NOTES DE COURS DE L3, TOPOLOGIE ET CALCUL DIFFÉRENTIEL, 2022-23

G. David

April 17, 2023

Contents

J'essaie de modifier ces notes provisoires de temps en temps, et de les mettre sur ecampus assez régulièrement. Je ne crois pas que je mettrai des numéros de version. Attention: même si ces notes sont distribuées, il restera plein de fautes de frappe. Ne jamais croire à fond quelque chose, sous pretexte que c'est imprimé même avec des jolies fontes.

Désolé, c'est relativement peu dense (sauf pour les rappels de topologie); du coup j'espère que ca sera facile à lire, et aussi que les fautes de frappe devraient être plus faciles à corriger.

Je remercie encore Filipa Caetano qui m'a laissé ses notes de cours très utiles.

1 Rappels de Topologie.

J'essaie de rappeler les choses générales dont on aura besoin, en insistant sur la notion de compacité. Cette année encore on a essayé d'être intelligents et de se répartir les rappels de topologie et d'espaces vectoriels normés, mais quelques répétitions ne sont pas exclues.

Ceci dit, ce qui suit est à mi-chemin entre ce que je ferai en cours (proablement juste le cas d'espaces métriques en pensant très fort à $\mathbb{R}^n$) et un cours de topologie (qui devrait prendre plus son temps; ce qui suit, pour un nouveau, est un peu raide).

1.1 Espaces topologiques

Définition 1.1. *Un espace topologique, c'est la donnée d'un ensemble E, et d'une classe $\mathcal{O}$ des ensembles ouverts de E. Donc on peut dire que c'est un couple $(E, \mathcal{O})$, avec les propriétés suivantes.*
1. *D'abord, $\mathcal{O}$ est composé de parties (sous-ensembles) de E, et de plus $\emptyset$ et E sont dans $\mathcal{O}$.*
2. *Ensuite, toute union d'ouverts de E est un ouvert de E. Autrement dit, si pour $i \in I$ (un ensemble d'indices) $U_i \in \mathcal{O}$, alors $\cup_{i\in I} U_i \in \mathcal{O}$.*
3. *Finalement toute intersection de deux ouverts de E est un ouvert de E. Autrement dit, si $U \in \mathcal{O}$ et $V \in \mathcal{O}$, alors $U \cap V \in \mathcal{O}$.*

Noter qu'alors toute intersection finie d'ouverts de E est un ouvert de E.

On verra le cas le plus important des espaces métriques, mais la notion d'espace topologique est utile même en dehors de ce cadre.

Rien n'interdit de définir plusieurs topologies sur une même ensemble E; alors les notions qui suivent (fermés, applications continues, etc.) seront aussi différentes selon la topologie choisie.

On verra la topologie standard dans $\mathbb{R}^n$ bientôt (dès qu'on aura une distance).

Exemples extrèmes: la topologie discrète sur E, où $\mathcal{O} = \mathcal{P}(E)$ (où $\mathcal{P}(E)$ est l'ensemble de toutes les parties de E), et la topologie grossière, où $\mathcal{O} = \{\emptyset, E\}$.

Définition 1.2. *Soit $(E, \mathcal{O})$, un espace topologique. On dit que F est un fermé de E quand $E \setminus F$ (son complémentaire dans E) est ouvert.*

Donc $\emptyset$ et E sont fermés, la classe $\mathcal{F}$ des fermés de E est stable par intersections quelconques et unions finies. Et bien sûr $A \subset E$ peut être ouvert, fermé, aucun des deux, ou les deux à la fois.

La définition qui vient naturellement avec celle d'espace topologique est la notion d'application continue.

Définition 1.3. *On se donne des espaces topologiques $(E_1, \mathcal{O}_1)$ et $(E_2, \mathcal{O}_2)$, et une application $f : E_1 \to E_2$. On dit que f est continue quand $f^{-1}(A) \in \mathcal{O}_1$ pour tout $A \in \mathcal{O}_1$.*

Attention, ce sont bien les images inverses $f^{-1}(A) = \big\{x \in E_1\,;\, f(x) \in A\big\}$.

Exercice. Vérifier que la composée $f \circ g$ est continue quand f et g sont continues (et dire sur quels espaces $(E_1, \mathcal{O}_1)$, $(E_2, \mathcal{O}_2)$, $(E_3, \mathcal{O}_3)$ on se place).

Exercice. Vérifier que $f : E_1 \to E_2$ est continue si et seulement si $f^{-1}(A)$ est une partie fermée de E_1 pour tout ensemble $A \subset E_2$ fermé.

Exercice(un peu plus désagréable). On dit que f est continue au point x quand pour tout ouvert V de E_2 contenant $f(x)$ on peut trouver un ouvert U de E_1 contenant x tel que

$f(U) \subset V$. Vérifier que f est continue sur E_1 si et seulement si elle est continue en tout point.

Exercice. Je vais essayer de ne pas trop parler de voisinages, mais pour référence, disons qu'un voisinage de $x \in E$ est n'importe quel ensemble $V \subset E$ qui contient un ouvert qui contient x. Pour un espace métrique (voir plus loin), c'est donc n'importe quel ensemble $V \subset E$ qui contient une boule ouverte centrée en x. Vérifier que f est continue au point x si et seulement pour tout voisinage V_2 de $f(x)$, il existe un voisinage V_1 de x tel que $f(V_1) \subset V_2$.

On utilisera aussi la notion d'homéomorphisme.

Définition 1.4. *On se donne deux espaces topologiques $(E_1, \mathcal{O}_1)$ et $(E_2, \mathcal{O}_2)$, et une application $f : E_1 \to E_2$. On dit que f est un homéomorphisme (de E_1 dans E_2) quand f est une bijection de E_1 dans E_2, et $f : E_1 \to E_2$ est continue, ET enfin $f^{-1} : E_2 \to E_1$ est continue.*

On verra bientôt comment ceci se traduit dans les espaces métriques, qui seront notre cas principal si ce n'est unique.

Sous-espace d'un espace topologique. La notion a un sens mais la définition peut vous paraître bizarre a priori. Si $(E, \mathcal{O})$ est une espace topologique et $F \subset E$ est n'importe quel sous-ensemble, on définit une topologie sur F en décidant que $U \subset F$ est un ouvert de F si et seulement si on peut trouver un ouvert $V \in \mathcal{O}$ tel que $U = V \cap F$. Vérification assez facile: les 3 propriétés de la définition des ouverts.

1.2 Espaces métriques

Définition 1.5. *Un espace métrique est un couple (E, d), où E est un ensemble et d une distance sur E. Ce qui signifie que d est une fonction de $E \times E$ dans $[0, +\infty[$ qui vérifie les conditions suivantes:*

(1.1)

$d(x, x) = 0$ *pour tout* $x \in E$ *et* $d(x, y) > 0$ *pour tout choix de* $x, y \in E$ *tels que* $x \neq y$.

(1.2) $$d(x, y) = d(y, x) \text{ pour tout choix de } x, y \in E$$

(on dit que d est symétrique), et enfin l'inégalité triangulaire:

(1.3) $$d(x, z) \leq d(x, y) + d(y, z) \text{ pour tout choix de } x, y, z \in E.$$

On note parfois la distance dist(x, y) au lieu de $d(x, y)$ pour éviter des confusions.

Exemple de base. Sur $\mathbb{R}$, la distance donnée par la valeur absolue, donc $d(x, y) = |y - x|$. Vérification facile.

Exemple de base. Sur $\mathbb{R}^n$, la distance euclidienne donnée par

$$d(x, y) = \Big\{ \sum_{i=1}^{n} (y_i - x_i)^2 \Big\}^{1/2},$$

où l'on a noté $x = (x_1, x_2, \ldots, x_n)$ et $y = (y_1, y_2, \ldots, y_n)$.
Exemple un peu plus général, mais encore standard. On se donne un espace vectoriel normé $(E, ||\,||)$. La distance naturellement associée est donnée par $d(x, y) = ||y - x||$. L'inégalité triangulaire pour d se déduit de celle pour $||\,||$, à savoir le fait que $||x+y|| \leq ||x|| + ||y||$ pour tous $x, y \in E$. Mais on perd un peu de l'info, à savoir le fait que les vecteurs peuvent être multipliés par des scalaires λ, et que $||\lambda x|| = |\lambda|||x||$ pour $x \in E$ et $\lambda \in E$.

Noter au passage qu'on a aussi une distance sur $\mathbb{C}$, venant du module (autrement dit, $d(x, y) = |y - x|$ pour $x, y \in \mathbb{C}$. Et cette distance coincide avec la distance euclidienne si on voit $\mathbb{C}$ comme étant pareil que $\mathbb{R}^2$. De même, on peut mettre la distance euclidienne (et dans ce cas on dit plutôt "Hilbertienne") sur $\mathbb{C}^n$ (même formule que pour $\mathbb{R}^n$, mais avec des modules), et encore notre dernier exemple marche aussi avec des espaces vectoriels normés sur $\mathbb{C}$.

Si (E, d) est une espace métrique et $F \subset E$, la restriction de d à $F \times F$ est une distance sur F.

On utilisera beaucoup les boules associées à d.

Définition 1.6. *Soient (E, d) un espace métrique, $x \in E$, et $r > 0$.*
La boule ouverte de centre x et de rayon r est l'ensemble $B(x, r) = \big\{y \in E\,;\, d(x, y) < r\big\}$.
La boule fermée de centre x et de rayon r est l'ensemble $\overline{B}(x, r) = \big\{y \in E\,;\, d(x, y) \leq r\big\}$.

Attention cependant: tout le monde n'a pas les mêmes notations. D'abord, certains préfèrent utiliser la notation $B_r(x) = B(x, r)$ et $\overline{B}_r(x) = \overline{B}(x, r)$. Mais aussi d'autres notent $B(x, r)$ la boule fermée et par exemple $U(x, r)$ pour la boule ouverte. On pourrait aussi mettre un rond au-dessus du B.

Passons à la topologie associée à une distance d. On se donne un espace métrique (E, d) et on doit dire qui est $\mathcal{O}$, la classe des ouverts (associée à d).

Définition 1.7. *Soit (E, d) un espace métrique. On dit que $U \subset E$ est ouvert quand pour tout $x \in U$, il existe $r > 0$ tel que $B(x, r) \subset U$.*

La vérification des propriétés de la définition 1.1 est facile. Aussi, si (E, d) est une espace métrique et $F \subset E$, on a dit que la restriction de d à $F \times F$ est une distance sur F, qui définit donc une topologie sur F. Il se trouve que c'est bien la même topologie sur F que celle qui vient de la topologie de E. Vérification pas très compliquée, mais passée sous silence; c'est le contraire qui aurait été surprenant.

Evidemment, il se peut que deux distances différentes sur le même ensemble E donnent la même topologie. Le cas le plus simple est celui de distances équivalentes. On dit que les distances d_1 et d_2 sur E sont équivalentes quand in existe une constante $C > 0$ telle que

$$d_1(x, y) \leq C d_2(x, y) \text{ et } d_2(x, y) \leq C d_1(x, y) \quad \text{pour tous } x, y \in E. \tag{1.4}$$

Bien sûr il est important que C ne dépend pas de x ni de y. On vérifie sans mal que si d_1 et d_2 sont équivalentes, alors elles définissent la même topologie sur E. Mais c'est aussi assez

facile de définir d'autres distances sur $\mathbb{R}$ que la distance euclidienne, par exemple, qui ne lui sont pas équivalentes, mais qui donnent la même topologie sur $\mathbb{R}$.
Exemples sur $\mathbb{R}$: $d_1(x,y) = \min(14, |x-y|)$ ou $d_2(x,y) = \sqrt{|x-y|}$ (une vérifcation à faire pour l'inégalité triangulaire).

Proposition 1.8. *Soient (E_1, d_1) et (E_2, d_2) deux espaces métriques et $f : E_1 \to E_2$. Alors f est continue si et seulement si pour tout $x \in E_1$ et tout $\varepsilon > 0$, il existe $\delta > 0$ tel que $f(B(x,\delta)) \subset B(y,\varepsilon)$.*

Et on dit que f est continue en x quand pour tout $\varepsilon > 0$, il existe $\delta > 0$ tel que $f(B(x,\delta)) \subset B(y,\varepsilon)$. Ce dernier signifie que $d(f(y), f(x)) \leq \varepsilon$ pour tout $y \in E_1$ tel que $d(y,x) < \delta$. Vous devez reconnaître la définition de la continuité en x que vous connaissez, et retrouver que continue signifie bien continue en tout point.

J'ai essayé de me débrouiller pour ne pas avoir à donner deux noms différents (comme d_{E_1} et d_{E_2}) aux deux distances, mais peut-être que j'aurais dû le faire.

Démonstration. Dans un sens on suppose la condition sur les boules satisfaite, on se donne un ouvert U, et veut montrer que $f^{-1}(U)$ est ouvert. On se donne $x \in f^{-1}(U)$; alors $f(x) \in U$, et comme U est ouvert il existe $\varepsilon > 0$ tel que $B(f(x),\varepsilon) \subset U$. Ensuite on trouve δ comme ci-dessus. Et on observe que $B(x,\delta) \subset f^{-1}(U)$ car pour tout $y \in B(x,\delta)$, $d(f(y), f(x)) < \varepsilon$ donc $f(y) \in B(f(x),\varepsilon) \subset U$. Donc $f^{-1}(U)$ contient une boule ouverte centrée en x, ce qu'il fallait vérifier pour dire que $f^{-1}(U)$ est ouvert.

Dans l'autre sens, on suppose f continue et on doit vérifier l'histoire des boules. On se donne donc $x \in E_1$ et $\delta > 0$; on veut trouver δ. On note que $f^{-1}(B(f(x),\varepsilon))$ est ouvert contenant x. Donc il contient une petite boule $B(x,\delta)$. On vérifie que donc $f(B(x,\delta)) \subset B(f(x),\varepsilon)$. □

1.3 Suites et convergence (dans un espaces métrique)

L'idée générale est que les suites permettent de répondre à toutes les questions de topologie dans un espace métrique. Affirmation à prendre avec un grain de sel; comprendre aussi qu'en fait, hors des espaces métriques, une bonne partie de ce qui suit devient faux.

D'abord, rappelons la définition de convergence pour une suite: si $\{x_k\}$, $k \in \mathbb{N}$, est une suite à valeurs dans (de points de) E, et $y \in E$, on dit que la suite $\{x_k\}$ converge vers y, et on écrit aussi $\lim_{k\to+\infty} x_k = y$, quand $\lim_{k\to+\infty} d(x_k, y) = 0$.

Maintenant, comment reconnaître un ensemble fermé?

Proposition 1.9. *Soient (E,d) un espace métrique et $F \subset E$. Alors F est fermé si et seulement si pour toute suite $\{x_k\}$ de points de F qui converge vers une limite $y \in E$, on a que $y \in F$.*

Evidemment il y a des suites qui ne convergent pas. On dit juste que si F est fermé et la suite converge, alors la limite est dans F. D'ailleurs commençons par vérifier dans ce sens. Supposons que non. Comme F est fermé, son complémentaire $E \setminus F$ est ouvert, donc

puisque $y \in E \setminus F$ est dedans, il existe $r > 0$ tel que $B(y, r) \subset E \setminus F$. Mais par définition de la limite, on sait que pour k assez grand, $d(x_k, y) < r$, donc $x_k \in B(y, r)$. Ceci contredit la définition de $B(y, r)$ (puisque $x_k \in F$) et prouve la partie directe.

Dans l'autre sens, supposons que F n'est pas fermé, donc $E \setminus F$ n'est pas ouvert. Donc il existe $y \in E \setminus F$ tel qu'aucune boule $B(y, r)$ n'est contenue dans $E \setminus F$. En particulier, pour tout k la boule $B(y, 2^{-k})$ n'est pas contenu dans $E \setminus F$, donc on peut trouver $x_k \in B(y, 2^{-k})$ qui est dans F. La suite $\{x_k\}$ est dans F, on vérifie aisément qu'elle converge vers y (puisque $d(x_k, y) < 2^{-k}$ pour tout k). Et la limite y n'est pas dans F, donc la propriété définissante est violée, comme on voulait. □

Noter que j'ai parlé de comment reconnaître un fermé. J'aurais pu aussi dire comment construire une adhérence. Rappelons que l'adhérence d'un ensemble $A \subset E$, notée $\overline{A}$, est le plus petit fermé (de E) qui contient A. Son existence est facile à prouver: prendre juste l'intersection de tous les fermés F qui contiennent A. C'est bien un fermé (qui contient A), puisqu'une intersection quelconque de fermés set fermée. Donc ici, à cause de la proposition, $\overline{A}$ est aussi l'ensemble des limtes de suites convergentes dont tous les termes sont dans A (vérifiez-le tranquillement).

J'en profite pour dire que l'intérieur de A, qui est le plus grand ensemble A° (je devrais peut-être mettre le petit cercle directement au-dessus de A, mais c'est à la limite de ce que je sais faire en *late*χ) qui est contenu dans A. C'est l'union des ouverts qui sont contenus dans A, et c'est aussi le complémentaire de l'adhérence de $E \setminus A$. Vérifications faciles et on y reviendra si on s'en sert.

Et aussi (suite à une question), je "rappelle" qu'un ensemble $A \subset E$ est dit dense dans E quand son adhérence est E tout entier. L'exemple emblématique est celui de $\mathbb{Q}$ dans $\mathbb{R}$.

Exercice. Vérifier que pour un espace métrique (E, d), la partie A est dense dans E si et seulement si, pour tout $x \in E$ et tout $\varepsilon > 0$, la boule $B(x, \varepsilon)$ contient au moins un point de A. Muni ce ce critère, vérifier que $\mathbb{Q}$ est dense dans $\mathbb{R}$.

Maintenant vérifions que la continuité des fonctions aussi se voit avec des suites.

Proposition 1.10. *Soient (E_1, d_1) et (E_2, d_2) deux espaces métriques, $f : E_1 \to E_2$, et $x \in E_1$. Alors f est continue au point x si et seulement si, pour toute suite $\{x_k\}$ qui converge vers x, on a $\lim_{k\to+\infty} f(x_k) = f(x)$.*

Et du coup, $f : E_1 \to E_2$ est continue (partout) si et seulement si pour toute suite $\{x_k\}$ qui converge, on a $\lim_{k\to+\infty} f(x_k) = f(\lim_{k\to+\infty} x_k)$.

Démonstration maintenant. Supposons f continue en x et soit une suite $\{x_k\}$ qui converge vers x. Vérifions que $\lim_{k\to+\infty} f(x_k) = f(x)$. Pour tout $\varepsilon > 0$, on sait qu'il existe $\delta > 0$ tel que $\text{dist}(f(y), f(x)) < \varepsilon$ dès que $\text{dist}(y, x) < \delta$. Mais on sait que pour tout ε, et ensuite δ choisi comme ci-dessus, $\text{dist}(x_k, x) < \delta$ pour k assez grand, donc $\text{dist}(f(x_k), f(x)) < \varepsilon$ pour k assez grand. On en déduit bien que $\text{dist}(f(x_k), f(x))$ tend vers 0, comme souhaité.

Dans l'autre sens, supposons la propriété vraie pour les suites et démontrons la continuité de f en x. Soit donc $\varepsilon > 0$, et cherchons $\delta > 0$ tel que $\text{dist}(f(y), f(x)) < \varepsilon$ dès que $\text{dist}(y, x) < \delta$. Supposons qu'on ne puisse pas trouver δ (et prouvons une contradiction). Donc pour tout $k \geq 0$, $\delta = 2^{-k}$ ne marche pas. Ce qui veut dire qu'il existe $x_k \in B(x, 2^{-k})$

tel que $f(x_k) \notin B(f(x), \varepsilon)$. Donc $\text{dist}(f(x_k), f(x)) \geq \varepsilon$. Alors $\{x_k\}$ converge bien vers x, et pourtant $\{f(x_k)\}$ ne converge pas vers $f(x)$, ce qui signifie bien que la propriété annonée sur les suite est fausse, contrairement à ce qu'on a dit. □

2 Compacité.

2.1 Définitions et propriétés de base

On commence par une définition générale dans les espaces topologiques. Vous n'en aurez pas spécialement besoin, mais je pense que c'est bon à savoir. Donc si vous voulez, vous pouvez supposer que E est métrique et prendre la propriété de Bolzano-Weierstrass comme définition. Ca sera juste un peu dommage mais il ne vous arrivera rien de mal.

Fin du cours 1, 2023

Définition 2.1. *Soit $(E, \mathcal{O})$ un espace topologique et soit $K \subset E$. On dit que K est compact quand pour tout revouvrement de K par des ouverts U_i, $i \in I$, on, peut trouver un sous-recouvremment fini de K. Ce qui veut dire: si $K \subset \cup_{i\in I} U_i$, alors il existe un ensemble fini $I_0 \subset I$ tel que $K \subset \cup_{i\in I_0} U_i$.*

C'est peut-être bizarre, mais c'est très utile! J'ai mis cette définition parce que le plus souvent, K esra une partie de $E = \mathbb{R}^n$.

Remarque (que vous pouvez bien passer dans un premier temps). Normalement j'aurais dû donner la définition pour un espace topologique $(K, \mathcal{O})$ (défini sans référence à un espace plus grand), en prenant juste $K = E$ dans la définition ci-dessus. Heureusement c'est pareil: il est facile de vérifier que quand $K \subset E$, il est compact (comme dans la définition 2.1) si et seulement si, muni de la topologie restreinte, il est compact (avec la définition intrinsèque suggérée juste ci-dessus).

Exemples les plus simples: $[0, 1] \subset \mathbb{R}$ est compact, alors que $]0, 1]$ et $\mathbb{R}$ ne le sont pas.

Exercice. Trouver un recouvrement de $\mathbb{R}$, et même un de $]0, 1]$, qui dont on ne peut pas extraire un sous-recouvrement fini. Donc $\mathbb{R}$, et $]0, 1]$ ne sont pas compacts.

Démonstration du fait que l'intervalle $[a, b] \subset \mathbb{R}$ est compact. Je ne ferai sans doute pas en cours, et à la place on utilisera Bolzano-Weierstrass, mais comme c'est assez simple et unne bonne illustration de la définition, je l'écris quand même (pas obligé de lire). On se donne donc une famille U_i, $i \in I$, d'ouverts de $\mathbb{R}$ qui recouvrent I, et on cherche un sous-recouvrement fini. Considérons la partie A de $[a, b]$ composée des points $x \in [a, b]$ tels qu'il existe une partie finie de I telle que les ouverts U_i, $i \in I_0$, recouvrent tout l'intervalle $[a, x]$. Bien entendu, I_0 a le droit de dépendre de x. Noter que A contient a (donc n'est pas vide), puisqu'il suffit de prendre un seul ouvert U_i (choisi pour contenir a).

Ensuite, utilisons la propriété fondamentale de $\mathbb{R}$: l'ensemble $A \subset \mathbb{R}$, qui n'est pas vide (il contient a) et est majoré (par b), a donc une borne supérieure. C'est-à-dire qu'il existe un point ξ qui est le plus petit majorant de A. Donc, $x \leq \xi$ pour $x \in A$, et ξ est le plus petit nombre qui a cette popriété. Noter que $a \leq \xi$ (puisque $a \in A$) et $\xi \leq b$ (puisque b est clairement un majorant de $A \subset [a, b]$.

Vérifions que $\xi = b$. D'abord, puisque $\xi \in [a, b]$, il est contenu dans l'un des U_i; disons U_{i_0}. En fait, puisque U_i est ouvert, il contient même un intervalle $J =]\xi - 2\varepsilon, \xi + 2\varepsilon[$, avec $\varepsilon > 0$. Comme ξ est le plus petit majorant de A, il existe un $x \in A$ tel que $x \geq \xi - \varepsilon$. Donc on peut recouvrir $[a, x]$ par un nombre fini de nos ouverts, et en ajotant U_{i_0}, on obtient même un recouvrement de $[a, \xi + \varepsilon]$. On en déduit que $\xi = b$ (sinon, il y a des $x \in A$ qui sont même strictement que ξ. Et de plus, on a trouvé un bon recouvrement de $[a, \xi] = [a, b]$, ce qui démontre le fait que $[a, b]$ est compact. □

Revenons au cours de notre exposé. On verra que dans l'espace $\mathbb{R}^n$ muni de la distance euclidienne (et en fait c'est seulement une propriété de la topologie associée: une autre distance qui donnerait la même topologie donnerait évidemment le même résultat), $K \subset \mathbb{R}^n$ est compact si et seulement si K est à la fois fermé (dans $\mathbb{R}^n$) et borné (il existe $R > 0$ tel que $K \subset B(0, R)$).

Mais la notion est encore utile dans des espaces normés de dimension infinie, par exemple. Ceci dit, la boule unité fermée de $\ell^2 = \ell^2(\mathbb{N})$ (le plus simple des espaces vectoriels normés de dimension infinie) n'est pas compacte. Voir en fin de chapitre, et vous savez peut-être déjà à cause d'autres cours.

Voici une version de la définition de "compact", cette fois avec des fermés, obtenue juste en passant aux complémentaires.

Proposition 2.2. *Soit $(E, \mathcal{O})$ un espace topologique et soit $K \subset E$. Alors K est compact si et seulement si, pour toute famille F_i d'ensembles fermés telle que $K \cap \left(\bigcap_{i \in I_0} F_i \right) \neq \emptyset$ pour toute partie finie I_0 de I, on a aussi $K \cap \left(\bigcap_{i \in I} F_i \right) \neq \emptyset$.*

On peut soit prendre les F_i fermés dans E comme suggéré par l'énoncé, soit travailler directement dans K et prendre des fermés dans K. C'est pareil puisque les fermés de K sont les intersections avec K de fermés dans E (c'est comme avec la défnition des ouverts, et la vérification à partir de la défnition des ouverts est facile).

Pour démontrer la partie directe, prendre des fermés F_i comme dans l'énoncé, puis leurs complémentaires U_i. Si l'intersection $K \cap \left(\bigcap_{i \in I} F_i \right)$ est vide, alors les U_i recouvrent, on sait qu'une union finie suffit à recouvrir, et alors en passant au complémentaire $K \cap \left(\bigcap_{i \in I_0} F_i \right) = \emptyset$, une contradiction.

Dans l'autre sens c'est pareil: on suppose la propriété sur les fermés vraie, on se donne un recouvrement par des ouverts U_i, on note $F_i = E \setminus U_i$, ce sont des fermés dont l'intersection ne rencontre pas K; par la propriété, il y a une partie finie I_0 de I telle que $K \cap \left(\cap_{i \in I_0} F_i \right) = \emptyset$ (sinon, l'intersection infinie aussi ne rencontre pas K), et donc l'union finie des complémentaires U_i recouvre K.

Corollaire 2.3. *Si $K \subset E$ est compact, et si on se donne une suite décroissante $\{F_i\}$, $i \in \mathbb{N}$ de fermés non vides dans K (donc $F_i \subset F_j$ pour $i > j$), alors $\cap_{i \in \mathbb{N}} F_i \neq \emptyset$.*

Démonstration: appliquer le critère précédent en notant que toute intersection finie de F_i est égale au dernier des F_i de la famille.

Remarque 2.4. Si E est un espace métrique, ou au moins un espace topologique séparé et si $K \subset E$ est compact, alors K est fermé (dans E).

Ca explique que $[0,1[$ ne soit pas compact.

Séparé signifie que pour $x \neq y$ dans E, on peut trouver un ouvert V_y qui contient y et un ouvert W_y qui contient x mais tels que $V_y \cap W_y = \emptyset$. Et à mon sens tous les espaces topologiques raisonables vérifient cela. C'est vrai en particulier dans le cas d'espaces métriques, où l'on peut prendre $V_y = B(y,r)$ et $W_y = B(x,r)$, avec $r = d(x,y)/3$.

Démonstration de la remarque. On suppose que K n'est pas fermé (et on va prouver qu'il n'est pas compact). On peut donc trouver x qui est dans l'adhérence de K (le plus petit fermé qui contient K, qui est l'intersection de tous les fermés qui contiennent K), mais qui n'est pas dans K.

Pour chaque $y \in K$, on sait que $y \neq x$. L'axiome de séparation donne un ouvert V_y qui contient y et un ouvert W_y qui contient x, tels que $V_y \cap W_y = \emptyset$.

Les V_y, $y \in K$, forment un recouvrement de K par des ouverts. Par compacité il existe donc une partie finie K_0 de K telle que $K \subset \cup_{y \in K_0} V_y$. Mais cette partie ne rencontre pas $W = \cap_{y \in K_0} W_y$, qui pourtant est un ouvert qui contient x. Mais x est dans l'adhérence $\overline{K}$, donc tout voisinage de x rencontre K. Autrement dit, W rencontre K, donc W rencontre $\cup_{y \in K_0} V_y$ qui est plus grand, une contradiction. □

Corollaire 2.5. *Supposons encore E métrique ou au moins séparé. Si $K \subset E$ est compact, et si $L \subset K$, alors L est compact si et seulement si L est fermé.*

On a déja vu que tout compact de E est fermé (à condition que E soit "séparé", c'est pour cela que je le mets); donc K est fermé (dans E), et aussi L s'il est compact. D'ailleurs, puisque K est fermé, L est fermé dans K si et seulement si il est fermé dans E (vérification simple, mais c'est faux si $K =]0,1[$ dans $E \subset \mathbb{R}$, puisque K est toujours fermé dans K).

Donc il reste la partie directe intéressante. On se donne L fermé dans K compact. On se donne par exemple un recouvrement de L par une famille d'ouverts U_i, $i \in I$. On ajoute $E \setminus L$ qui est ouvert, et on trouve un recouvrement de K. Par compacité, on peut recouvrir K par un nombre fini des U_i, plus peut-être $E \setminus L$. Ceci recouvre $L \subset K$ aussi, et comme $E \setminus L$ ne sert à rien sur cet ensemble, les U_i restants recouvrent L, comme souhaité. □

Cet énoncé nous simplifie la vie pour trouver de nouveaux compacts: dès qu'on aura montré que tout cube $Q = [-N,N]^n \subset \mathbb{R}^n$ est compact dans $\mathbb{R}^n$, ou en déduira que les ensembles fermés bornés dans $\mathbb{R}^n$ sont compacts. Voir ci-dessous.

2.2 Compacité et espaces métriques (Bolzano-Weierstrass, complétude, et précompacité)

Dans le cadre des espaces métriques, la compacité se caractérise bien par deux propriétés en apparence plus simples: la précompacité (plus complétude) et la propriété de Bolzano-Weierstrass. Je commence par la définition de BW, qu'on utilise plus souvent.

J'ai besoin de la notion de suite extraite.

Définition 2.6. *Soit E un ensemble et $\{x_k\}_{k\geq 0}$ une suite à valeur dans E. Autrement dit, pour tout $k \in \mathbb{N}$, on se donne un point $x_k \in E$. On peut aussi dire qu'une suite, c'est juste une application de $\mathbb{N}$ dans E. Une suite extraite de $\{x_k\}$ (on dit aussi une sous-suite de $\{x_k\}$), c'est la suite associée à une application de la forme $\ell \to x_{\varphi(\ell)}$, où φ est une application strictement croissante $\ell \to \varphi(\ell)$ de $\mathbb{N}$ dans $\mathbb{N}$.*

Désolé, c'est un peu lourd. J'ai décidé comme beaucoup qu'une suite, ça commence avec x_0, mais on aurait pu considérer qu'elle commence avec x_1, c'est-à-dire, prendre une application de $\mathbb{N}^*$ dans E. Finalement, prendre une sous-suite, ça consiste juste à ne garder que certains des x_k, puis à les renuméroter pour avoir une suite. Ainsi, si on commence les suite avec x_1, $\varphi(\ell)$ est juste le numéro du ℓ-ième point de la suite qu'on a gardé.

Noter que $\lim_{\ell\to+\infty} \varphi(\ell) = +\infty$ puisqu'en fait $\varphi(\ell) \geq \ell$ pour tout ℓ (par récurrence facile).

Définition 2.7. *Soit (E, d) un espace métrique et soit $K \subset E$. On dit que K a la propriété de Bolzano-Weierstrass quand pour toute suite $\{x_k\}$ à valeurs dans K, on peut extraire une sous-suite $\{x_{\varphi(\ell)}\}$, $\ell \in \mathbb{N}$, qui a une limite dans K.*

Donc d'une part il y a une limite, et d'autre part elle est dans K. Si on avait pris $E = K$ (définition intrinsèque), ce qui aurait été plus logique mais correspond moins à notre pratique où K sera systématiquement $\mathbb{R}^n$, on n'aurait eu qu'à se préoccuper de l'existence de la limite.

Le cas typique de cet histoire de sous-suites est dans l'espace compact $\{-1, 1\}$ avec jute deux points: on prend $x_k = (-1)^k$; alors la suite ne converge pas, mais en prenant par exemple la sous-suite $\{x_{2\ell}\}$ ou $\{x_{2\ell+1}\}$, on obtient des suites convergentes avec des limites qui peuvent être différentes.

La propriété de Bolzano-Weierstrass est bien pratique aussi, et caractérise la compacité.

Théorème 2.8. *Soit (E, d) un espace métrique et soit $K \subset E$. Alors K est compact si et seulement si il a la propriété de Bolzano-Weierstrass.*

Ca c'est un théorème! Mais on va attendre un peu pour parler de la démonstration. Parlons déjà de complétude.

Proposition 2.9. *Soit (K, d) un espace métrique compact. Alors K est complet.*

Souvent K est un sous-espace d'un espace métrique (E, d) plus grand, mais c'est inutile d'en parler dans l'énoncé. On doit montrer que K est complet, c.-à-d. que toute suite de Cauchy dans K converge. Soit donc $\{x_k\}$ une suite de Cauchy. Donc pour tout entier n, il existe $k_0 = k_0(n)$ tel que $d(x_k, x_\ell) \leq 2^{-k}$ pour $k, \ell \geq k_0$. On peut aussi choisir les entiers $k_0(n)$ de façon que $n \to k_0(n)$ soit une suite croissante, par exemple en remplaçant $k_0(n)$ par $k_1(n) = \max\{k_0(m)\,;\, 0 \leq m \leq n\}$.

Posons maintenant $F_n = \overline{B}(x_{k(n)}, 2^{-n+1})$; c'est un ensemble fermé, et $F_{n+1} \subset F_n$ puisque $d(x_{k(n+1}, k(x_n))) \leq 2^{-n}$. Enfin F_n est non vide, puisque $x_{k(n)} \in F_n$. Par le premier corollaire de la définition d'un compact, on peut trouver $x \in \cap_n F_n$.

Il reste à vérifier que $\lim_{k\to+\infty} x_k = x$. Soit $\varepsilon > 0$. Choisissons n tel que $2^{-n} < \varepsilon/4$. Par définition, $d(x, x_{k(n)}) \leq 2^{-n+1}$. Et en plus, pour $k \geq k(n)$, on a $d(x_k, x_{k(n)}) \leq 2^{-n}$. Donc aussi $d(x_k, x_{k(n)} \leq 2^{-n+2} < \varepsilon$. C'est ce qu'il fallait démontrer. □

Remarque: avec la propriété de Bolzano-Weierstrass, la démonstration est encore plus courte: on sait (sans doute) que pour toute suite de Cauchy $\{x_k\}$, la suite est convergente si et seulement si elle admet une sous-suite (suite extraite) convergente. Donc on se donne une suite de Cauchy, on utilise BW pour trouver une sous-suite convergente, et on en déduit que la suite initiale convergeait!

Il y a une autre notion utile, la "précompacité":

Définition 2.10. *Soit (E, d) un espace métrique. On dit que E est précompact (et les anglosaxons disent completely bounded) quand, pour tout $\varepsilon > 0$, on peut recouvrir E par un nombre fini de boules de rayon ε.*

Evidemment on ne dit pas combien le nombre $N(\varepsilon)$ de boules dont on a besoin pour recouvrir E vaut: il pourrait tendre vers $+\infty$ très vite quand ε tend vers 0. Noter aussi qu'au moins la notion ne change pas quand on remplace d par une distance équivalente.

Exemple: $[0,1]^n$ dans $\mathbb{R}^n$: prendre une grille $G = \frac{\varepsilon}{10\sqrt{n}}\mathbb{Z}$, et utiliser les boules de rayon ε centrées sur $G \cap [0,1]^n$. Evidemment, $[0,1[$ aussi marche avec la même démonstration. Quand E est un sous-ensemble d'un F (et on prend la restriction de la distance d définie sur F), on peut soit insister pour recouvrir E par des boules centrées sur E, soit autoriser aussi des boules centrées sur $F \setminus E$. On vérifie que les deux notions sont équivalentes (les boules centrées sur $F \setminus E$ sont soit inutiles quand elles sont loin de E, soit elles sont près de E et peuvent être remplacées par des boules à peine 2 fois plus grandes, et centrées sur E.

Contrexemple: $\mathbb{R}$: il n'est pas possible de le recouvrir avec un nombre fini de boules de rayon 1. Par contre, $]0,1[$ n'est pas compact pour l'autre raison (voir le thérème qui suit): il n'est pas complet parce qu'il y a des suites dans $]0,1[$ qui tendent vers 1 (dans $\mathbb{R}$), sont de Cauchy, mais n'ont pas de limite dans $]0,1[$. C'est amusant que $\mathbb{R}$ et $]0,1[$ ne sont pas compacts pour des raisons qui semblent différentes, alors qu'ils sont homéomorphes (équivalents pour la topologie). C'est que les distances correspondantes sont, elles, bien différentes. Par contre, la compacité est une notion purement topologique, donc c'est naturel que les deux soient de la même sorte (non compacts).

Théorème 2.11. *Soit (E, d) un espace métrique. Alors E est compact si et seulement si il est à la fois complet et précompact (totalement borné).*

Avec le théorème 2.8 sur BW, ceci nous fait les deux résultats principaux sur les espaces métriques compacts. Le théorème 2.8 (la propriété de BW) est sans doute le plus utile; mais le théorème 2.11 est un intermédiaire bien pratique.

Voici maintenant un plan pour la démonstration jointe des deux théorèmes. Partons de la propriété "E est compact". Elle implique que E est complet (vu ci-dessus), vérifions aussi qu'elle implique la précompacité.

On se donne donc $\varepsilon > 0$. On veut recouvrir E par un nombre fini de boules de rayon ε. On a déjà un recouvrement de E par les boules $B(x, \varepsilon)$, $x \in E$; ça fait beaucoup de boules, mais la compacité dit qu'on peut trouver un sous-recouvrement fini. Et voià.

Démonstration du fait que si E est complet et précompact, alors il vérifie la propriété de Bolzano-Weierstrass. On va utiliser un outil très pratique, qu'on appelle le procédé diagonal d'extraction de sous-suites.

Soit E complet et précompact. On commence par une préparation. Pour $k \geq 0$ entier, choisissons un recouvrement fini de E par des boites $B_{k,\ell} = B(x_{k,\ell}, 2^{-k})$, où $\ell \in L(k)$, un ensemble d'indices dépendant de k.

Ensuite on se donne une suite $\{x_n\}$ (j'ai déjà utilisé k). On part avec $k = 0$, et on choisit une boite B_{0,ℓ_0} qui contient x_n pour une infinité de valeurs de n. On peut la trouver, puisqu'il y a un nombre fini de boites $B_{0,\ell}$ et une infinité de valeurs de n. On en profite pour définir une première sous-suite, qu'on va noter $\{x_{\varphi_1(j)}\}$, $j \geq 1$, où en fait $\varphi_1(j)$ est le j-ème élément n tel que $x_n \in B_{0,\ell_0}$. Bref on jette tous les n tels que $x_n \notin B_{0,\ell_0}$ et on renumérote les autres. C'est un peu plus pratique de numéroter à partir de 1, alors on fait ça. Notons au passage que $\varphi_1(j) \geq j$ pour tout j.

On recommence à partir de la suite $x_{\varphi_1(j)}$. Il existe une boule B_{1,ℓ_1} parmi nos boules de génération $k = 1$ qui contient $x_{\varphi_1(j)}$ pour une infinité de j. On note maintenant $\varphi_2(j)$ le j-ème entier de la forme $\varphi_1(i)$ tel que $x_{\varphi_1(j)} \in B_{1,\ell_1}$. Ceci nous donne une suite extraite $x_{\varphi_2(j)}$, qui est en fait une sous-suite de la première. Donc en fait $x_{\varphi_2(j)} \in B_{0,\ell_0} \cap B_{1,\ell_1}$ pour tout $j \geq 1$, et comme précédemment $\varphi_2(j) \geq \varphi_1(j) \geq j$.

Et on recommence: pour tout $m \geq 1$ on construit une sous-suite $\varphi_{m+1}(j)$, qui est en fait une sous-suite de $\varphi_m(j)$, et qui vérifie aussi que pour tout $j \geq 1$, $x_{\varphi_{m+1}(j)} \in B_{0,\ell_0} \cap B_{1,\ell_1} \ldots \cap B_{m+1}, \ell_{m+1}$, où B_{m+1}, ℓ_{m+1} est une boite de génération $m+1$ qu'on choisit pour l'occasion.

Maintenant on utilise le procédé diagonal, qui consiste juste à définir une nouvelle application croissante ψ de N^* dans N^*, par $\psi(j) = \varphi_j(j)$. Bref, on prend le premier élément de la première suite, puis le secont élément de la seconde suite, et ainsi de suite. Noter que ψ est strictement croissante, parce que l'on a vu que $\varphi_m(j+1) \geq \varphi_m(j)$ pour tout m et tout j, et donc $\psi(m+1) = \varphi_{m+1}(m+1) \geq \varphi_m(m+1) > \varphi_m(m) = \psi(m)$. Ensuite, par ce qui précède, $x_{\psi(j)} = x_{\varphi_j(j)} \in B_{0,\ell_0} \cap B_{1,\ell_1} \ldots B_{j,\ell_j}$ pour tout j, ce qui implique que $x_{\psi(j)}$ est dans la boite B_{m,ℓ_m} pour tout $j \geq m$. On en déduit assez facilement que $\{x_{\psi(j)}\}$ est une suite de Cauchy (faites la vérification, en utilisant le fait que le rayon des boules B_{m,ℓ_m} tend vers 0). Par complétude, elle converge, et BW s'en déduit puisque $\{x_{\psi(j)}\}$ est une suite extraite. □

Démonstration du fait que BW implique la compacité (définie avec la propriété sur les recouvrements). Pas fait en cours (pour ne pas abuser), mais argument intéressant caractéristique de comment on travaille avec les compacts. On se donne un recouvrement de E par des ouverts U_i, $i \in I$, et on commence par démontrer que:

(2.1) il existe $\varepsilon_0 > 0$ tel que pour tout $x \in E$, $B(x, \varepsilon_0)$ est contenu dans l'un des U_i.

En fait, pour tout $x \in E$, notons $\varepsilon(x) > 0$ la borne sup des $\varepsilon > 0$ tels que $B(x, \varepsilon)$ est contenu dans l'un des ouverts U_i. On veut juste montrer que $\inf_{x \in E} \varepsilon(x) > 0$ (et alors on prend pour ε_0 la moitié de l'inf, disons). Donc supposons que l'inf est 0, et choisissons pour tout $k \geq 0$ un $x_k \in E$ tel que $\varepsilon(x) \leq 2^{-k}$. Puis par BW trouvons une suite extraite $\{x_{k(j)}\}$ qui converge vers une limite x. Par définition de $\varepsilon(x)$ comme une borne supérieure, il existe $i \in I$

tel que $B(x, \varepsilon(x)/2) \subset U_i$. Alors, pour tout j est assez grand pour que $d(x_{k(j)}, x) < \varepsilon(x)/4$, on a que $B(x_{k(j)}, \varepsilon(x)/4) \subset B(x, \varepsilon/2) \subset U_i$, ce qui implique que $\varepsilon(x_{k(j)}) \geq \varepsilon(x)/2$, ce qui donne la contradiction souhaitée. Donc on a (2.1).

Reste à trouver un recouvrement fini. Essayons de trouver des points $x \in E$ tels que les $B(x, \varepsilon_0)$ recouvrent E. Si E est vide, un seul ouvert (probablement $\emptyset$) suffit à le recouvrir. Ou même, le recouvrement vide avec zéro ouvert. Sinon, choisissons $x_1 \in E$ et posons $B_1 = B(x_1, \varepsilon_0)$. Si $E \subset B_1$, on a gagné, E est recouvert par B_1, qui par construction est contenu dans un des U_i. Sinon, prenons $x_2 \in E \setminus B_1$ et posons $B_2 = B(x_2, \varepsilon_0)$. Si $E \subset B_1 \cup B_2$, on a gagné (chaque B_i est contenue dans un U_j). Sinon, on continue.

Si à un moment on doit s'arrrêter, on a gagné puisqu'alors E est contenu dans une union finie de B_i, chacune contenue dans un U_j. Sinon, on trouve une suite (infinie) $\{x_i\}$, et chaque point est à distance au moins ε_0 de tous les précédnts. Ceci reste vrai pour n'importe quelle sous -suite de $\{x_i\}$, et il est facile de voir qu'une telle suite ne peut converger, puisque si x_∞ était la limite, on aurait $d(x_i, x_{i+1}) \leq d(x_i, x_\infty) + d(x_\infty, x_{i+1})$ tend vers 0.

Je crois qu'on a fini de tout démontrer. □

2.3 Des exemples

D'abord, dans $\mathbb{R}$, tout intervalle fermé borné est compact. C'est pour ça que souvent l'on appelle ça un intervalle compact.

Vérification la plus simple: il est complet car fermé dans $\mathbb{R}$, et précompact par découpage en petites boites (voir plus haut). Mais Bolzano-Weierstrass est aussi facile à démontrer (par dichotomie: couper l'intervalle en deux parties égales et trouver une sous-suite qui reste d'un seul coté, puis recommencer. C'est pareil que plus haut, mais un peu plus simple). Enfin, la propriété avec les recouvrements par des ouverts se démontre aussi directement: voir au-dessous de la définition 2.1 □

La première démonstration pour les intervalles marche pareil pour montrer que dans $\mathbb{R}^n$, tout produit d'intervalles compacts est compact. On va en déduire le résultat suivant.

Proposition 2.12. *Soit $K \subset \mathbb{R}^n$ (muni de la topologie usuelle). Alors K est compact si et seulement si K est fermé et borné.*

J'ai dit topologie usuelle, parce que la compacité est une histoire de topologie. Mais $\mathbb{R}^n$ est aussi muni de diverses distances naturelles (vues ci-dessus) toutes équivalentes, et qui donnent cette topologie. Prenons-en une, la distance euclidienne, et travaillons avec elle.

Borné signifie qu'il existe $R > 0$ tel que $K \subset B(0, R)$. [Avec une autre distance équivalente on aurait un autre R, mais K resterait borné. Par contre évitons $d(x, y) = \min(1, |x-y|)$ pour laquelle $\mathbb{R}$ est la boule (fermée) de centre 0 et de rayon 1 et $\mathbb{R}$ aurait l'air borné.]

La partie intéressante est maintenant facile: si K est fermé borné, et R tel que $K \subset B(0, R)$, alors K est aussi un sous-espace fermé de $K_R = [-R, R]^n$, qui est compact. On a vu qu'alors K est compact.

Dans l'autre sens, on a vu que si K est compact, K est fermé. Il est aussi borné, parce que sinon pour tout entier $k \geq 0$, on peut trouver $x_k \in K$ tel que $||x_k|| \geq 2^k$, et il est clair

que cela donne une suite $\{x_k\}$ dont aucune sous-suite ne peut converger (elles tendent toutes vers l'infini). □

Notons encore que dans $E = B(0,1) \subset \mathbb{R}^n$, il est faux que les compacts de E sont les fermés bornés de E, puisque E est borné et fermé dans E mais n'est pas compact.

Ensuite parlons un peu de produits. Initialement, j'ai mis ce paragraphe pour donner un exemple d'application de BW et du procédé diagonal. On se dirige vers des produits (finis ou dénombrables) d'espaces métriques compacts. Pour un produit fini d'espaces métriques $(E_1, d_1), \ldots, (E_n, d_n)$, il est facile de construire une distance d sur $E = E_1 \times \ldots \times E_n$. Voir un exercice fait en TD. On peut par exemple prendre

$$d(x,y) = \sum_{i=1}^{n} d_i(x_i, y_i), \tag{2.2}$$

où on a noté $x = (x_1, \ldots, x_n)$ et $y = (y_1, \ldots, y_n)$, mais on peut aussi prendre

$$d(x,y) = \sup_{1 \le i \le n} d_i(x_i, y_i), \tag{2.3}$$

ou plein d'autres distances intermédiaires. Il se trouve que toutes (celles qui sont raisonnables) sont équivalentes les unes aux autres, et aussi donnent la même topologie qui est celle qu'on mettrait sur l'espace topologique produit. Mais je ne définirai pas la topologie produit ici.

Ensuite on peut se poser la question d'un produit infini $E = \prod_{i \in I} E_i$, où chaque E_i est muni d'une distance d_i. Il se trouve que pour tout choix de I, même non dénombrable, on peut définit une topologie produit sur E (à partir de la topologie sur chaque E_i), mais je ne le ferai pas non plus. Et dans ce cadre, un produit d'espaces compact est compact, mais je ne ferai pas la démonstration, qui nous mènerait trop loin.

Par contre, on va s'occuper du cas d'un produit dénombrable (et on va même prendre $I = \mathbb{N}$), parce que c'est amusant.

Donc on se donne un espace métrique (E_i, d_i) pour $i \in \mathbb{N}$, et on commence par essayer de définir une distance sur $E = \prod_{i \in \mathbb{N}} E_i$. En fait il y en aurait plein mais on va en choisir une raisonnable. Voir aussi un exercice fait en TD, sur lequel je me reposerai pour ne pas faire la démonstration en cours. On commence par modifier chaque d_i pour la rendre bornée. On pose

$$\widetilde{d}_i(x,y) = \min(1, d_i(x,y)) \quad \text{pour } x, y \in E_i. \tag{2.4}$$

C'est encore une distance (vérification assez facile, voir TD), qui n'est pas forcément équivalente à d_i, mais qui définit quand même la même topologie (les petites boules sont les mêmes).

Pourquoi faire ce remplacement? C'est que maintenant l'on peut poser

$$d(x,y) = \sum_{i \in \mathbb{N}} 2^{-i} \widetilde{d}_i(x_i, y_i) \quad \text{pour } x = (x_1, \ldots, x_n, \ldots) \text{ et } y = (y_1, \ldots, y_n, \ldots) \text{ dans } E. \tag{2.5}$$

La série converge absolument (puisque $\widetilde{d}_i(x_i, y_i) \leq 1$), et ceci est une distance. De plus il se trouve qu'elle donne bien la bonne topologie sur E (que je n'ai pas définie). En tout cas, on a la propriété raisonnable (vérifiée en TD) que pour toute suite $\{x^k\}$ dans E et tout $x \in E$ (j'indexe par le haut pour pouvoir dire que les coordonnées de x^k sont les x^k_n),

$$\lim_{k\to+\infty} x^k = x \quad \text{si et seulement si} \quad \lim_{k\to+\infty} x^k_n = x_n \text{ pour tout } n. \tag{2.6}$$

Donc si la convergence das E, muni de la distance d, se vérifie coordonnée par coordonnée, comme dans $\mathbb{R}^n$ ou un produit fini.

Fin du cours 2, 2023

Muni de toutes ces définition et préliminaires, on a le théorème suivant.

Théorème 2.13. *Soit $E = \prod_{i\in I} E_i$, où $I = \mathbb{N}^*$ ou $i = \{1, 2, \dots n\}$, et chaque E_i est muni d'une distance d_i, et est supposé compact. Alors le produit E, muni de la distance ci-dessus, est aussi compact.*

On va utiliser la caractérisation par Bolzano-Weierstrass (mais en allant un peu vite dans les étapes, et encore plus en cours). Donc on se donne une suite $\{x^k\}$ dans E, et on veut en extraire une sous-suite $\{x^{\psi(j)}\}$ qui est convergente. On va procéder par extractions successives, pour faire converger la première coordonnée, puis la seconde, et ainsi de suite, avec une extraction diagonale à la fin si $I = \mathbb{N}^*$.

Notons x^k_m la m-ième coordonnée de x^k; donc $x^k_m \in E_m$ qui est compact. Commençons par extraire une première sous-suite $\{x^{\varphi_1(j)}\}$, où donc $\varphi_1 : \mathbb{N} \to \mathbb{N}$ et une application strictement croissante. On utilise BW sur le compact E_1 pour choisir φ_1 telle que la suite extraite $\{x_1^{\varphi_1(j)}\}$ converge.

Puis on recommence: on choisit une sous-suite de $\{x^{\varphi_1(j)}\}$, qu'on notera $\{x^{\varphi_2(j)}\}$ (donc en fait φ_2 est de la forme $\widetilde{\varphi}_2 \circ \varphi_1$), choisie pour que $\{x_2^{\varphi_2(j)}\}$ soit convergente. Noter que $\{x_1^{\varphi_2(j)}\}$, qui est une suite extraite de $\{x_1^{\varphi_1(j)}\}$, reste convergente (faites la vérification si vous n'avez pas vu ceci dans un autre cours).

Et ainsi de suite: pour tout m, on extrait une sous suite $\{x^{\varphi_{m+1}(j)}\}$ de la suite précédente $\{x^{\varphi_m(j)}\}$, de manière que $\{x_{m+1}^{\varphi_{m+1}(j)}\}$ soit convergente. Et on note que les $\{x_p^{\varphi_{m+1}(j)}\}$, $p \leq m$, restent convergentes.

Quand $I = \{1, 2, \dots n\}$ est fini, on s'arrête à $\{x^{\varphi_n(j)}\}$, qui est une sous-suite convergente parce que chacune de ses cordonnées $\{x_p^{\varphi_n(j)}\}$ est convergente (on a dit plus haut qu'une suite converge dans E si et seulement si chacune de ses coordonnées converge).

Quand $I = \mathbb{N}^*$, on a recours au procédé diagonal: on note $\psi(j) = \varphi_j(j)$. Donc $\psi(j)$ est le j-ième élément de la j-ième suite). On remarque que $\psi(j+1) = \varphi_{j+1}(j+1) \geq \varphi_j(j+1) > \varphi_j(j) = \psi(j)$, donc ψ est bien strictement croissante. Et même, pour chaque j, on peut montrer que pour $i \geq j$, $\psi(i) = \varphi_j(h(i))$ pour un certain $h(i)$ qui est une fonction strictement croissante de i (que je ne recalcule pas à partir des $\widetilde{\varphi}_\ell$ pour gagner du temps), de sorte que $\{x^{\psi(i)}\}$, $i \geq j$, est bien une suite extraite de $\{x^{\varphi_j(i)}\}$. Ainsi, la suite $\{x_j^{\psi(i)}\}$, $i \geq j$, est convergente par construction.

En fin de compte, chacune des coordonnées de $\{x^{\psi(i)}\}$ converge, et on vérifie assez facilement à parti de la définition de d donnée en (2.5) que cela implique que $\{x^{\psi(i)}\}$ converge dans le produit E. □

Donnons encore (rapidement) un exemple et un contre-exemple dans un espace vectoriel normé de dimension infinie. On prend le plus simple, à savoir l'espace $\ell^2 = \ell^2(\mathbb{R})$ des suites $\{x_j\}$, $j \geq 0$, à valeurs dans $\mathbb{R}$, et qui sont de carré sommable. Ce qui veut dire que

$$||x||^2 = \sum_{j\geq 0} x_j^2 < +\infty. \tag{2.7}$$

Vous verrez (ou avez vu) que ℓ^2 est un magnfique espace de Hilbert. Sa boule unité fermée $\overline{B}(0,1)$ n'est pas compacte. En effet, considérons pour tout $n \geq 0$ l'élément $e_n \in \overline{B}$ qui est la suite $(0,\ldots,1,0,\ldots)$ avec un 1 juste à la n-ième place et des 0 partout ailleurs. On vérifie aisément que $||e_m - e_n|| = \sqrt{2}$ pour tout $m \neq n$, de sorte qu'aucune suite extraite de $\{e_n\}$, $n \geq 0$, ne peut converger (et donc BW est faux sur $\overline{B}$).

En fait, dans un espace vectoriel normé de dimension infinie, la boule unité fermée n'est jamais compacte, presque pour la même raison (mais ici c'est explicite).

Par contre une grosse astuce consiste à changer de topologie: vous verrez sans doute que muni d'une topologie différente (appelée faible, ou $*$-faible), $\overline{B}$ redevient compact. Cette astuce est très utile.

Et aussi, le fermé plus petit (que je crois qu'on l'appelle boule de Hilbert mais je n'ai pas retrouvé facilement sur internet)

$$K = \left\{x \in \ell^1 \,;\, |x_j| \leq 1/(j+1) \text{ pour tout } j \geq 0\right\} \tag{2.8}$$

est bien compact (quand on le muni de la distance $d(x,y) = ||x-y||$). Ce qui se passe dans ce cas est que si vous arrivez à faire converger chaque coordonnée pour une suite dans K, la suite entière converge, parce que "le reste des coordonnées donne une norme aussi petite qu'on veut".

2.4 Fonctions continues sur un compact

Je devrais sans doute dire "restriction à un compact d'une fonction continue", parce que c'est souvent comme cela que les choses se présentent, mais bien entendu c'est pareil. Deux théorèmes de base, un général et un particulier.

Théorème 2.14. *Soient K un espace (topologique) compact, F un espace topologique, et $f : K \to F$ une application continue. Alors $f(K)$ est compact.*

Comme je n'ai pas dit métrique, c'est qu'il y a une démonstration avec des recouvrements. On se donne un recouvrement de $f(K)$ par des ouverts V_i, $i \in I$. On pose $U_i = f^{-1}(V_i)$. C'est un ouvert, et les U_i recouvrent bien K. Donc on peut trouver un sous-recouvrement fini: il existe $I_0 \subset I$ fini tel que $K \subset \cup_{i\in I_0} U_i$. Alors $f(K) \subset \cup_{i\in I_0} f(U_i)$. Mais $f(U_i) \subset f(K) \cap V_i$ (dans un cas pareil, vérifiez tranquillement: on prend un $y \in f(U_i)$, et il est bien dans V_i

puisque tout point de U_i s'envoie dans V_i). Donc on a trouvé un sous-recouvrement de $f(K)$ par un nombre fini de V_i, et la compacité de $f(K)$ s'en déduit. □

Remarque de plus: c'est pas parce que c'est simple que ça n'est pas utile! Voici la conséquence qu'on utilisera le plus.

Théorème 2.15. *Soient K un compact non vide et $f : K \to \mathbb{R}$ une application continue. Alors f est bornée et atteint ses bornes (définitions ci-dessous).*

Bornée signifie qu'il existe $R > 0$ tel que $|f(x)| \leq R$ pour tout $x \in K$. Dit de manière un peu plus précise, l'ensemble $f(K) = \{f(x)\,;\, x \in K\}$ est borné, ce qui veut dire, puisqu'il n'est pas vide, que sa borne supérieure et sa borne inférieure sont finies. En fait, posons

$$m = \inf_{x\in K} f(x) = \inf\{f(x)\,;\, x \in K\} \tag{2.9}$$

(la borne inférieure dont on vient de parler, qui est bien définie à cause de la précaution "$K \neq \emptyset$") et

$$M = \sup_{x\in K} f(x) = \sup\{f(x)\,;\, x \in K\}. \tag{2.10}$$

La première partie de l'énoncé est que $-\infty < m \leq M < +\infty$, et la seconde partie dit qu'il existe x_m et x_M dans K tels que $f(x_m) = m$ et $f(x_M) = M$.

Démonstration facile. Puisque $f(K)$ est compact par le théorème précédent, c'est un ensemble borné (donc $-\infty < m \leq M < +\infty$), et de plus fermé. Alors je prétends que $f(K)$ continent ses bornes m et M. Par exemple, pour m, par définition de la borne inférieure, pour tout $k \geq 0$ on peut trouver $\ell_k \in f(K)$ tel que $\ell_k \leq m + 2^{-k}$; le fait que $\ell_k \geq m$ est encore plus clair (toujours par définition de m), donc $m \leq \ell_k \leq m + 2^{-k}$. C'est clair que $\{\ell_k\}$ converge vers m, et puisque $f(K)$ est fermé, $m \in f(K)$. Alors il existe $x_m \in K$ tel que $f(x_m) = m$, comme annoncé, et on ferait pareil pour M. □

Remarque: ici j'ai fait le malin pour ne pas avoir à supposer que K est un espace métrique, en utilisant à la place que $\mathbb{R}$ est métrique. Mais si K est métrique, le plus classique pour trouver x_m, par exemple, est de prendre une suite minimisante $\{x_k\}$, c'est à dire une suite de points de K tels que $f(x_k)$ tend vers m, puis d'appliquer BW pour trouver une sous-suite de $\{x_k\}$ qui converge. Alors la limite x_m de cette sous-suite vérifie bien $f(x_m) = m$, parce que f est continue et que la sous-suite converge vers x_m.

Noter que pour $f : K \to \mathbb{R}^n$, on a le droit d'appliquer le théorème à chaque coordonnée de f (mais on trouve des points x_m et x_M qui dépendent de la coordonnée choisie), ou à $||f||$ (pour montrer que f est bornée et $||f|$ atteint ses bornes. Ceci parce que $x \to ||x||$ est une application continue de $\mathbb{R}^n$ dans $\mathbb{R}$ (et en fait pour tout choix de norme $||\,||$ raisonnable).

Et maintenant un corollaire souvent utilisé.

Corollaire 2.16. *Soient K un compact non vide et $f : K \to]0, +\infty[$ une application continue. Alors il existe $\varepsilon_0 > 0$ tel que $f(x) \geq \varepsilon_0$ pour tout $x \in K$.*

C'est facile, puisqu'on peut prendre $\varepsilon_0 = f(x_m)$ avec x_m comme ci-dessus. Mais c'est bon à savoir: si f est strictement positive et continue sur un compact, elle ne peut pas s'approcher trop près de 0.

Encore une conséquence classique: l'équivalence des normes sur $\mathbb{R}^n$. Notons $||\,||$ la norme euclidienne (pour fixer les idées).

Corollaire 2.17. *Pour toute norme N définie sur $\mathbb{R}^n$, il existe une constante $C > 0$ (qui dépend de N) telle que*

$$N(x) \leq C||x|| \quad \text{et } ||x|| \leq CN(x) \quad \text{pour tout } x \in \mathbb{R}^n. \tag{2.11}$$

Démonstration. D'abord, la première inégalité n'est qu'une simple majoration. Notons $(e_1, \dots, e_n)$ la base canonique de $\mathbb{R}^n$, et $x = (x_1, \dots, x_n)$ le point générique de $\mathbb{R}^n$. Alors, pour $x \in \mathbb{R}^n$,

$$N(x) = N(\sum_{i=1}^{n} x_i e_i) \leq \sum_{i=1} N(x_i e_i) \leq \sum_{i=1} |x_i| N(e_i). \tag{2.12}$$

La vie aurait été un peu plus simple si j'avais décidé de comparer N à la norme $\sum_i |x_i|$ (dite norme ℓ_1), mais tant pis, pour ne pas perdre trop dans les inégalités, appliquons Cauchy-Schwartz aux deux vecteurs v, de coordonnées $|x_i|$ et w, de coordonnées $N(e_i)$. On trouve

$$N(x) \leq \sum_{i=1} |x_i| N(e_i) = \langle v, w \rangle \leq ||v||||w|| = ||v|| \Big(\sum_{i=1}^{n} N(e_i) \Big)^{1/2}. \tag{2.13}$$

Le nombre $||w|| = \big(\sum_{i=1}^n N(e_i)\big)^{1/2}$ ne dépend que de N, pas de x, donc on a bien démontré la première partie de (2.11), et on peut prendre $C = ||w||$.

Une conséquence de ceci est que $N(x)$ est une fonction continue de x, puisque pour $x, y \in \mathbb{R}^n$,

$$|N(x) - N(y)| \leq N(x - y) \leq C||x - y|| = C \operatorname{dist}(x, y) \tag{2.14}$$

(où pour changer j'ai noté $\operatorname{dist}(x, y)$ la distance euclidienne).

Pour la seconde inégalité, on va utiliser le corollaire. Soit $S = \partial B(0, 1)$ la sphère unité (pour la norme euclidienne): on veut faire les calculs loin de 0. La fonction $x \to N(x)$, définie sur S est une fonction continue, à valeurs strictement positives (la norme de $x \neq 0$ n'est jamais nulle), donc il existe $\varepsilon_0 > 0$ (qui donc dépend de N) tel que $N(x) \geq \varepsilon_0$ pour tout $x \in S$. Et maintenant, pour tout $x \in \mathbb{R}^n \setminus \{0\}$, on écrit $x = ||x||\,\widetilde{x}$, avec $\widetilde{x} = x/||x||$, et par définition d'une norme $N(x) = N(|x|\,\widetilde{x}) = |x|N(\widetilde{x}) \geq \varepsilon_0 ||x||$. Ceci reste vrai pour $x = 0$, donc $N(x) \geq \varepsilon_0 ||x||$ pour tout $x \in \mathbb{R}^n$, ce qui donne $||x|| \leq \varepsilon^{-1} N(x)$, comme souhaité. □

J'ai dit toute les normes sur $\mathbb{R}^n$ sont équivalentes, et j'ai juste démontré l'équivalence de chacune à $||\,||$. Mais évidemment l'équivalence entre deux normes quelconques s'en déduit: étant donné deux normes N_1 et N_2, on sait que (2.11) a lieu pour N_1 avec une constante C_1 et pour N_2 avec une constante C_1, et on en déduit aussitôt que

$$N_1(x) \leq C_1 C_2 N_2(x) \quad \text{et } N_2(x) \leq C_1 C_2 N_1(x) \quad \text{pour tout } x \in \mathbb{R}^n. \tag{2.15}$$

3 Différentielle de fonctions défnies sur un ouvert de $\mathbb{R}^n$

Donc on se donne une fonction f définie sur un ensemble ouvert U de $\mathbb{R}^n$: pas la peine que f soit définie partout, et un ouvert c'est plus pratique pour ne pas avoir à se préoccuper du bord, au moins dans un premier temps. Et on va étudier f au voisinage d'un point $x \in \mathbb{R}^n$. On pense aux histoires de dérivabilité, etc., mais donc dans $\mathbb{R}^n$ et plus seulement $\mathbb{R}$.

J'essaie de noter $x = (x_1, ..., x_n)$ pour éviter des notations trop lourdes.

On commence avec le cas un peu plus simple où $f : U \to \mathbb{R}$, mais prendre des valeurs dans $\mathbb{C}$ ou dans un autre $\mathbb{R}^m$ ne changerait pas trop la théorie, juste les notations. On en parlera plus tard.

3.1 Dérivées partielles

Vous connaissez probablement. En tout cas c'est la notion la plus simple. On se place au point $z = (z_1, ..., z_n)$. On choisit une coordonnée j, et on note e_j le j-ème vecteur de la base canonique. On dit que la dérivée partielle $\frac{\partial f}{\partial x_j}$ existe au point z (drôle de notation mais x_0 n'est pas pratique parce qu'on a déjà des indices) si la limite

$$\frac{\partial f}{\partial x_j}(z) = \lim_{t \to 0} \frac{f(z + te_j) - f(z)}{t} \tag{3.1}$$

existe (et est finie). Bref on cherche la dérivée en 0 de l'application $t \to f(z+te_j)$. La fonction $t \to f(z + te_j)$ est parfois appelée fonction partielle. Bref, on sait dériver les fonctions d'une variable, et on essaie de s'y ramener.

Autre notation parfois plus pratique: $\partial_j f(z)$ au lieu de $\dfrac{\partial f}{\partial x_j}(z)$.

Définition: f a des dérivées partielles au point z si toutes les $\dfrac{\partial f}{\partial x_j}(z)$ existent.

C'est un peu étrange parce qu'on privilégie les axes alors qu'a priori on ne devrait pas. Mais on peut aussi définir une dérivée directionnelle dans la direction v, où $v \in \mathbb{R}^n$ (on autorise $v = 0$, mais prenez-le non nul pour que ce soit intéressant), par

$$\partial_v f(z) = \lim_{t \to 0} \frac{f(z + tv) - f(z)}{t} \tag{3.2}$$

(si elle existe). Donc le cas particulier de $v = e_j$ donne $\frac{\partial f}{\partial x_j}(z)$. Et le cas où $v = 0$ (je devrais dire $(0, \ldots, 0)$ mais je je ferai rarement) est autorisé, mais trivial: on trouve une fonction constante et une dérivée directionnelle nulle.

Notons au passage que puisque f est définie sur U qui est ouvert, et $z \in U$, il existe $r > 0$ tel que $B(z, r) \subset U$, donc en fait la fonction $t \mapsto f(z+tv)$ est définie au moins sur l'intervalle $]-\delta, \delta[$, où on choisit δ tel que $\delta||v|| \leq r$. C'est le bon cadre pour définir la dérivée ci-dessus.

Pour l'instant c'est simple, on sait calculer les dérivées directionneles dès qu'on sait dériver les fonctions d'une seule variable, mais on va voir que les dérivées partielles ne sont pas tout-à-fait suffisantes pour ce qu'on veut faire.

Remarque dans ce sens: la fonction f peut très bien avoir une dérivée directionnelle dans chaque direction v mais ne pas être continue en z. Evidemment, ça nous change de la dimension 1, et c'est un symptôme qu'il nous manque quelque chose dans la définition.

Exemple au point $z = 0$ et en dimension 2. Prenons $f(0, y) = 0$ et $f(x, y) = \dfrac{y^2}{x}$ pour $x \neq 0$. On calcule $f(tv)$ et on voit que $\partial_v f(0)$ existe et vaut 0 pour $v = (0, v_2)$ et v_2^2/v_1 pour $v = (v_1, v_2)$ avec $v_1 \neq 0$. Mais en calculant $f(t^2, t)$ on se rend compte que la limite en $t = 0$ vaut $1 \neq f(0)$. Et en fait (en prenant $z = (t^3, t)$ on voit que) f n'est même pas bornée au voisinage de 0.

L'exemple n'est pas si compliqué: on s'est débrouillé dans la formule pour que dans une direction fixée, f ne soit pas bien méchante. Mais que quand on change de direction, f se comporte très mal. On peut trouver d'autres exemples comme ça. Ce qui ne va pas dans la définition, c'est le manque d'uniformité, en fonction de la direction v, de la vitesse à laquelle la limite est atteinte. On s'en tirera mieux en demandant, comme dans la définition suivante, un développement limité d'ordre 1. Ou en demandant des dérivées partielles continues. Voir plus bas.

3.2 Définition de la différentielle

Ca sera ça la bonne notion. En plus (mais ce n'est pas notre préoccupation ici), c'est la notion qui permet de passer à f définie sur un espace de Banach (de dimension infinie). On se souvient que pour $f : \mathbb{R} \to \mathbb{R}$, l'existence de $f'(z)$ est une propriété d'approximation (à l'ordre 1) de f par une fonction affine. On va faire pareil pour $f : \mathbb{R}^n \to \mathbb{R}$.

On va noter $\mathcal{L} = \mathcal{L}(\mathbb{R}^n, \mathbb{R})$ l'ensemble des applications linéaires de $\mathbb{R}^n$ dans $\mathbb{R}$ (on dit aussi formes linéaires). Ici, comme on travaille sur un espace de dimension finie, pas besoin de se demander si l'on veut qe L soit bornée (= continue): c'est vrai automatiquement. Ce sont les applications $L : \mathbb{R}^n \to \mathbb{R}$ données par la formule $L(x) = \sum_{j=1}^n \alpha_j x_j$, pour certains coefficients $\alpha_j \in \mathbb{R}$. En fait, $\alpha_j = L(e_j)$, et la formule s'obtient en utilisant le fait que L est une application linéaire, puisqu'alors $L(x) = L\big(\sum_{j=1}^n x_j e_j\big) = \sum_{j=1}^n x_j L(e_j) = \sum_{j=1}^n \alpha_j x_j$.

Définition 3.1. *On dit que $f : U \subset \mathbb{R}^n \to \mathbb{R}$ est différentiable au point $z \in U$ quand il existe une forme linéaire $L \in \mathcal{L}(\mathbb{R}^n, \mathbb{R})$ telle que*

$$\lim_{y \in \mathbb{R}^n,\, y \to 0} \frac{|f(z+y) - f(z) - L(y)|}{||y||} = 0. \tag{3.3}$$

Comme vous n'avez peut-être pas trop l'habitude de faire converger des points $y \in \mathbb{R}^n$, je traduis. Mais c'est juste une histoire de limites dans un espace métrique. Donc (3.3) signifie que pour tout $\varepsilon > 0$, il existe $\delta > 0$ tel que $\frac{|f(z+y)-f(z)-L(y)|}{|y|} < \varepsilon$ pour tout $y \in B(0, \delta) \setminus \{0\} \subset \mathbb{R}^n$. Ou, si vous préférez, que $|f(z + y) - f(z) - L(y)| < \varepsilon ||y||$ pour tout $y \in B(0, \delta) \setminus \{0\} \subset \mathbb{R}^n$.

Si vous aimez la notation de Landau, vous pouvez aussi traduire (3.3) par

$$f(z + y) = f(z) + L(y) + o(||y||) \tag{3.4}$$

Pour moi, $o(||y||)$ est une notation pour désigner n'importe quelle fonction h telle que $|h(y)|/||y||$ tend vers 0 au point considéré; plus généralement, pour une fonction f, $o(f)$ serait n'importe quelle fonction h telle que $|h(y)|/|f(y)|$ tend vers 0, et avec juste un tout petit abus de langage, on écrira aussi $o(y)$ ci-dessus, même si y est à valeurs vectorielles. Autre manière de dire (3.4), en posant $x = z + y$:

$$f(x) = f(z) + L(x - z) + o(x - z) \tag{3.5}$$

quand x tend vers z.

Quand est différentiable au point $z \in U$, on dit que L (ci-dessus) est la différentielle de f au point z, et je noterai volontiers $L = Df(z)$ ou même $L = D_z f$, ce qui laisse de la place pour mettre le vecteur.

Notons tout de suite l'unicité de $Df(z)$ (quand elle existe). Ce serait assez facile à vérifier tout de suite (on a deux développements limités en z, et on peut les comparer), mais j'attends la prochaine proposition pour que ce soit encore plus simple.

Quand $n = 1$, c'est en fait pareil que la dérivabilité de $f : \mathbb{R} \to \mathbb{R}$, où l'on utilise l'application linéaire $L : \mathbb{R} \to \mathbb{R}$ donnée par $L(y) = f'(z)y$.

Donc la différentiabilité est une propriété d'approximation par une fonction affine. Une fonction affine, c'est une fonction de la forme $a(x) = \alpha_0 + L(x) = \alpha_0 + \sum_{j=1}^n \alpha_j x_j$, avec $L \in \mathcal{L}$.

On en reparlera, mais pour une fonction $f : \mathbb{R}^n \to \mathbb{R}^m$ (ou $\mathbb{C}$ si vous voulez), la définition est la même, sauf que maintenant $L \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ est une application linéaire de $\mathbb{R}^n$ dans $\mathbb{R}^m$. En fait, on peut aussi procéder coordonnée par coordonnée, au sens où f est différentiable en z si et seulement si chacune de ses coordonnées f_k est différentiable; l'application linéaire $L \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ est juste donnée par ses m coordonnées $L_1, ...L_m \in \mathcal{L}(\mathbb{R}^n, \mathbb{R})$, chacune est une forme linéaire, et est la différentielle du f_k correspondant. Vérification assez facile de tous ces détails.

Proposition 3.2. *Si f est différentiable en z, f a des dérivées partielles en z, et même a une dérivée directionnelle $\partial_v f(z)$ pour tout vecteur $v \in \mathbb{R}^n$, qui est donnée par la formule*

$$\partial_v f(z) = Df(z)(v).$$

Enfin, f est continue au point z.

Donc en particulier, les dérivées partielles existent et sont données par $\frac{\partial f}{\partial x_j}(z) = L(e_j) = \alpha_j$ (où comme plus haut $L(y) = \sum \alpha_j y_j$). Notons aussi que du coup $v \to \partial_v f(z)$ est linéaire.

Ainsi, dans le sens "différentiable" implique "existence de dérivées partielles" c'est bon. Et on a vu un exemple où les $v \to \partial_v f(z)$ existent, et l'application $v \to \partial_v f(z)$ n'est linéaire. Donc, f n'est pas différentiable en z. D'ailleurs, cet f n'est même pas continu en z.

Démonstration de la proposition. On se donne $v \in \mathbb{R}^n$, on pose $h(t) = f(z + tv)$, et on doit prouver que h est dérivable en 0, avec $h'(0) = L(v)$. Mais quand on prend $y = tv$ dans (3.3), on voit que

$$\frac{h(t) - h(0) - tL(v)}{t} = \frac{f(z + tv) - f(z) - L(tv)}{t} = \frac{f(z + tv) - f(z) - L(tv)}{||tv||} \frac{||tv||}{t},$$

avec un premier facteur qui tend vers 0 quand t tend vers 0 (par (3.3) et parce qu'alors tv tend vers 0), et un second facteur qui est borné (par $||v||$). On en bien déduit que $h'(0) = L(v)$. □

Démonstration de l'unicité de la différentielle $L = Df(z)$. Maintenant si f admet à la fois la différentielle L_1 et L_2 en z, la proposition dit que pour tout v et h comme ci-dessus, $h'(0) = L_1(v) = L_2(v)$, donc $L_1(v) = L_2(v)$ pour tout v, et $L_1 = L_2$ comme annoncé. On s'est juste ramené à l'unicité de la dérivée d'une fonction d'une variable, qu'on suppose connue, mais qui est facile à vérifier: cette fois on peut utiliser la définition usuelle de la dérivée et se ramener à l'unicité d'une limite dans $\mathbb{R}$. □

3.3 Gradient

La définition du gradient de f au point z est la suivante:

$$\nabla f(z) = \Big(\frac{\partial f}{\partial x_1}(z), \dots \frac{\partial f}{\partial x_n}(z)\Big), \tag{3.6}$$

qui est donc défini dès que les dérivées partielles au point z existent, et donc en particulier dès que f est différentiable en z.

Ceci dit, on ne pourra vraiment bien utiliser le gradient que quand on saura que f est différentiable en z.

J'ai noté $\nabla f(z)$ comme un vecteur-ligne, pour gagner de la place et parce que j'ai du mal avec les vecteurs-colonne en lateχ, mais c'est un abus de notation: quand on fera des vrais calculs matriciels, on se souviendra qu'en fait $\nabla f(z)$ est plutôt un vecteur colonne, ou en tout cas il faudra faire attention aux notations.

J'ai attendu un peu pour insister sur le fait que maintenant on va utiliser le produit scalaire. Je veux dire, pour l'instant on n'utilisait que la structure linéaire de $\mathbb{R}^n$ (c'est un espace vectoriel), et maintenant on va donc faire un peu d'algèbre hilbertienne élémentaire. Et c'est bien ce qu'on fera quand on utilisera le gradient.

Soit $L \in \mathcal{L} = \mathcal{L}(\mathbb{R}^n, \mathbb{R})$ (une forme linéaire). Si on était en dimension infinie (sur un espace de Hilbert), on demanderait continue, mais ici ce sera automatique.

Fin du cours 3, 2023

Ainsi $L(y) = \sum \alpha_j y_j$, avec $\alpha_j = L(e_j)$ comme plus haut. Et en notant $\alpha \in \mathbb{R}^n$ le vecteur (colonne) $(\alpha_1, ...\alpha_n)$ (y aussi est un vecteur colonne), on peut aussi écrire

$$L(y) = \langle y, \alpha \rangle = y \cdot \alpha = \langle y|\alpha \rangle \tag{3.7}$$

avec des notations différentes pour le produit scalaire (je crois que je choisirai souvent la première). C'est aussi le produit de la matrice transposée de y par la matrice de α; j'utiliserai cette dernière notation matricielle le mins posible, au moins jusqu'au moment où on composera et où on fera des formes quadratiques. Ce serait aussi une notation pratique (en faisant bien attention) si on considérait des fonctions $f : \mathbb{R}^n \to \mathbb{R}^m$.

Ici c'est simple et j'ai donné la formule directement, mais plus généralement le théorème de Riesz dit que dans un espace de Hilbert H, toute forme linéaire continue L (c.-à-d., toute

application linéaire continue de H dans $\mathbb{R}$ (ou $\mathbb{C}$ pour les Hilbert complexes)) est donnée par la formule (3.7) pour un vecteur $\alpha \in H$. Du coup, grâce à (3.7), on peut calculer la norme de L.

Proposition 3.3. *La norme de L (donnée par $L(y) = \sum \alpha_j y_j$) est*

$$|||L||| := \sup\big\{||L(y)|| \,;\, y \in \mathbb{R}^n \text{ avec } ||y|| \le 1\big\} \\ = \sup\big\{||L(y)|| \,;\, y \in \mathbb{R}^n \text{ avec } ||y|| = 1\big\} = ||\alpha||. \tag{3.8}$$

Evidemment, ici $||\cdot||$ est la norme euclidienne (et dans H Hilbert ce serait la norme).

Démonstration facile. Je le fais dans H Hilbert, avec $L(y) = \langle \alpha, y\rangle$, mais pensez à $H = \mathbb{R}^n$. La première partie est juste la définition de la norme d'une application linéaire continue (encore appelée opérateur borné).

Notons N le second sup; alors $N \le |||L|||$ puisquele sup porte sur moins de termes (la sphère est contenue dans la boule), mais aussi $|||L||| \le N$ parce que pour y tel que $||y|| \le 1$, soit $y = 0$ et alors $||L(y)|| = 0$, soit $y \neq 0$, et alors en posant $y' = y/||y||$, on a $L(y) = ||y||L(y') \le ||y||N \le N$ puisque y' est dans la sphère et $||y|| \le 1$.

Il reste à montrer que $|||L||| = ||\alpha||$. Notez au passage que j'ai écrit $||L(y)||$, mais ici puisque $L(y) \in \mathbb{R}$ (ou éventuellement $\mathbb{C}$), j'aurais plutôt du noter $|L(y)|$.

Maintenant l'inégalité principale $|L(y)| = |\langle \alpha, y\rangle| \le ||\alpha||||y||$ vient de Cauchy-Schwarz, et en prenant le sup on trouve $|||L||| \le ||\alpha||$. Enfin dans l'autre sens, soit $\alpha = 0$ et alors $|||L||| = 0$, soit $\alpha \neq 0$ et on essaie $y = \alpha/||\alpha||$, qui est dans la sphère unité, et on trouve que

$$|||L||| \ge |L(y)| = ||\alpha||^{-1}|L(\alpha)| = ||\alpha||^{-1}|\langle \alpha, \alpha\rangle| = ||\alpha||^{-1}||\alpha||^2 = ||\alpha||,$$

comme annoncé. □

Noter que dans le cas d'une application linéaire $L \in \mathcal{L}(\mathbb{R}^n, \mathbb{R})$, on savait déjà que L est bornée, mais maintenant notre estimation est plus précise puisqu'on a calculé sa norme. Et L est continue, et même $||\alpha||$-Lipschitzienne, puisque $|L(y) - L(w)| = |L(y - w)| \le |||L||||y - w| = ||\alpha|||y - w| = ||\alpha|| \operatorname{dist}(y, w)$.

Quand L est la différentielle de f au point z, la discussion ci-dessus donne

$$Df(z)(v) = \langle \nabla f(z), v\rangle \quad \text{pour tout } v \in \mathbb{R}^n,$$

donc (3.3) s'écrit aussi

$$\lim_{y \in \mathbb{R}^n,\, y \to 0} \frac{|f(z + y) - f(z) - \langle \nabla f(z), y\rangle|}{||y||} = 0. \tag{3.9}$$

3.4 Fonctions différentiables

On garde $f = U \to \mathbb{R}$, où U est un ouvert de $\mathbb{R}^n$. On a déjà défini la différentiabilité en un point $z \in U$.

On dit que f est <u>différentiable sur U</u> quand elle est différentiable en chaque point de U.

On dit que f est <u>de classe C^1</u> quand de plus la différentielle (ou si vous voulez $\nabla f(z)$) est une fonction continue sur U). Et on verrra dans un instant que l'existence sur U de dérivées partielles continues implique la différentiabilité partout.

Pour la différentiabilité d'ordre supérieur: on re-dérivera chacune des coordonnées du gradient, et ainsi de suite. Venons-en au critère qui nous permettra de prouver la différentiabilité de f sans effort la plupart du temps.

Théorème 3.4. *Soit f définie sur l'ouvert U. Si f a des dérivées partielles continues sur U, alors elle est différentiable (et donc de classe C^1) sur U.*
Et même, si f a des dérivées partielles sur un voisinage de $z \in U$, qui sont continues au point z, alors f est différentiable au point z.

C'est important. Je n'ai pas précisé "à valeurs dans $\mathbb{R}$" parce que le même théorème vaut pour $f : U \to \mathbb{R}^m$, en appliquant le théorème à chaque fonction coordonnée de f. Donc je me suis plaint très fort des dérivées partielles (prises toutes seules), mais juste en ajoutant leur continuité, c'est bon je suis content. Je veux dire, rapport au contre-exemple cité plus haut, on a juste ajouté la continuté des $\frac{\partial f}{\partial x_j}(z)$ et maintenant ça marche. C'est un prix bien faible à payer. Et bien sûr dès qu'on sait que f est différentiable, on a vu que la différentielle est donnée par les dérivées partielles, donc dès que celles-ci sont continues, f est bien de classe C^1.

Démonstration. On se donne $z \in U$, on définit L par $L(y) = \sum_j \frac{\partial f}{\partial x_j}(z) y_j$, et on doit vérifier que L est la différentielle de f en z (donc qu'on a par exemple (3.4)).

On va appliquer le théorème des accroissements finis n fois, en utilisant les fonctions partielles, puis il restera à estimer un peu les erreurs. On se donne aussi $y = (y_1, y_2, \dots, y_n) = y_1 e_1 + \dots + y_n e_n$, et on va calculer $f(z+y) - f(z)$.

Ce sera pratique d'utiliser les points intermédiaires $z_j = z + \sum_{0 \le i \le j} y_i e_j$. Noter que $z_0 = z$ et $z_n = z + y$. On écrit

$$f(z+y) - f(z) = \sum_{j=1}^{n} \Delta_j, \tag{3.10}$$

où

$$\Delta_j = f(z_j) - f(z_{j-1}). \tag{3.11}$$

Ensuite on calcule chaque Δ_j en utilisant le fait que $z_j = z_{j-1} + y_j e_j$, donc on ne fait bouger qu'une seule variable à la fois. On applique le théorème des accroissements finis à la fonction d'une variable $h_j : t \to f(y_{j-1} + t y_j e_j)$. Ainsi $\Delta_j = h_j(1) - h_j(0)$, la fonction h est dérivable sur $[0,1]$, sa dérivée est $h_j'(t) = y_j \partial_j f(y_{j-1} + t y_j e_j)$, et donc le théorème des accroissements finis donne

$$\Delta_j = h_j(1) - h_j(0) = h_j'(\xi_j) = y_j \partial_j f(y_{j-1} + \xi_j y_j e_j) \tag{3.12}$$

pour un certain $\xi_j \in [0,1]$. Comme on veut utiliser la continuité des $\partial_j f$, on écrit encore

$$\Delta_j = y_j \partial_j f(z) + \varepsilon_j y_j, \quad \text{avec } \varepsilon_j = \partial_j f(w_j) - \partial_j f(z), \text{ où } w_j = y_{j-1} + \xi_j y_j e_j \tag{3.13}$$

On regroupe, et on trouve

$$f(z+y) - f(z) = \sum_{j=1}^{n} \Delta_j = \sum_{j=1}^{n} y_j \partial_j f(z) + \sum_{j=1}^{n} \varepsilon_j y_j = L(y) + \sum_{j=1}^{n} \varepsilon_j y_j, \tag{3.14}$$

où $L(y) = \sum_{j=1}^{n} y_j \partial_j f(z)$ est l'application linéaire associée aux dérivées partielles de f en z. Il ne reste plus qu'à estimer

$$\sum_{j=1}^{n} \varepsilon_j y_j \leq ||y|| \sum_j |\varepsilon_j|. \tag{3.15}$$

Ensuite, quand y tend vers 0, tous les arguments $w_j = z_{j-1} + \xi_j y_j e_j = z + \sum_{i<j} y_i e_i + \xi_j y_j e_j$ tendent vers z, donc tous les $\varepsilon_j = \varepsilon_j(y)$ tendent vers 0 (puisque les $\partial_j f$ sont continues en z. Donc (3.14) donne bien le dévelopement limité souhaité de $f(z+y)$, comme en (3.4). □

On se souviendra bien du théorème. Le plus souvent, pour vérifier que f est différentiable sur U, on commencera par calculer les dérivées partielles de f (en général, c'est facile), et il ne restera plus qu'à vérifier qu'elles sont continues. Le retour à la définition de $L = Df(z)$ sera reservé à des cas rares.

Fin du cours 3, 2022

3.5 Graphes et hyperplans tangents

Juste un court paragraphe pour dire que la différentielle de f en z permet (de démonter l'existence et) de calculer le plan tangent au graphe de f au point $(z, f(z))$.

On va d'abord se placer dans les circonstances suivantes. On se donne $U \subset \mathbb{R}^n$ un ouvert, et $f : U \to \mathbb{R}^m$ une fonction. Pensez très fort au cas où $m = 1$, mais autant faire le cas général. On note $G = G_f$ le graphe de f. Donc

$$G_f = \big\{(x,y) \in \mathbb{R}^n \times \mathbb{R}^m \,;\, x \in U \text{ et } y = f(x)\big\} \subset \mathbb{R}^n \times \mathbb{R}^m \simeq \mathbb{R}^{n+m}. \tag{3.16}$$

Proposition 3.5. *Supposons que $f : U \to \mathbb{R}$ est différentiable en $x_0 \in U$, notons $L = Df(x_0)$ sa différentielle en x_0, et $G \subset \mathbb{R}^{n+1}$ le graphe de f. Soit encore P le plan de dimension n qui passe par $M_0 = (x_0, f(x_0))$ et d'équation $y = f(x_0) + L(x - x_0)$. Alors P est tangent à G en $M_0 = (x_0, f(x_0))$, au sens où*

$$\lim_{M \in G_f \,;\, M \to M_0} \frac{\operatorname{dist}(M, P)}{||M - M_0||} = 0. \tag{3.17}$$

Explications et démonstration.

Avant de commencer, deux remarques.

Dans le cas où f est juste à valeurs réelles, P est bien un hyperplan.

Et j'espère que (3.17) vous semble une définition raisonnable d'un plan tangent: on demande que $\mathrm{dist}(M,P) = o(\mathrm{dist}(M,M_0))$. Et toutes nos distances sont euclidiennes (mais passer à une autre distance équivalente ne changerait rien à l'affaire).

Revenons à la proposition. D'abord ré-écrivons P:

$$P = \big\{M = (x,y)\,;\, x \in \mathbb{R}^n \text{ et } y = y_0 + L(x - x_0)\big\}, \tag{3.18}$$

où j'ai noté $y_0 = f(x_0)$ pour gagner un peu de place. Vérifions que c'est le plan de dimension n qui passe par M_0 et dont le plan vectoriel associé est le n-plan engendré par les vecteurs $v_j = (e_j, Df_{x_0}(e_j))$, où $j = 1, 2, \dots, n$ et $(e_1, \dots, e_n)$ est la base canonique de $\mathbb{R}^n$.

D'abord, les e_j sont indépendants, donc les v_j aussi sont indépendants (regarder la projection sur $\mathbb{R}^n$ d'une combinaison linéaire des v_j; c'est la combinaison linéaire correspondante des e_j, donc c'est nul si et seulement si les coefficients sont nuls. Donc le plan vectoriel P' engendré par les v_j est bien de dimension n, et il reste à voir que $P = M_0 + P'$.

Soit $v \in P'$; on écrit $v = \sum_j \lambda_j v_j$, sa première coordonnée est $\sum_j \lambda_j e_j$, et sa seconde coordonnée (dans $\mathbb{R}^m$) est $\sum_j \lambda_j Df_{x_0}(e_j) = \sum_j \lambda_j L(e_j) = L(\sum_j \lambda_j e_j)$. Donc v vérifie l'équation $y = Lx$ et $M_0 + v$ vérifie l'équation $y = y_0 + L(x - x_0)$. En bref, $M_0 + P' \subset P$. Réciproquement, soit $(x,y) \in P$; donc $y = y_0 + L(x - x_0)$ et $(x', y') = (x - x_0, y - y_0)$ vérifie l'équation $y' = Lx'$. On écrit $x' \in \mathbb{R}^n$ dans la base $(e_1, \dots, e_n)$; on trouve $x' = \sum_j \lambda_j e_j$, et l'équation $y' = Lx'$ s'écrit maintenant $y' = \sum_j \lambda_j L(e_j)$. Donc $(x', y') = \sum_j \lambda_j v_j$, et $(x', y') \in P'$. Ainsi $P - M_0 \subset P'$, et ceci termine notre vérification.

Pour démontrer (3.17), notons $a(x) = f(x_0) + L(x - x_0) = y_0 + L(x - x_0)$; c'est l'approximation affine de f associée au développement limité de (3.5) (avec $z = x_0$); on sait donc que $||f(x) - a(x)|| = o(||x - x_0||)$ quand $x \in \mathbb{R}^n$ tend vers x_0.

Maintenant on prend un point $M \in G_f$, qu'on peut donc écrire $M = (x, f(x))$ pour un certain $x \in U$. Dire que M tend vers M_0, ca veut dire que $||M - M_0||$ tend vers 0, ou de manière équivalente que ses deux parties $x - x_0$ et $y - y_0$ (où $y_0 = f(x_0)$) tendent vers 0. En particulier donc, x tend vers x_0. Et maintenant il ne reste plus qu'à dire que puisque le point $N = N(M) = (x, a(x))$ est dans P,

$$\mathrm{dist}(M,P) \le \mathrm{dist}(M,N) = ||(x, f(x)) - (x, a(x))|| = ||f(x)) - a(x)|| = o(||x - x_0||) \tag{3.19}$$

puisque $M = (x, f(x))$ et en calculant. Mais $||M - M_0|| \ge ||x - x_0||$ (regardez juste les premières coordonnées), donc $o(||x - x_0||) = o(||M - M_0||)$, ou dit autrement $\frac{\mathrm{dist}(M,P)}{||M - M_0||} = \frac{\mathrm{dist}(M,P)}{||x - x_0||} \frac{||x - x_0||}{||M - M_0||}$ tend vers 0 et on en déduit bien (3.17). □

Encore deux mots sur le cas où f est juste à valeurs réelles, et donc P est un hyperplan. Alors l'équation de P peut aussi être écrite

$$y = f(x_0) + \langle x - x_0, \nabla f(x_0)\rangle, \tag{3.20}$$

puisque $Df_{x_0}(x-x_0) = \langle x-x_0, \nabla f(x_0)\rangle$.

On peut aussi définir P par le fait qu'il passe par $M_0 = (x_0, f(x_0))$ et par un vecteur normal $\vec{n}_0$ (orthogonal à P), puisque P est de codimension 1. Quand $n = 1$, P est engendré par le vecteur $(1, f'(x_0))$, et est orthogonal au vecteur $\vec{n} = (-f'(x_0), 1)$ (faites le calcul!). Si on veut un vecteur $\vec{n}_0$ de norme 1, prendre par exemple $\vec{n}_0 = (1 + f'(x_0)^2)^{-1/2}\,\vec{n}$. En dimension n générale, on doit trouver un vecteur de $\mathbb{R}^{n+1}$ orthogonal à chaque $(e_j, \frac{\partial f}{\partial x_j})$, et le vecteur

$$\vec{n} = \vec{n}(x_0) = (-\frac{\partial f}{\partial x_1}(x_0), \dots, -\frac{\partial f}{\partial x_n}(x_0), 1) = (-\nabla f(x_0), 1) \in \mathbb{R}^{n+1} \tag{3.21}$$

marche. Ensuite on peut prendre $\vec{n}_0 = \vec{n}/||\vec{n}|| = (1 + ||\nabla f||^2)^{-1/2}\vec{n}$ si on veut un vecteur normé.

On parlera plus tard de l'utilisation de Df et ∇f pour décrire les ensembles de niveaux d'une fonction f. Ce sera un peu du même genre. Ici, une équation du graphe est $F(x, y) = 0$, où $F(x, y) = y - f(x)$; le gradient de la fonction F, vue comme une fonction de $\mathbb{R}^n \times \mathbb{R} \simeq \mathbb{R}^{n+1}$, est en fait $-\vec{n}$ (au point M_0); on ne sera donc pas étonné plus tard que la tangente à l'ensemble d'équation $F(x, y) = 0$ soit orthogonale au gradient de F.

Enfin disons deux mots des courbes paramétrées [sans doute pas fait en cours]. On se donne $g : I \to \mathbb{R}^n$ une fonction définie sur un intervalle, disons ouvert, I. Ou on peut voir g comme le paramétrage d'une courbe, dont on note $\Gamma = g(I) \subset \mathbb{R}^n$ l'image.

Supposons carrément g dérivable (ou de classe C^1), ce ne sera pas ça la question. On s'attend à pouvoir dire que Γ a, au point $M_0 = g(x_0)$, une droite tangente dont la direction est $g'(x_0) \in \mathbb{R}^n$. Et c'est presque vrai. Ce qui est vrai c'est que quand x tend vers x_0, $g(x) = g(x_0) + (x - x_0)g'(x_0) + o(|x - x_0|)$.

Pour en déduire une histoire de droite tangente, il est raisonable de demander en plus que $g'(x_0) \neq 0$. Ceci sert à deux choses: premièrement, cela permet de définir la droite L qui passe par $g(x_0)$ et dont la direction est $g'(x_0)$. Mais aussi de prouver l'équivalent de (3.17), à savoir que

$$\lim_{x \to x_0} \frac{\text{dist}(g(x), L)}{||g(x) - g(x_0)||} = 0. \tag{3.22}$$

Je vous laisse faire, mais noter que grâce à notre hypothèse, on trouve que $\text{dist}(g(x), g(x_0)) = ||g(x) - g(x_0)|| \sim |x - x_0||g'(x_0)|$, ce qui permet bien de passer de notre information initiale, le fait que $\text{dist}(g(x), L) \leq \text{dist}(g(x), g(x_0) + (x - x_0)g'(x_0)) = o(|x - x_0|)$ (parce que $g(x_0) + (x - x_0)g'(x_0) \in L$), à ce qu'on veut. Donc dans notre histoire de graphes ci-dessus, c'était bien de pouvoir utiliser le fait que $||M - M_0|| \geq ||x - x_0||$.

Ensuite, il n'est pas exclu que la courbe Γ repasse une seconde fois par le même point $M_0 = g(x_0)$, avec une direction de tangente différente, et dans ce cas il faut discuter de ce qu'on entend par une droite tangente à Γ au point M_0.

3.6 Dérivées d'ordre 2, relation de Schwarz, matrice hessienne

Soient $U \subset \mathbb{R}^n$ un ouvert et $f : U \to \mathbb{R}$. On dit que f est $\underline{k \text{ fois différentiable en } z \in U}$ quand elle est $k-1$ fois différentiable dans un voisinage de z et quand toutes ses dérivées partielles d'ordre $k-1$ (ou si vous préférez sa différentielle d'ordre $k-1$, mais il faudrait formaliser) sont différentiables en z.

Et on dira que $\underline{f \text{ est de classe } C^k \text{ dans } U}$ lorsque toutes ses dérivées partielles d'ordre $k-1$ sont de classe C^1, et on se contentera souvent de ce cas pour éviter toute complication. L'avantage de C^k, c'est qu'on peut utiliser le théorème 3.4 ci-dessus pour prouver l'existence de la différentielle de chaque dérivée partielle d'ordre $k-1$ en calculant juste ses dérivées partielles et en vérifiant leur continuité.

On a pris f à valeurs dans $\mathbb{R}$ mais si elle est à valeurs dans $\mathbb{C}$ ou $\mathbb{R}^m$ on peut étudier chaque coordonnée séparément!

Ici on va regarder ce qui se passe avec $k = 2$, bien ranger les dérivées d'ordre 2 dans une matrice, et surtout démontrer la relation suivante très utile:

Théorème 3.6 (Schwartz). *Si f est de classe C^2 sur U, alors $\partial_i(\partial_j(f)) = \partial_j(\partial_i(f))$ sur U. Du coup on peut noter cette seconde dérivée $\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_i \partial x_j}$ ou $\partial^2_{ij} f$ sans risque de se tromper dans l'ordre des dérivations.*

Noter qu'on peut toujours appliquer ceci à des plus petits domaines V; c'est utile si on veut démontrer la formule seulement en z et qu'on sait seulement que f est de classe C^2 dans un petit voisinage V de z.

En fait on a écrit des hypothèses un peu trop fortes: il suffit que les deux dérivées partielles $\partial_i f$ et $\partial_j f$ existent sur U, puis que $\partial_j(\partial_i f)$ et $\partial_i(\partial_j f)$ existent toutes les deux dans U, puis que ces deux dérivées partielles d'ordre 2 soient toutes les deux continues en z. Et alors ces deux dérivées partielles d'ordre 2 sont égales au point z. C'est ce que donnera la démonstration. Mais autant ne pas s'embrouiller avec des énoncés trop subtils pour la pratique courante.

Prenez cet énoncé, avec l'hypothèse de classe C^2 comme une légère mise en garde: comme il sera souvent très difficile de faire des calculs sur les dérivées secondes sans avoir la relation de Schwarz, on conseille vivement de vérifier que f est de classe C^2 avant de commencer (presque toujours, c'est aussi simple).

Démonstration. On va faire écrire une expression de deux manières différentes, en appliquant le théorème des accroissements finis deux fois dans un ordre différent, puis passer à la limite.

Donnons nous $z \in U$ et deux numéros de variables i et j (avec $i \neq j$; si $i = j$ la formule est claire!), et calculons

$$\Delta = \Delta(s,t) = f(z + se_i + te_j) - f(z + se_i) - f(z + te_j) + f(z),$$

où comme d'habitude e_i, e_j sont les éléments de la base canonique et on se restreint à s et t petits, comme ça on reste dans U et tout est défini. C'est bien symétrique (faire un dessin avec 4 points d'un rectangle), et on verra plus tard qu'on pourrait faire un développement

limité de Δ (en appliquant la formule de Taylor) dont le terme principal serait justement avec la dérivée croisée; mais ignorons ceci et calculons Δ de deux façons différentes. En gros en regroupant les termes par paquets de 2 des deux manières logiques possibles.

D'abord, écrivons

$$\Delta = h_s(t) - h_s(0), \quad \text{avec } h_s(t) = f(z + se_i + te_j) - f(z + te_j). \tag{3.23}$$

On applique le théorème des accroissement finis à la fonction h_s, entre 0 et t, et on trouve

$$\Delta = th_s'(\xi) = t\partial_j f(z + se_i + \xi e_j) - t\partial_j f(z + \xi e_j), \tag{3.24}$$

pour un certain $\xi \in]0, t[$. La notation $\partial_j f$ pour $\frac{\partial f}{\partial x_j}$ fait gagner en place et en clarté. Pour ξ donné (même si on ne sait pas qui c'est!), on peut maintenant appliquer le théorème des accroissements finis à Δ, vu avec ξ fixe, mais comme fonction de s. La dérivée de cette fonction est $\partial_i(\partial_j f(z + se_i + \xi e_j))$, et donc on trouve qu'il existe $\eta \in [0, s]$ tel que

$$\Delta = ts\partial_i\partial_j f(z + \eta e_i + \xi e_j), \tag{3.25}$$

où on se souvient de faire attention à l'ordre dans lequel on a dérivé f deux fois dans la direction j puis i.

Bon, maintenant il y a un autre calcul, qui consiste en fait à échanger i et j et l'ordre dans lequel on calcule les différences. Je veux dire, on peut commencer par écrire (à la place de (3.23)) que

$$\Delta = g_t(s) - g_t(0), \quad \text{avec } g_t(s) = f(z + se_i + te_j) - f(z + se_i). \tag{3.26}$$

Puis on applique le théorème des accroissements finis à la fonction g_t sur $[0, s]$, ce qui donne un $\widetilde{\eta} \in [0, s]$ tel que

$$\Delta = sg_t'(\widetilde{\eta}) = s\partial_i f(z + \widetilde{\eta} e_i + te_j) - s\partial_i f(z + \widetilde{\eta} e_j), \tag{3.27}$$

puis on ré-applique le théorème des accroissements finis à cette fonction de t et on trouve $\widetilde{\xi} \in [0, t]$ tel que (comme pour (3.25))

$$\Delta = st\partial_j\partial_i f(z + \widetilde{\eta} e_i + \widetilde{\xi} e_j). \tag{3.28}$$

On a nos deux écritures, et maintenant on les compare. En fait, on pouvait prendre $s = t$, mais continuons encore avec s et t quelconques mais petits (et non nuls) tous les deux. On compare (3.25) et (3.28), on divise par st, et on trouve que

$$\partial_i\partial_j f(z + \eta e_i + \xi e_j) = \partial_j\partial_i f(z + \widetilde{\eta} e_i + \widetilde{\xi} e_j). \tag{3.29}$$

Ca a l'air dangereux à cause des quatre points "inconnus" $\xi, \eta, \widetilde{\xi}, \widetilde{\eta}$, mais quand s et t tendent tous les deux vers 0, ces quatre points tendent tous vers z. Donc on prend en fait n'importe quelle suite de couples (s, t), avec $s \neq 0$, $t \neq 0$, tels que s et t tendent vers 0 (évidemment,

prendre $s = t = 2^{-k}$ est le plus simple). Le membre de gauche tend vers $\partial_i \partial_j f(z)$ et le membre de droite vers $\partial_j \partial_i f(z)$. Donc ces nombres sont égaux. □

La démonstration avait un coté un peu inquiétant (qui sont ces 4 points?). Les choses seraient plus stables, mais plus longues (et avec en plus des hypothèses un peu plus fortes), en écrivant que chaque fonction ci-dessus est l'intégrale de se dérivée.

On continue les définitions en supposant que f est de classe C^2 (comme cela on peut appliquer l'identité de Schwarz sur l'égalité des deux dérivées croisées).

La matrice Hessienne au point z, que je noterai $H_f(z)$, ou $Hess_f(z)$, ou même par abus de langage $D^2 f(z)$, est la matrice carrée dont le terme général est

$$H_{i,j} = \frac{\partial^2 f}{\partial x_i \partial x_j}.$$

C'est donc une matrice symétrique (est Schwarz est bien pratique pour ne pas avoir à distinguer lignes et colonnes).

Ca tombe bien, puisqu'on sait que par algèbre linéaire, toute matrice symétrique est diagonalisable, et même sur une base orthonormée. On s'en servira notamment quand on utilisera les dérivées seconde pour savoir si f a un minimum ou un maximul (ou aucun des deux) en un point critique.

Fin du cours 4, 2022.

3.7 Formules de Taylor

On va se contenter de calculs aux ordres 1 et 2. On va tricher affreusement, en appliquant la formule de Taylor en une variable à la fonction $t \to f(x + th)$, où $x \in U \subset \mathbb{R}^n$ et $h \in \mathbb{R}^n$ est un vecteur.

a- Rappel de ce qu'on a en dimension 1.

On commence par l'ordre 1. On suppose systématiquement que f est de classe C^1 sur un intervalle qui contient x et $x + t$. Je dis ceci ainsi parce que je ne veux pas supposer que $t > 0$, même si souvent on démontre les formules dans ce cas, puis on dit que c'est pareil pour $t < 0$.

Formule des accroissements finis. On a déjà utilisé, mais rappelons qu'elle dit qu'il existe ξ compris entre x et $x + t$, tel que

$$f(x + t) = f(x) + tf'(\xi). \tag{3.30}$$

En fait, juste besoin de la dérivabilité entre x et $x+t$ et de la continuité de f en x et en $x+t$. Et ceci se démontre à partir du théorème de Rolle, qui dit que dans le cas où $f(x) = f(x+t)$, f' s'annule entre x et $x + t$. Le cas général s'obtient en retirant une fonction affine.

Conséquence (en supposant aussi que f' est continue en x):

$$f(x + t) = f(x) + tf'(x) + o(|t|) \quad \text{(quand } t \text{ tend vers } 0). \tag{3.31}$$

Et la formule de primitivation:

$$f(x+t) = f(x) + \int_0^t f'(x+u)du. \tag{3.32}$$

J'ai tendance à aimer celle-ci parce qu'elle est exacte donc on a l'impression qu'on ne pert pas d'info. Et le ξ mystérieux est toujours un peu stressant. Mais (3.30) est souvent juste plus simple!

Passons maintenant à la formule de Taylor à l'ordre 2. Pour nous simplifier un peu la vie, supposons que f est de classe C^2 sur l'intervalle borné par x et $x+t$. Alors

1. (Taylor-Lagrange): il existe ξ compris entre x et $x+t$ tel que

$$f(x+t) = f(x) + tf'(x) + \frac{t^2}{2}f''(\xi). \tag{3.33}$$

Si ma mémoire est bonne, se démontre avec une petite astuce en appliquant le théorème des accroissements finis à la bonne fonction auxiliaire.

2. (Taylor-Young): à x fixé et quand t tend vers 0,

$$f(x+t) = f(x) + tf'(x) + \frac{t^2}{2}f''(x) + o(t^2). \tag{3.34}$$

On suppose que f'' est continue en x et on déduit ceci de (3.33).

3. (Taylor avec reste intégral). Cette fois je prends $t > 0$ parce que j'ai peur de me tromper dans les signes:

$$f(x+t) = f(x) + tf'(x) + \int_0^t f''(x+u)(t-u)du. \tag{3.35}$$

On n'utilisera sans doute pas la fomule correspondante pour $t < 0$, mais se souvenir qu'on peut appliquer (3.35) à la fonction $t \to f(x-t)$, au point 0. Noter aussi, pour se rassurer, que $\int_0^t (t-u)du = t^2/2$ donc la formule est correcte quand f'' est constante. Enfin, (3.35), ainsi que les formules semblables aux ordres supérieurs, se démontre en intégrant par parties autant de fois que nécessaire.

Je n'écris pas les formules de Taylor aux ordres supérieurs parce que vous les connaissez, et qu'on n'en aura pas besoin.

Fin de cours 3 en 2021.

b- Formules de Taylor dans $U \subset \mathbb{R}^n$

D'abord à l'ordre 1. On commence par écrire une formule des accroissement finis valable pour $f : U \to \mathbb{R}$.

Proposition 3.7. *Soient $U \subset \mathbb{R}^n$ un ouvert, $x \in U$ et $h \in \mathbb{R}^n$ tels que le segment $[x, x+h]$ est contenu dans U, et $f : U \to \mathbb{R}$ une fonction. On suppose que f est différentiable en tout point de $[x, x+h]$. Alors il existe $\eta \in [x, x+h]$ tel que*

$$f(x+h) = f(x) + Df(\eta)(h), \tag{3.36}$$

où donc $Df(\eta)$, la différentielle de f au point η, est appliquée au vecteur h.

On démontre ceci en appliquant la formule des accroissements finis à la fonction d'une variable réelles g définie par

$$g(t) = f(x + th) \quad \text{pour } 0 \leq t \leq 1. \tag{3.37}$$

Vérifions que cette fonction est bien dérivable sur $[0, 1]$ (on n'aurait besoin que de la continuité en 0 et 0, mais bon. Pour la dérivabilité en t_0, on constante que par définition, $g'(t_0)$ est la dérivée directionnelle de f au point $x+t_0h$, dans la direction h, et qu'on a notée $\partial_h f(x+t_0h)$ dans le passé. Cette dérivée existe, parce que f est différentiable en $x + t_0h$, et est donnée par

$$g'(t) = \partial_h f(x + t_0h) = \sum_{j=1}^{n} h_i \frac{\partial f}{\partial x_j}(x + t_0h) \tag{3.38}$$

La formule des accroissements finis (3.30), appliquée à g, donne maintenant

$$f(x + h) - f(x) = g(1) - g(0) = g'(\xi) = \sum_{j=1}^{n} h_i \frac{\partial f}{\partial x_j}(x + t_0h) = Df(x + \xi h)(h), \tag{3.39}$$

pour un $\xi \in]0, 1[$. C'est ce qu'on voulait, avec $\eta = x + \xi h \in [x, x + h]$. □

Remarque: si on ajoute l'hypothèse que Df est aussi continue sur $[x, y]$, alors la dérivée g', qui est donnée par (3.38), est continue sur $[0, 1]$. Et alors on peut aussi utiliser la formule d'intégration (3.32), qui donne

$$f(x + h) - f(x) = g(1) - g(0) = \int_0^1 g'(t)dt = \sum_{j=1}^{n} h_i \int_0^1 \frac{\partial f}{\partial x_j}(x + th)dt \tag{3.40}$$

(où pour chaque i, la fonction de t à intégrer est bien continue, donc l'intégrale a un sens).

Notons encore que par contre, la proposition ne vaut pas, même quand $n = 1$, quand f est à valeurs dans $\mathbb{R}^m$ pour un $m \geq 2$. Autrement dit, le théorème des accroissements finis est faux en général pour une fonction (dérivable) $f : \mathbb{R} \to \mathbb{R}^m$.

En effet, il est exact qu'on a bien (3.30) pour chacune des coordonées f_i de f, mais pour chacune on trouve un point ξ_i, qui peut être différent des autres. Ainsi, pour $f(x) = e^{ix}$, disons entre 0 et 2π, on a que $f(2\pi) = f(0)$ et pourtant f' ne s'annule pas donc on ne peut pas avoir (3.30).

Passons maintenant à l'ordre 2. On suppose maintenant que $f : U \to \mathbb{R}$ est de classe C^2 dans un voisinage de $x \in U \subset \mathbb{R}^n$. Voici l'énoncé le plus simple: la formule de Taylor-Young dans ce cadre.

Théorème 3.8. *Soit $f : U \to \mathbb{R}$ de classe C^2 dans un voisinage de $x \in U \subset \mathbb{R}^n$. Alors pour $h = (h_1, \ldots, h_n) \in \mathbb{R}^n$ petit, on a*

$$f(x + h) = f(x) + Df_x(h) + \frac{1}{2}Q(h) + o(||h||^2), \tag{3.41}$$

ou $Df_x = Df(x)$ est la différentielle de f en x (une application linéaire de $\mathbb{R}^n$ dans $\mathbb{R}$, on en a parlé) et Q est la forme quadratique associée à la seconde dérivée de f en x, donnée par

$$\begin{aligned} Q(h) &= \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\partial^2 f}{\partial x_i \partial x_j}(x) h_i h_j \\ &= \sum_{i=1}^{n} \frac{\partial^2 f}{\partial x_i^2}(x) h_i^2 + 2 \sum_{1\le i<j\le n} \frac{\partial^2 f}{\partial x_i \partial x_j}(x) h_i h_j. \end{aligned} \tag{3.42}$$

J'ai mis la seconde ligne pour insister un peu. En pratique, elle va plus vite pour calculer puisque l'on n'a pas besoin d'écrire deux fois le même terme. Evidemment c'est pareil. Rappelons que

$$Df_x(h) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x) h_i \tag{3.43}$$

ou, avec des notations un peu plus ramassées,

$$Df_x(h) = \langle \nabla f(x), h\rangle. \tag{3.44}$$

De même, si on note H la matrice Hessienne de f au point x, donc la matrice carrée donnée par $H_{i,j} = \frac{\partial^2 f}{\partial x_i \partial x_j}(x)$, on peut aussi écrire en notation matricielle

$$Q(h) = h^t H h \tag{3.45}$$

où (cette fois c'est important) on a considéré h comme un vecteur colonne, puis écrit le transposé h^t qui est un vecteur ligne, et on a bien obtenu un nombre).

Reste à démontrer le théorème, et ensuite on va faire quelques remarques. En fait on va démontrer un peu plus.

On se donne $x \in U$ et $h \in \mathbb{R}^n$, assez petit pour que le segment $[x, x+h]$ est contenu dans U. Et on pose, comme ci-dessus,

$$g(t) = f(x+th) \quad \text{pour } 0 \le t \le 1. \tag{3.46}$$

On va vérifier que c'est une fonction de classe C^2, avec des dérivées qu'on va pouvoir écrire, et ensuite on appliquera le théorème de Taylor-Lagrange à g. En fait, on a déjà calculé en (3.38) que

$$g'(t) = \langle \nabla f(x+th), h\rangle = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x+th) h_i. \tag{3.47}$$

Et maintenant on voit que chaque fonction $g_i = \frac{\partial f}{\partial x_i}(x+th)$ est comme $g(t) = f(x+th)$ mais avec f remplacée par $\frac{\partial f}{\partial x_i}$. Donc sa dérivée en t est $g_i'(t) = \sum_{j=1}^{n} \frac{\partial^2 f}{\partial x_j \partial x_i}(x+th) h_j$. On multiplie par h_i, on somme, et on trouve

$$g''(t) = \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\partial^2 f}{\partial x_j \partial x_i}(x+th) h_i h_j \tag{3.48}$$

pour $0 \leq t \leq 1$. C'est parfait. On applique maintenant le théorème de Taylor-Lagrange à g entre 0 et 1 et on trouve qu'il existe $\xi \in (0,1)$ tel que

$$\begin{aligned} f(x+h) - f(x) &= g(1) - g(0) = g'(0) + \frac{1}{2}g''(\xi) \\ &= \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x)h_i + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\partial^2 f}{\partial x_j \partial x_i}(x+\xi h)h_i h_j. \end{aligned} \tag{3.49}$$

C'est donc un peu plus précis que le théorème, puiqu'on a une formule semi-explicite avec un point $x + \xi h$ sur le segment $[x, x+h]$. En tout cas le théorème s'en déduit puisque quand h tend vers 0, $x + \xi h$ tend vers x et donc chaque $\frac{\partial^2 f}{\partial x_j \partial x_i}(x + \xi h)$ tend vers le $\frac{\partial^2 f}{\partial x_j \partial x_i}(x)$ correspondant. □

C'est bien de se souvenir qu'on peut faire plus précis que le théorème si on en a besoin (voire écrire une formule de Taylor avec reste intégral, en fait même à tous les ordres si on en a le courage). Et aussi de se souvenir comment on a fait: le théorème en une seule variable suffit.

C'est bien d'avoir supposé que f est C^2 (en fait dans un voisinage de $[x, x+h]$), à la fois pour la limite finale, mais aussi pour ne pas se prendre la tête avec des dérivées croisées incalculables (si on n'a pas Schwarz).

Donc la dérivée seconde, on peut encore la maîtriser avec des matrices. Pour la dérivée troisième, on aurait une somme triple (la dérivée troisième est associée à une application tri-linéaire), et le plus simple alors serait de garder des sommes triples.

On va parler au prochain chapitre de fonctions f à valeurs dans $\mathbb{R}^m$, ou déjà dans $\mathbb{C}$. On verra que la règle simple, qui consiste à **prendre** les dérivées et écrire les formules **coordonnée par coordonnée**, marche très bien.

4 Différentiation des fonctions de $\mathbb{R}^n$ dans $\mathbb{R}^m$

Ce sera surtout une histoire de notations. Normalement, il n'y a pas grand-chose à dire, je campe sur mes positions comme quoi si $f = (f_1, \ldots, f_m)$ est à valeur dans $\mathbb{R}^m$, tout ce qu'on a à faire c'est étudier chaque f_j et éventuellement rassembler les résultats de manière agréable ou synthétique.

Donc on va surtout essayer ici de s'organiser un peu.

4.1 Généralités et composition

Notation: $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ sera l'espace des applications linéaires de $\mathbb{R}^n$ dans $\mathbb{R}^m$. On a vu et utilisé le cas où $m = 1$. Une fois qu'on a choisi les bases canoniques dans $\mathbb{R}^n$ et $\mathbb{R}^m$, tout $L \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ sera représentée par une matrice $M = M(L)$ à m lignes et n colonnes. De sorte que si $v \in \mathbb{R}^n$ est représenté par le vecteur colonne V, son image $L(v)$ est le vecteur colonne MV.

On se donne maintenant $f : U \subset \mathbb{R}^n \to \mathbb{R}^m$. On notera f_i sa i-ième coordonnée, $1 \leq i \leq m$. Et on identifie $f(x)$ à un vecteur colonne (avec m lignes). Les dérivées partielles de f, qu'on notera encore $\partial_j f$ ou $\frac{\partial f}{\partial x_j}$ ou $\partial_j f$, sont définies comme avant, sauf que maintenant ce sont des vecteurs de $\mathbb{R}^m$, qu'on notera comme des vecteurs colonne.

Quand elles existent, on les rangera dans une matrice M à m lignes et n colonnes, dont le terme général est donc $\partial_j f_i$ à la ligne i et à la colonne j. Je noterai cette matrice M, ou $M(x)$ (si on calcule au point x), ou $J_f(x)$, parce qu'on l'appelle aussi matrice jacobienne de f. Donc il n'y a pas de raison que ce soit une matrice carrée et quand $m = 1$ c'est une matrice à une ligne, qui serait donc plutôt la transposée de $\nabla f(x)$. Mais quand $m > 1$, je crois qu'il est maladroit d'essayer de parler du gradient: on a déjà assez de confusion comme ça.

Revenons à la théorie. On dit que f est différentiable en $z \in U$ quand il existe une application linéaire $L : \mathbb{R}^n \to \mathbb{R}^m$ telle qu'on ait le développement limité

$$f(z + y) = f(z) + L(y) + o(||y||) \tag{4.1}$$

quand y tend vers 0. Ca ressemble à ce qu'on a fait plus haut. Quand on regarde la coordonnée i, on trouve $f_i(z + y) = f_i(z) + L_i(y) + o(||y||)$, où L_i est la coordonnée i de L. Donc f et différentiable si et seulement si chacune de ses coordonnées l'est. On a élaboré un peu plus en cours (comment retrouver L quand on a les $\partial_j f_i$?). La matrice de L est bien M: à la colonne j, on a mis le vecteur $L(e_j)$.

On se souvient que $o(||y||)$ est juste une manière d'appeler une fonction g telle que $\lim_{y\to 0} \frac{g(y)}{||y||} = 0$.

Exercice de vérification conseillé: écrire une fonction simple $f : \mathbb{R}^2 \to \mathbb{R}^3$ et calculer la matrice J_f en un point (3 lignes, 2 colonnes).

Evidemment, $J_f(z)$ dépend de notre choix d'utiliser les bases canoniques. Si on changeait la base de $\mathbb{R}^m$, ou aussi de coordonnées dans $\mathbb{R}^n$, on aurait une autre matrice.

Alors que l'application linéaire L, que je noterai volontiers $Df(z)$, est plus intrinséque. Et son effet sur le vecteur y est noté $D_f(z)(y)$ (ou $J_f(z)y$) comme avant.

Important: le théorème comme quoi différentiable implique dérivées partielle, et la réciproque qui dit que si on a des dérivées partielles continues alors f est différentiable, restent vrais (c'est vrai coordonnée par coordonnée).

On définit les fonctions de classe C^1, les dérivées partielles successives, la classe C^k comme avant. D'ailleurs, déjà dans le cas de $f : \mathbb{R}^n \to \mathbb{R}$, la seconde dérivée est déjà la dérivée d'une fonction g à valeurs dans $\mathcal{L}(\mathbb{R}^n, \mathbb{R})$, qu'on peut identifier à $\mathbb{R}^n$ par le théorème de représentation de Riesz (en passant aux gradients). Quand on dit ça on mélange un peu, mais pas tant que ça.

Maintenant, avec toutes ces notations, on est en état de composer les dérivées. Ce qu'on n'avait pas encore fait officiellement (curieusement).

Théorème 4.1. *Soient $f : U \subset \mathbb{R}^n \to \mathbb{R}^m$ et $g : V \subset \mathbb{R}^m \to \mathbb{R}^p$ deux fonctions définies dans des ouverts. On suppose que f est différentiable en $x \in U$ (notons $L_f = Df(x)$), et g est différentiable en $f(x)$ (notons $L_g = Dg(f(x))$. Alors $g \circ f$ est différentiable en x, et sa différentielle en x est $L_g \circ L_f$.*

C'est logique en termes de composition. Vérifier toutes les flèches. Pour la démonstration, observer d'abord que comme f est différentiable en x, elle est continue aussi; donc comme V contient une petite boule ouverte centrée en $f(x)$, il existe $r > 0$ tel que $f(B(z,r)) \subset V$, donc $g \circ f$ est définie dans cette petite boule (et on peur parler de différentier $g \circ f$ en x). Ensuite on écrit les deux développements limités et ca marche (par composition des limites): avec les notations de l'énoncé,

$$f(x+h) - f(x) = L_f(h) + o(h) \tag{4.2}$$

où je gagne un peu de place en écrivant $o(h)$ au lieu de $o(||h||)$, et aussi en posant $y = f(x)$

$$g(y+\ell) - g(y) = L_g(\ell) + o(\ell). \tag{4.3}$$

Pour éviter de vous faire peur avec les notations (et un peu exceptionnellement), je traduis. Pour (4.2) il existe une fonction ε_1 définie dans une petite boule de $\mathbb{R}^n$ centrée en 0, telle que $\lim_{h\to 0} \varepsilon_1(h) = 0$ et $f(x+h) - f(x) = L_f(h) + ||h||\varepsilon_1(h)$, et de même, pour (4.3), il existe une fonction ε_2, telle que $\lim_{\ell\to 0} \varepsilon_2(\ell) = 0$ et $g(y+\ell) - g(y) = L_g(\ell) + ||\ell||\varepsilon_2(\ell)$ au voisinage de 0. Dans le premier cas, $h \to 0$ signifie que chacune des coordonnées de h tend vers 0, ou si vous préférez que $||h||$ tend vers 0; pareil pour ℓ. Et les fonctions ε_1 et ε_2 sont à valeurs vectorelles (donc chacune de leur coordonnée tend vers 0.

En tout cas, par (4.2) et (4.3),

$$\begin{aligned} g\circ f(x+h) - g\circ f(x) &= g(f(x+h)) - g(y) \\ &= L_g(f(x+h) - f(x)) + o(||f(x+h) - f(x)||) \\ &= L_g(L_f(h) + o(h)) + o(||f(x+h) - f(x)||) \\ &= L_g(L_f(h)) + L_g(o(h)) + o(||f(x+h) - f(x)||). \end{aligned} \tag{4.4}$$

Il ne reste plus qu'à remarquer que $L_g(o(h)) = o(h)$ parce que $||L_g(u)|| \leq C||u||$ pour tout u (L_g est une application linéaire, continue donc bornée, parce qu'on est en dimension finie), et que $||f(x+h) - f(x)|| = ||L_f(h) + o(h)|| \leq ||L_f(h)|| + o(h) \leq C||h|| + o(h)$, donc les deux derniers termes sont bien négligeables devant $||h||$.

Ou alors, en version développée, en posant $\ell = f(x+h) - y$, et en notant bien que par la continuité de f en x (qui suit de sa différentabilité en x), ℓ tend vers 0 quand h tend vers 0,

$$\begin{aligned} g\circ f(x+h) - g\circ f(x) &= g(f(x+h)) - g(y) = g(y+\ell) - g(y) \\ &= L_g(\ell) + ||\ell||\varepsilon_2(\ell) = L_g(f(x+h) - y) + ||\ell||\varepsilon_2(\ell) \\ &= L_g(f(x+h) - f(x)) + ||\ell||\varepsilon_2(\ell) \\ &= L_g(L_f(h) + ||h||\varepsilon_1(h)) + ||\ell||\varepsilon_2(\ell) \\ &= L_g(L_f(h)) + ||h||L_g(\varepsilon_1(h)) + ||\ell||\varepsilon_2(\ell). \end{aligned} \tag{4.5}$$

Le terme principal est $L_g(L_f(h)) = L_g \circ L_f(h)$, où $L_g \circ L_f$ est bien la différentielle présumée de $f \circ g$ au point x, et il ne reste plus qu'à vérifier que les deux termes suivants sont négligeables devant $||h||$. Pour le premier, $||h||L_g(\varepsilon_1(h))$, il est déjà tout écrit sous la forme souhaitée,

et il s'agit juste de voir que $L_g(\varepsilon_1(h))$ tend vers 0, parce que $\varepsilon_1(h)$ tend vers 0 et L_g, qui et une application linéaire (en dimension finie), est continue en 0. Pour le second, on observe que $\ell = f(x+h) - y = f(x+h) - f(x) = L_f(h) + ||h||\varepsilon_1(h)$, donc par inégalité triangulaire $||\ell|| \leq ||L_f(h)|| + ||h||\,||\varepsilon_1(h)|| \leq C||h|| + ||h||\,||\varepsilon_1(h)||$, où C est la nome de L_f. Donc pour $h \neq 0$,

$$\frac{||\ell||\varepsilon_2(\ell)}{||h||} \leq \frac{\big[C||h|| + ||h||\,||\varepsilon_1(h)\big]\varepsilon_2(\ell)}{||h||} = \big[C + ||\varepsilon_1(h)||\big]\varepsilon_2(\ell)$$

qui tend bien vers 0 comme demandé. Evidemment, c'est un peu normal que la démonstration explicite utilise les mêmes ingrédients que la démonstration courte. □

Fin du cours 5 en 2022 (toute fin de démonstration laissée à l'audience).

Note au passage: On a organisé la multiplication des matrices pour que $J_{g\circ f} = J_g J_f$ (garder l'écriture dans cet ordre, pas en échangeant). D'où la multiplication étrange (peut-être) des matrices.

Quelques exemples pour vérifier pour qu'on a compris.

Facile mais utile: quand f linéaire, sa différentielle est f. Quand f est affine, c'est la partie linéaire de f. Exemple: $f(x) = (2x_1 + 3x_2 + 5, 4x_1 + 6x_2 + 7)$, et alors $Df(x)(u,v) = (2u + 3v, 4u + 6v)$.

Les calculs sont plus simples quand la première fonction f va de $I \subset \mathbb{R}$ vers $U \subset \mathbb{R}^n$, et ensuite $g : U \to \mathbb{R}^m$. Dans ce cas la différentielle de f en x est juste donnée par multiplication par le vecteur $f'(x)$. Autrement dit, $Df(x)(u) = uf'(x)$ pour $u \in \mathbb{R}$, et ensuite on compose par $Dg(f(x))$, ce qui donne $D(g \circ f)(x)(u) = Dg(f(x))(Df(x)(u)) = Dg(f(x))(uf'(x)) = uDg(f(x))(f'(x))$ par linéarité. Oui, je sais, ça fait pas mal de parenthèses. Du coup, en notant que $D(g \circ f)(x)(u)$ est lui aussi donné par multipliation par un vecteur dérivée $(g \circ f)'(x)$, on trouve

$$(g \circ f)'(x) = Dg(f(x))(f'(x)) = \sum_{i=1}^{n} \frac{\partial g}{\partial x_i} f_i'(x)$$

en passant en coordonnées, et où chaque $\frac{\partial g}{\partial x_i}$ est un vecteur de $\mathbb{R}^m$. Probablement cette dernière formule est un peu plus claire.

Les dérivées partielles, et aussi la dérivée directionnelle $\partial_h g$, de g, sont des exemples de composition de ce genre, avec pour f la fonction $t \to x + th$ pour calculer une dérivée directionnelle, et $t \to x + te_i$ pour calculer $\frac{\partial g}{\partial x_i}$. Exercice: vérifier tout ceci pour $f(x_1, x_2) = 3x_1 + 2x_2 + 17x_1x_2$.

On prend $g : U \to \mathbb{R}$ différentiable en z. On pré-compose par $f : \mathbb{R} \to \mathbb{R}^n$ défnie par $f(t) = z + th$. La dérivée de la composée (en 0) est ce qu'on a noté $\partial_h g(z)$. Et c'est bien la composée de $t \to z + th$ et de $Df(z)$, comme annoncé.

Autre exemples: la composition avec une application linéaire (ou affine) $g : \mathbb{R}^m \to \mathbb{R}^p$. Juste composer la différentielle de f avec la partie linéaire de g. Précomposer avec un changement de variable $f : \mathbb{R}^n \to \mathbb{R}^n$: c'est pareil mais noter que la matrice de $Df(z)$ est carrée.

4.2 Inégalité dite des accroissements finis

On a dit au dernier chapitre que la formule des accroissements finis est fausse en général pour une fonction à valeurs dans $\mathbb{R}^m$, $m \geq 2$, mais on va quand même vérifier une inégalité, qui serait une conséquence de la formule des accroissements finis si elle était vraie, et qui est souvent suffisante. Je donne d'abord un énoncé simple, et après on commente.

Proposition 4.2. *On se donne un ouvert U de $\mathbb{R}^n$, et deux points $x, y \in U$ tels que $[x, y] \in U$. On suppose $f : U \to \mathbb{R}^m$ est différentiable en chaque point de $[x, y]$. Alors*

$$||f(y) - f(x)|| \leq ||y - x|| \sup_{\eta \in [x,y]} |||Df(\eta)|||. \tag{4.6}$$

On a noté $|||Df(\eta)|||$ la norme de la différentielle $Df(\eta)$, qui est un opérateur linéaire (borné) de $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$, et on a muni $\mathbb{R}^n$ et $\mathbb{R}^m$ chacun de sa norme euclidienne. On verra comment calculer mieux dans notre cas.

En tout cas on commence par appliquer la proposition 3.7, mais comme f n'est plus à valeurs réelle, on triche un peu en se ramenant à une fonction $f_e : U \to \mathbb{R}$. Astuce qu'on pourrait sans doute éviter ici, mais qui est bonne à connaître de tote façon. On se donne un vecteur unitaire $e \in \mathbb{R}^m$, à choisir plus tard, et on considère la fonction f_e définie par

$$f_e(z) = \langle f(z), e \rangle \text{ pour } z \in U. \tag{4.7}$$

Maintenant on est exactement dans la situation et on trouve (comme en (3.36), mais en posant $h = y - x$) que

$$f_e(y) = f_e(x) + Df_e(\eta)(y - x), \tag{4.8}$$

où η est un certain point du segment $[x, y]$. En plus, f_e est la composée de f et de la fonction $g : \mathbb{R}^m \to \mathbb{R}$ définie par $g(w) = \langle w, e \rangle$. Cette dernière application est linéaire, donc égale à sa différentielle, et donc par la formule de composition, $Df_e = g \circ Df$, ce qui en clair signifie que pour tout vecteur $v \in \mathbb{R}^n$,

$$Df_e(\eta)(v) = \langle Df(\eta)(v), e \rangle. \tag{4.9}$$

Du coup

$$||f_e(y) - f_e(x)|| = ||Df_e(\eta)(y - x)|| \leq ||Df(\eta)(y - x)|| \leq |||Df(\eta)||| \, ||y - x|| \tag{4.10}$$

parce que e est unitaire, puis par définition de $|||Df(\eta)|||$.

Ceci est vrai pour tout e unitaire; c'est le moment de choisir le e qui nous plait, et vous avez sans doute déjà vu ceci (pour calculer la norme de g). Si $f(y) - f(x) = 0$, le résultat est vrai trivialement. Sinon, on prend $e = \frac{f(y)-f(x)}{||f(y)-f(x)||}$. On note que $f_e(y) - f_e(x) = \langle f(y) - f(x), e \rangle = ||f(y) - f(x)||$, on utilise l'estimée (4.10) ci-dessus, et on trouve bien (4.6). □

Remarque: si on suppose de plus que Df est continue sur le segment $[x, y]$, on peut utiliser (3.40) au lieu de (3.36), et trouver que

$$f(y) = f(x) + \int_0^1 Df(x + t(y - x))(y - x)dt. \tag{4.11}$$

Pour ce calcul, il reste bien vrai pour $f : U \to \mathbb{R}^n$, parce qu'on peut faire le calcul coordonnée par coordonnée! On utilise maintenant le fait que

$$||Df(x + t(y - x))(y - x)|| \leq ||y - x|||||Df(x + t(y - x))|||$$

pour chaque t, puis on intègre par rapport à $t \in [0, 1]$, et on trouve que

$$||f(y) - f(x)|| \leq ||y - x|| \int_0^1 |||Df(x + t(y - x))|||. \tag{4.12}$$

C'est un peu mieux: une moyenne au lieu d'un sup. Mais j'ai triché un peu: j'ai utilisé le fait que quand on intègre un vecteur $V(t)$ (coordooné par coordonnée), on a bien que

$$\Big|\Big| \int_0^1 V(t)dt \Big|\Big| \leq \int_0^1 ||V(t)||dt. \tag{4.13}$$

C'est une forme intégrée de l'inégalité triangulaire, et qui serait vraie pour n'importe quelle norme dans $\mathbb{R}^m$, mais que vous n'avez sans doute jamais démontrée.

Dans le cas de la norme euclidienne, c'est plus facile, et on vient de voir l'astuce qui permet de le faire sans se fatiguer: pour démontrer (4.13), il suffit (comme ci-dessus) de démontrer que pour tout vecteur unitaire e, $\langle \int_0^1 V(t)dt, e\rangle \leq \int_0^1 ||V(t)||dt$. Mais $\langle \int_0^1 V(t)dt, e\rangle = \int_0^1 \langle V(t), e\rangle dt$ par linéarité (comprendre, par calcul direct avec la formule du produit scalaire), et comme $\langle V(t), e\rangle \leq |V(t)||$, on intègre et on conclut.

Signalons pour finir qu'en fait on n'a pas forcément besoin de calculer toute la norme de $Df(\eta)(y-x)$ (dans toutes les directions), mais seulement la taille de son effet sur $h = y - x$. Alors on peut l'écrire explicitement, puisque

$$Df(\eta)(y - x) = Df(\eta)(h) = \sum_{i=1}^n h_i \partial_i f(\eta) \tag{4.14}$$

Bon, c'est plus compliqué que prévu, puisque f a m coordonnées f_j, $1 \leq j \leq m$, et que pareillement $Df(\eta)(y - x)$ a m coordonnées, les $\sum_{i=1}^n h_i \partial_i f_j(\eta)$. Donc en fait

$$||Df(\eta)(y - x)||^2 = ||Df(\eta)(h)||^2 = \sum_{j=1}^m \Big\{ \sum_{i=1}^n h_i \partial_i f_j(\eta) \Big\}^2 \tag{4.15}$$

Ceci nous fait un bel exercice sur Cauchy-Schwarz. La somme intérieure vaut

$$\Big\{ \sum_{i=1}^n h_i \partial_i f_j(\eta) \Big\}^2 \leq \Big\{ \sum_{i=1}^n h_i^2 \Big\} \Big\{ \sum_{i=1}^n |\partial_i f_j(\eta)|^2 \Big\} = ||h||^2 \Big\{ \sum_{i=1}^n |\partial_i f_j(\eta)|^2 \Big\} \tag{4.16}$$

puis on recopie dans (4.18) et on trouve

$$||Df(\eta)(h)||^2 \leq ||h||^2 \sum_{j=1}^{m} \sum_{i=1}^{n} |\partial_i f_j(\eta)|^2 \tag{4.17}$$

Autrement dit,

$$|||Df(\eta)||| \leq \Big\{ \sum_{j=1}^{m} \sum_{i=1}^{n} |\partial_i f_j(\eta)|^2 \Big\}^{1/2}. \tag{4.18}$$

Souvent on peut faire mieux, mais c'est un bon début.

Bon, on s'écarte un peu. Signalons un corollaire simple de la proposition.

Corollaire 4.3. *Soit U un ouvert convexe de $\mathbb{R}^n$, et $f : U \to \mathbb{R}^m$ une fonction différentiable. On suppose que $|||Df(z)||| \leq M$ pour tout $z \in U$ (pour un certain $M \geq 0$). Alors f est Lipschitzienne sur U, avec une constante de Lipschiz au plus M, ce qui signifie que*

$$||f(x) - f(y)|| \leq M||x - y|| \quad \textit{pour } x, y \in U. \tag{4.19}$$

Convexe signifie que $[x, y] \subset U$ pour $x, y \in U$. Ca tombe bien, puisqu'alors on peut appliquer la proposition, et obtenir (4.6) avec un sup inférieur à M. On en déduit aussitôt (4.19). □

Corollaire 4.4. *Soit U un ouvert convexe de $\mathbb{R}^n$, et $f : U \to \mathbb{R}^m$ une fonction différentiable. On suppose que $Df(z) = 0$ pour tout $z \in U$. Alors f est constante sur U.*

Facile, prendre $M = 0$. □

Mais la vérité est que la bonne hypothèse sur U est que U est connexe, ce qui signifie que (c.à.d., si on peut écrire $U = U_1 \cup U_2$ comme une union disjointe de deux ouverts de U (donc ici, puisque U est ouvert, deux ouverts contenus dans U), alors $U_1 = \emptyset$ ou $U_2 = \emptyset$. Et aussi, il suffit par exemple que f ait des dérivées partielles nulles sur U pour que ça marche. Exercice: faire la démonstration. Indication: se donner $x_0 \in U$ et montrer que $U_1 = \big\{x \in U\,;\, f(x) = f(x_0)\big\}$ est un ouvert.

5 Théorème d'inversion locale

On commence par un court énoncé sur un intervalle qui donne une idée de ce qu'on voudrait. On se done une fonction $f : I \to \mathbb{R}$, définie sur un intervalle qui contient 0 (pour simplifier), et on suppose que $f'(0) > 0$. Quitte à regarder $-f$, on va supposer que f est de classe C^1 dans in petit voisinage de 0, et que de plus $f'(0) > 0$.

Posons $c = f'(0) > 0$. Alors, pour commencer, il existe un voisinage $J = [a, b]$ de 0 sur lequel $f'(x) \geq c/2$. Et du coup, par le théorème des accroissements finis (par exemple), $f(y) - f(x) \geq \frac{c}{2}(y - x)$ pour $x, y \in J$ tels que $y \geq x$. On sait qu'alors f (restreint à J est

une bijection (croissante) de J sur son image $[f(a), f(b)]$, et même, comme f' ne s'annule pas sr J, que la réciproque g est aussi de classe C^1, avec la dérivée g' donnée par la formule $g'(y) = \frac{1}{f'(g(y)}$.

On veut faire pareil avec une fonction définie dans un ouvert U de $\mathbb{R}^n$ (et à valeurs dans $\mathbb{R}^n$, on verra que prendre le même n est obligatoire). On ne pourra plus utiliser la croissance (la notion n'a pas vraiment de sens sur U), mais l'existence d'une différentielle continue au voisinage de 0 et injective en 0 sera en fait suffisante! Ce sera le théorème d'inversion locale.

Par contre, dans un intervalle, on a aussi un théorème global (si f est continue strictement croissante de $[a, b]$ dans $\mathbb{R}$, alors elle est bijective de $[a, b]$ dans $[f(a), f(b)]$, et même la réciproque est continue. L'histoire de la réciproque est une belle conséquence de la compacité de $[a, b]$, mais surtout l'énoncé utilise l'ordre sur $[a, b]$ (et son image). On n'aura pas d'énoncé global aussi simple pour $f : U \to \mathbb{R}^n$.

Mais on va commencer par des définitions pour introduire les notions pertinentes et préparer 'énoncé.

5.1 Difféomorphismes

Pour l'instant, faisons comme si on regardait juste la structure associée aux applications différentiables, ou plutôt de classe C^1, en regardant come d'habitude les histoires de composition, d'inversions, etc.

Pensez à la structure d'espace vectoriel, où les applications qui préservent la structure sont les applications linéaires, ou aussi la structure d'espace topologique, où les bonnes applications sont les applications continues.

Définition 5.1. *Soient U et V deux ouverts de $\mathbb{R}^n$. On dit que $f : U \to V$ est un difféomorphisme (respectivement, un difféomorphisme de classe C^1) si f est différentiable en tout point de U (respectivement, de classe C^1 sur U), est une bijection de U dans V, et si de plus la réciproque $f^{-1} : V \to U$ est différentiable en tout point de V (respectivement, de classe C^1 sur V).*

On se rappelle qu'un homéomorphisme de U dans V est une application continue bijective de U dans V telle que de plus $f^{-1} : V \to U$ est continue aussi. Ici on se restreint à des ouverts (pour pouvoir définir les différentielles en toute tranquilité), et on demande en plus que f et f^{-1} soient différentiables (respectivement, de classe C^1).

La notion correspondante pour des espaces vectoriels (ou de groupes, ou d'anneaux) serait celle d'isomorphisme.

On commence par des remarques simples. D'abord, si f et $g = f^{-1}$ sont toutes les deux différentiables, alors on peut dériver des deux identités $f \circ g = I$ et $g \circ f = I$. Plus précisément, $g(f(x)) = x$ pour tout $x \in U$, donc en dérivant (et en notant I_n l'identité sur $\mathbb{R}^n$,

$$Dg(f(x)) \circ Df(x) = I_n \text{ pour } x \in U.$$

Mais aussi, $f(g(y)) = y$ pour $y \in V$, donc en dérivant, et en supposant un instant que V est un ouvert de $\mathbb{R}^m$,

$$Df(g(y)) \circ Dg(y) = I_m \text{ pour } y \in V.$$

On applique ceci avec $y = f(x)$ et donc $g(y) = x$ et on trouve $Dg(y) \circ Df(x) = I_n$ et $Df(x) \circ Dg(y) = I_m$. Donc $Df(x)$ et $Dg(f(x))$ sont inverses l'une de l'autre. Comme vous savez, ceci implique entre autres que $m = n$, car si F est une application linéaire bijective de $\mathbb{R}^n$ dans $\mathbb{R}^m$, alors $m = n$. C'est le théorème du rang en algèbre linéaire (de dimension finie!), et c'est pour cela qu'on a pris les devants et demandé que $m = n$ avant de commencer.

Pour résumer, $Df(x)$ est inversible (pour tout $x \in U$), et son inverse (comme application linéaire) est $Dg(y)$ (où $y = f(x)$). Donc:

(5.1) si $f : U \to V$ est un difféomorphisme, $Df(x)$ est inversible pour tout $x \in U$, et $Df(x)^{-1} = D(f^{-1})(f(x))$.

Notons au passage qu'en fait qu'il existe un résultat semblable pour les homéomorphisme, souvent appelé l'invariance du domaine, qui dit que si $U \subset \mathbb{R}^n$ et $V \subset \mathbb{R}^m$ sont des ouverts homéomorphes (non vides), alors $n = m$. C'est un résultat de topologie nettement plus délicat à démontrer (mais vrai quand même!). Par contre il y a des applications continues surjectives de $[0, 1]$ sur $[0, 1] \times [0, 1] \subset \mathbb{R}^2$, par exemple (courbes de Peano).

Une petite remarque sur les restrictions: si $f : U \to V$ est un difféomorphisme (donc U et V sont des ouverts de $\mathbb{R}^n$), et $U' \subset U$ est un ouvert, alors la restriction de f à U' est un difféomorphisme de U' sur $V' = f(U')$. Démonstration assez facile, en faisant attention au début. D'abord on doit vérifier que V' est un ouvert, et pour cela on dit que V' est l'image réciproque par f^{-1} (qui est continue!) de l'ouvert U'. Donc V' est ouvert dans V, et c'est un donc un ouvert puisque V lui-même est ouvert (sinon, on aurait juste pu dire que $V' = V \cap W$, où W est un ouvert). Ensuite, on observe que la restriction de f^{-1} à V' fait un très bon inverse pour $f : U' \to V'$, et là les vérifications sont faciles (car la restriction à V' d'une fonction différentiable est différentiable).

Proposition 5.2. *Soit $f : U \to V$ un difféomorphisme. Si de plus f est de classe C^1, alors $f^{-1} : V \to U$ est également de classe C^1.*

Donc $f^{-1} : V \to U$ est un difféomorphisme de classe C^1. Je n'ai pas dit que U et V sont des ouverts du même $\mathbb{R}^n$, mais c'est vrai par définition. Bon, on a déjà vu que pour tout $x \in U$, f^{-1} est différentiable en $y = f(x)$, et

$$D(f^{-1})(y) = Df(x)^{-1} \tag{5.2}$$

ou, en termes de matrices, $J_{f^{-1}}(x) = J_f(y)$. Pour démontrer la proposition, il s'agit donc de démontrer que $Df(x)^{-1}$ est une fonction continue de $y \in V$, où $x = f^{-1}(y)$. Comme x est une fonction continue de y (et même diférentiable), et que $Df(x)$ est une fonction continue de x à valeurs dans l'ensemble $\mathcal{L}^*$ des applications linéaires inversibles de $\mathbb{R}^n$, il ne reste plus qu'à montrer le lemme suivant.

Lemme 5.3. *L'application H de $\mathcal{L}^*$ dans $\mathcal{L} = \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$, définie par $H(\varphi) = \varphi^{-1}$, est continue. Et aussi, notons M_n l'ensemble des matrices carrées de taille n, et $M_n^* \subset M_n$ l'ensemble des matrices carrées inversibles de taille n. L'application $\widehat{H} : M_n^* \to M_n$ qui à $A \in M_n$ associe A^{-1} est continue.*

Noter que tous ces ensembles (je veux dire $\mathcal{L}$, $\mathcal{L}^*$, M_n, M_n^*) sont vus comme des sous-ensembles d'espaces vectoriels $\mathbb{R}^k$ de dimensions finies, donc on y met la topologie de $\mathbb{R}^k$. Ce qui signifie qu'en fait on a mis des coordonnées sur tous ces espaces, et que la convergence est donnée par la convergence de chaque coordonnée.

Bon, il y a une association entre $\mathcal{L}$ et M_n: on a choisi la base canonique, et c'est l'application bijective qui à $\varphi \in \mathcal{L}$ associe sa matrice $M \in M_n$. Les deux sont des espaces vectoriels de dimension finie, donc munis de la topologie correspondante (et on peut même utiliser la norme d'opérateur). Et notre bijection de $\mathcal{L}$ dans M_n est linéaire, donc continue, ainsi que son inverse. Donc les deux énoncés du lemme sont en fait équivalents.

On va le démontrer pour les matrices, car ça a l'air plus simple et calculatoire. D'abord, l'application det qui envoie la matrice A sur son déterminant est continue (c'est une fonction polynomiale des coefficients de A). Donc M_n^* est un ouvert de M_n (l'image inverse de $\mathbb{R} \setminus \{0\}$ par cette application). Il reste à voir que $A \to A^{-1}$ est continue sur cet ouvert. Et à nouveau, pas de problème, c'est en fait une fraction rationnelle, en fonction des coefficients de A, avec un dénominateur $\det(A)$ qui ne s'annule pas. Quelque chose comme $a_{ij} = \frac{1}{\det(M)} (-1)^{i+j} \det(A_{ij})$, où A_{ij} est la matrice de cofacteurs où on a supprimé la ligne j et la colonne i à M, puis on a transposé tout ça. □

Finalement, on a obtenu $Dg(z)$ en composant trois fonctions, toutes les trois continues au point choisi. □

Remarque: pour montrer que Z est une partie ouverte de $\mathcal{L} = \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$, et que l'application $A \to A^{-1}$, de Z dans $\mathcal{L}$, est continue, on a une autre ressource, qui est de se ramener au cas de la matrice identique, et d'inverser par une série de Neuman. Je crois que je vais éviter, parce que la la partie "se ramener" semble un peu désagréable quand on entre dans les détails.

A ce stade, on est content de la notion de difféomorphisme, et on va enfin pouvoir passer au fameux théorème de d'inversion locale.

Fin du cours 6 en 2022 (mais en passant sous silence certaines choses sur les formules de Taylor).

5.2 Théorème d'inversion locale

Théorème 5.4. *Soient $U \subset \mathbb{R}^n$ un ouvert, $f : U \to \mathbb{R}^n$ une fonction, et $x_0 \in U$. On suppose que f est de classe C^1 sur U, et que $Df(x_0)$ est inversible. Alors il existe $r > 0$ tel que $B = B(x_0, r) \subset U$, et la restriction de f à B est un difféomorphisme de classe C^1 sur son image $f(B)$, qui est un ouvert.*

Noter qu'on a choisi de prendre une petite boule B parce que c'est plus joli. Mais si on a le théorème avec une boule B (ou même un ouvert B qui contient x_0), et si $B' \subset B$

est n'importe quel ouvert qui contient encore x_0, alors la conclusion reste vraie avec la restriction de f à B'. Donc par exemple, on aurait pu décider de choisir B tel que $f(B)$ soit un ouvert (en remplaçant B par l'image inverse d'une boule contenant $f(x_0)$ et contenue dans $f(B)$). Mais évidemment pas les deux à la fois: en principe l'image d'une boule n'est pas une boule. Pour cette remarque, on a bien utilisé le fait que la restriction à un ouvert d'un difféomorphisme est encore un difféomorphisme.

On pourrait supposer seulement que f est de classe C^1 sur un voisinage $\widetilde{U}$ de x_0; c'est pareil, parce que rien n'empêche d'appliquer le théorème à $\widetilde{U}$ au lieu de U.

Dire que que $Df(x_0)$ est inversible, c'est pareil que de dire qu'elle est injective, ou que le déterminant de $J_f(x_0)$ est non nul, ou que $Df(x_0)$ est surjective, puisqu'on a pris $f : U \to \mathbb{R}^n$. Donc (à cause du déterminant) c'est facile à vérifier!

J'ai demandé C^1, mais la démonstration marche aussi en supposant f différentiable dans un voisinage de 0, et Df continue en 0.

Contrairement à ce qui se passe pour $f : I \to \mathbb{R}$, l'invertibilité de $Df(x_0)$, même en tout point de U, n'implique pas l'injectivité globale de f, parce que f peut tourner lentement. Le plus simple est de considérer l'application $(r, \theta) \to (r\cos(\theta), r\sin(\theta))$, disons défini pour $\theta \in \mathbb{R}$ et $1 < r < 2$. On sait ce qui se passe: l'image est un anneau, qu'on parcourt une infinité de fois.

Vous pourriez objecter qu'il se passe quelque chose de bizarre en 0, alors on peut aussi cacher ceci à l'infini. Par exemple, toujours dans le plan, prenons $f(x, y) = (e^x\cos(y), e^x\sin(y))$ pour $(x, y) \in \mathbb{R}^2$. Cette application n'est pas injective (ajouter 2π à y), et pourtant sa différentielle est de rang 2 en tout point, puisque la matrice jacobienne est $\begin{pmatrix} e^x\cos(y) & e^x\sin(y) \\ -e^x\sin(y) & e^x\cos(y) \end{pmatrix}$, dont le déterminant est $e^{2x} > 0$ partout.

Le fait que $f : B(x_0, r) \to \mathbb{R}^n$ est ouverte, c'est-à dire que $f(W)$ est un ouvert pour tout $W \subset B(x_0, r)$ ouvert, sera une conséquence de l'invertibilité de $Df(x)$ pour tout $x \in B(x_0, r)$. Cette propriété vient avec le théorème et n'allait pas de soi a priori. Pensez à l'application $(x, y) \to (x, |y|)$ qui fait un pli.

On commence la démonstration par une <u>première réduction</u>: il suffit de démontrer le théorème quand $x_0 = 0$ et $y_0 = f(x_0) = 0$. En effet, si f est comme dans l'énoncé, remarquons que $\widetilde{f}$, définie par $\widetilde{f}(x) = f(x + x_0) - y_0$ vérifie aussi les hypothèses, avec en plus $x_0 = 0$ et $\widetilde{f}(0) = y_0 - y_0 = 0$. Noter au passage que $D\widetilde{f}(0) = Df(x_0)$ (qui est inversible). Maintenant, si on trouve $r > 0$ tel que $\widetilde{f}$ vérifie la conclusion, donc en particulier est un difféomorphisme sur son image (un ouvert), f aussi, puisque $f(x) = \widetilde{f}(x - x_0) + y_0$. L'image est celle de $B(0, r)$ par $\widetilde{f}$, translatée de y_0.

Pour cette réduction, et encore plus pour la suivante, on conseille vivement au lecteur de faire un petit diagramme avec des flèches pour vérifier qu'on a compris les domaines et espaces d'arrivée de toutes nos applications.

<u>Seconde réduction</u> (qui nous simplifiera quand même la vie): on peut supposer (en plus) que $Df(x_0) = Df(0) = I$ (l'application identique). En effet, si f est comme dans l'énoncé et si on note $L = Df(x_0)$, alors $\widetilde{f} = L^{-1} \circ f$ vérifie les mêmes hypothèses avec de plus

$D\widetilde{h}(x) = L^{-1} \circ Df(x)$, qui maintenant vaut I en 0. Si on sait démontrer le résultat pour $\widetilde{f}$, on en déduit pareil pour $f = L \circ \widetilde{f}$.

Voilà. On peut supposer qu'en plus des hypothèses, $x_0 = 0 = f(x_0)$ et $Df(x_0) = I$. On choisit maintenant r assez petit pour que

$$|||Df(x) - I||| \leq \varepsilon \quad \text{pour } x \in B(0,r), \tag{5.3}$$

où l'on peut choisir $\varepsilon > 0$ aussi petit qu'on veut, mais en fait $\varepsilon = 1/2$ marche déjà très bien. Je prends la norme euclidienne pour définir la norme d'opérateur $|||Df(x) - I|||$. Et l'existence de r vient de la continuité de Df en 0.

Montrons déjà sous cette hypothèse que f est injective sur $B(0,r)$, et même que

$$||f(x) - f(y)|| \geq (1-\varepsilon)||x - y|| \quad \text{pour } x, y \in B(0,r). \tag{5.4}$$

En effet, posons $g(x) = f(x) - x$; donc $|||Dg(x)||| \leq \varepsilon$ pour $x \in B(0,r)$, et la forme la plus simple du théorème des accroissements finis (prenez par exemple la proposition 4.2) dit que

$$||g(x) - g(y)|| \leq \varepsilon||x - y|| \quad \text{pour } x, y \in B(0,r). \tag{5.5}$$

Il reste à apliquer l'inégalité triangulaire: $f(x) - f(y) = (x - y) + (g(x) - g(y))$ donc

$$||f(x) - f(y)|| \geq ||x - y|| - ||g(x) - g(y)|| \geq (1-\varepsilon)||x - y||.$$

On n'a pas encore fini: il faut maintenant prouver que $f(B(x,r))$ est ouvert, ou au moins contient une boule ouverte $B(0,\rho)$. Et on va prendre $\rho = r/3$ et donc prouver que

$$B(0,\rho) \subset f(B(0,r)). \tag{5.6}$$

Autrement dit, étant donné $y \in B(0,\rho)$, on doit juste trouver $x \in B(0,r)$ tel que $f(x) = y$. Rien qu'une petite équation à résoudre.

On essaie brutalement, en utilisant le fait que $Df - I$ est petit. Ensuite on formalisera tout ceci. On pense qu'on tire au canon et qu'on essaie de viser la cible y, en modifiant x (les paramètres du tir). Donc $f(x)$ est l'endroit où le boulet tombe, on tire plusieurs fois, et essaye de corriger le tir au fur et à meure en otant de x l'erreur commise à chaque nouvel essai.

A l'essai $k \geq 0$, on notera x_k notre essai pour x, $f(x_k)$ le point où ça tombe, $e_k = y - f(x_k)$ l'erreur commise, et au cran suivant on essaie $x_{k+1} = x_k + e_k$. Redisons ceci.

Essai 0: on prend $x_0 = 0$, et on fait l'erreur $e_0 = y - f(x_0) = y$. Premier essai: on prend $x_1 = 0 + e_0 = y$. On trouve $f(x_1)$ au lieu de y; nouvelle erreur $e_1 = y - f(x_1) = y - f(y)$, qu'on espère plus petite que $e_0 = y$.

Puis par récurrence, en fait pour $k \geq 0$, on a déjà choisi x_k, tapé en $f(x_k)$, commis l'erreur $e_k = y - f(x_k)$, et on essaie $x_{k+1} = x_k + e_k$ comme annoncé plus haut. On se souvient qu'on peut continuer comme ceci tant que x_k ne sort pas de $B = B(0,r)$.

En fait la formule va marcher parce qu'on a (5.3) et qu'on cherche la formule la plus simple possible, mais en vrai ce qui marcherait le mieux (et même sans supposer au début

que $Df(x_0) = I$), ce serait de prendre $x_{k+1} = x_k + Df(x_k)^{-1}(e_k)$, parce que c'est une formule qui tomberait juste si Df était constante égale à $Df(x_k)$, et dont on imagine qu'elle donne la meilleure approximation du x cherché. Mais ici on veut juste s'en tirer avec le minimum de calculs (et en fait sans avoir à calculer ou connaître précisément $Df(x_k)^{-1}$), donc on prend $x_{k+1} = x_k + e_k$.

Commentaire sur la marine de grand-père: un coup trop près, puis un coup trop loin mais en notant bien ce qu'on a fait, puis on utilise les deux premiers essais pour étalonner, c.-à-d. déterminer la dérivée de la fonction f (on pense, d'une seule variable), et ensuite on vise pour de vrai.

Maintenant on estime les erreurs. On veut s'assurer qu'on tombe de plus en plus près, donc on suppose que x_k et x_{k+1} sont dans B, et on estime $e_{k+1} = y - f(x_{k+1})$, à savoir

$$e_{k+1} = y - f(x_{k+1}) = y - f(x_k + e_k) = f(x_k) + e_k - f(x_k + e_k). \tag{5.7}$$

A quoi on joue? On veut vérifier que e_{k+1} est plus petit que e_k. Si la différentielle de f était constante égale à I, on aurait que $f(x_k + e_k) = f(x_k) + e_k$ et l'erreur serait nulle. Ici, (5.3) dit que Df est proche de I, donc on s'attend à ce que e_{k+1} soit petit. Et justement, en reprenant $g(x) = f(x) - x$,

$$\begin{aligned} e_{k+1} = f(x_k) + e_k - f(x_k + e_k) &= [g(x_k) + x_k] + e_k - [g(x_k + e_k) + (x_k + e_k)] \\ &= g(x_k) - g(x_k + e_k). \end{aligned} \tag{5.8}$$

On a juste soustrait de force l'identité pour faire apparaître des petits termes. Et maintenant (5.5) dit que

$$||e_{k+1}|| \leq \varepsilon ||e_k||. \tag{5.9}$$

On est parti de $e_0 = y$. Et, tant que $0, x_1, \ldots, x_k$ et x_{k+1} restent dans $B(0, r)$, on trouve

$$||e_{k+1}|| \leq \varepsilon ||e_k|| \leq \ldots \varepsilon^k ||e_0|| = \varepsilon^k ||y|| = \varepsilon^k \rho. \tag{5.10}$$

Maintenant, où se trouvent les x_j? Toujours tant que x_k est resté dans B, on peut utiliser les estimations précédentes, et en fait $x_0 = 0$, $x_1 = x_0 + e_0$, $x_2 = x_1 + e_1$, et en sommant tout

$$x_k = e_0 + e_1 + \ldots + e_{k-1} \tag{5.11}$$

donc

$$||x_k|| \leq \sum_{j=0}^{k-1} ||e_j|| \leq \sum_{1 \leq j \leq k-1} \varepsilon^j \rho < \frac{\rho}{1 - \varepsilon} \leq 2\rho \tag{5.12}$$

(car on a pris $\varepsilon < 1/2$). Et du coup $x_{k+1} = x_k + e_k$ est bien défini, et vérifie aussi $||x_{k+1}|| < 2\rho$. Donc, par récurrence, tous les x_k sont bien définis, et $e_k \leq \varepsilon^k \rho$ pour tout k.

Ensuite, on déroule. La suite $\{x_k\}$ est de Cauchy, puisque pour $0 \leq k < \ell$,

$$x_\ell - x_k = \sum_{j=k}^{\ell-1} e_j$$

et donc $||x_\ell - x_k|| \leq \sum_{j=k}^{\ell-1} \varepsilon^j \rho \leq 2\varepsilon^k \rho$.

Donc $\{x_k\}$ est convergente. Elle a une limite $z \in \overline{B}(0, 2\rho) \subset B$ (par choix de $\rho = r/3$). Il faut vérifier que $f(z) = y$. Mais $||f(x_k) - y|| = ||e_k||$ tend vers 0, Comme z est la limite des x_k, et que l'application $x \to ||f(x) - y||$ est continue, on en déduit bien que $f(z) = y$.

On a presque fini. Sous les hypothèses du théorème, on a montré que f est injective sur une petite boule $B = B(x_0, r)$, et que $f(B)$ contient une petite boule $B(f(x_0), \rho)$ centrée en $f(x_0)$. Mais en tout point $x \in B$, les hypothèses du théorème sont satisfaites au point x (noter que $Df(x)$ est inversible, par (5.3)), donc $f(x)$ est un point de l'intérieur de $f(B)$. Donc $f(B)$ est bien ouvert. C'est ce qu'il nous manquait, parce que la différentiabilité de f^{-1}, et la continuité de $D(f^{-1})$, viennent de la proposition 5.2. Noter que le fait que $f^{-1} : f(B) \to B$ est 2-Lipschitzienne vient de (5.4). □

Rappelons que le fait que $f^{-1} : f(B) \to B$ est aussi différentiable, avec même une différentielle continue si f a une différentielle continue, a été vérifié au début du chapitre.

5.3 Le théorème (le plus célèbre) de point fixe

Bon, en y réfléchissant (et en étant un peu aidé), on se rend compte qu'on a en fait trouvé un point fixe d'une application contractante. Voyons ceci. Au moment d'essayer de résoudre l'équation $f(x) = y$, on peut poser

$$\Phi(x) = \Phi_y(x) = x - f(x) + y. \tag{5.13}$$

Alors résoudre l'équation $f(x) = y$ (avec $x \in B$), c'est exactement pareil que résoudre l'équation $\Phi(x) = x$ (avec $x \in B$), donc chercher un point fixe à l'application Φ. Et il existe un magnifique théorème qui fait ceci pour nous, et dont l'hypothèse principale est que Φ est contractante, à savoir qu'il existe $c \in (0, 1)$ tel que

$$||\Phi(x) - \Phi(z)|| \leq c||x - z|| \quad \text{pour } x, z \in B. \tag{5.14}$$

Or dans le cas présent,

$$\Phi(x) - \Phi(z) = (x - f(x) + y) - (z - f(z) + y) = g(z) - g(x),$$

où $g(x) = f(x) - x$ comme plus haut, donc (5.14) avec $c = \varepsilon$ se déduit de (5.5). Autrement dit, après une petite vérification de plus (le fait que $\Phi(B) \subset B$ pour pouvoir itérer), on aurait pu utiliser le théorème suivant au lieu de régler le canon à la main. Evidemment, la démonstration est presque la même.

Il est temps de passer au théorème de point fixe.

Théorème 5.5. *On se donne un ensemble fermé $F \subset \mathbb{R}^n$, et une application $\Phi : F \to F$. On suppose que Φ est contractante, ce qui signifie qu'il existe $c \in (0,1)$ telle que*

$$||\Phi(x) - \Phi(y)|| \leq c||x - z|| \quad \textit{pour } x, y \in F. \tag{5.15}$$

Alors Φ a un unique point fixe dans F: il existe un unique $x \in F$ tel que $\Phi(x) = x$.

De plus (et c'est souvent important), pour tout $x_0 \in F$, on peut définir une suite $\{x_k\}$ dans F par (la valeur de x_0 et) le fait que $x_{k+1} = \Phi(x_k)$ pour tout $k \geq 0$, et de plus la suite $\{x_k\}$ converge vers x.

Attention, $c < 1$ est important ($c = 1$ ne marche pas!).

Attention, le fait que $\Phi(F) \subset F$ est important aussi (souvent Φ est définie sur une partie plus grande de $\mathbb{R}^n$, et choisir F comme il faut est important.

Dans le cas précédent, on pouvait prendre $F = \overline{B}(0, r)$. Exercice: vérifier qu'alors la fonction $f = \Phi_y$ de (5.13) est bien contractante et verifie bien la condition $f(F) \subset F$. Ce doit être une partie des calculs faits plus haut.

En fait le théorème marche pareil (avec la même démonstration, en remplaçant seulement $||x - y||$ par la distance dans F) en supposant seulement que F est un espace métrique complet. C'est une méthode qui est beaucoup utilisée pour résoudre diverses équations; souvent la partie subtile de la démonstration consiste, pour un problème donné, à deviner quel est l'espace métrique à utiliser, en particulier quelle distance on peut y mettre pour que f soit contractante pour cette distance (rien n'oblige à prendre la distance euclidienne, même si elle existe!).

Bon, on passe à la démonstration. On va voir l'existence bientôt, mais notons que l'unicité est facile: si x et y sont tous les deux des points fixes, noter que $||\Phi(y) - \Phi(x)|| \leq c||x - y||$ par (5.15), donc $||x - y|| = ||\Phi(y) - \Phi(x)|| \leq c||x - y||$, donc $||x - y|| = 0$, donc $x = y$.

Pour l'existence, on se donne $x_0 \in F$, et on définit la suite $\{x_k\}$ comme dans l'énoncé. L'existence se fait facilement par récurrence, puisque $\Phi : F \to F$. Notre prochaine étape est d'estimer les distances, et c'est facile: pour $k \geq 1$,

$$||x_{k+1} - x_k|| = ||\Phi(x_k) - \Phi(x_{k-1})|| \leq c||x_k - x_{k-1}|| \tag{5.16}$$

par (5.15), donc par récurrence

$$||x_{k+1} - x_k|| \leq c^k||x_1 - x_0||. \tag{5.17}$$

Fin du cours 7, le 8 mars 2022

Ceci doit vous rappeler (5.9). Et maintenant, pour $\ell > k$,

$$||x_k - x_\ell|| \leq ||x_k - x_{k+1}|| + ||x_{k+1} - x_{k+2}|| + \dots ||x_{\ell-1} - x_\ell|| \leq c^k + c^{k+1} + \dots = \frac{c^k}{1-c} \tag{5.18}$$

Donc $\{x_k\}$ est une suite de Cauchy, et puisque F est complet (car c'est un fermé de $\mathbb{R}^n$), cette suite converge vers une limite x_∞.

Il ne reste plus qu'à vérifier que x_∞ est un point fixe (le point fixe cherché). Mais x_∞ est aussi la limite des x_{k+1} (si ça n'est pas clair, vérifiez-le à partir des définition de la convergence d'une suite) et donc

$$(5.19)\quad \begin{aligned} ||f(x_\infty) - x_\infty|| = ||f(\lim_{k\to+\infty} x_k) - x_\infty|| &= ||\lim_{k\to+\infty} f(x_k) - x_\infty|| \\ &= ||\lim_{k\to+\infty} x_{k+1} - x_\infty|| = ||x_\infty - x_\infty|| = 0 \end{aligned}$$

parce que f est continue au point x_∞. On reconnaît aussi le genre de calcul fait en fin de démonstration ci-dessus. □

6 Théorème des fonctions implicites

6.1 Introduction

Dans tout ce qui suit, F est une fonction de classe C^1 définie sur un domaine U de $\mathbb{R}^N$, et à valeurs dans $\mathbb{R}^n$. Avec $N \geq n$ et souvent $N > n$. Et on va s'intéresser aux ensembles de niveau de F. Cela veut dire qu'on se donne $w \in \mathbb{R}^n$, et on considère l'ensemble

$$(6.1)\qquad Z = Z_w = F^{-1}(w) = \big\{x \in U\,;\, f(x) = w\big\} \subset \mathbb{R}^N.$$

L'exemple de base est celui des ensembles de niveau dans une carte de l'IGN, où $N = 2$, $n = 1$, et donc $w \in \mathbb{R}$. On dit souvent lignes de niveau, parce que les ensembles de niveau sont le plus souvent des courbes. Mais pas toujours: penser aux lacs (l'ensemble de niveau peut être un disque), ou à ce qui se passe à un col (où dans le cas le plus simple, deux lignes de niveau se croisent).

L'exemple le plus parlant sera peut-être quand $N = 3$ (on se place dans $\mathbb{R}^3$), $n = 1$ (la fonction F est à valeurs dans $\mathbb{R}$), et les ensembles de niveaux qu'on regarde sont en principe de dimension 2. Par exemple les points de l'atmosphère qui sont à une température (ou une pression) donnée. Cette fois, dans le bons cas, on trouvera des surfaces.

Sans hypothèse supplémentaire, Z peut être n'importe quel ensemble fermé. En fait, si $Z \subset \mathbb{R}^N$ est fermé, on peut prendre $F(x) = \text{dist}(x, Z)$, et on a bien que $Z = \big\{x \in \mathbb{R}^n\,;\, F(x) = 0\big\}$. Vous objeterez que probablement F n'est pas de classe C^1, mais en fait en se débrouillant mieux on peut arranger cela.

Donc Z peut en principe être assez moche. Se souvenir aussi que quand on prend w hors de l'image de F, on obtient $Z_w = \emptyset$.

Il reste une propriété très importante (quoique triviale) des ensembles de niveaux: ils sont disjoints. Je veux dire, $Z_w \cap Z_{w'}$ quand $w \neq w'$. Si c'était faux, les cartes de l'IGN seraient assez illisibles.

Le résultat principal du chapitre est que, en supposant que la différentielle DF est de rang n au point $M_0 \in Z_w$, on arrivera à dire que Z_w est, près de Z_w, une gentille surface de dimension $N - n$; voir plus loin pour les détails.

Ce sera, comme pour le théorème d'inversion locale, seulement un résultat local: on n'aura une bonne description de Z_w que près de M_0.

Exemple 1, un très bon cas (global) mais peu instructif: $N = n = 1$ et $F : I \to J$ a une dérivée qui ne s'annule pas. Alors $F : I \to J$ est strictement croissante (ou décroissante), et chaque Z_w, $w \in J$, est juste un point.

Exemple 2, un bon cas (mais local) celui du théorème d'inversion locale. Prendre $N = n$, supposer que $DF(x_0)$ est inversible, et on trouve que Z_w est réduit à un point pour tout point z d'un petit voisinage de $F(z_0)$.

Mais vous avez deviné qu'on s'intéresse surtout au cas où $n < N$, pour que la géométrie soit plus intéressante. On s'attend à ce que, quand $n = 1$ (donc $F : U \to \mathbb{R}$), Z_w soit une surface de dimension $N - 1$, puis que quand $n = 2$, Z_w soit l'intersection de deux telles surfaces, donc une surface de dimension $N - 2$, et ainsi de suite.

On verra que, pour que la surface soit bien propre, quand $n = 1$ il est utile de demander que $\nabla F(M_0)$ soit non nul. Et que, pour que l'intersection des deux surfaces se fasse bien quand $F = (f_1, f_2)$ est à valeurs dans $\mathbb{R}^2$, il est mieux que $\nabla f_1(M_0)$ et $\nabla f_2(M_0)$ soient non seulement non nuls, mais aussi indépendants. Et ainsi de suite.

Exemple 3. Prenons $N = 2$, $n = 1$, $M_0 = 0$. En supposant que $\nabla F(0) \neq 0$, on verra que près de M_0, Z_w est une jolie courbe avec une tangente en M_0 orthogonale à $\nabla F(0)$. Mais, si nous oublions de demander que $\nabla F(0) \neq 0$, le plus joli qui peut arriver est que Z_w soit composé de plusieurs courbes qui se croisent. Par exemple, si $F(x, y) = xy$, on trouve l'union des deux axes; si $F(x, y) = x(x - y)(x - 17y))$ on trouve une union de trois droites. Si on se débrouille encore plus mal, F peut être constante au voisinage de M_0, et alors Z_w va contenir un disque. Et des choses bien plus horribles peuvent arriver.

Exemple 4. Déjà très proche du cas général. Prenons $N = 3$, $n = 1$, donc on s'attend à un Z_w de dimension 2. Et ce qu'on appellera une jolie surface, ce sera en fait le graphe d'une fonction φ de classe C^1, définie sur un domaine $I \subset \mathbb{R}^2$, et à valeurs dans $\mathbb{R}$. Donc l'ensemble $G_\varphi = \big\{(x, y) \in I \times \mathbb{R}\,;\, y : \varphi(x)\big\}$. Mais l'ensemble G_φ peut aussi être décrit (dans $I \times \mathbb{R}$) par l'équation $F(x, y) = 0$, où $F(x, y) = y - \varphi(x)$. Autrement dit, le cas particulier de $F(x, y) = y - \varphi(x)$ est un cas où on a directement la description qu'on cherche pour l'ensemble Z_0.

Exemple 5. $N = 3$, $n = 1$, et $F(x, y, z) = x^2 + y^2 + z^2$. Pour $w > 0$, la surface $Z_w = F^{-1}(w)$ est la sphère de rayon $\sqrt{w}$, et c'est bien une gentille surface. Notons que $\nabla f(x, y, z) = (2x, 2y, 2z)$ ne s'annule que quand $(x, y, z) = 0$, ce qui confirme que le cas de $w = 0$ est une peu spécial. Et aussi, autour d'un point de Z_w tel que $z > 0$, par exemple, on a bien une description de Z_w comme le graphe de la fonction $(x, y) \to \sqrt{w - x^2 - y^2}$, mais au point $(\sqrt{w}, 0, 0)$, par exemple, on a une tangente verticale et pas de bonne description de Z_w comme graphe sur le plan horizontal. Mais on aurait une bonne description dans d'autres coodonnées.

Enfin en plus de la description de la sphère comme ensemble de niveau, ou comme union de deux graphes, on peut aussi utiliser des représentations paramétriques différentes (par exemple, en coordonnées sphériques). Noter qu'écrire Z comme le graphe de $\varphi : \mathbb{R}^m \to \mathbb{R}^n$,

c'est aussi une représentation paramétrique, puisque le point courant de Z est alors $(x, \varphi(x))$, avec $x \in \mathbb{R}^m$.

Et encore: l'union de l'ensemble de niveau $Z_1 = F_1^{-1}(w_1)$ et $Z_2 = F_2^{-1}(w_2)$, où F_1, F_2 sont à valeurs réelles, est aussi un ensemble de niveau, à savoir $G^{-1}(0)$, où $G(x) = [F_1(x) - w_1][F_2(x) - w_2]$. Exercice: faire pareil quand F_1, F_2 sont à valeurs dans $\mathbb{R}^2$.

6.2 L'énoncé du théorème des fonctions implicites

J'essaie de donner l'énoncé directement, et on commentera ensuite. Avec les notations ci-dessus, on garde n (le nombre d'équations) et on note m la dimension de l'ensemble qu'on va décrire. Donc en fait $N = m + n$. On se placera donc sur $\mathbb{R}^m \times \mathbb{R}^n$, et on notera (x, y) les points de $\mathbb{R}^m \times \mathbb{R}^n$, avec $x \in R^m$ et $y \in \mathbb{R}^n$.

L'énoncé va être un peu long (mais il contient tout).

Théorème 6.1. *Soit U un ouvert de $\mathbb{R}^m \times \mathbb{R}^n$ et soit $F = U \to \mathbb{R}^n$ une fonction de classe C^1. Soit $(a, b) \in U$ un point tel que*

(6.2) *la restriction à $\mathbb{R}^n$ de la différentielle de F en (a, b) est inversible.*

Alors il existe un ouvert I de $\mathbb{R}^m$ qui contient a, un ouvert J de $\mathbb{R}^n$ qui contient b, et une application $\varphi : I \to J$ de classe C^1, avec les propriétés suivantes:

(6.3) $$I \times J \subset U,$$

(6.4) $$\varphi(a) = b$$

(6.5) *pour $(x, y) \in I \times J$, on a que $F(x, y) = F(a, b)$ si et seulement si $y = \varphi(x)$.*

Cet énoncé appelle plein de commentaires.

D'abord, traduisons la condition (6.2) d'inversibilité. La différentielle $DF(a, b)$ de F en (a, b) est une application linéaire de $R^m \times R^n$ vers $\mathbb{R}^n$. Notons la L pour simplifier. Donc $L(u, v) \in \mathbb{R}^n$ est défini pour $(u, v) \in \mathbb{R}^m \times \mathbb{R}^n$. Ce que j'appelle (léger abus de langage) sa restriction à $\mathbb{R}^n$ est l'application $L_y : \mathbb{R}^n \to \mathbb{R}^n$ définie par $L_y(v) = L(0, v)$. Comme les dimensions sont les mêmes, il y a effectivement une chance que L_y soit inversible, et c'est quelquechose qui se détermine en calculant un déterminant.

Quelle est la matrice de L_y dans les bases canoniques? C'est juste la matrice M dont la coordonée $M_{i,j}$ est $\frac{\partial F_j}{\partial y_i}(a, b)$ (une dérivée partielle de F_j). Cette matrice est juste obtenue à partir de celle de $DF(a, b)$ en oubliant toutes les dérivées partielles par rapport aux premières variables. C'est bien une matrice carrée, et (6.2) est vrai si et seulement si $\det(M) \neq 0$.

Fin approx du cours 7, le 7 mars 2023

Si par malheur ce n'était pas le cas, la suggestion est de trouver un autre choix des premières variables, c'est-à-dire de chercher une autre base de $\mathbb{R}^{m+n}$, dans laquelle le déterminant correspondant serait non nul. Voir plus bas, au paragraphe 6.5.

Le cas le plus simple du théorème est celui où $n = 1$ et $F : U \to \mathbb{R}$. Dans ce cas, $U \subset \mathbb{R}^m \times \mathbb{R}^n$, et la matrice 1×1 dans la condition (6.3) est juste le nombre $\frac{\partial F}{\partial y}(a, b)$, la dernière dérivée partielle de F. Et donc, si ce nombre est non nul, on trouve qu'autour du point $M_0 = (a, b, F(a, b))$, l'ensemble de niveau $Z_{F(a,b)}$ est le graphe d'une fonction Lipschitzienne de $\mathbb{R}^m$ dans $\mathbb{R}$.

L'exemple le plus simple avec $n \geq 2$ serait obtenu comme ceci. Prenons $m = 1$, $n = 2$, et une fonction $F : \mathbb{R}^3 \to \mathbb{R}^2$ au hasard. Par exemple, $F_1(x, y, z) = x + 2xy + 3y + z$ et $F_2(x, y, z) = x^2 + y^2 + 17z$. Ecrire la matrice de $DF_1(x, y, z)$. C'est une matrice avec 2 lignes, à savoir $(1 + 2y, 2x + 3, 1)$ et $(2x, 2y, 17)$. On oublie a première colonne (les dérivées par rapport à la première variable), et il reste une matrice carrée, on la calcule en $x = y = z = 0$, et on trouve une matrice dont la première ligne est $(3, 1)$ et la seconde ligne est $0, 17$. Le déterminant est $3 \times 17 \neq 0$, donc on peut appliquer le théorème au point $(0, 0, 0)$ et obtenir une description de l'ensemble de niveau $Z_{0,0} = \big\{(x, y, z)\,;\, F_1(x, y, z) = F_2(x, y, z) = 0\big\}$ autour de $(0, 0, 0)$. Comme étant le graphe d'une certaine fonction φ qui va d'un voisinage de 0 dans $\mathbb{R}$ vers $\mathbb{R}^2$; bref, une jolie courbe. Et il n'est pas étonnant que ce soit une courbe, qui est en fait l'intersection de deux surfaces.

Maintenant, quel rapport avec nos histoires de Z_w? On a pris un point de U, et finalement on regarde l'ensemble de niveau Z_w qui passe par (a, b). Ce qui nous oblige à prendre $w = F(a, b)$ et

$$Z_w = F^{-1}(w) = F^{-1}(F(a, b)) = \big\{(x, y) \in U\,;\, F(x, y) = F(a, b)\big\}. \tag{6.6}$$

Ensuite, on a dû localiser, c'est-à-dire se placer sur un ouvert plus petit, qu'on a pris sous la forme $U_0 = I \times J$ parce que notre graphe sera naturellement contenu dans un tel ensemble, et on a dit qu'on avait $\varphi : I \to J$ de classe C^1 telle qu'on ait (6.5). Notons alors Γ_φ le graphe de φ. Donc

$$\Gamma_\varphi = \big\{(x, y) \in I \times J\,;\, y = \varphi(x)\big\}. \tag{6.7}$$

Et maintenant (6.5) dit exactement que

$$Z_w \cap (I \times J) = \Gamma_\varphi \cap (I \times J). \tag{6.8}$$

Autrement dit, dans notre nouvel ouvert plus petit $U_0 = I \times J$, l'ensemble de niveau Z_w qui passe par (a, b) coincide avec le graphe de φ.

En fait on a dit (6.5) mais on pensait très fort à (6.8).

Ensuite, pourquoi a-t-on été obligé de se restreindre à un ouvert U_0 plus petit? La forme de U_0 comme produit est naturelle si on veut parler de graphe, mais surtout, un peu loin de (a, b), il se peut que Z_w devienne moche, ou tourne pour devenir vertical. Et aussi, sur le grand ouvert U, rien n'empêche que Z_w soit composé de deux graphes, ou plus. Voir l'exemple 5.

Le cas dégénéré où $m = 0$ ne nous intéresse pas beaucoup. Le théorème d'inversion locale donne bien que (a, b) est un point isolé de Z_w, $w = F(a, b)$ mais le théorème des fonctions

implicites en dit plus long: il dit comment l'unique point de Z_w dépend de w. D'ailleurs, la démonstration ci-dessous utilisera le théorème d'inversion locale, et si on regarde bien on verra qu'elle donne aussi la manière dont Z_w dépend de w, pour w proche de $F(a,b)$.

Enfin, pourquoi dire fonctions implicites? On pense à l'ensemble de niveau (disons $n=1$) comme donné par l'équation $F(x,y) = w_0$. On voit ceci comme une manière implicite d'associer y (tel que $F(x,y) = w_0$) à x, par opposition à la manière explicite qui consiste à dire que $y = \varphi(x)$. En tout cas, je pense que c'est une histoire comme ceci.

Et encore un dernier commentaire sur les ensembles de niveau. Dans l'énoncé, on a juste donné ue description de l'ensemble de niveau $Z_{w_0} = F^{-1}(w_0)$, avec $w_0 = F(a,b)$. Mais en fait, il permet d'obtenir aussi une description semblable de tous les ensembles de niveau voisins Z_w, où w est dans une petite boule de $\mathbb{R}^n$ centrée en w_0.

Cette remarque contient deux informations. D'abord, toutes les valeurs de w proches de w_0 sont prises par F; en fait, pour un tel w, ont même trouver $y = y_b \in \mathbb{R}^n$ proche de b tel que $F(a,y) = w$. En effet, on peut appliquer le théorème d'inversion locale à l'application $y \mapsto F(a,y)$ (maintenant d'un voisinage de b dans $\mathbb{R}^n$ et à valeurs dans $\mathbb{R}^n$), et c'est ainsi qu'on trouve y.

Et ensuite, pour avoir une représentation sympathique de Z_w, ils suffit maintenant de d'appliquer le théorème des fonctions implicites à F, mais au point (a, y_w). Comme les dérivées partielles de F sont continues, la condition de non-dégénerescence (4.2) est aussi vraie en ce point!

Donc maintenant on ne sera pas surpris que la démonstration ci-dessous repose sur le théorème d'inversion locale.

6.3 La démonstration du TFI

On se donne donc U, F, et (a,b) comme dans l'énoncé. On pose $w_0 = F(a,b)$; donc on veut une bonne description de $Z_{w_0} = F^{-1}(w_0)$.

On va appliquer le théorème d'inversion locale à la fonction auxiliaire $G : U \to \mathbb{R}^m \times R^n$ définie par

$$G(x,y) = (x, F(x,y)) \in \mathbb{R}^m \times \mathbb{R}^n \quad \text{pour } (x,y) \in U. \tag{6.9}$$

L'avantage est que maintenant l'image aussi est de dimension $N = m+n$. Evidemment G aussi est de classe C^1, puisqu'on a juste ajouté des coordonées de classe C^1. On commence par vérifier que

$$DG(a,b) : \mathbb{R}^m \times R^n \to \mathbb{R}^m \times \mathbb{R}^n \text{ est inversible.} \tag{6.10}$$

On regarde sa matrice P, qui est une matrice $(m+n) \times (m+n)$.

On commence par regarder la partie en haut de P, qui est une matrice à m colonnes, celle de l'application $(x,y) \to x \in \mathbb{R}^m$. On y trouve la matrice identique (à gauche) et la matrice nulle (à droite: ce sont les dérivée de la fonction $(x,y) \to x$ par rapport à y, qui sont nulles.

En bas, on retrouve la matrice de la différentielle $DF(a,b)$, qui est une matrice à n lignes (et toujours $m+n$ colonnes).

Pour (6.10), on doit calculer le déterminant de la matrice P, et on calcule en développant par rapport aux premières colonnes (à chaque fois, un seul terme non nul). On trouve que $\det(P) = \det(M)$, où M est la partie $n \times n$ qui reste en bas à droite, et qui est justement la matrice de la restriction de $DF(a,b)$ à $\mathbb{R}^n$, celle de (6.2). Donc (6.10) est exactement (6.2).

Le TIL donne un voisinage U_1 de (a,b), contenu dans U, et sur lequel G est un difféomorphisme. Notons $V_1 = G(U_1)$, un ouvert qui contient $G(a,b) = (a, F(a,b))$. Et encore $H = G^{-1}$, la réciproque de G, qui va de V_1 dans U_1.

Notons x' et w les variables dans V_1, et aussi $w_0 = F(a,b) \in \mathbb{R}^n$. Et aussi notons $H = (H_1, H_2)$ les deux coordonnées de H; en fait, par définition,

$$H_1(x', w) = x' \quad \text{pour } (x', w) \in V_1, \tag{6.11}$$

parce que comme $G(H(x',w)) = G \circ G^{-1}(x',w) = (x',w)$, en regardant la premiére coordonnée, on trouve bien que la première coordonnée de $H(x',w)$ est x'.

On commence à faire une dessin avec des flèches, et on note systématiqument $G = (G_1, G_2)$ et $H = (H_1, H_2)$, mais on se souviendra que G_1 et H_1 sont juste la première coordonnée (x ou x').

On note I_0 l'ensemble des $x \in \mathbb{R}^m$ tels que $(x, w_0) \in V_1$. Cet I_0 est un ouvert de $\mathbb{R}^m$ qui contient a puisque $G(a,b) = (a, F(a,b)) = (a, w_0)$ est en plein milieu de V_1. Et on définit φ sur I_0 par

$$\varphi(x) = H_2(x, w_0) \in \mathbb{R}^n. \tag{6.12}$$

C'est bien une application de classe C^1 définie dans I_0 (par composition). Voyons si elle fait ce qu'on veut. D'abord vérifions que le graphe de φ est bien contenu dans la ligne de niveau $F^{-1}(w_0)$. Pour $x \in I_0$, par définition $(x, w_0) \in V_1$, et comme $H = G^{-1}$, on trouve que

$$(x, w_0) = G \circ H(x, w_0). \tag{6.13}$$

On regarde la première coordonnée, et on voit que

$$x = G_1(H(x, w_0)) = H_1(x, w_0), \tag{6.14}$$

puisque G_1 se contente de recopier la première coordonnée. Ensuite on regarde la seconde coordonnée, et on voit que

$$w_0 = G_2(H(x,w_0)) = F(H_1(x,w_0), H_2(x,w_0)) = F(x, H_2(x,w_0)) = F(x, \varphi(x)) \tag{6.15}$$

par définition de φ. On aura vérifié au passage que tout est bien défini parce que $(x, w_0) \in V_1$. Donc tout point du graphe est bien un point de l'ensemble de niveau $F^{-1}(w_0) \subset U$.

On a encore besoin de la réciproque, et pour ceci on choisit notre voisinage final $I \times J \subset U_1$ de (a,b), encore plus petit.

D'abord, choisissons J, une petite boule de $\mathbb{R}^n$ centrée en b, assez petite pour que $\{a\} \times \overline{J} \subset U_1$. Ensuite, choisissons $I \subset I_0$, une boule centrée en a, et de rayon assez petit pour que $I \times J \subset U_1$ (prendre un rayon de boule plus petit que la moitié de la distance entre $\{a\} \times \overline{J}$ et le complémentaire de U_1), et aussi que $\varphi(x) \in J$ pour $x \in I$. C'est facile à assurer aussi, parce que φ et continue, et qu'au centre: $\varphi(a) = H_2(a, w_0) = b$ puisque $H(a, w_0) = (a, b)$ car $H = G^{-1}$ et $G(a, b) = (a, w_0)$. On se place dans $I \times J$, et on doit vérifier que dans cet ensemble le graphe de φ sur I et l'ensemble de niveau $Z_{w_0} = F^{-1}(w_0)$ coincident.

Notre dernière précaution, c'était pour s'assurer que $\varphi : I \to J$ comme on l'a demandé. Comme on a pris un petit J, il était logique de diminuer I en conséquence.

Ensuite, on a vu que tout point du graphe situé au-dessus de I est bien dans $I \times J$ et contenu dans Z_{w_0}.

Finalement on se donne $(x, y) \in Z_{w_0} \cap (I \times J)$ et on veut montrer que (x, y) est dans le graphe, donc que $y = \varphi(x)$. Par hypothèse, $(x, y) \in U_1$, et $G(x, y) = (x, F(x, y)) = (x, w_0)$ et par définition de l'inverse (sur V_1 qui contient (x, w_0)),

$$H(x, w_0) = G^{-1}(x, w_0) = G^{-1} \circ G(x, y) = (x, y). \tag{6.16}$$

On regarde la seconde coordonnée et on trouve que $H_2(x, w_0) = y$, donc par définition $\varphi(x) = y$ et (x, y) est dans le graphe. Ouf c'est fini! □

Fin du cours 8, le 15 mars 2022

S'il vous reste encore un peu d'énergie, vous pouvez constater que si l'on prend w assez proche de w_0, on peut encore définir $\varphi_w : I_2 \to \mathbb{R}^n$ comme en (6.12), mais avec $\varphi_w(x) = H_2(x, w)$. Et ceci donne une description locale de l'ensemble de niveau Z_w par le graphe de φ_w. C'est mieux que ce qu'on a dit, d'une part parce que l'on a aussi des descriptions des lignes de niveaux voisines de celle qui contient (a, b), mais aussi (surtout) parce qu'en plus la fonction φ_w qui donne Z_w dépend aussi de w de manière C^1. Et d'ailleurs on obtient aussi que les lignes de niveaux Z_w, w assez proches de w_0, ne sont pas vides!

Noter que si on veut seulement la description du Z_w qui passe par un point (a', b') proche de (a, b), il suffit d'appliquer le théorème des fonctions implicites en (a', b'); les hypothèses restent vérifiées pour (a', b') proche de (a, b), parce que le déterminant qu'on calcule est une fonction continue de (a', b'), donc reste non nul.

6.4 Calcul des dérivées partielles de φ dans le théorème 6.1

Paragraphe ajouté après coup en 2023, après m'être rendu compte qu'on s'en servait souvent. Plaçons-nous avec les hypothèses et notations du théorème 6.1. Donc on a trouvé une fonction φ telle que dans un voisinage de (a, b), $F(x, y) = F(a, b)$ si et seulement si $y = \varphi(x)$. Dit autrement, en notant $w = F(a, b)$ près de (a, b), notre ensemble de niveau $Z = \big\{(x, y)\,;\, F(x, y) = w\big\} = F^{-1}(w)$ est le graphe de φ. Ca c'est la partie difficile à démontrer.

Et maintenant on veut calculer les dérivées partielles $\frac{\partial \varphi_i}{\partial x_j}$. Juste en utilisant le fait que pour tout x proche de a, on a que $(x, \varphi(x)) \in Z_w$, donc que $F(x, \varphi(x)) = w$ (une constante).

Alors, puisqu'on sait que φ et F sont différentiables, on peut utiliser la formule de différentiabilité de la composée $x \mapsto (x, \varphi) \mapsto F(x, \varphi(x))$ pour dériver $F(x, \varphi(x))$ par rapport à n'importe quel x_j. On trouve que

$$\frac{\partial}{\partial x_j}\Big(F(x, \varphi(x))\Big) = 0 \tag{6.17}$$

(dans un voisinage de a) puisque notre application $x_j \mapsto F(x, \varphi(x))$ est constante. La notation est un peu bizarre; j'ai dû trop parler à des physiciens. Je devrais sans doute dire que je considère la fonction $\widetilde{F}(x) = F(x, \varphi(x))$ et que la dérivée $\frac{\partial}{\partial x_j}(\widetilde{F}(x))$ est nulle. Comme F est à valeurs dans $\mathbb{R}^n$, on trouve plus précisément que pour $1 \leq i \leq n$,

$$\frac{\partial}{\partial x_j}\Big(F_i(x, \varphi(x))\Big) = 0 \tag{6.18}$$

pour chaque coordonnée F_i de F. Mais en calculant la fonction composée,

$$\frac{\partial}{\partial x_j}\Big(F_i(x, \varphi(x))\Big) = \frac{\partial F_i}{\partial x_j}(x, \varphi(x)) + \sum_{\ell=1}^{n} \frac{\partial F_i}{\partial y_\ell}(x, \varphi(x))\, \frac{\partial \varphi_\ell}{\partial x_j}(x), \tag{6.19}$$

où l'on a additionné les effets des variations de chacune des coordonnées de $(x, \varphi(x))$ dans les variations de $\widetilde{F}(x)$. Pour ce qui est des variations de la première partie x, il y a un seul terme $\frac{\partial x_\ell}{\partial x_i}$ qui est non nul, celui avec $\ell = j$, et c'est pour ça qu'on a un seul terme au début. Conseil si vous n'aimez pas, faites le calcul avec une fonction F choisie au hasard. Bref à cause de (6.19), on trouve que pour tout choix de $j \in \{1, ..., m\}$ et de $i \in \{1, ..., n\}$, on a

$$\frac{\partial F_i}{\partial x_j}(x, \varphi(x)) + \sum_{\ell=1}^{n} \frac{\partial F_i}{\partial y_\ell}(x, \varphi(x))\, \frac{\partial \varphi_\ell}{\partial x_j}(x) = 0, \tag{6.20}$$

Travaillons avec j fixé. On se souvient qu'on veut calculer les n dérivées partielles $\frac{\partial \varphi_\ell}{\partial x_j}(x)$ en fonction des dérivées partielles de F. Or (6.20) donne n équations (une par indice i), donc si on a de la chance on peut calculer $\frac{\partial \varphi_\ell}{\partial x_j}(x)$ en fonction des $\frac{\partial F_i}{\partial x_j}(x, \varphi(x))$ et des $\frac{\partial F_i}{\partial y_\ell}(x, \varphi(x))$.

Bon, en fait on a de la chance: les coefficients importants $\frac{\partial F_i}{\partial y_\ell}(x, \varphi(x))$ (ceux qui correspondent au premier membre de l'équation linéaire) forment une matrice $n \times n$, qui est exactement au point (a, b) la matrice dont on a supposé en (6.2) qu'elle est inversible. Et, par continuité des dérivées partielles, plus le fait que l'ensemble des matrices inversible est ouvert, on trouve que cette matrice, même calcilée au point $(x, \varphi(x))$ comme ci-dessus, est inversible. Donc on peut bien calculer les $\frac{\partial \varphi_\ell}{\partial x_j}(x)$ en fonction des $\frac{\partial F_i}{\partial x_j}(x, \varphi(x))$ et des $\frac{\partial F_i}{\partial y_\ell}(x, \varphi(x))$.

Evidemment, il reste encore un peu d'implicite là-dedans, puisque dans la formule il reste l'argument $(x, \varphi(x))$ qu'on ne connaît pas. C'est pourquoi on vous demande seulement dans les exercices de calculer les $\frac{\partial \varphi_\ell}{\partial x_j}(a)$. Parce que cette fois, on sait que $\varphi(a) = b$, et on peut calculer toutes les dérivées partielles de F en (a, b), et donc on peut calculer les $\frac{\partial \varphi_\ell}{\partial x_j}(a)$. On ne se prive pas trop de vous le demander dans les exos.

6.5 Dernière remarque: condition pour pouvoir appliquer le théorème dans une base bien choisie

Dans des situations concrètes, il n'est pas clair du tout qu'il faille choisir les variables dans l'ordre d'apparition. Considèrons par exemple l'ensemble Z donné par

$$Z = \big\{(x_1, x_2, x_3) \in \mathbb{R}^3 \,;\, f(x_1, x_2, x_3) = w\big\}, \tag{6.21}$$

où f est à valeurs réelles, et $w \in \mathbb{R}$ est donné. on peut avoir à décider soi-même du choix de ce qu'on appelle les premières coordonnés (dans le théorème, c'était x_1 et x_2). On aura à faire ceci, par exemple, quand on aura à démontrer, dans le théorème 9.1, que l'ensemble de niveau S est localement un graphe.

Dans le cas de la sphère donnée par $x_1^2 + x_2^2 + x_3^2 = 1$, On voit bien que si on veut une représentation comme graphe près du point $M_0 = (1, 0, 0)$, on aura des ennuis avec le choix de $x = (x_1, x_2)$ et $y = x_3$. Le calcul donne d'ailleurs une dérivée par rapport à x_3 en M_0 égal à 0, qui ne vérifie donc pas les conditions du théorème. Par contre, pour représenter Z comme un graphe sur un des plans verticaux évidents, ça marche.

Reprenons le cas général de

$$Z = \big\{x \in \mathbb{R}^N \,;\, F(x) = w\big\}, \tag{6.22}$$

où $F : \mathbb{R}^N \to \mathbb{R}^n$, et $w \in \mathbb{R}^n$ est donné. On suppose que F est de classe C^1 au voisinage d'un point $z \in \mathbb{R}^N$ (tel que $F(z) = w$). Quand-est-ce que l'on peut trouver des coordonnées dans lesquelles on peut appliquer le théorème des fonctions implicites?

La réponse est assez simple: quand la différentielle $DF(z)$ (au point z, donc) est de rang n. Autrement dit, quand les N vecteurs $\nabla F_i(z)$, $1 \le i \le n$, engendrent $\mathbb{R}^n$. Comme $N > n$, on a quand même de bonne chances que ça arrive. Et si c'est le cas, cela veut dire que l'on peut trouver n vecteurs $V_i = \nabla F_i(z)$ parmi les N qui sont indépendants.

Alors on choisit comme seconde variable $y \in \mathbb{R}^n$ la variable obtenue comme combinaison de tous les x_i en question. Et donc comme variables initiales composant $x \in \mathbb{R}^{N-n}$, toutes les autres variables x_j restantes. Du coup la description de Z sera une écriture de chacune des x_i qui composent y en fonction des variables x_i restantes. La condition (6.2) est équivalente à l'indépendance des $V_i = \nabla F_i(z)$.

Exemple fait en cours: dans $\mathbb{R}^3$, l'intersection de deux sphères. Deux équations, une ligne de niveau (un cercle) qui peut se représenter comme le graphe d'une fonction de classe C^1 qui par exemple à x associe les deux autres coordonnées ($y(x)$ et $z(x)$) d'un point de Z. On doit parfois choisir l'axe de départ en fonction du point où l'on veut une bonne description dans un voisinage.

Encore une remarque. On a maintenant deux descriptions possibles d'une surface Z de codimension n (de dimension $N - n$); l'une comme graphe d'une fonction C^1, dans certaines coordonnées, l'autre comme ligne de niveau, comme en (6.22). Et on sait que quand DF est de rang n, ces deux descriptions sont (localement) équivalentes.

La description comme graphe donne l'existence d'un $(N-d)$-plan tangent P à Z (au point z considéré); on a vu ça au paragraphe 3.5 dans le cas où $n=1$, mais ça marche pareillement quand $n>1$. Donc on sait que P existe. Et on peut retrouver sa direction, dans la description de (6.22), à partir de $DF(z)$. Je le fais sans les détails. Ecrivons $P=z+P_0$, où maintenant P_0 est un plan vectoriel, et posons $L=DF(z)$. Par hypothèse, L est une application linéaire de rang n qui va de $\mathbb{R}^N$ dans $\mathbb{R}^n$. Le point central est que

$$L(v)=0 \text{ pour tout } v\in P_0. \tag{6.23}$$

Ceci se démontre directement en utilisant que tous le points de P (près de z) sont très proches de Z, et que F est constante sur Z. Ou, on peut aussi passer dans la représentation de Z comme un graphe, et calculer. Dans le calcul il y a le fait que pour $x\in\mathbb{R}^{N-n}=\mathbb{R}^m$, on peut calculer la différentielle de $F(x,\varphi(x))$ (qui est une fonction composée) en fonction de $DF(z)$ et de $D\varphi(x)$.

Maintenant, une autre manière de dire (6.23), c'est de dire que P_0 est contenu dans le noyau $Ker(L)$ de $L=DF(z)$. Or L est de rang n par hypothèse, donc par le théorème noyau-image (ou est-ce le théorème de la dimension), $Ker(L)$ est de dimension $N-d$. Et comme P_0 est de dimension $N-d$ aussi, on trouve que

$$P_0=Ker(L)=Ker(DF(z)). \tag{6.24}$$

En particulier, quand $n=1$, P_0 est de codimension 1, et comme $L(v)=\sum_{i=1}^N \frac{\partial F}{\partial x_i}(z)v_i$, on trouve que $v\in P_0$ si et seulement si $\sum_{i=1}^N \frac{\partial F}{\partial x_i}(z)v_i=0$ si et seulement si $v\perp\nabla F(z)$. Autrement dit, $\nabla F(z)$, qui est non nul par hypothèse (le rang de L est $n=1$), peut servir de vecteur orthogonal au plan tangent P.

En 2022: finalement j'ai fait un peu plus de détails en cours, en faisant un DL d'ordre 1 de $F(z+tv)$, en disant que $F(z+tv)-F(z)=o(t)$ quand v est un vecteur tangent (parce que $\text{dist}(z+tv,Z)=o(t)$ dans ce cas. Et on trouve (6.23) en comparant.

En 2023, j'ai juste commencé à dire qu'en supposant $a=b=0$ pour simplifier, alors à cause du théorème (TFI) et de ce qu'on a fait sur les graphes, on est déjà certain que notre ensemble de niveau Z a un plan tangent P de dimension m. Ensuite, on peut démontrer (je ne l'ai pas fait jusqu'au bout) que pour tout vecteur $v\in P$ (le plan tangent, qui passe par l'origine par hypothèse) on a bien (6.23). Une fois qu'on a vérifié ceci, on compte le nombre d'équations et on trouve que P est entièrement déterminé par cette équation. Voir ci-dessus mais les détails n'ont pas été faits en cours.

Fin du cours 8, le 14 mars 2023.

7 Optimisation-Recherche d'extrema

Ne voyez pas de signification trop profonde au mot "optimisation". On veut juste dire que dans plein de problèmes concrets, on aime bien choisir des solutions qui maximisent une fonction f des paramètres à déterminer, et où f est choisie en fonction du but à obtenir. Par exemple, f pourrait être la distance à une cible (réelle ou fictive), ou un gain.

Par exemple vous voulez choisir le site $x \in U \subset \mathbb{R}^2$ de construction d'une éoliene. Vous calculez une fonction $c(x)$ qui mesure le coût de l'instalation en x, une fonction $m(x)$ qui mesure le mécontentement des populations locales, une autre $i(x)$ qui mesure la quantité d'impôts locaux que vous pourrez en tirer, et vous choisissez x qui maximise la quantité $f(x) = i(x) - ac(x) - bm(x)$, où vous avez choisi $a > 0$ et $b > 0$ en fonction de vos inclinations générales.

Donc on aime bien trouver les extremas de fonctions f définies sur un ouvert U. Et le calcul (ou la recherche de ces points) sera plus facile si la fonction est différentiable. Voyons tout ceci plus en détails.

D'abord, les définitions.

Définition 7.1. *Soient E un ensemble et $f : E \to \mathbb{R}$ une fonction. On dit que f admet un maximum global en $x_0 \in E$ (ou par léger abus de notation, que x_0 est un maximum global pour f) quand $f(x_0) \geq f(x)$ pour tout $x \in E$.*

Si de plus E est muni d'une topologie, on dit que f admet un maximum local en $x_0 \in E$ (ou que x_0 est un maximum local pour f) quand il existe un voisinage V de x_0 dans E tel que $f(x_0) \geq f(x)$ pour tout $x \in E \cap V$.

Ecrire "pour tout $x \in E \cap V$" est un abus de précautions, puisque par définition V est inclus dans E. Le plus souvent, E est une partie de $\mathbb{R}^n$ et la topologie vient de $\mathbb{R}^n$. Dans ce cas, ou peut aussi dire que f admet un maximum local en $x_0 \in E$ quand il existe $r > 0$ tel que $f(x_0) \geq f(x)$ pour tout $x \in E \cap B(x_0, r)$.

Un minimum global (respectivement local) de f sur E est juste un minimum global (respectivement local) de $-f$ sur E.

Comme les théorèmes seront en gros les mêmes pour les deux, on gagnera parfois du temps en parlant d'extremum (pluriel extrema). Un extremum est un point de E qui est soit un minimum, soit un maximum (soit les deux, ce qui peut arriver si f est constante dans un voisinage de x_0).

7.1 Existence de minima globaux

On commence par rappeler le résultat le plus pratique à utiliser dans ce contexte: le théorème de Weierstrass.

Théorème 7.2. *Soit E un ensemble compact, et $f : E \to \mathbb{R}$ une fonction continue. Alors f est bornée et atteint ses bornes.*

Implicitement, E est muni d'une topologie; quand $E \subset \mathbb{R}^n$, ce sera toujours la topologie qui vient de la distance euclidienne (ou toute autre distance équivalente ou qui donne la même topologie), et donc dans ce cas E est un fermé borné de $\mathbb{R}^n$.

Donc f a au moins un maximum global et un minimum global.

Démonstration simple par Bolzano-Weierstrass (sur $\mathbb{R}$ si E n'est pas un espace métrique); voir plus haut.

Contre-exemple quand on oublie le mot fermé dans "fermé borné de $\mathbb{R}^n$": la fonction $||x||$ sur la boule unité ouverte de $\mathbb{R}^n$ n'admet pas de maximum.

Bon, justement, parfois E est un ouvert de $\mathbb{R}^n$ qui n'est pas compact, et on veut quand même utiliser le théorème de Weierstrass; c'est possible quand on arrive à démontrer que rien ne peut arriver près du bord.

Corollaire 7.3. *Soient U un espace topologique et $f : U \to \mathbb{R}$ une fonction continue. On suppose qu'il existe $x_0 \in U$ et un compact $K \subset U$ tels que $f(x) \leq f(x_0)$ pour tout $x \in U \setminus K$. Alors f est majorée dans U et même il existe $x_1 \in K$ tel que $f(x) \leq f(x_1)$ pour $x \in U$.*

Donc, f a au moins un maximum global et on peut le trouver dans K.

Exercice simple: formuler un résultat semblable pour trouver un minimum.

Démonstration: on commence par utiliser le théorème pour obtenir $x_1 \in K$ tel que $f(x_1) \geq f(x)$ pour tout $x \in K$. Puis on remarque que pour $x \in U \setminus K$, on a encore mieux, car $f(x) \leq f(x_0) \leq f(x_1)$ (puisque $x_0 \in K$). Evidemment, f est majorée sur U puisque $f(x) \leq f(x_1) < +\infty$ sur U. □

Un moyen simple d'obtenir que $f(x) \leq c$ hors d'un compact K est de demander que la limite de $f(x)$, quand x tend vers l'infini ou le bord de U, existe et est $< c$. Voyons juste des exemples.

Exemple 1. Soit P un polynome de degré 26 sur $\mathbb{R}$, tel que le coefficient de x^{26} de P est strictement positif. Alors P a au moins un minimum global sur $\mathbb{R}$.

On a pris 26 parce que 26 est pair, et que donc $\lim_{x\to\pm\infty} P(x) = +\infty$. On choisi $x_0 \in \mathbb{R}$ au hasard, puis $r > 0$ tel que $f(x) \geq f(x_0) + 1$ pour tout $x \in \mathbb{R} \setminus [-r, r]$. Puis on applique le corollaire.

Exemple 2 (exercice). On se donne un polynôme P non nul à coefficients complexes sur $\mathbb{C}$. Montrer qu'il existe $z_0 \in \mathbb{C}$ tel que $|P(z_0)| \leq |P(z)|$ pour $z \in \mathbb{C}$. C'est la première étape pour montrer (D'alembert) que P a au moins une racine (sinon, $|P(z_0)| > 0$ et on regarde un développement de $P(z)$ autour de z_0 pour obtenir une contradiction).

Exemple 3. La fonction $f : \mathbb{R}^n \to \mathbb{R}$, définie par $f(x) = \frac{x_1 \sin(x_1) + x_2 \cos(x_2)}{||x||^2}$ admet (au moins) un maximum et un minimum (globaux) sur $\mathbb{R}^n$. En effet, $\lim_{||x||\to+\infty} f(x) = 0$, donc il existe $r > 0$ tel que $|f(x)| \leq 10^{-10}$ pour $x \in \mathbb{R}^n$ tel que $||x|| \geq r$. On peut bien entendu choisir $r \geq 10$. Prenons alors $K = \overline{B}(0, r)$. Il ne reste plus qu'à trouver $x_0 \in K$ tel que $f(x_0) \geq 10^{-10}$ et $x_0' \in K$ tel que $f(x_0') \leq -10^{-10}$. Je vous laisse faire.

Exemple 4. Soit $U \subset \mathbb{R}^n$ un ouvert borné de $\mathbb{R}^n$ (vous pouvez penser à $U = B(0, 1)$ si vous voulez, mais bien sûr il sera facile de deviner qui doit être x_0). La fonction $x \mapsto f(x) = \operatorname{dist}(x, \mathbb{R}^n \setminus U)$ est bornée et atteint son maximum en (au moins) un point de U.

En effet, on prend un point $x_0 \in U$, puis on pose $\varepsilon = f(x_0) = \frac{1}{2} \operatorname{dist}(x_0, \mathbb{R}^n \setminus U) > 0$, et il ne reste plus qu'à vérifier que $K = \left\{x \in U\,;\ \operatorname{dist}(x, \mathbb{R}^n \setminus U) \geq \varepsilon\right\}$ est un compact puis d'appliquer le corollaire. Mais cet ensemble est borné (car U est borné), donc il suffit de voir qu'il est fermé. Soit donc x un point de l'adhérence de K; il existe une suite $\{x_k\}$ dans x qui tend vers x. Comme $x_k \in K$, $\operatorname{dist}(x_k, \mathbb{R}^n \setminus U) \geq \varepsilon$ et, comme la fonction distance à $\mathbb{R}^n \setminus U$ est Lipschitzienne donc continue (vérifiez!) on a bien $\operatorname{dist}(x, \partial U) \geq \varepsilon$. □

Notons encore que si on oublie de demander que U est borné, il se peut que f ne soit pas bornée: prendre $U = \mathbb{R}^n \setminus \overline{B}(0,1)$.

7.2 Points critiques

L'inconvénient des théorèmes d'existence ci-dessus est qu'ils ne disent pas où sont les extrema. La seconde chose de base à savoir dans ce sujet est que si la fonction f est différentiable, on a une condition nécessaire simple, à savoir que $Df(x_0) = 0$. Voyons l'énoncé.

Théorème 7.4. *Soient $U \subset \mathbb{R}^n$ un ouvert, $x_0 \in U$, et $f : U \to \mathbb{R}$. Supposons que f a un extremum local en x_0 et est différentiable en x_0. Alors $Df(x_0) = 0$.*

On a demandé que U soit ouvert, et il aurait suffi que x_0 soit un point intérieur de U, autrement dit que U contient une petite boule centrée en x_0. C'est nécessaire si on veut définir correctement (et utiliser) $Df(x_0)$. Si par exemple x_0 est au bord de U, il y a des moyens de finasser, mais en tout cas les choses ne sont pas si simples. Voir quand même le dernier chapitre (sur les extrema liés), puisque c'est de cela qu'il s'agit!

Le statut de théorème n'est pas dû à la difficulté du résultat, mais plutôt à son utilisation constante.

Vous connaissez ce résultat pour $f : \mathbb{R} \to \mathbb{R}$. On va juste déduire le théorème de ce cas particulier. On se donne U, x_0, et f comme dans le théorème, et on suppose pour simplifier la discussion que f a un minimum local en x_0. On fixe n'importe quel vecteur unitaire $v \in \mathbb{R}^n$, et on considère la même fonction f_v qui a servi à définir les dérivées directionnelles, définie par $f_v(t) = f(x_0 + tv)$. Par hypothèse, f_v est bien définie sur un voisinage de 0, elle est dérivable en 0 parce que f est différentiable en x_0 et donc a une dérivée directionnelle dans la direction v. De plus on a vu que $f_v'(0) = \partial_v f(x_0) = Df(x_0)(v)$. Voir la proposition 3.2.

Or f_v a un minimum local en 0 puisque f a un minimum local en x_0, donc par l'argument usuel (développement limité d'ordre 1 de f_v au voisinage de 0), $f_v'(0) = 0$. Donc $Df(x_0)(v) = 0$ pour tout v unitaire (on a pris unitaire, mais en fait on n'en a pas besoin). Donc $Df(x_0) = 0$ (si vous voulez garder "unitaire", noter que pour tout $v \neq 0$, $Df(x_0)(v) = ||v||Df(x_0)(v/||v||) = 0$. □

D'où la définition suivante, puisque la notion est utile.

Définition 7.5. *Un <u>point critique</u> de $f : U \to \mathbb{R}$ est un point x_0 intérieur à U, en lequel f est différentiable, et tel que $Df(x_0) = 0$. Ou donc $\nabla f(x_0) = 0$.*

J'ai ajouté intérieur pour le cas où vous autoriseriez U non ouvert. Bien entendu, le théorème ci-dessus est une condition nécessaire mais non suffisante. Il y a des tas de points critiques qui ne sont pas des extrema locaux. Voici les deux les plus emblémtiques.

Exemple 1. Sur $\mathbb{R}$, 0 est un point critique de $x \to x^3$ qui n'est ni minimum local ni maximum local.

Exemple 2. Sur $\mathbb{R}^2$, la forme la plus simple d'un point critique est la suivante: prendre $f(x,y) = ax^2 + by^2$, où par exemple (si l'on ne veut pas d'extremum) $a > 0$ et $b < 0$. L'origine

est bien un point critique, mais pas un extremum local: voir le dessin du graphe en forme de selle de cheval (ou de col de montagne). Evidemment on a de tels exemples aussi dans $\mathbb{R}^n$.
Fin du cours 9 du 22 (=23) Mars 2022.

En attendant, même si ce n'est pas une CNS, le théorème 7.4 permet de bien diminuer la liste des suspects.

7.3 Seconde dérivée

Pour décider de si un point critique est un extremum local, on va faire appel à la seconde dérivée. Ce sera un peu plus lourd, mais parfois bien utile. Mais cette fois encore on aura des conditions plus précises mais toujours pas nécessaires et suffisantes.

Pensez à ce qu'on fait en dimension 1: si $f'(x_0) = 0$ et si la dérivée seconde $f''(x_0)$ est strictement positive (négative), on a un minimum local (un maximum local), et même strict (atteint seulement en x_0). Par contre si $f''(x_0) = 0$, comme pour $x \to x^3$ en 0, on ne sait toujours pas dire a priori.

On va faire pareil dans $U \subset \mathbb{R}^n$, mais il faudra tenir compte qu'on a maintenant plusieurs directions où se déplacer (comme dans l'exemple 2 plus haut).

On se donne des notations pour le prochain théorème.

On se donne un ouvert $U \subset \mathbb{R}^n$ et un point $a \in U$.

On suppose que $f : U \to \mathbb{R}$ est de classe C^2 dans un voisinage de a; c'est sans doute un peu trop, mais on veut appliquer la formule de Taylor à l'ordre 2 et on ne veut pas d'ennui. On suppose que a est un point critique, c'est-à-dire que

$$\nabla f(a) = 0. \tag{7.1}$$

Et on note $D^2_a f$ la seconde dérivée de f au point a. C'est une forme bilinéaire sur $\mathbb{R}^n$ qui est définie par

$$D^2_a f(h,k) = \partial_h(\partial_k f)(a) = \sum_{i=1}^n \sum_{j=1}^n h_i k_j \frac{\partial^2 f}{\partial x_i \partial x_j}(a), \tag{7.2}$$

où dans la première partie j'ai mis deux dérivées directionnelles successives de f, calculées au point a. Une autre manière de dire est de définir la matrice Hessienne $H = H_f(a)$ (de f au point a), qui est la matrice carrée symétrique définie par ses coefficients

$$H_{i,j} = \frac{\partial^2 f}{\partial x_i \partial x_j}(a), \quad 1 \leq i,j \leq n, \tag{7.3}$$

puis de définir $D^2_a f$ par

$$D^2_a f(h,k) = k^t H h \tag{7.4}$$

pour $h, k \in \mathbb{R}^n$, où pour simplifier on a identifié h et k avec des vecteurs colonnes. On a noté k^t le transposé de k, qui est donc une matrice ligne. Noter que puisque H est symétrique, on a aussi $D^2_a f(h,k) = h^t H k = D^2_a f(k,h)$.

7.4 Intermède sur les matrices et formes bilinéaires symétriques

Avant de passer aux fonctions et dérivées (différentielles) d'ordre 2, on va maintenant dire quelques mots sur les formes bilinéaires, formes quadratiques, et diagonalisation des matrices symétriques. En particulier, il est bon de connaître le résultat suivant.

Proposition 7.6. *Soit $H \in M_n(\mathbb{R})$ une matrice symétrique à coefficients réels. Alors il existe une base ortonormée $(e_1, \dots, e_n)$ de $\mathbb{R}^n$ telle que la matrice de l'application linéaire associée à H dans cette base est une matrice diagonale.*

L'application linéaire associée à H est l'application qui envoie $h \in \mathbb{R}^n$ sur Hh, où l'on a encore identifié $\mathbb{R}^n$ à l'ensemble des matrices colonne. Et cet opérateur (on va l'appler L pour ne pas confondre) est auto-adjoint, au sens où

$$\langle L(u), v\rangle = \langle u, L(v)\rangle \quad \text{pour tous } u, v \in \mathbb{R}^n. \tag{7.5}$$

Vérification facile, puisque en écrivant le produit scalaire en termes de matrices,

$$\langle L(u), v\rangle = \langle Hu, v\rangle = (Hu)^t v = u^t H^t v = u^t H v = u^t L(v) = \langle u, L(v)$$

parce que H est symétrique. En fait, pour simplifier la démonstration de la proposition par récurrence, on va aussi démontrer en même temps le résultat équivalent suivant.

Proposition 7.7. *Soient E un espace de Hilbert de dimension finie sur $\mathbb{R}$ et $L : E \to E$ une application linéaire auto-adjointe (comme en (7.5)). Alors il existe une base ortonormée $(e_1, \dots, e_n)$ de E telle que la matrice de L dans cette base est une matrice diagonale.*

La seconde proposition implique la première: si H est une matrice symétrique, on a vu que l'application L associée est auto-adjointe, donc diagonalisable dans une BO de $\mathbb{R}^n$; c'est ce qu'on voulait. Dans l'autre direction, si on a la première propostion on peut l'applique à la matrice de L dans n'importe quelle BO de E; c'est facile de voir que cette matrice H et symétrique (mais comme on n'aura pas besoin de cette direction, je le laisse en exercice), donc on a une BO de vecteurs colonnes où H se diagonalise, et ceci donne une BO de vecteurs de E où L est diagonale.

Maintenant démontrons la seconde proposition. Soit L comme dans l'éoncé. On commence par vérifier que

$$\text{toutes les valeurs propres de } L \text{ sont réelles.} \tag{7.6}$$

Pour ceci on complexifie E et L. Pour E, on pose $\widehat{E} = \big\{v + iw\,;\, v, w \in E\big\}$. Pas la peine de définir un produit scalaire sur $\widehat{E}$. Puis on définit l'application linéaire $\widehat{L} : \widehat{E} \to \widehat{E}$ par $L(v + iw) = L(v) + iL(w)$. Je vous laisse vérifier que cette application est $\mathbb{C}$-linéaire sur $\widehat{E}$.

On prend une BO de E; ceci donne une base de $\widehat{E}$ aussi (vérification directe, l'indépendance se faisant à la main en prenant partie réelle et partie imaginaire d'une combinaison linéaire des vecteurs de base), et la matrice de $\widehat{L}$ dans cette base est la même que la matrice M de L (par définition de la matrice d'une application linéaire).

Supposons que $\lambda = a + ib$ est valeur propre de L (donc de M), avec $b \neq 0$. Alors $\det(L - \lambda I) = 0$ (par définition), donc $L - \lambda I$ n'est pas inversible, donc $L - \lambda I$ a un noyau (dans $\widetilde{E}$): il existe un vecteur non nul $\xi = v + iw \in \widehat{E}$ tel que $\widehat{L}(\xi) = \lambda\xi$. Alors

$$L(v) + iL(w) = \widehat{L}(v + iw) = \widehat{L}(\xi) = \lambda\xi = (a + ib)(v + iw) = av - bw + i(aw + bv).$$

En prenant les parties réelle et imaginaire, $L(v) = av - bw$ et $L(w) = aw + bv$. On déduit du premier que $\langle L(v), w\rangle = \langle av - bw, w\rangle$ Et du second que $\langle L(w), v\rangle = \langle aw + bv, v\rangle$. Mais les deux sont égaux car L est auto-adjoint (ouf, on a utilisé l'hypothèse!) donc après simplification $-b\langle w, w\rangle = b\langle v, v\rangle$. Comme $b \neq 0$, on simplifie et on trouve $||v||^2 + ||w^2|| = 0$. Ceci ne va pas non plus, on a pris $\xi \neq 0$. Donc on a (7.6).

Maintenant on va vérifier que si V est un espace vectoriel stable par L, alors son orthogonal aussi est stable par L. Soit donc V tel que $L(v) \in V$ pour tout $v \in V$. Soit $y \in V^\perp$ et vérifions que $L(y) \in V^\perp$, donc que $\langle L(y), v\rangle = 0$ pour tout $v \in V$. On écrit juste que

$$\langle L(y), v\rangle = \langle y, L(v)\rangle = 0 \tag{7.7}$$

parce que L est symétrique (cf (7.5)) et que $L(v) \in V$.

Pour finir la démonstration, on doit maintenant construire notre base orthonormale $(e_1, \ldots, e_n)$ de E où L est diagonale.

Comme le polynôme caractéristique n'a que des racines réelles, il en a au moins une, donc on peut trouver un premier vecteur propre unitaire e_1. Maintenant, puisque $V = Vect(e_1)$ est stable par L, son orthogonal aussi est stable. De plus, la restriction de L à $V^\perp$ est également auto-adjointe ((7.5) reste vrai pour $u, v \in V^\perp$), donc on peut recommencer. Ou décider de démontrer la proposition par récurrence sur la dimension de E et à se stade dire que par hypothèse de récurrence, L se diagonalise bien sur $V^\perp$, donc on a une BO de $V^\perp$ composée de vecteurs propres. On ajoute e_1 à cette base et on trouve la BO de E cherchée. □

Bon, revenons doucement aux dérivées seconde, mais avec un petit passage par les formes bilinéaire symétriques et de formes quadratiques. On a vu (sur l'xemple des dérivées secondes) que si H est une matrice symétrique $n \times n$ à coefficients réels, on peut défnir une forme bilinéaire symétrique sur $\mathbb{R}^n \times \mathbb{R}^n$, que je noterai B, par

$$B(u, v) = v^t Hu = \langle v, Hu\rangle \text{ pour } u, v \in \mathbb{R}^n. \tag{7.8}$$

Bilinéaire veut dire linéaire par rapport à chaque variable (quand on fixe l'autre) et symétrique signifie que $B(u, v) = B(v, u)$ pour $u, v \in \mathbb{R}^n$ et vient de ce que $H^t = H$. Ensuite, on peut aussi définir une forme quadratique Q, sur $\mathbb{R}^n$, par

$$Q(u) = B(u, u) = u^t Hu = \langle v, Hu\rangle. \tag{7.9}$$

Donc Q n'apporte rien de vraiment nouveau par rapport à Q, et par ailleurs si on connait Q on peut retrouver B, car le calcul donne $Q(u + v) = B(u + v, u + v) = B(u, u) + 2B(u, v) + B(v, v) = Q(u) + Q(v) + 2B(u, v)$, donc $2B(u, v) = Q(u + v) - Q(u) - Q(v)$.

Revenons à la proposition 7.6, qui dit qu'on peut trouver une base orthonormée $(e_1, \dots, e_n)$ de $\mathbb{R}^n$ composée de vecteurs propres pour la multiplication par H. Autrement dit, telle que

$$He_i = d_i e_i \quad \text{pour } 1 \leq i \leq n, \tag{7.10}$$

où les d_i sont des nombres réels. Vérifions qu'alors

$$B(e_i, e_j) = d_i \delta_{i,j} \text{ pour } 1 \leq i, j \leq n, \tag{7.11}$$

où $\delta_{i,j}$ est le symbole de Kronecker qui vaut 1 quand $i = j$ et 0 sinon. Par bilinéarité, ceci permet de retrouver les valeurs de $B(u, v)$ pour tout choix de $u, v \in \mathbb{R}^n$, puisque si $u = \sum u_i e_i$ et $v = \sum v_j e_j$, on trouve en développant (puis en éliminant les termes où $i \neq j$ qui sont nuls)

$$B(u, v) = \sum_{i=1}^{n} d_i x_i y_i \quad \text{et donc} \quad Q(u) = B(u, u) = \sum_{i=1}^{n} d_i x_i^2 = \sum_{i=1}^{n} d_i \langle u, e_i \rangle^2. \tag{7.12}$$

Bref, on a trouvé une BO $(e_1, \dots, e_n)$ de $\mathbb{R}^n$ dans laquelle l'expression de $B(u, v)$ ou $Q(u)$ est particulièrement simple, et ce sera bien pratique pour étudier la fonction $u \to Q(u)$.

Commentaires supplémentaires sur l'algèbre linéaire et les matrices de passage. Je me suis débrouillé pour ne pas en avoir besoin, mais en fait c'est bien pratique de savoir comment les changements de base affectent l'expression des applications linéaires et applications bilinéaires. Je crois que c'est juste pour votre culture: on va pouvoir se débrouiller très bien avec (7.10) et (7.12).

On se donne un espace vectoriel E de dimension finie, avec une ancienne base (disons, $(o_1, \dots, o_n)$) et une nouvelle base (disons, $(e_1, \dots, e_n)$ comme ci-dessus). Chaque vecteur $U \in E$ a des coordonnées u_i dans l'ancienne base x_i dans la nouvelle. Et je note aussi u le vecteur colonne des u_i et x le vecteur colonne des x_i. On note P la matrice de passage, dont la colonne j est le vecteur des (anciennes) coordonnées de e_j. Et on veut vérifier pour commencer que la relation entre u et x est

$$u = Px, \quad \text{ou de manière équivalente, } x = P^{-1}u. \tag{7.13}$$

(Mnémotechnique: se rappeler qu'on n'a rien sans rien, donc calculer les nouvelles coordonnées inclus le calcul de P^{-1}). C'est assez facile; on peut se contenter de vérifier que la formule de gauche vaut quand $U = e_j$, donc x est un vecteur avec une seule coordonnée non nulle (à la ligne j), et alors Px est bien le vecteur colonne des coordonnées de e_j.

Comment calcule-t-on la matrice de l'application linéaire L dans la nouvelle base, à partir de la matrice M de L dans l'ancienne? Avec les notations précédentes, les anciennes coordonnées de $L(U)$ sont Mu, donc les nouvelles sont $P^{-1}Mu = P^{-1}MPx$, et donc

$$\text{la matrice de } L \text{ dans la nouvelle base est } P^{-1}MP \tag{7.14}$$

(on utilise systématiquement le fait que la matrice détermine L, et réciproqement). Donc par exemple, dans la situation des propositions ci-dessus, $D = P^{-1}HP$, où D est la matrice diagonale ci-desus.

Maintenant comment calcule-t-on la matrice de l'application linéaire B dans la nouvelle base, à partir de la matrice H de B dans l'ancienne base? On se donne donc deux vecteurs U, V, on note u, v les anciennes coordonnées et x, y les nouvelles, on note que $u = Px$ et $v = Py$, et on calcule

$$B(U,V) = \langle Hu, v\rangle = v^t Hu = (Py)^t H(Px) = y^t(P^t HP)x$$

donc

$$\text{la matrice de } B \text{ dans la nouvelle base est } P^t MP \tag{7.15}$$

C'est pas tout-à-fait pareil, sauf qu'en fait, dans la proposition, on n'a pas pris n'importe quelle base $(e_1, \dots, e_n)$, mais une base orthonormée. Et dans ce cas il se trouve que P est une matrice orthogonale (c'est-à-dire que $P^tP = PP^t = I$ (la matrice identique) et donc que $P^{-1} = P^t$.

En effet, calculons le ceoefficient $c_{i,j}$ de P^tP. C' en fait le produit scalaire de la ligne i de P^t avec la colonne j de P. Mais la ligne i de P^t, c'est la colonne i de P, donc on est en train de calculer $\langle e_i, e_j\rangle = \delta_{i,j}$ (puisque $(e_1, \dots, e_n)$ est ue BO). Donc $P^tP = I$. Comme il s'agit de matrices carrées, on en déduit bien que P est inversible, que son inverse est P^{-1}, et on en déduit que $PP^t = I$ aussi.

Ce qui fait que comme on l'a calculé différemment, la matrice de l'opérateur de multiplication par H fans la base $(e_1, \dots, e_n)$ set truve être aussi matrice de l'opérateur bilinéaire B associé à H (toujours das cette même base), à savoir la matrice diagonale de coefficients d_j, $1 \le j \le n$.

Retour à la forme quadratique Q. On utilise le vocabulaire suivant sur les formes quadratiques.

Définition 7.8. *Soit Q la forme quadratique associée à la forme bilinéaire symétrique B sur $\mathbb{R}^n \times \mathbb{R}^n$ (donc $Q(u) = B(u,u)$). On note $d_1, \dots d_n$ les coefficients in (7.11). Et on dit que*

$$Q \text{ est définie positive quand } d_j > 0 \text{ pour tout } j, \tag{7.16}$$

$$Q \text{ est définie négative quand } d_j < 0 \text{ pour tout } j, \tag{7.17}$$

$$Q \text{ est dégénérée quand } d_j = 0 \text{ pour au moins un } j. \tag{7.18}$$

Il est important de noter que ces notions ne dépendent pas de la base $(e_1, \dots, e_n)$ choisie. En fait, on peut calculer directement sur la matrice diagonale D de (la multiplication par H) que le polynôme caractéristique de H, donc de D, est $P(\lambda) = \det(D - \lambda I) = \prod_{j=1}^n (d_j - \lambda)$, dont les racines sont les d_j. De sorte que H détermine le polynôme caractéristique, qui détermine la liste des d_j (je veux dire, à permutation près).

Ces propriétés se caractérisent donc bien quand on a calculé le polynôme caractéristique de H (mais parfois le calcul est désagréable). On vérifie aisément que Q est définie positive si et seulement si

$$Q(u) > 0 \quad \text{pour tout } u \in \mathbb{R}^n \setminus \{0\} \tag{7.19}$$

et aussi si et seulement si il existe $c > 0$ (en fait le plus petit des d_j) tel que

$$Q(u) \geq c||u||^2 \quad \text{pour tout } u \in \mathbb{R}^n. \tag{7.20}$$

Pareil en changent des signes pour définie négative.

Un peu plus généralement, on définit la signature de Q comme étant le couple (p, q), où p est le nombre de $d_i > 0$ (comptés avec multiplicité) et q le nombre de $d_i < 0$ (comptés avec multiplicité). Donc Q est dégénérée si et seulement si $p + q < n$.

On dit aussi que Q est positive (resp., négative) quand tous les d_j sont positifs ou nuls (resp., négatifs ou nuls). Mais attention, souvent la bonne notion est définie positive (ou définie négative), qui inclus la non-dégénérescence.

Et donc souvenons nous que dès la dimension $n = 2$, il y a des formes quadratiques non dégénérées, comme $Q(x, y) = x^2 - y^2$ qui ne sont ni positives ni négatives!

7.5 Retour à la dérivée seconde, Taylor, et les extrema locaux

On revient à f, définie et de classe C^2 sur $U \subset \mathbb{R}^n$, au point $a \in U$, à sa dérivée seconde $D^2_a f$, et sa matrice Hessienne H_a au point a.

On revient à la formule de Taylor démontrée au théorème 3.8, que je recopie directement avec juste le changement de notation (x devient a et h devient u).

Théorème 7.9. *Soit $f : U \to \mathbb{R}$ de classe C^2 dans un voisinage de $a \in U \subset \mathbb{R}^n$. Alors pour $u = (u_1, \ldots, u_n) \in \mathbb{R}^n$ petit, on a*

$$f(a + u) = f(a) + Df_a(u) + \frac{1}{2}Q(u) + o(||u||^2), \tag{7.21}$$

ou $Df_a = Df(a)$ est la différentielle de f en a et Q est la forme quadratique associée à la seconde dérivée de f en a, donnée par

$$\begin{aligned} Q(u) &= \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\partial^2 f}{\partial x_i \partial x_j}(a) u_i u_j \\ &= \sum_{i=1}^{n} \frac{\partial^2 f}{\partial x_i^2}(a) u_i^2 + 2 \sum_{1 \leq i < j \leq n} \frac{\partial^2 f}{\partial x_i \partial x_j}(x) u_i u_j. \end{aligned} \tag{7.22}$$

Avec les notations ci-dessus, Q est la forme quadratique associée à la forme bilinéaire $B(u, v) = \langle Hu, v\rangle$, où $H = H_a$ est la matrice Hessienne de f en a. Et on est prêt à tirer les conclusions qui sont maintenant faciles.

Corollaire 7.10. *Soient U, a, f comme ci-dessus. Supposons en outre que a est un point critique (que $\nabla f(a) = 0$). Alors*

- *Si Q est définie positive, alors f a un minimum local (strict) en a;*
- *Si Q est définie négative, alors f a un maximum local (strict) en a;*
- *Si au moins H a au moins une valeur propre $d_i > 0$ et au moins une valeur propre $d_i < 0$, alors f n'a ni minimum local en a, ni maximum local en a.*

Minimum local strict signifie qu'il existe $r > 0$ tel que $f(x) > f(a)$ pour $x \in B(a, r) \setminus \{a\}$. Pareil pour mximum local strict, mais on demande que $f(x) < f(a)$ pour $x \in B(a, r) \setminus \{a\}$.

Noter qu'il reste des cas où l'on ne peut pas conclure: quand Q est positive (ou négative), non définie positive ou négative. C'est le das déjà en dimension 1, par exemple à l'origine, pour $f(x) = x^3$, ou $f(x) = x^4$.

On rappelle que définie positive signifie que tous le d_i de la description ci-dessus sont > 0.

Et le cas typique du troisième cas est celui de la selle de cheval (ou du col) $f(x, y) = 2x^2 - 3y^2$, par exemple.

Démonstration facile: dans le premier cas, on a que $Q(u) \geq c||u||^2$ à cause de (7.20), et comme le terme d'erreur est au plus $\frac{c}{2}||u||^2$ pour u petit (et que le terme d'ordre 1 disparait), c'est le second terme qui gagne et qui donne le signe.

Dans le troisième cas, on trouve des u petits (prendre des multiples d'un u donné, donc en gros se restreindre à une droite dans U) où $Q(u) \geq c||u||^2$ donc $f(a+u) > f(u)$ par le même raisonnement que ci-dessus, et d'autres où au contraire $Q(u) \leq -c||u||^2$ donc $f(a+u) < f(u)$, ce qui exclut les deux types d'extrema. □

Dernière remarque. En principe on doit calculer le Hessien $H = H_a$, puis son polynôme caractéristique, puis regarder les signes des racines. Parfois on peut gagner un peu de temps de calcul. Par exemple, en dimension $n = 2$, quand on a calculé le polynôme caractéristique $P(\lambda) = \lambda^2 - 2b\lambda + d$, où en fait on peut vérifier que $d = \det(H)$ (la valeur de $P(\lambda)$ en 0) et $2b$ est en fait la trace de la matrice H, on peut parfois en déduire des infos sans résoudre, puisque d est le produit des racines et b leur somme. Donc par exemple $d < 0$ signifie racines de signes opposés, donc troisième cas, $d = 0$ signifie dégénéré (et on ne sait pas trop), et $d > 0$ signifie définie positive ou négative donc extremum local.

En dimension 24, $d < 0$ signifie toujours troisième cas, mais $d > 0$ donne encore plein de choix de signatures.

8 Convexité

Un sujet bien plus vaste que ce qu'on traitera ici. On s'intéressera surtout, après les définitions, aux extrema. De ce point de vue la convexité est une alternative à la compacité (quand on travaille en dimension infinie, ce qui ne sera pas le cas), et d'ailleurs même dans

le cas de $\mathbb{R}^n$, la convexité donne souvent l'unicité des minima et des moyens de les trouver raidement.

En tout cas la notion est importante, d'où le chapitre spécial!

8.1 Ensembles convexes

ici le 6 avril 2021, si j'ai bien noté

Définition 8.1. *Soit X un espace vectoriel et $E \subset X$ un ensemble. On dit que E est convexe quand $[x, y] \subset E$ pour tout choix de $x, y \in E$.*

Et comme plus haut $[x, y] = \{x + t(y - x)\,;\, 0 \leq t \leq 1\}$ est le segment entre x et y.

Observation sans démonstration (mais qui est parfois bien pratique). Quand E est une partie fermée de $\mathbb{R}^n$ (je dis $\mathbb{R}^n$ pour éviter toute discussion sur quelle topologie on met sur X et E, mais l'important est que $\{t \in [0, 1]\,;\, x + t(y - x) \in E\}$ soit fermé dans $[0, 1]$), si on sait que $(x + y)/2 \in E$ pour tout choix de $x, y \in E$, alors E est convexe.

Si on oublie "fermé", ceci ne marche pas: penser à $\mathbb{Q} \subset \mathbb{R}$ qui n'est pas convexe, mais stable par moyennes.

Dans l'autre sens, notons que si $E \subset X$ est convexe, alors il est aussi stable par prise de barycentres, au sens où pour tout entier $k > 0$, si $x_1, \dots x_k \in E$, et si les nombres $t_i \in [0, 1]$ sont tels que $\sum_{i=1}^k t_i = 1$, alors $\sum_{i=1}^k t_i x_i \in X$.

Démonstration facile par récurrence sur k. Quand $k = 0$ c'est la définition (noter que $t_1 x + t_2 y = (1 - t_2)x + t_2 y = x + t_2(y - x) \in E$ puisque $t_2 \in [0, 1]$ et $x, y \in E$). Ensuite, si la propriété est vraie pour k, et si on se donne $x = \sum_{i=1}^{k+1} t_i x_i$ comme ci-dessus, on distingue deux cas. Si $t_{k+1} = 1$, $x = x_{k+1} \in E$ et on est content. Autrement, on écrit que $x = t_{k+1}x_{k+1} + \sum_{i=1}^k t_i x_i = t_{k+1}x_{k+1} + (1 - t_{k+1})y$, où $y = (1 - t_{k+1})^{-1} \sum_{i=1}^k t_i x_i$, qui est bien dans E car les coefficients $(1 - t_{k+1})^{-1} t_i$ sont positifs et leur somme est 1.

Définition 8.2. *Soit X un espace vectoriel et $E \subset X$ un ensemble. L'enveloppe convexe de E est le plus petit ensemble convexe $conv(E)$ qui contient E. L'enveloppe convexe fermée de E est le plus petit ensemble convexe fermé $\overline{conv}(E)$ qui contient E.*

Les deux existent; en fait, $conv(E)$ est l'intersection de tous les ensembles convexes qui contiennent E (qui est convexe parce qu'une intersection de convexes est convexe), et pareillement $\overline{conv}(E)$ est l'intersection de tous les ensembles convexes fermés qui contiennent E. Il peut arriver que E soit fermé mais que $conv(E)$ ne le soit pas (partie du plan au-dessus du graphe de $x \to (1 + x^2)^{-1}$.

Dans $\mathbb{R}$, les ensembles convexes sont exactement les intervalles. C'est facile à vérifier (et on trouve directement les bornes comme inf et sup de l'ensemble E. Il reste juste à regarder si la borne est dans l'ensemble ou non.

Exemple de base. Soient X un espace vectoriel et N une norme sur X. Alors la boule unité $B(0, 1) = \{x \in X\,;\, N(x) < 1\}$ et la boule unité fermée $\overline{B}(0, 1) = \{x \in X\,;\, N(x) \leq 1\}$ sont convexes.

Par exemple, si $N(x) < 1$ et $N(y) < 1$, il se trouve que pour $0 \leq t \leq 1$, $N(x+t(y-x)) = N((1-t)x+ty) \leq N((1-t)x) + N(ty) \leq (1-t)N(x) + tN(y) < 1$. Pareil pour la boule fermée.

Et il est facile de voir que $E \cap F$ est convexe si E et F le sont, et pareil pour $\overline{E}$ (disons, dans $\mathbb{R}^n$ pour ne pas avoir à définir de topologie).

Je vous renvoie à la littérature pour toutes les histoires intéressantes concernant les propriétés de séparation des convexes et les semi-normes (Hahn Banach et plein d'autres). C'est très intéressant mais on n'a pas le temps.

8.2 Fonctions convexes (sur un ensemble convexe)

On va se placer sur $\mathbb{R}^n$ pour simplifier.

Définition 8.3. *Soit $E \subset \mathbb{R}^n$ un ensemble convexe, et $f : E \to \mathbb{R}$ une fonction. On dit que f est une fonction convexe (sur E) quand*

$$f((1-t)x+ty) \leq (1-t)f(x)+tf(y) \quad \text{pour tout choix de } x, y \in E \text{ et } 0 < t < 1. \tag{8.1}$$

On dit que f est strictement convexe quand

$$f((1-t)x+ty) < (1-t)f(x)+tf(y) \quad \text{pour tout choix de } x, y \in E,\ x \neq y \text{ et } 0 < t < 1. \tag{8.2}$$

Noter que l'on a trivialement $f((1-t)x+ty) = (1-t)f(x)+tf(y)$ quand $t = 0$ et $t = 1$, donc on a simplement exclu ces cas.

Il était utile de demander que E soit convexe, sinon $f((1-t)x+ty)$ n'aurait peut-être pas été défini.

Exemple: si f est affine, elle est convexe (on a même l'égalité dans (8.1)).

Dessin important: le point du graphe est en-dessous du segment. D'ailleurs, ...

Proposition 8.4. *Soit $E \subset \mathbb{R}^n$ un ensemble convexe, et $f : E \to \mathbb{R}$ une fonction. Alors f est convexe si et seulement si l'épigraphe $\mathcal{G}_+ = \big\{(x,y) \in E \times \mathbb{R}\,;\, y \geq f(x)\big\} \subset E \times \mathbb{R}$ est convexe.*

La proposition est encore vraie avec la version ouverte $\big\{(x,y) \in E \times \mathbb{R}\,;\, y > f(x)\big\}$, peut-être plus naturelle, de l'épigraphe. Vérification assez facile à partir des définitions; je vous laisse faire. □

Trois propriétés simples de stabilité. D'abord λf est convexe si f est convexe et $\lambda \geq 0$ (noter que $\lambda < 0$ ne marche pas).

Ensuite, $f + g$ est convexe si f et g sont convexes. En particulier, ajouter une constante, ou une fonction affine, à f ne change pas sa convexité.

Enfin le plus amusant: une borne supérieure (essentiellement quelconque) de fonctions convexes est encore convexe. Plus précisément, si les f_i, $i \in I$, sont convexes sur E, et

$f(x) = \sup_{i\in I} f_i(x)$ est fini pour tout $x \in \mathbb{R}$, alors f est convexe aussi. En effet, pour tout choix de x, y, t, (8.1) dit que pour chaque i,

$$f_i((1-t)x + ty) \leq (1-t)f_i(x) + tf_i(y) \leq (1-t)f(x) + tf(y)$$

(par définition de f). Donc en prenant la borne supérieure, $f((1-t)x+ty) \leq (1-t)f(x) + tf(y)$ aussi.

Noter au passage que la borne inférieure ne marche pas: sur $\mathbb{R}$, $f(x) = -|x|$ n'est pas convexe, et c'est l'inf de x et $-x$. Et il est également vrai que toute fonction convexe peut être obtenue comme borne supérieure de fonctions affines (ou si vous préférez des fonctions affines a telles que $a \leq f$).

On définirait concave, et strictement concave, comme ci-dessus mais avec l'autre signe. Disons simplement que $f = E \to \mathbb{R}$ est dite concave (strictement concave) quand $-f$ est convexe (strictement convexe). Et donc l'étude des fonctions concaves se déduit de celle des fonctions convexes.

8.3 Convexité et dérivée croissante pour f définie sur un intervalle.

C'est quand même en dimension 1 qu'on y voit le plus clair, alors on commence par f définie sur un intervalle. Aussi, désolé pour cette partie du cours, l'essentiel des démonstration se comprend avec des petits dessins!

Théorème 8.5. *Soient $I \subset R$ un intervalle, et $f : I \to \mathbb{R}$ une fonction convexe. En tout point x intérieur de I, f admet une demi-dérivée à gauche $f'_g(x)$ et une demi-dérivée à droite $f'_d(x)$. De plus, si $x < y$ sont deux points intérieurs de I,*

$$f'_g(x) \leq f'_d(x) \leq f'_g(y) \leq f'_d(y). \tag{8.3}$$

En particulier, quand f' existe, elle est croissante.

L'énoncé se prolonge de manière naturelle aux extrémités de I (quand elles existent), avec les mêmes démonstrations. Par exemple, si $I = [a, b]$, $f'_d(a) \in [-\infty, +\infty[$ et $f'_g(b) \in]-\infty, +\infty]$ existent, et $f'_d(a) \leq f'_g(x) \leq f'_d(x) \leq f'_g(b)$ pour tout point intérieur x de I. Le cas où $f_d(a) = -\infty$ et $f_g(b) = +\infty$ sont possibles (faire un dessin!).

Avant de commencer la démonstration, noter que cet énoncé ne change pas si on ajoute à f une fonction affine de son choix. Ceci peut permettre, si vous vous sentez perdu dans une démonstration, supposer par exemple que $f(x) = f(y) = 0$ et de simplifier des calculs. On va essayer de se débrouiller aussi sans faire cela.

Base de la démonstration pour $x, y \in I$, avec $x < y$ (pour ne pas se perdre dans les signes), on pose

$$\tau(x, y) = \frac{f(y) - f(x)}{y - x}. \tag{8.4}$$

C'est le taux d'accroissement, ou la pente de la droite qui passe par les deux points du graphe.

Lemme 8.6. *Soient I et f comme ci-dessus. Pour $x, y, z \in I$ tels que $x < y < z$, on a*

$$\tau(x,y) \leq \tau(x,z) \leq \tau(y,z). \tag{8.5}$$

On peut trouver $t \in (0,1)$ tel que $y = (1-t)x + tz$. Si on avait $f(y) = (1-t)f(x) + tf(z)$ (les trois points du graphe sont sur une droite), on aurait que $\tau(x,y) = \tau(x,z) = \tau(y,z)$. Maintenant $f(y)$ est plus petit que cela (ou égal), ce qui laisse $\tau(x,z)$ pareil, diminue $\tau(x,y)$, et augmente $\tau(y,z)$. D'où l'inégalité.

Si vous n'aimez pas cet argument, vous pouvez dire que le lemme ne change pas quand on ajoute une fonction affine a à f (les troix taux sont juste augmentés de a'), ce qui permet de se ramener au cas où $f(x) = f(z)$, donc $f(y) \leq f(x)$, et dans ce cas $\tau(x,z) = 0$, $\tau(x,y) \leq 0$ et $\tau(y,z) \geq 0$. □

Quand on fixe x et $y > x$ et qu'on fait tendre z vers y, on voit que $\tau(y,z)$ est une fonction croissante de z, qui admet donc une limite (finie ou $-\infty$) quand z tend vers y_+. Mais aussi comme $\tau(y,z) \geq \tau(x,y)$, cette limite est finie, et donc $f'_d(y)$ existe. L'existence de $f'_g(y) \in \mathbb{R}$ se démontre pareillement, en regardant (8.4) et en faisant tendre x vers y.

La première inégalité de (8.3), au point y, se voit en fixant y et en faisant tendre x et z vers y dans (8.5). Pour la seconde on démontre carrément que $f'_d(x) \leq \tau(x,y) \leq f'_g(y)$, où pour la première inégalité on regarde (8.5) avec des points $x < x' < y$ et on fait tendre x' vers x, et pour la seconde on applique (8.5) avec des points $x < y' < y$ et on fait tendre y' vers y. Enfin la dernière inégalité de (8.3) se voit en fixant y et en faisant tendre x et z vers y.

Je vous laisse faire le cas des extrémités éventuelles de I; on ne peut pas dire que les demi-dérivées sont finies parce qu'on n'a pas de point de l'autre coté pour borner les pentes. □

Fin du cours du ma 5 avril 22 en gros. Pas de détails sur la diagonalisation, ils savent.

Le théorème 8.5 a une sorte de réciproque.

Théorème 8.7. *Soient $I \subset R$ un intervalle, et $f : I \to \mathbb{R}$ une fonction dérivable. On suppose que f' est croissante sur I. Alors f est convexe.*

Il y aurait moyen de s'arranger avec des demi-dérivées et d'avoir une réciproque plus proche, mais je préfère ne pas avoir à gérer les demi-dérivées, et surtout pouvoir appliquer le théorème des accroissements finis.

Démonstration. On se donne x, z dans I et $y = (1-t)x + tz$ strictement entre les deux (sinon, l'inégalité (8.1) est vraie trivialement). On calcule les taux en utilisant le TAF: on trouve $\tau(x,y) = f'(\xi_1)$ et $\tau(y,z) = f'(\xi_2)$, avec $x < \xi_1 < y < \xi_2 < z$, donc $f'(\xi_1) \leq f'(\xi_2)$. Donc $\tau(x,y) \leq \tau(y,z)$ et je vous laisse vérifier que ceci veut dire que $f(y) \leq (1-t)f(x) + tf(z)$: à nouveau $\tau(x,y) = \tau(y,z)$ ssi $f(y) = (1-t)f(x) + tf(z)$, et diminuer la valeur de $f(y)$ fait augmenter $\tau(y,z) - \tau(x,y)$, et réciproquement. Ou alors ramenez vous au cas où $f(x) = f(z) = 0$ en soustrayant une fonction affine; dans ce cas les choses sont plus claires.

On finit ce paragraphe par un critère qui se généralisera mieux à plusieurs variables.

Théorème 8.8. *Soient $I \subset R$ un intervalle, et $f : I \to \mathbb{R}$ une fonction dérivable. Alors f est convexe si et seulement si*

$$f(y) \geq f(x) + (y-x)f'(x) \text{ pour tout choix de } x, y \in I. \tag{8.6}$$

Noter que (8.6) est trivial quand $x = y$.

C'est un peu plus bizarre écrit comme ça, mais (8.6) signifie juste que le point $(y, f(y))$ est au-dessus de la tangente au graphe au point $(x, f(x))$, donc ça a un sens géométrique.

On ne peut pas déduire le cas où $x > y$ formellement du cas où $x < y$, parce que le choix de $f'(x)$ brise la symétrie de (8.6), mais ce n'est pas un problème puisque les deux sont vrais.

On commence par supposer f convexe. Donc f' est croissante (par le théorème 8.5). Supposons d'abord que $x < y$; le théorème des accroissements finis dit que $f(y) = f(x)+(y-x)f'(\xi)$ avec un $\xi \in]x, y[$. Donc $f'(\xi) \geq f'(x)$ et on a (8.6). Si maintenant $x > y$, le théorème des accroissements finis dit que $f(x) - f(y) = (x-y)f'(\xi)$ pour un $\xi \in]y, x[$ donc tel que $f'(\xi) \leq f'(x)$. On en déduit encore (8.6) puisque $f(y)-f(x) = -(x-y)f'(\xi) \geq -(x-y)f'(x)$.

Passons à l'autre sens, et supposons cette fois qu'on a (8.6). On se donne $x < z$, et $y = (1-t)x+tz$ strictement entre les deux; on doit vérifier que $f(y) \leq (1-t)f(x)+tf(z)$ et comme précédemment il suffit de voir que $\tau(x, y) \leq \tau(y, z)$ (et si on veut on peut supposer que $f(x) = f(z) = 0$ si on veut simplifier). Mais (8.6) (en échangeant x et y) dit que $\tau(x, y) \leq f'(y)$, et (8.6) entre y et z dit que $\tau(y, z) \geq f'(y)$. On met les deux estimées ensemble et on conclut. □

8.4 Une variante en plusieurs dimensions.

Maintenant on considère $f : E \to \mathbb{R}$, mais avec un ensemble convexe $E \subset \mathbb{R}^n$.

On ne s'effraie pas: la convexité, c'est défini à partir de la restriction de f à des intervalles, donc comprendre les fonctions convexes d'une seule variable devrait nous mener assez loin. Formalisons ceci.

Proposition 8.9. *Soient $E \subset R^n$ un ensemble convexe et $f : E \to \mathbb{R}$ une fonction. Alors f est convexe si et seulement si, pour tout choix de $x, y \in E$, la restriction de f à $[x, y]$ est convexe.*

Ou encore, si et seulement si, pour tout choix de $x, y \in E$, la fonction $h_{x,y} : [0, 1] \to \mathbb{R}$ définie par $h_{x,y}(t) = f((1-t)x + ty)$ est convexe.

Démonstration facile en traduisant tranquillement les définitions. Je vous laisse faire.

Re-traduisons encore. On va se placer dans le cas où E est un ouvert Ω de $\mathbb{R}^n$. Pour tout $z \in \Omega$ et tout vecteur $v \in \mathbb{R}^n \setminus \{0\}$ (vous pouvez aussi le prendre unitaire), notons $L_{z,v} \subset \mathbb{R}^n$ la droite passant par x et de direction e. Donc $L_{z,v} = \{z + tv\,;\, t \in \mathbb{R}\}$. Puis notons $I_{z,v} = \{t \in \mathbb{R}\,;\, z+tv \in \Omega\}$. Donc ce sont les points de $\mathbb{R}$ qui correspondent à $L_{z,v} \cap \Omega$.

Comme on suppose Ω est ouvert, chaque $I_{z,v}$ est ouvert (image réciproque de l'ouvert Ω par l'application continue $t \mapsto z + tv$). Et c'est facile de voir que Ω est convexe si et seulement si chaque $I_{z,v}$ est un intervalle (une partie convexe de $\mathbb{R}$).

Et maintenant, on vérifie aussi aisément (je vous laisse faire) que $f : \Omega \to \mathbb{R}$ est convexe si et seulement si chaque fonction $f_{z,v}$ est convexe, où $f_{z,v} : I_{z,v} \to \mathbb{R}$ est définie par $f_{z,v}(t) = f(z + tv)$ pour $t \in I_{z,v}$. Bref, on prend les restrictions de f aux ensembles $L_{z,v} \cap \Omega$ (modulo l'identifcation de cet ensemble à un intervalle de $\mathbb{R}$) et on demande que cette fonction soit convexe.

Dans ce qui suit on utilisera les dérivées des $f_{z,v}$ pour essayer de caractériser la convexité de f.

Maintenant une version en plus grandes dimensions des théorèmes 8.5 et 8.8.

Théorème 8.10. *Soient Ω un ouvert convexe et $f : \Omega \to \mathbb{R}$ une fonction différentiable sur Ω. Alors sont équivalents*

(8.7) $$f \text{ est convexe sur } \Omega$$

(8.8) $$f(y) - f(x) \geq \langle \nabla f(x), y - x \rangle \quad \text{pour } x \neq y \in \Omega$$

(8.9) $$\langle \nabla f(y) - \nabla f(x), y - x \rangle \geq 0\rangle \quad \text{pour } x \neq y \in \Omega.$$

Fin du cours du 6 avril 2021. Il en restait 2 je crois.

Cette fois on a pris Ω ouvert, parce que sinon on ne sait pas ce que différentiable veut dire. Je crois que c'est plus facile à comprendre avec les gradients, mais voici les versions de (8.8) et (8.9) en termes de différentielle:

(8.10) $$f(y) - f(x) \geq Df(x)(y - x) \quad \text{pour } x \neq y \in \Omega$$

(8.11) $$[Df(y) - Df(x)](y - x) \geq 0\rangle \quad \text{pour } x \neq y \in \Omega.$$

Pour démontrer le théorème, on va utiliser les énoncés correspondants en une variable, et la proposition 8.9. Etant donnés $x, y \in \Omega$, $y \neq x$, on pose $v = y - x$ et on utilise la fonction $f_{x,v} : I_{x,v} \to \mathbb{R}$ définie plus haut, à savoir

(8.12) $$f_{x,v}(t) = f(x + tv) = f((1 - t)x + ty) \quad \text{pour } t \in I_{x,v} = \big\{t \in \mathbb{R}\,;\, x + tv \in \Omega\big\}.$$

Noter qu'ici $I_{x,v}$ contient $[0, 1]$ puisque $[x, y] \subset \Omega$. Du coup, puisque $I_{x,v}$ est ouvert, il contient même un intervalle plus grand $]-\delta, 1 + \delta[$ (avec $\delta > 0$).

La fonction $f_{x,v}$ est dérivable (par composition de fonctions différentiables), et sa dérivée est

(8.13) $$f'_{x,y}(t) = Df(x + tv)(v) = \langle \nabla f(x + tv), v\rangle = \langle \nabla f(x + t(y - x)), y - x\rangle.$$

Ou, si vous préférez les cordonnées,

(8.14) $$f'_{x,y}(t) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x + tv)v_i \quad (\text{avec } v = y - x, \text{ et où les } v_i \text{ sont ses coordonnées}).$$

Retour au théorème. D'abord on suppose f convexe. Alors $f_{x,y}$ est convexe (sur un intervalle $]-\delta, 1+\delta[$) pour tout choix de x, y.

Appliquons (8.6) à $f_{x,v}$ aux points 0 et 1. On trouve $f_{x,y}(1) \geq f_{x,y}(0) + f'_{x,y}(0)$ donc en traduisant avec (8.13), $f(y) \geq f(x) + Df(x)(y-x)$, ce qui donne (8.10) et (8.8). Maintenant appliquons à $f_{x,v}$ la version simple du théorème 8.5, qui dit que $f_{x,v}$ est croissante. On trouve $f'_{x,v}(1) \geq f'_{x,v}(0)$, ce qui se traduit par $Df(x + (y-x))(y-x) \geq Df(x)(y-x)$ ou encore $Df(y)(y-x) \geq Df(x)(y-x)$, ce qui donne (8.11) et (8.9).

Il reste à démontrer les réciproques. On suppose qu'on a (8.10) ou (8.11) (pour tout choix de $x \neq y \in \Omega$) et on veut montrer que f est convexe sur tout intervalle $[x, y]$, ou de manière équivalente que chaque $f_{x,v}$ est convexe sur $I_{x,v}$. Alors on fixe x et v, et on regarde ce que nous donne la condition (8.10) ou (8.11), prise aux points $\widetilde{x} = x + t_1 v$ et $\widetilde{y} = x + t_2 v$. Par exemple, pour (8.10), on trouve $f_{x,v}(t_1) - f_{x,v}(t_2) \geq Df(\widetilde{x})(\widetilde{y} - \widetilde{x}) = (t_2 - t_1) Df(\widetilde{x})(v) = (t_2 - t_1) f'_{x,v}(t_1)$. Ceci est vrai pour tous choix de $t_1 \neq t_2 \in I_{x,v}$, et le théorème 8.8 dit que $f_{x,y}$ est bien convexe sur $I_{x,v}$.

De même si on a (8.10), on trouve que $f'_{x,v}(t_2) \geq f'_{x,v}(t_1)$ pour $t_2 > t_1$, donc que $f'_{x,v}$ est croissante, donc que f est convexe (sur $I_{x,v}$), comme souhaité. □

Je ne sais pas trop comment post-commenter ceci. D'un coté, la démontration montre qu'il n'y a pas vraiment plus dans cet énoncé que dans ceux de dimension 1. D'un autre coté, on est vraiment en train de se débrouiller un peu avec une fonction de plusieurs variables, en écrivant des inégalités sur les gradients (mais quand même, toujours en faisant le produit scalaire avec $v = (y - x)$, donc on ne vise jamais dans deux directions à la fois).

8.5 Cararctérisation (partielle) par la dérivée seconde

Théorème 8.11. *Soient Ω un ouvert convexe et $f : \Omega \to \mathbb{R}$ une fonction de classe C^2 sur Ω. Alors f est convexe sur Ω si et seulement si*

$$D^2 f(x)(u, u) \geq 0 \quad \textit{pour tout } x \in \Omega \textit{ et tout } u \in \mathbb{R}^n. \tag{8.15}$$

Autrement dit, si et seulement si, pour tout $x \in \Omega$, la forme quadratique associée naturellement à la seconde dérivée de f en x est positive (mais pas forcément définie positive!). Ou encore, si on note $H = H_x$ la matrice Hessienne de f en x, ssi pour tout x, la forme quadratique $u \to u^t H u$ est positive sur $\mathbb{R}^n$ (vu comme un ensemble de vecteurs colonne).

J'ai demandé C^2 parce que c'est plus pratique. Ce que je veux vraiment, c'est pouvoir calculer la deuxième dérivée de f sur n'importe que segment, comme ci-dessous. Et comme je vais appliquer la relation de Schwarz pour ne pas me mélanger dans les dérivées seconde, j'utiliserai C^2.

Si on supposait que H est définie positive en x, on obtiendrait par la même démonstration ci-dessous que f est strictement convexe dans un voisinage de x.

Si ceci vous rappelle ce qu'on a fait pour les minima, c'est normal. A l'époque, on a supposé que $Df(x) = 0$ (on s'est placé en un point critique), et on a demandé si le graphe de f était au-dessus de son plan tangent. Ici on va faire un peu pareil.

Démonstration en dimension 1: on dit que pour $f : I \to \mathbb{R}$ de classe C^2, f est convexe ssi f' est croissante ssi $f'' \geq 0$. Et c'est bien vrai.

Et maintenant, passons à la démonstration en dimension n quelconque. Supposons f convexe. Alors pour tout $x \in \Omega$ et tout vecteur $v \in \mathbb{R}^n \setminus \{0\}$, la fonction $f_{x,v}$ définie par $f_{x,v}(t) = f(x+tv)$ (comme en (8.12)) est convexe dans un ouvert autour de 0. Elle est aussi de classe C^2, comme composition de fonctions de classe C^2, et on va re-calculer ses dérivées. On a vu en (8.14) que

$$f'_{x,v}(t) = Df(x+tv)(v) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x+tv)v_i = \sum_{i=1}^{n} v_i \partial_i f(x+tv),$$

où cette seconde notation pour les dérivées partielles sera un peu plus pratique. On re-dérive cette fonction en fonction de t (c'est le même calcul, mais où l'on a remplacé f par un des $\partial_i f$. On trouve

(8.16)

$$f''_{x,v}(t) = \sum_{i=1}^{n} v_i \Big\{ \sum_{j=1}^{n} \partial_j(\partial_i f))(x+tv)v_j \Big\} = \sum_{i=1}^{n}\sum_{j=1}^{n} \partial^2_{ij} f(x+tv)v_j v_i = D^2 f(x+tv)(v,v)$$

où l'on a utilisé la relation de Schwarz pour ne pas avoir à se péoccuper de l'ordre des dérivations, puis par définition de $D^2 f(x+tu)$.

Bon, la fonction $f_{x,v}$ est convexe, donc sa dérivée seconde est positive autour de 0, et on en déduit bien que $D^2 f(x+tv)(v,v) \geq 0$. Et comme ceci est vrai pour tout $x \in \Omega$ et tout $v \in \mathbb{R}^n \setminus \{0\}$, on a démontré la direction directe.

Pour la réciproque, on dit que si $D^2 f(x)(v,v) \geq 0$ partout, alors pour tout choix de x, v, et t comme ci-dessous, on a que la dérivée seconde de $f_{x,v}$ au point t est positive. On en déduit que la fonction $f_{x,v}$ ci-dessus est convexe sur l'intervalle $I_{x,v}$. Maintenant la proposition 8.9 montre que f est convexe. □

Remarque: dans une première version, j'avais calculé avec la formule de Taylor en dimension n, et fait des développements limités et utilisé le critère du théorème 8.10. C'est logique, mais comme à la fin tout se passe sur des segments, autant l'avouer et se ramener directement à la dimension 1.

Exemple le plus typique: une forme quadratique positive (sur $\mathbb{R}^n$), ou plus généralement (toujours sur $\mathbb{R}^n$) une fonction f de la forme

$$f(x) = \sum_{j=1}^{m} \Big| \sum_{i=1}^{n} \alpha_{i,j} x_j \Big|^{p_j}, \tag{8.17}$$

avec des coefficients réels $\alpha_{i,j}$ quelconques et des exposants $p_j > 1$. Mais le plus simple est de prendre $p_j = 2$ pour tout j. La démonstration est assez simple: une somme de fonctions convexes est convexe, et pour chacune des fonctions $\Big|\sum_{i=1}^{n} \alpha_{i,j} x_j\Big|^{p_j}$ on se ramène (à la fin des calculs que je ne fais pas) à montrer que $y \to |y|^p$ est convexe sur $\mathbb{R}$ dès que $p > 1$, par le

calcul direct de croissance de la première dérivée (si $p \geq 2$, on peut même calculer la seconde dérivée, mais autrement il y a un problème en 0).

En dimension 1 vous connaissez d'autres fonctions convexes fort utiles, comme la fonction e^x, certaines fonctions puissances (comme ci-dessus), ou la fonction $-\ln(x)$ (sur $\mathbb{R}_+^*$). Ca se voit tellement bien sur le graphe (on a probablement un logiciel dans le cerveau pour détecter la convexité) qu'on a peu de chances de se tromper.

9 Extrema liés; multiplicateurs de Lagrange

On veut maintenant chercher les minima (ou les maxima) d'une fonction f sur une partie $S \subset \Omega$ d'un ouvert de $\mathbb{R}^N$ (vous pouvez prendre $\Omega = \mathbb{R}^N$). Mais où la fonction $f : \Omega \to \mathbb{R}$ est définie (et bientôt différentiable) sur le grand ensemble Ω.

Pour bien insister, notons $f_{|S}$ la restriction de f à S. On cherche une condition nécessaire pour que f ait un minimum local en $w \in S$. Evidemment, un maximum local serait pareil.

Exemple 1. Prenons pour commencer l'exemple un peu trivial où $\Omega = \mathbb{R}^N$ et S est le sous-espace de dimension $m = N - n$ donné par

$$S = \big\{x = (x_1, \dots, x_N) \in \mathbb{R}^n \,;\, x_{m+1}, = x_{m+2} = \dots = x_N = 0\big\} \simeq \mathbb{R}^m \subset \mathbb{R}^N. \tag{9.1}$$

Je vais essayer de garder N comme dimension ambiante, m comme dimension de S, et n comme nombre d'équations, pour faciliter le lien avec le théorème des fonctions implicites. Mais des typos peuvent se produire ci-dessous.

Si $f_{|S}$ a un minimum local, disons en $w = 0 \in S$, on sait que la fonction $(x_1, \dots, x_m) \to f(x_1, \dots, x_m, 0, \dots, 0)$ a un minimum local en $0 \in \mathbb{R}^{m-n}$, donc que les dérivées partielles $\frac{\partial f}{\partial x_i}$, $1 \leq i \leq m$, s'annulent en 0. Par contre, les dérivées partielles $\frac{\partial f}{\partial x_i}$, $i > m$, font ce qu'elles veulent! Exemple où $N = 2$, $m = n = 1$, ce qui compte est $f(x_1, 0)$ et la condition est $\partial_1 f(x_1, 0) = 0$.

Exemple 2. Le cas suivant est quand S est paramétré par un ensemble V. Par exemple, par un ouvert V de $\mathbb{R}^m$ comme c'est le cas trivialement ci-dessus.

Donc on se donne $S \subset \Omega \subset \mathbb{R}^N$ et $f : \Omega \to \mathbb{R}$ comme ci-dessus, et maintenant aussi un ensemble V et $F : V \to S$. On se donne aussi $a \in V$ et $w = F(a) \in S$. Commencons par vérifier que

(9.2) Si $f_{|S}$ a un minimum global en w, alors $f \circ F : V \to \mathbb{R}$ a un minimum global en a.

C'est facile, puisque $f(z) \geq f(w)$ pour tout $z \in S$, donc aussi $f(F(b)) \geq f(w) = f \circ F(a)$ pour tout $b \in V$, puisque $z = F(b)$ est un point de S. Pour un minimum local, on a juste besoin en plus de la continuité de F en a:

(9.3) Si F est continue au point a et $f_{|S}$ a un minimum local en w, alors $f \circ F : V \to \mathbb{R}$ a un minimum local en a.

Cette fois l'hypothèse donne un voisinage W de w dans S tel que $f(z) \geq f(w)$ pour tout $z \in W$. Mais F est continue en a, donc il existe un voisnage V' de a dans V tel que $F(b) \in W$ pour tout $b \in V'$. Alors, pour tout $b \in V'$, on a bien que $z = F(b) \in W$, donc $f \circ F(b) = f(z) \geq f(w) = f \circ F(a)$, comme souhaité.

Maintenant on suppose de plus que f est différentiable en $w = F(a)$ (on a supposé que f est définie sur un voisinage de w, et pas seulement sur S, donc cette nouvelle hypothèse a un sens), et aussi que a est un point de l'intérieur de V et que F est différentiable en a (les deux vont un peu ensemble: on n'a pas défini différentiable autrement). Et on applique le critère connu: puisque $f \circ F$ est différentiable en a et a un minimum local en a, on trouve que $Df \circ F)(a) = 0$. Et comme $D(f \circ F)(a)(v) = Df(w)(DF(a)(v))$ pour tout vecteur $v \in \mathbb{R}^m$, on trouve que

$$Df(w)((DF(a)(v)) = 0 \text{ pour tout } v \in \mathbb{R}^m. \tag{9.4}$$

Comme ceci fait trop de parenthèses et que $Df(w)$ est donné par produit scalaire avec $\nabla f(w)$, (9.4) signifie plus simplement que

$$\nabla f(w) \perp DF(a)(v) \text{ pour tout } v \in \mathbb{R}^m. \tag{9.5}$$

Traduisons encore une fois. Pour chaque coordonnée i, $1 \leq i \leq m$, on peut prendre $v = e_i$, poser $\xi_i = DF(a)(e_i) \in \mathbb{R}^N$, et on a trouvé que (si $f_{|S}$ a un minimum local en $w = F(a)$ et de plus tout le monde est differentiable comme ci-dessus), on a les m conditions

$$\nabla f(w) \perp \xi_i \text{ pour } 1 \leq i \leq m. \tag{9.6}$$

On aurait pu prouver la même condition $\nabla f(w) \perp \xi_i$ en disant que l'application (définie sur un voisinage de 0 dans $\mathbb{R}$) $t \mapsto f(F(a + te_i))$ a un minimum local en $t = 0$, et en dérivant cette application. En effet la dérivée de cette application est $t \to DF(a + te_i)(e_i)$ (en tout t où tout le monde est différentiable, et ici on a juste besoin de $t = 0$).

Encore un mot: ξ_i est la dérivée partielle au point a de F, donc $\frac{\partial F}{\partial x_i}$, qui est donc un vecteur de $\mathbb{R}^N$ dont la coordonée j est $\frac{\partial F_j}{\partial x_i}(a)$. C'est légitime d'en faire le produit scalaire avec $\nabla f(w) \in \mathbb{R}^N$.

En gros fin du cours du mercredi 6 avril 22 (cours $n - 1$).

Exemple 3. Juste un peu plus précis que l'exemple précédent, où maintenant on suppose que S est en fait le graphe d'une fonction de classe C^1. Je veux dire: soit $\varphi : V \to \mathbb{R}^n$ une fonction de classe C^1 définie sur un ouvert de $\mathbb{R}^m$. Ensuite notons $S \subset \mathbb{R}^N = \mathbb{R}^m \times \mathbb{R}^n$ le graphe de φ. Plus précisément, prenons

$$S = \big\{(x, \varphi(x)) \in \mathbb{R}^m \times \mathbb{R}^n \, ; \, x \in V\big\}. \tag{9.7}$$

Sur $\mathbb{R}^N \simeq \mathbb{R}^m \times \mathbb{R}^n$ on notera volontiers x les m premières coordonnées, et y les n dernières.

C'est en fait le même exemple que ci-dessus, avec la fonction $F : V \to \mathbb{R}^N$ donnée par

$$F(x) = (x, \varphi(x)) \text{ pour } x \in V. \tag{9.8}$$

Et les vecteurs ξ_i (qui engendrent le plan tangent de dimension m) sont faciles à calculer, puisque

$$\xi_i = \frac{\partial F}{\partial x_i}(a) = \Big(e_i, \frac{\partial \varphi}{\partial x_i}(a)\Big) \tag{9.9}$$

(avec la première partie qui est un vecteur de $\mathbb{R}^m$ et la seconde qui est un vecteur de $\mathbb{R}^n$). Et donc, pour un minimum local de $f_{|S}$ en $w = (a, \varphi(a) \in S$, on a que (9.6) est satisfait avec ce choix des ξ_i. Ceci nous fait donc m équations que $\nabla f(w)$ doit satisfaire.

Et en plus, les ξ_i sont indépendants: juste regarder leur projections sur $\mathbb{R}^m$ (composée des premières coordonnées), et ce sont déjà des vecteurs indépendants, donc les équations peuvent s'écrire

$$\nabla f(w) \in X^\perp, \text{ où } X \text{ est l'espace vectoriel (de dimension } m) \text{ engendré par les } \xi_i$$

et $X^\perp$ est l'orthogonal, qui est de dimension n.

Après tous ces exemples on en vient à l'énoncé principal, qui est une condition nécessaire exactement comme dans (9.6), mais où l'on n'aura pas à calculer les ξ_i. Cette fois on se place dans le cas où S est donné par n équations (indépendantes au point w, donc dans un petit voisinage aussi).

Théorème 9.1. *On se donne $N = m + n$ (des entiers positifs). Soient $\Omega \subset \mathbb{R}^N$ un ouvert, puis $G = (g_1, \dots, g_n)$ une fonction différentiable de Ω dans $\mathbb{R}^n$. On note*

$$S = \Omega \cap G^{-1}(0) = \big\{z \in U\,;\, G(z) = 0\big\} = \big\{z \in U\,;\, g_1(z) = \dots = g_n(z) = 0\big\}. \tag{9.10}$$

Soit encore $f : \Omega \to \mathbb{R}$ une fonction différentiable. On se donne $w \in S$ et on suppose que la restriction de f à S a un minimum local en w.

On suppose enfin que les vecteurs $\nabla g_i(w)$, $1 \le i \le n$, sont indépendants.

Alors $\nabla f(w)$ est dans l'espace vectoriel engendré par les vecteurs $\nabla g_i(w)$, $1 \le i \le n$.

Commentaires:

La conclusion dit qu'il existe des nombres réels $\lambda_1, \dots, \lambda_n$ tels que

$$\nabla f(w) = \sum_{1 \le i \le n} \lambda_i \nabla g_i(w). \tag{9.11}$$

En fait, ces nombres sont uniques, puisque les $\nabla g_i(w)$ sont supposés indépendants. On appelle les λ_i les multiplicateurs de Lagrange.

Le cas où $n = 1$ est important. Alors S est une hypersurface donnée par une seule équation $g(z) = 0$, avec $\nabla g(w) \neq 0$. Dans ce cas on trouve que si f a un extremum (sur S) en w, alors $\nabla f(w)$ est un multiple de $\nabla g(w)$. Dit autrement, $\nabla f(w)$ est orthogonal à la direction du plan tangent à S en w, ou encore, $\nabla f(w)$ est un multiple du vecteur normal $n(w) = \nabla g(w)/|\nabla g(w)|$ qu'on a trouvé il y a longtemps. C'est pas trop surprenant: f a bien le droit de varier, mais seulement dans la direction orthogonale au plan tangent à S.

L'exemple 1 donné plus haut où $S = \mathbb{R}^m$ correspond à prendre

$$S = \big\{x \in \mathbb{R}^n \,;\, x_{m+1} = \ldots = x_N = 0\big\},$$

donc les g_i sont les dernières coordonnées, leur gradients sont les n derniers vecteurs de base, et on demande que $\nabla f(w)$ soit une combinaison linéaire de ces vecteurs-là. Autrement dit, on demande que les premières coordonnées de $\nabla f(w)$ soient toutes nulles, c.-à.-d., comme on l'a dit plus haut, que $\frac{\partial f}{\partial x_i} = 0$ pour $1 \leq i \leq m$.

L'essentiel de la démonstration du théorème va consister à voir que dans le cas de l'exemple 3 (où S le graphe de $\varphi : V \subset \mathbb{R}^m \to \mathbb{R}^n$), il se trouve que la conclusion (comme quoi $\nabla f(w)$ est dans l'espace vectoriel engendré par les vecteurs $\nabla g_i(w)$) correspond bien à ce qu'on a dit plus haut.

D'abord, les dimensions correspondent: on a trouvé que les équations (9.6) définissent un espace vectoriel $X^\perp$ de dimension $n = N - m$, et l'espace engendré par les $\nabla g_i(w)$ aussi est de dimension n (par hypothèse d'indépendance). Donc pour voir que ces deux espaces sont les mêmes, il suffit de voir que chacun des $\nabla g_i(w)$ est bien dans $X^\perp$, donc orthogonal à chaque ξ_i.

Evidemment, on pourrait faire le calcul, mais on est paresseux alors on va s'en dispenser. En effet par définition de S, chaque g_j est constante sur S, donc a un minimum local en w, donc vérifie la fameuse condition (9.6). Donc ∇g_j est orthogonal aux ξ_i. C'est exactement ce qu'on voulait.

Bref, le théorème est juste quand S est le graphe d'une fonction φ de classe C^1. Il ne reste plus qu'à voir que c'est en fait (modulo choix d'une autre base orthonormée sur $\mathbb{R}^N$) le cas général!

Et pour ceci on va fair appel au théorème des fonctions implicites (vous le sentiez venir, j'espère). Donc on se donne Ω, $S = \Omega \cap G^{-1}(0)$, f, et w comme dans l'énoncé, et on va commencer par choisir certaines coordonnées (qu'on appellera les premières coordonnées). Ceci doit ressembler au paragraphe 6.5.

On nous donne n vecteurs indépendants $V_j = \nabla g_j(w)$. Ce sont des vecteurs de $\mathbb{R}^N$, et quand on les met ensemble, on a donc une matrice $N \times n$ qui est de rang n. Ses coefficients sont les $v_{i,j} = \frac{\partial g_j}{\partial x_i}$, avec $1 \leq i \leq N$ et $1 \leq j \leq n$.

La matrice transposée est aussi de rang n, donc dedans (par le théorème du rang) il y a une matrice $n \times n$ qui est encore de rang m. Ceci signifie que l'on peut trouver n numéros de colonnes de la matrice transposée (de ligne dans la matrice des V_i), à savoir $i_1 < i_2 < \ldots i_m$ tels que la matrice carrée (appelons-la M) des $v_{i_\ell,j}$, où $1 \leq \ell, j \leq n$, est inversible. Et on rappelle que $v_{i,j}$ est la ligne i de $\nabla g_j(w)$, donc $v_{i,j} = \frac{\partial g_j}{\partial x_i}$.

Bon, pour comprendre ce qui se passe on va supposer que les coordonnées choisies sont les dernières, c.-à-d. que $i_\ell = m + \ell$. Celles qu'on a appelées $y \in \mathbb{R}^n$ dans le passé. Et donc dans ce cas, on sait que les dérivées de G par rapport aux dernières variables (donc y) forment une matrice inversible $n \times n$, ce qui est exactement la condition (6.2) dans le théorème des fonctions implicites.

Alors le théorème s'applique, on trouve que , dans un voisinage de w, $S = G^{-1}(0)$ est le graphe d'une fonction φ de classe C^1 comme ci-dessus, et du coup on est dans le cas particulier où le théorème a été démontré.

Il reste encore le cas où les coordonnées choisies $i_1, \dots i_n$ ne sont pas directement les dernières coordonnées de $\mathbb{R}^n$. Mais alors, il suffit en fait de changer de base dans $\mathbb{R}^n$, c.-à-d. de considerer l'écriture de S dans une autre base (où l'on a juste renuméroté les e_i pour que les e_{ℓ_i} soient les derniers). Dans cette base, S a la forme souhaitée, et on en déduit bien le théorème! Pas la peine de garder trace de qui sont ces coordonnées, ce qui nous intéresse est le résultat du théorème, et pas vraiment la valeur précise des ξ_i qu'on a choisi comme base de X (par contre X et $X^\perp$ nous intéressent, mais là on sait que $X^\perp$ est l'espace engendré par les ∇g_j). □

Fin du dernier cours le 12/4/22. Mais j'ai essayé de garder 40mn pour des exos de recherche d'extrema.

Encore une remarque: dans le cas des graphes, on a vu que les ξ_i décrits plus haut sont une base de l'espace vectoriel X parallèle au plan tangent à S en w (je prends des précautions, parce que pour moi X est un plan vectoriel de dimension $n - m$, alors que le plan tangent est le plan affine parallèle à X qui passe par w). Et donc les $\nabla g_j(w)$ sont m vecteurs qui engendrent l'orthogonal $X^\perp$.

Quand $m = 1$, on a une hypersurface avec un hyperplan tangent, et un seul vecteur normal (multiple de) $\nabla g(w)$. Et donc on a dit que $\nabla f(w)$ est un multiple de $\nabla g(w)$.

C'est aussi amusant de constater que dans cette histoire on peut changer G (par exemple en prenant $\widetilde{g}_j(z) = \sin(g_j(z)) + \sum_{i<j} 2^i g_i(z)$), tout en gardant le même S. Quand on fait ceci, typiquement on trouve d'autres $\nabla \widetilde{g}_j(w)$, mais si on n'a pas trop mal choisi les nouveaux g_j, on trouve toujours le même espace $X^\perp$ engendré, et donc la même condition nécessaire.

Exemple 4. Prenons $f : \mathbb{R}^2 \to \mathbb{R}^2$, donnée par $f(x, y) = xy$. On cherche les extrema de f sur S, la sphère unité.

Evidemment, on sait qu'il y en a (f est continue et S compacte). On calcule $\nabla f(x, y) = (y, x)$. Ici, on peut définir la sphère comme $g^{-1}(0)$, où $g(x, y) = x^2 + y^2 - 1$ (prendre la racine finirait par marcher, mais avec des calculs plus compliqués pour se débarasser des racines). Donc $\nabla g(x, y) = (2x, 2y)$. Ce vecteur ne s'annule pas sur S, donc le théorème s'applique.

La condition de Lagrange dit que $(y, x) = \lambda(x, y)$ pour un certain $\lambda \in \mathbb{R}$. Et on voit que $y = \lambda x = \lambda^2 y$ et $x = \lambda y = \lambda^2 x$. Comme $(x, y) \neq (0, 0)$, on trouve $\lambda^2 = 1$. De sorte que les seuls points possibles sont avec $x = y$ ou $x = -y$. Comme $x^2 + y^2 = 1$, il ne reste que les 4 points $(\pm\frac{\sqrt{2}}{2}, \pm\frac{\sqrt{2}}{2})$. Donc les bornes inférieure et supéieure de f sur S doivent être atteincte en certains de ces points. On en déduit que le maximum est $1/2$ (atteint en deux de ces points), et le minimum est $-1/2$ (atteint aux deux autres).

Evidemment, l'exemple est un peu trop simple: on aurait pu passer en polaire, étudier la fonction $\widetilde{f}(\theta) = \sin(\theta)\cos(\theta) = \frac{1}{2}\sin(2\theta)$, et trouver nos quatre extrema facilement.

Exemple 5. Un peu plus compliqué pour que vous ne pensiez pas qu'on se moque de vous. Cherchons à nouveau les extrema pour $f(x, y, z) = xyz$ sur la sphère unité S de $\mathbb{R}^3$.

On est gentil, on a pris un exemple avec $m = 1$, une seule fonction définissante $g(x, y, z) = x^2 + y^2 + z^2 - 1$, et donc on aura un seul mutiplicateur λ.

On calcule $\nabla g(x, y, z) = (2x, 2y, 2z)$ qui ne s'annule pas sur la sphère. En un extremum (x, y, z) de f sur S, le théorème dit qu'il existe $\lambda \in \mathbb{R}$ tel que $\nabla f = \lambda \nabla g$. Ici $\nabla f(x, y, z) = (yz, xz, yz)$. Donc on trouve $yz = 2\lambda x$, $xz = 2\lambda y$, et $xy = 2\lambda z$ (et en fait poser $\mu = 2\lambda$ aurait été plus simple).

Une première option est que $x = 0$ ou $y = 0$ ou $z = 0$. Par exemple, si $x = 0$, on trouve $yz = 0$, donc par exemple $x = y = 0$ (le cas où $x = z = 0$ serait pareil). Alors $z = \pm 1$, on voit directement que $\nabla f(x, y, z) = (yz, xz, yz) = (0, 0, 0)$ (donc la condition de Lagrange est satisfaite avec $\lambda = 0$, mais on peut aussi voir que f prend des valeurs strictement positives et négatives près de (x, y, z), ce point n'est pas un extremum local. Pareil pour tous les autres points tels que $xyz = 0$.

Il reste le cas où ni x, ni y, ni z ne sont nuls. Alors λ non plus (regarder n'importe quelle équation). On fait le produit des deux premières équations et on trouve en simplifiant $z^2 = 4\lambda^2$. On fait pareil avec les autres équations et on trouve $x^2 = y^2 = z^2$. Et comme la somme des trois fait 1, on trouve $x, y, z = \pm\frac{1}{\sqrt{3}}$, avec des signes variables. On pourrait vérifier que ce sont tous des solutions des équations, mais la vérité est qu'on en a assez peu pour calculer toutes les valeurs, et que suivant les signes toutes donnent la valeur $\pm 3^{-3/2}$.

Ici encore, on aurait pu passer en sphérique, et juste chercher les points critiques d'une fonction des deux variables φ, θ, mais déjà ça a l'air moins drôle.

Exemple 6 (un dernier pour la route). Soit maintenant à maximiser ou minimiser le même f, mais sur l'intersection de S et de la surface d'équation $x^2 + y^2 - z^2 = 0$. On essaie de ne pas tricher: on aurait pu commencer par calculer $z^2 = 1/2$ et ainsi gagner une dimension aisément, mais non, on essaie par Lagrange. On a maintenant une autre fonction g_2, de gradient $(2x, 2y, -2z)$. On trouve que $\nabla f = \lambda \nabla g + \mu \nabla g_2$, qui donne les équations $yz = 2\lambda x + 2\mu x$, $xz = 2\lambda y + 2\mu y$, et $xy = 2\lambda z - 2\mu z$, où λ et μ sont des réels inconnus aussi. Et cette fois c'est aussi raisonnable de dire qu'on a aussi $x^2 + y^2 + z^2 = 1$ et $x^2 + y^2 - z^2 = 0$. Ca fait 5 variables et 5 inconnues, on peut s'en sortir mais ca a l'air désagréable (éliminer z tout de suite aurait été plus simple). Bon, moi j'y renonce: je vois des astuces, mais justement parce que je sais qu'on peut éliminer z. Donc le théorème semble cool, mais attention au nombre d'équations.

Dit d'une autre manière, ici si on arrive (par un miracle, vous direz, qui correspond en gros à calculer z^2) à paramétrer notre intersection (je parie que c'est une union de deux cercles horizontaux), on trouvera un seul paramètre, et il ne nous suffira de dériver $f \circ F$ pour obtenir une seule équation et donc, en principe, un nombre fini de solutions potentielles parmi lesquelles trier.