Méthodes numériques : optimisation
L3 2016–2017 — 2e semestre

Amic Frouvelle

5 mai 2017

Table des matières

Introduction

Ces notes de cours sont une introduction aux méthodes numériques de résolution de problèmes d'optimisation. Ces problèmes d'optimisation sont omniprésents dès la modélisation dans l'essentiel des branches des applications des mathématiques, telles que la physique (minimisation d'énergie), l'industrie (optimisation de la qualité de production), l'économie (optimisation de plan de production et de distribution), la finance (optimisation de portefeuille), le traitement d'image, la biologie, etc.

D'autre part, de nombreux problèmes ne se formulent pas dès la modélisation par des problèmes d'optimisation, mais leur résolution numérique se fait souvent via la transformation du problème en un problème d'optimisation. L'exemple le plus classique, et qui servira de base à de nombreuses reprises dans ce cours est le suivant : si b est un vecteur de $\mathbb{R}^n$ et $A \in S_n^+(\mathbb{R})$ une matrice symétrique définie positive, alors résoudre $Ax + b = 0$ est équivalent à minimiser $\frac{1}{2}\langle x, Ax\rangle + \langle b, x\rangle$ sur $\mathbb{R}^n$.

L'objectif de ce cours est de donner un aperçu de certaines méthodes de résolution de problèmes de minimisation continue non-linéaires. Il s'agit des méthodes dites « méthodes de descente », qui reposent sur le principe suivant : on part d'un point $x \in \mathbb{R}^n$, et on le déplace pas à pas en prenant soin que la direction dans laquelle on se déplace à chaque étape est bien une direction de descente (dans laquelle la fonction à minimiser diminue localement). Comme on cherche à utiliser les informations locales au voisinage du point à chaque étape, on aura besoin de régularité sur la fonction, pour pouvoir par exemple utiliser une approximation numérique de son gradient. En effet, on verra que le gradient est un vecteur dont la norme est la plus grande pente et l'orientation est dirigée par cette direction de plus grande pente, donc un vecteur opposé au gradient donne une direction de descente. Suivant le type de problème que l'on cherche à étudier, on verra que certaines méthodes sont plus efficaces que d'autres.

Il existe d'autres types de problèmes d'optimisation qui se traitent différemment et qui ne sont pas l'objet de ce cours : problèmes d'optimisation linéaire (qui se traitent avec des méthodes différentes, comme la méthode du simplexe), problèmes d'optimisation discrète (lorsque les variables ne sont pas continues, qui sont en général plus difficiles).

Le plan du cours est le suivant : on commencera dans un premier chapitre par présenter des méthodes numériques pour résoudre des problèmes d'optimisation continue à une dimension (sur un intervalle de $\mathbb{R}$). On étudiera ensuite dans le deuxième chapitre les méthodes de descente de gradient (lorsque la direction de descente est simplement donnée par l'opposé du gradient). Le troisième chapitre portera sur les méthodes de gradient conjugué, et le dernier chapitre concernera la méthode de Newton et les méthodes dites « Quasi-Newton ».

Ces notes de cours sont inspirées du cours donné par François-Xavier Vialard les années précédentes, dont les notes sont disponibles sur sa page web : `http://www.ceremade.dauphine.fr/~vialard/CoursOptimisationMai.pdf`.

Pour le lecteur intéressé, une référence très complète est le livre (en anglais) de Jorge Nocedal et Stephen J. Wright : « Numerical Optimisation », présenté sur la page `http://users.iems.northwestern.edu/~nocedal/book/num-opt.html`.

Chapitre 1

Optimisation continue en une dimension

On considère une fonction f continue d'un intervalle $I \subset \mathbb{R}$ à valeurs réelles et on s'intéresse au problème suivant :

$$\inf_{x \in I} f(x), \tag{1.1}$$

c'est-à-dire que l'on cherche un réel $x \in I$ tel que pour tout $y \in I$, on ait $f(x) \leqslant f(y)$. On s'intéressera également aux solutions locales de (1.1), pour lesquelles on a $f(x) \leqslant f(y)$ sur un voisinage de x dans I.

1.1 Généralités

On rappelle sans démonstration quelques résultats de base sur les fonctions réelles, mais c'est un excellent exercice de se souvenir comment on les montre.

Proposition 1.1. *Conditions d'existence de minimiseurs.*

- *Si I est compact, alors il existe au moins une solution au problème* (1.1).
- *Si I est fermé et f est coercive sur I (c'est-à-dire que $f(x) \to +\infty$ lorsque $|x| \to \infty$ tout en restant dans I), alors il existe au moins une solution au problème* (1.1).
- *Si f est dérivable en un point x de l'intérieur de I qui est un minimiseur, alors $f'(x) = 0$.*
- *Si f est deux fois dérivable en un tel point, alors $f''(x) \geqslant 0$.*
- *Si f est deux fois dérivable en un point x de l'intérieur de I et si $f'(x) = 0$ et $f''(x) > 0$, alors f admet un minimum strict local au voisinage de x.*
- *Si f est convexe, tout point de minimum local est un minimum global.*
- *Si f est strictement convexe, alors il existe au plus une solution au problème* (1.1).

À part en utilisant de la convexité, il est difficile d'obtenir des conditions suffisantes qui garantissent l'existence d'un unique minimiseur. On cherchera donc souvent des minimiseurs locaux, et on essaye dans ce cas de se restreindre à un sous-intervalle de I sur lequel il n'y a pas plus d'un minimum local. Attention cependant, ce n'est pas toujours possible, un exemple de cas pathologique est la fonction $x \mapsto x^4 + x^2 \sin^2 \frac{1}{x}$, qui a un minimum local strict en 0, mais qui n'est pas isolé (quel que soit le voisinage de 0 qu'on considère, il contient au moins un autre minimum local).

La plupart des méthodes de résolution de problèmes d'optimisation (en tout cas toute celles que l'on va étudier) consistent à essayer de faire converger une suite de points vers un minimum local. Pour pouvoir comparer ces méthodes les unes aux autres, trois caractéristiques principales se distinguent.

— La robustesse : on veut que la méthode fonctionne dans le plus grand nombre de cas possible, pour le plus grand ensemble possible de points initiaux.
— La précision : on veut que la méthode, malgré les erreurs d'approximation et les erreurs d'arrondi accumulées lors de l'exécution de l'algorithme par un ordinateur (qui fait donc les calculs avec une précision donnée), donne un résultat proche du point de minimum recherché.
— La vitesse : on veut que la méthode converge rapidement vers un point de minimum.

Ces trois caractéristiques sont souvent difficiles à satisfaire en même temps : les méthodes les plus robustes seront souvent plus lentes... Elles peuvent être également subjectives, suivant les problèmes auxquels on s'intéresse, on pourra par exemple privilégier la précision à la vitesse ou le contraire. Enfin on s'intéressera à des caractéristiques analogues (par exemple le coût total d'exécution, pas seulement en temps, mais peut-être en espace mémoire, etc.).

On présente donc d'abord des notions de mesure de la « vitesse » à laquelle les suites convergent.

Définition 1.1. *Ordre de convergence d'une suite.*

Soit (x_k) une suite réelle, et x_∞ un réel.

— *On dit que (x_n) converge linéairement (ou à l'ordre 1) vers x_∞ s'il existe des constantes $C > 0$ et $\alpha \in]0,1[$ telles qu'à partir d'un certain rang, on ait :*

$$|x_k - x_\infty| \leqslant C\alpha^k. \tag{1.2}$$

— *On note r la borne inférieure des α qui satisfont la condition ci-dessus, et r est appelé taux de convergence linéaire de la suite.*
— *Si $r = 0$, on dit que la convergence est superlinéaire.*
— *Enfin, s'il existe des constantes $C > 0$, $\alpha \in]0,1[$ et $\beta > 1$ telles qu'à partir d'un certain rang, on ait :*

$$|x_k - x_\infty| \leqslant C\alpha^{\beta^k}, \tag{1.3}$$

alors on dit que la convergence est d'ordre au moins β. La borne supérieure de ces β est appelé l'ordre de convergence de la suite.

L'ordre de convergence est lié à la vitesse à laquelle le nombre de chiffres exacts augmente dans l'écriture décimale de x_n (par rapport aux décimales de x_∞). Dans le cas de la convergence linéaire, le nombre de décimales s'accroit d'un facteur constant à chaque étape (ce facteur dépendant de r : plus r est petit, et plus le nombre de décimales exactes ajoutées à chaque étape est grand). Dans le cas de convergence d'ordre β avec $\beta > 1$, le nombre de décimales exactes est multiplié par un facteur β à chaque étape.

Pour comparer deux méthodes, on comparera donc d'abord leur ordre de convergence (par abus de langage, on appelle ordre de convergence d'une méthode l'ordre de convergence des suites qu'elle génère). Si les deux sont linéaires, alors on peut comparer leur taux de convergence linéaire pour mesurer leur rapidité de convergence (plus le taux est petit, plus la vitesse est grande).

En pratique pour estimer la vitesse de convergence de suites, on ne passe pas directement par la définition 1.1, on utilisera souvent des critères qui permettent d'obtenir des résultats sur la vitesse de convergence. Le premier type de critère concerne le cas où on connaît la valeur de la limite x_∞.

Proposition 1.2. *Critère pour l'ordre de convergence d'une suite.*

Soit x_k une suite et x_∞ un réel.

— *S'il existe $\alpha \in]0,1[$ tel que pour tout k à partir d'un certain rang, on ait :*

$$|x_{k+1} - x_\infty| \leqslant \alpha |x_k - x_\infty|, \tag{1.4}$$

alors (x_k) converge linéairement vers x_∞, à un taux inférieur ou égal à α.

— *S'il existe $\beta > 1$ et $\gamma > 0$ tels que, à partir d'un certain rang k_0, on ait*

$$|x_{k+1} - x_\infty| \leqslant \gamma |x_k - x_\infty|^\beta, \tag{1.5}$$

et que pour un certain $k_1 \geqslant k_0$, on a $\gamma |x_{k_1} - x_\infty|^{\beta-1} < 1$ (ce qui est toujours le cas si on sait déjà par exemple que la suite converge), alors la suite (x_k) converge vers x_∞ avec un ordre de convergence supérieur ou égal à β.

La démonstration est laissée en exercice, en particulier pour le deuxième point :

Exercice 1.1. *Montrer que si il existe un $k_1 \geqslant k_0$, pour lequel on a $\gamma |x_{k_1} - x_\infty|^{\beta-1} < 1$, alors on obtient pour $k \geqslant k_1$:*

$$|x_k - x_\infty| \leqslant \gamma^{\frac{-1}{\beta-1}} \left(\gamma^{\frac{1}{\beta-1}} |x_{k_1} - x_\infty|\right)^{\beta^{k-k_1}},$$

et que cela donne bien une estimation de la forme (1.3)*, avec $\alpha = \left(\gamma^{\frac{1}{\beta-1}} |x_{k_1} - x_\infty|\right)^{\beta^{-k_1}} < 1$.*

Indication : *poser $\varepsilon_k = \gamma^{\frac{1}{\beta-1}} |x_k - x_\infty|$, puis observer que ε_k est décroissante à partir du rang k_1, et que $\varepsilon_{k+1} \leqslant (\varepsilon_k)^\beta$ pour $k \geqslant k_1$, pour pouvoir obtenir une estimation simple de ε_k par récurrence.*

Il faut savoir que la définition de l'ordre de convergence n'est pas parfaitement figée, suivant les références certains auteurs prennent pour définition le critère ci-dessus. En pratique (comme on le verra dans la proposition 1.4) il y a des cas où les suites satisfont une des estimations (1.2)-(1.3) mais pas le critère correspondant ci-dessus.

Dans beaucoup de cas, on souhaitera obtenir à la fois la convergence et l'ordre de convergence en estimant les différence entre deux points à deux étapes successives, sans connaître a priori la limite. Les critères suivants permettent d'obtenir cela.

Proposition 1.3. *Critères de convergence et d'estimation d'ordre. Soit (x_k) une suite réelle.*

— *Supposons qu'il existe $\alpha \in]0, 1[$ tel que pour tout k à partir d'un certain rang, on ait*

$$|x_{k+1} - x_k| \leqslant \alpha |x_k - x_{k-1}|.$$

Alors la suite (x_k) converge vers une limite x_∞ et la convergence est linéaire, le taux de convergence linéaire étant inférieur ou égal à α.

— *Supposons qu'il existe $\beta > 1$, $\gamma > 0$ et $k_0 \in \mathbb{N}$ tel que pour tout $k \geqslant k_0$ à partir d'un certain rang, on ait*

$$|x_{k+1} - x_k| \leqslant \gamma |x_k - x_{k-1}|^\beta.$$

On suppose de plus qu'il existe $k_1 \geqslant k_0$ tel que $\gamma |x_{k_1} - x_{k_1-1}|^{\beta-1} < 1$.
Alors la suite (x_k) converge vers une limite x_∞ et la convergence est d'ordre supérieur ou égal à β.

Démonstration. Notons d'abord que par récurrence, on a $|x_{i+1} - x_i| \leqslant \alpha^{i-k} |x_{k+1} - x_k|$ dès que $i \geqslant k \geqslant k_0$.

Pour la première partie, supposons que l'estimation soit vraie à partir du rang k_0. On a alors

$$x_k = x_{k_0} + \sum_{i=k_0}^{k-1} (x_{i+1} - x_i).$$

Pour chaque terme de la somme, on a $|x_{i+1} - x_i| \leqslant \alpha^{i-k_0} |x_{k_0+1} - x_{k_0}|$, qui est le terme général d'une série géométrique convergente. Donc la série dont la somme partielle est donnée ci-dessus

est absolument convergente, et converge vers une limite x_∞, c'est-à-dire que (x_k) converge bien vers x_∞. Ensuite on a pour tout $k \geqslant k_0$

$$|x_\infty - x_k| = \Big|\sum_{i=k}^{\infty}(x_{i+1}-x_i)\Big| \leqslant \sum_{i=k}^{\infty}|x_{i+1}-x_i|$$
$$\leqslant \sum_{i=k}^{\infty}\alpha^{i-k}|x_{k+1}-x_k| = \frac{1}{1-\alpha}|x_{k+1}-x_k| \qquad \leqslant \frac{\alpha^{k-k_0}}{1-\alpha}|x_{k_0+1}-x_{k_0}| = C\alpha^k,$$

avec $C = \alpha^{-k_0}(1-\alpha)|x_{k_0+1}-x_{k_0}|$, qui est exactement l'estimation (1.2) de la définition. On remarque qu'au cours du calcul, on a obtenu que $|x_\infty - x_k| \leqslant \frac{1}{1-\alpha}|x_{k+1}-x_k|$, ce qui fait le lien entre les deux types de critères : dans le cas de la convergence linéaire, si on a une bonne estimation des différences des itérées, cela se traduit en une bonne estimation de l'erreur totale.

Pour la deuxième partie, on observe d'abord par récurrence, en posant $\alpha = \gamma|x_{k_1}-x_{k_1-1}|^{\beta-1} < 1$, que la suite $(|x_k - x_{k-1}|)_{k\geqslant k_1}$ est décroissante et que $\gamma|x_k - x_{k-1}|^{\beta-1} \leqslant \alpha < 1$ pour tout $k \geqslant k_1$. On obtient donc $|x_{k+1}-x_k| \leqslant \alpha|x_k - x_{k-1}|$ et comme $\alpha < 1$ on applique la première partie pour avoir la convergence de (x_k) vers un réel (x_∞) (la convergence étant au moins linéaire). Il ne nous reste plus qu'à estimer proprement $|x_{k+1}-x_k|$ d'après la remarque précédente.

En posant $\delta_k = \gamma^{\frac{1}{\beta-1}}|x_k - x_{k-1}|$, on obtient

$$\delta_{k+1} \leqslant \gamma^{\frac{1}{\beta-1}} \cdot \gamma|x_k - x_{k-1}|^\beta = (\delta_k)^\beta,$$

et par récurrence on en déduit que $\delta_k \leqslant \delta_{k_1}^{\beta^{k-k_1}} = \left(\alpha^{\frac{1}{\beta-1}}\right)^{\beta^{k-k_1}}$. On obtient donc

$$|x_\infty - x_k| \leqslant \frac{1}{1-\alpha}|x_{k+1}-x_k| = \frac{1}{(1-\alpha)\gamma^{\frac{1}{\beta-1}}}\delta_{k+1} \leqslant C\left(\alpha^{\frac{\beta^{1-k_1}}{\beta-1}}\right)^{\beta^k},$$

qui est bien de la forme correspondant à la définition 1.1, puisque $\alpha^{\frac{\beta^{-k_1}}{\beta-1}} < 1$. □

Remarque 1.1. *Dans le cadre de ce critère, l'estimation $|x_k - x_\infty| \leqslant \frac{1}{1-\alpha}|x_{k+1}-x_k|$ est importante : elle signifie que le comportement de $|x_k - x_\infty|$ est « le même » que celui de la suites des différences $|x_{k+1}-x_k|$. En effet, on a déjà l'estimation dans l'autre sens par l'inégalité triangulaire : $|x_{k+1}-x_k| \leqslant |x_k - x_\infty| + |x_{k+1}-x_\infty|$.*

On utilisera ceci en pratique par exemple pour visualiser la vitesse de convergence de suites dont on ne connaît pas la limite, en traçant $|x_{k+1}-x_k|$ en fonction de k.

Remarque 1.2. *Toutes ces notions ont été présentées en dimension un, mais elles restent valables dans $\mathbb{R}^n$ en remplaçant la valeur absolue par la norme (peu importe laquelle), et les critères de convergence et de vitesse de convergence sont exactement les mêmes : on n'a utilisé que l'inégalité triangulaire dans toutes les démonstrations, et le fait qu'une série absolument convergente était convergente (ce qui est vrai dès que l'espace est complet).*

Nous allons rapidement illustrer ces notions avec un exemple de méthode qui fonctionne lorsque l'on connait la dérivée, la méthode de dichotomie.

Proposition 1.4. *Méthode de dichotomie (ou de dissection) pour la dérivée.*

On suppose que f est une fonction C^1 sur I et qu'il existe des points de I, a_0 et b_0 avec $a_0 < b_0$ tels que $f'(a_0) < 0 < f'(b_0)$.

La méthode de dichotomie est la suivante : par récurrence, on pose $c_n = \frac{1}{2}(a_n + b_n)$ le point milieu et on définit ensuite :

$$(a_{n+1}, b_{n+1}) = \begin{cases} (a_n, c_n) & \text{si } f'(c_n) > 0 \\ (c_n, b_n) & \text{si } f'(c_n) < 0 \\ (c_n, c_n) & \text{si } f'(c_n) = 0. \end{cases}$$

Alors les suites a_n, b_n, c_n convergent vers une limite ℓ qui vérifie $f'(\ell) = 0$, et la convergence est linéaire. Si la suite n'est pas constante à partir d'un certain rang (autrement dit si on n'est jamais dans le cas $f'(c_n) = 0$), alors le taux de convergence linéaire est $\frac{1}{2}$.

Démonstration. On observe que si on n'a jamais le cas $f'(c_n) = 0$, alors la longueur de l'intervalle $[a_n, b_n]$ est divisée par deux à chaque étape. Comme $c_n = a_{n+1}$ ou b_{n+1}, on a dans tous les cas $|c_{n+1} - c_n| = \frac{1}{2}|b_{n+1} - a_{n+1}| = \frac{1}{2^{n+1}}|b_1 - a_1|$, et donc $|c_{n+1} - c_n| = \frac{1}{2}|c_n - c_{n-1}|$, ce qui donne directement la convergence linéaire et le fait que le taux de convergence est inférieur ou égal à $\frac{1}{2}$, d'après le critère de la proposition 1.3.

Si ℓ est la limite de (c_n), alors comme $|a_n - c_n| = \frac{1}{2}|b_n - a_n| = \frac{1}{2^{n+1}}|b_0 - a_0|$, on obtient que $|a_n - \ell| \leqslant |a_n - c_n| + |c_n - \ell| \leqslant C\frac{1}{2^n}$. Donc (a_n) converge aussi vers ℓ, la convergence étant linéaire, et le taux de convergence linéaire plus petit que $\frac{1}{2}$. On obtient exactement la même chose pour b_n de la même manière. En passant à la limite, on obtient $f'(\ell) \geqslant 0$ et $f'(\ell) \leqslant 0$, donc on a bien $f'(\ell) = 0$.

En fait on a même que le taux de convergence de c_n est égal à $\frac{1}{2}$: s'il était strictement plus petit, on aurait, pour $\alpha < \frac{1}{2}$

$$|c_1 - c_0| = 2^n|c_{n+1} - c_n| \leqslant 2^n(|c_{n+1} - \ell| + |c_n - \ell|) \leqslant 2^n(C\alpha^{n+1} + C\alpha^n) = C(\alpha + 1)(2\alpha)^n \to 0,$$

ce qui est une contradiction.

Si le taux de convergence (par exemple pour a_n, la démonstration pour b_n se fait de la même manière) était strictement plus petit que $\frac{1}{2}$, on aurait $|a_n - \ell| \leqslant C\alpha^n$ qui serait plus petit que $|a_n - c_n| = \frac{1}{2^{n+1}}|b_0 - a_0|$ à partir d'un certain rang (par croissance comparée). Donc $\ell \in [a_n, c_n[$ à partir d'un certain rang, et donc on ne peut pas avoir $a_{n+1} = c_n$ (sinon on aurait $\ell < a_{n+1} \leqslant \ell$ puisque (a_n) est croissante). Donc on a $a_{n+1} = a_n$ à partir d'un certain rang, donc $\ell = a_n$ ce qui est absurde puisqu'on a toujours $f'(a_n) < 0$. □

Si par exemple f' est croissante sur I, on a alors obtenu un minimum global de f sur $[a_0, b_0]$.

Remarque 1.3. *Autres méthodes pour trouver un zéro de la dérivée. Les deux méthodes les plus connues pour trouver un zéro de la dérivée (et donc potentiellement un minimum de la fonction) sont la méthode de Newton et la méthode de la sécante.*

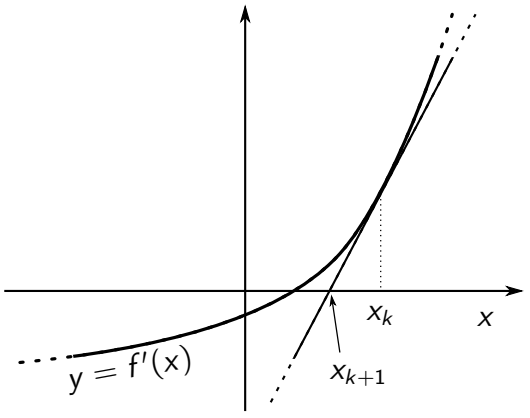

FIGURE 1.1 – Méthode de Newton : la *tangente* à la courbe $y = f'(x)$ en x_k coupe l'axe des abscisses en x_{k+1}.

La méthode de Newton nécessite d'avoir accès à la dérivée de f' (donc on voudra f au moins de classe C^2), les itérés sont donnés par la formule suivante (voir la figure 1.1 pour une interprétation

graphique) :

$$x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)}.$$

Contrairement à la méthode de dichotomie, on n'est pas assuré de la convergence de la suite des itérées. Cependant, si la fonction est de classe C^3 et si x_0 est suffisamment proche d'un minimum local x_* pour lequel $f''(x_*) > 0$, alors la suite des itérés converge vers x_* et la convergence est d'ordre 2. La méthode de la sécante (voir la figure 1.2 pour une interprétation graphique) est elle donnée par la formule

$$x_{k+1} = x_k - \frac{x_k - x_{k-1}}{f'(x_k) - f'(x_{k-1})} f'(x_k).$$

De même ici, on n'est pas assuré de la convergence de la suite (x_k). Cependant, si f est de classe C^3

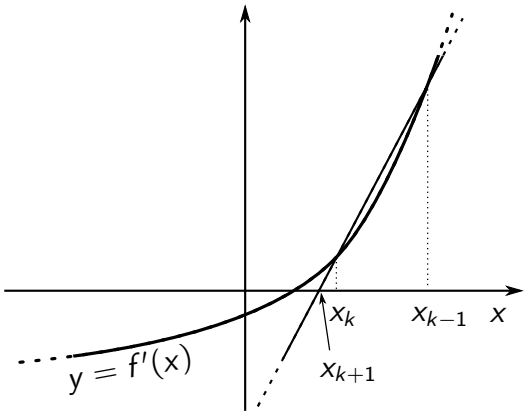

FIGURE 1.2 – Méthode de la sécante : la *sécante* à la courbe $y = f'(x)$ en x_k et x_{k-1} coupe l'axe des abscisses en x_{k+1}.

et si x_0 et x_1 sont suffisamment proches d'un minimum local x_* pour lequel $f''(x_*) > 0$, alors la suite converge vers x_* et l'ordre de convergence est $\varphi = \frac{1+\sqrt{5}}{2} \approx 1,618$ (le nombre d'or).

On définit enfin le coût d'une méthode comme le nombre d'évaluations de la fonction f pour obtenir un erreur inférieure à une tolérance donnée. Suivant le problème que l'on se donne, on peut être intéressé par une faible erreur sur la localisation du minimiseur, ou une faible erreur sur la valeur du minimum de la fonction.

En pratique, pour évaluer les méthodes, on cherchera plutôt a évaluer l'erreur en fonction du nombre d'évaluations plutôt que le contraire.

Définition 1.2. *Coût d'une méthode et ordre effectif. On définit l'ordre effectif comme suit, si l'erreur de la méthode par rapport au résultat recherché est ε_k lorsque le nombre d'évaluation de la fonction f lors de l'application de la méthode est k :*

- *On dit que l'ordre effectif est linéaire (d'ordre un) quand il existe $C > 0$ et $\alpha \in]0,1[$ tels que $\varepsilon_k \leqslant C\alpha^k$. Le taux de convergence linéaire effectif est la borne inférieure de tels α.*
- *On dit que l'ordre effectif est supérieur à β quand il existe $C > 0$, $\alpha \in]0,1[$ tels que $\varepsilon_k \leqslant C\alpha^{\beta^k}$. L'ordre effectif est la borne supérieure de tels β.*

Remarque 1.4. *Si chaque étape d'une méthode d'optimisation nécessite un nombre constant p d'évaluation de la fonction f, alors si la convergence de la suite des itérés de la méthode est linéaire, c'est aussi le cas de l'ordre effectif. Par contre si le taux de convergence linéaire est α, alors le taux effectif est $\alpha^{\frac{1}{p}}$. Si la convergence de la suite des itérés est d'ordre β, alors l'ordre effectif est $\beta^{\frac{1}{p}}$.*

Suivant les cas et les problèmes, si on est amené à évaluer la fonction ou ses dérivées, on peut supposer que le coût pour les évaluer n'est pas le même, et suivant la difficulté à estimer la dérivée par exemple, on peut s'intéresser au coût effectif en fonction du nombre d'évaluations de la dérivée plutôt que de la fonction.

Par exemple, si dans la méthode de Newton, on estime que le coût pour évaluer $f'(x_k)$ et $f''(x_k)$ correspond à deux évaluations de la fonction, alors l'ordre effectif de la méthode devient $\sqrt{2}$ soit environ 1.41, ce qui est moins bon que la méthode de la sécante, qui nécessite seulement d'évaluer une fois la fonction à chaque étape.

1.2 Méthode de la section dorée

Comme indiqué au début de la partie précédente, on cherchera souvent à se restreindre à un intervalle où il n'y a pas plusieurs minima locaux.

Définition 1.3. *Fonctions unimodales. Soit f une fonction continue sur $[a, b]$. On dit que f est unimodale s'il existe $x_* \in]a, b[$ tel que f soit strictement décroissante sur $[a, x_*]$ et strictement croissante sur $[x_*, b]$. On a donc un minimum local strict en x_* (c'est même l'unique minimum global sur $[a, b]$).*

Le terme « unimodal » peut avoir d'autres significations (en probabilités par exemple), mais on n'utilisera que cette définition-là.

1.2.1 Principe général : réduction du triplet

Tout comme il existe des paires de points admissibles pour la méthode de dichotomie (des points pour lesquel le signe de la fonction est différents) qui assurent d'avoir un zéro entre les deux, il existe des triplets de points qui assurent l'existence d'un minimum local.

Définition 1.4. *Triplet de points admissibles.*

Soit f une fonction continue sur un intervalle I, et trois réels $a < c < b$ de I. On dit que le triplet est admissible (pour le problème de minimisation de f) si on a $f(a) \geqslant f(c) \leqslant f(b)$.

Dans ce cas la fonction admet un minimum local sur $]a, b[$: elle admet un minimum global sur le compact $[a, b]$, et ce minimum ne peut être ni en a ni en b, sauf si $f(a) = f(c)$ ou $f(b) = f(c)$, mais dans ce cas il y a bien un minimum local en c.

Définition 1.5. *Algorithme général de réduction d'un triplet.*

Supposons qu'à une étape de l'algorithme on ait un triplet admissible $a < c < b$. L'itération suivante de l'algorithme consiste à se donner un quatrième point $d \in]a, b[$, avec $d \neq c$. On prend alors pour triplet suivant soit $\{a, c, d\}$ soit $\{c, d, b\}$, de telle sorte que ce soit un triplet admissible. Cela est toujours possible (faire des dessins) :

- *si $c < d$ et si $f(c) \leqslant f(d)$, le triplet (a, c, d) est admissible,*
- *si $c < d$ et si $f(c) \geqslant f(d)$, le triplet (c, d, b) est admissible,*
- *si $d < c$ et si $f(d) \leqslant f(c)$, le triplet (a, d, c) est admissible,*
- *si $d < c$ et si $f(d) \geqslant f(c)$, le triplet (d, c, b) est admissible.*

On obtient donc à chaque itération un nouveau triplet admissible.

On espère donc que pour une méthode donnée de choix du quatrième point à chaque étape, la taille du triplet (la différence entre les deux extrémités) tende au fur et à mesure vers 0.

Proposition 1.5. *Convergence vers le minimum local.*

Si l'algorithme de la définition fournit une suite de triplets (a_n, c_n, b_n) tels que $b_n - a_n \to 0$, et si la fonction est unimodale sur $[a_0, b_0]$ avec un minimum en x_*, alors les suites (a_n), (c_n) et (b_n) convergent vers x_*.

Démonstration. Il suffit de voir que les suites a_n et b_n sont adjacentes, donc elles convergent vers une limite ℓ (et donc (c_n) aussi par encadrement). D'autre part on a toujours $f(c_n) \leqslant f(b_n)$, avec $c_n < b_n$ et donc on ne peut pas avoir $b_n \leqslant x_*$ puisque f est strictement décroissante sur $[a, x_*]$. On a donc $b_n > x_*$ et donc à la limite $\ell \geqslant x_*$. De même en utilisant $f(a_n) \leqslant f(c_n)$, on obtient l'inégalité inverse et donc $\ell = x_*$. □

Remarque 1.5. *Pour initialiser l'algorithme, on a besoin d'un premier triplet admissible. On utilisera en particulier cet algorithme dans le chapitre suivant, pour minimiser des fonctions de la forme $h(t) = f(x + td)$ sur $]0, +\infty[$ où on sait seulement que h est strictement décroissante au voisinage de zéro. Il faut alors faire une première étape de recherche du triplet initial, par exemple en prenant $a = 0$ et $c = 1$, en diminuant d'abord c petit à petit jusqu'à ce qu'on ait $f(c) \leqslant f(a)$, puis en prenant $b \geqslant c$ et l'augmentant jusqu'à ce que $f(b) \geqslant f(c)$.*

1.2.2 Présentation de la méthode

La méthode de la section dorée consiste à s'arranger pour que la taille du triplet soit divisée d'un facteur constant à chaque étape. On s'aperçoit alors que cela contraint ce facteur à être $\varphi = \frac{1}{2}(1 + \sqrt{5})$, le nombre d'or.

Proposition 1.6. *Réduction constante de la taille du triplet.*

On suppose que l'algorithme de la définition 1.5 fournit une suite de triplets (a_n, c_n, b_n) tels que $b_{n+1} - a_{n+1} = \alpha(b_n - a_n)$ quel que soit le cas, alors on a deux possibilités pour c_n :

- *soit $c_n = a_n + \alpha(b_n - a_n)$, dans ce cas le quatrième point est placé en $d_n = a_n + (1-\alpha)(b_n - a_n)$,*
- *soit $c_n = a_n + (1-\alpha)(b_n - a_n)$, dans ce cas le quatrième point est placé en $d_n = a_n + \alpha(b_n - a_n)$.*

De plus le paramètre α vaut $\frac{1}{\varphi} = \frac{1}{2}(\sqrt{5} - 1)$.

Démonstration. Supposons par exemple que $c_n < d_n$. À l'étape suivante on a que (a_{n+1}, b_{n+1}) vaut (a_n, d_n) ou (c_n, b_n). On doit donc avoir $(d_n - a_n) = (b_n - c_n) = \alpha(b_n - a_n)$. On obtient donc la deuxième possibilité de la proposition. Si on avait $c_n > d_n$ on obtiendrait la première possibilité : $c_n = a_n + \alpha(b_n - a_n)$ et $d_n = a_n + (1 - \alpha)(b_n - a_n)$.

Si maintenant on suppose toujours que $c_n < d_n$, et qu'on est dans le cas $(a_{n+1}, c_{n+1}, b_{n+1}) = (a_n, c_n, d_n)$ on doit avoir soit $c_{n+1} = a_{n+1} + \alpha(b_{n+1} - a_{n+1})$, c'est à dire $c_n = a_n + \alpha(d_n - a_n)$, soit $c_n = a_n + (1-\alpha)(d_n - a_n)$. Mais ce dernier cas est exclus puisqu'on a d'après ce qui précède $c_n = a_n + (1 - \alpha)(b_n - a_n)$ et que $d_n < b_n$. On a donc

$$(1 - \alpha)(b_n - a_n) = c_n - a_n = \alpha(d_n - a_n) = \alpha^2(b_n - a_n),$$

ce qui donne $(1 - \alpha) = \alpha^2$ dont la solution positive est $\frac{1}{2}(\sqrt{5} - 1)$. □

En pratique, quitte à changer les noms des variables, et vu qu'on sait exactement où doivent être placés les points intérieurs, on écrira toujours le cas $c_n < d_n$. La méthode de la section dorée peut donc s'écrire comme suit :

Définition 1.6. *Méthode de la section dorée.*

On se donne une fonction f continue sur $[a_0, b_0]$. On pose $\alpha = \frac{1}{2}(\sqrt{5} - 1)$ et $c_0 = a_0 + (1 - \alpha)(b_0 - a_0)$ et $d_0 = a_0 + \alpha(b_0 - a_0)$. On calcule $f(a_0)$, $f(b_0)$, $f(c_0)$, et $f(d_0)$, et on suppose qu'un des triplets (a_0, c_0, b_0) ou (a_0, d_0, b_0) est admissible.

On définit les suites par récurrence :

— *si* $f(c_n) < f(d_n)$*, alors le triplet* (a_n, c_n, d_n) *est admissible, on pose* $(a_{n+1}, d_{n+1}, b_{n+1}) = (a_n, c_n, d_n)$ *et* $c_{n+1} = a_{n+1} + (1-\alpha)(b_{n+1} - a_{n+1})$*. On a simplement besoin de calculer* $f(c_{n+1})$*, puisque* $f(a_{n+1})$*,* $f(d_{n+1})$ *et* $f(b_{n+1})$ *sont déja connues.*

— *si* $f(c_n) \geqslant f(d_n)$*, alors le triplet* (c_n, d_n, b_n) *est admissible, on pose* $(a_{n+1}, c_{n+1}, b_{n+1}) = (c_n, d_n, b_n)$ *et* $d_{n+1} = a_{n+1} + \alpha(b_{n+1} - a_{n+1})$*. On a simplement besoin de calculer* $f(d_{n+1})$*, puisque* $f(a_{n+1})$*,* $f(c_{n+1})$ *et* $f(b_{n+1})$ *sont déja connues.*

Proposition 1.7. *Convergence de la méthode de la section dorée.*

Les suites (a_n)*,* (b_n)*,* (c_n) *et* (d_n) *convergent linéairement avec un taux de convergence* α *vers une limite* ℓ*. Si la fonction* f *est unimodale sur* $[a_0, b_0]$*, alors* f *admet son minimum en* ℓ*.*

Démonstration. Les suites (a_n) et (b_n) sont adjacentes et $b_n - a_n = \alpha^n(b_0 - a_0) \to 0$. On a donc $a_n \leqslant \ell \leqslant b_n$. Et

$$\max\{|a_n - \ell|, |b_n - \ell|, |c_n - \ell|, |d_n - \ell|\} \leqslant b_n - a_n = \alpha^n(b_0 - a_0),$$

ce qui donne la convergence linéaire vers ℓ des quatre suites, avec un taux inférieur ou égal à α. On montre que ce taux est exactement α comme dans la preuve de la méthode de dichotomie. Le fait que f admette son minimum en ℓ si f est unimodale est une conséquence de la proposition 1.5. □

Remarque 1.6. *Le taux effectif de convergence linéaire de cette méthode est* $\alpha \approx 0,618$*. En effet, on n'a besoin d'évaluer* f *qu'en un seul point à chaque itération. Si on n'a pas directement accès au calcul de la dérivée, et qu'on utilise la méthode de dichotomie présentée précédemment en approximant* f' *par différences finies, on évalue la fonction* f *en deux points différents à chaque itération, ce qui fait que le taux de convergence linéaire effectif est* $\sqrt{\frac{1}{2}} \approx 0,707$*. La méthode de la section dorée est donc plus efficace, et tout aussi robuste (on sait qu'elle converge dans tous les cas, et on sait exactement à quelle vitesse). On va voir dans les paragraphes suivants des méthodes pouvant converger bien plus rapidement, mais qui sont moins robustes. Tout comme la méthode de Newton ou la méthode de la sécante pour la recherche de zéro d'une fonction, elles ne convergent pas pour toute condition initiale, mais lorsqu'elles convergent, elles le font de manière superlinéaire.*

1.3 Autres méthodes

1.3.1 Par interpolations quadratiques successives

Lorsque l'on connaît la valeur de f en trois points $a < c < b$ constituant un triplet admissible, on peut espérer que la fonction s'approche d'une parabole, que l'on peut calculer par interpolation quadratique. Il est alors possible d'obtenir le minimum exact de cette parabole.

Exercice 1.2. *Si* $p(x)$ *est une fonction polynomiale de degré* 2 *telle que* $p(a) = f(a)$*,* $p(b) = f(b)$*, et* $p(c) = f(c)$*, alors montrer que* p' *s'annule au point* x_* *donné par*

$$x_* = F_2(a, b, c) = \frac{1}{2}\frac{(b^2 - c^2)f(a) + (c^2 - a^2)f(b) + (a^2 - b^2)f(c)}{(b - c)f(a) + (c - a)f(b) + (a - b)f(c)}. \tag{1.6}$$

On pourrait donc appliquer la méthode générale de réduction de triplet en prenant pour quatrième point ce point x_* (on peut montrer que si $a < c < b$ forment un triplet admissible, alors forcément $x_* \in]a, b[$). Cependant en général ceci ne fonctionne pas aussi bien qu'on pourrait l'espérer, même si on démarre proche du minimum local. On a par exemple le bord droit du triplet b_n qui reste constant à partir d'un certain rang, alors que les a_n et c_n restent à gauche du minimum local, et convergent linéairement vers le minimum.

Exercice 1.3. *Pour la fonction $f : x \mapsto x^2 + \frac{1}{4}x^3$, qui a un minimum local strict en 0, montrer que la fonction F_2 définie en (1.6) est donnée par*

$$F_2(a, b, c) = \frac{ab + bc + ca}{8 + 2(a + b + c)}.$$

En déduire que la méthode générale de réduction du triplet appliquée avec $d_n = F_2(a_n, b_n, c_n)$ et $a_0 = -\frac{\alpha}{2}$, $c_0 = -\frac{\alpha}{4}$ et $b_0 = \alpha$ pour $\alpha \in]0, 1]$ conduit toujours à $b_{n+1} = b_n = \alpha$, $a_n < 0$, et $c_n < 0$.

Évaluer numériquement cette suite et observer la convergence linéaire de a_n et c_n vers 0 à l'aide d'un graphique.

La méthode de réduction par interpolations quadratiques successives consiste donc à prendre trois points initiaux x_0, x_1 et x_2 et à prendre à chaque étape $x_{k+1} = F_2(x_k, x_{k-1}, x_{k-2})$. On n'obtient pas forcément à chaque étape un triplet admissible, il peut par exemple exister des étapes où x_k, x_{k-1} et x_{k-2} se situent tous d'un même côté d'un minimum local...

Toutefois, on peut obtenir que si f est de classe C^3 et que x_0, x_1 et x_2 sont suffisamment proches d'un minimum local x_* de f pour lequel $f''(x) > 0$, alors la suite (x_k) converge vers x_*, avec un ordre d'au moins $\beta \approx 1,3247$, correspondant à l'unique solution réelle de l'équation $\beta^3 = \beta + 1$.

Exercice 1.4. *Calculer numériquement la suite x_n dans le cas de la fonction f de l'exercice 1.3, lorsque $x_0 = 1$, $x_1 = -\frac{1}{2}$ et $x_2 = -\frac{1}{4}$. Observer la convergence superlinéaire de x_n vers 0 à l'aide d'un graphique.*

1.3.2 *Méthodes utilisant la dérivée*

Lorsque l'on a accès à la dérivée f et à la fonction f', il existe des méthodes plus efficaces que les méthodes précédentes. Tout d'abord pour les fonctions unimodales, on peut être sûr d'encadrer le minimum avec seulement deux points a et b dès que l'on a $f'(a) < 0 < f'(b)$.

On peut donc chercher des algorithmes qui réduisent seulement ce couple, en conservant cet encadrement. Par exemple on peut chercher le polynôme p de degré 3 qui satisfait $f(a) = p(a)$, $f(b) = p(b)$, $f'(a) = p'(a)$ et $f'(b) = p'(b)$. On prend alors comme nouveau point l'unique minimum local de p que l'on peut résoudre directement (c'est une équation de degré deux) et que l'on note $F_3(a, b)$. On peut montrer qu'il appartient à $]a, b[$ si on a $f'(a) < 0 < f'(b)$).

Exercice 1.5. *Montrer qu'on a*

$$p(x) = \frac{f(b)(x-a)^2(3b-a-2x) - f(a)(x-b)^2(3a-b-2x)}{(b-a)^3} + \frac{f'(b)(x-a)^2(x-b) + f'(a)(x-b)^2(x-a)}{(b-a)^2}.$$

Exercice 1.6. **(calculatoire).*

On pose $\Delta = \frac{f(b)-f(a)}{b-a}$, $\beta = f'(b) + f'(a) - 2\Delta$ et $\delta = (\beta - \Delta)^2 - f'(a)f'(b)$. Montrer que p' a deux racines réelles distinctes si et seulement si $\beta \neq 0$ et $\delta > 0$. Dans ce cas, montrer alors que la racine correspondant au minimum local de p est donnée par

$$F_3(a, b) = \frac{af'(b) + bf'(a) + (a+b)(\beta - \Delta) + |b-a|\sqrt{\delta}}{3\beta}.$$

Une méthode de réduction consisterait à éliminer un des points a ou b de sorte que l'on ait toujours une dérivée négative pour le point de gauche et positive pour le point de droite. Comme précédemment, une telle méthode ne fonctionne pas aussi bien qu'on peut l'espérer : une des deux suites a_n ou b_n peut devenir stationnaire à partir d'un certain rang, et la convergence de l'autre n'est alors pas mieux que linéaire.

Exercice 1.7. *Pour la fonction* $f : x \mapsto x^2 + \frac{1}{6}x^3 - \frac{1}{12}x^4$*, qui a un minimum local strict en* 0*, programmer la méthode proposée ci-dessus en prenant* $a_0 = -1$ *et* $b_0 = \frac{1}{2}$*, observer que l'on a toujours* $a_n = -1$*, et que la convergence de* b_n *vers* 0 *est linéaire à l'aide d'un graphique.*

Pour obtenir une convergence superlinéaire, on peut faire comme précédemment : on choisit x_0 et x_1 et on prend $x_{k+1} = F_3(x_k, x_{k-1})$. On peut montrer que si la fonction f est C^4 et que les points x_0 et x_1 sont suffisamment proches d'un minimum local x_* pour lequel $f''(x_*) > 0$ et $f^{(3)}(x_*) \neq 0$, alors la suite x_k converge quadratiquement vers x_*.

Exercice 1.8. *Calculer numériquement la suite* x_n *dans le cas de la fonction* f *de l'exercice 1.7, lorsque* $x_0 = -1$ *et* $x_1 = \frac{1}{2}$*. Observer la convergence superlinéaire de* x_n *vers* 0 *à l'aide d'un graphique.*

On peut aussi montrer que les deux méthodes coïncident lorsque $f^{(4)}(x_*) > 0$, dès que les points initiaux sont suffisamment proches de x_*.

Chapitre 2

Méthodes de descente de gradient

Le cadre de ce chapitre est le suivant. On se place dans $\mathbb{R}^n$, avec $n \geqslant 2$, et on a une fonction f de $\mathbb{R}^n$ dans $\mathbb{R}$, ou plus généralement de Ω dans $\mathbb{R}$, où Ω est un ouvert. Lorsqu'on utilisera la notation Ω, ce sera toujours pour parler d'un ouvert de $\mathbb{R}^n$.

On cherche un minimiseur local de f, c'est à dire un point x_* tel que pour tout x dans un voisinage de x_*, par exemple la boule $B(x_*, r)$ de centre x_* et de rayon $r > 0$ bien choisi, on ait $f(x_*) \leqslant f(x)$. Si l'inégalité est stricte dès que $x \neq x_*$, on dit alors que c'est un point de minimum local strict.

Proposition 2.1. *Rappels.*

- *Si f est différentiable en un point de minimum local x_*, alors $\nabla f(x_*) = 0$.*
- *Si de plus f est de classe C^2, alors la hessienne $H_f(x_*)$ de f au point x_* est une matrice symétrique positive.*
- *D'autre part, si f est C^2, si $\nabla f(x_*) = 0$ et $H_f(x_*)$ est symétrique définie positive, alors x_* est un point de minimum local strict.*

On cherchera donc à résoudre l'équation $\nabla f(x) = 0$. D'une manière plus générale, il existe des problèmes de la forme $g(x) = 0$, où cette fois g est une fonction de $\mathbb{R}^n$ (ou Ω) dans $\mathbb{R}^n$. Contrairement au cas unidimensionnel, il existe peu de méthodes efficaces pour résoudre ces problèmes numériquement (à part dans des cas très particuliers). Une des manières de s'attaquer au problème est de le transformer en un problème d'optimisation, en trouvant une fonction f telle que $\nabla f = g$, et en cherchant un minimum (ou un maximum) local de f par des méthodes d'optimisation numérique.

2.1 Présentation générale des méthodes de descente

Le principe de base des méthodes de descente est très simple : on fait comme sur une carte topographique, et on utilise les lignes de niveau pour se diriger vers le fond d'une vallée.

Faire un dessin, des courbes de niveaux d'une fonction simple, par exemple des ellipses pour bien observer un chemin correspondant à de la descente.

Définition 2.1. *Direction de descente.*

Soit f une fonction C^1 de Ω dans $\mathbb{R}$. On dit que $d \in \mathbb{R}^n$ est une direction de descente en x si la fonction auxiliaire d'une variable $h : t \mapsto f(x + td)$ vérifie $h'(0) < 0$.

Si d est une direction de descente, alors h est strictement décroissante au voisinage de 0 : pour tout $\alpha > 0$ suffisamment petit, on a $f(x + \alpha d) < f(x)$. Attention, ce n'est pas équivalent, il se pourrait bien que ce soit le cas avec $h'(0) = 0$. La notion de direction de descente est une notion un peu plus forte.

Définition 2.2. *Algorithme général pour les méthodes de descente.*

On se fixe un point initial $x_0 \in \Omega$. À chaque étape (notons k le numéro de l'étape) :

— *On choisit une direction de descente d_k,*

— *on choisit un pas α_k tel que $x_k + \alpha_k d_k \in \Omega$ et que $f(x_k + \alpha_k d_k) < f(x_k)$,*

— *on pose $x_{k+1} = x_k + \alpha_k d_k$.*

Les différentes méthodes de descente correspondent à des choix différents pour les directions de descente et les pas. On commence par la méthode la plus simple, où le pas est constant, et où la direction est celle de plus grande pente.

2.2 Descente de gradient à pas fixe

Proposition 2.2. *Interprétation géométrique du gradient.*

On munit $\mathbb{R}^n$ de sa norme euclidienne notée $\|\cdot\|$. Soit f une fonction C^1 de Ω dans $\mathbb{R}$. On a les résultats suivants :

— *L'opposé du gradient, $-\nabla f(x)$, est une direction de descente en x si $\nabla f(x) \neq 0$.*

— *La direction du gradient est la direction de plus forte pente (si $\|d\| = 1$, alors la pente dans la direction d correspond à la dérivée de h en 0).*

— *Le gradient est perpendiculaire aux lignes de niveau.*

Faire un dessin avec des lignes de niveau un peu allongées (ellipses bien aplaties) et illustrer ces affirmations. Expliquer que la direction du gradient n'est pas forcément bien alignée avec le minimum

Démonstration. Si $h(t) = f(x + td)$, par composition on a $h'(t) = \langle \nabla f(x + td), d\rangle$, donc en 0 on obtient $h'(0) = \langle \nabla f(x), d\rangle$. Donc si $d = -\nabla f(x)$ on obtient $h'(0) = -\|\nabla f(x)\|^2 < 0$ si $\nabla f(x) \neq 0$.

Si $\|d\| = 1$, par Cauchy-Schwarz, on obtient que $h'(0) = \langle \nabla f(x), d\rangle \leqslant \|\nabla f(x)\|\|d\|$, avec égalité si et seulement si ∇f et d sont positivement liés, autrement dit si $d = \frac{\nabla f(x)}{\|\nabla f(x)\|}$ et que donc $h'(0)$ est maximal lorsque d est aligné avec $\nabla f(x)$ (et minimal si d est aligné avec $-\nabla f(x)$, par abus on parle aussi de plus forte pente lorsqu'elle est négative, avec une valeur absolue maximale).

Enfin si $t \mapsto \gamma(t) \in \Omega$ (pour t dans un intervalle I fixé) est une ligne de niveau et si $x = \gamma(t_0)$ est un point de cette ligne de niveau, alors on a $f(\gamma(t)) = f(x)$ pour tout $t \in I$. En dérivant par rapport à t on obtient $\langle \nabla f(x), \gamma'(t_0)\rangle = 0$. Le vecteur $\gamma'(t_0)$ étant un vecteur directeur de la tangente de la courbe γ en $\gamma(t_0)$, donc $\nabla f(x)$ qui est orthogonal à la tangente à la courbe en x. C'est cela qu'on appelle être « perpendiculaire » aux lignes de niveau. □

Définition 2.3. *Algorithme de descente de gradient à pas fixe.*

On se fixe un pas $\alpha > 0$ et un point de départ $x_0 \in \mathbb{R}^n$. On pose alors

$$x_{k+1} = x_k - \alpha \nabla f(x_k). \tag{2.1}$$

On a pris ici $\mathbb{R}^n$ à la place de Ω pour s'éviter le cas où x_{k+1} n'appartiendrait pas à Ω. On se donne en général un critère d'arrêt pour une tolérance ε fixée, par exemple $\|\nabla f(x_k)\| \leqslant \varepsilon$ (le gradient est suffisamment proche de 0) ou $|f(x_{k+1}) - f(x_k)| \leqslant \varepsilon$ (la valeur de f ne diminue plus trop).

2.2.1 Étude du cas test

On regarde ce que donne la méthode dans le cas typique d'une fonction quadratique $f(x) = \frac{1}{2}\langle x, Ax\rangle + \langle b, x\rangle$ avec $A \in M_n(\mathbb{R})$ une matrice symétrique définie positive.

Le gradient en x est donné par $Ax+b$, donc on cherche à approcher la solution de l'équation $Ax+b=0$.

La méthode de descente de gradient à pas fixe est donc donnée par

$$x_{k+1} = x_k - \alpha(Ax + b).$$

Avant d'étudier la convergence, on a besoin d'un petit lemme classique concernant les matrices symétriques.

Lemme 2.1. *Soit M une matrice symétrique de taille n. Alors, si $\lambda_1, \dots \lambda_n$ sont les valeurs propres de M, on a, pour tout $x \in \mathbb{R}^n$, $\|Mx\| \leqslant \max_{1\leqslant i\leqslant n} |\lambda_i| \|x\|$.*

Démonstration. On choisit $e_1, \dots, e_n$ une base orthonormée de vecteurs propres de M, et si on décompose x dans cette base : $x = \sum_{i=1}^n x_i e_i$, alors on a $\|x\|^2 = \langle x, x\rangle = \sum_{i=1}^n x_i^2$.

On a donc aussi $\|Mx\|^2 = \langle Mx, Mx\rangle = \sum_{i=1}^n \lambda_i^2 x_i^2 \leqslant \max_{1\leqslant i\leqslant n} \lambda_i^2 \|x\|^2$. On obtient exactement le résultat voulu en prenant la racine carrée. □

Proposition 2.3. *On pose ℓ (resp. L) la plus petite (resp. grande) valeur propre de A.*

Si $\alpha \in]0, \frac{2}{L}[$, alors la convergence est linéaire avec un taux plus petit que $\max(|1-\alpha\ell|, |1-\alpha L|)$.

Le meilleur choix du pas pour rendre ce taux le plus petit possible est $\alpha = \dfrac{2}{\ell + L}$, et le taux est alors $\dfrac{L-\ell}{L+\ell}$.

Démonstration. On utilise le critère de convergence linéaire, en écrivant que

$$x_{k+1} - x_k = x_k - \alpha(Ax_k + b) - x_{k-1} - \alpha(Ax_{k-1} + b) = (I_n - \alpha A)(x_k - x_{k-1}).$$

Les valeurs propres de $I_n - \alpha A$ sont $1 - \alpha\lambda$ pour λ valeur propre de A, et donc on obtient

$$\|x_{k+1} - x_k\| \leqslant \max(|1 - \alpha\ell|, |1 - \alpha L|)\|x_k - x_{k-1}\|.$$

On a donc convergence linéaire avec le taux indiqué, dès que ce taux est strictement inférieur à 1, ce qui correspond à $\alpha \in]0, \frac{2}{\ell}[$ et $\alpha \in]0, \frac{2}{L}[$.

On obtient le minimum du taux lorsque les deux valeurs $|1 - \alpha\ell|$ et $|1 - \alpha L|$ sont égales (sinon on peut diminuer le max en modifiant légèrement α). Dès que $L > \ell$ la seule possibilité pour que ces deux valeurs soient égales est que $1-\alpha\ell = -(1-\alpha L)$, ce qui donne bien le résultat attendu $\alpha = \frac{2}{\ell+L}$, et le taux vaut alors

$$1 - \frac{2\ell}{\ell + L} = \frac{2L}{\ell + L} - 1 = \frac{L-\ell}{L+\ell}.$$

□

On observe ici que le taux peut être très proche de 1 lorsque $\ell \ll L$, on dit alors que la matrice est mal conditionnée.

2.2.2 Convergence de la méthode de gradient à pas fixe

On aimerait montrer un analogue de ce qu'on a obtenu pour le cas test dans le cas général. Suivant les hypothèses que l'on fait, on obtiendra des résultats de convergence plus ou moins forts.

L'hypothèse de base dont on a besoin est une certaine régularité sur le gradient de la fonction. Au minimum, on aura besoin qu'il soit Lipschitzien. On donne d'abord un premier outil d'estimation qui sera utile par la suite.

Proposition 2.4. *Estimation de second ordre.*

Supposons que $f : \Omega \to \mathbb{R}$ est de classe C^1 et que ∇f est L-Lipschitzienne sur Ω. Si x et y sont deux points de Ω tels que le segment $[x, y]$ soit inclus dans Ω, alors on a

$$f(y) \leqslant f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2. \tag{2.2}$$

Démonstration. On pose $h = y - x$, et on écrit la formule de Taylor à l'ordre 1 au point x (qui est valide dès que le segment $[x, x + h]$ est inclus dans Ω, ce qui est le cas ici puisque $x + h = y$) :

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

et on soustrait $\langle \nabla f(x), h \rangle = \int_0^1 \langle \nabla f(x), h \rangle \mathrm{d}t$ des deux côtés pour obtenir

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

À l'aide de l'inégalité de Cauchy-Schwarz et du fait que ∇f est L-Lipschitz sur Ω, on obtient

$$f(x + h) - f(x) - \langle \nabla f(x), h \rangle \leqslant \int_0^1 Lt\|h\|\|h\| \mathrm{d}t = \frac{L}{2}\|h\|^2,$$

qui est exactement l'inégalité souhaitée. □

On remarque ici qu'on n'a en fait eu besoin du fait que ∇f soit L-Lipschitz uniquement sur le segment $[x, y]$.

On peut donc montrer un premier résultat de convergence de la méthode de gradient à pas fixe.

Théorème 1. *Convergence de la méthode de gradient à pas fixe.*

Soit $f : \Omega \to \mathbb{R}$ une fonction C^1, bornée inférieurement. On se fixe un point $x_0 \in \Omega$ et on suppose les deux conditions suivantes :

— L'ensemble $S_0 = \{x \in \Omega, f(x) \leqslant f(x_0)\}$ est fermé dans $\mathbb{R}^n$.

— Le gradient ∇f est L-Lipschitz sur S_0.

Alors, si on choisit un pas $\alpha < \frac{2}{L}$, l'algorithme de descente de gradient donné en (2.1) *fournit effectivement une suite (x_k) correspondant à une méthode de descente au sens de la définition 2.2, et on a un résultat de convergence :*

— Pour tout $k \in \mathbb{N}$, $x_k \in \Omega$.

— La suite $(f(x_k))_{k\in\mathbb{N}}$ est décroissante, et converge vers une limite finie.

— La suite $(\nabla f(x_k))_{k\in\mathbb{N}}$ converge vers 0 dans $\mathbb{R}^n$.

Démonstration. On note tout d'abord que s'il existe un rang pour lequel $\nabla f(x_k) = 0$, alors la suite est stationnaire à partir de ce rang et il n'y a donc plus rien à montrer. On suppose donc que $\nabla f(x_k) \neq 0$ pour tout k, de sorte qu'on a bien affaire à une direction de descente dans tous les cas.

Montrons par récurrence que $x_k \in S_0$. On note I l'ensemble des $t \in]0, \frac{2}{L}]$ tels que le segment $[x_k, x_k - t\nabla f(x_k)]$ soit inclus dans S_0. Alors pour $t \in I$, on obtient d'après l'estimation de second ordre de la proposition 2.4 :

$$f(x_k - t\nabla f(x_k)) \leqslant f(x_k) - t\langle \nabla f(x_k), \nabla f(x_k) \rangle + t^2 \tfrac{L}{2}\|\nabla f(x_k)\|^2.$$

On a donc, puisque $t \leqslant \frac{2}{L}$,

$$f(x_k - t\nabla f(x_k)) - f(x_k) \leqslant -t\|\nabla f(x_k)\|^2(1 - \tfrac{1}{2}tL) \leqslant 0. \tag{2.3}$$

Un argument de connexité permet de montrer que $I =]0, \frac{2}{L}]$:

— d'une part, comme $\nabla f(x_k)$ est une direction de descente et que Ω est ouvert, I est non vide : pour tout t suffisamment petit, on a $x_k - t\nabla f(x_k) \in \Omega$ et $f(x_k - t\nabla f(x_k)) < f(x_k) \leqslant f(x_0)$ puisque $x_k \in S_0$ par hypothèse de récurrence. On peut donc définir sa borne supérieure $t_* \leqslant \frac{2}{L}$.
— Comme S_0 est fermé, si $t_n \in I$ et $t_n \to t_*$, alors le segment $[x_k, x_k - t_n\nabla f(x_k)]$ est inclus dans S_0 pour tout n, et on obtient à la limite que $x_k - t\nabla f(x_k) \in S_0$, donc le segment $[x_k, x_k - t\nabla f(x_k)]$ est inclus dans S_0, donc $t_* \in I$.
— Enfin si $t_* < \frac{2}{L}$, la dernière inégalité de l'estimation (2.3) est une inégalité stricte, ce qui permet d'obtenir que $f(x_k - t_*\nabla f(x_k)) < f(x_k)$. Comme Ω est ouvert, pour tout ε suffisamment petit on aurait $x_k - (t_* + \varepsilon)\nabla f(x_k) \in \Omega$ avec $f(x_k - (t_* + \varepsilon)\nabla f(x_k)) < f(x_k) \leqslant f(x_0)$ et donc $x_k - (t_* + \varepsilon)\nabla f(x_k) \in S_0$ pour tout ε suffisamment petit. Ce qui donnerait que $t_* + \varepsilon$ serait dans I pour tout ε suffisamment petit, en contradiction avec le fait que t_* est la borne supérieure.

On a donc bien l'estimation (2.3) pour tout $t \in [0, \frac{2}{L}]$, et en particulier pour $t = \alpha$ (on a supposé que $\alpha < \frac{2}{L}$), on obtient que $x_{k+1} \in S_0$. De plus, on a donc pour tout $k \in \mathbb{N}$

$$f(x_{k+1}) - f(x_k) \leqslant -\alpha\|\nabla f(x_k)\|^2(1 - \tfrac{1}{2}\alpha L) < 0,$$

ce qui nous donne que la suite $(f(x_k))$ est strictement décroissante, et converge vers une limite finie puisque f est bornée inférieurement. En passant à la limite dans les inégalités précédentes, on obtient bien que $\|\nabla f(x_k)\| \to 0$ lorsque $k \to \infty$. □

Remarque 2.1. *Ce résultat permet d'affirmer que si on se donne comme critère d'arrêt le fait que $\|\nabla f(x_k)\| < \varepsilon$ ou $|f(x_{k+1}) - f(x_k)| < \varepsilon$ pour ε une tolérance fixée, alors l'algorithme va bien terminer. Mais cela ne permet pas de dire que la suite x_k converge, et encore moins de dire qu'elle converge vers un minimum local (il est même possible qu'il n'y ait pas de minimum local, comme par exemple si $f(x) = \frac{1}{x}$ avec $\Omega =]0, +\infty[$, pour lequel le théorème s'applique).*

On peut raffiner ce théorème, en ajoutant des hypothèses pour obtenir des résultats plus précis, dans le but d'obtenir la convergence de la suite (x_k) vers un minimum local.

Corollaire 2.1. *Sous les hypothèses du Théorème 1, si on suppose de plus que S_0 est compact, alors la suite des $f(x_k)$ converge vers une valeur critique.*

Démonstration. On rappelle qu'une valeur critique de f est un réel c tel qu'il existe un point $x_* \in \Omega$ avec $f(x_*) = c$ et $\nabla f(x_*) = 0$. Par compacité, on peut extraire une suite $x_{\varphi(k)}$ qui converge dans S_0 vers un point x_∞. Par continuité de f et de ∇f, on obtient que $\lim_{k\to\infty} f(x_k) = \lim_{k\to\infty} f(x_{\varphi(k)}) = f(x_\infty)$, et $\|\nabla f(x_\infty)\| = \lim_{k\to\infty} \|\nabla f(x_{\varphi(k)})\| = \lim_{k\to\infty} \|\nabla f(x_k)\| = 0$, ce qui est bien ce qu'on voulait démontrer. □

Ici, on n'a toujours rien montré sur la convergence des x_k. On peut l'obtenir si on suppose des conditions de non-dégenerescence (on dit qu'un point critique est non-dégénéré si la Hessienne en ce point est inversible).

Corollaire 2.2. *Sous les hypothèses du Théorème 1, si S_0 est compact, si f est de classe C^2, et si tout point critique de f sur S_0 est non-dégénéré, c'est à dire si pour tout $x \in S_0$ on a*

$$\nabla f(x) = 0 \Rightarrow H_f(x) \text{ inversible},$$

alors la suite $(x_k)_{k\in\mathbb{N}}$ converge vers un point critique de la fonction f.

Démonstration. On va montrer que les points critiques sont des points isolés. Soit x un point critique. Alors, pour h suffisamment petit, on a $x + h \in \Omega$ et

$$\nabla f(x + h) = \nabla f(x) + \int_0^1 H_f(x + th)(h)\mathrm{d}t.$$

Donc comme $\nabla f(x) = 0$, on obtient (en écrivant $\|\cdot\|$ une norme d'opérateur pour les matrices)

$$\|\nabla f(x+h) - H_f(x)(h)\| = \left\| \int_0^1 (H_f(x+th) - H_f(x))(h) \mathrm{d}t \right\| \leqslant \sup_{t\in[0,1]} \|H_f(x+th) - H_f(x)\| \|h\|.$$

Pour $\delta > 0$ arbitraire et dès que h est suffisamment petit, on obtient donc par continuité de H_f :

$$\|\nabla f(x+h) - H_f(x)(h)\| \leqslant \delta \|h\| \leqslant \delta \|H_f^{-1}(x)\| \|H_f(x)(h)\| \leqslant \frac{1}{2} \|H_f(x)(h)\|.$$

Et donc en prenant par exemple $\delta = \frac{1}{2\|H_f^{-1}(x)\|}$, on obtient qu'il existe ε tel que dès que $\|h\| \leqslant \varepsilon$, on ait $\|\nabla f(x+h) - H_f(x)(h)\| \leqslant \frac{1}{2}\|H_f(x)(h)\|$, et donc

$$\|\nabla f(x+h)\| \geqslant \|H_f(x)(h)\| - \|\nabla f(x+h) - H_f(x)(h)\| \geqslant \tfrac{1}{2}\|H_f(x)(h)\|.$$

Par ailleurs, puisque $H_f(x)$ est inversible, dès que $h \neq 0$, l'inégalité nous donne $\|\nabla f(x+h)\| > 0$. Il n'y a donc aucun point critique autre que x dans la boule de centre x et de rayon ε.

Enfin, comme dans le corollaire 2.1 les valeurs d'adhérence de la suite x_k sont des points critiques, mais on a également par le théorème 1 que $\|x_{k+1} - x_k\| \to 0$. Montrons qu'elle a une seule valeur d'adhérence. Soit x_* une valeur d'adhérence, qui est un point critique, et pour lequel la boule $B(x_*, r)$ ne contient pas d'autre point critique. Soit $\varepsilon < \frac{r}{2}$, et supposons que la suite (x_k) rentre et sort une infinité de fois de la boule $B(x_*, \varepsilon)$, c'est à dire qu'on a une extraction $\varphi(k)$ telle que $x_{\varphi(k)} \in B(x*, \varepsilon)$ et $x_{\varphi(k)+1} \notin B(x*, \varepsilon)$. À partir d'un certain rang, on a $\|x_{\varphi(k)+1} - x_{\varphi(k)}\| \leqslant \varepsilon$ et donc $x_{\varphi(k)+1} \in \overline{B(x_*, 2\varepsilon)}$. Par compacité on peut donc extraire une sous suite $x_{\psi(k)}$ qui converge et qui vérifie $\varepsilon \leqslant \|x_{\psi(k)} - x_*\| \leqslant 2\varepsilon$, et donc à la limite on obtient un autre point critique dans $\overline{B(x_*, 2\varepsilon)} \subset B(x_*, r)$, ce qui est une contradiction. Donc la suite reste dans la boule $B(x_*, \varepsilon)$ à partir d'un certain rang, et ce quelque soit ε, donc la suite converge bien vers x_*. □

On n'a toujours pas de résultat de convergence vers un minimum local, et pour cela on va rajouter une hypothèse de convexité locale (sur la Hessienne, lorsque f est de classe C^2).

Pour exprimer cette hypothèse on va utiliser une notation de comparaison des matrices symétriques.

Définition 2.4. *Soit M une matrice symétrique et L un réel. On notera $M \geqslant 0$ pour dire que M est positive (au sens que la forme quadratique associée est positive, c'est à dire que pour tout $h \in \mathbb{R}^n$ on a $\langle h, Mh\rangle \geqslant 0$). Cela équivaut à ce que toutes les valeurs propres de M soient positives.*

De même, on notera $M \leqslant LI_n$ (resp. $LI_n \leqslant M$) si $LI_n - M$ (resp. $M - LI_n$) est positive. Cela équivaut à ce que toutes les valeurs propres de M soient inférieures (resp. supérieures) ou égales à L.

On comparera toujours les matrices à des multiples de l'identité, et il n'y a que dans ce cas-là que l'on peut conclure quelque chose sur les valeurs propres.

Théorème 2. *Convergence vers un minimum local dans le cas de convexité locale.*

Soit $f : \Omega \to \mathbb{R}$ une fonction C^2, bornée inférieurement. On se fixe un point $x_0 \in \Omega$ et on suppose les deux conditions suivantes, pour un certain $L > 0$:

— L'ensemble $S_0 = \{x \in \Omega, f(x) \leqslant f(x_0)\}$ est fermé dans $\mathbb{R}^n$.

— Pour tout $x \in S_0$, la Hessienne $H_f(x)$ vérifie $0 \leqslant H_f(x) \leqslant LI_n$.

Alors, si on choisit un pas $\alpha < \frac{2}{L}$, on a les résultats du Théorème 1, et si $\nabla f(x_0) \neq 0$, on a les deux possibilités suivantes :

— soit la suite (x_k) converge vers un minimum local de f,

— *soit* $\lim_{k\to\infty} \|x_k\| = +\infty$.

Démonstration. Les hypothèses sur la Hessienne donnent que ∇f est bien L-Lipschitz : en effet, toutes les valeurs propres de $H_f(x)$ sont entre 0 et L. Et donc d'après le lemme 2.1 pour tout h on a $\|H_f(x)(h)\| \leqslant L\|h\|$. En appliquant la formule de Taylor à l'ordre 1 à ∇f, on obtient directement que ∇f est bien L-Lipschitz, au moins sur tout segment $[x, y] \subset S_0$, ce qui est suffisant pour la preuve du Théorème 1.

Si on suppose que la suite x_k ne tend pas en norme vers $+\infty$, elle admet une sous-suite bornée. En effet la négation de $\|x_k\| \to \infty$ est qu'il existe $M > 0$ tel que pour tout $N_0 \in \mathbb{N}$, il existe $N \geqslant N_0$ tel que $\|x_k\| \leqslant M$, on peut donc construire l'extraction par récurrence en prenant $N_0 = \varphi(k) + 1$ et en posant $\varphi(k+1) = N$, et on a donc $\|x_{\varphi(k)}\| \leqslant M$ pour tout k.

On peut en extraire de nouveau par compacité une suite convergente vers un point x_∞, avec donc par continuité $\nabla f(x_\infty) = 0$ et $f(x_\infty) \leqslant f(x_1) < f(x_0)$ puisqu'on a supposé que $\nabla f(x_0) \neq 0$, et que donc on a $f(x_k) \leqslant f(x_1) < f(x_0)$ pour tout $k \geqslant 1$. On peut donc trouver un ε tel que $B(x_\infty, \varepsilon) \subset S_0$. On va montrer qu'à partir d'un certain rang la suite $\|x_k - x_\infty\|$ est décroissante et x_k reste dans $B(x_\infty, \varepsilon)$. Supposons que $x_k \in B(x_\infty, \varepsilon)$ (c'est vrai pour un k assez grand vu que x_∞ est une valeur d'adhérence des x_k). On écrit alors

$$\begin{aligned}\|x_{k+1} - x_\infty\| &= \|x_k - \alpha\nabla f(x_k) - x_\infty\| = \|x_k - x_\infty - \alpha(\nabla f(x_k) - \nabla f(x_\infty))\| \\ &= \left\| \int_0^1 (I_n - \alpha H_f(x_\infty + t(x_k - x_\infty))(x_k - x_\infty) \mathrm{d}t \right\| \\ &\leqslant \int_0^1 \max(1, |1 - \alpha L|)\|x_k - x_\infty\| \mathrm{d}t = \|x_k - x_\infty\|.\end{aligned}$$

En effet, les valeurs propres de $I_n - \alpha H_f(x)$ appartiennent à $[1 - \alpha L, 1]$ lorsque $x \in S_0$, et d'autre part on a $\alpha < \frac{2}{L}$, donc $1 - \alpha L > 1 - 2 = -1$, donc $|1 - \alpha L| < 1$.

La suite $\|x_k - x_\infty\|$ est donc décroissante à partir du moment où x_k rentre dans $B(x_\infty, \varepsilon)$, et sa limite est 0, puisque la suite extraite initiale converge vers x_∞. Donc la suite $(x_k)_{k\in\mathbb{N}}$ converge vers x_∞.

Enfin, puisque la fonction est convexe sur $B(x_\infty, \varepsilon)$ et que x_∞ est un point critique, alors c'est un minimum sur $B(x_\infty, \varepsilon)$, donc c'est bien un minimum local. □

Le dernier théorème de convergence que l'on va énoncer concerne le cas où on a encore plus d'information, lorsque la Hessienne est définie positive dans un voisinage du minimum local, et on obtiendra le même genre de résultat que dans l'étude du cas test $x \mapsto \frac{1}{2}\langle x, Ax\rangle + \langle b, x\rangle$.

Théorème 3. *Convergence linéaire sous condition d'ellipticité.*

Soit $f : \Omega \to \mathbb{R}$ une fonction C^2, bornée inférieurement. On se fixe un point $x_0 \in \Omega$ et on suppose les deux conditions suivantes, pour $0 < \ell \leqslant L$:

— *L'ensemble $S_0 = \{x \in \Omega, f(x) \leqslant f(x_0)\}$ est fermé dans $\mathbb{R}^n$.*

— *Pour tout $x \in S_0$, la Hessienne $H_f(x)$ vérifie $\ell I_n \leqslant H_f(x) \leqslant L I_n$ (on dit que H_f est elliptique sur S_0).*

Alors, si on choisit un pas $\alpha < \frac{2}{L}$, la suite (x_k) converge linéairement vers un point de minimum local strict de f, avec un taux inférieur ou égal à $r(\alpha) = \max(|1 - \ell\alpha|, |1 - L\alpha|)$.

Démonstration. On note $F(x) = x - \alpha\nabla f(x)$, de sorte que l'on a $x_{k+1} = F(x_k)$.

En notant que $F'(x) = I_n - \alpha H_f(x)$, les hypothèses nous donnent que $F'(x)$ a ses valeurs propres entre $1-\alpha L$ et $1-\alpha\ell$. D'après le lemme 2.1, on a donc que $\|F'(x)\| \leqslant \max(|1-\alpha\ell|, |1-\alpha L|) = r(\alpha)$. On est toujours dans le cadre des théorèmes précédents, et on a donc que $[x_k, x_{k-1}]$ est inclus dans S_0 pour $k \geqslant 1$. On applique la formule de Taylor à l'ordre 1 pour F :

$$\|x_{k+1} - x_k\| = \|F(x_k) - F(x_{k-1})\| \leqslant \sup_{x\in[x_k, x_{k-1}]} \|F'(x)\|\|x_k - x_{k-1}\| \leqslant r(\alpha)\|x_k - x_{k-1}\|.$$

Le critère de convergence du précédent chapitre nous donne donc que (x_k) converge linéairement à un taux inférieur ou égal à $r(\alpha)$. Sa limite, notée $x_* \in S_0$, est donc un point critique d'après les théorèmes de convergence précédents.

Comme la Hessienne est définie positive en x_*, on obtient que x_* est un point de minimum local strict. □

On a déjà montré dans l'étude du cas test que le meilleur choix de α pour minimiser $r(\alpha)$. On obtient que le minimum de $r(\alpha)$ est atteint lorsque $\alpha = \alpha_* = \frac{2}{\ell+L}$, et on a alors $r(\alpha_*) = \frac{L-\ell}{L+\ell}$.

En pratique, on n'a pas besoin d'avoir une estimation de la Hessienne sur tout S_0 pour avoir convergence linéaire, si l'on sait déjà que la suite converge vers un minimum x_* et si la Hessienne est définie positive au point x_*.

Exercice 2.1. *On suppose que la fonction f est de classe C^2, et que la méthode de gradient à pas fixe α produit une suite qui converge vers x_*, un point de minimum local strict tel que $\ell I_n \leqslant H_f(x_*) \leqslant L I_n$, avec $0 < \ell \leqslant L$ et $0 < \alpha < \frac{2}{L}$. Montrer que la convergence est linéaire avec un taux inférieur ou égal à $r(\alpha)$.*

2.3 Descente de gradient à pas optimal

On s'intéresse maintenant à ce qui se passe si on choisit le pas α_k de façon optimale, au sens où la fonction à optimiser diminue le plus possible.

Définition 2.5. *On dit que la méthode de descente donnée par*

$$x_{k+1} = x_k + \alpha_k d_k,$$

où d_k est une direction de descente, est une descente à pas optimal si le pas α_k minimise la fonction réelle $t \mapsto h(t) = f(x_k + t d_k)$ sur $\mathbb{R}_+$.

Dans le cas où $d_k = -\nabla f(x_k)$, on parle donc de descente de gradient à pas optimal.

On a tout d'abord une relation d'orthogonalité lorsque le pas est optimal

Proposition 2.5. *Si α_k est un pas optimal, alors $\nabla f(x_{k+1})$ et d_k sont orthogonaux.*

Démonstration. On a $h'(t) = \langle \nabla f(x_k + t d_k), d_k \rangle$. Comme α_k minimise h sur $\mathbb{R}_+$, on a $\alpha_k > 0$ (puisque $h'(0) < 0$, on a supposé que d_k était une direction de descente), le minimum est donc dans l'ouvert $]0, +\infty[$ et donc $h'(\alpha_k) = 0$, ce qui donne $\langle \nabla f(x_{k+1}), d_k \rangle = 0$. □

On peut donc donner une intérprétation géométrique de cette proposition.

Remarque 2.2. *Au point x_{k+1}, le gradient de f est orthogonal aux courbes de niveau de f (cf. Proposition 2.2), donc on peut en quelque sorte dire que la direction d_k est « tangente » à l'hypersurface de niveau de f passant par x_{k+1}. En dimension deux, il s'agit simplement de la ligne de niveau. On peut formaliser un peu plus tout cela à l'aide du théorème des fonctions implicites (lorsque le gradient est non-nul), mais l'important est de comprendre que tant que d_k n'est pas tangent à l'hypersurface au point $x_k + t d_k$, alors la direction d_k « traverse » cette hypersurface, de sorte que la valeur de f diminue strictement, donc t n'est pas un minimum de h dans cette direction.*

Remarque 2.3. *Dans le cas particulier de la descente de gradient à pas optimal, on obtient que les directions de descentes successives sont orthogonales : $\langle \nabla f(x_{k+1}), \nabla f(x_k) \rangle = 0$. La direction de descente $d_k = -\nabla f(x_k)$ est orthogonale à l'hypersurface de niveau au point de départ x_k et tangente à l'hypersurface de niveau au point d'arrivée x_{k+1}.*

Dessin illustratif en dimension deux.

Avantages et inconvénients de la méthode. Tout d'abord, le fait d'avoir un pas variable permet d'avoir une certaine souplesse que ne permettait pas la descente de gradient à pas fixe. Ensuite, la méthode à pas fixe imposait un pas suffisamment petit, ce qui peut poser des problèmes de lenteur au démarrage (bien que la vitesse de convergence soit linéaire, ce n'est qu'un résultat asymptotique), comme on peut l'observer sur des fonctions du type $x \mapsto 1-\frac{1}{1+\|x\|^2}$, qui ne décroissent pas suffisamment lorsque le point initial est loin du minimum. Ceci n'est plus un problème avec une méthode à pas optimal.

Du côté des inconvénients, à moins que l'on puisse faire le calcul exact du pas (ce qui est le cas par exemple pour le cas test quadratique $f(x) = \frac{1}{2}\langle x, Ax\rangle + \langle b, x\rangle$), on doit faire la recherche du pas optimal par une méthode d'optimisation unidimensionnelle. On connaît des méthodes robustes telles que la méthode de la section dorée, qui nous donne un taux de convergence fixe, mais dans tous les cas le pas optimal sera une approximation, qui peut parfois être coûteuse en terme du nombre d'évaluations de la fonction.

En pratique, le taux de convergence linéaire, pour la descente de gradient à pas optimal, n'est pas vraiment meilleur que celui de la descente de gradient à pas fixe (lorsque le pas est bien choisi), comme le montre la proposition suivante (admise).

Proposition 2.6. *Pour la fonction $f : x \mapsto \frac{1}{2}\langle x, Ax\rangle + \langle b, x\rangle$, avec A symétrique définie positive dont on note ℓ et L la plus petite et la plus grande valeur propre, si on note x_* l'unique solution de $Ax + b = 0$ et si (x_k) correspond à la suite des itérés de la descente de gradient à pas optimal, on a*

$$\|x_{k+1} - x_*\|_A \leqslant \frac{L-\ell}{L+\ell}\|x_k - x_*\|_A,$$

où $\|x\|_A = \sqrt{\langle x, Ax\rangle}$. De plus il existe des x_0 telle que l'inégalité précédente soit une égalité.

On obtient donc que le taux de convergence est proche de un lorsque la matrice est mal conditionnée (si $\ell \ll L$), même en prenant le pas optimal. En pratique plutôt que de choisir un pas optimal exact, on se contentera de satisfaire certains critères pour obtenir un pas « satisfaisant ».

2.3.1 Critères de choix de pas pour une recherche de pas approchée

La recherche d'un pas « satisfaisant » est souvent appelée *recherche linéaire* ou *linesearch* puisque cela revient à observer comment se comporte la fonction le long de la ligne donnée par la direction de descente.

On donne ici quelques critères souvent utilisés correspondant à se dire que la fonction à optimiser $h(t)$ décroit suffisamment entre 0 et α_k, ou que la pente a suffisamment réduit...

Définition 2.6. *Règle d'Armijo. On dit que α_k satisfait la règle d'Armijo pour le paramètre $c_1 \in]0, 1[$ si on a*

$$f(x_{k+1}) \leqslant f(x_k) + c_1\alpha_k\langle \nabla f(x_k), d_k\rangle.$$

Cela revient encore à dire que $h(\alpha_k) \leqslant h(0) + c_1\alpha_k h'(0)$. C'est toujours faisable en prenant α_k suffisamment petit, puisque $h(t) - h(0) \sim th'(0) < c_1 th'(0)$ pour t assez petit (on se rappelle que comme d_k est une direction de descente, on a $h'(0) < 0$).

Dessin en 1d de l'interprétation en terme de corde et de tangente en 0 pour h.

Cette règle n'empêche jamais de prendre un pas trop petit, et on rajoute souvent une autre condition permettant justement de prendre un pas suffisamment grand.

Une manière d'y parvenir consiste à demander que le gradient (si on y a accès) ait suffisamment diminué.

Définition 2.7. *Règle de Wolfe. On dit que α_k satisfait la règle de Wolfe pour les paramètres c_1 et c_2 (avec $0 < c_1 < c_2 < 1$) lorsqu'il satisfait la règle d'Armijo pour la constante c_1 et que l'on a de plus*

$$\langle \nabla f(x_{k+1}), d_k \rangle \geqslant c_2 \langle \nabla f(x_k), d_k \rangle.$$

Cela revient à dire que $h'(\alpha_k) \geqslant c_2 h'(0)$. Par exemple, si α_k est un pas optimal, on a $h'(\alpha_k) = 0$ et la règle est toujours satisfaite.

Dessin en 1d de l'interprétation en terme de tangentes en 0 et en t tel que la pente soit $c_2 h'(0)$ pour h.

En pratique on aimerait prendre c_2 suffisamment petit, de sorte de s'approcher du pas optimal, mais alors c'est plus coûteux de trouver un pas qui satisfait cette condition.

Algorithme pour satisfaire la règle de Wolfe. On peut donner un algorithme simple de dichotomie pour trouver un pas α_k qui satisfasse la règle de Wolfe, en encadrant le pas par un $\alpha_{\min}$ qui satisfait la règle d'Armijo, mais pas la deuxième condition, et un $\alpha_{\max}$ qui ne satisfait pas la règle d'Armijo.

- On part par exemple de $\alpha_{\min} = 1$ et on le divise par deux autant de fois que nécessaire jusqu'à ce qu'il satisfasse la règle d'Armijo.
- Si à ce moment-là $\alpha_{\min}$ satisfait la deuxième règle on a obtenu ce qu'on voulait et on s'arrête en renvoyant ce pas $\alpha_{\min}$ comme résultat, sinon on part de $\alpha_{\max} = \alpha_{\min}$, que l'on multiplie par deux autant de fois que nécessaire jusqu'à ce qu'il ne satisfasse plus la règle d'Armijo.
- Ensuite, par dichotomie, on pose $\alpha = \frac{1}{2}(\alpha_{\max} + \alpha_{\min})$. Si α satisfait les deux critères on peut s'arrêter et renvoyer ce pas α comme résultat, sinon on met à jour $\alpha_{\max}$ (resp. $\alpha_{\min}$) à la valeur α si α ne satisfait pas la règle d'Armijo (resp. satisfait la règle d'Armijo mais pas le deuxième critère), et on recommence cette étape.

Remarque 2.4. *Pour la deuxième étape, si on avait déjà divisé par deux au moins une fois $\alpha_{\min}$, il suffit de multiplier par deux une seule fois, sinon il se peut qu'on ait besoin de multiplier plusieurs fois.*

Proposition 2.7. *Si f est de classe C^1, et bornée inférieurement, alors l'algorithme donné ci-dessus s'arrête en un nombre fini d'étapes (et renvoie donc un pas satisfaisant la règle de Wolfe).*

Démonstration. L'algorithme consiste en trois boucles, il suffit donc de montrer que chacune s'arrête.

Pour la première, on a déjà remarqué que la règle d'Armijo était toujours satisfaite dès que le pas était suffisamment petit.

Pour la deuxième, si elle ne s'arrêtait pas on aurait l'existence de $\alpha_{\max}$ arbitrairement grand tel que $h(\alpha_{\max}) \leqslant h(0) + c_1 \alpha_{\max} h'(0)$, en contradiction avec le fait que h est bornée inférieurement (comme f), puisque $h'(0) < 0$.

Enfin si la troisième boucle ne s'arrêtait pas, on aurait l'existence de $\alpha_{\max}$ et $\alpha_{\min}$ aussi proches qu'on veut d'une valeur limite α_* et satisfaisant $h(\alpha_{max}) > h(0) + c_1 \alpha_{\max} h'(0)$ et $h(\alpha_{\min}) \leqslant h(0) + c_1 \alpha_{\min} h'(0)$. On obtient donc $h(\alpha_{\max}) - h(\alpha_{\min}) > c_1 h'(0)(\alpha_{\max} - \alpha_{\min})$, ce qui nous donnerait à la limite $h'(\alpha_*) \geqslant c_1 h'(0)$. Mais on aurait également $h'(\alpha_{\min}) < c_2 h'(0)$ et donc à la limite $h'(\alpha_*) \leqslant c_2 h'(0)$, ce qui donne $c_1 h'(0) \leqslant c_2 h'(0)$, en contradiction avec le fait que $h'(0) < 0$ et $0 < c_1 < c_2 < 1$. □

L'intérêt de ces règles est qu'elles permettent de démontrer des résultats théoriques de convergence, comme le théorème de Zoutendijk (admis).

Théorème 4. *Théorème de Zoutendijk. On suppose que ∇f est L-Lipschitz, que le pas α_k satisfait la règle de Wolfe, et que f est bornée inférieurement. Si on note θ_k l'angle entre $-\nabla f(x_k)$ et d_k, de telle sorte que $\cos\theta_k = \frac{\langle -\nabla f(x_k), d_k\rangle}{\|\nabla f(x_k)\|\|d_k\|}$ (on peut prendre $\theta_k = 0$ pour une descente selon la direction du gradient), alors on a*

$$\sum_k \cos^2\theta_k \|\nabla f(x_k)\| < +\infty.$$

Ce théorème nous donne donc directement que (sous les bonnes hypothèses) $\nabla f(x_k)$ converge en norme vers 0 dans le cas d'une descente de gradient dans lequel le pas satisfait la règle de Wolfe.

Exercice 2.2. *Règle de Wolfe, de Goldstein, et gradient à pas fixe.*

Quelles sont les conditions pour le que pas fixe de la méthode de descente de gradient satisfasse la règle de Wolfe (ou de Goldstein) dans le cas où $f(x) = \frac{1}{2}\langle x, Ax\rangle + \langle b, x\rangle$?

Chapitre 3

La méthode du gradient conjugué

La méthode du gradient conjugué est une méthode permettant de résoudre un problème d'optimisation quadratique correspondant au cas test des parties précédentes : minimiser $\frac{1}{2}\langle x, Ax\rangle + \langle b, x\rangle$ sur $\mathbb{R}^n$, avec A une matrice définie positive. Cela revient à résoudre $Ax + b = 0$.

La méthode a d'abord été introduite comme méthode de résolution théorique exacte de l'équation en un nombre fini d'étapes, puis a obtenu un regain d'intérêt lorsque l'on s'est aperçu que la convergence était bonne dès les premiers pas, et que c'était une méthode adaptée aux grandes dimensions, en particulier quand le calcul de Ax est peu coûteux (par exemple si A est une matrice creuse provenant de la discrétisation d'un opérateur comme le Laplacien).

On a vu précédemment que l'on ne pouvait pas obtenir une convergence très rapide lorsqu'on prenait la direction du gradient, même en prenant un pas optimal (qu'il est possible de calculer de manière exacte dans ce cas particulier), dans le cas où la matrice est mal conditionnée, ce qui arrive souvent en grande dimension. On va donc prendre des directions différentes, qui seront conjuguées au sens où les directions forment une base orthogonale pour le produit scalaire $(x, y) \mapsto \langle x, Ay\rangle$.

3.1 Définition et interprétation en terme de méthode de descente

On peut écrire la méthode de gradient conjugué comme une méthode de descente à pas optimal. Pour être cohérent avec les notations standard de la méthode, on notera ici p_k la direction de descente.

Définition 3.1. *Algorithme du gradient conjugué.*

On se donne la fonction $f : x \mapsto \frac{1}{2}\langle x, Ax\rangle + \langle b, x\rangle$ que l'on cherche à minimiser sur $\mathbb{R}^n$, avec A une matrice symétrique définie positive, et $b \in \mathbb{R}^n$.

On se donne $x_0 \in \mathbb{R}^n$, on pose $r_0 = Ax_0 + b = \nabla f(x_0)$ et $p_0 = -r_0 = -\nabla f(x_0)$.

Tant que $r_k \neq 0$ on pose :

$$x_{k+1} = x_k + \alpha_k p_k, \text{ où le pas est donné par } \alpha_k = -\frac{\langle p_k, r_k\rangle}{\langle p_k, Ap_k\rangle}.$$

On met à jour la direction de descente en posant

$$r_{k+1} = Ax_{k+1} + b = \nabla f(x_{k+1}) \quad \text{et} \quad p_{k+1} = -r_{k+1} + \beta_k p_k,$$

où le coefficient d'ajustement β_k est calculé de manière à ce que $\langle p_{k+1}, Ap_k\rangle = 0$: on dit que les directions p_k et p_{k+1} sont conjuguées pour A. Cela donne la formule suivante

$$\beta_k = \frac{\langle r_{k+1}, Ap_k\rangle}{\langle p_k, Ap_k\rangle}.$$

Proposition 3.1. *Si $r_k \neq 0$, le pas α_k est bien défini et strictement positif, et c'est le pas optimal au point x_k dans la direction de descente p_k.*

Démonstration. Montrons-le par récurrence. On suppose que α_k est bien défini et strictement positif (donc que $\langle p_k, Ap_k\rangle$), que p_k est une direction de descente (c'est bien le cas pour $k = 0$ puisque si $r_0 \neq 0$, alors p_0 est l'opposé du gradient) et enfin que α_k est le pas optimal pour cette direction de descente (on va voir par la suite que c'est bien le cas également quand $k = 0$). On suppose que $r_{k+1} \neq 0$.

D'après la proposition 2.5, comme α_k est un pas optimal on a $\langle \nabla f(x_{k+1}), p_k\rangle = 0$, c'est à dire $\langle r_{k+1}, p_k\rangle = 0$. En prenant le produit scalaire avec r_{k+1} dans la définition de p_{k+1}, on obtient donc $\langle r_{k+1}, p_{k+1}\rangle = -\|r_{k+1}^2\| < 0$ et donc $p_{k+1} \neq 0$. Comme A est symétrique définie positive, on obtient bien que $\langle p_{k+1}, Ap_{k+1}\rangle > 0$ et que donc α_{k+1} est bien défini et strictement positif.

En posant $h(t) = f(x_{k+1} + tp_{k+1})$, on obtient

$$\begin{aligned} h'(\alpha_{k+1}) = \langle \nabla f(x_{k+1} + \alpha_{k+1}p_{k+1}), p_{k+1}\rangle &= \langle A(x_{k+1} + \alpha_{k+1}p_{k+1}) + b, p_{k+1}\rangle \\ &= \langle r_{k+1}, p_{k+1}\rangle + \alpha_{k+1}\langle p_{k+1}, Ap_{k+1}\rangle = 0, \end{aligned}$$

donc α_{k+1} est bien le minimum de la fonction h (qui est un polynôme de degré 2 avec un coefficient dominant strictement positif $\langle p_{k+1}, Ap_{k+1}\rangle$), et la direction p_{k+1} est bien une direction de descente puisque $h'(0) < 0$ (sinon le minimum de la fonction serait atteint pour un $t < 0$). Ce calcul étant valable également pour la première étape, on obtient que α_0 est bien le pas optimal pour la direction de descente p_0, cela permet donc bien d'initialiser la récurrence. □

Remarque 3.1. *Si on avait pris $\beta_k = 0$ à la place de la formule donnée on aurait obtenu la méthode de descente de gradient à pas optimal.*

3.2 Présentation standard et convergence

On donne tout d'abord la présentation standard de l'algorithme, qui exploite certaines identités pour n'avoir à calculer qu'une seule fois par itération un produit entre A et un vecteur : le produit Ap_k (on l'utilise ensuite deux fois, dans le calcul de α_k et celui de r_{k+1}).

Proposition 3.2. *Forme standard de la boucle de gradient conjugué. Si $r_k \neq 0$, alors on a*

$$\begin{aligned} \alpha_k &= \frac{\|r_k\|^2}{\langle p_k, Ap_k\rangle}, \\ x_{k+1} &= x_k + \alpha_k p_k, \\ r_{k+1} &= r_k + \alpha_k Ap_k, \\ p_{k+1} &= -r_{k+1} + \frac{\|r_{k+1}\|^2}{\|r_k\|^2} p_k. \end{aligned}$$

Démonstration. On a déjà montré dans la preuve de la proposition 3.1 que $\langle r_k, p_k\rangle = -\|r_k\|^2$, et donc cela donne l'expression de α_k. L'expression de x_{k+1} est inchangée. Pour l'expression de r_{k+1}, il suffit de remarquer que $r_{k+1} - r_k = Ax_{k+1} + b - (Ax_k + b) = A(x_{k+1} - x_k)$, ce qui donne la formule. Enfin il reste à montrer que $\beta_k = \frac{\|r_{k+1}\|^2}{\|r_k\|^2}$.

On tout d'abord montrer que r_{k+1} et r_k sont orthogonaux. C'est vrai pour $k = 0$ puisqu'on a déjà vu que $\langle r_{k+1}, p_k\rangle = 0$ et que pour $k = 0$ on a $p_0 = -r_0$. Dans le cas où $k \geqslant 1$, on observe d'abord que $\langle r_{k+1}, p_{k-1}\rangle = \langle r_k + \alpha_k Ap_k, p_{k-1}\rangle = 0$ puisque l'on a $\langle r_k, p_{k-1}\rangle = 0$ et $\langle Ap_k, p_{k-1}\rangle = 0$. On peut donc écrire $\langle r_{k+1}, r_k\rangle = \langle r_{k+1}, -p_k + \beta_{k-1}p_{k-1}\rangle = 0$.

L'idée est ensuite d'écrire $Ap_k = \frac{1}{\alpha_k}(r_{k+1} - r_k)$ et de le remplacer dans l'expression du numérateur pour obtenir $\langle r_{k+1}, Ap_k \rangle = \frac{1}{\alpha_k}\|r_{k+1}\|^2$. En utilisant que $\langle r_{k+1}, p_k \rangle = 0$ et que $\langle p_k, r_k \rangle = -\|r_k\|^2$, le dénominateur est $\langle p_k, Ap_k \rangle = \frac{1}{\alpha_k}\langle p_k, r_{k+1} - r_k \rangle = -\frac{1}{\alpha_k}\langle p_k, r_k \rangle = \frac{1}{\alpha_k}\|r_k\|^2$, ce qui donne le résultat. □

Au cours des preuves des deux dernières propositions, on a obtenu les relations d'orthogonalité suivantes : $\langle p_{k+1}, Ap_k \rangle = 0$ (directions de descente conjuguées), $\langle r_{k+1}, p_k \rangle = 0$ (grâce au fait que α_k est un pas optimal), puis en faisant des manipulations on a également obtenu $\langle r_{k+1}, p_{k-1} \rangle = 0$ et enfin $\langle r_{k+1}, r_k \rangle = 0$ (les directions consécutives des gradients sont orthogonales).

On va en fait montrer que les directions de descente sont toutes conjuguées deux à deux pour la matrice A, et que les gradients r_k sont tous orthogonaux deux à deux pour le produit scalaire usuel. Par conséquents, ces familles forment des familles orthogonales et ne peuvent être non-nulles qu'un nombre de fois limité par la dimension de l'espace. Ceci nous donnera qu'en théorie, si tous les calculs sont exacts, l'algorithme s'arrête en moins de n étapes.

Théorème 5. *L'algorithme s'arrête en au plus n étapes et on a pour tout k, tant que l'algorithme n'a pas terminé (c'est à dire si $r_k \neq 0$), on a égalité entre les sous-espaces suivants :*

$$\mathrm{Vect}(r_0, Ar_0, \ldots, A^k r_0) = \mathrm{Vect}(r_0, r_1, \ldots, r_k) = \mathrm{Vect}(p_0, p_1, \ldots, p_k).$$

De plus, pour tout i tel que $0 \leqslant i \leqslant k$, on a les relations d'orthogonalité suivantes :

(i) $\langle p_{k+1}, Ap_i \rangle = 0$,
(ii) $\langle r_{k+1}, p_i \rangle = 0$,
(iii) $\langle r_{k+1}, r_i \rangle = 0$.

Démonstration. Montrons l'égalité des sous-espaces par récurrence sur k. L'initialisation vient de ce que $p_0 = -r_0$. Supposons que $\mathrm{Vect}(r_0, Ar_0, \ldots, A^k r_0) = \mathrm{Vect}(r_0, r_1, \ldots, r_k) = \mathrm{Vect}(p_0, p_1, \ldots, p_k)$. Pour montrer l'égalité entre les deux derniers sous-espaces au rang $k+1$, il suffit donc de montrer que $p_{k+1} \in \mathrm{Vect}(r_0, r_1, \ldots, r_k, r_{k+1})$ et que $r_{k+1} \in \mathrm{Vect}(p_0, p_1, \ldots, p_{k+1})$. On utilise l'équation définissant p_{k+1}, qui nous donne $p_{k+1} = -r_{k+1} + \beta_k r_k$, et on obtient donc directement que p_{k+1} appartient à $\mathrm{Vect}(r_0, r_1, \ldots, r_k, r_{k+1})$. En écrivant alors $r_{k+1} = -p_{k+1} + \beta_k r_k$ et en utilisant le fait que $r_k \in \mathrm{Vect}(p_0, p_1, \ldots, p_k)$ on obtient bien que $r_{k+1} \in \mathrm{Vect}(p_0, p_1, \ldots, p_{k+1})$. Ensuite pour montrer l'égalité $\mathrm{Vect}(r_0, Ar_0, \ldots, A^k r_0) = \mathrm{Vect}(r_0, r_1, \ldots, r_k, r_{k+1})$, il suffit également de montrer que $A^{k+1} r_0 \in \mathrm{Vect}(r_0, r_1, \ldots, r_k, r_{k+1})$ et que $r_{k+1} \in \mathrm{Vect}(r_0, Ar_0, \ldots, A^k r_0)$. Cette dernière condition provient du fait que $r_{k+1} = r_k + \alpha_k Ap_k$, et comme r_k et p_k sont dans $\mathrm{Vect}(r_0, Ar_0, \ldots, A^k r_0)$ par hypothèse de récurrence, on obtient que Ap_k est dans $\mathrm{Vect}(Ar_0, A^2 r_0, \ldots, A^k r_0, A^{k+1} r_0)$ puis que $r_{k+1} \in \mathrm{Vect}(r_0, Ar_0, A^2 r_0, \ldots, A^k r_0, A^{k+1} r_0)$. Pour l'inclusion réciproque, on écrit (par hypothèse de récurrence) que $A^k r_0 = \sum_{i=0}^{k} \lambda_i p_i$, et donc (on a vu précédement que $\alpha_k > 0$ tant que $r_k \neq 0$)

$$A^{k+1} r_0 = \sum_{i=0}^{k} \lambda_i A p_i = \sum_{i=0}^{k} \lambda_i \frac{r_{i+1} - r_i}{\alpha_i} \in \mathrm{Vect}(r_0, r_1, \ldots, r_k, r_{k+1}).$$

Montrons maintenant les deux premières relations d'orthogonalité, également par récurrence. On sait déjà qu'on a $\langle p_{k+1}, Ap_k \rangle = 0$ et $\langle r_{k+1}, p_k \rangle = 0$, ce qui donne l'initialisation pour $k = 0$.

Supposons les relations vraies au rang $k-1$, et montrons qu'elles sont vraies au rang k. On a simplement besoin de montrer les relations d'orthogonalité pour $i \leqslant k-1$, puisqu'on les connaît déjà pour $i = k$. On a $\langle r_{k+1}, p_i \rangle = \langle r_k + \alpha_k Ap_k, p_i \rangle = \langle r_k, p_i \rangle + \alpha_k \langle p_k, Ap_i \rangle$ qui est nul par hypothèse de récurrence. Ensuite $\langle p_{k+1}, Ap_i \rangle = \langle -r_{k+1} + \beta_k p_k, Ap_i \rangle = -\langle r_{k+1}, Ap_i \rangle$ par hypothèse de récurrence. Mais comme $p_i \in \mathrm{Vect}(r_0, Ar_0, \ldots, A^i r_0)$, on obtient que $Ap_i \in \mathrm{Vect}(Ar_0, A^2 r_0, \ldots, A^{i+1} r_0)$ qui est inclus dans $\mathrm{Vect}(p_0, p_1, \ldots, p_{i+1})$. Comme on vient de voir que $\langle r_{k+1}, p_j \rangle = 0$ pour $j \leqslant k$, on obtient donc que $\langle r_{k+1}, Ap_i \rangle = 0$ et donc $\langle p_{k+1}, Ap_i \rangle = 0$.

Enfin la dernière relation d'orthogonalité vient du fait que pour $i \leqslant k$, $r_i \in \mathrm{Vect}(p_0, p_1, \ldots, p_i)$, d'après l'égalité des sous-espaces. On obtient donc que $\langle r_{k+1}, r_i \rangle = 0$, puisqu'on vient de montrer que $\langle r_{k+1}, p_j \rangle = 0$ pour $0 \leqslant j \leqslant k$.

Enfin pour montrer que l'algorithme s'arrête en au plus n étapes, il suffit de remarquer que tant que $r_k \neq 0$, la famille $(r_i)_{0 \leqslant i \leqslant k}$ est donc une famille orthogonale et donc ne peut avoir qu'au plus n vecteurs non-nuls, en particulier on a donc $r_n = 0$. □

Exercice 3.1. *Montrer que x_k est le minimiseur de f sur l'espace affine $x_0 + \mathrm{Vect}(\mathrm{p}_0, \mathrm{p}_1, \ldots, \mathrm{p}_k)$.*

Remarque 3.2. *On remarque que le nombre d'itérations peut être plus petit que n si le sous-espace $\mathrm{Vect}(r_0, Ar_0, \ldots, A^{n-1}r_0)$ (appelé sous-espace de Krylov) est de dimension plus petite que n. Par exemple si r_0 se décompose sur m vecteurs propres : on obtient alors $\Pi_\lambda(A - \lambda)r_0 = 0$, où le produit est pris sur les m valeurs propres, ce qui donne que $A^m r_0 \in Vect(r_0, Ar_0, \ldots A^{m-1}r_0)$, et la dimension du sous-espace est donc inférieure ou égale à m. Donc en particulier si A a seulement m valeurs propres distinctes alors l'algorithme s'arrête toujours en m étapes.*

En théorie la méthode du gradient conjugué pourrait donc servir pour obtenir une solution exacte du problème $Ax + b = 0$. Cependant en pratique, cette méthode se comporte mal au niveau des erreurs d'arrondi (les vecteurs perdent les propriétés d'orthogonalité au fur et à mesure que les erreurs d'arrondi se cumulent). Elle n'a donc jamais été utilisée pour ça (d'autres méthodes sont connues pour résoudre des systèmes linéaires, qui ne nécessitent même pas que la matrice soit symétrique définie positive), et malgré son introduction dans les années 50, elle n'a pas attiré beaucoup l'attention. C'est seulement lorsqu'on s'est rendu compte que la convergence pour les premières itérations était bonne qu'il y a eu un regain d'intérêt pour cette méthode.

Pour de nombreux problèmes en effet, le coût de calcul de Ax n'est pas très grand (par exemple si beaucoup de coefficients sont nuls dans A, on ne fait pas les calculs associés), même si la dimension n est grande. On ne peut par contre pas se permettre de faire n itérations, ce serait beaucoup trop long, mais on cherche tout de même une solution approchée. On va voir dans la proposition suivante (admise) que le gradient conjugué fournit une méthode qui converge linéairement avec un taux meilleur que celui du gradient à pas optimal.

Proposition 3.3. *Si on note ℓ et L la plus petite et la plus grande valeur propre de A, et si x_* est l'unique solution de $Ax + b = 0$, alors on a*

$$\|x_k - x_*\|_A \leqslant 2\Big(\frac{\sqrt{L} - \sqrt{\ell}}{\sqrt{L} - \sqrt{\ell}}\Big)^k \|x_0 - x_*\|_A,$$

où $\|x\|_A = \sqrt{\langle x, Ax \rangle}$.

On peut comparer cette proposition avec la proposition 2.6, en supposant que le nombre de conditionnement $\kappa = \frac{L}{\ell}$ est grand. Pour se fixer les idées prenons $\frac{L}{\ell}$ de l'ordre de 100. Le taux de convergence linéaire pour la méthode de descente de gradient à pas optimal est $\frac{L-\ell}{L+\ell} = 1 - \frac{2}{\kappa+1} \approx 0,98$, ce qui veut dire qu'on gagne un facteur 2% à chaque itération. Par contre pour la méthode du gradient conjugué, le taux est $\frac{\sqrt{L}-\sqrt{\ell}}{\sqrt{L}+\sqrt{\ell}} = 1 - \frac{2}{\sqrt{\kappa}+1} \approx 0,82$, ce qui fait qu'on gagne cette fois-ci 18% à chaque itération. Plus le nombre de conditionnement est grand, plus on a intérêt à prendre la méthode de gradient conjugué par rapport à la descente de gradient à pas optimal, même si les deux méthodes deviennent de plus en plus lentes lorsque le nombre de conditionnement croît. En pratique on cherche donc souvent à essayer de résoudre un problème auxiliaire dont le nombre de conditionnement est bien plus faible, c'est ce qu'on appelle le préconditionnement, que l'on va voir dans la suite.

La deuxième remarque que l'on peut faire concernant les deux propositions 2.6 et 3.3 est que le facteur $\frac{L-\ell}{L+\ell}$ est un facteur que l'on gagne à chaque itération dans la méthode de descente de gradient

à pas optimal, alors qu'on ne peut pas assurer que l'on gagne un facteur $\frac{\sqrt{L}-\sqrt{\ell}}{\sqrt{L}+\sqrt{\ell}}$ pour la méthode de gradient conjugué à chaque étape, par exemple le premier pas de la méthode est exactement un pas de descente de gradient à pas optimal. C'est pour cela qu'on a un facteur 2 devant le terme d'erreur et qu'on compare l'erreur à l'étape k globalement par rapport à l'erreur initiale, et non pas l'erreur à l'étape $k+1$ par rapport à l'étape k.

3.3 Préconditionnement

L'idée du préconditionnement est de « changer » la matrice A en une autre matrice dont le nombre de conditionnement est meilleur. Par exemple si on connaît une matrice M « proche » de A en un certain sens, et qui est facile à inverser, alors on cherchera à résoudre $M^{-1}Ax + M^{-1}b = 0$ ce qui revient exactement au même. Si la matrice $M^{-1}A$ est proche de l'identité, on s'attend à ce qu'elle ait un meilleur conditionnement. Malheureusement si on veut appliquer les méthodes de descente, on n'est même pas sûr que la matrice $M^{-1}A$ est symétrique.

Cette première difficulté peut être évitée avec la remarque suivante : la matrice M si elle est symétrique, peut s'écrire de la forme EE^T (une des méthodes pour trouver une telle E s'appelle factorisation de Cholesky), et alors plutôt que de s'intéresser à la matrice $M^{-1}A$, on va s'intéresser à $E^{-1}A(E^{-1})^T$, qui ont en fait le même nombre de conditionnement.

Exercice 3.2. *Montrer que si $M = EE^T$, alors $M^{-1}A$ et $E^{-1}A(E^{-1})^T$ ont les mêmes valeurs propres.*

Le système $Ax + b = 0$ peut être transformé en le système $E^{-1}A(E^{-1})^T\hat{x} + E^{-1}b = 0$ en posant $\hat{x} = E^T x$. Si on résout pour $\hat{x}$, on peut alors ensuite résoudre pour x simplement en posant $x = (E^{-1})^T\hat{x}$. On peut donc appliquer la méthode de gradient conjugué à la matrice $E^{-1}A(E^{-1})^T$ et au vecteur $E^{-1}b$ pour trouver $\hat{x}$. Le problème de ceci est qu'il faut connaître la matrice E et savoir appliquer son inverse efficacement. Cependant, en réécrivant les différentes relations de la boucle de l'algorithme, et en remplaçant $\hat{x}_k$ par $E^T x_k$ dans les expressions, on obtient des simplifications, et on s'aperçoit qu'on a simplement besoin de connaitre et de savoir efficacement inverser la matrice EE^T, qui correspond à notre matrice M initiale !

Exercice 3.3. *En notant $\hat{x}_k$, $\hat{r}_k$, $\hat{p}_k$ les points, gradients, et directions de descente de la méthode de gradient conjugué appliquée à la résolution du système $E^{-1}A(E^{-1})^T\hat{x} + E^{-1}b = 0$ (ou à la minimisation de $\frac{1}{2}\langle\hat{x}, E^{-1}A(E^{-1})^T\hat{x}\rangle + \langle E^{-1}b, \hat{x}\rangle$), écrire la boucle standard de gradient conjugué. En remplaçant $\hat{x}_k$ par $E^T x_k$, $\hat{r}_k$ par $E^{-1}r_k$ et $\hat{p}_k$ par $E^T p_k$ (attention, ce n'est pas le même remplacement à chaque fois), observer qu'on obtient les relations suivantes (en partant de $r_0 = Ax_0 + b$, et $p_0 = -M^{-1}r_0$, où $M = EE^T$), dès que $r_k \neq 0$:*

$$\begin{aligned}
\alpha_k &= \frac{\langle r_k, M^{-1}r_k\rangle}{\langle p_k, Ap_k\rangle},\\
x_{k+1} &= x_k + \alpha_k p_k,\\
r_{k+1} &= r_k + \alpha_k A p_k,\\
p_{k+1} &= -M^{-1}r_{k+1} + \frac{\langle r_{k+1}, M^{-1}r_{k+1}\rangle}{\langle r_k, M^{-1}r_k\rangle}p_k.
\end{aligned}$$

On obtient donc une méthode où on n'évalue qu'une fois par boucle l'inverse de M appliqué à r_k, et où on n'applique qu'une fois par boucle la matrice A au vecteur p_k.

Le choix du préconditionneur M est extrêmement vaste, et dépend souvent fortement du problème considéré. Il permet d'obtenir des convergences bien plus rapide. En général, on utilise un préconditionnement la plupart du temps lorsque l'on effectue la méthode de gradient conjugué.

3.4 Adaptation au cas non linéaire

L'adaptation au cas non linéaire (au sens où on ne cherche plus à résoudre une équation linéaire $Ax + b = 0$, mais bien une équation de la forme $\nabla f(x) = 0$ avec $x \mapsto \nabla f(x)$ non linéaire)est assez directe, à partir du moment où on a bien interprété la méthode comme une méthode de descente à pas optimal, avec des directions obtenues à partir du gradient, mais en les modifiant légèrement de telles sorte qu'elles soient conjuguées.

Les différences principales se trouvent dans trois ingrédients :

— On ne peut pas calculer le pas optimal α_k directement, il faut utiliser une méthode unidimensionnelle.

— On ne peut plus calculer r_k en utilisant la formule à partir de p_k, il faut recalculer à chaque étape le gradient en x_k.

— On ne sait pas comment choisir β_k de telle sorte que les directions soient « conjuguées » dans un certain sens (il n'y a plus de matrice A).

Deux choix principaux ont été proposés pour le calcul de β_k : un premier choix, appelé méthode de Fletcher-Reeves, où on garde la même formule $\beta_k = \frac{\|r_{k+1}\|^2}{\|r_k\|^2}$, et un deuxième choix appelé méthode de Polak-Ribière, où on essaye d'avoir une approximation de ce que serait A (une matrice correspondant au Hessien) et où on essaye de garder la conjuguaison par rapport à ce A. Pour cela, on approxime Ap_k comme s'il était donné par la formule du cas linéaire : $Ap_k = \frac{r_{k+1}-r_k}{\alpha_k}$, et on écrit alors $\beta_k = \frac{\langle r_{k+1}, Ap_k\rangle}{\langle Ap_k, p_k\rangle} = \frac{\langle r_{k+1}-r_k, r_{k+1}\rangle}{\langle r_{k+1}-r_k, p_k\rangle}$. Encore une fois, si on suppose que le pas α_k est optimal, on peut obtenir la même relation d'orthogonalité $\langle p_k, r_{k+1}\rangle = 0$ qui donne que $\langle p_k, r_k\rangle = -\|r_k\|^2$, et on obtient donc l'expression $\beta_k = \frac{\langle r_{k+1}-r_k, r_{k+1}\rangle}{\|r_k\|^2}$.

Il a été montré que la méthode de Fletcher-Reeves converge globalement (mais parfois assez lentement), alors que certains cas pathologiques peuvent se produirent avec la méthode de Polak-Ribière pour des fonctions et des conditions initiales très particulières. En pratique cette dernière donne de meilleurs résultats, et c'est celle-là qui est utilisée, avec une modification consistant à prendre $\beta_k = 0$ si le calcul donne un nombre négatif, ce qui permet d'assurer la convergence (cela consiste à repartir « à zéro » à ce moment-là, et à oublier les précédentes directions de descentes).

En règle générale, étant donné que les directions conjuguées forment une famille libre, cela a du sens de redémarrer la méthode toutes les n itérations (surtout si n est petit), en prenant $\beta_k = 0$ pour un passage dans la boucle.

Enfin, il faut également choisir une méthode d'optimisation unidimensionnelle pour choisir le pas α_k, là aussi il y a plusieurs choix dépendant du contexte, et on peut également appliquer des règles du type de la règle de Wolfe pour s'assurer que le pas choisi est convenable.

En résumé, voici l'algorithme de gradient conjugué adapté au cas non-linéaire. On s'est donné par exemple ici une tolérance relative ε par rapport à la norme du gradient à la première itération.

Définition 3.2. *Algorithme du gradient conjugué non-linéaire.*

On se donne une fonction f que l'on cherche à minimiser sur $\mathbb{R}^n$.

On se donne $x_0 \in \mathbb{R}^n$, on pose $r_0 = \nabla f(x_0)$ et $p_0 = -r_0 = -\nabla f(x_0)$.

Tant que $\|r_k\| > \varepsilon\|r_0\|$ on pose :

$$x_{k+1} = x_k + \alpha_k p_k, \text{ où } \alpha_k \text{ est le pas optimal dans la direction } p_k.$$

On met à jour la direction de descente en posant

$$r_{k+1} = \nabla f(x_{k+1}) \quad \text{et} \quad p_{k+1} = -r_{k+1} + \beta_k p_k,$$

où le coefficient d'ajustement β_k est calculé par la formule suivante

$$\beta_k = \begin{cases} \frac{\|r_{k+1}\|^2}{\|r_k\|^2} & \text{pour la méthode de Fletcher–Reeves,} \\ \max(\frac{\langle r_{k+1}-r_k, r_{k+1}\rangle}{\|r_k\|^2}, 0) & \text{pour la méthode de Polak–Ribière.} \end{cases}$$

Chapitre 4

Méthodes de Newton et quasi-Newton

On cherche à étendre les méthodes de Newton et de la sécante présentées brièvement au chapitre 1 dans la remarque 1.3 dans le cas où la dimension est supérieure ou égale à 2.

4.1 La méthode de Newton pour l'optimisation dans $\mathbb{R}^n$

La méthode de Newton pour résoudre un problème de minimisation consiste à approximer la fonction par son développement de Taylor à l'ordre 2 autour du point courant et à choisir le nouveau point comme un minimiseur de ce développement de Taylor.

Le développement de Taylor au point x (en supposant que la fonction f est de classe C^2 au voisinage de x) est donné par

$$f(x+h) = f(x) + \langle h, \nabla f(x)\rangle + \frac{1}{2}\langle h, H_f(x)h\rangle + o(\|h\|^2).$$

La méthode de Newton consiste donc si le point courant est x, à prendre pour point suivant le point $x+h$ où h minimise la fonction quadratique $h \mapsto f(x) + \langle h, \nabla f(x)\rangle + \frac{1}{2}\langle h, H_f(x)h\rangle$. Cette fonction est de la forme $h \mapsto \frac{1}{2}\langle h, Ah\rangle + \langle b, h\rangle + c$ avec $A = H_f(x)$ et $B = \nabla f(x)$, et on sait que le gradient est $h \mapsto Ah + b$, la condition de point critique s'écrit donc $Ah + b = 0$. On sait que ce point critique est unique et que c'est un minimum local dans le cas où A est symétrique définie positive, et il est alors donné par $h = -A^{-1}b$ c'est-à-dire $h = -H_f(x)^{-1}\nabla f(x)$. On remarque que cette formule correspond en dimension un au terme $-\frac{f'(x)}{f''(x)}$, et on retrouve bien la méthode décrite au chapitre 1 dans la remarque 1.3.

En pratique, étant donné que le développement de f au voisinage de x n'est pas exact, on peut aussi simplement prendre ce h comme une direction de descente et choisir un pas différent de 1.

Définition 4.1. *Méthode de Newton à pas fixe ou variable.*

On se donne une fonction $f : \mathbb{R}^n \to \mathbb{R}$, on suppose que l'on a accès à son gradient et sa hessienne. On se donne $x_0 \in \mathbb{R}^n$. Les itérées de la méthode de Newton sont données par la formule suivante :

$$x_{k+1} = x_k + \alpha_k d_k, \text{ avec } d_k = -H_f(x_k)^{-1}\nabla f(x_k),$$

en supposant que la hessienne de f est bien inversible au point x_k. La méthode de Newton à pas fixe correspond au choix $\alpha_k = 1$ (attention, dans ce cadre dès qu'on parle de pas fixe, c'est forcément 1, et pas un $\alpha > 0$ arbitraire). La méthode de Newton à pas variable consiste à choisir $\alpha_k > 0$ selon une recherche de pas unidimensionnelle (on peut chercher à avoir un pas optimal, ou se contenter de satisfaire une règle telle que la règle de Wolfe).

Remarque 4.1. *On n'a pas supposé ici que la hessienne de f était définie positive au point x_k, et si ce n'est pas le cas, il se peut que cette méthode ne soit pas une méthode de descente.*

En fait la méthode de Newton à pas fixe peut être aussi vue comme une méthode de recherche de zéros d'une fonction $g : \mathbb{R}^n \to \mathbb{R}^n$, en écrivant $x_{k+1} = x_k - g'(x_k)^{-1}g(x_k)$. Cela correspond à écrire un développement de Taylor à l'ordre un pour g et à résoudre l'équation linéaire $g(x) + g'(x)h = 0$ plutôt que l'équation $g(x + h) = 0$. Dans le cas de l'optimisation, la fonction g qui nous intéresse est le gradient de f, et on retombe sur la méthode donnée.

Dans le cas où la hessienne est symétrique définie positive, alors $H_f(x_k)^{-1}$ est aussi définie positive, et on obtient bien que la méthode est une méthode de descente : on a $\langle \nabla f(x_k), d_k \rangle = -\langle \nabla f(x_k), H_f(x)^{-1} \nabla f(x_k) \rangle$ qui est strictement négatif si $\nabla f(x_k) \neq 0$.

On a un premier théorème utile pour la méthode de Newton à pas fixe, qui nous donne qu'au voisinage des points critiques non-dégénérés, la convergence a bien lieu, et que dans ce cas-là elle est quadratique. On va en fait montrer le théorème dans le cadre général de la méthode de Newton pour la recherche d'un zéro d'une fonction $g \in C^1(\mathbb{R}^n, \mathbb{R}^n)$.

Théorème 6. *Convergence locale quadratique pour la méthode de Newton générale de recherche d'un zéro. On se donne une fonction $g \in C^1(\mathbb{R}^n, \mathbb{R}^n)$, et on suppose que l'on a un point $x_* \in \mathbb{R}^n$ qui satisfait les trois critères suivants :*

— Zéro de g : $g(x_*) = 0$.

— Non-dégénéré : $g'(x_*)$, vue comme une matrice de $M_n(\mathbb{R})$, est inversible.

— Régularité : g' est Lipschitzienne au voisinage de x_*.

Alors il existe $\varepsilon > 0$ tel que pour toute initialisation x_0 dans $B(x_*, \varepsilon)$, les itérées de la méthode de Newton générale données par $x_{k+1} = x_k - g'(x_k)^{-1}g(x_k)$ convergent quadratiquement vers x_* : il existe une constante C telle que

$$x_k \to x_* \text{ quand } k \to \infty \text{ et pour } k \geqslant 0, \text{ on a } \|x_{k+1} - x_*\| \leqslant C\|x_k - x_*\|^2.$$

Démonstration. On a par les hypothèses (et comme $A \to A^{-1}$ est continue de $GL_n(\mathbb{R})$ (ouvert) dans $M_n(\mathbb{R})$), l'existence de $\varepsilon_0 > 0$, M et K tel que pour g' est M-Lipschitzienne sur $x \in B(x_*, \varepsilon_0)$, et que $\|g'(x)^{-1}\| \leqslant K$ pour tout $x \in B(x_*, \varepsilon)$.

On peut donc utiliser l'inégalité suivante, pour $x \in B(x_*, \varepsilon)$

$$\begin{aligned}\|g(x_*) - g(x) - g'(x)(x_* - x)\| &= \left\| \int_0^1 [g'(x + t(x_* - x)) - g'(x)](x_* - x) \mathrm{d}t \right\| \\ &\leqslant \int_0^1 M|t|\|x_* - x\|^2 \mathrm{d}t = \frac{M}{2}\|x - x_*\|^2.\end{aligned}$$

Comme $g(x_*) = 0$, on obtient donc que

$$x_{k+1} - x_* = x_k - x_* - g'(x_k)^{-1}(g(x_k) - g(x_*)) = g'(x_k)^{-1}[g(x_*) - g(x_k) - g'(x_k)(x_* - x_k)],$$

et donc si $x_k \in B(x_*, \varepsilon_0)$, on obtient

$$\|x_{k+1} - x_*\| \leqslant \|g'(x_k)^{-1}\| \frac{M}{2}\|x_k - x_*\|^2 \leqslant \frac{KM}{2}\|x_k - x_*\|^2.$$

Si on prend $\varepsilon = \min(\varepsilon_0, \frac{1}{KM})$, on a donc que pour $x_k \in B(x_*, \varepsilon)$, l'inégalité ci-dessus est vérifiée, et $\|x_k - x_*\| \leqslant \frac{1}{KM}$, donc on obtient $\|x_{k+1} - x_*\| \leqslant \frac{1}{2}\|x_k - x_*\|$, ce qui montre bien que $x_{k+1} \in B(x_*, \varepsilon)$, et que donc par récurrence dès que x_0 est dans $B(x_*, \varepsilon)$, alors tous les x_k y sont aussi, et convergent vers x_*. D'autre part l'inégalité précédente montre la convergence quadratique (on peut donc prendre $C = \frac{KM}{2}$ dans l'énoncé). □

On peut donc reformuler ce théorème dans le cadre de l'optimisation, pour la méthode de Newton à pas fixe (égal à 1).

Corollaire 4.1. *Convergence locale quadratique pour la méthode de Newton à pas fixe.*

On suppose que l'on a un point $x_* \in \mathbb{R}^n$ qui satisfait les trois critères suivants :

— Point critique : $\nabla f(x_*) = 0$.

— Non-dégénéré : $H_f(x_*)$ est inversible.

— Régularité : H_f est Lipschitzienne au voisinage de x_*.

Alors il existe $\varepsilon > 0$ tel que si la méthode de Newton à pas fixe est initialisée avec x_0 dans $B(x_*, \varepsilon)$, alors les itérées convergent quadratiquement vers x_* : il existe une constante C telle que

$$x_k \to x_* \text{ quand } k \to \infty \text{ et pour } k \geqslant 0, \text{ on a } \|x_{k+1} - x_*\| \leqslant C\|x_k - x_*\|^2.$$

En particulier ici, on n'a pas supposé que la matrice hessienne était définie positive, et le théorème montre bien que la méthode peut converger vers un point de maximum global, ou même un point selle. La condition de non-dégénerescence est également importante pour avoir la convergence quadratique, même pour un point de minimum global strict :

Exercice 4.1. *Montrer que la convergence vers 0 des itérées la méthode de Newton à pas fixe pour la minimisation de $x \mapsto x^4$ est linéaire dès que $x_0 \neq 0$.*

Si on veut un critère de convergence vers un minimum, une manière de faire est donc déjà d'avoir une hessienne toujours définie positive. En fait, ceci nous donne alors une fonction strictement convexe. On va voir que dans le cas convexe, on peut s'assurer de la convergence vers un minimum en prenant par exemple le pas optimal.

Théorème 7. *Convergence globale dans le cas convexe de la méthode de Newton à pas optimal.*

Supposons que f soit C^2. On se donne $x_0 \in \mathbb{R}^n$ et on suppose que $S_0 = \{x \in \mathbb{R}^n, f(x) \leqslant f(x_0)\}$ est convexe et que pour tout $x \in S_0$, on ait $0 < cI_n \leqslant H_f(x) \leqslant KI_n$.

Alors f est strictement convexe sur S_0, y admet un unique minimiseur x_*, et la suite des itérées $(x_k)_{k\in\mathbb{N}}$ de la méthode de Newton à pas optimal converge vers x_*.

De plus si H_f est M-Lipschitzienne dans un voisinage de x_* (et que la suite n'est pas stationnaire à partir d'un certain rang), alors

— Le pas optimal α_k converge vers 1.

— On a convergence quadratique de la suite (x_k) :

$$x_k \to x_* \quad \text{et} \quad \limsup_{k\to\infty} \frac{\|x_{k+1} - x_*\|}{\|x_k - x_*\|^2} \leqslant \frac{M}{c}.$$

Ce théorème est admis. Attention, ici le pas optimal (ou un critère pour choisir un pas convenable) est nécessaire pour avoir la convergence globale. Ce n'est pas le cas pour la méthode de Newton à pas fixe, comme le montre le simple exemple suivant dans $\mathbb{R}$ (en dimension un, une méthode à pas optimal converge en une seule itération...) :

Exercice 4.2. *Soit f la fonction $x \mapsto \sqrt{1+x^2}$. Si $|x_0| > 1$ montrer que toutes les hypothèses du théorème 7 sont satisfaites, mais que la suite des itérées de la méthode de Newton à pas fixe diverge. Montrer que si l'on se donne $0 < c_1 < c_2 < 1$ avec $c_1 < \frac{1}{2}$ et qu'on sélectionne un pas α_k qui satisfasse la règle de Wolfe, en commençant par tester si le cas $\alpha_k = 1$ convient, alors les itérées convergent vers 0 pour toute condition initiale, et à partir d'un certain rang on a tout le temps $\alpha_k = 1$ (et accessoirement la convergence est cubique, ceci est dû au fait que la dérivée troisième s'annule au point de minimum).*

De même, la condition dite « d'ellipticité » $H_f(x) \geqslant cI_n$ est nécessaire pour obtenir la convergence quadratique :

Exercice 4.3. *Si f est la fonction $(x,y) \mapsto |x|^3 + y^2$, montrer que la fonction est C^2, strictement convexe et que sa hessienne est Lipschitizienne sur $\mathbb{R}^2$, mais pas inversible au point de minimum $(0,0)$. En notant α_k le pas de la méthode de Newton à pas optimal et (x_k, y_k) les itérées, montrer que si on pose $t_k = 1 - \frac{1}{2}\alpha_k$ et $z_k = \frac{2y_k^2}{3|x_k|^3}$, on a alors $t_k|t_k| + (2t_k - 1)z_k = 0$. En déduire une expression de t_k en fonction de z_k puis exprimer z_{k+1} en fonction de t_k et de z_k. Programmer la méthode avec $x_0 = y_0 = 1$ et observer la convergence linéaire (mais pas quadratique).*

Ces deux derniers théorèmes mettent en avant les avantages de la méthode de Newton : convergence globale dans le cas convexe, très bonne vitesse de convergence.

Cependant, dans de nombreux cas, cette méthode n'est pas réalisable en pratique pour trois raisons majeures :

— Le calcul de $H_f(x)$, s'il n'est pas donné directement, peut être assez coûteux si la dimension est un peu grande (il faut évaluer $\partial_i\partial_j f$ pour $1 \leqslant i \leqslant j \leqslant n$, ce qui nécessite de l'ordre de n^2 approximations).

— Le calcul de $H_f(x)^{-1}\nabla f(x)$ est également coûteux (de l'ordre de n^3 opérations pour une méthode directe de décomposition, du type LU – ici on utilisera plutôt une décomposition de Cholesky, la matrice étant symétrique), à moins que $H_f(x)$ n'ait une forme particulière.

— Enfin, on a besoin d'avoir $H_f(x_k)$ définie positive à chaque étape, et ceci est une très forte contrainte : il se pourrait que la fonction n'ait qu'un minimum local, correspondant au minimum global, et aucun autre point critique, mais qu'en certains points la hessienne ne soit pas forcément définie positive.

Les méthodes dites de quasi-Newton permettent souvent de s'affranchir de ces inconvénients, ce sont des méthodes qui généralisent la méthode de la sécante en dimensions supérieures.

On peut aussi citer une dernière propriété intéressante qui sera aussi vérifiée pour les méthodes de quasi-Newton : la méthode de Newton est invariante par changements de coordonnées affines.

Exercice 4.4. *Soit (x_k) la suite des itérées de la méthode de Newton (à pas fixe ou variable) pour la fonction f. On se donne un changement de variable affine $\widetilde{x} = Mx + v$ dans $\mathbb{R}^n$ (où M est une matrice inversible de $M_n(\mathbb{R})$ fixée, et v un vecteur de $\mathbb{R}^n$ fixé) et un changement de variable affine dans $\mathbb{R}$ $\widetilde{y} = a\,y + b$ où $a > 0$ et $b \in \mathbb{R}$. On pose alors la fonction $\widetilde{f}$ correspondant à f avec ce changement de variable : $\widetilde{f}(Mx + v) = a\,f(x) + b$ (ou encore $\widetilde{f}(\widetilde{x}) = a\,f(M^{-1}(\widetilde{x} - v) + b)$.*

Montrer que les itérées $\widetilde{x}_k$ de la méthode de Newton (avec le même pas) pour la fonction $\widetilde{f}$ et partant de $\widetilde{x}_0 = Mx_0 + v$ sont données par x_k tels que $\widetilde{x_k} = Mx_k + v$.

Montrer que ce n'est en général pas le cas si on prend les méthodes de descente de gradient à pas fixe (ou à pas optimal), même si $a = 1$ (sauf si la matrice M est orthogonale).

On peut utiliser cette propriété pour construire des fonctions « invariantes » par changement d'échelle affine qui créent des pathologies : par exemple une fonction C^3, strictement convexe, avec un unique minimum en 0, mais pour laquelle il existe des conditions initiales aussi proche du minimum que l'on veut et pour lesquelles les itérées de la méthode de Newton à pas fixe divergent.

Exercice 4.5. ** On considère la fonction $g : \mathbb{R} \to \mathbb{R}$ définie par $g(x) = x^3(1 + \delta\cos(\ln|x|))$ (et $g(0) = 0$), et f une de ses primitives. Montrer que si $|\delta| < \frac{3}{\sqrt{10}}$, alors f est de classe C^3, strictement convexe, admettant un minimum global strict en 0.*

Montrer qu'il existe deux constantes $m > 1$ et $a > 0$ (indépendantes de δ) telles que pour tout $x \in \mathbb{R}$, on ait $f(m\,x) = a\,f(x)$, autrement dit on a $\widetilde{f} = f$ pour les changements de variables affines correspondant à $\widetilde{x} = m\,x$ et $\widetilde{y} = a\,y$.

Montrer qu'il existe $u \in \mathbb{R}$ tel que $g'(u) = 2u^2(1 - \frac{\sqrt{10}}{3}\delta)$. En déduire qu'il existe $\delta \in]0, \frac{3}{\sqrt{10}}[$ tel que si on effectue la méthode de Newton à pas fixe partant de $x_0 = u$, on ait $x_1 = -m\,u$. On se fixe un tel δ. En déduire que quel que soit $n \in \mathbb{N}$, si on part de $x_0 = \frac{1}{m^n}u$, la suite des itérées de la méthode de Newton à pas fixe diverge.

4.2 Les méthodes de quasi-Newton

Les méthodes de quasi-Newton consistent à appliquer la même formule que pour la méthode de Newton, mais en se donnant une approximation B_k de la hessienne $H_f(x_k)$ plutôt que de la calculer, et en mettant à jour cette approximation à chaque étape. Attention aux notations dans cette sous-partie, historiquement on utilise la notation B_k pour parler d'une approximation de la hessienne $H_f(x_k)$, et on utilisera la notation H_k pour parler d'une approximation de son inverse $H_f(x_k)^{-1}$.

En supposant que x_k et x_{k-1} sont assez proches, on peut écrire, puisque H_f est la différentielle de ∇f, que $\nabla f(x_{k-1}) = \nabla f(x_k) + H_f(x_k)(x_{k-1} - x_k) + o(\|x_{k-1} - x_k\|)$. La condition que l'on va imposer sur B_k est qu'elle satisfasse cette équation (où B_k remplace $H_f(x_k)$) où on néglige le reste en $o(\|x_{k-1} - x_k\|)$:

$$\nabla f(x_k) - \nabla f(x_{k-1}) = B_k(x_k - x_{k-1}). \tag{4.1}$$

Cette condition est appelée condition de la sécante, et on peut observer qu'en dimension 1 cela équivaut à $B_k = \frac{f'(x_k)-f'(x_{k-1})}{x_k-x_{k-1}}$. Cela correspond bien à la méthode de la sécante puisque l'inverse de la hessienne $\frac{1}{f''(x)}$ (pour la méthode de Newton) est bien remplacé par $\frac{x_k-x_{k-1}}{f'(x_k)-f'(x_{k-1})}$ dans la méthode de la sécante, et ce dernier terme est l'inverse de B_k. En dimension supérieure ou égale à deux, c'est plus compliqué, parce que l'on peut avoir beaucoup plus de solutions pour la matrice B_k satisfaisant la condition de la sécante (4.1) : il y a n équations, et n^2 coefficients à mettre dans la matrice B_k (ou même seulement $\frac{1}{2}n(n+1)$ si on veut une matrice B_k symétrique, mais on a toujours le problème que $\frac{1}{2}n(n+1) > n$ dès que $n \geqslant 2$). Pour simplifier les notations, on notera par la suite $s_k = x_{k+1} - x_k$ et $y_k = \nabla f(x_{k+1}) - \nabla f(x_k)$, de sorte que la condition de la sécante s'écrit simplement $y_{k-1} = B_k s_{k-1}$.

Une deuxième condition que l'on impose sur B_k est qu'elle soit symétrique définie positive. Cela implique que $\langle s_{k-1}, B_k s_{k-1}\rangle > 0$ (dès que $x_k \neq x_{k-1}$), autrement dit (si B_k satisfait la condition de la sécante) que $\langle s_{k-1}, y_{k-1}\rangle > 0$. Il se trouve que cette condition est équivalente à l'existence d'une telle matrice B_k, et on peut en fait assurer cette condition avec la règle de Wolfe (en fait, c'est à l'origine pour ce type de méthode que la règle de Wolfe a été en partie conçue) :

Exercice 4.6. *Montrer que si $d_k = \frac{1}{\alpha_k}(x_{k+1} - x_k)$ est une direction de descente au point x_k et que le pas α_k satisfait la deuxième condition de la règle de Wolfe (pour une constante $c_2 < 1$), alors on a que $\langle y_k, s_k\rangle = \langle x_{k+1} - x_k, \nabla f(x_{k+1}) - \nabla f(x_k)\rangle > 0$.*

Montrer alors qu'il existe une matrice symétrique définie positive B_{k+1} telle que la condition de la sécante (4.1) *soit satisfaite (on pourra commencer par la dimension* 2*, dans le cas où $s_k = \binom{1}{0}$, et se ramener à ce cas en utilisant un procédé de Gram-Schmidt).*

Cependant, même avec cette condition, il reste toujours trop de solutions à la condition de la sécante pour calculer B_k. Il existe une solution théorique d'approximation de la hessienne qui vérifie bien la condition de la sécante. Cela provient de la formule de Taylor avec reste intégral à l'ordre 1 pour la fonction ∇f entre x_{k-1} et x_k : si l'on pose

$$\overline{B}_k = \int_0^1 H_f(x_{k-1} + ts_{k-1})\mathrm{d}t, \tag{4.2}$$

alors on obtient que $\overline{B}_k$ est une matrice symétrique qui vérifie la formule de la sécante. On a même que si les matrices $H_f(x_{k-1} + ts_{k-1})$ sont définies positives, alors $\overline{B}_k$ l'est aussi. Mais cette formule pour $\overline{B}_k$ fait intervenir la hessienne, et on cherche donc à obtenir une approximation de $\overline{B}_k$ sans avoir justement à calculer la hessienne. D'autre part cela nécessiterait de faire une approximation numérique de l'intégrale, ce qui serait encore plus coûteux.

L'idée historique (introduite par Davidon en 1959, puis popularisée par Fletcher et Powell en 1963, cela donne ce qu'on appelle aujourd'hui la formule DFP) a d'abord été d'essayer de mettre à jour B_k pour qu'elle satisfasse cette condition de la sécante (4.1) sans trop la modifier par rapport à B_{k-1} : on choisit B_k une matrice définie positive satisfaisant cette condition, et minimisant la distance à B_{k-1} pour une norme matricielle bien choisie. Suivant le choix de la norme, on peut alors obtenir des formules directes pour B_k. Il reste alors à savoir calculer une direction de descente $d_k = -B_k^{-1}\nabla f(x_k)$ (qui est bien une direction de descente si B_k est symétrique définie positive), qui correspond à la formule de la direction de descente de Newton, modifiée en remplaçant la hessienne par B_k.

Ce calcul de l'inverse peut paraître coûteux au premier abord, mais en fait on peut également trouver une formule directe pour mettre à jour directement l'inverse de B_k, qui est habituellement noté H_k (attention encore une fois à la confusion possible avec ces notations qui sont assez standard : ici H_k est censé être une approximation de l'inverse de la hessienne $H_f(x_k)$). Cela donne une formule un peu compliquée que l'on ne donnera pas ici. En effet en 1970, simultanément et indépendamment, quatre chercheurs, Broyden, Fletcher, Goldfarb et Shanno, sont arrivés à une méthode très légèrement différente de celle qui vient d'être présentée, mais qui donne de meilleurs résultats. Leur idée est de choisir de minimiser directement la distance entre H_{k+1} et H_k tout en imposant que H_{k+1} reste symétrique, et satisfasse la condition de la sécante, qui se réécrit donc $s_k = H_{k+1}y_k$, avec les notations précédentes.

Exercice 4.7. *On se donne $\overline{B}$ une matrice symétrique définie positive vérifiant la condition de la sécante suivante : $y_k = \overline{B}s_k$ (on peut en trouver une dès que $\langle y_k, s_k\rangle > 0$ d'après l'exercice 4.6, ou utiliser la formule* (4.2) *en prenant $\overline{B} = \overline{B}_{k+1}$ si la hessienne est symétrique définie positive sur le segment $[x_k, x_{k+1}]$).*

On note $\|\cdot\|_{\overline{B}}$ la norme euclidienne sur $M_n(\mathbb{R})$ provenant du produit scalaire $\langle\cdot,\cdot\rangle_{\overline{B}}$ défini par $\langle G, H\rangle_{\overline{B}} = \mathrm{tr}(\overline{B}G\overline{B}H^T)$, et on s'intéresse au problème de minimisation suivant (contraint aux matrices H symétriques et vérifiant la condition de la sécante) :

$$\inf_{H \text{ sym. } t.q. Hy_k = s_k} \|H - H_k\|_{\overline{B}}.$$

Montrer que si $\overline{B}$ est définie positive, alors $\langle\cdot,\cdot\rangle_{\overline{B}}$ définit bien un produit scalaire (on pourra diagonaliser $\overline{B}$ et construire une matrice symétrique définie positive A telle que $A^2 = \overline{B}$). Montrer que le problème d'optimisation admet bien un unique minimum (penser à la projection orthogonale sur un sous-espace affine). Caractériser ce minimum par des propriétés d'orthogonalité, et en déduire que le minimum H_{k+1} est caractérisé par le fait que

$$\mathrm{tr}(H_{k+1}M) = \mathrm{tr}(H_kM) \text{ pour toute matrice } M \text{ symétrique telle que } Ms_k = 0. \tag{4.3}$$

En conclure que ce minimum H_{k+1} ne dépend pas de $\overline{B}$ (parmi toutes les matrices symétriques définies positives telles que $y_k = \overline{B}s_k$).

On tombe alors sur une autre formule, appelée formule BFGS en référence aux initiales des différents auteurs, qui ne dépend que de H_k, s_k et y_k que l'on donne ici sans plus de détails :

$$H_{k+1} = \left(I_n - \frac{s_k y_k^T}{\langle s_k, y_k\rangle}\right) H_k \left(I_n - \frac{y_k s_k^T}{\langle s_k, y_k\rangle}\right) + \frac{s_k s_k^T}{\langle s_k, y_k\rangle}. \tag{4.4}$$

On peut montrer que cette formule est bien la solution du problème de minimisation, et qu'elle donne bien une matrice symétrique définie positive pour H_{k+1} si H_k est elle-même symétrique définie positive et si $\langle s_k, y_k\rangle > 0$ (on a vu précédemment dans l'exercice 4.6 que ceci est assuré dès que l'on prend un pas α_k satisfaisant la deuxième condition de Wolfe).

Exercice 4.8. *Si on suppose qu'on a $\langle s_k, y_k \rangle > 0$, montrer que la formule BFGS* (4.4) *donne bien une matrice H_{k+1} qui est symétrique, vérifie la condition de la sécante et qui est la solution au problème d'optimisation de l'exercice 4.7. On pourra utiliser le fait que pour deux vecteurs u et v dans $\mathbb{R}^n$, on a $u^T v = \mathrm{tr}(uv^T) = \langle u, v \rangle$, et vérifier que H_{k+1} satisfait la caractérisation* (4.3)*. Si on suppose de plus que H_k est définie positive, montrer que pour $v \in \mathbb{R}^n$, on a $\langle v, H_{k+1} v \rangle = \langle w, H_k w \rangle + \frac{\langle s_k, v \rangle^2}{\langle s_k, y_k \rangle}$ avec $w = v - \frac{\langle s_k, v \rangle}{\langle s_k, y_k \rangle} y_k$, et que la matrice H_{k+1} est symétrique définie positive.*

On peut résumer ainsi la méthode BFGS :

Définition 4.2. *Méthode BFGS.*

On se donne une fonction f à minimiser, un point initial x_0, ainsi qu'une matrice initiale H_0 symétrique définie positive (en général, on prend l'identité, à moins que le problème ait une forme particulière).

Pour $k \geqslant 0$, on pose alors $d_k = -H_k \nabla f(x_k)$. On choisit un pas α_k satisfaisant la règle de Wolfe, en essayant d'abord le pas $\alpha_k = 1$, et on pose $x_{k+1} = x_k + \alpha_k d_k$. On pose $s_k = x_{k+1} - x_k$ et $y_k = \nabla f(x_{k+1}) - \nabla f(x_k)$, et on calcule alors H_{k+1} par la formule BFGS (4.4)*.*

On peut montrer que la méthode BFGS est globalement convergente dans le même cadre que la méthode de Newton (convexité, ellipticité), et qu'on garde l'avantage d'une convergence superlinéaire (comme dans le cas de la méthode de la sécante en dimension un) :

Théorème 8. *Convergence globale dans le cas convexe de la méthode BFGS.*

Supposons que f soit C^2. On se donne $x_0 \in \mathbb{R}^n$ et on suppose que $S_0 = \{x \in \mathbb{R}^n, f(x) \leqslant f(x_0)\}$ est convexe et que pour tout $x \in S_0$, on ait $0 < cI_n \leqslant H_f(x) \leqslant KI_n$.

Alors la suite des itérées $(x_k)_{k \in \mathbb{N}}$ de la méthode BFGS converge vers x_*, l'unique minimiseur de f sur S_0.

De plus si H_f est Lipschitzienne dans un voisinage de x_*, alors la convergence est super-linéaire.

Ce théorème est de nouveau admis. Ce qu'il faut retenir c'est que numériquement, la formule BFGS est celle qui est la plus utilisée de nos jours, de par son efficacité et sa stabilité numérique : les erreurs d'approximation de l'inverse du Hessien H_k s'atténuent au cours des itérations.

Il arrive dans certains cas qu'on ne puisse pas appliquer cette méthode pour des raisons de stockage, lorsque la dimension est trop grande (la matrice H_k ayant n^2 coefficients à stocker). On peut alors revenir aux méthodes de type gradient conjugué non-linéaires vues à la fin du chapitre précédent, mais il existe aussi des méthodes de quasi-Newton à mémoire limitée.