Notes d'optimisation différentiable

Hicham Janati
hicham.janati@inria.fr

Janvier 2020

Disclaimer : ces notes constituent un résumé et non un substitut du cours d'Optimisation différentiable.

1 Calcul différentiel

Objectifs :

1. Différentielle et gradient
2. Dérivées partielles et leur continuité
3. Chain rule
4. Hessienne et approximation de second ordre

notations

- On note $\langle x, y\rangle = \sum_{i=1}^n x_i y_i = x^\top y$ le produit scalaire Euclidien de $x, y \in \mathbb{R}^n$
- La transposée d'une matrice A est notée $A^\top$.
- La notation $x = o(h^p)$ est équivalente à $x = \|h\|^p \varepsilon(h)$ où ε est une fonction $\mathbb{R}^n \to \mathbb{R}$ continue en 0 et $\varepsilon(\mathbf{0}) = 0$.
- $\|.\|$ denote la norme Euclidienne : $\|x\| = \sqrt{\sum_i x_i^2}$.
- Pour une matrice symmétrique, $H \succ 0$ signifie que H est définie-positive, càd $x^\top H x > 0$ pour tout x non nul (ou encore toutes ses valeurs propres sont strictement positives).

Définition 1 (Différentielle et Jacobienne)**.** *Soit $f : \mathbb{R}^n \to \mathbb{R}^m$ et $x \in \mathbb{R}$. On dit que f est différentiable en x si et seulement s'il existe une application linéaire $J_x : \mathbb{R}^n \to \mathbb{R}^m$ tel que pour tout $h \in \mathbb{R}^n$:*

$$f(x + h) = f(x) + J_x(h) + o(h) \tag{1}$$

L'application J_x est dite différentielle de f en x.

Comme J_x est linéaire, elle peut être représentée par une matrice $J_f(x) \in \mathbb{R}^{m\times n}$ appelée Jacobienne de f et on a $J_f(x)_{ij} = \frac{\partial f_i}{\partial x_j}(x)$.

Exemple 1. *Soit $A \in \mathbb{R}^{m,n}$. La fonction linéaire $f : x \mapsto Ax$ est différentiable et sa hessienne est donnée par $J_f(x) = A$ pour tout $x \in \mathbb{R}^n$. En effet, $f(x + h) = A(x + h) = Ax + Ah = f(x) + Ah$.*

Définition 2 (Gradient d'une fonction scalaire)**.** *Soit $f : \mathbb{R}^n \to \mathbb{R}$ et $x \in \mathbb{R}$ une fonction différentiable. La différentielle de f en x est une application linéaire donc il existe $a_x \in \mathbb{R}^n$ tel que pour tout $h \in \mathbb{R}^n$:*

$$f(x + h) = f(x) + \langle a_x, h\rangle + o(h) \tag{2}$$

Le vecteur a_x est dit gradient de f en x et on note, pour tout x où f est différentiable : $\nabla f(x) = a_x$. En plus, les coordonnées de $\nabla f(x)$ sont données par les dérivées partielles de f : $\nabla f(x)_i = \frac{\partial f}{\partial x_i}(x)$.

*Réciproque : si les dérivées partielles $x \to \frac{\partial f}{\partial x_i}(x)$ existent et sont **continues**, alors f est différentiable et son gradient est donné par $\nabla f(x) = \left(\frac{\partial f}{\partial x_i}(x)\right)$.*

Exemple 2. *La fonction $f : x \mapsto \frac{1}{2}\|x\|^2$ est différentiable et $\nabla f(x) = x$. En effet, $\frac{1}{2}\|x+h\|^2 = \frac{1}{2}\|x\|^2 + \langle x, h\rangle + \frac{1}{2}\|h\|^2 = f(x) + \langle x, h\rangle + o(h)$.*

La continuité des dérivées partielles est nécessaire pour avoir la différentiabilité de f. Pour chercher la dérivée partielle d'une fonction f en x_0 suivant une direction $u \neq 0$, il suffit de calculer la limite :

$$\lim_{h\to 0} \frac{f(x_0+hu) - f(x_0)}{h} \tag{3}$$

L'exercice suivant montre que l'existence des dérivées partielles ne garantit pas la continuité de f – et donc sa différentiabilité.

Exercice 1 (Mi-parcours 2018)**.** *Montrer que les deux fonctions suivantes admettent des dérivées partielles en (0, 0) dans toutes les directions de $\mathbb{R}^2$ sans pour autant être continues en (0, 0).*

1. $f(x,y) = \begin{cases} y^2 \log(|x|) \text{ if } x \neq 0 \\ 0 \text{ if } x = 0 \end{cases}$
2. $f(x,y) = \begin{cases} \frac{x^2 y}{x^4+y^2} \text{ if } (x,y) \neq (0,0) \\ 0 \text{ if } (x,y) = (0,0) \end{cases}$

Proposition 1 (Chain rule)**.** *Soit $f : \mathbb{R}^n \to \mathbb{R}^m$ et $g : \mathbb{R}^m \to \mathbb{R}^p$ deux applications différentiables. Leur composée $h = gof$ est différentiable et sa Jacobienne est donnée par le produit matriciel des Jacobiennes :*

$$J_{gof}(x) = J_g(f(x))J_f(x) \tag{4}$$

Remark 1. Pour une fonction scalaire différentiable (càd à valeurs dans $\mathbb{R}$), la Jacobienne est la transposée du gradient. Dans la proposition ci-dessus, si g est scalaire ($p = 1$) alors h l'est aussi est on a :

$$\nabla h(x) = J_f(x)^\top \nabla g(f(x)) \tag{5}$$

Exemple 3 (Changement de variable linéaire)**.** *Dans la remarque ci-dessus, si $f : x \mapsto Ax$ avec $A \in \mathbb{R}^{m,n}$ alors : $\nabla h(x) = A^\top \nabla g(Ax)$.*

Exercice 2 (Les classiques)**.** *On dit que x est un point stationnaire (ou point critique) d'une fonction f différentiable en x si et seulement si $\nabla f(x) = 0$. Soit $a \in \mathbb{R}^n$, $b \in \mathbb{R}^m$ et $A \in \mathbb{R}^{m\times n}$. Soit une fonction différentable $g : \mathbb{R}^m \to \mathbb{R}$. Les fonctions suivantes sont-elles différentiables ? Donnez le gradient (Là où il existe) et les points critiques éventuels des fonctions de $\mathbb{R}^n$ dans $\mathbb{R}$ suivantes en fonction de a, b, A et ∇g.*

1. $x \mapsto \langle a, x\rangle = a^\top x$
2. $x \mapsto \|x\|^2$
3. $x \mapsto \|Ax - b\|^2$
4. $x \mapsto g(Ax)$
5. $x \mapsto \|x\|$

Définition 3 (Hessienne d'une fonction scalaire)**.** *Soit $f : \mathbb{R}^n \to \mathbb{R}$ et $x \in \mathbb{R}$ une fonction différentiable. On dit que f est deux fois différentiable en x si et seulement s'il existe une forme bilinéaire symétrique $S : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ tel que pour tout $h \in \mathbb{R}^n$:*

$$f(x+h) = f(x) + \langle \nabla f(x), h\rangle + \frac{1}{2} S_x(h,h) + o(h^2) \tag{6}$$

Comme S_x est bilinéaire symétrique, elle admet une représentation matricielle donnée par : $H_f(x)_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}(x)$. $H_f(x)$ est la Hessienne de f en x et (6) devient :

$$f(x+h) = f(x) + \nabla f(x)^\top h + \frac{1}{2} h^\top H_f(x) h + o(h^2) \tag{7}$$

Remark 2 (Hessienne comme Jacobienne). L'écriture matricielle permet de voir la Hessienne comme la Jacobienne de ∇f. Souvent, il est plus facile de retrouver la Hessienne à partir du gradient. S'il existe une application linéaire J_x telle que :

$$\nabla f(x+h) = \nabla f(x) + J_x(h) + o(h) \tag{8}$$

Alors $H_f(x) = J_x$.

Exercice 3. *Calculez la Hessienne des fonctions 1 - 2 - 3 de l'exercice 2.*

2 Optimisation sans contraintes

Objectifs :

1. Utiliser la coercivité pour montre l'existence de solution
2. Étude de points critiques avec les critères de premier et second ordre
3. Convexité et courbure d'une fonction

2.1 Existence

Soit $f : \mathbb{R}^n \to \mathbb{R}$ une fonction deux fois différentiable sur $\mathbb{R}^n$. On s'intéresse aux problème :

$$\min_{x \in \mathbb{R}^n} f(x) \tag{9}$$

Le problème (9) peut éventuellement ne pas admettre de solution, si par exemple f n'est pas minorée. Une condition suffisante pour qu'une solution existe est la coercivité

Définition 4 (Coercivité). *On dit que f est coercive si et seulement si :* $\lim_{\|x\| \to +\infty} f(x) = +\infty$

Proposition 2. *Si une fonction continue f est coercive, alors le problème (9) admet une solution.*

Exemple 4. *Soit $a \in \mathbb{R}^n$ La fonction $f : x \mapsto \|x\|^2 - \langle x, a\rangle$ est coercive. Par l'inégalité de Cauchy-Schwartz : $f(x) \geq \|x\|^2 - \|x\|\|a\| \to +\infty$ when $\|x\| \to +\infty$.*

2.2 Étude de points critiques

Souvent, on se contente de trouver des solutions locales au problème. S'il existe $x^\star$ tel que $f(x^\star) \leq f(x) \quad \forall x$ alors $x^\star$ est un minimiseur global de f, solution de (9). S'il existe $x^\star$ et $r > 0$ tel que $f(x^\star) \leq f(x) \quad \forall x \, \|x - x^\star\| \leq r$ alors $x^\star$ est un minimiseur local de f.

Proposition 3. *Soit $x^\star$ un minimiseur local de f alors $\nabla f(x^\star) = 0$.*

PROOF. Il existe un voisinage $\mathcal{N}$ de $x^\star$ tel que $\forall x \in \mathcal{N} \quad f(x^\star) \leq f(x)$. Soit $a \in \mathbb{R}^n$ et $t > 0$, on définit l'interpolation : $x_t = ta + x^\star$. On voit que si t est assez petit, $x_t \in \mathcal{N}$ et donc, pour t assez petit on peut écrire l'équation de premier ordre de f en x_t autour de $x^\star$:

$$\begin{aligned}
& f(x^\star) \leq f(x_t) \\
\Rightarrow\ & f(x^\star) \leq f(x^\star + ta) \\
\Rightarrow\ & f(x^\star) \leq f(x^\star) + \langle \nabla f(x^\star), ta\rangle + o(ta) \\
\Rightarrow\ & 0 \leq \langle \nabla f(x^\star), ta\rangle + o(t)\|a\| \\
\Rightarrow\ & 0 \leq \langle \nabla f(x^\star), a\rangle + \frac{o(t)}{t}\|a\| \\
\Rightarrow\ & 0 \leq \langle \nabla f(x^\star), a\rangle
\end{aligned}$$

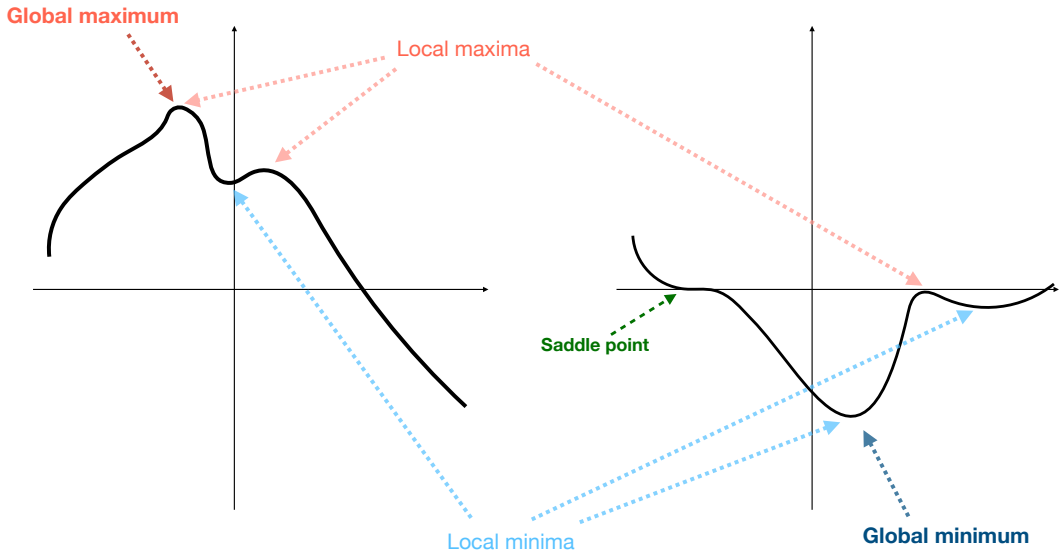

FIGURE 1 – Exemples de points critiques

où on a divisé par t avant de passer à la limite $t \to 0$.

Comme a est arbitraire dans $\mathbb{R}^n$, on a $\nabla f(x^\star) = 0$. □

On remplaçant f par $-f$, on voit que tout maximiseur local de f annule également le gradient de f. Les points annulant le gradient de f sont appelés *points critiques* ou *points stationnaires*. Pour connaitre leur nature, il faut aller au second ordre et évaluer la Hessienne de f :

Proposition 4. *Soit $x^\star$ un point critique de f. Alors :*

1. *$H_f(x^\star) \succ 0 \Rightarrow x^\star$ est un minimiseur local.*
2. *$H_f(x^\star) \prec 0 \Rightarrow x^\star$ est un maximiseur local.*

PROOF. Soit $a \in \mathbb{R}^n$ et $t > 0$. On a :

$$f(x^\star + ta) - f(x^\star) = \overbrace{\langle \nabla f(x^\star), ta \rangle}^{=0} + \frac{1}{2} t^2 a^\top H_f(x^\star) a + \|a\|^2 o(t^2)$$
$$\Rightarrow \frac{f(x^\star + ta) - f(x^\star)}{t^2} = \frac{1}{2} a^\top H_f(x^\star) a + \|a\|^2 o(1)$$

Donc pour t assez petit, le signe du membre de gauche est le signe de $a^\top H_f(x^\star) a$, et comme a est arbitraire on obtient 1 et 2. □

Exercice 4 (Examen 2018)**.** *Calculer le gradient et la Hessienne des fonctions suivantes. En déduire les points critiques des fonctions suivantes et déterminer leur nature*

1. $f : (x, y) \in \mathbb{R} \times \mathbb{R}_+^* \mapsto x^2 - \sqrt{y}$
2. $f : (x, y) \in \mathbb{R}_+^* \times \mathbb{R}_+^* \mapsto \sqrt{xy}$
3. $f : (x, y) \in \mathbb{R} \times \mathbb{R} \mapsto x^2 + y^2$

La proposition 3 donne une condition suffisante pour déterminer si un point critique est un minimiseur ou maximiseur local. Si en revanche la Hessienne a une valeur propre nulle, ce critère de deuxième ordre ne permet pas de déterminer la nature du point critique. En pratique, pour un problème de minimisation sans contraintes, après avoir énuméré tous les points critiques, il suffit d'évaluer f en ces points et comparer les valeurs obtenues : car si un minimiseur global existe, il doit être parmi ces points critiques. L'exercice suivant illustre cette situation.

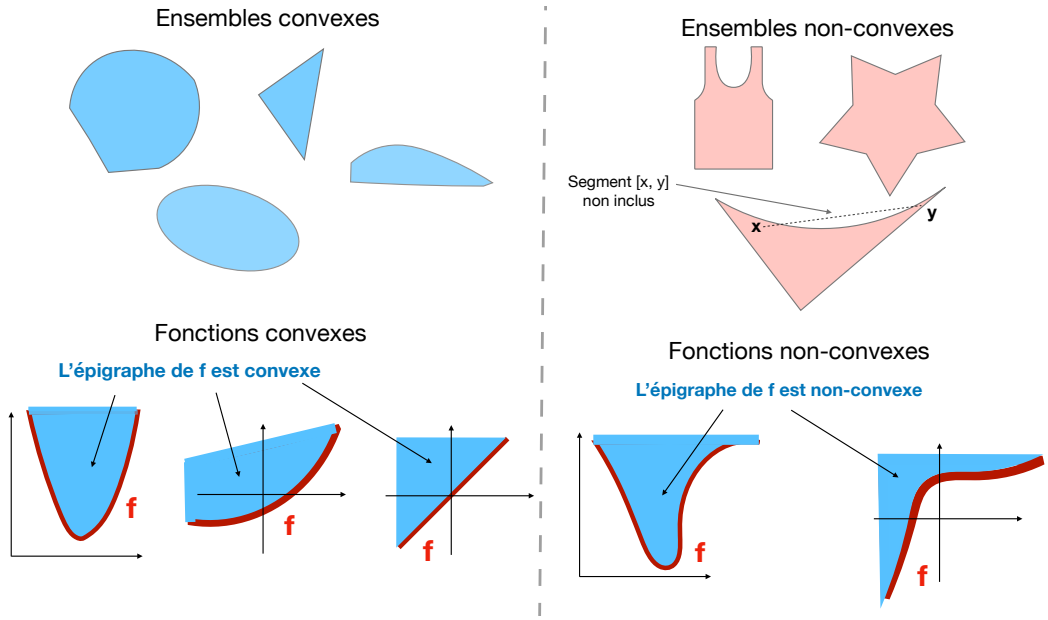

FIGURE 2 – Exemples d'ensembles et fonctions convexes

Exercice 5. *Soit $f : \mathbb{R}^2 \to \mathbb{R}$ telle que :*

$$f(x, y) = x^4 + y^4 - 2x^2 - 2y^2 + 4xy$$

On s'intéresse au problème

$$\min_{\mathbb{R}^2} f(x, y) \qquad (P)$$

1. *Montrez que (P) admet une solution*
2. *Résoudre (P)*

2.3 Cas d'une fonction convexe

Un ensemble C est convexe si et seulement si pour tout $x, y \in C$, le segment liant x à y (formellement tout point $tx + (1-t)y$ pour tout $t \in [0, 1]$) est inclus dans C. On dit qu'une fonction f est convexe si son épigraphe est convexe. L'épigraphe d'une fonction est tout simplement l'ensemble des points au-dessus de son graphe : $\text{epi}_f = \{(x_1, \ldots, x_n, y) \in \mathbb{R}^{n+1} | f(x) \leq y\}$. Cette définition admet d'autres formulations équivalentes :

Proposition 5. *Soit f une fonction deux fois différentiable de $\mathbb{R}^n$ dans $\mathbb{R}$. Les assertions suivantes sont équivalentes :*

1. *f est convexe*
2. *l'ensemble $epi_f = \{(x_1, \ldots, x_n, y) \in \mathbb{R}^{n+1} | f(x) \leq y\}$ est convexe.*
3. *$\forall (x, y) \quad \forall t \in [0, 1] f(tx + (1-t)y) \leq tf(x) + (1-t)f(y)$*
4. *$\forall (x, x_0) \quad f(x) \geq f(x_0) + \langle x - x_0, \nabla f(x_0) \rangle$ (f est supérieure à toutes ses tangeantes)*
5. *La hessienne de f est semi-définie positive pour tout x : $H_f(x) \succcurlyeq 0$.*

Lors que f est convexe, elle admet **au plus** un minimum global, atteint en potentiellement **plusieurs minimiseurs**. Si elle est strictement convexe, alors **s'il existe, ce minimiseur est unique**. En effet, l'approximation au premier ordre d'une fonction convexe permet de montrer que tout point critique est un minimiseur global. C'est donc une caractérisation des solutions du problème $\min f$.

Proposition 6. *Soit f une fonction convexe et différentiable de $\mathbb{R}^n$ dans $\mathbb{R}$. $x^\star$ est un minimiseur global de f si et seulement si $\nabla F(x^\star) = 0$.*

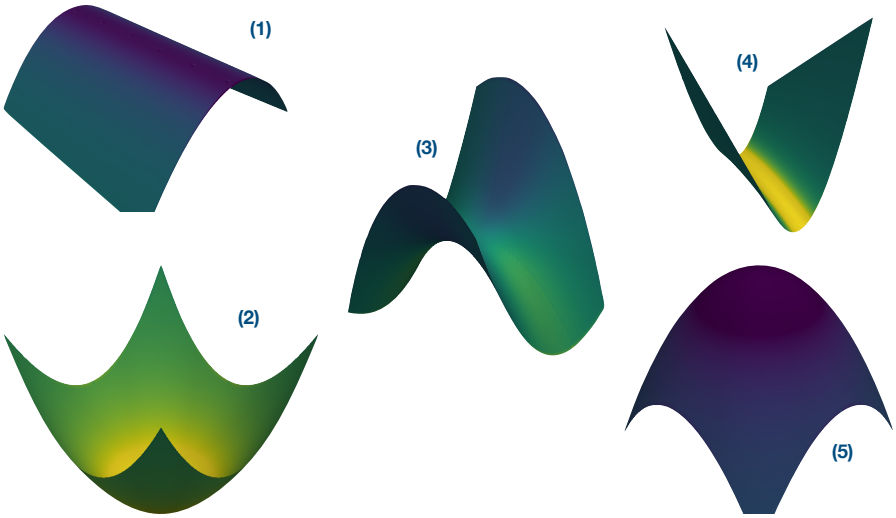

FIGURE 3 – Surfaces de fonctions quadratiques - Exercice 6

2.4 Fonction quadratique

La proposition 3 montre que l'étude des points critiques passe par l'étude de la fonction quadratique $h \to h^\top H_f h$. Les fonctions quadratiques donnent l'approximation de second ordre de toute fonction deux fois différentiable. Comprendre le lien entre la courbure d'une fonction quadratique, sa convexité et sa Hessienne est crucial en optimisation. Ceci est l'object de l'exercice suivant.

Exercice 6. *Soit S une matrice symmétrique dans $\mathbb{R}^{n,n}$ et $b \in \mathbb{R}^n$. On s'intéresse à la fonction quadratique : $f(x) = \frac{1}{2}x^\top Sx + b^\top x$.*

1. *Calculez le gradient et la Hessienne de f.*
2. *Quels sont les points critiques de f ? Discutez leur nature.*
3. *Montrez que si S a une valeur propre strictement négative, f ne peut pas être coercive.*
4. *On suppose que $S \succcurlyeq 0$. Trouvez une condition nécéssaire et suffisante sur b et S pour qu'un minimum de f existe.*
5. *Prenons $n = 2$. La figure 3 visualise la surface de f pour différentes matrices A. Déterminez le signe des valeurs propres de A dans les cas suivants :*