x}^\star) = \underline{0}$, il y a convergence locale et on peut montrer que la convergence est quadratique, i.e que*

$$\lim_{k\to\infty} \frac{\|\underline{x}^{(k+1)} - \underline{x}^\star\|}{\|\underline{x}^{(k)} - \underline{x}^\star\|^2} < \infty$$

*où $\|.\|$ représente une norme sur $\mathbb{R}^n$.*

**Preuve.** *Montrons que $\underline{\underline{\mathcal{J}}}_{\underline{F}}(\underline{x}^\star) = \underline{0}$. Pour simplifier les expressions, on notera $\{\alpha_{ij}(\underline{x})\}_{1\le i,j\le n}$ les coefficients de $\underline{\underline{\mathcal{J}}}_{\underline{f}}(\underline{x})^{-1}$. Par conséquent,*

$$\underline{F}(\underline{x}) = \underline{x} - \underline{\underline{\mathcal{J}}}_{\underline{f}}(\underline{x})^{-1}.\underline{f}(\underline{x}) = \begin{cases} F_1(\underline{x}) = x_1 - \sum_{j=1}^n \alpha_{1j}(\underline{x}) f_j(\underline{x}) \\ \quad \ldots = \ldots \\ F_i(\underline{x}^\star) = x_i - \sum_{j=1}^n \alpha_{ij}(\underline{x}) f_j(\underline{x}) \\ \quad \ldots = \ldots \\ F_n(\underline{x}^\star) = x_n - \sum_{j=1}^n \alpha_{mj}(\underline{x}) f_j(\underline{x}) \end{cases}$$

*Les coefficients de la matrice jacobienne associée à $\underline{F}$ en $\underline{x}^\star$ valent donc :*

$$\left(\underline{\underline{\mathcal{J}}}_{\underline{F}}(\underline{x}^\star)\right)_{ij} = \delta_{ij} - \sum_{j=1}^n \left[\partial_i \alpha_{ij}(\underline{x}^\star) f_j(\underline{x}^\star) + \alpha_{ij}(\underline{x}^\star)\partial_i f_j(\underline{x}^\star)\right]$$

*or pour tout $j \in [\![1;n]\!]$, $\underline{x}^\star$ est la racine de $f_j$ donc $f_j(\underline{x}^\star) = 0$ et $\sum_{j=1}^n \alpha_{ij}(\underline{x}^\star)\partial_i f_j(\underline{x}^\star) = \left(\underline{\underline{\mathcal{J}}}_{\underline{f}}^{-1}.\underline{\underline{\mathcal{J}}}_{\underline{f}}\right)_{ij} = \delta_{ij}$*

*Au final, $\left(\underline{\underline{\mathcal{J}}}_{\underline{F}}(\underline{x}^\star)\right)_{ij} = 0$ i.e $\underline{\underline{\mathcal{J}}}_{\underline{F}}(\underline{x}^\star) = \underline{0}$. On admettra le développement limité de $\underline{F}$, mais en ayant montré que le terme d'ordre $\underline{h} = 0$, même sans exprimer le terme d'ordre 2 on sait que la méthode est au moins d'ordre deux et que le point $\underline{x}^\star$ est super-attractif.*

### 6.6.2 Méthodes fondées sur le principe de Gauss-Seidel

#### Le principe de Gauss-Seidel

Soit $\underline{x}^{(0)} \in \mathbb{R}^n$ donné. Pour résoudre le système $\underline{f}(\underline{x}) = \underline{0}$, le principe de Gauss-Seidel consiste à construire $\underline{x}^{(k+1)}$ à partir de $\underline{x}^{(k)}$ en calculant $x_i^{(k+1)}$ (la $i^{\text{ème}}$ composante du vecteur $\underline{x}^{(k+1)}$) dans l'ordre croissant $(i = 1, 2, \ldots, n)$ comme solution d'une équation scalaire :

$$x_i^{(k+1)} = \xi \text{ solution de } f_i\left(x_1^{(k+1)}, x_2^{(k+1)}, \ldots, x_{i-1}^{(k+1)}, \xi, x_{i+1}^{(k)}, \ldots, x_n^{(k)}\right) = 0$$

Ce qui correspond à la résolution d'un système (a piori non-linéaire) de structure triangulaire inférieur :

$$\begin{cases} f_1(x_1^{(k+1)}, x_2^{(k)}, \ldots, x_n^{(k)}) = 0 \\ f_2(x_1^{(k+1)}, x_2^{(k+1)}, x_3^{(k)}, \ldots, x_n^{(k)}) = 0 \\ \ldots \\ f_i(x_1^{(k+1)}, \ldots, x_i^{(k+1)}, x_{i+1}^{(k)}, \ldots, x_n^{(k)}) = 0 \\ \ldots \\ f_n(x_1^{(k+1)}, \ldots, x_n^{(k+1)}) = 0 \end{cases}$$

**Remarque 6.13.**
Cet algorithme est très théorique et ne sert que de principe de base. En effet, si le système $\underline{f}(\underline{x}) = \underline{0}$ admet une solution, il se peut que certaines équations $f_i(x_1^{(k+1)}, \ldots, x_i^{(k+1)}, x_{i+1}^{(k)}, \ldots, x_n^{(k)}) = 0$ n'aient pas de solution et donc que $\underline{x}^{(k+1)}$ n'existe pas.

**Remarque 6.14.**
De plus, il est non rentable de résoudre exactement ou avec une grande précision chacune de ces équations, car $\underline{x}^{(k+1)}$ n'est qu'un résultat intermédiare.

**Remarque 6.15.**
Dans le cas linéaire, cet algorithme est cependant directement applicable et donne la méthode de Gauss-Seidel (voir chapitre 4).

#### Méthode de Gauss-Seidel-Newton

Cette méthode consiste à calculer $x_i^{(k+1)}$ avec un pas de Newton-Raphson en partant de $x_i^{(k)}$ : soit $\Phi : \xi \mapsto f_i(x_1^{(k+1)}, \ldots, x_{i-1}^{(k+1)}, \xi, x_{i+1}^{(k)}, \ldots, x_n^{(k)})$ qui a pour dérivée :

$$\Phi'(\xi) = \partial_i f_i(x_1^{(k+1)}, \ldots, x_{i-1}^{(k+1)}, \xi, x_{i+1}^{(k)}, \ldots, x_n^{(k)})$$

En appliquant la méthode de Newton-Raphson à $\Phi$, on a donc l'expression explicite de $x_i^{(k+1)}$ :

$$x_i^{(k+1)} = x_i^{(k)} - \frac{f_i(x_1^{(k+1)}, \ldots, x_{i-1}^{(k+1)}, x_i^{(k)}, x_{i+1}^{(k)}, \ldots, x_n^{(k)})}{\partial_i f_i(x_1^{(k+1)}, \ldots, x_{i-1}^{(k+1)}, x_i^{(k)}, x_{i+1}^{(k)}, \ldots, x_n^{(k)})}$$

Le passage de $\underline{x}^{(k)}$ à $\underline{x}^{(k+1)}$ définie par cette formule itérative ordonnée correspond à une fonction itérative $\underline{F}$ telle que $\underline{x}^{(k+1)} = \underline{F}(\underline{x}^{(k)})$ qui ne peut être explicitée. Cette méthode est très simple à programmer : on ne peut utiliser qu'un vecteur de mémoire $\underline{X}$ pour tous les itérés et programmer :

$$\forall i \in [\![1; n]\!], \quad X_i = X_i - \frac{f_i(X_1, X_2, \ldots, X_n)}{\partial_i f_i(X_1, X_2, \ldots, X_n)}$$

**Exercice 6.12.** Montrer que la méthode de Gauss-Seidel-Newton appliquée à un système linéaire $\underline{\underline{A}}.\underline{x} = \underline{b}$ redonne la formule de Gauss-Seidel :

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left( b_i - \sum_{j<i} a_{ij} x_j^{(k+1)} - \sum_{j>i} a_{ij} x_j^{(k)} \right)$$

## 6.7 Exercices supplémentaires

### 6.7.1 Exercices avancés

**Exercice 6.13.** Résoudre le système non-linéaire suivant grâce à la méthode de Newton :

$$\left\{ \begin{array}{lcl} x_2^3 - x_1^3 - x_1 - 1 & = & 0 \\ 2x_2^3 - 2x_1^3 - 2x_1 + x_2 + 1 & = & 0 \end{array} \right.$$

On commencera par évaluer $\underline{x}^\star$ et on choisira $\underline{x}^{(0)}$ en conséquence.

**Exercice 6.14.** On considère le système défini par :

$$\underline{F}(\underline{x}) = \begin{pmatrix} x_1^2 + x_2^2 - 1 \\ 2x_1 + x_2 - 1 \end{pmatrix} = \underline{0}$$

dont les deux points fixes sont :

$$\tilde{\underline{x}} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \quad ; \quad \underline{x}^\star = \begin{pmatrix} 4/5 \\ -3/5 \end{pmatrix}$$

On considère les méthodes de point fixe définies par $\underline{G}_i$ telles que :

$$\underline{G}_1(\underline{x}) = \begin{pmatrix} \frac{1-x_2}{2} \\ \sqrt{1-x_1^2} \end{pmatrix} \quad ; \quad \underline{G}_2(\underline{x}) = \begin{pmatrix} \frac{1-x_2}{2} \\ -\sqrt{1-x_1^2} \end{pmatrix}$$

Analyser la convergence de ces méthodes.

**Exercice 6.15.** On veut caractériser l'inverse de $\alpha > 0$ comme étant l'unique zéro des fonctions $f$ et $g$ définies par :

$$f(x) = \alpha - \frac{1}{x} \qquad g(x) = \alpha^3 - \frac{1}{x^3}$$

1. Expliciter la fonction $F$ (respectivement $G$) de l'itération de Newton pour la résolution de $f(x) = 0$ (resp. $g(x) = 0$).
2. On considère les suites $(x^{(k)})$ et $(y^{(k)})$ définies par :

$$\left\{ \begin{array}{l} x^{(0)} \\ x^{(k+1)} = F(x^{(k)}) \end{array} \right. \qquad \left\{ \begin{array}{l} y^{(0)} \\ y^{(k+1)} = G(y^{(k)}) \end{array} \right.$$

   Étudier la convergence de ces suites selon la valeur de $x^{(0)}$ et de $y^{(0)}$. Pour $x^{(k)}$, on pourra utiliser $r^{(k)} = 1 - \alpha x^{(k)}$.
3. On considère $\alpha = 3,141592$ et $x^{(0)} = 0,1$. Calculer l'inverse de $\alpha$ à $10^{-6}$ près avec les deux méthodes.

**Exercice 6.16.** Pour alléger les calculs de la méthode de Newton, on peut approximer $f'(x^{(k)})$ par :

$$\frac{f(x^{(k)}) - f(x^{(k-1)})}{x^{(k)} - x^{(k-1)}}$$

C'est la méthode dite de la *sécante*. Représenter graphiquement le principe de cette méthode dans $\mathbb{R}$, puis résoudre (3 itérations) l'équation suivante avec les deux méthodes :

$$f(x) = e^{-x} - x = 0$$

en prenant $x^{(0)} = 0$ et $x^{(1)} = 1$.
Comparer les vitesses de convergence de ces deux méthodes

### 6.7.2 TD

**Exercice 6.17.** On veut approcher numériquement $\ln 3$. Pour ce faire on va étudier la fonction :

$$f(x) = 1 - 3\,e^{-x}$$

1. Montrer que la fonction $f$ a une unique racine $\ell$ sur $\mathbb{R}$ et déterminer $k \in \mathbb{Z}$ tel que $\ell \in I = [k, k+1]$.
2. On cherche $\ell$ à l'aide de la suite $\left(x^{(n)}\right)_n$ définie par :

$$\begin{cases} x^{(0)} \in I \\ x^{(n+1)} = x^{(n)} + f\left(x^{(n)}\right) \end{cases}$$

   Cette suite converge-t-elle vers $\ell$ ?
3. Soit $\lambda \in \mathbb{R}$. On définit une nouvelle suite $\left(x^{(n)}\right)_n$ comme :

$$\begin{cases} x^{(0)} \in I \\ x^{(n+1)} = x^{(n)} + \lambda f\left(x^{(n)}\right) \end{cases}$$

   Déterminer les valeurs de $\lambda$ pour lesquelles cette suite converge vers $\ell$ ?
4. On cherche enfin la racine de $f(x)$ par la méthode de Newton.
   (i) Donner la fonction d'itération correspondante.
   (ii) Déterminer l'intervalle dans lequel on doit choisir $x^{(0)}$ pour que la convergence de la méthode soit assurée.
5. Déterminer $\ell$ à $10^{-5}$ près (test sur le résidu) par la méthode décrite à la question (3) pour $\lambda = -\frac{e}{2}$ et par la méthode de Newton en prenant $x^{(0)} = 1$.

**Exercice 6.18.** Soit $f$ la fonction de $I = ]0; 4[$ dans $\mathbb{R}$ définie par :

$$f(x) = 4\,e^{\frac{x}{4}} - 6$$

1. Montrer que la fonction $f$ possède un unique zéro $\bar{x}$ sur $I$.
2. On veut approcher numériquement $\bar{x}$ par les méthodes de points fixes :

$$x^{(k+1)} = \phi_i\left(x^{(k)}\right) \qquad i \in [\![1, 3]\!]$$

   où

$$\begin{cases} \phi_1(x) = x + f(x) \\ \phi_2(x) = x - \dfrac{f(x)}{64} \\ \phi_3 \text{ est la fonction d'itération de Newton.} \end{cases}$$

   Étudier la convergence de ces trois méthodes pour tout $x^{(0)} \in I$ et donner le cas échéant leur ordre de convergence.
3. Pour la méthode $\phi_2$, trouver $K_2 \in ]0, 1[$ tel que :

$$|x^{(k+1)} - \bar{x}| \leq K_2\,|x^{(k)} - \bar{x}|$$

4. Pour la méthode $\phi_3$, trouver $K_3 \in ]0, 1[$ tel que :

$$|x^{(k+1)} - \bar{x}| \leq K_3\,|x^{(k)} - \bar{x}|^2$$

5. Déterminer pour chaque méthode $\phi_i, i = 2, 3$, le nombre minimal d'itérations nécessaires pour avoir une erreur inférieure à $10^{-6}$ si on choisit $x^{(0)}$ tel que :

$$|x^{(0)} - \bar{x}| < 2$$

6. Calculer numériquement $\bar{x}$ à $10^{-6}$ près par la méthode de votre choix. On prendra $x^{(0)} = 1$ et on détaillera les itérations nécessaires pour obtenir le résultat.

### 6.7.3 Annales

**Exercice 6.19 (**Annale 2019**).** Soient $U \subset \mathbb{R}$ un ouvert non vide, $f : U \to \mathbb{R}$ une fonction de classe $\mathcal{C}^2(U)$ et $x^\star \in U$ une racine simple de $f$. Considérons la méthode itérative suivante :

$$G_f(x) = x - \frac{f(x)}{df(x)} \quad \text{où} \quad df(x) = \frac{1}{2}\left[\frac{f(x+f(x)) - f(x)}{f(x)} + \frac{f(x) - f(x-f(x))}{f(x)}\right]$$

1. Dans le but de déterminer l'ordre de convergence de la méthode, montrer le résultat intermédiaire suivant :

$$f(x^\star + h + f(x^\star + h)) - f(x^\star + h - f(x^\star + h)) = 2\,h\,f'(x^\star)^2\left(1 + \frac{3h}{2}\frac{f''(x^\star)}{f'(x^\star)} + o(h)\right)$$

2. En déduire l'ordre de la méthode itérative.
3. Comparer avec les méthodes de Newton-Raphson et de Steffensen.

**Exercice 6.20 (**Annale 2020**).** Soient $a \in \mathbb{R}_+^\star$ et $f_a$ la fonction définie sur $\mathbb{R}$ par $f_a(x) = x\,(x^2 - a^2)$. Soit $F_a$ une méthode itérative "artisanale", définie par $F_a(x) = x - f_a(x)$. Le but de cet exercice est de déterminer une partie des bassins d'attraction des points fixes de $F_a$.

**Préliminaires.** Déterminer les points fixes de $F_a$, puis appliquer le théorème d'Ostrowski pour chacun d'entre eux. Quelles conclusions peut-on en tirer ? Dans la suite on prendra $a = 1/\sqrt{2}$. Ce choix est-il judicieux (à justifier) ? A l'aide du logiciel de votre choix, tracer les courbes $x \mapsto F_a(x)$ et $x \mapsto x$.

**Variations de $F_a$** . Etablir le tableau de variations de $F_a$ en y ajoutant les valeurs remarquables suivantes :

— Déterminer les racines de $F_a$, notées $x_F^-, x_F^\star$ et $x_F^+$ telles que $x_F^- < x_F^\star < x_F^+$ (Expressions à déterminer)

— Montrer que :

$$\begin{aligned}
&\exists! x'_F > 0 / \; F_a(x'_F) = -a \quad ; \quad && F_a(-x'_F) = a \\
&\exists! x''_F > 0 / \; F_a(x''_F) = x_F^- \quad ; \quad && F_a(-x''_F) = x_F^+ \\
&\exists! x'''_F > 0 / \; F_a(x'''_F) = -x'_F \quad ; \quad && F_a(-x'''_F) = x'_F
\end{aligned}$$

vérifiant

$$-\infty < -x'''_F < -x''_F < -x'_F < x_F^- < -a < x_F^\star < a < x_F^+ < x'_F < x''_F < x'''_F < +\infty$$

On ne cherchera pas à déterminer les valeurs de $x'_F, x''_F, x'''_F$.

**Images réciproques et antécédents par $F_a$.** A l'aide du tableau de variations et du graphe de $F_a$, déterminer les intervalles ouverts :

— $\{I_i\}_{1\le i\le 4}$ tels que $F_a^{-1}(I_0 = ]0, a[) = I_0 \cup I_1 \cup I_2$, $F_a^{-1}(I_1) = I_3$ et $F_a^{-1}(I_2) = I_4$

— $\{J_i\}_{1\le i\le 4}$ tels que $F_a^{-1}(J_0 = ]-a, 0[) = J_0 \cup J_1 \cup J_2$, $F_a^{-1}(J_1) = J_3$ et $F_a^{-1}(J_2) = J_4$

Déterminer également les antécédents par $F_a$ de $\{x_F^\star, \pm a, x_F^-, x_F^+\}$.

**Bassins.** Déterminer les ouverts tels que $F_a(x) > x$ et $F_a(x) < x$. Déduire de tout ce qui précède les bassins d'attractions des points fixes de $F_a$ dans l'intervalle ouvert $]-x'''_F, x'''_F[$.

**Exercice 6.21 (**Annale rattrapage 2020**).** Soit $r \geq 2$ un entier, $U \subset \mathbb{R}$ un ouvert non vide, $f : U \to \mathbb{R}$ une fonction de classe $\mathcal{C}^r(\mathbb{R})$ et $x^\star \in U$ une racine simple de f, i.e $f(x^\star) = 0$ et $f'(x^\star) \neq 0$. On supposera de plus que $\forall 2 \leq i \leq r-1, f^{(i)}(x^\star) = 0$ et $f^{(r)}(x^\star) \neq 0$. Afin d'alléger les notations, on pourra utiliser la notation $c_r = \frac{1}{r!}\frac{f^{(r)}(x^\star)}{f'(x^\star)}$.

1. Ecrire le développement limité de $f$ autour de $x^\star$ à l'ordre $r$. En déduire les développements limités de $x \mapsto f'(x)$ et $x \mapsto f^2(x)$ autour de $x^\star$ (ordre à déterminer).
2. Démontrer l'ordre de convergence de la méthode de Newton-Raphson.

3. Démontrer que l'ordre de convergence de la méthode de Steffensen, dont on rappelle la définition :

$$S_f(x) = x - \frac{f^2(x)}{f(x + f(x)) - f(x)}$$

est d'ordre $r$ si $[1 + f'(x^\star)]^r \neq 1 + f'(x^\star)$

# Chapitre 7

# Rappels et compléments

## Sommaire

## 7.1 Rappels d'algèbre linéaire

**Proposition 7.1.** *Tout système linéaire possède soit zéro solution, soit une unique solution soit une infinité de solutions.*

**Définition 7.1.**
Un système est dit homogène si le second membre est nul, i.e qu'on considère le problème :

$$\text{Soit } \underline{\underline{A}} \in \mathcal{M}_{mn}(\mathbb{R}).\text{Trouver } \underline{x} \in \mathbb{R}^n \text{ tel que } \underline{\underline{A}}.\underline{x} = \underline{0} \tag{7.1}$$

$\underline{x}$ représente donc le noyau de la matrice $\underline{\underline{A}}$, qui contient au moins le vecteur nul. Par conséquent, tout système homogène contient au moins une solution, le vecteur nul $\underline{x} = \underline{0}$.

### 7.1.1 Systèmes carrés

**Théorème 7.1.**
*Une matrice $\underline{\underline{A}} \in \mathcal{M}_n(\mathbb{R})$ est dite inversible si et seulement si son déterminant est non nul.*

**Définition 7.2.**
Soit $\underline{\underline{A}} \in \mathcal{M}_n(\mathbb{K})$. On dit que $\underline{\underline{A}}$ est à diagonale strictement dominante si le module de chaque terme diagonal de $\underline{\underline{A}}$ est strictement supérieur à la somme des modules des autres termes de sa ligne :

$$\forall i \in [\![1; n]\!],\ |a_{ii}| > \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|$$

**Théorème 7.2.**
*Toute matrice à diagonale strictement dominante est inversible.*

**Proposition 7.2 (**Système homogène**).** *Considérons le système homogène* (7.1) *avec $m = n$. Si $\underline{\underline{A}} \in \mathcal{M}_n(\mathbb{R})$ est inversible, alors le système* (7.1) *possède une unique solution, la solution nulle. Sinon, le système* (7.1) *a une infinité de solutions. Si la dimension du noyau de $\underline{\underline{A}}$ vaut $k$, alors le rang du système vaut $n - k$.*

**Proposition 7.3 (**Système non homogène**).** *Considérons à présent le système non homogène* (4.1) *avec $m = n$. Si $\underline{\underline{A}} \in \mathcal{M}_n(\mathbb{R})$ est inversible, alors le système possède une unique solution $\underline{x} = \underline{\underline{A}}^{-1}.\underline{b}$. Sinon, pour que le système* (4.1) *possède au moins une solution il faut que rang($\underline{\underline{A}}$) =rang($\underline{\underline{Ab}}$), où $\underline{\underline{Ab}}$ est la matrice résultant de la concaténation de la matrice $\underline{\underline{A}}$ et du vecteur $\underline{b}$.*

**Exemple 7.1.** Soient la matrice $\underline{\underline{A}} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$ et les vecteurs $\underline{b}_1 = \begin{pmatrix} 1 \\ 4 \\ 2 \end{pmatrix}$, $\underline{b}_2 = \begin{pmatrix} 2 \\ 4 \\ 6 \end{pmatrix}$. Les systèmes $\underline{\underline{A}}.\underline{x} = \underline{b}_1$ et $\underline{\underline{A}}.\underline{x} = \underline{b}_2$ ont-ils des solutions ?

$\det(\underline{\underline{A}}) = 0$ donc $\underline{\underline{A}}$ n'est pas inversible donc le système admet soit aucune solution soit une infinité de solutions. Par combinaisons linéaires des lignes :

— $\ker(\underline{\underline{A}})=\text{Vect}\left\{\begin{pmatrix} -1 \\ 2 \\ -1 \end{pmatrix}\right\}$ donc $\text{rang}(\underline{\underline{A}}) = 2$

— $\ker(\underline{\underline{Ab}}_1)=\text{Vect}\left\{\begin{pmatrix} -1 \\ 2 \\ -1 \\ 0 \end{pmatrix}\right\}$ donc $\text{rang}(\underline{\underline{Ab}}_1) = 3$

— $\ker(\underline{\underline{Ab}}_2)=\text{Vect}\left\{\begin{pmatrix} -1 \\ 2 \\ -1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 2 \\ -3 \end{pmatrix}\right\}$ donc $\text{rang}(\underline{\underline{Ab}}_2) = 2 = \text{rang}(\underline{\underline{A}})$

Par conséquent, le système $\underline{\underline{A}}.\underline{x} = \underline{b}_1$ ne possède pas de solution, alors que le système $\underline{\underline{A}}.\underline{x} = \underline{b}_2$ en possède une infinité.

### 7.1.2 Systèmes non carrés

Deux cas apparaissent :

1. Si $m > n$, i.e s'il y a plus d'équations que d'inconnues : dans ce cas le système est dit **sur-déterminé**. En général ce système n'admet pas de solution, pour le prouver on procède comme précédemment par calcul des rangs des matrices $\underline{\underline{A}}$ et $\underline{\underline{Ab}}$. On verra par la suite qu'il est tout de même possible de cherche une solution approchée, par exemple avec la méthodes des moindres carrés. Ce type de problème se rencontre fréquemment en analyse de mesures expérimentales, par exemple lorsqu'on procède aux mesures de convergence en tunnel à l'aide de cordes.
2. Si $m < n$, i.e s'il y a moins d'équations que d'inconnues : le système est alors **sous-déterminé**, il y aura une infinité de solutions.

### 7.1.3 Déterminant et Formules de Cramer

**Définition 7.3 (**Déterminant**).**

Soit $\underline{\underline{A}} \in \mathcal{M}_n(\mathbb{R})$

— si $n = 1$, on appelle déterminant de $\underline{\underline{A}}$, noté $\Delta_A$, est le réel $a_{11} = \Delta_A$

— si $n > 1$, $\forall i \in [\![1; n]\!]$, on note $\Delta_i$ le déterminant de la matrice de taille $(n-1) \times (n-1)$ en enlevant la première colonne et la $i^{\text{ème}}$ ligne de la matrice initiale. Ainsi :

$$\Delta_A = \sum_{i=1}^{n} (-1)^{i+1} a_{i1} \Delta_i$$

Estimons le coût du calcul du déterminant d'une matrice de taille $n \times n$ selon cette procédure :

— $n$ calculs de déterminants de taille $(n-1) \times (n-1)$

— $2n$ multiplications

— $(n-1)$ additions

Puis pour chacun des déterminants de taille $(n-1) \times (n-1)$ :

— $(n-1)$ calculs de déterminants de taille $(n-2) \times (n-2)$

— $(n-1)$ multiplications

— $(n-2)$ additions

etc.

Approximativement, le coût de calcul du déterminant d'une matrice $n \times n$ est $O(n(n!))$ opérations. Soit environ $4.10^8$ opérations élémentaires $(+, -, \times, /)$ pour un système de dimension 10, et $4.9\,10^{19}$ pour un système de dimension 20. Pour se donner un ordre d'idée, en prenant un ordinateur avec un processeur de 2.7GHz et en faisant une approximation relativement fausse qui dirait que l'ordinateur peur effectuer $2.7 \times 10^9$ opérations par seconde, ce calcul prendrait environ 570 ans...

**Théorème 7.3 (**Méthode de Cramer**).**
*Soit le système non homogène* (4.1) *carré, avec* $\Delta_A \neq 0$. *Si on note* $(\delta_j)_{1\leq j\leq n}$ *le déterminant obtenu en remplaçant la* $j^{ème}$ *colonne de* $\underline{\underline{A}}$ *par le second membre* $\underline{b}$, *l'unique solution est :*

$$\forall j \in [\![1; n]\!], \quad x_j = \frac{\delta_j}{\Delta_A}$$

Le coût de calcul de la méthode de Cramer est donc directement lié au coût de calcul du déterminant. Cette méthode est donc attrayante théoriquement, mais en pratique on la réservera pour une matrice de dimension 2 voire 3. Pour les dimensions supérieures on préfèrera d'autres méthodes, par exemple le pivot de Gauss.

### 7.1.4 Généralités sur les matrices carrées

**Définition 7.4.**
$\underline{\underline{A}} \in \mathcal{M}_n(\mathbb{K})$ est dite
- hermitienne si $\underline{\underline{A}}^{\mathrm{H}} = \underline{\underline{A}}$
- normale si $\underline{\underline{A}}^{\mathrm{H}}.\underline{\underline{A}} = \underline{\underline{A}}.\underline{\underline{A}}^{\mathrm{H}}$
- unitaire si $\underline{\underline{A}}^{\mathrm{H}}.\underline{\underline{A}} = \underline{\underline{A}}.\underline{\underline{A}}^{\mathrm{H}} = \underline{\underline{I}}$

$\underline{\underline{A}} \in \mathcal{M}_n(\mathbb{R})$ est dite
- symétrique si ${}^t\underline{\underline{A}} = \underline{\underline{A}}$
- orthogonale si ${}^t\underline{\underline{A}}.\underline{\underline{A}} = \underline{\underline{A}}.{}^t\underline{\underline{A}} = \underline{\underline{I}}$

**Définition 7.5.**
Soit $\underline{\underline{M}} \in \mathcal{M}_n(\mathbb{K})$, on note $\underline{\underline{M}}^\star, \underline{\underline{M}}^{\mathrm{H}}$ ou $\underline{\underline{M}}^\dagger$ la matrice transconjuguée (ou adjointe) de $\underline{\underline{M}}$, i.e la matrice transposée de la matrice conjuguée de $\underline{\underline{M}}$ :

$$\underline{\underline{M}}^{\mathrm{H}} = {}^t\,\overline{\underline{\underline{M}}} \text{ telle que } \left(\underline{\underline{M}}^{\mathrm{H}}\right)_{ij} = \overline{M}_{ji}$$

**Remarque 7.1.**
Dans ce cours nous utiliserons la notation $\underline{\underline{M}}^{\mathrm{H}}$, plutôt que $\underline{\underline{M}}^\star$ (notation déjà utilisée) et $\underline{\underline{M}}^\dagger$ (utilisée dans le cadre de la mécanique quantique).

**Définition 7.6.**
Soit $\underline{\underline{A}} \in \mathcal{M}_n(\mathbb{K})$. On appelle rayon spectral de $\underline{\underline{A}}$, noté $\rho(\underline{\underline{A}})$, la quantité :

$$\rho(\underline{\underline{A}}) = \max_{1\leq i\leq n} |\lambda_i|$$

où $\{\lambda_i\}_{1\leq i\leq n}$ sont les valeurs propres de $\underline{\underline{A}}$.

### 7.1.5 Matrice définie positive

**Définition 7.7 (**Matrice ou vecteur positif**).**
Une matrice ou un vecteur sont dits positifs si toutes leurs composantes sont positives.

**Définition 7.8.**
Une matrice $\underline{\underline{A}} \in \mathcal{M}_n(\mathbb{R})$ symétrique est dite définie positive si elle vérifie l'une des propriétés suivantes :

1. $\forall \underline{x} \in \mathbb{R}^n / \underline{x} \neq \underline{0}$, ${}^t\underline{x}.\underline{\underline{A}}\underline{x} > 0$ ;
2. Toutes les valeurs propres de $\underline{\underline{A}}$ sont strictement positives ($\underline{\underline{A}}$ étant nécessairement diagonalisable) ;
3. La forme bilinéaire symétrique : $\begin{cases} \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R} \\ (\underline{x}, \underline{y}) \mapsto {}^t\underline{x}.\underline{\underline{A}}.\underline{y} \end{cases}$ définit un produit scalaire sur $\mathbb{R}^n$.

**Remarque 7.2.**
Une matrice définie positive est une matrice positive et inversible. Son déterminant est strictement positif.

**Définition 7.9.**
$\underline{\underline{A}} \in \mathcal{M}_n(\mathbb{R})$ symétrique est dite définie négative si $-\underline{\underline{A}}$ est définie positive.

**Théorème 7.4 (**Critère de Sylvester**).**
*Soient $\underline{\underline{A}} \in \mathcal{M}_n(\mathbb{R})$ une matrice symétrique de coefficients $\{a_{ij},\ 1 \leq i,j \leq n\}$ et pour tout $k \in [\![1;n]\!]$ on considère les matrices $\underline{\underline{A}}_k \in \mathcal{M}_k(\mathbb{R})$ telles que $\{(\underline{\underline{A}}_k)_{ij} = a_{ij},\ 1 \leq i,j \leq k\}$. $\underline{\underline{A}}$ est définie positive si et seulement si $\forall k \in [\![1;n]\!]$, $\det(\underline{\underline{A}}_k) > 0$.*

### 7.1.6 Matrice monotone

**Définition 7.10 (**Matrice monotone**).**
Une matrice $\underline{\underline{A}} \in \mathcal{M}_n(\mathbb{R})$ est monotone si $\underline{\underline{A}}$ est inversible et que $\underline{\underline{A}}^{-1} \geq 0$.

**Proposition 7.4.** *Une matrice $\underline{\underline{A}} \in \mathcal{M}_n(\mathbb{R})$ est monotone si et seulement si :*

$$\forall \underline{X} \in \mathbb{R}^N,\ \underline{\underline{A}}.\underline{X} \geq 0 \Rightarrow \underline{X} \geq 0$$

**Définition 7.11 (**M-matrice**).**
Une matrice $\underline{\underline{A}} \in \mathcal{M}_n(\mathbb{R})$ est une M-matrice si elle est monotone et si $a_{ij} \leq 0$ pour $i \neq j$.

**Proposition 7.5.** *Soit une matrice $\underline{\underline{A}} \in \mathcal{M}_n(\mathbb{R})$ telle que :*
- *$a_{ij} \leq 0$ pour $i \neq j$*
- *$\forall i \in [\![1;n]\!]$, $\sum_{1<j<n} a_{ij} > 0$*

*Alors $\underline{\underline{A}}$ est une M-matrice.*

Une autre proposition :

**Proposition 7.6.** *Soit une matrice $\underline{\underline{A}} \in \mathcal{M}_n(\mathbb{R})$ telle que :*
- *$a_{ij} \leq 0$ pour $i \neq j$*
- *$\forall i \in [\![1;n]\!]$, $\sum_{1<j<n} a_{ij} \geq 0$*
- *$\underline{\underline{A}}$ est inversible.*

*Alors $\underline{\underline{A}}$ est une M-matrice.*

**Exercice 7.1.** Montrer que la matrice $\underline{\underline{A}} = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}$ est une M-matrice.

### 7.1.7 Norme matricielle

**Définition 7.12.**

On dit que l'application $\|.\| : \mathcal{M}_n(\mathbb{K}) \to \mathbb{R}$ est une norme matricielle si :
- $\|.\|$ est une norme de l'espace vectoriel $\mathcal{M}_n(\mathbb{K})$
- $\forall(\underline{\underline{A}}, \underline{\underline{B}}) \in \mathcal{M}_n(\mathbb{K}) \times \mathcal{M}_n(\mathbb{K}), \quad \|\underline{\underline{A}}.\underline{\underline{B}}\| \leq \|\underline{\underline{A}}\| \, \|\underline{\underline{B}}\|$ (sous-multiplicativité)

**Définition 7.13.**

Soit $\|.\|$ une norme sur $E = \mathbb{K}^n$. La norme subordonnée à cette norme est l'application :

$$\begin{aligned} |||.||| : &\mathcal{M}_n(\mathbb{K}) \to \mathbb{R} \\ &\underline{\underline{A}} \mapsto \left|\left|\left|\underline{\underline{A}}\right|\right|\right| = \sup_{\underline{x} \neq \underline{0}} \frac{\|\underline{\underline{A}}.\underline{x}\|}{\|\underline{x}\|} = \sup_{\|\underline{x}\|<1} \frac{\|\underline{\underline{A}}.\underline{x}\|}{\|\underline{x}\|} = \sup_{\|\underline{x}\|=1} \|\underline{\underline{A}}.\underline{x}\| \end{aligned}$$

1. Une norme subordonnée à une norme sur $E$ est bien une norme et c'est une norme matricielle.
2. On a
$$\forall \underline{x} \in E, \ \|\underline{\underline{A}}.\underline{x}\| \leq \left|\left|\left|\underline{\underline{A}}\right|\right|\right| \|\underline{x}\|$$

De plus $\left|\left|\left|\underline{\underline{A}}\right|\right|\right|$ est le plus petit des réels $\lambda \in \mathbb{R} / \ \forall x \in E, \|\underline{\underline{A}}.\underline{x}\| \leq \lambda \|\underline{x}\|$.

**Théorème 7.5 (**Normes usuelles et leurs normes subordonnées**).**

*Soient $\|.\|_1$, $\|.\|_2$, et $\|.\|_\infty$ les normes usuelles sur $\mathbb{K}^n$ et $|||.|||_1$, $|||.|||_2$, et $|||.|||_\infty$ leur norme subordonnée respective. Les expressions de ces normes sont donc, pour $\underline{\underline{A}} \in \mathcal{M}_n(\mathbb{K})$ et $\underline{x} \in \mathbb{R}^n$ :*

$$\begin{cases} \|\underline{x}\|_1 = \sum_{i=1}^{n} |x_i| & \left|\left|\left|\underline{\underline{A}}\right|\right|\right|_1 = \max_{1 \leq j \leq n} \left( \sum_{i=1}^{n} |A_{ij}| \right) \\ \|\underline{x}\|_\infty = \max_{1 \leq i \leq n} |x_i| & \left|\left|\left|\underline{\underline{A}}\right|\right|\right|_\infty = \max_{1 \leq i \leq n} \left( \sum_{j=1}^{n} |A_{ij}| \right) \\ \|\underline{x}\|_2 = \sqrt{\sum_{i=1}^{n} |x_i|^2} & \left|\left|\left|\underline{\underline{A}}\right|\right|\right|_2 = \sqrt{\rho(\underline{\underline{A}}^{\mathrm{H}}.\underline{\underline{A}})} = \sqrt{\rho(\underline{\underline{A}}.\underline{\underline{A}}^{\mathrm{H}})} = \left|\left|\left|\underline{\underline{A}}^{\mathrm{H}}\right|\right|\right|_2 \end{cases}$$

*où $\rho(\underline{\underline{A}})$ est le rayon spectral de $\underline{\underline{A}}$, c'est-à-dire le maximum des valeurs propres en valeur absolue.*

Dans la suite on notera indifféremment, sans confusion possible, $\|.\|$ pour une norme vectorielle ou matricielle, qu'elle soit subordonnée ou non.

**Théorème 7.6.**

*Soit $\underline{\underline{A}} \in \mathcal{M}_n(\mathbb{K})$.*

1. *Si $\underline{\underline{A}}$ est hermitienne ou symétrique (donc normale), alors $\|\underline{\underline{A}}\|_2 = \rho(\underline{\underline{A}})$*
2. *Si $\underline{\underline{A}}$ est unitaire ou orthogonale (donc normale), alors $\|\underline{\underline{A}}\|_2 = 1$*

**Théorème 7.7.**

*Soit $\underline{\underline{A}} \in \mathcal{M}_n(\mathbb{K})$.*

1. *Soit $\|.\|$ une norme matricielle subordonnée ou non quelconque. Alors*

$$\rho(\underline{\underline{A}}) \leq \|\underline{\underline{A}}\|$$

2. *Soit $\varepsilon > 0$, il existe au moins une norme matricielle subordonnée telle que :*

$$\|\underline{\underline{A}}\| \leq \rho(\underline{\underline{A}}) + \varepsilon$$

**Définition 7.14 (**Norme de Frobenius**).**
Soit le produit scalaire défini par

$$\forall(\underline{\underline{A}}, \underline{\underline{B}}) \in \mathcal{M}_{mn}(\mathbb{K}) \times \mathcal{M}_{mn}(\mathbb{K}),\ \langle \underline{\underline{A}}, \underline{\underline{B}} \rangle_F = \mathrm{tr}\left(\underline{\underline{A}}^{\mathrm{H}}.\underline{\underline{B}}\right) = \mathrm{tr}\left(\underline{\underline{B}}^{\mathrm{H}}.\underline{\underline{A}}\right)$$

$\|.\|_F$ la norme de Frobenius, appelée également norme euclidienne matricielle, est la norme induite par ce produit scalaire :

$$\begin{aligned} \|.\|_F :& \mathcal{M}_{mn}(\mathbb{K}) \to \mathbb{R}^+ \\ & \underline{\underline{A}} \mapsto \|\underline{\underline{A}}\|_F = \sqrt{\mathrm{tr}\left(\underline{\underline{A}}^{\mathrm{H}}.\underline{\underline{A}}\right)} = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}|^2} \end{aligned}$$

**Remarque 7.3.**
La norme de Frobenius
- est une norme matricielle, qui est donc également une norme vectorielle ;
- mais ce n'est pas pas une norme subordonnée par une norme vectorielle (comme les normes matricielles $\|.\|_1$, $\|.\|_2$ ou $\|.\|_\infty$). On pourra en particuler vérifier que $\|\underline{\underline{I}}\|_F = \sqrt{n}$.

## 7.2 Continuité et dérivabilité des fonctions d'une variable réelle.

### 7.2.1 Définitions

**Définition 7.15.**
$f : I \subset \mathbb{R} \to \mathbb{R}$ est continue sur $I$ si et seulement si

$$\forall x \in I, \forall \varepsilon > 0, \exists \eta > 0 / \quad \forall y \in I, |x - y| \leq \eta \Rightarrow |f(x) - f(y)| \leq \varepsilon$$

**Définition 7.16.**
$f : I \subset \mathbb{R} \to \mathbb{R}$ est uniformément continue sur $I$ si et seulement si

$$\forall \varepsilon > 0, \exists \eta > 0 / \quad \forall (x, y) \in I \times I, |x - y| \leq \eta \Rightarrow |f(x) - f(y)| \leq \varepsilon$$

**Exemple 7.2.** La fonction $f : x \mapsto \sqrt{x}$ est uniformément continue sur $I = \mathbb{R}^+$. En effet,

$$\forall (x, y) \in I \times I, |\sqrt{x} - \sqrt{y}| \leq \sqrt{|x - y|}$$

car $0 \leq x \leq y \Rightarrow x \leq y \leq x + (y - x) + 2\sqrt{x(y - x)} = \left(\sqrt{x} + \sqrt{y - x}\right)^2$

**Proposition 7.7.** *Soit $f : I \subset \mathbb{R} \to \mathbb{R}$. Alors on a l'implication :*

$$f \text{ uniformément continue sur } I \Rightarrow f \text{ continue sur } I$$

**Remarque 7.4.**
Attention la réciproque est fausse dans le cas général.

**Contre-exemple 1.** Soit la fonction $f : \mathbb{R} \to \mathbb{R}$ telle que $f : x \mapsto x^2$. Montrons que $f$ est continue sur $\mathbb{R}$ mais pas uniformément continue sur $\mathbb{R}$.

$f$ continue sur $\mathbb{R}$ :

$$\forall x \in \mathbb{R}, \exists \eta > 0 / \, \forall y \in \mathbb{R}, \, |x-y| \leq \eta \Rightarrow |f(x) - f(y)| = |x-y| \, |x+y| \leq \eta \, |\eta + 2y| = \varepsilon$$

$f$ non uniformément continue sur $\mathbb{R}$ : Soit $\eta \in \mathbb{R}$ tel que $x = \eta + 1/\eta$ et $y = 1/\eta$. Ainsi $\forall \eta > 0$, $|x-y| \leq \eta$. De plus, $|f(x) - f(y)| = |\eta^2 + 2| > 2$. En prenant $\varepsilon = 2$,

$$\exists \varepsilon = 2 > 0 / \forall \eta > 0 / \quad \exists (x, y) \in \mathbb{R}^2, |x-y| \leq \eta \text{ et } |f(x) - f(y)| > \varepsilon$$

Donc $f$ n'est pas uniformément continue sur $\mathbb{R}$.

**Définition 7.17.**
$f : I \subset \mathbb{R} \to \mathbb{R}$ est dite $k$-lipschitzienne sur $I$ si et seulement si

$$\exists k > 0 / \, \forall (x, y) \in I \times I \, |f(x) - f(y)| \leq k \, |x - y|$$

Si $k \leq 1$ (resp. $k < 1$), alors $f$ est dite contractante (resp. strictement contractante).

**Exemple 7.3.** Toute fonction $\mathcal{C}^1$ sur un intervalle $I$ fermé borné est lipschitzienne sur $I$. (Voir théorème des accroissements finis)

**Contre-exemple 2.** Soit $\forall \varepsilon > 0$, $f : [0, \varepsilon] \to \mathbb{R}$ telle que $f : x \mapsto x^\alpha$, $|\alpha| < 1$ ? Alors $f$ n'est pas lipschitzienne car $\forall x > 0$, $\frac{f(x)-f(0)}{x-0} = x^{\alpha-1}$ non bornée au voisinage de 0.

**Proposition 7.8.** *Soit $f : I \subset \mathbb{R} \to \mathbb{R}$. Alors on a l'implication :*

$$f \, k - \textit{lispchitzienne sur } I \Rightarrow f \textit{ uniformément continue sur } I \Rightarrow f \textit{ continue sur } I$$

### 7.2.2 Théorèmes fondamentaux

**Théorème 7.8** (Heine)**.**
*Toute fonction $f$ à valeurs réelles continue sur un intervalle fermé borné de $\mathbb{R}$ est uniformément continue.*

**Théorème 7.9** (des valeurs intermédiaires)**.**
*Soient $I \subset \mathbb{R}$ un intervalle et $f : I \to \mathbb{R}$ une application continue. Alors $f(I)$ est un intervalle.*

**Théorème 7.10** (Rolle)**.**
*Si une fonction $f : [a, b] \to \mathbb{R}$ est continue sur $[a, b]$, dérivable sur $]a, b[$ et si $f(a) = f(b)$ alors il existe $c$ dans $]a, b[$ tel que $f'(c) = 0$.*

**Théorème 7.11** (accroissements finis)**.**
*Soit $f : [a, b] \to \mathbb{R}$ une fonction continue sur $[a, b]$ et dérivable sur $]a, b[$. Alors il existe $c \in ]a, b[$ tel que :*

$$f'(c) = \frac{f(b) - f(a)}{b - a}$$

### 7.2.3 Continuité et dérivabilité des fonctions limites

**Théorème 7.12.**
*Soit $(f_n)_{n\geq 0}$, une suite de fonctions continues sur un intervalle $I \subset \mathbb{R}$ dans $\mathbb{R}$. Si $(f_n)$ converge uniformément sur $I$ vers une certaine fonction $f$, alors $f$ est continue sur $I$.*

**Théorème 7.13.**
*Soit $(f_n)_{n\geq 0}$ une suite de fonctions de classe $\mathcal{C}^1$ d'un intervalle $I \subset \mathbb{R}$ dans $\mathbb{R}$. On suppose que (i) $(f_n)$ converge simplement sur $I$ vers $f$ et (ii) la suite $(f'_n)_{n\geq 0}$ converge uniformément sur $I$ vers une certaine fonction $g$. Alors $f$ est de classe $\mathcal{C}^1$ et $f' = g$.*

**Théorème 7.14.**
*Soient $E$ l'espace vectoriel réel des fonctions continues de $[0, 1]$ dans $\mathbb{R}$, muni de $\|.\|_\infty$, et $F$ l'ensemble des fonctions de $E$ dérivables nulle part. Alors (i) $(E, \|.\|_\infty)$ est un espace de Banach et (ii) $F$ est dense dans $E$.*

**Théorème 7.15.**
*Soient $(X, \tau, \mu)$ un espace mesuré et $f : \mathbb{R} \times X \to \mathbb{R}$ une fonction. Si les conditions suivantes sont vérifiées :*

1. *pour tout $t \in \mathbb{R}$, la fonction $x \mapsto f(t, x)$ est mesurable ;*
2. *pour presque tout $x \in X$, la fonction $t \mapsto f(t, x)$ est continue sur $\mathbb{R}$ ;*
3. *il existe une fonction $g \in L^1(X)$ telle que pour presque tout $x \in X$ :*

$$|f(t, x)| \leq g(x) \qquad t \in \mathbb{R}$$

*alors la fonction $t \mapsto \int_X f(t, x)\, d\mu$ est continue sur $\mathbb{R}$.*

**Théorème 7.16.**
*Soient $I \subset \mathbb{R}$ un intervalle et $f : I \times X \to \mathbb{R}$ une fonction. Si les conditions suivantes sont vérifiées :*

1. *pour tout $t \in I$, la fonction $x \mapsto f(t, x)$ est dans $L^1(X)$ ;*
2. *pour presque tout $x \in X$, la fonction $t \mapsto f(t, x)$ est dérivable sur $I$ ;*

*3. pour tout compact $K \in I$, il existe $g \in L^1(X)$ telle que pour presque tout $x \in X$*

$$|\partial_1 f(t,x)| \leq g(x) \qquad \forall t \in K$$

*alors la fonction $t \mapsto \int f(t,x)\,d\mu$ est dérivable sur $I$ et :*

$$\partial_1 \int f(t,x)\,d\mu = \int \partial_1 f(t,x)\,d\mu$$

## 7.3 Formules trigonométriques usuelles.

### 7.3.1 Trigonométrie circulaire.

**Formules d'addition et de différence des arcs.**

| $\cos(x) = \frac{e^{ix}+e^{-ix}}{2}$ | $\sin(x) = \frac{e^{ix}-e^{-ix}}{2\,i}$ |
|---|---|
| $\cos(a+b) = \cos(a)\,\cos(b) - \sin(a)\,\sin(b)$ | $\sin(a+b) = \sin(a)\,\cos(b) + \cos(a)\,\sin(b)$ |
| $\cos(a-b) = \cos(a)\,\cos(b) + \sin(a)\,\sin(b)$ | $\sin(a-b) = \sin(a)\,\cos(b) - \cos(a)\,\sin(b)$ |
| $\cos(2\,a) = \begin{cases} \cos^2(a) - \sin^2(a) \\ 2\,\cos^2(a) - 1 \\ 1 - 2\,\sin^2(a) \end{cases}$ | $\sin(2\,a) = 2\sin(a)\cos(a)$ |
| $\cos(3\,a) = -3\,\cos(a) + 4\,\cos^3(a)$ | $\sin(3\,a) = 3\,\sin(a) - 4\,\sin^3(a)$ |

| $\tan(x) = \frac{\sin(x)}{\cos(x)}$ | |
|---|---|
| $\tan(a+b) = \frac{\tan(a)+\tan(b)}{1-\tan(a)\,\tan(b)}$ | $\tan(2\,a) = \frac{2\,\tan(a)}{1-\tan^2(a)}$ |
| $\tan(a-b) = \frac{\tan(a)-\tan(b)}{1+\tan(a)\,\tan(b)}$ | $\tan(3\,a) = \frac{3\,\tan(a)-\tan^3(a)}{1-3\,\tan^2(a)}$ |

**Formules de Simpson.**

$$\cos(p) + \cos(q) = 2\cos\left(\tfrac{p+q}{2}\right)\cos\left(\tfrac{p-q}{2}\right) \qquad \cos(p) - \cos(q) = -2\sin\left(\tfrac{p+q}{2}\right)\sin\left(\tfrac{p-q}{2}\right)$$

$$\sin(p) + \sin(q) = 2\sin\left(\tfrac{p+q}{2}\right)\cos\left(\tfrac{p-q}{2}\right) \qquad \sin(p) - \sin(q) = 2\cos\left(\tfrac{p+q}{2}\right)\sin\left(\tfrac{p-q}{2}\right)$$

$$\tan(p) + \tan(q) = \frac{\sin(p+q)}{\cos(p)\,\cos(q)} \qquad \tan(p) - \tan(q) = \frac{\sin(p-q)}{\cos(p)\,\cos(q)}$$

**Formules de l'arc moitié.**

En posant $t = \tan\left(\frac{a}{2}\right)$ on a :

$$\sin(a) = \frac{2\,t}{1+t^2}, \quad \cos(a) = \frac{1-t^2}{1+t^2}, \quad \tan(a) = \frac{2t}{1-t^2}$$

**Relations entre les rapports trigonométriques d'un même arc.**

| $\cos^2(a) + \sin^2(a) = 1$ | |
|---|---|
| $\tan(a) = \frac{\sin(a)}{\cos(a)}$ | $\cot(a) = \frac{\cos(a)}{\sin(a)}$ |
| $1 + \tan^2(a) = \frac{1}{\cos^2(a)}$ | $1 + \cot^2(a) = \frac{1}{\sin^2(a)}$ |

**Fonctions circulaires réciproques.**

| $\arccos(x) + \arcsin(x) = \frac{\pi}{2}$ | $\arccos(-x) = \pi - \arccos(x)$ | $\arctan(x) + \operatorname{arccotan}(x) = \frac{\pi}{2}$ |
|---|---|---|

$$\arctan(a) + \arctan(b) = \begin{cases} \arctan\left(\dfrac{a+b}{1-a\,b}\right) & \text{si } a\,b < 1 \\ \dfrac{\pi}{2}\,\operatorname{sign}(a) & \text{si } a\,b = 1 \\ \arctan\left(\dfrac{a+b}{1-a\,b}\right) + \pi\,\operatorname{sign}(a) & \text{si } a\,b > 1 \end{cases}$$

### 7.3.2 Trigonométrie hyperbolique.

**Formules d'addition et de différence des arcs.**

| $\cosh(x) = \frac{e^x+e^{-x}}{2}$ | $\sinh(x) = \frac{e^x-e^{-x}}{2}$ |
|---|---|
| $\cosh(a-b) = \cosh(a)\,\cosh(b) - \sinh(a)\,\sinh(b)$ | $\sinh(a-b) = \sinh(a)\,\cosh(b) - \cosh(a)\,\sinh(b)$ |
| $\cosh(a+b) = \cosh(a)\,\cosh(b) + \sinh(a)\,\sinh(b)$ | $\sinh(a+b) = \sinh(a)\,\cosh(b) + \cosh(a)\,\sinh(b)$ |
| $\cosh(2\,a) = \begin{cases} \cosh^2(a) + \sinh^2(a) \\ 2\,\cosh^2(a) - 1 \\ 1 + 2\,\sinh^2(a) \end{cases}$ | $\sinh(2\,a) = 2\,\sinh(a)\,\cosh(a)$ |

| $\tanh(x) = \frac{\sinh(x)}{\cosh(x)}$ | |
|---|---|
| $\tanh(a-b) = \frac{\tanh(a)+\tanh(b)}{1-\tanh(a)\,\tanh(b)}$<br>$\tanh(a+b) = \frac{\tanh(a)+\tanh(b)}{1+\tanh(a)\,\tanh(b)}$ | $\tanh(2\,a) = \frac{2\,\tanh(a)}{1+\tanh^2(a)}$ |

**Formules de Simpson.**

$$\begin{array}{c|c} \cosh(p)+\cosh(q) = 2\cosh\left(\frac{p+q}{2}\right)\cosh\left(\frac{p-q}{2}\right) & \cosh(p)-\cosh(q) = 2\sinh\left(\frac{p+q}{2}\right)\sinh\left(\frac{p-q}{2}\right) \\ \sinh(p)+\sinh(q) = 2\sinh\left(\frac{p+q}{2}\right)\cosh\left(\frac{p-q}{2}\right) & \sinh(p)-\sinh(q) = 2\cosh\left(\frac{p+q}{2}\right)\sinh\left(\frac{p-q}{2}\right) \\ \tanh(p)+\tanh(q) = \frac{\sinh(p+q)}{\cosh(p)\,\cosh(q)} & \tanh(p)-\tanh(q) = \frac{\sinh(p-q)}{\cosh(p)\,\cosh(q)} \end{array}$$

**Formules de l'arc moitié.**

En posant $t = \tanh\left(\frac{a}{2}\right)$ on a :

$$\sinh(a) = \frac{2\,t}{1-t^2}, \quad \cosh(a) = \frac{1+t^2}{1-t^2}, \quad \tanh(a) = \frac{2t}{1+t^2}$$

**Relations entre les rapports trigonométriques d'un même arc.**

| $\cosh^2(a) - \sinh^2(a) = 1$ | |
|---|---|
| $\tanh(a) = \frac{\sinh(a)}{\cosh(a)}$<br>$1 - \tanh^2(a) = \frac{1}{\cosh^2(a)}$ | $\mathrm{cotanh}(a) = \frac{\cosh(a)}{\sinh(a)}$<br>$1 - \mathrm{cotanh}^2(a) = -\frac{1}{\sinh^2(a)}$ |

**Fonctions hyperboliques réciproques.**

$$\mathrm{argcosh}(x) = \ln\left(x + \sqrt{x^2-1}\right), \quad \mathrm{argsinh}(x) = \ln\left(x + \sqrt{x^2+1}\right), \quad \mathrm{argtanh}(x) = \ln\left(\frac{1+x}{1-x}\right)$$

## 7.4 Dérivées des fonctions usuelles.

| Fonction $f(x)$ | Dérivée $f'(x)$ | Domaine de définition | Domaine de dérivabilité |
|---|---|---|---|
| $k$ | $0$ | $\mathbb{R}$ | |
| $x^\alpha$ | $\alpha\, x^{\alpha-1}$ | $\mathbb{R}$ si $\alpha > 0$ ou $\mathbb{R}^*$ si $\alpha < 0$ | |
| $\exp(x)$ | $\exp(x)$ | $\mathbb{R}$ | |
| $a^x$ | $\ln(a)\, a^x$ | $\mathbb{R}$ avec $a > 0$ | |
| $\ln(\lvert x\rvert)$ | $\frac{1}{x}$ | $\mathbb{R}_+^*$ | |
| $\cos(x)$ | $-\sin(x)$ | $\mathbb{R}$ | |
| $\sin(x)$ | $\cos(x)$ | $\mathbb{R}$ | |
| $\tan(x)$ | $\frac{1}{\cos^2(x)} = 1 + \tan^2(x)$ | $\mathbb{R}\backslash\{\frac{\pi}{2} + \pi\,\mathbb{Z}\}$ | |
| $\cot(x)$ | $-\frac{1}{\sin^2(x)} = -\left(1 + \cot^2(x)\right)$ | $\mathbb{R}\backslash\{\pi\,\mathbb{Z}\}$ | |
| $\arccos(x)$ | $-\frac{1}{\sqrt{1-x^2}}$ | $[-1,1]$ | $]-1,1[$ |
| $\arcsin(x)$ | $\frac{1}{\sqrt{1-x^2}}$ | $[-1,1]$ | $]-1,1[$ |
| $\arctan(x)$ | $\frac{1}{1+x^2}$ | $\mathbb{R}$ | |
| $\cosh(x)$ | $\sinh(x)$ | $\mathbb{R}$ | |
| $\sinh(x)$ | $\cosh(x)$ | $\mathbb{R}$ | |
| $\tanh(x)$ | $\frac{1}{\cosh^2(x)} = 1 - \tanh^2(x)$ | $\mathbb{R}$ | |
| $\operatorname{cotanh}(x)$ | $-\frac{1}{\sinh^2(x)} = 1 - \operatorname{cotanh}^2(x)$ | $\mathbb{R}^*$ | |
| $\operatorname{argcosh}(x)$ | $\frac{1}{\sqrt{1+x^2}}$ | $\mathbb{R}$ | |
| $\operatorname{argsinh}(x)$ | $\frac{1}{\sqrt{1-x^2}}$ | $[-1,1]$ | $]-1,1[$ |
| $\operatorname{argtanh}(x)$ | $\frac{1}{1-x^2}$ | $]-1,1[$ | |

## 7.5 Primitives des fonctions usuelles.

| Fonction $f(x)$ | Primitive $F(x)$ (-constante) | Domaine de définition | Domaine d'intégrabilité |
|---|---|---|---|
| k | $k\,x$ | $\mathbb{R}$ | |
| $x^\alpha$ | $\frac{x^{\alpha+1}}{\alpha+1}$ | $\mathbb{R}$ si $\alpha > 0$ ou $\mathbb{R}^*$ si $\alpha \in \mathbb{R}_-^* \backslash \{-1\}$ | |
| $\frac{1}{x}$ | $\ln(\lvert x\rvert)$ | $\mathbb{R}^*$ | |
| $\ln(\lvert x\rvert)$ | $x\ln(\lvert x\rvert) - x$ | $\mathbb{R}^*$ | |
| $\exp(x)$ | $\exp(x)$ | $\mathbb{R}$ | |
| $a^x$ | $\frac{a^x}{\ln(a)}$ | $\mathbb{R}$ avec $a \in \mathbb{R}_+^*$ | $\mathbb{R}$ avec $a \in \mathbb{R}_+^* \backslash \{1\}$ |
| $\cos(x)$ | $\sin(x)$ | $\mathbb{R}$ | |
| $\sin(x)$ | $-\cos(x)$ | $\mathbb{R}$ | |
| $\tan(x)$ | $-\ln\lvert\cos(x)\rvert$ | $\mathbb{R}\backslash\{\frac{\pi}{2} + \pi\,\mathbb{Z}\}$ | |
| $\cot(x)$ | $\ln\lvert\sin(x)\rvert$ | $\mathbb{R}\backslash\{\pi\,\mathbb{Z}\}$ | |
| $\frac{1}{\cos(x)}$ | $\ln\lvert\tan(\frac{x}{2} + \frac{\pi}{4})\rvert$ | $\mathbb{R}\backslash\{\frac{\pi}{2} + \pi\,\mathbb{Z}\}$ | |
| $\frac{1}{\sin(x)}$ | $\ln\lvert\tan(\frac{x}{2})\rvert$ | $\mathbb{R}\backslash\{\pi\,\mathbb{Z}\}$ | |
| $\frac{1}{\cos^2(x)} = 1 + \tan^2(x)$ | $\tan(x)$ | $\mathbb{R}\backslash\{\frac{\pi}{2} + \pi\,\mathbb{Z}\}$ | |
| $\frac{1}{\sin^2(x)} = 1 + \cot^2(x)$ | $-\cot(x)$ | $\mathbb{R}\backslash\{\pi\,\mathbb{Z}\}$ | |
| $\sinh(x)$ | $\cosh(x)$ | $\mathbb{R}$ | |
| $\cosh(x)$ | $\sinh(x)$ | $\mathbb{R}$ | |
| $\tanh(x)$ | $\ln(\cosh(x))$ | $\mathbb{R}$ | |
| $\mathrm{cotanh}(x)$ | $\ln\lvert\sinh(x)\rvert$ | $\mathbb{R}$ | |
| $\frac{1}{\cosh(x)}$ | $2\,\arctan\left(\exp\left(x\right)\right)$ | $\mathbb{R}$ | |
| $\frac{1}{\sinh(x)}$ | $-2\,\arctan\left(\exp\left(x\right)\right)$ | $\mathbb{R}^*$ | $\mathbb{R}$ |
| $\frac{1}{\cosh^2(x)} = 1 - \tanh^2(x)$ | $\tanh(x)$ | $\mathbb{R}$ | |
| $\frac{1}{\sinh^2(x)} = \mathrm{cotanh}^2(x) - 1$ | $-\mathrm{cotanh}(x)$ | $\mathbb{R}^*$ | |
| $\frac{1}{1+x^2}$ | $\arctan(x)$ | $\mathbb{R}$ | |
| $\frac{1}{1-x^2}$ | $\mathrm{argtanh}(x) = \frac{1}{2}\ln\left(\frac{1+x}{1-x}\right)$ | $\mathbb{R}\backslash\{1\}$ | $]-1,1[$ |
| $\frac{1}{\sqrt{1-x^2}}$ | $\arcsin(x)$ | $]-1,1[$ | $[-1,1]$ |
| $\frac{1}{\sqrt{x^2-1}}$ | $\mathrm{argcosh}(x) = \ln\left(x + \sqrt{x^2-1}\right)$ | $\mathbb{R}\backslash\,[-1,1]$ | $\mathbb{R}$ |
| $\frac{1}{\sqrt{1+x^2}}$ | $\mathrm{argsinh}(x) = \ln\left(x + \sqrt{x^2+1}\right)$ | $\mathbb{R}$ | |

## 7.6 Développements limités usuels.

Notations :

— $D(a,r)$ représente la boule fermée de $\mathbb{C}$ centrée en $a$ et de rayon $r$
— $B_n$ est le $n^{\text{ième}}$ nombre de Bernoulli, défini par :

$$B_n = \sum_{k=0}^{n} \frac{1}{k+1} \sum_{j=0}^{k} (-1)^j \binom{k}{j} j^n \qquad , \qquad B_{2n} = (-1)^{n+1} \, |B_{2n}|$$

### 7.6.1 Binômes.

$$\forall x \in D(0,1),\ \frac{1}{1-x} = \sum_{n=0}^{+\infty} x^n$$

$$\forall x \in ]-1,1[,\ \forall \alpha \notin \mathbb{N},\ (1+x)^\alpha = 1 + \sum_{n=1}^{+\infty} \frac{\alpha\,(\alpha-1)\dots(\alpha-n+1)}{n!}\, x^n$$

$$\forall x \in \mathbb{R},\ \forall \alpha \in \mathbb{N},\ (1+x)^\alpha = 1 + \sum_{n=1}^{\alpha} \frac{\alpha\,(\alpha-1)\dots(\alpha-n+1)}{n!}\, x^n = \sum_{n=0}^{\alpha} \binom{\alpha}{n} x^n$$

### 7.6.2 Fonctions exponentielles et logarithmiques.

$$\forall x \in \mathbb{C},\ \exp(x) = \sum_{n=0}^{+\infty} \frac{x^n}{n!}$$

$$\forall x \in \mathbb{C},\ a^x = e^{x \ln a} = \sum_{n=0}^{+\infty} \frac{(\ln a)^n}{n!}\, x^n$$

$$\forall x \in ]-1,1],\ \ln(1-x) = -\sum_{n=1}^{+\infty} \frac{x^n}{n}$$

$$\forall x \in ]-1,1],\ \ln(1+x) = -\sum_{n=1}^{+\infty} \frac{(-x)^n}{n}$$

### 7.6.3 Fonctions trigonométriques et trigonométriques inverses.

$$\forall x \in \mathbb{C},\ \sin(x) = \sum_{n=0}^{+\infty} (-1)^n \frac{x^{2n+1}}{(2\,n+1)!}$$

$$\forall x \in \mathbb{C},\ \cos(x) = \sum_{n=0}^{+\infty} (-1)^n \frac{x^{2n}}{(2\,n)!}$$

$$\forall x \in \left]-\frac{\pi}{2}, \frac{\pi}{2}\right[,\ \tan(x) = \sum_{n=1}^{+\infty} \frac{2^{2n}(2^{2n}-1)}{(2n)!}\ |B_{2n}|\ x^{2n-1}$$

$$\forall x \in D(0,\pi) \setminus \{0\}, \cot(x) = \frac{1}{x} - \sum_{n=1}^{+\infty} \frac{2^{2n}}{(2n)!}\ |B_{2n}|\ x^{2n-1}$$

$$\forall x \in ]-1,1[,\ \arcsin(x) = \sum_{n=0}^{+\infty} \frac{(2n)!}{(n!2^n)^2}\ \frac{x^{2n+1}}{2n+1}$$

$$\forall x \in ]-1,1[,\ \arccos(x) = \frac{\pi}{2} - \arcsin x = \frac{\pi}{2} + \sum_{n=0}^{+\infty} \frac{(2n)!}{(n!2^n)^2}\ \frac{x^{2n+1}}{2n+1}$$

$$\forall x \in ]-1,1[, \arctan(x) = \sum_{n=0}^{+\infty} (-1)^n\ \frac{x^{2\,n+1}}{2\,n+1}$$

$$\forall x \in ]-1,1[, \operatorname{arccot}(x) = \frac{\pi}{2} - \arctan(x) = \frac{\pi}{2} - \sum_{n=0}^{+\infty} (-1)^n\ \frac{x^{2\,n+1}}{2\,n+1}$$

### 7.6.4 Fonctions hyperboliques et hyperboliques inverses.

$$\forall x \in \mathbb{C},\ \operatorname{sh}(x) = \sum_{n=0}^{+\infty} \frac{x^{2n+1}}{(2\,n+1)!}$$

$$\forall x \in \mathbb{C},\ \operatorname{ch}(x) = \sum_{n=0}^{+\infty} \frac{x^{2n}}{(2\,n)!}$$

$$\forall x \in \left]-\frac{\pi}{2}, \frac{\pi}{2}\right[, \operatorname{th}(x) = \sum_{n=1}^{+\infty} \frac{2^{2n}(2^{2n}-1)}{(2n)!} B_{2n} x^{2n-1}$$

$$\forall x \in ]0,\pi[, \coth(x) = \frac{1}{x} + \sum_{n=1}^{+\infty} \frac{2^{2n}}{(2n)!} B_{2n} x^{2n-1}$$

$$\forall x \in ]-1,1[,\ \operatorname{argsh}(x) = x + \sum_{n=1}^{+\infty} (-1)^n \frac{(2\,n)!}{(n!\,2^n)^2}\ \frac{x^{2n+1}}{2\,n+1}$$

$$\forall x \in ]-1,1[,\ \operatorname{argth}(x) = \sum_{n=0}^{+\infty} \frac{x^{2\,n+1}}{2\,n+1}$$

# Bibliographie


Allaire, G. (2012). *Analyse numérique et Optimisation.* Ellipses, 2ème édition.

Bastien, J. et Martin, J.-N. (2003). *Introduction à l'analyse numérique. Applications sous Matlab.* Dunod.

Crouzeix, M. et Mignot, A.-L. (1984). *Analyse numérique des équations différentielles.* Masson.

Durand, E. (1971). *Solutions numériques des équations algébriques, tomes I et II.* Masson, Paris.

Fortin, A. (2016). *Analyse numérique pour ingénieurs.* Presse internationales Polytechnique, Montréal, 5ème édition.

Gourdon, X. (2008). *Les maths en tête, Analyse.* Ellipses, Paris, 2ème édition.

Lakrib, M. (2017). *Analyse numérique - Cours et exercices résolus.* Ellipses.

Lascaux, P. et Théodor, R. (2000). *Analyse numérique matricielle appliquée à l'art de l'ingénieur, tomes I (méthodes directes) et II (méthodes itératives).* Masson, Paris.

Nougier, J.-P. (1987). *Méthode de calcul numérique.* Masson, Paris.

Popier, A. et Winterberger, O. (2006). *Equations différentielles, résumé de cours.*

Rotella, F. et Zambettakis, I. (2015). *Traitement des systèmes linéaires.* Ellipses.

Sibony, M. (1997a). *Analyse numérique, Volume 2. Approximations et équations différentielles.* Hermann.

Sibony, M. (1997b). *Analyse numérique, Volume 3. Itérations et approximations.* Hermann.

Sibony, M. et Mardon, J.-C. (1997). *Analyse numérique, Volume 1. Systèmes linéaires et non linéaires.* Hermann.

Voedts, J. (2001). *Cours de Mathematiques MP-MP*.* Ellipses.

Yger, A. (2015). *Calcul scientifique et symbolique - Éléments de cours illustrés par des TP guidés sous les environnements Maple, MATLAB ou Scilab, SAGE sous Python.* Ellipses.