Polycopié du cours :

OPTIMISATION CONVEXE (Première partie)

Edoardo Provenzi

Table des matières

Chapitre 1

Les outils algébriques pour la résolution du problème des moindres carrés

Dans ce chapitre initial on va introduire des outils algébriques, qu'on trouve souvent dans les applications pratiques, pour la résolution d'un problème d'optimisation très simple mais très répandu : le problème des moindres carrés.

1.1 Introduction aux outils algébriques de l'optimisation avec le problème des moindres carrés

Dans les applications des mathématiques on trouve souvent des systèmes d'équations sur-déterminés, i.e. avec un nombre d'équations supérieur au nombre des inconnues, ou sous-déterminés, i.e. avec un nombre d'équations supérieur au nombre des inconnues.

La raison est simple à comprendre : imaginons de devoir déterminer un vecteur x via des expériences et que chaque expérience donne les valeurs d'une équation linéaire satisfaite par x. Les erreurs dans la mesure imposent, quand il est possible, d'estimer x avec une quantité d'expériences supérieure à celle des inconnues. Cela correspond à un système sur-déterminé. Par contre, il est possible que la difficulté d'accès aux données (par exemple la mesure d'une particule qui tombe sur la Terre rarement) fait que le système aït moins d'équations que des inconnues, cette fois-ci on tombe dans le cas d'un système sous-déterminé.

Dans les deux cas, la solution exacte du système n'existe pas, il faut se contenter de calculer le vecteur $\bar{x}$ qui minimise, dans un sens à préciser, les erreurs expérimentales. Pour formaliser mathématiquement cela, on a à disposition le concept de distance entre vecteurs, i.e. la norme de leur différence. En réalité, on va voir que, plutôt que la norme de la différence, on utilisera la *norme au carré* pour des raisons qui seront claires plus tard. La minimisation de la norme au carré entre deux vecteurs est appelée une méthode des « moindres carrés ».

Allons introduire concrètement le problème des moindres carrés avec un exemple très simple. Imaginons de savoir qu'une quantité y dépend linéairement d'une autre quantité x, fixons 4 valeurs de x et allons mesurer les valeurs de y correspondants. Supposons d'obtenir la table suivante :

$\bar{x}$	$\bar{y}$
1	6
2	5
3	7
4	10

On veut trouver la droite d'équation $y = \alpha_1 + \alpha_2 x$ qui détermine le relation linéaire entre x et y. On sait que pour déterminer une droite il faut et il suffit un couple d'équations indépendantes pour α_1 et α_2, donc, avec le système (sur-déterminé) suivant

$$\begin{cases} \alpha_1 + \alpha_2 = 6 \\ \alpha_1 + 2\alpha_2 = 5 \\ \alpha_1 + 3\alpha_2 = 7 \\ \alpha_1 + 4\alpha_2 = 10 \end{cases}$$

on ne peut pas trouver une solution analytique à notre problème, en fait, par exemple, si on considère les deux premières équations, on obtient le système linéaire :

$$\begin{cases} \alpha_1 + \alpha_2 = 6 \\ \alpha_1 + 2\alpha_2 = 5 \end{cases}$$

qui est résolu par $\alpha_1 = 7$ et $\alpha_2 = -1$, mais le couple $(\alpha_1, \alpha_2) = (7, -1)$ n'est pas solution de l'équation $\alpha_1 + 3\alpha_2 = 7$ ni de l'équation $\alpha_1 + 3\alpha_2 = 10$!

Le fait que le système ne soit pas résoluble analytiquement ne veut pas dire qu'on doit renoncer à notre propos, comme on l'a dit avant, il faut changer le paradigme : on se contente de trouver la droite qui mieux approxime la relation de dépendance linéaire entre x et y. Cela implique d'établir un critère d'approximation : quand ce critère est la minimisation de la somme des erreurs quadratiques entre le côté de gauche et le côté de droite des équations du système, alors on parle d'un problème des **moindres carrés**.

Le but des sections suivantes est de montrer que ce problème peut être résolu avec de techniques d'algèbre linéaire relativement simples, élégants et rapides.

Le lecteur est fortement invité à lire l'appendice A, relative aux outils de l'algèbre linéaire, avant de continuer la lecture.

1.2 Résolution d'un système linéaire dans le sens des moindres carrés

Considérons un système linéaire avec m équations et n inconnues écrit sous forme matricielle $Ax = b$, où

$$A \in M_{m,n}(\mathbb{R}), \quad A = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \dots & a_{mn} \end{pmatrix} = (a_{ij})_{j=1,\dots,n}^{i=1,\dots,m}$$

est la matrice des coefficients du système,

$$x \in \mathbb{R}^n, \quad x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

est les vecteur des inconnues, et

$$b \in \mathbb{R}^m, \quad b = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix}$$

est le vecteur des donnés connus, typiquement mesurées.

Déf. 1.2.1 *Le système $Ax = b$ est **résoluble** s'il existe, au moins, une solution, i.e. un vecteur $\bar{x}$ tel que l'équation $A\bar{x} = b$ est une identité.*

Par définition, $Ax = b$ **est résoluble si et seulement si** $b \in$ **Im**(A).

Il est important de rappeler que la solution générale de $Ax = b$ est donnée par *la somme de la solution générale du système homogène associé*, i.e. $Ax = 0$, *et d'une solution particulière de* $Ax = b$. En fait, si x_0 est la solution générale de $Ax = 0$ et $\bar{x}$ est une solution particulière de $Ax = b$, i.e. $A\bar{x} = b$, alors :

$$A(x_0 + \bar{x}) = Ax_0 + A\bar{x} = 0 + b = b,$$

par conséquent, si $\ker(A) \neq \{0\}$, à chaque solution $\bar{x}$ de $Ax = b$ on peut rajouter une solution non nulle de $Ax = 0$, i.e. un vecteur qui appartient à $\ker(A)$, et obtenir une autre solution. Ceci a une conséquence importante : si $Ax = b$ est résoluble, alors, *soit il a une seule solution, soit il en a une infinité*, en fait, s'il existe $x_0 \neq 0$, $x_0 \in \ker(A)$, alors aussi $\lambda x_0 \in \ker(A)$ $\forall \lambda \in \mathbb{R}$, donc on peut construire une infinité de solutions différentes en faisant varier le coefficient λ.

Sous l'hypothèse que $b \in \mathrm{Im}(A)$, analysons les trois possibilités qu'on peut avoir :

- $\boxed{n > m}$: on a plus d'inconnues que d'équations, le système est dit **sous-déterminé**. Comme r =rank$(A) \leqslant \min(m, n) = m < n$, alors, grâce au théorème de nullité + rank (appendice A) $\dim(\ker(A)) = n - m > 0$, donc le système a un nombre infini de solutions, il existe un nombre $n-r$ d'inconnues libres, auxquelles on peut donner un valeur quelconque ;

- $\boxed{n = m}$: même nombre d'inconnues et d'équations, le système est dit **déterminé**. Dans ce cas, la condition nécessaire et suffisante pour l'unicité de la solution est $\mathrm{rank}(A) = n \Leftrightarrow \ker(A) = \{0\}$. Si $\mathrm{rank}(A) < n$, le système a un nombre infini de solutions ;

- $\boxed{n < m}$: plus d'équations que d'inconnues, les système est dit **sur-déterminé**. Si on a n équations indépendantes et les autres $m - n$ sont combinaisons linéaires des précédentes, i.e. si $\mathrm{rank}(A) = n$, alors on a l'unicité de la solution. Autrement, si $\mathrm{rank}(A) < n$, on a des inconnues libres et, donc, un nombre infini de solutions.

Jusqu'ici on a examiné les systèmes résolubles, supposons maintenant que $b \notin \mathrm{Im}(A)$, alors $Ax = b$ n'est pas résoluble de manière exacte, mais, comme $\mathrm{Im}(A)$ est un sous-espace vectoriel de $\mathbb{R}^m$, on a la tentation de remplacer b par le vecteur de $\mathrm{Im}(A)$ le plus proche à lui. Dans l'appendice A on démontre que ce vecteur est la projection orthogonale de b sur $\mathrm{Im}(A)$:

$$b' = P_{\mathrm{Im}(A)}b.$$

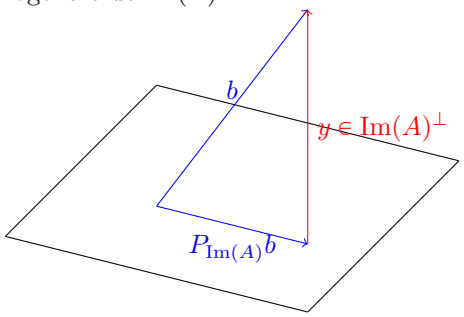
Allons examiner les propriétés du nouveau système linéaire $Ax = b'$.

Théorème 1.2.1 *La résolution du système* $Ax = P_{\mathrm{Im}(A)}b$ *est équivalente à la résolution du problème suivant :*

$$\arg\min_{x \in \mathbb{R}^n} \|Ax - b\|^2.$$

Interprétation du théorème : $A\bar{x} = P_{Im(A)}b$ si et seulement si $\bar{x}$ est le vecteur qui minimise la distance Euclidienne entre Ax et b, i.e. $\|A\bar{x} - b\| \leqslant \|Ax - b\|$ pour tout $x \in \mathbb{R}^n$.

Preuve. Dans l'appendice A on montre que l'erreur y entre b et $P_{\mathrm{Im}(A)}b$ appartient au complément orthogonale de $\mathrm{Im}(A)$:



$$y = b - P_{\mathrm{Im}(A)}b$$

$$b = P_{\mathrm{Im}(A)}b + y$$

$$Ax - b = \underbrace{\underbrace{Ax}_{\mathrm{Im}(A)} - \underbrace{P_{\mathrm{Im}(A)}b}_{\mathrm{Im}(A)}}_{\mathrm{Im}(A)} + \underbrace{-y}_{\mathrm{Im}(A)^\perp}$$

Vu que $Ax - P_{\mathrm{Im}(A)}b$ et y sont orthogonales et comme $\| - y\|^2 = \|y\|^2$, on peut appliquer le théorème de Pythagore généralisé (Annexe A) pour écrire :

$$\|Ax - b\|^2 = \underbrace{\|Ax - P_{\mathrm{Im}(A)}b\|^2}_{\geqslant 0 \ \forall x \in \mathbb{R}^n} + \|y\|^2,$$

donc $\arg\min_{x\in\mathbb{R}^n} \|Ax - b\|^2 = \bar{x}$ tel que $\|A\bar{x} - P_{\mathrm{Im}(A)}b\|^2 = 0$ (car $\|y\|^2$ est une constante par rapport à x), i.e. $A\bar{x} - P_{\mathrm{Im}(A)}b = 0$, d'où $A\bar{x} = P_{\mathrm{Im}(A)}b$.

En résumé, le fait que y, le résidu de la projection, soit perpendiculaire à $\mathrm{Im}(A)$, nous permet d'utiliser les théorème de Pythagore et la définie positivité de la norme pour obtenir le résultat. □

1.3 Les équations normales associées à un système linéaire

Maintenant qu'on a montré l'équivalence entre le nouveau système linéaire $Ax = P_{\mathrm{Im}(A)}b$ et la minimisation de la norme au carré de $Ax - b$, on se pose le problème de déterminer une méthode simple pour résoudre le nouveau problème. Cette méthode sera, automatiquement, une technique d'optimisation !

Encore une fois, les propriétés d'orthogonalité vont nous aider pour déterminer cette technique.

Théorème 1.3.1 *Soient $A \in M_{m,n}(\mathbb{R})$, $x \in \mathbb{R}^n$, $b \in \mathbb{R}^m$, alors :*

$$\boxed{A\bar{x} = P_{Im(A)}b \quad \Leftrightarrow \quad \bar{x} = \arg\min_{x\in\mathbb{R}^n} \|Ax - b\|^2 \quad \Leftrightarrow \quad A^tA\bar{x} = A^tb},$$

i.e. la résolution du système projeté $A\bar{x} = P_{Im(A)}b$, qu'on a démontré être équivalente à la solution du problème des moindres carrés $\arg\min_{x\in\mathbb{R}^n} \|Ax - b\|^2$, est équivalente à la résolution des équations $A^tA\bar{x} = A^tb$.

Preuve.

$\underline{A\bar{x} = P_{\mathrm{Im}(A)}b \implies A^tA\bar{x} = A^tb}$:

$$A\bar{x} = P_{\mathrm{Im}(A)}b \Leftrightarrow A\bar{x} - b = P_{\mathrm{Im}(A)}b - b \Leftrightarrow A\bar{x} - b \underset{y=b-P_{\mathrm{Im}(A)}b}{=} -y \in (\mathrm{Im}(A))^\perp \underset{A.1.5}{=} \ker(A^t),$$

mais $A\bar{x} - b \in \ker(A^t)$ veut dire que $A^t(A\bar{x} - b) = 0$, i.e. $A^tA\bar{x} - A^tb = 0$, donc $A^tA\bar{x} = A^tb$.

$\underline{A^tA\bar{x} = A^tb \implies A\bar{x} = P_{\mathrm{Im}(A)}b}$: $A^tA\bar{x} = A^tb$ implique $A^tA\bar{x} - A^tb = 0$, i.e. $A^t(A\bar{x} - b) = 0$, i.e. $A\bar{x} - b \in \ker(A^t) = \mathrm{Im}(A)^\perp$.

Comme $A\bar{x} \in \mathrm{Im}(A)$ et $b - A\bar{x} \in \mathrm{Im}(A)^\perp$, l'écriture $b = A\bar{x} + (b - A\bar{x})$ est la décomposition orthogonale de b sur $\mathrm{Im}(A)$. Grâce au théorème de la projection A.1.3 on sait que cette décomposition est unique, donc $A\bar{x} = P_{\mathrm{Im}(A)}b$. □

Déf. 1.3.1 *Les équations $A^tAx = A^tb$ sont dites* ***équations normales*** *associées au système $Ax = b$, elle s'appellent normales car elles descendent de l'orthogonalité (aussi dite normalité) entre $b - P_{Im(A)}b$ et $Im(A)$.*

Les équations normales sont obtenues simplement par produit à gauche de la matrice transposée de A aux deux côtés de l'équation $Ax = b$. Il est vraiment remarquable que cette opération, apparemment très simple, permet de transformer un système qui n'a pas forcément une solution, $Ax = b$ quand $b \notin \mathrm{Im}(A)$, dans un problème toujours résoluble, car $b' = P_{\mathrm{Im}(A)}b \in \mathrm{Im}(A)$!

Il faut souligner que c'est le système $Ax = P_{\mathrm{Im}(A)}b$ à être équivalent à $A^tAx = A^tb$ et que, en général, $Ax = b$ **n'est pas équivalent à** $A^tAx = A^tb$, car, comme souligné dans l'appendice A, si M est une matrice avec $\ker(M) \neq \{0\}$, alors $MN = MP$ n'implique pas $N = P$!

Donc, il faut bien se rappeler du fait que

$$A^t(A\bar{x} - b) = 0 \not\Longrightarrow A\bar{x} - b = 0\,!$$

1.4 Résolution des équations normales, la matrice pseudo-inverse de Moore-Penrose

Dans cette section on va entrer dans les détails de la résolution des équations normales. Encore une fois, on souligne qu'il est fortement conseillé de lire l'appendice A avant d'avancer dans la lecture.

1.4.1 A^tA inversible

Dans l'appendice A on montre que, pour toute matrice A $m \times n$, la matrice A^tA est une matrice carrée de dimension $n \times n$. On va commencer avec le cas le plus simple : imaginons que A **soit full rank**, i.e. $\mathrm{rank}(A) = n$, alors, comme on le démontre dans l'appendice A, A^tA est inversible, i.e. $\exists (A^tA)^{-1}$, et la solution des équations normales, i.e. du problème des moindres carrés, est obtenue très simplement comme ça :

$$\bar{x} = I\bar{x} = (A^tA)^{-1}A^tA\bar{x} \underset{A^tA\bar{x} = A^tb}{=} (A^tA)^{-1}A^tb$$

donc :

$$\boxed{\bar{x} = (A^tA)^{-1}A^tb} \text{ est la solution de } \arg\min_{x \in \mathbb{R}^n} \|Ax - b\|^2 \text{ quand } \exists (A^tA)^{-1}.$$

La caractérisation du projecteur $P_{\mathrm{Im}(A)}$ est immédiate :

$$A\bar{x} = P_{\mathrm{Im}(A)}b \quad \Leftrightarrow \quad A\bar{x} = A(A^tA)^{-1}A^tb,$$

donc, dans ce cas, $P_{\mathrm{Im}(A)} = A(A^tA)^{-1}A^t$, qu'on a démontré être un projecteur dans l'appendice A. Si la dimension de A est grande, la formule qu'on vient de déterminer est trop computationnellement couteuse à cause de l'inversion matricielle. On verra dans la suite des techniques plus efficaces.

1.4.2 A^tA non inversible, mais diagonale

Supposons maintenant que A^tA ne soit pas inversible, i.e. $\mathrm{rank}(A) < n$. Dans l'appendice A on a examiné les propriétés de A^tA, pour le moment on a besoin que du fait que A^tA est une matrice carrée de taille n réelle et symétrique.

Les matrices réelles symétriques sont toujours diagonalisables (en fait elles sont des matrices diagonales « en cachette »), en fait un théorème basique d'algèbre linéaire dit que si $M \in M_n(\mathbb{R})$, $M^t = M$, alors il existe une base orthonormée de $\mathbb{R}^n$ donnée par les vecteurs propres de M, si P est la matrice qui a comme colonnes les vecteurs de cette base, alors P est une matrice orthogonale, i.e. $P^{-1} = P^t$ et

$$P^{-1}MP = D,$$

où $D = \text{diag}(\lambda_1, \ldots, \lambda_n)$. Donc M, réelle et symétrique, est semblable à une matrice diagonale qui a les valeurs propres de M sur la diagonale.

Cette observation justifie la volonté de commencer l'analyse de la résolution des équations normales quand A^tA n'est pas inversible, mais diagonale, car l'extension au cas général sera très simple.

Soit, donc :

$$A^tA = \text{diag}(d_1, \ldots, d_r, 0, \ldots, 0),\ d_i \neq 0,\ i = 1, \ldots, r$$

(modulo une permutation des colonnes, on peut toujours représenter A^tA comme ça, i.e. mettre les valeurs non nulles de la diagonale en premier et les 0 après), bien évidemment $r < n$, car si $r = n$, A^tA serait inversible !

On reprend les équations normales :

$$A^tA\bar{x} = \underbrace{A^tb}_{=\bar{b}} \quad \Leftrightarrow \quad \begin{pmatrix} d_1 & & & & & \\ & \ddots & & & & \\ & & d_r & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{pmatrix} \begin{pmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_r \\ \vdots \\ \bar{x}_n \end{pmatrix} = \begin{pmatrix} \bar{b}_1 \\ \vdots \\ \bar{b}_r \\ \vdots \\ \bar{b}_n \end{pmatrix}$$

$$\Leftrightarrow \begin{cases} d_1\bar{x}_1 = \bar{b}_1 \\ \quad \vdots \\ d_r\bar{x}_r = \bar{b}_r \\ 0\bar{x}_j = \bar{b}_j \qquad j = r+1, ..., n \end{cases} \qquad \Leftrightarrow \begin{cases} \bar{x}_i = \frac{\bar{b}_i}{d_i} \qquad i = 1, ..., r \\ \bar{x}_j \textbf{ indéterminées} \qquad j = r+1, ..., n. \end{cases}$$

Un choix pour fixer les variables indéterminées est, par exemple, $\bar{x}_j = 0 \quad j = r+1, ..., n$, ce qui minimise la norme de $\bar{x}$.

Allons maintenant introduire une matrice très utilisée dans la résolution des problèmes de moindres carrés.

Déf. 1.4.1 *On appelle* ***matrice pseudo-inverse de Moore-Penrose***[1] *de $D \in M_n(\mathbb{R})$, la matrice*

$$D^+ = \text{diag}\left(\frac{1}{d_1}, \ldots, \frac{1}{d_r}, 0, \ldots, 0\right) \in M_n(\mathbb{R}).$$

Par calcul direct on voit que

$$\bar{x} = D^+\bar{b} \quad \Leftrightarrow \quad \bar{x} = D^+A^tb$$

formellement, c'est comme si D^+ était $(A^tA)^{-1}$ car

$$\begin{cases} A^tA\bar{x} = \bar{b} \\ \quad \bar{x} = D^+\bar{b} \end{cases}$$

mais $D^+ \neq (A^tA)^{-1}$ car A^tA n'est pas inversible ! En fait, par calcul direct, on obtient

$$D^+A^tA = \text{diag}(1, \ldots, 1, 0 \ldots, 0) \neq I_n,$$

où les valeurs 1 se répètent r fois.

En résumé : si A^tA est non inversible mais diagonale avec rang $r < n$, alors

$$\boxed{\bar{x} = D^+A^tb \text{ est solution de } Ax = P_{\text{Im}(A)}b \quad \Leftrightarrow \quad A^tAx = A^tb \quad \Leftrightarrow \quad \underset{x\in\mathbb{R}^n}{\arg\min}\|Ax - b\|^2}.$$

1. La définition de matrice pseudo-inverse de Moore-Penrose qu'on a donné est un cas particulier d'une théorie, celle des matrices pseudo-inverses, plus riche et compliquée. Néanmoins, on a voulu rester dans ce cadre pour ne pas compliquer inutilement la présentation et parce que la définition qu'on a donné est suffisante pour ce cours.

1.4.3 A^tA non inversible et non diagonale

Si A^tA n'est pas diagonale, on a vu qu'on peut la diagonaliser avec la matrice orthogonale P, $P^{-1} = P^t$, qui a comme colonnes une base orthonormée de $\mathbb{R}^n$ de vecteurs propres de A^tA : $P^tA^tAP = D$. Comme P est inversible, $\ker P = \{0\}$, donc on peut invoquer le théorème A.1.2 et écrire :

$$P^tA^tAP = D \quad \Leftrightarrow \quad PP^tA^tAPP^t = PDP^t \quad \Leftrightarrow \quad A^tA = PDP^t,$$

donc, en reconsidérant les équations normales, on peut écrire

$$A^tA\bar{x} = A^tb \quad \Leftrightarrow \quad PDP^t\bar{x} = A^tb \quad \Leftrightarrow \quad P^tPDP^t\bar{x} = P^tA^tb \quad \Leftrightarrow \quad DP^t\bar{x} = P^tA^tb,$$

qui est un problème très similaire à ce qu'on a examiné dans la section précédente, i.e $D\bar{x} = A^tb$, et qu'on a résolu avec la matrice pseudo-inverse de Moore-Penrose D^+. Pour revenir exactement au cas précédent on fait les changements de variable suivants :

$$\begin{cases} \tilde{\bar{x}} = P^t\bar{x} \\ \tilde{\bar{b}} = P^tA^tb \quad \text{(rappeler que } \bar{b} = A^tb\text{)}, \end{cases}$$

alors, en appliquant la même technique de la section précédente à $D\tilde{\bar{x}} = \tilde{\bar{b}}$, on arrive à écrire $\tilde{\bar{x}} = D^+\tilde{\bar{b}}$, or

$$P^t\bar{x} = D^+P^tA^tb \quad \Leftrightarrow \quad PP^t\bar{x} = PD^+P^tA^tb \quad \Leftrightarrow \quad \bar{x} = PD^+P^tA^tb.$$

En résumé : si A^tA est non inversible et non diagonale, alors

$$\boxed{\bar{x} = PD^+P^tA^tb \text{ est solution de } Ax = P_{\mathrm{Im}(A)}b \quad \Leftrightarrow \quad A^tAx = A^tb \quad \Leftrightarrow \quad \underset{x\in\mathbb{R}^n}{\arg\min}\|Ax - b\|^2},$$

où, encore, P est la matrice orthogonale qui a comme colonnes une base orthonormée de $\mathbb{R}^n$ de vecteurs propres de A^tA.

On voit que la solution nécessite 4 multiplications matricielles appliquées à un vecteur, si la dimension n du problème est élevée, cette quantité d'opérations peut poser des problèmes de coût algorithmique. Dans la prochaine section on va voir une solution moins coûteuse qui passe par une décomposition de A et non pas de A^tA.

1.5 La décomposition en valeurs singulières : SVD

On sait qu'une matrice réelle symétrique A de taille n peut toujours être diagonalisée via la transformation $A = PDP^t$, où P a sur les colonnes une base de $\mathbb{R}^n$ de vecteurs propres de A et D est une matrice diagonale, avec les valeurs propres de A déposés sur la diagonale.

Pour toute matrice $A \in M_{m,n}(\mathbb{R})$ quelconque, on a à disposition une formule, la décomposition en valeurs singulières, qui est un substitut très utile de la formule de diagonalisation.

On commence en rappelant que, dans l'appendice A, on a vu que A^tA est toujours semi-définie positive et que ses valeurs propres λ_i sont toutes $\geqslant 0$ $\forall A \in M_{m,n}(\mathbb{R})$, ce qui nous permet de définir ce qui suit.

Déf. 1.5.1 *On appelle* ***valeurs singulières*** *de $A \in M_{m,n}(\mathbb{R})$ toute les racines carrées des valeurs propres de la matrice A^tA, et on les notes σ_i :*

$$\sigma_i = \sqrt{\lambda_i} \quad avec \quad \lambda_i : valeurs\ propres\ de\ A^tA,$$

c'est courant d'écrire les valeurs singulières de A dans l'ordre décroissant $\sigma_1 \geqslant \sigma_2 \geqslant \ldots \geqslant \sigma_n$, cela étant toujours possible en permutant les colonnes de la matrice orthogonale P qui diagonalise A^tA.

La matrice diagonale dont les éléments diagonaux sont les valeurs singuliéres de A intervient dans une décomposition de A qu'on appelle SVD. Pour prouver cette décomposition, on doit d'abord introduire deux résultats préliminaires.

Lemme 1.5.1 *Hypothèses :*

- *$A \in M_{m,n}(\mathbb{R})$ avec valeurs singulières $\sigma_1, \ldots, \sigma_n$;*
- *$(v_1, \ldots, v_n)$: base orthonormée de $\mathbb{R}^n$ composée par des vecteurs propres de A^tA, avec valeurs propres λ_i, $i = 1, \ldots, n$.*

Alors :

1. $\|Av_i\| = \sigma_i$, $\forall i = 1, \ldots, n$;

2. $\langle Av_i, Av_j \rangle = \lambda_i \delta_{i,j}$, $i, j = 1, \ldots, n$. *En particulier,* $Av_i \perp Av_j$, $\forall i \neq j$.

Preuve.

1. $\|Av_i\|^2 = \langle Av_i, Av_i \rangle = \langle v_i, A^tAv_i \rangle = \langle v_i, \lambda_i v_i \rangle = \lambda_i \langle v_i, v_i \rangle = \lambda_i \|v_i\|^2 = \lambda_i$, car les vecteurs v_i sont unitaires. Donc : $\|Av_i\| = \sqrt{\lambda_i} = \sigma_i$, $\forall i = 1, \ldots, n$.

2. $\langle Av_i, Av_j \rangle = \langle v_i, A^tAv_j \rangle = \langle v_i, \lambda_j v_j \rangle = \lambda_j \langle v_i, v_j \rangle = \lambda_j \delta_{i,j}$ par orthonormalité. □

Lemme 1.5.2 *Sous les mêmes hypothèses du Lemme 1.5.1, et avec l'hypothèse supplémentaire que $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \lambda_r > \lambda_{r+1} = \ldots = \lambda_n = 0$, alors $rank(A) = r$ et :*

$$(Av_1, \ldots, Av_r) \text{ est une base orthogonale de } Im(A).$$

Preuve. L'hypothèse $\lambda_1, \ldots, \lambda_r > 0$ implique que $\sigma_1, \cdots, \sigma_r > 0$ et donc, Lemme 1.5.1 **1.**, $\|Av_1\|, \ldots, \|Av_r\| > 0$. La propriété **2.** du Lemme 1.5.1 implique que les vecteurs $(Av_1, \ldots, Av_r)$ sont linéairement indépendants dans $\mathrm{Im}(A)$ car non nuls et orthogonaux.

Par contre, $\lambda_{r+1}, \ldots, \lambda_n = 0$ implique $\sigma_{r+1}, \ldots, \sigma_n = 0$ et, encore grâce au Lemme 1.5.1 **1.**, $\|Av_{r+1}\| = \ldots = \|Av_n\| = 0$, i.e. $Av_{r+1} = \cdots = Av_n = 0$.

Soit maintenant $w \in \mathrm{Im}(A)$ quelconque, alors $\exists v \in \mathbb{R}^n$ tel que $w = Av$. Allons décomposer v sur la base $(v_1, \ldots, v_n)$: ils existent des scalaires réels c_i, $i = 1, \ldots, n$, tels que $v = \sum_{i=1}^n c_i v_i$, or, par linéarité, $w = Av = \sum_{i=1}^n c_i Av_i = \sum_{k=1}^r c_k Av_k$, vu que $Av_{r+1} = \cdots = Av_n = 0$. Par conséquent, $\mathrm{Im}(A) = \mathrm{span}(Av_1, \ldots, Av_r)$ et, par définition, $\mathrm{rank}(A) = r$. □

Avant d'énoncer et démontrer le théorème sur la décomposition en valeurs singulières on a besoin d'une dernière définition.

Déf. 1.5.2 *Une matrice $M = (m_{ij}) \in M_{m,n}(\mathbb{R})$ est dite* ***pseudo-diagonale*** *si $m_{ij} = 0$ toutefois que $i \neq j$.*

On observe qu'une matrice carrée pseudo-diagonale est diagonale tout-court. Par contre, si elle n'est pas carrée, on peut avoir de situations comme celles-ci :

$$M_1 = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \end{pmatrix}, \quad M_2 = \begin{pmatrix} 1 & 0 \\ 0 & 4 \\ 0 & 0 \end{pmatrix}.$$

Déf. 1.5.3 *Étant donnée une* ***matrice pseudo-diagonale*** *$M \in M_{m,n}(\mathbb{R})$, sa* ***matrice pseudo-inverse de Moore-Penrose*** *M^+ est la matrice de $M_{n,m}(\mathbb{R})$ obtenue en considérant la transposée $M^t \in M_{n,m}(\mathbb{R})$ de M et en remplaçant tous les éléments non-nuls d_i de la diagonale avec leurs inverses d_i^{-1}. Le résultat du produit matriciel M^+M est une matrice carrée $n \times n$ qui a 0 partout, sauf pour les valeurs 1 sur la diagonale, répétées $\min(m, n)$ fois.*

L'exigence de passer par la transposition est essentielle pour avoir une cohérence dimensionnelle. Par exemple, la pseudo-inverse de la matrice pseudo-diagonale $M = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \end{pmatrix}$ est $M^+ = \begin{pmatrix} 1/2 & 0 \\ 0 & 1/3 \\ 0 & 0 \end{pmatrix}$ et $M^+M = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$.

On a maintenant la possibilité de démontrer le théorème qui donne la décomposition en valeurs singulières. On rappelle que le symbole $O(N)$ représente l'ensemble des matrices orthogonales, i.e. matrices réelles carrés de taille N telles que $O^t = O^{-1}$.

Théorème 1.5.1 (SVD) *Soit $A \in M_{m,n}(\mathbb{R})$, $rang(A) = r$ et soient $\sigma_1 \geqslant \sigma_2 \geqslant \ldots \sigma_r > 0$ les valeurs singulières de A. Alors, ça vaut la décomposition en valeurs singulières (SVD) suivante :*

$$\boxed{A = U\Sigma V^t},$$

où $U \in O(m)$, $V \in O(n)$ et Σ est une matrice pseudo-diagonale réelle de taille $m \times n$ telle que :

$$\Sigma = \left(\begin{array}{ccc|c} \sigma_1 & & 0 & \\ & \ddots & & 0 \\ 0 & & \sigma_r & \\ \hline & 0 & & 0 \end{array}\right) = \mathrm{diag}(\sigma_1, \ldots, \sigma_r, 0, \ldots, 0).$$

La SVD ***n'est pas unique***[2]*, néanmoins, dans toute décomposition, les éléments non nuls σ_i de Σ sont les valeurs singulières de A.*

Preuve. A^tA est réelle symétrique, donc il existe une base orthonormée de $\mathbb{R}^n$ composée par de vecteurs propres de A^tA, qu'on écrit comme $(v_1, \cdots, v_n)$.

Le Lemme 1.5.2 garantit que $(Av_1, \cdots, Av_r)$ est une base de $\mathrm{Im}(A)$ qu'on peut normaliser en une base orthonormée de $\mathrm{Im}(A)$ comme ça :

$$(u_1, \ldots, u_r) = \left(\frac{Av_1}{\sigma_1}, \ldots, \frac{Av_r}{\sigma_r}\right),$$

i.e. $Av_k = \sigma_k u_k$, $k = 1, \ldots, r$, où on a utilisé la propriété **1.** du Lemme 1.5.2.

$(u_1, \ldots, u_r)$ peut être étendue à une base orthonormée $(u_1, \cdots, u_m)$ de $\mathbb{R}^m$, l'extension, évidemment, n'est pas unique.

2. Car, comme on le verra dans la preuve, les matrices U, V, en général, ne sont pas univoquement déterminée.

On définit les matrices orthogonales $V \in O(n)$ et $U \in O(m)$ comme les matrices ayant par colonnes sont les vecteurs des bases $(v_1, \cdots, v_n)$ et $(u_1, \cdots, u_m)$, respectivement :

$$U = \ (u_1 \mid \cdots \mid u_m) \ , \quad V = \ (v_1 \mid \cdots \mid v_n).$$

Comme $Av_k = \sigma_k u_k$ pour $k = 1, \ldots, r$ et $Av_j = 0$ pour $j = r+1, \cdots, n$ (par hypothèse sur le rang de A), alors $AV = (Av_1|\cdots|Av_r|0|\cdots|0) = (\sigma_1 u_1|\cdots|\sigma_r u_r|0|\cdots|0)$, donc, si on définit $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_r, 0, \ldots, 0)$, alors $U\Sigma = (\sigma_1 u_1|\cdots|\sigma_r u_r|0|\cdots|0) = AV$, or $U\Sigma V^t = AVV^t$ et, comme $V^{-1} = V^t$, $A = U\Sigma V^t$. □

Allons finalement utiliser la SVD pour résoudre les équations normale. Observons tout d'abord que

$A^tA = V\Sigma^t U^t U\Sigma V^t = V\Sigma^t\Sigma V^t = VDV^t$, où $D = \Sigma^t\Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_r^2, 0, \ldots, 0)$, qui a comme pseudo-inverse de Moore-Penrose la matrice $D^+ = \text{diag}\left(\frac{1}{\sigma_1^2}, \ldots, \frac{1}{\sigma_r^2}, 0, \ldots, 0\right)$.

On a donc $A^tA = VDV^t$ et, par la SVD, $A^t = (U\Sigma V^t)^t = V\Sigma^t U^t$ et alors les équations normales deviennent $A^tA\bar{x} = A^tb \iff VDV^t\bar{x} = V\Sigma^t U^t b$, comme V et V^t sont inversibles, on peut invoquer le théorème A.1.2 et écrire :

$$A^tA\bar{x} = A^tb \iff VDV^t\bar{x} = V\Sigma^t U^t b \iff V^tVDV^t\bar{x} = V^tV\Sigma^t U^t b \iff DV^t\bar{x} = \Sigma^t U^t b.$$

On utilise maintenant D^+ pour pseudo-inverser la dernière équation (comme on l'a fait dans la section 1.4.2) en obtenant $V^t\bar{x} = D^+\Sigma^t U^t b$, où :

$$\begin{aligned} D^+\Sigma^t &= \text{diag}\left(\frac{1}{\sigma_1^2}, \ldots, \frac{1}{\sigma_r^2}, 0, \ldots, 0\right) \text{diag}(\sigma_1, \ldots, \sigma_r, 0, \ldots, 0) \\ &= \text{diag}\left(\frac{1}{\sigma_1}, \ldots, \frac{1}{\sigma_r}, 0, \ldots, 0\right) \\ &= \Sigma^+, \end{aligned}$$

où $\Sigma^+ \in M_{n,m}(\mathbb{R})$ est la pseudo-inverse de la matrice pseudo-diagonale $\Sigma \in M_{m,n}$, qui, comme vu dans la definition 1.5.3, est obtenue en considérant la transposée $\Sigma^t \in M_{n,m}(\mathbb{R})$ de Σ et en remplaçant tous les éléments non-nuls σ_i avec leurs inverses σ_i^{-1}.

Donc, $V^t\bar{x} = \Sigma^+ U^t b \iff VV^t\bar{x} = V\Sigma^+ U^t b$. En résumé :

$$\boxed{\bar{x} = V\Sigma^+ U^t b \textbf{ est solution de } \underset{x\in\mathbb{R}^n}{\arg\min} \|Ax - b\|^2},$$

et, par définition, la pseudo-inverse de A est :

$$A^+ = V\Sigma^+ U^t.$$

Cette solution du problème des moindres carrés nécessite de 3 multiplications matricielles et non 4 comme dans la section précédente. En realité, comme la matrice Σ^+ est diagonale, il restent seulement 2 produits matriciels.

Grâce aux propriétés générales des matrices pseudo-inverses, il est possible de démontrer l'important théorème suivant, qui généralise les calculs qu'on vient de faire.

Théorème 1.5.2 *Soit $A \in M_{m,n}(\mathbb{R})$ et A^+ la matrice pseudo-inverse de Moore-Penrose de A, alors :*

- *A^+A représente le projecteur orthogonale de $\mathbb{R}^n$ sur Lignes(A) ;*
- *AA^+ représente le projecteur orthogonale de $\mathbb{R}^m$ sur Col(A) = Im(A) ;*
- *$\forall y \in \mathbb{R}^m$, $x^+ = A^+y$ est la seule solution des moindres carrés de $Ax = y$ qui appartient à Lignes(A) ;*
- *si rank$(A) = n$, alors $A^+ = (A^tA)^{-1}A^t$ est l'inverse gauche de A ;*
- *si rank$(A) = m$, alors $A^+ = A^t(AA^t)^{-1}$ est l'inverse droite de A ;*
- *si $n = m$ et A est inversible, alors $A^+ = A^{-1}$;*
- *si la SVD de A est $A = U\Sigma V^t$, alors sa pseudo-inverse A^+ est :*

$$A^+ = V\Sigma^+U^t,$$

en particulier, les valeurs singulières de A^+ sont les inverses de ceux de A.

1.5.1 SVD comme solution de norme minimale au problème des moindres carrés

Quand $A \in M_{m,n}(\mathbb{R})$ et $\text{rank}(A) < n$, les solutions du système $Ax = b$ dans le sens des moindres carrés sont infinie, précisément elles sont les vecteurs de la forme

$$x = A^+b + x_0, \quad x_0 \in \ker(A).$$

Parmi toutes ces solutions, celle donnée par $x^+ = A^+b$ **a norme minimale**, en fait $x^+ \in \text{Lignes}(A) = \text{Col}(A^t) = \text{Im}(A^t) = \ker(A)^\perp$, donc par le théorème de Pythagore :

$$\|x\|^2 = \|x^+\|^2 + \|x_0\|^2 \geqslant \|x^+\|^2,$$

donc x^+, parmi les solutions de moindre carré de $Ax = b$, est celle à distance minimale de l'origine.

Vu que, habituellement, on calcule A^+ via la SVD en écrivant $A^+ = V\Sigma^+U^t$, dans beaucoup d'ouvrages on dit que *la solution au problème des moindres carrés offerte par la SVD est celle optimale*, en faisant une liaison entre optimalité et minimalité de la norme de x^+.

Chapitre 2

Convexité

Le problème des moindres carrés examiné dans le chapitre 1 est particulièrement simple parce que la fonction qu'il faut minimiser est la norme Euclidienne au carré. Allons analyser plus en détail l'expression analytique de cette fonction : en $\mathbb{R}$ on a $f(x) = x^2$, en $\mathbb{R}^2$ on a $f(\vec{x}) = \|\vec{x}\|^2 = x^2 + y^2$, en $\mathbb{R}^n$ on a $f(\vec{x}) = \|x\|^2 = x_1^2 + \ldots + x_n^2$, dans le premier cas le graphe de f est représenté par une parabole en $\mathbb{R}^2$, dans le deuxième cas par la surface d'un paraboloïde en $\mathbb{R}^3$ et, dans le cas général, par un paraboloïde en $\mathbb{R}^{n+1}$.

La parabole et les paraboloïdes qui correspondent à la norme Euclidienne au carré ont la propriété d'avoir un seul point de minimum absolu, qui coïncide avec le sommet, comme dans la figure 2.1.

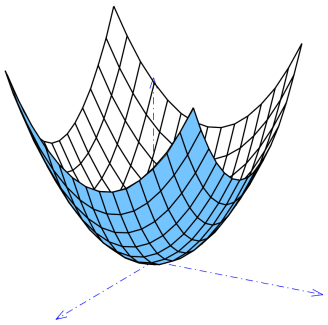


FIGURE 2.1 – La surface d'un paraboloïde qui représente une norme Euclidienne au carré.

Les paraboles et les paraboloïdes sont un cas particulier de fonctions convexes. Dans les sections qui suivent on va introduire les concepts les plus importantes relatifs à la convexité, tout en sachant que la présentation qu'on fera n'est que le début d'une théorie très riche et toujours en développement.

Pour rendre le discours le plus simple possible, on gardera toujours dans l'esprit l'idée de vouloir généraliser la propriété clé des paraboles et des paraboloïdes d'avoir un seul minimum.

Le lecteur est fortement invité à lire l'appendice B, relative aux outils de calcul différentiel, avant de continuer la lecture.

2.1 Ensembles et fonctions convexes

Il y a deux propriétés géométriques des paraboles qui vont nous aider à comprendre comment introduire le concept de convexité : le comportement des sécantes et des tangentes.

On va commencer par les sécantes : dans la figure 2.2 on peut voir que, pour une parabole, la droite sécante en deux points du graphe a toujours une ordonnée supérieure à celle des points du

graphe de la parabole. Si on veut étendre cette propriété à une fonction f définie sur un domaine

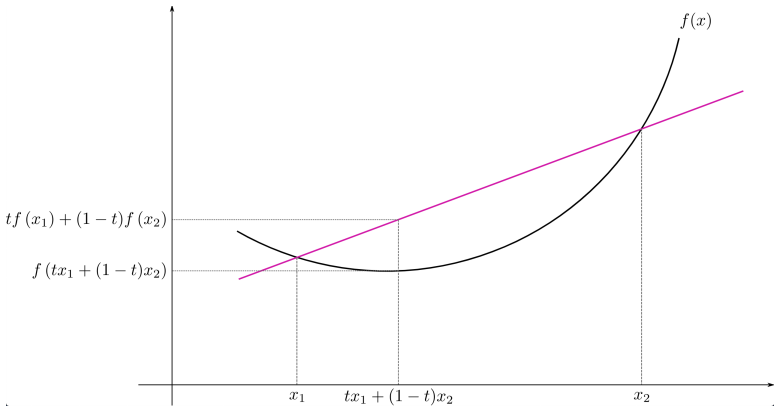


FIGURE 2.2 – Propriété géométrique des sécantes à une parabole.

de $\mathbb{R}^n$ à valeurs réels on doit tout d'abord s'assurer du fait que le segment de la droite qui connecte deux points du domaine reste dans le domaine lui-même. Pour traiter ce problème on a besoin d'introduire les définitions suivantes.

Déf. 2.1.1 *On appelle droite passant par deux éléments distincts x et y de $\mathbb{R}^n$ l'ensemble (infini) défini par*

$$r_{x,y} = \{tx + (1-t)y \ , \ t \in \mathbb{R}\}.$$

On appelle ***segment de droite passant par deux éléments*** *distincts x et y de $\mathbb{R}^n$ l'ensemble (borné) défini par*

$$[x, y] := \{tx + (1-t)y \ , \ t \in [0, 1]\}.$$

Un élément ξ du segment $[x, y]$ s'écrit aussi sous la forme $\xi = y + t(x - y)$. On peut interpréter ξ comme la somme d'une *point initial* y et d'une *direction* $x - y$ pondérée par le paramètre t, qui donne la fraction du chemin reliant y et x où ξ se trouve, en fait, comme t varie de 0 à 1, ξ varie de y à x comme le montre la figure 2.3.

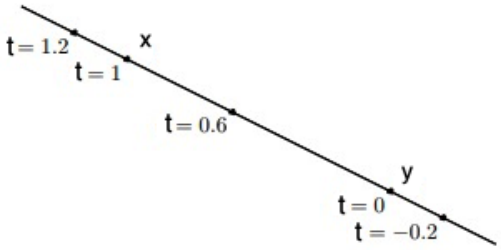


FIGURE 2.3 – La droite reliant x et y est décrite par l'équation paramétrique $tx + (1-t)y$ où $t \in \mathbb{R}$. Le segment reliant x et y est la portion de cette droite qui correspond à $t \in [0, 1]$.

Si l'on remplace dans la définition précédente $[0, 1]$, respectivement, par les intervalles $[0, 1[$, $]0, 1]$ et $]0, 1[$ on définit les segments $[x, y[$, $]x, y]$ et $]x, y[$.

Déf. 2.1.2 *Une partie $E \subset \mathbb{R}^n$ est dite* ***étoilée*** *par rapport à un élément x_0 de E si pour tout x appartenant à E le segment $[x_0, x]$ appartient à E.*

Autrement dit, une partie est étoilée par rapport à un élément x_0 si tout segment d'extrémité x_0 et un élément de cette partie sont inclus dans cette partie. Une interprétation intuitive consiste à dire

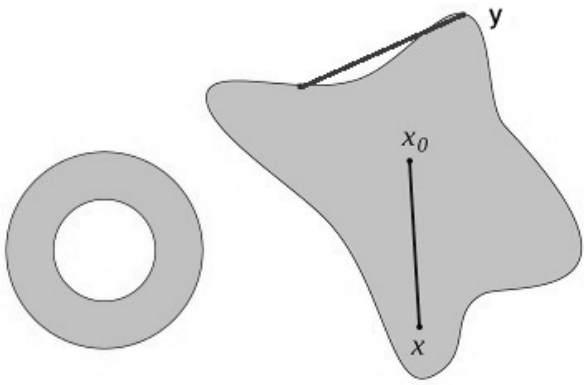


FIGURE 2.4 – Exemples simples de parties étoilée et non étoilée. *À gauche* : une couronne qui n'est pas une partie étoilée (par rapport à aucun élément). *À droite* : il s'agit d'une partie étoilée par rapport à x_0 mais qui n'est pas étoilée par rapport à y.

que, dans une pièce étoilée, il y a toujours une personne pouvant regarder toutes les personnes de la pièce. En figure 2.4 on montre ce concept sous forme graphique en $\mathbb{R}^2$.

Déf. 2.1.3 (Ensemble convexe) *Soit C un sous-ensemble de $\mathbb{R}^n$. On dit que C est un* ***sous-ensemble convexe*** *de $\mathbb{R}^n$ si, pour tout $x, y \in C$, le segment $[x, y]$ appartient à C.*

Dans la suite, par abus de langage et en absence d'ambiguïté, on parlera simplement d'un convexe pour désigner un sous-ensemble convexe de $\mathbb{R}^n$. Un interprétation intuitive consiste à dire que, dans un pièce convexe, deux personnes peuvent toujours s'apercevoir. Dans ce sens, un convexe est donc une pièce sans recoin. On montrera des exemples explicites d'ensembles convexes dans la section 2.1.4. En figure 2.1 on montre un exemple de partie convexe et non convexe en $\mathbb{R}^2$.

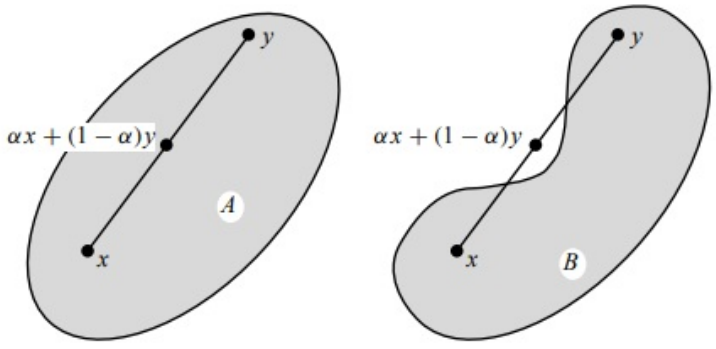


FIGURE 2.5 – Exemples simples de parties convexes et non convexes. A est convexe, B est non convexe.

On a tous les éléments pour définir le concept de fonction convexe, qui va généraliser celui de parabole, en traduisant mathématiquement la relation entre les points d'une parabole et ceux du segment de la droite sécante à la parabole en deux points distincts.

Déf. 2.1.4 (Fonction convexe (concave) et strictement convexe (concave)) *$f : C \to \mathbb{R}$ une fonction définie sur un convexe C en $\mathbb{R}^n$. On dit que f est convexe si :*

$$\forall x, y \in C \quad f(tx + (1-t)y) \leqslant tf(x) + (1-t)f(y) \qquad \forall t \in [0, 1].$$

f est dite strictement convexe si ça vaut la condition suivante[1] *:*

$$\forall x, y \in C \quad f(tx + (1-t)y) < tf(x) + (1-t)f(y) \qquad \forall t \in]0, 1[,$$

$f : C \to \mathbb{R}$, $C \subseteq \mathbb{R}^n$ est concave (strictement concave) si $-f$ est convexe (strictement convexe).

On montrera des exemples explicites de fonctions convexes dans la section 2.2.

Les **fonctions affines**, i.e. celles qui peuvent être écrites comme ça

$$f(x) = \langle a, x \rangle + b = a^t x + b, \qquad a, b \in \mathbb{R}^n,$$

i.e. par la somme d'une forme linéaire et d'une constante, sont toutes et seules les fonctions qui sont convexes et concaves (non strictement) au même temps car elles satisfont l'inégalité non stricte de convexité et de concavité avec une égalité (la preuve dans la section 2.2).

Dans la figure 2.6 on peut voir l'exemple d'une fonction convexe mais non strictement, son graphe montre que les fonctions convexes peuvent avoir une infinité de minima.

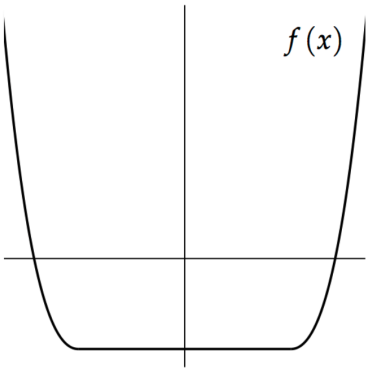


FIGURE 2.6 – Exemple d'une fonction convexe mais non strictement convexe.

2.1.1 Caractérisation au premier ordre de la convexité et ses conséquences

Considérons maintenant le relation géométrique entre droites tangentes et parabole : dans la figure 2.7 (gauche) on peut voir que la droite tangente à une parabole dans un point de son graphe a toujours une ordonnée inférieure à celle du points du graphe de la parabole (à l'inverse des sécantes). Dans la même figure à droite on voit la version 3D avec le plan tangent à la surface d'un paraboloïde.

Comme on veut que les fonctions convexes soient une généralisation des paraboles et paraboloïdes, on attend que la propriété ci-dessus soit respectée par une fonction convexe. Le théorème suivant montre que ceci n'est pas seulement vrai, mais la propriété géométrique qu'on vient d'examiner caractérise toutes et seules les fonctions convexes et donc, comme pour toute caractérisation, elle pourrait être utilisée comme définition alternative de fonction convexe, ce qui est très utile quand l'inégalité qui définit la convexité d'une fonction n'est pas simple à vérifier.

Théorème 2.1.1 (Caractérisation au premier ordre de la convexité d'une fonction) *Soit $f : C \to \mathbb{R}$, C convexe en $\mathbb{R}^n$, f dérivable au moins une fois sur C. Alors :*

$$f \text{ est convexe} \iff f(y) - f(x) \geqslant \langle \nabla f(x), y - x \rangle \qquad \forall x, y \in C. \tag{2.1}$$

De plus, f est strictement convexe $\iff$ $f(y) - f(x) > \langle \nabla f(x), y - x \rangle$ $\qquad \forall x, y \in C.$

1. Observer que $t = 0$ et $t = 1$ ne sont pas considérés car, sinon, on aurait $f(y) < f(y)$ quand $t = 0$ et $f(x) < f(x)$ quand $t = 1$.

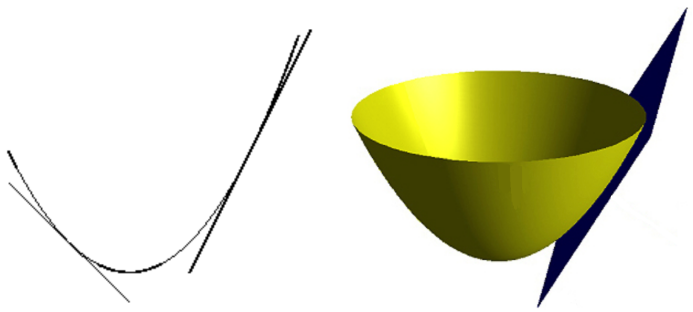


FIGURE 2.7 – Propriété géométrique des tangentes à une parabole et du plan tangent à un paraboloïde.

Avant de démontrer le théorème, allons l'interpréter géométriquement :

- Si $n = 1$, alors la thèse du théorème dit que $f(y) - f(x) \geqslant f'(x)(y - x)$, i.e. $f(y) \geqslant f(x) + f'(x)(y - x)$, à gauche on trouve les valeurs des ordonnés du graphe de f, tandis que à droite on trouve les valeurs de l'ordonné sur la droite tangente au graphe de f en x. Ce qu'on cherchait.
- Si $n = 2$, alors, en développant le produit scalaire, on trouve $f(y) \geqslant \partial_{x_1} f(x_1, x_2)(y_1 - x_1) + \partial_{x_2} f(x_1, x_2)(y_2 - x_2)$, cette fois-ce à droite on trouve les valeurs des ordonnés du plan tangent au graphe de f en $x = (x_1, x_2)$, définit par l'équation $z = \partial_{x_1} f(x)(y_1 - x_1) + \partial_{x_2} f(x)(y_2 - x_2)$, encore une fois, ceci est cohérent avec la propriété géométrique qu'on voulait.
- Plus en général, l'équation $z = \langle \nabla f(x), y - x \rangle$ définit l'hyperplan tangent à la surface de f en x, i.e. son approximation au premier ordre, l'inégalité de la thèse du théorème est traduite souvent avec l'expression suivante : *l'hyperplan tangent est un minorant affine* en chaque point de la surface d'une fonction convexe.

Preuve.

$\boxed{\Rightarrow}$: comme f est convexe ça vaut que

$$\forall x, y \in C \quad f(tx + (1 - t)y) \leqslant tf(x) + (1 - t)f(y) \qquad \forall t \in [0, 1],$$

on peut réécrire $tx + (1 - t)y = y + t(x - y)$ et $tf(x) + (1 - t)f(y) = f(y) + t(f(x) - f(y))$, en obtenant

$$\forall x, y \in C \quad f(y + t(x - y)) \leqslant f(y) + t(f(x) - f(y)) \qquad \forall t \in [0, 1].$$

Pour tout $t \in]0, 1]$ (on considérera 0 comme cas limite) on peut diviser par t les deux côtés

$$\forall x, y \in C \quad \frac{f(y + t(x - y))}{t} \leqslant \frac{f(y)}{t} + f(x) - f(y) \qquad \forall t \in]0, 1],$$

i.e.

$$f(x) - f(y) \geqslant \frac{f(y + t(x - y)) - f(y)}{t},$$

en passant à la limite $t \to 0$ au deux côtés, et comme $f(x) - f(y)$ ne dépend pas de t, on obtient

$$f(x) - f(y) \geqslant \lim_{t \to 0} \frac{f(y + t(x - y)) - f(y)}{t}$$

grâce à la dérivabilité de f (qui est dans les hypothèses) la limite existe et elle donne $D_{x-y} f(y)$, la dérivée directionnelle de f en direction du vecteur $x - y$ calculée dans le point y. Grâce au théorème du gradient (B.1) on peut réécrire $D_{x-y} f(y) = \langle \nabla f(y), x - y \rangle$, donc

$$f(x) - f(y) \geqslant \langle \nabla f(y), x - y \rangle \qquad \forall x, y \in C,$$

x, y étant arbitraires, on peut les échanger, en obtenant

$$f(y) - f(x) \geqslant \langle \nabla f(x), y - x \rangle \qquad \forall x, y \in C,$$

qui est l'implication directe du théorème.

$\boxed{\Leftarrow}$: si ça vaut $f(y) - f(x) \geqslant \langle \nabla f(x), y - x \rangle \ \forall x, y \in C$, alors, en particulier, on peut considérer le point $z = tx + (1-t)y$, qui appartient à C par convexité, et écrire les deux inégalités suivantes

$$f(x) - f(z) \geqslant \langle \nabla f(z), x - z \rangle, \tag{2.2}$$

$$f(y) - f(z) \geqslant \langle \nabla f(z), y - z \rangle. \tag{2.3}$$

On va multiplier les deux côtés de (2.2) par t et celles de (2.3) par $1-t$ (on observe que, comme ces quantités sont positives, l'ordre des inégalités ne change pas) :

$$tf(x) - tf(z) \geqslant t\langle \nabla f(z), x - z \rangle,$$

$$(1-t)f(y) - (1-t)f(z) \geqslant (1-t)\langle \nabla f(z), y - z \rangle.$$

La somme des côtés gauches des deux dernières inégalités est supérieure à la somme des côtés droits, i.e.

$$tf(x) - tf(z) + (1-t)f(y) - (1-t)f(z) \geqslant t\langle \nabla f(z), x - z \rangle + (1-t)\langle \nabla f(z), y - z \rangle,$$

un peu de maquillage mathématique :

$$tf(x) - \cancel{tf(z)} + (1-t)f(y) - f(z) + \cancel{tf(z)} \geqslant \langle \nabla f(z), t(x-z) \rangle + \langle \nabla f(z), (1-t)(y-z) \rangle$$

$$tf(x) + (1-t)f(y) - f(z) \geqslant \langle \nabla f(z), tx - \cancel{tz} + y - z - ty + \cancel{tz} \rangle$$

$$tf(x) + (1-t)f(y) - f(z) \geqslant \langle \nabla f(z), tx + (1-t)y - z \rangle.$$

Mais, par définition, $z = tx + (1-t)y$, donc $tx + (1-t)y - z = 0$ et alors $\langle \nabla f(z), tx + (1-t)y - z \rangle = 0$, ce qui implique

$$tf(x) + (1-t)f(y) - f(z) \geqslant 0 \iff tf(x) + (1-t)f(y) \geqslant f(z) \iff tf(x) + (1-t)f(y) \geqslant f(tx + (1-t)y),$$

i.e. $f(tx + (1-t)y) \leqslant tf(x) + (1-t)f(y) \ \forall x, y \in C$ et $\forall t \in [0, 1]$, qui est la définition de convexité de f. Donc l'implication inverse est prouvée.

La preuve de l'affirmation par rapport à la convexité stricte est laissé comme (simple) exercice. □

Le théorème précédent a beaucoup de conséquences importantes, peut être, la plus importante de toutes est la suivante, qui devrait faire comprendre clairement l'intérêt vers la convexité dans la théorie de l'optimisation.

Théorème 2.1.2 (Fermat (1637)) *Soit $f : C \to \mathbb{R}$, C convexe et ouvert*[2] *en $\mathbb{R}^n$, f convexe et différentiable au moins une fois sur C. Alors :*

$$x^* = \arg\min_{x \in C} f(x) \iff \nabla f(x^*) = 0, \tag{2.4}$$

i.e. pour une fonction convexe dérivable au moins une fois sur un ouvert, la condition nécessaire de stationnarité $\nabla f(x^*) = 0$ pour l'existence des extrema dévient une condition nécessaire et suffisante pour l'existence de minima. Dit d'une manière encore plus directe : les points stationnaires d'une fonction convexe et dérivable sur un ouvert sont des minima.

2. On souligne l'hypothèse de l'ouverture de C pour la validité du théorème.

Preuve. On sait que $\nabla f(x^*) = 0$ est nécessaire pour avoir $x^* = \arg\min_{x\in C} f(x)$, montrons qu'elle est aussi suffisante sous les conditions du théorème. Pour cela, tout ce qu'on doit faire est d'utiliser la caractérisation au premier ordre de la convexité d'une fonction en remplaçant x avec x^* et $\nabla f(x^*)$ avec 0 dans l'éq. (2.1) :

$$f(y) - f(x^*) \geqslant \langle \nabla f(x^*), y - x^* \rangle = \langle 0, y - x^* \rangle = 0 \qquad \forall y \in C,$$

i.e. $f(y) \geqslant f(x^*)\ \forall y \in C$, i.e. $x^* = \underset{x\in C}{\arg\min} f(x)$. □

Une deuxième conséquence du théorème 2.1.1 est une autre caractérisation de la convexité, la monotonie de la dérivée première [3] (en 1D) ou du gradient (en dimension supérieure à 1). On commence avec la dimension 1.

Théorème 2.1.3 *Soit $f :]a, b[\to \mathbb{R}$, $a, b \in \mathbb{R}$, $a < b$, f dérivable en $]a, b[$. Alors :*

$$f \text{ est convexe} \iff f' :]a, b[\to \mathbb{R} \text{ est monotone croissante}$$

De plus, f est strictement convexe $\iff$ f' est strictement croissante.
Si, de plus, f est dérivable deux fois sur $]a, b[$, alors :

$$f \text{ est convexe} \iff f' :]a, b[\to \mathbb{R} \text{ est monotone croissante} \iff f''(x) \geqslant 0 \qquad \forall x \in]a, b[,$$

et pareil pour la stricte convexité.

Preuve.

$\boxed{\Rightarrow}$: soient f une fonction convexe et $\forall x_1, x_2 \in C$ une couple de points qui satisfont la relation d'ordre suivante : $x_2 \geqslant x_1$. Pour démontrer que f' est croissante, il faut démontrer qu'elle préserve la relation d'ordre, i.e. que $f'(x_2) \geqslant f'(x_1)$. Pour arriver à ça, on utilise la caractérisation (2.1) pour les couples de points (x_1, y) et (x_2, y), avec $y \in C$ arbitraire :

$$f(y) \geqslant f(x_1) + f'(x_1)(y - x_1), \tag{2.5}$$

$$f(y) \geqslant f(x_2) + f'(x_2)(y - x_2) \tag{2.6}$$

comme y est arbitraire, on peut choisir $y = x_2$ dans la (2.5), en obtenant :

$$f(x_2) \geqslant f(x_1) + f'(x_1)(x_2 - x_1)$$

qui donne des informations significatives quand $x_2 \neq x_1$, i.e. $x_2 - x_1 > 0$, en fait, dans ce cas, l'inégalité précédente dévient

$$\frac{f(x_2) - f(x_1)}{x_2 - x_1} \geqslant f'(x_1)$$

Également, on peut choisir $y = x_1$ dans la (2.6), en obtenant

$$f(x_1) \geqslant f(x_2) + f'(x_2)(x_1 - x_2) \iff f(x_1) \geqslant f(x_2) - f'(x_2)(x_2 - x_1)$$

d'où, en considérant encore le cas significatif $x_1 \neq x_2$:

$$f'(x_2) \geqslant \frac{f(x_2) - f(x_1)}{x_2 - x_1}.$$

En résumé :

$$f'(x_2) \geqslant \frac{f(x_2) - f(x_1)}{x_2 - x_1} \geqslant f'(x_1),$$

3. Penser encore à la parabole peut aider à représenter graphiquement la monotonie de la dérivée. Par simplicité considérons $f(x) = (x - s)^2$, alors $f'(x) = 2(x - s)$, qui tend vers $-\infty$ quand $x \to -\infty$, elle augment vers 0 quand $x = s$ (le sommet) et elle augmente encore vers $+\infty$ quand $x \to +\infty$.

i.e. $f'(x_2) \geqslant f'(x_1)$, i.e. f' est croissante. Si f' est dérivable, nous savons que f' est croissante sur $]a,b[$ si et seulement si sa dérivée première est positive, i.e. $(f')' = f'' \geqslant 0$ sur $]a,b[$.

$\boxed{\Leftarrow}$: soit f' croissante, $x \in]a,b[$ fixé, et considérons la fonction auxiliaire

$$g(y) = f(y) - f'(x)(y-x) - f(x), \qquad y \in]a,b[.$$

La dérivée première de g par rapport à sa variable y est : $g'(y) = f'(y) - f'(x)$, mais vu que f' est croissante, $g'(y) \lesseqgtr 0$ si $y \lesseqgtr x$, i.e. x est un minimum global pour g.

Allons calculer la valeur de g dans son minimum $g(x) = f(x) - f'(x)(x-x) - f(x) = 0$, i.e. la valeur minimale atteinte par g est 0, par conséquent $g(y) \geqslant 0$ $\forall y \in]a,b[$, mais alors, par définition de g, $\forall y \in]a,b[$ ça vaut : $f(y) - f'(x)(y-x) - f(x) \geqslant 0$, i.e. $f(y) - f(x) \geqslant f'(x)(y-x)$, que, par (2.1) est équivalent à la convexité de f.

La preuve de l'affirmation par rapport à la convexité stricte est laissé comme (simple) exercice. □

Pour étendre ce résultat aux dimensions supérieures à 1 on a besoin d'un résultat (très important) intermédiaire, une ultérieure caractérisation de la convexité, qu'on va examiner dans la section suivante.

2.1.2 La convexité est une propriété unidimensionnelle : caractérisation de la convexité via monotonie

Dans cette section on va examiner formaliser un fait qui devrait déjà être clair depuis la définition de fonction convexe : la convexité est une propriété unidimensionnelle.

La formalisation passe par fixer un convexe $C \subseteq \mathbb{R}^n$, deux points $x, y \in C$ et considérer le sous ensemble de $\mathbb{R}$ définit comme ça

$$D_{x,y} = \{t \in \mathbb{R} \ : \ x + ty \in C\} \subseteq \mathbb{R}, \tag{2.7}$$

i.e. $D_{x,y}$ contient les valeurs de t tels que le segment de la droite d'équation $z = x + ty$, qui passe par x en direction de y, est inclus dans le convexe C.

Si $f : C \subseteq \mathbb{R}^n \to \mathbb{R}$ est une fonction quelconque, on peut définir la fonction réelle de variable réelle (et donc unidimensionnelle) suivante

$$\begin{array}{rccl} g: & D_{x,y} \subseteq \mathbb{R} & \longrightarrow & \mathbb{R} \\ & t & \longmapsto & g(t) = f(x+ty), \end{array}$$

i.e. g prends les valeurs de f restreinte au segment de la droite d'équation $z = x + ty$ contenu en C. Avec un abus de langage plutôt fréquent, mais tout à fait déchiffrable à la lumière de ce qu'on vient de décrire avec rigueur, on dit que g est la restriction de f à la direction y en C.

Le théorème suivant formalise le caractère unidimensionnel de la convexité.

Théorème 2.1.4 *Avec les notations ci-dessus, f est (strictement) convexe $\iff$ g est (strictement) convexe.*

Preuve.

$\boxed{\Rightarrow}$: par l'absurde, supposons que f soit convexe et que g ne le soit pas, i.e. que

$$\forall \lambda \in [0,1],\ \exists t, s \in D_{x,y} \text{ tels que : } \quad g(\lambda t + (1-\lambda)s) > \lambda g(t) + (1-\lambda) g(s)$$

i.e., par définition de g,

$$f(x + [\lambda t + (1-\lambda)s]y) > \lambda f(x+ty) + (1-\lambda) f(x+sy), \tag{2.8}$$

mais :

$$\begin{aligned} x + [\lambda t + (1-\lambda)s]y &= x + \lambda ty + sy - \lambda sy \\ &= (\text{maquillage mathématique : } \pm \lambda x) \\ &= \lambda x + \lambda ty + x + sy - \lambda x - \lambda sy \\ &= \lambda(x + ty) + (1-\lambda)(x + sy). \end{aligned}$$

Si on écrit $x + ty \equiv \xi$ et $x + sy \equiv \eta$, on peut reformuler l'inégalité (2.8) comme ça :

$$\forall \lambda \in [0,1] \ : \quad f(\lambda\xi + (1-\lambda)\eta) > \lambda f(\xi) + (1-\lambda)f(\eta),$$

mais, vu que $t, s \in D_{x,y}$, ξ et η représentent deux points arbitraires de C, par définition de $D_{x,y}$, donc la dernière inégalité qu'on a écrit est absurde, car elle est contraire à la convexité de f.

$\boxed{\Leftarrow}$: c'est pratiquement la même preuve à l'inverse, pour varier un peu on n'utilise pas l'argument par l'absurde. g est convexe si $\forall t, s \in D_{x,y}$ et $\forall \lambda \in [0,1]$, $g(\lambda t + (1-\lambda)s) \leqslant \lambda g(t) + (1-\lambda)g(s)$, i.e.

$$f(x + [\lambda t + (1-\lambda)s]y) \leqslant \lambda f(x + ty) + (1-\lambda)f(x + sy). \tag{2.9}$$

On observe que

$$\begin{aligned} f(x + [\lambda t + (1-\lambda)s]y) &= f(x + \lambda ty + sy - \lambda sy) \\ &= (\text{maquillage mathématique : } \pm \lambda x) \\ &= f(\lambda x + \lambda ty + x + sy - \lambda x - \lambda sy) \\ &= f(\lambda(x + ty) + (1-\lambda)(x + sy)). \end{aligned}$$

Posons, comme avant, $x + ty \equiv \xi$ et $x + sy \equiv \eta$, alors on peut réécrire (2.9) comme ça :

$$\forall \xi, \eta \in C, \quad f(\lambda\xi + (1-\lambda)\eta) \leqslant \lambda f(\xi) + (1-\lambda)f(\eta) \qquad \forall \lambda \in [0,1],$$

i.e. la convexité de f.

La preuve de l'affirmation par rapport à la convexité stricte est laissé comme (simple) exercice. □

Maintenant on peut étendre le théorème 2.1.3 à dimensions supérieures.

Corollaire 2.1.1 *Soit $f : C \subseteq \mathbb{R}^n$, C convexe et ouvert, f différentiable au moins une fois sur C. Alors f est (strictement) convexe $\iff$ la restriction de ∇f sur $D_{x,y}$ est (strictement) croissante pour n'importe quel choix de $x, y \in C$.*

Preuve. Ce résultat est une conséquence directe du théorème précédent et de la caractérisation (2.1.3) de la convexité (et de la convexité stricte). En fait, ça suffit d'observer que la restriction du gradient de f sur l'axe unidimensionnel défini par la droite qui a la direction du vecteur $y \in C$ et qui passe par x est la dérivée première de la fonction réelle de variable réelle $g(t) = f(x + ty)$, $t \in D_{x,y}$.

On sait que f est convexe $\iff$ g est convexe (théorème 2.1.4), i.e. monotone croissante (théorème 2.1.3), d'ici la thèse. □

2.1.3 Caractérisation au second ordre de la convexité

Dans cette section of va donner une autre caractérisation de la convexité très importante et utile, cette fois-ci sous l'hypothèse d'existence de la dérivée seconde. Dans la preuve, on profitera pour voir en action l'artillerie mathématique qu'on vient de développer dans les sections précédentes.

Théorème 2.1.5 *Soit $f : C \subseteq \mathbb{R}^n \to \mathbb{R}$, C convexe et ouvert, et soit f deux fois différentiable sur C, alors :*

$$f \text{ est convexe} \iff \text{la matrice Hessienne } Hf(x) \text{ est semi-définie positive } \forall x \in C.$$

De plus, f est strictement convexe $\iff$ $Hf(x)$ est définie positive $\forall x \in C$.

Preuve. Si $n = 1$, alors[4] C est un intervalle ouvert de $\mathbb{R}$ et la matrice Hessienne de f en x est simplement la dérivée deuxième de f : $f''(x)$. Grâce au théorème 2.1.3, on sait que f est convexe sur $C \iff f'$ est croissante sur C, mais, comme C est un intervalle, ceci est équivalent à dire que $f''(x) \geqslant 0 \ \forall x \in C$.

L'extension au cas $n > 1$ est faite à l'aide du théorème 2.1.4. L'argument est le suivant :

- la convexité de f est équivalente à celle des ses restrictions unidimensionnelles $g(t) = f(x+ty)$, $\forall x, y \in C$, $t \in D_{x,y}$, où $D_{x,y}$ est défini en (2.7) ;
- on vient de démontrer que g est convexe $\iff$ $g''(t) \geqslant 0 \ \forall t \in D_{x,y}$;
- grâce à la formule (B.4), on peut vérifier (c'est un exercice utile qu'on invite à faire...) que $g''(t) = \frac{1}{2}\langle Hf(x+ty)y, y\rangle$. Donc, la positivité de $g''(t)$ est équivalente à la semi-définie positivité de Hf.

La preuve de l'affirmation par rapport à la convexité stricte est laissé comme (simple) exercice. □

Exemple. Allons voir un exemple d'utilisation de ce théorème, qui est particulièrement efficace quand $n = 2$. Soit $f(x, y) = \frac{1}{xy}$ définie sur $C = \{(x, y) \in \mathbb{R}^2, x, y > 0\}$. Déterminer si f est convexe.

C est évidemment convexe. Allons utiliser la caractérisation au deuxième ordre : la matrice Hessienne de f s'écrit

$$Hf(x, y) = \frac{1}{xy}\begin{pmatrix} \frac{2}{x^2} & \frac{1}{xy} \\ \frac{1}{xy} & \frac{2}{y^2} \end{pmatrix}$$

$Hf(x, y)$ est une matrice réelle et symétrique, donc elle est diagonalisable. Si $P(x, y)$ est la matrice qui a sur les colonnes les valeurs propres de $Hf(x, y)$, alors la matrice $P(x, y)Hf(x, y)P^{-1}(x, y)$ a sur la diagonale les deux valeurs propres λ_1, λ_2 de $Hf(x, y)$ et 0 ailleurs.

Nous rappelons que $\det(Hf(x, y)) = \det(P(x, y)Hf(x, y)P^{-1}(x, y)) = \lambda_1 \cdot \lambda_2$ et aussi que $\mathrm{Tr}(Hf(x, y)) = \mathrm{Tr}(P(x, y)Hf(x, y)P^{-1}(x, y)) = \lambda_1 + \lambda_2$, vu que le déterminant et la trace d'une matrice sont invariants par changement de base.

Par calcul direct :

$$\det Hf(x, y) = \frac{3}{x^4y^4} = \lambda_1 \cdot \lambda_2 > 0,$$

donc les deux valeurs propres de Hf ont le même signe et ils sont non nuls. De plus :

$$\mathrm{Tr}\ Hf(x, y) = \frac{2}{xy}\left(\frac{1}{x^2} + \frac{1}{y^2}\right) = \lambda_1 + \lambda_2 > 0,$$

ce qui implique que $Hf(x, y)$ a deux valeurs propres strictement positifs, donc Hf est définie positive et alors f est strictement convexe.

4. Un argument élégant pour démontrer que $f''(x) \geqslant 0$ implique la convexité de f dans le cas $n = 1$ passe par la formule de Taylor au deuxième ordre avec reste de Lagrange, qui dit qu'il existe ξ appartenant au segment de droite entre x et y, tel que $f(y) \underset{y\to x}{=} f(x) + f'(x)(y - x) + \frac{1}{2}f''(\xi)(y - x)^2$. Si $f''(x) \geqslant 0 \ \forall x \in C$, alors $\frac{1}{2}f''(\xi)(y - x)^2 \geqslant 0$ et alors $f(y) - f(x) \geqslant f'(x)(y - x)$, i.e. la caractérisation de la convexité au premier ordre.

2.1.4 Exemples d'ensembles convexes

Commençons avec des exemples élémentaires.

1. Dans $\mathbb{R}$ les ensembles convexes sont exactement les intervalles.

2. Une droite reliant deux éléments de $\mathbb{R}^n$ est un ensemble convexe.

3. Les sous-espaces vectoriels de $\mathbb{R}^n$ sont convexes.

4. Dans $\mathbb{R}^n$, $\overline{U_r(c)} = \{x \in \mathbb{R}^n \ : \ \|x - c\| \leqslant r\}$ le **voisinage fermé de centre** $c \in \mathbb{R}^n$ **et de rayon** $r > 0$ **associée à une norme quelconque** $\|\ \|$ est un ensemble convexe. Donc, en particulier, les cercles, les sphères, les carrés et les cubes sont convexes.
   En effet, en utilisant l'homogénéité et l'inégalité triangulaire, si $\|x - c\| \leqslant r$, $\|y - c\| \leqslant r$ et $t \in [0, 1]$, on a

   $$\|tx + (1-t)y - c\| = \|t(x-c) + (1-t)(y-c)\| \leqslant t\|x-c\| + (1-t)\|y-c\| \leqslant tr + (1-t)r = r.$$

   En figure 2.8 on peut voir des exemples avec trois normes. Le résultat reste vrai aussi pour les voisinages ouverts $U_r(c)$.

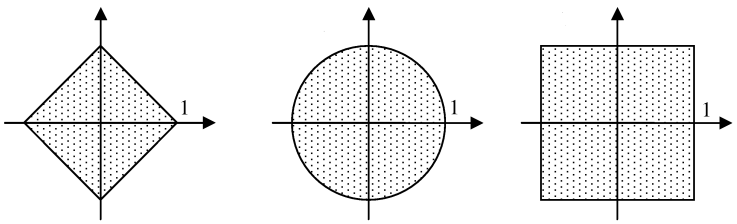


FIGURE 2.8 – Exemples de voisinages unités fermés dans $\mathbb{R}^2$. *À gauche* : au sens de la norme $\|\|_1$, *au milieu* : au sens de la norme Euclidienne $\|\|_2$, *à droite* : au sens de la norme $\|\|_\infty$. La définition de ces normes sera rappelée dans la section 2.2.1.

Allons maintenant à examiner des exemples plus compliqués et très importantes pour l'optimisation.

n-Simplexes.

Soient $x_0, \ldots, x_n \in \mathbb{R}^n$ affinement indépendantes, i.e. $x_1 - x_0, \ldots, x_n - x_0$ sont linéairement indépendantes, alors un n-simplexe est un ensemble définit comme ça

$$S = \left\{ x \in \mathbb{R}^n \ : \ x = \sum_{i=0}^{n} \alpha_i x_i, \ \alpha_i \geqslant 0 \ \forall i = 1, .., n, \ \ \sum_{i=1}^{n} \alpha_i = 1 \right\}.$$

Cas particuliers : un 2-simplexe est un triangle et un 3-simplexe est un tétraèdre, comme le montre la figure 2.9.

Cône convexe

Les cônes jouent un rôle important dans la formulation des contraintes d'inégalités.

Déf. 2.1.5 *Un ensemble C est appelée* ***cône*** *si pour tout x appartenant à C et $t \geqslant 0$, tx appartient à C.*

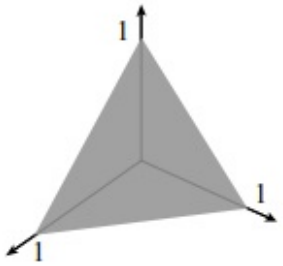


FIGURE 2.9 – Simplexe dans $\mathbb{R}^3$: tétraèdre.

Géométriquement, un cône est la surface obtenue par l'union de demi-droites qui ont une origine commune, l'**apex**, c'est-à-dire le plus haut sommet, et qui connectent l'apex avec une courbe (dite *directrice*) différente de l'apex et non nécessairement fermé. Un cône n'est pas nécessairement un convexe comme le montre la figure 2.10.

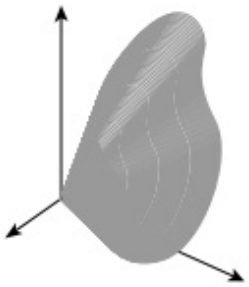


FIGURE 2.10 – Exemple d'un cône non convexe en $\mathbb{R}^3$.

Déf. 2.1.6 *Un **cône convexe** est un ensemble qui est à la fois un cône et un convexe.*

On laisse comme exercice de montrer la caractérisation suivante des cônes convexes.

Théorème 2.1.6 *Un ensemble C est un cône convexe si et seulement s'il est stable par rapport aux combinaisons linéaires avec coefficients non négatifs de ses éléments.*

Les cônes convexes nous donnent la possibilité de montrer des *exemples importants de convexes non bornés* : dans la figure 2.11, les éléments s'écrivant sous la forme $t_1x_1 + t_2x_2$ avec $t_1, t_2 \geqslant 0$ sont les éléments appartenant au domaine d'apex 0 et délimité par les demi-droites passant, respectivement, par x_1 et x_2.

Un exemple simple de cône convexe est l'**orthant positif** : il s'agit de l'ensemble

$$\mathbb{R}^n_+ := \{x \in \mathbb{R}^n | x \geqslant 0\},$$

où la notation $x \geqslant 0$ signifie que pour tout $i \in \{1, ..n\}$, la composante x_i de x est $\geqslant 0$ (figure 2.12).

Hyperplans et demi-planes

Déf. 2.1.7 *On appelle **hyperplan** tout sous-ensemble de $\mathbb{R}^n$ définit par*

$$H_{s,r} = \{x \in \mathbb{R}^n \ : \ s^t x = \langle s, x \rangle = r\}$$

avec $r \in \mathbb{R}$ et $s \in \mathbb{R}^n$.

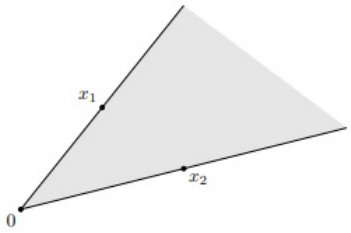


FIGURE 2.11 – Le domaine contient tous les éléments de la forme $t_1x_1 + t_2x_2$ avec $t_1, t_2 \geqslant 0$. L'apex 0 correspond à $t_1 = t_2 = 0$.

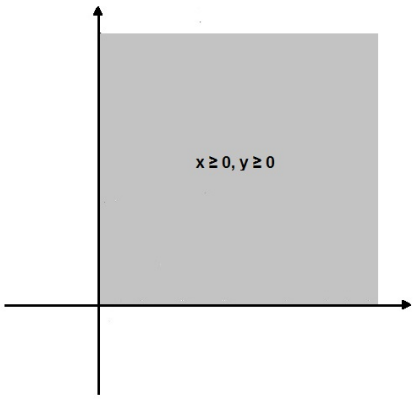


FIGURE 2.12 – Orthant positif dans $\mathbb{R}^2$.

Géométriquement, il s'agit de l'ensemble d'éléments dont le produit scalaire avec un vecteur donné s, appelé **vecteur normal**, reste constant. La constante r détermine le décalage de l'hyperplan affine par rapport à l'origine. Analytiquement, l'hyperplan est la solution de l'équation linéaire $\langle s, x\rangle = r$ d'inconnue x.

Démontrons que $H_{s,r}$ est un convexe : il faut prouver que si $x, y \in H_{s,r}$ et $t \in [0, 1]$, alors $tx + (1-t)y \in H_{s,r}$, i.e. $\langle s, tx + (1-t)y\rangle = r$.

Comme $x, y \in H_{s,r}$, $\langle s, x\rangle = \langle s, y\rangle = r$, donc $\forall t \in [0, 1]$:

$$\langle s, tx\rangle = t\langle s, x\rangle = tr,$$

$$\langle s, (1-t)y\rangle = (1-t)\langle s, y\rangle = (1-t)r,$$

donc, si on fait la somme des côtés gauche et droite des deux dernières équations, on obtient : $\langle s, tx\rangle + \langle s, (1-t)y\rangle = \langle s, tx + (1-t)y\rangle = tr + (1-t)r = r$.

On va expliquer maintenant pourquoi s est dit vecteur normal. Observons maintenant que, si a est un élément quelconque fixé de $H_{s,r}$, on a :

$$H_{s,r} = \{x \in \mathbb{R}^n \ : \ \langle s, x\rangle = \langle s, a\rangle\} = \{x \in \mathbb{R}^n \ : \ \langle s, x - a\rangle = 0\} = \{x \in \mathbb{R}^n \ : \ x - a \in H_{s,0}\}.$$

Fixons s et faisons varier r dans $\mathbb{R}\backslash\{0\}$: les hyperplans $H_{s,r}$ sont les translations (par les vecteurs a) de l'hyperplan $H_{s,0}$. Or $H_{s,0}$ est le sous espace vectoriel des vecteurs perpendiculaires à s que l'on note par $\{s\}^\perp$, on peut donc écrire

$$H_{s,r} = a + \{s\}^\perp,$$

avec $a \in H_{s,r}$. En figure 2.13 on montre la représentation graphique bidimensionnelle qu'on trouve habituellement dans les livres de ce qu'on vient de dire.

Un hyperplan affine divise $\mathbb{R}^n$ en deux sous-ensembles :

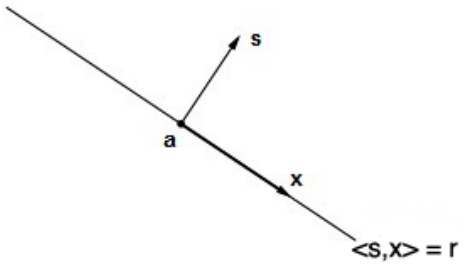


FIGURE 2.13 – Un hyperplan de $\mathbb{R}^2$ de vecteur normal s et a un élément de cet hyperplan. Pour tout élément x de l'hyperplan, $x - a$ est orthogonal à s.

Déf. 2.1.8 *On appelle* ***demi-plan fermé*** *un sous-ensemble de* $\mathbb{R}^n$ *de la forme*

$$H^+_{s,r} := \{x \in \mathbb{R}^n \ : \ \langle s, x\rangle \geqslant r\}$$

ou

$$H^-_{s,r} := \{x \in \mathbb{R}^n \ : \ \langle s, x\rangle \leqslant r\}$$

avec $r \in \mathbb{R}$ *et* $s \in \mathbb{R}^n$.

De même, si a un élément quelconque de l'hyperplan associé, i.e. vérifiant $\langle s, a\rangle = r$, alors $\langle s, x\rangle - \langle s, a\rangle = \langle s, x - a\rangle$ qui est $\geqslant 0$ $\forall x \in H^+_{s,r}$ et $\leqslant 0$ $\forall x \in H^-_{s,r}$, donc les ensembles $H^+_{s,r}$ et $H^-_{s,r}$ peuvent s'écrire, respectivement, sous la forme

$$H^+_{s,r} = \{x \in \mathbb{R}^n \ : \ \langle s, x - a\rangle \geqslant 0\},$$

$$H^-_{s,r} = \{x \in \mathbb{R}^n \ : \ \langle s, x - a\rangle \leqslant 0\}.$$

Ceci nous permet d'interpréter géométriquement $H^-_{s,r}$ comme l'ensemble composé par s plus tout vecteur faisant un angle obtus $\pi/2 \leqslant \vartheta \leqslant \pi$ avec s et $H^+_{s,r}$ comme l'ensemble composé par s plus tout vecteur faisant un angle aigu $0 \leqslant \vartheta \leqslant \pi/2$ avec s.

Dans les figures 2.14 et 2.15 on montre la représentation graphique de cela en 2D.

Les demi-plans sont des ensembles convexes (on fait la preuve dans un de deux cas). Soient $x, y \in H^-_{s,r}$ et $t \in [0, 1]$, montrons que $tx + (1 - t)y$ appartient à $H^-_{s,r}$: $\langle s, tx + (1 - t)y\rangle = t\langle s, x\rangle + (1 - t)\langle s, y\rangle$. Comme $t \in [0, 1]$, $(1 - t) \in [0, 1]$ et $t\langle s, x\rangle + (1 - t)\langle s, y\rangle \leqslant tr + (1 - t)r = r$.

S'il l'on remplace l'inégalité large dans la définition de demi-plan par une inégalité stricte, on obtient la définition de **demi-plan ouvert**.

Puisque le produit scalaire $. \mapsto \langle s, .\rangle$ est une application continue, un demi-plan ouvert (resp. fermé) est un ouvert (resp. fermé) de l'espace $\mathbb{R}^n$ muni de sa topologie usuelle. Un demi-plan ouvert est l'intérieur du demi-plan fermé correspondant et un demi-plan fermé est l'adhérence du demi-plan ouvert correspondant.

Ensembles de sous-niveau

Soit $C \subseteq \mathbb{R}^n$ un convexe et $f : C \to \mathbb{R}$ une fonction convexe. Pour tout $\lambda \in \mathbb{R}$ on définit l'ensemble de λ-sous-niveau de f comme ceci

$$S_\lambda = \{x \in C \ : \ f(x) \leqslant \lambda\},$$

i.e. les points du domaine de f tels que leurs images sont $\leqslant$ à λ, comme le montre la figure 2.16.

Montrons que S_λ est convexe : soient $x, y \in S_\lambda$, il faut montrer que $\forall t \in [0, 1]$, $tx + (1 - t)y \in C$, i.e. que $f(tx + (1 - t)y) \leqslant \lambda$:

$$f(tx + (1 - t)y) \underset{\text{(convexité)}}{\leqslant} tf(x) + (1 - t)f(y) \underset{x,y \in S_\lambda}{\leqslant} t\lambda + (1 - t)\lambda = \lambda.$$

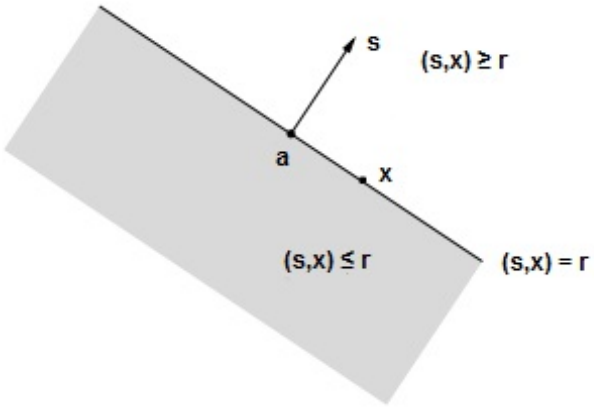


FIGURE 2.14 – L'hyper plan affine divise $\mathbb{R}^2$ en deux demi-plans : un demi-plan (en gris) de $\mathbb{R}^2$ d'équation $\langle s, x\rangle \leqslant r$ se situant dans la direction de $-s$ et celui déterminé par $\langle s, x\rangle \geqslant r$ se situant dans la direction de s.

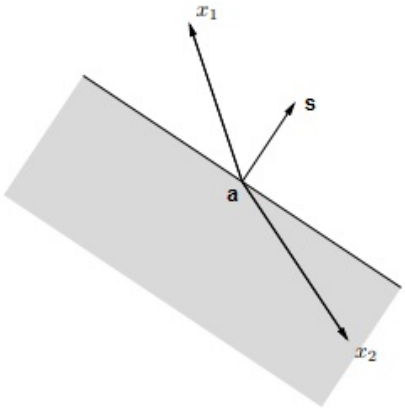


FIGURE 2.15 – Le vecteur $x_1 - a$ fait un angle aigu avec s, il n'appartient pas donc à l'hyperplan $H^-_{s,r}$. Le vecteur $x_2 - a$ fait un angle obtus avec s donc il y appartient.

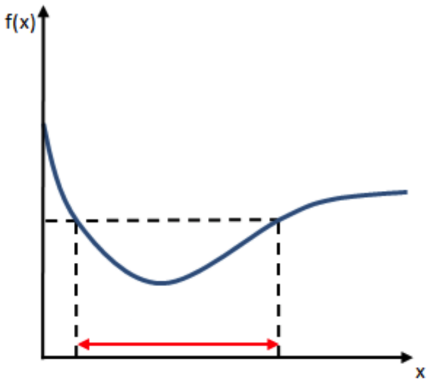


FIGURE 2.16 – En rouge on voit l'intervalle qui représente un ensemble de sous niveau en 2D.

Les ensembles de sur-niveau, définis de la manière qu'on peut imaginer, ne sons pas convexes.

Ellipsoïdes

Déf. 2.1.9 *Soit $c \in \mathbb{R}^n$ et $P \in M(n,\mathbb{R})$ une matrice définie positive. On appelle* ***ellipsoïde*** *de centre c tout sous-ensemble de $\mathbb{R}^n$ de la forme*

$$\mathcal{E} := \{x \in \mathbb{R}^n \ : \ \langle x-c, P(x-c)\rangle \leqslant 1\}.$$

Comme on l'a vu dans l'Annexe 1, une matrice définie positive P peut être écrite comme ça $P = A^tA$, pour une matrice opportune $A \in M(n,\mathbb{R})$, donc la condition $\langle x-c, P(x-c)\rangle \leqslant 1$ dévient

$$\langle x-c, A^tA(x-c)\rangle = \langle A(x-c), A(x-c)\rangle = \|A(x-c)\|^2 \leqslant 1,$$

qui montre que, si $A = \frac{1}{r}I_n$, alors on obtient une hypersphère de rayon r et de centre c comme cas particulier, vue que, dans ce cas, $\|A(x-c)\|^2 \leqslant 1$ est équivalent à $\|x-c\|^2 \leqslant r^2$.

La matrice P détermine comment l'ellipsoïde s'étend à partir du centre c dans chaque direction. Les valeurs propres λ_i de la matrice P donnent les longueurs de ses demi-axes : $1/\sqrt{\lambda_i}$.

Le fait qu'un ellipsoïde est un ensemble convexe suit simplement de la condition $\|A(x-c)\|^2 \leqslant 1$.

Matrices symétriques définies positive

On termine avec un exemple très abstrait, mais qui a beaucoup d'applications. L'ensemble

$$S_n^+ = \{A \in M(n,\mathbb{R}) \ : \ A^t = A, A \text{ semi-définie positive}\},$$

est un cône convexe en $\mathbb{R}^{n^2}$, en fait, il est clair que la multiplication par un coefficient réel positif ne change pas la symétrie ou la définie positivité d'une matrice ; de plus, si $A, B \in S_n^+$ et $t \in [0,1]$, alors, pour tout $x \in \mathbb{R}^n$:

$$\langle x, (tA + (1-t)B)x\rangle = t\langle x, Ax\rangle + (1-t)\langle x, Bx\rangle \geqslant 0,$$

grâce au fait que A, B sont définies positive.

Polyèdres

Déf. 2.1.10 *Soient A une matrice réelle de taille $m \times n$ et b un vecteur de $\mathbb{R}^m$. Un polyèdre est un sous-ensemble de $\mathbb{R}^n$ qui s'écrit sous la forme*

$$P := \{x \in \mathbb{R}^n \ : \ Ax \leqslant b\}$$

Si on identifie la j-ème ligne de la matrice A avec un vecteur de $\mathbb{R}^n$ et on la note A^j et, de même, si on note avec b^j la j-ième composante du vecteur b, alors l'ensemble P s'écrit sous la forme

$$P := \{x \in \mathbb{R}^n \ : \ \langle A^j, x\rangle \leqslant b^j, \ j = 1, .., m\},$$

il s'agit donc d'une intersection finie de demi-plans fermés de $\mathbb{R}^n$ (voir figure 2.17). Comme on le verra dans la section suivante, l'intersection d'ensembles convexes est encore un convexe, ce qui donne la prouve de la proposition suivante.

Théorème 2.1.7 *Un polyèdre est un ensemble convexe de $\mathbb{R}^n$.*

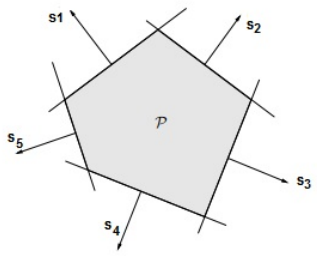


FIGURE 2.17 – Le polyèdre $\mathcal{P}$ est l'intersection des demi-plans de vecteurs normaux $s_1, .., s_5$.

2.1.5 Opérations qui préservent la convexité des ensembles

Allons examiner des opérations qui préservent la convexité des ensembles et leurs conséquences. Commençons pas la suivante.

Théorème 2.1.8 *Soit $(C_i)_{i=1,\dots,n}$ une famille de convexes de $\mathbb{R}^n$. Alors leur intersection $\bigcap_{i=1,\dots,n} C_i$ est un convexe.*

Preuve. Presque immédiate : si $x, y \in \bigcap_{i=1,\dots,n} C_i$, alors, par convexité, $tx + (1-t)y \in C_i$, pour tout $t \in [0,1]$ et $i = 1, \dots, n$, donc $tx + (1-t)y \in \bigcap_{i=1,\dots,n} C_i$, i.e. la convexité de l'intersection. □

Cependant, l'union de convexes n'est pas en général un convexe. Par exemple les segments $[0,1]$ et $[2,3]$ sont des convexes de $\mathbb{R}$, mais $[0,1] \cup [2,3]$ n'est pas un convexe car pour tout $t \in]0,1[$, $t \cdot 1 + (1-t) \cdot 2 = 2 - t$ n'appartient pas à $[0,1] \cup [2,3]$.

Théorème 2.1.9 *Soient $(C_i)_{ii}n1,..,N$ une famille finie de convexes de $\mathbb{R}^{n_i}$. Alors leur produit cartésien $C_1 \times .. \times C_N$ est un convexe de $\mathbb{R}^{n_1} \times .. \times \mathbb{R}^{n_N}$.*

On laisse la simple preuve du théorème comme exercice.

Théorème 2.1.10 *Soit $A : \mathbb{R}^n \to \mathbb{R}^m$ une application affine, alors :*

1. *si C est un convexe de $\mathbb{R}^n$ alors l'image directe de C par A, notée $A(C)$, est un convexe de $\mathbb{R}^m$;*
2. *si D est un convexe de $\mathbb{R}^m$ alors l'image réciproque de C par A, notée $A^{-1}(D)$, est un convexe de $\mathbb{R}^n$.*

Preuve. Il est suffisant d'observer que, si x, y sont deux éléments de $\mathbb{R}^n$, par affinité de A, l'image du segment de droite $[x, y]$ par A est le segment de droite $[A(x), A(y)] \subseteq \mathbb{R}^m$, ceci prouve que $A(C)$ est convexe, mais aussi que $A^{-1}(C)$ l'est, car si x, y sont deux éléments de $\mathbb{R}^n$ tels que $A(x)$ et $A(y)$ sont deux éléments du convexe D, alors tout élément du segment $[x, y]$ a son image dans $[A(x), A(y)] \subseteq D$. □

Des conséquences directes des résultats ci-dessus sont le suivants ($C \subseteq \mathbb{R}^n$).

1. L'**opposé** $-C = \{-x,\ x \in C\}$ d'un convexe C est un convexe.
2. Le **translaté** $a + C = \{a + x,\ x \in C\}$ d'un convexe C par un vecteur a de $\mathbb{R}^n$ est un convexe.
3. L'**homothétie** αC d'un convexe C de rapport $\alpha \in \mathbb{R}$ est un convexe.
4. La **somme vectorielle** de convexes $C_1, C_2 \subseteq \mathbb{R}^n$, $C_1 + C_2 = \{x_1 + x_2,\ x_1 \in C_1,\ x_2 \in C_2\}$, est un convexe.

5. Plus généralement, si $C_1, C_2 \subseteq \mathbb{R}^n$ sont convexes et si α_1 et α_2 sont deux réels, alors $\alpha_1 C_1 + \alpha_2 C_2 = \{\alpha_1 x_1 + \alpha_2 x_2,\ x_1 \in C_1,\ x_2 \in C_2\}$ est un convexe. En fait, il s'agit de l'image directe du convexe $C_1 \times C_2$ par l'application affine $A : (x_1, x_2) \in \mathbb{R}^n \times \mathbb{R}^n \mapsto \alpha_1 x_1 + \alpha_2 x_2 \in \mathbb{R}^n$.

2.2 Comment détecter la convexité de fonctions : fonctions convexes standards et opérations qui préservent leur convexité

Dans la suite du cours, on verra que, dans un problème d'optimisation, détecter la convexité de la fonction objectif et des éventuelles contraintes est cruciale : on verra que les problèmes avec cette propriété possèdent des caractéristiques théoriques très agréables (par exemple, on a vu que les conditions locales nécessaires d'optimalité sont suffisantes pour fonctions convexes) et, ce qui est beaucoup plus important, les problèmes convexes peuvent être résolus efficacement (dans le sens théorique et, dans une certaine mesure, dans le sens pratique de ce mot), ce qui n'est pas, malheureusement, le cas pour des problèmes non convexes généraux. C'est pourquoi il est si important de savoir comment détecter la convexité d'une fonction donnée. On a bien sûr la possibilité d'utiliser les caractérisations au premier et deuxième ordre qu'on a vu, mais on peut aussi suivre la procédure suivante.

Le plan de notre recherche est typique dans le cadre mathématique et c'est exactement ce qu'on utilise en analyse pour détecter la continuité d'une fonction : ça serait vraiment un désastre si chaque fois que nous devons prouver la continuité d'une fonction, nous étions obligés d'utiliser la définition « $\varepsilon - \delta$ » ! Ce qu'on fait c'est d'utiliser cette définition sur les fonctions élémentaires de l'analyse, nos « matières premières », e sur les opérations élémentaires entre elles, nos « outils premiers », comme l'addition, la multiplication, la composition, etc., mais après que cet effort soit fait une seule fois, nous n'avons normalement aucune difficulté à prouver la continuité d'une fonction donnée : il suffit de démontrer qu'elle peut être obtenue, en nombre fini d'étapes, de nos matières premières en appliquant nos outils premiers, i.e. les règles de combinaison qui préservent la continuité. Typiquement, cette démonstration est effectuée par un mot simple « évident » ou même est assumée par défaut.

C'est exactement le cas avec la convexité. Ici nous devons également préciser la liste d'un certain nombre de fonctions convexes standards et d'opérations qui préservent la convexité.

2.2.1 Les fonctions convexes standards

On invite à prouver les premières 7 propriétés de la liste suivante.

- e^x
- $-\log x$
- x^a, avec $x > 0$ et $a \geqslant 1$ ou $a \leqslant 0$.
- $-x^a$, avec $x > 0$ et $a \in [0, 1]$
- $|x|^a$, $x \in \mathbb{R}$, $a \geqslant 1$
- $x \log x$, $x > 0$
- $(x-b)^+ = \max\{x-b, 0\}$ et $(x-b)^- = -\min\{x-b, 0\}$ sont convexes $\forall x, b \in \mathbb{R}^n$
- Les fonctions **affines**, et donc, en particulier, les fonctions **linéaires**, sont convexes, en fait, si $f(x) = \langle a, x\rangle + b$, $a, b, x \in \mathbb{R}^n$, alors $\forall t \in [0, 1]$:

$$\begin{aligned} f(tx + (1-t)y) &= \langle a, tx + (1-t)y\rangle + b = t\langle a, x\rangle + (1-t)\langle a, y\rangle + \underbrace{b + tb - tb}_{tb+(1-t)b} \\ &= t(\langle a, x\rangle + b) + (1-t)(\langle a, y\rangle + b) = tf(x) + (1-t)f(y). \end{aligned}$$

- Une fonction $f : \mathbb{R}^n \to \mathbb{R}$, **positivement homogène de degré 1 et sous-linéaire**, i.e. :

$$f(tx) = tf(x), \quad \forall t \geqslant 0, \text{ et, } \quad f(x+y) \leqslant f(x) + f(y) \qquad \forall x, y \in \mathbb{R}^n,$$

est convexe. En fait : $\forall t \in [0,1]$, $\forall x, y \in \mathbb{R}^n$

$$f(tx + (1-t)y) \underset{\text{sous-lin.}}{\leqslant} f(tx) + (1-t)f(y) \underset{\text{homog.}}{\leqslant} tf(x) + (1-t)f(y).$$

- Comme cas particulier du cas précédent on obtient le très important résultat que : **toutes les normes sont des fonctions convexes** $\| \ \| : \mathbb{R}^n \to \mathbb{R}_0^+$, car elles sont positivement homogènes de degré 1 et, grâce à l'inégalité triangulaire, sous-linéaires. Dans la figure (2.18) on peut voir la représentation graphique en 2D du voisinage de rayon 1 centré en 0 engendré par les normes-ℓ :

$$U_0^\ell(1) = \left\{ x \in \mathbb{R}^2 \ : \ \|x\|_\ell = \sqrt[\ell]{|x_1|^\ell + |x_2|^\ell} = 1 \right\}, U_0^\infty(1) = \left\{ x \in \mathbb{R}^2 \ : \ \|x\|_\infty = \max_{i=1,2} |x_i| = 1 \right\}.$$

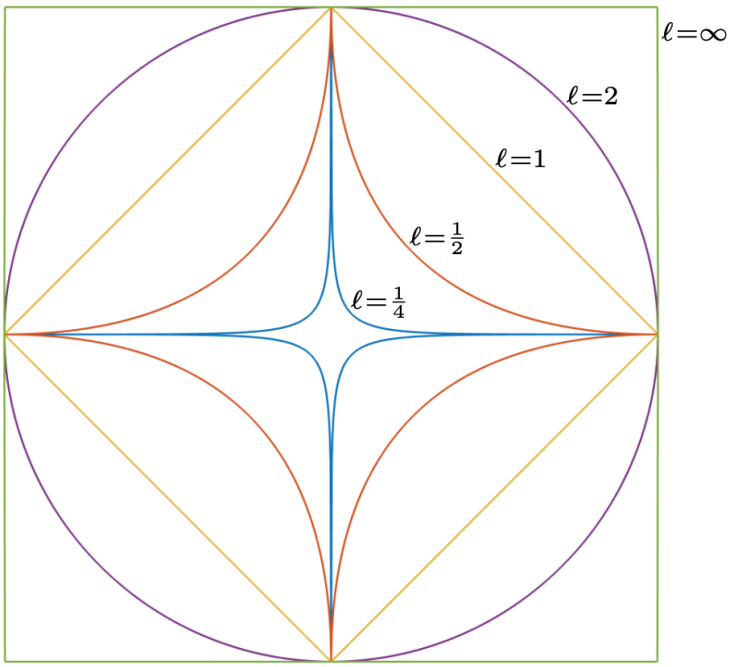


FIGURE 2.18 – Voisinage de rayon 1 centré en 0 engendré par les normes-ℓ en $\mathbb{R}^2$.

Exercice : vérifier que $f(x) = x^2$ est une fonction strictement convexe sur $\mathbb{R}$.

Il faut vérifier que $f(tx+(1-t)y) < tf(x)+(1-t)f(y)$ $\forall t \in]0,1[$ $\forall x, y \in \mathbb{R}$, $x \neq y$, i.e. $[tx+(1-t)y]^2 < tx^2 + (1-t)y^2$.

$$[tx + (1-t)y]^2 \underset{?}{\leq} tx^2 + (1-t)y^2$$

$$t^2x^2 + (1-t)^2y^2 + 2t(1-t)xy \underset{?}{\leq} tx^2 + (1-t)y^2$$

$$(t^2 - t)x^2 + [(1-t)^2 + (1-t)]y^2 + 2t(1-t)xy \underset{?}{\leq} 0$$

$$t(t-1)x^2 + (1-t)[(1-t) + 1]y^2 + 2t(1-t)xy \underset{?}{\leq} 0$$

$$t(t-1)x^2 + t(t-1)y^2 - t(t-1)2xy \underset{?}{\leq} 0$$

$$t(t-1)[x^2 + y^2 - 2xy] \underset{?}{\leq} 0$$

$$\underset{(>0)}{t}\ \underset{(<0)}{(t-1)}\underset{(>0)}{(x-y)^2} \underset{\text{oui !}}{<} 0.$$

2.2.2 Opérations qui préservent la convexité de fonctions

- Si f est une fonction convexe et $k \geqslant 0$, alors kf est une fonctions convexe. La preuve est laissé comme exercice.

- La combinaison conique de fonctions convexes est une fonction convexe, i.e. si $f_1, \ldots, f_n : \mathbb{R}^n \to \mathbb{R}$ sont des fonctions convexes et $c_1, \ldots, c_n \geqslant 0$, alors $\sum_{i=1}^{n} c_i f_i$ est une fonction convexe. On fait la preuve pour $n = 2$, le cas général suit par l'induction. Soient $f, g : \mathbb{R}^n \to \mathbb{R}$ convexes, $a, b \geqslant 0$ et soit $h = af + bg$, alors $\forall t \in [0, 1]$

$$\begin{aligned} h(tx + (1-t)y) &= af(tx + (1-t)y) + bg(tx + (1-t)y) = (af \text{ et } bg \text{ convexes car } a, b \geqslant 0) \\ &\leqslant a[tf(x) + (1-t)f(y)] + b[tg(x) + (1-t)g(y)] = (\text{réarrangement}) \\ &= t[af(x) + bg(x)] + (1-t)[af(y) + bg(y)] \\ &= th(x) + (1-t)h(y). \end{aligned}$$

- Si on fait une combinaison linéaire de fonctions convexes avec des coefficients de signe alterne, alors la fonction qu'on obtient *peut être convexe*, mais on ne peut pas le garantir en général (ça dépend de la relation entre les coefficients et l'expression analytique des fonctions).

- La *composition* d'une fonction convexe *avec une fonction affine* est encore une fonction convexe : $A \in M(n, \mathbb{R})$, $b \in \mathbb{R}^n$, $f : \mathbb{R}^n \to \mathbb{R}$ convexe, alors $g(x) = f(Ax + b)$ $\forall x \in \mathbb{R}^n$ est convexe.

- Si $f_1, f_2 : \mathbb{R}^n \to \mathbb{R}$ sont convexes, alors $f = \max(f_1, f_2)$ est convexe. On verra la preuve de ce résultat dans la section 2.3.

- La *composition* d'une fonction convexe *avec une fonction convexe et croissante* est encore une fonction convexe : $C \subseteq \mathbb{R}^n$ ensemble convexe, $f : C \to \mathbb{R}$ convexe, $\phi : f(C) \to \mathbb{R}$ convexe et croissante, alors $\phi \circ f : C \to \mathbb{R}$ est convexe.

2.2.3 L'interprétation analytique du problème des moindres carrés

Allons voire une conséquence importante de ces propriétés : pour toute matrice $A \in M_{m,n}(\mathbb{R})$ et tout vecteur $b \in \mathbb{R}^n$, la fonction

$$\begin{array}{rccc} f_{A,b}: & \mathbb{R}^n & \longrightarrow & \mathbb{R} \\ & x & \longmapsto & f_{A,b}(x) = \frac{1}{2}\|Ax-b\|^2, \end{array}$$

est convexe, comme on le voit dans le diagramme suivant,

$$\begin{array}{ccccccccc} \mathbb{R}^n & \longrightarrow & \mathbb{R}^n & \longrightarrow & \mathbb{R}_0^+ & \longrightarrow & \mathbb{R}_0^+ & \longrightarrow & \mathbb{R}_0^+ \\ x & \longmapsto & Ax-b & \longmapsto & \|Ax-b\| & \longmapsto & \|Ax-b\|^2 & \longmapsto & \frac{1}{2}\|Ax-b\|^2, \end{array}$$

elle est obtenue par composition entre une fonction affine, une fonction convexe (la norme Euclidienne), une fonction convexe et croissante (sur $\mathbb{R}_0^+$!) et par multiplication d'un coefficient positif.

Comme on l'a vu dans la section B.2.2, formule (B.2.5), $\nabla f_{A,b}(x) = A^t(Ax-b)$, donc les équations de Euler-Lagrange pour la fonction $f_{A,b}$ sont :

$$\nabla f_{A,b}(\bar{x}) = 0 \iff A^t(A\bar{x}-b) = 0 \iff A^tA\bar{x} = A^tb,$$

i.e. le point stationnaire $\bar{x}$ de la fonction $f_{A,b}(x) = \frac{1}{2}\|Ax-b\|^2$ est les solutions des équations normales associées au système linéaire $Ax = b$ qu'on sait être la solution du système dans le sens des moindres carrés.

Vu que $f_{A,b}$ est convexe, les points stationnaires de $f_{A,b}$ sont de minima, ça montre l'interprétation analytique au problème des moindres carrés, qui se rajoute à l'interprétation géométrique et algébrique, comme résumé ci-dessous.

Les trois interprétations alternatives du problème des moindres carrés

$$\bar{x} = \underset{x\in\mathbb{R}^n}{\arg\min}\, \|Ax-b\|^2$$

— **Interprétation géométrique** : $A\bar{x} = P_{\mathrm{Im}(A)}b$, i.e. résolution du système linéaire projeté sur l'espace image de A ;

— **Interprétation algébrique** : $A^tA\bar{x} = A^tb$, i.e. résolution des équations normales ;

— **Interprétation analytique** : $\nabla\left(\frac{1}{2}\|A\bar{x}-b\|^2\right) = 0$, i.e. résolution des équations de Euler-Lagrange associées à la fonction $f_{A,b}(x) = \frac{1}{2}\|Ax-b\|^2$.

2.3 Lien entre ensembles convexes et fonctions convexes : épigraphe et hypographe, enveloppe convexe

Dans cette section on va formaliser le lien entre fonctions et ensembles convexes. Commençons par rappeler que le graphe d'une fonction $f : \Omega \subseteq \mathbb{R}^n \to \mathbb{R}$, est le sous-ensemble de $\mathbb{R}^{n+1}$ défini par :

$$\text{graphe}(f) = \{(x, y) \in \Omega \times \mathbb{R} \ : \ y = f(x)\}.$$

Le graphe de f est un sous-ensemble de l'ensemble suivant, qui joue un rôle très important dans la théorie de l'optimisation.

Déf. 2.3.1 (Épigraphe) *Soit $f : \Omega \subseteq \mathbb{R}^n \to \mathbb{R}$, on appelle épigraphe de f le sous-ensemble de $\mathbb{R}^{n+1}$ défini par :*

$$\text{Epi}(f) = \{(x, \lambda) \in \Omega \times \mathbb{R} \ : \ \lambda \geqslant f(x)\}.$$

Le nom dérive du fait que $\varepsilon\pi\iota$ veut dire « au-dessus », qui fait référence au fait que les valeurs de λ dans la deuxième entrée des coordonnés de l'épigraphe sont au-dessus du graphe de f, comme le montre la figure 2.19.

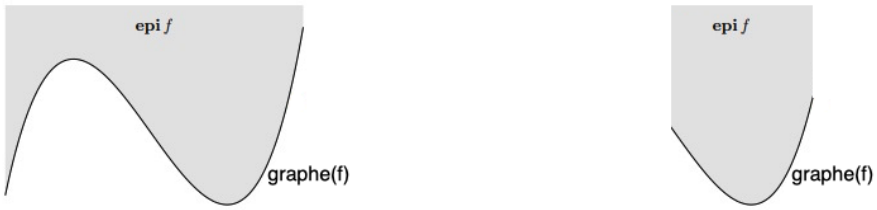


FIGURE 2.19 – Le graphe des fonctions est dessinée en trait foncé. L'épigraphe est constitué de la partie grise et du graphe de la fonction.

La figure montre aussi que l'épigraphe d'une fonction convexe est un ensemble convexe et que celui d'une fonction non convexe n'est pas un ensemble convexe. Le résultat suivant montre que ceci n'est pas un hasard, mais la règle.

Théorème 2.3.1 *Soit $f : C \subseteq \mathbb{R}^n \to \mathbb{R}$ une fonction et C un ensemble convexe. Alors f est convexe (en tant que fonction) si et seulement si Epi(f) est convexe (en tant que ensemble).*

Avant de démontrer le théorème, il faut le commenter : d'un côté, à l'aide de la définition d'épigraphe, on peut construire, à partir d'une fonction convexe f, l'ensemble convexe Epi(f), réciproquement, si $\Omega \subset \mathbb{R}^n \times \mathbb{R}$ est l'épigraphe d'une certaine fonction convexe, alors on obtient cette fonction via

$$f(x) = \inf_{(x,\lambda)\in\Omega} \lambda,$$

car les valeurs du graphe de f sont les minimiseurs des valeurs de λ dans l'épigraphe.

Ceci montre un lien étroit entre les fonctions convexes et les ensembles convexes.

Preuve.

$\boxed{\Rightarrow}$ Supposons que f soit convexe et montrons que Epi(f) est convexe : soient $(x, \lambda), (y, \rho) \in$ Epi(f), il faut démontrer que, pour tout $t \in [0, 1]$, $t(x, \lambda)+(1-t)(y, \rho) = (tx+(1-t)y, t\lambda+(1-t)\rho) \in$ Epi(f), mais ça, par définition, est vrai si et seulement si $t\lambda + (1 - t)\rho \geqslant f(tx + (1 - t)y)$.

Comme f est convexe, on a

$$f(tx + (1 - t)y) \leqslant tf(x) + (1 - t)f(y) \leqslant t\lambda + (1 - t)\rho \text{ (par définition d'épigraphe).}$$

$\boxed{\Leftarrow}$ Supposons que $\mathrm{Epi}(f)$ soit convexe et montrons que f est convexe : on rappelle que le graphe de f est inclus dans son épigraphe, donc $\forall x, y \in C$, $(x, f(x)), (y, f(y)) \in \mathrm{Epi}(f)$, comme on a supposé que $\mathrm{Epi}(f)$ est convexe, pour tout $t \in [0,1]$ ça vaut que $t(x, f(x)) + (1-t)(y, f(y)) \in \mathrm{Epi}(f)$, d'où $(tx + (1-t)y, tf(x) + (1-t)f(x)) \in \mathrm{Epi}(f)$, c'est à dire $f(tx + (1-t)y) \leqslant tf(x) + (1-t)f(x)$. $\square$

Grâce à ce théorème la preuve du fait que, si f_1, f_2 sont convexes, alors $f = \max(f_1, f_2)$ est convexe, est immédiate. En fait, f est une fonction convexe $\iff$ $\mathrm{Epi}(f)$ est un ensemble convexe, mais l'épigraphe de f est l'intersection de l'épigraphe de f_1 et de f_2, qui sont deux ensembles convexes car f_1, f_2 sont convexes. Comme l'intersection d'ensembles convexe est encore un ensemble convexe, $\mathrm{Epi}(f)$ est convexe et donc f est convexe.

Allons renverser l'ordre...

Déf. 2.3.2 (Hypographe) *Soit $f : \Omega \subseteq \mathbb{R}^n \to \mathbb{R}$, on appelle hypographe de f le sous-ensemble de $\mathbb{R}^{n+1}$ défini par :*

$$\mathrm{Hypo}(f) = \{(x, \lambda) \in \Omega \times \mathbb{R} \ : \ \lambda \leqslant f(x)\}.$$

Le nom dérive du fait que hypo veut dire « en dessous », qui fait référence au fait que les valeurs de λ dans la deuxième entrée des coordonnés de l'épigraphe sont en dessous du graphe de f.

On invite le lecteur à démontrer le résultat suivant.

Théorème 2.3.2 *Soit $f : C \subseteq \mathbb{R}^n \to \mathbb{R}$ une fonction et C un ensemble convexe. Alors f est concave (en tant que fonction) si et seulement si Hypo(f) est convexe (en tant que ensemble).*

2.4 Enveloppe convexe, combinaisons linéaires convexes et inégalité de Jensen

Considérons les points dans la figure 2.20 : on peut les envelopper dans un nombre infini d'ensembles convexes, mais le plus petit ensemble convexe qui les enveloppe tous est celui dessiné.

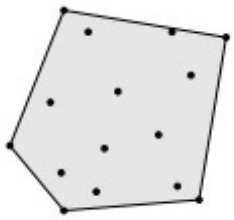


FIGURE 2.20 – Exemples d'enveloppe convexe de quinze éléments de $\mathbb{R}^2$.

On formalise cette observation avec la définition suivante.

Déf. 2.4.1 (Enveloppe convexe d'un ensemble) *L'**enveloppe convexe**, en anglais **convex hull**, d'un ensemble $S \subseteq \mathbb{R}^n$, notée conv (C) ou hull (C), est l'intersection*[5] *de tous les ensembles convexes contenant S. Il s'agit donc du plus petit convexe contenant S. Évidemment, si C est déjà convexe, $C \equiv conv(C)$.*

Comme d'habitude, on veut donner une caractérisation plus opérationnelle d'enveloppe convexe, par exemple pour savoir comment dessiner l'enveloppe convexe d'un ensemble. Cette caractérisation est faite via les combinaisons convexes de n vecteurs de $\mathbb{R}^n$, qu'on va définir ci-dessous et que généralisent le concept de combinaison convexe de deux vecteurs.

Déf. 2.4.2 (Combinaison convexe) *Soient $x_1, \ldots, x_n$ des éléments de $\mathbb{R}^n$ et $\lambda_1, \ldots, \lambda_n$ des réels $\geqslant 0$ tels que $\lambda_1 + \ldots + \lambda_n = 1$. On dit que $x = \lambda_1 x_1 + \ldots + \lambda_n x_n$ est une **combinaison convexe** des éléments $x_1, \ldots, x_n$. Plus généralement, si S est un sous-ensemble de $\mathbb{R}^n$ on dit que $x \in \mathbb{R}^n$ est combinaison convexe d'éléments de S s'il existe un nombre fini d'éléments de S dont x soit une combinaison convexe.*

Une combinaison convexe n'est rien d'autre donc qu'une *moyenne pondérée* et une combinaison convexe de deux éléments n'est rien d'autre que le segment qui les relie, car, dans ce cas $\lambda_1 + \lambda_2 = 1 \iff \lambda_2 = 1 - \lambda_1$.

La proposition suivante donne une importante caractérisation d'un ensemble convexe via les combinaisons convexes de ses éléments.

Théorème 2.4.1 *$C \subseteq \mathbb{R}^n$ est convexe si et seulement si C contient toutes les combinaisons convexes de ses éléments.*

Preuve.

$\boxed{\Leftarrow}$: si un ensemble C contient toutes les combinaisons convexes de ses éléments, alors, comme on l'a vu ci-dessus, il contient, en particulier, tous les segments reliant deux de ces éléments, qui est la définition de convexité.

$\boxed{\Rightarrow}$: soient C un ensemble convexe et y un élément de C s'écrivant sous la forme $y = \sum_{i=1}^{n} \lambda_i x_i$ avec les λ_i des réels positifs vérifiant $\sum_{i=1}^{n} \lambda_i = 1$ et les $x_i \in C$, montrons que y est un élément de C,

5. Cette notion est bien définie car on sait que l'intersection de convexes est un convexe.

i.e. que C contient une arbitraire combinaison convexe de ses éléments. Puisque la sommes des λ_i vaut 1, il existe au moins un $\lambda_i > 0$, quitte à réindexer, supposons que $\lambda_1 > 0$.

On considère la construction suivante :

$$z_1 = \frac{\lambda_1}{\lambda_1 + \lambda_2} x_1 + \frac{\lambda_2}{\lambda_1 + \lambda_2} x_2 \quad (\in C)$$

$$z_2 = \frac{\lambda_1 + \lambda_2}{\lambda_1 + \lambda_2 + \lambda_3} z_1 + \frac{\lambda_3}{\lambda_1 + \lambda_2 + \lambda_3} x_3 \quad (\in C)$$

$$z_{n-1} = \frac{\lambda_1 + \lambda_2 + \ldots + \lambda_{n-1}}{\sum_{i=1}^{n} \lambda_i} z_{n-2} + \frac{\lambda_n}{\sum_{i=1}^{n} \lambda_i} x_n \underset{\sum_{i=1}^{n} \lambda_i = 1}{=} (\lambda_1 + \lambda_2 + \ldots + \lambda_{n-1}) z_{n-2} + \lambda_n x_n \quad (\in C).$$

Avec un instant de réflexion sur la structure des z_k, $k = 1, \ldots, n-1$ (considérer, par exemple, $n = 3$), on constate que $z_{n-1} = y$, donc $y \in C$. □

On est prêt pour démontrer la caractérisation de l'enveloppe convexe d'un ensemble.

Théorème 2.4.2 *L'enveloppe convexe d'un ensemble S, conv(S), coïncide avec l'ensemble $\hat{S}$ de toutes les combinaisons convexes d'éléments de S.*

Preuve.

$\boxed{\text{conv}(S) \subseteq \hat{S}}$: d'un côté, conv(S) est, par définition, le plus petit convexe qui contient S, de l'autre côté, le théorème 2.4.1 dit que $\hat{S}$ est convexe et, bien sûr, $S \subseteq \hat{S}$, car tout élément de S est identifiable comme une combinaison convexe de lui même avec un seul coefficient : $\lambda = 1$! Donc, conv$(S) \subseteq \hat{S}$.

$\boxed{\text{conv}(S) \supseteq \hat{S}}$: conv(S) est un convexe qui contient S, alors, d'après la proposition 2.4.1, il contient toutes les combinaisons convexes d'éléments de S, d'où $\hat{S} \subseteq \text{conv}(S)$. □

En résumé, lorsqu'un ensemble $S \subset \mathbb{R}^n$ n'est pas convexe, on peut considérer le convexe le plus similaire à lui : son enveloppe convexe, qu'on peut construire en connectant avec des segments de droite (issus, justement, de combinaisons convexes, comme prescrit par le théorème qu'on vient de démontrer !) les points extrêmes de l'ensemble original, comme on peut le voir dans la figure 2.21.

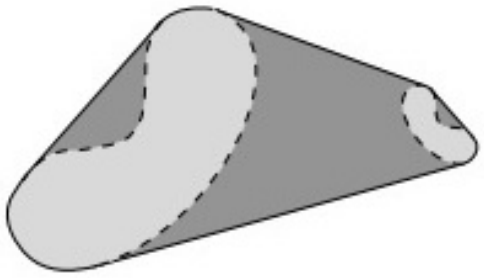


FIGURE 2.21 – Exemples d'enveloppe convexe d'un ensemble non-convexe de $\mathbb{R}^2$.

D'un point de vu théorique, on constate donc que l'enveloppe convexe d'un ensemble S de $\mathbb{R}^n$ peut être construit de deux manières : une construction « interne », en considérant les combinaisons convexes d'éléments de S, et une construction « externe », en considérant l'intersection des convexes contenant S.

Bien que la construction par dessus semble plus naturelle, on utilise souvent en pratique la deuxième, car la description de toutes les combinaisons convexes d'un ensemble peut parfois être compliquée.

On termine cette section avec la relation entre fonctions convexes et combinaisons convexes.

Théorème 2.4.3 (Inégalité de Jensen) *Soit $C \subset \mathbb{R}^n$ un ensemble convexe et $f : C \to \mathbb{R}$ une fonction convexe. Soient $x_i \in C$, $\lambda_i \geqslant 0$, $i = 1, \ldots, n$, $\sum_{i=1}^{n} \lambda_i = 1$, alors :*

$$f\left(\sum_{i=1}^{n} \lambda_i x_i\right) \leqslant \sum_{i=1}^{n} \lambda_i f(x_i).$$

Avant de démontrer ce résultat, on observe que, quand $n = 2$, l'inégalité de Jensen coïncide avec la définition de convexité. Donc, ce théorème dit que satisfaire cette inégalité pour $n = 2$ est équivalent à la satisfaire pour n quelconque, fini.

Preuve. La preuve la plus simple passe par l'épigraphe : les points $(x_i, f(x_i)) \in C \times \mathbb{R}$ appartiennent au graphe de f et donc à son épigraphe, qu'on sait être convexe car f est convexe par hypothèse, donc, grâce au théorème 2.4.1, la combinaison convexe de points de $C \times \mathbb{R}$

$$\sum_{i=1}^{n} \lambda_i (x_i, f(x_i)) = \left(\sum_{i=1}^{n} \lambda_i x_i, \sum_{i=1}^{n} \lambda_i f(x_i)\right)$$

appartient encore à $\mathrm{Epi}(f)$.

Par définition d'épigraphe, ceci implique que

$$\sum_{i=1}^{n} \lambda_i f(x_i) \geqslant f\left(\sum_{i=1}^{n} \lambda_i x_i\right).$$

□

2.5 Fonctions convexes à valeurs dans $\overline{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$

En optimisation, il est parfois intéressant de travailler sur des fonction pouvant prendre des valeurs infinies. Par exemple, quand on travaille sur le problème d'optimisation avec contraintes $\min_{x \in C} f(x)$ (C est un sous-ensemble de $\mathbb{R}^n$), il est parfois utile de le remplacer par le problème d'optimisation sans contraintes $\min_{x \in \mathbb{R}^n} \tilde{f}(x)$ où $\tilde{f}$ prend les mêmes valeurs que f sur C et la valeur $+\infty$ sur le complémentaire de C.

Cette astuce permet de traiter au même temps les problèmes avec ou sans optimisation. D'où, on prend souvent $C = \mathbb{R}^n$ et on autorise f à prendre les valeurs $\pm\infty$.

Quand on autorise f à prendre des valeurs infinies, il y a un ensemble de définitions de l'analyse convexe qu'il faut connaître et qu'on résume ci-dessous.

Déf. 2.5.1 (Fonction indicatrice) *On appelle fonction indicatrice de l'ensemble convexe $C \subseteq \mathbb{R}^n$ la fonction suivante :*

$$\begin{array}{rccl} I_C : & C & \longrightarrow & \{0, +\infty\} \\ & x & \longmapsto & I_C(x) = \begin{cases} 0 & si\ x \in C \\ +\infty & si\ x \notin C. \end{cases} \end{array}$$

Il est clair que *minimiser la fonction $f : C \to \mathbb{R}$ sur C est équivalent à minimiser sur $\mathbb{R}^n$ la fonction $\tilde{f} : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, $\tilde{f} = f + I_C$*, ou, plus précisément :

$$\begin{array}{rccl} \tilde{f} : & \mathbb{R}^n & \longrightarrow & \mathbb{R} \cup \{+\infty\} \\ & x & \longmapsto & \tilde{f}(x) = \begin{cases} f(x) & \text{si } x \in C \\ +\infty & \text{si } x \notin C. \end{cases} \end{array}$$

Déf. 2.5.2 (Domaine effectif) *On appelle domaine effectif, ou simplement domaine, de la fonction $f : \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$ l'ensemble de points $x \in \mathbb{R}^n$ tels que $f(x) \neq +\infty$. On écrit $dom(f)$.*

On admet aussi $-\infty$ dans cette définition pour permettre à $\mathrm{dom}(f)$ d'être convexe lorsque f l'est.

Une fonction identiquement égale à $+\infty$ présente peu d'intérêt. On se limite donc aux fonctions suivantes.

Déf. 2.5.3 (Fonction propre) *$f : \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$ est dite propre si elle n'est pas identiquement égale à $+\infty$, i.e. si $dom(f) \neq \varnothing$.*

Déf. 2.5.4 (Fonction coercive) *$f : \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$ est dite coercive si elle tend vers $+\infty$ quand sa variable tend vers l'infini, i.e. $\lim_{\|x\| \to +\infty} f(x) = +\infty$.*

2.5.1 Les minima locaux d'une fonction convexe propre sont des minima globaux

Le résultat suivant souligne encore plus clairement l'importance de la convexité dans la théorie de l'optimisation.

Théorème 2.5.1 *Soit $f : C \subseteq \mathbb{R}^n \to \bar{\mathbb{R}}$, C convexe, une fonction convexe et propre. S'il existe un minimiseur local $x^* \in \mathbb{R}^n$ de f, alors x^* est aussi un minimiseur global.*

De plus, l'ensemble $Argmin_C(f) \subseteq \mathbb{R}^n$ de tous les minimiseurs locaux (et donc globaux) de f sur C est convexe.

Pour terminer, si f est strictement convexe, elle peut avoir un seul minimiseur (global).

Donc, si $f : C \subseteq \mathbb{R}^n \to \mathbb{R}$ est convexe et propre, elle *peut* avoir seulement de minima globaux, qui peuvent être une infinité (par exemple, un plateaux où f est constante à la valeur minimale) ; mais si f est strictement convexe, alors elle peut avoir seulement un point di minimum, nécessairement, global.

Preuve. Par l'absurde : supposons que $x^* \in C$ soit un minimiseur local de f en C, si x^* n'est pas un minimiseur global, alors il existe $\bar{x} \in C$ tel que $f(\bar{x}) < f(x^*)$, $f(\bar{x}) < +\infty$ car f est propre. Alors, par convexité de C, le segment de la droite qui connecte $\bar{x}$ avec x^*, i.e. $tx^* + (1-t)\bar{x}$, $t \in [0,1]$, est inclus en C et, si on applique la fonction f, par convexité on peut écrire

$$f(tx^* + (1-t)\bar{x}) \leqslant tf(x^*) + (1-t)f(\bar{x}).$$

Analysons cette dernière expression :

— quand $t = 0$ le majorant est $f(\bar{x})$;
— quand $0 < t < 1$ le majorant est une combinaison convexe de $f(\bar{x})$ et $f(x^*)$;
— quand $t = 1$ le majorant $f(x^*)$.

Si on élimine $t = 1$ on a la possibilité d'écrire la majoration suivante :

$$f(tx^* + (1-t)\bar{x}) < f(x^*) \qquad \forall t \in [0,1[.$$

Cette écriture est en contradiction avec l'hypothèse que x^* soit un minimum local pour f, en fait, l'existence d'un segment continu de points de C dans lesquels f prend des valeurs strictement inférieures à $f(x^*)$ empêche l'existence d'un voisinage $U(x^*)$ dans lequel $f(x^*) \leqslant f(\xi)$ $\forall \xi \in U(x^*)$.

Pour montrer que $\operatorname{Argmin}_C(f)$ est un sous-ensemble convexe de $\mathbb{R}^n$ il suffit de noter $\lambda \equiv \min\limits_{x \in C} f(x)$ et d'observe que $\operatorname{Argmin}_C(f)$ est l'ensemble de λ-sous-niveau de f, que l'ont sait être convexe.

Terminons avec le cas de la stricte convexité. Par l'absurde, supposons qu'il existe une couple de points x_1^*, x_2^* qui soient minima (nécessairement globaux, pour ce que l'on vient de démontrer) pour f, en particulier, on observe que : $f(x_1^*) = f(x_2^*) = \min\limits_{x \in C} f(x)$ (sinon, un des deux ne serait pas un minimum !). Par convexité de C, le point au milieu entre x_1^* et x_2^*, i.e.

$$\xi = \frac{x_1^* + x_2^*}{2} = \frac{1}{2}x_1^* + \frac{1}{2}x_2^* = \frac{1}{2}x_1^* + \left(1 - \frac{1}{2}\right)x_2^*$$

appartient à C, si on applique f à ξ on obtient, par convexité stricte :

$$f(\xi) = f\left(\frac{1}{2}x_1^* + \left(1 - \frac{1}{2}\right)x_2^*\right) < \frac{1}{2}f(x_1^*) + \frac{1}{2}f(x_2^*) = \frac{1}{2}\min_{x \in C} f(x) + \frac{1}{2}\min_{x \in C} f(x) = \min_{x \in C} f(x),$$

i.e. $f(\xi) < \min\limits_{x \in C} f(x)$, ce qui est une évidente contradiction. □

2.5.2 Semicontinuité inférieure et existence des minima des fonctions convexes

Dans la section précédente on a vu que, pour une fonction convexe et propre, si un point est un minimum local, alors il est automatiquement un minimum global. Néanmoins, on n'a pas garanti l'existence d'un tel point.

On va introduire ici une condition technique (due au mathématicien Réne Baire) qui garantit l'existence des minima pour les fonctions convexes.

Déf. 2.5.5 (Semicontinuité inférieure) *$f : C \subseteq \mathbb{R}^n \to \bar{\mathbb{R}}$ est semicontinue inférieurement* (SCI) *en $x \in C$ si*

$$\forall \varepsilon > 0,\ \exists \text{ un voisinage } U(x) \text{ tel que } f(y) > f(x) - \varepsilon \quad \forall y \in U(x),$$

i.e.

$$\liminf_{y \to x} f(y) \geqslant f(x) \iff \liminf_{n \to +\infty} f(x_n) \geqslant f(x) \quad \forall (x_n)_{n \in \mathbb{N}} \text{ tel que } x_n \underset{n \to +\infty}{\longrightarrow} x.$$

$f : C \subseteq \mathbb{R}^n \to \mathbb{R}$ est semicontinue inférieurement sur C si elle l'est dans tous les points de C.

Théorème 2.5.2 *Soit $f : C \subseteq \mathbb{R}^n \to \bar{\mathbb{R}}$, C convexe, une fonction*

- *convexe ;*
- *propre ;*
- *SCI sur C.*

Alors, f admet au moins un minimiseur (nécessairement global) dans C. Si on remplace la convexité de f avec la convexité stricte, alors f admet un seul minimiseur (global) dans C.

La preuve de ce théorème n'est pas difficile, mais un peu technique et on préfère l'admettre pour avancer plus rapidement avec le cours.

Appendices

Annexe A

Un très bref rappel d'algèbre linéaire

Dans cette appendice, on va rappeler les concepts d'algèbre linéaire dont on a besoin dans le cours. L'exposition des concepts est volontairement rapide parce qu'on imagine que les lecteurs ont déjà une base d'algèbre linéaire et parce qu'il y a une grande quantité de livres excellentes sur le sujet qui peuvent (doivent...) être consultés comme complément à ces notes. Seulement les preuves des théorèmes qui apportent des éléments d'intérêt pour les cours seront reproduites.

A.1 Généralités

On commence par rappeler que le produit scalaire Euclidien de $\mathbb{R}^n$ est défini comme ça :

$$\forall x, y \in \mathbb{R}^n \ : \ \langle x, y \rangle = x^t y = \sum_{i=1}^{n} x_i y_i,$$

où x^t est le vecteur transposé de x. On rappelle que deux vecteurs $x, y \in \mathbb{R}^n$ sont *orthogonaux* quand $\langle x, y \rangle = 0$. Deux vecteurs orthogonaux sont linéairement indépendants.

La norme Euclidienne, ou norme-2, de $x \in \mathbb{R}^n$, est :

$$\|x\| = \sqrt{\langle x, x \rangle} = \sqrt{x^t x} = \left(\sum_{i=1}^{n} x_i^2 \right)^{1/2}.$$

On utilisera souvent ces deux faits :

— Le seul vecteur orthogonale à tous les autres est le vecteur nul ;

— $\forall x \in \mathbb{R}^n$, $\|x\| = 0 \implies x = 0$, qui entraine $\|x - y\| = 0 \implies x - y = 0$, i.e. $x = y$, ceci donne une technique (standard) pour montrer l'égalité de deux vecteurs via l'analyse de la norme de leur différence.

Étant donné un opérateur linéaire $A : \mathbb{R}^n \to \mathbb{R}^m$, fixé une base de $\mathbb{R}^n$ et de $\mathbb{R}^m$, on peut lui associer univoquement une matrice [1], qu'on écrit encore avec le symbole A :

$$A \in M_{m,n}(\mathbb{R}), \quad A = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \dots & a_{mn} \end{pmatrix} = (a_{ij})_{j=1,\dots,n}^{i=1,\dots,m},$$

i est l'indice des lignes et j l'indice des colonnes, donc la matrice A a :

1. On écrit avec $M_{m,n}(\mathbb{R})$ l'ensemble des matrices $m \times n$ à coefficients réels.

— m vecteurs lignes $L_1, \dots, L_m \in \mathbb{R}^n$ (m comme la dimension de l'espace d'arrivé de l'opérateur A) ;

— n vecteurs colonnes $C_1, \dots, C_n \in \mathbb{R}^m$ (n comme la dimension du domaine de l'opérateur A) ;

— La matrice transposée $A^t \in M_{n,m}(\mathbb{R})$ de la matrice $A \in M_{m,n}(\mathbb{R})$ est la matrice que a par colonnes les lignes de A et par lignes les colonnes de A ;

— $A \in M_{n,n}(\mathbb{R})$ est dite symétrique si $A^t = A$, i.e. l'échange de lignes et colonnes ne change pas la matrice : $a_{ij} = a_{ji}\ \forall i, j = 1, \dots, n$.

Vu l'univocité de l'association entre un opérateur linéaire et sa matrice (une fois fixé les bases du domaine et de l'espace d'arrivé), dorénavant on utilisera le même symbole pour un opérateur linéaire et sa matrice associée et chaque définition relative à un opérateur linéaire sera automatiquement étendue à sa matrice associée.

Un opérateur linéaire $A : \mathbb{R}^n \to \mathbb{R}^n$ est dit *endomorphisme*, et sa matrice associée est carrée. Dans ce cas, le produit scalaire entre x et Ay, $x, y \in \mathbb{R}^n$, peut être écrit comme $\langle x, Ay\rangle = x^t Ay$, qui est une formule qu'on utilisera dans le cours.

Notation : on écrit avec $L(\mathbb{R}^n, \mathbb{R}^m)$ l'espace des opérateurs linéaires de $\mathbb{R}^n$ à $\mathbb{R}^m$ et avec $\mathrm{End}(\mathbb{R}^n)$ l'espace des endomorphismes de $\mathbb{R}^n$.

On rappelle deux sous-espaces très importantes pour un opérateur $A \in L(\mathbb{R}^n, \mathbb{R}^m)$:

$$\ker(A) = \{x \in \mathbb{R}^m \ :\ Ax = 0\} \subseteq \mathbb{R}^n \qquad \textbf{Noyau de } A$$

$$\mathrm{Im}(A) = \{y \in \mathbb{R}^n \ :\ \exists x \in \mathbb{R}^n \ :\ y = Ax\} \subseteq \mathbb{R}^m \qquad \textbf{Image de } A.$$

On appelle :

— **Nullité** de A : $\dim(\ker A)$ =nul(A) ;

— **Rang** de A : $\dim(\mathrm{Im}A)$ =rank(A).

La nullité et le rang sont liés par le célèbre résultat suivant.

Théorème A.1.1 (Théorème de nullité + rang) $\forall A \in L(\mathbb{R}^n, \mathbb{R}^m)$: $\boxed{nul(A) + rank(A) = n}$.

Un opérateur linéaire $A : \mathbb{R}^n \to \mathbb{R}^m$ est injectif si $\forall x_1, x_2 \in \mathbb{R}^n$, $x_1 \neq x_2$ implique $A(x_1) \neq A(x_2)$. Il est connu [2] que cette propriété est équivalente à la condition $\ker(A) = \{0\}$, dans ce cas $\mathrm{nul}(A) = 0$ et $\mathrm{rank}(A) = n$.

Par conséquent, si A est un endomorphisme, i.e. $n = m$, alors l'injectivité de A implique que $\mathrm{rank}(A) = n$ (en Anglais on dit que A est *full rank*), i.e. la surjectivité de A, i.e. si A est un endomorphisme injectif, alors il est automatiquement bijectif, et donc inversible, i.e. il existe l'opérateur inverse de A, $A^{-1} : \mathbb{R}^n \to \mathbb{R}^n$ tel que $A^{-1}A(x) = AA^{-1}(x) = x$ pour tout $x \in \mathbb{R}^n$.

L'un des intérêts principaux de l'organisation des vecteurs lignes ou colonnes dans une matrice est la possibilité de réaliser deux opérations qui ne sont pas définies pour les vecteurs : le produit et l'inversion.

Donnés deux matrices $A \in M_{m,n}(\mathbb{R})$ et $B \in M_{n,p}(\mathbb{R})$, on peut définir la matrice produit $C = AB \in M_{m,p}(\mathbb{R})$ comme la matrice donc l'element de position (i, j) est le produit scalaire Euclidien de la ligne i de A avec la colonne j de B. Le produit matriciel est l'opération algébrique qui traduit la composition entre applications linéaires associées aux matrices.

C'est un bon exercice de vérifier les affirmations suivantes :

2. La prevue est très simple : si A est injective, alors, comme $A(0) = 0$, si $x \neq 0$, alors $Ax \neq 0$, i.e. $\ker(A) = \{0\}$, vice-versa, soit $\ker(A) = \{0\}$ et, par l'absurde, soient $x_1 \neq x_2$ tels que $A(x_1) = A(x_2)$, alors $A(x_1) - A(x_2) = 0$, i.e. par linéarité, $A(x_1 - x_2) = 0$, mais alors $x_1 - x_2 \in \ker(A)$, mais on avait supposé que $\ker(A) = \{0\}$, et alors $x_1 - x_2 = 0$, i.e. $x_1 = x_2$, ce qui est en contradiction avec l'hypothèse que $x_1 \neq x_2$.

— La ligne i de C est le produit matriciel de la ligne i de A (matrice $1 \times n$) avec la matrice B (matrice $n \times p$). Le résultat est un vecteur ligne à p composantes, $i = 1, \ldots, m$, donc on a bien une matrice $m \times p$;

— La colonne j de C est le produit matriciel de la matrice A (matrice $m \times n$) avec la colonne j de B (matrice $n \times 1$). Le résultat est un vecteur colonne à m composantes, $j = 1, \ldots, p$, donc on a bien une matrice $m \times p$.

Quand on écrira le produit de deux matrices sans spécifier leur dimensions, on supposera toujours que les dimensions sont cohérentes pour permettre la bonne définition du produit.

Le produit matriciel est associatif $(AB)C = A(BC)$, distributif à droite et à gauche, $(A_1 + A_2)B = A_1B + A_2B$ et $A(B_1 + B_2) = AB_1 + AB_2$, homogène, $\alpha(AB) = (\alpha A)B = A(\alpha B)\ \forall \alpha \in \mathbb{R}$. Néanmoins, **le produit matriciel n'est pas commutatif** : $AB \neq BA$, en général. La transposé d'une matrice produit est le produit des transposées, mais en sens inverse : $(AB)^t = B^t A^t$.

En plus, **pour le produit matriciel il ne vaut pas la loi d'élimination**, i.e. le produit de deux matrices peut être nul sans qu'au moins une des deux matrices soit nécessairement nulle, par exemple : $\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$ le produit est la matrice nulle, mais les deux matrices facteur ne sont pas nulles !

Une conséquence très importante est que, **en général, pour les matrices, l'équation $AB = AC$ n'implique pas $B = C$**, ceci est dû au fait que $AB = AC$ est équivalent à $AB - AC = 0$, i.e. par distributivité, $A(B - C) = 0$, mais, comme on l'a vu ci-dessus, cette équation n'implique pas, en général, que $B - C = 0$, i.e. que $B = C$!

Néanmoins, il existe une exception, décrite par le théorème suivant.

Théorème A.1.2 *Si* $\ker(A) = \{0\}$*, alors* $AB = AC$ *implique* $B = C$*.*

Preuve. Soit vrai que $AB = AC$ alors $AB - AC = 0$ et $A(B - C) = 0$. Il est utile d'interpréter chaque colonne fixée du produit matriciel $A(B - C)$ comme le résultat du produit matriciel de A avec une colonne fixée de $B - C$. Le fait que $A(B - C) = 0$ est traduit par le fait que toute colonne de la matrice $B - C$ appartiennent à $\ker(A)$, qui, par hypothèse, est $\{0\}$, donc toutes les colonnes de $B - C$ sont nulles, i.e. $B = C$. □

Un corollaire immédiat est le suivant.

Corollaire A.1.1 *Si A est une matrice carrée de dimension n full rank, i.e. rank$(A) = n$, alors $AB = AC$ implique $B = C$.*

Venons maintenant à l'inversion : si A est une matrice carrée de taille n, alors A est inversible si existe une matrice B carré de taille n telle que : $AB = BA = I_n$, où I_n (ou simplement I quand la spécification de la dimension n'est pas importante) est la *matrice identité* de dimension n, qui a zéro partout, sauf sur la diagonale, où elle a 1 dans chaque position. On écrit $B = A^{-1}$. L'inverse d'une matrice produit est le produit des inverses, mais en sens contraire : $(AB)^{-1} = B^{-1}A^{-1}$. Condition nécessaire et suffisante pour l'inversibilité d'une matrice carré est que son déterminant soit $\neq 0$.

Si $A \in M_{m,n}(\mathbb{R})$, alors on peut définir l'inverse droite et gauche comme ça :

— $B \in M_{n,m}(\mathbb{R})$ est l'**inverse gauche** de A si $BA = I_n$

— $B \in M_{n,m}(\mathbb{R})$ est l'**inverse droite** de A si $AB = I_m$.

En général, si $n \neq m$ l'inverse gauche et droite ne coïncident pas. Par contre, l'inverse d'une matrice carrée, si elle existe, est unique.

Allons maintenant à rappeler la relation entre rank et colonnes d'une matrice non nécessairement carrée. Pour cela, allons développer le produit Ax, avec A matrice $m \times n$ et x vecteur colonne de

dimension n :

$$Ax = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \dots & a_{mn} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} a_{11}x_1 + \ldots + a_{1n}x_n \\ \vdots \\ a_{m1}x_1 + \ldots + a_{mn}x_n \end{pmatrix} = \begin{pmatrix} a_{11} \\ \vdots \\ a_{m1} \end{pmatrix} x_1 + \ldots + \begin{pmatrix} a_{1n} \\ \vdots \\ a_{mn} \end{pmatrix} x_n,$$

$\forall j = 1, \ldots, n$, on écrit la j-ième colonne de A comme ça

$$C_j \in \mathbb{R}^m, \quad C_j = \begin{pmatrix} a_{1j} \\ \vdots \\ a_{mj} \end{pmatrix},$$

alors

$$Ax = C_1x_1 + \ldots + C_nx_n,$$

c'est-à-dire, l'image d'un vecteur Ax est combinaison linéaire des colonnes de A et les coefficients de la combinaison linéaire sont les composantes du vecteur x.

Donc [3] :

$$\mathrm{Im}(A) = \mathrm{span}(C_1, \ldots, C_n) \equiv \mathrm{Col}(A) \subseteq \mathbb{R}^m$$

et, par conséquent, $\mathrm{rank}(A)$ coïncide avec nombre de colonnes linéairement indépendantes de A.

Il est possible de démontrer que la dimension de l'espace vectoriel engendré par les vecteurs lignes de A, $\mathrm{span}(L_1, \ldots, L_m) \equiv \mathrm{Lignes}(A) \subseteq \mathbb{R}^n$, coïncide avec celle de l'espace vectoriel engendré par les vecteurs colonnes de A, donc le $\mathrm{rank}(A)$ peut être définit aussi comme le nombre de lignes linéairement indépendantes de A. Ceci implique nécessairement que :

$$\mathrm{rank}(A^t) = \mathrm{rank}(A) \leqslant \min(n, m).$$

Allons maintenant à prendre en considération la présence du produit scalaire Euclidien en $\mathbb{R}^n$.

Donné un sous-espace vectoriel $V \subseteq \mathbb{R}^n$, on appelle complément orthogonale de V le sous-espace de $\mathbb{R}^n$ définit comme ça

$$V^\perp = \{x \in \mathbb{R}^n \ : \ \langle x, y \rangle = 0 \ \forall y \in V\},$$

évidemment $V \cap V^\perp = \{0\}$. En dimension finie l'orthogonalisation est une *involution*, i.e. $V^{\perp\perp} = V$.

Théorème A.1.3 (Théorème de la projection) *Pour tout sous-espace vectoriel V de $\mathbb{R}^n$ ça vaut que :*

$$\mathbb{R}^n = V \oplus V^\perp,$$

i.e. tout vecteur $x \in \mathbb{R}^n$ peut être écrit d'une manière unique comme $x = P_V(x) + P_{V^\perp}(x)$, où $P_V(x)$ est la projection orthogonale de x sur V et $P_{V^\perp}(x)$ est la projection orthogonale de x sur $V^\perp$.

On caractérisera dans la section A.2 les opérateurs de projection.

On a déjà vu que $\mathrm{Col}(A)$ coïncide avec $\mathrm{Im}(A)$ pour toute matrice $A \in M_{m,n}(\mathbb{R})$, en utilisant le complément orthogonale, on peut montrer la relation entre $\mathrm{Lignes}(A)$ et $\ker(A)$.

Théorème A.1.4 *Pour toute matrice $A \in M_{m,n}(\mathbb{R})$, c'est vrai que :*

$$\ker(A) = \mathrm{Lignes}(A)^\perp \quad \textit{et} \quad \mathrm{Lignes}(A) = \ker(A)^\perp.$$

3. Donné un ensemble de vecteurs $(v_1, \ldots, v_n)$, on écrit avec $\mathrm{span}(v_1, \ldots, v_n)$ l'espace vectoriel engendré par ces vecteurs, i.e. l'espace vectoriel dont les éléments sont toutes les combinaisons linéaires des vecteurs $(v_1, \ldots, v_n)$.

Preuve. Soient $L_1, \dots, L_m \in \mathbb{R}^n$ les vecteurs ligne de A, alors :

$$Ax = \begin{pmatrix} L_1 \\ \vdots \\ L_m \end{pmatrix} x = \begin{pmatrix} \langle L_1, x \rangle \\ \vdots \\ \langle L_m, x \rangle \end{pmatrix},$$

donc $x \in \ker(A)$, i.e. $Ax = 0 \iff \langle L_i, x \rangle = 0 \ \forall i = 1, \dots, m \iff x$ est orthogonale à toutes les lignes de $A \iff x \perp \text{Lignes}(A)$, donc $\ker(A) = \text{Lignes}(A)^\perp$. Si on considère le complément orthogonale de cette relation on obtient : $\ker(A)^\perp = \text{Lignes}(A)^{\perp\perp} = \text{Lignes}(A)$. □

Une propriété très importante des vecteurs orthogonaux est exprimé par la généralisation du **théorème de Pythagore** suivante : si $x \perp y$, alors $\|x + y\|^2 = \|x\|^2 + \|y\|^2$.

Grâce à la présence du produit scalaire, on peut associer à chaque opérateur $A \in L(\mathbb{R}^n, \mathbb{R}^m)$ un seul opérateur $A^\dagger \in L(\mathbb{R}^m, \mathbb{R}^n)$ qui satisfait cette propriété :

$$\langle Ax, y \rangle = \langle x, A^\dagger y \rangle, \qquad \forall x, y \in \mathbb{R}^n,$$

de plus, la matrice associée à $A^\dagger$ est la transposée de la matrice associée à A (par rapport au choix des mêmes bases, bien évidemment). On appelle $A^\dagger$ l'**opérateur transposé** ou **adjoint** de l'opérateur A.

Pour tout opérateur $A \in L(\mathbb{R}^n, \mathbb{R}^m)$, $A^\dagger$ satisfait :

— $(A^\dagger)^\dagger = A$

— $\langle A^\dagger x, y \rangle = \langle x, Ay \rangle$, $\forall x, y \in \mathbb{R}^n$

— Si, en plus, A est un endomorphisme inversible, alors : $(A^{-1})^\dagger = (A^\dagger)^{-1}$.

Il existe une relation extrêmement importante entre le noyau de A et l'image de $A^\dagger$ et vice-versa, comme dit par le théorème suivant.

Théorème A.1.5 *Pour tout opérateur $A \in L(\mathbb{R}^n, \mathbb{R}^m)$ ça vaut que :*

$$\boxed{\ker(A^\dagger) = (\text{Im}(A))^\perp} \quad \textit{et} \quad \boxed{\text{Im}(A) = (\ker(A^\dagger))^\perp}.$$

Preuve. Le théorème est un corollaire immédiat du théorème A.1.4, en fait, pour toute matrice $A \in M_{m,n}(\mathbb{R})$, $\text{Col}(A) = \text{Lignes}(A^t)$ et $\text{Lignes}(A) = \text{Col}(A^t)$, comme Col $(A) = \text{Im}(A)$, les formules de ce théorèmes sont une simple réécriture de celles du théorème A.1.4. Néanmoins, on veut proposer une preuve alternative, qui a l'avantage de pouvoir être étendue sans difficulté à la dimension infinie, différemment de la précédente.

$\underline{\ker(A^\dagger) \subseteq (\text{Im}(A))^\perp}$: soit $x \in \ker(A^\dagger)$, i.e. $A^\dagger x = 0$, alors $\forall y \in \mathbb{R}^n$:

$$0 = \langle 0, y \rangle = \langle A^\dagger x, y \rangle = \langle x, Ay \rangle,$$

i.e. x est orthogonale à tous les éléments de l'image de A, ce qui prouve que $\ker(A^\dagger) \subseteq (\text{Im}(A))^\perp$.

$\underline{(\text{Im}(A))^\perp \subseteq \ker(A^\dagger)}$: soit $y \in (\text{Im}(A))^\perp$, i.e. $\langle y, Ax \rangle = 0 \ \forall x \in \mathbb{R}^n$, mais alors $\langle A^\dagger y, x \rangle = 0 \ \forall x \in \mathbb{R}^n$, ce qui est possible seulement si $A^\dagger y = 0$ car le seul vecteur orthogonal à tous les autres est le vecteur nul, mais alors $y \in \ker(A^\dagger)$.

La deuxième formule descend de la première tout simplement en considérant le complément orthogonale aux deux côtés et en rappelant que le biorthogonale d'un sous-espace de dimension finie est le sous-espace même. □

La composition entre A et son adjoint $A^\dagger$ génère deux opérateurs très importantes dans la théorie de l'optimisation : $A^\dagger A$ et $AA^\dagger$. Tout d'abord, on observe que si $A \in L(\mathbb{R}^n, \mathbb{R}^m)$, la matrice associée à A est $m \times n$, alors $AA^\dagger$ est associé à une matrice carrée $m \times m$ et $A^\dagger A$ est associé à une matrice carrée $n \times n$. Le fait de travailler avec des endomorphismes associés à des matrices carrées à déjà un avantage évitent : la possibilité d'examiner l'inversion de ces opérateurs et de ces matrices.

Allons examiner d'abord les propriétés de $A^\dagger A$ via l'analyse de sa matrice associée $A^t A$.

Tout d'abord : $A^t A$ est une matrice **symétrique**, en fait :

$$(A^t A)^t \underset{(MN)^t = N^t M^t}{=} A^t A^{tt} = A^t A.$$

La même propriété vaut pour AA^t. Toute matrice symétrique a des valeurs propres réels grâce à un résultat standard de l'algèbre linéaire.

On rappelle un concept important.

Déf. A.1.1 *$M \in M(n, \mathbb{R})$ est dite **semi-définie positive** si M est symétrique et si :*

$$\langle Mx, x \rangle \geqslant 0 \qquad \forall x \in \mathbb{R}^n.$$

*Si ça vaut l'inégalité stricte $\forall x \in \mathbb{R}^n \backslash \{0\}$, alors M est dite **définie positive**.*

Théorème A.1.6 *Soit $M \in M(n, \mathbb{R})$ une matrice symétrique. Alors, les affirmations suivantes sont équivalentes :*

1. *M est semi-définie positive ;*
2. *Toutes les valeurs propres de M sont $\geqslant 0$;*
3. *$M = N^t N$ pour une matrice opportune $N \in M(n, \mathbb{R})$.*

Preuve.

$\boxed{1) \Rightarrow 2)}$: soit $\lambda \in \mathbb{R}$ un valeur propre de M, alors :

$$0 \leqslant \langle Ax, x \rangle = \langle \lambda x, x \rangle = \lambda \langle x, x \rangle = \lambda \|x\|^2,$$

ce qui implique $\lambda \geqslant 0$.

$\boxed{2) \Rightarrow 3)}$: Comme M est réelle et symétrique, elle est diagonalisable, i.e. il existe une matrice orthogonale P, i.e. $P^{-1} = P^t$, telle que : $PMP^t = D$, où $D = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$, avec $\lambda_i \geqslant 0$: i-ème valeur propre de M.

Vu que les éléments de la diagonale de D sont $\geqslant 0$ par hypothèse, on peut écrire $D = C^2$, où $C = \mathrm{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n})$.

Mais alors : $PMP^t = D$ peut être réécrit comme $M = P^t DP = P^t CCP$, si on écrit $N = CP$, alors $N^t = P^t C^t = P^t C$ vue que la transposée d'une matrice diagonale est elle-même, donc $M = N^t N$.

$\boxed{3) \Rightarrow 1)}$: pour tout $x \in \mathbb{R}^n$,

$$\langle Mx, x \rangle = \langle N^t Nx, x \rangle = \langle Nx, Nx \rangle = \|Nx\|^2 \geqslant 0,$$

i.e. M est semi-définie positive. □

On laisse la (simple) preuve de ce théorème au lecteur.

Corollaire A.1.2 *Soit $M \in M(n, \mathbb{R})$ une matrice symétrique. Alors, M est définie positive $\iff$ toutes les valeurs propres de M sont positives.*

Reconsidérons la matrice A^tA et allons à examiner ses propriétés.

Théorème A.1.7 *Pour toute matrice $A \in M(n, \mathbb{R})$, la matrice A^tA est semi-définie positive.*

Preuve. On a déjà vu que A^tA est symétrique, de plus, $\forall x \in \mathbb{R}^n$:

$$\langle A^tAx, x\rangle = \langle Ax, Ax\rangle = \|Ax\|^2 \geqslant 0.$$

□

En tant que matrice semi-définie positive, les valeurs propres de A^tA sont $\geqslant 0$. Ceci implique qu'on peut calculer leurs racines carrées.

Déf. A.1.2 *Le valeurs singulières d'une matrice $A \in M(n, \mathbb{R})$ sont les racines carrées des valeurs propres de la matrice A^tA :*

$$\{\sqrt{\lambda},\ \lambda :\ \textit{valeur propre de } A^tA\} = \ \boldsymbol{\textit{Valeurs singulières de } A}.$$

Les valeurs singulières de A seront utilisées pour introduire la technique de décomposition en valeurs singulières : SVD.

Théorème A.1.8 *Soit $A \in M_{m,n}(\mathbb{R})$ quelconque, alors* $\boxed{\ker(A^tA) = \ker(A)}$.

Preuve.

$\underline{\ker(A) \subseteq \ker(A^tA)}$: soit $x \in \ker(A)$, alors $A^t(Ax) = A^t0 = 0$, donc $x \in \ker(A^tA)$.

$\underline{\ker(A^tA) \subseteq \ker(A)}$: soit $x \in \ker(A^tA)$, alors $A^t(Ax) = 0$, i.e. $Ax \in \ker(A^t)$, mais, grâce au théorème A.1.5, ceci est équivalent à $Ax \in (\mathrm{Im}(A))^\perp$, mais Ax est, par définition, un élément de $\mathrm{Im}(A)$, donc $Ax \in \mathrm{Im}(A) \cap (\mathrm{Im}(A))^\perp = \{0\}$, c'est-à-dire $Ax = 0$ et donc $x \in \ker(A)$. □

Théorème A.1.9 *Soit $A \in M_{m,n}(\mathbb{R})$ quelconque, alors* $\boxed{\mathrm{Im}(A^tA) = \mathrm{Im}(A^t)}$.

Preuve. $\mathrm{Im}(A^t) = \ker(A)^\perp = \ker(A^tA)^\perp \underset{(A.1.4)}{=} \mathrm{Lignes}(A^tA)$, donc $\mathrm{Im}(A^t) = \mathrm{Lignes}(A^tA)$, mais A^tA est symétrique, donc $\mathrm{Lignes}(A^tA) = \mathrm{Col}(A^tA) = \mathrm{Im}(A^tA)$ et alors $\mathrm{Im}(A^t) = \mathrm{Im}(A^tA)$. □

Corollaire A.1.3 *Soit $A \in M_{m,n}(\mathbb{R})$ quelconque, alors* $\boxed{\mathrm{rank}(A^tA) = \mathrm{rank}(A)}$.

Preuve. La thèse suit du fait que $\mathrm{rank}(A^t) = \mathrm{rank}(A)$ et que $\mathrm{rank}(A) = \dim \mathrm{Im}(A)$. □

La propriété $\ker(A) = \ker(A^tA)$ et le fait que A^tA soit un endomorphisme permettent de *caractériser l'inversibilité de A^tA avec une condition sur A* :

$$\boxed{A^tA \text{ inversible } \iff A \text{ est full rank } \iff \ker(A) = \{0\}}.$$

En plus, dans ce cas, A^tA est **définie positive**, i.e. $\langle A^tAx, x\rangle > 0\ \forall x \neq 0$, en fait, comme vu avant, A^tA est toujours semi-définie positive, car $\langle A^tAx, x\rangle = \langle Ax, Ax\rangle = \|Ax\|^2$, mais $\|Ax\| = 0 \iff Ax = 0 \iff x = 0$ car $\ker(A) = \{0\}$. Comme vu avant, ceci est équivalent au fait que, **quand A^tA est inversible**, A^tA **a seulement des valeurs propres strictement positives**.

A.2 Projecteurs

Les opérateurs de projection, ou projecteurs, jouent un rôle fondamentale dans pratiquement toutes les disciplines des mathématiques pures et appliquées, dans le cours on aura l'occasion d'apprécier leur importance aussi dans le champ de l'optimisation. Il est donc essentiel un rappel de ces opérateurs.

On commence par rappeler qu'une famille de n vecteurs orthogonaux non nuls de $\mathbb{R}^n$ est dite **base orthogonale** de $\mathbb{R}^n$. Si, en plus, la famille est orthonormée, i.e. chaque vecteur a norme unitaire, alors on l'appelle **base orthonormée** de $\mathbb{R}^n$. Une base orthonormée $(u_1, \ldots, u_n)$ peut être caractérisée via la relation suivante :

$$\langle u_i, u_j \rangle = \delta_{i,j} = \begin{cases} 1 & \text{si } i = j \\ 0 & \text{si } i \neq j \end{cases},$$

$\delta_{i,j}$ est dit *symbole de Kronecker*.

On rappelle que, pour déterminer les composantes d'un vecteur par rapport à une base quelconque on doit résoudre un système linéaire de n équations avec n inconnues. Par contre, si on a une base orthogonale ou orthonormale, les composantes sont déterminées par des produits scalaires, en fait on peut démontrer que, si $B = (u_1, \ldots, u_n)$ une base **orthogonale** de $\mathbb{R}^n$, alors :

$$\boxed{v = \sum_{i=1}^{n} \frac{\langle v, u_i \rangle}{\|u_i\|^2} u_i},$$

en particulier, si B est une base **orthonormée**, alors :

$$\boxed{v = \sum_{i=1}^{n} \langle v, u_i \rangle u_i}.$$

On observe que la résolution d'un système linéaire de n équations avec n inconnues nécessite, en général, beaucoup plus d'opérations que le calcul de produits scalaires, ceci montre un des avantages de connaître une base orthogonale de $\mathbb{R}^n$.

Interprétation géométrique du théorème : le théorème qu'on vient de démontrer est la généralisation du théorème de décomposition d'un vecteur dans le plan $\mathbb{R}^2$ ou dans l'espace $\mathbb{R}^3$ sur la base canonique des vecteurs unitaires des axes. Pour simplifier, on considère le cas de $\mathbb{R}^2$ comme dans la figure A.1.

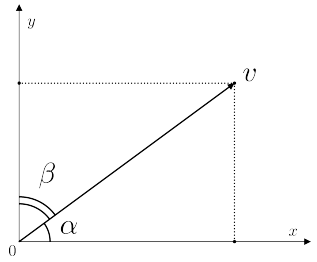


FIGURE A.1 – Représentation graphique du théorème de décomposition sur la base canonique en $\mathbb{R}^2$.

Si $\hat{\imath}$ et $\hat{\jmath}$ sont respectivement les vecteurs unitaires des axes x et y, alors le théorème de décomposition dit que :

$$v = \underbrace{\|v\| \cos\alpha}_{\langle v, \hat{\imath} \rangle} \hat{\imath} + \underbrace{\|v\| \cos\beta}_{\langle v, \hat{\jmath} \rangle} \hat{\jmath} = \langle v, \hat{\imath} \rangle \hat{\imath} + \langle v, \hat{\jmath} \rangle \hat{\jmath},$$

qui est un cas particulier du théorème ci-dessus.

Dans l'espace Euclidien $\mathbb{R}^2$, il est clair que le produit scalaire d'un vecteur v avec un vecteur unitaire u réalise la projection orthogonale de v dans la direction donnée par u.

De la même manière, on peut définir la projection orthogonale p d'un vecteur de $\mathbb{R}^3$ sur le plan engendré par deux vecteurs unitaires comme la somme des projections orthogonales p_1 et p_2 sur les deux vecteurs unitaires considérés séparément, comme il est montré dans la figure A.2.

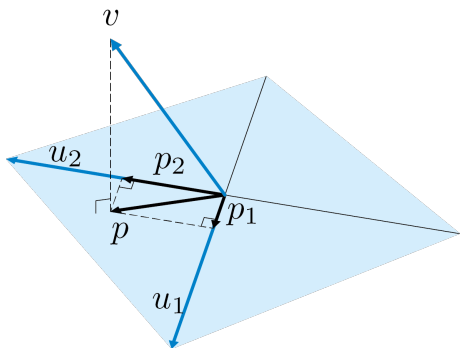


FIGURE A.2 – Projection orthogonale p d'un vecteur de $\mathbb{R}^3$ sur le plan engendré par deux vecteurs unitaires.

On considère maintenant une famille orthogonale $F = \{u_1, \ldots, u_m\}$, $m \leqslant n$, de vecteurs non nuls : $u_i \neq 0 \ \forall i = 1, \ldots, m$. On étend d'une façon naturelle la définition de la **projection orthogonale d'un vecteur** $v \in V$ **sur** $S =$**span**(F) comme ceci :

$$\boxed{P_S(v) = \sum_{i=1}^{m} \frac{\langle v, u_i \rangle}{\|u_i\|^2} u_i},$$

il faut noter que la présence de la norme au carré est due au fait qu'il faut *normaliser deux fois* u_i. On définit l'**opérateur de projection orthogonale** ou **projecteur** sur S comme l'application (évidemment) linéaire :

$$\begin{array}{rccl} P_S : & \mathbb{R}^n & \longrightarrow & S \subseteq V \\ & v & \longmapsto & P_S(v) = \sum_{i=1}^{m} \frac{\langle v, u_i \rangle}{\|u_i\|^2} u_i, \end{array}$$

$P_S v$ est une combinaison linéaire des vecteurs $u_1, \ldots, u_m$. Le théorème suivant montre que la projection orthogonale définie ci-dessus a toutes les propriétés de la projection orthogonale en $\mathbb{R}^2$ et $\mathbb{R}^3$.

Théorème A.2.1 *Avec les notations ci-dessus :*

1) *Si $s \in S$ alors $P_S(s) = s$, i.e. l'action de P_S sur les vecteurs de S est l'identité. Par conséquent $P_S^2 = id_S$;*

2) *$\forall v \in \mathbb{R}^n$ et $s \in S$, le vecteur* ***résidu*** *de la projection, i.e. $v - P_S(v)$, est $\perp$ à S :*

$$\langle v - P_S(v), s \rangle = 0 \quad \Longleftrightarrow \quad v - P_S(v) \perp s \ ;$$

3) *$\forall v \in \mathbb{R}^n$ et $s \in S$:*

$$\|v - P_S(v)\| \leqslant \|v - s\|$$

et l'égalité vaut si et seulement si $s = P_S(v)$.

Observation importante : la propriété 3) dit que, **parmi tous les vecteurs de S, le vecteur qui minimise la fonction distance à v, i.e. $d : \mathbb{R}^n \times \mathbb{R}^n \to [0, +\infty[$, $d(v, s) = \|v - s\|$, est la projection orthogonale** $P_S(v)$: $P_S(v) = \text{argmin}_{s \in S}\ d(v, s)$.

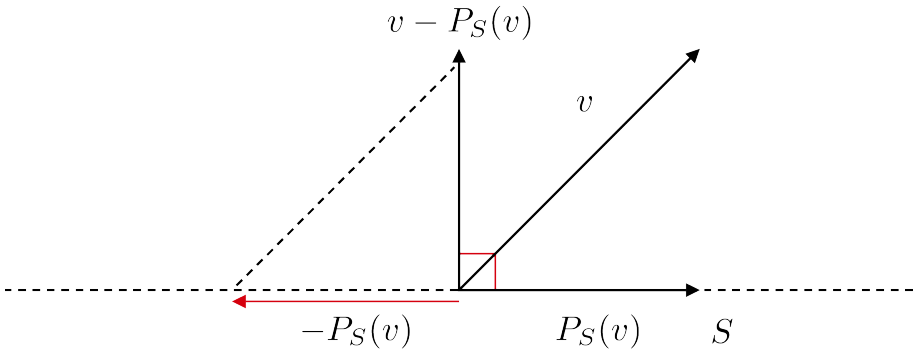


FIGURE A.3 – Visualisation de la propriété 2) en $\mathbb{R}^2$.

Par ailleurs, la propriété 2) est la généralisation d'un fait géométrique qu'on peut visualiser très facilement en $\mathbb{R}^2$, comme dans la figure A.3.

Preuve de 1) : Soit $s \in S$, i.e. $s = \sum_{j=1}^{m} \alpha_j u_j$, alors :

$$P_S(s) = \sum_{i=1}^{m} \frac{\langle \sum_{j=1}^{m} \alpha_j u_j, u_i \rangle}{\|u_i\|^2} u_i \underset{(u_i \perp u_j \ \forall i \neq j)}{=} \sum_{i=1}^{m} \frac{\alpha_i \langle u_i, u_i \rangle}{\|u_i\|^2} u_i = \sum_{i=1}^{m} \alpha_i u_i = s.$$

Par conséquent, $\forall v \in \mathbb{R}^n$, $P_S^2(v) = P_S(P_S(v)) = P_S(v)$ car $P_S(v) \in S$, donc $P_S^2 = id_S$.

Preuve de 2) : On commence par considérer encore le produit scalaire de $P_S(v)$ avec un vecteur quelconque u_j, $j \in \{1, \ldots, m\}$ fixé :

$$\langle P_S(v), u_j \rangle = \sum_{i=1}^{m} \frac{\langle v, u_i \rangle}{\|u_i\|^2} \langle u_i, u_j \rangle \underset{(u_i \perp u_j \ \forall i \neq j)}{=} \frac{\langle v, u_j \rangle}{\|u_j\|^2} \langle u_j, u_j \rangle = \langle v, u_j \rangle$$

d'où

$$\langle v, u_j \rangle - \langle P_S(v), u_j \rangle = 0 \underset{\text{linéarité de } \langle\ ,\ \rangle}{\iff} \langle v - P_S(v), u_j \rangle = 0 \quad \forall j \in \{1, ..., m\}.$$

Maintenant, si $s \in S$, alors $\exists\ \alpha_1, \ldots, \alpha_m$ tels que $s = \sum_{j=1}^{m} \alpha_j u_j$, donc

$$\langle v - P_S(v), s \rangle = \langle v - P_S(v), \sum_{j=1}^{m} \alpha_j u_j \rangle = \sum_{j=1}^{m} \alpha_j \underbrace{\langle v - P_S(v), u_j \rangle}_{=0} = 0,$$

et la propriété 2) est prouvée.

Preuve de 3) : il est utile d'écrire la différence $v - s$ comme ceci : $v - P_S(v) + P_S(v) - s$. La propriété 2) nous dit que $v - P_S(v) \perp S$, par contre, $P_S(v), s \in S$ donc $P_S(v) - s \in S$. Par conséquent : $(v - P_S(v)) \perp (P_S(v) - s)$.
En utilisant la généralisation du théorème de Pythagore on peut alors écrire :

$$\|v - s\|^2 = \|v - P_S(v) + P_S(v) - s\|^2 = \|v - P_S(v)\|^2 + \underbrace{\|P_S(v) - s\|^2}_{\geqslant 0} \geqslant \|v - P_S(v)\|^2,$$

ce qui implique : $\|v - s\| \geqslant \|v - P_S(v)\|\ \forall v \in V,\ s \in S$.

Bien évidemment, $\|P_S(v) - s\|^2 = 0$ si et seulement si $s = P_S(v)$, et dans ce cas on a $\|v - s\|^2 = \|v - P_S(v)\|^2$. □

Comme conséquence du théorème qu'on vient de prouver, on peut dire que la formule $v = v - s + s$ pour tout $v \in V$ et $s \in S$ cache une information importante : $v - s$ est orthogonale à s.

Terminons en montrant comment on peut *réaliser les opérateur de projection sous forme matricielle.* On commence avec la projection sur un seul axe donné par le vecteur $u \in \mathbb{R}^n$ et après on va généraliser le discours.

Soit P_u le projecteur orthogonale sur l'axe $u \in \mathbb{R}^n$, alors, si on l'applique à n'importe quel vecteur $v \in \mathbb{R}^n$, on obtient un multiple de u, i.e. $P_u v = \alpha u$, avec $\alpha \in \mathbb{R}$. Allons maintenant utiliser le fait que le vecteur résidu $v - P_u v$ est orthogonal à u :

$$0 = \langle u, v - P_u v \rangle = u^t(v - P_u v) = u^t(v - \alpha u) = u^t v - \alpha u^t u,$$

donc $u^t v = \alpha u^t u$, ce qui permet de caractériser α comme ça : $\alpha = u^t v / u^t u$. On utilise cette information pour déterminer P_u :

$$P_u v = \alpha u \underset{\alpha \in \mathbb{R}}{=} u\alpha = u\frac{u^t v}{u^t u} = \frac{1}{u^t u} u(u^t v).$$

Par l'associativité du produit matriciel, ça vaut que $u(u^t v) = (uu^t)v$, où uu^t est à interpréter comme la matrice $n \times n$ obtenue comme produit matricielle de u vu comme une matrice colonne $n \times 1$ et de u^t vu comme une matrice ligne $1 \times n$. Donc on peut écrire : $P_u v = \frac{uu^t}{u^t u} v$, pour tout vecteur v, i.e.

$$\boxed{P_u = \frac{uu^t}{u^t u}} \qquad \|u\| \neq 1,$$

en particulier, si $u^t u = \|u\|^2 = 1$, i.e. si u est un vecteur unitaire, alors :

$$\boxed{P_u = uu^t} \qquad \|u\| = 1.$$

Considérons maintenant un sous-espace S de $\mathbb{R}^n$ engendré par une famille de vecteurs non nuls et orthogonaux entre eux qu'on indique avec $(u_1, \ldots, u_m)$, $m < n$. On a vu que la projection orthogonale sur S d'un vecteur $v \in \mathbb{R}^n$, dans ce cas, est un vecteur de S, i.e. il est obtenu par combinaison linéaire des vecteurs $u_1, \ldots, u_m$. Si on place les coefficients $\alpha_1, \ldots, \alpha_m$ de la combinaison linéaire dans un vecteur colonne et **on utilise les générateurs de S comme colonnes d'une matrice** A de type $n \times m$, i.e.

$$A = \begin{pmatrix} u_1^1 & & u_m^1 \\ \vdots & \vdots & \vdots \\ u_1^n & & u_m^n \end{pmatrix},$$

alors on peut réécrire la projection orthogonale comme ça :

$$P_S v = \underbrace{\begin{pmatrix} u_1^1 & & u_m^1 \\ \vdots & \vdots & \vdots \\ u_1^n & & u_m^n \end{pmatrix}}_{n \times m} \underbrace{\begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_m \end{pmatrix}}_{m \times 1} \equiv \underbrace{A\vec{\alpha}}_{n \times 1} = \begin{pmatrix} \alpha_1 u_1^1 + \ldots + \alpha_m u_m^1 \\ \vdots \\ \alpha_1 u_1^n + \ldots + \alpha_m u_m^n \end{pmatrix} = \alpha_1 u_1 + \ldots + \alpha_m u_m, \quad \forall v \in \mathbb{R}^n.$$

La formule montre que $P_S v \in \mathrm{Im}(A)$, par contre on sait que le vecteur résidu $v - P_S v \in (\mathrm{Im}(A))^\perp = \ker(A^t)$, donc :

$$0 = A^t(v - P_S v) = A^t(v - A\vec{\alpha}) = A^t v - A^t A\vec{\alpha},$$

i.e. $A^tA\vec{\alpha} = A^tv$. Allons se concentrer sur A^tA, qui est une matrice $m \times m$ écrite comme ça :

$$A^tA = \begin{pmatrix} u_1^1 & & u_1^n \\ \vdots & \vdots & \vdots \\ u_m^1 & & u_m^n \end{pmatrix} \begin{pmatrix} u_1^1 & & u_m^1 \\ \vdots & \vdots & \vdots \\ u_1^n & & u_m^n \end{pmatrix} = \operatorname{diag}(\|u_1\|^2, \ldots, \|u_m\|^2),$$

vu que les vecteurs $u_1, \ldots, u_m$ sont orthogonaux entre eux et donc tous les éléments hors de la diagonale sont nuls car obtenus via le produit scalaire de vecteurs orthogonaux ! Sur la diagonal on trouve la norme au carré des vecteurs $u_1, \ldots, u_m$, toutes ces normes sont strictement positives, car nous avons supposé que ces vecteurs sont non nuls. Donc A^tA est une matrice inversible et alors nous pouvons obtenir $\vec{\alpha}$ comme ça :

$$\vec{\alpha} = (A^tA)^{-1}A^tv$$

et, comme $P_Sv = A\vec{\alpha} = A(A^tA)^{-1}A^tv$ pour tout vecteur $v \in \mathbb{R}^n$, on obtient la représentation matricielle de P_S :

$$\boxed{P_S = A(A^tA)^{-1}A^t} \qquad \text{colonnes de } A \text{ : générateurs } \textbf{orthogonaux} \text{ de } S.$$

Il faut observer que les matrices A et A^t ne sont pas carrés, donc il ne faut pas tomber dans la tentation de simplifier la formule précédente comme ceci : $AA^{-1}(A^t)^{-1}A^t = I$! On observe aussi que la matrice $(A^tA)^{-1}$ joue un rôle analogue à celui du facteur $(u^tu)^{-1}$ dans le cas mono-dimensionnel, en fait :

$$P_u = u(u^tu)^{-1}u^t \qquad \text{vs.} \qquad P_S = A(A^tA)^{-1}A^t.$$

Si u était unitaire, le facteur u^tu pouvait être simplifié, dans ce cas, si $u_1, \ldots, u_m$ est une famille orthonormale, alors $A^tA = \operatorname{diag}(1, \ldots, 1) = I_m$ et donc la formule pour le projecteur orthogonale devient :

$$\boxed{P_S = AA^t} \qquad \text{colonnes de } A \text{ : générateurs } \textbf{orthonormés} \text{ de } S.$$

Il est possible de démontrer que les projecteurs orthogonaux P en $\mathbb{R}^n$ peuvent être caractérisés par la chaine d'égalités suivante :

$$P = P^2 = P^t,$$

allons vérifier que le projecteur $A(A^tA)^{-1}A^t$ possède ces propriétés :

$$(A(A^tA)^{-1}A^t)^t = A^{tt}((A^tA)^{-1})^tA^t = A((A^tA)^t)^{-1}A^t = A(A^tA^{tt})^{-1}A^t = A(A^tA)^{-1}A^t,$$

et

$$(A(A^tA)^{-1}A^t)^2 = (A(A^tA)^{-1}A^t)(A(A^tA)^{-1}A^t) = A(A^tA)^{-1}(A^tA)(A^tA)^{-1}A^t = A(A^tA)^{-1}A^t.$$

La vérification de ces propriétés pour AA^t est encore plus facile et elle est laissée par exercice.

Annexe B

Un très bref rappel sur les espaces métriques et le calcul différentiel en $\mathbb{R}^n$

Dans cette deuxième appendice on va présenter un court résumé des concepts fondamentales des espaces métriques et du calcul différentiel en $\mathbb{R}^n$. Malgré le fait que les résultats et les algorithmes présentés dans le cours sont développés pour $\mathbb{R}^n$ et ses parties, on a néanmoins voulu présenter quelque détail de la théorie des espaces métriques vu leur grande utilité dans l'optimisation.

B.1 Espaces métriques

En $\mathbb{R}^n$ on peut mesurer la distance entre deux points $x, y \in \mathbb{R}^n$, par exemple, grâce à la norme Euclidienne : $d(x, y) = \|x - y\|$. Dans les applications des mathématiques on peut devoir traiter des espaces plus compliqués que $\mathbb{R}^n$, et même de dimension infinie, il est donc essentiel d'abstraire et de généraliser la notion de distance, la bonne définition est la suivante.

Déf. B.1.1 *Soit X un ensemble quelconque et $d : X \times X \to [0, +\infty)$. d est une* ***distance*** *ou* ***métrique*** *sur X si :*

1. $\forall x, y \in X,\ d(x, y) \geqslant 0$ *et* $d(x, y) = 0 \Leftrightarrow x = y$ *(positivité)*
2. $\forall x, y \in X, d(x, y) = d(y, x)$ *(symétrie)*
3. $\forall x, y, z \in X, d(x, z) \leqslant d(x, y) + d(y, z)$ *(inégalité triangulaire)*

La couple (X, d) est dite un ***espace métrique****.*

Une ***suite***[1] ***à valeurs en*** *$E \subseteq (X, d)$ est une fonction $\varphi : \mathbb{N} \to X$, $n \mapsto \varphi(n) = x_n$. Souvent on identifie la suite avec son codomaine, dans ce cas on écrira $(x_n)_{n \in N} \subseteq E$. L'ensemble $\mathbb{N}$ peut être remplacé par un autre ensemble dénombrable. Par exemple si on remplace $\mathbb{N}$ avec $\mathbb{Z}$ on parle de suites bilatérales.*

Une suite $(x_n)_{n \in \mathbb{N}} \subset X$ est convergente à la limite $L \in X$ si :

$$\forall \varepsilon > 0 \ \exists N_\varepsilon \in \mathbb{N} \text{ tel que } n > N_\varepsilon \implies d(x_n, L) < \varepsilon.$$

On écrit $x_n \underset{n \to +\infty}{\longrightarrow} L$, ou $L = \lim\limits_{n \to +\infty} x_n$.

1. Les suites émergent d'une manière naturelle en optimisation quand on considère les algorithmes itératifs pour approcher la solution à un problème trop compliqué pour être résolu d'une manière analytique.

Étant donné un point $x_0 \in (X, d)$ quelconque, on appelle ***voisinage de centre*** x_0 ***et rayon*** $r \in \mathbb{R}$, $r > 0$ *l'ensemble des points de X qui ont une distance de x_0 inférieure à r, i.e.*

$$U_r(x_0) = \{x \in X \ : \ d(x, x_0) < r\},$$

une autre notation habituelle qu'on trouve dans les livres est $B_r(x_0)$. Si d est la distance Euclidienne et $X = \mathbb{R}^n$, on appelle le voisinage une « boule ».

Le concept de limite est intimement lié avec le concept défini ci dessous.

Déf. B.1.2 *On dit que $x_0 \in (X, d)$ est un* ***point d'adhérence****, ou un point d'****accumulation****, ou un point* ***limite*** *pour une partie $E \subseteq X$ si pour tout $r > 0$, il existe un voisinage $U_r(x_0)$ qui contient des éléments de E différents de x_0.*

Interprétation de la définition : si on fait un « zoom » autour d'un point d'adhérence, avec n'importe quel niveau d'agrandissement, on voit toujours de points de E différents de x_0.

Déf. B.1.3 *On dit que $E \subset (X, d)$ est une partie fermé si E contient tous ses points d'adhérence.*

Un exemple de partie fermé en $\mathbb{R}$ est n'importe quel intervalle $[a, b]$, $a, b \in \mathbb{R}$, $a < b$, mais $[a, b[$ n'est pas fermé.

Déf. B.1.4 *On appelle* ***fermeture*** *de $E \subset (X, d)$ l'intersection de toutes les parties fermés qui contiennent E, ou, d'une manière équivalente, le plus petit sous-ensemble fermé de X qui contient E.*

Pour les intérêts de l'optimisation, il est utile de caractériser la fermeture d'une partie et les points d'adhérence à travers des suites, allons revoir les concepts et les résultats qui permettent de formaliser cette affirmation.

Déf. B.1.5 *Considérons une suite croissante ψ à valeurs en $\mathbb{N}$, i.e. $\psi : \mathbb{N} \to \mathbb{N}$, $k \mapsto \psi(k) = n_k$, $n_k < n_{k+1}$ $\forall k \in \mathbb{N}$. On appelle* ***suite extraite*** *(ou sous-suite) de la suite φ la suite composée $\varphi \circ \psi : \mathbb{N} \to \mathbb{N} \to E$, $k \mapsto n_k \mapsto x_{n_k}$. Comme pour les suites, souvent on identifie la suite extraite avec son codomaine $(x_{n_k})_{k \in \mathbb{N}}$.*

Déf. B.1.6 *Une suite $(x_n)_{n \in \mathbb{N}} \subseteq E$ est bornée si $\exists x_0 \in E$ et $r > 0$ tels que : $\{x_n \ n \in \mathbb{N}\} \subseteq U_r(x_0)$.*

Interprétation : les éléments d'une suite bornée sont tous contenus dans le voisinage d'un élément de E si on choisit un rayon suffisamment grande, mais fini !

On observe que **toute suite convergente est bornée**, en fait, par définition, si L est la limite de la suite, alors $\forall \varepsilon > 0$ et $\forall n \geqslant N_\varepsilon$: $\{x_n \ n \in \mathbb{N}\} \subseteq U_\varepsilon(L)$. Considérons

$$\tilde{r} = \max\{d(x_n, L), \ n = 0, 1, \ldots, N_\varepsilon - 1\},$$

alors, si on définit $r = \max(\varepsilon, \tilde{r})$, il est clair que $\{x_n \ n \in \mathbb{N}\} \subseteq U_r(L)$.

Le théorème suivant dit que les points d'adhérence sont comme des aimant pour les suites...

Théorème B.1.1 *Soit $L \in X$ un point d'adhérence pour $E \subset (X, d)$, alors il existe une suite $(x_n)_{n \in \mathbb{N}} \subseteq E$ qui converge vers L.*

Autrement dit : tout point d'adhérence de E est la limite d'une suite à valeurs en E.

On montre la preuve de ce théorème car elle permet de comprendre, via un argument classique et très élégant, la signification de la définition de point d'adhérence, qui peut être un peu obscure au tout début.

Preuve. La démonstration est constructive, i.e. on va construire la suite qui converge vers L. L'idée à la base de la preuve consiste en utiliser la propriété définitoire de point d'adhérence, i.e. le fait qu'on peut toujours trouver au moins un point de E différent de L en chaque voisinage de L : considérons la suite de rayons $r = 1, 1/2, \ldots, 1/n \ldots$, alors, pour tout $n \in \mathbb{N}$, il existe $x_n \neq L$ tel que $d(x_n, L) < 1/n$.

On va montrer que ceci implique la convergence de $(x_n)_{n\in\mathbb{N}}$ vers L : pour tout $\varepsilon > 0$ fixé, on va définir [2] :

$$N_\varepsilon = \left\lfloor \frac{1}{\varepsilon} \right\rfloor + 1 > \frac{1}{\varepsilon} \implies \frac{1}{N_\varepsilon} < \frac{1}{\frac{1}{\varepsilon}} = \varepsilon$$

alors, pour tout $n \geqslant N_\varepsilon$ ça vaut que $\frac{1}{n} \leqslant \frac{1}{N_\varepsilon} < \varepsilon$ et donc, en résumé, on a démontré que :

$$\forall \varepsilon > 0 \; \exists N_\varepsilon > 0 \; : \; n \geqslant N_\varepsilon \implies d(x_n, L) < \frac{1}{n} < \varepsilon,$$

qui est la définition de convergence de $(x_n)_{n\in\mathbb{N}}$ vers L. □

Ce dernier théorème nous donne la possibilité de caractériser l'adhérence d'un espace métrique : quand on écrit $X = \overline{E}$ ça veut dire que pour tout $x \in X$ il existe une suite $(x_n)_{n\in\mathbb{N}} \subseteq E$ telle que $x = \lim_{n\to+\infty} x_n$. On dit aussi que E **est dense en** X.

B.1.1 Le théorème de Bolzano-Weierstrass

On va examiner ici la relation entre limites de suites et des suites extraites. Tout d'abord, on observe qu'une suite non-convergente peut admettre des suites extraites convergentes, l'exemple le plus simple est probablement le suivant :

$(x_n) = (-1)^n$ qui n'est pas convergente, mais qui admet la suite extraite

$(x_{2n}) = (-1)^{2n} \equiv 1$ qui converge à 1 en tant que suite constante.

Néanmoins, si une suite est convergente vers une limite L, alors toutes ses suites extraites sont obligées à être convergentes vers la même limite L. La preuve de cette affirmation est immédiate : si $L = \lim_{n\to+\infty} x_n$ alors $\forall \varepsilon > 0 \; \exists N_\varepsilon > 0 \; : \; n \geqslant N_\varepsilon \implies d(x_n, L) < \varepsilon$, mais alors, comme $(n_k)_{k\in\mathbb{N}}$ est une suite croissante, il existe un $K_\varepsilon > 0$ tel que $n_{K_\varepsilon} \geqslant N_\varepsilon$ et alors, pour tout $k \geqslant n_{K_\varepsilon}$ ça vaut que $d(x_{n_k}, L) < \varepsilon$, i.e. $L = \lim_{k\to+\infty} x_{n_k}$.

Donc, en résumé, si une suite est convergente, alors toutes ses suites extraites sont convergentes à la même limite, si elle n'est pas convergente, alors elle *peut* avoir des suites extraites convergentes. Le théorème suivant dit que, si le codomaine d'une suite admet un point d'adhérence, alors la suite *doit* avoir une suite extraite qui converge vers ce même point.

Théorème B.1.2 *Soit $(x_n)_{n\in\mathbb{N}}$ une suite à valeurs en (X, d). Si la partie $E = \{x_n, \; n \in \mathbb{N}\} \subseteq X$ admet un point d'adhérence L, alors il existe une suite extraite de $(x_n)_{n\in\mathbb{N}}$ qui converge vers L.*

Preuve. On utilise le même argument de la preuve précédente : grâce à la définition de point d'adhérence, il existe

$$\begin{aligned} x_{n_1} \text{ tel que } \; & d(x_{n_1}, L) < 1 \\ x_{n_2} \text{ tel que } \; & d(x_{n_2}, L) < \frac{1}{2} \\ & \vdots \\ x_{n_k} \text{ tel que } \; & d(x_{n_k}, L) < \frac{1}{k} \end{aligned}$$

2. On rappelle que $\lfloor \xi \rfloor$ est la partie entière, i.e. le plus petit nombre entier non supérieur à ξ.

donc $\lim_{k\to+\infty} x_{n_k} = L$. □

La conséquence du théorème précédent est que, si on garantit l'existence d'un point d'adhérence pour la suite, alors on garantit automatiquement l'existence d'une suite extraite convergente. En dimension infinie ce problème est plutôt délicat, par contre, en dimension finie, une condition suffisante est garantie par le célèbre théorème qui suit, donc on assumera la preuve.

Théorème B.1.3 (Théorème de Bolzano-Weierstrass) *Soit $E \subseteq \mathbb{R}^n$ une partie borné (i.e. contenue dans le voisinage d'un point de $\mathbb{R}^n$) et infinie (i.e. avec un nombre infini d'éléments), alors E admet un point d'adhérence.*

Corollaire B.1.1 *Toute suite bornée à valeurs en $\mathbb{R}^n$ admet une suite extraite convergente.*

Preuve. Conséquence directe des deux derniers théorèmes. Supposons que la suite soit constante après un certain $\bar{n} \in \mathbb{N}$: alors elle est sûrement bornée et convergente (vers la constante même), donc toutes ses suites extraites sont convergentes.

Supposons maintenant que la suite soit non constante et borné. Comme la suite n'est pas constante, elle est composé par un nombre infini d'éléments, et donc, en tant que partie de $\mathbb{R}^n$, elle est bornée et infinie. Par conséquent, elle admet un point d'adhérence par le théorème de Bolzano-Weierstrass, ce qui implique l'existence d'une suite extraite convergente par le théorème B.1.2. □

B.2 Éléments de calcul différentiel en $\mathbb{R}^n$ pour l'optimisation

Dans cette section on rappelle les éléments de calcul différentiel en $\mathbb{R}^n$ qui sont indispensables pour l'optimisation. On assumera pratiquement toutes les preuves des résultats qu'on va citer, car on imagine que le lecteur a déjà eu la possibilité de les voir dans les cours canoniques d'analyse.

Pour commencer, le calcul différentiel est développé d'abord pour les ensembles ouverts, dont on rappelle la définition ci-dessous.

Déf. B.2.1 *$E \subseteq \mathbb{R}^n$ est dit **ouvert** si :*

$$\forall x_0 \in E \; \exists r > 0 \text{ tel que } U_r(x_0) \subseteq E,$$

i.e. pour tout élément d'une partie ouverte E de $\mathbb{R}^n$, on peut trouver un voisinage de rayon positif composé que par des éléments de E même. La raison pour laquelle cette propriété est si importante en analyse est que *l'opération basique du calcul différentiel est la perturbation de la position d'un point*, le fait d'avoir un entier voisinage de chaque point de E composé par des éléments de E *permet d'opérer des perturbations dans toutes les directions, sans devoir se préoccuper de « sortir » de E.*

Déf. B.2.2 *La partie*

$$\partial E = \{x_0 \in E \text{ tels que } \exists r > 0 \text{ tel que } U_r(x_0) \cap E \neq \varnothing \text{ et } U_r(x_0) \cap E^c \neq \varnothing\}$$

*est dite **frontière de** E, où E^c est le complémentaire de E, i.e. $E^c = \mathbb{R}^n \backslash E$.*

Donc la frontière de E est composée par les points de $\mathbb{R}^n$ qui ont de voisinages qui intersectent d'une manière non triviale E et son complémentaire E^c.

Déf. B.2.3 *E est **fermé** si son complémentaire E^c est ouvert, et vice-versa.*

B.2.1 Dérivée directionnelle, partielle, gradient et ligne de niveau

On commence par une définition préliminaire.

Déf. B.2.4 *Soit $u \in R^n$, $\|u\| = 1$, un vecteur unitaire. On appelle droite en $\mathbb{R}^n$ passant par le point $x_0 \in \mathbb{R}^n$ en direction de u l'ensemble défini comme ça :*

$$r_{x_0,u} = \{x \in \mathbb{R}^n \ : \ x = x_0 + tu, \ t \in \mathbb{R}\}.$$

Dans les définitions suivantes on va considérer une fonction $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$, où D est le domaine de f, qu'on va supposer ouvert.

Déf. B.2.5 *Soit $u \in \mathbb{R}^n$ un vecteur unitaire, $\|u\| = 1$. La dérivée directionnelle de f en $x_0 \in D$ en direction direction u est la valeur de la limite suivante, si elle existe et elle est finie :*

$$\boxed{D_u f(x_0) = \lim_{\varepsilon \to 0} \frac{f(x_0 + \varepsilon u) - f(x_0)}{\varepsilon}}.$$

La dérivée directionnelle est l'expression de la vitesse de variation d'une fonction quand on se déplace d'un point en suivant la droite passant par le point en direction d'un vecteur unitaire fixé.

Comme la définition est faite en utilisant l'opération de limite, qui est linéaire, la dérivée directionnel est linéaire elle-même.

Si $n = 2$, on peut expliciter simplement la définition : si $x_0 = (x_0, y_0)$ et $u = (u_1, u_2)$, $\sqrt{u_1^2 + u_2^2} = 1$, alors :

$$D_u f(x_0) = \lim_{\varepsilon \to 0} \frac{f(x_0 + \varepsilon u_1, y_0 + \varepsilon u_2) - f(x_0, y_0)}{\varepsilon}.$$

En $\mathbb{R}^n$ on a n directions privilégiées, celles de la base canonique $e_i(j) = \delta_{i,j}$, $i, j = 1, \ldots, n$, les dérivées directionnelles calculées par rapport aux vecteurs de la base canonique ont un nom et un symbole particulier.

Déf. B.2.6 *On appelle **derivée partielle** selon l'axe i, $i = 1, \ldots, n$, de f en $x_0 \in D$ la valeur de la limite suivante, si elle existe et elle est finie :*

$$D_{e_i} f(x_0) \equiv \frac{\partial f}{\partial x_i}(x_0) = \lim_{\varepsilon \to 0} \frac{f(x_0 + \varepsilon e_i) - f(x_0)}{\varepsilon},$$

plus explicitement, comme $x_0 + \varepsilon e_i = (x_0, \ldots, x_i + \varepsilon, \ldots, x_n)$, on peut écrire

$$\boxed{\frac{\partial f}{\partial x_i}(x_0) = \lim_{\varepsilon \to 0} \frac{f(x_0, \ldots, x_i + \varepsilon, \ldots, x_n) - f(x_0, \ldots, x_i, \ldots, x_n)}{\varepsilon}}.$$

Notations alternatives : $\partial_{x_i} f(x_0)$, $f_{x_i}(x_0)$.

Donc, quand on calcule la dérivée partielle i-ème, seulement la composante i varie, les autres doivent être considérées comme fixes.

L'interprétation géométrique des dérivées partielles est une conséquence directe de celle de dérivée d'une fonction d'une variable réelle. Pour visualiser cela, considérons le cas $n = 2$ et fixons avant $x = x_0$ et après $y = y_0$, ce qu'on obtient sont deux courbes sur la surface définie par l'équation $z = f(x, y)$, comme dans la figure B.2.1.

Les dérivées partielles représentent la peinte des droites tangentes en chaque point à la courbe mentionnée ci-dessus.

Déf. B.2.7 *Les dérivées partielles de $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$ en $x_0 \in D$ peuvent être utilisées comme composantes d'un vecteur de $\mathbb{R}^n$ qu'on appelle **gradient de f en** x_0 :*

$$\boxed{\nabla f(x_0) \equiv grad f(x_0) = \left(\frac{\partial f}{\partial x_1}(x_0), \ldots, \frac{\partial f}{\partial x_n}(x_0) \right)}.$$

On peut considérer $\nabla f(x_0)$ comme un vecteur colonne ou ligne, selon les nécessités.

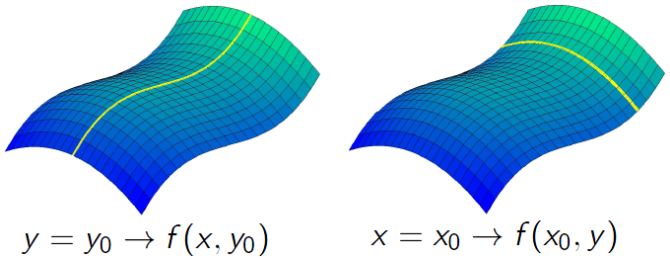


FIGURE B.1 – Variations d'une seule variable pour une fonction de deux variable.

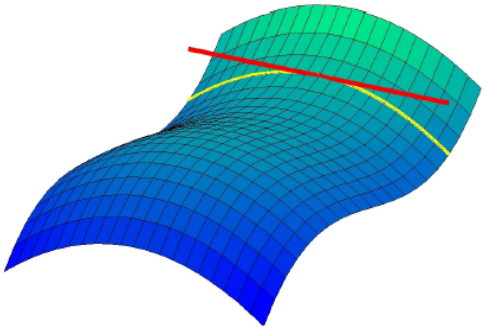


FIGURE B.2 – Représentent géométrique d'un dérivée partielle comme peinte de la droite tangente.

Exemple : calculer le gradient de $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x, y) = \log(1+x^2+y^2)$ dans un point arbitraire de coordonnées (x, y). Avant tout on calcule les dérivées partielles en (x, y) : $f_x(x, y) = 2x/(1+x^2+y^2)$, $f_y(x, y) = 2y/(1 + x^2 + y^2)$, alors :

$$\nabla f(x, y) = \left(\frac{2x}{1 + x^2 + y^2}, \frac{2y}{1 + x^2 + y^2} \right).$$

Le gradient n'est pas simplement une forme compacte pour organiser les dérivées partielles, en fait il contient une information géométrique très importante, comme dit par le théorème suivant.

Théorème B.2.1 *Soient $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$, $x_0 \in D$, $u \in \mathbb{R}^n$, $\|u\| = 1$. Alors :*

$$\boxed{D_u f(x_0) = \langle \nabla f(x_0), u \rangle.} \tag{B.1}$$

Grâce à la formule (B.1) on peut calculer le gradient via la dérivée directionnelle ou vice-versa, selon l'utilité et la simplicité de calcul. Allons voir un exemple : calculer la dérivée directionnelle de $f(x, y) = x^2 e^{-y}$ en (x, y) dans la direction de $u = \left(1/\sqrt{2}, 1/\sqrt{2}\right)$. Le calcul direct donne :

$$D_u f(x, y) = \lim_{h \to 0} \frac{f(x + hu_x, y + hu_y) - f(x, y)}{h}$$

i.e.

$$\lim_{h \to 0} \frac{\left(x + \frac{h}{\sqrt{2}}\right)^2 e^{-\left(y + \frac{h}{\sqrt{2}}\right)} - x^2 e^{-y}}{h},$$

qui est une limite difficile à calculer, pour cela on utilise la formule (B.1) après avoir calculé le gradient : $f_x(x, y) = 2xe^{-y}$, $f_y(x, y) = -x^2 e^{-y}$, donc $\nabla f(x, y) = (2xe^{-y}, -x^2 e^{-y})$ et

$$D_u f(x, y) = \langle \nabla f(x, y), u \rangle = \frac{2xe^{-y}}{\sqrt{2}} + \frac{(-x^2 e^{-y})}{\sqrt{2}} = \frac{xe^{-y}}{\sqrt{2}}(2 - x).$$

L'utilisation inverse de la formule, i.e. le calcul du gradient via la dérivée directionnelle, sera montré dans la section suivante.

On termine cette section avec la signification géométrique de la formule (B.1). On rappelle que, pour toute couple de vecteurs $v, w \in \mathbb{R}^n$

$$\langle v, w \rangle = \|v\|\|w\| \cos\alpha,$$

où α est l'angle le plus petit entre les deux vecteurs.

D'après cette observation, on peut réécrire (B.1) comme ceci :

$$D_u f(x_0) = \|\nabla f(x_0)\|\|u\| \cos\alpha, \qquad \alpha = \text{angle entre } \nabla f(x_0) \text{ et } u,$$

mais $\|u\| = 1$, donc :

$$\boxed{D_u f(x_0) = \|\nabla f(x_0)\| \cos\alpha}.$$

Le cosinus est une fonction bornée entre -1 et +1, donc :

$$\boxed{-\|\nabla f(x_0)\| \leqslant D_u f(x_0) \leqslant +\|\nabla f(x_0)\|}, \qquad \forall u \in \mathbb{R}^n.$$

Il y a trois situations remarquables pour la valeur de la fonction cosinus : quand elle prend sa valeur inférieure -1, sa valeur supérieure +1 et quand elle s'annule. Ces trois situations correspondent, respectivement, au fait que la dérivée directionnelle atteint sa valeur minimale (la plus négative), sa valeur maximale et que elle soit nulle. Allons examiner la signification de ces trois situations.

— $\cos\alpha = +1$ si et seulement si $\nabla f(x_0)$ et u sont parallèles ($\nabla f(x_0) \parallel u$), i.e. $\alpha = 0$. Donc, **la direction de plus rapide croissance** de la function f par rapport au point x_0 es celle du gradient de f en x_0. Le vecteur unitaire qui représente cette direction est :

$$\boxed{u_{\text{max. croissance}} = \frac{\nabla f(x_0)}{\|\nabla f(x_0)\|}}.$$

— De la même manière, $\cos\alpha = -1$ si et seulement si $\nabla f(x_0)$ et u son anti-parallèles, i.e. $\alpha = \pi$. Ceci implique que **la direction de plus rapide décroissance** de la function f par rapport au point x_0 es celle opposée au gradient de f en x_0. Le vecteur unitaire qui représente cette direction est :

$$\boxed{u_{\text{max. décroissance}} = -\frac{\nabla f(x_0)}{\|\nabla f(x_0)\|}}.$$

— $D_u f(x_0)$: s'il y a eu un déplacement, le gradient ne peut pas être nul, donc,$\|\nabla f(x_0)\| \neq 0$ et, comme $\|u\| = 1$, la seule possibilité d'avoir $D_u f(x_0)$ est que $\cos\alpha = 0$. Ceci est possible seulement si $\nabla f(x_0)$ et u son orthogonaux($\nabla f(x_0) \perp u$).

La dernière option qu'on a examiné nous permet d'introduire un concept très importante en *optimisation sous contraintes.*

Déf. B.2.8 *On appelle* ***ligne de niveau*** λ *de la fonction* $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$ *l'ensemble*

$$C_f(\lambda) = \{x \in D \ : \ f(x) = \lambda\}.$$

Comme f est constante sur une ligne de niveau, la dérivée directionnelle de f en un point x_0 calculé par rapport au vecteur u tangent à la ligne de niveau de f qui passe par x_0 est nulle. Mais on vient de voir que la nullité de la dérivée directionnelle correspond à l'orthogonalité entre la direction de dérivation et le vecteur gradient de f en x_0, donc les lignes de niveau de f peuvent être définies d'une manière équivalente comme les lignes dont le vecteur tangent est orthogonale au gradient de f en chaque point. Avec un langage pas formel, mais qui a le don de la synthèse, on dit habituellement que « *le gradient est orthogonale aux lignes de niveau* ». La figure B.2.1 visualise ce concept.

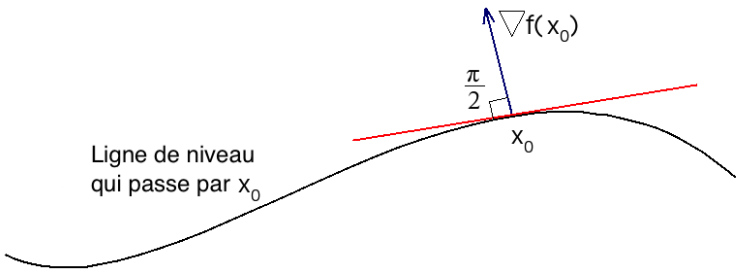


FIGURE B.3 – Relation ligne de niveau et gradient.

B.2.2 Calcul de quelque gradient utile pour l'optimisation via la dérivée directionnelle

Le calcul de la dérivée directionnelle de fonctions qui dépendent de la norme au carré ou du produit scalaire est particulièrement simple, comme on va le voir dans les exemples suivants. Les calculs de cette section seront utilisés souvent dans le cours.

Avant de commencer avec les calculs, on rappelle que, pour tout $a, b \in \mathbb{R}^n$:

$$\|a+b\|^2 = \langle a+b, a+b\rangle = \langle a,a\rangle + \langle a,b\rangle + \langle b,a\rangle + \langle b,b\rangle,$$

par symétrie du produit scalaire Euclidien réel, on obtient

$$\|a+b\|^2 = \|a\|^2 + \|b\|^2 + 2\langle a,b\rangle.$$

Théorème B.2.2 *Si* $f(x) = \|x\|^2$ *alors* $\nabla f(x) = 2x \ \forall x \in \mathbb{R}^n$.

Preuve. Par calcul direct :

$$\begin{aligned} D_u f(x) &= \lim_{\varepsilon\to 0} \frac{f(x+\varepsilon u) - f(x)}{\varepsilon} = \lim_{\varepsilon\to 0} \frac{\|x+\varepsilon u\|^2 - \|x\|^2}{\varepsilon} \\ &= \lim_{\varepsilon\to 0} \frac{\|x\|^2 + \|\varepsilon u\|^2 + 2\langle x, \varepsilon u\rangle - \|x\|^2}{\varepsilon} \\ &= \lim_{\varepsilon\to 0} \frac{\varepsilon^2\|u\|^2 + 2\varepsilon\langle x,u\rangle}{\varepsilon} \\ &= \lim_{\varepsilon\to 0} \left(\varepsilon\|u\|^2 + 2\langle x,u\rangle\right) = 2\langle x,u\rangle. \end{aligned}$$

Grâce à (B.1), $D_u f(x) = \langle \nabla f(x), u\rangle = 2\langle x,u\rangle$, i.e. $\langle \nabla f(x), u\rangle = \langle 2x, u\rangle$, or $\langle \nabla f(x) - 2x, u\rangle = 0$ pour toutes les directions u, mais cela est possible si et seulement si $\nabla f(x) - 2x = 0$, i.e. $\nabla f(x) = 2x$. □

Observation sur les dimensions : $f : \mathbb{R}^n \to \mathbb{R}$, $f(x) = \|x\|^2$, $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$, $\nabla f(x) = 2x$!

Théorème B.2.3 *Si* $f_a(x) = \|x-a\|^2$ *alors* $\nabla f_a(x) = 2(x-a) \ \forall x, a \in \mathbb{R}^n$.

Interprétation : le calcul du gradient de la fonction norme au carré de $x \in \mathbb{R}^n$ (et de ses translations) est, formellement, identique au calcul de la dérivée première d'une fonction de variable réelle au carré (et de ses translations).

Preuve. Par calcul direct :

$$
\begin{aligned}
D_u f_a(x) &= \lim_{\varepsilon \to 0} \frac{\|x + \varepsilon u - a\|^2 - \|x - a\|^2}{\varepsilon} \\
&= \lim_{\varepsilon \to 0} \frac{\|(x-a) + \varepsilon u\|^2 - \|x - a\|^2}{\varepsilon} \\
&= \lim_{\varepsilon \to 0} \frac{\|x-a\|^2 + \|\varepsilon u\|^2 + 2\langle x - a, \varepsilon u\rangle - \|x-a\|^2}{\varepsilon} \\
&= \lim_{\varepsilon \to 0} \frac{\varepsilon^2 \|u\|^2 + \varepsilon \langle 2(x-a), u\rangle}{\varepsilon} \\
&= \lim_{\varepsilon \to 0} \left(\varepsilon \|u\|^2 + \langle 2(x-a), u\rangle\right) = \langle 2(x-a), u\rangle.
\end{aligned}
$$

Le même argument de la preuve précédente amène à écrire $\langle \nabla f(x) - 2(x-a), u\rangle = 0$ pour toutes les directions u, i.e. $\nabla f(x) = 2(x-a)$. □

Théorème B.2.4 *Si $f_w(x) = \langle w, x\rangle$ alors $\nabla f_w(x) = w \ \forall x, w \in \mathbb{R}^n$.*

Interprétation : le calcul du gradient de la fonction produit scalaire entre deux vecteurs de $\mathbb{R}^n$ est, formellement, identique au calcul de la dérivée première de la fonction produit d'une variable réelle par un scalaire.
Preuve. Par calcul direct :

$$
\begin{aligned}
D_u f_w(x) &= \lim_{\varepsilon \to 0} \frac{\langle w, x + \varepsilon u\rangle - \langle w, x\rangle}{\varepsilon} \\
&= \lim_{\varepsilon \to 0} \frac{\langle w, x\rangle + \varepsilon \langle w, u\rangle - \langle w, x\rangle}{\varepsilon} \\
&= \lim_{\varepsilon \to 0} \frac{\varepsilon \langle w, u\rangle}{\varepsilon} \\
&= \langle w, u\rangle.
\end{aligned}
$$

Donc $\langle \nabla f(x) - w, u\rangle = 0$ pour toutes les directions u, i.e. $\nabla f(x) = w$. □

Théorème B.2.5 *Si $f_{A,b}(x) = \frac{1}{2}\|Ax - b\|^2$ alors $\nabla f_{A,b}(x) = A^t(Ax - b) \ \forall x \in \mathbb{R}^n$, $b \in \mathbb{R}^m$ et pour toute matrice $A \in M_{m,n}(\mathbb{R})$.*

Preuve. Calculons $f_{A,b}(x + \varepsilon u)$:

$$
\begin{aligned}
f_{A,b}(x + \varepsilon u) &= \frac{1}{2}\|A(x + \varepsilon u) - b\|^2 = \frac{1}{2}\|(Ax - b) + \varepsilon Au\|^2 \\
&= \frac{1}{2}\left(\|Ax - b\|^2 + \varepsilon^2 \|Au\|^2 + 2\varepsilon \langle Ax - b, Au\rangle\right).
\end{aligned}
$$

Donc :

$$
\begin{aligned}
D_u f_{A,b}(x) &= \lim_{\varepsilon \to 0} \frac{\|Ax - b\|^2 + \varepsilon^2 \|Au\|^2 + 2\varepsilon \langle Ax - b, Au\rangle - \|Ax - b\|^2}{2\varepsilon} \\
&= \lim_{\varepsilon \to 0} \frac{\varepsilon^2 \|Au\|^2 + 2\varepsilon \langle Ax - b, Au\rangle}{2\varepsilon} \\
&= \lim_{\varepsilon \to 0} \left(\frac{\varepsilon \|Au\|^2}{2} + \langle Ax - b, Au\rangle\right) = \langle Ax - b, Au\rangle \\
&= \langle A^t(Ax - b), u\rangle.
\end{aligned}
$$

Donc $\langle \nabla f_{A,b}(x) - A^t(Ax - b), u\rangle = 0$ pour toutes les directions u, i.e. $\nabla f_{A,b}(x) = A^t(Ax - b)$. □

B.2.3 Les points stationnaires et les équations de Euler-Lagrange

Rappelons les définitions d'extrema d'une fonction de plusieurs variables réelles :

Déf. B.2.9 *Soit $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$ et $x_0 \in D$, alors on dit que*
- *x_0 est un point de **minimum globale** pour f si $f(x_0) \leqslant f(x)\ \forall x \in D$;*
- *x_0 est un point de **maximum globale** pour f si $f(x_0) \geqslant f(x)\ \forall x \in D$;*
- *(x_0) est un point de **minimum locale** pour f s'il existe un voisinage $U(x_0)$ tel que $f(x_0) \leqslant f(x)\ \forall x \in U(x_0)$;*
- *(x_0) est un point de **maximum locale** pour f s'il existe un voisinage $U(x_0)$ tel que $f(x_0) \geqslant f(x)\ \forall x \in U(x_0)$.*

*Un point de minimum ou de maximum est appelé un **extremum**.*

Une représentation graphique est offerte dans la figure B.2.3.

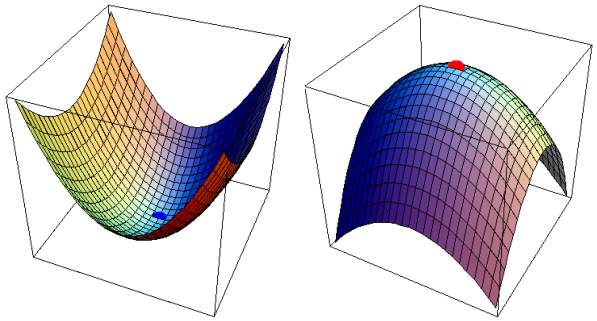


FIGURE B.4 – Exemples de minimum et maximum d'une fonction de deux variables réelles.

Déf. B.2.10 *On appelle $x_0 \in D$ un **point stationnaire** pour une fonction $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$ si $\nabla f(x_0) = \vec{0}$. L'équation $\nabla f(x_0) = \vec{0}$ correspond aux système de n équations qui impose l'annulation des n dérivés partielles de f en x_0, qui sont appelées **équations de Euler-Lagrange**.*

Le résultat suivant est l'équivalent du théorème de Fermat sur les extrema pour les fonctions de plusieurs variables réelles.

Théorème B.2.6 (Fermat en n dimensions) *Si $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$ est partiellement dérivable en $x_0 \in D$ et si x_0 es un extremum pour f, alors : $\nabla f(x_0) = \vec{0}$.*

Par conséquent, les extrema de f peuvent se trouver dans :

- Les points de frontière de D ;
- Les points où f n'est pas dérivable ;
- Les points stationnaires de f.

La condition de stationnarité est seulement nécessaire, pour devenir suffisante elle a besoin d'être accompagné par des autres conditions, notamment la convexité, comme on le montre dans le chapitre 2. La figure B.2.3 montre un cas emblématique : un point « selle », qui est un maximum par rapport à une direction et un minimum par rapport à une autre. Un point selle est stationnaire sans être un extremum.

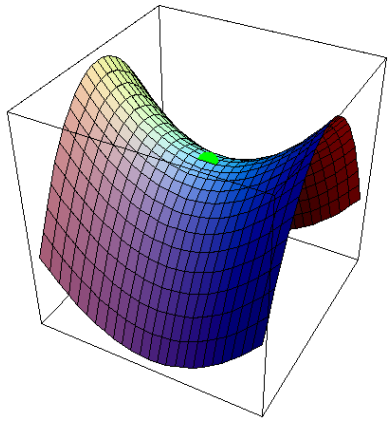


FIGURE B.5 – Un point selle : maximum par rapport à une direction, minimum par rapport à une autre direction.

B.2.4 La matrice Jacobienne

Dans cette section on va examiner l'extension du concept de gradient aux fonctions à valeurs vectoriels : $f : D \subseteq \mathbb{R}^n \to \mathbb{R}^m$, $f = (f_1, \ldots, f_m)$, où $f_i : D \subseteq \mathbb{R}^n \to \mathbb{R}$, $i = 1, \ldots, m$, sont les fonctions composantes, qui sont à valeur scalaires et pour lesquelles on peut définir les dérivées partielles comme on l'a fait avant, i.e.

$$\frac{\partial f_i}{\partial x_j}(x_0) = \lim_{h \to 0} \frac{f_i(x_0 + he_j) - f_i(x_0)}{h}, \qquad i = 1, \ldots, m,\ j = 1, \ldots, n,$$

$$\nabla f_i(x_0) = \left(\frac{\partial f_i}{\partial x_1}(x_0), \ldots, \frac{\partial f_i}{\partial x_n}(x_0) \right), \qquad i = 1, \ldots, m.$$

Si on fait varier les indices i et j on obtient $m \cdot n$ dérivées partielles, qui peuvent être organisées en m vecteurs gradient à n composantes.

Exemple : calculer les dérivées partielles et les gradients des fonctions composantes de la fonction suivante :

$$\begin{array}{rccc} f: & \mathbb{R}^3 & \longrightarrow & \mathbb{R}^2 \\ & (x,y,z) & \longmapsto & f(x,y,z) = (x+y+z,\ xyz^3). \end{array}$$

Comme $n = 3$ y $m = 2$, on va avoir 6 dérivées partielles et 2 vecteurs gradient avec 3 composantes. Les fonctions composantes sont : $f_1(x, y, z) = x + y + z$, $f_2(x, y, z) = xyz^3$, donc :

$$\frac{\partial f_1}{\partial x}(x,y,z) = \frac{\partial f_1}{\partial y}(x,y,z) = \frac{\partial f_1}{\partial z}(x,y,z) = 1,$$

$$\frac{\partial f_2}{\partial x}(x,y,z) = yz^3,\ \frac{\partial f_2}{\partial y}(x,y,z) = xz^3,\ \frac{\partial f_2}{\partial z}(x,y,z) = 3xyz^2.$$

Les gradients des fonctions composantes sont :

$$\nabla f_1(x,y,z) = (1,1,1),\ \nabla f_2(x,y,z) = (yz^3,\ xz^3,\ 3xyz^2).$$

Les $m \cdot n$ dérivées partielles de $f : D \subseteq \mathbb{R}^n \to \mathbb{R}^m$ peuvent être disposées dans une matrice dite **Jacobienne**, une matrice $m \times n$ avec lignes données par les gradients des fonctions composantes :

$$J_f(x) = \begin{pmatrix} \nabla f_1(x) \\ \vdots \\ \nabla f_m(x) \end{pmatrix} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(x) & \cdots & \frac{\partial f_1}{\partial x_n}(x) \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1}(x) & \cdots & \frac{\partial f_m}{\partial x_n}(x) \end{pmatrix}.$$

Retenir que :
— Le nombre de colonnes de $J_f(x)$ est la dimension du domaine de f ;
— Le nombre de lignes de $J_f(x)$ est la dimension du codomaine de f.

Exemple de matrice Jacobienne : $f(x,y,z) = (x+y+z,\, xyz^3)$, alors :

$$J_f(x,y,z) = \begin{pmatrix} \nabla f_1(x,y,z) \\ \nabla f_2(x,y,z) \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 \\ yz^3 & xz^3 & 3xyz^2 \end{pmatrix}.$$

B.2.5 La matrice Hessienne

Comme pour les fonctions d'une seule variable réelle, on peut définir les dérivées partielles d'ordre supérieure. Par exemple, considérons une fonction de deux variables $f(x,y)$:

$$\begin{array}{rccc} \frac{\partial f}{\partial x}: & \mathbb{R}^2 & \longrightarrow & \mathbb{R} \\ & (x,y) & \longmapsto & \frac{\partial f}{\partial x}(x,y) \end{array}$$

$$\begin{array}{rccc} \frac{\partial f}{\partial y}: & \mathbb{R}^2 & \longrightarrow & \mathbb{R} \\ & (x,y) & \longmapsto & \frac{\partial f}{\partial y}(x,y), \end{array}$$

si on dérive partiellement une autre fois on obtient :

$$\frac{\partial}{\partial x}\frac{\partial f}{\partial x}(x,y) = \frac{\partial^2 f}{\partial x^2}(x,y) = f_{xx}(x,y)$$
$$\frac{\partial}{\partial y}\frac{\partial f}{\partial x}(x,y) = \frac{\partial^2 f}{\partial y \partial x}(x,y) = f_{yx}(x,y)$$
$$\frac{\partial}{\partial x}\frac{\partial f}{\partial y}(x,y) = \frac{\partial^2 f}{\partial x \partial y}(x,y) = f_{xy}(x,y)$$
$$\frac{\partial}{\partial y}\frac{\partial f}{\partial y}(x,y) = \frac{\partial^2 f}{\partial y^2}(x,y) = f_{yy}(x,y).$$

On définit :
— $f_{xx}(x,y)$, $f_{yy}(x,y)$: **dérivées partielles d'ordre 2 pures** ;
— $f_{yx}(x,y)$, $f_{xy}(x,y)$: **dérivées partielles d'ordre 2 mixtes**.

On peut itérer le processus de dérivation jusqu'à l'ordre que l'on veut.

Si, au lieu de deux variables on a $n > 2$ variables, alors la technique pour obtenir les dérivées partielles d'ordre supérieure est la même. Il y a n^2 dérivées partielles d'ordre 2 dans ce cas. Heureusement, un résultat très connu nous aide dans le calcul de ces dérivées.

Théorème B.2.7 (Théorème de Schwarz) *Si les dérivées partielles d'ordre 1 de $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$ existent et sont **dérivables** en un voisinage de x_0, alors les dérivées partielles d'ordre 2 de f dans le même voisinage existent et coïncident.*

Comme pour les dérivées partielles d'ordre 1, il existe une structure algébrique très importante dans laquelle on peut placer les dérivées partielles d'ordre 2.

Déf. B.2.11 *La matrice suivante est dite **matrice Hessienne** en x_0 de la fonction f (évidemment supposée être 2 fois partiellement dérivable en x_0) :*

$$H_f(x,y) = \begin{pmatrix} f_{xx}(x,y) & f_{yx}(x,y) \\ f_{x,y}(x,y) & f_{yy}(x,y) \end{pmatrix} \qquad n = 2$$

$$H_f(x,y) = \begin{pmatrix} f_{x_1x_1}(x,y) & f_{x_1x_2}(x,y) & \dots & f_{x_1x_n}(x,y) \\ f_{x_2x_1}(x,y) & f_{x_2x_2}(x,y) & \dots & f_{x_2x_n}(x,y) \\ \vdots & \vdots & \ddots & \vdots \\ f_{x_nx_1}(x,y) & f_{x_nx_2}(x,y) & \dots & f_{x_nx_n}(x,y) \end{pmatrix} \qquad n \text{ arbitraire.}$$

Sous les hypothèses du théorème de Schwarz **la matrice Hessienne est symétrique**.

Exemple : $f(x,y) = \sin(x^2 y)$,

$$f_x(x,y) = 2xy\cos(x^2y), \text{ dérivable partout}$$

$$f_y(x,y) = x^2\cos(x^2y), \text{ dérivable partout}$$

ça vaut le théorème de Schwarz, donc :

$$f_{xx}(x,y) = 2y\cos(x^2y) - 4x^2y^2\sin(x^2y)$$

$$f_{xy}(x,y) = f_{xy}(x,y) = 2x\cos(x^2y) - 2x^3y\sin(x^2y)$$

$$f_{yy}(x,y) = -x^4\sin(x^2y)$$

el alors la matrice Hessienne de f dans le point (x,y) est :

$$H_f(x,y) = \begin{pmatrix} 2y\cos(x^2y) - 4x^2y^2\sin(x^2y) & 2x\cos(x^2y) - 2x^3y\sin(x^2y) \\ 2x\cos(x^2y) - 2x^3y\sin(x^2y) & -x^4\sin(x^2y) \end{pmatrix}.$$

B.2.6 La formule de Taylor pour fonctions de plusieurs variables

Rappelons la formule de Taylor à l'ordre 1 pour fonctions d'une seule variable réelle : si f est dérivable en un voisinage de x_0 avec dérivée première continue, alors ça vaut :

$$f(x) \underset{x\to x_0}{=} f(x_0) + f'(x_0)(x - x_0) + o(x - x_0),$$

où :

$$\lim_{x\to x_0} \frac{o(\|x - x_0\|)}{\|x - x_0\|} = 0.$$

L'expression $\underset{x\to x_0}{=}$ veut dire qu'il existe un voisinage de x_0 dans lequel la formule est valide. L'interprétation de la formule de Taylor au premier ordre est d'importance fondamentale : elle dit qu'il existe un voisinage de x_0 dans lequel la fonction f peut être approchée par la fonction linéaire $y = f(x_0) + f'(x_0)(x - x_0)$, i.e. la droite tangente au graphe de f en x_0, et que l'erreur qu'on fait avec cette approximation, mesuré par le terme $o(x - x_0)$ (« o petit »), est négligeable par rapport à la distance Euclidienne entre x et x_0, i.e. $\|x - x_0\|$.

Si $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$ est une fonction de n variables, on doit remplacer la dérivée première par le gradient : si f est partiellement dérivable 1 fois dans un voisinage de $x_0 \in D$, avec dérivée partielles d'ordre 1 continues, alors ça vaut la formule de Taylor à l'ordre 1 suivante :

$$f(x) \underset{x\to x_0}{=} f(x_0) + \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x_0)(x_i - x_{0,i}) + o\left(\sqrt{\sum_{i=i}^{n}(x_i - x_{0,i})^2}\right),$$

qui peut être écrite dans une forme compacte grâce au gradient et au produit scalaire Euclidien :

$$\boxed{f(x) \underset{x\to x_0}{=} f(x_0) + \langle \nabla f(x_0), (x - x_0)\rangle + o(\|x - x_0\|)}. \tag{B.2}$$

Déf. B.2.12 *L'équation :*

$$\boxed{z = f(x_0) + \langle \nabla f(x_0), (x - x_0)\rangle},$$

*définit l'**hyperplan tangent** à la surface de f en x_0.*

Si $n = 2$ l'équation du plan tangent est :

$$\boxed{z = f(x_0, y_0) + \frac{\partial f}{\partial x}(x_0, y_0)(x - x_0) + \frac{\partial f}{\partial y}(x_0, y_0)(y - y_0)},$$

L'interprétation de la formule de Taylor pour une fonction de n variables est la suivante : il existe un voisinage de x_0 dans lequel la fonction f peut être approchée par la fonction linéaire $z = f(x_0) + \langle \nabla f(x_0), (x - x_0) \rangle$, i.e. l'hyperplan tangent à la surface de f en x_0, et l'erreur qu'on fait avec cette approximation est négligeable par rapport à la distance Euclidienne entre x et x_0.

Il est connu que la fonction valeur absolu $f(x) = |x| = \sqrt{x^2}$ n'est pas dérivable en $x_0 = 0$, en fait dans ce point on ne peut pas définir d'une manière unique une droite tangente à la fonction valeur absolu. L'extension à 2 variables de ce cas est la fonction $f(x_1, x_2) = \sqrt{x_1^2 + x_2^2} = \|x\|$, qui n'est pas partiellement dérivable en $(0, 0)$, qui est le sommet du cône décrit par cette fonction. La généralisation à n variables est simple : $f(x_1, x_2) = \sqrt{x_1^2 + \ldots + x_n^2} = \|x\|$.

Observation importante : le fait de pouvoir approcher localement f à travers d'une fonction linéaire nous permet d'utiliser les outils de l'algèbre linéaire pour obtenir des informations sur l'action de f. Le prix à payer est que cette approximation est précise seulement dans un voisinage d'un point, dès qu'on sort de ce voisinage on doit répéter le processus d'approximation linéaire par rapport à un deuxième point. Celle-ci est la raison pour laquelle les méthodes numériques basés sur les approximations linéaires des fonctions ont besoin de plusieurs étapes d'itération avant d'arriver à un bon résultat.

Si $f : D \subseteq \mathbb{R}^n \to \mathbb{R}^m$, alors la formule de Taylor à l'ordre 1 doit être écrite à l'aide de la matrice Jacobienne :

$$\boxed{f(x) \underset{x \to x_0}{=} f(x_0) + J_f(x_0)(x - x_0) + o(\|x - x_0\|)}, \tag{B.3}$$

la formule a les dimensions correctes si on représente $f(x)$ comme un vecteur colonne $m \times 1$, alors le produit matriciel $J_f(x_0)(x - x_0)$ a dimensions $(m \times n) \times (n \times 1) = m \times 1$ et $o(\|x - x_0\|)$ est aussi un vecteur colonne $m \times 1$.

Rappelons aussi la formule de Taylor à l'ordre 2 pour une fonction d'une seule variable réelle : si f est 2 fois dérivable avec continuité dans un voisinage de x_0, alors ça vaut :

$$f(x) \underset{x \to x_0}{=} f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2} f''(x_0)(x - x_0)(x - x_0) + o((x - x_0)^2),$$

où $o((x - x_0)^2)$ es un erreur négligeable par rapport à $(x - x_0)^n$, i.e.

$$\frac{o((x - x_0)^2)}{(x - x_0)^2} \underset{x \to x_0}{\to} 0.$$

La généralisation à n variables est faite à l'aide de la matrice Hessienne pour remplacer la dérivée seconde : si $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$ est partiellement dérivable 2 fois avec continuité dans un voisinage de $x_0 \in D$, alors ça vaut la formule :

$$\boxed{f(x) \underset{x \to x_0}{=} f(x_0) + \langle \nabla f(x_0), (x - x_0) \rangle + \frac{1}{2} \langle H_f(x_0)(x - x_0), (x - x_0) \rangle + o(\|x - x_0\|^2)}, \tag{B.4}$$

qui montre que $\frac{1}{2} f''(x_0)(x - x_0)^2$ en dimension supérieure à 1 est remplacée par le terme $\frac{1}{2} \langle H_f(x_0)(x - x_0), (x - x_0) \rangle$. Si $n = 2$, alors on peut écrire explicitement cette formule comme ça :

$$\begin{aligned} f(x, y) \underset{(x,y) \to (x_0, y_0)}{=} & f(x_0, y_0) + f_x(x_0, y_0)(x - x_0) + f_y(x_0, y_0)(y - y_0) \\ & + \frac{1}{2} \left(f_{xx}(x_0, y_0)(x - x_0)^2 + 2 f_{xy}(x_0, y_0)(x - x_0)(y - y_0) + f_{yy}(x_0, y_0)(y - y_0)^2 \right) \\ & + o((x - x_0)^2 + (y - y_0)^2). \end{aligned}$$

Les termes d'ordre supérieur dans la formule de Taylor rajoutent des détails plus fins par rapport à l'approximation linéaire de f, comme on peut le voir dans les figures B.2.6 et B.2.6.

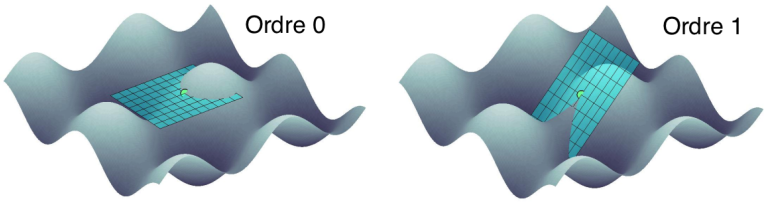


FIGURE B.6 – Approximations d'ordre 0 et 1

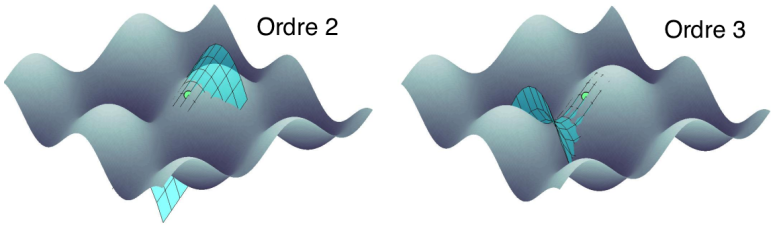


FIGURE B.7 – Approximations d'ordre 2 et 3