

Formes quadratiques, déterminants et calcul différentiel : application à l'économie

Yannick VINCENT
13 juin 2026

Table des matières

I. Formes quadratiques et déterminants	2
1. Définitions et notations	2
2. Formes quadratiques définies positives ou définies négatives	3
3. Formes quadratiques positives ou négatives	4
4. Autre démonstration du critère de Sylvester généralisé	5
5. Formes quadratiques définies positives ou définies négatives sur un sous-espace	7
6. Formes quadratiques positives ou négatives sur un sous-espace	11
II. Extrema, convexité et concavité	14
1. Extrema	14
2. Convexité et concavité : généralités	16
3. Cas des fonctions définies sur $\mathbb{R}$	17
4. Convexité et différentiabilité	18
5. Convexité et extrema	20
6. Résultats équivalents pour la concavité	20
7. Exemples	21
III. Quasi-convexité et quasi-concavité	23
1. Généralités	23
2. Cas des fonctions définies sur $\mathbb{R}$	24
3. Quasi-convexité et différentiabilité	25
4. Résultats équivalents pour la quasi-concavité	28
5. Exemples	29
6. Synthèse des exemples	32
IV. Extrema sous contraintes	33
1. Multiplicateurs de Lagrange	33
2. Reformulation avec le Lagrangien et conditions d'ordre deux	37

Quelques idées en guise d'introduction

- Ce texte provient d'une observation. D'une part, il est possible d'énoncer un critère permettant d'assurer qu'une fonction est quasi-convexe, portant sur les mineurs de la matrice Hessienne bordée (par le gradient) de la fonction considérée (voir Corollaire 44). D'autre part, la matrice Hessienne du Lagrangien est, en elle-même, une matrice bordée. S'agit-il d'une coïncidence ou existe-t-il une raison plus profonde reliant ces deux résultats. Leur démonstration permettent en fait de comprendre qu'ils utilisent tous les deux un même résultat : le critère de Hotelling-Mann permettant d'assurer qu'une forme quadratique restreinte à un sous-espace vectoriel est définie positive (ou semi-définie positive pour le critère de Hotelling-Mann généralisé).

La première partie vise à présenter des résultats sur les formes quadratiques, dont le critère de Hotelling-Mann et sa généralisation. La seconde partie traite de la convexité quand la troisième partie aborde le cas des fonctions quasi-convexes. La dernière partie présente enfin l'optimisation sous-contraintes. Comme nous l'avons expliqué, c'est donc dans les parties 3 et 4 que nous utiliserons les résultats établis dans la première partie. Pour cette raison, certains résultats de la partie 2 portant sur la convexité, ne constituant pas le cœur de ce texte, sont démontrés de manière succincte.

- Pour ce qui est de la quasi-convexité, les résultats portant sur la matrice Hessienne sont assez intéressants mais on pourra noter quelques différences avec le cas de la convexité (voir Théorème 32). Ils ne nécessitent pas seulement que la fonction soit deux fois différentiable mais qu'elle soit de classe $\mathcal{C}^2$. De plus, il n'est pas possible d'énoncer une CNS simple pour qu'une fonction soit quasi-convexe. Enfin, pour ce qui est des conditions suffisantes, on pourra retenir qu'elles nécessitent, soit que le gradient ne s'annule pas, soit que, s'il s'annule en un point, la matrice Hessienne de la fonction (et non pas la matrice Hessienne bordée) soit définie positive en ce point (voir le lemme de Crouzeix : Lemme 41).
- La figure page 32 permet de faire la synthèse des exemples traités sur la convexité et la quasi-convexité. Cela donne un aperçu des différents cas possibles (convexe mais pas strictement quasi-convexe, etc.).
- Une partie concernant les applications à l'économie reste à écrire (preuve de la stricte quasi-concavité des fonctions de Cobb-Douglas notamment).

I. Formes quadratiques et déterminants

1. Définitions et notations

Dans toute la suite, on considère une forme quadratique réelle q sur un $\mathbb{R}$ espace vectoriel E de dimension finie n . On note :

- $b : E \times E \longrightarrow \mathbb{R}$ la forme bilinéaire symétrique associée à q , c'est-à-dire l'unique forme bilinéaire symétrique telle que pour tout $x \in E$, $b(x, x) = q(x)$.
- Si $\mathcal{B} = (e_1, \dots, e_n)$ est une base de E ,

$$\text{Mat}_{\mathcal{B}}(q) = (b(e_i, e_j))_{1 \leq i, j \leq n}.$$

On montre alors que si $\mathcal{B}$ et $\mathcal{B}'$ sont deux bases de E et si P est la matrice de passage de $\mathcal{B}$ vers $\mathcal{B}'$, alors

$$\text{Mat}_{\mathcal{B}'}(q) = {}^t P \text{Mat}_{\mathcal{B}}(q) P$$

et on dit que les matrices $\text{Mat}_{\mathcal{B}}(q)$ et $\text{Mat}_{\mathcal{B}'}(q)$ sont congruentes.

- Si $A \subset E$ est une partie de E ,

$$A^\perp = \left\{ x \in E, \quad \forall y \in A, \quad b(x, y) = 0 \right\}.$$

- Par définition, $\ker q = E^\perp$. En identifiant $E \simeq \mathbb{R}^n$, et si $\mathcal{B}$ est la base canonique de $\mathbb{R}^n$, on a $\ker(q) = \ker(\text{Mat}_{\mathcal{B}}(q))$.
- Si $\ker(q) = \{0\}$, on dit que q est non-dégénérée. Autrement dit, q est non dégénérée si $|H| \neq 0$.
- Si pour tout $x \in E$, $q(x) = 0 \implies x = 0$, on dit que q est définie.
- Il est clair que si q est définie, alors q est non dégénérée. La réciproque est fausse.
- Si pour tout $x \in E$, $q(x) \geq 0$, on dit que q est positive (idem pour négative).
- Si q est une forme quadratique non-dégénérée, on peut montrer que pour tout sous-espace vectoriel F de E ,

$$F \oplus F^\perp = E.$$

- Si $A \in \mathcal{M}_n(\mathbb{R})$ est une matrice symétrique, et si q et q' sont deux formes quadratiques dont la matrice est A , alors q est positive si, et seulement si, q' l'est. Dans ce cas, A est dite positive (idem pour définie positive).
- Si A est une matrice, on notera $|A|$ son déterminant. On définit enfin les mineurs principaux d'une matrice de la façon suivante :

Définition 1 – Mineurs principaux

- Soit $A = (a_{ij})_{1 \leq i, j \leq n} \in \mathcal{M}_n(\mathbb{R})$ et soit $I \subset \{1, \dots, n\}$ avec $I \neq \emptyset$. On note $A_I = (a_{ij})_{i \in I, j \in I}$ et on dit que le réel $|A_I|$ est un **mineur principal** de A .
- On dit que le mineur principal $|A_I|$ est d'**ordre** $\text{Card}(I)$ (que l'on notera aussi $\#I$).
- Soit $k \in \{1, \dots, n\}$. On note $A_k = A_{\{1, \dots, k\}}$ et le réel $|A_k|$ est appelé un **mineur principal dominant** de A .

Remarque. Si $A \in \mathcal{M}_n(\mathbb{R})$ et si $1 \leq k \leq n$, il y a $\binom{n}{k}$ mineurs principaux de A d'ordre k . Au total, A admet donc

$$\sum_{k=1}^n \binom{n}{k} = 2^n - 1 \text{ mineurs principaux.}$$

D'un point de vue algorithmique, pour montrer qu'une forme quadratique est positive ou définie positive, on peut utiliser la réduction de Gauss des formes quadratiques. Les parties qui suivent donnent en fait d'autres caractérisations par les déterminants des sous-matrices de la forme quadratique. Ils sont bien entendus beaucoup moins efficaces en terme de complexité de calculs que la réduction de Gauss.

2. Formes quadratiques définies positives ou définies négatives

Théorème 1 – Critère de Sylvester

Soit q une forme quadratique réelle sur E et soit $\mathcal{B} = (e_1, \dots, e_n)$ une base de E . On note $H = \text{Mat}_{\mathcal{B}}(q)$. Les trois assertions suivantes sont équivalentes :

- (i) q est définie positive;
- (ii) pour tout $I \subset \{1, \dots, n\}$ avec $I \neq \emptyset$, $|H_I| > 0$.
- (iii) pour tout $k \in \{1, \dots, n\}$, $|H_k| > 0$.

Démonstration. [CALDERO et PERONIER, 2022]

(i) $\implies$ (ii) : Supposons que q est définie positive. On a alors $|H| > 0$. En effet, l'algorithme de réduction de Gauss permet de montrer qu'il existe une base orthonormée pour q . Autrement dit, il existe $P \in \text{GL}_n(\mathbb{R})$ (la matrice de passage de la base $\mathcal{B}$ vers cette base orthonormée) telle que $H = {}^t P I_n P$, et on a donc $|H| = |{}^t P P| = |P|^2 > 0$.

Pour $1 \leq k \leq n$ et pour $I = \{i_1, \dots, i_k\} \subset \{1, \dots, n\}$, on considère ensuite la restriction de q au sous-espace vectoriel $\text{Vect}(e_{i_1}, \dots, e_{i_k})$. Il s'agit d'une forme quadratique définie positive et sa matrice est H_I . On en déduit donc que $|H_I| > 0$.

(ii) $\implies$ (iii) est évident.

(iii) $\implies$ (i) : On prouve le résultat par récurrence sur $n = \dim(E)$.

Le résultat est clairement vrai si $n = 1$.

Supposons maintenant qu'il est vrai au rang $n - 1 \geq 1$ et montrons qu'il l'est au rang n .

On considère ainsi E une forme quadratique q sur E de dimension n telle que, pour tout $1 \leq k \leq n$, $|H_k| > 0$. Soit $V = \text{Vect}(e_1, \dots, e_{n-1})$. Comme q est non dégénérée, $V \oplus V^\perp = E$ et on a donc $\dim(V^\perp) = 1$. Par conséquent, il existe $e \in E$ non nul tel que l'on ait la décomposition orthogonale suivante :

$$E = V \oplus \mathbb{R}e.$$

D'après l'hypothèse de récurrence, la restriction de q à V est définie positive. De plus, $q(e) > 0$. En effet, la matrice $K = \text{Mat}_{(e_1, \dots, e_{n-1}, e)}(q)$ est diagonale par blocs, avec $K = \begin{pmatrix} H_{n-1} & 0 \\ 0 & q(e) \end{pmatrix}$. Ainsi, en notant P la matrice de passage de $\mathcal{B}$ vers $(e_1, \dots, e_{n-1}, e)$, on a $K = {}^t P H P$ et on voit donc que $|K| > 0$. Il en résulte que :

$$q(e) = \frac{|K|}{|H_{n-1}|} > 0.$$

Finalement, q est définie positive car pour tout $v \in V$ et tout $\alpha \in \mathbb{R}$,

$$q(v + \alpha e) = q(v) + \alpha^2 q(e) \geq 0$$

et $q(v + \alpha e)$ uniquement si $q(v) = 0$ et si $\alpha = 0$, c'est-à-dire si $v + \alpha e = 0$. □

Corollaire 2

Soit q une forme quadratique réelle sur E et soit $\mathcal{B}$ une base de E . On note $H = \text{Mat}_{\mathcal{B}}(q)$. Alors, q est définie positive si, et seulement si, pour tout $k \in \{1, \dots, n\}$, $(-1)^k |H_k| > 0$.

Démonstration. En posant $\tilde{q} = -q$, on voit que q est définie négative si, et seulement si, $\tilde{q}$ est définie négative. De plus, si $H = \text{Mat}_{\mathcal{B}}(q)$ et $\tilde{H} = \text{Mat}_{\mathcal{B}}(\tilde{q})$, on a $\tilde{H} = -H$ et donc, pour tout $k \in \{1, \dots, n\}$, $|\tilde{H}_k| = (-1)^k |H_k|$. On obtient donc le résultat en appliquant le théorème 1 à $\tilde{q}$. □

3. Formes quadratiques positives ou négatives

Proposition 3 – Critère de Sylvester généralisé

Soit q une forme quadratique réelle et soit $\mathcal{B}$ une base de E . On note $H = \text{Mat}_{\mathcal{B}}(q)$. Alors, q est positive si, et seulement si, pour tout $I \subset \{1, \dots, n\}$ avec $I \neq \emptyset$, $|H_I| \geq 0$.

Démonstration. $\Rightarrow$ Supposons que q est positive. On commence par montrer, comme dans la preuve du théorème 1 que $|H| \geq 0$. On note ensuite $\mathcal{B}_1 = (e_i)_{i \in I}$, $V_1 = \text{Vect}(\mathcal{B}_1)$ et q_1 la restriction de q à V_1 (qui est positive). On a alors $H_1 = \text{Mat}_{\mathcal{B}_1}(q_1)$ et on en déduit que $|H_1| \geq 0$.

$\Leftarrow$ Réciproquement, on va montrer par récurrence sur $n = \dim(E)$ que si l'on suppose que pour tout $I \subset \{1, \dots, n\}$ avec $I \neq \emptyset$, on a $|H_I| \neq 0$, alors q est positive. La propriété est clairement vraie si $n = 1$. Supposons donc qu'elle le soit jusqu'au rang $n - 1 \geq 1$ et montrons qu'alors elle est vraie au rang n .

On distingue deux cas :

- Si $|H| > 0$, on procède comme pour la preuve du critère de Sylvester : on note $V = \text{Vect}(e_1, \dots, e_{n-1})$ et comme q est non dégénérée, il existe $e \neq 0$ tel que l'on ait la somme directe orthogonale $V \oplus \mathbb{R}e = E$. En notant $K = \text{Mat}_{(e_1, \dots, e_{n-1}, e)}(q)$ on a $K = \begin{pmatrix} H_{n-1} & 0 \\ 0 & q(e) \end{pmatrix}$. Cette matrice étant congruente à H , on a $|K| > 0$. De plus, comme $|K| = |H_{n-1}|q(e)$ et que $|H_{n-1}| \geq 0$, on en déduit que $|H_{n-1}| > 0$ et que $q(e) > 0$. Par suite q est définie positive (et donc positive).
- Si $|H| = 0$, cela signifie que q est une forme quadratique dégénérée et qu'il existe $e \in \text{Ker}(q) \setminus \{0\}$. D'après le théorème de la base incomplète, il existe $1 \leq i_1 < i_2 < \dots < i_{n-1} \leq n$ tels que la famille $\mathcal{B} = (e_{i_1}, e_{i_2}, \dots, e_{i_{n-1}}, e)$ soit une base de E . En notant $K = \text{Mat}_{\mathcal{B}}(q)$, on a $K = \begin{pmatrix} H_{\{i_1, \dots, i_{n-1}\}} & 0 \\ 0 & 0 \end{pmatrix}$. On note de plus $V = \text{Vect}(e_{i_1}, \dots, e_{i_{n-1}})$ et on considère la restriction de q à V (qui a pour matrice $H_{\{i_1, \dots, i_{n-1}\}}$). D'après l'hypothèse de récurrence, cette restriction est une forme quadratique positive. Finalement, q est positive car pour tout $v \in V$ et $\alpha \in \mathbb{R}$,

$$q(v + \alpha e) = q(v) \geq 0.$$

□

Remarque. Le premier cas de la démonstration (où $|H| > 0$) prouve par ailleurs que toute forme quadratique positive est définie si, et seulement si, elle est non dégénérée.

Corollaire 4

Soit q une forme quadratique réelle et soit $\mathcal{B}$ une base de E . On note $H = \text{Mat}_{\mathcal{B}}(q)$. Alors, q est négative si, et seulement si, pour tout $I \subset \{1, \dots, n\}$ avec $I \neq \emptyset$, $(-1)^{\#I} |H_I| \geq 0$.

Démonstration. La preuve est similaire à celle du Corollaire 2

□

Remarque. La matrice $H = \begin{pmatrix} 0 & 0 \\ 0 & -1 \end{pmatrix}$ montre qu'il n'est pas possible d'espérer caractériser les formes quadratiques positives en considérant uniquement les mineurs principaux dominants comme c'est le cas pour les formes quadratiques définies positives. Le résultat suivant permet toutefois de raffiner la proposition 3 afin d'obtenir une condition portant sur un nombre restreint de mineurs principaux.

Proposition 5

Soit q une forme quadratique réelle non nulle et soit $\mathcal{B}$ une base de E . On note $H = \text{Mat}_{\mathcal{B}}(q)$. Soit $m = \max\{\#I, |H_I| \neq 0\}$. Alors q est positive si, et seulement si, il existe $I_0 \subset \{1, \dots, n\}$ tel que $\#I_0 = m$ et tel que pour tout $I \subset I_0$, $|H_I| > 0$.

Remarque. On a vu qu'une forme quadratique positive non dégénérée est nécessairement définie positive. Par conséquent, l'hypothèse « pour tout $I \subset I_0$, $|H_I| > 0$ » peut aisément être remplacée par l'hypothèse équivalente « $|H_{I_0}| > 0$ et pour tout $I \subsetneq I_0$, $|H_I| \geq 0$ ».

Démonstration. Supposons que $\#I_0 = m$ et que pour tout $I \subset I_0$, $|H_I| > 0$. Cela signifie que la restriction de q à $V = \text{Vect}(e_i)_{i \in I}$ est définie positive. De plus, le rang d'une matrice étant l'ordre maximal d'un mineur non nul, on a $\text{rg}(H) \geq m$. En fait, on a même $\text{rg}(H) = m$ car il est possible de construire une famille libre de $n - m$ vecteurs de $\text{Ker}(q)$ (cela découle de la construction établie dans la preuve de la proposition 3 et du fait que si $\#I > m$, alors $|H_I| = 0$). Au final, en notant $(f_{m+1}, \dots, f_n)$ une base du noyau de q et en notant $\mathcal{B}$ la base obtenue en concaténant la famille $(e_i)_{i \in I}$ et la famille $(f_{m+1}, \dots, f_n)$, on a

$$\text{Mat}_{\mathcal{B}}(q) = \begin{pmatrix} H_I & 0 \\ 0 & 0 \end{pmatrix}.$$

Cela finit de prouver que q est bien positive. $\square$

Corollaire 6

Soit q une forme quadratique réelle non nulle et soit $\mathcal{B}$ une base de E . On note $H = \text{Mat}_{\mathcal{B}}(q)$. Soit $m = \max \{ \#I, |H_I| \neq 0 \}$. Alors q est négative si, et seulement si, il existe $I_0 \subset \{1, \dots, n\}$ tel que $\#I_0 = m$ et tel que pour tout $I \subset I_0$, $(-1)^{\#I} |H_I| > 0$.

4. Autre démonstration du critère de Sylvester généralisé

Il est possible de démontrer la proposition 3 comme une sorte de corollaire du critère de Sylvester (Théorème 1). On utilise pour cela les deux lemmes suivants. Le premier (Lemme 7) fait effectivement le lien entre forme quadratique positive et forme quadratique définie positive quand le second (Lemme 8) est un résultat plus calculatoire. Cette façon de procéder, présentée dans [??] aura en outre l'avantage de fonctionner dans le cas de la généralisation du critère de Hotelling-Mann (Théorème 9).

Lemme 7

Soit $H \in \mathcal{M}_n(\mathbb{R})$ une matrice symétrique. H est positive si, et seulement si, pour tout $\alpha > 0$, la matrice $H + \alpha I_n$ est symétrique définie positive.

Démonstration. [DEBREU, 1952]

La preuve est immédiate par passage à la limite. $\square$

On peut alors proposer une autre démonstration de la proposition 3 grâce au lemme suivant.

Lemme 8

Pour tout $H \in \mathcal{M}_n(\mathbb{R})$, pour tout $\alpha \in \mathbb{R}$,

$$|H + \alpha I_n| = \alpha^n + \sum_{k=0}^{n-1} \left(\sum_{\substack{I \subset \{1, \dots, n\} \\ \#I = n-k}} |H_I| \right) \alpha^k$$

Démonstration. [DEBREU, 1952]

Soit $H = (a_{i,j})_{1 \leq i,j \leq n} \in \mathcal{M}_n(\mathbb{R})$ et $\alpha \in \mathbb{R}$. On notera :

- $S(I)$ le groupe des permutations de l'ensemble I et, plus particulièrement, S_n le groupe symétrique de l'ensemble $\{1, \dots, n\}$;
- $\mathcal{P}(\{1, \dots, n\})$ l'ensemble des parties de $\{1, \dots, n\}$ et $\mathcal{P}_k(\{1, \dots, n\})$ l'ensemble des parties de $\{1, \dots, n\}$ ayant k éléments;
- Pour $\sigma \in S_n$ une permutation, $\text{Fix}(\sigma) = \{k \in \{1, \dots, n\}, \sigma(k) = k\}$ l'ensemble des points fixes de σ ;
- $\Sigma_k^n = \{\sigma \in S_n, \# \text{Fix}(\sigma) = k\}$ l'ensemble des permutations ayant exactement k points fixes;
- $\Sigma_{\geq k}^n = \bigcup_{j=k}^n \Sigma_j^n$ l'ensemble des permutations ayant au moins k points fixes.

On rappelle de plus la convention : $\prod_{i \in \emptyset} x_i = 1$.

D'après la formule de Leibniz ¹,

$$\begin{aligned}
 |H + \alpha I_n| &= \sum_{\sigma \in S_n} \varepsilon(\sigma) \prod_{i=1}^n (a_{i,\sigma(i)} + \alpha \delta_{i,\sigma(i)}) \\
 &= \sum_{j=0}^n \sum_{\sigma \in \Sigma_j^n} \varepsilon(\sigma) \left(\prod_{i \notin \text{Fix}(\sigma)} a_{i,\sigma(i)} \right) \left(\prod_{i \in \text{Fix}(\sigma)} (a_{i,\sigma(i)} + \alpha) \right) \\
 &= \sum_{j=0}^n \sum_{\sigma \in \Sigma_j^n} \varepsilon(\sigma) \left(\prod_{i \notin \text{Fix}(\sigma)} a_{i,\sigma(i)} \right) \left(\sum_{k=0}^j \left(\sum_{\substack{S \subset \text{Fix}(\sigma) \\ \#S=j-k}} \prod_{i \in S} a_{i,\sigma(i)} \right) \alpha^k \right) \\
 &= \sum_{j=0}^n \sum_{\sigma \in \Sigma_j^n} \varepsilon(\sigma) \left(\sum_{k=0}^j \left(\sum_{\substack{S \subset \text{Fix}(\sigma) \\ \#S=j-k}} \left(\prod_{i \notin \text{Fix}(\sigma)} a_{i,\sigma(i)} \prod_{i \in S} a_{i,\sigma(i)} \right) \right) \alpha^k \right) \\
 &= \sum_{k=0}^n \sum_{j=k}^n \left(\sum_{\sigma \in \Sigma_j^n} \varepsilon(\sigma) \left(\sum_{\substack{S \subset \text{Fix}(\sigma) \\ \#S=j-k}} \left(\prod_{i \notin \text{Fix}(\sigma)} a_{i,\sigma(i)} \prod_{i \in S} a_{i,\sigma(i)} \right) \right) \alpha^k \right) \\
 &= \sum_{k=0}^n \left(\sum_{j=k}^n \sum_{\sigma \in \Sigma_j^n} \sum_{\substack{S \subset \text{Fix}(\sigma) \\ \#S=j-k}} \varepsilon(\sigma) \prod_{i \in \text{Fix}(\sigma)^c \cup S} a_{i,\sigma(i)} \right) \alpha^k \\
 &= \sum_{k=0}^n \left(\sum_{\sigma \in \Sigma_{\geq k}^n} \sum_{\substack{S \subset \text{Fix}(\sigma) \\ \#S=\#\text{Fix}(\sigma)-k}} \varepsilon(\sigma) \prod_{i \in \text{Fix}(\sigma)^c \cup S} a_{i,\sigma(i)} \right) \alpha^k \\
 &= \sum_{k=0}^n \left(\sum_{(\sigma, I) \in \mathcal{P}_1} \varepsilon(\sigma) \prod_{i \in I} a_{i,\sigma(i)} \right) \alpha^k
 \end{aligned}$$

où $\mathcal{P}_1 = \{(\sigma, I) \mid \sigma \in \Sigma_{\geq k}^n, S \subset \text{Fix}(\sigma) \text{ et } \#S = \#\text{Fix}(\sigma) - k\} \subset S_n \times \mathcal{P}(\{1, \dots, n\})$.

Or, en posant $\mathcal{P}_2 = \{(\sigma, I) \mid \sigma \in S_n, I \in \mathcal{P}_{n-k}(\{1, \dots, n\}) \text{ et } \sigma|_I = \text{Id}\}$, on montre facilement que $\mathcal{P}_1 = \mathcal{P}_2$. En effet :

- Si $(\sigma, I) \in \mathcal{P}_1$, alors $I = \text{Fix}(\sigma)^c \cup S$ donc $I^c = \text{Fix}(\sigma) \cap S^c$ et on a bien, par définition de $\text{Fix}(\sigma)$, $\sigma|_{I^c} = \text{Id}$. De plus, comme $S \subset \text{Fix}(\sigma)$, on a $\#I = \#\text{Fix}(\sigma)^c + \#S = n - k$. Cela prouve bien que $(\sigma, I) \in \mathcal{P}_2$.
- Inversement, si $(\sigma, I) \in \mathcal{P}_2$, il est clair que $\sigma \in \Sigma_{\geq k}^n$. De plus, en posant $S = \text{Fix}(\sigma) \setminus I^c = \text{Fix}(\sigma) \cap I$, on a bien $S \subset \text{Fix}(\sigma)$ et $S \cup \text{Fix}(\sigma)^c = I$, ce qui signifie que $(\sigma, I) \in \mathcal{P}_1$.

Ainsi,

$$|H + \alpha I_n| = \sum_{k=0}^n \left(\sum_{(\sigma, I) \in \mathcal{P}_2} \varepsilon(\sigma) \prod_{i \in I} a_{i,\sigma(i)} \right) \alpha^k$$

Pour $(\sigma, I) \in \mathcal{P}_2$, on identifiera naturellement σ à un élément de $S(I)$. Remarquons que sa signature dans $S(I)$ est identique à celle dans S_n . On obtient alors, en adoptant la convention $|H_\emptyset| = 1$:

$$\begin{aligned}
 |H + \alpha I_n| &= \sum_{k=0}^n \left(\sum_{\substack{I \subset \{1, \dots, n\} \\ \#I=n-k}} \sum_{\sigma \in S(I)} \varepsilon(\sigma) \prod_{i \in I} a_{i,\sigma(i)} \right) \alpha^k \\
 &= \sum_{k=0}^n \left(\sum_{\substack{I \subset \{1, \dots, n\} \\ \#I=n-k}} |H_I| \right) \alpha^k
 \end{aligned}$$

□

Démonstration. (de la proposition 3) [DEBREU, 1952]

1. On entend, par « formule de Leibniz », pour $A = (a_{i,j})_{1 \leq i,j \leq n} \in \mathcal{M}_n(\mathbb{R})$, $|A| = \sum_{\sigma \in S_n} \varepsilon(\sigma) \prod_{i=1}^n a_{i,\sigma(i)}$ où $\varepsilon(\sigma)$ désigne la signature de la permutation σ .

- $\Rightarrow$ Supposons que q est positive. Soit $I \subset \{1, \dots, n\}$ avec $I \neq \emptyset$. Alors la restriction de q à $\text{Vect}(e_i)_{i \in I}$ est positive. Par conséquent, si on note $m = \#I$, on sait, d'après le lemme 7, que pour tout $\alpha > 0$, $H_I + \alpha I_m$ est définie positive. D'après le critère de Sylvester, on en déduit que pour tout $\alpha > 0$, $|H_I + \alpha I_m| > 0$. Par continuité du déterminant, cela implique nécessairement que $|H_I| \geq 0$.
- $\Leftarrow$ Réciproquement, supposons que pour tout $I \subset \{1, \dots, n\}$ avec $I \neq \emptyset$, $|H_I| \geq 0$. Alors, d'après le lemme 8, pour tout $\alpha > 0$, pour tout $1 \leq k \leq n$:

$$|(H + \alpha I_n)_k| = \alpha^k + \sum_{i=0}^{k-1} \alpha^i \sum_{\substack{I \subset \{1, \dots, k\} \\ \#I = k-i}} |H_I|$$

Or, tous les $|H_I|$ sont positifs par hypothèse et on en déduit donc que pour tout $\alpha > 0$, pour tout $1 \leq k \leq n$, $|(H + \alpha I_n)_k| > 0$. D'après le critère de Sylvester, cela signifie exactement que $H + \alpha I_n$ est définie positive. D'après le lemme 7, on en déduit que H est positive. $\square$

5. Formes quadratiques définies positives ou définies négatives sur un sous-espace

Soit $n \geq 2$. On considère un $\mathbb{R}$ espace vectoriel E muni d'une base $\mathcal{B} = (e_1, \dots, e_n)$ et on note $\mathcal{B}^* = (e_i^*)_{1 \leq i \leq n}$ la base duale de la base $\mathcal{B}$. On considère alors une famille libre $(\varphi_k)_{1 \leq k \leq m}$ de formes linéaires (pour $1 \leq m \leq n-1$) et on pose $F = \bigcap_{k=1}^m \text{Ker}(\varphi_k)$. On note de plus $B = (b_{i,j})_{1 \leq i,j \leq n} = {}^t \text{Mat}_{\mathcal{B}^*}(\varphi_k) \in \mathcal{M}_{m,n}(\mathbb{R})$. Autrement dit,

B est une matrice de rang m et pour tout $X = {}^t(x_1, \dots, x_n) \in \mathbb{R}^n$, $\sum_{i=1}^n x_i e_i \in F$ si, et seulement si, $BX = 0$.

Définition 2 – Matrice bordée

Soit $H \in \mathcal{M}_n(\mathbb{R})$ et soit $B \in \mathcal{M}_{m,n}(\mathbb{R})$. On appelle **matrice bordée** de H par B , la matrice

$$\overline{H}^B = \begin{pmatrix} 0 & B \\ {}^t B & H \end{pmatrix} \in \mathcal{M}_{m+n}(\mathbb{R}).$$

Avant d'énoncer un critère pour qu'une forme quadratique restreinte à un sous-espace soit définie positive, on donne également la définition suivante qui généralise la notion de sous-matrice au cas de matrices non carrées.

Définition 3

Soit $A = (a_{i,j})_{1 \leq i \leq p, 1 \leq j \leq q} \in \mathcal{M}_{p,q}(\mathbb{R})$.

- Pour $I \subset \{1, \dots, p\}$ et $J \subset \{1, \dots, q\}$ avec $I \neq \emptyset$ et $J \neq \emptyset$, on pose $A_{I,J} = (a_{i,j})_{i \in I, j \in J}$ la matrice formée par les lignes et les colonnes dont les indices appartiennent respectivement à I et J .
- Pour $1 \leq m \leq p$ et $1 \leq n \leq q$, on note $A_{m,n} = A_{\{1, \dots, m\}, \{1, \dots, n\}} = (a_{i,j})_{1 \leq i \leq m, 1 \leq j \leq n}$ la matrice formée par les p premières lignes et les q premières colonnes de A .
- Pour $1 \leq n \leq \min(p, q)$, on note $A_n = A_{n,n}$ la matrice formée par les n premières lignes et colonnes de A .

On peut alors énoncer le critère suivant.

Théorème 9 – Critère de Hotelling–Mann

Soit $n \geq 2$. Soit q une forme quadratique sur E (de dimension n) et soit $\mathcal{B}$ une base de E . On note $H = \text{Mat}_{\mathcal{B}}(q)$. On reprend les notations ci-dessus et on suppose que $|B_m| \neq 0$. Alors la restriction de q à l'espace F est définie positive (c'est à dire que pour tout $x \in F \setminus \{0\}$, $q(x) > 0$) si, et seulement si, pour tout $m+1 \leq k \leq n$, $(-1)^m \left| \overline{H}_k^{B_{m,k}} \right| > 0$.

Remarque.

- On rappelle que, par définition, pour $m+1 \leq k \leq n$, on a $\overline{H}_k^{B_{m,k}} = \begin{pmatrix} 0 & B_{m,k} \\ {}^t B_{m,k} & H_k \end{pmatrix} = (\overline{H}^B)_{m+k}$.

- Le nom « Hotelling-Mann » n'est pas standard pour désigner ce résultat. Nous avons toutefois décidé de retenir cette appellation car [DEBREU, 1952] explique que Hotelling a établi le résultat dans le cas d'un hyperplan ($m = 1$) avant que Mann ne le généralise en toute dimension.

Pour démontrer le critère de Hotelling-Mann, on commence par établir deux lemmes.

Lemme 10

Avec les notation du théorème 9, la restriction de q à F est définie positive si, et seulement si, il existe $\lambda_0 \in \mathbb{R}$ tel que pour tout $\lambda \geq \lambda_0$ et pour tout $X \in \mathbb{R}^n \setminus \{0\}$,

$$X^t H X + \lambda^t X^t B B X > 0.$$

Autrement dit, en identifiant E à $\mathbb{R}^n$, la restriction de q à $\text{Ker}(B) \setminus \{0\}$ est définie positive si, et seulement si, il existe $\lambda \in \mathbb{R}$ tel que $H + \lambda^t B B$ est une matrice symétrique définie positive.

Démonstration. (du lemme 10) [DEBREU, 1952]

- S'il existe λ tel que $H + \lambda^t B B$ soit définie positive, il est clair que pour tout $X \in \text{Ker}(B) \setminus \{0\}$, $^t X H X > 0$, et donc que la restriction de q à F est définie positive.
- Réciproquement, supposons que H est définie positive. Montrons l'existence d'un λ_0 vérifiant la propriété annoncée par le lemme. Il s'agit en fait de prouver que la fonction θ définie pour tout $X \in \mathbb{R}^n \setminus \text{Ker}(B)$ par

$$\theta(X) = -\frac{^t X H X}{^t X^t B B X} = -\frac{^t X H X}{^t (B X) B X}$$

est majorée (il suffira alors de considérer $\lambda_0 > \sup \theta$). Or, il est clair que pour tout $\lambda \in \mathbb{R}^*$, pour tout $X \in \mathbb{R}^n \setminus \text{Ker}(B)$, $\theta(\lambda X) = \theta(X)$. Il suffit donc de montrer que θ est majorée sur $V = \mathbb{S}^{n-1} \cap \text{Ker}(B)$ (où $\mathbb{S}^{n-1} = \{X \in \mathbb{R}^n, ^t X X = 1\}$ désigne la sphère unité de $\mathbb{R}^n$). En raisonnant par l'absurde, supposons au contraire que θ ne soit pas majorée sur V . Il existerait ainsi une suite $X_k \in V$ telle que $\theta(X_k) \xrightarrow[k \rightarrow +\infty]{} +\infty$.

Comme $\mathbb{S}^{n-1}$ est compact, il existerait donc une sous-suite X_{k_l} convergeant vers $X \in \mathbb{S}_{n-1}$.

Si $X \in V$, alors $\theta(X_k) \rightarrow \theta(X)$ par continuité de θ sur V .

Sinon, $X \notin V$, ce qui signifie que $X \in \text{Ker}(B)$. Dans ce cas, sachant que $^t X H X > 0$, et que pour tout k , $^t X_k^t B B X_k = ^t (B X_k) (B X_k) > 0$, on voit que $\theta(X_k) \xrightarrow[k \rightarrow +\infty]{} -\infty$.

Dans tous les cas, on voit que la convergence de $\theta(X_k)$ vers $+\infty$ est impossible, ce qui conclut la démonstration. □

Lemme 11

Soient $H \in \mathcal{M}_n(\mathbb{R})$ et $B \in \mathcal{M}_{m,n}(\mathbb{R})$. $|H + \lambda^t B B|$ est un polynôme en λ de degré au plus m . Le coefficient devant λ^m de ce polynôme est :

$$(-1)^m |\overline{H}^B|.$$

Remarque. Attention, il se peut que le coefficient devant λ^m soit nul, c'est-à-dire que l'on ait $|\overline{H}^B| = 0$.

Démonstration. (du lemme 11) [DEBREU, 1952]

On a l'égalité matricielle suivante :

$$\begin{pmatrix} -I_m & B \\ \lambda^t B & H \end{pmatrix} \begin{pmatrix} I_m & B \\ 0 & I_m \end{pmatrix} = \begin{pmatrix} -I_m & 0 \\ \lambda^t B & H + \lambda^t B B \end{pmatrix}.$$

Puis, en passant au déterminant :

$$\begin{vmatrix} -I_m & B \\ \lambda^t B & H \end{vmatrix} = (-1)^m |H + \lambda^t B B|.$$

Par conséquent,

$$|H + \lambda^t B B| = (-1)^m \begin{vmatrix} -I_m & B \\ \lambda^t B & H \end{vmatrix}$$

et il en résulte, d'après la formule de Leibniz, que $|H + \lambda^t B B|$ est bien un polynôme en λ de degré au plus m . Par ailleurs, on voit à l'aide de cette formule de Leibniz que le coefficient devant λ^m est obtenu comme

une somme de termes correspondant aux permutations laissant stable l'ensemble $\{m+1, \dots, n\}$. Ainsi, pour toute matrice $A \in \mathcal{M}_m(\mathbb{R})$, le coefficient devant λ^m de $\begin{vmatrix} -I_m & B \\ \lambda^t B & H \end{vmatrix}$ est le même que celui de $\begin{vmatrix} A & B \\ \lambda^t B & H \end{vmatrix}$. En particulier, pour $A = 0$, on voit que le coefficient devant λ^m de $|H + \lambda^t BB|$ est le même que celui de $\begin{vmatrix} 0 & B \\ \lambda^t B & H \end{vmatrix}$. Or, par multilinéarité du déterminant, on a

$$\begin{vmatrix} 0 & B \\ \lambda^t B & H \end{vmatrix} = \lambda^m \overline{H^B},$$

d'où l'on déduit le résultat annoncé. $\square$

On peut maintenant démontrer le critère de Hotelling-Mann.

Démonstration. (du Théorème 9) [DEBREU, 1952]

- $\Rightarrow$ Supposons que la restriction de q à F est définie positive. On commence par montrer que $(-1)^m \overline{H^B} > 0$ et le cas des sous-mineurs principaux se traitera ensuite de manière similaire. D'après le lemme 10, il existe $\lambda_0 \in \mathbb{R}$ tel que pour tout $\lambda \geq \lambda_0$, pour tout $X \in \mathbb{R}^n \setminus \{0\}$, ${}^t X H X + \lambda^t X^t B B X > 0$. Par conséquent, d'après le critère de Sylvester (Théorème 1), on a :

$$\text{pour tout } \lambda \geq \lambda_0, |H + \lambda^t B B| > 0 \quad (*).$$

Commençons par montrer que $\overline{H^B} \neq 0$, autrement dit que la matrice $\overline{H^B}$ est inversible.

Soit $X = \begin{pmatrix} x \\ y \end{pmatrix} \in \text{Ker}(\overline{H})$ (avec $x \in \mathbb{R}^m$ et $y \in \mathbb{R}^n$). On a $\begin{cases} B y = 0 \\ {}^t B x + H y = 0 \end{cases}$. En multipliant la deuxième équation par ${}^t y$ à gauche, on obtient :

$$\begin{aligned} & {}^t y^t B x + {}^t y H y = 0 \\ \text{donc} & \quad {}^t (B y) x + {}^t y H y = 0 \\ \text{donc} & \quad {}^t y H y = 0 \end{aligned}$$

En identifiant F à $\text{Ker}(B)$, la restriction de q à $\text{Ker}(B)$ est, par hypothèse, définie positive. On en déduit donc que $y = 0$ et, par suite, que ${}^t B x = 0$ puis que $x = 0$. Pour justifier l'implication ${}^t B x = 0 \Rightarrow x = 0$, il suffit de voir que ${}^t B_m$ est inversible (car B_m l'est), donc que $\text{rg}({}^t B) \geq m$, puis que, d'après le théorème du rang, $\dim(\text{Ker}({}^t B)) = m - \text{rg}({}^t B) = 0$. Cela conclut la preuve du fait que $\text{Ker}(\overline{H}) = \{0\}$ et donc de l'inversibilité de $\overline{H^B}$.

Ainsi, comme $\overline{H^B}$ est inversible, $|H + \lambda^t B B|$ est un polynôme de degré exactement m et l'égalité (*)

implique nécessairement que le coefficient dominant, à savoir $(-1)^m \overline{H^B}$, est strictement positif.

Soit finalement $m+1 \leq k \leq n$. On considère la forme quadratique $\tilde{q}$ définie comme étant la restriction de q sur $\text{Vect}((e_i)_{1 \leq i \leq k})$. En notant $\tilde{\mathcal{B}} = (e_i)_{1 \leq i \leq k}$, on a $\text{Mat}_{\tilde{\mathcal{B}}}(\tilde{q}) = H_k$. On note de plus $\tilde{B} = B_{m,k}$. Toujours en identifiant E à $\mathbb{R}^n$, il est immédiat de vérifier que $\tilde{q}$ restreinte à $\text{Ker}(\tilde{B})$ est définie positive. En effet, si ${}^t(x_1, \dots, x_k) \in \text{Ker}(\tilde{B}) \setminus \{0\}$, alors ${}^t(x_1, \dots, x_k, 0, \dots, 0) \in \text{Ker}(B)$ et on a donc $\tilde{q}\left(\sum_{i=1}^k x_i e_i\right) =$

$$q\left(\sum_{i=1}^k x_i e_i\right) > 0.$$

Ainsi, comme la restriction de $\tilde{q}$ à $\text{Ker}(\tilde{B})$ est définie positive et que $|\tilde{B}_m| = |B_m|$, on a, d'après ce qui a été fait précédemment,

$$(-1)^m \begin{vmatrix} 0 & \tilde{B} \\ {}^t \tilde{B} & H_k \end{vmatrix} > 0.$$

Cela signifie exactement que $(-1)^m \overline{H_k^{B_{m,k}}} > 0$.

- $\Leftarrow$ Réciproquement, supposons que pour tout $k \geq m+1$, $(-1)^m \overline{H_k^{B_{m,k}}} > 0$. D'après le lemme 10, on sait qu'il suffit de montrer qu'il existe $\lambda_0 \in \mathbb{R}$ tel que pour tout $\lambda \geq \lambda_0$, la matrice $H + \lambda^t B B$ est définie positive. D'après le critère de Sylvester, cela équivaut d'ailleurs au fait que tous les mineurs principaux dominants de cette matrice sont strictement positifs. Or, $(H + \lambda^t B B)_j = H_j + \lambda({}^t B B)_j =$

$H_j + \lambda {}^t B_{j,m} B_{m,j}$. Il s'agit donc de montrer que pour tout $1 \leq j \leq n$, il existe $\lambda_0 \in \mathbb{R}$ tel que pour tout $\lambda \geq \lambda_0$, $|H_j + \lambda {}^t(B_{m,j})B_{m,j}| > 0$.

On distingue alors les cas selon la valeur de j .

- Si $j \geq m+1$: d'après le lemme 11, $|H_j + \lambda {}^t(B_{m,j})B_{m,j}|$ a le même signe, pour λ suffisamment grand, que $(-1)^m \left| \overline{H_j}^{B_{m,j}} \right|$. Or,

$$(-1)^m \left| \overline{H_j}^{B_{m,j}} \right| = (-1)^m \left| \begin{array}{cc} 0 & B_{m,j} \\ {}^t B_{m,j} & H_j \end{array} \right| = (-1)^m \left| \begin{pmatrix} 0 & B \\ {}^t B & H \end{pmatrix}_{m+j} \right| = (-1)^m \left| \overline{H_j}^{B_{m,j}} \right|,$$

et on sait que $(-1)^m \left| \overline{H_j}^{B_{m,j}} \right| > 0$ par hypothèse.

- Si $1 \leq j \leq m$: d'après le calcul fait dans le lemme 11,

$$|H_j + \lambda {}^t(B_{m,j})B_{m,j}| = (-1)^m \left| \begin{array}{cc} -I_m & B_{m,j} \\ \lambda {}^t(B_{m,j}) & H_j \end{array} \right| = \lambda^j (-1)^m \left| \begin{array}{cc} -I_m & B_{m,j} \\ {}^t(B_{m,j}) & \frac{1}{\lambda} H_j \end{array} \right|.$$

Il s'agit donc de montrer que pour λ suffisamment grand, $(-1)^m \left| \begin{array}{cc} -I_m & B_{m,j} \\ {}^t(B_{m,j}) & \frac{1}{\lambda} H_j \end{array} \right| > 0$. Or, lorsque λ tend vers $+\infty$, $\frac{1}{\lambda} H_j$ tend vers 0 et il suffit donc de montrer, par continuité du déterminant, que

$$(-1)^m \left| \begin{array}{cc} -I_m & B_{m,j} \\ {}^t(B_{m,j}) & 0 \end{array} \right| > 0.$$

Or, la forme quadratique $\hat{q}$ définie sur $\mathbb{R}^m$ par $\hat{q}(Y) = -{}^t Y Y$ est définie négative et elle l'est, a fortiori sur $\text{Ker}({}^t(B_{m,j}))$. Ainsi, en utilisant le corollaire 12 qui suit², on en déduit (pour $k = j$)

que $(-1)^{2j} \left| \begin{array}{cc} 0 & {}^t(B_{m,j}) \\ B_{m,j} & -I_m \end{array} \right| > 0$, ce qui, en passant à la transposée, donne bien le résultat souhaité.

□

Corollaire 12

Soit q une forme quadratique et soit $\mathcal{B}$ une base de E . On note $H = \text{Mat}_{\mathcal{B}}(q)$. On reprend les notations ci-dessus et on suppose que $|B_m| \neq 0$. Alors la restriction de q à l'espace F est définie négative si, et seulement si, pour tout $m+1 \leq k \leq n$, $(-1)^k \left| \overline{H_k}^{B_{m,k}} \right| > 0$.

Démonstration. [DEBREU, 1952]

On pose $\tilde{q} = -q$ et $\tilde{B} = -B$. Ainsi, q restreinte à $\text{Ker}(B)$ est définie négative si, et seulement si, $\tilde{q}$ restreinte à $\text{Ker}(\tilde{B}) = \text{Ker}(B)$ est définie positive. En notant $\tilde{H} = \text{Mat}_{\mathcal{B}}(\tilde{q})$, on a $\tilde{H}^B = -\tilde{H}^{\tilde{B}}$. Ainsi,

La restriction de q à $\text{Ker}(B)$ est définie négative $\iff$ la restriction de $\tilde{q}$ à $\text{Ker}(\tilde{B})$ est définie positive

$$\iff \forall m+1 \leq k \leq n, \quad (-1)^m \left| \overline{\tilde{H}_k}^{\tilde{B}_{m,k}} \right| > 0$$

$$\iff \forall m+1 \leq k \leq n, \quad (-1)^m \left| -\overline{H_k}^{B_{m,k}} \right| > 0$$

$$\iff \forall m+1 \leq k \leq n, \quad (-1)^{k+2m} \left| \overline{H_k}^{B_{m,k}} \right| > 0$$

$$\iff \forall m+1 \leq k \leq n, \quad (-1)^k \left| \overline{H_k}^{B_{m,k}} \right| > 0$$

□

Remarque. Dans le cas où B est de rang m mais que l'on n'a pas $|B_m| \neq 0$, il est toujours possible, de se ramener à un tel cas en permutant les coordonnées.

Exemple 1. Sur $E = \mathbb{R}^3$, on définit la forme quadratique q_α par

$$q_\alpha({}^t(x_1, x_2, x_3)) = x_1^2 + x_2^2 - x_3^2.$$

2. La condition nécessaire du Corollaire découle simplement de la condition nécessaire du théorème déjà démontré ci-dessus.

En notant $\mathcal{B} = (e_1, e_2, e_3)$ la base canonique de $\mathbb{R}^3$, on a ainsi, $H_\alpha = \text{Mat}_{\mathcal{B}}(q_\alpha) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{pmatrix}$.

On pose de plus $F = \{^t(x_1, x_2, x_3) \in \mathbb{R}^3, x_3 = 0\} \simeq \text{Ker}(B)$ (où $B = \begin{pmatrix} 0 & 0 & 1 \end{pmatrix}$). Il est clair que q_α est définie positive sur F . En revanche, il n'est pas possible d'utiliser le critère de Hotelling-Mann (Théorème 9) pour le justifier car $|B_1| = 0$. On a d'ailleurs, $\left| \left(\overline{H}^B \right)_3 \right| = 0$.

Comme indiqué dans la remarque ci-dessus, on va donc permuter les coordonnées. On pose $P = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$ puis on définit la forme quadratique q_β par

$$q_\beta(^t(y_1, y_2, y_3)) = q_\alpha \circ P^{-1}(y_1, y_2, y_3) = y_2^2 + y_3^2 - y_1^2$$

On a ainsi

$$H_\beta = \text{Mat}_{\mathcal{B}}(q_\beta) = \begin{pmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Il est clair que q_α restreinte à $\text{Ker}(B)$ est définie positive si, et seulement si, q_β restreinte à $P(\text{Ker}(B))$ est définie positive. Or,

$$P(\text{Ker}(B)) = P(\text{Vect}(e_1, e_2)) = \text{Vect}(e_2, e_3) = \{^t(x_1, x_2, x_3) \in \mathbb{R}^3, x_1 = 0\} = \text{Ker}(\tilde{B}).$$

$$\text{où } \tilde{B} = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix}.$$

On a

$$\overline{H}_\beta^{\tilde{B}} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

et on voit que l'on peut bien appliquer le critère de Hotelling-Mann à q_β (plus précisément, on a $\left| \left(\overline{H}_\beta^{\tilde{B}} \right)_3 \right| = -1$ et $\left| \left(\overline{H}_\beta^{\tilde{B}} \right)_4 \right| = -1$, ce qui indique que q_β restreinte à $\text{Ker}(\tilde{B})$ est définie positive).

6. Formes quadratiques positives ou négatives sur un sous-espace

De manière similaire au lemme 7, on peut démontrer le lemme suivant par simple passage à la limite :

Lemme 13

Soit $H \in \mathcal{M}_n(\mathbb{R})$ une matrice symétrique et $B \in \mathcal{M}_{m,n}(\mathbb{R})$. Pour tout $X \in \text{Ker}(B)$, ${}^tXHX \geq 0$ si, et seulement si, pour tout $\alpha > 0$, pour tout $X \in \text{Ker}(B) \setminus \{0\}$, ${}^tXHX + \alpha {}^tXX > 0$.

On peut alors caractériser les formes quadratiques dont la restriction à $\text{Ker}(B)$ est positive :

Proposition 14 – Critère de Hotelling-Mann généralisé

Soit q une forme quadratique et soit $\mathcal{B}$ une base de E . On note $H = \text{Mat}_{\mathcal{B}}(q)$. On reprend les notations de la partie précédente et on suppose que $|B_m| \neq 0$. Alors la restriction de q à l'espace $F \simeq \text{Ker}(B)$ est positive si, et seulement si, pour tout $I \subset \{1, \dots, n\}$ tel que $m+1 \leq \#I \leq n$, $(-1)^m \left| \overline{H}_I^{B_{\{1, \dots, m\}, I}} \right| \geq 0$.

Remarque. On rappelle que, par définition, $\overline{H}_I^{B_{\{1, \dots, m\}, I}} = \begin{pmatrix} 0 & B_{\{1, \dots, m\}, I} \\ {}^tB_{\{1, \dots, m\}, I} & H_I \end{pmatrix}$.

Démonstration. (de la proposition 14) [DEBREU, 1952]

- $\Rightarrow$ Supposons que la restriction de q à $F \simeq \text{Ker}(B)$ est positive et considérons un ensemble $I \subset \{1, \dots, n\}$ tel que $m+1 \leq \#I \leq n$. A fortiori, la restriction de q à $F \cap \text{Vect}(e_i)_{i \in I}$ est positive. Autrement dit, en notant q_I la restriction de q à $\text{Vect}(e_i)_{i \in I}$, on voit que la restriction de q_I à $F \cap \text{Vect}(e_i)_{i \in I}$ est

positive. Ainsi, en faisant l'identification $E \simeq \mathbb{R}^n$, et d'après le lemme 13, la restriction de $H_I + \alpha I_{\#I}$ à $F \cap \text{Vect}(e_i)_{i \in I}$ est également définie positive pour tout $\alpha > 0$. Or, pour tout $X = \sum_{i \in I} x_i e_i \in \text{Vect}(e_i)_{i \in I}$,

$$X \in F \iff B_{\{1, \dots, m\}, I} X = 0.$$

D'après le critère de Hotelling-Mann (Théorème 9), pour $k = \#I + m$:

$$(-1)^m \left| \overline{(H_I + \alpha I_{\#I})}^{B_{\{1, \dots, m\}, I}} \right| > 0.$$

Cette égalité étant vraie pour tout $\alpha > 0$, on en déduit, par passage à la limite, quand α tend vers 0, que

$$(-1)^m \left| \overline{H_I}^{B_{\{1, \dots, m\}, I}} \right| \geq 0.$$

- $\Leftarrow$ Réciproquement, supposons que pour tout $I \subset \{1, \dots, n\}$ tel que $m+1 \leq \#I \leq n$, $(-1)^m \left| \overline{H_I}^{B_{\{1, \dots, m\}, I}} \right| \geq 0$ et montrons qu'alors q restreinte à $\text{Ker}(B)$ est positive. D'après le lemme 13 et le critère de Hotelling, il suffit de prouver que pour tout $k \in \{2m+1, \dots, n+m\}$, pour tout $\alpha > 0$, $(-1)^m \left| \overline{(H + \alpha I_n)}^B_k \right| > 0$. Autrement dit, on veut prouver que pour tout $k \in \{m+1, \dots, n\}$, pour tout $\alpha > 0$,

$$(-1)^m \left| \overline{(H + \alpha I_n)}^B_{\{1, \dots, m\}, \{1, \dots, k\}} \right| > 0.$$

Or, pour $k \in \{m+1, \dots, n\}$ et $\alpha > 0$, on peut démontrer que :

$$(-1)^m \left| \overline{(H + \alpha I_n)}^B_{\{1, \dots, m\}, \{1, \dots, k\}} \right| = \sum_{l=0}^{k-m} \left(\sum_{\substack{\{1, \dots, m\} \subset I \subset \{1, \dots, k+m\} \\ \#I = k+m-l}} (-1)^m \left| \overline{H_k}^{B_{m,k}} \right|_I \right) \alpha^l$$

La preuve de cette égalité est similaire à celle du lemme 8. Nous remettons toutefois les détails à plus tard (voir lemme 16).

Considérons maintenant $k \in \{m+1, \dots, n\}$ et $\alpha > 0$ fixés. D'après l'égalité précédente, il suffit de montrer que pour tout $\{1, \dots, m\} \subset I \subset \{1, \dots, k+m\}$ tel que $2m \leq \#I \leq k+m$,

$$(-1)^m \left| \left(\overline{H_k}^{B_{m,k}} \right)_I \right| \geq 0$$

et que l'un de ces termes, au moins, est strictement positif.

Pour $\{1, \dots, m\} \subset I \subset \{1, \dots, k+m\}$, on pose $J \subset \{1, \dots, k\}$ tel que $I = \{1, \dots, m\} \cup (m+J)$ (où $m+J = \{m+j, j \in J\}$). On a alors

$$\left(\overline{H_k}^{B_{m,k}} \right)_I = \left(\overline{H_J} \right)^{B_{\{1, \dots, m\}, J}}$$

et on est donc ramené à montrer que pour tout $J \subset \{1, \dots, k\}$ tel que $m \leq \#J \leq k$,

$$(-1)^m \left| \left(\overline{H_J} \right)^{B_{\{1, \dots, m\}, J}} \right| \geq 0$$

et que l'un de ces termes, au moins, est strictement positif.

Or, par hypothèse, si $m+1 \leq \#J \leq n$, on a $(-1)^m \left| \left(\overline{H_J} \right)^{B_{\{1, \dots, m\}, J}} \right| \geq 0$.

De plus, dans le cas où $\#J = m$, comme $B_{\{1, \dots, m\}, J}$ est une matrice carrée de taille m :

$$(-1)^m \left| \left(\overline{H_J} \right)^{B_{\{1, \dots, m\}, J}} \right| = (-1)^m \left| \begin{array}{cc} 0 & B_{\{1, \dots, m\}, J} \\ {}^t B_{\{1, \dots, m\}, J} & H_{I^c} \end{array} \right| = \left| \begin{array}{cc} {}^t B_{\{1, \dots, m\}, J} & H_J \\ 0 & B_{\{1, \dots, m\}, J} \end{array} \right| = |B_{\{1, \dots, m\}, J}|^2 \geq 0.$$

On remarquera que l'égalité $\left| \begin{array}{cc} 0 & B_{\{1, \dots, m\}, J} \\ {}^t B_{\{1, \dots, m\}, J} & H_J \end{array} \right| = (-1)^m \left| \begin{array}{cc} {}^t B_{\{1, \dots, m\}, J} & H_J \\ 0 & B_{\{1, \dots, m\}, J} \end{array} \right|$ est simplement obtenue en permutant les m premières lignes avec les m dernières lignes de la matrice.

Enfin, dans le cas particulier où $J = \{1, \dots, m\}$, on a même $(-1)^m \left| \left(\overline{H_J} \right)^{B_{\{1, \dots, m\}, J}} \right| = |B_m|^2 > 0$ par hypothèse.

Au final, l'ensemble des inégalités mentionnées permettent d'assurer que tout $k \in \{m+1, \dots, n\}$, pour tout $\alpha > 0$, $(-1)^m \left| \overline{(H + \alpha I_n)}^B_{\{1, \dots, m\}, \{1, \dots, k\}} \right| > 0$, ce qui conclut la démonstration.

□

On établit de même le résultat suivant.

Proposition 15

Soit q une forme quadratique et soit $\mathcal{B}$ une base de E . On note $H = \text{Mat}_{\mathcal{B}}(q)$. On reprend les notations de la partie précédente et on suppose que $|B_m| \neq 0$. Alors la restriction de q à l'espace $F \simeq \text{Ker}(B)$ est positive si, et seulement si, pour tout $I \subset \{1, \dots, n\}$ tel que $m+1 \leq \#I \leq n$, $(-1)^{\#I} \left| \overline{H_I}^{B(1, \dots, m), I} \right| \geq 0$.

Il reste enfin à démontrer le lemme suivant que nous avons utilisé dans la preuve du critère de Hotelling-Mann généralisé (proposition 14).

Lemme 16

Pour tout $H \in \mathcal{M}_n(\mathbb{R})$, pour tout $B \in \mathcal{M}_{m,n}(\mathbb{R})$ (avec $1 \leq m \leq n-1$) et pour tout $\alpha \in \mathbb{R}$,

$$\left| \overline{(H + \alpha I_n)}^B \right| = \sum_{l=0}^{n-m} \left(\sum_{\substack{\{1, \dots, m\} \subset I \subset \{1, \dots, n+m\} \\ \#I = m+n-l}} \left| \left(\overline{H}^B \right)_I \right| \right) \alpha^l$$

Démonstration. On note $\overline{H}^B = (a_{i,j})_{1 \leq i,j \leq n+m}$. On pose de plus

$$\tilde{S}_{n+m} = \left\{ \sigma \in S_{n+m} \text{ tel que } \sigma(\{1, \dots, m\}) \subset \{m+1, \dots, m+n\} \right\}.$$

Ainsi défini, pour tout $\sigma \notin \tilde{S}_{n+m}$, il existe $i \in \{1, \dots, m\}$ tel que $a_{i, \sigma(i)} = 0$. On en déduit donc que :

$$\left| \overline{(H + \alpha I_n)}^B \right| = \sum_{\sigma \in \tilde{S}_{n+m}} \varepsilon(\sigma) \prod_{i=1}^{n+m} (a_{i, \sigma(i)} + \alpha \delta_{i, \sigma(i)}).$$

On remarque de plus que si $\sigma \in \tilde{S}_{n+m}$, σ admet au plus $n-m$ points fixes. On a donc :

$$\left| \overline{(H + \alpha I_n)}^B \right| = \sum_{j=0}^{n-m} \sum_{\sigma \in \tilde{\Sigma}_j} \varepsilon(\sigma) \left(\prod_{i \notin \text{Fix}(\sigma)} a_{i, \sigma(i)} \right) \left(\prod_{i \in \text{Fix}(\sigma)} a_{i, \sigma(i)} + \alpha \right),$$

où l'on a posé

$$\tilde{\Sigma}_j = \left\{ \sigma \in \tilde{S}_{n+m}, \# \text{Fix}(\sigma) = j \right\}.$$

Le même calcul que celui effectué dans la preuve du lemme 8 permet d'obtenir l'expression suivante :

$$\left| \overline{(H + \alpha I_n)}^B \right| = \sum_{l=0}^{n-m} \left(\sum_{\sigma \in \tilde{\Sigma}_{\geq l}} \sum_{\substack{S \subset \text{Fix}(\sigma) \\ \#S = \# \text{Fix}(\sigma) - l}} \varepsilon(\sigma) \prod_{i \in \text{Fix}(\sigma)^c \cup S} a_{i, \sigma(i)} \right) \alpha^l,$$

où l'on a noté $\tilde{\Sigma}_{\geq l} = \bigcup_{j=l}^{n-m} \tilde{\Sigma}_j$. La fin de la preuve est, là aussi, similaire à celle du lemme 8 : on pose

$$\mathcal{P}_1 = \left\{ (\sigma, \text{Fix}(\sigma)^c \cup S) \text{ tel que } \sigma \in \tilde{\Sigma}_{\geq l}, S \subset \text{Fix}(\sigma) \text{ et } \#S = \# \text{Fix}(\sigma) - l \right\}$$

et

$$\mathcal{P}_2 = \left\{ (\sigma, I) \text{ tel que } \sigma \in \tilde{S}_{n+m}, \{1, \dots, m\} \subset I \subset \{1, \dots, n+m\}, \#I = n+m-l \text{ et } \sigma|_{I^c} = \text{Id} \right\},$$

et on a alors $\mathcal{P}_1 = \mathcal{P}_2$.

Il en résulte que

$$\begin{aligned}
 \left| \overline{(H + \alpha I_n)^B} \right| &= \sum_{l=0}^{n-m} \left(\sum_{(\sigma, l) \in \mathcal{P}_1} \varepsilon(\sigma) \prod_{i \in I} a_{i, \sigma(i)} \right) \alpha^l \\
 &= \sum_{l=0}^{n-m} \left(\sum_{(\sigma, l) \in \mathcal{P}_2} \varepsilon(\sigma) \prod_{i \in I} a_{i, \sigma(i)} \right) \alpha^l \\
 &= \sum_{l=0}^{n-m} \left(\sum_{\substack{\{1, \dots, m\} \subset I \subset \{1, \dots, n+m\} \\ \#I = m+n-l}} \sum_{\sigma \in S(I)} \varepsilon(\sigma) \prod_{i \in I} a_{i, \sigma(i)} \right) \alpha^l \\
 &= \sum_{l=0}^{n-m} \left(\sum_{\substack{\{1, \dots, m\} \subset I \subset \{1, \dots, n+m\} \\ \#I = m+n-l}} \left| \left(\overline{H}^B \right)_I \right| \right) \alpha^l
 \end{aligned}$$

□

II. Extrema, convexité et concavité

Dans les parties II. et III., on considère un $\mathbb{R}$ espace vectoriel normé E de dimension finie n , muni d'une base $\mathcal{B} = (e_1, \dots, e_n)$. On considère de plus une fonction $f : U \subset E \rightarrow \mathbb{R}$ où U est un ouvert de E . Dans le cas où f est une fois différentiable, on notera $\nabla f(x) = {}^t \text{Mat}_{\mathcal{B}}(d_x f) \in \mathcal{M}_{n,1}(\mathbb{R})$ le gradient de f au point $x \in U$ et on adoptera la notation $\partial_i f(x)$ pour désigner la i^{e} dérivée partielle de f au point x . Aussi, dans le cas où f est deux fois différentiable, on notera $Hf(x) = \text{Mat}(d_x^2 f) \in \mathcal{M}_n(\mathbb{R})$ la matrice Hessienne de f au point $x \in U$. Rappelons que si f est deux fois différentiable, le théorème de Schwarz assure que $Hf(x)$ est symétrique pour tout $x \in U$.

Remarque. Les définitions et propriétés générales peuvent être énoncées dans le cas de la dimension infinie. Le lien que nous ferons ensuite avec la première partie et l'utilisation de matrices nécessite en revanche de se placer sur un espace de dimension finie.

On rappelle enfin le théorème des fonctions implicites que nous utiliserons à plusieurs reprises dans cette partie et dans les suivantes.

Théorème 17 – Théorème des fonctions implicites

Soit E_1, E_2 et E_3 trois $\mathbb{R}$ espaces vectoriels de dimension finie, $U \subset E_1 \times E_2$ un ouvert et $f : U \rightarrow E_3$ une fonction de classe $\mathcal{C}^1$. Soit $x^* = (x_1^*, x_2^*) \in U$ tel que $f(x^*) = 0$ et tel que $d_{2,x^*} f$ (la différentielle partielle de f par rapport à x_2 en x^*) soit un isomorphisme. Alors il existe un voisinage ouvert $V \subset E_1$ de x_1^* , un voisinage ouvert $W \subset U$ de x^* et une fonction $\varphi : V \rightarrow E_2$ tels que, pour tout $(x_1, x_2) \in U$:

$$(x_1, x_2) \in W \text{ et } f(x_1, x_2) = 0 \iff x_1 \in V \text{ et } x_2 = \varphi(x_1).$$

De plus, pour tout $x_1 \in V$,

$$d_{x_1} \varphi = - (d_{2, (x_1, \varphi(x_1))} f)^{-1} \circ d_{1, (x_1, \varphi(x_1))} f.$$

1. Extrema

Définition 4 – Extremum global

Soit $D \subset E$ un ouvert et soit $f : D \rightarrow \mathbb{R}$ une fonction. Soit $x_0 \in D$.

- On dit que f admet un maximum en x_0 si pour tout $x \in D$, $f(x) \leq f(x_0)$;
- On dit que f admet un minimum en x_0 si pour tout $x \in D$, $f(x) \geq f(x_0)$;

Remarque. Par abus de langage, et lorsqu'il n'y a pas d'ambiguïté, on dira même que x_0 est un maximum (respectivement minimum) de f .

Définition 5 – Extremum local

Soit $D \subset E$ un ouvert et soit $f : D \rightarrow \mathbb{R}$ une fonction. Soit $x_0 \in D$. On dit que f admet un maximum (resp. minimum) local en x_0 s'il existe un voisinage ouvert U de x_0 tel que $f|_{D \cap U}$ admet un maximum (resp. minimum) en x_0 .

Proposition 18

Soit $f : U \rightarrow \mathbb{R}$ une fonction définie sur un ouvert $U \subset E$ et soit $x_0 \in U$.

- Si f est différentiable en x_0 et admet un minimum local en x_0 , alors $d_{x_0} f \equiv 0$.
- Si f est deux fois différentiable en x_0 et admet un minimum local en x_0 , alors pour tout $h \in E$, $d_{x_0}^2 f(h, h) \geq 0$.
- Réciproquement, si f est deux fois différentiable en x_0 , si $d_{x_0} f \equiv 0$ et si pour tout $h \in E$, $d_{x_0}^2 f(h, h) > 0$, alors f admet un minimum local en x_0 .

Remarque.

- La condition $d_{x_0}^2 f(h, h) \geq 0$ (resp. $d_{x_0}^2 f(h, h) > 0$) signifie exactement que la forme quadratique $h \in E \rightarrow d_{x_0}^2 f(h, h) \in \mathbb{R}$ est positive (resp. définie positive). On peut donc traduire cette condition par une condition portant sur les mineurs $Hf(x)$.
- La même proposition peut être énoncée pour un maximum local en changeant la condition $d_{x_0}^2 f(h, h) \geq 0$ par $d_{x_0}^2 f(h, h) \leq 0$.
- Attention, la proposition « si $d_{x_0} f \equiv 0$ et si pour tout $h \in E$, $d_{x_0}^2 f(h, h) \geq 0$, alors f admet un minimum local en x_0 » est fausse. On pourra, pour s'en convaincre, considérer la fonction $f : x \in \mathbb{R} \rightarrow x^3$ en $x_0 = 0$.

Démonstration. [BENZONI-GAVAGE, 2021]

Dans le cas où $E = \mathbb{R}$, il s'agit d'un résultat classique que l'on peut démontrer en considérant les développements limités de f à l'ordre 1 et 2.

Dans le cas général où E est de dimension n , on considère une boule $B(x_0, r) \subset U$ et on fixe $h \in B(0, r)$. On applique alors le résultat de la dimension 1 à la fonction $g : t \in [0, 1] \rightarrow f(x_0 + th)$ pour laquelle $g'(0) = d_{x_0} f(h)$ et $g''(0) = d_{x_0}^2 f(h, h)$. $\square$

On rappelle maintenant la définition d'ensemble convexe (voir définition 6) afin d'énoncer, dans le cas où une fonction f est définie sur un convexe, une condition nécessaire pour qu'un point x_0 soit un minimum de f .

Définition 6 – Ensemble convexe

Soit $C \subset E$ un ensemble. On dit que C est convexe si, pour tous $x, y \in C$, pour tout $t \in [0, 1]$, $x + t(y - x) \in C$.

Proposition 19

Soit $f : U \rightarrow \mathbb{R}$ une fonction différentiable et soit $C \subset U$ un ensemble convexe. Si $x_0 \in C$ est un minimum de $f|_C$, alors pour tout $x \in C$, $d_{x_0} f(x - x_0) \geq 0$.

Remarque. En identifiant E avec $\mathbb{R}^n$, la condition $d_{x_0} f(x - x_0) \geq 0$ s'écrit $\nabla f(x_0) \cdot (x - x_0) \geq 0$. Cela signifie que l'ensemble C est inclus dans le demi-espace délimité par $x_0 + \nabla f(x_0)^\perp$ et vers lequel pointe $\nabla f(x_0)$. On remarquera par ailleurs que cette proposition n'apporte une véritable information que dans le cas où $\nabla f(x_0) \neq 0$.

Démonstration. [BENZONI-GAVAGE, 2021]

Soit $x \in C$. On a :

$$f(x_0 + t(x - x_0)) = f(x_0) + t d_{x_0} f(x - x_0) + t\tau(t)$$

où τ est une fonction à valeur dans $\mathbb{R}$ qui tend vers 0 en 0.

Si on avait $d_{x_0} f(x - x_0) < 0$, on aurait $f(x_0 + t(x - x_0)) < f(x_0)$ pour $t > 0$ suffisamment petit et x_0 ne serait donc pas un minimum de $f|_C$. $\square$

2. Convexité et concavité : généralités

Définition 7 – Fonction convexe

Une fonction $f : C \rightarrow \mathbb{R}$ (où $C \subset E$ est un ensemble convexe), est dite convexe si pour tous $x, y \in C$, pour tout $t \in [0; 1]$:

$$f(tx + (1 - t)y) \leq tf(x) + (1 - t)f(y).$$

De plus, si l'inégalité ci-dessus est stricte pour tout $t \in]0; 1[$, on dit que f est strictement convexe.

Définition 8 – Fonction concave

Une fonction $f : C \rightarrow \mathbb{R}$, où C est un ensemble convexe, est dite concave si $-f$ est convexe, c'est-à-dire si pour tous $x, y \in C$, pour tout $t \in [0; 1]$:

$$f(tx + (1 - t)y) \geq tf(x) + (1 - t)f(y).$$

De plus, si l'inégalité ci-dessus est stricte pour tout $t \in]0; 1[$, on dit que f est strictement concave.

Définition 9 – Épigraphe et hypographe

Soit A un ensemble et soit $f : A \rightarrow \mathbb{R}$ une fonction.

- On appelle **épigraphe** de f l'ensemble

$$\text{epi}(f) = \{(x, z) \in A \times \mathbb{R}, f(x) \leq z\};$$

- On appelle **hypographe** de f l'ensemble

$$\text{hyp}(f) = \{(x, z) \in A \times \mathbb{R}, f(x) \geq z\}.$$

On démontre sans difficulté la proposition suivante :

Proposition 20

Soit $f : C \rightarrow \mathbb{R}$ une fonction (où $C \subset E$ est un ensemble convexe). Alors :

- f est convexe si, et seulement si, $\text{epi}(f)$ est convexe.
- f est concave si, et seulement si, $\text{hyp}(f)$ est convexe.

Remarque. On en déduit également que si $f : C \rightarrow \mathbb{R}$ est convexe, alors pour tout $z \in \mathbb{R}$, $\{x \in C, f(x) \leq z\}$ est convexe. La réciproque est néanmoins fautive et nous verrons d'ailleurs par la suite que les fonctions pour lesquelles les ensembles $\{x \in C, f(x) \leq z\}$ sont convexes sont appelées *quasi-convexes*.

En définissant la notion d'ensemble strictement convexe (voir ci-dessous), on peut également caractériser les fonctions strictement convexes (resp. strictement concaves) à l'aide de l'épigraphe (resp. l'hypographe).

Définition 10 – Ensemble strictement convexe

Soit $C \subset E$ un ensemble. On dit que C est strictement convexe si, pour tous $x, y \in C$, pour tout $t \in]0; 1[$, $x + t(y - x) \in \overset{\circ}{C}$ où $\overset{\circ}{C}$ désigne l'intérieur de C .

Proposition 21

Soit $f : C \rightarrow \mathbb{R}$ une fonction (où $C \subset E$ est un ensemble convexe). Alors :

- f est strictement convexe si, et seulement si, $\text{epi}(f)$ est strictement convexe.
- f est strictement concave si, et seulement si, $\text{hyp}(f)$ est strictement convexe.

Démonstration. La preuve découle directement du fait que la frontière de l'épigraphe et de l'hypographe est $\{(x, z) \in C \times \mathbb{R}, f(x) = z\}$. □

3. Cas des fonctions définies sur $\mathbb{R}$

Proposition 22 – Lemme des trois cordes

Soit $I \subset \mathbb{R}$ un intervalle et $f : I \rightarrow \mathbb{R}$ une fonction. f est convexe si, et seulement si, pour tous $x, y, z \in I$ avec $x < y < z$,

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

Remarque. Il suffit même que l'une des deux inégalités soit vérifiée pour que f soit convexe.

Démonstration. Les deux inégalités sont en fait équivalentes. Plus précisément, elles sont toutes les deux équivalentes à l'inégalité suivante :

$$f(y) \leq \frac{z - y}{z - x} f(x) + \frac{y - x}{z - x} f(z). \quad (1)$$

Par ailleurs, en posant $t = \frac{z - y}{z - x}$, on a $1 - t = \frac{y - x}{z - x}$ et $tx + (1 - t)z = y$. On en déduit ainsi que l'inégalité 1 est vérifiée pour tous $x < y < z$ si, et seulement si, pour tous $x < z$ et $t \in [0; 1]$,

$$f(tx + (1 - t)z) \leq tf(x) + (1 - t)f(z),$$

c'est-à-dire si, et seulement si, f est convexe. □

On déduit alors les résultats suivants du lemme des trois cordes.

Corollaire 23

Soit $f : I \subset \mathbb{R} \rightarrow \mathbb{R}$ une fonction convexe. Alors f admet une dérivée à gauche et une dérivée à droite en tout point intérieur à I . En particulier, f est continue en tout point intérieur à I .

Corollaire 24

- Soit $f : I \subset \mathbb{R} \rightarrow \mathbb{R}$ dérivable sur I . Alors f est convexe si, et seulement si, f' est croissante sur I .
- Soit $f : I \subset \mathbb{R} \rightarrow \mathbb{R}$ deux fois dérivable sur I . Alors f est convexe si, et seulement si, f'' est positive sur I .

Démonstration. Commençons par montrer le premier point. Si f est convexe, on montre que f' est croissante par simple passage à la limite dans le lemme des trois cordes. Réciproquement, supposons par l'absurde que f' est croissante mais que f n'est pas convexe. Il existerait donc $x < y < z$ tels que

$$\frac{f(y) - f(x)}{y - x} > \frac{f(z) - f(x)}{z - x}.$$

En appliquant le théorème des accroissements finis sur $]x, y[$ et sur $]y, z[$, il existerait donc $u \in]x, y[$ et $v \in]y, z[$ tels que $f'(u) = \frac{f(y) - f(x)}{y - x}$ et $f'(v) = \frac{f(z) - f(x)}{z - x}$. Le fait que $u < v$ mais que $f'(u) > f'(v)$ est contradictoire avec la croissance de f' .

Le deuxième point est ensuite évident car f' est croissante si, et seulement si, f'' est positive. □

Avec le lemme des trois cordes, on peut également démontrer le théorème suivant que l'on établira par ailleurs en toute dimension dans la partie suivante.

Proposition 25

Soit $f : I \subset \mathbb{R} \rightarrow \mathbb{R}$ dérivable sur I . Alors f est convexe si, et seulement si, pour tous $x, y \in I$,

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

Remarque. Il est possible d'énoncer le lemme des trois cordes pour une fonction strictement convexe en remplaçant les inégalités larges par des inégalités strictes. On en déduit alors qu'une fonction dérivable sur un intervalle $I \subset \mathbb{R}$ est strictement convexe si, et seulement si, f' est strictement croissante. De plus, en utilisant le fait qu'une fonction définie sur un intervalle I est strictement croissante si, et seulement si, sa dérivée est positive sur I et strictement positive sur le complémentaire d'un ensemble dont l'intérieur est vide, on en déduit le résultat suivant.

Proposition 26

Soit $f : I \subset \mathbb{R} \rightarrow \mathbb{R}$ deux fois dérivable sur I . Alors f est strictement convexe si, et seulement si, il existe $J \subset I$ avec $\overset{\circ}{J} = \emptyset$, tel que pour tout $x \in I \setminus J$, $f''(x) > 0$ et tel que pour tout $x \in J$, $f''(x) \geq 0$.

Remarque. Pour chaque proposition de cette partie (propositions 22 à 26), il est possible d'énoncer des résultats analogues pour des fonctions concaves. Il suffit pour cela de changer le sens des inégalités. Les preuves de ces résultats se font simplement en appliquant à la fonction $-f$ les résultats relatifs à la convexité.

4. Convexité et différentiabilité**Théorème 27 – Condition d'ordre un**

Soit $U \subset \mathbb{E}$ un ouvert et soit $f : U \rightarrow \mathbb{R}$ une fonction différentiable. Soit $C \subset U$ un ensemble convexe. Alors $f|_C$ est convexe si, et seulement si pour tous $x, y \in C$:

$$f(y) \geq f(x) + d_x f(y - x).$$

De plus, f est strictement convexe si, et seulement si, pour tous $x, y \in C$ tels que $x \neq y$:

$$f(y) > f(x) + d_x f(y - x).$$

Remarque. Ce théorème traduit le fait que le graphe d'une fonction convexe est situé en dessous de ses hyperplans tangents.

Démonstration. [BENZONI-GAVAGE, 2021]

Supposons que $f|_C$ est convexe. Alors, par définition de la convexité de f , pour tous $x, y \in C$ et pour tout $t \in]0; 1[$:

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

En faisant tendre t vers 0, on obtient : $d_x f(y - x) \leq f(y) - f(x)$, ce qui est l'inégalité annoncée.

Réciproquement, supposons que pour tous $x, y \in C$ $f(y) \geq f(x) + d_x f(y - x)$. Montrons qu'alors f est convexe. On peut en fait écrire les deux inégalités suivantes pour $x, y \in C$ et $t \in [0; 1]$:

$$f(x) \geq f(x + t(y - x)) + d_{x+t(y-x)} f(-t(y - x))$$

et

$$f(y) \geq f(x + t(y - x)) + d_{x+t(y-x)} f((1 - t)(y - x)).$$

En prenant la combinaison convexe de ces deux inégalités, on obtient :

$$(1 - t)f(x) + tf(y) \geq f(x + t(y - x)),$$

ce qui signifie précisément que f est convexe.

Supposons enfin que f est strictement convexe. Pour démontrer l'inégalité stricte annoncée, on observe que, pour $0 < t < \lambda < 1$:

$$x + t(y - x) = \left(1 - \frac{t}{\lambda}\right)x + \frac{t}{\lambda}(x + \lambda(y - x)).$$

Ainsi, par stricte convexité,

$$\begin{aligned} \frac{f(x + t(y - x)) - f(x)}{t} &< \frac{\left(1 - \frac{t}{\lambda}\right)f(x) + \frac{t}{\lambda}f(x + \lambda(y - x)) - f(x)}{t} \\ &< \frac{f(x + \lambda(y - x)) - f(x)}{\lambda} < f(y) - f(x) \end{aligned}$$

On obtient ensuite le résultat en gardant λ fixé et en faisant tendre t vers 0.

La réciproque consistant à démontrer que l'inégalité stricte implique la stricte convexité se démontre de la même façon. Il suffit ici de remplacer les inégalités larges de la preuve précédente par des inégalités strictes. $\square$

Théorème 28 – Condition d'ordre deux

Soit $U \subset E$ un ouvert et soit $f : U \rightarrow \mathbb{R}$ une fonction deux fois différentiable. Soit $C \subset U$ un ensemble convexe. Alors $f|_C$ est convexe si, et seulement si pour tous $x, y \in C$:

$$d_x^2 f(y - x, y - x) \geq 0.$$

De plus, si, pour tous $x, y \in C$ tels que $x \neq y$:

$$d_x^2 f(y - x, y - x) > 0,$$

alors f est strictement convexe.

Remarque.

- Attention, la réciproque de la dernière proposition est fautive : il se peut que f soit strictement convexe et qu'il existe $x \neq y$ tels que $d_x^2 f(y - x, y - x) = 0$. Pour s'en convaincre, il suffit de considérer la fonction $x \in \mathbb{R} \mapsto x^4$ dont la différentielle d'ordre deux est nulle en 0.
- Naturellement, on peut énoncer des résultats similaires pour caractériser les fonctions concaves en changeant le sens des inégalités.

Démonstration. [BENZONI-GAVAGE, 2021]

Supposons que f est deux fois différentiable et convexe. Soit $x \in C$ fixé. On pose $g : y \in C \mapsto f(y) - d_x f(y)$. Comme $g - f$ est une fonction linéaire, il est clair que g est convexe. De plus, d'après le théorème 27, x est un minimum global de $g|_C$. Cela implique nécessairement que $d_x g$ est identiquement nulle.

On pose alors $\Phi : t \in [0; 1] \mapsto g(x + t(y - x))$. On a $\Phi(0) = g(x)$, $\Phi'(0) = 0$ et $\Phi''(0) = d_x^2 f(y - x, y - x)$. Avec une formule de Taylor-Young à l'ordre 2 :

$$\begin{aligned} \Phi(t) &= \Phi(0) + t\Phi'(0) + \frac{t^2}{2}\Phi''(0) + o(t^2) \\ &= g(x) + \frac{t^2}{2}d_x^2 f(y - x, y - x) + o(t^2) \end{aligned}$$

En supposant par l'absurde que $d_x^2 f(y - x, y - x) < 0$, il existerait $\varepsilon > 0$ tel que pour tout $0 < t < \varepsilon$, $\Phi(t) < g(x)$. Cela contredirait le fait que x est un minimum de $g|_C$.

Réciproquement, on suppose que pour tous $x, y \in C$, $d_x^2 f(y - x, y - x) \geq 0$.

On considère $x, y \in C$ fixés et on applique la formule des accroissements finis à la fonction

$$\Psi : \begin{cases} [0; 1] & \longrightarrow \mathbb{R} \\ t & \longmapsto f(x + t(y - x)) + (1 - t)d_{x+t(y-x)} f(y - x) \end{cases}$$

Il existe $t_0 \in]0; 1[$ tel que $\Psi(1) - \Psi(0) = \Psi'(t_0)$. Cela signifie que :

$$\begin{aligned} f(y) - f(x) - d_x f(y - x) &= (1 - t_0)d_{x+t_0(y-x)}^2 f(y - x, y - x) \\ &= \frac{1}{1 - t_0}d_{x+t_0(y-x)}^2 f(y - (x + t_0(y - x)), y - (x + t_0(y - x))) \geq 0 \end{aligned}$$

On en conclut que f est convexe en appliquant le théorème 27.

La même preuve dans le cas où l'inégalité est stricte permet de montrer alors que f est strictement convexe. $\square$

On peut alors traduire le théorème 32 à l'aide de la matrice Hessienne de f de la façon suivante³.

Proposition 29

Soit $f : U \rightarrow \mathbb{R}$ définie sur un ouvert convexe U et supposée deux fois différentiable. Ainsi,

- f est convexe si, et seulement si, pour tout $x \in U$, sa matrice Hessienne, notée $Hf(x)$ est (semi-définie) positive.
- S'il existe un ensemble discret $V \subset U$ (cela signifie que V ne contient pas de point d'accumulation) tel que pour tout $x \in U \setminus V$, $Hf(x)$ est définie positive et tel que pour tout $x \in V$, $Hf(x)$ est (semi-définie) positive, alors f est strictement convexe.

3. La matrice Hessienne $Hf(x)$ est définie comme la matrice de la forme bilinéaire $d_x^2 f$ dans la base $\mathcal{B}$.

Démonstration. Le deuxième point se prouve à l'aide de la proposition 26. En effet, pour montrer que f est strictement convexe, il suffit de montrer que pour tous $x, y \in U$ avec $x \neq y$, la fonction $g : t \mapsto f(x + t(y - x))$ est strictement convexe sur $[0, 1]$. Or, pour tout $t \in [0, 1]$, $g''(t) = d_{x+t(y-x)}^2 f(y - x, y - x)$. Le fait que la matrice Hessienne est définie positive sauf éventuellement sur un ensemble discret où elle est seulement positive a pour conséquence que $g'(t) > 0$ sauf éventuellement sur un ensemble d'intérieur vide (où $g'(t)$ peut être nul). La proposition 26 permet donc de conclure que g est bien strictement convexe. $\square$

Remarque. On sait, d'après la première partie, que le caractère positif ou défini positif de la matrice Hessienne de f (cette matrice est symétrique lorsque f est deux fois différentiable) est donné par le signe des mineurs.

5. Convexité et extrema

Proposition 30

Soit $f : U \rightarrow \mathbb{R}$ une fonction définie sur un ouvert $U \subset E$ et soit $C \subset U$ un ensemble convexe.

- Si $f|_C$ est convexe et admet un minimum local dans C , alors c'est un minimum global;
- Si $f|_C$ est strictement convexe, alors elle admet au plus un minimum dans C et c'est un minimum strict;
- Si f est différentiable, si $f|_C$ est convexe et s'il existe $x_0 \in C$ tel que $d_{x_0} f(x - x_0) \geq 0$ pour tout $x \in C$, alors x_0 est un minimum de $f|_C$.

Remarque. Le dernier point est en fait la réciproque de la proposition 19 dans le cas où la fonction est convexe.

Démonstration. [BENZONI-GAVAGE, 2021]

- Supposons que f admette un minimum local en x_0 . Pour tout $x \in C$, pour tout $t \in [0, 1]$,

$$f(x_0 + t(x - x_0)) - f(x_0) \leq t(f(x) - f(x_0)).$$

Le membre de gauche est positif pour t suffisamment petit. On en déduit donc que $f(x) - f(x_0) \geq 0$ pour tout $x \in C$, ce qui signifie exactement que $f(x_0)$ est un minimum global.

- La démonstration est identique au point 1 en remplaçant l'inégalité large par une inégalité stricte.
- Le troisième point découle immédiatement du théorème 27.

$\square$

6. Résultats équivalents pour la concavité

En appliquant les théorèmes concernant la convexité à la fonction $-f$, on obtient les résultats suivants.

Théorème 31 – Condition d'ordre un

Soit $U \subset E$ un ouvert et soit $f : U \rightarrow \mathbb{R}$ une fonction différentiable. Soit $C \subset U$ un ensemble convexe. Alors $f|_C$ est concave si, et seulement si pour tous $x, y \in C$:

$$f(y) \leq f(x) + d_x f(y - x).$$

De plus, f est strictement concave si, et seulement si, pour tous $x, y \in C$ tels que $x \neq y$:

$$f(y) < f(x) + d_x f(y - x).$$

Théorème 32 – Condition d'ordre deux

Soit $U \subset E$ un ouvert et soit $f : U \rightarrow \mathbb{R}$ une fonction deux fois différentiable. Soit $C \subset U$ un ensemble convexe. Alors $f|_C$ est concave si, et seulement si pour tous $x, y \in C$:

$$d_x^2 f(y - x, y - x) \leq 0.$$

De plus, si, pour tous $x, y \in C$ tels que $x \neq y$:

$$d_x^2 f(y - x, y - x) < 0,$$

alors f est strictement concave.

Proposition 33

Soit $f : U \rightarrow \mathbb{R}$ définie sur un ouvert convexe $U \subset E$ et supposée deux fois différentiable. Ainsi,

- f est concave si, et seulement si, pour tout $x \in U$, sa matrice Hessienne, notée $Hf(x)$ est (semi-définie) négative.
- S'il existe un ensemble discret $V \subset U$ tel que pour tout $x \in U \setminus V$, $Hf(x)$ est définie négative et tel que pour tout $x \in V$, $Hf(x)$ est (semi-définie) négative, alors f est strictement concave.

Proposition 34

Soit $f : U \rightarrow \mathbb{R}$ une fonction définie sur un ouvert $U \subset E$ et soit $C \subset U$ un ensemble convexe. Alors :

- Si $f|_C$ est concave et admet un maximum local dans C , alors c'est un maximum global ;
- Si $f|_C$ est strictement concave, alors elle admet au plus un maximum dans C et c'est un maximum strict ;
- Si f est différentiable, si $f|_C$ est convexe et s'il existe $x_0 \in C$ tel que $d_{x_0} f(x - x_0) \leq 0$ pour tout $x \in C$, alors x_0 est un maximum de $f|_C$.

7. Exemples

Exemple 2. Considérons la fonction $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ définie par :

$$f(x_1, x_2) = e^{x_1} + e^{x_2} + e^{x_1 + x_2}$$

Un simple calcul permet de montrer que la matrice hessienne de f au point $x \in \mathbb{R}^2$ est :

$$Hf(x) = \begin{pmatrix} e^{x_1} + e^{x_1 + x_2} & e^{x_1 + x_2} \\ e^{x_1 + x_2} & e^{x_2} + e^{x_1 + x_2} \end{pmatrix}$$

Les mineurs principaux dominants de $Hf(x)$ sont :

$$|Hf(x)_1| = e^{x_1} + e^{x_1 + x_2} > 0, \quad |Hf(x)_1| = |Hf(x)| = e^{x_1 + x_2} + e^{x_1 + 2x_2} + e^{2x_1 + x_2} > 0$$

Par conséquent, d'après le critère de Sylvester, la matrice Hf est définie positive, ce qui prouve, d'après la proposition 29 que f est strictement convexe. A fortiori, f est également convexe.

Exemple 3. Considérons la fonction $g : \mathbb{R}^2 \rightarrow \mathbb{R}$ définie par :

$$g(x_1, x_2) = (x_1 + 2x_2 - 3)^2$$

La matrice Hessienne de g est donnée par :

$$\forall x \in \mathbb{R}^2, \quad Hg(x) = \begin{pmatrix} 2 & 4 \\ 4 & 8 \end{pmatrix}$$

Pour tout $x \in \mathbb{R}^2$, les mineurs principaux de $Hg(x)$ sont :

$$|Hg(x)_1| = 2 \geq 0, \quad :|Hg(x)_1| = 8 \geq 0, \quad :|Hg(x)_{1,2}| = |Hg(x)| = 2 \cdot 8 - 4 \cdot 4 = 0 \geq 0.$$

Par conséquent, pour tout $x \in \mathbb{R}^2$, $Hg(x)$ est positive (d'après la proposition 3) et la fonction g est donc convexe sur $\mathbb{R}^2$ (d'après la proposition 29).

En revanche, la fonction g n'est pas strictement convexe. Pour le montrer, considérons les points $x = (x_1, x_2) = (3, 0)$ et $y = (y_1, y_2) = (1, 1)$. On a $g(x) = g(y) = 0$. De plus, pour tout $t \in [0; 1]$,

$$g(x + t(y - x)) = g(3 - 2t, t) = 0,$$

ce qui prouve effectivement que g n'est pas strictement convexe.

Remarque. Attention, le fait que le déterminant de la matrice Hessienne soit nulle n'est pas suffisant pour prouver qu'une fonction n'est pas strictement convexe. C'est par exemple ce que montre l'exemple suivant.

Exemple 4. Considérons la fonction $h : \mathbb{R}^2 \longrightarrow \mathbb{R}$ définie par

$$h(x_1, x_2) = x_1^4 + x_2^4.$$

La matrice Hessienne de h est donnée par

$$\forall x \in \mathbb{R}^2, \quad Hh(x) = \begin{pmatrix} 12x_1^2 & 0 \\ 0 & 12x_2^2 \end{pmatrix}.$$

Ainsi, la matrice Hessienne est positive pour tout $(x_1, x_2) \in \mathbb{R}^2$ mais pour $(x_1, x_2) \in T = \{(x_1, x_2) \in \mathbb{R}^2; x_1 = 0 \text{ ou } x_2 = 0\}$, elle est seulement (semi-définie) positive. Comme l'ensemble T n'est pas discret, on ne peut pas en conclure, avec la proposition 30, que h est strictement convexe.

Pourtant, on va montrer que h est bien strictement convexe. Cela résulte du fait que la fonction $x \in \mathbb{R} \rightarrow x^4$ est strictement convexe d'après la proposition 26.

Ainsi, en considérant $x = (x_1, x_2) \in \mathbb{R}^2$, $y = (y_1, y_2) \in \mathbb{R}^2$ et $t \in]0; 1[$, on a :

$$h((1-t)x + ty) = ((1-t)x_1 + ty_1)^4 + ((1-t)x_2 + ty_2)^4 < (1-t)(x_1^4 + x_2^4) + t(y_1^4 + y_2^4) = (1-t)h(x) + th(y).$$

Cela prouve bien que h est strictement convexe.

En fait, on voit à l'aide de cet exemple que la condition suffisante donnée par la proposition 30 n'est pas une condition nécessaire. Ce qui est réellement important est plutôt que, pour tout segment, lorsque la matrice Hessienne n'est pas définie positive en certains points, le noyau de la Hessienne ne soit pas dans la même direction que le segment considéré. On pourra d'ailleurs relire la preuve du théorème 32 et de la proposition 30 pour s'en convaincre.

Ici, dans l'exemple de la fonction h , on peut distinguer quatre cas :

- Si le segment $[x, y]$ ne contient aucun point de l'ensemble T , alors pour tout $z \in [x, y]$, la matrice $Hh(z)$ est définie positive.
- Si le segment $[x, y]$ ne contient qu'un seul point de l'ensemble T , alors il n'existe qu'un seul point du segment pour lequel la matrice Hessienne n'est pas définie positive.
- Si le segment $[x, y]$ est inclus dans la droite d'équation $x_1 = 0$ (sa direction est $\text{Vect}(e_2)$), alors pour tout $z \in [x, y]$, le noyau de $Hh(z)$ est $\text{Vect}(e_1)$. On voit ainsi que $d_z^2 h(e_2, e_2) > 0$.
- De même, si le segment $[x, y]$ est inclus dans la droite d'équation $x_2 = 0$ (sa direction est $\text{Vect}(e_1)$), alors pour tout $z \in [x, y]$, le noyau de $Hh(z)$ est $\text{Vect}(e_2)$. On voit ainsi que $d_z^2 h(e_1, e_1) > 0$.

On donne enfin un exemple de fonction qui n'est pas convexe. Sachant que la proposition 29 donne une CNS sur la matrice Hessienne pour qu'une fonction soit convexe, il est toujours possible de démontrer qu'une fonction n'est pas convexe en déterminant le signe des mineurs.

Exemple 5. Considérons la fonction $l : U \subset \mathbb{R}^2 \longrightarrow \mathbb{R}$ définie sur $U = \mathbb{R} \times (\mathbb{R}_*^+)$ par

$$l(x_1, x_2) = \frac{x_1}{x_2}.$$

La matrice Hessienne de l est donnée par :

$$\forall x \in U, \quad Hl(x) = \begin{pmatrix} 0 & -\frac{1}{x_2^2} \\ -\frac{1}{x_2^2} & \frac{2x_1}{x_2^3} \end{pmatrix}.$$

Ainsi, pour tout $x \in U$, $|Hl(x)| = -\frac{1}{x_2^4} < 0$. Cela implique que $Hl(x)$ n'est ni positive, ni négative et la fonction l n'est donc, ni convexe, ni concave sur U . On voit même que pour tout convexe $C \subset U$, f est ni convexe, ni concave sur C .

III. Quasi-convexité et quasi-concavité

1. Généralités

Définition 11 – Fonction quasi-convexe

Une fonction $f : C \rightarrow \mathbb{R}$, où C est un ensemble convexe, est dite quasi-convexe si pour tous $x, y \in C$, pour tout $t \in [0; 1]$:

$$f(tx + (1-t)y) \leq \max\{f(x), f(y)\}.$$

De plus, si l'inégalité ci-dessus est stricte pour tout $t \in]0; 1[$, on dit que f est strictement quasi-convexe.

Définition 12 – Fonction quasi-concave

Une fonction $f : C \rightarrow \mathbb{R}$, où C est un ensemble convexe, est dite quasi-concave si $-f$ est quasi-convexe, c'est-à-dire si pour tous $x, y \in C$, pour tout $t \in [0; 1]$:

$$f(tx + (1-t)y) \geq \min\{f(x), f(y)\}.$$

De plus, si l'inégalité ci-dessus est stricte pour tout $t \in]0; 1[$, on dit que f est strictement quasi-concave.

On démontre sans difficulté les propositions suivantes :

Proposition 35

Soit $f : C \rightarrow \mathbb{R}$ une fonction (où $C \subset E$ est un ensemble convexe).

- Si f est (strictement) convexe, alors f est (strictement) quasi-convexe.
- Si f est (strictement) concave, alors f est (strictement) quasi-concave.

Proposition 36

Soit $f : C \rightarrow \mathbb{R}$ une fonction (où $C \subset E$ est un ensemble convexe). Alors :

- f est (strictement) quasi-convexe si, et seulement si, pour tout $z \in \mathbb{R}$, $\{x \in C, f(x) \leq z\}$ est (strictement) convexe.
- f est (strictement) quasi-concave si, et seulement si, pour tout $z \in \mathbb{R}$, $\{x \in C, f(x) \geq z\}$ est (strictement) convexe.

Remarque. À ce stade, on peut noter une différence importante entre semi-continuité et convexité. En effet, pour une fonction $f : X \rightarrow \mathbb{R}$ définie sur un espace topologique X , l'épigraphe est fermé si, et seulement si, pour tout $z \in \mathbb{R}$, les ensembles de niveaux $\{x \in X, f(x) \leq z\}$ sont fermés. Dans ce cas, on dit que la fonction f est semi-continue inférieurement. À l'inverse, le fait que l'épigraphe soit convexe n'est pas équivalent au fait que les ensembles de niveaux soient convexes. Cela nécessite la distinction entre les notions de convexité et de quasi-convexité.

Remarque. Il existe certaines situations dans lesquelles on est assuré de la quasi-convexité ou de la quasi-concavité. On peut, entre autres, citer les cas suivants où $C \subset \mathbb{R}^n$ désigne un ensemble convexe :

- Si $f : C \rightarrow \mathbb{R}$ est convexe (concave) et si $g : \mathbb{R} \rightarrow \mathbb{R}$ est croissante, alors $g \circ f$ est quasi-convexe (concave).
- Si $f : C \rightarrow \mathbb{R}$ est convexe (concave) et si $g : C \rightarrow \mathbb{R}$ est affine et strictement positive sur C , alors $\frac{f}{g}$ est quasi-convexe (quasi-concave).

Les démonstrations de ces résultats ne posent pas de difficulté particulière.

On énonce enfin une proposition permettant de mieux cerner la différence entre quasi-convexité et stricte quasi-convexité.

Proposition 37

Soit $f : C \rightarrow \mathbb{R}$ une fonction quasi-convexe (où $C \subset E$ est un ensemble convexe). Les deux assertions suivantes sont équivalentes :

- f est strictement quasi-convexe ;
- f est non constante sur tout segment non réduit à un point.

Remarque. Autrement dit, les fonctions quasi-convexes qui ne sont pas strictement quasi-convexes sont les fonctions pour lesquelles une courbe de niveau contient un segment. Cela peut se produire lorsqu'une courbe de niveau est localement une partie de droite ou lorsqu'une courbe de niveau est d'intérieur non vide.

Démonstration.

- Si f est constante sur un segment non réduit à un point, il est clair que f ne sera pas strictement quasi-convexe.
- Inversement, supposons que f n'est pas strictement quasi-convexe et montrons qu'il segment alors un segment sur lequel f est constante.
Comme f n'est pas strictement quasi-convexe, il existe $x, y \in C$ avec $x_1 \neq y_1$ et $x_0 \in]x_1, y_1[$ tels que $f(x_0) \geq \max\{f(x_1), f(y_1)\}$. Comme f est toutefois supposée quasi-convexe, on a nécessairement

$$f(x_0) = \max\{f(x_1), f(y_1)\}.$$

On peut supposer de plus, sans perte de généralité, que $f(x_1) \leq f(y_1)$.

On va montrer que pour tout $x \in]x_0, y_1[$, $f(x) = f(y_1)$.

D'une part, comme f est quasi-convexe, on a, pour tout $x \in]x_0, y_1[$, $f(x) \leq f(y_1)$.

D'autre part, si $x \in]x_0, y_1[$, on a $x_0 \in]x_1, x[$.

Par quasi-convexité de f , on en déduit que $f(y_1) = f(x_0) \leq f(x)$.

Au final, on a donc $f(x) = f(y_1)$ pour tout $x \in]x_0, y_1[$ et f est bien constante sur un segment non réduit à un point. □

2. Cas des fonctions définies sur $\mathbb{R}$

Proposition 38

Une fonction $f : C \subset \mathbb{R} \rightarrow \mathbb{R}$ (où C est un intervalle) est quasi-convexe si, et seulement si, f est monotone où f est successivement décroissante puis croissante (dans le sens où il existe $x_0 \in C$ tel que f est décroissante sur $] -\infty; x_0] \cap C$ et croissante sur $[x_0; +\infty[\cap C$).

Démonstration. Supposons que f soit quasi-convexe. On pose $I = \{x \in C \mid \exists y \in C, y > x \text{ et } f(y) < f(x)\}$ et $J = C \setminus I = \{x \in C \mid \forall y \in C, y > x \implies f(y) \geq f(x)\}$. Par définition, f est croissante sur J .

Montrons maintenant que f est également décroissante sur I . Par définition de I , si $x_1, x_2 \in I$ avec $x_1 < x_2$, il existe $y \in C$ tel que $y > x_2$ et tel que $f(y) < f(x_2)$. Or, f étant quasi-convexe, et sachant que $x_2 \in]x_1; y[$, on a : $f(x_2) \leq \max\{f(y); f(x_1)\}$. Comme $f(y) < f(x_2)$, cela implique que $f(x_2) \leq f(x_1)$, achevant ainsi la preuve du fait que f est décroissante sur I .

On montre enfin, par des arguments similaires, que I et J sont convexes, et que, pour tout $x \in I$ et $y \in J$, $x \leq y$. En définitive :

- Si $I = \emptyset$, alors f est croissante sur C ;
- Si $J = \emptyset$, alors f est décroissante sur C ;
- Si $I \neq \emptyset$ et $J \neq \emptyset$, alors f est décroissante puis croissante.

Réciproquement, il est immédiat de vérifier que les fonctions monotones ainsi que les fonctions décroissantes puis croissantes sont bien quasi-convexes. □

Remarque.

- On montre de la même manière que $f : C \subset \mathbb{R} \rightarrow \mathbb{R}$ est quasi-concave si, et seulement si, f est monotone ou f est croissante puis décroissante.
- Les fonctions $f : \mathbb{R} \rightarrow \mathbb{R}$ à la fois quasi-convexes et quasi-concaves sont donc les fonctions monotones.

3. Quasi-convexité et différentiabilité

Théorème 39 – Condition d'ordre un – Arrow - Enthoven

Soit $U \subset E$ un ouvert et soit $f : U \rightarrow \mathbb{R}$ une fonction différentiable. Soit $C \subset U$ un ensemble convexe. Alors $f|_C$ est quasi-convexe si, et seulement si pour tous $x, y \in C$:

$$f(x) \geq f(y) \implies d_x f(y - x) \leq 0.$$

En supposant de plus que pour tout $x \in C$, $d_x f$ n'est pas identiquement nulle, $f|_C$ est strictement quasi-convexe si, et seulement si pour tous $x, y \in C$ avec $x \neq y$:

$$f(x) \geq f(y) \implies d_x f(y - x) < 0.$$

Démonstration. [AVRIEL et al., 1988, CAMBINI et MARTEIN, 2008]

- $\implies$ Supposons que f est quasi-convexe. Soient $x, y \in C$ tels que $f(x) \geq f(y)$. Pour tout $t \in]0; 1[$,

$$f(x + t(y - x)) \leq f(x).$$

Par conséquent,

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

En faisant tendre t vers 0, on en déduit que $d_x f(y - x) \leq 0$.

- $\impliedby$ Réciproquement, supposons par l'absurde que l'implication suivante est vraie pour tous $x, y \in C$:

$$f(x) \geq f(y) \implies d_x f(y - x) \leq 0,$$

mais que f n'est pas quasi-convexe. Cela signifie, par définition, qu'il existe $x, y \in C$ et $t_0 \in]0; 1[$ tels que $f(x + t_0(y - x)) > \max\{f(x); f(y)\}$. Quitte à permuter les rôles de x et de y , on peut supposer $f(x) \geq f(y)$. On pose alors $t_1 = \sup \left\{ t \in [0; t_0] ; f(x + t(y - x)) \leq f(x) \right\}$.

Par continuité de f , on a $0 \leq t_1 < t_0$. De plus, $f(x + t_1(y - x)) = f(x)$ et pour tout $t \in]t_1; t_0[$, $f(x + t(y - x)) > f(x) \geq f(y)$. Cette dernière inégalité implique que, par hypothèse, $d_{x+t(y-x)} f(y - (x + t(y - x))) \leq 0$ pour tout $t \in]t_1; t_0[$. Plus simplement, on a donc $d_{x+t(y-x)} f(y - x) \leq 0$ pour tout $t \in]t_1; t_0[$.

Autrement dit, la fonction $g : t \in [0; 1] \mapsto f(x + t(y - x))$ est décroissante sur $[t_1; t_0]$. En effet, g est dérivable et pour tout $t \in]t_1; t_0[$, $g'(t) = d_{x+t(y-x)} f(y - x) \leq 0$.

Au final, la décroissance de g est contradictoire avec le fait que $g(t_1) < g(t_0)$. □

On s'intéresse maintenant aux conditions d'ordre deux pour la quasi-convexité. Rappelons que dans le cas de la convexité, on a établi une CNS sur la différentielle d'ordre deux pour qu'une fonction soit convexe (voir théorème 32). Nous ne pourrions pas en faire de même pour la quasi-convexité. Seul le cas où le gradient de la fonction est non nul en tout point permettra d'énoncer une CNS. Dans le cas général, on se contentera de traiter les conditions nécessaires et les conditions suffisantes de manière séparées. C'est l'objet des deux lemmes qui suivent. Le premier (lemme de Arrow-Enthoven-Avriel) donne une condition nécessaire pour qu'une fonction soit quasi-convexe, quand le second (lemme de Crouzeix) énonce une condition suffisante.

Lemme 40 – Arrow - Enthoven - Avriel

Soit $C \subset E$ un ouvert convexe et soit $f : C \rightarrow \mathbb{R}$ une fonction quasi-convexe de classe $\mathcal{C}^2$. Alors pour tout $x \in C$, pour tout $v \in E$:

$$d_x f(v) = 0 \implies d_x^2 f(v, v) \geq 0.$$

Remarque. • On notera les spécificités des hypothèses du lemme 40 : il faut que le convexe considéré C soit lui-même ouvert (et pas seulement inclus dans un ouvert sur lequel f est définie) et que f ne soit pas seulement deux fois différentiable mais bien de classe $\mathcal{C}^2$. Il en sera de même pour tous les résultats relatifs aux conditions de quasi-convexité d'ordre deux que nous énoncerons.

- La condition « $d_x f(v) = 0 \implies d_x^2 f(v, v) \geq 0$ » n'est pas suffisante pour qu'une fonction soit quasi-convexe (elle l'est uniquement si le gradient ne s'annule pas). On pourra, pour s'en convaincre, considérer le cas de la fonction $f : x \in \mathbb{R} \rightarrow -x^4$. Cette fonction n'est pas quasi-convexe mais elle vérifie bien la condition mentionnée plus haut.

Démonstration. [AVRIEL et al., 1988, CAMBINI et MARTEIN, 2008]

On considère une fonction f de classe $\mathcal{C}^2$ et quasi-convexe sur un ouvert convexe C . Supposons par l'absurde qu'il existe $x_0 \in C$ et $v \in E$ tels que $d_{x_0} f(v) = 0$ et $d_{x_0}^2 f(v, v) < 0$. L'objectif est de trouver x_1 et x_2 tels que $x_0 \in [x_1; x_2]$ mais tels que $f(x_0) > \max\{f(x_1), f(x_2)\}$.

Par continuité de $d^2 f$, il existe $\eta > 0$ tel que pour tout $x \in B(x_0, \eta)$, $d_x^2 f(v, v) < 0$.

On considère t_0 tel que $0 < t_0 \|v\| < \eta$.

Écrivons maintenant une formule de Taylor Lagrange à l'ordre 2 appliquée à la fonction $g : t \mapsto f(x_0 + tv)$ entre 0 et t_0 : il existe $t_1 \in]0; t_0[$ tel que

$$f(x_0 + t_0 v) = f(x_0) + t_0 d_{x_0} f(v) + \frac{t_0^2}{2} d_{x_0 + t_1 v}^2 f(v, v).$$

Comme $d_{x_0} f(v) = 0$ et que $d_{x_0 + t_1 v}^2 f(v, v) < 0$, on en déduit que $f(x_0 + t_0 v) < f(x_0)$.

On montre de la même manière que $f(x_0 - t_0 v) < f(x_0)$. Par conséquent, en posant $x_1 = x_0 - t_0 v \in C$ et $x_2 = x_0 + t_0 v \in C$, on a $x_0 = \frac{x_1 + x_2}{2}$ mais $f(x_0) > \max\{f(x_1), f(x_2)\}$. Cela contredit le caractère quasi-convexe de la fonction f . $\square$

Lemme 41 – Crouzeix

Soit $C \subset E$ un ouvert convexe et soit $f : C \rightarrow \mathbb{R}$ une fonction de classe $\mathcal{C}^2$. On suppose de plus que f vérifie les deux assertions suivantes :

- H1 : pour tout $x \in C$, pour tout $v \in E$: $d_x f(v) = 0 \implies d_x^2 f(v, v) \geq 0$
- H2 : pour tout $x \in C$ tel que $d_x f \equiv 0$, pour tout $v \in E$, $d_x^2 f(v, v) > 0$.

Alors f est quasi-convexe.

Remarque. Comme nous l'avons indiqué précédemment, l'hypothèse H_1 n'est pas suffisante pour qu'une fonction soit quasi-convexe. C'est d'ailleurs pour cela que l'on a ajouté l'hypothèse H_2 dans le lemme de Crouzeix ci-dessus. L'ensemble des hypothèses H_1 et H_2 ne sont en revanche pas nécessaires pour qu'une fonction soit quasi-convexe. Il suffit, pour s'en convaincre, de considérer une fonction constante : elle est naturellement quasi-convexe mais ne vérifie pas l'hypothèse H_2 .

Démonstration. [AVRIEL et al., 1988, CAMBINI et MARTEIN, 2008, CROUZEIX, 1980]

Supposons que f vérifie les hypothèses du lemme. On suppose de plus, par l'absurde, que f n'est pas quasi-convexe. Cela signifie qu'il existe $x_1, y_1 \in C$ tel que $M := \max\{f(x_1 + t(y_1 - x_1)); t \in [0; 1]\} > \max\{f(x_1), f(y_1)\}$. On considère $t_0 = \max\{t \in [0; 1]; f(x_1 + t(y_1 - x_1)) = M\}$ le plus grand t pour lequel le maximum est atteint, et on note $x_0 = x_1 + t_0(y_1 - x_1)$. Par définition de t_0 , on a ainsi :

$$\forall t_0 < t \leq 1, \quad f(x_1 + t(y_1 - x_1)) < f(x_0) = M.$$

Dans la suite, on notera $\nabla f(x) = \sum_{i=1}^n d_x f(e_i) e_i$ le gradient de f au point x exprimé dans la base $(e_1, \dots, e_n)$.

Aussi, par définition de x_0 , la dérivée de la fonction $t \mapsto f(x_0 + t(y_1 - x_1))$ s'annule en 0, ce qui se traduit également par le fait que ${}^t(y_1 - x_1) \nabla f(x_0) = 0$.

On distingue alors deux cas :

- Si $\nabla f(x_0) = 0$, on considère la fonction Φ définie sur un voisinage I de 0 (dans $\mathbb{R}$) par

$$\Phi(t) = f(x_0 + t(y_1 - x_1)).$$

On a $\Phi'(0) = 0$ et $\Phi''(0) = d_{x_0}^2 f(y_1 - x_1) > 0$ par hypothèse.

Ainsi, avec un développement de Taylor-Young appliqué à la fonction Φ en 0, il existe ε définie sur I , telle que $\lim_{h \rightarrow 0} \varepsilon(h) = 0$ et telle que

$$\forall t \in I, \quad \Phi(t) = \Phi(0) + \frac{t^2}{2} \Phi''(0) + t^2 \varepsilon(t).$$

On en déduit qu'il existe $\eta > 0$ tel que pour tout $0 < t < \eta$, $\Phi(t) > \Phi(0)$. Autrement dit, pour tout $0 < t < \eta$, $f(x_0 + t(y_1 - x_1)) > f(x_0)$, ce qui est absurde par définition de x_0 .

- Si $\nabla f(x_0) \neq 0$, on considère la fonction Ψ définie sur un voisinage de 0 (dans $\mathbb{R}^2$) par

$$\Psi(\alpha, \beta) = f(x_0 + \alpha(y_1 - x_1) + \beta \nabla f(x_0)).$$

Comme $\frac{\partial \Psi}{\partial \beta}(0; 0) = \langle f(x_0), \nabla f(x_0) \rangle = \|\nabla f(x_0)\| > 0$, on peut appliquer le théorème des fonction implicites : il existe une fonction $g : I \subset \mathbb{R} \rightarrow \mathbb{R}$ de classe $\mathcal{C}^2$ (car Ψ est évidemment de classe $\mathcal{C}^2$) définie sur un voisinage de 0 telle que $g(0) = 0$ et telle que :

$$\Psi(\alpha, \beta) = 0 \iff \beta = g(\alpha).$$

Ainsi, pour tout $\alpha \in I$,

$$f(x_0 + \alpha(y_1 - x_1) + g(\alpha)\nabla f(x_0)) = f(x_0). \quad (2)$$

L'objectif est maintenant de montrer que, quitte à restreindre l'intervalle I , pour tout $\alpha \in I$, $g'(\alpha) \leq 0$. Pour cela, on va prouver que g est concave (localement, autour de 0) et que $g'(0) = 0$. L'inégalité résultera alors du fait que $g(0) = 0$. Une fois cela établi, la dernière étape de la démonstration consistera ensuite à écrire un développement de Taylor, appliqué à la fonction Ψ . En dérivant (2) (par rapport à α), on obtient :

$${}^t \nabla f(x_0 + \alpha(y_1 - x_1) + g(\alpha)\nabla f(x_0)) \left((y_1 - x_1) + g'(\alpha)\nabla f(x_0) \right) = 0. \quad (3)$$

Pour $\alpha = 0$, on a, plus particulièrement : ${}^t \nabla f(x_0)(y_1 - x_1) + g'(0)\|\nabla f(x_0)\|^2 = 0$. Comme ${}^t \nabla f(x_0)(y_1 - x_1) = 0$ et que $\|\nabla f(x_0)\|^2 > 0$, on en déduit que $g'(0) = 0$.

De plus, en dérivant de nouveau l'égalité (3.), on a :

$${}^t \left(g'(\alpha)\nabla f(x_0) + (y_1 - x_1) \right) d_{z(\alpha)}^2 f \left(g'(\alpha)\nabla f(x_0) + (y_1 - x_1) \right) + g''(\alpha){}^t \nabla f(z(\alpha)) = 0 \quad (4)$$

où $z(\alpha) = x_0 + \alpha(y_1 - x_1) + g(\alpha)\nabla f(x_0)$. Or, on sait d'après que $g'(\alpha)\nabla f(x_0) + (y_1 - x_1) \in \text{Ker}(d_{z(\alpha)}f)$. Cela implique, par hypothèse, que

$${}^t \left(g'(\alpha)\nabla f(x_0) + (y_1 - x_1) \right) d_{z(\alpha)}^2 f \left(g'(\alpha)\nabla f(x_0) + (y_1 - x_1) \right) \geq 0$$

et il en résulte donc l'inégalité suivante :

$$g''(\alpha){}^t \nabla f(z(\alpha)) \leq 0.$$

Comme $\|\nabla f(x_0)\|^2 > 0$, on sait, par continuité du gradient, que $\|\nabla f(z(\alpha))\|^2 > 0$ sur un voisinage de 0. Autrement dit, quitte à restreindre I , $g''(\alpha) \leq 0$ pour tout $\alpha \in I$. Cela signifie que g est concave sur I et on en déduit, comme annoncé, que $g(\alpha) \leq 0$ pour tout $\alpha \in I$.

On remarque par ailleurs que, pour tout $\alpha \in I \cap]0; +\infty[$, $g(\alpha) < 0$. En effet, si l'on avait $g(\alpha) = 0$ pour $\alpha \in I \cap]0; +\infty[$, cela contredirait la définition de t_0 et le fait que

$$\forall t_0 < t \leq 1, \quad f(x_1 + t(y_1 - x_1)) < f(x_0) = M.$$

Finalement, en écrivant un développement de Taylor appliqué à la fonction Ψ :

$$\forall \alpha \in I, \quad f(x_0) = f(x_0 + \alpha(y_1 - x_1) + g(\alpha)\nabla f(x_0)) = f(x_0 + \alpha(y_1 - x_1)) + g(\alpha){}^t \nabla f(x_0 + \alpha(y_1 - x_1))\nabla f(x_0) + o(g(\alpha)).$$

Pour $\alpha > 0$ suffisamment petit, on a vu que $g(\alpha) < 0$. De plus, ${}^t \nabla f(x_0 + \alpha(y_1 - x_1))\nabla f(x_0)$ par continuité du gradient. Il en résulte donc que, pour $\alpha > 0$, suffisamment petit, $f(x_0) < f(x_0 + \alpha(y_1 - x_1))$. Cela est absurde par définition de x_0 (et de t_0).

□

Comme conséquence du lemme de Arrow-Enthoven-Avriel et du lemme de Crouzeix, on obtient la CNS suivante de quasi-convexité lorsque le gradient de la fonction ne s'annule pas.

Corollaire 42 – Condition d'ordre deux

Soit $C \subset E$ un ouvert convexe et soit $f : C \rightarrow \mathbb{R}$ une fonction de classe $\mathcal{C}^2$. On suppose de plus que pour tout $x \in C$, $d_x f$ n'est pas identiquement nulle. Alors f est quasi-convexe si, et seulement si, pour tout $x \in C$, pour tout $v \in E$:

$$d_x f(v) = 0 \implies d_x^2 f(v, v) \geq 0.$$

Démonstration. Le corollaire 42 est une conséquence immédiate des deux lemmes qui précèdent (lemmes 40 et 41). $\square$

Remarque. Une autre façon de formuler le corollaire 42 est que, sous les hypothèses dudit théorème, f est quasi-convexe si, et seulement si, pour tout $x \in C$, la forme quadratique $v \mapsto d_x^2 f(v, v)$ restreinte à $\ker(d_x f)$ est positive. On peut donc appliquer le critère donné par la proposition 14 pour déterminer si une fonction est quasi-convexe.

On donne enfin une condition suffisante, résultant du lemme de Crouzeix et permettant de montrer qu'une fonction est strictement quasi-convexe.

Corollaire 43

Soit $C \subset E$ un ouvert convexe et soit $f : C \rightarrow \mathbb{R}$ une fonction de classe $\mathcal{C}^2$. Si pour tout $x \in C$, pour tout $v \in E \setminus \{0\}$:

$$d_x f(v) = 0 \implies d_x^2 f(v, v) > 0,$$

alors f est strictement quasi-convexe.

Remarque. L'hypothèse « pour tout $x \in C$, pour tout $v \in E \setminus \{0\} : d_x f(v) = 0 \implies d_x^2 f(v, v) > 0$ » implique en fait la seconde hypothèse du lemme de Crouzeix : « pour tout $x \in C$ tel que $d_x f \equiv 0$, pour tout $v \in E$, $d_x^2 f(v, v) > 0$ ».

Démonstration. [AVRIEL et al., 1988]

Il est clair, d'après le lemme de Crouzeix (lemme 41) que f est quasi-convexe. Supposons par l'absurde qu'elle n'est pas strictement quasi-convexe. D'après la proposition 37, il existe un intervalle non trivial sur lequel f est constante.

On écrit alors un développement de Taylor en n'importe quel point x_2 intérieur à cet intervalle :

$$f(x_2 + t(y_1 - x_1)) = f(x_2) + t d_{x_2} f(y_1 - x_1) + \frac{t^2}{2} d_{x_2}^2 f(y_1 - x_1, y_1 - x_1) + o(t^2)$$

La contradiction vient du fait que $d_{x_2} f(y_1 - x_1) = 0$ (car la fonction $t \mapsto f(x_2 + t(y_1 - x_1))$ est constante au voisinage de 0) et donc que $d_{x_2}^2 f(y_1 - x_1, y_1 - x_1) > 0$ par hypothèse. Cela est absurde car on aurait alors, au voisinage de 0, $f(x_2 + t(y_1 - x_1)) > f(x_2)$. $\square$

Corollaire 44

Soit E un $\mathbb{R}$ espace vectoriel de dimension finie $n \geq 2$. Soit $C \subset E$ un ouvert convexe et soit $f : C \rightarrow \mathbb{R}$ une fonction de classe $\mathcal{C}^2$. On suppose que pour tout $x \in C$, pour tout $2 \leq k \leq n$, $\left| \overline{Hf(x)}_k^{-1} \nabla f(x)_{1,k} \right| < 0$. Alors f est strictement quasi-convexe sur C .

Démonstration. [AVRIEL et al., 1988]

Si l'on veut appliquer le corollaire 43, il est nécessaire de prouver que, pour tout $x \in C$, la forme quadratique $d_x^2 f_x$ restreinte à $\text{Ker } d_x f \simeq \text{Ker}({}^t \nabla f(x))$ est définie positive.

La condition $\left| \overline{Hf(x)}_k^{-1} \nabla f(x) \right| < 0$ (pour $2 \leq k \leq n$) implique que pour tout $x \in C$, $\nabla f(x) \neq 0$. Quitte à permuter les coordonnées, on peut d'ailleurs toujours supposer que $\partial_1 f(x) \neq 0$ (voir exemple 1, page 10). On en déduit alors le résultat par application du critère de Hotelling-Mann (Théorème 9) et du corollaire 43. $\square$

4. Résultats équivalents pour la quasi-concavité

Théorème 45

Soit $U \subset E$ un ouvert et soit $f : U \rightarrow \mathbb{R}$ une fonction différentiable. Soit $C \subset U$ un ensemble convexe. Alors $f|_C$ est quasi-concave si, et seulement si pour tous $x, y \in C$:

$$f(x) \geq f(y) \implies d_x f(y - x) \geq 0.$$

En supposant de plus que pour tout $x \in C$, $d_x f$ n'est pas identiquement nulle, $f|_C$ est strictement quasi-concave si, et seulement si pour tous $x, y \in C$ avec $x \neq y$:

$$f(x) \geq f(y) \implies d_x f(y - x) > 0.$$

Lemme 46 – Arrow - Enthoven - Avriel

Soit $C \subset E$ un ouvert convexe et soit $f : C \rightarrow \mathbb{R}$ une fonction quasi-concave de classe $\mathcal{C}^2$. Alors pour tout $x \in C$, pour tout $v \in E$:

$$d_x f(v) = 0 \implies d_x^2 f(v, v) \leq 0.$$

Lemme 47 – Crouzeix

Soit $C \subset E$ un ouvert convexe et soit $f : C \rightarrow \mathbb{R}$ une fonction quasi-convexe de classe $\mathcal{C}^2$. On suppose de plus que f vérifie les deux assertions suivantes :

- pour tout $x \in C$, pour tout $v \in E$: $d_x f(v) = 0 \implies d_x^2 f(v, v) \leq 0$
- pour tout $x \in C$ tel que $d_x f \equiv 0$, pour tout $v \in E$, $d_x^2 f(v, v) < 0$.

Alors f est quasi-concave.

Corollaire 48

Soit $C \subset E$ un ouvert convexe et soit $f : C \rightarrow \mathbb{R}$ une fonction de classe $\mathcal{C}^2$. On suppose de plus que pour tout $x \in C$, $d_x f$ n'est pas identiquement nulle. Alors f est quasi-concave si, et seulement si, pour tout $x \in C$, pour tout $v \in E$:

$$d_x f(v) = 0 \implies d_x^2 f(v, v) \leq 0.$$

Corollaire 49

Soit $C \subset E$ un ouvert convexe et soit $f : C \rightarrow \mathbb{R}$ une fonction de classe $\mathcal{C}^2$. On suppose de plus que pour tout $x \in C$, $d_x f$ n'est pas identiquement nulle. Si pour tout $x \in C$, pour tout $v \in E \setminus \{0\}$:

$$d_x f(v) = 0 \implies d_x^2 f(v, v) < 0,$$

alors f est strictement quasi-concave.

Corollaire 50

Soit E un $\mathbb{R}$ espace vectoriel de dimension finie $n \geq 2$. Soit $C \subset E$ un ouvert convexe et soit $f : C \rightarrow \mathbb{R}$ une fonction de classe $\mathcal{C}^2$. On suppose que pour tout $x \in C$, pour tout $2 \leq k \leq n$, $(-1)^{k-1} \left| \overline{Hf(x)}_k^{\nabla f(x)_{1,k}} \right| > 0$. Alors f est strictement quasi-concave sur C .

5. Exemples

Pour chacun des exemples suivants, on présentera les calculs de la matrice Hessienne bordée (que l'on notera plus simplement $\bar{H}$ afin de déterminer si les fonctions données sont quasi-convexe (quasi-concave) ou non. On donnera par ailleurs, lorsque cela est possible, d'autres arguments permettant d'assurer cette quasi-convexité (quasi-concavité). On verra que ces arguments sont en général plus simples et moins calculatoires.

Exemple 6. On considère la fonction f définie sur $\mathbb{R}^2$ par

$$f(x_1, x_2) = \ln(1 + x_1^2 + x_2^2).$$

Un simple calcul montre que le gradient et la matrice Hessienne de f sont :

$$\forall x \in \mathbb{R}^2, \quad \nabla f(x) = \begin{pmatrix} \frac{2x_1}{r} \\ \frac{2x_2}{r} \end{pmatrix} \quad \text{et} \quad Hf(x) = \frac{1}{r^2} \begin{pmatrix} 2r - 4x_1^2 & -4x_1 x_2 \\ -4x_1 x_2 & 2r - 4x_2^2 \end{pmatrix},$$

où l'on a posé $r = 1 + x_1^2 + x_2^2$. Par conséquent, pour tout $x \in \mathbb{R}^2$,

$$|Hf(x)| = \frac{4(1 - x_1^2 - x_2^2)}{r^3}.$$

On voit donc que $|\mathrm{H}f(x)|$ peut être négatif (si $x_1^2 + x_2^2 > 1$) et la matrice Hessienne n'est donc ni positive, ni négative pour de tels $x \in \mathbb{R}^2$. On en déduit que la fonction f n'est ni convexe, ni concave sur $\mathbb{R}^2$.

Le lemme Crouzeix va néanmoins permettre de montrer que f est (strictement) quasi-concave. En effet, la matrice Hessienne bordée est :

$$\forall x \in \mathbb{R}^2, \quad \overline{\mathrm{H}f}(x) = \frac{1}{r^2} \begin{pmatrix} 0 & 2rx_1 & 2rx_2 \\ 2rx_1 & 2r - 4x_1^2 & -4x_1x_2 \\ 2rx_2 & -4x_1x_2 & 2r - 4x_2^2 \end{pmatrix}$$

Le calcul du déterminant permet d'obtenir, pour tout $x \in \mathbb{R}^2$,

$$|\overline{\mathrm{H}f}(x)| = -\frac{8}{r^3}(x_1^2 + x_2^2).$$

Ainsi, pour tout $x \neq (0, 0)$, on a $\nabla f(x) \neq 0$ et $|\overline{\mathrm{H}f}(x)| \leq 0$.

De plus, si $x = (0, 0)$, la matrice Hessienne est $\mathrm{H}f(0, 0) = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$ qui est bien définie positive. Les deux conditions du lemme de Crouzeix (lemme 41) sont ainsi vérifiées et on peut donc conclure que f est quasi-convexe. Sachant que l'on a même $|\mathrm{H}f(x)| < 0$ pour tout $x \neq (0, 0)$, le corollaire 43 permet de conclure que f est strictement quasi-convexe.

Finalemt, notons que cette propriété de quasi-convexité était en fait plus facile à justifier en remarquant que la fonction $(x_1, x_2) \mapsto x_1^2 + x_2^2$ est convexe et que la fonction $\ln$ est croissante. Il s'agit alors d'une simple application de la remarque 1..

Exemple 7. Soit g la fonction définie par

$$g(x_1, x_2) = x_1 x_2,$$

que l'on considérera alternativement définie sur $\mathrm{U} = \mathbb{R}^2$ et $\mathrm{V} = (\mathbb{R}_*^+)^2$.

La matrice Hessienne de g est donnée par :

$$\forall x \in \mathbb{R}^2, \quad \mathrm{H}g(x) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Comme $\mathrm{H}g(x)$ n'est ni négative, ni positive, la fonction g n'est ni convexe, ni concave, ni sur U , ni sur V . On va de plus montrer qu'elle est strictement quasi-concave sur V bien qu'elle ne soit pas quasi-concave sur U . En effet, la matrice Hessienne bordée est :

$$\forall x \in \mathbb{R}^2, \quad \overline{\mathrm{H}g}(x) = \begin{pmatrix} 0 & x_2 & x_1 \\ x_1 & 0 & 1 \\ x_2 & 1 & 0 \end{pmatrix}.$$

Ainsi, pour tout $x \in \mathrm{V}$, $|\overline{\mathrm{H}g}(x)| = x_1^2 + x_2^2 > 0$. Le corollaire du lemme de Crouzeix (corollaire 50) permet donc de conclure que g est strictement quasi-concave sur $\mathrm{V} = (\mathbb{R}_*^+)^2$.

Il est en revanche impossible d'appliquer le lemme de Crouzeix sur U . En effet, pour $x = (0, 0)$, on a $\nabla g(x) = 0$.

Or, la matrice Hessienne en ce point est $\mathrm{H}g(0, 0) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ qui n'est pas définie négative. Les hypothèses du lemme de Crouzeix ne sont donc pas vérifiées. La fonction g n'est d'ailleurs pas quasi-concave sur $\mathrm{U} = \mathbb{R}^2$. Il suffit pour cela de considérer les points $x = (-1, -1)$ et $y = (1, 1)$. Le point $(0, 0)$ appartient au segment $[xy]$ mais $g(0, 0) = 0 > \min\{g(x), g(y)\}$.

Remarquons que l'on pouvait par ailleurs obtenir la stricte quasi-concavité de g sur V à partir des ensembles de niveau (voir proposition 36). En effet, pour tout $z \in \mathbb{R}$, pour tout $(x_1, x_2) \in \mathrm{V}$,

$$\begin{aligned} g(x_1, x_2) \geq z &\iff x_1 x_2 \geq z \\ &\iff x_2 \geq \frac{z}{x_1} \end{aligned}$$

et l'ensemble $\{(x_1, x_2) \in \mathrm{V}, x_2 \geq \frac{z}{x_1}\}$ est un ensemble strictement convexe car la fonction $x \mapsto \frac{z}{x}$ est strictement convexe.

Exemple 8. On considère la fonction h définie sur $\mathbb{R}^2$ par :

$$h(x_1, x_2) = \begin{cases} e^{-\left((x_1^2 + x_2^2)^{\frac{1}{2}} - 1\right)^2} & \text{si } (x_1^2 + x_2^2)^{\frac{1}{2}} \geq 1 \\ 1 & \text{sinon} \end{cases}$$

Il est classique de vérifier que la fonction h est $\mathcal{C}^\infty$ sur $\mathbb{R}^2$. On va de plus montrer que h est quasi-concave sur $\mathbb{R}^2$.

La fonction h étant constante sur le disque⁴ $D(0, 1)$, il est clair qu'il est impossible d'utiliser le lemme de Crouzeix pour assurer la quasi-concavité de h . On va plutôt raisonner à partir des ensembles de niveau en distinguant plusieurs cas.

- Pour tout $z \leq 0$, on a $\{(x_1, x_2) \in \mathbb{R}^2; \quad h(x_1, x_2) \geq z\} = \mathbb{R}^2$
- Pour tout $z > 1$, on a $\{(x_1, x_2) \in \mathbb{R}^2; \quad h(x_1, x_2) \geq z\} = \emptyset$
- Pour tout $0 < z \leq 1$, pour tout $(x_1, x_2) \in \mathbb{R}^2 \setminus D(0, 1)$:

$$\begin{aligned} h(x_1, x_2) \geq z &\iff e^{-\left((x_1^2 + x_2^2)^{\frac{1}{2}} - 1\right)^2} \geq z \\ &\iff -\left((x_1^2 + x_2^2)^{\frac{1}{2}} - 1\right)^2 \geq \ln(z) \\ &\iff \left((x_1^2 + x_2^2)^{\frac{1}{2}} - 1\right)^2 \leq \ln\left(\frac{1}{z}\right) \\ &\iff (x_1^2 + x_2^2)^{\frac{1}{2}} - 1 \leq \left(\ln\left(\frac{1}{z}\right)\right)^{1/2} \quad \text{car pour tout } (x_1, x_2) \in \mathbb{R} \setminus D(0, 1), \quad (x_1^2 + x_2^2)^{\frac{1}{2}} - 1 \geq 0 \\ &\iff (x_1^2 + x_2^2)^{\frac{1}{2}} \leq 1 + \left(\ln\left(\frac{1}{z}\right)\right)^{1/2} \end{aligned}$$

Par ailleurs, pour tout $(x_1, x_2) \in D(0, 1)$, pour tout $0 < z \leq 1$, on a $h(x_1, x_2) = 1 \geq z$. On en déduit donc que pour tout $0 < z \leq 1$,

$$\{(x_1, x_2) \in \mathbb{R}^2; \quad h(x_1, x_2) \geq z\} = \overline{D}\left(0; 1 + \left(\ln\left(\frac{1}{z}\right)\right)^{1/2}\right).$$

Dans tous les cas, les ensembles de niveaux sont donc convexes et on peut ainsi conclure que h est quasi-concave. Comme h est constante sur $D(0, 1)$, il est en revanche clair qu'elle n'est pas strictement quasi-concave.

Exemple 9. On considère la fonction $l : U \subset \mathbb{R}^2 \longrightarrow \mathbb{R}$ définie sur $U = \mathbb{R} \times (\mathbb{R}_*^+)$ par

$$l(x_1, x_2) = \frac{x_1}{x_2}.$$

On a déjà montré, dans l'exemple 5, que cette fonction n'est ni concave, ni convexe. On va toutefois montrer qu'elle est quasi-concave et quasi-convexe.

La matrice Hessienne bordée est :

$$\forall x \in U, \quad \overline{Hl}(x) = \begin{pmatrix} 0 & \frac{1}{x_2} & -\frac{x_1}{x_2^2} \\ \frac{1}{x_2} & 0 & -\frac{1}{x_2^2} \\ -\frac{x_1}{x_2^2} & -\frac{1}{x_2^2} & \frac{x_1}{x_2} \end{pmatrix}.$$

Un simple calcul permet d'obtenir : $|\overline{Hl}(x)| = 0$ pour tout $x \in U$.

Comme $\partial_1 l(x) \neq 0$ pour tout $x \in U$, on en déduit, d'après le critère de Hotelling généralisé (proposition 14), que la forme quadratique $v \mapsto d_x^2 l(v, v)$ restreinte à $\text{Ker}(d_x l)$ est positive. Par suite, d'après le lemme de Crouzeix pour la quasi-convexité ainsi que celui pour la quasi-concavité (lemmes 41 et 47), la fonction l est donc à la fois quasi-convexe et quasi-concave.

4. Dans toute la suite, on notera $D(0, 1)$ le disque ouvert de $\mathbb{R}^2$ de centre x et de rayon r et $\overline{D}(x, r)$ son adhérence.

La fonction l n'est toutefois ni strictement quasi-convexe, ni strictement quasi-concave. Par exemple, pour $x = (0, 1)$ et $y = (0, 2)$ on a pour tout $t \in [0 ; 1]$,

$$f(tx + (1 - t)y) = 0 = \max\{f(x), f(y)\} = \min\{f(x), f(y)\}.$$

Les propriétés de quasi-convexité et de quasi-concavité de la fonction l peuvent se retrouver facilement en considérant les ensembles de niveaux ou encore en remarquant que $l = \frac{l_1}{l_2}$ avec $l_1 : (x_1, x_2) \rightarrow x_1$ et $l_2 : (x_1, x_2) \rightarrow x_2$. Le résultat découle alors du fait que l_1 est convexe (et concave) et que l_2 est affine et strictement positive sur U (voir remarque 1.).

On termine enfin cette partie en donnant un exemple d'une fonction convexe (donc, *a fortiori* quasi-convexe), mais non strictement quasi-convexe.

Exemple 10. On considère la fonction w définie sur $\mathbb{R}^2$ par

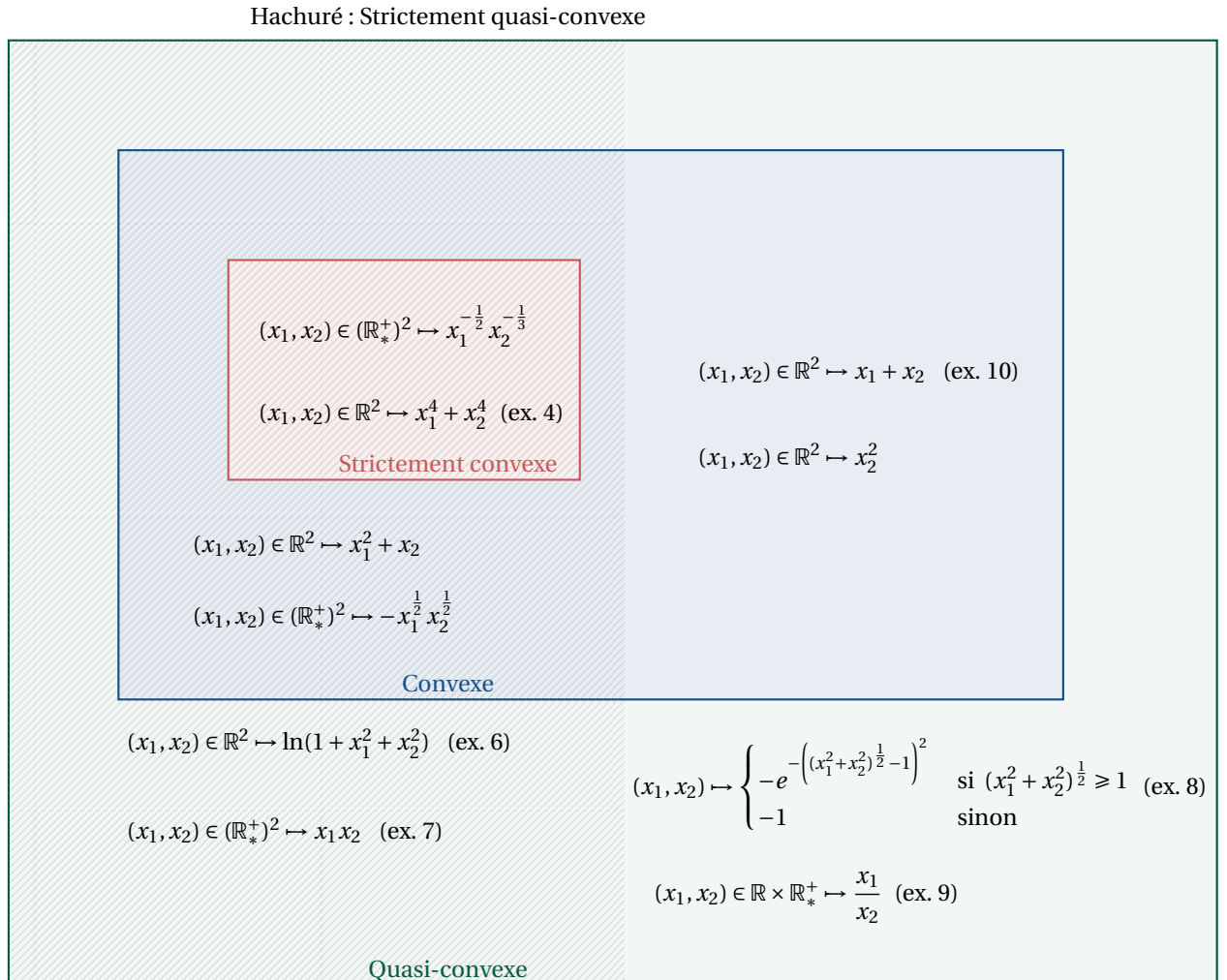
$$w(x_1, x_2) = x_1 + x_2.$$

La matrice Hessienne de w est en tout point de $\mathbb{R}^2$. Cette matrice étant positive, la fonction w est donc convexe sur $\mathbb{R}^2$ (elle est de plus concave).

En revanche, w n'est pas strictement quasi-convexe. Pour le justifier, il suffit de remarquer que, pour $x = (2, 2)$, $y = (4, 0)$ et $z = \frac{x+y}{2} = (3, 1)$, on a $f(z) = 4 = \max\{f(x), f(y)\}$.

6. Synthèse des exemples

On a représenté sur le dessin ci-dessous les ensembles des fonctions strictement convexes, convexes, etc. en respectant les relations d'inclusion et en indiquant différents exemples. Les exemples que nous n'avons pas traités précédemment ne posent pas de difficulté particulière.



IV. Extrema sous contraintes

Dans toute cette partie, E désigne un $\mathbb{R}$ espace vectoriel de dimension $n \geq 2$.

1. Multiplicateurs de Lagrange

Définition 13

Soit $U \subset E$ un ouvert, soit $f : U \rightarrow \mathbb{R}$ une fonction de classe $\mathcal{C}^1$ et, pour $1 \leq i \leq m$, soit $g_i : U \rightarrow \mathbb{R}$ des fonctions de classe $\mathcal{C}^1$. On note $C = \{x \in U, g_i(x) = 0, \forall 1 \leq i \leq m\}$. On dit que f admet un minimum (resp. maximum) en $x^* \in C$ sous les contraintes $g_i(x) = 0$ (pour $1 \leq i \leq m$) si la restriction de f à C admet un minimum (resp. maximum) en x^* .

Théorème 51 – Multiplicateurs de Lagrange

Soit $U \subset E$ un ouvert, soit $f : U \rightarrow \mathbb{R}$ une fonction de classe $\mathcal{C}^1$ et, pour $1 \leq i \leq m$, soit $g_i : U \rightarrow \mathbb{R}$ des fonctions de classe $\mathcal{C}^1$. Si f admet un extremum local en x^* sous les contraintes $g_i(x) = 0$ (pour $1 \leq i \leq m$) et si les formes linéaires $d_{x^*} g_1, \dots, d_{x^*} g_m$ sont linéairement indépendantes, alors il existe $(\lambda_i^*)_{1 \leq i \leq m}$ (appelés **multiplicateurs de Lagrange**) tels que :

$$d_{x^*} f + \sum_{i=1}^m \lambda_i^* d_{x^*} g_i = 0.$$

Remarque. On propose trois démonstrations. La première traite seulement du cas de la dimension 2 (avec une seule contrainte). Cette preuve est moins technique que les deux autres, tout en étant assez instructive. La seconde généralise le résultat en développant les mêmes idées en dimension n quelconque. On présentera enfin une dernière démonstration qui aura principalement deux intérêts. Elle sera notamment valable dans le cas des espaces de dimension infinie et elle permettra par ailleurs de démontrer le théorème ???. L'idée principale de cette troisième preuve est de ne pas travailler dans la base canonique comme nous l'avons fait dans la deuxième preuve mais plutôt dans la base $(\nabla g_i(x^*))$ qui est naturellement mieux adaptée au problème que nous considérons.

Démonstration. (Preuve 1 du théorème 51) On considère le cas $n = 2$ et $m = 1$. On notera dans la suite de la démonstration $x^* = (x_1^*, x_2^*)$.

Comme $\nabla g(x^*) \neq 0$, on peut supposer (quitte à permuter x_1 et x_2) que $\partial_2 g(x^*) \neq 0$. D'après le théorème des fonctions implicites, il existe un ouvert $W \subset U \subset \mathbb{R}^2$, un intervalle ouvert $V \subset \mathbb{R}$ et une fonction $\varphi : V \rightarrow \mathbb{R}$ telle que

$$(x_1, x_2) \in W \text{ et } g(x_1, x_2) = 0 \iff x_1 \in V \text{ et } x_2 = \varphi(x_1).$$

On sait que l'on a de plus, pour tout $x_1 \in V$,

$$\varphi'(x_1) = -\frac{\partial_1 g(x_1, \varphi(x_1))}{\partial_2 g(x_1, \varphi(x_1))}. \quad (5)$$

On considère ensuite la fonction $\Psi : x_1 \in V \rightarrow f(x_1, \varphi(x_1)) \in \mathbb{R}$. Alors Ψ admet un extremum local en x_1^* . Cela découle directement du fait que f admet un extremum local sous la contrainte $g_1(x) = 0$.

On a donc, nécessairement, $\Psi'(x_1^*) = 0$.

En calculant la dérivée de Ψ et en utilisant l'expression (5), on en déduit que

$$\partial_1 f(x^*) - \partial_2 f(x^*) \frac{\partial_1 g(x^*)}{\partial_2 g(x^*)} = 0.$$

Cela implique que $\nabla f(x^*) = \lambda \nabla g_1(x^*)$ avec $\lambda = \frac{\partial_2 f(x^*)}{\partial_2 g(x^*)}$. □

Démonstration. (Preuve 2 du théorème 51) On fait l'identification $E \simeq \mathbb{R}^n$.

Le cas $m > n$ est impossible car la famille $d_{x_0} g_1, \dots, d_{x_0} g_m$ serait nécessairement liée. De plus, si $m = n$, cette même famille est alors nécessairement une base de l'espace des formes linéaires sur E et le résultat est alors

évident. On peut donc supposer que $m < n$.

On pose $G = {}^t(g_1, \dots, g_m) : U \longrightarrow \mathbb{R}^m$. La matrice Jacobienne de G est alors

$$\forall x \in U, \quad \nabla G(x) = \begin{pmatrix} \partial_1 g_1(x) & \cdots & \cdots & \partial_n g_1(x) \\ \vdots & & & \vdots \\ \partial_1 g_m(x) & \cdots & \cdots & \partial_n g_m(x) \end{pmatrix}.$$

On sait, par hypothèse, que $\nabla G(x^*)$ est de rang m . On peut donc en extraire une sous matrice inversible de taille $m \times m$. Quitte à effectuer une permutation des variables, on peut supposer, sans perte de généralité, que $(\partial_j g_i(x^*))_{n-m+1 \leq i, j \leq n}$ est inversible.

D'après le théorème des fonctions implicites appliqué à la fonction G (qui est de classe $\mathcal{C}^1$ et qui vérifie bien les hypothèses du théorème des fonctions implicites), il existe un voisinage ouvert de $V \subset \mathbb{R}^{n-m}$ de $(x_1^*, \dots, x_{n-m}^*)$ et un voisinage ouvert $W \subset \mathbb{R}^n$ de $(x_1^*, \dots, x_n^*)$ et une fonction $\varphi : V \longrightarrow \mathbb{R}^m$ tels que, pour tout $(x_1, \dots, x_n) \in U$:

$$(x_1, \dots, x_n) \in W \text{ et } G(x_1, \dots, x_n) = 0_{\mathbb{R}^m} \iff (x_1, \dots, x_{n-m}) \in V \text{ et } (x_{n-m+1}, \dots, x_n) = \varphi(x_1, \dots, x_{n-m}).$$

On pose, pour tout $(x_1, \dots, x_{n-m}) \in V$,

$$\psi(x_1, \dots, x_{n-m}) = f\left((x_1, \dots, x_{n-m}), \varphi(x_1, \dots, x_{n-m})\right).$$

Ainsi, ψ admet un extremum local en $(x_1^*, \dots, x_{n-m}^*)$ (car f admet un extremum local en $(x_1^*, \dots, x_n^*)$ sous les contraintes $g_i(x_1, \dots, x_n) = 0$). Par conséquent, d'après la proposition 18, on a $\nabla \psi(x_1^*, \dots, x_{n-m}^*) = 0$. En utilisant la formule de dérivation de la composée et en notant $\varphi = (\varphi_1, \dots, \varphi_m)$, cela signifie exactement que, pour tout $1 \leq i \leq n-m$:

$$\partial_i f(x_1^*, \dots, x_n^*) + \sum_{j=1}^m \partial_{n-m+j} f(x_1^*, \dots, x_n^*) \partial_j \varphi(x_1^*, \dots, x_{n-m}^*) = 0. \quad (6)$$

De plus, pour tout $1 \leq k \leq m$, la fonction $(x_1, \dots, x_{n-m}) \longmapsto g_k((x_1, \dots, x_{n-m}), \varphi(x_1, \dots, x_{n-m}))$ est identiquement nulle au voisinage de $(x_1^*, \dots, x_{n-m}^*)$. Son gradient en $(x_1^*, \dots, x_{n-m}^*)$ est *a fortiori* nul et par un calcul similaire au précédent, on a : pour tout $1 \leq k \leq m$, pour tout $1 \leq i \leq n-m$:

$$\partial_i g_k(x_1^*, \dots, x_n^*) + \sum_{j=1}^m \partial_j \varphi(x_1^*, \dots, x_{n-m}^*) \partial_{n-m+j} g_k(x_1^*, \dots, x_n^*) = 0. \quad (7)$$

On considère maintenant la matrice suivante de taille $(m+1) \times n$:

$$M = \begin{pmatrix} \partial_1 f(x^*) & \cdots & \cdots & \partial_n f(x^*) \\ \partial_1 g_1(x^*) & \cdots & \cdots & \partial_n g_1(x^*) \\ \vdots & & & \vdots \\ \partial_1 g_m(x^*) & \cdots & \cdots & \partial_n g_m(x^*) \end{pmatrix}$$

D'après les relations (6) et (7), les $n-m$ premières colonnes de M s'expriment linéairement en fonction des m dernières. Par conséquent, $rg(M) \leq m$ et donc $rg({}^t M) \leq m$.

Comme la famille $(\nabla g_i(x^*))_{1 \leq i \leq m}$ est libre, c'est donc que $\nabla f(x^*) \in \text{Vect}\{\nabla g_i(x^*)\}_{1 \leq i \leq m}$, ce qui conclut la démonstration. $\square$

Démonstration. (Preuve 3 du théorème 51) [BENZONI-GAVAGE, 2021]

On considère des fonction $f, g_1, \dots, g_m$ vérifiant les hypothèses du théorème. En faisant l'identification $E \simeq \mathbb{R}^n$, on note $K = \text{Vect}\{\nabla g_i(x^*)\}_{1 \leq i \leq m}$ et $H = K^\perp$. On a $E \simeq H \oplus K \simeq H \times K$. On définit alors

$$G : \begin{cases} H \times K & \longrightarrow \mathbb{R}^m \\ (h, k) & \longmapsto \begin{pmatrix} g_1(x^* + h + k) \\ \vdots \\ g_m(x^* + h + k) \end{pmatrix} \end{cases}$$

La différentielle partielle de G par rapport à la seconde variable en $(0,0)$ est donnée par :

$$\forall l \in K, \quad d_{2,(0,0)}G(l) = \begin{pmatrix} \langle \nabla g_1(x^*), l \rangle \\ \vdots \\ \langle \nabla g_m(x^*), l \rangle \end{pmatrix}.$$

Ainsi, $d_{2,(0,0)}G$ est une application linéaire de K dans $\mathbb{R}^m$ et en prenant $(\nabla g_i(x^*))_{1 \leq i \leq m}$ comme base de K et la base canonique de $\mathbb{R}^m$, la matrice de cette application linéaire est $M = (\langle \nabla g_i(x^*), \nabla g_j(x^*) \rangle)_{1 \leq i, j \leq m}$. Cette matrice M est en fait la matrice du produit scalaire canonique de $\mathbb{R}^m$ exprimé dans la base $(\nabla g_i(x^*))_{1 \leq i \leq m}$. M est donc congruente à l'identité et elle est par conséquent inversible. Finalement, cela signifie que $d_{2,(0,0)}G$ est un isomorphisme et on peut ainsi appliquer le théorème des fonctions implicites à la fonction G (qui est de classe $\mathcal{C}^1$ comme composée de fonctions de classe $\mathcal{C}^1$). Il existe donc $V \subset H$ un voisinage ouvert de 0 et $W \subset H \times K$ un voisinage ouvert de $(0,0)$ et une fonction $\varphi : V \rightarrow K$ tels que

$$(h, k) \in W \text{ et } G(h, k) = 0 \iff h \in V \text{ et } k = \varphi(h).$$

Ainsi, comme f admet un extremum local en x^* sous les contraintes $g_1(x) = 0, \dots, g_m(x) = 0$, la fonction $\Psi : h \in V \mapsto f(x^* + h + \varphi(h))$ admet un extremum local en 0 . On a donc nécessairement $d_0\Psi = 0$. On en déduit alors, en différentiant la fonction Ψ que

$$d_{x^*}f(h + d_0\varphi(h)) \quad \text{pour tout } h \in H. \quad (8)$$

De plus, par construction de φ , $d_0\varphi$ est identiquement nulle sur H (ce qui est équivalent à dire que $\nabla\varphi(0) = 0$). En effet, comme φ est à valeurs dans $K = \text{Vect}\{\nabla g_i(x^*)\}_{1 \leq i \leq m}$, il suffit de montrer que pour tout $h \in H$, $\langle \nabla\varphi(0), \nabla g_i(x^*) \rangle = 0$ pour tout $1 \leq i \leq m$. Ceci s'obtient en fait en différentiant la fonction $h \mapsto g_i(x^* + h + \varphi(h))$ qui est identiquement nulle sur V et en remarquant que pour tout $h \in H$, pour tout $1 \leq i \leq m$, $\langle \nabla g_i(x^*), h \rangle = 0$ (cela résulte immédiatement de la définition de H).

Au final, en utilisant la relation 8, on obtient que d_0f est identiquement nulle sur H . Autrement dit, $\langle \nabla f(x^*), h \rangle = 0$ pour tout h et on en conclut donc que $\nabla f(x^*) \in H^\perp = K$, ce qui conclut la démonstration. $\square$

Proposition 52 – Condition suffisante de minimum sous contraintes

Soit $U \subset E$ un ouvert, soit $f : U \rightarrow \mathbb{R}$ une fonction de classe $\mathcal{C}^2$ et, pour $1 \leq i \leq m$, soit $g_i : U \rightarrow \mathbb{R}$ des fonctions de classe $\mathcal{C}^2$ et soit $x^* \in U$ tel que pour tout $1 \leq i \leq m$, $g_i(x^*) = 0$. On suppose que les formes linéaires $d_{x^*}g_1, \dots, d_{x^*}g_m$ sont linéairement indépendantes et qu'il existe par ailleurs $(\lambda_i^*)_{1 \leq i \leq m}$ tels que

$$d_{x^*}f + \sum_{i=1}^m \lambda_i^* d_{x^*}g_i = 0$$

et tels que la forme quadratique

$$d_{x^*}^2f + \sum_{i=1}^m \lambda_i^* d_{x^*}^2g_i$$

est définie positive sur le sous espace $H = \left(\text{Vect}\{d_{x^*}g_i\}_{1 \leq i \leq m} \right)^\perp$.

Alors f admet un minimum local en x^* sous les contraintes $g_i(x) = 0$ (pour $1 \leq i \leq m$).

Démonstration. [BENZONI-GAVAGE, 2021]

On reprend les notations de la preuve 3 du théorème 51 et on rappelle que l'on note $C = \{x \in U, g_i(x) = 0, \forall 1 \leq i \leq m\}$. On note de plus

$$Z = C \cap (x^* + W) = \{x^* + h + k, (h, k) \in W, G(h, k) = 0\} = \{x^* + h + \varphi(h), h \in V\}.$$

L'objectif est de montrer que, quitte à restreindre W , on a $f(x) \geq f(x^*)$ pour tout $x \in Z$. Cela est d'ailleurs équivalent de montrer que, quitte à restreindre V , on a, pour tout $h \in V$, $f(x^* + h + \varphi(h)) \geq f(x^*)$.

En appliquant le théorème de Taylor Young à la fonction $f + \sum_{i=1}^m \lambda_i^* g_i$ en x^* (où les λ_i^* sont les multiplicateurs de Lagrange donnés par le théorème 51), on obtient :

$$f(x^* + l) + \sum_{i=1}^m \lambda_i^* g_i(l) = f(x^*) + \frac{1}{2} A(l, l) + o(\|l\|^2)$$

où $A = d_{x^*}^2 f + \sum_{i=1}^m \lambda_i^* d_{x^*}^2 g_i$.

En prenant $l = h + \varphi(h)$, on a :

$$f(x^* + h + \varphi(h)) = f(x^*) + \frac{1}{2} A(h + \varphi(h), h + \varphi(h)) + o(\|h + \varphi(h)\|^2).$$

Plus précisément, cela signifie qu'il existe $\varepsilon : V \rightarrow \mathbb{R}$ telle que $\lim_{h \rightarrow 0} \varepsilon(h) = 0$ et telle que pour tout $h \in V$:

$$f(x^* + h + \varphi(h)) = f(x^*) + \frac{1}{2} A(h + \varphi(h), h + \varphi(h)) + \|h + \varphi(h)\|^2 \varepsilon(h).$$

On est ainsi ramené à prouver que, au voisinage de 0,

$$\frac{1}{2} A(h + \varphi(h), h + \varphi(h)) + \|h + \varphi(h)\|^2 \varepsilon(h) > 0.$$

Par hypothèse, A est définie positive sur H . Comme H est de dimension finie, on sait même que A est coercive : il existe $\alpha > 0$ tel que pour tout $h \in H$, $A(h, h) \geq \alpha \|h\|^2$. De plus, d'après la preuve du théorème 51, on sait que $\varphi(0) = 0$ et que $d_0 \varphi = 0$. On en déduit donc que pour tout $c > 0$, il existe $\eta > 0$ tel que pour tout $h \in V \cap B(0, \eta)$, $\|\varphi(h)\| \leq c \|h\|$. En prenant $c = \min\left(\frac{1}{2}, \frac{\alpha}{4\|A\|}\right)$, on obtient les inégalités suivantes :

- $A(h, h) \geq \alpha \|h\|^2$
- $|A(h, \varphi(h))| \leq \frac{\alpha}{4} \|h\|^2$
- $|A(\varphi(h), \varphi(h))| \leq \frac{\alpha}{4} \|h\|^2$

Par conséquent, pour tout $h \in V \cap B(0, \eta)$,

$$\begin{aligned} \frac{1}{2} A(h + \varphi(h), h + \varphi(h)) &= \frac{1}{2} (A(h, h) + 2A(h, \varphi(h)) + A(\varphi(h), \varphi(h))) \\ &\geq \frac{\alpha}{8} \|h\|^2 \end{aligned}$$

Enfin, comme pour tout $h \in V \cap B(0, \eta)$,

$$\frac{1}{2} \|h\| \leq \|h + \varphi(h)\| \leq \frac{3}{2} \|h\|,$$

on voit qu'il existe $0 < \eta' \leq \eta$ tel que pour tout $h \in V \cap B(0, \eta')$,

$$\|h + \varphi(h)\|^2 |\varepsilon(h)| \leq \frac{\alpha}{16} \|h\|^2.$$

Au final, on aura donc, pour tout $h \in V \cap B(0, \eta')$,

$$\frac{1}{2} A(h + \varphi(h), h + \varphi(h)) + \|h + \varphi(h)\|^2 \varepsilon(h) \geq \frac{\alpha}{16} \|h\|^2 > 0,$$

ce qui conclut la démonstration. □

Proposition 53 – Condition suffisante de maximum sous contraintes

Soit $U \subset E$ un ouvert, soit $f : U \rightarrow \mathbb{R}$ une fonction de classe $\mathcal{C}^2$ et, pour $1 \leq i \leq m$, soit $g_i : U \rightarrow \mathbb{R}$ des fonctions de classe $\mathcal{C}^2$ et soit $x^* \in U$ tel que pour tout $1 \leq i \leq m$, $g_i(x^*) = 0$. On suppose que les formes linéaires $d_{x^*} g_1, \dots, d_{x^*} g_m$ sont linéairement indépendantes et qu'il existe par ailleurs $(\lambda_i^*)_{1 \leq i \leq m}$ tels que

$$d_{x^*} f + \sum_{i=1}^m \lambda_i^* d_{x^*} g_i = 0$$

et tels que la forme quadratique

$$d_{x^*}^2 f + \sum_{i=1}^m \lambda_i^* d_{x^*}^2 g_i$$

est définie négative sur le sous espace $H = \left(\text{Vect} \{ d_{x^*} g_i \}_{1 \leq i \leq m} \right)^\perp$.

Alors f admet un maximum local en x^* sous les contraintes $g_i(x) = 0$ (pour $1 \leq i \leq m$).

Démonstration. Il suffit d'appliquer la proposition 52 à la fonction $-f$. □

2. Reformulation avec le Lagrangien et conditions d'ordre deux**Définition 14**

Soit $U \subset E$ un ouvert, soit $f : U \rightarrow \mathbb{R}$ une fonction de classe $\mathcal{C}^1$ et, pour $1 \leq i \leq m$, soit $g_i : U \rightarrow \mathbb{R}$ des fonctions de classe $\mathcal{C}^1$. On appelle **Lagrangien** de f sous les contraintes $g_1, \dots, g_m$, l'application

$$\mathcal{L} : \begin{cases} \mathbb{R}^m \times U & \longrightarrow \\ \left((\lambda_1, \dots, \lambda_m), x \right) & \longmapsto f(x) + \sum_{i=1}^m \lambda_i g_i(x) \end{cases}$$

Le Lagrangien nous permet alors de reformuler le théorème 51.

Théorème 54 – Multiplieurs de Lagrange

Soit $U \subset E$ un ouvert, soit $f : U \rightarrow \mathbb{R}$ une fonction de classe $\mathcal{C}^1$ et, pour $1 \leq i \leq m$, soit $g_i : U \rightarrow \mathbb{R}$ des fonctions de classe $\mathcal{C}^1$. Si f admet un minimum local en x^* sous les contraintes $g_i(x) = 0$ (pour $1 \leq i \leq m$) et si les formes linéaires $d_{x^*} g_1, \dots, d_{x^*} g_m$ sont linéairement indépendantes, alors il existe $(\lambda_i^*)_{1 \leq i \leq m}$ tels que :

$$d_{(\lambda_1^*, \dots, \lambda_m^*, x^*)} \mathcal{L} \equiv 0.$$

Remarque. Autrement dit, la condition $d_{(\lambda_1^*, \dots, \lambda_m^*, x^*)} \mathcal{L} \equiv 0$ signifie que $(\lambda_1^*, \dots, \lambda_m^*, x^*)$ est un point stationnaire du Lagrangien.

Démonstration. Le résultat découle immédiatement du fait que

$$\nabla \mathcal{L}(\lambda_1, \dots, \lambda_m, x) = {}^t \left(g_1(x), \dots, g_m(x), \nabla f(x) + \sum_{i=1}^m \lambda_i \nabla g_i(x) \right)$$

et de l'application du théorème 51. □

Théorème 55 – Condition suffisante de minimum sous contraintes

Soit $U \subset E$ un ouvert, soit $f : U \rightarrow \mathbb{R}$ une fonction de classe $\mathcal{C}^2$ et, pour $1 \leq i \leq m$, soit $g_i : U \rightarrow \mathbb{R}$ des fonctions de classe $\mathcal{C}^2$. Soit (λ^*, x^*) un point stationnaire du Lagrangien $\mathcal{L}$. On suppose de plus que $(-1)^m \left| \left(H\mathcal{L}(\lambda^*, x^*) \right)_k \right| > 0$ pour tout $2m+1 \leq k \leq m+n$. Alors x^* est un minimum local sous les contraintes $g_i(x) = 0$ (pour $1 \leq i \leq m$).

Démonstration. On commence par remarquer que si (x^*, λ^*) est un point stationnaire du Lagrangien, alors $g_i(x^*) = 0$ pour tout $1 \leq i \leq m$ et $d_{x^*} f + \sum_{i=1}^m \lambda_i^* d_{x^*} g_i = 0$.

Ainsi, en notant $A = d_{x^*}^2 f + \sum_{i=1}^m \lambda_i^* d_{x^*}^2 g_i$, on sait qu'il suffit de montrer que A est définie positive sur

$\left(\text{Vect}\{d_{x^*} g_i\}_{1 \leq i \leq m}\right)^\perp$ pour pouvoir conclure avec la proposition 52.

Or, $\left(\text{Vect}\{d_{x^*} g_i\}_{1 \leq i \leq m}\right)^\perp \simeq \text{Ker}(B)$ avec $B = {}^t(\nabla g_1(x^*), \dots, \nabla g_m(x^*))$.

D'après le critère de Hotelling-Mann, il suffit donc de vérifier que les mineurs de la matrice bordée $\begin{pmatrix} 0 & B \\ {}^t B & \text{Mat}(A) \end{pmatrix}$ sont de même signe de $(-1)^m$. Cette condition est vérifiée par hypothèse car

$$\begin{pmatrix} 0 & B \\ {}^t B & \text{Mat}(A) \end{pmatrix} = H\mathcal{L}(\lambda^*, x^*).$$

□

Théorème 56 – Condition suffisante de maximum sous contraintes

Soit $U \subset E$ un ouvert, soit $f : U \rightarrow \mathbb{R}$ une fonction de classe $\mathcal{C}^2$ et, pour $1 \leq i \leq m$, soit $g_i : U \rightarrow \mathbb{R}$ des fonctions de classe $\mathcal{C}^2$. Soit (λ^*, x^*) un point stationnaire du Lagrangien $\mathcal{L}$. On suppose de plus que $(-1)^{k-m} \left| \left(H\mathcal{L}(\lambda^*, x^*) \right)_k \right| > 0$ pour tout $2m+1 \leq k \leq m+n$. Alors x^* est un maximum local sous les contraintes $g_i(x) = 0$ (pour $1 \leq i \leq m$).

Démonstration. Là aussi, il suffit d'appliquer le théorème précédent à la fonction $-f$. □

Exemple 11. Soit $p > 0$. On va montrer que parmi tous les triangles de périmètre $2p$, le triangle équilatéral de côtés $x_1 = x_2 = x_3 = \frac{2p}{3}$ est le triangle dont l'aire est maximale.

On rappelle la formule de Héron : si T est un triangle dont les côtés sont x_1 , x_2 et x_3 et dont le périmètre est $2p = x_1 + x_2 + x_3$, alors l'aire de T est :

$$\mathcal{A} = \sqrt{p(p-x_1)(p-x_2)(p-x_3)}.$$

On considère la fonction $f : (x_1, x_2, x_3) \in \mathbb{R}^3 \rightarrow p(p-x_1)(p-x_2)(p-x_3)$ et la fonction $g : (x_1, x_2, x_3) \in \mathbb{R}^3 \rightarrow x_1 + x_2 + x_3 - 2p$. Le problème consiste donc à déterminer le maximum de f sur l'ensemble $\mathcal{C} = \{(x_1, x_2, x_3) \in \mathbb{R}^3; g(x_1, x_2, x_3) = 0, \forall i \in \{1, 2, 3\}, 0 \leq x_i \leq p\}$ et de montrer qu'il est atteint pour $x_1 = x_2 = x_3 = \frac{2p}{3}$. Comme f est une fonction continue et que $\mathcal{C}$ est un ensemble compact, il est clair que f admet un maximum et un minimum sur $\mathcal{C}$. On peut remarquer que pour tout $(x_1, x_2, x_3) \in \partial \mathcal{C}$, on a $f(x_1, x_2, x_3) = 0$ et que pour tout $(x_1, x_2, x_3) \in \mathring{\mathcal{C}}$, $f(x_1, x_2, x_3) > 0$. Cela signifie donc que le maximum est atteint à l'intérieur de $\mathcal{C}$. Par conséquent, le maximum de f sur $\mathcal{C}$ est un maximum (au moins) local de f sous la contrainte $g(x_1, x_2, x_3) = 0$.

On est donc ramené à déterminer le maximum de f sous la contrainte $g(x_1, x_2, x_3) = 0$.

Le Lagrangien est défini par :

$$\mathcal{L}(\lambda, x_1, x_2, x_3) = p(x-x_1)(p-x_2)(p-x_3) + \lambda(x_1 + x_2 + x_3 - 2p).$$

On a donc

$$\nabla \mathcal{L}(\lambda, x_1, x_2, x_3) = \begin{pmatrix} x_1 + x_2 + x_3 - 2p \\ \lambda - p(p-x_2)(p-x_3) \\ \lambda - p(p-x_1)(p-x_3) \\ \lambda - p(p-x_1)(p-x_2) \end{pmatrix}.$$

Une simple résolution permet de montrer que $\nabla \mathcal{L}(\lambda, x_1, x_2, x_3) = 0 \iff x_1 = x_2 = x_3 = \frac{2p}{3}$.

Le Lagrangien admet ainsi un unique point stationnaire et il s'agit du maximum que nous cherchons.

Nous avons déjà justifié du fait que le point stationnaire trouvé est en fait un maximum local du problème sous contrainte. Nous allons maintenant retrouver ce résultat à l'aide de la matrice Hessienne du Lagrangien (voir théorème 56).

En fait, si on note (λ^*, x^*) le point stationnaire du Lagrangien, on a

$$H\mathcal{L}(\lambda^*, x^*) = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & \frac{p^2}{3} & \frac{p^2}{3} \\ 1 & \frac{p^2}{3} & 0 & \frac{p^2}{3} \\ 1 & \frac{p^2}{3} & \frac{p^2}{3} & 0 \end{pmatrix}.$$

Les deux mineurs à calculer sont donc

$$|(H\mathcal{L}(\lambda^*, x^*))_3| = \begin{vmatrix} 0 & 1 & 1 \\ 1 & 0 & \frac{p^2}{3} \\ 1 & \frac{p^2}{3} & 0 \end{vmatrix} = \frac{2p^2}{3} > 0$$

et

$$|(H\mathcal{L}(\lambda^*, x^*))_4| = \begin{vmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & \frac{p^2}{3} & \frac{p^2}{3} \\ 1 & \frac{p^2}{3} & 0 & \frac{p^2}{3} \\ 1 & \frac{p^2}{3} & \frac{p^2}{3} & 0 \end{vmatrix} = -\frac{p^2}{3} < 0$$

Ainsi, comme le nombre de conditions est $m = 1$, on a :

- pour $k = 3$, $(-1)^{3-1} |(H\mathcal{L}(\lambda^*, x^*))_3| > 0$
- pour $k = 4$, $(-1)^{4-1} |(H\mathcal{L}(\lambda^*, x^*))_4| > 0$

Les hypothèses du théorème 56 sont vérifiées et on retrouve bien le fait que le point stationnaire est un maximum local sous contrainte.

Références

- AVRIEL, M., DIEWERT, W., SCHAIBLE, S., & ZANG, I. (1988). *Generalized Concavity*. Society for Industrial; Applied Mathematics.
- BENZONI-GAVAGE, S. (2021). *Calcul différentiel et équations différentielles : Cours et exercices corrigés* (2e). Dunod.
- CALDERO, P., & PERONIER, M. (2022). *Carnet de voyage en Algérie*. Calvage et Mounet.
- CAMBINI, A., & MARTEIN, L. (2008). *Generalized Convexity and Optimization: Theory and Applications*. Springer Berlin Heidelberg.
- CROUZEIX, J. P. (1980). On second order conditions for quasiconvexity. *Mathematical Programming*, 18, 349-352.
- DEBREU, G. (1952). Definite and semidefinite quadratic forms. *Econometrica*, 20(2), 295-300.
- KOJIC, V. (2019). A Note on Concavity Conditions of Cobb–Douglas and CES Production Function with at Least Two Inputs. *Studies in Microeconomics*.