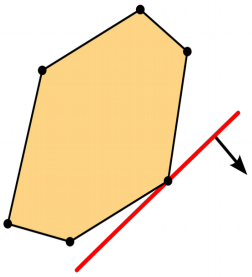
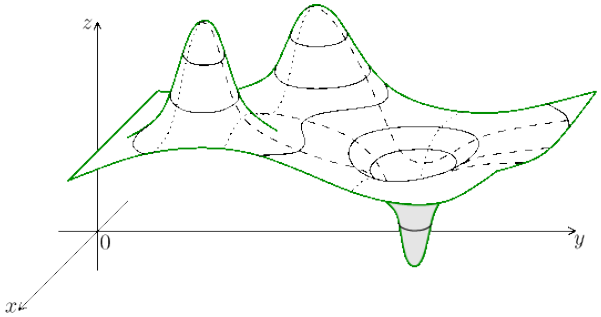
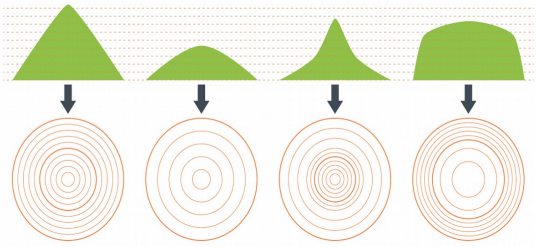
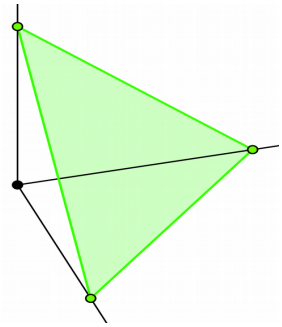
OPTIMISATION

S. LE BORGNE









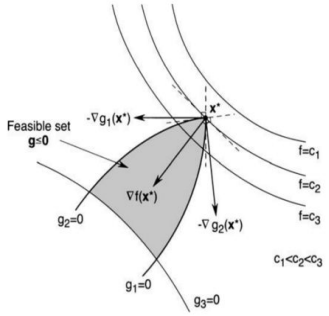


Figure 21.1 Illustration of the Karush-Kuhn-Tucker (KKT) theorem.

Table des matières

1 Introduction

Ce cours est destiné à des étudiants de licence mention MiaSHS en troisième année. Le public visé a donc suivi des cours d'algèbre linéaire et de calcul différentiel en deux ou trois variables. Nous introduisons peu de nouvelles notions mathématiques. Le but que nous nous fixons est de consolider les acquis mathématiques des étudiants et d'étudier des problèmes, des exemples. On décrit aussi plusieurs algorithmes (mais aucun programme).

1.1 Repères historiques

Le calcul différentiel est une invention scientifique majeure, un outil extraordinaire. Son histoire a fait l'objet d'innombrables études. Je ne vais pas me risquer dans un domaine que je connais mal, mais puisque nous utiliserons le calcul différentiel comme outil permettant de trouver les extrema d'une fonction, je vais expliquer ce que Fermat savait déjà faire en 1638 sans calcul différentiel qui n'existait pas encore. Un lecteur moderne peut facilement reformuler le raisonnement en utilisant les dérivées. Il serait intéressant de lire les mathématiques qu'écrivait Fermat. Vous trouverez facilement le texte orginal complet. En voici le début :

Je veux par ma méthode *couper la ligne* AC (*fig.* 18) *donnée en telle sorte au point* B, *que le solide compris soubs le quarré*

Fig. 18.

de AB *et la ligne* BC *soit le plus grand* de tous les solides descrits de mesme sorte, en coupant AC en quelque autre point que ce soit .

Posons en notes que la ligne AC s'appelle B et la ligne AB inconnue A, BC sera $B - A$. Il faudra donc que le solide $Aq.$ in $B - Ac.$ satisface à la question.

Prenons derechef au lieu de A, $A + E$; le solide qui se fera du quarré de $A + E$ et de $B - A - E$ sera :

$$B \text{ in } Aq. + B \text{ in } Eq. + B \text{ in } A \text{ in } E \text{ bis} - Ac. - A \text{ in } Eq. \text{ ter} - Aq. \text{ in } E \text{ ter} - Ec.$$

Ce que fait Fermat dans les quatre dernières lignes ci-dessus est de calculer la même quantité en ajoutant un E (qu'il annulera plus tard, après avoir divisé par E : c'est une façon de calculer la limite d'un rapport). Décrivons son raisonnement en utilisant des notations modernes (ce qui pose problème, mais faisons-le quand même).

La question posée au début est la suivante : on se donne un segment $[AC]$ de longueur b et on cherche à placer un point B sur segment à distance a de A de façon à rendre maximum la quantité $a^2(b-a)$ (le volume de solide de base un carré de côté a et de hauteur $b-a$). Cette quantité est égale à $a^2b - a^3$ ($Aq.$ désigne le carré de A, *in* le produit, $Ac.$ le cube de A). Si on remplace a par $a + \epsilon$ et qu'on développe l'expression, on obtient

$$\begin{aligned}(a+\epsilon)^2 b - (a+\epsilon)^3 &= (a^2 + 2a\epsilon + \epsilon^2)b - (a^3 + 3a^2\epsilon + 3a\epsilon^2 + \epsilon^3) \\ &= ba^2 + b\epsilon^2 + 2ba\epsilon - a^3 - 3a^2\epsilon - 3a\epsilon^2 - \epsilon^3\end{aligned}$$

(Fermat écrit *bis* et *ter* pour la multiplication par deux et trois). Que fait ensuite Fermat ? D'abord la différence des deux expressions, comme si elles étaient égales bien qu'elles ne le soient pas. Cela donne

$$(a+\epsilon)^2 b - (a+\epsilon)^3 - a^2(b-a) = b\epsilon^2 + 2ba\epsilon - 3a^2\epsilon - 3a\epsilon^2 - \epsilon^3.$$

Il reprend ensuite l'idée que la différence est nulle et en déduit que la somme des termes marqués d'un signe $+$ est égale à celle des termes marqués d'un signe $+$:

$$b\epsilon^2 + 2ba\epsilon = 3a^2\epsilon + 3a\epsilon^2 + \epsilon^3.$$

Il divise ensuite par ϵ (autant de fois que possible, dit-il, une fois ici) :

$$b\epsilon + 2ba = 3a^2 + 3a\epsilon + \epsilon^2.$$

Cela fait, il efface tous les termes où intervient encore une puissance de ϵ :

$$2ba = 3a^2$$

Il dit alors qu'il ne faut plus faire des comparaisons feintes (*comme si ϵ était nul*) mais une vraie équation. Il réécrit enfin l'équation

$$2b = 3a$$

et conclut

Reuenons à nostre question et diuisons AC au point B en sorte que

AC soit à AB comme 3 à 2,

ie dis que le solide du quarré AB en BC sera le plus grand de tous ceux qui peuvent semblablement estre descris sur la ligne AC en quelque autre section que ce soit.

Voilà. Plusieurs affirmations ne sont pas justifiées comme nous demandons à un étudiant de les justifier. Mais Fermat ne se trompe pas ; le maximum du volume qu'il cherche est bien celui qu'il dit et sa technique (associée à d'autres considérations) est une très bonne technique pour résoudre de tels problèmes.

Le calcul différentiel a été considérablement développé au dix-huitième siècle. Citons deux mathématiciens importants : Euler (1707-1783), Lagrange (1736-1813). Au dix-neuvième siècle le développement s'est poursuivi, en particulier les raisonnements d'analyse ont été rendus plus rigoureux (citons Cauchy (1789-1857) et Weierstrass (1815-1897)).

Les succès considérables du calcul différentiel en physique ont pu donner à certains savants l'impression que l'homme avait alors découvert toutes les lois de la nature et que sa capacité à prédire le futur n'était plus limité que par ses possibilités de calculs.

« *Nous devons envisager l'état présent de l'univers comme l'effet de son état antérieur, et comme la cause de celui qui va suivre. Une intelligence qui, pour un instant donné, connaîtrait toutes les forces dont la nature est animée et la situation respective des êtres qui la composent, si d'ailleurs elle était assez vaste pour soumettre ces données à l'analyse, embrasserait dans la même formule les mouvements des plus grands corps de l'univers et ceux du plus léger atome : rien ne serait incertain pour elle, et l'avenir, comme le passé, serait présent à ses yeux.* »Pierre-Simon Laplace Essai philosophique sur les probabilités (1814)

On a depuis découvert des phénomènes qui tempèrent ce point de vue.

« *Une cause très petite, qui nous échappe, détermine un effet considérable que nous ne pouvons pas ne pas voir, et alors nous disons que cet effet est dû au hasard...Mais, lors même*

que les lois naturelles n'auraient plus de secret pour nous, nous ne pourrons connaître la situation initiale qu'approximativement. Si cela nous permet de prévoir la situation ultérieure avec la même approximation, c'est tout ce qu'il nous faut, nous dirons que le phénomène a été prévu, qu'il est régi par des lois ; mais il n'en est pas toujours ainsi, il peut arriver que de petites différences dans les conditions initiales en engendrent de très grandes dans les phénomènes finaux... »(Poincaré. 1908).

Dans les années 20 les physiciens ont inventé la mécanique quantique. Au niveau atomique les meilleurs modèles ne sont plus déterministes...

1.2 Les mathématiques en économie

Les mathématiques sont beaucoup utilisées par une partie des économistes, en particulier dans les modèles microéconomiques. Le vocabulaire, les concepts mathématiques font partie du langage de certains économistes.

« *Admettre que l'individu puisse se contenter d'une consommation nulle de tous les biens apparaît moins satisfaisant, puisque c'est négliger l'existence d'un minimum vital physique ou sociologique que l'économiste devrait reconnaître. Cependant supposer que le vecteur nul appartient à X simplifiera les démonstrations, ce qui m'est ici une justification suffisante.* »
E. Malinvaud, Leçons de théorie microéconomique.

« *In order to insure that optimality positions do not lie at "corners", it is necessary to place some restrictions on the second derivatives of the cost function. (...) Reasonable men will often differ on the amount of damages or benefits caused by different activities. To some, any wage rates set by competitive labor markets are permissible, while others, rates below a certain minimum are violations of basic rights ; to some, gambling, prostitution and even abortion should be freely available to anyone willing to pay market price, while to others, gambling is sinful and abortion is murder.* » G. Becker, Crime and punishment.

Comme le montrent ces deux exemples les phrases des économistes sont parfois des mélanges qui peuvent soulever des questions, ou des objections. Il n'est pas impossible que les lignes citées ci-dessus soient des raccourcis pris dans un souci pédagogique de simplification. La signification des modèles mathématiques introduits en économie est souvent sujet à polémique.

Voici une position sur la question qui me semble assez saine.

« *Basically, this course is about a certain class of economic concepts and models. Although we will be studying formal concepts and models, they will always be given an interpretation. An economic model differs substantially from a purely mathematical model in that it is a combination of a mathematical model and its interpretation. The names of the mathematical objects are an integral part of an economic model. When mathematicians use terms such as "field" or "ring" which are in everyday use, it is only for the sake of*

convenience. When they name a collection of sets a "filter," they are doing so in an associative manner ; in principle, they could call it "ice cream cone." When they use the term "good ordering" they are not making an ethical judgment. In contrast to mathematics, interpretation is an essential ingredient of any economic model. The word "model" sounds more scientific than "fable" or "fairy tale" but I don't see much difference between them. The author of a fable draws a parallel to a situation in real life and has some moral he wishes to impart to the reader. The fable is an imaginary situation which is somewhere between fantasy and reality. Any fable can be dismissed as being unrealistic or simplistic but this is also the fable's advantage. Being something between fantasy and reality, a fable is free of extraneous details and annoying diversions. In this unencumbered state, we can clearly discern what cannot always be seen from the real world. On our return to reality, we are in possession of some sound advice or a relevant argument that can be used in the real world. We do exactly the same thing in economic theory. Thus, a good model in economic theory, like a good fable, identifies a number of themes and elucidates them. We perform thought exercises which are only loosely connected to reality and which have been stripped of most of their real-life characteristics. However, in a good model, as in a good fable, something significant remains. One can think about this book as an attempt to introduce the characters that inhabit economic fables. Here, we observe the characters in isolation. In models of markets and games, we further investigate the interactions between the characters. »Ariel Rubinstein, Lecture Notes in Microeconomic Theory.

Le sens du modèle économique est très discuté. L'adéquation au réel n'est pas nécessairement le but recherché mais plutôt parfois l'exploration des possibles ou la découverte de modes de pensée.

L'un des succès de l'application du calcul différentiel à l'économie est la démonstration de théorème d'existence d'équilibres pour des marchés sous différentes conditions. Le problème a été posé par Walras à la fin du 19ème siécle. Walras ne disposait pas de l'outil mathématiques adéquat pour démontrer le théorème d'existence d'un équilibre qu'il avait en vue. Ce n'est que presque cent ans plus tard que Debreu a établi le théorème. Mais Walras postulait aussi que la loi du marché conduisait naturellement à une position d'équilibre stable et optimale. Les économistes ont tenté d'établir un tel résultat. En vain, car ce n'est pas ce qui se produit (résultats de la fin des années 70 de Sonnenschein et Debreu).

« *Plus personne ne s'intéresse au problème de Walras.* » Bernard Maris.

Ce qu'écrit Bernard Maris est probablement plus un souhait (et l'expression de son propre désintérêt) qu'une description objective.

L'un des représentants du courant libéral de l'économie est l'autrichien Hayek. Sa position est particulière car il est souvent plus considéré comme un idéologue que comme un économiste. C'était le cas pendant sa carrière en tout cas. Il a connu son heure de gloire au moment des révolutions libérales qu'ont constituées les gouvernements de Reagan aux

États unis et Thatcher en Grande Bretagne. Ce n'est peut-être pas le laissez-faire qu'il défend. Il prône l'adoption par les agents économiques des règles morales en vigueur dans la société : ces règles morales contiennent selon lui le résultat de multiples tentatives d'organisation effectuées par les générations précédentes ; quand bien même elles ne sont pas rationnellement fondées elles sont meilleures que celles qu'un petit groupe d'hommes pourraient inventer ; ne pas connaître les raisons de leurs efficacité ne doit pas nous en écarter. Il use par exemple de l'argument suivant contre les progressistes (qui critiquent fréquemment l'individualisme des libéraux) : ne vouloir régler sa conduite que sur sa raison c'est ne pas attacher d'importance à ce qui est au plus haut point social, la morale, c'est donc se montrer particulièrement individualiste.

L'utilisation de mathématiques en économie est aujourd'hui souvent associée à une position de droite. Il n'est pas impossible que cela corresponde à peu près à la réalité actuelle des sciences économiques. Cela ne signifie pas qu'une orientation politique soit inscrite dans les mathématiques (ni même dans les applications des mathématiques ce qui n'exclut pas que les mathématiques puissent être appliquées pour justifier des positions de droite ; le cas d'Hayek montre que certains n'ont pas besoin de mathématiques pour être de droite). On trouvera aussi des utilisateurs des mathématiques souvent classés à gauche (Gaël Giraud par exemple). Sur ces questions on pourra se reporter aux travaux cités en références ([2], [3], [8], [14], [17]).

1.3 Commentaires sur la notion d'optimisation

Le mot optimisation est d'origine latine. C'est le superlatif de bonus qui signifie bon.

Bonus melior optimus
Paruus minor minimus
Magnus major maximus

En mathématiques chercher un optimum revient le plus souvent à chercher un minimum ou un maximum. Autrement dit on passe de bon à petit ou grand. Reste beaucoup de latitude : petit ou grand en quel sens ?

Distinction qualitatif/quantitatif. On dit parfois que l'utilisation de calcul pour maximiser ou minimiser suppose de s'intéresser uniquement à des données quantitatives. C'est souvent le cas effectivement. Cela dit il n'est pas toujours si facile de faire une différence entre données qualitatives et quantitatives (il est possible par exemple de coder une image par des nombres). Il n'est pas toujours impossible de manipuler des données qualitatives grâce à l'introduction de données quantitatives. Par exemple les algorithmes d'apprentissage profond permettent de faire de la reconnaissance d'image (performante) à partir de méthodes quantitatives.

Critique de l'utilitarisme (https://fr.wikipedia.org/wiki/Utilitarisme). L'utilitarisme fait usage des notions d'optimisation. Malgré le fait que l'utilitarisme ne doive pas

nécessairement être compris dans un sens étroit ou synonyme d'exploitation des humains (https://fr.wikipedia.org/wiki/économie_du_bien-être), il est souvent dévalorisé ([14]). La question est compliquée et on peut regretter l'importance trop grande accordée à la notion d'utilité ou bien discuter les différentes sortes d'utilité possibles, de leurs importances respectives, des échelles de temps auxquelles elles sont liées, etc... Mesurer l'utilité est sans doute assez difficile. Il est parfois reproché d'aller au plus facile, de tout réduire à l'argent. Une tendance serait de ne voir d'utilité qu'à ce qui serait monnayable, qu'à ce qui pourrait rapporter de l'argent. Ce n'est sans doute pas faire injure à ces critiques de remarquer qu'il n'est pas rare que ceux-là mêmes qui dénoncent l'argent roi expliquent en même temps qu'ils souffrent d'un manque de moyens (et reconnaissent donc l'importance de l'argent ; il semble plus difficile d'admettre que l'argent est important et qu'on puisse en avoir assez ou trop). La question du partage (quantitatif) des ressources est souvent associée à la question de la justice sociale (notion qualitative ?).

Paul Veyne écrit dans son autobiographie : « Mon père, fils de paysan nanti du certificat d'études primaires avec mention très bien, était devenu un négociant enrichi et réputé. (...) il se désintéressa de moi, tout en restant libéral à mon égard : à trente ans, grâce à lui, je roulais en Mercedes, car il avait le geste large, le pourboire facile, et sa largesse avait fait son succès de négociant auprès de ses clients et fournisseurs. »Le père de Paul Veyne a-t-il optimisé sa largesse ?

1.4 Classification des problèmes d'optimisation

Programmation linéaire.
Optimisation convexe.
Optimisation différentielle (ou lisse).
Optimisation SDP.
Optimisation non-différentielles (ou non lisse).
Optimisation multicritère (ou multiobjectif).
Optimisation en dimension infinie.
Optimisation discrète (ou combinatoire).
Sans contraintes. Avec contraintes.

1.5 Conseils de travail

Un premier conseil : travailler. Ne pas avoir peur d'en mourir comme les étudiants d'Yvette Guilbert.

Ils se remirent à l'étude
Avec acharnement

Ils se remirent à l'étude
Avec acharnement

N'avaient pas l'habitude
Sont morts au bout d'un an

Les choses ne se mettent en place que lentement. Remarques sur l'apprentissage (qui ne valent pas que pour les mathématiques) :
1) des choses qui paraissent très difficiles sont atteintes par tous ceux qui acceptent de suivre un apprentissage et un entraînement adéquats, ,
2) le temps nécessaire à l'apprentissage peut-être long (des années...),
3) la différence entre ceux qui ont accepté de travailler et les autres saute aux yeux (même de ceux qui ne sont pas compétents ; a fortiori de ceux qui le sont),
4) il n'est pas nécessaire de commencer très tôt pour réussir (mais quand on commence tard certains, plus jeunes, sont plus avancés ; est-ce un problème ?).

Les épreuves d'évaluation ne comporteront aucun piège. Ce que je souhaite vous apprendre figure dans les livres, est accessible. Les exercices posés aux devoirs surveillés seront destinés à vérifier que vous savez manier les notions vues en cours (et, hormis les notions vues les années précédentes nécessaires, uniquement cela). Les exercices seront des versions peu différentes d'exercices déjà faits. Cela ne signifie pas qu'il vous sera facile de les réussir. Cela signifie qu'il sera facile de le faire à ceux qui ont compris le cours. La plupart des notions du cours sont accessibles à tous ceux qui accepteront de passer du temps à les travailler.

1.6 Notations

$\mathbb{R}^d$ désigne l'espace vectoriel des d-uplets de nombres réels.

$\langle x, y\rangle$ désigne le produit usuel dans $\mathbb{R}^d$.

$\frac{\partial \phi}{\partial x}$: dérivée partielle de ϕ par rapport à x. Cela signifie que x est l'une des variables dont dépend ϕ.

Si ϕ est une fonction de $\mathbb{R}^d$ dans $\mathbb{R}$ nous noterons généralement $\frac{\partial \phi}{\partial x_i}$ la dérivée partielle de ϕ par rapport à sa i-ème variable.

$\nabla\phi$ est le vecteur gradient de ϕ c'est-à-dire le vecteur formé des dérivées partielles de ϕ. Nous écrirons ce vecteur en colonne :

$$\nabla\phi(x) = \begin{pmatrix} \frac{\partial \phi}{\partial x_1}(x) \\ \frac{\partial \phi}{\partial x_2}(x) \\ \vdots \\ \frac{\partial \phi}{\partial x_d}(x) \end{pmatrix}.$$

Il n'est pas rare qu'on sous-entende le point où on calcule les dérivée (pour alléger les

notations) :

$$\nabla\phi(x) = \begin{pmatrix} \frac{\partial\phi}{\partial x_1} \\ \frac{\partial\phi}{\partial x_2} \\ \vdots \\ \frac{\partial\phi}{\partial x_d} \end{pmatrix}.$$

Hessϕ est la matrice hessienne de ϕ :

$$\text{Hess}\phi(x) = \begin{pmatrix} \frac{\partial^2\phi}{\partial x_1^2} & \frac{\partial^2\phi}{\partial x_2\partial x_1} & \cdots & \frac{\partial^2\phi}{\partial x_d\partial x_1} \\ \frac{\partial^2\phi}{\partial x_1\partial x_2} & \frac{\partial^2\phi}{\partial x_2^2} & \cdots & \frac{\partial^2\phi}{\partial x_d\partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2\phi}{\partial x_1\partial x_d} & \frac{\partial^2\phi}{\partial x_2\partial x_d} & \cdots & \frac{\partial^2\phi}{\partial x_d^2} \end{pmatrix}.$$

Il faut faire attention aux noms des variables qui peuvent provoquer des confusions. En dimensions deux et trois il est fréquent que x, y désignent des variables (coordonnées de vecteur). En dimension plus grande x, y désigneront le plus souvent des vecteurs.

Nous utiliserons aussi les notations classiques de la statistique descriptive. Si $x_1, \ldots, x_n$, $y_1, \ldots, y_n$ sont des nombres on notera

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2$$

$$\sigma_x^2 = \overline{x^2} - \bar{x}^2$$

$$\overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$$

$$cov(x,y) = \overline{xy} - \overline{x}.\overline{y}$$

1.7 Définitions

Définition 1.1. *Soient D une partie de $\mathbb{R}^n$, f une fonction définie sur D à valeurs dans $\mathbb{R}$ et $x_0 \in D$.*
On dit que f est majorée sur D s'il existe $M \in \mathbb{R}$ tel que pour tout $x \in D$ on ait $f(x) \leq M$.
On dit que f est minorée sur D s'il existe $m \in \mathbb{R}$ tel que pour tout $x \in D$ on ait $f(x) \geq m$.
On dit que f est bornée sur D si elle est à la fois majorée et minorée. Cela revient à dire qu'il existe $M \geq 0$ tel que pour tout $x \in D$ on ait $|f(x)| \leq M$.
On dit que f a un minimum local en x_0 s'il existe $r > 0$ tel que pour tout $x \in B(x_0, r)$ on a $f(x) \geq f(x_0)$.

On dit que f a un maximum local en x_0 s'il existe $r > 0$ tel que pour tout $x \in B(x_0, r)$ on a $f(x) \leq f(x_0)$.
On dit que f a un minimum local strict en x_0 s'il existe $r > 0$ tel que pour tout $x \in B(x_0, r)$, $x \neq x_0$, on a $f(x) > f(x_0)$.
On dit que f a un maximum local strict en x_0 s'il existe $r > 0$ tel que pour tout $x \in B(x_0, r)$, $x \neq x_0$, on a $f(x) < f(x_0)$.
On dit que f a un minimum en x_0 si pour tout $x \in D$ on a $f(x) \geq f(x_0)$.
On dit que f a un maximum en x_0 si pour tout $x \in D$ on a $f(x) \leq f(x_0)$.
On dit que f a un minimum strict en x_0 si pour tout $x \in D$, $x \neq x_0$, on a $f(x) > f(x_0)$.
On dit que f a un maximum strict en x_0 si pour tout $x \in D$, $x \neq x_0$, on a $f(x) < f(x_0)$.

Définitions
Soient D une partie de $\mathbb{R}^n$, f une fonction définie sur D à valeurs dans $\mathbb{R}$.
Si f est majorée, on appelle borne supérieure de f sur D le nombre réel noté $\sup_D f$ ou $\sup_{x \in D} f(x)$ défini par :

$$\forall x \in D \quad f(x) \leq \sup_D f, \quad \forall M < \sup_D f, \ \exists x \in D, \ f(x) > M.$$

Si f est minorée, on appelle borne inférieure de f sur D le nombre réel noté $\inf_D f$ ou $\inf_{x \in D} f(x)$ défini par :

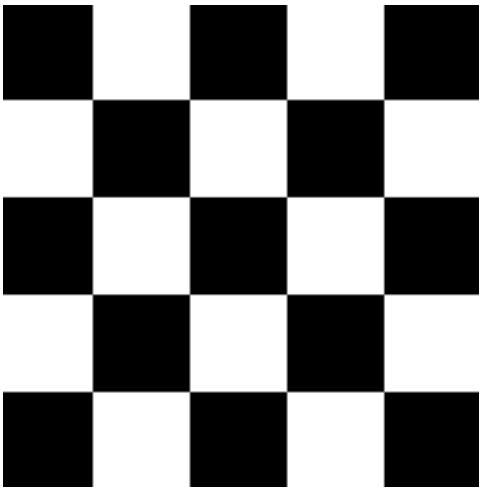
$$\forall x \in D \quad f(x) \geq \inf_D f, \quad \forall m > \inf_D f, \ \exists x \in D, \ f(x) < m.$$

Si f est majorée sur D, on dit que f atteint sa borne supérieure s'il existe $x \in D$ tel que $f(x) = \sup_D f$.
Si f est minorée sur D, on dit que f atteint sa borne inférieure s'il existe $x \in D$ tel que $f(x) = \inf_D f$.

2 Des exemples

2.1 Le théorème des quatre couleurs

Considérons un damier :



Vous voyez facilement que deux couleurs suffisent pour le colorier de sorte que deux régions contiguës ne soient jamais de la même couleur. Vous pourrez vérifier que si vous tracez des frontières uniquement avec des droites (infinies) et des cercles, alors deux couleurs vous suffiront. Ce n'est pas le cas en général. Cinq couleurs sont utilisées pour la carte ci-dessous.

Peut-on faire avec moins de cinq ? Au milieu du dix-neuvième siècle on a émis la conjecture que quatre couleurs suffisaient toujours. Cela a été démontré en 1976. La démonstration est très très longue et utilise l'ordinateur. Aucun humain n'a traité l'ensemble de toutes

les configurations possibles. *Ici on cherche à minimiser le nombre de couleurs pour colorier une carte.* Le problème n'a pas d'intérêt pratique. On connaît de toute façon la solution (la difficulté était de démontrer que ce qu'on pensait était vrai).

2.2 Droites de régression

2.2.1 Covariance et coefficient de corrélation linéaire, droites de régression

Ici on cherche à minimiser la distance d'un nuage de points à une droite. Soient $(x_i, y_i)_{i=1}^n$ une série statistique double. On cherche à approcher le nuage de points défini par cette série par une droite. Comment le faire ? Évidemment il est toujours possible de tracer une droite au jugé. On cherche une méthode systématique. Une méthode qui puisse être programmée. Plusieurs possibilités existent. Une manière raisonnable de procéder : on définit une quantité reflétant la "distance" du nuage de point à une droite quelconque, puis on cherche la meilleure droite au sens de cette distance, celle qui est à "distance" minimale.

Une droite est définie par son équation : $y = ax + b$.

Quelques quantités qui peuvent mesurer l'éloignement du nuage de points à cette droite :

$$F_1(a,b) = \sum_{i=1}^{n} |y_i - ax_i - b|$$

$$F_2(a,b) = \sum_{i=1}^{n} (y_i - ax_i - b)^2$$

$$F_3(a,b) = \sum_{i=1}^{n} (y_i - ax_i - b)^4$$

On pourrait aussi minimiser la somme des distances des points à la droite. Quelle quantité cela donne-t-il ? La distance du point (x_i, y_i) à la droite d'équation $y = ax + b$ est donnée par

$$\frac{|y_i - ax_i - b|}{\sqrt{1 + a^2}}.$$

La quantité à minimiser dans ce cas est donc

$$F_4(a,b) = \sum_{i=1}^{n} \frac{|y_i - ax_i - b|}{\sqrt{1 + a^2}}.$$

La quantité la plus manipulable est la deuxième. C'est avec celle-ci que nous allons travailler. Retenons que ce n'est pas le seul choix possible. Remarquons aussi la chose suivante : quelle que soit la quantité que vous cherchez à minimiser la droite que vous obtenez est sans intérêt si le nuage de points est très mal approché par une droite, correspond à la droite contenant le nuage de points si les points du nuage sont alignés.

Chercher la droite la plus proche du nuage de points au sens de F_2 c'est chercher a et b tels que F_2 soit minimale. Il s'agit donc de minimiser une fonction de deux variables. On peut attaquer ce type de problème comme pour les fonctions d'une variable en dérivant. Ici on peut dériver par rapport à a ou à b. Comme dans le cas des fonctions d'une variables, trouver les points où la dérivée s'annule n'est pas suffisant pour affirmer que la fonction est minimale en un point donné (par exemple la dérivée est nulle aussi en un point où la fonction est maximale). Nous emploierons donc une autre méthode (plus élémentaire) pour obtenir une réponse complète.

$$\frac{\partial F_2}{\partial a} = \sum_{i=1}^{n} -2x_i(y_i - ax_i - b)$$

$$\frac{\partial F_2}{\partial b} = \sum_{i=1}^{n} -2(y_i - ax_i - b)$$

Pour que ces deux dérivées soient nulles, il faut donc qu'on ait

$$\sum_{i=1}^{n} -2x_i(y_i - ax_i - b) = 0$$

$$\sum_{i=1}^{n} -2(y_i - ax_i - b) = 0$$

soit

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} y_i = a\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} b$$

En divisant par n on obtient les deux conditions suivantes :

$$\bar{xy} = a\overline{x^2} + b\bar{x} \text{ et } \bar{y} = a\bar{x} + b$$

En multipliant la deuxième par $\bar{x}$ et en la soustrayant à la première on obtient

$$\bar{xy} - \bar{x}\bar{y} = a(\overline{x^2} - \bar{x}^2) \text{ et } \bar{y} = a\bar{x} + b$$

soit

$$Cov(x, y) = aVar(x) = a\sigma_x^2 \text{ et } \bar{y} = a\bar{x} + b$$

Si la série x n'est pas constante alors le système a une solution unique

$$a = \frac{Cov(x, y)}{\sigma_x^2},\ b = \bar{y} - \frac{Cov(x, y)}{\sigma_x^2}\bar{x}.$$

Remarquons que si $\sigma_x^2 = 0$ alors x est constante : le nuage de point est sur une droite verticale ; cette droite a pour équation $x = cste$ (pas une équation de la forme $y = ax + b$).

Dans le cas où $\sigma_x^2 \neq 0$ les valeurs trouvées de a et b sont effectivement celles qui minimisent F_2. Nous allons maintenant le montrer.

$$\begin{aligned}
F_2(a,b) &= \sum_{i=1}^{n}(y_i - ax_i - b)^2 \\
&= \sum_{i=1}^{n}(y_i - ax_i - b - \bar{y} + a\bar{x} + b + \bar{y} - a\bar{x} - b)^2 \\
&= \sum_{i=1}^{n}((y_i - \bar{y}) - a(x_i - \bar{x}) + \bar{y} - a\bar{x} - b)^2 \\
&= \sum_{i=1}^{n}((y_i - \bar{y}) - a(x_i - \bar{x}))^2 + \sum_{i=1}^{n}(\bar{y} - a\bar{x} - b)^2 \\
&\qquad +2(\bar{y} - a\bar{x} - b)\sum_{i=1}^{n}((y_i - \bar{y}) - a(x_i - \bar{x}))
\end{aligned}$$

La dernière égalité étant obtenue en développant le carré à l'intérieur du signe $\sum$. Comme

$$\sum_{i=1}^{n}(y_i - \bar{y}) = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n}\bar{y} = n\bar{y} - n\bar{y} = 0$$

(et une égalité analogue pour x) le dernier des trois termes précédents est nul. Transformons maintenant le premier de ces trois termes :

$$\begin{aligned}
\sum_{i=1}^{n}((y_i - \bar{y}) - a(x_i - \bar{x}))^2 &= \sum_{i=1}^{n}(y_i - \bar{y})^2 + a^2\sum_{i=1}^{n}(x_i - \bar{x})^2 - 2a\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) \\
&= n\sigma_y^2 + na^2\sigma_x^2 - 2naCov(x,y) \\
&= n\left(\sigma_y^2 + \left(a\sigma_x - \frac{Cov(x,y)}{\sigma_x}\right)^2 - \frac{Cov(x,y)^2}{\sigma_x^2}\right)
\end{aligned}$$

On a donc obtenu l'égalité

$$F_2(a,b) = n\sigma_y^2 - n\frac{Cov(x,y)^2}{\sigma_x^2} + n\left(a\sigma_x - \frac{Cov(x,y)}{\sigma_x}\right)^2 + n(\bar{y} - a\bar{x} - b)^2.$$

Le deuxième terme contient une partie qui ne dépend ni de a ni de b : $n\sigma_y^2 - n\frac{Cov(x,y)^2}{\sigma_x^2}$ à laquelle on ajoute une somme de carrés. Les carrés étant positifs on peut dire que

$$F_2(a,b) \geq n\sigma_y^2 - n\frac{Cov(x,y)^2}{\sigma_x^2}.$$

De plus il y a un seul choix de a et b qui annule les carrés que l'on ajoute à cette quantité pour obtenir $F_2(a,b)$. Ce sont les nombres a et b vérifiant :

$$a\sigma_x - \frac{Cov(x,y)}{\sigma_x} = 0 \text{ et } \bar{y} - a\bar{x} - b = 0.$$

En conclusion les valeurs de a et b qui rendent $F_2(a,b)$ minimale sont donc données par

$$a = \frac{Cov(x,y)}{\sigma_x^2} \text{ et } \bar{y} - a\bar{x} - b = 0.$$

La valeur minimale de F_2 est $n\left(\sigma_y^2 - \frac{Cov(x,y)^2}{\sigma_x^2}\right)$. À quelle condition cette valeur est-elle nulle ?

Proposition 2.1. *Soient $x = (x_1, \ldots, x_n)$ et $y = (y_1, \ldots, y_n)$ deux suites de n nombres réels. On a*

$$|Cov(x,y)| \leq \sigma_x \sigma_y,$$

avec égalité si et seulement si les points de coordonnées (x_i, y_i) sont alignés.

Démonstration Pour obtenir l'inégalité il suffit d'appliquer l'inégalité de Cauchy-Swcharz aux deux séries $x - \bar{x}$ et $y - \bar{y}$. On obtient

$$\left|\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\right| \leq \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2},$$

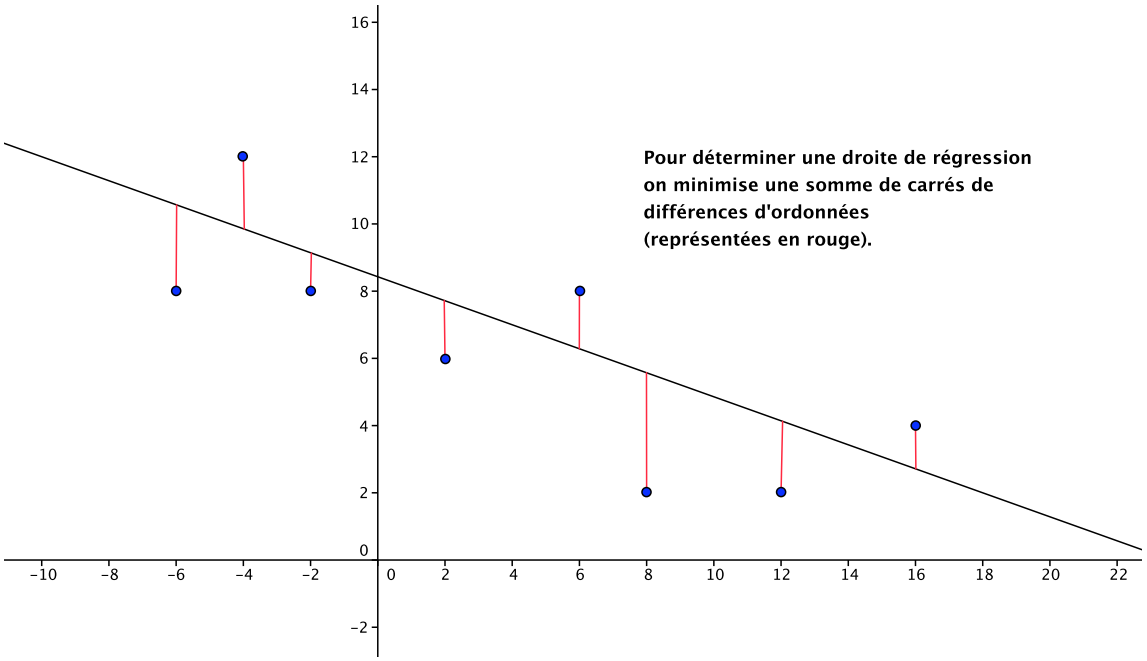
qui est exactement ce qu'on veut montrer. Le théorème affirme que l'inégalité n'est une égalité que s'il existe λ tel que, pour tout i on ait, $y_i - \bar{y} = \lambda(x_i - \bar{x})$ soit $y_i = \lambda x_i + \bar{y} - \lambda\bar{x}$. Cela signifie que les points (x_i, y_i) sont tous sur la droite d'équation $y = \lambda x + \bar{y} - \lambda\bar{x}$. □

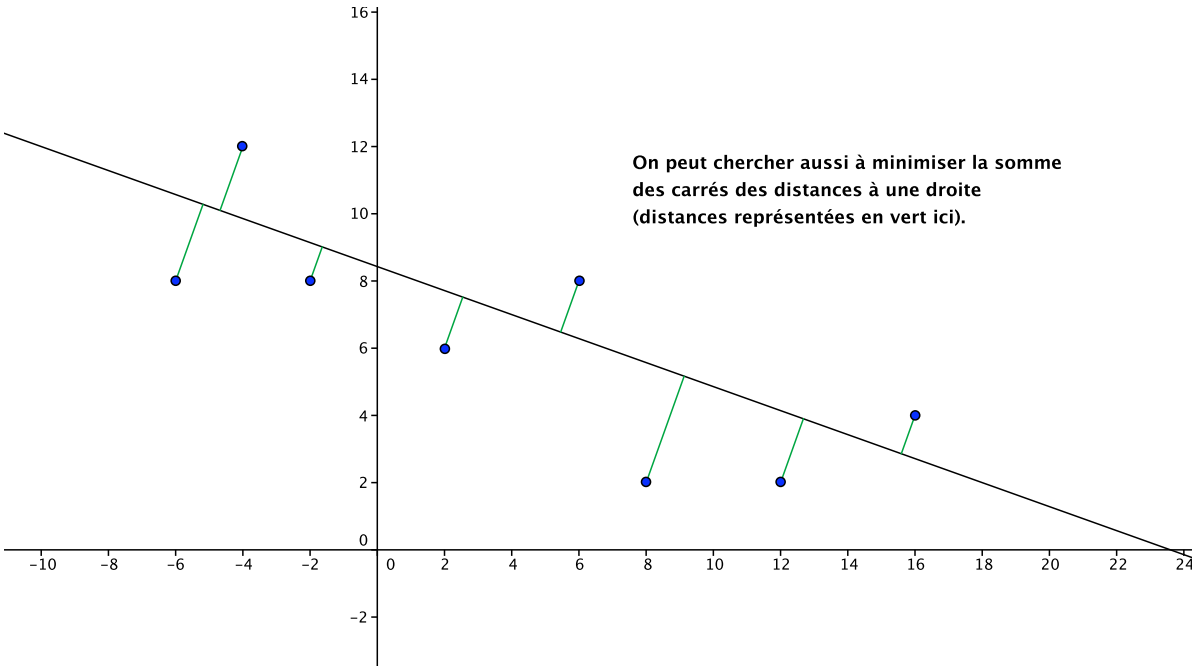
Quelques exemples. Différentes formes de nuages de points.

2.2.2 Un premier pas vers l'analyse en composantes principales

En grande dimension (quand on associe de nombreux caractères à chaque individu de la population) la représentation du nuage de points est impossible. On cherche alors à en obtenir une image en dimension 2 la moins déformée possible. Cette baisse de dimension est ce que nous avons fait avec la droite de régression : trouver la droite qui représente aussi bien que possible une nuage de points dans le plan (la définition de « aussi bien que possible »peut varier ; nous avons choisi une méthode des moindres carrés). D'autres critères peuvent être utilisés (nous avons par exemple vu qu'existaient deux droites de réfression). Dans tous les cas il convient d'accompagner le calcul de ces approximations d'un indicateur de la qualité de l'approximation obtenue (le coefficient de corrélation dans le cas de la droite de régression). Nous allons présenter une autre façon de procéder qui est très employée en dimension supérieure sous le nom d'analyse en composantes principales. En pratique cette méthode permet d'obtenir la meilleure image plane d'un nuage de points placé dans un espace de grande dimension. Elle n'est donc pas appliquée en dimension 2. Ce qui suit est une présentation de la méthode sur un cas simple.

Nous cherchons à minimiser, non plus des sommes de différences d'ordonnées ou d'abscisse, mais la somme des carrés des distances à une droite (et cherchons une droite qui minimise cette somme).





Nous allons le faire en nous limitant aux droites qui passent par le point moyen du nuage de points. La somme des carrés des distances au point point moyen est

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 + (y_i - \bar{y})^2.$$

Grâce au théorème de Pythagore on peut écrire cette somme comme la somme des carrés des distances des points à leurs projetés orthogonaux sur la droite plus la somme des

carrés des distances de ces projetés au point moyen. Minimiser la somme des carrés des distances des points à la droite revient donc à maximiser la somme des carrés des distances des projetés au point moyen. Soit $(\cos(t), \sin(t))$ un vecteur directeur unitaire de la droite. La somme à maximiser (on cherche t pour que cette somme soit maximale) est :

$$\sum_{i=1}^{n}(\cos(t)(x_i - \bar{x}) + \sin(t)(y_i - \bar{y}))^2.$$

Réécrivons cette somme :

$$\begin{aligned}
&\sum_{i=1}^{n}(\cos(t)(x_i - \bar{x}) + \sin(t)(y_i - \bar{y}))^2 \\
&= \sum_{i=1}^{n}\cos^2(t)(x_i - \bar{x})^2 + \sum_{i=1}^{n}\sin^2(t)(y_i - \bar{y})^2 + 2\sum_{i=1}^{n}\sin(t)\cos(t)(x_i - \bar{x})(y_i - \bar{y}) \\
&= n\left(\cos^2(t)\sigma_x^2 + \sin^2(t)\sigma_y^2 + 2\sin(t)\cos(t)cov(x,y)\right).
\end{aligned}$$

Il suffit maintenant d'étudier les variations de cette fonction de t pour déterminer ses extrema (oublions le facteur n qui n'a aucune importance pour l'étude des variations). Sa dérivée vaut

$$-2\sin(t)\cos(t)\sigma_x^2 + 2\sin(t)\cos(t)\sigma_y^2 + 2\cos(2t)cov(x,y) = \sin(2t)(\sigma_y^2 - \sigma_x^2) + 2\cos(2t)cov(x,y).$$

Cette dérivées est nulle si

$$\tan(2t) = -\frac{2cov(x,y)}{\sigma_y^2 - \sigma_x^2},$$

à condition que $\sigma_y^2 - \sigma_x^2$ ne soit pas nul. Cette égalité définit deux nombres t entre $-\pi/2$ et $\pi/2$ où la fonction qui nous intéresse est minimale et maximale (différentes possibilités suivant les signes). Si la covariance $cov(x,y)$ et la différence $\sigma_y^2 - \sigma_x^2$ sont nulles alors la fonction est constante. Si la différence $\sigma_y^2 - \sigma_x^2$ est nulle mais la covariance ne l'est pas alors la valeur maximale est atteinte pour t égal à $\pi/4$ si la covariance est positive, $-\pi/4$ si la covariance est négative.

Pour neutraliser les effets dûs au choix des unités il est fréquent de réduire et normaliser les données x et y. On se ramène alors au nuage de points défini par

$$\left(\frac{x_i - \bar{x}}{\sigma_x}, \frac{y_i - \bar{y}}{\sigma_y}\right)$$

pour lequel les variances apparaissant dans le calcul précédent sont toutes deux égales à 1 et la covariance vaut $\rho(x,y)$ le coefficient de corrélation de x et y. Ce que nous avons vu est que (si $\rho(x,y) < 0$ par exemple) la droite minimisant la somme des carrés distances à la droite est la deuxième bissectrice. Autrement dit on approche le nuage de points par la droite d'équation

$$\frac{y - \bar{y}}{\sigma_y} = -\frac{x - \bar{x}}{\sigma_x}.$$

Revenant au nuage de points de départ cela donne la droite

$$y = -\frac{\sigma_y}{\sigma_x}x + \frac{\sigma_y}{\sigma_x}\bar{x} - \bar{y}.$$

Nous obtenons ainsi une quatrième droite approchant le nuage de point (après la droite de régression de y sur x, la droite de régression de x sur y, la droite minimisant la somme des carrés des distances). L'inégalité de Cauchy-Schwarz donne ici ($\rho(x, y) < 0$)

$$0 > \rho(x, y) > -\sigma_x\sigma_y.$$

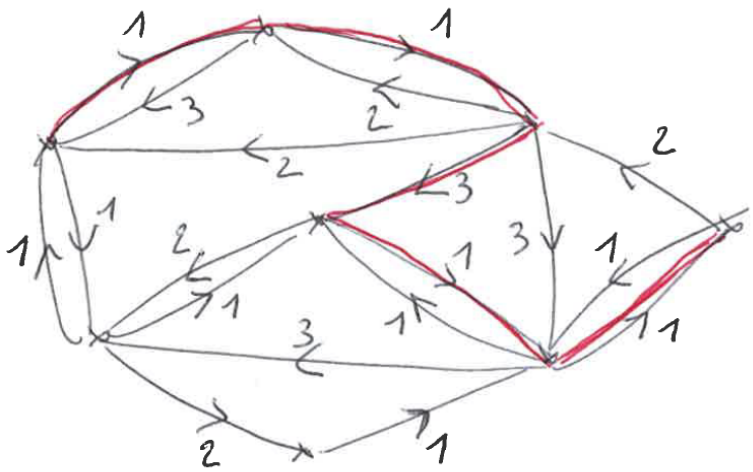
On en déduit l'encadrement

$$\frac{\sigma_y^2}{\rho(x, y)} < -\frac{\sigma_y}{\sigma_x} < \frac{\rho(x, y)}{\sigma_x^2}.$$

Autrement dit la quatrième droite dont nous avons parlé est entre les deux droites de régression.

2.3 L'algorithme de Dijkstra

Ici on cherche à trouver le chemin le plus court dans un graphe et dans le temps le plus court possible. On se donne un graphe dont les flèches ont un poids (ou une longueur) donné par un nombre réel strictement positif. Un chemin dans ce graphe est une suite de flèches qui se suivent. On appelle longueur d'un chemin la somme des longueurs des flèches qui composent ce chemin.



Par exemple le chemin rouge tracé ci-dessus a pour longueur 1+1+3+1+1=7. On cherche à déterminer des chemins de longueurs minimales dans un tel graphe. Pour les graphes de grandes tailles, il est impossible de calculer les longueurs de tous les chemins joignant

deux points donnés. Une méthode efficace de calcul de chemins les plus courts est donnée par l'algorithme de Dijkstra.

Notations

n : nombres de sommets du graphes

$s_i : i = 1, \ldots, n$, les sommets du graphe

l_{ij} : longueur de la flèche joignant s_i à s_j (par convention $l_{ij} = +\infty$ si aucune arête ne joint s_i à s_j).

On se donne un sommet D. L'algorithme fournit la liste de tous les chemins de longueurs minimales de D aux autres sommets du graphe.

Initialisation

* $\mathcal{Z} = \{D\}$
* $\mathcal{C} = \emptyset$
* Soit i_0 tel que $D = s_{i_0}$. Pour i tel que $s_i \neq D$ $(i \neq i_0)$ $\lambda_i = l_{i_0 i}$, $C_i = (D)$.

Boucle

Tant que $\mathcal{Z}$ ne contient pas tous les sommets du graphe faire :

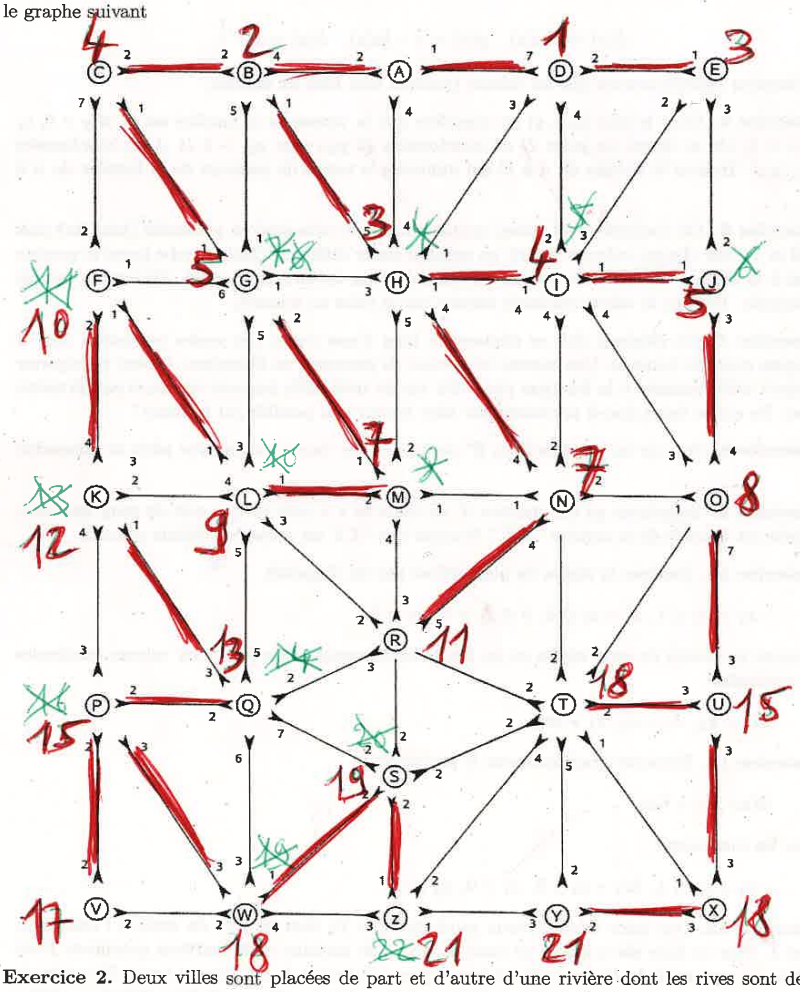
* Trouver $\lambda_{i_m} = \min_{i \,:\, s_i \notin \mathcal{Z}} \lambda_i$
* Ajouter s_{i_m} à C_{i_m}
* Ajouter s_{i_m} à $\mathcal{Z}$
* Ajouter $(C_{i_m}; \lambda_{i_m})$ à $\mathcal{C}$
* Pour i tel que $s_i \notin \mathcal{Z}$ faire : si $\lambda_i > \lambda_{i_m} + l_{i_m i}$
  + remplacer λ_i par $\lambda_{i_m} + l_{i_m i}$,
  + remplacer C_i par C_{i_m}.

À la fin $\mathcal{C}$ contient l'ensemble des chemins les plus courts de D à tous les autres sommets du graphes avec les longueurs correspondantes. Un exemple avec les traces de calculs de l'algorithme de Dijkstra :

Licence MiaSHS OPTIMISATION *2018-2019*

Feuille 1

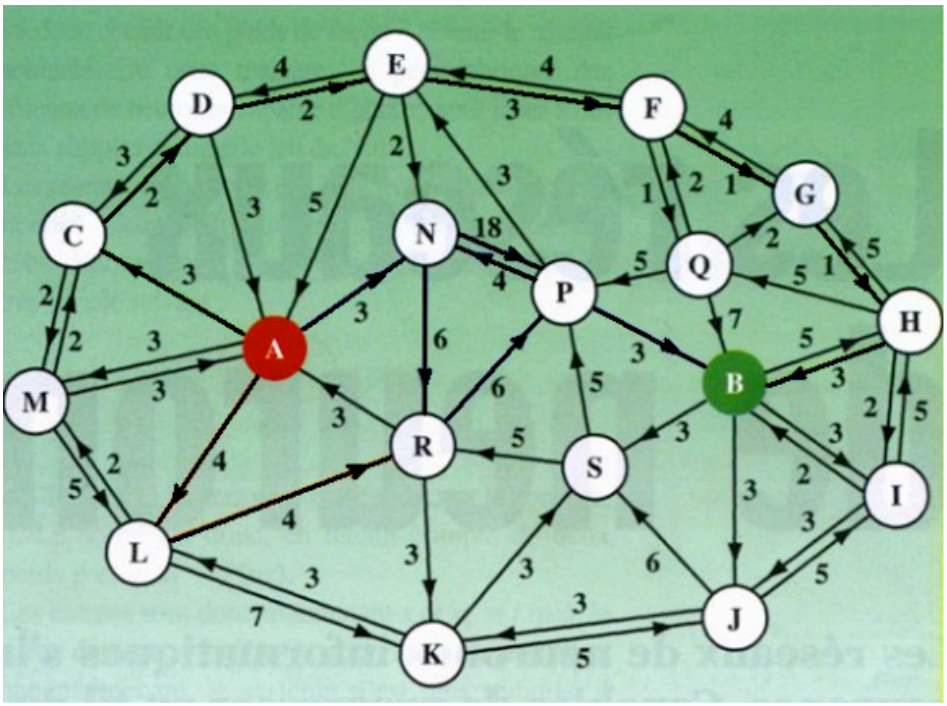
Exercice 1. Utiliser l'algorithme de Dijkstra pour trouver le plus court chemin de A à Z dans le graphe suivant



Exercice 2. Deux villes sont placées de part et d'autre d'une rivière dont les rives sont des droites parallèles. Où placer un pont (perpendiculaire aux rives) pour que la distance routière séparant les deux villes soit minimale ?

1

Un exemple pour s'entraîner :



2.4 L'algorithme de Héron

Le nombre $\sqrt{2}$ n'est pas rationnel. Supposons qu'il le soit. Alors il existe p et q deux nombres entiers premiers entre eux tels que $\sqrt{2} = p/q$. On a alors $2q^2 = p^2$. On en déduit que 2 divise p i.e. qu'on peut écrire $2p'$. On a alors $2q^2 = 2^3 p'^2$ soit $q^2 = 2p'^2$. Le nombre q est donc pair lui aussi. Contradiction car on avait supposé que p et q étaient premiers entre eux.

La racine cubique de 2 n'est pas rationnelle non plus. Supposons que $2^{1/3}$ soit rationnelle. Alors il existe p et q deux nombres entiers premiers entre eux tels que $2^{1/3} = p/q$. On a alors $2q^3 = p^3$. On en déduit que 2 divise p i.e. qu'on peut écrire $2p'$. On a alors $2q^3 = 2^3 p'^3$ soit $q^3 = 4p'^2$. Le nombre q est donc pair lui aussi. Contradiction car on avait supposé que p et q étaient premiers entre eux.

Comment obtenir des valeurs approchées rationnelles de $\sqrt{2}$? On peut procéder par dichotomie. Le nombre $\sqrt{2}$ est entre 1 et 2 car $1^2 < 2 < 2^2$. On calcule $1,5^2 = 2,25$ donc $1 < \sqrt{2} < 1,5$; puis $1,25^2 = 1,5625$ donc $1,25 < \sqrt{2} < 1,5$; etc... à chaque nouveau calcul on divise par 2 la longueur d'un intervalle dans lequel se trouve $\sqrt{2}$. Très bien. On peut faire mieux.

Partons d'un rectangle de côtés 1 et 2 et essayons de le rendre plus carré. On fabrique un rectangle de côtés $\frac{1+2}{2}$ et $2/\frac{1+2}{2}$ soit $3/2$ et $4/3$; on obtient ainsi un rectangle d'aire 2 qui est plus proche d'un carré. On recommence en prenant la moyenne arithmétique des longueurs des deux côtés $(3/2 + 4/3)/2 = 17/12$ et $2/(17/12) = 24/17$ et on continue de la même façon $(17/12 + 24/17)/2 = 577/408$ etc... Regardons le nombre que nous

avons obtenu 1,4142157. La valeur approchée de racine de 2 à 10^{-7} près est 1,4142136. En trois étapes nous avons obtenu cinq bonnes décimales. Faisons un calcul supplémentaire : $(577/408 + 816/577)/2$ est à peu près égale à 1,4142135623746900918718. Une valeur approchée de $\sqrt{2}$ avec 21 décimales est 1,4142135623730951454746 : nous avons donc obtenu onze bonne décimales. La méthode d'approximation de $\sqrt{2}$ que nous sommes en train de décrire est l'algorithme de Héron : à chaque nouvelle étape on double le nombre de bonnes décimales obtenues.

Pour obtenir une précision de 10^{-100} il faudrait donc faire quatre étapes supplémentaires (alors qu'il en faudrait plus de trois cents avec la méthode de dichotomie).

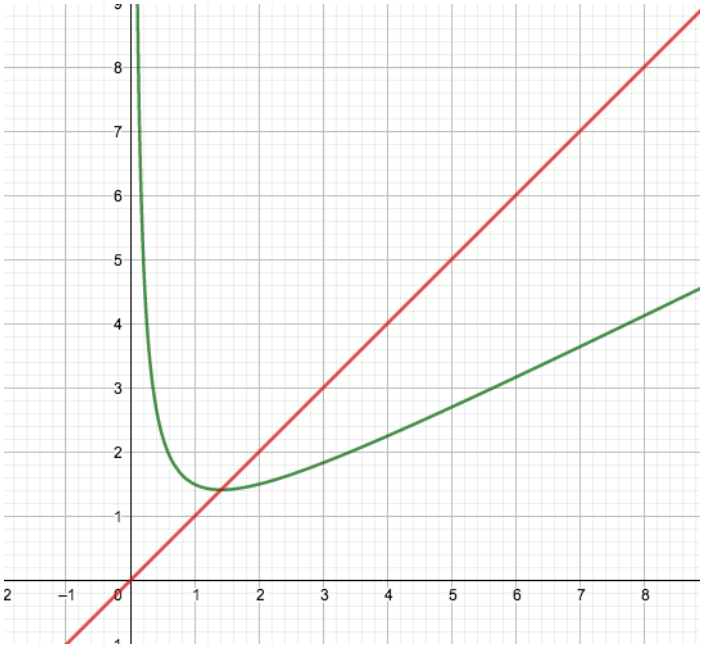
Essayons de comprendre pour quoi la vitesse de convergence est si grande pour l'algorithme de Héron. Appelons a_n la valeur approchée obtenu après le $n$ème calcul. On a

$$a_0 = 1, a_{n+1} = (a_n + 2/a_n)/2.$$

C'est une suite récurrente définie au moyen de la fonction

$$f(x) = (x + 2/x)/2.$$

Traçons le graphe de cette fonction :



La seule solution de $f(x) = x$ (pour $x > 0$) est $x = \sqrt{2}$ (abscisse de l'intersection du graphe de f avec la première bissectrice (tracée en rouge ici). Ce qui explique la grande vitesse de convergence de notre suite vers $\sqrt{2}$ est le fait que le graphe de f a une tangente horizontale au points d'abscisse $\sqrt{2}$:

$$f'(x) = (1 - 2/x^2)/2 \; ; \; f'(\sqrt{2}) = 0.$$

Nous y reviendrons quand nous parlerons de l'algorithme de Newton.

Bien souvent lorsqu'on cherche un point où une fonction atteint un extremum on cherche les points où sa (ou ses) dérivée(s) s'annule(nt). Il est le plus souvent impossible de résoudre explicitement l'équation ou le système d'équations obtenu : il est alors très important de disposer d'algorithmes très rapides donnant les solutions recherchées avec précision. *On cherche à minimiser le temps (ou le nombre d'opérations, ou l'énergie,...) nécessaire pour résoudre un problème.*

2.5 La méthode hongroise pour les problèmes d'affectation

Ici on cherche à résoudre un problème d'affectation optimale le plus rapidement possible. Imaginons le problème suivant. On a quatre tâches à effectuer et quatre machines pour faire le travail. Mais les coûts pour faire ces tâches sont différents pour chaque machine. On veut affecter les tâches de sorte que le coût total soit le plus petit possible. Comment résoudre simplement cette question ? On peut faire la liste de toutes les affectations possibles, calculer les coûts correspondants et adopter les affectations donnant le plus petit total. Pour deux machines et deux tâches deux possibilités sont à considérer, six pour trois machines et trois tâches, puis 24, 120... $n!$ possibilités pour n machines et n tâches. Si n est grand cette méthode est impraticable. La méthode hongroise permet de résoudre le problème avec un nombre d'opération de l'ordre de n^3.

Décrivons cette méthode sur quelques exemples.

Remarquons que si dans le tableau tous les nombres sont positifs ou nuls et qu'on peut trouver cinq zéros sur cinq lignes et cinq colonnes différentes alors en décidant l'affectation correspondante on obtient un coût total nul qui est nécessairement optimal. On indique en rouge une affectation optimale dans le tableau suivant.

	Tâche 1	Tâche 2	Tâche 3	Tâche 4	Tâche 5
Machine 1	2	4	0	1	0
Machine 2	2	3	1	0	3
Machine 3	4	0	5	3	6
Machine 2	0	1	4	3	5
Machine 5	5	2	0	2	0

Parfois (comme ici) plusieurs affectations optimales existent.

	Tâche 1	Tâche 2	Tâche 3	Tâche 4	Tâche 5
Machine 1	2	4	0	1	0
Machine 2	2	3	1	0	3
Machine 3	4	0	5	3	6
Machine 2	0	1	4	3	5
Machine 5	5	2	0	2	0

Considérons maintenant un exemple sans zéro.

	Tâche 1	Tâche 2	Tâche 3	Tâche 4	Tâche 5
Machine 1	2	4	8	9	9
Machine 2	2	3	9	9	3
Machine 3	4	10	5	3	6
Machine 4	2	6	4	3	5
Machine 5	5	2	8	2	9

Remarquons qu'on ne modifie pas le problème d'affectation si on ajoute ou retranche un même nombre aux coûts associés à une même machine ou bien à une même tâche. Ceci permet de faire apparaître des zéros dans le tableau (au moins un par ligne et un par colonne) en soustrayant à chaque ligne le plus petit nombre de cette ligne, puis en faisant de même pour les colonnes. Ici en enlevant 2 à la première ligne, 2 à la deuxième, 3 à la troisième, 2 à la quatrième et 2 à la cinquième on obtient le tableau suivant (je n'indique plus les numéros de tâches et de machines).

0	2	6	7	7
0	1	7	7	1
1	7	2	0	3
0	4	2	1	3
3	0	6	0	7

On procède de la même façon pour les colonnes (seules la troisième et la cinquième sont concernées).

0	2	4	7	6
0	1	5	7	0
1	7	0	0	2
0	4	0	1	2
3	0	4	0	6

On a obtenu une affectation optimale.

0	2	4	7	6
0	1	5	7	0
1	7	0	0	2
0	4	0	1	2
3	0	4	0	6

Revenons au problème de départ. L'unique affectation optimale est indiquée en rouge dans

le tableau suivant.

	Tâche 1	Tâche 2	Tâche 3	Tâche 4	Tâche 5
Machine 1	2	4	8	9	9
Machine 2	2	3	9	9	3
Machine 3	4	10	5	3	6
Machine 4	2	6	4	3	5
Machine 5	5	2	8	2	9

Et le coût minimal est $2 + 3 + 3 + 4 + 2 = 14$.

En général les opérations décrite sur l'exemple précédent ne suffisent pas à donner une réponse. Considérons un nouvel exemple.

7	2	1	9	4
9	6	9	5	5
8	8	3	1	8
7	9	4	2	2
4	3	7	4	8

Enlevons à chaque ligne son plus petit élément.

6	1	0	8	3
4	1	4	0	0
7	7	2	0	7
5	7	2	0	0
1	0	4	1	5

Faisons de même sur les colonnes (seule la première est concernée).

5	1	0	8	3
3	1	4	0	0
6	7	2	0	7
4	7	2	0	0
0	0	4	1	5

On ne peut pas trouver cinq zéros sur cinq lignes et cinq colonnes différentes. Une façon de s'en rendre compte est de remarquer qu'on peut tracer des traits horizontaux ou verticaux passant par tous les zéros obtenus en nombre strictement inférieur à cinq (ici quatre, trois colonnes une ligne).

5	1	~~0~~	~~8~~	~~3~~
3	1	~~4~~	~~0~~	~~0~~
6	7	~~2~~	~~0~~	~~7~~
4	7	~~2~~	~~0~~	~~0~~
~~0~~	~~0~~	~~4~~	~~1~~	~~5~~

Formalisons le problème.

Définition 2.2. *Soit A une matrice $n \times n$. Disons que deux coefficients de cette matrice sont indépendants s'ils ne sont ni sur la même ligne, ni sur la même colonne.*

Proposition 2.3. *Le nombre maximal de 0 indépendants dans une matrice est égal au nombre minimal de traits qu'il faut pour recouvrir tous les 0 de la matrice.*

Démonstration S'il est possible de recouvrir tous les 0 avec k traits alors il ne peut pas y avoir plus de k 0 indépendants (car si considère $k+1$ 0, deux au moins sont sur un même trait, donc ne sont pas indépendants). Le nombre maximal de 0 indépendants dans une matrice est donc inférieur ou égal au nombre minimal de traits qu'il faut pour recouvrir tous les 0 de la matrice. Il reste à montrer que si le nombre maximal de 0 indépendants est k alors il est possible de recouvrir tous les 0 avec k traits. Donnons-nous k 0 indépendants. Disons que ces 0 sont noirs. Les autres 0 ne sont pas indépendants de ces k 0 (sinon k ne serait pas le nombre maximal de 0 indépendants). Classons ces autres 0 en deux couleurs : les rouges qui sont sur la même ligne qu'un noir et aussi sur la même colonne qu'un (autre) noir ; les verts qui sont sur la même ligne qu'un noir, ou bien sur la même colonne qu'un noir (mais pas les deux). Supposons qu'il existe un 0 vert. Traçons à partir de ce 0 vert tous les chemins dans la matrice en joignant des zéros entre eux quand ils sont sur la même ligne ou la même colonne. Admettons que le 0 de départ soit sur la même ligne qu'un 0 noir. Alors il est le seul 0 de sa colonne. Tous les chemins construits à partir de ce zéro commence par un morceau de ligne. Supposons que m 0 noirs soient atteints par de tels chemins. Alors tous les 0 situés sur les colonnes de ces m 0 sont aussi sur les lignes de ces m 0. Pourquoi ? Parce que sinon on pourrait les atteindre en prolongeant un chemin déjà tracé par un morceau de colonne. En enlevant les m lignes et les m colonnes des m 0 noirs atteints on obtient un tableau plus petit avec $k-m$ 0 noirs indépendants (et ce nombre est maximal). Les m lignes enlevées contiennent tous les 0 dépendants d'un des m atteints. Pour cela il faut vérifier qu'un chemin ne peut pas se terminer par un morceau de colonne avec un bout 0 vert. Ce n'est pas le cas : si on avait une suite 0000000 par exemple, on pourrait la transformer en 0000000 et on aurait un 0 noir de plus ce qui contredirait la maximalité de k. On raisonne de la même façon si le 0 vert de départ est sur la même colonne qu'un 0 noir (en échageant ligne et colonne). S'il n'y a pas de 0 vert on procède de même en partant d'un sommet noir (avec des chemins uniquement composés de 0 noirs et rouges en alternance). On obtient des cycles. La réunion des lignes contenant tous les 0 noirs atteints contient aussi tous les 0 situés sur les colonnes contenant ces 0 noirs atteints (cas le plus simple : un 0 noir isolé (aucun autre 0 sur sa ligne ni sur sa colonne)). En enlevant ces lignes et ces colonnes on est ramené au même problème avec un tableau plus petit. On peut donc procéder par récurrence sur la taille des tableaux considérés. L'initialisation (tableau 1×1) ne pose pas de problème ! □

Dans ce cas la méthode hongroise indique de procéder de la façon suivante : on enlève la plus petite des valeurs non couvertes (ici les valeurs qui ne sont pas barrées en bleu) aux lignes qui ne sont pas couvertes, et on ajoute cette même valeur aux colonnes couvertes.

Ici on obtient

4	0	0	8	3
2	0	4	0	0
5	6	2	0	7
3	6	2	0	0
0	0	5	2	6

Une autre façon de décrire les opérations précédentes : on enlève la plus petite des valeurs non couvertes aux valeurs non couvertes et on ajoute cette valeur aux coefficients situés aux intersections des lignes et des colonnes couvertes.

Ces opérations ne changent pas le problème d'affectation (puisqu'on ajoute ou retranche la même quantité à des lignes ou colonnes entières). Dans notre exemple nous avons résolu notre problème d'affectation.

4	0	0	8	3
2	0	4	0	0
5	6	2	0	7
3	6	2	0	0
0	0	5	2	6

S'il avait encore été possible de recouvrir tous les zéros avec moins de cinq lignes ou colonnes, on aurait réitéré les opérations précédentes jusqu'à ce qu'il ne soit plus possible de recouvrir les zéros obtenus avec moins de cinq lignes ou colonnes (c'est-à-dire jusqu'à avoir un affectation optimale donnée par cinq zéros placés sur cinq colonnes différentes et sur cinq lignes différentes).

2.6 Exemples de problèmes faisant intervenir le hasard

Une compagnie aérienne a N places dans ses avions qu'elle vend à un prix P. Ses clients ne se présentent pas à l'embarquement avec probabilité p. Elle décide de vendre plus de places que ce dont elle dispose avec la politique suivante : si un client ne se présente pas il perd le prix de son billet, si un client se présente mais n'a pas de place la compagnie l'indemnise à hauteur de λP ($\lambda > 1$). La compagnie souhaite qu'avec une probabilité supérieure à 95% son chiffre d'affaire soit supérieur à NP. Combien doit-elle vendre de places au maximum (on supposera que N est grand et on utilisera une approximation par la loi normale) ? Est-ce ce critère qu'utiliserait une compagnie pour décider comment elle fixe ses prix ? Application numérique : $N = 5000$, $p = 0,1$, $\lambda = 4$.

Une compagnie d'assurance se propose d'assurer 100000 clients contre le vol. Les sommes en euros (la plupart du temps nulles) $X_1, ..., X_{100000}$ qu'aura à rembourser chaque année la compagnie aux clients sont des v.a. indépendantes d'espérance 75 et d'écart type 750. Quelle somme cette compagnie d'assurance doit-elle faire payer à chaque client par an pour que ses frais évalués à 1,5 millions d'euros soient couverts avec une probabilité supérieure

ou égale à 0,999 ? (On utilisera sans justification l'approximation par la loi normale.)

On cherche à minimiser un prix (pour être compétitif) tout en ne risquant pas la faillite. [5]

La somme que la compagnie doit verser à ses assurés est $S = X_1 + \ldots + X_{100000}$. Il faut y ajouter ses frais $1,5.10^6$. Ce qu'elle perçoit est $100000c$ ou c est le montant de la cotisation de chaque assuré. Les cotisations des assurés couvrent les dépenses si

$$100000c \geq X_1 + \ldots + X_{100000} + 1,5.10^6.$$

La question posée devient : comment doit on fixer c pour que

$$\mathbb{P}(100000c \geq X_1 + \ldots + X_{100000} + 1,5.10^6) \geq 0,999 \ ?$$

Nous allons réécrire cette probabilité de façon à pouvoir appliquer le théorème limite central.

$$\begin{aligned}
&\mathbb{P}(100000c \geq X_1 + \ldots + X_{100000} + 1,5.10^6) \\
&\quad = \mathbb{P}(X_1 + \ldots + X_{100000} \leq 100000c - 1,5.10^6) \\
&\quad = \mathbb{P}(\frac{X_1 + \ldots + X_{100000} - 100000.75}{\sqrt{100000}.750} \leq \frac{100000c - 1,5.10^6 - 100000.75}{\sqrt{100000}.750}) \\
&\quad \simeq \mathbb{P}(Y \leq \frac{100000c - 1,5.10^6 - 100000.75}{\sqrt{100000}.750}),
\end{aligned}$$

où Y suit la loi normale centrée réduite. Car d'après le théorème limite central la variable

$$\frac{X_1 + \ldots + X_{100000} - 100000.75}{\sqrt{100000}.750}$$

suit à peu près la loi normale centrée réduite. Notons Φ la fonction de répartition de la loi normale centrée réduite. On souhaite que

$$\Phi(\frac{100000c - 1,5.10^6 - 100000.75}{\sqrt{100000}.750}) \geq 0,999.$$

Cela est vérifié si

$$\frac{100000c - 1,5.10^6 - 100000.75}{\sqrt{100000}.750} \geq \Phi^{-1}(0,999),$$

c'est-à-dire si

$$c \geq 75 + \frac{\sqrt{100000}.750.\Phi^{-1}(0,999) + 1,5.10^6}{100000}.$$

3 Optimisation linéaire, convexes

Une partie du contenu de cette section est reprise d'un cours de Martin Grötschel (Lineare Optimierung (ADM II)) disponible sur internet http://www.zib.de/groetschel/teaching/materials.html.

3.1 Un exemple

Dans une raffinerie on décompose du pétrole brut en appliquant des procédés physiques ou chimiques afin d'obtenir de nouveaux produits. Ce qu'on obtient dépend du procédé employé. On admet qu'une raffinerie fournit trois types de composants : du pétrole lourd (noté S comme lourd), du pétrole moyen (M), du pétrole léger (L). Elle dispose de deux procédés différents dont les coûts (énergie, amortissement des machines, travail) et les résultats sont les suivants (pour dix unités de pétrole brut) :
le procédé 1, fournit 2 unités de L, 2 unités de M, 1 unité de L pour un coût de 3 unités monétaires ;
le procédé 2, fournit 1 unités de L, 2 unités de M, 4 unité de L pour un coût de 5 unités monétaires.

La raffinerie doit satisfaire une commande de 3 unités de S, 5 unités de M et 4 unités de L et souhaite le faire au coût le plus bas possible. On suppose que les deux procédés fonctionnent de manière indépendante et qu'ils peuvent utiliser une quantité quelconque de pétrole brut.

Notons x_1 la quantité de pétrole brut utilisée par le procédé 1 (en dizaine d'unité, $x_1 = 1,5$ signifie qu'on consomme 15 unités de brut par le procédé 1). On note de même x_2 la quantité consommée par l'utilisation du procédé 2.

Supposons qu'on utilise les procédés 1 et 2 avec des quantités de brut $10x_1$ et $10x_2$, alors les quantités produites sont :

$$2x_1 + x_2 \text{ pour S},\ 2x_1 + 2x_2 \text{ pour M},\ x_1 + 4x_2 \text{ pour L}.$$

Pour que la commande soit satisfaite il faut donc qu'on ait

$$2x_1 + x_2 \geq 3,\ 2x_1 + 2x_2 \geq 5,\ x_1 + 4x_2 \geq 4.$$

à ces contraintes il faut bien sûr ajouter le fait que x_1 et x_2 ne peuvent être négatives :

$$x_1 \geq 0,\ x_2 \geq 0.$$

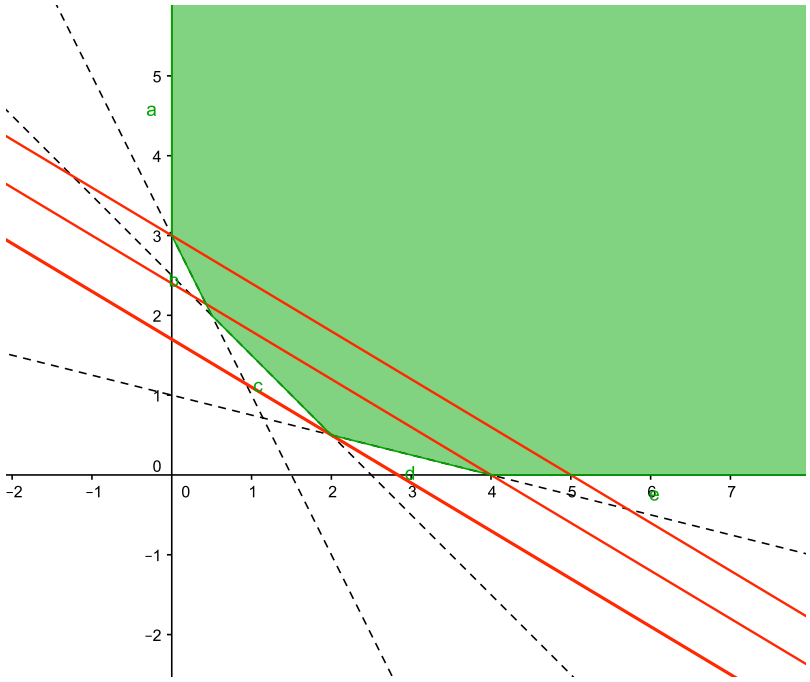
On cherche donc des valeurs de x_1 et x_2 qui respectent ces contraintes telles que $3x_1 + 5x_2$ soit le plus petit possible. C'est ce qu'on appelle un problème de programmation linéaire ou d'optimisation linéaire. On l'écrit de la façon suivante :

$$\min 3x_1 + 5x_2,$$

$$s.c.\ \ 2x_1 + x_2 \geq 3,\ 2x_1 + 2x_2 \geq 5,\ x_1 + 4x_2 \geq 4,\ x_1 \geq 0,\ x_2 \geq 0.$$

La fonction $3x_1 + 5x_2$ s'appelle la fonction objectif. Tout vecteur (x_1, x_2) respectant les contraintes s'appelle une solution admissible du problème. Ici on peut résoudre graphiquement le problème. On considère le plan muni d'un repère et on appelle x_1 et x_2 l'abscisse

et l'ordonnée d'un point. Chacune des contraintes indique que le points (x_1, x_2) est d'un côté d'un droite (qui délimite deux demi-plans). Par exemple $2x_1 + 2x_2 \geq 5$ signifie que le point (x_1, x_2) est au-dessus de la droite d'équation $2x_1 + 2x_2 = 5$. L'ensemble des points respectant les cinq contraintes est représenté en vert sur la figure ci-dessous.



Les lignes de niveau de la fonction objectif sont les ensembles $3x_1 + 5x_2 = t$. Ce sont des droites parallèles. Trois telles droites sont représentées en rouge ci-dessus. Ce qu'on cherche ce sont les point de l'ensemble vert de niveau minimal (pour la fonction objectif). Plus une droite rouge est haute, plus le niveau auquel elle correspond est élevé. On cherche donc la droite rouge le plus bas possible qui touche l'ensemble vert. Cette droite est représentée sur le schéma : elle passe par le point de coordonnées $(2, 1/2)$. On en déduit que le minimum recherché est $3.2 + 5.1/2 = 17/2$ atteint pour $x_1 = 2$ et $x_2 = 1/2$.

On peut dire que les dessins faits montrent que nous avons bien résolu notre problème. Mais si le nombre de variables concernées est plus grand (par exemple 333 ce qui est tout à fait possible) alors ce type de visualisation sera beaucoup plus difficile (car notre intuition géométrique en dimension 333 est très limitée). Il nous faut donc développer des méthodes systématiques valables en toute dimension qui ne reposent pas sur notre vision en dimension 2 ou 3.

Une notion naturelle en programmation linéaire est celle de programme dual. Introduisons trois variables positives ou nulles y_1, y_2, y_3, multiplions chacune des contraintes

$$2x_1 + x_2 \geq 3,\ 2x_1 + 2x_2 \geq 5,\ x_1 + 4x_2 \geq 4.$$

respectivement par y_1, y_2, y_3 et additionnons les. Nous obtenons

$$(2y_1 + 2y_2 + y_3)x_1 + (y_1 + 2y_2 + 4y_3)x_2 \geq 3y_1 + 5y_2 + 4y_3.$$

Si on suppose en plus que

$$2y_1 + 2y_2 + y_3 \leq 3 \text{ et } y_1 + 2y_2 + 4y_3 \leq 5,$$

alors le membre de gauche de cette inégalité est inférieure à la fonction objectif. La fonction objectif est donc supérieure à $3y_1+5y_2+4y_3$ dans ce cas. Posons alors le problème d'optimisation

$$\max 3y_1 + 5y_2 + 4y_3$$

$$s.c.\ 2y_1 + 2y_2 + y_3 \leq 3\ y_1 + 2y_2 + 4y_3 \leq 5\ y_1 \geq 0\ y_2 \geq 0\ y_3 \geq 0.$$

Le calcul que nous avons fait montre que si les contraintes des deux problèmes linéaires sont satisfaites alors toute valeur de la fonction objectif du deuxième problème est inférieure à toute valeur de la fonction objectif du premier problème. Conclusion : si on trouve une valeur commune aux deux fonctions objectif (les contraintes étant respectées) alors cette valeur commune est le minimum recherché du premier problème et le maximum recherché du deuxième problème. Ce deuxième problème est appelé le problème dual du premier.

Le langage des matrices fournit une manière concise d'écrire ces problèmes. Introduisons les notations suivantes :

$$A = \begin{pmatrix} 2 & 1 \\ 2 & 2 \\ 1 & 4 \end{pmatrix} \quad x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \quad c = \begin{pmatrix} 3 \\ 5 \end{pmatrix} \quad y = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} \quad b = \begin{pmatrix} 3 \\ 4 \\ 5 \end{pmatrix}.$$

Alors le premier problème s'écrit

$$\min\langle c, x\rangle\ s.c.\ Ax \geq b, x \geq 0,$$

où il faut comprendre qu'un vecteur est supérieur ou égal à un autre si chacune des coordonnées du premier est supérieure à la coordonnée correspondante du deuxième. Le deuxième problème s'écrit

$$\max\langle b, y\rangle\ s.c.\ {}^tAy \leq c, y \geq 0.$$

3.2 Un exemple de jeu matriciel

Considérons le jeu suivant : deux joueurs, A et B. A met dans son poing une pièce de 10 centimes ou une pièce de 20 centimes. B doit deviner ce que A a dans sa main. S'il devine, il gagne la pièce. Sinon il doit donner une pièce de la même valeur à A. On peut décrire dans une matrice les gains réalisés par A en fonction des différentes actions possibles :

	B choisit 10	B choisit 20
A choisit 10	-10	10
A choisit 20	20	-20

La matrice associée au jeu est

$$\begin{pmatrix} -10 & 10 \\ 20 & -20 \end{pmatrix}$$

Quelles stratégie doivent adopter les joueurs ? Si les joueurs jouent une seule fois, on ne sait pas très bien quoi répondre. On pourrait penser que A a intérêt à choisir 20 centimes pour gagner plus. Mais alors B pourrait suivre le même raisonnement et se dire que A a sans doute choisi 20 pour gagner plus et donc qu'il a lui même intérêt à choisir 20. Mais alors A peut faire toutes ces réflexions et choisir 10 pour cette raison. B peut très bien penser à tout ça aussi et choisir 10 pour être malin à un niveau supérieur. Alors A peut décider de choisir 20 parce que etc... Ce jeu de renvoi est sans fin. A ne doit pas se dire que, tant qu'à faire, autant choisir 20 puisqu'au moins il gagnera plus s'il gagne parce B se dirait etc...

Si A et B jouent de manière répétée alors il est possible de définir des stratégies. Chacun des joueurs va choisir une valeur au hasard. Sa stratégie consistera à choisir avec une certaine probabilité 10 ou 20 à chaque partie.

Supposons que chaque joueur choisisse 10 avec probabilité $1/2$, 20 avec probabilité $1/2$.

L'espérance du gain de A est alors

$$1/4(-10) + 1/4(10) + 1/4(20) + 1/4(-20) = 0.$$

En moyenne A ne gagnera rien et B non plus.

Supposons que B pense que A va choisir $1/2$, $1/2$. Peut-il choisir différemment pour changer le cours des choses ? S'il choisit 10 avec probabilité y, 20 avec probabilité $1 - y$ l'espérance du gain de A est

$$1/2y(-10) + 1/2(1-y)(10) + 1/2y(20) + 1/2(1-y)(-20) = 1/2(20y - 10) = 5(2y - 1).$$

On voit alors qu'en choisissant $y = 0$ c'est-à-dire toujours la pièce de 10 centimes B peut rendre l'espérance du gain de A égal à -5 ; autrement dit il gagnera en moyenne 5 centimes par partie si le nombre de parties est grand. Evidemment s'il voit que B choisit toujours 20, A changera sa stratégie et arrêtera de choisir avec probabilité $1/2$, $1/2$. B pourrait donc essayer de masquer son comportement en choisissant quand même 10 de temps en temps (en prenant y égal à $1/3$ par exemple il gagnera en moyenne $5/3$; s'il veut gagner plus, il a intérêt à diminuer la valeur de y mais A s'apercevra qu'il choisit souvent 20 ; s'il veut vraiment passer inaperçu il peut augmenter y mais plus il s'approchera de $1/2$ moins il gagnera).

Se passe-t-il quelque chose de similaire pour A ? Quand B choisit avec probabilité $1/2$, $1/2$ peut-il adapter sa stratégie pour grandir l'espérance de son gain ? S'il choisit 10 avec probabilité x, 20 avec probabilité $1 - x$ l'espérance de son gain est

$$1/2x(-10) + 1/2x(10) + 1/2(1-x)(20) + 1/2(1-x)(-20) = 0.$$

Autrement dit quoi qu'il fasse il ne pourra obtenir une espérance de gain positive.

Peut-il adopter une stratégie qui lui assure à lui aussi de ne rien perdre (en moyenne) quelle que soit la stratégie de B ? Oui.

Pour trouver comment A peut y arriver, nous allons exprimer l'espérance du gain de A si A choisit 10 avec probabilité x, 20 avec probabilité $1-x$ et B choisit 10 avec probabilité y, 20 avec probabilité $1-y$. Cette espérance vaut

$$F(x,y) = -10xy + 10x(1-y) + 20(1-x)y - 20(1-x)(1-y)$$

En développant les produits on obtient :

$$F(x,y) = -60xy + 30x + 40y - 20.$$

On peut écrire aussi cette quantité matriciellement :

$$F(x,y) = \begin{pmatrix} x & (1-x) \end{pmatrix} \begin{pmatrix} -10 & 10 \\ 20 & -20 \end{pmatrix} \begin{pmatrix} y \\ (1-y) \end{pmatrix}$$

On cherche un point critique de F c'est-à-dire un point où les deux dérivées de F s'annulent :

$$\frac{\partial F}{\partial x} = -60y + 30,$$
$$\frac{\partial F}{\partial y} = -60x + 40.$$

Les deux dérivées s'annulent si $y = 1/2$, $x = 2/3$. On remarque les deux égalité suivantes

$$\begin{pmatrix} 2/3 & 1/3 \end{pmatrix} \begin{pmatrix} -10 & 10 \\ 20 & -20 \end{pmatrix} = \begin{pmatrix} 0 & 0 \end{pmatrix}$$

$$\begin{pmatrix} -10 & 10 \\ 20 & -20 \end{pmatrix} \begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

On en déduit que pour tout x, y (entre 0 et 1)

$$F(2/3,y) = \begin{pmatrix} 2/3 & 1/3 \end{pmatrix} \begin{pmatrix} -10 & 10 \\ 20 & -20 \end{pmatrix} \begin{pmatrix} y \\ (1-y) \end{pmatrix} = \begin{pmatrix} 0 & 0 \end{pmatrix} \begin{pmatrix} y \\ (1-y) \end{pmatrix} = 0,$$

$$F(x,1/2) = \begin{pmatrix} x & (1-x) \end{pmatrix} \begin{pmatrix} -10 & 10 \\ 20 & -20 \end{pmatrix} \begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix} = \begin{pmatrix} x & (1-x) \end{pmatrix} \begin{pmatrix} 0 \\ 0 \end{pmatrix} = 0.$$

Conclusion : en choisissant les probabilités 2/3, 1/3, A est assuré qu'en moyenne le gain de B sera nul ; en choisissant les probabilités 1/2, 1/2, B est assuré qu'en moyenne le gain de A sera nul.

Exercice 1. *Colombes et faucons*
Pour essayer de comprendre les stratégies de confrontation adoptées au sein d'espèces animales, divers modèles ont été imaginés. En voici un très simple. Une population est composée de deux types d'individus qu'on désigne par les noms de colombes et faucons. En certaines occasion deux individus peuvent se disputer un gain G (nourriture, espace,...).

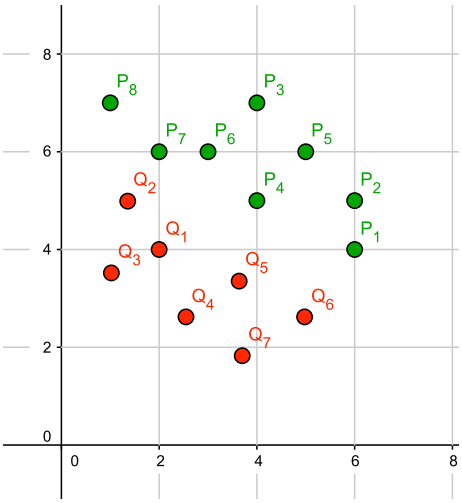
a. *Si deux colombes se disputent le gain, chacune parade et l'une abandonne à partir d'un certain moment. Le gain moyen pour chacune est $G/2$ (on considère qu'une colombe l'emporte une fois sur deux dans ce genre de rencontre).*
b. *Si deux faucons se disputent le gain, une escalade se produit qui aboutit à un combat. Le vainqueur emporte le gain G, mais le vaincu se voit imposer une perte C (un gain $-C$; $C > G$). On considère là aussi qu'un faucon gagne une fois sur deux ses combats contre d'autres faucons. Le gain moyen est donc $(G - C)/2$.*
c. *Si une colombe rencontre un faucon, elle refuse le combat. Son gain est nul, celui du faucon est G.*

Supposons que dans la population la proportion des colombes soit x, celle des faucons $1-x$.

1. *Quels sont les gains moyens des faucons et des colombes ?*
2. *Pour maximiser son gain moyen vaut-il mieux être un faucon ou une colombe ?*
3. *Quelle répartition (quelle valeur de x) rend indifférent le choix entre le comportement de colombe ou faucon ?*
4. *Commenter (faire varier G et C).*

3.3 Exemples de problèmes liés à la programmation linéaire

Séparation de points



$$\max \delta,$$

$$s.c.\ \forall i = 1, \ldots, 7\ y(q_i) \leq ax(q_i) + b + \delta, \forall i = 1, \ldots, 8\ y(p_i) \geq ax(p_i) + b - \delta.$$

On peut chercher à séparer les deux ensembles de points par autre chose que des droites (des paraboles, des courbes de degré trois,...) ou encore résoudre ce genre de problème en dimension plus grande.

Alternative à la droite de régression linéaire

On cherche à minimiser la quantité

$$\sum_{i=1}^{n} |y_i - ax_i - b|$$

Une solution analytique comme celle que nous utilisons pour minimiser

$$\sum_{i=1}^{n} (y_i - ax_i - b)^2$$

qui mène à la droite de régression est difficile à décrire simplement. Mais donner une solution algorithmique basée sur une présentation du problème comme programme linéaire est possible :

$$\min l_1 + l_2 + \ldots + l_n,$$
$$s.c.\ \forall i = 1, \ldots, n\ l_i \geq y_i - ax_i - b,\ l_i \geq -y_i + ax_i + b.$$

3.4 Éléments d'algèbre linéaire et de géométrie affine

Les problèmes de programmation linéaire sont résolus par des opérations très classiques sur les matrices (essentiellement la méthode du pivot de Gauss). Les contraintes et la fonction à optimiser sont de nature linéaire. Pour bien comprendre les problèmes et leurs résolutions l'outil matriciel (calculatoire et théorique) est essentiel. Considérons les matrices suivantes :

$$A = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 \\ 2 & 1 & 0 & 1 & 0 \\ -1 & 1 & 0 & 0 & 1 \end{pmatrix} \quad x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} \quad c = \begin{pmatrix} 1 \\ 2 \\ 0 \\ 0 \\ 0 \end{pmatrix} \quad b = \begin{pmatrix} 4 \\ 10 \\ 5 \end{pmatrix}.$$

Les vecteurs de la base canonique de $\mathbb{R}^5$ sont

$$e_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \quad e_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} \quad e_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} \quad e_4 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{pmatrix} \quad e_5 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}.$$

Les vecteurs colonnes de A sont les résultats de la multiplication de A par ces vecteurs de la base canonique. Par exemple

$$A.e_2 = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 \\ 2 & 1 & 0 & 1 & 0 \\ -1 & 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}.$$

L'image de l'application linéaire définie par la multiplication par A est le sous-espace vectoriel engendré par ses vecteurs colonne. Le rang de la matrice A est la dimension de cette image. Cette dimension ne peut pas excéder la dimension de l'espace d'arrivée (*i.e.* le nombre de lignes de la matrice, ici 3), ni le nombre de vecteurs colonne (ici 5 ; la dimension de l'espace d'arrivée). Le rang d'une matrice de taille $m \times n$ est donc au plus $\min(m, n)$. On dit qu'une matrice $m \times n$ est de rang plein (ou maximal) si son rang est $\min(m, n)$.

Définition 3.1. *Soit P une partie de $\mathbb{R}^d$. On appelle sous-espace affine engendré par P le plus petit sous-espace affine contenant P.*

Définition 3.2. *Un sous-espace affine de $\mathbb{R}^d$ est un ensemble de la forme*

$$x + E = \{x + u \in \mathbb{R}^d \ / u \in E\}$$

où x est un point de $\mathbb{R}^d$ et E un sous-espace vectoriel de $\mathbb{R}^d$: c'est le sous-espace affine de direction E passant par x. On appelle dimension d'un sous-espace affine la dimension du sous-espace vectoriel définissant sa direction ($\dim E$).

Si y et z appartiennent tous les deux à $x + E$ alors $y - z$ appartiennent à E. Pour définir complètement un sous-espace affine il suffit de se donner x et une base de E, ou encore $\dim E + 1$ points $x_0, \ldots, x_{\dim E}$ tels que les vecteurs $x_i - x_0$ engendrent E. On dit que de tels points sont affinement indépendants et que $x_0 + E$ est l'enveloppe affine de la famille $(x_0, \ldots, x_{\dim E})$.

Définition 3.3. *Soit P une partie de E. On appelle enveloppe affine de P l'ensemble*

$$\{\sum_{i=1}^{l} \alpha_i x_i \ / \ l \in \mathbb{N}^*, x_1, \ldots, x_l \in P, \sum_{i=1}^{l} \alpha_i = 1\}.$$

Soient u et v deux vecteurs appartenant tous les deux à $\mathbb{R}^d$. On dit que u est inférieur à v si chaque coordonnée de u est inférieure à v : pour tout $i = 1, \ldots, d$, $u_i \leq v_i$ [1]

1. Ce n'est pas une relation d'ordre totale : deux vecteurs ne sont pas forcément comparables. Par exemple $(1, 2)$ et $(2, 1)$ ne sont pas comparables dans $\mathbb{R}^2$ au sens de cette relation $\leq$.

Soit A une matrice $m \times n$ et $b \in \mathbb{R}^m$. Notons $a_{i,j}$ les coefficients de A. L'ensemble $C(A,b) = \{x \ : \ Ax \leq b\}$ est l'ensemble des x tels que, pour tout i allant de 1 à m on ait

$$\sum_{j=1}^{n} a_{i,j}x_j \leq b_i.$$

Pour i allant de 1 à m, notons l_i le vecteur de $\mathbb{R}^n$ dont les coordonnées sont les $a_{i,j}$:

$$l_i = (a_{i,1}, a_{i,2}, \ldots, a_{i,n}).$$

Les inégalités plus haut s'écrivent

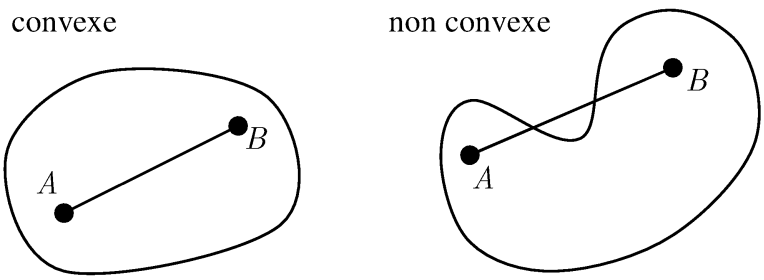
$$\langle {}^t l_i, x\rangle \leq b_i.$$

L'ensemble $\{x \ / \ \langle {}^t l_i, x\rangle \leq b_i\}$ est un demi-espace dont le bord est l'hyperplan *affine* $\{x \ / \ \langle {}^t l_i, x\rangle = b_i\}$. L'ensemble $C(A,b)$ apparaît ainsi comme l'intersection des m demi-espaces $\{x \ / \ \langle {}^t l_i, x\rangle \leq b_i\}$ (il faut que les m inégalités soient satisfaites c'est-à-dire que x doit être dans les m demi-espaces à la fois).

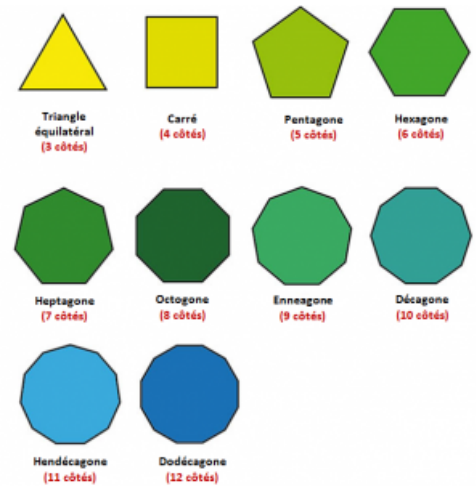
Nous allons étudier les propriétés des ensembles de contraintes qui sont appelés des polyèdres. Ce sont des exemples de parties dites convexes d'espaces vectoriels. Commençons par étudier les convexes en général. Ce sont des ensembles apparaissant souvent en optimisation.

3.5 Propriétés des convexes fermés

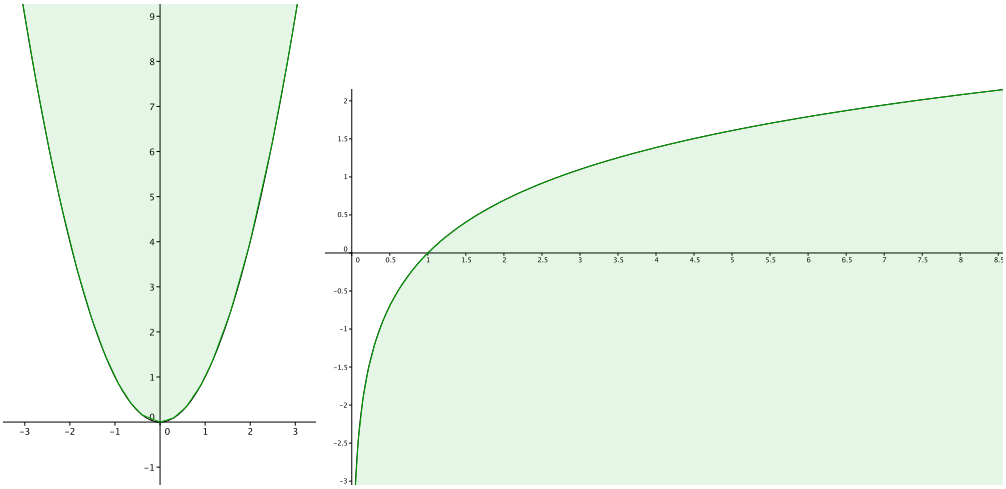
Définition 3.4. *Soit C une partie de $\mathbb{R}^d$. On dit que C est convexe si lorsque x et y sont deux points de C, alors le segment joignant ces deux points $[x,y] = \{tx + (1-t)y \ / \ t \in [0,1]\}$ est lui aussi inclus dans C.*



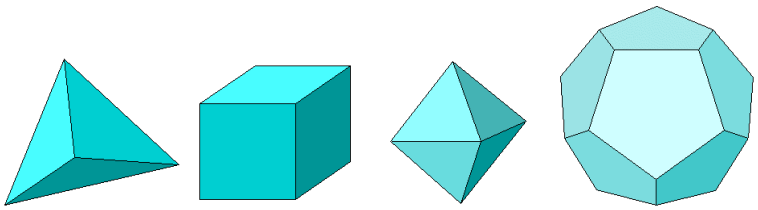
Des exemples en dimension 2.



La partie située au dessus da la parabole ou en-dessous du graphe de la fonction ln :



Des exemples en dimension 3.



En dimension quelconque, un polyèdre $C(A, b)$ est convexe. Soient deux points x et y tels que $Ax \leq b$ et $Ax \leq b$, et λ appartenant à $[0, 1]$. Par linéarité (ou distributivité du produit) on a

$$A(\lambda x + (1 - \lambda)y) = \lambda Ax + (1 - \lambda)Ay.$$

Par hypothèse Ax et Ay sont inférieurs ou égaux à b, et comme λ et $(1 - \lambda)$ sont positifs ou nuls, on en déduit

$$A(\lambda x + (1 - \lambda)y) \leq \lambda b + (1 - \lambda)b = b,$$

ce qui signifie que $\lambda x + (1-\lambda)y$ appartient à $C(A,b)$.

Soit C un ensemble convexe fermé inclus dans $\mathbb{R}^d$. C'est aussi une partie convexe de $Aff(C)$. On appelle intérieur relatif de C son intérieur comme sous-ensemble de $Aff(C)$. On a

$$C = \overline{Int_r(C)},$$

alors que l'intérieur de C dans $\mathbb{R}^d$ peut-être vide. Le bord relatif de C est $\partial_r(C) = C \setminus Int_r(C)$.

Définition 3.5. *Un point x est dit barycentre à coefficients positifs de deux points y et z s'il existe $\alpha \in]0,1[$ tel que $x = \alpha y + (1-\alpha)z$.*

Une face d'un convexe C de $\mathbb{R}^d$ est une partie F de C telle que si $x \in F$ est barycentre à coefficients positifs de deux points y et z de C alors y et z appartiennent aussi à F.

On appelle point extrémal d'un convexe fermé C une face réduite à un point c'est-à-dire un point x de C tel que si $x = \alpha y + (1-\alpha)z$ avec $\alpha \in]0,1[$, y et z dans C alors $x = y = z$.

Lorsque C est un polyèdre on appelle sommets ses points extrémaux. Comment trouver les sommets d'un polyèdre $C(A,b)$? En a-t-il ? Si le rang de A n'est pas n alors $C(A,b)$ contient une droite et n'a pas de sommet. En effet, dans ce cas le noyau de l'application $\phi_A : x \mapsto Ax$ est de dimension supérieure ou égale à 1 car la formule du rang donne

$$\dim Ker\phi_A + \dim Im\phi_A = n,$$

donc si $\dim Im\phi_A < n$ alors $\dim Ker\phi_A \geq 1$. Or si $x \in C(A,b)$ et $z \in Ker\phi_A$ alors $A(x+z) = Ax + Az = Ax + 0 = Ax \leq b$. Cela signifie que $x+z$ appartient à $C(A,b)$. Si $Ker\phi_A$ n'est pas réduit à 0, $C(A,b)$ contient donc une droite (à moins qu'il ne soit vide auquel cas il n'a pas non plus de point extrémaux).

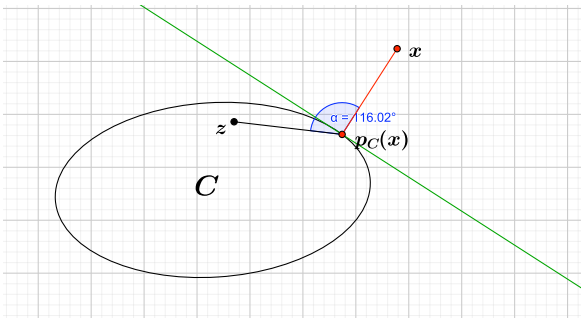
Théorème 3.6. *(projection sur un convexe) Soit C un convexe fermé inclus dans $\mathbb{R}^d$. Pour tout x n'appartenant pas à C il existe un unique point $p_C(x)$ de C minimisant la distance de x aux points de C c'est-à-dire tel que*

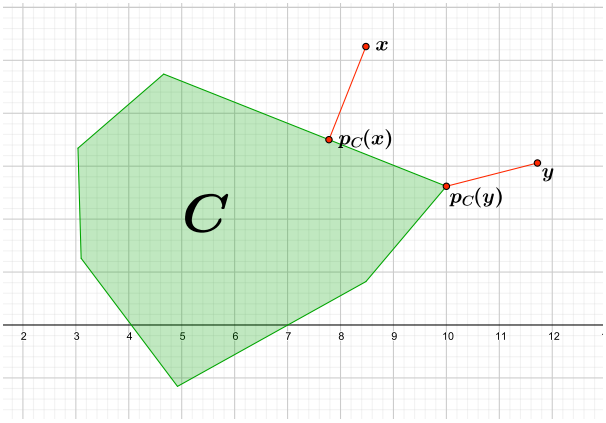
$$d(x, p_C(x)) = \min\{d(x,y) \ / \ y \in C\}.$$

Ce point $p_C(x)$ est appelé projection de x sur C. Il est caractérisé par la propriété que pour tout $z \in C$ on a

$$\langle x - p_C(x), z - p_C(x)\rangle \leq 0.$$

(ce qui signifie que l'angle fait par les vecteurs $x - p_C(x)$ et $z - p_C(x)$ est obtus).





Démonstration Soient $x \notin C$ et $y \in C$. L'ensemble $K = C \cap \overline{B(x, d(x,y)+1)}$ est un convexe fermé borné (fermé et convexe car intersection de deux ensembles fermés et convexes, borné car inclus dans une boule). L'ensemble $\{d(x,z) \ / \ z \in C\}$ est inclus dans $\mathbb{R}_+$. Il admet donc une borne inférieure qui est par définition la distance de x à C. Il existe une suite d'éléments de C telle que $d(y_n, x) \to_{n\to+\infty} d(x, C)$. À partir d'un certain rang y_n appartient à K (car à partir d'un certain rang $d(y_n, x) < d(y, x) + 1$). Comme K est compact, on peut extraire de la suite (y_n) une sous-suite convergente. Soit y_∞ la limite de cette suite extraite. On a $d(y_\infty, x) = d(x, C)$.

Supposons qu'on ait deux points y_∞ et z_∞ qui vérifient tous les deux

$$d(y_\infty, x) = d(x, C) \text{ et } d(z_\infty, x) = d(x, C).$$

Calculons $d(x, (y_\infty + z_\infty)/2)$

$$\begin{aligned} d(x, (y_\infty + z_\infty)/2)^2 &= \|x - \frac{y_\infty + z_\infty}{2}\|^2 \\ &= \|\frac{x - y_\infty}{2} + \frac{x - z_\infty}{2}\|^2 \\ &= \frac{1}{4}\left(\|x - y_\infty\|^2 + \|x - z_\infty\|^2 + 2\langle x - y_\infty, x - z_\infty\rangle\right). \end{aligned}$$

L'inégalité de Cauchy-Schwarz donne

$$\langle x - y_\infty, x - z_\infty\rangle \leq \|x - y_\infty\|\|x - z_\infty\|$$

avec égalité si $x - y_\infty$ est un multiple positif de $x - z_\infty$. On en déduit que

$$d(x,(y_\infty + z_\infty)/2)^2 \leq \frac{1}{4}\left(\|x-y_\infty\|^2 + \|x-z_\infty\|^2 + 2\|x-y_\infty\|\|x-z_\infty\|\right) = d(x,C)^2$$

avec égalité si $x - y_\infty$ est un multiple positif de $x - z_\infty$. Mais, comme par hypothèse $\|x - y_\infty\| = \|x - z_\infty\| = d(x, C)$, la seule possibilité est que $x - y_\infty = x - z_\infty$, autrement dit que $y_\infty = z_\infty$. Comme C est convexe, $(y_\infty + z_\infty)/2$ appartient à C. Par définition de $d(x, C)$ la distance $d(x, (y_\infty + z_\infty)/2)$ est supérieure ou égale à $d(x, C)$. On est donc dans le cas d'égalité.

On a montré l'existence d'un unique point de C qui soit à distance $d(x, C)$ de x : on peut appeler ce point $p_C(x)$.

Expliquons maintenant comment voir que les angles entre $x-p_C(x)$ et les vecteurs $z-p_C(x)$ pour $z \in C$ dont obtus. Remarquons que comme C est convexe pour tout $\lambda \in [0, 1]$, le point $(1 - \lambda)p_C(x) + \lambda z$ appartient à C. Cela entraîne que pour tout $\lambda \in [0, 1]$ on a

$$d(x,(1-\lambda)p_C(x) + \lambda z) \geq d(x, C) = d(x, p_C(x)).$$

En exprimant ces distances comme des normes et en prenant le carré on obtient

$$\|x - ((1-\lambda)p_C(x) + \lambda z)\|^2 \geq \|(x - p_C(x)\|^2,$$

ce qui s'écrit encore

$$\|(x - p_C(x)) + \lambda(p_C(x) - z))\|^2 \geq \|(x - p_C(x)\|^2.$$

En écrivant le premier membre comme un produit scalaire et en utilisant la bilinéarité on en déduit

$$\|x - p_C(x)\|^2 + \lambda^2\|p_C(x) - z\|^2 + 2\lambda\langle x - p_C(x), p_C(x) - z\rangle \geq \|x - p_C(x)\|^2,$$

ce qui donne

$$\lambda^2\|p_C(x) - z\|^2 + 2\lambda\langle x - p_C(x), p_C(x) - z\rangle \geq 0.$$

Si on suppose maintenant $\lambda > 0$ on peut diviser par λ. On obtient que pour tout $\lambda \in]0, 1]$ on a

$$\lambda\|p_C(x) - z\|^2 + 2\langle x - p_C(x), p_C(x) - z\rangle \geq 0.$$

En faisant tendre λ vers 0, on voit que ceci n'est possible que si $\langle x-p_C(x), p_C(x)-z\rangle \geq 0$, ce qui est la même chose que $\langle x - p_C(x), z - p_C(x)\rangle \leq 0$. □

Le théorème de projection fournit ce qu'on appelle des hyperplans et des demi-espaces d'appui pour le convexe C. Si $y \notin C$ alors

$$H = \{u \in \mathbb{R}^d \ / \ \ \langle u, y - p_C(y)\rangle = \langle p_C(y), y - p_C(y)\rangle\},$$

l'hyperplan orthogonal à $y - p_C(y)$ passant par $p_C(y)$, sépare $\mathbb{R}^d$ en deux demi-espaces dont l'un contient C :

$$C \subset \{u \in \mathbb{R}^d \ / \ \langle u, y - p_C(y)\rangle \leq \langle p_C(y), y - p_C(y)\rangle\}.$$

L'application p est continue. Plus précisément, si y et z sont deux points n'appartenant pas à C, alors

$$\|p_C(y) - p_C(z)\| \leq \|y - z\|.$$

C'est une conséquence de la propriété sur l'angle donnée plus haut :

$$\begin{aligned}
\|p_C(y) - p_C(z)\|^2 &= \langle p_C(y) - p_C(z), p_C(y) - p_C(z)\rangle \\
&= \langle (p_C(y) - y) + (y - z) + (z - p_C(z)), p_C(y) - p_C(z)\rangle \\
&= \langle p_C(y) - y, p_C(y) - p_C(z)\rangle + \langle y - z, p_C(y) - p_C(z)\rangle \\
&\qquad + \langle z - p_C(z), p_C(y) - p_C(z)\rangle \\
&\leq \langle y - z, p_C(y) - p_C(z)\rangle \\
&\leq \|y - z\|\|p_C(y) - p_C(z)\|
\end{aligned}$$

La première inégalité vient du fait que les deux produits scalaires $\langle p_C(y) - y, p_C(y) - p_C(z)\rangle$ et $\langle z - p_C(z), p_C(y) - p_C(z)\rangle$ sont négatifs ou nuls. La deuxième est l'inégalité de Cauchy-Schwarz.

Théorème 3.7. *(Minkowski) Soit C un convexe fermé borné inclus dans $\mathbb{R}^d$. Alors C est l'enveloppe convexe de ses points extrémaux.*

Théorème 3.8. *Soit C un convexe fermé inclus dans $\mathbb{R}^d$ ne contenant pas de droite. Alors C est l'enveloppe convexe de ses points extrémaux et de ses demi-droites extrémales.*

Démonstration Le théorème de Minkowski est une conséquence du deuxième résultat, si l'on remarque qu'un convexe fermé ne contenant pas de demi-droite est borné. Montrons le deuxième théorème par récurrence sur la dimension du convexe.

Les convexes fermés de dimension 1 sont les segments, les demi-droites et les droites. Si on suppose que C est de dimension 1 et ne contient pas de droite alors c'est un segment ou une demi-droite. Si c'est un segment alors il est l'enveloppe convexe de ses deux extrémités qui sont ses deux points extrémaux. Si c'est une demi-droite alors il coïncide évidemment avec l'enveloppe convexe de sa demi-droite extrémale.

Soit k un entier naturel. Supposons que le résultat soit vrai pour tout convexe fermé de dimension inférieur ou égal à k. Montrons qu'il l'est alors pour tout convexe de dimension $k+1$. Soit C un convexe fermé de dimension $k+1$ ne contenant pas de demi-droite. Le convexe C est inclus dans un espace $\mathbb{R}^d$ qui n'est pas nécessairement $\mathbb{R}^{k+1}$ mais quitte à remplacer $\mathbb{R}^d$ par $Aff(C)$ on peut supposer que C est inclus dans $\mathbb{R}^{k+1}$ et d'intérieur non vide. Soit x un point du bord de C. Il existe $y \notin C$ tel que $x = p_C(y)$. L'hyperplan

$$H = \{u \in \mathbb{R}^{k+1} \ / \ \langle u, y - p_C(y)\rangle = \langle p_C(y), y - p_C(y)\rangle\}$$

est un hyperplan d'appui pour C : H contient $p_C(y)$ et C est inclus dans l'un des demi-espaces définis pas H :

$$C \subset \{u \in \mathbb{R}^{k+1} \ / \ \langle u, y - p_C(y)\rangle \leq \langle p_C(y), y - p_C(y)\rangle\}.$$

L'ensemble $H \cap C$ est donc un convexe fermé (car intersection de deux convexes fermés), non vide (car il contient $p_C(y)$), de dimension inférieure ou égale à k (car il est inclus dans H). Par hypothèse de récurrence $H \cap C$ est l'enveloppe convexe de ses points extrémaux et de ses demi-droites extrémales. Or les points extrémaux de $H \cap C$ sont extrémaux dans C. Cela provient du fait que H est un hyperplan d'appui de C : tous les points de C sont du même côté de H. Si x est un point extrémal de $H \cap C$ et barycentre à coefficients positifs de deux points y et z de C alors y et z doivent être sur H (sinon l'un serait d'un côté de H, l'autre de l'autre côté, ce qui est impossible). On en déduit que y et z sont donc dans $H \cap C$. Or x est extrémal dans $H \cap C$ donc $y = z = x$. Nous avons donc montré que les points du bord de C étaient dans l'enveloppe convexe de ses points extrémaux et de ses demi-droites extrémales... à condition d'expliquer pourquoi il existe un point $y \notin C$ tel que $p_C(y) = x$.

Prenons une suite (y_k) de points qui ne soient pas dans C telle que $\lim y_k = x$. Alors $p(y_k)$ tend vers x. Notons $\mathbb{S}^k(x, 1)$ la sphère de rayon 1 centrée en x et z_k l'intersection de la demi-droite $[p(y_k), y_k)$ avec $\mathbb{S}^k(x, 1)$. Alors (z_k) est une suite de points à distance 1 de x qui sont projetés sur les points $p(y_k)$ qui tendent vers x. Comme la sphère $\mathbb{S}^k(x, 1)$ est compacte, on peut extraire de (z_k) une suite convergente vers une limite z. Alors $d(z, x) = 1$, $z \notin C$ et (par continuité de p) $p(z) = x$.

Considérons maintenant un point x de l'intérieur de C. Notons $\mathbb{S}^k$ la sphère unité de $\mathbb{R}^{k+1}$ (c'est-à-dire l'ensemble des vecteurs de norme 1). Pour $v \in \mathbb{S}^k$ considérons la demi-droite

$$D_v = \{x + \lambda v \ / \ \lambda \geq 0\},$$

et la partie de $\mathbb{S}^k$ définie par

$$E = \{v \in \mathbb{S}^k \ / \ D_v \subset C\}.$$

Par hypothèse E ne contient pas deux vecteurs opposés (car C ne contient pas de droite). Comme C est fermé, E est fermé aussi. Les deux ensembles E et $-E$ sont donc deux parties fermées disjointes de $\mathbb{S}^k$. Ce sont donc deux parties compactes disjointes et comme telles elles sont à distance positive l'une de l'autre. On en déduit que leur réunion ne peut pas être égale à $\mathbb{S}^k$. Prenons $v_0 \in \mathbb{S}^k \setminus (E \cup -E)$. Alors $D_{v_0} \not\subset C$ et $D_{-v_0} \not\subset C$. Cela signifie qu'il existe λ et μ positifs tels que

$$x + \lambda v_0 \in \partial C \text{ et } x - \mu v_0 \in \partial C.$$

Autrement dit x est sur un segment joignant deux points du bord de C. Ces deux points sont dans l'enveloppe convexe des points extrémaux et des demi-droites extrémales de C (d'après la première partie de la démonstration), donc x aussi. □

Proposition 3.9. *Soit C un convexe fermé inclus dans $\mathbb{R}^d$. Il est compact si et seulement s'il ne contient pas de demi-droite.*

Démonstration Une demi-droite n'est pas bornée donc si C contient une demi-droite, C n'est pas borné. Supposons que C ne soit pas bornée. Montrons qu'il contient une demi-droite. Comme C n'est pas borné, il existe une suite $(x_k)_{k\geq 0}$ telle que $\lim \|x_k\| = +\infty$. Considérons alors la suite des vecteurs

$$u_k = \frac{x_k - x_0}{\|x_k - x_0\|}.$$

Ce sont des vecteurs de norme 1 donc des éléments de $\mathbb{S}^k$. Or $\mathbb{S}^k$ est compacte, donc on peut extraire de $(u_k)_{k\geq 0}$ une suite convergente $(u_{k_j})_{j\geq 0}$ vers une limite l appartenant à $\mathbb{S}^k$. Pour tout j, x_{k_j} appartient à C, donc $x_0 + u_{k_j}$ aussi (si j est assez grand, car c'est un point du segment $[x_0, x_{k_j}]$). Comme C est fermé, on en déduit que $x_0 + l$ appartient à C. Maintenant on peut voir que tout point de la forme $x_0 + \lambda l$, avec $\lambda \geq 0$, est limite de $x_0 + \lambda u_{k_j}$ (qui appartient à C si j est assez grand) donc est dans C. Conclusion : C contient la demi-droite $\{x_0 + \lambda l \ / \ \lambda \in \mathbb{R}_+\}$. □

3.6 Les polyèdres

Théorème 3.10. *Un polyèdre a un nombre fini de faces. En particulier il a un nombre fini de sommets. S'il ne contient pas de droite il contient un nombre fini de demi-droites extrémales.*

Corollaire : un polyèdre compact est l'enveloppe convexe d'un nombre fini de points. Un polyèdre ne contenant pas de droites est l'enveloppe convexe d'un nombre fini de points et de demi-droites.

Question : Un polyèdre peut-il contenir une droite et avoir un sommet ?

Considérons un polyèdre du type suivant (ensemble de contraintes d'un programme linéaire sous forme standard)

$$C(A, b) = \{x \in \mathbb{R}^n \ / \ Ax = b, \ x \geq 0\},$$

où A est une matrice $m \times n$. Notons r le rang de A. Si $r < m$ alors, soit le système est incompatible, et dans ce cas $C(A, b)$ est vide, soit le système est équivalent à un système $A'x = b'$ avec une matrice $r \times n$ de rang r (en enlevant $m - r$ lignes superflues). Nous supposerons maintenant que $m \leq n$ et que A est de rang m.

Si $m = n$ le système $Ax = b$ a une unique solution, $C(A, b)$ est un point si $A^{-1}b \geq 0$, et vide sinon. Le cas le plus intéressant est $m < n$: $C(A, b)$ est l'intersection de $\{x \ / \ x \geq 0\}$ et d'un sous-espace affine de dimension $n - m$: s'il n'est pas vide, alors pour tout point x_0 de $C(A, b)$, on a

$$C(A, b) = [x_0 + \ker \phi_A] \cap \{x \ / \ x \geq 0\}.$$

Théorème 3.11. *(lemme de Farkas) Soient $b, a_1, \ldots, a_k$ $k+1$ vecteurs dans $\mathbb{R}^d$. Désignons par $C(a_1, \ldots, a_k)$ le cône engendré par les vecteurs a_i. Le vecteur b appartient à $C(a_1, \ldots, a_k)$ si et seulement si*

$$\bigcap_{i=1}^{k} \{x \ / \ \langle x, a_i \rangle \leq 0\} \subset \{x \ / \ \langle x, b \rangle \leq 0\}.$$

Démonstration Dire que b appartient à $C = C(a_1, \ldots, a_k)$ est dire qu'il existe des nombres $\lambda_i \geq 0$ tels que $b = \sum_{i=1}^{k} \lambda_i a_i$. On a alors

$$\langle x, b \rangle = \sum_{i=1}^{k} \lambda_i \langle x, a_i \rangle.$$

On en déduit que si tous les produits scalaires $\langle x, a_i \rangle$ sont négatifs ou nuls, il en est de même de $\langle x, b \rangle$. Cela signifie que l'inclusion

$$\bigcap_{i=1}^{k} \{x \ / \ \langle x, a_i \rangle \leq 0\} \subset \{x \ / \ \langle x, b \rangle \leq 0\}.$$

est satisfaite. On souhaite maintenant montrer que si l'inclusion est vraie alors b appartient au cône C. On procède par contraposée. Considérons un vecteur b qui n'appartient pas à C. Alors, comme C est un convexe fermé (à justifier), on peut séparer b et C par un hyperplan : il existe un vecteur v tel que $\langle v, b \rangle > 0$ et $\langle v, w \rangle \leq 0$ pour tout élément w de C (en particulier $\langle v, a_i \rangle \leq 0$ pour tout i) : l'inclusion n'est donc pas satisfaite.

Pour v on peut prendre $b - p_C(v)$; l'orthogonal de ce vecteur est un hyperplan d'appui contenant 0 car C contient les vecteurs ($\alpha p_C(v)$ pour $\alpha \geq 0$ car C est un cône). □

3.7 Formes canoniques et standard

On peut distinguer un problème primal et un problème dual. C'est le point de vue de celui à qui se pose le problème qui détermine lequel est primal, lequel est dual.

Formes canoniques

Problème primal : $\min\langle c, x \rangle$ sous contraintes $Ax \geq b$, $x \geq 0$.

Problème dual : $\max\langle b, y \rangle$ sous contraintes ${}^tAy \leq c$, $y \geq 0$.

Formes standard

Problème primal : $\max\langle c, x \rangle$ sous contraintes $Ax \leq b$.

Problème dual : $\min\langle b, y \rangle$ sous contraintes ${}^tAy = c$, $y \geq 0$.

Cette dernière forme est la forme adaptée pour appliquer l'algorithme du simplexe.

A est une matrice $m \times n$, x, c des vecteurs $n \times 1$, y des vecteurs $m \times 1$.

Supposons que x et y satisfont les contraintes des problèmes primal et dual sous la forme standard. On a alors

$$\langle c, x\rangle =^t ({}^tAy)x =^t yAx \leq^t yb = \langle b, y\rangle,$$

la première égalité est vraie car y satisfait les contraintes du programme dual, la deuxième est un calcul de transposée de produit de matrice, l'inégalité est vraie car x satisfait les contraintes du problème primal et car $y \geq 0$.

Ce que montre le calcul précédent est que si les contraintes sont vérifiées, toutes les valeurs de la fonction objectif $\langle c, x\rangle$ sont inférieures ou égales à celles, $\langle b, y\rangle$, du programme dual. On en déduit que si on trouve une valeur commune aux deux fonctions objectif en deux points dans les ensembles définis par les contraintes alors on a résolu les deux problèmes en même temps.

Si l'ensemble défini par les contraintes du problème primal n'est pas vide mais que la fonction objectif n'est pas majorée sur cet ensemble alors nécessairement l'ensemble défini par les contraintes du problème dual est vide (et on a une affirmation symétrique).

Si aucun des deux ensembles définis par les contraintes n'est vide alors la fonction objectif du problème primal est majorée donc a une borne supérieure, celle du problème dual est minorée, donc a une borne inférieure. Ces bornes sont-elles atteintes ? Sont-elles égales ?

Considérons l'ensemble $C(A, b) = \{x \ : \ Ax \leq b\}$. C'est une intersection de m demi-espaces : on appelle un tel ensemble un polyèdre.

3.8 Méthode du simplexe

L'algorithme du simplexe est un algorithme de résolution des problèmes d'optimisation linéaire. Il a été introduit par George Dantzig à partir de 1947. C'est probablement le premier algorithme permettant de minimiser une fonction sur un ensemble défini par des inégalités. De ce fait, il a beaucoup contribué au démarrage de l'optimisation numérique. L'algorithme du simplexe a longtemps été la méthode la plus utilisée pour résoudre les problèmes d'optimisation linéaire. Depuis les années 1985-90, il est concurrencé par les méthodes de points intérieurs, mais garde une place de choix dans certaines circonstances (en particulier si l'on a une idée des contraintes d'inégalité actives en la solution). [2]

Le nom de l'algorithme est dérivé de la notion de simplexe et a été suggéré par Motzkin2. En réalité, l'algorithme n'utilise pas de simplexes, mais certaines interprétations de l'ensemble admissible du problème renvoient au concept de simplexe.

2. Wikipédia ; https ://fr.wikipedia.org/wiki/Algorithme_du_simplexe

3.8.1 L'algorithme

On se donne un programme linéaire sous forme standard

$$\max\langle c, x\rangle \quad s.c.\ Ax = b,\ x \geq 0.$$

La matrice A est de taille $m \times n$ avec $m \leq n$, les vecteurs c et x dans $\mathbb{R}^n$, le vecteur b dans $\mathbb{R}^m$. On suppose que A est de rang maximal m. Appelons $C(A, b)$ l'ensemble des vecteurs x satisfaisant les contraintes

$$C(A, b) = \{x \in \mathbb{R}^n \ / \ Ax = b,\ x \geq 0\}.$$

Pour décrire l'algorithme il est pratique d'introduire certaines notations. Les sommets de $C(A, b)$ sont calculés à partir de matrices $m \times m$ obtenues en choisissant m colonnes parmi les n colonnes de A. On décrira un tel choix en indiquant quelles colonnes sont choisies (B) et quelles colonnes ne le sont pas (N). On se donnera deux familles d'indices B et N

$$B = (p_1, \ldots, p_m) \in \{1, \ldots, n\}^m,\ N = (q_1, \ldots, q_{n-m}) \in \{1, \ldots, n\}^{n-m}$$

telles que les p_i soient tous différents des q_j (et vice-versa) et telles que en prenant les p_i et les q_i on obtienne tous les indices de 1 à n. On notera A_B la matrice obtenue à partir de A en ne conservant que les colonnes dont les numéros sont dans B, A_N a matrice obtenue à partir de A en ne conservant que les colonnes dont les numéros sont dans N. La matrice A_N est une matrice $m \times (n - m)$, la matrice A_B est une matrice carrée $m \times m$.

Lorsque B est donné et $x \in \mathbb{R}^n$ on note x_B le vecteur de $\mathbb{R}^m$ dont les coordonnées sont celles de x de numéros dans B, x_N le vecteur de $\mathbb{R}^{n-m}$ dont les coordonnées sont celles de x de numéros dans N.

Si A_B est inversible on dit que c'est une matrice de base de A (la famille B elle même est alors souvent qualifiée de base...). Dans ce cas le vecteur x défini par $x_B = A_B^{-1}b$, $x_N = 0$ est dit solution de base. Si A_B est une base les x_j pour j dans B sont dites variables de base, les x_k pour k dans N sont dites variables hors base.

Si A_B est une base, on dit que A_B et la solution de base sont admissibles si $A_B^{-1}b \geq 0$. Une base admissible est dite dégénérée si certaines des coordonnées de $A_B^{-1}b$ sont nulles, non dégénérée si $A_B^{-1}b > 0$.

Proposition 3.12. *Le polyèdre $C(A, b)$ ne contient pas de droites. Il est donc égal à l'enveloppe convexe de l'ensemble de ses points extrémaux et de ses demi-droites extrémales.*

Démonstration Une droite ne peut pas être incluse dans l'ensemble $\{x \ / \ x \geq 0\}$ a fortiori pas dans $C(A, b)$. Montrons qu'une droite contient toujours des points dont certaines coordonnées sont négatives. Soit D une droite. Elle peut être représentée paramétriquement par un point et un vecteur directeur

$$D = \{a + \lambda v \ / \ \lambda \in \mathbb{R}\},$$

où a et v sont deux vecteurs et v un vecteur directeur de D n'est pas nul. Soit i_0 tel que v_{i_0} ne soit pas nul. Suivant le signe de v_{i_0} on a

$$\lim_{\lambda\to+\infty} a_{i_0}+\lambda v_{i_0}=+\infty \text{ et } \lim_{\lambda\to-\infty} a_{i_0}+\lambda v_{i_0}=-\infty$$

ou

$$\lim_{\lambda\to+\infty} a_{i_0}+\lambda v_{i_0}=-\infty \text{ et } \lim_{\lambda\to-\infty} a_{i_0}+\lambda v_{i_0}=+\infty.$$

Dans tous les cas pour certaines valeurs de λ la coordonnée numéro i_0 de $a+\lambda v$ est négative.□

En particulier $C(A,b)$ a des sommets. Soit $c\in\mathbb{R}^n$. Supposons que $\langle c,x\rangle$ soit majorée sur $C(A,b)$, alors le problème a une solution et le maximum est atteint en un sommet de $C(A,b)$.

Proposition 3.13. *Les sommets du polyèdre $C(A,b)$ sont les solutions de base admissibles, c'est-à-dire des points de la forme (x_B,x_N) avec $x_B=A_B^{-1}b$, $x_N=0$ où $A_B^{-1}b\geq 0$.*

Démonstration Supposons que x soit un sommet de $C(A,b)$. Le point x ne peut pas être à coordonnées strictement positives. En effet pour tout élément y du noyau de A on a $A(x+y)=Ax+Ay=b+0=b$. Si $x>0$ alors pour y tout petit dans le noyau de A on a

$$A(x+y)=A(x-y)=b,\ x+y>0,\ x-y>0.$$

Et dans ce cas x est le milieu du segment $[x-y,x+y]$ dont les extrémités sont dans $C(A,b)$ et distinctes : x n'est pas extrémal (x n'est pas un sommet). On peut dire mieux grâce à un raisonnement analogue : au moins $d-m$ coordonnées de x sont nulles. Supposons que $m+1$ coordonnées de x soient strictement positives : disons $x_{i_1}>0$,...$x_{i_{m+1}}>0$. Comme $\dim(\ker A)=d-m$, $ker(A)\cap Vect(e_{i_1},\dots,e_{i_{m+1}})\neq\{0\}$. Pour y tout petit dans $ker(A)\cap Vect(e_{i_1},\dots,e_{i_{m+1}})\neq\{0\}$ on a

$$A(x+y)=A(x-y)=b,\ x+y>0,\ x-y>0,$$

et x n'est pas un sommet. Reste à voir qu'on peut trouver une famille B à m éléments telle que $x_B=A_B^{-1}b$. ? ? ? ?

Il faut aussi montrer que les points de la forme $x=(x_B=A_B^{-1}b,x_N=0)$ sont des sommets. Supposons que $x=x'+x''$ avec x' et x'' dans $C(A,b)$. Alors on doit avoir $x'_N=x''_N=0$ (car x' et x'' sont à coordonnées positives ou nulles) et donc $A_Bx'=A_Bx''=b$ ce qui signifie que x, x' et x'' sont égaux ; le points x est bien extrémal. □

Le convexe $C(A,b)$ ne contient pas de droite donc est égal à l'enveloppe convexe de ses points extrémaux et de ses demi-droites extrémales. Notons s_i pour i allant de 1 à r les sommets de $C(A,b)$. Les extrémités des demi-droites sont des points extrémaux. On peut numéroter les sommets de telle façon que les premiers soient des extrémités de demi-droites extrémales. Si $C(A,b)$ contient s demi-droites extrémales elles sont de la forme

$$\{s_i+\lambda v_i\ /\ \lambda\geq 0\}.$$

Soit x un élément de $C(A,b)$. Comme $C(A,b)$ est l'enveloppe convexe de ses points extrémaux et de ses demi-droites extrémales, il existe des nombres positifs ou nuls α_i, λ_j, $i=1,\dots,r$, $j=1,\dots,s$, tels que $\sum_{i=1}^{r}\alpha_i = 1$ et

$$x = \sum_{j=1}^{s}\alpha_j(s_j+\lambda_j v_j) + \sum_{i=s+1}^{r}\alpha_i s_i.$$

Or $\langle c, v_j\rangle \leq 0$ pour tout j. En effet sinon on aurait

$$\lim_{\lambda\to+\infty}\langle c, s_i+\lambda v_j\rangle = +\infty$$

ce qui contredirait le fait que $\langle c,x\rangle$ est majorée sur $C(A,b)$ (nous nous sommes placé sous cette hypothèse). Mais alors on a

$$\begin{aligned}
\langle c,x\rangle &= \langle c, \sum_{j=1}^{s}\alpha_j(s_j+\lambda_j v_j) + \sum_{i=s+1}^{r}\alpha_i s_i\rangle \\
&= \sum_{j=1}^{s}\alpha_j(\langle c,s_j\rangle + \lambda_j\langle c,v_j\rangle) + \sum_{i=s+1}^{r}\alpha_i\langle c,s_i\rangle \\
&\leq \sum_{j=1}^{s}\alpha_j\langle c,s_j\rangle + \sum_{i=s+1}^{r}\alpha_i\langle c,s_i\rangle \\
&= \sum_{i=1}^{r}\alpha_i\langle c,s_i\rangle \\
&\leq \left(\sum_{i=1}^{r}\alpha_i\right)\max_i\langle c,s_i\rangle \\
&= \max_i\langle c,s_i\rangle
\end{aligned}$$

ce qui montre que le maximum de la fonction $\langle c,x\rangle$ est atteint en un sommet.

Proposition 3.14. *Soit A une matrice $m\times n$ avec $m\leq n$ de rang m. Soit B une famille d'indices compris entre 1 et n telle que A_B soit inversible. Alors sont équivalents :*
– $Ax = b$
– $x_B = A_B^{-1}b - A_B^{-1}A_N x_N$.

Démonstration L'égalité

$$(A_B, A_N)\begin{pmatrix} x_B \\ x_N\end{pmatrix} = b$$

s'écrit $A_B x_B + A_N x_N = b$. En multipliant par A_B^{-1} on obtient

$$x_B = A_B^{-1}b - A_B^{-1}A_N x_N.$$

□

Proposition 3.15. *On se donne $B = (p_1, \ldots, p_m)$, $N = (q_1, \ldots, q_{n-m})$ tels que A_B soit une matrice de base. Posons*

$$\overline{A} = A_B^{-1} A_N = (\overline{a}_{rs})_{r=1\ldots,m;s=1,\ldots n-m}, \quad \overline{b} = A_B^{-1} b$$

Si $\overline{a}_{rs} \neq 0$ alors, pour $B' = (p_1, \ldots, p_{r-1}, q_s, p_{r+1}, \ldots, p_m)$, $A_{B'}$ est inversible et $A_{B'}^{-1} = EA_B^{-1}$ où

$$E = \begin{pmatrix} 1 & 0 & \ldots & 0 & \eta_1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & \ldots & 0 & \eta_2 & 0 & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 & \eta_{r-1} & 0 & \ldots & 0 & 0 \\ 0 & 0 & \ldots & 0 & \eta_r & 0 & \ldots & 0 & 0 \\ 0 & 0 & \ldots & 0 & \eta_{r+1} & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & 0 & \eta_{m-1} & 0 & \ldots & 1 & 0 \\ 0 & 0 & \ldots & 0 & \eta_m & 0 & \ldots & 0 & 1 \end{pmatrix}$$

où $\eta_r = 1/\overline{a}_{rs}$ et si $i \neq r$, $\eta_i = -\overline{a}_{is}/\overline{a}_{rs}$ (seule la $r$ème colonne différencie E de l'identité).

L'élément $\overline{a}_{rs}$ s'appelle un élément pivot.

Si x et x' sont les solutions de bases associées à B et B' on a

$$x'_{p_i} = \overline{b}_i - \frac{\overline{a}_{is}}{\overline{a}_{rs}} \overline{b}_r, \text{ si } i \neq r, x'_{q_s} = \frac{\overline{b}_r}{\overline{a}_{rs}}, \ x'_j = 0 \text{ sinon}$$

Démonstration La matrice $A_{B'}$ est obtenue à partir de A_B en remplaçant sa r-éme colonne par la s-ème colonne de A_N. Notons $(e_1, \ldots, e_n)$ la base canonique de $\mathbb{R}^n$. Si $i \neq r$ on a

$$A_B e_{p_i} = A_{B'} e_{p_i}$$

(ces vecteurs sont les vecteurs colonne de A_B et $A_{B'}$ qui coïncident). On en déduit que $A_{B'}^{-1} A_B$ est l'identité sur $Vect\{e_{p_i} \ / \ i \neq r\}$. D'autre part $A_{B'} e_{q_s}$ est la colonne numéro q_s de A c'est-à-dire

$$A_{B'} e_{q_s} = A_N e_{q_s}.$$

On en déduit que $A_B^{-1} A_{B'} e_{q_s}$ vaut

$$A_B^{-1} A_{B'} e_{q_s} = A_B^{-1} A_N e_{q_s},$$

autrement dit $A_B^{-1} A_{B'} e_{q_s}$ est la colonne numéro s de $\overline{A}$. Nous avons ainsi identifié les m

vecteurs colonne de $A_B^{-1}A_{B'}$ ce qui permet d'écrire la matrice :

$$A_B^{-1}A_{B'} = \begin{pmatrix} 1 & 0 & \dots & 0 & \overline{a}_{1s} & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & \overline{a}_{2s} & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & \overline{a}_{(r-1)s} & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & \overline{a}_{rs} & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & \overline{a}_{(r+1)s} & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & \overline{a}_{(m-1)s} & 0 & \dots & 1 & 0 \\ 0 & 0 & \dots & 0 & \overline{a}_{ms} & 0 & \dots & 0 & 1 \end{pmatrix}$$

La matrice E de l'énoncé est l'inverse de cette matrice. On se convaincra qu'elle a bien la forme annoncée. □

Proposition 3.16. *On se donne un programme linéaire sous forme standard et A_B une matrice de base admissible. On a*

$$\langle c, x\rangle =^t c_B A_B^{-1} b + ({}^t c_N -^t c_B A_B^{-1} A_N) x_N.$$

On appelle le vecteur ${}^t\overline{c} =^t c_N - {}^t c_B A_B^{-1} A_N$ le vecteur des coût réduits. S'il est à coordonnées négatives ou nulles alors $x_B = A_B^{-1}b$, $x_N = 0$ fournit une solution optimale au problème. On a une réciproque partielle : si $A_B^{-1}b$ est à coordonnées strictement positives et fournit une solution optimale alors le vecteur des coûts réduits associé est à coordonnées négatives ou nulles.

Démonstration Considérons un point y tel que $Ay = b$ et $y \geq 0$. On a vu qu'alors on a l'égalité $y_B = A_B^{-1}b - A_B^{-1}A_N y_N$; en prenant le produit scalaire par c on obtient

$$\begin{aligned} \langle c, y\rangle &= \langle c_B, y_B\rangle + \langle c_N, y_N\rangle \\ &= \langle c_B, A_B^{-1}b - A_B^{-1}A_N y_N\rangle + \langle c_N, y_N\rangle \\ &= \langle c_B, A_B^{-1}b\rangle - \langle c_B, A_B^{-1}A_N y_N\rangle + \langle c_N, y_N\rangle \\ &= \langle c_B, A_B^{-1}b\rangle - \langle {}^t(A_B^{-1}A_N)c_B, y_N\rangle + \langle c_N, y_N\rangle \\ &= \langle c_B, A_B^{-1}b\rangle + \langle c_N -^t (A_B^{-1}A_N)c_B, y_N\rangle \\ &= \langle c_B, A_B^{-1}b\rangle + \langle \overline{c}, y_N\rangle. \end{aligned}$$

Si $\overline{c} \leq 0$ on a $\langle \overline{c}, y_N\rangle \leq 0$ car $y \geq 0$ donc

$$\langle \overline{c}, y_N\rangle \leq \langle c_B, A_B^{-1}b\rangle = \langle c_B, A_B^{-1}b\rangle + \langle c_N, 0_N\rangle,$$

où l'on noté 0_N le vecteur dont les coordonnées d'indice dans N sont nulles. Or le vecteur $(A_B^{-1}b, 0_N)$ est un sommet de $C(A, b)$. Nous avons montré que si $\overline{c} \leq 0$ alors le sommet associé à B est un point où la fonction objectif est maximale.

Supposons maintenant que le sommet $x = (A_B^{-1}b, 0_N)$ associé à B est un point où la fonction objectif est maximale. De l'inégalité

$$\langle c, y\rangle \leq \langle c, x\rangle$$

on tire

$$\langle c_B, A_B^{-1}b\rangle + \langle \overline{c}, y_N\rangle \leq \langle c_B, A_B^{-1}b\rangle + 0, \text{ donc } \langle \overline{c}, y_N\rangle \leq 0.$$

Que x ne soit pas dégénéré signifie que $A_B^{-1}b > 0$. Pour $\lambda > 0$ suffisamment petit on a donc $A_B^{-1}b \geq \lambda A_B^{-1}A_N e_i$. Posons $x_N^\lambda = \lambda e_i$, $x_B^\lambda = A_B^{-1}b - \lambda A_B^{-1}A_N e_i$, $x^\lambda = (x_B^\lambda, x_N^\lambda)$. On a $Ax^\lambda = b$, $x^\lambda \geq 0$ et

$$\langle c, x^\lambda\rangle = \langle c_B, A_B^{-1}b\rangle - \langle c_B, \lambda A_B^{-1}A_N e_i\rangle + \langle c_N, \lambda e_i\rangle = \langle c_B, A_B^{-1}b\rangle + \langle c_N - {}^t(A_B^{-1}A_N)c_B, \lambda e_i\rangle$$

autrement écrit

$$\langle c, x^\lambda\rangle = \langle c, x\rangle + \langle \overline{c}, \lambda e_i\rangle = \langle c, x\rangle + \lambda \overline{c}_i.$$

Comme $\langle c, x^\lambda\rangle \leq \langle c, x\rangle$ cela entraîne que $\overline{c}_i$ est négatif ou nul. □

Proposition 3.17. *On se donne un programme linéaire sous forme standard et A_B une matrice de base admissible. Introduisons des notations $\overline{A} = A_B^{-1}A_N$, $\overline{b} = A_B^{-1}b$, $\overline{c}$ le vecteur des coûts réduits. On suppose qu'il existe $q_s \in N$ tel que $\overline{c}_s > 0$.*

Si $\overline{A}_{.s} \leq 0$ alors $x \mapsto \langle c, x\rangle$ n'est pas majorée sur l'ensemble $C(A, b)$.

Si $\overline{A}_{.s} \not\leq 0$ considérons $\lambda_0 = \min\{\overline{b}_i/\overline{a}_{is} \ / \ i = 1, \ldots, m, \ \overline{a}_{is} > 0\}$ et prenons r tel que $\overline{b}_r/\overline{a}_{rs} = \lambda_0$. Alors pour $B' = (p_1, \ldots, p_{r-1}, q_s, p_{r+1}, \ldots, p_m)$, $x' = (A_{B'}^{-1}b, 0_{N'})$ est un sommet où la fonction objectif est plus grande (au sens large) qu'en $x = (A_B^{-1}b, 0_N)$. Si $x = (A_B^{-1}b, 0_N)$ nest pas dégénéré on a $\langle c, x'\rangle > \langle c, x\rangle$.

Démonstration Nous avons vu que

$$x'_{p_i} = \overline{b}_i - \frac{\overline{a}_{is}}{\overline{a}_{rs}}\overline{b}_r, \text{ si } i \neq r, x'_{q_s} = \frac{\overline{b}_r}{\overline{a}_{rs}}, \ x'_j = 0 \text{ sinon.}$$

Comme $\overline{b}_r/\overline{a}_{rs} \leq \overline{b}_i/\overline{a}_{is}$, on a

$$x'_{p_i} = \overline{b}_i - \frac{\overline{a}_{is}}{\overline{a}_{rs}}\overline{b}_r \geq \overline{b}_i - \frac{\overline{a}_{is}}{\overline{a}_{is}}\overline{b}_i = 0 \text{ si } \overline{a}_{is} > 0,$$

et

$$x'_{p_i} = \overline{b}_i - \frac{\overline{a}_{is}}{\overline{a}_{rs}}\overline{b}_r \geq \overline{b}_i \text{ si } \overline{a}_{is} = 0.$$

Par ailleurs $x'_{q_s} = \overline{b}_r/\overline{a}_{rs}$ et les autres coordonnées de x' sont nulles, donc x' est un sommet de $C(A, b)$.

Calculons maintenant la fonction objectif en x' :

$$\begin{aligned}
\langle c, x' \rangle &= \langle c_{B'}, x'_{B'} \rangle \\
&= \langle c_{B'}, A_{B'}^{-1} b \rangle \\
&= \langle c_{B'}, E A_B^{-1} b \rangle \\
&= \langle {}^t E c_{B'}, A_B^{-1} b \rangle \\
&= \langle {}^t E c_{B'}, x_B \rangle \\
&= \sum_{i \neq r} c_{p_i} x_{p_i} + \sum_{i \neq r} c_{p_i} \eta_i x_{p_r} + c_{q_s} \eta_r x_{p_r} \\
&= \sum_{i} c_{p_i} x_{p_i} - c_{p_r} x_{p_r} + \sum_{i \neq r} c_{p_i} \eta_i x_{p_r} + c_{q_s} \eta_r x_{p_r} \\
&= \langle c_B, x_B \rangle + \left(\langle c_{B'}, \eta \rangle - c_{p_r} \right) x_{p_r}.
\end{aligned}$$

Or

$$\begin{aligned}
0 < \overline{c}_s &= (c_N - {}^t (A_B^{-1} A_N) c_B)_s = c_{q_s} - \langle \overline{A}_{\cdot s}, c_B \rangle \\
&= c_{q_s} - \sum_i c_{p_i} \overline{a}_{is} = c_{q_s} - \sum_{i \neq r} c_{p_i} \overline{a}_{is} - c_{p_r} \overline{a}_{rs} \\
&= \overline{a}_{rs} \left(c_{q_s} / \overline{a}_{rs} - \sum_{i \neq r} c_{p_i} \overline{a}_{is} / \overline{a}_{rs} - c_{p_r} \right) \\
&= \overline{a}_{rs} \left(c_{q_s} \eta_r + \sum_{i \neq r} c_{p_i} \eta_i - c_{p_r} \right) \\
&= \overline{a}_{rs} \left(\langle c_{B'}, \eta \rangle - c_{p_r} \right).
\end{aligned}$$

On en déduit que $\langle c_{B'}, \eta \rangle - c_{p_r}$ est strictement positif, et comme x_{p_r} l'est aussi (car on suppose que le sommet n'est pas dégénéré), on a

$$\langle c, x' \rangle > \langle c_B, x_B \rangle = \langle c, x \rangle.$$

□

3.8.2 Exemples (sous forme de tableaux)

En pratique l'algorithme du simplexe fait partie des fonctions classiques des programmes de calculs usuels (comme les tableurs). Pour décrire comment il fonctionne nous allons le décrire sous la forme de tableaux successifs sur quelques exemples. Considérons le programme linéaire suivant

$$\max x_1 + 2x_2$$

sous contraintes

$$x_1 \leq 4,\ 2x_1 + x_2 \leq 10,\ -x_1 + x_2 \leq 5,\ x_1, x_2 \geq 0.$$

Nous introduisons des variables supplémentaires x_3, x_4, x_5 pour mettre le programme sous forme standard :

$$\max x_1 + 2x_2$$

sous contraintes

$$x_1 + x_3 = 4,\ 2x_1 + x_2 + x_4 = 10,\ -x_1 + x_2 + x_5 = 5,\ x_1, x_2, x_3, x_4, x_5 \geq 0.$$

Sous forme matricielle ce programme linéaire s'écrit :

$$\max\langle c, x\rangle \ s.c.\ Ax = b, x \geq 0,$$

où x est dans $\mathbb{R}^5$, c et b dans $\mathbb{R}^3$,

$$A = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 \\ 2 & 1 & 0 & 1 & 0 \\ -1 & 1 & 0 & 0 & 1 \end{pmatrix} \quad x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} \quad c = \begin{pmatrix} 1 \\ 2 \\ 0 \\ 0 \\ 0 \end{pmatrix} \quad b = \begin{pmatrix} 4 \\ 10 \\ 5 \end{pmatrix}.$$

à partir de ces données nous construisons le tableau suivant que nous allons transformer à partir des règles de pivotage définies plus haut.

1	2	0	0	0	0	
1	0	1	0	0	4	x_3
2	1	0	1	0	10	x_4
-1	1	0	0	1	5	x_5

Les traits du tableau définissent quatre zones : la ligne en haut à gauche est le vecteur des coûts réduits, en bas à gauche on reconnaît A, en bas à droite b et des noms de variables. Les noms de variables indiquent quel sommet du polyèdre est considéré. Ici c'est le sommet dont les coordonnées sont $(0, 0, 4, 10, 5)$. En ce sommet la fonction objectif est nulle : c'est la signification du 0 figurant en haut à droite.

Pour choisir comment nous allons transformer ce tableau on considère les coordonnées positives du vecteur des coûts réduits figurant (en rouge ci-dessous) et nous choisissons un pivot parmi les coefficients positifs dans les colonnes correspondantes (en vert).

1	**2**	0	0	0	0	
1	0	1	0	0	4	x_3
2	**1**	0	1	0	10	x_4
-1	**1**	0	0	1	5	x_5

Ici on peut choisir l'une ou l'autre des deux premières colonnes. Si l'on choisit la première on calcule $4/1 = 4$ et $10/2 = 5$ pour pivot il faut prendre le 1 (en violet ci-dessous)

1	2	0	0	0	0	
1	0	1	0	0	4	x_3
2	1	0	1	0	10	x_4
-1	1	0	0	1	5	x_5

Si l'on choisit la deuxième colonne on calcule $10/1 = 10$ et $5/1 = 5$ pour pivot il faut prendre le deuxième 1 (en violet ci-dessous)

1	2	0	0	0	0	
1	0	1	0	0	4	x_3
2	1	0	1	0	10	x_4
-1	**1**	0	0	1	5	x_5

Comment expliquer le critère de choix du pivot ? C'est expliqué plus haut dans le cas général. Comprenons-le sur notre exemple. Rappelons que l'algorithme est une façon de passer de sommet en sommet d'un polyèdre contenu dans l'ensemble des points à coordonnées positives. Choisir le pivot comme on le fait assure que les coordonnées de la partie en bas à droite dans le tableau restent positives (donc qu'on reste dans le polyèdre) quand on applique le pivot de Gauss. Considérons notre exemple.

1	2	0	0	0	0	
1	0	1	0	0	**4**	x_3
2	1	0	1	0	**10**	x_4
-1	1	0	0	1	5	x_5

Pour faire apparaître un 0 à la place du 2 nous retranchons deux fois la deuxième ligne à la troisième ; pour le dernier coefficient cela donne $10 - 2 \times 4$ qui est positif. Dire que cette quantité est positive est équivalent à dire que $4/1$ est inférieur à $10/2$. Si on essaye de faire le calcul avec 2 comme pivot, pour remplacer le 1 par 0 on retranchera la moitié de la troisième ligne à la deuxième ; le dernier coefficient de la ligne obtenue sera : $4 - 10/2 = -1 < 0$! Un nombre négatif ici ne sera jamais pris comme pivot car pour obtenir 1 il faudrait diviser toute la ligne par ce pivot négatif et le dernier coefficient deviendrait négatif ; on serait sorti du polyèdre. Remarquons enfin que pour obtenir 0 à la place des éléments négatifs de la colonne du pivot, il faut ajouter un multiple positif de la ligne du pivot (la dernière coordonnée reste donc bien positive). Pour les 0 de la colonne il n'y a rien à faire (donc là aussi la dernière coordonnées reste positive).

Revenons à l'algorithme. Supposons qu'on ait choisit le pivot 1 de la première colonne. On utilise ce pivot pour annuler tous les coefficients de la colonne correspondante en faisant des opérations sur les lignes (première ligne moins la deuxième, troisième moins deux fois la deuxième, quatrième plus la deuxième). On obtient

0	2	-1	0	0	-4	
1	0	1	0	0	4	x_1
0	1	-2	1	0	2	x_4
0	1	1	0	1	9	x_5

On indique en dernière colonne que le nouveau sommet obtenu a ses coordonnées x_1, x_4, x_5 différentes de 0 et x_2 x_3 sont nulles (car on a pivoté sur le coefficient correspondant à x_1 là où précédemment on considérait x_3). Les coordonnées de la base considérée à une étape de l'algorithme sont les trois pour lesquelles les trois colonnes donnent la matrice identité. Le nombre en haut à droite est égal à l'opposé de la valeur de la fonction objectif au sommet considéré : ici cette valeur est donc 4 (on est passé de 0 à 4). On continue de la même façon.

0	**2**	-1	0	0	-4	
1	0	1	0	0	4	x_1
0	**1**	-2	1	0	2	x_4
0	**1**	1	0	1	9	x_5

0	2	-1	0	0	-4	
1	0	1	0	0	4	x_1
0	**1**	-2	1	0	2	x_4
0	1	1	0	1	9	x_5

0	0	3	-2	0	-8	
1	0	1	0	0	4	x_1
0	1	-2	1	0	2	x_2
0	0	3	-1	1	7	x_5

0	0	3	-2	0	-8	
1	0	1	0	0	4	x_1
0	1	-2	1	0	2	x_2
0	0	**3**	-1	1	7	x_5

0	0	0	-1	-1	-15	
1	0	0	1/3	-1/3	5/3	x_1
0	1	0	1/3	2/3	20/3	x_2
0	0	1	-1/3	1/3	7/3	x_3

Tous les coûts réduits sont négatifs. Nous avons trouvé le sommet qui maximise la fonction objectif : $(5/3, 20/3, 7/3, 0, 0)$, et la valeur du maximum : 15.

Si on avait choisi la deuxième colonne on aurait obtenu :

1	2	0	0	0	0	
1	0	1	0	0	4	x_3
2	1	0	1	0	10	x_4
-1	**1**	0	0	1	5	x_5

3	0	0	0	-2	-10	
1	0	1	0	0	4	x_3
3	0	0	1	-1	5	x_4
-1	1	0	0	1	5	x_2

3	0	0	0	-2	-10	
1	0	1	0	0	4	x_3
3	0	0	1	-1	5	x_4
-1	1	0	0	1	5	x_2

0	0	0	-1	-1	-15	
1	0	1	-1/3	1/3	7/3	x_3
1	0	0	1/3	-1/3	5/3	x_1
0	1	0	1/3	2/3	20/3	x_2

C'est bien la même solution que celle que nous avions trouvé en choisissant la première colonne. Nous n'avons pas parcouru le même chemin de sommets pour parvenir à l'optimum mais nous aboutissons finalement au même point.

Considérons un autre exemple (celui de l'introduction) :

$$\min 3x_1 + 5x_2$$

sous contraintes

$$2x_1 + x_2 \geq 3,\ 2x_1 + 2x_2 \geq 5,\ x_1 + 4x_2 \geq 4,\ x_1, x_2 \geq 0.$$

Considérons le problème dual

$$\max 3y_1 + 5y_2 + 4y_3$$

sous contraintes

$$2y_1 + 2y_2 + y_3 \leq 3,\ y_1 + 2y_2 + 4y_3 \leq 5,\ y_1, y_2, y_3 \geq 0.$$

La version standard de ce problème sous forme canonique est

$$\max 3y_1 + 5y_2 + 4y_3$$

sous contraintes

$$2y_1 + 2y_2 + y_3 + s_1 = 3,\ y_1 + 2y_2 + 4y_3 + s_2 = 5,\ y_1, y_2, y_3, s_1, s_2 \geq 0.$$

Construisons le tableau correspondant et appliquons l'algorithme (pour chaque étape j'indique seulement le pivot choisi puis le résultat des calculs correspondants)

3	5	4	0	0	0	
2	2	1	1	0	3	s_1
1	2	4	0	1	5	s_2

3	5	4	0	0	0	
2	2	1	1	0	3	s_1
1	2	**4**	0	1	5	s_2

2	3	0	0	-1	-5	
7/4	3/2	0	1	0	7/4	s_1
1/4	1/2	1	0	1/4	5/4	x_3

2	3	0	0	-1	-5	
7/4	3/2	0	1	0	7/4	s_1
1/4	1/2	1	0	1/4	5/4	x_3

0	9/7	0	-8/7	-1	-7	
1	6/7	0	4/7	0	1	x_1
0	2/7	1	-1/7	1/4	1	x_3

0	9/7	0	-8/7	-1	-7	
1	**6/7**	0	4/7	0	1	x_1
0	2/7	1	-1/7	1/4	1	x_3

-3/2	0	0	-14/7	-1	-17/2	
7/6	1	0	2/3	0	7/6	x_2
-1/3	0	1	-2/21	1/4	2/3	x_3

L'algorithme est fini : la fonction objectif est maximale au sommet $(0, 7/6, 2/3, 0, 0)$ et sa valeur maximale est $17/2$.

Voyons comment sur un exemple :

$$\max x_1 - x_2 + 2x_3 + x_4 + 3x_5 + 2x_6$$

sous contraintes

$$2x_1 - 3x_2 + x_4 + x_6 = 3,\ -x_1 + 2x_2 + x_3 + x_5 = 1,\ -3x_1 - 5x_2 - x_5 - 2x_6 = -4,$$

$$x_1, x_2, x_3, x_4, x_5, x_6 \geq 0.$$

On introduit trois variables artificielles s_1, s_2, s_3 et on pose le problème

$$\max -s_1 - s_2 - s_3$$

sous contraintes

$$2x_1 - 3x_2 + x_4 + x_6 + s_1 = 3,\ -x_1 + 2x_2 + x_3 + x_5 + s_2 = 1,\ 3x_1 + 5x_2 + x_5 + 2x_6 = 4,$$

$$x_1, x_2, x_3, x_4, x_5, x_6, s_1, s_2, s_3 \geq 0.$$

(on arrange les signes pour que le vecteur dont les coordonnées x soient nulle et les coordonnées s soient celles de b multipliées par -1 si elles sont négatives (pour que le vecteur obtenu soit ≥ 0)). On applique l'algorithme jusqu'à ce que les coordonnées s_i soient nulles. Sous forme de tableau il s'écrit

0	0	0	0	0	0	-1	-1	-1	0	
2	-3	0	1	0	1	**1**	0	0	3	s_1
-1	2	1	0	1	0	0	**1**	0	1	s_2
3	5	0	0	1	2	0	0	**1**	4	s_3

On veut que les nombres verts soient des pivots : il faut annuler les -1 de la première ligne en lui ajoutant les trois autres. On obtient

4	4	1	1	2	3	0	0	0	8	
2	-3	0	1	0	1	**1**	0	0	3	s_1
-1	2	1	0	1	0	0	**1**	0	1	s_2
3	5	0	0	1	2	0	0	**1**	4	s_3

Cela signifie que le point $(0, 0, 0, 0, 0, 0, 3, 1, 4)$ est un sommet du polyèdre défini par les contraintes sur $x_1, x_2, x_3, x_4, x_5, x_6, s_1, s_2, s_3$. On va maintenant appliquer l'algorithme du simplexe jusqu'à ce que les coordonnées correspondant aux s_i soient nulles (si cela ne se produit pas alors cela signifie que le polyèdre défini par les contraintes de départ est vide).

4	4	1	1	2	3	0	0	0	8	
2	-3	0	**1**	0	1	1	0	0	3	s_1
-1	2	1	0	1	0	0	1	0	1	s_2
3	5	0	0	1	2	0	0	1	4	s_3

2	7	1	0	2	2	-1	0	0	5	
2	-3	0	1	0	1	1	0	0	3	x_4
-1	2	1	0	1	0	0	1	0	1	s_2
3	5	0	0	1	2	0	0	1	4	s_3

2	7	1	0	2	2	-1	0	0	5	
2	-3	0	1	0	1	1	0	0	3	x_4
-1	2	**1**	0	1	0	0	1	0	1	s_2
3	5	0	0	1	2	0	0	1	4	s_3

3	5	0	0	1	2	-1	-1	0	4	
2	-3	0	1	0	1	1	0	0	3	x_4
-1	2	1	0	1	0	0	1	0	1	x_3
3	5	0	0	1	2	0	0	1	4	s_3

3	5	0	0	1	2	-1	-1	0	4	
2	-3	0	1	0	1	1	0	0	3	x_4
-1	2	1	0	**1**	0	0	1	0	1	x_3
3	5	0	0	1	2	0	0	1	4	s_3

4	3	-1	0	0	2	-1	-2	0	3	
2	-3	0	1	0	1	1	0	0	3	x_4
-1	2	1	0	1	0	0	1	0	1	x_5
4	3	-1	0	0	2	0	0	-1	3	s_3

4	3	-1	0	0	2	-1	-2	0	3	
2	-3	0	1	0	1	1	0	0	3	x_4
-1	2	1	0	1	0	0	1	0	1	x_5
4	3	-1	0	0	2	0	0	-1	3	s_3

0	0	0	0	0	0	-1	-2	1	0	
0	-9/2	1/2	1	0	0	1	0	1/2	3/2	x_4
0	11/4	3/4	0	1	1/2	0	1	-1/4	7/4	x_5
1	3/4	-1/4	0	0	1/2	0	0	-1/4	3/4	x_1

Nous avons maintenant un sommet du polyèdre de départ $(3/4, 0, 0, 3/2, 7/4, 0)$. On revient au problème de départ avec ce sommet.

1	-1	2	1	3	2	0	
0	-9/2	1/2	1	0	0	3/2	x_4
0	11/4	3/4	0	1	1/2	7/4	x_5
1	3/4	-1/4	0	0	1/2	3/4	x_1

0	-11/2	-1/2	0	0	0	-15/2	
0	-9/2	1/2	1	0	0	3/2	x_4
0	11/4	3/4	0	1	1/2	7/4	x_5
1	3/4	-1/4	0	0	1/2	3/4	x_1

Les coûts réduits sont négatifs ou nuls. La fonction objectif est optimale au sommet $(3/4, 0, 0, 3/2, 7/4, 0)$ et la valeur maximale est $15/2$.

3.8.3 Cyclage

Le phénomème de cyclage peut survenir lorsque l'algorithme aboutit à un sommet dégénéré : il peut alors proposer un changement de base qui ne modifie pas la valeur de la fonction objectif et revenir au coup suivant à la même base. L'algorithme boucle alors sur une valeur qui n'est pas optimale. Voici un exemple pour lequel ce phénomène se produit.

$$\max 4/5x_1 - 18x_2 - x_3 - x_4$$

sous contraintes

$$16/5x_1 - 84x_2 - 12x_3 + 8x_4 \leq 0,\ 1/5x_1 - 5x_2 - 2/3x_3 + 1/3x_4 \leq 0,\ x_1 \leq 1\ x_1, x_2, x_3, x_4 \geq 0.$$

$$\max 4/5x_1 - 18x_2 - x_3 - x_4$$

sous contraintes

$$16/5x_1 - 84x_2 - 12x_3 + 8x_4 + x_5 = 0,\ 1/5x_1 - 5x_2 - 2/3x_3 + 1/3x_4 + x_6 = 0,\ x_1 + x_7 = 1$$

$$x_1, x_2, x_3, x_4, x_4, x_6, x_7 \geq 0.$$

4/5	-18	-1	-1	0	0	0	0	
16/5	-84	-12	8	1	0	0	0	x_5
1/5	-5	-2/3	1/3	0	1	0	0	x_6
1	0	0	0	0	0	1	1	x_7

0	3	2	-3	-1/4	0	0	0	
1	-105/4	-15/4	5/2	5/16	0	0	0	x_1
0	**1/4**	1/12	-1/6	-1/16	1	0	0	x_6
0	105/4	15/4	-5/2	-5/16	0	1	1	x_7

0	0	1	-1	1/2	-12	0	0	
1	0	**5**	-15	-25/4	105	0	0	x_1
0	1	1/3	-2/3	-1/4	4	0	0	x_2
0	0	-5	15	25/4	-105	1	1	x_7

-1/5	0	0	2	7/4	-33	0	0	
1/5	0	1	-3	-5/4	21	0	0	x_3
-1/15	1	0	**1/3**	1/6	-3	0	0	x_2
1	0	0	0	0	0	1	1	x_7

1/5	-6	0	0	3/4	-15	0	0	
-2/5	9	1	0	**1/4**	-6	0	0	x_3
-1/5	3	0	1	1/2	-9	0	0	x_4
1	0	0	0	0	0	1	1	x_7

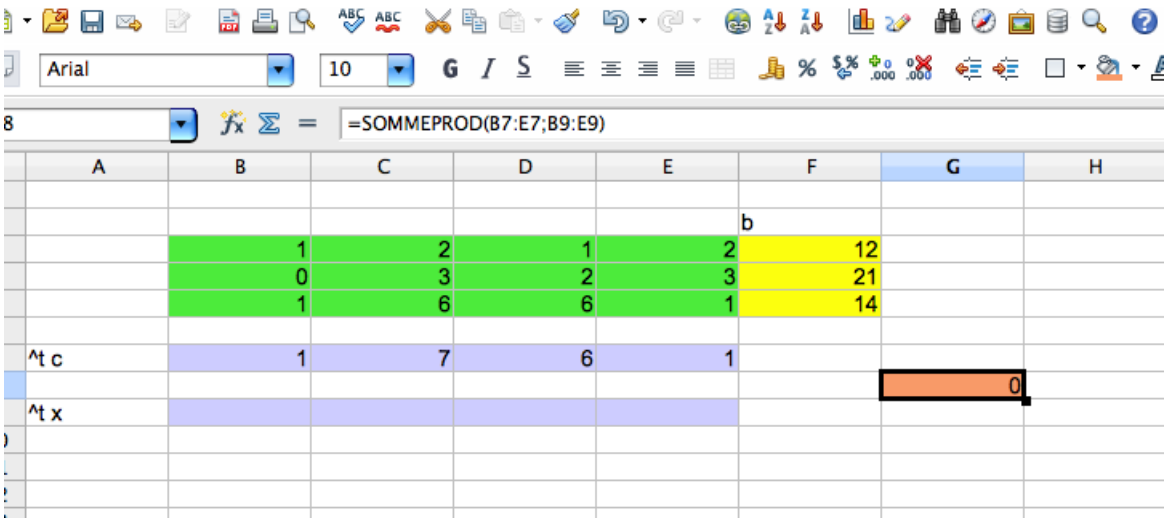
7/5	-33	-3	0	0	3	0	0	
-8/5	36	4	0	1	-24	0	0	x_5
3/5	-15	-2	1	0	**3**	0	0	x_4
1	0	0	0	0	0	1	1	x_7

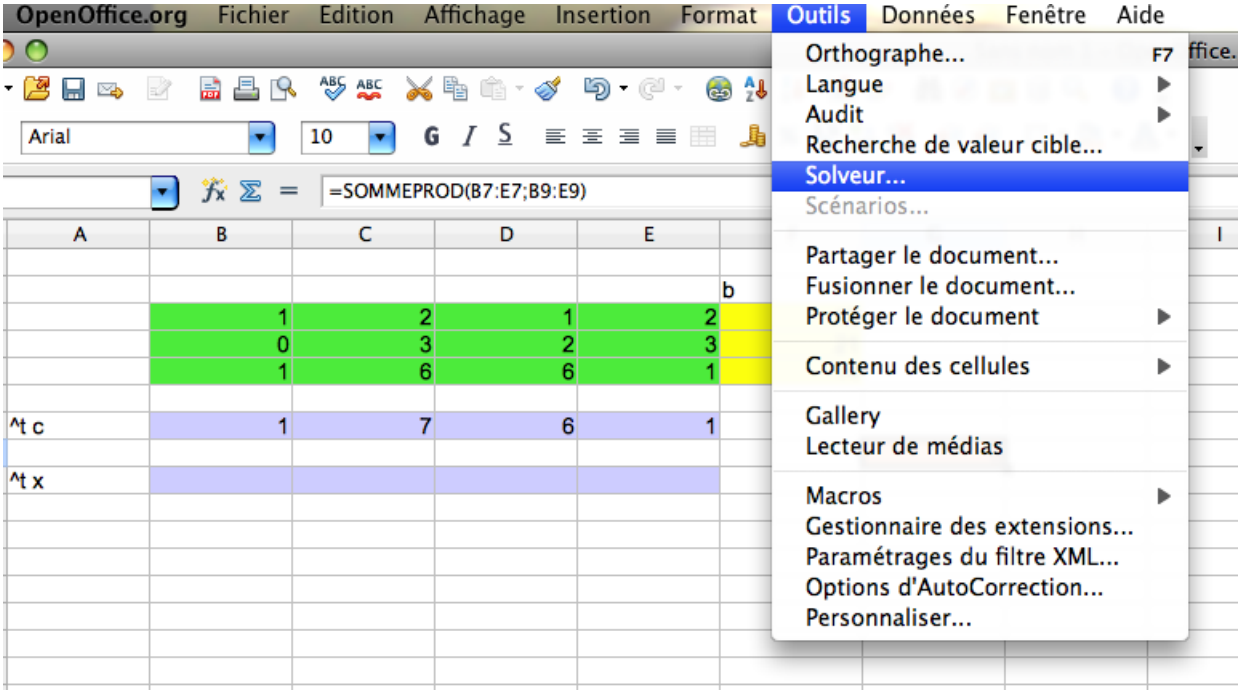
4/5	-18	-1	-1	0	0	0	0	
16/5	-84	-12	8	1	0	0	0	x_5
1/5	-5	-2/3	1/3	0	1	0	0	x_6
1	0	0	0	0	0	1	1	x_7

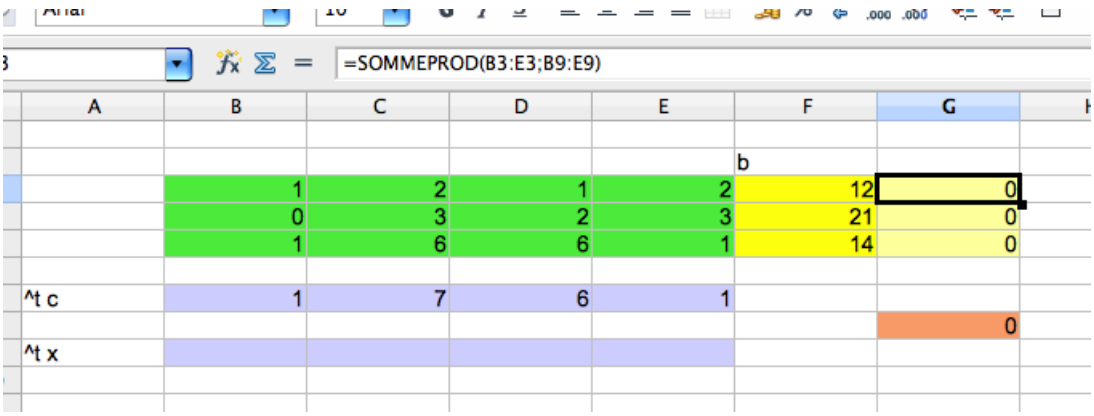
Nous sommes revenus au tableau de départ ! Tout ça pour rien... On dit qu'il y a eu cyclage. Cela peut se produire quand les sommets parcourus sont dégénérés.

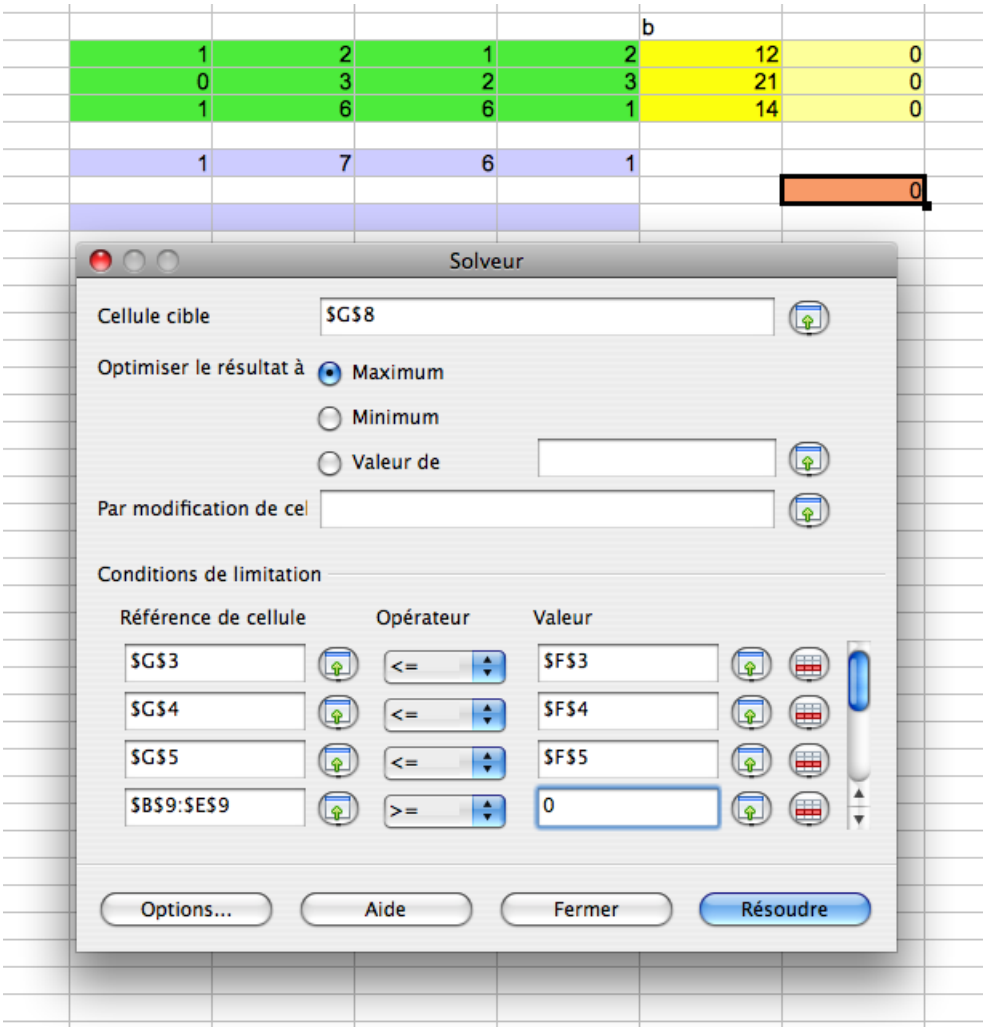
3.9 Utilisation du solveur d'un tableur

Ce que nous avons montré est que les problèmes de programmation linéaire pouvaient se résoudre grâce à des opérations simples sur des tableaux et que des algorithmes de résolution existaient. La résolution de tels problèmes peut se faire grâce à un tableur. Je recopie quelques photos d'écran qui donnent une idée de la méthode à suivre. Vous trouverez facilement plus de détails sur internet...




OpenOffice.org Fichier Edition Affichage Insertion Format Outils Données Fenêtre Aide
Arial 10
=SOMMEPROD(B7:E7;B9:E9)
Orthographe... F7
Langue
Audit
Recherche de valeur cible...
Solveur...
Scénarios...
Partager le document...
Fusionner le document...
Protéger le document
Contenu des cellules
Gallery
Lecteur de médias
Macros
Gestionnaire des extensions...
Paramétrages du filtre XML...
Options d'AutoCorrection...
Personnaliser...
A B C D E
b
1 2 1 2
0 3 2 3
1 6 6 1
^t c 1 7 6 1
^t x


=SOMMEPROD(B3:E3;B9:E9)
A B C D E F G
b
1 2 1 2 12 0
0 3 2 3 21 0
1 6 6 1 14 0
^t c 1 7 6 1
0
^t x



3.10 Jeux matriciels

On se donne une matrice A $m \times n$ qui donne les gains du joueur 1 (et les pertes du joueur 2) en fonction de leurs stratégies. On pose

$$g(x, y) = \sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij} x_i y_j.$$

On considère la fonction comme définie sur les simplexes de probabillité. La quantité est l'espérance du gain du jouer 1 lorsqu'il choisit sa stratégie en utilisant le vecteur de probabilité x et lorsque le joueur 2 choisit sa stratégie en utilisant le vecteur de probabilité y. On définit les deux quantité suivantes

$$\underline{V} = \sup_x \inf_y g(x, y)$$

$$\overline{V} = \inf_y \sup_x g(x, y)$$

Théorème 3.18. *Les bornes définissant* $\underline{V}$ *et* $\overline{V}$ *sont atteintes (ce sont des* min *et* max*) et on a*

$$\underline{V} = \overline{V}$$

Démonstration L'inégalité

$$\underline{V} \leq \overline{V}$$

s'obtient facilement. On a évidemment

$$\forall x, \forall y \quad \inf_s g(x,s) \leq g(x,y).$$

On en déduit

$$\forall y, \quad \sup_t \inf_s g(t,s) \leq \sup_t g(t,y),$$

puis

$$\sup_t \inf_s g(t,s) \leq \inf_y \sup_t g(t,y).$$

L'égalité contraire est plus difficile à montrer. Considérons un nombre r strictement inférieur à $\overline{V}$. Considérons l'ensemble

$$C = \cup_{y\in\Delta_2}[g(e_1,y),+\infty[\times[g(e_2,y),+\infty[\ldots\times[g(e_m,y),+\infty[$$

où les e_i sont les éléments de la base canonique. L'ensemble C est convexe (c'est la linéarité par rapport à la deuxième variable de g qui permet de l'affirmer) et fermé. Comme par hypothèse on a $\overline{V} > r$

$$\forall y, \ \exists x \quad g(x,y) > r.$$

Comme $x \mapsto g(x,y)$ est une forme linéaire, son maximum sur le simplexe des x possibles est atteint en l'un des sommets du simplexe (c'est-à-dire en l'un des e_i). On en déduit

$$\forall y \ \exists i \ g(e_i,y) > r.$$

Le vecteur $w = (r,r,\ldots,r)$ n'appartient donc pas à C. On peut séparer w de C par un hyperplan : il existe $u \neq 0$ tel que

$$\forall c \in C, \ \langle u,w\rangle < \langle u,c\rangle$$

Les éléments de C ont pour coordonnées des nombres c_i de la forme $c_i = g(e_i,y) + d_i$, avec $d_i \geq 0$. On a donc

$$\forall y \in \Delta_2, \ i, \ d_i \geq 0, \quad r\sum_{j=1}^m u_j \leq \sum_{i=1}^m u_i(g(e_i,y)+d_i).$$

En faisant tendre les nombres d_i vers $+\infty$ on voit que les nombre u_i sont positifs ou nuls (sinon on peut faire tendre le membre de droite vers $-\infty$). La somme des u_i positifs ou

nuls mais pas tous nuls est donc strictement positive. En divisant par cette somme on obtient

$$r \leq \sum_{i=1}^{m} \frac{1}{\sum_{j=1}^{m} u_j} u_i (g(e_i, y) + d_i)$$

et (par linéarité de g en la première variable)

$$r \leq g(\sum_{i=1}^{m} \frac{u_i}{\sum_{j=1}^{m} u_j} e_i, y) + \sum_{i=1}^{m} \frac{u_i}{\sum_{j=1}^{m} u_j} d_i.$$

Pour des d_i nuls, on obtient

$$\forall y \in \Delta_2,\ r \leq g(\sum_{i=1}^{m} \frac{u_i}{\sum_{j=1}^{m} u_j} e_i, y).$$

On a donc trouvé un x dans Δ_1 ($x = \sum_{i=1}^{m} \frac{u_i}{\sum_{j=1}^{m} u_j} e_i$) tel que

$$\forall y,\ r \leq g(x, y).$$

On en déduit que

$$\max_x \min_y g(x, y) \geq r.$$

Ce que nous venons de faire montre que pour tout r tel que $r < \overline{V}$ on a $r \leq \underline{V}$. Cela signifie que $\underline{V} \geq \overline{V}$. □Par définition des min et max, il existe x_* tel que, quel que soit y, $g(x_*, y) \geq \underline{V}$. De même, il existe y_* tel que, quel que soit x, $g(x, y_*) \leq \overline{V}$. Comme $\underline{V} = \overline{V}$ on obtient donc $g(x_*, y_*) = \underline{V} = \overline{V}$ et

$$\forall x,\ \forall y,\ g(x, y_*) \leq g(x_*, y_*) \leq g(x_*, y).$$

En choisissant la stratégie x, le premier joueur s'assure un gain minimal de $g(x_*, y_*)$ et en choisissant y_*, le deuxième joueur s'assure une perte minimale de $g(x_*, y_*)$.

Le théorème précédent montre que n'importe quel jeu matriciel a une solution. Mais comment trouver une solution ? La programmation linéaire permet de le faire. Voyons comment.

4 Optimisation différentielle

4.1 Optimisation libre

Une page de Lagrange :

DE MAXIMIS ET MINIMIS. 5

A est négatif il sera un *maximum;* si $A = 0$ on suivra les règles données (1).

4. Les variables contenues dans Z soient deux, savoir t et u; alors

$$d^2Z = A\,dt^2 + 2B\,dt\,du + C\,du^2.$$

Il paraît au premier aspect bien difficile de connaître si cette expression d^2Z doit être positive ou négative, sans qu'on ait le rapport de dt à du, qui n'est pas donné; car, puisqu'en changeant ce rapport la fonction d^2Z doit aussi varier, il semble indubitable qu'elle pourra aussi passer du positif au négatif, et du négatif au positif, pendant que les quantités A, B, C restent les mêmes. Qu'on donne cependant à la proposée

$$A\,dt^2 + 2B\,dt\,du + C\,du^2$$

cette forme

$$A\left(dt + \frac{B\,du}{A}\right)^2 + \left(C - \frac{B^2}{A}\right)du^2;$$

et on verra que, comme les carrés $\left(dt + \frac{B\,du}{A}\right)^2$ et du^2 ont toujours le même signe +, toute la quantité sera nécessairement positive si les deux coefficients A et $C - \frac{B^2}{A}$ sont positifs, et au contraire elle deviendra négative, lorsque ceux-ci seront tous deux négatifs, quel que soit le rapport de dt à du. On aura donc pour le cas du *minimum*

$$A > 0, \quad C - \frac{B^2}{A} > 0,$$

savoir

$$C > \frac{B^2}{A} \quad \text{ou} \quad CA > B^2,$$

ce qui donne de même

$$C > 0;$$

à moins donc que les quantités A, B, C n'aient ces conditions

$$A > 0, \quad C > 0 \quad \text{et} \quad AC > B^2,$$

4.1.1 En dimension 1

La majoration précédente s'écrit donc ici :

$$|f(x+h) - f(x) - hf'(x) - f''(x)h^2/2| \leq h^3 \max_{s\in[x,x+h]} |f'''(s)|/6.$$

Notons M un majorant de la dérivée troisième de f sur l'intervalle d'étude de f et supposons qu'en x la dérivée de f soit nulle. Supposons que $f''(x)$ ne soit pas nul, par exemple

qu'il soit positif. Prenons $h > 0$, on a alors :

$$f(x) + f''(x)h^2/2 - |h|^3 M/6 \leq f(x+h) \leq f(x) + f''(x)h^2/2 + |h|^3 M/6,$$

soit

$$f(x) + h^2(f''(x)/2 - |h|M/6) \leq f(x+h) \leq f(x) + h^2(f''(x)/2 + |h|M/6).$$

Alors, pour $|h|$ suffisamment petit, $f''(x)/2 - |h|M/6$ est positif. Par exemple si $0 < |h| < f''(x)/M$, $f''(x)/2 - |h|M/6 > f''(x)/2 - f''(x)/M.M/6 = f''(x)/3 > 0$. On a alors, pour tout $h \in [-f''(x)/M, f''(x)/M]$, $f(x) + h^2 f''(x)/3 \leq f(x+h)$. Autrement dit sur l'intervalle $[x - f''(x)/M, x + f''(x)/M]$, f atteint sa plus petite valeur en x. C'est exactement dire que f a un minimum local en x.

La recherche des **extrema locaux** pour une fonction d'une variable :

(i) On recherche les points critiques ($f'(x) = 0$).

(ii) On étudie la dérivée seconde f'' si a est un point critique et si
$f''(a) > 0$ il y a un minimum local,
$f''(a) < 0$ il y a un maximum local,
$f''(a) = 0$ il faut approfondir l'étude.

4.1.2 En dimension 2

Définition 4.1. *Soit $f(x, y)$ une fonction de classe $\mathcal{C}^2$. La* **matrice hessienne** *de f en (x_0, y_0) est la matrice* $\text{Hess}(x_0, y_0) = \begin{pmatrix} A & B \\ B & C \end{pmatrix}$

où $A = \dfrac{\partial^2 f}{\partial x^2}(x_0\,,\,y_0)$, $B = \dfrac{\partial^2 f}{\partial x\,\partial y}(x_0\,,\,y_0)$, $C = \dfrac{\partial^2 f}{\partial y^2}(x_0\,,\,y_0)$.

Exemple
Calculer la matrice Hessienne de $f(x\,,\,y) = 4xy - x^4 - y^4$.

Théorème 4.2. *(Formule de Taylor à l'ordre 2, en $X = (x_0, y_0)$)*

$$f(X+H) = f(X) + \nabla f(X) \cdot H + \frac{1}{2} H^t \ \text{Hess}(x_0, y_0)\ H + \|H\|^2 \varepsilon(H)$$

Théorème 4.3. *Soient $f(x\,,\,y)$ de classe $\mathcal{C}^2$ et $(x_0\,,\,y_0)$ un point critique. Soit $\triangle = AC - B^2 = \det(\text{Hess}(x_0, y_0))$. Alors :*
si $\triangle > 0$ et $A > 0$, f a un minimum local en $(x_0\,,\,y_0)$
si $\triangle > 0$ et $A < 0$, f a un maximum local en $(x_0\,,\,y_0)$
si $\triangle < 0$, f n'a ni maximum ni minimum, elle a un point selle
si $\triangle = 0$ on ne peut conclure (avec le seul développement à l'ordre 2).

La marche à suivre pour étudier les extrema d'une fonction différentiable sur un compact de $\mathbb{R}^2$ est la suivante.

Soit f une fonction différentiable définie sur un compact K de $\mathbb{R}^2$. Comme f est différentiable, elle est continue. Elle est donc bornée sur K et atteint ses bornes.

En pratique (dans les exercices que je vous demanderai de résoudre en particulier) la fonction f sera donnée par une formule valable sur un certain sous-ensemble de $\mathbb{R}^2$ et le compact K sera inclus dans cet ensemble de définition.

On mène l'étude des extrema de f en plusieurs étapes. La première est d'étudier l'existence d'extrema locaux de f à l'intérieur de K. C'est pour cette étude qu'on utilisera le développement de Taylor à l'ordre 2 donné ci-dessus.

Mais cette étude n'est pas suffisante. Il faut aussi regarder ce qui se passe sur le bord de K. Pour cela on procède autrement.

Exemple
Etude de $f(x\,,\,y) = x^2 - y^2$ sur $K = \{(x\,,\,y) \in \mathbb{R}^2 \,/\, x^2 + y^2 \leqslant 1\}$. On procède de la manière suivante :

(i) On cherche les points critiques et les extrema locaux dans $Int(K)$.
On trouve un seul point stationnaire en $(0,0)$. Mais en $(0,0)$ f a un point selle. La fonction n'a donc pas d'extremum à l'intérieur de K. Mais comme K est compact et f est continue sur K, f est bornée sur K et atteint ses bornes sur K. Ce sera donc sur le bord de K.

(ii) On analyse f sur ∂K.
Une possibilité ici est de paramétrer le bord de K : le cercle de rayon 1 centré en $(0,0)$. On obtient : $f(\cos t, \sin t) = \cos^2 t - \sin^2 t = \cos(2t)$. On peut alors étudier les variations de cette fonction. On obtient qu'elle est maximum égale à 1 lorsque $2t$ est égal à 0 modulo 2π, minimum égale à -1 lorsque $2t$ vaut π modulo 2π. La fonction f atteint donc son maximum 1 aux points $(1,0)$ et $(-1,0)$ de K, son minimum -1 aux points $(0,1)$ et $(0,-1)$.

4.1.3 En dimension d

Théorème 4.4. *Soit f une fonction différentiable définie sur $\mathbb{R}^d$ à valeurs dans $\mathbb{R}$. Si f a un extremum local en x alors $\nabla f(x) = 0$.*

Les extrema libres sont à chercher parmi les points critiques.

Théorème 4.5. *Soit f une fonction deux fois différentiable définie sur $\mathbb{R}^d$ à valeurs dans $\mathbb{R}$. Supposons que f ait un point critique en x. Alors :*
si Hess$f(x)$ *est définie positive, f a un minimum local en x,*
si Hess$f(x)$ *est définie négative, f a un maximum local en x,*

si $\mathrm{Hess} f(x)$ *a des valeurs propres positives et des valeurs propres négatives,* f *a un point selle en* x*,*
dans les autres cas, les dérivées d'ordres 1 et 2 ne suffisent pas à déterminer la nature du point critique.

Pour déterminer la nature d'un point critique, on est donc amené à étudier le signe d'une forme quadratique. Donnons trois méthodes pour le faire. Imaginons que la hessienne soit la matrice

$$A = \begin{pmatrix} 2 & 3 & 1 \\ 3 & 4 & 1 \\ 1 & 1 & 2 \end{pmatrix}.$$

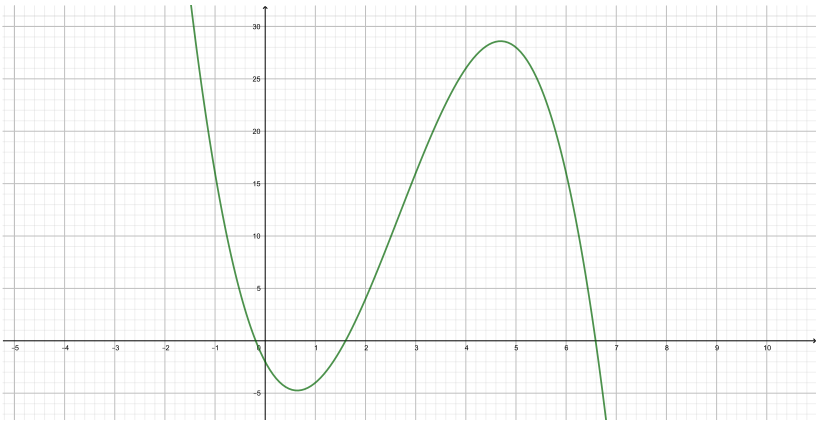
Déterminer le signe de la forme quadratique associée à cette matrice revient à déterminer les signes de ses valeurs propres.

Méthode 1 : calculer le polynôme caractéristique, en déduire les signes de ses racines.

Ici :

$$\begin{vmatrix} 2-x & 3 & 1 \\ 3 & 4-x & 1 \\ 1 & 1 & 2-x \end{vmatrix} = -x^3 + 8x^2 - 9x - 2 := \phi(x).$$

Une étude de fonction permet ensuite de déterminer les signes des racines de ϕ (même si on ne parvient pas à exprimer ces racines).



La courbe montre ici que ϕ a une racine négative et deux racines positives. Une étude de fonction le confirmerait.

Méthode 2 : décomposition en somme et différence de carrés de formes linéaires indépendantes (décomposition de Gauss). On calcule la forme quadratique associée à A :

$$\begin{pmatrix} x & y & z \end{pmatrix} \begin{pmatrix} 2 & 3 & 1 \\ 3 & 4 & 1 \\ 1 & 1 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = 2x^2 + 4y^2 + 2z^2 + 6xy + 2xz + 2yz.$$

Méthode 3 : calculer les mineurs principaux.

Les mineurs principaux sont les déterminants des matrices de tailles 1×1, 2×2, 3×3... obtenues en ne conservant que le coefficients de la première ligne et première colonne, des deux premières lignes et deux premières colonnes, des trois premières lignes et trois premières colonnes,... Ici par exemple les mineurs principaux sont

$$2, \begin{vmatrix} 2 & 3 \\ 3 & 4 \end{vmatrix}, \begin{vmatrix} 2 & 3 & 1 \\ 3 & 4 & 1 \\ 1 & 1 & 2 \end{vmatrix},$$

c'est-à-dire les nombres

$$2, -1, -2.$$

On en déduit que A a des valeurs propres positives et des valeurs propres négatives. C"est une conséquence de la proposition suivante.

Proposition 4.6. *Soit A une matrice symétrique réelle. Si tous les mineurs principaux de A sont strictement positifs alors A est définie positive. Si les mineurs principaux sont alternativement négatifs puis positifs, en commençant par négatif (sans jamais s'annuler) alors A est définie négative. Dans les autres cas A n'est ni définie positive ni définie négative.*

4.2 Convexité, concavité

Définition 4.7. *Soit C une partie de $\mathbb{R}^d$. On dit que C est convexe si lorsque x et y sont deux points de C, alors le segment joignant ces deux points $[x, y] = \{tx + (1-t)y \ / \ t \in [0, 1]\}$ est lui aussi inclus dans C.*

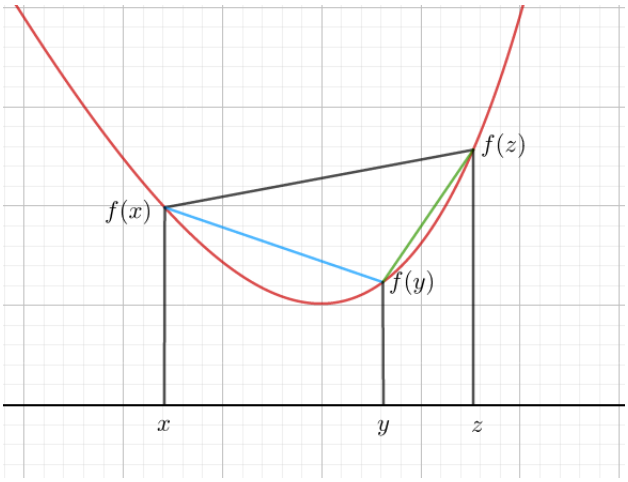
Définition 4.8. *Soit f une fonction définie sur une partie convexe C de $\mathbb{R}^d$. On dit que la fonction f est convexe si l'une des deux propriétés équivalentes suivantes est satisfaite :*
- *Pour tous x, y dans C, tout $t \in [0, 1]$, on a*

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

- *la partie de $\mathbb{R}^{d+1}$ définie par $\{(x, u) \in \mathbb{R}^d \times \mathbb{R} \ / \ u \geq f(x)\}$ est convexe (dans $\mathbb{R}^{d+1}$).*

En dimension 1

Caractérisation de la convexité au moyen d'inégalité sur les taux d'accroissement.



Proposition 4.9. *Soit f une fonction convexe définie sur un intervalle I et trois points x, y, z trois points de cet intervalle rangés dans l'ordre $x < y < z$. Alors*

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

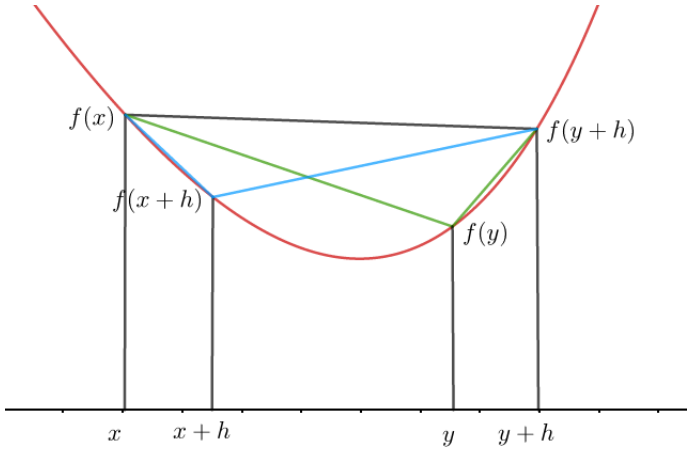
Les trois quotients apparaissant dans l'encadrement précédent sont les pentes des segments dessinés en bleu, noir et vert respectivement sur la figure ci-dessus.

D'après la définition d'une fonction convexe le point $(y, f(y))$ est en dessous du segment $[(x, f(x)); (z, f(z))]$. Cela entraîne l'encadrement des pentes énoncé.

Conséquences.

Théorème 4.10. *Si f est dérivable, alors f est convexe si et seulement si la dérivée est croissante. Si f est deux fois dérivable, alors f est convexe si la dérivée seconde est positive.*

Soit $x < y$ et $h > 0$ tel que $x + h < y$. On peut représenter la situation sur un dessin.



On applique la proposition précédente dans les triangles "noir et bleu" et "noir et vert". On obtient les encadrements

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

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

En particulier on a

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

En faisant tendre h vers 0 on obtient

$$f'(x) \leq f'(y).$$

La fonction f' est donc bien croissante.

Proposition 4.11. *Soit f une fonction convexe dérivable sur un intervalle I. Pour tous éléments x, y de I, on a*

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

Le théorème des accroissements finis assure l'existence d'un nombre θ appartenant à l'intervalle $]x, y[$ (ou $]y, x[$) tel que

$$f(y) = f(x) + f'(\theta)(y-x).$$

Si $x < y$, alors $x < \theta < y$ et (comme f' est croissante) $f'(x) \leq f'(\theta)$, donc (comme $y - x > 0$) $f'(x)(y-x) \leq f'(\theta)(y-x)$ et $f(y) \geq f(x) + f'(x)(y-x)$. Si $x > y$, alors $x > \theta > y$ et (comme f' est croissante) $f'(x) \geq f'(\theta)$, donc (comme $y - x < 0$) $f'(x)(y-x) \leq f'(\theta)(y-x)$ et $f(y) \geq f(x) + f'(x)(y-x)$.

Théorème 4.12. *Soit f une fonction convexe définie sur une partie convexe C. Si f a un minimum local en x_0 de C alors f a un minimum global en ce point.*

C'est un corollaire de la proposition précédente. Si $f'(x) = 0$ alors pour tout y on a

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

En dimension d.

Proposition 4.13. *Soit f une fonction définie sur une partie convexe C de $\mathbb{R}^d$ à valeurs dans $\mathbb{R}$. La fonction f est convexe si pour tous points x, y de C, la fonction ϕ définie sur $[0,1]$ par*

$$\phi(t) = f(tx + (1-t)y) = f(y + t(x-y))$$

est convexe.

Soient α, s, t trois points de $[0,1]$. On vérifie que les relations suivantes sont vraies (l'inégalité parce que f est convexe par hypothèse).

$$\begin{aligned}\phi(\alpha s+(1-\alpha)t) &= f(y+(\alpha s+(1-\alpha)t)(x-y))\\ &= f(\alpha y+(1-\alpha)y+(\alpha s+(1-\alpha)t)(x-y))\\ &= f(\alpha(y+s(x-y))+(1-\alpha)(y+t(x-y)))\\ &\leq \alpha f(y+s(x-y))+(1-\alpha)f(y+t(x-y))\\ &= \alpha\phi(s)+(1-\alpha)\phi(t).\end{aligned}$$

Supposons maintenant que les fonctions ϕ ainsi définies soient convexes. Pour tous x, y dans C et α dans $[0,1]$ on a

$$\begin{aligned}f(\alpha x+(1-\alpha)y) &= \phi(\alpha)\\ &= \phi(\alpha.1+(1-\alpha).0)\\ &\leq \alpha\phi(1)+(1-\alpha)\phi(0)\\ &= \alpha f(x)+(1-\alpha)f(y).\end{aligned}$$

Supposons que f soit différentiable deux fois. Alors ϕ l'est aussi. Calculons la dérivée de ϕ en utilisant la règle de la dérivation en chaîne :

$$\frac{d}{dt}\phi(t)=\frac{d}{dt}f(y+t(x-y))=\sum_{i=1}^{d}(x_i-y_i)\frac{\partial f}{\partial x_i}(y+t(x-y))=\langle\nabla f(y+t(x-y)),x-y\rangle.$$

Si f est convexe, ϕ aussi et on a

$$\phi(0)\geq\phi(1)+\phi'(1)(0-1)=\phi(1)-\phi'(1)$$

soit

$$f(y)\geq f(x)+\langle\nabla f(x),y-x\rangle.$$

Théorème 4.14. *Soit f une fonction convexe différentiable définie sur une partie compacte C de $\mathbb{R}^d$ à valeurs dans $\mathbb{R}$. Si x est un point critique de f alors f a un minimum global en x.*

En dérivant une deuxième fois on obtient

$$\frac{d^2}{dt^2}\phi(t)={}^t(x-y)\mathrm{Hess}f(y+t(x-y))(x-y).$$

On a vu que pour que ϕ soit convexe il faut que ϕ'' soit positive. Cela signifie que f est convexe si pour tous x, y

$${}^t(x-y)\mathrm{Hess}f(y+t(x-y))(x-y)\geq 0.$$

Autrement dit f est convexe si sa matrice hessienne est positive.

Théorème 4.15. *Soit f une fonction deux fois différentiable définie sur une partie compacte C de $\mathbb{R}^d$ à valeurs dans $\mathbb{R}$. Pour que f soit convexe il faut et il suffit que sa matrice hessienne soit positive en tout point de C.*

Rappelons que dire que la matrice hessienne est positive est dire que ses valeurs propres sont positives ou nulles.

4.3 Deux algorithmes

4.3.1 Méthode de Newton

Dans le paragraphe précédent nous avons vu comment trouver des conditions théoriques permettant de caractériser des extremums de fonctions. En pratique il faut résoudre des systèmes d'équations à plusieurs variables. Par exemple pour trouver les points critiques d'une fonction. Si f est une fonction différentiable de $\mathbb{R}^d$ dans $\mathbb{R}$, f a un point critique si son gradient est nul. Trouver les points critique revient donc à résoudre un système de d équations à d inconnues. Ce système n'est généralement pas linéaire et il est la plupart du temps impossible d'obtenir des formules donnant les solutions au moyen des fonctions classiques (polynômes, fractions, logarithmes, racines, exponentielles,...). On obtient des valeurs approchées grâce à des algorithmes et des ordinateurs. Nous allons donner deux exemples issus de la méthode de Newton. Trouver les points critiques d'une fonction f est trouver les points où la fonction ∇f (de $\mathbb{R}^d$ dans $\mathbb{R}^d$ s'annule). Commençons par le cas de la dimension 1.

En dimension 1.

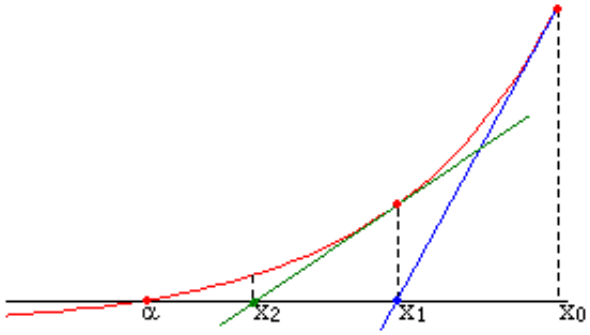
La méthode de Newton est une méthode itérative. On choisit un point x_0 dans l'ensemble de définition de f et on pose

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

Si (x_k) est bien définie (pour tout k) et converge vers x_* alors

$$x_* = x_* - \frac{f(x*)}{f'(x*)}, \text{ donc } f(x_*) = 0,$$

à condition que $f'(x_*) \neq 0$.



Nous avons déjà vu un exemple d'algorithme de Newton dans un cas très particulier : la méthode de Héron pour approcher $\sqrt{2}$. Dans ce cas la fonction considérée est $f(x) = x^2 - 2$ on obtient

$$x_{k+1} = x_k - \frac{x_k^2 - 2}{2x_k} = \frac{x_k^2 + 2}{2x_k} = \frac{1}{2}(x_k + \frac{2}{x_k}).$$

En enlevant $\sqrt{2}$ on obtient

$$x_{k+1} - \sqrt{2} = \frac{1}{2}(x_k + \frac{2}{x_k}) - \frac{1}{2}(\sqrt{2} + \frac{2}{\sqrt{2}}) = \frac{(x_k - \sqrt{2})^2}{2x_k}.$$

Supposons que x_0 soit supérieur à $\sqrt{2}$, alors pour tout k, x_k est supérieur à $\sqrt{2}$,

$$x_{k+1} - x_k = \frac{1}{2}(x_k + \frac{2}{x_k}) - x_k = \frac{1}{2}(\frac{2}{x_k} - x_k) = \frac{1}{2}\frac{2 - x_k^2}{x_k} < 0.$$

La suite (x_k) est donc convergente car décroissante minorée. De plus elle converge vers une solution de $l = \frac{1}{2}(l + \frac{2}{l})$ c'est-à-dire $\pm\sqrt{2}$. Comme les termes de (x_k) sont positifs c'est forcément $\sqrt{2}$. Si x_0 est inférieur à $-\sqrt{2}$

Imaginons que f ait une racine en x_* : $f(x_*) = 0$ que $f'(x_*)$ et $f''(x_*)$ soient différents de 0. Ecrivons le développement limité à l'ordre 2 entre x et x_* :

$$0 = f(x_*) = f(x) + f'(x)(x_* - x) + \frac{f''(\theta)}{2}(x_* - x)^2$$

$$x_{k+1} = x_k + \frac{f'(x_k)(x_* - x_k) + \frac{f''(\theta)}{2}(x_* - x_k)^2}{f'(x_k)} = x_* + \frac{f''(\theta)}{2f'(x_k)}(x_* - x_k)^2.$$

En dimension d.

La méthode de Newton en dimension d est un procédé itératif analogue. On se donne une fonction f de $\mathbb{R}^d$ dans $\mathbb{R}^d$ et on cherche un point où f s'annule. On se donne un point x_0 et on pose

$$x_{k+1} = x_k - f'(x_k)^{-1} f(x_k).$$

Ici $f'(x_k)$ est une matrice $d \times d$. Il peut être très coûteux en temps de calcul d'inverser une matrice $d \times d$ (et peu utile car si l'algorithme converge $f'(x_k)$ ne change plus beaucoup

lorsqu'on approche de la limite). On préfère généralement inverser moins souvent des matrices. Voici un résultat théorique qui assure la convergence de l'algorithme de Newton généralisé sous certaines conditions.

Théorème 4.16. *Soit f une fonction de $\mathbb{R}^d$ dans $\mathbb{R}^d$. Supposons que $f(x_*) = 0$ et que $f'(x_*)$, noté A, est inversible. Donnons-nous une suite de matrices inversibles A_k telle que, pour un certain nombre $\lambda < 1/2$, on ait*

$$\sup_k \|A_k - A\| \leq \frac{\lambda}{\|A^{-1}\|}.$$

Alors il existe un rayon $r > 0$, tel que si $x_0 \in B(x_*, r)$ et on pose, pour tout k,

$$x_{k+1} = x_k - A_k^{-1} f(x_k),$$

alors la suite (x_k) converge vers x_*. De plus la convergence est géométrique : il existe $\beta < 1$ tel que, pour tout k on ait

$$\|x_k - x_*\| \leq \beta^k \|x_0 - x_*\|.$$

Dans le cas de la recherche d'un point critique d'une fonction ϕ de $\mathbb{R}^d$ dans $\mathbb{R}$, on peut appliquer le théorème précédent avec la fonction $\nabla\phi$ de $\mathbb{R}^d$ dans $\mathbb{R}^d$. Ce théorème se reformule alors de la façon suivante.

Théorème 4.17. *Soit ϕ une fonction de $\mathbb{R}$ dans $\mathbb{R}^d$. Supposons que $\nabla\phi(x_*) = 0$ et que $\mathrm{Hess}\phi(x_*)$, noté A, est inversible. Donnons-nous une suite de matrices inversibles A_k telle que, pour un certain nombre $\lambda < 1/2$, on ait*

$$\sup_k \|A_k - A\| \leq \frac{\lambda}{\|A^{-1}\|}.$$

Alors il existe un rayon $r > 0$, tel que si $x_0 \in B(x_*, r)$ et on pose, pour tout k,

$$x_{k+1} = x_k - A_k^{-1} \nabla\phi(x_k),$$

alors la suite (x_k) converge vers x_*. De plus la convergence est géométrique : il existe $\beta < 1$ tel que, pour tout k on ait

$$\|x_k - x_*\| \leq \beta^k \|x_0 - x_*\|.$$

4.3.2 Méthode de descente

Commençons par décrire le principe de la méthode sur un exemple très (trop) simple.

On considère la fonction $f : x \mapsto x^2$. On sait évidemment que le minimum de cette fonction est 0 atteint en $x = 0$. Qu'est-ce que la méthode de descente de gradient ici ? On part d'un point u_0, disons -2, on fixe un pas h, disons $1/4$.

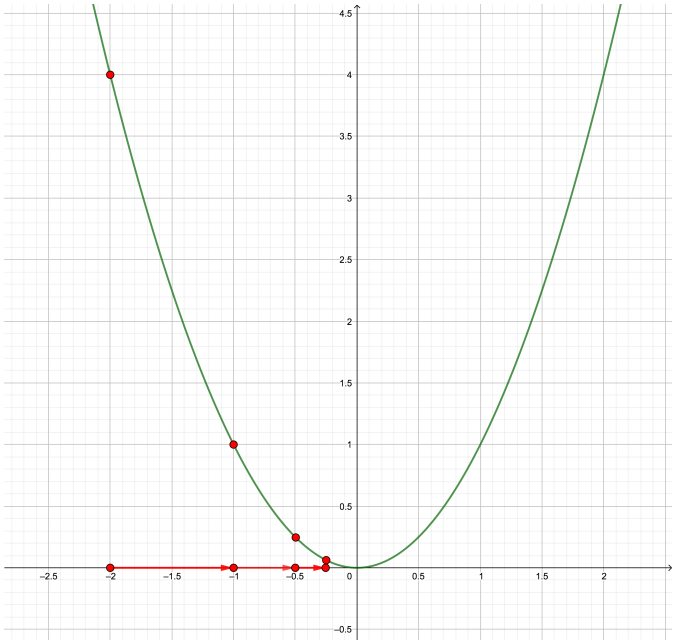
La définition de la dérivée de f en un point x donne l'approximation suivante (valable lorsque t est petit).

$$f(x+t) \simeq f(x) + tf'(x)$$

En prenant t de la forme $t = -hf'(x)$ on obtient

$$f(x - hf'(x)) \simeq f(x) - hf'(x)^2.$$

Ceci montre que pour h petit la valeur de f en $x - hf'(x)$ est plus petite que la valeur de f en x. Ceci donne une façon de trouver des valeurs de f de plus en plus petite. Lorsque la situation est favorable on pourra ainsi construire des suites qui approchent un point où la valeur de f est minimale. Ici on passe de $u_0 = -2$ à $u_1 = -2 - hf'(-2) = -2 - 1/4 * (-4) = -1$, $u_2 = -1 - hf'(-1) = -1 - 1/4 * (-2) = -1/2$, etc...



Ici les formules explicites sont très simples et on peut décrire précisément la suite u_n.

$$u_{n+1} = u_n - hf'(u_n) = u_n - 1/4 * 2u_n = u_n/2.$$

La suite u_n s'approche de 0 : à chaque nouveau pas la distance à zéro est divisée par deux.

Prenons un autre exemple en dimension quelconque. On considère la fonction

$$f(x) = c + \langle b, x\rangle + \frac{1}{2}\langle Ax, x\rangle$$

où c est une constante réelle $b \in \mathbb{R}^d$, A est une matrice symétrique définie positive. On peut calculer le gradient de f en un point x

$$\nabla f(x) = b + Ax.$$

L'unique minimum de la fonction (strictement convexe) f est obtenu au point x où le gradient s'annule, c'est-à-dire

$$x = -A^{-1}b.$$

Voyons ce que donne l'algorithme de descente ici.

$$f(x+t) \simeq f(x) + \langle \nabla f(x), t \rangle.$$

En prenant $t = -\alpha \nabla f(x)$ on obtient

$$f(x+t) \simeq f(x) + \langle \nabla f(x), -\alpha \nabla f(x) \rangle = f(x) - \alpha \|\nabla f(x)|^2.$$

La suite donnée par l'algorithme de descente est définie par récurrence

$$u_{n+1} = u_n - \alpha \nabla f(x) = u_n - \alpha(b + Au_n).$$

Nous allons comparer les distances de $d(u_{n+1}, -A^{-1}b)$ et $d(u_n, -A^{-1}b)$ ($-A^{-1}b$ est le point où f atteint son minimum). Pour cela exprimons le différence entre u_{n+1} et $-A^{-1}b$ en fonction de u_n. On obtient :

$$\begin{aligned} u_{n+1} + A^{-1}b &= u_n - \alpha(b + Au_n) + A^{-1}b \\ &= u_n - \alpha A(u_n + A^{-1}b) + A^{-1}b \\ &= u_n + A^{-1}b - \alpha A(u_n + A^{-1}b) \\ &= (Id - \alpha A)(u_n + A^{-1}b) \end{aligned}$$

Si A est une multiple de l'identité on peut avoir l'optimum en un pas : il suffit de choisir α pour que $Id - \alpha A$ soit nulle. Si A est diagonale appelons $\lambda_1, \ldots, \lambda_d$ ses coefficients diagonaux (tous positifs). Les cordonnées de $u_n + A^{-1}b$ sont multipliées par les nombres $1 - \alpha\lambda_i$. On souhaite rendre tous ces nombres le plus petit possible. Ils sont tous dans l'intervalle $[1 - \alpha\lambda_M, 1 - \alpha\lambda_m]$ où λ_m est le plus petit des nombres λ_i et λ_M le plus grand. Pour que les valeurs absolues de ces nombres soient le plus petit possible il faut choisir α pour que 0 soit au milieu de l'intervalle c'est-à-dire tel que $1 - \alpha\lambda_M = -(1 - \alpha\lambda_m)$ soit $\alpha^{-1} = \frac{\lambda_m + \lambda_M}{2}$. Tous les nombres $1 - \alpha\lambda_i$ sont alors inférieurs en valeur absolue à

$$\rho = \frac{\lambda_M - \lambda_m}{\lambda_M + \lambda_m} < 1,$$

et on a

$$d(u_n, -A^{-1}b) \leq \rho^n d(u_0, -A^{-1}b).$$

La convergence vers le minimum se fait donc à vitesse exponentielle. On remarque que le nombre ρ est d'autant plus petit que les nombres λ_i sont proches les uns des autres. Il peut être intéressant de modifier un problème pour obtenir des λ_i aussi proches que possibles les uns des autres (pour améliorer la vitesse de convergence). Dans le cas général A est diagonalisable dans une base orthonormée et... ? ? ? ?

4.3.3 Méthode du gradient conjugué

Article Discussion

Lire Modifier Modifier le code Voir l'historique

En analyse numérique, la méthode du gradient conjugué est un algorithme pour résoudre des systèmes d'équations linéaires dont la matrice est symétrique définie positive. Cette méthode, imaginée en 1950 simultanément par Cornelius Lanczos, Eduard Stiefel et Magnus Hestenes1, est une méthode itérative qui converge en un nombre fini d'itérations (au plus égal à la dimension du système linéaire). Toutefois, son grand intérêt pratique du point de vue du temps de calcul vient de ce qu'une initialisation astucieuse (dite « préconditionnement ») permet d'aboutir en seulement quelques passages à une estimation très proche de la solution exacte : c'est pourquoi, en pratique, on se borne à un nombre d'itérations bien inférieur au nombre d'inconnues.

La méthode du gradient biconjugué fournit une généralisation pour les matrices non symétriques.

4.4 Sous-ensembles de $\mathbb{R}^n$ et fonctions

Si f une fonction de deux variables son graphe est une surface incluse dans $\mathbb{R}^3$:

$$\{(x, y, f(x, y)) \ / \ (x, y) \in \mathbb{R}^2\}.$$

Une telle surface s'appelle une nappe paramétrée (par f).

Si f est une fonction partout dérivable alors son graphe admet en chaque point un plan tangent. Pour trouver des vecteurs appartenant au plan tangent en $(x_0, y_0, f(x_0, y_0))$ traçons deux courbes sur la surface dans des directions données par les coordonnées :

$$t \mapsto (x_0 + t, y_0, f(x_0 + t, y_0)) \quad t \mapsto (x_0, y_0 + t, f(x_0, y_0 + t))$$

et calculons les coordonnées de leurs vecteurs tangents à l'instant $t = 0$. On obtient

$$(1, 0, \frac{\partial f}{\partial x}(x_0, y_0)) \quad (0, 1, \frac{\partial f}{\partial y}(x_0, y_0)).$$

Ce sont deux vecteurs indépendants tangents à deux courbes tracées sur la nappe. Le plan tangent à la nappe paramétrée est le plan passant par (x_0, y_0) de direction engendrée par ces deux vecteurs. pour obtenir une équation de ce plan on peut utiliser le produit vectoriel. Un vecteur normal au plan est donné par

$$(1, 0, \frac{\partial f}{\partial x}(x_0, y_0)) \wedge (0, 1, \frac{\partial f}{\partial y}(x_0, y_0)) = (-\frac{\partial f}{\partial x}(x_0, y_0), -\frac{\partial f}{\partial y}(x_0, y_0), 1)$$

L'équation du plan tangent est donnée par :

$$-\frac{\partial f}{\partial x}(x_0, y_0)(x - x_0) - \frac{\partial f}{\partial y}(x_0, y_0)(y - y_0) + (z - f(x_0, y_0)) = 0$$

Plus généralement une nappe paramétrée est un ensemble décrit par deux paramètres

$$\{(f_1(s,t), f_2(s,t), f_3(s,t)) \ / \ (s,t) \in \mathbb{R}^2\}.$$

Les vecteurs

$$(\frac{\partial f_1}{\partial s}(s_0,t_0), \frac{\partial f_2}{\partial s}(s_0,t_0), \frac{\partial f_3}{\partial s}(s_0,t_0)) \quad (\frac{\partial f_1}{\partial t}(s_0,t_0), \frac{\partial f_2}{\partial t}(s_0,t_0), \frac{\partial f_3}{\partial t}(s_0,t_0))$$

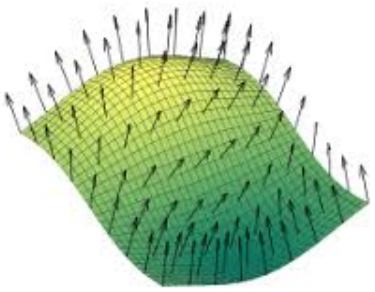
sont des vecteurs tangents à la surface au point image de (s_0,t_0). S'ils sont indépendants le paramétrage définit une surface en (s_0,t_0) et un vecteur normal à la surface est donné par le produit vectoriel des vecteurs précédents

$$(\frac{\partial f_2}{\partial s}\frac{\partial f_3}{\partial t} - \frac{\partial f_3}{\partial s}\frac{\partial f_2}{\partial t}, \frac{\partial f_3}{\partial s}\frac{\partial f_1}{\partial t} - \frac{\partial f_1}{\partial s}\frac{\partial f_3}{\partial t}, \frac{\partial f_1}{\partial s}\frac{\partial f_2}{\partial t} - \frac{\partial f_2}{\partial s}\frac{\partial f_1}{\partial t})(s_0,t_0)$$

Si on note (a,b,c) les coordonnées de ce vecteur l'équation du plan tangent à la nappe paramétrée au point image de (s_0,t_0) est

$$a(x - f_1(s_0,t_0)) + b(y - f_2(s_0,t_0)) + c(z - f_3(s_0,t_0)) = 0.$$

C'est simplement écrire que les vecteurs AM et (a,b,c) sont orthogonaux lorsque M a pour coordonnées (x,y,z) et A est le point image de (s_0,t_0), $(f_1(s_0,t_0), f_2(s_0,t_0), f_3(s_0,t_0))$.



Donnons-une fonction de deux variables f. Que dire des ensembles $\{(x,y) \ / \ f(x,y) = c\}$? On les appelle les courbes de niveaux de la fonction f. C'est dire qu'on s'attend à ce que ces ensembles soient des courbes.

Ce n'est toutefois pas toujours le cas. Si par exemple f est constante égale à 0, alors les courbes de niveau sont toutes vides sauf la courbe de niveau 0 qui est égale à $\mathbb{R}^2$ tout entier.

Si f est différentiable et son gradient n'est pas nul en (x_0,y_0) alors la courbe de niveau $f(x_0,y_0)$ définit bien une courbe au voisinage de (x_0,y_0). Cette courbe est régulière et l'équation de sa tangente est donnée par le gradient :

$$\frac{\partial f}{\partial x}(x_0,y_0)(x - x_0) + \frac{\partial f}{\partial y}(x_0,y_0)(y - y_0) = 0.$$

Par exemple, la tangente en $(1/\sqrt{3}, \sqrt{2/3})$ à la courbe de niveau 1 de la fonction $f(x, y) = x^2 + y^2$ est la droite d'équation :

$$2/\sqrt{3}(x - 1/\sqrt{3}) + 2\sqrt{2/3}(y - \sqrt{2/3}) = 0.$$

On a de la même façon les équations des plans tangents aux surfaces de niveaux de fonctions de trois variables dérivables au voisinage de points en lesquels le gradient n'est pas nul.

Par exemple le plan tangent à la surface $xyz = 1$ au point $(1/2, 1, 2)$ a pour équation :

$$2(x - 1/2) + (y - 1) + 1/2(z - 2) = 0.$$

Pourquoi l'équation du plan tangent est-elle donnée de cette façon par les dérivées partielle ? Voici une explication. Traçons une courbe $x(t) = (x_1(t), \ldots, x_n(t))$ sur la surface d'équation $f = 0$ passant en x_0 en $t = 0$ ($f(x_0) = 0$: le point appartient à la surface). On a alors, pour tout t,

$$c = f(x(t)).$$

En dérivant chaque membre de cette égalité (grâce à la règle de dérivation en chaîne pour le membre de droite), on obtient :

$$0 = \sum_{k=1}^{n} x_k'(t) \frac{\partial f}{\partial x_k}(x(t)),$$

ce qui donne, en $t = 0$,

$$0 = \sum_{k=1}^{n} x_k'(0) \frac{\partial f}{\partial x_k}(x_0).$$

Cela signifie que les vecteurs tangents en x_0 aux courbes tracées sur la surface sont orthogonaux au gradient de f en x_0. Le plan tangent est donc orthogonal au gradient ; on en déduit que son équation est

$$\sum_{k=1}^{n} \frac{\partial f}{\partial x_k}(x_0)(x_k - x_{0,k}) = 0.$$

Soient U un ouvert de $\mathbb{R}^n$, F une application de U dans $\mathbb{R}^n$ et $V = F(U) \subset \mathbb{R}^n$.

Définition 4.18. *F est* **inversible** *sur U s'il existe une application G de V dans $\mathbb{R}^n$ telle que $G \circ F = \mathbf{1}_U$ et $F \circ G = \mathbf{1}_V$.*

Exemples

(1) $f : \mathbb{R} \to \mathbb{R}$ avec $f(x) = x^3$

(2) $f : \mathbb{R} \to \mathbb{R}$ avec $f(x) = x^2$

(3) Si $A \in \mathbb{R}^n$, soit F de $\mathbb{R}^n$ dans $\mathbb{R}^n$ avec $F(X) = X + A$.

(4) $U = \{(r\,,\,\theta) \;/\; r > 0\,,\, 0 < \theta < \pi\}$
$F(r\,,\,\theta) = (r\,\cos\theta\,,\, r\,\sin\theta)$

Définition 4.19. *F de $\mathbb{R}^n$ dans $\mathbb{R}^n$ est* **localement inversible** *en $X \in \mathbb{R}^n$ s'il existe des ouverts U et V avec $X \in U$ et $F(X) \in V$ et $F(U) = V$ tel que F est inversible sur U.*

Théorème 4.20. *(d'inversion locale)*
Soient f définie sur un domaine de $\mathbb{R}^n$ à valeurs dans $\mathbb{R}^n$ de classe $\mathcal{C}^1$ et x_0 un point intérieur à D. Alors si $f'(x_0)$ est inversible (en tant qu'application linéaire) f est localement inversible en x_0. Si g désigne son inverse locale, g est aussi de classe C^1 et en $y = f(x)$ on a $g'(y) = f'(x)^{-1}$ (l'exposant désigne ici l'opération d'inversion d'une matrice).

Une démonstration de ce théorème est donnée en annexe.

Soit $f : \mathbb{R}^2 \to \mathbb{R}$. On considère la courbe de niveau $\{f(x\,,\,y) = 0\} = L_0\,.$

Définition 4.21. *On dit que la fonction $y = \varphi(x)$ est* **définie implicitement par** *$f(x\,,\,y) = 0$ si $f(x\,,\,\varphi(x)) = 0$, c'est-à-dire si $(x\,,\,\varphi(x)) \in L_0\,.$*
Alors on dit que $y = \varphi(x)$ est une **fonction implicite** *de $f(x\,,\,y) = 0$.*

Exemple
$f(x\,,\,y) = \ln(xy) - \sin x \quad$ avec $xy > 0$
$f(x\,,\,y) = x^2 + y^2 - 1$

Théorème 4.22. *(des fonctions implicites)*
Soient $f : \mathbb{R}^2 \to \mathbb{R}$ une fonction de classe $\mathcal{C}^1$ et $(x_0\,,\,y_0)$ un point tel que $f(x_0\,,\,y_0) = 0$.
Si $\dfrac{\partial f}{\partial y}(x_0\,,\,y_0) \neq 0$ alors :

(i) Il existe une fonction implicite $y = \varphi(x)$ de classe $\mathcal{C}^1$, définie sur un intervalle ouvert $B(x_0\,,\,\varepsilon)$, tel que $f(x\,,\,\varphi(x)) = 0$ et $y_0 = \varphi(x_0)$.

(ii) La dérivée de φ est donnée par $\varphi'(x) = \dfrac{-\frac{\partial f}{\partial x}(x\,,\,\varphi(x))}{\frac{\partial f}{\partial y}(x\,,\,\varphi(x))}$ en tout point où $\dfrac{\partial f}{\partial y}(x\,,\,\varphi(x)) \neq 0$.

Démonstration C'est une conséquence du théoréme d'inversion locale. Soit f une fonction C^1 de deux variables et (x_0, y_0) tel que $f(x_0, y_0) = 0$ et $\frac{\partial f}{\partial y}(x_0, y_0) \neq 0$. Considérons la fonction F définie par

$$F(x, y) = (x, f(x, y)).$$

La matrice jacobienne de F est

$$\begin{pmatrix} 1 & 0 \\ \frac{\partial f}{\partial x} & \frac{\partial f}{\partial y} \end{pmatrix}.$$

Par hypothèse $\frac{\partial f}{\partial y}$ ne s'annule pas en (x_0, y_0). La matrice $F'(x_0, y_0)$ est donc inversible et d'après le théorème d'inversion locale, F est localement inversible en (x_0, y_0) : il existe $r > 0$ tel que F soit une bijection de la boule $B = B((x_0, y_0), r)$ sur son image et l'application inverse, appelons la G est C^1 sur l'ouvert $F(B)$. Écrivons $G(s,t) = (G_1(s,t), G_2(s,t))$ les coordonnées de G. Comme G est l'inverse de F on a, pour tout (s,t) dans $F(B)$ (en utilisant la définition de F) :

$$(s,t) = F(G_1(s,t), G_2(s,t)) = (G_1(s,t), f(G_1(s,t), G_2(s,t))).$$

On a donc les égalités : $G_1(s,t) = s$ et $f(s, G_2(s,t)) = t$. Les points (x,y) de B pour lesquels $f(x,y) = 0$ sont les points dont l'image par F est de la forme $(x,0)$. Ce sont donc les points $G(x,0)$ pour $(x,0)$ dans $F(B)$, soit encore, d'après la forme de l'application G, les points $(x, G_2(x,0))$ pour $(x,0)$ dans $F(B)$. Or $F(B)$ est un ouvert contenant $(x_0, 0)$. Il existe donc $\alpha > 0$ tel que, pour $x \in]x_0 - \alpha, x_0 + \alpha[$, $(x,y) \in B$, l'équation $f(x,y) = 0$ équivaut à $y = G_2(x,0)$. Il suffit d'écrire $\phi(x) = G_2(x,0)$ pour voir qu'on a bien établi le résultat souhaité. □

Exemple : étude au point $(1\,,\,1)$ de $f(x\,,\,y) = x^2\,y + 3y^3\,x^4 - 4$

Théorème 4.23. *Si $f\colon \mathbb{R}^n \to \mathbb{R}$ est de classe $\mathcal{C}^1$ et si $\dfrac{\partial f}{\partial x_n}(X_0) \neq 0$ alors :*

(i) La fonction implicite $x_n = \varphi(x_1 \ldots x_{n-1})$ existe sur une boule ouverte $B((x_{1,0} \ldots x_{n-1,0})\,,\,\varepsilon)$ et on a : $f(x_1 \ldots x_{n-1}\,,\,\varphi(x_1 \ldots x_{n-1})) = 0$.

(ii) $$\frac{\partial \varphi}{\partial x_i} = \frac{-\frac{\partial f}{\partial x_i}(x_1 \ldots x_{n-1}\,,\,\varphi(x_1 \ldots x_{n-1}))}{\frac{\partial f}{\partial x_n}(x_1 \ldots x_{n-1}\,,\,\varphi(x_1 \ldots x_{n-1}))}$$

4.5 Extrema liés (multiplicateur de Lagrange)

4.5.1 Une seule contrainte

Il s'agit de trouver les extrema de $f(x\,,\,y\,,\,z)$ lorsque $(x\,,\,y\,,\,z)$ appartient à une surface S définie par $g(x\,,\,y\,,\,z) = C$.

Exemple
Maximiser $x^2\,y^2\,z^2$ lorsque $x^2 + y^2 + z^2 = 1$.

Définition 4.24. *Un point $P = (x_0\,,\,y_0\,,\,z_0)$ est un minimum (resp. maximum) local pour f, lié à la contrainte $g(x\,,\,y\,,\,z) = C$ si :*

(i) $g(P) = C$

(ii) Il existe $\varepsilon > 0$ tel que $f(P) \leqslant f(Q)$ (resp. $f(P) \geqslant f(Q)$) pour tout $Q \in S \cap B(P\,,\,\varepsilon)$.

Théorème 4.25. *(de Lagrange)*
Soit $f(x\,,\,y\,,\,z)$ et $g(x\,,\,y\,,\,z)$ de classe $\mathcal{C}^1$ telle que $\nabla g \neq 0$ sur S.
Alors si f admet un **extremum lié** *en $(x_0\,,\,y_0\,,\,z_0)$ on a : $\nabla f(x_0\,,\,y_0\,,\,z_0) = \lambda\,\nabla g(x_0\,,\,y_0\,,\,z_0)$ où $\lambda \in \mathbb{R}$ est appelé multiplicateur de Lagrange.*

Démonstration Ce qui suit n'est qu'une idée de démonstration qu'il faudrait préciser.
Soit $(x(t), y(t), z(t))$ une courbe tracée sur la surface S passant en (x_0, y_0, z_0) en $t = 0$. Pour que f soit maximale en (x_0, y_0, z_0) sur S il faut en particulier que $t \mapsto f(x(t), y(t), z(t)$ soit maximale en $t = 0$. Pour que ce soit le cas il faut que cette fonction ait une dérivée nulle en 0. La dérivée de cette fonction se calcule en appliquant la règle de dérivation en chaîne, $x'(t)\frac{\partial f}{\partial x} + y'(t)\frac{\partial f}{\partial y} + z'(t)\frac{\partial f}{\partial z}$. On obtient la condition

$$x'(0)\frac{\partial f}{\partial x}(x_0\,,\,y_0\,,\,z_0) + y'(0)\frac{\partial f}{\partial y}(x_0\,,\,y_0\,,\,z_0) + z'(0)\frac{\partial f}{\partial z}(x_0\,,\,y_0\,,\,z_0) = 0.$$

Par ailleurs, comme S est définie par $g = C$ le plan tangent à S en (x_0, y_0, z_0) a pour équation :

$$(x - x_0)\frac{\partial g}{\partial x}(x_0\,,\,y_0\,,\,z_0) + (y - y_0)\frac{\partial g}{\partial y}(x_0\,,\,y_0\,,\,z_0) + (z - z_0)\frac{\partial g}{\partial z}(x_0\,,\,y_0\,,\,z_0) = 0.$$

On a aussi

$$x'(0)\frac{\partial g}{\partial x}(x_0\,,\,y_0\,,\,z_0) + y'(0)\frac{\partial g}{\partial y}(x_0\,,\,y_0\,,\,z_0) + z'(0)\frac{\partial g}{\partial z}(x_0\,,\,y_0\,,\,z_0) = 0.$$

Lorsqu'on fait varier $(x(t), y(t), z(t))$ le vecteur $(x'(0), y'(0), z'(0))$ prend toutes les valeurs de la direction du plan tangent à S en $(x_0\,,\,y_0\,,\,z_0)$. Cela signifie que $\nabla f(x_0\,,\,y_0\,,\,z_0)$ est orthogonal à tous les vecteurs orthogonaux à $\nabla g(x_0\,,\,y_0\,,\,z_0)$ ou encore que $\nabla f(x_0\,,\,y_0\,,\,z_0)$ et $\nabla g(x_0\,,\,y_0\,,\,z_0)$ ont mêmes vecteurs orthogonaux. Cela entraîne que $\nabla f(x_0\,,\,y_0\,,\,z_0)$ et $\nabla g(x_0\,,\,y_0\,,\,z_0)$ sont colinéaires. □

Remarque
Si P est un extremum lié, on a $\nabla f(P)$ parallèle à $\nabla g(P)$. La réciproque n'est pas vraie. Nous avons une condition nécessaire mais pas suffisante. C'est l'équivalent de la nullité de la dérivée pour les extrema libres : en un extremum libre la dérivée est nulle mais la dérivée peut être nulle sans que la fonction ait un extremum (penser à $x \mapsto x^3$ en $x = 0$).

Exemple
Sur l'exemple précédent on montre la méthode de résolution.

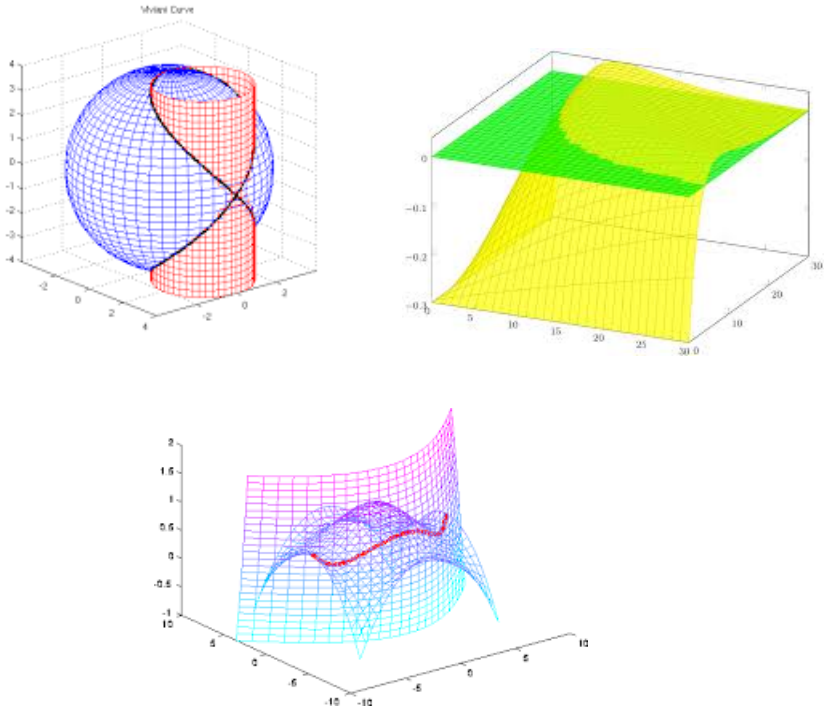
Le théorème est encore vrai en dimension plus grande.

Théorème 4.26. *(de Lagrange)*
Soit f et g deux fonctions de $\mathbb{R}^d$ dans $\mathbb{R}$ de classe $\mathcal{C}^1$ telle que $\nabla g \neq 0$ sur l'ensemble S défini par $g = C$ (c'est une hypersurface).
Alors si f admet un **extremum lié** *en x_0 on a : $\nabla f(x_0) = \lambda \nabla g(x_0)$ où $\lambda \in \mathbb{R}$ est appelé multiplicateur de Lagrange.*

4.5.2 Plusieurs contraintes

On cherche les extrema d'une fonction f sur l'ensemble S défini par $g_1 = g_2 = \ldots = g_k = 0$, toutes les fonctions considérées étant de classe C^1. Pour que les choses marchent bien il

faut faire l'hypothèse suivante : en tout point de S les gradients des fonctions g_i sont linéairement indépendants.



Théorème 4.27. *(de Lagrange)*
Soit f et $g_1, \ldots, g_k$ $k+1$ fonctions de $\mathbb{R}^d$ dans $\mathbb{R}$ de classe $\mathcal{C}^1$ telles que les vecteurs $\nabla g_1, \ldots, \nabla g_k$, soit indépendants sur sur l'ensemble S défini par $g_1 = \ldots = g_k = 0$.
Alors si f admet un **extrema lié** *sur S en x_0 le vecteur $\nabla f(x_0)$ est combinaison linéaire des vecteurs $\nabla g_i(x_0)$: il existe $\lambda_1, \ldots, \lambda_k$ tels que $\nabla f(x_0) = \sum_{i=1}^k \lambda_i \nabla g_i(x_0)$. Les nombres $\lambda_1, \ldots, \lambda_k$ sont appelés multiplicateurs de Lagrange.*

Démonstration On montre de la même manière que précédemment $\nabla f(x_0)$ est orthogonal à tous les vecteurs orthogonaux à tous les vecteurs $\nabla g_i(x_0)$. Attention, c'est là qu'il faut utiliser l'hypothèse d'indépendance sur les gradients. Considérons une base de $\mathbb{R}^d$ constituée de $v_1 = \nabla g_1(x_0), \ldots, v_k = \nabla g_k(x_0)$ complétée par une famille $(v_{k+1}, \ldots, v_d)$ de vecteurs orthogonaux entre eux et orthogonaux à tous les $\nabla g_i(x_0)$. Écrivons $\nabla f(x_0)$ dans cette base : $\nabla f(x_0) = \sum_{i=1}^d \lambda_i v_i$. Comme les vecteurs v_i sont orthogonaux deux à deux, pour tout j on a

$$\begin{aligned} \langle \nabla f(x_0), v_j \rangle &= \langle \sum_{i=1}^d \lambda_i v_i, v_j \rangle \\ &= \sum_{i=1}^d \lambda_i \langle v_i, v_j \rangle \\ &= \lambda_j \langle v_j, v_j \rangle. \end{aligned}$$

Mais, par ailleurs, on a vu que, pour $j > k$, on avait $\langle \nabla f(x_0), v_j \rangle = 0$. On en déduit donc que, pour $j > k$, λ_j est nul. $\square$

Caractérisation des fonctions convexes et concaves à partir des dérivées secondes.

4.5.3 Gradient projeté

4.6 Conditions de Karush-Kuhn-Tucker

Dans le théorème de Karush, Kuhn et Tucker, la contrainte K est de la forme

$$K = \{x \in \mathbb{R}^d / g_i(x) = 0, i \in I, h_j(x) \leq 0, j \in J\}$$

où $I = \{1, ..., p\}$ indexe les contraintes d'égalité et $J = \{1, ..., l\}$ indexe les contraintes d'inégalité. Les fonctions g_i et h_i sont toutes supposées de classe C^1 de $\mathbb{R}^d$ dans $\mathbb{R}$. Pour tout $x \in K$, on appelle contraintes actives (ou saturées) les indices $j \in \{1, ..., l\}$ tels que $h_j(x) = 0$:

$$J(x) = \{j \in \{1, ..., l\} \ / \ h_j(x) = 0\}.$$

Théorème 4.28. *Si un point x^* est un minimum du problème, alors il existe un vecteur $(\lambda_0, \ldots, \lambda_m, \mu_1, \ldots, \mu_l)$ de norme 1 tel que*

$$\lambda_0 \nabla f(x^*) + \sum_{j=i}^{p} \lambda_i \nabla g_i(x^*) + \sum_{j=1}^{l} \mu_j \nabla h_j(x^*) = 0.$$

Les nombres $\lambda_0, \ldots, \lambda_p, \mu_1, \ldots, \mu_l$ sont appelés multiplicateurs de Lagrange. On a $\lambda_0 \geq 0$ et $\mu_j \geq 0$, et $\mu_j = 0$ si $h_j < 0$.

Démonstration Pour simplifier les notations on peut supposer que $x = 0$ et que $f(x) = 0$ (cela ne diminue pas la généralité de la démonstration car il suffit de faire une translation dans l'espace de départ et une translation dans l'espace d'arrivée pour se ramener à ce cas). Supposons que pour $r \leq q$ $h_1, \ldots, h_r$ soient actives, et que h_j soit inactive si $j > q$ (là encore on ne diminue pas la généralité de la démonstration, il suffit de changer la numérotation des h_i pour se ramener à un cas de ce type). Soit δ tel que d'une part $h_j(y) < 0$ si $\|y\| < \delta$ et d'autre part $f(y) \geq 0$ si $\|y\| < \delta$ et si les contraintes sont satisfaites. Posons $h_i^+ = \max(h_i, 0)$.

Pour tout $\epsilon \in]0, \delta[$, il existe $\alpha > 0$ tel que, pour tout y de norme ϵ, on ait

$$f(y) + \|y\|^2 + \alpha \sum_{i=1}^{p} g_i(y)^2 + \alpha \sum_{j=1}^{r} h_j^+(y)^2 > 0 \ (*).$$

En effet supposons le contraire, c'est-à-dire qu'il existe $\epsilon \in]0, \delta[$ tel que pour tout $\alpha > 0$, il existe y de norme ϵ tel que

$$f(y) + \|y\|^2 + \alpha \sum_{i=1}^{p} g_i(y)^2 + \alpha \sum_{j=1}^{r} h_j^+(y)^2 \leq 0.$$

Considérons alors une suite (α_k) de nombres strictement positifs tendant vers $+\infty$ (par exemple $\alpha_k = k$). Pour tout k il existe y_k de norme ϵ tel que

$$f(y_k) + \|y_k\|^2 + \alpha_k \sum_{i=1}^{p} g_i(y_k)^2 + \alpha_k \sum_{j=1}^{r} h_j^+(y_k)^2 \leq 0.$$

Quitte à extraire une sous-suite on peut supposer que y_k converge vers y_∞ (car la sphère de rayon ϵ est compacte). Par continuité on a alors

$$\sum_{i=1}^{p} g_i(y_k)^2 + \sum_{j=1}^{r} h_j^+(y_k)^2 \leq 0,$$

donc

$$\sum_{i=1}^{p} g_i(y_k)^2 + \sum_{j=1}^{r} h_j^+(y_k)^2 = 0,$$

ce qui signifie que y_∞ satisfait les contraintes. Par hypothèse on a donc $f(y_\infty) \geq 0$ (car f a 0 pour minimum sur la boule de rayon ϵ centrée en 0). Mais par ailleurs, pour tout k on a

$$f(y_k) + \|y_k\|^2 \leq 0, \text{ donc } f(y_k) \leq -\|y_k\|^2 = -\epsilon^2, \text{ et } f(y_\infty) \leq -\epsilon^2.$$

Nous avons donc une contradiction.

Posons

$$F(y) = f(y) + \|y\|^2 + \alpha \sum_{i=1}^{p} g_i(y)^2 + \alpha \sum_{j=1}^{r} h_j^+(y)^2 > 0,$$

où α vérifie $(*)$ et cherchons le minimum de F sur la boule fermée $\overline{B(0,\epsilon)}$. Comme $F(0) = 0$ (0 satisfait les contraintes) et $F(y) > 0$ si $\|y\| = \epsilon$, le minimum de F sur $\overline{B(0,\epsilon)}$ se trouve à l'intérieur de $\overline{B(0,\epsilon)}$. C'est donc un extremum libre. En ce minimum, le gradient de F s'annule. Appelons u le point où F est minimale sur $\overline{B(0,\epsilon)}$. On a :

$$\nabla F(u) = \nabla f(u) + 2u + \alpha \sum_{i=1}^{p} 2g_i(u)\nabla g_i(u) + \alpha \sum_{j=1}^{r} 2h_j^+(u)\nabla h_j(u) = 0$$

En divisant cette égalité par la norme du vecteur $(1, 2\alpha g_1(u), \ldots, 2\alpha h_1^+(u), \ldots)$, on obtient

$$\lambda_0(\nabla f(u) + 2u) + \sum_{i=1}^{p} \lambda_i \nabla g_i(u) + \sum_{j=1}^{r} \mu_j \nabla h_j(u) = 0$$

où λ_0 et les μ_j sont positifs ou nuls et le vecteur $(\lambda_0, \lambda_1, \ldots, \mu_1, \ldots)$ est de norme 1.

Considérons alors une suite ϵ_n tendant vers 0. Par le raisonnement précédent on construit une suite (u_n) tendant vers 0 et des vecteurs de norme 1 $(\lambda_0^{(n)}, \lambda_1^{(n)}, \ldots, \mu_1^{(n)}, \ldots)$ où $\lambda_0^{(n)}$ et les $\mu_j^{(n)}$ sont positifs ou nuls, telle que pour tout n on ait

$$\lambda_0^{(n)}(\nabla f(u_n) + 2u_n) + \sum_{i=1}^{p} \lambda_i^{(n)} \nabla g_i(u_n) + \sum_{j=1}^{r} \mu_j^{(n)} \nabla h_j(u_n) = 0.$$

Comme la sphère unité est compacte, quitte à extraire une sous-suite, on peut supposer que le vecteur $(\lambda_0^{(n)}, \lambda_1^{(n)}, \ldots, \mu_1^{(n)}, \ldots)$ a une limite $(\lambda_0^{(\infty)}, \lambda_1^{(\infty)}, \ldots, \mu_1^{(\infty)}, \ldots)$. À la limite on obtient

$$\lambda_0^{(\infty)} \nabla f(0) + \sum_{i=1}^{p} \lambda_i^{(\infty)} \nabla g_i(0) + \sum_{j=1}^{r} \mu_j^{(\infty)} \nabla h_j(0) = 0.$$

□

Définition 4.29. *On dit que la contrainte K est qualifiée en un point x^* si, pour tous λ_i, $\mu_j \geq 0$ tels que*

$$\sum_{i=1}^{p} \mu_i h_i(x^*) = 0$$

et

$$\sum_{i=1}^{p} \lambda_i \nabla g_i(x^*) + \sum_{j=1}^{l} \mu_j \nabla h_j(x^*) = 0,$$

tous les λ_i et μ_j sont nécessairement nuls.

Proposition 4.30. *Dans le théorème précédent, si les contraintes sont qualifiées en x^*, on peut prendre $\lambda_0 = 1$.*

Démonstration Pour prendre $\lambda_0 = 1$, il suffit de diviser l'égalité par λ_0. Le seul obstacle possible est que λ_0 soit nul. Ce cas est exclu si la contrainte est qualifiée (car le vecteur $(\lambda_0, \lambda_1, \ldots, \mu_1, \ldots)$ n'est pas nul). □

Si les gradients $\nabla g_i(x^*)$ sont linéairement indépendants et s'il existe un vecteur v orthogonal à tous les $\nabla g_i(x^*)$ tel que $\langle v, \nabla h_j(x^*) \rangle < 0$ pour toutes les contraintes h_j active en x^*, alors les contraintes sont qualifiées en x^*. En effet en prenant le produit scalaire par v la deuxième égalité de la définition donne

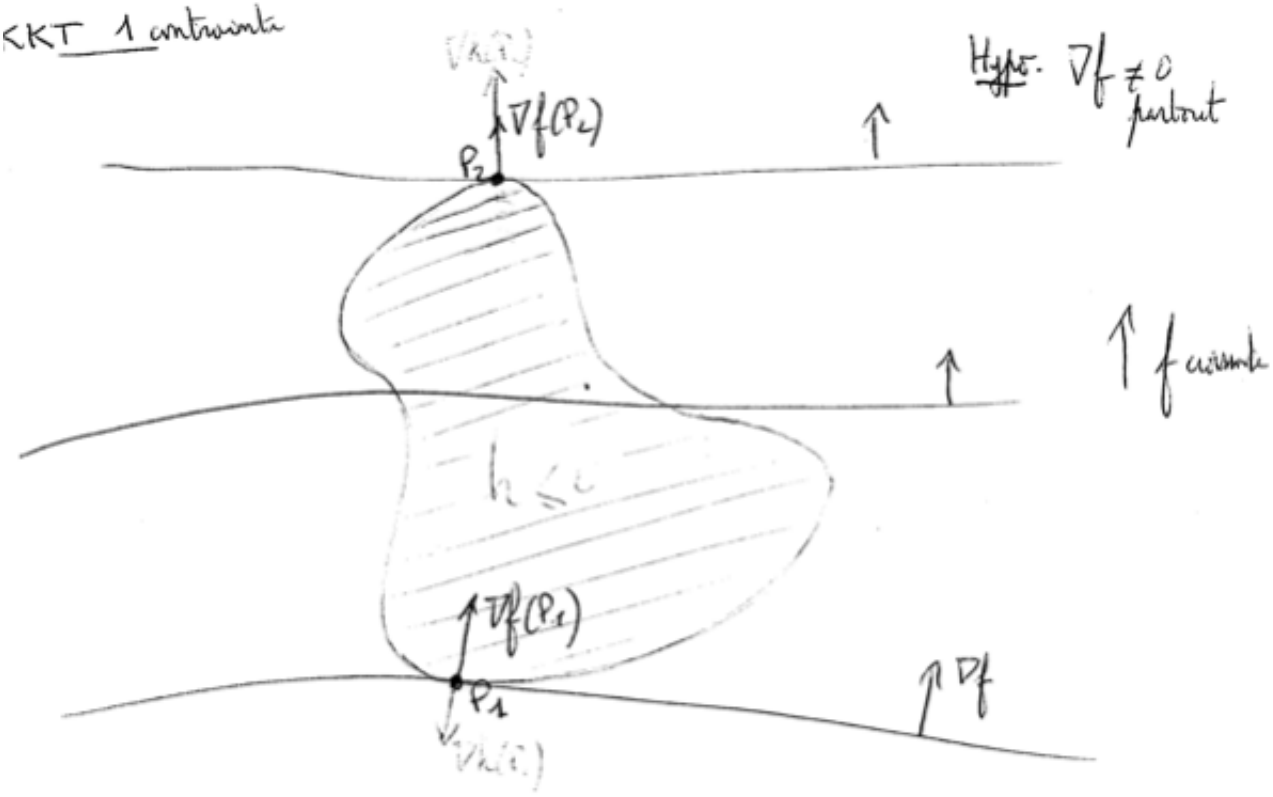
$$\sum_{i=1}^{p} \lambda_i \langle v, \nabla g_j(x^*) \rangle + \sum_{j=1}^{l} \mu_j \langle v, \nabla h_j(x^*) \rangle = 0,$$

donc

$$\sum_{j=1}^{l} \mu_j \langle v, \nabla h_j(x^*) \rangle = 0,$$

comme tous les nombres $\langle v, \nabla h_j(x^*) \rangle$ sont négatifs et les μ_j sont positifs, cela entraîne que les μ_j sont tous nuls. Par ailleurs si les vecteurs $\nabla g_j(x^*)$ sont indépendants, les λ_i sont nuls aussi. □

KKT 1 contrainte



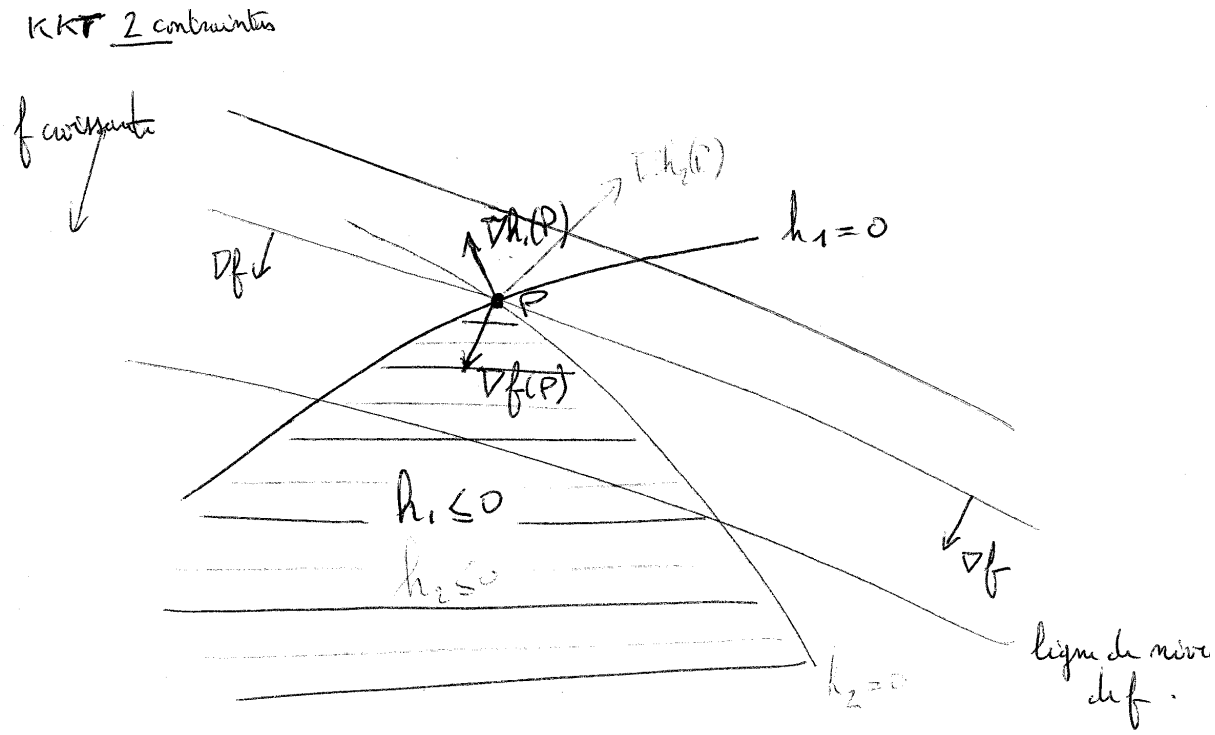
En P_1 f est minimale sous contrainte $h \leq 0$

$\nabla f(P_1)$ et $\nabla h(P_1)$ sont colinéaires de sens opposés.

En P_2 f est maximale sous contrainte $h \leq 0$

$\nabla f(P_2)$ et $\nabla h(P_2)$ sont colinéaires de même sens.

Remarque : Sur cet exemple f atteint ses bornes sur $h \leq 0$ en deux points du bord $h = 0$. (on dit que la contrainte $h \leq 0$ est active en ces points). En général f peut aussi être extrémale en un point de $h < 0$ (en un tel point le gradient de f est nul).



En P est minimum sous contraintes $h_1 \leq 0$, $h_2 \leq 0$

$\nabla f(P)$ est combinaison linéaire de $\nabla h_1(P)$ et $\nabla h_2(P)$ avec des coefficients négatifs.

$$\nabla f(P) + \mu_1 \nabla h_1(P) + \mu_2 \nabla h_2(P) = 0$$

avec $\mu_1, \mu_2 > 0$.

5 Quelques exemples de problèmes liés à l'optimisation

5.1 Les théorèmes de Condorcet et d'Arrow

Le but de ce paragraphe est de mettre en évidence les problèmes que pose l'agrégation de préférences. Que faire quand on a trouvé un optimum pour différents critères ? La réponse peut-être difficile voire inextricable.

Le paradoxe de Condorcet est facile à énoncer. Trois votants disent leurs préférences pour

trois candidats (A, B, C). Ces préférences sont données dans le tableau suivant.

	Votant 1	Votant 2	Votant 3
Premier choix	A	B	C
Deuxième choix	B	C	A
Troisième choix	C	A	B

On voit qu'une majorité préfère A à B, une majorité préfère B à C et... une majorité préfère C à A. Il paraît impossible de définir une préférence globale à partir des préférences individuelles.

Arrow a donné un résultat du même type. Un ensemble d'états sociaux est proposé à un ensemble d'agents sociaux. Comment définir le meilleur état social global à partir des préférences individuelles ? Arrow définit des propriétés qui semblent souhaitables.

* Deux états sociaux sont comparables.
* Si l'état x est préféré à y et y à z alors x est préféré à z.
* Pour tout système de préférences individuelles on peut définir un système de préférences global.
* Si x est préféré à y (globalement) alors x est aussi préféré à y si on modifie les changements individuels en faisant remonter x.
* Savoir si x est préféré à y ou le contraire ne dépend que de x et y.
* Si x est préféré à y par tous les individus, alors x est (globalement) préféré à y.
* Le système global n'est pas imposé : pour toute paire x, y, la préférence sociale de x sur y dépend effectivement des préférences individuelles.
* Le système n'est pas dictatorial : le choix ne coïncide pas avec le choix de l'un des individus.

Théorème 5.1. *Il n'existe pas de sytème de préférences social satisfaisant aux conditions précédentes dès que le nombre d'agents est plus grand que 1 et le nombre d'états sociaux plus grand que 2.*

5.2 Équilibre de Walras, équilibre de Nash

Existence d'un optimum en théorie du consommateur. Voir Malinvaud page 24. On suppose qu'un consommateur peut acheter des biens numérotés de 1 à l. Le bien numéroté h à un prix à l'unité p_h. Fonction d'utilité. Si le consommateur achète les quantités $x_1, \ldots, x_l$ de ces biens (x_i quantité de bien i achetée) alors l'utilité attachée est exprimée au moyen d'une fonction S par $S(x_1, \ldots, x_l)$. L'objectif est de maximaliser S. Mais existent des contraintes : chaque x_i doit être positif ou nul. La somme totale dépensée doit être inférieure au revenu du consommateur : R.

Je n'ai pas eu le temps de donner des détails sur le théorème de Debreu. Je me contente de recopier quelques images d'un article d'Arrow et Debreu qui montrent que les notions introduites dans le cours sont celles qui permettent à ces auteurs d'énoncer leur résultat.

Comme souvent en économie l'unanimité n'est pas de mise : « La construction de Debreu n'a aucune valeur scientifique, tant elle est totalement étrangère au monde de l'expérience. »Maurice Allais

The function $A_\iota(\bar{a}_\iota)$ is said to be *continuous* at $\bar{a}_\iota^0$ if for every $a_\iota^0 \in A_\iota(\bar{a}_\iota^0)$ and every sequence $\{\bar{a}_\iota^n\}$ converging to $\bar{a}_\iota^0$, there is a sequence $\{a_\iota^n\}$ converging to a_ι^0 such that $a_\iota^n \in A_\iota(\bar{a}_\iota^n)$ for all n. Again, if $A_\iota(\bar{a}_\iota)$ were a single-valued function, this definition would coincide with the ordinary definition of continuity.

2.5. LEMMA: *If, for each ι, $\mathfrak{A}_\iota$ is compact and convex, $f_\iota(\bar{a}_\iota, a_\iota)$ is continuous on $\mathfrak{A}$ and quasi-concave[9] in a_ι for every $\bar{a}_\iota$, $A_\iota(\bar{a}_\iota)$ is a continuous function whose graph is a closed set, and, for every $\bar{a}_\iota$, the set $A_\iota(\bar{a}_\iota)$ is convex and non-empty, then the abstract economy $[\mathfrak{A}_1, \cdots, \mathfrak{A}_\nu, f_1, \cdots, f_\nu, A_1(\bar{a}_1), \cdots, A_\nu(\bar{a}_\nu)]$ has an equilibrium point.*

This lemma generalizes Nash's theorem on the existence of equilibrium points for games [14]. It is a special case of the Theorem in [5], when taken in conjunction with the Remark on p. 889.[10]

3.3.0. Unfortunately, the Lemma stated in 2.5 is not directly applicable to E, since the action spaces are not compact.

Let

$$\hat{X}_i = \{x_i \mid x_i \in X_i, \text{ there exist } x_{i'} \in X_{i'} \text{ for each } i' \neq i \text{ and } y_j \in Y_j \text{ for each } j \text{ such that } z \leqq 0\},$$

$$\hat{Y}_j = \{y_j \mid y_j \in Y_j, \text{ there exist } x_i \in X_i \text{ for each } i, y_{j'} \in Y_{j'} \text{ for each } j' \neq j \text{ such that } z \leqq 0\}.$$

$\hat{X}_i$ is the set of consumption vectors available to individual i if he had complete control of the economy but had to take account of resource limitations. $\hat{Y}_j$ has a similar interpretation. We wish to prove that these sets are all bounded. It is clear that an E equilibrium x_i^* must belong to $\hat{X}_i$ and that an E equilibrium y_j^* must belong to $\hat{Y}_j$.

3.3.4. Now introduce a new abstract economy $\bar{E}$, identical with E in 3.1., except that X_i is replaced by $\bar{X}_i$ and Y_j by $\bar{Y}_j$ everywhere. Let $\bar{A}_i(\bar{x}_i)$ be the resultant modification of $A_i(\bar{x}_i)$ (See 3.1.0.). It will now be verified that all the conditions of the Lemma are satisfied for this new abstract economy.

Je vais m'arrêter plus longuement sur la notion d'équilibre de Nash pour les jeux concaves (je reprends une démonstration de Geanakoplos).

Commençons par un théorème que nous ne démontrerons pas.

Théorème 5.2. *(Brouwer) Toute application continue d'une partie convexe compacte de $\mathbb{R}^D$ dans elle-même admet un point fixe.*

Un jeu est défini par la donnée de N joueurs à qui on associe N ensembles de stratégies possibles $\Sigma_1, \ldots, \Sigma_N$ qui sont des parties compactes convexes d'espaces $\mathbb{R}^d$ et de fonctions de gains $u_1, \ldots, u_N$ définie respectivement sur Σ le produit $\Sigma \times \ldots \times \Sigma_N$ à valeurs dans $\mathbb{R}$. Le jeu est dit concave si, pour tout n, la fonctions u_n est concave en la $n$ème variable (quelles que soient la façons dont on fixe les autres).

Un équilibre de Nash pour un tel jeu est un élément x^* de Σ tel que, pour tout n, on ait :

$$u_n(x^*) \geq u_n(x_1^*, \ldots, x_{n-1}^*, x_n, x_{n+1}^*, \ldots, x_N^*).$$

Théorème 5.3. *Tout jeu concave admet un équilibre de Nash.*

Démonstration Fixons un x dans Σ et considérons la fonction

$$\psi_{n,x} \ : \ y_n \mapsto u_n(x_1, \ldots, x_{n-1}, y_n, x_{n+1}, \ldots, x_N) - \|y_n - x_n\|^2.$$

C'est une fonction continue strictement concave définie sur le compact convexe Σ_n. Comme elle est continue, elle atteint donc son maximum en un point de Σ, et comme elle est strictement concave ce point est unique. Notons le $\phi_n(x)$. Montrons que cette application ϕ_n définie sur Σ à valeurs dans Σ_n est elle-même continue. Soit $(x_k)_k$ une suite de points de Σ convergeant vers x_∞. Comme les fonctions u_n sont continues, pour tout y_n dans Σ_n, $\psi_{n,x_k}(y_n)$ tend vers $\psi_{n,x_\infty}(y_n)$ et le maximum de ψ_{n,x_k} converge vers celui de ψ_{n,x_∞}. Considérons alors la suite $(\phi_n(x_k))_k$. On veut montrer qu'elle converge vers $\phi_n(x_\infty)$. Supposons que ce ne soit pas le cas, alors on peut trouver une sous-suite $\big(\phi_n(x_{k_j})\big)$ convergeant vers un autre point z_∞ de Σ_n (car Σ_n est compact). Mais alors ψ_{n,x_∞} atteint son maximum en deux point distincts : $\phi_n(x_\infty)$ et z_∞. Contradiction, car ψ_{n,x_∞} est strictement concave.

On considère alors la fonction ϕ de Σ dans Σ, définie par :

$$\phi(x) = (\phi_1(x), \ldots, \phi_N(x)).$$

C'est une application continue du convexe compact Σ dans lui-même. D'après le théorème de Brouwer, elle a donc un point fixe (au moins) x^*. Nous allons maintenant voir que x^* est un équilibre de Nash. Supposons que ce ne soit pas le cas. Alors pour un certain n on peut trouver x_n dans Σ_n tel que

$$u_n(x^*) < u_n(x_1^*, \ldots, x_{n-1}^*, x_n, x_{n+1}^*, \ldots, x_N^*).$$

Appelons δ la différence

$$\delta = u_n(x_1^*, \ldots, x_{n-1}^*, x_n, x_{n+1}^*, \ldots, x_N^*) - u_n(x^*).$$

Comme u_n est strictement concave en la $n$ème variable, on a, pour $t \in]0, 1[$,

$$\begin{aligned}
&u_n(x_1^*, \ldots, x_{n-1}^*, tx_n + (1-t)x_n^*, x_{n+1}^*, \ldots, x_N^*) \\
&\quad > tu_n(x_1^*, \ldots, x_{n-1}^*, x_n, x_{n+1}^*, \ldots, x_N^*) + (1-t)u_n(x^*) \\
&\quad > u_n(x^*) + t\left(u_n(x_1^*, \ldots, x_{n-1}^*, x_n, x_{n+1}^*, \ldots, x_N^*) - u_n(x^*)\right) \\
&\quad = u_n(x^*) + t\delta.
\end{aligned}$$

Par ailleurs, on a

$$\begin{aligned}
&\psi_{n,x^*}(tx_n + (1-t)x_n^*) \\
&\quad = u_n(x_1, \dots, x_{n-1}, tx_n + (1-t)x_n^*, x_{n+1}, \dots, x_N) - \|tx_n + (1-t)x_n^* - x_n^*\|^2 \\
&\quad > u_n(x^*) + t\delta - \|t(x_n - x_n^*)\|^2 \\
&\quad > u_n(x^*) + t\delta - t^2\|x_n - x_n^*\|^2.
\end{aligned}$$

Pour t suffisamment petit, cette dernière quantité est strictement supérieure à $u_n(x^*)$, ce qui est contraire à la définition de ϕ_n. En effet par définition $x_n^* = \phi_n(x^*)$ signifie que ψ_{n,x^*} est maximale en x_n^*. □

5.3 Régression linéaire

Sur une collection d'individus on observe des variables y_i et des variables explicatives (ou régresseurs) x_i, $i = 1, ...n$, chaque paire (y_i, x_i) représentant une expérience. On les arrange dans un tableau de la façon suivante :

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} 1 & x_{12} & \dots & x_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n2} & \dots & x_{np} \end{pmatrix}.$$

x_i est donc un vecteur ligne. On notera $x_{.j}$ la j-ème colonne. On convient généralement que le premier régresseur est la constante, mais ça n'est pas obligatoire. Pour tout vecteur β, on définit l'erreur quadratique moyenne d'ajustement (ou erreur résiduelle) $S(\beta)$ comme

$$S(\beta)^2 = \frac{1}{n}\|y - X\beta\|^2 = \frac{1}{n}\sum_i (y_i - x_i\beta)^2.$$

Le vecteur des coefficients de regression calculé aux moindres carrés ordinaires est

$$\hat{\beta} = \arg\min_{\beta} S(\beta)$$

Proposition 5.4. *Si la matrice ${}^tXX$ est inversible, alors $\hat{\beta}$ est donné par*

$$\hat{\beta} = ({}^tXX)^{-1}.{}^tXy.$$

D'abord nous allons montrer que S a un point critique en $\hat{\beta}$ puis qu'elle a un minimum global en $\hat{\beta}$.

Calculons les dérivées partielles de S :

$$\begin{aligned}
\frac{\partial S}{\partial \beta_j} &= \frac{1}{n}\frac{\partial}{\partial \beta_j}\|y - X\beta\|^2 \\
&= \frac{1}{n}\frac{\partial}{\partial \beta_j}\sum_i (y_i - x_i\beta)^2 \\
&= \frac{1}{n}\frac{\partial}{\partial \beta_j}\sum_i (y_i - \sum_k x_{ik}\beta_k)^2 \\
&= \frac{1}{n}\sum_i \frac{\partial}{\partial \beta_j}(y_i - \sum_k x_{ik}\beta_k)^2 \\
&= \frac{1}{n}\sum_i (-2x_{ij})(y_i - \sum_k x_{ik}\beta_k) \\
&= \frac{-2}{n}\sum_i y_i x_{ij} + \frac{2}{n}\sum_i\sum_k x_{ij}x_{ik}\beta_k \\
&= \frac{-2}{n}({}^tXy)_j + \frac{2}{n}\sum_k (\sum_i x_{ij}x_{ik})\beta_k \\
&= \frac{-2}{n}({}^tXy)_j + \frac{2}{n}\sum_k ({}^tXX)_{jk}\beta_k \\
&= \frac{2}{n}(({}^tXX\beta)_j - ({}^tXy)_j) \\
&= \frac{2}{n}(({}^tXX\beta - {}^tXy)_j).
\end{aligned}$$

Autrement dit

$$\nabla S = \frac{2}{n}({}^tXX\beta - {}^tXy).$$

Le gradient s'annule lorsque ${}^tXX\beta - {}^tXy = 0$ c'est-à-dire $\beta = ({}^tXX)^{-1}({}^tXy)$.

Pour montrer que S a un minimum en son point critique on peut montrer que S est coercive (comme il y a une base de $\mathbb{R}^p$ formée de vecteurs ligne de X, si la norme de β est très grande au moins un produit de l'un de ces vecteurs ligne par β est très grand ce qui entraîne que $S(\beta)$ est très grand).

On peut aussi calculer la hessienne de S.

$$\begin{aligned}
\frac{\partial^2 S}{\partial \beta_l \partial \beta_j} &= \frac{-2}{n}\sum_i x_{ij}\frac{\partial}{\partial \beta_l}(y_i - \sum_k x_{ik}\beta_k) \\
&= \frac{2}{n}\sum_i x_{ij}x_{il} \\
&= \frac{2}{n}({}^tXX)_{jl}.
\end{aligned}$$

Cela signifie que la hessienne de S est $2\ {}^tXX/n$, matrice définie positive (pourquoi cette matrice est-elle définie positive ?). La fonction S est donc strictement convexe ; elle a un minimum en son unique point critique.

Autre argument pour montrer que le minimum est bien atteint en $({}^tXX)^{-1}({}^tXy)$. On cherche à minimiser $\|y - X\beta\|^2$. Ecrivons :

$$\|y-X\beta\|^2 = \|(y-X\hat\beta)+X(\beta-\hat\beta)\|^2 = \|y-X\hat\beta\|^2+\|X(\beta-\hat\beta)\|^2+2\langle y-X\hat\beta, X(\beta-\hat\beta\rangle.$$

Nous allons montré dans un instant que le produit scalaire $\langle y - X\hat\beta, X(\beta - \hat\beta)\rangle$ est nul. On a donc

$$\|y - X\beta\|^2 = \|(y - X\hat\beta) + X(\beta - \hat\beta)\|^2 = \|y - X\hat\beta\|^2 + \|X(\beta - \hat\beta)\|^2$$

qui est minimale quand $\|X(\beta - \hat\beta)\|^2 = 0$ (une quantité positive ou nulle est minimale lorsqu'elle est nulle). Or $\|X(\beta-\hat\beta)\|^2 = 0$ si et seulement si $X(\beta-\hat\beta)$ ou encore $\beta-\hat\beta = 0$ (pourquoi ?) soit $\beta = \hat\beta$.

Calculons donc le produit scalaire $\langle y - X\hat\beta, X(\beta - \hat\beta)\rangle$.

$$\begin{aligned}\langle y - X\hat\beta, X(\beta - \hat\beta)\rangle &= \langle y, X(\beta - \hat\beta)\rangle - \langle X\hat\beta, X(\beta - \hat\beta)\rangle \\ &= \langle {}^tXy, \beta - \hat\beta\rangle - \langle {}^tXX\hat\beta, \beta - \hat\beta\rangle \\ &= \langle {}^tXy - {}^tXX\hat\beta, \beta - \hat\beta\rangle\end{aligned}$$

Or

$${}^tXX\hat\beta = {}^tXX({}^tXX)^{-1}.{}^tXy = {}^tXy.$$

On a donc ${}^tXy - {}^tXX\hat\beta = 0$ et le produit scalaire est $\langle 0, \beta - \hat\beta\rangle = 0$.

Autre façon de faire le calcul.

$$\begin{aligned}S(\beta) &= \frac{1}{n}\|y - X\beta\|^2 \\ &= \frac{1}{n}\langle y - X\beta, y - X\beta\rangle \\ &= \frac{1}{n}\left(\|y\|^2 - 2\langle y, X\beta\rangle + \langle X\beta, X\beta\rangle\right) \\ &= \frac{1}{n}\left(\|y\|^2 - 2\langle {}^tXy, \beta\rangle + \langle {}^tXX\beta, \beta\rangle\right)\end{aligned}$$

On en déduite que le gradient de S en β est $\frac{2}{n}\left({}^tXX\beta - {}^tXy\right)$ et sa hessienne $2\ {}^tXX/n$. Le gradient est nul lorsque

$${}^tXX\beta - {}^tXy = 0, \text{ soit } \beta = ({}^tXX)^{-1}({}^tXy),$$

et comme la hessienne est constante définie positive S est strictement convexe. Elle a donc un minimum global en son unique point critique.

5.4 Analyse en composantes principales

On se donne n vecteurs $x_1, \ldots, x_n$ dans $\mathbb{R}^d$. Ces vecteurs représentent des données recueillies sur n individus (d nombres par individus). Ces n vecteurs définissent un nuage de

points dans $\mathbb{R}^d$. On veut projeter ce nuage de points sur une droite ou un plan en perdant le moins d'information possible. Nous supposons que les données sont centrées c'est-à-dire que la moyenne des vecteurs x_1 est le vecteur nul.

Commençons par chercher une droite Δ telle que la somme des distances de chaque x_i à Δ au carré soit minimale. Appelons p_Δ la projection sur Δ. On cherche à minimiser

$$\sum_{i=1}^{n} \|x_i - p_\Delta(x_i)\|^2.$$

Le théorème de Pythagore donne

$$\|x_i\|^2 = \|x_i - p_\Delta(x_i)\|^2 + \|p_\Delta(x_i)\|^2.$$

On a donc

$$\sum_{i=1}^{n} \|x_i - p_\Delta(x_i)\|^2 = \sum_{i=1}^{n} \|x_i\|^2 - \sum_{i=1}^{n} \|p_\Delta(x_i)\|^2,$$

et, comme $\sum_{i=1}^n \|x_i\|^2$ ne dépend pas de Δ, minimiser $\sum_{i=1}^n \|x_i - p_\Delta(x_i)\|^2$ revient à maximiser $\sum_{i=1}^n \|p_\Delta(x_i)\|^2$. Ecrivons qu'une droite est déterminée par un vecteur directeur unitaire v : $\Delta = \mathbb{R}v$. Cela permet de donner une expression analytique à la quantité que nous cherchons à maximiser car $p_\Delta(x_i) = \langle x_i, v\rangle v$. On cherche v de norme 1 tel que

$$\phi(v) = \sum_{i=1}^{n} \langle x_i, v\rangle^2.$$

C'est un problème d'optimisation sous contrainte ($\|v\|^2 = 1$). La fonction ϕ est un forme quadratique. Pour obtenir la matrice symétrique (positive par définition de ϕ (somme de carrés)) associée à ϕ, il suffit d'écrire le produits scalaire sous forme de produit de matrices :

$$\langle x_i, v\rangle^2 = {}^t v x_i {}^t x_i v.$$

Cela donne

$$\phi(v) = \sum_{i=1}^{n} \langle x_i, v\rangle^2 = \sum_{i=1}^{n} {}^t v x_i {}^t x_i v = {}^t v \left(\sum_{i=1}^{n} x_i {}^t x_i \right) v,$$

soit, en posant $A = \sum_{i=1}^n x_i {}^t x_i$:

$$\phi(v) = \langle v, Av\rangle.$$

Remarque : si nous appelons X la matrice $d \times n$ dont les vecteurs colonnes sont les x_i. On peut vérifier que l'on a

$$A = X.{}^t X,$$

qui est une matrice définie positive si X est de rang d (ce qui nécéssite $n \geq d$).

Les gradients de ϕ et de la contrainte au point v sont

$$\nabla\phi(v) = 2Av, \ \nabla\|v\|^2 = 2v.$$

Le gradient de la contrainte ne s'annule pas sur l'ensemble $\|v\|^2 = 1$, on peut donc appliquer le théorème des extrema liés : en un point où ϕ est maximale sous la contrainte, il existe λ tel que

$$\nabla\phi(v) = \lambda\nabla\|v\|^2 \text{ soit } Av = \lambda v.$$

(Remarque : grâce au calcul précédent on peut démontrer que les matrices réelles symétriques sont diagonalisables. Résultat qui apparaît alors comme une conséquence du théorème des extrema liés)

Ce que nous avons obtenu est que ϕ ne peut être maximale qu'en un vecteur propre de A. La valeur de ϕ en un vecteur propre de A unitaire est la valeur propre associée. On en déduit que $\phi(v)$ est maximale si v est un vecteur propre de A associé à la plus grande valeur propre de A.

Nous pouvons alors énoncer la proposition suivante.

Proposition 5.5. *Les droites qui approchent le mieux le nuage de points (au sens des moindres carrés pour une projection orthogonale) sont les droites propres associées à la plus grande valeur propre de la matrice $A = X.{}^tX$. Si la dimension de l'espace propre associé à la plus grande valeur propre est 1 alors une seule droite donne la meilleure approximation.*

Intéressons-nous maintenant à la meilleure approximation par un plan. Nous allons montrer la proposition suivante.

Proposition 5.6. *Les plans qui approchent le mieux le nuage de points (au sens des moindres carrés pour une projection orthogonale) sont les plans obtenus comme somme de deux droites propres associées aux deux plus grandes valeurs propres de la matrice $A = X.{}^tX$. Si la multiplicité de la plus grande valeur propre est au moins 2 alors on prend deux droites propres associées à cette plus grande valeur propre. Si la somme des dimensions des espaces propres associés aux deux plus grandes valeurs propre est 2 alors un seul plan donne la meilleure approximation.*

On cherche deux vecteurs unitaires v_1 v_2 orthogonaux tels que la somme

$$\sum_{i=1}^{n} \|x_i - p(x_i)\|^2,$$

où p est la projection orthogonale sur $Vect(v_1, v_2)$. On montre que

$$p(x_i) = \langle x_i, v_1\rangle v_1 + \langle x_i, v_2\rangle v_2.$$

Là encore le théorème de Pythagore donne

$$\|x_i\|^2 = \|x_i - p(x_i)\|^2 + \|p(x_i)\|^2,$$

et minimiser $\sum_{i=1}^n \|x_i - p(x_i)\|^2$ revient à maximiser $\sum_{i=1}^n \|p(x_i)\|^2$. La dépendance en v_1 et v_2 est cachée dans le p :

$$\psi(v_1, v_2) = \sum_{i=1}^n \|p(x_i)\|^2 = \sum_{i=1}^n \|\langle x_i, v_1\rangle v_1 + \langle x_i, v_2\rangle v_2\|^2 = \sum_{i=1}^n \langle x_i, v_1\rangle^2 + \sum_{i=1}^n \langle x_i, v_2\rangle^2.$$

On cherche donc à maximiser $\psi(v_1, v_2) = \phi(v_1) + \phi(v_2)$ sous la contrainte v_1 et v_2 sont orthogonaux. En prenant pour v_1 et v_2 deux vecteurs propres de A unitaires associés aux deux plus grandes valeurs propres de A on obtient la somme de ces deux plus grandes valeurs propres pour $\psi(v_1, v_2)$. On ne peut pas faire mieux. Comment le voir ?

Numérotons les valeurs propres de A par valeurs descendantes : $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d$ et fixons une famille $u_1, \ldots, u_d$ de vecteurs propres unitaires associés à ces valeurs propres ($Au_i = \lambda_i u_i$). Le maximum de $\phi(v) = \langle v, Av\rangle$ sur l'hyperplan $u_1^\perp$ est λ_2. Considérons alors deux vecteurs v_1 et v_2 unitaires et orthogonaux. Le plan $Vect(v_1, v_2)$ et l'hyperplan $u_1^\perp$ s'intersectent au moins le long d'une droite. Prenons une base orthonormée w_1, w_2 de $Vect(v_1, v_2)$ telle que w_1 appartiennent à $u_1^\perp$. On a alors $\phi(w_1) \leq \lambda_2$ et

$$\psi(v_1, v_2) = \psi(w_1, w_2) = \phi(w_1) + \phi(w_2) \leq \lambda_2 + \phi(w_2) \leq \lambda_2 + \lambda_1,$$

car λ_1 est le maximum de ϕ sur la sphère unité.

On montre de manière analogue que, pour tout $k = 1, \ldots, d-1$, le sous-espace vectoriel de dimension k qui approche le mieux le nuage de points est $Vect(u_1, \ldots, u_k)$. Les droites $\mathbb{R}u_i$ s'appellent les composantes principales ou les axes principaux.

5.5 Réseaux de neurones

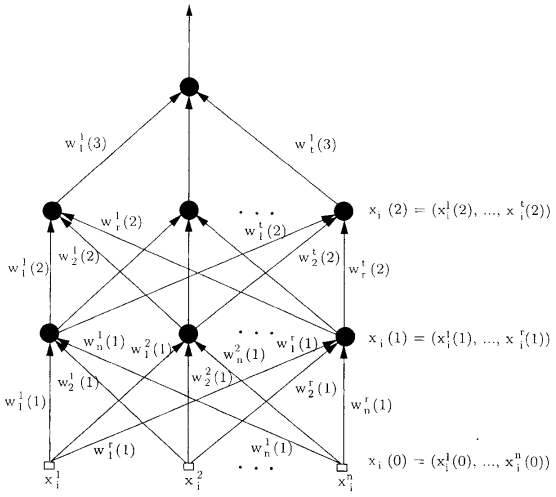
Un réseau de neurones à $m+1$ couches est défini de la façon suivante.

* La couche 0 décrit le vecteur d'entrée $x(0)$ à n coordonnées ;
* La couche numéro k est la donnée d'un vecteur à n_k coordonnées. Pour $k < m$, n_k peut être n'importe quel entier.
* La m-ème couche donne un nombre ($n_m = 1$).
* On calcule les coordonnées de la $k+1$-ème couche à partir de celle de la k-ème couche. On se donne une matrice M_{k+1} de taille $n_{k+1} \times n_k$. On calcule le vecteur image de la k-ème couche (considéré comme un vecteur colonne) en le multipliant par M_{k+1} :

  $$u(k+1) = M_{k+1} x(k),$$

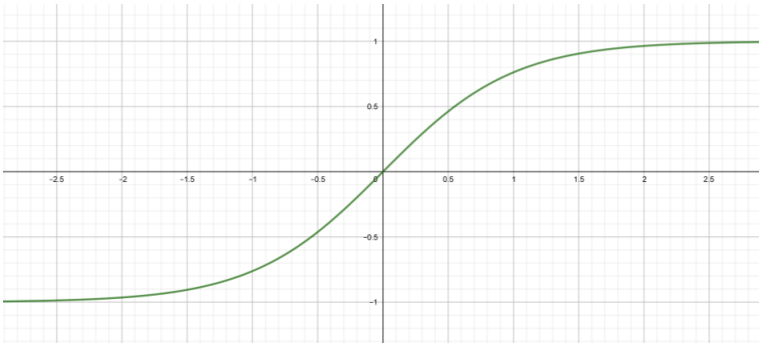
  puis on applique une fonction sigmoïde S à chaque coordonnée de $u(k+1)$ pour obtenir $x(k+1)$:

  $$x(k+1)_j = S(u(k+1)_j).$$



Une fonction sigmoïde est une fonction régulière croissante tendant vers 1 en $+\infty$ et vers -1 en $-\infty$. Par exemple on peut prendre

$$S(t) = \tanh(t).$$



Un réseau de neurones est une machine qui associe un nombre $x(m)$ à un vecteur d'entrée $x(0)$ à n coordonnées.

Imaginons que l'on puisse décrire la réaction souhaitée à une situation décrite par n paramètres au moyen d'un nombre. Dans le cas d'une réponse "oui" ou "non" ce nombre peut être 0 (pour "oui" par exemple) ou 1 (pour "non"). On veut apprendre à un réseau de neurone à répondre correctement. On donne des couples $(x^{(i)}(0), y^{(i)})$ où $y^{(i)}$ (i allant de 1 à l où l est le nombre d'expériences faites) est la bonne réponse à apporter à la situation décrite par le vecteur $x^{(i)}(0)$. On cherche à fixer les coefficients des matrices M_k pour que la différence entre la réponse que donne le réseau de neurone et la bonne réponse soit la plus petite possible. Autrement dit on cherche à résoudre le problème d'optimisation sous contrainte

$$\min \sum_{i=1}^{l} (y^{(i)} - x^{(i)}(m))^2$$

sous contraintes

$$u^{(i)}(k+1) = M_{k+1}x^{(i)}(k),\ x^{(i)}(k+1)_j = S(u^{(i)}(k+1)_j), \forall i = 1, \dots, l,\ \forall k = 0, \dots, m-1.$$

On peut appliquer le théorème des extrema liés. On ne peut pas résoudre explicitement le système obtenu. On applique une méthode de descente de gradient (stochastique) pour obtenir une valeur qu'on pense raisonnable. Pour plus de détails on pourra consulter [18]. Un point très important ici : les dimensions dans lesquels se font les calculs sont très grandes. En reconnaissance d'image par exemple n est de l'ordre du million.

6 Généralités sur les fonctions de plusieurs variables

6.1 Topologie de $\mathbb{R}^d$

Définition 6.1. *Si $x = (x_1 \dots x_d)$ et $y = (y_1 \dots y_d)$ sont deux vecteurs de $\mathbb{R}^d$, on définit leur* **produit scalaire** *par :*

$$\langle x, y\rangle = x_1y_1 + \cdots + x_dy_d$$

Définition 6.2. *On appelle* **norme** *de x (ou longueur) $\|x\| = \langle x, x\rangle^{1/2}$ et la* **distance** *entre deux vecteurs $d(x\,,\,y) = \|x - y\|$.*

Théorème 6.3. *Le produit scalaire vérifie l'***inégalité de Cauchy-Schwarz** *$\langle x, y\rangle^2 \leqslant \|x\|^2\,\|y\|^2$ avec égalité si et seulement si x et y sont colinéaires.*

Théorème 6.4. *La norme définie précédemment s'appelle* **norme euclidienne** *et vérifie :*

(i) $\|x\| = 0$ *si et seulement si* $x = 0$

(ii) $\|x\| > 0$ *si* $x \neq 0$

(iii) $\|\alpha\, x\| = |\alpha|\,\|x\|$

(iv) $\|x + y\| \leqslant \|x\| + \|y\|$

L'inégalité de Cauchy Schwarz permet aussi de définir l'angle géométrique entre deux vecteurs : comme $\langle x, y\rangle \leq \|x\|\|y\|$, si aucun des deux vecteurs n'est nul, alors le quotient $\dfrac{\langle x, y\rangle}{\|x\|\,\|y\|}$ est un nombre compris entre -1 et 1.

Définition 6.5. *L'***angle** *entre deux vecteurs non nuls est $\theta \in [0\,,\,\pi]$ vérifiant $\cos\theta = \dfrac{\langle x, y\rangle}{\|x\|\,\|y\|}$.*

Le coefficient de corrélation en statistique à deux variables est le cosinus de l'angle formé par les deux vecteurs de $\mathbb{R}^n$ définis par les deux vecteurs de données (vecteur des $x_i - \bar{x}$ et vecteur des $y_i - \bar{y}$).

Définition 6.6. *x et y de $\mathbb{R}^d$ sont dits* **orthogonaux** *lorsque $\langle x, y\rangle = 0$.*

Définition 6.7. *(plan dans $\mathbb{R}^3$)*
Soient $A = (x_0\,,\, y_0\,,\, z_0)$ un point de $\mathbb{R}^3$ et $N = (a\,,\, b\,,\, c)$ un vecteur non nul.
Le plan passant par A et orthogonal à N est $P = \{x \in \mathbb{R}^3 / (x - A) \cdot N = 0\}$.

Définition 6.8. *Soient $a \in \mathbb{R}^d$ et $r > 0$.*
On appelle $B(a\,,\, r) = \{x \in \mathbb{R}^d \,/\, \|x - a\| < r\}$ la **boule ouverte** *de centre a et de rayon r.*

Exemple
Dans $\mathbb{R}$, $\mathbb{R}^2$ ou $\mathbb{R}^3$ on retrouve les intervalles, les disques, les boules ouvertes.

Proposition 6.9. *Soient $A \subset \mathbb{R}^d$, $a \in \mathbb{R}^d$.*
Alors une des trois conditions suivantes est vérifiée :

(i) $\exists r > 0$ *tel que* $B(a\,,\, r) \subset A$

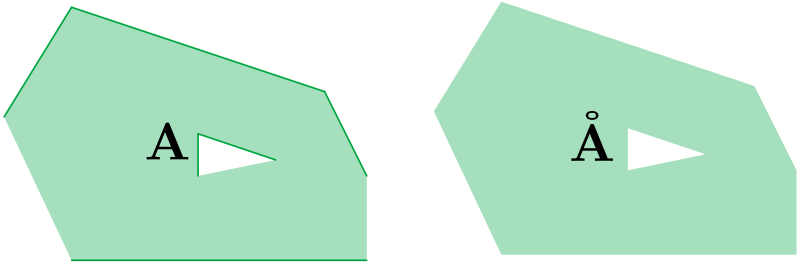
(ii) $\exists r > 0$ *tel que* $B(a\,,\, r) \subset A^c$ *où* $A^c = \mathbb{R}^d \setminus A$

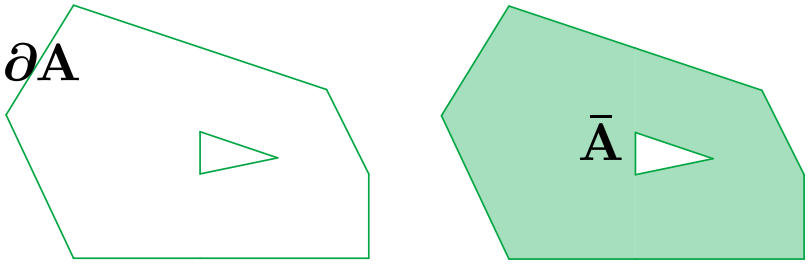
(iii) $\forall r > 0\,,\, B(a\,,\, r)$ *contient des points de A et de A^c.*

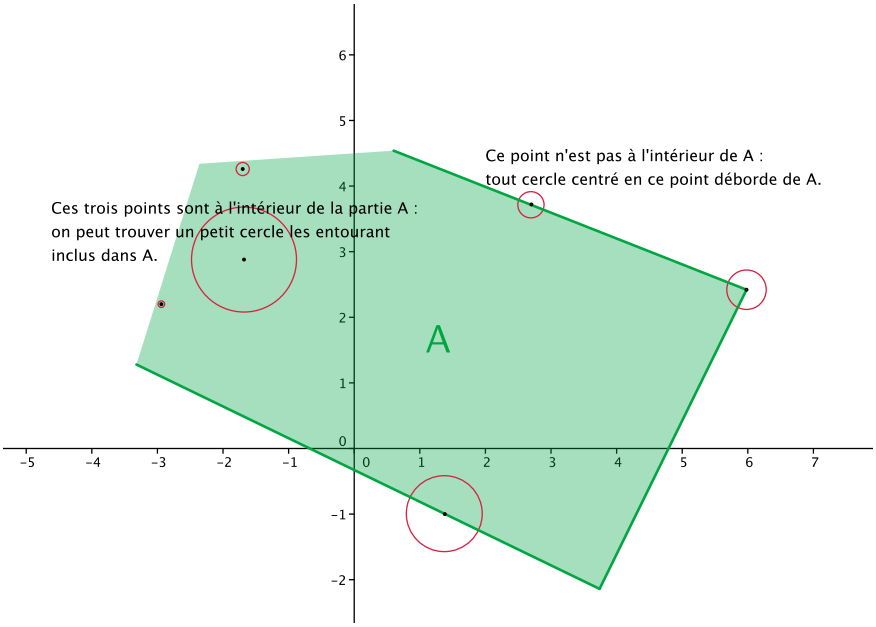
Définition 6.10. *L'***intérieur** *de A (noté $int(A)$ ou $\mathring{A}$) est l'ensemble des points de $\mathbb{R}^d$ vérifiant (i).*
*L'***extérieur** *de A (noté ext A) est l'ensemble des points de $\mathbb{R}^d$ vérifiant la condition (ii).*
La **frontière** *de A (notée ∂A) est l'ensemble des points de $\mathbb{R}^d$ vérifiant la condition (iii).*
La **fermeture** *de A (notée $\overline{A}$) est la réunion de A et de ∂A.*

Exemples dans $\mathbb{R}^2$
$A = \{x \in \mathbb{R}^2 \,/\, \|x\| < 1\}$
$A = \{(n\,,\, 0) \,/\, n \in \mathbb{Z}\}$

Les quatre dessins qui suivent représentent une partie A, son intérieur, sa frontière, son adhérence. à chaque fois la partie A est coloriée en vert. Pour indiquer si les points du bord font ou non partie de l'ensemble considéré, le bord est surligné en vert plus foncé ou non.







Définition 6.11. *Un ensemble A de $\mathbb{R}^d$ est :*

(i) **ouvert** *si* $\forall a \in A\,,\ \exists r > 0$ *tel que* $B(a\,,\,r) \subset A$

(ii) **fermé** *si* A^c *est ouvert.*

Proposition 6.12. *A est ouvert si et seulement si* $\overset{\circ}{A} = A$.
A est fermé si et seulement si $\overline{A} = A$.

Exemples

$A_1 = \{(x\,,\,y)\,/\,x^2 + y^2 < 1\}$ est ouvert.
$A_2 = \{(x\,,\,y)\,/\,x^2 + y^2 \leqslant 1\}$ est fermé.
$A_3 = A_1 \cup \{(1\,,\,0)\}$ n'est ni ouvert ni fermé.
$]0\,,\,1[\subset \mathbb{R}$ est ouvert dans $\mathbb{R}$.
$]0\,,\,1[\times\{0\} \subset \mathbb{R}^2$ n'est ni ouvert ni fermé.
$[0\,,\,1] \subset \mathbb{R}$ est fermé dans $\mathbb{R}$.
$[0\,,\,1] \times \{0\} \subset \mathbb{R}^2$ est fermé dans $\mathbb{R}^2$.

Proposition 6.13. *1. $\mathbb{R}^d$ et $\varnothing$ sont ouverts (et donc aussi fermés).*

2. *Toute réunion d'ouverts est un ouvert.*
3. *Toute intersection finie d'ouverts est un ouvert.*

Définition 6.14. *Une suite dans $\mathbb{R}^d$ est une famille de vecteurs $x_k = (x_1^{(k)}, \ldots, x_d^{(k)})$ indexée par l'ensemble des entiers naturels $(x_k)_{k\in\mathbb{N}}$. Chaque terme de la suite x_k est un vecteur avec ses n coordonnées.*

Définition 6.15. *Une suite $(x_k)_{k\in\mathbb{N}}$* **converge** *dans $\mathbb{R}^d$ vers $b \in \mathbb{R}^d$ si $\forall \varepsilon > 0\,, \exists N \in \mathbb{N}$ tel que $k \geqslant N$ entraîne $\|x_k - b\| < \varepsilon$.*

De manière équivalente on peut définir la convergence d'une suite de vecteurs (x_k) par la convergence de chacune des suites réelles données par les coordonnées $x_i^{(k)}$, i allant de 1 à n, k variant dans $\mathbb{N}$ (les suites des coordonnées sont indexées par k et il y en a n : $(x_i^{(k)})_{k\in\mathbb{N}}$).

Une autre façon de dire que la suite (x_k) tend vers b est de dire que la suite réelle de nombre positifs ou nuls $(d(x_k, b))_{k\in\mathbb{N}}$ tend vers 0.

Remarques

1. On dit que b est **la limite** de la suite (x_k) et on note $x_k \to b$.
2. $x_k \to b$ si et seulement si $\forall \varepsilon > 0$ la boule $B(b\,, \varepsilon)$ contient toute la suite sauf un nombre fini de x_k.

Proposition 6.16. *A est fermé si et seulement si pour toute suite convergente contenue dans A et convergente, la limite est dans A.*

Cette proposition fournit un critère pour démontrer qu'un ensemble A n'est pas fermé : il suffit de trouver une suite de points de A convergeant vers un point n'appartenant pas à A.

Théorème 6.17. *Soit (x_k) une suite bornée. Il existe une sous-suite de (x_k) convergeant dans $\mathbb{R}^d$.*

Définition 6.18. *$X \subset \mathbb{R}^d$ est compact si X est fermé et borné (borné veut dire qu'il existe $R > 0$ tel que $X \subset B(0\,, R)$).*

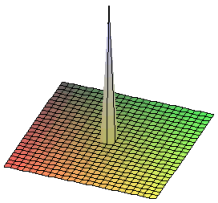
Exemples
$[0, 23]$ est un compact dans $\mathbb{R}$.
$\{(x, y)\ /x^2 + (y-2)^2 \leq 6\}$ est un compact de $\mathbb{R}^2$.
$[2, 3] \times [1, 3] \times [5, 7]$ est un compact dans $\mathbb{R}^3$.

Théorème 6.19. *(Bolzano-Weierstrass)*
Soit $X \subset \mathbb{R}^d$ compact.
Alors toute suite $(x_k) \subset X$ contient une sous-suite (x_{l_k}) qui converge vers un point de X.

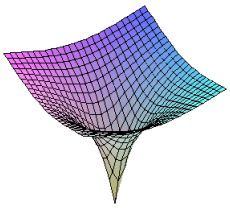
Définition 6.20. *Une fonction f définie sur un sous-ensemble D de $\mathbb{R}^d$ à valeurs dans $\mathbb{R}$ s'appelle fonction numérique de n variables.*
D est le domaine de définition de f.
$\{f(x) \,/\, x \in D\}$ est l'image de f.
$\{(x\,,\, f(x)) \,/\, x \in D\} \subseteq \mathbb{R}^d \times \mathbb{R}$ est appelé graphe de f.

Exemples [3]

$$f(x\,,\, y) = \frac{1}{\sqrt{x^2 + y^2}}$$



$$f(x\,,\, y\,,\, z) = Ln(1 + x^2 + y^2)$$



Définition 6.21. *Soient D et E deux parties de $\mathbb{R}^d$ telles que $D \subset E$ et f et g deux fonctions définies respectivement sur D et E. On dit que g est un prolongement de f à E si pour tout $x \in D$ on a $f(x) = g(x)$. Dans cette situation, on dit aussi que f est la restriction de g à D.*

Définition 6.22. *Une fonction $f : D \to \mathbb{R}$ (où $D \subset \mathbb{R}^d$) a pour limite b en x_0 si $x_0 \in \overline{D}$ et si $\forall \varepsilon > 0\,,\, \exists \delta > 0$ tel que :*
$x \in D \;\;,\;\; \|x - x_0\| < \delta \Rightarrow |f(x) - b| < \varepsilon.$

Notation
Dans ce cas $b = \lim\limits_{x \to x_0} f(x)$.

3. Les images données sont obtenues avec le logiciel Maple.

Définition 6.23. *(i)* $f : D \to \mathbb{R}$ *est* **continue en** $x_0 \in D$ *si et seulement si* $\lim_{x \to x_0} f(x) = f(x_0)$.

(ii) f *est* **continue sur** D *si et seulement si elle est continue en tout point de* D.

Théorème 6.24. *Soit* $f : \mathbb{R}^d \to \mathbb{R}$. *Les propositions suivantes sont équivalentes :*

(i) f *est continue sur* $\mathbb{R}^d$.

(ii) $\forall b \in \mathbb{R}^d$, $\forall x_p$ *avec* $x_p \to b$ *on a :* $f(x_p) \to f(b)$ *dans* $\mathbb{R}$.

(iii) $\forall \theta$ *ouvert de* $\mathbb{R}$, $f^{-1}(\theta) = \{x \in \mathbb{R}^d \,/\, f(x) \in \theta\}$ *est un ouvert de* $\mathbb{R}^d$.

(iv) $\forall F$ *fermé de* $\mathbb{R}$, $f^{-1}(F) = \{x \in \mathbb{R}^d \,/\, f(x) \in F\}$ *est un fermé de* $\mathbb{R}^d$.

Remarque
Si $D \neq \mathbb{R}^d$, il faut modifier les points (iii) et (iv), et dire que $f^{-1}(\theta)$ est un ouvert de D et $f^{-1}(F)$ est un fermé de D.

Exemple. Malinvaud. Continuité de la fonction d'utilité et relations de préférence (p 17-18).

Théorème 6.25. *Si* f *et* g *sont continues alors* $f + g$, fg , $\dfrac{f}{g}$ *et* $f \circ g$ *sont continues lorsqu'elles sont définies.*

Remarque : lorsqu'on ne précise pas continu en tel ou tel point, il faut comprendre continu en tout point de l'ensemble où la fonction est définie.

Exemple d'application de ce résultat Comme $|x - x'| \leq \|(x, y) - (x', y')\|$ et $|y - y'| \leq \|(x, y) - (x', y')\|$, les applications définies par $(x, y) \mapsto x$, $(x, y) \mapsto y$ sont continues sur $\mathbb{R}^2$.
D'après le théorème précédent les applications définies par $(x, y) \mapsto x + y$, $(x, y) \mapsto xy$, puis $(x, y) \mapsto x^2 + 3xy$ et toutes les fonctions polynôme en deux variables x et y sont continues sur $\mathbb{R}^2$.
De la même façon toutes les fractions rationnelles en deux variables sont continues là où elles sont définies.

Théorème 6.26. *Soit* $f : X \to \mathbb{R}$ *(avec* X *compact) continue. Alors :*

(i) f *est bornée sur* X.

(ii) f *atteint ses bornes inférieure et supérieure.*

6.2 Dérivées des fonctions de plusieurs variables

Les fonctions de plusieurs variables sont des fonctions de chacune de leurs variables. Si elles sont dérivables par rapport à chaque variables comme fonctions d'une variable alors elles admettent des dérivées partielles. Le calcul des dérivées partielles se fait donc comme

le calcul des dérivées des fonctions réelles de la variable réelles (les autres variables sont considérées comme des constantes). Mais en dimension supérieure, dire qu'une fonction est dérivable, n'est pas seulement dire qu'elle a des dérivées partielles. On dit qu'une fonction f est dérivable ou différentiable en un point si elle a une bonne approximation linéaire (ou affine) en ce point. Lorsque f et g sont dérivables alors les propriétés habituelles sont vérifiées. Mais $f'(x)$ n'est pas un nombre mais une matrice. Dans la formule de la dérivation des fonctions composées par exemple l'ordre a alors une grande importance.

Remarque : la plupart du temps on écrit les variables d'une fonction à plusieurs variables en ligne mais dans l'écriture du développement de Taylor on considère des vecteurs colonnes.

Soit $f : \mathbb{R} \to \mathbb{R}$. La dérivée de f en x, si elle existe, est : $f'(x) = \lim\limits_{h\to 0} \dfrac{f(x+h) - f(x)}{h}$.

Définition 6.27. *Soit $f : \mathbb{R}^d \to \mathbb{R}$. On définit la dérivée partielle de f par rapport à x_i par*

$$\lim_{h\to 0} \frac{f(x_1\,,\,\ldots\,,\,x_i+h\,,\,x_{i+1}\,,\,\ldots\,,\,x_d) - f(x_1\,,\,\ldots\,,\,x_d)}{h}$$

si cette limite existe.

Notation

Cela se note $\dfrac{\partial f}{\partial x_i}(x_1\,,\,\ldots\,,\,x_d)$, $f_{x_i}(x_1\,,\,\ldots\,,\,x_d)$, $D_i f(x_1\,,\,\ldots\,,\,x_d)$.

Dans le cas de deux variables on a :

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

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

$\dfrac{\partial f}{\partial x}(x_0\,,\,y_0)$ est la pente de la tangente à la courbe $z = f(x\,,\,y_0)$ en $(x_0\,,\,y_0)$.

Définition 6.28. *Soit $f : \mathbb{R}^d \to \mathbb{R}$ une fonction ayant des dérivées partielles. Son gradient en x, noté $\nabla f(x)$ est le vecteur* $\nabla f(x) = \left(\dfrac{\partial f}{\partial x_1}(x)\,,\,\ldots\,,\,\dfrac{\partial f}{\partial x_d}(x)\right)$.

Remarque

Le gradient peut être considéré comme un vecteur de $\mathbb{R}^d$ mais aussi comme une matrice $1 \times n$.

Théorème 6.29. *Si f et g sont deux fonctions de $\mathbb{R}^d$ dans $\mathbb{R}$ avec des gradients, alors :*

(i) $\nabla(f+g) = \nabla f + \nabla g$

(ii) $\nabla(cf) = c\,\nabla f$ *où* $c \in \mathbb{R}$

Théorème 6.30. *Si g est définie sur $\mathbb{R}$ par $g(t) = f(r(t))$ où f est $\mathcal{C}^1$ de $\mathbb{R}^d$ dans $\mathbb{R}$ et r dérivable de $\mathbb{R}$ dans $\mathbb{R}^d$, alors g est dérivable et on a $g'(t) = \nabla f(r(t)) \cdot r'(t)$.*

Théorème 6.31. *Soient $f : \mathbb{R}^d \to \mathbb{R}$ de classe $\mathcal{C}^1$, $x = (x_1 , \dots , x_d)$, $h = (h_1 , \dots , h_d)$. Alors il existe $\theta \in]0 , 1[$ tel que $f(x + h) - f(x) = \nabla f(x + \theta h) \cdot h$.*

Définition 6.32. *Soient $f : \mathbb{R}^d \to \mathbb{R}$ de classe $\mathcal{C}^1$ et $v \in \mathbb{R}^d$.*
La dérivée selon le vecteur v en x est définie par $D_v f(x) = \lim_{t \to 0} \dfrac{f(x + tv) - f(x)}{t}$.

Remarque
Si $v = e_i$, on retrouve $\dfrac{\partial f}{\partial x_i}$.

Proposition 6.33. *On a $D_v f(x) = \nabla f(x) \cdot v$.*

Interprétation géométrique du gradient
La variation de f est la plus forte dans la direction de $\nabla f(x)$.

Cas des fonctions d'une variable

(i) f est dérivable en x_0 si $\lim_{h \to 0} \dfrac{f(x_0 + h) - f(x_0)}{h}$ existe.
Sa valeur ℓ est notée $f'(x_0)$.

(ii) On peut, de manière équivalente, écrire $\lim_{h \to 0} \dfrac{f(x_0 + h) - f(x_0) - \ell h}{h} = 0$.
On remarque que $h \to L(h) = \ell h$ est une application linéaire de $\mathbb{R}$ dans $\mathbb{R}$, que l'on appelle **différentielle** de f en x_0 et que l'on note $df(x_0)$.

(iii) Si f est dérivable en x_0 , alors pour h petit : $f(x_0+h)$ est "voisin" de $f(x_0)+f'(x_0)h$.
Donc $h \to f(x_0) + f'(x_0)h$ est une application affine qui "approche" $f(x_0 + h)$.

Définition 6.34. *f est différentiable en x s'il existe une application linéaire $L : \mathbb{R}^d \to \mathbb{R}$ telle que :*

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

L'application L est **la différentielle de f en** *x et se note $df(x)$.*

Remarque
Comme dans le cas $n = 1$ on a $f(x + h)$ "voisin" de $f(x) + df(x) \cdot h$, on a $f(x + h)$ est "approché" par l'application affine $f(x) + df(x) \cdot h$.
La différentielle, lorsqu'elle existe, est unique.

Proposition 6.35. *Si f est différentiable en x, alors ses dérivées partielles existent et on a :*

$$\begin{aligned} df(x) \cdot h &= \frac{\partial f}{\partial x_1}(x)\, h_1 + \dots + \frac{\partial f}{\partial x_d}(x)\, h_d \\ &= \nabla f \cdot h \end{aligned}$$

Remarque
La matrice de l'application linéaire $df(x)$ dans la base canonique est le gradient $\nabla f(x)$.

Proposition 6.36. *Si f est différentiable en x alors f est continue en x.*

L'existence des dérivées partielles de f n'implique pas la différentiabilité.

Théorème 6.37. *Si f admet des dérivées partielles et si elles sont continues alors f est différentiable. On dit que f est de classe $\mathcal{C}^1$.*

Proposition 6.38. *Si f et g sont différentiables on a :*

(i) $d(f+g)(x) = df(x) + dg(x)$

(ii) $d(\lambda f)(x) = \lambda\, df(x)$

(iii) $d(fg)(x) = f(x)\, dg(x) + g(x)\, df(x)$

(iv) $d\left(\dfrac{f}{g}\right)(x) = \dfrac{g(x)\, df(x) - f(x)\, dg(x)}{g^2(x)}$

Les dérivées partielles $\dfrac{\partial f}{\partial x_i}(x_1, \ldots, x_d)$ sont des fonctions de $x_1, \ldots, x_d$, et il arrive souvent qu'elles soient elles-mêmes dérivables.

Définition 6.39. *On écrit, lorsqu'elle existe,* $\dfrac{\partial^2 f}{\partial x_i\, \partial x_j} = \dfrac{\partial}{\partial x_i}\left(\dfrac{\partial f}{\partial x_j}\right)$ *et on dit qu'il s'agit d'une* **dérivée partielle seconde** *de f.*

Exemple
$f\colon \mathbb{R}^2 \to \mathbb{R},\ (x,y) \mapsto x^3y^4$. Alors $\frac{\partial^2 f}{\partial x \partial y}(x,y) = 12x^2y^3 = \frac{\partial^2 f}{\partial y \partial x}(x,y)$.

Théorème 6.40. *(Schwarz)*
Si les déirvées partielles $\dfrac{\partial f}{\partial x_i}$, $\dfrac{\partial^2 f}{\partial x_i\, \partial x_j}$ *existent et sont continues dans une boule autour de $(a_1 \ldots a_d)$ alors :*

$$\frac{\partial^2 f}{\partial x_i\, \partial x_j}(a) = \frac{\partial^2 f}{\partial x_j\, \partial x_i}(a)$$

Dans l'espace d'arrivée $\mathbb{R}^m$ on remplace les habituelles valeurs absolues par des normes.

Définition 6.41. *F de $\mathbb{R}^d$ dans $\mathbb{R}^m$ est* **différentiable** *en $x \in \mathbb{R}^d$ s'il existe une* **application linéaire** *L de $\mathbb{R}^d$ dans $\mathbb{R}^m$ telle que :*

$$\lim_{\|h\| \to 0} \frac{F(x+h) - F(x) - L \cdot h}{\|h\|} = 0\,.$$

L est la **différentielle** *de F en x et se note : $dF(x)$.*

Théorème 6.42. *F est différentiable en x si et seulement si ses composants sont différentiables et on a :*

$$dF(x)\cdot h \;=\; (\nabla f_1(x)\cdot h\,,\,\ldots\,,\,\nabla f_m(x)\cdot h)\,.$$

Définition 6.43. *La matrice*

$$\begin{bmatrix} \frac{\partial f_1}{\partial x_1}(x) & \cdots & \frac{\partial f_1}{\partial x_d}(x) \\ \cdots & \cdots & \cdots \\ \frac{\partial f_m}{\partial x_1}(x) & \cdots & \frac{\partial f_m}{\partial x_d}(x) \end{bmatrix}$$

est la matrice de $dF(x)$ et est appelée **matrice jacobienne** *de F en x et se note : $J(F)(x)$.*

Théorème 6.44. *Si F a des composantes de classe $\mathcal{C}^1$ alors elles sont différentiables et F est également différentiable.*

Proposition 6.45. *Si F de $\mathbb{R}^d$ dans $\mathbb{R}^m$ est linéaire, alors $dF(x) = F$.*

Proposition 6.46. *Si F est différentiable en x alors F est continue en x.*

Si F est une fonction de $\mathbb{R}^d$ dans $\mathbb{R}^m$, si G est une fonction de $\mathbb{R}^m$ dans $\mathbb{R}^q$, alors $G\circ F$ est une fonction de $\mathbb{R}^d$ dans $\mathbb{R}^q$.

Théorème 6.47. *Si F est différentiable en x, et si G est différentiable en $F(x)$, alors $G\circ F$ est différentiable en x et on a :*

$$d(G\circ F)(x) \;=\; dG(F(x))\circ dF(x)\,.$$

Le résultat théorique

Soient $f : \mathbb{R}^d \to \mathbb{R}$ et $g : \mathbb{R}^p \to \mathbb{R}^d$ deux fonctions différentiables. Écrivons $h = f \circ g$. D'après la règle de dérivation des fonctions composées nous avons (comme pour les fonctions de $\mathbb{R}$ dans $\mathbb{R}$) :

$$h'(x) = (f\circ g)'(x) = f'(g(x)).g'(x).$$

La fonction $f\circ g$ est une fonction de $\mathbb{R}^p$ dans $\mathbb{R}$. Sa dérivée est donc un vecteur ligne à p colonnes, la transposée de son gradient :

$$h'(x) = \left(\frac{\partial h}{\partial x_1} \quad \frac{\partial h}{\partial x_2} \quad \ldots \quad \frac{\partial h}{\partial x_p}\right).$$

La fonction g est une fonction de $\mathbb{R}^p$ dans $\mathbb{R}^n$. Sa dérivée est la matrice $n\times p$ composée des vecteurs transposés des gradients des coordonnées de g. Si $g(x) = (g_1(x), g_2(x), \ldots, g_2(x))$ (on devrait écrire ce vecteur en colonne si on voulait se conformer en toute rigueur aux choix du cours) la dérivée de g s'écrit :

$$g'(x) = \begin{pmatrix} \frac{\partial g_1}{\partial x_1} & \frac{\partial g_1}{\partial x_2} & \cdots & \frac{\partial g_1}{\partial x_p} \\ \frac{\partial g_2}{\partial x_1} & \frac{\partial g_2}{\partial x_2} & \cdots & \frac{\partial g_2}{\partial x_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g_n}{\partial x_1} & \frac{\partial g_n}{\partial x_2} & \cdots & \frac{\partial g_n}{\partial x_p} \end{pmatrix}.$$

Pour simplifier la présentation appelons $g = (g_1, g_2, \ldots, g_n)$ un point de $\mathbb{R}^n$. C'est un abus de notation, g ne désigne pas ici la fonction g mais un vecteur, un point dans $\mathbb{R}^n$. La dérivée de f en un point g est donnée par la transposée de son gradient :

$$f'(g) = \begin{pmatrix} \frac{\partial f}{\partial g_1} & \frac{\partial f}{\partial g_2} & \cdots & \frac{\partial f}{\partial g_n} \end{pmatrix}.$$

L'égalité matricielle $h'(x) = (f \circ g)'(x) = f'(g(x)).g'(x)$ signifie donc :

$$\begin{pmatrix} \frac{\partial h}{\partial x_1} & \frac{\partial h}{\partial x_2} & \cdots & \frac{\partial h}{\partial x_p} \end{pmatrix} = \begin{pmatrix} \frac{\partial f}{\partial g_1} & \frac{\partial f}{\partial g_2} & \cdots & \frac{\partial f}{\partial g_n} \end{pmatrix} \begin{pmatrix} \frac{\partial g_1}{\partial x_1} & \frac{\partial g_1}{\partial x_2} & \cdots & \frac{\partial g_1}{\partial x_p} \\ \frac{\partial g_2}{\partial x_1} & \frac{\partial g_2}{\partial x_2} & \cdots & \frac{\partial g_2}{\partial x_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g_n}{\partial x_1} & \frac{\partial g_n}{\partial x_2} & \cdots & \frac{\partial g_n}{\partial x_p} \end{pmatrix}.$$

Autrement dit pour tout $i = 1, \ldots, p$ on a

$$\frac{\partial h}{\partial x_i} = \sum_{k=1}^{n} \frac{\partial f}{\partial g_k} \frac{\partial g_k}{\partial x_i}.$$

Attention ! Quand g_k apparaît au dénominateur cela signifie seulement que l'on prend la dérivée de f par rapport à sa $k$ième variable. Quand il apparaît au numérateur g_k désigne la $k$ième coordonnée de g : c'est alors une fonction.

Un exemple

Prenons $f : \mathbb{R}^3 \to \mathbb{R}$ et $g : \mathbb{R}^2 \to \mathbb{R}^3$ deux fonctions différentiables définies par

$$f(x, y, z) = 2xy - 3(x + z),$$

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

On demande de calculer les dérivées partielles de la fonction de deux variables $h = f \circ g$. Pour se ramener au théorème général et ne pas s'embrouiller, il est préférable de changer les noms des variables dans l'expression de f :

$$f(g_1, g_2, g_3) = 2g_1g_2 - 3(g_1 + g_3).$$

La formule de dérivation en chaîne donne alors

$$\frac{\partial h}{\partial x} = \frac{\partial f}{\partial g_1}\frac{\partial (x + y^4)}{\partial x} + \frac{\partial f}{\partial g_2}\frac{\partial (y - 3x^2)}{\partial x} + \frac{\partial f}{\partial g_3}\frac{\partial (2x^2 - 3y)}{\partial x},$$

$$\frac{\partial h}{\partial y} = \frac{\partial f}{\partial g_1}\frac{\partial (x + y^4)}{\partial y} + \frac{\partial f}{\partial g_2}\frac{\partial (y - 3x^2)}{\partial y} + \frac{\partial f}{\partial g_3}\frac{\partial (2x^2 - 3y)}{\partial y}.$$

Pour $\frac{\partial h}{\partial x}$, on obtient :

$$\frac{\partial h}{\partial x} = (2g_2 - 3).1 + 2g_1.(-6x) + (-3).4x$$

Exprimée en fonction de x et y cette dérivée s'écrit :

$$\frac{\partial h}{\partial x} = 2y - 6x^2 - 3 - 12x(x + y^4) - 12x = -12xy^4 - 18x^2 + 2y - 12x - 3.$$

Je vous laisse le calcul de la deuxième dérivée partielle de h en exercice.
Remarque. On peut aussi écrire les choses sous la forme :

$$\frac{\partial h}{\partial x} = \frac{\partial f}{\partial x}\frac{\partial(x + y^4)}{\partial x} + \frac{\partial f}{\partial y}\frac{\partial(y - 3x^2)}{\partial x} + \frac{\partial f}{\partial z}\frac{\partial(2x^2 - 3y)}{\partial x},$$

mais c'est un peu risqué. Il ne faut surtout pas oublier de prendre les valeurs des dérivées partielles de f au point $(x + y^4, y - 3x^2, 2x^2 - 3y)$.

6.3 La formule de Taylor à l'ordre 2

6.3.1 En dimension 1

Donnons-nous une fonction d'une variable F définie sur $[0, 1]$ et trois fois continument dérivable. Soient u et v deux fonctions C^1 sur un intervalle I. La formule de dérivation d'un produit donne :

$$(u.v)' = u'.v + u.v' \text{ soit } u.v' = (u.v)' - u'.v.$$

Les fonctions apparaissant dans ces égalités étant continues elles sont intégrables sur les intervalles bornés de I. Si a et x sont des éléments de I on a donc

$$\int_a^x u(t)v'(t)dt = \int_a^x u(t)v(t)dt - \int_a^x u'(t)v(t)dt = u(x)v(x) - u(a)v(a) - \int_a^x u'(t)v(t)dt.$$

On note souvent

$$u(x)v(x) - u(a)v(a) = [u(t)v(t)]_a^x$$

En intégrant deux fois par parties la première des expressions suivantes on obtient :

$$\begin{aligned} F(1) - F(0) &= \int_0^1 F'(s)ds \\ &= [-(1 - s)F'(s)]_0^1 + \int_0^1 F''(s)(1 - s)ds \\ &= F'(0) + [-F''(s)(1 - s)^2/2]_0^1 + \int_0^1 F'''(s)(1 - s)^2/2ds \\ &= F'(0) + F''(0)/2 + \int_0^1 F'''(s)(1 - s)^2/2ds \end{aligned}$$

De cette égalité on déduit la majoration suivante :

$$|F(1) - F(0) - F'(0) - F''(0)/2| \leq \max_{s\in[0,1]} |F'''(s)|/6.$$

Appliquons ce résultat à la fonction $F : s \mapsto f(x+sh)$ lorsque f est une fonction trois fois continument dérivable sur $\mathbb{R}$. On a $F'(s) = hf'(x+sh)$, $F''(s) = h^2 f''(x+sh)$, $F'''(x+sh) = h^3 f'''(x+sh)$, $F(0) = f(x)$, $F(1) = f(x+h)$. La majoration précédente s'écrit donc ici :

$$|f(x+h) - f(x) - hf'(x) - f''(x)h^2/2| \leq h^3 \max_{s\in[x,x+h]} |f'''(s)|/6.$$

Notons M un majorant de la dérivée troisième de f sur l'intervalle d'étude de f et supposons qu'en x la dérivée de f soit nulle. Supposons que $f''(x)$ ne soit pas nul, par exemple qu'il soit positif. Prenons $h > 0$, on a alors :

$$f(x) + f''(x)h^2/2 - |h|^3 M/6 \leq f(x+h) \leq f(x) + f''(x)h^2/2 + |h|^3 M/6,$$

soit

$$f(x) + h^2(f''(x)/2 - |h|M/6) \leq f(x+h) \leq f(x) + h^2(f''(x)/2 + |h|M/6).$$

6.3.2 En dimension 2

Soit f une fonction de $\mathbb{R}^2$ dans $\mathbb{R}$ trois fois différentiable et (x,y) un point de $\mathbb{R}^2$. On introduit un accroissement (h,k) qu'il faut penser petit, et on applique le résultat énoncé plus haut à la fonction $F : s \mapsto f(x+sh, y+sk)$. Il nous faut calculer les dérivées de F. Pour cela on applique la règle de dérivation en chaîne :

$$F'(s) = h\frac{\partial f}{\partial x}(x+sh, y+sk) + k\frac{\partial f}{\partial y}(x+sh, y+sk) = \langle \nabla f(x+sh, y+sk), \binom{h}{k} \rangle,$$

$$\begin{aligned} F''(s) &= h^2\frac{\partial^2 f}{\partial x^2}(x+sh, y+sk) + hk\frac{\partial^2 f}{\partial x \partial y}(x+sh, y+sk) \\ &\quad + kh\frac{\partial^2 f}{\partial x \partial y}(x+sh, y+sk) + k^2\frac{\partial^2 f}{\partial y^2}(x+sh, y+sk) \\ &= (h,k)\mathrm{Hess} f(x+sh, y+sk)\binom{h}{k} \end{aligned}$$

La dérivée troisième fait apparaître des termes de degré 3 en h et k.

On obtient ainsi la formule de Taylor à l'ordre 2 pour les fonctions de deux variables.

Proposition 6.48. *Soit f une fonction de $\mathbb{R}^2$ dans $\mathbb{R}$ trois fois différentiable et (x,y) un point de $\mathbb{R}^2$. On a*

$$f(x+h, y+k) = f(x,y) + (h,k)\nabla f(x,y) + \frac{1}{2}(h,k)\mathrm{Hess} f(x,y)\binom{h}{k} + o(\|(h,k)\|^2).$$

6.3.3 Etude de certaines surfaces quadratiques

On cherche à étudier les polynômes quadratiques de la forme $z = Lx^2 + 2Mxy + Ny^2$.

Pour ce faire nous allons utiliser l'identité remarquable vue au collège : $(x+\alpha)^2 = x^2 + 2\alpha x + \alpha^2$ en écrivant quand il le faudra $x^2 + 2\alpha x = (x+\alpha)^2 - \alpha^2$. Si L n'est pas nul, on peut écrire

$$\begin{aligned} Lx^2 + 2Mxy + Ny^2 &= L(x^2 + 2Mxy/L + Ny^2/L) \\ &= L((x + My/L)^2 - M^2y^2/L^2 + Ny^2/L) \\ &= L((x + My/L)^2 + (NL - M^2)y^2/L^2). \end{aligned}$$

Ceci permet de voir que si $NL - M^2$ est strictement positif alors dans la parenthèse on trouve la somme de deux carrés et le graphe est un paraboloïde elliptique ("tourné" vers le haut si $L > 0$, vers le bas si $L < 0$). Si $NL - M^2$ est strictement négatif alors dans la parenthèse on trouve la différence de deux carrés et le graphe est un paraboloïde hyperbolique.
Si N n'est pas nul, en procédant de la même façon, on peut écrire

$$\begin{aligned} Lx^2 + 2Mxy + Ny^2 &= N(Lx^2/N + 2Mxy/N + y^2) \\ &= N(Lx^2/N + (y + Mx/N)^2 - M^2x^2/N^2) \\ &= N((LN - M^2)x^2/N^2 + (y + Mx/N)^2). \end{aligned}$$

On obtient que si $LN - M^2$ est strictement positif alors le graphe est un paraboloïde elliptique ("tourné" vers le haut si $N > 0$, vers le bas si $N < 0$). Si $NL - M^2$ est strictement négatif alors dans la parenthèse on trouve la différence de deux carrés et le graphe est un paraboloïde hyperbolique.
Dans le cas où L et N ne sont pas nuls tous les deux, ces deux façons de faire donnent bien les mêmes résultats. En effet, si $LN - M^2$ est strictement positif alors en particulier $LN > 0$ autrement dit L et N ont le même signe (l'orientation du paraboloïde elliptique est bien déterminée de la même façon).
Si L et N sont tous les deux nuls, et M est différent de 0, alors le graphe est un paraboloïde hyperbolique. Pour le voir il suffit d'écrire (encore une identité remarquable !) :

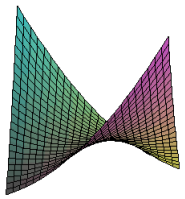
$$2Mxy = M(x+y)^2/2 - M(x-y)^2/2.$$

Quel que soit le signe de M nous avons une différence de deux carrés.
Enfin reste le cas où $LN - M^2$ est nul. On a alors

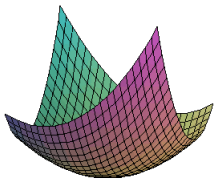
$$Lx^2 + 2Mxy + Ny^2 = L(x + My/L)^2$$

(quand $L \neq 0$ par exemple), et le graphe est un cylindre parabolique.

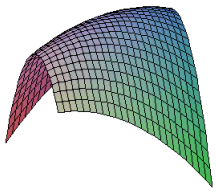
Exemples :
• $2x^2 - 6xy + y^2$
On calcule $2.1 - 3^2 = -7 < 0$. Le graphe est un paraboloïde hyperbolique :



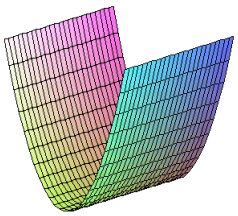
• $2x^2 - 2xy + 3y^2$
On calcule $2.3 - 2^2 = 2 > 0$ et $2 > 0$. Le graphe est un paraboloïde elliptique tourné vers le haut :



• $-x^2 + 2xy - 3y^2$
On calcule $(-1).(-3) - 1^2 = 1 > 0$ et $-1 < 0$. Le graphe est un paraboloïde elliptique tourné vers le bas :



• $x^2 + 2xy + y^2$
On calcule $1.1 - 1^2 = 0$. Le graphe est un cylindre parabolique :



Remarque : comme vous le voyez, il n'est pas toujours facile de reconnaître un paraboloïde hyperbolique ou un paraboloïde elliptique sur l'image donnée par l'ordinateur.

Proposition 6.49. *Il existe des coordonnées orthogonales X , Y , Z dans lesquelles $Z = k_1 X^2 + k_2 Y^2$.*

6.3.4 En dimension d

Soit f une fonction de $\mathbb{R}^d$ dans $\mathbb{R}$ trois fois différentiable et $x = (x_1, \ldots, x_d)$ un point de $\mathbb{R}^d$. On introduit un accroissement $h = (h_1, \ldots, h_d)$ qu'il faut penser petit, et on applique le résultat énoncé plus haut à la fonction $F : s \mapsto f(x+sh) = f(x_1 + sh_1, \ldots, x_d + sh_d)$. Il nous faut calculer les dérivées de F. Pour cela on applique la règle de dérivation en chaîne :

$$F'(s) = \sum_{i=1}^{d} h_i \frac{\partial f}{\partial x_i}(x+sh) = \langle \nabla f(x+sh), h \rangle,$$

$$\begin{aligned} F''(s) &= \sum_{i,j} h_i h_j \frac{\partial^2 f}{\partial x_j \partial x_i}(x+sh) \\ &= {}^t h \mathrm{Hess} f(x+sh) h \\ &= \langle \mathrm{Hess} f(x+sh) h, h \rangle \end{aligned}$$

La dérivée troisième fait apparaître des termes de degré 3 en les coordonnées de h.

On obtient ainsi la formule de Taylor à l'ordre 2 pour les fonctions de deux variables.

Proposition 6.50. *Soit f une fonction de $\mathbb{R}^d$ dans $\mathbb{R}$ trois fois différentiable et x un point de $\mathbb{R}^d$. On a*

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

Références

[1] D. Barnichon, *Mathématiques et statistiques appliquées à l'économie*, Bréal

[2] R. Boudon, *Les mathématiques en sociologie*,

[3] P. Bourdieu, *Méditations pascaliennes*,

[4] P. Cardaliaguet, *Optimisation et programmation dynamique*,

[5] A. Charpentier & M. Denuit, *Mathématiques de l'assurance non-vie*,

[6] P. G. Ciarlet, *Introduction à l'analyse numérique matricielle et à l'optimisation*,

[7] J.-C Culioli, *Introduction à l'optimisation*,

[8] A. Desrosières, *La politique des grands nombres histoire de la raison statistique...*,

[9] O. Ferrier, *Maths pour économistes*.

[10] J.-B. Hiriart-Urruty, *Les mathématiques du mieux faire*,

[11] P. Jensen, *Pourquoi la société ne se laisse pas mettre en équations*, Seuil.

[12] K. Lange, *Optimization*,

[13] E. Malinvaud, *Leçons de théorie micro-économique*, Dunod, Paris 1969

[14] N. Ordine, *L'utilité de l'inutile*,

[15] C.P. Simon & L. Blume, *Mathématiques pour économistes*,

[16] D. Smets, *Programmation linéaire et optimisation*,

[17] A. Supiot, *La Gouvernance par les nombres*, Fayard (une conférence donnée sur le sujet https://www.youtube.com/watch?v=q72RTYDtkY8)

[18] V. Vapnik, *The nature of statistical learning theory*, Springer