Introduction à l'optimisation non linéaire

Esquisse de cours

Sommaire

1. Introduction

Considérations générales

De nombreux problèmes nécessitent de minimiser (ou de maximiser) une fonction :

- minimiser la distance entre des points et une courbe donnés
- trouver l'état d'équilibre d'un système mécanique (minimiser E_{pot})
- trouver l'état d'équilibre d'un gaz, d'un mélange (maximiser l'entropie)
- minimiser le coût, ...

Points à souligner : formulation mathématique, optimisation continue ou discrète, optimisation **avec ou sans contraintes**, globale ou locale — le cas convexe —, déterministe ou stochastique.

Quelques exemples :

Problème no 1. On considère les points $(1,1)$, $(2,0)$, $(4,2)$ et $(2,3)$ dans $\mathbb{R}^2$. Il existe un cercle maximal inscrit dans le polygone déterminé par les quatre points. Peut-on trouver le centre et le rayon d'un tel cercle ?

Problème no 2 (Sylvester, 1857). On considère n points (x_1, y_1), ..., (x_n, y_n) dans $\mathbb{R}^2$. Quel est le centre et rayon du plus petit cercle contenant les points ?

Problème no 3. Trouver le minimum de $(x-2)^2 + (y-1)^2$ avec les contraintes $x^2 - y \leq 0$ et $x + y \leq 2$.

Une application pratique du deuxième problème serait le placement d'une centrale d'alarme pour un certain nombre de maisons, tel que la distance maximale entre la centrale et les maisons soit minimale.

En générale un problème d'optimisation dans $\mathbb{R}^n$ a la forme suivante : étant donné $S \subset \mathbb{R}^n$ et $f : S \to \mathbb{R}$, déterminer f^* et éventuellement $x^* \in S$ tel que

$$f^* := \inf\{f(x) \mid x \in S\} = f(x^*).$$

Les algorithmes d'optimisation sont itératifs ; caractéristiques souhaitées :

- robustesse (large spectre de problèmes)
- efficacité (temps de calcul et mémoire utilisée)
- précision (sensibilité réduite au erreurs des données).

Ces caractéristiques peuvent être contradictoires, par exemple sur les problèmes en très grandes dimensions.

Rappels et notations

La dérivée partielle de $f : U \subset \mathbb{R}^n \to \mathbb{R}$ en a par rapport à x_j est la limite

$$\frac{\partial f}{\partial x_j} = \lim_{\delta \to 0} \frac{f(a + \delta e_j) - f(a)}{\delta}$$

où $e_j = (0, \ldots, 0, 1, 0, \ldots, 0)$.

Proposition 1.1. *Soit $f : U \to \mathbb{R}^m$ et soit $a \in U$. Si f est différentiable en a alors les dérivées partielles $\dfrac{\partial f_\alpha}{\partial x_j}(a)$ existent et la matrice J dans la définition de la différentiabilité de f en a est*

$$J = \left(\frac{\partial f_\alpha}{\partial x_j}(a)\right)_{1 \leq j \leq n, 1 \leq \alpha \leq m}.$$

(La matrice apparaissant dans la formule s'appelle la matrice jacobienne de f en a, notée $J(f, a)$ ou $f'(a)$.

Démonstration. Dans la formule $f(a+h) - f(a) = Ch + \|h\|\,\varepsilon(h)$ on prend $h = \delta e_j$. Alors

$$\begin{pmatrix} \frac{\partial f_1}{\partial x_j} \\ \vdots \\ \frac{\partial f_m}{\partial x_j} \end{pmatrix} = \lim_{\delta \to 0} \frac{f(a+\delta e_j) - f(a)}{\delta} = \lim_{\delta \to 0} (Ce_j + \varepsilon(\delta e_j)) = \begin{pmatrix} c_{1j} \\ \vdots \\ c_{mj} \end{pmatrix}.$$

□

La réciproque est vraie si on demande aux dérivées partielles d'être continues. De plus, si les dérivées d'ordres deux existent et sont continues, elles sont symétriques (lemme de Schwarz). Si les dérivées partielles existent et sont continues, alors la fonction f est dite de classe $\mathcal{C}^1$. En particulier la fonction $x \mapsto f'(x)$ est une fonction continue.

Pour $f : U \to \mathbb{R}$ fonction de classe $\mathcal{C}^1$ ou $\mathcal{C}^2$, on note

$$\bigtriangledown f(x) = \begin{pmatrix} \frac{\partial f}{\partial x_1}(x) \\ \vdots \\ \frac{\partial f}{\partial x_n}(x) \end{pmatrix} \in \mathbb{R}^n \qquad \text{le gradient de } f \text{ en } x$$

et

$$\bigtriangledown^2 f(x) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \vdots & & & \\ \frac{\partial f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{pmatrix} \qquad \text{la hessienne de } f \text{ en } x.$$

On a la formule de Taylor

$$f(x+u) = f(x) + \bigtriangledown f(x)^T u + \frac{1}{2}\, u^T \bigtriangledown^2 f(x+tu)\, u$$

où t dépend de x et u. Les théorèmes ci-dessous s'ensuivent :

Théorème 1.2 (conditions nécessaires)**.** *Si $x^* \in U$ est un minimum local de f, alors $\bigtriangledown f(x^*) = 0$ et $\bigtriangledown^2 f(x^*)$ est semi-définie positive.*

Théorème 1.3 (conditions suffisantes)**.** *Si $\bigtriangledown f(x^*) = 0$ et $\bigtriangledown^2 f(x^*)$ est définie positive, alors $x^* \in U$ est un minimum local strict de f.*

2. Fonctions convexes

Introduction

Soit $C \subset \mathbb{R}^n$ un sous-ensemble convexe. Une fonction $f : C \to \mathbb{R}$ est dite *fonction convexe* si pour tous $x, y \in C$ et pour tout $\lambda \in [0, 1]$,

$$f((1-\lambda)x + \lambda y) \leq (1-\lambda)f(x) + \lambda f(y).$$

La fonction f est dite *strictement convexe* si l'inégalité ci-dessus est stricte dès que $x \neq y$ et $\lambda \neq 0, 1$. À toute fonction $f : C \to \mathbb{R}$ on associe l'*épigraphe* de f défini par $\mathrm{epi}(f) = \{(x, y) \mid f(x) \leq y\}$.

Lemme 2.1. *La fonction $f : C \to \mathbb{R}$ est convexe si et seulement si $\mathrm{epi}(f) \subset \mathbb{R}^{n+1}$ est un sous-ensemble convexe.*

Démonstration. Si $\mathrm{epi}(f)$ est convexe, alors pour tous $x, y \in C$ et tout $\lambda \in [0, 1]$,

$$(1-\lambda)(x, f(x)) + \lambda(y, f(y)) = ((1-\lambda)x + \lambda y, (1-\lambda)f(x) + \lambda f(y)) \in \mathrm{epi}(f),$$

donc $(1-\lambda)f(x) + \lambda f(y) \geq f((1-\lambda)x + \lambda y)$, c'est-à-dire f est une fonction convexe. Réciproquement, on suppose f convexe et on considère (x_1, y_1) et (x_2, y_2) deux points dans $\mathrm{epi}(f)$. Alors $f(x_j) \leq y_j$ et

$$f((1-\lambda)x_1 + \lambda x_2) \leq (1-\lambda)f(x_1) + \lambda f(x_2) \leq (1-\lambda)y_1 + \lambda y_2,$$

donc $(1-\lambda)(x_1, y_1) + \lambda(x_2, y_2) \in \mathrm{epi}(f)$. □

Corollaire 2.2. *Si $\lambda \geq 1$, alors*

$$f((1-\lambda)x + \lambda y) \geq (1-\lambda)f(x) + \lambda f(y)$$

pour tous $x, y \in C$ tels que $(1-\lambda)x + \lambda y \in C$.

Proposition 2.3 (L'inégalité de Jensen)**.** *Soit $f : C \to \mathbb{R}$ une fonction convexe et soient $\lambda_j \geq 0$ tels que $\lambda_1 + \cdots + \lambda_r = 1$. Pour tous $x_1, \ldots, x_r \in C$ on a*

$$f(\lambda_1 x_1 + \cdots + \lambda_r x_r) \leq \lambda_1 f(x_1) + \cdots + \lambda_r f(x_r).$$

Si f est strictement convexe et $\lambda_j > 0$, alors

$$f(\lambda_1 x_1 + \cdots + \lambda_r x_r) = \lambda_1 f(x_1) + \cdots + \lambda_r f(x_r)$$

entraine $x_1 = \cdots = x_r$.

2.1. Fonctions convexes d'une variable

L'ensemble $C \subset \mathbb{R}$ étant convexe, est un intervalle.

Théorème 2.4. *Une fonction convexe $f : [a, b] \to \mathbb{R}$ est continue sur (a, b).*

Démonstration. On démontre que f est continue en un point $x_0 \in (a,b)$. On peut supposer que x_0 est l'origine et que la valeur en x_0 est aussi 0 – en remplaçant la fonction convexe f par $f - f(x_0)$.

D'abord on affirme qu'on peut supposer qu'il existe un $h > 0$ tel que $f(x) \geq 0$ pour tout $x \in [x_0, x_0 + h]$. Pour justifier l'affirmation, on fait deux remarques : 1) si $f(x_0 \pm h) < 0$ alors $f(x_0 \mp h) > 0$. Plus précisément,

$$f(x_0 \pm h) + f(x_0 \mp h) \geq 2f(x_0) = 0. \tag{2.1}$$

2) si $f(x_0 + h) < 0$, alors pour tout $x \in]x_0, x_0 + h]$, on a $f(x) < 0$. L'affirmation est vérifiée ou bien par f ou bien par $x \mapsto f(x_0 - (x - x_0))$

Par conséquent, on prend $h > 0$ tel que $f(x) \geq 0$ pour tout $x \in [x_0, x_0 + h]$. Alors, pour $t \in [0,1]$,

$$x_t = x_0 + th = (1-t)x_0 + t(x_0 + h)$$

et, par convexité,

$$0 \leq f(x_t) \leq (1-t)f(x_0) + tf(x_0 + h) = tf(x_0 + h).$$

Donc $f(x) \to 0$ quand $x \to x_0$ avec $x > x_0$. Pour la continuité à gauche, on applique (2.1) et la continuité à droite. □

Lemme 2.5. *Soit $f : [a,b] \to \mathbb{R}$ une fonction convexe. Alors, si $a < x < b$,*

$$\frac{f(x) - f(a)}{x - a} \leq \frac{f(b) - f(a)}{b - a} \leq \frac{f(b) - f(x)}{b - x}$$

2.2. Fonctions différentiables convexes — une variable

Une fonction convexe peut ne pas être différentiable ; par exemple la fonction valeur absolue.

Théorème 2.6. *Une fonction différentiable, $f : (a,b) \to \mathbb{R}$, est convexe si et seulement si f' est croissante. Si f' et strictement croissante, alors f est strictement convexe.*

Démonstration. Si f est convexe, alors en utilisant le lemme 2.5, si $a < x < y < b$ on a

$$\frac{f(x) - f(x-h)}{h} \leq \frac{f(y) - f(x)}{y - x} \leq \frac{f(y+h) - f(y)}{h}$$

et en faisant $h \to 0^+$ on obtient

$$f'(x) \leq \frac{f(y) - f(x)}{y - x} \leq f'(y).$$

Réciproquement, on suppose que f' est croissante. On considère $a < x < y < b$ et $0 < \lambda < 1$. Alors, en appliquant le théorème des accroissements finis, on a

$$f((1-\lambda)x + \lambda y) - f(x) = \lambda f'(\xi_0)(y - x)$$

et

$$f(y) - f((1-\lambda)x + \lambda y) = (1-\lambda)f'(\xi_1)(y - x).$$

Donc, en posant $x_\lambda = (1-\lambda)x + \lambda y$,

$$\begin{aligned} f(x_\lambda) - [(1-\lambda)f(x) + \lambda f(y)] &= (1-\lambda)(f(x_\lambda) - f(x)) + \lambda(f(x_\lambda) - f(y)) \\ &= \lambda(1-\lambda)(y-x)(f'(\xi_0) - f'(\xi_1)) \\ &\leq 0, \end{aligned}$$

c'est-à-dire f est convexe. □

Corollaire 2.7. *Si f est deux fois différentiable, alors f est convexe si et seulement si $f''(x) \geq 0$ pour tout x. De plus, f est strictement convexe si $f''(x) > 0$ pour tout x.*

Démonstration. Si f est convexe, alors d'après le théorème 2.6 f' est croissante. Donc f'' est positive ou nulle. Réciproquement, on applique l'autre implication du théorème. □

Théorème 2.8. *Soit f différentiable sur (a,b). Alors f est convexe si et seulement si*

$$f(y) \geq f(x) + f'(x)(y-x)$$

pour tout $x, y \in (a,b)$.

En appliquant ce théorème on remarque que si $f'(x_0) = 0$, alors x_0 est un minimum global de f. On constate que les fonctions convexe sont très particulières.

Démonstration. Si f est convexe on applique le théorème 2.6 et le théorème des accroissements finis pour obtenir l'inégalité. Réciproquement, si on a l'inégalité – le graphe de f est au-dessus de la tangente en tout point –, on suppose $x < y$ et on a

$$f'(x) \leq \frac{f(y) - f(x)}{y - x} = \frac{f(x) - f(y)}{x - y} \leq f'(y),$$

en appliquant l'inégalité en x et puis en y. □

2.3. Fonctions différentiables convexes — plusieurs variables

Théorème 2.9. *Soit $U \subset \mathbb{R}^n$ un ouvert convexe et soit $f : U \to \mathbb{R}$ une fonction convexe de classe $\mathcal{C}^2$. La fonction f est convexe si et seulement si $\bigtriangledown^2 f(x)$ est positive pour tout $x \in U$. De plus, si $\bigtriangledown^2 f(x)$ est définie positive, alors f est strictement convexe.*

Démonstration. Si f est convexe on utilise la restriction de f à des droites passant par le point qu'on étudie. Réciproquement, on considère deux points $a, b \in U$ et on applique de nouveau le corollaire 2.7. □

Exemple 2.10 (Méthode des moindres carrés de Gauss)**.** Soient (x_1, y_1), $\ldots$,(x_n, y_n) des points dans le plan et soit $y = \alpha x + \beta$ l'équation d'une droite quelconque. On cherche à minimiser la fonction

$$f(\alpha, \beta) = \sum_{j=1}^{n} (y_j - \alpha x_j - \beta)^2.$$

(On cherche la droite qui approche le mieux les n points du point de vue des carrées des distances.) La fonction f est positive et polynomiale. On a

$$\frac{\partial f}{\partial \alpha} = -2\sum_{j=1}^{n} x_j(y_j - \alpha x_j - \beta) \quad \text{et} \quad \frac{\partial f}{\partial \beta} = -2\sum_{j=1}^{n}(y_j - \alpha x_j - \beta)$$

et donc

$$\bigtriangledown^2 f(\alpha, \beta) = 2\begin{pmatrix} \sum_{j=1}^{n} x_j^2 & \sum_{j=1}^{n} x_j \\ \sum_{j=1}^{n} x_j & n \end{pmatrix}.$$

Mais, pour une matrice A de taille 2×2, la condition de positivité est équivalente à $a_{11} \geq 0$ et $\det(A) \geq 0$. Mais

$$\det(\bigtriangledown^2 f(\alpha, \beta)) = n\sum_{j=1}^{n} x_j^2 - \left(\sum_{j=1}^{n} x_j\right)^2 \geq 0$$

avec égalité si et seulement si $x_1 = \cdots = x_n$ par l'inégalité de Cauchy-Bouniakovski-Schwarz. Donc en supposant que les points ne sont pas alignés, f est strictement convexe d'après le corollaire 2.7. Elle admet un minimum global et le point (α^*, β^*) est obtenu en résolvant $\bigtriangledown f = 0$. On obtient

$$\alpha^* = \frac{n\sum_j x_j y_j - (\sum_j x_j)(\sum_j y_j)}{n\sum_j x_j^2 - (\sum_j x_j)^2}$$

$$\beta^* = \frac{(\sum_j x_j^2)(\sum_j y_j) - (\sum_j x_j)(\sum_j x_j y_j)}{n\sum_j x_j^2 - (\sum_j x_j)^2.}$$

3. Optimisation sans contraintes en dimension 1 — le cas différentiable

La méthode classique de minimisation est la *descente de gradient*: on commence avec un point x_0, par exemple donné par l'utilisateur, et on descend le long de la plus grande pente locale.

Remarque. La plupart des algorithmes trouvent des minimums locaux et non pas le minimum absolu du problème. Ils se répartissent en deux groupes: ceux qui n'utilisent pas la dérivée de f (qui seront présentés dans la section suivante) et ceux qui l'utilisent. Dans ce cas, une grand partie du calcul numérique consiste à approcher (ou trouver) la dérivée $f'(x)$.

Remarque 3.1 (La méthode de Newton-Raphson)**.** Soit $f : \mathbb{R} \to \mathbb{R}$ et soit ξ tel que $f(\xi) = 0$. On veut calculer une valeur approchée de ξ. En faisant $\xi = x_0 + h$ dans la définition ci-dessus, on a

$$f(\xi) - f(x_0) = f'(x_0)(\xi - x_0) + (\xi - x_0)\varepsilon(\xi - x_0),$$

c'est-à-dire

$$-f(x_0) \approx f'(x_0)(\xi - x_0).$$

On pose x_1 l'élément pour lequel on a l'égalité ci-dessus,

$$-f(x_0) = f'(x_0)(x_1 - x_0)$$

c'est-à-dire $x_1 = x_0 - f(x_0)/f'(x_0)$, pourvu que $f'(x_0) \neq 0$. En itérant, on obtient une suite qui converge vers ξ dans certains cas. On remarque que la construction du point x_1 est donnée par l'intersection de la tangente au graphe de f en x_0 avec l'abscisse.

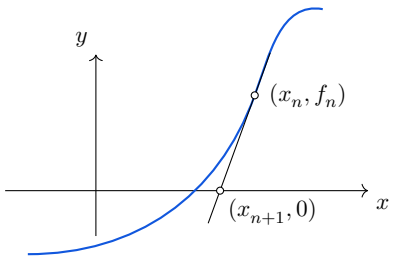


Figure 1: La méthode de Newton pour approcher une solution de l'équation $f(x) = 0$.

On a vu que la dérivée peut être utilisée pour trouver les extrémums locaux d'une fonction. Un point x_0 est dit point critique de f si $f'(x_0) = 0$.

Proposition 3.2 (Fermat)**.** *Soit $f : (a, b) \to \mathbb{R}$ une fonction différentiable. Si x_0 est un extrémum local de f, alors x_0 est un point critique.*

Théorème 3.3 (des accroissements finis)**.** *Soit $f : [a, b] \to \mathbb{R}$ continue et dérivable sur (a, b). Alors il existe $\xi \in (a, b)$ tel que*

$$f'(\xi) = \frac{f(b) - f(a)}{b - a}.$$

En changeant les notations, le théorème des accroissements finis devient

$$f(x+u) = f(x) + f'(x+tu)u \quad \text{avec } t \in [0,1].$$

Démonstration. On considère la fonction $g(x) = \dfrac{f(b)-f(a)}{b-a}(x-a) - f(x) + f(a)$. □

Corollaire 3.4. *f est croissante si et seulement si $f'(x) \geq 0$ pour tout $x \in (a,b)$. Si $f'(x) > 0$ pour tout $x \in (a,b)$ alors f est strictement croissante.*

Théorème 3.5. *Si $f : (a,b) \to \mathbb{R}$ est $n+1$ fois dérivable sur (a,b) et si $x \neq x_0 \in (a,b)$, alors il existe ξ entre x_0 et x tel que*

$$f(x) - T_{n,x_0}(x) = \frac{f^{n+1}(\xi)}{(n+1)!}(x-x_0)^{n+1}.$$

Démonstration. On considère la constante A telle que

$$f(x) - T_{n,x_0}(x) = A(x-x_0)^{n+1}.$$

On pose $g(t) = f(x) - T_{n,t}(x) - A(x-t)^{n+1}$. Alors

$$g'(t) = -\frac{f^{n+1}(t)}{n!}(x-t)^n + (n+1)A(x-t)^n$$

et comme $g(x) = g(x_0) = 0$, il existe ξ entre x_0 et x avec les propriétés désirées. □

Exercice 3.1. Montrer que

$$e - \left(1 + \frac{1}{1!} + \cdots + \frac{1}{n!}\right) < \frac{1}{n!}.$$

Corollaire 3.6. *Soit x_0 un point critique de f, fonction de classe $\mathcal{C}^{n+1}$, telle que*

$$f''(x_0) = \cdots = f^{n-1}(x_0) = 0 \quad \text{et} \quad f^n(x_0) \neq 0.$$

Les affirmations suivantes sont vraies :

- *Si n est pair et $f^{(n)}(x_0) > 0$, alors x_0 est un point de minimum.*
- *Si n est pair et $f^{(n)}(x_0) < 0$, alors x_0 est un point de maximum.*
- *Si n est impair, x_0 n'est pas un extrémum local.*

À partir de ce corollaire, nous en déduisons des conditions d'ordre 1 et 2 nécessaires (et suffisantes) pour qu'un point x^* soit un minimum local de f (où f est fonction d'une variable ou de plusieurs variables).

On se propose de minimiser $f : \mathbb{R} \to \mathbb{R}$ et on suppose qu'on connait ∇f ou qu'on peut l'approcher convenablement. Pour obtenir une approximation de x^*, un minimum local de f, on applique la descente de gradient (un procédé itérative). Partant d'un point x_0 choisi arbitrairement, un algorithme de descente cherche à générer une suite $(x_n)_{n\in\mathbb{N}}$ telle que

$$f(x_{n+1}) \leq f(x_n).$$

Pour la *descente de gradient*, l'idée est d'utiliser la relation

$$x_{n+1} = x_n + \delta_n\,(-\bigtriangledown f(x_n)),$$

où $\bigtriangledown f(x_n) = f'(x_n)$ et $\delta_n > 0$.

Pour justifier la viabilité de la méthode, remarquons que si on pose $\varphi(t) = f(0 - t\bigtriangledown f(0))$, alors, pour $t > 0$ suffisamment petit,

$$\varphi(t) - \varphi(0) = \varphi'(0)t + o(t) = f'(0)\Big(- \bigtriangledown f(0)\Big)t + o(t) = -\bigtriangledown f(0)^2 t + o(t) < 0.$$

Deux questions se posent : 1) Comment choisir δ_n ? et 2) Quel est le critère d'arrêt ?

Remarque. La méthode de descente se généralisent pour les problèmes en plusieurs dimensions avec en plus, la question portant sur le choix de la direction de descente. Pour la descente de gradient on prend $u_n = -\bigtriangledown f(x_n)$; mais on peut aussi considérer une direction u_n telle que $\bigtriangledown f(x_n)^T\, u_n < 0$.

Remarque 3.7 (Tests d'arrêt / de convergence)**.** En pratique un test d'arrêt doit être choisi pour garantir que l'algorithme s'arrête toujours après un nombre fini d'itérations et que le dernier point calculé soit suffisamment proche de x^*, un minimum local du problème.

Si $\varepsilon > 0$ est l précision demandée, alors les conditions d'arrêt sont les suivantes :

- $|\bigtriangledown f(x_n)| < \varepsilon$ (critère basé sur la condition nécessaire d'optimalité du premier ordre)
- $|x_{n+1} - x_n| < \varepsilon|x_n|$ (stagnation de la solution)
- $|f(x_{n+1}) - f(x_n)| < \varepsilon|f(x_n)|$ (stagnation de la valeur courante)
- $n <$ NIterMax (nombre d'itérations dépassant un seuil fixé)

Tout algorithme de descente de gradient ne converge pas vers un point optimal (local). Par exemple, pour la fonction $f(x) = x^2$, en prenant, pour tout $n \in \mathbb{N}$,

$$\delta_n = \frac{1}{2^{n+1}\bigtriangledown f(x_n)},$$

et en partant avec $x_0 = 2$, on obtient que $x_n \geq 0$ pour tout n et donc,

$$\begin{aligned} x_n &= x_{n-1} - \frac{1}{2^n \bigtriangledown f(x_{n-1})}\,\bigtriangledown f(x_{n-1}) = x_{n-1} - \frac{1}{2^n} \\ &= x_0 - \sum_{k=1}^{n} \frac{1}{2^k} = 2 - \frac{1}{2}\frac{1 - \frac{1}{2^n}}{1 - \frac{1}{2}} = 1 + \frac{1}{2^n} \longrightarrow 1. \end{aligned}$$

Pour caractériser la vitesse de convergence d'un algorithme, on introduit les notions suivantes qui mesurent l'évolution de l'erreur commise $|x_n - x^*|$: convergence linéaire, superlinéaire et d'ordre $p \in \mathbb{N}^*$. Explicitement, on a, respectivement,

$$\frac{|x_{n+1} - x^*|}{|x_n - x^*|} \longrightarrow \tau \text{ (avec } 0 < \tau < 1), \quad \frac{|x_{n+1} - x^*|}{|x_n - x^*|} \longrightarrow 0 \quad \text{et} \quad \frac{|x_{n+1} - x^*|}{|x_n - x^*|^p} \longrightarrow \tau.$$

Revenons aux choix de δ_n. Par la suite on décrit les trois grandes variations de cette méthode.

Descente de gradient à pas fixe. L'idée la plus simple est de prendre $\delta_n = \delta$, c'est-à-dire choisir le même pas à chaque itération. L'algorithme est appelé *la descente de gradient à pas fixe*. Difficile de dire à priori quelle valeur prendre pour δ. En général δ est beaucoup plus petit que la taille du domaine de recherche ; voir le théorème 3.8 qui suggère que la descente à pas fixe marche si la fonction n'a pas de variation brusque.

Théorème 3.8 (Convergence de la méthode du gradient à pas fixe)**.** *On suppose que la fonction $f : \mathbb{R} \to \mathbb{R}$ est de classe $\mathcal{C}^1$, qu'elle est **fortement convexe** de paramètre $\kappa > 0$ et qu'il existe $L > 0$ tel que pour tous $x_1, x_2 \in I$, on ait*

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

Alors si $\delta < 2\kappa/L^2$, la méthode du gradient à pas fixe converge vers l'unique solution du problème de minimisation.

Démonstration. Comme f est fortement convexe (donc strictement convexe), elle admet un unique point x^* de minimum. On veut trouver des conditions entraînant la convergence de la suite $(x_n)_n$ définie par $x_{n+1} = x_n - \delta f'(x_n)$ vers x^*. On considère l'application

$$\psi(x) = \psi_\delta(x) = x - \delta f'(x).$$

Le point x^* est fixe pour ψ. Il suffit de voir qu'elle est contractante. On a

$$\begin{aligned}
|\psi(b) - \psi(a)|^2 &= \Big((b-a) - \delta(f'(b) - f'(a))\Big)^2 \\
&= (b-a)^2 - 2\delta(b-a)(f'(b) - f'(a)) + \delta^2(f'(b) - f'(a))^2 \\
&\leq (b-a)^2 - 2\delta\kappa(b-a)^2 + \delta^2 L^2 (b-a)^2 \\
&= (1 - 2\delta\kappa + \delta^2 L^2)(b-a)^2.
\end{aligned}$$

Il s'ensuit que φ est contractante si $-2\kappa + \delta L^2 < 0$. □

Remarque. Une fonction $f : I \to \mathbb{R}$ est dite *fortement convexe de paramètre $\kappa > 0$* si pour tous $a < b \in I$ et tout $\lambda \in [0,1]$ on a

$$f((1-\lambda)a + \lambda b) \leq (1-\lambda)f(a) + \lambda f(b) - \frac{\kappa}{2}(1-\lambda)\lambda(b-a)^2.$$

En notant $x = (1-\lambda)a + \lambda b$, cette propriété devient

$$f(x) \leq \frac{b-x}{b-a}\,f(a) + \frac{x-a}{b-a}\,f(b) - \frac{\kappa}{2}\,(x-a)(b-x).$$

Avec des raisonnements similaires à ceux développés pour les fonctions convexe, on démontre les inégalités

$$\frac{f(x) - f(a)}{x-a} \leq \frac{f(b) - f(a)}{b-a} - \frac{\kappa}{2}\,(b-x)$$

et

$$\frac{f(b) - f(a)}{b-a} - \frac{\kappa}{2}\,(x-a) \leq \frac{f(b) - f(x)}{b-x}$$

d'où, pour tout $a < b$, l'inégalité $f'(b) - f'(a) \geq \kappa(b-a)$ utilisée dans la preuve du théorème 3.8.

Exemple. On considère la fonction $f(x) = -e^{-x^2}$, la précision $\varepsilon = 10^{-5}$, $N_{max} = 10000$ et $x_{ini} = 3$. Alors
- pour $\delta = 8 \cdot 10^{-3}$, l'algorithme converge en 358 pas
- pour $\delta = 10^{-2}$, l'algorithme ne converge pas
- pour $\delta = 2 \cdot 10^{-2}$, l'algorithme ne converge pas.

DESCENTE DE GRADIENT À PAS VARIABLE. La deuxième idée est de considérer un pas variable, c'est-à-dire d'essayer de minimiser "suffisamment" la fonction à chaque itération en faisant une *recherche linéaire* de δ_n. Les conditions que doit satisfaire le pas δ_n sont regroupées sous le nom de *conditions de Wolfe* et constituent les fondamentaux de la recherche linéaire moderne :

$$f(x + \delta u) \leq f(x) + c_1 \delta \bigtriangledown f(x)^T u \tag{3.1}$$
$$\bigtriangledown f(x + \delta u)^T u \geq c_2 \bigtriangledown f(x)^T u \tag{3.2}$$

avec $0 < c_1 < c_2 < 1$. (La première condition assure que le pas δ n'est pas trop grand ; la deuxième qu'il n'est ni trop petit.)

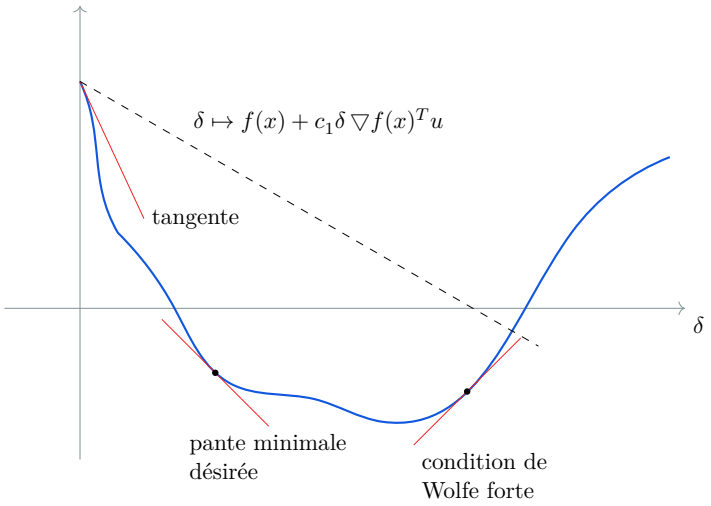


Figure 2: Les conditions de descente et de courbure de Wolfe

Lemme 3.9. *Soit $f : \mathbb{R} \to \mathbb{R}$ une fonction $\mathcal{C}^1$, bornée inférieurement. Soient $0 < c_1 < c_2 < 1$. Montrer que pour tout $a \in \mathbb{R}$ il existe un intervalle $I = I_a \subset]0, +\infty[$ de longueur > 0 tel que tout $t \in I$ vérifie les conditions de Wolfe forte en a, c'est-à-dire*

$$f(a - tf'(a)) \leq f(a) - c_1 t\, f'(a)^2$$
$$|f'(a - tf'(a))| \leq c_2 |f'(a)|.$$

Démonstration. Comme f est bornée inférieurement et comme $0 < c_1 < 1$, la droite définie par $y = f(a) - c_1 t\, f'(a)^2$ intersecte le graphe de $\varphi(t) = f(a - t\, f'(a))$ en un point $t_0 > 0$. D'après le théorème des accroissements finis, il existe $\theta \in]0, t_0[$ tel que $\varphi(t_0) - \varphi(0) = \varphi'(\theta) t_0$, c'est-à-dire

$$-t_0 f'(a) f'(a - \theta f'(a)) = f(a - t_0\, f'(a)) - f(a) = f(a) - c_1 t_0\, f'(a)^2 - f(a) = -c_1 t_0\, f'(a)^2.$$

Donc $f'(a) f'(a - \theta f'(a)) = c_1\, f'(a)^2 < c_2\, f'(a)^2$. On obtient l'intervalle par la continuité de f' et car $f'(a - \theta f'(a))$ et $f'(a)$ ont le même signe. □

Théorème 3.10 (Zoutendiejk)**.** *Soit $f : \mathbb{R}^n \to \mathbb{R}$ de classe $\mathcal{C}^1$, bornée inférieurement et $(x_n)_n$ une suite de descente de gradient obtenue en choisissant à chaque étape un pas satisfaisant les conditions de Wolfe (3.1) et (3.2).*

S'il existe $L > 0$ tel que

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$$

pour tout $x, y \in f^{-1}(]-\infty, f(x_0)])$, alors

$$\sum_{n\geq 0} \|\nabla f(x_n)\|^2 < +\infty.$$

Démonstration. En utilisant (3.2) et la définition de x_{n+1}, on a

$$(\nabla f_{n+1}^T - \nabla f_n^T)\nabla f_n \leq (c_2 - 1)\|\nabla f_n\|^2,$$

donc

$$(\nabla f_n^T - \nabla f_{n+1}^T)\nabla f_n \geq (1 - c_2)\|\nabla f_n\|^2$$

Mais, en usilisant la diéfinition de x_{k+1} dans la dernière égalité,

$$(\nabla f_n^T - \nabla f_{n+1}^T)\nabla f_n \leq \|\nabla f_n^T - \nabla f_{n+1}^T\| \, \|\nabla f_n\| \leq L\|x_k - x_{k+1}\| \, \|\nabla f_n\| = \delta_n L\|\nabla f_n\|^2.$$

Donc

$$(1 - c_2)\|\nabla f_n\|^2 \leq \delta_n L\|\nabla f_n\|^2$$

c'est-à-dire

$$\delta_n \geq \frac{1 - c_2}{L}.$$

En substituant cette inégalité dans la condition (3.1), on obtient

$$f_{n+1} \leq f_n - c_1\delta_n \|\nabla f_n\|^2 \leq f_n - \frac{c_1(1 - c_2)}{L} \|\nabla f_n\|^2.$$

En prenant la somme de ces relations pour tous les entiers $\leq n$, il s'ensuit que

$$f_{n+1} \leq f_0 - \frac{c_1(1 - c_2)}{L} \sum_{k=0}^{n} \|\nabla f_k\|^2.$$

Donc $\sum_{k=0}^{n} \|\nabla f_k\|^2$ est bornée supérieurement. □

Corollaire 3.11. *L'algorithme de descente de gradient avec recherche linéaire (en dimension 1) avec un pas de Wolfe se termine en un nombre fini de pas.*

Par exemple, pour la fonction de l'exemple précédent, $f(x) = e^{-x^2}$, en commençant avec $\delta = .1$, l'algorithme finit en une vingtaine de pas.

Pour la fonction de l'exemple précédent, en commençant avec $\delta = .1$, l'algorithme finit en une vingtaine de pas.

Descente de gradient à pas variable de Wolfe

Arguments : f, x_{ini}, δ, ε, N_{max}
But : approcher une solution x^* de $\bigtriangledown f = 0$ un utilisant la descente de gradient avec un pas de Wolfe

---

$x_0 \leftarrow x_{ini}$, calculer $\bigtriangledown f_0$
while $|\bigtriangledown f_0| \geq \varepsilon$ et $N < N_{max}$
    $\varphi(t) = f(x_0 - t\bigtriangledown f_0)$
    $\delta_0 \leftarrow \text{lineSearch}(\varphi, \delta)$
    $x_0 \leftarrow x_0 - \delta_0 \bigtriangledown f_0$
    calculer $\bigtriangledown f_0$ $N \leftarrow N + 1$
return x_0

Méthode de Newton. La troisième idée est d'utiliser la dérivée seconde, c'est-à-dire à regarder la condition d'optimalité d'ordre deux. L'algorithme porte le nom d'après Newton, car c'est l'algorithme de la tangente pour la fonction $\bigtriangledown f$. On a

$$f(x_n + h) = f(x_n) + \bigtriangledown f(x_n)h + \frac{1}{2}\bigtriangledown^2 f(x_n)h^2 + o(h^2)$$

et donc $\bigtriangledown f(x_n + h) = \bigtriangledown f(x_n) + \bigtriangledown^2 f(x_n)h + o(h)$. Pour minimiser f dans le voisinage de x_n, on cherche une solution de l'équation $\bigtriangledown f(x_n + h) = 0$. On obtient

$$h^* = -\frac{\bigtriangledown f(x_n)}{\bigtriangledown^2 f(x_n)} + \frac{o(h)}{\bigtriangledown^2 f(x_n)}$$

et on prend $\delta_n = \frac{1}{\bigtriangledown^2 f(x_n)}$. Quand $\bigtriangledown^2 f(x_n) > 0$ — on veut que la suite $(x_n)_n$ se rapproche de x^*, un point de minimum local de f —, la formule de récurrence devient

$$x_{n+1} = x_n + \frac{1}{\bigtriangledown^2 f(x_n)}(-\bigtriangledown f(x_n)). \tag{3.3}$$

Exemple. La formule (3.3) est utilisable seulement si x_0 est dans la partie convexe de f autour du minimum local ! On peut considérer $f(x) = \sin(x^2)$ entre 0 et 4. Alors pour $x_0 = 1.5$, la formule (3.3) produit le point de maximum $x^* = 1.2533143$, et pour $x_0 = 2$, le point de minimum $x^* = 2.1708044$.

Remarque 3.12. La formule (3.3) représente la méthode (itérative) de Newton pour résoudre l'équation $\bigtriangledown f(x) = 0$ — condition nécessaire pour un point de minimum. En plusieurs dimensions on obtient une formule similaire, sinon un peu plus compliquée car $\bigtriangledown f(p)$ est un vecteur et $\bigtriangledown^2 f(p)$ est une matrice carrée, la hessienne. Il existe trois types de coûts associés à la méthode de Newton : de dérivation, de calcul, de stockage — dans sa forme classique, on a besoin du calcul d'une dérivée d'ordre deux, de la résolution d'un système linéaire et du stockage d'une matrice.

Explicitement, si

$$Q(p) := f(x_n) + \bigtriangledown f(x_n)^T p + \frac{1}{2} p^T \bigtriangledown^2 f(x_n)\, p$$

est le polynôme de degré 2 associé à f en p, à chaque itération, la méthode de Newton approche $f(x)$ par Q, minimise Q comme fonction de p et pose

$$x_{n+1} = x_n + p_0 \quad \text{où} \quad \bigtriangledown^2 f(x_n)\, p_0 = -\bigtriangledown f(x_n).$$

La méthode de Newton représente "la méthode idéale" de minimisation. Elle peut ne pas marcher (si la hessienne n'est pas définie positive par exemple) ou demander trop de calculs. Virtuellement, toutes les autres méthodes de descente qui utilisent le gradient, sont des versions de la méthode de Newton.

4. Optimisation sans contraintes en dimension 1 — le cas non différentiable

On recherche le point de minimum local sans utiliser le gradient. Les méthodes sont itératives (de descente) et utilisent l'évaluation de la fonction à optimiser en certains points de son domaine. Il existe plusieurs idées menant à des algorithmes itératifs Les données nécessaires à chaque itération (et donc à l'initialisation) sont de trois types :

— trois points ordonnés tels que les valeurs correspondantes de la fonction satisfassent certaines inégalités ⇝ algorithme basé sur la dichotomie et algorithme basé sur l'interpolation quadratique
— deux points ⇝ algorithme de Nelder-Mead
— un point et un réel strictement positif ⇝ algorithme basé sur la recherche de motif.

La dichotomie. L'idée la plus simple est celle semblable à la méthode de bissection pour trouver un zéro d'une fonction réelle. On décrit l'algorithme : On commence par encadrer le minimum entre trois points $a < b < c$ tels que $f(b) < f(a)$ et $f(b) < f(c)$. Par la suite on applique les étapes illustrées dans l'algorithme ci-dessous. On remarque qu'à chaque étape, le choix du point suivant est fait dans le sous-intervalle le plus grand.

Dichotomie

Arguments : a, b, c tels que $f(a), f(c) > f(b)$
But : Trouver un point de minimum local entre a et b.

---

calculer p le milieu du plus long des intervalles $[a, b]$ et $[b, c]$
while $c - a > \varepsilon$
    construire le nouveau triplet (a, b, c) — à partir de a, b, c et p — en comparant $f(p)$ avec les valeurs $f(a)$, $f(b)$ et $f(c)$; il faut que $f(a) > f(b)$ et $f(c) > f(b)$
    calculer le nouveau p
return b et $f(b)$

L'interpolation quadratique. Une variante de l'algorithme précédent est suggérée par le lemme suivant. Il faut remarquer que pour initialiser l'algorithme, on suppose connus les points initiaux a, b et c pour lesquels $f(a) > f(b)$ et $f(c) > f(b)$ — de cette manière l'abscisse du sommet de la parabole sera contenue entre a et c. Même si f n'est pas quadratique, on utilise le calcul du lemme pour trouver un nouvel encadrement du minimum local x^* de f . Il faut imaginer un algorithme itérative... Voir la figure 3.

Lemme 4.1. *Si f est une fonction quadratique et on connait ses valeurs en a, b et c, alors son minimum est atteint en*

$$x^* = b + \frac{1}{2}\frac{(c-b)^2(f(b)-f(a)) + (a-b)^2(f(b)-f(c))}{(c-b)(f(b)-f(a)) - (a-b)(f(b)-f(c))}.$$

Démonstration. On écrit f en utilisant le polynôme d'interpolation de Lagrange : la fonction

de degré 2 passant par $(a, f(a))$, $(b, f(b))$ et $(c, f(c))$ est donnée par

$$\begin{aligned} f(x) &= \sum f(a)\,\frac{(x-b)(x-c)}{(a-b)(a-c)} \\ &= \sum \frac{f(a)}{(a-b)(a-c)}\,x^2 - \sum \frac{f(a)(b+c)}{(a-b)(a-c)}\,x + \cdots . \end{aligned}$$

Le signe somme signifie “prendre toutes les expressions obtenues par permutations circulaires” du triplet (a, b, c). Donc

$$\begin{aligned} x^* &= \frac{1}{2}\frac{f(a)(b^2-c^2)+f(b)(c^2-a^2)+f(c)(a^2-b^2)}{f(a)(b-c)+f(b)(c-a)+f(c)(a-b)} \\ &= b + \frac{1}{2}\frac{1}{f(a)(b-c)+f(b)(c-a)+f(c)(a-b)}\,[f(a)(b^2-c^2) \\ &\quad + f(b)(c^2-a^2)+f(c)(a^2-b^2)-2f(a)(b^2-bc)-2f(b)(bc-ba)-2f(c)(ab-b^2)]. \end{aligned}$$

En ajoutant $f(b)b^2 - f(b)b^2$ au numérateur, on obtient

$$\begin{aligned} x^* &= b + \frac{1}{2}\frac{-f(a)(b-c)^2+f(b)(b-c)^2-f(b)(a-b)^2+f(c)(a-b)^2}{f(a)(b-c)+f(b)(c-a)+f(c)(a-b)} \\ &= b + \frac{1}{2}\frac{(c-b)^2(f(b)-f(a))+(a-b)^2(f(b)-f(c))}{(c-b)(f(b)-f(a))-(a-b)(f(b)-f(c))}. \end{aligned}$$

□

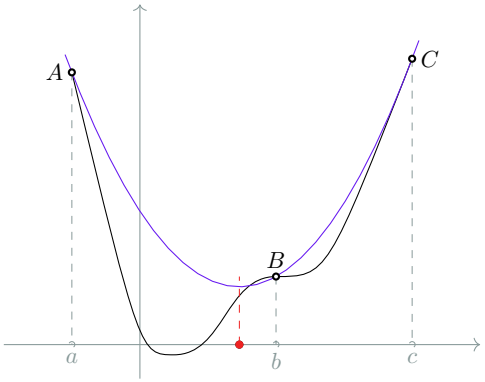


Figure 3: La courbe bleu est la parabole par les trois points A, B et C. Le point rouge, correspondant au minimum de la parabole, donne la coordonnée qui prendra la place de b pour l'étape suivante (a ne change pas et c devient b) de l'algorithme itérative cherchant le minimum de f entre a et c.

Remarque. La méthode parabolique ne fonctionne pas pour des fonctions linéaires. (Pourquoi ?) Donc si on est trop prés du minimum, la fonction devient localement linéaire et la méthode arrête de converger. Le triplet (A, B, C) devient constant. Si cela arrive, il suffit de considérer que la solution est donnée par B.

Exemple. On considère la fonction $f(x) = \sin(x^2)$ pour $x \in [0, 4]$. On commence avec $(a, b, c) = (0, 2, 4)$. On obtient, en appliquant la méthode parabolique, les évolutions illustrées dans la figure 4.

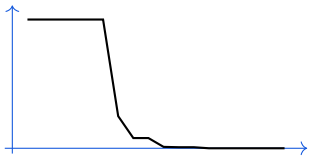


Figure 4: L'abscisse mesure le nombre d'itérations. Pour la dernière, $k = 18$, on a la valeur ...

Algorithme de Nelder et Mead. Par rapport aux deux algorithmes précédents, cet algorithme travaille, à chaque itération, avec un intervalle (un simplexe de points en plusieurs dimensions, c'est-à-dire trois en dimension 2, 4 en dimension 3,... qui doivent être affinement indépendants). Le critère d'arrêt sera donné par la taille de l'intervalle.

En dimension 1 on a l'algorithme décrit ci-dessous.

L'algorithme de Nelder-Mead

Arguments : f et le simplexe $[x_0x_1]$, $0 < c < 1 \leq r, d$
But : arriver "en descente" vers un simplexe tel que $|x_1 - x_0| < \varepsilon$

ordonner les points x_0 et x_1 tels que $f_0 \leq f_1$
while $|x_0 - x_1| > \varepsilon$
    $x_r \leftarrow x_0 - r(x_1 - x_0)$ # *vers la réflexion*
    if $f_r < f_0$
        $x_d \leftarrow x_0 + d(x_r - x_0)$ # *vers la dilatation*
        if $f_e < f_r$
            dilatation du simplexe $[x_0x_1] \leftarrow [x_0x_d]$
        else
            réflexion du simplexe par rapport à x_0, c'est-à-dire $[x_0x_1] \leftarrow [x_0x_r]$
    else if $f_0 \leq f_r < f_1$
        $x_c = x_0 + c(x_r - x_0)$ # *vers la contraction*
        if $f_c < f_r$
            contraction du simplexe $[x_0x_1] \leftarrow [x_0x_c]$
        else
            rétrécissement du simplexe par rapport à x_0
    else
        rétrécissement du simplexe par rapport à x_0
    ordonner les (nouveaux) points x_0 et x_1 tels que $f_0 \leq f_1$
return $[x_0x_1]$

Il faut préciser le sens de l'instruction *effectuer le rétrécissement du simplex* $[x_1, x_2]$ *centré en* x_1 et aussi qui sont les valeurs des constantes positives r, e et c. Le rétrécissement dépend d'une constante positive $\sigma < 1$ et signifie le remplacement de x_2 par le point $x_1 + \sigma(x_2 - x_1)$, c'est-à-dire

$$x_2 \leftarrow x_1 + \sigma(x_2 - x_1).$$

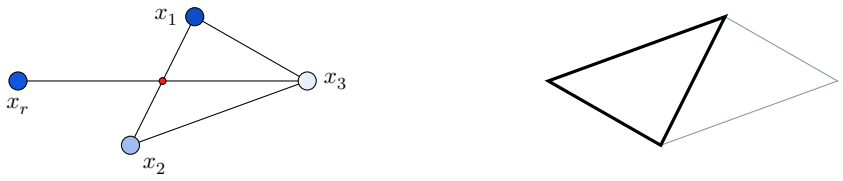
Chaque constante est associée à une transformation du simplexe initial $[x_1, x_2]$. Les noms des opérations et les valeurs habituelles des constantes sont donnés plus bas. On indique

graphiquement ces transformations en dimension 2. (Il faut remarquer que la première configuration ne peut pas apparaître en dimension 1 car x_1, $\bar{x}$ et x_2 représentent le même point.) Le point rouge est le centroïde des points $x_1, \ldots, x_n$ en dimension n. En général, on a

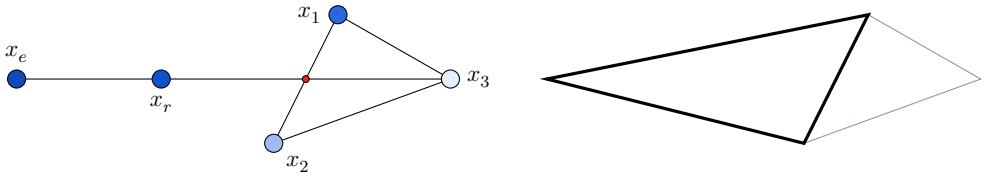
$$r > 0, \quad e > 1, \quad e > r, \quad \text{and} \quad 0 < c, \sigma < 1.$$

Remarque 4.2 (Nelder-Mead en plusieurs dimensions)**.** En plusieurs variables, les transformations du simplex lors des itérations sont les suivantes :

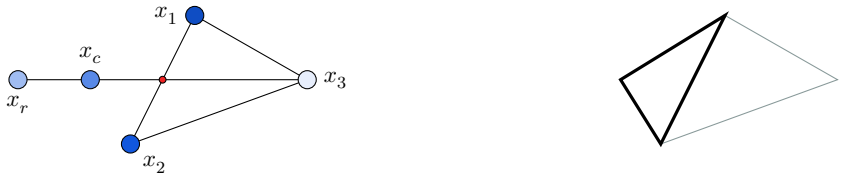
réflexion : $r = 1$



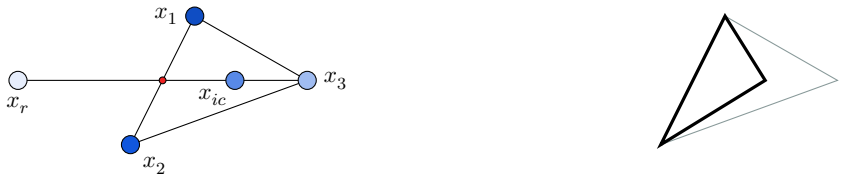
expansion : $e = 2$



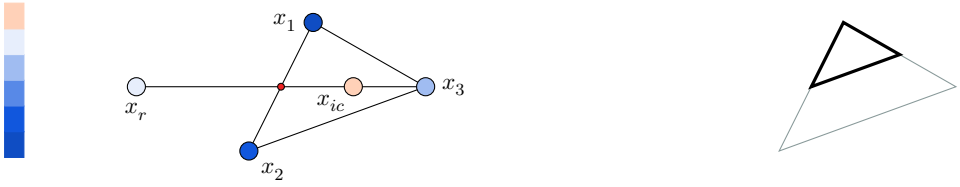
contraction extérieure : $c = 1/2$



contraction intérieure : $c = 1/2$ (n'apparaît pas en dimension 1)



rétrécissement : $\sigma = 1/2$



Remarque. Il y a un facteur 10 entre l'algorithme de Brent (parabolique) et celui de Nelder-Mead appliqués à la fonction

$$f(x) = \frac{1}{8}\,(35x^4 - 30x^2 + 3),$$

avec les points initiaux $x_1 = -1.2$ et $x_2 = -1$.

5. Fonctions différentiables à plusieurs variables. La méthode de Newton

5.1. La méthode de Newton (ou Newton-Raphson) en plusieurs variables

Si $F : \mathbb{R}^n \to \mathbb{R}^n$ est une fonction $\mathcal{C}^1$, on peut trouver une solution de l'équation $F(x) = 0$ en appliquant une variante en plusieurs variables de la méthode de Newton développée en une variable ; la formule de récurrence devient

$$x_{k+1} = x_k - \mathrm{Jac}(F, x_k)^{-1} F(x_k).$$

Exemple 5.1. On considère la fonction F définie par

$$\begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} y^2 - x^3 + x \\ y^3 - x^2 \end{pmatrix}.$$

La relation de récurrence devient

$$\begin{pmatrix} x_{n+1} \\ y_{n+1} \end{pmatrix} = \begin{pmatrix} x_n \\ y_n \end{pmatrix} - \begin{pmatrix} -3x_n^2 + 1 & 2y_n \\ -2x_n & 3y_n^2 \end{pmatrix}^{-1} \begin{pmatrix} y_n^2 - x_n^3 + x_n \\ y_n^3 - x_n^2 \end{pmatrix}.$$

Si on commence avec $(x_0, y_0) = (1, 1)$, alors la suite des approximations s'approche de $(1.46107, 1.2879)$ dans quelques pas. Si on commence avec $(x_0, y_0) = (-1, 1)$, alors la suite approche $(-0.47107, 0.60542)$. Ces valeurs approchées représentent des zéros du système formé par les équations $y^2 - x^3 + x = 0$ et $y^3 - x^2 = 0$. Mais si on commence avec $(x_0, y_0) = (-1, -1)$, alors $(x_1, y_1) = (-1.5, 0)$, et ce point est un point selle de F.

5.2. Extrémums locaux en plusieurs variables et méthodes de descente

Si $f : U \to \mathbb{R}$ est une fonction différentiables de plusieurs variables. La matrice jacobienne dans ce cas est une matrice ligne qu'on peut interpréter comme un vecteur, c'est-à-dire $J(f)^T$. En faisant cette interprétation, on utilise la notation $\bigtriangledown f := J(f)^T$; alors la valeur de l'application linéaire $f'(a)$ en v est donnée par $f'(a)v = \bigtriangledown f \cdot v = \bigtriangledown f^T v$.

Définition. Un point $a \in U$ est dit point critique de f si $\bigtriangledown f(a) = 0$.

Si un point $a \in U$ n'est pas un point critique de f, alors on peut utiliser le vecteur gradient pour se déplacer dans le voisinage de a et rendre f strictement plus petite ou plus grande.

Lemme-Définition 5.2. *Soit $u \in \mathbb{R}^n$ tel que $\bigtriangledown f(a) \cdot u < 0$, où $a \in U$, alors*

$$f(a + \lambda u) < f(a)$$

pour tout $\lambda > 0$ suffisamment petit. Un tel u est appelé ***direction de descente****.*

Démonstration. On écrit la définition de la différentiabilité de f en a :

$$f(a + u) - f(a) = \bigtriangledown f(a) \cdot u + \|u\| \, \varepsilon(u).$$

En remplaçant u par λu, on obtient

$$f(a+\lambda u) - f(a) = \lambda(\bigtriangledown f(a)\cdot u + \|u\|\,\varepsilon(\lambda u)) < 0$$

pour $\lambda > 0$ suffisamment petit, car $\varepsilon(\lambda u)$ est continue et nulle à l'origine. □

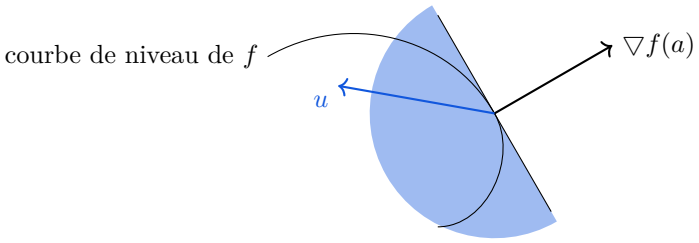


Figure 5: Le vecteur u est une direction de descente en a.

Si on regarde le comportement de la fonction f dans le voisinage d'un point x, on remarque que la direction dans laquelle la fonction décroit le plus rapidement est

$$u = -\bigtriangledown f(x).$$

Pour justifier cette affirmation, nous utilisons le théorème de Taylor. On a, pour un vecteur u de norme 1,

$$f(x+tu) = f(x) + t\bigtriangledown f(x)^T u + \frac{1}{2}\,t^2 u^T \bigtriangledown^2 f(x+\theta u)u$$

avec $\theta \in]0,t[$. Le taux de variation de f en x dans la direction u est donné par le coefficient de t, c'est-à-dire $\bigtriangledown f(x)^T u$. La direction de la plus rapide descente est donnée par la solution de

$$\min_{\|u\|=1} \bigtriangledown f(x)^T u = \min_{\|u\|=1} \|f(x)\|\,\|u\|\cos\alpha = \min_{0\le\alpha\le 2\pi} \|f(x)\|\cos\alpha,$$

avec α l'angle déterminé par les vecteurs $\bigtriangledown f(x)$ et u. On obtient la solution $\alpha = \pi$, c'est-à-dire la valeur minimale est atteinte en $u = -\bigtriangledown f(x)$.

Le lemme 5.2, comme dans le cas d'une variable. suggère une méthode pour approcher un extrémum local d'une fonction $f : U \to \mathbb{R}$. Si $\bigtriangledown f(x_0) \neq 0$, on a vu qu'un choix naturel pour u, la direction de descente, est $-\bigtriangledown f(x_0)$. Il reste à choisir λ, le pas, pour construire $x_1 = x_0 + \lambda u$ tel que $f(x_1) < f(x_0)$. Cette méthode itérative apparait en 1847 dans les travaux de Cauchy. Le lemme 5.2 ne dit rien en ce qui concerne le choix de λ. Pour trouver un bon λ, on peut

— choisir λ constant $\rightsquigarrow$ l'algorithme de descente de gradient à pas constant
— choisir λ tel que les conditions de Wolfe soient satisfaites $\rightsquigarrow$ l'algorithme de descente de gradient à pas variable
— choisir $\lambda = \operatorname{argmin}_{t\ge 0} f(x_0 + tu)$ (le premier point de minimum local) $\rightsquigarrow$ l'algorithme de descente de gradient à pas optimal.

L'algorithme de descente de gradient (sous toutes ses formes), se trouve au pôle opposé par rapport à la *méthode de Newton* pour la recherche d'un zéro du gradient de f — la suite $(x_n)_n$ est définie par

$$x_{n+1} = x_n - (\bigtriangledown^2 f_n)^{-1}\cdot \bigtriangledown f_n,$$

c'est-à-dire on modifie le terme x_n par la solution du système linéaire

$$\bigtriangledown^2 f_n \cdot u = -\bigtriangledown f_n.$$

Une condition nécessaire pour appliquer la méthode de Newton est que la hessienne soit définie positive. Entre ces deux pôles s'échelonnent les autres algorithmes qui essayent d'éviter la lenteur de la descente de gradient d'une part, et le calcul de la Hessienne et la sensibilité liée au point initial de l'autre.

descente de gradient GC BFGS Newton

La descente de gradient remplace la hessienne $\nabla^2 f_n$ par la matrice identité. Elle s'applique mal en pratique (la converge peut être très lente, parfois même en dessous de la précision de l'arithmétique de la machine) même si elle évite tous les coûts (dérivation, système linéaire, stockage) de la méthode de Newton.

5.3. Descente de gradient

Pour trouver $\lambda_n \geq 0$ — le pas — lors de la descente de gradient, on peut utiliser un pas variable ou un pas optimal. Pour le cas variable, on a, ou bien le choix donné par l'utilisation de la dérivée en dimension 1, ou bien l'algorithme parabolique (de Brent par exemple), ou l'utilisation d'un pas dans de la méthode de Newton en dimension 1 quand la deuxième dérivée est strictement positive. Plus précisément, comme

$$f(x_n - t\nabla f(x_n)) = g(t) \approx f(x_n) + t\nabla f(x_n)^T u + \frac{t^2}{2}\, u^T \nabla^2 f(x_n) u,$$

où $u = -\nabla f(x_n)$, en dérivant par rapport à t, on arrive à la condition

$$t\, u^T \nabla^2 f(x_n) u = -\nabla f(x_n)^T u.$$

Donc

$$\lambda = \frac{\|\nabla f(x_n)\|^2}{\nabla f(x_n)^T\, \nabla^2 f(x_n)\, \nabla f(x_n)}.$$

Comme on a précisé plus haut, ce choix est acceptable si le dénominateur est positif.

Remarque 5.3. On dispose d'un résultat général de convergence due à Zontendijk pour les algorithmes de descente avec "pas de Wolfe": Si ∇f est Lipschitzienne et bornée inférieurement, et si $x_{k+1} = x_k + \lambda_k u_k$ avec u_k direction de descente en x_k et λ_k qui vérifie les conditions de Wolfe, alors la série

$$\sum_k \cos^2(-\nabla f_k, u_k)\, \|\nabla f_k\|^2$$

converge.

Pour le choix du pas optimal, c'est-à-dire en cherchant λ_n le premier minimum local de $t \mapsto f(x_n - t\nabla f(x_n))$ (voir l'algorithme ci-dessous), la suite (x_n) a un comportement en "zig-zag" car les directions de descentes successives sont orthogonales. Même si cette descente est la plus profonde, la vitesse de convergence peut être très faible dû à ce comportement.

Exemple 5.4. On considère $f(x, y) = \frac{1}{2}\, x^2 + \frac{a}{2}\, y^2$, avec $a > 1$. Dans ce cas (quadratique), on peut expliciter la suite $(p_n)_n$ construite en appliquant la descente de gradient à pas optimal.

On a $\bigtriangledown f(x,y) = (x\ ay)$. Alors λ_n est le minimum de

$$t \longmapsto f(p_n - t\bigtriangledown f(p_n)) = \frac{1}{2}\,(x_n - tx_n)^2 + \frac{a}{2}\,(y_n - aty_n)^2.$$

On obtient

$$\lambda_n = \frac{x_n^2 + a^2y_n^2}{x_n^2 + a^3y_n^2}\left(= \frac{p_n^T A^2 p_n}{p_n^T A^3 p_n} \quad \text{en général}\right).$$

Donc

$$p_{n+1} = \begin{pmatrix} x_{n+1} \\ y_{n+1} \end{pmatrix} = \frac{(a-1)\,x_n y_n}{x_n^2 + a^3 y_n^2} \begin{pmatrix} a^2 y_n \\ -x_n \end{pmatrix}.$$

Il est facile de vérifier l'orthogonalité des gradients successifs :

$$\bigtriangledown f(p_n)^T \bigtriangledown f(p_{n+1}) = \text{const} \begin{pmatrix} x_n & ay_n \end{pmatrix} \begin{pmatrix} a^2 y_n \\ -ax_n \end{pmatrix} = 0.$$

Pour étudier la convergence de la suite $(p_n)_n$, on applique la norme $u \mapsto N_a(u) = \sqrt{u^T Au}$ aux termes de la suite. On obtient

$$N_a(p_{n+1})^2 = \left(\frac{(a-1)\,x_n y_n}{x_n^2 + a^3 y_n^2}\right)^2 (a^4 y_n^2 + a x_n^2) = \underbrace{\frac{a(a-1)^2 x_n^2 y_n^2}{(x_n^2 + a^3 y_n^2)(x_n^2 + a y_n^2)}}_{E(x_n^2, y_n^2)}\, N_a(p_n)^2.$$

On doit borné supérieurement l'expression $E(x_n^2, y_n^2)$ qui est homogène en x_n^2 et y_n^2. Il suffit d'étudier la variation d'une de-homogénéisation. On prend $x_n^2 = 1$ et $y_n^2 = s/a^2$; la fonction

$$h(s) = E(1, s/a^2) = \frac{a(a-1)^2 s/a^2}{(1+as)(1+s/a)} = \frac{(a-1)^2 s}{(1+as)(a+s)}$$

atteint son maximum global sur $]0, +\infty[$ en $s = 1$. Donc

$$N_a(p_{n+1}) \leq \frac{a-1}{a+1}\, N_a(p_n).$$

Dans le cas général (c'est-à-dire en dimension n), la constante s'exprime en fonction des valeurs propres minimale et maximale de la matrice A.

a	p_0	nombre d'itérations
$1 \cdot 11$	$(11, \frac{1}{2\cdot 1})$	47
$4 \cdot 11$	$(11, \frac{1}{2\cdot 2})$	264
$9 \cdot 11$	$(11, \frac{1}{2\cdot 3})$	502
$16 \cdot 11$	$(11, \frac{1}{2\cdot 4})$	666
$25 \cdot 11$	$(11, \frac{1}{2\cdot 5})$	772

Table 1: En variant a dans la fonction quadratique $f(x,y) = \frac{1}{2}\,x^2 + \frac{a}{2}\,y^2$(et le point initial en conséquence, c'est-à-dire en restant sur la courbe de niveau de même valeur) on obtient les donnés numériques ci-dessous. La condition d'arrêt est contrôlée par $\varepsilon = 10^{-4}$. La convergence vers le minimum global est de plus en plus lente.

Exemple 5.5. On considère $f(x,y) = x^2y^2(x^2+y^2-3)$. Alors

$$\bigtriangledown f = 2\big(xy^2(2x^2+y^2-3) \quad x^2y(x^2+2y^2-3)\big)^T$$

Les points critiques de f $(\pm 1, \pm 1)$ et les axes de coordonnées. On remarque que la valeur minimale globale de f est -1 !

Descente de gradient à pas optimal

Arguments : f et x_0
But : approcher une solution x^* de $\bigtriangledown f = 0$

$\bigtriangledown f_0 = \bigtriangledown f(x_0)$
while $\|\bigtriangledown f_n\| > \varepsilon$
    $t_n \leftarrow \operatorname{argmin}_{t\geq 0} f(x_n - t\bigtriangledown f_n)$
    $x_{n+1} \leftarrow x_n - t_n \bigtriangledown f_n$
    $\bigtriangledown f_{n+1} = \bigtriangledown f(x_{n+1})$
return x_{n+1}

5.4. La méthode du gradient conjugué

La méthode du gradient conjugué résout le problème d'oscillation du chemin de descente autour de la direction du minimum par le rajout d'un terme de friction ; chaque pas dépend des deux dernières valeurs du gradient. Comme conséquence, les angles de la trajectoire proches de $\pi/2$ sont diminués. La méthode est une généralisation directe de la méthode établie dans le cas quadratique dont le but est la convergence en un nombre de pas inférieur à la dimension du domaine de définition de la fonction.

Soit $f(x) = \frac{1}{2}\, x^T Ax - b^T x$, $x \in \mathbb{R}^n$. On cherche une méthode de descente convergeant en un nombre de pas $\leq n$ à $x^* = A^{-1}b$. Par la suite on utilise les notations $g_k = \bigtriangledown f(x_k) = Ax_k - b$ et u_k pour le gradient et la direction de descente en x_k, respectivement. On pose $u_0 = -g_0$ et, à chaque itération, le pas λ_k est le minimum de

$$t \longmapsto f(x_k + tu_k) = \frac{1}{2}\, u_k^T Au_k\, t^2 + u_k^T(Ax_k - b)\, t + \Big(\frac{1}{2}\, x_k^T Ax_k - x^T b\Big).$$

Donc[1]

$$\lambda_k = -\frac{u_k^T(Ax_k - b)}{u_k^T Au_k} = -\frac{u_k^T g_k}{u_k^T Au_k} = \frac{g_k^T g_k}{u_k^T Au_k}. \qquad (\sharp_\lambda)$$

Dans le cas $x \in \mathbb{R}^2$, on veut choisir la direction de descente u_1 telle que que $x_2 = x^*$.
On a

$$x_1 = x_0 + \lambda_0 u_0 \quad \left(= x_0 + \frac{g_0^T g_0}{u_0^T Au_0}\, u_0 \qquad \text{(non utilisé)}\right)$$

donc, en multipliant avec A,

$$g_1 = g_0 + \lambda_0\, Au_0.$$

[1] La dernière égalité est évidente pour $k = 0$. Elle sera justifiée en général dans la proposition 5.6.

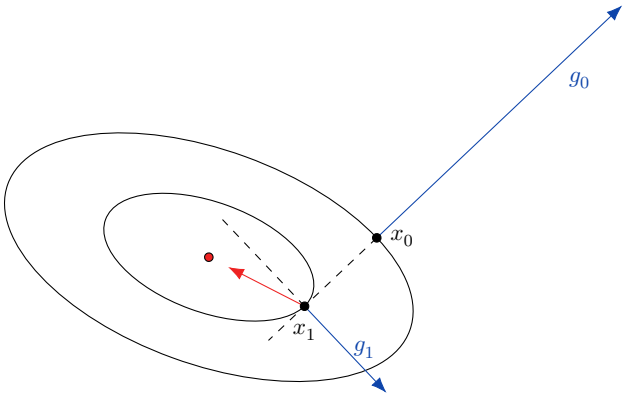


Figure 6: Le gradient conjugué dans le cas quadratique en dimension 2. L'algorithme converge en deux pas. La direction de descente u_1 est représentée en rouge l'angle décrit par u_0 et u_1 est plus petit que $\pi/2$.

On cherche le vecteur $u_1 = -g_1 + \beta u_0$ tel que

$$x_2 = x_1 + \lambda_1 u_1 = x^* = A^{-1}b.$$

En multipliant par A cette relation, on obtient

$$Ax_1 - b + \lambda_1 Au_1 = 0$$

c'est-à-dire

$$g_1 + \lambda_1 A(-g_1 + \beta u_0) = 0.$$

Comme g_0 et g_1 sont orthogonaux, on arrive à

$$0 = g_0^T g_1 + \lambda_1(-g_0^T A g_1 + \beta g_0^T A u_0) = \lambda_1(-g_0^T A g_1 + \beta g_0^T A u_0)$$

donc $\beta = \frac{g_0^T A g_1}{g_0^T A u_0}$. En utilisant la relation qui définit g_1, on a $Au_0 = \frac{1}{\lambda_0}(g_1 - g_0)$ et donc

$$\beta = \frac{g_0^T A g_1}{g_0^T A u_0} = -\frac{g_1^T(g_1 - g_0)}{(g_1^T - g_0^T)g_0} = \frac{g_1^T g_1}{g_0^T g_0}. \qquad (\sharp_\beta)$$

car $g_1^T g_0 = 0$.

Remarque. Il est important de comprendre la signification géométrique du choix du vecteur u_1, c'est-à-dire de la forme de $\beta = \beta_0$. On a, en utilisant la forme de λ_0,

$$u_0^T A u_1 = -u_0^T A g_1 + \beta u_0^T A u_0 = -\frac{1}{\lambda_0}(g_1 - g_0)^T g_1 + \beta u_0^T A u_0 = -\frac{u_0^T A u_0}{g_0^T g_0}\, g_1^T g_1 + \beta u_0^T A u_0 = 0,$$

car g_1 et g_0 sont orthogonaux. Donc, en dimension 2, la condition $x_2 = x^*$ est équivalente à la A-orthogonalité de u_0 et de u_1.

La discussion précédente suggère le lemme ci-dessous ; l'algorithme du gradient conjugué s'ensuivra.

Proposition 5.6. *Soit $f : \mathbb{R}^n \to \mathbb{R}$ définie par $f(x) = \frac{1}{2}\, x^T A x - b^T x$. Soient les suites $(x_k)_k$, $(g_k)_k$, $(u_k)_k$, $(\lambda_k)_k$ et $(\beta_k)_k$ construites par récurrence par*

- $x_0 \in \mathbb{R}^n$, $g_0 = \bigtriangledown f(x_0)$, $u_0 = -g_0$
- $\lambda_k = \dfrac{g_k^T g_k}{u_k^T A u_k}$, $x_{k+1} = x_k + \lambda_k u_k$ *et* $g_{k+1} = g_k + \lambda_k A u_k$
- $\beta_k = \dfrac{g_{k+1}^T g_{k+1}}{g_k^T g_k}$ *et* $u_{k+1} = -g_{k+1} + \beta_k u_k$.

Alors

(I) $g_{k+1} = \bigtriangledown f(x_{k+1})$

(II) $\mathrm{Vect}(g_0, \ldots, g_k) = \mathrm{Vect}(u_0, \ldots, u_k) =: E_k$ *pour tout* k

(III) $g_{k+1} \perp E_k$ *et* $u_{k+1} \perp_A E_k$ *pour tout* k *tant que* $E_k \neq \mathbb{R}^n$

(IV) *la suite* $(x_k)_k$ *est obtenue par un algorithme de descente à pas optimal, et d'après* (III), *elle converge en au plus* n *pas.*

Démonstration. Pour le premier point, on a

$$\bigtriangledown f(x_{k+1}) = A(x_k + \lambda_k u_k) - b = (Ax_k - b) + \lambda_k A u_k = g_{k+1}.$$

Pour le deuxième, si E_{k-1} est engendré par les familles $\{g_0, \ldots, g_{k-1}\}$ et $\{u_0, \ldots, u_{k-1}\}$, comme $u_k = -g_k + \beta_{k-1} u_{k-1}$, on a

$$\mathrm{Vect}(u_0, \ldots, u_k) = \mathrm{Vect}(g_k, E_{k-1}) = \mathrm{Vect}(g_0, \ldots, g_k).$$

Pour le troisième point, d'après l'hypothèse de récurrence, on a

$$\begin{aligned} g_{k+1}^T g_k &= (g_k + \lambda_k A u_k)^T g_k = g_k^T g_k + \lambda_k u_k^T A g_k \\ &= g_k^T g_k + \lambda_k u_k^T A(-u_k + \beta_{k-1} u_{k-1}) = g_k^T g_k - \lambda_k u_k^T A u_k = 0 \end{aligned}$$

et

$$\begin{aligned} u_{k+1}^T A u_k &= (-g_{k+1} + \beta_k u_k)^T A u_k = -g_{k+1}^T A u_k + \beta_k u_k^T A u_k \\ &= -g_{k+1}^T \frac{1}{\lambda_k}(g_{k+1} - g_k) + \beta_k u_k^T A u_k = \frac{1}{\lambda_k}(-g_{k+1}^T g_{k+1} + \beta_k \lambda_k u_k^T A u_k) \\ &= \frac{1}{\lambda_k}\left(- g_{k+1}^T g_{k+1} + \frac{g_{k+1}^T g_{k+1}}{g_k^T g_k} \frac{g_k^T g_k}{u_k^T A u_k} u_k^T A u_k \right) = 0. \end{aligned}$$

Le dernier point est évident car, dès que $\dim E_k = n$, les x_k stationnent. □

La méthode (ou l'algorithme) du gradient conjugué est une conséquence directe de la proposition précédente.

Dans le cas non linéaire, l'algorithme est le même sauf pour le calcul de λ_k et celui de g_{k+1}. Pour λ_k. il faut minimiser la restriction de f à la semi-droite passant par x_k de direction u_k. L'algorithme devient

Méthode du gradient conjugué (le cas quadratique)

Arguments : f, x_0, ε
But : Trouver un minimum local de f, fonction quadratique, en utilisant une descente le long de directions construites a partir du gradient "modifie".

$g_0 \leftarrow Ax_0 - b$ et $u_0 \leftarrow -g_0$
while $g_k > \varepsilon$

$$\lambda_k \leftarrow \frac{g_k^T g_k}{u_k^T A u_k}$$
$$x_{k+1} \leftarrow x_k + \lambda_k u_k$$
$$g_{k+1} \leftarrow g_k + \lambda_k A u_k$$
$$\beta_k \leftarrow \frac{g_{k+1}^T g_{k+1}}{g_k^T g_k} \quad \text{et} \quad u_{k+1} \leftarrow -g_{k+1} + \beta_k u_k$$

return x_{k+1}

5.5. La méthode quasi-Newton BFGS

Une autre méthode de descente, plus proche de la méthode de Newton, est la méthode BFGS. Cette méthode fait partie du groupe des méthodes *quasi-Newton* qui évitent le calcul de la matrice Hessienne par l'utilisation d'une approximation de celle-ci.

Comment calculer une approximation H_{k+1} en connaissant x_k, x_{k+1}, ∇f_k et ∇f_{k+1} ? En utilsant le développement limité de ∇f au voisinage de x_{k+1}, on obtient

$$\nabla f_k = \nabla f_{k+1} + \nabla^2 f_{k+1}(x_k - x_{k+1}) + o(x_k - x_{k+1}).$$

Donc

$$\nabla^2 f_{k+1}(x_k - x_{k+1}) \approx \nabla f_k - \nabla f_{k+1}$$

En général, on construit une approximation H_{k+1} comme solution de l'équation

$$\nabla f_{k+1} - \nabla f_k = H_{k+1}(x_{k+1} - x_k) \tag{5.1}$$

appelée équation de la sécante ou *de quasi-Newton*. Comme dans la méthode de Newton on utilise l'inverse de $\nabla^2 f_{n+1}$, il est plus utile de construire une approximation B_{k+1} de cette inverse, c'est-à-dire de considérer l'équation

$$B_{k+1}(\nabla f_{k+1} - \nabla f_k) = x_{k+1} - x_k.$$

Les équations sont sous-déterminées (n équations réelles avec n^2 inconnues). Même si on demande que B_{k+1} soit symétrique et définie positive, on a un nombre infini de solutions. L'idée proposée par Broyden en 1965, est de construire H_{k+1} comme solution du problème

$$\operatorname{argmin}_H \|H - H_k\|^2 \quad \text{quand } H \text{ vérifie} \quad \begin{cases} \nabla f_{k+1} - \nabla f_k = H(x_{k+1} - x_k) \\ H^T = H \end{cases}$$

pour une certaine norme matricielle.

La méthode la plus connue, BFGS, basée sur la norme

$$\|A\| = \|WAW\|_2$$

Mèthode du gadient conjugué (le cas général)

Arguments : f, x_0, ε
But : Trouver un minimum local de f en utilisant une descente le long de directions construites a partir du grandient "modifie".

---

calculer $u_0 = -\nabla f(x_0)$
while $\|\nabla f(x_k)\| > \varepsilon$
  chercher λ_k en minimisant $f(x_k + \lambda u_k)$ avec $\lambda \geq 0$
  $x_{k+1} \leftarrow x_k + \lambda_k u_k$
  calculer $\nabla f(x_{k+1})$
  $\beta_k \leftarrow \dfrac{\|\nabla f(x_{k+1})\|^2}{\|\nabla f(x_k)\|^2}$
  $u_{k+1} \leftarrow -\nabla f(x_{k+1}) + \beta_k u_k$
return x_{k+1}

où $\|\cdot\|_2$ est la norme de Frobenius et W est telle que W^2 est symétrique inversible vérifiant (5.1), c'est-à-dire $\nabla f_{k+1} - \nabla f_k = W^2(x_{k+1} - x_k)$. Pour énoncer le résultat, on utilise les notations

$$s_k = x_{k+1} - x_k \quad \text{et} \quad y_k = \nabla f_{k+1} - \nabla f_k.$$

La formule BFGS de l'itération est

$$H_{k+1} = H_k - \frac{(H_k s_k)(H_k s_k)^T}{s_k^T H_k s_k} + \frac{y_k y_k^T}{y_k^T s_k}.$$

Alors

$$B_{k+1} = \left(I - \frac{s_k y_k^T}{y_k^T s_k}\right) B_k \left(I - \frac{y_k s_k^T}{y_k^T s_k}\right) + \frac{s_k s_k^T}{y_k^T s_k}.$$

Remarque. Le calcul de B_{k+1} comme inverse de H_{k+1} est justifié par la formule de Sherman-Morrison : si A et $A + uv^T$ sont inversibles, alors $(A + uv^T)^{-1} = A^{-1} - \dfrac{A^{-1}uv^T A^{-1}}{1 + v^T A^{-1} u}$. Si A est symétrique, alors cette formule s'écrit comme

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

C'est sous cette forme qu'il faut utiliser la formule pour calculer l'inverse de H_{k+1} en deux étapes — étant donnée la forme de H_{k+1}.

Lemme 5.7. *Si $H_k > 0$ et H_{k+1} est obtenue par la formule BFGS, alors $H_{k+1} > 0$ si et seulement si $y_k^T s_k > 0$.*

Démonstration. Si $H_{k+1} > 0$, alors $s_k^T y_k = s_k^T H_{k+1} s_k > 0$. Dans l'autre sens, il suffit de montrer que si B est symétrique et définie positive, et si $u^T v > 0$, alors

$$C = B - \frac{(Bu)(Bu)^T}{u^T B u}$$

est “presque” définie positive (le vecteur v apparaîtra plus tard). On calcule

$$w^T C w = w^T B w - \frac{w^T B u u^T B w}{u^T B u} = \frac{1}{u^T B u}\big((u^T B u)(w^T B w) - (w^T B u)^2\big).$$

D'après l'inégalité de Cauchy-B-Schwarz, $w^T C w \geq 0$ avec égalité si et seulement si w et u sont liés. Il suffit de voir alors que le terme non-utilisé de notre formule, c'est-à-dire $uu^T/v^T v$, donne la positivité recherchée. □

La condition $y_k^T s_k > 0$ est satisfaite dès lors qu'on utilise la recherche de Wolfe pour le calcul du pas.

La méthode BFGS

Arguments : x_0, f, ε
But : Approcher un minimum local de f en utilisant la descente par la méthode quasi-Newton BFGS ; le pas de descente est déterminé avec une recherche de Wolfe.

$H_0 \leftarrow I$ et $u_0 \triangledown - \triangledown f_0$
trouver λ_0 par une recherche de Wolfe dans la direction u_0
while $\|\triangledown f_k\| > \varepsilon$
    $s_k \leftarrow \lambda_k u_k$
    $x_{k+1} = x_k + s_k$
    $y_k \leftarrow \triangledown f_{k+1} - \triangledown f_k$
    $H_{k+1} \leftarrow H_k - \frac{(H_k s_k)(H_k s_k)^T}{s_k^T H_k s_k} + \frac{y_k y_k^T}{y_k^T s_k}$
    $u_{k+1} \leftarrow -H_{k+1}^{-1} \triangledown f_{k+1}$ *# plus intéressant avec B_{k+1}*
    trouver λ_{k+1} par une recherche de Wolfe dans la direction u_{k+1}
return x_k

5.6. La méthode des moindres carrés

La fonction à minimiser est définie par

$$f(x) = \frac{1}{2}\sum_{j=1}^{m} r_j^2(x) = \frac{1}{2}\, r^T r \quad \text{où} \quad r = \begin{pmatrix} r_1 & \dots & r_m \end{pmatrix}^T.$$

Si on note par $J(x)$ la matrice jacobienne de r (de taille $m \times n$), alors

$$\bigtriangledown f = J^T r = \begin{pmatrix} \bigtriangledown r_1 & \dots & \bigtriangledown r_m \end{pmatrix} r, \quad \text{car} \quad \frac{\partial f}{\partial x_i} = \sum_{j=1}^{m} \frac{\partial r_j}{\partial x_i}\, r_j$$

et

$$\bigtriangledown^2 f = J^T J + \sum_{j=1}^{m} r_j \bigtriangledown^2 r_j.$$

En pratique, une fois la matrice J calculée, on obtient $\bigtriangledown f$ et la première partie de la hessienne. Assez souvent, le terme $J^T J$ est plus important par rapport au deuxième, ou bien car le modèle est presque linéaire autour du point optimal, ou bien car les résiduels (r_j) sont petits.

Dans le cas linéaire, c'est-à-dire quand $r(x) = Jx + r_0$, alors $\bigtriangledown f(x) = J^T(Jx + r_0)$ et $\bigtriangledown^2 f(x) = J^T J$. La fonction f est convexe *(ceci n'est pas toujours vrai dans le cas non linéaire)* et le (un si f n'est pas strictement convexe) point optimal est donné par le système

$$J^T J x = -J^T r_0.$$

Dans le cas non linéaire, la méthode la plus simple est celle de Newton-Gauss dans laquelle on obtient la direction de descente u_k en résolvant

$$J_k^T J_k u = -J_k^T r(x_k) = -\bigtriangledown f_k,$$

c'est-à-dire on considère seulement la première partie de la hessienne (voir la discussion ci-dessus).

6. Optimisation avec contraintes

6.1. Les multiplicateurs de Lagrange

Quelle est la valeur minimale de la fonction $f(x,y) = x + y$ si $x^2 + y^2 = 1$? C'est un exemple d'un problème de minimisation sujet à une contrainte non-linéaire. En général, en deux variables par exemple, on veut étudier le problème

$$\min\{f(x,y) \mid g(x,y) = 0\}. \tag{6.1}$$

Si X est le sous-ensemble des points déterminé par $g(x,y) = 0$, on étudie la variation de f quand (x,y) parcourt X. Un minimum local de f sur X est un point $\boldsymbol{a} \in X$ tel que

$$f(\boldsymbol{w}) \geq f(\boldsymbol{a}) = c_0$$

pour tout $\boldsymbol{w} = (x,y) \in U \cap X$, où U est un voisinage de $\boldsymbol{a}$. Intuitivement, on cherche la plus petite valeur c telle que les courbes définies par $f(x,y) = c$ et $g(x,y) = 0$ s'intersectent à peine en $\boldsymbol{a}$. De point de vue géométrique, les deux courbes doivent avoir une tangente commune, c'est-à-dire il existe $\lambda \in \mathbb{R}$ tel que

$$g(\boldsymbol{a}) = 0 \quad \text{et} \quad \bigtriangledown f(\boldsymbol{a}) + \lambda \bigtriangledown g(\boldsymbol{a}) = 0.$$

Une façon différente de formuler les équations ci-dessus pour l'exemple considéré précédemment, est de dire que le point $(\boldsymbol{a}, \lambda) \in \mathbb{R}^3$ est un point critique de la fonction lagrangienne

$$L(x,y,\lambda) = f(x,y) + \lambda g(x,y).$$

Le cas de deux variables. Pour transformer les idées esquissées ci-dessus en résultats mathématiques on utilise le théorème des fonctions implicites. Pour illustrer ce théorème dans l'exemple considéré précédemment, on veut exprimer y comme fonction de x dans un voisinage de $(0,1)$; on pose $y(x) = \sqrt{1-x^2}$. On a $g(x, y(x)) = 0$ pour tout $x \in]-1,1[$. Cette présentation est possible car $\dfrac{\partial g}{\partial y}(0,1) \neq 0$. Une telle présentation de la courbe $g(x,y) = 0$ dans le voisinage de $(-1,0)$ comme graphe d'une fonction en x est impossible !

On suppose que $\boldsymbol{a} = (x_0, y_0)$ est un minimum local pour le problème (6.1) avec

$$\frac{\partial g}{\partial y}(\boldsymbol{a}) \neq 0.$$

En exprimant la courbe $g = 0$ comme le graphe de $y = \varphi(x)$ dans un voisinage de $\boldsymbol{a}$ — en particulier $\varphi(x_0) = y_0$ —, le minimum local $\boldsymbol{a}$ pour f avec la contrainte $g = 0$ devient un minimum local, x_0 pour

$$F(x) = f(x, \varphi(x)).$$

On a

$$F'(x_0) = \frac{\partial f}{\partial x}(\boldsymbol{a}) + \frac{\partial f}{\partial y}(\boldsymbol{a})\varphi'(x_0) = 0$$

et

$$\frac{\partial g}{\partial x}(\boldsymbol{a}) + \frac{\partial g}{\partial y}(\boldsymbol{a})\varphi'(x_0) = 0,$$

car $g(x, \varphi(x)) \equiv 0$. Donc, en posant $\boldsymbol{v} = (1, \varphi'(x_0))$, on a

$$\bigtriangledown f(\boldsymbol{a}) \cdot \boldsymbol{v} = 0 \quad \text{et} \quad \bigtriangledown g(\boldsymbol{a}) \cdot \boldsymbol{v} = 0,$$

c'est-à-dire les vecteurs $\bigtriangledown f(\boldsymbol{a})$ et $\bigtriangledown g(\boldsymbol{a})$ sont liés. il existe donc un multiplicateur de Lagrange $\lambda \in \mathbb{R}$ tel que

$$\bigtriangledown f(\boldsymbol{a}) + \lambda \bigtriangledown g(\boldsymbol{a}) = 0.$$

Exemple 6.1. Pour minimiser $f(x, y) = x + y$ avec $g(x, y) = x^2 + y^2 - 1 = 0$, on a

$$\bigtriangledown f = (1, 1) \quad \text{et} \quad \bigtriangledown g = (2x, 2y),$$

donc on cherche à résoudre le système

$$\begin{cases} 1 + 2\lambda x & = 0 \\ 1 + 2\lambda y & = 0 \\ x^2 - y^2 - 1 & = 0. \end{cases}$$

On obtient $\lambda = \pm\sqrt{2}/2$ et les points

$$\pm \begin{pmatrix} \sqrt{2}/2 \\ \sqrt{2}/2 \end{pmatrix}.$$

Les signes correspondent aux minimum et maximum globaux du problème.

Les calculs en trois variables. En trois variables, on veut minimiser la fonction $f(x, y, z) = x+y+z$ sous les contraintes $x^2+y^2+z^2 = 1$ et $x-y+2z = 1$. On veut exprimer y et z comme fonctions de x, c'est-à-dire expliciter l'ensemble des contraintes X. On a $y = -1 + x + 2z$ et $x^2 + y^2 + z^2 = 1$ devient

$$0 = x^2 + (-1 + x + 2z)^2 + z^2 - 1 = 5z^2 + 4(x - 1)z + 2x(x - 1).$$

Donc

$$z = \frac{2(1-x)}{5} \pm \frac{\sqrt{4 + 2x - 6x^2}}{5} \quad \text{et} \quad y = -\frac{1-x}{5} \pm \frac{2\sqrt{4 + 2x - 6x^2}}{5}$$

pour $x \in [-2/3, 1]$. On prend le premier paramétrage donné par

$$y = -\frac{1-x}{5} + \frac{2\sqrt{4 + 2x - 6x^2}}{5} \quad \text{et} \quad z = \frac{2(1-x)}{5} + \frac{\sqrt{4 + 2x - 6x^2}}{5}.$$

Le problème de minimisation de f devient le problème de minimisation de la fonction d'une variable

$$\gamma : x \mapsto f(x, y(x), z(x)) = \frac{1 + 4x}{5} + \frac{3\sqrt{4 + 2x - 6x^2}}{5}$$

Mais

$$\gamma'(x) = \frac{4}{5} + \frac{3}{5} \frac{1 - 6x}{\sqrt{4 + 2x - 6x^2}}$$

et $\gamma' = 0$ mène à

$$84x^2 - 28x - 11 = 0$$

avec les solution $\frac{1}{6} \pm \frac{1}{21}\sqrt{70}$. On obtient le point optimal $x^* = \frac{1}{6} - \frac{1}{21}\sqrt{70}$ et on calcule $\gamma(x_0)$ pour avoir la valeur optimale.

Dans cet exemple on a pu expliciter la paramétrisation de X et finir le calcul de façon classique — la recherche d'un point extrémal pour une fonction définie sur un ouvert. Mais le plus souvent, il est impossible d'obtenir une forme explicite pour la paramétrisation. Si

$$g(x, y, z) = \begin{pmatrix} g_1(x, y, z) \\ g_2(x, y, z) \end{pmatrix},$$

alors $X : g(x, y, z) = \mathbf{0}$. Par le théorème des fonctions implicites, dans le voisinage d'un point $\boldsymbol{a}$ tel que

$$\det\left(\frac{\partial g}{\partial(y,z)}(\boldsymbol{a})\right) = \begin{vmatrix} \frac{\partial g_1}{\partial y}(\boldsymbol{a}) & \frac{\partial g_1}{\partial z}(\boldsymbol{a}) \\ \frac{\partial g_2}{\partial y}(\boldsymbol{a}) & \frac{\partial g_2}{\partial z}(\boldsymbol{a}) \end{vmatrix} \neq 0$$

il existe une paramétrisation $x \mapsto \varphi(x) \in \mathbb{R}^2$ tel que dans un voisinage de $\boldsymbol{a}$,

$$g(x, y, z) = \mathbf{0} \quad \text{si et seulement si} \quad (y, z) = \varphi(x).$$

(X est vue comme le graphe de la fonction $\varphi : I \to \mathbb{R}^2$ dans un voisinage de $\boldsymbol{a}$.) On suppose que $\boldsymbol{a} = (x_0, y_0, z_0)$ est un minimum local pour le problème avec

$$\det\left(\frac{\partial g}{\partial(y,z)}(\boldsymbol{a})\right) \neq 0.$$

Le minimum local $\boldsymbol{a}$ pour f avec la contrainte $G = \mathbf{0}$ devient un minimum local, x_0 pour

$$F(x) = (f \circ (1_X, \varphi))(x) = f(x, \varphi(x)) = f(x, \varphi_1(x), \varphi_2(x)).$$

On a

$$F'(x_0) = \frac{\partial f}{\partial x}(\boldsymbol{a}) + \frac{\partial f}{\partial y}(\boldsymbol{a})\varphi_1'(x_0) + \frac{\partial f}{\partial z}(\boldsymbol{a})\varphi_2'(x_0) = 0$$

en utilisant le fait que x_0 est un minimum de F, et

$$\frac{\partial g_j}{\partial x}(\boldsymbol{a}) + \frac{\partial g_j}{\partial y}(\boldsymbol{a})\varphi_1'(x_0) + \frac{\partial g_j}{\partial z}(\boldsymbol{a})\varphi_2'(x_0)$$

pour $j = 1, 2$ car $g_j(x, \varphi_1(x), \varphi_2(x)) \equiv 0$. Donc, en posant $\boldsymbol{v} = (1, \varphi_1'(x_0), \varphi_2'(x_0))$, on a

$$\bigtriangledown f(\boldsymbol{a}) \cdot \boldsymbol{v} = 0 \quad \text{et} \quad \bigtriangledown g_j(\boldsymbol{a}) \cdot \boldsymbol{v} = 0,$$

c'est-à-dire le vecteur $\bigtriangledown f(\boldsymbol{a})$ appartient au sous-espace engendré par les deux vecteurs linéairement indépendants $\bigtriangledown g_1(\boldsymbol{a})$ et $\bigtriangledown g_2(x_0)$. On conclut qu'*il existe les multiplicateurs de Lagrange* $\lambda_j \in \mathbb{R}$ *tels que*

$$\bigtriangledown f(\boldsymbol{a}) + \lambda_1 \bigtriangledown g_1(\boldsymbol{a}) + \lambda_2 \bigtriangledown g_2(\boldsymbol{a}) = 0.$$

Remarque. En appliquant les multiplicateurs de Lagrange à cet exemple on obtient

$$\bigtriangledown f = (1, 1, 1) \quad \text{et} \quad \bigtriangledown g_1 = (2x, 2y, 2z), \quad \bigtriangledown g_2 = (1, -1, 2)$$

donc on cherche à résoudre le système ayant 5 équations et 5 inconnues

$$1 - 2\lambda_1 x - \lambda_2 = 0 \qquad 1 - 2\lambda_1 y + \lambda_2 = 0 \qquad 1 - 2\lambda_1 z - 2\lambda_2 = 0$$

$$x^2 + y^2 + z^2 = 1 \quad \text{et} \quad x - y + 2z = 1.$$

En utilisant les trois premières équations on obtient les formules

$$x = \frac{1 - \lambda_2}{2\lambda_1} \qquad y = \frac{1 + \lambda_2}{2\lambda_1} \qquad z = \frac{1 - 2\lambda_2}{2\lambda_1}$$

qui avec la contrainte linéaire mènent à $\lambda_1 = 1 - 3\lambda_2$. La deuxième contrainte devient alors

$$1 = x^2 + y^2 + z^2 = \left(\frac{1 - \lambda_2}{2 - 6\lambda_2}\right)^2 + \left(\frac{1 + \lambda_2}{2 - 6\lambda_2}\right)^2 + \left(\frac{1 - 2\lambda_2}{2 - 6\lambda_2}\right)^2$$

avec les solutions $\lambda_2 = \dfrac{10 \pm \sqrt{70}}{30}$. On obtient $\lambda_1 = \mp\dfrac{\sqrt{70}}{10}$ et donc

$$x_0 = \frac{1}{6} \mp \frac{1}{21}\sqrt{70}, \qquad y_0 = -\frac{1}{6} \mp \frac{2}{21}\sqrt{70}, \qquad z_0 = \frac{1}{3} \mp \frac{1}{42}\sqrt{70}$$

avec $f(x_0, y_0, z_0) = \dfrac{1}{3} \mp \dfrac{1}{6}\sqrt{70}$. Le minimum de f restreinte à X est $\dfrac{1}{3} - \dfrac{1}{6}\sqrt{70}$ et il est atteint en

$$(x_0, y_0, z_0) = \left(\frac{1}{6} - \frac{1}{21}\sqrt{70}, -\frac{1}{6} - \frac{2}{21}\sqrt{70}, \frac{1}{3} - \frac{1}{42}\sqrt{70}\right).$$

Remarque. De point de vue géométrique, la droite tangente à la courbe X doit être contenue dans le plan tangent à la surface, c'est-à-dire il existe $\lambda_j \in \mathbb{R}$ tels que

$$g_j(\boldsymbol{a}) = 0 \quad \text{et} \quad \bigtriangledown f(\boldsymbol{a}) \in \mathrm{Vect}(\bigtriangledown g_1(\boldsymbol{a}), \bigtriangledown g_2(\boldsymbol{a})).$$

Une façon différente de formuler les équations ci-dessus est de dire que le point $(\boldsymbol{a}, \lambda_1, \lambda_2) \in \mathbb{R}^3 \times \mathbb{R}^2$ est un point critique de la fonction lagrangienne

$$L(x, y, z, \lambda_1, \lambda_2) = f(x, y, z) + \lambda_1 g_1(x, y, z) + \lambda_2 g_2(x, y, z).$$

Le cas général. On étudie le problème

$$\min\{f(\boldsymbol{x}) \mid \boldsymbol{x} \in X\},$$

où $f : \mathbb{R}^n \to \mathbb{R}$ et X est le sous-ensemble de $\mathbb{R}^n$ défini par le système (non linéaire)

$$\begin{cases} g_1(\boldsymbol{x}) = 0 \\ \vdots \\ g_r(\boldsymbol{x}) = 0. \end{cases}$$

Théorème 6.2. *On suppose que $\boldsymbol{a} \in X$ est un point d'extrémum local de f et que $\bigtriangledown g_1(\boldsymbol{a}), \ldots, \bigtriangledown g_r(\boldsymbol{a})$ sont linéairement indépendants. Alors il existe des scalaires $\lambda_1, \ldots, \lambda_r$ tels que le point $(\boldsymbol{a}, \lambda_1, \ldots, \lambda_r) \in \mathbb{R}^{n+q}$ soit critique pour le lagrangien*

$$L(\boldsymbol{x}, \boldsymbol{\lambda}) = f(\boldsymbol{x}) + \sum_{j=1}^{q} \lambda_j g_j(\boldsymbol{x}).$$

Démonstration. La condition d'indépendance linéaire dit que le sous-ensemble X est une sous-variété de dimension $n-r$ dans un voisinage de $\boldsymbol{a}$. D'après le théorème des fonctions implicites, on peut paramétrer X dans un voisinage de $\boldsymbol{a}$, c'est-à-dire il existe une application $\theta : U \subset \mathbb{R}^{n-r} \to \mathbb{R}^n$ telle que

1. il existe un ouvert $V \subset \mathbb{R}^n$, voisinage de $\boldsymbol{a}$, avec $\theta(U) = V \cap X$,
2. θ est injective,
3. θ' est de rang maximal pour tout $\boldsymbol{t} \in U$ et
4. $g_j \circ \theta \equiv 0$ pour tout $j = 1, \ldots, r$.

Pour finir, on considère

$$F = f \circ \theta$$

et on obtient

$$\bigtriangledown F(a) \cdot \boldsymbol{u} = \bigtriangledown f(a) \cdot \theta'(\boldsymbol{t}_0)(\boldsymbol{u}) = 0$$

pour tout $\boldsymbol{u} \in \mathbb{R}^{n-r}$. En utilisant les identités $g_j \circ \theta \equiv 0$, on a, pour tout $j = 1, \ldots, r$,

$$\bigtriangledown g_j(\boldsymbol{a}) \cdot \theta'(\boldsymbol{t}_0)(\boldsymbol{u}) = 0.$$

D'après la propriété 3), le sous-espace $K = \mathrm{im}(\theta'(\boldsymbol{t}_0))$ est de dimension $n - r$. D'après les hypothèses, $\bigtriangledown g_j(\boldsymbol{a})$ correspondent à r équations linéaires indépendantes sur $\mathbb{R}^n$; elles découpent un sous-espace de dimension $n - r$ qui contient K, donc qui s'identifie à K. Il s'en suit que l'équation correspondante à $\bigtriangledown f(\boldsymbol{a})$ est une combinaison linéaire de celles induites par les $\bigtriangledown g_j(\boldsymbol{a})$. □

Remarque. La propriété 4. dit que X est le graphe d'une application localement autour de $\boldsymbol{a}$. Par exemple, si

$$\begin{vmatrix} \dfrac{\partial g_1}{\partial x_1}(\boldsymbol{a}) & \cdots & \dfrac{\partial g_1}{\partial x_r}(\boldsymbol{a}) \\ \vdots & & \\ \dfrac{\partial g_r}{\partial x_1}(\boldsymbol{a}) & \cdots & \dfrac{\partial g_r}{\partial x_r}(\boldsymbol{a}) \end{vmatrix} \neq 0$$

alors on peut exprimer les coordonnées $x_1, \ldots, x_r$ en fonction des coordonnées $x_{r+1}, \ldots, x_n$. On obtient, pour $j = 1, \ldots, r$,

$$x_j = \varphi_j(x_{r+1}, \ldots, x_n)$$

et X est le graphe de $\overline{\boldsymbol{x}} \mapsto (\varphi_1(\overline{\boldsymbol{x}}), \ldots, \varphi_r(\overline{\boldsymbol{x}}))$ où $\overline{\boldsymbol{x}} = (x_{r+1}, \ldots, x_n)$. L'application θ apparaissant dans la preuve est donnée par

$$\theta(\overline{\boldsymbol{x}}) = (\varphi_1(\overline{\boldsymbol{x}}), \ldots, \varphi_r(\overline{\boldsymbol{x}}), \overline{\boldsymbol{x}}).$$

Remarque. La méthode du pivot de Gauss est un exemple du théorème des fonctions implicites. Les paramètres qui donnent la forme générale de la solution sont les composantes de $\overline{\boldsymbol{x}}$ ci-dessus.

Exemple 6.3. Trouver le minimum et le maximum de $f(x, y, z) = x - 2y + 2z$ sous les contraintes

$$\begin{cases} x^2 + y^2 + x^2 = 1 \\ x + y + z = 0. \end{cases}$$

On a $\bigtriangledown f = (1, -2, 1)$, on remarque que les deux équations découpent une sous-variété de dimension 1 (une courbe lisse) et on arrive au système

$$\begin{cases} x^2 + y^2 + x^2 & = 1 \\ x + y + z & = 0 \\ 2\lambda x + \mu & = 1 \\ 2\lambda y + \mu & = -2 \\ 2\lambda z + \mu & = 2. \end{cases}$$

En utilisant les quatre dernières équations, on obtient $\mu = 1/3$. Par la suite on obtient $\lambda = \pm\sqrt{13/6}$. Pour $\lambda = \sqrt{13/6}$, on obtient

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \frac{1}{\sqrt{78}} \begin{pmatrix} 2 \\ -\sqrt{7} \\ \sqrt{5} \end{pmatrix}.$$

On voit géométriquement que ce point est un point de maximum pour f sous les contraintes.

Exercice 6.1. Démontrer l'inégalité des moyennes en utilisant les multiplicateurs de Lagrange.

Exercice 6.2. Trouver le maximum global de $x^2 + y^2$ restreint à $x^2 + xy + y^2 = 4$. Peut-on trouver une interprétation géométrique à ce problème ?

Exercice 6.3. Une boite rectangulaire a les côtés x, y et z. Quel est le volume maximal de la boite si on suppose que (x, y, z) vérifie

$$\frac{x}{a} + \frac{y}{b} + \frac{z}{c} = 1,$$

avec $a, b, c > 0$.

Exercice 6.4. Une société commerciale doit produire une boite d'un volume de 2 mètres cubes. Les matériaux utilisés pour les différentes faces sont différents : le coût du mètre carré est de 1 \$ le pour les côtés et de 1.5 \$ pour les parties supérieure et inférieure. Trouver les dimension de la boite qui minimisent le coût.

7. Optimisation convexe

Soient $f, g_1, \ldots, g_m : U \to \mathbb{R}$ des fonctions différentiables. On considère le problème d'optimisation

$$\min\{f(x) \mid g_1(x) \leq 0, \ldots, g_m(x) \leq 0\} \tag{7.1}$$

(c'est-à-dire les contraintes sont des inégalités).

7.1. Le théorème de Karush-Kuhn-Tucker

Définition. Le système

$$\begin{aligned}
&\lambda_i \geq 0\\
&g_i(x_0) \leq 0\\
&\lambda_i g_i(x_0) = 0\\
&\bigtriangledown f(x_0) + \lambda_1 \bigtriangledown g_1(x_0) + \cdots + \lambda_m \bigtriangledown g_m(x_0) = 0
\end{aligned}$$

dans les inconnues $x_0 \in \mathbb{R}^n$ et $\lambda_1, \ldots, \lambda_m \in \mathbb{R}$ est appelé le système des conditions Karush-Kuhn-Tucker (KKT) pour le problème d'optimisation (7.1)

Théorème 7.1. *Soit $S \subset U$ le sous-ensemble défini par les contraintes du problème (7.1). On suppose que $x_0 \in S$ est un point optimal.*

1. *Si $\bigtriangledown g_i(x_0)$ sont linéairement indépendants, alors les conditions KKT sont satisfaites en x_0 pour des λ_i convenables.*
2. *Si g_i sont convexes et S admet un point* **strictement** *intérieur, alors les conditions KKT sont satisfaites en x_0 pour des λ_i convenables.*
3. *Si g_i sont affines, alors les conditions KKT sont satisfaites en x_0 pour des λ_i convenables.*

Théorème 7.2. *Si f et les g_i sont convexes et si les conditions KKT sont satisfaites en x_0 pour certains λ_i, alors x_0 est un point optimal.*

Démonstration. Si tous les λ_i sont nuls, alors $\bigtriangledown f(x_0) = 0$ et x_0 est un minimum global. Supposons $\lambda_1, \ldots, \lambda_r > 0$ et $\lambda_{r+1} = \cdots = \lambda_m = 0$. Alors $g_j(x_0) = 0$ pour tout $1 \leq j \leq r$, et l'équation du gradient devient

$$\bigtriangledown f(x_0) + \lambda_1 \bigtriangledown g_1(x_0) + \cdots + \lambda_r \bigtriangledown g_r(x_0) = 0.$$

Si x_0 n'est pas un minimum pour le problème, alors il existe $x \in S$ tel que $f(x) < f(x_0)$. Mais

$$f(x) \geq f(x_0) + \bigtriangledown f(x_0)(x - x_0),$$

donc $\bigtriangledown f(x_0)(x - x_0) < 0$. Comme les g_i sont des fonctions convexes, $[x_0 x] \subset S$. Donc $g_j(y) \leq 0$ pour tout $y \in [x_0 x]$. Alors $\bigtriangledown g_j(x_0)(x - x_0) \leq 0$. On arrive à une contradiction. □

7.2. Une méthode basée sur des points intérieurs

On considère un problème d'optimisation $\min\{f(x) \mid x \in S\}$ avec

$$S = \{x \in \mathbb{R}^n \mid g_j(x) \leq 0, 1 \leq j \leq m\}.$$

On suppose que toutes les fonctions sont différentiables et *convexes*, que S est compact et qu'il admet un point strictement intérieur. Alors le problème admet une solution optimale.

Pour la trouver, on pose

$$B(x) = -\sum_{j=1}^{m} \log(-g_j(x))$$

qui est une fonction convexe sur $S^\circ \neq \emptyset$, car on a supposé l'existence d'un point strictement intérieur. Pour voir ceci, on a le fait suivant : si $g : C \to]-\infty, 0[$ est convexe, alors $G(x) = -\ln(-g(x))$ est convexe. Pour le justifier, on calcule

$$\nabla G = -\frac{1}{g} \nabla g$$

et

$$\nabla^2 G = -\frac{1}{g} \nabla^2 g + \frac{1}{g^2} \nabla g (\nabla g)^T.$$

Il s'ensuit que $\nabla^2 G \geq 0$.

On considère la famille de fonctions convexes

$$f_\varepsilon(x) = f(x) + \varepsilon B(x),$$

pour $\varepsilon > 0$. Un minimum global de f_ε est atteint en $x_\varepsilon \in S^\circ$ car $B(x) \to +\infty$ si $x \to \partial S$.

Théorème 7.3. *Soit $f_\varepsilon(x_\varepsilon) = \min\{f_\varepsilon(x) \mid x \in S^\circ\}$ et soit $f^* = \min\{f(x) \mid x \in S\}$. Alors*

$$0 \leq f(x_\varepsilon) - f^* \leq m\,\varepsilon,$$

où m est le nombre de termes dans l'expression de B. Si le problème initial admet une solution unique x^*, alors $x_\varepsilon \to x^*$.

Démonstration. Comme x_ε est un minimum intérieur de f_ε, on a

$$0 = \nabla f_\varepsilon(x_\varepsilon) = \nabla f(x_\varepsilon) + \varepsilon \nabla B(x_\varepsilon) = \nabla f(x_\varepsilon) - \sum_j \frac{\varepsilon}{g_j(x_\varepsilon)} \nabla g_j(x_\varepsilon).$$

On en déduit que x_ε est un minimum pour le lagrangien

$$f(x) + \sum_{j=1}^{m} \lambda_j g_j(x)$$

avec

$$\lambda_j = -\frac{\varepsilon}{g_j(x_\varepsilon)} > 0.$$

Alors

$$\begin{aligned} f^* &\geq f(x^*) + \sum_j \lambda_j\, g_j(x^*) \geq \min\{f(x) + \sum_j \lambda_j\, g_j(x) \mid x \in S\} \\ &= f(x_\varepsilon) + \sum_j \lambda_j\, g_j(x_\varepsilon) = f(x_\varepsilon) - m\varepsilon. \end{aligned}$$

Comme $f^* \leq f(x_\varepsilon)$, on conclut que $f(x_\varepsilon) \to f^*$.

Si le problème initial admet une solution unique, alors, en prenant $\varepsilon = 1/k$, on obtient une suite $(x_k)_k$ telle que

$$x_k \in S \quad \text{et} \quad \lim_{k \to +\infty} f(x_k) = f(x^*).$$

Si $(x_k)_k$ ne tend pas vers x^*, alors il existe $\varepsilon_0 > 0$ et une sous-suite $(x_{k_n})_n$ telle que $\|x_{k_n} - x^*\| \geq \varepsilon_0$. Comme S est compact, on extrait une sous-suite convergente à a dans $S \setminus B_a(\varepsilon_0)$, et on arrive à la contradiction $f(a) = f(x^*)$. □

7.3. La méthode des points intérieurs pour des contraintes linéaires

Si le sous-ensemble S est défini par des contraintes linéaires, c'est-à-dire si S est le polyèdre convexe,

$$S : Ax \leq b.$$

Alors,

$$B(x) = -\sum_{j=1}^{m} \log(b_j - A_{j\bullet}x)$$

et le problème de minimisation pour $f_\varepsilon(x)$ à l'intérieur de S devient un problème classique pour une fonction convexe de plusieurs variables pour laquelle, pour calculer $\bigtriangledown f_\varepsilon(x)$ on a

$$\frac{\partial f_\varepsilon}{\partial x_i}(x) = \frac{\partial f}{\partial x_i}(x) + \varepsilon \sum_{j=1}^{m} \frac{a_{ji}}{b_j - A_{j\bullet}x}.$$

Donc

$$\bigtriangledown f_\varepsilon(x) = \bigtriangledown f(x) + \varepsilon \Big(\frac{1}{b - Ax}\Big)^T A$$

et

$$\bigtriangledown^2 f_\varepsilon(x) = \bigtriangledown^2 f(x) + \varepsilon A^T \operatorname{diag}\Big(\frac{1}{(b - Ax)^2}\Big) A.$$

Dans la méthode du gradient en une dimension, il suffit de faire la bissection, car la recherche est contrôlée par la borne

$$t_0 = \sup_{t \geq 0} \{x_0 + tv \in S^\circ\}.$$

Bibliographie

[1] F. de Gournay, A. Rondepierre, Introduction à l'optimisation numérique. INSA Toulouse

[2] N. Lauritzen, Undergraduate convexity. World Scientific

[3] S. Nash, A. Sofer, Linear and non-linear programming. McGraw-Hill International Editions

[4] J. Nocedal, S. J. Wright, Numerical optimization. Springer

[5] scipy.optimize, `https://docs.scipy.org/doc/scipy/reference/tutorial/optimize.html`