Un peu de calcul différentiel

Résumé

Optimiser une fonction objectif convexe par une démarche itératives réside dans le fait de déterminer une direction dans laquelle chercher le prochain itéré, celui-ci devant nécessairement avoir une valeur objectif plus petite. Géométriquement cela se caractérise par le fait de d'identifier les hyperplans d'appui aux sous-niveaux d'une fonction objectif ; ils indiquent une direction de recherche pour minimiser celle-ci. La définition de ces hyperplans d'appui s'effectue par une étude locale des fonctions objectifs, une étude qui généralise l'apport de la dérivée d'une fonction numérique à la compréhension du comportement de celle-ci en un point.

Table des matières

1 Contexte

Optimiser consiste à chercher les éventuelles plus grandes ou plus petites valeurs que peuvent prendre des fonctions à valeurs numériques ; des fonctions qui représentent un coût, une erreur ou un gain, le plus souvent. Cela sous-entend être en mesure de détecter la géométrie des graphes des fonctions en jeu, là où elles prennent des formes de puits, de bosses, quand elles sont croissantes ou non etc.
Les mathématiques inventent rarement quand cela ne semble pas nécessaire, du moins elles étreignent les généralisations faciles des démarches déjà éprouvées. Que savons-nous au sujet de l'étude des fonctions, notamment de l'étude de leurs extrema ? Si $f : I \to \mathbb{R}$ est une fonction définie sur un intervalle ouvert $I \subset \mathbb{R}$ vous savez déjà, dans l'hypothèse où f est dérivable, que le signe de la dérivée sera déterminant pour étudiant les variations de f. Vous savez de plus que les zéros de la dérivée (les points critiques) sont proteurs d'informations quant aux éventuels extrema. En supposant de plus que f deux fois dérivable, vous êtes en mesure d'éventuellement différencier un point où f atteint un maximum local d'un point où elle atteint un minimum local.
Ce succès [1] nous pousse à étendre les notions de dérivée ou de points critiques à des fonctions objectifs qui ne sont plus définies uniquement sur des parties, mes des parties de $\mathbb{R}^n$ pour $n > 1$. On est face aux deux questions suivantes :

— *quoi étendre exactement ?*
— *comment s'adapte les propriétés qu'on étend à un contexte multivarié ?*

La réponse à la première question s'articule autour du choix suivant : on étend le caractère de meilleure approximation affine locale de la fonction à l'étude. Pour ce qui est de la seconde, on est rapidement face à la difficulté suivante : en dimension 1, donner une direction relative équivaut à donner un signe, on va vers la gauche avec un signe $-$ et à droite avec un signe $+$. En dimension supérieure cette aspect particulièrement simplificateur n'existe plus. Il faut désormais enrichir l'extension de la notion de dérivée et de ses propriétés pour s'adapter à un cadre intrinséquemment vectoriel. Pour cela il nous faudra

— définir des manières de signifier qu'on est proche d'un point et donc de mesurer une distance ;
— d'introduire des éléments d'algèbre linéaire permettant, en s'associant au point précédent, de définir la notion de meilleure approximation affine d'une fonction localement ;
— transformer l'interprétation *signe de la dérivée* en une interprétation proprement géométrique adaptable à un contexte vectoriel.

C'est différents éléments sont le sujet de cette fiche de travail.

2 Attendus

Voici les éléments qu'on souhaiterez vous voir intégrer à la suite du travail d'étude qu'aura nécessiter cette fiche :

1. une connaissance des différentes normes standard en dimension finie et notamment de la forme des boules qu'elles décrivent
2. une culture des pathologiques qui peuvent apparaître dans les phénomènes de limite, continuité ou de comparaison locale de fonctions multivariées
3. un savoir-faire quant aux calculs de différentielles de fonctions, de dérivées partielles et de gradients (notamment des différentielles et dérivées partielles de fonctions composées)
4. une maîtrise de la notion de gradient d'une fonction multivariées et à valeurs dans $\mathbb{R}$ et de son interprétation géométrique
5. une compréhension aboutie du cas des problèmes d'optimisation quadratiques sans contraintes
6. une compréhension de l'apport de la différentielle seconde à l'étude local des fonctions multivariées
7. une connaissance des critères de convexité de premier et second ordre.

1. En réalité relatif, mais l'humanité s'accroche au peu de succès qu'elle a eu.

3 Normes sur $\mathbb{R}^n$

Une norme sur un espace vectoriel apporte une manière de mesurer la *longueur* d'un vecteur tout en respectant un minimum la structure vectorielle. Elle permet en particulier de définir une notion de distance entre deux points de $\mathbb{R}^n$ par la *longueur* du vecteur qui les relie. Pouvoir changer de mesure de longueur suivant les problèmes qu'on attaque est crucial lorsque l'on s'attaque à des problèmes d'optimisation. À la fois pour pouvoir modéliser les problèmes en jeu : *comment mesure-t-on la différence entre deux mots ?* Ou pour accélérer la convergence de certains algorithmes d'apprentissages.

Définition 3.1. Une norme sur $\mathbb{R}^n$ est une application $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}_+$ telle que :

1. $\|x\| = 0 \Leftrightarrow x = 0$;
2. $\forall \lambda \in \mathbb{R}$, $\forall x \in \mathbb{R}^n$, $\|\lambda x\| = |\lambda| \|x\|$ (relation d'*homogéinité*) ;
3. $\forall x \in \mathbb{R}^n$, $\forall y \in \mathbb{R}^n$, $\|x + y\| \leq \|x\| + \|y\|$ (*Inégalité triangulaire*).

Question 3-1 Interpréter chacune des propriétés suivantes avec vos propres mots.

L'inégalité triangulaire donne lieu à une autre inégalité, appelée *inégalité triangulaire inversée* qui peut parfois être utile [2] :

$$\forall x, y \in \mathbb{R}^n, \qquad \big|\|x\| - \|y\|\big| \leq \|x - y\|.$$

Question 3-2 Représenter cette inégalité géométriquement. Essayer de la déduire de l'inégalité triangulaire.

Les trois normes les plus fréquentes d'utilisation sont les trois suivantes, elles sont respectivement qualifiées de normes 1, 2 et infinie.

1. $\forall x \in \mathbb{R}^n$, $\|x\|_1 = \sum_{i=1}^n |x_i|$;
2. $\forall x \in \mathbb{R}^n$, $\|x\|_2 = \sqrt{\sum_{i=1}^n x_i^2} = \sqrt{x^T x}$;
3. $\forall x \in \mathbb{R}^n$, $\|x\|_\infty$, $\max\{|x_i| 1 \leq i \leq n\}$.

Les définitions des normes 1 et 2 vont pouvoir se généraliser pour tout $p \geq 1$ [3], par

$$\forall x \in \mathbb{R}^n, \qquad \|x\|_p = \left(\sum |x_i|^p\right)^{\frac{1}{p}}.$$

La norme $\|\cdot\|_p$ est communément appelée la ***norme p***. Montrer le fait que c'est une norme est uniquement délicat pour ce qui est de l'inégalité triangulaire, la preuve de ce fait se base sur des inégalités de convexités dites de Hölder, on n'abordera pas cette preuve dans ce cours. Les normes p ne seront que très localement utilisées en dehors des cas standards des normes 1, 2 et $+\infty$, il est cela dit usuel d'en connaître la définition.
À partir d'une norme on va être en mesure de définir :

Une notion de *distance* entre deux points de $\mathbb{R}^n$ par la norme du vecteur qui relie les deux points en question. Plus formellement étant donné une norme $\|\cdot\|$ on note

$$\forall x, y, \qquad d_{\|\cdot\|}(x, y) = \|x - y\|.$$

Dans le cas des normes p on se contente d'indiquer p en indice.

2. Il est très probable qu'elle ne fasse que partie de votre culture ...
3. Il est tout à fait naturel de se poser la question de savoir pourquoi $p \geq 1$! :)

Question 3-3 Représenter graphiquement les distances 1, 2 et $+\infty$ entre deux vecteurs de $\mathbb{R}^2$.

Une notion de *voisinage* d'un point de $\mathbb{R}^n$ au sens de la norme utilisée. Cette notion est éminemment liée à la première, on l'isole ici parce qu'elle apporte un point de vue auquel vous n'avez pas encore été confrontés. Étant donné un nombre réel $\varepsilon > 0$ on va qualifier ε-voisinage d'un point $x \in \mathbb{R}^n$ tous les points à distance (au sens de la norme utilisée) au plus ε de x. Dans le jargon, on appelle ***boule ouverte de rayon ε et centrée en x*** cette notion, pour une norme ambiante $\|\cdot\|$ on note

$$B_{\|\cdot\|}(x, \varepsilon) = \{y \in \mathbb{R}^n \mid d_{\|\cdot\|}(x, y) < \varepsilon\}$$

La ***boule fermée centrée en x et de rayon ε*** est la notion correspondante avec les points à distance ε incluses,

$$\overline{B}_{\|\cdot\|}(x, \varepsilon) = \{y \in \mathbb{R}^n \mid d_{\|\cdot\|}(x, y) \leq \varepsilon\}.$$

Dans le cas des normes p on se contente d'indexer les boules par p.
La notion de ε-voisinage permet d'exprimer des phénomènes de passage à la limite et d'études locales. Dire qu'une suite $(u_k)_{k\in\mathbb{N}}$ de points de $\mathbb{R}^n$ converge vers $\ell \in \mathbb{R}^n$, au sens d'une norme $\|\cdot\|$, correspond au fait de dire que pour tout ε-voisinage $B(\ell, \varepsilon)$ de ℓ, il existe un rang N à partir duquel tous les éléments de la suite u_k sont dans $B(\ell, \varepsilon)$.

Question 3-4
- Dessiner les boules unités $\overline{B}_p(\underline{0}, 1)$ pour $p \in \{1, 2, \infty\}$.
- Montrer que les ε-voisinages d'une norme $\|\cdot\|$ sur $\mathbb{R}^n$ sont convexes.
- La définition d'un équivalent à une norme p pour $p < 1$ définit-il une norme ?

On reprend plus en détails quelques extensions des notions liées aux études locales de fonctions ou suites dans $\mathbb{R}$.

Exemple 3.1 (Convergence d'une suite.)**.** Une suite $(u_k)_{k\in\mathbb{N}}$ dans $\mathbb{R}^n$ converge vers un point $\ell \in \mathbb{R}^n$, au sens d'une norme $\|\cdot\|$ si

$$\forall \varepsilon > 0, \exists N \in \mathbb{N}, \quad k \geq N \Rightarrow \|u_k - \ell\| < \varepsilon.$$

Cette défintion se traduit telle quelle par :

$$\forall \varepsilon > 0, \exists N \in N \quad k \geq N \Rightarrow u_k \in B(\ell, \varepsilon).$$

Ou encore : pour tout $\varepsilon > 0$ il existe un rang N à partir duquel tous les éléments de la suite u_k sont dans le ε-voisinage de ℓ pour la norme $\|\cdot\|$.
Dans le cas de la norme infinie, la définition précédente s'écrit explicitement comme

$$\forall \varepsilon > 0, \exists N \in \mathbb{N}, \quad k \geq N \Rightarrow \max_{1\leq i\leq n} \|u_{k,i} - \ell_i\| < \varepsilon.$$

où $u_{k,i}$ (resp. ℓ) est la composante le long de la coordonnée i du vecteur u_k (resp. ℓ_i). Cette définition peut encore être réécrite

$$\forall \varepsilon > 0, \exists N \in \mathbb{N}, \quad k \geq N \Rightarrow \forall i \in \{1, \ldots, n\}, \|u_{k,i} - \ell_i\| < \varepsilon.$$

Un peu de réflexion permet de voir que cette dernière est équivalente

$$\forall i \in \{1, \ldots, n\}, \forall \varepsilon > 0, \exists N \in \mathbb{N}, \quad k \geq N \Rightarrow \|u_{k,i} - \ell_i\| < \varepsilon.$$

Autrement dit, la suite (u_k) converge vers ℓ, au sens de la norme infinie, si et seulement si, chacune des suites **numériques** $(u_{k,i})$ converge vers ℓ_i.

Exemple 3.2 (Limite d'une fonction en un point.)**.** On considère deux normes $\|\cdot\|_\alpha$ et $\|\cdot\|_\beta$ respectivement sur $\mathbb{R}^n$ et $\mathbb{R}^m$. Une fonction $f : (\mathbb{R}^n, \|\cdot\|_\alpha) \to (\mathbb{R}^m, \|\cdot\|_\beta)$ a une limite $\ell \in \mathbb{R}^m$ en un point $a \in \mathbb{R}^n$ si

$$\forall \varepsilon > 0, \exists \eta > 0, \quad x \neq a \text{ et } \|x - a\|_\alpha < \eta \Rightarrow \|f(x) - \ell\|_\beta < \varepsilon.$$

Chose qu'on peut encore exprimer par

$$\forall \varepsilon > 0, \exists \eta > 0, \quad x \in B_{\|\cdot\|_\alpha}(a, \eta) \setminus \{a\} \Rightarrow f(x) \in B_{\|\cdot\|_\beta}(\ell, \varepsilon),$$

ou encore, pour tout ε-voisinage B de ℓ pour la norme $\|\cdot\|_\alpha$ il existe un η-voisinage de a, a étant exclu, dont tout élément a une image dans B.

Exemple 3.3 (Continuité d'une fonction en un point.)**.** La continuité d'une fonction $f : (\mathbb{R}^n, \|\cdot\|_\alpha) \to (\mathbb{R}^m, \|\cdot\|_\beta)$ en un point $a \in \mathbb{R}^n$ s'écrit

$$\forall \varepsilon > 0, \exists \eta > 0, \quad \|x - a\|_\alpha < \eta \Rightarrow \|f(x) - f(a)\|_\beta < \varepsilon.$$

Autrement dit pour tout ε-voisinage B de $f(a)$ pour la norme $\|\cdot\|_\alpha$ il existe un η-voisinage de a dont tout élément a une image dans B.

Question 3-5

- Donner des exemples de fonctions continues (pour les normes de votre choix) de $\mathbb{R}^2$ dans $\mathbb{R}$.
- Comment peut-on en construire d'autres à partir de celles-ci ? Que dire de l'addition, la multiplication, le quotient ou la composition de fonctions continues ?
- Comment tester la continuité de fonctions à plusieurs variables en se ramenant au cas des fonctions numériques ?
- Est-ce que ça marche tous le temps ?

Exemple 3.4 (Comparaisons en un point.)**.** Une fonction $f : (\mathbb{R}^n, \|\cdot\|_\alpha) \to (\mathbb{R}^m, \|\cdot\|_\beta)$ est un o en un point $a \in \mathbb{R}^n$ d'une fonction g, sur les mêmes espaces et par rapport aux mêmes normes, s'il existe une fonction $\epsilon : \mathbb{R}^n \to \mathbb{R}$ telle que $f = \epsilon g$ avec $\lim_{t \to a} \epsilon(t) = 0$. Quand g n'est pas nulle sur un voisinage de a la définition précédente est équivalente au fait que

$$\lim_{t \to a} \frac{\|f(t)\|_\alpha}{\|g(t)\|_\beta} = 0.$$

On écrit dans ce cas que f est un $o_a(g)$. On laisse tomber le point a quand celui-ci est clair du contexte. Pour obtenir la notion d'équivalence en un point a, il suffit de remplacer la limite de ϵ ou de quotient précédente par 1. Pour obtenir la notion de $\mathcal{O}$ il suffit d'attendre ϵ ou du quotient d'être borné au voisinage de a.

Question 3-6 Justifier les affirmations suivantes :

1. $\langle h, h\rangle$ est un $o_{\underline{0}}(h)$;
2. $\sin\big(\langle h, h\rangle\big) \sim_{\underline{0}} \langle h, h\rangle$.

À ce stade on est en droit de se demander si le choix des normes avec lesquelles on travaille a une incidence sur les limites des suites qu'on étudie (une même suite convergerait pour une norme et pas pour une autre ?) ou les fonctions qu'on regarde (une même fonction serait continue pour une norme et pas pour une autre ?) etc. Le théorème suivant permet de se rassurer sur ce point (du moins en dimension finie).

Définition 3.2. Deux normes $\|\cdot\|_\alpha$ et $\|\cdot\|_\beta$ sur $\mathbb{R}^n$ sont dites ***équivalentes*** s'il existe des constants c, $C \in \mathbb{R}_+^*$ telles que

$$\forall x \in \mathbb{R}^n, \qquad c\|x\|_\alpha \leq \|x\|_\beta \leq C\|x\|_\alpha.$$

En reprenant les exemples précédents, il est relativement simple de se convaincre que deux normes équivalentes donnent des suites de mêmes natures, les mêmes fonctions continues et les mêmes comparaisons.

Question 3-7 Montrer que les normes 1, 2 et ∞ sont équivalentes.

Théorème 3.1. *Toutes les normes sur $\mathbb{R}^n$ sont équivalentes.*

Remarque 1. La preuve de ce théorème est en dehors du périmètre de ce cours. Elle nécessite certaines notions de topologie qui vous manquent ; notamment la notion de compacité. On se contentera du résultat.

Le fait que les normes soient équivalentes en terme de questionnement topologique (les études des phénomènes locaux) ne signifie pas qu'utiliser l'une ou l'autre en modélisation revient au même.

Question 3-8 Quelles normes utiliseriez-vous pour modéliser les problématiques suivantes :

1. Le calcul de la distance que doit parcourir un oiseau entre deux coordonnées GPS pas trop éloignées.
2. Le calcul de la distance que parcours un touriste à Manhattan entre deux musée.
3. Le calcul du nombre de modifications nécessaires (lettre à lettre) pour changer un mot en un autre.

4 Différentiabilité et différentielle en un point

Cette section a pour objectif de vous apporter les éléments théoriques nécessaires pour pouvoir étendre les outils de calcul différentiel en dimension 1, dérivation et études d'extrema locaux, au cas de la dimension supérieure.

4.1 Dérivabliité et dérivée : quelques rappels

On se donne dans ce paragraph une fonction $f : I \to \mathbb{R}$ sur un intervalle ouvert $I \subset \mathbb{R}$ et un point $a \in I$. On rappelle que f est dite ***dérivable*** en $a \in I$ si le taux d'accroissement

$$\frac{f(a+h) - f(a)}{h}$$

a une limite *finie* quand h tend vers 0. Sous une telle condition on appelle ***nombre dérivé*** de f en a cette limite, il est noté $f'(a)$. Si f est dérivable en tout point de I on appelle ***dérivée*** de f la fonction $f' : I \to \mathbb{R}$ qui à un point $x \in I$ associe le nombre dérivée de f en x, $f'(x)$. Par abus, quand f est dérivable en un point on parle de ***dérivée de f en a***. Le nombre dérivé de f en a quand il existe a une interprétation géométrique : la droite $D_{a,h}$

$$y = \left[\frac{f(a+h) - f(a)}{h}\right](x - a) + f(a)$$

est une droite sécante au graphe de f passant par les points $(a, f(a))$ et $(a+h, f(a+h))$. Dire que f est dérivable en a signifie que ces droites sécantes admettent une droite limite $T_{f,a}$ qu'on appelle droite tangent au graphe de f en $(a, f(a))$. Le coefficient directeur de cette droite est $f'(a)$ et son équation est

$$y = f'(a)(x - a) + f(a).$$

Décrire la dérivabilité d'une fonction f en un point a par une condition sur le taux d'accroissement n'est pas adapté si l'on souhaite étendre la notion de dérivée à des fonctions entre espaces vectoriels de dimensions supérieures. La raison principale en est le fait que le quotient, correspondant au taux d'accroissement, perd de ces propriétés dès qu'on cherche à l'étendre au cas de vecteurs à plusieurs entrées. Une manière d'aborder plus sereinement l'extension de la question de dérivabilité à la dimension supérieure est de passer par l'existence d'un développement limité à l'ordre 1.

Proposition 4.1. *Une fonction $f : I \to \mathbb{R}$ définie sur un intervalle ouvert $I \subset \mathbb{R}$ est dérivable en un point a si et seulement si il existe un nombre réel α tel que, pour tout h dans un voisinage de* 0

$$f(a+h) = f(a) + \alpha h + o_0(h).$$

Sous une telle condition α est le nombre dérivée de f en a et donc $\alpha = f'(a)$.

On dit dans ce contexte que la fonction affine $f(a) + f'(a)h$, comme fonction de h, est la meilleure approximation affine de f au voisinage de a. Le terme en o est un terme qui encapsule des ordres de grandeurs en h^ℓ pour $\ell > 1$. La preuve de cette proposition ne présente pas de difficultés particulières, et n'est pas d'intérêt pour nous au-delà de l'intuition qu'elle peut nous apporter. Elle vous est laissé en exercice, si vous en ressentez le besoin.

Remarque 2. Dans la suite de ce document on explicite les notations de Landau (o, $\mathcal{O}$) pour éviter toute confusion. On écrira ainsi, avec les notations de la proposition précédente :

$$f(a+h) = f(a) + \alpha h + \epsilon(h)h.$$

avec $\epsilon(h) \xrightarrow[h\to 0]{} 0$.
La description précédente peut également être reformulée sous la forme suivante :

Proposition 4.2. *Une fonction $f : I \to \mathbb{R}$ définie sur un intervalle ouvert $I \subset \mathbb{R}$ est dérivable en un point a si et seulement si il existe une application linéaire $\lambda \in \mathcal{L}(\mathbb{R}, \mathbb{R})$ telle que, pour tout h dans un voisinage de* 0

$$f(a+h) = f(a) + \lambda(h) + \epsilon(h)h,$$

où $\epsilon(h) \xrightarrow[h\to 0]{} 0$. Sous une telle condition λ est la fonction $h \mapsto f'(a)h$.

Pour comprendre le lien entre ces deux manière de voir il faut se rappeler que toute application linéaire de $\mathbb{R}$ dans lui même est de la forme $h \mapsto \alpha h$.
C'est cette dernière formulation qui va nous permettre à l'aide des outils introduits d'étendre la notion de dérivabilité ou de dérivée au cas de la dimension supérieure.

4.2 Définition de la différentiabilité et de la différentielle

Définition 4.1. Soit $f : U \to \mathbb{R}^m$ une fonction définie sur un ouvert $U \subset \mathbb{R}^n$ et soit $a \in U$. On dit que f est ***différentiable*** en a s'il existe une application linéaire $\lambda \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ telle que pour tout h dans un voisinage de $0 \in \mathbb{R}^n$

$$f(a+h) = f(a) + \lambda(h) + \epsilon(h)\|h\|$$

où ϵ est une application d'un voisinage de 0 dans $\mathbb{R}^n$ vers $\mathbb{R}^m$ telle que $\epsilon(h) \xrightarrow[h\to 0]{} 0$.

Quand f est différentiable en a on appelle ***différentielle de f en*** a il n'existe qu'une seul application linéaire qui satisfait les propriétés ci-dessus satisfaites par λ. C'est une fonction de h dont la définition dépend de f et du point a, elle est notée dans ce cours $\mathbf{D}f(a)$. Ainsi, $\mathrm{D}f(a)$ est une application linéaire de $\mathbb{R}^n$ dans $\mathbb{R}^m$ qui prend en paramètre h. On écrit avec les notations précédentes :

$$f(a+h) = f(a) + \mathrm{D}f(a)(h) + \epsilon(h)\|h\|$$

où $\epsilon(h) \xrightarrow[h\to 0]{} 0$.
Remarque 3. Le choix de la norme de h dans la définition de la différentielle n'a pas d'importance ; toutes les normes étants équivalentes quelque soit la norme prise les limites ne changeront pas.

Question 4-9

1. Expliquer, avec vos propres mots, comment retrouver la définition de dérivabilité d'une fonction de $\mathbb{R}$ dans $\mathbb{R}$ en un point.

2. Calculer la différentielle de l'application sur $\mathbb{R}$, $f(x) = \frac{\sin(x)}{x^2+1}$ en tout point x. Quelle est la différence avec la dérivée de f.

Question 4-10 Calculer en développant les expressions $f(X+h)$ les différentielles en tout point X des applications suivantes :

1. la fonction $f : \mathbb{R}^n \to \mathbb{R}^m$ donnée par $f(X) = AX + b$ où $A \in \mathcal{M}_{m,n}(\mathbb{R})$ et $b \in \mathbb{R}^m$;
2. la fonction $f : \mathbb{R}^n \to \mathbb{R}$ donnée par $f(X) = X^T AX$ pour A une matrice symétrique dans $\mathcal{M}_n(R)$;
3. la fonction $f : \mathcal{M}_n(\mathbb{R}) \to \mathbb{R}$ donnée par $f(X) = \mathrm{tr}(X)^2$;
4. $f : \mathcal{M}_n(\mathbb{R}) \to \mathcal{M}_n(\mathbb{R})$ donnée par l'expression $B \mapsto \mathrm{tr}(AB)B$ où A est une matrice carrée de taille (n, n).

Tout comme le cas de la dimension 1 il est particulièrement utile de comprendre comment se comporte la différentiabilité et la différentielle vis-à-vis des opérations algébriques. Les propriétés que vous connaissez dans le cas de la dimension 1 se généralisent plutôt bien.

Proposition 4.3. *Soient f, $g : U \to \mathbb{R}^m$ deux fonctions définies sur un ouvert $U \subset \mathbb{R}^n$ et différentiables en $a \in U$. Alors,*

1. *pour tout $\lambda \in \mathbb{R}$, la fonction $f + \lambda g$ est différentiable en a et*

$$\mathrm{D}(f + \lambda g)(a) = \mathrm{D}f(a) + \lambda\, \mathrm{D}g(a)$$

2. *la fonction $\langle f, g\rangle$ est différentiable en a et*

$$\mathrm{D}(\langle f, g\rangle)(a) = \langle \mathrm{D}f(a), g(a)\rangle + \langle f(a), \mathrm{D}g(a)\rangle.$$

Ces deux relations sont des relations fonctionnelles, c'est-à-dire que des deux côté des égalités on a des fonctions en h. Plus précisément, pour tout h dans un voisinage de 0

$$\mathrm{D}(f + \lambda g)(a)(h) = \mathrm{D}f(a)(h) + \lambda\, \mathrm{D}g(a)(h)$$
$$\mathrm{D}(\langle f, g\rangle)(a)(h) = \langle \mathrm{D}f(a)(h), g(a)\rangle + \langle f(a), \mathrm{D}g(a)(h)\rangle.$$

Question 4-11 Calculer la différentielle en tout point de la fonction $f : \mathbb{R}^n \to \mathbb{R}$ donnée par

$$f(X) = \langle AX + b, \mathrm{tr}(A)X\rangle$$

avec $A \in \mathcal{M}_n(\mathbb{R})$ et $b \in \mathbb{R}^m$.

La relation qui vous permet de dériver des fonctions composées s'étend également au contextes de plus grandes dimensions pour donner la proposition suivante :

Proposition 4.4. *Soient $f : U \to \mathbb{R}^m$ et $g : V \to \mathbb{R}^p$ des fonctions respectivement définies sur les ouverts $U \subset \mathbb{R}^n$ et $V \subset \mathbb{R}^m$. On considère un point $a \in U$ d'image $f(a) \in V$. Si f est différentiable en a et g est différentiable en $f(a)$ alors $g \circ f$ est différentiable en a et pour tout $h \in \mathbb{R}^n$ dans un voisinage de* 0

$$\mathrm{D}(g \circ f(a)(h) = \mathrm{D}g\big(f(a)\big)(\mathrm{D}f(a)(h))$$

ou encore

$$\mathrm{D}(g \circ f(a) = \mathrm{D}g\big(f(a)\big) \circ \mathrm{D}f(a).$$

Question 4-12 Calculer la différentielle des fonctions suivantes :

1. $g : \mathbb{R}^n \to \mathbb{R}$ donnée par l'expression $X \mapsto 1/(X^T X + 1)$.
2. $f : \mathbb{R}^n \to \mathbb{R}$ donnée par l'expressoin $X \mapsto \cos^2(X^T AX)$, où A est une matrice carrée symétrique de taille (n, n) ;

5 Jacobienne et gradient de fonctions en un point

Toute application linéaire dans $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ peut être représentée par sa matrice dans les bases canoniques de $\mathbb{R}^n$ et $\mathbb{R}^m$. Dans ce cas, si f désigne un élément dans $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ et $M \in \mathcal{M}_{m,n}(\mathbb{R})$ sa matrice dans les bases canoniques, on retrouve l'image $f(x)$ d'un vecteur $x \in \mathbb{R}^n$ par f en effectuant le produit Mx. La différentielle en un point d'une application différentiable en ce même point est également une application linéaire de $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$, elle peut donc être représentée par son écriture matricielle. Comprendre cette écriture matricielle dans le détail est le sujet de cette section.

5.1 Définition de la jacobienne (gradient) d'une fonction différentiable en un point

La définition suivante est une définition temporaire, elle ne couvre, telle quelle, qu'une seule partie de la définition admise généralement.

Définition 5.1 (*Temporaire*)**.** On appelle ***matrice jacobienne en a*** d'une fonction différentiable $f : U \to \mathbb{R}^m$, en un point $a \in U$ d'un ouvert $U \subset \mathbb{R}^n$ la matrice dans les bases canoniques de $\mathbb{R}^n$ et $\mathbb{R}^m$ de $\mathrm{D}f(a)$. La matrice jacobienne de f en a est notée $\mathcal{J}_f(a)$.

Avec les notations précédentes et en écrivant les vecteurs dans les coordonnées usuelles de la base canonique, on aurait pour tout h dans un voisinage de 0

$$f(a + h) = f(a) + \mathcal{J}_f(a)h + \epsilon(h)\|h\|$$

où $\epsilon(h) \xrightarrow[h\to 0]{} 0$. Ici, $\mathcal{J}_f(a)h$ désigne le produit matriciel de $\mathcal{J}_f(a)$ par le vecteur h.
Les propriétés de la différentielle décrites au cours de la section précédente se répercutent sur les jacobiennes correspondantes, on a donc :

1. Si f, $g : U \to \mathbb{R}^m$ sont deux fonctions définies sur un ouvert $U \subset \mathbb{R}^n$ et différentiables en $a \in U$. Alors,
   (a) pour tout $\lambda \in \mathbb{R}$, la fonction $f + \lambda g$ est différentiable en a et
   $$\mathcal{J}_{f+\lambda g}(a) = \mathcal{J}_f(a) + \lambda \mathcal{J}_g(a)$$
   (b) la fonction $\langle f, g\rangle$ est différentiable en a et
   $$\mathcal{J}_{\langle f,g\rangle}(a) = g(a)^T \mathcal{J}_f(a) + f(a)^T \mathcal{J}_g(a)\,^4.$$
2. Si $f : U \to \mathbb{R}^m$ et $g : V \to \mathbb{R}^p$ sont des fonctions respectivement définies sur les ouverts $U \subset \mathbb{R}^n$ et $V \subset \mathbb{R}^m$. On considère un point $a \in U$ d'image $f(a) \in V$. Si f est différentiable en a et g est différentiable en $f(a)$ alors $g \circ f$ est différentiable en a
   $$\mathcal{J}_{g\circ f}(a) = \mathcal{J}_g\big(f(a)\big) \times \mathcal{J}_f(a).$$

Question 5-13 Identifier les matrices jacobiennes aux points considérés des fonctions dans les questions de la section précédente.

4. Vous ne voyez pas ourquoi ? Posez une question !

Dans le cas d'une fonction $f : U \to \mathbb{R}$ définie sur un ouvert $U \subset \mathbb{R}^n$ et différentiable en un point $a \in U$ la jacobienne $\mathcal{J}_f(a)$ est une matrice ligne dans $\mathcal{M}_{1,n}(\mathbb{R})$.

Définition 5.2. Soit $f : U \to \mathbb{R}$ une fonction définie sur un ouvert $U \subset \mathbb{R}^n$ et $a \in U$. Quand la jacobienne de f en a existe, on appelle ***gradient de f en*** a la vecteur colonne transposée de celle-ci, il est noté $\nabla f(a)$. De manière plus concise

$$\nabla f(a) = \mathcal{J}_f(a)^T.$$

Dans ce cadre, pour h dans un voisinage de 0 on peut écrire

$$f(a+h) = f(a) + \nabla f(a)^T h + \epsilon(h)\|h\|$$

où $\epsilon(h) \xrightarrow[h\to 0]{} 0$.

Question 5-14 Identifier les gradients des fonctions de la section précédente quand cela fait sens.

Une fonction $f : U \to \mathbb{R}^m$ définie sur un ouvert $U \subset \mathbb{R}^n$ n'est que la concaténation de m fonctions appelées ***composantes de f***. En effet une fonction telle que f est décrite par ses fonctions *coordonnées*, $f = (f_1, \dots, f_n)$ où pour tout $i \in \{1, \dots n\}$, $f_i : U \to \mathbb{R}$. Par exemple, si on considère la fonction affine

$$f : \begin{pmatrix} x \\ y \end{pmatrix} \longmapsto \begin{pmatrix} 1 & 2 \\ -1 & 3 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

on est en train de parler de la fonction f de $\mathbb{R}^2$ dans $\mathbb{R}^2$ dont les fonctions coordonnées sont respectivement données par

$$\begin{aligned} f_1(x, y) &= x + 2y \\ f_2(x, y) &= -x + 3y. \end{aligned}$$

En ayant cela en tête, on peut exprimer la jacobienne de f en un point à l'aide des gradients de ses composantes f_i pour $i \in \{1, \dots, n\}$. En effet, dire que f est différentiable en un point $a \in U$ signifie que pour tout $i \in \{1, \dots, n\}$ f_i est différentiable en a car composée des fonctions f et $(x_1, \dots, x_n) \mapsto x_i$ qui sont toutes deux différentiables, respectivement en a et en tout point. On a donc, pour chaque $i \in \{1, \dots, n\}$ un voisinage de $0 \in \mathbb{R}^n$ sur lequel

$$f_i(a+h) = f_i(a) + \mathrm{D}f_i(a)(h) + \varepsilon_i(h)\|h\|$$

où $\epsilon(h) \xrightarrow[h\to 0]{0}$. Ainsi, en écrivant les choses vectoriellement, il existe un voisinage de $0 \in \mathbb{R}^n$ sur lequel

$$f(a+h) = \begin{pmatrix} f_1(a+h) \\ \vdots \\ f_n(a+h) \end{pmatrix} = \begin{pmatrix} f_1(a) \\ \vdots \\ f_n(a) \end{pmatrix} + \begin{pmatrix} \mathrm{D}f_1(a) \\ \vdots \\ \mathrm{D}f_n(a) \end{pmatrix} (h) + \begin{pmatrix} \epsilon_1(h) \\ \vdots \\ \epsilon_b(h) \end{pmatrix} \|h\|$$

Dans l'égalité de droite, le dernier terme est négligeable au voisinage de 0 en h, le second est linéaire en h. On en obtient donc la proposition suivante :

Proposition 5.1. *Avec les notations précédentes*

$$\mathrm{D}f(a) = \begin{pmatrix} \mathrm{D}f_1(a) \\ \vdots \\ \mathrm{D}f_n(a) \end{pmatrix} = \begin{pmatrix} \nabla f_1(a)^T \\ \vdots \\ \nabla f_n(a)^T \end{pmatrix}$$

On peut en particulier ramener l'étude d'une fonction à valeur dans $\mathbb{R}^m$ à celle de fonctions à valeurs dans $\mathbb{R}$.

5.2 Dérivées partielles

Nous avons entamés dans la section précédente l'étude sur la structure de la matrice jacobienne d'une application différentiable $f : U \to \mathbb{R}^m$ sur un ouvert $U \subset \mathbb{R}^n$. Les lignes de celle-ci correspondent aux gradients des m fonctions coordonnées de f. On poursuit désormais cette étude dans le but d'analyser plus finement les coefficients de la jacobienne. D'après ce qui précède on peut se limiter au cas des fonctions $f : U \to \mathbb{R}$ définies sur un ouvert $U \subset \mathbb{R}^m$, on se limite donc à ces cas.
Par définition, la jacobienne de f en un point $a \in U$ correspond à la matrice de sa différentielle en a, $\mathrm{D}f(a)$, dans les bases canoniques. Pour rappelle $\mathrm{D}f(a)$ peut-être représentée par une matrice ligne ayant n entrées, son i-ème entrée est donc l'image par $\mathrm{D}f(a)$ du i-ème élément e_i de la base canonique de $\mathbb{R}^n$; plus formellement

$$\nabla f(a)_i = \mathrm{D}f(a)(e_i).$$

Ce second membre peut en réalité être calculé de manière relative explicite si l'on choisit d'écrire f en coordonnées [5], voici comment s'y prendre : quitte à prendre $t \in \mathbb{R}$ assez proche de 0, te_i dans un voisinage de 0 dans $\mathbb{R}^n$ pour lequel on peut écrire

$$f(a + te_i) = f(a) + \mathrm{D}f(a)(te_i) + \epsilon(te_i)\|te_i\|$$

avec $\epsilon(te_i) \xrightarrow[t\to 0]{} 0$. Comme la différentielle est une application linéaire, pour $t \neq 0$ on peut écrire

$$\mathrm{D}f(a)(e_i) = \frac{f(a + te_i) - f(a)}{t} \mp \epsilon(te_i)\|e_i\|$$

On obtient ainsi la relation

$$\nabla f(a)_i = \mathrm{D}f(a)(e_i) = \lim_{t\to 0} \frac{f(a + te_i) - f(a)}{t}.$$

Si l'on écrit cela explicitement en coordonnées (avec les précautions d'usage pour la validité des indices), en notant $a = (a_1, \ldots, a_n)$ on vient d'écrire

$$\nabla f(a)_i = \mathrm{D}f(a)(e_i) = \lim_{t\to 0} \frac{f(a_1, \ldots, a_{i-1}, a_i + t, a_{i+1}, \ldots, a_n) - f(a_1, \ldots, a_i, \ldots, a_n)}{t}. \tag{1}$$

On regarde donc le taux d'accroissement de la fonction f, comme si l'on fixait toute les variables sauf la i-ème. Ce qu'on peut encore exprimer de la manière suivante : pour chaque $i \in \{1, \ldots, n\}$ on note φ_i la fonction partielle de f en $a = (a_1, \ldots, a_n)$, c'est-à-dire la fonction donnée par

$$\varphi_i(x) = f_i(a_1, \ldots, x, \ldots, a_n),$$

avec cette notation la relation (1) s'écrit

$$\nabla f(a)_i = \mathrm{D}f(a)(e_i) = \varphi_i'(a_i).$$

Définition 5.3. Soient $f : U \to \mathbb{R}$ une fonction définie sur un ouvert $U \subset \mathbb{R}^n$ et $i \in \{1, \ldots, n\}$. On dit que f admet une ***dérivée partielle*** par rapport à la i-ème variable en un point $a = (a_1, \ldots, a_n)$ si la fonction partielle φ_i en a admet une dérivée en a_i. Dans ce cas on note $\frac{\partial f}{\partial x_i}(a)$ cette dérivée en a, elle est appelé ***dérivée partielle de f en a par rapport à la i-ème variable***.

Proposition 5.2. *Soit $f : U \to \mathbb{R}^n$ une fonction définie sur un ouvert $U \subset \mathbb{R}^n$ et admettant un gradient en $a \in U$. Le gradient de f en a est donné par la relation*

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

5. C'est rigolo pour résoudre des exos, mais dans la pratique on ne se retrouve pas souvent avec ce type d'écriture.

Sous les conditions de la propositions ci-dessus, dans le cas plus général d'une fonction $f : U \to \mathbb{R}^m$, on trouve l'expression suivante de la matrice jacobienne de f en a

$$\mathcal{J}_f(a) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix}.$$

Question 5-15 Expliciter le gradient, en tout point ou cela fait sens, des expressions différentiables suivantes :

1. $f(x, y) = e^{xy}(x + y)$;
2. $g(x, y, z) = (x + y\ln(z), xyz)$;
3. $h(x, y, z) = \left(\dfrac{x^2 \sin(xy)}{z^2 + 2}, \tan(xyz)\right)$;
4. $f(x, y, z) = \dfrac{x + y + z}{x + y^2 + 1}$;
5. $g(x, y) = \dfrac{\cos(xy)}{\sqrt{x^2 + y^2}}$;
6. $h(x, y, z) = \exp(xy - z)$.

Question 5-16 Soit $f : \mathbb{R}^3 \to \mathbb{R}$ une fonction différentiable en tout point de $\mathbb{R}^3$. On considère la fonction $g : \mathbb{R}^3 \to \mathbb{R}$ donnée pour tout $(x, y, z) \in \mathbb{R}^3$ par

$$g(x, y, z) = f(x - y, y - z, z - x).$$

La fonction g est différentiable en tout point car composée de fonction différentiables. Montrer qu'on a la relation

$$\frac{\partial g}{\partial x}(a, b, c) + \frac{\partial g}{\partial y}(a, b, c) + \frac{\partial g}{\partial z}(a, b, c) = 0$$

pour tout point $(a, b, c) \in \mathbb{R}^3$.

5.3 Dérivées directionnelles

Les dérivées partielles sont un cas particulier de ce qu'on appelle les ***dérivées directionnelles***. Pour rappel, si $f : U \to \mathbb{R}$ est une fonction définie en $a \in U$, pour chaque $i \in \{1, \ldots, n\}$ on a étudié le rapport

$$\frac{f(a + te_i) - f(a)}{t}$$

autrement dit le taux d'accroissement de f dans la direction indiquée par e_i.

Définition 5.4. Soit $f : U \to \mathbb{R}$ une fonction définie en $a \in U$. On dit que f admet une ***dérivée directionnelle suivant le vecteur*** $v \in \mathbb{R}^n$ si le taux d'accroissement

$$\frac{f(a + tv) - f(a)}{t}$$

admet une limite finie quand $t \to 0$. Si tel est le cas on note $\partial_v f(a)$ le nombre réelle correspondant à cette limite.

Remarque 4. Avec les notation ci-dessus, vous pourriez retrouver la notation ∂_{e_i} ou encore, par abus, ∂_i pour la i-ème dérivée partielle $\frac{\partial}{\partial x_i}$ introduite au cours de la section précédente.

Question 5-17 On considère la fonction définie sur $\mathbb{R}^2$ par $f(x,y) = \frac{x^2\cos(y)}{y+\sin^2(x)+1}$. Pour tout $\vartheta \in \mathbb{R}$ on désigne par w_ϑ le vecteur (ϑ, ϑ^2).

1. Calculer la dérivée directionnelle $\partial_{w_\vartheta} f(0,0)$ en fonction de ϑ ;
2. Étudier cette fonction de ϑ.
3. Pouvez-vous interpréter vos résultats géométriquement ?

5.4 Retour sur la définition de jacobienne

On revient sur la définition *temporaire* de jacobienne donnée définition (5.1). La définition de dérivée partielle donnée définition (5.3) ne présuppose rien de la différentiabilité de la fonction en jeu, cela est vrai de toute dérivée directionnelle.

Définition 5.5 (Jacobienne)**.** Soient $f : U \to \mathbb{R}^m$ une fonction définie sur un ouvert $U \subset \mathbb{R}^n$ et $a \in U$. Si $f = (f_1, \dots, f_m)$ admet des dérivées partielles en a, on appelle ***jacobienne*** de f en a la matrice

$$\mathcal{J}_f(a) = \left(\frac{\partial f_i}{\partial x_j}(a)\right)_{\substack{1\le i\le m\\ 1\le j\le n}}.$$

Dans le cas où f est à valeur dans $\mathbb{R}$ le gradient de f en a est défini comme la transposée de sa jacobienne, donc sous la seule condition d'existence de dérivées partielles en a.

Proposition 5.3. *Soit $f : U \to \mathbb{R}^m$ une fonction définie sur un ouvert $U \subset \mathbb{R}^n$ et $a \in U$. Si f admet des dérivées partielles en a qui sont continues au voisinage de a alors f est différentiable en a, de différentielle continue.*

On ne peut pas s'abstraire de la continuité des dérivées partielles dans notre cas La réciproque à cette proposition est fausse, il existe des fonctions admettant des dérivées partielles sans pour autant différentiables. Le plus simple de ces exemples est le suivant : on considère la fonction $f : \mathbb{R}^2 \to \mathbb{R}$ définie par

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

On peut montrer que cette fonction est continue en $(0,0)$ (elle est lipschitzienne de rapport 1 en $(0,0)$) mais non différentiable en ce point (un peu technique). Elle admet pourtant des dérivées partielle nulles en $(0,0)$, celles-ci sont cependant non continues en $(0,0)$.

Remarque 5. Imaginer remplacer l'existence de dérivées partielles par l'existence de dérivées directionnelles dans toutes les directions, dans le but d'obtenir un résultat de différentiabilité sans contraindre la continuité des dérivées directionnelles, n'est également pas suffisant. Pour vous en convaincre il vous suffit d'étudier l'exemple suivant : on considère la fonction $g : \mathbb{R}^2 \to \mathbb{R}$ définie par

$$g(x,y) = \begin{cases} x & \text{si } y = x, \\ 0 & \text{sinon} \end{cases}$$

Vous pourrez constater que la restriction le long de chaque droite passant par $(0,0)$ est une fonction réelle différentiable en ce point, mais la collection des droites tangentes au graphe de cette fonction en un point ne correspond pas à un plan tangent au graphe de g.

La figure (1) résume les relations qu'on vient de décrire ici. Pour rappel, dans la situation où une fonction f est différentiable en un point a, la différentielle en ce point a une matrice dans les bases canoniques donnée par la jacobienne. Cette dernière pourrait cependant exister sans pour autant que f soit différentiable en a, cependant si les dérivées partielles (donc les fonctions coordonnées de la jacobienne) sont continue alors f est différentiable de différentielle décrite dans la base canonique par la jacobienne. On termine cette section sur un problème qui vous permettra, suivant la démarche que vous choisissez, d'aborder les différentes notions que vous avez pu étudier à ce jour. C'est un exemple **essentiel** sur lequel vous serez toujours attendus.

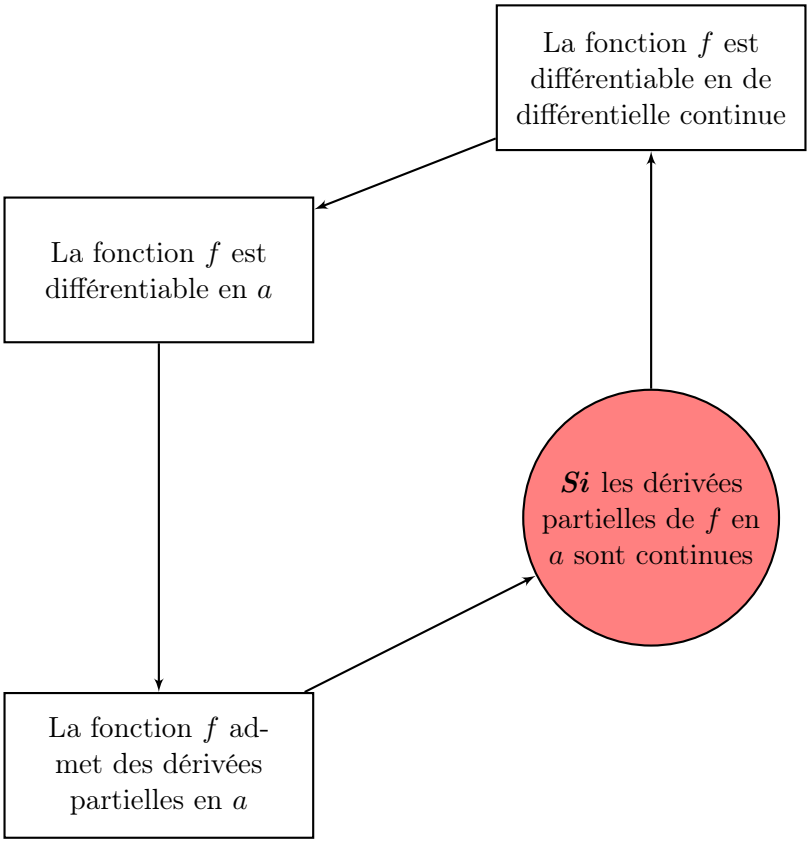

FIGURE 1 – Résumé des relations entre existence des dérivées partielles et différentiabilité

Question 5-18

1. Qu'est ce que le problème des *moindres carrés* [a] ?
2. Quel est le lien avec la régression linéaire ?
3. Quelle est la jacobienne de la fonction objectif que vous obtenez ?
4. Pourriez-vous étendre la formulation à des régressions plus générales ?

a. Oui, c'est une recherche bibliographique !

6 Espace tangent à une partie de $\mathbb{R}^n$

On consacre cette section au coeur de notre sujet : *les propriétés géométriques du gradient au coeur des méthodes itératives d'optmisation.* On y fait le choix délibéré de n'apporter que le nécessaire de formalisme et de preuves ; la thématique abordée nécessiterait un savoir-faire technique et des connaissance que nous n'avons pas à ce stade, notamment le théorème des fonctions implicites. Cela n'est cependant pas un frein à une compréhension correcte des phénomènes étudiés.

Hypothèse 6.1. On se limite uniquement aux parties suivantes de $\mathbb{R}^n$: les graphes de fonctions et les courbes de niveaux.

Une généralisation de la démarche à d'autres types de parties nous emmenerez vers la notion de sous-variété de $\mathbb{R}^n$, objet qu'on n'a ni les moyens ni le besoin d'appréhender ici.

Hypothèse 6.2. On suppose de plus que toute nos fonctions sont des fonctions différentiables d'ouverts de $\mathbb{R}^n$ dans $\mathbb{R}$. La notion d'espace tangent n'a pas de sens sans hypothèses de différentiabilité. Se limiter à un espace d'arrivée égal à $\mathbb{R}$ nous permet de se concentrer sur nos cas d'usages.

La notion d'espace tangent, qu'on aborde par la suite, est une notion qui sera attachée à *un point* d'une partie de $\mathbb{R}^n$, donnée implicitement par une fonction différentiable ou comme graphe d'une fonction différentiable. Vous avez eu l'occasion d'en voir une instance avec la notion de droite tangente au graphe d'une fonction en un point de celui-ci. Pour rappel, si $f : \mathbb{R} \to \mathbb{R}$ est une fonction différentiable la droite tangente au graphe de f au point $(a, f(a))$ est donnée (implicitement) par l'équation :

$$y - f'(a)(x - a) - f(a) = 0$$

ou encore, paramétriquement, par les points de $\mathbb{R}^2$ de la forme :

$$(a + t, f'(a)t + f(a)) = (a, f(a)) + t(1, f'(a))$$

pour tout $t \in \mathbb{R}$. La droite tangente correspond au cas de dimension 1 de l'espace tangent qu'on va introduire, remarquez qu'ici c'est un espace affine, en réalité on qualifie d'espace tangent uniquement la partie *linéaire* des description précédentes, c'est-à-dire

$$y - f'(a)x = 0$$

ou

$$(t, f'(a)t) = t(1, f'(a))$$

en gardant à l'esprit que ces des espaces vectoriels attachés au point $(a, f(a))$. On peut donc retrouver l'espace affine décrit ci-dessus à l'aide de ces deux informations.
De manière similaire à ce que nous avions introduits pour généraliser la notion de différentielle, nous allons devoir effectuer un pas de côté pour trouver une définition de droite tangente adaptable au cas de dimension supérieure.

Parfum d'espace tangent. On va retenir une idée très *physicienne* pour définir la notion d'espace tangent. Imaginez un instant que vous avez une partie V de $\mathbb{R}^n$ (pour simplifiez imaginer une courbe dans $\mathbb{R}^2$) et un point $p \in A$. On cherche à savoir comment on pourrait définir une notion d'espace affine tangent à V en p. Une physicienne ferait la chose suivante : elle prendrait un object ponctuel B soumis à des forces le contraignant à avoir une trajectoire contenue dans V et passant par p. Au moment où cet objet passe par p on élimine les contraintes sur B. La direction que prend B devrait être une direction tangente à V en p.
Pour formaliser l'idée précédente il nous suffit uniquement de conserver la notion de trajectoire de B, l'objet B en lui même n'a pas d'importance hors le fait de donner un nom à cette trajectoire. De plus, on ne s'intéresse qu'à ce qui se passe autour du point $p \in V$, on peut donc se limite à étudier la trajectoire de B autour de a. Pour résumer, et en ayant effectuer quelques simplifications de forme qui ne réduisent pas la généralité de notre étude, on s'intéresse donc aux courbes $\gamma :]-1, 1[\to V$ telle que $\gamma(0) = p$. La courbe γ correspond donc ici au vecteur position de l'objet B étudié, une direction tangente serait donc la direction instantanée que prend B en p, donc la dérivée[6] de γ **en 0**. Il nous faut donc s'imposer d'étudier les courbes γ **différentiables** en 0.
Ainsi, l'espace tangent à V en p devrait être donné par toutes les directions instantanée que prennent les courbes différentiables γ (d'image dans V et passant par p) au point 0 ; donc les dérivées de ses courbes en 0.

Définition 6.1. On appelle donc ***espace tangent à V en p***, qu'on note $\mathcal{T}_{V,p}$, le sous-ensemble :

$$\mathcal{T}_{V,p} = \{\gamma'(0) \mid \forall \gamma :]-1, 1[\to V, \gamma(0) = p\}.$$

L'espace affine qu'on représente géométriquement étant $p + \mathcal{T}_{V,p}$.

En supposant que cette définition correspond bien à notre intuition, elle vient avec son lot de questions, parmi celles-ci :

6. Fomellement le gradient de γ, c'est le vecteur des dérivées des composantes de γ. Par abus il est traditionnel de le noter avec un $'$, en raison du fait que le domaine de départ est dans $\mathbb{R}$.

— en quoi est-ce que $\mathcal{T}_{V,p}$ est un espace vectoriel ?
— quelle est la dimension de $\mathcal{T}_{V,p}$ si celui-ci est un espace vectoriel ?
— comment se donne-t-on $\mathcal{T}_{V,p}$ dans la pratique ?

On va répondre aux deux première questions dans le cas de parties décrites comme graphes de fonctions, on argumentera par la suite de la généralisation au cas des lieux de $\mathbb{R}^n$ décrits implicitement. La dernière partie aura surtout un sens dans notre contexte dans le cadre des lieux décrits implicitement.

Question 6-19 Est-ce que l'hypothèse d'avoir des courbes γ dont le domaine est $]-1,1[$ est importante ? Pouvez-vous relâcher cette contrainte ?

6.1 Espace tangent à un graphe

On se donne une fonction $f : U \to \mathbb{R}$ définie sur un ouvert $U \subset \mathbb{R}^n$ et un point $a \in U$. On note dans la suite G le graphe de f. Tout point de G s'écrit sous la forme $(x, f(x))$ pour $x \in U$. On souhaite étudier l'espace tangent à G au point $p = (a, f(a))$. Pour cela on commence par s'intéresser à l'étude des courbes différentiables définies sur $]-1,1[$, à valeurs dans G et passant par p en 0. Une telle courbe γ d'image dans G s'écrit sous la forme [7], pour tout $t \in]-1,1[\to \mathbb{R}$,

$$\gamma(t) = \big(\alpha(t), f(\alpha(t))\big).$$

avec $\alpha(0) = a$.

En quoi est-ce $\mathcal{T}_{G,p}$ est un espace vectoriel ? Pour montrer que $\mathcal{T}_{G,p}$ est un espace vectoriel il suffit de montrer qu'on a un sous-espace vectoriel de $\mathbb{R}^{n+1}$. En effet, avec les notations précédentes on est en train d'étudier

$$\mathcal{T}_{G,p} = \{\gamma'(0) \mid \forall \gamma :]-1,1[\to G,\ \gamma(0) = p\}.$$

or $\gamma'(0) = \big(\alpha'(0), \nabla f(a)^T \alpha'(0)\big)$ est composé en première entrée d'un élément de $\mathbb{R}^n$ et dans la seconde d'un nombre réel, donc $\mathcal{T}_{G,p} \subset \mathbb{R}^{n+1}$.
Désormais pour montrer que $\mathcal{T}_{G,p}$ est un espace vectoriel il suffit de montrer que la somme de deux vecteurs dans celui-ci est encore la dérivée d'une courbe dans G et passant par a en 0, de même que pour le produit d'un tel vecteur par un scalaire.

— Soient $\gamma_1 = (\alpha_1, f \circ \alpha_1)$ et $\gamma_2 = (\alpha_2, f \circ \alpha_2)$ deux courbes différentiables de G passant par a en 0, on souhaite savoir si $\gamma_1'(0) + \gamma_2'(0)$ est la directione instantanée en 0 d'une courbe différentiable dans G passant par a en 0. La courbe

$$\gamma(t) = \frac{\gamma_1(2t) + \gamma_2(2t)}{2}$$

restreinte à $]-1/2, 1/2[$, répond à cette question.
— Soit $\lambda \in \mathbb{R}$, on souhaite savoir si $\lambda\gamma_1'(0)$ est la direction instantanée en 0 d'une courbe différentiable dans G passant par a en 0. La courbe

$$\gamma(t) = \gamma_1(\lambda t)$$

éventuellement après restriction, répond à cette question.

En conclusion l'espace tangent $\mathcal{T}_{G,p}$ est bien, en toute généralité, un sous-espace vectoriel de $\mathbb{R}^{n+1}$.

Quelle est la dimension de $\mathcal{T}_{G,p}$? C'est un sous-espace vectoriel de dimension n, on commence par justifier que c'est de dimension au moins n. Pour vérifier que $\dim \mathcal{T}_{G,p} \geq n$ il suffit de prendre les courbes qui approchent le point a par les différentes n directions canoniques disponibles. Pour tout $i \in \{1, \ldots, n\}$, on note α_i la courbe donnée pour tout $t \in]-1,1[$ par

$$\alpha_i(t) = (a_1, \ldots, t + a_i, \ldots, a_n).$$

7. Sinon, elle n'est pas dans G ...

Chaque α_i est une courbe différentiable passant par a en 0, elle approche a suivant un trajet porté par la i-ème coordonnée. La dérivée de α_i en 0 est donnée par le vecteur de la base canonique e_i. Si on note $\gamma_i = (\alpha_i, f \circ \alpha_i)$ on obtient en 0

$$\gamma_i'(0) = \big(e_i, \nabla f(a)^T e_i\big) = \big(e_i, \nabla f(a)_i\big) = \left(e_i, \frac{\partial f}{\partial x_i}(a)\right). \tag{2}$$

En raison de la non dépendance linéaire des e_i il est immédiat de voir que les $\gamma_i'(0)$ sont linéairement indépedants, $\mathcal{T}_{G,p}$ est donc au moins de dimension n.
Suite au travail précédent quant à la structure d'espace vectoriel de $\mathcal{T}_{G,p}$, toute les combinaison linéaire de vecteurs de la forme donnée équation 2 est encore un élément dans $\mathcal{T}_{G,p}$, pour tout $v \in \mathbb{R}^n$,

$$\big(v, \nabla f(a)^T v\big) \in \mathcal{T}_{G,p},$$

on peut encore exprimer cela par le fait que l'image de l'application linéaire de $\mathbb{R}^n \to \mathbb{R}^{n+1}$ donnée par $\phi : v \mapsto \langle v, \nabla f(a)^T v\rangle$ est incluse dans $\mathcal{T}_{G,p}$. Cette image est au plus de dimension n [8], de plus toute courbe $(\alpha, f \circ \alpha)$ donne une direction instantanée en p est de la forme

$$\big(\alpha'(0), \nabla f(a)^T \alpha'(0)\big)$$

donc dans l'image que ϕ. On vient donc de justifier l'égalité

$$\mathrm{Im}(\phi) = \big\{\big(v, \nabla f(a)^T v\big) \mid v \in \mathbb{R}^n\big\} = \mathcal{T}_{G,p}.$$

La dimension de $\mathcal{T}_{G,p}$ étant au moins n et celle de $\mathrm{Im}(\phi)$ au plus n on en déduit que $\mathcal{T}_{G,p}$ est de dimension n.

Comment cela s'articule dans notre exemple de dimension 1 ? On revient sur le cas du graphe d'une fonction $f : I \to \mathbb{R}$ définie sur un intervalle ouvert $I \subset \mathbb{R}$. D'après le travail qu'on vient d'effectuer l'espace tangent au graphe de f en un point $p = (a, f(a))$, qu'on note encore G, est un sous-espace vectoriel de dimension 1, correspondant aux directions instantanées des courbes différentiables passant par $(a, f(a))$ en 0. Un premier exemple d'une telle courbe est la courbe γ donnée par $\gamma : t \mapsto (t + a, f(t + a))$. La direction instantanée de γ en 0 est donnée par $(1, f'(a))$. Pour des raisons de dimensions, on obtient l'égalité

$$\mathrm{Vect}\big((1, f'(a))\big) = \mathcal{T}_{G,p}.$$

Donc

$$\mathcal{T}_{G,p} = \{\lambda(1, f'(a)) \mid \lambda \in \mathbb{R}\}. \tag{3}$$

Ce qui nous ramène à ce que vous connaissez déjà dans ce cas. À noter qu'on représente graphiquemenent $p + \mathcal{T}_{G,p}$.

Question 6-20 Décrire les espaces tangents des graphes des fonctions suivantes en tout point.

1. La fonction $f : \mathbb{R} \to \mathbb{R}$ donnée par $f(x) = x^2$.
2. La fonction $g : \mathbb{R} \to \mathbb{R}$ donnée par $g(x) = \frac{x}{1+x^2}$

Dessiner les espaces tangents aux graphes de ces fonctions respectivement au point $(2, 4)$ et $(1, 1/2)$.

Question 6-21 Décrire les espaces tangents des graphes des fonctions suivantes en tout point :

1. La fonction $f : \mathbb{R}^2 \to \mathbb{R}$ donnée par $f(x, y) = ax^2 + by^2$ pour $(a, b) \in \big(\mathbb{R}_+^*\big)^2$
2. La fonction $g : \mathbb{R}^2 \to \mathbb{R}$ donnée par $g(x, y) = \cos(xy)$.

8. Pourquoi ?

Question 6-22 On considère la fonction $f : \mathbb{R}^3 \to \mathbb{R}$ donnée par $f(x, y, z) = x^2 + 2y^2 - z$. Déterminer les points du graphe de f ayant un espace tangent parallèle à l'hyperplan

$$x + y - z - w = 0.$$

Peut-on avoir un hyperplan tangent au graphe de f d'équation $x = 0$? Pouvez vous trouver d'autres hyperplan satisfaisant cette condition ?

6.2 Espace tangent à une courbe de niveau

On se donne une fonction $f : U \to \mathbb{R}$ différentiable sur un ouvert $U \subset \mathbb{R}^n$. On note $\mathcal{C}$ la courbe de niveau 0 de f et $a \in \mathcal{C}$. Pour rappel on peut ramener toute courbe de niveau d'une fonction à une courbe de niveau 0 d'un translaté de celle-ci.
À la différence avec le cas d'un graphe, on n'a pas d'écriture explicite des courbes paramétrées passant par $a \in \mathcal{C}$, la seule contrainte qu'on peut en attendre correspond au fait d'annueler f. Plus précisément, on s'intéresse aux courbes différentiables $\gamma :]-1, 1[\to U$ telles que

— $\gamma(0) = a$
— $f(\gamma(t)) = 0$ pour tout $t \in]-1, 1[$.

En dérivant la seconde relation on obtient la relation

$$\nabla f(a)^T \gamma'(0) = 0. \tag{4}$$

Autrement dit, les directions instantanées en a des courbes de type γ sont orthogonales au gradient de f, $\nabla f(a)$. Ce qu'on peut encore exprimer par l'inclusion

$$\nabla f(a)^\perp \supset \mathcal{T}_{C,a}.$$

En particulier l'espace tangent $\mathcal{T}_{C,a}$ est inclus dans un sous-espace vectoriel de dimension $n - 1$ si $\nabla f(a)$ n'est pas le vecteur nul.
On arrive ici à la limite de ce qu'on peut faire sans préciser le lien entre écriture paramétrique (comme graphe d'une fonction) et implicite (comme courbe de niveau). Nous avons vu à ce stade que le graphe d'une fonction $g : U \subset \mathbb{R}^n \to \mathbb{R}$ peut être décrit comme le lieu des zéros de $h(x, y) = y - g(x)$. Nous avons également vu qu'une courbe de niveau n'était pas nécessairement le graphe d'une fonction, chose qu'on peut estimer regrettable, car on serait dans ce cas que $\mathcal{T}_{C,a}$ serait de dimension $n - 1$. On peut encore se permettre d'espérer une telle conséquence grâce au résultat (fondamental) suivant :

Fait 6.3. Si $f : U \to \mathbb{R}$ est une fonction différentiable de différentielle non nulle en un point $a \in \mathcal{C}_{f,0} \subset \mathbb{R}^n$ alors $\mathcal{C}_{f,0}$ est, localement en a, donnée par le graphe d'une fonction de $\mathbb{R}^{n-1} \to \mathbb{R}$.

Ce résultat, quand il est plus formalisé, est appelé ***théorème des fonctions implicites***. Désormais, armé de celui-ci et du travail effectué au cours de la section 6.1 on peut affirmer que [9]

$$\mathcal{T}_{\mathcal{C},a} = \nabla f(a)^\perp.$$

Ainsi, le gradient de f décrit une équation de l'espace tangent à $\mathcal{T}_{\mathcal{C},a}$ en a :

$$\mathcal{T}_{\mathcal{C},a} : \quad \nabla f(a)^T x = 0.$$

Il faut décaler cette équation en a si l'on souhaite représenter $a + \mathcal{T}_{\mathcal{C},a}$.

9. Il reste un détail qu'on n'a pas aborder pour être rigoureux, qui peut le trouver ?

Question 6-23 Représenter graphiquement les espaces tangents au courbes de niveau 1 des fonctions suivantes en tout point :

1. La fonction $f : \mathbb{R}^2 \to \mathbb{R}$ donnée par $f(x, y) = ax^2 + by^2$ pour $(a, b) \in \left(\mathbb{R}_+^*\right)^2$
2. La fonction $g : \mathbb{R}^2 \to \mathbb{R}$ donnée par $g(x, y) = \cos(xy)$.

6.3 En conclusion

Soient $f : U \to \mathbb{R}$ une fonction différentiable sur un ouvert $U \subset \mathbb{R}^n$ et $a \in U$. On note $\mathcal{C}$ la courbe de niveau de f passant par a et G le graphe f. Pour rappel $\mathcal{C} \subset \mathbb{R}^n$ et $G \subset \mathbb{R}^{n+1}$.
L'espace tangent à $\mathcal{C}$ au point a est un hyperplan vectoriel (de dimension $n-1$) ayant le $\nabla f(a)$ comme vecteur orthogonal. On peut donc écrire

$$\mathcal{T}_{\mathcal{C},a} : \quad \nabla f(a)^T x = 0.$$

Le vecteur gradient est donc orthogonal à la ligne de niveau de f passant par a, au sens où il est orthogonal à l'espace tangent à celle-ci au point a.
L'espace tangent au graphe de G en un point $p = (a, f(a))$ est un hyperplan vectoriel de $\mathbb{R}^{n+1}$ (donc de dimension n) qu'on peut décrire paramétriquement comme l'espace vectoriel engendré par les vecteurs

$$\left(e_i, \frac{\partial f}{\partial x_i}(a)\right) = (e_i, \nabla f(a)_i)$$

ou encore de manière équivalente comme l'image de l'application $v \mapsto (v, \nabla f(a)^T v)$. Pour se donner cet hyperplan de manière implicite, ce qui est de moindre coût dans le cas d'un hyperplan, il suffit de se rappeler que G peut être décrit par l'équation

$$G : \quad f(x) - y = 0$$

avec les variables $x \in \mathbb{R}^n$ et $y \in \mathbb{R}$. En notant $g(x, y) = f(x) - y$ on a

$$\mathcal{T}_{G,p} : \quad \begin{pmatrix} \nabla f(a) \\ -1 \end{pmatrix}^T \begin{pmatrix} x \\ y \end{pmatrix} = 0 \iff \nabla f(a)^T x - y = 0$$

En particulier le vecteur colonne $(\nabla f(a), -1)$ est orthogonal à $\mathcal{T}_{G,p}$ et en donne donc une équation.

Question 6-24 Trouver les points sur le paraboloïde $z = 4x^2 + y^2$ où le plan tangent est parallèle au plan $x + 2y + z = 6$. Faire de même avec le plan $3x + 5y - 2z = 5$.

Question 6-25 Un étudiant malheureux trouve pour plan tangent à la surface donnée par $z = x^4 - y^2$ au point $(2, 3, 7)$ la réponse

$$z = 4x^3(x - 2) - 2y(y - 3) + 7.$$

1. Sans calcul, pourquoi est-ce faux ?
2. Donner la réponse correcte.
3. D'où venait la confusion de l'étudiant ?

Question 6-26 Dans le cas de la fonction $f : \mathbb{R}^2 \to \mathbb{R}$ donnée par $f(x, y) = ax^2 + by^2$ pour $(a, b) \in \left(\mathbb{R}_+^*\right)^2$ où se trouve le point réalisant le minimum de f ? Dans quelle direction pointe le gradient en un point quelconque des courbes de niveaux de f ? Quelle rôle joue le plan tangent ?

7 Gradient et minimisation

On sait désormais que le gradient d'une fonction $f : U \subset \mathbb{R}^n \to \mathbb{R}$ en un point $a \in \mathcal{C}_{f,r}$ où f est différentiable donne un vecteur orthogonal à l'espace tangent en $\mathcal{C}_{f,r}$ au point a. En particulier, cet espace tangent est décrit par l'équation

$$\nabla f(a)^T x = 0.$$

L'objectif de cette section sera d'approfondir encore notre étude du lien entre gradient et point minimaux, notamment d'une fonction convexe. On va étudier en un premier temps la notion de point critique d'une fonction différentiable, puis celle du comportement d'une fonction différentiable vis-à-vis de la minimisation quand on est en un point non critique pour enfin arriver au résultat central suivant : ***le gradient d'une fonction convexe en un point x, définit un hyperplan d'appui au sous-courbe de niveau de f passant par x***. Il faut entendre ce résultat comme un résultat de dichotomie, l'espace tangent permet de couper le domaine de f en deux, on ne peut minimiser f que de l'un de ces deux côté, celui opposé au gradient.

7.1 Points critiques

On rappelle qu'une fonction $f : U \subset \mathbb{R}^n \to \mathbb{R}$ définie sur un ouvert de U admet un ***extremum local*** en un point $a \in U$ s'il existe un voisinage V de a sur lequel f sa plus grande, ou plus petite valeur, en a. Dans la suite, et sauf mention explicite du contraire, on se limitera à l'étude des cas de ***minimum locaux***. On dit donc que f admet un ***minimum local*** en un point $a \in U$ s'il existe un voisinage V de a tel que, pour tout $y \in V$, $f(y) \geq f(a)$.

Proposition 7.1. *Soit $f : U \subset \mathbb{R}^n \to \mathbb{R}$ une fonction différentiable sur U. On suppose que f atteint un extremum local en a. Alors $\nabla f(a) = 0$* [10].

Démonstration. On va prouver ce résultat dans le cas d'un minimum local, le cas d'un maximum se traite de manière similaire.
On suppose que a est un point où f atteint un minimum local. Il existe donc un voisinage V de a sur lequel la plus petite valeur de f est $f(a)$. Comme f est différentiable, quitte à rétrécir V et avec les notations habituelles, pour tout $h \in V$

$$f(a+h) = f(a) + \nabla f(a)^T h + \epsilon(h)\|h\|.$$

Donc, pour tout $h \in V$

$$\nabla f(a)^T h = f(a+h) - f(a) - \epsilon(h)\|h\|.$$

Comme $\epsilon(h)\|h\| \xrightarrow[h\to 0]{} 0$ il existe un voisinage de 0 tel que $\nabla f(a)^T h$ est du signe de $f(a+h) - f(a)$, donc (par hypothèse) de signe positif [11]. Or une application linéaire ne peut être positive sur tout un voisinage de 0, sauf si elle est nulle. En effet, si $\nabla f(a)$ définie une application linéaire non nulle sur une boule ouverte de la forme $B(0,\eta)$, il existe $v \in B(0,\eta)$ tel que $\nabla f(a)^T v \neq 0$. Mais dans ce cas $\nabla f(a)^T(-v) \neq 0$ et est de signe opposé à $\nabla f(a)^T v$, en particulier $\nabla f(a)$ ne peut pas être de signe constant sur $B(0,\eta)$. Par contraposition $\nabla f(a)$ ne peut être que nul. □

Cette proposition est **aucunement** une équivalence.

Question 7-27 En s'inspirant d'exemples de l'analyse réelle en dimension 1 montrez donner des exemples de fonctions ayant un point a où la dérivée s'annule sans pour autant que a ne soit un point extrémal local.

10. Attention au fait que $0 \in \mathbb{R}^n$
11. Il y a un petit raccourci ici, n'hésitez pas à poser la question.

Quand on étudie un problème d'optimisation on cherche à identifier des potentiels points où la fonction objectif f atteint un ***minimum global***, c'est-à-dire des points $a \in U$ où, pour tout $yinU$, $f(y) \geq f(a)$. Ce type de point n'existe pas toujours, par exemple si f n'est pas minorée sur U, mais quand celui-ci existe, c'est nécessairement un minimum local au sens ci-dessus car on suppose ici travailler sur un ouvert.
On appelle ***point critique*** d'une fonction $f : U \subset \mathbb{R}^n \to \mathbb{R}$ différentiable sur un ouvert U tout point $a \in U$ tel que $\nabla f(a) = 0$. Le ***lieu critique*** de f est l'ensemble des points critiques de f. On peut reformuler la discussion à ce stade par la proposition suivante.

Proposition 7.2. *Soit $f : U \subset \mathbb{R}^n \to \mathbb{R}$ une fonction définie et différentiable sur un ouvert U. Si f admet un extremum global en un point a, alors a est un point critique de f.*

Question 7-28

1. Déterminer les lieux critiques des fonctions :
   (a) $f : (x, y) \mapsto x^2 + 2y^2$
   (b) $g : (x, y) \mapsto \cos(xy)$
   (c) $u : (x, y) \mapsto y^3x^3$
2. Décomposer le lieu critique suivant le types de points rencontrés :

Question 7-29 Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice symétrique. Déterminer, éventuellement en fonction de A, le lieu critique des fonctions donnée par

1. $X \mapsto AX + b$
2. $X \mapsto X^TAX$.

Question 7-30 Un problème d'optimisation quadratique sans contrainte est un problème d'optimisation qui prend la forme

$$\text{minimser } X^TPX + QX + r$$

avec P une matrice symétrique de taille (n, n), Q une matrice ligne de taille $(1, n)$ et r une constante.

1. Décrire le lieu critique de $X^TPX + QX + r$.
2. Quel est le lien avec la régression linéaire ?

7.2 Direction de descente

Suite à la section précédente, on pourrait imaginer la démarche suivante pour étudier tout problème d'optimisation sans contraintes donné par une fonction objectif $f : U \subset \mathbb{R}^n \to \mathbb{R}$ différentiable :

— calculer le gradient de f
— déterminer le lieu critique de f
— identifier la décomposition du lieu critique suivant le type d'extrema éventuels.

Cette démarche fait tout à fait sens dans certains cas, elle est cependant inopérante dans la majeure partie des cas ; la détermination du lieu critique d'une fonction, quand elle est possible, est calculatoirement trop coûteuse. Pour cette raison, une autre stratégie est envisageable, on y choisit de perdre de la précision et gagner en capacité de calcul. Cette stratégie suit l'idée suivante : on commence par un point initial $x_0 \in U$, à partir de ce point on souhaite trouver un nouvel itéré x_1 pour lequel la valeur objectif $f(x_0) \geq f(x_1)$. On reproduit ce schéma itérativement pour construire une suite (x_i) de points dans U tels que $f(x_i) \geq f(x_{i+1})$. L'algorithme s'arrête dès qu'on trouve un point *convenable* pour point optimal, cette question délicate sera étudiée plus en détail en seconde partie de cours.

La question qu'on traite ici cependant, et qui est au coeur de l'ensemble des algorithmes itératifs, est de savoir dans quelle direction, à partir de x_i, rechercher l'itéré x_{i+1} ?

Répondre à cette question, signifie d'analyser plus en détail la relation entre gradient, espace tangent et sous-niveaux de la fonction objectif étudiée. Celle-ci est résumée dans la proposition suivante.

Proposition 7.3. *Soit $f : U \subset \mathbb{R}^n \to \mathbb{R}$ une fonction différentiable sur un ouvert U. Soit x un point de f sur la courbe de niveau r de f. Si x n'est pas un point critique de f, il existe un point $x^+ = x + \Delta x$ tel que $f(x) \geq f(x^+)$, obtenu par un déplacement Δx à l'opposé de la direction de $\nabla f(x)$.*

Démonstration. Par hypothèse, il existe un voisinage W de 0 tel que pour tout $h \in W$,

$$f(x+h) = f(x) + \nabla f(x)^T h + \epsilon(h)\|h\|$$

avec les notations habituelles. Comme x appartient à $\mathcal{C}_{f,\leq r}$, on peut ré-écrire cette équation sous la forme

$$f(x+h) = r + \nabla f(x)^T h + \epsilon(h)\|h\|.$$

Comme $\nabla f(x) \neq 0$, la direction $\delta x = -\nabla f(x)$ satisfait

$$\nabla f(x)^T \delta x = -\|\nabla f(x)\|^2 < 0.$$

Ainsi pour $t \in \mathbb{R}_+^*$ assez proche de 0, $f(x + t\delta x) - r$ est du signe de $\nabla f(x)^T(t\delta x)$ car, par définition, $\epsilon(t\delta x)\|t\delta x\| \xrightarrow[t\to 0]{} 0$. Il existe donc un $t \in \mathbb{R}_+^*$ pour lequel $\Delta x = t\delta x$ vérifie $f(x + \Delta x) < f(x)$ et $\Delta x = -t\nabla f(x)$. . □

La preuve de la proposition précédente nous indique qu'en un point $x \in U$ tel que $\nabla f(x) \neq 0$, il existe $t_x > 0$ tel que

$$f(x - t_x \nabla f(x)) < f(x).$$

La relation

$$x_{i+1} = x_i - t_i \nabla f(x)$$

nous permettrait donc de créer une suite (x_i) d'itérés pour lesquels $f(x_{i+1}) < f(x_i)$. Cela aurait du sens tant qu'on ne tombe pas sur un point critique x^*, qui ne pourrait d'en ce cas qu'être, ou bien un point selle ou alors un minimum local [12]. Cette belle image ne correspond cependant pas à la difficulté qu'on croise en réalité :

- on a aucune information sur les t_i, ce qu'on appelle le ***learning rate***
- on a aucune garantie sur le fait que x^* soit un minimum global, notre condition d'arrêt ne nous donne donc aucune garantie quand à la validité du résultat de sortie.

Ces deux points sont centraux dans les problématiques d'implémentation, et d'interprétation des résultats des descentes de gradient. Sujet de la seconde partie du cours.

Question 7-31 Au regard des éléments apportés dans cette section analysées les directions du gradient en différents points du domaine des fonctions suivantes :

1. La fonction donnée par $x \mapsto x\sin(x)$.
2. La fonction donnée par $(x, y) \mapsto \cos(xy)$.

7.3 Apport de la convexité

On garde la notation précédente d'une fonction $f : U \subset \mathbb{R}^n \to \mathbb{R}$ différentiable sur un ouvert $U \subset \mathbb{R}^n$. Nous sommes désormais dans la situation suivante : à partir d'un point initial $x_0 \in U$ où le gradient ne s'annule pas, on peut imaginer construire une suite (x_i) (finie) de points dans U satisfaisants $f(x_{i+1} \leq f(x_i)$, tant que les indices font sens. Cette construction peut se faire via des choix *ad hoc* de *learning rate* pour chacune des itérations. Le dernier itéré x^* obtenu est :

12. Pourquoi ?

— le dernier point obtenu par fatigue
— un point critique de f.

Dans les deux cas on est dans une situation où x^* n'est pas garanti d'être un minimum local et encore moins un minimum global. Malgré ces défauts, cette démarche est celle conservé dans tous l'essentiel des algorithmes d'apprentissage sans contraintes, notamment pour entraîner des réseaux de neurones. Elle s'accompagne cependent d'heuristique qui permettent d'espérer trouver de meilleurs points d'arrêts.

Cette démarche reste cependant particulièrement risquée, on est théoriquement dans une situation où trouver un minimum local itérativement n'est pas garanti. Il existe cependant un contexte dans lequel la démarche précédente garantie, aux questions d'implémentations prêt, le fait d'atteindre un (le) minimum global quand celui-ci existe : la **convexité** de f. Cette hypothèse n'est pas hors propos car une partie des modèles classiques de ML sont convexe, par exemple les problèmes de régressions et les SVMs. Cette hypothèse de convexité de f vient apporter deux propriétés :

— un point critique de f réalise un point optimal pour f
— le gradient de f en un point x d'une courbe de niveau $C_{f,r}$ de f définit un hyperplan d'appui au sous-niveau de $C_{f,r}$, en particulier on peut ignorer le demi-espace porté par le gradient de f pour la recherche de l'itéré x^+.

Les deux points précédents découlent d'un même résultat, portant sur la caractérisation de premier ordre des fonctions convexes.

Proposition 7.4. *Une fonction $f : U \subset \mathbb{R}^n \to \mathbb{R}$ une fonction définie sur un ouvert convexe $U \subset \mathbb{R}^n$ est convexe si, et seulement si, pour tout x, $y \in U$,*

$$f(y) - f(x) \geq \nabla f(x)^T (y - x). \tag{5}$$

Démonstration. Supposons f convexe et montrons la relation 5. Dire que f est convexe signifie que pour tout $t \in [0, 1]$ et tout x, $y \in U$

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

On va chercher à réécrire cette inégalité de façon à faire apparaître le membre de gauche de l'inégalité 5 à droite dans l'expression ci-dessus. Un peu de travail permet de trouver :

$$\frac{f\big((1-t)x + ty\big) - f(x)}{t} \leq f(y) - f(x).$$

En réécrivant le membre de gauche de façon à baser l'argument de f en x on a

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

On reconnaît à gauche de cette inégalité la dérivée directionnelle de f en x le long du vecteur $(y - x)$. En prenant la limite quand $t \to 0$ on obtient

$$\nabla f(x)^T (y - x) \leq f(y) - f(x)$$

qui est l'inégalité recherchée.

Supposons, inversement que l'équation (5) est satisfaite. On va montrer que cela implique la convexité de f. Soient x, y deux points de U, on note $z_t = (1-t)x + ty$ un point sur le segment reliant x à y. En appliquant l'inégalité 5 aux deux entrees (x, z_t) et (y, z_t) on obtient les deux inégalités :

$$f(x) - f(z_t) \geq \nabla f(z_t)^T (x - z_t)$$
$$f(y) - f(z_t) \geq \nabla f(z_t)^T (y - z_t)$$

En multipliant la première ligne par $(1-t)$ et la seconde par t puis en sommant, on obtient

$$(1-t)f(x) + tf(y) - f(z_t) \geq \nabla f(z_t)^T \big(\underbrace{(1-t)x + ty - z_t)}_{=0} \big) = 0$$

D'où

$$f\big((1-t)x+ty\big) = f(z_t) \leq (1-t)f(x) + tf(y).$$

□

Remarque 6. La caractérisation (5) a une interprétation géométrique simple : l'hyperplan tangent au graphe de f en un point $(x, f(x))$ est un hyperplan d'appui. Cette caractérisation de la convexité qu'on a déjà vue, se traduit en un point $x \in U$ par

$$\begin{pmatrix} \nabla f(x) \\ -1 \end{pmatrix}^T \begin{pmatrix} y-x \\ f(y)-f(x) \end{pmatrix} \leq 0$$

qui donne, une fois la multiplication matricielle explicitée, la relation (5). On rappelle que le vecteur $(\nabla f(x), -1)$ est un vecteur normal définissant l'hyperplan tangent au graphe de f au point $(x, f(x))$. Notez au passage le fait que $\nabla f(x)$, x et y sont des vecteurs dans $\mathbb{R}^n$; l'écriture ci-dessus est donc une écriture par blocs.

On décline dans la suite les corollaires de ce résultats pour affiner notre analyse du cas convexe, quand on s'intéresse aux questions d'optimisation.

Points critiques et convexité La proposition 7.4 n'est pas nécessaire pour montrer qu'une fonction convexe non localement constante admet un unique point minimal local. De même celle-ci n'est pas nécessaire pour montrer qu'une fonction convexe ne peut avoir de maximum locaux qui ne soient pas des minimums locaux (et dans ce cas on est localement constant). Celle-ci permet cependant de traiter de tous les cas de points critiques qu'on peut rencontrer, notamment ceux où la différentielle seconde s'annule. Des preuves de ces faits dans le cas général, non nécessairement différentiable sont à votre disposition en annexe.
Le résultat suivant est un des deux résultats fondamentaux de ce cours. Il est à l'origine même du qualificatif *convexe* composant le titre de cours.

Corollaire 7.5 (F1)**.** *Soit $f : U \subset \mathbb{R}^n \to \mathbb{R}$ une fonction convexe, alors tout point critique de f est un point optimal pour f*

Démonstration. Si x est un point critique pour f, c'est-à-dire que le $\nabla f(x) = 0$, d'après l'inégalité de convexité, pour tout $y \in U$,

$$f(y) - f(x) \geq 0$$

Donc pour tout $y \in U$, $f(y) \geq f(x)$. □

Géométrie du gradient d'une fonction convexe La proposition 7.4 peut également s'interpréter, avec les notations qui précèdente, de la manière suivante : si $x \in \mathcal{C}_{f,\leq r}$ alors le gradient de f en x définit un hyperplan d'appui aux sous-niveaux $\mathcal{C}_{f,\leq r}$. Autrement dit, si x est de niveau r, tout point de valeur objective plus petite que r se trouve dans le demi-espace définit par $\nabla f(x)$, à l'opposé de celui-ci ; le demi-espace négatif.

Corollaire 7.6 (F2)**.** *Si $f : U \subset \mathbb{R}^n \to \mathbb{R}$ est une fonction convexe et différentiable sur un ouvert U, alors le gradient de f en un point x définit un hyperplan d'appui aux sous-niveau de f de niveau $f(x)$.*

Démonstration. On se place en un point x de niveau r. D'après l'inégalité de convexité, pour tout $y \in \mathcal{C}_{f,r}$

$$f(y) - f(x) \geq \nabla f(x)^T (y - x).$$

Comme par définition $f(y) \leq r$ et $f(x) = r$ on en déduit, pour tout $y \in \mathcal{C}_{f,\geq r}$

$$\nabla f(x)^T (y - x) \leq 0.$$

Ce qui est la définition même d'avoir $\nabla f(x)$ définir un hyperplan d'appui à $\mathcal{C}_{f,\leq r}$. □

On en vient donc à l'image géométrique que l'on souhaite garder en mémoire : pour une fonction convexe f le gradient de f en un point x définit un hyperplan d'appui aux sous-courbe de niveaux de f de niveau $f(x)$. Autrement dit, l'espace tangent à la courbe de niveau $f(x)$ passant par x sépare l'espace en deux, uniquement l'un des deux demi-espaces qu'il définit permet de minimiser f, celui à l'opposé de la direction protée par le gradient en x. Accompagné du résultat 7.6, cela garantit le fait que d'aller dans la direction opposé au gradient nous amène nécessairement vers le minimum ***global*** de f dès que f est minorée. Cette garatnie nous manquait dans le cas non nécessairement convexe.

Question 7-32 Revenir sur les exemples des fonctions
- $(x, y) \mapsto ax^2 + by^2$ où $(a, b) \in \mathbb{R}^{\mathbf{2}}$
- $(x, y) \mapsto \cos(xy)$

pour constater la réalisation ou non des affirmation précédentes.

Question 7-33 Sous quelles conditions est-ce qu'une fonction de la forme

$$X^T PX + QX + r$$

où P et Q sont des matrices carrées de tailles (n, n) et $r \in \mathbb{R}$, est convexe ?

8 Hessienne d'une fonction

L'étude menée jusqu'à présent est une étude dite *de premier ordre* ; on s'intéresse à l'approximation locale d'une fonction au voisinage d'un point par une fonction affine. Malgré sa simplicité, cela est amplement suffisant pour l'essentiel des cas pratiques en ML, les descentes de gradients se basent essentiellement sur cette approximation pour choisir une direction de mise à jour.
Dans cette section on va se permettre d'aller un peu plus loin quant à l'approximation faite des fonctions qu'on étudie ; l'approximation *de second ordre*. Celle-ci est à l'origine de ***l'algorithme de Newton*** qui permet une réduction du nombre d'itérations pour converger vers un point optimal d'une fonction [13]. Cette approximation de second ordre, va nous inviter à généraliser la notion de dérivée seconde, elle nous donnera en contrepartie les mêmes outils que dans le cas unidimensionnel pour analyser les types de points critiques qu'on rencontre. Bien entendu, cela nécessite encore de l'algèbre linéaire.

Définition 8.1. Soit $f : U \subset \mathbb{R}^n \to \mathbb{R}$ une fonction différentiable en tout point d'un ouvert U, de gradient continue. On dit que f est 2-fois différentiable en un point a du domaine de f si la fonction gradient $x \mapsto \nabla f(x)$ est différentiable en a. Dans ce cas on appelle ***hessienne*** de f en a la jacobienne de ∇f, elle est notée $H_f(a)$.

Le gradient d'une fonction f est composée des dérivées partielles de f

$$\nabla f(x) = \left(\frac{\partial f}{\partial x_1}(x), \ldots, \frac{\partial f}{\partial x_n} \right).$$

La jacobienne de ∇f est donnée par les dérivées partielles par rapport aux mêmes variables des coordonnées de ∇f, donc des dérivées partielles de f, explicitement :

$$H_f(a) = \begin{pmatrix} \frac{\partial}{\partial x_1}\left(\frac{\partial f}{\partial x_1}\right)(a) & \cdots & \frac{\partial}{\partial x_n}\left(\frac{\partial f}{\partial x_1}\right)(a) \\ \vdots & \ddots & \vdots \\ \frac{\partial}{\partial x_1}\left(\frac{\partial f}{\partial x_n}\right)(a) & \cdots & \frac{\partial}{\partial x_n}\left(\frac{\partial f}{\partial x_n}\right)(a) \end{pmatrix}$$

13. Cela se fait au prix d'un coût plus élevé de chaque itération.

Cette écriture met en évidence le fait que la hessienne d'une fonction $f : U \subset \mathbb{R}^n \to \mathbb{R}$ est une matrice carrée de taille n. Pour simplifier les notations on pose

$$\frac{\partial^2 f}{\partial x_i^2} = \frac{\partial}{\partial x_i}\left(\frac{\partial f}{\partial x_i}\right)$$
$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial}{\partial x_i}\left(\frac{\partial f}{\partial x_j}\right)$$

Écriture qui ramène l'écriture de la hessienne de f à l'expression :

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

Dans le cas où les dérivées partielles secondes sont continues, le théorème de Schwarz permet d'affirmer que $H_f(a)$ est symétrique. Autrement dit, indépendamment de l'ordre avec lequele on effectue la dérivation partielle on retrouve le même résultat. Avec les notations précédentes :

$$\forall i, j \in \{1, \ldots, n\} \qquad \frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}.$$

Hypothèse 8.1. Sauf mention explicite du contraire, on suppose par la suite que la hessienne est symétrique.

Dans le cas de premier ordre, la matrice jacobienne de f en un point représente la matrice d'une application linéaire. La matrice hessienne d'une fonction f en un point ne représente plus une application linéaire mais une application bilinéaire. On ne rentrera pas dans les détails de ces aspects qui nous écarterais du coeur de ce cours. Il nous suffit, pour nos besoin, d'être en mesure de développer une fonction au second ordre, ces éléments sont résumés dans la proposition suivante. Vous êtes renvoyés aux annexes pour une rapide explication de ce phénomène.

Proposition 8.2. *Soit $f : U \subset \mathbb{R}^n \to \mathbb{R}$ une application 2-fois différentiable en un point a de l'ouvert U. Alors pour h dans un voisinage de 0*

$$f(a+h) = f(a) + \nabla f(a)h + \frac{1}{2}h^T H_f(a)h + \epsilon(h)\|h\|^2. \tag{6}$$

avec $\epsilon(h) \xrightarrow[h\to 0]{} 0$.

Comme vous pouvez le constater dans l'expression la hessienne de f est associé au terme donnée par

$$\frac{1}{2}h^T H_f(a)h$$

c'est désormais une expression bilinéaire en h et non linéaire.

Question 8-34 Soient $f : \mathbb{R} \to \mathbb{R}$ une fonction 2-fois dérivable et P une matrice symétrique. Quelle est la différentielle seconde, en tout point, de l'application f_P définie sur $\mathbb{R}^n$ par $x \mapsto f(x^T P x)$?

Question 8-35 Calculer les hessiennes en tout point des fonctions à valeurs réelles suivantes :

1. $f(x, y) = x^2 + xy + y^2 + \dfrac{x^3}{4}$;
2. $f(x, y) = x^3 + y^3$;
3. $f(x, y) = x^2 y - \dfrac{x^2}{2} - y^2$;

4. $f(x,y,z) = x^2 + 3y^2 - z^2 + 2xy - 5yz$;
5. $f(x,y,z) = \ln\left(e^x + e^y + e^z\right)$.

Tout comme c'est le cas en dimension 1 pour la dérivée seconde, la hessienne d'une fonctoin permet de déterminer si une fonction de $\mathbb{R}^n$ dans $\mathbb{R}$ est convexe, plus précisément

Proposition 8.3. *Soit $f : U \to \mathbb{R}$ une fonction 2 fois différentiable définie sur un ouvert convexe $U \subset \mathbb{R}^n$. Alors f est convexe si et seulement si*

$$\forall x \in U, \quad H_f(x) \geq 0. \tag{7}$$

Remarque 7. La positivité de $H_f(x)$ est à prendre ici au sens des application bilinéaire, il faut donc que pour tout $h \in \mathbb{R}^n$, $h^T H_f(x) h \geq 0$. De manière équivalente ($H_f(x)$ est par hypothèse symétrique) doit avoir des valeurs propres positives.

On retrouve à l'aide de cette caractérisation, dans le cas 2-fois différentiable, le fait qu'un point critique d'une fonction convexe ne peut être qu'un minimum local.

Question 8-36 Les fonction de la question (8-35) sont-elles convexes ? Justifier.

Le critère de convexité locale (et de second ordre) donné ici, se généralise pour permettre l'étude des types de points critiques qu'on peut obtenir, toujours sous l'hypothèse d'être 2-fois différentiable. On va revenir pour en parler sur la notion de positivité d'une application bilinéaire.
Pour rappel, toute application bilinéaire $\phi : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ est donnée par une matrice A dans laquelle pour tout h_1, $h_2 \in \mathbb{R}^n$

$$\phi(h_1, h_2) = h_1^T A h_2.$$

Dans notre cas on aura uniquement affaire à des applications bilinéaires symétriques, donc la matrice A le sera également. Pour plus de détails sur la question des forme bilinéaires on vous renvoie 1à l'annexe C.
On dit qu'une application bilinéaire $\phi : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ est

- positive si $\forall h \in \mathbb{R}^n$, $\phi(h,h) \geq 0$
- négative si $\forall h \in \mathbb{R}^n$, $\phi(h,h) \leq 0$
- définie positive si $\forall h \in \mathbb{R}^n$, $h \neq 0$, $\phi(h,h) > 0$
- définie négative si $\forall h \in \mathbb{R}^n$, $h \neq 0$, $\phi(h,h) < 0$.

Sous l'hypothèse de ϕ symétrique, la matrice associée A est une matrice symétrique réelle, elle est donc diagonalisable. Les définitions précédentes s'expriment sous la forme suivante : ϕ est

- (définie) positive si les valeurs propres de A sont (strictement) positives
- (définie) négative si les valeurs propres de A sont (strictement) négatives.

Proposition 8.4. *Soit $f : U \subset \mathbb{R}^n \to \mathbb{R}$ une fonction 2-fois différentiable en un point a. On suppose que a est un point critique de f, si $H_f(a) \neq 0$ est*

- *positive alors a est un minimum local*
- *négative alors a est un maximum local*
- *ni positive ni négative alors a est un point selle.*

Remarque 8. Dans le cas où $H_f(a) = 0$ on est incapable de dire ce qui peut se passer sans une étude plus fine. Prenez pour exemples les fonctions $x \mapsto x^3$ et $x \mapsto x^4$ en 0.

Question 8-37 Utiliser *Geogebra* pour tracer les graphes des fonctions :

1. $(x,y) \mapsto \pm(x^2 + y^2)$
2. $(x,y) \mapsto x^2 - y^2$

Question 8-38 Trouver un critère simple, à l'aide du déterminent et la trace, pour déterminer si un point critique est un maximum local, minimum local ou point selle. Est-ce que ce critère marche en dimension trois ?

A Questions de convexités

On choisit de regrouper dans cette section quelques preuves qui affinent l'analyse générale de l'apport de la convexité à la situation d'optimisation. Même si l'on ne suppose pas la différentiabilité d'une fonction convexe il n'en reste qu'elle ne peut pas avoir plusieurs minimums locaux, elle ne peut pas non plus avoir des maximums locaux qui ne soient pas des minimum locaux (auquel cas on parle d'une fonction localement constante au voisinage d'un tel maximum).

Lemme A.1. *Soit $f : U \subset \mathbb{R}^n \to \mathbb{R}$ une fonction convexe sur un ouvert $U \subset \mathbb{R}^n$. Alors f ne peut avoir un maximum local qui ne soit pas un minimum local.*

Démonstration. Dire que a est un maximum local qui n'est pas un minimum local signifie qu'il existe un voisinage V de a sur lequel f prend sa plus grande valeur en a et de façon à ce que tout voisinage de a contient un point ayant une image par a strictement plus petite que $f(a)$. En particulier, pour tout $x \in V$, $f(x) \leq f(a)$ et il existe $b \in V$ tel que $f(b) < f(a)$. On note B la boule fermée centrée en a et de rayon $\|b-a\|$, le point b est un point du bord de B. On note c le point opposé à b par rapport à a. Avec cette écriture on a

$$a = \frac{b+c}{2} = \left(1 - \frac{1}{2}\right) b + \frac{1}{2} c.$$

et a est donc de la forme $(1-t)b + tc$ pour $t = 1/2$ [14]. Comme $f(b) < f(a)$ en notant $t = 1/2$ on a l'inégalité

$$(1-t)f(b) + tf(c) < (1-t)f(a) + tf(a) = f(a) = f\big((1-t)b + tc\big).$$

En gardant les deux bouts de l'expression on vient de contredire le fait que f est convexe. Ainsi, f ne peut avoir de maximum local en a qui n'est pas également un minimum local ; f est dans ce cas localement constant au voisinage de a. □

Lemme A.2. *Soit $f : U \subset \mathbb{R}^n \to \mathbb{R}$ une fonction convexe. Soient x_1, $x_2 \in U$ deux points où f admet des minimums locaux, alors f est constante sur le segment de droite reliant x_1 à x_2.*

Démonstration. TBA □

Corollaire A.3. *Une fonction convexe admet au plus un minimum local ; s'il existe c'est un minimum global.*

Démonstration. TBA □

B Différentielle seconde et hessienne

Le cours laisse de côté les raisons pour lesquelles la hessienne d'une fonction objectif représente la matrice d'une application bilinéaire, et non linéaire. On revient sur ce point dans cette section, il est à votre disposition pour les éventuels curieux.

Définition B.1. On dit qu'une fonction $f : U \to \mathbb{R}^m$ sur un ouvert $U \subset \mathbb{R}^n$ est 2 fois différentiable en $a \in U$ si

- f est différentiable au voisinage de a ;
- la fonction $x \mapsto \mathrm{D}f(x)$,définie sur un voisinage ouvert V de a, est une fonction différentiable de V dans $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ [15].

14. Dans le jargon, on dit que a est combinaison convexe de b et c.
15. Qu'on identifie à $\mathbb{R}^{n \times m}$.

Quand elle existe, la différentielle seconde de f en a, c'est-à-dire la différentielle de $\mathrm{D}f$ en a, est notée $\mathrm{D}^2 f(a)$.

Il est important de comprendre quel type d'objet est la différentielle seconde. Avec les notations précédentes et en supposant f 2 fois différentiable en a, la différentielle seconde de f en a est un élément de $\mathcal{L}\big(\mathbb{R}^n, \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)\big)$. Essayons de comprendre les objets de cet ensemble. Soit $\phi \in \mathcal{L}\big(\mathbb{R}^n, \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)\big)$. On est donc face à une application

$$\begin{array}{rccc} \phi : & \mathbb{R}^n & \longrightarrow & \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m) \\ & h_1 & \longmapsto & \phi(h_1) : h_2 \mapsto \phi(h_1)(h_2) \end{array}$$

Par définition l'application ϕ est linéaire « à la fois en h_1 et en h_2 ». Faison l'identification consistant à écrire, par abus, $\phi(h_1, h_2)$ au lieu de $\phi(h_1)(h_2)$. Sous cette identification, ϕ devient donc une application bilinéaire de $\mathbb{R}^n \times \mathbb{R}^n$ dans $\mathbb{R}^m$. Donc $\mathrm{D}^2 f(a)$ est une application bilinéaire de $\mathbb{R}^n \times \mathbb{R}^n$ dans $\mathbb{R}^m$. En réalité on a plus, mais la preuve de ce résultat dépasse nos objectifs pour ce cours.

Théorème B.1. *Soit $f : U \to \mathbb{R}^m$ une fonction 2 fois différentiable en un point a de l'ouvert $U \subset \mathbb{R}^n$. Alors $\mathrm{D}^2 f(a)$ est une application bilinéaire symétrique sur $\mathbb{R}^n$.*

Dans la pratique on s'intéresse essentiellement aux valeurs de la différentielle seconde le long de la diagonale, c'est-à-dire pour les couples de la forme (h, h) avec $h \in \mathbb{R}^n$. La raison en est la suivante :

Proposition B.2. *Soit $f : U \to \mathbb{R}^m$ une application 2 fois différentiable en un point a de l'ouvert U. On a dans ce cas*

$$f(a+h) = f(a) + \mathrm{D}f(a)h + \frac{1}{2}\mathrm{D}^2 f(a)(h,h) + \epsilon(h)\|h\|^2 \tag{8}$$

pour h dans un voisinage de 0 et $\epsilon(h) \xrightarrow[t\to 0]{} h$.

L'expression (8) est le développement limité à l'ordre 2 de f au voisinage de a.
Quand une fonction $f : U \to \mathbb{R}$ pour $U \subset \mathbb{R}^n$ est donnée par une expression explicite, on calcule souvent de manière explicite une écriture « en coordonnées » de la différentielle seconde : la hessienne. Pour en inférer la définition on procède comme dans le cas de la jacobienne.
Soit $f : U \to \mathbb{R}$ une fonction 2 fois différentiable en un point a de l'ouvert $U \subset \mathbb{R}^n$. On note $h = \sum_{i=1}^n h_i e_i$ un vecteur dans $\mathbb{R}^n$. On a

$$\mathrm{D}^2 f(a)(h,h) = \sum_{i,j=1}^n h_i h_j \, \mathrm{D}^2 f(a)(e_i, e_j)$$

par définition $\mathrm{D}^2 f(a)(e_i, e_j) = \partial_{e_j}\big(t \mapsto \mathrm{D}f(t)(e_i)\big)(a)$, d'où

$$\begin{aligned} \sum_{i,j=1}^n h_i h_j \, \mathrm{D}^2 f(a)(e_i, e_j) &= \sum_{i,j=1}^n h_i h_j \partial_{e_j}\big(t \mapsto \mathrm{D}f(t)(e_i)\big)(a) \\ &= \sum_{i,j=1}^n h_i h_j \partial_{e_j}\left(t \mapsto \frac{\partial f}{\partial x_i}(t)\right)(a) \end{aligned}$$

en notant $\partial_{e_j}\left(t \mapsto \frac{\partial f}{\partial x_i}(t)\right)(a)$ par $\frac{\partial^2 f}{\partial x_j \partial x_i}(a)$, on obtient l'expression

$$\mathrm{D}^2 f(a)(h,h) = \sum_{i,j=1}^n h_i h_j \frac{\partial^2 f}{\partial x_j \partial x_i}(a).$$

Ce qu'on peut encore représenter matriciellement comme

$$\mathrm{D}^2 f(a)(h,h) = h^T \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2}(a) & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n}(a) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1}(a) & \cdots & \frac{\partial^2 f}{\partial {x_n}^2}(a) \end{pmatrix} h$$

où $\partial^2 f / \partial x_i^2$ est une écriture simplifiée pour $(\partial^2 f / \partial x_i \partial x_i)$. La matrice dans le membre de droite de l'égalité est la ***hessienne*** de f au point a.

C Les formes quadratiques

Les formes bilinéaires symétriques donnent naissance aux formes quadratiques. Celles-ci donnent des programmes d'optimisations qu'on résout assez bien par les méthodes introduites dans ce cours. On passe d'un contexte linéaire dans le cas des programmes linéaires à celui de degré 2.

Définition C.1. Soit E un espace vectoriel sur $\mathbb{R}$. On dit que $Q : E \to \mathbb{R}$ est une forme quadratique sur E s'il existe une forme bilinéaire symétrique $\phi : E \times E \to \mathbb{R}$ telle que pour tout $x \in E$, $Q(x) = \phi(x, x)$.

Cette définition permet déjà de voir qu'étant donné un produit scalaire $\langle \cdot, \cdot \rangle$ sur E, l'application $x \mapsto \langle x, x \rangle$ est une forme quadratique ; c'est la norme au carré (la norme associée au produit scalaire). Une forme quadratique Q sur $\mathbb{R}^n$ est donc la donnée d'une matrice symétrique P telle que

$$\forall x \in \mathbb{R}^n, \quad Q(x) = x^T P x.$$

Exemple C.1. La matrice carrée $P = \begin{pmatrix} 1 & 2 \\ 2 & 0 \end{pmatrix}$ donne lieu à la forme quadratique Q l'expression pour $(x, y) \in \mathbb{R}^2$ est

$$Q\big((x, y)\big) = x^2 + 4xy.$$

On remarque que l'expression obtenue est un polynôme de degré total [16] 2. Ce fait est général : étant donné une forme quadratique non nulle Q sur $\mathbb{R}^n$, l'expression $Q(x)$ est un polynôme de degré total 2 en les coordonnées du vecteur $x \in \mathbb{R}^n$.

Remarque 9. Par abus, dans le cas de $\mathbb{R}^n$, ces adjectifs sont accolés à la matrice symétrique P telle que $Q(x) = x^T P x$.

Question 3-39 Donner, quand possible, un exemple de forme quadratique sur $\mathbb{R}^2$ qui soit (n'oubliez pas de justifier vos réponses) :
- positive, non définie ;
- définie positive ;
- définie non positive ni négative.

Dessiner un croquis illustrant le graphe de la fonction quadratique donnée dans chacun des cas [a].

a. Vous êtes autorisés (encouragés ?) à utiliser `matplotlib`.

On admet par la suite le résultat important suivant :

Théorème C.1. *Toute matrice symétrique $P \in M_n(\mathbb{R})$ est diagonalisable sur $\mathbb{R}$ dans une base orthonormée.*

Ce résultat, important, est valable sous des conditions beaucoup plus larges sur le corps de base ; on le montre en partie à l'aide d'un algorithme dit de Gauss.

Question 3-40 À l'aide du déterminant d'une matrice carrée, donner une condition sur une forme quadratique Q sur $\mathbb{R}^n$ afin que celle-ci soit :
- définie ;
- non positive et non négative.

Pouvez-vous en dire plus dans le cas des formes quadratiques sur $\mathbb{R}^2$? Sait-on, dans ce cas, quand une forme quadratique est positive ? Comment différencier le cas positif du cas négatif ? Formaliser le cas des formes quadratiques sur $\mathbb{R}^2$.

16. C'est le degré du polynôme obtenu en remplaçant toutes les variables par une unique variable t.

Question 3-41 Suffit-il qu'une matrice symétrique ait des coefficients positifs pour que la forme quadratique associée le soit ?

Question 3-42 Donner des conditions suffisantes sur la positivité de la matrice associée à une forme quadratique pour que celle-ci soit

- convexe ou concave ;
- ni l'un ni l'autre.

Ces conditions sont-elles nécessaires ?