Systèmes non linéaires et optimisation

MAIN 3
Polytech Sorbonne

Année 2022-2023

Fabien Vergnet

fabien.vergnet@sorbonne-universite.fr

Révisions version 2023-2024

Version février 2024, Marie Postel, marie.postel@sorbonne-universite.fr

- Changement de signe du terme linéaire dans la forme quadratique $f(x) = \frac{1}{2}x.\cdot(Ax)+b\cdot x$ (au lieu de $-b\cdot x$ précédemment), dans le paragraphe "Méthode de gradient conjugué".

Révisions version 2024-2025

Version février 2025, Marie Postel, marie.postel@sorbonne-universite.fr

- Correction formule de Taylor Lagrange page 6

Table des matières

Chapitre 1

Résolution de systèmes non linéaires

1.1 Objectifs

Dans ce chapitre, nous considérons une application f continue de $\mathbb{R}^N$ dans $\mathbb{R}^N$ et nous cherchons à résoudre l'équation

$$\begin{cases} x^* \in \mathbb{R}^N, \\ f(x^*) = 0_{\mathbb{R}^N}, \end{cases} \tag{1.1}$$

Il est important de noter que les équations de la forme $g(x) = C$, où C est un vecteur fixé de $\mathbb{R}^N$, ou encore les équations du type $g(x) = h(x)$ peuvent évidement se récrire sous la forme du système (1.1).

En général, on ne peut pas donner d'expression analytique de la solution x^* du système (1.1). Il faut donc passer par un calcul approché de la solution x^*. Dans tous les cas, nous considèrons des méthodes itératives, c'est-à-dire que nous construisons des suites $(x_n)_{n\in\mathbb{N}}$ de manière à ce que

$$x_n \to x^* \quad \text{quand } n \to \infty.$$

Dans ce chapitre, nous présentons les deux classes de méthodes les plus utilisées en pratique pour résoudre le problème (1.1) : les méthodes de point fixe et les méthodes de Newton.

1.2 Ordre de convergence

Dans cette partie, nous introduisons la notion d'ordre de convergence pour évaluer la vitesse à laquelle une suite converge. Pour cela, nous avons besoin de définir l'erreur pour une suite (x_n) de $\mathbb{R}^N$ convergeant vers un certain x^*. L'erreur au rang n est donnée par

$$e_n = x_n - x^*.$$

Définition 1.1 (Ordre de convergence)**.** La convergence de la suite (x_n) vers x^* est d'ordre $p \in \mathbb{N}^*$ s'il existe $C > 0$ et $n_0 \in \mathbb{N}$ tel que $\forall\, n \geq n_0$,

$$\|e_{n+1}\| \leq C\|e_n\|^p.$$

Si $p = 1$, il faut que $C < 1$ et on parle de convergence linéaire. Si $p = 2$, on parle de convergence quadratique.

Définition 1.2. On suppose qu'il existe $\beta \in [0, 1]$ tel que

$$\lim_{n\to+\infty} \frac{\|e_{n+1}\|}{\|e_n\|} = \beta.$$

- Si $\beta = 0$, on dit que la convergence vers x^* est sur-linéaire.

- Si $\beta = 1$, on dit que la convergence vers x^* est sous-linéaire.

Remarque 1.3.
1. Si dans cette définition, β appartient à $]0, 1[$, alors la convergence est (exactement) linéaire, c'est-à-dire qu'il existe $0 < C_1 < C_2 < 1$ et $n_0 \in \mathbb{N}$ tels que, pour tout $n \geq n_0$
$$C_1\|e_n\| \leq \|e_{n+1}\| \leq C_2\|e_n\|.$$
La minoration de $\|e_{n+1}\|$ par $C_1\|e_n\|$ montre qu'on ne peut pas avoir mieux que de la convergence linéaire.
2. Toute convergence avec un ordre $p > 1$ correspond à une convergence sur-linéaire.
3. La suite $\left(\frac{1}{n^2}\right)$ a une convergence vers 0 sous-linéaire, la suite $\left(\frac{1}{2^n}\right)$ a une convergence vers 0 linéaire (avec $\beta = 1/2$) et la suite $\left(\frac{1}{n!}\right)$ a une convergence sur-linéaire.

Revenons à la méthode de point fixe. La suite $(x_n)_n$ définie par (1.2) satisfait :

$$\|e_{n+1}\| = \|g(x_n) - g(x^*)\| \leq k\|e_n\|$$

avec $k < 1$. La convergence de cette méthode est donc linéaire.

Proposition 1.4 (Estimation de l'erreur)**.** *Si la suite (x_n) converge linéairement vers x^*, alors, il existe $C > 0$, $\alpha < 1$, $n_0 \in \mathbb{N}$ tel que, pour tout $n \geq n_0$*

$$\|e_n\| \leq C\alpha^n.$$

Si la suite (x_n) converge quadratiquement vers x^*, alors, il existe $C > 0$, $\alpha < 1$, $n_0 \in \mathbb{N}$ tel que, pour tout $n \geq n_0$

$$\|e_n\| \leq C\alpha^{2^n}.$$

Démonstration. Si la suite (x_n) converge linéairement, alors il existe $\alpha < 1$ et $n_0 \in \mathbb{N}$ tel que $\forall\, n \geq n_0$, $\|e_{n+1}\| \leq \alpha\|e_n\|$. Par récurrence, on montre facilement que, pour tout $n \geq n_0$

$$\|e_n\| \leq \alpha^{n-n_0}\|e_{n_0}\|$$

ce qui donne le résultat voulu avec $C = \|e_{n_0}\|\alpha^{-n_0}$.

Si la suite (x_n) converge quadratiquement, alors il existe $C_1 > 0$ et $n_0 \in \mathbb{N}$ tel que $\forall\, n \geq n_0$, $\|e_{n+1}\| \leq C_1\|e_n\|^2$. Par récurrence, on montre alors que, pour tout $n \geq n_0$

$$\|e_n\| \leq C_1^{2^{n-n_0}-1}\|e_{n_0}\|^{2^{n-n_0}} = \frac{1}{C_1}\alpha^{2^n}$$

avec $\alpha = (C_1\|e_{n_0}\|)^{2^{-n_0}}$ qu'on peut toujours supposer plus petit que 1 quitte à prendre n_0 plus grand pour avoir $C_1\|e_{n_0}\| < 1$.

□

1.3 Méthode de point fixe

Définition 1.5. Soit g une fonction. Un point fixe de g est un élément x tel que $g(x) = x$.

Dans cette partie, nous cherchons à résoudre le système non linéaire (1.1), c'est-à-dire trouver $x^* \in \mathbb{R}^N$ tel que $f(x^*) = 0$ en utilisant une méthode de point fixe.

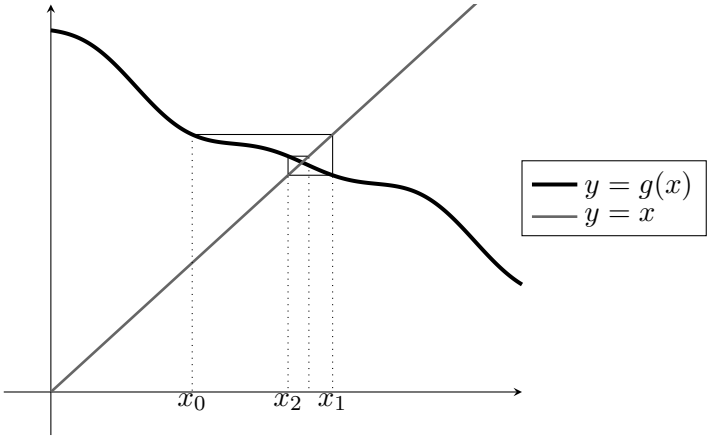

FIGURE 1.1 – En dimension $N = 1$, pour construire graphiquement le terme x_{n+1} à partir du terme x_n, on prend le segment horizontal entre $g(x_n)$ et la droite $y = x$, puis le segment vertical jusqu'à l'axe des abscisses.

Proposition 1.6. *Soit $f : \mathbb{R}^N \to \mathbb{R}^N$ et $g : \mathbb{R}^N \to \mathbb{R}^N$ définie par $g(x) = f(x) + x$. Un point $x^\star$ de $\mathbb{R}^N$ est solution de l'équation $f(x^\star) = 0$ si et seulement si $x^\star$ est un point fixe de g.*

Définissons par récurrence la suite

$$x_{n+1} = g(x_n) \tag{1.2}$$

pour un x_0 donné. Si la suite (x_n) converge, alors, en passant à la limite dans cette expression, on voit que la limite est un point fixe de g. Comme nous allons le voir plus bas, la convergence de la suite (x_n) dépend de la fonction g et du choix de la condition initiale x_0. À titre d'illustration, nous représentons graphiquement, sur la Figure 1.1., la construction d'une telle suite pour une fonction g donnée et pour $N = 1$.

Définition 1.7. Un point fixe $x \in \mathbb{R}^N$ d'une fonction $g : \mathbb{R}^N \to \mathbb{R}^N$ est dit attractif s'il existe un voisinage V de x tel que pour tout x_0 dans V, la suite définie par $x_{n+1} = g(x_n)$ converge vers x. Dans le cas contraire, le point est dit répulsif.

Munissons $\mathbb{R}^N$ d'une norme qu'on note $\|\cdot\|$. Rappelons quelques définitions et résultats sur les espaces vectoriels normés.

Le résultat suivant donne un résultat de convergence pour la suite définie par (1.2).

Théorème 1.8 (Théorème de Picard)**.** *Soit F un fermé de $\mathbb{R}^N$ et soit $g : F \subset \mathbb{R}^N \to \mathbb{R}^N$ une application telle que $g(F) \subset F$. On suppose que g est contractante c'est-à-dire qu'il existe $k \in]0, 1[$ tel que:*

$$\forall x, y \in F, \qquad \|g(x) - g(y)\| \leq k\|x - y\|. \tag{1.3}$$

Alors il existe un unique $x^* \in F \subset \mathbb{R}^N$ tel que $g(x^*) = x^*$ et, pour tout $x_0 \in F$, la suite définie par (1.2) converge vers x^* (c'est-à-dire que $x^\star$ est un point fixe attractif).

De plus, il existe une constante C (dépendant du choix de x_0 et de la fonction g) telle que

$$\|x_n - x^*\| \leq Ck^n. \tag{1.4}$$

Démonstration. On va procéder en trois étapes :

1. montrer que si le point fixe existe, il est unique ;
2. montrer que la suite (x_n) est une suite de Cauchy ;

3. montrer que la limite de la suite (x_n) est un point fixe de g.

Etape 1: Soient x_1^* et x_2^* deux points fixes de g, alors en utilisant (1.3), on obtient

$$\|x_1^* - x_2^*\| \leq k\|x_1^* - x_2^*\|.$$

Cette dernière inégalité n'est vérifiée que si $\|x_1^* - x_2^*\| = 0$.

Etape 2: On a, pour tout $n \in \mathbb{N}^*$,

$$\|x_{n+1} - x_n\| = \|g(x_n) - g(x_{n-1})\| \leq k\|x_n - x_{n-1}\|,$$

donc on obtient par récurrence que pour tout $n \in \mathbb{N}$,

$$\|x_{n+1} - x_n\| \leq k^n\|x_1 - x_0\|.$$

Ceci donne en particulier que, pour tout $n \in \mathbb{N}$, pour tout $p \in \mathbb{N}^*$

$$\begin{aligned}
\|x_{n+p} - x_n\| &\leq \sum_{q=n}^{n+p-1} \|x_{q+1} - x_q\| \\
&\leq k^n \left(\sum_{q=0}^{p-1} k^q \right) \|x_1 - x_0\| \\
&= k^n \frac{1-k^p}{1-k} \|x_1 - x_0\| \\
&\leq \frac{k^n}{1-k} \|x_1 - x_0\|. \qquad (1.5)
\end{aligned}$$

Comme la suite $n \mapsto k^n$ tend vers 0, on en déduit que la suite (x_n) est de Cauchy. Comme $\mathbb{R}^N$ est un espace vectoriel de dimension fini, il est complet. De plus, F est également complet en tant que sous-espace fermé d'un espace complet donc la suite (x_n) converge vers un certain $x^* \in F \subset \mathbb{R}^N$.

Etape 3: En passant à la limite $n \to \infty$ dans (1.2) (ce qui est possible car g est continue), on obtient $x^* = g(x^*)$, donc x^* est un point fixe de g.

Enfin, pour montrer (1.4), si on laisse tendre p vers $+\infty$ dans (1.5), on obtient

$$\|x^* - x_n\| \leq \frac{k^n}{1-k} \|g(x_0) - x_0\|.$$

La constante C dans l'inégalité (1.4) est donnée par $C = \frac{1}{1-k}\|g(x_0) - x_0\|$, et ne dépend que de g (k ne dépend que de g) et de x_0. □

Remarque 1.9. On peut également obtenir une estimation de type (1.4) en utilisant les notions définies dans la section précédente. Pour $n \in \mathbb{N}$, on a, d'après la définition 1.1,

$$e_{n+1} = \|g(x_n) - x^\star\| = \|g(x_n) - g(x^\star)\| \leq k\|x_n - x^\star\| \leq k\|e_n\|,$$

ce qui signifie que la convergence de la méthode de point fixe est linéaire car $0 < k < 1$. Par suite, d'après la proposition 1.4 sur l'estimation d'erreur, on sait qu'il existe $C > 0$ telle que, pour tout $n \in \mathbb{N}$

$$\|e_n\| \leq Ck^n.$$

La différence avec l'estimation obtenue dans la preuve du théorème précédent est que nous n'avons pas de formule explicite pour cette constante C.

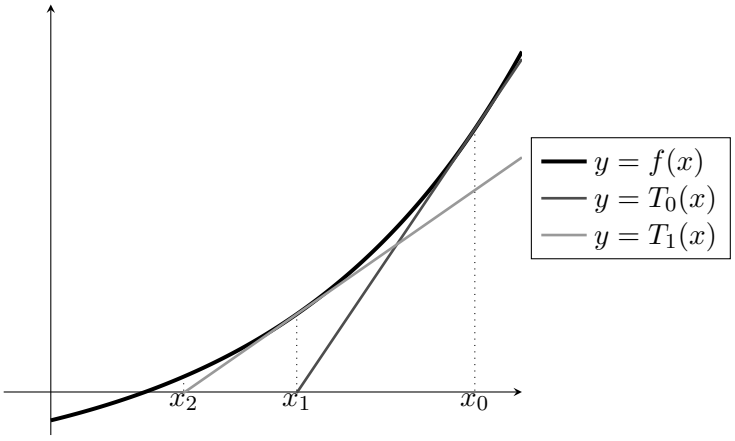

FIGURE 1.2 – En dimension $N = 1$, x_{n+1} est obtenu en considérant l'intersection de la tangente au graphe au point $(x_n, f(x_n))$, notée T_n, avec l'axe des abscisses. Graphiquement, le point x_3 est confondu avec la solution x^*. Comme nous allons le voir (Théorème 1.14), la méthode converge très rapidement.

Remarque 1.10. Ce résultat est vrai de façon plus générale dans un espace de Banach (on rappelle qu'un espace de Banach est un espace vectoriel normé complet, c'est-à-dire tel que toute suite de Cauchy est convergente et que tout espace vectoriel normé de dimension finie est un espaces de Banach).

1.4 Méthode de Newton

Soit $f \in \mathcal{C}^1(\mathbb{R}^N; \mathbb{R}^N)$. Pour tout point $x \in \mathbb{R}^{\mathbb{N}}$, on note $J_f(x)$ la matrice jacobienne de f au point x, définie par

$$J_f(x)_{ij} = \frac{\partial f_i}{\partial x_j}, \quad \forall i, j \in \{1, \cdots, N\}.$$

De plus, on suppose qu'il existe $x^* \in \mathbb{R}^N$ tel que

$$f(x^*) = 0, \qquad J_f(x^*) \in \mathcal{M}_N(\mathbb{R}) \text{ est inversible.} \tag{1.6}$$

Dans ce cas, la méthode de Newton est une méthode particulièrement efficace pour approcher la solution x^*.

1.4.1 Méthode de Newton en dimension 1

En dimension $N = 1$, la méthode de Newton consiste à construire la suite par la relation de récurrence suivante : pour x_0 donné, on a

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, \quad n \geq 0. \tag{1.7}$$

Dans la Figure 1.2, nous donnons une interprétation de la construction de la suite définie ci-dessus : x_{n+1} est l'unique solution de l'équation

$$T_n(x) = 0$$

où l'application affine $T_n : x \mapsto f(x_n) + f'(x_n)(x - x_n)$ est la tangente à f au point x_n. La suite définie par la méthode de Newton converge de façon quadratique vers x^* à condition de partir de x_0 suffisamment proche de x^* (c'est ce qu'on appelle la convergence locale). Le résultat précis est énoncé dans le Théorème 1.14 en dimension quelconque.

Remarque 1.11. Si $f'(x^*) = 0$ et $f''(x^*) \neq 0$, on aura tout de même convergence locale vers x^* mais uniquement linéairement (ce résultat est démontré dans la feuille de TD 1).

1.4.2 Méthode de Newton en dimension quelconque

On définit la méthode de Newton par l'algorithme suivant:

$$\begin{cases} x_0 \in \mathbb{R}^N \\ x_{n+1} = x_n - (J_f(x_n))^{-1} f(x_n). \end{cases} \tag{1.8}$$

On voit dans cette définition qu'il faut que $J_f(x_n)$ soit inversible pour pouvoir définir x_{n+1}. Le lemme qui suit affirme que $J_f(x)$ est inversible pour x proche de x^* :

Lemme 1.12. *Soit $f \in \mathcal{C}^1(\mathbb{R}^N;\mathbb{R}^N)$, et soit $x^* \in \mathbb{R}^N$ tel que $J_f(x^*)$ soit inversible. Alors il existe $\alpha > 0$ tel que pour tout $x \in B(x^*,\alpha)$, $J_f(x)$ est inversible. De plus, il existe C ne dépendant que de α et de f tel que*

$$\|(J_f(x))^{-1}\| \leq C, \, \forall \, x \in B(x^*,\alpha).$$

Démonstration. On a la formule suivante :

$$A^{-1} = \frac{1}{\det A} \left(\mathrm{com}(A)\right)^t.$$

Comme $J_f(x^*)$ est inversible, $\det(J_f(x^*)) \neq 0$. La fonction f est de classe $\mathcal{C}^1$ donc J_f est une fonction continue. Cela assure que $x \mapsto \det(J_f(x))$ est continue et donc, il existe $\alpha > 0$ tel que $|\det(J_f(x))| \geq \dfrac{|\det(J_f(x^*))|}{2}$, pour tout $x \in \overline{B(x^*,\alpha)}$. La continuité de J_f assure aussi que $x \mapsto \mathrm{com}(J_f(x))$ est continue donc bornée sur le compact $\overline{B(x^*,\alpha)}$.

On en déduit ainsi que $x \mapsto (J_f(x))^{-1}$ est bornée sur $\overline{B(x^*,\alpha)}$ et donc aussi sur $B(x^*,\alpha)$. □

Lemme 1.13. *Si $f \in \mathcal{C}^2(\mathbb{R}^N;\mathbb{R}^P)$, alors*

$$\|f(x+h) - f(x) - Df(x)h\| \leq \frac{1}{2} \sup_{z\in[x,x+h]} \|D^2 f(z)\| \|h\|^2.$$

Démonstration. Soit $x, h \in \mathbb{R}^N$. On considère les fonctions $g_i(t) = f_i(x+th)$ avec $i = 1,\ldots,P$. D'après la formule de Taylor-Lagrange [1] pour tout $|t| \leq 1$ il existe $|\zeta_i| < |t|$ telle que

$$\begin{aligned} g_i(t) &= g_i(0) + g_i'(0)t + g_i''(\zeta_i)\frac{t^2}{2} \\ f_i(x+th) &= f_i(x) + t\langle \nabla f_i(x), h\rangle + \frac{t^2}{2}\langle Hf_i(x+\zeta_i h)h, h\rangle \end{aligned}$$

D'où on obtient l'inégalité

$$f_i(x+h) - f_i(x) + \langle \nabla f_i(x), h\rangle\| \leq \frac{1}{2} \sup_{y\in[x,x+h]} \|Hf_i(y)\|_{\mathcal{L}(\mathbb{R}^N;\mathbb{R}^N)} \|h\|^2,$$

puis en passant au max sur $i = 1,\ldots P$

$$\|f(x+h) - f(x) + Df(x)h\| \leq \frac{1}{2} \sup_{y\in[x,x+h]} \|D^2 f(y)\|_{\mathcal{L}(\mathbb{R}^N;\mathbb{R}^N\times\mathbb{R}^P)} \|h\|^2.$$

1. Attention la formule $f(x+h) = f(x) + f'(x)h + f''(\zeta)\dfrac{h^2}{2}$ n'est valable que pour $f : \mathbb{R} \to \mathbb{R}$.

On notera ici, qu'il faut comprendre le segment $[x, x+h]$ comme le sous-ensemble de $\mathbb{R}^N$ défini par $\{y \in \mathbb{R}^N; \exists \lambda \in [0,1], y = (1-\lambda)x + \lambda(x+h)\}$. □

On énonce maintenant le théorème de convergence de la méthode de Newton :

Théorème 1.14. *Soit $f \in \mathcal{C}^2(\mathbb{R}^N; \mathbb{R}^N)$ et soit x^* qui satisfait (1.6), alors il existe $\eta > 0$ tel que, si $x_0 \in B(x^*, \eta)$, la méthode de Newton décrite par (1.8) converge vers x^*. De plus, il existe $\beta > 0$ tel que, pour n suffisamment grand,*

$$\|x_{n+1} - x^*\| \leq \beta \|x_n - x^*\|^2. \tag{1.9}$$

Ainsi, la suite $(x_n)_n$ converge de manière quadratique vers x^*.

Démonstration. Tout d'abord, d'après le Lemme 1.12, il existe $\alpha > 0$ et $C_1 > 0$ tels que, pour tout $x \in B(x^*, \alpha)$, $J_f(x)$ est inversible et

$$\left\| (J_f(x))^{-1} \right\| \leq C_1. \tag{1.10}$$

On se place au rang n et on suppose que x_n est bien défini et que $x_n \in B(x^*, \eta)$ pour un certain $0 < \eta \leq \alpha$ qui sera déterminé plus tard. D'après ce qui précède, $J_f(x_n)$ est inversible et donc x_{n+1} est bien défini. On a

$$x_{n+1} - x^* = x_n - x^* - (J_f(x_n))^{-1} (f(x_n) - f(x^*))$$

car $f(x^*) = 0$. Ainsi

$$\|x_{n+1} - x^*\| \leq \| (J_f(x_n))^{-1} \| \, \|J_f(x_n)(x_n - x^*) - (f(x_n) - f(x^*))\|$$

Or, grâce au Lemme 1.13, il existe $C_2 > 0$ tel que

$$\|f(x^*) - f(x_n) - J_f(x_n)(x^* - x_n)\| \leq C_2 \|x_n - x^*\|^2.$$

En utilisant (1.10) pour $x = x_n$, on en déduit

$$\|x_{n+1} - x^*\| \leq C_1 C_2 \|x_n - x^*\|^2.$$

Pour avoir $\|x_{n+1} - x^*\| < \eta$, on définit η par $\eta = \min\left(\alpha, \frac{1}{C_1 C_2}\right)$.

De cette façon, on obtient par récurrence que, si $x_0 \in B(x^*, \eta)$, alors pour tout $n \in \mathbb{N}$, x_n est bien défini et x_n est dans la boule $B(x^*, \eta)$. Et on a montré (1.9) avec $\beta = C_1 C_2$. □

Remarque 1.15.

1. La convergence quadratique obtenue ici est très rapide : formellement, cela signifie que, au moins pour n grand, lorsque l'on passe de x_n à x_{n+1}, on double le nombre de décimales exactes pour chaque coordonnée $(x_n)_i$.
2. La méthode de Newton est une méthode locale : elle ne converge que sous la condition que x_0 est bien choisi, au sens où il doit être suffisamment proche de la solution x^*. Des méthodes (comme la méthode de dichotomie en dimension 1) permettent de choisir un tel x_0.
3. Cette méthode nécessite le calcul de la matrice $J_f(x_n)$, qui n'est pas forcément facile à obtenir. Certaines variantes de la méthode de Newton permettent d'éviter cela, mais on perd en général en vitesse de convergence. C'est le principe des méthodes présentées dans le paragraphe suivant.

Remarque 1.16. Lorsqu'on approche la solution d'une équation différentielle donnée par

$$x'(t) = f(t, x(t)),\ t \in [0, T]$$

par un schéma implicite, on est amené à résoudre à chaque pas de temps un problème non linéaire. Par exemple, si on considère le schéma d'Euler implicite, la solution exacte est approchée au temps t_{k+1} par x_{k+1} qui est défini implicitement par

$$x_{k+1} = x_k + \Delta t f(t_{k+1}, x_{k+1}).$$

A moins que f ne soit linéaire par rapport à la variable x, on doit résoudre ce système non linéaire en x_{k+1}. On utilise très souvent une méthode de Newton en initialisant les itérations à x_k.

Remarque 1.17. La condition $f \in \mathcal{C}^2(\mathbb{R}^N; \mathbb{R}^N)$ est suffisante mais non nécessaire. Si $f \in \mathcal{C}^1(\mathbb{R}^N; \mathbb{R}^N)$ on peut également montrer la convergence de la méthode de Newton, mais sous des hypothèses plus difficiles à vérifier en pratique.

1.4.3 Variantes de la méthode de Newton

Méthode de la sécante

Cette méthode est utilisée en dimension 1. L'idée est de remplacer dans (1.7) $f'(x_n)$ par le taux d'accroissement

$$\frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}.$$

On définit alors l'algorithme suivant :

$$\begin{cases} x_0, x_1 \neq x_0 \\ x_{n+1} = x_n - \dfrac{f(x_n)(x_n - x_{n-1})}{f(x_n) - f(x_{n-1})}. \end{cases} \tag{1.11}$$

Si on reprend l'interprétation géométrique de la méthode de Newton faite au paragraphe 1.4.1, on voit que la tangente à f au point x_n est remplacée par la droite qui relie les points $(x_{n-1}, f(x_{n-1}))$ et $(x_n, f(x_n))$ qui est donnée par l'équation :

$$y = f(x_n) + \frac{f(x_{n-1}) - f(x_n)}{x_{n-1} - x_n}(x - x_n).$$

L'avantage de cette méthode par rapport à la méthode de Newton est qu'il n'est plus nécessaire de calculer la dérivée, ce qui peut être compliqué d'un point de vue numérique, par exemple quand la fonction f est donnée implicitement. Pour cette méthode, on a un résultat de convergence locale :

Théorème 1.18. *Soit $f \in \mathcal{C}^2(\mathbb{R}; \mathbb{R})$ et $x^* \in \mathbb{R}$ tel que $f(x^*) = 0$ et $f'(x^*) \neq 0$. Il existe $\eta > 0$ tel que, si x_0, $x_1 \in]x^* - \eta, x^* + \eta[$, la méthode de la sécante décrite par (1.11) converge vers x^*. De plus, si $f''(x^*) \neq 0$, il existe $C > 0$, $q \in]0, 1[$, $n_0 \in \mathbb{N}$ tel que, pour tout $n \geq n_0$*

$$\|x_{n+1} - x^*\| \leq C q^{\alpha^n} \tag{1.12}$$

avec $\alpha = \frac{1+\sqrt{5}}{2}$.

Ainsi la convergence est moins rapide que pour la méthode de Newton (la méthode est "d'ordre α").

Méthode de Newton généralisée (ou "quasi Newton")

Le principe de la méthode de Newton généralisée en dimension quelconque est de remplacer $J_f(x_n)$ dans (1.8) par une approximation $A_n \in \mathbb{M}_N(\mathbb{R})$.

On considère ainsi l'algorithme suivant :

$$\begin{cases} x_0 \in \mathbb{R}^N \\ x_{n+1} = x_n - (A_n)^{-1} f(x_n). \end{cases}$$

On montre alors que si la suite de matrice (A_n) est choisie suffisamment proche de $J_f(x^*)$, la suite x_n converge linéairement.

Par exemple, en s'inspirant de la méthode de la sécante en dimension 1 ci-dessus, on cherche une matrice A_n qui, x_n et x_{n-1} étant connus, vérifie la condition

$$A_n(x_n - x_{n-1}) = f(x_n) - f(x_{n-1}). \tag{1.13}$$

Dans le cas où $N > 1$, la condition (1.13) ne permet pas de déterminer A_n de façon unique. Une façon de faire, appelée méthode de Broyden, consiste à chercher $A_n \in \mathbb{M}_N(\mathbb{R})$ qui vérifie également la condition

$$A_n v = A_{n-1} v, \quad \forall v \in \mathbb{R}^N \text{ tel que } v \cdot (x_n - x_{n-1}) = 0.$$

Sous ces deux conditions, on peut alors construire l'approximation A_n à chaque itération

$$A_n = A_{n-1} + \frac{f(x_n) - f(x_{n-1}) - A_{n-1}(x_n - x_{n-1})}{\|x_n - x_{n-1}\|^2} \otimes (x_n - x_{n-1}).$$

En pratique, on initialise généralement $A_0 = I_N$, la matrice identité de taille N.

Chapitre 2

Optimisation

2.1 Introduction

2.1.1 Problème général

Dans ce chapitre, nous nous intéressons à la résolution numérique de problèmes d'optimisation. Nous considèrons une fonction continue $f : \mathbb{R}^N \to \mathbb{R}$ et une partie K de $\mathbb{R}^N$ et nous cherche alors à résoudre le problème d'optimisation suivant :

$$\inf_{x \in K} f(x).$$

La fonction f est appelée **fonction coût** et l'ensemble K qui contient les contraintes du problème est l'ensemble des **éléments admissibles**. Si $K = \mathbb{R}^N$, on parle d'optimisation sans contraintes. Sinon, on parle d'optimisation sous contraintes.

Un élément $x^* \in K$ tel que

$$f(x^*) = \inf_{x \in K} f(x)$$

est appelé **minimiseur** de f sur K. Nous utiliserons la notation

$$x^* \in \operatorname*{argmin}_K f.$$

Remarque 2.1. Il est possible que le problème de minimisation n'admette pas de minimiseur comme dans les cas suivants :

$$f(x) = x \text{ et } K =]0, +\infty[$$

ou

$$f(x) = e^x \text{ et } K = \mathbb{R}.$$

Dans ce chapitre, nous allons chercher à répondre en particulier aux questions suivantes : l'élément x^* existe-t-il ? Est-il unique ? Comment le caractériser ? Il est en effet indispensable d'étudier ces questions avant de (passer du temps à) chercher à approcher numériquement x^*.

Définition 2.2. On appelle suite minimisante de f dans K une suite $(x_n)_{n \in \mathbb{N}}$ d'éléments de K telle que

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

On peut toujours construire une telle suite minimisante par définition de la borne inférieure. Mais cette suite n'admet pas toujours de limite dans K (comme dans les exemples précédents).

En règle générale, le problème d'optimisation est mal posé et il est nécessaire de faire des hypothèses restrictives sur la fonction f et l'ensemble K. La plupart des résultats qu'on va voir dans ce chapitre se placent dans le cadre de l'optimisation convexe.

2.1.2 Exemples de problèmes d'optimisation

Les problèmes d'optimisation interviennent dans des domaines d'application extrêmement variés. Nous donnons ici quelques exemples pour en donner un rapide aperçu.

Problème de reconstruction de signal

En traitement du signal, les techniques de débruitage consistent à éliminer les perturbations du signal mesuré. Ce problème peut se formuler simplement comme un problème d'optimisation : le signal débruité $x \in \mathbb{R}^N$ doit d'une part être proche du signal mesuré $x_{mes} \in \mathbb{R}^N$ et d'autre part être régulier. Ces deux critères sont pris en compte en considérant une fonctionnelle du type :

$$f(x) = \|x - x_{mes}\|^2 + \lambda\phi(x)$$

où $\lambda > 0$ est un paramètre à déterminer et ϕ est une fonction de régularisation. Un exemple de définition pour la fonction ϕ est le suivant :

$$\phi(x) = \sum_{i=1}^{N-1} |x_{i+1} - x_i|^2.$$

Si ϕ est petit, cela signifie que les variations des coordonnées de x sont faibles donc que le signal est assez régulier.

Etat d'équilibre

En mécanique, un état d'équilibre stable correspond à un minimiseur de l'énergie potentielle.

Apprentissage automatique (machine learning)

Dans les techniques d'apprentissage à partir d'un grand nombre de données, les méthodes de classification consistent à attribuer une classe à chaque donnée. Certaines méthodes de classification linéaire (régression logistique, Support Vector Machine) qui consistent à séparer les données par des hyperplans reposent sur des techniques d'optimisation.

Calage de paramètres dans un modèle

Lorsqu'on modélise un problème par des équations, celles-ci font intervenir des paramètres dont la valeur numérique n'est pas connue. Des mesures du système permettent de déterminer leurs valeurs en résolvant un problème d'optimisation : on cherche les paramètres qui permettent de minimiser l'écart entre la solution donnée par le modèle et les mesures.

2.2 Analyse mathématique

2.2.1 Premiers résultats

Définition 2.3. Soit $f : F \mapsto \mathbb{R}$ une fonction continue sur F un ensemble fermé non borné de $\mathbb{R}^N$. On dit que f est coercive si

$$f(x) \to +\infty \text{ quand } \|x\| \to +\infty.$$

Théorème 2.4. *Soit $f : F \mapsto \mathbb{R}$ une fonction continue sur F un ensemble fermé non vide de $\mathbb{R}^N$. Si F n'est pas borné, on suppose de plus que f est coercive. Alors f admet un minimiseur sur F.*

Démonstration. Une fonction continue atteint ses bornes sur un ensemble compact, donc si le fermé F est borné, f admet un minimum sur F. Si F n'est pas borné, on prend un élément x_0 arbitraire de F et on pose $M = f(x_0) + 1$. Comme $f(x) \to +\infty$ lorsque $\|x\| \to +\infty$, il existe $A > 0$ tel que, pour tout $x \in F$

$$\|x\| > A \Rightarrow f(x) > M.$$

Ainsi, comme $\{x \in F / \|x\| \leq A\}$ est un compact de $\mathbb{R}^N$, f atteint sa borne inférieure sur cet ensemble et c'est le minimum sur F tout entier.

□

2.2.2 Caractérisation d'un minimiseur local dans le cas sans contraintes

Définition 2.5. On dit que $x^* \in K$ est un minimiseur local de f s'il existe $\varepsilon > 0$ tel que, pour tout $x \in K$ satisfaisant $\|x - x^*\| < \varepsilon$, on a

$$f(x^*) \leq f(x).$$

Si l'inégalité est stricte dès que $x \neq x^*$, on parle de minimiseur local strict.

Théorème 2.6. *Soit f une fonction différentiable de $\mathbb{R}^N$ dans $\mathbb{R}$. Si x^* est un minimiseur local de f sur $\mathbb{R}^N$, alors*

$$\nabla f(x^*) = 0. \tag{2.1}$$

Si on suppose de plus que f est de classe $\mathcal{C}^2$, alors la matrice hessienne en x^* (qui est symétrique) est positive, c'est-à-dire

$$h \cdot (H_f(x^*)h) \geq 0, \, \forall \, h \in \mathbb{R}^N.$$

Démonstration. Comme x^* est un minimiseur local de f, il existe $\epsilon_0 > 0$ tel que, $\forall \, x \in B(x^*, \epsilon_0)$, $f(x) \geq f(x^*)$. Ainsi, si on note $(e_i)_{1 \leq i \leq N}$ la base canonique de $\mathbb{R}^N$, on a, pour tout $1 \leq i \leq N$, pour tout $|\epsilon| < \epsilon_0$,

$$f(x^* + \epsilon e_i) \geq f(x^*).$$

On en déduit que, pour tout $0 < \epsilon < \epsilon_0$,

$$\frac{f(x^* + \epsilon e_i) - f(x^*)}{\epsilon} \geq 0.$$

En passant à la limite quand ϵ tend vers 0, cela implique que $\frac{\partial f}{\partial x_i}(x^*) \geq 0$. De même, pour tout $-\epsilon_0 < \epsilon < 0$,

$$\frac{f(x^* + \epsilon e_i) - f(x^*)}{\epsilon} \leq 0$$

ce qui implique, en passant à la limite quand ϵ tend vers 0, $\frac{\partial f}{\partial x_i}(x^*) \leq 0$.
On en déduit donc que $\frac{\partial f}{\partial x_i}(x^*) = 0$ pour tout $1 \leq i \leq N$.

Démontrons maintenant la deuxième partie de ce théorème. Soit $h \in \mathbb{R}^N$. En faisant un développement de Taylor à l'ordre 2 au point x^*, on a, pour tout $\epsilon > 0$

$$f(x^* + \epsilon h) = f(x^*) + \epsilon^2 h \cdot (H_f(x^*)h) + o(\epsilon^2 \|h\|^2). \tag{2.2}$$

Or, d'après la première partie de la preuve et en choisissant ϵ tel que $0 < \epsilon \|h\| < \epsilon_0$ on obtient $f(x^* + \epsilon h) \geq f(x^*)$, ce qui implique alors

$$\epsilon^2 h \cdot (H_f(x^*)h) + o(\epsilon^2) \geq 0,$$

c'est-à-dire $h \cdot (H_f(x^*)h) \geq 0$. □

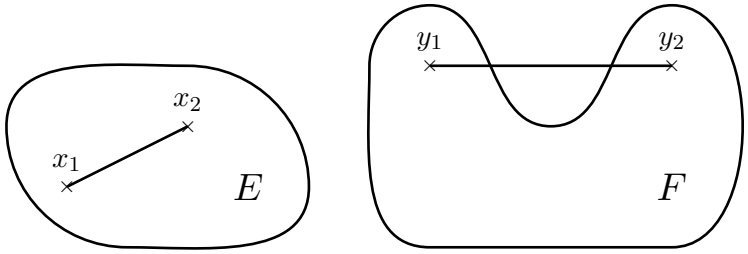

Figure 2.1 – Un ensemble E est convexe si pour tout couple de point $(x_1, x_2) \in E^2$, le segment $[x_1, x_2]$ est inclus dans E. En particulier, l'ensemble F représenté sur la figure du bas est non convexe, car il existe un couple de points $(y_1, y_2) \in F^2$ tel que le segment $[y_1, y_2]$ ne soit pas inclus dans F.

Remarque 2.7. D'après ce théorème, un minimiseur fait donc partie des solutions de l'équation (2.1). Cette équation est appelée l'équation d'Euler associée au problème de minimisation de f sur $\mathbb{R}^N$ et les solutions de cette équation sont appelés points critiques de f.

Remarque 2.8. La réciproque de ce théorème est fausse. Par exemple, 0 est un point critique de la fonction $x \mapsto -x^2$ mais ce n'est pas un minimiseur de cette fonction (c'est un maximiseur). De même, la dérivée de la fonction $x \mapsto x^3$ s'annule en 0, pourtant ce n'est un minimiseur local, c'est un point d'inflexion.

Théorème 2.9. *Soit f de classe $\mathcal{C}^2$ sur $\mathbb{R}^N$. On suppose que $\nabla f(x^*) = 0$ et que $H_f(x^*)$ est définie positive, c'est-à-dire*

$$h \cdot (H_f(x^*)h) > 0, \forall\, h \in \mathbb{R}^N \setminus \{0\}.$$

Alors f admet un minimum local en x^*.

Démonstration. On utilise la formule de Taylor de f à l'ordre 2 au point x^*. □

Remarque 2.10. En changeant f en $-f$ dans les théorèmes précédents, on déduit des résultats analogues (mais avec des signes opposés) lorsque f admet un maximum local en x^*.

Remarque 2.11. Comme on va le voir dans le paragraphe suivant (Proposition 2.17), si on ajoute l'hypothèse que f est convexe, être solution de l'équation d'Euler $\nabla f(x^*) = 0$ est suffisant pour en déduire que x^* est un minimiseur global. La propriété de convexité donne donc un cadre dans lequel la résolution du problème de minimisation se formule simplement.

2.2.3 Cas d'une fonction convexe

Définition 2.12 (ensemble convexe)**.** Un ensemble E de $\mathbb{R}^N$ est dit convexe si pour tout $x_1, x_2 \in E$, et pour tout $t \in [0, 1]$, l'élément $tx_1 + (1-t)x_2 \in E$ (voir Figure 2.1).

Définition 2.13 (fonction convexe)**.** Soit f une fonction continue de $\mathbb{R}^N$ dans $\mathbb{R}$. On dit que la fonction f est *convexe* si, pour tout $x_1, x_2 \in \mathbb{R}^N$, et pour tout $t \in [0, 1]$, on a

$$f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2). \tag{2.3}$$

On dit que la fonction est *strictement convexe* si pour tout $x_1, x_2 \in \mathbb{R}^N$ distincts, et pour tout $t \in]0, 1[$, on a

$$f(tx_1 + (1-t)x_2) < tf(x_1) + (1-t)f(x_2). \tag{2.4}$$

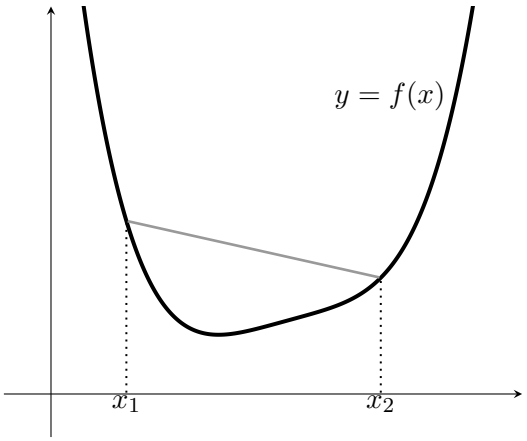

Figure 2.2 – Graphiquement, la signification de l'équation (2.3) est la suivante : le graphe de la fonction f est en dessous de ses cordes.

Proposition 2.14. *Soit $f : \mathbb{R}^N \mapsto \mathbb{R}$ de classe $\mathcal{C}^1$ sur $\mathbb{R}^N$. Alors les assertions suivantes sont équivalentes*

1. *f est convexe ;*
2. *Pour tout x, $y \in \mathbb{R}^N$,*
$$f(y) \geq f(x) + \nabla f(x) \cdot (y - x). \tag{2.5}$$
3. *Pour tout x, $y \in \mathbb{R}^N$,*
$$(\nabla f(x) - \nabla f(y)) \cdot (x - y) \geq 0. \tag{2.6}$$

De plus, f est strictement convexe si et seulement si les inégalités ci-dessus sont strictes dès que $x \neq y$.

Proposition 2.15. *Soit $f : \mathbb{R}^N \mapsto \mathbb{R}$ de classe $\mathcal{C}^2$ sur $\mathbb{R}^N$.*

1. *La fonction f est convexe si et seulement si, pour tout $x \in \mathbb{R}^N$, sa matrice hessienne en x $H_f(x)$ est positive.*
2. *Si, pour tout $x \in \mathbb{R}^N$, la matrice $H_f(x)$ est symétrique définie positive, alors f est strictement convexe.*

Remarque 2.16. La réciproque de la deuxième propriété est fausse. Par exemple, $f(x) = x^4$ est strictement convexe mais f'' s'annule en 0.

Proposition 2.17. *Soit $f : \mathbb{R}^N \mapsto \mathbb{R}$ une fonction convexe de classe $\mathcal{C}^1$ sur $\mathbb{R}^N$. Si on a $\nabla f(x^*) = 0$, alors x^* est un minimiseur global de f.*

Démonstration. Soit $x^* \in \mathbb{R}^N$ tel que $\nabla f(x^*) = 0$. Or, comme f est convexe et de classe $\mathcal{C}^1$ sur $\mathbb{R}^N$, on a, d'après la propriété 2 de la Proposition 2.14, que pour tout $x \in \mathbb{R}^N$

$$f(x) \geq f(x^*) + \nabla f(x^*) \cdot (x - x^*) \geq f(x^*).$$

□

Théorème 2.18 (minimisation d'une fonction strictement convexe)**.** *Soit $f : \mathbb{R}^N \mapsto \mathbb{R}$ une fonction de classe $\mathcal{C}^1$ strictement convexe et coercive. Alors f admet un unique minimiseur. De plus, ce minimiseur $x^* = \operatorname{argmin} f$ est caractérisé par l'équation*

$$\nabla f(x^*) = 0.$$

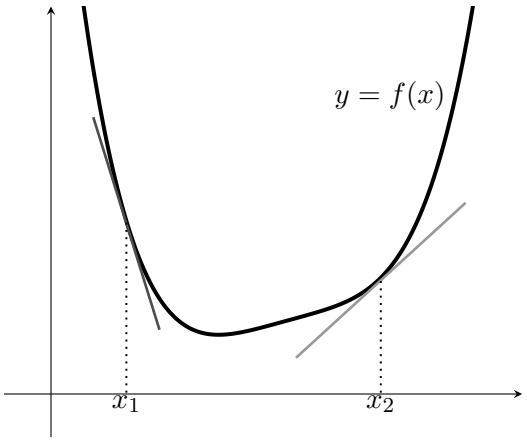

Figure 2.3 – L'équation (2.5) traduit le fait que le graphe d'une fonction convexe f est toujours au-dessus de ses tangentes. En dimension $N = 1$, l'équation (2.6) traduit le fait que la dérivée d'une fonction (strictement) convexe est (strictement) croissante. La Proposition 2.15 traduit le fait que la dérivée seconde d'une fonction convexe est positive. De plus, si celle-ci est strictement positive, alors la fonction est strictement convexe.

Démonstration. Comme f est coercive, d'après le Théorème 2.4, f admet un minimiseur sur $\mathbb{R}^N$. La stricte convexité implique que ce minimiseur est unique. En effet, supposons qu'il existe deux minimiseurs distincts $x_1^* \neq x_2^*$. Alors, par stricte convexité, si on définit $x_3^* = \frac{x_1^*+x_2^*}{2}$, on a

$$f(x_3^*) < \frac{1}{2}(f(x_1^*) + f(x_2^*)) = f(x_1^*),$$

ce qui contredit l'hypothèse que x_1^* est un minimiseur. On a donc un unique minimiseur qu'on note x^*.

Le Théorème 2.6 et la Proposition 2.17 complètent la preuve.

□

Remarque 2.19. L'hypothèse de coercivité est nécessaire. Par exemple, la fonction $x \mapsto e^x$ est strictement convexe, mais n'admet pas de minimiseur sur $\mathbb{R}$.

2.3 Algorithmes de descente

Dans cette partie, nous nous intéressons à un premier type de méthodes numériques pour la résolution de problèmes de minimisation. Comme nous allons le voir, les méthodes de descente (de gradient) reposent sur le calcul du gradient de la fonction à minimiser afin de déterminer une direction de descente. Toutes les méthodes présentées ici correspondent au cas sans contraintes, i.e. lorsque $K = \mathbb{R}^N$.

Définition 2.20 (direction de descente)**.** Soit $f : \mathbb{R}^N \mapsto \mathbb{R}$ une fonction continue, et soit $x \in \mathbb{R}^N$.

1. On dit que $w \in \mathbb{R}^N \setminus \{0\}$ est une direction de descente en x s'il existe $\alpha_0 > 0$ tel que
$$\forall \alpha \in [0, \alpha_0], \qquad f(x + \alpha w) \leq f(x).$$
2. On dit que $w \in \mathbb{R}^N \setminus \{0\}$ est une direction de descente stricte en x s'il existe $\alpha_0 > 0$ tel que
$$\forall \alpha \in]0, \alpha_0], \qquad f(x + \alpha w) < f(x).$$

Le résultat suivant donne une condition suffisante pour avoir une direction de descente.

Proposition 2.21. *Soit $f : \mathbb{R}^N \mapsto \mathbb{R}$ une fonction de classe $\mathcal{C}^1$, et soit un élément $x \in \mathbb{R}^N$ tel que $\nabla f(x) \neq 0$. S'il existe $w \in \mathbb{R}^N$ tel que*

$$\nabla f(x) \cdot w < 0, \tag{2.7}$$

alors w est une direction de descente stricte en x.

Démonstration. On a :

$$f(x + \alpha w) = f(x) + \alpha \nabla f(x) \cdot w + o(\alpha).$$

Donc si la condition (2.7) est satisfaite, $f(x + \alpha w) < f(x)$ pour α assez petit. □

On en déduit immédiatement le résultat suivant.

Proposition 2.22. *Soit $f : \mathbb{R}^N \mapsto \mathbb{R}$ une fonction de classe $\mathcal{C}^1$ et soit $x \in \mathbb{R}^N$ tel que $\nabla f(x) \neq 0$. Alors $(-\nabla f(x))$ est une direction de descente stricte en x.*

Nous pouvons ainsi définir le principe général d'un algorithme de descente.

Définition 2.23 (méthode de descente)**.** Une méthode de descente pour la minimisation d'une fonction $f : \mathbb{R}^N \mapsto \mathbb{R}$ consiste à construire une suite $(x_n)_{n\in\mathbb{N}}$ de la manière suivante:

1. Trouver une direction de descente w_n en x_n.
2. Définir $x_{n+1} = x_n + \alpha_n w_n$, où α_n est "bien choisi".

D'après la Proposition 2.22, un bon choix de de direction de descente est $w_n = -\nabla f(x_n)$ (lorsque f est de classe $\mathcal{C}^1$). Cela amène aux algorithmes présentés dans les deux paragraphes suivants.

2.3.1 Descente de gradient à pas fixe

Soit $f \in \mathcal{C}^1(\mathbb{R}^N; \mathbb{R})$, alors la méthode de descente de gradient à pas fixe (ou constant) s'écrit :

$$\begin{cases} x_0 \in \mathbb{R}^N, \\ x_{n+1} = x_n - \alpha \nabla f(x_n) \end{cases} \tag{2.8}$$

où $\alpha > 0$ est ce qu'on appelle le pas de la méthode.

Cette méthode ne converge pas toujours, même dans le cas où la fonction f est strictement convexe. On a le résultat suivant qui assure la convergence sous certaines conditions sur f et sous la condition que le pas α est suffisamment petit :

Théorème 2.24. *Soit f une fonction de classe $\mathcal{C}^1(\mathbb{R}^N; \mathbb{R})$ telle que*

1. il existe $\beta > 0$ tel que

$$\forall x, y \in \mathbb{R}^N, \qquad (\nabla f(x) - \nabla f(y)) \cdot (x - y) \geq \beta \|x - y\|^2;$$

2. il existe $M > 0$ tel que

$$\forall x, y \in \mathbb{R}^N, \qquad \|\nabla f(x) - \nabla f(y)\| \leq M \|x - y\|.$$

Alors, la suite $(x_n)_n$ donnée par (2.8) converge vers l'unique minimiseur $x^* = \operatorname{argmin}_{\mathbb{R}^N} f$ sous la condittion $0 < \alpha < \frac{2\beta}{M^2}$.

Démonstration. En premier lieu, nous allons montrer que la première hypothèse du théorème implique que f est strictement convexe et coercive.

Soit $x, y \in \mathbb{R}^N$. Nous introduisons la fonction ϕ définie de $[0,1]$ dans $\mathbb{R}$ par $\phi(t) = f(x + t(y-x))$. Alors,

$$f(y) - f(x) = \phi(1) - \phi(0) = \int_0^1 \phi'(t)dt = \int_0^1 \nabla f(x + t(y-x)) \cdot (y-x)dt.$$

On en déduit que

$$f(y) - f(x) - \nabla f(x) \cdot (y-x) = \int_0^1 (\nabla f(x + t(y-x)) - \nabla f(x)) \cdot (y-x)dt.$$

En utilisant la première hypothèse du théorème avec $x = x + t(y-x)$ et $y = x$, et le fait que $t \in [0,1]$, cela entraine

$$\beta t^2 \|x-y\|^2 \leq (\nabla f(x+t(y-x)) - \nabla f(x)) \cdot (t(y-x)) \leq (\nabla f(x+t(y-x)) - \nabla f(x)) \cdot (y-x).$$

En remplaçant dans l'équation précédente, on obtient alors

$$f(y) - f(x) - \nabla f(x) \cdot (y-x) \geq \beta \|y-x\|^2 \int_0^1 t^2 dt = \frac{\beta}{3}\|y-x\|^2 > 0 \quad \text{si } y \neq x, \tag{2.9}$$

ce qui prouve la stricte convexité de f.

Concernant la coercivité de f, écrivons l'inégalité (2.9) pour $x = 0$:

$$f(y) \geq f(0) + \nabla f(0) \cdot y + \frac{\beta}{3}\|y\|^2.$$

Comme $\nabla f(0) \cdot y \geq -\|\nabla f(0)\|\|y\|$ (d'après l'inégalité de Cauchy-Schwartz), on a donc

$$f(y) \geq f(0) + \|y\| \left(\frac{\beta}{3}\|y\| - \|\nabla f(0)\| \right) \to +\infty \quad \text{quand} \quad \|y\| \to +\infty.$$

Ainsi, d'après le Théorème 2.18, f admet un unique minimiseur dans $\mathbb{R}^N$.

Montrons maintenant la convergence de la suite construite par l'algorithme du gradient à pas fixe en nous ramenant à un algorithme de point fixe. On pose $h(x) = x - \alpha \nabla f(x)$. L'algorithme du gradient à pas fixe est alors un algorithme de point fixe pour la fonction h, car

$$x_{n+1} = x_n - \alpha \nabla f(x_n) = h(x_n).$$

On a évidement que $h(\mathbb{R}^N) \subset \mathbb{R}^N$. Il reste donc à montrer que la fonction h est contractante. Pour tout $x, y \in \mathbb{R}^N$

$$\begin{array}{rcl} \|h(x) - h(y)\|^2 & = & \|x - \alpha\nabla f(x) - y + \alpha \nabla f(y)\|^2, \\ & = & (x - y + \alpha(\nabla f(y) - \nabla f(x))) \cdot (x - y + \alpha(\nabla f(y) - \nabla f(x))), \\ & = & \|x-y\|^2 + \alpha^2 \|f(x) - f(y)\|^2 - 2\alpha(x-y) \cdot (f(y) - f(x)), \\ & \leq & \|x-y\|^2 + \alpha^2 M^2 \|x-y\|^2 - 2\beta\alpha\|x-y\|^2, \\ & \leq & (M^2\alpha^2 - 2\beta\alpha + 1)\|x-y\|^2. \end{array}$$

Ainsi, la fonction h est contractante si et seulement si $M^2\alpha^2 - 2\beta\alpha + 1 < 1$, c'est-à-dire si $0 < \alpha < \dfrac{2\beta}{M^2}$.

Nous concluons alors avec le Théorème de Picard 1.8 qui nous assure la convergence linéaire de la suite définie par cet algorithme vers l'unique point fixe de h, noté x^*, qui vérifie

$$x^* - \alpha \nabla f(x^*) = x^*, \quad \text{c'est-à-dire} \quad \nabla f(x^*) = 0.$$

Comme f est convexe, la proposition 2.17 implique alors que x^* est l'unique minimiseur de f. □

Remarque 2.25. Si f est de classe $\mathcal{C}^2$ et s'il existe $0 < \beta \leq M$ tels que

$$\forall x \in \mathbb{R}^N, \qquad \text{spec}\,(H_f(x)) \subset [\beta, M],$$

alors les hypothèses du Théorème 2.24 sont satisfaites.

2.3.2 Descente de gradient à pas optimal

L'idée de cette méthode est de choisir à chaque étape le meilleur pas. À chaque étape de la méthode de descente, nous allons donc déterminer un pas α_n tel que la valeur de la fonction à minimiser soit la plus petite possible à l'étape $n+1$. Soit $f \in \mathcal{C}^1(\mathbb{R}^N;\mathbb{R})$, alors la méthode de descente de gradient à pas optimal s'écrit :

$$\begin{cases} x_0 \in \mathbb{R}^N, \\ x_{n+1} = x_n - \alpha_n \nabla f(x_n), \end{cases} \tag{2.10}$$

où α_n est solution du problème de minimisation unidimensionnel

$$\alpha_n = \underset{\mathbb{R}}{\text{argmin}}\ \left(\alpha \mapsto f\left(x_n - \alpha \nabla f(x_n)\right)\right). \tag{2.11}$$

Remarque 2.26. Si f est strictement convexe et coercive et si $x_n \neq x^*$ (ce qui assure que $\nabla f(x_n) \neq 0$), la fonction

$$\alpha \mapsto f\left(x_n - \alpha \nabla f(x_n)\right) \tag{2.12}$$

est strictement convexe et coercive, donc, d'après le Théorème 2.18, α_n est défini de manière unique.

Remarque 2.27. Puisque α_n minimise (2.12), il annule la dérivée de cette fonction. On a ainsi

$$\nabla f\left(x_n - \alpha_n \nabla f(x_n)\right) \cdot \nabla f(x_n) = 0$$

ce qui donne l'orthogonalité entre les vecteurs $\nabla f(x_{n+1})$ et $\nabla f(x_n)$. La méthode de gradient à pas optimal avance "en zigzag".

Pour la méthode de gradient à pas optimal, on a le résultat de convergence suivant :

Théorème 2.28. *Soit $f \in \mathcal{C}^1(\mathbb{R}^N;\mathbb{R})$ une fonction strictement convexe et coercive. Alors, la suite (x_n) construite par l'algorithme* (2.10)-(2.11) *converge vers $x^* = \underset{\mathbb{R}^N}{\text{argmin}}\ f$.*

2.3.3 Méthode de gradient conjugué : résolution de systèmes linéaires

Cette méthode est une méthode de résolution de système linéaire pour des matrices dans l'ensemble $S_N^{++}(\mathbb{R})$ c'est-à-dire dans l'ensemble des matrices symétriques définies positives. La méthode s'appuie sur le résultat suivant :

Proposition 2.29. *Résoudre le système $Ax = -b$ pour $A \in S_N^{++}(\mathbb{R})$ et $b \in \mathbb{R}^N$ équivaut à déterminer l'unique minimiseur de*

$$f(x) = \frac{1}{2} x \cdot (Ax) + b \cdot x. \tag{2.13}$$

Démonstration. Montrons que $\nabla f(x) = Ax + b$. On peut faire un calcul de dérivées partielles mais il faut faire attention aux indices ! Le plus simple est de remarquer que le gradient de f doit satisfaire, en utilisant la définition A.1 de la différentielle : pour tout $h \in \mathbb{R}^N$,

$$f(x+h) = f(x) + \nabla f(x) \cdot h + o(\|h\|).$$

et de développer l'expression de $f(x+h)$ pour identifier ces trois termes :

$$\begin{aligned} f(x+h) &= \frac{1}{2}(x+h)\cdot(A(x+h)) + b\cdot(x+h) \\ &= \frac{1}{2}x\cdot(Ax) + \frac{1}{2}h\cdot(Ax) + \frac{1}{2}x\cdot(Ah) + \frac{1}{2}h\cdot(Ah) + b\cdot x + b\cdot h \\ &= \frac{1}{2}x\cdot(Ax) + h\cdot(Ax) + \frac{1}{2}h\cdot(Ah) + b\cdot x + b\cdot h \\ &= f(x) + (Ax+b)\cdot h + \frac{1}{2}h\cdot(Ah). \end{aligned}$$

On voit que $\frac{1}{2}(Ah)\cdot h = o(\|h\|)$ et donc $\nabla f(x)\cdot h = (Ax+b)\cdot h$, pour tout $h \in \mathbb{R}^N$. Cela implique donc que $\nabla f(x) = Ax + b$.
Ainsi, pour tout x, $y \in \mathbb{R}^N$

$$(\nabla f(x) - \nabla f(y))\cdot(x-y) = A(x-y)\cdot(x-y) > 0$$

dès que $x \neq y$, car la matrice A est symétrique définie positive. Cette inégalité stricte implique, d'après la Proposition 2.14, que la fonction f est strictement convexe. D'autre part, comme A est définie positive, il existe $\alpha > 0$ tel que,

$$\forall\, x \in \mathbb{R}^N,\ (Ax)\cdot x \geq \alpha\|x\|^2.$$

Ainsi, en utilisant l'inégalité de Cauchy-Schwarz, il vient

$$f(x) \geq \left(\frac{\alpha}{2}\|x\| - \|b\|\right)\|x\| \quad \to +\infty \quad \text{quand} \quad \|x\| \to +\infty,$$

ce qui prouve que f est coercive. D'après le Théorème 2.18, f admet donc un unique minimiseur dans $\mathbb{R}^N$, noté x^*, caractérisé par la propriété $\nabla f(x) = 0$. Dans notre cas, ce minimiseur vérifie également $Ax^* + b = 0$. □

On va ainsi appliquer une méthode de minimisation à la fonction (2.13) pour déterminer une solution du système $Ax = -b$. Dans la méthode du gradient conjugué qu'on présente ici, comme pour la méthode de gradient à pas optimal, à partir d'une direction de descente w_n, on choisit le pas optimal c'est-à-dire le pas α_n qui minimise la fonction

$$T : \alpha \mapsto f\left(x_n + \alpha w_n\right). \tag{2.14}$$

Comme nous l'avons vu précédemment, la fonction f est strictement convexe et coercive. Cela implique alors que la fonction T, définie de $\mathbb{R}$ dans $\mathbb{R}$, est également strictement convexe et coercive. D'après le Théorème 2.18, la fonction T admet donc un unique minimiseur caractérisé par l'équation $T'(\alpha) = 0$. Pour calculer la dérivée de T, développons l'expression $T(\alpha+h)$ pour tout $\alpha, h \in \mathbb{R}$, afin d'identifier sa différentielle (et donc sa dérivée).

$$\begin{aligned} T(\alpha+h) &= f(x_n + \alpha w_n + h w_n), \\ &= f(x_n + \alpha w_n) + h w_n\cdot(A(x_n + \alpha w_n)) + hb\cdot w_n + o(h). \end{aligned}$$

On remarque alors que, $DT(\alpha)h = hw_n\cdot(A(x_n+\alpha w_n)) + hb\cdot w_n$ pour tout $h \in \mathbb{R}$, ce qui implique que $T'(\alpha) = DT(\alpha) = w_n\cdot(A(x_n+\alpha w_n)) + b\cdot w_n$. Il s'agit maintenant de résoudre l'équation $T'(\alpha) = 0$, dont la solution, α_n, est donnée par

$$\alpha_n = \frac{r_n\cdot w_n}{w_n\cdot(Aw_n)} \tag{2.15}$$

où

$$r_n = -\nabla f(x_n) = -b - Ax_n$$

est appelé le résidu au rang n.

La différence avec la méthode de gradient à pas optimal réside dans le choix des directions de descente. Comme expliqué dans la Remarque 2.27, dans la méthode de gradient à pas optimal, deux directions de descente successives sont orthogonales entre elles. Ici, on va choisir des directions non nulles (w_n) telle que deux directions successives sont $A-$orthogonales ou $A-$conjuguées c'est-à-dire qu'elles satisfont :

$$w_n \cdot (Aw_{n-1}) = 0 \tag{2.16}$$

ce qui revient à avoir des directions orthogonales pour le produit scalaire associé à A.

A l'étape 0, il est naturel de choisir la direction opposée du gradient :

$$w_0 = -\nabla f(x_0) = r_0.$$

A l'étape $n \geq 1$, nous choisissons la direction de descente w_n comme combinaison linéaire de r_n et de w_{n-1}, de manière à ce que w_n soit orthogonal à w_{n-1} pour le produit scalaire associé à la matrice A. Nous introduisons alors un réel t_n, tel que la direction de descente w_n s'écrit

$$w_n = r_n + t_n w_{n-1}.$$

Il est alors facile de montrer que la contrainte d'orthogonalité $w_n \cdot (Aw_{n-1}) = 0$ impose le choix du paramètre t_n suivant :

$$t_n = -\frac{r_n \cdot (Aw_{n-1})}{w_{n-1} \cdot (Aw_{n-1})}. \tag{2.17}$$

Nous pouvons alors montrer que les éléments de la suite des directions de descente ainsi construits sont deux à deux A-orthogonaux.

Lemme 2.30. *Soit $m \in \mathbb{N}^*$. Supposons que pour tout $0 \leq k \leq m$, $\nabla f(x_k) \neq 0$. Alors, pour tout $0 \leq k \leq m$, $w_k \neq 0$ et w_k est une direction de descente au point x_k. De plus*

$$(Aw_k) \cdot w_l = 0, \quad \textit{pour tout } k \neq l.$$

Par ailleurs, cette méthode a l'avantage de donner la solution exacte en au plus N itérations.

Théorème 2.31. *Si $A \in S_N^{++}(\mathbb{R})$, l'algorithme du gradient conjugué converge en au plus N itérations vers la solution du système $Ax = -b$.*

En résumé, l'algorithme du gradient conjugué est le suivant.

Algorithme (Gradient conjugué)

- Initialisation : $x_0 \in \mathbb{R}^N$ et $w_0 = r_0$
- Pour $n \geq 0$,
  1. On pose
     - $x_{n+1} = x_n + \alpha_n w_n$ où α_n est donné par (2.15)
     - $r_{n+1} = -b - Ax_{n+1}$ (cette quantité est appelée *résidu*)
     - $w_{n+1} = r_{n+1} + t_{n+1} w_n$ où t_{n+1} est donné par (2.17)
  2. Si $\|r_{n+1}\| < \epsilon$, l'algorithme s'arrête (x_{n+1} est proche de x^*).

2.4 Moindres carrés non-linéaires

2.4.1 Présentation du problème

On va ici se placer dans un cadre où on a plus de contraintes que de variables inconnues. Cela revient à considérer une fonction

$$f : \left\{ \begin{array}{rcl} \mathbb{R}^P & \to & \mathbb{R}^Q \\ x = (x_1, \ldots, x_P)^t & \mapsto & (f_1(x), \ldots, f_Q(x))^t \end{array} \right.$$

pour $Q > P$ et à chercher une solution du problème : $f(x) = 0$. Cela n'est pas possible en général même pour des fonctions f très simples : par exemple, si f est une fonction affine injective, en général, on n'a pas de solution qui satisfait l'équation de façon exacte mais on cherche une solution au sens des moindres carrés.

Rappel 2.32. (Moindres carrés linéaires) On cherche à résoudre un système linéaire du type $Ax = b$ pour $b \in \mathbb{R}^Q$ et A une matrice injective de $\mathcal{M}_{Q,P}(\mathbb{R})$ avec $Q > P$. Le problème : trouver $x \in \mathbb{R}^P$ tel que

$$\|Ax - b\| = \min_{y \in \mathbb{R}^P} \|Ay - b\|$$

admet une unique solution donnée par l'équation normale

$$A^t A x = A^t b.$$

On va reprendre le principe de la méthode des moindres carrés dans le cas non linéaire et chercher à résoudre le problème de minimisation :

$$\left\{ \begin{array}{l} \text{Trouver } x^* \in \mathbb{R}^P \text{ tel que} \\ \|f(x^*)\|^2 = \min\limits_{x \in \mathbb{R}^P} \|f(x)\|^2, \end{array} \right. \tag{2.18}$$

où $\|\cdot\|$ désigne la norme issue du produit scalaire canonique sur $\mathbb{R}^Q$ et où f est une fonction de $\mathcal{C}^1(\mathbb{R}^P; \mathbb{R}^Q)$. On suppose que :

$$\forall x \in \mathbb{R}^P, \qquad J_f(x) \in \mathcal{M}_{Q,P}(\mathbb{R}) \text{ est injective.} \tag{2.19}$$

Cela assure au passage que $(J_f(x))^t J_f(x)$ est symétrique définie positive, donc inversible.

On note

$$g : \left\{ \begin{array}{rcl} \mathbb{R}^P & \to & \mathbb{R} \\ x & \mapsto & \|f(x)\|^2 \end{array} \right.$$

On suppose que

$$g \text{ est strictement convexe et coercive.} \tag{2.20}$$

Le problème (2.18) admet alors une unique solution x^* qui est caractérisée par le fait que c'est un point critique de g c'est-à-dire :

$$\nabla g(x^*) = 0. \tag{2.21}$$

Exprimons ∇g en fonction de f. L'application g est la composée $N \circ f$ de f et de N défini par

$$N(y) = \|y\|^2, \, \forall \, y \in \mathbb{R}^Q$$

donc, d'après le Théorème A.8, on a

$$\forall x \in \mathbb{R}^P, \forall h \in \mathbb{R}^P, \qquad \nabla g(x) \cdot h = DN(f(x)) Df(x) h. \tag{2.22}$$

Or, il est facile de vérifier que pour tout $y \in \mathbb{R}^Q$, pour tout $z \in \mathbb{R}^Q$, on a

$$DN(y) z = 2y \cdot z.$$

On déduit alors que (2.22) se réécrit

$$\nabla g(x) \cdot h = 2f(x) \cdot (Df(x)h),$$

ou encore, avec le formalisme matriciel

$$\nabla g(x) \cdot h = 2f(x) \cdot (J_f(x)h),$$

puis en utilisant la définition de la matrice transposée, on obtient

$$\forall x \in \mathbb{R}^P, \forall h \in \mathbb{R}^P, \qquad \nabla g(x) \cdot h = 2(J_f(x)^t f(x)) \cdot h.$$

Comme cette relation est vraie pour tout h, on a nécessairement

$$\nabla g(x) = 2\left(J_f(x)\right)^t f(x). \tag{2.23}$$

La relation (2.21) devient donc

$$\left(J_f(x^*)\right)^t f(x^*) = 0. \tag{2.24}$$

On est donc ramené à chercher les zéros de la fonction $(J_f)^t f$ de $\mathbb{R}^P$ dans $\mathbb{R}^P$. On pourrait déterminer une solution de ce problème par la méthode de Newton mais cela ferait apparaître la différentielle seconde de la fonction f. Afin d'éviter cela, on introduit l'algorithme de minimisation de Gauss-Newton.

2.4.2 Algorithme de Gauss-Newton

Rappelons que, dans le cas où $P = Q$, la méthode de Newton pour résoudre de façon approchée $f(x) = 0$ consiste à faire le calcul suivant :

$$x_{n+1} = x_n + \delta x_n$$

où δx_n est l'unique solution de $-J_f(x_n)\delta x_n = f(x_n)$.

La méthode de Gauss-Newton va généraliser la méthode de Newton en reprenant cet algorithme mais maintenant on va résoudre le système linéaire $-J_f(x_n)\delta x_n = f(x_n)$ qui est de taille $Q \times P$ au sens des moindres carrés. Ainsi δx_n est donné par :

$$\delta x_n = -\left[\left(J_f(x_n)\right)^t J_f(x_n)\right]^{-1} \left(J_f(x_n)\right)^t f(x_n). \tag{2.25}$$

Cette expression fait apparaître $\left[\left(J_f(x_n)\right)^t J_f(x_n)\right]^{-1} \left(J_f(x_n)\right)^t$ le pseudo-inverse de $J_f(x_n)$. La méthode de Gauss-Newton va ainsi définir une suite de solutions approchées $(x_n)_n$ par

$$\begin{cases} x_0 \in \mathbb{R}^P \\ x_{n+1} = x_n - \left[\left(J_f(x_n)\right)^t J_f(x_n)\right]^{-1} \left(J_f(x_n)\right)^t f(x_n). \end{cases} \tag{2.26}$$

On remarque que grâce à l'hypothèse (2.19), on peut toujours définir x_{n+1} à partir de x_n, et la suite est donc bien définie.

Proposition 2.33. *On suppose (2.19) et (2.20). Soit $x_n \in \mathbb{R}^P$, alors la direction δx_n définie par (2.25) satisfait*

$$\nabla g(x_n) \cdot \delta x_n \leq 0. \tag{2.27}$$

Si $x_n \neq x^*$ alors

$$\nabla g(x_n) \cdot \delta x_n < 0. \tag{2.28}$$

Donc, d'après la Proposition 2.21, δx_n est une direction de descente pour g en x_n.

Démonstration. D'après l'expression (2.23) de ∇g et la définition (2.25) de δx_n, en posant

$$y_n = (J_f(x_n))^t f(x_n),$$

on obtient

$$\nabla g(x_n) \cdot \delta x_n = -2y_n \cdot \left(\left[(J_f(x_n))^t J_f(x_n) \right]^{-1} y_n \right).$$

La matrice $(J_f(x_n))^t J_f(x_n)$ est symétrique définie positive, donc son inverse aussi, on obtient (2.27). Comme x^* est l'unique point critique de g, on sait que, si $x_n \neq x^*$, $\nabla g(x_n) \neq 0$ donc $y_n \neq 0$. Comme la matrice $\left[(J_f(x_n))^t J_f(x_n) \right]^{-1}$ est symétrique définie positive, on obtient (2.28). □

On n'est pas assuré que la méthode de Gauss-Newton converge vers x^*, même si on prend x_0 très proche de x^*. Cependant, la proposition suivante permet d'assurer que si la suite $(x_n)_n$ converge, alors la limite est un point critique de g.

Proposition 2.34. *On suppose que (2.19) et (2.20) sont satisfaites. Si la suite $(x_n)_n$ définie par (2.26) converge, alors sa limite est x^*.*

Démonstration. Soit $\overline{x}$ la limite de la suite (x_n), alors en passant à la limite dans (2.26), on obtient que

$$\left[(J_f(\overline{x}))^t J_f(\overline{x}) \right]^{-1} (J_f(\overline{x}))^t f(\overline{x}) = 0.$$

Comme la matrice $\left[(J_f(\overline{x}))^t J_f(\overline{x}) \right]^{-1}$ est injective, on a

$$(J_f(\overline{x}))^t f(\overline{x}) = \nabla g(\overline{x}) = 0,$$

et $\overline{x}$ est un point critique de g. Par unicité du point critique de g, on a donc $\overline{x} = x^*$. □

2.4.3 Méthode de Gauss-Newton à pas optimal

Dans les cas où la méthode de Gauss-Newton ne converge pas, on peut modifier l'algorithme précédent en définissant la suite (x_n) de la manière suivante. On suppose toujours que (2.19) et (2.20) sont satisfaites et que, à l'étape n, $x_n \neq x^*$.

- Étape 1 : calculer la direction de descente δx_n donnée par (2.25).
- Étape 2 : minimiser g sur la droite passant par x_n et de direction δx_n, ce qui revient à trouver α_n tel que

  $$\alpha_n = \underset{\alpha \in \mathbb{R}}{\operatorname{argmin}} \left(g\left(x_n + \alpha \delta x_n \right) \right).$$

  Comme, d'après la Proposition 2.33, la direction δx_n est une direction de descente stricte et que g est strictement convexe, alors on sait que ce minimum sera atteint pour un unique $\alpha_n > 0$.
- Étape 3 : choisir $x_{n+1} = x_n + \alpha_n \delta x_n$ puis retourner à l'étape 1 si $x_{n+1} \neq x^*$.

On a alors le résultat de convergence suivant.

Théorème 2.35 (convergence de la méthode de Gauss-Newton à pas optimal)**.** *On suppose (2.19) et (2.20). Alors, pour tout x_0 dans $\mathbb{R}^P$, la suite $(x_n)_n$ converge vers x^*.*

2.5 Optimisation sous contraintes

Dans cette partie, on s'intéresse à la minimisation d'une fonction $f \in \mathcal{C}^1(\mathbb{R}^N, \mathbb{R})$ sur un ensemble K strictement inclus dans $\mathbb{R}^N$, c'est-à-dire qu'on cherche $x^* \in K$ tel que $f(x^*) \leq f(x)$ pour tout $x \in K$. La contrainte dans ce problème, est alors de supposer que $x \in K$. Deux cas sont étudiés dans la suite :

- le cas de P contraintes d'égalité : il existe $g \in \mathcal{C}^1(\mathbb{R}^N; \mathbb{R}^P)$ tel que

$$K = \{x \in \mathbb{R}^N |\ g(x) = 0\}.$$

- le cas de P contraintes d'inégalité : il existe $g = (g_1, \ldots, g_P)^t \in \mathcal{C}^1(\mathbb{R}^N; \mathbb{R}^P)$ tel que

$$K = \left\{x \in \mathbb{R}^N |\ g_i(x) \leq 0, \text{ pour tout } 1 \leq i \leq P\right\}.$$

Dans les deux cas, on a le résultat d'existence et d'unicité suivant.

Théorème 2.36. *Existence et unicité Soit $f \in \mathcal{C}(\mathbb{R}^N; \mathbb{R}^N)$ une fonction strictement convexe et K un sous-ensemble convexe fermé de $\mathbb{R}^N$. Si K est borné ou si f est coercive, alors il existe un unique élément x^* de K qui minimise la fonction f, c'est-à-dire tel que*

$$f(x^*) = \min_{x \in K} f(x).$$

2.5.1 Optimisation sous contraintes d'égalités

Théorème 2.37 (théorème des extréma liés)**.** *Soit $f \in \mathcal{C}^1(\mathbb{R}^N; \mathbb{R})$ et P contraintes d'égalités notées $g = (g_1, \ldots, g_P)^t \in \mathcal{C}^1(\mathbb{R}^N; \mathbb{R}^P)$. Notons*

$$K = \left\{x \in \mathbb{R}^N |\ g(x) = 0\right\} = \left\{x \in \mathbb{R}^N |\ g_i(x) = 0, \text{ pour tout } 1 \leq i \leq P\right\}.$$

On suppose que $x^* \in K$ minimise f sur K et que la matrice jacobienne $J_g(x^*) \in \mathcal{M}_{P,N}(\mathbb{R})$ est surjective, c'est-à-dire que la famille $(\nabla g_1(x^*), \ldots, \nabla g_P(x^*))$ est libre, alors il existe P inconnues $\lambda_1, \ldots, \lambda_P \in \mathbb{R}$, appelés multiplicateurs de Lagrange*, tels que*

$$\nabla f(x^*) + \sum_{i=1}^{P} \lambda_i \nabla g_i(x^*) = 0. \tag{2.29}$$

Il est important de noter que, d'après le Théorème 2.37, si $x^* \in K$ est un minimiseur de f, alors la condition (2.29) est forcément vérifiée. En revanche, si la condition (2.29) est vérifiée en un point $\overline{x}$, alors cela n'implique pas que $\overline{x}$ est un minimiseur (de même que, quand $K = \mathbb{R}^N$, si $\nabla f(x^*) = 0$ on n'a pas en général que x^* est un minimiseur). La condition (2.29) est nécessaire, mais n'est pas suffisante en général. Elle sera tout de même suffisante dans certains cas particuliers, comme le montre le résultat qui suit :

Proposition 2.38. *Si g est une application affine de $\mathbb{R}^N$ dans $\mathbb{R}^P$ et si f est strictement convexe et coercive, alors l'unique point vérifiant la relation (2.29) est le point $x^* = \operatorname{argmin}_K f$.*

Pour résoudre (2.29), il faut alors trouver $(x_1^*, \ldots, x_N^*, \lambda_1, \ldots, \lambda_P)^t \in \mathbb{R}^{N+P}$ satisfaisant les $N + P$ équations non-linéaires

$$\begin{cases} \dfrac{\partial f(x^*)}{\partial x_j} + \displaystyle\sum_{i=1}^{P} \lambda_i \frac{\partial g_i(x^*)}{\partial x_j} = 0, & 1 \leq j \leq N \\ g_i(x^*) = 0, & 1 \leq i \leq P. \end{cases}$$

On renvoie donc au Chapitre 1 pour les méthodes de résolution d'un tel problème.

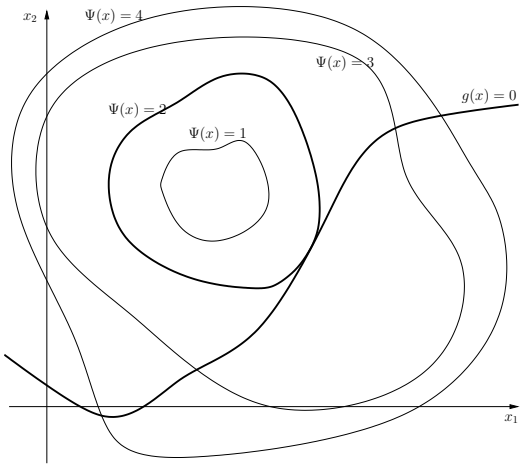

FIGURE 2.4 – Dire que $J_g(x^*)$ est surjective équivaut à dire que $(J_g(x^*))^t$ est injective. La relation (2.29) signifie que $\nabla f(x^*) \in \mathrm{Im}(J_g(x^*))$, ce qui se réécrit $\nabla f(x^*) \in (\mathrm{Ker}(J_g(x^*)))^\perp$. Au niveau géométrique, dans le cas $p = 1$, cela signifie qu'au point x^*, les lignes de niveau de f sont tangentes à l'ensemble $\{g(x) = 0\}$.

2.5.2 Optimisation sous contraintes d'inégalités

Théorème de Kuhn-Tucker

Dans cette partie, l'ensemble K sur lequel on souhaite minimiser la fonction f est défini par

$$K = \{x \in \mathbb{R}^N \mid g_i(x) \leq 0 \text{ pour tout } 1 \leq i \leq P\},$$

où les fonctions g_i sont régulières. On cherche, comme précédemment, à minimiser f dans K c'est-à-dire que l'on cherche $x^* \in K$ tel que

$$x^* = \underset{x \in K}{\mathrm{argmin}} f(x). \tag{2.30}$$

Supposons que x^* existe et que pour tout $i \in \{1, \ldots, P\}$, $g_i(x^*) < 0$. Alors si les applications g_i sont continues, il existe $\eta > 0$ tel que pour tout $x \in B(x^*, \eta)$, on a $g_i(x) < 0$. On a donc que $f(x^*) \leq f(x)$ pour tout $x \in B(x^*, \eta)$. On est alors ramené à un problème de minimisation sans contraintes et si f est différentiable en x^* on a donc $\nabla f(x^*) = 0$.

Le théorème ci-dessous donne un équivalent du théorème des extréma liés 2.37 pour le cas de contraintes inégalités. Il y est stipulé que seuls les multiplicateurs de Lagrange λ_i correspondant aux contraintes saturées, i.e. $g_i(x^*) = 0$, peuvent être non nuls. De plus, ils sont tous positifs (ou nuls).

Théorème 2.39 (théorème de Kuhn-Tucker)**.** *Soit $f \in \mathcal{C}^1(\mathbb{R}^N; \mathbb{R})$, et pour $i \in \{1, \ldots, P\}$, soit $g_i \in \mathcal{C}^1(\mathbb{R}^N; \mathbb{R})$. Soit $x^* \in K$ solution du problème* (2.30)*. On définit l'ensemble des indices $I(x^*) = \{i \in \{1, \ldots, P\} \mid g_i(x^*) = 0\}$ et on suppose que la famille $(\nabla g_i(x^*))_{i \in I(x^*)}$ est libre. Alors, pour $i \in I(x^*)$, il existe $\lambda_i \in \mathbb{R}_+$ tel que*

$$\nabla f(x^*) + \sum_{i \in I(x^*)} \lambda_i \nabla g_i(x^*) = 0.$$

La conclusion du théorème de Kuhn-Tucker peut se réécrire

$$\begin{cases} \dfrac{\partial f(x^*)}{\partial x_j} + \displaystyle\sum_{i=1}^{P} \lambda_i \frac{\partial g_i}{\partial x_j}(x^*) = 0, & 1 \leq j \leq N, \\ \lambda_i g_i(x^*) = 0, & 1 \leq i \leq P, \\ \lambda_i \geq 0, & 1 \leq i \leq P, \\ g_i(x^*) \leq 0, & 1 \leq i \leq P. \end{cases} \tag{2.31}$$

On obtient donc un système de $N + P$ égalités et $2P$ inégalités.

Rappel sur les projections

Définition 2.40 (projection sur un convexe fermé)**.** Soit $K \subset \mathbb{R}^N$ un ensemble convexe fermé, alors, pour tout $x \in \mathbb{R}^N$, il existe un unique élément $p_K(x) \in K$ tel que

$$\forall y \in K, \qquad \|x - p_K(x)\| \leq \|x - y\|. \tag{2.32}$$

Cet élément est appelé projeté (orthogonal) de x sur K. Il est caractérisé par

$$x_0 = p_K(x) \quad \Leftrightarrow \quad \forall y \in K, \quad (x - x_0) \cdot (y - x_0) \leq 0. \tag{2.33}$$

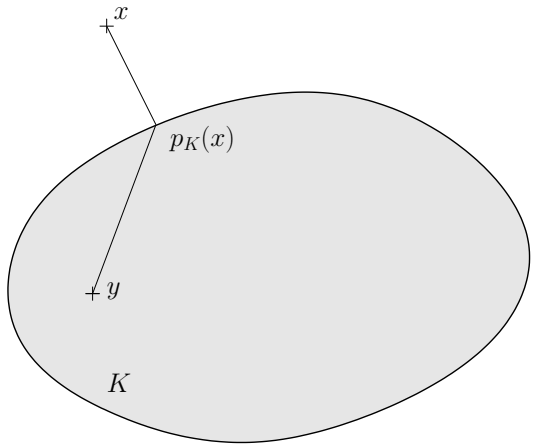

Figure 2.5 – Comme cela est écrit dans la condition (2.32), le point $p_K(x)$ est le point de K qui est le plus proche de x. La condition (2.33) exprime le fait que c'est le seul point de K tel que l'angle en $p_K(x)$ entre les vecteurs $(x - p_K(x))$ et $(y - p_K(x))$ soit obtus pour tout $y \in K$. On remarque que si $x \in K$, alors $p_K(x) = x$.

Proposition 2.41. *Soit K un convexe fermé de $\mathbb{R}^N$, alors l'application*

$$p_K : \left\{ \begin{array}{rcl} \mathbb{R}^N & \to & K \\ x & \mapsto & p_K(x) \end{array} \right.$$

est continue, et vérifie

$$\forall x, y \in K, \qquad \|p_K(x) - p_K(y)\| \leq \|x - y\|. \tag{2.34}$$

On donne un cas important où l'on peut calculer explicitement la projection sur K d'un élément $x \in \mathbb{R}^N$.

Proposition 2.42. *Soit $K = \left\{(x_1, \ldots, x_N)^t \in \mathbb{R}^N / \, x_i \geq 0, \, \forall i \in \{1, \ldots, N\}\right\}$. Alors,*

$$\forall x = (x_1, \ldots, x_N)^t \in \mathbb{R}^N, \qquad p_K(x) = \left(\max(x_1, 0), \ldots, \max(x_N, 0)\right).$$

Algorithme de descente de gradient à pas fixe avec projection

On suppose dans ce paragraphe que l'ensemble $K = \{x \in \mathbb{R}^N \ | g_i(x) \leq 0, \, 1 \leq i \leq P\}$ est convexe. On définit alors l'algorithme de descente de gradient à pas fixe avec projection sur K (GPFK) de la manière suivante :

Algorithme (GPFK)

- Initialisation : $x_0 \in K$.

- Itération : On suppose qu'on connait $x_n \in K$.
  1. On calcule $\nabla f(x_n)$.
  2. On calcule $\tilde{x}_{n+1} = x_n - \alpha \nabla f(x_n)$. Alors on n'a aucune assurance que $\tilde{x}_{n+1} \in K$.
  3. On projette $\tilde{x}_{n+1}$ sur K :
  $$x_{n+1} = p_K(\tilde{x}_{n+1}),$$
  où p_K est l'application définie à la Définition 2.40.

Proposition 2.43. *Supposons que la suite (x_n) converge vers $\overline{x}$ lorsque $n \to \infty$. Alors $\overline{x}$ est solution du problème* (2.30).

Théorème 2.44. *Soit $f \in \mathcal{C}^1(\mathbb{R}^N;\mathbb{R})$ et K un convexe fermé non vide. On suppose que*

a) il existe $\beta > 0$ tel que

$$\forall x, y \in \mathbb{R}^n, \qquad (\nabla f(x) - \nabla f(y)) \cdot (x - y) \geq \beta \|x - y\|^2;$$

b) il existe M tel que

$$\forall x, y \in \mathbb{R}^n, \qquad \|\nabla f(x) - \nabla f(y)\| \leq M \|x - y\|.$$

Alors la suite (x_n) de l'algorithme GPFK converge vers l'unique solution x^* du problème (2.30), *sous l'hypothèse*

$$0 < \alpha < \frac{2\beta}{M^2}.$$

Annexe A

Rappels de calcul différentiel

Dans cette partie, nous notons $\mathcal{L}(\mathbb{R}^N, \mathbb{R}^P)$ l'ensemble des applications linéaires de $\mathbb{R}^N$ dans $\mathbb{R}^P$ et $\mathcal{M}_{P,N}(\mathbb{R})$ l'ensemble des matrices à P lignes et N colonnes. On rappelle qu'une application linéaire de $\mathcal{L}(\mathbb{R}^N, \mathbb{R}^P)$ peut être representée par une matrice de $\mathcal{M}_{P,N}(\mathbb{R})$. On munit $\mathbb{R}^N$ d'une norme notée $\|\cdot\|$.

Définition A.1 (Différentielle d'une application en un point)**.** Soit Ω un ouvert de $\mathbb{R}^N$, et soit $f : \Omega \to \mathbb{R}^P$ une application. On dit que f est *différentiable* en $x \in \Omega$ s'il existe une application linéaire de $\mathcal{L}(\mathbb{R}^N, \mathbb{R}^P)$, notée $Df(x)$, et appelée *différentielle* de f en x, telle que, pour tout $h \in \mathbb{R}^N$ tel que $x + h \in \Omega$,

$$f(x+h) = f(x) + Df(x)h + \|h\|\varepsilon(h), \qquad \text{avec } \varepsilon(h) \to 0 \text{ quand } h \to 0.$$

On note aussi :

$$f(x+h) = f(x) + Df(x)h + o(\|h\|).$$

Il découle de la définition de la différentielle en un point que si une fonction f est différentiable en x, alors elle est continue en x (la réciproque est bien entendu fausse en général). De plus, la différentielle $Df(x)$ de f en x, si elle existe, est unique.

Remarque A.2. Si f est une fonction de $\mathbb{R}$ dans $\mathbb{R}$, la différentielle coïncide avec la dérivée. Plus précisément, l'application $Df(x)$ est donnée par $Df(x)h = f'(x)h$, $\forall\, h \in \mathbb{R}$.

Définition A.3 (dérivée partielle)**.** Soit f une application définie par

$$f : \left\{ \begin{array}{rcl} \mathbb{R}^N & \to & \mathbb{R}^P \\ x = (x_1, \dots x_N)^t & \mapsto & f(x) = (f_1(x_1, \dots, x_N), \dots, f_P(x_1, \dots, x_N))^t \end{array} \right.$$

Pour $1 \leq i \leq P$ et $1 \leq j \leq N$, on définit la dérivée partielle $\frac{\partial f_i}{\partial x_j} : \mathbb{R}^N \to \mathbb{R}$ par

$$\frac{\partial f_i}{\partial x_j}(x) = \lim_{h_j \to 0} \frac{f_i(x_1, \dots, x_j + h_j, \dots, x_N) - f_i(x_1, \dots, x_j, \dots, x_N)}{h_j} \tag{A.1}$$

là où cette limite existe.

Proposition A.4. *Si l'application $f : \mathbb{R}^N \to \mathbb{R}^P$ est différentiable en un point $x \in \mathbb{R}^N$, alors toutes les dérivées partielles $\frac{\partial f_i}{\partial x_j}(x)$, $i \in \{1, \dots, P\}$, $j \in \{1, \dots, N\}$ sont définies, c'est-à-dire que les limites* (A.1) *existent. De plus, la différentielle $Df(x)$ de f en x est caractérisée par*

$$Df(x) : \left\{ \begin{array}{rcl} \mathbb{R}^N & \to \mathbb{R}^P & \\ h = (h_1, \dots, h_N)^t & \mapsto & \left(\sum_{j=1}^N \frac{\partial f_1}{\partial x_j}(x) h_j, \dots, \sum_{j=1}^N \frac{\partial f_P}{\partial x_j}(x) h_j \right)^t. \end{array} \right. \tag{A.2}$$

L'application $Df(x)$ est une application linéaire de $\mathbb{R}^N$ dans $\mathbb{R}^P$. Elle peut donc être représentée par une matrice, appelée matrice *Jacobienne* de f, notée $J_f(x)$, appartenant à $\mathcal{M}_{P,N}(\mathbb{R})$. L'expression (A.2) signifie que

$$J_f(x)_{ij} = \frac{\partial f_i}{\partial x_j}(x). \tag{A.3}$$

Définition A.5 (Application différentielle)**.** Soit Ω un ouvert de $\mathbb{R}^N$ tel que, pour tout $x \in \Omega$, f soit différentiable en x. Alors on appelle *application différentielle* de f la fonction

$$Df : \left\{ \begin{array}{ccc} \Omega & \to & \mathcal{L}(\mathbb{R}^N, \mathbb{R}^P) \\ x & \mapsto & Df(x). \end{array} \right.$$

Lorsque Df est continue, on dit que f est *continument différentiable*, et on note $f \in \mathcal{C}^1(\Omega; \mathbb{R}^P)$.

Proposition A.6. *Soit Ω un ouvert de $\mathbb{R}^N$. Les deux assertions suivantes sont équivalentes :*

1. *les applications $x :\mapsto \frac{\partial f_i}{\partial x_j}(x)$, pour $i \in \{1, \dots, P\}$ et $j \in \{1, \dots, N\}$, sont continues de Ω dans $\mathbb{R}$;*
2. *l'application f est* continument différentiable *de Ω dans $\mathbb{R}^P$.*

Exemple A.7. On considère la fonction suivante :

$$f : \left\{ \begin{array}{ccc} \mathbb{R}^2 & \to & \mathbb{R}^3 \\ x = (x_1, x_2)^t & \mapsto & f(x) = \big(x_1 x_2^2,\, x_1 + x_2,\, \cos(x_1 x_2)\big)^t \end{array} \right.$$

Alors les dérivées partielles de f sont données par

$$\frac{\partial f_1}{\partial x_1}(x) = x_2^2,\ \frac{\partial f_1}{\partial x_2}(x) = 2x_1x_2,\ \frac{\partial f_2}{\partial x_1}(x) = 1,\ \frac{\partial f_2}{\partial x_2}(x) = 1,$$
$$\frac{\partial f_3}{\partial x_1}(x) = -x_2 \sin(x_1x_2),\ \frac{\partial f_3}{\partial x_2}(x) = -x_1 \sin(x_1x_2).$$

D'après la proposition A.6, comme les dérivées partielles de f sont continues sur $\mathbb{R}^2$, $f \in \mathcal{C}^1(\mathbb{R}^2; \mathbb{R}^3)$ et Df est donnée par la proposition A.4 : pour tout $x = (x_1, x_2) \in \mathbb{R}^2$, pour tout $h = (h_1, h_2) \in \mathbb{R}^2$,

$$Df(x)(h) = \Big(x_2^2 h_1 + 2x_1x_2h_2,\, h_1 + h_2,\, -\sin(x_1x_2)(x_2h_1 + x_1h_2)\Big)^t$$

et la matrice jacobienne de f est

$$J_f(x) = \begin{pmatrix} x_2^2 & 2x_1x_2 \\ 1 & 1 \\ -\sin(x_1x_2)x_2 & -\sin(x_1x_2)x_1 \end{pmatrix}.$$

On donne ci-dessous un théorème qui est l'analogue en dimension plus grande que 1 de la formule de la dérivée d'une fonction composée

$$(g \circ f)'(x) = f'(x)g'(f(x)).$$

Théorème A.8 (Différentielle de fonctions composées)**.** *Soit $f \in \mathcal{C}^1(\mathbb{R}^N, \mathbb{R}^P)$ et $g \in \mathcal{C}^1(\mathbb{R}^P, \mathbb{R}^Q)$, alors $g \circ f \in \mathcal{C}^1(\mathbb{R}^N, \mathbb{R}^Q)$ et, pour tout $x \in \mathbb{R}^N$, pour tout $h \in \mathbb{R}^N$,*

$$D(g \circ f)(x)h = Dg(f(x))Df(x)h.$$

Ceci se traduit sur les matrices jacobiennes par

$$J_{g \circ f}(x) = J_g(f(x))J_f(x).$$

L'application $x \mapsto Df(x)$ est à valeurs dans $\mathcal{L}(\mathbb{R}^N, \mathbb{R}^P)$, qui peut être identifié à $\mathcal{M}_{P,N}(\mathbb{R})$. On peut donc étudier la dérivabilité de cette fonction.

Définition A.9 (différentielle seconde)**.** Un fonction f est dite deux fois différentiable en un point x si elle est continuement différentiable sur $B(x, \eta)$ pour un $\eta > 0$, et s'il existe une application, notée $D^2 f(x)$, linéaire de $\mathbb{R}^N$ dans $\mathcal{L}(\mathbb{R}^N, \mathbb{R}^P) \simeq \mathcal{M}_{P,N}(\mathbb{R})$ telle que, pour tout $h \in B(0, \eta)$,

$$Df(x+h) = Df(x) + D^2 f(x)h + o(\|h\|).$$

L'application f est dite de classe $\mathcal{C}^2$ si $x \mapsto D^2 f(x)$ est continue.

Théorème A.10 (Inégalité des accroissements finis)**.** *Soit $f \in \mathcal{C}^1(\mathbb{R}^N; \mathbb{R}^P)$. Pour tout x, $y \in \mathbb{R}^N$, on a*

$$\|f(y) - f(x)\| \leq \sup_{z \in [x,y]} \|Df(z)\| \|x - y\|.$$

Attention, dans ce résultat, on ne différencie pas les différentes normes qui interviennent. On a en fait :

$$\|f(y) - f(x)\|_P \leq \sup_{z \in [x,y]} |||Df(z)||| \, \|x - y\|_N$$

où $\|\cdot\|_N$ est une norme sur $\mathbb{R}^N$, $\|\cdot\|_P$ est une norme sur $\mathbb{R}^P$ et $|||\cdot|||$ est la norme matricielle subordonnée sur $\mathcal{M}_{P,N}(\mathbb{R})$, c'est-à-dire : pour $A \in \mathcal{M}_{P,N}(\mathbb{R})$,

$$|||A||| = \sup_{x \in \mathbb{R}^N \setminus \{0\}} \frac{\|Ax\|_P}{\|x\|_N} = \sup_{x \in \mathbb{R}^N, \|x\|_N = 1} \|Ax\|_P.$$

Pour $N = P = 1$, on a un résultat plus fort donné par le théorème des accroissements finis :

Proposition A.11 (Théorème des accroissements finis)**.** *Soit $f \in \mathcal{C}^1(\mathbb{R}; \mathbb{R})$. Pour tout x, $y \in \mathbb{R}$, il existe $z \in]x, y[$ tel que*

$$f(y) - f(x) = f'(z)(y - x).$$

On considère maintenant le cas particulier où f est à valeurs dans $\mathbb{R}$, c'est-à-dire $P = 1$. Alors la matrice Jacobienne J_f, définie par (A.3) est une matrice ligne, donc la transposée d'un vecteur-colonne de $\mathbb{R}^N$. Ceci justifie la définition du gradient ci-dessous.

Définition A.12 (Gradient d'une application)**.** Soit $f \in \mathcal{C}^1(\mathbb{R}^N, \mathbb{R})$, alors, pour tout $x \in \mathbb{R}^N$, il existe un unique vecteur de $\mathbb{R}^N$, appelé *gradient* de f en x, et noté $\nabla f(x)$, tel que,

$$\forall h = (h_1, \dots h_N)^t \in \mathbb{R}^N, \qquad Df(x)h = J_f(x)h = \nabla f(x) \cdot h.$$

On a alors

$$(\nabla f(x))_i = \frac{\partial f}{\partial x_i}(x).$$

Interprétation On a

$$f(x+h) = f(x) + \nabla f(x) \cdot h + o(\|h\|).$$

Le vecteur $\nabla f(x)$ donne la direction dans laquelle la fonction varie le plus vite, et est orienté des faibles valeurs vers les grandes valeurs. Pour toute constante K, on définit l'ensemble de niveau Γ_K par

$$\Gamma_K = \{x \in \mathbb{R}^N | f(x) = K\}.$$

Soit $x_0 \in \mathbb{R}^N$. Si $x \in \Gamma_{f(x_0)}$, alors

$$0 = \nabla f(x_0) \cdot (x - x_0) + o(\|x - x_0\|)$$

et donc $\nabla f(x_0) \cdot \left(\dfrac{x - x_0}{\|x - x_0\|} \right) \to 0$ lorsque $x \in \Gamma_{f(x_0)}$ tend vers x_0. Cela signifie que le vecteur $\nabla f(x_0)$ est orthogonal à $\Gamma_{f(x_0)}$. On en déduit en particulier que, si f est constante sur un ouvert ω, alors $\nabla f = 0$ sur l'ouvert ω.

Puisque $D^2 f(x)$ est un élément de $\mathcal{L}(\mathbb{R}^N; \mathcal{L}(\mathbb{R}^N; \mathbb{R})) \approx \mathcal{L}(\mathbb{R}^N; \mathbb{R}^N)$, il peut être représenté par une matrice de $\mathcal{M}_N(\mathbb{R})$. Cette matrice est appelée *matrice Hessienne* de f et on la note H_f. Les coefficients de cette matrice sont alors donnés par

$$H_f(x)_{ij} = \frac{\partial}{\partial x_i} \frac{\partial f}{\partial x_j}(x) = \frac{\partial^2 f}{\partial x_i \partial x_j}(x).$$

Théorème A.13 (Théorème de Schwarz)**.** *Soit $f \in \mathcal{C}^2(\mathbb{R}^N; \mathbb{R})$, alors pour tout $i, j \in \{1, \ldots, N\}$, et pour tout $x \in \mathbb{R}^N$, on a*

$$\frac{\partial}{\partial x_i} \frac{\partial f}{\partial x_j}(x) = \frac{\partial}{\partial x_j} \frac{\partial f}{\partial x_i}(x).$$

La matrice Hessienne $H_f(x)$ est donc symétrique.

Exemple A.14. On considère la fonction f de $\mathbb{R}^4$ dans $\mathbb{R}$ définie par

$$f(x_1, x_2, x_3, x_4) = x_1 x_3 + x_4 e^{2x_3}.$$

Alors

$$\nabla f(x) = (x_3,\, 0,\, x_1 + 2x_4 e^{2x_3},\, e^{2x_3})^t$$

et

$$H_f(x) = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 4x_4 e^{2x_3} & 2e^{2x_3} \\ 0 & 0 & 2e^{2x_3} & 0 \end{pmatrix}.$$