

Approximations locales d'une fonction différentiable de plusieurs variables

Guillaume Lecué¹

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. 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. Quand la fonction a plus de régularité, on peut donner des approximations 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 vectoriels réels normés (e.v.n.). Soit U un ouvert de E et $f : U \rightarrow F$ 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 $h \rightarrow 0$,

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

Soit $g : E \rightarrow F$. On rappelle " $g(h) = o(\|h\|_E)$ quand $h \rightarrow 0$ " signifie que $g(h)/\|h\|_E \rightarrow 0$ quand $h \rightarrow 0$. Autrement dit $g(h)/\|h\|_E = o(1)$ quand $h \rightarrow 0$. 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 $h \rightarrow 0$. En particulier pour tout $x \in E$, on aurait $df'_a(\lambda x) - df_a(\lambda x) = o(\|\lambda x\|_E)$ quand $\lambda \rightarrow 0$ et par linéarité, $df'_a(x) - df_a(x) = o(1)$ quand $\lambda \rightarrow 0$, donc $df'_a(x) = df_a(x)$).

2. Si f est différentiable en a alors elle est continue en a (en effet, $df_a(h) \rightarrow 0$ quand $h \rightarrow 0$, car df_a est continue, ainsi $f(a + h) \rightarrow f(a)$ quand $h \rightarrow 0$).

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 $f : \mathbb{R} \rightarrow \mathbb{R}$, 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 $df_a(h) = f'(a)h$.

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

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 $T : (E, \|\cdot\|_E) \rightarrow (F, \|\cdot\|_F)$ 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 $x, y \in E$, $\|Tx - Ty\|_F = \|T(x - y)\|_F \leq L \|x - y\|_E$. 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 $x = (x_i)_{i=1}^n \in \mathbb{R}^n$

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

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

Définition 1.3 Soient $(E, \|\cdot\|_E)$ et $(F, \|\cdot\|_F)$ deux espaces vectoriels réels normés (e.v.n.). Soit U un ouvert de E et $f : U \rightarrow F$ 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,

$$df : \begin{cases} U & \rightarrow \mathcal{L}_c(E, F) \\ a & \rightarrow df_a \end{cases}$$

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 $\|T\| = \|T\|_{E \rightarrow F} = \sup_{x \neq 0} \|Tx\|_F / \|x\|_E$. 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 $\|RT\|_{E \rightarrow G} \leq \|R\|_{F \rightarrow G} \|T\|_{E \rightarrow F}$.

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

$$f : \begin{cases} \mathbb{R}^{n \times n} \times \mathbb{R}^{n \times n} & \rightarrow \mathbb{R}^{n \times n} \\ (A, B) & \rightarrow AB. \end{cases}$$

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 $(H, K) \rightarrow 0$ 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 $(H, K) \rightarrow 0$. Donc f est différentiable et $df_{(A,B)} = \varphi_{(A,B)}$. Par ailleurs, la différentielle de f , $df : (A, B) \rightarrow df_{(A,B)}$ 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 C^1 . En fait, f est une fonction polynomiale, elle est donc C^∞ .

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

$$f : \begin{cases} \mathbb{R}^n & \rightarrow \mathbb{R} \\ x & \rightarrow \|y - Ax\|_2^2. \end{cases}$$

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 $h \rightarrow 0$ 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 $A : \ell_2^n \rightarrow \ell_2^n$. Donc la différentielle de f est $df : x \in \mathbb{R}^n \rightarrow (df_x : h \rightarrow \langle -2A^\top(y - Ax), h \rangle)$.

Quand $f : U \rightarrow \mathbb{R}$ (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 $f : U \rightarrow \mathbb{R}$ 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, 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 $f, g : U \rightarrow F$ 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 $f : U \rightarrow V$ et $g : V \rightarrow G$. On suppose que f est différentiable en a et g en $f(a)$ alors $g \circ f : U \rightarrow G$ est différentiable en a et

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

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

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

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

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

Ainsi,

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

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

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

■

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 $df_x(h) = dg_{k(x)}(dk_x(h)) = dg_{Ax-y}(Ah) = \langle 2(Ax - y), Ah \rangle = \langle 2A^\top(Ax - y), h \rangle$.

Exemple 1.9 La chain rule se transcrit en terme de gradient. Si $f : H \rightarrow \mathbb{R}$ et $g : \mathbb{R} \rightarrow \mathbb{R}$ alors pour tout $x \in H$,

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

En effet, 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 $f, g : U \rightarrow \mathbb{R}$ deux fonctions différentiables en $a \in U$. 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 $\nabla(fg)(a) = \nabla f(a)g(a) + f(a)\nabla g(a)$.

1.2 Dérivées partielles

Définition 1.11 Soient E, F deux e.v.n., U un ouvert de E , $f : U \rightarrow F$, $a \in U$ et $v \in E$. Si la fonction $t \in \mathbb{R} \rightarrow f(a + tv)$ est différentiable en 0, on dit que f **admet une dérivée partielle en a selon v** et on note

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

Proposition 1.12 Soit U un ouvert de E et $f : U \rightarrow F$. On suppose 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 $t \rightarrow 0$, on a $f(a + tv) = f(a) + df_a(tv) + o(\|tv\|_E)$. Alors, quand $t \neq 0$ et $t \rightarrow 0$,

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

Autrement dit,

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

■

Exemple 1.13 Cependant, la réciproque à la Proposition 1.12 est fausse. On peut le voir sur l'exemple

$$f := \begin{cases} \mathbb{R}^2 & \rightarrow \mathbb{R} \\ (x, y) & \rightarrow \begin{cases} y^2/x & \text{si } x \neq 0 \\ y & \text{si } x = 0 \end{cases} \end{cases}$$

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 $f(0,0)$.

Soient $(u, v) \in \mathbb{R}^2$ et $t \neq 0$. On a

$$\frac{f((0,0) + t(u, v)) - f(0,0)}{t} = \frac{f(tu, tv)}{t} = \begin{cases} (tv)^2/(t^2u) & \text{si } tu \neq 0 \\ tv/t & \text{si } tu = 0 \end{cases} = \begin{cases} v^2/u & \text{si } u \neq 0 \\ v & \text{si } u = 0. \end{cases}$$

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

$$\partial_{(u,v)} f(0,0) = \begin{cases} v^2/u & \text{si } u \neq 0 \\ v & \text{si } u = 0. \end{cases}$$

Notation : Quand $f : U \rightarrow F$ 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

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

Proposition 1.14 Soit $f : U \rightarrow \mathbb{R}$ 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 1.12, 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 . 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 1.13).

Exemple 1.15 Soient $a_1, \dots, a_m \in \mathbb{R}^n$ et $b_1, \dots, b_m \in \mathbb{R}$. Calculer le gradient de

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

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

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

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

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

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

Dans l'Exemple 1.13, on a vu qu'une fonction peut admettre des dérivée partielle dans toutes les directions sans pour autant être différentiable. Cependant, on supposant l'existence ET la continuité de toutes les dérivées partielles d'ordre i , la réciproque est vraie.

Proposition 1.16 Soit $f : U \rightarrow \mathbb{R}$ où U est un ouvert de \mathbb{R}^n . Si toutes les dérivées partielles de f existent sur U et sont continues 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 $h \rightarrow 0$, 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 $x \rightarrow a$ 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

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

On rappelle 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

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

de sorte que $y_0 = a$ et $y_n = x$, ainsi y_0, y_1, \dots, 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 $k = 1, \dots, n$,

$$g_k : \begin{cases} [a_k, x_k] & \rightarrow F \\ t & \rightarrow g(x_1, \dots, x_{k-1}, t, a_{k+1}, \dots, a_n) \end{cases}$$

On a pour tout $t \in (a_k, x_k)$, $g'_k(t) = \partial_k g(x_1, \dots, x_{k-1}, t, a_{k+1}, \dots, a_n)$ et donc, d'après (1.1), $\|g'_k(t)\|_F \leq \epsilon$. Le théorème de l'inégalité 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 $k = 1, \dots, n$ 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 $x \rightarrow a$. ■

En conclusion, si f admet des dérivées partielles continues alors f est dérivable. La réciproque est fautive et peut se voir sur l'exemple suivant :

$$f : \begin{cases} \mathbb{R} & \rightarrow & \mathbb{R} \\ x & \rightarrow & \begin{cases} x^2 \sin(1/x) & \text{si } x \neq 0 \\ 0 & \text{si } x = 0. \end{cases} \end{cases}$$

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

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

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

On a utilisé de manière fondamentale le théorème des inégalités des accroissements finis dans la preuve de la Proposition 1.16. On le rappelle maintenant.

Théorème 1.17 (Inégalité des accroissements finis) Soient E et F deux e.v.n. et U un ouvert de E . Soit $f : U \rightarrow F$. 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 $f : U \rightarrow \mathbb{R}$ 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 $c \in (0, 2\pi)$, $df_c(0 - 2\pi) \neq 0$.

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 la généraliser cette notion à des fonctions à valeurs dans \mathbb{R}^m , c'est l'idée de la matrice Jacobienne.

Définition 1.18 Soit U un ouvert de \mathbb{R}^n et $f : U \rightarrow \mathbb{R}^m$ 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 $i = 1, \dots, m$, $f_i : U \rightarrow \mathbb{R}$ 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

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

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 1.19 1. Quand $f : \mathbb{R}^n \rightarrow \mathbb{R}$ est différentiable en a , on a $df_a(h) = \langle \nabla f(a), h \rangle$ et si $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ est différentiable en a , on a $df_a(h) = (\langle \nabla f_i(a), h \rangle)_{i=1}^m$.
2. Si $f : \mathbb{R}^n \rightarrow \mathbb{R}$, $J(f)(a) = (\nabla f(a))^\top$.

Le Jacobien joue un rôle essentiel dans le théorème de changement de base. Avant de rappeler ce résultat, on rappelle, que $\varphi : U \rightarrow V$ 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 1.20 Soient U et V deux ouverts de \mathbb{R}^n et $\varphi : U \rightarrow V$ un \mathcal{C}^1 -difféomorphisme de U sur V . Soit $f : V \rightarrow \mathbb{R}$ 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 $\mathbb{R}^2 \setminus \{(x, 0) : x \geq 0\}$ est représenté en polaire par $\Delta \subset \mathbb{R}_+^* \times (0, 2\pi)$ alors pour toute fonction f intégrable sur D , on a

$$\int_D f(x, y)dx dy = \int_\Delta f(r \cos(\theta), r \sin(\theta))r dr d\theta.$$

On obtient ce résultat en appliquant le Théorème 1.20 à la fonction

$$\varphi : \begin{cases} \Delta & \rightarrow D \\ (r, \theta)^\top & \rightarrow (r \cos(\theta), r \sin(\theta))^\top \end{cases}$$

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 1.20. De manière informelle, on “pose” $(x, y) = \varphi(r, \theta)$ et on écrit $dx dy = |\det J(\varphi)(r, \theta)| r dr d\theta$.

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

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

On peut trouver la preuve en page 329 de [Gourdon, Analyse].

Proposition 1.21 (chain rule) *Soient V un ouvert de \mathbb{R}^m et U un ouvert de \mathbb{R}^n . Soient $\varphi : V \rightarrow U$ et $f : U \rightarrow \mathbb{R}$. Soit $a \in V$. On suppose que φ est différentiable en a et f est différentiable en $\varphi(a)$. On a alors*

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

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

$$\begin{aligned} \langle \nabla(f \circ \varphi)(a), h \rangle &= df_{\varphi(a)}(d\varphi_a(h)) = \langle \nabla f(\varphi(a)), d\varphi_a(h) \rangle = \langle \nabla f(\varphi(a)), J(\varphi)(a)h \rangle \\ &= \langle J(\varphi)(a)^\top \nabla f(\varphi(a)), h \rangle. \end{aligned}$$

■

Exemple 1.22 *Soit $f : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ 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*

$$\varphi : \begin{cases} \mathbb{R}_+^* \times (0, 2\pi) & \rightarrow \mathbb{R}^2 \setminus \{(x, 0) : x \geq 0\} \\ (r, \theta)^\top & \rightarrow (r \cos(\theta), r \sin(\theta))^\top \end{cases}$$

et $F = f \circ \varphi$. Montrer que $\Delta f = \partial_r^2 F + r^{-1} \partial_r F + r^{-2} \partial_\theta^2 F$.

2 Dérivées d’ordres supérieurs – Formules de Taylor

But : Dans la section précédente, on a seulement cherché 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 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. Sous réserve d’existence, on peut définir par récurrence une notion de dérivée partielle d’ordre p par la relation

$$\partial_{i_p} \cdots \partial_{i_1} f = \frac{\partial^p f}{\partial_{i_p} \cdots \partial_{i_1}} = \frac{\partial}{\partial_{i_p}} \left(\frac{\partial^{p-1} f}{\partial_{i_{p-1}} \cdots \partial_{i_1}} \right). \quad (2.1)$$

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

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 2.2 (Théorème de Schwarz) Soient U un ouvert de \mathbb{R}^2 et $f : U \rightarrow \mathbb{R}$ 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 $\partial_{xy}^2 f = \partial_x \partial_y f = \partial_y \partial_x f$.

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

$$f := \begin{cases} \mathbb{R}^2 & \rightarrow \mathbb{R} \\ (x, y) & \rightarrow \begin{cases} xy \frac{x^2 - y^2}{x^2 + y^2} & \text{si } (x, y) \neq (0, 0) \\ 0 & \text{sinon.} \end{cases} \end{cases}$$

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 $x \rightarrow \partial_y f(x, 0)$ existe sur un voisinage de 0 et que $\lim_{x \rightarrow 0} (\partial_y f(x, 0) - \partial_y f(0, 0))/x$ existe. On a pour tout $x, y \neq 0$, quand $y \rightarrow 0$,

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

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

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

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 $\partial_y \partial_x f(0, 0) = -1$. ■

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

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

où

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

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 2.3 (Formule de Taylor) Soit $f : U \subset \mathbb{R}^q \rightarrow \mathbb{R}^p$ de classe \mathcal{C}^k sur U . Soit $x \in U$ et $h \in \mathbb{R}^q$ tels que $[x, x + h] \subset U$. Alors

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

où le terme de reste est donné par :

1. *Taylor-Lagrange pour $p = 1$: il existe $\theta \in (0, 1)$ tel que*

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

2. *Taylor avec reste intégral :*

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

3. *Taylor-Young : $\text{reste} = o(\|h\|_2^k)$.*

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. Dans ce cas, le terme d'approximation linéaire fait apparaître le gradient et le terme du second ordre fait apparaître la **Hessienne** : pour $f : U \subset \mathbb{R}^p \rightarrow \mathbb{R}$ de classe \mathcal{C}^2 , on a

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

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

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

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

Exemple 2.4 *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 $x \rightarrow \nabla f(x)$ en x : $J(\nabla f)(x) = \nabla^2 f(x) = 2A^\top A$.

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.

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

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

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 à minimiser reste constant. Comme indiqué dans le chapitre 2 du premier, on préfère représenter géométriquement une fonction $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ par plusieurs de ces lignes de niveau dans \mathbb{R}^2 plutôt que par un graphique en 3D.

On donne maintenant quelques outils qui permettent de représenter le gradient d'une fonction en fonction de ces lignes de niveau.

Définition 3.2 Soit S un sous-ensemble de \mathbb{R}^p non vide. Soit $x \in S$ et $d \in \mathbb{R}^p$. On dit que d est un **vecteur unitaire tangent à S en x** quand il existe une suite $(x_n)_n$ d'éléments de S convergente vers x telle que

$$d = \lim_{n \rightarrow \infty} \frac{x_n - x}{\|x_n - x\|_2}.$$

Par ailleurs, on dit qu'un vecteur v est **orthogonal à S en x** quand pour tout vecteur unitaire tangent d à S en x on a $\langle d, v \rangle = 0$.

Par exemple, quand $S = \mathbb{R}^p$ alors, tous les éléments de $\mathcal{S}_2^{p-1} = \{x \in \mathbb{R}^p : \|x\|_2 = 1\}$ sont des vecteurs unitaire tangent de S en chacun de ces points. Si S est un hyper-plan ou un demi-espace affine alors l'ensemble de ses vecteurs unitaires tangents est donné par $(S - x) \cap \mathcal{S}_2^{p-1}$ pour n'importe quel point $x \in S$.

Proposition 3.3 Soit $f : U \rightarrow \mathbb{R}$ 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 $\mathcal{L}_f(x)$.

Preuve. Soit d un vecteur unitaire tangent à $\mathcal{L}_f(x)$ en x . On veut montrer que $\langle \nabla f(x), d \rangle = 0$. Par définition, il existe une suite $(x_n)_n$ d'éléments de $\mathcal{L}_f(x)$ telle que $d = \lim (x_n - x) / \|x_n - x\|$. On a alors

$$\langle d, \nabla f(x) \rangle = \lim_{n \rightarrow +\infty} \left\langle \frac{x_n - x}{\|x_n - x\|_2}, \nabla f(x) \right\rangle.$$

Par ailleurs, on a pour tout n

$$f(x_n) = f(x) + \langle x_n - x, \nabla f(x) \rangle + o(\|x_n - x\|_2)$$

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

$$\langle x_n - x, \nabla f(x) \rangle + o(\|x_n - x\|_2) = 0$$

autrement dit $\lim_{n \rightarrow +\infty} \langle (x_n - x) / \|x_n - x\|_2, \nabla f(x) \rangle = 0$ ■

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. Il se trouve que c'est le gradient (renormalisé) de f en ce point. En effet, si $f : U \rightarrow \mathbb{R}$ est différentiable en x alors la direction $v \in \mathcal{S}_2^{p-1}$ de plus forte pente est celle qui maximise

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

Or $\max_{v \in \mathcal{S}_2^{p-1}} \langle \nabla f(x), v \rangle$ 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 . C'est aussi la direction de plus forte pente de f en ce point x .

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

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

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 normal à la courbe de niveau $f(x^*)$ de F_{x^*} , la meilleure approximation affine de f en x^* .

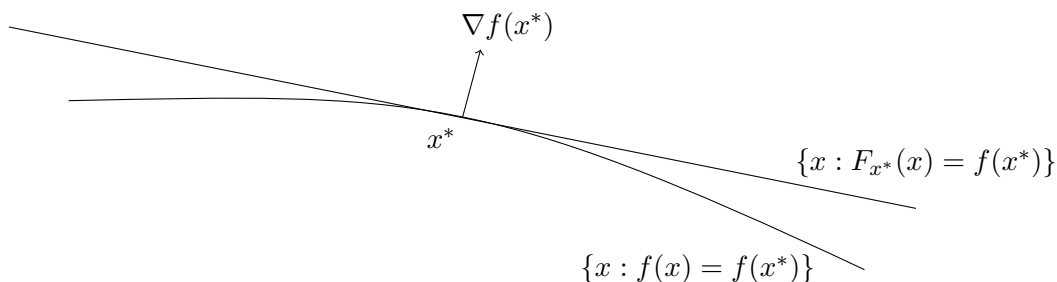


FIGURE 1 – Le gradient d'une fonction f en un point x^* est orthogonal à la courbe de niveau $f(x^*)$ de f et à la courbe de niveau $f(x^*)$ de la meilleure approximation affine de f en x^* . C'est de plus la direction de plus forte pente de f en ce point x^* .

On peut alors représenter une fonction de $\mathbb{R}^2 \rightarrow \mathbb{R}$ par quelques-une 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.

```
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
```

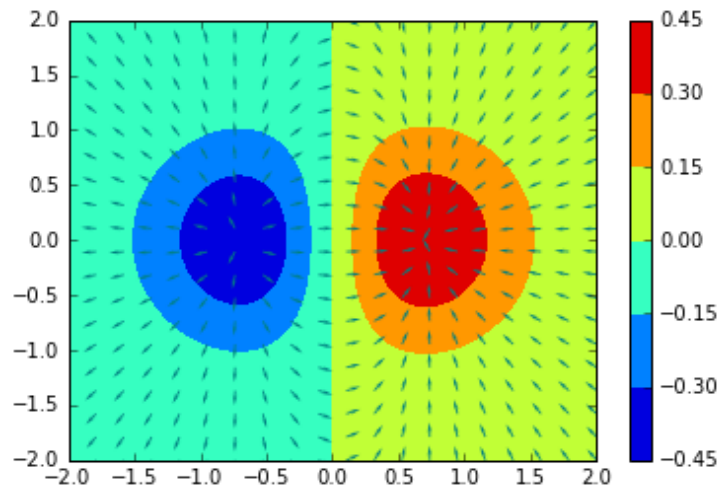


FIGURE 2 – Représentation d’une fonction $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ par quelques unes de ses lignes de niveau et de ses gradients.

```
VN = V/speed
quiver = plt.quiver(X, Y, UN, VN, color='Teal', headlength=7)
```