Approximations locales d'une fonction différentiable de plusieurs variables

Guillaume Lecué[1]

On a vu, au chapitre précédent, une manière de résoudre les problèmes d'optimisation basé sur des approximations locales d'une fonction et d'un ensemble d'une manière informelle. Dans ce chapitre, on décrit le comportement local d'une fonction de $\mathbb{R}^n$ dans $\mathbb{R}$. Dans le cas différentiable, c'est le gradient qui porte toute cette information. On insistera donc sur cette notion fondamentale aussi bien d'un point de vue théorique que pratique. Quand la fonction a plus de régularité, on peut donner des approximations locales d'une fonction à des ordres plus grands. C'est le but des formules de Taylor qu'on présente aussi ici et qui sont aussi utiles pour résoudre des problèmes d'optimisation.

1 Différentielle, gradient, dérivées partielles et Jacobienne

On introduit dans cette section, plusieurs outils pour l'approximation du premier ordre d'une fonction différentiable.

1.1 Différentielle et gradient

Idée : Différencier une fonction en un point, c'est chercher la meilleure approximation affine de cette fonction en ce point.

Définition 1.1 *Soient $(E, \|\cdot\|_E)$ et $(F, \|\cdot\|_F)$ deux espaces espaces vectoriels réels normés (e.v.n.). Soit U un ouvert de E et $f : U \to F$ une fonction. Soit $a \in U$. On dit que f* ***est différentiable (ou dérivable) en*** *a quand il existe une application linéaire df_a continue de E dans F telle que, quand $h \to 0$,*

$$f(a+h) = f(a) + df_a(h) + o(\|h\|_E).$$

Soit $g : E \to F$. On rappelle "$g(h) = o(\|h\|_E)$ quand $h \to 0$" signifie que $g(h)/\|h\|_E \to 0$ quand $h \to 0$. Autrement dit $g(h)/\|h\|_E = o(1)$ quand $h \to 0$. On a donc aussi $o(\|h\|_E) = \|h\|_E\, o(1)$.

Remarque 1.2

1. *Si df_a existe alors elle est unique (sinon, on aurait pour une autre application linéaire df'_a que $df'_a(h) - df_a(h) = o(\|h\|_E)$ quand $h \to 0$. En particulier pour tout $x \in E$, on aurait $df'_a(\lambda x) - df_a(\lambda x) = o(\|\lambda x\|_E)$ quand $\lambda \to 0$ et par linéarité, $df'_a(x) - df_a(x) = o(1)$ quand $\lambda \to 0$, donc $df'_a(x) = df_a(x)$).*
2. *Si f est différentiable en a alors elle est continue en a (en effet, $df_a(h) \to 0$ quand $h \to 0$, car df_a est continue, ainsi $f(a+h) \to f(a)$ quand $h \to 0$).*

1. CREST, ENSAE. Bureau 3029, 5 avenue Henry Le Chatelier. 91 120 Palaiseau. Email : guillaume.lecue@ensae.fr.

3. *Dans le cas de fonctions définies sur $\mathbb{R}$ et à valeurs dans $\mathbb{R}$, on retrouve bien la définition usuelle de dérivée. Soit $f : \mathbb{R} \to \mathbb{R}$, on a "$f$ est différentiable en a" si et seulement si "f est dérivable (au sens habituel) en a" et dans ce cas, on a $df_a(h) = f'(a)h$.*
4. *En dimension infinie, la notion de différentiation dépend des normes sur E et F. En dimension finie, comme toutes les normes sont équivalentes, cette notion est indépendante des normes choisies sur E et F.*
5. *Si $T : (E, \|\cdot\|_E) \to (F, \|\cdot\|_F)$ est une application linéaire alors il y a équivalence entre*
   a) T est borné, i.e. il existe $L > 0$ tel que pour tout $x \in E, \|Tx\|_F \leq L \|x\|_E$.
   b) T est continue sur E
   c) T est continue en 0.
   En effet, a) implique b) car pour tout $x, y \in E, \|Tx - Ty\|_F = \|T(x-y)\| \leq L \|x-y\|_E$. Alors T est Lipschitz donc continue. b) implique c) est triviale. Pour c) implique a), on a $T(0) = 0$ et donc, par continuité, il existe $\delta > 0$ tel que pour tout $x \in E$ si $\|x\|_E \leq \delta$ alors $\|Tx\|_F \leq 1$. Soit $y \in E$, on note $x = \delta y / \|y\|_E$ alors $\|x\|_E = \delta$ alors $\|Tx\|_F \leq 1$ et par linéarité, $\|Ty\|_F \leq \delta^{-1} \|y\|_E$.
6. *En dimension finie, toutes les applications linéaires sont continues. En effet, soit T une application linéaire de $\mathbb{R}^n$ dans F. On va montrer que T est bornée. Comme toutes les normes sont équivalentes sur $\mathbb{R}^n$, on choisit de munir $\mathbb{R}^n$ de la norme $\|\cdot\|_1$. On note $(e_i)_{i=1}^n$ la base canonique de $\mathbb{R}^n$. On a pour tout $x = (x_i)_{i=1}^n \in \mathbb{R}^n$*

$$\|Tx\|_F = \left\| \sum_{i=1}^n x_i T e_i \right\|_F \leq \sum_{i=1}^n |x_i| \, \|Te_i\|_F \leq L \|x\|_1$$

   où $L = \max_{1 \leq i \leq n} \|Te_i\|_F$. Donc T est borné et donc continue.

Définition 1.3 *Soient $(E, \|\cdot\|_E)$ et $(F, \|\cdot\|_F)$ deux espaces espaces vectoriels réels normés (e.v.n.). Soit U un ouvert de E et $f : U \to F$ une fonction. On dit que f* ***est différentiable sur*** *U quand f est différentiable en tout point de U. Si tel est le cas, on appelle* ***différentielle****, l'application,*

$$df : \left\{ \begin{array}{ccc} U & \to & \mathcal{L}_c(E,F) \\ a & \to & df_a \end{array} \right.$$

où $\mathcal{L}_c(E,F)$ est l'espace vectoriel réel des applications linéaires continues de E dans F. De plus, si df est continue sur U (pour E muni de $\|\cdot\|_E$ et $\mathcal{L}_c(E,F)$ muni de la norme d'opérateur de $(E, \|\cdot\|_E)$ dans $(F, \|\cdot\|_F)$) alors on dit que f ***est de classe*** *$\mathcal{C}^1$.*

On rappelle que la norme d'opérateur de $T \in \mathcal{L}_c(E,F)$ est définie par $\|T\| = \|T\|_{E \to F} = \sup_{x \neq 0} \|Tx\|_F / \|x\|_E$. Par définition, on a pour tout $x \in E, \|Tx\|_F \leq \|T\| \, \|x\|_E$. Par ailleurs, si $R \in \mathcal{L}_c(F,G)$ alors $\|RT\|_{E \to G} \leq \|R\|_{F \to G} \|T\|_{E \to F}$.

Exemple 1.4 *Montrer que f est différentiable sur $\mathbb{R}^{n \times n} \times \mathbb{R}^{n \times n}$ et calculer sa différentielle pour*

$$f : \left\{ \begin{array}{ccc} \mathbb{R}^{n \times n} \times \mathbb{R}^{n \times n} & \to & \mathbb{R}^{n \times n} \\ (A, B) & \to & AB. \end{array} \right.$$

Soit $(A, B) \in \mathbb{R}^{n \times n} \times \mathbb{R}^{n \times n}$. On a

$$\begin{aligned} f((A,B) + (H,K)) &= f(A+H, B+K) = AB + AK + HB + HK \\ &= f(A,B) + \varphi_{(A,B)}(H,K) + HK \end{aligned}$$

où $\varphi_{(A,B)}(H,K) = AK + HB$ est une fonction linéaire (continue). Il reste alors à montrer que $HK = o(\|(H,K)\|)$ quand $(H,K) \to 0$ pour une certaine norme $\|\cdot\|$ sur $\mathbb{R}^{n\times n} \times \mathbb{R}^{n\times n}$. On peut, par exemple, prendre $\|(H,K)\| = \max(\|H\|, \|K\|)$ où $\|H\| = \sup_{x\neq 0} \|Hx\|_2 / \|x\|_2$ est la norme d'opérateur de ℓ_2^n sur lui-même. On a $\|HK\| \leq \|H\| \|K\| \leq \|(H,K)\|^2$ et donc $HK = o(\|(H,K)\|)$ quand $(H,K) \to 0$. Donc f est différentiable et $df_{(A,B)} = \varphi_{(A,B)}$. Par ailleurs, la différentielle de f, $df : (A,B) \to df_{(A,B)}$ est une application linéaire de l'espace de dimension finie $\mathbb{R}^{n\times n} \times \mathbb{R}^{n\times n}$ à valeur dans l'e.v.n. $\mathcal{L}_c(\mathbb{R}^{n\times n} \times \mathbb{R}^{n\times n}, \mathbb{R}^n)$, elle est donc continue et donc f est de classe $\mathcal{C}^1$. En fait, f est une fonction polynomiale, elle est donc $\mathcal{C}^\infty$.

Exemple 1.5 *Soit $y \in \mathbb{R}^n$ et $A \in \mathbb{R}^{n\times n}$. Calculer la différentielle de*

$$f : \left\{ \begin{array}{ccc} \mathbb{R}^n & \to & \mathbb{R} \\ x & \to & \|y - Ax\|_2^2 . \end{array} \right.$$

Soit $x \in \mathbb{R}^n$ et $h \in \mathbb{R}^n$. On a

$$f(x+h) = \|y - Ax\|_2^2 - 2\langle y - Ax, Ah\rangle + \|Ah\|_2^2 = f(x) + \varphi_x(h) + \|Ah\|_2^2$$

où $\varphi_x(h) = \langle -2A^\top(y - Ax), h\rangle$ est une application linéaire (et continue en h). Il suffit de montrer que $\|Ah\|_2^2 = o(\|h\|_2)$ quand $h \to 0$ pour montrer que φ_x est la différentielle de f en x. On a

$$\|Ah\|_2^2 \leq \|A\|^2 \|h\|_2^2 = o(\|h\|_2)$$

où $\|A\|$ est la norme d'opérateur de $A : \ell_2^n \to \ell_2^n$. Donc la différentielle de f est $df : x \in \mathbb{R}^n \to (df_x : h \to \langle -2A^\top(y - Ax), h\rangle)$.

Quand $f : U \to \mathbb{R}$ (où U est un ouvert de E) est une fonction différentiable à valeurs réelles alors la différentielle associée est une fonction linéaire de E dans $\mathbb{R}$. Ces fonctions sont appelées des **formes linéaires**. Il se trouve que les formes linéaires continues d'un espace de Hilbert ont une représentation particulière due au théorème de Riesz : "Pour toute forme linéaire continue T sur un espace de Hilbert H il existe un unique $a \in H$, tel que $\forall x \in H, T(x) = \langle a, x\rangle$". C'est grâce à cette représentation qu'on peut définir le gradient d'une fonction différentiable à valeurs réelles et définie sur un ouvert d'un Hilbert.

Définition 1.6 *Soit H un espace de Hilbert. Soit U un ouvert de H et $f : U \to \mathbb{R}$ une application différentiable en $a \in U$. On note $df_a \in \mathcal{L}_c(H, \mathbb{R})$ sa différentielle en a. Par le théorème de représentation de Riesz, il existe un unique vecteur de H, noté $\nabla f(a)$, tel que $df_a(h) = \langle \nabla f(a), h\rangle$ pour tout $h \in H$. Le vecteur $\nabla f(a)$ est appelé **gradient de f en** a.*

Dans l'exemple 1.5, on a montré que $df_x(h) = \langle -2A^\top(y - Ax), h\rangle$. On voit alors directement que le gradient de f en x est $\nabla f(x) = -2A^\top(y - Ax)$.

Idée : Le gradient est la notion la plus importante de ce cours aussi bien d'un point de vue théorique que algorithmique. Il apparaît dans tous les principaux résultats et dans tous les algorithmes du cours.

La différentielle est linéaire : si $f, g : U \to F$ où U est un ouvert de E et f et g sont deux fonctions différentiables en $a \in U$ alors $\lambda f + g$ est différentiable en a et $d(\lambda f + g)_a = \lambda df_a + dg_a$. Une propriété connue sous le nom de **chain rule** concerne la différentielle d'une composée de deux fonctions.

Proposition 1.7 (Chain rule) *Soient E, F, G des e.v.n., $U \subset E$ un ouvert et $V \subset F$ un ouvert. Soit $f : U \to V$ et $g : V \to G$. On suppose que f est différentiable en a et g en $f(a)$ alors $g \circ f : U \to G$ est différentiable en a et*

$$d(g \circ f)_a = dg_{f(a)} \circ df_a.$$

Preuve. Quand $h \to 0$, on a

$$(g \circ f)(a+h) = g(f(a+h)) = g(f(a) + df_a(h) + o(\|h\|_E)) = g(f(a) + h')$$

où $h' = df_a(h) + o(\|h\|_E) \to 0$ quand $h \to 0$ car df_a est continue en 0. On a donc quand $h \to 0$,

$$(g \circ f)(a+h) = g(f(a)) + dg_{f(a)}(h') + o(\|h'\|_F).$$

De plus, par linéarité et continuité de $dg_{f(a)}$ en 0, on a, quand $h \to 0$,

$$dg_{f(a)}(o(\|h\|_E)) = \|h\|_E \, dg_{f(a)}(o(1))) = o(\|h\|_E)$$

et donc

$$\begin{aligned} dg_{f(a)}(h') &= dg_{f(a)}(df_a(h) + o(\|h\|_E)) \\ &= dg_{f(a)}(df_a(h)) + dg_{f(a)}(o(\|h\|_E)) = dg_{f(a)}(df_a(h)) + o(\|h\|_E). \end{aligned}$$

De plus, df_a est continue, elle est donc bornée, càd $\|df_a(h)\|_F \leq L \|h\|_E$; on a alors

$$o(\|h'\|_F) = o(\|df_a(h)\|_F) + o(o(\|h\|_E)) = o(\|h\|_E).$$

■

Pour retrouver la chain rule de manière informelle, on peut se rappeler que si $h \to 0$ alors

$$g(f(a+h)) \approx g(f(a) + df_a(h)) \approx g(f(a)) + dg_{f(a)}(df_a(h))$$

et donc on reconnait $d(g \circ f)_a(h) = dg_{f(a)}(df_a(h))$.

Exemple 1.8 *On reprend l'Exercice 1.5, en regardant f comme étant la composée $f = g \circ k$ où $g(z) = \|z\|_2^2$ et $k(x) = Ax - y$. On a $dg_z(h) = \langle 2z, h \rangle$ et $dk_x(h) = Ah$. On a alors $df_x(h) = dg_{k(x)}(dk_x(h)) = dg_{Ax-y}(Ah) = \langle 2(Ax-y), Ah \rangle = \langle 2A^\top (Ax-y), h \rangle$.*

On utilise souvent la chain rule pour calculer des gradients car en voyant une fonction comme la composée de fonctions dont on peut calculer facilement le gradient, on peut en déduire le gradient de la fonction elle-même grâce à la chain rule. La chain rule se transcrit ainsi en terme de gradient.

Proposition 1.9 *Soit H un espace de Hilbert. Si $f : H \to \mathbb{R}$ et $g : \mathbb{R} \to \mathbb{R}$ sont différentiables alors pour tout $x \in H$,*

$$\nabla(g \circ f)(x) = g'(f(x)) \nabla f(x).$$

Preuve. Pour tout $h \in H$, $d(g \circ f)_x(h) = \langle \nabla(g \circ f)(x), h \rangle$ et

$$d(g \circ f)_x(h) = dg_{f(x)}(df_x(h)) = g'(f(x)) \langle \nabla f(x), h \rangle = \langle g'(f(x)) \nabla f(x), h \rangle.$$

■

Exemple 1.10 *Soient $f, g : U \to \mathbb{R}$ deux fonctions différentiables en $a \in U$ (où U est un ouvert d'un espace de Hilbert). Alors le produit est différentiable et*

$$\nabla(fg)_a = \nabla f(a)g(a) + f(a)\nabla g(a).$$

En effet, on peut voir le produit fg comme la composition $(fg)(a) = \varphi(\psi(a))$ où $\psi(a) = (f(a), g(a))$ et $\varphi(\alpha, \beta) = \alpha\beta$. On a $d(fg)_a = d\varphi_{\psi(a)} \circ d\psi_a$ et comme

$$d\varphi_{(\alpha,\beta)}(h, k) = \alpha k + h\beta \text{ et } d\psi_a = (df_a, dg_a)$$

on a bien $d(fg)_a = df_a g(a) + f(a)dg_a$ et donc $\nabla(fg)(a) = \nabla f(a)g(a) + f(a)\nabla f(a)$.

Exemple 1.11 *Soit H un espace de Hilbert munit de sa norme $\|\cdot\|_2$, $a \in H$ et $f : x \in H \to \|a - x\|_2$. On a pour tout $x \in H\backslash\{a\}$,*

$$\nabla f(x) = \frac{x - a}{\|x - a\|_2}.$$

En effet, on écrit f comme la composée de deux fonctions $f = \varphi \circ F$ où $F : x \in H \to \|a - x\|_2^2$ et $\varphi : t \in \mathbb{R}_+^* \to \sqrt{t}$. On a pour tout $x \neq a$, $\nabla F(x) = -2(a - x)$ et tout $t > 0$, $\varphi'(t) = 1/(2\sqrt{t})$. Donc

$$\nabla f(x) = \varphi'(F(x))\nabla F(x) = -2(a - x)/(2\|a - x\|_2).$$

Exemple 1.12 *On peut calculer la dérivé d'une fonction réciproque en utilisant la chain rule. En effet, si $f : U \to \mathbb{R}$ est une fonction $\mathcal{C}^1$ inversible d'inverse notée $f^{-1} : f(U) \to U$ alors f^{-1} est aussi $\mathcal{C}^{-1}$. On obtient sa dérivée de la manière suivante : pour tout $x \in U$, on a $f^{-1}(f(x)) = x$ alors en dérivant cette égalité, par la chain rule, on obtient $(f^{-1})'(f(x))f'(x) = 1$ et donc pour tout $y \in f(U)$, en écrivant $x = f^{-1}(y)$, on a $(f^{-1})'(y) = 1/f'(f^{-1}(y))$.*

Par exemple, si $f = \cos :]0, \pi[\to]-1, 1[$ alors $f^{-1} = \arccos :]-1, 1[\to]0, \pi[$ et donc par la fomule de dérivation d'une fonction inverse

$$\arccos'(y) = \frac{1}{\cos'(\arccos(y))} = \frac{-1}{\sqrt{1 - y^2}}$$

où on a utiliser que $\cos' = -\sin$ et $\cos^2(x) + \sin^2(x) = 1$.

1.2 Dérivées partielles

Idée : Le gradient est la notion la plus importante de ce cours. Il faut donc savoir se le représenter (on verra ça dans la Section 3) et le calculer. On a déjà vu deux techniques de calcul de gradient (faire un dévéloppement limité au premier ordre de $f(a + h)$ et la chain rule). Ici on donne une autre technique de calcul du gradient basé sur les dérivées partielles. Sous l'hypothèse d'existence et continuité des dérivées partielles, on verra que le gradient est le vecteur des dérivées partielles. Celà signifie que pour dériver une fonction de n variables il suffit de dériver n fonctions de $\mathbb{R}$ dans $\mathbb{R}$. Etant donné qu'on est plus habitués à travailler avec des fonctions uni-variées, cette méthode est très pratique.

Définition 1.13 *Soient E, F deux e.v.n., U un ouvert de E, $f : U \to F$, $a \in U$ et $v \in E$. Si la fonction $t \in \mathbb{R} \to f(a + tv)$ est différentiable en 0, on dit que f* ***admet une dérivée partielle en a selon v*** *et on note*

$$f'_v(a) = \partial_v f(a) = \lim_{\substack{t \to 0 \\ t \neq 0}} \frac{f(a + tv) - f(a)}{t}.$$

Proposition 1.14 *Soit U un ouvert de E et $f : U \to F$. On suppose que f est différentiable en $a \in U$; alors f admet une dérivée partielle en a selon toute les directions $v \in E$ et $\partial_v f(a) = df_a(v)$.*

Preuve. Quand $t \to 0$, on a $f(a+tv) = f(a) + df_a(tv) + o(\|tv\|_E)$. Alors, quand $t \neq 0$ et $t \to 0$,

$$\frac{f(a+tv) - f(a)}{t} = df_a(v) + o(\|v\|_E) = df_a(v) + o(1).$$

Autrement dit,

$$\lim_{\substack{t \to 0 \\ t \neq 0}} \frac{f(a+tv) - f(a)}{t} = df_a(v).$$

■

Notation : Quand $f : U \to F$ où U est un ouvert de $\mathbb{R}^n$ et si $(e_1, \dots, e_n)$ est la base canonique de $\mathbb{R}^n$ alors, si f est dérivable en $a \in U$ dans la direction e_i, on dit que f admet une **dérivée partielle d'indice i en a** et on note

$$\partial_i f(a) = f'_{e_i}(a) = \partial_{e_i} f(a).$$

Les dérivées partielles d'une fonction jouent un rôle important car elles permettent de calculer le gradient d'une fonction en se ramenant à des calculs de dérivées de fonctions d'une seule variable.

Proposition 1.15 *Soit $f : U \to \mathbb{R}$ où U est un ouvert de $\mathbb{R}^n$. On suppose que f est dérivable en $a \in U$. Alors f admet des dérivées partielles d'indice i en a pour tout $i \in \{1, \dots, n\}$ et on a*

$$\nabla f(a) = (\partial_i f(a))_{i=1}^n .$$

Preuve. D'après la Proposition 1.14, pour tout $i = 1, \dots, n$

$$\partial_i f(a) = df_a(e_i) = \langle \nabla f(a), e_i \rangle.$$

Donc, la i-ième coordonnée de $\nabla f(a)$ est $\partial_i f(a)$. ■

Autrement dit, le gradient (quand il existe) a pour coordonnées les dérivées partielles d'ordre i de f.

Exemple 1.16 *Soient $a_1, \dots, a_m \in \mathbb{R}^n$ et $b_1, \dots, b_m \in \mathbb{R}$. Calculer le gradient de*

$$f : x \in \mathbb{R}^n \to \log \left(\sum_{j=1}^m \exp\left(b_j + \langle x, a_j \rangle\right) \right).$$

On a pour tout $x \in \mathbb{R}^n$, $f(x) = \log(g(x))$ et d'après la chain rule,

$$\nabla f(x) = (\log)'(g(x)) \nabla g(x) = \frac{\nabla g(x)}{g(x)} \text{ où } g(x) = \sum_{j=1}^m \exp\left(b_j + \langle x, a_j \rangle\right).$$

De plus $\nabla g(x) = (\partial_i g(x))_{i=1}^n$ et on vérifie que, pour $a_j = (a_{ji})_{i=1}^n$,

$$\partial_i g(x) = \sum_{j=1}^m a_{ji} \exp\left(b_j + \langle x, a_j \rangle\right).$$

On note $z = \left(\exp\left(b_j + \langle x, a_j \rangle\right)\right)_{j=1}^m$ et $A = [a_1 | \cdots | a_m] \in \mathbb{R}^{n \times m}$ et $\mathbf{1} = (1)_{j=1}^m$ alors $\nabla f(x) = Az / \langle \mathbf{1}, z \rangle$.

Cependant, il se peut que les dérivées partielles d'ordre i existent pour tout $i \in \{1, \dots, n\}$ sans pour autant que le gradient existe (voir Exemple 1.17).

Exemple 1.17 *La Proposition 1.14 montre que si f est différentiable alors elle admet des dérivées partielles dans toutes les directions. Cependant, la réciproque à la Proposition 1.14 est fausse. On peut le voir sur l'exemple suivant :*

$$f := \left\{ \begin{array}{ccc} \mathbb{R}^2 & \to & \mathbb{R} \\ (x,y) & \to & \left| \begin{array}{ll} y^2/x & si\ x \neq 0 \\ y & si\ x = 0 \end{array} \right. \end{array} \right.$$

On montre que f est dérivable en $(0,0)$ dans toutes les directions mais f n'est pas continue en $(0,0)$.

En effet, f n'est pas continue en $(0,0)$ car $f(0,0) = 0$ mais $f(1/n^2, 1/n) = 1$ pour tout $n \in \mathbb{N}^*$ alors $(f(1/n^2, 1/n))_n$ ne converge pas vers $f(0,0)$.

Montrons maintenant que f admet des dérivées partielles en $(0,0)$ dans toutes les directions. Soient $(u,v) \in \mathbb{R}^2$ et $t \neq 0$. On a

$$\frac{f((0,0)+t(u,v)) - f(0,0)}{t} = \frac{f(tu,tv)}{t} = \left| \begin{array}{ll} (tv)^2/(t^2u) & si\ tu \neq 0 \\ tv/t & si\ tu = 0 \end{array} \right. = \left| \begin{array}{ll} v^2/u & si\ u \neq 0 \\ v & si\ u = 0. \end{array} \right.$$

Donc, f est dérivale en $(0,0)$ selon (u,v) et on a

$$\partial_{(u,v)} f(0,0) = \left| \begin{array}{ll} v^2/u & si\ u \neq 0 \\ v & si\ u = 0. \end{array} \right.$$

Dans l'Exemple 1.17, on a vu qu'une fonction peut admettre des dérivées partielles dans toutes les directions sans pour autant être différentiable (même pas continue). Cependant, en supposant l'existence ET la continuité de toutes les dérivées partielles d'ordre i, la réciproque devient vraie.

Proposition 1.18 *Soit $f : U \to \mathbb{R}$ où U est un ouvert de $\mathbb{R}^n$. Si toutes les dérivées partielles de f existent sur U et sont continues (en tant que fonctions $\partial_i f : U \to \mathbb{R}$) alors f est différentiable sur U et pour tout $a \in U, \nabla f(a) = (\partial_i f(a))_{i=1}^n$.*

Preuve. Soit $a \in U$. Montrons que quand $h \to 0$, on a $f(a+h) = f(a) + \langle \nabla, h \rangle + o(\|h\|)$ où $\nabla = (\partial_i f(a))_{i=1}^n$ ce qui équivaut à montrer que, si $x \to a$ alors

$$g(x) - g(a) = o(\|x - a\|) \text{ où } g(x) = f(x) - \langle \nabla, x \rangle.$$

Soit $\epsilon > 0$. Comme les dérivées partielles de f en a sont continues, il existe $\alpha > 0$ tel que $B_1(a,\alpha) \subset U$ et pour tout $x \in B_1(a,\alpha)$, on a

$$\|\partial_i g(x)\|_F = \|\partial_i f(x) - \partial_i f(a)\|_F \leq \epsilon. \tag{1.1}$$

On rappel que $B_1(a,\alpha)$ est la boule de centre a et de rayon α pour la norme ℓ_1^n.

Soit $x \in B_1(a,\alpha)$. On pose

$$y_0 = a = (a_1, \dots, a_n)^\top \text{ et } \forall k = 1, \dots, n, y_k = (x_1, \cdots, x_k, a_{k+1}, \cdots, a_n)$$

de sorte que $y_0 = a$ et $y_n = x$, ainsi $y_0, y_1, \cdots, y_n$ forment un "chemin" de a vers x en ne changeant qu'une seule coordonnée de a en une de x à chaque pas. On pose pour tout $k = 1, \dots, n$,

$$g_k : \left\{ \begin{array}{ccc} [a_k, x_k] & \to & F \\ t & \to & g(x_1, \cdots, x_{k-1}, t, a_{k+1}, \cdots, a_n) \end{array} \right.$$

On a pour tout $t \in (a_k, x_k)$, $g_k'(t) = \partial_k g(x_1, \cdots, x_{k-1}, t, a_{k+1}, \cdots, a_n)$ et donc, d'après (1.1), $\|g_k'(t)\|_F \leq \epsilon$. Le théorème de l'inégalités des accroissements finis donne ensuite

$$\|g_k(a_k) - g_k(x_k)\|_F \leq \epsilon |a_k - x_k|.$$

(Attention, ici on applique le théorème des inégalités des accroissements finis et pas celui des accroissements finis qui est faux pour les fonctions à valeurs dans un e.v.n. quelconque – on rappelle ce théorème dans la suite). Par ailleurs, on a $g_k(a_k) = g(y_{k-1})$ et $g_k(x_k) = g(y_k)$ pour tout $k = 1, \cdots, n$ et donc

$$\begin{aligned}\|g(x) - g(a)\|_F &\leq \left\| \sum_{k=1}^n g(y_k) - g(y_{k-1}) \right\|_F \leq \sum_{k=1}^n \|g(y_k) - g(y_{k-1})\|_F \\ &= \sum_{k=1}^n \|g_k(a_k) - g_k(x_k)\|_F \leq \epsilon \|a - x\|_1 .\end{aligned}$$

On a donc bien $g(x) - g(a) = o(\|x - a\|_1)$ quand $x \to a$. ∎

En conclusion, si f admet des dérivées partielles continues alors f est dérivable. La réciproque est fausse et peut se voir sur l'exemple suivant d'une fonction dérivable sur $\mathbb{R}$ dont la dérivée n'est pas continue en 0 :

$$f : \left\{ \begin{array}{ccc|cc} \mathbb{R} & \to & & \mathbb{R} & \\ x & \to & & x^2 \sin(1/x) & \text{si } x \neq 0 \\ & & & 0 & \text{si } x = 0. \end{array} \right.$$

En effet, f est différentiable en 0 et $f'(0) = 0$ car pour tout $h \neq 0$ et $h \to 0$,

$$|f(0 + h) - f(0)| = h^2 \sin(1/h) = o(h).$$

Pour tout $x \neq 0$, $f'(x) = 2x \sin(1/x) - \cos(1/x)$ donc f' n'est pas continue en 0.

On a donc les deux implications :

différentiable $\Rightarrow$ les dérivées partielles existent

les dérivées partielles existent et sont continues $\Rightarrow$ différentiable

et dans les deux cas $\nabla f = (\partial_i f)$. Mais aucune réciproque n'est vraie. Par ailleurs, on voit maintenant qu'une fonction est $\mathcal{C}^1$ si et seulement si les dérivées partielles existent et sont continues.

Exemple 1.19 *On considère la fonction déterminant* $\det : A \in \mathbb{R}^{n \times n} \to \det(A)$. *Montrer* det *est de classe* $\mathcal{C}^1$ *et calculer sa différentielle en la matrice identité* I_d.

La fonction det *est un fonction polynomiale, c'est donc une fonction de classe* $\mathcal{C}^\infty$. *Soit* $H \in \mathbb{R}^{n \times n}$ *une matrice de valeurs propres* $\lambda_1, \ldots, \lambda_n$ *(potentiellement complexes). Soit* $t > 0$. *La matrice* $I_d + tH$ *a pour valeurs propres* $1 + t\lambda_1, \ldots, 1 + t\lambda_n$ *et donc quand* $t \to 0$

$$\det(I_d + tH) = \prod_{i=1}^n (1 + t\lambda_i) = 1 + t \sum_{i=1}^n \lambda_i + o(t) = \det(I_d) + t\mathrm{Tr}(H) + o(t).$$

On en déduit donc que la dérivée partielle de det *en* I_d *dans la direction* H *est* $\mathrm{Tr}(H)$. *Par ailleurs,* det *est* $\mathcal{C}^1$ *donc ses dérivées partielles aussi et donc* $\nabla \det(I_d) = (\mathrm{Tr}(E_{ij}))_{1 \leq i,j \leq n} = I_d$ *et*

$$d_{I_d} \det(H) = \left\langle \nabla \det(I_d), H \right\rangle = \sum_i H_{ii} = \mathrm{Tr}(H).$$

1.3 Matrice Jacobienne et Jacobien

La notion de gradient est propre aux fonctions à valeurs réelles (et définie sur un ouvert d'un Hilbert). Il est cependant possible de généraliser cette notion à des fonctions à valeurs dans $\mathbb{R}^m$, c'est l'idée de la matrice Jacobienne.

Définition 1.20 *Soit U un ouvert de $\mathbb{R}^n$ et $f : U \to \mathbb{R}^m$ une fonction différentiable en un point $a \in U$. On note $(e_j)_{j=1}^n$ la base canonique de $\mathbb{R}^n$ et par $(e'_i)_{i=1}^m$ celle de $\mathbb{R}^m$. On écrit $f = \sum_{i=1}^m f_i e'_i$ où pour tout $i = 1, \dots, m$, $f_i : U \to \mathbb{R}$ est la i-ième composante de f. La différentielle de f en a est une application linéaire de $\mathbb{R}^n$ dans $\mathbb{R}^m$, elle peut donc se représenter par une matrice $J(f)(a) \in \mathbb{R}^{m\times n}$, appelée* ***matrice Jacobienne de f en*** *a, donnée par*

$$J(f)(a) = (\partial_j f_i(a))_{\substack{1\le j\le n \\ 1\le i\le m}} = \begin{pmatrix} \nabla f_1(a)^\top \\ \vdots \\ \nabla f_m(a)^\top \end{pmatrix}.$$

On a alors $df_a(h) = J(f)(a)h$. Quand $m = n$, le déterminant de $J(f)(a)$ est appelé ***Jacobien*** *de f en a.*

Remarque 1.21
1. *Quand $f : \mathbb{R}^n \to \mathbb{R}$ est différentiable en a, on a $df_a(h) = \langle \nabla f(a), h\rangle$ et si $f : \mathbb{R}^n \to \mathbb{R}^m$ est différentiable en a, on a $df_a(h) = \big(\langle \nabla f_i(a), h\rangle\big)_{i=1}^m$.*
2. *Si $f : \mathbb{R}^n \to \mathbb{R}$, $J(f)(a) = (\nabla f(a))^\top$ (on rappelle que $u^\top v = \langle u, v\rangle$ pour tout $u, v \in \mathbb{R}^n$).*

Le Jacobien joue un rôle essentiel dans le théorème de changement de base. Avant de rappeler ce résultat, on rappelle, que $\varphi : U \to V$ est un **$\mathcal{C}^1$-difféomorphisme** quand φ est bijective, φ est $\mathcal{C}^1$ et φ^{-1} est $\mathcal{C}^1$. Il est cependant, en général, difficile de calculer φ^{-1} et de montrer qu'elle est $\mathcal{C}^1$. On utilise plutôt la caractérisation suivante d'un $\mathcal{C}^1$-difféomorphisme : si φ est une bijection $\mathcal{C}^1$ telle que $d\varphi_x$ est inversible en tout point x (càd $\det(J(\varphi)(x)) \neq 0$) alors φ est un $\mathcal{C}^1$-difféomorphisme.

Théorème 1.22 *Soient U et V deux ouverts de $\mathbb{R}^n$ et $\varphi : U \to V$ un $\mathcal{C}^1$-difféomorphisme de U sur V. Soit $f : V \to U$ une fonction intégrable sur V alors*

$$\int_V f(v)dv = \int_U f(\varphi(u))|\det J(\varphi)(u)|du.$$

L'exemple classique est celui du changement de base des coordonnées cartésiennes aux coordonnées polaires. En effet, si un domaine ouvert D de $\mathbb{R}^2\backslash\{(x, 0) : x \geq 0\}$ est représenté en polaire par $\Delta \subset \mathbb{R}_+^* \times (0, 2\pi)$ alors pour toute fonction f intégrable sur D, on a

$$\int_D f(x, y)dxdy = \int_\Delta f(r\cos(\theta), r\sin(\theta))rdrd\theta. \tag{1.2}$$

On obtient ce résultat en appliquant le Théorème 1.22 à la fonction

$$\varphi : \left\{ \begin{array}{ccc} \Delta & \to & D \\ (r, \theta)^\top & \to & (r\cos(\theta), r\sin(\theta))^\top \end{array} \right.$$

On vérifie bien que φ est une bijection par définition des ensembles D et Δ (Δ est l'écriture en coordonnées polaires de D), qu'elle est $\mathcal{C}^1$ et que comme $\det J(\varphi)(r, \theta) = r(\cos^2(\theta) + \sin^2(\theta)) = r \neq 0$ où

$$J(\varphi)(r, \theta) = \begin{pmatrix} \cos(\theta) & -r\sin(\theta) \\ \sin(\theta) & r\cos(\theta) \end{pmatrix},$$

on a bien que φ est un $\mathcal{C}^1$-difféomorphisme ; on peut donc appliquer le Théorème 1.22. De manière informelle, on "pose" $(x,y)=\varphi(r,\theta)$ et on écrit $dxdy = |\det J(\varphi)(r,\theta)| r dr d\theta$.

Un exemple classique d'application de cette formule est de montrer que

$$\int_{\mathbb{R}} \exp(-x^2)dx = \sqrt{\pi}. \tag{1.3}$$

En effet, pour tout $a>0$, on note

$$D_a = \{(x,y)\in\mathbb{R}^2 : x^2+y^2 \leq a^2\} \text{ et } C_a = [-a,a]^2$$

ainsi que les intégrales associées

$$I_a = \int D_a \exp(-(x^2+y^2))dxdy \text{ et } J_a = \int C_a \exp(-(x^2+y^2))dxdy.$$

En passant en coordonnées polaires (càd en appliquant la formule (1.2)), on obtient

$$I_a = \left(\int_0^a \exp(-r^2) r dr\right)\left(\int_0^{2\pi} d\theta\right) = \pi(1-\exp(-a^2)).$$

Par ailleurs, on a

$$J_a = \left(\int_{-a}^a \exp(-x^2)dx)\right)\left(\int_{-a}^a \exp(-y^2)dy)\right) = \left(\int_{-a}^a \exp(-x^2)dx)\right)^2.$$

Or $D_a \subset C_a \subset D_{\sqrt{2}a}$ et la fonction a intégrée est positive donc $I_a \leq J_a \leq I_{\sqrt{2}a}$. Autrement dit, pour tout $a>0$,

$$\pi(1-\exp(-a^2)) \leq \left(\int_{-a}^a \exp(-x^2)dx)\right)^2 \leq \pi(1-\exp(-2a^2)).$$

En passant à la limite quand $a\to+\infty$, on obtient bien le résultat (1.3).

Le théorème de changement de variables (Théorème 1.22) permet aussi de retrouver le lien entre volume et déterminant (même si ce lien a été historiquement établi avant le théorème de changement de variable ; c'est d'ailleurs, ce lien qui permet d'intuiter le théorème de changement de variable, voir (1.4)). Soit $(x_1,\dots,x_n)$ une famille libre de $\mathbb{R}^n$. On note $A=[x_1|\cdots|x_n]$ la matrice ayant pour colonnes les vecteurs $x_1,\dots,x_n$. On note par Δ l'image du cube $[0,1]^n$ par A :

$$\Delta = \left\{\sum_{j=1}^n \lambda_j x_j : 0\leq\lambda_j\leq 1\right\} = A[0,1]^n.$$

Comme $(x_1,\dots,x_n)$ est une famille libre, $\det(A)\neq 0$ et donc $\varphi : u\in]0,1[^n \to Au\in\mathring{\Delta}$ est un $\mathcal{C}^1$-difféomorphisme dont la matrice jacobienne est donnée par A. On peut alors appliquer le théorème de changement de variables à la fonction $f\equiv 1$ pour obtenir

$$\mathrm{vol}(\Delta) = \int_{\mathring{\Delta}} dx = \int_{]0,1[^n} |\det(A)| d(v) = |\det(A)|.$$

On a donc bien retrouver que le volume du parallélépipède engendré par les vecteurs $x_1,\dots,x_n$ est donnée par la valeur absolue du déterminant de A. On peut d'ailleurs, retrouver intuitivement le Théorème 1.22 à partir de cette observation. En effet, pour appliquer le Théorème 1.22, on écrit

parfois de manière informelle $v=\varphi(u)$ et donc $dv=|\det(J(\varphi)(u))|du$. Cette dernière égalité peut être envisagée sur l'angle de la modification du volume 'infinitésimal' de $u+[0,du1]\times\cdots[0,du_n]$ par φ, localement en u pour $v=\varphi(u)$: $\varphi(u+[0,du1]\times\cdots\times[0,du_n])\sim\varphi(u)+J(\varphi)(u)[0,du1]\times\cdots\times[0,du_n]$. On a donc en approchant l'intégral par une somme infinie sur des pavés que

$$\begin{aligned}&\int_V f(v)dv\sim\sum_v f(v)dv\sim\sum_v f(v)\text{vol}\left(v+[0,dv1]\times\cdots\times[0,dv_n]\right)\\&\sim\sum_u f(\varphi(u))\text{vol}\left(\varphi\left(u+[0,du1]\times\cdots[0,du_n]\right)\right)\sim\sum_u f(\varphi(u))\text{vol}\left(J(\varphi)(u)[0,du1]\times\cdots[0,du_n]\right)\\&=\sum_u f(\varphi(u))|\det(J(\varphi)(u)|du\sim\int_U f(\varphi(u))|\det(J(\varphi)(u)|du\end{aligned}\tag{1.4}$$

où on a bien utilisé que le volume de $J(\varphi)(u)[0,du1]\times\cdots[0,du_n]$ est donné par $|\det(J(\varphi)(u)|du$ où on écrit que $[0,du1]\times\cdots[0,du_n]=[0,1]du$.

On fini cette section avec la chain rule pour la composition de fonction sur des e.v.n. de dimensions finies. On commence avec une fonction $\mathbb{R}^m\xrightarrow{f}\mathbb{R}^n\xrightarrow{g}\mathbb{R}$. On a vu la forme générale de la chain rule pour $E\xrightarrow{f}F\xrightarrow{g}G$ qui est $d(g\circ f)_a=dg_{f(a)}\circ df_a$, puis celle pour $H\xrightarrow{f}\mathbb{R}\xrightarrow{g}\mathbb{R}$ donnée par $\nabla(g\circ f)(a)=g'(f(a))\nabla f(a)$.

Proposition 1.23 (chain rule) *Soient V un ouvert de $\mathbb{R}^m$ et U un ouvert de $\mathbb{R}^n$. Soient $f:V\to U$ et $g:U\to\mathbb{R}$. Soit $a\in V$. On suppose que f est différentiable en a et g est différentiable en $f(a)$. On a alors*

$$\nabla(g\circ f)(a)=(J(f)(a))^\top\nabla g(f(a))$$

et pour tout $j\in\{1,\dots,m\}$,

$$\partial_j(g\circ f)(a)=\sum_{i=1}^n\partial_j f_i(a)\partial_i g(a)$$

où on note $f:x\in V\to(f_i(x))_{i=1}^n\in V$ les fonctions coordonnées de f.

Preuve. Par la chain rule, on a $d(g\circ f)_a=dg_{f(a)}\circ df_a$. Alors, pour tout $h\in\mathbb{R}^m$, on a

$$\begin{aligned}\left\langle\nabla(g\circ f)(a),h\right\rangle&=dg_{f(a)}(df_a(h))=\left\langle\nabla g(f(a)),df_a(h)\right\rangle=\left\langle\nabla g(f(a)),J(f)(a)h\right\rangle\\&=\left\langle J(f)(a)^\top\nabla g(f(a)),h\right\rangle.\end{aligned}$$

Pour la deuxième assertion, il suffit de voir que $\partial_j(g\circ f)(a)$ est la $j-$ième coordonnée de $\nabla(g\circ f)(a)$. On obtient donc la formule en identifiant la $j-$ième coordonnée du produit $(J(f)(a))^\top\nabla g(f(a))$. ■

La matrice Jacobienne permet d'écrire la chain rule de manière condensée et facile à retenir (on privilégiera donc cette formule aux autres). On écrit maintenant la chain rule 'version Jacobienne'. Ce résultat est plus général que celui des Proposition 1.9 et Proposition 1.23 puisse qu'il permet de les retrouver.

Proposition 1.24 (chain rule) *Soient V un ouvert de $\mathbb{R}^m$ et U un ouvert de $\mathbb{R}^n$. Soient $f:V\to U$ et $g:U\to\mathbb{R}^p$. Soit $a\in V$. On suppose que f est différentiable en a et g est différentiable en $f(a)$. On a alors*

$$J(g\circ f)(a)=J(g)(f(a))\times J(f)(a).$$

Preuve. Par la chain rule, on a $d(g\circ f)_a=dg_{f(a)}\circ df_a$. Alors, pour tout $h\in\mathbb{R}^m$, on a

$$J(g\circ f)(a)h=d(g\circ f)_a=dg_{f(a)}(df_a(h))=J(g)(f(a))df_a(h)=J(g)(f(a))J(f)(a)h$$

■

On retrouve bien la Proposition 1.23 à partir de la Proposition 1.24 car si g est à valeurs réelles alors $J(g)(f(a)) = (\nabla g(f(a)))^\top$ et $J(g \circ f)(a) = (\nabla(g \circ f)(a))^\top$. On peut donc se rappeler uniquement de la chain rule sous la forme suivante ainsi que la définition de la matrice Jacobienne :

Chain rule : $J(g \circ f)(a) = J(g)(f(a)) \times J(f)(a)$ et la matrice Jacobienne d'une fonction a pour vecteurs lignes les transposés des gradients des fonctions coordonnées : pour $f : x \in \mathbb{R}^n \to (f_i(x))_{i=1}^m \in \mathbb{R}^m$ et $a \in \mathbb{R}^n$

$$J(f)(a) = \begin{pmatrix} \nabla f_1(a)^\top \\ \vdots \\ \nabla f_m(a)^\top \end{pmatrix} \in \mathbb{R}^{m \times n}.$$

1.4 Accroissements finis et théorème fondamental de l'analyse

On a utilisé le théorème des inégalités des accroissements finis dans la preuve de la Proposition 1.18. On le rappelle maintenant.

Théorème 1.25 (Inégalité des accroissements finis) *Soient E et F deux e.v.n. et U un ouvert de E. Soit $f : U \to F$. Soit $a, b \in U$ tel que $[a, b] \subset U$. On suppose que*

1. *f est continue sur $[a, b]$*
2. *f est différentiable en tout point de (a, b)*
3. *il existe C tel que pour tout $x \in (a, b)$, $|||df_x||| \leq C$ où $||| \cdot |||$ est la norme d'opérateur de E dans F.*

Alors $\|f(a) - f(b)\|_F \leq C \|a - b\|_E$.

Attention, il ne faut pas confondre "inégalité des accroissements finis" et "égalité des accroissements finis". En effet, l'égalité des accroissements finis concerne une fonction à valeurs dans $\mathbb{R}$ et dit que si $f : U \to \mathbb{R}$ est différentiable alors pour tout $a, b \in U$, il existe $c \in (a, b)$ tel que $f(a) = f(b) + df_c(a - b)$. Ce résultat est faux si f est à valeurs dans un e.v.n. quelconque. On peut, par exemple, voir que $f(x) = (\cos(x), \sin(x))$ vérifie $f(0) = f(2\pi)$ mais pour tout $c \in (0, 2\pi), df_c(0 - 2\pi) \neq 0$.

L'inégalité des accroissement finis est un résultat classique du calcul différentielle. Il existe un autre résultat aussi classique qu'on appelle le théorème fondamentale de l'analyse qui s'énonce de la manière suivante. On rappelle avant qu'un **chemin** dans un ouvert U de E est une fonction γ continue sur un intervalle compact $[a, b]$ de $\mathbb{R}$ tel que $\gamma([a, b]) \subset U$. On dit que le chemin est $\mathcal{C}^1$ quand la restriction de γ à l'intervalle ouvert $]a, b[$ est $\mathcal{C}^1$. On rappelle aussi qu'un e.v.n. est complet quand toute suite de Cauchy converge (c'est par exemple, le cas des e.v. de dimension finie).

Théorème 1.26 *Soient E et F deux e.v.n. tel que F est complet et U un ouvert de E. Soit $f : U \to F$ de classe $\mathcal{C}^1$ et $\gamma : [a, b] \to U$ un chemin de classe $\mathcal{C}^1$ de U. On a*

$$f(\gamma(b)) - f(\gamma(a)) = \int_a^b df_{\gamma(t)}(\gamma'(t))dt$$

où on note par $\gamma'(t)$ l'unique élément de E tel que si $h \to 0$ alors $\gamma(t + h) = \gamma(t) + \gamma'(t)h + o(h)$.

En particulier, pour tout $x, y \in U$ quand $\gamma : t \in [0,1] \to (1-t)x + ty$, on a pour $f : U \to F$ de classe $\mathcal{C}^1$,

$$f(y) - f(x) = \int_0^1 df_{(1-t)x+ty}(y-x)dt.$$

2 Dérivées d'ordres supérieurs, Hessienne et Formules de Taylor

But : Dans la section précédente, on a seulement chercher à approcher f par une fonction affine. Cette approximation sera utile pour les conditions du premier ordre en optimisation. En se servant des dérivées d'ordres supérieurs, quand elles existent, on obtient plus d'information sur les points "critiques" obtenus par des conditions du premier ordre. Par exemple, en regardant la forme de la meilleure approximation quadratique, on pourra identifier si un point critique est un minimum ou un maximum (local).

Dans ce cours, on n'utilisera que des dérivées partielles au plus du second ordre. Néanmoins, on présente les définitions et résultats pour tous les ordres. On commence par rappeler les notations pour les dérivées partielles d'ordres supérieurs. L'idée est la suivante : si $f : U \to \mathbb{R}$ est différentiable sur U, un ouvert de $\mathbb{R}^n$, alors on a pour tout $a \in U$, quand $h \to 0$,

$$f(a+h) = f(a) + \langle \nabla f(a), h \rangle + o(\|h\|)$$

et $\nabla f(a) = (\partial_i f(a))_{i=1}^n$. Les dérivées partielles $\partial_i f$ sont les fonctions coordonnées du gradient. Elles sont aussi des fonctions de U dans $\mathbb{R}$, comme f. On peut alors voir si elles sont différentiables sur U. Si c'est le cas alors on a pour tout $a \in U$ et $h \to 0$,

$$(\partial_i f)(a+h) = (\partial_i f)(a) + \langle \nabla(\partial_i f)(a), h \rangle + o(\|h\|)$$

où $\nabla(\partial_i f)(a) = (\partial_j(\partial_i f)(a))$ est le gradient de $(\partial_i f)$ et $(\partial_j(\partial_i f))_{j=1}^n$ sont les dérivées partielles de $\partial_i f$. Sous réserve d'existence, on définit par récurrence une notion de dérivée partielle d'ordre p par la relation

$$\partial_{i_p}\partial_{i_{p-1}} \cdots \partial_{i_2}\partial_{i_1} f = \partial_{i_p}\partial_{i_{p-1}} \cdots \partial_{i_2}(\partial_{i_1} f) \tag{2.1}$$

pour tout $i_1, \ldots, i_p \in \{1, \ldots, n\}$.

Par exemple, si $x \to \partial_i f(x)$ existe dans un voisinage ouvert de $a \in U$ et que

$$t \in \mathbb{R}^* \to \frac{\partial_i f(a + te_j) - \partial_i f(a)}{t}$$

(où on note $(e_j)_{j=1}^n$ la base canonique de $\mathbb{R}^n$) admet une limite quand $t \to 0$ alors, cette limite est notée $\partial_j \partial_i f(a)$.

On peut définir des classes de fonctions selon l'existence et la continuité de ses dérivées partielles.

Définition 2.1 *Soit U un ouvert de $\mathbb{R}^n$ et $f : U \to \mathbb{R}$. On dit que f est de* ***classe*** *$\mathcal{C}^p$ si toutes ses dérivées partielles jusqu'à l'ordre p existent et sont continues sur U.*

Remarque 2.2 *Dans le chapitre précédent, on a vu que $f : U \to \mathbb{R}$ est une fonction de classe $\mathcal{C}^1$ quand sa différentielle $a \to df_a$ est continue de $\mathbb{R}^n$ dans $\mathcal{L}_c(\mathbb{R}^n, \mathbb{R})$. Ici on définit la notion de classe $\mathcal{C}^p$ uniquement à partir des dérivées partielles. La première définition est donc plus forte vue que*

$$|\partial_i f(a+h) - \partial_i f(a)| = |df_{a+h}(e_i) - df_a(e_i)| \leq \|df_{a+h} - df_a\|_{\mathbb{R}^n \to \mathbb{R}} \|h\|_2$$

où $\|\cdot\|_{\mathbb{R}^n\to\mathbb{R}}$ est la norme d'opérateur de $\mathbb{R}^n$ munit de la norme Euclidienne $\|\cdot\|_2$ dans $\mathbb{R}$ munit de la valeur absolue. Alors si df_a est continue en a, les dérivées partielles le sont aussi. La réciproque est aussi vraie ici en dimension finie, car on a

$$df_a(h) = \langle \nabla f(a), h\rangle = \sum_{j=1}^{n} h_j \partial_j f(a)$$

et donc

$$\|df_{a+h} - df_a\|_{\mathbb{R}^n\to\mathbb{R}} = \sup_{\|h\|_2=1} \Big| \sum_{j=1}^{n} h_j \big(\partial_j f(a+h) - \partial_j f(a)\big)\Big| = \sqrt{\sum_{j=1}^{n} \big(\partial_j f(a+h) - \partial_j f(a)\big)^2}.$$

Donc la continuité des dérivées partielles $\partial_j f(\cdot)$ en a implique celle de df en a.

Dans la définition (2.1), l'ordre de différentiation a de l'importance : dériver d'abord par rapport à la i-ième variable et ensuite par rapport à la j-ième variable n'est, a priori, pas la même chose que de dériver d'abord par rapport à la j-ième variable et ensuite par rapport à la i-ième variable. Cependant, sous une hypothèse de continuité des dérivées partielles, il se trouve que l'ordre de différentiation n'est en fait pas important. C'est le théorème de Schwarz rappelé maintenant.

Théorème 2.3 (Théorème de Schwarz) *Soient U un ouvert de $\mathbb{R}^2$ et $f : (x,y) \in U \to \mathbb{R}$ telle que f admet des dérivées partielles $\partial_x\partial_y f$ et $\partial_y\partial_x f$ sur U et qui sont continues sur U. Alors $\partial_x\partial_y f = \partial_y\partial_x f$ et dans ce cas, on notera $\partial^2_{xy} f = \partial_x\partial_y f = \partial_y\partial_x f$.*

En particulier, pour U un ouvert de $\mathbb{R}^n$ et $f : U \to \mathbb{R}$ de classe $\mathcal{C}^2$ on a pour tout $x \in U$ et tout $i,j \in \{1,\dots,n\}$ $\partial_i\partial_j f(x) = \partial_j\partial_i f(x)$. Ainsi, la matrice $\nabla^2 f(x) = (\partial^2_{ij} f(x))_{1\le i,j\le n}$ où $\partial^2_{ij} f(x) = \partial_i\partial_j f(x) = \partial_j\partial_i f(x)$ est symétrique. Cette matrice porte le nom de **Hessienne**. C'est l'objet central utilisé pour l'approximation au second ordre d'une fonction (de classe $\mathcal{C}^2$) localement. De même que le gradient est l'objet central pour son approximation au premier.

Définition 2.4 *Soit U un ouvert de $\mathbb{R}^n$ et $f : U \to \mathbb{R}$ une fonction de classe $\mathcal{C}^2$. Pour tout $x \in U$, la matrice $\nabla^2 f(x) = (\partial^2_{ij} f(x))_{1\le i,j\le n}$ est appelée la **Hessienne** de f en x.*

Sans la condition de continuité sur les dérivées partielles, le résultat du Théorème de Schwarz n'est plus vrai. Un contre-exemple est donné par

$$f := \left\{ \begin{array}{ccc|cc} \mathbb{R}^2 & \to & & \mathbb{R} & \\ (x,y) & \to & \Big| & xy\frac{x^2-y^2}{x^2+y^2} & \text{si } (x,y)\neq(0,0) \\ & & & 0 & \text{sinon.} \end{array} \right.$$

On montre que $\partial_x\partial_y f(0,0)$ et $\partial_y\partial_x f(0,0)$ existent mais sont différentes. Pour montrer que $\partial_x\partial_y f(0,0)$ existe, il est nécessaire et suffisant de montrer que $x \to \partial_y f(x,0)$ existe sur un voisinage ouvert de 0 et que $\lim_{x\to 0}(\partial_y f(x,0) - \partial_y f(0,0))/x$ existe. Dans ce cas, on aura

$$\partial_x\partial_y f(0,0) = \lim_{x\to 0, x\neq 0} \frac{\partial_y f(x,0) - \partial_y f(0,0)}{x}.$$

On doit donc calculer $\partial_y f(x,0)$ pour tout $x \neq 0$ dans un voisinnage de 0 et $\partial_y f(0,0)$.

On a pour tout $x, y \neq 0$, quand $y \to 0$,

$$\frac{f(x,y)-f(x,0)}{y} = x\frac{x^2-y^2}{x^2+y^2} \to x$$

donc $\partial_y f(x,0) = x$ et pour $x = 0$, on a pour tout $y \neq 0$, quand $y \to 0$,

$$\frac{f(0,y)-f(0,0)}{y} = 0 = x.$$

Donc pour tout $x, \partial_y f(x,0) = x$ et donc $\partial_x \partial_y f(0,0) = 1$. Par ailleurs, $f(y,x) = -f(x,y)$ donc $\partial_y \partial_x f(x,y) = -\partial_x \partial_y f(y,x)$ et ainsi $\partial_y \partial_x f(0,0) = -\partial_x \partial_y f(0,0) = -1$. ■

La notation "puissance symbolique n-ième" est donnée pour une fonction $f : U \subset \mathbb{R}^q \to \mathbb{R}^p$ de classe $\mathcal{C}^n$ par

$$\left[\sum_{i=1}^{q} h_i \partial_i f(a)\right]^{[n]} = \sum_{i_1+\cdots+i_q=n} \left(\frac{n!}{i_1!\cdots i_q!}\right) h_1^{i_1}\cdots h_q^{i_q} \frac{\partial^n f}{\partial^{i_1} x_1 \cdots \partial^{i_q} x_q}(a)$$

où

$$\frac{\partial^n f}{\partial^{i_1} x_1 \cdots \partial^{i_q} x_q}(a) = \left(\frac{\partial^n f_j}{\partial^{i_1} x_1 \cdots \partial^{i_q} x_q}(a)\right)_{j=1}^{p}.$$

Différencier en x une fonction f est équivalent à chercher la meilleure approximation affine de cette fonction en x. On peut aussi chercher la meilleure approximation quadratique, cubique, etc. Ces approximations sont données par les formules de Taylor avec différentes expression sur l'erreur d'approximation appelée *reste*.

Théorème 2.5 (Formule de Taylor) *Soit $f : U \subset \mathbb{R}^q \to \mathbb{R}^p$ de classe $\mathcal{C}^k$ sur U. Soit $x \in U$ et $h \in \mathbb{R}^q$. Alors*

$$f(x+h) = f(x) + \left[\sum_{i=1}^{q} h_i \partial_i f(x)\right]^{[1]} + \frac{1}{2!}\left[\sum_{i=1}^{q} h_i \partial_i f(x)\right]^{[2]} + \cdots + \frac{1}{(k-1)!}\left[\sum_{i=1}^{q} h_i \partial_i f(x)\right]^{[k-1]} + reste$$

où le terme de reste est donné par :

1. *Taylor-Lagrange pour $p = 1$: si $[x, x+h] \subset U$ il existe $\theta \in (0,1)$ tel que*

$$reste = \frac{1}{k!}\left[\sum_{i=1}^{q} h_i \partial_i f(x+\theta h)\right]^{[k]}$$

2. *Taylor avec reste intégral : si $[x, x+h] \subset U$ alors*

$$reste = \int_0^1 \frac{(1-t)^{k-1}}{(k-1)!}\left[\sum_{i=1}^{q} h_i \partial_i f(x+th)\right]^{[k]} dt$$

3. *Taylor-Young : quand $h \to 0$, $reste = o\left(\|h\|_2^k\right)$.*

Dans ce cours, on utilisera principalement la meilleure approximation quadratique d'une fonction à valeurs dans $\mathbb{R}$, c'est-à-dire la formule de Taylor à l'ordre 2 pour $p = 1$. Dans ce cas, le terme d'approximation linéaire fait apparaître le gradient et le terme du second ordre fait apparaître la **Hessienne** : pour $f : U \subset \mathbb{R}^n \to \mathbb{R}$ de classe $\mathcal{C}^2$ et $x \in U$, on a par la formule de Taylor-Young

$$f(x+h) = f(x) + \langle \nabla f(x), h \rangle + \frac{1}{2} h^\top \nabla^2 f(x) h + o(\|h\|^2) \tag{2.2}$$

où $\nabla f(x)$ est le gradient et $\nabla^2 f(x)$ la Hessienne de f en x définis par

$$\nabla f(x) = (\partial_i f(x))_{i=1}^n \text{ et } \nabla^2 f(x) = \left(\partial_{ij}^2 f(x)\right)_{i,j=1}^n .$$

Comme f est de classe $\mathcal{C}^2$, on peut appliquer le théorème de Schwarz et obtenir que la matrice Hessienne de f est symétrique (i.e. $\nabla^2 f(x) = (\nabla^2 f(x))^\top$).

On peut aussi écrire les deux autres formules de Taylor dans ce cadre. Soit U un ouvert de $\mathbb{R}^n$, $x \in U$ et $f : U \subset \mathbb{R}^n \to \mathbb{R}$ de classe $\mathcal{C}^2$. On a pour tout $h \in \mathbb{R}^n$ tel que $[x, x+h] \subset U$, d'après la formule de Taylor-Lagrange qu'il existe $\theta \in (0,1)$ tel que

$$f(x+h) = f(x) + \langle \nabla f(x), h \rangle + \frac{1}{2} h^\top \nabla^2 f(x+\theta h) h$$

et d'après la formule de Taylor avec reste intégrale

$$f(x+h) = f(x) + \langle \nabla f(x), h \rangle + \int_0^1 (1-t) h^\top \nabla^2 f(x+th) h dt.$$

Les formules de Taylor-Lagrange et Taylor avec reste intégrale ont l'avantage d'être exactes et sont parfois utiles pour démontrer des résultats. Cependant c'est la formule de Taylor-Young qui donne du sens aux objets : c'est le gradient et la Hessienne qui donnent la meilleure approximation quadratique d'une fonction $\mathcal{C}^2$ en un point. C'est cette formule qu'il est bon d'avoir en tête lorsqu'on manipule ces deux objets.

Exemple 2.6 *On peut par exemple chercher le gradient et la Hessienne dans un cas trivial d'une fonction quadratique. Dans ce cas, Taylor d'ordre deux est exacte (la meilleure approximation quadratique de f étant f elle-même, étant donné que f est quadratique). On a alors pour $f(x) = \|Ax - y\|_2^2$ que*

$$f(x+h) = f(x) + \langle \nabla f(x), h \rangle + \frac{1}{2} h^\top \nabla^2 f(x) h$$

où $\nabla f(x) = 2A^\top(Ax - y)$ et $\nabla^2 f(x) = 2A^\top A$.

Au passage, on voit que la Hessienne de f en x est la Jacobienne de $x \to \nabla f(x)$ en x : $J(\nabla f)(x) = \nabla^2 f(x) = 2A^\top A$ car $\nabla f(x+h) - \nabla f(x) = 2A^\top Ah = J(\nabla f(x))h$ et donc $J(\nabla f)(x) = 2A^\top A$. C'est en fait un résultat général qu'on énonce dans le résultat suivant.

Proposition 2.7 *Soit U un ouvert de $\mathbb{R}^n$ et $f : U \to \mathbb{R}$ de classe $\mathcal{C}^2$. On a $J(\nabla f) = \nabla^2 f$ (càd la matrice Jacobienne du gradient est la Hessienne).*

Preuve. Comme f est $\mathcal{C}^2$, $\nabla f : U \to \mathbb{R}^n$ est $\mathcal{C}^1$ alors sa matrice Jacobienne a pour entrées les dérivées partielles de ses fonctions coordonnées. Or $\nabla f = (\partial_i f)_{i=1}^n$ alors ces fonctions coordonnées sont les $\partial_i f$ pour $i = 1, \ldots, n$. On a donc $J(\nabla f) = (\partial_j \partial_i f)_{1 \le i,j \le n}$. On conclut avec le théorème de Schwarz et la définition de la Hessienne. ∎

Exemple 2.8 *Soit $f : \mathbb{R}^2 \to \mathbb{R}^2$ de classe $\mathcal{C}^2$. Le Laplacien de f est $\Delta f = \partial_x^2 f + \partial_y^2 f$. On veut écrire le Laplacien de f en coordonnées polaires. Soit*

$$\varphi : \left\{ \begin{array}{ccc} \mathbb{R}_+^* \times (0, 2\pi) & \to & \mathbb{R}^2 \backslash \{(x, 0) : x \geq 0\} \\ (r, \theta)^\top & \to & (r\cos(\theta), r\sin(\theta))^\top \end{array} \right.$$

et $F = f \circ \varphi$. Montrer que pour tout $(r, \theta) \in \mathbb{R}_+^* \times (0, 2\pi)$,

$$\Delta f(r\cos\theta, \theta\sin\theta) = \partial_r^2 F(r, \theta) + r^{-1}\partial_r F(r, \theta) + r^{-2}\partial_\theta^2 F(r, \theta).$$

On a d'après la chain rule que $\nabla F(r, \theta) = \nabla(f \circ \varphi)(r, \theta) = (J(\varphi)(r, \theta))^\top \nabla f(\varphi(r, \theta))$ et donc

$$\left(\begin{array}{c} \partial_r F(r, \theta) \\ \partial_\theta F(r, \theta) \end{array} \right) = \nabla F(r, \theta) = \left(\begin{array}{c} \cos\theta\partial_x f(r\cos\theta, \theta\sin\theta) + \sin\theta\partial_y f(r\cos\theta, \theta\sin\theta) \\ -r\sin\theta\partial_x f(r\cos\theta, \theta\sin\theta) + r\cos\theta\partial_y f(r\cos\theta, \theta\sin\theta) \end{array} \right).$$

On en déduit que

$$(\partial_x f) \circ \varphi(r, \theta) = \partial_x f(r\cos\theta, \theta\sin\theta) = \cos\theta\partial_r F(r, \theta) - \frac{\sin\theta}{r}\partial_\theta F(r, \theta) := g_1(r, \theta)$$

et

$$(\partial_y f) \circ \varphi(r, \theta) = \partial_y f(r\cos\theta, \theta\sin\theta) = \sin\theta\partial_r F(r, \theta) + \frac{\cos\theta}{r}\partial_\theta F(r, \theta) := g_2(r, \theta).$$

où on voit $\partial_x f : \mathbb{R}^2 \to \mathbb{R}$ et de même $\partial_y f : \mathbb{R}^2 \to \mathbb{R}$.

En utilisant que $g_1 = (\partial_x f) \circ \varphi$, la chain rule nous donne que

$$\nabla g_1(r, \theta) = (J(\varphi)(r, \theta))^\top \nabla(\partial_x f)(\varphi(r, \theta)).$$

De même en utilisant que $g_2 = (\partial_y f) \circ \varphi$, la chain rule nous donne que

$$\nabla g_2(r, \theta) = (J(\varphi)(r, \theta))^\top \nabla(\partial_y f)(\varphi(r, \theta)).$$

On en déduit que

$$\left(\begin{array}{c} \partial_r g_1(r, \theta) \\ \partial_\theta g_1(r, \theta) \end{array} \right) = \nabla g_1(r, \theta) = \left(\begin{array}{c} \cos\theta\partial_x^2 f(r\cos\theta, \theta\sin\theta) + \sin\theta\partial_{xy}^2 f(r\cos\theta, \theta\sin\theta) \\ -r\sin\theta\partial_x^2 f(r\cos\theta, \theta\sin\theta) + r\cos\theta\partial_{xy} f(r\cos\theta, \theta\sin\theta) \end{array} \right)$$

et

$$\left(\begin{array}{c} \partial_r g_2(r, \theta) \\ \partial_\theta g_2(r, \theta) \end{array} \right) = \nabla g_2(r, \theta) = \left(\begin{array}{c} \cos\theta\partial_{xy}^2 f(r\cos\theta, \theta\sin\theta) + \sin\theta\partial_y^2 f(r\cos\theta, \theta\sin\theta) \\ -r\sin\theta\partial_{xy}^2 f(r\cos\theta, \theta\sin\theta) + r\cos\theta\partial_y f(r\cos\theta, \theta\sin\theta) \end{array} \right)$$

En même temps, on sait que $g_1(r, \theta) = \cos\theta\partial_r F(r, \theta) - \frac{\sin\theta}{r}\partial_\theta F(r, \theta)$. On a donc

$$\nabla g_1(r, \theta) = \left(\begin{array}{c} \cos\theta\partial_r^2 F(r, \theta) + \frac{\sin\theta}{r^2}\partial_\theta F(r, \theta) - \frac{\sin\theta}{r}\partial_{r\theta}^2 F(r, \theta) \\ -\sin\theta\partial_r F(r, \theta) + \cos\theta\partial_{r\theta}^2 F(r, \theta) - \frac{\cos\theta}{r}\partial_\theta F(r, \theta) - \frac{\sin\theta}{r}\partial_\theta^2 F(r, \theta) \end{array} \right).$$

De même on a $g_2(r, \theta) = \sin\theta\partial_r F(r, \theta) + \frac{\cos\theta}{r}\partial_\theta F(r, \theta)$. On en déduit donc que

$$\nabla g_2(r, \theta) = \left(\begin{array}{c} \sin\theta\partial_r^2 F(r, \theta) - \frac{\cos\theta}{r^2}\partial_\theta F(r, \theta) + \frac{\cos\theta}{r}\partial_{r\theta}^2 F(r, \theta) \\ \cos\theta\partial_r F(r, \theta) + \sin\theta\partial_{r\theta}^2 F(r, \theta) - \frac{\sin\theta}{r}\partial_\theta F(r, \theta) + \frac{\cos\theta}{r}\partial_\theta^2 F(r, \theta) \end{array} \right).$$

On en déduit le résultat en calculant

$$\Delta f(r\cos\theta, \theta\sin\theta) = \partial_x^2 f(r\cos\theta, \theta\sin\theta) + \partial_y^2 f(r\cos\theta, \theta\sin\theta).$$

3 Représentation géométrique du gradient

La plupart des théorèmes en optimisation ont une interprétation géométrique qu'il est bon de connaître pour pouvoir les retrouver (et les comprendre). Une des premières choses qu'il faut savoir bien visualiser est le gradient d'une fonction. Pour cela, on introduit quelques notions de géométrie. On commence par rappeler la notion de lignes de niveau d'une fonction (déjà vue au premier chapitre) et d'ensemble de niveau.

Définition 3.1 *Soit $f : U \to \mathbb{R}$. Soit $\alpha \in \mathbb{R}$. La **ligne de niveau** α **de** f est*

$$\mathcal{L}_f(\alpha) = \{x \in U : f(x) = \alpha\}.$$

*L'**ensemble de niveau** α **de** f est*

$$\mathcal{L}_f(\leq \alpha) = \{x \in U : f(x) \leq \alpha\}.$$

Les courbes de niveaux d'une fonction qu'on cherche à minimiser (càd une fonction objectif) sont des bons "repères géométriques" car c'est le long de ces surfaces que le critère a minimiser reste constant. Comme indiqué dans la dernière section du premier chapitre, on préférera représenter géométriquement une fonction $f : \mathbb{R}^2 \to \mathbb{R}$ par plusieurs de ces lignes de niveau dans $\mathbb{R}^2$ plutôt que par un graphique en $3D$ dans $\mathbb{R}^3$.

On donne maintenant quelques outils de géométrie différentielle qui permettent de représenter le gradient d'une fonction en fonction de ces lignes et ensembles de niveau. On retrouvera ces outils plus tard lorsqu'on cherchera à décrire localement un ensemble par un cône.

Définition 3.2 *Soit S un sous-ensemble de $\mathbb{R}^n$ non vide. Soit $x \in S$ et $v \in \mathbb{R}^p$. On dit que v est un **vecteur tangent à** S **en** x quand il existe deux suites $(\lambda_k)_k \subset \mathbb{R}_+^*$ et $(v_k)_k \subset \mathbb{R}^n$ telles que $(\lambda_k) \downarrow 0$, $x + \lambda_k v_k \in S$ et $v = \lim_k v_k$.*

*On dit qu'un vecteur $v^\perp$ est **orthogonal à** S **en** x quand pour tout vecteur v tangent à S en x on a $\langle v, v^\perp \rangle = 0$. On dit qu'un vecteur v^* est **normal à** S **en** x quand pour tout vecteur v tangent à S en x on a $\langle v, v^* \rangle \leq 0$.*

Par exemple, quand $S = \mathbb{R}^p$ alors tous les éléments de $\mathbb{R}^p$ sont des vecteurs tangents de S en chacun de ces points. Plus généralement, si S est un espace affine alors l'ensemble de ses vecteurs tangents est donné par $(S - x) = \{s - x : s \in S\}$ pour n'importe quel point $x \in S$. En effet, il suffit de prendre $v \in (S-x)$, $\lambda_k = 1/k$ et $v_k = v$ pour tout k, on a bien $(\lambda_k) \downarrow 0$, $x + \lambda_k v_k \in S$ car $S - x$ est un espace linéaire et $v_k \to v$. Réciproquement, si v est tangent à S en x alors comme $v_k \in S - x$ (car $S - x$ est positivement homogène càd si $u \in (S - x)$ et $\lambda > 0$ alors $\lambda x \in S - x$) et que $S - x$ est fermé, on a aussi sa limite qui est dans $S - x$.

On peut voir les espaces tangents comme des approximations du premier ordre des ensembles. En effet, si $x \in S$ alors la meilleure approximation d'ordre 0 de S en x est donné par x (c'est comme pour une fonction f, sa meilleure approximation en x à l'ordre 0 est donné par $f(x)$). La meilleure approximation affine de S en x est donnée par $x + T_S(x)$ où $T_S(x)$ est l'ensemble des vecteurs tangents à S en x. On cherche à approcher localement l'ensemble S en x par un ensemble de la forme $x +$ *un cône* (un cône étant un ensemble positivement homogène càd si v est dans le cône alors λv pour $\lambda \geq 0$ est aussi dans le cône). En particulier, quand S est un espace affine ou un demi-espace affine et que x est sur le bord de S, alors sa meilleure approximation affine c'est lui-même car $S = x + (S - x)$ et $S - x$ est un espace linéaire (donc en particulier c'est un cône). En particulier, dans ce cas l'approximation est exacte vue qu'on a égalité entre S et son approximation $x + (S - x)$ et c'est une relation globale car elle est vraie même loin de x (l'endroit où on veut approcher S). Donc l'ensemble des vecteurs tangents à S en x (un élément du bord de S) est donné par $S - x$.

Proposition 3.3 *Soit $f : U \to \mathbb{R}$ où U est un ouvert de $\mathbb{R}^n$. Soit $x \in U$. On suppose que f est différentiable en x. Alors $\nabla f(x)$ est orthogonal à la courbe de niveau $f(x)$ de f, càd à $\mathcal{L}_f(f(x))$. Le gradient $\nabla f(x)$ est normal à l'ensemble de niveau $f(x)$ de f, càd à $\mathcal{L}_f(\leq f(x))$.*

Preuve. Soit v un vecteur tangent à $\mathcal{L}_f(f(x))$ en x. Montrons que $\langle v, \nabla f(x)\rangle = 0$. Par définition, il existe deux suites $(\lambda_k)_k \subset \mathbb{R}_+^*$ et $(v_k)_k \subset \mathbb{R}^n$ telles que $(\lambda_k) \downarrow 0$, $x + \lambda_k v_k \in \mathcal{L}_f(f(x))$ et $v = \lim_k v_k$. On a alors

$$\langle v, \nabla f(x)\rangle = \lim_{k\to+\infty} \langle v_k, \nabla f(x)\rangle.$$

Par ailleurs, comme $v_k \to v$, $(v_k)_k$ est une suite bornée et comme $(\lambda_k) \downarrow 0$, on a que $\lambda_k v_k \to 0$. Ainsi, quand $k \to +\infty$,

$$f(x + \lambda_k v_k) = f(x) + \langle \lambda_k v_k, \nabla f(x)\rangle + o(\lambda_k)$$

et comme $x + \lambda_k v_k \in \mathcal{L}_f(f(x))$, on a $f(x + \lambda_k v_k) = f(x)$. Alors

$$\langle \lambda_k v_k, \nabla f(x)\rangle = o(\lambda_k)$$

autrement dit $\lim_k \langle v_k, \nabla f(x)\rangle = 0$.

Pour la normalité du gradient de f en x à l'ensemble de niveau $f(x)$ de f, on procède comme précédemment sauf qu'on a seulement $f(x + \lambda_k v_k) \leq f(x)$ (au lieu d'avoir l'égalité comme précédemment). ∎

Par ailleurs, pour un problème d'optimisation, une bonne façon de voir le gradient de f en x est de se placer en x et de chercher la direction de plus forte pente, càd la direction qu'il faut prendre partant de x pour faire croître le plus f. Il se trouve que c'est le gradient (renormalisé càd $\nabla f(x)/\|\nabla f(x)\|_2$) de f en ce point qui indique la direction de plus forte pente. En effet, si $f : U \to \mathbb{R}$ est différentiable en $x \in U$ (où U est un ouvert de $\mathbb{R}^n$) alors la direction $v \in S_2^{n-1}$ de plus forte pente est celle qui maximise

$$\lim_{t\to 0} \frac{f(x + tv) - f(x)}{t}.$$

Or, on a vu que cette quantité est la dérivée partielle de f en x dans la direction v et qu'elle peut s'écrire en fonction du gradient.

$$\lim_{t\to 0} \frac{f(x + tv) - f(x)}{t} = \partial_v f(x) = \langle \nabla f(x), v\rangle.$$

Par ailleurs, $\max_{v\in S_2^{n-1}} \langle \nabla f(x), v\rangle$ est atteint en $\nabla f(x)/\|\nabla f(x)\|_2$. Il faut donc bien aller dans la direction du gradient $\nabla f(x)$ pour accroître le plus la valeur de f à partir de x.

Conclusion : Le gradient d'une fonction f en un point x est orthogonal à la courbe de niveau $f(x)$ de f et normal à l'ensemble de niveau $f(x)$ de f. C'est aussi la direction de plus forte pente de f partant de ce point x.

Un dernier mot sur l'approximation d'une fonction par une fonction affine. On a vu que si $f : U \to \mathbb{R}$ est différentiable en $x^* \in U$ alors, la meilleure approximation affine de f en x^* est donnée par

$$F_{x^*} : x \in \mathbb{R}^n \to f(x^*) + \langle \nabla f(x^*), x - x^*\rangle.$$

On a vu au cours précédent que la ligne de niveau $f(x^*)$ de F_{x^*} est donnée par

$$\mathcal{L}_{F_{x^*}}(f(x^*)) = \{x \in \mathbb{R}^n : \langle \nabla f(x^*), x - x^* \rangle = 0\} = x^* + \text{vect}(\nabla f(x^*))^\perp,$$

c'est donc l'espace affine passant par x^* et de vecteur normal $\nabla f(x^*)$. Donc $\nabla f(x^*)$ est aussi un vecteur orthogonal (et donc aussi normal) à la courbe de niveau $f(x^*)$ de F_{x^*}, la meilleure approximation affine de f en x^*.

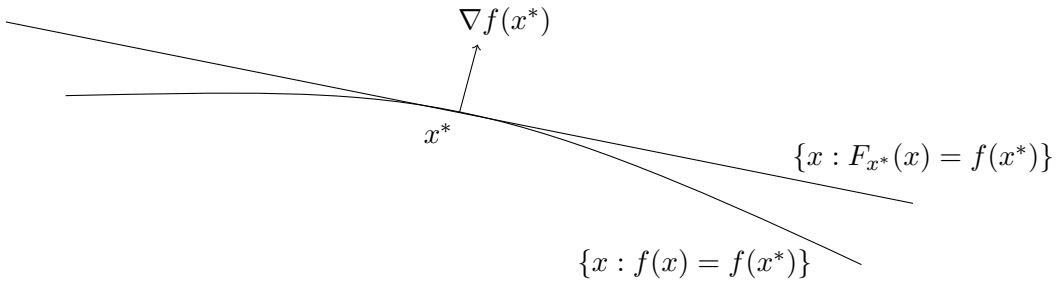

FIGURE 1 – Le gradient d'une fonction f en un point x^* est orthogonal à la courbe de niveau $f(x^*)$ de f et à la courbe de niveau $f(x^*)$ de la meilleure approximation affine de f en x^*. C'est de plus la direction de plus forte croissance de f partant du point x^*.

On peut alors représenter une fonction de $\mathbb{R}^2 \to \mathbb{R}$ par quelques-unes de ses lignes de niveaux et quelques-uns de ces gradients. Sous python, on utilise la méthode `contourf` pour tracer les lignes de niveau et la méthode `quiver` pour le champ de vecteur gradient. On rappelle qu'un champs de vecteurs sur $\mathbb{R}^n$ est une application de $\mathbb{R}^n$ dans lui-même. Dans ce cas, il est bon de voir l'espace de départ $\mathbb{R}^n$ comme un ensemble de points alors que l'espace d'arrivée $\mathbb{R}^n$ est plutôt vu comme un ensemble de vecteurs. Le gradient d'une fonction $f : \mathbb{R}^n \to \mathbb{R}$ est bien un champs de vecteurs en tant qu'application de $\mathbb{R}^n$ dans $\mathbb{R}^n$ (un exemple est donné dans la Figure 2).

```
from matplotlib.pyplot import cm
import numpy as np
import matplotlib.pyplot as plt

# Contour Plot
X, Y = np.mgrid[-2:2:100j, -2:2:100j]
Z = X*np.exp(-(X**2 + Y**2))
cp = plt.contourf(X, Y, Z)
cb = plt.colorbar(cp)

# Vector Field
Y, X = np.mgrid[-2:2:20j, -2:2:20j]
U =(1 - 2*(X**2))*np.exp(-((X**2)+(Y**2)))
V = -2*X*Y*np.exp(-((X**2)+(Y**2)))
speed = np.sqrt(U**2 + V**2)
UN = U/speed
VN = V/speed
quiv = plt.quiver(X, Y, UN, VN, color='Teal', headlength=7)
```

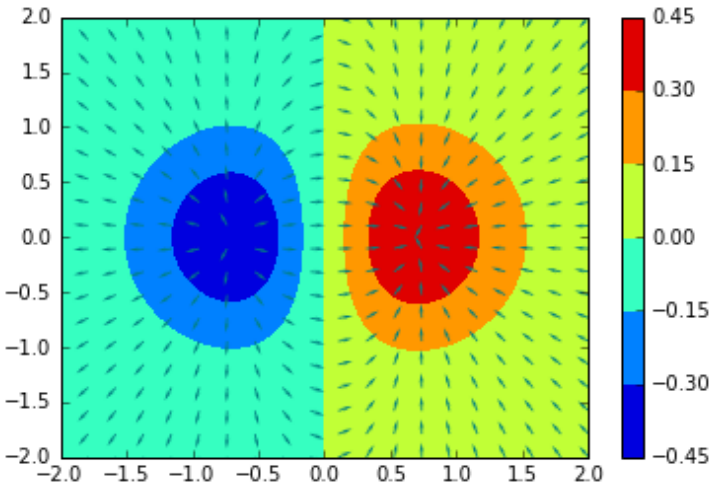

FIGURE 2 – Représentation d'une fonction $f : \mathbb{R}^2 \to \mathbb{R}$ par quelques unes de ses lignes de niveau et de ses gradients.