Résolution numérique de problèmes nonlinéaires

M1 Calcul Scientifique 2021/2022

Andrea Natale

Table des matières

Ces notes de cours sont basées sur des notes précédentes de Benoit Merlet, ainsi que sur ces notes de cours de Quentin Mérigot, et les livres “Iterative Methods for Linear and Nonlinear Equations” de C.T. Kelly et “Scientific Computing with Ordinary Differential Equations” de Peter Deuflhard et Folkmar Bornemann.

1 Résolution numérique d'EDOs

Dans ce chapitre, nous étudions la résolution numérique du *problème de Cauchy* suivant : étant donnée une fonction $f : \Omega \subset \mathbb{R}\times\mathbb{R}^d \to \mathbb{R}^d$, on cherche une fonction $y : J \to \mathbb{R}^d$, où $J \subset \mathbb{R}$ est un intervalle contenant $t_0 \in J$, qui vérifie

$$\begin{cases} y'(t) = f(t, y(t)) \quad \forall t \in J\,, \\ y(t_0) = y_0\,, \end{cases} \tag{1.0.1}$$

Selon les propriétés et la régularité de f, ce problème peut, dans certains cas, ne pas admettre de solution, ou bien les solutions peuvent être définies uniquement localement. Même lorsqu'une unique solution existe, celle-ci peut ne pas admettre d'expression analytique, ce qui oblige à construire des approximations numériques. En particulier, étant donné $t_0 < t_1 < \ldots < t_n < \ldots$, une solution numérique du problème de Cauchy (1.0.1) est une suite $(y_n)_{n\geq 0}$ telle que y_n est une approximation de $y(t_n)$, où y est une solution.

1.1 Quelques rappels sur les EDOs

Nous commençons par rappeler quelques propriétés du problème de Cauchy (1.0.1) ainsi que des résultats d'existence et d'unicité des solutions. Tout d'abord, il est facile de montrer l'équivalence suivante (exercice) : soit $f \in C(\Omega;\mathbb{R}^d)$, $J \subset \mathbb{R}$ un intervalle contenant $t_0 \in J$ et $y \in C(J;\mathbb{R}^d)$. Alors :

$$\begin{cases} y \in C^1(J,\mathbb{R}^d)\,, \\ y'(t) = f(t, y(t)) \quad \forall t \in J\,, \\ y(t_0) = y_0\,, \end{cases} \quad \Leftrightarrow \quad y(t) = y_0 + \int_{t_0}^{t} f(s, y(s))\,\mathrm{d}s \quad \forall\, t \in J\,. \tag{1.1.1}$$

Cette équivalence nous sera utile, car elle permet de transformer le problème de la résolution numérique de (1.0.1) en un problème de quadrature.

Si la fonction f ne dépend pas de t, le système est dit *autonome* ; en revanche, si f dépend de t de manière non triviale, le système est dit *non autonome*.

Exemple (Équations de Lotka-Volterra). Les équations de Lotka-Volterra, qui décrivent l'évolution des populations de proies y_1 et de prédateurs y_2 en

interaction, constituent le système autonome suivant :

$$\begin{cases} y_1' = \alpha y_1 - \beta y_1 y_2 \,, \\ y_2' = \delta y_1 y_2 - \gamma y_2 \,. \end{cases}$$

où $\alpha, \beta, \delta, \gamma \in [0, \infty)$.

On peut toujours transformer un problème de Cauchy non autonome en un problème autonome en ajoutant une dimension. Plus précisément, $y : J \to \mathbb{R}^d$ résout le problème de Cauchy (1.1.1) si et seulement si $z : t \in J \mapsto (t, y(t)) \in \mathbb{R}^{d+1}$ résout le problème autonome suivant :

$$\begin{cases} z'(t) = g(z(t)) \quad \forall t \in J \,, \\ z(t_0) = z_0 \,, \end{cases} \tag{1.1.2}$$

où $g((t, y)) \coloneqq (1, f(t, y))$ pour tout $(t, y) \in \Omega \subset \mathbb{R} \times \mathbb{R}^d$ et $z_0 \coloneqq (t_0, y_0) \in \Omega \subset \mathbb{R} \times \mathbb{R}^d$.

Une solution $y \in C^1(J; \mathbb{R}^d)$ du problème de Cauchy (1.1.1) peut être prolongée s'il existe un intervalle $(t_-, t_+) \supset J$ et une solution $\tilde{y} \in C^1((t_-, t_+); \mathbb{R}^d)$ telle que $\tilde{y}|_J = y$. On dit que y peut être prolongée jusqu'au bord de Ω si l'une des conditions suivantes est vérifiée pour $t_* \in \{t_-, t_+\}$:

1. $|t_*| = \infty$;
2. $|t_*| < \infty$ et $\lim_{t \to t_*} \operatorname{dist}(y(t), \partial\Omega) = 0$;
3. $|t_*| < \infty$ et $\lim_{t \to t_*} \|y(t)\| = \infty$.

L'existence d'une solution au problème de Cauchy est assurée sous des conditions très faibles sur f. En particulier, on a le résultat suivant :

Théorème 1.1.1 (Peano)**.** *Soit $f \in C(\Omega; \mathbb{R}^d)$. Alors, le problème de Cauchy (1.0.1) avec $(t_0, y_0) \in \Omega$ admet au moins une solution $y \in C^1(J; \mathbb{R}^d)$, et cette solution peut être prolongée jusqu'au bord de Ω.*

Pour assurer l'unicité des solutions, une régularité supplémentaire de f est nécessaire (trouver un contre-exemple). La condition clé dans ce contexte est celle de lipschitzianité.

Rappel: Fonctions lipschitziennes

Définition 1.1.2 (Fonction lipschitzienne)**.** Soit $\varphi : U \subset \mathbb{R}^n \to \mathbb{R}^m$. On dit que φ est une fonction *lipschitzienne* avec une constante de Lipschitz L (ou L-lipschitzienne) si et seulement si

$$\|\varphi(x) - \varphi(y)\| \leq L\|x - y\| \,, \quad \forall x, y \in U \,.$$

On dit que φ est *localement lipschitzienne* si, pour tout $x \in U$, il existe un

voisinage $V \subset U$ de x tel que φ est lipschitzienne sur V.

Exercice 1.1.3. Montrez que si $\varphi : U \subset \mathbb{R}^n \to \mathbb{R}^m$ est différentiable sur un ouvert convexe U, et $|||D\varphi(x)||| \leq L$ pour tout $x \in U$, où $|||\cdot|||$ est la norme induite sur $\mathcal{M}_{mn}(\mathbb{R})$ par une norme $\|\cdot\|$ sur $\mathbb{R}^d$, alors φ est L-lipschitzienne par rapport à cette norme.

Pour assurer l'unicité d'une solution globale sur un ouvert Ω, il suffit que f soit localement lipschitzienne par rapport à sa deuxième variable, c'est-à-dire qu'il existe, pour tout $(t_0, y_0) \in \Omega$, un voisinage $\Omega' \subset \Omega$ de (t_0, y_0) et une constante $L > 0$ tels que

$$\|f(t, y_1) - f(t, y_2)\| \leq L\|y_1 - y_2\|\,, \quad \forall (t, y_1), (t, y_2) \in \Omega'\,.$$

On obtient alors le théorème suivant :

Théorème 1.1.4 (Picard-Lindelöf)**.** *Soit $f : \Omega \subset \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^d$ une fonction continue et localement lipschitzienne par rapport à sa deuxième variable. Alors, il existe un intervalle J tel que $t_0 \in J$ et une fonction $y \in C^1(J; \mathbb{R}^d)$ qui résout le problème de Cauchy (1.0.1). Cette solution peut être prolongée jusqu'au bord de Ω et elle est unique.*

Rappel: Flot d'un champ de vecteurs Étant données des conditions initiales $(t_0, y_0) \in \Omega$, dans les hypothèses du Théorème 1.1.4, on peut construire une solution $y \in C^1(J_{max}(t_0, y_0); \mathbb{R}^d)$ (unique) définie sur un intervalle maximal $J_{max}(t_0, y_0)$ (qui dépend de la condition initiale). Soit $U \subset \mathbb{R}^d$ un ouvert, $J \subset \mathbb{R}$ un intervalle et $\varepsilon > 0$, en supposant que pour toute condition initiale $(t_0, y_0) \in J \times U$, $(t_0 - \varepsilon, t_0 + \varepsilon) \subset J_{max}(t_0, y_0)$, on peut définir pour tout $\tau \in (-\varepsilon, \varepsilon)$ une application $\Phi_\tau : J \times U \to U$, qu'on appelle le *flot* du champ de vecteurs f, par

$$\Phi_\tau(t_0, y_0) \coloneqq y(t_0 + \tau)$$

où y résout le problème de Cauchy associé avec f et les conditions initiales $y(t_0) = y_0$.

Exercice 1.1.5. Montrer que si $f : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^d$ vérifie les hypothèses du Théorème 1.1.4, et en plus il existe une constante $M > 0$ telle que :

$$\|f(t, y)\| \leq M(1 + \|y\|) \quad \forall\, (t, y) \in \mathbb{R} \times \mathbb{R}^d\,,$$

alors $J_{max}(t_0, y_0) = \mathbb{R}$ pour toute condition initiale $(t_0, y_0) \in \mathbb{R} \times \mathbb{R}$ et le flot de f est bien défini pour tout $\tau \in \mathbb{R}$.

Dans le cas d'un système autonome où $f : U \subset \mathbb{R}^d \to \mathbb{R}^d$, le choix du temps initial t_0 ne change pas la solution au temps $t_0 + \tau$, où $\tau \in \mathbb{R}$ est suffisamment petit. Dans ce cas, on omet l'argument t_0 et on appelle le *flot* de f l'application

$\Phi_\tau : U \to U$ définie par

$$\Phi_\tau(y_0) \coloneqq y(\tau)$$

où y résout le problème de Cauchy associé avec f et les conditions initiales $y(0) = y_0$.

1.2 Rappels sur la méthode d'Euler explicite

Pour simplicité, on considère ici le cas où $f : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^d$ et on suppose que f vérifie les hypothèses du Théorème 1.1.4. La méthode d'Euler est basée sur l'observation suivante : étant donnée une solution $y \in C^1([t_0, T]; \mathbb{R}^d)$ du problème de Cauchy (1.0.1), pour $\tau > 0$ suffisamment petit,

$$y(t_0 + \tau) = y(t_0) + y'(t_0)\tau + o(\tau) = y(t_0) + f(t_0, y_0)\tau + o(\tau)\,, \qquad (1.2.1)$$

où $o(\tau)$ désigne une fonction telle que $\|o(\tau)\|/\tau \to 0$ lorsque $\tau \to 0$ et $\tau \neq 0$.

Soit $t_0 < t_1 < \ldots < t_N = T$ une décomposition de l'intervalle $[t_0, T]$ et $\tau_n \coloneqq t_{n+1} - t_n$. Le développement limité de y suggère de construire une solution numérique $(y_n)_{n=0}^N$ associée à cette décomposition, successivement pour $n = 0, \ldots, N-1$, grâce au schéma suivant dit d'*Euler explicite* :

$$y_{n+1} = \tilde{\Phi}_{\tau_n}(t_n, y_n)\,, \quad \tilde{\Phi}_\tau(t, y) \coloneqq y + f(t, y)\tau$$

pour tout $(t, y, \tau) \in \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}$.

En vue de la formulation intégrale (1.1.1), la méthode d'Euler explicite peut être aussi interprétée comme le résultat de l'application de la formule de quadrature des rectangles à gauche pour évaluer l'intégrale

$$\int_{t_n}^{t_{n+1}} f(s, y(s))\,\mathrm{d}s \approx f(t_n, y(t_n))\tau_n\,. \qquad (1.2.2)$$

Ce point de vue justifie aussi la définition suivante de méthode à un pas, qui généralise la méthode d'Euler explicite, et où on remplace l'intégrale (1.2.2) par une fonction générale.

Définition 1.2.1 (Méthode à un pas)**.** Une *méthode à un pas* est une méthode pour laquelle y_{n+1} dépend seulement de t_n, y_n et du pas τ_n. Elle est définie par une application $\Psi : \mathbb{R} \times \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^d$, et on pose, pour $0 \leq n \leq N-1$,

$$y_{n+1} = y_n + \tau_n \Psi(t_n, y_n, \tau_n)\,.$$

Soit Φ le flot du champ de vecteur f. Pour étudier la convergence de la méthode d'Euler explicite, on observe que :

$$\begin{aligned}\varepsilon_{n+1} \coloneqq \|y(t_{n+1}) - y_{n+1}\| \leq \|\Phi_{\tau_n}(t_n, y(t_n)) - \tilde{\Phi}_{\tau_n}(t_n, y(t_n))\| \\ + \|\tilde{\Phi}_{\tau_n}(t_n, y(t_n)) - \tilde{\Phi}_{\tau_n}(t_n, y_n)\|. \quad (1.2.3)\end{aligned}$$

Ceci montre que l'erreur au temps t_{n+1}, dénotée ε_{n+1}, peut s'exprimer comme la somme de deux contributions : un premier terme qui dépend de la consistance de l'approximation $\tilde{\Phi}$ du flot exact Φ, et un deuxième terme qui mesure comment l'erreur se propage si on applique la méthode numérique avec deux conditions initiales distinctes.

Supposons que $y \in C^2([t_0, T]; \mathbb{R}^d)$. Par la formule de Taylor-Lagrange, l'erreur de consistance au temps t_{n+1} pour la méthode d'Euler explicite est donnée par

$$\begin{aligned}\varepsilon_{n+1}^{cons} &\coloneqq \|\Phi_{\tau_n}(t_n, y(t_n)) - \tilde{\Phi}_{\tau_n}(t_n, y(t_n))\| \\ &= \|y(t_{n+1}) - y(t_n) - f(t_n, y(t_n))\tau_n\| = \frac{\tau_n^2}{2}\|y''(\xi)\|\end{aligned}$$

pour un certain $\xi \in (t_n, t_{n+1})$.

L'erreur propagée au temps t_{n+1} est :

$$\varepsilon_{n+1}^{prop} \leq \tau_n \|y(t_n) - y_n\|.$$

La combinaison des deux erreurs donne la borne finale de l'erreur de la méthode d'Euler explicite :

$$\varepsilon_{n+1} \leq \frac{\tau_n^2}{2}\|y''(\xi)\| + \tau_n\|y(t_n) - y_n\|.$$

1.2.1 Formulation générale des méthodes RK

Si l'on souhaite utiliser une formule de quadrature plus précise pour obtenir une meilleure approximation, en général, on aura besoin de $s \geq 1$ évaluations du terme à droite du problème de Cauchy à des instants différents. Autrement dit, pour chaque pas de temps, il faut estimer

$$f(t_n + c_i\tau, y(t_n + c_i\tau)),$$

avec des coefficients $c_i \in \mathbb{R}$, pour $1 \leq i \leq s$. Dans les méthodes de Runge-Kutta explicites, ces évaluations sont réalisées de manière itérative, en estimant $y(t_n + c_i\tau)$ à l'aide des approximations déjà calculées de $f(t_n +$

$c_j\tau, y(t_n + c_j\tau))$ pour $1 \leq j < i$. Cela conduit à la méthode de Runge-Kutta explicite à s stades : pour $n = 0, \ldots, N-1$,

$$\begin{aligned} k_1 &= f(t_n, y_n), \\ k_2 &= f(t_n + c_2\tau, y_n + a_{21}k_1\tau), \\ &\vdots \\ k_s &= f(t_n + c_s\tau, y_n + \sum_{j=1}^{s-1} a_{sj}k_j\tau), \\ y_{n+1} &= y_n + \sum_{i=1}^{s} b_i k_i \tau. \end{aligned}$$

Les coefficients b_i et c_i sont choisis respectivement comme les nœuds et les poids d'une formule de quadrature, de sorte que si f ne dépend pas de y,

$$\int_{t_n}^{t_{n+1}} f(t)\,\mathrm{d}t \approx \sum_{i=1}^{s} b_i f(t_n + c_i\tau).$$

Pour les méthodes de Runge-Kutta explicites, on demande généralement que

$$\sum_{j=1}^{i-1} a_{ij} = c_i \quad \text{pour } 2 \leq i \leq s. \tag{1.2.4}$$

Ces conditions sont suffisantes (mais pas nécessaires) pour que la méthode obtenue soit au moins d'ordre un. En effet, les conditions (1.2.4) impliquent que chaque k_i est une approximation d'ordre deux de $y'(t_n + c_i\tau)$, c'est-à-dire que si y et f sont suffisamment régulières [1]

$$k_i = f(t_n + c_i\tau, y(t_n + c_i\tau)) + \mathcal{O}(\tau^2),$$

et ainsi la formule de quadrature est justifiée.

Dans les méthodes de Runge-Kutta implicites, une ou plusieurs étapes de la méthode ne peuvent pas être formulées de manière explicite pour k_i, ce qui implique qu'il faut résoudre un système d'équations (non linéaires) pour avancer la méthode. La formulation générale des méthodes Runge-Kutta est donc la suivante :

1. Cela peut être montré par induction. En particulier, puisque f est Lipschitzienne en y, il suffit de montrer que $y_n + \sum_{j=1}^{i-1} a_{i-1,j}k_j$ est une approximation d'ordre deux de $y(t_n + c_i\tau)$.

Définition 1.2.2. Une méthode de Runge-Kutta à s stades est définie comme suit : pour $0 \leq n \leq N-1$,

$$k_i = f(t_n + c_i\tau, y_n + \sum_{j=1}^{s} a_{ij}k_j\tau), \quad 1 \leq i \leq s,$$

$$y_{n+1} = y_n + \sum_{i=1}^{s} b_i k_i \tau.$$

Les coefficients a_{ij}, b_i, c_i, qui définissent la méthode, sont organisés dans le *tableau de Butcher* :

$$\begin{array}{c|ccc} c_1 & a_{11} & \cdots & a_{1s} \\ \vdots & \vdots & & \vdots \\ c_s & a_{s1} & \cdots & a_{ss} \\ \hline & b_1 & \cdots & b_s \end{array}, \quad \text{où} \quad \sum_{j=1}^{s} a_{ij} = c_i, \ \ 1 \leq i \leq s.$$

La méthode est dite *explicite* si $a_{ij} = 0$ pour $j \geq i$, et *implicite* si ce n'est pas le cas.

Les méthodes de Runge-Kutta sont des méthodes à un pas selon la définition 1.2.1, où la fonction Ψ est définie de manière récursive (et implicitement pour les méthodes implicites). De même, on peut aussi définir une application $\tilde{\Phi}_\tau : (t_n, y_n) \to y_{n+1}$, qui représente le flot discret associé à la méthode.

En suivant le même raisonnement que pour Euler explicite, on arrive au résultat suivant :

Théorème 1.2.3. *Soit $f \in C(\mathbb{R} \times \mathbb{R}^d; \mathbb{R}^d)$ localement Lipschitzienne par rapport à sa deuxième variable, et soit $y \in C^1([t_0, T], \mathbb{R}^d)$ l'unique solution du problème de Cauchy $y' = f(t, y)$ sur l'intervalle $[t_0, T]$ et vérifiant $y(t_0) = y_0$. Soit $(y_n)_{n=0}^N$ la solution numérique de la méthode à un pas : pour $1 \leq n \leq N-1$,*

$$y_{n+1} = \tilde{\Phi}_\tau(t_n, y_n) = y_n + \tau\Psi(t_n, y_n, \tau),$$

et supposons que

1. *il existe des constantes $D > 0$ et $p > 0$, telles que pour tout $0 \leq n \leq N-1$,*

$$\varepsilon_n^{cons} = \|\Phi_\tau(t_n, y(t_n)) - \tilde{\Phi}_\tau(t_n, y(t_n))\| \leq D\tau^{p+1};$$

2. *$\Psi(t_n, \cdot, \tau)$ est L-Lipschitzienne.*

Alors, la méthode converge à l'ordre p : il existe une constante $C > 0$ indépendante de τ, telle que

$$\|y(T) - y_N\| \leq C\tau^p. \tag{1.2.5}$$

1.3 Problèmes raides et analyse de stabilité linéaire

Des estimations d'erreur globale du type (1.2.5) ne peuvent pas être le seul critère pour evaluer une méthode numérique. En effet, même pour des méthodes convergentes, les solutions numériques qu'on obient peuvent être qualitativement très différentes des solution exactes.

Pour illustrer ce phénomène, on considère le problème autonome où on cherche $y : [0, \infty) \to \mathbb{R}$ tel que

$$y' = \lambda y\,, \quad y(0) = y_0\,,$$

où $\lambda < 0$, dont l'unique solution est donnée par $y(t) = y_0 \exp(\lambda t)$. Ce problème est *asymptotiquement stable* par rapport aux données initiales, dans le sens que si on considère la solution $\tilde{y}(t) = \tilde{y}_0 \exp(-\lambda t)$ corréspondant aux conditions initiales perturbés $\tilde{y}(0) = \tilde{y}_0$, on a

$$|\tilde{y}(t) - y(t)| = \exp(\lambda t)|\tilde{y}_0 - y_0| \to 0\,, \quad \text{lorsque} \;\; t \to \infty\,.$$

La méthode d'Euler explicite appliquée à cette équation avec un pas de temps $\tau > 0$, donne la solution discrète

$$y_n = (1 + \lambda\tau)^n y_0\,,$$

pour $n \geq 1$. Puisque $\lambda < 0$, $y(t) \to 0$ lorsque $t \to \infty$. Par contre la solution numérique peut exhiber trois comportements distinctes :

1. si $\tau < 2|\lambda|^{-1}$, $y_n \to 0$ lorsque $n \to \infty$, ce qui est consistante avec la solution exacte ;
2. si $\tau = 2|\lambda|^{-1}$, $y_n = y_0$ pour tout $n \geq 1$;
3. si $\tau > 2|\lambda|^{-1}$, $|y_n| \to \infty$ lorsque $n \to \infty$.

Si $|\lambda| \gg 1$, ceci veut dire que même pour $\tau \ll 1$, si $\tau > |\lambda|^{-1}$, le comportement de la solution discrète est completement différent de celui de la solution exacte, au moins après un nombre de pas de temps suffisant. De plus, la solution est instable, puisque si $(\tilde{y}_n)_n$ est la solution numérique associée à des conditions initiales perturbés $\tilde{y}_0$, on a

$$|\tilde{y}_n - y_n| \to \infty\,, \quad \text{lorsque} \;\; t \to \infty\,,$$

même si $\tilde{y}_0$ est arbitrairement proche à y_0. Pour éviter ça on doit imposer une restriction sur le pas de temps ($\tau < |\lambda|^{-1}$) qui est completement indépendante de la convergence de la méthode.

De façon générale, on appelle raides ("stiff") les problèmes pour lesquelles on doit imposer des "restrictions contraignantes" sur le pas de temps d'une certaine méthode numérique pour obtenir une "bonne approximation" de la solution. La définition est nécessairement très imprécise, puisque la défintion de ce que c'est un bonne approximation, ou même qu'est-ce qui représente une restriction sur le pas de temps acceptable ou pas, dépend fortement de l'application considérée.

1.3.1 Stabilité linéaire

Soit $A \in \mathcal{M}_d(\mathbb{C})$ et considèrons le problème de Cauchy, pour $y : [0, \infty) \to \mathbb{C}^d$,

$$y'(t) = Ay(t)\,, \quad y(0) = y_0\,. \tag{1.3.1}$$

Ce problème admet une unique solution qui peut être exprimé comme suit :

$$y(t) = \exp(tA)y_0\,, \quad \text{où} \quad \exp(tA) \coloneqq \sum_{n=0}^{\infty} \frac{(tA)^n}{n!}\,. \tag{1.3.2}$$

De façon équivalente, on peut dire que le flot du champs de vécteur $f(y) = Ay$ est une application linéare $\Phi_\tau \in \mathcal{L}(\mathbb{C}^d, \mathbb{C}^d)$ et on peut l'identifier avec la matrice $\exp(A\tau)$.

Rappel: Réduction de Jordan Soit $A \in \mathcal{M}_d(\mathbb{C})$. Il existe une matrice inversible V telle que $A = V\Lambda V^{-1}$, où Λ est une matrice diagonale par blocs avec k blocs de taille $k_i \times k_i$ et avec $\sum_i k_i = d$,

$$\Lambda = \begin{bmatrix} \Lambda(\lambda_1) & & \\ & \ddots & \\ & & \Lambda(\lambda_k) \end{bmatrix} \text{ et } \Lambda(\lambda_i) = \begin{bmatrix} \lambda_i & 1 & & \\ & \ddots & \ddots & \\ & & \lambda_i & 1 \\ & & & \lambda_i \end{bmatrix} \in \mathcal{M}_{k_i}(\mathbb{C})\,,$$

où $\lambda_i \in \sigma(A)$, le spectre de A, pour tout $1 \leq i \leq k$. Si la muliticiplité algébrique et géométrique de $\lambda \in \sigma(A)$ sont égaux à m alors on aura exactement m blocs de Λ associés à λ de diménsion égale à un.

On rappelle qu' une solution $y : [t_0, \infty) \to \mathbb{C}^d$ du problème $y' = f(t, y)$, $y(t_0) = y_0$ est dite *stable* (au sens de Lyapunov) ssi pour tout $\varepsilon > 0$ il

existe $\delta > 0$ tel que : pour tout $\tilde{y}_0 \in B_\delta(y_0)$ il existe une unique solution $\tilde{y} : [t_0, \infty) \to \mathbb{C}^d$ du même problème avec $y(0) = \tilde{y}_0$ et

$$\|y(t) - \tilde{y}(t)\| \leq \varepsilon \quad \forall\, t \geq t_0;$$

la solution y est dite *asymptotiquement stable* si en plus

$$\|y(t) - \tilde{y}(t)\| \to 0 \quad \text{lorsque} \quad t \to \infty.$$

Dans les cas discret, on utilisera des définitions analogues obtenues en remplacent la variable continue $t \geq t_0$ par l'indice $n \in \mathbb{N}$.

La réduction de Jordan de la matrice A nous permet de characteriser la stabilité asymptotique des solutions du système (1.3.1) comme suit :

Proposition 1.3.1. *L'unique solution $y : [0, \infty) \to \mathbb{C}^d$ du système linéaire $y' = Ay$, $y(0) = y_0$, est asymptotiquement stable ssi $\mathrm{Re}(\lambda) < 0$ pour tout $\lambda \in \sigma(A)$, et en particulier dans ce cas $y(t) \to 0$ lorsque $t \to \infty$ pour tout $y_0 \in \mathbb{C}^d$. Elle est stable ssi : $\mathrm{Re}(\lambda) \leq 0$ pour tout $\lambda \in \sigma(A)$, et si pour tout $\lambda \in \sigma(A)$ avec $\mathrm{Re}(\lambda) = 0$ la multiciplité algébrique et géométrique de λ sont égaux.*

Démonstration. Soit $A = V^{-1}\Lambda V$ la réduction de Jordan de A. Pour toute matrice inversible M, $M^{-1}\exp(tA)M = \exp(tM^{-1}AM)$, donc on peut considerer que le cas $A = \Lambda$. De plus, puisque $\exp(\Lambda) = \mathrm{diag}(\exp(\Lambda(\lambda_1)), \ldots, \exp(\Lambda(\lambda_k)))$ il suffit de considerer le cas

$$A = \lambda I + N\,, \quad \text{où} \quad N = \begin{bmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & 0 & 1 \\ & & & 0 \end{bmatrix} \in \mathcal{M}_d(\mathbb{R})\,. \tag{1.3.3}$$

Puisque $N^d = 0$, on obtient

$$\exp(tA) = \exp(t\lambda) \sum_{i=0}^{d-1} \frac{(tN)^i}{i!} \;\Rightarrow\; |||\exp(tA)||| \leq |\exp(t\lambda)| p(t)$$

où $p(t) = \sum_{i=0}^{d-1} \frac{t^i |||N|||^i}{i!}$. Puisque $p(t)$ est un polynôme en t, pour tout $\varepsilon > 0$ il existe $C_\varepsilon > 0$ telle que $p(t) \leq C_\varepsilon \exp(\varepsilon t)$ pour tout $t \geq 0$. Donc, on obtient

$$\|y(t)\| \leq C_\varepsilon |\exp(t(\lambda + \varepsilon))| \quad \forall\, t \geq 0\,.$$

Si $\mathrm{Re}(\lambda) < 0$ ceci implique que $y(t) \to 0$, ce qui montre la stabilité asymptotique dans ce cas. S'il existe un valeur propre λ_i de A avec $\mathrm{Re}(\lambda) = 0$ et

$k_i = 1$, alors on peut se ramener au cas scalaire pour montrer la stabilité de la solution.

Pour montrer la nécessité des conditions, on observe que si λ est une valeur propre de A avec $\mathrm{Re}(\lambda) > 0$ et avec vecteur propre $v \in \mathbb{C}^d$, alors la solution $\tilde{y}$ de l'EDO avec conditions initiales $\tilde{y}(0) = y_0 + \varepsilon v$ est donnée par :

$$\tilde{y}(t) = y(t) + \varepsilon \exp(\lambda t) v$$

pour tout $\varepsilon > 0$. On en déduit que la solution y est unstable.

De façon similaire, si $\mathrm{Re}(\lambda) = 0$ et la multiciplité algébrique et géométrique de λ ne sont pas égaux, il suffit de considerer le cas où A est donné par (1.3.3) et $d > 1$. On choisit $v = e_2 = (0, 1, 0, \ldots, 0)$ où $\{e_i\}_i$ est la base canonique de $\mathbb{C}^d$, et $\tilde{y}$ comme ci-dessus. Alors,

$$\tilde{y}(t) = \exp(t\lambda) \sum_{i=0}^{d-1} \frac{(tN)^i}{i!} (y_0 + \varepsilon e_2) = y(t) + \exp(t\lambda)\varepsilon(e_2 + te_1)$$

ce qui montre que y est unstable. □

1.3.2 Stabilité linéaire des méthodes RK

Considérons une méthode de Runge-Kutta à s stades, définie par la matrice des coefficients $\mathcal{A} = (a_{ij})_{ij}$, les nœuds $c = (c_i)_i$ et les poids $b = (b_i)_i$. En appliquant cette méthode au problème scalaire $y' = \lambda y$, avec $y_0 \in \mathbb{C}$, on obtient pour $0 \leq n \leq N - 1$,

$$\begin{aligned} k &= \mathbf{1}\lambda y_n + \tau\lambda\mathcal{A}k\,, \\ y_{n+1} &= y_n + \tau b \cdot k\,, \end{aligned}$$

où $k \in \mathbb{C}^s$, et $\mathbf{1} = (1, \ldots, 1) \in \mathbb{R}^s$. En résolvant le système pour k, on peut écrire $y_{n+1} = R(\lambda\tau)y_n$, où

$$R(z) = 1 + zb \cdot \left((\mathrm{Id} - z\mathcal{A})^{-1}\mathbf{1}\right)\,. \tag{1.3.4}$$

Cette fonction est une fonction rationnelle de z [2], et elle est appelée *fonction de stabilité* de la méthode.

En généralisant ce raisonnement au cas vectoriel, on obtient que le flot discret $\tilde{\Phi}_\tau$ associé à une méthode de Runge-Kutta appliquée au système (1.3.1) est aussi une application linéaire (comme le flot exact dans (1.3.2)) et

2. On peut vérifier cela en appliquant la formule de Cramer pour calculer l'inverse de $(\mathrm{Id} - z\mathcal{A})$.

peut être identifié avec une fonction rationnelle R (donnée par (1.3.4)) de la matrice τA. Plus précisément, étant donnés deux polynômes $P, Q : \mathbb{C} \to \mathbb{C}$, et la fonction rationnelle $R(z) := P(z)/Q(z)$, si $Q(A)$ est inversible, on définit

$$R(A) = Q(A)^{-1} P(A)\,.$$

Ce fait est résumé dans le lemme suivant.

Lemme 1.3.2. *Soit $\tilde{\Phi}_\tau$ le flot associé à une méthode de Runge-Kutta à s stades appliquée au problème (1.3.1). Alors, il existe une fonction rationnelle $R(z) = P(z)/Q(z)$, avec $P(0) = Q(0) = 1$, telle que $\tilde{\Phi}_\tau = R(\tau A)$; si la méthode est explicite, on peut prendre $Q(z) = 1$ et $R(z)$ est un polynôme de degré au plus égal à s.*

Exemple. Si on applique la méthode d'Euler implicite à l'équation $y' = \lambda y$, on a $y^{n+1} - y^n = \lambda\tau y^{n+1}$. En posant $y^{n+1} = R(\tau\lambda) y^n$, on obtient la fonction de stabilité

$$R(z) = \frac{1}{1-z}\,.$$

La solution numérique $(y_n)_n$ vérifie donc

$$y_{n+1} = B y_n$$

où $B \in \mathcal{M}_d(\mathbb{C})$. La stabilité (asymptotique) des solutions de ce système par rapport aux données initiales peut être étudiée de façon similaire au cas du système (1.3.1). On a le résultat suivant :

Proposition 1.3.3. *Les itérations $y_{n+1} = B y_n$ sont asymptotiquement stables par rapport aux données initiales y_0, si et seulement si le rayon spectral $\rho(B) < 1$, et en particulier dans ce cas, $y_n \to 0$ lorsque $n \to \infty$ pour tout $y_0 \in \mathbb{C}^d$. Elles sont stables si et seulement si : $\rho(B) \leq 1$ et pour tout $\lambda \in \sigma(B)$ avec $|\lambda| = 1$, la multiplicité algébrique et géométrique de λ sont égales.*

Démonstration. On observe que $y_n = B^n y_0$, et par la réduction de Jordan de B, il suffit de considérer le cas

$$B = \lambda I + N\,, \quad \text{où} \quad N = \begin{bmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & 0 & 1 \\ & & & 0 \end{bmatrix} \in \mathcal{M}_d(\mathbb{R})\,.$$

Dans ce cas, puisque $N^d = 0$, on observe que pour $n \geq d-1$,

$$B^n = \lambda^n I + \begin{pmatrix} n \\ 1 \end{pmatrix} \lambda^{n-1} N + \ldots + \begin{pmatrix} n \\ d-1 \end{pmatrix} \lambda^{n-(d-1)} N^{d-1}.$$

Si $\lambda = 0$, on a terminé. Sinon, on observe que $\|B^n\| \leq |\lambda|^n p(n)$, où

$$p(n) \coloneqq 1 + \begin{pmatrix} n \\ 1 \end{pmatrix} |\lambda|^{-1} \|N\| + \ldots + \begin{pmatrix} n \\ d-1 \end{pmatrix} |\lambda|^{-(d-1)} \|N\|^{d-1}.$$

Le terme $p(n)$ est un polynôme de degré $d-1$ en n. Donc, pour tout $\theta > 1$, il existe une constante $C_\theta > 0$ telle que $p(n) \leq C_\theta \theta^n$ pour $n \geq d-1$. Finalement, on obtient $\|B^n\| \leq C|\theta\lambda|^n$, et si $|\lambda| < 1$, on peut choisir $\theta > 1$ tel que $|\theta\lambda| < 1$, ce qui montre la première partie de la proposition. La stabilité et la nécessité des conditions peuvent être montrées de façon analogue à la preuve de la Proposition 1.3.1. □

Lemme 1.3.4. *Soient P et Q deux polynômes premiers entre eux, et $R = P/Q$. Si $A \in \mathcal{M}_d(\mathbb{C})$, alors $R(A)$ est bien défini si et seulement si $Q(\lambda) \neq 0$ pour tout $\lambda \in \sigma(A)$. Dans ce cas,*

$$\sigma(R(A)) = R(\sigma(A))$$

Démonstration. Par la réduction de Jordan de la matrice $A = V\Lambda V^{-1}$, et puisque pour tout polynôme P et toute matrice inversible M,

$$P(MAM^{-1}) = MP(A)M^{-1},$$

il suffit de prouver le résultat pour la matrice Λ. De plus, puisque $P(\Lambda) = \operatorname{diag}(P(\Lambda(\lambda_1)), \ldots, P(\Lambda(\lambda_k)))$, il suffit de considérer le cas

$$A = \lambda I + N, \quad \text{où} \quad N = \begin{bmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & 0 & 1 \\ & & & 0 \end{bmatrix}.$$

On observe que si P est de degré p, on remplace $P(\lambda I + N)$ par son développement limité en λI, ce qui donne

$$P(\lambda I + N) = P(\lambda) I + \sum_{i=1}^{p} \frac{P^{(i)}(\lambda)}{i!} N^i = P(\lambda) I + N_P,$$

où N_P est une matrice triangulaire supérieure avec diagonale nulle. De même, $Q(\lambda I + N) = Q(\lambda)I + N_Q$, et Q est inversible si et seulement si $Q(\lambda) \neq 0$. Dans ce cas, on a

$$Q(\lambda I + N)^{-1} = Q(\lambda)^{-1} I + N_{Q^{-1}},$$

où $N_{Q^{-1}}$ est aussi triangulaire supérieure à diagonale nulle. Ainsi, il existe une matrice N_R triangulaire supérieure à diagonale nulle telle que

$$R(\lambda I + N) = R(\lambda)I + N_R,$$

ce qui donne le résultat. □

On introduit la *région de stabilité* d'une méthode de Runge-Kutta avec fonction de stabilité R, l'ensemble

$$\mathcal{S} \coloneqq \{z \in \mathbb{C} \,:\, |R(z)| \leq 1\}.$$

Définition 1.3.5 (A-stabilité)**.** Une méthode de Runge-Kutta avec région de stabilité $\mathcal{S}$ est dite *A-stable* si et seulement si :

$$\mathcal{S} \supseteq \mathbb{C}^- \coloneqq \{z \in \mathbb{C} \,:\, \mathrm{Re}(z) \leq 0\}.$$

D'après le principe du maximum du module pour les fonctions holomorphes, on a que $|R(z)| = 1$ exactement sur le bord de la région de stabilité. En effet, si $|R(z_0)| = 1$ pour $z_0 \in \mathring{\mathcal{S}}$, alors nécessairement z_0 est en maximum pour $|R(\cdot)|$, mais cela est possible seulement si $|R(\cdot)|$ était une constante. Donc, on en déduit que pour une méthode A-stable avec fonction de stabilité R,

$$\mathrm{Re}(\lambda) < 0 \quad \forall \lambda \in \sigma(A) \Rightarrow \tau\lambda \in \mathring{\mathcal{S}}, \quad \forall \lambda \in \sigma(A) \Rightarrow \rho(R(\tau A)) < 1,$$

et cette suite d'implications reste vraie si on remplace les inégalités strictes par des inégalités larges.

De plus, la preuve du lemme 1.3.4 montre que la structure par blocs de la réduction de Jordan de la matrice τA est préservée par l'application de la fonction rationnelle R. Ce fait, combiné à la Proposition 1.3.3, implique que les méthodes A-stables héritent la stabilité (asymptotique) du problème linéaire. Plus précisément, on a le résultat suivant :

Théorème 1.3.6. *Soit $(y_n)_n$ une solution discrète obtenue par l'application d'une méthode de Runge-Kutta avec fonction de stabilité R, c'est-à-dire $y_{n+1} = R(\tau A)y_n$ pour tout $n \geq 0$. Alors $(y_n)_n$ est (asymptotiquement) stable pour tout $\tau > 0$ et toute matrice $A \in \mathcal{M}_d(\mathbb{C})$ pour laquelle le système linéaire $y' = Ay$, $y(0) = y_0$, est (asymptotiquement) stable, si et seulement si la méthode est A-stable.*

Exemple. La fonction de stabilité de la méthode d'Euler implicite est donnée par $R(z) = (1-z)^{-1}$. Elle est A-stable puisque la région de stabilité $\mathcal{S}$ est l'ensemble complémentaire de la boule ouverte centrée en $z = 1$ et de rayon égal à 1.

Exemple. La fonction de stabilité associée à la méthode des trapèzes est $R(z) = \frac{1+z/2}{1-z/2}$. Elle est A-stable puisque la région de stabilité est donnée par $\mathcal{S} = \mathbb{C}^-$.

En vertu du théorème 1.3.6, lorsque la région de stabilité est bornée, pour garantir que la discrétisation hérite les propriétés de stabilité du système d'origine, on doit imposer des restrictions sur le pas. Notamment, on demande

$$\tau_n \leq \tau^c \coloneqq \inf \{\tau > 0 : \exists \lambda \in \sigma(A) \text{ tel que } \tau\lambda \notin \mathcal{S}\},$$

où τ^c est généralement appelé **pas critique** ou caractéristique. On remarque que cette quantité dépend de la méthode, mais aussi du problème qu'on est en train de discrétiser : en particulier, pour un problème raide, $\tau^c \ll 1$.

Remarque 1.3.7 (A-stabilité des méthodes explicites)**.** *On remarque que :*

1. *une méthode explicite ne peut pas être A-stable ! En effet, la fonction de stabilité d'une méthode explicite est un polynôme $P(z)$ et on a*

$$\lim_{|z|\to\infty} |P(z)| = \infty,$$

   ce qui montre que la région de stabilité $\mathcal{S}$ doit être bornée (et donc $\tau^c < \infty$) ;

2. *pour des problèmes raides, on privilégiera donc des méthodes implicites.*

Remarque 1.3.8 (Fonctions de stabilité et ordre de la méthode)**.** *Pour qu'une méthode RK avec fonction de stabilité R soit consistante, il faut nécessairement que $R(z) \approx \exp(z)$ (le flot exact de l'équation $y'(t) = y(t)$ étant défini par $\Phi_t(y_0) \coloneqq \exp(t)y_0$). Plus précisément, on observe que :*

1. *pour une méthode d'ordre p, on a toujours*

$$R(z) = 1 + z + \ldots + \frac{z^p}{p!} + \mathcal{O}(|z|^{p+1}),$$

   et pour une méthode explicite d'ordre p avec $p = s$ (s étant le nombre de stades), l'égalité est exacte sans le résidu $\mathcal{O}(|z|^{p+1})$;

2. *pour toute méthode consistante, on a* $R(0) = 1$*, et donc, par le principe du maximum des fonctions holomorphes :*

$$0 \in \partial\mathcal{S},$$

où $\partial\mathcal{S}$ *dénote le bord de la région de stabilité.*

1.3.3 Quelques remarques sur la stabilité non-linéaire des méthodes RK

Revenons au problème de Cauchy (autonome) :

$$\begin{cases} y'(t) = f(y(t)) \\ y(0) = y_0 \end{cases} \tag{1.3.5}$$

où $f \in C^1(\mathbb{R}^d, \mathbb{R}^d)$. Le concept de stabilité linéaire peut être étendu à ce système autour de ses points fixes, c'est-à-dire des points $y^* \in \mathbb{R}^d$ tels que $f(y^*) = 0$. En effet, on a le résultat suivant :

Théorème 1.3.9. *Soit* $f \in C^1(\mathbb{R}^d, \mathbb{R}^d)$ *et* $y^* \in \mathbb{R}^d$ *tel que* $f(y^*) = 0$*. Si*

$$\max\{\mathrm{Re}(\lambda) \,:\, \lambda \in \sigma(Df(y^*))\} < 0$$

alors la solution stationnaire $y(t) = y^*$ *est asymptotiquement stable.*

Quand on applique une méthode de RK au système (1.3.5), on peut vérifier que si y^* est un point fixe, alors la solution discrète qu'on obtient avec $y_0 = y^*$ est aussi stationnaire, c'est-à-dire $y_n = y^*$ pour tout $n \geq 0$. Pour pouvoir caractériser sa stabilité, on observe que si on linéarise le système (1.3.5) autour de y^*, on obtient le système suivant :

$$\begin{cases} z'(t) = Df(y^*)(z(t) - y^*) \\ z(0) = z_0 \end{cases}$$

dont la solution exacte est $z(t) = y^* + \exp(Df(y^*))(z_0 - y^*)$. De façon similaire, en appliquant une méthode de RK avec fonction de stabilité R à ce système, on obtient la solution discrète définie par les itérations :

$$z_n = y^* + R(\tau Df(y^*))(z_0 - y^*)\,.$$

Cette solution est différente de celle qu'on obtient en appliquant la méthode de RK directement au système non linéaire, mais on peut l'utiliser pour déduire des conditions sur le pas τ qui garantissent la stabilité des solutions discrètes.

Théorème 1.3.10. *Dans les hypothèses du Théorème 1.3.9, si*

$$\tau < \tau^c \coloneqq \inf\{\tau > 0 \,:\, \exists\, \lambda \in \sigma(Df(y^*)) \text{ tel que } \tau\lambda \notin \mathcal{S}\}$$

alors la solution discrète $y_n = y^*$ qu'on obtient en appliquant une méthode de RK avec région de stabilité $\mathcal{S}$ est asymptotiquement stable.

Remarque 1.3.11. *Malgré le résultat du Théorème 1.3.10, le concept de A-stabilité n'est pas suffisant pour garantir qu'une méthode de RK produise des solutions stables en toute généralité, et il est dangereux de se baser uniquement sur ce critère lorsqu'on considère des systèmes non linéaires. Il existe cependant d'autres critères (comme la B-stabilité, la symplecticité, etc.) qui permettent de caractériser le comportement des solutions discrètes même loin des points stationnaires.*

2 Résolution numérique de systèmes non-linéaires

Soit $F : U \subset \mathbb{R}^d \to \mathbb{R}^d$. On souhaite résoudre numériquement le problème : trouver $x \in U$ tel que

$$F(x) = 0. \tag{2.0.1}$$

Ce type de problème apparait naturellement dans le contexte de la discrétisation des EDOs pour les méthodes dites implicites, où, à chaque pas de temps, on doit résoudre un problème du type (2.0.1).

En général, s'il existe une solution x^* du problème (2.0.1), on ne peut pas en donner une formule explicite et on se contente de chercher une approximation x^ε de x^*. L'objectif de ce chapitre sera donc de construire des méthodes qui, étant donné une tolérance $\varepsilon > 0$, nous permettent de trouver

$$x^\varepsilon \in U : \exists x^* \in U \text{ avec } F(x^*) = 0 \text{ et } \|x - x^*\| \leq \varepsilon.$$

L'idée de base sera de ramener la solution du problème (2.0.1) à un problème de point fixe $x = \Phi(x)$, ce qui suggère l'algorithme suivant :

$$x^{n+1} = \Phi(x^n),$$

où $\Phi : U \subset \mathbb{R}^d \to \mathbb{R}^d$ est une fonction qui définit la méthode.

2.1 Types de convergence

Un critère utile pour classifier les différentes méthodes que nous allons introduire est donné par l'ordre de convergence par rapport à une norme $\|\cdot\|$ sur $\mathbb{R}^d$.

Définition 2.1.1. Soit $(x^n)_n$ une suite avec $x_n \in \mathbb{R}^d$ et $x^* \in \mathbb{R}^d$. On dit que x^n converge vers x^*

1. à l'ordre 1 (linéairement), si et seulement si il existe $\lambda \in [0, 1)$ tel que pour n suffisamment grand,

$$\|x^{n+1} - x^*\| \leq \lambda \|x^n - x^*\|; \tag{2.1.1}$$

2. à l'ordre 2 (quadratiquement), si et seulement si il existe $C \geq 0$ tel que pour n suffisamment grand,

$$\|x^{n+1} - x^*\| \leq C\|x^n - x^*\|^2;$$

3. à l'ordre $\alpha > 1$, si et seulement si il existe $C \geq 0$ tel que pour n suffisamment grand,

$$\|x^{n+1} - x^*\| \leq C\|x^n - x^*\|^\alpha. \tag{2.1.2}$$

Remarque 2.1.2. *La convergence linéaire dans une norme n'implique pas la convergence linéaire dans une autre norme ! Il suffit de considérer la suite* $(x_n)_n$ *avec* $x_n \in \mathbb{R}^2$ *définie par*

$$x_n = \frac{1}{2^n}(\cos(n\pi/2), \sin(n\pi/2)).$$

Qu'est-ce qui se passe si on choisit la norme euclidienne ou la norme $\|(x_1, x_2)\|_2' := \sqrt{x_1^2 + 4x_2^2}$ *?*

En revanche, si une suite converge à l'ordre $\alpha > 1$ *dans une norme* $\|\cdot\|$*, elle converge avec le même ordre pour toute autre norme sur* $\mathbb{R}^d$*.*

Dans le cas linéaire, la convergence vers la limite x^* est géometrique. Par exemple, si (2.1.1) est valide pour tout $n \geq n_0$ on a

$$\|x^{n+1} - x^*\| \leq \lambda^{n-n_0+1}\|x^{n_0} - x^*\| = C\lambda^{n+1}\,.$$

Une convergence d'ordre $\alpha > 1$ est dite super-linéaire. Plus en général, on dit que $x^n \to x^*$ de façon super-linéaire ssi :

$$\lim_{n\to\infty} \frac{\|x^{n+1} - x^*\|}{\|x^n - x^*\|} = 0.$$

Une méthode qui converge de façon super-linéaire, converge aussi de façon linéaire. En effet, la convergence est de plus en plus rapide pour α de plus en plus grand. Plus précisement, si equation (2.1.2) est verifié pour tout $n \geq n_0$, et si on définit $u^n := -\log(\|x^n - x^*\|) - \log(C)/(\alpha - 1)$, on obtient

$$u^{n+1} \geq \alpha u^n \geq \alpha^{n-n_0+1}u^{n_0} = \beta\alpha^n$$

où $\beta := u^{n_0}\alpha^{-n_0} > 0$. On en déduit que pour $n \geq n_0$,

$$\|x^{n+1} - x^*\| \leq \frac{1}{C^{\alpha-1}} \exp(-\beta\alpha^{n+1}) = C'\lambda^{\alpha^{n+1}}\,,$$

où $C' > 0$ et $\lambda = \exp(-\beta) < 1$.

Pourtant, souvent il est preferable d'utiliser une méthode qui converge avec un ordre bas, plutot qu'une méthode d'ordre élevé. La raison est que le nombre d'opérations nécessaires pour calculer x^{n+1} à partir de x^n peut varier énormement entre differentes méthodes. En effet, à parité de nombre d'opérations, on peut arriver plus rapidement à une approximation x^ε de la solution avec une méthode d'ordre bas. On verra des exemples concrets de ce phenomène dans la suite.

2.2 Méthodes itératives

Typiquement, une méthode itérative pour la résolution du problème (2.0.1) prend la forme suivante :

$$x^{n+1} = \Phi(x^n) \tag{2.2.1}$$

où $\Phi : U \subset \mathbb{R}^d \to \mathbb{R}^d$ est une fonction qui définit la méthode.

Le résultat de convergence de la méthode est connu sous le nom de théorème du point fixe de Picard. Avant de l'énoncer, on rappelle tout d'abord la définition de fonction contractante.

Définition 2.2.1 (Fonction contractante)**.** On dit que $\varphi : U \subset \mathbb{R}^n \to \mathbb{R}^n$ est une fonction contractante ssi il existe une constante $\lambda < 1$ telle que

$$\|\varphi(x) - \varphi(y)\| \leq \lambda \|x - y\|\,, \quad \forall x, y \in U\,,$$

c'est-à-dire ssi φ est une fonction lipschitzienne avec une constante de Lipschitz $\lambda < 1$.

Théorème 2.2.2. *Soit $\Phi : U \subset \mathbb{R}^d \to \mathbb{R}^d$ une fonction contractante avec constante de Lipschitz $\lambda < 1$ et telle que $\Phi(x) \in \overline{B}$ pour tout $x \in \bar{B} \subseteq U$. Alors, il existe un unique point fixe $x^* \in \overline{B}$ de Φ, c'est-à-dire $\Phi(x^*) = x^*$, et la suite $(x^n)_n$ définie par (2.2.1) converge linéairement pour tout $x^0 \in \overline{B}$, et en particulier, pour tout $n \geq 0$:*

$$\|x^{n+1} - x^*\| \leq \lambda \|x^n - x^*\|\,.$$

L'hypothèse que Φ soit contractante sur son domaine est généralement trop restrictive pour les méthodes qu'on considérera dans la suite. Par contre, si Φ est suffisamment régulière, on peut utiliser le théorème du point fixe pour déduire un résultat de convergence local à partir du comportement de Φ autour d'un point fixe x^*.

Supposons que $\Phi : U \subset \mathbb{R}^d \to \mathbb{R}^d$ est différentiable sur un ouvert $U \subset \mathbb{R}^d$. Alors, pour tout $x \in U$,

$$\Phi(x) = x^* + D\Phi(x^*)(x - x^*) + o(\|x - x^*\|),$$

où $D\Phi(x^*) \in \mathcal{L}(\mathbb{R}^d, \mathbb{R}^d)$ est la dérivée de Φ en x^*. On identifiera $D\Phi(x^*)$ avec sa représentation matricielle dans $\mathcal{M}_d(\mathbb{R})$ par rapport à la base canonique, c'est-à-dire la matrice jacobienne de Φ en x^*.

On peut caractériser la contractivité de Φ autour de x^* en utilisant le rayon spectral de $D\Phi(x^*)$, noté $\rho(D\Phi(x^*))$. On a le lemme suivant :

Lemme 2.2.3. *Soit $A \in \mathcal{M}_d(\mathbb{R})$. Pour tout $\varepsilon > 0$, il existe une norme $\|\cdot\|_\varepsilon$ sur $\mathbb{R}^d$ telle que la norme induite sur $\mathcal{M}_d(\mathbb{R})$, notée $|||\cdot|||_\varepsilon$, satisfait*

$$|||A|||_\varepsilon \leq \rho(A) + \varepsilon.$$

Supposons que $\rho(D\Phi(x^*)) < 1$. Alors, il existe $\varepsilon > 0$ tel que $\rho(D\Phi(x^*)) < 1 - 2\varepsilon$. Le lemma ci-dessus implique qu'il existe une norme $\|\cdot\|_\varepsilon$ telle que $|||D\Phi(x^*)|||_\varepsilon \leq 1 - \varepsilon$. De plus,

$$\|\Phi(x) - x^*\|_\varepsilon \leq |||D\Phi(x^*)|||_\varepsilon \|x - x^*\|_\varepsilon + o(\|x - x^*\|)\,.$$

Par definition de $o(\|x - x^*\|)$ il existe un $\eta > 0$ tel que pour tout $x \in \overline{B_\eta(x^*)} \subset V$ on a $o(\|x - x^*\|) \leq \varepsilon \|x - x_\varepsilon\|_\varepsilon$. Finalement on obtient :

$$\|\Phi(x) - x^*\|_\varepsilon \leq (\rho(D\Phi(x^*)) + 2\varepsilon)\|x - x^*\|_\varepsilon.$$

Cela montre que $\Phi(x) \in \overline{B_\eta(x^*)}$ pour tout $x \in \overline{B_\eta(x^*)}$ et que la suite définie par (2.2.1) converge linéairement vers $x*$. On a donc le résultat résultat suivant.

Théorème 2.2.4. *Soit $\Phi : U \subset \mathbb{R}^d \to \mathbb{R}^d$. Suppose que $x^* \in U$ est un point fixe, que Φ est différentiable en x^* et que $\rho(D\Phi(x^*)) < 1$. Alors, il existe $\eta > 0$ tel que pour tout $x^0 \in \overline{B_\eta(x^*)} \subset U$ la suite $(x^n)_n$ verifient (2.2.1) est bien définie et pour $n \to \infty$ elle converge linéairement vers x^*.*

Remarque 2.2.5. *Dans les mêmes hypothèses, on obtient aussi qu'il existe une constante $\lambda < 1$ et une norme $\|\cdot\|$ sur $\mathbb{R}^d$, telles que :*

$$\|x^{n+1} - x^n\| \leq \lambda \|x^n - x^{n-1}\|\,. \tag{2.2.2}$$

A noter que (2.2.2) ne peut pas etre deduite comme consequence de la convergence linéaire de la suite $(x^n)_n$. Par exemple, la suite $x^n = s(n)/(n+1)$,

où $s(n) = 1$ *si* $n \in \{0, 1, 4, 5, 8, 9, \ldots\}$ *et* $s(n) = -1$ *autrement, converge linéarement vers* $x^* = 0$ *mais elle ne vérifie pas* (2.2.2)*. De la même manière, l'inegalité* (2.2.2) *n'implique pas la convergence linéare : on peut considérer la suite définie par* $x^0 = 0$*,* $x^1 = 1$*,* $x^2 = 1/3$*, et* $x^n = x^{n \bmod 3}(2/3)^{n \bmod 3}$ *pour* $n \geq 3$*, qui vérifie* (2.2.2) *avec* $\lambda = 2/3$ *mais ne converge pas linéairement.*

Supposons que $D\Phi(x^*) = 0$ et que Φ est deux fois différentiable en x^*. Dans ce cas, par le théorème précedent, la suite $(x^n)_n$ converge linéairement vers x^*, mais on peut montrer que la convergence est quadratique des que x^n devient suffisamment proche de x^*. Tout d'abord on observe que, pour tout x dans un voisinage V de x^*,

$$\Phi(x) = x^* + \frac{1}{2}D^2\Phi(x^*)(x - x^*, x - x^*) + o(\|x - x^*\|^2)\,,$$

où $D^2\Phi(x^*) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ est la dérivé seconde de Φ en x^*. Il existe une constante $C > 0$ telle que $\|D^2\Phi(x^*)(h, h)\| \leq C\|h\|^2$ pour tout $h \in \mathbb{R}^d$. Soit $\eta > 0$ tel que $o(\|x - x^*\|^2) \leq C\|x - x^*\|^2/2$ pour tout $x \in \overline{B}_\eta(x^*) \subset V$. Alors, on obtient pour tout $x \in \overline{B}_\eta(x^*)$,

$$\|\Phi(x) - x^*\| \leq C\|x - x^*\|^2\,.$$

Soit $\delta \coloneqq \min(1/(2C), \eta)$, alors pour tout $x \in \overline{B}_\delta(x^*)$, on a $\Phi(x) \in \overline{B}_\delta(x^*)$. En effet

$$\|\Phi(x) - x^*\| \leq C\|x - x^*\|^2 \leq \frac{1}{2}\|x - x^*\|\,.$$

Cela implique que la suite définie par (2.2.1), avec $x^0 \in \overline{B_\delta(x^*)}$ vérifie pour tout $n \geq 0$,

$$\|x^{n+1} - x^*\| \leq C\|x^n - x^*\|^2\,.$$

On a donc le résultat suivant.

Théorème 2.2.6. *Soit* $\Phi : U \subset \mathbb{R}^d \to \mathbb{R}^d$*. Suppose que* $x^* \in U$ *est un point fixe, que* Φ *est deux fois différentiable en* $x^* \in U$*, et que* $D\Phi(x^*) = 0$*. Alors, il existe* $\delta > 0$ *tel que pour tout* $x^0 \in \overline{B_\delta(x^*)} \subset U$ *la suite* $(x^n)_n$ *verifient* (2.2.1) *est bien définie et pour* $n \to \infty$ *elle converge quadratiquement vers* x^**.*

2.2.1 Critères d'arrêt

Revenons au problème $F(x) = 0$. Puisque nous ne connaissons pas sa solution exacte x^*, nous ne pouvons pas calculer l'erreur $\|x^n - x^*\|$ à l'itération n de la méthode itérative choisie pour le résoudre. Cela pose le problème

de déterminer si x^n represente une bonne approximation de x^* ou, ce qui est équivalent, de choisir un *critère d'arrêt* pour la méthode. En général, le choix de ce critère dépend à la fois du problème et de la méthode itérative utilisée.

Critère 1 : Condition sur le résidue $\|F(x^n)\|$. Supposons que F soit différentiable en x^* et que $DF(x^*)$ soit inversible. Dans un voisinage de la solution x^*,

$$DF(x^*)^{-1}F(x) + DF(x^*)^{-1}o(\|x - x^*\|) = x - x^* .$$

Pour x suffisamment proche à x^* on peut supposer

$$\|DF(x^*)^{-1}o(\|x - x^*\|)\| \leq \|x - x^*\|/2.$$

On en déduit que

$$\frac{1}{2}\|x - x^*\|\left|\left|\left|DF(x^*)^{-1}\right|\right|\right|^{-1} \leq \|F(x)\| \leq 2|||DF(x^*)|||\|x - x^*\|$$

où la deuxième inégalité est aussi une consequence du developpement limité de F autour de x^*. Pour x^n suffisamment proche à x^*, on a donc le critère suivant :

$$\|F(x^n)\| \leq \frac{1}{2}\left|\left|\left|DF(x^*)^{-1}\right|\right|\right|^{-1}\varepsilon \quad \Longrightarrow \quad \|x^n - x^*\| \leq \varepsilon .$$

Critère 2 : Condition sur le pas $\|x^{n+1}-x^n\|$ (cas linéaire). Supposons que $(x^n)_n$ converge linéairement vers x^* avec une constante de decroissance $\lambda < 1$. On a alors

$$\|x^n - x^*\| \leq \|x^n - x^{n+1}\| + \lambda\|x^n - x^*\| ,$$

qui implique :

$$\|x^n - x^*\| \leq \frac{1}{1-\lambda}\|x^{n+1} - x^n\|$$

On obtient le critère suivant :

$$|x^{n+1} - x^n\| \leq (1-\lambda)\varepsilon \quad \Longrightarrow \quad \|x^n - x^*\| \leq \varepsilon .$$

Critère 3 : Condition sur le pas $\|x^{n+1}-x^n\|$ (cas super-linéaire). Si la suite $(x^n)_n$ converge vers x^* à l'ordre $\alpha > 1$, pour n suffisamment grand

$$\|x^n - x^*\| \leq \|x^{n+1} - x^n\| + C\|x^n - x^*\|^\alpha .$$

Puisque $\alpha > 1$, pour n suffisamment grand, on peut estimer $C\|x^n - x^*\|^\alpha \leq \|x^n - x^*\|/2$, ce qui donne le critère suivant :

$$\|x^{n+1} - x^n\| \leq \frac{\varepsilon}{2} \quad \Longrightarrow \quad \|x^n - x^*\| \leq \varepsilon .$$

2.3 Méthode de point fixe

On reformule le problème $F(x) = 0$ comme un problème de point fixe $\Phi(x) = x$, où $\Phi(x) = x - MF(x)$ et $M \in \mathcal{M}_d(\mathbb{R})$ est une matrice inversible. En effet, on a

$$\Phi(x) = x \quad \Longleftrightarrow \quad F(x) = 0\,.$$

Ici, on considère le cas plus simple où $M = p\,\mathrm{Id}$, avec $p \in \mathbb{R}$ et où Id est l'identité sur $\mathbb{R}^d$. Dans ce cas, la méthode de point fixe est définie par la formule :

$$x^{n+1} = \Phi(x^n)\,, \quad \Phi(x) = x - pF(x)\,. \tag{2.3.1}$$

Rappel: Fonctions monotones

Définition 2.3.1 (Fonction monotone/α-monotones)**.** Soit $\varphi : U \subset \mathbb{R}^d \to \mathbb{R}^d$. On dit que φ est *monotone* ssi

$$\langle \varphi(x) - \varphi(y), x - y\rangle \geq 0 \quad \forall x, y \in U\,.$$

On dit que φ est α-*monotone* ssi il existe une constante $\alpha > 0$ telle que

$$\langle \varphi(x) - \varphi(y), x - y\rangle \geq \alpha\|x - y\|^2 \quad \forall x, y \in U\,.$$

Exercice 2.3.2. Montrer les affirmations suivantes :

1. soit $\varphi \in C^1(\mathbb{R})$, alors φ est monotone ssi $\varphi' \geq 0$;
2. soit $f \in C^1(\mathbb{R}^d, \mathbb{R})$, alors $\nabla f \in C^0(\mathbb{R}^d, \mathbb{R}^d)$ est α-monotone ssi f est α-convexe, c'est à dire

$$f((1-t)x + ty) \leq (1-t)f(x) + tf(y) - \alpha\frac{t(1-t)}{2}\|x - y\|^2\,,$$

pour tout $t \in [0, 1]$.

Trouver une fonction monotone $F : \mathbb{R}^2 \to \mathbb{R}^2$ qui ne soit pas le gradient d'une fonction convexe.

Si F est L-lipschitzienne et α-monotone, et différentiable en x^*, on obtient

$$L\|h\|^2 \geq DF(x^*)(h, h) \geq \alpha\|h\|^2 \quad \forall\, h \in \mathbb{R}^d\,,$$

ce qui implique que

$$(1 - pL)\|h\|^2 \leq h^T D\Phi(x^*)h = h^T(\mathrm{Id} - pJF(x^*))h \leq (1 - p\alpha)\|h\|^2\,.$$

Si on suppose que $D\Phi(x^*)$ est symmétrique, en appliquant le Théorème 2.2.4 on obtien la convergence linéaire (locale) de la méthode si $p \in (0, 2/L)$. On

peut aussi obtenir un résultat similaire en demandant moins de régularité sur F en x^*.

Théorème 2.3.3. *Soit $F : \mathbb{R}^d \to \mathbb{R}^d$ une fonction L-lipschitzienne et α-monotone, et soit $x^* \in \mathbb{R}^d$ tel que $F(x^*) = 0$. Notons $p^* := 2\alpha/L^2$. Si $p \in (0, p^*)$, alors pour tout $x^0 \in \mathbb{R}^d$ la suite $(x^n)_n$ définie par (2.3.1) converge linéairement vers x^*.*

Démonstration. On montre que $\Phi : x \mapsto x - pF(x)$ est contractante. Pour tout $x, y \in \mathbb{R}^d$,

$$\begin{aligned}\|\Phi(x) - \Phi(y)\|^2 &= \|x - y\|^2 - 2p\langle F(x) - F(y), x - y\rangle + \|F(x) - F(y)\|^2 \\ &\leq (1 - 2p\alpha + pL^2)\|x - y\|^2\end{aligned}$$

et on observe que $0 < (1 - 2p\alpha + pL^2) < 1$ pour $0 < p < p^*$. □

Remarque 2.3.4. *Si F est L-lipschitzienne et α-monotone, alors $\alpha \leq L$ et en particulier $2/L \geq 2\alpha/L^2 = p^*$.*

Exemple (Méthodes de Runge-Kutta implicite). Revenons à la résolution des EDOs avec des méthodes de Runge-Kutta. Considèrons le système linéaire $y'(t) = Ay(t)$ et appliquons une méthode implicite avec un pas de temps τ. À chaque pas de temps on doit résoudre un système linéaire dans la forme :

$$Q(\tau A)x = P(\tau A)y^n\,,$$

puisque le flot discret est donné par une fonction rationnelle de τA, $R(\tau A) = Q(\tau A)^{-1}P(\tau A)$. On essaye de résoudre ce système avec un méthode de point fixe en chosissant $\Phi(x) = x - Q(\tau A)x + P(\tau A)y^n$, ce qui donne les itérations de Richardson. Pour ce problème linéaire, les itérations convergent ssi

$$\rho(I - Q(\tau A)) < 1$$

ou, en introduisant le polynome $Q_0(z) = 1 - Q(z)$, $\rho(Q_0(\tau A)) < 1$, ou encore par un des résultat du premier chapitre

$$|Q_0(\lambda\tau)| < 1\,, \quad \forall \lambda \in \sigma(A).$$

L'ensemble $\{z \in \mathbb{C} : |Q_0(z)| < 1\}$ étant borné, cette condition implique que pour obtenir une méthode convergente

$$\tau \leq \mathcal{O}(\rho(A)^{-1}) \quad \text{lorsque}\, \rho(A) \to \infty\,.$$

On a donc à nouveau une réstriction sur le pas de temps pour un problème raide même en utilisant une méthode implicite ! Ceci suggère que les méthode de point fixe ne sont pas appropriées pour ce problème.

2.4 Méthode de Newton

La méthode de Newton consiste à approcher à chaque itération la fonction F par son developpement limité en x^n, et à choisir x^{n+1} comme un zéro de ce développement. Plus précisement, supposons que x^n est une approximation de x^*, solution de $F(x) = 0$. Si F est différentiable en x^n, on obtient

$$F(x) = F(x^n) + DF(x^n)(x - x^n) + o(\|x - x^n\|)\,, \tag{2.4.1}$$

Cela suggère de choisir x^{n+1} comme solution de l'équation $F(x^n)+DF(x^n)(x-x^n) = 0$. Si $DF(x^n)$ est inversible, on obtient

$$x^{n+1} = \Phi(x^n)\,, \quad \Phi(x) \coloneqq x - DF(x)^{-1}F(x)\,. \tag{2.4.2}$$

Remarque 2.4.1. *Notons que :*

1. *la suite (2.4.2) est bien définie lorsque $DF(x^n)$ est inversible. Si $DF(x^n)$ n'est pas inversible ou mal conditionnée on arrête l'algorithme ;*
2. *à chaque étape de l'algorithme on doit calculer $DF(x^n)$ et résoudre le système linéaire $DF(x^n)y^n = F(x^n)$: les deux operations peuvent être très couteuses en temps de calcul.*

En remplacent x par x^* dans (2.4.1) ainsi que la définition de x^{n+1} on obtient

$$\|x^{n+1} - x^*\| \leq DF(x^n)^{-1} o(\|x^n - x^*\|)\,.$$

Si $(DF)^{-1}$ est borné uniformement dans un voisinage de x^*, cette inégalité implique la convergence de la linéaire de la méthode.

Supposons que F est de classe C^2 et $DF(x^*)$ est inversible. Alors $DF(x)$ est inversible dans un voisinage de x^* et Φ est bien définie dans le même voisinage. De plus, $DF(x)^{-1}$ est une fonction de classe C^1 et on a pour tout $h \in \mathbb{R}^d$,

$$\langle D(DF(x)^{-1}), h\rangle = DF(x)^{-1} \circ \langle D^2F(x), h\rangle \circ DF(x)^{-1}\,, \tag{2.4.3}$$

où $\langle D(DF^{-1}, h\rangle \in \mathcal{L}(\mathbb{R}^d, \mathbb{R}^d)$ et $\langle D^2F(x), h\rangle \in \mathcal{L}(\mathbb{R}^d, \mathbb{R}^d)$ sont les dérivés directionnelles des fonctions $x \mapsto DF(x)^{-1}$ et $x \mapsto DF(x)$ respectivement. Puisque $F(x^*) = 0$, on deduit que pour tout $h \in \mathbb{R}^d$,

$$\langle D\Phi(x^*), h\rangle = h - \langle D(DF^{-1})(x^*), h\rangle F(x^*) - \langle DF^{-1}(x^*)DF(x^*), h\rangle = 0\,.$$

On est donc dans les hypothèse du Théorème (2.2.6), ce qui implique la convergence quadratique locale de l'algorithme.

Pour donner un résultat de convergence plus générale avec des hypothèses moins restrictives on rappelle tout d'abord le lemme suivant.

Lemme 2.4.2. *Soit $M \in \mathcal{M}_d(\mathbb{R})$ tel que $|||M||| < 1$ alors $Id - M$ est inversible et on a :*

$$|||(Id - M)^{-1}||| \leq \frac{1}{1 - |||M|||}\,.$$

Supposons que DF est γ-lipschitzienne dans un voisinage $\overline{B_\eta(x^*)}$, avec $\eta > 0$, d'une solution x^*, et que $DF(x^*)$ est inversible. Alors prenant $M = Id - DF(x^*)^{-1}DF(x)$, si on trouve un voisinage de x^* tel que $|||M||| \leq 1/2$ alors par le lemme précedent, $DF(x)$ est inversibe, $|||DF(x)^{-1}DF(x^*)||| \leq 2$ et par le fait que $|||\cdot|||$ est une norme matricielle,

$$|||DF(x)^{-1}||| \leq 2|||DF(x^*)^{-1}|||\,. \tag{2.4.4}$$

En effet, on a

$$|||M||| \leq |||DF(x^*)^{-1}|||\,|||DF(x^*) - DF(x)||| \leq |||DF(x^*)^{-1}|||\gamma\|x - x^*\|\,.$$

Donc (2.4.4) est vérifié si $x \in \overline{B_\delta(x^*)}$ avec $\delta \leq \min(\eta, 1/(2\gamma|||DF(x^*)^{-1}|||))$.

Soit $x^n \in B_\delta(x^*)$. La fonction Φ qui définie la méthode de Newton est bien définie en x^n et on peut réécrire x^{n+1} comme suit

$$x^{n+1} = x^n - DF(x^n)^{-1}\int_0^1 DF(x^* + t(x^n - x^*))(x^n - x^*)\,\mathrm{d}t\,.$$

On en deduit que

$$x^{n+1} - x^* = DF(x^n)^{-1}\int_0^1 (DF(x^n) - DF(x^* + t(x^n - x^*))(x^n - x^*)\,\mathrm{d}t\,.$$

En utilisant l'inégalité (2.4.4), on obtient

$$\|x^{n+1} - x^*\| \leq |||DF(x^*)^{-1}|||\gamma\|x^n - x^*\|^2 \leq \frac{\delta}{2}\,,$$

puisque $\delta \leq 1/(2\gamma|||DF(x^*)^{-1}|||)$. Cette dernière inégalité montre que la méthode est bien définie pour tout $x^0 \in \overline{B}_\delta(x^*)$ et que en plus la convergence est quadratique. Plus précisement, on a le résultat suivant.

Théorème 2.4.3. *Soit $F : U \subset \mathbb{R}^d \to \mathbb{R}^d$. Suppose que DF est γ-lipschitzienne dans un voisinage d'un point $x^* \in U$ tel que $F(x^*) = 0$, et que $DF(x^*)$ est inversible. Notons $K := |||DF(x^*)^{-1}|||$. Alors, il existe $\delta > 0$ tel que pour tout $x^0 \in \overline{B_\delta(x^*)}$, la méthode de Newton définie par (2.4.2) converge vers x^* quadratiquement et avec constante $K\gamma$.*

Remarque 2.4.4. *Si on définit* $\alpha^n := -\log_{10}(K\gamma\|x^n - x^*\|)$, *on a* $\alpha^{n+1} \geq 2\alpha_n$. *Par example, si* $d = 1$ *et* $K\gamma \approx 1$, *cela veut dire que pour* x^n *suffisamment proche à* x^*, *le nombre de decimales correctes dans l'approximation* x^n *double à chaque itération.*

Remarque 2.4.5 (Échec de convergence/convergence linéaire)**.** *La convergence de la méthode de Newton est quadratique seulement dans un voisinage de* x^* *et dans les hypothèses du théorème précendent. En particulier :*

1. *l'initialisation de la méthode est une etape très importante puisque la suite* $(x^n)_n$ *peut ne pas converger si* $\|x^0 - x^*\|$ *est trop grand. Généralement on utilise donc une prémière méthode plus robuste pour determiner* x^0 *;*
2. *si* $DF(x^*)$ *n'est pas inversible la convergence est seulement linéaire.*

2.5 Méthode de la sécante (cas $d = 1$)

On considère ici le cas $d = 1$, et on cherche donc une solution de l'équation $f(x) = 0$ où $f : U \subset \mathbb{R} \to \mathbb{R}$. Dans la méthode de la sécante, l'approximation x^{n+1} à l'étape $n + 1$ est définie comme le zéro de la ligne droite passant par les points $(x^n, f(x^n))$ et $(x^{n-1}, f(x^{n-1}))$, c'est-à-dire :

$$x \in \mathbb{R} \mapsto f(x^n) + \frac{f(x^{n-1}) - f(x^n)}{x^{n-1} - x^n}(x - x^n)\,.$$

On obtient donc la méthode suivante : étant donnés $x_0, x_1 \in \mathbb{R}$ avec $f(x_0) \neq f(x_1)$, calculer pour $n \geq 1$,

$$x^{n+1} = x^n - \frac{x^{n-1} - x^n}{f(x^{n-1}) - f(x^n)} f(x^n)\,. \tag{2.5.1}$$

En adaptant le raisonnement précedent pour la méthode de Newton, on obtient le résultat suivant.

Théorème 2.5.1. *Soit* $f : U \subset \mathbb{R} \to \mathbb{R}$. *Suppose que* f' *est* γ*-lipschitzienne dans un voisinage d'un point* $x^* \in U$ *tel que* $f(x^*) = 0$, *et que* $f'(x^*) \neq 0$. *Alors, il existe* $\delta > 0$ *tel que pour tout* $x^0, x^1 \in \overline{B_\delta(x^*)}$, *la méthode de la sécante définie par* (2.6.3) *converge vers* x^* *super-linéairement, et en outre ils existent des constantes* $C > 0$ *et* $\lambda < 1$ *telles que*

$$|x^n - x^*| \leq C\lambda^{\alpha^n}$$

où $\alpha = (1 + \sqrt{5})/2$ *est le nombre d'or et il vérifie* $\alpha^2 - \alpha - 1 = 0$.

Démonstration. Notons $b^n \coloneqq (f(x^{n-1}) - f(x^n))/(x^{n-1} - x^n)$ et supposons que $b_n \neq 0$. Par le théorème fondamental du calcul,

$$b^n = \int_0^1 f'(x^n + t(x^{n-1} - x^n))\mathrm{d}t$$

Notons $K \coloneqq 1/|f'(x^*)|$. Soit $\eta > 0$ tel que $f'(x) > f'(x^*)/2$ pour tout $x \in [x^* - \eta, x^* + \eta]$ dans le cas où $f'(x^*) > 0$, ou $f'(x) < f'(x^*)/2$ pour tout $x \in [x^* - \eta, x^* + \eta]$ dans le cas où $f'(x^*) < 0$. Si $x^n, x^{n-1} \in [x^* - \eta, x^* + \eta]$, alors

$$|b^n| \geq \frac{1}{2K}\,.$$

Puisque $f(x^*) = 0$, on peut réécrire les itérations de la méthode comme suit :

$$\begin{aligned} x^{n+1} - x^* &= x^n - x^* - \frac{1}{b^n}\int_0^1 f'(x^* + t(x^n - x^*))(x^n - x^*)\mathrm{d}t \\ &= \frac{x^n - x^*}{b^n}\left(\int_0^1 f'(x^n + t(x^{n-1} - x^n))\mathrm{d}t - \int_0^1 f'(x^* + t(x^n - x^*))\mathrm{d}t\right)\,. \end{aligned}$$

Notons $e^n \coloneqq |x^n - x^*|$ l'erreur. En utilisant le fait que f' est γ-lipschizienne, on obtient pour $x^n, x^{n-1} \in [x^* - \delta, x^* + \delta]$ et $0 < \delta < \eta$,

$$\begin{aligned} e^{n+1} &\leq |b_n|^{-1}|x^n - x^*|\gamma\int_0^1 |(1-t)(x^n - x^*) + t(x^{n-1} - x^n)|\mathrm{d}t \\ &\leq 2Ke^n\gamma\left(\frac{|x^n - x^*|}{2} + \frac{|x^{n-1} - x^n|}{2}\right) \qquad (2.5.2) \\ &\leq 2K\gamma((e^n)^2 + \frac{1}{2}e^n e^{n-1}) \end{aligned}$$

En prenant $\delta \leq 1/(6K\gamma)$ on obtient $e^{n+1} \leq e^n/2 \leq \delta/2$ et donc la méthode est bien définie (puisque $\delta < \eta$, $|b^{n+1}| > (2K)^{-1}$) et elle converge linéairement vers x^*. Par l'inégalité précedente on vérifie facilement que la convergence est aussi super-linéaire. De plus, on a

$$e^{n+1} \leq 2K\gamma e^n e^{n-1}$$

qui peut etre réécrite en termes de $u^n \coloneqq -\log(e^n) - \log(2K\gamma)$ comme suit :

$$u^{n+1} \geq u^n + u^{n-1}$$

Puisque $u^n \to \infty$ lorsque $n \to \infty$, on peut choisir $n_0 \geq 1$ tel que $u_{n_0-1} \geq 1$ et $u_{n_0} \geq \alpha > 1$. Alors,

$$u^{n_0+1} \geq \alpha + 1 = \alpha^2$$

Par récurrence, on obtient que pour tout $k \geq 0$

$$u^{n_0+k} \geq \alpha^{k+1}$$

ou en définissant $\beta = \alpha^{-n_0+1} > 0$ on a $u^n \geq \beta\alpha^n$ pour tout $n \geq n_0$. Cela implique

$$e^n \leq \frac{1}{C} \exp(-\beta\alpha^n)\,.$$

□

Remarque 2.5.2. *La méthode de la sécante se comporte comme une méthode d'ordre $\alpha = (1+\sqrt{5})/2$, mais à différence de la méthode de Newton à chaque itération on doit effectuer une unique évaluation de f et aucune évaluation de f'. Pour cette raison, à parité de nombre d'évalutions, la méthode de la sécante se comporte comme une méthode d'ordre $\alpha^2 = \alpha + 1 > 2$ et dans ce sens elle est meilleur de la méthode de Newton.*

2.6 Méthode de Broyden

On revient à la résolution du problème $F(x) = 0$ où $F : U \subset \mathbb{R}^d \to \mathbb{R}^d$. Comme pour la méthode de la sécante, l'idée des méthode de *quasi-Newton* est de remplacer $DF(x^n)^{-1}$ par une approximation $B^n \in \mathcal{M}_d(\mathbb{R})$ construite itérativement. Plus précisement, on consruit la suite $(x^n)_n$ en posant :

$$x^{n+1} = x^n - B^n F(x^n)\,.$$

Dans la méthode de Broyden, B^n est construite en imposant la *condition de la sécante inverse*

$$B^n \Delta F^n = \Delta x^n \tag{2.6.1}$$

où $\Delta x^n := x^n - x^{n-1}$ et $\Delta F^n := F(x^n) - F(x^{n-1})$. La condition (2.6.1) à une unique solution dans le cas $d = 1$, et dans ce cas on obtient la méthode de la sécante. Pour $d > 1$, par contre, elle ne détermine pas B^n de façon unique : en effet on a un système de d équations et d^2 inconnus. Pour obtenir une unique solution, on fixe $B^0 \in \mathcal{M}_d(\mathbb{R})$ (par exemple, $B^0 = \mathrm{Id}$) et on cherche B^n satisfaisant (2.6.1) et tel que $\|B^n - B^{n-1}\|_F$ soit minimale, où $\|\cdot\|_F$ dénote la norme de Frobenius. En posant, $C^n := B^n - B^{n-1}$ on doit résoudre le problème

$$C^n = \mathrm{argmin}_C \left\{ \frac{1}{2}\|C\|_F^2\,;\ C\Delta F^n = b^n \right\}$$

où $b^n := \Delta x^n - B^{n-1}\Delta F^n$. Par les conditions d'Euler-Lagrange, si C^n est un minimiseur, il existe $\lambda \in \mathbb{R}^d$ tel que pour tout $1 \leq i, j \leq d$,

$$C^n_{ij} = \lambda_i \Delta F^n_j \,,$$

où en interprétantλ et ΔF^n comme des vecteurs colonne, $C^n = \lambda(\Delta F^n)^T$. Puisque le problème de minimisation est convexe cette condition est aussi suffisante pour que C^n soit un minimiseur. On utilise la contrainte $C\Delta F^n = b^n$ pour déterminer λ, et on obtient :

$$\lambda = \frac{1}{\|\Delta F^n\|^2} b^n \,,$$

ce qui implique que

$$B^n = B^{n-1} + \frac{1}{\|\Delta F^n\|^2}(\Delta x^n - B^{n-1}\Delta F^n)(\Delta F^n)^T \,. \tag{2.6.2}$$

On remarque que :

1. $B^n - B^{n-1}$ est une matrice de rang égal à 1 ;
2. $B^n e = B^{n-1} e$ pour tout $e \in \{\Delta F^n\}^\perp$.

En effet ses deux condition déterminent uniquement B^n et elle sont équivalentes à la formule (2.6.2).

Cette formule donne lieu à la méthode connue sous le nom de "Bad Broyden method". La raison de ce nom est que quand cette méthode à été proposé par Broyden, les résultats numériques observés étaient pires que ceux obtenus par une méthode alternative décrite ci-dessous et nommée "Good Broyden method". Pour cette dernière méthode, au lieu de construire directement une approximation de $(DF(x^n))^{-1}$, on propose d'abord une approximation H^n de $DF(x^n)$. Celle-ci doit vérifier la *condition de la sécante*

$$H^n \Delta x^n = \Delta F^n \,. \tag{2.6.3}$$

Comme dans le cas précedent, cette condition ne détermine pas H^n de façon unique, et on définit alors H^n comme la matrice qui réalise le minimum de $\|H^n - H^{n-1}\|_F$ sous la contrainte (2.6.3). Finalement, on obtient

$$H^n = H^{n-1} + \frac{1}{\|\Delta x^n\|^2}(\Delta F^n - H^{n-1}\Delta x^n)(\Delta x^n)^T. \tag{2.6.4}$$

Comme pour la méthode précedente, on a que :

1. $H^n - H^{n-1}$ est une matrice de rang égal à 1 ;

2. $H^n e = H^{n-1} e$ pour tout $e \in \{\Delta x^n\}^\perp$;

et ses deux condition déterminent uniquement H^n et elle sont équivalentes à la formule (2.6.4). Le premier point nous permet d'avoir une caractérisation explicite de la matrice $B^n = (H^n)^{-1}$ qui est nécessaire pour définir la méthode. On a le lemme suivant

Lemme 2.6.1 (Formule de Sherman–Morrison–Woodbury)**.** *Soit $A \in \mathcal{M}_d(\mathbb{R})$ une matrice inversible et soient $u, v \in \mathbb{R}^d$ deux vecteurs colonnes tels que $1 + v^T A^{-1} u \neq 0$. Alors $A + uv^T$ est inversible et on a*

$$(A + uv^T)^{-1} = A^{-1} - \frac{A^{-1} u v^T A^{-1}}{1 + v^T A^{-1} u}\,.$$

On obtient :

$$B^{n+1} = B^n + \frac{1}{(\Delta x^n)^T B^n \Delta F^n} (\Delta x^n - B^n \Delta F^n)(\Delta x^n)^T B^n\,.$$

3 Optimisation dans $\mathbb{R}^d$, descente de gradient et Newton

Soit $f : U \subset \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$. Dans ce chapitre, nous cherchons à résoudre le problème de minimisation

$$\inf_{x \in U} f(x)\,. \tag{3.0.1}$$

Une solution de ce problème est un vecteur $x^* \in U$, tel que $f(x^*) \leq f(x)$ pour tout $x \in U$. On appellera ces solutions les minimiseurs globaux de f. Évidemment, ce problème a un sens seulement si f est minorée sur U, sinon la valeur du problème (3.0.1) est $-\infty$ et il n'admet pas de solutions.

Pour résoudre numériquement ce problème, on construira des *suites minimisantes*, c'est-à-dire des suites $(x_n)_n$ avec $x_n \in U$, telles que

$$\lim_{n \to \infty} f(x_n) = \inf_{x \in U} f(x)\,.$$

Telles suites existent toujours, sans hypothèses supplémentaires sur f. En général, on s'intéressera aussi à construire des minimiseurs (ou solutions) locales du problème (3.0.1). Plus précisément, un *minimiseur local* de f (ou aussi une *solution locale* de (3.0.1)) est un vecteur $x^* \in U$, pour lequel il existe $r > 0$ tel que $f(x^*) \leq f(x)$ pour tout $x \in U \cap B_r(x^*)$.

Dans la suite, on construira des algorithmes qui nous permettent de calculer numériquement des approximations d'une solution x^* (locale ou globale, si elle existe) du problème (3.0.1). Plus précisément, pour $\varepsilon > 0$, on cherchera $x^\varepsilon \in U$ qui satisfait l'un des critères suivants :

1. Erreur sur la solution : $|x^\varepsilon - x^*| < \varepsilon$, où $x^* \in \operatorname{argmin}\{f(x);\, x \in U\}$;
2. Erreur sur la fonction objectif : $f(x^\varepsilon) \leq \min\{f(x);\, x \in U\} + \varepsilon$.

3.1 Quelques rappels d'optimisation convexe

L'existence de solutions du problème de minimisation (3.0.1) repose sur deux propriétés : la compacité des suites minimisantes pour pouvoir définir une limite x^* ; et la (semi-)continuité de la fonction f.

Définition 3.1.1 (Semi-continuité)**.** La fonction $f : U \subset \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ est semi-continue inférieurement en $\bar{x} \in U$, ssi pour toute suite $(x_n)_n$ avec $x_n \in U$ et $x_n \to \bar{x}$, $f(\bar{x}) \leq \liminf_n f(x_n)$. La fonction f est dite semi-continue inférieurement (s.c.i.) ssi elle est s.c.i. en tout $x \in U$.

On peut montrer (exercice) que f est s.c.i. en x tel que $f(x) < \infty$ ssi

$$\forall \varepsilon \geq 0 \, \exists \delta \geq 0 \text{ tel que } x \in B_\delta(\bar{x}) \Rightarrow f(\bar{x}) \leq f(x_n) + \varepsilon \, .$$

Exemple (Fonction indicatrice)*.* Soit $U \subset \mathbb{R}^d$, la fonction indicatrice de U est la fonction $I_U : \mathbb{R}^d \to [0, \infty]$ définie par

$$I_U \coloneqq \begin{cases} 0 & \text{si } x \in U \, , \\ +\infty & \text{si } x \notin U. \end{cases}$$

La fonction indicatrice d'un ensemble U est s.c.i. ssi l'ensemble U est fermé.

Définition 3.1.2 (Coercivité)**.** La fonction $f : U \subset \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ est coercive ssi pour tout $\alpha \in \mathbb{R}$, l'ensemble

$$\{x \in U ; \, f(x) \leq \alpha\}$$

est borné.

De façon équivalente, $f : \mathbb{R}^d \to \mathbb{R}$ est coercive ssi $f(x) \to +\infty$ lorsque $\|x\| \to \infty$.

Exemple.

1. L'indicatrice I_U est coercive ssi U est borné ;
2. S'il existe $p \geq 1$, $C > 0$ et $D \in \mathbb{R}$ tels que

$$f(x) \geq C\|x\|^p + D, \quad \forall \, x \in \mathbb{R}^d \, ,$$

   alors f est coercive. En général, si g est une fonction coercive sur $\mathbb{R}^d$ et $f \geq g$, alors f est aussi coercive ;
3. Si $f : U \subset \mathbb{R}^d \to \mathbb{R}$ est coercive et $g : U \subset \mathbb{R}^d \to \mathbb{R}$ est bornée inférieurement, alors $f + g$ est coercive (exercice).

Proposition 3.1.3 (Existence de minimiseurs globaux)**.** *Soit $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ et supposons qu'il existe $x \in \mathbb{R}^d$ tel que $f(x) < \infty$. Si f est coercive et s.c.i., alors il existe au moins un minimiseur global de f.*

Définition 3.1.4 (Fonctions convexes/strictement convexes/α-convexes)**.** Soit C un convexe et $\alpha > 0$. Une fonction $f : C \subseteq \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ est dite :

— **convexe** ssi

$$f((1-t)x+ty) \leq (1-t)f(x)+tf(y) \tag{3.1.1}$$

pour tout $t \in [0,1]$ et pour tout $x, y \in C$ tels que $f(x), f(y) < \infty$;

— **strictement convexe** ssi

$$f((1-t)x+ty) < (1-t)f(x)+tf(y) \tag{3.1.2}$$

pour tout $t \in (0,1)$ et pour tout $x, y \in C$ tels que $f(x), f(y) < \infty$ et $x \neq y$;

— α**-convexe** ssi

$$f((1-t)x+ty) \leq (1-t)f(x)+tf(y) - \alpha\frac{t(1-t)}{2}\|x-y\|^2 \tag{3.1.3}$$

pour tout $t \in [0,1]$ et pour tout $x, y \in C$ tels que $f(x), f(y) < \infty$.

Remarque 3.1.5. *α-convexe $\Rightarrow$ strictement convexe $\Rightarrow$ convexe.*

Rappel: Caractérisation différentielle de la convexité Si f est suffisamment régulière on peut characteriser les propriétés de convexité d'un point de vue différentielle. En particulier, soit C un ouvert convexe et $f : C \subseteq \mathbb{R}^d \to \mathbb{R}$:

1. si f est de classe C^1. Alors f est convexe ssi la fonction

$$g : t \in [0,1] \mapsto f((1-t)x+ty)$$

est convexe, ce qui implique

$$\begin{aligned} f \text{ convexe} \quad &\Leftrightarrow \quad f(y) \geq f(x) + \langle \nabla f(x), y-x\rangle \quad \forall x, y \in C \\ &\Leftrightarrow \quad \langle \nabla f(x) - \nabla f(y), x-y\rangle \geq 0 \quad \forall x, y \in C. \end{aligned}$$

De façon similaire,

$$\begin{aligned} f \text{ str. convexe} \quad &\Leftrightarrow \quad f(y) > f(x) + \langle \nabla f(x), y-x\rangle \quad \forall x, y \in C,\ x \neq y \\ &\Leftrightarrow \quad \langle \nabla f(x) - \nabla f(y), x-y\rangle > 0 \quad \forall x, y \in C,\ x \neq y; \end{aligned}$$

$$\begin{aligned} f\ \alpha\text{-convexe} \quad &\Leftrightarrow\ f(y) \geq f(x) + \langle \nabla f(x), y-x\rangle + \alpha\frac{\|y-x\|^2}{2} \quad \forall x, y \in C, \\ &\Leftrightarrow\ \langle \nabla f(x) - \nabla f(y), x-y\rangle \geq \alpha\|y-x\|^2 \quad \forall x, y \in C; \end{aligned}$$

2. si f est de classe C^2 :

$$f \text{ convexe} \quad \Leftrightarrow \quad D^2 f(x) \succeq 0 \quad \forall x \in C\,,$$

où on a identifié $D^2 f(x)$ avec la matrice Hessienne de f : c'est-à-dire la matrice symmétrique $(\partial_{x_i}\partial_{x_j} f)_{ij}$ dont les composantes sont les dérivés patielles de f par rapport à la base canonique de $\mathbb{R}^d$.

$$f \text{ str. convexe} \quad \Leftarrow \quad D^2 f(x) \succ 0 \quad \forall x \in C\,.$$

$$f \ \alpha\text{-convexe} \quad \Leftrightarrow \quad D^2 f(x) \succeq \alpha \mathrm{Id} \quad \forall x \in C\,.$$

Si f est une fonction convexe alors l'ensemble $\{x : f(x) \leq \alpha\}$ est convexe pour tout $\alpha \in \mathbb{R}$ (exercice). Ceci implique que

Proposition 3.1.6. *Soit C un convexe. L'ensemble des minimiseurs d'une fonction convexe $f : C \subseteq \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ est convexe ou vide. Si f est strictement convexe, cet ensemble est consitué d'un seul point ou vide.*

3.2 Méthodes de descente

Les méthodes de descente pour résoudre le problème de minimisation (3.0.1), peuvent s'écrire comme suit : étant donné $x^0 \in U$,

$$x^{n+1} = x^n + \tau^n d^n \tag{3.2.1}$$

pour tout $n \geq 0$, où $\tau^n > 0$ est le *pas* à l'itération n est $d^n \in \mathbb{R}^d$ est une *diréction de descente* qui est choisie de façon que $x^{n+1} \in U$ et

$$f(x^{k+1}) \leq f(x^k)\,.$$

Définition 3.2.1. Soit $f \in C^0(U, \mathbb{R})$. On appelle direction de descente en $x \in \mathbb{R}^d$, tout vecteur $d \in \mathbb{R}^d$ pour lequel il existe $\bar{\tau} > 0$ tel que pour tout $0 < \tau < \bar{\tau}$, $f(x + d\tau) < f(x)$.

Si $f : U \subset \mathbb{R}^d \to \mathbb{R}$ est différentiable en tout $x \in U$, avec U un ouvert, alors tout $d \in \mathbb{R}^d$ tel que $d \cdot \nabla f(x) < 0$ est une direction de descente en x. En effet, pour $\varepsilon > 0$ suffisamment petit, considerons la fonction

$$g : s \in (-\varepsilon, \varepsilon) \to g(s) \coloneqq f(x + sd)\,.$$

Cette fonction est différentiable en $s = 0$, et donc

$$g(s) = g(0) + g'(0)s + o(s)\,.$$

Par définition de g,

$$g'(0) = \frac{\mathrm{d}}{\mathrm{d}s}\Big|_{s=0} f(x + sd) = \langle \nabla f(x), d \rangle < 0$$

Donc

$$f(x+sd) = f(x) + \langle \nabla f(x), d\rangle s + o(s)\,.$$

Il existe $\delta > 0$ tel que pour $s < \delta$ on peut prendre $o(s) < -\frac{\langle \nabla f(x), d\rangle}{2} s$ et donc on a pour tout $0 < s < \delta$

$$f(x+sd) \leq f(x) + \frac{\langle \nabla f(x), d\rangle}{2} s < f(x). \tag{3.2.2}$$

La méthode de descente de gradient consiste à choisir $d^n = -\nabla f(x^n)$ dans (3.2.2).

3.2.1 Descente de gradient à pas constant

La méthode de descente la plus simple est celle à pas constante, c'est-à-dire $\tau^n = \tau$ pour tout $n \geq 0$. Ici on montre quelques résultats de convergence pour cette méthode.

Supposons que $f : \mathbb{R}^d \to \mathbb{R}$ est de classe C^1, que ∇f est L-lipschitzienne et que f est bornée inférieurement :

$$\inf_x f(x) > -\infty\,.$$

Alors on a pour tout $x, y \in \mathbb{R}^d$

$$f(y) = f(x) + \int_0^1 \langle \nabla f((1-t)x + ty), y - x\rangle \mathrm{d}t$$

En choisissant $y = x^{n+1}$ et $x = x^n$ on obtient :

$$\begin{aligned} f(x^{n+1}) = f(x^n) + \langle \nabla f(x^n), x^{n+1} - x^n\rangle + \\ \int_0^1 \langle \nabla f((1-t)x^n + tx^{n+1}) - \nabla f(x^n), x^{n+1} - x^n\rangle \mathrm{d}t \end{aligned} \tag{3.2.3}$$

On observe que

$$\left| \int_0^1 \langle \nabla f((1-t)x^n + tx^{n+1}) - \nabla f(x^n), x^{n+1} - x^n\rangle \mathrm{d}t \right| \leq L \frac{\|x^{n+1} - x^n\|^2}{2}$$

De plus pour la méthode considérée $x^{n+1} - x^n = -\tau \nabla f(x^n)$, donc

$$\begin{aligned} f(x^{n+1}) &\leq f(x^n) - \tau \|\nabla f(x^n)\|^2 + \tau^2 L \frac{\|\nabla f(x^n)\|^2}{2} \\ &= f(x^n) - \tau(1 - \frac{L\tau}{2}) \|\nabla f(x^n)\|^2 \end{aligned} \tag{3.2.4}$$

Donc si $\tau < 2/L$ la suite $f(x^n)$ est strictement décroissante. De plus,

$$\tau(1 - \frac{L\tau}{2})\|\nabla f(x^n)\|^2 \leq f(x^n) - f(x^{n+1})$$

et en sommant ces estimations pour $n = 0, \ldots, N$

$$\tau(1 - \frac{L\tau}{2})\sum_{n=0}^{N}\|\nabla f(x^n)\|^2 \leq f(x^0) - f(x^{N+1})$$

puisque f est bornée inférieurement $f(x^{N+1}) \geq m := \inf_x f(x)$, et en considérant la limite $N \to \infty$

$$\tau(1 - \frac{L\tau}{2})\sum_{n=0}^{\infty}\|\nabla f(x^n)\|^2 \leq f(x^0) - m$$

Donc $\nabla f(x^n) \to 0$ lorsque $n \to \infty$. On a donc montré le résultat suivant :

Théorème 3.2.2. *Soit $f \in C^1(\mathbb{R}^d, \mathbb{R})$ bornée inférieurement et supposons que ∇f est L-lipschitzienne. Alors la suite $(x^n)_n$ méthode de descente de gradient à pas constante $\tau < 2/L$ satisfait les propriétés suivantes :*

1. *$f(x^n)$ est strictement décroissante ;*
2. *$\nabla f(x^n) \to 0$ lorsque $n \to \infty$.*

Corollary 3.2.3. *Dans les mêmes hypothèses que dans le théorème 3.2.2 on a aussi que*

1. *si f est coercive alors il existe une sous-suite de $(x^n)_n$ qui converge vers x^* un point critique de f ;*
2. *si f est coercive et strictement convexe alors $x^n \to x^*$, l'unique minimum globale de f.*

Reprenons les arguments ci-dessus avec l'hypothèse supplémentaire que f est convexe et qu'elle admet au moins un minimiseur x^*. Dans ce cas, l'inegalité (3.2.4) implique

$$\begin{aligned} f(x^{n+1}) &\leq f(x^n) - \tau(1 - \frac{L\tau}{2})\|\nabla f(x^n)\|^2 \\ &\leq f(x^*) - \langle \nabla f(x^n), x^* - x^n\rangle - \tau(1 - \frac{L\tau}{2})\|\nabla f(x^n)\|^2 \end{aligned}$$

Si $\tau \leq 1/L$, on a

$$\begin{aligned} f(x^{n+1}) &\leq f(x^*) - \langle \nabla f(x^n), x^* - x^n \rangle - \frac{\tau}{2}\|\nabla f(x^n)\|^2 \\ &\leq f(x^*) + \frac{1}{2\tau}(\|x^n - x^*\|^2 - \|x^* - x^n - \tau \nabla f(x^n)\|^2) \\ &= f(x^*) + \frac{1}{2\tau}(\|x^n - x^*\|^2 - \|x^* - x^{n+1}\|^2) \end{aligned}$$

et en sommant sur n, on trouve

$$\sum_{k=1}^{n+1} f(x^k) \leq (n+1) f(x^*) + \frac{1}{2\tau}(\|x^0 - x^*\|^2 - \|x^* - x^{n+1}\|^2)$$

Puisque $f(x^k)$ est decroissante, $f(x^{n+1}) \leq f(x^k)$ pour tout $k \leq n+1$, donc

$$f(x^{n+1}) - f(x^*) \leq \frac{1}{2(n+1)\tau}\|x^0 - x^*\|^2$$

Ce type d'arguments peut se généraliser pour tout $\tau < 2/L$ (voir par exemple le Théorème 2.1.14 dans le livre "Lectures on convex optimization" de Yurii Nesterov). En plus si f est α convexe on a aussi

$$\frac{\alpha}{2}\|x^{n+1} - x^*\|^2 \leq f(x^{n+1}) - f(x^*) \leq \frac{1}{2\tau}(\|x^n - x^*\|^2 - \|x^* - x^{n+1}\|^2).$$

ce qui implique

$$\|x^{n+1} - x^*\|^2 \leq \frac{\tau}{\tau\alpha + 1}\|x^n - x^*\|^2$$

On a donc le résultat suivant :

Théorème 3.2.4. *Soit $f \in C^1(\mathbb{R}^d, \mathbb{R})$ convexe et supposons que ∇f et L-lipschitzienne. Soit x^* un minimiseur de f. Alors, il existe une constante $C > 0$ telle que*

$$f(x^k) - f(x^*) \leq \frac{C}{k}$$

en plus si f est α-convexe x^k converge vers x^* linéairement.

Remarque 3.2.5. *La méthode de descente de gradient à pas constant est équivalente à une méthode de point fixe du type étudié dans le chapitre précedent. Par exemple, si f est α-convexe, ∇f est α monotone et l'application $x \mapsto x - \tau \nabla f(x)$ est contractante pour tout $0 < \tau < 2\alpha/L^2$*

De plus, toujour dans le cas où la fonction minimisée est α-convexe, on peut quantifier l'erreur sur le minimiseur et sur le minimum par l'erreur sur les conditions d'optimalité $\nabla f(x) = 0$:

Lemme 3.2.6. *Soit $f \in C^1(\mathbb{R}^d, \mathbb{R})$ une fonction α-convexe et $x^* \in \mathbb{R}^d$ son unique minimiseur. Alors, pour tout $x \in \mathbb{R}^d$*

$$\|x - x^*\| \leq \frac{2\|\nabla f(x)\|}{\alpha}, \tag{3.2.5}$$

$$f(x) - f(x^*) \leq \frac{\|\nabla f(x)\|^2}{2\alpha}. \tag{3.2.6}$$

Démonstration. Pour montrer la prémière inégalité on utilise la caractérisation différentielle de l'α-convexité : pour tout $x \in \mathbb{R}^d$

$$f(x^*) \geq f(x) + \langle \nabla f(x), x - x^* \rangle + \alpha \frac{\|x - x^*\|^2}{2}, \tag{3.2.7}$$

Et puisque $f(x) \geq f(x^*)$ on montre le résultat par inégalité triangulaire. Pour montrer la deuxième inégalité on déduit de (3.2.7) que

$$f(x) - f(x^*) + \alpha \frac{\|x - x^*\|^2}{2} \leq \|\nabla f(x)\| \|x - x^*\| \leq \frac{\|\nabla f(x)\|^2}{2\alpha} + \alpha \frac{\|x - x^*\|^2}{2}.$$

□

3.2.2 Choix du pas

Pas optimal (linesearch) Une fois déterminée une diréction de descente, le pas optimale est la solution du problème de minimisation :

$$\tau^n = \arg\min_{s \geq 0} g(s) \quad g(s) \coloneqq f(x^n + s d^n)\,.$$

Dans le cas où f est strictemet convexe et coercive, par exemple, ce problème admet une unique solution. Si d'un coté ce choix est optimale pour une certaine diréction de descente, en général, le cout computationnel de la méthode est beaucoup plus élévé que dans le cas du pas constant, puisque à chaque itération on doit résoudre un problème d'optimisation où l'évaluation de la fonction a minimiser ainsi que son gradient equivaut à l'évaluation de f et de son gradient, respectivement.

Rebroussement (backtracking) Une strategie différente consiste en changer le pas selon le comportement locale de la fonction, simplement pour éviter des pas trop grands où la fonction varie rapidement. Le rebroussement linéaire consiste en réduire le pas jusqu'à que une condition de descente est vérifiée : par exemple, $\beta \in (0,1)$, on posera

$\tau^n = 1$
while $f(x^n + \tau^n d^n) \geq f(x^n)$ **do**
    $\tau^n = \beta\tau^n$
end while

Condition de Armijo Soit $g(s) \coloneqq f(x^n + sd^n)$. On observe que

$$g'(0) = \nabla f(x^n) \cdot d^n$$

Et donc la droite tangente à g en $s = 0$ est donnéee par

$$y(s) = f(x^n) + s\nabla f(x^n) \cdot d^n$$

Si f est convexe et différentiable en x^n, g est convexe et on a toujours $g(s) \geq y(s)$. Par contre, pour tout $c \in (0,1)$ il existe $s > 0$ tel que

$$g(s) < f(x^n) + cs\nabla f(x^n) \cdot d^n$$

En effet, puisque d^n est une diréction de descente $\nabla f(x^n) \cdot d^n < 0$ et donc

$$f(x^n + sd^n) = f(x^n) + s\nabla f(x^n) \cdot d^n + o(s)$$

et pour s suffisamment petit, on peut prendre $o(s) < -(1-c)f(x^n) \cdot d^n$, ce qui implique

$$f(x^n + sd^n) < f(x^n) + cs\nabla f(x^n) \cdot d^n$$

Définition 3.2.7. On dit que le pas τ^n satisfait la *condition d'Armijo*, si pour $c_A \in (0,1)$ donné,

$$f(x^n + \tau^n d^n) < f(x^n) + c_A\tau^n\nabla f(x^n) \cdot d^n \tag{3.2.8}$$

Pour satisfaire cette condition on peut utiliser une procedure de rebroussement, comme dans le cas précedent :

Algorithm 1 (Condition d'Armijo)

$\tau^n = 1$
while $f(x^n + \tau^n d^n) \geq f(x^n) + c_A \tau^n \nabla f(x^n) \cdot d^n$ **do**
  $\tau^n = \beta \tau^n$
end while

Conditions de Wolfe Dans certain cas des pas trop petits peuvent emmener à un echec de convergence de l'algorithme de descente à pas variable :

Exemple. Supposons de vouloir minimiser $f(x) = x^2/2$ avec direction de descente $d^n = -f'(x^n)/|f'(x^n)| = -\text{sign}(f'(x^n))$ et $\tau^n = 1/2^{n+1}$. Si $x^0 = 2$ on peut verifier que $x^n \to 1$ alors que 0 est l'unique minimiseur de f.

Puisque d^n est une direction de descente $g'(0) < 0$. Par contre si f est régulière, et pour le pas optimal τ^*, $g'(\tau^*) = 0$. Pour imposer une borne inferieur sur τ^n on demande alors que $g'(\tau^n)$ soit suffisamment plus large de $g'(0)$, c'est-à-dire on demande $g'(\tau^n) > cg'(0)$ pour $c \in (0, 1)$.

Définition 3.2.8. On dit que le pas τ^n satisfait les *conditions de Wolfe*, si pour $0 < c_A < c_W < 1$ donnés,

$$f(x^n + \tau^n d^n) < f(x^n) + c_A \tau^n \nabla f(x^n) \cdot d^n \tag{3.2.9}$$

$$\nabla f(x^n + \tau^n d^n) \cdot d^n > c_W \nabla f(x^n) \cdot d^n \,. \tag{3.2.10}$$

Remarque 3.2.9. *La condition de Wolfe* (3.2.10) *n'est pas vérifiée pour* $\tau^n = 0$ *: si* f *est de classe* C^1*, sont gradient est continu et cela implique que la condition n'est pas vérifiée pour des pas trop petits. De plus, si* $f : \mathbb{R}^d \to \mathbb{R}$ *est bornée inferieurement, coercive et de classe* C^1*, il existe toujours un ouvert de* $\mathbb{R}$ *où la condition est vérifiée.*

Un algorithme possible pour determiner un pas admissible respectant les conditions de Wolfe est le suivant :

Remarque 3.2.10. *On observe que :*

1. *L'algorithme 2 est basée sur l'idée (simple) de recherche par dichotomie ;*
2. *On peut montrer qu'il converge après un nombre fini d'itérations - mais ceci peut etre élévé en pratique ;*
3. *Des alternative plus performantes sont possibles : par exemple on peut utiliser à chaque itération les approximation précedentes grace à des interpolations polynomiales de* f *...*

Algorithm 2 (Conditions de Wolfe)

$\tau^n = 1$, $\tau^+ = \infty$, $\tau^- = 0$
while Conditions de Wolfe pas vérifiées **do**
  if $f(x^n + \tau^n d^n) \geq f(x^n) + c_A \tau^n \nabla f(x^n) \cdot d^n$ **then**
    $\tau^+ = \tau^n$, $\tau^n = (\tau^+ + \tau^-)/2$
  else if $\nabla f(x^n + \tau^n d^n) \cdot d^n \leq c_W \nabla f(x^n) \cdot d^n$ **then**
    $\tau^- = \tau^n$,
    if $\tau^+ < \infty$ **then**
      $\tau^n = (\tau^+ + \tau^-)/2$
    else
      $\tau^n = 2\tau^-$
    end if
  end if
end while

3.2.3 Descente de gradient preconditionné à rebroussement

Dans cette section on s'intéresse à des méthode de descente où la direction de descente à l'itération n est construite en multipliant le gradient par une matrice symmétrique définie positive $B^n \in \mathcal{M}_d(\mathbb{R})$, c'est-à-dire on choisi :

$$d^n = -B^n \nabla f(x^n).$$

Si f est différentiable, par l'équation (3.2.2), on voit facilement que d^n est une diréction de descente en x^n. Nous considerons l'algorithme suivant : étant donné $x^0 \in \mathbb{R}^d$, pour $n \geq 0$,

$$\begin{aligned} d^n &= -B^n \nabla f(x^n) \\ \tau^n &\leftarrow \text{Algorithme 1 avec } c_A \in (0, 1/2] \\ x^{n+1} &= x^n + \tau^n d^n \end{aligned} \tag{3.2.11}$$

où on calcule τ^n grace à l'algorithme de rebroussement présenté dans la séction precedente et donc vérifie la condition d'Armijo.

Lemme 3.2.11. *Soit $f \in C^2(\mathbb{R}^d)$ et $B^n = (A^n)^2$ où A^n est une matrice symmétrique définie-positive, telle que pour tout $x \in \mathbb{R}^d$,*

$$A^n D^2 f(x) A^n \preceq L\mathrm{Id}.$$

Si $0 < \tau^n < 1/L$, alors τ^n il vérifie la condition d'Armijo (3.2.9) avec $c_A \in (0, 1/2]$. En particulier, en particulier le pas τ^n obtenu par l'alogrithme

de rebroussement 1 avec $\beta < 1$ vérifie :

$$\tau^n \geq \min(1, \frac{\beta}{L})$$

Démonstration. Par le developpement de Taylor de f dans x^n avec reste de Lagrange, il existe $t \in [0, s]$ tel que

$$f(x^n + sd^n) = f(x^n) + s\langle \nabla f(x^n), d^n \rangle + \frac{s^2}{2}\langle D^2 f(x^n + td^n)d^n, d^n \rangle$$

De plus,

$$\langle D^2 f(x^n + td^n)d^n, d^n \rangle \leq -L\langle d^n, \nabla f(x^n) \rangle$$

Donc

$$f(x^n + sd^n) \leq f(x^n) + s(1 - \frac{Ls}{2})\langle d^n, \nabla f(x^n) \rangle \,.$$

Donc la condition est vérifié si $s < 1/L$. Si τ^n ne vérifie pas la condition d'Armijo on devra faire une itération supplémentaire de l'algorithme de rebroussement, d'où la borne sur τ^n. □

Remarque 3.2.12. *Lemme 3.2.11 nous permet de quantifier le nombre d'itérations nécessaires pour que l'algorithme 1 converge. En effet, si on itialise l'algorithme avec l'initialisation $\tau^n = 1$ à chaque itération, on τ^n vérifie la condition d'Armijo après k itérations avec $\beta^k < 1/L$, c'est-à-dire $k > \log(L)/\log(1/\beta)$.*

Dans les hypothèses du lemme précedent, si τ^n vérifie la condition d'Armijo alors

$$\begin{aligned} f(x^{n+1}) &\leq f(x^n) - \tau^n c_A \langle d^n, \nabla f(x^n) \rangle \\ &= f(x^n) - \tau^n c_A \|A^n \nabla f(x^n)\|^2 \\ &\leq f(x^n) - \delta c_A \|A^n \nabla f(x^n)\|^2 \end{aligned} \tag{3.2.12}$$

où $\delta \coloneqq \min(1, \frac{\beta}{L})$. En raisonnant comme pour la preuve du Théorème 3.2.2, ceci implique que si f est bornée inférieurement

$$A^n \nabla f(x^n) \to 0 \quad \text{lorsque} \quad n \to \infty.$$

Supposons maintenant que, en plus, f est coercive et qu'il existe $\alpha > 0$ tel que

$$A^n D^2 f(x) A^n \succeq \alpha \mathrm{Id}.$$

Alors, la fonction $y \mapsto g(y) \coloneqq f(A^n y)$ est α-convexe et coercive : elle admet un unique minimiseur $y^* \in \mathbb{R}^d$ et par le lemme (3.2.6), pour tout $y \in \mathbb{R}^d$,

$$g(y) - g(y^*) \leq \frac{\|\nabla g(y)\|^2}{2\alpha}\,.$$

La fonction f admet aussi un unique minimiseur $x^* = (A^n)^{-1}y^*$ et puisque $\nabla g(y) = A^n \nabla f(A^n y)$ on a, pour tout $x \in \mathbb{R}^d$,

$$f(x) - f(x^*) \leq \frac{\|A^n \nabla f(x)\|^2}{2\alpha}\,.$$

Donc, par (3.2.12) :

$$f(x^{n+1}) - f(x^*) \leq (1 - 2\alpha c_A \delta)(f(x^n) - f(x^*))$$

On a donc montré le résultat suivant :

Théorème 3.2.13. *Soit $f \in C^2(\mathbb{R}^d)$ coercive et bornée inferieurement. Soit $A^n \in \mathcal{M}_d(\mathbb{R})$ une matrice symmétrique définie-positive pour tout $n \geq 0$, et suppose qu' ils existent $0 < \alpha < L$ tels que pour tout $n \geq 0$ et pour tout $x \in \mathbb{R}^d$*

$$\alpha \mathrm{Id} \preceq A^n D^2 f(x) A^n \preceq L\mathrm{Id}\,.$$

Alors les itérées $(x^n)_n$ de l'algorithme (3.2.11) convergent vers l'unique minimiseur de f et de plus $f(x^n)$ converge linéairement vers $f(x^*)$:

$$f(x^{n+1}) - f(x^*) \leq \left[1 - 2\alpha c_A \min(1, \frac{\beta}{L})\right](f(x^n) - f(x^*))$$

Remarque 3.2.14. *On peut accelerer la convergence de la méthode si on choisi A^n de sorte que $L/\alpha \approx 1$. En particulier, le choix $(A^n)^2 = [D^2 f(x^n)]^{-1}$ donne lieu à la méthode de Newton.*

3.3 Méthode de Newton

Soit $f : \mathbb{R}^d \to \mathbb{R}$ de classe C^2. L'idée de la méthode de Newton est la suivante : à l'itération $n+1$ de la méthode, on définit x^{n+1} comme le minimiseur de la fonction

$$f(x^n) + \langle \nabla f(x^n), x - x^n\rangle + \frac{1}{2}\langle D^2 f(x^n)(x - x^n), x - x^n\rangle.$$

Si $D^2 f(x^n)$ est définie-positive ce problème admet une unique solution et en particulier :

$$d^n = -[D^2 f(x^n)]^{-1} \nabla f(x^n)$$
$$x^{n+1} = x^n + d^n$$

La méthode qu'on obtient équivaut à appliquer la méthode de Newton introduite dans le chapitre précedent aux conditions d'optimalité du problème d'origine, c'est-à-dire au système nonlinéaire

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

Le meme résultat de convergence quadratique est donc valide dans ce cas. Plus précisement, supposons que $f : \mathbb{R}^d \to \mathbb{R}$ est de classe C^2 et telle que $D^2 f$ est γ-lipschitzienne, et que x^* est un minimum local de f tel que

$$D^2 f(x^*) \succeq \alpha \mathrm{Id},$$

avec $\alpha > 0$.

Lemme 3.3.1. *Soit $f \in C^2(\mathbb{R}^d)$ telle que $D^2 f$ est γ-lipschitzienne. Alors, pour tout $x, y \in \mathbb{R}^d$*

$$D^2 f(x) - \gamma \|x - y\| \mathrm{Id} \preceq D^2 f(y) \preceq D^2 f(x) + \gamma \|x - y\| \mathrm{Id}$$

Démonstration. On pose $G \coloneqq D^2 f(x) - D^2 f(y)$, et on observe que $|||G||| \leq \gamma \|x - y\|$. Puisque G est une matrice symmétrique et $|||\cdot|||$ dénote la norme induite par la norme euclidienne, si on denote par λ_i ces valeur propres on obtient :

$$|\lambda_i| \leq \max_i |\lambda_i| = |||G||| \leq \gamma \|x - y\|,$$

ce qui est équivalent à $-\gamma \|x - y\| \mathrm{Id} \preceq G \preceq \gamma \|x - y\| \mathrm{Id}$. □

Le dernier lemme implique que

$$D^2 f(x) \succeq D^2 f(x^*) - \gamma \|x - x^*\| \mathrm{Id} \succeq (\alpha - \gamma \|x - x^*\|) \mathrm{Id}$$

et donc $D^2 f(x)$ reste inversible dans un voisinage de x^*, ou plus précisement pour tout $x \in \mathbb{R}^d$ tel que $\|x - x^*\| < \alpha/\gamma$. Dans le meme voisinage, on a aussi

$$\left|\left|\left|[D^2 f(x)]^{-1}\right|\right|\right| \leq (\alpha - \gamma \|x - x^*\|)^{-1}. \tag{3.3.1}$$

Pour determiner l'erreur de convergence locale, on procède comme dans le chapitre précedent : si x^n est suffisamment près de x^*, $D^2 f(x^n)$ est inversible et on peut écrire

$$x^{n+1} = x^n - [D^2 f(x^n)]^{-1} \nabla f(x^n)$$

donc

$$\begin{aligned} x^{n+1} - x^* &= x^n - x^* - [D^2 f(x^n)]^{-1} \nabla f(x^n) \\ &= x^n - x^* - [D^2 f(x^n)]^{-1} \int_0^1 D^2 f(x^* + s(x^n - x^*))(x^n - x^*) \mathrm{d}s \\ &\leq \left\| [D^2 f(x_n)]^{-1} \right\| \frac{\gamma}{2} \|x^n - x^*\|^2 . \end{aligned}$$

Par (3.3.1) on obtient :

$$\|x^{n+1} - x^*\| \leq \frac{\gamma \|x^n - x^*\|^2}{2(\alpha - \gamma \|x^n - x^*\|)}$$

Si $\|x^n - x^*\| \leq 2\alpha/(3\gamma)$ le terme à droite est borné par $\|x^n - x^*\|$ et donc $\|x^{n+1} - x^*\| \leq 2\alpha/(3\gamma)$. On a montré le résultat suivant :

Théorème 3.3.2. *Soit $f \in C^2(\mathbb{R}^d)$ telle que $D^2 f$ est γ-lipschitzienne et soit x^* un minimum local de f tel que*

$$D^2 f(x^*) \succeq \alpha \mathrm{Id},$$

avec $\alpha > 0$. Alors, si $\|x^0 - x^*\| \leq 2\alpha/(3\gamma)$ les itérées de la méthodes de Newton $(x^n)_n$ sont bien-définies et elles convergent quadratiquement vers x^*.

Remarque 3.3.3. *La méthode de Newton souffre de deux inconvénients majeurs :*

1. *La convergence (quadratique) est seulement locale. Ceci est le cas même si la fonction qu'on minimise est α-convexe (Figure 3.1), ce qui est pas le cas avec la méthode du gradient avec un pas suffisamment petit. Généralement, il est nécessaire d'utiliser une méthode plus robuste d'ordre plus bas pour approcher un minimiseur, mais il est recommendable d'utiliser Newton quand on est suffisamment près du minimiseur.*
2. *À chaque itération on doit calculer $D^2 f(x^n)$ et résoudre un système linéaire : les deux opérations peuvent etre très couteuses du point de vue computationnel.*

Une stratégie possible pour améliorer le comportement globale de l'algorithme consiste en utiliser un pas variable : l'algorithme qu'on obtient est connu sous le nom de *méthode de Newton amortie* (damped Newton method). Par exemple, on pose :

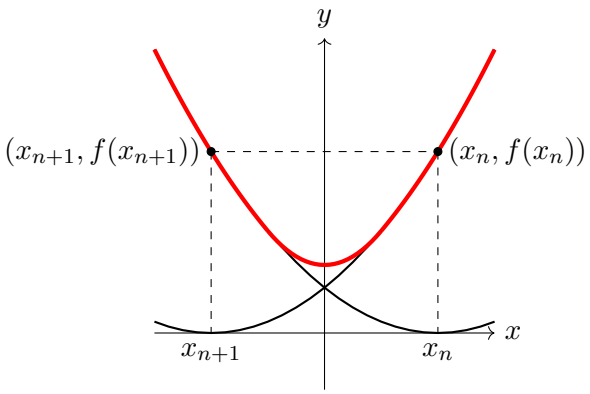

FIGURE 3.1 – Un exemple où la méthode de Newton ne converge pas pour la minimisation d'une fonction α-convexe (courbe rouge). Cette fonction coincide dehors d'un voisinage de zéro avec deux paraboles dont les minima sont situés en x_n et $x_{n+1} = -x_n$ et telle que $f(x_n) = f(x_{n+1})$.

$$\begin{aligned} &d^n = -[D^2 f(x^n)]^{-1} \nabla f(x^n) \\ &\tau^n \leftarrow \text{Rebroussement (e.g., algorithme 1)} \\ &x^{n+1} = x^n + \tau^n d^n \end{aligned} \tag{3.3.2}$$

Il faut observer que $\tau^n = 1$ est l'unique choix qui permet d'obtenir la convergence quadratique près d'un minimiseur ! Pour cette raison, on choisira $\tau^n = 1$ pour initialiser l'algorithme de rebroussement.

3.4 Quasi-Newton

L'idée des méthodes de quasi-Newton est de remplacer $[D^2 f(x^n)]^{-1}$ par une version approchée. À l'itération $n+1$ de la méthode, on définit x^{n+1} comme le minimiseur de la fonction

$$f(x^n) + \langle \nabla f(x^n), x - x^n \rangle + \frac{1}{2} \langle H^n (x - x^n), x - x^n \rangle.$$

où $H^n \in \mathcal{M}_d(\mathbb{R})$ est construite de façon itérative de sorte que

$$H^n \approx D^2 f(x^n).$$

Si $H^n \succ 0$, on obtient donc

$$x^{n+1} = x^n - [H^n]^{-1} \nabla f(x^n)\,.$$

Comme pour la méthode de Broyden, pour construire H^n, on demande que $H^n = (H^n)^T \succ 0$ et

$$H^n(x^n - x^{n-1}) = \nabla f(x^n) - \nabla f(x^{n-1})\,.$$

De façon équivalente, on peut définir $B^n = (H^n)^{-1}$ de sorte que $B^n = (B^n)^T \succ 0$ et

$$x^n - x^{n-1} = B^n(\nabla f(x^n) - \nabla f(x^{n-1}))\,.$$

Ces conditions ne déterminent pas H^n ni B^n de façon unique. Souvent, on impose en plus une condition sur le rang de $H^{n+1} - H^n$. La méthode considérée comme la plus robuste d'un point de vue computationnel est la méthode BFGS (Broyden-Fletcher-Goldfarb-Shanno) où $H^{n+1} - H^n$ est une matrice de rang 2, et l'inverse peut être calculé explicitement grâce à la formule de Woodbury.

On observe que :

1. l'algorithme rentre dans le cadre des méthodes de gradient préconditionné : on peut appliquer le résultat de convergence dans le Théorème 3.2.13 ;
2. on peut montrer un résultat de convergence locale superlinéaire ;
3. le comportement de l'algorithme en termes de convergence globale ne peut pas être meilleur que celui de Newton : on utilise généralement une version amortie qui peut être formulée comme pour Newton (3.3.2).