



Département Sciences du Numérique

Calcul différentiel - Optimisation sans contraintes - Premiers algorithmes

O. Cots, J. Gergaud, S. Gratton, D. Ruiz et E. Simon

16 septembre 2021

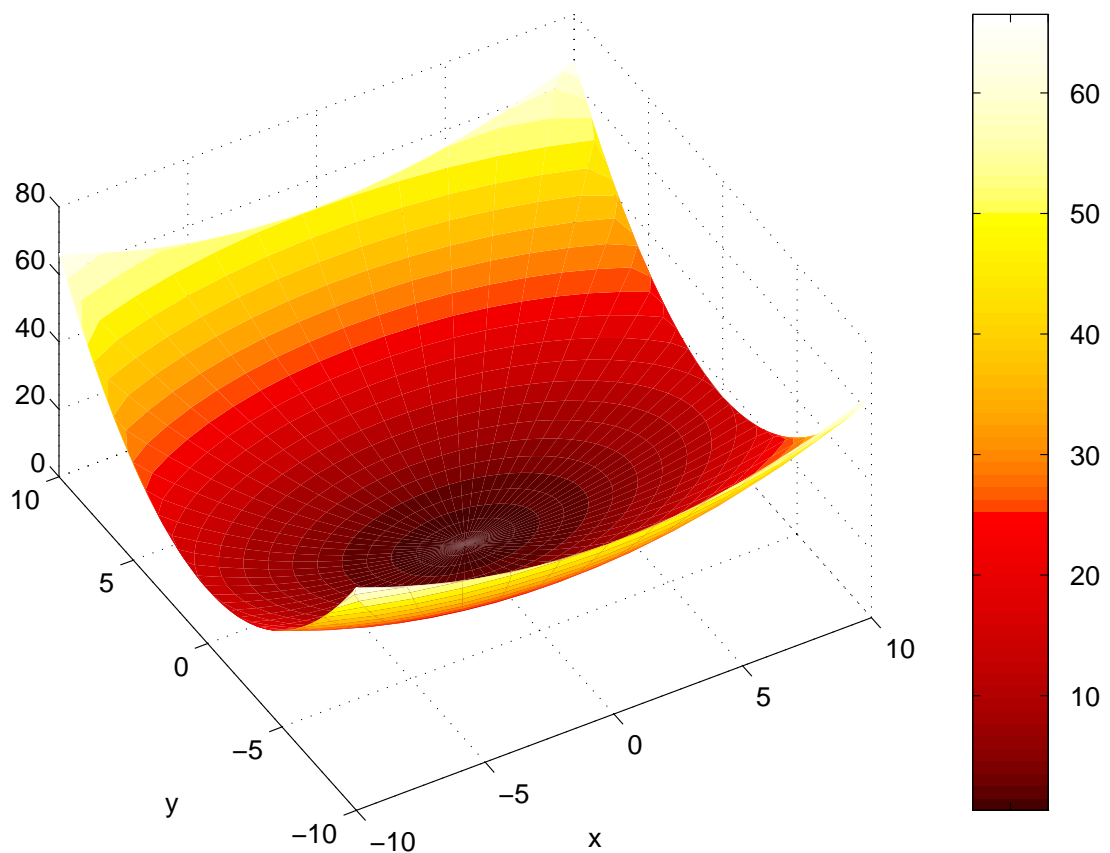


Table des matières

Introduction

Position du problème

Optimiser, c'est rechercher parmi un ensemble C de choix possibles le meilleur (s'il existe!). Si f est une application d'un ensemble E dans F . On note le problème

$$(P) \begin{cases} \min f(x) \\ x \in C \subset E. \end{cases}$$

Il faut donc pour cela pouvoir comparer 2 choix et donc avoir une structure d'ordre sur l'ensemble F . On prendra toujours $F = \mathbb{R}$. Suivant les domaines d'applications :

- E s'appelle l'ensemble des stratégies, des états, des paramètres, l'espace ;
- C est l'ensemble des contraintes ;
- f est la fonction coût, économique ou le critère, l'objectif.

Une fois le problème bien défini, il se pose deux questions. La première est de savoir si (P) admet une solution. Si la réponse est positive, il nous faut trouver la ou les solutions. Suivant la nature de l'ensemble E les réponses sont plus ou moins faciles. Si E est fini, l'existence de solution est évidente, mais le calcul est difficile si le nombre d'éléments est grand. Par contre si $E = \mathbb{R}^n$ ou est de dimension infinie la question de l'existence de solution est moins triviale, mais si les fonctions sont dérivables il est "plus" facile de calculer la solution.

Exemples et définitions

1.1 Exemples

1.1.1 Cas continu et de dimension finie



FIGURE 1.1 – *Pierre de Fermat, né vers 1601, à Beaumont-de-Lomagne, près de Montauban, et mort le 12 janvier 1665 à Castres.*

Exemple 1.1.1 (Principe de Fermat). Pierre de Fermat est un juriste et mathématicien français, surnommé « le prince des amateurs ». On lui doit entre autre le principe de Fermat qui dit que la lumière se propage d'un point à un autre sur des trajectoires telles que la durée du parcours soit minimale. Il imagina aussi pour la solution des problèmes, une méthode, dite de maximis et minimis, qui le fait regarder comme le premier inventeur du calcul différentiel dont il est un précurseur : il est le premier à utiliser la formule (sinon le concept) du nombre dérivé!¹

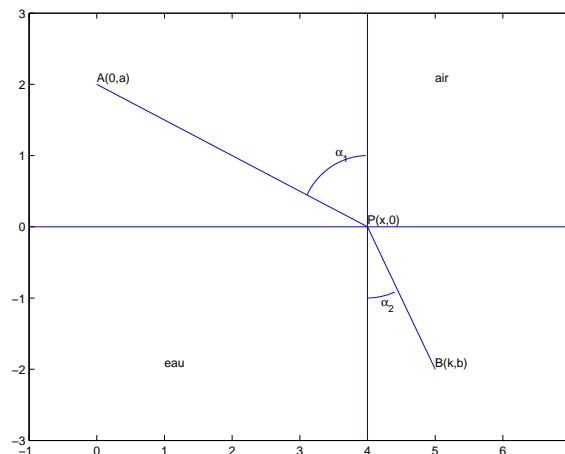


FIGURE 1.2 – Principe de Fermat.

On s'intéresse ici à la trajectoire d'un rayon lumineux d'un point $A(0, a)$ vers un point $B(k, b)$ situés dans deux milieux homogènes différents (cf. la figure ??). Nous allons grâce au principe de Fermat retrouver la loi de la réfraction. On suppose pour cela que la trajectoire d'un rayon lumineux dans un

1. http://fr.wikipedia.org/wiki/Pierre_de_Fermat.

milieu homogène est un segment de droite (ce qui peut aussi se démontrer grâce au principe de Fermat via le calcul des variations, cf. l'exemple ??, qui est un problème d'optimisation en dimension infinie!).

On note P , de coordonnées $(x, 0)$, le point d'impact du rayon lumineux sur la surface du changement de milieu et c_1 et c_2 les vitesses de la lumière dans l'air et dans l'eau. Le temps de parcours entre les points A et B est donc

$$T(x) = \frac{1}{c_1} \sqrt{a^2 + x^2} + \frac{1}{c_2} \sqrt{b^2 + (k - x)^2}.$$

Le problème est alors ici de trouver le point P (c'est-à-dire $x^* \in \mathbb{R}$) tel que

$$T(x^*) \leq T(x) \forall x \in \mathbb{R} \iff (P) \left\{ \begin{array}{l} \min T(x) \\ x \in \mathbb{R}. \end{array} \right.$$

On peut ici tracer cette fonction (cf. la figure ??).

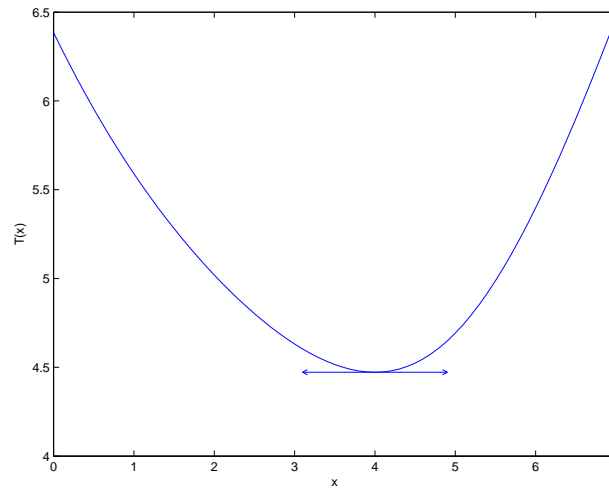


FIGURE 1.3 – Fonction T .

Une condition nécessaire de solution de (P) est $T'(x) = 0$ (cf. la figure ??). Ce qui donne ici

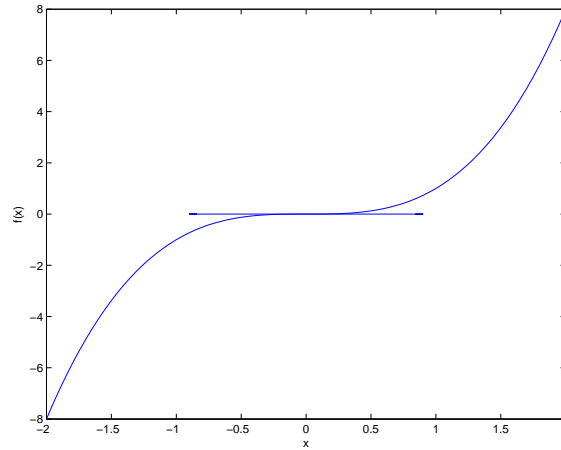
$$\begin{aligned} T'(x) &= \frac{x}{c_1 \sqrt{a^2 + x^2}} + \frac{-(k - x)}{c_2 \sqrt{b^2 + (k - x)^2}} = 0 \\ \iff \frac{x}{c_1 \sqrt{a^2 + x^2}} &= \frac{(k - x)}{c_2 \sqrt{b^2 + (k - x)^2}} \\ \iff \frac{\sin \alpha_1}{c_1} &= \frac{\sin \alpha_2}{c_2} \\ \iff n_1 \sin \alpha_1 &= n_2 \sin \alpha_2. \end{aligned}$$

□

Remarque 1.1.1. Nous retrouvons dans ce cas les lois de Descartes² ou de Snell.

Remarque 1.1.2. La condition $T'(x) = 0$ n'est qu'une condition nécessaire, en effet si nous considérons la fonctionnelle réelle $f(x) = x^3$ nous avons $f'(0) = 0$ mais 0 n'est pas un minimum de f (cf. l'exemple Figure ??).

2. Associer les noms de Fermat et Descartes est surprenant pour qui connaît les confrontations scientifiques virulentes qui les opposèrent. Les étudiants intéressés peuvent voir la vidéo ([?]) où se rendre au musée Pierre de Fermat de Beaumont de Lomagne, ville natale de P. de Fermat près de Toulouse.

FIGURE 1.4 – $f'(0) = 0$ et 0 n'est pas un minimum.

Exemple 1.1.2 (Datation par le carbone 14). Le carbone radioactif ^{14}C est produit dans l'atmosphère par l'effet des rayons cosmiques sur l'azote atmosphérique. Il est oxydé en $^{14}\text{CO}_2$ et absorbé sous cette forme par les organismes vivants qui, par suite, contiennent un certain pourcentage de carbone radioactif relativement aux carbones ^{12}C et ^{13}C qui sont stables. On suppose que la production de carbone ^{14}C atmosphérique est demeurée constante durant les derniers millénaires. On suppose d'autre part que, lorsqu'un organisme meurt, ses échanges avec l'atmosphère cessent et que la radioactivité due au carbone ^{14}C décroît suivant la loi exponentielle suivante :

$$A_{(A_0, \lambda)}(t) = A_0 e^{-\lambda t}$$

où λ est une constante positive, t représente le temps en année et $A(t)$ est la radioactivité exprimée en nombre de désintégrations par minute et par gramme de carbone. On désire estimer les paramètres A_0 et λ par la méthode des moindres carrés. Pour cela on analyse les troncs (le bois est un tissu mort) de très vieux arbres *Sequoia gigantea* et *Pinus aristata*. Par un prélèvement effectué sur le tronc, on peut obtenir (cf. table ??) :

- son âge t en année, en comptant le nombre des anneaux de croissance,
- sa radioactivité A en mesurant le nombre de désintégration.

t_i	500	1000	2000	3000	4000	5000	6300
A_i	14.5	13.5	12.0	10.8	9.9	8.9	8.0

TABLE 1.1 – Données.

Notre but est ici de trouver les valeurs des paramètres A_0 et λ pour que la fonction $A_{(A_0, \lambda)}(t)$ "colle" au mieux aux données.



Ici les instants t_i et les valeurs A_i , pour $i = 1, \dots, 7$ sont connus. Ce sont les valeurs des paramètres A_0 et λ que l'on cherche. Le statut de A_0 et des A_1, \dots, A_7 , n'est donc pas le même.

Si on donne des valeurs aux paramètres, nous pouvons calculer les quantités appelées résidus (cf. la Fig. ?? pour les valeurs des paramètres $A_0 = 20$ et $\lambda = 0.0002$)

$$r_i(A_0, \lambda) = A_i - A_{(A_0, \lambda)}(t_i) = A_i - A_0 e^{-\lambda t_i}.$$

Par suite nous pouvons calculer la quantité

$$f(A_0, \lambda) = \frac{1}{2} \sum_{i=1}^n (A_i - A_0 e^{-\lambda t_i})^2.$$

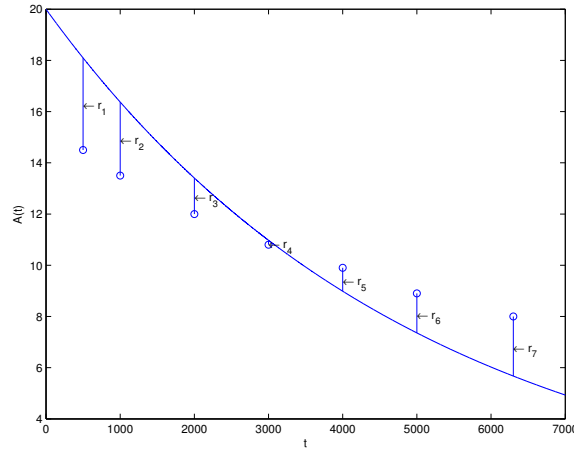


FIGURE 1.5 – Résidus $r(20, 0.0002)$ pour le problème de datation par le carbone 14.

Cette quantité est la somme des carrés des longueurs des résidus.

Plus cette quantité sera faible, plus notre courbe sera proche de nos points expérimentaux. Estimer les paramètres A_0 et λ par les moindres carrés, c'est rechercher la valeur solution du problème d'optimisation suivant :

$$(P) \begin{cases} \min f(A_0, \lambda) = \frac{1}{2} \sum_{i=1}^n (A_i - A_0 e^{-\lambda t_i})^2 \\ (A_0, \lambda) \in \mathbb{R}^2. \end{cases}$$

□

Remarque 1.1.3. • Dans l'exemple précédent on peut aussi écrire : $f(\beta) = \frac{1}{2} \|r(\beta)\|^2$ où

$$\begin{aligned} r : \mathbb{R}^2 &\longrightarrow \mathbb{R}^7 \\ \beta = (A_0, \lambda) &\longmapsto \begin{pmatrix} r_1(\beta) \\ \vdots \\ r_7(\beta) \end{pmatrix} \end{aligned}$$

avec

$$r_i(\beta) = A_i - A_0 e^{-\lambda t_i}$$

et où $\|\cdot\|$ est la norme euclidienne.

- Minimiser $f(\beta)$ est équivalent à minimiser $\alpha f(\beta)$ avec $\alpha > 0$. Le terme $\frac{1}{2}$ est mis ici afin de ne pas avoir le terme 2 lorsque l'on dérive la fonction $f(\beta)$.
- On peut aussi prendre comme critère :
 - $f(\beta) = \|r(\beta)\|_1 = \sum_{i=1}^n |r_i(\beta)|$;
 - $f(\beta) = \|r(\beta)\|_\infty = \max_{i=1, \dots, n} |r_i(\beta)|$.

Remarque 1.1.4. Cet exemple est un exemple important d'un problème d'estimations de paramètres dans un modèle par les moindres carrés. Nous en verrons beaucoup d'ordre dans ce cours.

Définition 1.1.1

On appelle problème aux moindres carrés tout problème qui s'écrit

$$(P) \begin{cases} \min f(\beta) = \frac{1}{2} \|r(\beta)\|^2 \\ \beta \in \mathbb{R}^p. \end{cases}$$

où r est une fonction de \mathbb{R}^p à valeurs dans \mathbb{R}^n .

le problème au moindres carrés est dit linéaire si r est une fonction affine : $r(\beta) = y - X\beta$.

Exercice 1.1.3. Régression linéaire simple

Soit n points expérimentaux $M_i = (x_i, y_i)$ pour $i = 1, \dots, n$. On considère le modèle suivant : $y(x, \beta) = \beta_0 + \beta_1 x$.


1. On veut estimer les paramètres par les moindres carrés. Écrire le problème sous la forme :

$$\begin{cases} \text{Min} & f(\beta) = \frac{1}{2} \|r(\beta)\|^2 = \frac{1}{2} \|y - X\beta\|^2 \\ \beta \in \mathbb{R}^p \end{cases}$$

On donnera les valeurs de X et de y et à quoi correspond β .

2. On souhaite maintenant trouver la meilleure droite au sens des moindres carrés qui passe par l'origine. Écrire le problème d'optimisation.

□

 **Exercice 1.1.4** (Courbe étalon). La première étape d'un dosage radioimmunologique consiste à établir une courbe étalon. Un dosage repose sur l'hypothèse qu'une hormone et son *isotope marqué* se comportent de façon équivalente vis-à-vis de leur anticorps spécifique : lorsque l'on met en présence une quantité déterminée d'anticorps, une quantité déterminée d'hormone radioactive et une quantité variable d'hormone froide, la dose de complexe anticorps-hormone marquée en fin de réaction est d'autant plus faible que la quantité d'hormone froide est importante. Néanmoins, la relation qui existe entre la dose d'hormone froide mise en réaction et la radioactivité de complexe extrait n'est pas stable et doit être appréciée dans chaque situation expérimentale. C'est l'objet de l'établissement de la courbe étalon, à partir d'une gamme de dilutions connues d'une quantité déterminée de l'hormone à doser. La table ?? donne les données recueillies pour une telle courbe dans le cas d'un dosage du cortisol : on a mesuré la radioactivité du complexe (en coups par minutes ou cpm). On considère le modèle suivant :

$$y(x, \beta) = \beta_2 + \frac{\beta_1 - \beta_2}{(1 + \exp(\beta_3 + \beta_4 x))^{\beta_5}}. \quad (1.1)$$

Dose en ng/.1 ml	Réponse en c.p.m.			
0	2868	2785	2849	2805
0	2779	2588	2701	2752
0.02	2615	2651	2506	2498
0.04	2474	2573	2378	2494
0.06	2152	2307	2101	2216
0.08	2114	2052	2016	2030
0.1	1862	1935	1800	1871
0.2	1364	1412	1377	1304
0.4	910	919	855	875
0.6	702	701	689	696
0.8	586	596	561	562
1	501	495	478	493
1.5	392	358	399	394
2	330	351	343	333
4	250	261	244	242
100	131	135	134	133

TABLE 1.2 – Données pour un dosage de Cortisol

On désire estimer les paramètres par les moindres carrés (attention, il y a pour chaque dose 4 observations de y). On notera $(x_i)_{i=1,\dots,16}$ (respectivement $(y_{i,j})_{i=1,\dots,16;j=1,\dots,4}$) les éléments de la première colonne (respectivement des 4 dernières colonnes) de la table ?? et $r_{i,j}(\beta)$ le résidu liés au point $(x_i, y_{i,j})$.

1. Écrire le résidu lié au point (0.04,2378).

2. (i) Quelle est la dimension du vecteur des paramètres β .

(ii) Quel est le nombre de points n ?

3. Écrire le problème d'optimisation des paramètres par les moindres carrés. □

Exemple 1.1.5 (Modèle de Kaplan). On désire étudier la diffusion d'une drogue dans un organe d'un corps donné. La drogue est injectée par intraveineuse dans le sang à l'instant $t_0 = 0$. On modélise le système par un modèle à compartiments (cf. la figure ??).

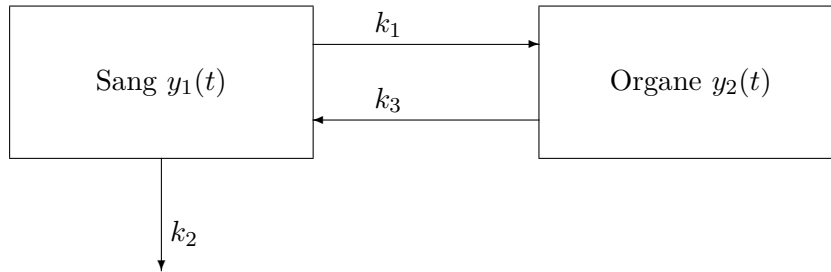


FIGURE 1.6 – Modèle par compartiments.

Les concentrations dans le sang, mesurées à différents instants, sont données à la table ??.

t_i	y_{i1}	t_i	y_{i1}
0.25	215.6	3.00	101.2
0.50	189.2	4.00	88.0
0.75	176.0	6.00	61.6
1.00	162.8	12.00	22.0
1.50	138.6	24.00	4.4
2.00	121.0	48.00	0.0

TABLE 1.3 – Données pour l'exemple de Kaplan.

Le système d'équations différentielles décrivant le modèle est alors

$$(EDO) \begin{cases} \frac{dy_1}{dt} = \dot{y}_1(t) = -(k_1 + k_2)y_1(t) + k_3y_2(t) \\ \frac{dy_2}{dt} = \dot{y}_2(t) = k_1y_1(t) - k_3y_2(t) \\ y_1(0) = c_0 \\ y_2(0) = 0. \end{cases}$$

On désire estimer les paramètres c_0, k_1, k_2 et k_3 par les moindres carrés. Posons $\beta = (c_0, k_1, k_2, k_3)$, alors pour toute valeur de β , on peut intégrer le système d'équations différentielles ordinaires à condition initiale (EDO). Notons $(y_1(t\beta), y_2(t\beta))$ cette solution. Par suite on peut calculer les n résidus

$$r_i(\beta) = y_{i1} - y_1(t_i\beta).$$

Ces résidus sont visualisés sur la figure ?? . Nous estimerons alors le paramètre β en résolvant le problème d'optimisation aux moindres carrés

$$(P) \begin{cases} \min f(\beta) = \frac{1}{2} \sum_{i=1}^n r_i^2(\beta) = \frac{1}{2} \|r(\beta)\|^2 \\ \beta \in \mathbb{R}^4. \end{cases}$$

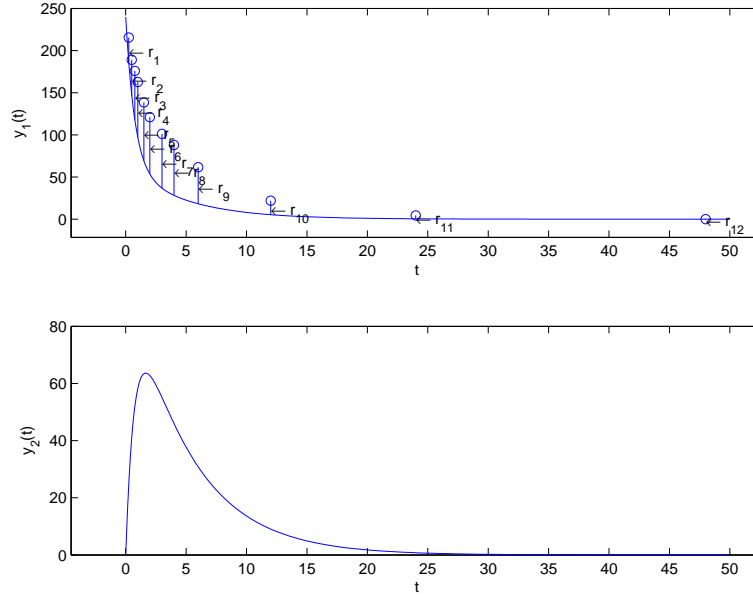


FIGURE 1.7 – Critère des moindres carrés pour le modèle de Kaplan.

□

Exemple 1.1.6. On veut mesurer la liaison entre 2 gènes dominants, l'un contrôlant la couleur d'une fleur, rouge (R) est dominant sur blanc (b), et l'autre la taille, grand (G) est dominant sur petit (p). Dans la descendance F_2 , issu de deux populations homozygotes de phénotype $[RG]$ et $[bp]$, on a étudié $n = 3839$ plantes. On a obtenu les résultats de la table ??.

Phénotypes	$[RG]$	$[Rp]$	$[bG]$	$[bp]$
Effectifs observés	1997	906	904	32

TABLE 1.4 – Données de Sir R.A. Fisher.

Le problème est d'estimer, à partir de ces données le taux de recombinaison r . Ici la population F_1 est hétérozygote de génotype Rb, Gp . Nous avons donc les probabilités de la table ?? pour les différents gamètes possibles et les différents croisements possibles.

Par suite nous avons dans la population F_2 la loi suivante pour la variable aléatoire phénotype X

$$\begin{aligned} X : F_2 &\longrightarrow \{[RG], [Rp], [bG], [bp]\} \\ 1 \text{ plante} &\longmapsto \text{son phénotype,} \end{aligned}$$

$$\begin{aligned} P(X = [RG]) &= \frac{1}{4}(3 - 2r + r^2) = \frac{2 + \theta}{4} \\ P(X = [Rp]) &= \frac{1}{4}(2r - r^2) = \frac{1 - \theta}{4} \\ P(X = [bG]) &= \frac{1}{4}(2r - r^2) = \frac{1 - \theta}{4} \\ P(X = [bp]) &= \frac{1}{4}(1 - r)^2 = \frac{\theta}{4} \end{aligned}$$

$\varphi : \sigma$	RG	bp	Rp	bG
	$\frac{1}{2}(1-r)$	$\frac{1}{2}(1-r)$	$\frac{1}{2}r$	$\frac{1}{2}r$
RG	$\frac{1}{4}(1-r)^2$	$\frac{1}{4}(1-r)^2$	$\frac{1}{4}r(1-r)$	$\frac{1}{4}r(1-r)$
bp	$\frac{1}{4}(1-r)^2$	$\frac{1}{4}(1-r)^2$	$\frac{1}{4}r(1-r)$	$\frac{1}{4}r(1-r)$
Rp	$\frac{1}{4}r(1-r)$	$\frac{1}{4}r(1-r)$	$\frac{1}{4}r^2$	$\frac{1}{4}r^2$
bG	$\frac{1}{4}r(1-r)$	$\frac{1}{4}r(1-r)$	$\frac{1}{4}r^2$	$\frac{1}{4}r^2$

TABLE 1.5 – Probabilités pour la descendance F_2 .

où $\theta = (1-r)^2 \in [\frac{1}{4}; 1]$.

Définissons maintenant le vecteur aléatoire de dimension 4

$$(A, B, C, D) : F_2^n \longrightarrow \mathbb{R}^4$$

(nb de plantes de phénotypes $[RG]$,
 nb de plantes de phénotypes $[Rp]$,
 nb de plantes de phénotypes $[bG]$,
 nb de plantes de phénotypes $[bp]$).

n plantes \longmapsto

On suppose la population F_2 de taille infinie, donc la loi de ce vecteur aléatoire est une loi multinomiale

$$\begin{aligned}
 L(a, b, c, d; \theta) &= P((A, B, C, D) = (a, b, c, d)) \\
 &= \frac{n!}{a!b!c!d!} P(X = [RG])^a P(X = [Rp])^b P(X = [bG])^c P(X = [bp])^d \\
 &= \frac{n!}{a!b!c!d!} \left(\frac{2+\theta}{4}\right)^a \left(\frac{1-\theta}{4}\right)^{b+c} \left(\frac{\theta}{4}\right)^d.
 \end{aligned}$$

L s'appelle la vraisemblance³. L'estimation de θ par le maximum de vraisemblance consiste alors à rechercher la valeur de θ solution du problème de maximisation suivant

$$(P) \begin{cases} \max L(1997, 906, 904, 32; \theta) \\ \theta \in [\frac{1}{4}; 1]. \end{cases}$$

□

Exemple 1.1.7. Un fermier désire déterminer les quantités de lisier de porc et d'engrais composé à étendre sur 20 ha de prairie de façon à optimiser le coût total de la fertilisation. Le coût et la composition du lisier et de l'engrais sont donnés à la table ??.

	coût (par tonne)	composition chimique ($kg t^{-1}$)		
		azote	phosphate	potasse
lisier	25 francs	6	1.5	4
engrais	1300 francs	250	100	100

TABLE 1.6 – Coûts et compositions des engrais.

Le fermier veut appliquer au moins $75 kg ha^{-1}$ d'azote, $25 kg ha^{-1}$ de phosphate et $35 kg ha^{-1}$ de potasse. Il ne peut appliquer le lisier qu'à un taux maximum de $8 t/heure$ et l'engrais qu'à un taux maximum de $0.4 t/heure$. Il ne peut de plus consacrer pour ce travail qu'un maximum de 25 heures.

3. likelihood en anglais.

Appelons x_1 (respectivement x_2) la quantité en tonnes de lisier (respectivement d'engrais) étendu. Le problème est alors d'obtenir un coût minimum, c'est-à-dire que l'on cherche à minimiser $25x_1 + 1300x_2$. Mais nous avons aussi les contraintes suivantes :

$$\begin{array}{ll}
 x_1 \geq 0 & \text{non négativité de } x_1 \\
 x_2 \geq 0 & \text{non négativité de } x_2 \\
 6x_1 + 250x_2 \geq 75 \times 20 = 1500 & \text{contrainte sur l'azote} \\
 1.5x_1 + 100x_2 \geq 500 & \text{contrainte sur le phosphate} \\
 4x_1 + 100x_2 \geq 700 & \text{contrainte sur la potasse} \\
 (1/8)x_1 + (1/0.4)x_2 \leq 25 & \text{contrainte de temps.}
 \end{array}$$

En résumé nous avons le problème suivant à résoudre :

$$(P) \left\{ \begin{array}{l} \min f(x) = 25x_1 + 1300x_2 \\ x_1 \geq 0 \\ x_2 \geq 0 \\ 6x_1 + 250x_2 \geq 75 \times 20 = 1500 \\ 1.5x_1 + 100x_2 \geq 500 \\ 4x_1 + 100x_2 \geq 700 \\ (1/8)x_1 + (1/0.4)x_2 \leq 25. \end{array} \right.$$

□

Exemple 1.1.8 (Gestion de portefeuille [?]). La théorie de la sélection optimale de portefeuille a été développée par Harry Markowitz, prix Nobel d'économie en 1990, dans les années 1950. On considère un investisseur qui a une quantité fixée d'argent à sa disposition pour investir dans n actifs différentes (actions, stocks, ...) dont le retour est aléatoire. Pour chaque actif, on suppose connu son espérance mathématique μ_i , sa variance σ_i^2 . On suppose aussi connu pour deux actifs i et j leur coefficient de corrélation linéaire ρ_{ij} . On note x_i la proportion investie dans l'actif i . On peut donc calculer les espérance mathématique et variance résultant d'un portefeuille $x = (x_1, \dots, x_n)$

$$\begin{aligned}
 E(x) &= \mu^T x \\
 Var(x) &= x^T Q x,
 \end{aligned}$$

où Q est la matrice des covariances, $q_{ij} = \rho_{ij}\sigma_i\sigma_j$. Le portefeuille sera dit efficace si, pour une variance fixée, il a la plus grande espérance mathématique. C'est à dire s'il est solution du problème d'optimisation

$$(P) \left\{ \begin{array}{l} \max E(x) \\ Var(x) = V \\ \sum_{i=1}^n x_i = 1 \\ x \geq 0. \end{array} \right.$$

On peut aussi s'intéresser au problème (MVO)⁴ de Markowitz.

$$(MVO) \left\{ \begin{array}{l} \min Var(x) \\ E(x) \geq R \\ \sum_{i=1}^n x_i = 1 \\ x \geq 0. \end{array} \right.$$

Ces deux formulations sont en fait équivalentes.

□

1.1.2 Problèmes en nombres entiers

Exemple 1.1.9 (Problème du sac à dos de Knapsack). Un alpiniste veut mettre dans son sac à dos un maximum de 16 kg de ravitaillement. Il peut choisir un certain nombre d'unités de trois produits

Produits	I	II	III
Poids	2	5	7
Valeurs	4	10	15

TABLE 1.7 – Poids unitaires et valeurs énergétiques unitaires.

différents. Le poids unitaire en kilogrammes et la valeur énergétique unitaire des ces produits sont connus et donnés dans la table (??).

Le problème pour l'alpiniste est de savoir ce qu'il doit emporter pour avoir une valeur totale en calories maximale sans dépasser les 16 kg.

Si nous notons x_1, x_2 et x_3 les nombres d'unités à emporter des articles I, II et III, le problème s'écrit

$$(P) \begin{cases} \max & 4x_1 + 10x_2 + 15x_3 \\ & 2x_1 + 5x_2 + 7x_3 \leq 16 \\ & (x_1, x_2, x_3) \in \mathbb{N}^3. \end{cases}$$

□

Exemple 1.1.10 (cf. [?]). Dans un service hospitalier, les malades i attendent d'être opérés. Le malade i a besoin d'une durée d'opération D_i . D'autre part, compte tenu des disponibilités des chirurgiens, la somme des durées des opérations possibles chaque jours j de la période étudiée est connue et égale à T_j . On veut minimiser la somme des pénalités d'attente pour les différents malades. On note :

- $x_{ij} = 1$ si le malade i est opéré le jour j ;
- $x_{ij} = 0$ si le malade i n'est pas opéré le jour j ;
- c_{ij} la pénalité du malade i s'il est opéré le jour j . c_{ij} est une fonction croissante de j .

Le problème s'écrit alors :

$$(P) \begin{cases} \min & f(x) = \sum_i \sum_j c_{ij} x_{ij} \\ & \sum_i D_i x_{ij} \leq T_j \quad \forall j \text{ limitation des possibilités opératoire du jour } j \\ & \sum_j x_{ij} = 1 \quad \forall i \text{ Le malade } i \text{ est opéré une fois et une seule} \\ & x_{ij} = 0 \text{ ou } 1 \text{ l'opération est effectuée en une fois.} \end{cases}$$

□

Exemple 1.1.11 (Alignement de séquences). Soit 2 séquences *CTGTATC* et *CTATAATCCC*. On désire trouver le "meilleur" alignement possible. À chaque alignement, est associé un score (simple ici) suivant : pour chaque position on associe 0 si les 2 bases sont identiques, +1 si les deux bases sont différentes et +3 s'il y a un "trou". On effectue ensuite la somme. La figure (??) donne un exemple de la fonction score S .

$$\begin{array}{cccccccccccc} C & T & A & T & - & A & A & - & T & C & C & C \\ - & - & C & T & G & T & A & T & C & - & - & - \\ 3 & 3 & 1 & 0 & 3 & 1 & 0 & 3 & 1 & 3 & 3 & 3 & = & 24 \end{array}$$

FIGURE 1.8 – Exemple de calcul d'un score.

Le problème est alors de résoudre le problème d'optimisation suivant

$$(P) \begin{cases} \min S(\text{alignement}) \\ \text{pour tous les alignements possibles.} \end{cases}$$

Remarque 1.1.5. la difficulté est ici de construire l'ensemble de tous les alignements possibles. Ceci se fait de la façon suivante. Supposons que l'on soit à la position i , alors pour aller à la position $i + 1$, nous avons trois possibilités :

- avancer d'un nucléotide pour les 2 séquences ;
- avancer d'un nucléotide pour la séquence S_1 et mettre un "trou" pour la séquence S_2 ;
- avancer d'un nucléotide pour la séquence S_2 et mettre un "trou" pour la séquence S_1 .

Nous pouvons ainsi construire un arbre permettant d'avoir tous les alignements possibles.

□

1.1.3 Problème en dimension infinie

Exemple 1.1.12 (Problème de la brachistochrone). Le problème de la brachistochrone⁵ fut posé par Jean Bernoulli⁶ (cf. la figure ??) en 1696 et est considéré comme le problème fondateur du calcul des variations.

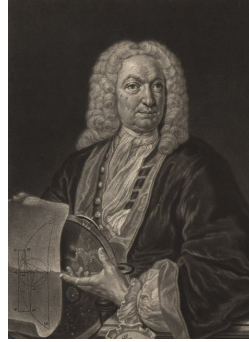


FIGURE 1.9 – Jean Bernoulli 27 juillet 1667 – 1er janvier 1748.

Ce problème consiste en la recherche dans un plan vertical du chemin reliant 2 points P_0 et P_f de ce plan, suivant lequel un corps M entraîné par son propre poids effectuera le trajet de P_0 à P_f en un temps minimum. On suppose qu'il n'y a pas de frottement. Introduisons dans le plan un système de coordonnées (x, y) pour lequel P_0 ait comme coordonnées $(0, 0)$ et $P_f(x_f, y_f)$, $x_f > 0$ et $y_f < 0$. Supposons que $y(\cdot)$ est la fonction qui donne l'équation de la courbe joignant les points P_0 et P_f . Les lois de la mécanique nous disent que le module de la vitesse v en $(x, y(x))$ ne dépend pas de la forme de la courbe $y(\cdot)$ sur $[0, x]$, mais seulement de l'ordonnée $y(x)$, et que cette vitesse est égale à $\sqrt{2g(-y(x))}$, où g est l'accélération gravitationnelle. Si on note s l'abscisse curviligne, le temps pour parcourir l'élément $ds = \sqrt{dx^2 + dy^2}$ est alors $ds/\sqrt{2g(-y(x))}$. Posons

$$T : C^1([0, x_f], \mathbb{R}) \longrightarrow \mathbb{R}$$

$$y(\cdot) \longmapsto \int_0^{x_f} \frac{\sqrt{1 + y'(x)^2}}{\sqrt{2g(-y(x))}} dx.$$

Le problème s'écrit alors

$$(P) \begin{cases} \min T(y(\cdot)) \\ y(0) = 0 \\ y(x_f) = y_f. \end{cases}$$

La solution de ce problème est visualisée à la figure ??.

□

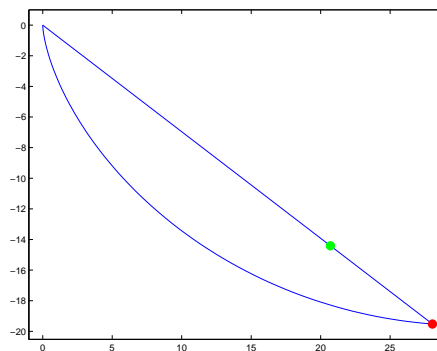


FIGURE 1.10 – Brachistochrone.

5. Le mot brachistochrone vient du grec *brakhisto* qui signifie *le plus court* et de *chronos* qui signifie *temps*.

6. http://fr.wikipedia.org/wiki/Jean_Bernoulli

Exemple 1.1.13 (Transfert orbital). On désire transférer un satellite S d'une orbite initiale (celle où la fusée Ariane l'a "posé") vers l'orbite géostationnaire (cf. la figure ??), le moteur du satellite étant un moteur à poussée faible.

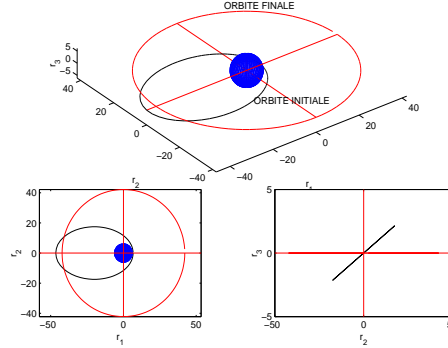


FIGURE 1.11 – Transfert orbital.

Le satellite est considéré comme un point matériel et on note $r(t) \in \mathbb{R}^3$ la position, $v(t) \in \mathbb{R}^3$ la vitesse, $m(t)$ la masse du satellite et $T(t) \in \mathbb{R}^3$ la poussée du moteur. L'équation du mouvement, provenant des équations de Newton est alors

$$\begin{aligned}\dot{r}(t) &= v(t) \\ \dot{v}(t) &= -\frac{\mu r(t)}{\|r(t)\|^3} + \frac{T(t)}{m(t)} \\ \dot{m}(t) &= -\beta \|T(t)\|,\end{aligned}$$

où μ est la constante gravitationnelle de la Terre et $\beta = 1/g_0 Isp$ est une constante positive (g_0 est l'accélération gravitationnelle terrestre à la surface de la Terre et Isp est une constante caractéristique du moteur appelé impulsion spécifique). À l'instant initial les position, vitesse et masse du satellite sont connues et à l'instant terminal t_f le satellite doit être sur l'orbite géostationnaire à une position et vitesse $(r(t_f), v(t_f)) = (r_f, v_f)$ fixés. Bien évidemment la poussée du moteur est bornée

$$\|T(t)\| \leq T_{max}.$$

L'objectif est alors de trouver une loi de commande du moteur qui réalise le transfert et qui minimise le temps de transfert. On peut aussi s'intéresser à la maximisation de la masse finale (dans ce cas le temps de transfert doit-être fixé). Si on normalise le contrôle $u(t) = T(t)/T_{max}$ alors le problème s'écrit pour la maximisation de la masse finale (ou la minimisation de la consommation)

$$(P) \begin{cases} \min & J(u) = \int_0^{t_f} \|u(t)\| dt \\ \dot{r}(t) = v(t) & \text{p.p. dans } [0, t_f], \quad t_f \text{ fixé} \\ \dot{v}(t) = -\mu r(t)/\|r(t)\|^3 + \frac{T_{max}}{m(t)} u(t) \\ \dot{m}(t) = -\beta T_{max} \|u(t)\| \\ (r(t), v(t), m(t)) \in A \\ \|u(t)\| \leq 1 \\ r(0), v(0), m(0) \text{ fixé} \\ r(t_f), v(t_f) \text{ fixé}, \end{cases}$$

ce problème est un problème de contrôle optimal et l'inconnue est la commande, donc une fonction u , ici de $[0, t_f]$ à valeurs dans \mathbb{R}^3 . □

Remarque 1.1.6. Ces problèmes d'optimisation en dimension infinie seront traités dans le cours de contrôle optimal en deuxième année majeure mathématiques appliquées.

1.2 Problème d'optimisation

1.2.1 Définitions

Définition 1.2.1 – Ensemble convexe

Un sous ensemble C d'un espace vectoriel est dit convexe si pour tout $(x, y) \in C^2$ le segment $[x, y] = \{\alpha x + (1 - \alpha)y, \alpha \in [0, 1]\}$ est inclus dans C .

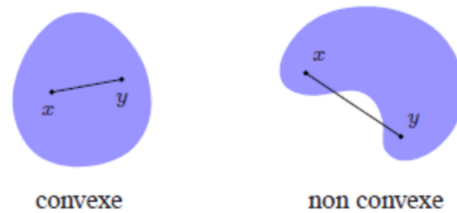


FIGURE 1.12 – Ensemble convexe et non convexe.

Définition 1.2.2 – Fonction convexe

Une fonction f de $C \subset E$ à valeurs dans \mathbb{R} , E espace vectoriel, est convexe si et seulement si elle vérifie :

- (i) C est convexe ;
- (ii)

$$\forall (x, y) \in C^2, \quad \forall \alpha \in [0, 1], \quad f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha)f(y).$$

Dans la cas $n = 1$, ceci signifie que le graphe de la fonction f est toujours sous la corde, cf. la figure (??).

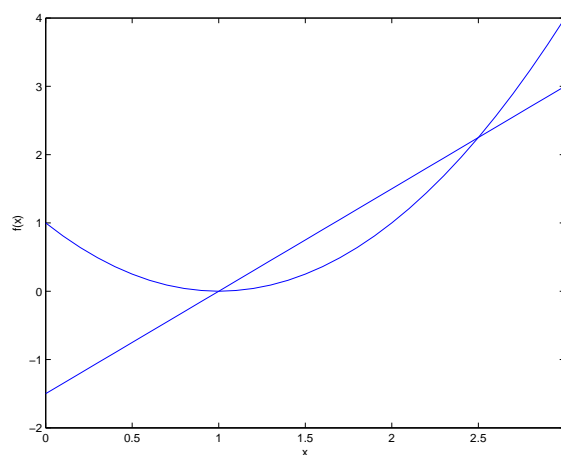


FIGURE 1.13 – Fonction convexe.

Définition 1.2.3 – Problème d'optimisation sans contraintes

On appelle problème d'optimisation sans contraintes en dimension finie tout problème (P) consistant en la recherche d'un minimum d'une fonctionnelle f définie sur \mathbb{R}^n . On notera ce problème

sous la forme suivante :

$$(P) \begin{cases} \min f(x) \\ x \in \mathbb{R}^n \end{cases}$$

où $f : \mathbb{R}^n \rightarrow \mathbb{R}$ sera donnée.

Remarque 1.2.1. Résoudre le problème (P) revient à rechercher le point x^* de \mathbb{R}^n tel que $f(x^*) \leq f(x) \forall x \in \mathbb{R}^n$.

Remarque 1.2.2. Un problème de maximisation se ramène très facilement à un problème de minimisation :

$$\max f(x) \iff \min(-f(x))$$

Définition 1.2.4 – Problème d'optimisation avec contraintes

On appelle problème d'optimisation avec contraintes tout problème (P) consistant en la recherche d'un minimum sur un ensemble C inclus dans \mathbb{R}^n d'une fonctionnelle f définie sur \mathbb{R}^n . On notera ce problème sous la forme suivante :

$$(P) \begin{cases} \min f(x) \\ x \in C \subset \mathbb{R}^n \end{cases}$$

où $f : \mathbb{R}^n \rightarrow \mathbb{R}$ sera donnée.

Remarque 1.2.3. Dans la pratique C sera défini de la façon suivante :

$$C = \{x \in \mathbb{R}^n / g_i(x) \leq 0 \ i = 1, \dots, m \text{ et } h_l(x) = 0 \ l = 1, \dots, p\} \quad (1.2)$$

et nous écrirons (P) sous la forme

$$(P) \begin{cases} \min f(x) \\ g_i(x) \leq 0 \ i = 1, \dots, m \\ h_l(x) = 0 \ l = 1, \dots, p \end{cases}$$

Définition 1.2.5 – Optimisation non différentiable

On appelle problème d'optimisation non différentiable un problème d'optimisation où les fonctions qui interviennent ne sont pas dérivables.

Remarque 1.2.4. On ne traitera dans ce cours que des problèmes d'optimisation différentiables.

Définition 1.2.6 – Problème d'optimisation convexe

Un problème d'optimisation est dit convexe si et seulement si la fonction f est convexe et l'ensemble des contrainte C est convexe.

Remarque 1.2.5. Si C est définie par (??) et si les fonctions g_i sont convexes et les fonctions h_l sont affines, alors C est convexe. Attention, la réciproque est fausse.

Définition 1.2.7 – Problème aux moindres carrés

On appelle problème aux moindres carrés un problème d'optimisation sans contraintes où la fon-

tionnelle f est de la forme suivante :

$$f(\beta) = \frac{1}{2} \|r(\beta)\|^2 = \frac{1}{2} (r(\beta)|r(\beta)) = \frac{1}{2} \sum_{i=1}^n r_i^2(\beta)$$

Le problème est dit aux moindres carrés linéaires si la fonction r est affine :

$$\begin{aligned} r : \mathbb{R}^p &\longrightarrow \mathbb{R}^n \\ \beta &\longmapsto y - X\beta \end{aligned}$$

où X matrice de type (n, p) et y un élément de \mathbb{R}^n .

Remarque 1.2.6. L'exemple (??) est un problème aux moindres carrés non linéaire.

Définition 1.2.8 – Problème linéaire

Un problème d'optimisation est dit linéaire si et seulement si les fonctions f , g_i , et h_l sont affines.

Remarque 1.2.7. L'exemple (??) est un problème linéaire.

Définition 1.2.9 – Optimum global, optimum local

Soit (P) un problème d'optimisation sans contraintes.

- (i) x^* est un minimum global $\iff x^*$ est la solution de (P)
- (ii) x^* est un minimum local faible \iff il existe $\varepsilon > 0$ tel que x^* est la solution de (P') où

$$(P') \left\{ \begin{array}{l} \min f(x) \\ \|x - x^*\| < \varepsilon \end{array} \right.$$

- (iii) x^* est un minimum local fort si

$$\forall \mathbf{x} \in B(\mathbf{x}^*, \varepsilon) = \{\mathbf{x} \in \mathbb{R}^n / \|\mathbf{x} - \mathbf{x}^*\| < \varepsilon\}, \mathbf{x} \neq \mathbf{x}^*, f(\mathbf{x}^*) < f(\mathbf{x}).$$

Dans le cas où $n = 1$, $\|x - x^*\|$ devient $|x - x^*|$ et par suite nous avons

$$\|x - x^*\| < \varepsilon \iff |x - x^*| < \varepsilon \iff x^* - \varepsilon < x < x^* + \varepsilon,$$

(cf. la figure ??).

Remarque 1.2.8. On dit que x^* est un minimum alors que c'est $f(x^*)$ qui est un minimum. Il s'agit d'un abus de langage que nous emploierons systématiquement.

Remarque 1.2.9. On appelle optimisation globale la recherche d'un optimum global. Un algorithme globalement convergent est lui un algorithme qui converge vers un minimum local quel que soit le point de départ.

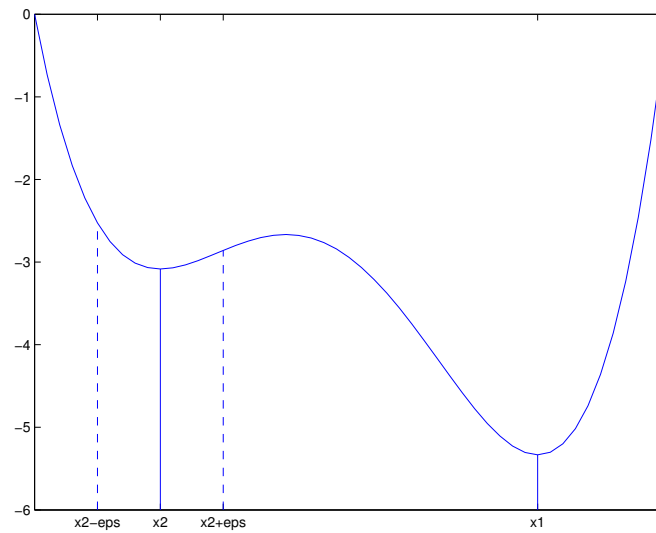


FIGURE 1.14 – x^2 est un minimum local fort, x^1 est un minimum global

1.2.2 Classification

Considérons le problème d'optimisation suivant :

$$(P) \begin{cases} \min f(x) \\ x \in C \subset E \end{cases}$$

Suivant la nature des ensembles C et E et de la fonction f nous avons différents types de problème d'optimisation. La figure ?? donne une classification des problèmes d'optimisation (nous n'étudierons dans ce cours que les parties en bleu).

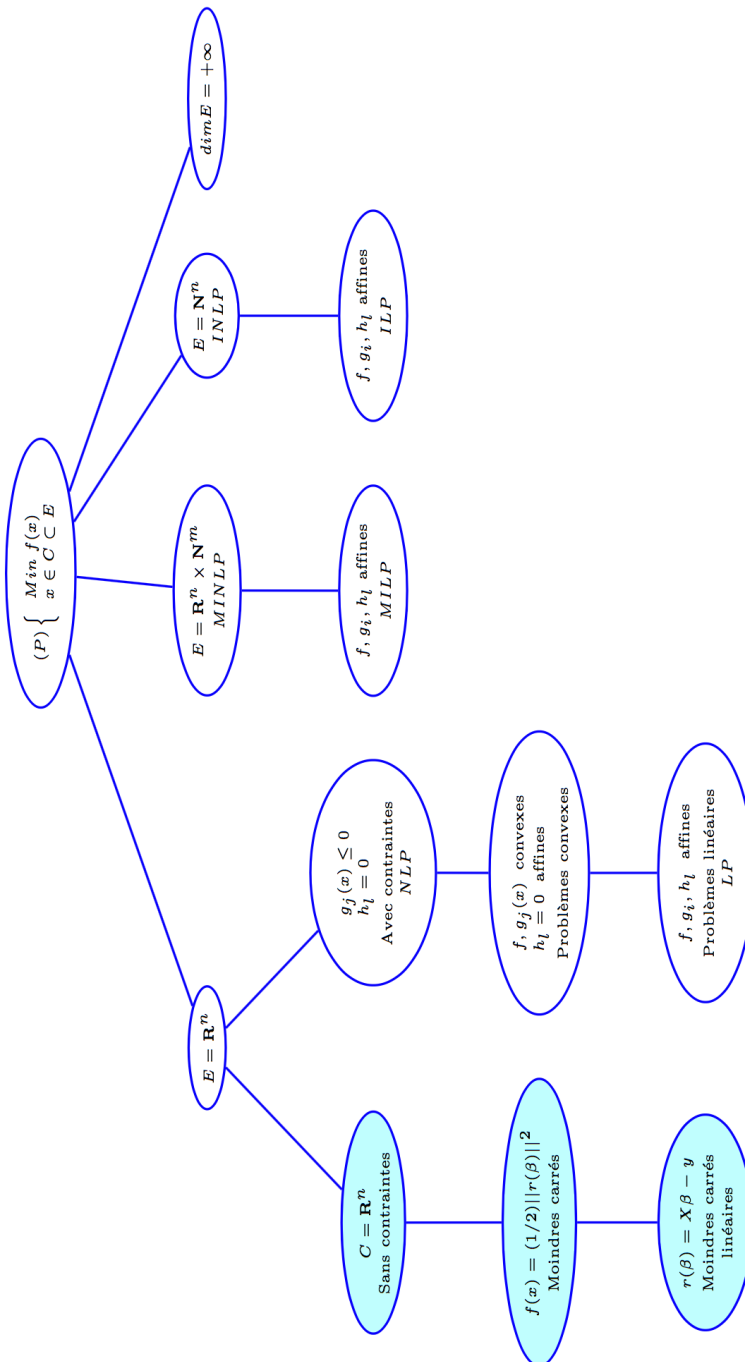



FIGURE 1.15 – Classification des problèmes d'optimisation, on a en bleu ce qui sera vu en cours.

1.3 Exercices

 **Exercice 1.3.1** (Géoréférence d'une image satellite). On dispose d'une image satellite que l'on désire recaler par rapport à une carte géographique que l'on a à notre disposition. Pour cela on définit n points, appelés points d'amer, que l'on peut parfaitement faire correspondre sur la carte et sur l'image satellite. On prend par exemple un croisement de route, un point particulier sur une rivière... Concrètement on a donc à notre disposition n coordonnées (x_i, y_i) des n points d'amer sur la carte et n coordonnées (x'_i, y'_i) de ces mêmes points sur l'image satellite. On choisit d'exprimer ces coordonnées :

- en pixels pour les (x'_i, y'_i) (coordonnées $(0, 0)$ pour le coin inférieur gauche) ;
- en mètres relativement à un référentielle géodésique particulier pour les (x_i, y_i) , via une carte IGN par exemple.

On a par exemple les données suivantes :


Numéros	x_i	y_i	x'_i	y'_i
1	252	2661	458805	1831634
2	235	2603	458157	1830577
\vdots	\vdots	\vdots	\vdots	\vdots
23	1021	2254	471301	1819574

En pratique l'image satellite est déformée par rapport à la réalité. Cette déformation a plusieurs origines : satellite non verticale par rapport à la prise de vue, présence de nuages dans l'atmosphère, ... En conséquence on suppose que l'on peut écrire :

$$\begin{cases} x = \gamma_0 + \gamma_1 x' + \gamma_2 y' + \gamma_3 x'^2 + \gamma_4 x' y' + \gamma_5 y'^2 \\ y = \delta_0 + \delta_1 x' + \delta_2 y' + \delta_3 x'^2 + \delta_4 x' y' + \delta_5 y'^2 \end{cases}$$

On désire estimer les paramètres par les moindres carrées

1. Pour l'estimation des paramètres $\gamma = (\gamma_0, \gamma_1, \dots, \gamma_5)$ quelles sont les données ?
2. Écrire le problème d'estimation par les moindres carrés linéaires de γ .
3. Mêmes questions pour δ . □

 **Exercice 1.3.2** (Réseaux de neurones). On s'intéresse ici à la modélisation via les réseaux de neurones. Un neurone formel est une fonction paramétrée par $n + 1$ paramètres w_1, \dots, w_n, θ

$$g: \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R} \longrightarrow \mathbb{R}$$

$$(x, w, \theta) \longmapsto g(x, w, \theta) := \sigma\left(\sum_{i=1}^n w_i x_i + \theta\right)$$

où σ est une fonction donnée qui s'appelle une fonction d'activation. Chaque paramètre w_i s'appelle le poids synaptique associé au signal d'entrée x_i .

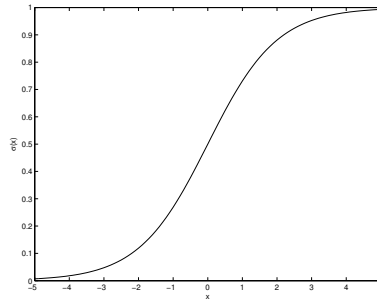
On prendra dans la suite, sauf mention contraire, comme fonction σ la fonction sigmoïde (cf. la figure ??)

$$\sigma: \mathbb{R} \longrightarrow \mathbb{R}$$

$$x \longmapsto \frac{1}{1 + e^x}.$$

On a à notre disposition K points $x^k \in \mathbb{R}^n$ et $y^k \in \mathbb{R}$, on appelle apprentissage du neurone l'estimation par les moindres carrés des paramètres du neurone.

1. Écrire le problème au moindres carrés qui définit l'apprentissage. On donnera en particulier la fonction résidus r en précisant clairement l'espace de départ et l'espace d'arrivée.

FIGURE 1.16 – *Fonction sigmoïde.*

2. Ce problème est-il un problème aux moindres carrés linéaires ? Si oui, on donnera la matrice X .
3. Si on prend comme fonction d'activation σ l'identité le problème au moindres carrés devient-il linéaire ? Si oui, on donnera la matrice X .
4. On considère maintenant le modèle d'une couche de m neurones, c'est-à-dire un ensemble de m neurones g_i ayant la même fonction d'activation σ . Une couche est donc une fonction de \mathbb{R}^n à valeurs dans \mathbb{R}^m dépendant de paramètres w_{ij}, θ_j pour $i = 1, \dots, n$ et $j = 1, \dots, m$.

On a à notre disposition K points $x^k \in \mathbb{R}^n$ et $y^k \in \mathbb{R}^m$, on appelle apprentissage l'estimation par les moindres carrés des paramètres du réseau de neurones formé d'une couche de m neurones.

Écrire le problème au moindres carrés qui définit l'apprentissage. On donnera en particulier la fonction résidus r en précisant clairement l'espace de départ et l'espace d'arrivée. \square

Formes bilinéaires et quadratiques

Dans ce chapitre vous découvrirez :

- L'étude des formes quadratiques et des formes bilinéaires (Il s'agit d'une extension des notions de produit scalaire)
- Une application des notions de diagonalisation aux matrices symétriques et aux endomorphismes symétriques.

2.1 Forme bilinéaire – Matrice d'une forme bilinéaire

2.1.1 Formes bilinéaires

Définition 2.1.1 – Formes bilinéaires

Soit E un espace vectoriel réel de dimension finie. On appelle forme bilinéaire sur E , toute application f de $E \times E$ dans \mathbb{R} vérifiant les propriétés suivantes, pour tous vecteurs \mathbf{u} , $\tilde{\mathbf{u}}$, \mathbf{v} , et $\tilde{\mathbf{v}}$ de E et tout scalaire λ de \mathbb{R} :

$$\begin{aligned} f(\mathbf{u} + \tilde{\mathbf{u}}, \mathbf{v}) &= f(\mathbf{u}, \mathbf{v}) + f(\tilde{\mathbf{u}}, \mathbf{v}) & f(\lambda \mathbf{u}, \mathbf{v}) &= \lambda f(\mathbf{u}, \mathbf{v}) \\ f(\mathbf{u}, \mathbf{v} + \tilde{\mathbf{v}}) &= f(\mathbf{u}, \mathbf{v}) + f(\mathbf{u}, \tilde{\mathbf{v}}) & f(\mathbf{u}, \lambda \mathbf{v}) &= \lambda f(\mathbf{u}, \mathbf{v}) \end{aligned}$$

f est en fait linéaire par rapport à chacune de ses deux variables.

Définition 2.1.2 – Forme bilinéaire symétrique

Soit E un espace vectoriel réel de dimension finie, et soit f une forme bilinéaire sur E . On dit que f est symétrique si, pour tous vecteurs \mathbf{x} et \mathbf{y} de E , on a :

$$f(\mathbf{x}, \mathbf{y}) = f(\mathbf{y}, \mathbf{x}).$$

2.1.2 Représentation matricielle d'une forme bilinéaire

Soit $\mathcal{B} = \{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n\}$ une base de E . Toute forme bilinéaire f est entièrement déterminée par la connaissance des réels $f(\mathbf{e}_i, \mathbf{e}_j)$, pour tout $1 \leq i, j \leq n$. En effet, soient $\mathbf{x} = \sum_{i=1}^n x_i \mathbf{e}_i$ et $\mathbf{y} = \sum_{i=1}^n y_i \mathbf{e}_i$ deux vecteurs de E . Par linéarité à gauche, et à droite, on peut écrire, après développement :

$$f(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^n \sum_{j=1}^n x_i y_j f(\mathbf{e}_i, \mathbf{e}_j).$$

Introduisons alors $\mathbf{X} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ et $\mathbf{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$, les vecteurs de \mathbb{R}^n formés des composantes de \mathbf{x} et \mathbf{y} dans la base \mathcal{B} , et \mathbf{A} la matrice des coefficients $f(\mathbf{e}_i, \mathbf{e}_j)$,

$$\mathbf{A} = \begin{pmatrix} f(\mathbf{e}_1, \mathbf{e}_1) & \dots & f(\mathbf{e}_1, \mathbf{e}_n) \\ \vdots & \ddots & \vdots \\ f(\mathbf{e}_n, \mathbf{e}_1) & \dots & f(\mathbf{e}_n, \mathbf{e}_n) \end{pmatrix}.$$

En utilisant ces notations, on peut alors écrire la valeur de $f(\mathbf{x}, \mathbf{y})$ en terme du produit matriciel suivant :

$$f(\mathbf{x}, \mathbf{y}) = \mathbf{X}^T \mathbf{A} \mathbf{Y}.$$

Propriété : Si f est une forme bilinéaire symétrique sur E , alors la matrice associée à f dans une base quelconque de E est symétrique.

2.1.3 Exemple dans \mathbb{R}^3

$$\begin{aligned} f(\mathbf{x}, \mathbf{y}) &= x_1 y_1 + 2x_2 y_2 + 3x_3 y_3 + x_1 y_3 + x_3 y_1 + x_2 y_3 \\ &= (x_1, x_2, x_3) \begin{pmatrix} 1 & 0 & 1 \\ 0 & 2 & 1 \\ 1 & 0 & 3 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}. \end{aligned}$$

2.2 Formes quadratiques

Définition 2.2.1 – Formes quadratiques

On appelle forme quadratique associée à la forme bilinéaire f , l'application q définie de E dans \mathbb{R} par :

$$\forall \mathbf{x} \in E, \quad q(\mathbf{x}) = f(\mathbf{x}, \mathbf{x}).$$

Remarques :

- On a aussi, en utilisant la matrice \mathbf{A} de f dans une base \mathcal{B} de E :

$$q(\mathbf{x}) = \mathbf{X}^T \mathbf{A} \mathbf{X},$$

où \mathbf{X} est le vecteur des coordonnées de \mathbf{x} dans la base \mathcal{B} . Ainsi, \mathbf{A} représente aussi la matrice de la forme quadratique q dans la base \mathcal{B} .

- Par contre, la représentation matricielle d'une forme quadratique n'est pas unique. En effet, pour une forme quadratique donnée, il existe plusieurs formes bilinéaires qui peuvent lui être associées.

Exemple : Dans \mathbb{R}^3 :

$$\begin{aligned} f(\mathbf{x}, \mathbf{y}) &= x_1 y_1 - 2x_2 y_2 + 3x_3 y_3 + x_1 y_3 + x_3 y_1 + 4x_2 y_3 + 4x_3 y_2 \\ &= (x_1, x_2, x_3) \begin{pmatrix} 1 & 0 & 1 \\ 0 & -2 & 4 \\ 1 & 4 & 3 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}. \end{aligned}$$

La forme quadratique associée est

$$q(\mathbf{x}) = x_1^2 - 2x_2^2 + 3x_3^2 + 2x_1 x_3 + 8x_2 x_3 \quad \text{soit} \quad q(\mathbf{x}) = (x_1, x_2, x_3) \begin{pmatrix} 1 & 0 & 1 \\ 0 & -2 & 4 \\ 1 & 4 & 3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}.$$

Mais on a aussi, du point de vue matriciel :

$$q(\mathbf{x}) = (x_1, x_2, x_3) \begin{pmatrix} 1 & 0 & 2 \\ 0 & -2 & 8 \\ 0 & 0 & 3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = (x_1, x_2, x_3) \begin{pmatrix} 1 & 0 & 0 \\ 0 & -2 & 0 \\ 2 & 8 & 3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}.$$

Propriétés :

- Pour un vecteur $\mathbf{u} \in E$ donné, $q(\mathbf{u})$ est un polynôme homogène de degré 2. Ainsi, tout polynôme homogène de degré 2 par rapport aux coordonnées d'un vecteur \mathbf{u} de E peut correspondre à une forme quadratique q .
- En outre, à la question “*existe-t-il une forme bilinéaire symétrique dont q soit la forme quadratique et si oui, est-elle unique ?*”, la réponse est “oui”.

Voici comment procéder : il suffit pour cela d'écrire la matrice $\mathbf{A} = (a_{ij})$ associée à ce polynôme homogène de degré 2 en plaçant, sur la diagonale, les coefficients a_{ii} correspondant aux termes en x_i^2 , et sur les termes hors diagonaux a_{ij} et a_{ji} la moitié des coefficients des termes en $x_i x_j$.

- Enfin, si à une même forme quadratique q , on peut effectivement associer diverses formes bilinéaires f (de matrice associée \mathbf{A}_f dans une base \mathcal{B} fixée), ces formes bilinéaires ont toutes en commun **la même partie symétrique** :

$$s(\mathbf{u}, \mathbf{v}) = \frac{f(\mathbf{u}, \mathbf{v}) + f(\mathbf{v}, \mathbf{u})}{2}, \quad \text{de matrice associée} \quad \frac{\mathbf{A}_f + \mathbf{A}_f^T}{2} \quad \text{indépendante de } f.$$

Exemple : Dans \mathbb{R}^3 :

$$q(\mathbf{x}) = 5x_1^2 + 12x_2^2 - 6x_3^2 - 8x_2x_3 + 5x_3x_1 - x_2x_1,$$

la forme matricielle symétrique associée étant

$$q(\mathbf{x}) = (x_1, x_2, x_3) \begin{pmatrix} 5 & -1/2 & 5/2 \\ -1/2 & 12 & -4 \\ 5/2 & -4 & -6 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}.$$

2.2.1 Propriétés

Soit f une forme bilinéaire symétrique sur E , et q la forme quadratique associée. Pour tous vecteurs \mathbf{u} et \mathbf{v} de E et tout scalaire λ , on a :

- $q(\lambda\mathbf{u}) = f(\lambda\mathbf{u}, \lambda\mathbf{u}) = \lambda^2 q(\mathbf{u})$: q n'est pas linéaire.
- $f(\mathbf{u}, \mathbf{v}) = \frac{1}{4} (q(\mathbf{u} + \mathbf{v}) - q(\mathbf{u} - \mathbf{v}))$.
- $f(\mathbf{u}, \mathbf{v}) = \frac{1}{2} (q(\mathbf{u} + \mathbf{v}) - q(\mathbf{u}) - q(\mathbf{v}))$.
- Pour une forme quadratique q donnée, la forme bilinéaire symétrique f qui lui est associée est aussi appelée forme polaire de q .
- On définit deux ensembles : Le noyau de q : $N(q) = \{\mathbf{y} \in E, \forall \mathbf{x} \in E, f(\mathbf{x}, \mathbf{y}) = 0\}$
le cône isotrope : $I(q) = \{\mathbf{x} \in E, q(\mathbf{x}) = 0\}$. Sauf cas particulier, ce n'est pas un espace vectoriel, mais un cône, c'est à dire un sous ensemble de vecteurs C tel que si $x \in C$ alors pour tout scalaire λ , $\lambda x \in C$.

On a $N(q) \subset I(q)$.

- q est dite non dégénérée si $N(q) = \{\mathbf{0}\}$.
- q est dite semi-définie positive ssi $\forall \mathbf{x} \in E, q(\mathbf{x}) \geq 0$.
- q est dite semi-définie négative ssi $-q$ est semi-définie positive.
- q est dite indéfinie ssi q n'est ni semi-définie positive, ni semi-définie négative.
- q est dite définie positive si $\forall \mathbf{x} \in E, q(\mathbf{x}) \geq 0$ et $q(\mathbf{x}) = 0 \Rightarrow \mathbf{x} = \mathbf{0}$.
- En dimension finie : $\dim E = \dim N(q) + \text{rang}(q)$ le rang de q est par définition le rang de la matrice de q .

2.3 Formes quadratiques définies positives

2.3.1 Produit scalaire

On rappelle que un **produit scalaire** sur un \mathbb{R} -espace vectoriel E est une forme **bilinéaire, symétrique, et définie positive**. La définie positivité d'une forme bilinéaire f sur E correspond en fait à la définie positivité de sa forme quadratique, à savoir :

$$\forall \mathbf{u} \in E, q(\mathbf{u}) = f(\mathbf{u}, \mathbf{u}) \geq 0 \quad \text{et} \quad q(\mathbf{u}) = f(\mathbf{u}, \mathbf{u}) = 0 \Leftrightarrow \mathbf{u} = \mathbf{0}.$$

Ainsi, sur un même espace vectoriel E , à toute forme quadratique q définie positive, on peut associer un produit scalaire sur E en considérant la forme bilinéaire symétrique f associée à q (la forme polaire de q). Pour un tel un produit scalaire f , $f(\mathbf{u}, \mathbf{v})$ pourra aussi être noté $\langle \mathbf{u}, \mathbf{v} \rangle$.

Proposition 2.3.1

Soit E un \mathbb{R} -espace vectoriel de dimension finie, et soit q une forme quadratique définie positive sur E . Alors, la forme polaire de q , qui est une forme bilinéaire symétrique (ou à symétrie hermitienne si le corps de référence est \mathbb{C}) définie positive sur E , constitue un produit scalaire sur E , et pour la norme associée, E est un espace EUCLIDIEN.

Remarques :

- Une façon de vérifier la définie positivité d'une forme quadratique q donnée consiste à la décomposer en une somme de carrés de termes du premier degré.
- Une autre façon de vérifier la définie positivité d'une forme quadratique q consiste à rechercher les valeurs propres de la matrice symétrique représentant q et à vérifier qu'elles sont bien toutes positives strictement.

2.3.2 Exemples

(i) Dans \mathbb{R}^3 , soit la forme quadratique q définie par

$$q(\mathbf{u}) = x^2 + 6xy + 4yz + 14y^2 + z^2,$$

avec $\mathbf{u} = \begin{pmatrix} x \\ y \\ z \end{pmatrix}$. Voyons si q est définie positive. Pour ce faire, décomposons q en somme de trois carrés dans \mathbb{R} :

$$\begin{aligned} q(\mathbf{u}) &= x^2 + 6xy + 4yz + 14y^2 + z^2 \\ &= (x + 3y)^2 - 9y^2 + 4yz + 14y^2 + z^2 = (x + 3y)^2 + 5y^2 + 4yz + z^2 \\ &= (x + 3y)^2 + 5\left(y + \frac{2}{5}z\right)^2 - \frac{4}{5}z^2 + z^2 = (x + 3y)^2 + 5\left(y + \frac{2}{5}z\right)^2 + \frac{1}{5}z^2. \end{aligned}$$

Cette somme de carrés dans \mathbb{R} est positive, donc la forme quadratique q est semi-définie positive ($\forall \mathbf{u} \in E, q(\mathbf{u}) \geq 0$). De plus :

$$\begin{aligned} q(\mathbf{u}) = (x + 3y)^2 + 5\left(y + \frac{2}{5}z\right)^2 + \frac{1}{5}z^2 = 0 &\Leftrightarrow \begin{cases} x + 3y = 0 \\ y + \frac{2}{5}z = 0 \\ z = 0 \end{cases} \\ &\Leftrightarrow x = y = z = 0 \\ &\Leftrightarrow \mathbf{u} = \mathbf{0}. \end{aligned}$$

Bilan : cette forme quadratique est bien définie positive, et la forme bilinéaire symétrique associée

$$\begin{aligned} f(\mathbf{x}, \mathbf{y}) &= x_1 y_1 + 3x_1 y_2 + 3x_2 y_1 + 2x_2 y_3 + 2x_3 y_2 + 14x_2 y_2 + x_3 y_3 \\ &= (x_1, x_2, x_3) \begin{pmatrix} 1 & 3 & 0 \\ 3 & 14 & 2 \\ 0 & 2 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} \end{aligned}$$

définit bien un produit scalaire sur \mathbb{R}^3 .

(ii) Soit la forme quadratique $q(\mathbf{x}) = x_1^2 - 2x_2^2 + 2x_2 x_1$, avec $\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$. Voyons si q est définie

positive. Un rapide coup d'oeil nous permet de penser que le terme en $-2x_2^2$, terme en carré à coefficient négatif, risque de poser problème quant à la définie positivité, ne serait-ce que parce qu'on peut l'isoler (ou le sélectionner) en prenant $x_1 = 0$. En effet, il est facile de vérifier que q est même **indéfinie**, c'est à dire qu'il existe des vecteurs \mathbf{x} pour lesquels $q(\mathbf{x}) > 0$, et des vecteurs \mathbf{y} pour lesquels $q(\mathbf{y}) < 0$. Par exemple, $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$ étant la base canonique de E ,

$$q(\mathbf{e}_1) = 1, \quad \text{et} \quad q(\mathbf{e}_2) = -2.$$

2.4 Diagonalisation des endomorphismes symétriques

2.4.1 Introduction

E étant un espace vectoriel euclidien, le produit scalaire sur E sera noté $\langle \mathbf{u}, \mathbf{v} \rangle$. Soit g un endomorphisme de E dont la matrice est symétrique dans la base canonique de E , $\{\mathbf{e}_1, \dots, \mathbf{e}_n\}$. Regardons si g est diagonalisable.

Prenons un exemple : Soit $E = \mathbb{R}^3$ muni du produit scalaire canonique, et g de matrice

$$\mathbf{A} = \begin{pmatrix} -1 & 1 & 1 \\ 1 & -1 & 1 \\ 1 & 1 & -1 \end{pmatrix}.$$

Les valeurs propres de g sont 1 et -2 de multiplicités respectives 1 et 2, les espaces propres associés étant :

$$V_1 = \text{Vect} \left\{ \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \right\} \quad \text{et} \quad V_{-2} = \text{Vect} \left\{ \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix} \right\}.$$

et g est donc diagonalisable.

On remarque que ces deux espaces V_1 et V_{-2} sont orthogonaux, c'est à dire tout vecteur de l'un est orthogonal à tout vecteur de l'autre. De plus, on peut choisir une base orthonormée pour écrire la matrice diagonale de g . Il suffit, dans un premier temps, d'orthogonaliser la base de V_{-2} , de dimension 2, en appliquant le procédé de SCHMIDT. On obtient :

$$V_{-2} = \text{Vect} \left\{ \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}, \begin{pmatrix} 1/2 \\ -1 \\ 1/2 \end{pmatrix} \right\}.$$

Enfin, il ne reste plus qu'à normaliser les vecteurs $\left\{ \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}, \begin{pmatrix} 1/2 \\ -1 \\ 1/2 \end{pmatrix} \right\}$.

Bilan : Dans la base orthonormée

$$\left\{ \frac{1}{\sqrt{3}} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}, \sqrt{\frac{2}{3}} \begin{pmatrix} 1/2 \\ -1 \\ 1/2 \end{pmatrix} \right\},$$

la matrice de l'endomorphisme g s'écrit :

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & -2 \end{pmatrix}.$$

2.4.2 Généralisation

Proposition 2.4.1

On démontre les résultats suivants :

- Tout endomorphisme symétrique d'un espace euclidien est diagonalisable.
- Ses valeurs propres sont réelles.
- Les espaces propres sont deux à deux orthogonaux.
- Il existe toujours une base orthonormée formée de vecteurs propres.

Remarques :

- Il est intéressant de diagonaliser dans une base orthonormée de vecteurs propres car alors, la matrice de passage \mathbf{U} de la base canonique initiale à la nouvelle base orthonormée vérifie

$$\mathbf{U}^{-1} = \mathbf{U}^T.$$

- Le fait que, dans un espace euclidien, tout endomorphisme symétrique se diagonalise dans une base orthonormale de vecteurs propres s'écrit en termes d'algèbre linéaire sous la forme :

$$\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T, \text{ avec } \mathbf{U}^T\mathbf{U} = \mathbf{U}\mathbf{U}^T = \mathbf{I} \text{ et } \mathbf{\Lambda} = \text{diag}(\lambda_i)_{1 \leq i \leq n}.$$

C'est d'ailleurs l'un des principaux intérêts des notations matricielles, à savoir d'exprimer de manière très concise des propriétés ou des transformations.

2.5 Diagonalisation d'une forme quadratique

On peut associer à toute forme quadratique q sur un \mathbb{R} -espace vectoriel euclidien E une forme bilinéaire symétrique f . De manière équivalente, cette forme bilinéaire symétrique peut être représentée sous forme matricielle par la matrice symétrique \mathbf{A} des coefficients $f(\mathbf{e}_i, \mathbf{e}_j)$, où les \mathbf{e}_k sont les vecteurs de la base canonique par exemple. De manière plus explicite, on a en effet :

$$\forall \mathbf{x}, \mathbf{y} \in E, \quad f(\mathbf{x}, \mathbf{y}) = \mathbf{X}^T \mathbf{A} \mathbf{Y},$$

\mathbf{X} et \mathbf{Y} étant les vecteurs des composantes de \mathbf{x} et \mathbf{y} dans la base $\mathcal{B} = (\mathbf{e}_k)_{1 \leq k \leq n}$.

La matrice \mathbf{A} étant symétrique, elle est diagonalisable dans une base orthonormée de vecteurs propres ($\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$, avec $\mathbf{U}^T = \mathbf{U}^{-1}$), et dans cette base de vecteurs propres, la matrice \mathbf{A} devenant $\mathbf{\Lambda} = \text{diag}(\lambda_i)_{1 \leq i \leq n}$, la forme quadratique q se transforme alors en somme élémentaire de carrés :

$$\forall \mathbf{x} \in E, \quad q(\mathbf{x}) = \mathbf{Z}^T \mathbf{\Lambda} \mathbf{Z} = \sum_{i=1}^n \lambda_i z_i^2,$$

où les z_i , $i = 1, \dots, n$, sont les composantes de \mathbf{x} dans la base des vecteurs propres :

$$\mathbf{x} = \sum_{i=1}^n z_i \mathbf{u}_i.$$

Cette dernière égalité peut aussi s'écrire matriciellement sous la forme :

$$\mathbf{X} = \mathbf{U}\mathbf{Z} = \mathbf{U}(\mathbf{U}^T \mathbf{X}),$$

avec $\mathbf{Z} = \mathbf{U}^T \mathbf{X}$ le vecteur des composantes z_i .

Remarques :

- Il est à noter que $z_i = \mathbf{u}_i^T \mathbf{X}$ n'est rien d'autre que le produit scalaire du $i^{\text{ème}}$ vecteur propre de \mathbf{A} (i.e. la $i^{\text{ème}}$ colonne de \mathbf{U}) avec le vecteur \mathbf{x} . Cela correspond au calcul des composantes d'un vecteur dans une base orthonormée donnée, que l'on obtient effectivement par produit scalaire avec les vecteurs de cette base.
- D'un point de vue géométrique, l'écriture de q sous la forme $\sum_{i=1}^n \lambda_i z_i^2$ signifie simplement que la forme quadratique q se décompose en paraboles élémentaires, dirigées selon les axes des vecteurs propres \mathbf{u}_i , et de courbures respectives λ_i .
- De manière équivalente, on peut aussi dire que les iso-contours $q(\mathbf{x}) = C^{\text{ste}}$ sont des coniques dans \mathbb{R}^n dont les axes principaux correspondent aux vecteurs propres de la matrice \mathbf{A} associée à la forme quadratique q .
- **Cas particulier :** si la forme quadratique q est définie positive, alors les valeurs propres λ_i ci-dessus sont nécessairement toutes strictement positives, et les iso-contours $q(\mathbf{x}) = C^{\text{ste}}$ correspondent alors à des hyper-ellipsoïdes dans \mathbb{R}^n .

Par exemple, $\lambda_1 z_1^2 + \lambda_2 z_2^2 = C$, avec $\lambda_1 > 0$ et $\lambda_2 > 0$, est l'équation d'une ellipse dans \mathbb{R}^2 , et l'équation

$$\lambda_1 z_1^2 + \lambda_2 z_2^2 + \lambda_3 z_3^2 = C,$$

avec $\lambda_{1,2,3}$ strictement positifs, représenterait une surface dans \mathbb{R}^3 du type “ballon de rugby”.

Illustration graphique : La figure ?? ci-dessous illustre la forme géométrique d'une nappe quadratique, à savoir le dessin dans \mathbb{R}^3 d'une forme quadratique de \mathbb{R}^2 dans \mathbb{R} , où (x, y) jouent le rôle de $\mathbf{x} \in \mathbb{R}^2$ et $z = \frac{1}{2} \mathbf{x}^T \mathbf{A} \mathbf{x}$ (avec \mathbf{A} matrice 2×2 symétrique définie positive).

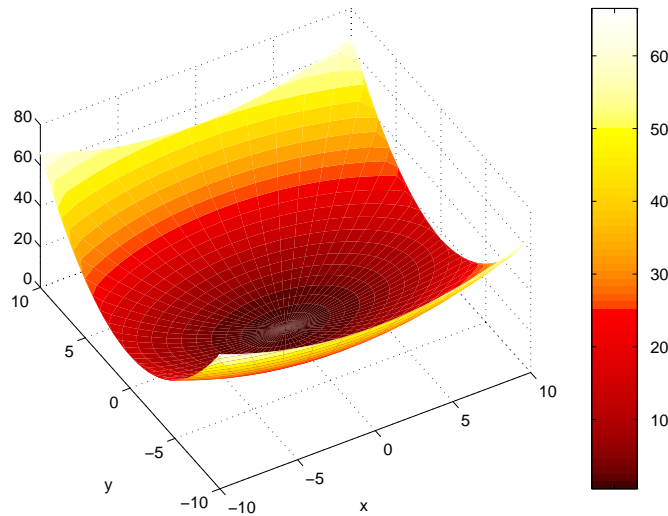


FIGURE 2.1 – Un exemple de forme quadratique en dimension 2

2.6 Compléments

Définition 2.6.1 – Fonctionnelle quadratique généralisée

On appelle **Fonctionnelle quadratique généralisée** toute application f de \mathbb{R}^n dans \mathbb{R} sous la forme :

$$\forall \mathbf{x} \in \mathbb{R}^n, f(\mathbf{x}) = \frac{1}{2} \mathbf{x}^T \mathbf{A} \mathbf{x} - \mathbf{b}^T \mathbf{x} + c,$$

où \mathbf{A} est une matrice de $\mathcal{M}_n(\mathbb{R})$, \mathbf{b} un vecteur de \mathbb{R}^n , et c une constante réelle.

On appelle **terme quadratique** associé à la fonctionnelle f le terme $\frac{1}{2} \mathbf{x}^T \mathbf{A} \mathbf{x}$.

Remarque : On peut toujours se ramener au cas où la matrice \mathbf{A} est symétrique, car on a :

$$\forall \mathbf{u} \in \mathbb{R}^n, \mathbf{u}^T \mathbf{A} \mathbf{u} = \mathbf{u}^T \left(\frac{\mathbf{A} + \mathbf{A}^T}{2} \right) \mathbf{u}.$$

Illustration graphique : Considérons le cas $n = 2$ et posons

$$Q = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix}, \quad A = Q \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} Q^T, \quad b = \begin{pmatrix} 1 & 2 \end{pmatrix}.$$

La figure ?? illustre les formes des courbes de niveaux de q pour les divers choix suivants pour les valeurs propres de A :

- (i) $\lambda_1 = 1, \lambda_2 = 3/2$ (la matrice A est définie positive) ;
- (ii) $\lambda_1 = 1, \lambda_2 = -3/2$ (la matrice A est indéfinie) ;
- (iii) $\lambda_1 = -1, \lambda_2 = -3/2$ (la matrice A est définie négative).

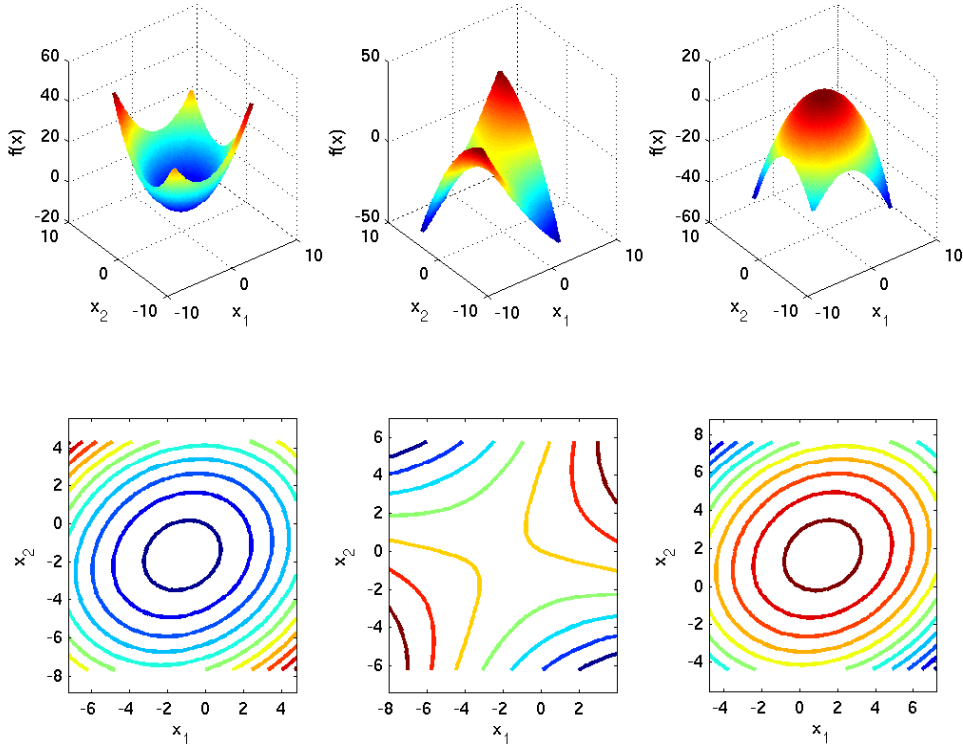
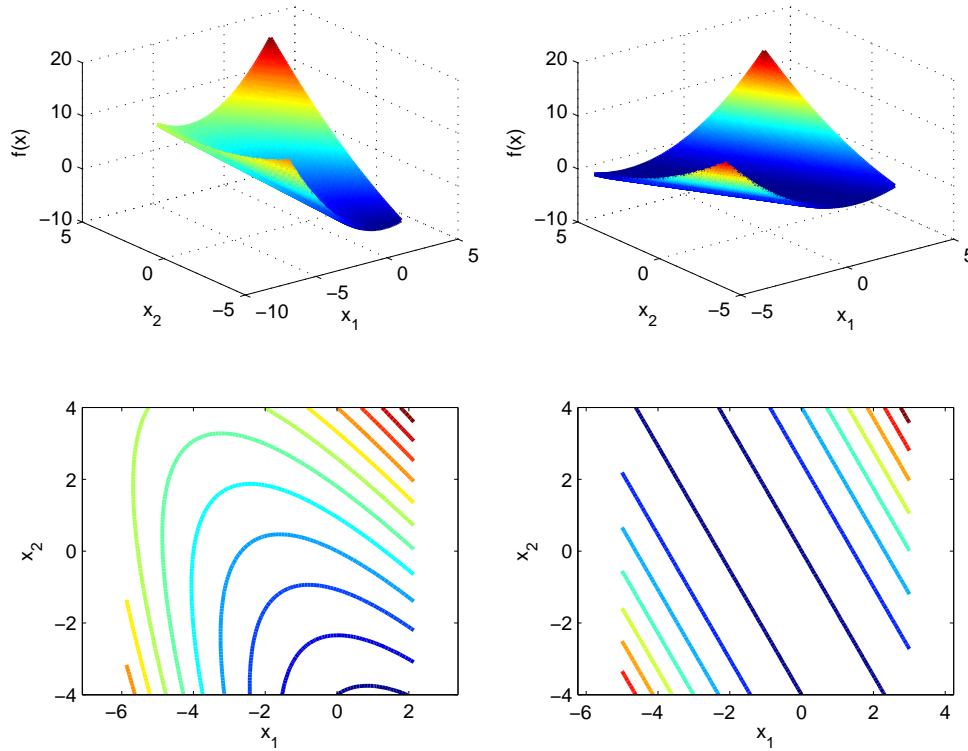


FIGURE 2.2 – Cas où $\text{rank}(A) = 2$ et $\theta = \pi/6$.

Supposons maintenant que le rang de A est 1 ($\lambda_2 = 0$ par exemple). La figure ?? illustre alors la forme des courbes de niveaux de q pour $\lambda_1 = 1, \lambda_2 = 0$ et $b = \begin{pmatrix} 1 & 2 \end{pmatrix}^T$ dans le premier cas, et $b = \begin{pmatrix} \cos(\theta) & \sin(\theta) \end{pmatrix}^T$ dans le deuxième cas (le vecteur b est dans l'image de la matrice A).

FIGURE 2.3 – Cas où $\text{rank}(A) = 1$ et $\theta = \pi/6$.**Définition 2.6.2**

Si la matrice \mathbf{A} introduite dans la définition précédente est **symétrique définie positive**, alors la fonctionnelle quadratique généralisée est dite **associée à une forme quadratique définie positive sur \mathbb{R}^n** .

Dans le cas des fonctionnelles quadratiques généralisées associées à une forme quadratique définie positive sur \mathbb{R}^n , on a le résultat suivant :

Proposition 2.6.3

Soit $f : \mathbb{R}^n \rightarrow \mathbb{R}$ une fonctionnelle quadratique généralisée associée à une forme quadratique définie positive sur \mathbb{R}^n . Alors :

- (i) f admet un minimum global unique sur \mathbb{R}^n , noté $\hat{\mathbf{x}}$.
- (ii) $\hat{\mathbf{x}}$ est l'unique solution du système linéaire $\mathbf{A}\mathbf{x} = \mathbf{b}$.
- (iii) $\hat{\mathbf{x}}$ est le centre de symétrie d'un réseau d'ellipsoïdes homothétiques (E_α) définis par :

$$E_\alpha = \{\mathbf{x} \in \mathbb{R}^n / f(\mathbf{x}) \leq \alpha\}, \quad \forall \alpha \geq f(\hat{\mathbf{x}}).$$

- (iv) Le minimum de f sur \mathbb{R}^n vaut

$$f(\hat{\mathbf{x}}) = -\frac{1}{2}\mathbf{b}^T \mathbf{A}^{-1} \mathbf{b} + c.$$

- (v) Pour tout vecteur $\mathbf{x} \in \mathbb{R}^n$ on a :

$$f(\mathbf{x}) = f(\hat{\mathbf{x}}) + \frac{1}{2}(\mathbf{x} - \hat{\mathbf{x}})^T \mathbf{A}(\mathbf{x} - \hat{\mathbf{x}}).$$

Remarque : On peut noter qu'il y a une certaine dualité entre la recherche du minimum d'une fonctionnelle quadratique associée à une forme quadratique définie positive et la résolution d'un système linéaire associé à une matrice symétrique définie positive.

Remarque : La matrice \mathbf{A} étant symétrique définie positive, elle définit une norme sur \mathbb{R}^n par l'égalité

$$\|\mathbf{x}\|_{\mathbf{A}} = \sqrt{\mathbf{x}^T \mathbf{A} \mathbf{x}} .$$

Cette norme est dite ellipsoïdale, car les iso-contours $\{\|\mathbf{x}\|_{\mathbf{A}} = C^{ste}\}$ forment des ellipsoïdes dans \mathbb{R}^n (c.f. figure ??). De plus, cette norme est associée au produit scalaire suivant :

$$(\mathbf{x}, \mathbf{y})_{\mathbf{A}} = \mathbf{x}^T \mathbf{A} \mathbf{y} .$$

Différentiabilité – Convexité

3.1 Dérivées de fonctions à plusieurs variables

Rappelons tout d'abord qu'une fonction d'une seule variable réelle à valeurs dans \mathbb{R} est dérivable en un point x de \mathbb{R} s'il existe un nombre réel a noté $f'(x)$ tel que :

$$\lim_{t \rightarrow 0} \frac{f(x+t) - f(x) - at}{t} = 0 .$$

Cette définition s'étend de façon naturelle dans le cas de fonctions de n variables réelles, et de manière plus générale dans le cas de fonctions définies sur un espace vectoriel normé E et à valeurs dans un espace vectoriel normé F .

3.1.1 Dérivée première

Définition 3.1.1 – Différentiabilité au sens de Frêchet

Soient E et F deux espaces vectoriels normés. Soit f une application définie sur le domaine $D \subset E$ et à valeurs dans F . L'application f est dite **F-différentiable** (ou différentiable au sens de Frêchet, ou encore différentiable au sens fort) en un point \mathbf{x} de l'intérieur du domaine D , s'il existe un opérateur linéaire continu $f'(\mathbf{x})$ de E dans F ($f'(\mathbf{x}) \in \mathcal{L}(E, F)$), tel que

$$\forall \mathbf{h} \in E , \quad f(\mathbf{x} + \mathbf{h}) = f(\mathbf{x}) + f'(\mathbf{x}) \cdot \mathbf{h} + \|\mathbf{h}\|_E \varepsilon(\mathbf{h}) , \quad \text{avec} \quad \lim_{\|\mathbf{h}\|_E \rightarrow 0} \|\varepsilon(\mathbf{h})\|_F = 0 . \quad (3.1)$$

Remarque : On dira aussi que f est dérivable au point \mathbf{x} sans autre précision pour signifier qu'elle est différentiable au sens de Frêchet (au sens fort).

Proposition 3.1.2

La F -dérivée de f au point \mathbf{x} (qui correspond à l'opérateur $f'(\mathbf{x}) \in \mathcal{L}(E, F)$ dans la définition ci-dessus), si elle existe, est unique.

Proposition 3.1.3

Si l'application f est F -différentiable (dérivable) au point \mathbf{x} , elle est alors continue au point \mathbf{x} .

Définition 3.1.4

Soient E et F deux espaces vectoriels normés, et $\Omega \subset E$ un ouvert de E . On dit que l'**application** $f : \Omega \subset E \rightarrow F$ est **dérivable dans** Ω si elle est dérivable en tout point \mathbf{x} de Ω . On peut alors définir l'application

$$f' : \mathbf{x} \in \Omega \subset E \rightarrow f'(\mathbf{x}) \in \mathcal{L}(E, F) ,$$

appelée **application dérivée de** f . Si l'application dérivée $f' : \Omega \subset E \rightarrow \mathcal{L}(E, F)$ est continue, on dit que l'**application** $f : \Omega \subset E \rightarrow F$ est **(une fois) continûment dérivable dans** Ω , et on écrit

$$f \in \mathcal{C}^1(\Omega) .$$

Théorème 3.1.5 – Différentiabilité des applications composées

Soient E, F , et G , trois espaces vectoriels normés. Soit $f : \Omega \subset E \rightarrow F$ une application dérivable en un point $\mathbf{x} \in \Omega$ (Ω ouvert de E), et soit $g : \widetilde{\Omega} \subset F \rightarrow G$ une application dérivable au point $\mathbf{y} = f(\mathbf{x}) \in \widetilde{\Omega}$ ($\widetilde{\Omega}$ ouvert de F). On suppose $f(\Omega) \subset \widetilde{\Omega}$. Alors l'application composée

$$g \circ f : \Omega \subset E \rightarrow G$$

est dérivable au point $\mathbf{x} \in \Omega$ et

$$\forall \mathbf{h} \in E, \quad (g \circ f)'(\mathbf{x}) \cdot \mathbf{h} = g'(f(\mathbf{x})) \cdot (f'(\mathbf{x}) \cdot \mathbf{h}).$$

3.1.2 Dérivée seconde**Définition 3.1.6**

Soit $f : \Omega \subset E \rightarrow F$ une application dérivable sur l'ouvert $\Omega \subset E$. Si l'application dérivée

$$f' : \Omega \subset E \rightarrow \mathcal{L}(E, F)$$

est elle-même dérivable (i.e. F -différentiable) en un point $\mathbf{x} \in \Omega$, sa dérivée, notée

$$f''(\mathbf{x}) \stackrel{\text{déf}}{=} (f')'(\mathbf{x}) \in \mathcal{L}(E, \mathcal{L}(E, F)) ,$$

est appelée **dérivée seconde de l'application f au point \mathbf{x}** , et on dit que **l'application f est deux fois dérivable au point \mathbf{x}** .

Notation : Il est facile de remarquer que l'application

$$\mathbf{B} : (\mathbf{h}, \mathbf{k}) \in E \times E \rightarrow ((f''(\mathbf{x}) \cdot \mathbf{h}) \cdot \mathbf{k}) \in F ,$$

est linéaire séparément en chacune des variables \mathbf{h} et \mathbf{k} , et est de ce fait **bilinéaire**. En d'autres termes, il existe une interprétation naturelle permettant d'identifier l'application dérivée seconde de f au point \mathbf{x} , $f''(\mathbf{x}) \in \mathcal{L}(E, \mathcal{L}(E, F))$, à une application de l'espace $\mathcal{L}_2(E \times E, F)$, espace des applications bilinéaires continues de $E \times E$ dans F . On écrira alors

$$\forall \mathbf{h}, \mathbf{k} \in E, \quad f''(\mathbf{x})(\mathbf{h}, \mathbf{k}) = (f''(\mathbf{x}) \cdot \mathbf{h}) \cdot \mathbf{k}.$$

Proposition 3.1.7

Si l'application $f : \Omega \subset E \rightarrow F$ est deux fois dérivable au point \mathbf{x} de l'ouvert $\Omega \subset E$, alors l'application dérivée seconde de f au point \mathbf{x} est une **application bilinéaire symétrique** en ce sens que

$$\forall \mathbf{h}, \mathbf{k} \in E, \quad f''(\mathbf{x})(\mathbf{h}, \mathbf{k}) = f''(\mathbf{x})(\mathbf{k}, \mathbf{h}).$$

Définition 3.1.8

On dit que **l'application $f : \Omega \subset E \rightarrow F$ est deux fois dérivable dans Ω** si elle est deux fois dérivable en tout point \mathbf{x} de Ω . On peut alors définir **l'application dérivée seconde de f**

$$f'' : \mathbf{x} \in \Omega \subset E \rightarrow f''(\mathbf{x}) \in \mathcal{L}_2(E \times E, F) .$$

Si cette dernière application est continue, l'application f est dite **deux fois continûment dérivable dans Ω** , et on écrit

$$f \in \mathcal{C}^2(\Omega) .$$

Remarque : En ce qui concerne le **calcul** effectif des dérivées secondes, on utilise le résultat suivant, qui permet de se ramener à des calculs de dérivées premières :
 étant donné deux vecteurs quelconques $\mathbf{h}, \mathbf{k} \in E$, l'élément $f''(\mathbf{x})(\mathbf{h}, \mathbf{k}) \in F$ est égal à la dérivée au point $\mathbf{x} \in \Omega$ de l'application $\mathbf{v} \in \Omega \rightarrow f'(\mathbf{v}) \cdot \mathbf{k} \in F$, appliquée au vecteur \mathbf{h} .

3.1.3 Formule des accroissements finis - Formules de Taylor

Théorème 3.1.9 – Formules de Taylor pour les applications une fois dérivables

Soient $f : \Omega \subset E \rightarrow F$ et $[\mathbf{a}, \mathbf{a} + \mathbf{h}]$ un segment fermé contenu dans Ω (Ω ouvert de E).

(i) Si f est dérivable en \mathbf{a} , alors

$$f(\mathbf{a} + \mathbf{h}) = f(\mathbf{a}) + f'(\mathbf{a}) \cdot \mathbf{h} + \|\mathbf{h}\|_E \varepsilon(\mathbf{h}) \quad , \quad \text{avec} \quad \lim_{\|\mathbf{h}\|_E \rightarrow 0} \|\varepsilon(\mathbf{h})\|_F = 0 \quad .$$

(ii) **Formule des accroissements finis :** si $f \in \mathcal{C}^0(\Omega)$ et f est dérivable en tout point du segment ouvert $] \mathbf{a}, \mathbf{a} + \mathbf{h} [$, alors

$$\|f(\mathbf{a} + \mathbf{h}) - f(\mathbf{a})\|_F \leq \sup_{\mathbf{x} \in] \mathbf{a}, \mathbf{a} + \mathbf{h} [} \|f'(\mathbf{x})\|_{\mathcal{L}(E, F)} \|\mathbf{h}\|_E \quad .$$

(iii) **Formule de Taylor-Maclaurin :** si $f \in \mathcal{C}^0(\Omega)$ et f est dérivable en tout point du segment ouvert $] \mathbf{a}, \mathbf{a} + \mathbf{h} [$, et si $F = \mathbb{R}$, alors

$$\exists \theta \in]0, 1[\quad \text{tel que} \quad f(\mathbf{a} + \mathbf{h}) = f(\mathbf{a}) + f'(\mathbf{a} + \theta \mathbf{h}) \cdot \mathbf{h} \quad .$$

(iv) **Formule de Taylor avec reste intégral :** si $f \in \mathcal{C}^1(\Omega)$ et F est un espace complet, alors

$$f(\mathbf{a} + \mathbf{h}) = f(\mathbf{a}) + \int_0^1 \{f'(\mathbf{a} + t\mathbf{h}) \cdot \mathbf{h}\} dt \quad .$$

Théorème 3.1.10 – Formules de Taylor pour les applications deux fois dérivables

Soit $f : \Omega \subset E \rightarrow F$ et $[\mathbf{a}, \mathbf{a} + \mathbf{h}]$ un segment fermé contenu dans Ω (Ω ouvert de E).

(i) **Formule de Taylor-Young :** si f est dérivable dans Ω , et si f est deux fois dérivable au point \mathbf{a} , alors

$$f(\mathbf{a} + \mathbf{h}) = f(\mathbf{a}) + f'(\mathbf{a}) \cdot \mathbf{h} + \frac{1}{2} f''(\mathbf{a})(\mathbf{h}, \mathbf{h}) + \|\mathbf{h}\|_E^2 \varepsilon(\mathbf{h}) \quad , \quad \text{avec} \quad \lim_{\|\mathbf{h}\|_E \rightarrow 0} \|\varepsilon(\mathbf{h})\|_F = 0 \quad .$$

(ii) **Formule des accroissements finis généralisée :** si $f \in \mathcal{C}^1(\Omega)$ et f est deux fois dérivable en tout point du segment ouvert $] \mathbf{a}, \mathbf{a} + \mathbf{h} [$, alors

$$\|f(\mathbf{a} + \mathbf{h}) - f(\mathbf{a}) - f'(\mathbf{a}) \cdot \mathbf{h}\|_F \leq \frac{1}{2} \sup_{\mathbf{x} \in] \mathbf{a}, \mathbf{a} + \mathbf{h} [} \|f''(\mathbf{x})\|_{\mathcal{L}_2(E \times E, F)} \|\mathbf{h}\|_E^2 \quad .$$

(iii) **Formule de Taylor-Maclaurin :** si $f \in \mathcal{C}^1(\Omega)$ et f est deux fois dérivable en tout point du segment ouvert $] \mathbf{a}, \mathbf{a} + \mathbf{h} [$, et si $F = \mathbb{R}$, alors

$$\exists \theta \in]0, 1[\quad \text{tel que} \quad f(\mathbf{a} + \mathbf{h}) = f(\mathbf{a}) + f'(\mathbf{a}) \cdot \mathbf{h} + \frac{1}{2} f''(\mathbf{a} + \theta \mathbf{h})(\mathbf{h}, \mathbf{h}) \quad .$$

(iv) **Formule de Taylor avec reste intégral :** si $f \in \mathcal{C}^2(\Omega)$ et F est un espace complet, alors

$$f(\mathbf{a} + \mathbf{h}) = f(\mathbf{a}) + f'(\mathbf{a}) \cdot \mathbf{h} + \int_0^1 (1-t) \{f''(\mathbf{a} + t\mathbf{h})(\mathbf{h}, \mathbf{h})\} dt \quad .$$

Remarques :

- Alors que la formule (i) du théorème ?? est exactement la définition de la différentiabilité première, la formule (i) du théorème ?? n'est pas égale à la définition de la différentiabilité seconde en un point.
- On sait qu'il existe au moins un nombre $\theta \in]0, 1[$ tel que les formules de Taylor-Maclaurin soient vraies, mais en général on n'a aucun autre renseignement sur θ ; on rappelle au passage qu'il est indispensable de se restreindre au cas $F = \mathbb{R}$ pour les formules (3).
- Pour que les formules (iv) aient un sens, il faut savoir intégrer les fonctions impliquées dans ces formules, et c'est pourquoi on suppose que ces fonctions sont continues et que l'espace F est complet.

3.1.4 Dimension finie et dérivées partielles

La donnée d'une application

$$f : \Omega \subset E \rightarrow F = \mathbb{R}^p$$

revient à se donner p applications composantes $f_i : \Omega \subset E \rightarrow \mathbb{R}$, $1 \leq i \leq p$, de telle façon que

$$\forall \mathbf{x} \in E, \quad f(\mathbf{x}) = \begin{pmatrix} f_1(\mathbf{x}) \\ f_2(\mathbf{x}) \\ \vdots \\ f_p(\mathbf{x}) \end{pmatrix}.$$

Proposition 3.1.11

On établit facilement que l'application f est dérivable en un point $\mathbf{a} \in \Omega$ si et seulement si chaque application composante l'est aussi, et on a alors :

$$f'(\mathbf{a}) = \begin{pmatrix} f'_1(\mathbf{a}) \\ f'_2(\mathbf{a}) \\ \vdots \\ f'_p(\mathbf{a}) \end{pmatrix} \quad \text{avec } f'_i(\mathbf{a}) \in \mathcal{L}(E, \mathbb{R}), \quad 1 \leq i \leq p.$$

Considérons ensuite une application

$$f : \Omega \subset \mathbb{R}^n \rightarrow \mathbb{R}$$

où Ω est un ouvert de \mathbb{R}^n (on dira aussi que f est une **fonctionnelle**). Soit \mathbf{a} un point de Ω de composantes (a_1, a_2, \dots, a_n) , et soit $k \in \{1, 2, \dots, n\}$ l'un des indices.

Par définition de la topologie produit, il existe un ouvert $\Omega_k \subset \mathbb{R}$ tel que tous les points de composantes $(a_1, \dots, a_{k-1}, x_k, a_{k+1}, \dots, a_n)$ appartiennent à l'ouvert Ω lorsque le réel x_k appartient à l'ouvert Ω_k . Par suite, on peut étudier la dérivabilité éventuelle sur $\Omega_k \subset \mathbb{R}$ de l'**application partielle**

$$\begin{aligned} \Omega_k \subset \mathbb{R} &\longrightarrow \mathbb{R} \\ x_k &\longrightarrow f(a_1, \dots, a_{k-1}, x_k, a_{k+1}, \dots, a_n) \end{aligned}$$

Si cette application est dérivable (au sens classique des fonctions de \mathbb{R} dans \mathbb{R}) au point $a_k \in \Omega_k \subset \mathbb{R}$, on note

$$\frac{\partial f}{\partial x_k}(\mathbf{a}) \in \mathbb{R}$$

sa dérivée, appelée **dérivée partielle de la fonctionnelle f au point \mathbf{a} , par rapport à la k -ième variable**.

Remarque : L'espace $\mathcal{L}(\mathbb{R}, \mathbb{R})$ (matrices 1×1) s'identifiant à \mathbb{R} , les dérivées partielles $\frac{\partial f}{\partial x_k}(\mathbf{a})$ de la fonctionnelle $f : \Omega \subset \mathbb{R}^n \rightarrow \mathbb{R}$ peuvent être effectivement considérées comme des nombres réels.

Proposition 3.1.12

Si une fonctionnelle

$$f : \Omega \subset \mathbb{R}^n \rightarrow \mathbb{R}$$

est dérivable en un point $\mathbf{a} \in \Omega$, on établit aisément qu'elle possède des dérivées partielles par rapport à chacune des variables et que, de plus,

$$f'(\mathbf{a}) \cdot \mathbf{h} = \sum_{k=1}^n \frac{\partial f}{\partial x_k}(\mathbf{a}) h_k, \text{ pour tout } \mathbf{h} = \begin{pmatrix} h_1 \\ h_2 \\ \vdots \\ h_n \end{pmatrix} \in \mathbb{R}^n \quad (3.2)$$



La réciproque est par contre inexacte. On a cependant le résultat suivant :

Proposition 3.1.13

Une fonctionnelle $f : \Omega \subset \mathbb{R}^n \rightarrow \mathbb{R}$ est continûment dérivable dans Ω si et seulement si les dérivées partielles $\frac{\partial f}{\partial x_k}(\mathbf{a})$, $1 \leq k \leq n$, existent en tout point $\mathbf{a} \in \Omega$ et si les applications dérivées partielles :

$$\frac{\partial f}{\partial x_k} : \mathbf{a} \in \Omega \longrightarrow \frac{\partial f}{\partial x_k}(\mathbf{a}) \in \mathbb{R}$$

sont continues dans Ω .

Supposons enfin que $E = \mathbb{R}^n$ et que $F = \mathbb{R}^p$, de sorte qu'une application $f : \Omega \subset \mathbb{R}^n \rightarrow \mathbb{R}^p$ est déterminée par la donnée de p fonctionnelles $f_i : \Omega \subset \mathbb{R}^n \rightarrow \mathbb{R}$, $1 \leq i \leq p$, de n variables chacune. Alors la relation

$$\mathbf{k} = f'(\mathbf{a}) \cdot \mathbf{h}, \text{ avec } \mathbf{h} = \begin{pmatrix} h_1 \\ h_2 \\ \vdots \\ h_n \end{pmatrix} \in \mathbb{R}^n \text{ et } \mathbf{k} = \begin{pmatrix} k_1 \\ k_2 \\ \vdots \\ k_p \end{pmatrix} \in \mathbb{R}^p$$

s'écrit sous la forme matricielle suivante

$$\begin{pmatrix} k_1 \\ k_2 \\ \vdots \\ k_p \end{pmatrix} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(\mathbf{a}) & \frac{\partial f_1}{\partial x_2}(\mathbf{a}) & \cdots & \frac{\partial f_1}{\partial x_n}(\mathbf{a}) \\ \frac{\partial f_2}{\partial x_1}(\mathbf{a}) & \frac{\partial f_2}{\partial x_2}(\mathbf{a}) & \cdots & \frac{\partial f_2}{\partial x_n}(\mathbf{a}) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_p}{\partial x_1}(\mathbf{a}) & \frac{\partial f_p}{\partial x_2}(\mathbf{a}) & \cdots & \frac{\partial f_p}{\partial x_n}(\mathbf{a}) \end{pmatrix} \begin{pmatrix} h_1 \\ h_2 \\ \vdots \\ h_n \end{pmatrix}.$$

La matrice “ p lignes, n colonnes” ci-dessus représente donc l'application linéaire

$$f'(\mathbf{a}) \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^p).$$

En ce qui concerne l'expression de la matrice dérivée d'applications composées en dimension finie, on a le résultat suivant en relation avec le théorème ?? :

Proposition 3.1.14

Soit $f : \Omega \subset \mathbb{R}^n \rightarrow \mathbb{R}^m$ une application dérivable en un point $\mathbf{a} \in \Omega$ (Ω ouvert de \mathbb{R}^n), et soit $g : \widetilde{\Omega} \subset \mathbb{R}^m \rightarrow \mathbb{R}^p$ une application dérivable au point $\mathbf{b} = f(\mathbf{a}) \in \widetilde{\Omega}$ ($\widetilde{\Omega}$ ouvert de \mathbb{R}^m). On suppose $\Omega \subset f(\Omega)$. Alors l'application composée

$$h = g \circ f : \Omega \subset \mathbb{R}^n \rightarrow \mathbb{R}^p$$

est dérivable au point $\mathbf{a} \in \Omega$ et la dérivée de l'application h au point \mathbf{a} s'exprime matriciellement de la façon suivante :

$$\begin{pmatrix} \frac{\partial h_1}{\partial x_1}(\mathbf{a}) & \cdots & \frac{\partial h_1}{\partial x_n}(\mathbf{a}) \\ \vdots & \ddots & \vdots \\ \frac{\partial h_p}{\partial x_1}(\mathbf{a}) & \cdots & \frac{\partial h_p}{\partial x_n}(\mathbf{a}) \end{pmatrix} = \begin{pmatrix} \frac{\partial g_1}{\partial y_1}(\mathbf{b}) & \cdots & \frac{\partial g_1}{\partial y_m}(\mathbf{b}) \\ \vdots & \ddots & \vdots \\ \frac{\partial g_p}{\partial y_1}(\mathbf{b}) & \cdots & \frac{\partial g_p}{\partial y_m}(\mathbf{b}) \end{pmatrix} \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(\mathbf{a}) & \cdots & \frac{\partial f_1}{\partial x_n}(\mathbf{a}) \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1}(\mathbf{a}) & \cdots & \frac{\partial f_m}{\partial x_n}(\mathbf{a}) \end{pmatrix}.$$

Pour terminer, précisons quelques *notations* particulières aux fonctionnelles $f : \Omega \subset \mathbb{R}^n \rightarrow \mathbb{R}$. En tout point $\mathbf{a} \in \Omega$ où cette application est une fois, ou deux fois, dérivable, on introduit le *vecteur* $\nabla f(\mathbf{a}) \in \mathbb{R}^n$ et la *matrice* $\nabla^2 f(\mathbf{a}) \in \mathcal{M}_n(\mathbb{R})$ définis respectivement par les relations

$$f'(\mathbf{a}) \cdot \mathbf{h} = \langle \nabla f(\mathbf{a}), \mathbf{h} \rangle \quad \text{pour tout } \mathbf{h} \in \mathbb{R}^n,$$

$$f''(\mathbf{a})(\mathbf{h}, \mathbf{k}) = \langle \nabla^2 f(\mathbf{a})\mathbf{h}, \mathbf{k} \rangle = \langle \mathbf{h}, \nabla^2 f(\mathbf{a})\mathbf{k} \rangle \quad \text{pour tout } \mathbf{h}, \mathbf{k} \in \mathbb{R}^n.$$

$\langle \cdot, \cdot \rangle$ désignant comme d'habitude le produit scalaire euclidien sur \mathbb{R}^n . Le vecteur

$$\nabla f(\mathbf{a}) = \begin{pmatrix} \frac{\partial f}{\partial x_1}(\mathbf{a}) \\ \vdots \\ \frac{\partial f}{\partial x_n}(\mathbf{a}) \end{pmatrix}$$

s'appelle le **gradient** de l'application f au point \mathbf{a} , et la matrice (symétrique)

$$\nabla^2 f(\mathbf{a}) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1}(\mathbf{a}) & \frac{\partial^2 f}{\partial x_1 \partial x_2}(\mathbf{a}) & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n}(\mathbf{a}) \\ \frac{\partial^2 f}{\partial x_2 \partial x_1}(\mathbf{a}) & \frac{\partial^2 f}{\partial x_2 \partial x_2}(\mathbf{a}) & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n}(\mathbf{a}) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1}(\mathbf{a}) & \frac{\partial^2 f}{\partial x_n \partial x_2}(\mathbf{a}) & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_n}(\mathbf{a}) \end{pmatrix}$$

s'appelle le **Hessien** de l'application f au point \mathbf{a} .

Pour illustrer ces considérations, voici trois façons équivalentes d'écrire (par exemple) la formule de Taylor-Young pour une fonctionnelle $f : \Omega \subset \mathbb{R}^n \rightarrow \mathbb{R}$ deux fois dérivable :

$$f(\mathbf{a} + \mathbf{h}) = f(\mathbf{a}) + f'(\mathbf{a}) \cdot \mathbf{h} + \frac{1}{2} f''(\mathbf{a})(\mathbf{h}, \mathbf{h}) + \|\mathbf{h}\|_2^2 \varepsilon(\mathbf{h})$$

$$f(\mathbf{a} + \mathbf{h}) = f(\mathbf{a}) + \langle \nabla f(\mathbf{a}), \mathbf{h} \rangle + \frac{1}{2} \langle \nabla^2 f(\mathbf{a})\mathbf{h}, \mathbf{h} \rangle + \langle \mathbf{h}, \mathbf{h} \rangle \varepsilon(\mathbf{h})$$

$$f(\mathbf{a} + \mathbf{h}) = f(\mathbf{a}) + (\nabla f(\mathbf{a}))^T \mathbf{h} + \frac{1}{2} \mathbf{h}^T \nabla^2 f(\mathbf{a}) \mathbf{h} + \mathbf{h}^T \mathbf{h} \varepsilon(\mathbf{h}).$$

3.2 Convexité des fonctionnelles

3.2.1 Ensembles convexes - fonctionnelles convexes

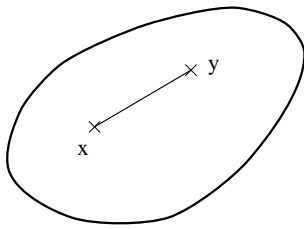
On indique dans ce paragraphe quelques propriétés de base d'une classe très importante de fonctionnelles.

Définition 3.2.1 – Ensembles convexes

L'ensemble D_0 est dit **convexe** si et seulement si

$$\forall \mathbf{x} \in D_0, \forall \mathbf{y} \in D_0, \forall \alpha \in [0, 1] \subset \mathbb{R} \text{ on a } \alpha \mathbf{x} + (1 - \alpha) \mathbf{y} \in D_0 .$$

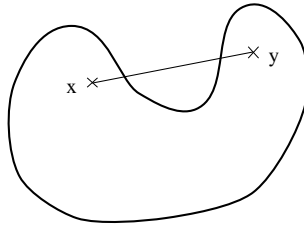
Remarque :



autrement dit, si $\mathbf{x} \in D_0$ et $\mathbf{y} \in D_0$, alors le segment qui joint ces deux points est également contenu dans D_0 , le segment $[\mathbf{x}, \mathbf{y}]$ étant défini par

$$\mathbf{z} \in [\mathbf{x}, \mathbf{y}] \iff \exists \alpha \in [0, 1] \text{ t.q. } \mathbf{z} = \alpha \mathbf{x} + (1 - \alpha) \mathbf{y} .$$

Exemple d'ensemble non convexe



Remarque : la notion d'ensemble convexe correspond en fait à une propriété de régularité du domaine D_0 considéré

Définition 3.2.2 – Fonctionnelles convexes

Une fonctionnelle $f : D_0 \subset E \rightarrow \mathbb{R}$ est **convexe** sur le domaine convexe $D_0 \subset E$ (E espace vectoriel normé) si

$$\forall \mathbf{x}, \mathbf{y} \in D_0, \forall \alpha \in [0, 1], \quad f(\alpha \mathbf{x} + (1 - \alpha) \mathbf{y}) \leq \alpha f(\mathbf{x}) + (1 - \alpha) f(\mathbf{y}) .$$

La fonctionnelle f est **strictement convexe** sur le domaine convexe D_0 si

$$\forall \mathbf{x}, \mathbf{y} \in D_0, \mathbf{x} \neq \mathbf{y}, \forall \alpha \in]0, 1[, \quad f(\alpha \mathbf{x} + (1 - \alpha) \mathbf{y}) < \alpha f(\mathbf{x}) + (1 - \alpha) f(\mathbf{y}) .$$

La fonctionnelle f est **uniformément convexe** sur le domaine convexe D_0 si il existe une constante $c > 0$ telle que

$$\begin{aligned} \forall \mathbf{x}, \mathbf{y} \in D_0, \forall \alpha \in]0, 1[, \\ \alpha f(\mathbf{x}) + (1 - \alpha) f(\mathbf{y}) - f(\alpha \mathbf{x} + (1 - \alpha) \mathbf{y}) \geq c \alpha (1 - \alpha) \|\mathbf{x} - \mathbf{y}\|_E^2 . \end{aligned}$$

Remarques :

- (i) Il est clair que la *convexité uniforme* entraîne la *convexité stricte* qui à son tour entraîne la *convexité*.
- (ii) La convexité indique une certaine régularité de la fonctionnelle. En dimension finie, par exemple, la convexité peut induire des propriétés de continuité (c.f. proposition suivante).

Proposition 3.2.3

Soit $f : D_0 \subset \mathbb{R}^n \rightarrow \mathbb{R}$ une fonctionnelle convexe sur l'ouvert convexe $D_0 \subset \mathbb{R}^n$. Alors f est continue sur D_0 .

Les définitions de base de la convexité (large, stricte, ou uniforme) peuvent parfois s'avérer d'un emploi peu commode. Le but des paragraphes qui suivent est de mettre en avant des propriétés qui s'y rapportent, exploitant la différentiabilité d'une fonctionnelle, et plus faciles à manipuler.

3.2.2 Convexité et dérivée première**Théorème 3.2.4 – Caractérisation de la convexité à l'aide de la dérivée première**

On suppose que la fonctionnelle $f : \Omega \subset E \rightarrow \mathbb{R}$ est dérivable sur un sous-ensemble convexe $D_0 \subset \Omega$. On a alors :

(i) f est convexe sur D_0 si et seulement si

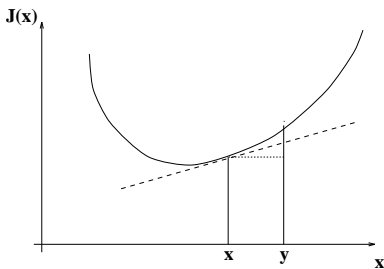
$$\forall \mathbf{x}, \mathbf{y} \in D_0, \quad f(\mathbf{y}) - f(\mathbf{x}) \geq f'(\mathbf{x}) \cdot (\mathbf{y} - \mathbf{x}).$$

(ii) f est strictement convexe sur D_0 si et seulement si

$$\forall \mathbf{x}, \mathbf{y} \in D_0, \quad \mathbf{x} \neq \mathbf{y}, \quad f(\mathbf{y}) - f(\mathbf{x}) > f'(\mathbf{x}) \cdot (\mathbf{y} - \mathbf{x}).$$

(iii) La fonctionnelle f est uniformément convexe sur D_0 si et seulement si il existe une constante $c > 0$ telle que

$$\forall \mathbf{x}, \mathbf{y} \in D_0, \quad f(\mathbf{y}) - f(\mathbf{x}) \geq f'(\mathbf{x}) \cdot (\mathbf{y} - \mathbf{x}) + c \|\mathbf{y} - \mathbf{x}\|_E^2.$$

Interprétation géométrique

L'interprétation géométrique de

$$\forall \mathbf{x}, \mathbf{y} \in D_0, \quad f(\mathbf{y}) - f(\mathbf{x}) \geq f'(\mathbf{x}) \cdot (\mathbf{y} - \mathbf{x}),$$

est que le graphe de la fonctionnelle convexe f est toujours au dessus de son plan tangent en un point quelconque du domaine D_0 .

Définition 3.2.5

Soit une fonctionnelle $f : \Omega \subset E \rightarrow \mathbb{R}$ dérivable sur l'ouvert Ω .

L'application dérivée $f' : \Omega \subset E \rightarrow \mathcal{L}(E, \mathbb{R})$ est dite **monotone sur le sous-ensemble** $D_0 \subset \Omega$ si et seulement si

$$\forall \mathbf{x}, \mathbf{y} \in D_0, \quad (f'(\mathbf{y}) - f'(\mathbf{x})) \cdot (\mathbf{y} - \mathbf{x}) \geq 0.$$

L'application dérivée f' est dite **strictement monotone sur le sous-ensemble** $D_0 \subset \Omega$ si et seulement si

$$\forall \mathbf{x}, \mathbf{y} \in D_0, \quad \mathbf{x} \neq \mathbf{y}, \quad (f'(\mathbf{y}) - f'(\mathbf{x})) \cdot (\mathbf{y} - \mathbf{x}) > 0.$$

L'application dérivée f' est dite **fortement monotone sur le sous-ensemble** $D_0 \subset \Omega$ si et seulement si il existe une constante $c > 0$ telle que

$$\forall \mathbf{x}, \mathbf{y} \in D_0, \quad (f'(\mathbf{y}) - f'(\mathbf{x})) \cdot (\mathbf{y} - \mathbf{x}) \geq 2c \|\mathbf{y} - \mathbf{x}\|_E^2.$$

Proposition 3.2.6 – Relations entre convexité et monotonie de la dérivée première

On suppose que la fonctionnelle $f : \Omega \subset E \rightarrow \mathbb{R}$ est dérivable sur Ω . On a alors :

- (i) La fonctionnelle f est convexe sur le sous-ensemble convexe $D_0 \subset \Omega$ si et seulement si l'application dérivée f' est monotone sur D_0 .
- (ii) La fonctionnelle f est strictement convexe sur le sous-ensemble convexe $D_0 \subset \Omega$ si et seulement si l'application dérivée f' est strictement monotone sur D_0 .
- (iii) La fonctionnelle f est uniformément convexe sur le sous-ensemble convexe $D_0 \subset \Omega$ si et seulement si l'application dérivée f' est fortement monotone sur D_0 (la constante $c > 0$ intervenant dans la définition de la convexité uniforme correspondant à la constante $c > 0$ introduite dans la définition de la forte monotonie de la dérivée).

3.2.3 Convexité et dérivée seconde**Théorème 3.2.7 – Relations entre convexité et positivité de la dérivée seconde**

On suppose que la fonctionnelle $f : \Omega \subset E \rightarrow \mathbb{R}$ est deux fois dérivable dans un ouvert Ω de l'espace vectoriel normé E , et soit D_0 une partie convexe de Ω .

- (i) La fonctionnelle f est convexe sur le sous-ensemble convexe $D_0 \subset \Omega$ si et seulement si

$$\forall \mathbf{x}, \mathbf{y} \in D_0, \quad f''(\mathbf{x})(\mathbf{y} - \mathbf{x}, \mathbf{y} - \mathbf{x}) \geq 0.$$

- (ii) Si

$$\forall \mathbf{x}, \mathbf{y} \in D_0, \quad \mathbf{x} \neq \mathbf{y}, \quad f''(\mathbf{x})(\mathbf{y} - \mathbf{x}, \mathbf{y} - \mathbf{x}) > 0,$$

alors la fonctionnelle f est strictement convexe sur D_0 .

- (iii) La fonctionnelle f est uniformément convexe sur le sous-ensemble convexe $D_0 \subset \Omega$ si et seulement si il existe une constante $c > 0$ telle que

$$\forall \mathbf{x}, \mathbf{y} \in D_0, \quad f''(\mathbf{x})(\mathbf{y} - \mathbf{x}, \mathbf{y} - \mathbf{x}) \geq 2c \|\mathbf{y} - \mathbf{x}\|_E^2.$$



La condition (ii) ci-dessus n'est qu'une condition suffisante, la réciproque étant inexacte.

Théorème 3.2.8 – Cas Ω ouvert convexe

On suppose que la fonctionnelle $f : \Omega \subset E \rightarrow \mathbb{R}$ est deux fois dérivable dans un ouvert convexe Ω de l'espace vectoriel normé E .

- (i) La fonctionnelle f est convexe sur le sous-ensemble convexe Ω si et seulement si $\forall \mathbf{x} \in \Omega$, $f''(\mathbf{x})$ est semi-définie positive.
- (ii) Si $\forall \mathbf{x} \in \Omega$, $f''(\mathbf{x})$ est définie positive, alors la fonctionnelle f est strictement convexe sur Ω .
- (iii) En dimension finie, la stricte convexité est équivalente à l'uniforme convexité.

Exercice 3.2.1.

1. Soit $f : \mathbb{R}^2 \rightarrow \mathbb{R}$, $f(x) = ax_1^2 + 2bx_1x_2 + cx_2^2$. Donner une condition nécessaire et suffisante sur $(a, b, c) \in \mathbb{R}^3$ pour que f soit convexe sur \mathbb{R}^2 .

2. L'application $g : \mathbb{R}^2 \rightarrow \mathbb{R}$, $g(x) = x_1^2 + x_1x_2 + x_2^2 + 3|x_1 + x_2 + 3|$ est-elle convexe, voire strictement convexe, sur \mathbb{R}^2 ?

3. L'application $h : \mathbb{R}^n \rightarrow \mathbb{R}$, $h(x) = e^{\frac{1}{2}(x_1^2 + \dots + x_n^2)}$ est-elle convexe, voire strictement convexe, sur \mathbb{R}^n ?

□

Existence de solution, unicité de solution

4.1 Introduction

Les problèmes d'optimisation où l'ensemble C est fini admettent toujours une solution, par contre, ceci n'est pas toujours le cas si C a un nombre infini d'éléments. Par exemple, le problème d'optimisation où la fonctionnelle à minimiser est $f(x) = 1/x$ et l'ensemble des contraintes est $C = \{x \in \mathbb{R}, x > 0\}$, n'admet pas de solution. En effet $f(x) > 0$ pour tout x dans C et pour tout $\varepsilon > 0$, il existe $x > 1/\varepsilon$ tel que $f(x) < \varepsilon$. Il est donc préférable, avant de vouloir calculer la solution, de s'assurer que le problème en admet une.

4.2 Existence de solution

4.2.1 Problèmes avec contraintes

Théorème 4.2.1

Soit (P) un problème d'optimisation avec contraintes $C \subset E$. Si f est continue et C est un compact non vide, alors le problème (P) admet une solution.

► C'est une application immédiate du théorème qui dit que l'image d'un compact par une application continue dans un espace séparé est un compact. ■

Remarque 4.2.1. On rappelle que, en dimension finie, un ensemble C est compact si et seulement si C est fermé et borné.

Exemple 4.2.1. Considérons le problème suivant :

$$(P) \begin{cases} \min & f(x) \\ x \in & [0, 1] \end{cases}$$

où f est la fonction suivante :

$$\begin{aligned} f : [0, 1] &\longrightarrow \mathbb{R} \\ 0 &\longmapsto 1 \\ x &\longmapsto x \end{aligned}$$

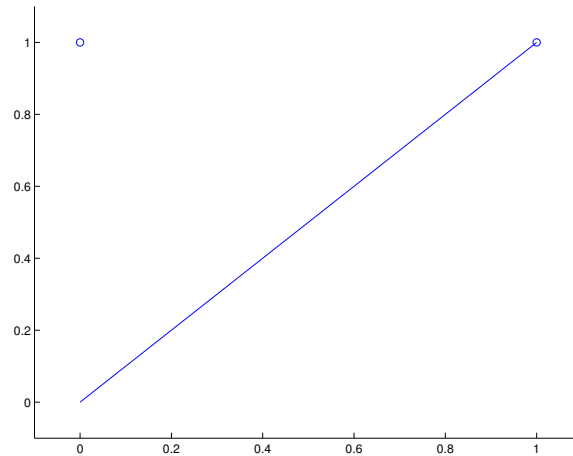
Ce problème n'admet pas de solution. L'hypothèse du théorème (??) qui n'est pas vérifiée est la continuité de f . □

Exemple 4.2.2. Considérons le problème suivant :

$$(P) \begin{cases} \min & f(x) = \frac{1}{x} \\ x \in & [1, 5] \end{cases}$$

- (i) f est continue ;
- (ii) $[1, 5]$ est un fermé et borné, donc un compact de \mathbb{R} .

Par suite ce problème admet une solution. □

FIGURE 4.1 – Exemple où f non continue

Exemple 4.2.3. Considérons le problème suivant :

$$(P) \begin{cases} \min f(x) = \frac{1}{x} \\ x \in]1, 5] \end{cases}$$

Ce problème a une solution, mais les hypothèses du théorème ne sont pas vérifiées. □

Exemple 4.2.4. Considérons le problème suivant :

$$(P) \begin{cases} \min f(x) = \frac{1}{x} \\ x \in [1, 5[\end{cases}$$

Ce problème n'admet pas de solution, $C = [1, 5[$ n'est pas fermé. □

4.2.2 Problème sans contraintes

Définition 4.2.2

Une fonction $f : E \rightarrow \mathbb{R}$, E espace vectoriel normé, est dite 0-coercive si et seulement si

$$f(x) \longrightarrow +\infty \text{ quand } \|x\| \longrightarrow +\infty. \quad (4.1)$$

Théorème 4.2.3

Soit (P) un problème d'optimisation avec contraintes où f est une fonction de \mathbb{R}^n à valeurs dans \mathbb{R} et C est un fermé non vide. Si f est continue et 0-coercive, alors le problème admet une solution.

► Soit $(x_k)_{k \in \mathbb{N}}$ une suite minimisante de points de C , c'est-à-dire une suite de point de C telle que $\lim_{k \rightarrow +\infty} f(x_k) = \inf_{x \in \mathbb{R}^n} f(x) = \mu < +\infty$. Montrons que cette suite est bornée. Sinon il existe une sous-suite $(x_{n_k})_{n_k}$ telle que $\|x_{n_k}\|$ tende vers $+\infty$ lorsque n_k tend vers $+\infty$ et donc, comme f est 0-coercive, $\lim_{n_k \rightarrow +\infty} f(x_{n_k}) = +\infty$, ce qui est impossible.

Par suite il existe un réel $R > 0$ tel que la suite $(x_k)_{k \in \mathbb{N}}$ soit contenue dans $C \cap B_f(0, R)$ qui est un fermé borné de \mathbb{R}^n ; c'est donc un compact dont on peut extraire une sous-suite qui converge vers x^* . Mais f est continue, et donc $f(x^*) = \mu$ et x^* est une solution du problème d'optimisation. ■

Remarque 4.2.2. Le théorème précédent s'applique si le problème d'optimisation est sans contraintes car dans ce cas $C = E$.

4.3 Cas convexe

Théorème 4.3.1

Si C est un convexe de E espace vectoriel normé et si f est une fonction de C à valeurs dans \mathbb{R} convexe, alors l'ensemble des solutions est soit vide soit un ensemble convexe de E .

► Supposons que l'ensemble des solutions ne soit pas vide. Soient x et y deux solutions alors $f(x) = f(y)$ car ($f(x) \leq f(y)$ et $f(y) \leq f(x)$). Par suite, pour tout $\alpha \in]0, 1[$, nous avons

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

En conséquence $\alpha x + (1 - \alpha)y$ est aussi une solution. ■

Théorème 4.3.2

Si C est un convexe de E espace vectoriel normé et si f est une fonction de C à valeurs dans \mathbb{R} strictement convexe, alors il existe au plus un point x^ minimisant f sur C .*

► Supposons qu'il existe deux solutions x_1 et x_2 . Pour $\alpha \in]0, 1[$, on pose $x_\alpha = \alpha x_1 + (1 - \alpha)x_2$, alors, puisque f est strictement convexe on a

$$f(x_\alpha) < \alpha f(x_1) + (1 - \alpha)f(x_2) = f(x_1) = f(x_2),$$

ce qui est impossible. ■

Théorème 4.3.3

Si C est un convexe de E espace vectoriel normé et si f est une fonction de C à valeurs dans \mathbb{R} convexe, alors tout minimum local x^ de f sur C est un minimum global de f sur C .*

► Soit x^* un minimum local de f sur C . Il existe donc $\eta > 0$ tel que pour tout $x \in C \cap B(x^*, \eta)$, $f(x^*) \leq f(x)$. Supposons maintenant qu'il existe dans C un point y tel que $f(y) < f(x^*)$. Alors, puisque f est convexe, on a pour tout $\alpha \in]0, 1[$

$$\begin{aligned} f(x^* + \alpha(y - x^*)) &= f((1 - \alpha)x^* + \alpha y) \leq (1 - \alpha)f(x^*) + \alpha f(y) \\ &< (1 - \alpha)f(x^*) + \alpha f(x^*) = f(x^*). \end{aligned}$$

Mais pour α suffisamment proche de 0, $x^* + \alpha(y - x^*) \in B(x^*, \eta)$, d'où la contradiction. ■

Condition nécessaire, condition suffisante de solution

Cas sans contraintes et cas de contraintes convexes

5.1 Condition du premier ordre

5.1.1 Cas sans contraintes

Théorème 5.1.1

Soient Ω un ouvert d'un espace vectoriel normé E et f une application de Ω à valeurs dans \mathbb{R} . Si f admet un minimum local en x^* et si f est dérivable en x^* alors on a l'équation parfois appelée équation d'Euler

$$f'(x^*) = 0. \quad (5.1)$$

► Soit $h \in E$, comme Ω est ouvert, il existe $\eta > 0$ tel que la fonction

$$\begin{aligned} \varphi :]-\eta, \eta[&\longrightarrow \mathbb{R} \\ t &\longmapsto \varphi(t) = f(x^* + th) \end{aligned}$$

soit bien définie. φ est dérivable en 0 et $\varphi'(0) = f'(x^*) \cdot h$. Mais x^* est un minimum local de f , donc 0 est un minimum local de φ , par suite on a

$$0 \geq \lim_{t \rightarrow 0^-} \frac{\varphi(t) - \varphi(0)}{t} = \varphi'(0) = \lim_{t \rightarrow 0^+} \frac{\varphi(t) - \varphi(0)}{t} \geq 0.$$

Ainsi, pour tout h , $\varphi'(0) = f'(x^*) \cdot h = 0$. ■

Définition 5.1.2 – Point critique

Un point qui vérifie $f'(x) = 0$ est dit un point critique et sa valeur en f , $f(x)$ une valeur critique.

5.1.2 Cas de contraintes convexes

Théorème 5.1.3

Soient Ω un ouvert d'un espace vectoriel normé E et $f : \Omega \rightarrow \mathbb{R}$ à valeur dans \mathbb{R} . Soit $C \subset \Omega$ convexe. Si f admet un minimum local en x^* sur C et si f est dérivable en x^* alors on a l'inéquation d'Euler

$$\forall y \in C, f'(x^*) \cdot (y - x^*) \geq 0. \quad (5.2)$$

► Soit $y \in C$, alors la fonction

$$\begin{aligned} \varphi : [0, 1] &\longrightarrow \mathbb{R} \\ t &\longmapsto \varphi(t) = f(x^* + t(y - x^*)) \end{aligned}$$

est bien définie et admet une dérivée à droite en 0 $\varphi'^+(0) = f'(x^*) \cdot (y - x^*)$. Mais 0 est un minimum local de φ et donc $\varphi(0) \leq \varphi(t)$ pour t suffisamment proche de 0. Par suite

$$\varphi'^+(0) = \lim_{t \rightarrow 0^+} \frac{\varphi(t) - \varphi(0)}{t} \geq 0. \quad \blacksquare$$

Remarque 5.1.1. (i) Si C est un sous espace affine ($C = x_0 + V$, avec V sous-espace vectoriel de E) alors l'inéquation d'Euler (??) devient

$$\forall h \in V, f'(x^*) \cdot h = 0$$

(ii) Si $C = E$ alors l'inéquation d'Euler (??) devient l'équation d'Euler (??).

5.1.3 Problèmes convexes

Théorème 5.1.4

Soit $f : \Omega \subset E \rightarrow \mathbb{R}$, Ω ouvert d'un espace vectoriel normé E et soit $C \subset \Omega$ convexe. On suppose que f est convexe sur C et dérivable en tout point de C , alors les conditions suivantes sont équivalentes.

- (i) x^* est un minimum global de f sur C .
- (ii) x^* est un minimum local de f sur C .
- (iii) Pour tout $y \in C$, $f'(x^*) \cdot (y - x^*) \geq 0$.

► (??) \Rightarrow (??) est évident.

(??) \Rightarrow (??) est le théorème (??) précédent.

(??) \Rightarrow (??) ?

f est convexe, par suite (cf. le théorème ??) nous avons grâce à (??)

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

■

Remarque 5.1.2. Si C est un ouvert convexe (??) est équivalent à $f'(x^*) = 0$. L'équation d'Euler est donc dans ce cas une condition nécessaire et suffisante de solution.

Corollaire 5.1.5

On considère le problème au moindres carrés linéaire

$$(P) \begin{cases} \min f(\beta) = \frac{1}{2} \|y - X\beta\|^2 \\ \beta \in \mathbb{R}^p. \end{cases}$$

Alors β^* est une solution de (P) si et seulement si ce point vérifie les équations normales

$$X^T X \beta = X^T y. \quad (5.3)$$

► le problème est un problème convexe et on a $\nabla f(\beta) = X^T X \beta - X^T y$. ■

5.2 Conditions du deuxième ordre

5.2.1 Condition nécessaire

Théorème 5.2.1 – Condition nécessaire du deuxième ordre

Soit Ω un ouvert d'un espace vectoriel normé E et $f : \Omega \rightarrow \mathbb{R}$. Si x^* est un minimum local de f et si f est deux fois dérivable en x^* alors $f''(x^*)$ est semi-définie positive.

► Soit $h \neq 0$ un vecteur quelconque de E . x^* est un minimum local de f , donc $f'(x^*) = 0$ et il existe $\theta_0 > 0$ tel que pour tout $0 < \theta < \theta_0$ on ait $f(x^*) \leq f(x^* + \theta h)$. f étant deux fois dérivable en x^* on a

par Taylor-Young

$$\begin{aligned} f(x^* + \theta h) - f(x^*) &= f'(x^*) \cdot h + \frac{\theta^2}{2} f''(x^*) \cdot (h, h) + \|\theta h\|^2 \varepsilon(\theta h) \\ &= \frac{\theta^2}{2} (f''(x^*) \cdot (h, h) + 2\|h\|^2 \varepsilon(\theta h)) \geq 0. \end{aligned}$$

En divisant par θ^2 et en passant à la limite dans le membre de droite on en déduit que $f''(x^*) \cdot (h, h) \geq 0$. Ceci étant vrai pour tout h , on obtient le résultat. ■

Remarque 5.2.1. Il ne s'agit bien que d'une condition nécessaire (prendre $f(x) = x^3$).

5.2.2 Condition suffisante

Définition 5.2.2

Soit $B \in \mathcal{L}_2(E, \mathbb{R})$ une forme bilinéaire symétrique définie sur E , espace vectoriel normé.

(i) B est dite semi-définie positive si et seulement si pour tout $h \in E$

$$B(h, h) \geq 0.$$

(ii) B est dite définie positive si et seulement si pour tout $h \in E, h \neq 0$

$$B(h, h) > 0.$$

(iii) B est uniformément définie positive ou elliptique si et seulement si il existe $c > 0$ tel que pour tout $h \in E$

$$B(h, h) \geq c\|h\|^2.$$

Remarque 5.2.2. Si E est un espace vectoriel de dimension fini il y a équivalence entre la définie positivité et l'ellipticité.

Théorème 5.2.3 – Condition suffisante du deuxième ordre

Soit Ω un ouvert d'un espace vectoriel normé E et $f : \Omega \rightarrow \mathbb{R}$ dérivable sur Ω .

- (i) Si x^* est un point de Ω tel que $f'(x^*) = 0$, f deux fois dérivable en x^* et $f''(x^*)$ elliptique, alors x^* est un minimum local de f .
- (ii) Si f est deux fois dérivable sur Ω et s'il existe une boule $B(x^*, \eta) \subset \Omega$ telle que, pour tout $x \in B(x^*, \eta)$, $f''(x)$ est semi-définie positive et si $f'(x^*) = 0$, alors x^* est un minimum local de f .

► (i) La formule de Taylor-Young nous permet d'écrire pour tout h suffisamment petit

$$\begin{aligned} f(x^* + h) - f(x^*) &= \frac{1}{2} f''(x^*) \cdot (h, h) + \|h\|^2 \varepsilon(h) \\ &\geq \frac{1}{2} (c + 2\varepsilon(h)) \|h\|^2. \end{aligned}$$

Par suite il existe $\eta > 0$ tel que pour tout h tel que $\|h\| < \eta$ on ait

$$f(x^* + h) - f(x^*) > 0.$$

- (ii) Si pour tout $x \in B(x^*, \eta)$, $f''(x)$ est semi-positive, f est convexe sur l'ouvert $B(x^*, \eta)$. Par suite, comme $f'(x^*) = 0$, x^* est un minimum local de f . ■

Remarque 5.2.3. Pour les conditions nécessaires de solution du premier ordre et du deuxième ordre, on exploite le développement de Taylor-Young le long de toute direction h , combiné avec le fait que l'on a un minimum dans cette direction. Pour la condition suffisante du deuxième ordre, cette seule information, pourtant relative à toute direction h donnée, est insuffisante (cf. l'exercice ??, extrait de [?] page 50).

5.3 Exercices

 **Exercice 5.3.1.** On considère la fonction


$$\begin{aligned} f : \mathbb{R}^2 &\longrightarrow \mathbb{R} \\ x &\longmapsto f(x) = 3x_1^4 - 4x_1^2x_2 + x_2^2. \end{aligned}$$

1. Montrer que l'unique point critique de f est $\bar{x} = (0 \ 0)$.
2. La condition nécessaire de solution du deuxième ordre est-elle vérifiée ?
3. La condition suffisante de solution du deuxième ordre est-elle vérifiée ?
4. On fixe maintenant $d \in \mathbb{R}^2$ et on considère la fonction

$$\begin{aligned} \varphi : \mathbb{R} &\longrightarrow \mathbb{R} \\ t &\longmapsto \varphi(t) = f(\bar{x} + td). \end{aligned}$$

Montrer que $t = 0$ est un minimum local de φ .

5. Calculer $f(x_1, 2x_1^2)$. Conclusion (faire le lien avec la remarque ??). □


 **Exercice 5.3.2.** Attention à l'intuition dans \mathbb{R} , cf. [?] page 52 !

1. On considère une fonction f de \mathbb{R} à valeurs dans \mathbb{R} dérivable en tout point. On suppose que f admette un minimum local en \bar{x} et que \bar{x} est l'unique point critique de f . Démontrer que \bar{x} est un minimum global de f .

2. On considère maintenant la fonction

$$\begin{aligned} f : \mathbb{R}^2 &\longrightarrow \mathbb{R} \\ x &\longmapsto f(x) = 2x_1^3 + 3e^{2x_2} - 6x_1e^{x_2} \end{aligned}$$

Montrer que $\bar{x} = (1 \ 0)$ est l'unique point critique de f , que \bar{x} est un minimum local de f , mais que f n'admet pas de minimum global. □

 **Exercice 5.3.3.** Donner un exemple de fonction f (de \mathbb{R} dans \mathbb{R}), deux fois dérivable, ayant un minimum strict en un point \bar{x} et telle que dans toute boule $\mathcal{B}(\bar{x}, \rho)$ il existe un point $x \in \mathcal{B}(\bar{x}, \rho)$ vérifiant $f''(x) < 0$. □

Problèmes aux moindres carrés

6.1 Introduction

L'objectif de ce chapitre est de résoudre les problèmes aux moindres carrés non-linéaires.

$$(P_1) \begin{cases} \min f(\beta) = \frac{1}{2} \|r(\beta)\|^2 \\ \beta \in \mathbb{R}^p. \end{cases}$$

Dans le cas où l'application r est affine, le problème est linéaire et sera résolu facilement. Si r n'est pas affine, il nous faut développer des algorithmes pour calculer une solution. La branche des mathématiques qui s'occupe de ces questions pour les problèmes d'optimisation continues s'appelle l'optimisation continue et sera abordée en deuxième année. Nous allons ici nous limiter à des algorithmes de base pour les problèmes aux moindres carrés, qui sont l'algorithme de Newton et l'algorithme de Gauss-Newton.

6.2 Les moindres carrés linéaires

6.2.1 Rappels

Considérons le problème aux moindres carrés linéaire

$$(P2) \begin{cases} \min f(\beta) = \frac{1}{2} \|y - X\beta\|^2 \\ \beta \in \mathbb{R}^p. \end{cases}$$

Ce problème admet une solution. En effet, ce problème est équivalent à résoudre

$$(P3) \begin{cases} \min g(\gamma) = \frac{1}{2} \|y - \gamma\|^2 \\ \gamma \in \text{Im } X \subset \mathbb{R}^n. \end{cases}$$

Comme $\text{Im } X$ est un fermé et que g est 0-coercive, le théorème ???.?? démontre l'existence d'une solution.

Remarque 6.2.1. Le problème $(P??)$ est en fait le problème de la projection orthogonale du vecteur y sur $\text{Im } X$. Il possède une unique solution car g est strictement convexe. Par contre le problème initial $(P??)$ possède une ou une infinité de solutions suivant que le rang de X est p ou est strictement inférieur à p .

Le problème $(P??)$ est un problème convexe et différentiable, par suite une solution est caractérisée par la condition nécessaire d'ordre 1, qui conduit au système d'équations suivant, aussi appelé dans ce cas *équations normales* :

$$\nabla f(\beta) = X^T X \beta - X^T y = 0. \quad (6.1)$$

Remarque 6.2.2. Considérons ici la matrice X comme l'expression d'une application linéaire de \mathbb{R}^p à valeurs dans \mathbb{R}^n . on a alors

$$\begin{aligned} \mathbb{R}^p &= \text{Ker } X^\perp \oplus \text{Ker } X & \mathbb{R}^n &= \text{Im } X \oplus \text{Im } X^\perp = \text{Im } X \oplus \text{Ker}(X^T) \\ \beta^* &= X^+ y & \gamma^* &= \text{Proj}_{\text{Im } X}(y) \\ X^+ X &= \text{Proj}_{(\text{Ker } X)^\perp} & X X^+ &= \text{Proj}_{\text{Im } X} \\ \text{Si } \text{rank}(X) &= p, \text{ alors } \text{Ker } X &= \{\vec{0}\} \text{ et } X^+ &= (X^T X)^{-1} X^T \end{aligned}$$

6.2.2 Application : approximation d'une fonction au sens des moindres carrés

Le problème de l'approximation d'une fonction f sur un intervalle $I \subset \mathbb{R}$ est fondamentalement différent de celui de l'interpolation. Il consiste à remplacer la fonction f considérée par une autre fonction $\mathcal{P}(x)$ (en général plus régulière, et facile à manipuler) de sorte que la distance entre f et \mathcal{P} soit aussi petite que possible.

On peut chercher par exemple un polynôme de bas degré, qui approche la fonction f en un sens à préciser sur l'intervalle I , ce qui diffère du problème d'interpolation qui consiste à trouver un polynôme de degré en général élevé qui coïncide au maximum avec la fonction f .

La notion de distance entre les fonctions f et \mathcal{P} est bien évidemment fondamentale dans la définition du procédé d'approximation. On pourra par exemple distinguer :

- (i) **l'approximation au sens de la convergence uniforme**, où il s'agit de minimiser

$$\max_{x \in I} |f(x) - \mathcal{P}(x)| = \|f - \mathcal{P}\|_{\infty} ,$$

(problème ne relevant pas des moindres carrés)

- (ii) **l'approximation en moyenne quadratique** où il s'agit de minimiser la quantité

$$\int_I (f(x) - \mathcal{P}(x))^2 dx = \|f - \mathcal{P}\|_{L^2(I)}^2 ,$$

- (iii) **l'approximation au sens des moindres carrés discrets**, utile lorsque f n'est connue que de manière discrète (c'est à dire sur un ensemble fini de points $x_i \in I$, $1 \leq i \leq m$) ; cette approximation consiste alors à minimiser la quantité

$$\sum_{i=1}^m (f(x_i) - \mathcal{P}(x_i))^2 .$$

6.2.2.1 Approximation en moyenne quadratique

On se propose d'approcher f sur l'intervalle $I = [a, b]$ par une fonction $\mathcal{P}(x)$, où \mathcal{P} est une combinaison linéaire d'un ensemble de n fonctions données $u_j \in L^2(I)$, $j = 1, \dots, n$,

$$\mathcal{P}(x) = \sum_{j=1}^n \beta_j u_j(x) ,$$

les coefficients β_j étant donc les inconnues à déterminer de façon à minimiser la quantité

$$f(\beta_1, \dots, \beta_n) = \|f - \mathcal{P}\|_{L^2(I)}^2 = \int_a^b (f(x) - \sum_{j=1}^n \beta_j u_j(x))^2 dx .$$

Remarque : Si on souhaite réaliser une approximation polynômiale de f , il suffira de choisir

$$u_j(x) = x^{j-1} , \quad j = 1, 2, \dots, n .$$

Proposition 6.2.1

Soit $\beta \in \mathbb{R}^n$ le vecteur des coefficients β_j , $j = 1, \dots, n$. Une condition nécessaire et suffisante pour que β réalise le minimum de la fonctionnelle $f(\beta)$, est que β soit solution du système linéaire :

$$\sum_{j=1}^n (u_j | u_i) \beta_j = (f | u_i) , \quad i = 1, \dots, n ,$$

avec

$$(u_j|u_i) = \int_a^b u_j(x) u_i(x) dx \quad \text{et} \quad (f|u_i) = \int_a^b f(x) u_i(x) dx.$$

6.2.2.2 Approximation au sens des moindres carrés discrets

Soit f une fonction dont on connaît la valeur sur un sous ensemble fini de points x_1, \dots, x_m . On se propose alors d'approcher la fonction f par une fonction \mathcal{P} de telle façon que la quantité

$$\sum_{i=1}^m (f(x_i) - \mathcal{P}(x_i))^2,$$

soit minimale.

De plus, on cherche \mathcal{P} sous la forme

$$\mathcal{P}(x) = \sum_{j=1}^n \beta_j u_j(x);,$$

où les u_j sont des fonctions connues et les $\{\beta_j\}_{j=1}^n$ sont les inconnues à déterminer. On suppose de plus que $m > n$, pour ne pas être ramené à un problème d'interpolation. Ce problème correspond au *lissage d'une fonction f donnée par une combinaison linéaire de fonctions quelconques*.

Introduisons alors la matrice X à m lignes et n colonnes, ainsi que les vecteurs $\beta \in \mathbb{R}^n$ et $y \in \mathbb{R}^m$ définis par

$$X = \begin{pmatrix} u_1(x_1) & \cdots & u_n(x_1) \\ u_1(x_2) & \cdots & u_n(x_2) \\ \vdots & & \vdots \\ u_1(x_m) & \cdots & u_n(x_m) \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix}, \quad \text{et} \quad y = \begin{pmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_m) \end{pmatrix}.$$

En utilisant ces notations, on a alors le résultat suivant :

Proposition 6.2.2

Il est facile de voir que le problème d'approximation décrit ci-dessus se ramène en fait à minimiser par rapport à $\beta \in \mathbb{R}^n$ la quantité

$$f(\beta) = \|X\beta - y\|_2^2,$$

et donc que le vecteur de coefficients $\beta \in \mathbb{R}^n$ recherché correspond à la solution au sens des moindres carrés du système linéaire surdéterminé

$$X\beta = y.$$

De plus, si les points x_i , $i = 1, \dots, m$, et les fonctions u_j , $j = 1, \dots, n$, sont choisis de telle façon que la matrice X soit de rang maximal, alors le problème d'approximation au sens des moindres carrés discrets précédent admet une solution unique $\beta \in \mathbb{R}^n$, solution du système linéaire

$$X^T X \beta = X^T y.$$

6.3 La méthode de Newton

6.3.1 Introduction

L'algorithme de Newton est à la base des algorithmes d'optimisation implémentés dans les bibliothèques numériques. Mais, avant de considérer le cas d'un problème d'optimisation, nous allons nous intéresser au problème de la résolution d'un système d'équations non linéaires à n équations et n inconnues.

6.3.2 Algorithme de Newton pour résoudre $f(x) = 0$

6.3.2.1 Résolution d'une équation : cas de la dimension 1

le problème est ici de résoudre numériquement une équation $f(x) = 0$ où la fonction f est une fonction réelle de la variable réelle. nous supposons de plus que cette fonction est dérivable. on considère alors l'algorithme suivant :

Algorithme 6.1.

Initialisation :

choisir $x^{(0)} \in \mathbb{R}$

choisir $\varepsilon > 0$ et $MaxIter$

$k := 0$

Corps :

répéter

Résoudre $f(x^{(k)}) + f'(x^{(k)})(x - x^{(k)}) = 0$, soit $x^{(k+1)}$ la solution

$k := k + 1$

jusqu'à $(|f(x^{(k)})| < \varepsilon(|f(x^{(0)}| + 1))$ ou $(k = MaxIter)$

Remarque 6.3.1. (i) ici $f'(x^{(k)})$ appartient à \mathbb{R} .

(ii) l'algorithme peut se “bloquer” si $f'(x^{(k)}) = 0$, et donc dans ce cas l'algorithme ne fournit pas de solution.

(iii) cet algorithme ne converge pas toujours.

Illustration graphique 6.3.2. l'intersection de la tangente à f en $x^{(k)}$ avec l'axe des abscisses (cf. figure ??) est donnée par la solution de $f(x^{(k)}) + f'(x^{(k)})(x - x^{(k)}) = 0$, soit :

$$x = x^{(k)} - \frac{1}{f'(x^{(k)})} \cdot f(x^{(k)})$$

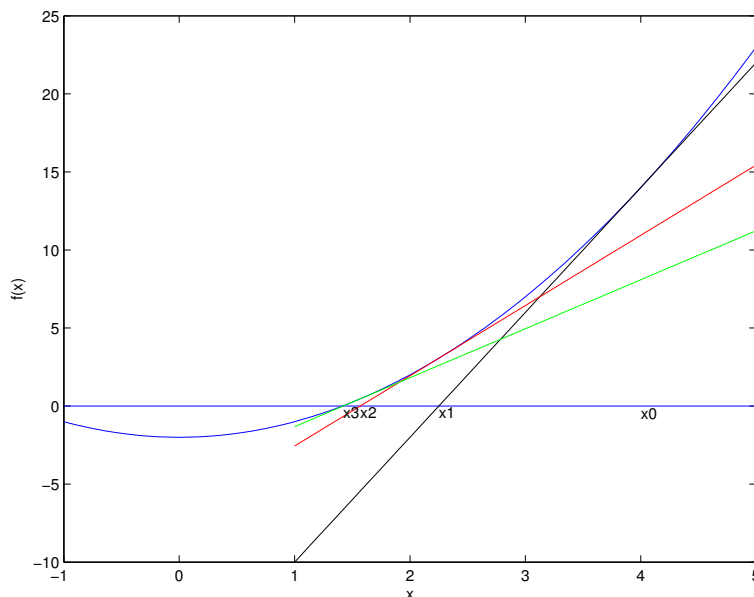


FIGURE 6.1 – Algorithme de Newton.

6.3.3 Résolution d'équations : cas de la dimension n

De la même façon qu'en dimension 1 on obtient l'algorithme en calculant $x^{(k+1)}$ à partir de $x^{(k)}$ donné en annulant la "meilleure" approximation affine de la fonction f au voisinage de $x^{(k)}$, c'est-à-dire en résolvant le système linéaire à n équations et à n inconnues suivant :

$$f(x^{(k)}) + J_f(x^{(k)})(x - x^{(k)}) = 0$$

qui admet une unique solution si $J_f(x^{(k)})$ est inversible.

Remarque 6.3.3. On présente très souvent l'algorithme de Newton sous la forme de la mise à jour du point courant donnée par ???. Cette équation est très utile pour la théorie, mais est bien évidemment à bannir pour une implémentation informatique.

$$x^{(k+1)} := x^{(k)} - [J_f(x^{(k)})]^{-1} f(x^{(k)}) \quad (6.2)$$

Remarque 6.3.4. lorsque $n = 1$, $f'(x^{(k)})$ est un réel et $[f'(x^{(k)})]^{-1} = 1/f'(x^{(k)})$, et nous retrouvons la mise à jour exposée précédemment.

en conclusion nous obtenons l'algorithme suivant :

Algorithme 6.2. [algorithme de newton]

Initialisation :

choisir $x^{(0)} \in \mathbb{R}^n$

choisir $\varepsilon > 0$ et $MaxIter$

$k := 0$

Corps :

répéter


Résoudre $f(x^{(k)}) + J_f(x^{(k)})(x - x^{(k)}) = 0$, soit $x^{(k+1)}$ la solution

$k := k + 1$

jusqu'à $(\|f(x^{(k)})\| < \varepsilon(\|f(x^{(0)})\| + 1))$ ou $(k = MaxIter)$


Remarque 6.3.5. (i) cet algorithme se "bloque" si $J_f(x^{(k)})$ n'est pas inversible.

(ii) le test d'arrêt $\|f(x^{(k)})\| < \varepsilon\|f(x^{(0)})\|$ signifie en fait que toutes les composantes de $f(x^{(k)})$ sont "proches" de 0 en relatif.

 **Exercice 6.3.1.** On considère la fonction

$$\begin{aligned} \text{soit } f : \mathbb{R} &\longrightarrow \mathbb{R} \\ x &\longmapsto x^2 - a \text{ avec } a > 0. \end{aligned}$$

1. Donner l'itération de Newton pour résoudre $f(x) = 0$. □

 **Exercice 6.3.2.** On considère la fonction

$$\begin{aligned} f : \mathbb{R}^2 &\longrightarrow \mathbb{R}^2 \\ \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} &\longmapsto \begin{pmatrix} x_1 + x_2 - 3 \\ x_1^2 + x_2^2 - 9 \end{pmatrix} \end{aligned}$$

alors $f(x) = 0$ si et seulement si $x = (0, 3)^T$ ou $x = (3, 0)^T$.

1. Appliquer l'algorithme de Newton (??) en partant du point $x^{(0)} = (1, 5)^T$ et en prenant $\varepsilon = 0,6$. □

6.3.4 Convergence

Théorème 6.3.1

Soit f une fonction définie sur un ouvert Ω de \mathbb{R}^n à valeurs dans \mathbb{R}^n de classe C^2 dans $B(x^*, r) \subset \Omega$ et x^* un point de Ω tel que $f(x^*) = 0$. On suppose que $f'(x^*)$ est inversible, alors il existe $\varepsilon > 0$ tel que pour tout point $x^{(0)} \in B(x^*, \varepsilon)$, l'algorithme de Newton est bien défini et la suite des itérés $(x_k)_k$ converge vers x^* . De plus la convergence est quadratique, c'est-à-dire qu'il existe $c > 0$ tel que

$$\|x^{(k+1)} - x^*\| \leq c \|x^{(k)} - x^*\|^2 \quad (6.3)$$

Avant de voir la démonstration, voyons ce que signifie l'équation (??) (exemple provenant de [?] page 23). À la table (??) on voit que la convergence est plus rapide pour la fonction $f_1(x) = x^2 - 1$ qui vérifie les hypothèses du théorème que pour la fonction $f_2(x) = (x - 1)^2$ qui ne vérifie pas que $f'(1)$ soit inversible. La convergence quadratique signifie en pratique que si on est suffisamment près de la solution et si à une itération k on a p décimales qui sont exactes, on aura à l'itération $k + 1$, $2p$ décimales qui seront exactes.¹

	$f_1(x) = x^2 - 1$	$f_2(x) = (x - 1)^2$
x_0	2	2
x_1	1.25	1.5
x_2	1.025	1.25
x_3	1.0003048780488	1.125
x_4	1.0000000464611	1.0625
x_5	1.0	1.03125

TABLE 6.1 – Convergence quadratique pour f_1 et linéaire pour f_2 .

Démontrons maintenant le théorème.

► L'ensemble $\{x \in \Omega, J_f(x) \text{ inversible}\} = \mathbb{C}_\Omega\{x \in \Omega, \det \circ J_f(x) = 0\}$ est un ouvert car c'est le complémentaire de l'antécédent d'un fermé par une application continue. Par suite il existe $\varepsilon_1 > 0$ tel que pour tout $x \in B(x^*, \varepsilon_1)$, $J_f(x)$ soit inversible. Comme f est C^2 l'application qui à $x \in B(x^*, \varepsilon_1)$ associe $\|[J_f(x)]^{-1}\|$ est continue, on en déduit que pour $0 < \varepsilon_2 < \varepsilon_1$ l'image par cette application de $\overline{B(x^*, \varepsilon_2)}$ est un compact. Par suite, il existe $\beta > 0$ tel que pour tout $x \in \overline{B(x^*, \varepsilon_2)}$ on a $\|[J_f(x)]^{-1}\| \leq \beta$.

$$\|x^{(k+1)} - x^*\| = \|x^{(k)} - x^* - [J_f(x^{(k)})]^{-1} f(x^{(k)})\| \quad (6.4)$$

$$\leq \|[J_f(x^{(k)})]^{-1}\| \|f(x^{(k)}) - f(x^*) - J_f(x^{(k)})(x^{(k)} - x^*)\| \quad (6.5)$$

$$\leq \frac{1}{2} \|[J_f(x^{(k)})]^{-1}\| \sup_{x \in B_f(x^*, \varepsilon_2)} \|\nabla^2 f(x)\| \|x^{(k)} - x^*\|^2. \quad (6.6)$$

Mais f est C^2 par suite en posant $\sup_{x \in B_f(x^*, \varepsilon_2)} \|\nabla^2 f(x)\| = \gamma$ on obtient

$$\|x^{(k+1)} - x^*\| \leq \frac{1}{2} \beta \gamma \|x^{(k)} - x^*\|^2. \quad (6.7)$$

Posons maintenant $\varepsilon = \min(\varepsilon_2, 1/(\beta\gamma))$ et prenons $x^{(0)} \in B(x^*, \varepsilon)$, on obtient

$$\|x^{(1)} - x^*\| \leq \frac{1}{2} \|x^{(0)} - x^*\| \leq \varepsilon/2.$$

Donc $x^{(1)} \in B(x^*, \varepsilon)$ et par récurrence

$$\|x^{(k)} - x^*\| \leq \frac{1}{2^k} \|x^{(0)} - x^*\|.$$

1. Ajouter le graphique, cf. Daniel.

Par suite $x^{(k)} \in B(x^*, \varepsilon)$ et la suite des itérés de Newton existe et converge vers x^* . Quand à la convergence quadratique elle vient de l'inéquation (??). ■

6.3.5 Application aux problèmes d'optimisation

Nous rappelons que le problème qui nous intéresse ici est le suivant :

$$(P) \begin{cases} \min f(x) \\ x \in \mathbb{R}^n \end{cases}$$

et nous avons vu (théorème (??)) qu'une condition nécessaire de solution est $f'(x) = 0$. Cette condition nous conduit donc tout naturellement vers la recherche d'un zéro de l'équation $g(x) = 0$ avec $g(x) = \nabla f(x)$. Il nous suffit donc d'appliquer l'algorithme de Newton à cette fonction g . A chaque itération nous aurons donc à résoudre le système linéaire suivant :

$$\nabla^2 f(x^{(k)})(x - x^{(k)}) + \nabla f(x^{(k)}) = 0. \quad (6.8)$$

Remarque 6.3.6. Si $\nabla^2 f(x^{(k)})$ n'est pas inversible alors l'algorithme se bloque.

Remarque 6.3.7. Nous avons vu dans la formule de Taylor-Young à l'ordre 2 (Chapitre ??, Théorème ??) que la meilleure approximation quadratique de la fonctionnelle f au voisinage du point $x^{(k)}$ est donnée par :

$$q(x) = f(x^{(k)}) + (\nabla f(x^{(k)})|x - x^{(k)}) + \frac{1}{2}(\nabla^2 f(x^{(k)})(x - x^{(k)})|x - x^{(k)})$$

et nous avons $\nabla q(x) = \nabla f(x^{(k)}) + \nabla^2 f(x^{(k)})(x - x^{(k)})$ et $\nabla^2 q(x) = \nabla^2 f(x^{(k)})$. Par suite si $\nabla^2 f(x^{(k)})$ est définie positive q est convexe et rechercher le minimum de $q(x)$ sur \mathbb{R}^n est équivalent à résoudre l'équation $\nabla q(x) = 0$. Mais cette dernière équation est équivalente à l'itération de Newton pour résoudre $\nabla f(x) = 0$. En conclusion notre algorithme recherche à chaque itération, lorsque $\nabla^2 f(x^{(k)})$ est définie positive, le minimum de l'approximation à l'ordre 2 de la fonctionnelle f .

6.4 Résolution des problèmes aux moindres carrés non linéaires

6.4.1 Algorithme de Newton

La fonction à optimiser s'écrit ici

$$f(\beta) = \frac{1}{2} \|r(\beta)\|^2 = \frac{1}{2} \sum_{i=1}^n r_i^2(\beta).$$

Nous avons donc

$$\begin{aligned} \nabla f(\beta) &= \sum_i r_i(\beta) \nabla r_i(\beta) = J_r(\beta)^T r(\beta) \\ \nabla^2 f(\beta) &= \sum_i r_i(\beta) \nabla^2 r_i(\beta) + \sum_i \nabla r_i(\beta) \nabla r_i(\beta)^T \\ &= S(\beta) + J_r(\beta)^T J_r(\beta) \end{aligned}$$

L'itération de l'algorithme de Newton s'écrit donc

$$\beta^{(k+1)} = \beta^{(k)} - [S(\beta^{(k)}) + J_r(\beta^{(k)})^T J_r(\beta^{(k)})]^{-1} J_r(\beta^{(k)})^T r(\beta^{(k)}) \quad (6.9)$$

6.4.2 Algorithme de Gauß-Newton

L'idée est ici de linéariser les résidus autour du point $\beta^{(k)}$ et ainsi de se ramener à un problème aux moindres carrés linéaire. Posons $s = \beta - \beta^{(k)}$, on cherche donc à chaque itération à résoudre

$$(P_k) \begin{cases} \min_{s \in \mathbb{R}^p} f_k(s) = \frac{1}{2} \|r(\beta^{(k)}) + J_r(\beta^{(k)})s\|^2 \end{cases}$$

ce qui est équivalent à résoudre les équations normales

$$J_r(\beta^{(k)})^T J_r(\beta^{(k)})s + J_r(\beta^{(k)})^T r(\beta^{(k)}) = 0.$$

Donc, si $J_r(\beta^{(k)})^T J_r(\beta^{(k)})$ est inversible, on peut écrire l'itération de Gauß-Newton :

$$\beta^{(k+1)} = \beta^{(k)} - [J_r(\beta^{(k)})^T J_r(\beta^{(k)})]^{-1} J_r(\beta^{(k)})^T r(\beta^{(k)}). \quad (6.10)$$

Remarque 6.4.1. (i) La différence entre les deux algorithmes réside dans l'absence du terme $S(\beta)$ dans l'équation (??), terme qui contient les matrices hessiennes des résidus.

(ii) Il y a deux avantages à l'algorithme de Gauß-Newton par rapport à l'algorithme de Newton :

- On n'a pas besoin de calculer les matrices hessiennes des résidus ;
- Contrairement à l'algorithme de Newton, on peut toujours trouver une solution $\beta^{(k)}$ à (P_k) .

6.4.3 Exemples

Exemple 6.4.1. Considérons le problème suivant

$$(P) \begin{cases} \min_{x \in \mathbb{R}^2} f(x) = \frac{1}{2}((x_1^2 - x_2)^2 + (1 - x_1)^2) \end{cases}$$

La figure ?? visualise la fonctionnelle à minimiser. Calculons le gradient de f et sa dérivée seconde :

$$\begin{aligned} \nabla f(x) &= \begin{pmatrix} 2x_1^3 - 2x_1x_2 + x_1 - 1 \\ -x_1^2 + x_2 \end{pmatrix}, \\ \nabla^2 f(x) &= \begin{pmatrix} 6x_1^2 - 2x_2 + 1 & -2x_1 \\ -2x_1 & 1 \end{pmatrix}. \end{aligned}$$

Prenons $x^{(0)} = (0 \ 1)^T$. À la première itération, nous avons à résoudre le système linéaire

$$\begin{aligned} \nabla^2 f(x^{(0)})(x - x^{(0)}) + \nabla f(x^{(0)}) &= 0 \iff \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 - 1 \end{pmatrix} + \begin{pmatrix} -1 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \\ &\iff \begin{cases} -x_1 + 0x_2 &= 1 \\ 0x_1 + x_2 &= 0 \end{cases} \end{aligned}$$

D'où

$$x^{(1)} = \begin{pmatrix} -1 \\ 0 \end{pmatrix}.$$

À la deuxième itération, nous avons à résoudre le système linéaire

$$\begin{aligned} \nabla^2 f(x^{(1)})(x - x^{(1)}) + \nabla f(x^{(1)}) &= 0 \iff \begin{pmatrix} 7 & 2 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} x_1 + 1 \\ x_2 \end{pmatrix} + \begin{pmatrix} -4 \\ -1 \end{pmatrix} \\ &\iff \begin{cases} 7x_1 + 2x_2 &= -3 \\ 2x_1 + x_2 &= -1 \end{cases} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \end{aligned}$$

D'où

$$x^{(2)} = \begin{pmatrix} -\frac{1}{3} \\ -\frac{1}{3} \end{pmatrix}.$$

La figure ?? donne les courbes de niveaux de la fonction à minimiser ainsi que les itérés successifs.

□

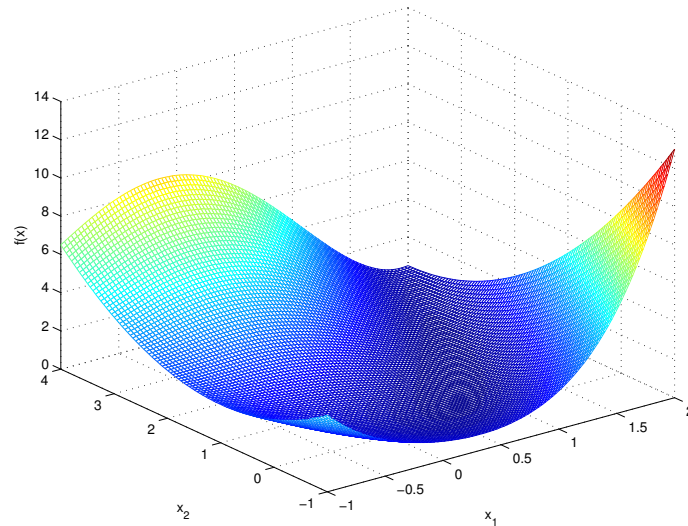
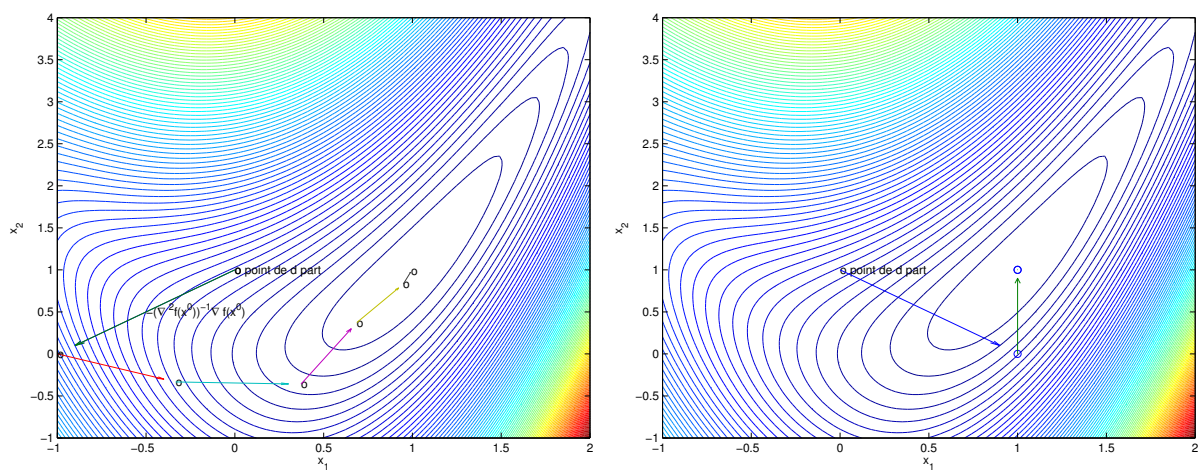
FIGURE 6.2 – Fonction f à minimiser pour l'exemple ??.

FIGURE 6.3 – Itérés de l'algorithme de Newton (Gauche), Gauß-Newton (Droite) pour l'exemple ??.

