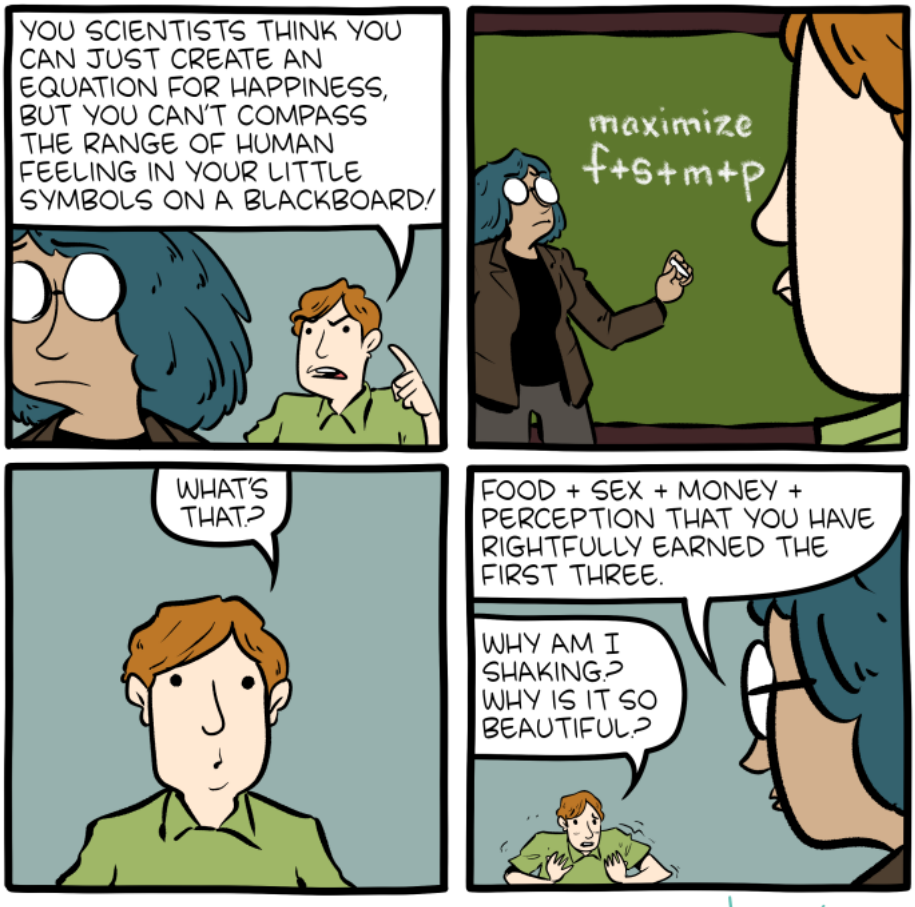
Université Paris-Saclay

OPTIMISATION ET CALCUL DIFFERENTIEL

Amaury Freslon

Département de Mathématiques d'Orsay

2024 – 2025



AVANT-PROPOS

Ce qui m'attriſte n'eſt pas la penſée de toutes les bêtiſes que j'ai pu dire dans ma vie,
mais de toutes celles qui me reſtent à dire – et peut-être à écrire.

J. Green, Ce qui reſte de jour

Ce document eſt un support pour le cours intitulé *Calcul Différentiel et Optimisation* donné à l'Université Paris-Saclay. Il s'adresse à des étudiants suivant un cursus menant à un double diplôme en Mathématiques et dans l'un des disciplines suivantes : Économie, Informatique, Physique, Sciences de la vie. C'eſt pourquoi, sans sacrifier la rigueur mathématique, nous nous sommes efforcés de mettre l'accent sur l'optimisation et ses interactions avec certaines de ces disciplines.

Ces notes couvrent l'intégralité du programme du cours, mais ne sont qu'un aperçu de plusieurs vaſtes champs des mathématiques. Pour les personnes qui souhaiteraient aller plus loin, voici quelques conseils bibliographiques.

- Pour un traitement approfondi du calcul différentiel, une référence incontournable eſt le grand classique [9]. Ce livre donne également quelques ouvertures vers la géométrie différentielle (à peine esquissée dans ces notes), pour laquelle on pourra par ailleurs consulter [7].
- Concernant l'optimisation, le champ eſt vaſte ! On peut néanmoins conseiller [2], qui traite la plupart des sujets abordés dans ce cours tout en faisant la part belle aux aſpects algorithmiques, ainsi que la seconde partie de l'ouvrage [1].
- L'optimisation linéaire eſt un sujet en soit, qui eſt par exemple traité en détails dans [12]. Plus généralement il exiſte de nombreux ouvrages sur l'optimisation convexe, et l'on peut notamment recommander [6].
- L'optimisation eſt également souvent traitée dans les manuels de mathématiques pour l'économie. Nous nous sommes particulièrement appuyés sur deux tels ouvrages, à savoir [5] et [10] pour certains exemples et certaines interprétations.
- Un autre débouché naturel de l'optimisation eſt la *recherche opérationnelle*, dont on trouvera une description détaillée dans [12] et [13].

Profitons de cet avant-propos pour remercier toutes les personnes qui ont fait des remarques sur des versions préliminaires de ce document, et en particulier Victor Dubois pour sa relecture attentive. Mentionnons pour conclure que les figures planes de ce texte ont été réalisées avec GeoGebra tandis que les images de surfaces sont issues de 3D-XplorMath.

TABLE DES MATIÈRES

CHAPITRE 1

INTRODUCTION

Usus magister est optimus [1]

Locution latine

L'objectif de ce texte est d'aborder des outils fondamentaux pour résoudre des problèmes dits d'*optimisation*. De quoi s'agit-il ? Il s'agit d'une situation où l'on cherche une solution la meilleure possible [2] étant données certaines contraintes. Ce genre de problème est omniprésent dans de nombreuses disciplines et notamment – à titre d'exemples –

- En économie, que ce soit au niveau des acteurs individuels (une entreprise cherche à gagner le plus d'argent possible) ou à un niveau plus global (trouver un système de prix qui satisfasse à la fois les vendeurs et les acheteurs).
- En biologie, dans l'élaboration des stratégies comportementales de certains animaux qui doivent trouver le moyen de se nourrir au mieux et de se reproduire le plus possible tout en protégeant les individus plus jeunes par exemple.
- En physique, où beaucoup de lois peuvent être déduites de principes affirmant que l'évolution d'un système sera, parmi tous les possibles, celle qui minimise le temps (Principe de Fermat) ou une autre quantité appelé *action* (Principe de Maupertuis).

Un problème d'optimisation est donc, avant tout, un problème de la vie réelle qu'il s'agit de résoudre. Pour cela, on passe par une phase de modélisation qui va produire une fonction f. Cette fonction mesure la quantité jugée critique dans le problème de départ – on l'appelle souvent *fonction objectif* – et l'optimisation va consister à trouver un *extremum* de la fonction, c'est-à-dire un *minimum* ou un *maximum*. La fonction f dépend des paramètres du problème, qui peuvent être nombreux. En particulier, f peut avoir plus d'une variable, et c'est pourquoi nous allons étudier dans ce cours les fonctions dites "de plusieurs variables". Quant aux différentes contraintes du problème, elles se traduisent par des relations entre les différentes variables qui doivent être satisfaites.

1. *La pratique est le meilleur des maîtres.*
2. C'est précisément le sens du mot *optimus, a, um* en latin : c'est le superlatif de *bonus, a, um* qui signifie bon.

Le cadre étant posé, nous allons maintenant présenter plusieurs problèmes d'optimisation issus de domaines différents. L'objectif est double. D'une part, nous voulons montrer que l'optimisation est omniprésente et qu'il est donc pertinent – voire indispensable – de savoir gérer ces questions et d'autre part, nous voulons introduire différents types de problèmes d'optimisation qui nécessiteront des techniques différentes ou plus ou moins élaborées, afin de motiver la suite de ce texte.

1.1 Deux problèmes élémentaires

S'il n'y a pas de solution, c'est qu'il n'y a pas de problème.

J. Rouxel, Les Shadocks

Nous allons commencer par présenter deux problèmes d'optimisation relativement élémentaires, au sens où il est possible de les résoudre complètement (ce que nous allons d'ailleurs faire) avec des outils de première année de Licence. Cela permettra néanmoins de préciser ce que nous entendons par le vocable "problème d'optimisation" et d'illustrer quelques aspects importants des techniques d'optimisation que nous développerons dans la suite.

1.1.1 Apis mellifica

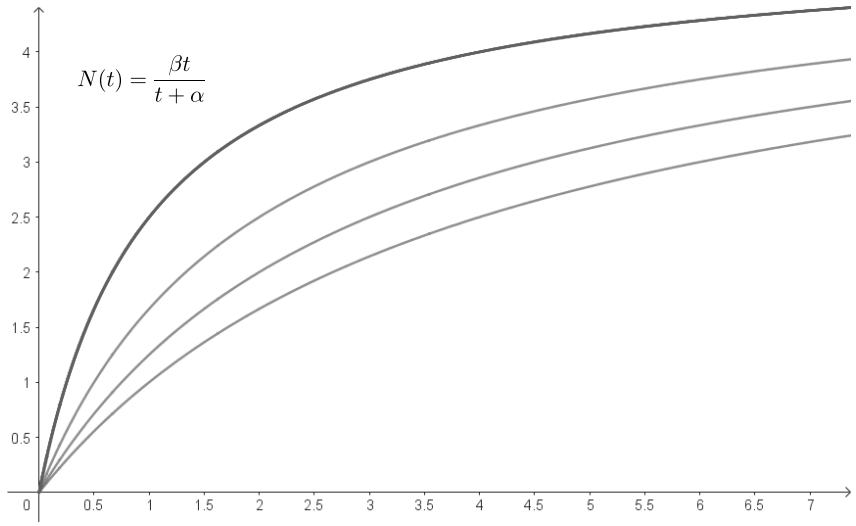
Les abeilles, comme chacun sait, butinent les fleurs. Plus précisément, elles extraient de certaines fleurs une substance appelée *nectar*. Le problème, c'est que moins il reste de nectar dans la fleur, plus il est difficile de l'aspirer. Par conséquent, il est plus profitable pour une abeille de ne pas butiner tout le nectar de chaque fleur pour avoir le temps d'aborder plus de fleurs. Les abeilles se trouvent alors face à un problème d'optimisation : quel est le temps de butinage optimal par fleur pour maximiser la quantité total de nectar butiné dans une journée ?

Pour rendre les choses plus précises, nous allons les modéliser sommairement. Tout d'abord, nous supposerons que notre abeille se trouve dans un champ rempli de fleurs toutes identiques et réparties de façon uniforme [3]. Ainsi, le temps de trajet entre une fleur et la suivante sera constant et nous le noterons τ. Pour une fleur donnée, la quantité de nectar que l'abeille est parvenue à aspirer au bout d'un temps t est donné par

$$N(t) = \frac{\beta t}{t + \alpha}.$$

On voit en faisant tendre t vers $+\infty$ que β est la quantité totale de nectar contenu dans une fleur. Quant à α, c'est une constante qui mesure la difficulté à aspirer le nectar. Voici (avec $\beta = 5$) le graphe de la fonction N pour $\alpha \in \{1, 2, 3, 4, 5\}$:

3. On voit qu'il faudrait ici introduire en général un cadre probabiliste (répartition des différentes fleurs, distribution de la quantité de nectar dans chaque fleur, ...). On parlera alors d'*optimisation stochastique*, un sujet que nous n'aurons malheureusement pas le temps d'aborder.



Il faut maintenant prendre en compte le trajet entre deux fleurs. Admettons que l'abeille butine pendant un temps total fixé dans la journée que nous noterons T, exprimé dans la même unité que τ, par exemple la seconde. Si elle décide de consacrer t seconde au butinage de chaque fleur, le nombre total n de fleurs visitées par jour sera

$$n = \frac{T}{t+\tau}.$$

Le nectar récolté par fleur étant $N(t)$, on conclut que la quantité totale de nectar ramassé dans une journée est donnée par la fonction

$$\boxed{f(t) = N(t)\frac{T}{t+\tau} = T\frac{\beta t}{(t+\alpha)(t+\tau)}.}$$

L'abeille doit donc résoudre le *problème d'optimisation* suivant : pour quelle valeur de t la fonction f est-elle maximale ?

On le sait, pour trouver le maximum d'une fonction il faut commencer par savoir où sa dérivée s'annule. Ici, on a

$$\begin{aligned} f'(t) &= T\frac{\beta(t+\tau)(t+\alpha) - \beta t(2t+\tau+\alpha)}{(t+\tau)^2(t+\alpha)^2} \\ &= T\frac{-\beta t^2 + \beta\alpha\tau}{(t+\tau)^2(t+\alpha)^2} \\ &= T\frac{\beta(\alpha\tau - t^2)}{(t+\tau)^2(t+\alpha)^2}. \end{aligned}$$

Comme on ne considère que des valeurs positives de t, cette fonction ne s'annule qu'en $t = \sqrt{\alpha\tau}$. On en déduit le tableau de variations suivant :

t	0	$\sqrt{\alpha\tau}$	$+\infty$
$f(t)$	0 ↗	$f_{\max}$	↘ 0

Nous avons ainsi résolu le problème de l'abeille et trouvé le temps optimal à passer sur chaque fleur pour que le rendement soit le meilleur possible, à savoir

$$\boxed{t_{\text{opt}} = \sqrt{\alpha\tau}}$$

Remarque 1.1.1. On remarquera que le temps optimal ne dépend pas de β. Ceci est assez logique, puisqu'on a supposé que ce nombre était le même pour toutes les fleurs.

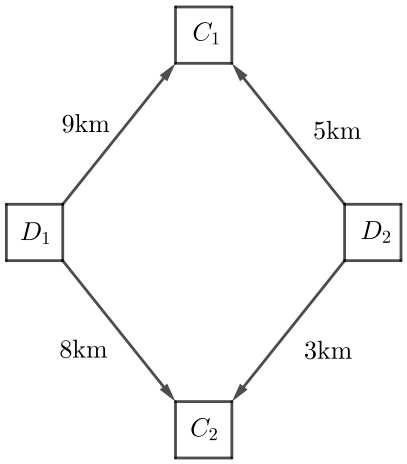
Ceci est évidemment un cadre simplifié. En particulier, nous ne considérons qu'une seule variable, à savoir le temps. Pour rendre le modèle plus réaliste, on pourrait faire varier les coefficients α et β en fonction de la fleur. Où alors, on pourrait prendre pour $N(t)$ une fonction plus compliquée. Pour ce qui nous concerne, peu importe. L'essentiel est de retenir les deux point suivants :

- De très nombreuses situations, que ce soit en biologie comme ci-dessus, en informatique (voir Section 1.2), en économie (voir Section 1.3) ou en physique (voir 1.4) peuvent se formuler comme des problèmes d'optimisation.
- Pour résoudre ces problèmes d'optimisation, on utilise la dérivation et des propriétés d'analyse réelle. Pour résoudre des problèmes d'optimisation faisant intervenir plusieurs variables, il faudra donc une généralisation de ces outils : c'est ce qu'on appelle le *calcul différentiel*.

Dans le Chapitre 2, nous généraliserons les techniques d'analyse réelle – comme la dérivation – aux fonctions de plusieurs variables afin d'avoir un moyen de localiser les extrema d'une fonction, comme nous l'avons fait ci-dessus. Cependant, avant de nous lancer dans cette aventure, nous allons donner quelques autres exemples de problèmes d'optimisation qui font, cette fois, intervenir plusieurs variables.

1.1.2 Le salaire de la peur

Voici un second exemple qui, quoique simpliste dans sa formulation, illustre bien le genre de questions que la logistique peut poser. Une entreprise effectuant de gros chantiers possède deux dépôts D_1 et D_2 où sont rangés ses camions, à raison de 8 camion dans D_1 et 6 camions dans D_2. Elle doit effectuer simultanément deux chantiers C_1 et C_2 situés à des distance des dépôts indiquées sur le schéma suivant :



Le chantier C_1 nécessite 4 camions tandis que le chantier C_2 en nécessite 7. Le problème de l'entreprise est d'affecter ses camions aux deux chantiers de façon à minimiser la distance totale parcourue, puisque celle-ci est proportionnelle aux dépenses (carburant, usure) engendrées. Autrement dit, si on note x_1 le nombre de camions envoyés de D_1 à C_1, x_2 le nombre de camions envoyés de D_1 à C_2, x_3 le nombre de camions envoyés de D_2 à C_1 et x_4 le nombre de camions envoyés de D_2 à C_2, on cherche à minimiser la fonction

$$f : (x_1, x_2, x_3, x_4) \mapsto 9x_1 + 8x_2 + 5x_3 + 3x_4.$$

Ainsi présenté, le problème d'optimisation n'a aucun intérêt. En effet, la fonction f n'a ni maximum ni minimum puisque par exemple

$$\begin{aligned} f(x_1, 0, 0, 0) &\underset{x_1 \to +\infty}{\longrightarrow} +\infty \\ f(x_1, 0, 0, 0) &\underset{x_1 \to -\infty}{\longrightarrow} -\infty \end{aligned}$$

Mais on ne veut pas minimiser la fonction f sur $\mathbf{R}^4$, on veut la minimiser sur une partie de $\mathbf{R}^4$ vérifiant les *contraintes* suivantes :

$$\begin{aligned} x_1, x_2, x_3, x_4 &\geqslant 0 \\ x_1 + x_2 &\leqslant 8 \\ x_3 + x_4 &\leqslant 6 \\ x_1 + x_3 &= 4 \\ x_2 + x_4 &= 7 \end{aligned}$$

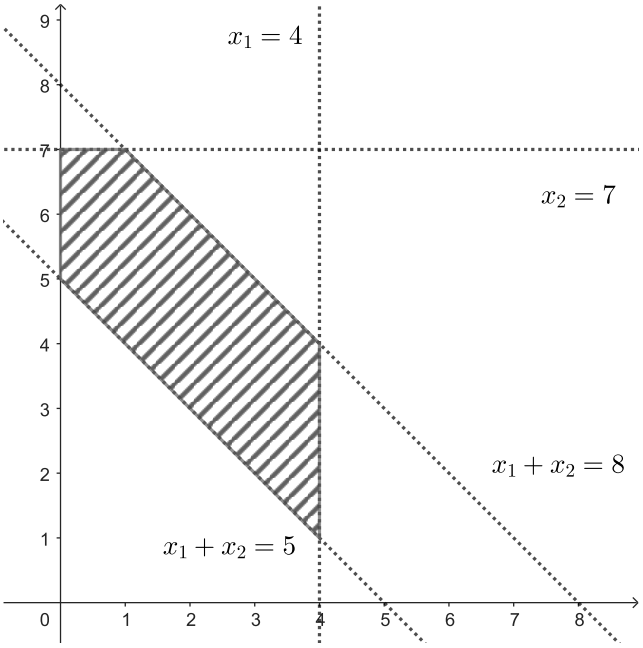
On peut remarquer que les deux dernières égalités permettent de réduire le nombre de variables du problème à deux, par exemple x_1 et x_2. On cherche alors à minimiser

$$F : (x_1, x_2) \mapsto 4x_1 + 5x_2 + 41$$

avec les contraintes

$$\begin{aligned} x_1, x_2 &\geqslant 0 \\ x_1 + x_2 &\leqslant 8 \\ -x_1 - x_2 &\leqslant -5 \\ x_1 &\leqslant 4 \\ x_2 &\leqslant 7 \end{aligned}$$

Les contraintes définissent une zone du plan qu'on peut dessiner :



Nous sommes ici face à un problème plus difficile que le précédent. Pas tant parce qu'il a deux variables, mais à cause des contraintes. En effet, rien ne sert a priori de chercher des points où les "dérivées partielles[4]" de F s'annulent, puisqu'elles sont constantes. Il faudra faire intervenir d'une façon ou d'une autre les contraintes pour résoudre ce problème. Ceci fera l'objet du chapitre 3, mais nous pouvons déjà faire quelques remarques importantes :

- Il faudrait déjà savoir si le minimum que nous cherchons existe bien. C'est le cas pour une raison topologique : la zone hachurée délimitée par les contraintes est un *compact* de $\mathbf{R}^2$. Comme F est continue, elle admet un maximum et un minimum sur ce compact.

- Si le minimum était atteint à l'*intérieur* de la zone hachurée, alors on aurait un minimum local. Il est facile de se convaincre que comme F est affine, elle n'a pas de tel minimum local. Ainsi, le point que nous recherchons est sur le *bord*.

- Le bord peut se décomposer en cinq segments. Sur chacun de ces segments, soit x_1 est constante, soit on peut exprimer x_2 en fonction de x_1, de sorte qu'on se ramène à chercher à chaque fois le minimum d'une fonction d'une variable, ce qui peut se faire à l'aide d'un tableau de variations.

- Il suffit ensuite de prendre le minimum de ces cinq minima pour résoudre le problème.

En appliquant la méthode précédente (ce que nous ne ferons pas explicitement ici car c'est un peu pédestre, une autre approche sera présentée au Paragraphe 3.3.2), on trouve $x_1 = 4$ et $x_2 = 1$, ce qui donne $x_3 = 0$ et $x_4 = 6$. La distance totale parcourue par les camions est alors

$$f(4, 1, 0, 6) = 62\text{km}.$$

4. Quel que soit le sens de cette expression, que nous définirons bientôt (voir Paragraphe 2.2.1).

Remarquons que, de façon peu intuitive, cette solution optimale fait emprunter le trajet le plus long à 4 camions. Notons aussi que si le problème peut être résolu à la main – bien que nous ne l'ayons pas fait – c'est parce qu'il a une forme relativement simple : la fonction à minimiser est affine. On parle dans ce cas d'*optimisation linéaire* ou parfois de *programmation linéaire*[5]. La première approche générale de l'optimisation linéaire a été développée par Kantorovich[6] en 1939 dans le cadre de travaux en économie. Ces travaux lui valurent le Prix de la Banque de Suède en sciences économiques en mémoire d'Alfred Nobel – généralement appelé improprement Prix Nobel d'économie – en 1975, partagé avec Koopmans[7] qui avait travaillé sur le même sujet de façon indépendante à la même époque. Nous aborderons quelques aspects de l'optimisation linéaire au Paragraphe 3.3.

1.2 Les petites cellules grises

L'homme, on a dit qu'il était fait de cellules et de sang.
Mais en réalité, il est comme un feuillage [...]
il faut que le vent passe pour que ça chante.

J. Giono, Que ma joie demeure (1935)

Qu'est-ce qu'une neurone ? Du point de vue biologique, un objet compliqué que nous ne tenterons pas de décrire, mais qui a deux caractéristiques importantes :

- Un neurone peut recevoir des signaux sous forme d'impulsion électrique.
- Un neurone est une cellule *excitable* : sous certaines conditions, elle émettra elle-même des impulsions électriques en réponse au stimulus.

Ces deux aspects ont conduit les informaticiens à modéliser un neurone comme une composée de deux fonctions. Plus précisément, un *neurone* (parfois appelé *perceptron*) est un objet prenant en entrée un vecteur à n coordonnées et produisant en sortie un nombre selon la procédure suivante :

1. Le vecteur $x = (x_1, \cdots, x_n)$ est transformé en un nombre

$$\overline{x} = a_0 + \sum_{i=1}^{n} a_i x_i.$$

   Ici, les coefficients $(a_i)_{1 \leqslant i \leqslant n}$ ne dépendent que du neurone et a_0 est appelé le *biais*. Cette opération est la *partie affine*.

5. Ce terme n'a rien à voir avec l'informatique mais traduit le fait qu'on cherche à résoudre ce qu'on appelle parfois un *programme d'optimisation linéaire*. Cette terminologie est très usitée en anglais.

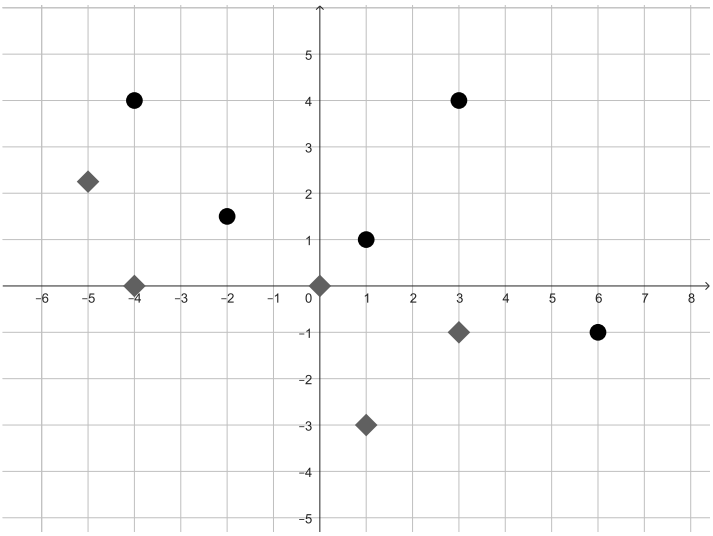
6. Leonid Kantorovich (1912–1986) : Mathématicien et économiste russe, qui a contribué de façon importante à l'analyse fonctionnelle et à la théorie de l'optimisation. Ses travaux n'étaient pas seulement théoriques, puisqu'il a joué un rôle important dans la planification de l'économie russe au début de l'ère socialiste, ainsi que dans la logistique militaire pendant la seconde guerre mondiale, et notamment au siège de Leningrad.

7. Tjalling Koopmans (1910–1985) : Mathématicien et économiste néerlando-américain qui s'est beaucoup intéressé à l'optimisation en économie, appliquée à l'établissement des prix et l'allocation des ressources.

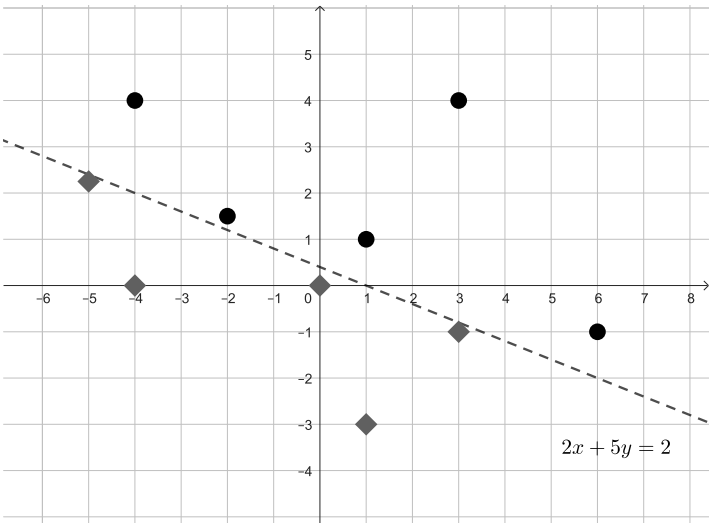
2. Le résultat produit par la partie affine est ensuite utilisé comme argument d'une *fonction d'activation* φ. Idéalement, φ est la fonction de Heaviside [8] – aussi connue sous le nom de "fonction créneau" – c'est-à-dire que

$$\varphi(\overline{x}) = \begin{cases} 1 & \text{si} \quad \overline{x} \geqslant 0 \\ 0 & \text{si} \quad \overline{x} < 0 \end{cases}$$

Un tel neurone peut être utilisé comme une sorte de "porte logique" pour effectuer une tâche simple. Par exemple, imaginons que nous voulions fabriquer un neurone qui permette de distinguer les ronds des carrés dans la grille ci-dessous.



On remarque que la droite d'équation $2x + 5y = 2$ sépare le plan en deux demi-plans qui ne contiennent chacun qu'un type de point :



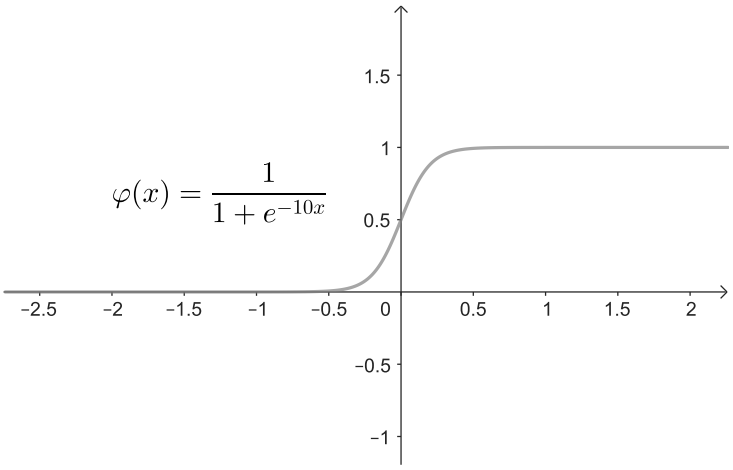
8. Oliver Heaviside (1850–1925) : Physicien et mathématicien autodidacte, il s'est beaucoup intéressé à l'étude mathématique des circuits électriques. Outre la définition de la fonction qui porte son nom, on lui doit l'idée d'utiliser les nombres complexes (notamment l'impédance) pour l'étude des circuits électriques. C'est aussi lui qui a donné aux Équations de Maxwell leur forme la plus courante en utilisant des opérateurs vectoriels (Maxwell avait encore huit équations au lieu des quatre couramment enseignées aujourd'hui).

Or, appartenir à un demi-plan, c'est vérifier une inégalité du type $f(x,y) \geqslant 0$ avec f une fonction affine. Ainsi, si la partie affine du neurone est

$$(x_1, x_2) \mapsto 2x_1 + 5x_2 - 2,$$

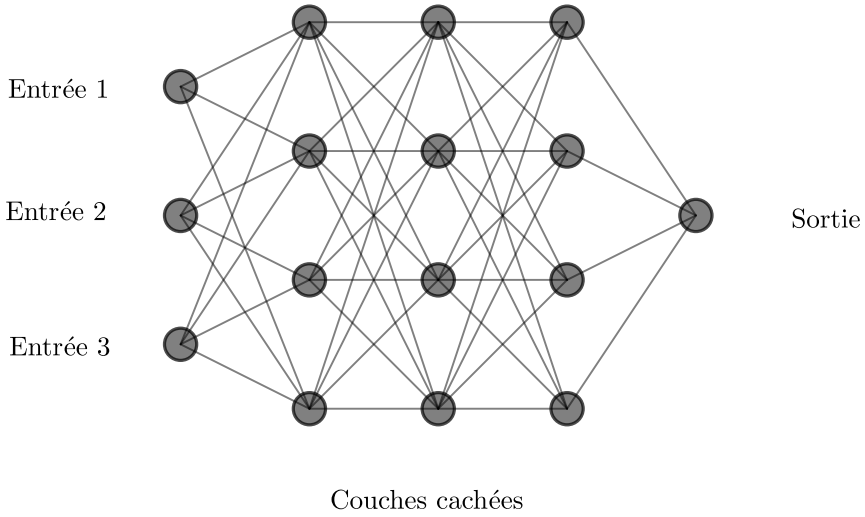
alors si on donne au neurone les coordonnées d'un point, il renverra 1 si le point est un rond et 0 si c'est un carré.

Mais en pratique il peut être utile d'avoir à la place de la fonction de Heaviside une fonction plus régulière qui "ressemble" à un créneau. On parle alors de *fonction sigmoïde*, et en voici un exemple :



Cela permet de modéliser une marge d'erreur du neurone, qui au lieu de répondre Oui ou Non à une question peu par exemple répondre Oui à 80%.

Bien qu'un neurone soit un objet intéressant, ses possibilités sont limitées. Pour améliorer le concept, il faut alors combiner plusieurs neurones : c'est un *réseau de neurones*. Un tel réseau est habituellement organisé en un nombre fini m de couches, la couche numéro k comportant n_k neurones dont les entrées sont donnée par des vecteurs formés par les sorties de la couche précédente. Schématiquement, cela ressemble au dessin suivant :



Si $x(0)$ désigne le vecteur d'entrée de la première couche, alors le vecteur de sortie de la couche k est

$$x(k) = \varphi \circ F_k\,(x(k-1))$$

où F_k est une fonction affine donnée par les coefficients des neurones de la couche k.

Nous arrivons maintenant au cœur du problème. On souhaite configurer les coefficients des neurones afin que le réseau puisse accomplir une tâche prescrite, par exemple reconnaître un motif sur des images. Pour cela, on *entraîne* le réseau en lui donnant des vecteurs d'entrée $x(0)_1, \cdots, x(0)_\ell$ où ℓ est le nombre d'expériences. Les résultats sont des vecteurs $x(m)_1, \cdots, x(m)_\ell$ qu'on veut comparer avec les "bonnes" réponses attendues $y_1, \cdots, y_\ell$. On doit donc résoudre le problème d'optimisation consistant à minimiser la fonction [9]

$$\sum_{i=1}^{\ell} \left\| y_i - x(m)_i \right\|^2.$$

Tel qu'il est écrit ci-dessus, le problème n'a pas grand sens. Pour qu'il en ait, il faut rajouter des *contraintes*, à savoir

$$\begin{aligned} x(1)_i &= \varphi \circ F_1\,(x(0)_i) \\ &\vdots \\ x(m)_i &= \varphi \circ F_m\,(x(m-1)_i) \end{aligned}$$

pour tout $1 \leqslant i \leqslant \ell$.

Nous avons donc à nouveau un problème d'optimisation sous contraintes. De plus, la fonction à optimiser a ici une forme particulièrement simple puisqu'elle est donnée par une norme au carré. On parle alors de *problème quadratique*. L'avantage de ces problèmes est qu'ils sont des cas particuliers bien compris d'une famille plus générale : les problèmes d'optimisation *convexes*. On dispose donc d'outils plus développés et plus puissants pour les attaquer. En particulier, il existe des algorithmes performants pour

9. Dans tout ce texte, la notation $\|\cdot\|$ désigne la norme euclidienne sur $\mathbf{R}^n$.

approcher numériquement le minimum qui nous intéresse. Nous aborderons le thème de la convexité et de son utilité en optimisation au Chapitre 4. Précisons néanmoins, pour être tout à fait honnêtes, que l'entraînement d'un réseau de neurones n'est pas en général un problème convexe, parce que les contraintes ne le sont pas.

1.3 Le juste prix

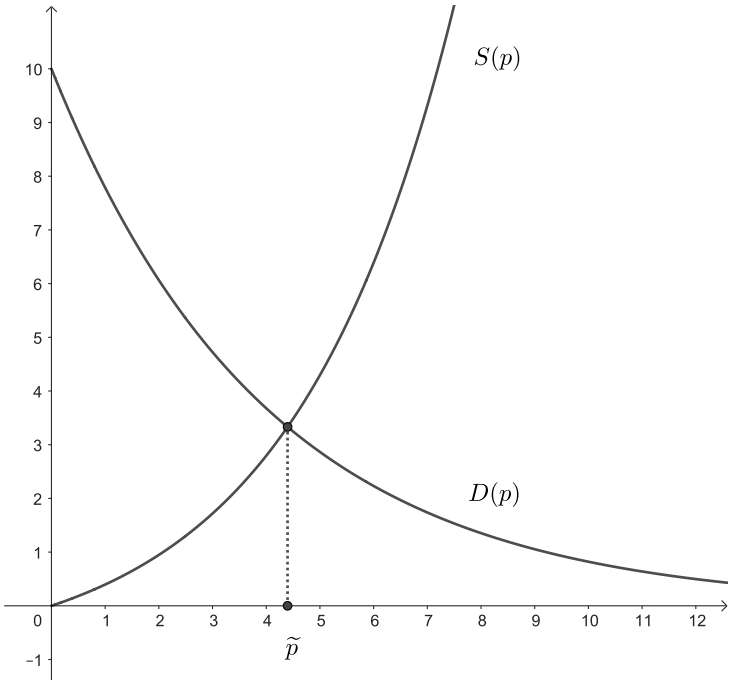
Les politiques grecs ne reconnaissent d'autre force que celle de la vertu.
Ceux d'aujourd'hui ne vous parlent que de manufactures,
de commerce, de finances, de richesses et de luxe même.

Montesquieu, De l'esprit des lois

Nous allons maintenant présenter un autre type de problème d'optimisation, cette fois dans le domaine de l'économie. L'un des problèmes fondamentaux de l'économétrie est celle de l'équilibre entre l'offre et la demande. Pour simplifier, considérons un marché isolé, c'est-à-dire un ensemble d'agents qui vendent ou achètent un unique bien. Il y a une certaine quantité de biens disponibles – c'est l'offre – et les acheteurs désirent acquérir une autre quantité du même bien – c'est la demande. Ces deux quantités dépendent du prix du bien : s'il est très cher, il y aura moins de demande et donc beaucoup d'offre puisque les stocks s'écouleront mal. On modélise le problème par deux fonctions, l'une décrivant l'offre [10] $S(p)$ en fonction du prix, et l'autre décrivant la demande $D(p)$. Un *prix d'équilibre* est une valeur $\widetilde{p}$ qui vérifie l'égalité

$$S(\widetilde{p}) = D(\widetilde{p}).$$

Dans notre exemple, les fonctions S et D ont souvent la forme ci-dessous, qui rend l'existence d'un prix d'équilibre assez intuitive.



10. Notée S à cause du mot *supply* en anglais, mais aussi parce que O est une notation peu pratique.

Le problème de l'équilibre global consiste à étendre les considérations précédentes à des marchés plus complexes sur lesquels plusieurs biens sont échangés. Plus précisément, on considère un marché global constitué d'un nombre fini n de *commodités*, qui peuvent être de deux natures : les biens ou les services (par exemple du travail). À chaque agent du système correspond un vecteur $x \in \mathbf{R}^n$, son *panier*, dont les coordonnées correspondent à la quantité de chaque bien qu'il possède. Si une coordonnée correspond à un service, alors elle est négative. À la i-ème commodité est associé un prix p_i, de sorte que le prix total d'un élément $x \in \mathbf{R}^n$ est

$$\sum_{i=1}^{n} p_i x_i = \langle p, x \rangle,$$

où p est le *vecteur des prix* de coordonnées $(p_i)_{1 \leqslant i \leqslant n}$.

Pour échanger des biens, il faut les produire. On considère donc des *unités de productions* qui sont chacune décrites par une partie $Y \subset \mathbf{R}^n$. Chaque élément de Y représente un *plan de production* : les coordonnées positives sont les biens qu'on produit et les composantes négatives sont ceux que l'on consomme. Les hypothèse générales sur le fonctionnement économique du système se traduisent en conditions sur les unités de production, à savoir : si $Y_1, \cdots, Y_m$ désignent les m unités de production du marché, alors

- Si $y \in Y_j$ et $t \in [0;1]$, alors $ty \in Y_j$, pour tout $1 \leqslant j \leqslant m$. Cela signifie qu'on peut réaliser des *changements d'échelles décroissants des plans de production* : si on divise tous les moyens, on produit alors les mêmes biens mais en quantité divisée par le même facteur. Cela signifie aussi que $0 \in Y_j$ de sorte qu'il est possible de ne rien produire.

- Pour tout $1 \leqslant j \leqslant m$, si $y, y' \in Y_j$, alors $y + y' \in Y_j$. En effet, si l'on peut produire certains biens en consommant certaines ressources, alors en consommant la somme des ressources on peut produire la somme des biens. On remarquera que cette hypothèse peut poser questions : n'y a-t-il pas de limites absolues aux quantités qu'on peut produire (capacité de production des usines, capacité de stockage, disponibilité des matières premières) ?

- Soit
$$Y = \sum_{i=1}^{m} Y_j = \left\{ y_1 + \cdots + y_m \mid y_j \in Y_j \text{ pour tout } 1 \leqslant j \leqslant m \right\},$$
alors Y ne contient aucun élément dont toutes les composantes soit strictement positives. En effet, on pourrait alors produire sans rien consommer.

- On a $Y \cap (-Y) = \{0\}$. Cela signifie que l'économie n'est pas réversible : s'il est possible de produire globalement des commodités en en consommant d'autres, il n'est pas possible de faire exactement l'inverse.

Remarque 1.3.1. Les deux premières hypothèses peuvent se résumer dans l'assertion suivante : Y_j est un *cône convexe de sommet* 0.

Mais pour que le marché fonctionne, il faut surtout des *consommateurs*. Chacun est décrit par une partie $X \subset \mathbf{R}^n$ qui représente les souhaits du consommateur, c'est-à-dire les paniers qu'il aimerait détenir. Pour comparer ces différents souhaits, on dispose d'une *fonction d'utilité*

$$u : X \to \mathbf{R}_+$$

telle que $u(x') > u(x)$ si le consommateur préfère le panier x' au panier x. Ces parties satisfont certaines hypothèses :

- X est convexe, fermée et minorée au sens où il existe x_0 tel que $x \geqslant x_0$ (au sens où l'inégalité est vérifiée pour chaque coordonnée) pour tout $x \in X$. Cette dernière condition signifie que tout le monde a un minimum d'exigences.
- Pour tout $x \in X$, il existe $x' \in X$ tel que $u(x') > u(x)$. Autrement dit, on n'en veut toujours plus : il n'y a pas de satiété du consommateur.
- La fonction u est convexe (voir le Chapitre 4).

Enfin, pour éviter que la monnaie ne s'épuise, il faut que les unités de production profitent d'une manière ou d'une autre aux consommateurs. Ceci est incarné par des nombres $\alpha_{X,Y} \in [0;1]$ représentant la fraction de *dividendes* de l'unité de production Y que touche le consommateur X. On doit bien sûr avoir

$$\sum_X \alpha_{X,Y} = 1$$

pour toute unité de production Y. Il est important de remarquer que ce cadre signifie que nous sommes dans une économie fictive entièrement capitaliste : les moyens de productions appartiennent entièrement aux agents qui en touchent donc des dividendes.

On considère un marché global avec des unités de production $Y_1, \cdots, Y_m$ et des consommateurs $(X_1, u_1), \cdots, (X_\ell, u_\ell)$ vérifiant toutes les hypothèses ci-dessus. Un état du marché est donné par

$$(x_1, \cdots, x_\ell, y_1, \cdots, y_m, p),$$

où les $\ell + m$ premières coordonnées désignent l'état des acteurs et $p \in \mathbf{R}_+^n$ désigne l'ensemble des prix des commodités. Pour pouvoir lancer la machine économique, il faut que les consommateurs puissent initialement s'appauvrir. On suppose que pour chaque $1 \leqslant j \leqslant \ell$, le consommateur correspondant démarre avec un panier ζ_i pour lequel il existe $x \in X_i$ vérifiant $x_i < \zeta_i$.

L'existence d'un prix pour lequel l'ensemble du marché est à l'équilibre est un sujet de réflexion pour les économistes depuis que le cadre précédent a été pour la première fois posé (quoique pas en ces termes) par Walras [11] au XIX^e siècle. Cependant, la rigueur mathématique n'était pas toujours au rendez-vous : Walras considérait par exemple qu'un système comportant autant d'équations que d'inconnues possède nécessairement une unique solution. Or, nous savons que ceci n'est même pas vrai quand le système est linéaire ... L'un de résultats les plus importants dans le domaine fut obtenu en par Arrow [12] et Debreu [13].

11. Léon Walras (1834–1910) : économiste français, il est l'un des premiers à s'intéresser aux problèmes d'équilibre en économie et à tenter de les formaliser, sans pour autant être en mesure de déterminer leur existence. On lui doit aussi une métaphore qui a fait florès en économie, le *tâtonnement Walrassien*.

12. Kenneth J. Arrow (1921–2017) : économiste américain, considéré comme l'un des fondateurs de l'*école néoclassique*. Outre ses travaux sur la théorie de l'équilibre général, il a également donné une version rigoureuse et générale du Paradoxe de Condorcet, aujourd'hui connu sous le nome de Théorème d'impossibilité d'Arrow.

13. Gérard Debreu (1921–2004) : mathématicien et économiste français. Il a apporté ses compétences mathématiques à plusieurs domaines de l'analyse économique et a contribué à les développer en important les techniques modernes, notamment d'analyse convexe.

Théorème 1.3.2 (Arrow-Debreu) Sous les hypothèses précédentes, il existe un prix $\widetilde{p} \in \mathbf{R}^n$ tel que

- Pour tout $1 \leqslant j \leqslant m$, la fonction
$$y \in Y_j \mapsto \langle \widetilde{p}, y \rangle$$
admet un maximum en un point $\widetilde{y_j}$.
- Pour tout $1 \leqslant i \leqslant \ell$ la fonction u_j admet un maximum en un point $\widetilde{x_j}$ sous la contrainte
$$\langle \widetilde{p}, x \rangle \leqslant \langle \widetilde{p}, \zeta_i \rangle + \sum_{j=1}^{m} \alpha_{ij} \langle \widetilde{p}, \widetilde{y_j} \rangle$$
- L'état du marché satisfait l'*équilibre de l'offre et de la demande* :
$$\sum_{i=1}^{\ell} x_i - \sum_{j=1}^{m} y_j - \sum_{i=1}^{\ell} \zeta_i \leqslant 0.$$
- L'état du marché satisfait la *Loi de Walras*
$$\sum_{i=1}^{\ell} \langle p, x_i \rangle - \sum_{j=1}^{m} \langle p, y_j \rangle - \sum_{i=1}^{\ell} \langle p, \zeta_i \rangle = 0.$$

Remarque 1.3.3. La Loi de Walras exprime une contrainte budgétaire globale du marché et est un aspect essentiel de cette approche du problème. En particulier, elle signifie que les prix des commodités ne dépendent que du marché lui-même, et pas d'autres biens auxquels les agents n'auraient pas accès.

La preuve de ce résultat d'optimisation est loin d'être facile [14] et sort du cadre de ce cours (on pourra par exemple consulter [3] pour un exposé complet). D'ailleurs, on remarquera qu'aucune hypothèse de régularité n'a été faite sur les fonctions, ce qui prévient a priori tout recourt au calcul différentiel. Ceci n'est toutefois qu'apparent, et on peut démontrer ce théorème et bien d'autres du même type en utilisant le calcul différentiel, comme illustré dans [11].

Notons enfin que notre description du marché fait appel à une hypothèse très importante dans les problèmes d'optimisation, à savoir la convexité. Le fait que les fonctions d'utilités u_i soient convexes est un point essentiel et la preuve ne fonctionne plus sans elle ! La justification micro-économique de cette hypothèse est donc essentielle pour assurer sa pertinence. Profitons-en pour ajouter que l'étude des fonctions convexes est un domaine riche et important des mathématiques. En fait, beaucoup de résultat d'optimisation que nous allons voir sont encore vrais pour des fonctions qui ne sont pas différentiables, pour peu qu'elles soient convexes. L'un des raisons principales de ce phénomène est que les fonctions convexes sont en sens toujours "presque différentiables".

14. Elle a valu le Prix de la Banque de Suède en sciences économiques en mémoire d'Alfred Nobel à ses auteurs.

Remarque 1.3.4. On pourra aussi remarquer que toutes les équations intervenant dans le modèle ci-dessus sont *homogènes de degré* 1, c'est-à-dire que si on multiplie toutes les variables par un scalaire λ, alors les relations sont multipliées par λ dans chaque membre. En particulier, cela veut dire qu'on peut toujours décider de normaliser une partie des variables. En général, on suppose donc que la somme des prix vaut 1.

1.4 Thème et variations

Le monde n'est qu'une branloire pérenne.
Toutes choses y branlent sans cesse : la terre, les rochers du Caucase, les pyramides d'Égypte, & du branle public & du leur.
Montaigne, Essais

Nous avons mentionné plusieurs problèmes d'optimisation, et tenté de suggérer la façon dont les outils connus pour les fonctions d'une variable pourraient être étendus pour les étudier. Ce n'est pas la seule méthode, comme nous l'avons évoqué pour le Théorème 1.3.2, mais c'en est une redoutablement efficace quand elle s'applique. Avant de la décrire dans les chapitres qui suivent, nous aimerions revenir sur ce *calcul différentiel* en donnant quelques éléments de son histoire.

1.4.1 Une brève histoire du calcul différentiel

Il est difficile de distinguer dans l'histoire des mathématiques l'origine d'une idée ou d'une théorie particulière parmi toutes les tentatives plus ou moins fructueuses. En ce qui concerne le calcul différentiel, on peut néanmoins mentionner un problème mathématique proche qui a intéressé les mathématiciens pendant longtemps : le *problème de la quadrature*. Il s'agit de calculer l'aire d'une partie du plan délimitée par une ou plusieurs courbes, c'est-à-dire de l'exprimer en termes de "carrés" dont le côté fait une unité de longueur, d'où le nom. En langage moderne, cela revient essentiellement à calculer des intégrales. Pour ce faire, on est assez naturellement mené à considérer des limites ou des quantités "infiniment petites", comme le fameux symbole "dx" de l'intégration le rappelle. On voit donc que le problème de la quadrature mène à des questions qui sont au fondement du calcul différentiel.

De nombreuses techniques ont été élaborées pour aborder le problème de la quadrature. Citons en particulier la *Méthode d'exhaustion* qui consiste à encadrer la partie du plan à laquelle on s'intéresse par des polygones dont on sait calculer l'aire. Due peut-être à Eudoxe [15], la méthode a été reprise et raffinée par Archimède [16] qui en fit un grand usage, d'une part pour la quadrature de la parabole, et d'autre part pour calculer des approximations du nombre π. Cette méthode a également un intérêt théorique puisque c'est grâce à elle qu'Euclide [17] démontre que la circonférence du cercle

15. Eudoxe de Cnide (408 – 355) : mathématicien, astronome, médecin et philosophe grec. Il a été disciple de Platon, et s'est beaucoup intéressé avec ce dernier au mouvement des astres. Sa recherche d'une description l'a mené à la théorie dite des *sphères homocentriques* qui sera reprise jusqu'aux travaux de Ptolémée. Il est également le premier à calculer qu'une année dure 365, 25 jours.

16. Archimède (287–212) : l'un des plus grands scientifiques de l'antiquité, dont la vie est malheureusement peu connue. Ses travaux en géométrie sont fondateurs sur de nombreux sujets : calculs de longueurs, d'aires et de volumes, courbes paramétrées, coniques. Il a également laissé une œuvre considérable en physique (statique des fluides, balances, mécanique).

17. Euclide (vers 300) : on ne sait presque rien de sa vie. Il a transmis à la postérité une fraction de son ouvrage monumental, les Éléments, qui sont souvent considérés comme l'un de premiers traités de mathématiques. De nombreux sujets y sont abordés, de l'arithmétique à la géométrie.

est proportionnelle au diamètre et qu'Archimède démontre que l'aire du disque est proportionnelle au carré du diamètre.

Un second problème mathématique qui a passionné les mathématiciens au tournant du XVII^e siècle est le *problème des tangentes*. Il s'agit, étant donnée une courbe, de déterminer ses tangentes, c'est-à-dire d'en trouver une description géométrique ou une équation. En langage moderne, ceci consiste essentiellement à calculer le nombre dérivé d'une fonction. Plusieurs mathématiciens développèrent des méthodes plus ou moins rigoureuses et plus ou moins justifiées pour ce faire. En particulier, Fermat [18] utilise une méthode analytique inspirée de ses travaux sur la détermination des maxima et des minima : le calcul différentiel naît ici de l'optimisation ! Dans cette méthode dite d'*adégalisation* (terme qu'il emprunte à Diophante [19]), on part de $f(x+e)$, auquel on retranche ensuite $f(x)$. On observe (par exemple dans le cas d'une fonction polynomiale) qu'on parvient alors à factoriser e. En le simplifiant et en posant $e = 0$, on obtient la pente de la tangente ! Comme on le voit, ceci revient à calculer la limite du taux d'accroissement dans le cas où ce dernier peut se simplifier et donc se calculer directement.

Mais la concurrence est rude à l'époque et Descartes [20] engage une polémique avec Fermat sur la paternité de la méthode. Vers la même époque, un autre mathématicien, Barrow [21], fait une contribution importante en remarquant par des raisonnements géométriques que le problème des tangentes et le problème de la quadrature sont inverses l'un de l'autre, mais sans parvenir à en tirer pleinement parti.

C'est finalement avec une nouvelle polémique entre deux autres grandes figures que se jouera l'élaboration du calcul différentiel tel que nous le connaissons aujourd'hui. En effet, dans la même période, Newton [22] et Leibniz [23] développent un cadre théorique qui permet de réorganiser tous les résultats précédents et de donner des méthodes algorithmiques générales et efficaces pour calculer les tangentes ou les aires, tout en mettant en évidence la réciprocité des deux opérations. La querelle de priorité entre les deux personnages, dans laquelle se sont engagés nombres de scientifiques de

18. Pierre de Fermat (1601–1665) : magistrat, scientifique et homme de lettres français qui a beaucoup contribué aux mathématiques et à la physique. Outre ses travaux en optique et son "petit" théorème d'arithmétique, on lui doit une *formule des tangentes* avant le développement du calcul différentiel, ainsi qu'un "grand" théorème qui ne fut démontré que trois siècles après sa mort.

19. Diophante d'Alexandrie (entre le I^e et le IV^e siècle avant notre ère) : on ne sait pratiquement rien de la vie de ce mathématicien grec, si ce n'est qu'il a vécu à Alexandrie. On ne possède que des fragments de son œuvre, notamment l'Arithmétique qui a eu une grande influence sur les mathématiques arabes et occidentales, mais dont les marges étaient trop étroites pour que Fermat y écrivît la démonstration de son célèbre théorème.

20. René Descartes (1596–1650) : philosophe, mathématicien et physicien français. Il a introduit les notations qui portent son nom pour repérer les points dans le plan ou dans l'espace, et a publié des ouvrages de mécanique et d'optique. Son œuvre philosophique a eu une postérité au moins égale à son œuvre mathématique.

21. Isaac Barrow (1630–1677) : philologue, mathématicien et théologien anglais. Outre ses travaux précurseurs pour le calcul différentiel, il est connu pour avoir été le professeur de Newton puisqu'il occupait la prestigieuse *Lucasian chair* de l'Université de Cambridge. On lui doit également une édition des éléments d'Euclide.

22. Isaac Newton (1642–1727) : on ne présente plus l'un des physiciens et mathématiciens majeurs de l'histoire européenne. Ses Principa Mathematica Philosophiæ Naturalis eurent un retentissement important pour l'exposé qu'elles donnent d'une théorie complète de la gravitation et des lois fondamentales de la mécanique.

23. Gottfried-Wilhelm Leibniz (1646–1716) : philosophe et scientifique qui fut une figure intellectuelle marquante du XVII^e siècle par ses contributions importantes dans des domaines très variés des sciences : philosophie, philologie, histoire, droit, logique, mathématiques. Il a introduit des notations qui ont perduré jusqu'à aujourd'hui, notamment les symboles $\int$ et $\mathrm{d}x$.

l'époque, est un problème historique complexe que nous n'aborderons pas. Disons simplement que du point de vue du vocabulaire et des notations, c'est finalement Leibniz qui est passé à la postérité.

1.4.2 Du calcul différentiel à l'optimisation

Nous avons présenté quelques étapes de l'histoire du calcul différentiel au regard des deux problèmes mathématiques de la quadrature et des tangentes, mais il ne s'agit que de l'une des facettes du sujet. En effet, pour établir rigoureusement le calcul différentiel, il a fallu ensuite une bonne notion de limite, qui manquait toujours chez Newton et Leibniz. Cette nécessité n'était pas que mathématique, mais également physique.

En effet, la mécanique moderne décrit le mouvement des objets à partir de quantités *instantanées* comme l'accélération ou la vitesse. Si l'on étudie par exemple le mouvement d'un point matériel se déplaçant au cours du temps, il est facile de définir sa vitesse moyenne sur un intervalle de temps $[t, t+\delta t]$: elle sera donnée par le rapport entre la distance parcourue et la durée de l'intervalle

$$v = \frac{x(t+\delta t) - x(t)}{\delta t}.$$

Sous cette forme, on voit bien qu'il est naturel de chercher à comprendre ce qui se passe lorsque la durée δt de l'intervalle devient "arbitrairement" ou "infiniment" petite. Décrire rigoureusement ce phénomène revient à définir le nombre dérivée, et donc le calcul différentiel. Cette approche est notamment celle de Newton, à l'aide de laquelle il formule ses célèbres lois qui permettent de décrire entièrement la mécanique classique. Ceci mène naturellement à décrire les phénomènes physiques par des équations différentielles, comme la Troisième Loi de Newton :

$$mx''(t) = F.$$

Ce n'est pas la seule approche possible. En effet, Fermat propose par exemple en optique le Principe de moindre temps, stipulant que la lumière suit entre deux points le trajet le plus rapide. Il s'agit a priori d'une approche très différente de l'approche géométrique usuelle, mais elle permet de retrouver toutes les lois de l'optique géométrique (voir le Paragraphe 2.2.2 pour un exemple). Ce qui nous intéresse ici, c'est que le problème mathématique sous-jacent est un problème d'optimisation.

Une approche similaire est possible en mécanique. Maupertuis a ainsi proposé un Principe de moindre action : on peut associer à tout système mécanique une fonction appelée *action*, de sorte que la trajectoire du système soit celle qui minimise l'action. Introduisons quelques notations pour rendre ceci plus précis. On considère par exemple un point matériel dont la position à un instant donné est $x(t) \in \mathbf{R}^3$. On suppose qu'on connait la position du point l'instant t_0, qu'on note x_0, et la position à un instant $t_1 > t_0$, que l'on note x_1. Si le point est soumis à une force dérivant d'un potentiel $V : \mathbf{R}^3 \to \mathbf{R}$, alors l'action est

$$S = \int_{t_0}^{t_1} \frac{1}{2} m \|x'(t)\|^2 - V(x(t)) \mathrm{d}t.$$

Ainsi écrite, on ne voit pas de quoi l'action S peut être fonction. Elle l'est en fait de la *trajectoire*. Si $\gamma : [t_0; t_1] \mapsto \mathbf{R}^3$ est une fonction dérivable telle que $\gamma(t_0) = x_0$ et $\gamma(t_1) = x_1$, alors l'action correspondante est

$$S(\gamma) = \int_{t_0}^{t_1} \frac{1}{2} m \|\gamma'(t)\|^2 - V(\gamma(t)) \mathrm{d}t.$$

Le Principe de moindre action affirme alors que la trajectoire réelle du point sera donnée par la courbe γ qui minimise cette action [24].

La formulation via l'action permet de remplacer la formulation à partir d'équations différentielles en une formulation d'optimisation. Cela dit, ce problème d'optimisation n'est pas de la même forme que ceux que nous avons présenté précédemment. En effet, la variable sur laquelle on chercher à optimiser la fonction n'est pas un vecteur d'un espace vectoriel, mais une courbe. Ceci signifie que les outils que nous allons développer ci-après ne s'appliquent pas dans ce cadre. Il s'agit d'une théorie différente, qu'on appelle parfois *calcul des variations*. Même si nous ne l'aborderons pas, il est important de signaler qu'il s'agit d'une approche centrale dans la physique moderne, par exemple en relativité générale avec l'*action d'Einstein-Hilbert*. Il est également possible de transformer le Principe de moindre action en optimisation d'une autre fonctionnelle appelée le *Lagrangien* du système. Cette dernière formulation admet une forme de dualité avec une autre fonctionnelle appelée *Hamiltionnien*. L'étude de ces formulations est parfois connue sous le nom de *mécanique analytique*.

24. Cette formulation historique est en fait erronée pour plusieurs raisons. La première est qu'il n'y a pas forcément unicité de l'extremum. La seconde est que la trajectoire ne correspond pas nécessairement à un extremum global, mais seulement à un point critique de l'action.

CHAPITRE 2

À LA RECHERCHE DES EXTREMA

Trouver n'est rien. Le difficile est de s'ajouter ce que l'on trouve.

P. Valéry, Monsieur Teste

L'objectif de ce chapitre est simple : trouver les extrema d'une fonction de plusieurs variables. Pour ce faire, nous allons procéder en trois étapes, afin de répondre aux trois questions suivantes :

- Comment savoir qu'une fonction a un extremum ?
- Si elle en a, comment les localiser ?
- Si on a localisé un extremum, comment identifier sa nature (minimum ou maximum) ?

Comme nous allons le voir, pour répondre à ces questions nous allons devoir développer toute une théorie généralisant l'analyse réelle : c'est cette théorie qu'on appelle le *calcul différentiel*.

2.1 Existence

L'existence précède l'essence.

J.-P. Sartre, L'existentialisme est un humanisme.

Comme expliqué ci-dessus, nous commençons par la question la plus terre à terre : comment peut-on être sûr qu'une fonction donnée admet un ou plusieurs extrema ? Curieusement, le problème n'a rien à voir avec le calcul différentiel, mais plutôt avec la topologie ... un terme qui peut évoquer des souvenirs contrastés. Nous n'allons pas vraiment faire de rappels détaillés de topologie ici, mais simplement redonner les définitions et résultats fondamentaux dont nous aurons besoin dans la suite.

2.1.1 Souvenirs, souvenirs

Fonctions à valeurs vectorielles

La grande nouveauté dans ce texte par rapport à ce qui a été vu les années passées est de considérer des fonctions ayant plusieurs variables. Mais cela nous poussera aussi à considérer des fonctions dont l'espace d'arrivée est une partie d'un espace vectoriel. Heureusement, ceci ne pose pas de difficulté particulière, mais nous allons tout de même expliquer brièvement pourquoi. Pour un intervalle I de $\mathbf{R}$ et un entier m, on considère une fonction $f : I \to \mathbf{R}^m$. Pour tout $x \in I$, $f(x)$ est un vecteur qu'on peut écrire en coordonnées de la façon suivante :

$$f(x) = \begin{pmatrix} f_1(x) \\ \vdots \\ f_m(x) \end{pmatrix}.$$

Ceci définit m fonctions $f_i : I \to \mathbf{R}$ pour $1 \leqslant i \leqslant m$ qui décrivent complètement la fonction f et sont appelées *(fonctions) coordonnées de f*. Ainsi, les propriété usuelles sur la fonction f s'adaptent à ce contexte en les considérant pour toutes les fonctions f_i. Notamment,

- La fonction f est *continue* en un point x si et seulement si f_i est continue pour tout $1 \leqslant i \leqslant m$;
- La fonction f est *dérivable* en un point x si et seulement si f_i est dérivable pour tout $1 \leqslant i \leqslant m$. De plus, on note alors

$$f'(x) = \begin{pmatrix} f_1'(x) \\ \vdots \\ f_m'(x) \end{pmatrix} \in \mathbf{R}^m.$$

- La fonction f est de classe $\mathcal{C}^k$ si et seulement si f_i est de classe $\mathcal{C}^k$ pour tout $1 \leqslant i \leqslant m$.
- La fonction f est intégrable si et seulement si f_i est intégrable pour tout $1 \leqslant i \leqslant m$. De plus, $\displaystyle\int_a^b f(t)\mathrm{d}t$ est le vecteur de coordonnées $\displaystyle\int_a^b f_i(t)\mathrm{d}t$.

Remarque 2.1.1. Profitons-en pour rappeler qu'une fonction $f : I \to \mathbf{R}$ est de classe $\mathcal{C}^1$ si elle est dérivable et si sa dérivée est continue, et de classe $\mathcal{C}^k$ pour $k \geqslant 2$ si elle est dérivable si sa dérivée est de classe $\mathcal{C}^{k-1}$.

Topologie des espaces vectoriels normés

Pour parler de continuité – et plus encore de dérivabilité – dans un cadre plus vaste que celui des intervalles de $\mathbf{R}$, on a inventé la topologie. Les notions de topologie nécessaires, qui ont déjà été vues l'an dernier, seront rappelées en temps utiles. Pour l'instant, redonnons simplement les plus fondamentales d'entre elles.

On considère $\mathbf{R}^n$ muni de la norme euclidienne, notée $\|\cdot\|$. Rappelons que pour un vecteur $x = (x_1, \cdots, x_n)$,

$$\|x\| = \sqrt{\sum_{i=1}^{n} x_i^2}.$$

Le fait que cette quantité est une norme signifie – outre le fait qu'elle est à valeurs positives – les trois propriétés suivantes :

- **Homogénéïté :** $\|\lambda x\| = |\lambda|\|x\|$ pour tout $\lambda \in \mathbf{R}$ et $x \in \mathbf{R}^n$;
- **Séparation :** $\|x\| = 0$ si et seulement si $x = 0$;
- **Inégalité triangulaire :** $\|x + y\| \leqslant \|x\| + \|y\|$.

Remarque 2.1.2. Dans $\mathbf{R}^n$, toutes les normes sont *équivalentes*, ce qui signifie que dans tout ce que nous allons faire on pourrait remplacer la norme euclidienne par n'importe quelle autre norme. Nous avons choisi la norme euclidienne parce que c'est la plus intuitive et souvent la plus pratique.

La norme permet de mesurer la distance entre deux éléments de $\mathbf{R}^n$, et notamment de quantifier le fait qu'ils soient "proches" l'un de l'autre. La définition la plus pratique pour ce faire est celle de boule ouverte.

Définition 2.1.3. Soit $x \in \mathbf{R}^n$ et $r > 0$. La *boule ouverte de centre x et de rayon r* est l'ensemble

$$B(x, r) = \{y \in \mathbf{R}^n \mid \|x - y\| < r\}.$$

Nous aurons souvent besoin de considérer une partie de $\mathbf{R}^n$, typiquement le domaine de définition d'une fonction, dans lequel on peut trouver des boules centrées autour de n'importe quel point. Ceci correspond à la notion centrale de la topologie.

Définition 2.1.4. Une partie $U \subset \mathbf{R}^n$ est dite *ouverte* si pour tout $x \in U$, il existe $r > 0$ tel que $B(x, r) \subset U$.

Le résultat suivant n'est pas standard mais sera utilisé de nombreuses fois. Il est donc raisonnable de le démontrer une fois pour toutes. La preuve est élémentaire, mais c'est un bon moyen de vérifier qu'on a compris les définitions.

Proposition 2.1.5. *Soit $x \in \mathbf{R}^n$ et $r > 0$. On note $x = (x_1, \cdots, x_n)$. Alors, existe $\delta > 0$ tel que pour tout $1 \leqslant i \leqslant n$ et pour tout $t \in]x_i - \delta; x_i + \delta[$,*

$$(x_1, \cdots, x_{i-1}, t, x_{i+1}, \cdots, x_n) \in B(x, r).$$

Démonstration. Il suffit de calculer

$$\begin{aligned} \|(x_1, \cdots, x_n) - (x_1, \cdots, x_{i-1}, t, x_{i+1}, \cdots, x_n)\| &= \sqrt{t^2} \\ &= |t| \end{aligned}$$

pour constater que $\delta = r$ convient. ∎

La topologie, c'est avant tout du vocabulaire. Au point que, quand on donne un nom à une partie, on en donne aussi un à son complémentaire.

Définition 2.1.6. Une partie de $\mathbf{R}^n$ est dite *fermée* si son complémentaire est une partie ouverte.

L'exemple le plus important est l'analogue de la boule ouverte définie plus haut. Pour $x \in \mathbf{R}^n$ et $r > 0$, la *boule fermée* de centre x et de rayon r est l'ensemble

$$B_f(x, r) = \{y \in \mathbf{R}^n \mid \|x - y\| \leqslant r\}.$$

Montrer que cette partie est effectivement fermée est un exercice profitable.

Quand une partie n'est pas ouverte, on peut essayer de se ramener à la partie ouverte la plus "proche" :

Définition 2.1.7. L'*intérieur* d'une partie U de $\mathbf{R}^n$ est le plus grand ouvert contenu dans U, c'est-à-dire la réunion de tous les ouverts contenus dans U.

On peut de même considérer la partie fermée la plus "proche" d'une partie donnée : l'*adhérence* de U est le plus petit fermé contenant U, c'est-à-dire l'intersection de tous les fermés contenant U.

Limites et continuité

Le but de la topologie est de permettre de définir des limites. Dans le cas d'une seule variable, la définition fait intervenir, outre les incontournables δ et ϵ, des valeurs absolues. En fait, il s'agit simplement de la norme euclidienne sur $\mathbf{R}$. Donc, en remplaçant les valeurs absolues par des normes, on obtient les généralisation des notions usuelles, que nous rappelons maintenant. Soit $f : U \to \mathbf{R}^m$ une fonction, avec U une partie de $\mathbf{R}^n$, et soit $x \in U$.

- On dit que f *a pour limite* $\ell \in \mathbf{R}^m$ *au point* x si pour tout $\epsilon > 0$, il existe $\delta > 0$ tel que pour tout $y \in B(x, \delta) \cap U$, $\|f(y) - \ell\| < \epsilon$.
- On dit que f est *continue en* x si f a pour limite $f(x)$ en x.
- On dit que f est *continue sur* U si f est continue en tout point de U.

La continuité a de nombreux intérêts, notamment en lien avec les suites. Ceci provient du résultat suivant dont nous ne rappellerons pas non plus la démonstration.

Proposition 2.1.8 (Caractérisation séquentielle de la continuité)**.** *Soit $f : U \to \mathbf{R}^m$ une fonction et $x \in U$. Alors, f est continue en x si et seulement si pour toute suite $(x_k)_{k\in\mathbf{N}}$ qui converge*[1] *vers x, la suite $(f(x_k))_{k\in\mathbf{N}}$ converge vers $f(x)$.*

Démonstration. Supposons tout d'abord que f est continue en x, et soit $(x_n)_{n\in\mathbf{N}}$ une suite de U qui converge vers x. Si $\epsilon > 0$ est fixé, alors par continuité de f il existe δ tel que $\|f(y) - f(x)\| < \epsilon$ pour tout $y \in B(x, \delta) \cap U$. D'autre part, la convergence de la suite $(x_n)_{n\in\mathbf{N}}$ assure l'existence d'un range N tel que pour tout $n \geqslant N$, $\|x_n - x\| < \delta$. Autrement dit, pour tout $n \geqslant N$, $x_n \in B(x, \delta) \cap U$ et par conséquent $\|f(x_n) - f(x)\| < \epsilon$. Nous venons ainsi de montrer que $(f(x_n))_{n\in\mathbf{N}}$ converge vers $f(x)$.

Réciproquement supposons que f n'est pas continue en x. Alors, la négation de la définition de la continuité donne l'existence d'un $\epsilon > 0$ tel que pour tout $\delta > 0$, il existe $y \in B(x, \delta) \cap U$ vérifiant $\|f(y) - f(x)\| > \epsilon$. En particulier, pour tout $n \in \mathbf{N}$, il existe un élément $x_n \in B(x, 1/(n+1)) \cap U$ tel que $\|f(x_n) - f(x)\| > \epsilon$. Par définition, on a pour tout $n \in \mathbf{N}$

$$\|x_n - x\| < \frac{1}{n+1} \underset{n\to+\infty}{\longrightarrow} 0$$

donc la suite $(x_n)_{n\in\mathbf{N}}$ converge vers x. Néanmoins, la suite $(f(x_n))_{n\in\mathbf{N}}$, elle, ne converge pas vers $f(x)$ puisque $\|f(x_n) - f(x)\| > \epsilon$ pour tout $n \in \mathbf{N}$. ∎

À l'aide de cette caractérisation, il est facile de vérifier que la continuité jouit d'un certain nombre de propriétés de stabilité qui permettent de montrer facilement qu'une fonction donnée est continue. En voici les principales :

1. Une suite $(x_n)_{n\in\mathbf{N}}$ de $\mathbf{R}^n$ converge vers une limite x si la suite $\|x_n - x\|$ tend vers 0.

- Les fonctions coordonnée $x \mapsto x_i$ sont continues.
- Si $f, g : U \to \mathbf{R}^m$ sont continues, alors leur somme $f + g$ est également continue.
- Si $f, g : U \to \mathbf{R}^m$ sont continues, alors leur produit fg est également continu.
- Si $f : U \to \mathbf{R}$ est continue et ne s'annule pas, alors $1/f$ est également continue.

Il suit par exemple de ces propriétés que toute fonction polynomiale en les coefficients (c'est-à-dire formée de sommes et de produits de fonctions coordonnées) est continue. Puisque nous utiliserons sans cesse ce résultat, nous allons lui donner un énoncé en bonne et due forme.

Nous aurons besoin dans la suite d'un résultat important concernant la continuité pour les applications linéaires.

Théorème 2.1.9 Soit n et m deux entier. Toute application $f : \mathbf{R}^n \to \mathbf{R}^m$ dont les coordonnées sont des polynômes en les variables est continue.

2.1.2 Le miracle de la compacité

Nous allons maintenant aborder la notion topologique qui permet de montrer l'existence d'extrema.

Définition 2.1.10. Une partie K de $\mathbf{R}^n$ est *compacte* si pour toute suite $(x_k)_{k\in\mathbf{N}}$ d'éléments de K, il existe une sous-suite $(x_{\phi(k)})_{k\in\mathbf{N}}$ qui converge vers un élément de K.

La définition précédente n'est en fait pas conçue pour être vérifiée, mais pour être utilisée. Mais alors, comment savoir qu'une partie est compacte ? C'est l'objet du prochain résultat, qui donne un critère très simple utilisant la notion suivante : une partie K de $\mathbf{R}^n$ est dite *bornée* s'il existe $M > 0$ tel que pour tout $x \in K$, $\|x\| < M$.

Remarque 2.1.11. On vérifie sans peine qu'une partie est bornée si et seulement si elle est contenue dans une boule (ouverte ou fermée).

Proposition 2.1.12. *Une partie K de $\mathbf{R}^n$ est compacte si et seulement si elle est à la fois fermée et bornée.*

D'accord, mais quel intérêt ? Il suffit de voir le résultat suivant pour comprendre que nous aurons souvent affaire à la compacité.

Théorème 2.1.13 (Théorème des bornes atteintes) Soit K une partie compacte de $\mathbf{R}^n$ et $f : K \to \mathbf{R}$ une fonction continue. Alors, f admet un maximum global et un minimum global : il existe $x_+, x_- \in K$ tel que pour tout $x \in K$,

$$f(x_-) \leqslant f(x) \leqslant f(x_+).$$

Démonstration. Traitons le cas du maximum, celui du minimum se démontrant de la même manière, *mutatis mutandis*. L'ensemble $f(K) \subset \mathbf{R}$ admet, comme toute partie non-vide de $\mathbf{R}$, une borne supérieure M dans $\mathbf{R} \cup \{+\infty\}$. De plus, par la caractérisation séquentielle de la borne supérieure, il existe une suite $(x_k)_{k\in\mathbf{N}}$ d'éléments de K telle que $f(x_k) \to M$. Soit $(x_{\phi(k)})_{k\in\mathbf{N}}$ une sous-suite qui converge vers une limite $x_+ \in K$, dont l'existence est assurée par la compacité de K. Alors, f étant continue, on peut appliquer la Proposition 2.1.8 pour conclure que

$$f(x_+) = \lim_{k\to+\infty} f(x_{\phi(k)}) = M.$$

Ceci prouve deux choses : $M \neq +\infty$ et M est atteinte, ce qu'il fallait démontrer. ■

Faisons halte pour reprendre les questions énoncées au début de ce chapitre. La première était de savoir s'il existe des extrema. Nous venons de voir un critère pour cela, qui est de nature purement *topologique*. Aucun besoin de calcul différentiel, de dérivation ou quoi que ce soit de ce genre. Il s'agit là d'un phénomène assez général en optimisation : les trois questions fondamentales se regroupent en deux parties en fonction des outils utilisés.

1. Démontrer l'existence d'extrema est un problème topologique ;
2. Trouver et identifier les extrema est un problème qui ne peut se résoudre de façon purement topologique en général et nécessite des outils plus avancés (calcul différentiel, analyse convexe, formes quadratiques).

Si le problème de l'existence d'extrema est de nature topologique, il ne requiert pas nécessairement de se placer sur un espace compact comme dans le Théorème 2.1.13. D'ailleurs, il est courant de chercher à optimiser une fonction sur des ensembles qui ne sont pas bornés. Pour cela, il existe heureusement des moyens, dont voici celui qui nous sera la plus utile. La démonstration sera faite en TD et est par conséquent omise.

Théorème 2.1.14 Soit $f : \mathbf{R}^n \to \mathbf{R}$ une fonction telle que $f(x) \to +\infty$ quand $\|x\| \to +\infty$. Alors, f admet un minimum global.

2.2 Localisation

J'élis un lieu flagrant et nul comme l'ossuaire des saisons.

Saint-John Perse, Exil

Nous avons maintenant un critère permettant de savoir qu'il existe des extrema, mais cela ne dit pas où les chercher. Pour aller plus loin, mettons-nous d'abord d'accord sur ce que nous entendons par extremum. Bien sûr, un extremum global est facile à définir.

Définition 2.2.1. Soit $f : U \to \mathbf{R}$ une fonction et $x \in U$. On dit que f admet

- Un *minimum global* en x si pour tout $y \in U$, $f(y) \geqslant f(x)$.
- Un *maximum global* en x si pour tout $y \in U$, $f(y) \leqslant f(x)$.

De plus, si l'inégalité précédente est stricte pour tout $y \neq x$, alors le minimum/maximum est dit *strict*. Si f admet un maximum ou un minimum global en x, on dit alors que f a un *extremum global* en x.

Le problème de cette définition, c'est qu'en pratique il est difficile de détecter directement des extrema globaux. Ce que l'on peut espérer de mieux, c'est de trouver des extrema *locaux* et c'est pour donner un sens à ce terme que la topologie est utile. En effet, un maximum local, par exemple, est un point où la valeur de la fonction est "plus grande" qu'en tous les points "suffisamment proches". Avec la notion de boule ouverte, il est facile de donner un sens précis à cette idée.

Définition 2.2.2. Soit $f : U \to \mathbf{R}$ une fonction et $x \in U$. On dit que f admet

- Un *minimum local* en x s'il existe $r > 0$ tel que pour tout $y \in B(x,r) \cap U$, $f(y) \geqslant f(x)$.
- Un *maximum local* en x s'il existe $r > 0$ tel que pour tout $y \in B(x,r) \cap U$, $f(y) \leqslant f(x)$.

De plus, si l'inégalité précédente est stricte pour tout $y \neq x$, alors le minimum/maximum est dit *strict*. Si f admet un maximum ou un minimum local en x, on dit alors que f a un *extremum local* en x.

Remarque 2.2.3. On pourrait aussi dire : f a un extremum local s'il existe $r > 0$ tel que f a un extremum global sur $B(x,r) \cap U$.

2.2.1 Le premier théorème

Commençons par rappeler un résultat bien connu. Si $f : I \to \mathbf{R}$ est une fonction dérivable d'une variable qui admet un extremum local en un point x, alors $f'(x) = 0$. Ainsi, pour trouver les extrema de f, il suffit de trouver les points auxquels la dérivée s'annule. Alors, pourquoi ne pas essayer de faire de même avec plusieurs variables ? Avant de s'embarquer sur cette piste, un commentaire s'impose. Nous allons considérer des fonctions de la forme $f : U \to \mathbf{R}$, où U est une partie de $\mathbf{R}^n$. Ainsi, si $x = (x_1, \cdots, x_n) \in U$, on peut voir f comme une fonctions possédant une variable vectorielle et on écrira dans ce cas $f(x)$ pour la valeur de f en x. Ou bien, on peut voir f comme une fonction possédant n variables scalaires, et on notera alors le même nombre $f(x_1, \cdots, x_n)$. La façon la plus naturelle d'étendre la notion de dérivation aux fonctions de plusieurs variables est alors tout simplement de considérer chaque variable séparément. Il est tout à fait remarquable que, malgré quelques subtilités techniques que nous aborderons ci-après, cette idée soit essentiellement la bonne et permette d'étendre tous les outils d'optimisation de la dimension un. Commençons donc par donner une définition pour fixer les idées.

Définition 2.2.4. Soit $f : U \to \mathbf{R}^m$ une fonction et $x = (x_1, \cdots, x_n)$ un point de U. La *dérivée partielle de f par rapport à la i-ème coordonnée* au point x est la dérivée de la fonction

$$f_{x,i} : t \mapsto f(x_1, \cdots, x_{i-1}, t, x_{i+1}, \cdots, x_n)$$

au point x_i. On la note

$$\boxed{\frac{\partial f}{\partial x_i}(x) = f'_{x,i}(x_i).}$$

Si f admet une dérivée partielle par rapport à la i-ième coordonnée en tout point de U, on notera

$$\frac{\partial f}{\partial x_i} : U \to \mathbf{R}^m$$

la fonction correspondante.

Faisons quelques commentaires.

- Tout d'abord, comme U est ouvert, il existe $r > 0$ tel que la boule $B(x,r)$ de centre x et de rayon r soit contenue dans U. La Proposition 2.1.5 donne donc $\delta > 0$ tel que

  $$(x_1, \cdots, x_{i-1}, t, x_{i+1}, \cdots, x_n) \in B(x,r) \subset U$$

  pour tout $t \in]x_i - \delta; x_i + \delta[$. Autrement dit, la fonction $f_{x,i}$ est définie sur l'intervalle $]x_i - \delta; x_i + \delta[$ et cela a donc bien un sens de considérer la dérivée de la fonction $f_{x,i}$ au point x_i.

- On peut reformuler la définition comme suit : « la dérivée partielle par rapport à la i-ième coordonnée est la dérivée de la fonction obtenue en fixant toutes les coordonnées sauf la i-ième. »
- On peut également donner une expression plus explicite et littérale de la dérivée partielle en utilisant la définition du nombre dérivé d'une fonction en un point :

$$\frac{\partial f}{\partial x_i}(x) = \lim_{t \to x_i} \frac{f(x_1, \cdots x_{i-1}, t, x_{i+1}, \cdots, x_n) - f(x)}{t - x_i}.$$

- Notons $(e_1, \cdots, e_n)$ la base canonique de $\mathbf{R}^n$. Alors,

$$f_{x,i}(t) = f(x + (t - x_i)e_i)$$

et sa dérivée en x_i coïncide avec la dérivée en 0 de la fonction $t \mapsto f(x + te_i)$. On pourrait donc plus généralement considérer un vecteur $v \in \mathbf{R}^n$ et s'intéresser à la dérivée, si elle existe, de la fonction

$$t \mapsto f(x + tv).$$

On obtient alors la *dérivée directionnelle de f dans la direction v*, une notion que nous n'utiliserons pas.

- Il existe dans la littérature plusieurs notations pour désigner les dérivées partielles. On peut par exemple rencontrer $\partial_{x_i} f$ ou $\partial_i f$. Nous nous efforcerons ici de garder la notation fractionnaire de la Définition 2.2.4.

Avant d'aller plus loin, remarquons que notre définition des dérivées partielles concerne des fonctions à valeurs vectorielles. On pourrait se demander pourquoi nous ne nous restreignons pas aux fonctions à valeurs dans $\mathbf{R}$, puisque ce sont celles qui vont nous intéresser. Le fait est que, même en partant d'une telle fonction, nous allons être amenés à considérer à partir d'elles d'autres fonctions qui seront, cette fois, à valeurs vectorielles (voir par exemple le Paragraphe 2.3 ci-dessous). Il est donc indispensable de travailler dès maintenant dans cette généralité.

Notre objectif principal est, rappelons-le, d'être capable de détecter les extrema d'une fonction par une propriété d'annulation de la dérivée. Or, nous avons déjà atteint cet objectif ! Le tout est de s'en rendre compte ...

Théorème 2.2.5 (Localisation des extrema) Soit $f : U \to \mathbf{R}$ une fonction qui admet des dérivées partielles par rapport à toutes les coordonnées. Si x est un point auquel f admet un extremum, alors

$$\frac{\partial f}{\partial x_i}(x) = 0$$

pour tout $1 \leqslant i \leqslant n$.

Démonstration. Fixons une coordonnée $1 \leqslant i \leqslant n$ et soit, par la Proposition 2.1.5, $\delta > 0$ tel que pour tout $t \in]x_i - \delta; x_i + \delta[$, on ait $(x_1, \cdots, x_{i-1}, t, x_{i+1}, \cdots, x_n) \in U$. Alors, la fonction $f_{x,i}$ a, par hypothèse, une extremum local en $t = x_i$. Donc, par le résultat dans le cas d'une variable, sa dérivée s'annule en x_i. Autrement dit,

$$\frac{\partial f}{\partial x_i}(x) = 0,$$

ce qu'il fallait démontrer.

Afin d'être plus complet, nous allons rappeler la démonstration du cas d'une variable. Supposons que $f_{x,i}$ ait un maximum local en x_i, le cas d'un minimum se traitant de la même façon, *mutatis mutandis*. Rappelons qu'il existe $\delta > 0$ tel que $f_{x,i}$ est définie sur l'intervalle $]x_i - \delta; x_i + \delta[$. Si $-\delta < t < 0$, alors comme $f(t) \leqslant f(x_i)$, on a

$$\frac{f_{x,i}(x_i + t) - f_{x,i}(x_i)}{t} \geqslant 0$$

et en faisant tendre t vers 0 on conclut que $f'_{x,i}(x_i) \geqslant 0$. Par ailleurs, si $0 < t < \delta$, on a toujours $f(t) \leqslant f(x_i)$, donc

$$\frac{f_{x,i}(x_i + t) - f_{x,i}(x_i)}{t} \geqslant 0.$$

En faisant tendre t vers 0, on obtient cette fois-ci $f'_{x,i}(x_i) \leqslant 0$. En combinant les deux inégalités, on trouve finalement que $f'_{x,i}(x_i) = 0$, ce qu'il fallait démontrer. ∎

2.2.2 Un exemple éclairant

L'histoire de la physique a retenu le nom de Fermat comme celui de l'un des premiers à avoir formulé une loi physique sous forme de problème d'optimisation. S'intéressant à l'optique géométrique, c'est-à-dire à l'étude de la propagation d'un rayon lumineux dans différents milieux, il remarque que si l'on connait le point de départ et le point d'arriver du rayon, alors la trajectoire est toujours celle qui minimise le temps du trajet : c'est le Principe de moindre temps [2].

Nous allons regarder ce que ce principe nous dit dans le cas du phénomène de *réfraction*. On considère une partie de l'espace formée de deux milieux homogènes séparés par une vitre plane dont on néglige l'épaisseur. Un rayon lumineux part d'un point P_1 dans le premier milieu et arrive à un point P_2 dans le second milieu. Quelle est sa trajectoire ? Tant que le rayon est dans un milieu, le plus court chemin est un segment de droite. Par conséquent, la trajectoire du rayon sera formée de deux segments $[P_1Q]$ et $[QP_2]$, où Q est un point de la vitre. Le problème est simplement de trouver ce point Q.

Nous allons commencer par définir un repère qui nous simplifiera la vie. Si $\mathcal{P}$ est le plan de la vitre, on fixe un repère orthonormé de sorte que $\mathcal{P}$ soit le plan d'équation $z = 0$. On peut également choisir les axes de sorte que les coordonnées du point d'arrivée P_2 soient $(x_2, 0, z_2)$. Enfin, en fixant convenablement l'origine on peut faire en sorte que les coordonnées du point de départ P_1 soient $(0, 0, z_1)$. Quant au point Q, ses coordonnées sont $(x, y, 0)$ et nous aurons donc deux variables données par les deux premières coordonnées.

Si v_1 désigne la vitesse de la lumière dans le premier milieu (la vitesse est constante puisque le milieu est homogène) et v_2 la vitesse de la lumière dans le second milieu, le temps de parcours global est

$$T(x, y) = \frac{1}{v_1}\sqrt{x^2 + y^2 + z_1^2} + \frac{1}{v_2}\sqrt{(x_2 - x)^2 + y^2 + z_2^2}.$$

2. Comme le fait remarquer R. Feynman dans [4], le nom est trompeur, car il ne s'agit pas en général du trajet prenant le moins de temps (penser au cas d'une réflexion) mais plutôt d'une trajectoire « telle qu'il y a beaucoup d'autres trajectoires voisines qui prennent presque exactement le *même* temps », au sens où la différence de temps sera nulle au premier ordre. Toutefois, dans notre cas, cela correspond effectivement au trajet prenant le moins de temps.

Pour trouver un minimum de T, on calcule ses dérivées partielles :

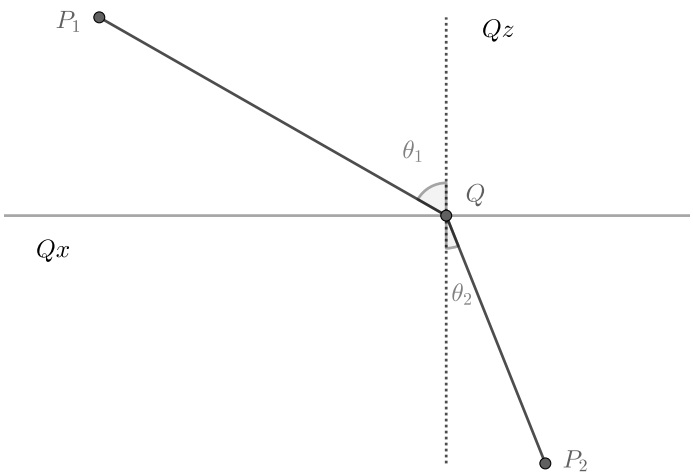
$$\frac{\partial T}{\partial x}(x,y) = \frac{x}{v_1\sqrt{x^2+y^2+z_1^2}} + \frac{x-x_2}{v_2\sqrt{(x_2-x)^2+y^2+z_2^2}}$$
$$\frac{\partial T}{\partial y}(x,y) = \frac{y}{v_1\sqrt{x^2+y^2+z_1^2}} + \frac{y}{v_2\sqrt{(x_2-x)^2+y^2+z_2^2}}.$$

Si le second terme s'annule, alors $y = 0$. Autrement dit, le rayon se déplace nécessairement dans l'unique plan $\mathcal{P}'$ orthogonal à $\mathcal{P}$ contenant P_1 et P_2 ! Cela n'a rien d'évident et faisait avant le Principe de Fermat partie des "lois de l'optique géométrique".

On peut maintenant simplifier la première expression pour trouver

$$\frac{\partial T}{\partial x}(x,0) = \frac{x}{v_1\sqrt{x^2+z_1^2}} + \frac{x-x_2}{v_2\sqrt{(x_2-x)^2+z_2^2}}.$$

Pour interpréter cette équation, un dessin s'impose. Nous représentons ci-dessus la trajectoire du rayon dans le plan $\mathcal{P}'$.



On voit que les coordonnées de P_1 dans le repère (Q, Qx, Qz) sont

$$(-QP_1\sin(\theta_1), QP_1\cos(\theta)_1),$$

où θ_1 est l'angle formé par (QP_1) et Qx. De même, les coordonnées de P_2 sont

$$(QP_2\sin(\theta_2), QP_2\cos(\theta_2)).$$

D'autre part, l'abscisse de P_1 dans le repère (Q, Qx, Qz) est égale à $-x$, tandis que celle de P_2 est égale à $x_2 - x$. Ainsi,

$$\frac{\partial T}{\partial x}(x,y) = \frac{1}{v_1}\frac{QP_1\sin(\theta_1)}{QP_1} + \frac{1}{v_2}\frac{-QP_2\sin(\theta_2)}{QP_2}.$$

Par conséquent, l'annulation de la première dérivée partielle s'écrit

$$\frac{\sin(\theta_1)}{v_1} = \frac{\sin(\theta_2)}{v_2}.$$

En notant $n_i = c/v_i$ l'*indice* de chaque milieu, on retrouve la Loi de Snell [3]-Descartes

$$\boxed{n_1 \sin(\theta_1) = n_2 \sin(\theta_2).}$$

2.2.3 Les dérivées partielles

Le Théorème 2.2.5 montre que les dérivées partielles seront l'outil central de notre étude des extrema. Il est donc utile de rassembler des informations et des résultats sur elles qui nous simplifieront le travail ultérieurement. Au vu de leur définition, les règles de calcul usuelles pour les dérivées s'appliquent également au dérivées partielles, notamment

- Si f et g admettent une dérivée partielle par rapport à la i-ième coordonnée, alors leur somme également, et
$$\frac{\partial(f+g)}{\partial x_i} = \frac{\partial f}{\partial x_i} + \frac{\partial g}{\partial x_i}.$$
- Si f et g sont à valeurs réelles et admettent une dérivée partielle par rapport à la i-ième coordonnée, alors leur produit également, et
$$\frac{\partial(f \times g)}{\partial x_i} = \frac{\partial f}{\partial x_i} g + f \frac{\partial g}{\partial x_i}.$$
- En particulier, toute fonction dont les coordonnées sont polynomiales en les variables admet des dérivées partielles par rapport à chaque coordonnée.
- Si f est à valeurs réelles, ne s'annule pas et admet une dérivée partielle par rapport à la i-ème coordonnée, alors son inverse également, et
$$\frac{\partial(1/f)}{\partial x_i} = -\frac{1}{f^2} \frac{\partial f}{\partial x_i}.$$

La seule difficulté provient de la composition. En effet, il nous est maintenant possible de composer des fonctions à valeurs vectorielles, et l'analogue de la formule usuelle $(g \circ f)' = f' \times g' \circ f$ n'est pas clair dans ce cadre. Et de fait, il n'y a tout simplement pas de formule, car il n'est pas vrai que si f et g admettent des dérivées partielles, alors $g \circ f$ également. Voici un exemple qui illustre ce phénomène quelque peu déroutant au premier abord.

Exemple 2.2.6. Considérons la fonction $f : \mathbf{R} \to \mathbf{R}^2$ définie par

$$f(x) = (x, x)$$

et la fonction $g : \mathbf{R}^2 \to \mathbf{R}$ définie par

$$(x,y) \mapsto \begin{cases} \dfrac{xy}{x^2+y^2} & \text{si } (x,y) \neq (0,0) \\ 0 & \text{si } (x,y) = (0,0) \end{cases}$$

3. Willebrord Snell van Royen (1580–1626) : physicien et mathématicien néerlandais, il a développé et raffiné plusieurs méthodes numériques, ce qui lui a permis de donner une approximation de π à 25 décimales près ou encore de donner le premier calcul du rayon terrestre par triangulation. Il n'a jamais publié la loi d'optique qui porte son nom, mais Huygens lui en a donné la paternité contre Descartes.

Il est clair que f est dérivable en 0. La fonction g, quant à elle, admet des dérivées partielles en $(0,0)$. En effet,

$$\frac{g(0+h,0)-g(0,0)}{h} = 0 \underset{h\to 0}{\longrightarrow} 0$$

et

$$\frac{g(0,0+h)-g(0,0)}{h} = 0 \underset{h\to 0}{\longrightarrow} 0$$

donc

$$\frac{\partial g}{\partial x}(0,0) = 0 = \frac{\partial g}{\partial y}(0,0).$$

De plus, $(g\circ f)(0) = 0$, mais pour tout $h \neq 0$,

$$(g\circ f)(h) = \frac{h^2}{h^2+h^2} = \frac{1}{2}.$$

Ainsi $g\circ f : \mathbf{R} \to \mathbf{R}$ est une fonction constante égale à 1/2 sauf en 0 où elle vaut 0. Une telle fonction n'est pas dérivable en 0.

Ceci est l'une des manifestations d'un problème profond lié aux dérivées partielles, que nous réglerons en imposant systématiquement une hypothèse de régularité plus forte sur les fonctions (voir le Paragraphe 2.3.2).

2.2.4 Un exemple productif

Une des notions les plus importantes en économétrie est celle de *fonction de production*, introduite par Wicksteed[4] en 1894 pour décrire la production d'une entité économique en fonction de ses *facteurs de production*, c'est-à-dire ce qu'elle utilise dans le processus (ressources, argent, travail, ...). Un modèle empirique simple utilisé pour décrire cette fonction est celui proposé par Cobb[5] et Douglas[6]. L'idée est de grouper tous les facteurs de production en deux variables : la capital K (pour *Kapital* en allemand) et le travail L (pour *Labour* en anglais). En comparant avec les données disponibles à l'époque[7], on obtient par une méthode de moindre carrés la *fonction de Cobb-Douglas*

$$P(K,L) = 1,01 \times K^{1/4} L^{3/4}.$$

Comme il s'agit d'un produit, on peut utiliser la formule bien connue pour calculer les dérivées partielles de P, ce qui donne

$$\frac{\partial P}{\partial K}(K,L) = 0,2525\left(\frac{L}{K}\right)^{3/4} \quad \& \quad \frac{\partial P}{\partial L}(K,L) = 0,7575\left(\frac{K}{L}\right)^{1/4}.$$

Ces dérivées partielles sont appelées en économie *productivités marginales*. Elles montrent l'impact du changement de chacun des facteurs de production sur la production globale de l'entité.

4. Philip Wicksteed (1844–1927) : économiste anglais appartenant au courant dit *marginaliste*. Il a introduit les fonctions de productions et les quantités "marginales" associées afin de décrire quantitativement leur rôle dans les équilibres économiques.

5. Charles Cobb (1875 – 1949) : économiste et mathématicien américain.

6. Paul Howard Douglas (1892 – 1976) : économiste et homme politique américain, il a notamment été sénateur de l'Illinois sous la bannière du Parti Démocrate.

7. Le travail de Cobb et Douglas exploite les statistiques de la production manufacturière américaine au niveau macroéconomique, entre 1899 et 1922.

La notion de productivité marginale peut être raffinée par celle d'*élasticité de production*. Il s'agit de comparer la productivité marginale non pas à une variation absolue du paramètre, mais à une variation relative. De même que la dérivée partielle par rapport à L décrit la productivité marginale de L, c'est-à-dire l'impact d'une variation infinitésimal de L, le rapport P/L décrit la *productivité moyenne* de L qui est une autre mesure du poids de L dans la production. L'élasticité de production est obtenue en comparant ces deux quantités :

$$e_K = \frac{1}{P/K}\frac{\partial P}{\partial K}(K,L) = \frac{1}{4} \quad \& \quad e_L = \frac{1}{P/L}\frac{\partial P}{\partial L}(K,L) = \frac{3}{4}.$$

On voit ici l'un des points importants de ce modèle : l'élasticité de production se retrouve directement dans les exposants. Autrement dit, on peut ajuster les paramètres de la fonction pour choisir les valeurs des élasticités de production.

2.3 Identification

L'œil : la fenêtre de l'âme ; le centre de la beauté du visage ;
le point où se concentre l'identité d'un individu.

Milan Kundera, L'identité

Nous avons maintenant une méthode pour localiser les extrema d'une fonction. Mais il ne s'agit que d'une condition nécessaire, rien ne garantit qu'elle soit suffisante. D'ailleurs, elle ne l'est pas, et cela se voit déjà avec une seule variable.

Exemple 2.3.1. Soit $f : \mathbf{R} \to \mathbf{R}$ donnée par $f(x) = x^3$. Alors, $f'(0) = 0$, mais f n'a pas d'extremum local en 0 puisque $f(x) > 0$ si $x > 0$ et $f(x) < 0$ si $x < 0$.

Afin de ne pas confondre extrema et points où les dérivées partielles s'annulent, nous allons – une fois n'est pas coutume – donner un nom à ces derniers. Avant cela, introduisons une notation pratique. Si $f : U \to \mathbf{R}$ est une fonction admettant des dérivées partielles par rapport à chaque coordonnée, on peut regrouper ces dérivées partielles pour former un vecteur

$$\nabla f(x) = \begin{pmatrix} \dfrac{\partial f}{\partial x_1}(x) \\ \vdots \\ \dfrac{\partial f}{\partial x_n}(x) \end{pmatrix}$$

Ce vecteur est appelé *gradient de f au point x*, et la fonction associée est aussi appelée *gradient de f*. On constate que le gradient est non seulement une fonction de plusieurs variables, mais également une fonction à valeurs vectorielles

$$\nabla f : U \to \mathbf{R}^n.$$

Ceci montre, comme nous l'évoquions plus haut, la nécessité de ne pas se restreindre à $\mathbf{R}$ pour l'espace d'arrivée.

Définition 2.3.2. Soit $f : U \to \mathbf{R}$ une fonction qui admet des dérivées partielles par rapport à chaque coordonnée. Un *point critique* de f est un point $x \in U$ tel que

$$\boxed{\nabla f(x) = 0}$$

2.3.1 La stratégie

Comment donc identifier, parmi les points critiques, les extrema ? Et comment déterminer leur nature (minimum ou maximum) ? Dans le cas d'une variable, cela se fait en utilisant un développement limité. Pour plus de clarté, nous allons tout d'abord rappeler ce résultat ainsi que sa preuve.

Proposition 2.3.3. *Soit I un intervalle de $\mathbf{R}$ et $f : I \to \mathbf{R}$ une fonction deux fois dérivables. Soit $x \in I$ tel que $f'(x) = 0$. Alors,*

- *Si $f''(x) > 0$, f admet un minimum local strict en x.*
- *Si $f''(x) < 0$, f admet un maximum local strict en x.*

Remarque 2.3.4. L'énoncé indique déjà que cette ce critère ne pourra pas toujours nous aider, puisqu'il ne peut détecter que des extrema stricts.

Démonstration. La preuve la plus simple consiste à exploiter le tableau de variations de f autour du point x. Malheureusement, cette notion ne s'étend pas à plusieurs variables et nous allons donc donner une démonstration plus abstraite et généralisable.

Le point de départ de la preuve est d'écrire le développement limité de f au second ordre en x grâce à la Formule de Taylor-Young, ce qui donne

$$\begin{aligned} f(x+h) - f(x) &= hf'(x) + \frac{h^2}{2} f''(x) + o(h^2) \\ &= \frac{h^2}{2} f''(x) + o(h^2). \end{aligned}$$

Cette formule est valable pour tout $h \in]-\delta;\delta[$ pour un certain $\delta > 0$. La *notation de Landau* $o(h^2)$ désigne une fonction de la forme $h^2\varepsilon(h)$, où $\varepsilon(h) \to 0$ quand $h \to 0$. On peut donc écrire l'égalité précédente sous la forme

$$f(x+h) - f(x) = h^2 \left(\frac{f''(x)}{2} + \varepsilon(h) \right).$$

Supposons par exemple $f''(x) > 0$, l'autre cas se traitant – *mutatis mutandis* – de la même façon. Le terme entre parenthèses dans le membre de droite tend vers $f''(x)/2$ quand h tend vers 0, et il existe par conséquent $\delta' > 0$ tel qu'en posant $\delta'' = \min(\delta, \delta')$, on a pour tout $h \in]-\delta'';\delta''[$

$$\frac{f''(x)}{2} + \varepsilon(h) > \frac{f''(x)}{4} > 0.$$

Ainsi, si $y \in]x-\delta'';x+\delta''[\cap I$, on a

$$\begin{aligned} f(y) - f(x) &= f(x + (y-x)) - f(x) \\ &= (y-x)^2 \left(\frac{f''(x)}{2} + \varepsilon(h) \right) \\ &> (y-x)^2 \frac{f''(x)}{4} \\ &> 0. \end{aligned}$$

Autrement dit, f admet un minimum local strict en x. ∎

L'outil crucial dans la démonstration précédente est le développement limité à l'ordre deux. Notre objectif est donc d'obtenir quelque chose d'analogue pour les fonctions de plusieurs variables.

2.3.2 Développement limité

Premier ordre

Commençons – *chi va piano, va sano* – par nous intéresser au développement limité à l'ordre un. Il s'agit, pour une fonction $f : I \to \mathbf{R}$, de l'égalité

$$f(x+h) - f(x) = hf'(x) + o(h).$$

Ce qui est caché dans cette formule, c'est le fait que la dérivée de f donne une approximation de f par une fonction linéaire au voisinage du point x, à savoir la fonction $h \mapsto hf'(x)$. Il y a un candidat naturel pour une telle application linéaire quand la fonction a plusieurs variables : il suffit d'utiliser les dérivées partielles !

Définition 2.3.5. Soit $f : U \to \mathbf{R}^m$ et un point $x \in U$ auquel f admet des dérivées partielles selon toutes les coordonnées. La *différentielle* de f au point x est l'application linéaire $D_x f : \mathbf{R}^n \to \mathbf{R}^m$ définie par

$$\boxed{D_x(f)(h_1, \cdots, h_n) = \sum_{i=1}^{n} h_i \frac{\partial f}{\partial x_i}(x)}$$

Remarque 2.3.6. Bien que f ne soit définie que sur une partie ouverte U de $\mathbf{R}^n$, la différentielle est une application linéaire qui est donc toujours définie sur tout l'espace vectoriel $\mathbf{R}^n$. Géométriquement, cela provient du fait que $\mathbf{R}^n$ est *l'espace tangent* à U au point x. Nous reverrons cela en temps voulu au Paragraphe 3.1.3.

On sait bien que toute application linéaire peut se représenter, une fois des bases fixées dans les espaces vectoriels de départ et d'arrivée, par une matrice. Si on note f_i les fonctions coordonnées de f, pour $1 \leqslant i \leqslant m$, alors la matrice de $D_x f$ dans les bases canoniques est

$$J_f(x) = \left(\frac{\partial f_i}{\partial x_j}\right)_{1 \leqslant i \leqslant m, 1 \leqslant j \leqslant n}.$$

Cette matrice permet facilement de représenter la différentielle de f et est donc pratique pour les calculs. On l'appelle *matrice Jacobienne de f au point x*. Si $m = 1$, cette matrice est une colonne, qui n'est autre que le gradient $\nabla f(x)$ de f au point x. Autrement dit, pour une fonction à valeurs scalaires on a

$$\boxed{D_x f(h) = \langle \nabla f(x), h \rangle}$$

Nous voudrions donc maintenant obtenir une égalité de la forme

$$f(x+h) = f(x) + D_x f(h) + o(h),$$

mais encore faut-il donner un sens au symbole o. Le problème est qu'on ne peut pas supposer qu'il s'agit d'une fonction de la forme $h\varepsilon(h)$ car h est maintenant un vecteur et cette écriture n'a pas nécessairement de sens. La solution la plus élémentaire est de faire intervenir la norme : le symbole $o(h)$ désignera une fonction de la forme

$$h \mapsto \|h\| \varepsilon(h),$$

où $\varepsilon : \mathbf{R}^n \to \mathbf{R}^m$ est une fonction qui tend vers 0 quand son argument tend vers 0.

Nous avons tous les outils en main, mais maintenant va apparaître l'une des spécificités les plus notables du cas de plusieurs variables. En effet, l'existence des dérivées

partielles ne garantit pas en général qu'on puisse écrire un développement limité au premier ordre ! Le problème est le même que pour la composition, comme le montre l'exemple suivant.

Exemple 2.3.7. Reprenons la fonction $f : \mathbf{R}^2 \to \mathbf{R}^2$ de l'Exemple 2.2.6, qui était définie par

$$(x,y) \mapsto \begin{cases} \dfrac{xy}{x^2+y^2} & \text{si } (x,y) \neq (0,0) \\ 0 & \text{si } (x,y) = (0,0) \end{cases}$$

Nous avons déjà vu qu'elle admet des dérivées partielles en $(0,0)$.

Pour montrer que f n'admet pas de développement limité en $(0,0)$, observons que si un tel développement existe, alors f doit être continue en $(0,0)$. En effet, toute application linéaire sur un espace vectoriel de dimension finie est continue (Théorème 2.1.9), donc on aurait

$$\begin{aligned} \|f(x+h) - f(x)\| &\leqslant \|D_x f(h)\| + \|o(h)\| \\ &\underset{h\to 0}{\longrightarrow} 0. \end{aligned}$$

Or, la fonction f n'est pas continue en $(0,0)$ puisque $f(1/n, 1/n) = 1/2$ ne tend pas vers $0 = f(0,0)$ quand n tend vers $+\infty$, contredisant la Proposition 2.1.8.

Heureusement, la situation n'est pas complètement désespérée. En effet, si l'on est prêt à faire une hypothèse un peu plus forte sur les dérivées partielles, alors tout marche.

Définition 2.3.8. Une fonction $f : U \to \mathbf{R}^m$ est de *classe* $\mathcal{C}^1$ si elle admet des dérivées partielles par rapport à chaque variable, et si ces dernières sont des fonctions continues.

Proposition 2.3.9 (Développement limité au premier ordre)**.** *Soit $f : U \to \mathbf{R}^m$ une fonction de classe $\mathcal{C}^1$ et $x \in U$. Alors, pour tout $h \in \mathbf{R}^n$ tel que $x + h \in U$,*

$$\boxed{f(x+h) - f(x) = D_x f(h) + o(h)} \tag{2.1}$$

Démonstration. La preuve n'est pas vraiment compliquée mais les notations sont un peu lourdes. Inspirez profondément, on y va ! Tout d'abord, il faut comprendre qu'à cause de la présence de $o(h)$, cet énoncé ne donne ne fait que la valeur d'une limite quand h tend vers 0. Ainsi, il suffit de démontrer l'égalité pour h dans une boule ouverte centrée en 0. Et de fait, comme U est ouvert il existe $\delta > 0$ tel que $B(x,\delta) \subset U$, donc $x + h \in U$ dès que $h \in B(0,\delta)$.

Écrivons $x = (x_1, \cdots, x_n)$ et $h = (h_1, \cdots, h_n)$, et décomposons astucieusement la quantité qui nous intéresse de la façon suivante :

$$\begin{aligned} f(x+h) - f(x) &= \sum_{i=1}^{n} f(x_1 + h_1, \cdots, x_i + h_i, x_{i+1}, \cdots, x_n) - f(x_1 + h_1, \cdots, x_{i-1} + h_{i-1}, x_i, \cdots, x_n) \\ &= \sum_{i=1}^{n} \int_0^{h_i} \frac{\partial f}{\partial x_i}(x_1 + h_1, \cdots, x_{i-1} + h_{i-1}, x_i + t, x_{i+1}, \cdots, x_n)\,\mathrm{d}t \end{aligned}$$

Pour comparer cette quantité à $D_x f(h)$, on peut écrire

$$\begin{aligned} D_x f(h) &= \sum_{i=1}^{n} h_i \frac{\partial f}{\partial x_i}(x_1, \cdots, x_n) \\ &= \sum_{i=1}^{n} \int_0^{h_i} \frac{\partial f}{\partial x_i}(x_1, \cdots, x_n) \mathrm{d}t. \end{aligned}$$

Ainsi, la quantité qui nous intéresse, à savoir $f(x+h) - f(x) - D_x f(h)$, est égale à

$$\sum_{i=1}^{n} \int_0^{h_i} \left(\frac{\partial f}{\partial x_i}(x_1 + h_1, \cdots, x_{i-1} + h_{i-1}, x_i + t, x_{i+1}, \cdots, x_n) - \frac{\partial f}{\partial x_i}(x_1, \cdots, x_n) \right) \mathrm{d}t$$

Soit $\epsilon > 0$. Alors, pour tout $1 \leqslant i \leqslant n$, il existe par continuité $\delta_i > 0$ tel que $\|y - x\| < \delta_i$ implique

$$\left\| \frac{\partial f}{\partial x_i}(y) - \frac{\partial f}{\partial x_i}(x) \right\| < \epsilon.$$

Si l'on pose $\delta = \min_i(\delta_i)$, alors pour $\|h\| < \delta$ toutes les intégrandes sont plus petites en norme que ϵ. Autrement dit,

$$\begin{aligned} \|f(x+h) - f(x) - D_x f(h)\| &< \sum_{i=1}^{n} |h_i| \epsilon \\ &= \epsilon \sum_{i=1}^{n} |h_i| \\ &\leqslant \epsilon \sqrt{n} \sqrt{\sum_{i=1}^{n} h_i^2} \\ &= \epsilon \sqrt{n} \|h\|, \end{aligned}$$

où la dernière inégalité est une application [8] de l'Inégalité de Cauchy-Schwarz.

La preuve est maintenant terminée, mais nous allons prendre le temps de nous en convaincre. Soit $\epsilon > 0$ et $\epsilon' = \epsilon/n$. Soit δ associé comme ci-dessus à ϵ'. Alors, pour tout h tel que $\|h\| < \delta$, on a

$$\frac{1}{\|h\|} (f(x+h) - f(x) - D_x f(h)) < \epsilon.$$

Autrement dit, le membre de gauche est une fonction qui tend vers 0 quand $\|h\|$ tend vers 0. Ceci signifie que $f(x+h) - f(x) - D_x f(h)$ est le produit de $\|h\|$ par une fonction qui tend vers 0 quand $\|h\| \to 0$, ce qui conclut. ∎

Remarque 2.3.10. Si $f : U \to \mathbf{R}^m$ est une fonction, et si pour un point $x \in U$ il existe [9] une application linéaire $D_x f$ vérifiant l'Équation (2.1), alors f est dite *différentiable en* x. Il n'est pas nécessaire pour cela qu'elle soit de classe $\mathcal{C}^1$, mais cette condition est très souvent valide et facile à vérifier dans la pratique. Nous nous y restreindrons donc dans la suite.

8. On peut en effet observer, en notant ξ le vecteur dont toutes les coordonnées sont égales à 1, que pour tout vecteur $k \in \mathbf{R}^n$,

$$\sum_{i=1}^{n} k_i = \langle \xi, k \rangle \leqslant \|\xi\| \|k\| = \sqrt{n} \|k\|.$$

Il suffit ensuite d'appliquer ce résultat au vecteur de coordonnées $k_i = |h_i|$.

9. C'est un exercice simple mais profitable sur les développements limités vectoriels de vérifier qu'une telle application, si elle existe, est nécessairement unique.

Le développement limité n'est pas qu'une propriété intéressante des fonctions de classe $\mathcal{C}^1$, c'est aussi un moyen utile d'en calculer les dérivées partielles. En effet, il est parfois plus simple et plus naturel de calculer le développement limité d'une fonction que de calculer des dérivées partielles. Nous allons expliciter ceci sous forme d'une proposition.

Proposition 2.3.11. *Soit $f : U \to \mathbf{R}^m$ une fonction. On suppose que f admet un développement limité au premier ordre au sens suivant : il existe une application linéaire $L : \mathbf{R}^n \to \mathbf{R}^m$ tel que pour tout h tel que $x + h \in U$,*

$$f(x+h) - f(x) = L(h) + o(h).$$

Alors, f admet des dérivées partielles par rapport à toutes les coordonnées au point x, et si $(e_i)_{1\leqslant i\leqslant n}$ désigne la base canonique de $\mathbf{R}^n$, on a

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

Démonstration. Soit I un intervalle de $\mathbf{R}$ tel que $x + te_i \in U$ pour tout $t \in I$ (qui existe par la Proposition 2.1.5). On a alors

$$\begin{aligned}\frac{f(x+te_i) - f(x)}{t} &= \frac{1}{t}L(te_i) + \frac{1}{t}\|te_i\|\varepsilon(te_i)\\ &= L(e_i) + \frac{|t|}{t}\varepsilon(te_i).\end{aligned}$$

Comme ε tend vers 0 quand son argument tend vers 0, il suit que le taux d'accroissement converge vers $L(e_i)$, ce qu'il fallait démontrer. ∎

Avant de poursuivre, nous allons faire un détour pour répondre à une question évoquée précédemment, à savoir celle du calcul des dérivées partielles d'une fonction composée. Pour y voir plus clair, commençons par fixer les notations. Nous allons considérer une fonction $f : U \to \mathbf{R}^m$, où U est un ouvert de $\mathbf{R}^n$, et $g : V \to \mathbf{R}^p$, où V est un ouvert de $\mathbf{R}^m$. Pour que la composition ait un sens, nous allons supposer que $f(U) \subset V$.

Proposition 2.3.12 (Dérivées partielles composées)**.** *Supposons f et g de classe $\mathcal{C}^1$. Alors, avec les notations ci-dessus, $g \circ f$ est de classe $\mathcal{C}^1$, et sa différentielle en un point $x \in U$ est donnée par*

$$D_x(g \circ f) = \big(D_{f(x)}g\big) \circ (D_x f).$$

Par conséquent, en notant $(f_j)_{1\leqslant j\leqslant m}$ les fonctions coordonnées de f et $(g_k)_{1\leqslant k\leqslant p}$ les fonctions coordonnées de g, on a pour tout $x \in U$,

$$\frac{\partial (g\circ f)_k}{\partial x_i}(x) = \sum_{j=1}^{m} \frac{\partial g_k}{\partial x_j}(f(x))\,\frac{\partial f_j}{\partial x_i}(x).$$

Démonstration. L'idée de la preuve est de composer les développements limités des deux fonctions. Plus précisément, si $r > 0$ est tel que $B(x,r) \subset U$, alors pour tout $h \in B(0,r)$,

$$\begin{aligned}(g\circ f)(x+h) &= g\left(f(x+h)\right)\\ &= g\left(f(x) + D_x f(h) + \|h\|\varepsilon_1(\|h\|)\right).\end{aligned}$$

En posant $h' = D_x f(h) + \|h\|\varepsilon_1(\|h\|)$, on a alors

$$\begin{aligned} g(f(x) + D_x f(h) + \|h\|\varepsilon_1(\|h\|)) &= g(f(x) + h') \\ &= g(f(x)) + D_{f(x)}g(h') + \|h'\|\varepsilon_2(\|h'\|) \\ &= (g \circ f)(x) + \left(D_{f(x)}g\right) \circ (D_x f)(h) \\ &+ D_{f(x)}g(\|h\|\varepsilon_1(\|h\|)) + \|h\|\varepsilon_2(\|D_x f(h) + \|h\|\varepsilon_1(\|h\|)\|). \end{aligned}$$

Pour conclure, il faut montrer que les deux derniers termes sont de la forme $o(h)$. Pour

$$D_{f(x)}g(\|h\|\varepsilon_1(\|h\|)) = \|h\| D_{f(x)}g(\varepsilon_1(\|h\|)),$$

cela découle de la continuité [10] de $D_{f(x)}g$. Pour

$$\|h\|\varepsilon_2(\|D_x f(\|h\|) + \|h\|\varepsilon_1(\|h\|)\|),$$

cela suit de la continuité de $D_x f$.

Pour en déduire la formule concernant les dérivées partielles, il suffit d'écrire l'égalité avec des matrices. En effet, on a alors

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

et le résultat découle de la formule pour le produit de matrices. ■

Ce résultat vient avec un bonus : il permet également de comprendre ce qui se passe dans le cas de la réciproque d'une application bijective. Attention cependant, il y a une subtilité. Montrer que la réciproque d'une bijection $\mathcal{C}^1$ admet des dérivées partielles n'est pas facile (voir le Théorème B.5). Pour le moment, nous allons simplifier les choses en supposant que la réciproque est continue. Ce n'est pas grand-chose, mais cela simplifie la vie.

Proposition 2.3.13. *Soit $f : U \to \mathbf{R}^n$ une application bijective sur son image $f(U)$, de classe $\mathcal{C}^1$ et dont la réciproque est continue, et soit $x \in U$. Si $D_x f$ est inversible, alors la réciproque f^{-1} de f est de classe $\mathcal{C}^1$, et sa différentielle est donnée par*

$$D_{f(x)}(f^{-1}) = (D_x f)^{-1}.$$

Démonstration. Avant de démontrer formellement le résultat, nous allons vérifier que la formule proposée pour la différentielle de f^{-1} est bien cohérente. Supposons donc f^{-1} de classe $\mathcal{C}^1$, et considérons la composée $(f^{-1}) \circ f = \mathrm{Id} : U \to U$. En calculant la différentielle au point x avec la formule de la Proposition 2.3.12, on trouve

$$\begin{aligned} \mathrm{Id} &= D_x \mathrm{Id} \\ &= D_{f(x)}(f^{-1}) \circ D_x f, \end{aligned}$$

d'où l'égalité $D_{f(x)}(f^{-1}) = (D_x f)^{-1}$.

Passons maintenant à la démonstration. En fixant $y = f(x) \in f(U)$, nous devons montrer que la quantité

$$f^{-1}(y + h) - f^{-1}(y) - (D_x f)^{-1}(h)$$

10. N'oublions pas qu'en dimension finie, toutes les applications linéaires sont continues par le Théorème 2.1.9.

est un $o(h)$. Pour ce faire, remarquons que

$$\begin{aligned} y+h &= f(x)+h \\ &= f(x)+(D_xf)\circ(D_xf)^{-1}(h) \\ &= f(x+(D_xf)^{-1}(h))+o\Big((D_xf)^{-1}(h)\Big). \end{aligned}$$

De plus, par continuité de f^{-1}, on aura [11]

$$f^{-1}\Big(f(x+(D_xf)^{-1}(h))+o\Big((D_xf)^{-1}(h)\Big)\Big) = f^{-1}\Big(f(x+(D_xf)^{-1}(h))\Big)+o\Big(o\Big((D_xf)^{-1}(h)\Big)\Big).$$

Comme la différentielle (et donc son inverse) est linéaire, on peut enlever le $o(\cdot)$ extérieur, ce qui en réinjectant dans notre première égalité donne finalement

$$\begin{aligned} f^{-1}(y+h)-f^{-1}(y)-(D_xf)^{-1}(h) &= f^{-1}\Big(f(x+(D_xf)^{-1}(h))+o\Big((D_xf)^{-1}(h)\Big)\Big)-f^{-1}(y)-(D_xf)^{-1}(h) \\ &= f^{-1}\Big(f(x+(D_xf)^{-1}(h))\Big)+o\Big((D_xf)^{-1}(h)\Big)-f^{-1}(y)-(D_xf)^{-1}(h) \\ &= x+(D_xf)^{-1}(h)+o\Big((D_xf)^{-1}(h)\Big)-f^{-1}(y)-(D_xf)^{-1}(h) \\ &= x-f^{-1}(y)+o\Big((D_xf)^{-1}(h)\Big) \\ &= o\Big((D_xf)^{-1}(h)\Big). \end{aligned}$$

Il suffit maintenant d'observer que $\|(D_xf)^{-1}(h)\| \leqslant \|(D_xf)^{-1}\|\|h\|$ pour conclure. ∎

Il est bon de préciser un point qui est implicite dans l'énoncé précédent : comme la différentielle $D_xf : \mathbf{R}^n \to \mathbf{R}^m$ est une application linéaire, elle ne peut être bijective que si les espaces vectoriels de départ et d'arrivée sont isomorphes, autrement dit si $n = m$. Il n'y a donc pas de faute de frappe !

Second ordre

Maintenant que les développements limités sont pour nous un peu moins mystérieux, nous pouvons nous attaquer à l'ordre deux. Pour cela, il faut tout d'abord définir une notion de dérivée partielle d'ordre supérieur. Aucun souci : il suffit de prendre les dérivées partielles des dérivées partielles. Les notations s'étendent facilement à ce cadre : la dérivée partielle de $\dfrac{\partial f}{\partial x_i}$ par rapport à la variable x_j sera notée

$$\frac{\partial^2 f}{\partial x_j \partial x_i}.$$

Cela fait beaucoup de fonctions. Par exemple, dans le cas d'une fonction à valeurs scalaires $f : U \to \mathbf{R}$, on dispose de n dérivées partielles qui, étant chacune une fonction définie sur U, auront n dérivées partielles. Cela nous donne un total de n^2 fonctions pour décrire le comportement au second ordre. De même qu'il était commode d'organiser les dérivées partielles en une application linéaire – la différentielle – de même nous allons organiser les dérivées partielles secondes. Comme on dispose de deux familles d'indices, on peut utiliser deux vecteurs de $\mathbf{R}^n$ comme variables, ce qui donne

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

11. Il est important de se convaincre que la continuité d'une fonction F en un point x est équivalente à l'égalité $F(x+h) = F(x)+o(h)$, c'est-à-dire à un "développement limité à l'ordre 0".

On le voit, il est plus naturel de penser dans ce cadre à une application utilisant deux vecteurs de $\mathbf{R}^n$ qu'à une application sur $\mathbf{R}^{2n}$. De plus, l'application ci-dessus est linéaire en h quand k est fixé, et linéaire en k quand h est fixé. C'est donc une application *bilinéaire*.

Définition 2.3.14. Soit $f : U \to \mathbf{R}^m$ une fonction admettant des dérivées partielles secondes par rapport à toutes les coordonnées. Sa *différentielle seconde* au point x est l'application bilinéaire $D_x^2 f : \mathbf{R}^n \times \mathbf{R}^n \to \mathbf{R}^m$ définie par

$$\boxed{D_x^2 f(h,k) = \sum_{i,j=1}^{n} h_i k_j \frac{\partial^2 f}{\partial x_j \partial x_i}.}$$

Nous pouvons maintenant donner la forme du développement limité à l'ordre deux pour les fonctions de plusieurs variables. Pour cela, il nous faut une notion de fonction de classe $\mathcal{C}^2$, mais tant que nous y sommes, nous allons définir les fonctions de classe $\mathcal{C}^k$ pour tout entier k.

Définition 2.3.15. Une fonction $f : U \to \mathbf{R}^m$ est dite *de classe* $\mathcal{C}^k$ si elle admet des dérivées partielles et que ces dernières sont de classe $\mathcal{C}^{k-1}$. Si f est de classe $\mathcal{C}^k$ pour tout entier k, alors elle est dite *de classe* $\mathcal{C}^\infty$.

Toutes les propriétés de permanence usuelles sont encore valables ici. Résumons-les :

- Si $f, g : U \to \mathbf{R}^m$ sont de classe $\mathcal{C}^k$, alors $f+g$ est de classe $\mathcal{C}^k$.
- Si $f, g : U \to \mathbf{R}$ sont de classe $\mathcal{C}^k$, alors $f \times g$ est de classe $\mathcal{C}^k$.
- Si $f : U \to \mathbf{R}$ ne s'annule pas et est de classe $\mathcal{C}^k$, alors $1/f$ est de classe $\mathcal{C}^k$.
- Si $f : U \to \mathbf{R}^m$ et $g : V \to \mathbf{R}^p$ sont de classe $\mathcal{C}^k$ et $f(U) \subset V$, alors $g \circ f$ est de classe $\mathcal{C}^k$.

Théorème 2.3.16 (Développement limité au second ordre) Soit $f : U \to \mathbf{R}$ une fonction de classe $\mathcal{C}^2$ et soit $x \in U$. Alors, pour tout $h \in \mathbf{R}^n$ tel que $x + h \in U$,

$$\boxed{f(x+h) - f(x) = D_x f(h) + \frac{1}{2} D_x^2 f(h,h) + o(\|h\|^2).} \tag{2.2}$$

Démonstration. La preuve est très simple car elle repose sur une technique que nous utiliserons constamment : se ramener à une seule variable en prenant des *segments*. Bien sûr, comme dans la démonstration de la Proposition 2.3.9, il suffit de travailler avec un h suffisamment petit. Concrètement, soit $r > 0$ tel que $B(x,r) \subset U$. Alors, pour tout $t \in [0;1]$ et $h \in B(0,r)$, on a

$$\begin{aligned} \|x + th - x\| &= \|th\| \\ &= |t|\|h\| \\ &< r. \end{aligned}$$

Ainsi, on peut considérer la fonction

$$\phi : t \in [0;1] \mapsto f(x+th).$$

Si f est de classe $\mathcal{C}^2$, alors ϕ l'est également. Par conséquent, on a [12]

$$\phi(1) - \phi(0) = \phi'(0) + \int_0^1 (1-s)\phi''(s)\mathrm{d}s.$$

On a $\phi(1) = f(x+h)$ et $\phi(0) = f(x)$, donc pour conclure il nous faut maintenant calculer $\phi'(t)$ et $\phi''(t)$. On utilise pour cela la Proposition 2.3.12. En effet, ϕ est la composée de f et de la fonction $g : [0;1] \to \mathbf{R}^n$ définie par

$$g(t) = x + th,$$

dont la dérivée n'est autre que la fonction (vectorielle) constante $g' : t \mapsto h$. Ainsi,

$$\begin{aligned}\phi'(t) &= D_{g(t)}f\left(g'(t)\right)\\ &= D_{g(t)}f(h).\end{aligned}$$

Écrivons les choses un peu plus explicitement pour ne pas nous perdre :

$$\phi'(t) = \sum_{i=1}^n h_i \frac{\partial f}{\partial x_i}(g(t)),$$

donc

$$\begin{aligned}\phi''(t) &= \sum_{i=1}^n h_i D_{g(t)} \frac{\partial f}{\partial x_i}(g'(t))\\ &= \sum_{i=1}^n h_i D_{g(t)} \frac{\partial f}{\partial x_i}(h)\\ &= \sum_{i=1}^n h_i \left(\sum_{j=1}^n h_j \frac{\partial^2 f}{\partial x_j \partial x_i}(g(t)) \right)\\ &= \sum_{i,j=1}^n h_i h_j \frac{\partial^2 f}{\partial x_j \partial x_i}(g(t))\\ &= D^2_{g(t)}f\,(h,h)\\ &= D^2_{x+th}f(h,h).\end{aligned}$$

En remplaçant on trouve alors

$$\begin{aligned}f(x+h) - f(x) - D_x f(h) - \frac{1}{2}D_x^2 f(h,h) &= g(1) - g(0) - g'(0) - \frac{1}{2}g''(0)\\ &= \int_0^1 (1-s)D^2_{x+sh}f(h,h)\mathrm{d}s - \frac{1}{2}D_x^2 f(h,h)\\ &= \int_0^1 (1-s)D^2_{x+sh}f(h,h)\mathrm{d}s - \int_0^1 (1-s)D_x^2 f(h,h)\mathrm{d}s\\ &= \int_0^1 (1-s)\left(D^2_{x+sh}f(h,h) - D_x^2 f(h,h)\right)\mathrm{d}s\\ &= \sum_{i,j=1}^n h_i h_j \int_0^1 (1-s)\left(\frac{\partial^2 f}{\partial x_j \partial x_i}(x+sh) - \frac{\partial^2 f}{\partial x_j \partial x_i}(x)\right)\mathrm{d}s.\end{aligned}$$

12. Il s'agit de la Formule de Taylor avec reste intégral à l'ordre deux, mais l'égalité peut aussi se vérifier directement à l'aide d'une simple intégration par parties.

Soit $\epsilon > 0$. Par continuité des dérivées partielles secondes, il existe pour tous $1 \leqslant i, j \leqslant n$ un réel $\delta_{i,j} > 0$ tel que pour tout $y \in B(x, \delta_{i,j})$,

$$\left\| \frac{\partial^2 f}{\partial x_j \partial x_i}(y) - \frac{\partial^2 f}{\partial x_j \partial x_i}(x) \right\| \leqslant \epsilon.$$

Si $\delta = \min_{i,j} \delta_{i,j} > 0$, alors pour $\|h\| < \min(\delta, r)$ et $s \in [0;1]$, on a

$$x + sh \in B(x, \delta) \subset B(x, \delta_{ij}),$$

donc

$$\begin{aligned} \left\| f(x+h) - f(x) - D_x f(h) - \frac{1}{2} D_x^2 f(h,h) \right\| &\leqslant \sum_{i,j=1}^{n} |h_i||h_j| \int_0^1 (1-s)\epsilon \mathrm{d}s \\ &= \sum_{i,j=1}^{n} |h_i||h_j|\epsilon \\ &\leqslant n\|h\|^2 \epsilon, \end{aligned}$$

la dernière inégalité découlant de l'Inégalité de Cauchy-Schwarz comme dans la démonstration de la Proposition 2.3.9. Autrement dit, le membre de gauche est un $o(\|h\|^2)$, ce que nous voulions démontrer. ∎

Dernier ordre

À titre indicatif, nous allons conclure cette partie par une formule de développement limité à un ordre quelconque. Pour cela, il est clair qu'il faudra faire intervenir les dérivées partielles d'ordre supérieur de la fonction f, ce qui ne pose pas vraiment de problème. Pour alléger les formules, nous poserons

$$D_x^{(k)} f(h) = \sum_{i_1, \cdots, i_k = 1}^{n} h_{i_1} \cdots h_{i_k} \frac{\partial^k f}{\partial x_{i_1} \cdots \partial x_{i_k}}.$$

On a alors :

Proposition 2.3.17. *Soit $f : U \to \mathbf{R}^m$ une fonction de classe $\mathcal{C}^k$ et soit $x \in U$. Alors, pour tout $h \in \mathbf{R}^n$ tel que $x + h \in U$,*

$$\boxed{f(x+h) - f(x) = \sum_{j=0}^{k} \frac{1}{j!} D_x^{(j)} f(h) + o(\|h\|^k)}$$

Démonstration. La preuve est la même que pour le Théorème 2.3.16. Soit donc $r > 0$ tel que $B(x,r) \subset U$ et soit $h \in B(0,r)$. Alors, $x + h \in B(x,r) \subset U$ et de plus la fonction $g : [0;1] \to \mathbf{R}^m$ définie par

$$g(t) = f(x + th)$$

est bien définie et de classe $\mathcal{C}^k$. On peut donc lui appliquer la Formule de Taylor avec reste intégral :

$$g(1) - g(0) = \sum_{j=1}^{k-1} \frac{1}{j!} g^{(j)}(0) + \int_0^1 \frac{(1-s)^{k-1}}{(k-1)!} g^{(k)}(s) \mathrm{d}s.$$

Une récurrence simple permet de calculer les dérivées successives de g :

$$g^j(s) = D^j_{x+sh} f\,(h)$$

et on en déduit que

$$\begin{aligned} f(x+h) - f(x) - \sum_{j=1}^{k} \frac{1}{j!} D_x^{(j)} f(h) &= g(1) - g(0) - \sum_{j=1}^{k-1} \frac{1}{j!} g^{(j)}(0) - \frac{1}{k!} g^{(k)}(0) \\ &= \int_0^1 \frac{(1-s)^{k-1}}{(k-1)!} D^k_{x+sh}(h)\mathrm{d}s - \frac{1}{k!} D^k_x f(h) \\ &= \int_0^1 \frac{(1-s)^{k-1}}{(k-1)!} \Big(D^k_{x+sh}(h) - D^k_x f(h) \Big) \mathrm{d}s \\ &= \sum_{i_1,\cdots,i_k=1}^{n} h_{i_1} \cdots h_{i_k} \\ &\int_0^1 \frac{(1-s)^{k-1}}{(k-1)!} \left(\frac{\partial^k f}{\partial x_{i_1} \cdots \partial x_{i_k}}(x+sh) - \frac{\partial^k f}{\partial x_{i_1} \cdots \partial x_{i_k}}(x) \right) \mathrm{d}s. \end{aligned}$$

On montre enfin – en utilisant la continuité des dérivées partielles d'ordre k – que ce terme est bien négligeable devant $\|h\|^k$, exactement de la même façon que dans la démonstration du Théorème 2.3.16. ■

2.3.3 Condition suffisante d'extremum

Pour une fonction d'une variable, la condition nécessaire pour qu'un point critique soit un extremum est que la dérivée seconde en ce point soit non nulle. Mais ce qui est plus important, comme nous l'avons rappelé dans la Proposition 2.3.3, c'est que le signe de la dérivée seconde permet alors de savoir si l'on a un minimum ou un maximum. Dans le cas de plusieurs variables, la notion de signe n'est pas claire puisque la différentielle seconde $D^2_x f$ est maintenant une forme bilinéaire de $\mathbf{R}^n \times \mathbf{R}^n$ dans $\mathbf{R}$. Cela dit, il existe précisément une notion de positivité pour les formes bilinéaires.

En effet, si $B : \mathbf{R}^n \times \mathbf{R}^n \to \mathbf{R}$ est bilinéaire, on peut lui associer une *forme quadratique* $Q : \mathbf{R}^n \to \mathbf{R}$ par la formule

$$Q(h) = B(h,h).$$

C'est exactement la quantité qui apparaît dans le développement limité à l'ordre deux, c'est donc cette forme quadratique qu'il faut comprendre. Dans certains cas, elle est de signe constant, et l'on peut alors dire quelle est positive ou négative. Nous aurons néanmoins besoin d'une notion légèrement plus forte, que nous rappelons.

Définition 2.3.18. Une forme quadratique $Q : \mathbf{R}^n \to \mathbf{R}$ est dite *définie positive*[13] s'il existe $\lambda > 0$ tel que

$$Q(h) \geqslant \lambda \|h\|^2$$

pour tout $h \in \mathbf{R}^n$. Elle est dite *définie négative* s'il existe $\mu > 0$ tel que

$$Q(h) \leqslant -\mu \|h\|^2$$

pour tout $h \in \mathbf{R}^n$.

13. La définition que nous donnons ici est en fait celle d'une forme quadratique *coercive*. Cela dit, en dimension finie, les deux notions sont équivalentes et nous nous permettrons donc cet abus terminologique.

Avec cette notion, la stratégie pour les fonctions d'une variable expliquée plus haut s'étend sans difficulté.

Théorème 2.3.19 (Condition suffisante d'extrema) Soit $f : U \to \mathbf{R}$ une fonction de classe $\mathcal{C}^2$ et soit x un point critique de f. Si la forme quadratique

$$\boxed{Q_x : h \mapsto D_x^2 f(h,h)}$$

est définie positive, alors f admet un minimum local strict en x. De même, si Q_x est définie négative, alors f admet un maximum local strict en x.

Démonstration. Supposons f définie positive, l'autre cas se traitant avec la même preuve, *mutatis mutandis*. Soit $r > 0$ tel que $B(x,r) \subset U$ et h tel que $0 < \|h\| < r$. Comme x est un point critique, le premier terme du développement limité de f en x s'annule et on a donc par le Théorème 2.3.16

$$\begin{aligned} f(x+h) - f(x) &= \frac{1}{2} D_x^2 f(h,h) + o(\|h\|^2) \\ &= \frac{1}{2} Q_x(h) + \|h\|^2 \varepsilon(h) \\ &= \|h\|^2 \left(\frac{1}{2\|h\|^2} Q_x(h) + \varepsilon(h) \right) \\ &\geqslant \|h\|^2 \left(\frac{\lambda}{2} + \varepsilon(h) \right) \end{aligned}$$

La preuve se conclut maintenant exactement comme dans le cas d'une variable, mais nous allons quand même donner les détails. Le terme entre parenthèses dans le membre de droite tend vers $\lambda/2$ quand $h \to 0$, donc il existe $\delta > 0$ tel que

$$\frac{\lambda}{2} + \varepsilon(h) \geqslant \frac{\lambda}{4}$$

dès que $\|h\| < r' = \min(\delta, r)$. Alors, si $y \in B(x,r')$, on a en posant $h = y - x$ que $\|h\| < r'$ et donc que

$$\begin{aligned} f(y) - f(x) &= f(x+h) - f(x) \\ &\geqslant \|h\|^2 \frac{\lambda}{4} \\ &> 0. \end{aligned}$$

Autrement dit, f admet un minimum local strict en x, ce qu'il fallait démontrer. ∎

L'intérêt du résultat précédent est qu'il fournit une condition *suffisante* pour déterminer la nature d'un point critique : si la condition est vérifiée, alors c'est une extremum local. Il existe également une condition nécessaire – mais non suffisante – utilisant le développement limité à l'ordre deux. Bien qu'elle soit peu utile en pratique, nous la donnons maintenant. Une forme quadratique $Q : \mathbf{R}^n \to \mathbf{R}$ est dite *positive* si $Q(h) \geqslant 0$ pour tout $h \in \mathbf{R}^n$ et négative si $Q(h) \leqslant 0$ pour tout $h \in \mathbf{R}^n$.

Proposition 2.3.20. *Soit $f : U \to \mathbf{R}$ une fonction de classe $\mathcal{C}^2$ et soit x un point critique de f. Si f admet un minimum local en x, alors la forme quadratique Q_x est positive. Si f admet un maximum local en x, alors la forme quadratique Q_x est négative.*

Démonstration. Le raisonnement commence comme dans la preuve du Théorème 2.3.19. Il existe $r > 0$ tel que pour tout $h \in B(0,r)$,

$$f(x+h) - f(x) = \|h\|^2 \left(\frac{Q_x(h,h)}{\|h\|^2} + \varepsilon(h) \right).$$

Si x est un minimum local, cette quantité est positive, donc

$$\frac{Q_x(h,h)}{\|h\|^2} + \varepsilon(h) \geqslant 0.$$

Pour exploiter cette inégalité, nous allons recourir à une astuce qui consiste à écrire h sous la forme th' avec $t \in [0;1]$. En effet, si $h' \in B(0,r)$ et $t \in [0;1]$, alors $h = th' \in B(0,r)$ et l'inégalité précédente devient

$$\begin{aligned} 0 &\leqslant \frac{Q_x(th',th')}{\|th'\|^2} + \varepsilon(th') \\ &= \frac{Q_x(h',h')}{\|h'\|^2} + \varepsilon(th'). \end{aligned}$$

En faisant tendre t vers 0, $\varepsilon(th') \to 0$ et on obtient donc

$$\frac{Q_x(h',h')}{\|h'\|^2} \geqslant 0,$$

d'où le résultat. ∎

Entre les deux conditions que nous venons de présenter, il y a encore plusieurs possibilités. On peut parfois affirmer qu'il n'y a pas d'extremum local en un point (voir la Proposition 2.3.25 ci-après), mais le plus souvent on ne peut déterminer la nature du point critique en considérant simplement les dérivées partielles secondes. On parle alors de *cas dégénéré*, et il faudrait pousser le développement limité plus loin pour tâcher de comprendre ce qu'il se passe, mais il n'y a pas de critère général simple.

2.3.4 Lien avec les dérivées partielles

Le Théorème 2.3.19 nous donne un critère pour vérifier qu'un point critique est un extremum. Cela dit, ce critère n'est pas très explicite puisqu'il faut déterminer une propriété de la forme quadratique Q_x, et qu'il n'est pas humainement possible de calculer $Q_x(h)$ pour tous les $h \in \mathbf{R}^n$ afin d'en vérifier le signe. Comment s'y prendre alors ? Il faut relier Q_x aux dérivées partielles !

Une question de symétrie

Nous avons vu qu'une forme quadratique est définie à partir d'une forme bilinéaire B. Cette dernière peut se représenter à l'aide d'un matrice $M(B)$ dont les coefficients sont, en notant $(e_i)_{1 \leqslant i \leqslant n}$ la base canonique de $\mathbf{R}^n$,

$$M(B)_{ij} = B(e_i, e_j).$$

Ceci permet bien de retrouver B puisque étant donnés deux vecteurs x et x' on aura par bilinéarité

$$\begin{aligned} B(x,x') &= B\left(\sum_{i=1}^n x_i e_i, \sum_{j=1}^n x_j e_j\right) \\ &= \sum_{i,j=1}^n x_i x'_j B(e_i,e_j) \\ &= \sum_{i,j=1}^n M(B)_{ij} x_i x'_j. \end{aligned}$$

Une autre façon d'écrire cette égalité est

$$B(x,x') = \langle M(B)x', x\rangle$$

et par conséquent

$$Q(x) = \langle M(B)x, x\rangle.$$

Sous cette dernière forme, on voit que la condition d'être définie positive ou négative pourrait être obtenue en comprenant les valeurs propres de $M(B)$. En effet, admettons qu'il existe une base orthonormée $\mathcal{B} = (v_1, \cdots, v_n)$ de $\mathbf{R}^n$ dont tous les vecteurs sont propres pour $M(B)$, c'est-à-dire qu'il existe $\lambda_1, \cdots, \lambda_n \in \mathbf{R}$ tels que pour tout $1 \leqslant i \leqslant n$,

$$M(B)v_i = \lambda_i v_i.$$

Alors, pour tout $x = \sum_i x_i v_i$,

$$\begin{aligned} Q(x) &= \langle M(B)x, x\rangle \\ &= \left\langle M(B)\left(\sum_{i=1}^n x_i v_i\right), \sum_{i=1}^n x_i v_i\right\rangle \\ &= \sum_{i,j=1}^n x_i x_j \left\langle M(B)v_i, v_j\right\rangle \\ &= \sum_{i,j=1}^n x_i x_j \left\langle \lambda_i v_i, v_j\right\rangle \\ &= \sum_{i=1}^n x_i^2 \lambda_i. \end{aligned}$$

En particulier, si toutes les valeurs propres de $M(B)$ sont strictement négatives, alors Q est définie négative et si toutes ses valeurs propres sont strictement positives, alors Q est définie positive.

Le raisonnement précédent n'a qu'une faiblesse : comment assurer que $M(B)$ admet une base orthonormée de vecteurs propres ? L'algèbre linéaire nous fournit heureusement une réponse : il suffit que la matrice $M(B)$ soit *symétrique*. L'un des miracles du calcul différentiel est que c'est (presque) toujours le cas ! Ce résultat est généralement connu sous le nom de Théorème de Schwarz [14] et nous allons maintenant l'énoncer et le démontrer.

14. Hermann Schwarz (1843–1921) : mathématicien allemand, élève de Weierstrass, qui a contribué notablement à l'analyse à une et plusieurs variables, réelles ou complexes, et à la géométrie différentielle. On lui doit la version générale du l'inégalité dont il partage le nom avec Augustin-Louis Cauchy (1789–1857).

Théorème 2.3.21 (Théorème de Schwarz) Soit $f : U \to \mathbf{R}^m$ une fonction de classe $\mathcal{C}^2$ et soit $x \in U$. Alors, pour tout $1 \leqslant i, j \leqslant n$,

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

Démonstration. Nous allons commencer par simplifier légèrement le problème en nous ramenant à seulement deux variables. Fixons un point $x \in U$ ainsi que deux indices $i < j$ compris entre 1 et n. Si $r > 0$ est tel que $B(x, r) \subset U$, et si

$$(t_1, t_2) \in B((x_i, x_j), r),$$

alors

$$\left\|(x_1, \cdots, x_{i-1}, t_1, x_{i+1}, \cdots, x_{j-1}, t_2, x_{j+1}, \cdots, x_n) - (x_1, \cdots, x_n)\right\| = \sqrt{t_1^2 + t_2^2} < r$$

Ainsi, en posant $V = B((x_i, x_j), \sqrt{r/2}) \subset \mathbf{R}^2$, on peut considérer la fonction $F : V \to \mathbf{R}$ définie par

$$F : (t_1, t_2) \mapsto f(x_1, \cdots, x_{i-1}, t_1, x_{i+1}, \cdots, x_{j-1}, t_2, x_{j+1}, \cdots, x_n).$$

Cette fonction est bien évidemment de classe $\mathcal{C}^2$.

On considère maintenant un rectangle $[a; b] \times [c; d]$ contenu dans V. Alors,

$$\begin{aligned}\int_a^b \left(\int_c^d \frac{\partial^2 F}{\partial t_2 \partial t_1}(s_1, s_2) \mathrm{d}s_2 \right) \mathrm{d}s_1 &= \int_a^b \frac{\partial F}{\partial t_1}(s_1, d) - \frac{\partial F}{\partial t_1}(s_1, c) \mathrm{d}s_1 \\ &= F(b, d) - F(a, d) - F(b, c) + F(a, c)\end{aligned}$$

et

$$\begin{aligned}\int_c^d \left(\int_a^b \frac{\partial^2 F}{\partial t_1 \partial t_2}(s_1, s_2) \mathrm{d}s_1 \right) \mathrm{d}s_2 &= \int_c^d \frac{\partial F}{\partial t_2}(b, s_2) - \frac{\partial F}{\partial t_2}(a, s_2) \mathrm{d}s_2 \\ &= F(b, d) - F(a, d) - F(b, c) + F(a, c) \\ &= \int_a^b \left(\int_c^d \frac{\partial^2 F}{\partial t_2 \partial t_1}(s_1, s_2) \mathrm{d}s_2 \right) \mathrm{d}s_1.\end{aligned}$$

Pour poursuivre le calcul, il suffit de remarquer que les intégrales peuvent être échangées : c'est une version élémentaire [15] du Théorème de Fubini. On obtient alors

$$\begin{aligned}\int_a^b \left(\int_c^d \frac{\partial^2 F}{\partial t_2 \partial t_1}(s_1, s_2) \mathrm{d}s_2 \right) \mathrm{d}s_1 &= \int_c^d \left(\int_a^b \frac{\partial^2 F}{\partial t_1 \partial t_2}(s_1, s_2) \mathrm{d}s_1 \right) \mathrm{d}s_2 \\ &= \int_a^b \left(\int_c^d \frac{\partial^2 F}{\partial t_1 \partial t_2}(s_1, s_2) \mathrm{d}s_2 \right) \mathrm{d}s_1.\end{aligned}$$

Posons

$$G : (s_1, s_2) \mapsto \frac{\partial^2 F}{\partial t_2 \partial t_1}(s_1, s_2) - \frac{\partial^2 F}{\partial t_1 \partial t_2}(s_1, s_2)$$

15. On peut démontrer facilement ce résultat dans le cadre de l'intégrale de Riemann, voir l'Appendice A.

et supposons que $G(x_i, x_j) \neq 0$. Si par exemple $G(x_i, x_j) > 0$, alors il existe par continuité de G un réel $\delta > 0$ tel que $B((x_i, x_j), \delta) \subset V$ et

$$G(t_1, t_2) \geqslant \frac{G(x_i, x_j)}{2} > 0$$

pour tout $(t_1, t_2) \in B((x_i, x_j), \delta)$. Comme $B((x_i, x_j), \delta)$ est ouverte, elle contient un rectangle $[a; b] \times [c; d]$ contenant (x_i, x_j). Or, le calcul précédent montre que

$$\int_c^d \int_a^b G(s_1, s_2) \mathrm{d}s_1 \mathrm{d}s_2 = 0.$$

Comme G est continue et de signe constant, cela implique $G = 0$ et donc en particulier $G(x_i, x_j) = 0$, ce qui donne enfin le résultat. ∎

Remarque 2.3.22. On peut étendre par récurrence ce résultat aux dérivées partielles d'ordre supérieur de la façon suivante : pour tous $1 \leqslant i_1, \cdots, i_k \leqslant n$ et toute permutation $\sigma \in \mathfrak{S}_k$, on a

$$\frac{\partial^k f}{\partial x_{i_1} \cdots \partial x_{i_k}} = \frac{\partial^k f}{\partial x_{i_{\sigma(1)}} \cdots \partial x_{i_{\sigma(k)}}}.$$

Armés de ce résultat, nous pouvons appliquer la méthode expliquée au début de cette partie pour caractériser le fait que la différentielle seconde soit définie positive ou négative. Profitons de l'occasion pour donner une définition supplémentaire.

Définition 2.3.23. Soit $f : U \to \mathbf{R}$ une fonction de classe $\mathcal{C}^2$. La matrice de la forme quadratique $D_x^2 f$ dans la base canonique est appelée *matrice Hessienne de f au point x* et notée $H_f(x)$. Il s'agit d'une matrice symétrique dont les coefficients sont

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

Mentionnons pour l'anecdote que bien que le nom de "Hessienne" fasse référence à Hesse [16], il a en fait été introduit par Sylvester [17]. Bref, on a maintenant :

Proposition 2.3.24. *Soit $f : U \to \mathbf{R}$ une fonction de classe $\mathcal{C}^2$. Pour $x \in U$, la forme quadratique $D_x^2 f$ est*

- *Définie positive si et seulement si toutes les valeurs propres de $H_f(x)$ sont strictement positives.*
- *Définie négative si et seulement si toutes les valeurs propres de $H_f(x)$ sont strictement négatives.*

Et si on n'est dans aucune de ces deux situations ? Il y a alors deux cas de figure possibles. Si les valeurs propres sont toutes de même signe, mais que ce signe n'est pas strict, alors on ne peut rien dire : c'est un cas dégénéré. S'il y a des valeurs propres de signes opposés, alors le point critique n'est pas un extremum. Montrons cette dernière assertion.

16. Ludwig Hesse (1811–1874) : mathématicien allemand qui malgré l'objet qui lui est attaché en calcul différentiel a surtout travaillé en algèbre, notamment sur la théories des invariants.

17. James Sylvester (1814 - 1897) : mathématicien anglais qui a contribué de façon importante au développement de l'algèbre (linéaire) moderne et entre autres des formes quadratiques. On lui doit l'introduction de plusieurs termes et notations des mathématiques modernes, et en particulier du mot "matrice".

Proposition 2.3.25. *Soit $f : U \to \mathbf{R}$ une fonction de classe $\mathcal{C}^2$ et soit $x \in U$ un point critique de f. Si $H_f(x)$ a au moins une valeur propre strictement positive et une valeur propre strictement négative, alors f n'a pas d'extremum local au point x.*

Démonstration. Soit $\lambda > 0$ une valeur propre de $H_f(x)$ et soit $v \in \mathbf{R}^n$ un vecteur propre associé, qu'on peut choisir de sorte que $\|v\| = 1$. Si $r > 0$ est tel que $B(x,r) \subset U$, alors pour tout $t \in]-r;r[$, $x + tv \in U$ et

$$\begin{aligned} f(x+tv) - f(x) &= D_x f(tv) + \frac{1}{2} D_x^2 f(tv, tv) + o(\|tv\|^2) \\ &= \frac{\lambda t^2}{2} + o(t^2) \\ &= t^2 \left(\frac{\lambda}{2} + \varepsilon(t) \right). \end{aligned}$$

Le même argument que dans la preuve du Théorème 2.3.19 montre alors qu'il existe $r' > 0$ tel que pour tout $t \in]-r';r'[$,

$$f(x+tv) - f(x) > 0.$$

Si maintenant $\mu < 0$ est une autre valeur propre et w un vecteur propre associé de norme 1. Le même raisonnement montre qu'il existe $r'' > 0$ tel que pour tout $t \in]-r'';r''[$,

$$f(x+tw) - f(x) < 0.$$

Ainsi, on peut trouver des points aussi proches qu'on veut de x auxquels f prend des valeurs plus grandes et plus petites que $f(x)$. Autrement dit, f n'a pas d'extremum local au point x. ∎

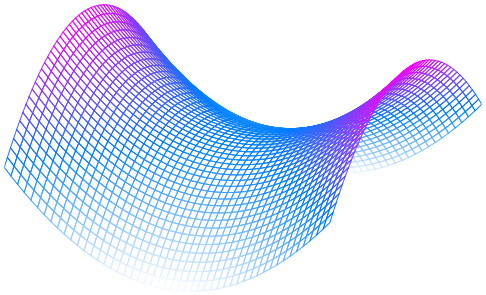
Ce cas mérite un nom, et donc un définition.

Définition 2.3.26. Soit $f : U \to \mathbf{R}$ une fonction de classe $\mathcal{C}^2$ et soit $x \in U$ un point critique de f. Si $H_f(x)$ a au moins une valeur propre strictement positive et une valeur propre strictement négative, alors x est appelé *point selle* [18].

Exemple 2.3.27. On considère la fonction $f : \mathbf{R}^2 \to \mathbf{R}$ définie par $f(x) = x^2 - y^2$. Un calcul immédiat donne

$$\nabla f(x,y) = \begin{pmatrix} 2x \\ -2y \end{pmatrix} \quad \& \quad H_f(x,y) = \begin{pmatrix} 2 & 0 \\ 0 & -2 \end{pmatrix}$$

On en déduit que $(0,0)$ est un point critique. De plus, $H_f(0,0)$ a pour valeurs propres 2 et -2, donc l'origine est un point selle. Si l'on dessine la surface d'équation $z = f(x,y)$, on voit apparaître un *paraboloïde hyperbolique* dont la forme en "selle de cheval" illustre bien le concept :



18. Il existe une terminologie concurrente, à savoir *point col*. Libre à chacun de choisir selon ses goûts entre l'équitation et l'alpinisme.

Concluons par une remarque qui, bien qu'anodine en apparence, cache un phénomène mathématique fascinant : pour vérifier qu'on a un extremum local, il suffit de connaître le comportement de la fonction dans un nombre fini de directions. De plus, ces directions sont nécessairement orthogonales les unes aux autres, ce qui facilite le travail !

Le cas de deux variables

Pour conclure cette partie, nous allons traduire le critère précédent de façon plus concrète pour des fonctions de deux variables. En effet, dans ce cas on peut facilement déterminer la nature d'un point singulier à l'aide des dérivées partielles. Pour voir comment, remarquons tout d'abord que la matrice Hessienne, étant symétrique, s'écrit simplement

$$H_f(x) = \begin{pmatrix} a & b \\ b & a \end{pmatrix}$$

Son polynôme caractéristique est

$$P(X) = X^2 - 2aX + a^2 - b^2,$$

dont les racines sont

$$a \pm |b|.$$

Comme les racines du polynôme caractéristique sont exactement les valeurs propres, on voit qu'il suffit de comparer les valeurs de a et $|b|$.

On peut également procéder autrement, en partant de l'observation que la somme des racines est égale à

$$2a = \mathrm{Tr}(H_f(x))$$

et que leur produit est égal à

$$a^2 - b^2 = \det(H_f(x)).$$

Puisque seul le signe des valeurs propres nous intéresse, le signe de ces deux dernière quantités doit suffire. Et en effet,

Proposition 2.3.28. *Soit U un ouvert de $\mathbf{R}^2$ et $f : U \to \mathbf{R}$ une fonction de classe $\mathcal{C}^2$. Soit $x \in U$ un point critique de f. Alors,*

1. *Si $\det(H_f(x)) < 0$, alors x est un point selle.*
2. *Si $\det(H_f(x)) > 0$ et $\mathrm{Tr}(H_f(x)) > 0$, alors x est un minimum local.*
3. *Si $\det(H_f(x)) > 0$ et $\mathrm{Tr}(H_f(x)) < 0$, alors x est un maximum local.*

Démonstration. Il suffit, comme expliqué plus haut, de rapprocher chaque information du signe des valeurs propres.

1. Si le produit des valeurs propres est strictement négatif, alors elles sont de signes opposés.
2. Si le produit des valeurs propres est strictement positif, alors elles sont de même signe. Ce signe est le même que leur somme, d'où les deux derniers cas.
3. Voir ci-dessus.

■

On peut même faire encore un peu plus simple, en utilisant ce que l'histoire a retenu sous le nom de *notations de Monge* [19]. Il s'agit simplement de poser

$$r = \frac{\partial^2 f}{\partial x^2} \quad ; \quad s = \frac{\partial^2 f}{\partial y \partial x} \quad ; \quad t = \frac{\partial^2 f}{\partial y^2}.$$

Alors, on peut donner une alternative à la Proposition 2.3.28 de la façon suivante :

Proposition 2.3.29. *Soit $f : U \to \mathbf{R}$ une fonction de classe $\mathcal{C}^2$ et soit $x \in U$ un point critique. Avec les notations ci-dessus,*

- *Si $rt - s^2 < 0$, alors on a un point selle ;*
- *Si $rt - s^2 > 0$ et $r > 0$ alors on a un minimum local ;*
- *Si $rt - s^2 > 0$ et $r < 0$ alors on a un maximum local.*

Démonstration. Le premier cas a déjà été traité dans la Proposition 2.3.28. De plus, si $rt - s^2 > 0$, alors on sait déjà que $H_f(x)$ est soit définie positive, soit définie négative. Comme

$$Q_x(h_1, 0) = rh_1^2,$$

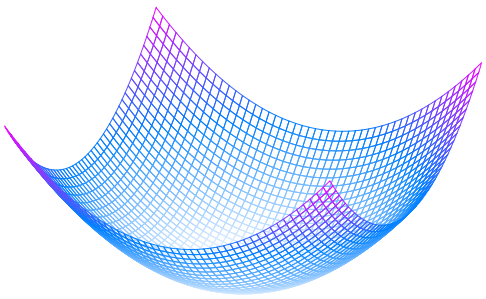
le signe de r suffit à trancher. ■

Pour conclure, nous allons donner quelques exemples. Le cas d'un point selle a déjà été illustré dans l'Exemple 2.3.27, nous allons donc donner un exemple où les deux valeurs propres sont de même signe strict.

Exemple 2.3.30. On considère la fonction $f : \mathbf{R}^2 \to \mathbf{R}$ définie par

$$f(x, y) = x^2 + y^2.$$

Alors, l'unique point critique est $(0, 0)$. De plus, la Hessienne $H_f(0, 0)$ n'est autre que $2\mathrm{I}_2$. On a donc un minimum local en ce point. On peut alors vérifier directement que ce minimum est global.



Nous poursuivons avec deux exemples de point critique dégénéré.

19. Gaspard Monge (1746 – 1818) : mathématicien français qui fut l'un des plus importants de son époque. Il a beaucoup travaillé en analyse et en géométrie. Il est aussi l'un des fondateurs de l'*École Polytechnique*.

Exemple 2.3.31. On considère la fonction $f : \mathbf{R}^2 \to \mathbf{R}$ définie par

$$f(x,y) = x^4 + y^3 - 3y - 2.$$

On a

$$\nabla f(x,y) = \begin{pmatrix} 4x^3 \\ 3y^2 - 3 \end{pmatrix}$$

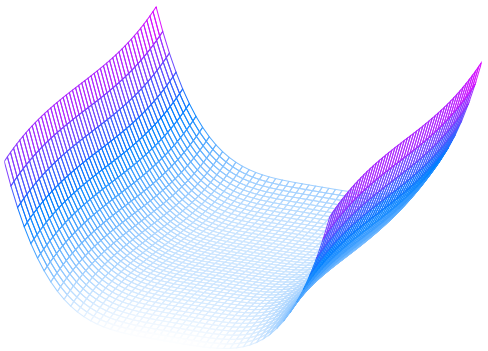
qui s'annule en $(0,1)$ et en $(0,-1)$. De plus,

$$H_f(0,\pm 1) = \begin{pmatrix} 0 & 0 \\ 0 & \pm 6 \end{pmatrix}$$

qui a toujours une valeur propre nulle et on ne peut donc pas conclure. Néanmoins, il est possible de déterminer la nature de ces points critiques en poussant le développement limité un peu plus loin. En effet, on a, pour $(h,k) \in \mathbf{R}^2$,

$$\begin{aligned} f(h,1+k) - f(0,1) &= h^4 + 1 + 3k + 3k^2 + k^3 - 3 - 3k - 2 - (-4) \\ &= h^4 + 3k^2 + k^3 \\ &= h^4 + k^2(3+k). \end{aligned}$$

Pour $|k| < 3$, cette quantité est positive, donc on a bien un minimum local. Cependant, comme $f(0,-n)$ tend vers $-\infty$, ce n'est pas un minimum global.



Exemple 2.3.32. On considère la fonction $f : \mathbf{R}^2 \to \mathbf{R}$ définie par

$$f(x,y) = x^2 + y^3.$$

À nouveau, l'unique point critique est $(0,0)$. Cependant, on a maintenant

$$H_f(0,0) = \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix}.$$

Comme cette matrice a une valeur propre nulle, on ne peut pas conclure. De fait, en observant que $f(0,y) < 0$ pour tout $y < 0$ tandis que $f(x,0) > 0$ pour tout x, on voit que l'origine n'est ni un maximum local ni un minimum local.

2.4 La vraie vie

One's real life is so often the life that one does not lead.

O. Wilde, *L'envoi* in Rose leaf and apple leaf

Les résultats précédents sont satisfaisants mais restent très théoriques. En pratique, comment peut-on trouver les extrema d'une fonction ? Par exemple, pour trouver les points critiques il faut résoudre l'équation $\nabla f(x) = 0$, mais il s'agit d'un système de n équations qui ne sont en général pas linéaires, donc pour lesquelles il n'existe pas nécessairement de méthode de résolution exacte. Il faut par conséquent chercher des solutions approchées à l'aide de méthodes numériques. Nous allons en donner ici un exemple, dont nous discuterons la pertinence et les limites.

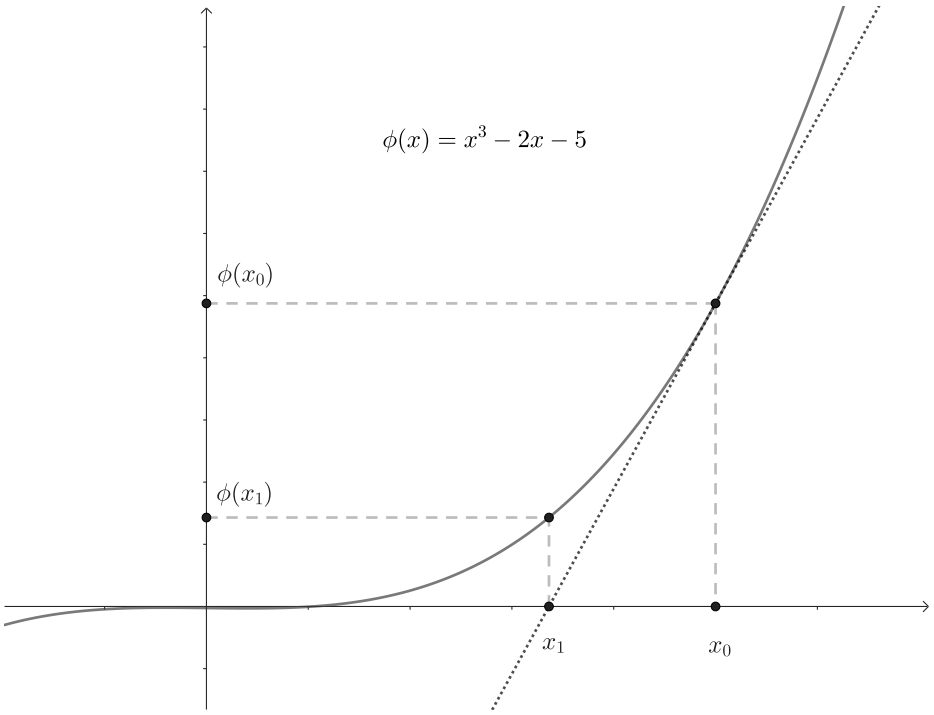
2.4.1 Illustration avec une variable

Le problème numérique qui nous intéresse est de trouver la solution à une équation de la forme $\phi(x) = 0$, où $\phi : U \to \mathbf{R}^n$ est une fonction de n variables. En effet, il suffira ensuite d'appliquer la méthode à la fonction $\phi = \nabla f$.

Commençons par considérer une fonction $\phi : I \to \mathbf{R}$, disons de classe $\mathcal{C}^2$, pour laquelle on veut résoudre l'équation $\phi(x) = 0$. L'idée est de partir d'un point $x_0 \in I$ et de "suivre la pente" de ϕ. Autrement dit, si $\phi(x_0) > 0$ et ϕ est croissante au voisinage de x_0, alors son zéro devrait se trouver "à gauche" de x_0 tandis que si ϕ est décroissante, son zéro devrait se trouver "à droite". Plus précisément, l'équation de la tangente au point x_0 est

$$y = \phi'(x_0) + (x - x_0)\phi'(x_0)$$

et nous pouvons donc suivre la tangente jusqu'à trouver son intersection x_1 avec l'axe des abscisses. Celui-ci sera alors "du bon côté" de x_0. Voici une illustration sur un exemple utilisé par l'inventeur de cette méthode, qui n'est autre que Newton [20] :



20. Si la méthode que nous allons exposer est généralement appelée Méthode de Newton et a été effectivement décrite par ce dernier dans un ouvrage publié en 1711 mais écrit en 1669, une version simplifiée en avait déjà été donnée en 1690 par J. Raphson. On parle parfois pour cette raison de Méthode de Newton-Raphson.

Formellement, il suffit de trouver x tel que $y = 0$ dans l'équation de la tangente, ce qui donne

$$x_1 = x_0 - \frac{\phi(x_0)}{\phi'(x_0)}.$$

Un autre façon de comprendre la définition de x_1 est la suivante : s'il existe $\widetilde{x}$ tel que $\phi(\widetilde{x}) = 0$, alors le développement limité de ϕ au point x_0 au premier ordre nous dit que $\phi(\widetilde{x})$ est à peu près égal à

$$\phi(x_0) + (x_0 - \widetilde{x})\phi'(x_0).$$

Si ce développement était exact, comme $\phi(\widetilde{x}) = 0$ on aurait alors $\widetilde{x} = x_1$.

Remarquons que si $\phi'(x_0) = 0$, alors la définition n'a guère de sens. Cela signifie que la méthode ne fonctionnera que si l'on peut garantir que ϕ' ne s'annule pas, au moins sur un voisinage de $\widetilde{x}$. On construit alors par récurrence une suite $(x_k)_{k\in\mathbf{N}}$ en posant

$$x_{k+1} = x_k - \frac{\phi(x_k)}{\phi'(x_k)}$$

Pour élémentaire que soit cette idée, elle est efficace en un certain sens que nous allons maintenant préciser. Avant cela, remarquons que si la suite $(x_k)_{k\in\mathbf{N}}$ converge vers une limite $\widetilde{x}$, cette limite est nécessairement un zéro de ϕ. En effet, en passant à la limite dans la relation de récurrence on trouve

$$\widetilde{x} = \widetilde{x} - \frac{\phi(\widetilde{x})}{\phi'(\widetilde{x})},$$

ce qui donne bien $\phi(\widetilde{x}) = 0$.

Proposition 2.4.1 (Méthode de Newton à une variable)**.** *Soit $\phi : I \to \mathbf{R}$ une fonction de classe $\mathcal{C}^2$ et $\widetilde{x} \in I$ tel que $\phi(\widetilde{x}) = 0$ et $\phi'(\widetilde{x}) \neq 0$. Alors, il existe $\delta > 0$ tel que pour tout $x_0 \in]\widetilde{x} - \delta; \widetilde{x} + \delta[$, la suite $(x_k)_{k\in\mathbf{N}}$ converge vers $\widetilde{x}$.*

Démonstration. Tout d'abord, il existe $r > 0$ tel que pour tout $x \in [\widetilde{x} - r; \widetilde{x} + r] \subset I$, $\phi'(x) \neq 0$. On utilise maintenant la Formule de Taylor-Lagrange à l'ordre deux pour la fonction ϕ, qui donne un élément c_k situé entre $\widetilde{x}$ et x_k tel que

$$\phi(\widetilde{x}) - \phi(x_k) = (\widetilde{x} - x_k)\phi'(x_k) + \frac{(\widetilde{x} - x_k)^2}{2}\phi''(c_k).$$

On en déduit, comme $\phi(\widetilde{x}) = 0$, que

$$x_k - \widetilde{x} = \frac{\phi(x_k)}{\phi'(x_k)} + \frac{\phi''(c_k)}{2\phi'(x_k)}(x_k - \widetilde{x})^2.$$

L'astuce pour poursuivre est de remarquer que

$$\frac{\phi(x_k)}{\phi'(x_k)} = x_k - x_{k+1}.$$

En effet, il découle de cette égalité et de la précédente que

$$\begin{aligned} |x_{k+1} - \widetilde{x}| &= |(x_{k+1} - x_k) + (x_k - \widetilde{x})| \\ &= \left|\frac{\phi''(c_k)}{2\phi'(x_k)}(x_k - \widetilde{x})^2\right|. \end{aligned}$$

La fonction ϕ' est continue sur $[\widetilde{x}-r;\widetilde{x}+r]$ – qui est compact – et ne s'y annule pas, donc elle a un minimum $m>0$. De plus, la fonction ϕ'' est continue sur ce même intervalle, donc elle y admet un maximum M. Ainsi, si $x_k \in [\widetilde{x}-r;\widetilde{x}+r]$, alors

$$|x_{k+1}-\widetilde{x}| \leqslant \left(\frac{2M}{m}\right)^2 |x_k-\widetilde{x}|^2.$$

Pour conclure, posons

$$\delta = \min\left(\frac{m}{2M}, \left(\frac{m}{2M}\right)^2, r\right)$$

et observons que si $x_k \in]\widetilde{x}-\delta;\widetilde{x}+\delta[$, alors comme $\delta \leqslant r$, le calcul précédent donne

$$\begin{aligned} |x_{k+1}-\widetilde{x}| &< \left(\frac{2M}{m}\right)^2 \delta^2 \\ &< \delta, \end{aligned}$$

où on a utilisé le fait que $\delta < (m/2M)^2$ est équivalent à $(2M\delta/m)^2 < \delta$. En particulier,

$$|x_{k+1}-\widetilde{x}| < \delta \leqslant r.$$

Ainsi, si $x_0 \in]\widetilde{x}-\delta;\widetilde{x}+\delta[$, alors $x_k \in]\widetilde{x}-\delta;\widetilde{x}+\delta[$ et les calculs précédents donnent par une récurrence simple

$$|x_k-\widetilde{x}| \leqslant \left(\frac{2M}{m}|x_0-\widetilde{x}|\right)^{2k}.$$

Comme

$$\frac{2M}{m}|x_0-\widetilde{x}| < 1,$$

on conclut que la suite $(x_k)_{k\in\mathbf{N}}$ converge vers $\widetilde{x}$. ∎

2.4.2 Méthode de Newton à plusieurs variables

Notre objectif est maintenant d'étendre la méthode précédente au cas d'une fonction de plusieurs variables. Mais il ne faut pas perdre de vue que le but est d'appliquer cette méthode au gradient de f, qui est une fonction à valeurs dans $\mathbf{R}^n$. Nous nous concentrerons donc sur une fonction $\phi : U \to \mathbf{R}^n$, avec U un ouvert de $\mathbf{R}^n$. La formule de récurrence à une variable n'a plus de sens ici puisqu'on ne peut diviser par la différentielle de f. Cela dit, si cette dernière est inversible, alors on peut multiplier par son inverse, ce qui revient au même. L'idée la plus naturelle est donc, étant donné un point de départ $x_0 \in U$, est de définir une suite par récurrence en posant

$$\boxed{x_{k+1} = x_k - J_\phi(x_k)^{-1}\phi(x_k).}$$

On voit alors que la condition n'est plus que la dérivée soit non-nulle, mais qu'elle soit inversible. Il est possible, en suivant le schéma de la preuve de la Proposition 2.4.1 de démontrer un résultat analogue en dimension n :

> **Théorème 2.4.2** (Méthode de Newton à plusieurs variables) Soit $\phi : U \to \mathbf{R}^n$ une fonction de classe $\mathcal{C}^2$ et $\widetilde{x} \in U$ tel que $\phi(\widetilde{x}) = 0$ et $J_\phi(\widetilde{x})$ est inversible. Alors, il existe $\delta > 0$ tel que pour tout $x_0 \in B(\widetilde{x},\delta)$, la suite $(x_k)_{k\in\mathbf{N}}$ converge vers $\widetilde{x}$.

Démonstration. Tout d'abord, l'ensemble des matrices inversibles formant un ouvert de $\mathrm{M}_n(\mathbf{R})$, il existe $r > 0$ tel que toute matrice $M \in B_f(J_\phi(\widetilde{x}), r)$ est inversible. Comme de plus la fonction $x \mapsto J_\phi(x)$ est continue, on en déduit qu'il existe $r' > 0$ tel que si $x \in B(\widetilde{x}, r') \subset U$, alors $J_\phi(x) \in B(J_\phi(\widetilde{x}), r)$ est inversible.

Supposons maintenant que, pour un indice k, $x_k \in B(\widetilde{x}, r')$ et considérons la fonction [21]

$$\psi : t \in [0;1] \mapsto \phi\left(x_k + t(\widetilde{x} - x_k)\right).$$

Elle est de classe $\mathcal{C}^2$ et la Formule de Taylor-Lagrange à l'ordre deux donne

$$\psi(1) - \psi(0) = \psi'(0) + \frac{1}{2}\psi''(c_k)$$

pour un certain $c_k \in [0;1]$. En utilisant la formule pour les dérivées partielles d'une composée (Proposition 2.3.12), on trouve (on sera attentif au fait que la différentielle seconde est ici une application bilinéaire de $\mathbf{R}^n \times \mathbf{R}^n$ dans $\mathbf{R}^n$)

$$\psi'(t) = J_\phi\left(x_k + t(\widetilde{x} - x_k)\right)(\widetilde{x} - x_k) \quad \& \quad \psi''(t) = D^2_{x_k + t(\widetilde{x} - x_k)}\phi\,(\widetilde{x} - x_k, \widetilde{x} - x_k),$$

dont on déduit, en posant $y_k = x_k + c_k(\widetilde{x} - x_k)$,

$$\phi(\widetilde{x}) - \phi(x_k) = J_\phi(x_k)(\widetilde{x} - x_k) + \frac{1}{2}D^2_{y_k}\phi\,(\widetilde{x} - x_k, \widetilde{x} - x_k).$$

Cette équation peut s'écrire sous la forme

$$\begin{aligned}\widetilde{x} - x_k &= -J_\phi(x_k)^{-1}\phi(x_k) - \frac{1}{2}J_\phi(x_k)^{-1}D^2_{y_k}\phi\,(\widetilde{x} - x_k, \widetilde{x} - x_k)\\ &= x_{k+1} - x_k - \frac{1}{2}J_\phi(x_k)^{-1}D^2_{y_k}\phi\,(\widetilde{x} - x_k, \widetilde{x} - x_k).\end{aligned}$$

On obtient alors

$$\begin{aligned}\|\widetilde{x} - x_{k+1}\| &= \left\|\frac{1}{2}J_\phi(x_k)^{-1} \circ D^2_{y_k}\phi\,(\widetilde{x} - x_k, \widetilde{x} - x_k)\right\|\\ &\leqslant \|\widetilde{x} - x_k\|^2 \frac{\|J_\phi(x_k)^{-1}\|\|D^2_{y_k}\phi\,(\widetilde{x} - x_k, \widetilde{x} - x_k)\|}{2}\end{aligned}$$

Comme ϕ est de classe $\mathcal{C}^2$ et que l'inversion de matrice est continue, les fonctions $x \mapsto \|J_\phi(x)^{-1}\|$ et $x \mapsto \|H_\phi(x)\|$ admettent sur le compact $B_f(\widetilde{x}, r'/2) \subset B(x, r/2) \subset U$ respectivement un minimum m et un maximum M. Il suit que

$$\|\widetilde{x} - x_{k+1}\| \leqslant \frac{M}{2m}\|\widetilde{x} - x_k\|^2.$$

La preuve se conclut alors exactement comme celle de la Proposition 2.4.1. ∎

Remarque 2.4.3. On remarquera que la convergence de la suite est *quadratique*, au sens où à chaque étape, la distance entre x_{k+1} et $\widetilde{x}$ est majorée – à une constante près – par le carré de la distance entre x_k et $\widetilde{x}$. Ceci permet d'évaluer concrètement le nombre d'itérations nécessaires pour parvenir à une approximation d'une précision donnée.

Un exemple d'implémentation de cette méthode en Python est donné en appendice de ce texte, au Paragraphe D.1.

21. On notera que cette définition a bien un sens car $\|x_k + t(\widetilde{x} - x_k) - \widetilde{x}\| = \|(1-t)(x_k - \widetilde{x})\| < r'$ et par conséquent $x_k + t(\widetilde{x} - x_k) \in B(\widetilde{x}, r') \subset U$.

2.4.3 Personne n'est parfait

La méthode de Newton que nous venons de présenter a l'avantage de la simplicité. Toutefois, elle présente aussi des inconvénients qui peuvent s'avérer rédhibitoires.

Il faut partir à point

Le premier problème, c'est que le Théorème 2.4.2 n'est pas explicite concernant le point de départ. En effet, il nous dit simplement que si on démarre assez près du point critique, alors l'algorithme convergera. Tout se joue donc sur le choix du point de départ. Voici quelques exemples de comportements indésirables qui peuvent se produire si le point de départ est "mal choisi". Comme on le notera, tous les problèmes surviennent déjà dans le cas d'une seule variable.

Exemple 2.4.4. Considérons la fonction $f : \mathbf{R} \to \mathbf{R}$ définie par

$$f(x) = -x^3 + 4x^2 - 2x + 2$$

et partons de $x_0 = 0$. On a $f'(x) = -3x^2 + 8x - 2$, d'où

$$x_1 = 0 - \frac{2}{-2} = 1$$

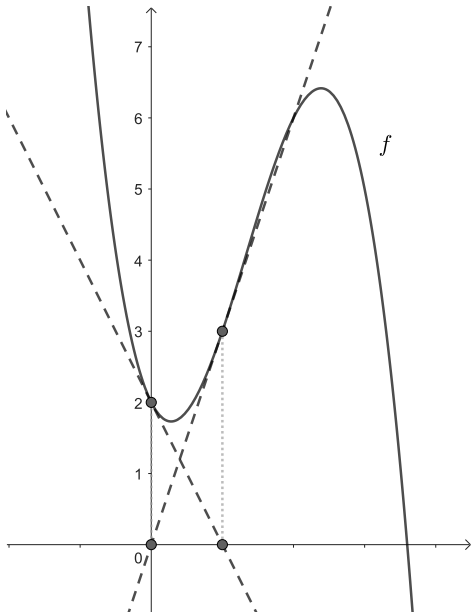
puis

$$x_2 = 1 - \frac{3}{3} = 0.$$

Nous sommes retombés sur le point x_0. Autrement, dit la suite oscille entre deux valeurs : $x_{2k} = 0$ et $x_{2k+1} = 1$ pour tout $k \in \mathbf{N}$. En particulier, cette suite ne converge pas vers l'unique solution de l'équation $f(x) = 0$, qui est [22]

$$\widetilde{x} = \frac{1}{3}\left(4 + \sqrt{10} + \sqrt[3]{100}\right).$$

Voici une illustration :



22. Ceci se vérifie par un calcul direct.

Les choses peuvent aussi diverger plus franchement, c'est-à-dire que la suite $(x_k)_{k\in\mathbf{N}}$ peut tout simplement tendre vers l'infini.

Exemple 2.4.5. On considère la fonction $f = \arctan : \mathbf{R} \to \mathbf{R}$. La suite $(x_k)_{k\in\mathbf{N}}$ vérifie alors la relation de récurrence

$$x_{k+1} = g(x_k),$$

où

$$g(x) = x - (1 + x^2)\arctan(x).$$

Posons $h : x \mapsto g(x)+x$ – en remarquant qu'il s'agit d'une fonction impaire – et étudions son signe. On a

$$h'(x) = -2x\arctan(x) + 1.$$

L'étude du signe de cette fonction mène au tableau de variations suivant où $0 < \alpha < 1$ est un réel que nous n'aurons pas besoin de calculer explicitement :

x	$-\infty$		$-\alpha$		α		$+\infty$
$h(x)$	$+\infty$	↘	$h(-\alpha)$	↗	$h(\alpha)$	↘	$-\infty$

On en déduit par le Théorème des valeurs intermédiaires qu'il existe un unique réel [23] $z \in]\alpha, +\infty[$ tel que $h(z) = -z$ (et comme h est impaire, on a également $h(-z) = z$).

Si $x \in]z; +\infty[$, alors $h(x) < h(z)$ par décroissance de h, donc $g(x) < -x$. Autrement dit, $|x_{k+1}| > |x_k|$. Le même résultat est vrai en supposant $x \in]-\infty; -z[$ et on en déduit que si $x_0 > z$, alors la suite $(|x_k|)_{k\in\mathbf{N}}$ est strictement croissante. Supposons qu'elle converge vers un réel y. On aura alors par continuité $|y| = |g(y)|$, donc $|y| = |z|$. Or, par croissance stricte de la suite,

$$|y| > |x_0| > |z|,$$

une contradiction. Ainsi, dans ce cas la suite $(|x_k|)_{k\in\mathbf{N}}$ tend vers $+\infty$.

Pour conclure, voici un exemple qui montre la sensibilité de la méthode au choix de x_0.

Exemple 2.4.6. On considère la fonction $f : \mathbf{R} \to \mathbf{R}$ définie par

$$f(x) = x^3 - 3x^2 + 2x.$$

On peut observer que ce polynôme se factorise par x, ce qui permet d'écrire

$$f(x) = x(x^2 - 3x + 2) = x(x-1)(x-2).$$

Ainsi, la méthode de Newton a trois limites possibles, à savoir 0, 1 et 2. Des simulations numériques montrent que pour $x_0 = 1,4$, la méthode converge vers 1 tandis que pour $x_0 = 1,5$, on a $x_1 = 0$. Ainsi, la suite "saute" dans ce cas par dessus 1 pour converger vers 0. Cela montre déjà qu'on ne peut pas prédire quel zéro on va obtenir.

Mais il y a plus. En effet, on peut pousser les calculs plus loin pour constater que pour $x_0 = 1,4447$ la méthode converge vers 1 tandis que pour $x_0 = 1,4448$ elle

23. Une recherche par dichotomie donne l'approximation $z \approx 1,3917$.

converge vers 0. Y aurait-il une valeur de "transition" quelque part dans l'intervalle $[1,4447; 1,4448]$? Prenons maintenant

$$x_0 = 1 + \frac{1}{\sqrt{5}} \approx 1,447214.$$

Alors, $x_1 = 1 - 1/\sqrt{5}$ et $x_2 = 1 + 1/\sqrt{5}$. Autrement dit, la suite oscille et ne converge pas. Ainsi, le comportement de la méthode de Newton par rapport au point de départ peut être de nature *chaotique*.

Le coût de la vie

Dans le cas de n variable, la méthode de Newton pose aussi des problèmes d'implémentation. En effet, contrairement au cas d'une seule variable, il faut à chaque étape de la récurrence calculer une nouvelle matrice Jacobienne et surtout l'inverser. Or l'inversion de matrice est une opération coûteuse, sans parler des erreurs d'arrondis qu'elle est susceptible de produire. La méthode de Newton est donc délicate à mettre en pratique.

Pour remédier à ce problème, on peut tenter de remplacer à chaque étape la nouvelle valeur de la Jacobienne par une "correction" de la valeur précédente. Nous allons présenter ici un méthode de ce type, appelée Méthode de Broyden [24]. L'idée est de remplacer la dérivation par une différence finie. Pour une fonction d'une variable $\phi : \mathbf{R} \to \mathbf{R}$, cela signifie remplacer $\phi'(x_{k+1})$ par

$$\Delta_{k+1}\phi = \frac{\phi(x_{k+1}) - \phi(x_k)}{x_{k+1} - x_k}.$$

Malheureusement, cette expression n'a pas de sens pour une fonction de plusieurs variables. Elle en a un cependant si on multiplie l'égalité par $(x_{k+1} - x_k)$ puisqu'on a alors

$$(\Delta_{k+1}\phi)(x_{k+1} - x_k) = \phi(x_{k+1}) - \phi(x_k).$$

Ceci suggère de chercher à fabriquer à chaque étape une matrice J_{k+1} telle que

$$J_{k+1}(x_{k+1} - x_k) = \phi(x_{k+1}) - \phi(x_k).$$

Pour cela, il existe une façon simple : en supposant que J_k a déjà été construite, on pose [25]

$$J_{k+1} = J_k + \frac{1}{\|x_{k+1} - x_k\|^2}\left(\phi(x_{k+1}) - \phi(x_k) - J_k(x_{k+1} - x_k)\right)(x_{k+1} - x_k)^t.$$

On vérifie alors directement que

$$\begin{aligned} J_{k+1}(x_{k+1} - x_k) &= J_k(x_{k+1} - x_k) \\ &\quad + \frac{1}{\|x_{k+1} - x_k\|^2}\left(\phi(x_{k+1}) - \phi(x_k) - J_k(x_{k+1} - x_k)\right)(x_{k+1} - x_k)^t(x_{k+1} - x_k) \\ &= J_k(x_{k+1} - x_k) + \phi(x_{k+1}) - \phi(x_k) - J_k(x_{k+1} - x_k) \\ &= \phi(x_{k+1}) - \phi(x_k). \end{aligned}$$

24. Charles G. Broyden (1933–2011) : physicien et mathématicien américain, spécialiste des problèmes d'optimisation. Il fut l'un des pionniers du développements des méthodes quasi-Newton (voir Remarque 2.4.11).

25. L'exposant t désigne une transposée, de sorte que $(x_{k+1} - x_k)^t$ est un vecteur ligne, c'est-à-dire la matrice d'une forme linéaire.

Quoique tout cela soit fort intéressant, et simplifie le problème du calcul des Jacobiennes successives (en admettant qu'on puisse effectivement les remplacer par J_k), l'obstacle majeur reste le calcul des inverses. Et c'est là que la méthode de Broyden est vraiment intéressante. En effet, J_{k+1} est obtenu en ajoutant à J_k une matrice de rang un. Or, il existe une formule simple pour calculer l'inverse d'une telle matrice, et la voici.

Lemme 2.4.7 (Formule de Sherman-Morrison)**.** *Soit $A \in \mathrm{M}_n(\mathbf{R})$ une matrice inversible et $u, v \in \mathbf{R}^n$. Si $1 + v^t A^{-1} u \neq 0$, alors la matrice $B = A + uv^t$ est inversible et de plus,*

$$B^{-1} = A^{-1} - \frac{A^{-1} u v^t A^{-1}}{1 + v^t A^{-1} u}.$$

Démonstration. Supposons $1 + v^t A^{-1} u \neq 0$. Il suffit alors de calculer le produit

$$\begin{aligned} B\left(A^{-1} - \frac{A^{-1} u v^t A^{-1}}{1 + v^t A^{-1} u}\right) &= \mathrm{Id} - \frac{u v^t A^{-1}}{1 + v^t A^{-1} u} + u v^t A^{-1} - u v^t \frac{A^{-1} u v^t A^{-1}}{1 + v^t A^{-1} u} \\ &= \mathrm{Id} + u v^t A^{-1} - \frac{1}{1 + v^t A^{-1} u} u \left(1 + v^t A^{-1} u\right) v^t A^{-1} \\ &= \mathrm{Id} + u v^t A^{-1} - u v^t A^{-1} \\ &= \mathrm{Id}. \end{aligned}$$

■

Remarque 2.4.8. On peut également, en raisonnant sur les déterminants, montrer que si $1 + v^t A^{-1} u = 0$ alors la matrice B n'est pas inversible.

La conclusion de cette histoire, c'est que si on peut montrer qu'en remplaçant $J_\phi(x_k)$ par J_k la méthode de Newton converge, alors on aura une façon beaucoup plus efficace de l'implémenter. Un tel résultat est démontrable, mais la preuve n'est pas simple et nous entraînerait loin. Nous allons néanmoins donner un énoncé précis. Rappelons les notations pour plus de facilité. Pour $x_0 \in U$, on définit une suite $(x_k)_{k\in\mathbf{N}}$ d'éléments de U et une suite $(H_k)_{k\in\mathbf{N}}$ d'éléments de $\mathrm{M}_n(\mathbf{R})$ par les relations

$$\begin{cases} H_0 = J_\phi(x_0)^{-1} \\ x_{k+1} = x_k - H_k \phi(x_k) \\ H_{k+1} = H_k + \dfrac{(x_{k+1} - x_k) - H_k(\phi(x_{k+1}) - \phi(x_k))}{(x_{k+1} - x_k)^t H_k (\phi(x_{k+1}) - \phi(x_k))} (x_{k+1} - x_k)^t H_k \end{cases}$$

Remarque 2.4.9. Comme on le voit ci-dessus, il n'est même pas nécessaire de calculer la suite $(J_k)_{k\in\mathbf{N}}$.

Supposons que la suite $(x_k)_{k\in\mathbf{N}}$ converge vers une limite $\widetilde{x}$ et que la suite $(H_k)_{k\in\mathbf{N}}$ converge vers une limite H inversible. Alors, on a en passant à la limite dans la relation de récurrence

$$\widetilde{x} = \widetilde{x} - H\phi(\widetilde{x}),$$

d'où $\phi(\widetilde{x}) = 0$. Ainsi, la limite sera bien un zéro de la fonction de départ.

Théorème 2.4.10 (Convergence de la méthode de Broyden) Soit $\phi : U \to \mathbf{R}$ une fonction de classe $\mathcal{C}^2$ et $\widetilde{x} \in U$ tel que $\phi(\widetilde{x}) = 0$ et $J_\phi(\widetilde{x})$ est inversible. Alors, il existe $\delta > 0$ tel que pour tout $x_0 \in B(\widetilde{x}, \delta)$, la suite $(x_k)_{k\in\mathbf{N}}$ converge vers $\widetilde{x}$.

Une implémentation de cette méthode en Python est donnée en appendice, au Paragraphe D.2. On pourra la comparer avec celle de la méthode de Newton.

Remarque 2.4.11. La Méthode de Broyden est un exemple de méthode dite *quasi-Newton*, c'est-à-dire dans laquelle on remplace les dérivées successives par des approximations afin de réduire les calculs.

On pourrait s'interroger sur le fait que cette méthode ne règle pas le problème du point de départ. C'est vrai, mais dans la pratique ce n'est pas forcément gênant. On peut en effet procéder de la façon suivante :

- On choisit un point de départ "raisonnable" (il existe des méthodes heuristiques [26] pour cela) ;
- On fait tourner l'algorithme pour calculer beaucoup de valeurs de la suite ;
- Si la suite ne semble pas converger, on change de point de départ ;
- Si la suite semble converger, on calcule autant de chiffres significatifs que souhaité en poursuivant la récurrence, puis on injecte cette approximation de la limite dans ϕ pour vérifier que le résultat est proche de 0.

26. On entend par là des méthodes dont le succès n'est pas assuré par une démonstration rigoureuse, mais dont des raisonnements plus lâches suggèrent qu'elles peuvent être efficaces.

CHAPITRE 3

LES CONTRAINTES

La volonté est tellement libre de sa nature, qu'elle ne peut jamais être contrainte.

R. Descartes, Les passions de l'âme

Nous avons jusqu'à maintenant traité des extrema d'une fonction $f : U \to \mathbf{R}$ sur tout son domaine de définition, qui était un ouvert. Nous avons également expliqué dans le Chapitre 1 que la plupart des problèmes concrets d'optimisation comportent des contraintes et que dans ce cas les critères du Chapitre 2 ne sont pas directement applicables. Nous avons illustré au Paragraphe 1.1.2 comment aborder ces problèmes avec contrainte :

1. On cherche d'abord un extremum à l'intérieur du domaine (qui est un ouvert et où l'on peut donc appliquer les méthodes du Chapitre 2).

2. On cherche ensuite les extrema sur le bord.

Le but de ce chapitre est de développer des outils pour aborder le second point, ou les deux en même temps.

3.1 Égalité

On doit puiser avec une cuiller ce qui est égal dans les choses.

W. Benjamin, Haschich à Marseille

3.1.1 Motivation

Considérons le second point ci-dessus. On recherche non pas le maximum de f sur U, mais son maximum sur un ensemble de la forme

$$\mathcal{S} = \{x \in \mathbf{R}^n \mid g(x) = 0\}$$

pour une fonction $g : \mathbf{R}^n \to \mathbf{R}$. Donnons pour plus de facilité une définition.

Définition 3.1.1. Soit U un ouvert de $\mathbf{R}^n$ et $f : U \to \mathbf{R}$. Soit également $g : U \to \mathbf{R}$. On dit que f a un *extremum local en x sous la contrainte $g = 0$* si la restriction de f à

$$\mathcal{S} = \{x \in U \mid g(x) = 0\}$$

a un extremum local en x. On définit de même les extrema stricts et globaux sous contrainte.

Un tel problème sera appelé *problème d'optimisation sous contraintes d'égalité*.

Remarque 3.1.2. Plus concrètement, f a, disons, un minimum local en x sous la contrainte $g = 0$ s'il existe $r > 0$ tel que pour tout $y \in B(x, r) \cap \mathcal{S}$, $f(y) \geqslant f(x)$.

Au Paragraphe 1.1.2, nous avions utilisé pour résoudre le problème le fait que le bord était constitué de segments et qu'il était possible de paramétrer chacun de ces segments de façon à ce qu'une des variables s'exprime en fonction des autres. Ceci permet de ramener le problème à la recherche d'un extremum pour une fonction d'une seule variable. De même, si l'on était capable de trouver un ouvert $V \subset \mathbf{R}^{n-1}$ et une fonction

$$\varphi : V \to \mathbf{R}$$

telle que $g(x) = 0$ si et seulement si $x_n = \varphi(x_1, \cdots, x_{n-1})$, alors le problème se ramènerait à la recherche d'extrema de la fonction

$$\widetilde{f} : (x_1, \cdots, x_{n-1}) \in V \mapsto f(x_1, \cdots, x_{n-1}, \varphi(x_1, \cdots, x_{n-1})).$$

La mauvaise nouvelle, c'est qu'une telle fonction n'existe pas en général, et qu'il est même facile de trouver un contre-exemple.

Exemple 3.1.3. Considérons la fonction $g : \mathbf{R}^2 \to \mathbf{R}$ définie par

$$g(x, y) = x^2 + y^2 - 1,$$

de sorte que l'ensemble $\mathcal{S}$ des points où g s'annule est le cercle de centre 0 et de rayon 1. Supposons qu'il existe une fonction $\varphi : I \to \mathbf{R}$ telle que $g(x, y) = 0$ si et seulement si $y = \varphi(x)$. Alors, $\mathcal{S}$ serait le graphe de la fonction φ, et en particulier son intersection avec une droite d'équation $x = k$ contiendrait au plus un point, pour tout $k \in \mathbf{R}$. Or l'intersection de $\mathcal{S}$ avec la droite d'équation $x = 1/\sqrt{2}$ contient les deux points $(1/\sqrt{2}, 1/\sqrt{2})$ et $(1/\sqrt{2}, -1/\sqrt{2})$.

La bonne nouvelle, c'est qu'une telle fonction existe quand même toujours "localement", et que cela est suffisant pour résoudre notre problème. Mais il s'agit d'un résultat non-trivial qui va nécessiter un petit peu de travail. Avant de nous lancer, regardons de plus près l'Exemple 3.1.3 ci-dessus.

Soit (x_0, y_0) un point du cercle. On voudrait trouver un arc de cercle contenant (x_0, y_0) qui puisse être le graphe d'une fonction $\varphi : I \to \mathbf{R}$. On dispose déjà des fonctions $\varphi_+, \varphi_- :]-1; 1[\to \mathbf{R}$ définies par

$$\varphi_\pm(x) = \pm\sqrt{1 - x^2}.$$

Ces applications sont de classe $\mathcal{C}^1$ et leurs dérivées ne s'annulent pas, ce sont donc des $\mathcal{C}^1$-*difféomorphismes* [1]. Leurs graphes sont des demi-cercles ouverts qui contiennent

1. Rappelons qu'un $\mathcal{C}^1$-difféomorphisme est une application bijective de classe $\mathcal{C}^1$ dont la réciproque est également de classe $\mathcal{C}^1$. Les propriétés de dérivation d'une fonction réciproque assurent qu'en effet, une fonction de classe $\mathcal{C}^1$ dont la dérivée ne s'annule pas est un $\mathcal{C}^1$-difféomorphisme.

à eux deux tous les point du cercle excepté $(-1,0)$ et $(1,0)$. Ainsi, les autres points appartiennent "localement" à des graphes de fonctions.

Qu'en est-il par exemple du point $(1,0)$? Il suffit en fait de tourner la figure! En effet, après rotation d'angle $\pi/4$, ce point devient le point $(0,1)$, pour lequel on a une description locale à l'aide d'un graphe. Plus concrètement, au lieu de considérer le graphe de φ_+, à savoir $\{(x,\varphi_+(x)) \mid x \in]-1;1[\}$, on peut considérer l'ensemble

$$\{(\varphi_+(y),y) \mid y \in]-1;1[\},$$

qui contient bien $(1,0)$.

Ainsi, *quitte à échanger les coordonnées*, on peut "recouvrir" le cercle par des graphes de fonctions. En d'autres termes, le cercle est "localement" un graphe de fonction à permutation près des coordonnées. Les fonctions en questions sont alors des "paramétrages locaux" du cercle.

3.1.2 Le théorème des fonctions implicites

Notre objectif est maintenant de trouver des conditions – si possible simples à exprimer – sur une fonction g pour que l'ensemble $\mathcal{S} = \{x \mid g(x) = 0\}$ puisse être "recouvert" par des graphes de fonctions comme dans le cas du cercle. Il s'avère que quoique remarquable, un tel résultat n'est pas si difficile à obtenir, et peut se démontrer en utilisant simplement des résultats sur les fonctions d'une variable.

Cependant, pour y voir plus clair, nous allons commencer par traiter le cas d'une seule contrainte. Et dans ce cas, la condition nécessaire pour pouvoir exprimer une coordonnée en fonction des autres est particulièrement simple. Pour avoir une intuition de son origine, considérons une fonction $g : U \to \mathbf{R}$ et imaginons que nous puissions la remplacer par le début de son développement limité au premier ordre pour un certain $\widetilde{x} \in U$ tel que $g(\widetilde{x}) = 0$, c'est-à-dire qu'on ait, pour tout $y \in U$ tel que $g(y) = 0$,

$$\begin{aligned} 0 &= g(y) \\ &= g(\widetilde{x}) + D_{\widetilde{x}}g(y-\widetilde{x}) \\ &= \langle \nabla g(\widetilde{x}), y-\widetilde{x}\rangle \\ &= \sum_{i=1}^{n} (y_i - \widetilde{x}_i)\frac{\partial g}{\partial x_i}(\widetilde{x}) \end{aligned}$$

Si par exemple la dérivée partielle de g par rapport à la dernière variable ne s'annule pas, alors on peut écrire

$$\begin{aligned} y_n &= \widetilde{x}_n - \left(\frac{\partial g}{\partial x_n}(\widetilde{x})\right)^{-1} \sum_{i=1}^{n-1} (y_i - \widetilde{x}_i)\frac{\partial g}{\partial x_i}(\widetilde{x}) \\ &= \varphi(y_1, \cdots, y_{n-1}). \end{aligned}$$

Autrement dit, on peut exprimer une des coordonnées de y en fonction des autres.

En général, on sait que g coïncide "presque" avec le début de son développement limité si l'on est "suffisamment proche" de $\widetilde{x}$, et le calcul précédent a donc une chance d'être encore valable, à condition qu'au moins une des dérivées partielles de g soit non-nulle. Cette dernière condition peut s'écrire de façon plus synthétique $\nabla g(\widetilde{x}) \neq 0$, et nous allons maintenant voir qu'elle est effectivement suffisante pour obtenir le résultat souhaité.

Proposition 3.1.4. *Soit $g : U \to \mathbf{R}$ une fonction de classe $\mathcal{C}^k$ et $\widetilde{x} \in U$ tel que $g(\widetilde{x}) = 0$ et*

$$\nabla g(\widetilde{x}) \neq 0.$$

*Alors, **quitte à permuter les coordonnées**, il existe un ouvert $V \subset \mathbf{R}^{n-1}$ contenant $(\widetilde{x}_1, \cdots, \widetilde{x}_{n-1})$, un ouvert $W \subset \mathbf{R}$ contenant $\widetilde{x}_n$ et une fonction de classe $\mathcal{C}^k$*

$$\varphi : V \to W$$

tels que pour tout $x \in V \times W \subset U$,

$$g(x) = 0 \iff x_n = \varphi(x_1, \cdots, x_{n-1}).$$

Démonstration. L'hypothèse sur le gradient de g signifie qu'il existe au moins une dérivée partielle qui est non-nulle au point $\widetilde{x}$. Quitte à changer la numérotation des coordonnées, on peut supposer – ce que nous ferons – qu'il s'agit de la dernière. Autrement dit, dans toute la preuve nous utiliserons le fait que

$$\frac{\partial g}{\partial x_n}(\widetilde{x}) \neq 0.$$

Si la démonstration n'est pas très ardue, elle n'est pas complètement intuitive pour autant. Nous allons donc procéder par étapes pour plus de clarté. Remarquons avant de commencer que l'hypothèse de l'énoncé nous dit que la dérivée partielle de g par rapport à la n-ième coordonnée est soit strictement positive, soit strictement négative au point $\widetilde{x}$. Quitte à remplacer g par son opposée – ce qui ne change pas l'ensemble $\mathcal{S}$ – on peut supposer sans perte de généralité que

$$\frac{\partial g}{\partial x_n}(\widetilde{x}) > 0.$$

► *Existence de la fonction φ*

Par continuité de la dérivée partielle de g par rapport à la n-ième coordonnée, il existe $r > 0$ tel que [2] pour tout $x \in B(\widetilde{x}, r) \subset U$,

$$\frac{\partial g}{\partial x_n}(x) > 0.$$

Nous allons maintenant "découper la boule en deux". Plus précisément, posons

$$\widetilde{\chi} = (\widetilde{x}_1, \cdots, \widetilde{x}_{n-1}) \in \mathbf{R}^{n-1}.$$

Alors, si $\chi \in B_1 = B(\widetilde{\chi}, r/2)$ et $y \in]\widetilde{x}_n - r/2; \widetilde{x}_n + r/2[$, en notant $(\chi, y) = (\chi_1, \cdots, \chi_{n-1}, y)$ on a par l'Inégalité triangulaire

$$\|(\chi, y) - \widetilde{x}\| \leqslant \|\chi - \widetilde{\chi}\| + \|y - \widetilde{x}_n\| < r.$$

Par conséquent, pour $\chi \in B_1$, la fonction

$$y \in]\widetilde{x}_n - r/2; \widetilde{x}_n + r/2[\mapsto g(\chi, y)$$

2. On sait également qu'il existe $r' > 0$ tel que $B(\widetilde{x}, r') \subset U$ parce que U est un ouvert. En remplaçant r par $\min(r, r')$ on peut donc supposer comme nous le faisons que $B(\widetilde{x}, r) \subset U$.

est strictement croissante. De plus, on sait que $g(\widetilde{\chi}, \widetilde{x_n}) = 0$, donc par croissance stricte,

$$g\left(\widetilde{\chi}, \widetilde{x_n} - \frac{r}{4}\right) < 0 \quad \& \quad g\left(\widetilde{\chi}, \widetilde{x_n} + \frac{r}{4}\right) > 0.$$

Par continuité de g, il existe donc $\delta, \delta' > 0$ tels que pour tout $\chi \in B(\widetilde{\chi}, \delta) \subset U$

$$g\left(\chi, \widetilde{x_n} - \frac{r}{4}\right) < 0$$

et pour tout $\chi \in B(\widetilde{\chi}, \delta')$,

$$g\left(\chi, \widetilde{x_n} + \frac{r}{4}\right) > 0.$$

Ainsi, en posant $r' = \min(r/2, \delta, \delta')$, on a pour tout $\chi \in B' = B(\widetilde{\chi}, r')$ que

$$g\left(\chi, \widetilde{x_n} - \frac{r}{4}\right) < 0 \quad \& \quad g\left(\chi, \widetilde{x_n} - \frac{r}{4}\right) > 0.$$

Alors, par le Théorème des valeurs intermédiaires appliqué à la fonction $y \mapsto g(\chi, y)$, il existe un unique $\varphi(\chi) \in]\widetilde{x_n} - r/4; \widetilde{x_n} + r/4[$ tel que $g(\chi, \varphi(\chi)) = 0$. En posant $V = B'$ et $W =]\widetilde{x_n} - r/4; \widetilde{x_n} + r/4[$, on a bien l'existence de la fonction de l'énoncé.

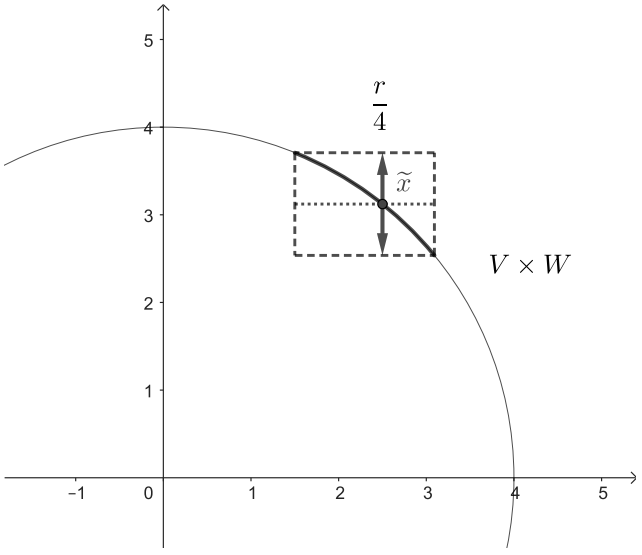
Avant d'aller plus loin, nous allons tenter d'illustrer le raisonnement précédent dans le cas de deux variables. On considère la fonction $g : \mathbf{R}^2 \to \mathbf{R}$ définie par

$$g(x_1, x_2) = x_1^2 + x_2^2 - 16.$$

Dans la figure ci-après, on a représenté le cercle d'équation $g(x_1, x_2) = 0$, ainsi qu'un point $\widetilde{x}$ vérifiant

$$\frac{\partial g}{\partial x_2}(\widetilde{x}) > 0.$$

On peut constater sur le dessin que si l'on se déplace horizontalement à l'intérieur du rectangle, alors on pourra toujours ensuite atteindre le cercle en se déplaçant d'au plus $r/4$ verticalement, soit vers le haut soit vers le bas. Ainsi, l'arc de cercle en gras est bien le graphe d'une fonction de la variable x_1.



▶ *Continuité de la fonction φ*

Nous allons d'abord montrer que φ est continue au point $\widetilde{\chi}$. Pour ce faire, considérons $\epsilon > 0$. En prenant $r < 4\epsilon$ dans la démonstration de la première partie, on voit qu'il existe $\delta > 0$ tel que pour tout $\chi \in B(\widetilde{\chi}, \delta)$,

$$\varphi(\chi) \in]\widetilde{x}_n - \epsilon; \widetilde{x}_n + \epsilon[,$$

donc φ est bien continue en $\widetilde{\chi}$. Si maintenant on considère un point quelconque $\chi \in V$, posons $\widehat{x} = (\chi, \varphi(\chi))$. Alors, la fonction g vérifie

$$g(\widehat{x}) = 0 \quad \& \quad \frac{\partial g}{\partial x_n}(\widehat{x}) \neq 0,$$

donc on peut appliquer la première partie de la preuve pour obtenir l'existence d'une fonction $\widehat{\varphi} : \widehat{V} \times \widehat{W} \to U$ qui est continue en χ. Il suffit maintenant de remarquer que pour tout $\overline{\chi} \in V \cap \widehat{V}$, on a $g(\overline{\chi}, \widehat{\varphi}(\overline{\chi})) = 0$ donc $\widehat{\varphi}(\overline{\chi}) = \varphi(\overline{\chi})$. Ainsi, φ et $\widehat{\varphi}$ coïncident sur cet ouvert qui contient χ donc φ est continue en χ.

▶ *Régularité de la fonction φ*

In nous reste à montrer que φ est de classe $\mathcal{C}^k$, et l'essentiel du travail est en fait de montrer qu'elle admet des dérivées partielles, ce que nous allons maintenant faire. Soit $1 \leqslant i \leqslant n-1$, soit $\chi \in V$ et soit (par la Proposition 2.1.5) $\delta > 0$ tel que pour tout $t \in]-\delta; \delta[$, $\chi + te_i \in V$. On pose, pour un tel t,

$$a = (\chi, \varphi(\chi)) \quad \& \quad b_t = (\chi + te_i, \varphi(\chi + te_i)).$$

On remarque que ces deux points appartiennent à $V \times W \subset B(\widetilde{x}, r)$, donc que pour tout $s \in [0;1]$, $a + s(b_t - a) \in B(\widetilde{x}, r)$. On peut donc considérer la fonction

$$h : s \in [0;1] \mapsto g\left(a + s(b_t - a)\right).$$

Il s'agit d'une fonction de classe $\mathcal{C}^1$, et le Théorème des accroissements finis affirme l'existence d'un $s_0 \in]0;1[$ tel que

$$h(1) - h(0) = h'(s_0).$$

En appliquant la Proposition 2.3.12 pour dérivée la fonction composée h, on trouve

$$h'(s) = D_{a+s(b_t-a)}g(b_t - a) = \langle \nabla g(a + s(b_t - a)), b_t - a \rangle,$$

ce qui donne, avec $c = a + s_0(b - a)$,

$$\begin{aligned} g(b_t) - g(a) &= \sum_{i=1}^{n} \frac{\partial g}{\partial x_i}(c)(b_t - a)_i \\ &= \frac{\partial g}{\partial x_i}(c)t + \frac{\partial g}{\partial x_n}(c)\left(\varphi(\chi + te_i) - \varphi(\chi)\right). \end{aligned}$$

Comme $g(a) = 0 = g(b_t)$ par construction, on en déduit que

$$\frac{\varphi(\chi + te_i) - \varphi(\chi)}{t} = -\left(\frac{\partial g}{\partial x_n}(c)\right)^{-1} \frac{\partial g}{\partial x_i}(c).$$

Par continuité de φ, $b_t \to a$ quand $t \to 0$, donc $c \to \widetilde{x}$. Il suit que φ admet une dérivée partielle par rapport à la i-ème coordonnée, et que

$$\frac{\partial \varphi}{\partial x_i}(\chi) = -\left(\frac{\partial g}{\partial x_n}(\chi, \varphi(\chi))\right)^{-1} \frac{\partial g}{\partial x_i}(\chi, \varphi(\chi)).$$

Si g est de classe $\mathcal{C}^k$, alors le membre de droite est de classe $\mathcal{C}^{k-1}$, donc φ est de classe $\mathcal{C}^k$, ce qui conclut la preuve. ∎

Exemple 3.1.5. Illustrons ce résultat dans le cas du cercle. On pose $g(x,y) = x^2 + y^2 - 1$. Alors, $\nabla g(x,y) = (2x, 2y)$ et ce vecteur ne s'annule qu'au point $(0,0)$. Or $g(0,0) \neq 0$, donc on peut appliquer la Proposition 3.1.4 en tout point de l'ensemble $\{x \in \mathbf{R}^n \mid g(x) = 0\}$, qui n'est autre que le cercle de centre 0 et de rayon 1. Nous avons déjà vu que tout point du cercle est contenu dans un graphe de fonction lui-même contenu dans le cercle, mais nous retrouvons ici ce résultat presque sans aucun calcul.

Le problème désormais, c'est qu'en général il n'y a pas de raison pour qu'il n'y ait qu'une seule contrainte. Il faudrait donc considérer m fonctions

$$g_1, \cdots, g_m : U \to \mathbf{R},$$

qui définissent une partie

$$\mathcal{S} = \{x \in U \mid g_i(x) = 0 \text{ pour tout } 1 \leqslant i \leqslant m\}$$

sur laquelle on cherche des extrema. Comme précédemment, on cherche une fonction $\varphi : V \to W \subset \mathbf{R}^m$, où V est un ouvert, telle que $g_i(x) = 0$ pour tout x si et seulement si $x = (\chi, \varphi(\chi))$. Seulement, comme nous avons maintenant non pas une mais m contraintes, il nous faudra peut-être utiliser plus de variables pour φ. Ainsi, V sera plutôt un ouvert de $\mathbf{R}^{n-m}$.

Cela dit, cette réflexion ne vaudra que si les contraintes sont vraiment "indépendantes" les unes des autres, et toute la difficulté est de comprendre quelle condition exprime cette indépendance. Dans le cas d'une variable, nous avions deviné que la contrainte était "non-triviale" si $\nabla g(\widetilde{x}) \neq 0$ en assimilant g à son développement limité au premier ordre. On peut faire de même ici : on aurait alors, pour $g(y) = 0$,

$$\begin{aligned} 0 &= g(y) \\ &= g(\widetilde{x}) + D_{\widetilde{x}} g(y - \widetilde{x}) \\ &= D_{\widetilde{x}} g(y) - D_{\widetilde{x}} g(\widetilde{x}). \end{aligned}$$

On aboutit alors à l'équation $D_{\widetilde{x}} g(y) = D_{\widetilde{x}} g(\widetilde{x})$. Celle-ci nous permet-elle d'exprimer certaines des coordonnées de y en fonction des autres ? Oui, car $D_{\widetilde{x}} g$ est une application linéaire de $\mathbf{R}^n \to \mathbf{R}^m$, et l'équation précédente signifie que les coordonnées de y sont solutions d'un système linéaire de m équations. Cela dit, m équations ne signifient pas qu'on peut simplifier m variables, il faut encore pour cela que les équations soient indépendantes, ou ce qui revient au même, que $D_{\widetilde{x}} g$ soit de rang m. Or, les colonnes de la matrice de $D_{\widetilde{x}} g$ dans la base canonique ne sont autres que les gradients $\nabla g_i(\widetilde{x})$ des contraintes. Ainsi, la bonne généralisation de notre hypothèse est l'indépendance linéaire des gradients.

Toutes ces considération mènent au résultat suivant, qui est l'un des plus fondamentaux du calcul différentiel.

Théorème 3.1.6 (Théorème des fonctions implicites) Soit U un ouvert de $\mathbf{R}^n$ et $g_1, \cdots, g_m : U \to \mathbf{R}$ des fonctions de classe $\mathcal{C}^k$. Soit $\widetilde{x} \in U$ tel que $g_i(\widetilde{x}) = 0$ pour tout $1 \leqslant i \leqslant m$. Si la famille de vecteurs

$$\nabla g_1(\widetilde{x}), \cdots, \nabla g_m(\widetilde{x}) \in \mathbf{R}^n$$

est libre, alors – **quitte à permuter les coordonnées** – il existe un ouvert $V \subset \mathbf{R}^{n-m}$ contenant $(\widetilde{x}_1, \cdots, \widetilde{x}_{n-m})$, un ouvert $W \subset \mathbf{R}^m$ contenant $(\widetilde{x}_{n-m+1}, \cdots, \widetilde{x}_n)$ et une fonction de classe $\mathcal{C}^k$

$$\varphi : V \to W$$

tels que pour tout $x \in V \times W \subset U$ on ait

$$g_1(x) = \cdots = g_m(x) = 0 \iff (x_{n-m+1}, \cdots, x_n) = \varphi(x_1, \cdots, x_{n-m}).$$

Les mots en gras dans l'énoncé sont importants. Ce que dit le théorème, c'est qu'on peut exprimer m coordonnées en fonction des $n-m$ autres. Mais ces coordonnées ne sont pas forcément les m dernières dans la base canonique, comme nous l'avons vu dans l'exemple du cercle (il faut parfois exprimer x en fonction de y).

Démonstration. L'idée de la preuve est de procéder par récurrence à l'aide de la Proposition 3.1.4. Ceci nécessite d'exprimer d'abord une variable en fonction des autres et nous allons donc diviser la démonstration en deux étapes pour plus de clarté.

Nous allons procéder par récurrence, avec l'hypothèse de récurrence suivante :

H_m : « Le Théorème est vrai quand le nombre de fonctions est exactement égal à m. »

Nous avons déjà démontré H_1 : c'est exactement le contenu de la Proposition 3.1.4. Supposons donc H_m vraie, et considérons le cas $m+1$.

► *Simplification d'une variable*

Nous allons d'abord appliquer la Proposition 3.1.4 à l'une des coordonnées pour la faire disparaître, de sorte qu'il suffira ensuite d'utiliser H_m. Pour ce faire, remarquons d'abord que $\nabla g_{m+1}(\widetilde{x}) \neq 0$ puisqu'il fait partie d'une famille libre, donc qu'au moins une de ses coordonnées est non nulle. Quitte à permuter les coordonnées, on peut donc supposer que

$$\frac{\partial g_{m+1}}{\partial x_n}(\widetilde{x}) \neq 0.$$

Il va nous falloir quelques notations pour nous y retrouver. Posons donc

$$\widetilde{\chi} = (\widetilde{x}_1, \cdots, \widetilde{x}_{n-m}) \quad \& \quad \widetilde{y} = (\widetilde{x}_{n-m+1}, \cdots, \widetilde{x}_n).$$

De plus, pour $y = (y_1, \cdots, y_{m+1}) \in \mathbf{R}^{m+1}$, on notera $y' = (y_1, \cdots, y_m)$. Avec tout ceci, on a

$$g_1(\widetilde{\chi}, \widetilde{y}', \widetilde{x}_n) = 0 \quad \& \quad \frac{\partial g_1}{\partial x_n}(\widetilde{\chi}, \widetilde{y}', \widetilde{x}_n) \neq 0,$$

donc par la Proposition 3.1.4 il existe $V' \subset \mathbf{R}^{n-1}$ ouvert contenant $(\widetilde{\chi}, \widetilde{y}')$ et $W' \subset \mathbf{R}$ ouvert contenant $\widetilde{x}_n$ et une fonction $\psi : V' \to W'$ de classe $\mathcal{C}^k$ telle que pour tout $x = (\chi, y', x_n) \in V' \times W'$,

$$g_{m+1}(x) = 0 \iff x_n = \psi(\chi, y').$$

► *Application de la récurrence*

Nous savons maintenant exprimer une des variables en fonction des autres, ce qui nous permet de nous ramener au cas de m variables. Pour ce faire, nous allons réduire le nombre de variables des fonctions du problème, en posant pour tout $1 \leqslant i \leqslant m$

$$\widetilde{g_i} : (x_1, \cdots, x_{n-1}) \in V' \mapsto g_i(x_1, \cdots, x_{n-1}, \psi(x_1, \cdots, x_{n-1})).$$

Comme $\psi(\widetilde{x}_1, \cdots, \widetilde{x}_{n-1}) = \widetilde{x}_n$, on sait déjà que toutes les fonctions ci-dessus s'annulent au point $(\widetilde{x}_1, \cdots, \widetilde{x}_{n-1})$. Il reste donc à vérifier que leurs gradients sont libres pour pouvoir appliquer l'hypothèse de récurrence. Pour ce faire, nous allons calculer leurs dérivées partielles : pour $1 \leqslant i \leqslant m$ et $1 \leqslant j \leqslant n-1$, la dérivation des fonctions composées (Proposition 2.3.12) donne

$$\begin{aligned}\frac{\partial \widetilde{g_i}}{\partial x_j}(x,y) &= \frac{\partial g_i}{\partial x_j}(x,y,\psi(x,y)) + \frac{\partial \psi}{\partial x_j}(x,y,\psi(x,y))\frac{\partial g_i}{\partial x_n}(x,y,\psi(x,y)) \\ &= \frac{\partial g_i}{\partial x_j}(x,y,\psi(x,y)) - \frac{\partial g_{m+1}}{\partial x_j}(x,y,\psi(x,y))\left(\frac{\partial g_{m+1}}{\partial x_n}(x,y,\psi(x,y))\right)^{-1}\frac{\partial g_i}{\partial x_n}(x,y,\psi(x,y))\end{aligned}$$

Autrement dit,

$$\nabla \widetilde{g_i}(\widetilde{x}_1, \cdots, \widetilde{x}_{n-1}) = \nabla g_i(\widetilde{x}) - \left(\frac{\partial g_{m+1}}{\partial x_n}(\widetilde{x})\right)^{-1}\frac{\partial g_i}{\partial x_n}(\widetilde{x})\nabla g_{m+1}(\widetilde{x})$$

et ces vecteurs forment une famille libre. En effet, si $\lambda_1, \cdots, \lambda_m \in \mathbf{R}$ vérifient

$$\sum_{i=1}^{m} \lambda_i \nabla \widetilde{g_i}(\widetilde{x}_1, \cdots, \widetilde{x}_{n-1}) = 0,$$

alors on a

$$\sum_{i=1}^{m} \lambda_i \nabla g_i(\widetilde{x}) - \left(\sum_{i=1}^{m} \lambda_i \left(\frac{\partial g_{m+1}}{\partial x_{n+m}}\right)^{-1}\frac{\partial g_i}{\partial x_n}\right)\nabla g_{m+1}(\widetilde{x}) = 0.$$

La liberté de la famille $\nabla g_1(\widetilde{x}), \cdots, \nabla g_{m+1}(\widetilde{x})$ implique alors $\lambda_i = 0$ pour tout $1 \leqslant i \leqslant m$. On peut donc appliquer l'hypothèse de récurrence au rang m pour obtenir des ouverts

$$(\widetilde{x}_1, \cdots, \widetilde{x}_{n-m}) \in V'' \subset \mathbf{R}^{(n-1)-m} \quad \& \quad (\widetilde{x}_{(n-1)-m+1}, \cdots, \widetilde{n-1}) \subset W'' \subset \mathbf{R}^m$$

et une application $\phi : V'' \to W''$ de classe $\mathcal{C}^k$ tels que pour tout $(\chi, y) \in V'' \times W'' \subset V'$,

$$\widetilde{g_1}(x,y) = \cdots = \widetilde{g_m}(x,y) = 0 \iff y = \phi(x).$$

► *Conclusion*

Posons

$$V = V'' \quad \& \quad W = (\phi(V'') \times \psi(V'' \times \phi(V''))) \subset W'' \times W'.$$

Alors, pour $x = (\chi, y', x_n) \in V \times W \subset U$ tel que $g_1(x) = \cdots = g_{m+1}(x) = 0$, on a $x_n = \psi(\chi, y')$ et donc $y' = \phi(\chi)$. Autrement dit,

$$(y', x_n) = (\phi(\chi), \psi(\chi, \phi(\chi))).$$

En définissant $\varphi : V'' \to W'' \times W'$ est définie par

$$\varphi(\chi) = (\chi, \psi(\chi, \phi(\chi))),$$

qui est de classe $\mathcal{C}^k$ par construction, a donc montré que $g_1(x) = \cdots = g_{m+1}(x) = 0$ implique

$$(x_1, \cdots, x_{n-(m+1)}) = \varphi(x_{n-(m+1)+1}, \cdots, x_n).$$

L'implication réciproque étant évidente, on a bien démontré le résultat voulu. ■

Remarque 3.1.7. Le Théorème 3.1.6 possède un proche parent, le Théorème d'inversion locale. Ces énoncés sont en général inséparables, mais nous n'aurons pas besoin du second de la suite et c'est pourquoi nous ne l'aborderons pas ici. Néanmoins, vu son importance pour le calcul différentiel, la géométrie et les mathématiques en général, nous ne pouvons l'occulter complètement. Nous renvoyons donc à l'Appendice B pour une démonstration du Théorème d'inversion locale ainsi que de son compagnon, le Théorème d'inversion globale.

Comme son nom le suggère, le Théorème 3.1.6 nous assure l'existence de la fonction φ mais ne nous la donne pas explicitement. En général il n'existe pas de moyen de la connaître, excepté au point $\widetilde{x}$. En effet, la dernière condition de l'énoncé assure que

$$\varphi(\widetilde{x}_1, \cdots, \widetilde{x}_{n-m}) = (\widetilde{x}_{n-m+1}, \cdots, \widetilde{x}_n).$$

On peut de fait aller plus loin et également connaître les valeurs des dérivées partielles de φ en ce point. Afin de les exprimer plus simplement, nous allons définir deux applications linéaires $D_{x,1}g : \mathbf{R}^{n-m} \to \mathbf{R}^m$ et $D_{x,2}g : \mathbf{R}^m \to \mathbf{R}^m$ de la façon suivante :

$$D_{x,1}g(h_1, \cdots, h_{n-m}) = \sum_{i=1}^{n-m} h_i \frac{\partial g}{\partial x_i}(x)$$

$$D_{x,2}g(h_{n-m+1}, \cdots, h_m) = \sum_{i=n-m+1}^{n} h_i \frac{\partial g}{\partial x_i}(x)$$

Ces application sont appelées *différentielles partielles* puisqu'elles se définissent comme la différentielle mais en se restreignant à certaines coordonnées

Proposition 3.1.8. *Sous les hypothèses du Théorème 3.1.6, l'application linéaire $D_{\widetilde{x},2}g$ est inversible et*

$$\frac{\partial \varphi}{\partial x_j}(\widetilde{x}_1, \cdots, \widetilde{x}_{n-m}) = -(D_{\widetilde{x},2}g)^{-1}\left(\frac{\partial g}{\partial x_j}(\widetilde{x})\right),$$

où $g = (g_1, \cdots, g_m)$.

Démonstration. Notons $g : U \to \mathbf{R}^m$ la fonction dont les coordonnées sont $g_1, \cdots, g_m$, et remarquons tout d'abord que nous sommes dans la situation où, quitte à échanger les coordonnées, les m dernières coordonnées de chacun des gradients $\nabla g_1(\widetilde{x}), \cdots \nabla g_m(\widetilde{x})$ forment une famille libre. Comme les vecteurs formés pas ces coordonnées forment les colonnes de la matrice de $D_{\widetilde{x},2}g \in \mathcal{L}(R^m, \mathbf{R}^m)$ dans la base canonique, on conclut que cette dernière est inversible.

Considérons la fonction $h : V \to \mathbf{R}$ définie par

$$h(x_1, \cdots, x_{n-m}) = g\left(x_1, \cdots, x_{n-m}, \varphi(x_1, \cdots, x_{n-m})\right).$$

D'après la dernière condition du Théorème 3.1.6, la fonction h est identiquement nulle sur V. En particulier, ses dérivées partielles sont nulles. Or, pour $1 \leqslant i \leqslant n-m$, on a par la Proposition 2.3.12

$$\begin{aligned}\frac{\partial h_i}{\partial x_j}(x_1,\cdots,x_{n-m}) = &\frac{\partial g_i}{\partial x_j}(x_1,\cdots,x_{n-m},\varphi(x_1,\cdots,x_{n-m}))\\ &+\sum_{k=1}^{m}\frac{\partial \varphi_k}{\partial x_j}(x_1,\cdots,x_{n-m})\frac{\partial g_i}{\partial x_{n-m+k}}(x_1,\cdots,x_{n-m},\varphi(x_1,\cdots,x_{n-m})).\end{aligned}$$

Comme h est nulle en $(\widetilde{x}_1,\cdots,\widetilde{x}_{n-m})$ et que $\varphi(\widetilde{x}_1,\cdots,\widetilde{x}_{n-m}) = (\widetilde{x}_{n-m+1},\cdots,\widetilde{x}_n)$, cette égalité peut s'écrire

$$0 = \frac{\partial g}{\partial x_j}(\widetilde{x}_1,\cdots,\widetilde{x}_n) + D_{\widetilde{x},2}g\left(\frac{\partial \varphi}{\partial x_j}(\widetilde{x}_1,\cdots,\widetilde{x}_{n-m})\right).$$

et le résultat suit. ∎

Dans le cas particulier où $m = 1$, $D_{\widetilde{x},2}g$ est simplement la multiplication par la dernière dérivée partielle de g, et on peut donc reformuler le résultat précédent sous la forme suivante :

$$\frac{\partial \varphi}{\partial x_j}(\widetilde{x}_1,\cdots,\widetilde{x}_{n-m}) = -\left(\frac{\partial g}{\partial x_n}(\widetilde{x}_1,\cdots,\widetilde{x}_n)\right)^{-1}\frac{\partial g}{\partial x_j}(\widetilde{x}_1,\cdots,\widetilde{x}_n).$$

On le voit bien, la méthode de la preuve peut s'adapter pour calculer les dérivées partielles d'ordre supérieur. Néanmoins, les calculs deviennent vite lourds et nous n'en aurons de toute façon pas besoin dans la suite.

3.1.3 Lagrange à la rescousse

Condition du premier ordre

Armés du Théorème 3.1.6, nous pouvons reprendre notre étude des extrema sous contraintes en suivant le plan esquissé au début de ce chapitre. Pour cela, nous allons chercher à exploiter le critère du Théorème 2.2.5. Et pour mieux comprendre nous allons commencer par le cas où il y a une seule contrainte g. Supposons donc, avec les même notations que précédemment, qu'on ait un point $\widetilde{x} \in \mathcal{S}$ auquel f admet un extremum local. Supposons également g de classe $\mathcal{C}^1$ et $\partial g/\partial x_n(x) \neq 0$.

On dispose de la fonction

$$\widetilde{f} : (x_1,\cdots,x_{n-1}) \mapsto f(x_1,\cdots,x_{n-1},\varphi(x_1,\cdots,x_{n-1}))$$

et ses dérivées partielles doivent s'annuler au point $\widetilde{x}$ d'après le Théorème 2.2.5. En utilisant la relation pour la dérivée partielle d'une fonction composée, ainsi que les calculs de la Proposition 3.1.8, on trouve alors pour tout $1 \leqslant i \leqslant n-1$

$$\begin{aligned}\frac{\partial \widetilde{f}}{\partial x_i}(\widetilde{x}) &= \frac{\partial f}{\partial x_i}(\widetilde{x}) + \frac{\partial \varphi}{\partial x_i}(\widetilde{x}_1,\cdots,\widetilde{x}_{n-1})\frac{\partial f}{\partial x_n}(\widetilde{x})\\ &= \frac{\partial f}{\partial x_i}(\widetilde{x}) - \frac{\partial g/\partial x_i(\widetilde{x})}{\partial g/\partial x_n(\widetilde{x})}\frac{\partial f}{\partial x_n}(\widetilde{x}).\end{aligned}$$

Si cette quantité est nulle, alors on a

$$\frac{\partial g}{\partial x_n}(\widetilde{x})\frac{\partial f}{\partial x_i}(\widetilde{x}) = \frac{\partial g}{\partial x_i}(\widetilde{x})\frac{\partial f}{\partial x_n}(\widetilde{x}).$$

Que signifient ces relations ? Si $\partial g/\partial x_i(\widetilde{x}) = 0$, alors comme la dérivée partielle de g par rapport à la n-ième variable n'est pas nulle, on doit avoir $\partial f/\partial x_i = 0$. Sinon, en posant

$$\lambda = \frac{\partial f/\partial x_n}{\partial g/\partial x_n},$$

on a

$$\frac{\partial f}{\partial x_i} = \lambda \frac{\partial g}{\partial x_i}.$$

Cette relation étant trivialement vérifiée quand les deux membres sont nuls, on a qu'elle est valable pour tout $1 \leqslant i \leqslant n$. On peut résumer ce que nous venons de montrer de la façon suivante : il existe $\lambda \in \mathbf{R}$ tel que

$$\nabla f(\widetilde{x}) = \lambda \nabla g(\widetilde{x}).$$

Nous avons donc une condition suffisante d'extremum sous contrainte d'égalité !

Tout cela est bien beau, mais comme nous l'avons vu dans l'exemple de la Section 1.1.2, les contraintes peuvent être données par plusieurs fonctions $g_1, \cdots, g_m$. Dans ce cas, l'analogue naturel de l'égalité ci-dessus serait que $\nabla f(x)$ soit une combinaison linéaire des $\nabla g_i(x)$. De fait, c'est le cas et il s'agit d'un résultat très important du à Lagrange [3].

Théorème 3.1.9 (Condition nécessaire d'extremum sous contrainte) Soit $f : U \to \mathbf{R}$ une fonction de classe $\mathcal{C}^1$, soient $g_1, \cdots, g_m : U \to \mathbf{R}$ des fonctions également de classe $\mathcal{C}^1$. Supposons que la restriction de f à l'ensemble

$$\mathcal{S} = \{x \in U \mid g_i(x) = 0 \text{ pour tout } 1 \leqslant i \leqslant m\}.$$

admet un extremum local en un point $\widetilde{x}$. Si la famille de vecteurs

$$\nabla g_1(\widetilde{x}), \cdots, \nabla g_m(\widetilde{x})$$

est libre, alors il existe $\lambda_1, \cdots, \lambda_m \in \mathbf{R}$ tels que

$$\boxed{\nabla f(\widetilde{x}) = \sum_{i=1}^{m} \lambda_i \nabla g_i(\widetilde{x})}$$

Les scalaires $\lambda_1, \cdots, \lambda_m$ sont appelés *multiplicateurs de Lagrange*.

Démonstration. Le Théorème 3.1.6 nous donne une fonction implicite $\varphi : V \to W$ telle que la fonction

$$\widetilde{f} : (x_1, \cdots, x_{n-m}) \mapsto f\left(x_1, \cdots, x_{n-m}, \varphi(x_1, \cdots, x_{n-m})\right).$$

3. Joseph-Louis Lagrange (1736–1813) : mathématicien et physicien d'origine italienne et naturalisé français. Il est l'un des plus important scientifiques du XVIII^e siècle. Ses travaux mathématiques portent sur tous les domaines de l'algèbre et de l'analyse, des plus théoriques au plus appliqués. Ses travaux en sciences physiques, quant à eux, ont eu une important cruciale, notamment sa Mécanique analytique.

a un extremum local au point $\widetilde{x}$. Nous allons maintenant différentier cette expression, et pour rendre les calculs plus clairs nous allons poser

$$\Phi(x_1, \cdots, x_{n-m}) = (x_1, \cdots, x_{n-m}, \varphi(x_1, \cdots, x_{n-m})),$$

de sorte que $\widetilde{f} = f \circ \Phi$. Alors, en utilisant la Proposition 3.1.8,

$$\begin{aligned} 0 &= D_{\widetilde{x}}\widetilde{f} \\ &= D_{\Phi(\widetilde{x}_1, \cdots, \widetilde{x}_{n-m})} f \circ D_{\widetilde{x}}\Phi \\ &= D_{\Phi(\widetilde{x}_1, \cdots, \widetilde{x}_{n-m}),1} f + D_{\Phi(\widetilde{x}_1, \cdots, \widetilde{x}_{n-m}),2} f \circ D_{(\widetilde{x}_1, \cdots, \widetilde{x}_{n-m})}\varphi \\ &= D_{\widetilde{x},1} f - D_{\widetilde{x},2} f \circ (D_{\widetilde{x},2} g)^{-1} \circ D_{\widetilde{x},1} g. \end{aligned}$$

Appliquée à un vecteur $h \in \mathbf{R}^{n-m}$, l'égalité précédente devient

$$\begin{aligned} \langle \nabla_1 f(\widetilde{x}), h \rangle &= \left\langle \nabla_2 f(\widetilde{x}), (D_{\widetilde{x},2} g)^{-1} \circ D_{\widetilde{x},1} g(h) \right\rangle \\ &= \left\langle \left[(D_{\widetilde{x},2} g)^{-1} \circ D_{\widetilde{x},1} g \right]^t (\nabla_2 f(\widetilde{x})), h \right\rangle, \end{aligned}$$

où $\nabla_1 f(\widetilde{x})$ est le vecteur formé des $n - m$ première coordonnées de $\nabla f(\widetilde{x})$ et $\nabla_2 f(\widetilde{x})$ le vecteur formé des m dernières. Comme ceci doit être valable pour tout h, les deux vecteurs à gauche des produits scalaires sont égaux. Autrement dit, $\nabla f(\widetilde{x})$ appartient à l'espace vectoriel

$$\mathcal{V} = \left\{ \xi \in \mathbf{R}^n \mid (\xi_1, \cdots, \xi_{n-m}) = \left(D_{\widetilde{x},2} g^{-1} \circ D_{\widetilde{x},1}(g) \right)^t (\xi_{n-m+1}, \cdots, \xi_n) \right\}.$$

Les coordonnées d'un élément de $\mathcal{V}$ sont déterminées par les m dernières, donc c'est un espace vectoriel de dimension au plus m. Par ailleurs, pour $1 \leqslant i \leqslant m$, on a $g_i \circ \Phi = 0$, donc cette fonction a en particulier un extremum local en $\widetilde{x}$, ce qui par le calcul précédent montre que $\nabla g_i(\widetilde{x}) \in \mathcal{V}$. Comme

$$\nabla g_1(\widetilde{x}), \cdots, \nabla g_m(\widetilde{x})$$

est une famille libre, il suit que $\dim(\mathcal{V}) \geqslant m$. Ainsi, $\dim(\mathcal{V}) = m$ et les vecteurs précédents forment une base de $\mathcal{V}$. Par conséquent, $\nabla f(\widetilde{x}) \in \mathcal{V}$ est combinaison linéaire de ces derniers, ce qu'il fallait démontrer. ∎

Donnons un petit exemple pour illustrer la méthode et son utilité.

Exemple 3.1.10. On considère un consommateur qui veut posséder deux biens x_1 et x_2. Sa satisfaction vis-à-vis des quantités de chaque bien qu'il possède est donnée par une fonction d'utilité u que nous prendrons de type Cobb-Douglas :

$$u(x_1, x_2) = x_1^{\alpha} x_2^{1-\alpha}.$$

Si les prix p_1 et p_2 des biens sont fixés, et si le consommateur veut atteindre une valeur d'utilité u_0 fixée (qu'on peut interpréter comme son niveau de vie idéal), quel est le moyen le plus économique de réaliser son objectif ?

Il s'agit de minimiser la fonction

$$p(x_1, x_2) = p_1 x_1 + p_2 x_2$$

sous la contrainte $u(x_1, x_2) = u_0$. Une façon pratique d'utiliser le Théorème 3.1.9 est de considérer la fonction

$$\begin{aligned}\mathcal{L}(x_1, x_2, \lambda) &= p(x_1, x_2) - \lambda\left(u(x_1, x_2) - u_0\right) \\ &= p_1 x_1 + p_2 x_2 - \lambda x_1^{\alpha} x_2^{1-\alpha}\end{aligned}$$

et de chercher ses points critiques. On obtient alors le système

$$\left\{\begin{array}{ccc} p_1 & = & \lambda\alpha x_1^{\alpha-1} x_2^{1-\alpha} \\ p_2 & = & \lambda(1-\alpha) x_1^{\alpha} x_2^{-\alpha} \\ x_1^{\alpha} x_2^{1-\alpha} & = & u_0 \end{array}\right.$$

En divisant la première équation par la dernière, on trouve

$$\frac{p_1}{u_0} = \frac{\lambda\alpha}{x_1}$$

tandis qu'en divisant la seconde équation par la dernière on trouve

$$\frac{p_2}{u_0} = \frac{\lambda(1-\alpha)}{x_2}.$$

On voit donc que pour que λ existe, on doit avoir

$$\frac{p_1 x_1}{\alpha} = \frac{p_2 x_2}{1-\alpha}.$$

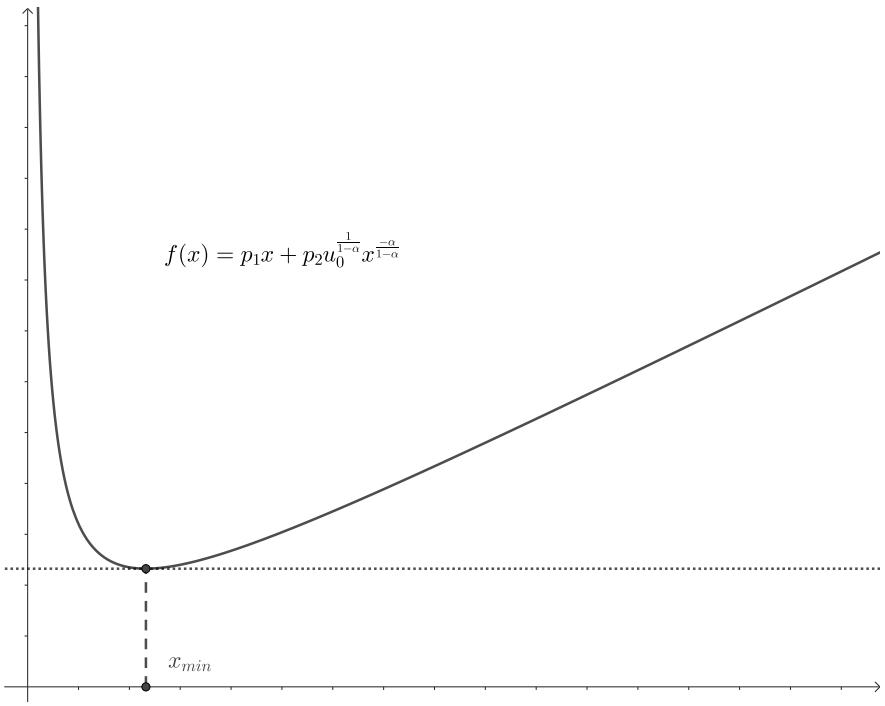
En réinjectant dans u on trouve alors

$$\left(\frac{\alpha}{1-\alpha}\right)^{1-\alpha}\left(\frac{p_1}{p_2}\right)^{1-\alpha} x_1^{\alpha} x_1^{1-\alpha} = u_0$$

et donc

$$x_1 = \left(\frac{1-\alpha}{\alpha}\right)^{1-\alpha} u_0 \left(\frac{p_2}{p_1}\right)^{1-\alpha} \quad \& \quad x_2 = \left(\frac{1-\alpha}{\alpha}\right)^{\alpha} u_0 \left(\frac{p_1}{p_2}\right)^{\alpha}.$$

Bien sûr, nous n'avons trouvé ici qu'un point critique sous contraintes, et rien de garantit a priori qu'il s'agisse d'un maximum local et encore moins global. C'est néanmoins bien le cas, comme le suggère l'allure de la fonction obtenue en exprimant x_2 en fonction de x_1 grâce à la contrainte :



Pour le démontrer, on ne peut recourir à un argument de compacité car

$$\{(x_1, x_2) \in \mathbf{R}_+^2 \mid u(x_1, x_2) = u_0\}$$

n'est pas borné. D'ailleurs, il n'y a qu'un seul point critique sous contrainte, donc f n'a pas de maximum. On peut néanmoins observer que $f(x) \to +\infty$ quand $\|x\| \to +\infty$, et ceci suffit à assurer l'existence d'un minimum par le Théorème 2.1.14.

Une interprétation géométrique

Nous avons maintenant un critère relativement simple pour détecter les points critiques sous contraintes d'égalité d'une fonction. Mais que signifie-t-il ? On pourrait bien sûr se contenter de le considérer comme une condition algébrique utile bien que mystérieuse, mais il y a en fait plus : il y a derrière une interprétation géométrique riche. Pour l'expliquer, revenons à l'idée de recouvrement par des graphes que nous avons évoquée au début de ce chapitre. Pour ce faire, précisons d'abord notre vocabulaire.

Définition 3.1.11. Soit $\varphi : U \subset \mathbf{R}^{n-p} \to \mathbf{R}^p$ une fonction. Son *graphe* est l'ensemble

$$\mathcal{S}_\varphi = \{(x, \varphi(x)) \mid x \in U\} \subset \mathbf{R}^n.$$

Le Théorème des fonctions implicites 3.1.6 nous dit précisément que si les gradients des fonctions g_i sont linéairement indépendants en un point x, alors il existe un ouvert $V \subset \mathbf{R}^n$ contenant x tel que $\mathcal{S} \cap V = \mathcal{S}_\varphi$, où φ est la fonction implicite. Ceci suggère que ces conditions sont en général les bonnes pour avoir un objet géométrique sur lequel on peut travailler, et mène ainsi à la définition suivante :

Définition 3.1.12. Une *sous-variété de* $\mathbf{R}^n$ *de codimension* p *et de classe* $\mathcal{C}^k$ est une partie $\mathcal{S} \subset \mathbf{R}^n$ telle que pour tout $x \in \mathcal{S}$, **quitte à permuter les coordonnées**, il existe

- Un ouvert V de $\mathbf{R}^{n-p}$;

- Un ouvert W de $\mathbf{R}^p$;

- Une application $\varphi : V \to W$ de classe $\mathcal{C}^k$ telle que $x \in V \times W$ et

$$\mathcal{S} \cap (V \times W) = \mathcal{S}_\varphi.$$

Autrement dit, une sous-variété est localement le graphe d'une fonction à permutation près des coordonnées. Tout l'enjeu de la géométrie différentielle est de comprendre la structure globale des sous-variétés alors que la définition ne donne que des informations locales. Pour ce qui nous concerne, nous voyons donc que trouver les extrema d'une fonction f sous des contraintes d'égalité revient à trouver les extrema d'une fonction f quand elle parcourt les points d'une sous-variété. Quand, dans un problème de géométrie, on se restreint à des points appartenant à un objet particulier (droite, cercle) on dit parfois que les points sont *liés* à l'objet. Ceci explique pourquoi un extremum sous contrainte est parfois appelé *extremum lié*. On parlera de même de *point critique lié*.

Le Théorème des fonctions implicites 3.1.6 peut alors s'exprimer géométriquement.

Corollaire 3.1.13. *Soient $g_1, \cdots, g_m : U \to \mathbf{R}$ des fonctions de classe $\mathcal{C}^k$ telles qu'en tout point x de*

$$\mathcal{S} = \{x \in U \mid g_i(x) = 0 \text{ pour tout } 1 \leqslant i \leqslant m\},$$

la famille de vecteurs

$$\nabla g_1(x), \cdots, \nabla g_m(x)$$

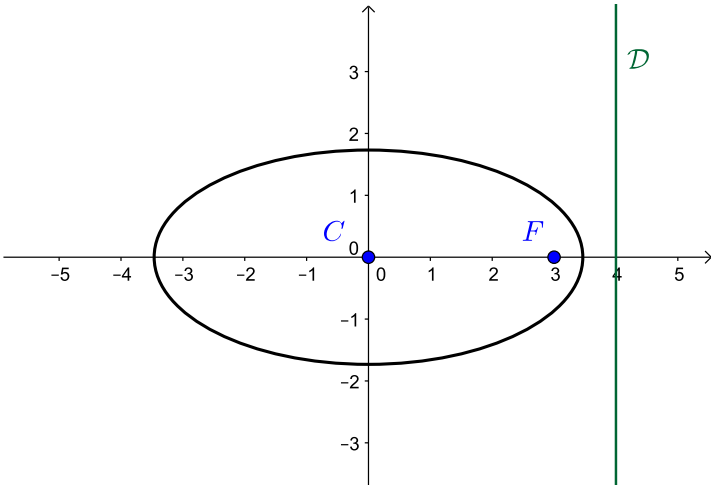
est libre. Alors, $\mathcal{S}$ est une sous-variété de codimension m et de classe $\mathcal{C}^k$.

Illustrons ceci par quelques exemples.

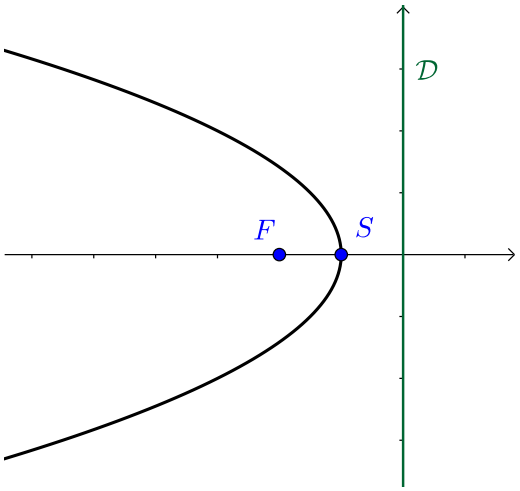
Exemple 3.1.14 (Coniques)**.** Tout d'abord, si on a une seule fonction g et deux variables, on obtient alors une courbe plane. Dans le cas d'une fonction polynomiale du second degré (qui est l'exemple le plus simple puisque dans le cas du premier degré on obtient une droite), les courbes obtenues seront des *coniques*. Il en existe trois types que nous listons ci-dessous. Elles sont accompagnées d'un dessin les représentant avec leur *foyer* et leur *directrice*, qui permettent de les définir de manière purement géométrique [4].

- L'*ellipse* d'équation $\dfrac{x_1^2}{a^2} + \dfrac{x_2^2}{b^2} = 1$,

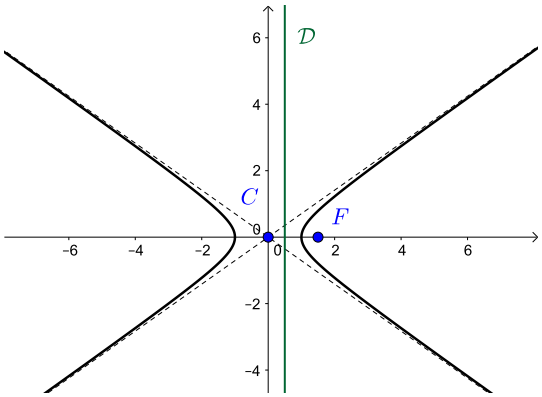
4. Étant donnés une droite $\mathcal{D}$, un point F extérieur à cette droite et une réel $e > 0$, la conique de foyer F, de directrice $\mathcal{D}$ et d'excentricité e est l'ensemble des points M du plan tels que $d(M, F) = ed(M, \mathcal{D})$.



- La *parabole* d'équation $x_2^2 = -2px_1$,



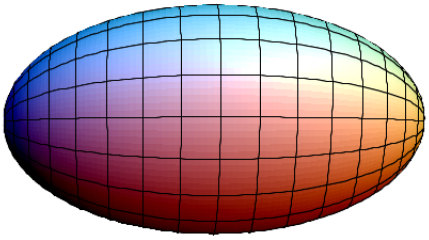
- L'hyperbole d'équation $\dfrac{x_1^2}{a^2} - \dfrac{x_2^2}{b^2} = 1$,



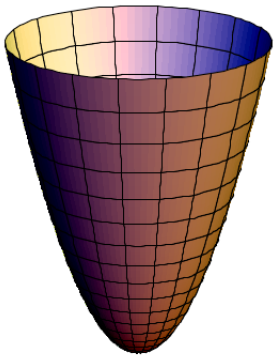
Dans le cas de l'ellipse et de l'hyperbole, le gradient de la fonction g ne s'annule qu'en $(0,0)$, qui ne vérifie pas l'équation de la conique tandis que pour la parabole, la première coordonnée du gradient n'est jamais nulle. Ainsi, dans tous les cas, on a une sous-variété par le Corollaire 3.1.13.

Exemple 3.1.15 (Quadriques)**.** Si l'on a maintenant une fonction et deux inconnues, on obtient une surface dans $\mathbf{R}^3$. Si la fonction est un polynôme de degré deux, la surface sera une *quadrique*. En voici quelques exemples :

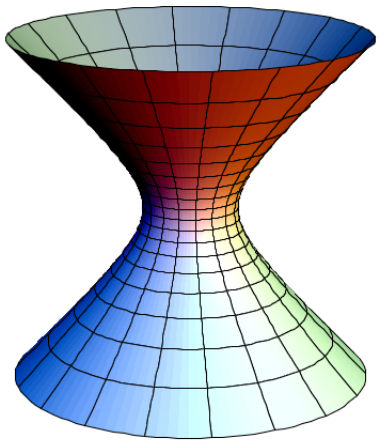
- L'*ellipsoïde* d'équation $\frac{x_1^2}{a^2} + \frac{x_2^2}{b^2} + \frac{x_3^2}{c^2} = 1$,



- Le *paraboloïde elliptique* d'équation $\frac{x_1^2}{a^2} + \frac{x_2^2}{b^2} = x_3$,



- L'*hyperboloïde à une nappe* d'équation $\frac{x_1^2}{a^2} + \frac{x_2^2}{b^2} - \frac{x_3^2}{c^2} = 1$,



Comme dans le cas des coniques, on vérifie facilement que ce sont des sous-variétés.

Les quadriques précédentes admettent des généralisations en dimension quelconque. Nous n'en donnerons qu'une des plus simples.

Exemple 3.1.16. Considérons la fonction $g : \mathbf{R}^n \to \mathbf{R}$ définie, pour $r > 0$, par

$$g(x_1, \cdots, x_n) = \sum_{i=1}^{n} x_i^2 - r^2.$$

La i-ème coordonnée de $\nabla g(x)$ n'est autre que $2x_i$, donc en tout point distinct de l'origine, $\nabla g(x) \neq 0$. Ainsi, $\mathcal{S}$ est une sous-variété, et l'on aura reconnu la sphère de rayon r centrée sur l'origine.

Abordons maintenant un exemple un peu moins visuel.

Exemple 3.1.17. On considère l'ensemble $\mathrm{SL}_n(\mathbf{R})$ des matrices carrées de taille n et de déterminant 1. Pour le décrire à l'aide d'une fonction, il semble naturel de définir $g : \mathrm{M}_n(\mathbf{R}) \to \mathbf{R}$ par

$$g(M) = \det(M) - 1.$$

La différentielle de g est donnée par [5]

$$D_M g(H) = \mathrm{Tr}(\mathrm{Com}(H)^t H).$$

En particulier, pour $H = E_{ij}$ – la matrice dont le coefficient (ij) vaut 1 et tous les autres 0 – on trouve

$$\frac{\partial g}{\partial M_{ij}}(M) = D_M g(E_{ij}) = \mathrm{Com}(M)_{ji}.$$

Or, ce dernier terme est un multiple du déterminant d'une matrice extraite de taille $n-1$ de M. Comme M est inversible si elle est dans SL_n, au moins l'un de ces déterminants est non nul. Autrement dit, $\nabla g(M) \neq 0$ et SL_n est donc bien une sous-variété.

Nous avons donc notre intuition géométrique, mais *quid* de la condition du Théorème 3.1.9 ? Pour l'expliquer, il nous faut comprendre le lien entre la géométrie de $\mathcal{S}$ et les gradients des fonctions g_i. Nous allons pour cela faire appel à la notion d'*espace tangent*.

5. Ceci a été vu en TD.

Qu'est-ce que la tangente au graphe d'une fonction ? C'est, en un point donné $\widetilde{x}$, la droite qui "approche" le mieux la courbe, au sens où si $y = ax+b$ est l'équation de cette droite, alors

$$f(x) - (ax + b) = o(x - \widetilde{x}).$$

Le développement limité au premier ordre de f au point $\widetilde{x}$ pouvant s'écrire

$$f(x) - [f(\widetilde{x}) + f'(\widetilde{x})(x - \widetilde{x})] = o(x - \widetilde{x}),$$

on trouve immédiatement l'équation de la tangente. Pour trouver un sous-espace de $\mathbf{R}^n$ qui "approche" la sous-variété, il faudra de même partir du développement limité de φ au premier ordre. Rappelons qu'un sous-espace vectoriel de dimension d peut être paramétré par la fonction

$$\psi : (t_1, \cdots, t_d) \in \mathbf{R}^d \mapsto \sum_{i=1}^{d} t_i v_i,$$

où $(v_i)_{1 \leqslant i \leqslant d}$ est une base du sous-espace. Or, le développement limité de φ au premier ordre au point $\widetilde{x}$ s'écrit

$$\varphi(x) - \varphi(\widetilde{x}) - \sum_{i=1}^{n-p} (x_i - \widetilde{x}_i) \frac{\partial \varphi}{\partial x_i}(\widetilde{x}) = o(x - \widetilde{x}).$$

Le sous-espace que nous recherchons est donc engendré par les dérivées partielles de φ. Nous sommes maintenant prêts à donner une définition, en n'oubliant pas que dans le cas d'une variable par exemple, la tangente est un objet géométrique, c'est-à-dire le graphe de la fonction $x \mapsto ax + b$. Pour alléger les notations, nous utiliserons la différentielle de φ au lieu des dérivées partielles, en se souvenant que

$$D_{\widetilde{x}}\varphi(x - \widetilde{x}) = \sum_{i=1}^{n-p} (x_i - \widetilde{x}_i) \frac{\partial \varphi}{\partial x_i}(\widetilde{x}).$$

Définition 3.1.18. Sois $\mathcal{S}$ une sous-variété, $\widetilde{x} \in \mathcal{S}$ et φ tel que $\mathcal{S}$ coïncide au voisinage de $\widetilde{x}$ avec le graphe de φ à permutation des coordonnées près. Alors, l'*espace tangent* à $\mathcal{S}$ au point $\widetilde{x}$ est le sous-espace affine de $\mathbf{R}^n$ passant par $(x, \varphi(x))$ et dirigé par le graphe de $D_{\widetilde{x}}\varphi$, c'est-à-dire par le sous-espace vectoriel

$$T_x\mathcal{S} = \mathbf{R}^{n-p} \times D_{\widetilde{x}}\varphi(\mathbf{R}^{n-p})$$

Il s'agit d'un sous-espace affine de dimension $n - p$.

Dans la suite, nous oublierons l'aspect affine et appellerons par abus $T_x\mathcal{S}$ l'espace tangent à $\mathcal{S}$ au point x.

Remarque 3.1.19. On sera peut-être gêné de ce que la définition précédente dépend de φ, alors qu'elle est sensée décrire un objet géométrique associé à la sous-variété $\mathcal{S}$. Évidemment, il s'agit d'un faux problème dans la mesure où l'espace tangent ne dépend pas vraiment du choix de φ, mais encore faut-il le démontrer. Ceci est fait dans l'Appendice C, mais le résultat ci-dessous montre déjà que dans le cas d'une sous-variété définie par des fonctions $g_1, \cdots, g_p$, l'espace tangent ne dépend que des gradients de ces fonctions au point considéré.

Proposition 3.1.20. *Soit U un ouvert de $\mathbf{R}^n$ et $g_1, \cdots, g_m : U \to \mathbf{R}^n$ des fonctions de classe $\mathcal{C}^1$ tels que pour tout $x \in U$, les vecteurs*

$$\nabla g_1(x), \cdots, \nabla g_m(x)$$

forment une famille libre. Alors, l'espace tangent à la sous-variété

$$\mathcal{S} = \{x \in U \mid g_1(x) = \cdots = g_m(x) = 0\}$$

en un point x est l'orthogonal de l'espace engendré par les vecteurs $\nabla g_i(x)$.

Démonstration. Considérons le sous-espace vectoriel

$$\mathcal{V} = \left\{\xi \in \mathbf{R}^n \mid (\xi_1, \cdots, \xi_{n-m}) = \left(D_{\widetilde{x},2}g^{-1} \circ D_{\widetilde{x},1}(g)\right)^t (\xi_{n-m+1}, \cdots, \xi_n)\right\}$$

introduit dans la démonstration du Théorème 3.1.9. Il est engendré par les gradients des fonctions g_i au point $\widetilde{x}$, donc il suffit [6] de montrer que $\mathcal{V}$ est orthogonal à $T_x\mathcal{S}$. Soit donc $\xi \in \mathcal{V}$ et $\chi \in \mathbf{R}^{n-p}$. Alors,

$$\begin{aligned}\langle \xi, (\chi, D_{\widetilde{x}}\varphi(\chi))\rangle &= \langle (\xi_1, \cdots, \xi_{n-m}), \chi\rangle + \langle (\xi_{n-m+1}, \cdots, \xi_n), D_{\widetilde{x}}\varphi(\chi)\rangle \\ &= \langle (\xi_1, \cdots, \xi_{n-m}), \chi\rangle - \left\langle \left(D_{\widetilde{x},2}g^{-1} \circ D_{\widetilde{x},1}(g)\right)^t (\xi_{n-m+1}, \cdots, \xi_n), \chi \right\rangle \\ &= 0.\end{aligned}$$

■

Exemple 3.1.21. Reprenons l'Exemple 3.1.16 de la sphère. Comme $\nabla g(x) = 2x$, l'espace tangent en x est le sous-espace affine passant par x et dirigé par $x^\perp$. Dans le cas du cercle ($n = 2$), on retrouve bien la caractérisation de la tangente comme droite orthogonale au rayon (qui est dirigé par le vecteur x).

Le critère du Théorème 3.1.9 peut alors se reformuler de la façon suivante :

Corollaire 3.1.22. *Soit $f : U \to \mathbf{R}$ une fonction de classe $\mathcal{C}^1$ et $\mathcal{S}$ une sous-variété de $\mathbf{R}^n$ de classe $\mathcal{C}^1$ et de codimension p. Si la restriction de f a un extremum local en un point $\widetilde{x}$, alors $\nabla f(\widetilde{x})$ est orthogonal à l'espace tangent à $\mathcal{S}$ au point $\widetilde{x}$.*

Remarquons que s'il n'y a aucune contrainte, $\mathcal{S} = U$ est un ouvert et son espace tangent est $\mathbf{R}^n$. Comme seul le vecteur nul est orthogonal à $\mathbf{R}^n$, on retrouve alors la condition du Théorème 2.2.5. Il est possible de donner une démonstration géométrique directe du Corollaire 3.1.22, à partir de laquelle on retrouvera la version analytique donnée dans le Théorème 3.1.9. Néanmoins, cela nécessite une définition plus abstraite de l'espace tangent que nous ne voulons pas introduire ici. Nous donnerons donc simplement l'intuition de l'argument, et renvoyons à l'Appendice C pour une démonstration complète et rigoureuse.

L'idée principale est la suivante : de façon générale, le gradient d'une fonction en un point donne la direction dans laquelle elle croît le plus vite. En effet, le développement limité de f à l'ordre 1 en $\widetilde{x}$ s'écrit

$$f(x) - f(\widetilde{x}) = \langle \nabla f(\widetilde{x}), x - \widetilde{x}\rangle + o(\|x - \widetilde{x}\|)$$

et le produit scalaire est le plus grand possible quand $x - \widetilde{x}$ est positivement colinéaire à $\nabla f(\widetilde{x})$, c'est-à-dire quand x est obtenu à partir de $\widetilde{x}$ par translation dans la direction de $\nabla f(\widetilde{x})$.

Le problème, c'est que $\mathcal{S}$ n'est pas "plate" : si on translate $\widetilde{x}$, on risque d'en sortir. Supposons néanmoins que $\nabla f(\widetilde{x})$ appartienne à l'espace tangent. Au voisinage du point

6. Pour des raisons de dimension.

$\widetilde{x}$, on peut "replier" l'espace tangent sur $\mathcal{S}$, de la même façon qu'on peut enrouler une droite tangente autour d'un cercle. Alors, il sera possible de se déplacer très légèrement sur $\mathcal{S}$ "dans la direction de $\nabla f(\widetilde{x})$. On obtiendra des points de $\mathcal{S}$ arbitrairement proches de $\widetilde{x}$ sur lesquels f prends une valeur plus grande. De même, en suivant $-\nabla f(\widetilde{w})$, on obtiendra des valeurs plus petites. Autrement dit, il n'y aura pas d'extremum local en $\widetilde{x}$.

Si maintenant $\nabla f(\widetilde{x})$ n'est pas dans l'espace tangent, on peut le décomposer en deux parties : une dans l'espace tangent et une qui lui est orthogonal. En se déplaçant comme dans le paragraphe précédente le long de la composante dans l'espace tangent, on arrivera à la même conclusion. En effet, seul entre en compte le produit scalaire de $\nabla f(\widetilde{x})$ avec $x-\widetilde{x}$, et celui-ci ne dépend à la limite que de la composante de $\nabla f(\widetilde{x})$ dans l'espace tangent. Ainsi, la seule possibilité pour avoir un extremum local est que $\nabla f(\widetilde{x})$ n'ait pas de composante dans l'espace tangent, c'est-à-dire qu'il lui soit orthogonal.

Exemple 3.1.23. Soit $A \in \mathrm{M}_n(\mathbf{R})$ une matrice symétrique, et considérons la fonction $f : \mathbf{R}^n \to \mathbf{R}$ définie par

$$f(x) = \langle Ax, x\rangle.$$

Un calcul élémentaire permet d'obtenir le gradient de f :

$$\begin{aligned} f(x+h) &= \langle Ax, x\rangle + \langle Ax, h\rangle + \langle Ah, x\rangle + \langle Ah, h\rangle \\ &= f(x) + \langle Ax, h\rangle + \langle h, Ax\rangle + o(\|h\|) \\ &= f(x) + 2\langle Ax, h\rangle + o(\|h\|) \end{aligned}$$

donc $\nabla f(x) = 2Ax$. On s'intéresse maintenant à la restriction de cette fonction à la sphère de l'Exemple 3.1.16. Comme elle est compacte, f y admet des extrema globaux – donc locaux – sous contraintes. En particulier, f a au moins un point critique sous contrainte. D'après l'Exemple 3.1.21, en un tel point $\widetilde{x}$, $\nabla f(\widetilde{x})$ doit être orthogonal à l'espace tangent, donc colinéaire au vecteur $\widetilde{x}$. Autrement dit, il existe $\lambda \in \mathbf{R}$ tel que

$$2A\widetilde{x} = \lambda\widetilde{x}.$$

Ainsi, nous venons de prouver que toute matrice symétrique a un vecteur propre, ce qui par récurrence montre que toute matrice symétrique est diagonalisable !

Condition du second ordre

Nous avons maintenant à notre disposition un analogue du Théorème 2.2.5 qui nous permet de localiser les extrema potentiels, mais qu'en est-il du Théorème 2.3.19 qui permet de déterminer la nature des points critiques liés ? Il s'avère qu'on a aussi un résultat analogue et, bien que l'énoncé soit long, il s'agit de la façon la plus naturelle d'étendre le résultat du Théorème 2.3.19.

Théorème 3.1.24 (Condition suffisante d'extremum sous contrainte) Soit $f : U \to \mathbf{R}$ une fonction de classe $\mathcal{C}^2$, soient $g_1, \cdots, g_m : U \to \mathbf{R}$ des fonctions également de classe $\mathcal{C}^2$. Soit $\widetilde{x}$ un point de l'ensemble

$$\mathcal{S} = \{x \in U \mid g_i(x) = 0 \text{ pour tout } 1 \leqslant i \leqslant m\}$$

pour lequel il existe $\lambda_1, \cdots, \lambda_m \in \mathbf{R}$ tels que

$$\nabla f(\widetilde{x}) = \sum_{i=1}^{m} \lambda_i \nabla g_i(\widetilde{x}).$$

Si la famille de vecteurs
$$\nabla g_1(\widetilde{x}), \cdots, \nabla g_n(\widetilde{x})$$
est libre et si la forme quadratique
$$Q = D^2_{\widetilde{x}} f - \sum_{i=1}^{m} \lambda_i D^2_{\widetilde{x}} g_i$$
est définie positive sur
$$E = \bigcap_{i=1}^{m} \ker(D_{\widetilde{x}} g_i),$$
alors f admet un minimum local strict sous contraintes d'égalité au point $\widetilde{x}$. De même, si Q est définie négative sur E, alors f admet un maximum local strict sous contrainte au point $\widetilde{x}$.

Démonstration. Nous reprenons les notations de la démonstration du Théorème 3.1.9, l'idée étant similaire : nous allons montrer que sous les hypothèses de l'énoncé, la différentielle seconde de $\widetilde{f}$ est définie positive (l'autre cas se traitant de même, *mutatis mutandis*), ce qui donnera le résultat puisqu'on sait déjà que $\widetilde{x}$ est un point critique pour $\widetilde{f}$ (c'est le contenu de la preuve du Théorème 3.1.9). Pour ce faire, commençons par calculer les dérivées partielles secondes de $\widetilde{f}$. Nous avons déjà établi que pour $1 \leqslant i \leqslant n-m$,
$$\frac{\partial \widetilde{f}}{\partial x_i} = \sum_{k=1}^{n} \left(\frac{\partial f}{\partial x_k} \circ \Phi \right) \frac{\partial \Phi_k}{\partial x_i},$$
donc pour tous $1 \leqslant i, j \leqslant n-m$,
$$\frac{\partial^2 \widetilde{f}}{\partial x_j \partial x_i} = \sum_{k,\ell=1}^{n} \left(\frac{\partial^2 f}{\partial x_\ell \partial x_k} \circ \Phi \right) \frac{\partial \Phi_\ell}{\partial x_j} \frac{\partial \Phi_k}{\partial x_i} + \sum_{k=1}^{n} \left(\frac{\partial f}{\partial x_k} \circ \Phi \right) \frac{\partial^2 \Phi_k}{\partial x_j \partial x_i}.$$
Considérons maintenant, pour $1 \leqslant p \leqslant m$, la fonction $\widetilde{g}_p = g_p \circ \Phi$. La même formule que ci-dessus est valable pour les dérivées partielles secondes de $\widetilde{g}_p$, mais cette fonction est identiquement nulle par définition de Φ. Par conséquent, on a
$$\sum_{k=1}^{n} \left(\frac{\partial g_p}{\partial x_k} \circ \Phi \right) \frac{\partial^2 \Phi_k}{\partial x_j \partial x_i} = - \sum_{k,\ell=1}^{n} \left(\frac{\partial^2 g_p}{\partial x_\ell \partial x_k} \circ \Phi \right) \frac{\partial \Phi_\ell}{\partial x_j} \frac{\partial \Phi_k}{\partial x_i}.$$
Nous pouvons maintenant faire appel à nouveau au lien entre le gradient de f et ceux des contraintes pour écrire
$$\begin{aligned}
\sum_{k=1}^{n} \frac{\partial f}{\partial x_k} \circ \Phi \times \frac{\partial^2 \Phi_k}{\partial x_j \partial x_i} &= \sum_{k=1}^{n} \sum_{p=1}^{m} \lambda_p \frac{\partial g_p}{\partial x_k} \circ \Phi \times \frac{\partial^2 \Phi_k}{\partial x_j \partial x_i} \\
&= \sum_{p=1}^{m} \lambda_p \sum_{k=1}^{n} \left(\frac{\partial g_p}{\partial x_k} \circ \Phi \right) \frac{\partial^2 \Phi_k}{\partial x_j \partial x_i} \\
&= - \sum_{p=1}^{m} \lambda_p \sum_{k,\ell=1}^{n} \left(\frac{\partial^2 g_p}{\partial x_\ell \partial x_k} \circ \Phi \right) \frac{\partial \Phi_\ell}{\partial x_j} \frac{\partial \Phi_k}{\partial x_i}
\end{aligned}$$
qui donne finalement
$$\frac{\partial^2 \widetilde{f}}{\partial x_j \partial x_i} = \sum_{k,\ell=1}^{n} \left(\frac{\partial^2 f}{\partial x_\ell \partial x_k} \circ \Phi \right) \frac{\partial \Phi_\ell}{\partial x_j} \frac{\partial \Phi_k}{\partial x_i} - \sum_{p=1}^{m} \lambda_p \sum_{k,\ell=1}^{n} \left(\frac{\partial^2 g_p}{\partial x_\ell \partial x_k} \circ \Phi \right) \frac{\partial \Phi_\ell}{\partial x_j} \frac{\partial \Phi_k}{\partial x_i}.$$

En évaluant cette égalité au point $\widetilde{x}$ et en factorisant on conclut alors que

$$\frac{\partial^2 \widetilde{f}}{\partial x_j \partial x_i}(\widetilde{x}) = \sum_{k,\ell=1}^{n} \left(\frac{\partial^2 f}{\partial x_\ell \partial x_k}(\widetilde{x}) - \sum_{p=1}^{m} \lambda_p \sum_{k,\ell=1}^{n} \frac{\partial^2 g_p}{\partial x_\ell \partial x_k}(\widetilde{x}) \right) \frac{\partial \Phi_\ell}{\partial x_j}(\widetilde{x}) \frac{\partial \Phi_k}{\partial x_i}(\widetilde{x}).$$

Matriciellement, cette égalité devient

$$H_{\widetilde{f}}(\widetilde{x}) = (J_\Phi(\widetilde{x}_1, \cdots, \widetilde{x}_{n-m}))^t \left(H_f(\widetilde{x}) - \sum_{i=1}^{p} \lambda_i H_{g_i}(\widetilde{x}) \right) (J_\Phi(\widetilde{x}_1, \cdots, \widetilde{x}_{n-m})).$$

On voit alors [7] que $H_{\widetilde{f}}(\widetilde{x})$ sera définie positive (respectivement définie négative) dès que la forme quadratique du milieu est définie positive (respectivement définie négative) sur l'image de $J_\Phi(\widetilde{x}_1, \cdots, \widetilde{x}_{n-m})$. Or, nous avons déjà montré lors de la démonstration du Théorème 3.1.9 que

$$D_{(\widetilde{x}_1, \cdots, \widetilde{x}_{n-m})}\Phi = \mathrm{Id} - (D_{\widetilde{x},2} g)^{-1} \circ (D_{\widetilde{x},1} g).$$

Par conséquent, on a $\mathrm{Ker}(D_{(\widetilde{x}_1, \cdots, \widetilde{x}_{n-m})}\Phi)^t = \mathcal{V}$, où $\mathcal{V}$ est l'espace vectoriel de la démonstration du Théorème 3.1.9, à savoir

$$\mathcal{V} = \left\{ \xi \in \mathbf{R}^n \mid (\xi_1, \cdots, \xi_{n-m}) = \left(D_{\widetilde{x},2} g^{-1} \circ D_{\widetilde{x},1}(g) \right)^t (\xi_{n-m+1}, \cdots, \xi_n) \right\}.$$

Mais alors [8],

$$\mathrm{Im}(D_{(\widetilde{x}_1, \cdots, \widetilde{x}_{n-m})}\Phi) = \left(\mathrm{Ker}(D_{(\widetilde{x}_1, \cdots, \widetilde{x}_{n-m})}\Phi)^t \right)^\perp = \mathcal{V}^\perp.$$

Or, puisque $\mathcal{V}$ est engendré par les gradients des fonctions g_i, on a [9]

$$\begin{aligned} \mathcal{V}^\perp &= \bigcap_{i=1}^{m} \nabla g_i(\widetilde{x})^\perp \\ &= \bigcap_{i=1}^{m} \ker(D_{\widetilde{x}} g_i) \\ &= E, \end{aligned}$$

ce qui conclut la preuve. ∎

Remarque 3.1.25. On peut également démontrer un analogue de la Proposition 2.3.20 : s'il y a un maximum local sous contrainte, alors la forme quadratique Q doit être positive sur E, et s'il s'agit d'un minimum alors elle doit être négative.

7. En effet, si A est un matrice symétrique et M une matrice quelconque, et si $\langle A\xi, \xi \rangle > 0$ pour tout $\xi \in \mathrm{Im}(M)$, alors pour tout $\xi \in \mathbf{R}^n$ on a $\langle M^t A M \xi, \xi \rangle = \langle AM\xi, M\xi \rangle > 0$.

8. Rappelons que pour toute application linéaire $T : \mathbf{R}^{n_1} \to \mathbf{R}^{n_2}$, on a $\mathrm{Im}(T) = \mathrm{Ker}(T^t)^\perp$. En effet, si $x \in \mathrm{Im}(T)$ et $y \in \mathrm{Ker}(T^t)^\perp$, alors $x = T(z)$ et donc $\langle x, y \rangle = \langle T(z), y \rangle = \langle z, T^t(y) \rangle = 0$. Réciproquement, si $y \in \mathrm{Im}(T)^\perp$, alors pour tout $z \in \mathbf{R}^{n_1}$ on a $\langle z, T^t(y) \rangle = \langle T(z), y \rangle = 0$, donc $T^t(y) = 0$.

9. Nous utilisons ici le résultat suivant d'algèbre linéaire : Si un sous-espace vectoriel $V \subset \mathbf{R}^n$ est engendré par des vecteurs $v_1, \cdots, v_d$, alors $V^\perp = \bigcap_{i=1}^{d} v_i^\perp$. En effet, si $x \in V^\perp$, alors en particulier $\langle x, v_i \rangle = 0$ pour tout $1 \leqslant i \leqslant d$. Réciproquement, si $x \in \bigcap_{i=1}^{d} v_i^\perp$ et $y \in V$, alors $y = \sum_{i=1}^{d} \lambda_i v_i$, donc

$$\langle x, y \rangle = \sum_{i=1}^{d} \lambda_i \langle x, v_i \rangle = 0.$$

Le Théorème 3.1.24 n'est pas d'un emploi aisé, puisqu'il faut réussir à comprendre la forme quadratique donnée par les multiplicateurs de Lagrange sur l'espace $\ker(D_{\widetilde{x}}g)$ qui peut être ardu à décrire. Nous ne donnerons qu'un exemple, dans lequel les choses se passent sympathiquement, et qui est le prolongement de l'Exemple 3.1.23.

Exemple 3.1.26. Soit $A \in \mathrm{M}_n(\mathbf{R})$ une matrice symétrique et $f : \mathbf{R}^n \to \mathbf{R}$ la fonction définie par

$$f(x) = \langle Ax, x\rangle.$$

Nous avons vu à l'Exemple 3.1.23 que tout extremum local de f sur la sphère

$$\mathcal{S} = \{x \in \mathbf{R}^n \mid \|x\|^2 = 1\}$$

correspond à un vecteur propre. Soit donc $\widetilde{x}$ un tel vecteur propre et λ la valeur propre associée, qui n'est autre que le multiplicateur de Lagrange correspondant. En posant $g(x) = \|x\|^2$, on a

$$D^2_{\widetilde{x}}g(h,h) = 2\|h\|^2.$$

D'autre part, puisque $\nabla f(x) = 2Ax$, on a $D^2_{\widetilde{x}}f(h,h) = \langle 2Ah, h\rangle$. Ainsi,

$$D^2_{\widetilde{x}}f(h,h) - \lambda D^2_{\widetilde{x}}g(h,h) = 2\langle (A - \lambda \mathrm{I}_n)h, h\rangle.$$

La matrice $A' = A - \lambda \mathrm{I}_n$ est symétrique – donc diagonalisable en base orthonormée – et ses valeurs propres sont les nombres de la forme $\mu - \lambda$, où μ est une valeur propre de A. On constate ici que A' n'est jamais définie positive, puisque $\lambda - \lambda = 0$ est valeur propre. C'est pourquoi la restriction à E dans l'énoncé du Théorème 3.1.24 est essentielle.

Ainsi, pour que le Théorème 3.1.24 s'applique, il faut que tous les vecteurs propres orthogonaux à x correspondent à une valeur propre μ telle que $\mu - \lambda > 0$. Autrement dit, il faut λ soit la plus petite valeur propre de A et soit de multiplicité 1. Alors, f a un minimum local strict en x et on peut montrer que réciproquement, si f a un minimum local en x alors λ doit être la plus petite valeur propre. On établit de même que f a un maximum local strict si λ est la plus grande valeur propre et est de multiplicité 1.

3.2 Inégalités

Tous les animaux sont égaux
mais certains sont plus égaux que d'autres.

G. Orwell, La ferme des animaux

Le problème évoqué au paragraphe 1.1.2 du Chapitre 1 ne comportait pas que des contraintes d'égalité. De fait, ces dernières étaient linéaires et pouvaient donc directement être utilisées pour réduire le nombre de variables. Les autres contraintes étaient données par des *inégalités*. Il serait donc utile pour nous d'avoir un résultat similaire au Théorème 3.1.9 permettant de détecter les extrema sous contraintes d'inégalité.

Une stratégie possible pour aborder des contraintes d'inégalité est la suivante : le domaine défini par les inégalités est la réunion disjointe de son intérieur - qui est ouvert - et de son bord [10]. En admettant qu'une fonction admette un extremum local sur ce domaine, nous avons donc deux possibilités :

10. On pourra se reporter au premier TD pour la définition rigoureuse du bord d'une partie de $\mathbf{R}^n$, mais pour toute cette discussion cette notion n'est en fait pas nécessaire, et le terme "bord" peut être pris dans un sens intuitif.

- S'il est atteint sur l'intérieur, on peut l'étudier à l'aide des résultats des Paragraphes 2.2 et 2.3 du Chapitre 2 ;
- Sinon il est atteint sur le bord et on peut essayer d'utiliser le Théorème 3.1.9.

Cette stratégie fonctionne dans certains cas, mais peut se heurter à plusieurs problèmes. En particulier, décrire le bord par des contraintes d'égalité n'est pas toujours une bonne idée. En effet, imaginons par exemple que nous ayons trois contraintes

- $h_1(x,y) = x \leqslant 0$;
- $h_2(x,y) = y \leqslant 0$;
- $h_3(x,y) = x + y + 1 \leqslant 0$.

et qu'un extremum local soit atteint en un point $\widetilde{x}$ tel que $h_1(\widetilde{x}) = 0 = h_2(\widetilde{x})$ mais $h_3(\widetilde{x}) < 0$. Les vecteurs $\nabla h_1(x,y) = (1,0)$ et $\nabla h_2(x,y) = (0,1)$ sont biens libres en tout point, mais

$$\nabla h_3(x,y) = \nabla h_1(x,y) + \nabla h_2(x,y)$$

donc si l'on cherche simplement les points critiques liées sous les conditions $h_1, h_2, h_3 = 0$, on ne peut appliquer le Théorème 3.1.9 et à raison puisqu'il n'y aura même pas nécessairement de solution. Nous avons par conséquent besoin d'une façon efficace de "tester" les différentes combinaisons de contraintes qui s'annulent ou non en un potentiel extremum, et c'est ce que nous allons maintenant développer.

3.2.1 Plus fort que Lagrange

Pour ce faire, il va nous falloir introduire un peu de vocabulaire. On considère toutjours une fonction $f : U \to \mathbf{R}$, mais cette fois-ci on cherche ses extrema dans une partie de $\mathbf{R}^n$ de la forme [11]

$$\mathcal{D} = \{x \in U \mid h_i(x) \leqslant 0 \text{ pour tout } 1 \leqslant i \leqslant p\},$$

pour des fonctions $h_1, \cdots, h_p : U \to \mathbf{R}$. Ces conditions sont plus larges que celles des contraintes d'égalité, donc il est a priori "plus difficile" d'être un extremum dans ce cas. Cela signifie que s'il existe un analogue du Théorème 3.1.9 pour les contraintes d'inégalité, il doit y avoir des conditions supplémentaires sur les multiplicateurs. Et de fait, il existe une condition assez simple pour cela, que nous allons maintenant donner.

Avant de commencer néanmoins, nous allons avoir besoin d'une notion pratique. Comme nous l'avons évoqué, l'un des enjeux du problème est de déterminer quelles inégalités sont strictement vérifiées ou non aux extrema. Pour en parler, un peu de vocabulaire s'impose.

Définition 3.2.1. Soient $h_1, \cdots, h_p : U \to \mathbf{R}$ des fonctions et $x \in U$. On dit que la contrainte h_i est *active* au point x si $h_i(x) = 0$. Autrement, on dit que la contrainte h_i est *inactive*.

Intuitivement, si une fonction f admet un extremum en un point $\widetilde{x}$, seules les contraintes qui sont actives en $\widetilde{x}$ devraient être prises en compte. En effet, si $h_j(\widetilde{x}) < 0$,

11. Il est clair que toute inégalité est équivalente à une inégalité de la forme $h(x) \leqslant 0$.

alors la même inégalité est vraie dans un voisinage de $\widetilde{x}$, et par conséquent f a toujours un extremum local si l'on retire la j-ième contrainte. Montrons cela proprement, en remarquant que quitte à renuméroter les fonctions, on peut supposer que c'est la première contrainte qui est inactive.

Lemme 3.2.2. *Soient $f, h_1, \cdots, h_p : U \to \mathbf{R}$ des fonctions de classe $\mathcal{C}^1$ et soit $\widetilde{x} \in U$ un point tel que f admet un extremum local en $\widetilde{x}$ sous les contraintes $h_j(x) \leqslant 0$ pour tout $1 \leqslant j \leqslant p$. Si $h_1(\widetilde{x}) < 0$, alors f admet un extremum local en $\widetilde{x}$ sous les contraintes $h_j(x) \leqslant 0$ pour $2 \leqslant j \leqslant p$.*

Démonstration. Notons, pour fixer les idées,

$$\mathcal{D} = \{x \in U \mid h_j(x) \leqslant 0 \text{ pour tout } 1 \leqslant j \leqslant p\}$$
$$\mathcal{D}' = \{x \in U \mid h_j(x) \leqslant 0 \text{ pour tout } 2 \leqslant j \leqslant p\}.$$

Puisque la restriction de f à $\mathcal{D}$ a un extremum local en $\widetilde{x}$, il existe $r_1 > 0$ tel que pour tout $y \in \mathcal{D} \cap B(\widetilde{x}, r_1)$, $f(y) \geqslant f(\widetilde{x})$. Par ailleurs, par continuité de h_1, il existe $r_2 > 0$ tel que pour tout $y \in B(\widetilde{x}, r_2)$,

$$h_1(y) \leqslant \frac{h_1(x)}{2} < 0.$$

Ceci implique que $\mathcal{D}' \cap B(\widetilde{x}, r_2) \subset \mathcal{D}$. Par conséquent, si $r = \min(r_1, r_2)$, on a pour tout

$$y \in \mathcal{D}' \cap B(\widetilde{x}, r) \subset \mathcal{D} \cap B(\widetilde{x}, r_1)$$

que $f(y) \geqslant f(\widetilde{x})$. Autrement dit, f a un minimum local en $\widetilde{x}$ sous les contraintes $h_j(x) \leqslant 0$ pour tout $2 \leqslant j \leqslant p$. ∎

Nous pouvons maintenant établir notre critère de localisation des extrema sous contraintes d'inégalité. Pour alléger la présentation, nous n'énoncerons le résultat que dans le cas d'un minimum. Pour un maximum, il suffira de se souvenir qu'un maximum de f n'est autre qu'un minimum de $-f$.

Proposition 3.2.3. *Soit $f : U \to \mathbf{R}$ une fonction de classe $\mathcal{C}^1$, $h_1, \cdots, h_p : U \to \mathbf{R}$ des fonctions également de classe $\mathcal{C}^1$ et $\widetilde{x} \in U$. On suppose qu'il existe $\xi_0 \in \mathbf{R}^n$ tel que pour tout $1 \leqslant j \leqslant p$ **tel que la contrainte h_j est active au point $\widetilde{x}$,***

$$\left\langle \nabla h_j(\widetilde{x}), \xi_0 \right\rangle < 0.$$

Si la restriction de f à $\mathcal{D}$ admet un minimum *local en $\widetilde{x}$, alors il existe $\mu_1, \cdots, \mu_p \leqslant 0$ tels que*

$$\boxed{\nabla f(\widetilde{x}) = \sum_{j=1}^{p} \mu_j \nabla h_j(\widetilde{x})}$$

De plus, on a $\mu_j = 0$ si la contrainte h_j est inactive au point $\widetilde{x}$.

La condition sur les produits scalaires est cruciale. On dit que les contraintes sont *qualifiées en $\widetilde{x}$* si un tel vecteur ξ_0 existe. Quant au fait que le résultat requiert des coefficients négatifs, cela provient du résultat suivant, qui est une forme simplifiée – mais suffisante pour nous – d'un résultat fondamental d'analyse convexe appelé Lemme de Farkas [12].

12. Gyulia Farkas de Kisbarnak (1847–1930) : mathématicien et physicien hongrois, il a cherché à prolonger le travail de Lagrange sur l'optimisation sous contraintes en prenant en compte les inégalités, ce qui l'a entre autres choses mené à démontrer le lemme qui porte son nom.

Lemme 3.2.4. *Soient $u, v_1, \cdots, v_p \in \mathbf{R}^n$ tels que pour tout $x \in \mathbf{R}^n$, si $\langle v_j, x\rangle \leqslant 0$ pour tout $1 \leqslant j \leqslant p$, alors $\langle u, x\rangle \geqslant 0$. Alors, il existe $\mu_1, \cdots, \mu_p \in \mathbf{R}_-$ tels que*

$$u = \sum_{j=1}^{p} \mu_j v_j.$$

Nous allons donner d'abord la preuve de ce lemme, puis démontrer en l'utilisant la Proposition 3.2.3.

Démonstration du Lemme 3.2.4. On procède par récurrence sur p, avec l'hypothèse suivante :

$$H_p : \text{« Soient } u, v_1, \cdots, v_p \in \mathbf{R}^n \text{ tels que pour tout } x \in \mathbf{R}^n, \text{ si } \langle x, v_i\rangle \leqslant 0 \text{ pour tout } 1 \leqslant i \leqslant p, \text{ alors } \langle u, x\rangle \geqslant 0. \text{ Alors, il existe } \mu_1, \cdots, \mu_p \in \mathbf{R}_- \text{ tels que } u = \textstyle\sum_{j=1}^{p} \mu_j v_j. \text{ »}$$

Pour initialiser, considérons deux vecteurs $u, v \in \mathbf{R}^n$ tels que pour tout $x \in \mathbf{R}^n$, $\langle v, x\rangle \leqslant 0$ implique $\langle u, x\rangle \geqslant 0$. Supposons qu'il n'existe pas de réel $\mu \in \mathbf{R}_-$ tel que $u = \mu v$, et remarquons qu'il n'existera alors pas non plus de réel $\mu \in \mathbf{R}_+^*$ tel que $u = \mu v$ puisqu'alors l'hypothèse ne serait pas vérifiée pour $x = -v$. Autrement dit, u et v ne sont pas colinéaires. Par conséquent, si u' est le projeté orthogonal de u sur $v^\perp$, on a $\langle v, -u'\rangle = 0$ et $\langle u, -u'\rangle = -\|u'\|^2 < 0$, une contradiction.

On suppose désormais H_p vraie et on considère des vecteurs $u, v_1, \cdots, v_{p+1} \in \mathbf{R}^n$. Notons P la projection orthogonale sur $v_{p+1}^\perp$ et observons que pour tout $x \in v_{p+1}^\perp$, on a $\langle P(u), x\rangle = \langle u, x\rangle$ et $\langle P(v_i), x\rangle = \langle v_i, x\rangle$. Ainsi, si $\langle P(v_i), x\rangle \leqslant 0$ pour tout $1 \leqslant i \leqslant p$, alors $\langle P(u), x\rangle \geqslant 0$. On peut donc appliquer H_p pour conclure qu'il existe $\mu_1, \cdots, \mu_p \in \mathbf{R}_-$ tels que

$$P(u) = \sum_{j=1}^{p} \mu_j P(v_j).$$

Posons maintenant

$$u' = u - \sum_{j=1}^{p} \mu_j v_j.$$

Comme $P(u') = 0$, $u' \in \mathrm{Vect}(v_{p+1})$ donc il existe $\mu_{p+1} \in \mathbf{R}$ tel que $u' = \mu_{p+1} v_{p+1}$, c'est-à-dire

$$u = \sum_{j=1}^{p+1} \mu_j v_j.$$

Si $\mu_{p+1} \leqslant 0$, alors la démonstration est terminée. Supposons donc $\mu_{p+1} > 0$. Alors, étant donné $x \in \mathbf{R}^n$ tel que $\langle v_i, x\rangle \leqslant 0$ pour tout $1 \leqslant i \leqslant p$,

- Si $\langle v_{p+1}, x\rangle \leqslant 0$, alors par hypothèse $\langle u, x\rangle \geqslant 0$;
- Si $\langle v_{p+1}, x\rangle > 0$, alors $\langle u, x\rangle > 0$ d'après la décomposition ci-dessus.

Ainsi, dans tous les cas $\langle u, x\rangle \geqslant 0$. On peut donc simplement appliquer H_p pour conclure qu'il existe $\mu_1', \cdots, \mu_p' \in \mathbf{R}_-$ tels que

$$u = \sum_{j=1}^{p} \mu_j' v_j$$

et conclure en posant $\mu_{p+1}' = 0$. ∎

Nous sommes maintenant armés pour démontrer notre résultat d'optimisation sous contraintes d'inégalité.

Démonstration de la Proposition 3.2.3. D'après le Lemme 3.2.2, on peut supposer [13] que $h_j(\widetilde{x}) = 0$ pour tout $1 \leqslant j \leqslant p$ ou – avec la terminologie introduite plus haut – supposer que toutes les contraintes sont actives. C'est la raison pour laquelle les hypothèses ne font intervenir que les contraintes actives.

Soit maintenant $\epsilon > 0$ et $\xi_0 \in \mathbf{R}^n$ tel que $\langle \nabla h_i, \xi_0 \rangle < 0$ pour tout $1 \leqslant i \leqslant p$. Pour un vecteur $\xi_1 \in \mathbf{R}^n$ vérifiant

$$\langle \nabla h_i, \xi_1 \rangle \leqslant 0$$

pour tout $1 \leqslant i \leqslant p$, on pose

$$\xi : t \mapsto \widetilde{x} + t(\xi_1 + \epsilon\xi_0).$$

En remarquant que $\xi(0) = \widetilde{x}$, nous allons montrer que pour t assez petit, $\xi(t) \in \mathcal{D}$. De fait, la fonction h_i étant de classe $\mathcal{C}^1$, on a par la Proposition 2.3.9

$$\begin{aligned} h_i(\xi(t)) &= h_i(\widetilde{x}) + \langle \nabla h_i(\widetilde{x}), t(\xi_1 + \epsilon\xi_0) \rangle + o(t) \\ &= t\left(\langle \nabla h_i(\widetilde{x}), \xi_1 + \epsilon\xi_0 \rangle + \varepsilon(t)\right) \end{aligned}$$

où $\varepsilon(t) \to 0$ quand $t \to 0$. Soit $\delta_i > 0$ tel que pour tout $t \in]0;\delta_i[$,

$$\langle \nabla h_i(\widetilde{x}), \xi_1 + \epsilon\xi_0 \rangle + \varepsilon(t) \leqslant \frac{\langle \nabla h_i(\widetilde{x}), \xi_1 + \epsilon\xi_0 \rangle}{2}$$

(un tel t existe car le membre de droite tend vers $\langle \nabla h_i(\widetilde{x}), \xi_1 + \epsilon\xi_0 \rangle < 0$). Alors, pour ces mêmes valeurs de t,

$$\begin{aligned} h_i(\xi(t)) &\leqslant t\frac{\langle \nabla h_i(\widetilde{x}), \xi_1 + \epsilon\xi_0 \rangle}{2} \\ &\leqslant t\epsilon\frac{\langle \nabla h_i(\widetilde{x}), \xi_0 \rangle}{2} \\ &< 0. \end{aligned}$$

En particulier, $h_i(\xi(t)) < 0$. On posant $\delta = \min_i \delta_i$, on donc que $\xi(t) \in \mathcal{D}$ pour tout $t \in]0;\delta[$.

Revenons maintenant à notre problème d'origine. Pour tout $t \in]0;\delta[$, on a

$$\frac{f(\xi(t)) - f(\widetilde{x})}{t} \geqslant 0.$$

En faisant tendre t vers 0, on en déduit que

$$\langle \nabla f(\widetilde{x}), \xi_1 + \epsilon\xi_0 \rangle \geqslant 0.$$

Cette inégalité est valable pour tout $\epsilon > 0$, donc en faisant tendre ϵ vers 0 on en déduit que

$$\langle \nabla f(\widetilde{x}), \xi_1 \rangle \geqslant 0.$$

Nous pouvons maintenant appliquer le Lemme 3.2.4 à $\nabla f(\widetilde{x})$ et à la famille de vecteurs $-\nabla h_1(\widetilde{x}), \cdots, -\nabla h_p(\widetilde{x})$ pour conclure. ∎

Ce résultat appelle plusieurs commentaires importants.

13. On dit alors que les contraintes sont *saturées en* $\widetilde{x}$.

- La condition de qualification est plus "souple" que la condition d'indépendance linéaire des gradients du Théorème 3.1.9. En effet, elle est vérifiée par exemple pour deux fonctions h_1 et h_2 telles que $\nabla h_1(x) = \nabla h_2(x)$.
- La condition $\mu_j \leqslant 0$ peut s'interpréter géométriquement dans le cadre des sous-variétés. Par exemple, si on a une seule contrainte d'inégalité $h_j(x) \leqslant 0$ et que $\nabla h_j(x) \neq 0$ pour tout x, le signe de μ_j signifie que non seulement le gradient de f doit être orthogonal au plan tangent, mais qu'il doit de plus "entrer" dans le domaine $\mathcal{D} = \{x \in U \mid h_j(x) \leqslant 0\}$. Ceci permet donc d'éliminer certains points critiques liés qui ne peuvent correspondre à des extrema.
- La dernière ligne de l'énoncé peut s'écrire de façon plus concise

  $$\mu_j h_j(\widetilde{x}) = 0.$$

  Cette condition est parfois dite de *complémentarité* et elle est essentielle pour l'utilisation du résultat. En effet, elle permet de "tester" la nullité ou non des multiplicateurs pour voir quelles contraintes peuvent être actives et donc quels gradients peuvent apparaître avec un coefficient non nul (un exemple est donné au Paragraphe 3.2.2 ci-dessous).

Remarque 3.2.5. Nous pourrions être tentés de jouer les malins et d'utiliser la Proposition 3.2.3 pour obtenir une preuve plus simple du Théorème 3.1.9. En effet, une contrainte d'égalité $g = 0$ est équivalente à deux contraintes d'inégalité, à savoir $g \leqslant 0$ et $g \geqslant 0$. Néanmoins, de telles contraintes ne seront jamais qualifiées. En effet, si $\xi_0 \in \mathbf{R}^n$, on ne peut avoir à la fois $\langle \nabla g(\widetilde{x}), \xi_0 \rangle < 0$ et $\langle \nabla(-g)(\widetilde{x}), \xi_0 \rangle < 0$ puisque ces deux nombres sont opposés. Ainsi, la qualification des contraintes est essentielle et restreint l'utilisation du résultat à des contraintes d'inégalité non-triviales.

Pour remarquable qu'il soit, le résultat précédent est insuffisant. En effet, idéalement, nous voulons pouvoir gérer à la fois des contraintes d'égalité et d'inégalité, ce qu'on appelle parfois – dans un grand élan d'imagination – des *contraintes mixtes*. Pour cela, on ne peut juxtaposer le Théorème 3.1.9 et la Proposition 3.2.3, il faut comprendre comment ils interagissent. Heureusement, le Théorème 3.1.24 nous donne une idée : la bonne solution devrait être d'exprimer les conditions de la Proposition 3.2.3 non pas sur tout l'espace mais sur le sous-espace vectoriel E qui est le noyau de la différentielle de la fonction g encodant les contraintes d'égalité.

De fait, cette solution est la bonne. Elle fut exprimée pour la première fois par Karush [14] en 1939, mais son résultat ne fut pas très remarqué. Il fut redécouvert par Kuhn [15] et Tucker [16] en 1951 et porte donc le nom de Condition de Karush-Kuhn-Tucker, que nous abrégerons en KKT.

Théorème 3.2.6 (Condition KKT) Soit $f : U \to \mathbf{R}$ une fonction de classe $\mathcal{C}^1$, soient $g_1, \cdots, g_m, h_1, \cdots, h_p : U \to \mathbf{R}$ des fonctions également de classes $\mathcal{C}^1$ et $\widetilde{x} \in \mathcal{D}$. On suppose que les vecteurs

$$\nabla g_1(\widetilde{x}), \cdots, \nabla g_m(\widetilde{x})$$

14. William Karush (1917–1997) : mathématicien américain, il a obtenu les conditions d'extrema sous contrainte d'inégalités dans son mémoire de Master, qu'il n'a pas publié. Il a travaillé plus tard pour le *Projet Manhattan* avant de devenir militant pacifiste.

15. Harold W. Kuhn (1925–2014) : mathématicien canadien, spécialiste entre autres de théorie des jeux. On lui doit une étude détaillée d'une version simplifiée du jeu de Poker, appelé *Poker de Kuhn*.

16. Albert W. Tucker (1905–1995) : mathématicien canadien au multiples centres d'intérêts (géométrie différentielle, théorie des jeux, optimisation). La *Mathematical Optimisation Society* décerne tous les trois ans un prix en son honneur.

forment une famille libre et qu'il existe

$$\xi_0 \in E = \bigcap_{i=1}^{m} \ker D_{\widetilde{x}} g_i$$

tel que pour tout $1 \leqslant j \leqslant p$ **tel que la contrainte h_j est active au point $\widetilde{x}$,**

$$\left\langle \nabla h_j(\widetilde{x}), \xi_0 \right\rangle < 0.$$

Si f admet un *minimum* local en $\widetilde{x}$ sous les contraintes

- $g_i = 0$ pour tout $1 \leqslant i \leqslant m$;
- $h_j \leqslant 0$ pour tout $1 \leqslant j \leqslant p$;

alors il existe $\lambda_1, \cdots, \lambda_m \in \mathbf{R}$ et $\mu_1, \cdots, \mu_p \leqslant 0$ tels que

$$\boxed{\nabla f(\widetilde{x}) = \sum_{i=1}^{m} \lambda_i \nabla g_i(\widetilde{x}) + \sum_{j=1}^{p} \mu_j \nabla h_j(\widetilde{x})}$$

De plus, $\mu_j = 0$ si la contrainte h_j est inactive au point $\widetilde{x}$.

Démonstration. En nous inspirant des notations de la preuve de la Proposition 3.2.3, posons $\widetilde{\chi} = (\widetilde{x}_1, \cdots, \widetilde{x}_{n-m})$ et pour $\xi_1 \in E$ et $t > 0$,

$$\xi(t) = \widetilde{\chi} + t(\widetilde{\xi_1} + \epsilon \widetilde{\xi_0}),$$

où $\widetilde{\xi_i} \in \mathbf{R}^n$ est le vecteur formé des $n-m$ premières coordonnées de ξ_i pour $i = 0, 1$. Pour construire maintenant une fonction dont les valeurs sont des vecteurs satisfaisant les contraintes d'égalité, il suffit de poser – en notant φ la fonction implicite donnée par le Théorème 3.1.6 appliquée à $g_1, \cdots, g_m$ qui vérifient bien les hypothèses –

$$\zeta(t) = (\xi(t), \varphi(\xi(t)) \in \mathbf{R}^n.$$

Attention cependant : cette expression n'a pas de sens pour tout t. Néanmoins, comme $\xi(0) = \widetilde{x}$ et que $t \mapsto \xi(t)$ est continue car linéaire, il existe $\delta > 0$ tel que pour $t \in [0; \delta[$, $\xi(t) \in V$ et la fonction ζ est alors bien définie sur $[0; \delta[$.

Il nous faut maintenant vérifier que pour t suffisamment petit, $\zeta(t)$ vérifie les contraintes d'inégalité. Nous allons pour cela distinguer suivant qu'elles sont actives ou non. Si $h_j(\widetilde{x}) < 0$, alors il existe δ_j tel que $h_j(x) < 0$ pour tout $x \in B(\widetilde{x}, \delta_j)$. Par continuité de ζ, il existe δ_j' tel que $\zeta(t) \in B(\widetilde{x}, \delta_j)$ pour tout $t \in [0; \delta_j'[$ et en définissant δ' comme le minimum de δ et de tous les δ_j pour les contraintes inactives, on a que $h_j(\zeta(t)) < 0$ si $t \in [0; \delta'[$. Supposons maintenant que $h_j(\widetilde{x}) = 0$. Le point clef est le calcul suivant, reposant sur la Proposition 3.1.8 et où nous notons comme d'habitude g la fonction de coordonnées $(g_1, \cdots, g_m)$ et $\Phi : V \to \mathbf{R}$ la fonction définie par $\Phi(z) = (z, \varphi(z))$:

$$\begin{aligned}
\zeta(t) &= \Phi(\xi(t)) \\
&= \Phi\left(\widetilde{\chi} + t(\widetilde{\xi_1} + \epsilon \widetilde{\xi_0})\right) \\
&= \Phi(\widetilde{\chi}) + D_{\widetilde{\chi}}\Phi\left(t(\widetilde{\xi_1} + \epsilon \widetilde{\xi_0})\right) + o(t) \\
&= \widetilde{x} - \left(D_{\widetilde{x},2} g\right)^{-1} \circ D_{\widetilde{x},1} g\left(t(\widetilde{\xi_1} + \epsilon \widetilde{\xi_0})\right) + o(t)
\end{aligned}$$

Pour poursuivre le calcul, rappelons que par définition, $E = \ker D_{\widetilde{x}}g$. Par conséquent, si $\eta \in E$, alors

$$\begin{aligned} D_{\widetilde{x},1}g(\eta_1,\cdots,\eta_{n-m}) &= \sum_{i=1}^{n-m} \eta_i \frac{\partial g}{\partial x_i}(\widetilde{x}) \\ &= \sum_{i=1}^{n} \eta_i \frac{\partial g}{\partial x_i}(\widetilde{x}) - \sum_{i=n-m+1}^{n} \eta_i \frac{\partial g}{\partial x_i}(\widetilde{x}) \\ &= D_{\widetilde{x}}g(\xi) - D_{\widetilde{x},2}g(\eta) \\ &= -D_{\widetilde{x},2}g(\eta). \end{aligned}$$

On en déduit alors que

$$\zeta(t) = \widetilde{x} + t(\xi_1 + \epsilon\xi_0) + o(t)$$

d'où

$$\begin{aligned} h_j(\zeta(t)) &= h_j(\widetilde{x}) + \langle \nabla h_j(\widetilde{x}), t(\xi_1 + \epsilon\xi_0)\rangle + o(t) \\ &= t\langle \nabla h_j(\widetilde{x}), \xi_1 + \epsilon\xi_0\rangle + o(t). \end{aligned}$$

Le même raisonnement que dans la démonstration de la Proposition 3.2.3 montre alors que si $\langle \nabla h_j(\widetilde{x}), \xi_1\rangle \leqslant 0$, pour t assez petit on aura $h_j(\zeta(t)) \leqslant 0$.

Nous pouvons maintenant reprendre exactement le même raisonnement que dans la démonstration de la Proposition 3.2.3 pour conclure que

$$\langle \nabla f(\widetilde{x}), \xi_1\rangle \geqslant 0.$$

Soit P_E la projection orthogonale sur E. On a alors

$$\left\langle \nabla h_j(\widetilde{x}), \xi_1 \right\rangle = \left\langle P_E\left(\nabla h_j(\widetilde{x})\right), \xi_1 \right\rangle \quad \& \quad \langle \nabla f(\widetilde{x}), \xi_1\rangle = \langle P_E(\nabla f(\widetilde{x})), \xi_1\rangle,$$

donc on peut appliquer la Proposition 3.2.3 pour conclure à l'existence de $\mu_1,\cdots,\mu_p \leqslant 0$ tels que

$$P_E(\nabla f(\widetilde{x})) = \sum_{j=1}^{p} \mu_j P_E\left(\nabla h_j(\widetilde{x})\right).$$

Pour conclure, posons

$$X = \nabla f(\widetilde{x}) - \sum_{j=1}^{p} \mu_j \nabla h_j(\widetilde{x}).$$

C'est un vecteur de $\mathbf{R}^n$ et $P_E(X) = 0$. Par conséquent, $X \in E^\perp$ et ce dernier espace est engendré par les gradients des fonctions $\nabla g_i(\widetilde{x})$, d'où le résultat. ∎

3.2.2 Une page de publicité

Le Théorème 3.2.6 n'est pas évident à utiliser. En effet, il faut bien comprendre ce qu'il dit. Sa subtilité est de permettre de tester facilement les diverses configurations de contraintes qui sont actives ou non. Un exemple permettra de mieux s'y retrouver.

On considère une entreprise qui produit un bien. Son objectif est de maximiser son revenu – c'est-à-dire les ventes – tout en maintenant son profit au-dessus d'une certaine valeur p_0. Pour ce faire, elle va faire de la publicité afin de doper les ventes.

Plus précisément, on suppose que la production de x unités du bien coûte $C(x)$ et que le revenu qu'on en tire à l'aide d'un coût de publicité a est $R(x,a)$. Le problème est donc de maximiser la fonction R sous les contraintes

$$\begin{cases} R(x,a) - C(x) - a & \geqslant & p_0 \\ x & \geqslant & 0 \\ a & \geqslant & 0 \end{cases}$$

Supposons qu'il existe un maximum $(\widetilde{x},\widetilde{a})$ avec $\widetilde{x} > 0$, de sorte que la contrainte $x \geqslant 0$ est inactive et regardons ce qu'impose la qualification des contraintes. Il ne nous reste que deux contraintes actives, $h_1(x) = p_0 - R(x,a) + C(x) + a$ et $h_2(x,a) = -a$, dont les gradients sont

$$\nabla h_1(\widetilde{x},\widetilde{a}) = \begin{pmatrix} -\dfrac{\partial R}{\partial x}(\widetilde{x},\widetilde{a}) + C'(\widetilde{x}) \\ -\dfrac{\partial R}{\partial a}(\widetilde{x},\widetilde{a}) + 1 \end{pmatrix} \quad \& \quad \nabla h_2(\widetilde{x},\widetilde{a}) = \begin{pmatrix} 0 \\ -1 \end{pmatrix}$$

et il faut donc trouver un vecteur $\xi_0 = (\xi_1, \xi_2)$ tel que

$$\begin{aligned} \xi_1 \left(C'(\widetilde{x}) - \frac{\partial R}{\partial x}(\widetilde{x},\widetilde{a}) \right) + \xi_2 \left(1 - \frac{\partial R}{\partial a}(\widetilde{x},\widetilde{a}) \right) &< 0 \\ -\xi_2 &< 0 \end{aligned}$$

La première équation peut s'écrire

$$\xi_1 \left(C'(\widetilde{x}) - \frac{\partial R}{\partial x}(\widetilde{x},\widetilde{a}) \right) < \xi_2 \left(\frac{\partial R}{\partial a}(\widetilde{x},\widetilde{a}) - 1 \right)$$

qui donne en divisant (puisque $\xi_2 \neq 0$)

$$\frac{\xi_1}{\xi_2} \left(C'(\widetilde{x}) - \frac{\partial R}{\partial x}(\widetilde{x},\widetilde{a}) \right) < \left(\frac{\partial R}{\partial a}(\widetilde{x},\widetilde{a}) - 1 \right).$$

La seule situation pour laquelle on ne peut réaliser cette inégalité est si

$$C'(\widetilde{x}) = \frac{\partial R}{\partial x}(\widetilde{x},\widetilde{a}) \quad \& \quad \frac{\partial R}{\partial a}(\widetilde{x},\widetilde{a}) \leqslant 1.$$

Nous supposerons dorénavant que cette condition n'est pas remplie.

Il faut maintenant être précautionneux : la Proposition 3.2.3 s'applique à un minimum, et nous considérons ici un maximum. Il faut donc prendre l'opposé de la fonction qui nous intéresse, c'es-à-dire la fonction $-R$. Alors, il existe $\mu_1, \mu_2 \leqslant 0$ tels que

$$\begin{cases} -\frac{\partial R}{\partial x}(\widetilde{x},\widetilde{a}) & = & -\mu_1 \frac{\partial R}{\partial x}(\widetilde{x},\widetilde{a}) + \mu_1 C'(x) \\ -\frac{\partial R}{\partial a}(\widetilde{x},\widetilde{a}) & = & -\mu_1 \frac{\partial R}{\partial a}(\widetilde{x},\widetilde{a}) + \mu_1 - \mu_2 \\ \mu_1 \left(p_0 - R(\widetilde{x},\widetilde{a}) + C(\widetilde{x}) + \widetilde{a} \right) & = & 0 \\ -\mu_2 a & = 0 & \end{cases}$$

Considérons tout d'abord la seconde égalité. Elle peut s'écrire

$$\mu_2 - \mu_1 = (1 - \mu_1) \frac{\partial R}{\partial a}(\widetilde{x},\widetilde{a}).$$

Supposons que la dérivée partielle de R par rapport à a est toujours strictement positive, ce qui paraît raisonnable (plus de publicité augmente les ventes). Alors, comme $1-\mu_1>0$, on a $\mu_2-\mu_1>0$ et puisque $\mu_1,\mu_2\leqslant 0$ on doit finalement avoir $\mu_1<0$. Mais alors, la troisième condition donne

$$R(\widetilde{x},\widetilde{a})-C(\widetilde{x})-\widetilde{a}=p_0.$$

Autrement dit, s'il y a un maximum il est atteint quand le profit est minimal !

On peut même dire un peu plus. En effet, le coût doit être une fonction strictement croissante (plus on produit plus ça coûte) de x, donc $C'(\widetilde{x})>0$. Comme $\mu_1<0$, on conclut de la première équation que

$$\frac{\partial R}{\partial x}(\widetilde{x},\widetilde{a})=\frac{\mu_1}{\mu_1-1}C'(x)>0.$$

Ainsi, le revenu marginal est strictement positif à l'optimum, ce qui signifie que produire un peu plus que $\widetilde{x}$ augmenterait le revenu. Ceci peut paraître surprenant puisqu'on est sensé être à un maximum. Néanmoins, si on note $P(x,a)$ le profit, on a

$$\begin{aligned}\frac{\partial P}{\partial x}(\widetilde{x},\widetilde{a})&=\frac{\partial R}{\partial x}(\widetilde{x},\widetilde{a})-C'(x)\\&=\frac{1}{1-\mu_1}C'(x)\\&=-C'(x)\\&<0.\end{aligned}$$

Ainsi, produire plus réduirait le profit, le faisant passer sous le seuil imposé. C'est pour cela qu'on a bien un maximum quand les contraintes sont prises en compte.

3.2.3 Qualification linéaire des contraintes

L'exemple précédent montre que vérifier la qualification des contraintes n'est pas toujours chose aisée. Souvent, il est possible d'aller plus vite en vérifiant une condition plus forte, dite de *qualification linéaire des contraintes*.

Définition 3.2.7. Soient $g_1,\cdots,g_m,h_1,\cdots,h_p:U\to\mathbf{R}$ des fonctions de classes $\mathcal{C}^1$ et $x\in U$ tel que $g_i(x)=0$ pour tout $1\leqslant i\leqslant m$ et $h_j(x)\leqslant 0$ pour tout $1\leqslant j\leqslant p$. On dit que la qualification linéaire des contraintes est vérifiée au point x si la famille de vecteurs

$$P_E(\nabla h_1(x)),\cdots,P_E(\nabla h_p(x))$$

est libre, où

$$E=\bigcap_{1\leqslant i\leqslant m}\ker(D_xg_i)$$

Si l'on oublie les projections sur l'espace E, il s'agit en fait de la condition du Théorème 3.1.9. Comme nous l'avions évoqué plus haut, les conditions de la Proposition 3.2.3 sont moins fortes que la qualification linéaire des contraintes. Néanmoins, il arrive souvent que les contraintes vérifient les conditions de qualification linéaire, qui sont plus simple à établir. Pour le démontrer, nous allons nous appuyer sur un résultat classique d'algèbre linéaire dû à Gram [17], que nous allons d'abord démontrer au cas où vous n'auriez pas eu le plaisir de le rencontrer sur le chemin de vos études.

17. Jorgen Pedersen Gram (1850–1916) : mathématicien danois qui, en plus de ses contributions importantes à l'algèbre linéaire comme le procédé d'orthonormalisation qui porte son nom, a travaillé la majeure partie de sa vie dans des compagnies d'assurance.

Proposition 3.2.8. *Soient $v_1, \cdots, v_p$ des vecteurs de $\mathbf{R}^n$. On pose $M_{ij} = \langle v_i, v_j \rangle$ et on note $M \in \mathrm{M}_p(\mathbf{R})$ la matrice donnée par ces coefficients. Si la famille de vecteurs est libre, alors M est inversible.*

Démonstration. Notons

$$A = \begin{pmatrix} v_1 & | & \cdots & | & v_p \end{pmatrix} \in \mathrm{M}_{pn}(\mathbf{R})$$

la matrice dont les colonnes sont les vecteurs $v_1, \cdots, v_p$. On a alors l'égalité $M = A^t A$:

$$\begin{aligned}(A^t A)_{ij} &= \sum_{k=1}^{n} A_{ki} A_{kj} \\ &= \sum_{k=1}^{n} (v_i)_k (v_j)_k \\ &= \langle v_i, v_j \rangle \\ &= M_{ij}.\end{aligned}$$

Il suffit donc de montrer que si la famille est libre, alors $\ker(M) = \ker(A^t A) = \{0\}$. Et de fait, si $x \in \ker(A^t A)$, alors

$$\begin{aligned}0 &= \langle A^t A x, x \rangle \\ &= \langle Ax, Ax \rangle \\ &= \|Ax\|^2,\end{aligned}$$

donc $x \in \ker(A)$. Cela signifie que

$$\sum_{i=1}^{n} x_i v_i = 0,$$

donc que $x = 0$. Ainsi, $\ker(M) = 0$, ce qu'il fallait démontrer. ∎

Nous pouvons maintenant montrer que la qualification linéaire des contraintes implique la qualification au sens de la Proposition 3.2.3.

Proposition 3.2.9. *Si la qualification linéaire des contraintes est vérifiée en un point x, alors les contraintes sont qualifiées en x au sens de la Proposition 3.2.3.*

Démonstration. Posons, pour $1 \leqslant j \leqslant p$, $w_j = P_E(\nabla h_j(x))$. Il suffit maintenant de trouver un vecteur $\xi_0 \in E$ tel que $\langle w_i, \xi_0 \rangle = -1$. Pour ce faire, nous allons le chercher sous la forme particulière

$$\xi_0 = \sum_{j=1}^{p} \alpha_j w_j.$$

La condition s'écrit alors, pour tout $1 \leqslant j \leqslant p$,

$$\sum_{j'=1}^{p} \alpha_{j'} \langle w_{j'}, w_j \rangle = -1.$$

Si $M \in \mathrm{M}_p(\mathbf{R})$ est la matrice de coefficients $M_{jj'} = \langle w_{j'}, w_j \rangle$ et si $\eta \in \mathbf{R}^p$ est le vecteur dont toutes les coordonnées sont égales à -1, alors l'équation devient $M\alpha = \eta$. Or, M est la *matrice de Gram* de la famille de vecteurs $(w_1, \cdots, w_p)$. Comme la matrice de Gram d'une famille libre est inversible par la Proposition 3.2.8, on conclut qu'un tel ξ_0 existe, ce qu'il fallait démontrer. ∎

3.3 Le cas linéaire

La mort est un processus rectiligne.

D. Pennac, La petite marchande de prose

Avant de poursuivre, faisons une pause pour voir ce que le Théorème 3.2.6 nous dit sur le problème évoqué au Paragraphe 1.1.2.Pour cela, commençons par remarquer qu'il a une structure très particulière : il est affine ! En effet, tant la fonction à minimiser que les contraintes sont données par des fonctions affines. On imagine que pour ce genre de problème les choses doivent être simples, mais une petite surprise nous attend. Avant tout, un brin de vocabulaire : il y a un abus systématique qui veut qu'un problème dont toutes les fonctions et contraintes sont affines soit appelé *linéaire*. Nous nous plierons avec résignation à cette terminologie consacrée par l'usage.

3.3.1 Le KKT amélioré

La fonction f, si elle est affine, est de la forme $x \mapsto \langle u, x\rangle + b$ pour un certain vecteur $u \in \mathbf{R}^n$ et un réel $b \in \mathbf{R}$. De même, on a $g_i(x) = \langle v_i, x\rangle - c_i$ et $h_j(x) = \langle w_j, x\rangle - d_j$. La condition du Théorème 3.2.6 devient alors

$$u = \sum_{i=1}^{m} \lambda_i v_i + \sum_{j=1}^{p} \mu_j w_j.$$

La variable x a disparu ! Pour y voir plus clair, introduisons quelques notations bien pratiques. On définit des matrices à l'aide des vecteurs v_i et w_j de la façon suivante,

$$G = \left(\begin{array}{c} v_1^t \\ \hline \vdots \\ \hline v_m^t \end{array}\right) \in \mathrm{M}_{mn}(\mathbf{R})$$

$$H = \left(\begin{array}{c} w_1^t \\ \hline \vdots \\ \hline w_p^t \end{array}\right) \in \mathrm{M}_{pn}(\mathbf{R})$$

de sorte que les contraintes peuvent s'écrire $Gx - c = 0$ et $Hx - d \leqslant 0$, l'inégalité étant comprise coordonnée par coordonnée. La condition ci-dessus devient alors

$$u = G^t\lambda + H^t\mu.$$

Cette condition, réunie aux contraintes, s'avère être en fait suffisante pour l'existence d'un extremum. Autrement dit, le Théorème 3.2.6 devient non plus un critère, mais une équivalence dans le cas linéaire.

Théorème 3.3.1 (KKT pour les problèmes linéaires) Un point $\widetilde{x} \in \mathbf{R}^n$ réalise un minimum global d'une fonction linéaire f sous les contraintes linéaires $g_i = 0$ et $h_j \leqslant 0$ si et seulement s'il existe $\lambda \in \mathbf{R}^m$ et $\mu \in \mathbf{R}^p_-$ tels que pour tout $1 \leqslant j \leqslant p$,

$$\left\{\begin{array}{rcl} u & = & G^t\lambda + H^t\mu \\ G\widetilde{x} & = & c \\ H\widetilde{x} & \leqslant & d \\ \mu_j\langle w_j, \widetilde{x}\rangle & = & \mu_j d_j \end{array}\right.$$

‖

Démonstration. Il y a trois informations essentielles dans cet énoncé :

- Aucune qualification des contraintes n'est requise pour que la condition nécessaire du Théorème 3.2.6 soit valide.
- La conclusion qui n'est que nécessaire dans le Théorème 3.2.6 est aussi suffisante.
- Tout extremum local est global.

Les deux derniers points se démontrent en fait en même temps, et nous allons donc diviser la démonstration en deux parties.

► *Fi de la qualification*

Supposons qu'on dispose d'un minimum local sous contraintes en un point $\widetilde{x}$ et notons

$$J = \{1 \leqslant j \leqslant p \mid h_j(\widetilde{x}) = 0\}$$

l'ensemble des indices des contraintes actives au point $\widetilde{x}$. L'idée est d'essayer d'appliquer directement le Lemme 3.2.4 sans passer par les constructions de la démonstration de la Proposition 3.2.3.

Pour ce faire, considérons un vecteur $\xi_0 \in \mathbf{R}^n$ satisfaisant

1. $\langle v_i, \xi_0 \rangle = 0$ pour tout $1 \leqslant i \leqslant m$.
2. $\langle w_j, \xi_0 \rangle \leqslant 0$ pour tout $j \in J$.

Alors, nous affirmons que pour t suffisamment petit, le vecteur $\xi(t) = \widetilde{x} + t\xi_0$ est dans $\mathcal{D}$. En effet, si $j \notin J$, alors $h_j(\widetilde{x}) < 0$, donc il existe $r_j > 0$ tel que $h_j(x) < 0$ pour tout $x \in B(\widetilde{x}, r_j)$. En posant

$$r = \min_{j \notin J} r_j,$$

on a que $h_j(x) < 0$ pour tout $j \notin J$ et donc en particulier $h_j(\xi(t)) < 0$ pour $t \in [0; r[$. De plus, pour les mêmes t, on a pour tout $1 \leqslant i \leqslant m$

$$\begin{aligned} g_i(\xi(t)) &= \langle v_i, \widetilde{x} + t\xi_0 \rangle - c_i \\ &= \langle v_i, \widetilde{x} \rangle - c_i + t\langle v_i, \xi_0 \rangle \\ &= g_i(\widetilde{x}) + t\langle v_i, \xi_0 \rangle \\ &= 0 \end{aligned}$$

et pour tout $j \in J$,

$$\begin{aligned} h_j(\xi(t)) &= \langle w_j, \widetilde{x} + t\xi_0 \rangle - d_j \\ &= \langle w_j, \widetilde{x} \rangle - d_j + t\langle w_j, \xi_0 \rangle \\ &= h_j(\widetilde{x}) + t\langle w_j, \xi_0 \rangle \\ &= t\langle w_j, \xi_0 \rangle \\ &\leqslant 0. \end{aligned}$$

Ainsi, $\xi(t) \in \mathcal{D}$ pour tout $t \in [0; r[$.

Nous sommes maintenant proches du but. En effet, supposons que $\langle u, \xi_0 \rangle < 0$. Alors, pour $t \in [0; r[$,

$$\begin{aligned} f(\xi(t)) &= \langle u, \widetilde{x} + t\xi_0 \rangle + b \\ &= \langle u, \widetilde{x} \rangle + b + t\langle u, \xi_0 \rangle \\ &= f(\widetilde{x}) + t\langle u, \xi_0 \rangle \\ &< f(\widetilde{x}), \end{aligned}$$

ce qui contredit le fait que f admet un minimum local en $\widetilde{x}$. Ainsi, on doit avoir $\langle u, \xi_0 \rangle \geqslant 0$. Pour résumer, nous avons montré que pour tout vecteur

$$\xi_0 \in E = \bigcap_{1 \leqslant i \leqslant m} v_i^{\perp},$$

si $\langle w_j, \xi_0 \rangle \leqslant 0$ pour tout $j \in J$, alors $\langle u, \xi_0 \rangle \geqslant 0$. On conclut alors comme dans la démonstration de la Proposition 3.2.3, en appliquant le Lemme 3.2.4 à $P_E(u)$ et à la famille de vecteurs $(P_E(w_j))_{j \in J}$.

► *Suffisance des conditions KKT et globalité du minimum*

Supposons maintenant qu'on dispose d'un point $\widetilde{x}$ vérifiant les conditions de l'énoncé et supposons de plus que toutes les contraintes sont actives (ce qui ne fait pas perdre de généralité d'après le Lemme 3.2.2). Les deuxième et troisième équations garantissent qu'il satisfait les contraintes. De plus, on a pour tout $x \in \mathcal{D}$,

$$\begin{aligned} f(x) - f(\widetilde{x}) &= \langle u, x - \widetilde{x} \rangle \\ &= \sum_{i=1}^{m} \lambda_i \langle v_i, x - \widetilde{x} \rangle + \sum_{j=1}^{p} \mu_j \langle w_j, x - \widetilde{x} \rangle \\ &= \sum_{i=1}^{m} \lambda_i \langle v_i, x \rangle - \sum_{i=1}^{m} \lambda_i \langle v_i, \widetilde{x} \rangle + \sum_{j=1}^{p} \mu_j \langle w_j, x - \widetilde{x} \rangle \\ &= \sum_{i=1}^{m} \lambda_i c_i - \sum_{i=1}^{m} \lambda_i c_i + \sum_{j=1}^{p} \mu_j \langle w_j, x - \widetilde{x} \rangle \\ &= \sum_{j=1}^{p} \mu_j \langle w_j, x - \widetilde{x} \rangle \\ &= \sum_{j=1}^{p} \mu_j \langle w_j, x \rangle - \sum_{j=1}^{p} \mu_j \langle w_j, \widetilde{x} \rangle \\ &= \sum_{j=1}^{p} \mu_j \langle w_j, x \rangle - \mu_j d_j \\ &\geqslant 0, \end{aligned}$$

donc f a un minimum *global* sous contraintes au point $\widetilde{x}$. ■

3.3.2 Vers les sommets

Le Théorème 3.3.1 ramène le problème d'optimisation à un problème matriciel qui reste à résoudre, ce qui n'est pas nécessairement facile. Ceci sort du cadre de ce texte,

mais il existe des méthodes pour aborder ce genre de problème. Notons cependant qu'il était plus ou moins intuitif dès le début qu'il s'agissait d'un problème d'algèbre linéaire. Plus précisément, les contraintes d'inégalité définissent des demi-plans et on cherche donc à optimiser f sur une intersection de demi-espaces. S'il y en a suffisamment, cette intersection est une partie bornée (et donc compacte) de l'espace appelée *polytope*. Dans ce cas, la méthode esquissée au Paragraphe 1.1.2 fonctionne :

- La compacité assure l'existence d'un extremum ;
- La fonction f étant linéaire, son gradient et constant et non nul, donc il n'y a pas d'extremum à l'intérieur ;
- On est donc ramené à optimiser f sur chaque *face*.

La notion de face se définit assez simplement : il suffit de relâcher une des contraintes.

Définition 3.3.2. Soit P un polytope défini par des formes linéaires $h_1, \cdots, h_p$, c'est-à-dire

$$P = \{x \in \mathbf{R}^n \mid h_j(x) \leqslant 0 \text{ pour tout } 1 \leqslant j \leqslant p\}.$$

Pour $1 \leqslant j \leqslant p$, la j-ième face est définie comme

$$F_j = \{x \in P \mid h_j(x) = 0\}.$$

Le troisième point ci-dessus revient à affirmer que le *bord* de P, qui est l'ensemble de tous les points de P qui ne sont pas dans l'intérieur de P, est la réunion des faces. Puisque ce n'est pas très difficile, nous allons le démontrer.

Proposition 3.3.3. *Soit P un polytope définit par des formes linéaires $h_1, \cdots, h_p$ et soit ∂P son bord. Alors,*

$$\partial P = \bigcup_{j=1}^{p} F_j.$$

Démonstration. Si on note F l'union des faces, on peut écrire $P = Q \cup F$, où

$$Q = \{x \in \mathbf{R}^n \mid h_j(x) < 0 \text{ pour tout } 1 \leqslant j \leqslant p\}.$$

Il suffit donc de montrer que Q est l'intérieur de P. Nous allons commencer par montrer que Q est ouvert. En effet, si $x \in Q$, posons

$$\epsilon = \min_{1 \leqslant j \leqslant p} -\frac{h_j(x)}{2} > 0.$$

Alors, par continuité de h_j, il existe $\delta_j > 0$ tel que pour tout $y \in B(x, \delta_j)$,

$$h_j(y) \leqslant h_j(x) + \epsilon = \frac{h_j(x)}{2} < 0.$$

En posant $\delta = \min_j \delta_j$, on a $B(x, \delta) \subset Q$ et Q est bien un ouvert.

Si maintenant $x \in F$ et si j est tel que $h_j(x) = 0$, on a par définition $\langle w_j, x \rangle - d_j = 0$. Mais alors, pour tout $\delta > 0$, on a

$$y = x + \delta \frac{w_j}{\|w_j\|^2} \in B(x, \delta)$$

et

$$\begin{aligned}\langle w_j, y\rangle - d_j &= \langle w_j, x\rangle - d_j + \delta \\ &> 0.\end{aligned}$$

Autrement dit, $B(x,\delta)$ n'est jamais contenue dans P, ce qui signifie que x n'est pas intérieur à P. ■

Expliquons maintenant comment ces notions élémentaires permettent de développer une méthode efficace de recherche d'extremum. Il suffit de remarquer qu'une face d'un polytope est à nouveau un polytope. De plus, on peut voir F_j comme un polytope de l'espace $\ker(h_j)$, qui est de dimension $n-1$. Ainsi, en itérant le raisonnement présenté un peu plus haut, on voit que tout extremum doit être atteint sur une face d'un polytope de dimension 0 ... ce qu'on appelle un *sommet du polytope*. Résumons ceci :

Théorème 3.3.4 (Solution-sommet) Étant donné un problème d'optimisation linéaire, s'il existe une extremum alors il est atteint sur un sommet.

Démonstration. On procède par récurrence avec l'hypothèse de récurrence suivante

H_N : « Si un problème d'optimisation linéaire en dimension n admet une solution, alors il admet une solution qui est un sommet. »

Pour l'initialisation, nous allons démontrer H_1. Nous avons donc une fonction affine $f : x \mapsto ax+b$, avec $a \neq 0$, à optimiser sur un ensemble P défini par des contraintes. S'il y a des contraintes d'égalité, alors P est soit vide soit réduit à un point, auquel cas il n'y a rien à faire. Sinon, les contraintes d'inégalités sont de la forme $x \geqslant \alpha$ ou $x \leqslant \beta$, donc on cherche à optimiser f sur un intervalle. Comme f y est strictement monotone, si elle a un extremum il est atteint à l'une des extrémités de l'intervalle, ce qui prouve H_1.

Supposons maintenant H_n vraie et considérons un problème d'optimisation linéaire en dimension $n+1$. Supposons qu'il existe un extremum, qui est alors nécessairement global. Comme $\nabla f = u \neq 0$, cet extremum ne peut être atteint sur l'intérieur de P. Par conséquent, il est atteint sur ∂P. D'après la Proposition 3.3.3, le bord est la réunion des faces, donc l'extremum est atteint sur l'une des faces, disons F_j. On considère alors le problème d'optimisation obtenu en restreignant toutes les fonctions au sous-espace affine d'équation $h_j = 0$. Il s'agit d'un problème d'optimisation linéaire en dimension n, et si le minimum de f est atteint sur F_j, alors il est égal au minimum de cette restriction (et de même pour un maximum). Par H_n, l'extremum est donc atteint sur un sommet de F_j. On conclut en remarquant que les sommets de F_j sont des sommets de P. ■

Ceci ramène le problème à un ensemble fini, mais il n'est pas stratégique d'essayer de calculer la valeur de f sur tous les sommets du polytope, parce que leur nombre a tendance à croître exponentiellement avec les dimensions n et m du problème. Il faut être un peu plus malin et essayer de s'arranger pour ne pas avoir besoin de regarder tous les sommets. La méthode la plus utilisée pour cela est la Méthode du simplexe, dont nous allons donner l'idée. Le point de départ est de se restreindre non pas uniquement aux sommets, mais aux sommets et aux arêtes, ce qui nécessite d'abord de définir ce terme.

Définition 3.3.5. Soit P un polytope. Une 1-*face* de P est tout simplement une face. De plus, pour tout $1 \leqslant j \leqslant n-1$ une $j+1$-face de P est une face d'une j-face de P. Les n-faces de P sont ses sommets, et les $(n-1)$-faces de P sont appelées les *arêtes* de P.

Nous dirons désormais que deux sommets sont *adjacents* s'ils sont les faces d'une arête. Partons d'un sommet S quelconque du polytope P et supposons qu'il existe un sommet S' adjacent à S tel que $f(S') \leqslant f(S)$. Si on considère la restriction de f à l'arête reliant S à S', on a une fonction affine sur un segment dont la valeur à l'extrêmité droite du segment est inférieure à sa valeur à l'extrêmité gauche. Autrement dit, la fonction doit être décroissante. Ainsi, il suffit de chercher une direction, parmis les arêtes dont S est un sommet, selon laquelle f décroît, et de regarder le sommet suivant dans cette direction. D'accord, mais que se passe-t-il si f croît le long de toutes les arêtes ? C'est ici que la linéarité s'avère utile : S sera alors automatiquement un minimum local, donc global, ce qui signifie que l'algorithme peut s'arrêter.

Proposition 3.3.6. *Soit S un sommet tel que pour tout sommet S' adjacent à S, $f(S') \geqslant f(S)$. Alors, f a un minimum local sous contraintes au point S.*

Comme nous n'utiliserons pas ce résultat dans la suite, et que la démonstration nécessiterait d'anticiper sur quelques éléments de géométrie convexe, nous l'omettrons.

Nous pouvons maintenant décrire schématiquement un algorithme :

1. On part d'un sommet S quelconque de P ;
2. Étant donné le somme S actuel, on cherche une arête partant de S le long de laquelle f décroît ;
   - S'il n'y en a pas, l'algorithme s'arrête et on renvoie S.
   - S'il y en a une, on change S en S', le sommet adjacent à S le long de cette arête.
3. On réitère.

Ce que l'on a gagné, c'est que l'on ne teste pas tous les sommets puisqu'on suit à chaque fois une direction décroissante. Autrement dit, quand on arrive à un sommet, on élimine immédiatement tous ses voisins qui ont une valeur inférieure. Néanmoins, cela peut être encore assez inefficace. Pour améliorer l'algorithme, il vaudrait mieux suivra la direction dans laquelle f diminue le plus vite. Le problème c'est que cette direction ne correspond pas forcément à une arête. Mais au fond, cela n'est pas très grave : on peut suivre la direction "à travers" le polytope jusqu'à ce qu'on ressorte par un autre côté. On parle alors de *méthode de points intérieurs*. La plus célèbre d'entre elles, proposée par Dantzig [18], est appelée *méthode du simplexe* et s'avère à ce jour l'une des plus efficaces.

Pour conclure, revenons simplement à l'exemple du Paragraphe 1.1.2 du Chapitre 1. Le polytope défini par les contraintes est ici un polygone qui possède cinq sommets que nous noterons $S_1, \cdots, S_5$ en les numérotant en sens direct [19] depuis $S_1 = (0,7)$. On

18. Georges Bernard Dantzig (1914−−2005) : mathématicien et informaticien américain, il est notamment célèbre pour son invention de l'algorithme du simplexe, et plus généralement pour ses travaux en optimisation (linéaire, non-linéaire et stochastique) ainsi qu'en analyse numérique matricielle. La *Mathematical Optimization Society* décerne tous les trois ans un prix en son honneur.

19. Rappelons que le sens trigonométrique direct est le sens inverse des aiguilles d'une montre.

calcule les valeurs correspondantes de F :

$$\begin{aligned} F(S_1) &= 76 \\ F(S_2) &= 66 \\ F(S_3) &= 62 \\ F(S_4) &= 77 \\ F(S_5) &= 80 \end{aligned}$$

On trouve que le minimum est atteint au point $S_3 = (4,1)$, ce qui donne

$$(x_1, x_2, x_3, x_4) = (4, 1, 0, 6)$$

et $f_{\min} = 62$, comme affirmé.

Concluons en ajoutant que le calcul différentiel nous apprend comment approcher une fonction donnée par une fonction linéaire (sa différentielle). On peut donc imaginer, dans un problème général d'optimisation sous contraintes, d'approcher à la fois f et ses contraintes par des fonctions linéaires afin de se ramener à un problème linéaire. Il s'agit de la technique de ... *linéarisation* !

3.4 Dualité

Le pathétique de l'amour consiste dans une dualité insurmontable des êtres.

E. Levinas, Le temps et l'autre

Avant de poursuivre notre étude, faisons une pause pour reformuler ce que nous avons démontré. Nous avons vu que pour étudier les points critiques sous contraintes, il fallait faire intervenir des scalaire $\lambda_1, \cdots, \lambda_m$ et $\mu_1, \cdots, \mu_p$, appelés *multiplicateurs de Lagrange*. On peut synthétiser cela à l'aide d'une seule fonction, ce qui permet d'exprimer les résultats de façon plus compacte. Nous appellerons dans cette partie *problème d'optimisation sous contraintes* la donnée d'une fonction $f : U \to \mathbf{R}$ qu'on cherche à minimiser sous des contraintes d'égalité données par des fonctions $g_1, \cdots, g_m : U \to \mathbf{R}$ et des contraintes d'inégalité données par des fonctions $h_1, \cdots, h_p : U \to \mathbf{R}$.

3.4.1 Le Lagrangien

Définition 3.4.1. Le *Lagrangien* d'un problème d'optimisation sous contraintes est la fonction $\mathcal{L} : U \times \mathbf{R}^m \times \mathbf{R}^p_- \to \mathbf{R}$ définie par

$$\begin{aligned} \mathcal{L}(x_1, \cdots, x_n, \lambda_1, \cdots, \lambda_m, \mu_1, \cdots, \mu_p) = f(x_1, \cdots, x_n) \\ - \sum_{i=1}^{m} \lambda_i g(x_1, \cdots, x_n) \\ - \sum_{j=1}^{p} \mu_j h_j(x_1, \cdots, x_n). \end{aligned}$$

On peut alors reformuler le Théorème 3.1.9 de la façon suivante : si f admet un minimum local sous contraintes au point $\widetilde{x}$, alors il existe $(\lambda_1, \cdots, \lambda_m)$ et $(\mu_1, \cdots, \mu_p)$ tels que $\mathcal{L}$ a un point critique en

$$(\widetilde{x_1}, \cdots, \widetilde{x_n}, \lambda_1, \cdots, \lambda_m, \mu_1, \cdots, \mu_p).$$

On remarquera qu'être un point critique pour $\mathcal{L}$ correspond à $n+m+p$ équations données par chaque dérivée partielle, ce qui semble surdéterminer le problème. Il n'en est en fait rien, car

$$\frac{\partial \mathcal{L}}{\partial \lambda_i} = g_i$$
$$\frac{\partial \mathcal{L}}{\partial \mu_j} = h_j$$

Ainsi, l'annulation des dérivées partielles par rapport aux coordonnées λ_i et μ_j traduit simplement les contraintes du problème (n'oublions pas que $h_j(\widetilde{x}) = 0$ si $\mu_j \neq 0$).

La formulation à l'aide du Lagrangien présente un avantage : elle met en lumière un nouveau problème d'optimisation lié au précédent. Pour le voir remarquons que pour tout $x \in U$,

$$\sup_{\lambda \in \mathbf{R}^m, \mu \in \mathbf{R}^p_-} \mathcal{L}(x, \lambda, \mu) = f(x) - \inf_{\lambda \in \mathbf{R}^m, \mu \in \mathbf{R}^p_-} \left(\sum_{i=1}^{m} \lambda_i g(x) + \sum_{j=1}^{p} \mu_j h_j(x) \right).$$

La borne inférieure vaut souvent $-\infty$. Cependant, si x satisfait les contraintes, alors $g_i(x) = 0$ et $h_j(x) \leqslant 0$, donc comme $\mu_j \leqslant 0$, la borne inférieur vaut tout simplement 0. Autrement dit, pour tout $x \in U$ et en notant comme précédemment $\mathcal{D}$ l'ensemble des points vérifiant les contraintes, on a

$$\sup_{\lambda \in \mathbf{R}^m, \mu \in \mathbf{R}^p_-} \mathcal{L}(x, \lambda, \mu) = \begin{cases} f(x) & \text{si} \quad x \in \mathcal{D} \\ +\infty & \text{si} \quad x \notin \mathcal{D} \end{cases}$$

Ainsi, notre problème d'optimisation consiste à trouver

$$\inf_{x \in U} \sup_{\lambda \in \mathbf{R}^m, \mu \in \mathbf{R}^p_-} \mathcal{L}(x, \lambda, \mu).$$

Les contraintes se trouvent incorporées dans le fait que la fonction à minimiser vaut de toute façon $+\infty$ dès qu'elles ne sont pas vérifiées.

Il est alors naturel de s'intéresser au problème d'optimisation consistant à chercher

$$\sup_{\lambda \in \mathbf{R}^m, \mu \in \mathbf{R}^p_-} \inf_{x \in U} \mathcal{L}(x, \lambda, \mu).$$

Pour cela on définit une fonction

$$f^*(\lambda, \mu) = \inf_{x \in U} \mathcal{L}(x, \lambda, \mu)$$

qu'on va chercher à maximiser.

Définition 3.4.2. Le problème ci-dessus est appelé *problème dual* du problème d'origine. Réciproquement, le problème d'origine est appelé *problème primal*.

Nous n'avons a priori pas fait grand chose. En fait, l'espoir dans la définition du problème dual est que sa résolution puisse nous donner des informations sur le problème primal. Il s'avère – et c'est tout à fait remarquable – que c'est le cas ! La manifestation la plus simple de ce phénomène est la suivante :

Proposition 3.4.3 (Dualité de Lagrange faible). *Pour tous x, λ et μ satisfaisant les contraintes, on a*

$$f(x) \geqslant f^*(\lambda, \mu).$$

En particulier, si f admet un minimum local sous contraintes au point $\widetilde{x}$, et si f^* admet un maximum local sous contraintes en $(\widetilde{\lambda}, \widetilde{\mu})$, alors

$$\boxed{f(\widetilde{x}) \geqslant f^*(\widetilde{\lambda}, \widetilde{\mu})}$$

De plus, s'il y a égalité alors les deux extrema sont globaux.

Démonstration. Comme x satisfait les contraintes, on a $g_i(x) = 0$ pour tout $1 \leqslant i \leqslant m$, $h_j(x) \leqslant 0$ pour tout $1 \leqslant j \leqslant p$. Comme de plus $\mu_j \leqslant 0$ on a

$$\begin{aligned} f(x) &= \mathcal{L}(x, \lambda, \mu) + \sum_{i=1}^{m} \lambda_i g_i(x) + \sum_{j=1}^{p} \mu_j h_j(x) \\ &= \mathcal{L}(x, \lambda, \mu) + \sum_{j=1}^{p} \mu_j h_j(x) \\ &\geqslant \mathcal{L}(x, \lambda, \mu) \\ &\geqslant f^*(\lambda, \mu). \end{aligned}$$

La seconde partie de l'énoncé suit immédiatement.

Quant à la troisième, s'il existait $\widetilde{x}'$ tel que $f(\widetilde{x}') < f(\widetilde{x})$, alors on aurait

$$\begin{aligned} f^*(\widetilde{\lambda}, \widetilde{\mu}) &\leqslant f(\widetilde{x}') \\ &< f(\widetilde{x}) \\ &= f^*(\widetilde{\lambda}, \widetilde{\mu}), \end{aligned}$$

une contradiction. ∎

La différence

$$S = f^*(\widetilde{\lambda}, \widetilde{\mu}) - f(\widetilde{x})$$

mesure en un sens la différence entre les deux problèmes d'optimisation. La plus petite valeur possible de S est appelé *saut de dualité* du problème. La Proposition 3.4.3 affirme que $S \geqslant 0$. Si $S > 0$, le problème dual "perd" de l'information par rapport au problème primal. Par contre, si $S = 0$, cela signifie que résoudre le problème dual suffit à résoudre le problème primal. On dit dans ce cas que la problème satisfait la *dualité de Lagrange forte*.

3.4.2 Dualité linéaire

Tout ceci est bien beau, mais pour être utile il faut que le problème dual soit plus facile à résoudre que le problème primal. Et c'est bien le cas ! En effet, le problème dual est toujours un problème de *maximisation concave* ou de façon équivalente – en changeant tous les signes – de *minimisation convexe*, et il existe pour ce type de problème des méthodes très efficaces (voir le Paragraphe 4.2.3). Ceci fera l'objet du prochain chapitre, mais en attendant, nous allons regarder d'un peu plus près le cas où le problème d'origine est linéaire. Considérons donc une fonction linéaire $f : x \mapsto \langle u, x \rangle + b$

que l'on cherche à minimiser sous les contraintes $Gx = c$ et $Hx \leqslant d$, où $G \in \mathrm{M}_{mn}(\mathbf{R})$ et $H \in \mathrm{M}_{pn}(\mathbf{R})$ et l'inégalité est comprise coordonnée par coordonnée. Le Lagrangien du problème est

$$\begin{aligned}\mathcal{L}(x,\lambda,\mu) &= \langle u, x\rangle + b - \langle \lambda, Gx - c\rangle - \langle \mu, Hx - d\rangle \\ &= \langle \lambda, c\rangle + \langle \mu, d\rangle + \langle u - G^t\lambda - H^t\mu, x\rangle + b\end{aligned}$$

Il faut maintenant minimiser cette quantité sur x. Or, si $u - G^t\lambda - H^t\mu \neq 0$, alors le minimum vaut $-\infty$. Ainsi, la fonction duale ne sera bien définie que si $u = G^t\lambda + H^t\mu$. Dans ce cas, le terme en x vaut 0. On conclut que le problème dual consiste à maximiser la fonction $f^* : (\lambda, \mu) \mapsto \langle \lambda, c\rangle + \langle \mu, d\rangle + b$ sous les contraintes

$$\begin{cases} u &= G^t\lambda + H^t\mu \\ \mu &\leqslant 0 \end{cases}$$

Autrement dit : le dual d'un problème linéaire est encore un problème linéaire ! Mais alors, il est légitime de s'interroger sur son rapport au problème initial. Et là encore, les problèmes linéaires se comportent de façon remarquable !

Proposition 3.4.4 (Bidualité pour les problèmes linéaires)**.** *Le problème dual du problème dual d'un problème linéaire est le problème primal.*

Démonstration. Nous allons nous ramener à un problème de minimisation en gardant les contraintes mais en considérant la fonction $-f^*$. Le nouveau Lagrangien s'écrit (avec une barre pour le distinguer du premier et en notant α et β les multiplicateurs)

$$\begin{aligned}\overline{\mathcal{L}}(\lambda,\mu,\alpha,\beta) &= -\langle \lambda, c\rangle - \langle \mu, d\rangle - b - \langle \alpha, G^t\lambda + H^t\mu - u\rangle - \langle \beta, \mu\rangle \\ &= \langle u, \alpha\rangle - b - \langle \lambda, c + G\alpha\rangle - \langle \mu, d + \beta + H\alpha\rangle.\end{aligned}$$

Pour que cette quantité ait une borne inférieure différente de $-\infty$, il faut tout d'abord que $c + G\alpha = 0$. De plus, comme $\mu \leqslant 0$, il faut que $d + \beta + H\alpha \leqslant 0$ également, sinon le produit scalaire ne sera pas borné inférieurement. Cette dernière inégalité peut s'écrire

$$H\alpha \leqslant d + \beta$$

On conclut donc que le problème dual revient à maximiser $f^{**} : (\alpha, \beta) \mapsto \langle u, \alpha\rangle - b$ sous les contraintes

$$\begin{cases} G\alpha &= -c \\ H\alpha &\geqslant -d - \beta \\ \beta &\leqslant 0 \end{cases}$$

Les deux dernière inégalités donnent $H\alpha \geqslant -d$. Réciproquement, si α réalise le minimum de $\alpha \mapsto \langle u, \alpha\rangle - b$ sous les contraintes $G\alpha = -c$ et $H\alpha \leqslant -d$, alors en prenant $\beta = 0$ on a une solution du système. Pour comparer ce problème au problème primal, nous allons une fois de plus prendre l'opposé de la fonction f^{**} afin d'avoir un problème de minimisation. Il s'agit donc de minimiser $\alpha \mapsto -\langle u, \alpha\rangle + b$, et en posant $x = -\alpha$ on trouve donc que $-f^{**}(\alpha) = f(x)$ et que les contraintes sont

$$\begin{cases} Gx &= c \\ Hx &\leqslant d \end{cases}$$

Ainsi, nous avons retrouvé le problème primal, comme annoncé. ∎

Pour conclure, nous allons utiliser la caractérisation des solutions d'un problème linéaire donnée dans le Théorème 3.3.1 pour améliorer la Proposition 3.4.3.

Théorème 3.4.5 (Dualité de Lagrange forte pour les problèmes linéaires) Le saut de dualité d'un problème linéaire est nul. Autrement dit, si le problème primal a une solution, alors il existe $\widetilde{x}, \widetilde{\lambda}, \widetilde{\mu}$ tels que

$$f(\widetilde{x}) = f^*(\widetilde{\lambda}, \widetilde{\mu}).$$

De plus, on peut prendre pour $\widetilde{\lambda}, \widetilde{\mu}$ les multiplicateurs associés à $\widetilde{x}$.

Démonstration. Soit $\widetilde{x} \in \mathcal{D}$ un point auquel f admet un minimum (nécessairement global). Alors, on sait par le Théorème 3.3.1 qu'il existe $\widetilde{\lambda}$ et $\widetilde{\mu}$ tels que

$$u = G^t\widetilde{\lambda} + H^t\widetilde{\mu}$$

et $\langle \widetilde{\mu}, H\widetilde{x} \rangle = \langle \widetilde{\mu}, d \rangle$ par complémentarité. On calcule alors

$$\begin{aligned} f(\widetilde{x}) &= \left\langle G^t\widetilde{\lambda} + H^t\widetilde{\mu}, \widetilde{x} \right\rangle + b \\ &= \langle \widetilde{\lambda}, G\widetilde{x} \rangle + \langle \widetilde{\mu}, H\widetilde{x} \rangle + b \\ &= \langle \widetilde{\lambda}, c \rangle + \langle \widetilde{\mu}, H\widetilde{x} \rangle + b \\ &= \langle \widetilde{\lambda}, c \rangle + \langle \widetilde{\mu}, d \rangle + b \\ &= f^*(\widetilde{\lambda}, \widetilde{\mu}) + b, \end{aligned}$$

où pour passer de la deuxième à la troisième ligne on a utilisé le fait que $\widetilde{x}$ vérifie les contraintes d'égalité. On conclut alors par la dernière partie de la Proposition 3.4.3. ∎

3.4.3 Un intermède ludique

La dualité de Lagrange forte pour les problèmes linéaires a plusieurs applications intéressantes. Nous allons en donner une remarquable, qui est un des résultats fondamentaux de la théorie des jeux. Pour cela, il nous faut d'abord introduire un peu de contexte.

Jeux à deux joueurs

Nous considérons ici le cas d'un jeu à deux joueurs, que nous nommerons pour plus de commodité Alice (abrégé en A) et Bob (abrégé en B). Le jeu sera dit *fini*, dans la mesure ou chaque joueur n'a qu'un nombre fini d'actions (qu'on appelle *stratégies*) possibles. C'est le cas par exemple des jeux de société, dont l'un des exemples les plus simples est le jeu de *Pierre-Feuille-Ciseaux*. Si on note n le nombre d'actions possibles de A et m le nombre d'actions possibles de B, alors le résultat du jeu peut être donné par une matrice de taille $n \times m$. En effet, si on note M cette matrice, alors M_{ij} sera le gain de A si A joue la stratégie i et B joue la stratégie j. Et qu'en est-il du gain de B ? Là encore nous allons faire une hypothèse simplificatrice : le jeu sera *à somme nulle*. Cela signifie que le gain de B est toujours exactement l'opposé du gain de A. La matrice des gains de B est donc $-M$.

Tout ceci étant posé, le problème fondamental de la théorie des jeux consiste à savoir si l'un au moins des deux joueurs peut posséder une stratégie qu'on pourrait qualifier d'optimale dans un certain sens. Par exemple, dans le cas de *Pierre-Feuille-Ciseaux*, si A choisit de jouer *Pierre* à chaque coup, alors B n'a qu'à décidé de jouer *Feuille* pour être sûr de gagner. Dans ce cas, A va être forcé de changer de stratégie

et de jouer *Ciseaux*, mais B réagira à son tour en décidant de jouer *Pierre* ... Bref, les choses vont tourner en rond et chaque joueur perdra et gagnera alternativement. Une façon d'éviter ce comportement cyclique est d'introduire un peu d'aléa dans la partie : chaque joueur peut choisir "au hasard" quel signe il utilisera, par exemple en lançant un dé. En particulier, si les deux joueurs choisissent leur symbole uniformément, le gain moyen de A sera de 0, et de même pour B.

Pour formaliser un peu tout ça, nous allons fixer une fois pour toute les stratégies possibles. Il y en aura n pour A et m pour B. Un choix aléatoire de stratégie pour A est alors donné par des nombres $x_1, \cdots, x_n \geqslant 0$ tels que

$$\sum_{i=1}^{n} x_i = 1.$$

De même, une stratégie aléatoire pour B est donnée par des nombres $y_1, \cdots, y_m$ tels que

$$\sum_{j=1}^{m} y_j = 1.$$

Qui dit aléatoire dit moyenne (ou espérance), et c'est donc ces quantités qu'il faut considérer pour choisir une bonne stratégie. Imaginons maintenant que deux tels stratégies aléatoires $x = (x_1, \cdots, x_n)$ et $y = (y_1, \cdots, y_n)$ soient choisies. Alors, le gain moyen de A sera donné par

$$\sum_{i=1}^{n} \sum_{j=1}^{m} x_i y_j M_{ij} = \langle x, My \rangle.$$

Le théorème de von Neumann

Nous arrivons maintenant au concept central de la théorie des jeux : une bonne stratégie pour A sera une stratégie qui maximise son *plancher de gain moyen*. Qu'est-ce à dire ? Si A joue la stratégie x, son plus petit gain sera en moyenne

$$G(x) = \min_y \langle x, My \rangle,$$

où le minimum porte sur toutes les stratégies aléatoires de B. Ainsi, la meilleur stratégie pour A sera (si elle existe) la stratégie $\widetilde{x}$ qui vérifie

$$G(\widetilde{x}) = \max_x \min_y \langle x, My \rangle.$$

On parle parfois – pour des raisons évidentes – de *stratégie maximin*.

Et B dans tout ça ? Va-t-il chercher la même chose ? Oui, sauf qu'il faut se rappeler que le gain de B est l'opposé du gain de A puisque le jeu est à somme nulle. Ainsi, son plus petit gain quant il joue la stratégie y est

$$G(y) = \min_x -\langle x, My \rangle = -\max_x \langle x, My \rangle.$$

Par conséquent, la meilleurs stratégie pour B sera (si elle existe) la stratégie $\widetilde{y}$ qui vérifie

$$G(\widetilde{y}) = \min_y \max_x \langle x, My \rangle.$$

On l'appelle – oh surprise ! – la *stratégie minimax*.

Le résultat que nous allons maintenant aborder, est que les stratégies minimax et maximin existent, et qu'en plus elles donnent les mêmes espérances de gains au deux joueurs ! L'idée de la démonstration est de se ramener à deux problèmes d'optimisation linéaires en dualité. Pour ce faire, nous allons utiliser un petit résultat élémentaire que nous donnons maintenant.

Lemme 3.4.6. *Soit $x \in \mathbf{R}^n$ et $M \in \mathrm{M}_n(\mathbf{R})$. Alors, si $\mathcal{B}$ désigne l'ensemble des vecteurs $y \in \mathbf{R}_+^n$ tels que $\sum y_i = 1$, on a*

$$\min_{y \in \mathcal{B}} \langle x, My \rangle = \min_{1 \leqslant i \leqslant n} (M^t x)_i$$

Démonstration. Tout d'abord, en prenant $y = e_i$ – le i-ème vecteur de la base canonique – on a pour tout $1 \leqslant i \leqslant n$.

$$\begin{aligned} \langle x, Me_i \rangle &= \langle M^t x, e_i \rangle \\ &= (M^t x)_i. \end{aligned}$$

Ceci montre que le membre de droite dans l'égalité de l'énoncé est supérieur au membre de gauche, et il reste à montrer l'inégalité inverse. Pour ce faire, écrivons

$$y = \sum_{i=1}^{n} y_i e_i,$$

de sorte que

$$\begin{aligned} \langle x, My \rangle &= \langle M^t x, y \rangle \\ &= \sum_{i=1}^{n} \langle M^t x, y_i e_i \rangle \\ &= \sum_{i=1}^{n} y_i \langle M^t x, e_i \rangle \\ &= \sum_{i=1}^{n} y_i (M^t x)_i \end{aligned}$$

Comme $y_i \geqslant 0$ pour tout $1 \leqslant i \leqslant n$, on peut alors minorer la somme :

$$\begin{aligned} \sum_{i=1}^{n} y_i (M^t x)_i &\geqslant \sum_{i=1}^{n} y_i \min_{1 \leqslant i \leqslant n} (M^t x)_i \\ &= \left(\min_{1 \leqslant i \leqslant n} (M^t x)_i \right) \sum_{i=1}^{n} y_i \\ &= \min_{1 \leqslant i \leqslant n} (M^t x)_i \end{aligned}$$

Nous avons ainsi obtenu l'inégalité inverse, ce qui conclut la preuve. ∎

Nous pouvons maintenant démontrer ce qu'on appelle le Théorème du minimax, originellement démontré par von Neumann [20] et Morgenstern [21] :

20. John von Neumann (1903 – 1957) : mathématicien et physicien américain d'origine hongroise qui a été l'un des plus brillants du XX^e siècle. Il a contribué de façon fondamentale à de nombreux domaines des mathématiques et notamment l'analyse fonctionnelle, la logique et la théorie des jeux. Il a également été l'un des acteurs du développement de la mécanique quantique et l'un des développeurs de la bombe atomique. Il fut aussi l'un des pionniers de l'informatique.

21. Oskar Morgenstern (1902 – 1977) : mathématicien et économiste allemand puis américain. Outre

Théorème 3.4.7 Étant donné un jeu à deux joueurs à somme nulle, on a

$$\boxed{\max_x \min_y \langle x, My\rangle = \min_y \max_x \langle x, My\rangle}$$

Démonstration. La démonstration se fait en plusieurs étapes, que nous allons distinguer.

▶ *Équivalence avec un problème d'optimisation linéaire*

Considérons le problème suivant : on cherche à minimiser la fonction $f : \mathbf{R}^n \to \mathbf{R}$ définie par

$$x \mapsto \sum_{i=1}^{n} x_i$$

sous les contraintes

$$\begin{cases} M^t x & \geqslant & \mathbf{1} \\ x_i & \geqslant & 0 \end{cases},$$

où $\mathbf{1} \in \mathbf{R}^n$ désigne le vecteur dont tous les coefficients sont égaux à 1.

Quel rapport avec ce qui nous intéresse ? Regardons-y de plus près en considérant un réel

$$0 < t \leqslant \max_x \min_y \langle x, My\rangle.$$

Alors, il existe par définition $x \in \mathbf{R}^n_+$ vérifiant $f(x) = 1$ et

$$\min_y \langle x, My\rangle \geqslant t.$$

D'après le Lemme 3.4.6, cette inégalité peut s'écrire

$$M^t x \geqslant t\mathbf{1},$$

ce qui en posant $x' = x/t$ nous donne $x' \in \mathbf{R}^n_+$ vérifiant $f(x') = 1/t$ et $M^t x' \geqslant \mathbf{1}$. Ainsi, par définition, on a – en notant $\mathcal{D}$ l'ensemble des points vérifiant les contraintes –

$$\frac{1}{t} \geqslant \min_{\mathcal{D}} f.$$

Chacun des arguments précédents est réversible, ce qui permet de déduire qu'on a bien

$$\frac{1}{\min_{\mathcal{D}} f} = \max_x \min_y \langle x, My\rangle.$$

▶ *Problème dual*

Nous avons ramené le calcul du maximin à un problème d'optimisation linéaire, et nous allons maintenant faire de même avec le minimax. Cela n'est pas très difficile

ses travaux appliquant la théorie des jeux à l'économie, il s'est aussi intéressé aux problèmes des mesures en statistiques économiques, ainsi qu'à la stabilité des marchés financiers.

en s'inspirant des calculs précédents. Considérons donc la fonction $g : \mathbf{R}^m \to \mathbf{R}$ définie par

$$g(y) = \sum_{i=1}^{m} y_i,$$

et intéressons-nous cette fois au problème de *maximisation* de g sous les contraintes

$$\begin{cases} My & \leqslant & \mathbf{1} \\ y_i & \geqslant & 0 \end{cases}.$$

Le même calcul que précédemment montre alors l'égalité

$$\frac{1}{\max\limits_{\mathcal{D}'} g} = \min_y \max_x \langle x, My \rangle.$$

► *Conclusion*

Nous allons maintenant montrer que les deux problèmes d'optimisation linéaire que nous avons construits sont duaux l'un de l'autre. Considérons donc le premier problème et intéressons-nous à son Lagrangien. Comme il y a deux contraintes d'inégalités, nous noterons $\mu \in \mathbf{R}^m_-$ et $\mu' \in \mathbf{R}^n_-$ respectivement les multiplicateurs associés à chacune d'elle. En remarquant que $f(x) = \langle \mathbf{1}, x \rangle$, on voit que

$$\begin{aligned} \mathcal{L}(x, \mu, \mu') &= f(x) - \langle \mu, \mathbf{1} - M^t x \rangle - \langle \mu', -x \rangle \\ &= \langle \mathbf{1}, x \rangle - \langle \mu, \mathbf{1} - M^t x \rangle - \langle \mu', -x \rangle \\ &= \langle \mathbf{1} + M\mu + \mu', x \rangle - \langle \mu, \mathbf{1} \rangle. \end{aligned}$$

Pour que cette quantité ait une borne inférieure finie par rapport à x, il faut et il suffit que $\mathbf{1} + M\mu + \mu' = 0$. Ainsi, le problème dual revient à maximiser la fonction $\mu \mapsto -\langle \mu, \mathbf{1} \rangle$ sous les contraintes

$$\begin{cases} \mathbf{1} + M\mu + \mu' & = & 0 \\ \mu & \leqslant & 0 \\ \mu' & \leqslant & 0 \end{cases}.$$

La première et la dernière contrainte peuvent être regroupées pour donner $\mathbf{1} + M\mu \geqslant 0$, c'est-à-dire $-M\mu \leqslant \mathbf{1}$. Pour conclure, il suffit alors de poser $y = -\mu$. En effet, on doit alors maximiser la fonction $y \mapsto \langle y, \mathbf{1} \rangle$ sous les contraintes

$$\begin{cases} My & \leqslant & \mathbf{1} \\ y & \geqslant & 0 \end{cases},$$

qui est exactement le second problème d'optimisation linéaire que nous avons étudié.

D'après le Théorème 3.4.5, il existe donc $\widetilde{x}$ et $\widetilde{y}$ tels que

$$g(\widetilde{y}) = \max_{\mathcal{D}'} g = \min_{\mathcal{D}} f = f(\widetilde{x}),$$

et le résultat suit. ■

Une question d'équilibre

On peut-être sceptique quant au fait que les stratégies données par le Théorème 3.4.7 soient vraiment les plus intéressantes. Pour appuyer cette affirmation, nous allons maintenant en donner une autre propriété intéressante, qui découle de la démonstration que nous avons faites.

Corollaire 3.4.8. *Il existe des stratégies mixtes $\widetilde{x}$ et $\widetilde{y}$ pour A et B respectivement telle que pour tout x, y*

$$\langle x, M\widetilde{y}\rangle \leqslant \langle \widetilde{x}, M\widetilde{y}\rangle \leqslant \langle \widetilde{x}, My\rangle.$$

Démonstration. Considérons la fonction $\varphi : x \mapsto \min_{1\leqslant i\leqslant n}(M^t x)_i$. Elle est continue, donc admet un maximum qui est atteint en un point $\widetilde{x}$. De plus, nous avons montré dans le Lemme 3.4.6 que

$$\varphi(\widetilde{x}) = \min_y \langle \widetilde{x}, My\rangle.$$

De même, la fonction $\psi : y \mapsto \max_{1\leqslant i\leqslant n}(My)_i$ est continue et admet donc un minimum qui est atteint en un point $\widetilde{y}$. Le Théorème 3.4.7 peut alors s'écrire

$$\varphi(\widetilde{x}) = \psi(\widetilde{y}).$$

Ainsi, pour tout y on a

$$\begin{aligned}\langle \widetilde{x}, My\rangle &\geqslant \min_y \langle \widetilde{x}, My\rangle \\ &= \varphi(\widetilde{x}) \\ &= \psi(\widetilde{y}) \\ &= \max_x \langle , M\widetilde{y}\rangle \\ &\geqslant \langle \widetilde{x}, M\widetilde{y}\rangle,\end{aligned}$$

ce qui donne le membre de droite de l'inégalité. Le membre de gauche s'obtient par un calcul similaire. ∎

Ce résultat admet une interprétation en termes de jeu. Supposons que A et B aient adopté respectivement les stratégies $\widetilde{x}$ et $\widetilde{y}$. À un moment donné, A envisage de changer de stratégie, en espérant peut-être mieux contrer la stratégie de B. Mais le Corollaire 3.4.8 lui indique que son gain moyen sera alors inférieur. Autrement dit, tant que B ne change rien, A n'a aucun intérêt à changer. Le même raisonnement fonctionne dans l'autre sens, si bien qu'on a une situation qui est en un sens *équilibrée*. C'est cette observation qui à mené Nash [22] à développer une notion d'équilibre aujourd'hui centrale en théorie des jeux :

Définition 3.4.9. Dans un jeu avec un nombre arbitraire de joueurs, un *équilibre de Nash* est un choix de stratégie mixte pour chaque joueur tel que si un seul joueur change de stratégie, alors son gain moyen diminue.

L'avantage de cette notion est qu'elle a un sens pour n'importe que nombre de joueurs, tandis que l'approche du minimax ne fonctionne qu'avec deux joueurs : on prend déjà le minimum sur y et le maximum sur x, que ferait-on sur z ? Elle permet également de se passer de l'hypothèse de somme nulle du jeu, même si cela n'est pas vraiment un problème : tout jeu est équivalent à un jeu à somme nulle avec un joueur de plus (la banque). L'un des résultats les plus célèbres de Nash est l'existence d'un équilibre dans une très grande généralité. Ce résultat nous mènerait malheureusement trop loin de notre sujet, mais on pourra en savoir plus en consultant par exemple [14].

22. John Forber Nash (1928 – 2015) : mathématicien américain qui a obtenu des résultats importants en géométrie différentielle, en analyse des équations aux dérivées partielles et en théorie des jeux. Ses travaux remarquables lui ont valu à la fois le Prix de la Banque de Suède en sciences économiques en mémoire d'Alfred Nobel et le Prix Abel.

CHAPITRE 4

CONVEXITÉ

Les Barbares s'avançaient lentement, pour ne point s'essouffler, en battant la terre avec leurs pieds ; le centre de l'armée punique formait une courbe convexe

G. Flaubert, Salammbô

La notion de convexité est déjà occasionnellement apparue plusieurs fois dans ce texte, par exemple dans le cas de l'optimisation linéaire au Paragraphe 3.3 du Chapitre 3. Il est donc temps de nous pencher plus sérieusement dessus afin de voir ce qu'elle apporte. Dans ce chapitre, nous allons introduire une notion de convexité pour les fonctions de plusieurs variables, qui généralise celle connue dans le cas d'une variable. Comme nous allons le voir, la convexité est intéressante pour plusieurs raisons :

- Elle permet d'améliorer sensiblement les résultats d'optimisation du Chapitre 2.
- Elle permet d'utiliser des méthodes d'approximation plus efficaces.
- Beaucoup de problèmes concrets portent, en fait, sur des fonctions convexes.

Sur le troisième point, on peut par exemple mentionner les problèmes dits de *moindre carrés*, et plus généralement d'*optimisation quadratique*, dont nous parlerons au Paragraphe 4.3.

Avant de nous lancer, indiquons que la convexité est une propriété extrêmement riche, même en l'absence de régularité. Par cela, nous voulons dire que beaucoup de résultats d'optimisation sont valables pour des fonctions convexes qui ne sont pas différentiables, ce qui est parfois utile en pratique dans des situations où il ne serait pas naturel de supposer certaines données lisses. Toutefois, quand les fonctions considérées sont de classe $\mathcal{C}^1$, alors le travail que nous avons déjà accompli permet relativement facilement d'obtenir des théorèmes puissants pour l'optimisation des fonctions convexes.

4.1 La descente

Plonger au fond du gouffre, Enfer ou Ciel, qu'importe ? Au fond de l'inconnu pour trouver du nouveau

C. Baudelaire, *Voyage* in Les fleurs du mal.

Nous avons vu à la fin du Chapitre 2 une méthode numérique pour chercher les extrema d'une fonction : la Méthode de Newton. Elle présentait cependant plusieurs inconvénients. L'un d'entre eux est qu'elle nécessite d'abord de calculer le gradient ∇f de f, puis sa différentielle – c'est-à-dire la Hessienne H_f de f – puisque c'est en fait à cette fonction que la méthode s'applique. On peut donc se demander s'il n'existerait pas une procédure plus directe qui puisse s'appliquer directement à f.

Une famille de méthodes de ce type est donnée par les méthodes dites de *descente de gradient*. L'idée est la suivante : on part d'un point x_0 et on se déplace en suivant la pente donnée par le gradient, ce qui donne un point

$$x_1 = x_0 - t_1 \nabla f(x_0).$$

On définit ainsi une suite par une formule de récurrence de la forme

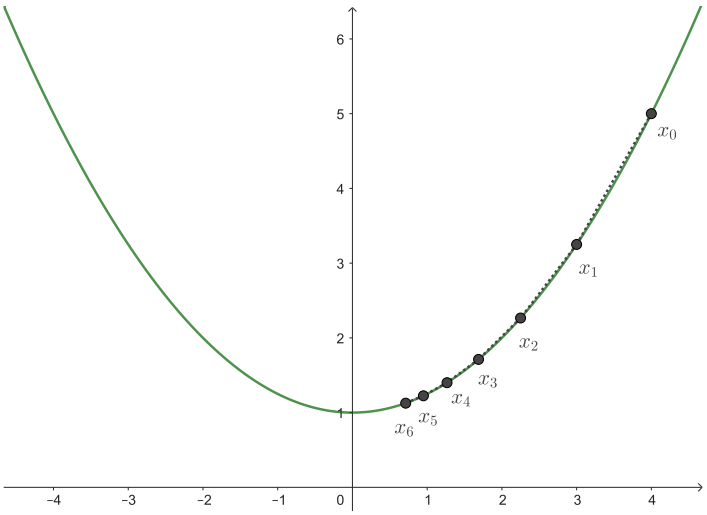
$$x_{k+1} = x_k - t_k \nabla f(x_k), \tag{4.1}$$

où $t_k > 0$ est appelé le *pas* de la méthode à l'étape k.

Pour comprendre l'intérêt de cette méthode, regardons ce qui se passe pour une fonction d'une variable. On considère donc une fonction $f : I \to \mathbf{R}$ possédant un minimum local strict en un point $\widetilde{x} \in I$. Le développement limité de f au point $\widetilde{x}$ s'écrit

$$f(x) = f(\widetilde{x}) + \frac{(x-\widetilde{x})^2}{2} f''(\widetilde{x}) + o((x-\widetilde{x})^2),$$

donc si $f''(\widetilde{x}) > 0$, sur un petit intervalle $[\widetilde{x}-\delta;\widetilde{x}+\delta]$ sur lequel on peut négliger le dernier terme, la fonction ressemble au dessin suivant (sur lequel on a représenté à titre d'exemple les points de coordonnées $(x_k, f(x_k))$ pour un pas $t_k = 0.5$ avec $x_0 = 4$).



On voit donc que si $x \in [\widetilde{x}-\delta;\widetilde{x}+\delta]$, alors on a deux possibilités :

- Si $x > \widetilde{x}$, $f'(x) > 0$ donc $x - tf'(x) < x$;
- Si $x < \widetilde{x}$, $f'(x) < 0$ donc $x - tf'(x) > x$.

Ainsi, $-f'(x)$ nous indique toujours la "bonne direction", celle qui nous rapproche de $\widetilde{x}$. Plus précisément encore, elle indique la direction dans la quelle f décroît, donc celle qui mène au minimum.

Il en est de même dans le cas de plusieurs variables. En effet, en utilisant le développement limité au premier ordre (Proposition 2.3.9) on voit que pour tout vecteur $v \in \mathbf{R}^n$,

$$\begin{aligned} f(x-tv) - f(x) &= \langle \nabla f(x), -tv \rangle + o(\|x-\widetilde{x}\|) \\ &= -t\langle \nabla f(x), v \rangle + o(\|x-\widetilde{x}\|). \end{aligned}$$

Si on oublie le second terme $o(\|x-\widetilde{x}\|)$, alors pour que l'écart entre $f(x-tv)$ et $f(x)$ soit le plus petit possible, il faut que le produit scalaire $\langle \nabla f(x), v \rangle$ soit le plus grand possible, et l'Inégalité de Cauchy-Schwarz nous dit que ceci se produit exactement quand v est positivement colinéaire à $\nabla f(x)$. Ainsi, le gradient de f indique ce qu'on appelle la *direction de plus grande pente*.

La suite donnée par la relation (4.1) converge-t-elle vers $\widetilde{x}$? En général non, mais comme le suggère la figure ci-dessus, il y a une chance que cela fonctionne si la fonction a une allure suffisamment sympathique. C'est ici que la notion de convexité s'avère utile. En effet, une fonction convexe d'une variable est une fonction qui ressemble à celle de la figure. Pour une telle fonction, on devine que la méthode fonctionne, pour peu qu'on choisisse habilement le nombre t_k à chaque étape. Ceci suggère donc de développer une notion de convexité pour les fonctions de plusieurs variables, ce que nous allons faire de ce pas.

4.2 La théorie

Vuelve la noche cóncava que descifró Anaxágoras
Vuelve a mi carne humana la eternidad constante
Y el recuerdo ¿el proyecto? de un poema incesante :
"Lo supieron los arduos alumnos de Pitágoras [1] *..."*

G.-L. Borges, *La noche cíclica*

4.2.1 Définition

Nous allons maintenant développer une notion de fonction convexe à plusieurs variables. Suivant l'intuition donné par le cas scalaire, pour une fonction $f : U \to \mathbf{R}$ on voudrait demander que soit satisfaite la condition suivante :

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

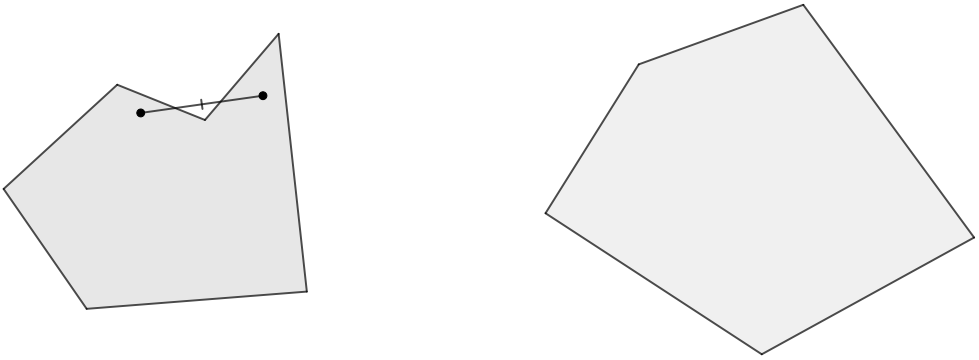
Il y a néanmoins un problème : le point $x+t(y-x)$ n'a aucune raison d'appartenir à U, et le membre de gauche peut donc très bien ne pas avoir de sens. Pour y remédier, il nous faut ainsi d'abord une notion de partie convexe de $\mathbf{R}^n$.

Définition 4.2.1. Une partie $U \subset \mathbf{R}^n$ est dite *convexe* si pour tous $x, y \in U$ et $t \in [0;1]$,

$$x + t(y-x) \in U.$$

1. Revient la nuit concave que déchiffra Anaxagore
Revient dans ma chair humaine l'éternité constante
Et le souvenir – le projet ? – incessant d'un poème :
"Les élèves assidus de Pythagore le savaient ..."

Autrement dit, si U contient deux points, alors elle contient tout le segment délimité par ces deux points. Quelques dessins permettent de saisir cette notion somme toute assez intuitive. Saurez-vous dire lequel de ces deux polygones est convexe et lequel ne l'est pas ?



Un exemple fondamental de partie convexe de $\mathbf{R}^n$ est donné par les boules.

Proposition 4.2.2. *Les boules ouvertes et fermées sont convexes.*

Démonstration. C'est un résultat que nous avons déjà établi sans le dire au cours de certaines démonstrations. Soit $x \in \mathbf{R}^n$ et $r > 0$. Si $y, y' \in B(x,r)$ et $t \in [0;1]$, alors

$$\begin{aligned} \|y + t(y'-y) - x\| &= \|(1-t)(y-x) + t(y'-x)\| \\ &\leqslant (1-t)\|y-x\| + t\|y'-x\| \\ &< (1-t)r + tr \\ &= r \end{aligned}$$

donc $y + t(y'-y) \in B(x,r)$. Si $y, y' \in B_f(x,r)$, le même calcul est valable excepté qu'à l'avant-dernière ligne on a une inégalité large, ce qui montre précisément que

$$y + t(y'-y) \in B_f(x,r).$$

■

Avant de passer aux fonctions, mentionnons un résultat pratique que nous avons, en fait, déjà utilisé de façon discrète : il s'agit de la généralisation de l'Inégalité des accroissements finis. Rappelons d'abord le résultat pour les fonctions d'une variable. Si $f : I \to \mathbf{R}$ est dérivable et si sa dérivée est majorée par une constante M, alors pour tout $a, b \in I$,

$$|f(b) - f(a)| \leqslant M|b-a|.$$

Proposition 4.2.3 (Inégalité des accroissements finis)**.** *Soit U un ouvert* convexe *de $\mathbf{R}^n$ et $f : U \to \mathbf{R}$ une fonction de classe $\mathcal{C}^1$ telle qu'il existe $M > 0$ vérifiant*

$$\|\nabla f(x)\| \leqslant M$$

pour tout $x \in U$. Alors, pour tous $x, y \in U$,

$$\boxed{\|f(x) - f(y)\| \leqslant M\|x-y\|}.$$

Démonstration. Il suffit de considérer la fonction $g : [0;1] \to \mathbf{R}$ définie par

$$g(t) = f\left(x + t(y-x)\right),$$

qui a un sens précisément parce que U est convexe. La dérivation des fonctions composées (Proposition 2.3.12) donne

$$g'(t) = \langle \nabla f\left(x + t(y-x)\right), y - x\rangle$$

dont on déduit

$$\begin{aligned} |g'(t)| &= \left|\langle \nabla f\left(x + t(y-x)\right), y-x\rangle\right| \\ &\leqslant \left\|\nabla f\left(x + t(y-x)\right)\right\| \|y-x\| \\ &\leqslant M\|y-x\|. \end{aligned}$$

Il ne reste plus qu'à appliquer l'Inégalité des accroissements finis (pour les fonctions d'une variable) à g pour obtenir

$$\begin{aligned} |f(y) - f(x)| &= |g(1) - g(0)| \\ &\leqslant M\|y-x\| \times |1-0| \\ &= M\|x-y\|. \end{aligned}$$

■

Nous pouvons maintenant définir la convexité pour les fonctions de plusieurs variables.

Définition 4.2.4. Soit U une partie convexe de $\mathbf{R}^n$. Une fonction $f : U \to \mathbf{R}$ est dite *convexe* si pour tous $x, y \in U$ et $t \in [0;1]$,

$$\boxed{f\left(x + t(y-x)\right) \leqslant f(x) + t\left(f(y) - f(x)\right).}$$

Si l'inégalité précédente est stricte dès que $x \neq y$ et $t \in]0;1[$, alors f sera dite *strictement convexe*.

Remarque 4.2.5. L'*épigraphe* d'une fonction $f : U \to \mathbf{R}$ est l'ensemble

$$E(f) = \{(\lambda, x) \in \mathbf{R} \times U \mid \lambda \geqslant f(x)\}.$$

On voit aisément que f est convexe si et seulement si son épigraphe est une partie convexe de $\mathbf{R} \times \mathbf{R}^n \simeq \mathbf{R}^{n+1}$. En effet, si f est convexe et $(\lambda, x), (\mu, y) \in E(f)$, alors pour tout $t \in [0;1]$,

$$\begin{aligned} \lambda + t(\mu - \lambda) &= (1-t)\lambda + t\mu \\ &\geqslant (1-t)f(x) + tf(y) \\ &= f(x) + t(f(y) - f(x)) \\ &\geqslant f(x + t(y-x)), \end{aligned}$$

donc $(\lambda + t(\mu - \lambda), x + t(y-x)) \in E(f)$. Réciproquement, si $E(f)$ est convexe et $x, y \in U$, alors pour tout $t \in [0;1]$ on a $(f(x) + t(f(y) - f(x)), x + t(y-x)) \in E(f)$, donc

$$f(x) + t(f(y) - f(x)) \geqslant x + t(y-x)$$

et f est convexe.

Avant d'aller plus loin, faisons une petite pause pour nous poser une question oiseuse. La Définition 4.2.4 n'utilise que deux points, mais on pourrait vouloir faire des combinaisons de trois points ou plus. Qu'en est-il alors ? En fait, cela ne change pas la notion considérée.

Proposition 4.2.6 (Inégalité de Jensen). *Soit $f : U \to \mathbf{R}$ une fonction convexe. Alors, pour tous points $x_1, \cdots, x_n \in U$ et tous $0 < t_1, \cdots, t_n < 1$ tels que*

$$\sum_{i=1}^{n} t_i = 1,$$

on a

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

Démonstration. Nous allons démontrer ce résultat par récurrence, en notant H_n l'inégalité pour n points. Comme H_1 est triviale, nous allons commencer par montrer H_2, qui n'est pas exactement la définition de la convexité que nous avons donnée dans la Définition 4.2.4. Pour cela, remarquons que comme $t_1 + t_2 = 1$, on a $t_1 = 1 - t_2$. Par conséquent,

$$t_1 x_1 + t_2 x_2 = x_1 + t_2(x_2 - x_1)$$

et donc par définition de la convexité

$$\begin{aligned} f(t_1 x_1 + t_2 x_2) &\leqslant f(x_1) + t_2(f(x_2) - f(x_1)) \\ &= (1 - t_2) f(x_1) + t_2 f(x_2) \\ &= t_1 f(x_1) + t_2 f(x_2). \end{aligned}$$

Supposons maintenant H_n vraie pour un $n \geqslant 2$, et considérons $n+1$ points. L'idée est de regrouper partiellement [2] les points pour n'en avoir plus que deux et pouvoir appliquer H_2. Ceci se fait comme suit : on pose $t'_2 = t_{n+1}$ et

$$t'_1 = \sum_{i=1}^{n} t_i = 1 - t'_2.$$

En posant $x'_2 = x_n$ et

$$x'_1 = \sum_{i=1}^{n} \frac{t_i}{t'_1} x_i,$$

on a

$$\begin{aligned} f\left(\sum_{i=1}^{n+1} t_i x_i\right) &= f\left(t'_1 x'_1 + t'_2 x'_2\right) \\ &\leqslant t'_1 f(x'_1) + t'_2 f(x'_2) \\ &= t'_1 f(x'_1) + t_{n+1} f(x_{n+1}). \end{aligned}$$

Si l'on pose maintenant, pour $1 \leqslant i \leqslant n$, $s_i = t_i / t'_1$, on a

$$\sum_{i=1}^{n} s_i = 1.$$

2. Les personnes amatrices de géométrie affine y reconnaitront la notion de *barycentre partiel*.

On peut donc appliquer H_n pour conclure que

$$\begin{aligned} t_1' f(x_1') &= t_1' f\left(\sum_{i=1}^{n} s_i x_i\right) \\ &\leqslant \sum_{i=1}^{n} t_1' s_i f(x_i) \\ &= \sum_{i=1}^{n} t_i f(x_i). \end{aligned}$$

■

Le terme "convexe" va rarement sans son alter ego "concave". Une petite précaution s'impose néanmoins : il n'existe pas de notion de partie concave d'un espace vectoriel. Quant aux fonctions, il s'agit d'un simple changement de signe :

Définition 4.2.7. Soit $U \subset \mathbf{R}^n$ une partie convexe. Une fonction $f : U \to \mathbf{R}$ est dite *(strictement) concave* si la fonction $-f$ est (strictement) convexe.

Il est clair que tout énoncé sur les fonctions convexes se traduit sans peine en un énoncé sur les fonctions concaves. Nous ne mentionnerons donc pas ce dernier afin de ne pas alourdir le texte ou les énoncés.

4.2.2 Propriétés des fonctions convexes

Les fonctions convexes sont importantes parce qu'elle jouissent de propriétés remarquables. Par conséquent, il est important d'être capable de reconnaître rapidement une telle fonction. Pour cela, quelques propriétés de permanence peuvent s'avérer utiles.

Proposition 4.2.8. *Soit $U \subset \mathbf{R}^n$ une partie convexe.*

1. *Si $f, g : U \to \mathbf{R}$ sont convexes, alors $f + g$ est convexe.*
2. *Si $f : U \to \mathbf{R}$ est convexe et $\lambda > 0$, alors λf est convexe.*
3. *Si $f : U \to \mathbf{R}$ est convexe et $g : f(U) \to \mathbf{R}$ est convexe* et croissante, *alors $g \circ f$ est convexe.*
4. *Si $f : U \to \mathbf{R}$ est convexe et $g : \mathbf{R}^n \to \mathbf{R}^n$ est affine, alors $f \circ g$ est convexe.*

Démonstration. Il suffit à chaque fois de vérifier directement la définition. Soient donc $x, y \in \mathbf{R}^n$ et $t \in [0;1]$.

1. On a

$$\begin{aligned} [f+g](x+t(y-x)) &= f(x+t(y-x)) + g(x+t(y-x)) \\ &\leqslant f(x) + t(f(x) - f(y)) + g(x) + t(g(y) - g(x)) \\ &= [f+g](x) + t([f+g](y) - [f+g](x)). \end{aligned}$$

2. On a

$$\begin{aligned} [\lambda f](x+t(y-x)) &\leqslant \lambda(f(x) + t(f(y) - f(x)) \\ &= [\lambda f](x) + t([\lambda f](y) - [\lambda f](x)) \end{aligned}$$

3. On a $f(x+t(y-x)) \leqslant f(x)+t(f(y)-f(x))$ par convexité de f, donc par croissance de g

$$\begin{aligned}[g\circ f](x+t(y-x)) &\leqslant g\left(f(x)+t(f(y)-f(x))\right)\\ &\leqslant g(f(x))+t\left(g(f(y))-g(f(x))\right)\\ &= [g\circ f](x)+t\left([g\circ f](y)-[g\circ f](x)\right).\end{aligned}$$

4. On peut écrire $g(x)=Ax+b$ pour une matrice $A\in \mathrm{M}_n(\mathbf{R})$ et $b\in\mathbf{R}^n$. On a alors

$$\begin{aligned}[f\circ g](x+t(y-x)) &= f\left(A(x+t(y-x))+b\right)\\ &= f\left(Ax+b+t(Ax-Ay)\right)\\ &= f\left(Ax+b+t((Ay+b)-(Ax+b))\right)\\ &= f\left(g(x)+t(g(y)-g(x))\right)\\ &\leqslant f(g(x))+t\left(f(g(y))-f(g(x))\right)\\ &= [f\circ g](x)+t\left([f\circ g](y)-[f\circ g](x)\right).\end{aligned}$$

■

Remarque 4.2.9. L'hypothèse sur la croissance de g dans le troisième point de l'énoncé ci-dessus est indispensable. En effet, considérons les fonctions $f : x\mapsto x^2$ et $g : x\mapsto e^{-x}$. Elles sont toutes deux convexes, mais leur composée

$$g\circ f : x\mapsto e^{-x^2}$$

n'est pas convexe. On peut le voir en observant que sa dérivée seconde est égale à $2(1-x^2)e^{-x^2}$ qui prend des valeurs strictement négative (nous généraliserons au cas de plusieurs variables cette caractérisation de la convexité un peu plus loin dans la Proposition 4.2.25). Bien entendu, la fonction g n'est pas croissante.

Nous avons vu que pour faire de l'optimisation, il faut en général des propriétés de régularité sur les fonctions considérées telles que la continuité ou le caractère $\mathcal{C}^1$. Comme nous l'avons déjà mentionné, certaines de ces propriétés sont "presque" automatiques pour les fonctions convexes. Même si nous n'en aurons pas besoin ultérieurement, nous allons donner un exemple concernant la continuité. Afin de simplifier un peu le texte, nous dirons dans la suite qu'un point $x\in\mathbf{R}^n$ est *combinaison convexes* de points $a_1,\cdots,a_m$ s'il existe $0\leqslant t_1,\cdots,t_m\leqslant 1$ tels que

$$\sum_{i=1}^m t_i = 1 \quad \& \quad x=\sum_{i=1}^m t_i a_i.$$

Proposition 4.2.10. *Soit $U\subset\mathbf{R}^n$ une partie convexe et $f : U\to\mathbf{R}$ une fonction convexe. Si x est à l'intérieur de U, alors f est continue en x.*

Démonstration. La preuve recèle quelques subtilités et nous allons donc procéder par étapes. Pour commencer, fixons un point x à l'intérieur de U.

► *Restriction à un hypercube*

Comme x est intérieur à U, il existe $r>0$ tel que $B(x,r)\subset U$. L'idée est mainentnat de montrer que f est bornée sur la boule $B(x,r)$. Néanmoins, pour une fonction convexe

il est plus aisé de travailler avec des polytopes qu'avec des boules. Or, il y en a un assez simple qui fera l'affaire : un *hypercube*. Plus précisément, en notant $x = (x_1, \cdots, x_n)$, nous allons nous intéresser à l'ensemble

$$C = \prod_{i=1}^{n} \left[x_i - \frac{r}{\sqrt{2n}}; x_i + \frac{r}{\sqrt{2n}} \right].$$

La première étape est de montrer que $C \subset U$. Pour cela, prenons $y \in C$ et observons que toutes les coordonnées de $y - x$ sont inférieures en valeur absolue à $r/\sqrt{2n}$. Par conséquent,

$$\|y - x\|^2 \leqslant \sum_{i=1}^{n} \frac{r^2}{2n} \leqslant \frac{r^2}{2} < r^2.$$

Quel est l'intérêt de cet ensemble C ? C'est qu'il est convexe mais n'a qu'un nombre fini de *sommets*. Plus concrètement, les sommets de C sont les points dont toutes les coordonnées sont de la forme $x_i \pm r/\sqrt{2n}$ et il y en a donc 2^n. On notera désormais $S = \{a_1, \cdots, a_{2^n}\}$ cet ensemble fini.

► *Hypercube et points extrémaux*

L'étape suivante est de montrer que f est bornée sur l'hypercube C. Pour cela, il nous faut établir un fait important sur ce dernier : tout point de C est une combinaison convexe de ses sommets. Ceci peut aisément se démontrer par récurrence sur la dimension n. En effet, pour $n = 1$, C est un segment et tout point d'un segment est combinaison convexe de ses extrémités. Si le résultat est vrai pour un certain $n \geqslant 1$, considérons un point y de C. Il existe $t \in [0;1]$ tel que

$$y_1 = t\left(x_1 - \frac{r}{\sqrt{2n}}\right) + (1-t)\left(x_1 + \frac{r}{\sqrt{2n}}\right).$$

Les deux points y^+ et y^- obtenus en remplaçant la première coordonnée de y par $x_1 \pm r/\sqrt{2n}$ vérifient donc $y = ty^- + (1-t)y^+$. De plus,

$$y' = (y_2, \cdots, y_n) \in C' = \prod_{i=2}^{n+1} \left[x_i - \frac{r}{\sqrt{2n}}; x_i + \frac{r}{\sqrt{2n}} \right],$$

donc par l'hypothèse de récurrence, il existe des réels $0 \leqslant s_1, \cdots s_{2^{n-1}} \leqslant 1$ dont la somme fait 1 et tels que – en utilisant l'Inégalité de Jensen de la Proposition 4.2.6 –

$$y' = \sum_{j=1}^{2^{n-1}} s_j b_j,$$

où $(b_j)_{1 \leqslant j \leqslant 2^{n-1}}$ désigne l'ensemble des sommets de C'. Pour conclure, il suffit d'observer, en notant b_j^+ (respectivement b_j^-) le vecteur de $\mathbf{R}^n$ obtenu à partir de b_j en ajoutant comme première coordonnée $x_i - \frac{r}{\sqrt{2n}}$ (respectivement $x_i + \frac{r}{\sqrt{2n}}$), que

$$\begin{aligned} y &= ty^- + (1-t)y^+ \\ &= t \sum_{j=1}^{2^{n-1}} s_j b_j^- + (1-t) \sum_{j=1}^{2^{n-1}} s_j b_j^+ \\ &= \sum_{i=1}^{2^n} t_i a_i, \end{aligned}$$

où $t_i = ts_i$ si la première coordonnée de a_i est $x_i - \frac{r}{\sqrt{2n}}$ et $t_i = (1-t)s_i$ sinon. Il reste à verifier que

$$\begin{aligned}\sum_{i=1}^{2^n} t_i &= t\sum_{j=1}^{2^{n-1}} s_j + (1-t)\sum_{j=1}^{2^{n-1}} s_j \\ &= t + (1-t) \\ &= 1.\end{aligned}$$

► *La fonction f est localement bornée*

Pour borner f, nous allons poser $M = \max_S f$. Alors, si $y \in C$, il existe $0 \leqslant t_1 \leqslant \cdots \leqslant t_{2^n} \leqslant 1$ tels que, par la Proposition 4.2.6,

$$\begin{aligned}f(y) &= f\left(\sum_{i=1}^{n} t_i a_i\right) \\ &\leqslant \sum_{i=1}^{n} t_i f(a_i) \\ &\leqslant \sum_{i=1}^{n} t_i M \\ &= M.\end{aligned}$$

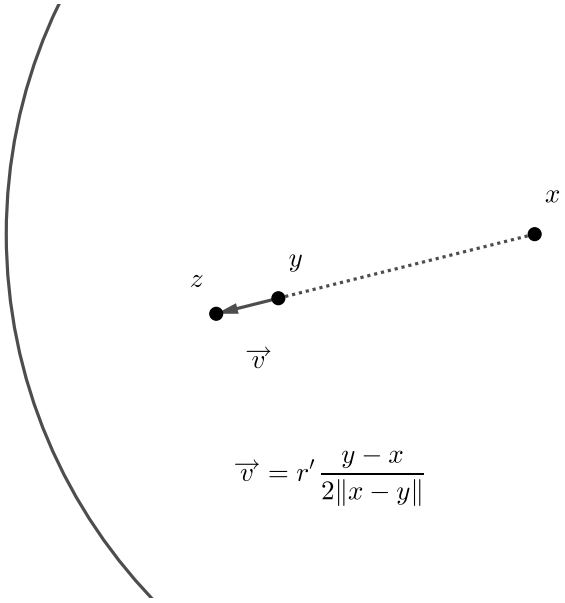
Ainsi, $f(y) \leqslant M$ pour tout $y \in C$. Si on pose $r' = r/\sqrt{2n}$, alors $B(x,r') \subset C$ donc f est bornée sur $B(x,r')$, ce que nous voulions prouver.

► *La fonction f est continue*

Considérons maintenant un point $y \in B(x,r'/2)$ et recourons à une astuce : il s'agit de construire un élément de $B(x,r')$ dont la bornitude nous permettra de majorer $f(x) - f(y)$. La solution la plus simple pour cela est de poser

$$z = y + \frac{r'}{2\|y-x\|}(y-x).$$

Géométriquement, z est obtenu en partant de y et en se déplaçant dans la direction opposée à x, mais pas trop de façon à rester dans la boule $B(x,r')$. Voici une illustration en deux dimensions :



En effet, on a bien

$$\begin{aligned}\|z-x\| &\leqslant \|y-x\| + \frac{r'}{2\|y-x\|}\|y-x\| \\ &< \frac{r'}{2} + \frac{r'}{2} \\ &= r'.\end{aligned}$$

De plus, en posant $\alpha = r'/2\|y-x\|$, on peut écrire $u = y + \alpha(y-x)$ et donc

$$y = \frac{1}{\alpha+1} z + \frac{\alpha}{\alpha+1} x.$$

Comme $\alpha > 0$, les deux coefficients ci-dessus sont positifs et leur somme fait 1. On peut donc appliquer la convexité de f pour obtenir

$$f(y) \leqslant \frac{1}{\alpha+1} f(z) + \frac{\alpha}{\alpha+1} f(x).$$

En soustrayant $f(x)$ aux deux membres on trouve finalement

$$\begin{aligned}f(y) - f(x) &\leqslant \frac{1}{\alpha+1}(f(z) - f(x)) \\ &\leqslant \frac{2M}{\alpha+1} \\ &= 2M \frac{2\|y-x\|}{r' + 2\|y-x\|} \\ &\leqslant \frac{4M}{r'}\|y-x\|.\end{aligned}$$

En échangeant les rôles de x et y, on obtient la même majoration pour $f(x) - f(y)$, ce qui permet de conclure que

$$\big|f(y) - f(x)\big| \leqslant \frac{4M}{r'}\|y-x\|.$$

Nous pouvons maintenant montrer la continuité : pour tout $\epsilon > 0$, si

$$y \in B\left(x, \min\left(r', \frac{4M\epsilon}{r'}\right)\right)$$

alors $|f(y) - f(x)| \leqslant \epsilon$ d'après ce qui précède, ce qu'il fallait démontrer. ∎

Remarque 4.2.11. On pourra noter que nous avons montré un résultat un peu plus fort : une fonction convexe est *localement lipschitzienne* sur l'intérieur de son domaine de définition.

4.2.3 Les théorèmes améliorés

Le grand intérêt des fonctions convexes, c'est qu'en s'y restreignant on obtient des version améliorées des théorèmes généraux sur les extrema de fonctions différentiables. Par exemple, les points critiques – qui sont faciles à détecter – ne sont pas toujours des extrema. Eh bien pour les fonctions convexes, si ! Cela dit, la preuve nécessite d'abord de mieux comprendre le lien entre convexité et dérivées partielles.

Pour une fonction d'une variable $f : I \to \mathbf{R}$ dérivable, on sait bien que la convexité est équivalente à la croissance de la fonction dérivée f'. Dans le cas de plusieurs variables, un résultat analogue est valable, à condition de comprendre ce que la croissance signifie.

Proposition 4.2.12. *Soit U un ouvert convexe de $\mathbf{R}^n$ et $f : U \to \mathbf{R}$ une fonction de classe $\mathcal{C}^1$. Alors, f est convexe si et seulement si pour tous $x, y \in U$,*

$$\boxed{f(y) - f(x) \geqslant \langle \nabla f(x), y - x \rangle.}$$

De plus, f est strictement convexe si et seulement si cette inégalité est stricte dès que $x \neq y$.

Démonstration. Il s'agit en fait simplement d'une version "infinitésimale" de la définition de la convexité, c'est-à-dire une version obtenue en faisant habilement tendre t vers 0. Voici comment procéder. Si f est convexe on a, pour tout $t \in]0;1]$,

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

ce qui peut s'écrire

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

En divisant par t, cette inégalité devient

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

Posons $g : t \in [0;1] \mapsto f(x + t(y - x))$. Il s'agit d'une fonction de classe $\mathcal{C}^1$ et le membre de droite ci-dessus en est un taux d'accroissement en 0. Si on peut la prolonger en une fonction de classe $\mathcal{C}^1$ sur $]-\delta;1]$ pour un $\delta > 0$, alors on pourra passer à la limite et obtenir $g'(0)$. Pour ce faire, remarquons que U est ouvert, donc qu'il existe $r > 0$ tel que $B(x, r) \subset U$. Alors, en posant $\delta = r/\|y - x\|$, on a bien

$$\left\|x + t(y - x) - x\right\| < r$$

pour $t \in]-\delta;\delta[$ et on peut donc prolonger g. Quand $t \to 0$, le membre de droite tend donc vers $g'(0)$ qui par dérivation des fonctions composées (Proposition 2.3.12) est égal à $\langle \nabla f(x), y - x) \rangle$, et le résultat suit.

Réciproquement, nous allons utiliser l'inégalité de l'énoncé entre les points x et $x + t(y - x)$ d'une part, et $x + t(y - x)$ et y d'autre part. Cela donne pour les premiers points

$$\begin{aligned} f(x) &\geqslant f\left(x + t(y-x)\right) + \left\langle \nabla f\left(x + t(y-x)\right), x - (x + t(y-x))\right\rangle \\ &= f\left(x + t(y-x)\right) - t\left\langle \nabla f\left(x + t(y-x)\right), y - x\right\rangle \end{aligned}$$

et de même pour les seconds

$$\begin{aligned} f(y) &\geqslant f(x + t(y-x)) + \left\langle \nabla f\left(x + t(y-x)\right), y - (x + t(y-x))\right\rangle \\ &= f(x + t(y-x)) + (1-t)\left\langle \nabla f\left(x + t(y-x)\right), y - x\right\rangle. \end{aligned}$$

Notre objectif étant de faire disparaître les gradients, nous allons multiplier la première inégalité par $(1 - t)$, la seconde par t, et additionner pour trouver

$$(1-t)f(x) + tf(y) \geqslant f\left(x + t(y-x)\right).$$

■

Avec ce résultat, on peut facilement montrer que points critiques et extrema coïncident.

Théorème 4.2.13 (Points critiques des fonctions convexes) Soit U un ouvert convexe de $\mathbf{R}^n$ et $f : U \to \mathbf{R}$ une fonction convexe de classe $\mathcal{C}^1$. Si x est un point critique de f, alors f a un *minimum global* en x.

Démonstration. Il suffit d'appliquer la Proposition 4.2.12 : comme x est un point critique, on a pour tout $y \in U$,

$$\begin{aligned} f(y) - f(x) &\geqslant \langle \nabla f(x), y - x\rangle \\ &= 0, \end{aligned}$$

donc $f(y) \geqslant f(x)$. ■

Il n'y a bien sûr pas unicité en général – il suffit de penser à une fonction constante sur $\mathbf{R}$ – mais on peut l'obtenir en renforçant un peu l'hypothèse sur f :

Lemme 4.2.14. *Soit $f : U \to \mathbf{R}$ une fonction* strictement *convexe. Alors, si f admet un minimum global en un point x, ce minimum est unique.*

Démonstration. Soit y un point tel que $f(y) = f(x)$. Alors, en prenant par exemple $t = 1/2$ dans la définition de la convexité on a

$$f\left(\frac{x+y}{2}\right) \leqslant \frac{f(x) + f(y)}{2} = f(x).$$

Comme l'inégalité ne peut être stricte, on doit donc avoir $x = y$ par convexité stricte de f. ■

Une conséquence du résultat précédent est que tout minimum local d'une fonction convexe est en fait global ! Il se trouve que l'on n'a pas besoin de supposer la fonction de classe $\mathcal{C}^1$ ni d'utiliser le calcul différentiel pour montrer ce résultat. À titre de curiosité, et pour montrer la force et l'intérêt de l'analyse convexe non différentiable, nous en donnons une démonstration.

Proposition 4.2.15. *Soit $U \subset \mathbf{R}^n$ une partie convexe et soit $f : U \to \mathbf{R}$ une fonction convexe. Si f a un minimum local en un point $x \in U$, alors c'est un minimum global.*

Démonstration. Par définition d'un minimum local, il existe $\delta > 0$ tel que si $y \in U$ vérifie $\|x-y\| < \delta$, alors $f(x) \leqslant f(y)$. Si maintenant on considère un point $z \in U$ quelconque, posons, pour $t \in [0;1]$,

$$z_t = x + t(z-x).$$

Alors,

$$\begin{aligned} \|z_t - x\| &= \|t(z-x)\| \\ &= t\|z-x\|. \end{aligned}$$

Ainsi, pour $t_0 = \delta/2\|z-x\|$, on a $f(z_{t_0}) \geqslant f(x)$. On conclut ensuite en utilisant la convexité :

$$\begin{aligned} t_0 f(z) &= f(x) + t_0(f(z) - f(x)) - (1-t_0)f(x) \\ &\geqslant f(z_{t_0}) - (1-t_0)f(x) \\ &\geqslant f(x) - (1-t_0)f(x) \\ &= t_0 f(x). \end{aligned}$$

■

Le Théorème 4.2.13 affirme que la condition d'annulation du gradient, qui n'est que nécessaire pour l'existence d'un extremum, est également suffisante. Ce résultat reste encore vrai sous contraintes, fournissant ainsi une réciproque au Théorème 3.2.6. De fait, nous avons déjà vu cela dans le cas particulier des fonctions affines, qui sont bien évidemment convexes. Nous donnons ici un résultat plus général, en avertissant néanmoins qu'il contient, outre la convexité de la fonction f à minimiser, des hypothèses fortes sur les fonctions définissant les contraintes. En effet, si l'on veut pouvoir exploiter la convexité de f, alors il faut que l'ensemble $\mathcal{D}$ sur lequel on cherche à l'optimiser soit lui-même convexe. Pour les contraintes d'inégalité, cela se traduit facilement :

Lemme 4.2.16. *Soient $h_1, \cdots, h_p : U \to \mathbf{R}$ des fonctions convexes. Alors, l'ensemble*

$$\mathcal{D} = \{x \in U \mid h_j(x) \leqslant 0 \text{ pour tout } 1 \leqslant j \leqslant p\}$$

est convexe.

Démonstration. Il suffit de le faire pour une seule fonction, puisqu'une intersection de parties convexes est convexe [3]. Or, si $h(x), h(y) \leqslant 0$, alors on a pour tout $t \in [0;1]$

$$\begin{aligned} h(x + t(y-x)) &\leqslant f(x) + t(f(y) - f(x)) \\ &= (1-t)f(x) + tf(y) \\ &\leqslant 0. \end{aligned}$$

■

Pour les contraintes d'égalité, par contre, les choses se compliquent. En effet, il est très rare qu'un ensemble de la forme $\{x \in U \mid g(x) = 0\}$ soit convexe. Le seul cas où il est facile de s'assurer que c'est le cas est celui des fonctions affines. En effet, l'ensemble considéré est alors un sous-espace affine, donc une partie convexe. Nous supposerons par conséquent que les contraintes d'égalité sont toutes de cette forme.

3. Ceci a été montré en TD.

Théorème 4.2.17 (Points critiques sous contraintes des fonctions convexes) Soit U un ouvert convexe de $\mathbf{R}^n$ et $f : U \to \mathbf{R}$ une fonction convexe de classe $\mathcal{C}^1$. Soient également $g_1, \cdots, g_m : U \to \mathbf{R}$ des fonctions **affines** et $h_1, \cdots, h_p : U \to \mathbf{R}$ des fonction **convexes** de classe $\mathcal{C}^1$. Soit $\widetilde{x} \in \mathcal{D}$ tel qu'il existe $\lambda_1, \cdots, \lambda_m \in \mathbf{R}$ et $\mu_1, \cdots, \mu_p \in \mathbf{R}_-$ vérifiant

$$\nabla f(\widetilde{x}) = \sum_{i=1}^{m} \lambda_i \nabla g_i(\widetilde{x}) + \sum_{j=1}^{p} \mu_j \nabla h_j(\widetilde{x}).$$

Alors, f admet en $\widetilde{x}$ un minimum global sous contraintes.

Démonstration. L'idée est la même que pour la seconde partie de la démonstration du Théorème 3.3.1. En effet, pour tout $x \in \mathcal{D}$ on a

$$\begin{aligned} f(x) - f(\widetilde{x}) &\geqslant \langle \nabla f(\widetilde{x}), x - \widetilde{x} \rangle \\ &= \sum_{i=1}^{m} \lambda_i \langle \nabla g_i(\widetilde{x}), x - \widetilde{x} \rangle + \sum_{j=1}^{p} \mu_j \langle \nabla h_j(\widetilde{x}), x - \widetilde{x} \rangle \\ &= \sum_{i=1}^{m} \lambda_i \langle v_i, x \rangle + \sum_{i=1}^{m} \lambda_i \langle v_i, \widetilde{x} \rangle + \sum_{j=1}^{p} \mu_j \langle \nabla h_j(\widetilde{x}), x - \widetilde{x} \rangle \\ &= \sum_{i=1}^{m} \lambda_i c_i - \sum_{i=1}^{m} \lambda_i c_i + \sum_{j=1}^{p} \mu_j \langle \nabla h_j(\widetilde{x}), x - \widetilde{x} \rangle \\ &= \sum_{j=1}^{p} \mu_j \langle \nabla h_j(\widetilde{x}), x - \widetilde{x} \rangle \end{aligned}$$

La convexité de h_j donne alors

$$\langle \nabla h_j(\widetilde{x}), x - \widetilde{x} \rangle \leqslant h_j(x) - h_j(\widetilde{x}).$$

Si $h_j(\widetilde{x}) \neq 0$, alors la contrainte est inactive et $\mu_j = 0$. Sinon, on a

$$h_j(x) - h_j(\widetilde{x}) = h_j(x) \leqslant 0.$$

Comme ce terme est multiplié par $\mu_j \leqslant 0$, on trouve finalement

$$f(x) - f(\widetilde{x}) \geqslant 0,$$

ce qu'il fallait démontrer. ∎

Remarque 4.2.18. Cet énoncé, quoique proche de celui du Théorème 3.3.1, n'est pas tout à fait du même ordre. En effet, nous n'affirmons pas ici que les conditions KKT sont valables indépendamment de toute qualification des contraintes. Par contre, il est bien vrai que les conditions KKT sont suffisantes et que tout minimum local est global.

Ce dernier résultat suggère de regarder de plus près le comportement du Lagrangien sous les mêmes hypothèses. Et en effet, on peut alors améliorer la dualité faible donnée par la Proposition 3.4.3 en un résultat de dualité forte. Il s'agit d'un résultat important pour l'optimisation convexe, du à Slater [4].

4. Morton L. Slater (1921–2002) : mathématicien américain ayant travaillé dans de nombreuses branches de l'analyse : topologie, théorie de la mesure, équations différentielles, optimisation. Il est surtout resté dans l'histoire pour le résultat d'optimisation convexe qui porte son nom.

Théorème 4.2.19 (Dualité forte de Lagrange pour les problèmes convexes) Soit U un ouvert convexe de $\mathbf{R}^n$ et $f : U \to \mathbf{R}$ une fonction convexe de classe $\mathcal{C}^1$. Soient également $g_1, \cdots, g_m : U \to \mathbf{R}$ des fonctions **affines** et $h_1, \cdots, h_p : U \to \mathbf{R}$ des fonction **convexes** de classe $\mathcal{C}^1$. On suppose que

- Il existe $\widehat{x} \in \mathcal{D}$ tel que $h_j(\widehat{x}) < 0$ pour tout $1 \leqslant j \leqslant p$;
- Le point 0 est intérieur à $g(U) \subset \mathbf{R}^m$.

Si f admet un minimum (nécessairement global) sous contraintes en un point $\widetilde{x} \in \mathcal{D}$, alors il existe $\widetilde{\lambda} \in \mathbf{R}^n$ et $\widetilde{\mu} \in \mathbf{R}^p_-$ tels que

$$\boxed{f(\widetilde{x}) = \mathcal{L}(\widetilde{x}, \widetilde{\lambda}, \widetilde{\mu}) = f^*(\widetilde{\lambda}, \widetilde{\mu})}$$

En particulier, f admet un minimum global sous contraintes et les multiplicateurs correspondants donnent un maximum global sous contrainte du problème dual.

La démonstration de ce théorème nécessite un résultat général sur les parties convexes de $\mathbf{R}^n$ que nous allons d'abord énoncer et démontrer. On pourra noter sa ressemblance avec le Lemme de Farkas (Lemme 3.2.4) qui était crucial pour la démonstration du Théorème KKT (Théorème 3.2.6).

Lemme 4.2.20. *Soit $\mathcal{C} \subset \mathbf{R}^n$ une partie convexe ne contenant pas 0. Alors, il existe un vecteur $u \neq 0$ tel que*

$$\langle u, x \rangle \geqslant 0$$

pour tout $x \in \mathcal{C}$.

Démonstration. La preuve, quoique simple, recèle une subtilité. En effet, nous allons avoir besoin de travailler avec l'adhérence $\overline{\mathcal{C}}$ de $\mathcal{C}$, et bien que $0 \notin \mathcal{C}$, il se pourrait qu'il soit dans l'adhérence. Néanmoins, même dans ce cas, il existe une suite $(x_k)_{k \in \mathbf{N}}$ d'éléments de $\mathbf{R}^n$ telle que $x_k \to 0$ et $x_k \notin \overline{\mathcal{C}}$ pour tout $k \in \mathbf{N}$.

Posons $F_k : x \in \mathcal{C} \mapsto \|x - x_k\|$. Si $\mathcal{C}$ est bornée, il en est de même pour son adhérence qui est alors compacte. Dans ce cas, F_k étant continue elle admet un minimum global. Si $\mathcal{C}$ n'est pas bornée, alors F_k tend vers $+\infty$ quand $\|x\| \to +\infty$, donc une fois encore F_k admet un minimum global. Ainsi, dans tous les cas on dispose d'un point $\widetilde{x}_k$ qui réalise le minimum.

Nous allons maintenant montrer que pour tout $x \in \overline{\mathcal{C}}$,

$$\langle x - \widetilde{x}_k, \widetilde{x}_k - x_k \rangle \geqslant 0.$$

Pour ce faire, posons, pour $t \in [0;1]$, $x_t = \widetilde{x}_k + t(x - \widetilde{x}_k)$. Alors, $x_t \in \overline{\mathcal{C}}$ donc par minimalité on a $\|x_t - x_k\| \geqslant \|\widetilde{x}_k - x_k\|$. En élevant au carré cette égalité on obtient

$$\|\widetilde{x}_k - x_k\|^2 + 2t\langle x - \widetilde{x}_k, \widetilde{x}_k - x_k \rangle + t^2 \|x - \widetilde{x}_k\|^2 \geqslant \|\widetilde{x}_k - x_k\|^2,$$

ce qui en simplifiant donne

$$2t\langle x - \widetilde{x}_k, \widetilde{x}_k - x_k \rangle \geqslant -t^2 \|x - \widetilde{x}_k\|^2.$$

En divisant par t puis en faisant tendre ce dernier vers 0, on obtient l'inégalité annoncée.

Pour conclure, posons

$$u_k = \frac{\widetilde{x_k} - x_k}{\|\widetilde{x_k} - x_k\|}.$$

La suite $(u_k)_{k\in\mathbf{N}}$ est bornée car tous ses éléments sont de norme 1. Par conséquent, il existe une sous-suite $(u_{\phi(k)})_{k\in\mathbf{N}}$ qui converge vers un vecteur u qui vérifie $\|u\| = 1$ et donc en particulier $u \neq 0$. De plus,

$$\begin{aligned}\langle x, u_k\rangle &\geqslant \langle \widetilde{x_k}, u_k\rangle \\ &= \langle \widetilde{x_k} - x_k + x_k, u_k\rangle \qquad\qquad = \langle x_k, u_k\rangle + \|\widetilde{x_k} - x_k\| \\ &\geqslant \langle x_k, u_k\rangle\end{aligned}$$

Comme on a, d'après l'Inégalité de Cauchy-Schwarz,

$$\begin{aligned}|\langle x_k, u_k\rangle| &\leqslant \|x_k\|\|u_k\| \\ &= \|x_k\| \\ &\to 0,\end{aligned}$$

on en passant à la limite que

$$\langle x, u\rangle \geqslant 0$$

pour tout $x \in \overline{\mathcal{C}}$, et donc en particulier pour tout $x \in \mathcal{C}$. ∎

Remarque 4.2.21. Ce résultat admet une interprétation géométrique intéressante : l'hyperplan (sous-espace de codimension 1) formé de tous les vecteurs orthogonaux à u "sépare" $\mathcal{C}$ de l'origine. En effet, tous les points de $\mathcal{C}$ sont dans le même demi-espace délimité par cet hyperplan, tandis que l'origine appartient, elle, à l'hyperplan lui-même. Il existe de nombreux résultats de séparation de convexes reposant sur les mêmes idées, et qui peuvent s'avérer utiles en optimisation.

Remarque 4.2.22. Indiquons comment ce résultat permet de retrouver le Lemme de Farkas 3.2.4. On vérifie sans peine que l'ensemble

$$\mathcal{C} = \left\{ \sum_{j=1}^{p} \mu_j v_j \mid \mu_j \leqslant 0 \right\} \subset \mathbf{R}^n$$

est convexe, et qu'il en est donc de même pour $\mathcal{C}' = \mathcal{C} - u = \{x - u \mid x \in \mathcal{C}\}$. Alors, si $u \notin \mathcal{C}$, $0 \notin \mathcal{C}'$ donc il existe un vecteur $w \in \mathbf{R}^n$ tel que $\langle y, w\rangle \geqslant 0$ pour tout $y \in \mathcal{C}'$, ou de façon équivalente,

$$\langle x, w\rangle \geqslant \langle u, w\rangle$$

pour tout $x \in \mathcal{C}$. Si donc $\langle x, u\rangle < 0$ dès que $\langle x, v_j\rangle \geqslant 0$ pour tout $1 \leqslant j \leqslant p$, l'inégalité précédente ne peut avoir lieu et $u \in \mathcal{C}$.

Nous sommes maintenant prêts à démontrer la dualité forte pour les problèmes convexes.

Démonstration du Théorème 4.2.19. Commençons par quelques notations pour alléger un peu les formules : nous allons maintenant utiliser les fonctions $\widetilde{f} : x \mapsto f(x) - f(\widetilde{x})$ ainsi que $g = (g_1, \cdots, g_m) : U \to \mathbf{R}^m$ et $h = (h_1, \cdots, h_p) : U \to \mathbf{R}^p$. Nous allons nous intéresser à l'ensemble

$$\mathcal{C} = \{(X, Y, Z) \in \mathbf{R} \times \mathbf{R}^m \times \mathbf{R}^p \mid X > \widetilde{f}(x), Y = g(x), Z \geqslant h(x) \text{ pour un } x \in U\}.$$

Faisons deux observations :

- Parce que f et h sont convexes et que g est affine, l'ensemble $\mathcal{C}$ est lui-même convexe ;
- On a $(0,0,0) \notin \mathcal{C}$. En effet, on aurait sinon un point $x \in U$ tel que $f(x) < f(\widetilde{x})$, $g(x) = 0$ et $h(x) \leqslant 0$, ce qui contredit la minimalité de $\widetilde{x}$ sous contraintes.

D'après le Lemme 4.2.20, il existe donc $u \in \mathbf{R} \times \mathbf{R}^m \times \mathbf{R}^p$ non nul tel que

$$\langle u, (X, Y, Z) \rangle \geqslant 0$$

pour tout $(X, Y, Z) \in \mathcal{C}$. Pour $x \in U$, on a $(\widetilde{f}(x) + 1/N, g(x), h(x)) \in \mathcal{C}$, donc en passant à la limite quand $N \to +\infty$ on conclut que

$$\langle u, (\widetilde{f}(x), g(x), h(x)) \rangle \geqslant 0.$$

Ainsi, en notant $\alpha, \beta_1, \cdots, \beta_m, \gamma_1, \cdots, \gamma_p$ les coefficients de $u \in \mathbf{R}^{1+m+p}$, on a

$$\alpha f(x) + \sum_{i=1}^{m} \beta_i g_i(x) + \sum_{j=1}^{p} \gamma_j h_j(x) \geqslant \alpha f(\widetilde{x}).$$

Supposons pour l'instant $\alpha > 0$ et $\gamma_1, \cdots, \gamma_p \geqslant 0$. Alors, en divisant par α et en posant $\widetilde{\lambda}_i = -\beta_i/\alpha$ et $\widetilde{\mu}_j = -\gamma_j/\alpha$ on a

$$f(x) - \sum_{i=1}^{m} \widetilde{\lambda}_i g_i(x) - \sum_{j=1}^{p} \widetilde{\mu}_j h_j(x) \geqslant f(\widetilde{x}).$$

En prenant la borne inférieure du membre de droite sur x, on en déduit

$$f^*(\widetilde{\lambda}, \widetilde{\mu}) \geqslant f(\widetilde{x}).$$

Comme l'inégalité opposée est vraie par la Proposition 3.4.3, on conclut qu'il y a égalité, ce qui donne la dualité forte. Pour voir que cette valeur coïncide avec celle du Lagrangien, remarquons que comme $\widetilde{\mu}_j \leqslant 0$ et $h_j(x) \leqslant 0$, leur produit est positif et par conséquent leur somme également. Or, prenant $x = \widetilde{x}$ dans l'égalité ci-dessus et en se souvenant que $g_i(\widetilde{x}) = 0$ pour tout $1 \leqslant i \leqslant m$, on voit que

$$\sum_{j=1}^{p} \widetilde{\mu}_j h_j(\widetilde{x}) \leqslant 0$$

et que par conséquent cette quantité est nulle, ce qui donne finalement

$$f(\widetilde{x}) = f(\widetilde{x}) - \sum_{i=1}^{m} \widetilde{\lambda}_i g_i(\widetilde{x}) - \sum_{j=1}^{p} \widetilde{\mu}_j h_j(\widetilde{x}) = \mathcal{L}(\widetilde{x}, \widetilde{\lambda}, \widetilde{\mu}).$$

Néanmoins, il nous reste à vérifier que $\alpha, \gamma_1, \cdots, \gamma_p \geqslant 0$ et $\alpha \neq 0$. Supposons pour commencer $\alpha < 0$. Alors, comme $(\widetilde{f}(x) + N, g(x), h(x)) \in \mathcal{C}$ pour tout $N \in \mathbf{N}$, on a

$$\alpha \widetilde{f}(x) + N\alpha + \sum_{i=1}^{m} \beta_i g_i(x) + \sum_{j=1}^{p} \gamma_j h_j(x) \geqslant 0,$$

ce qui donne une contradiction quand $N \to +\infty$. Ainsi, $\alpha \geqslant 0$ et le même raisonnement montre que $\gamma_j \geqslant 0$ pour tout $1 \leqslant j \leqslant p$. Supposons maintenant que $\alpha = 0$. Par hypothèse, il existe $\widehat{x} \in U$ tel que $g(\widehat{x}) = 0$ et $h(\widehat{x}) < 0$. Or, si $\alpha = 0$ on a

$$\begin{aligned} 0 &\leqslant \alpha f(\widehat{x}) - \alpha f(\widehat{x}) + \sum_{i=1}^{m} \beta_i g_i(\widehat{x}) + \sum_{j=1}^{p} \gamma_j h_j(\widehat{x}) \\ &= 0 - 0 + \sum_{i=1}^{m} 0 + \sum_{j=1}^{p} \gamma_j h_j(\widehat{x}) \\ &= \sum_{j=1}^{p} \gamma_j h_j(\widehat{x}) \end{aligned}$$

Comme $h_j(\widehat{x}) < 0$ et $\gamma_j \geqslant 0$ pour tout $1 \leqslant j \leqslant p$, on doit donc avoir $\gamma_j = 0$. Mais alors, on a pour tout $x \in U$ que

$$\sum_{i=1}^{m} \beta_i g_i(x) \geqslant 0.$$

Mais comme par hypothèse 0 est intérieur à $g(U)$, il existe $t > 0$ tel que $-t\beta \in g(U)$, où $\beta = (\beta_1, \cdots, \beta_m)$. Si x est un antécédent de $t\beta$ par g, alors

$$-t\|\beta\|^2 = \sum_{i=1}^{m} \beta_i g_i(x) \geqslant 0,$$

qui implique $\beta = 0$ et donc $u = 0$, une contradiction. Ainsi, $\alpha > 0$ et la preuve est terminée. ■

Nous avons mentionné à la Remarque 4.2.18 que si les conditions KKT sont toujours suffisantes pour un problème convexe, rien ne garanti qu'elles soient nécessaires puisque les contraintes n'ont pas de raison d'être qualifiées en général. Néanmoins, l'énoncé du Théorème 4.2.19 contient une condition qui ressemble à une qualification, à savoir l'existence du point $\widehat{x}$. Il est donc légitime de se demander si cette condition implique la qualification des contraintes. Il s'avère que c'est le cas, comme nous allons maintenant le montrer.

Proposition 4.2.23. *Soit U un ouvert convexe de $\mathbf{R}^n$ et $f : U \to \mathbf{R}$ une fonction convexe de classe $\mathcal{C}^1$. Soient également $g_1, \cdots, g_m : U \to \mathbf{R}$ des fonctions **affines** et $h_1, \cdots, h_p : U \to \mathbf{R}$ des fonction **convexes** de classe $\mathcal{C}^1$. On suppose qu'il existe $\widehat{x} \in \mathcal{D}$ tel que $h_j(\widehat{x}) < 0$ pour tout $1 \leqslant j \leqslant p$. Alors, les contraintes sont qualifiées en tout point de $\mathcal{D}$.*

Démonstration. Soit $x \in \mathcal{D}$ et $1 \leqslant j \leqslant p$ tel que la contrainte h_j est active au point x, c'est-à-dire que $h_j(x) = 0$. Comme h_j est convexe et de classe $\mathcal{C}^1$, on a par la Proposition 4.2.12 l'inégalité

$$\begin{aligned} h_j(\widehat{x}) &\geqslant h_j(x) + \langle \nabla h_j(x), x - \widehat{x} \rangle \\ &= \langle \nabla h_j(x), x - \widehat{x} \rangle. \end{aligned}$$

Ainsi, en posant $\xi_0 = x - \widehat{x}$, on a bien

$$\langle \nabla h_j(x), \xi_0 \rangle \leqslant h_j(\widehat{x}) < 0.$$

■

Rappelons qu'en général, la dualité faible de la Proposition 3.4.3 nous dit qu'on peut minorer les extrema de f par les maxima de f^*. L'espoir que nous avions évoqué était que f^* soit dans certains cas plus simple à étudier que f. De fait, f^* est en un sens toujours plus facile, parce qu'elle est toujours concave.

Proposition 4.2.24. *Pour tout problème d'optimisation, la fonction duale f^* est concave.*

Démonstration. La fonction f^* est définie comme la borne inférieure des fonctions

$$f_x : (\lambda, \mu) \mapsto f(x) - \sum_{i=1}^{m} \lambda_i g_i(x) - \sum_{j=1}^{p} \mu_j h_j(x).$$

Chacune de ces fonctions, étant linéaire, est en particulier concave (et également convexe, mais cela ne nous est d'aucune utilité ici). Or, une borne inférieure de fonctions concaves est toujours concave. En effet, pour $\lambda, \lambda' \in \mathbf{R}^m$ et $\mu, \mu' \in \mathbf{R}^p_-$ (on remarquera que $\mathbf{R}^p_-$ est convexe) et $t \in [0;1]$, on a

$$\begin{aligned} f^*((\lambda,\mu) + t(\lambda' - \lambda, \mu' - \mu)) &\leqslant f_x((\lambda,\mu) + t(\lambda' - \lambda, \mu' - \mu)) \\ &\leqslant f_x(\lambda,\mu) + t\left(f_x(\lambda',\mu') - f_x(\lambda,\mu)\right) \end{aligned}$$

et en passant à la borne inférieure dans le membre de droite, on obtient le résultat souhaité. ∎

En particulier, maximiser f^* revient à minimiser $-f^*$, et nous savons que si un minimum local existe et que les contraintes d'égalité sont affines et les contraintes d'inégalités convexes, alors il est global.

4.2.4 Une autre caractérisation de la convexité

Nous avons donné dans la propriété 4.2.12 une caractérisation des fonctions convexes de classe $\mathcal{C}^1$ utilisant le gradient. Dans le cas d'une variable, s'il est vrai que les fonctions dérivables qui sont convexes sont celles dont la dérivée est croissante, il est souvent plus facile de vérifier que la dérivée seconde – pour peu qu'elle existe – est positive. Cela est également possible dans le cas de plusieurs variables.

Proposition 4.2.25. *Soit U un ouvert convexe de $\mathbf{R}^n$ et $f : U \to \mathbf{R}$ une fonction de classe $\mathcal{C}^2$. Alors, f est convexe si et seulement si sa matrice Hessienne $H_f(x)$ est* positive *pour tout $x \in U$. De plus, si $H_f(x)$ est* définie positive, *alors f est strictement convexe.*

Remarque 4.2.26. Il n'est pas vrai que si f est strictement convexe, alors $H_f(x)$ est définie positive pour tout x. En effet, la fonction d'une variable $f : x \mapsto x^4$ est strictement convexe, mais $f''(0) = 0$.

Démonstration. Supposons f convexe et prenons $x \in U$. Comme U est ouvert, il existe $r > 0$ tel que $B(x,r) \subset U$ et on considère un élément h tel que $\|h\| < r$, de sorte que $x + th \in B(x,r) \subset U$ pour tout $t \in [0;1]$.

Nous allons avoir besoin d'une variante de la caractérisation donnée par la Propriété 4.2.12. En fait nous allons la "symétriser" : en échangeant les rôles de x et y on obtient

$$f(x) - f(y) \geqslant \langle \nabla f(y), x - y \rangle$$

et en additionnant à l'inégalité originale on trouve

$$\langle \nabla f(y) - \nabla f(x), y - x \rangle \geqslant 0.$$

En appliquant maintenant cette inégalité à $y = x + th$, on voit que

$$\langle \nabla f(x + th) - \nabla f(x), th \rangle \geqslant 0$$

ce qui en divisant par t donne

$$\left\langle \frac{\nabla f(x + th) - \nabla f(x)}{t}, h \right\rangle \geqslant 0.$$

Quand $t \to 0$, le terme de gauche dans le produit scalaire tend, par dérivation des fonctions composées (Proposition 2.3.12), vers $\langle H_f(x)h, h \rangle$, d'où le résultat.

Réciproquement, pour $x, y \in U$, on considère la fonction $g : [0;1] \to \mathbf{R}$ définie par

$$g(t) = f\left(x + t(y - x)\right).$$

On peut alors appliquer l'Égalité de Taylor-Lagrange pour conclure qu'il existe $c \in [0;1]$ tel que

$$g(1) = g(0) + g'(0) + \frac{1}{2} g''(c).$$

Par dérivation des fonctions composées, on a $g'(0) = \langle \nabla f(x), y - x \rangle$ et

$$g''(c) = \left\langle H_f(x + c(y - x))(y - x), y - x \right\rangle \geqslant 0.$$

Par conséquent,

$$\begin{aligned} f(y) &= g(1) \\ &\geqslant g(0) + g'(0) \\ &= f(x) + \langle \nabla f(x), y - x \rangle. \end{aligned}$$

D'après la Proposition 4.2.12, ceci implique que f est convexe. ∎

4.3 Optimisation quadratique

Le carré est un triangle qui a réussi,
ou une circonférence qui a mal tourné.

P. Dac, L'os à Moelle

Nous allons maintenant aborder une famille d'exemples extrêmement importante. Rappelons-nous que dans le Chapitre 3, nous avions considéré le cas "le plus simple" d'optimisation sous contrainte, à savoir le cas où toutes les fonctions sont affines. Ici, nous allons nous pencher sur le cas le plus simple après le cas linéaire, à savoir le cas où la fonction à optimiser est *quadratique*. Commençons par donner un sens à ce terme.

Définition 4.3.1. Une fonction $f : \mathbf{R}^n \to \mathbf{R}$ est dite quadratique s'il existe une matrice $A \in \mathrm{M}_n(\mathbf{R})$, un vecteur $b \in \mathbf{R}^n$ et $c \in \mathbf{R}$ tels que pour tout $x \in \mathbf{R}^n$,

$$f(x) = \langle Ax, x \rangle + \langle b, x \rangle + c.$$

Remarque 4.3.2. La constante c dans l'expression de f ne joue pas de rôle significatif pour nous. En effet, si f a un extremum, alors $f - c$ également, et au même point. Ainsi, pour alléger les notations, nous supposerons généralement $c = 0$ dans la suite.

4.3.1 Convexité

Avant d'aller plus loin, donnons un exemple de première importance : la fonction

$$f : x \mapsto \|x - y\|^2,$$

pour $y \in \mathbf{R}^n$ fixé. Il suffit en effet de développer la norme au carré pour trouver

$$f(x) = \langle x, x\rangle - 2\langle x, y\rangle + \|y\|^2,$$

ce qui donne $A = \mathrm{I}_n$, $b = -2y$ et $c = \|y\|^2$. Nous avons déjà rencontré cette fonction au Paragraphe 1.2 du Chapitre 1 sous une forme un peu différente. En effet, nous cherchions alors à optimiser une fonction de la forme

$$\sum_{i=1}^{\ell} \left\|y_i - x(m)_i\right\|^2,$$

mais en plaçant les vecteurs $x(m)_i$ les uns sous les autres pour former un vecteur $X(m)$ et en procédant de même pour former un vecteur Y, la somme ci-dessus est bien égale à $\|Y - X(m)\|^2$. Ainsi, l'optimisation quadratique contient en particulier ce qu'on appelle les *méthodes de moindre carré*.

Allons maintenant plus loin en essayant d'utiliser les résultats que nous avons pour étudier les problèmes d'optimisation quadratiques. Pour cela, il paraît raisonnable de commencer par calculer le gradient et la Hessienne de la fonction f (étant polynomiale, elle est de classe $\mathcal{C}^\infty$, donc ces objets ont un sens).

Proposition 4.3.3. *Soit f une fonction quadratique donnée par une matrice A et un vecteur b. On a alors*

$$\nabla f(x) = (A + A^t)x + b \quad \& \quad H_f(x) = A + A^t.$$

Démonstration. Il suffit de calculer, pour $x, h \in \mathbf{R}^n$,

$$\begin{aligned} f(x+h) &= \langle Ax + Ah, x + h\rangle + \langle b, x + h\rangle \\ &= \langle Ax, x\rangle + \langle b, x\rangle + \langle Ax, h\rangle + \langle Ah, x\rangle + \langle b, h\rangle + \langle Ah, h\rangle \\ &= f(x) + \langle Ax, h\rangle + \langle h, A^t x\rangle + \langle b, h\rangle + \langle Ah, h\rangle \\ &= f(x) + \langle (A + A^t)x + b, h\rangle + \langle Ah, h\rangle \end{aligned}$$

En observant que l'application

$$h \mapsto \langle (A + A^t)x + b, h\rangle$$

est linéaire tandis que

$$\begin{aligned} |\langle Ah, h\rangle| &\leqslant \|Ah\| \times \|h\| \\ &\leqslant \|A\| \times \|h\|^2 \end{aligned}$$

de sorte que ce dernier terme est négligeable devant $\|h\|$, on conclut que nous avons un développement limité au premier ordre, et que par conséquent $\nabla f(x) = (A + A^t)x + b$.

Le gradient étant linéaire, sa différentielle – qui n'est autre que la Hessienne de f – se calcule aisément :

$$\begin{aligned} f(x+h) - f(x) &= (A + A^t)(x + h) + b - ((A + A^t)x + b) \\ &= (A + A^t)h \end{aligned}$$

donc $H_f(x) = A + A^t$. ∎

La remarque importante à faire sur ce résultat, c'est que les quantités qui nous intéressent ne dépendent pas vraiment de A mais plutôt de $A + A^t$. Or, cette matrice a la propriété intéressante d'être toujours symétrique ! Peut-on donc toujours supposer A symétrique ? La réponse est oui, en remarquant simplement que pour tout $x \in \mathbf{R}^n$,

$$\begin{aligned}\langle Ax, x\rangle &= \frac{1}{2}\langle Ax, x\rangle + \frac{1}{2}\langle Ax, x\rangle \\ &= \frac{1}{2}\langle Ax, x\rangle + \frac{1}{2}\langle x, A^t x\rangle \\ &= \frac{1}{2}\langle Ax, x\rangle + \frac{1}{2}\langle A^t x, x\rangle \\ &= \frac{1}{2}\langle (A + A^t)x, x\rangle.\end{aligned}$$

Partant de ce constat, nous écrirons désormais toujours une fonction quadratique sous la forme

$$f : x \mapsto \frac{1}{2}\langle Ax, x\rangle + \langle b, x\rangle,$$

avec $A \in \mathrm{M}_n(\mathbf{R})$ une matrice *symétrique*. Attention ! Le facteur 1/2 fait que l'on a désormais

$$\boxed{\nabla f(x) = Ax + b \quad \& \quad H_f(x) = A.}$$

En particulier nous pouvons maintenant aisément caractériser la convexité d'une fonction quadratique.

Corollaire 4.3.4. *Soit f une fonction quadratique. Alors,*

- *Elle est convexe si et seulement si A est positive ;*
- *Elle est concave si et seulement si A est négative.*

De plus, si le signe des valeurs propres de A est strict, alors la convexité (respectivement la concavité) de f est stricte également.

Remarque 4.3.5. Si A admet des valeurs propres strictement positives et strictement négatives, il est aisé en considérant les vecteurs propres associés et en s'inspirant de la démonstration de la Proposition 2.3.25, de voir que f n'a pas d'extremum global. Ce sont donc essentiellement les cas convexe et concave qui sont intéressants du point de vue de l'optimisation quand il n'y a pas de contraintes.

Avant d'aller plus loin, écrivons explicitement ce que nos résultats précédents nous disent dans le cadre d'une optimisation sans contrainte. Si A est positive, alors la fonction quadratique f possède un minimum global si et seulement si l'équation $Ax = b$ admet une solution. En effet, $Ax = b$ si et seulement si $\nabla f(x) = 0$, et le Théorème 4.2.13 assure que cette condition est suffisante pour avoir un extremum global. On remarque en particulier que le minimum est unique si A est inversible. Par contre, si A n'est pas surjective il peut ne pas exister, et si A n'est pas injective il y aura soit aucun minimum soit une infinité. Résumons ceci dans le cas le plus favorable.

Proposition 4.3.6. *Soit $A \in \mathrm{M}_n(\mathbf{R})$ une matrice* définie positive, *soit $b \in \mathbf{R}^n$ et soit $c \in \mathbf{R}$. Alors, la fonction quadratique $f : \mathbf{R}^n \to \mathbf{R}$ définie par*

$$f(x) = \frac{1}{2}\langle Ax, x\rangle + \langle b, x\rangle + c$$

admet un unique minimum global au point $\widetilde{x} = -A^{-1}b$, dont la valeur est

$$f(\widetilde{x}) = c - \frac{1}{2}\left\langle b, A^{-1}b\right\rangle.$$

Démonstration. Comme $\nabla f(x) = Ax+b$ et que A est inversible, on voit que $\widetilde{x}$ est l'unique point critique. Par convexité, c'est donc un minimum global, qui est nécessairement unique. La valeur de $f(\widetilde{x})$ se calcule ensuite directement :

$$\begin{aligned} f(\widetilde{x}) &= \frac{1}{2}\langle A\widetilde{x}, \widetilde{x}\rangle + \langle b, \widetilde{x}\rangle + c \\ &= \frac{1}{2}\langle A(-A^{-1}b), -A^{-1}b\rangle + \langle b, -A^{-1}b\rangle + c \\ &= \frac{1}{2}\langle b, A^{-1}b\rangle - \langle b, A^{-1}b\rangle + c \\ &= c - \frac{1}{2}\langle b, A^{-1}b\rangle. \end{aligned}$$

■

4.3.2 La méthode d'Uzawa

Nous avons vu au Paragraphe 3.3 que la forme particulière des fonctions linéaires rendait leur optimisation sous contraintes linéaire abordable à l'aide d'algorithmes spécialement conçus, comme l'algorithme du simplexe. Dans le cas quadratique, il est également possible de donner des algorithmes très efficaces en exploitant la forme particulière de la fonction et nous allons en donner un exemple ci-dessous. La fonction quadratique f est comme toujours donnée par

$$f : x \mapsto \frac{1}{2}\langle Ax, x\rangle + \langle b, x\rangle,$$

avec A une matrice symétrique que nous supposerons *définie positive*. De plus, nous avons vu que pour garantir la convexité de l'ensemble $\mathcal{D}$, il fallait en général supposer les contraintes d'égalité affines. D'un point de vue théorique, de telles contraintes peuvent être enlevées puisqu'elles consistent simplement en une restriction à un sous-espace affine. Par conséquent, nous supposerons pour simplifier qu'il n'y a pas de contraintes d'égalité.

Quant aux contraintes d'inégalité, elles doivent être convexes pour que la convexité de f puisse être utilisée. Mais pour alléger un peu la démonstration de la convergence de la méthode, nous allons nous restreindre au cas de contraintes affines. Ainsi, notre cadre sera le suivant : nous considérerons des vecteurs $w_1, \cdots, w_p \in \mathbf{R}^n$ et les contraintes correspondantes $h_j(x) = \langle w_j, x\rangle - d_j \leqslant 0$. Comme dans le cas linéaire, ces contraintes peuvent être regroupées dans une matrice H pour s'écrire $Hx \leqslant d$, où l'inégalité est comprise coefficient par coefficient. Avec ces notations, les conditions KKT deviennent

$$\left\{\begin{array}{rcl} A\widetilde{x} + b - H^t\mu & = & 0 \\ H\widetilde{x} & \leqslant & d \\ \mu & \leqslant & 0 \\ \mu_j h_j(\widetilde{x}) & = & 0 \end{array}\right.$$

L'idée de la méthode que nous allons présenter est que, comme A est inversible, on peut aussi bien chercher à approcher μ que $\widetilde{x}$. Cela revient en un sens à essayer de résoudre numériquement le problème dual, ce qui est équivalent à la résolution du problème primal grâce à la dualité forte donnée par le Théorème 4.2.19. Nous ne sommes donc pas très avancés, à moins de tenter de jouer les équilibristes en approchant simultanément $\widetilde{x}$ et μ ! En effet, la première équation des conditions KKT nous

donne le moyen de calculer x_k en fonction de μ_k, puisque A est inversible. Mais comment calculer ensuite l'approximation suivante μ_{k+1} de μ en fonction de x_k ? Pour le voir, il nous faut reformuler les conditions de complémentarité. Elles nous disent que pour tout $1 \leqslant j \leqslant p$, le coefficient μ_j est nul dès que le coefficient $(Hx - d)_j$ ne l'est pas. On peut reformuler cette condition au niveau du vecteur $\mu - (H\widetilde{x} - d)$ de la façon suivante :

Lemme 4.3.7. *Les conditions $\mu_j h_j(\widetilde{x}) = 0$ sont équivalentes aux conditions suivantes pour tout $1 \leqslant j \leqslant p$:*

- *Si $(\mu - (H\widetilde{x} - d))_j > 0$, alors $\mu_j = 0$;*
- *Si $(\mu - (H\widetilde{x} - d))_j \leqslant 0$, alors $(H\widetilde{x} - d)_j = 0$.*

Démonstration. Dans le premier cas, on doit avoir $(H\widetilde{x} - d)_j \neq 0$ et donc $\mu_j = 0$. Dans le second cas, si on avait $(H\widetilde{x} - d)_j < 0$, alors on aurait $\mu_j = 0$ et donc $(\mu - (H\widetilde{x} - d))_j > 0$, d'où le résultat. ∎

Telles qu'exprimées ci-dessus, les conditions de complémentarité ne sont pas vraiment plus simples. Mais cette écriture suggère une façon plus compacte de les présenter, grâce à une application que nous allons maintenant introduire.

Définition 4.3.8. Pour un réel x, on pose $x^- = \min(x, 0)$. Si $\mu \in \mathbf{R}^p$, on définit alors $P(\mu) \in \mathbf{R}^p_-$ comme étant le vecteur de coordonnées $(\mu_j^-)_{1 \leqslant j \leqslant p}$.

Le Lemme 4.3.7 peut alors s'exprimer comme suit : les conditions de complémentarité sont équivalentes à

$$\mu = P\left(\mu - (H\widetilde{x} - d)\right)$$

L'application P est appelée *projection sur* $\mathbf{R}^p_-$ (attention, il ne s'agit pas d'une application linéaire) et elle vérifie des propriétés fort sympathiques, dont voici un échantillon.

Proposition 4.3.9. *Pour tout $\mu, \mu' \in \mathbf{R}^p_-$, on a*

$$\langle \mu - P(\mu), \mu' - P(\mu) \rangle \leqslant 0. \tag{4.2}$$

De plus, la projection P est contractante *au sens où*

$$\|P(\mu) - P(\mu')\| \leqslant \|\mu - \mu'\|.$$

Démonstration. Nous allons d'abord montrer que l'inégalité (4.2) est bien vérifiée. Pour cela, il faut calculer

$$\langle \mu - P(\mu), \mu' - P(\mu) \rangle = \sum_{i=1}^{n} (\mu_i - \mu_i^-)(\mu_i' - \mu_i^-)$$

Or, par définition, $\mu_i - \mu_i^- = 0$ si $\mu_i \leqslant 0$ et μ_i sinon. Par conséquent, les seuls termes non nuls dans la somme sont ceux pour lesquels $\mu_i \geqslant 0$, c'est-à-dire $\mu_i^- = 0$. Ainsi, en posant $\mu_i^+ = \mu_i - \mu_i^-$,

$$\begin{aligned} \langle \mu - P(\mu), \mu' - P(\mu) \rangle &= \sum_{i=1}^{n} \mu_i^+ \mu_i' \\ &\leqslant 0. \end{aligned}$$

Montrons maintenant la seconde propriété. On a

$$\begin{aligned}\langle \mu - \mu', P(\mu) - P(\mu')\rangle &= \langle \mu - P(\mu) + P(\mu) - P(\mu') + P(\mu') - \mu', P(\mu) - P(\mu')\rangle \\ &= \langle \mu - P(\mu), P(\mu) - P(\mu')\rangle + \langle P(\mu) - P(\mu'), P(\mu) - P(\mu')\rangle \\ &+ \langle P(\mu') - \mu', P(\mu) - P(\mu')\rangle \\ &= \langle \mu - P(\mu), P(\mu) - P(\mu')\rangle + \|P(\mu) - P(\mu')\|^2 \\ &+ \langle P(\mu') - \mu', P(\mu) - P(\mu')\rangle\end{aligned}$$

Les deux produits scalaires du membre de droite sont négatifs par l'inégalité (4.2), donc

$$\langle \mu - \mu', P(\mu) - P(\mu')\rangle \geqslant \|P(\mu) - P(\mu')\|^2.$$

En utilisant l'Inégalité de Cauchy-Schwarz pour le membre de gauche, on trouve alors

$$\begin{aligned}\|P(\mu) - P(\mu')\|^2 &\leqslant \langle \mu - \mu', P_(\mu) - P(\mu')\rangle \\ &\leqslant \|P(\mu) - P(\mu')\|\|\mu - \mu'\|\end{aligned}$$

et le résultat en découle en divisant par $\|P(\mu) - P(\mu')\|$. ∎

Ceci suggère la méthode suivante pour approcher le minimum $\widetilde{x}$ du problème : on part d'un $\mu_0 \in \mathbf{R}^p_-$ et on pose

$$\left\{\begin{array}{rcl} Ax_k + b & = & H^t\mu_k \\ \mu_{k+1} & = & P\left(\mu_k - (Hx_k - d)\right)\end{array}\right.$$

Le seul inconvénient de cette idée, c'est qu'elle ne fonctionne pas ! En effet, la suite $(x_k)_{k\in\mathbf{N}}$ ci-dessus n'a aucune raison de converger, même pour des problèmes très sympathiques. Le problème vient du fait que la quantité soustraite à μ_k dans la formule définissant μ_{k+1} est trop grande. Mais tout n'est pas perdu ! En effet, rappelons-nous que dans le cas de la Méthode de descente de gradient, on ne suivait pas naïvement le gradient, mais on ajustait la distance parcourue à l'aide d'un *pas* t_k. On peut ici faire exactement de même, ce qui revient à poser

$$\left\{\begin{array}{rcl} Ax_k + b & = & H^t\mu_k \\ \mu_{k+1} & = & P\left(\mu_k - t_k(Hx_k - d)\right)\end{array}\right.$$

L'avantage de cette idée est qu'elle fonctionne bien et converge sous peu d'hypothèses. Il s'agit d'un résultat dû à Uzawa [5], qui a donné son nom à la méthode. Nous allons ici considérer le cas où le pas t est constant.

Théorème 4.3.10 (Convergence de la méthode d'Uzawa) On suppose que les contraintes sont qualifiées en $\widetilde{x}$. Soit $\lambda_0 > 0$ la plus petite valeur propre de A. Alors, pour tout

$$t \in]0; 2\lambda_0/\|H\|^2[,$$

la suite $(x_k)_{k\in\mathbf{N}}$ converge vers $\widetilde{x}$.

5. Hirofumi Uzawa (1928–2014) : économiste japonais, il s'est notamment intéressé à la formalisation des mécanismes d'économie politique. Il a également contribué à l'étude mathématique de l'optimisation linéaire et non-linéaire, avec comme résultat le plus célèbre la preuve de la convergence de l'algorithme qui porte aujourd'hui son nom.

||

Démonstration. Nous allons tout d'abord considérer la suite $\|\mu_k - \mu\|^2$ et montrer qu'elle décroît. Pour cela, on écrit en utilisant le point précédent

$$\begin{aligned}\|\mu_{k+1} - \mu\|^2 &= \|P(\mu_k - t(Hx_k - d)) - P(\mu - t(H\widetilde{x} - d))\|^2 \\ &\leqslant \|\mu_k - t(Hx_k - d) - \mu + t(H\widetilde{x} - d)\|^2 \\ &= \|\mu_k - \mu - tH(x_k - \widetilde{x})\|^2 \\ &= \|\mu_k - \mu\|^2 - 2t\langle \mu_k - \mu, H(x_k - \widetilde{x})\rangle + t^2\|H(x_k - \widetilde{x})\|^2.\end{aligned}$$

On peut facilement majorer le dernier terme en écrivant

$$\|H(x_k - \widetilde{x})\| \leqslant \|H\|\|x_k - \widetilde{x}\|.$$

Quant au second, on a [6]

$$\begin{aligned}\langle \mu_k - \mu, H(x_k - \widetilde{x})\rangle &= \langle H^t\mu_k - H^t\mu, x_k - \widetilde{x}\rangle \\ &= \langle Ax_k + b - (A\widetilde{x} + b), x_k - \widetilde{x}\rangle \\ &= \langle A(x_k - \widetilde{x}), x_k - \widetilde{x}\rangle \\ &\geqslant \lambda_0\|x_k - \widetilde{x}\|^2.\end{aligned}$$

Ainsi,

$$\begin{aligned}\|\mu_{k+1} - \mu\|^2 &\leqslant \|\mu_k - \mu\|^2 - 2t\lambda_0\|x_k - \widetilde{x}\|^2 + t^2\|H\|^2\|x_k - \widetilde{x}\|^2 \\ &= \|\mu_k - \mu\|^2 - \gamma\|x_k - \widetilde{x}\|^2\end{aligned}$$

avec

$$\gamma = t\left(2\lambda_0 - t\|H\|^2\right).$$

Si $0 < t < 2\lambda_0/\|H\|^2$, alors $\gamma > 0$ et on a bien $\|\mu_{k+1} - \mu\| \leqslant \|\mu_k - \mu\|$.

En reprenant la fin des calculs précédents, on obtient l'inégalité

$$0 \leqslant \gamma\|x_k - \widetilde{x}\|^2 \leqslant \|\mu_{k+1} - \mu\|^2 - \|\mu_k - \mu\|^2.$$

Or, la suite $\|\mu_k - \mu\|^2$ est décroissante et comme elle est minorée par 0, elle converge vers une limite ℓ. Mais alors,le membre de droite dans l'inégalité ci-dessus tend vers $\ell - \ell = 0$, donc par encadrement on a

$$\|x_k - \widetilde{x}\| \underset{k\to+\infty}{\longrightarrow} 0,$$

ce qu'il fallait démontrer. ∎

Une implémentation de cette méthode en Python est donnée en appendice au Paragraphe D.3.

Remarque 4.3.11. Le même résultat reste valable si les contraintes $h_1, \cdots, h_p$ sont seulement supposées convexes, mais la preuve est alors plus compliquée et nous avons par préféré nous restreindre à ce cadre plus élémentaire.

6. Comme expliqué au Paragraphe 2.3.4, si toutes les valeurs propres de A sont supérieures à $\lambda_0 \geqslant 0$, alors on a $\langle Ax, x\rangle \geqslant \lambda_0\|x\|^2$ pour tout $x \in \mathbf{R}^n$.

Remarque 4.3.12. On notera que la démonstration implique que la suite $(\mu_k)_{k\in\mathbf{N}}$ est bornée, mais pas qu'elle converge. De fait, cela n'est pas vrai en général. Par contre, si H est surjective, alors la matrice HH^t est inversible et on a donc

$$\mu_k = (HH^t)^{-1}(Ax_k + b),$$

ce qui donne la convergence.

Remarque 4.3.13. Il est intéressant de remarquer que rien ne garantit que x_k satisfasse les contraintes. Comme $\mathcal{D}$ est parfois appelé le *domaine de faisabilité*, cette méthode est dite *infaisable*. Ceci est un inconvénient non négligeable, puisqu'elle fournit une solution approchée que n'est pas forcément utilisable en pratique dans la mesure où elle peut violer certaines contraintes.

4.4 D'un bon pas

Chaque fois que la science avance d'un pas,
c'est qu'un imbécile la pousse, sans faire exprès.

E. Zola, La joie de vivre

En guise de conclusion, nous allons revenir à la méthode d'approximation décrite au début de ce chapitre. Notre intuition est que dans le cas d'une fonction convexe, cette méthode devrait converger vers un minimum de la fonction si les "pas" t_n sont intelligemment choisis. Nous allons maintenant montrer que c'est le cas, d'abord dans le cadre général des fonctions convexes, puis en spécialisant aux problèmes quadratiques.

4.4.1 L'importance d'être constant

Nous allons traiter, en toute généralité, le choix de pas le plus simple : une constante. On parle dans ce cas de *descente de gradient à pas constant*.

Concrètement, on fixe un point $x_0 \in U$, un réel strictement positif $t \in \mathbf{R}_+^*$ et on définit une suite $(x_k)_{k\in\mathbf{N}}$ par récurrence de la façon suivante :

$$x_{k+1} = x_k - t\nabla f(x_k).$$

Nous allons maintenant démontrer que, sous certaines conditions raisonnables sur f, cette suite converge effectivement vers un minimum. Voici un premier résultat dans cette direction.

Proposition 4.4.1. *Soit U un ouvert convexe de $\mathbf{R}^n$ et $f : U \to \mathbf{R}$ une fonction convexe de classe $\mathcal{C}^1$ qui admet un minimum (nécessairement global) en un point $\widetilde{x} \in U$. On suppose qu'il existe $\gamma > 0$ tel que pour tous $x, y \in \mathbf{R}^n$,*

$$\|\nabla f(x) - \nabla f(y)\| \leqslant \gamma\|x - y\|.$$

On suppose de plus qu'il existe $\alpha > 0$ tel que pour tous $x, y \in \mathbf{R}^n$,

$$\langle \nabla f(x) - \nabla f(y), x - y\rangle \geqslant \alpha\|x - y\|^2$$

Alors, si $0 < t < 2\alpha/\gamma^2$, la suite $(x_k)_{k\in\mathbf{N}}$ converge.

Avant de nous lancer dans la preuve, faisons deux commentaires sur les hypothèses de cet énoncé :

- La première hypothèse dit que la fonction ∇f est *lipschitzienne*. Si cette propriété n'a pas de raison d'être vérifiée en général, elle le sera souvent au moins localement. En effet, si f est de classe $\mathcal{C}^2$, alors sa différentielle seconde est bornée sur tout compact. Comme les boules fermées sont convexes, le Théorème des accroissements finis (Théorème 4.2.3) implique que ∇f est lipschitzienne sur $B_f(\widetilde{x}, r)$ pour tout $r > 0$. Il "suffit" donc de prendre le point de départ x_0 dans cette boule et de s'assurer que la suite n'en sort pas.
- La seconde hypothèse est parfois appelée *ellipticité* de f. Comme nous le verrons ci-dessous, elle est en fait une conséquence de la première. Nous avons cependant préféré donner cet énoncé d'abord, parce que la preuve en sera plus simple avec ces deux hypothèses.

Démonstration. Commençons par calculer, en nous rappelant que $\nabla f(\widetilde{x}) = 0$,

$$\begin{aligned} x_{k+1} - \widetilde{x} &= x_k - t\nabla f(x_k) - \widetilde{x} \\ &= x_k - t\nabla f(x_k) - \widetilde{x} + t\nabla f(\widetilde{x}) \\ &= x_k - \widetilde{x} - t(\nabla f(x_k) - \nabla f(\widetilde{x})) \end{aligned}$$

On a alors

$$\begin{aligned} \|x_{k+1} - \widetilde{x}\|^2 &= \|x_k - \widetilde{x}\|^2 + t^2\|\nabla f(x_k) - \nabla f(\widetilde{x})\|^2 \\ &\quad - 2t\langle \nabla f(x_k) - \nabla f(\widetilde{x}), x_k - \widetilde{x}\rangle \\ &\leqslant \|x_k - \widetilde{x}\|^2 + t^2\gamma^2\|x_k - \widetilde{x}\|^2 - 2t\alpha\|x_k - \widetilde{x}\|^2 \\ &= (t^2\gamma^2 - 2t\alpha + 1)\|x_k - \widetilde{x}\|^2 \\ &= \beta(t)\|x_k - \widetilde{x}\|^2. \end{aligned}$$

Il suit par récurrence que pour tout $n \geqslant 1$,

$$\|x_k - \widetilde{x}\| \leqslant \sqrt{\beta(t)}^k \|x_0 - \widetilde{x}\|.$$

Il reste donc simplement à montrer que dans l'intervalle que nous avons choisi pour t, $\beta(t) < 1$. Or $t \mapsto \beta(t)$ est une fonction polynomiale de degré 2, et

$$\beta(0) = 1 = \beta\left(\frac{2\alpha}{\gamma^2}\right),$$

donc $\beta(t) < 1$ sur $]0; 2\alpha/\gamma^2[$. ∎

Remarque 4.4.2. On voit ici en quoi le choix du pas t, est critique et difficile. Si t est trop grand, la convergence de la méthode n'est pas assurée. Mais si t est trop petit, c'est-à-dire proche de 0, alors $\beta(t)$ est proche de 1 et la convergence est alors a priori très lente.

Comme annoncé un peu plus haut, la seconde hypothèse de la Proposition 4.4.1 est en fait redondante, bien que cela n'ait rien d'évident. Néanmoins, si nous le démontrons, alors nous obtiendrons un meilleur énoncé de convergence pour la méthode du gradient à pas constant. Par conséquent, nous allons maintenant énoncer le résultat général, et sa démonstration consistera à montrer que si le gradient de f est lipschitzien, alors f est elliptique.

Théorème 4.4.3 (Convergence de la descente de gradient à pas constant) Soit U un ouvert convexe de $\mathbf{R}^n$ et $f : U \to \mathbf{R}$ une fonction convexe de classe $\mathcal{C}^1$ qui admet un minimum (nécessairement global) en un point $\widetilde{x} \in U$. On suppose qu'il existe $\gamma > 0$ tel que pour tous $x, y \in \mathbf{R}^n$,

$$\|\nabla f(x) - \nabla f(y)\| \leqslant \gamma \|x - y\|.$$

Alors, si $t < 1/\gamma^3$, la suite $(x_k)_{k\in\mathbf{N}}$ converge.

Démonstration. Nous avons vu au cours de la démonstration de la Proposition 4.2.25 que pour une fonction convexe f, on a toujours

$$\langle \nabla f(x) - \nabla f(y), x - y\rangle \geqslant 0.$$

Il s'agit donc de raffiner cette inégalité en exploitant le caractère lipschitzien de ∇f. Dans un premier temps, nous allons montrer que

$$f(z) \leqslant f(y) + \langle \nabla f(y), x - y\rangle + \frac{\gamma}{2}\|x - y\|^2$$

Pour cela, définissons pour $x, y \in U$ une fonction $g : [0;1] \to \mathbf{R}$ par

$$g(t) = g(x + t(y - x)).$$

Alors, g est de classe $\mathcal{C}^1$, donc par le Théorème fondamental de l'analyse,

$$\begin{aligned} g(1) &= g(0) + \int_0^1 g'(s)\mathrm{d}s \\ &= g(0) + g'(0) + \int_0^1 g'(s) - g'(0)\mathrm{d}s. \end{aligned}$$

Par dérivation des fonctions composées, on a $g'(s) = \langle \nabla f(x + s(y - x)), y - x\rangle$, d'où

$$\begin{aligned} g'(s) - g'(0) &= \langle \nabla f(x + s(y - x)) - f(x), y - x\rangle \\ &\leqslant \|\nabla f(x + s(y - x)) - f(x)\|\|y - x\| \\ &\leqslant \gamma\|s(y - x)\|\|y - x\| \\ &= s\gamma\|y - x\|^2. \end{aligned}$$

Ainsi,

$$g(1) \leqslant g(0) + g'(0) + \frac{\gamma}{2}\|y - x\|^2,$$

ce qui en revenant à la définition donne

$$f(y) \leqslant f(x) + \langle \nabla f(x), y - x\rangle + \frac{\gamma}{2}\|y - x\|^2.$$

En appliquant l'inégalité précédente à $y = x$ et $z = x - \nabla f(x)/\gamma$, on obtient

$$\begin{aligned} f\left(x - \frac{\nabla f(x)}{\gamma}\right) &\leqslant f(x) + \left\langle \nabla f(x), -\frac{\nabla f(x)}{\gamma}\right\rangle + \frac{1}{2\gamma}\|\nabla f(x)\|^2 \\ &= f(x) - \frac{1}{2\gamma}\|\nabla f(x)\|^2. \end{aligned}$$

Le membre de gauche est plus grand que $f(\widetilde{x})$ par définition d'un minimum global, donc on a finalement

$$f(\widetilde{x}) \leqslant f(x) - \frac{1}{2\gamma}\|\nabla f(x)\|^2.$$

Pour conclure, nous allons maintenant recourir à une astuce. Les points x et y étant fixés, on définit deux fonctions $f_1, f_2 : U \to \mathbf{R}$ par

$$\begin{cases} f_1(z) &= f(z) - \langle \nabla f(x), z\rangle \\ f_2(z) &= f(z) - \langle \nabla f(y), z\rangle \end{cases}$$

Ces deux fonctions sont convexes – en effet,

$$\begin{aligned} \langle \nabla f_1(z), z' - z\rangle &= \langle \nabla f(z), z' - z\rangle - \langle \nabla f(x), z' - z\rangle \\ &\leqslant \langle f(z') - f(z) - \langle \nabla f(x), z' - z\rangle \\ &= f_1(z') - f_1(z) \end{aligned}$$

ce qui suffit par la Proposition 4.2.12, et de même pour f_2 – et ont un minimum global en x et y respectivement, d'où

$$\begin{cases} f_1(x) \leqslant f_1(y) - \dfrac{1}{2\gamma}\|\nabla f_1(y)\|^2 \\ f_2(y) \leqslant f_2(x) - \dfrac{1}{2\gamma}\|\nabla f_2(x)\|^2 \end{cases}$$

qui s'écrit aussi

$$\begin{cases} f(x) - f(y) \leqslant \langle \nabla f(x), x\rangle - \langle \nabla f(x), y\rangle - \dfrac{1}{2\gamma}\|\nabla f(y) - \nabla f(x)\|^2 \\ f(y) - f(x) \leqslant \langle \nabla f(y), y\rangle - \langle \nabla f(y), x\rangle - \dfrac{1}{2\gamma}\|\nabla f(x) - \nabla f(y)\|^2 \end{cases}$$

En sommant ces deux inégalités on obtient le résultat dont nous avons besoin, à savoir

$$\langle \nabla f(x) - \nabla f(y), x - y\rangle \geqslant \frac{1}{2\gamma}\|x - y\|^2.$$

On conclut simplement en remplaçant α par $1/2\gamma$ dans l'énoncé de la Proposition 4.4.1. ∎

Une implémentation de cette méthode en Python est donnée en appendice au Paragraphe D.4.1.

Remarque 4.4.4. La preuve donne même un peu plus : s'il existe $0 < a < b < 1/\gamma^3$ tels que pour tout $k \in \mathbf{N}$, $t_k \in [a; b]$, alors la suite $(x_k)_{k\in\mathbf{N}}$ converge vers $\widetilde{x}$.

4.4.2 Optimiser l'optimisation

On pourra objecter à ce qui précède que le choix d'un pas constant est une solution de facilité. En effet, à chaque étape, parmi tous les pas possibles, certains sont "meilleurs" que d'autres au sens où pour t_k, t_k' on peut avoir

$$f(x_k - t_k\nabla f(x_k)) < f(x_k + t_k'\nabla f(x_k)).$$

Il serait donc probablement plus efficace de choisir, à chaque étape un pas qui donne la valeur de f la plus petite possible : un *pas optimal* ! Des ennuis se profilent néanmoins à l'horizon : comment savoir si un tel pas optimal existe et surtout comment le trouver ? Si l'on a recours à un algorithme, alors on multiplie les complexité des deux méthodes, ce qui risque de largement compenser le gain obtenu par le choix plus judicieux du pas.

Nous allons ici nous concentrer sur un cas pour lequel cette méthode est efficace, et c'est celui des fonctions quadratiques. Pourquoi ? Parce qu'alors, on sait résoudre explicitement le problème de minimisation à chaque pas et qu'il n'y a donc pas besoin de combiner avec une seconde méthode. Plus concrètement, nous allons nous concentrer sur une fonction $f : \mathbf{R}^n \to \mathbf{R}$ définie par

$$f(x) = \frac{1}{2}\langle Ax, x\rangle + \langle b, x\rangle$$

et, pour assurer l'existence d'un minimum, nous supposerons A définie positive. Une fois fixé un point de départ $x_0 \in \mathbf{R}^n$, nous allons chercher un pas optimal, c'est-à-dire une valeur de t qui minimise la quantité

$$f(x_k - t\nabla f(x_k)).$$

La première chose à faire est de s'assurer que cette valeur existe. Bien sûr, si $\nabla f(x_k) = 0$, alors on a trouvé le minimum et on peut arrêter l'itération. Nous allons donc supposer que ce n'est pas le cas.

Lemme 4.4.5. *On suppose que $\nabla f(x_k) \neq 0$. Alors, il existe une valeur optimale t_k de t, qui est donnée par*

$$t_k = \frac{\|\nabla f(x_k)\|^2}{\langle A\nabla f(x_k), \nabla f(x_k)\rangle}.$$

Démonstration. Considérons la fonction $g : \mathbf{R} \to \mathbf{R}$ définie par

$$g(t) = f(x_k - t\nabla f(x_k)).$$

Elle est de classe $\mathcal{C}^\infty$, et sa dérivée est donnée par

$$\begin{aligned} g'(t) &= -\langle \nabla f(x_k - t\nabla f(x_k)), \nabla f(x_k)\rangle \\ &= -\langle A(x_k - t\nabla f(x_k)) + b, Ax_k + b\rangle \\ &= -\|Ax_k + b\|^2 + t\langle A\nabla f(x_k), Ax_k + b\rangle \\ &= t\langle A\nabla f(x_k), \nabla f(x_k)\rangle - \|\nabla f(x_k)\|^2. \end{aligned}$$

On voit qu'il y a un unique point critique qui est donné par la formule de l'énoncé. Comme d'autre part g est une fonction polynomiale du second degré, ce point critique est un extremum global. Enfin, $g(t) \to +\infty$ quand $t \to +\infty$, donc le point critique doit être un minimum global, ce qui conclut la preuve. ∎

Nous pouvons maintenant définir rigoureusement notre méthode itérative : avec les notations précédentes, on pose simplement

$$x_{k+1} = x_k - \frac{\|\nabla f(x_k)\|^2}{\langle A\nabla f(x_k), \nabla f(x_k)\rangle}(Ax_k + b).$$

Ce qui est remarquable c'est que, contrairement au cas du gradient à pas constant, la convergence n'est pas seulement locale mais valable pour toute valeur initiale x_0. La

démonstration sera plus claire si nous introduisons d'abord quelques notations. Étant donnée une matrice définie positive A, on pose [7]

$$\|x\|_A = \langle Ax, x\rangle.$$

Comme nous allons le voir, il est assez naturel de travailler avec cette quantité, mais il faut alors être capable de la comparer à la norme usuelle [8]. En remarquant que A^{-1} est également définie positive, et donc que $\|x\|_{A^{-1}}$ a un sens, on peu obtenir l'inégalité dont nous aurons besoin, à savoir

Lemme 4.4.6. *Soient λ_1 et λ_n respectivement la plus petite et la plus grande valeur propre de A. Alors, pour tout $x \in \mathbf{R}^n$,*

$$\frac{\|x\|^4}{\|x\|_A^2\|x\|_{A^{-1}}^2} \geqslant \frac{\lambda_1}{\lambda_n}.$$

Démonstration. Soit $(v_1, \cdots, v_n)$ une base orthonormée de vecteurs propres de A. Si $x \in \mathbf{R}^n$, il se décompose sur la base précédente de la façon suivante :

$$x = \sum_{i=1}^{n} x_i v_i.$$

On a par conséquent

$$\begin{aligned}
\|x\|_A^2 &= \sum_{i=1}^{n} \lambda_i x_i^2 \\
&\leqslant \sum_{i=1}^{n} \lambda_n x_i^2 \\
&= \lambda_n \sum_{i=1}^{n} x_i^2 \\
&= \lambda_n \|x\|^2.
\end{aligned}$$

Notons que $(v_1, \cdots, v_n)$ est également une base orthonormée de vecteurs propres pour A^{-1} et que sa plus grande valeur propre est λ_1^{-1}. Donc, le même calcul donne $\|x\|_{A^{-1}} \leqslant \lambda_1^{-1}\|x\|^2$, de sorte que finalement

$$\begin{aligned}
\frac{\|x\|^4}{\|x\|_A^2\|x\|_{A^{-1}}^2} &= \frac{\|x\|^2}{\|x\|_A^2}\frac{\|x\|^2}{\|x\|_{A^{-1}}^2} \\
&\geqslant \frac{\|x\|^2}{\lambda_n\|x\|^2}\frac{\|x\|^2}{\lambda_1^{-1}\|x\|^2} \\
&= \frac{\lambda_1}{\lambda_n}.
\end{aligned}$$

■

7. On peut sans peine vérifier que la fonction $(x, y) \mapsto \langle Ax, y\rangle$ est symétrique et définie positive sur $\mathbf{R}^n$, et que par conséquent elle donne bien lieu à une norme comme le suggère la notation.

8. Rappelons que toutes les normes étant équivalentes sur $\mathbf{R}^n$, il est de fait possible de comparer $\|x\|_A$ à $\|x\|$.

Remarque 4.4.7. En se rappelant que la norme d'une matrice symétrique positive est égale à sa plus grande valeur propre, on voit que la constante apparaissant dans l'énoncé est $\|A\|\|A^{-1}\|$. Cette quantité est appelé le *conditionnement* de la matrice A et il s'agit généralement du paramètre critique pour la vitesse de convergence des algorithmes d'optimisation. Une matrice est dite *bien conditionnée* si son conditionnement est proche de 1.

Nous pouvons maintenant montrer la convergence de la méthode.

Théorème 4.4.8 On suppose A définie positive. Alors, la méthode de gradient à pas optimal converge quel que soit le point de départ x_0.

Démonstration. En utilisant la notation $\|\cdot\|_A$, le pas optimal peut s'écrire

$$t_k = \frac{\|\nabla f(x_k)\|^2}{\|\nabla f(x_k)\|_A^2}.$$

L'idée est de chercher à évaluer $\|x_k - \widetilde{x}\|_A$ plutôt que la distance en norme usuelle. En effet, on a

$$\begin{aligned}\|x_{k+1} - \widetilde{x}\|_A^2 &= \langle A(x_{k+1} - \widetilde{x}), x_{k+1} - \widetilde{x}\rangle \\ &= \langle A(x_k - t_k\nabla f(x_k)) - A\widetilde{x}, x_k - t_k\nabla f(x_k) - \widetilde{x}\rangle \\ &= \langle A(x_k - \widetilde{x}), x_k - \widetilde{x}\rangle - 2t_k\langle A(x_k - \widetilde{x}), \nabla f(x_k)\rangle + t_k^2\langle A\nabla f(x_k), \nabla f(x_k)\rangle\end{aligned}$$

En remarquant que $A\widetilde{x} = -b$ puisque $\nabla f(\widetilde{x}) = 0$, on peut poursuivre le calcul

$$\begin{aligned}\|x_{k+1} - \widetilde{x}\|_A^2 &= \|x_k - \widetilde{x}\|_A^2 - 2t_k\|\nabla f(x_k)\|^2 + t_k^2\|\nabla f(x_k)\|_A^2 \\ &= \|x_k - \widetilde{x}\|_A^2 - 2\frac{\|\nabla f(x_k)\|^4}{\|\nabla f(x_k)\|_A^2} + \frac{\|\nabla f(x_k)\|^4}{\|\nabla f(x_k)\|_A^2} \\ &= \|x_k - \widetilde{x}\|_A^2 - \frac{\|\nabla f(x_k)\|^4}{\|\nabla f(x_k)\|_A^2}.\end{aligned}$$

Afin d'aller plus loin, nous allons insérer une troisième norme : celle associée à A^{-1}. En effet, on a

$$\begin{aligned}\|x_k - \widetilde{x}\|^2 &= \langle A(x_k - \widetilde{x}), x_k - \widetilde{x}\rangle \\ &= \langle A(x_k - \widetilde{x}), A^{-1}A(x_k - \widetilde{x})\rangle \\ &= \|A(x_k - \widetilde{x})\|_{A^{-1}}^2 \\ &= \|\nabla f(x_k)\|_{A^{-1}}^2.\end{aligned}$$

Ainsi,

$$\begin{aligned}\|x_{k+1} - \widetilde{x}\|_A^2 &= \|x_k - \widetilde{x}\|_A^2 - \frac{\|\nabla f(x_k)\|^4}{\|\nabla f(x_k)\|_A^2\|\nabla f(x_k)\|_{A^{-1}}^2}\|x_k - \widetilde{x}\|_A^2 \\ &= \|x_k - \widetilde{x}\|_A^2\left(1 - \frac{\|\nabla f(x_k)\|^4}{\|\nabla f(x_k)\|_A^2\|\nabla f(x_k)\|_{A^{-1}}^2}\right).\end{aligned}$$

Il ne reste plus pour conclure qu'à appliquer le Lemme 4.4.6 pour obtenir

$$\|x_{k+1} - \widetilde{x}\|_A^2 \leqslant \|x_k - \widetilde{x}\|_A^2\left(1 - \frac{\lambda_1}{\lambda_n}\right).$$

Comme on a par une récurrence immédiate

$$\|x_{k+1} - \widetilde{x}\|_A^2 \leqslant \|x_0 - \widetilde{x}\|_A^2 \left(1 - \frac{\lambda_1}{\lambda_n}\right)^k,$$

on conclut que $\|x_k - \widetilde{x}\|_A \to 0$.

Nous n'avons pas tout à fait terminé, puisque nous n'avons pas montré que

$$\|x_k - x\| \to 0.$$

Néanmoins, ceci se fait sans difficulté. En effet, nous savons que pour tout $x \in \mathbf{R}^n$, $\|x\|_A \geqslant \lambda_1 \|x\|$, et donc

$$\|x_k - \widetilde{x}\| \leqslant \frac{1}{\lambda_1} \|x_k - x\|_A \to 0.$$

■

On remarquera que si A est un multiple de l'identité (c'est-à-dire si $\lambda_1 = \lambda_n$) alors le taux de convergence est nul. Autrement dit, l'algorithme converge en une itération. Cela n'est pas surprenant ! En effet, si $A = \lambda \mathrm{I}_n$, alors $t_k = \lambda^{-1}$ donc

$$x_1 = x_0 - \lambda^{-1}(Ax_0 + b) = -\lambda^{-1} b,$$

qui est bien l'unique solution de l'équation $Ax = -b$.

En général, on voit que le taux de convergence est d'autant meilleur que la quantité λ_1/λ_n est proche de 1, c'est-à-dire d'autant meilleur que A est bien conditionnée. Une implémentation de cette méthode en Python est donnée en appendice au Paragraphe D.4.2.

ANNEXE A

UN THÉORÈME DE FUBINI "FACILE"

Nam consuetudo et exercitatio facilitates maxime parit [1]

M. F. Quintilianus, De institutione oratoria

Nous avons utilisé, dans la démonstration du Théorème 2.3.21, une propriété permettant d'intervertir des intégrales. Si le Théorème de Fubini permet bien sûr de justifier cette interversion, il existe en fait une justification beaucoup plus simple. Cela est dû au fait que la fonction considérée est continue et que les intégrales portent sur des segments. Nous allons donc énoncer et démontrer un résultat d'interversion qui suffit à nos besoins et ne nécessite que des propriétés classiques de l'intégrale de Riemann.

Soit $f : U \to \mathbf{R}$ une fonction continue et soit $R = [a;b] \times [c;d]$ un rectangle contenu dans U. Avant de démontrer le résultat, nous allons vérifier que l'égalité a bien un sens, c'est-à-dire que chaque intégrande est bien intégrable au sens de Riemann. Cela sera facile car elles sont tout simplement continues.

Lemme A.1. *Les fonctions*

$$\mathcal{F}_x : y \mapsto \int_a^b f(x,y)\mathrm{d}x$$
$$\mathcal{F}_y : x \mapsto \int_c^d f(x,y)\mathrm{d}y$$

sont continues, donc intégrables.

Démonstration. Nous ne traiterons que la première fonction, la preuve pour la seconde est la même, *mutatis mutandis*. Soit $\epsilon > 0$. La fonction f étant continue sur le rectangle $[a;b] \times [c;d]$ qui est compact, elle est uniformément continue. Il existe donc $\eta > 0$ tel que dès que $\|(x,y) - (x',y')\| < \eta$, on a

$$|f(x,y) - f(x',y')| < \epsilon/(b-a).$$

1. C'est surtout par l'habitude et l'exercice qu'on acquiert la facilité

Alors, étant donné $y_0 \in [c;d]$, pour tout $y \in]y_0-\eta; y_0+\eta[\cap [c;d]$ on a

$$\begin{aligned}|\mathcal{F}_x(y) - \mathcal{F}_x(y_0)| &\leqslant \int_a^b |f(x,y) - f(x,y_0)| \mathrm{d}x \\ &\leqslant \int_a^b \frac{\epsilon}{b-a} \mathrm{d}x \\ &= \epsilon,\end{aligned}$$

ce qu'il fallait démontrer. ∎

Nous sommes maintenant en mesure de démontrer le résultat utilisé dans la démonstration du Théorème 2.3.21 de façon – comme promis – élémentaire.

Théorème A.2 On a l'égalité

$$\int_a^b \left(\int_c^d f(x,y) \mathrm{d}y \right) \mathrm{d}x = \int_c^d \left(\int_a^b f(x,y) \mathrm{d}x \right) \mathrm{d}y.$$

Démonstration. Le Lemme A.1 assure que les deux termes de l'égalité sont bien définis. De plus, on a, en utilisant une somme de Riemann et la linéarité de l'intégrale,

$$\begin{aligned}\int_a^b \left(\int_c^d f(x,y) \mathrm{d}y \right) \mathrm{d}x &= \int_a^b \mathcal{F}_y \mathrm{d}x \\ &= \lim_{n\to+\infty} \frac{1}{n+1} \sum_{k=0}^{n} \mathcal{F}_y \left(a + k\frac{b-a}{n} \right) \\ &= \lim_{n\to+\infty} \frac{1}{n+1} \sum_{k=0}^{n} \int_c^d f\left(a + k\frac{b-a}{n}, y \right) \mathrm{d}y \\ &= \lim_{n\to+\infty} \int_c^d \frac{1}{n+1} \sum_{k=0}^{n} f\left(a + k\frac{b-a}{n}, y \right) \mathrm{d}y.\end{aligned}$$

Posons, pour $n \in \mathbf{N}$,

$$g_n(y) = \frac{1}{n} \sum_{k=1}^{n} f\left(a + k\frac{b-a}{n}, y \right).$$

Nous allons montrer que la suite $(g_n)_{n\in\mathbf{N}}$ converge uniformément vers $\mathcal{F}_x$, ce qui permettra de conclure.

Soit $\epsilon > 0$. En posant, pour $0 \leqslant k \leqslant n$,

$$a_k = a + k\frac{b-a}{n}$$

et utilisant la Relation de Chasles, on a

$$\begin{aligned}\left| g_n(y) - \mathcal{F}_x(y) \right| &= \left| \frac{1}{n+1} \sum_{k=0}^{n-1} \int_{a_k}^{a_{k+1}} f\left(a + k\frac{b-a}{n}, y \right) - f(x,y) \, \mathrm{d}y \right| \\ &\leqslant \frac{1}{n+1} \sum_{k=0}^{n-1} \int_{a_k}^{a_{k+1}} \left| f\left(a + k\frac{b-a}{n}, y \right) - f(x,y) \right| \mathrm{d}y\end{aligned}$$

Comme f est continue sur le rectangle R qui est compact, elle est *uniformément continue* par le Théorème de Heine. Par conséquent, il existe $\eta > 0$ tel que $\|(x,y)-(x',y')\| < \eta$ implique

$$|f(x,y)-f(x',y')| < \epsilon.$$

Soit N tel que $(b-a)/N < \eta$. Alors, dans chaque terme de la somme, l'intégrande est majorée par ϵ. Par conséquent, pour tout $n \geqslant N$,

$$\begin{aligned}\left|g_n(y)-\mathcal{F}_x(y)\right| &\leqslant \frac{1}{n+1}\sum_{k=0}^{n-1}\int_{a_k}^{a_{k+1}} \epsilon \mathrm{d}y \\ &= \epsilon\frac{n}{n+1} \\ &< \epsilon.\end{aligned}$$

Ceci montre que $\|g_n-\mathcal{F}_x\|_\infty < \epsilon$ pour tout $n \geqslant N$, donc que la suite converge uniformément. Le Théorème de convergence uniforme pour l'intégrale de Riemann donne alors le résultat. ■

ANNEXE B

L'ENVERS DE L'ENDROIT : LES THÉORÈMES D'INVERSION

La mémoire, c'est l'imagination à l'envers.

D. Pennac, La fée carabine

Nous avons mentionné au Paragraphe 3.1.2 du Chapitre 3 que le Théorème des fonctions implicites 3.1.6 possède un proche parent connu sous le nom de Théorème d'inversion locale. Nous allons maintenant discuter un peu plus en détails ce dernier, ainsi que son inséparable comparse le Théorème d'inversion globale.

Le problème de départ est assez naturel : existe-t-il un moyen simple de s'assurer qu'une application $f : U \to \mathbf{R}^m$ de classe $\mathcal{C}^k$ est une bijection ? Et si possible une bijection dont l'inverse est également de classe $\mathcal{C}^k$? Suivant un procédé éprouvé des mathématiques, il est sage de commencer par aborder le problème localement. Autrement dit, étant donné $x \in U$, à quelle condition existe-t-il un ouvert $V \subset U$ contenant x et un ouvert W dans $\mathbf{R}^m$ contenant $f(x)$ tel que f soit une bijection de V sur W dont la réciproque est de classe $\mathcal{C}^k$? Avant d'aller plus loin remarquons l'hypothèse que W soit un ouvert. Cela est indispensable si on veut que la réciproque soit au moins continue. En effet, W sera alors la préimage par f^{-1} de V, qui est ouvert !

Maintenant, commençons par chercher une condition nécessaire. Si nous cherchons dans notre mémoire (ou dans le début de ce texte), nous voyons que pour calculer la différentielle d'une fonction réciproque f^{-1}, il nous faut supposer que la différentielle de f au point correspondant est inversible. Cela dit, ceci pourrait n'être qu'une condition suffisante pratique pour les calculs. Vérifions qu'il n'en est rien, en commençant par introduire une terminologie bien pratique.

Définition B.1. Une application $f : U \to \mathbf{R}^m$ est un *$\mathcal{C}^k$-difféomorphisme* si elle est de classe $\mathcal{C}^k$, bijective, et que sa réciproque f^{-1} est également de classe $\mathcal{C}^k$.

Nous pouvons maintenant démontrer notre condition nécessaire, qui au passage nous donnera également une information sur la dimension de l'espace d'arrivée. Il est fructueux d'accorder un instant d'attention à sa proximité – mais surtout ses différences – avec le Corollaire **??**.

Proposition B.2. *Soit $f : U \to \mathbf{R}^m$ un $\mathcal{C}^k$-difféomorphisme. Alors, pour tout $x \in U$, l'application linéaire $D_x f$ est inversible, d'inverse $D_{f(x)}(f^{-1})$. De plus, on doit avoir $m = n$.*

Démonstration. On sait que $(f^{-1}) \circ f = \mathrm{Id}$, donc en différentiant au point x on trouve

$$\mathrm{Id} = D_x\mathrm{Id} = D_{f(x)}(f^{-1}) \circ D_x f.$$

Ainsi, puisqu'on est en dimension finie, $D_x f$ est inversible d'inverse $D_{f(x)}(f^{-1})$. Par ailleurs, puisque l'application $D_x f : \mathbf{R}^n \to \mathbf{R}^m$ est inversible, ses espaces de départ et d'arrivée doivent avoir la même dimension, autrement dit $n = m$. ∎

Nous arrivons au moment où naît en nous un espoir : la condition nécessaire pourrait-elle être suffisante ? Pour avoir une meilleur intuition, considérons le cas d'une seule variable. On a alors une fonction $f : I \to \mathbf{R}$ et un point $x \in I$ tel que $f'(x) \neq 0$. Si f est de classe $\mathcal{C}^1$, la continuité de f' assure l'existence d'un intervalle $J \subset I$ sur lequel f' est de signe constant strict. Alors, f est strictement monotone sur J, donc inversible ! De plus, le résultat classique sur les fonction bijectives dérivables dont la dérivée ne s'annule pas assure que f^{-1} sera dérivable, et même de classe $\mathcal{C}^1$. Nous voilà convaincus que notre condition nécessaire est suffisante. Le démontrer dans le cas de plusieurs variables n'est cependant pas chose aisée ... à moins d'utiliser habilement le Théorème des fonctions implicites !

Théorème B.3 (Théorème d'inversion locale) Soit $f : U \to \mathbf{R}^n$ une fonction de classe $\mathcal{C}^k$ et soit $x \in U$ tel que $D_x f$ est inversible. Alors, il existe un ouvert $V \subset U$ contenant x et un ouvert $W \subset \mathbf{R}^m$ contenant $f(x)$ tels que $f(V) = W$ et la restriction de f à V est un $\mathcal{C}^k$-difféomorphisme.

Démonstration. Attention, astuce droit devant ! Considérons les applications

$$g_i : \mathbf{R}^{2n} \to \mathbf{R}$$

définies par

$$g_i(z_1, \cdots, z_{2n}) = z_i - f_i(z_{n+1}, \cdots, z_{2n}).$$

pour $1 \leqslant i \leqslant n$. En posant $\widetilde{x} = (f(x), x)$, on observe que $g_i(\widetilde{x}) = 0$. De plus, les n dernières coordonnées de ∇g_i ne sont autres que les dérivées partielles de f_i, tands que les n premières sont les colonnes de la matrice identité. Par conséquent, la matrice dont les colonnes sont les ∇g_i est :

$$J_g(\widetilde{x}) = \begin{pmatrix} \mathrm{I}_n & 0 \\ 0 & J_f(\widetilde{x}) \end{pmatrix}.$$

Ainsi, comme $J_f(x)$ est inversible, la famille de vecteurs

$$\nabla g_{n+1}(\widetilde{x}), \cdots, \nabla g_{2n}(\widetilde{x})$$

est libres. On peut donc appliquer le Théorème des fonctions implicites : il existe un ouvert W contenant $f(x)$, un ouvert V contenant x et une application $\varphi : W \to V$ de classe $\mathcal{C}^k$ telle que pour tous $(z, z') \in W \times V$ et tout $1 \leqslant i \leqslant n$,

$$g_i(z, z') = 0 \iff z' = \varphi(z).$$

L'équation de gauche est équivalente à $z = f(z')$, donc φ est une réciproque de f, ce qui démontre le théorème. ∎

Nous avons obtenu ci-dessus une solution "locale" à notre problème, qui est de plus optimale au sens où nous avons une équivalence : f est un difféomorphisme au voisinage de x si et seulement si $D_x f$ est inversible. L'étape suivante est de s'interroger

sur la possibilité d'un critère global. Pour cela, il est sain de chercher en premier lieu ce qui peut faillir dans la caractérisation précédente. Autrement dit : à quoi ressemble une fonction qui est une bijection au voisinage de tout point mais qui n'est pas une bijection globale ? Il est en fait assez simple de donner un exemple.

Exemple B.4. On considère $U =]0;1[\cup]2;3[$ qui est un ouvert de $\mathbf{R}$ et la fonction $f : U \to \mathbf{R}$ définie par

$$f(x) = \begin{cases} x & \text{si} \quad x \in]0;1[\\ x-2 & \text{si} \quad x \in]2;3[\end{cases}$$

Cette fonction est de classe $\mathcal{C}^\infty$ et sa différentielle en tout point est l'application linéaire identité de $\mathbf{R}$ dans $\mathbf{R}$, qui est inversible. Pourtant, f n'est pas un difféomorphisme, puisqu'elle n'est pas injective : $f(1/2) = f(3/2)$.

Pour se débarrasser du contre-exemple précédent, une solution consiste à ajouter l'hypothèse d'injectivité. Après tout, elle est nécessaire pour avoir un difféomorphisme. Il s'avère que cela suffit à conclure.

> **Théorème B.5** (Théorème d'inversion globale) Soit $f : U \to \mathbf{R}^n$ une application de classe $\mathcal{C}^k$. Si f est injective, et si pour tout $x \in U$, $D_x f$ est inversible. Alors $f(U)$ est un ouvert et $f : U \to f(U)$ est un $\mathcal{C}^k$-difféomorphisme.

Démonstration. Comme le suggère l'énoncé, nous allons commencer par montrer que $f(U)$ est ouvert. Soit donc $y \in f(U)$ et $x \in U$ tel que $f(x) = y$. D'après le Théorème B.3, il existe un ouvert $V \subset U$ contenant x et un ouvert W contenant $f(x)$ tel que la restriction de $f : V \to W$ soit un $\mathcal{C}^k$-difféomorphisme. En particulier, $W = f(V) \subset f(U)$ donc $f(U)$ est ouvert.

Comme f est supposée injective, l'application $f : U \to f(U)$ est bijective. De plus, si $y = f(x) \in f(U)$ et si les ouverts V et W sont comme précédemment, alors la restriction de f^{-1} à W coïncide avec l'application donnée par le Théorème B.3, qui est de classe $\mathcal{C}^k$. Ainsi, f^{-1} est de classe $\mathcal{C}^k$ au voisinage de chaque point, donc elle est bien globalement de classe $\mathcal{C}^k$, ce qui conclut la preuve. ∎

Remarque B.0.1. Nous venons de répondre à une question laissée en suspend à la Proposition 2.3.13 : si f est bijective de classe $\mathcal{C}^1$ avec une différentielle inversible, alors sa réciproque est aussi de classe $\mathcal{C}^1$, sans avoir besoin de supposer qu'elle est continue. Ce qui est caché derrière ce résultat, c'est la preuve de la continuité de la fonction implicite dans la démonstration de la Proposition 3.1.4. Ce n'est donc pas du tout trivial !

Concluons par un exemple d'application de ce résultat. On définit une fonction $\phi :]-\pi;\pi[\times \mathbf{R}_+^* \to \mathbf{R}^2$ par

$$\phi(\theta, r) = (r\cos(\theta), r\sin(\theta)).$$

Si $f : \mathbf{R}^2 \to \mathbf{R}$ est une fonction, alors $f \circ \phi$ est une expression de f *en coordonnées polaires*. Autrement dit, ϕ permet de faire un changement de variable des coordonnées cartésiennes aux coordonnées polaires. Ce que nous allons montrer, c'est que ce changement de variables est réversible. Pour cela, il faut montrer deux choses :

- **Injectivité.** Soient $\theta_1, \theta_2 \in]-\pi;\pi[$ et $r_1, r_2 \in \mathbf{R}_+^*$ tels que $\phi(\theta_1, r_1) = \phi(\theta_2, r_2)$. En remarquant que $\|\phi(\theta, r)\| = r$, on voit que $r_1 = r_2$. On en déduit alors en considérant les coordonnées que $\cos(\theta_1) = \cos(\theta_2)$ et $\sin(\theta_1) = \sin(\theta_2)$. Cette égalité peut s'écrire plus synthétiquement sous la forme $e^{i\theta_1} = e^{i\theta_2}$. Or, la fonction $t \mapsto e^{it}$ est injective sur $]-\pi;\pi[$, donc $\theta_1 = \theta_2$ et nous avons bien montré que ϕ est injective.

- **Inversibilité de la différentielle.** Calculons la matrice jacobienne de ϕ en un point quelconque :

$$J_\phi(\theta, r) = \begin{pmatrix} -r\sin(\theta) & \cos(\theta) \\ r\cos(\theta) & \sin(\theta) \end{pmatrix}.$$

Son déterminant est égal à

$$-r\sin(\theta)^2 - r\cos(\theta)^2 = -r,$$

qui est non-nul. Ainsi, par le Théorème B.5, ϕ est un $\mathcal{C}^1$-difféomorphise. Par conséquent une fonction $f : \mathbf{R}^2 \to \mathbf{R}$ est de classe $\mathcal{C}^1$ si et seulement si $f \circ \phi$ l'est.

ANNEXE C

TANGENCE

Je m'étais mis à choyer immodérément les mots pour l'espace qu'ils admettent autour d'eux, pour leurs tangences avec d'autres mots innombrables.

A. Breton, Manifeste du surréalisme

Nous allons dans ce bref appendice donner quelques éléments pour mieux appréhender la notion d'espace tangent à une sous-variété, introduite au Paragraphe 3.1.3 du Chapitre 3. Rappelons pour plus de confort la définition de sous-variété que nous avions alors introduite :

Définition C.0.1. Une *sous-variété de* $\mathbf{R}^n$ *de codimension* p *et de classe* $\mathcal{C}^k$ est une partie $\mathcal{S} \subset \mathbf{R}^n$ telle que pour tout $x \in \mathcal{S}$, **quitte a permuter les coordonnées**, il existe

- Un ouvert U de $\mathbf{R}^{n-p}$;
- Un ouvert V de $\mathbf{R}^p$;
- Une application $\varphi : U \to V$ de classe $\mathcal{C}^k$ telle que $x \in U \times V$ et

$$\mathcal{S} \cap (U \times V) = \mathcal{S}_\varphi.$$

Si $\mathcal{S}$ désigne une sous-variété de $\mathbf{R}^n$ de codimension p et de classe $\mathcal{C}^1$, nous avons défini pour tout point x de $\mathcal{S}$ un sous-espace vectoriel $T_x\mathcal{S}$ en utilisant une fonction φ de la définition. Le problème, c'est qu'il y a souvent plusieurs choix possibles de fonction φ, et qu'il n'est pas clair a priori qu'elles vont toutes donner le même sous-espace vectoriel.

Pour montrer que c'est bien le cas, nous allons donner une autre description de cet *espace tangent* qui n'utilisera cette fois que $\mathcal{S}$ en tant qu'ensemble. Cette description repose sur une notion de *vecteur tangent* que nous introduisons maintenant.

Définition C.1. Un vecteur $v \in \mathbf{R}^n$ est dit *tangent* à $\mathcal{S}$ en un point x s'il existe $\epsilon > 0$ et une fonction continue $c :]-\epsilon;\epsilon[\to \mathcal{S}$ dérivable telle que

- $c(0) = x$;
- $c'(0) = v$.

Le choix de la lettre c n'est pas anodin, elle a pour but d'évoquer un *chemin*. En effet, l'idée cachée derrière l'objet que nous venons de définir provient de la physique : il s'agit de suivre le déplacement d'un point le long d'un chemin et de regarder ses "vecteurs vitesses", qui sont les vecteurs $c'(t)$. Ainsi, les vecteurs tangents au point x sont les vecteurs qui peuvent décrire la vitesse instantanée en x d'un point se déplaçant sur $\mathcal{S}$.

Il n'est pas évident sur la définition précédente que les vecteurs vitesse forment un sous-espace vectoriel de $\mathbf{R}^n$, ni encore moins qu'ils ont quelque chose avoir avec $T_x\mathcal{S}$. Et pourtant ...

Théorème C.2 Soit $\mathcal{S}$ une sous-variété de $\mathbf{R}^n$ de codimension p et de classe $\mathcal{C}^1$. Alors, pour tout $x \in \mathcal{S}$, les vecteurs tangents au point x forment un espace vectoriel qui est isomorphe à $T_x\mathcal{S}$ (qui ne dépend donc pas du choix de φ).

Démonstration. Soit $x \in \mathcal{S}$. On fixe $U \subset \mathbf{R}^{n-p}$, $V \subset \mathbf{R}^p$ des ouverts tels que $x \in U \times V$ ainsi que $\varphi : U \to V$ de classe $\mathcal{C}^1$ telle que $\mathcal{S} \cap (U \times V) = \mathcal{S}_\varphi$.

Soit v un vecteur tangent à $\mathcal{S}$ au point x et c la fonction correspondante. Si on note $(c_1, \cdots, c_n)$ les coordonnées de la fonction c, posons $\widetilde{c} = (c_1, \cdots, c_{n-p})$. Comme $U \times V$ est un ouvert contenant $x = c(0)$ et que c est en particulier continue, il existe $\epsilon' > 0$ tel que $c(t) \in U \times V$ pour tout $t \in]-\epsilon'; \epsilon'[$. Par conséquent, pour de tels t on a

$$c(t) = (\widetilde{c}(t), \varphi \circ \widetilde{c}(t)).$$

En dérivant cette égalité en $t = 0$, on obtient alors

$$c'(0) = (\widetilde{c}'(0), D_x\varphi(\widetilde{c}(0))) \in T_x\mathcal{S}.$$

Réciproquement, soit $v \in T_x\mathcal{S}$. Il existe donc $x \in \mathbf{R}^{n-p}$ tel que

$$v = (w, D_x\varphi(w)).$$

Soit maintenant $\widetilde{x} = (x_1, \cdots, x_{n-p}) \in U$ et $r > 0$ tel que $B(\widetilde{x}, r) \subset U$. En posant $\epsilon = r/\|w\|$, on a que pour tout $t \in]-\epsilon; \epsilon[$,

$$\|(\widetilde{x} + tw) - \widetilde{x}\| < r,$$

donc que $\widetilde{x} + tw \in U$. Par conséquent, la fonction $c :]-\epsilon; \epsilon[\to \mathbf{R}^n$ définie par

$$c(t) = (\widetilde{x} + tw, \varphi(\widetilde{x} + tw))$$

est à valeurs dans $\mathcal{S}$ et $c(0) = (\widetilde{x}, \varphi(\widetilde{x})) = x$. Pour conclure, il suffit de dériver en 0, ce qui donne

$$c'(0) = (w, D_{\widetilde{x}}\varphi(w)) = v.$$

■

Donnons un exemple dans l'un des cas les plus simples.

Exemple C.0.2. On considère un point M du cercle de centre 0 et de rayon 1 de coordonnées (x_0, y_0). Un chemin passant par M est donc une fonction $c :]-\epsilon; \epsilon[\to \mathbf{R}^2$ de classe $\mathcal{C}^1$ telle que, en notant $c(t) = (x(t), y(t))$, on ait $c(0) = M$ et pour tout $t \in]-\epsilon; \epsilon[$,

$$x(t)^2 + y(t)^2 = 1.$$

Si on dérive cette relation, on obtient

$$2x'(t)x(t) + 2y'(t)y(t) = 0.$$

En simplifiant les facteurs 2, on peut écrire cette égalité sous la forme $\langle c'(t), c(t)\rangle = 0$. Ceci est vrai en particulier en $t = 0$, et donc $\langle c'(0), \overrightarrow{OM}\rangle = 0$. Ainsi, tout vecteur tangent au cercle au point M est orthogonal au rayon OM. Nous retrouvons bien encore une fois la caractérisation usuelle de la tangente en un point du cercle.

On peut d'ailleurs, montrer que réciproquement, tout vecteur orthogonal au rayon est un vecteur tangent. En effet, on sait déjà qu'on dispose d'au moins un tel vecteur : au voisinage de M, le cercle peut être décrit comme le graphe d'une fonction (quitte à permuter les coordonnée), autrement dit comme l'image d'un chemin. Comme la fonction est un difféomorphisme, sa dérivée n'est pas nulle et fournit donc un vecteur tangent $v \neq 0$. Si maintenant w est un vecteur orthogonal à $\overrightarrow{OM}$, alors il existe $\alpha \in \mathbf{R}$ tel que $w = \alpha v$. On vérifie par un calcul direct que si on définit le chemin $c_\alpha :]-\epsilon/\alpha; \epsilon/\alpha[\to \mathbf{R}^2$ par $c_\alpha(t) = c(\alpha t)$, alors

$$\begin{aligned} c'_\alpha(0) &= \alpha c'(0) \\ &= \alpha v \\ &= w. \end{aligned}$$

Muni de cette description de l'espace tangent, il nous est maintenant possible de donner une démonstration directe rigoureuse du lien entre le gradient d'une fonction sur une sous-variété et les extrema locaux (Corollaire 3.1.22) sans passer par le Théorème 3.1.9. Cette démonstration avait été évoquée de façon heuristique après le Corollaire 3.1.22.

Démonstration directe du Corollaire 3.1.22. Notons v le projeté orthogonal de $\nabla f(\widetilde{x})$ sur l'espace tangent $T_{\widetilde{x}}\mathcal{S}$ et supposons que $v \neq 0$. Il existe alors $\epsilon > 0$ et une fonction $c :]-\epsilon; \epsilon[\to \mathcal{S}$ de classe $\mathcal{C}^1$ telle que $c(0) = \widetilde{x}$ et $c'(0) = v$. Par continuité de c en 0, il existe $\epsilon' \leqslant \epsilon$ tel que $c(]-\epsilon'; \epsilon'[\subset U \cap \mathcal{S}$. On peut alors considérer la fonction $\phi = f \circ c :]-\epsilon'; \epsilon'[\to \mathcal{S}$. Le développement limité de ϕ en 0 s'écrit

$$\begin{aligned} \phi(t) &= \phi(0) + t\phi'(0) + o(t) \\ &= f(\widetilde{x}) + t\langle \nabla f(\widetilde{x}), v\rangle + o(t) \\ &= f(\widetilde{x}) + t\|v\|^2 + t\varepsilon(t) \\ &= f(\widetilde{x}) + t(\|v\|^2 + \varepsilon(t)) \end{aligned}$$

Comme $\varepsilon(t) \to 0$ quand $t \to 0$, il existe $\delta > 0$ tel que pour tout $t \in]-\delta; \delta[$, $|\varepsilon(t)| \leqslant \|v\|^2/2$. Par conséquent, pour tout $t \in]0; \delta[$,

$$\phi(t) \geqslant f(\widetilde{x}) + t\frac{\|v\|^2}{2} > f(\widetilde{x})$$

tandis que pour tout $t \in]-\delta; 0[$,

$$\phi(t) \leqslant f(\widetilde{x}) + t\frac{\|v\|^2}{2} < f(\widetilde{x}).$$

Ainsi, il existe des points de $\mathcal{S} \cap U$ arbitrairement proches de $\widetilde{x}$ pour lesquels f prend des valeurs strictement supérieures à $f(\widetilde{x})$ et des valeurs strictement inférieures) $f(\widetilde{x})$. Autrement dit, f n'a pas d'extremum local en $\widetilde{x}$. ∎

ANNEXE D

IN CAUDA VENENUM

Pour qui sont ces serpents qui sifflent sur vos têtes ?

J. RACINE, Andromaque

Tout au long de ce texte, nous avons décrit des méthodes numériques pour approcher l'optimum d'une fonction. Nous avons également donné des critères de convergence pour ces méthodes, mais nous n'avons pas parlé de leur implémentation concrète. La raison principale en est que cela nous aurait éloigné de notre propos, et se serait mieux prêté à des expérimentations sur machine qu'à un cours en amphithéâtre. Néanmoins, il serait incongru de ne pas donner au moins un aperçu des algorithmes évoqués. C'est pourquoi nous allons les décrire maintenant, en utilisant le langage PYTHON [1].

Pour cela, commençons par quelques généralités. Les algorithmes que nous avons évoqués sont tous *itératifs*, au sens où ils construisent par récurrence une suite qui converge vers le point recherché. L'implémentation sera donc construite autour d'une boucle `for`. Par ricochet, ceci impose de définir à l'avance un nombre maximal d'itération. Il sera dans la suite toujours noté `N`. Ainsi, nous calculerons les termes de la suite $(x_n)_{n\in\mathbf{N}}$ jusqu'à x_N. Néanmoins, il n'est pas forcément nécessaire de faire autant de calculs. En effet, dans la pratique, ce que l'on souhaite c'est obtenir une approximation de la solution avec une précision donnée. Pour cela, on se fixe une *tolérance* ϵ. Mais alors, il est inutile de poursuivre les calculs une fois qu'on est parvenu à une solution "à ϵ près". C'est la raison pour laquelle nos algorithmes incluront une *condition d'arrêt* : à chaque itération, on commence par vérifier si la précision souhaitée a été atteinte. On a alors deux options :

- Si c'est le cas, on peut terminer immédiatement en renvoyant la valeur trouvée ;
- Sinon, on poursuit l'opération.

Assez naturellement, ceci s'implémentera par une instruction conditionnelle avec une boucle `if`, incluant comme condition une comparaison avec la tolérance qui sera notée `epsilon`.

1. L'apprentissage et l'utilisation de PYTHON ne sont pas l'objet de ce texte et nous prendrons donc pour acquise une connaissance au moins sommaire du langage. Nous renvoyons par exemple le lecteur à [8] pour une introduction.

Un nouveau problème surgit alors : il se pourrait qu'au bout des N itérations, on ne soit pas parvenu à la tolérance souhaitée. La procédure est alors mise en échec, et il faut prévoir un cas d'exception pour renvoyer l'information. Autrement dit, on retournera alors un message indiquant que l'algorithme n'a pas pu conclure.

Concluons par un aspect plus concret : Python dispose de nombreuses librairies contenant des outils divers en fonction des besoins. Dans notre cas, il faudra pouvoir manipuler des vecteurs et des matrices et ceci peut se faire à l'aide de la librairie `numpy`. Il faudra donc commencer ainsi :

```
import numpy as np
```

D.1 Méthode de Newton

La Méthode de Newton a été vue au Paragraphe 2.4.2 du Chapitre 2. De même qu'elle était simple à expliquer, elle est aisée à mettre en œuvre. Son argument est a priori une fonction $f : \mathbf{R}^n \to \mathbf{R}^n$, mais nous aurons également besoin de sa différentielle. Nous prendrons donc ces deux fonctions, `f` et `df` comme arguments. Il nous faudra également un point de départ `xo` ainsi comme toujours qu'une tolérance `epsilon` et un nombre maximal d'itérations `N`.

Le but de l'algorithme est de trouver un point x tel que $f(x) = 0$, en construisant une suite par la formule de récurrence

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

Il est donc assez raisonnable de faire porter la condition d'arrêt sur la norme $\|f(x_n)\|$. Ceci mène au code suivant :

```
def Newton(f, df, xo, epsilon, N):
  xn = xo
  for n in range(N):
    if norm(f(xn)) < epsilon:
      return(xn)
    xn = xn - dot(np.linalg.inv(df(xn)), f(xn))
  return("Echec")
```

D.2 Méthode de Broyden

La Méthode de Broyden est une amélioration de la précédente qui a été décrite au Paragraphe 2.4.3 du Chapitre 2. L'idée principale est d'utiliser une approximation de l'inverse de la Jacobienne $J_f(x_n)^{-1}$ calculée itérativement pour éviter les calculs d'inverses. L'implémentation est une légère modification de la précédente – gardant en particulier les mêmes arguments et la même condition d'arrêt – avec cependant un détail : le calcul de la nouvelle approximation H_{n+1} utilise non seulement la valeur de x_{n+1} (calculée avec H_n), mais également l'ancienne valeur x_n. En effet, les relations de

récurrence sont

$$\begin{cases} x_{k+1} = & x_k - H_k \phi(x_k) \\ H_{k+1} = H_k + \dfrac{(x_{k+1}-x_k) - H_k(\phi(x_{k+1}) - \phi(x_k))}{(x_{k+1}-x_k)^t H_k(\phi(x_{k+1}) - \phi(x_k))} (x_{k+1}-x_k)^t H_k \end{cases}$$

Il faut donc garder celle-ci en mémoire le temps de faire le calcul. C'est pourquoi on introduit ici une variable `xnplusun` qui servira au calcul avant de mettre à jour la valeur de `xn`. De plus, afin d'alléger la rédaction, le dénominateur de la Formule de Sherman-Morrison est calculé séparément.

```
def Broyden(f, df, xo, epsilon, N):
  Hn = np.linalg.inv(df(xo))
  xn = xo
  for n in range(N):
    if norm(f(xn)) < epsilon
      return(xn)
    xnplusun = xn - dot(Hn, f(xn))
    an = dot(transpose(xnplusun - xn), dot(Hn, f(xnplusun) - f(xn)))
    Hn = Hn + ((xnplusn - xn - dot(Hn, f(xnplusun) - f(xn))/an)*Hn
    xn = xnplusun
  return("Echec")
```

D.3 Algorithme d'Uzawa

L'algorithme d'Uzawa, détaillé au Paragraphe 4.3.2, ne cherche pas les zéros d'une fonction mais le minimum d'une fonction quadratique sous contraintes d'inégalité. Ceci signifie que le choix de la condition d'arrêt n'est pas aussi immédiat. En effet, puisqu'on ne connaît pas le point $\widetilde{x}$ auquel le minimum est atteint, on ne peut calculer $\|\widetilde{x} - x_n\|$ pour estimer la qualité de l'approximation. Une solution est d'utiliser les conditions KKT, et de s'arrêter quand ces dernières sont satisfaites à ϵ près, c'est-à-dire quand

$$\left\| Ax_n + b - H^t \mu_n \right\| < \epsilon.$$

Pour implémenter l'Algorithme d'Uzawa, nous aurons également besoin de la projection P sur $\mathbf{R}^p_-$. Celle-ci peut se définir sans peine à l'aide de la fonction `min` de Python, qui utilise le même ordre sur les vecteurs que nous (c'est-à-dire la comparaison coordonnée par coordonnée).

```
def P(x):
  min(o, x)
```

Cela dit, on peut également directement utiliser la fonction `min` dans l'algorithme, ce que nous allons faire. Regardons maintenant les arguments de la fonction. Il nous faudra la matrice `A` et le vecteur `b` définissant la forme quadratique, ainsi que la matrice `H` et le vecteur `d` définissant les contraintes. Nous aurons également besoin d'un vecteur de départ `muo` ainsi que d'un pas `t`[2]. Enfin, les indispensables tolérance `epsilon` et nombre maximal d'itérations `N`.

2. Inclure le pas comme variable permet de facilement tester plusieurs valeurs, afin par exemple d'améliorer la vitesse de convergence si l'algorithme échoue.

Rappelons la formule de récurrence utilisée :

$$\left\{ \begin{array}{rcl} x_k & = & A^{-1}\left(H^t \mu_k - b\right) \\ \mu_{k+1} & = & P\left(\mu_k - t_k(Hx_k - d)\right) \end{array} \right.$$

Avant de donner le code, ajoutons un détail : il est préférable de calculer dès le début les matrices $B = A^{-1}$ et $K = H^t$ afin de ne pas refaire ce calcul à chaque fois [3].

```
def Uzawa(A, b, H, d, muo, t, epsilon, N):
  B = np.linalg.inv(A)
  K = H.transpose()
  mun = muo
  xn = np.dot(B, np.dot(K, mun)-b)
  for n in range(N):
    if norm(np.dot(A, xn) + b - np.dot(K, mun)) < epsilon:
      return(xn)
    mun = min(o, mun - t*(np.dot(K, xn) - d))
    xn = np.dot(B, np.dot(K, mun)-b)
  return("Echec")
```

D.4 Descente de gradient

D.4.1 À pas constant

Rappelons tout d'abord que le but de cette méthode est de chercher un point critique d'une fonction $f : \mathbf{R}^n \to \mathbf{R}$ en partant d'un point $x_0 \in \mathbf{R}^n$ et en définissant par récurrence une suite par la relation

$$x_{n+1} = x_n - t\nabla f(x_n),$$

où $t > 0$ est le pas de la méthode. Comme dans le cas précédent, le pas t sera inclus comme variable de l'algorithme afin de pouvoir aisément l'ajuster.

Pour tester la précision d'une itération donnée, le plus simple est de calculer la norme du gradient $\|\nabla f(x_n)\|$. En effet, l'algorithme converge vers un point critique de f, qui vérifie donc $\nabla f(x) = 0$. La fonction f étant supposée de classe $\mathcal{C}^1$, ∇f sera petit dans un voisinage de x, donc on peut décider de s'arrêter lorsque $\|\nabla f(x_n)\| < \epsilon$. On obtient alors l'algorithme suivant, qui ne nécessite d'ailleurs pas de prendre f en argument mais seulement son gradient `gradf`, ainsi bien sûr que le pas `t`, le point de départ `xo`, la tolérance `epsilon` et le nombre d'itérations `N`. Remarquons que le gradient est utilisé deux fois dans l'itération : pour vérifier la condition d'arrêt, puis pour calculer la nouvelle valeur de x_n. Il est donc judicieux de le calculer une seule fois, au début de la boucle, et de stocker sa valeur dans une variable temporaire `grad` qu'on appellera ensuite.

```
def GradienAPasConstant(gradf, t, xo, epsilon, N):
  xn = xo
  for n in range(N):
    grad = gradf(xn)
    if norm(grad) < epsilon:
```

3. On pourrait aussi envisager de ne jamais calculer A^{-1} et d'appeler un algorithme pour résoudre numériquement l'équation $Ax_n + b = H^t \mu_n$ à chaque étape, si ceci s'avère pertinent.

```
      return(xn)
    xn = xn - t*grad
  print("Erreur")
```

D.4.2 À pas optimal

Pour la méthode du gradient à pas optimal, le seul changement est que l'on doit à chaque étape calculer le nouveau pas. Dans le cas quadratique, qui est le seul auquel nous nous sommes intéressés, ce pas se calcule facilement grâce à une formule explicite, ce qui simplifie l'algorithme. De plus, la donnée du problème étant réduite à la matrice symétrique définie positive A et au vecteur b, ce sont ces deux quantités qui apparaîtront comme variables plutôt que le gradient la fonction f. Cela permettra de simplifier également le calcul du gradient, qu'on prendre soin de n'effectuer qu'une seule fois par itération à l'aide d'une variable auxiliaire comme précédemment.

```
def GradientAPasOptimal(A, b, xo, epsilon, N):
  xn = xo
  for n in range(N):
    grad = dot(A, xn) - b
    if norm(grad) < epsilon:
      return(xn)
    tn = (norm(grad)^4)/(dot(dot(A, grad), grad)^2)
    xn = xn - tn*grad
  print("Erreur")
```

BIBLIOGRAPHIE

[1] P. Ciarlet & J.-L. Lions – *Introduction à l'analyse numérique matricielle et à l'optimisation*, Sciences Sup, Dunod, 2007.

[2] J.-C. Culioli – *Introduction à l'optimisation*, Ellipses, 2012.

[3] G. Debreu – *Theory of value. An axiomatic analysis of economic equilibrium*, Yale University Press, 1959.

[4] R. Feynman, R. Leighton & M. Sands – *Le cours de physique de Feynman : Mécanique quantique*, Dunod, 2018.

[5] M. Hoy, J. Livernois, C. McKenna, R. Rees & T. Stengos – *Mathematics for Economics*, MIT press, 2022.

[6] H.-U. J.-B. – *Optimisation et analyse convexe*, EDP Sciences, 2023.

[7] J. Lafontaine – *Introduction aux variétés différentielles*, EDP sciences, 1996.

[8] V. Le Goff – *Apprenez à programmer en Python (4ème édition)*, Eyrolles, 2022.

[9] F. Rouvière – *Petit guide de calcul différentiel à l'usage de la licence et de l'agrégation*, Cassini, 2009.

[10] C. Simon & L. Blume – *Mathematics for economists*, Norton, New York, 1994.

[11] S. Smale – « Global analysis and economics », *Handbook of mathematical economics* 1 (1981), p. 331–370.

[12] P. Thegem – *Recherche opérationnelle, tome 1 – méthodes d'optimisation*, Ellipses, 2012.

[13] — , *Recherche opérationnelle, tome 2 – gestion de production, méthodes aléatoires, aide multicritère*, Ellipses, 2013.

[14] M. Yildizoglu – *Introduction à la théorie des jeux*, Dunod, 2011.