

# Méthodes numériques : optimisation

Amic Frouvelle

19 avril 2015

# Table des matières

<b>1</b>	<b>Optimisation continue en une dimension</b>	<b>3</b>
1.1	Généralités . . . . .	3
1.2	Méthode de la section dorée . . . . .	8
1.2.1	Principe général : réduction du triplet . . . . .	8
1.2.2	Présentation de la méthode . . . . .	9
1.3	Autres méthodes . . . . .	10
1.3.1	Par interpolations quadratiques successives . . . . .	10
1.3.2	*Méthodes utilisant la dérivée* . . . . .	11
<b>2</b>	<b>Méthodes de descente de gradient</b>	<b>13</b>
2.1	Présentation générale des méthodes de descente . . . . .	13
2.2	Descente de gradient à pas fixe . . . . .	14
2.2.1	Étude du cas test . . . . .	14
2.2.2	Convergence de la méthode de gradient à pas fixe . . . . .	15
2.3	Descente de gradient à pas optimal . . . . .	20
2.3.1	Critères de choix de pas pour une recherche de pas approchée . . . . .	21
<b>3</b>	<b>La méthode du gradient conjugué</b>	<b>24</b>
3.1	Définition et interprétation en terme de méthode de descente . . . . .	24
3.2	Présentation standard et convergence . . . . .	25
3.3	Préconditionnement . . . . .	28
3.4	Adaptation au cas non linéaire . . . . .	29
<b>4</b>	<b>Méthodes de Newton et quasi-Newton</b>	<b>30</b>
4.1	La méthode de Newton dans $\mathbb{R}^n$ . . . . .	30
4.2	Les méthodes de quasi-Newton . . . . .	32

# 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, de pouvoir par exemple approximer numériquement son gradient. En effet, on sait que le gradient est un vecteur dont la norme est la plus grande pente et l'orientation est dirigée par cette 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 s'il reste un peu de temps, on s'intéressera à des méthodes numériques pour résoudre des problèmes d'optimisation avec contraintes.

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>.

# 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 :

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

c'est-à-dire que l'on cherche un réel  $x \in I$  tel que pour tout  $y \in I$ , on ait  $f(x) \leq f(y)$ . On s'intéressera également aux solutions locales de (1.1), pour lesquelles on a  $f(x) \leq f(y)$  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  $f(x) \rightarrow +\infty$  lorsque  $|x| \rightarrow \infty$  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  $f''(x) \geq 0$ .
- 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 strictement convexe, alors il existe au plus une solution au problème (1.1), et tout minimum local est en fait l'unique minimum global.

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 évaluer l'efficacité de ces méthodes, on veut mesurer la « vitesse » à laquelle ces suites convergent.

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

Soit  $x_k$  une suite convergente vers une limite  $x_\infty$ .

— On dit que la convergence de  $(x_k)$  est linéaire (ou d'ordre 1) s'il existe  $\alpha \in ]0, 1[$  tel que pour tout  $k$  à partir d'un certain rang, on ait :

$$|x_{k+1} - x_\infty| \leq \alpha |x_k - x_\infty|. \quad (1.2)$$

— Si la convergence est linéaire, on définit le taux de convergence linéaire  $r$  par

$$r = \limsup_{k \rightarrow \infty} \frac{|x_{k+1} - x_\infty|}{|x_k - x_\infty|}.$$

C'est également la borne inférieure des  $\alpha$  qui vérifient la condition précédente.

— Si  $r = 0$ , on dit alors que la convergence est superlinéaire.

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

$$|x_{k+1} - x_\infty| \leq \gamma |x_k - x_\infty|^\beta, \quad (1.3)$$

alors on dit que la convergence de la suite est (au moins) d'ordre  $\beta$ . L'ordre de convergence est défini comme la borne supérieure des  $\beta$  satisfaisant cette condition.

**Remarque 1.1.** *La condition dans la définition de l'ordre de convergence linéaire implique bien que la suite converge dans tous les cas : si la condition (1.2) est vérifiée à partir du rang  $k_0$ , on a pour tout  $k \geq k_0$  :*

$$|x_k - x_\infty| \leq \alpha^{k-k_0} |x_{k_0} - x_\infty| \rightarrow 0 \text{ (car } \alpha \in ]0, 1[),$$

soit une estimation de la forme

$$|x_k - x_\infty| \leq C \alpha^k. \quad (1.4)$$

Il faut faire par contre attention pour la condition pour les ordres plus grands que 1 : si la condition (1.3) est satisfaite à partir d'un certain rang, on n'est pas assuré de la convergence de la suite vers  $x_\infty$ . Cependant, s'il existe  $k_1 \geq k_0$  tel que  $x_{k_1}$  est suffisamment proche de  $x_\infty$ , on a alors convergence, et on peut obtenir une estimation de la forme

$$|x_k - x_\infty| \leq C \alpha^{\beta^k}, \quad (1.5)$$

avec  $\alpha < 1$  (voir exercice ci-dessous).

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

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

et que cela donne bien une estimation de la forme (1.5), avec  $\alpha = \left( \gamma^{\frac{1}{\beta-1}} |x_{k_1} - x_\infty| \right)^{\beta-1} < 1$ .

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  $x_\infty$ ). Dans le cas de la convergence linéaire, le nombre de décimales s'accroît 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  $\beta$  avec  $\beta > 1$ , le nombre de décimales exactes est multiplié par un facteur  $\beta$  à chaque étape.

En pratique (comme on le verra dans la proposition 1.3) il y a des cas où les suites semblent converger « à la même vitesse » que les estimations (1.4)-(1.5) de la remarque 1.1 mais ne rentrent pas dans le cadre de la définition précédente. On étend alors légèrement les définitions.

**Définition 1.2.** *Extension de la définition de l'ordre de convergence d'une suite.*

Soit  $x_k$  une suite, et  $x_\infty$  un réel.

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

$$|x_k - x_\infty| \leq C\alpha^k.$$

— On note  $r$  la borne inférieure des  $\alpha$  qui satisfont la condition ci-dessus, et  $r$  est appelé par extension le taux de convergence linéaire de la suite.

— Si  $r = 0$ , on dit aussi 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 :

$$|x_k - x_\infty| \leq C\alpha^{\beta^k},$$

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

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. Les critères suivants permettent d'obtenir cela.

**Proposition 1.2.** *Critères pratiques de convergence et d'estimation d'ordre.*

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

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

Alors la suite  $(x_k)$  converge vers une limite  $(x_\infty)$  et la convergence est linéaire (au sens de la définition 1.2), le taux de convergence linéaire étant inférieur ou égal à  $\alpha$ .

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

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

On suppose de plus qu'il existe  $k_1 \geq k_0$  tel que  $\gamma|x_{k_1} - x_{k_1-1}|^{\beta-1} < 1$ .

Alors la suite  $(x_k)$  converge vers une limite  $(x_\infty)$  et la convergence est d'ordre supérieur ou égal à  $\beta$  (au sens de la définition 1.2).

*Démonstration.* 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  $|x_{i+1} - x_i| \leq \alpha^{i-k}|x_{k_0+1} - x_{k_0}|$ , 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  $x_\infty$ , c'est-à-dire que  $(x_k)$  converge bien vers  $x_\infty$ . Ensuite on a pour tout  $k \geq k_0$

$$\begin{aligned} |x_\infty - x_k| &= \left| \sum_{i=k}^{\infty} (x_{i+1} - x_i) \right| \leq \sum_{i=k}^{\infty} |x_{i+1} - x_i| \\ &\leq \sum_{i=k}^{\infty} \alpha^{i-k_0} |x_{k_0+1} - x_{k_0}| = \frac{\alpha^{k-k_0}}{1-\alpha} |x_{k_0+1} - x_{k_0}| = C\alpha^k, \end{aligned}$$

avec  $C = \alpha^{-k_0}(1-\alpha)|x_{k_0+1} - x_{k_0}|$ , qui est exactement l'estimation (1.4) de la définition.

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  $(|x_k - x_{k-1}|)_{k \geq k_1}$  est décroissante et que  $\gamma|x_k - x_{k-1}|^{\beta-1} \leq \alpha < 1$  pour tout  $k \geq k_1$ . On obtient donc  $|x_{k+1} - x_k| \leq \alpha|x_k - x_{k-1}|$  et comme  $\alpha < 1$  on applique la première partie pour avoir la convergence de  $(x_k)$  vers un réel  $(x_\infty)$  (la convergence étant au moins linéaire). Ensuite en posant  $y_k = \gamma^{\frac{1}{\beta-1}}|x_k - x_{k-1}|$ , on obtient

$$y_{k+1} \leq \gamma^{\frac{1}{\beta-1}} \cdot \gamma|x_k - x_{k-1}|^\beta = (y_k)^\beta,$$

et par récurrence on en déduit que  $y_k \leq y_{k_1}^{\beta^{k-k_1}} = (\alpha^{\frac{1}{\beta-1}})^{\beta^{k-k_1}}$ . Enfin on écrit

$$\begin{aligned} |x_\infty - x_k| &= \left| \sum_{i=k}^{\infty} (x_{i+1} - x_i) \right| \leq \sum_{i=k}^{\infty} |x_{i+1} - x_i| \\ &\leq \sum_{i=k}^{\infty} \alpha^{i-k} |x_{k+1} - x_k| = \frac{1}{1-\alpha} |x_{k+1} - x_k| = \frac{1}{(1-\alpha)\gamma^{\frac{1}{\beta-1}}} y_k \leq C \left( \alpha^{\frac{\beta-k_1}{\beta-1}} \right)^{\beta^k}, \end{aligned}$$

qui est bien de la forme correspondant à la définition 1.2, puisque  $\alpha^{\frac{\beta-k_1}{\beta-1}} < 1$ . □

Pour comparer deux méthodes, on comparera donc d'abord leur ordre de convergence. 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).

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

**Proposition 1.3.** *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 : on pose  $c_n = \frac{1}{2}(a_n + b_n)$  le point milieu. Par récurrence, on pose 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  $\ell$  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 1.2. En fait on a même que le taux de convergence est égal à  $\frac{1}{2}$  : s'il était strictement plus petit, on aurait, pour  $\alpha < \frac{1}{2}$

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

ce qui est une contradiction.

Si  $\ell$  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  $|a_n - \ell| \leq |a_n - c_n| + |c_n - \ell| \leq C\frac{1}{2^n}$ . Donc  $(a_n)$  converge aussi vers  $\ell$ , la convergence étant linéaire, et le taux de convergence linéaire plus petit que  $\frac{1}{2}$ . Si le taux de convergence était strictement plus

petit, on aurait  $|a_n - \ell| \leq C\alpha^n$  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  $\ell < a_{n+1} \leq \ell$  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  $f'(a_n) < 0$ .

On obtient exactement la même chose pour  $b_n$  de la même manière. En passant à la limite, on obtient  $f'(\ell) \geq 0$  et  $f'(\ell) \leq 0$ , donc on a bien  $f'(\ell) = 0$ .  $\square$

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

**Remarque 1.2.** *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.*

*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 :*

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

*Contrairement à la méthode de dichotomie, on n'est pas assuré de la convergence de la suite des itérés. 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 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  $C^3$  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 à évaluer l'erreur en fonction du nombre d'évaluations plutôt que le contraire.

**Définition 1.3.** *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  $\varepsilon_k$  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  $\varepsilon_k \leq C\alpha^k$ . Le taux de convergence linéaire effectif est la borne inférieure de tels  $\alpha$ .*
- *On dit que l'ordre effectif est supérieur à  $\beta$  quand il existe  $C > 0$ ,  $\alpha \in ]0, 1[$  tels que  $\varepsilon_k \leq C\alpha^{\beta k}$ . L'ordre effectif est la borne supérieure de tels  $\beta$ .*

**Remarque 1.3.** *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  $\alpha$ , alors le taux effectif est  $\alpha^{\frac{1}{p}}$ . Si la convergence de la suite des itérés est d'ordre  $\beta$ , 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 1.4.** *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 lesquels le signe de la fonction est différent) qui assurent d'avoir un zéro entre les deux, il existe des triplets de points qui assurent l'existence d'un minimum local.

**Définition 1.5.** *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  $f(a) \geq f(c) \leq f(b)$ .

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 1.6.** *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  $f(c) \leq f(d)$ , le triplet  $(a, c, d)$  est admissible,
- si  $c < d$  et si  $f(c) \geq f(d)$ , le triplet  $(c, d, b)$  est admissible,
- si  $d < c$  et si  $f(d) \leq f(c)$ , le triplet  $(a, d, c)$  est admissible,
- si  $d < c$  et si  $f(d) \geq f(c)$ , 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 1.4.** *Convergence vers le minimum local.*

Si l'algorithme de la définition fournit une suite de triplets  $(a_n, c_n, b_n)$  tels que  $b_n - a_n \rightarrow 0$ , 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  $\ell$  (et donc  $(c_n)$  aussi par encadrement). D'autre part on a toujours  $f(c_n) \leq f(b_n)$ , avec  $c_n < b_n$  et donc on ne peut pas avoir  $b_n \leq x_*$  puisque  $f$  est strictement décroissante sur  $[a, x_*]$ . On a donc  $b_n > x_*$  et donc à la limite  $\ell \geq x_*$ . De même en utilisant  $f(a_n) \leq f(c_n)$ , on obtient l'inégalité inverse et donc  $\ell = x_*$ .  $\square$

**Remarque 1.4.** 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  $f(c) \leq f(a)$ , puis en prenant  $b \geq c$  et l'augmentant jusqu'à ce que  $f(b) \geq f(c)$ .

## 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 1.5.** Réduction constante de la taille du triplet.

On suppose que l'algorithme de la définition 1.4 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  $\alpha$  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  $c_{n+1} = a_{n+1} + \alpha(b_{n+1} - c_{n+1})$ , 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 exclu 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  $\frac{1}{2}(\sqrt{5} - 1)$ .  $\square$

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 1.7.** 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 déjà connues.

— si  $f(c_n) \geq f(d_n)$ , 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 déjà connues.

**Proposition 1.6.** *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  $\alpha$  vers une limite  $\ell$ . Si la fonction  $f$  est unimodale sur  $[a_0, b_0]$ , alors  $f$  admet son minimum en  $\ell$ .

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

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

ce qui donne la convergence linéaire vers  $\ell$  des quatre suites, avec un taux inférieur ou égal à  $\alpha$ . On montre que ce taux est exactement  $\alpha$  comme dans la preuve de la méthode de dichotomie. Le fait que  $f$  admette son minimum en  $\ell$  si  $f$  est unimodale est une conséquence de la proposition 1.4.  $\square$

**Remarque 1.5.** *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*

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

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  $n \geq 2$ , et on a une fonction  $f$  de  $\mathbb{R}^n$  dans  $\mathbb{R}$ , ou plus généralement de  $\Omega$  dans  $\mathbb{R}$ , où  $\Omega$  est un ouvert. Lorsqu'on utilisera la notation  $\Omega$ , 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  $f(x_*) \leq f(x)$ . 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  $\Omega$ ) 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  $\Omega$  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  $\alpha_k$  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  $\Omega$  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  $h'(0) = \langle \nabla f(x), d \rangle \leq \|\nabla f(x)\| \|d\|$ , avec égalité si et seulement si  $\nabla f$  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  $\gamma$  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.  $\square$

**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

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

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

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

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  $x \in \mathbb{R}^n$ ,  $\|Mx\| \leq \max_{1 \leq i \leq n} |\lambda_i| \|x\|$ .*

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

**Proposition 2.3.** *On pose  $\ell$  (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  $\alpha = \frac{2}{\ell + L}$ , et le taux est alors  $\frac{L - \ell}{L + \ell}$ .*

*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  $\lambda$  valeur propre de  $A$ , et donc on obtient

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

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

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  $\alpha$ ). 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

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

$\square$

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

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

*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  $\Omega$ , ce qui est le cas ici puisque  $x + h = y$ ) :

$$f(x + h) = f(x) + \int_0^1 \langle \nabla f(x + th), h \rangle dt,$$

et on soustrait  $\langle \nabla f(x), h \rangle = \int_0^1 \langle \nabla f(x), h \rangle dt$  des deux côtés pour obtenir

$$f(x + h) - f(x) - \langle \nabla f(x), h \rangle = \int_0^1 \langle \nabla f(x + th) - \nabla f(x), h \rangle dt.$$

À l'aide de l'inégalité de Cauchy-Schwarz et du fait que  $\nabla f$  est  $L$ -Lipschitz sur  $\Omega$ , on obtient

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

qui est exactement l'inégalité souhaitée. □

On remarque ici qu'on n'a en fait eu besoin du fait que  $\nabla f$  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  $f : \Omega \rightarrow \mathbb{R}$  une fonction  $C^1$ , bornée inférieurement. On se fixe un point  $x_0 \in \Omega$  et on suppose les deux conditions suivantes :

- L'ensemble  $S_0 = \{x \in \Omega, f(x) \leq f(x_0)\}$  est fermé dans  $\mathbb{R}^n$ .
- Le gradient  $\nabla f$  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 :

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

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

$$f(x_k - t\nabla f(x_k)) - f(x_k) \leq -t\|\nabla f(x_k)\|^2(1 - \frac{1}{2}tL) \leq 0. \quad (2.3)$$

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  $\Omega$  est ouvert,  $I$  est non vide : pour tout  $t$  suffisamment petit, on a  $x_k - t\nabla f(x_k) \in \Omega$  et  $f(x_k - t\nabla f(x_k)) < f(x_k) \leq f(x_0)$  puisque  $x_k \in S_0$  par hypothèse de récurrence. On peut donc définir sa borne supérieure  $t_* \leq \frac{2}{L}$ .
- Comme  $S_0$  est fermé, si  $t_n \in I$  et  $t_n \rightarrow t_*$ , 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  $\Omega$  est ouvert, pour tout  $\varepsilon$  suffisamment petit on aurait  $x_k - (t_* + \varepsilon) \nabla f(x_k) \in \Omega$  avec  $f(x_k - (t_* + \varepsilon) \nabla f(x_k)) < f(x_k) \leq f(x_0)$  et donc  $x_k - (t_* + \varepsilon) \nabla f(x_k) \in S_0$  pour tout  $\varepsilon$  suffisamment petit. Ce qui donnerait que  $t_* + \varepsilon$  serait dans  $I$  pour tout  $\varepsilon$  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  $k \in \mathbb{N}$

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

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  $\|\nabla f(x_k)\| \rightarrow 0$  lorsque  $k \rightarrow \infty$ .  $\square$

**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  $\varepsilon$  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  $x_\infty$ . Par continuité de  $f$  et de  $\nabla f$ , on obtient que  $\lim_{k \rightarrow \infty} f(x_k) = \lim_{k \rightarrow \infty} f(x_{\varphi(k)}) = f(x_\infty)$ , et  $\|\nabla f(x_\infty)\| = \lim_{k \rightarrow \infty} \|\nabla f(x_{\varphi(k)})\| = \lim_{k \rightarrow \infty} \|\nabla f(x_k)\| = 0$ , ce qui est bien ce qu'on voulait démontrer.  $\square$

Ici, on n'a toujours rien montré sur la convergence des  $x_k$ . On peut l'obtenir si on suppose des conditions de non-dégénérescence (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

$$\nabla f(x + h) = \nabla f(x) + \int_0^1 H_f(x + th)(h) dt.$$

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

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

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

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

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

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

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  $\varepsilon$ .

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  $\|x_{k+1} - x_k\| \rightarrow 0$ . 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  $\|x_{\varphi(k)+1} - x_{\varphi(k)}\| \leq \varepsilon$  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  $\varepsilon \leq \|x_{\psi(k)} - x_*\| \leq 2\varepsilon$ , 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  $\varepsilon$ , donc la suite converge bien vers  $x_*$ .  $\square$

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  $M \geq 0$  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  $\langle h, Mh \rangle \geq 0$ ). Cela équivaut à ce que toutes les valeurs propres de  $M$  soient positives.

De même, on notera  $M \leq LI_n$  (resp.  $LI_n \leq M$ ) 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  $f : \Omega \rightarrow \mathbb{R}$  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  $S_0 = \{x \in \Omega, f(x) \leq f(x_0)\}$  est fermé dans  $\mathbb{R}^n$ .

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

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  $\lim_{k \rightarrow \infty} \|x_k\| = +\infty$ .

*Démonstration.* Les hypothèses sur la Hessienne donnent que  $\nabla f$  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  $\|H_f(x)(h)\| \leq L\|h\|$ . En appliquant la formule de Taylor à l'ordre 1 à  $\nabla f$ , on obtient directement que  $\nabla f$  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  $\|x_k\| \rightarrow \infty$  est qu'il existe  $M > 0$  tel que pour tout  $N_0 \in \mathbb{N}$ , il existe  $N \geq N_0$  tel que  $\|x_k\| \leq M$ , 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  $\|x_{\varphi(k)}\| \leq M$  pour tout  $k$ .

On peut en extraire de nouveau par compacité une suite convergente vers un point  $x_\infty$ , avec donc par continuité  $\nabla f(x_\infty) = 0$  et  $f(x_\infty) \leq f(x_1) < f(x_0)$  puisqu'on a supposé que  $\nabla f(x_0) \neq 0$ , et que donc on a  $f(x_k) \leq f(x_1) < f(x_0)$  pour tout  $k \geq 1$ . On peut donc trouver un  $\varepsilon$  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  $x_\infty$  est une valeur d'adhérence des  $x_k$ ). On écrit alors

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

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  $x_\infty$ . Donc la suite  $(x_k)_{k \in \mathbb{N}}$  converge vers  $x_\infty$ .

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

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

- L'ensemble  $S_0 = \{x \in \Omega, f(x) \leq f(x_0)\}$  est fermé dans  $\mathbb{R}^n$ .
- Pour tout  $x \in S_0$ , la Hessienne  $H_f(x)$  vérifie  $\ell I_n \leq H_f(x) \leq L I_n$  (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  $\|F'(x)\| \leq \max(|1 - \alpha \ell|, |1 - \alpha L|) = r(\alpha)$ . 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  $k \geq 1$ . On applique la formule de Taylor à l'ordre 1 pour  $F$  :

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

Le critère de convergence du précédent chapitre (qui s'adapte sans souci dans  $\mathbb{R}^n$ , on utilise seulement la complétude et le fait qu'une série normalement convergente converge) 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.  $\square$

On a déjà montré dans l'étude du cas test que le meilleur choix de  $\alpha$  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  $\alpha$  produit une suite qui converge vers  $x_*$ , un point de minimum local strict tel que  $\ell I_n \leq H_f(x_*) \leq L I_n$ , avec  $0 < \ell \leq L$  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  $\alpha_k$  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  $\alpha_k$  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  $\alpha_k$  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  $\alpha_k$  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  $\langle \nabla f(x_{k+1}), d_k \rangle = 0$ .  $\square$

On peut donc donner une interprétation 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 le taux de convergence soit bien linéaire), comme on peut l'observer sur des fonctions du type  $x \mapsto 1 - \frac{1}{\|x\|^2}$ , qui ne décroissent pas suffisamment lorsque le point initial est loin du minimum.

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  $\ell$  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*

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

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  $h(t)$  décroît suffisamment entre 0 et  $\alpha_k$ , ou que la pente a suffisamment réduit...

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

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

Cela revient encore à dire que  $h(\alpha_k) \leq h(0) + c_1 \alpha_k h'(0)$ . C'est toujours faisable en prenant  $\alpha_k$  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.

**Définition 2.7.** *Règle de Goldstein.* On dit que  $\alpha_k$  satisfait la règle de Goldstein 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

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

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

Cette règle permet de garder un pas « assez grand ». En pratique on choisit souvent  $c_1 < \frac{1}{2}$  et  $c_2 > \frac{1}{2}$ , ce qui permet de ne pas écarter le pas optimal dans le cas où la fonction  $h$  est bien approximée par un polynôme de degré 2.

**Exercice 2.2.** Montrer que si la fonction  $f$  est quadratique et que  $c_1 < \frac{1}{2} < c_2$ , alors le pas optimal satisfait la règle de Goldstein.

Une autre condition correspondant au fait de garder un pas « assez grand » consiste à demander que le gradient (si on y a accès) ait suffisamment diminué.

**Définition 2.8.** Règle de Wolfe. On dit que  $\alpha_k$  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

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

Cela revient à dire que  $h'(\alpha_k) \geq c_2 h'(0)$ . Par exemple, si  $\alpha_k$  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 ou de Goldstein.** On peut donner un algorithme simple de dichotomie pour trouver un pas  $\alpha_k$  qui satisfasse la règle de Wolfe ou de Goldstein, 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 il satisfait la deuxième règle on a obtenu ce qu'on voulait et on s'arrête, 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. 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, mais on est sûr qu'au bout d'un moment la règle d'Armijo ne sera pas satisfaite, sinon  $f$  ne serait pas bornée inférieurement.
- Ensuite, par dichotomie, on pose  $\alpha = \frac{1}{2}(\alpha_{\max} + \alpha_{\min})$ . Si  $\alpha$  satisfait les deux critères on peut s'arrêter, sinon on met à jour  $\alpha_{\max}$  (resp.  $\alpha_{\min}$ ) à la valeur  $\alpha$  si  $\alpha$  ne satisfait pas la règle d'Armijo (resp. satisfait la règle d'Armijo mais pas le deuxième critère).

Avec cette méthode, on est sûr que l'algorithme va s'arrêter. En effet sinon on aurait l'existence de  $\alpha_{\max}$  et  $\alpha_{\min}$  aussi proches qu'on veut d'une valeur limite  $\alpha_*$  et satisfaisant :

- Dans le cas de la règle de Goldstein,  $h(\alpha_{\max}) > h(0) + c_1 \alpha_{\max} h'(0)$  et  $h(\alpha_{\min}) < h(0) + c_2 \alpha_{\min} h'(0)$ , soit à la limite  $c_1 \alpha_* h'(0) \leq c_2 \alpha_* h'(0)$ , en contradiction avec  $h'(0) > 0$  et  $c_1 < c_2$ .
- Dans le cas de la règle de Wolfe,  $h(\alpha_{\max}) > h(0) + c_1 \alpha_{\max} h'(0)$  et  $h(\alpha_{\min}) \leq h(0) + c_1 \alpha_{\min} h'(0)$ , soit  $h(\alpha_{\max}) - h(\alpha_{\min}) > c_1 h'(0)(\alpha_{\max} - \alpha_{\min})$ , ce qui nous donnerait à la limite  $h'(\alpha_*) \geq c_1 h'(0)$ . Mais on a également  $h'(\alpha_{\min}) < c_2 h'(0)$  et donc à la limite  $c_1 h'(0) \leq c_2 h'(0)$ , qui donne la même contradiction.

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  $\nabla f$  est  $L$ -Lipschitz, que le pas  $\alpha_k$  satisfait la règle de Wolfe, et que  $f$  est bornée inférieurement. Si on note  $\theta_k$  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 2.3.** *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  $\beta_k$  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  $\alpha_k$  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  $\alpha_k$  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  $\alpha_k$  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  $\alpha_k$  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  $\alpha_{k+1}$  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  $\alpha_{k+1}$  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  $\alpha_0$  est bien le pas optimal pour la direction de descente  $p_0$ , cela permet donc bien d'initialiser la récurrence.  $\square$

**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  $\alpha_k$  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  $\alpha_k$ . 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  $k \geq 1$ , 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  $\langle r_{k+1}, r_k \rangle = \langle r_{k+1}, -p_k + \beta_k p_{k-1} \rangle = 0$ .

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  $\langle r_{k+1}, Ap_k \rangle = \frac{1}{\alpha_k} \|r_{k+1}\|^2$ . Puis on écrit le dénominateur  $\langle p_k, Ap_k \rangle = \frac{1}{\alpha_k} \langle p_k, r_{k+1} - r_k \rangle = -\frac{1}{\alpha_k} \langle p_k, r_k \rangle = \frac{1}{\alpha_k} \|r_k\|^2$  (on avait déjà montré que  $\langle r_{k+1}, p_k \rangle = 0$  et que  $\langle p_k, r_k \rangle = -\|r_k\|^2$ ), ce qui nous donne le résultat.  $\square$

Au cours des deux dernières preuves, 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  $\alpha_k$  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 conséquent, 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 :*

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

De plus, pour tout  $i$  tel que  $0 \leq i \leq k$ , 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.* On montre l'égalité des sous-espaces par récurrence sur  $k$ . Le cas  $k = 0$  est évident vu que  $p_0 = -r_0$ . Supposons donc  $\text{Vect}(r_0, Ar_0, \dots, A^{k-1} r_0) = \text{Vect}(r_0, r_1, \dots, r_k) = \text{Vect}(p_0, p_1, \dots, p_k)$ . Pour montrer l'égalité entre les deux derniers sous-espaces au rang  $k + 1$ , il suffit donc de montrer que  $p_{k+1} \in \text{Vect}(r_0, r_1, \dots, r_k, r_{k+1})$  et que  $r_{k+1} \in \text{Vect}(p_0, p_1, \dots, p_k)$ . 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  $p_{k+1} \in \text{Vect}(r_0, r_1, \dots, r_k, r_{k+1})$ . En écrivant alors  $r_{k+1} = -p_{k+1} + \beta_k r_k$  et en utilisant le fait que  $r_k \in \text{Vect}(p_0, p_1, \dots, p_k)$  on obtient bien que  $r_{k+1} \in \text{Vect}(p_0, p_1, \dots, p_k)$ .

Ensuite pour montrer l'égalité  $\text{Vect}(r_0, Ar_0, \dots, A^k r_0) = \text{Vect}(r_0, r_1, \dots, r_k, r_{k+1})$ , il suffit également de montrer que  $A^{k+1} r_0 \in \text{Vect}(r_0, r_1, \dots, r_k, r_{k+1})$  et que  $r_{k+1} \in \text{Vect}(r_0, Ar_0, \dots, A^k r_0)$ . 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  $\text{Vect}(r_0, Ar_0, \dots, A^k r_0)$  par hypothèse de récurrence, on obtient que  $Ap_k \in \text{Vect}(Ar_0, A^2 r_0, \dots, A^k r_0, A^{k+1} r_0)$  puis que  $r_{k+1} \in \text{Vect}(r_0, Ar_0, A^2 r_0, \dots, A^k r_0, A^{k+1} r_0)$ . Pour l'inclusion réciproque, on écrit que  $A^k r_0 = \sum_{i=0}^k \lambda_i p_i$  par hypothèse de récurrence, et donc (on a vu précédemment que  $\alpha_k > 0$  tant que  $r_k \neq 0$ )

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

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  $i \leq k - 1$ , 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  $p_i \in \text{Vect}(r_0, Ar_0, \dots, A^i r_0)$ , on obtient que  $Ap_i \in \text{Vect}(Ar_0, A^2 r_0, \dots, A^{i+1} r_0) \subset \text{Vect}(p_0, p_1, \dots, p_{i+1})$ . Comme on vient de voir que  $\langle r_{k+1}, p_j \rangle = 0$  pour  $j \leq k$ , 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  $i \leq k$ ,  $r_i \in \text{Vect}(p_0, p_1, \dots, p_i)$ , 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  $0 \leq j \leq k$ .

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  $(r_i)_{0 \leq i \leq k}$  est donc une famille orthogonale et donc ne peut avoir qu'au plus  $n$  vecteurs non-nuls, en particulier on a donc  $r_n = 0$ .  $\square$

**Exercice 3.1.** Montrer que  $x_k$  est le minimiseur de la fonction  $f$  sur l'espace affine  $x_0 + \text{Vect}(p_0, p_1, \dots, p_k)$ .

**Remarque 3.2.** On remarque que le nombre d'itération peut être plus petit que  $n$  si l'espace  $\text{Vect}(r_0, Ar_0, \dots, A^{n-1}r_0)$  (appelé sous-espace de Krylov) est de dimension plus petite que  $n$ . Par exemple si  $A$  a seulement  $p$  valeurs propres distinctes alors il existe un polynôme de degré  $p$  annihilant  $A$ , et la dimension du sous-espace est donc inférieure ou égale à  $p$ . De même si  $r_0$  se décompose sur un nombre  $p$  de valeurs propres distinctes, alors on peut montrer que le sous-espace est également de dimension inférieure ou égale à  $p$ .

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  $\ell$  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

$$\|x_k - x_*\|_A \leq 2 \left( \frac{\sqrt{L} - \sqrt{\ell}}{\sqrt{L} + \sqrt{\ell}} \right)^k \|x_0 - x_*\|_A,$$

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 preconditionnement 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 connaître 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  $r_k \neq 0$  :*

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

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 preconditionneur  $M$  est extrêmement vaste, et dépend souvent fortement du problème considéré. Il permet d'obtenir des convergences bien plus rapide. En général, on utilise un preconditionnement 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 telle sorte qu'elles soient conjuguées.

Les différences principales se trouvent dans trois ingrédients :

- On ne peut pas calculer le pas optimal  $\alpha_k$  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  $\beta_k$  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  $\beta_k$  : 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 conjugaison 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  $\alpha_k$  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 produire la méthode de Polak-Ribière dans des cas très particuliers. 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  $\alpha_k$ , 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  $\varepsilon$  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  $\beta_k$  est calculé par la formule suivante*

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

# 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 1.2 dans le cas où la dimension est supérieure ou égale à 2.

### 4.1 La méthode de Newton 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 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 1.2.

En pratique, étant donné que le développement de  $f$  au voisinage de  $x$  n'est pas exact, on peut alors 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  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ , 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  $\alpha_k = 1$ . 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 vue comme une méthode de recherche de zéros d'une fonction  $g : \mathbb{R}^n \rightarrow \mathbb{R}^n$ , en écrivant  $x_{k+1} = x_k - Dg(x_k)^{-1}g(x_k)$ . Cela correspond à écrire un développement de Taylor à l'ordre un pour  $g$  et à résoudre l'équation linéaire  $g(x) + Dg(x)h = 0$  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  $\langle \nabla f(x_k), d_k \rangle = -\langle \nabla f(x_k), H_f(x_k)^{-1} \nabla f(x_k) \rangle$  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 que la convergence a lieu au voisinage des points critiques non-dégénérés, et que dans ce cas-là elle est quadratique.

**Théorème 6.** *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

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

Ce théorème est admis, en fait, on l'obtient également pour la méthode de Newton générale de recherche de zéros d'une fonction  $g : \mathbb{R}^n \rightarrow \mathbb{R}^n$  indiquée dans la remarque précédente. 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.

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  $S_0 = \{x \in \mathbb{R}^n, f(x) \leq f(x_0)\}$  est convexe et que pour tout  $x \in S_0$ , on ait  $0 < cI_n \leq H_f(x) \leq KI_n$ .

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  $x_*$ , alors

- Le pas optimal  $\alpha_k$  converge vers 1.
- On a convergence quadratique de la suite  $(x_k)$  : à partir d'un certain rang, on a

$$\frac{\|x_{k+1} - x_*\|}{\|x_k - x_*\|} \leq \frac{M}{c}.$$

De même, ce théorème est admis. Attention, ici le pas optimal est nécessaire pour 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 4.1.** 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 diverge.

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



**Exercice 4.2.** Si  $f$  est la fonction  $(x, y) \mapsto x^4 + 6y^2$ , montrer que la fonction est strictement convexe, mais que la hessienne n'est pas inversible au point de minimum  $(0, 0)$ .

Programmer la méthode de Newton à pas optimal en l'initialisant en  $(1, 1)$  et observer que la convergence est linéaire et non 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  $1 \leq i \leq j \leq n$ , 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.

## 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. En supposant que  $x_k$  et  $x_{k-1}$  sont assez proches, on peut écrire  $\nabla f(x_{k-1}) = \nabla f(x_k) + H_f(x_k)(x_{k-1} - x_k) + o(\|x_{k-1} - x_k\|)$ , puisque  $H_f$  est la différentielle de  $\nabla f$ . 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  $o(\|x_{k-1} - x_k\|)$  :

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

Cette condition est appelée condition de la sécante, et on peut observer qu'en dimension 1 cela équivaut à  $B_k = \frac{f'(x_{k-1}) - f'(x_k)}{x_{k-1} - x_k}$ . Cela correspond bien à la méthode de la sécante puisque l'inverse de la hessienne  $\frac{1}{f''(x)}$  est bien remplacé par  $\frac{x_{k-1} - x_k}{f'(x_{k-1}) - f'(x_k)}$  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).

Une deuxième condition que l'on impose sur  $B_k$  est qu'elle soit définie positive. Cela implique que  $\langle x_{k-1} - x_k, B_k(x_{k-1} - x_k) \rangle > 0$ , autrement dit que  $\langle x_k - x_{k-1}, \nabla f(x_{k-1}) - \nabla f(x_k) \rangle > 0$ , et il se trouve que l'on peut 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 conçue). Cependant, même avec cette condition, il reste toujours trop de solutions à la condition de la sécante pour calculer  $B_k$ .

L'idée historique 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 modifier  $B_k$  : on choisit  $B_{k+1}$  une matrice définie positive satisfaisant cette condition, et minimisant la distance à  $B_k$  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  $-B_k^{-1}\nabla f(x_k)$ , 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 des formules directes pour mettre à jour directement l'inverse de  $B_k$ , qui est habituellement noté  $H_k$  (attention 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  $H_f(x_k)$ ). En 1970, simultanément et indépendamment, quatre chercheurs, Broyden, Fletcher, Goldfarb et Shanno, sont arrivés à une méthode légèrement différente de celle qui vient d'être présentée, en choisissant directement de minimiser la distance entre  $H_{k+1}$  et  $H_k$  tout en imposant que  $H_{k+1}$  reste symétrique définie positive, et satisfasse la condition de la sécante, qui se réécrit  $x_{k+1} - x_k = H_{k+1}(\nabla f(x_{k+1}) - \nabla f(x_k))$ , et que l'on abrège en général  $s_k = H_{k+1}y_k$ , avec  $s_k = x_{k+1} - x_k$  et  $y_k = \nabla f(x_{k+1}) - \nabla f(x_k)$ . 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 :

$$H_{k+1} = \left( I_n - \frac{s_k y_k^T}{\langle s_k, y_k \rangle} \right) H_k \left( I_n - \frac{y_k s_k^T}{\langle s_k, y_k \rangle} \right) + \frac{s_k s_k^T}{\langle s_k, y_k \rangle}. \quad (4.2)$$

On peut montrer que cette formule 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 peut résumer ainsi la méthode BFGS :

**Définition 4.2.** *Méthode BFGS.*

*On se une fonction  $f$  à minimiser, un point initial  $x_0$ , ainsi qu'une matrice initiale  $H_0$  symétrique définie positive.*

*Pour  $k \geq 0$ , on pose alors  $d_k = -H_k \nabla f(x_k)$ . On choisit un pas  $\alpha_k$  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 (4.2).*

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  $S_0 = \{x \in \mathbb{R}^n, f(x) \leq f(x_0)\}$  est convexe et que pour tout  $x \in S_0$ , on ait  $0 < cI_n \leq H_f(x) \leq KI_n$ .*

*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.