Calcul Différentiel et Optimisation

Cours de L2, 2024-2025
Université Paris-Dauphine

David Gontier & Isabelle Catto

(version du 19 mai 2025).

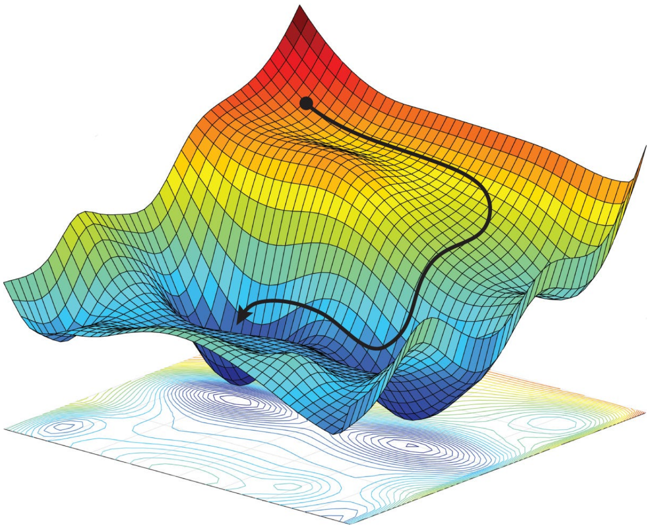

TABLE DES MATIÈRES

INTRODUCTION

Ces notes de cours ont été écrites dans le cadre du cours *Calcul différentiel et optimisation*[1] (L2, Université Paris-Dauphine).

Le but de ce cours est l'étude de fonctions à plusieurs variables, de type $f(x_1, \cdots, x_d)$, et en particulier leur optimisation. Nous définissons la notion de continuité et de différentiabilité pour ces fonctions, et montrons que les optimiseurs de f sont des points critiques. Nous utilisons cette propriété pour trouver les optimiseurs de f.

---

[1] Page web du cours ici.

CHAPITRE 1

TOPOLOGIE DANS LES ESPACES VECTORIELS NORMÉS

1.1 Premiers rappels et définitions

Dans ce chapitre, nous introduisons des concepts de base de topologie, définissons les fonctions continues, et donnons des critères simples d'existence d'optimiseurs pour les fonctions continues.

1.1.1 Rappels sur les espaces vectoriels normés

Pour commencer, on se place dans un cadre général, et on s'intéresse aux fonctions définies de E dans F, où E et F sont des **espaces vectoriels normés**. Dans la suite, on prendra $E = \mathbb{R}^d$ et $F = \mathbb{R}^m$, mais nous préférons énoncer les premières propriétés et définitions dans un cadre plus abstrait.

On rappelle qu'un ensemble E est un $\mathbb{R}$–espace vectoriel s'il est muni :

- d'un **élément neutre**, noté $\mathbf{0} \in E$,
- d'une **addition** de $E \times E$ dans E, notée $+$. Si $\mathbf{x}, \mathbf{y} \in E$ sont des vecteurs, alors $\mathbf{x} + \mathbf{y} \in E$ est encore un vecteur,
- d'une **multiplication par un scalaire** $\mathbb{R} \times E$ dans E, notée "$\cdot$". Si $\lambda \in \mathbb{R}$ et $\mathbf{x} \in E$, alors $\lambda \cdot \mathbf{x}$ (ou $\lambda \mathbf{x}$) est encore un vecteur.

qui vérifient les règles de calculs usuelles (que nous ne rappelons pas).

Définition 1.1 : Normes

Une **norme** sur E est une application $\mathcal{N} : E \to \mathbb{R}_+$ telle que :

- pour tout $\mathbf{x} \in E$, on a $\mathcal{N}(\mathbf{x}) = 0$ si et seulement si $\mathbf{x} = \mathbf{0}$,
- (homogénéité) Pour tout $\mathbf{x} \in E$ et tout $\lambda \in \mathbb{R}$, on a $\mathcal{N}(\lambda \mathbf{x}) = |\lambda| \mathcal{N}(\mathbf{x})$,
- (inégalité triangulaire) $\mathcal{N}(\mathbf{x} + \mathbf{y}) \leq \mathcal{N}(\mathbf{x}) + \mathcal{N}(\mathbf{y})$.

En prenant $\mathbf{z} = \mathbf{x} + \mathbf{y}$ dans l'inégalité triangulaire, on a $\mathcal{N}(\mathbf{z}) \leq \mathcal{N}(\mathbf{x}) + \mathcal{N}(\mathbf{z} - \mathbf{x})$, donc $\mathcal{N}(\mathbf{z}) - \mathcal{N}(\mathbf{x}) \leq \mathcal{N}(\mathbf{z} - \mathbf{x})$. Quitte à changer $\mathbf{x} \longleftrightarrow \mathbf{z}$, on a aussi $\mathcal{N}(\mathbf{x}) - \mathcal{N}(\mathbf{z}) \leq \mathcal{N}(\mathbf{z} - \mathbf{x})$, et donc

$$\forall \mathbf{x}, \mathbf{y} \in E, \qquad |\mathcal{N}(\mathbf{x}) - \mathcal{N}(\mathbf{y})| \leq \mathcal{N}(\mathbf{x} - \mathbf{y}). \qquad \text{(inégalité triangulaire inverse).}$$

Un espace vectoriel E muni d'une norme $\mathcal{N}$ est appelé un **espace vectoriel normé**. On le note $(E, \mathcal{N})$ pour préciser avec quelle norme nous travaillons. Des normes différentes peuvent induire des résultats différents, comme nous le verrons. Un concept important est celui de *normes équivalentes*.

Définition 1.2 : Normes équivalentes

On dit que deux normes $\mathcal{N}_1$ et $\mathcal{N}_2$ sur E sont **équivalentes** s'il existe $\alpha > 0$ telle que

$$\forall \mathbf{x} \in E, \qquad \mathcal{N}_1(\mathbf{x}) \le \alpha \mathcal{N}_2(\mathbf{x}), \quad \text{et} \quad \mathcal{N}_2(\mathbf{x}) \le \alpha \mathcal{N}_1(\mathbf{x}).$$

Proposition 1.3. *Si $\mathcal{N}_1$ et $\mathcal{N}_2$ sont équivalentes, et si $\mathcal{N}_2$ et $\mathcal{N}_3$ sont équivalentes, alors $\mathcal{N}_1$ et $\mathcal{N}_3$ sont équivalentes.*

On dit que *être équivalent* est une *relation d'équivalence.*

Démonstration. Soit $\alpha > 0$ et $\beta > 0$ tel que, pour tout $\mathbf{x} \in E$, on a

$$\mathcal{N}_1(\mathbf{x}) \le \alpha \mathcal{N}_2(\mathbf{x}), \quad \mathcal{N}_2(\mathbf{x}) \le \alpha \mathcal{N}_1(\mathbf{x}), \quad \text{et} \quad \mathcal{N}_2(\mathbf{x}) \le \beta \mathcal{N}_3(\mathbf{x}), \quad \mathcal{N}_3(\mathbf{x}) \le \beta \mathcal{N}_2(\mathbf{x}).$$

Alors, on a, pour tout $\mathbf{x} \in E$,

$$\mathcal{N}_1(\mathbf{x}) \le \alpha \mathcal{N}_2(\mathbf{x}) \le \alpha\beta \mathcal{N}_3(\mathbf{x}), \quad \text{et} \quad \mathcal{N}_3(\mathbf{x}) \le \beta \mathcal{N}_2(\mathbf{x}) \le \beta\alpha \mathcal{N}_1(\mathbf{x}),$$

ce qui montre que $\mathcal{N}_1$ et $\mathcal{N}_3$ sont équivalentes, en prenant la constante $\gamma := \alpha\beta > 0$. □

Exercice 1.4

Montrer que $\mathcal{N}_1$ et $\mathcal{N}_2$ sont deux normes équivalentes ssi il existe $0 < \alpha \le \beta < \infty$ telles que

$$\forall \mathbf{x} \in E, \qquad \beta \mathcal{N}_1(\mathbf{x}) \le \mathcal{N}_2(\mathbf{x}) \le \alpha \mathcal{N}_1(\mathbf{x}).$$

(on note que cette formule n'est pas symétrique. C'est pourquoi nous préférons la version symétrique pour la définition.)

Il est d'usage de noter $\| \cdot \| =: \mathcal{N}$ les normes. Nous utiliserons les deux notations.

1.1.2 Rappels sur les suites

Soit $(E, \| \cdot \|)$ un $\mathbb{R}$-espace vectoriel normé. Une suite de E est une application $\mathbb{N} \ni n \mapsto \mathbf{x}_n \in E$. Dans la suite, on écrira $(\mathbf{x}_n)_{n\in\mathbb{N}} \in E$ ou simplement $(\mathbf{x}_n) \in E$ pour parler d'une suite de E.

Convergence de suites

Définition 1.5 : Convergence

Soit $(\mathbf{x}_n)$ une suite de E, et soit $\mathbf{x}_* \in E$. On dit qu'une suite $(\mathbf{x}_n)$ converge vers $\mathbf{x}_*$ si

$$\lim_{n\to\infty} \| \mathbf{x}_n - \mathbf{x}_* \| = 0.$$

Cela signifie que pour tout $\varepsilon > 0$, il existe $N \in \mathbb{N}$ telle que

$$\forall n \ge N, \quad \| \mathbf{x}_n - \mathbf{x}_* \| < \varepsilon.$$

On écrira parfois simplement $\mathbf{x}_n \to \mathbf{x}_*$ ou $\lim_{n\to\infty} \mathbf{x}_n = \mathbf{x}_*$. On fera attention cependant que ces écritures sont trompeuses, car elles n'indiquent pas avec quelle norme on travaille !

Exercice 1.6

Soit $E := C^0([0,1])$ l'ensemble des fonctions continues de $[0,1]$ dans $\mathbb{R}$. Soit $f_n(x) := x^n$.
On note $\|f\|_1 := \int_0^1 f(s)\mathrm{d}s$, et $\|f\|_\infty := \max\{f(x),\ x \in [0,1]\}$.
a/ Montrer que $\| \cdot \|_1$ et $\| \cdot \|_\infty$ sont des normes sur E.
b/ Montrer que (f_n) converge vers $f_*(x) = 0$ pour la norme $\| \cdot \|_1$, mais pas pour la norme $\| \cdot \|_\infty$.

Cependant, lorsque les normes sont équivalentes, les convergences sont les mêmes, comme le montre le résultat suivant.

Théorème 1.7

Soit $\mathcal{N}_1$ et $\mathcal{N}_2$ deux normes équivalentes sur E. Alors, pour toute suite $(\mathbf{x}_n)$ de E et tout $\mathbf{x}_* \in E$, la suite $(\mathbf{x}_n)$ converge vers $\mathbf{x}_*$ pour la norme $\mathcal{N}_1$ ssi elle converge vers $\mathbf{x}_*$ pour la norme $\mathcal{N}_2$.

Démonstration. Si $(\mathbf{x}_n)$ converge vers $\mathbf{x}_*$ pour la norme $\mathcal{N}_1$, on a $\mathcal{N}_1(\mathbf{x}_n - \mathbf{x}_*) \to 0$. Donc

$$0 \le \mathcal{N}_2(\mathbf{x}_n - \mathbf{x}_*) \le \alpha \mathcal{N}_1(\mathbf{x}_n - \mathbf{x}_*) \xrightarrow[n\to\infty]{} 0,$$

ce qui montre que $(\mathbf{x}_n)$ converge vers $\mathbf{x}_*$ pour la norme $\mathcal{N}_2$. La réciproque est obtenue en échangeant les rôles de $\mathcal{N}_1$ et $\mathcal{N}_2$. ☐

On dit que les normes équivalentes définissent la même **topologie**, car elles induisent les mêmes notions de convergence.

Suites bornées

Définition 1.8 : Suite bornée

On dit qu'une suite $(\mathbf{x}_n)$ de E est bornée s'il existe $M \in \mathbb{R}_+$ tel que, pour tout $n \in \mathbb{N}$, on a $\|\mathbf{x}_n\| \le M$.

On rappelle le résultat classique suivant.

Théorème 1.9 : Les suites convergentes sont bornées

Si $(\mathbf{x}_n)$ est une suite convergente, alors $(\mathbf{x}_n)$ est une suite bornée.

Démonstration. Soit $\mathbf{x}_*$ la limite de la suite. Pour $\varepsilon = 1$, il existe $N \in \mathbb{N}$ tel que $\|\mathbf{x}_n - \mathbf{x}_*\| < 1$ pour tout $n \ge N$. En particulier, pour $n \ge N$, on a $\|\mathbf{x}_n\| \le \|\mathbf{x}_*\| + \|\mathbf{x}_* - \mathbf{x}_n\| \le \|\mathbf{x}_*\| + 1 =: R_1$. Les termes $\mathbf{x}_n$ avec $n \ge N$ sont donc bornés par R_1. Par ailleurs, soit $R_2 := \max_{1 \le i \le N} \|\mathbf{x}_i\|$. C'est le maximum d'un nombre finis de valeurs, donc $R_2 < \infty$. Les N premiers termes de la suite sont bornés par R_2. Au final, tous les termes de la suite sont bornés par $R := \max\{R_1, R_2\}$. ☐

Sous-suites

Définition 1.10 : Sous-suite

Une **extraction** est une fonction $\phi : \mathbb{N} \to \mathbb{N}$ qui est strictement croissante (donc injective).
Une suite $(\mathbf{y}_n)$ est une **sous suite** de $(\mathbf{x}_n)$ si il existe une extraction ϕ telle que $\mathbf{y}_n = \mathbf{x}_{\phi(n)}$.

Exercice 1.11

Soit ϕ une extraction. Montrer que $\phi(n) \ge n$.
Soit ϕ_1 et ϕ_2 deux extractions. Montrer que $\phi_1 \circ \phi_2$ est encore une extraction.

Exemple 1.12. *Considérons la suite* $x_n = (-1)^n$. *Alors en prenant l'extraction* $\phi(n) := 2n$, *on voit que la suite constante* $y_n = x_{2n} = 1$ *est une sous-suite de* x_n. *De même en prenant l'extraction* $\phi(n) := 2n + 1$, *la suite constante* $y_n = -1$ *est une sous-suite de* x_n.

Définition 1.13 : Valeur d'adhérence

Soit $(\mathbf{x}_n)$ une suite de E, et soit $\mathbf{x}_* \in E$. On dit que $\mathbf{x}_*$ est une **valeur d'adhérence** de la suite $(\mathbf{x}_n)$ s'il existe une sous-suite de $(\mathbf{x}_n)$ qui converge de $\mathbf{x}_*$. Autrement dit, il existe une extraction $\phi : \mathbb{N} \to \mathbb{N}$ telle que $\lim_{n\to\infty} \mathbf{x}_{\phi(n)} = \mathbf{x}_*$.

Dans l'exemple précédent, les points -1 et 1 sont des valeurs d'adhérence de la suite $(-1)^n$. Une suite peut avoir une infinité de valeurs d'adhérence, comme le montre l'exercice suivant.

Exercice 1.14

Soit (x_n) une suite de variables aléatoires indépendantes identiquement distribuées selon la loi uniforme de $[0, 1]$. Montrer que pour tout $a \in [0, 1]$, a est une valeur d'adhérence de la suite (x_n), presque sûrement.

Théorème 1.15

Si $(\mathbf{x}_n)$ converge vers $\mathbf{x}_*$, alors toute sous-suite de $(\mathbf{x}_n)$ converge aussi vers $\mathbf{x}_*$.
En particulier, $\mathbf{x}_*$ est la seule valeur d'adhérence de la suite $(\mathbf{x}_n)$.

Démonstration. Soit $\varepsilon > 0$, et soit $N \in \mathbb{N}$ tel que, pour tout $n \geq N$, on a $\|\mathbf{x}_n - \mathbf{x}_*\| < \varepsilon$. Soit ϕ une extraction quelconque, et soit $\mathbf{y}_n = \mathbf{x}_{\phi(n)}$ une sous-suite de $(\mathbf{x}_n)$. Alors, pour tout $n \geq N$, on a $\phi(n) \geq n \geq N$, donc $\|\mathbf{y}_n - \mathbf{x}_*\| = \|\mathbf{x}_{\phi(n)} - \mathbf{x}_*\| < \varepsilon$. □

1.2 Topologie dans les espaces vectoriels normés

Dans la suite, on travaille dans l'espace vectoriel normé $(E, \|\cdot\|)$.

1.2.1 Définitions

Nous commençons avec des définitions.

Définition 1.16 : Boule ouverte, boule fermée

La **boule ouverte** de centre $\mathbf{x}_0 \in E$ et de rayon $r > 0$ est le sous-ensemble de E défini par

$$\mathcal{B}(\mathbf{x}_0, r) := \{\mathbf{x} \in E, \quad \|\mathbf{x} - \mathbf{x}_0\| < r\}.$$

La **boule fermée** de centre $\mathbf{x}_0 \in E$ et de rayon $r > 0$ est le sous-ensemble de E défini par

$$\overline{\mathcal{B}(\mathbf{x}_0, r)} := \{\mathbf{x} \in E, \quad \|\mathbf{x} - \mathbf{x}_0\| \leq r\}.$$

La seule différence entre la boule ouverte et fermée est que la *sphère* $\|\mathbf{x} - \mathbf{x}_0\| = r$ est incluse dans la boule fermée, mais pas dans la boule ouverte. Par exemple, lorsque $E = \mathbb{R}$ avec $\|x\| := |x|$ (valeur absolue), on a

$$\mathcal{B}(0, 1) =]-1, 1[, \qquad \text{et} \qquad \overline{\mathcal{B}(0, 1)} = [-1, 1].$$

Dans ce cours, nous adoptons la convention anglo-saxonne, et notons $(-1, 1)$ pour $]-1, 1[$, ou bien $(-1, 0]$ pour $]-1, 0]$, etc.

Définition 1.17 : Ouverts, Fermés

Un sous–ensemble $A \subset E$ est **ouvert** si, pour tout $\mathbf{x} \in A$, il existe $r > 0$ telle que $\mathcal{B}(\mathbf{x}, r) \subset A$.
Un sous–ensemble $A \subset E$ est **fermé** si son complémentaire est ouvert, *i.e.* si $\mathbb{R}^d \setminus A$ est ouvert.

Par exemple, l'ensemble $\emptyset$ est ouvert (on rappelle que toute proposition commençant par "*pour tout* $\mathbf{x} \in \emptyset$, ..." est vraie). L'ensemble E est ouvert (prendre $r = 1$ par exemple). En prenant les complémentaires, on obtient que E et $\emptyset$ sont aussi des ensembles fermés (en fait, on peut montrer que ce sont les seuls dans E).

Théorème 1.18

Pour tout $\mathbf{x}_0 \in E$ et pour tout $r > 0$, la boule ouverte $\mathcal{B}(\mathbf{x}_0, r)$ est un ensemble ouvert.
Pour tout $\mathbf{x}_0 \in E$ et pour tout $r > 0$, la boule fermée $\overline{\mathcal{B}(\mathbf{x}_0, r)}$ est un ensemble fermé.

Démonstration. Soit $\mathbf{x} \in \mathcal{B}(\mathbf{x}_0, r)$ et soit $r' := \|\mathbf{x} - \mathbf{x}_0\|$. Par définition de $\mathcal{B}(\mathbf{x}_0, r)$, on a $r' < r$. Soit $\varepsilon > 0$ suffisamment petit pour que $r' + \varepsilon < r$. Nous prétendons que $\mathcal{B}(\mathbf{x}, \varepsilon) \subset \mathcal{B}(\mathbf{x}_0, r)$.
En effet, soit $\mathbf{y} \in \mathcal{B}(\mathbf{x}, \varepsilon)$, de sorte que $\|\mathbf{x} - \mathbf{y}\| < \varepsilon$. En utilisant l'inégalité triangulaire, on a

$$\|\mathbf{x}_0 - \mathbf{y}\| \leq \|\mathbf{x}_0 - \mathbf{x}\| + \|\mathbf{x} - \mathbf{y}\| \leq r' + \varepsilon < r.$$

Donc $\mathbf{y} \in \mathcal{B}(\mathbf{x}_0, r)$. Ceci étant vrai pour tout $\mathbf{y} \in \mathcal{B}(\mathbf{x}, \varepsilon)$, on a $\mathcal{B}(\mathbf{x}, \varepsilon) \subset \mathcal{B}(\mathbf{x}_0, r)$.

Pour la deuxième partie, nous remarquons que le complémentaire de $\overline{\mathcal{B}(\mathbf{x}_0, r)}$ est l'ensemble

$$\mathbb{R}^d \setminus \overline{\mathcal{B}(\mathbf{x}_0, r)} = \{\mathbf{x} \in \mathbb{R}^d, \quad \|\mathbf{x} - \mathbf{x}_0\| > r\}.$$

Soit $\mathbf{x} \in \mathbb{R}^d \setminus \overline{\mathcal{B}(\mathbf{x}_0, r)}$. On pose comme avant $r' := \|\mathbf{x} - \mathbf{x}_0\| > r$ et $\varepsilon > 0$ tel que $r' - \varepsilon > r$. On répète les arguments précédent en utilisant cette fois l'inégalité triangulaire inverse :

$$\|\mathbf{x}_0 - \mathbf{y}\| \geq \|\mathbf{x}_0 - \mathbf{x}\| - \|\mathbf{x} - \mathbf{y}\| \geq r' - \varepsilon > r.$$

$\square$

Exercice 1.19

Dans la preuve précédente, montrer qu'on peut prendre $\varepsilon = \frac{1}{2}(r - r') > 0$.
Soit $\mathbf{x} \neq \mathbf{y}$ dans $\mathbb{R}^d$. Monter qu'il existe $\varepsilon > 0$ tel que $\mathcal{B}(\mathbf{x}, \varepsilon) \cap \mathcal{B}(\mathbf{y}, \varepsilon) = \emptyset$.

En pratique, on sera intéressé à des ouverts autour d'un point $\mathbf{x}$ donné. Cette notion est tellement importante qu'on lui donne une définition.

Définition 1.20 : Voisinage de $\mathbf{x}$

On dit que $A \subset E$ est un **voisinage** de $\mathbf{x} \in E$ si $\mathbf{x} \in A$ et si A est un ensemble ouvert.

Exercice 1.21

Montrer que si A est un voisinage de $\mathbf{x}$, alors il existe $r > 0$ tel que $\mathcal{B}(\mathbf{x}, r) \subset A$.

1.2.2 Convergence de suites avec des ouverts/fermés

Grâce à la notion de boule ouverte, on peut traduire la notion de convergence en terme "topologique", de la manière suivante.

Théorème 1.22 : Caractérisation de la limite

Un point $\mathbf{x}_*$ est la limite de la suite $(\mathbf{x}_n)$ ssi pour tout $\varepsilon > 0$, il existe $N \in \mathbb{N}$ tel que pour tout $n \geq N$ on a $\mathbf{x}_n \in \mathcal{B}(\mathbf{x}_*, \varepsilon)$.

Autrement dit, si $\mathbf{x}_*$ est la limite de la suite $\mathbf{x}_n$ (pour la norme $\|\cdot\|$), alors quelque soit le voisinage de $\mathbf{x}_*$, toute la suite est dans ce voisinage à partir d'un certain rang.

Théorème 1.23 : Caractérisation des valeurs d'adhérence

Un point $\mathbf{x}_*$ est valeur d'adhérence de la suite $(\mathbf{x}_n)$ ssi pour tout $\varepsilon > 0$ et pour $N \in \mathbb{N}$, il existe $n \geq N$ tel que $\mathbf{x}_n \in \mathcal{B}(\mathbf{x}_*, \varepsilon)$.

Autrement dit, $\mathbf{x}_*$ est valeur d'adhérence si pour tout voisinage de $\mathbf{x}_*$, il y a une infinité d'éléments de la suite $(\mathbf{x}_n)$ qui appartiennent à ce voisinage.

Démonstration. Supposons que $\mathbf{x}_*$ vérifie la propriété que pour tout $\varepsilon > 0$ et pour tout $N \in \mathbb{N}$, il existe $n \geq N$ tel que $\mathbf{x}_n \in \mathcal{B}(\mathbf{x}_*, \varepsilon)$. Prenons une suite $\varepsilon_n > 0$ de nombres positifs qui tend vers 0 (par exemple $\varepsilon_n = \frac{1}{n}$). On construit l'extraction par induction. On prend $N_0 = 0$, et on considère un entier n_1 tel que $\mathbf{x}_{n_1} \in \mathcal{B}(\mathbf{x}_*, \varepsilon_1)$. On pose $\phi(1) = n_1$. Ensuite, on pose $N_1 = n_1 + 1$, et on considère un entier $n_2 \geq N_1 > n_1$ tel que $\mathbf{x}_{n_2} \in \mathcal{B}(\mathbf{x}_*, \varepsilon_2)$. On pose $\phi(2) = n_2$. Et ainsi de suite.

Par construction, on a $n_1 < n_2 < \cdots$, donc la fonction ϕ est strictement croissante. C'est bien une extraction. De plus, on a $\mathbf{x}_{\phi(n)} \in \mathcal{B}(\mathbf{x}_*, \varepsilon_n)$, donc $0 \leq \|\mathbf{x}_{\phi(n)} - \mathbf{x}_*\| < \varepsilon_n$. La suite ε_n tend vers 0, donc $\|\mathbf{x}_{\phi(n)} - \mathbf{x}_*\|$ aussi, ce qui prouve le théorème. ☐

Nous terminons cette section avec un résultat qui fait le lien entre les ensembles fermés, et les convergences de suites.

Théorème 1.24 : Caractérisation des fermés

Soit A est un ensemble fermé de E, et soit $(\mathbf{x}_n) \subset A$ une suite de A qui converge dans $\mathbb{R}^d$ vers un certain $\mathbf{x}_*$. Alors $\mathbf{x}_* \in A$.

Démonstration. Supposons par l'absurde que $\mathbf{x}_* \notin A$. Alors $\mathbf{x}_* \in A^c$ qui est un ensemble ouvert. En particulier, il existe $\varepsilon > 0$ tel que $\mathcal{B}(\mathbf{x}_*, \varepsilon) \cap A = \emptyset$. Cela implique que pour tout $\mathbf{x} \in A$, on a $\|\mathbf{x} - \mathbf{x}_*\| \geq \varepsilon > 0$. En particulier, on a $\|\mathbf{x}_n - \mathbf{x}_*\| \geq \varepsilon$, qui ne tend pas vers 0. Contradiction. ☐

1.3 L'espace $\mathbb{R}^d$

Nous nous intéressons maintenant au cas où $E = \mathbb{R}^d$. Dans la suite, on note $\|\cdot\|$ la *norme euclidienne* de $\mathbb{R}^d$. On rappelle que celle-ci est définie par

$$\forall \mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix} \in \mathbb{R}^d, \quad \|\mathbf{x}\| := \|\mathbf{x}\|_{\mathbb{R}^d} = \sqrt{x_1^2 + \cdots + x_d^2}.$$

Cette norme est aussi parfois notée $\|x\|_2$. On peut munir $\mathbb{R}^d$ d'autres normes, et les définitions suivantes font encore sens. On démontrera dans la suite qu'en dimension finie (notre cas ici, d est la dimension), toutes les normes sont équivalentes.

1.3.1 Convergence composante par composante

Dans notre cas de l'espace $\mathbb{R}^d$, on peut vérifier la convergence dans $\mathbb{R}^d$ (convergence de vecteurs), en vérifiant la convergence de chaque composante (convergence de nombres réels).

Théorème 1.25 : Convergence composante par composante

Soit $(\mathbf{x}_n)$ une suite de $\mathbb{R}^d$, et soit $\mathbf{x}_* \in \mathbb{R}^d$. On note $\mathbf{x}_n = (x_{1,n}, \cdots x_{d,n})^T$ et $\mathbf{x}_* = (x_{1,*}, \cdots x_{d,*})^T$. Alors $(\mathbf{x}_n)$ converge vers $\mathbf{x}_*$ ssi pour tout $1 \leq i \leq d$, la suite **réelle** $(x_{j,n})$ converge vers $x_{j,*}$.

Démonstration. Supposons pour commencer que $(\mathbf{x}_n)$ converge vers $\mathbf{x}_*$. Soit $1 \le i \le d$. On a

$$0 \le |x_{i,n} - x_{i,*}| \le \left(\sum_{j=1}^{d} |x_{j,n} - x_{j,*}|^2 \right)^{1/2} = \|\mathbf{x}_n - \mathbf{x}_*\|.$$

Comme $(\mathbf{x}_n)$ converge vers $\mathbf{x}_*$, le terme de droite converge vers 0 lorsque $n \to \infty$. Donc $|x_{i,n} - x_{i,*}|$ aussi, ce qui montre que $(x_{i,n})$ converge vers $x_{i,*}$ (dans $\mathbb{R}$).
Supposons maintenant que $x_{j,n} - x_{j,*}$ converge vers 0 pour tout $1 \le j \le d$. Alors $\lim_{n\to\infty} |x_{j,n} - x_{j,*}| = 0$. En prenant le carré, en sommant, et en utilisant les propriétés des limites, on obtient

$$\lim_{n\to\infty} \|\mathbf{x}_n - \mathbf{x}_*\|^2 = \lim_{n\to\infty} \sum_{j=1}^{n} |x_{j,n} - x_{j,*}|^2 = \sum_{j=1}^{n} \lim_{n\to\infty} |x_{j,n} - x_{j,*}|^2 = 0.$$

□

1.3.2 Théorèmes de Bolzano Weierstrass

On rappelle le théorème de Bolzano–Weierstrass classique, dans $\mathbb{R}$.

Théorème 1.26 : Bolzano–Weierstrass dans $\mathbb{R}$.

Si $(x_n) \subset \mathbb{R}$ est une suite **réelle** bornée, alors (x_n) admet une valeur d'adhérence.

Démonstration. Nous donnons deux preuves de ce théorème important.

Première preuve (par dichotomie). Comme (x_n) est bornée, il existe $R > 0$ telle que $|x_n| < R$, donc $-R < x_n < R$ pour tout n. On considère les sous ensembles d'indices

$$I_1 = \{n \in \mathbb{N}, \quad x_n \in (-R, 0]\}, \quad J_1 = \{n \in \mathbb{N}, \quad x_n \in (0, R]\}.$$

Comme $-R < x_n < R$ pour tout n, on a $I_1 \cup J_1 = \mathbb{N}$, donc au moins un de ces deux ensembles contient une infinité de termes de la suite. On note $(a_1, b_1]$ cet interval (donc $a_1 = -R$ et $b_1 = 0$ si I_1 est de cardinal infini, et $a_1 = 0$ et $b_1 = R$ sinon). On recoupe cet intervalle en 2, et on pose

$$I_2 = \left\{ n \in \mathbb{N}, \quad x_n \in [a_1, \frac{1}{2}(a_1 + b_1)] \right\}, \quad J_2 = \left\{ n \in \mathbb{N}, \quad x_n \in (\frac{1}{2}(a_1 + b_1), b_1] \right\},$$

de sorte que $I_2 \cup J_2 = (a_1, b_1]$, donc au moins un de ces deux ensembles est de cardinal infini, et on note $(a_2, b_2]$ cet intervalle. On continue de même, et on construit deux suites (a_j) et (b_j) tel que

- l'intervalle $(a_j, b_j]$ contient une infinité de termes de la suite (x_n) ;
- la suite (a_j) est croissante, et la suite (b_j) est décroissante ;
- On a $|b_j - a_j| = 2^{-j} R \to 0$ as $j \to 0$.

D'après le théorème d'encadrement, les suites (a_j) et (b_j) converge vers la même limite, notée $x_* \in \mathbb{R}$, et on a $x_* \in [a_j, b_j]$ pour tout $j \in \mathbb{N}$. Montrons que x_* est une valeur d'adhérence de la suite (x_n). Pour tout $\varepsilon > 0$, il existe j tel que $2^{-j} R < \varepsilon$. En particulier, pour ce j, il existe un n (en fait une infinité) tel que $x_n \in (a_j, b_j]$ et $x_* \in [a_j, b_j]$, donc $|x_n - x_*| \le |b_j - a_j| = R 2^{-j} < \varepsilon$. On conclut avec le Théorème 1.23.

Deuxème preuve. Comme (x_n) est bornée, il existe $R > 0$ telle que $|x_n| < R$. Soit $\nu = [-R, R] \to \mathbb{N} \cup \{\infty\}$ la fonction définie par

$$\forall t \in [-R, R], \quad \nu(t) := \mathrm{Card}\, \{n \in \mathbb{N}, \quad x_n < t\}.$$

Le nombre $\nu(t)$ compte le nombre d'éléments de la suite (x_n) qui sont plus petits que t. On s'autorise à prendre la valeur ∞ lorsqu'il y a une infinité d'éléments de la suite plus petits que t.

La suite $\nu(t)$ est croissante, et vérifie $\nu(-R) = 0$ et $\nu(R) = \infty$. On pose

$$t_* := \inf \{ t \in [-R, R], \quad \forall s > t, \quad \nu(s) = \infty \} .$$

Montrons que t_* est une valeur d'adhérence de (x_n). Par définition de t_*, pour tout $\varepsilon > 0$, on a $\nu(t_* + \varepsilon) = \infty$ et $\nu(t_* - \varepsilon) < \infty$. En particulier, il y a une infinité d'éléments de la suite dans l'intervalle $(t_* - \varepsilon, t_* + \varepsilon)$, et on conclut avec le Théorème 1.23. ☐

Remarque 1.27. *Le nombre t_* construit dans la deuxième preuve s'appelle la* **liminf** *de la suite (x_n). C'est la plus petite valeur d'adhérence de la suite (x_n).*

La deuxième preuve utilise fortement le caractère **ordonné** de $\mathbb{R}$. Dans le cas $\mathbb{R}^d$, on peut répéter l'argument "coordonnée par coordonné". On obtient le théorème suivant.

Théorème 1.28 : Bolzano–Weierstrass dans $\mathbb{R}^d$.

Si $(\mathbf{x}_n) \subset \mathbb{R}^d$ est une suite (vectorielle) bornée, alors $(\mathbf{x}_n)$ admet une valeur d'adhérence.

Démonstration. Si $(\mathbf{x}_n)$ est bornée dans $\mathbb{R}^d$, alors la suite réelle $(x_{1,n})$ (première composante) est bornée dans $\mathbb{R}$. D'après le théorème de Bolzano–Weierstrass dans $\mathbb{R}$, on peut trouver une extraction $\phi_1 : \mathbb{N} \to \mathbb{N}$ et une limite $y_1 \in \mathbb{R}$ tel que $(x_{1,\phi_1(n)}) \to y_1$.

On considère maintenant la suite extraite réelle $\mathbf{x}_{2,\phi_1(n)}$ (deuxième composante). Cette suite est bornée dans $\mathbb{R}$. D'après le théorème de Bolzano–Weierstrass dans $\mathbb{R}$, on peut trouver une extraction $\phi_2 : \mathbb{N} \to \mathbb{N}$ et une limite $y_2 \in \mathbb{R}$ tel que $x_{2,\phi_1(\phi_2(n))} \to y_2$(attention à l'ordre des extractions). De plus, par le Théorème 1.15 (toute sous-suite d'une suite convergente converge vers la même limite), on a encore $x_{1,\phi_1(\phi_2(n))} \to y_1$. Après cette seconde extraction, les deux premières composantes convergent.

On continue le raisonnement, et on construit les extractions ϕ_1, ϕ_2,..., ϕ_d. On pose enfin

$$\phi(n) := \phi_1 \circ \phi_2 \circ \cdots \circ \phi_d(n) = \phi_1 \left(\phi_2 \cdots (\phi_d(n)) \right) .$$

Alors, par construction, la j-ème coordonnée de $\mathbf{x}_{\phi(n)}$ converge vers y_j. Donc la suite $\mathbf{x}_{\phi(n)}$ converge vers $\mathbf{y} := (y_1, \cdots, y_d)^T$. ☐

Ce résultat est faux en général en dimension infinie (que nous n'étudierons pas dans ce cours). Une raison est que la composée d'une infinité d'extraction n'est pas forcément bien définie. Par exemple, si $\phi(n) := n + 1$, on a $\phi^{(k)}(n) = n + k$ (composée k fois), et $\phi^{(\infty)}$ n'a aucun sens.

Exercice 1.29

Soit E l'ensemble des suites réelles bornées. Pour $\mathbf{u} = (u_n) \in E$, on pose $\|\mathbf{u}\|_E := \sup_{n \in \mathbb{N}} |u_n|$.
a/ Montrer que $(E, \| \cdot \|_E)$ est un $\mathbb{R}$-espace vectoriel normé (appelé souvent $\ell^\infty(\mathbb{N}, \mathbb{R})$.
b/ Pour $k \in \mathbb{N}$, soit $\mathbf{u}_k \in E$ la suite dont les k premiers termes valent 0, et tous les suivants valent 1. Montrer que pour tout k, $\|\mathbf{u}_k\|_E = 1$, et que $k \neq k'$, on a aussi $\|\mathbf{u}_k - \mathbf{u}_{k'}\| = 1$.
c/ En déduire que la suite $(\mathbf{u}_k)$ de E n'admet pas de sous-suite convergente.

CHAPITRE 2

OPTIMISATION : EXISTENCE ET UNICITÉ

2.1 Fonctions continues

Dans la suite, on se fixe deux espaces vectoriels normés, $(E, \| \cdot \|_E)$ et $(F, \| \cdot \|_F)$, et on s'intéresse aux fonctions continues de $(E, \| \cdot \|_E)$ dans $(F, \| \cdot \|_F)$.

Nous rappelons qu'une fonction de E dans F devrait être notée $(f, \mathcal{D}_f)$, où $\mathcal{D}_f \subset E$ est le **domaine de définition** de f. C'est un sous-ensemble de E qui vérifie que l'évaluation $f(\mathbf{x})$ est bien définie pour tout $\mathbf{x} \in \mathcal{D}_f$. Cependant, on utilise en pratique la notation f (sans expliciter son domaine). Cela peut parfois (mais rarement) porter à confusion.

Définition 2.1 : Fonction continue, première définition

Soit $(f, \mathcal{D}_f)$ une fonction de E dans F. Soit $\mathbf{x}_* \in \mathcal{D}_f$. On dit que f est continue en $\mathbf{x}_*$ si :

$$\forall \varepsilon > 0, \quad \exists \delta > 0, \quad \forall \mathbf{y} \in \mathcal{D}_f, \quad \| \mathbf{y} - \mathbf{x}_* \|_E < \delta \implies \| f(\mathbf{x}_*) - f(\mathbf{y}) \|_F < \varepsilon.$$

On dit que f est continue sur $K \subset \mathcal{D}_f$ si f est continue en tout point de K.

Voici une deuxième définition possible, et équivalente (le fait que ces deux définitions soient équivalentes est laissé en exercice).

Définition 2.2 : Fonction continue, deuxième définition

Soit $(f, \mathcal{D}_f)$ une fonction de E dans F. Soit $\mathbf{x}_* \in \mathcal{D}_f$. On dit que f est continue en $\mathbf{x}_*$ si, pour toute suite $(\mathbf{x}_n)_n$ de $\mathcal{D}_f$ qui converge vers $\mathbf{x}_*$ (dans E), la suite $(f(\mathbf{x}_n))_n$ converge vers $f(\mathbf{x}_*)$ (dans F).

Remarque 2.3. *La notion de continuité dépend a priori des normes qu'on choisit sur E et sur F.*

Avant de donner des exemples explicites, nous faisons une petite digression. Avec cette définition, une fonction f de $\mathbb{Z}$ dans $\mathbb{R}$ est toujours continue (pourquoi ?)... Considérons les fonctions $(f_1, \mathbb{R})$ et $(f_2, \mathbb{Q})$ définies par $f_1(x) = f_2(x) = \mathbb{1}(x \in \mathbb{Q})$. Alors, quand bien même f_1 et f_2 ont la même définition, la fonction f_2 est continue, mais pas la fonction f_1... Cela vient du fait qu'on a restreint le domaine de définition de f_1. *La notion de continuité dépend donc du choix du domaine.* Dire « *f est continue en $\mathbf{x}_0$* » est un abus de langage. On devrait plutôt dire « *$(f, \mathcal{D}_f)$ est continue en $\mathbf{x}_0$* ». Cependant, cela alourdit les notations, et on supposera dans la suite qu'on travaille toujours avec le **domaine maximal** de f.

Voici quelques exemples.

- (**addition**) La fonction $A : \mathbb{R}^2 \to \mathbb{R}$ définie par $A(x_1, x_2) := x_1 + x_2$ est continue (car la somme des limites est égale à la limite des sommes).
- (**multiplication**) La fonction $M : \mathbb{R}^2 \to \mathbb{R}$ définie par $P(x_1, x_2) := x_1 x_2$ est continue (car le produit des limites est égale à la limite des produit).
- (**projection**) Pour tout $1 \le j \le d$, la fonction $P_j : (x_1, \cdots, x_d) \mapsto x_j$ est continue (cf. Théorème 1.25).
- (**norme**) Une norme $\mathcal{N}(\cdot)$ sur $E \to \mathbb{R}$ est continue dans $(E, \mathcal{N}(\cdot))$. On dit qu'une nome est continue pour sa topologie. En effet, si $\mathbf{x}_n$ converge vers $\mathbf{x}_*$, on a, avec l'inégalité triangulaire inverse,
$$0 \le |\mathcal{N}(\mathbf{x}_n) - \mathcal{N}(\mathbf{x}_*)| \le \mathcal{N}(\mathbf{x}_n - \mathbf{x}_*) \xrightarrow[n\to\infty]{} 0.$$

Exercice 2.4

On dit qu'une fonction f de $(E, \|\cdot\|_E)$ dans $(F, \|\cdot\|_F)$ est M–Lipschitz si

$$\forall \mathbf{x}, \mathbf{y} \in E, \qquad \|f(\mathbf{x}) - f(\mathbf{y})\|_F \le M \|\mathbf{x} - \mathbf{y}\|_E.$$

1/ Montrer qu'une fonction Lipschitz est continue.
2/ Montrer qu'une norme $\mathcal{N}(\cdot)$ est 1–Lipschitz de $(E, \mathcal{N}(\cdot))$ dans $(\mathbb{R}, |\cdot|)$.

Théorème 2.5 : La composition de fonctions continues est continue

Soit $(f, \mathcal{D}_f)$ une fonction de $(E, \|\cdot\|_E)$ dans $(F, \|\cdot\|_F)$, et soit $(g, \mathcal{D}_g)$ une fonction de $(F, \|\cdot\|_F)$ dans $(G, \|\cdot\|_G)$. On suppose que $f(\mathcal{D}_f) \subset \mathcal{D}_g$. Alors $(g \circ f, \mathcal{D}_f)$ est continue de $\mathcal{D}_f$ dans G.

Démonstration. Soit $\mathbf{x}_* \in \mathcal{D}_f$ fixé, et soit $(\mathbf{x}_n)$ une suite de $\mathcal{D}_f$ qui converge vers $\mathbf{x}_*$. Comme f est continue, la suite $\mathbf{y}_n := f(\mathbf{x}_n)$ converge vers $\mathbf{y}_* := f(\mathbf{x}_*)$. Comme g est continue, la suite $g(\mathbf{y}_n)$ converge vers $g(\mathbf{y}_*)$. Donc $g(f(\mathbf{x}_n)) \to g(f(\mathbf{x}_*))$. On en déduit que $g \circ f$ est continue en $\mathbf{x}_*$. Ceci étant vrai pour tout $\mathbf{x}_* \in \mathcal{D}_f$, $g \circ f$ est continue sur $\mathcal{D}_f$. ☐

2.1.1 Continuité des fonctions usuelles

Le Théorème 2.5 permet de composer les fonctions continues. Par applications successives de ce théorème, on en déduit que (presque) toutes les fonctions usuelles, et leur composition, sont continues sur leur domaine maximal de définition.

En pratique, lorsqu'on voudra démontrer qu'une fonction f donnée est continue, on pourra se contenter de calculer son domaine de définition, et d'écrire

La fonction f est continue sur son domaine de définition comme composée de fonctions usuelles continues sur leur domaine de définition.

Il faudra tout de même préciser le domaine (maximal) de définition de f.

2.1.2 Prolongement par continuité

Définition 2.6 : Prolongement par continuité

Soit $(f, \mathcal{D}_f)$ une fonction de E dans F, et soit $\mathbf{x}_* \in E \backslash \mathcal{D}_f$. On dit que f admet un prolongement par continuité en $\mathbf{x}_*$ s'il existe une limite $\ell \in F$ telle que, pour toute suite $(\mathbf{x}_n)$ de $\mathcal{D}_f \cup \{\mathbf{x}_*\}$ qui converge vers $\mathbf{x}_*$, on a $f(\mathbf{x}_n) \to \ell$.
Dans ce cas, l'extension de f par continuité en $\mathbf{x}_*$ est la fonction $(\widetilde{f}, \mathcal{D}_{\widetilde{f}})$ avec $\mathcal{D}_{\widetilde{f}} := \mathcal{D}_f \cup \{\mathbf{x}_*\}$ définie par

$$\forall \mathbf{x} \in \mathcal{D}_{\widetilde{f}}, \quad \widetilde{f}(\mathbf{x}) = \begin{cases} f(\mathbf{x}) & \text{if} \quad \mathbf{x} \in \mathcal{D}_f \\ \ell \quad \text{if} & \mathbf{x} = \mathbf{x}_*. \end{cases}$$

Si f était une fonction continue sur $\mathcal{D}_f$, alors la nouvelle fonction $\widetilde{f}$ est continue sur $\mathcal{D}_{\widetilde{f}}$.

2.1.3 Un exemple important

Considérons l'exercice (classique) suivant.

Exercice 2.7

La fonction $f : \mathbb{R}^2 \to \mathbb{R}$, définie sur $\mathbb{R}^2 \setminus \{\mathbf{0}\}$ par

$$f(x, y) := \frac{xy}{x^2 + y^2}$$

admet-elle un prolongement par continuité en 0 ?

On remarque que f est continue sur son domaine de définition $\mathbb{R}^2 \setminus \{\mathbf{0}\}$ comme composée de fonctions usuelles. Supposons que f admette un prolongement par continuité en $\mathbf{0}$, avec limite $\ell \in \mathbb{R}$. Alors, en prenant la suite $\mathbf{x}_n := (\frac{1}{n}, 0)$, on obtient $f(\mathbf{x}_n) \to 0$, donc $\ell = 0$. Cependant, en prenant la suite $\mathbf{y}_n := (\frac{1}{n}, \frac{1}{n})$, on a $f(\mathbf{y}_n) \to \frac{1}{2}$, donc $\ell = \frac{1}{2}$, contradiction. Ceci montre que f n'admet pas de prolongement par continuité en $\mathbf{0}$.

Cet exemple est assez surprenant, car les applications partielles $x \mapsto f(x, 0)$ et $y \mapsto (0, y)$ admettent des prolongements par continuité sur $\mathbb{R}$ (ce sont les fonctions nulles), mais pas la fonction f. Cela vient du fait que dans $\mathbb{R}^2$, on peut approcher le point $\mathbf{0}$ de plein de façons différentes, par seulement selon les axes $x = 0$ et $y = 0$.

Par exemple, pour tout $\theta \in \mathbb{R}$, la fonction $g_\theta : \mathbb{R}_* \to \mathbb{R}$ définie par $g(r) := f(r\cos(\theta), r\sin(\theta))$ admet un prolongement par continuité en $r = 0$ avec $g_\theta(0) = \cos(\theta)\sin(\theta)$. Donc la fonction f initiale peut se prolonger par continuité le long de toute droite passant par $\mathbf{0}$. En revanche, elle admet des limites différentes selon la droite qu'on considère...

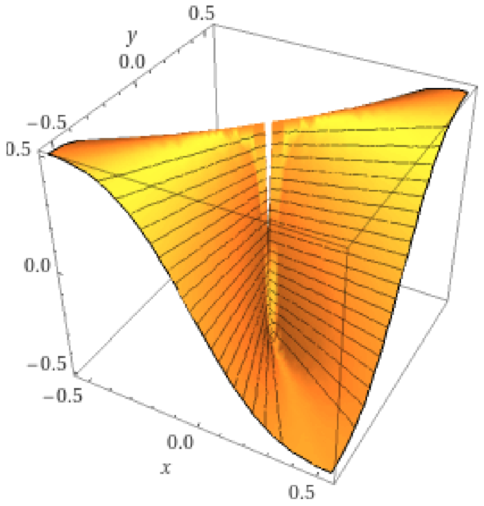

FIGURE 2.1 – La surface de f de l'Exercice 2.7. Il faut imaginer une *nappe pincée*.

2.1.4 Fonctions à m composantes

Dans la suite, on étudiera parfois des fonctions f de $\mathbb{R}^d$ à valeurs dans $\mathbb{R}^m$. Ici, d est le nombre de **variables** (d pour *dimension*), et m est le nombre de **composantes** de la fonction f. Une telle fonction est de la forme

$$f(\mathbf{x}) = f(x_1, \cdots, x_d) = \begin{pmatrix} f_1(\mathbf{x}) \\ f_2(\mathbf{x}) \\ \vdots \\ f_m(\mathbf{x}) \end{pmatrix} = \begin{pmatrix} f_1(x_1, \cdots, x_d) \\ f_2(x_1, \cdots, x_d) \\ \vdots \\ f_m(x_1, \cdots, x_d) \end{pmatrix}.$$

Il est important de noter que les composantes de f sont *indépendantes*. On peut imaginer une situation par exemple où f_1 est une température, f_2 une pression, f_3 un nombre de pommes, etc. En appliquant le Théorème 1.25, on obtient le résultat suivant.

Théorème 2.8

Soit $(f, \mathcal{D}_f)$ une fonction de $\mathbb{R}^d$ dans $\mathbb{R}^m$. Alors f est continue en $\mathbf{x}_* \in \mathcal{D}_f$ ssi toutes ses composantes sont continues en $\mathbf{x}_*$. En particulier, f est continue sur $K \subset \mathcal{D}_f$ ssi toutes ses composantes sont continues sur K.

Il n'y a donc pas de difficultés supplémentaires à étudier des fonctions à plusieurs composantes : on se ramène au cas des fonctions à valeurs réelles.

2.2 Optimisation, premières définitions

Optimiser une fonction, c'est trouver sa plus petite et/ou sa plus grande valeur. En particulier,

On ne peut optimiser qu'une fonction à valeurs réelles !

C'est à dire que l'espace d'arrivée de f doit être $\mathbb{R}$ ($m = 1$ composante) ! En effet, on doit pouvoir comparer les valeurs $f(\mathbf{x})$ et $f(\mathbf{y})$, et pouvoir écrire $f(\mathbf{x}) \le f(\mathbf{y})$ ou $f(\mathbf{x}) > f(\mathbf{y})$. Or nous n'avons une notion d'ordre que pour des nombres réelles (on ne peut pas comparer deux nombres complexes, ni deux vecteurs). Dans la suite, on s'intéresse donc exclusivement aux fonctions à valeurs dans $\mathbb{R}$.

2.2.1 Premières définitions

Nous commençons avec quelques définitions.

Infimum et minimum d'un sous-ensemble de $\mathbb{R}$.

Définition 2.9 : Infimum d'un ensemble

Soit $I \subset \mathbb{R}$ un sous ensemble *quelconque* de $\mathbb{R}$.
Un nombre $m \in \mathbb{R}$ est un **minorant** de I si pour tout $x \in I$, on a $m \le x$.
L'ensemble I est **borné inférieurement** si I admet au moins un minorant.
L'**infimum** de I est le plus grand des minorants. C'est le nombre $m_* \in \mathbb{R}$, noté $\inf I$, tel que, si m est un minorant de I, alors $m_* \ge m$.
Par convention, si I n'est pas borné inférieurement, on note $\inf I := -\infty$.

On définit de même la notion de **majorant**, **borné supérieurement** et de **supremum** de I. On pourra remarquer de manière cocasse que définir l'infimum comme *le plus grand des minorants* suppose qu'on sait déjà définir le supremum (*le plus grand*) d'un ensemble, et que donc cette définition tourne en rond...
En fait, il est classique de montrer que l'ensemble des minorants de I est un intervalle de la forme $(-\infty, M]$ (ou bien $\emptyset$ si I n'est pas bornée inférieurement) pour un certain M unique, et on définit l'infimum de I comme étant ce nombre M.

Exemples :

- Si $I = \mathbb{R}$, on a $\inf I = -\infty$ et $\sup I = +\infty$.
- Si $I = (a, b)$, on a $\inf I = a$ et $\sup I = b$.
- Si $I = [a, b]$, on a $\inf I = a$, et $\sup I = b$.

Dans le premier cas, l'infimum et le supremum de K ne sont pas dans I, et dans le second cas, ils sont dans I. Ceci motive la définition suivante.

Définition 2.10

Soit $m_* := \inf I$ l'infimum de I. On dit que m_* est un **minimum** de I si $m_* \in I$.

Ainsi, a n'est pas un minimum de (a, b), mais est un minimum de $[a, b]$. Par convention, si $m_* = -\infty$, m_* n'est pas un minimum. On définit de même la notion de **maximum** de K.

Théorème 2.11 : Caractérisation de l'infimum avec des suites

Soit $I \subset \mathbb{R}$ un ensemble non vide quelconque.
Si I est non borné inférieurement, alors il existe une suite (x_n) de I qui diverge vers $-\infty$.
Si I est borné inférieurement, et si $m \in \mathbb{R}$ est un minorant de I, alors m est l'infimum de I ssi il existe une suite $(x_n) \in I$ qui converge vers m_*.

Démonstration. Nous prouvons le deuxième point seulement. Soit m_* l'infimum de K. Comme m est un minorant, on a $m \le m_*$. Si $m < m_*$, alors pour tout suite (x_n) de I, on a $m < m_* \le x_n$, donc $|x_n - m| > |m_* - m|$, et la suite (x_n) ne peut pas converger vers m. Au contraire, si $m = m_*$ est l'infimum, alors, pour tout $n \in \mathbb{N}^*$, le nombre $m_* + \frac{1}{n}$ n'est pas un minorant de K. Donc il existe $x_n \in I$ avec $m_* \le x_n \le m_* + \frac{1}{n}$. Ceci construit une suite (x_n) de I qui converge vers m_*. ☐

Exercice 2.12

Montrer que si I est fermé et borné inférieurement, alors I admet un minimum.

Infimum et minimum d'une fonction réelle.

Les notions précédentes s'appliquent pour des fonctions. Soit $K \subset \mathbb{R}^d$ un sous-ensemble *quelconque* de $\mathbb{R}^d$, et soit $f : K \to \mathbb{R}$ une fonction réelle. On note l'image de f par

$$I := f(K) = \{f(\mathbf{x}), \, \mathbf{x} \in K\} \qquad \subset \mathbb{R}.$$

Avec ces notations, on appelle "l'infimum de f sur K" le nombre $\inf I$. On le note

$$\inf_K f \quad \text{ou} \quad \inf \{f(\mathbf{x}), \ \mathbf{x} \in K\}.$$

On note de même le supremum de f par $\sup_K f$.

Définition 2.13 : Minimum et minimiseur d'une fonction

Soit $m_* := \inf\{f(\mathbf{x}), \ \mathbf{x} \in K\}$ l'infimum de f. On dit que m_* est le **minimum** de f si m_* est le minimum de I. Dans ce cas, il existe $\mathbf{x}_* \in K$ tel que $f(\mathbf{x}_*) = m_*$. Tout point $\mathbf{x}_* \in K$ vérifiant $f(\mathbf{x}_*) = \inf_K f$ est appelé un **minimiseur** de f sur K.

On définit pareillement la notion de **maximum** et de **maximiseur** de f sur K.

Exemples

- Soit $K := \mathbb{R}^d \setminus \{\mathbf{0}\}$, et soit $f : K \to \mathbb{R}$ définie par $f(\mathbf{x}) = \|\mathbf{x}\|$. Alors 0 est l'infimum de f sur K, mais 0 n'est pas un minimiseur, car $\mathbf{0} \notin K$.

- Soit $K = \mathbb{R}$, et $f : \mathbb{R} \to \mathbb{R}$ définie par $f(x) = \sin^2(x)$. Alors 0 est l'infimum de f. C'est un minimum de f, et l'ensemble des minimiseurs de f est l'ensemble $\pi\mathbb{Z}$.

Grâce au Théorème 2.11, on a une caractérisation de l'infimum de f sur K.

Théorème 2.14

Soit m un minorant de f sur K. Alors m est l'infimum de f sur K ssi il existe une suite $(\mathbf{x}_n)$ de K telle que $f(\mathbf{x}_n) \to m$.

Définition 2.15 : Suite minimisante

Soit $m_* := \inf_K f$ l'infimum de f sur K. Une **suite minimisante** de f sur K est une suite $(\mathbf{x}_n)$ de K telle que $f(\mathbf{x}_n) \to m_*$.

On remarquera que la suite $(\mathbf{x}_n)$ ne converge pas forcément. Seules les images $f(\mathbf{x}_n)$ convergent...

2.2.2 Toutes les normes de $\mathbb{R}^d$ sont équivalentes

Théorème 2.16 : En dimension finie, toutes les normes sont équivalentes

Toutes les normes de $\mathbb{R}^d$ sont équivalentes.

Démonstration. Soit $\mathcal{N}$ une norme sur $\mathbb{R}^d$. Il suffit de montrer que $\mathcal{N}$ est équivalente à la norme euclidienne $\| \cdot \|$. Pour commencer, en notant $(\mathbf{e}_i)_{1 \le i \le n}$ la base canonique de $\mathbb{R}^d$, et en notant $\mathbb{R}^d \ni \mathbf{x} = \sum_{i=1}^d x_i \mathbf{e}_i$ avec $x_i \in \mathbb{R}$, on a

$$\forall \mathbf{x} \in \mathbb{R}^d, \quad \mathcal{N}(\mathbf{x}) = \mathcal{N}\left(\sum_{i=1}^d x_i \mathbf{e}_i\right) \le \sum_{i=1}^d \mathcal{N}(x_i \mathbf{e}_i) = \sum_{i=1}^d |x_i| \mathcal{N}(\mathbf{e}_i) \le \left(\sum_{i=1}^d |x_i|^2\right)^{1/2} \left(\sum_{i=1}^d \mathcal{N}(\mathbf{e}_i)^2\right)^{1/2}.$$

On a utilisé l'inégalité de Cauchy-Schwarz pour la dernière inégalité. Ceci montre que

$$\forall \mathbf{x} \in \mathbb{R}^d, \quad \mathcal{N}(\mathbf{x}) \le \alpha_1 \|\mathbf{x}\|, \quad \text{avec} \quad \alpha_1 := \left(\sum_{i=1}^d \mathcal{N}(\mathbf{e}_i)^2\right)^{1/2}.$$

En particulier, avec l'inégalité triangulaire inverse, on a

$$|\mathcal{N}(\mathbf{x}) - \mathcal{N}(\mathbf{y})| \le \mathcal{N}(\mathbf{x} - \mathbf{y}) \le \alpha_1 \|\mathbf{x} - \mathbf{y}\|,$$

ce qui montre que $\mathcal{N}(\cdot)$ est α_1–Lipschitz de $(\mathbb{R}^d, \| \cdot \|)$ dans $(\mathbb{R}, | \cdot |)$, donc continue *pour la norme euclidienne*, c'est à dire de $(\mathbb{R}^d, \| \cdot \|)$ dans $\mathbb{R}$. Ce résultat est à comparer avec celui de l'Exercice 2.4 (et de l'exemple précédent), qui nous dit que $\mathcal{N}(\cdot)$ est continue de $(\mathbb{R}^d, \mathcal{N}(\cdot))$ dans $\mathbb{R}$, i.e. pour sa propre norme.

On pose

$$m := \inf \left\{ \mathcal{N}(\mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^d, \ \|\mathbf{x}\| = 1 \right\}, \tag{2.1}$$

et soit $(\mathbf{x}_n)$ une suite minimisante pour ce problème, c'est à dire qui vérifie $\|\mathbf{x}_n\| = 1$ et $\mathcal{N}(\mathbf{x}_n) \to m$. La suite $(\mathbf{x}_n)$ est bornée dans $\mathbb{R}^d$. D'après le théorème de Bolzano-Weierstrass, il existe une sous-suite $(\mathbf{x}_{\phi(n)})$ et un élément $\mathbf{x}_* \in \mathbb{R}^d$ telle que $\mathbf{x}_n$ converge vers $\mathbf{x}_*$ pour la norme $\| \cdot \|$. Par continuité de $\| \cdot \|$ et de $\mathcal{N}(\cdot)$ pour la norme euclidienne, on obtient

$$\|\mathbf{x}_*\| = \lim_{n \to \infty} \|\mathbf{x}_n\| = 1, \quad \text{et} \quad \mathcal{N}(\mathbf{x}_*) = \lim_{n \to \infty} \mathcal{N}(\mathbf{x}_n) = m.$$

Ainsi, m est un minimum du problème d'optimisation (2.1), et $\mathbf{x}_*$ est un minimiseur. Comme $\|\mathbf{x}_*\| = 1$, on en déduit que $\mathbf{x}_* \neq \mathbf{0}$, et donc $m = \mathcal{N}(\mathbf{x}_*) > 0$.

Pour conclure, on remarque que pour tout $\mathbf{x} \in \mathbb{R}^d \setminus \{\mathbf{0}\}$, on a

$$\mathcal{N}(\mathbf{x}) = \mathcal{N}\left(\|\mathbf{x}\| \frac{\mathbf{x}}{\|\mathbf{x}\|} \right) = \|\mathbf{x}\| \mathcal{N}\left(\frac{\mathbf{x}}{\|\mathbf{x}\|} \right) \geq \|\mathbf{x}\| m.$$

Cette inégalité est encore vraie avec $\mathbf{x} = \mathbf{0}$. Ceci montre que

$$\forall \mathbf{x} \in \mathbb{R}^d, \quad \|\mathbf{x}\| \leq \alpha_2 \mathcal{N}(\mathbf{x}), \quad \text{avec} \quad \alpha_2 := \frac{1}{m} > 0.$$

En prenant $\alpha := \max\{\alpha_1, \alpha_2\}$, on trouve que $\mathcal{N}$ et $\| \cdot \|$ sont équivalentes.

□

2.3 Optimisation de fonctions continues, existence

Dans cette section, nous donnons des critères simples pour qu'un problème d'optimisation ait une solution.

2.3.1 Existence de minimiseurs dans le cas fermé borné

Nous commençons avec l'important théorème de Weierstrass.

Théorème 2.17 : Théorème de Weierstrass

Soit $K \subset \mathbb{R}^d$ un ensemble **fermé** et **borné**, et soit $f : K \to \mathbb{R}$ une fonction **continue** sur K. Alors f atteint son minimum sur K, c'est à dire qu'il existe $\mathbf{x}_* \in K$ tel que $f(\mathbf{x}_*) = \inf_K f$.

Démonstration. Soit $m_* := \inf_K f$, et soit $(\mathbf{x}_n)$ une suite minimisante de f.
La suite $(\mathbf{x}_n)$ est dans K, qui est borné. D'après le théorème de Bolzano–Weierstrass, on peut en extraire une sous-suite $\mathbf{x}_{\phi(n)}$ qui converge vers un certain $\mathbf{x}_* \in \mathbb{R}^d$.
La fonction f est continue, donc $f(\mathbf{x}_{\phi(n)}) \to f(\mathbf{x}_*)$. Par ailleurs, comme $(\mathbf{x}_n)$ est une suite minimisante, on a aussi $f(\mathbf{x}_{\phi(n)}) \to m_*$. Par unicité de la limite, on en déduit que $f(\mathbf{x}_*) = m_*$.
De plus, l'ensemble K est fermé, et $\mathbf{x}_{\phi(n)}$ est une suite de K qui converge vers $\mathbf{x}_*$, donc $\mathbf{x}_* \in K$. Ainsi, $\mathbf{x}_*$ est un minimiseur de f sur K.

□

En particulier, comme f est définie sur K, $f(\mathbf{x}_*) \in \mathbb{R}$ ne prend pas la valeur $-\infty$. On en déduit que, sous les mêmes hypothèses, la fonction f est bornée inférieurement.

Évidemment, le théorème reste vrai si on remplace "minimum" par "maximum".

2.3.2 Existence de minimiseurs pour les fonctions continues coercives

Dans la pratique, on étudie parfois des fonctions $f : \mathbb{R}^d \to \mathbb{R}$ définie sur tout l'espace $\mathbb{R}^d$. L'espace $\mathbb{R}^d$ n'est pas borné, et on ne peux pas utiliser le théorème de Weierstrass. On utilise plutôt la notion de **coercivité**.

Définition 2.18 : Fonctions coercives

Soit $f : \mathbb{R}^d \to \mathbb{R}$. On dit que f est **coercive**, si, pour tout $M \geq \mathbb{R}$, il existe $R > 0$ tel que

$$\forall \mathbf{x} \in \mathbb{R}^d, \quad \|\mathbf{x}\| > R \implies f(\mathbf{x}) > M.$$

Cela signifie qu'en dehors d'une grande boule $\mathcal{B}(\mathbf{0}, R)$, la fonction f prend des valeurs très grandes, plus grandes qu'une valeur M arbitraire. On note parfois $\lim_{\|\mathbf{x}\| \to \infty} f(\mathbf{x}) = \infty$ lorsque f est coercive.

Exercice 2.19

Soit $\alpha > 0$. Montrer que la fonction $f : \mathbb{R}^d \to \mathbb{R}$ définie par $f(\mathbf{x}) = \|\mathbf{x}\|^\alpha$ est coercive.

Théorème 2.20 : Existence de minimiseurs pour les fonctions continues coercives

Soit $f : \mathbb{R}^d \to \mathbb{R}$ une fonction **continue** et **coercive**. Alors f atteint son minimum : il existe $\mathbf{x}_* \in \mathbb{R}^d$ tel que $f(\mathbf{x}_*) = \inf_{\mathbb{R}^d} f$.

Démonstration. Soit m_* l'infimum de f, et soit $(\mathbf{x}_n)$ une suite minimisante de f, i.e. tel que $f(\mathbf{x}_n) \to m_*$. Comme l'infimum est un minorant, on a $m_* \leq f(\mathbf{0})$. Par coercivité de f il existe $R > 0$ tel que pour tout $\mathbf{x} \in \mathbb{R}^d \setminus \overline{\mathcal{B}(\mathbf{0}, R)}$, on a $f(\mathbf{x}) > f(\mathbf{0}) + 1 > m_*$. En particulier, la suite $(\mathbf{x}_n)$ est dans la boule fermée (bornée) $\overline{\mathcal{B}(\mathbf{0}, R)}$ à partir d'un certain rang. On s'est ramené au cas du fermé borné $K = \overline{\mathcal{B}(\mathbf{0}, R)}$, et on conclut comme pour le théorème de Weierstrass. ☐

2.3.3 Unicité des minimiseurs pour les fonctions strictement convexes

Dans les sections précédentes, nous avons trouvé des critères simples pour qu'une fonction admette des minimiseurs. Nous cherchons maintenant un critère simple pour que ce minimiseur soit **unique**. Une bonne notion est celle de **convexité**.

Définition 2.21 : Ensemble convexe

Soit $K \subset \mathbb{R}^d$ un ensemble quelconque. On dit que K est **convexe**, si, pour tout $\mathbf{x}, \mathbf{y} \in K$ et tout $0 \leq t \leq 1$, le point $t\mathbf{x} + (1-t)\mathbf{y}$ est dans K.

On remarquera que le point $t\mathbf{x} + (1-t)\mathbf{y}$ est une combinaison *barycentrique* des points $\mathbf{x}$ et $\mathbf{y}$, qui parcourt le segment $[\mathbf{x}, \mathbf{y}]$ lorsque t parcourt l'intervalle $[0, 1]$. Autrement dit, K est convexe si, pour toute paire de points $\mathbf{x}, \mathbf{y}$ dans K, le segment $[\mathbf{x}, \mathbf{y}]$ est inclus dans K. On remarquera qu'il n'existe pas de notion *d'ensemble concave...*

Définition 2.22 : Fonction convexe

Soit $K \subset \mathbb{R}^d$ un ensemble convexe, et soit $f : K \to \mathbb{R}$ une fonction définie sur K.
On dit que f est **convexe** si

$$\forall \mathbf{x}, \mathbf{y} \in K, \quad \forall t \in [0,1], \quad f(t\mathbf{x} + (1-t)\mathbf{y}) \leq t f(\mathbf{x}) + (1-t) f(\mathbf{y}).$$

On dit que f est **strictement** convexe si

$$\forall \mathbf{x} \neq \mathbf{y} \in K, \quad \forall t \in (0,1), \quad f(t\mathbf{x} + (1-t)\mathbf{y}) < t f(\mathbf{x}) + (1-t) f(\mathbf{y}).$$

Lorsque la fonction $-f$ est convexe, on dit que la fonction est **concave**. Pour les fonctions concaves ou strictement concaves, les inégalités sont dans l'autre sens. Nous n'utiliserons cependant pas cette notion dans ce cours, car nous sommes intéressés surtout par la minimisation de fonctions.

Exercice 2.23

Soit $f : \mathbb{R}^d \to \mathbb{R}$ une fonction convexe, et soit $E \in \mathbb{R}$. Montrer que l'ensemble $K_E := \left\{ \mathbf{x} \in \mathbb{R}^d, \ f(\mathbf{x}) \leq E \right\}$ est un ensemble convexe de $\mathbb{R}^d$.

Exemples

— L'ensemble $\mathbb{R}^d$ est convexe.

— Si $\mathcal{N} : \mathbb{R}^d \to \mathbb{R}$ est une norme, alors $\mathcal{N}$ est convexe. En effet, on a, avec l'inégalité triangulaire,

$$\mathcal{N}(t\mathbf{x} + (1-t)\mathbf{y}) \leq \mathcal{N}(t\mathbf{x}) + \mathcal{N}((1-t)\mathbf{y}) = t\mathcal{N}(\mathbf{x}) + (1-t)\mathcal{N}(\mathbf{y}).$$

— Les boules ouvertes et fermées sont des ensembles convexes. En effet, ces ensembles sont de la forme $\overline{\mathcal{B}}(0,E) = \left\{\mathbf{x} \in \mathbb{R}^d,\ \mathcal{N}(\mathbf{x}) \leq E\right\}$ (cf exercice précédent).

Exercice 2.24

Montrer que si K_1 et K_2 sont deux ensembles convexes, alors l'intersection $K_1 \cap K_2$ est encore convexe. Est-ce que l'union $K_1 \cup K_2$ est convexe ?

Théorème 2.25 : Unicité du minimiseur pour les fonctions strictement convexes

Soit $K \subset \mathbb{R}^d$ un ensemble convexe, et soit $f : K \to \mathbb{R}$ une fonction strictement convexe sur K. Alors, si f admet un minimiseur, celui-ci est unique.

Démonstration. Supposons que f admette deux minimiseurs $\mathbf{x}_0 \neq \mathbf{x}_1$ dans K. Donc $f(\mathbf{x}_0) = f(\mathbf{x}_1) = m_* := \inf_K f$. On considère le point $\mathbf{x}_* := \frac{1}{2}\mathbf{x}_0 + \frac{1}{2}\mathbf{x}_1$. Comme K est convexe, on a $\mathbf{x}_* \in K$, et comme f est strictement convexe, on a

$$f(\mathbf{x}_*) = f\left(\frac{1}{2}\mathbf{x}_0 + \frac{1}{2}\mathbf{x}_1\right) < \frac{1}{2}f(\mathbf{x}_0) + \frac{1}{2}f(\mathbf{x}_1) = m_*.$$

Le point $\mathbf{x}_*$ de K vérifie $f(\mathbf{x}_*) < m_*$. Ceci contredit la définition de m_* comme minimum. Donc $\mathbf{x}_0 = \mathbf{x}_1$, et le minimiseur est unique. ☐

Pour conclure ce chapitre, on retiendra que, pour une fonction continue à valeurs réelles,
— **coercive** $\Longrightarrow$ il existe au moins un minimiseur,
— **strictement convexe** $\Longrightarrow$ il existe au plus un minimiseur.

CHAPITRE 3

FONCTIONS DIFFÉRENTIABLES

3.1 Applications linéaires

3.1.1 L'espace vectoriel des applications linéaires bornées

Avant de définir la dérivée de fonctions à plusieurs variables, nous faisons un rappel sur les applications linéaires. Dans la suite, on considère $(E, \|\cdot\|_E)$ et $(F, \|\cdot\|_F)$ deux $\mathbb{R}$–espaces vectoriels normés.

Définition 3.1 : Application linéaire

Soit $L : E \to F$. On dit que L est une **application linéaire** de E dans F si :
- pour tous $\mathbf{x}, \mathbf{y} \in E$, on a $L(\mathbf{x}+\mathbf{y}) = L(\mathbf{x}) + L(\mathbf{y})$,
- pour tout $\lambda \in \mathbb{R}$ et tout $\mathbf{x} \in E$, on a $L(\lambda \mathbf{x}) = \lambda L(\mathbf{x})$.

On dit que cette application linéaire est **bornée** s'il existe $M > 0$ tel que pour tout $\mathbf{x} \in E$ avec $\|\mathbf{x}\|_E = 1$, on a $\|L(\mathbf{x})\|_F \leq M$.

On note $\mathcal{L}(E, F)$ l'ensemble des applications linéaires bornées de E dans F.

Exemples :
- L'application nulle $L(\mathbf{x}) = \mathbf{0}_F$ est linéaire. Elle est bornée, avec $M = 0$.
- L'identité $L : E \to E$ définie par $L(\mathbf{x}) = \mathbf{x}$ est linéaire, bornée, avec $M = 1$.
- Si $E = \mathbb{R}^n$ et $F = \mathbb{R}^m$. Alors pour toute matrice $A \in \mathcal{M}_{m,n}(\mathbb{R})$, l'application $L_A : \mathbf{x} \mapsto A\mathbf{x}$ est linéaire. Elle est bornée (cf plus tard).

Si L est bornée, on a que, pour tout $\mathbf{x} \in E \setminus \{\mathbf{0}\}$, en remarquant $\frac{\mathbf{x}}{\|\mathbf{x}\|_E}$ est de norme 1,

$$\|L(\mathbf{x})\|_F = \left\| L\left(\|\mathbf{x}\|_E \frac{\mathbf{x}}{\|\mathbf{x}\|_E} \right) \right\|_F = \|\mathbf{x}\|_E \left\| L\left(\frac{\mathbf{x}}{\|\mathbf{x}\|_E} \right) \right\|_F \leq M \|\mathbf{x}\|_E.$$

Ceci est vrai aussi pour $\mathbf{x} = \mathbf{0}$. Le raisonnement précédent qui consiste à remonter une inégalité sur la **sphère** de E à tout l'espace E s'appelle un argument **d'homogénéité**. Le plus petit $M > 0$ qui vérifie cette inégalité pour tout $\mathbf{x} \in E$ s'appelle la **norme d'opérateur** de L, notée $\|L\|_{\text{op}}$. Autrement dit, si $L \in \mathcal{L}(E, F)$,

$$\|L\|_{\text{op}} := \inf \left\{ \|L(\mathbf{x})\|_F, \quad \mathbf{x} \in E, \ \|\mathbf{x}\|_E = 1 \right\}.$$

On peut écrire l'inégalité utile suivante :

$$\boxed{\forall L \in \mathcal{L}(E, F), \quad \forall \mathbf{x} \in E, \quad \|L(\mathbf{x})\|_F \leq \|L\|_{\text{op}} \|\mathbf{x}\|_E.} \tag{3.1}$$

Théorème 3.2

L'espace $\mathcal{L}(E,F)$ muni de la norme $\|\cdot\|_{\mathrm{op}}$ est un espace vectoriel.

Démonstration. On vérifie directement que $\mathcal{L}(E,F)$ est un espace vectoriel, avec $(L_1+L_2)(\mathbf{x}) := L_1(\mathbf{x})+L_2(\mathbf{x})$, et $(\lambda L)(\mathbf{x}) := \lambda L(\mathbf{x})$. Montrons que $\|\cdot\|_{\mathrm{op}}$ est une norme. Pour commencer, si $\|L\|_{\mathrm{op}}=0$, alors l'inégalité (3.1) implique que $L(\mathbf{x})=\mathbf{0}$ pour tout $\mathbf{x}\in E$, donc L est l'application nulle. Ensuite, on vérifie directement que $\|\lambda L\|_{\mathrm{op}} = |\lambda|\,\|L\|_{\mathrm{op}}$. Il reste à vérifier l'inégalité triangulaire. On a, pour tout $\mathbf{x}\in E$ avec $\|\mathbf{x}\|_E=1$,

$$\|(L_1+L_2)(\mathbf{x})\|_F \le \|L_1(\mathbf{x})\|_F + \|L_2(\mathbf{x})\|_F \le \|L_1\|_{\mathrm{op}} + \|L_2\|_{\mathrm{op}}.$$

En prenant l'infimum sur $\mathbf{x}\in E$ avec $\|\mathbf{x}\|_E=1$, on obtient bien $\|L_1+L_2\|_{\mathrm{op}} \le \|L_1\|_{\mathrm{op}} + \|L_2\|_{\mathrm{op}}$. □

Théorème 3.3

Une application linéaire $L : E\to F$ est continue ssi elle est bornée.

Attention, ce théorème n'est vrai que pour des applications linéaires !

Démonstration. Supposons que L soit bornée. Soit $(\mathbf{x}_n)$ une suite de E qui converge vers $\mathbf{x}_*$ dans E. On a

$$\|L(\mathbf{x}_n)-L(\mathbf{x}_*)\|_F = \|L(\mathbf{x}_n-\mathbf{x}_*)\| \le \|L\|_{\mathrm{op}}\|\mathbf{x}_n-\mathbf{x}_*\| \xrightarrow[n\to\infty]{} 0.$$

Ceci montre que L est continue, en tant que fonction de E dans F (en fait, L est même 1–Lipschitz de E dans F).

Supposons au contraire que L ne soit pas bornée. Dans ce cas, il existe une suite $(\mathbf{x}_n)\in E$ avec $\|\mathbf{x}_n\|_E=1$ et $\|L(\mathbf{x}_n)\|_F\to\infty$. On pose $\mathbf{y}_n := \frac{\mathbf{x}_n}{\|L(\mathbf{x}_n)\|_F}$. On a $\|\mathbf{y}_n\|_E = \frac{1}{\|L(\mathbf{x}_n)\|_F}\to 0$, donc la suite $(\mathbf{y}_n)$ tend vers $\mathbf{0}_E$ dans E. En revanche, on a

$$\|L(\mathbf{y}_n)\|_F = \left\| L\left(\frac{\mathbf{x}_n}{\|L(\mathbf{x}_n)\|_F}\right)\right\|_F = \frac{1}{\|L(\mathbf{x}_n)\|_F}\|L(\mathbf{x}_n)\|_F = 1.$$

Or $\mathbf{y}_n\to\mathbf{0}_E$. On en déduit que L n'est pas continue en $\mathbf{0}_E$, donc a fortiori n'est pas continue. □

Remarque 3.4. *Le raisonnement précédent montre que si L est continue en $\mathbf{0}$, alors L est bornée, donc L est continue sur tout E. Ainsi, pour une application linéaire L, les trois assertions suivantes sont équivalentes :*

- *L est continue sur tout E,*
- *L est continue en $\mathbf{0}$,*
- *L est bornée.*

3.1.2 Le cas de la dimension finie

Dans le cas où $E=\mathbb{R}^d$ et $F=\mathbb{R}^m$ sont de dimension finie, on peut dire plus de choses.

Théorème 3.5

Une application linéaire $L:\mathbb{R}^d\to\mathbb{R}^m$ est automatiquement bornée, donc continue.

Démonstration. Notons $(\mathbf{e}_1, \cdots, \mathbf{e}_d)$ la base canonique de $\mathbb{R}^d$. On a, pour $\mathbf{x} = \sum_{j=1}^d x_j \mathbf{e}_j \in \mathbb{R}^d$,

$$\| L(\mathbf{x}) \|_{\mathbb{R}^m} = \left\| \sum_{j=1}^d x_j L(\mathbf{e}_j) \right\|_{\mathbb{R}^m} \le \sum_{j=1}^d |x_j| \, \| L(\mathbf{e}_j) \|_{\mathbb{R}^m} \le \left(\sum_{j=1}^d |x_j|^2 \right)^{1/2} \left(\sum_{j=1}^d \| L(\mathbf{e}_j) \|_{\mathbb{R}^m}^2 \right)^{1/2}.$$

On a utilisé l'inégalité de Cauchy-Schwarz pour la dernière inégalité. On reconnaît la norme de $\mathbf{x}$ dans le membre de droite. Ainsi,

$$\forall \mathbf{x} \in \mathbb{R}^d, \ \| \mathbf{x} \|_{\mathbb{R}^d} = 1, \qquad \| L(\mathbf{x}) \|_{\mathbb{R}^m} \le M, \quad \text{avec} \quad M = \left(\sum_{j=1}^d \| L(\mathbf{e}_j) \|_{\mathbb{R}^m}^2 \right)^{1/2} < \infty.$$

□

Représentation matricielle

Soit $(\mathbf{e}_1, \cdots, \mathbf{e}_d)$ la base canonique de $E = \mathbb{R}^d$, et soit $(\mathbf{f}_1, \cdots, \mathbf{f}_m)$ celle de $F = \mathbb{R}^m$ (en fait, le raisonnement suivant marche avec n'importe quelle base).

Pour tout $1 \le j \le d$, le vecteur $L(\mathbf{e}_j)$ appartient à F, et peut être décomposé dans la base $(\mathbf{f}_1, \cdots, \mathbf{f}_m)$. On note a_{ij} les coefficients correspondants, c'est à dire

$$L(\mathbf{e}_j) = \sum_{i=1}^m a_{ij} \mathbf{f}_i \in F \rightsquigarrow \begin{pmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{mj} \end{pmatrix} \in \mathbb{R}^m.$$

Soit $A = (a_{ij})_{1 \le i \le m, 1 \le j \le d}$ la matrice dont les vecteurs colonnes sont les $L(\mathbf{e}_i)$ pour $1 \le i \le d$, c'est à dire

$$A = (L(\mathbf{e}_1), L(\mathbf{e}_2), \cdots, L(\mathbf{e}_d)) = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1d} \\ a_{21} & a_{22} & \cdots & a_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{md} \end{pmatrix} \in \mathcal{M}_{m \times d}(\mathbb{R}).$$

On dit que A est la **matrice représentative de L dans les bases $(\mathbf{e}_1, \cdots, \mathbf{e}_d)$ et $(\mathbf{f}_1, \cdots, \mathbf{f}_m)$**. On notera en effet que la matrice A dépend du choix de la base choisie (alors que l'application L n'en dépend pas).

L'importance de cette matrice, et du calcul matriciel en général, vient du fait que, pour tout $\mathbf{x} = \sum x_j \mathbf{e}_j$, on a

$$L(\mathbf{x}) = L\left(\sum_{j=1}^d x_j \mathbf{e}_j \right) = \sum_{j=1}^d x_j L(\mathbf{e}_j) = \sum_{j=1}^d \sum_{i=1}^m a_{ij} x_j \mathbf{f}_i = \sum_{i=1}^m \left(\sum_{j=1}^d a_{ij} x_j \right) \mathbf{f}_i.$$

Autrement dit, si $\mathbf{y} = L(\mathbf{x})$, alors les coefficients y_j de $\mathbf{y}$ (dans la base canonique de F) sont obtenus à partir des coefficients x_i de $\mathbf{x}$ (dans la base canonique de E) via la formule

$$y_i = \sum_{j=1}^d a_{ij} x_j, \quad \text{ou encore,} \quad \mathbf{y} = A\mathbf{x}.$$

où la dernière égalité est la multiplication matrice–vecteur de A par $\mathbf{x}$. On remarquera qu'il y a un léger abus de notation ici. En effet, l'égalité $\mathbf{y} = L(\mathbf{x})$ ne dépend pas de la base choisie ($\mathbf{x} \in E$ et

$\mathbf{y} \in F$ sont des vecteurs abstraits), alors que dans l'égalité $\mathbf{y} = A\mathbf{x}$, on a posé

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix} \in \mathbb{R}^d, \quad \text{et} \quad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix} \in \mathbb{R}^m,$$

et cette notation vectorielle dépend du choix de la base de $E = \mathbb{R}^d$ et de $F = \mathbb{R}^m$. Dans la suite, on fera cet abus de notation, avec la convention qu'on travaille toujours dans la base canonique.

On en déduit le résultat suivant.

Théorème 3.6

L'espace vectoriel $\mathcal{L}(\mathbb{R}^d, \mathbb{R}^m)$ est en bijection avec l'espace vectoriel $\mathcal{M}_{m\times d}(\mathbb{R})$, qui est en bijection avec $\mathbb{R}^{m\times d}$. En particulier, $\mathcal{L}(\mathbb{R}^d, \mathbb{R}^m)$ est de dimension finie, de dimension $d \times m$, et toutes ses normes sont équivalentes.

3.2 Fonctions différentiables

3.2.1 Retour en dimension $d = 1$

En dimension $d = 1$, on rappelle la définition de la dérivabilité d'une fonction. On dit qu'une fonction $f : \mathbb{R} \to \mathbb{R}$ est **dérivable** en $x_0 \in \mathbb{R}$ si la limite

$$\lim_{h\to 0} \frac{f(x_0 + h) - f(x_0)}{h}$$

existe. Si c'est le cas, on note $f'(x_0) \in \mathbb{R}$ cette limite. Malheureusement, cette définition ne peut pas s'étendre au cas des dimensions supérieures. En effet, on aurait naturellement $\mathbf{x}_0 \in \mathbb{R}^d$ et $\mathbf{h} \in \mathbb{R}^d$, or on ne peut pas diviser par un vecteur $\mathbf{h}$...

Pour contourner cette difficulté, on réécrit la définition précédente d'une autre manière. Posons

$$\mathcal{E}(h) := \frac{f(x_0 + h) - f(x_0)}{h} - f'(x_0).$$

Dire que le ratio converge vers $f'(x)$ lorsque $h \to 0$, c'est dire que $\lim_{h\to 0} E(h) = 0$. Ainsi, on a, en multipliant par h et en ré-ordonnant les termes,

$$f(x_0 + h) = f(x_0) + f'(x_0)h + h\mathcal{E}(h).$$

On remarque que le dernier terme est un $o(h)$ (cf plus tard pour la définition de $o(\cdot)$). Ainsi, la définition de la dérivabilité en dimension 1 est équivalente à dire que f admet un **développement limité**, ou **développement de Taylor**, à l'ordre 1, de la forme

$$f(x_0 + h) = f(x_0) + f'(x_0)h + o(h).$$

En remarquant que la fonction $h \mapsto f'(x_0)h$ est une fonction linéaire de $\mathbb{R}$ dans $\mathbb{R}$, on comprend comment étendre la définition aux dimensions supérieures.

3.2.2 Définition de la différentiabilité

On se place dans le cadre général des espaces vectoriels normés $(E, \|\cdot\|_E)$ et $(F, \|\cdot\|_F)$ pour énoncer la définition. Dans la suite, on écrira $f : U \subset E \to F$ pour indiquer que U est un ouvert de E et que f est définie sur U (donc $U \subset \mathcal{D}_f \subset E$).

Définition 3.7 : Différentiabilité en un point.

Soit $f : U \subset E \to F$, et soit $\mathbf{x}_0 \in U$. On dit que f est **différentiable** en $\mathbf{x}_0$ s'il existe une application linéaire bornée $\mathcal{L}(E, F)$ telle que

$$f(\mathbf{x}_0 + \mathbf{h}) = f(\mathbf{x}_0) + L(\mathbf{h}) + o(\mathbf{h}).$$

Si c'est le cas, L est unique, est notée $\mathrm{d}f_{\mathbf{x}_0}$, et est appelée **différentielle** de f en $\mathbf{x}_0$.
Si f est différentiable en tout point de U, on dit que que f est différentiable sur U.

La formule

$$f(\mathbf{x}_0 + \mathbf{h}) = f(\mathbf{x}_0) + \mathrm{d}f_{\mathbf{x}_0}(\mathbf{h}) + o(\mathbf{h})$$

s'appelle parfois la **formule de Taylor** (à l'ordre 1).

Définition 3.8 : Notation de Landau

Une fonction $o(\mathbf{h})$ est une fonction de la forme $\|\mathbf{h}\|\mathcal{E}(\mathbf{h})$ avec $\mathcal{E}(\mathbf{h}) \to \mathbf{0}$ lorsque $\mathbf{h} \to \mathbf{0}$.

Une fonction $\mathcal{E}$ vérifie $\mathcal{E}(\mathbf{h}) \to \mathbf{0}$ lorsque $h \to \mathbf{0}$ ssi elle est continue en $\mathbf{0}$ avec $\mathcal{E}(\mathbf{0}) = \mathbf{0}$.

On insiste sur le fait que $\mathrm{d}f_{\mathbf{x}_0}$ est une application linéaire.

$\star$ **Exemple (fonctions de $\mathbb{R}$ dans $\mathbb{R}$).** Dans l'exemple précédent $d = 1$, la différentielle de $f : \mathbb{R} \to \mathbb{R}$ au point $x_0 \in R$ est l'application linéaire $\mathrm{d}f_{x_0} \in \mathcal{L}(\mathbb{R}, \mathbb{R})$ définie par

$$\forall h \in \mathbb{R}, \quad \mathrm{d}f_{x_0}(h) := f'(x_0)h.$$

En quelque sorte, la **dérivée** $f'(x_0)$ est le coefficient réel qui représente la **différentielle** $\mathrm{d}f_{x_0}$.

$\star$ **Exemple (fonctions linéaires) :** Si $L \in \mathcal{L}(E, F)$ est une application linéaire bornée/continue, alors L est différentiable sur E, avec, pour tout $\mathbf{x}_0 \in E$, $\mathrm{d}L_{\mathbf{x}_0} = L$. En effet, on a

$$L(\mathbf{x}_0 + \mathbf{h}) = L(\mathbf{x}_0) + L(\mathbf{h}) \qquad + o(\mathbf{h}), \quad \text{avec} \quad o(\mathbf{h}) = \mathbf{0}.$$

$\star$ **Exemple (forme quadratique) :** Soit $A \in \mathcal{M}_{d,d}(\mathbb{R})$, et $f : \mathbb{R}^d \to \mathbb{R}$ définie par $f(\mathbf{x}) := \langle \mathbf{x}, A\mathbf{x} \rangle$ (produit scalaire Euclidien). Calculons la différentielle de f. On a

$$f(\mathbf{x} + \mathbf{h}) = \langle \mathbf{x} + \mathbf{h}, A(\mathbf{x} + \mathbf{h}) \rangle = \langle \mathbf{x}, A\mathbf{x} \rangle + \langle \mathbf{x}, A\mathbf{h} \rangle + \langle \mathbf{h}, A\mathbf{x} \rangle + \langle \mathbf{h}, A\mathbf{h} \rangle.$$

Le premier terme est $f(\mathbf{x})$. Le second et le troisième sont linéaires en $\mathbf{h}$ (donc compose la différentielle). Montrons que le dernier terme est un $o(\mathbf{h})$. On pose

$$\mathcal{E}(\mathbf{h}) := \begin{cases} \frac{\langle \mathbf{h}, A\mathbf{h} \rangle}{\|\mathbf{h}\|} & \text{si} \quad \mathbf{h} \neq \mathbf{0} \\ 0 & \text{si} \quad \mathbf{h} = \mathbf{0}. \end{cases},$$

de sorte que $\langle \mathbf{h}, A\mathbf{h} \rangle = \|\mathbf{h}\|\mathcal{E}(\mathbf{h})$. Montrons que $\mathcal{E}$ est continue en $\mathbf{0}$ avec $\mathcal{E}(\mathbf{0}) = 0$. Pour cela, on utilise l'inégalité de Cauchy-Schwarz et l'inégalité de la norme d'opérateur de A, et on a

$$|\langle \mathbf{h}, A\mathbf{h} \rangle| \leq \|\mathbf{h}\| \cdot \|A\mathbf{h}\| \leq \|\mathbf{h}\| \cdot \|A\|_{\mathrm{op}} \cdot \|\mathbf{h}\| = \|A\|_{\mathrm{op}} \|\mathbf{h}\|^2.$$

Ceci montre que $|\mathcal{E}(\mathbf{h})| \leq \|A\|_{\mathrm{op}} \|\mathbf{h}\|$, donc $\mathcal{E}$ est bien continue en $\mathbf{0}$ avec $\mathcal{E}(\mathbf{0}) = 0$. Ainsi, notre développement précédent est de la forme $f(\mathbf{x} + \mathbf{h}) = f(\mathbf{x}) + \mathrm{d}f_{\mathbf{x}}(\mathbf{h}) + o(\mathbf{h})$ avec, *par identification*,

$$\mathrm{d}f_{\mathbf{x}} : \mathbb{R}^d \to \mathbb{R}, \quad \text{définie par} \quad \mathrm{d}f_{\mathbf{x}}(\mathbf{h}) := \langle \mathbf{x}, A\mathbf{h} \rangle + \langle \mathbf{h}, A\mathbf{x} \rangle.$$

Théorème 3.9 : Différentiabilité implique continuité

Si $f : U \subset E \to F$ est différentiable en $\mathbf{x}_0 \in E$, alors f est continue en $\mathbf{x}_0$.

Démonstration. Soit $(\mathbf{x}_n)$ une suite qui converge vers $\mathbf{x}_0$. Posons $\mathbf{h}_n := \mathbf{x}_n - \mathbf{x}_0$, de sorte que $\mathbf{x}_n = \mathbf{x}_0 + \mathbf{h}_n$. La suite $\mathbf{h}_n$ tend vers $\mathbf{0}$ dans E. On a

$$\|f(\mathbf{x}_0+\mathbf{h}_n) - f(\mathbf{x}_0)\|_F = \|\mathrm{d}f_{\mathbf{x}_0}(\mathbf{h}_n) + \|\mathbf{h}_n\|_E \mathcal{E}(\mathbf{h}_n)\|_F \le \|\mathrm{d}f_{\mathbf{x}_0}\|_{\mathrm{op}} \|\mathbf{h}_n\|_E + \|\mathbf{h}_n\|_E \|\mathcal{E}(\mathbf{h}_n)\|_F.$$

Comme $\mathcal{E}$ est continue en $\mathbf{0}$ avec $\mathcal{E}(\mathbf{0}_E) = \mathbf{0}_F$, et que $\mathbf{h}_n \to \mathbf{0}$, le terme de droite converge vers 0 lorsque $n \to \infty$. Ainsi, on a $f(\mathbf{x}_n) \to f(\mathbf{x}_0)$, comme souhaité. □

3.2.3 Dérivées directionnelles et dérivées partielles

Soit $f : U \subset E \to F$, soit $\mathbf{x}_0 \in U$ et soit $\mathbf{h} \in E$ une **direction** quelconque de E. On s'intéresse à la fonction réelle, $g_{\mathbf{x}_0,\mathbf{h}} : \mathbb{R} \to F$ définie par

$$g_{\mathbf{x}_0,\mathbf{h}}(t) := f(\mathbf{x}_0 + t\mathbf{h}).$$

Cette fonction est bien définie dans un voisinage de $t = 0$. On peut dériver g, en tant que fonction ayant une seule variable. Cela motive la définition suivante.

Définition 3.10

Si $g_{\mathbf{x}_0,\mathbf{h}}$ est dérivable en $t = 0$, on dit que f admet une **dérivée directionnelle** au point $\mathbf{x}_0 \in U$, dans la direction $\mathbf{h} \in E$. On la note

$$\partial_{\mathbf{h}} f(\mathbf{x}_0) := \frac{\partial f}{\partial \mathbf{h}}(\mathbf{x}_0) := g'_{\mathbf{x}_0,\mathbf{h}}(0) = \lim_{t\to 0} \frac{f(\mathbf{x}_0 + t\mathbf{h}) - f(\mathbf{x}_0)}{t} \qquad \in F.$$

Dans le cas important où $E = \mathbb{R}^d$ et $\mathbf{h} = \mathbf{e}_i$ est le i-ème vecteur de la base canonique de $\mathbb{R}^d$, on parle plutôt de **dérivée partielle**. On écrit dans ce cas

$$\partial_i f(\mathbf{x}_0) := \frac{\partial f}{\partial x_i}(\mathbf{x}_0) := \frac{\partial f}{\partial \mathbf{e}_i}(\mathbf{x}_0) := \lim_{t\to 0} \frac{f(x_1, \cdots, x_{i-1}, x_i + t, x_{i+1}, \cdots, x_d) - f(x_1, \cdots, x_d)}{t}.$$

En pratique, pour calculer cette dérivée, on voit les variables $x_1, x_2, \cdots, x_{i-1}, x_{i+1}, \cdots, x_d$ comme fixées, et on dérive “normalement” dans la variable x_i. Par exemple, pour la fonction $f(x_1, x_2) = x_1 \mathrm{e}^{-x_2}$, on a

$$(\partial_1 f)(x_1, x_2) = \mathrm{e}^{-x_2} \quad \text{et} \quad (\partial_2 f)(x_1, x_2) = -x_1 \mathrm{e}^{-x_2}.$$

Théorème 3.11

Si f est différentiable en $\mathbf{x}_0$, alors, pour tout $\mathbf{h} \in E$, on a

$$\frac{\partial f}{\partial \mathbf{h}}(\mathbf{x}_0) = \mathrm{d}f_{\mathbf{x}_0}(\mathbf{h}) \qquad \in F.$$

Démonstration. On utilise la formule de Taylor pour f. On a

$$f(\mathbf{x}_0 + t\mathbf{h}) = f(\mathbf{x}_0) + \mathrm{d}f_{\mathbf{x}_0}(t\mathbf{h}) + o(t\mathbf{h}) = f(\mathbf{x}_0) + t\mathrm{d}f_{\mathbf{x}_0}(\mathbf{h}) + |t| \cdot \|\mathbf{h}\| \mathcal{E}(t\mathbf{h})$$

pour une certaine fonction $\mathcal{E}$ qui tend vers 0 en 0. Ainsi,

$$\frac{f(\mathbf{x}_0 + t\mathbf{h}) - f(\mathbf{x}_0)}{t} = \frac{t\mathrm{d}f_{\mathbf{x}_0}(\mathbf{h}) + |t| \cdot \|\mathbf{h}\| \mathcal{E}(t\mathbf{h})}{t} = \mathrm{d}f_{\mathbf{x}_0}(\mathbf{h}) \pm \mathcal{E}(t\mathbf{h}) \xrightarrow[t\to 0]{} \mathrm{d}f_{\mathbf{x}_0}(\mathbf{h}).$$

□

Ainsi, si f est différentiable en $\mathbf{x}_0$, alors toutes les dérivées directionnelles $\partial_{\mathbf{h}} f$ existent en $\mathbf{x}_0$. La réciproque est fausse en général, comme le montre l'exemple suivant.

Exemple 3.12. *Reprenons la fonction* $f(x,y) := \frac{xy}{x^2+y^2}$ *vue dans l'Exercice 2.7. Pour tout* $x \in \mathbb{R}$, *on a*

$$\frac{\partial f}{\partial x}(x,0) = \lim_{h\to 0} \frac{f(x+h,0)-f(x,0)}{h} = 0,$$

et de même $\frac{\partial f}{\partial y}(0,y) = 0$. *En particulier, on a* $\frac{\partial f}{\partial x}(0,0) = \frac{\partial f}{\partial y}(0,0) = 0$, *mais* f *n'est pas différentiable en* $(0,0)$. *Elle n'est même pas continue en* $(0,0)$ *(cf Exercice 2.7).*

3.2.4 Matrice Jacobienne

On se place maintenant dans le cas où $E = \mathbb{R}^d$ et $F = \mathbb{R}^m$ (le cas utile en pratique). Dans ce cas, l'application $\mathrm{d}f_{\mathbf{x}_0} \in \mathcal{L}(E,F)$ peut être représentée par une matrice de $\mathcal{M}_{m\times d}(\mathbb{R})$, dans les bases canoniques.

Définition 3.13 : Jacobienne

La matrice de $\mathcal{M}_{m\times d}(\mathbb{R})$ qui représente $\mathrm{d}f_{\mathbf{x}_0}(\mathbf{h})$ dans les bases canoniques s'appelle **matrice Jacobienne**, et est notée

$$J_f(\mathbf{x}_0) \in \mathcal{M}_{m\times d}(\mathbb{R}).$$

Par définition de la matrice représentative, c'est la matrice telle que

$$\forall \mathbf{h} \in \mathbb{R}^d, \quad \mathrm{d}f_{\mathbf{x}_0}(\mathbf{h}) = J_f(\mathbf{x}_0)\mathbf{h} \quad \in \mathbb{R}^d.$$

Dans le membre de gauche, on a une application linéaire $\mathrm{d}f_{\mathbf{x}_0}$ qui agit sur le vecteur $\mathbf{h} \in E$, et dans le membre de droite, on a la multiplication de la matrice $J_f(\mathbf{x}_0)$ et du vecteur $\mathbf{h} \in \mathbb{R}^d$.

Dans la suite on note

$$f(\mathbf{x}) = \begin{pmatrix} f_1(x_1,\cdots,x_d) \\ f_2(x_1,\cdots,x_d) \\ \vdots \\ f_m(x_1,\cdots,x_d) \end{pmatrix},$$

les composantes de f dans la base canonique de $F = \mathbb{R}^m$.

Théorème 3.14 : Identification de la Jacobienne

Si $f : U \subset \mathbb{R}^d \to \mathbb{R}^m$ est différentiable en $\mathbf{x}_0 \in U$, alors les coefficients de la matrice Jacobienne $J_f(\mathbf{x}_0)$ sont les dérivées partielles des f_i. Plus exactement,

$$J_f(\mathbf{x}_0) = \left(\partial_{x_1} f, \cdots, \partial_{x_d} f\right) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_d} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_d} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_d} \end{pmatrix} (\mathbf{x}_0) \qquad \in \mathcal{M}_{m\times d}(\mathbb{R}).$$

La preuve du théorème découle directement de la définition de $\partial_{x_i} f$.

Remarque 3.15. *Les composantes de* f *ne se parlent pas. Chaque composante "vit" dans sa ligne ! On raisonne* composante par composante.

3.2.5 Le gradient

On se place enfin dans le cas où $F = \mathbb{R}$, c'est à dire dans le cas où $f = f(x_1,\cdots,x_d)$ est une fonction à valeurs réelles. On rappelle que ce sont les fonctions qu'on peut optimiser.

Dans ce cas, f n'a qu'une seule composante, et on a

$$J_f(\mathbf{x}_0) = \big(\partial_1 f, \cdots, \partial_d f\big) \quad \in \mathcal{M}_{1,d}(\mathbb{R}).$$

La Jacobienne est donc un **vecteur ligne**.

Définition 3.16 : Gradient

Soit $f : U \subset \mathbb{R}^d \to \mathbb{R}$ une fonction réelle différentiable en $\mathbf{x}_0 \in U$. Le **gradient** de f en $\mathbf{x}_0$ est le **vecteur** de $\mathbb{R}^d$, noté $\nabla f(\mathbf{x}_0)$ dont les composantes sont

$$\nabla f(\mathbf{x}_0) := \begin{pmatrix} \partial_1 f \\ \partial_2 f \\ \vdots \\ \partial_d f \end{pmatrix} (\mathbf{x}_0).$$

Avec cette définition, on remarque que le gradient est la transposée de la Jacobienne. L'importance de ce vecteur vient de la constatation suivante. On rappelle que le **produit scalaire** de $\mathbb{R}^d$ est défini par

$$\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d, \qquad \langle \mathbf{x}, \mathbf{y} \rangle := \sum_{i=1}^d x_i y_i = \begin{pmatrix} x_1 & x_2 & \cdots & x_d \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_d \end{pmatrix} = \mathbf{x}^T \mathbf{y} \qquad \in \mathbb{R}.$$

Alors on peut écrire

$$\mathrm{d}f_{\mathbf{x}_0}(\mathbf{h}) = J_f(\mathbf{x}_0)\mathbf{h} = \langle \nabla f(\mathbf{x}_0), \mathbf{h} \rangle.$$

L'avantage de la dernière expression est qu'elle est de nouveau indépendante du choix de la base de $\mathbb{R}^d$. Le vecteur gradient est un vecteur de $\mathbb{R}^d$ bien défini (qui se trouve avoir les $\partial_i f$ comme coefficients dans la base canonique). En fait, on peut retenir que

Le gradient de f en $\mathbf{x}_0$ est le vecteur qui pointe vers la plus grande pente montante de f.

Cela vient de l'inégalité de Cauchy-Schwarz :

$$f(\mathbf{x}_0 + \mathbf{h}) = f(\mathbf{x}_0) + \langle \nabla f(\mathbf{x}_0), \mathbf{h} \rangle + o(\mathbf{h}) \le f(\mathbf{x}_0) + \| \nabla f(\mathbf{x}_0) \|_{\mathbb{R}^d} \cdot \| \mathbf{h} \|_{\mathbb{R}^d} + o(\mathbf{h}),$$

avec égalité ssi $\mathbf{h} = \alpha \nabla f(\mathbf{x}_0)$, avec $\alpha > 0$. Nous reviendrons sur ce point important lorsqu'on optimisera des fonctions.

3.2.6 Pour aller encore plus loin (*)

Une autre façon d'écrire les choses est la suivante.

Définition 3.17 : Base duale

Pour $1 \le i \le d$, on pose $\mathrm{d}x_i : \mathbb{R}^d \to \mathbb{R}$ l'application linéaire définie par

$$\mathrm{d}x_i(\mathbf{x}) = x_i.$$

La famille $(\mathrm{d}x_1, \cdots, \mathrm{d}x_d)$, qui est une famille de $\mathcal{L}(\mathbb{R}^d, \mathbb{R})$, s'appelle parfois la **base duale** de $(\mathbf{e}_1, \cdots, \mathbf{e}_d)$.

Autrement dit, $\mathrm{d}x_i$ est l'application linéaire qui prend un vecteur $\mathbf{x} \in \mathbb{R}^d$ en entrée, et retourne sa i-ème composante réelle $x_i \in \mathbb{R}$.

Avec la notation de la base duale, on peut écrire

$$\boxed{\mathrm{d}f_{\mathbf{x}_0} = \frac{\partial f}{\partial x_1}\mathrm{d}x_1 + \cdots + \frac{\partial f}{\partial x_d}\mathrm{d}x_d.}$$

Il s'agit d'une égalité au sens de $\mathcal{L}(\mathbb{R}^d, \mathbb{R}^m)$, c'est à dire au sens des applications linéaires :

$$\forall \mathbf{h} \in \mathbb{R}^m, \qquad \mathrm{d}f_{\mathbf{x}_0}(\mathbf{h}) = \frac{\partial f}{\partial x_1}h_1 + \cdots + \frac{\partial f}{\partial x_d}h_d \qquad \in \mathbb{R}^m.$$

Remarque 3.18. *En dimension $d = 1$, on a $\mathrm{d}f = f'(x)\mathrm{d}x$. Le $\mathrm{d}x$ qui apparaît est celui dans les intégrales. En fait, la notation $\int f'(x)\mathrm{d}x$ signifie $\int \mathrm{d}f_{\mathbf{x}}$: on intégre des* **formes différentielles**.

3.2.7 Règle de la chaîne

Théorème 3.19 : Règle de la chaîne

Soient $f : U \subset E \to F$ et $g : V \subset F \to G$ deux fonctions, soit $\mathbf{x}_0 \in U$ telle que f est différentiable en $\mathbf{x}_0$, et g est différentiable en $f(\mathbf{x}_0) \in V$. Alors $g \circ f : U \subset E \to G$ est différentiable en $\mathbf{x}_0$, et

$$\mathrm{d}(g \circ f)(\mathbf{x}_0) = \mathrm{d}g_{f(\mathbf{x}_0)} \circ \mathrm{d}f_{\mathbf{x}_0}.$$

Dans le cas où $E = \mathbb{R}^d$, $F = \mathbb{R}^m$ et $G = \mathbb{R}^p$, on peut écrire de manière équivalente le produit matriciel des Jacobiennes

$$J_{g \circ f}(\mathbf{x}_0) = J_g(f(\mathbf{x}_0)) \cdot J_f(\mathbf{x}_0).$$

La composée à droite est la composée d'applications linéaires. On note que $\mathrm{d}f_{\mathbf{x}_0} \in \mathcal{L}(E, F)$ et que $\mathrm{d}g_{f(\mathbf{x}_0)} \in \mathcal{L}(F, G)$, donc la composée $\mathrm{d}g_{f(\mathbf{x}_0)} \circ \mathrm{d}f_{\mathbf{x}_0}$ est bien une application linéaire bornée de $\mathcal{L}(E, G)$.

Remarque 3.20. *Les matrices ne commutent pas (ni les applications linéaires). L'ordre est important dans ce théorème. Lorsque f et g sont des fonctions de $\mathbb{R}$ dans $\mathbb{R}$, on a*

$$(g \circ f)'(x) = g'(f(x)) \cdot f'(x) = f'(x) \cdot g'(f(x)),$$

mais dans le cas général, seule la première expression est valide.

La démonstration de ce résultat est immédiate, il suffit d'écrire les développements de Taylor.

3.3 Fonctions continûment différentiables

Nous avons vu la notion de fonctions différentiables sur un ouvert $U \subset E$. Nous nous demandons maintenant quand ces différentielles sont continues.

Définition 3.21

Soit $f : U \subset E \to F$ une fonction différentiable sur u, et soit $U \ni \mathbf{x} \mapsto \mathrm{d}f_{\mathbf{x}} \in \mathcal{L}(E, F)$ sa différentielle. On dit que f est **continûment différentiable**, ou **de classe** C^1, sur U si l'application $\mathbf{x} \mapsto \mathrm{d}f_{\mathbf{x}}$ est continue, en tant qu'application de $U \subset E$ dans $\mathcal{L}(E, F)$.
On note $C^1(U, F)$ l'ensemble des fonctions continûment différentiable sur U.

Remarquons que l'expression $\mathrm{d}f_{\mathbf{x}}(\mathbf{h})$ dépend à la fois de $\mathbf{x}$ et de $\mathbf{h}$. Dans les section précédentes, nous avons vu que à $\mathbf{x}$ fixé, l'application $\mathbf{h} \mapsto \mathrm{d}f_{\mathbf{x}}(\mathbf{h})$ était continue (elle était même linéaire !). Maintenant, on regarde plutôt l'application $\mathbf{x} \mapsto \mathrm{d}f_{\mathbf{x}}(\mathbf{h})$, qui n'a *a priori* aucune raison d'être continue (et encore moins linéaire).

$\star$ **Exemple (fonctions de $\mathbb{R}$ dans $\mathbb{R}$.** Dans le cas des fonctions $f : \mathbb{R} \to \mathbb{R}$, on a vu que la différentielle était $\mathrm{d}f_x(h) = f'(x)h$. Cette expression est continue et linéaire en h. Cependant, elle est continue en x ssi $x \mapsto f'(x)$ est continue. On retrouve la notion "classique".

3.3.1 Théorème «fondamental» de l'analyse

On rappelle le résultat suivant, valide pour une fonction $f : \mathbb{R} \to \mathbb{R}$ de classe C^1.

Théorème 3.22 : Théorème fondamental de l'analyse, première forme

Soit $f \in C^1(\mathbb{R}, \mathbb{R})$. Pour tout $a < b$, on a

$$f(b) - f(a) = \int_a^b f'(s)\mathrm{d}s.$$

Notons que comme f est de classe C^1, on a f' continue, et l'intégrale est bien définie (intégrale de Riemann pour les fonctions continues).

Nous en donnons une deuxième forme, équivalente, et très pratique aussi.

Théorème 3.23 : Théorème fondamentale de l'analyse, deuxième forme

Soit $f \in C^1(\mathbb{R}, \mathbb{R})$. Pour tout $x \in \mathbb{R}$ et tout $h \in \mathbb{R}$, on a

$$f(x+h) - f(x) = h \int_0^1 f'(x+th)\mathrm{d}t.$$

Démonstration. On applique le théorème précédent avec $a = x$ et $b = x + h$, et on a

$$f(x+h) - f(x) = \int_x^{x+h} f'(s)\mathrm{d}s.$$

On fait ensuite le changement de variable $s = s(t) = x + th$, donc $\mathrm{d}s = h\mathrm{d}t$. □

On étend maintenant ce théorème aux fonctions à plusieurs variables. Soit $f : C^1(U, F)$ une fonction de classe C^1 sur $U \subset E$. Alors pour $\mathbf{x} \in U$ et $\mathbf{h} \in E$, on introduit la fonction $g : \mathbb{R} \to F$, définie par

$$g_{\mathbf{x},\mathbf{h}}(t) := f(\mathbf{x} + t\mathbf{h}).$$

Comme f est de classe C^1, on a g de classe C^1. En appliquant le théorème fondamental de l'analyse à la fonction g, on obtient

$$g_{\mathbf{x},\mathbf{h}}(1) - g_{\mathbf{x},\mathbf{h}}(0) = \int_0^1 g'_{\mathbf{x},\mathbf{h}}(\mathbf{x}+t\mathbf{h})\mathrm{d}t = \int_0^1 \mathrm{d}f_{\mathbf{x}+t\mathbf{h}} \cdot \mathbf{h}\,\mathrm{d}t,$$

où on a utilisé le Théorème 3.11 pour la dernière égalité. On en déduit le théorème suivant.

Théorème 3.24 : Théorème fondamental de l'analyse

Soit $f \in C^1(U, F)$ avec $U \subset E$ ouvert convexe. Pour tout $x \in U$ et tout $\mathbf{h} \in E$ tel que $\mathbf{x}+\mathbf{h} \in U$, on a

$$f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) = \left(\int_0^1 \mathrm{d}f_{\mathbf{x}+t\mathbf{h}}\mathrm{d}t\right) \cdot \mathbf{h}.$$

Dans le cas où $E = \mathbb{R}^d$ et $F = \mathbb{R}^m$, on peut aussi écrire

$$f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) = \left(\int_0^1 J_f(\mathbf{x}+t\mathbf{h})\mathrm{d}t\right) \times \mathbf{h}.$$

Enfin, dans le cas où $m = 1$ (f est une fonction réelle), on a aussi

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

On rappelle qu'on peut intégrer des matrices et des vecteurs, en intégrant "composante par composant".

3.3.2 Théorème des accroissements finis

On commence par rappeler le théorème de la moyenne.

Lemme 3.25. *Si $g \in C^0(\mathbb{R}, \mathbb{R})$ est continue à valeurs* **réelles***, alors pour tout $a < b$, il existe $c \in [a, b]$ telle que*

$$\frac{1}{b-a}\int_a^b g(s)\mathrm{d}s = g(c).$$

Démonstration. La fonction g est continue sur le compact $[a, b]$, donc est atteint son minimum m et son maximum M en x_m et x_M respectivement, avec $x_m, x_M \in [a, b]$. On a $m \le g(x) \le M$ pour tout $x \in [a, b]$, donc

$$m \le \frac{1}{|b-a|}\int_a^b g(s)\mathrm{d}s \le M.$$

D'après le théorème des valeurs intermédiaires pour les fonctions continues, il existe un point $c \in [x_m, x_M]$ telle que

$$\frac{1}{|b-a|}\int_a^b g(s)\mathrm{d}s = g(c),$$

ce qui conclut la preuve. ☐

En prenant $g(t) = \langle \nabla f(\mathbf{x} + t\mathbf{h}), \mathbf{h} \rangle$ dans le troisième cas du Théorème 3.24, on obtient le résultat suivant.

Théorème 3.26 : Théorème des accroissements finis dans $\mathbb{R}^d$

Soit $f : \mathbb{R}^d \to \mathbb{R}$ une fonction réelle (m = 1) de classe C^1. Alors, pour tout $\mathbf{x} \in \mathbb{R}^d$ et $\mathbf{h} \in \mathbb{R}^d$, il existe $t_* \in [0, 1]$ telle que

$$f(\mathbf{x} + \mathbf{h}) - f(\mathbf{x}) = \langle \nabla f(\mathbf{x} + t_*\mathbf{h}), \mathbf{h} \rangle.$$

Il est important ici que la fonction soit réelle, c'est à dire qu'elle n'a qu'une seule composante. Un résultat similaire pour des fonctions f avec plusieurs composantes ne peut pas être vraie. Si $f = (f_1, f_2)$ est une fonction à deux composantes, on peut appliquer le théorème des accroissements finies à f_1 et f_2, et trouver des paramètre t_1 et t_2 pour ces deux composantes. En revanche, il n'y a aucune raison a priori pour que $t_1 = t_2$...

3.3.3 Caractérisation des fonctions de classe C^1

Soit $f : \mathbb{R}^d \to \mathbb{R}^m$ une fonction de classe C^1. Alors toutes les dérivées partielles sont continues, et en particulier, la fonction Jacobienne $\mathbf{x} \mapsto J_f(\mathbf{x})$ est continue (de $\mathbb{R}^d$ dans $\mathcal{M}_{m \times d}(\mathbb{R})$), ou, de manière équivalente, toutes les dérivées partielles sont continues. En fait, la réciproque est vraie.

Théorème 3.27 : Caractérisation des fonctions C^1

Soit $f : U \subset \mathbb{R}^d \to \mathbb{R}^m$. La fonction f est de classe C^1 sur U ssi toutes les dérivées partielles $\frac{\partial f}{\partial x_i}$ existent et sont continues sur U.

On rappelle que ce n'est pas parce que les dérivées partielles de f en $\mathbf{x}_0$ existent que f est différentiable en $\mathbf{x}_0$ (cf Exemple 3.12). En fait, la partie difficile de ce théorème est de montrer que f est différentiable sur U. Une fois qu'on a montré cela, on pourra définir la Jacobienne $J_f(\mathbf{x})$. Ses composantes sont $\frac{\partial f_j}{\partial x_i}(\mathbf{x})$, qui sont continues en $\mathbf{x}$, donc $\mathbf{x} \mapsto J_f(\mathbf{x})$ est continue ce qui prouve directement que f est C^1.

Démonstration. D'après la remarque précédente, il suffit de montrer que f est différentiable sur U. On prend $m = 1$ pour simplifier (on travaille composante par composante), et $d = 2$. Soit $\mathbf{x}_0 \in U$, et soit $\varepsilon > 0$ tel que $\mathcal{B}(\mathbf{x}_0, \varepsilon) \subset U$ (on rappelle que U est ouvert). Pour $\mathbf{h} \in \mathcal{B}(\mathbf{0}, \varepsilon)$, on a

$$\begin{aligned} f(\mathbf{x}_0 + \mathbf{h}) - f(\mathbf{x}_0) &= (f(\mathbf{x}_0 + h_1\mathbf{e}_1 + h_2\mathbf{e}_2) - f(\mathbf{x}_0 + h_1\mathbf{e}_1)) + (f(\mathbf{x}_0 + h_1\mathbf{e}_1) - f(\mathbf{x}_0)) \\ &= \left[\int_0^1 \partial_2 f(\mathbf{x}_0 + h_1\mathbf{e}_1 + th_2\mathbf{e}_2)\mathrm{d}t\right] h_2 + \left[\int_0^1 \partial_1 f(\mathbf{x}_0 + th_1\mathbf{e}_1)\mathrm{d}t\right] h_1 \\ &= \partial_2 f(\mathbf{x}_0)h_2 + \partial_1 f(\mathbf{x}_0)h_1 + \|\mathbf{h}\|\mathcal{E}(\mathbf{h}), \end{aligned}$$

où $\mathcal{E}(\cdot)$ est définie par

$$\frac{1}{\|\mathbf{h}\|}\left(\left[\int_0^1 \{\partial_2 f(\mathbf{x}_0 + h_1\mathbf{e}_1 + th_2\mathbf{e}_2) - \partial_2 f(\mathbf{x}_0)\}\,\mathrm{d}t\right] h_2 + \left[\int_0^1 \{\partial_1 f(\mathbf{x}_0 + th_1\mathbf{e}_1) - \partial_1 f(\mathbf{x}_0)\}\,\mathrm{d}t\right] h_1\right).$$

On laisse en exercice de montrer que $\mathcal{E}$ tend vers 0 lorsque $\mathbf{h} \to \mathbf{0}$, et on en déduit que f est différentiable en $\mathbf{x}_0$, et que sa différentielle est l'application linéaire

$$\mathrm{d}f_{\mathbf{x}_0} : \mathbf{h} \mapsto \partial_1 f(\mathbf{x}_0)h_1 + \partial_2 f(\mathbf{x}_0)h_2 = \langle \nabla f(\mathbf{x}_0), \mathbf{h}\rangle\,.$$

Ceci étant vrai pour $\mathbf{x}_0 \in U$, on a que f est différentiable sur U. De plus, sa différentielle est continue (en $\mathbf{x}$), car $\mathbf{x} \mapsto \nabla f(\mathbf{x})$ est continue, donc f est de classe C^1. □

CHAPITRE 4

FONCTIONS DE CLASSE C^2

On s'intéresse maintenant aux dérivées supérieures de f. Dans le cas de la dérivée première, on a vu que l'objet mathématique qui apparaissait naturellement était les applications linéaires. Pour les dérivées secondes, on fera naturellement apparaître des applications quadratiques.

4.1 Fonctions deux fois différentiables

4.1.1 Premières définitions

Définition 4.1 : Différentielles supérieures

Soit $f : U \subset E \to F$ une fonction de classe C^1 sur U, de différentielle $\mathrm{d}f = U \to \mathcal{L}(E, F)$. On dit que f est deux fois différentiable sur U si $\mathrm{d}f$ est différentiable sur U. On dit que f est de classe C^2 sur U si $\mathrm{d}f$ est de classe C^1 sur U.

Par induction, on peut dire que f est de classe C^k sur U si $\mathrm{d}f$ est de classe C^{k-1} sur U.

On remarquera que $\mathrm{d}f$ est une fonction de E dans $F_1 := \mathcal{L}(E, F)$. Dans le cas où $E = \mathbb{R}^d$ et $F = \mathbb{R}^m$, cet espace F_1 est de dimension $d \times m$. Il faut vérifier que $\mathbf{x} \mapsto \mathrm{d}f_{\mathbf{x}}$ est de classe C^1 en tant que fonction de d variables et $d \times m$ composantes. On note $\mathrm{d}^2 f$ la différentielle de $\mathrm{d}f$. Si $\mathrm{d}f$ est différentiable en $\mathbf{x}$, on a, par définition

$$\mathrm{d}f_{\mathbf{x}+\mathbf{h}} = \mathrm{d}f_{\mathbf{x}} + \mathrm{d}^2 f_{\mathbf{x}}(\mathbf{h}) + \|\mathbf{h}\|\mathcal{E}(\mathbf{h}),$$

qui est une égalité dans $F_1 = \mathcal{L}(E, F)$. En particulier, $\mathcal{E}(\mathbf{h})$ est ici une fonction linéaire de $\mathcal{L}(E, F)$, avec $\|\mathcal{E}(\mathbf{h})\|_{\mathrm{op}} \to 0$ lorsque $\mathbf{h} \to \mathbf{0}$. De plus, l'application $\mathbf{h} \ni \mathbb{R}^d \mapsto \mathrm{d}^2 f_{\mathbf{x}}(\mathbf{h}) \in \mathcal{L}(E, F)$ est linéaire par définition d'une différentielle. En fait, on a $\mathrm{d}^2 f_{\mathbf{x}} \in \mathcal{L}(E, \mathcal{L}(E, F))$, qui est un espace de dimension $d \times d \times m$. En remarquant que deux formes linéaires sont égales ssi elles coïncident sur tous les arguments, on peut écrire de manière équivalente

$$\forall \mathbf{k} \in \mathbb{R}^d, \quad \mathrm{d}f_{\mathbf{x}+\mathbf{h}}(\mathbf{k}) = \mathrm{d}f_{\mathbf{x}}(\mathbf{k}) + \mathrm{d}^2 f_{\mathbf{x}}(\mathbf{h})(\mathbf{k}) + \|\mathbf{h}\|\mathcal{E}(\mathbf{h})(\mathbf{k}).$$

L'application $(\mathbf{h}, \mathbf{k}) \mapsto \mathrm{d}^2 f_{\mathbf{x}}(\mathbf{h})(\mathbf{k})$ est linéaire en $\mathbf{h}$ et en $\mathbf{k}$, donc est une application bilinéaire. On écrit souvent $\mathrm{d}^2 f_{\mathbf{x}}(\mathbf{h}, \mathbf{k})$ cette application bilinéaire de $\mathbb{R}^d \times \mathbb{R}^d$ dans $\mathbb{R}^m$. En anticipant un peu sur la suite, cette application bilinéaire est moralement

$$\mathrm{d}^2 f_{\mathbf{x}}(\mathbf{h}, \mathbf{k}) = \sum_{i=1}^{d} \sum_{j=1}^{d} \frac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{x}) h_i k_j.$$

4.1.2 Dérivées croisées

Dans la suite, on s'intéressera principalement au cas où $E = \mathbb{R}^d$ et $F = \mathbb{R}$ (une seule composante). Dans le cas où $F = \mathbb{R}^m$, on pourra travailler *composante par composante.*

Définition 4.2 : Dérivées partielles d'ordre supérieure

Soit $f : U \subset \mathbb{R}^d \to \mathbb{R}$ une fonction deux fois différentiables. Pour $1 \le i, j \le d$ on note

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{x}) := \partial^2_{x_i x_j} f(\mathbf{x}) := \frac{\partial}{\partial x_i}\left(\frac{\partial f}{\partial x_j}\right)(\mathbf{x}).$$

Par exemple, si $f(x, y) = \cos(x)\mathrm{e}^{y^2}$, on a $\partial_x f(x, y) = -\sin(x)\mathrm{e}^{y^2}$, et $\partial^2_{yx} f(x, y) = -2y\sin(x)\mathrm{e}^{y^2}$.

4.1.3 Fonctions de classe C^2

On rappelle le Théorème 3.27 qui dit qu'une fonction est C^1 si toutes les dérivées partielles sont continues. En appliquant ce résultat à la fonction $\mathrm{d}f$, on obtient la caractérisation suivante.

Théorème 4.3 : Caractérisation des fonctions de classe C^2

Soit $f : U \subset \mathbb{R}^d \to \mathbb{R}$. La fonction f est de classe C^2 sur U ssi toutes les dérivées secondes existent et sont continues sur U.

On remarquera qu'il y a d^2 fonctions $\partial^2_{x_i x_j} f$ lorsque i et j varient entre 1 et d. De même, on peut vérifier qu'une fonction f est de classe C^3 si toutes les dérivées troisièmes $\partial^3_{x_i x_j x_k} f(\mathbf{x})$ sont continues, et il y a d^3 telles fonctions. Et ainsi de suite.

On dit qu'une fonction f est de classe C^∞ si elle est de classe C^k pour tout $k \in \mathbb{N}$. En pratique, toutes les fonctions *usuelles* sont C^∞ sur leur domaine (ouvert) de définition.

Le théorème suivant (admis) affirme que l'ordre de dérivation dans les dérivées croisées n'importe pas dès que f est suffisamment régulière.

Théorème 4.4 : Théorème de Schwarz (admis)

Si $f : U \subset \mathbb{R}^d \to \mathbb{R}$ est de classe C^2 sur U, alors, pour tout $\mathbf{x} \in U$, on a

$$\forall 1 \le i, j \le d, \qquad \frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{x}) = \frac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{x}).$$

L'Exercice suivant montre qu'il n'y pas toujours égalité lorsque f n'est que deux fois différentiable (au lieu de C^2).

Exercice 4.5

Soit $f : \mathbb{R}^2 \to \mathbb{R}$ définie par

$$f(x, y) := \frac{x^3 y}{x^2 + y^2}, \qquad f(0, 0) = 0.$$

Montrer que $\partial^2_{xy} f \neq \partial^2_{yx} f$.

4.1.4 La Hessienne

Définition 4.6 : Matrice Hessienne

Soit $f : U \subset \mathbb{R}^d \to \mathbb{R}$ une fonction de classe C^2. La Hessienne de f est la matrice de taille $d \times d$ contenant les dérivées croisées de f. Explicitement,

$$H_f(\mathbf{x}) := \begin{pmatrix} \dfrac{\partial^2 f}{\partial x_1^2} & \dfrac{\partial^2 f}{\partial x_2 \partial x_1} & \cdots & \dfrac{\partial^2 f}{\partial x_d \partial x_1} \\ \dfrac{\partial^2 f}{\partial x_1 \partial x_2} & \dfrac{\partial^2 f}{\partial x_2^2} & \cdots & \dfrac{\partial^2 f}{\partial x_d \partial x_1} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial^2 f}{\partial x_1 \partial x_d} & \cdots & \cdots & \dfrac{\partial^2 f}{\partial x_d^2} \end{pmatrix} (\mathbf{x}).$$

D'après le théorème de Schwarz, cette matrice est **symétrique**. On pourra aussi remarquer que

la Hessienne est la Jacobienne du gradient.

En effet, on rappelle que

$$\nabla f(\mathbf{x}) = \begin{pmatrix} \partial_{x_1} f \\ \vdots \\ \partial_{x_d} f \end{pmatrix} (\mathbf{x})$$

qui peut-être vue comme une fonction de $\mathbb{R}^d$ dans $\mathbb{R}^d$ (le gradient est un vecteur colonne). En prenant la Jacobienne de cette application, on retrouve la Hessienne de f.

4.1.5 Formule de Taylor-Young à l'ordre 2

Nous avons vu que la formule de Taylor à l'ordre 1 permettait de **définir** la différentielle, puis la notion d'être C^1, etc. Nous démontrons maintenant une formule de Taylor à l'ordre supérieur.

Théorème 4.7 : Formule de Taylor à l'ordre 2

Si $f : \mathbb{R}^d \to \mathbb{R}$ est de classe C^2, alors

$$f(\mathbf{x} + \mathbf{h}) = f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{h} \rangle + \frac{1}{2} \langle \mathbf{h}, H_f(\mathbf{x}) \mathbf{h} \rangle + o(\mathbf{h}^2). \tag{4.1}$$

On retrouve la partie linéaire en $\mathbf{h}$ dans $\langle \nabla f(\mathbf{x}), \mathbf{h} \rangle$. La partie suivante est une partie quadratique en $\mathbf{h}$, dans le sens où $\mathbf{h} \mapsto \frac{1}{2} \langle \mathbf{h}, H_f(\mathbf{x}) \mathbf{h} \rangle$ est quadratique en $\mathbf{h}$. Ici, la notation $o(\mathbf{h}^2)$ est un abus de langage (que signifie un vecteur au carré ?) qui signifie une fonction de type

$$o(\mathbf{h}^2) = \|\mathbf{h}\|^2 \mathcal{E}(\mathbf{h}), \qquad \text{avec} \quad \mathcal{E}(\mathbf{h}) \xrightarrow[\mathbf{h} \to \mathbf{0}]{} 0.$$

Démonstration. On commence par appliquer le théorème fondamental de l'analyse à la fonction f (Théorème 3.24), et on obtient

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

Par ailleurs, en appliquant le théorème fondamental à la fonction ∇f, et en rappelant que la Hessienne est la Jacobienne du gradient, on a

$$\nabla f(\mathbf{x} + t\mathbf{h}) - \nabla f(\mathbf{x}) = \left(\int_0^1 H_f(\mathbf{x} + st\mathbf{h}) \mathrm{d}s \right) \times t\mathbf{h}.$$

On en déduit que

$$\begin{aligned} f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) &= \langle \nabla f(\mathbf{x}), \mathbf{h} \rangle + \int_0^1 \int_0^1 \langle \mathbf{h}, H_f(\mathbf{x}+st\mathbf{h}) t\mathbf{h} \rangle \mathrm{d}s\mathrm{d}t \\ &= \langle \nabla f(\mathbf{x}), \mathbf{h} \rangle + \int_0^1 \int_0^1 \langle \mathbf{h}, H_f(\mathbf{x}) t\mathbf{h} \rangle \mathrm{d}s\mathrm{d}t + \int_0^1 \int_0^1 \langle \mathbf{h}, \left[H_f(\mathbf{x}+st\mathbf{h}) - H_f(\mathbf{x}) \right] t\mathbf{h} \rangle \mathrm{d}s\mathrm{d}t. \end{aligned}$$

Dans la première intégrale, on a $\int_0^1 t\mathrm{d}t = \frac{1}{2}$. Le dernier terme est de la forme

$$\|\mathbf{h}\|^2 \mathcal{E}(\mathbf{h}), \quad \text{avec} \quad \mathcal{E}(\mathbf{h}) = \int_0^1 \int_0^1 \left\langle \frac{\mathbf{h}}{\|\mathbf{h}\|}, \left[H_f(\mathbf{x}+st\mathbf{h}) - H_f(\mathbf{x}) \right] t \frac{\mathbf{h}}{\|\mathbf{h}\|} \right\rangle \mathrm{d}s\mathrm{d}t$$

et vérifie, par l'inégalité de Cauchy-Schwarz et la propriété des normes opérateurs,

$$|\mathcal{E}(\mathbf{h})| \leq \int_0^1 \int_0^1 \| H_f(\mathbf{x}+st\mathbf{h}) - H_f(\mathbf{x}) \|_{\mathrm{op}} t \mathrm{d}s\mathrm{d}t,$$

qui tend vers 0 lorsque $\mathbf{h} \to \mathbf{0}$ par continuité de la Hessienne. Finalement, on a bien trouvé

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

□

4.2 Application à l'optimisation de fonctions

La formule de Taylor du Théorème 4.7 donne la meilleure approximation de f par une forme quadratique. Or il est facile de déterminer les minima des formes quadratiques. Dans la suite, nous rappelons les propriétés des formes quadratiques, et appliquons les résultats pour la recherche de minima.

4.2.1 Rappels sur les matrices symétriques

On a vu que la Hessienne d'une fonction de classe C^2 était une matrice symétrique. On rappelle qu'une matrice symétrique $A \in \mathcal{M}_{d \times d}(\mathbb{R})$ est diagonalisable. Cela signifie qu'il existe une matrice orthogonale $U \in O(d)$ et une matrice diagonale $D = \mathrm{diag}(\lambda_1, \cdots, \lambda_d)$ telle que

$$A = UDU^T.$$

Comme $U^T U = U U^T = \mathbb{I}_d$, on a $AU = UD$. En écrivant $U = (\mathbf{u}_1, \cdots, \mathbf{u}_d)$ les colonnes de U, la relation $AU = UD$ montre que

$$\forall 1 \leq i \leq d, \quad A\mathbf{u}_i = \lambda_i \mathbf{u}_i,$$

autrement dit, les vecteurs $\mathbf{u}_i$ sont des vecteurs propres de A pour les valeurs propres λ_i. De plus, la relation $U^T U = \mathbb{I}_d$ montre que les vecteurs $(\mathbf{u}_1, \cdots, \mathbf{u}_d)$ forment une base orthonormale. Autrement dit, on a trouvé une base orthonormale composée de vecteurs propres de A.

Définition 4.8 : Matrices positives, négatives, ...

On note $\mathcal{S}_d(\mathbb{R})$ l'ensemble des matrices symétrique de taille $d \times d$. Pour $A \in \mathcal{S}_d(\mathbb{R})$, on note $\lambda_1(A) \leq \lambda_2(A) \leq \cdots \leq \lambda_d(A)$ les valeurs propres de A, rangées dans l'ordre croissant.
On dit que A est **dégénérée** si $\lambda_j(A) = 0$ pour un certain $1 \leq j \leq d$.
On dit que A est **positive** si $\lambda_j(A) \geq 0$ pour tout $1 \leq j \leq d$.
On dit que A est **définie positive** si $\lambda_j(A) > 0$ pour tout $1 \leq j \leq d$.
On note $\mathcal{S}_d^+(\mathbb{R})$ l'ensemble des matrices positives, et $\mathcal{S}_d^{++}(\mathbb{R})$ l'ensemble des matrices définies positives.

On définit de même les matrices négatives et définies négatives. Une matrice A est définie positive ssi elle est positive et non dégénérée. On remarquera que A est dégénérée ssi $\mathrm{Ker}(A) \neq \{0\}$ (car 0 est une valeur propre de A) ssi A est non inversible.

Théorème 4.9 : Caractérisation des matrices positives

Soit $A \in \mathcal{S}_d(\mathbb{R})$. On a toujours l'inégalité

$$\forall \mathbf{x} \in \mathbb{R}^d, \quad \langle \mathbf{x}, A\mathbf{x} \rangle \geq \lambda_1(A) \|\mathbf{x}\|^2,$$

avec égalité ssi $\mathbf{x} \in \mathrm{Ker}(A - \lambda_1(A))$.
En particulier, A est positive ssi $\langle \mathbf{x}, A\mathbf{x} \rangle \geq 0$ pour tout $\mathbf{x} \in \mathbb{R}^d$, et A est définie positive ssi $\langle \mathbf{x}, A\mathbf{x} \rangle > 0$ pour tout $\mathbf{x} \in \mathbb{R}^d \setminus \{\mathbf{0}\}$.

Démonstration. Soit $\mathbf{x} \in \mathbb{R}^d$. On décompose $\mathbf{x}$ dans la base des $(\mathbf{u}_i)$, et on écrit

$$\mathbf{x} = x_1 \mathbf{u}_1 + x_2 \mathbf{u}_2 + \cdots + x_d \mathbf{u}_d.$$

En appliquant A, et en utilisant le fait que les $\mathbf{u}_i$ sont des vecteurs propres, on obtient

$$A\mathbf{x} = x_1 \lambda_1 \mathbf{u}_1 + x_2 \lambda_2 \mathbf{u}_2 + \cdots + x_d \lambda_d \mathbf{u}_d.$$

Ainsi, en utilisant que les $(\mathbf{u}_i)$ sont orthonormés, on obtient

$$\langle \mathbf{x}, A\mathbf{x} \rangle = \sum_{i=1}^{d} |x_i|^2 \lambda_i \geq \lambda_1 \sum_{i=1}^{d} |x_i|^2 = \lambda_1 \|\mathbf{x}\|^2.$$

☐

En règle générale, pour déterminer si une matrice A est positive, il faut la diagonaliser pour trouver ses valeurs propres, ce qui peut être fastidieux. Dans le cas des matrices 2×2, on a un critère simple pour savoir si une matrice est positive.

Théorème 4.10 : Matrices positives de taille 2×2

Une matrice $A \in \mathcal{S}_2(\mathbb{R})$ est positive ssi sa trace et son déterminant sont positifs. Elle est non dégénérée si, de plus, son déterminant est non nul.

Démonstration. On a $\mathrm{Tr}(A) = \lambda_1 + \lambda_2$ (somme des valeurs propres) et $\det(A) = \lambda_1 \lambda_2$ (produit des valeurs propres). Or deux nombres sont positifs ssi la somme et le produit de ces nombres sont positifs. Le résultat suit. ☐

Attention : le résultat est faux en dimension $d \geq 3$. Par exemple, la matrice

$$A = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 3 \end{pmatrix}$$

vérifie $\mathrm{Tr}(A) = 1 \geq 0$ et $\det(A) = 3 > 0$, mais a deux valeurs propres négatives. En dimension supérieur, on pourra vérifier qu'une matrice est définie positive en appliquant le critère de Sylvester (admis).

Théorème 4.11 : Critère de Sylvester en dimension quelconque

Si $A \in \mathcal{S}_d(\mathbb{R})$, alors A est définie positive ssi tous ses mineurs principaux dominants sont strictement positifs. En d'autres termes, si $A = (a_{i,j})_{1 \leq i,j \leq d} \in \mathcal{S}_d(\mathbb{R})$, on note $A_k = (a_{i,j})_{1 \leq i,j \leq k} \in \mathcal{S}_k(\mathbb{R})$. Alors $A \in \mathcal{S}_d^{++}(\mathbb{R})$ si et seulement si $\det(A_k) > 0$ pour tout $k \in \{1, \ldots, d\}$.

4.2.2 Caractérisation des points critiques non dégénérés

Nous sommes enfin prêts pour trouver des critères "simples" pour déterminer les minima des fonctions.

Théorème 4.12 : Les minimiseurs sont des points critiques

Soit $f : \mathbb{R}^d \to \mathbb{R}$ une fonction de classe C^2. Si $\mathbf{x}_*$ est un minimiseur local de f, alors

$$\nabla f(\mathbf{x}_*) = \mathbf{0}, \quad \text{et} \quad H_f(\mathbf{x}_*) \geq 0.$$

Démonstration. Soit $\varepsilon > 0$ tel que $\mathbf{x}_*$ est un minimiseur de f dans $\mathcal{B}(\mathbf{x}_*, \varepsilon)$. Soit $\mathbf{h} \in \mathbb{R}^d$ quelconque. Pour tout $0 \leq t < \varepsilon \|\mathbf{h}\|^{-1}$, on a $\|t\mathbf{h}\| \leq |t| \|\mathbf{h}\| \leq \varepsilon$, et donc $t\mathbf{h} \in \mathcal{B}(\mathbf{0}, \varepsilon)$. Ainsi

$$f(\mathbf{x}_*) \leq f(\mathbf{x}_* + t\mathbf{h}) = f(\mathbf{x}_*) + t\langle \nabla f(\mathbf{x}_*), \mathbf{h} \rangle + \frac{1}{2} t^2 \langle \mathbf{h}, H_f(\mathbf{x}_*)\mathbf{h} \rangle + t^2 \mathbf{h}^2 \mathcal{E}(t\mathbf{h}).$$

avec $\mathcal{E}(t\mathbf{h}) \to 0$ lorsque $t \to 0$. En simplifiant par $f(\mathbf{x}_*)$ et en divisant par t, on obtient

$$\langle \nabla f(\mathbf{x}_*), \mathbf{h} \rangle + \frac{1}{2} t \langle \mathbf{h}, H_f(\mathbf{x}_*)\mathbf{h} \rangle + t\mathbf{h}^2 \mathcal{E}(t\mathbf{h}) \geq 0.$$

En prenant la limite $t \to 0$, on obtient $\langle \nabla f(\mathbf{x}_*), \mathbf{h} \rangle \geq 0$. Ceci devant être vrai pour tout $\mathbf{h} \in \mathbb{R}^d$, on en déduit que $\nabla f(\mathbf{x}_*) = \mathbf{0}$ (On utilise ici la non dégénérescence du produit scalaire).

En reprenant l'inégalité précédente, on trouve

$$\frac{1}{2} t \langle \mathbf{h}, H_f(\mathbf{x}_*)\mathbf{h} \rangle + t\mathbf{h}^2 \mathcal{E}(t\mathbf{h}) \geq 0.$$

On redivise par t, et on prend la limite $t \to 0$ pour obtenir cette fois que $\langle \mathbf{h}, H_f(\mathbf{x}_*)\mathbf{h} \rangle \geq 0$. Ceci étant vrai pour tout $\mathbf{h} \in \mathbb{R}^d$, on conclut que $H_f(\mathbf{x}_*)$ est une matrice symétrique positive. ☐

Cela motive d'étudier les points $\mathbf{x}_*$ tel que $\nabla f(\mathbf{x}_*) = \mathbf{0}$. Comme on le verra dans la suite, il s'agit d'une famille de d équations à d inconnues.

Définition 4.13 : Points critiques, points critiques non dégénérés

Soit $f : U \subset \mathbb{R}^d \to \mathbb{R}$ une fonction de classe C^2 sur U.
On dit que $\mathbf{x}_*$ est un **point critique** de f si $\nabla f(\mathbf{x}_*) = \mathbf{0}$.
Un point critique est dit **non dégénéré** si de plus $H_f(\mathbf{x}_*)$ est non dégénérée ($\det H_f(\mathbf{x}_*) \neq 0$).
Un point critique qui n'est ni un minimiseur local, ni un maximiseur local est un **point selle**.

La troisième définition stipule qu'il existe 3 types de points critiques : les minimiseurs, les maximiseurs, et les points selles.

En combinant cette définition avec le résultat précédent, on remarque que, si $\mathbf{x}_*$ est un minimiseur non dégénéré, alors $H_f(\mathbf{x}_*) > 0$ ($H_f(\mathbf{x}_*)$ est défini positif). On s'intéresse maintenant à la *réciproque* du résultat précédent.

Théorème 4.14

Soit $f : \mathbb{R}^d \to \mathbb{R}$ une fonction de classe C^2. Si $\mathbf{x}_* \in \mathbb{R}^d$ est tel que

$$\nabla F(\mathbf{x}_*) = \mathbf{0}, \quad \text{et} \quad H_f(\mathbf{x}_*) > 0,$$

alors $\mathbf{x}_*$ est un minimiseur local non dégénéré de F.

On remarquera qu'on demande $H_f(\mathbf{x}_*)$ d'être définie positive dans le deuxième cas pour pouvoir conclure. Dans le cas $d=1$ par exemple, les fonctions définies par

$$f_1(x) = x^3, \quad \text{et} \quad f_2(x) = x^4$$

vérifient toutes les deux $f'(0) = f''(0) = 0$. Dans le premier cas, le point 0 n'est pas un minimiseur de f_1, mais 0 est un minimiseur de f_2.

Démonstration. Soit $\lambda_1 > 0$ la plus petite valeur propre de $H_f(\mathbf{x}_*)$. On a

$$f(\mathbf{x}_* + \mathbf{h}) \geq f(\mathbf{x}_*) + \frac{1}{2}\lambda_1 \|\mathbf{h}\|^2 + \|\mathbf{h}\|^2 \mathcal{E}(\mathbf{h}).$$

On rappelle que $\mathcal{E}$ est continue en $\mathbf{h} = \mathbf{0}$ avec $\mathcal{E}(\mathbf{0}) = 0$. Soit $\varepsilon > 0$ suffisamment petit pour que $\mathcal{E}(\mathbf{h}) > -\frac{1}{2}\lambda_1$ pour tout $\|\mathbf{h}\| < \varepsilon$. Alors, pour tout $\mathbf{h} \in \mathcal{B}(\mathbf{0}, \varepsilon) \setminus \{\mathbf{0}\}$, on a

$$f(\mathbf{x}_* + \mathbf{h}) \geq f(\mathbf{x}_*) + \frac{1}{4}\lambda_1 \|\mathbf{h}\|^2 > f(\mathbf{x}_*),$$

ce qui prouve que $\mathbf{x}_*$ est un minimiseur strict dans $\mathcal{B}(\mathbf{x}_*, \varepsilon)$. □

4.2.3 Un exemple de recherche de minimiseurs

On conclut cette section en expliquant comment étudier une fonction à plusieurs variables en pratique. On s'intéressera principalement aux fonctions à deux variables. Un exercice typique est le suivant.

Exercice 4.15 : Exemple type

Étudier les points critiques de la fonction

$$f(x,y) := (x^2 + 2y^2)\mathrm{e}^{-(x^2+y^2)}.$$

Étude de la régularité

La fonction f est de classe C^∞ sur $\mathbb{R}^2$ comme composée de fonctions usuelles C^∞ sur leur domaine de définition.

Calculs des dérivées premières

On a

$$\frac{\partial f}{\partial x}(x,y) = \mathrm{e}^{-(x^2+y^2)}\left[2x - 2x(x^2+2y^2)\right] = 2x\mathrm{e}^{-(x^2+y^2)}\left[1 - x^2 - 2y^2\right].$$

De même,

$$\frac{\partial f}{\partial y}(x,y) = \mathrm{e}^{-(x^2+y^2)}\left[4y - 2y(x^2+2y^2)\right] = 2y\mathrm{e}^{-(x^2+y^2)}\left[2 - x^2 - 2y^2\right].$$

Recherche des points critiques

Pour trouver les points critiques, on résout le système

$$\begin{cases} \partial_x f(x,y) & = 0 \\ \partial_y f(x,y) & = 0. \end{cases}, \quad \iff \quad \begin{cases} x\left[1 - x^2 - 2y^2\right] = 0 \\ y\left[2 - x^2 - 2y^2\right] = 0, \end{cases} \quad \iff \quad \begin{cases} x = 0 \quad \text{ou} \quad x^2 + 2y^2 = 1 \\ y = 0 \quad \text{ou} \quad x^2 + 2y^2 = 2. \end{cases}$$

On en déduit qu'il y a 5 points critiques, à savoir les points

$$(0,0), \quad (1,0), \quad (-1,0), \quad (0,1) \quad \text{et} \quad (0,-1).$$

Calcul de la Hessienne

On a

$$\begin{aligned}\frac{\partial^2 f}{\partial x^2}(x,y) &= \mathrm{e}^{-(x^2+y^2)}\left[2-6x^2-4y^2-2x(2x-2x^3-4xy^2)\right],\\ \frac{\partial^2 f}{\partial y^2}(x,y) &= \mathrm{e}^{-(x^2+y^2)}\left[4-2x^2-12y^2-2y(4y-2yx^2-4y^3)\right],\\ \frac{\partial^2 f}{\partial x\partial y}(x,y) = \frac{\partial^2 f}{\partial x\partial y}(x,y) &= \mathrm{e}^{-(x^2+y^2)}\left[-8xy-2y(x^2-2x^3-4xy^2)\right],\end{aligned}$$

où on a utilisé le Lemme de Schwarz pour l'égalité $\partial^2_{xy}f = \partial^2_{yx}f$. L'expression de la hessienne est donc assez complexe. Cependant, on veut l'évaluer uniquement aux points critiques.

Étude des points critiques

On trouve

$$H_f(0,0)=\begin{pmatrix}2&0\\0&4\end{pmatrix},\quad H_f(1,0)=H_f(-1,0)=\mathrm{e}^{-1}\begin{pmatrix}-4&0\\0&2\end{pmatrix},\quad H_f(0,-1)=H_f(0,1)=\mathrm{e}^{-1}\begin{pmatrix}-2&0\\0&-8\end{pmatrix}.$$

Dans cet exemple, les hessiennes sont diagonales, et il est facile de déterminer la nature des points critiques.

- La hessienne au point $(0,0)$ est positive, donc $(0,0)$ est un **minimum local**. De plus, il est facile de voir que $f>0$ et que $f(0,0)=0$. Donc $(0,0)$ est un minimum global !
- La hessienne aux points $(\pm 1,0)$ ont une valeur propre positive et une valeur propre négative. Ces points sont donc des **points selles**.
- La hessienne aux points $(0,\pm 1)$ est négative, donc ces points sont des **maxima locaux**. En fait, on peut montrer que ce sont des maxima globaux, mais c'est plus difficile. Cela vient du fait que la fonction f tend vers 0 en l'infini.

4.2.4 Un autre exemple type

Exercice 4.16 : Exemple type (bis)

Étudier les points critiques de la fonction

$$f(x,y)=x^3+3xy^2-15x-12y.$$

Étude de la régularité

La fonction f est de classe C^∞ sur $\mathbb{R}^2$, car f est un polynôme.

Calculs des dérivées premières

On a

$$\frac{\partial f}{\partial x}(x,y)=3x^2+3y^2-15,\quad \text{et}\quad \frac{\partial f}{\partial y}(x,y)=6xy-12.$$

Recherche des points critiques

Pour trouver les points critiques, on résout le système

$$\begin{cases}\partial_x f(x,y) &= 0\\ \partial_y f(x,y) &= 0\end{cases}\iff\begin{cases}x^2+y^2=5\\ xy=2\end{cases}\iff\begin{cases}(x+y)^2=9\\ (x-y)^2=1\end{cases}\iff\begin{cases}x+y=\pm 3\\ x-y=\pm 1\end{cases}$$

Il y a donc 4 points critiques possibles, suivant les signes qu'on met. On trouve les 4 points critiques

$$(1,2)\quad(-1,-2)\quad(2,1)\quad(-2,-1).$$

Calcul de la Hessienne

On trouve directement

$$H_f(x,y) = 6 \begin{pmatrix} x & y \\ y & x \end{pmatrix}.$$

On remarque que la trace $\mathrm{Tr}(H_f) = 12x$ est du signe de x, et que $\det(H_f) = 36(x^2 - y^2)$ est positif si $|x| > |y|$ et négatif si $|x| < |y|$. On trouve donc

- Pour le point $(1, 2)$, la hessienne a une valeur propre positive, et une valeur propre négative. Donc le point $(1, 2)$ est un point selle.
- Pour le point $(-1, -2)$, idem.
- Pour le point $(2, 1)$, la hessienne est positive, donc $(2, 1)$ est un minimum local de f.
- Pour le point $(-2, -1)$, la hessienne est négative, donc $(-2, -1)$ est un maximum local de f.

4.3 Minimisation de fonctions usuelles

Dans cette section, nous présentons des résultats théoriques pour la minimisation des fonctions quadratiques. Nous appliquons ensuite la méthode pour la minimisation aux moindres carrés.

4.3.1 Minimisation des fonctions quadratiques

On dit qu'une fonction $Q : \mathbb{R}^d \to \mathbb{R}$ est **quadratique** si elle est de la forme

$$\boxed{Q(\mathbf{x}) := \frac{1}{2}\langle \mathbf{x}, A\mathbf{x}\rangle - \langle \mathbf{b}, \mathbf{x}\rangle.}$$

avec $A \in \mathcal{M}_d(\mathbb{R})$ une matrice carrée, et $\mathbf{b} \in \mathbb{R}^d$ un vecteur de $\mathbb{R}^d$. En dimension 1, on retrouve les paraboles, de la forme $\frac{1}{2}ax^2 - bx$. Le choix du facteur $\frac{1}{2}$ et du signe devant b est évidemment arbitraire, mais permet des calculs plus agréables dans la suite. Pour commencer, on note qu'on peut toujours supposer A symétrique.

Lemme 4.17. *Soit $A \in \mathcal{M}_d(\mathbb{R})$, soit $\widetilde{A} := \frac{1}{2}(A + A^T)$, et soit $\widetilde{Q}(\mathbf{x}) := \frac{1}{2}\langle \mathbf{x}, \widetilde{A}\mathbf{x}\rangle - \langle \mathbf{b}, \mathbf{x}\rangle$. Alors $\widetilde{Q} = Q$. En particulier, on peut toujours représenter une fonction quadratique avec une matrice A symétrique.*

Démonstration. Posons $B := A - \widetilde{A} = \frac{1}{2}(A - A^T)$, de sorte que $A = \widetilde{A} + B$. On pourra remarquer que B est anti-symétrique : $B^T = -B$. En particulier, on a, pour tout $\mathbf{x} \in \mathbb{R}^d$,

$$\langle \mathbf{x}, B\mathbf{x}\rangle = \langle B^T\mathbf{x}, \mathbf{x}\rangle = -\langle B\mathbf{x}, \mathbf{x}\rangle = -\langle \mathbf{x}, B\mathbf{x}\rangle,$$

et donc $\langle \mathbf{x}, B\mathbf{x}\rangle = 0$. Ainsi, $\langle \mathbf{x}, A\mathbf{x}\rangle = \langle \mathbf{x}, (\widetilde{A} + B)\mathbf{x}\rangle = \langle \mathbf{x}, \widetilde{A}\mathbf{x}\rangle$. □

Dans la suite, on se restreint uniquement aux matrices $A \in \mathcal{S}_d(\mathbb{R})$ symétriques.

Lemme 4.18. *Soit $A \in \mathcal{S}_d$ et $\mathbf{b} \in \mathbb{R}^d$, et soit Q la forme quadratique associée. Alors $Q(\cdot)$ est coercive ssi $A \in \mathcal{S}_d^{++}(\mathbb{R})$ ssi $Q(\cdot)$ est strictement convexe.*

Démonstration. Pour commencer, un calcul direct montre que (le faire !)

$$Q((1-t)\mathbf{x} + t\mathbf{y}) - (1-t)Q(\mathbf{x}) - tQ(\mathbf{y}) = -t(1-t)Q(\mathbf{x} - \mathbf{y}) = -t(1-t)\langle (\mathbf{x} - \mathbf{y}), A(\mathbf{x} - \mathbf{y})\rangle.$$

Ainsi, le membre de gauche est strictement négatif pour tout $0 < t < 1$ et tout $\mathbf{x} \neq \mathbf{y}$ ssi $\langle \mathbf{h}, A\mathbf{h}\rangle > 0$ pour tout $\mathbf{h} \neq \mathbf{0}$, c'est à dire ssi $A \in \mathcal{S}_d^{++}$. Par ailleurs, si $A \in \mathcal{S}_d^{++}$, on a, avec le Théorème 4.9 et Cauchy-Schwarz,

$$Q(\mathbf{x}) \geq \frac{1}{2}\lambda_1(A)\|\mathbf{x}\|^2 - \|\mathbf{b}\| \cdot \|\mathbf{x}\|.$$

Comme $\lambda_1(A) > 0$, la fonction $X \mapsto \frac{1}{2}\lambda_1(A)X^2 - \|\mathbf{b}\|X$ est coercive sur $\mathbb{R}$, donc $Q(\cdot)$ est coercive sur $\mathbb{R}^d$. Réciproquement, si $A \notin \mathcal{S}_d^{++}$, il existe un vecteur unitaire $\mathbf{u}$ tel que $\langle \mathbf{u}, A\mathbf{u} \rangle \le 0$. En particulier, on a

$$Q(t\mathbf{u}) \le -t\langle \mathbf{b}, \mathbf{u} \rangle,$$

qui tend vers $-\infty$ lorsque t tend vers $+\infty$ ou $-\infty$ (dépendant du signe de $\langle \mathbf{b}, \mathbf{u} \rangle$), ou vers 0 si $\langle \mathbf{b}, \mathbf{u} \rangle = 0$. Dans tous les cas, on a trouvé une direction dans laquelle Q ne tend pas vers $+\infty$ à l'infini, donc $Q(\cdot)$ n'est pas coercive. □

Les fonctions quadratiques fournissent donc un prototype parfait de fonctions à minimiser : si $A \in \mathcal{S}_d^{++}$, alors Q est strictement convexe et coercive, donc admet un unique minimiseur. Le théorème suivant caractérise ce minimiseur.

Théorème 4.19

Soit $A \in \mathcal{S}_d^{++}$, $\mathbf{b} \in \mathbb{R}^d$, et soit $Q(\cdot)$ la fonction quadratique associée.
La gradient de $Q(\cdot)$ est $\nabla Q(\mathbf{x}) := A\mathbf{x} - \mathbf{b}$ et la hessienne de $Q(\cdot)$ est $H_Q(\mathbf{x}) = A$. En particulier, l'unique minimiseur de $Q(\cdot)$ est

$$\mathbf{x}_* := A^{-1}\mathbf{b},$$

et c'est un minimum global non dégénéré.

Démonstration. Pour calculer le gradient, on remarque que

$$\begin{aligned} Q(\mathbf{x}+\mathbf{h}) &= \frac{1}{2}\langle \mathbf{x}+\mathbf{h}, A(\mathbf{x}+\mathbf{h}) \rangle - \langle \mathbf{b}, \mathbf{x}+\mathbf{h} \rangle \\ &= \left(\frac{1}{2}\langle \mathbf{x}, A\mathbf{x} \rangle - \langle \mathbf{b}, \mathbf{x} \rangle \right) + \left(\frac{1}{2}(\langle \mathbf{x}, A\mathbf{h} \rangle + \langle \mathbf{h}, A\mathbf{x} \rangle) - \langle \mathbf{b}, \mathbf{h} \rangle \right) + \frac{1}{2}\langle \mathbf{h}, A\mathbf{h} \rangle \\ &= Q(\mathbf{x}) + \langle \frac{1}{2}(A + A^T)\mathbf{x} - \mathbf{b}, \mathbf{h} \rangle + o(\mathbf{h}). \end{aligned}$$

On a utilisé le fait que $|\langle \mathbf{h}, A\mathbf{h} \rangle| \le \|A\|_{\mathrm{op}} \|\mathbf{h}\|^2 = O(\mathbf{h})$. En utilisant le fait que A est symétrique, donc que $A = A^T$, et par identification, on trouve bien que $\nabla Q(\mathbf{x}) = A\mathbf{x} - \mathbf{b}$. De plus, comme la hessienne est la jacobienne du gradient, on trouve que $H_Q(\mathbf{x}) = A$.
Comme $A \in \mathcal{S}_d^{++}$ est non dégénérée, elle est inversible. En particulier, l'équation $\nabla Q(\mathbf{x}) = \mathbf{0}$ s'écrit $A\mathbf{x} = \mathbf{b}$, et n'a qu'une seule solution $\mathbf{x}_* = A^{-1}\mathbf{b}$. La Hessienne en ce point est $A \in \mathcal{S}_d^{++}$, donc $\mathbf{x}_*$ est un minimum local non dégénéré.
Par ailleurs, on a déjà montré que Q est convexe, donc ce minimiseur est, en fait, global. □

Le point tout à fait remarquable du théorème précédent est qu'on peut résoudre le système linéaire $A\mathbf{x} = \mathbf{b}$ en *minimisation la fonctionnelle quadratique* Q. On a résolu un problème d'algèbre avec une méthode analytique !

On fera cependant attention que cette méthode ne marche que pour des matrices A symétriques.

4.3.2 Minimisation aux moindres carrés

Un exemple important est donné par la *minimisation aux moindres carrés*. L'idée est la suivante : on se donne un ensemble de données de la forme $(\mathbf{x}_j, y_j)$, $1 \le j \le N$ (N est le nombre de données), avec $\mathbf{x}_i \in \mathbb{R}^d$ et $y_j \in \mathbb{R}$, et on cherche une fonction f telle que $f(\mathbf{x}_j) = y_j$.

A priori, on peut choisir f n'importe comment... mais cela n'apporte pas forcément d'informations utiles. Dans la minimisation aux moindres carrés *linéaire*, on cherche f de la forme

$$f_{\boldsymbol{\alpha}}(\mathbf{x}) = \sum_{k=1}^{p} \alpha_k \phi_k(\mathbf{x}),$$

où les fonctions $(\phi_k)_{1\le k\le p}$ sont fixées à l'avance. Notre but est alors d'optimiser les coefficients $(\alpha_1, \cdots, \alpha_p)$. Le nombre p est le nombre de *paramètres* à optimiser. En toute généralité, il n'est pas possible de trouver un f qui réalise exactement l'égalité $f(\mathbf{x}_j) = y_j$. On cherche plutôt à minimiser l'erreur quadratique

$$F(\boldsymbol{\alpha}) := \sum_{j=1}^{N} |y_j - f_{\boldsymbol{\alpha}}(\mathbf{x}_j)|^2.$$

Le terme *moindres carrés* vient du choix de cette erreur, avec la moyenne des carrés. Notre problème d'optimisation s'écrit donc

$$\inf \left\{ F(\boldsymbol{\alpha}), \quad \boldsymbol{\alpha} \in \mathbb{R}^p \right\},$$

C'est un problème de minimisation d'une fonction F de $\mathbb{R}^p$ dans $\mathbb{R}$.

Voici quelques exemples de fonctions ϕ_k usuelles.

— **Régression polynomiale**. Dans le cas $d = 1$ ($\mathbf{x}_i = x_i \in \mathbb{R}$), on peut prendre $\phi_k(x) = x^{k-1}$. Dans ce cas, f est un polynôme de degré $p - 1$. Ce cas inclut la régression affine, on cherche le meilleur couple $(\alpha_0, \alpha_1) \in \mathbb{R}^2$ (donc $p = 2$ paramètres à optimiser), qui minimise l'erreur

$$\inf \left\{ \sum_{j=1}^{N} |y_j - \alpha_1 x_j - \alpha_0|^2, \qquad (\alpha_0, \alpha_1) \in \mathbb{R}^2 \right\}.$$

— **Régression linéaire**. Prenons $p = N$, et $\phi_k(\mathbf{x}) = x_k$ (la k-ème coordonnée de $\mathbf{x}$). Dans ce cas, on peut écrire

$$f_{\boldsymbol{\alpha}}(\mathbf{x}) = \sum_{k=1}^{p} \alpha_k x_k = \langle \boldsymbol{\alpha}, \mathbf{x} \rangle \quad \text{avec} \quad \boldsymbol{\alpha} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_p \end{pmatrix}.$$

et le problème devient

$$\inf \left\{ \sum_{j=1}^{N} |y_j - \langle \boldsymbol{\alpha}, \mathbf{x}_j \rangle|^2, \quad \boldsymbol{\alpha} \in \mathbb{R}^p \right\}.$$

Afin de résoudre le problème de minimisation, on commence par ré-écrire le problème sous forme matricielle. On pose

$$A := \begin{pmatrix} \phi_1(\mathbf{x}_1) & \phi_2(\mathbf{x}_1) & \cdots & \phi_p(\mathbf{x}_1) \\ \phi_1(\mathbf{x}_2) & \phi_2(\mathbf{x}_2) & \cdots & \phi_p(\mathbf{x}_2) \\ \vdots & \vdots & \cdots & \vdots \\ \phi_1(\mathbf{x}_N) & \phi_2(\mathbf{x}_1) & \cdots & \phi_p(\mathbf{x}_N) \end{pmatrix} \in \mathcal{M}_{N\times p}, \qquad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} \in \mathbb{R}^N \quad \text{et} \quad \boldsymbol{\alpha} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_p \end{pmatrix} \in \mathbb{R}^p.$$

Avec ces notations, on remarque que

$$F(\boldsymbol{\alpha}) = \|\mathbf{y} - A\boldsymbol{\alpha}\|^2.$$

Ce problème s'exprime donc très simplement avec des matrices : on cherche le vecteur $\boldsymbol{\alpha} \in \mathbb{R}^p$ tel que $A\boldsymbol{\alpha}$ soit proche de $\mathbf{y}$, pour la norme euclidienne de $\mathbb{R}^N$. On note que la matrice A est fixée : elle ne dépend que des données $\mathbf{x}_j$ et du choix des fonctions ϕ_k. Ainsi, on s'est ramené au problème suivant. Soit $A \in \mathcal{M}_{N\times p}(\mathbb{R})$ et $\mathbf{y} \in \mathbb{R}^N$ fixés, on veut résoudre

$$\boxed{\inf\{\|\mathbf{y} - A\boldsymbol{\alpha}\|^2, \quad \boldsymbol{\alpha} \in \mathbb{R}^p\}.}$$

Résolvons ce problème de minimisation. Pour commencer, on a

$$F(\boldsymbol{\alpha}) := \frac{1}{2}\|\mathbf{y} - A\boldsymbol{\alpha}\|^2 = \frac{1}{2}\langle \mathbf{y} - A\boldsymbol{\alpha}, \mathbf{y} - A\boldsymbol{\alpha} \rangle = \frac{1}{2}\left(\|\mathbf{y}\|^2 - 2\langle A\boldsymbol{\alpha}, \mathbf{y} \rangle + \|A\boldsymbol{\alpha}\|^2 \right).$$

L'application $\boldsymbol{\alpha} \mapsto \|\mathbf{y} - A\boldsymbol{\alpha}\|^2$ est de classe C^∞, car c'est un polynôme en $\boldsymbol{\alpha}$. Calculons le gradient et la hessienne de cette application. On a

$$\begin{aligned} F(\boldsymbol{\alpha} + \mathbf{h}) &= F(\boldsymbol{\alpha}) - \langle A\mathbf{h}, \mathbf{y} \rangle + \langle A\mathbf{h}, A\boldsymbol{\alpha} \rangle + o(\mathbf{h}) \\ &= F(\boldsymbol{\alpha}) + \langle \mathbf{h}, A^T A\boldsymbol{\alpha} - A^T\mathbf{y} \rangle + o(\mathbf{h}). \end{aligned}$$

Ainsi, par identification, le gradient de F est

$$\nabla F(\boldsymbol{\alpha}) := A^T A\boldsymbol{\alpha} - A^T\mathbf{y}.$$

Lemme 4.20. *La matrice $A^T A \in \mathcal{M}_{p,p}(\mathbb{R})$ est symétrique positive.*
Elle est inversible (donc définie positive) ssi $A \in \mathcal{M}_{N\times p}(\mathbb{R})$ est injective. En particulier, si c'est le cas, on a $N \geq p$.

Démonstration. On remarque que si $A^T A\boldsymbol{\alpha} = \mathbf{0}$, alors $0 = \langle \boldsymbol{\alpha}, A^T A\boldsymbol{\alpha} \rangle = \|A\boldsymbol{\alpha}\|^2$, et donc $A\boldsymbol{\alpha} = \mathbf{0}$. □

Un cadre naturel pour ce problème est donc quand le nombre de données est plus grand que le nombre de paramètres. Dans la suite on supposera que A est injective.

Définition 4.21 : Pseudo–inverse de A

Soit $A \in \mathcal{M}_{N,p}(\mathbb{R})$ injective. Le pseudo-inverse de A est la matrice $A^\dagger \in \mathcal{M}_{p,N}(\mathbb{R})$ définie par

$$A^\dagger := (A^T A)^{-1} A^T.$$

Cette matrice vérifie

$$A^\dagger A = (A^T A)^{-1} A^T A = \mathbb{I}_p.$$

En revanche, $AA^\dagger$ est une matrice de taille $N \times N$ de rang p. Si $N > p$, cette matrice n'est pas inversible (et en particulier n'est pas l'identité).

Exercice 4.22

Montrer que si $N = p$ et si A est injective, alors A est inversible, et $A^\dagger = A^{-1}$.

Dans le cas où A est injective, on en déduit qu'il y a un unique point critique pour la fonction F, donnée par $\nabla F(\boldsymbol{\alpha}) = \mathbf{0}$, c'est à dire

$$\boldsymbol{\alpha}_* := A^\dagger \mathbf{y}.$$

Calculons maintenant la hessienne. On rappelle que la hessienne est la jacobienne du gradient. Comme le gradient est linéaire en $\boldsymbol{\alpha}$, on trouve directement

$$H_F(\boldsymbol{\alpha}) := A^T A.$$

Cette matrice ne dépend pas du point $\boldsymbol{\alpha}$. On a montré que si A est injective, alors $A^T A$ est définie positive. On en déduit directement que $\boldsymbol{\alpha}_*$ est un minimum local. De plus, comme F est coercive (pourquoi ?) et admet un unique point critique, c'est un minimiseur global.

Pour faire le lien avec la section précédente (minimisation d'une fonction quadratique), on pourra remarquer que la fonction

$$Q(\boldsymbol{\alpha}) := F(\boldsymbol{\alpha}) - \frac{1}{2}\|\mathbf{y}\|^2 = \frac{1}{2}\langle \boldsymbol{\alpha}, A^T A\boldsymbol{\alpha} \rangle - \langle A^T\mathbf{y}, \boldsymbol{\alpha} \rangle,$$

Comme $\|\mathbf{y}\|$ est constante du problème, les minimiseurs de F et de Q sont les mêmes. Or $Q(\cdot)$ est une fonction quadratique de la forme $Q(\boldsymbol{\alpha}) = \frac{1}{2}\langle \boldsymbol{\alpha}, \widetilde{A}\boldsymbol{\alpha} \rangle - \langle \widetilde{\mathbf{b}}, \mathbf{x} \rangle$ avec $\widetilde{A} = A^T A$ et $\widetilde{\mathbf{b}} = A^T\mathbf{y}$.

4.4 Fonctions convexes

On rappelle qu'une fonction $f : \mathbb{R}^d \to \mathbb{R}$ est convexe si

$$\boxed{\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d, \quad \forall 0 \le t \le 1, \qquad f(t\mathbf{y} + (1-t)\mathbf{x}) \le t f(\mathbf{y}) + (1-t) f(\mathbf{x}).}$$

En posant $\mathbf{h} = \mathbf{y} - \mathbf{x}$, de sorte que $\mathbf{y} = \mathbf{x} + \mathbf{h}$, c'est équivalent à l'inégalité

$$\forall \mathbf{x}, \mathbf{h} \in \mathbb{R}^d, \quad \forall 0 \le t \le 1, \quad f(\mathbf{x} + t\mathbf{h}) \le f(\mathbf{x}) + t\left[f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x})\right].$$

On insiste sur le fait qu'une fonction f convexe est forcément à valeur réelle, puisqu'il faut comparer des valeurs de f dans la définition de convexité. On s'intéresse dans ce chapitre aux fonctions convexes qui sont différentiables, ou deux fois différentiables. On aimerait trouver un critère simple sur les dérivées de f pour déterminer si f est convexe.

4.4.1 Fonctions convexes de classe C^1

Définition 4.23 : Fonctions monotones

Soit $F : \mathbb{R}^d \to \mathbb{R}^d$. On dit que F est **monotone** si

$$\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d, \quad \langle F(\mathbf{x}) - F(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \ge 0.$$

On remarquera que cette définition ne fait sens que pour des fonctions F ayant le même nombre de variables que de composantes. En dimension $d = 1$, la condition s'écrit

$$\forall x, y \in \mathbb{R}, \quad (F(x) - F(y)) \times (x - y) \ge 0.$$

Cela signifie que $F(x) \le F(y)$ ssi $x \le y$, et on trouve que F est monotone ssi F est croissante (on remarquera que la formulation précédente est indépendante de *l'orientation* de la droite réelle $\mathbb{R}$). L'importance de cette définition vient du résultat suivant.

Théorème 4.24 : Caractérisation des fonctions convexes de classe C^1

Soit $f : \mathbb{R}^d \to \mathbb{R}$ une fonction de classe C^1. Alors f est convexe ssi ∇f est monotone.

En dimension 1, on retrouve que f est convexe ssi f' est croissante.

Démonstration. Supposons pour commencer que f est convexe. On pose $\mathbf{h} = \mathbf{y} - \mathbf{x}$, et la deuxième inégalité de convexité avec $t > 0$ s'écrit

$$\frac{f(\mathbf{x} + t\mathbf{h}) - f(\mathbf{x})}{t} \le f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) = f(\mathbf{y}) - f(\mathbf{x}).$$

En passant à la limite $t \to 0$, et en utilisant que f est différentiable, on trouve

$$\langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \le f(\mathbf{y}) - f(\mathbf{x}).$$

En inversant les rôles de $\mathbf{x}$ et $\mathbf{y}$, on a aussi $\langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \le f(\mathbf{x}) - f(\mathbf{y})$. En additionnant ces deux inégalités, on trouve comme souhaité que ∇f est monotone, c'est à dire

$$\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \ge 0.$$

Réciproquement, supposons que ∇f soit monotone. En particulier, on a, pour tout $s \ge 0$, et tout $0 \le t \le 1$, que

$$\langle \nabla f(\mathbf{x} + s\mathbf{h}) - \nabla f(\mathbf{x} + ts\mathbf{h}), s(1-t)\mathbf{h} \rangle \ge 0, \quad \text{ou encore} \quad \langle \nabla f(\mathbf{x} + s\mathbf{h}), \mathbf{h} \rangle \ge \langle \nabla f(\mathbf{x} + ts\mathbf{h}), \mathbf{h} \rangle.$$

En intégrant cette relation pour s entre 0 et 1, on obtient que pour tout $0 \le t \le 1$, on a

$$f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) \ge \frac{f(\mathbf{x} + t\mathbf{h}) - f(\mathbf{x})}{t},$$

ce qui montre que f est convexe. □

4.4.2 Fonctions convexes de classe C^2

On s'intéresse maintenant aux fonctions convexes de classe C^2. Leur caractérisation est particulièrement simple, grâce au résultat suivant.

Théorème 4.25 : Caractérisation des fonctions convexes de classe C^2

Soit $f : \mathbb{R}^d \to \mathbb{R}$ une fonction de classe C^2. Alors f est convexe ssi sa Hessienne $H_f(\mathbf{x})$ est une matrice positive pour tout $\mathbf{x} \in \mathbb{R}^d$.

Dans le cas uni–dimensionnel $d = 1$, on retrouve que f est convexe ssi $f'' \geq 0$.

Démonstration. Supposons f convexe. Alors, ∇f est monotone, et on a, pour tout $t \neq 0$ et tout $\mathbf{x}, \mathbf{h} \in \mathbb{R}^d$,

$$\langle \nabla f(\mathbf{x} + t\mathbf{h}) - \nabla f(\mathbf{x}), t\mathbf{h} \rangle \geq 0, \quad \text{et donc} \quad \left\langle \frac{\nabla f(\mathbf{x} + t\mathbf{h}) - \nabla f(\mathbf{x})}{t}, \mathbf{h} \right\rangle \geq 0.$$

On rappelle que la Hessienne est la Jacobienne du gradient. En prenant la limite $t \to 0$, on a donc

$$\langle H_f(\mathbf{x})\mathbf{h}, \mathbf{h} \rangle \geq 0.$$

Ceci étant vrai pour tout $\mathbf{h} \in \mathbb{R}^d$, on en déduit que $H_f(\mathbf{x}) \in \mathcal{S}_d^+(\mathbb{R})$ pour tout $\mathbf{x} \in \mathbb{R}^d$.

Réciproquement, supposons que $\langle H_f(\mathbf{x})\mathbf{h}, \mathbf{h} \rangle \geq 0$ pour tout $\mathbf{x}, \mathbf{h} \in \mathbb{R}^d$. En particulier, on a

$$\langle H_f(\mathbf{x} + s\mathbf{h})\mathbf{h}, \mathbf{h} \rangle \geq 0.$$

En intégrant pour s entre 0 et 1, on obtient $\langle \nabla f(\mathbf{x} + \mathbf{h}) - \nabla f(\mathbf{x}), \mathbf{h} \rangle \geq 0$, ce qui montre que ∇f est monotone, donc que f est convexe. ☐

4.4.3 Points critiques des fonctions convexes

La caractérisation précédente des fonctions convexes permettent de démontrer que résultat suivant.

Théorème 4.26 : Les points critiques des fonctions convexes sont des minimiseurs

Soit $f : \mathbb{R}^d \to \mathbb{R}$ une fonction convexe de classe C^2. Si $\mathbf{x}_* \in \mathbb{R}^d$ est un point critique de f, alors $\mathbf{x}_*$ est un minimum global de f.

Démonstration. On pourra remarquer que $H_f(\mathbf{x}_*) \in \mathcal{S}_d^+(\mathbb{R})$ donc $\mathbf{x}_*$ est déjà un bon candidat pour être au moins un minimum local. Notons par exemple que si $H_f(\mathbf{x}_*) \in \mathcal{S}_d^{++}(\mathbb{R})$ est définie positive, alors $\mathbf{x}_*$ est un minimum local non dégénérée de f. Pour montrer que $\mathbf{x}_*$ est un minimum **global**, on utilise que ∇f est monotone, et donc

$$f(\mathbf{y}) - f(\mathbf{x}_*) = \int_0^1 \langle \nabla f(\mathbf{x}_* + t\mathbf{h}), \mathbf{h} \rangle \mathrm{d}t \geq \int_0^1 \langle \nabla f(\mathbf{x}_*), \mathbf{h} \rangle \mathrm{d}t = 0.$$

On a utilisé que $\nabla f(\mathbf{x}_*) = 0$ dans la dernière égalité. Ceci montre que $f(\mathbf{y}) \geq f(\mathbf{x}_*)$ pour tout $\mathbf{y} \in \mathbb{R}^d$, donc $\mathbf{x}_*$ est un minimiseur global. ☐

4.5 Méthode de gradient à pas constant

Nous avons vu comment trouver théoriquement des minima locaux, en étudiant les points critiques. Dans cette section, nous introduisons un algorithme simple, appelé *méthode de gradient à pas constant* pour trouver numériquement ces minima. La plupart des algorithmes connus pour la recherche d'optima sont des variations plus ou moins sophistiqués de cet algorithmes.

L'idée de la méthode est de générer une suite $(\mathbf{x}_n)$ qui converge vers un minimiseur local de f. Comme son nom l'indique, dans la méthode de *descente de gradient à pas constant*, on construit la suite par récurrence avec $\mathbf{x}_0$ un vecteur initial quelconque, puis

$$\boxed{\mathbf{x}_{n+1} = \mathbf{x}_n - \tau \nabla F(\mathbf{x}_n).} \qquad \text{(Itération du gradient à pas constant).} \tag{4.2}$$

Le paramètre $\tau > 0$ est le **pas** de la méthode. Le terme "pas constant" vient du fait que ce paramètre τ ne dépend de n. On remarque que la suite ne dépend que du gradient de f.

L'idée de la méthode est simple à comprendre : comme le gradient pointe vers la "direction qui monte le plus", on va chercher de l'autre côté (d'où le signe $-$ dans (4.2)). Dans cette section, on s'intéresse au choix de τ, qui est le seul paramètre du modèle : lorsque τ est trop petit, on fait des trop petits pas, et l'algorithme nécessite beaucoup d'itérations : la suite reste trop longtemps sur place. Lorsque τ est trop grand, la suite $(\mathbf{x}_n)$ oscille à côté de $\mathbf{x}_*$ sans le trouver. Il est donc important de bien choisir le pas τ.

4.5.1 Étude de la convergence pour une fonction quadratique

Dans cette section, nous étudions en détails les propriétés de l'algorithme de gradient à pas constant, dans le cas où la fonction est une fonction quadratique, de la forme

$$Q(\mathbf{x}) := \frac{1}{2}\langle \mathbf{x}, A\mathbf{x}\rangle - \langle \mathbf{b}, \mathbf{x}\rangle,$$

où $A \in \mathcal{S}_d^{++}$ est une matrice symétrique définie positive, et $\mathbf{b} \in \mathbb{R}^d$. On rappelle que la gradient de Q est donné par $\nabla Q(\mathbf{x}) := A\mathbf{x} - \mathbf{b}$. Le résultat principal de cette section est donné dans le lemme suivant.

Lemme 4.27 : Vitesse de convergence du gradient à pas constant, cas quadratique

Soit $A \in \mathcal{S}_d^{++}$ et $\mathbf{b} \in \mathbb{R}^d$, et soit $Q(\mathbf{x}) := \frac{1}{2}\langle \mathbf{x}, A\mathbf{x} - \langle \mathbf{b}, \mathbf{x}\rangle$. Soit $0 < \lambda_1 \le \cdots \le \lambda_d$ les valeurs propres de A rangées en ordre croissant. Soit enfin la suite $(\mathbf{x}_n)$ définie par $\mathbf{x}_0 \in \mathbb{R}^d$ et

$$\mathbf{x}_{n+1} = \mathbf{x}_n - \tau \nabla Q(\mathbf{x}_n) = \mathbf{x}_n - \tau(A\mathbf{x}_n - \mathbf{b}). \tag{4.3}$$

Alors la suite $(\mathbf{x}_n)$ converge vers $\mathbf{x}^* := A^{-1}\mathbf{b}$ si (et seulement si) $\tau\lambda_d < 2$. Dans ce cas, il existe on a

$$\|\mathbf{x}_n - \mathbf{x}_*\| \le \|\mathbf{x}_0 - \mathbf{x}_*\| \alpha^n, \qquad \text{avec} \quad \alpha := \max\{|1-\tau\lambda_1|, |1-\tau\lambda_d|\} \qquad \text{(taux de convergence).}$$

Le pas optimal est $\tau_* = \frac{2}{\lambda_1 + \lambda_d}$, et le taux vaut dans ce cas $\alpha_* = \frac{\lambda_d - \lambda_1}{\lambda_d + \lambda_1}$.

Démonstration. On pose $\mathbf{r}_n := \mathbf{x}_n - \mathbf{x}^*$ (l'erreur à l'itération n). Comme $\mathbf{x}^* = A^{-1}\mathbf{b}$, on a

$$\mathbf{r}_{n+1} = \mathbf{x}_{n+1} - \mathbf{x}^* = \mathbf{x}_n - \mathbf{x}^* - \tau A(\mathbf{x}_n - \mathbf{x}^*) = (\mathbb{I} - \tau A)(\mathbf{x}_n - \mathbf{x}^*) = (\mathbb{I} - \tau A)\mathbf{r}_n.$$

On a donc $\|\mathbf{r}_{n+1}\| \le \|\mathbb{I} - \tau A\|_{\mathrm{op}} \|\mathbf{r}_n\|$. La matrice $\mathbb{I} - \tau A$ est symétrique, et ses valeurs propres sont $1 - \tau\lambda_d \le \cdots \le 1 - \tau\lambda_1$. Ainsi, on a

$$\|\mathbf{r}_{n+1}\| \le \left(\max\{|1-\tau\lambda_1|, |1-\tau\lambda_d|\}\right) \|\mathbf{r}_n\| =: \alpha \|\mathbf{r}_n\|.$$

On en déduit que $(\mathbf{r}_n)$ converge vers $\mathbf{0}$, et donc $(\mathbf{x}_n)$ converge vers $\mathbf{x}^*$, si et seulement si $-1 < 1-\tau\lambda_d$ (condition dans le lemme) et si $1-\tau\lambda_1 < 1$ (ce qui est toujours vrai, car $\tau > 0$ et $\lambda_1 > 0$ par hypothèse).

Le taux optimal est atteint lorsque τ minimise $\tau \mapsto \max\{|1-\tau\lambda_1|, |1-\tau\lambda_d|\}$. Un calcul élémentaire montre que le minimum est atteint lorsque $|1-\tau\lambda_1| = |1-\tau\lambda_d|$, ou encore $1-\tau\lambda_1 = \tau\lambda_d - 1$, soit $\tau = \frac{2}{\lambda_1+\lambda_d}$. □

Le lemme 4.27 montre qu'on peut espérer une convergence rapide lorsque $\lambda_d \approx \lambda_1$, i.e. lorsque les valeurs propres de A sont du même ordre de grandeur. On dit dans ce cas que A est **bien conditionnée**.

4.5.2 Étude de la convergence dans le cas général

On s'intéresse maintenant aux propriétés de convergence dans le cas général. Dans ce cours, on ne s'intéressera qu'au cas où F est une fonction de classe C^2, et $\mathbf{x}^*$ est un minimiseur local de F *non dégénéré*, c'est à dire que $H_F(\mathbf{x}^*)$ est symétrique *définie* positive. On commence par un lemme utile d'algèbre linéaire.

Lemme 4.28

Pour tout $A, B \in \mathcal{S}_d(\mathbb{R})$, on a

$$|\lambda_1(A) - \lambda_1(B)| \le \|A-B\|_{\mathrm{op}} \quad \text{et} \quad |\lambda_d(A) - \lambda_d(B)| \le \|A-B\|_{\mathrm{op}}. \tag{4.4}$$

Démonstration. Soit $\mathbf{x}_B \in \mathbb{S}^{d-1}$ tel que $\lambda_1(B) = \langle \mathbf{x}_B, B\mathbf{x}_B\rangle$. D'après le principe du min/max, on a

$$\lambda_1(A) \le \langle \mathbf{x}_B, A\mathbf{x}_B\rangle = \langle \mathbf{x}_B, (A-B)\mathbf{x}_B\rangle + \langle \mathbf{x}_B, B\mathbf{x}_B\rangle \le \|A-B\|_{\mathrm{op}} + \lambda_1(B).$$

En inversant le rôle de A et B, on obtient (4.4). □

Autrement dit, les applications $A \mapsto \lambda_1(A)$ et $A \mapsto \lambda_d(A)$ sont 1-Lipschitz, donc continues. On en déduit le lemme suivant.

Lemme 4.29 : Continuité de la Hessienne

Soit $F : \mathbb{R}^d \to \mathbb{R}$ une fonction de classe C^2, et soit $\mathbf{x}^*$ un minimiseur non dégénéré de F. Soit $0 < \lambda_1 \le \cdots \le \lambda_d$ les valeurs propres de $H_F(\mathbf{x}^*)$. Alors, pour tout $0 < \ell < \lambda_1$ et tout $L > \lambda_d$, il existe $\varepsilon > 0$ tel que

$$\forall \mathbf{x} \in \mathcal{B}(\mathbf{x}^*, \varepsilon), \quad \ell \le H_F(\mathbf{x}) \le L.$$

On peut maintenant énoncer le résultat de convergence de la méthode du gradient à pas constant.

Théorème 4.30 : Vitesse de convergence du gradient à pas constant, cas général

Avec les mêmes notations que le Lemme 4.29, pour tout $0 < \tau < 2/L$ et pour tout $\mathbf{x}_0 \in \mathcal{B}(\mathbf{x}^*, \varepsilon)$, la suite définie par les itérations du gradient à pas constant (4.2) converge linéairement vers $\mathbf{x}^*$, à taux au plus $\max\{|1-\tau\ell|, |1-\tau L|\}$.

Démonstration. D'après la formule de Taylor à l'ordre 1 avec reste intégrale appliquée pour ∇F, et comme $\nabla F(\mathbf{x}^*) = \mathbf{0}$, on a, pour tout $\mathbf{x} \in \mathcal{B}(\mathbf{x}^*, \varepsilon)$,

$$\nabla F(\mathbf{x}) = \nabla F(\mathbf{x}^*) + \int_0^1 \left[H_F(\mathbf{x}^* + t(\mathbf{x}-\mathbf{x}^*))\right](\mathbf{x}-\mathbf{x}^*)\mathrm{d}t = \int_0^1 \left[H_F(\mathbf{x}^* + t(\mathbf{x}-\mathbf{x}^*))\right](\mathbf{x}-\mathbf{x}^*)\mathrm{d}t.$$

Supposons qu'à l'étape $n \in \mathbb{N}$, on a $\mathbf{x}_n \in \mathcal{B}(\mathbf{x}^*, \varepsilon)$, alors d'après la formule d'itération (4.2), on a

$$\begin{aligned}(\mathbf{x}_{n+1} - \mathbf{x}^*) = (\mathbf{x}_n - \mathbf{x}^*) - \tau \nabla F(\mathbf{x}_n) &= (\mathbf{x}_n - \mathbf{x}^*) - \tau \int_0^1 H_F(\mathbf{x}^* + t(\mathbf{x}_n - \mathbf{x}^*))(\mathbf{x}_n - \mathbf{x}^*) \mathrm{d}t \\ &= \left(1 - \tau \int_0^1 H_F(\mathbf{x}^* + t(\mathbf{x}_n - \mathbf{x}^*)) \mathrm{d}t \right) (\mathbf{x}_n - \mathbf{x}^*).\end{aligned}$$

La matrice entre parenthèse est une matrice symétrique dont les valeurs propres sont comprises entre $1 - \tau L$ et $1 - \tau \ell$. On en déduit que

$$\left\| \mathbf{x}_{n+1} - \mathbf{x}^* \right\| \le \alpha \left\| \mathbf{x}_n - \mathbf{x}^* \right\|, \quad \text{avec} \quad \alpha := \max \left\{ |1 - \tau \ell|, |1 - \tau L| \right\}.$$

Ainsi, si $\tau < 2/L$, on a $0 < \alpha < 1$. On en déduit premièrement que $\mathbf{x}_{n+1} \in \mathcal{B}(\mathbf{x}^*, \varepsilon)$ (et donc, par une récurrence immédiate, que $\mathbf{x}_n \in \mathcal{B}(\mathbf{x}^*, \varepsilon)$ pour tout $n \in \mathbb{N}$), puis que $|\mathbf{x}_n - \mathbf{x}^*| \le \alpha |\mathbf{x}_{n-1} - \mathbf{x}^*| \le \cdots \le \alpha^n |\mathbf{x}_0 - \mathbf{x}^*|$. ☐

On a ainsi démontré la convergence linéaire de la méthode du gradient à pas constant. D'après le Théorème 4.30 et un raisonnement similaire au Lemme 4.27, on voit que le taux optimal est atteint pour un pas $\tau \approx 2/(\lambda_1 + \lambda_d)$. Dans ce cas, on peut s'attendre à une convergence linéaire à taux $\frac{\lambda_d - \lambda_1}{\lambda_d + \lambda_1}$. On retrouve le fait que la convergence est rapide si la hessienne $H_F(\mathbf{x}^*)$ est bien conditionnée.

4.6 Visualisation des courbes de niveaux

4.6.1 Interprétation des points selles

CHAPITRE 5

THÉORÈME D'INVERSION LOCALE

Dans ce chapitre, on s'intéresse à la description des surfaces définies implicitement. Notre but est d'obtenir des théorèmes de la forme suivante : si $f(\mathbf{x}) = \mathbf{X}$, et si $\mathrm{d}f_{\mathbf{x}}$ est une application linéaire inversible, alors pour tout $\mathbf{Y}$ proche de $\mathbf{X}$, il existe un unique $\mathbf{y}$ proche de $\mathbf{x}$ tel que $f(\mathbf{y}) = \mathbf{Y}$.

5.1 Théorème de point fixe

On commence avec le théorème de points fixes, qui est un outil puissant pour démontrer l'existence et l'unicité de solutions.

Définition 5.1 : Fonctions Lipschitz

Soit E, F deux espaces vectoriels normés, et soit $f : E \to F$. On dit que f est L–Lipschitz si

$$\forall \mathbf{x}, \mathbf{y} \in E, \qquad \|f(\mathbf{x}) - f(\mathbf{y})\|_F \leq L\|\mathbf{x} - \mathbf{y}\|_E.$$

On dit que f est **contractante** si $E = F$, et si elle est α–Lipschitz pour un certain $0 \leq \alpha < 1$.

On pourra démontrer que les fonctions Lipschitz sont continues.

Théorème 5.2 : Théorème de point fixe

Soit $f : \mathbb{R}^d \to \mathbb{R}^d$ une application contractante. Alors f a un unique point fixe : il existe un unique $\mathbf{x}_* \in \mathbb{R}^d$ tel que $f(\mathbf{x}_*) = \mathbf{x}_*$.

On remarquera que l'espace de départ de f doit être le même que son espace d'arrivée afin que l'équation $f(\mathbf{x}_*) = \mathbf{x}_*$ ait un sens.

Démonstration. La preuve de ce théorème n'est pas difficile, mais elle est hors-programme, car elle nécessite la notion de **suites de Cauchy**, et de **complétude** de l'espace $\mathbb{R}^d$.

Existence. Posons $\mathbf{x}_0 \in \mathbb{R}^d$ un point quelconque et $\mathbf{x}_{n+1} = f(\mathbf{x}_n)$. On a, en utilisant le fait que f est α–Lipschitz, et par une récurrence immédiate, que pour tout $n \in \mathbb{N}$,

$$\|\mathbf{x}_{n+1} - \mathbf{x}_n\| = \|f(\mathbf{x}_n) - f(\mathbf{x}_{n-1})\| \leq \alpha\|\mathbf{x}_{n-1} - \mathbf{x}_{n-2}\| \leq \alpha^2\|\mathbf{x}_n - \mathbf{x}_{n-1}\| \leq \cdots \leq \alpha^n\|\mathbf{x}_1 - \mathbf{x}_0\|.$$

En particulier, pour tout $n, p \in \mathbb{N}$,

$$\begin{aligned}
\|\mathbf{x}_{n+p} - \mathbf{x}_n\| &\leq \|\mathbf{x}_{n+p} - \mathbf{x}_{n+p-1}\| + \|\mathbf{x}_{n+p-1} - \mathbf{x}_{n+p-2}\| + \cdots + \|\mathbf{x}_{n+1} - \mathbf{x}_n\| \\
&\leq (\alpha^{n+p-1} + \alpha^{n+p-2} + \cdots + \alpha^n)\,\|\mathbf{x}_1 - \mathbf{x}_0\| = \alpha^n \left(\alpha^{p-1} + \alpha^{p-2} + \cdots + 1\right) \|\mathbf{x}_1 - \mathbf{x}_0\| \\
&\leq \alpha^n \left(\sum_{k=0}^{\infty} \alpha^k\right) \|\mathbf{x}_1 - \mathbf{x}_0\| = \frac{\alpha^n}{1-\alpha} \|\mathbf{x}_1 - \mathbf{x}_0\|. \qquad (5.1)
\end{aligned}$$

On a utilisé le fait que $0 \leq \alpha < 1$ pour sommer la série géométrique. On en déduit que

$$\lim_{n\to\infty} \sup_{p\in\mathbb{N}} \|\mathbf{x}_{n+p} - \mathbf{x}_n\| = 0,$$

autrement dit, la suite $(\mathbf{x}_n)$ est de Cauchy dans l'espace $\mathbb{R}^d$, qui est complet. En particulier, il existe $\mathbf{x}_* \in \mathbb{R}^d$ tel que $\mathbf{x}_n \to \mathbf{x}_*$ dans $\mathbb{R}^d$. Par continuité de f, on a $f(\mathbf{x}_*) = \lim_{n\to\infty} f(\mathbf{x}_n) = \lim_{n\to\infty} \mathbf{x}_{n+1} = \mathbf{x}_*$, donc $\mathbf{x}_* \in \mathbb{R}^d$ est un point fixe de f.

Unicité. Soit $\mathbf{x}_*$ et $\mathbf{y}_*$ deux points fixes de f. On a

$$\|\mathbf{x}_* - \mathbf{y}_*\| = \|f(\mathbf{x}_*) - f(\mathbf{y}_*)\| \leq \alpha\|\mathbf{x}_* - \mathbf{y}_*\|, \quad \text{donc} \quad 0 \leq (1-\alpha)\|\mathbf{x}_* - \mathbf{y}_*\| \leq 0,$$

ce qui implique $\mathbf{x}_* = \mathbf{y}_*$, car $0 \leq \alpha < 1$.

□

5.2 Théorème d'inversion locale

5.2.1 Difféomorphismes

Dans la suite, on considère des fonctions d'un ouvert $U \subset \mathbb{R}^d$ dans $V \subset \mathbb{R}^n$, où U et V sont des ouverts. On écrira $f : U \subset \mathbb{R}^d \to V \subset \mathbb{R}^n$ pour insister sur ce point.

Définition 5.3 : Difféomorphisme

Soit $U \subset \mathbb{R}^d$, $V \subset \mathbb{R}^n$ et $f : U \to V$. On dit que f est un **difféomorphisme** de U dans V si

- f est bijective de U dans V,
- f est de classe C^1 sur U,
- f^{-1} (fonction inverse) est de classe C^1 sur V.

Comme f est bijective de U dans V, on a forcément $V = f(U)$ et $U = f^{-1}(V)$. Lorsque f est f^{-1} sont de classe C^k, on parle parfois de C^k–difféomorphisme. Si f est un difféomorphisme avec $U = \mathbb{R}^d$ dans $V = f(\mathbb{R}^d)$, on parle de **difféomorphisme global**.

Lemme 5.4

Si $f : U \subset \mathbb{R}^d \to V \subset \mathbb{R}^n$ est un difféomorphisme local, alors $d = n$.

Cela signifie qu'il n'existe pas de difféomorphismes entre des espaces qui n'ont pas la même dimension.

Démonstration. Pour tout $\mathbf{x} \in U$, on a $f^{-1}(f(\mathbf{x})) = \mathbf{x}$. En dérivant, et en utilisant la règle de la chaîne, on obtient

$$\forall \mathbf{x} \in U \subset \mathbb{R}^d, \quad J_{f^{-1}}(f(\mathbf{x}))J_f(\mathbf{x}) = \mathbb{I}_{\mathbb{R}^d}.$$

De même, pour tout $\mathbf{y} \in V$, on a $f(f^{-1}(\mathbf{y})) = \mathbf{y}$, donc

$$\forall \mathbf{y} \in V \subset \mathbb{R}^n, \quad J_f(f^{-1}(\mathbf{y}))J_{f^{-1}}(\mathbf{y}) = \mathbb{I}_{\mathbb{R}^n}.$$

Ainsi, pour $\mathbf{x}_0 \in U$ et en posant $\mathbf{y}_0 := f(\mathbf{x}_0) \in V$, on a

$$J_{f^{-1}}(\mathbf{y}_0)J_f(\mathbf{x}_0) = \mathbb{I}_{\mathbb{R}^d}, \quad \text{et} \quad J_f(\mathbf{x}_0)J_{f^{-1}}(\mathbf{y}_0) = \mathbb{I}_{\mathbb{R}^n}.$$

On en déduit que la matrice $J_f(\mathbf{x}_0)$, de taille $n \times d$, est inversible, d'inverse $J_{f^{-1}}(\mathbf{y}_0)$. En particulier, elle est carrée, donc $d = n$. ☐

Définition 5.5 : Difféomorphisme local en un point

Soit $f : \mathbb{R}^d \to \mathbb{R}^d$ et soit $\mathbf{a} \in \mathbb{R}^d$. On dit que f est un difféomorphisme local en $\mathbf{a}$ si il existe un voisinage U de $\mathbf{a}$ et un voisinage V de $f(\mathbf{a})$ tel que f soit un difféomorphisme de U dans V.

Exemple :

- La fonction $f : \mathbb{R} \to \mathbb{R}$ définie par $f(x) = x^3$ est bijective, d'inverse $f^{-1}(x) = x^{1/3}$. La fonction f est de classe C^1 sur $\mathbb{R}$, mais f^{-1} est de classe C^1 sur $\mathbb{R}^*$. On en déduit que f est un C^1 difféomorphisme local en a pour tout $a \in \mathbb{R}^*$.
- Si $f : \mathbb{R}^d \to \mathbb{R}^d$ est un difféomorphisme global, alors f est un difféomorphisme local en tout point $\mathbf{a} \in \mathbb{R}^d$ (prendre $U = V = \mathbb{R}^d$). Cependant, la réciproque n'est pas vraie. Par exemple, la fonction $f(x, y) := (\mathrm{e}^x \cos(y), \mathrm{e}^x \sin(y))$ est un difféomorphisme local en tout point (voir plus tard), mais n'est pas un difféomorphisme global, car f n'est pas injective (donc pas bijective). En effet, on a $f(x, y + 2\pi) = f(x, y)$.
- Soit $L \in \mathcal{L}(\mathbb{R}^d, \mathbb{R}^d)$ une application linéaire. L est un difféomorphisme global ssi L est inversible.

5.2.2 Théorème d'inversion locale

On peut maintenant énoncer le théorème d'inversion locale.

Théorème 5.6 : Théorème d'inversion locale

Soit $f : \mathbb{R}^d \to \mathbb{R}^d$ une fonction de classe C^1, et soit $\mathbf{a} \in \mathbb{R}^d$. Alors f est un difféomorphisme local en $\mathbf{a}$ ssi $J_f(\mathbf{a})$ est inversible.

On rappelle que la matrice $J_f(\mathbf{a})$ est inversible ssi l'application linéaire $\mathrm{d}f_{\mathbf{a}}$ est inversible. Ainsi, localement, une application **non linéaire** f est une *bijection locale* (au sens difféomorphisme) en $\mathbf{a}$ ssi son linéarisé $\mathrm{d}f_{\mathbf{a}}$ l'est.

On donne deux preuves de ce résultat. La première est spécifique à la dimension 1, mais a le mérite d'être facile à comprendre. La seconde donne la preuve dans le cas général. On rappelle que si f est difféomorphisme local, alors J_f est toujours inversible. On se concentre sur la réciproque : on suppose que $J_f(\mathbf{a})$ (donc $\mathrm{d}f_{\mathbf{a}}$) est inversible, et on veut montrer que f est un difféomorphisme local en $\mathbf{a}$.

Première preuve, dans le cas $d = 1$. Dans le cas unidimensionnel on a $J_f(a) = f'(a)$, qui est inversible ssi $f'(a) \neq 0$. On suppose $f'(a) > 0$ (la preuve pour $f'(a) < 0$ est similaire). Par continuité de f', il existe $\varepsilon > 0$ tel que $f'(x) > 0$ pour tout $x \in (a - \varepsilon, a + \varepsilon)$. Ainsi, f est strictement croissante et continue de $U := (a - \varepsilon, a + \varepsilon)$ dans $V = (f(a - \varepsilon), f(a + \varepsilon))$. On en déduit que f est bijective de U dans V. De plus, f est de classe C^1 par hypothèse.

On admettra la régularité de f^{-1}.

☐

Deuxième preuve, dans le cas général. Afin de simplifier les notations, on travaillera avec $g(\mathbf{h}) := f(\mathbf{a} + \mathbf{h}) - f(\mathbf{a})$, qui vérifie $g(\mathbf{0}) = \mathbf{0}$, et $\mathrm{d}g_{\mathbf{0}} = \mathrm{d}f_{\mathbf{a}}$ inversible. La fonction g est une version translatée de f, de sorte que f est un difféomorphisme local en $\mathbf{a}$ ssi g est un difféomorphisme local en $\mathbf{0}$.

Notre but est de montrer que g est localement surjective autour de $\mathbf{0}$: pour tout $\mathbf{y}$ proche de $\mathbf{0}$, il existe un unique $\mathbf{x}$ proche de $\mathbf{0}$ tel que l'équation $g(\mathbf{x}) = \mathbf{y}$ a une solution. Ceci montrera que g est surjective localement autour de $\mathbf{0}$, avec $\mathbf{x} = g^{-1}(\mathbf{y})$.

Soit $\mathbf{y} \in \mathbb{R}^d$ quelconque. On pose

$$\Phi_{\mathbf{y}}(\mathbf{x}) := \mathbf{x} - (\mathrm{d}g_{\mathbf{0}})^{-1}\left(g(\mathbf{x}) - \mathbf{y}\right). \tag{5.2}$$

On remarque que $\mathbf{x}$ est un point fixe de $\Phi_{\mathbf{y}}$ ssi $g(\mathbf{x}) = \mathbf{y}$. Ainsi, on se ramène à l'étude de point fixe de $\Phi_{\mathbf{y}}$. L'idée de la preuve est de montrer que pour tout $\mathbf{y}$ suffisamment petit, $\Phi_{\mathbf{y}}$ est une application contractante.

L'application $\Phi_{\mathbf{y}}$ est de classe C^1, avec

$$\mathrm{d}(\Phi_{\mathbf{y}})_{\mathbf{x}} = \mathbb{I}_d - (\mathrm{d}g_{\mathbf{0}})^{-1}(\mathrm{d}g_{\mathbf{x}}).$$

En particulier, en $\mathbf{y} = \mathbf{0}$, on a $\mathrm{d}(\Phi_{\mathbf{0}})_{\mathbf{0}} = \mathbb{I}_d - (\mathrm{d}g_{\mathbf{0}})^{-1}(\mathrm{d}g_{\mathbf{0}}) = 0$ qui est l'application nulle. Soit $0 < \alpha < 1$. Par continuité de $(\mathbf{x}, \mathbf{y}) \mapsto \mathrm{d}(\Phi_{\mathbf{y}})_{\mathbf{x}}$, on en déduit qu'il existe $\varepsilon > 0$ tel que

$$\forall \mathbf{x} \in \mathcal{B}(\mathbf{0}, \varepsilon), \quad \forall \mathbf{y} \in \mathcal{B}(\mathbf{0}, \varepsilon), \qquad \| \mathrm{d}(\Phi_{\mathbf{y}})_{\mathbf{x}} \| < \alpha.$$

On en déduit que pour tout $\mathbf{y} \in \mathcal{B}(\mathbf{0}, \varepsilon)$, l'application $\Phi_{\mathbf{y}}$ est contractante de $\mathcal{B}(0, \varepsilon)$ dans lui-même. D'après le théorème de point fixe, il existe un unique point fixe de $\Phi_{\mathbf{y}}$ dans $\mathcal{B}(0, \varepsilon)$. Ceci définit une application g^{-1} de $V := \mathcal{B}(0, \varepsilon)$ dans $U = g^{-1}(V)$, et montre que g est bijective de U dans V.

De nouveau, on admet la régularité de g^{-1}. ☐

5.2.3 Algorithme de Newton (*)

La fonction $\Phi_{\mathbf{y}}$ définie précédemment, est utilisée dans l'algorithme de Newton pour résoudre des équations de type $f(\mathbf{x}) = \mathbf{0}$. Dans le cas où $f = \nabla F$, résoudre l'équation $f(\mathbf{x}) = \mathbf{0}$ permet de trouver les points critiques de F (dont, en particulier, les minima locaux). Nous énonçons l'algorithme de Newton seulement dans le cas où $f = \nabla F$, qui est le cas qui nous intéresse.

Soit $F : \mathbb{R}^d \to \mathbb{R}$ une fonction de classe C^3, soit $\mathbf{x}_* \in \mathbb{R}^d$ un point critique de F, et soit $\mathbf{x}_0 \in \mathbb{R}^d$ un point «proche» de $\mathbf{x}_*$. On introduit la suite (de Newton)

$$\mathbf{x}_{n+1} = \mathbf{x}_n - \left[H_F(\mathbf{x}_n)\right]^{-1} \nabla F(\mathbf{x}_n) \qquad \text{(Algorithme de Newton).} \tag{5.3}$$

En comparant avec (5.2), on voit que $\mathbf{x}_{n+1} = \Phi_{\mathbf{0}}(\mathbf{x}_n)$ avec $g = \nabla F : \mathbb{R}^d \to \mathbb{R}^d$. Le théorème suivant assure que $\mathbf{x}_n$ converge très rapidement vers $\mathbf{x}_*$.

Théorème 5.7

Avec les mêmes notations, si $\mathbf{x}_*$ est un point critique non dégénéré de F, alors il existe $M > 0$ et $\varepsilon > 0$ avec $M\varepsilon < 1$ tel que pour tout $\mathbf{x}_0 \in \mathcal{B}(\mathbf{x}_*, \varepsilon)$, la suite définie par (5.3) vérifie $\mathbf{x}_n \in \mathcal{B}(\mathbf{x}_*, \varepsilon)$ pour tout n, et il

$$\| \mathbf{x}_n - \mathbf{x}_* \| \le M^{-1} \alpha^{2^n}, \qquad \text{avec} \quad \alpha := M \| \mathbf{x}_0 - \mathbf{x}_* \| < 1.$$

Démonstration. Nous présentons la preuve en dimension $d = 1$ pour simplifier. Pour commencer, on remarque que comme x_* est un point critique non dégénéré, $H_F(x_*) = F''(x_*)$ est inversible. En particulier, par continuité de F'', il existe $\varepsilon' > 0$ tel que $F''(x) \neq 0$ pour tout $x \in \mathcal{B}(x_*, \varepsilon')$. On pose

$$\Phi(x) := x - \frac{F'(x)}{F''(x)},$$

qui est bien définie de $\mathcal{B}(x_*, \varepsilon')$ dans $\mathbb{R}$. On remarque que la suite de Newton vérifie $x_{n+1} = \Phi(x_n)$. De plus, on a $\Phi(x_*) = x_*$. Enfin, on a

$$\Phi'(x) = 1 - \frac{F''(x)F''(x) - F'(x)F'''(x)}{|F''(x)|^2} = \frac{F'(x)F'''(x)}{|F''(x)|^2},$$

et, en particulier, $\Phi'(x_*) = 0$. Ainsi, en utilisant le théorème de Taylor–Young à l'ordre 2, on a

$$x_{n+1} - x_* = \Phi(x_n) - \Phi(x_*) = \underbrace{\Phi'(x_*)}_{=0}(x_n - x_*) + \left[\int_0^1 (1-t) F''(x_* + t(x_n - x_*)) \mathrm{d}t \right] (x_n - x_*)^2.$$

En particulier, avec $M := \sup_{y \in \mathcal{B}(x_*, \varepsilon')} |F''(y)|$, on a

$$|x_{n+1} - x_*| \le M |x_n - x_*|^2.$$

On pose $\varepsilon := \min\{\varepsilon', \frac{\varepsilon'}{M}\}$, de sorte que si $x_n \in \mathcal{B}(x_*, \varepsilon)$, on a $x_{n+1} \in \mathcal{B}(x_*, \varepsilon)$ aussi, et donc toute la suite est dans $\mathcal{B}(x_*, \varepsilon)$. Une récurrence immédiate donne alors

$$\begin{aligned} |x_{n+1} - x_*| &\le M|x_n - x_*|^2 \le M \left(M |x_{n-1} - x_*|^2 \right)^2 = M^{1+2} |x_{n-1} - x_*|^4 \\ &\le M^{1+2} \left(M |x_{n-2} - x_*|^2 \right)^4 = M^{1+2+4} |x_{n-2} - x_*|^8 \\ &\le \cdots \le M^{1+2+4+\cdots+2^{n-1}} |x_0 - x_*|^{2^n} \le M^{2^n - 1} |x_0 - x_*|^{2^n} = M^{-1} \left(M |x_0 - x_*| \right)^{2^n}. \end{aligned}$$

□

Si on compare avec l'algorithme du gradient à pas constant, la vitesse de convergence est beaucoup plus rapide (vitesse en α^{2^n} au lieu de α^n). On dit que l'algorithme de Newton converge **quadratiquement** ! De plus, l'algorithme de Newton permet de trouver n'importe quel point critique, dont des point selles (alors que l'algorithme du gradient ne trouve que des minima locaux). Cependant, l'algorithme de Newton nécessite :

— d'avoir une initialisation $\mathbf{x}_0$ proche du point qu'on cherche $\mathbf{x}_*$;

— d'inverser la Hessienne $H_F(\mathbf{x}_n)$ à chaque étape. La hessienne étant de taille $d \times d$, cette inversion peut être très coûteuse numériquement si d est très grand.

5.2.4 Théorème des fonctions implicites

Une des formes les plus utilisées du théorème d'inversion locale est la suivante. Dans la suite, on considère des fonctions à deux variables $f(\mathbf{x}, \mathbf{y})$ avec $\mathbf{x} \in \mathbb{R}^p$ et $\mathbf{y} \in \mathbb{R}^d$ (chacune des variables peut avoir plusieurs composantes). On notera $\partial_{\mathbf{y}} f$ la Jacobienne de f selon la variable $\mathbf{y}$ uniquement. Explicitement, si f a n composantes, $\partial_{\mathbf{y}} f$ est une matrice de taille $n \times d$, de composante

$$\partial_{\mathbf{y}} f = \begin{pmatrix} \dfrac{\partial f_1}{\partial y_1} & \dfrac{\partial f_1}{\partial y_2} & \cdots & \dfrac{\partial f_1}{\partial y_d} \\ \vdots & \vdots & \vdots & \vdots \\ \dfrac{\partial f_n}{\partial y_1} & \dfrac{\partial f_n}{\partial y_2} & \cdots & \dfrac{\partial f_n}{\partial y_d} \end{pmatrix}.$$

Théorème 5.8 : Théorème des fonctions implicites

Soit $f = f(\mathbf{x}, \mathbf{y}) : \mathbb{R}^p \times \mathbb{R}^d \to \mathbb{R}^d$ une fonction de classe C^1. Soit $(\mathbf{x}_0, \mathbf{y}_0) \in \mathbb{R}^p \times \mathbb{R}^d$ un point tel que $f(\mathbf{x}_0, \mathbf{y}_0) = \mathbf{0}$ et tel que $\partial_{\mathbf{y}} f(\mathbf{x}_0, \mathbf{y}_0)$ soit inversible.
Alors il existe un voisinage U de $\mathbf{x}_0$ dans $\mathbb{R}^p$, un voisinage V de $\mathbf{y}_0$ et une fonction $\varphi : U \to V$ de classe C^1 tel que

$$\forall (\mathbf{x}, \mathbf{y}) \in U \times V, \quad f(\mathbf{x}, \mathbf{y}) = \mathbf{0} \quad \text{ssi} \quad \mathbf{y} = \varphi(\mathbf{x}).$$

Nous admettons la preuve, qui est une conséquence directe du théorème d'inversion locale.

5.3 Surfaces

Dans cette section, nous introduisons la notion de **surface** (ou **variété**, ou **sous-variété**) de dimension n dans $\mathbb{R}^d$.

5.3.1 Première définition et exemples

Définition 5.9

Soit $S \subset \mathbb{R}^d$. On dit que S est une **surface de dimension** n si, pour tout $\mathbf{a} \in S$, il existe

- un ouvert U de $\mathbb{R}^d$ contenant $\mathbf{a}$;
- un ouvert B de $\mathbb{R}^d = \mathbb{R}^n \times \mathbb{R}^{d-n}$ contenant $\mathbf{0}$;
- un difféomorphisme local $\alpha : U \to V$ vérifiant $\alpha(\mathbf{a}) = \mathbf{0}$

tel que, pour tout $\mathbf{x} \in U$,

$$\mathbf{x} \in S \quad \Longleftrightarrow \quad \alpha(\mathbf{x}) \in \mathbb{R}^n \times \{\mathbf{0}\}.$$

La dernière proposition indique que les $n - d$ dernières coordonnées de α sont nulles. Une autre façon de dire est que (on rappelle que $U = \alpha^{-1}(V)$)

$$S \cap U = \alpha^{-1}((\mathbb{R}^n \times \{\mathbf{0}\}) \cap V).$$

Ainsi, localement, S est la déformation du sous-espace vectoriel $\mathbb{R}^n \times \{\mathbf{0}\} \subset \mathbb{R}^d$ par le difféomorphisme α^{-1}. Une surface de dimensions n est donc un ensemble de points qui ressemble en tout point à un hyperplan affine de dimension n, dans le sens par exemple où on peut définir un hyperplan tangent à S en tout point de S.

Par convention,

- $S \in \mathbb{R}^d$ est une surface de dimension 0 ssi S est un ensemble de points sans accumulation (pour tout $\mathbf{x} \in \mathbb{R}^d$, il existe $\varepsilon > 0$ tel que l'intersection $S \cap \mathcal{B}(\mathbf{x}, \varepsilon)$ soit au plus d'un point).
- $S \in \mathbb{R}^d$ est une surface de dimension d ssi S est un ouvert de $\mathbb{R}^d$.

Voici quelques exemples de sous-ensembles $S \subset \mathbb{R}^d$ qui ne sont **pas** des surfaces.
***** TODO *****.

Théorème 5.10 : Les graphes sont des surfaces.

Soit $f : \mathbb{R}^n \to \mathbb{R}^p$ une fonction de classe C^1. On pose $d := n + p$, et

$$G(f) := \{(\mathbf{x}, f(\mathbf{x})), \quad \mathbf{x} \in \mathbb{R}^n\} \subset \mathbb{R}^d, \quad \text{(graphe de } f\text{)}.$$

Alors $G(f)$ est une surface de dimension n dans $\mathbb{R}^d$.

Démonstration. On écrit $\mathbb{R}^d = \mathbb{R}^n \times \mathbb{R}^p$, et on écrit $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^d$ avec $\mathbf{x} \in \mathbb{R}^n$ et $\mathbf{y} \in \mathbb{R}^p$. On introduit les fonctions $\alpha : \mathbb{R}^d \to \mathbb{R}^d$ et $\beta : \mathbb{R}^d \to \mathbb{R}^d$, définies par

$$\alpha(\mathbf{x}, \mathbf{y}) := \begin{pmatrix} \mathbf{x} \\ \mathbf{y} - f(\mathbf{x}) \end{pmatrix} \quad \text{et} \quad \beta(\mathbf{x}, \mathbf{y}) := \begin{pmatrix} \mathbf{x} \\ \mathbf{y} + f(\mathbf{x}) \end{pmatrix}$$

Comme f est de classe C^1, les fonctions α et β sont de classe C^1. De plus, il est facile de voir que $(\mathbf{X}, \mathbf{Y}) = \alpha(\mathbf{x}, \mathbf{y})$ ssi $(\mathbf{x}, \mathbf{y}) = \beta(\mathbf{X}, \mathbf{Y})$, donc $\beta = \alpha^{-1}$. Ainsi, α est un difféomorphisme (global), d'inverse β. Par ailleurs, on a

$$\alpha(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^n \times \{\mathbf{0}\} \Longleftrightarrow \mathbf{y} - f(\mathbf{x}) = \mathbf{0} \Longleftrightarrow (\mathbf{x}, \mathbf{y}) \in G(f).$$

Ceci montre que $G(f)$ est une surface de dimension n dans $\mathbb{R}^d$. $\square$

5.3.2 Submersions

En pratique, le critère donné précédemment pour vérifier que $S \subset \mathbb{R}^d$ est une surface est inadéquat : il faut construire un difféomorphisme local en chaque point de S... Nous allons maintenant définir des surfaces *par contraintes*, ce qui, en pratique, est beaucoup plus utile. Pour cela, nous définisson la notion de submersion.

Définition 5.11 : Submersion

Soit $g : \mathbb{R}^d \to \mathbb{R}^p$ de classe C^1. On dit que g est une **submersion** en $\mathbf{a} \in \mathbb{R}^d$ si $\mathrm{d}g_{\mathbf{a}} \in \mathcal{L}(\mathbb{R}^d, \mathbb{R}^p)$ est une application linéaire surjective (donc, en particulier, $d \geq p$).

Cela est équivalent à dire que la matrice jacobienne $J_g(\mathbf{a}) \in \mathbb{R}^p \times \mathbb{R}^n$ est **surjective**, donc que ses colonnes engendrent tout l'espace $\mathbb{R}^p$.

Théorème 5.12 : Le cas $p = 1$

Soit $g : \mathbb{R}^d \to \mathbb{R}$ de classe C^1. Alors g est submersion en $\mathbf{a} \in \mathbb{R}^d$ ssi $\nabla g(\mathbf{a}) \neq \mathbf{0}$.

Démonstration. On a, par définition du gradient, $\mathrm{d}g_{\mathbf{a}}(\mathbf{h}) := \langle \nabla g(\mathbf{a}), \mathbf{h} \rangle$. Si $\nabla g(\mathbf{a}) = \mathbf{0}$, $\mathrm{d}g_{\mathbf{a}}$ est l'application linéaire nulle, et n'est pas surjective. Réciproquement, si $\nabla g(\mathbf{a}) \neq \mathbf{0}$, alors, pour tout $X \in \mathbb{R}$, on a, avec $\mathbf{h} := X \frac{\nabla g(\mathbf{a})}{\|\nabla g(\mathbf{a})\|^2}$, que $\mathrm{d}g_{\mathbf{a}}(\mathbf{h}) = X$, donc $\mathrm{d}g_{\mathbf{a}}$ est surjective. ☐

L'importance des submersions est donnée par le résultat suivant (parfois appelé le théorème des fonctions implicites géométrique).

Théorème 5.13 : les ensembles de niveau des submersions sont des surfaces.

Soit $g : \mathbb{R}^d \to \mathbb{R}^p$ de classe C^1, et soit $\boldsymbol{\lambda} \in \mathbb{R}^p$. On pose

$$S_{\boldsymbol{\lambda}} := g^{-1}(\{\boldsymbol{\lambda}\}) = \left\{ \mathbf{x} \in \mathbb{R}^d, \quad g(\mathbf{x}) = \boldsymbol{\lambda} \right\}.$$

Si, pour tout $\mathbf{a} \in S_{\boldsymbol{\lambda}}$, g est une submersion en $\mathbf{a}$, alors $S_{\boldsymbol{\lambda}}$ est une surface de dimension $n = d - p$ dans $\mathbb{R}^d$.

Démonstration. On fait la preuve dans le cas $\boldsymbol{\lambda} = \mathbf{0}$ (on peut se ramener à ce cas en considérant $g - \boldsymbol{\lambda}$). Soit $\mathbf{a} \in S_{\mathbf{0}}$. On veut construire un difféomorphisme α en $\mathbf{a}$...

Comme g est une submersion en $\mathbf{a}$, la matrice $J_g(\mathbf{a}) \in \mathbb{R}^p \times \mathbb{R}^d$ est surjective. Donc les d colonnes de $J_g(\mathbf{a})$ engendrent tout l'espace $\mathbb{R}^p$, et on peut trouver un sous-ensemble de p colonnes qui forment une base de $\mathbb{R}^p$ (on extrait une sous-matrice de taille $p \times p$ inversible). Quitte à échanger les vecteurs de la base canonique, on peut supposer que ce sont les p dernières colonnes. Ainsi, en notant $\mathbb{R}^d \ni \mathbf{x} = (\mathbf{z}, \mathbf{y}) \in \mathbb{R}^{d-p} \times \mathbb{R}^p$, et $g(\mathbf{x}) = g(\mathbf{z}, \mathbf{y}) : \mathbb{R}^{d-p} \times \mathbb{R}^p$, on a $\partial_{\mathbf{y}} g \in \mathcal{M}_{p \times p}$ inversible.

On a $g(\mathbf{a}) = \mathbf{0}$, et $\partial_{\mathbf{y}} g(\mathbf{a})$ inversible. On peut donc appliquer le Théorème des fonctions implicites. On en déduit qu'il existe une fonction $\phi : \mathbb{R}^{d-p} \to \mathbb{R}^p$ de classe C^1 tel que, localement autour de $\mathbf{a}$, $g(\mathbf{x}, \mathbf{y}) = \mathbf{0}$ ssi $\mathbf{y} = \phi(\mathbf{a})$. Cela signifie exactement que, localement autour de $\mathbf{a}$, l'ensemble $S_{\boldsymbol{\lambda}}$ est le graphe de ϕ. On en déduit que $S_{\boldsymbol{\lambda}}$ est une surface, cf Théorème 5.10. ☐

L'interprétation de $S_{\boldsymbol{\lambda}}$ est la suivante. On note $g = (g_1, g_2, \cdots, g_p)$ les composantes de g. L'ensemble $S_{\boldsymbol{\lambda}}$ est l'ensemble des points $\mathbf{x}$ qui vérifient

$$g_1(\mathbf{x}) = \lambda_1, \quad g_2(\mathbf{x}) = \lambda_2, \quad \cdots, \quad g_p(\mathbf{x}) = \lambda_p,$$

c'est donc l'ensemble des points $\mathbf{x}$ qui vérifient un ensemble de p contraintes scalaires. Plus il y a de contraintes, moins il y a de points. Cela explique pourquoi la dimension de $S_{\boldsymbol{\lambda}}$ est $d - p$ (= *nombre* *de variables - nombres d'équations*). Le fait que g soit une submersion assure en quelque sorte que ces contraintes ne sont pas redondantes.

Exemple 5.14. *Soit* $g_\lambda : \mathbb{R}^3 \to \mathbb{R}^2$ *définie par*

$$g_\lambda(x, y, z) = \begin{pmatrix} x^2 + y^2 + z^2 - 1 \\ x - \lambda \end{pmatrix}.$$

On s'intéresse à l'ensemble $C_\lambda = g_\lambda^{-1}(\{\mathbf{0}\})$*. C'est l'ensemble des points qui vérifient* $x^2 + y^2 + z^2 = 1$ *(donc sur la sphère de rayon 1), et* $x = \lambda$ *(donc sur un hyperplan). L'ensemble* C_λ *est donc l'intersection entre une sphére et un plan. Ainsi,*

- *Si* $|\lambda| < 1$*, l'intersection est un cercle (surface de dimension* 1*) ;*
- *Si* $|\lambda| = 1$*, l'intersection est un point (surface de dimension 0) ;*
- *Si* $|\lambda| > 1$*, l'intersection est vide (pas une surface).*

Par ailleurs on a

$$J_{g_\lambda}(x, y, z) = \begin{pmatrix} 2x & 2y & 2z \\ 1 & 0 & 0 \end{pmatrix},$$

qui est surjective, sauf si $y = z = 0$*. Les seuls points de la sphère vérifiant* $y = z = 0$ *sont* $(\pm 1, 0, 0)$*, qui sont sur* C_λ *ssi* $\lambda = \pm 1$*. On retrouve que si* $|\lambda| < 1$*,* g_λ *est une submersion sur tout* C_λ*, donc que* C_λ *est une surface de dimension* $1 = 3 - 2$*.*

5.3.3 Quelques exemples de surfaces

La sphère. On pose

$$\mathbb{S}^{d-1} := \left\{ \mathbf{x} \in \mathbb{R}^d, \quad \|\mathbf{x}\| = 1 \right\}.$$

En posant $g(\mathbf{x}) := \|\mathbf{x}\|^2$, de $\mathbb{R}^d$ dans $\mathbb{R}$, qui est de classe C^1, on voit que $\mathbb{S}^{d-1} = g^{-1}(1)$. On a $\nabla g(\mathbf{x}) = 2\mathbf{x}$, qui est nulle si et seulement si $\mathbf{x} = \mathbf{0}$. Or le point $\mathbf{0}$ n'est pas dans $\mathbb{S}^{d-1}$ (car $\|\mathbf{0}\| = 0 \neq 1$), donc g est une submersion sur $\mathbb{S}^{d-1}$.

On en déduit que $\mathbb{S}^{d-1}$ est une surface de dimension $d - 1$ dans $\mathbb{R}^d$ (d'où la notation $\mathbb{S}^{d-1}$).

Les hyperplans affines. On pose

$$P := \left\{ A\mathbf{x} + \mathbf{b}, \quad \mathbf{x} \in \mathbb{R}^d \right\},$$

où $A \in \mathcal{M}_{n\times d}$ et $\mathbf{b} \in \mathbb{R}^n$. P est le graphe de la fonction $f(\mathbf{x}) := A\mathbf{x} + \mathbf{b}$ de classe C^1 de $\mathbb{R}^d$ dans $\mathbb{R}^n$, donc P est une surface de dimension n dans $\mathbb{R}^d$.

On peut définir les hyperplans par contraintes aussi. Soit $B \in \mathcal{M}_{p\times d}$ une matrice surjective, et $\boldsymbol{\lambda} \in \mathbb{R}^p$. On pose

$$P' := \left\{ \mathbf{x} \in \mathbb{R}^d, \quad B\mathbf{x} = \boldsymbol{\lambda} \right\}.$$

C'est l'ensemble de niveau $g^{-1}(\boldsymbol{\lambda})$ avec $g(\mathbf{x}) := B\mathbf{x}$. Comme g est linéaire, on a $\mathrm{d}g_{\mathbf{a}} = g$, qui est surjective car B l'est. Donc g est une submersion globale, et P' est une surface de dimension $d - p$.

Dans le cas où $B \in \mathcal{M}_{1\times d}$ est un vecteur ligne, et en posant $\mathbf{v} := B^\perp$, on a $P' = \left\{ \langle \mathbf{v}, \mathbf{x} \rangle = \lambda, \ \mathbf{x} \in \mathbb{R}^d \right\}$, qui est l'hyperplan perpendiculaire à $\mathbf{v}$, et passant par $\lambda \frac{\mathbf{v}}{\|\mathbf{v}\|^2}$.

Les surfaces de niveaux. Soit $g : \mathbb{R}^d \to \mathbb{R}$ une fonction à valeurs réelles, et soit $P := \left\{ \mathbf{x} \in \mathbb{R}^d, \ \nabla g(\mathbf{x}) = \mathbf{0} \right\}$ l'ensemble de ses points critiques.

Alors, les seuls ensembles de niveau de g qui ne sont pas des surfaces sont les ensembles qui passent par un de ces points critiques :

$$\forall \lambda \in \mathbb{R}, \quad \text{si} \quad S_\lambda \cap P = \emptyset, \quad \text{alors} \quad S_\lambda \text{ est une surface.}$$

[IMAGE TODO]

5.3.4 Plan tangent à une surface.

Soit $S \subset \mathbb{R}^d$ une surface de dimension n. Soit $\gamma : \mathbb{R} \to \mathbb{R}^d$ (= courbe paramétrée). On dit que γ est un chemin tracé sur S si, pour tout $t \in \mathbb{R}$, on a $\gamma(t) \in S$.

Définition 5.15 : Espace tangent, plan tangent.

Soit $\mathbf{a} \in S$, l'**espace tangent** à S au point $\mathbf{a}$, noté $T_{\mathbf{a}}S$, est l'ensemble des vecteurs $\gamma'(0)$, où $\gamma : \mathbb{R} \to \mathbb{R}^d$ est un chemin tracé sur S avec $\gamma(0) = \mathbf{a}$.
Le **plan affine tangent** à la surface S en $\mathbf{a}$ est $\mathbf{a} + T_{\mathbf{a}}S$.

[IMAGE TODO]

Théorème 5.16

Si $S \subset \mathbb{R}^d$ est une surface de dimension n, alors, pour tout $\mathbf{a} \in S$, $T_{\mathbf{a}}S$ est un hyperplan affine de dimension d.

Démonstration. Soit $\alpha : (\mathbb{R}^d, \mathbf{a}) \to (\mathbb{R}^d, \mathbf{0})$ le difféomorphisme qui prouve que S est une surface. Si $\gamma : \mathbb{R} \to \mathbb{R}^d$ est tracée sur S, on a $\alpha(\gamma(t)) \in \mathbb{R}^n \times \{\mathbf{0}\}$ pour tout t. En différentiant, on trouve que

$$\forall t \in \mathbb{R}, \quad \mathrm{d}\alpha_{\gamma(t)} \cdot \gamma'(t) \in \mathbb{R}^n \times \{\mathbf{0}\}.$$

En particulier, si $\gamma(0) = \mathbf{a}$, on a $\mathrm{d}\alpha_{\mathbf{a}} \cdot \gamma'(0) \in \mathbb{R}^n \times \{\mathbf{0}\}$. Comme $\mathrm{d}\alpha_{\mathbf{a}}$ est inversible (car α est un difféomorphisme), on en déduit que $\gamma'(0) \in (\mathrm{d}\alpha_{\mathbf{a}})^{-1}(\mathbb{R}^n \times \mathbf{0})$, qui est de dimension n. En fait, on a *****TODO*****

$$T_{\mathbf{a}}S := \mathbf{a} + (\mathrm{d}\alpha_{\mathbf{a}})^{-1}(\mathbb{R}^n \times \mathbf{0}),$$

qui est de dimension n. □

Dans le cas des submersions, il est simple de caractériser ce plan tangent.

Théorème 5.17 : Caractérisation du plan tangent

Soit $g : \mathbb{R}^d \to \mathbb{R}^p$ de classe C^1, et soit $S := g^{-1}(\boldsymbol{\lambda})$ une surface où g est une submersion. Alors

$$\forall \mathbf{a} \in S, \quad T_{\mathbf{a}}S = \operatorname{Ker} \mathrm{d}g_{\mathbf{a}}.$$

Démonstration. Si $\gamma : \mathbb{R} \to \mathbb{R}^d$ est tracé sur S, on a $g(\gamma(t)) = \boldsymbol{\lambda}$ pour tout $t \in \mathbb{R}$. En différenciant, on trouve $\mathrm{d}g_{\gamma(t)}\gamma'(t) = \mathbf{0}$. Si $\gamma(0) = \mathbf{a}$, on obtient $\mathrm{d}g_{\mathbf{a}} \cdot \gamma'(0) = \mathbf{0}$, ce qui prouve que $T_{\mathbf{a}}S \subset \operatorname{Ker} \mathrm{d}g_{\mathbf{a}}$.
Par ailleurs, comme $\mathrm{d}g_{\mathbf{a}}$ est surjectif, le théorème du rang montre que

$$\dim \operatorname{Ker} \mathrm{d}g_{\mathbf{a}} = d - \operatorname{rg}(\mathrm{d}g_{\mathbf{a}}) = d - p.$$

Par ailleurs, on sait que $\dim T_{\mathbf{a}}S = d - p$ (c'est la dimension de S). Donc, comme les dimensions sont égales et que $T_{\mathbf{a}}S \subset \operatorname{Ker} \mathrm{d}g_{\mathbf{a}}$, on a égalité $T_{\mathbf{a}}S = \operatorname{Ker} \mathrm{d}g_{\mathbf{a}}$. □

On montre facilement qu'un hyperplan coïncide avec son plan tangent affine en tout point.

Dans le cas important où $p = 1$, on en déduit que

$$T_{\mathbf{a}}S = \left\{\mathbf{x} \in \mathbb{R}^d, \quad \langle \nabla g(\mathbf{a}), \mathbf{x} \rangle = 0\right\} = (\nabla g(\mathbf{a}))^{\perp}.$$

Ainsi, *le gradient est toujours orthogonal aux courbes de niveaux*, dans le sens où il est orthogonal aux plans tangents aux courbes de niveaux.

CHAPITRE 6

OPTIMISATION SOUS CONTRAINTE ÉGALITÉ

Dans ce chapitre, nous nous intéressons à l'optimisation sous contraintes, de la forme

$$\min\left\{f(\mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^d, \quad g(\mathbf{x}) = \mathbf{0}\right\},$$

où f est une fonction de $\mathbb{R}^d$ dans $\mathbb{R}$ (toujours à valeurs réelles, sinon on ne peut pas optimiser), et g est une fonction de $\mathbb{R}^d$ dans $\mathbb{R}^p$. Le nombre p s'appelle le **nombre de contraintes**. En fait, en écrivant $g = (g_1, \cdots, g_p)^T$ les composantes de g, où chaque g_j est une fonction de $\mathbb{R}^d$ dans $\mathbb{R}$, alors on a

$$g(\mathbf{x}) = \mathbf{0} \quad \text{ssi} \quad \begin{cases} g_1(\mathbf{x}) = 0 \\ \vdots \\ g_p(\mathbf{x}) = 0 \end{cases}.$$

La condition $g(\mathbf{x}) = \mathbf{0}$ est donc la réunion des p contraintes scalaires $g_j(\mathbf{x}) = 0$. En fait, grâce au chapitre précédent, on sait que la contrainte $g(\mathbf{x})$ est une manière d'écrire $\mathbf{x} \in S \subset \mathbb{R}^d$, où S est une surface de dimension $n = d - p$ dans $\mathbb{R}^d$. C'est la bonne façon d'écrire les choses : on demandera dans le suite à ce que g soit une submersion.

Ainsi, notre but est de déterminer le minimum de f sur la surface S. Autrement dit, on cherche

$$\min f|_S := \min\{f(\mathbf{x}), \ \mathbf{x} \in S\}.$$

6.1 Extrema liés, Euler–Lagrange

Le résultat principal de ce chapitre est donné par le théorème suivant.

Théorème 6.1 : Extrema liés

Soit $f : \mathbb{R}^d \to \mathbb{R}$ une fonction de classe C^1, et soit S une surface d'équation $g = \mathbf{0}$, où $g = (g_1, \cdots, g_p)$ est une submersion de $\mathbb{R}^d$ dans $\mathbb{R}^p$. Alors si $\mathbf{x}_* \in S$ est un minimum de f sur S, il existe $\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_p)^T \in \mathbb{R}^p$ tel que $\nabla f(\mathbf{x}_*) = (J_g(\mathbf{x}_*))^T \boldsymbol{\lambda}$, qui s'écrit aussi

$$\nabla f(\mathbf{x}_*) = \lambda_1 \nabla g_1(\mathbf{x}_*) + \lambda_2 \nabla g_2(\mathbf{x}_*) + \cdots + \lambda_p \nabla g_p(\mathbf{x}_*).$$

Démonstration. Soit $\gamma : \mathbb{R} \to S$ un chemin tracé sur S, passant par $\mathbf{x}_*$ en $t = 0$. Comme $\mathbf{x}_*$ est le minimum de f sur S, la fonction $f(\gamma(t))$ atteint son minimum en $t = 0$. En dérivant par rapport à t, et en utilisant la règle de la chaîne, on obtient

$$\langle \nabla f(\mathbf{x}_*), \gamma'(0) \rangle = 0.$$

Ceci étant vrai pour tout chemin tracé sur S avec $\gamma(0) = \mathbf{x}_*$, on en déduit que

$$\nabla f(\mathbf{x}_*) \in (T_{\mathbf{x}_*}S)^\perp .$$

Nous avons montré au Théorème 5.17 que $T_{\mathbf{x}_*}S = \operatorname{Ker} \mathrm{d}g_{\mathbf{x}_*} = \operatorname{Ker} J_g(\mathbf{x}_*)$. Nous rappelons le résultat (classique) suivant (nous en donnons la démonstration à la fin du paragraphe).

Lemme 6.2. *Pour toute matrice $A \in \mathcal{M}_{p\times d}$, on a* $(\operatorname{Ker} A)^\perp = \operatorname{Im} A^T$.

Ainsi, on a $\nabla f(\mathbf{x}_*) \in \operatorname{Im}(J_g(\mathbf{x}_*))^T$, donc $\nabla f(\mathbf{x}_*)$ est de la forme $\nabla f(\mathbf{x}_*) = (J_g(\mathbf{x}_*))^T \boldsymbol{\lambda}$ pour un certain $\boldsymbol{\lambda} \in \mathbb{R}^p$. En remarquant que les colonnes de $(J_g(\mathbf{x}_*))^T$ sont exactement $(\nabla g_1, \cdots, \nabla g_p)$, on obtient bien

$$\nabla f(\mathbf{x}_*) = \lambda_1 \nabla g_1(\mathbf{x}_*) + \lambda_2 \nabla g_2(\mathbf{x}_*) + \cdots + \lambda_p \nabla g_p(\mathbf{x}_*).$$

□

Preuve du Lemme 6.2. Soit $A \in \mathcal{M}_{p\times d}$, de sorte que $A^T \in \mathcal{M}_{d\times p}$. Soit $\mathbf{x} \in \operatorname{Im} A^T \subset \mathbb{R}^d$ de la forme $\mathbf{x} = A^T\mathbf{v}$ pour un certain $\mathbf{v} \in \mathbb{R}^p$. Alors, pour tout $\mathbf{y} \in \operatorname{Ker} A \subset \mathbb{R}^d$, on a

$$\langle \mathbf{y}, \mathbf{x}\rangle_{\mathbb{R}^d} = \langle \mathbf{y}, A^T\mathbf{v}\rangle_{\mathbb{R}^d} = \langle A\mathbf{y}, \mathbf{v}\rangle_{\mathbb{R}^p} = 0.$$

Ainsi, tout $\mathbf{x} \in \operatorname{Im} A^T$ est orthogonal à $\operatorname{Ker} A$, donc $\operatorname{Im} A^T \subset (\operatorname{Ker} A)^\perp$. Par ailleurs, le théorème du rang affirme que

$$\dim \operatorname{Im} A + \dim \operatorname{Ker} A = d.$$

De plus, on a $\dim \operatorname{Im} A = \dim \operatorname{Im} A^T$ (les rangs de A et A^T sont égaux). Ceci montre que $\dim \operatorname{Im} A^T = d - \dim \operatorname{Ker} A = \dim(\operatorname{Ker} A)^\perp$. Ainsi, on a égalité $\operatorname{Im} A^T = (\operatorname{Ker} A)^\perp$ comme souhaité. □

Remarque 6.3. *Le Lemme 6.2 montre en particulier que A est surjective ssi A^T est injective, et que A est injective ssi A^T est surjective.*

Les nombres $(\lambda_1, \cdots, \lambda_p)$ apparaissant dans le théorème sont appelés les **multiplicateurs de Lagrange**. Il y a un multiplicateur par contrainte (scalaire). L'équation

$$\nabla f(\mathbf{x}_*) = \lambda_1 \nabla g_1(\mathbf{x}_*) + \lambda_2 \nabla g_2(\mathbf{x}_*) + \cdots + \lambda_p \nabla g_p(\mathbf{x}_*) \tag{6.1}$$

s'appelle parfois **équation d'Euler–Lagrange**, ou **équation des extrema liés**. Elle exprime le fait que le gradient de f est une combinaison linéaire des gradients des contraintes.

Un point $(\mathbf{x}_*, \boldsymbol{\lambda}) \in \mathbb{R}^d \times \mathbb{R}^p$ vérifiant

$$\begin{cases} \nabla f(\mathbf{x}_*) & = \lambda_1 \nabla g_1(\mathbf{x}_*) + \lambda_2 \nabla g_2(\mathbf{x}_*) + \cdots + \lambda_p \nabla g_p(\mathbf{x}_*) \\ g(\mathbf{x}_*) & = \mathbf{0} \end{cases} \tag{6.2}$$

s'appelle un **point critique** pour le problème de minimisation sous contrainte. Tout comme l'optimisation classique, on va chercher les minima de f sur S parmi les points critiques. On remarquera qu'il s'agit d'un système avec $d + p$ équations et $d + p$ inconnues (car il faut trouver $\mathbf{x}_* \in \mathbb{R}^d$ et $\boldsymbol{\lambda} \in \mathbb{R}^p$).

6.1.1 Intuition géométrique

D'après l'équation d'Euler-Lagrange, si $(\mathbf{x}_*, \boldsymbol{\lambda}_*)$ est un point critique sous contrainte, le gradient de f en $\mathbf{x}_*$ est orthogonal au plan tangent à la surface définie par la contrainte. Donc la courbe de niveau $f(\mathbf{x}_*)$ de f (c'est-à-dire la courbe de niveau de f qui contient le point $\mathbf{x}_*$)[1] est tangente à la surface $\mathcal{S}$ au point $\mathbf{x}_*$: leurs hyperplans tangents coincident au point $\mathbf{x}_*$.

TODO

---

[1] Cette courbe de niveau est unique sinon g ne serait pas une submersion au voisinage de $\mathbf{x}_*$. En d'autres termes, le théorème des fonctions implicites ou, ce qui est équivalent, le théorème d'inversion locale, ne s'appliquerait pas au point $\mathbf{x}_*$.

6.1.2 Existence des minimiseurs

Concernant l'existence d'optimiseurs de problèmes sous contraintes, elle est souvent assez simple. On a par exemple le résultat suivant.

Théorème 6.4

Soit $f : \mathbb{R}^d \to \mathbb{R}$ une fonction continue. Si S est une surface **fermée** et **bornée**, alors f atteint son minimum sur S.

C'est une simple application du théorème de Weierstrass **??**. Si S est d'équation $g = \mathbf{0}$ avec g continue, alors $S = g^{-1}(\{\mathbf{0}\})$ est automatiquement fermée (c'est l'image réciproque du fermé $\{\mathbf{0}\}$ par l'application continue g). De plus, il est souvent simple de vérifier que S est bornée. Cela est le cas par exemple dès que g est coercive.

6.2 Le Lagrangien

Afin de se ramener à un problème de mininimisation sans contrainte, il est commode d'introduire la fonction suivante de $\mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}$, appelée **Lagrangien** associé au problème de minimisation, et définie par

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) - \langle \boldsymbol{\lambda}, g(\mathbf{x}) \rangle = f(\mathbf{x}) - \sum_{i=1}^{p} \lambda_i \, g_i(\mathbf{x}).$$

Le Lagrangien possède la même régularité que les fonctions f et g par rapport à la variable $\mathbf{x} \in \mathbb{R}^d$, et c'est une application linéaire, donc indéfiniment différentiable, par rapport à la variable $\boldsymbol{\lambda} \in \mathbb{R}^p$. En particulier, $\mathcal{L}$ est bien une application de classe $\mathcal{C}^1$ sur $\mathbb{R}^d \times \mathbb{R}^p$.

On observe que tout point critique sous contrainte $(\mathbf{x}_*, \boldsymbol{\lambda}_*)$, c'est-à-dire toute solution $(\mathbf{x}_*, \boldsymbol{\lambda}_*)$ du système (6.2), est un point critique du Lagrangien, et réciproquement. En effet, on a

$$\begin{cases} \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) & = \nabla f(\mathbf{x}) - J_g(\mathbf{x})^T \boldsymbol{\lambda}; \\ \nabla_{\boldsymbol{\lambda}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) & = g(\mathbf{x}). \end{cases}$$

Ainsi, l'équation $\nabla_{\mathbf{x}} \mathcal{L} = \mathbf{0}$ est équivalente aux équations d'Euler–Lagrange, et l'équation $\nabla_{\boldsymbol{\lambda}} \mathcal{L} = \mathbf{0}$ exprime la contrainte $g = \mathbf{0}$.

L'avantage du Lagrangien est qu'il donne des conditions simples pour déterminer si un point critique (au sens des problèmes d'optimisation sous contraintes) est un minimum.

6.2.1 Caractérisation d'un minimum avec le Lagrangien

Dans la suite, on s'intéresse à un point critique $(\mathbf{x}_*, \boldsymbol{\lambda}_*)$ du Lagrangien $\mathcal{L}$. En particulier, on a $\mathbf{x}_* \in S$. L'idée principale de cette section est qu'on peut déterminer si $\mathbf{x}_*$ est un minimum local de f sur S en étudiant l'application partielle

$$\ell_{\boldsymbol{\lambda}_*}(\mathbf{x}) := \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}_*).$$

C'est à dire qu'on fixe le multiplicateur de Lagrange $\boldsymbol{\lambda}_*$, et qu'on regarde la fonction $\mathbf{x} \mapsto \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}_*)$.

Théorème 6.5 : Caractérisation d'un minimum avec le Lagrangien

Soit $f : \mathbb{R}^d \to \mathbb{R}$ une fonction de classe C^1, et soit S une surface d'équation $g = \mathbf{0}$, où g est une submersion de $\mathbb{R}^d$ dans $\mathbb{R}^p$. Soit $\mathcal{L} : \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}$ le Lagrangien associé et soit $(\mathbf{x}_*, \boldsymbol{\lambda}_*)$ un point critique de $\mathcal{L}$. On pose $\ell_{\boldsymbol{\lambda}_*}(\mathbf{x}) := \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}_*)$.
Si $\mathbf{x}_*$ est un minimum local de $\ell_{\boldsymbol{\lambda}_*}$, alors $\mathbf{x}_*$ est un minimum local de f sur S.
Si $\mathbf{x}_*$ est un minimum global de $\ell_{\boldsymbol{\lambda}_*}$, alors $\mathbf{x}_*$ est un minimum global de f sur S.

Attention, la réciproque est fausse en général : il existe des cas où $\mathbf{x}_*$ est un minimum de f, mais n'est pas un minimum de $\ell_{\boldsymbol{\lambda}_*}$comme l'illustre l'exemple ci-dessous.

Exemple 6.6. *On considère la fonction f définie sur $\mathbb{R}^2$ par $f(x,y) = -xy$ que l'on souhaite minimiser sous la contrainte $x + y = 2$.*

En explicitant la contrainte, on est ramené à optimiser la fonction $\phi(x) = x(x-2)$ sur $\mathbb{R}$, qui atteint un minimum global en $x = 1$ de valeur -1. Donc le point $(1,1)$ est un minimum global de f sous la contrainte.

En appliquant la méthode du Lagrangien, on vérifie facilement que $(1,1)$ avec $\lambda = -1$ est la seule solution des équations d'Euler–Lagrange. Or, le Lagrangien vaut $\mathscr{L}(x,y) = xy+(x+y-2)$, qui n'est ni localement convexe ni localement concave en $(1,1)$ concave, et qui n'est ni majorée ni minorée sur $\mathbb{R}^2$. On peut démontrer en revanche que $(1,1)$ est un maximum local de f par la méthode du Lagrangien en linéarisant la contrainte (voir le théorème 6.9 ci-dessous).

Démonstration. Soit $(\mathbf{x}_*, \boldsymbol{\lambda}_*)$ un point critique de $\mathcal{L}$ tel que $\mathbf{x}_*$ soit un minimum local de $\ell_{\boldsymbol{\lambda}_*}$. Alors, pour tout $\mathbf{x} \in S$ proche de $\mathbf{x}_*$, on a (on utilise que $g(\mathbf{x}) = 0$, car $\mathbf{x} \in S$, et que $g(\mathbf{x}_*) = \mathbf{0}$, car $\nabla_{\boldsymbol{\lambda}} \mathcal{L}(\mathbf{x}_*, \boldsymbol{\lambda}_*) = 0$)

$$f(\mathbf{x}) = f(\mathbf{x}) - \boldsymbol{\lambda}_* g(\mathbf{x}) = \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}_*) \geq \mathcal{L}(\mathbf{x}_*, \boldsymbol{\lambda}_*) = f(\mathbf{x}_*) - \boldsymbol{\lambda}_* g(\mathbf{x}_*) = f(\mathbf{x}_*).$$

Ainsi, $\mathbf{x}_*$ est un minimum local de $f|_S$, comme voulu. La preuve dans le cas global est similaire. ☐

On donne deux applications classiques du théorème précédent.

Théorème 6.7 : Le cas où $\mathcal{L}$ est convexe

Avec les mêmes notations que le Théorème 6.5, si $\ell_{\boldsymbol{\lambda}_*}$ est convexe, alors $\mathbf{x}_*$ est un minimum global de f sur S.

Démonstration. Comme l'application $\ell_{\boldsymbol{\lambda}_*}$ est convexe, et que $\mathbf{x}_*$ est un point critique de cette application (car $\nabla \ell_{\boldsymbol{\lambda}_*}(\mathbf{x}_*) = \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}_*, \boldsymbol{\lambda}_*) = \mathbf{0}$), $\mathbf{x}_*$ est l'unique minimiseur de $\ell_{\boldsymbol{\lambda}_*}$. Le résultat découle alors du Théorème 6.5. ☐

Théorème 6.8 : Condition à l'ordre 2

Avec les mêmes notations que le Théorème 6.5, et en supposant que f et g sont de classe $\mathcal{C}^2$, si la Hessienne $H := H_{\ell_{\boldsymbol{\lambda}_*}}(\mathbf{x}_*)$ est définie positive, alors $\mathbf{x}_*$ est un minimum local de f sur S.

La Hessienne est selon la variable $\mathbf{x}$ uniquement.

6.2.2 Raffinement de la condition d'ordre 2

La condition d'ordre 2 énoncée dans le Théorème 6.8 peut être améliorée en faisant une étude locale plus précise au voisinage d'un point critique sous contrainte $\mathbf{x}_*$. En effet, il suffit d'étudier le signe des accroissements $f(\mathbf{x}_* + \mathbf{h}) - f(\mathbf{x}_*)$ pour $\mathbf{h}$ suffisamment petit pour que $\mathbf{x}_*$ et $\mathbf{x}_* + \mathbf{h}$ appartiennent à S, ce qui revient à restreindre $\mathbf{h}$ à l'espace tangent à la surface S au point $\mathbf{x}_*$.

Théorème 6.9 : Linéarisation de la contrainte

Avec les mêmes notations que le Théorème 6.5, et en supposant que f et g sont de classe $\mathcal{C}^2$, si la forme quadratique $\mathrm{d}^2\ell_{\boldsymbol{\lambda}_*}(\mathbf{h}, \mathbf{h})$ est strictement positive pour tout $\mathbf{h} \in T_{\mathbf{x}_*}S \setminus \{\mathbf{0}\}$, alors $\mathbf{x}_*$ est un minimum local de f sur S.

Remarque 6.10 (Linéarisation de la contrainte)**.** *Cette propriété est particulièrement utile lorsque* $\mathbf{x}_*$ *est un point selle du Lagrangien* ℓ_{λ_*}*. Il suffit alors d'étudier le signe de la forme quadratique associée à la matrice hessienne de* ℓ_{λ_*} *sur l'espace tangent à la surface* S *au point* $\mathbf{x}_*$*, c'est-à-dire sur le sous-espace vectoriel* $(T_{\mathbf{x}_*}S)^\perp$*.*

Démonstration. Nous reprenons la preuve du Théorème 6.1 qui établit les conditions nécessaires pour que $\mathbf{x}_* \in S$ soit un minimum local de f sur S, en supposant cette fois que $\gamma : \mathbb{R} \to S$ est un chemin de classe $\mathcal{C}^2$ tracé sur S et passant par $\mathbf{x}_*$ en $t = 0$. En particulier, $g(\gamma(t)) = 0$ pour tout t au voisinage de 0, ce qui impose $\gamma'(0) \in T_{\mathbf{x}_*}S$. En dérivant successivement la fonction $g \circ \gamma$ par rapport à t, on obtient d'abord

$$J_g(\gamma(t))\,\gamma'(t) = 0,$$

puis

$$J_g(\gamma(t))\gamma''(t) + \mathrm{d}^2 g_{\gamma(t)}\big(\gamma'(t), \gamma'(t)\big) = 0,$$

en appliquant la règle de la chaîne. En particulier, en $t = 0$,

$$J_g(\mathbf{x}_*)\,\gamma'(0) = 0 \quad \text{ et } J_g(\mathbf{x}_*)\gamma''(0) + \mathrm{d}^2 g_{\mathbf{x}_*}\big(\gamma'(0), \gamma'(0)\big) = 0. \tag{6.3}$$

On définit la fonction $\varphi := f \circ \gamma$ qui est de classe C^2 au voisinage de 0. Comme $\mathbf{x}_* = \gamma(0)$ est un minimum local de f sur S, la fonction φ atteint son minimum en $t = 0$. En particulier, on en déduit, comme dans la preuve du Théorème 6.1, que

$$\varphi'(0) = \langle \nabla f(\mathbf{x}_*), \gamma'(0) \rangle = 0, \tag{6.4}$$

et que $\nabla f(\mathbf{x}_*)$ est de la forme $\nabla f(\mathbf{x}_*) = (J_g(\mathbf{x}_*))^T\, \boldsymbol{\lambda}_*$ pour un certain $\boldsymbol{\lambda}_* \in \mathbb{R}^p$. En reportant cette dernière expression dans (6.3), on trouve en particulier que

$$\langle \nabla f(\mathbf{x}_*), \gamma''(0) \rangle = \nabla f(\mathbf{x}_*)^T \gamma''(0) = \boldsymbol{\lambda}_*^T J_g(\mathbf{x}_*)\gamma''(0) = -\boldsymbol{\lambda}_*^T \mathrm{d}^2 g_{\mathbf{x}_*}\big(\gamma'(0), \gamma'(0)\big). \tag{6.5}$$

Pour que φ atteigne un minimum local en 0, il suffit que $\varphi''(0) > 0$. Or, en appliquant deux fois la règle de la chaîne, on obtient

$$\varphi'(t) = \big\langle \nabla f(\gamma(t)), \gamma'(t) \big\rangle,$$

puis

$$\varphi''(t) = \big\langle \nabla f(\gamma(t)), \gamma''(t) \big\rangle + \big\langle \gamma'(t), H_f(\gamma(t))\,\gamma'(t) \big\rangle.$$

Donc, en utilisant (6.5),

$$\begin{aligned}
\varphi''(0) &= \big\langle \nabla f(\mathbf{x}_*), \gamma''(0) \big\rangle + \big\langle \gamma'(0), H_f(\mathbf{x}_*)\gamma'(0) \big\rangle \\
&= \big\langle \gamma'(0), H_f(\mathbf{x}_*)\gamma'(0) \big\rangle - \lambda_*^T \mathrm{d}^2 g_{\mathbf{x}_*}\big(\gamma'(0), \gamma'(0)\big) \\
&= \mathrm{d}^2 \ell_{\boldsymbol{\lambda}_*}\big(\gamma'(0), \gamma'(0)\big),
\end{aligned} \tag{6.6}$$

en utilisant la définition du Lagrangien.

Ceci étant vrai pour tout chemin tracé sur S avec $\gamma(0) = \mathbf{x}_*$, on en déduit la condition suffisante énoncée dans le théorème puisque $\gamma'(0)$ décrit le sous-espace vectoriel $T_{\mathbf{x}_*}S \setminus \{\mathbf{0}\}$. □

Reprenons l'exemple 6.6 en appliquant le théorème 6.9. La linéarisation de la contrainte s'écrit $h + k = 0$. Or, la forme quadratique associée à la matrice hessienne de $\mathscr{L}$ vaut $q_{\mathscr{L}}(h, k) = -hk$. En particulier $q_{\mathscr{L}}(h, -h) = h^2 > 0$ si $h \neq 0$. Le point $(1, 1)$ est donc un minimum local de $f(x, y) = xy$ sous la contrainte $x + y = 2$.

6.3 Exemples

6.3.1 Minimisation quadratique sur un espace affine

Soit P un espace affine d'équation $E := \left\{ \mathbf{x} \in \mathbb{R}^d, \quad C\mathbf{x} = \boldsymbol{\nu} \right\}$, avec $C \in \mathcal{M}_{p \times d}$ une matrice surjective, et $\boldsymbol{\nu} \in \mathbb{R}^p$ (donc E est un espace affine de dimension $n = d - p$ dans $\mathbb{R}^d$, cf Section **??**). En particulier, pour $p = 1$, E est un hyperplan. Pour $p = d - 1$, E est une droite (que l'on écrit ainsi comme l'intersection de $d - 1$ hyperplans).

On s'intéresse au problème de minimisation

$$\min \left\{ \frac{1}{2}\langle \mathbf{x}, A\mathbf{x} \rangle - \langle \mathbf{b}, \mathbf{x} \rangle, \quad \mathbf{x} \in E \right\},$$

où $A \in \mathcal{S}_d^{++}$ est une symétrique définie positive, et $\mathbf{b} \in \mathbb{R}^d$. On a donc posé $f(\mathbf{x}) := \frac{1}{2}\langle \mathbf{x}, A\mathbf{x} \rangle - \langle \mathbf{b}, \mathbf{x} \rangle$, et $g(\mathbf{x}) := C\mathbf{x} - \boldsymbol{\nu}$.

Les équations d'Euler–Lagrange s'écrivent

$$\begin{cases} A\mathbf{x} - \mathbf{b} = C^T \boldsymbol{\lambda}, \\ C\mathbf{x} = \boldsymbol{\nu} \end{cases}, \quad \text{ou encore} \quad \begin{pmatrix} A & -C^T \\ C & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{x} \\ \boldsymbol{\lambda} \end{pmatrix} = \begin{pmatrix} \mathbf{b} \\ \boldsymbol{\nu} \end{pmatrix} \qquad \in \mathbb{R}^d \times \mathbb{R}^p.$$

On note $M := \begin{pmatrix} A & -C^T \\ C & \mathbf{0} \end{pmatrix}$ la matrice par blocs apparaissant. On prétend que la matrice M est inversible (cf ci-dessous). Ainsi, il y a un seul critique, donné par

$$\begin{pmatrix} \mathbf{x}_* \\ \boldsymbol{\lambda}_* \end{pmatrix} := M^{-1} \begin{pmatrix} \mathbf{b} \\ \boldsymbol{\nu} \end{pmatrix}.$$

Pour montrer que ce point critique est le minimiseur, on peut remarquer que le Lagrangien est de la forme

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) = \frac{1}{2}\langle \mathbf{x}, A\mathbf{x} \rangle - \langle \mathbf{b}, \mathbf{x} \rangle - \langle \boldsymbol{\lambda}, C\mathbf{x} - \boldsymbol{\nu} \rangle.$$

En particulier, pour tout $\boldsymbol{\lambda}$ fixé (et en particulier pour $\boldsymbol{\lambda} = \boldsymbol{\lambda}_*$, l'application $\mathbf{x} \mapsto \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda})$ est convexe (car A est définie positive, et les autres termes sont linéaires en $\mathbf{x}$). Ainsi, $\mathbf{x}_*$ est le minimum global de f sur S.

Montrons que M est inversible. Pour cela, on montre que M est injective. Si $M(\mathbf{x}, \boldsymbol{\lambda})^T = \mathbf{0}$, on a $A\mathbf{x} = C^T \boldsymbol{\lambda}$ et $C\mathbf{x} = \mathbf{0}$. En prenant le produit scalaire de la première équation avec $\mathbf{x}$, on obtient

$$\langle \mathbf{x}, A\mathbf{x} \rangle = \langle \mathbf{x}, C^T \boldsymbol{\lambda} \rangle = \langle C\mathbf{x}, \boldsymbol{\lambda} \rangle = \mathbf{0}.$$

Comme A est définie positive, cela implique $\mathbf{x} = \mathbf{0}$. La première équation donne $C^T \boldsymbol{\lambda} = \mathbf{0}$. Or C est surjective, donc C^T est injective (cf Remarque 6.3), et on en déduit que $\boldsymbol{\lambda} = \mathbf{0}$.

Cas particulier : projection sur un espace affine. Soit $\mathbf{y} \in \mathbb{R}^d$. La projection de $\mathbf{y}$ sur E, est par définition le point de E le plus proche de $\mathbf{y}$ (pour la norme Euclidienne). On cherche donc à minimiser sur E la fonction

$$f_1(\mathbf{x}) := \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|^2 = \frac{1}{2}\|\mathbf{x}\|^2 - \langle \mathbf{y}, \mathbf{x} \rangle + \frac{1}{2}\|\mathbf{y}\|^2.$$

En remarquant que le terme $\frac{1}{2}\|\mathbf{y}\|^2$ est indépendant de $\mathbf{x}$ (et donc ne contribue pas à la minimisation), on en déduit qu'il suffit de minimiser la fonction $f(\mathbf{x}) := f_1(\mathbf{x}) - \frac{1}{2}\|\mathbf{y}\|^2$. Cette fonction est de la forme $\langle \mathbf{x}, A\mathbf{x} \rangle - \langle \mathbf{b}, \mathbf{x} \rangle$ avec $A = \mathbb{I}_d$ et $\mathbf{b} = \mathbf{y}$.

Ainsi, on peut calculer une projection sur un espace affine en inversant la matrice $M = \begin{pmatrix} \mathbb{I}_d & -C^T \\ C & \mathbf{0} \end{pmatrix}$ (ce qui est facile à faire numériquement par exemple).

Si E est un hyperplan, on peut choisir $C = \mathbf{v}^T$ avec $\mathbf{v}$ un vecteur unitaire orthogonal à E. Comme $p = 1$, $\lambda \in \mathbb{R}$. Les équations d'Euler–Lagrange s'écrivent alors

$$\begin{cases} \mathbf{x} - \mathbf{y} = \lambda\, \mathbf{v}, \\ \mathbf{v}^T \mathbf{x} = \nu. \end{cases}$$

La première équation montre que $\mathbf{x} - \mathbf{y}$ est colinéaire à $\mathbf{v}$ donc orthogonal à E. Le système se résout explicitement : en prenant le produit scalaire par $\mathbf{v}$ de la première équation et en utilisant la deuxième équation, on obtient que $\lambda = \nu - \mathbf{v}^T\mathbf{y}$, puis, en reportant dans la première équation, $\mathbf{x} = \mathbf{y} + (\nu - \mathbf{v}^T\mathbf{y})\, \mathbf{v}$. En particulier, si $y \in E$, on retrouve bien le fait que $\mathbf{x} = \mathbf{y}$ puisque $\lambda = 0$. La distance de $\mathbf{y}$ à E vaut $\|\mathbf{x} - \mathbf{y}\|$, et on trouve l'expression explicite $d(\mathbf{y}; E) = \|\mathbf{x} - \mathbf{y}\| = |\nu - \mathbf{v}^T\mathbf{y}|$.

6.3.2 Projection sur une sphère

On considère $\mathbb{S}^{d-1} := \left\{\mathbf{x} \in \mathbb{R}^d,\ \|\mathbf{x}\| = 1\right\}$, et, pour $\mathbf{y} \in \mathbb{R}^d$ fixé, on regarde le problème de minimisation

$$\min \left\{\|\mathbf{x} - \mathbf{y}\|, \quad \mathbf{x} \in \mathbb{S}^{d-1}\right\}.$$

Autrement dit, on cherche le point de $\mathbb{S}^{d-1}$ le plus proche de $\mathbf{y}$. Pour commencer, comme $\mathbb{S}^{d-1}$ est fermé et borné, ce minimum existe. Le même raisonnement montre que le maximum existe aussi.

On calcule les points critiques, avec $f(\mathbf{x}) := \|\mathbf{x} - \mathbf{y}\|^2$ et $g(\mathbf{x}) := \|\mathbf{x}\|^2 - 1$. On cherche $(\mathbf{x}, \lambda) \in \mathbb{R}^d \times \mathbb{R}$ solution de

$$\begin{cases} \mathbf{x} - \mathbf{y} & = \lambda\mathbf{x} \\ \|\mathbf{x}\|^2 & = 1 \end{cases} \iff \begin{cases} (1 - \lambda)\mathbf{x} & = \mathbf{y} \\ \|\mathbf{x}\|^2 & = 1 \end{cases}.$$

Si $\mathbf{y} = \mathbf{0}$, on doit avoir $\lambda = 1$, et dans ce cas, tous les points $(\mathbf{x}, 0)$ sont des points critiques. Ils sont tous les minima, car $\mathbf{y} = \mathbf{0}$ est à distance 1 de tous les points de la sphère $\mathbb{S}^{d-1}$.

Si $\mathbf{y} \neq \mathbf{0}$, on a $\lambda \neq 1$, et la première équation montre que $\mathbf{x}$ est colinéaire à $\mathbf{y}$, et de norme 1. On a donc deux points critiques pour f, à savoir

$$\mathbf{x}_+ := \frac{\mathbf{y}}{\|\mathbf{y}\|}, \quad \text{et} \quad \mathbf{x}_- := -\frac{\mathbf{y}}{\|\mathbf{y}\|}.$$

Il est facile de vérifier que $\|\mathbf{y} - \mathbf{x}_+\| < \|\mathbf{y} - \mathbf{x}_-\|$. On en déduit que $\mathbf{x}_+$ est le minimum global, et $\mathbf{x}_-$ est le maximum global.

6.3.3 Plus grande et de la plus petite valeur propre d'une matrice symétrique

Soit $A \in \mathcal{S}_d$ une matrice symétrique et soient $\lambda_1 \leq \cdots \leq \lambda_d$ ses d valeurs propres réelles ordonnées par ordre croissant. Nous allons montrer que

$$\lambda_1 = \min_{1 \leq i \leq d} \lambda_i = \inf \left\{\frac{\langle \mathbf{x}, A\mathbf{x}\rangle}{\|\mathbf{x}\|^2}, \quad \mathbf{x} \in \mathbb{R}^d \setminus \{\mathbf{0}\}\right\} = \inf \left\{\langle \mathbf{x}, A\mathbf{x}\rangle, \quad \mathbf{x} \in \mathbb{S}^{d-1}\right\}$$

et

$$\lambda_d = \max_{1 \leq i \leq d} \lambda_i = \sup \left\{\frac{\langle \mathbf{x}, A\mathbf{x}\rangle}{\|\mathbf{x}\|^2}, \quad \mathbf{x} \in \mathbb{R}^d \setminus \{\mathbf{0}\}\right\} = \sup \left\{\langle \mathbf{x}, A\mathbf{x}\rangle, \quad \mathbf{x} \in \mathbb{S}^{d-1}\right\}.$$

Dans chacune des deux formules ci-dessus, la dernière égalité vient du fait que

$$\forall \mathbf{x} \in \mathbb{R}^d \setminus \{\mathbf{0}\}, \quad \frac{\langle \mathbf{x}, A\mathbf{x}\rangle}{\|\mathbf{x}\|^2} = \left\langle \frac{\mathbf{x}}{\|\mathbf{x}\|}, A\left(\frac{\mathbf{x}}{\|\mathbf{x}\|}\right)\right\rangle \quad \text{avec } \frac{\mathbf{x}}{\|\mathbf{x}\|} \in \mathbb{S}^{d-1}.$$

Les quotients $\frac{\langle \mathbf{x}, A\mathbf{x}\rangle}{\|\mathbf{x}\|^2}$ sont appelés les *quotients de Rayleigh* de la matrice A.

On cherche donc à minimiser la fonction quadratique f définie dans $\mathbb{R}^d$ par $f(\mathbf{x}) := \langle x, A\mathbf{x}\rangle$ sur la sphère unité. On définira la fonction g comme dans le paragraphe ci-dessus par $g(\mathbf{x}) = \|\mathbf{x}\|^2$.

La fonction f est continue de $\mathbb{R}^d$ à valeurs dans $\mathbb{R}$ et $\mathbb{S}^{d-1}$ est un ensemble fermé et borné dans $\mathbb{R}^d$ Donc, par le théorème de Weierstrass, il existe au moins un point x_- de minimum global de f dans $\mathbb{S}^{d-1}$ et un point x_+ de maximum global de f dans $\mathbb{S}^{d-1}$. De plus, la fonction f est de classe $\mathcal{C}^1$ dans $\mathbb{R}^d$ avec $\nabla f(\mathbf{x}) = A\,\mathbf{x}$.

On cherche donc les points critiques de f sous la contrainte $g(\mathbf{x}) = 1$, c'est-à-dire, en vertu du théorème des extrema liés, les couples $(\mathbf{x}, \lambda) \in \mathbb{R}^d \times \mathbb{R}$ qui satisfont

$$\begin{cases} A\mathbf{x} & = \lambda\mathbf{x}; \\ \|\mathbf{x}\|^2 & = 1. \end{cases}$$

La première équation exprime le fait que λ est une valeur propre de A associée au vecteur propre $\mathbf{x}$ de $\mathbb{R}^d$ (normalisé par $\|\mathbf{x}\| = 1$ par la deuxième équation). En combinant les deux équations, on obtient

$$\lambda = \lambda\langle \mathbf{x}, \mathbf{x}\rangle = \lambda\langle \mathbf{x}, A\mathbf{x}\rangle = f(\mathbf{x}).$$

Donc le minimum global est atteint lorsque le multiplicateur de Lagrange est égal à la plus petite valeur propre, et le maximum global lorsqu'il est égal à la plus grande.

Les points critiques du Lagrangien sont en nombre infini : ce sont tous les couples de la forme $(\mathbf{x}, \lambda)$ où λ est une valeur propre de A et $\mathbf{x}$ un vecteur propre associé, de norme égale à 1.

Le Lagrangien associé à λ_1 est une fonction convexe car c'est une forme quadratique semi-définie positive, et le Lagrangien associé à λ_d est une fonction concave car c'est une forme quadratique semi-définie négative.

Si A admet une valeur propre λ telle que $\lambda_1 < \lambda < \lambda_d$, alors tout vecteur propre $\mathbf{x}_\lambda$ associé à λ, de norme 1, est un point selle du Lagrangien (le démontrer).

6.4 Introduction à l'optimisation sous contraintes d'inégalités

On s'intéresse enfin à des problèmes d'optimisation sous la forme

$$\inf \{f(\mathbf{x}), \quad g(\mathbf{x}) \le 0\},$$

où f et g sont des fonctions continues de $\mathbb{R}^d$ dans $\mathbb{R}$. On peut aussi considérer le cas avec p contraintes d'inégalités, du type $g_1(\mathbf{x}) \le 0$, $g_2(\mathbf{x}) \le 0$, ..., $g_p(\mathbf{x}) \le 0$, mais nous nous restreignons à une seule contrainte dans cette section, par simplicité.

6.4.1 Résolution au cas par cas

De nouveau, on peut considérer l'ensemble

$$K := \left\{\mathbf{x} \in \mathbb{R}^d, \quad g(\mathbf{x}) \le 0\right\},$$

auquel cas notre problème de minimisation est de trouver $\inf f|_K$.

Exercice 6.11

Montrer que si g est continue, alors K est un fermé.
Montrer que si g est coercive, alors K est borné.
Montrer que si g est une fonction convexe, alors K est un ensemble convexe.

Cet ensemble K peut se décomposer en son **intérieur** $\mathring{K}$, et sa frontière ∂K, définis respectivement par

$$\mathring{K} := \left\{\mathbf{x} \in \mathbb{R}^d, \quad g(\mathbf{x}) < 0\right\}, \quad \text{et} \quad \partial K := \left\{\mathbf{x} \in \mathbb{R}^d, \quad g(\mathbf{x}) = 0\right\}.$$

Par continuité de g, l'ensemble $\mathring{K}$ est un ouvert de $\mathbb{R}^d$, et l'ensemble ∂K est une surface de co–dimension 1 (dès que g est une submersion). On a évidemment

$$\inf \{f(\mathbf{x}),\ \mathbf{x} \in K\} = \min \{I_1, I_2\}, \quad \text{avec} \quad I_1 := \inf \left\{f(\mathbf{x}),\ \mathbf{x} \in \mathring{K}\right\}, \quad I_2 := \inf \{f(\mathbf{x}),\ \mathbf{x} \in \partial K\}.$$

On est ramené à l'étude du minimum de f sur l'ouvert $\overset{\circ}{K}$, et sur la surface ∂K.

Lemme 6.12

Si $\mathbf{x}_* \in K$ est un minimiseur local de $f|_K$, et si λ_* est le multiplicateur de Lagrange associé, alors $\lambda_* \le 0$, et on a toujours

$$\lambda_* g(\mathbf{x}_*) = 0 \quad \text{dans le sens} \quad \lambda_* = 0 \quad \text{ou} \quad g(\mathbf{x}_*) = 0.$$

Démonstration. De deux choses l'une. Soit $\mathbf{x}_* \in \overset{\circ}{K}$, soit $\mathbf{x}_* \in \partial K$.
Si $\mathbf{x}_* \in \overset{\circ}{K}$, alors $\nabla f(\mathbf{x}_*) = \mathbf{0}$, c'est à dire $\lambda_* = 0$ (donc $\lambda_* g(\mathbf{x}_*) = 0$). Il reste à montrer que λ_* est négatif.
Comme g est négatif sur K, et positif sur K^c, le gradient de g pointe vers l'extérieur de K. De même, comme f prend des valeurs plus grande que $f(\mathbf{x}_*)$ dans K, le gradient de f pointe vers l'intérieur de K. Ainsi, les directions de $\nabla f(\mathbf{x}_*)$ et $\nabla g(\mathbf{x}_*)$ sont opposées, c'est à dire $\lambda_* \le 0$.

Si $\mathbf{x}_* \in \partial K$, alors $\mathbf{x}_*$ est aussi le minimiseur de f sur la surface ∂K. On peut appliquer la théorie d'optimisation sous contrainte égalité, et déduire que $\nabla f(\mathbf{x}_*) = \lambda_* \nabla g(\mathbf{x}_*)$ pour un $\lambda_* \in \mathbb{R}$. Comme $\mathbf{x}_* \in \partial K$, on a $g(\mathbf{x}_*) = 0$, donc $\lambda_* g(\mathbf{x}_*) = 0$ ☐

On en déduit aussi un «algorithme» naïf pour trouver $\inf f|_K$.

Algorithme de recherche d'optima sous contrainte d'une inégalité $g(\mathbf{x}) < 0$.

1/ Calcul de I_1.
1a) on cherche les points critiques de f, solution de $\nabla f(\mathbf{x}) = \mathbf{0}$.
1b) parmi ces points, on cherche ceux qui vérifient la contrainte $g(\mathbf{x}) < 0$.
1c) et parmi ceux qui reste, on sélectionne celui qui donne la plus petite valeur de f.
$\Longrightarrow$ on trouve I_1.

2/ Calcul de I_2.
2a) on cherche les points critiques de f sur ∂K, solution de $\nabla f(\mathbf{x}) = \lambda g(\mathbf{x})$.
2b) parmi ces points, on cherche celui qui donne la plus petite valeur de f.
$\Longrightarrow$ on trouve I_2.

3/ On compare I_1 et I_2 pour conclure.

6.4.2 Dualité Lagrangienne

La méthode précédente est difficilement applicable dans le cas où il y a p contraintes $g_1(\mathbf{x}) \le 0$, ..., $g_p(\mathbf{x}) \le 0$. Il faut en effet considérer tous les cas, de type

$$\left\{ \mathbf{x} \in \mathbb{R}^d, \quad g_1(\mathbf{x}) < 0, \; g_2(\mathbf{x}) = 0, \; g_3(\mathbf{x}) < 0, \quad \cdots \right\},$$

c'est à dire regarder pour chaque contrainte si elle est saturée (ou **satisfaite**), ou non. Cela donne 2^p cas à étudier, ce qui est impraticable lorsque p est grand.

Nous donnons une autre méthode de résolution, basée sur le Lagrangien. On étudie de nouveau le cas avec une seule contrainte par simplicité (donc un seul multiplicateur de Lagrange $\lambda \in \mathbb{R}$). On pose

$$\mathcal{L}(\mathbf{x}, \lambda) := f(\mathbf{x}) - \lambda g(\mathbf{x}).$$

Dans la suite, on étudiera le Lagrangien en restreignant λ à être **négatif**. On remarquera que pour $\lambda \le 0$ et $g(\mathbf{x}) \le 0$, on a $\mathcal{L}(\mathbf{x}, \lambda) \le f(\mathbf{x})$. Plus exactement, on a le résultat suivant (où on ne suppose plus $g(\mathbf{x}) \le 0$).

Proposition 6.13. *On a, pour tout* $\mathbf{x} \in \mathbb{R}^d$ *fixé,*

$$\sup \{\mathcal{L}(\mathbf{x}, \lambda),\ \lambda \le 0\} = \begin{cases} f(\mathbf{x}) & \text{si} \quad g(\mathbf{x}) \le 0 \\ +\infty & \text{sinon.} \end{cases}$$

Démonstration. Si $g(\mathbf{x}) \le 0$, alors $-\lambda g(\mathbf{x}) \le 0$ pour tout $\lambda \le 0$, avec égalité ssi $\lambda = 0$. Dans ce cas, le supremum est atteint pour $\lambda = 0$, et on trouve $f(\mathbf{x})$.
Si $g(\mathbf{x}) > 0$, alors $-\lambda g(\mathbf{x}) > 0$ tend vers $+\infty$ lorsque $\lambda \to -\infty$, donc le supremum est $+\infty$. ☐

On en déduit le résultat suivant.

Lemme 6.14 : Problème primal

On a

$$\inf \{f(\mathbf{x}), \quad g(\mathbf{x}) \le 0\} = \inf_{\mathbf{x} \in \mathbb{R}^d} \sup_{\lambda \le 0} \mathcal{L}(\mathbf{x}, \lambda). \qquad (P)$$

On appelle (P) la valeur de ce minimum (P comme problème **primal**).

On définit le problème **dual** comme le problème

$$\sup_{\lambda \le 0} \inf_{\mathbf{x} \in \mathbb{R}^d} \mathcal{L}(\mathbf{x}, \lambda). \qquad (D).$$

On a interverti l'ordre $\inf \sup$ et $\sup \inf$. Il n'y *a priori* aucune raison pour que (D) et (P) soit relié. Et pourtant...

Exercice 6.15

Soit $f(x, y) := xy$. Montrer que f a un unique point critique, qui est un point selle, puis montrer que

$$\inf_{x \in \mathbb{R}} \sup_{y \in \mathbb{R}} f(x, y) = \sup_{y \in \mathbb{R}} \inf_{x \in \mathbb{R}} f(x, y) = 0.$$

Lemme 6.16 : Dualité faible

On a toujours l'inégalité $(D) \le (P)$.

Démonstration. Soit $(\mathbf{x}_0, \lambda_0) \in \mathbb{R}^d \times \mathbb{R}_+$ un point quelconque. On a

$$\inf_{\mathbf{x} \in \mathbb{R}^d} \mathcal{L}(\mathbf{x}, \lambda_0) \le \mathcal{L}(\mathbf{x}_0, \lambda_0) \le \sup_{\lambda \ge 0} \mathcal{L}(\mathbf{x}_0, \lambda).$$

En passant au sup en $\lambda_0 \le 0$ (le terme de droite n'en dépend pas), et à l'inf en $\mathbf{x}_0 \in \mathbb{R}^d$ (le terme de gauche n'en dépend pas), on trouve

$$\sup_{\lambda_0 \le 0} \inf_{\mathbf{x} \in \mathbb{R}^d} \mathcal{L}(\mathbf{x}, \lambda_0) \le \inf_{\mathbf{x}_0 \in \mathbb{R}^d} \sup_{\lambda_0 \le 0} \mathcal{L}(\mathbf{x}, \lambda),$$

ce qui est exactement l'inégalité voulue. ☐

On dit qu'on a la **dualité forte** si on a l'égalité $(D) = (P)$. Cette dualité n'est pas toujours vrai, il faut des hypothèses sur f et g.

Théorème 6.17 : Dualité forte, cas convexe

Si f et g sont deux fonctions de classe C^1 et convexes de $\mathbb{R}^d$ dans $\mathbb{R}$, et s'il existe un minimiseur $\mathbf{x}_* \in K$ à $f|_K$ (c'est à dire $f(\mathbf{x}_*) = (P)$), alors $(P) = (D)$.

Démonstration. Pour tout $\lambda \leq 0$, la fonction partielle $\ell_\lambda(\mathbf{x}) := f(\mathbf{x}) - \lambda g(\mathbf{x})$ est convexe sur $\mathbb{R}^d$. Soit $\lambda_* \leq 0$ le multiplicateur d'Euler–Lagrange associé à $\mathbf{x}_*$ (on rappelle que $\lambda_* = 0$ si $\mathbf{x}_* \in \overset{\circ}{K}$). Comme $\mathbf{x}_*$ est un point critique de ℓ_{λ_*}, qui est convexe, c'est le minimum de ℓ_{λ_*}. Ainsi

$$(D) = \sup_{\lambda \leq 0} \inf_{\mathbf{x} \in \mathbb{R}^d} \mathcal{L}(\mathbf{x}, \lambda) \geq \mathcal{L}(\mathbf{x}, \lambda_*) = \ell_{\lambda_*}(\mathbf{x}) \geq \ell_{\lambda_*}(\mathbf{x}_*) = f(\mathbf{x}_*) - \lambda_* g(\mathbf{x}_*) = f(\mathbf{x}_*) = (P),$$

où on a utilisé que $\lambda_* g(\mathbf{x}_*) = 0$. ☐

Le problème dual a plusieurs avantages comparés au problème primal. Par exemple, dans le cadre du théorème précédent, pour tout $\lambda \leq 0$, le problème d'optimisation

$$d(\lambda) := \inf_{\mathbf{x} \in \mathbb{R}^d} \mathcal{L}(\mathbf{x}, \lambda) = \inf_{\mathbf{x} \in \mathbb{R}^d} \ell_\lambda(\mathbf{x})$$

est un problème d'optimisation sans contrainte d'une fonction convexe. Cette optimisation est particulièrement facile à implémenter numériquement (donc $d(\lambda)$ se calcule rapidement). On est ramené ensuite à optimiser $d(\lambda)$ pour $\lambda \leq 0$, ce qui est un problème d'optimisation à une variable sur $\mathbb{R}_-$. De plus, la dérivée de d est facile à calculer, comme le montre le résultat suivant.

Lemme 6.18

Avec les mêmes hypothèses que précédemment, on suppose que pour tout $\lambda \leq 0$, la fonction ℓ_λ a un minimiseur $\mathbf{x}_\lambda \in \mathbb{R}^d$, et que $\lambda \mapsto \mathbf{x}_\lambda$ est de classe C^1. Alors

$$d'(\lambda) = -g(\mathbf{x}_\lambda).$$

Démonstration. Comme $\mathbf{x}_\lambda$ est un point critique de ℓ_λ, on a $\nabla \ell_\lambda(\mathbf{x}_\lambda) = \mathbf{0}$, c'est à dire

$$\forall \lambda \leq 0, \qquad \nabla f(\mathbf{x}_\lambda) - \lambda \nabla g(\mathbf{x}_\lambda) = \mathbf{0}.$$

Par ailleurs, comme x_λ est l'optimum, on a $d(\lambda) = \mathcal{L}(x_\lambda, \lambda)$ pour tout λ. En dérivant et en utilisant la règle de la chaîne, on obtient

$$\begin{aligned} d'(\lambda) &= \partial_{\mathbf{x}} \mathcal{L}(\mathbf{x}_\lambda, \lambda) \frac{\partial}{\partial \lambda} \mathbf{x}_\lambda + \partial_\lambda \mathcal{L}(\mathbf{x}_\lambda, \lambda) \\ &= \left[\nabla f(\mathbf{x}_\lambda) - \lambda \nabla g(\mathbf{x}_\lambda) \right] \frac{\partial}{\partial \lambda} \mathbf{x}_\lambda - g(\mathbf{x}_\lambda), \end{aligned}$$

et le terme entre crochet s'annule par l'équation précédente. ☐