Université Paris Cité
UFR de Mathématiques et Informatique
45, rue des Saints-Pères, 75006, Paris.

M1 MMA 2023-2024
Optimisation

Feuille de TD n°1

Différentiabilité, gradient, hessienne.

Exercice 1 (Échauffement)
Soit $n, p \in \mathbb{N}^*$ et $f : \mathbb{R}^n \to \mathbb{R}^p$.

1. Que pouvez-vous dire de l'assertion suivante : f est différentiable sur $\mathbb{R}^n$ si et seulement si il existe $\varphi : \mathbb{R}^n \to \mathbb{R}$ tel que pour tout $x \in \mathbb{R}^n$, $h \in \mathbb{R}^n$, $f(x+h) = f(x) + \varphi(h) + o(h)$.
2. On suppose $n = p = 1$ et f dérivable sur $\mathbb{R}$. Montrer alors que f est différentiable sur $\mathbb{R}$ et exprimer la différentielle de f en tout point $x \in \mathbb{R}$ en fonction de f'.

   Quelle est la différentielle de la fonction cos sur $\mathbb{R}$?
3. Soit $f : \mathbb{R}^n \to \mathbb{R}^p$. L'application gradient ∇f est-elle bien définie ?
4. Exprimer le gradient de log en tout $x \in \mathbb{R}$.

Correction.

1. Elle est fausse ! Il y a beaucoup de choses qui ne vont pas dans cette assertion.
   — Tout d'abord l'idée fondamentale de la différentiabilité est de généraliser à $\mathbb{R}^n$ la notion de dérivée classique des fonctions de la variable réel. On est donc intéressé par l'existence de la limite pour $x \in \mathbb{R}^n$ et $h \in \mathbb{R}^n$ de
   $$\lim_{t \to 0} \frac{f(x+th) - f(x)}{t}$$
   Or ici, pour tout $t \neq 0$, on a $\frac{f(x+th)-f(x)}{t} = \frac{1}{t}\varphi(th) + \frac{1}{t}o(th)$ et donc $\frac{1}{t}\varphi(h)$ n'a aucune raison d'admettre une limite lorsque $t \to 0$ sachant les hypothèses faites sur φ.
   > Rappel : $\frac{1}{t}o(th)$ tend vers 0 lorsque $t \to 0$ puisque cette quantité se réécrit de manière équivalente comme $\|th\|\,\varepsilon(th)$ où $\varepsilon : \mathbb{R}^n \to \mathbb{R}^p$ est une fonction admettant une limite nulle en 0.
   — Il y a un soucis de dimension dans la définition de φ. C'est une fonction qui doit forcément être à valeurs dans $\mathbb{R}^p$ puisque l'on écrit $f(x+h) = f(x) + \varphi(h) + \dots$ avec $f(x+h), f(x) \in \mathbb{R}^p$.
   — Le concept de différentiabilité en un $x \in \mathbb{R}$ impose l'approximation de f au voisinage de x par une fonction linéaire, puisqu'il vient généraliser le concept de dérivabilité d'une fonction $g : \mathbb{R} \to \mathbb{R}$ qui revient à l'approximer en chaque point, avec une erreur d'ordre 1 (le $o(h)$), par une droite affine (dont la pente est le nombre dérivée). La fonction $\varphi : \mathbb{R}^n \to \mathbb{R}^p$ doit donc être une application linéaire !

   La version corrigée, en l'état, de l'assertion est donc : f est différentiable sur $\mathbb{R}^n$ si et seulement si il existe $\varphi : \mathbb{R}^n \to \mathbb{R}^p$, une application linéaire, telle que pour tout $x \in \mathbb{R}^n$, $h \in \mathbb{R}^n$, $f(x+h) = f(x) + \varphi(h) + o(h)$.
   — Enfin dans cette définition φ est valable pour tout $x \in \mathbb{R}^n$. Cela impose nécessairement que la fonction f est affine. En effet, pour tout $x \in \mathbb{R}^n$, $t \in \mathbb{R}$ on a $f(0+tx) = f(0) + \varphi(tx) + o(tx)$. Soit $x \in \mathbb{R}^n$ fixé. Posons $g : t \in \mathbb{R} \mapsto f(tx) \in \mathbb{R}^p$. Alors g est dérivable sur $\mathbb{R}$, de dérivée pour tout $t \in \mathbb{R}$, $g'(t) = \varphi(x) \in \mathbb{R}^p$. On déduit donc que g' est constante sur $\mathbb{R}$ donc en particulier continue. Ainsi g est une fonction $\mathcal{C}^1$ sur $\mathbb{R}$. Par le théorème fondamental de l'analyse (formule de Taylor reste intégrale à l'ordre 1)
   $$\forall t \in \mathbb{R}, \quad g(t) = g(0) + \int_0^t g'(s)\mathrm{d}s = g(0) + \int_0^t \varphi(x)ds = g(0) + t\varphi(x).$$
   On a donc montré que pour tout $t \in \mathbb{R}$, $f(tx) = f(0) + t\varphi(x)$. En particulier $t = 1$, on a $f(x) = f(0) + \varphi(x)$. Comme x est quelconque, on a déduit finalement que f est une fonction affine.

La bonne assertion est donc finalement : f est différentiable sur $\mathbb{R}^n$ si et seulement pour tout $x \in \mathbb{R}^n$, si il existe $\varphi_x : \mathbb{R}^n \to \mathbb{R}^p$, une application linéaire, telle que pour tout $x \in \mathbb{R}^n$, $h \in \mathbb{R}^n$, $f(x+h) = f(x) + \varphi(h) + o(h)$.

> A noter qu'en toute généralité on impose également que l'application linéaire φ_x soit continue. Mais comme nous sommes ici (et dans tout le reste du cours) dans le cadre d'espace vectoriel normé de dimension finie, tout application linéaire est automatiquement continue.

> L'application φ_x est appelée différentielle de f en x et on la note en général $\mathrm{d}f(x)$, même s'il existe de nombreuses notations concurrentes !

> Ouverture : il existe de nombreuses notions de "dérivabilité". Ici la notion de différentiabilité présentée, qui est la probablement la plus "classique", est la différentiabilité dite au sens de Fréchet [a]. C'est celle qui généralise la notion de dérivabilité des fonctions définies sur $\mathbb{R}$.
> Mais donc il est possible d'alléger certaines hypothèses pour obtenir des concepts plus généraux de dérivabilité. Par exemple on peut uniquement imposer l'existence en un $x \in \mathbb{R}$ d'une limite du taux d'accroissement $\frac{f(x+th)-f(x)}{t}$ lorsque $t \to 0$ pour tout $h \in \mathbb{R}^n$. C'est-à-dire de dérivées directionnelles en un point x, dans toutes les directions $h \in \mathbb{R}^n$. Les limites, notées pour tout $h \in \mathbb{R}^n$, $Df(x,h)$, définissent une application $Df(x,\cdot)h \in \mathbb{R}^n \mapsto Df(x,h)$. Mais ici $Df(x,\cdot)$ n'est pas supposé linéaire (continue). Si c'était le cas, on retomberait sur la différentiabilité au sens de Fréchet en $x \in \mathbb{R}^n$ et pour tout $x \in \mathbb{R}^n$, on aurait $Df(x,h) = df(x)(h)$.
> Une fonction f qui vérifie cette condition en $x \in \mathbb{R}^n$ est dite différentiable au sens de Gâteaux [b] en x.
>
> a. https://en.wikipedia.org/wiki/Fréchet_derivative
> b. https://en.wikipedia.org/wiki/Gateaux_derivative

2. Soit $x \in \mathbb{R}$. Comme f est dérivable en x, f admet un développement limité en x à l'ordre 1. Ainsi pour tout $h \in \mathbb{R}$, $f(x+h) = f(x) + f'(x)h + o(h)$. Comme $h \in \mathbb{R} \mapsto f'(x)h$ est linéaire (continue), on a bien f différentiable sur en x et pour tout $h \in \mathbb{R}$, $df(x)(h) = f'(x)h$. C'est vrai pour tout $x \in \mathbb{R}$, donc f est différentiable sur $\mathbb{R}$.

   D'après ce qui précède, comme cos est dérivable sur $\mathbb{R}$, elle est différentiable sur $\mathbb{R}$ et on a pour tout $x \in \mathbb{R}$, tout $h \in \mathbb{R}$, $d(\cos)(x)(h) = (-\sin(x))h = -\sin(x)h$.

3. Seulement si $p = 1$, et bien sûr si f différentiable sur $\mathbb{R}^n$.

   Pour rappel la définition du gradient de f en $x \in \mathbb{R}^n$ repose sur le fait que $h \in \mathbb{R}^n \mapsto df(x)(h)$ soit une *forme linéaire*, c'est-à-dire une application linéaire de $\mathbb{R}^n$ dans $\mathbb{R}$. Car alors on peut utiliser le théorème de représentation de Riesz et représenter $h \in \mathbb{R}^n \mapsto df(x)(h)$ grâce à un produit scalaire. En effet il existe alors un unique $v \in \mathbb{R}^n$ tel que pour tout $h \in \mathbb{R}^n$, $df(x)(h) = \langle v, h \rangle$. Comme ce v est unique, on lui donne un nom et une notation dans ce contexte : le gradient de f en x et $v := \nabla f(x)$.

   Si $p > 1$, alors $h \in \mathbb{R}^n \mapsto df(x)(h)$ n'est plus une forme linéaire et donc on ne peut plus représenter cette application à travers un produit scalaire.

   Ainsi si $p = 1$ et f différentiable sur $\mathbb{R}^n$, alors pour tout $x \in \mathbb{R}$, $\nabla f(x)$ est bien défini et donc l'application gradient $\nabla f : x \in \mathbb{R}^n \mapsto \nabla f(x) \in \mathbb{R}^p$ l'est également.

4. L'application log est dérivable sur $\mathbb{R}_+^*$, donc différentiable sur $\mathbb{R}_+^*$ et pour tout $x \in \mathbb{R}_+^*$, tout $h \in \mathbb{R}$, $d(\log)(x)(h) = \left(\frac{1}{x}\right) h = \left\langle \frac{1}{x}, h \right\rangle$ avec le produit scalaire définie sur $\mathbb{R}$ qui correspond simplement au produit. Ainsi par identification $\nabla(\log)(x) = \frac{1}{x}$. C'est le nombre dérivé de la fonction log en x.

Exercice 2 (Vers une interprétation géométrique du gradient)
Soit $f : \mathbb{R}^2 \to \mathbb{R}$ la fonction définie pour tout $x \in \mathbb{R}^2$, par $f(x) = \|Ax\|_2^2$, où

$$A = \begin{pmatrix} 4 & 0 \\ 0 & 1 \end{pmatrix}.$$

1. Quelle est la régularité de la fonction f ?
2. Déterminer pour tout $x \in \mathbb{R}^2$, $h \in \mathbb{R}^2$, $\mathrm{d}f(x)(h)$, puis $\nabla f(x)$.
3. Déterminer pour tout $x \in \mathbb{R}^2$, $h, k \in \mathbb{R}^2$, $\mathrm{d}^2 f(x)(h,k)$, puis $\nabla^2 f(x)$.
4. a) Représenter graphiquement l'ensemble $L_1 = \{x \in \mathbb{R}^2 : f(x) = 1\}$, appelée ensemble de niveau 1 de f.

   b) Choisissez $x \in L_1$ et représenter le vecteur $\nabla f(x)$ en ce point. Que remarquez-vous ?

Correction.

1. La fonction f est polynomiale en les coefficients du vecteur $x \in \mathbb{R}^2$, donc f est $\mathcal{C}^\infty$ sur $\mathbb{R}^2$.

2. f est donc différentiable sur $\mathbb{R}^n$. Pour trouver sa différentielle, déterminons son DL à l'ordre 1. Soit $x \in \mathbb{R}^2$ et $h \in \mathbb{R}^2$. On a

$$f(x+h) = \|A(x+h)\|_2^2 = \langle Ax + Ah,\, Ax + Ah\rangle = f(x) + 2\langle Ax,\, Ah\rangle + \underbrace{\|Ah\|_2^2}_{=o_{h\to 0}(\|h\|)},$$
$$= f(x) + \left\langle 2A^TAx,\, h\right\rangle + o(\|h\|).$$

Par identification de la partie linéaire en h de cette expression on déduit que $df(x)(h) = \left\langle A^TAx,\, h\right\rangle$, puis $\nabla f(x) = 2A^TAx$. Dans la suite nous simplifions cette expression puisque comme A est une matrice symétrique, on a $A^TA = A^2$.

3. Je vous propose une réponse qui mêle des rappels de cours sur la manière dont est défini la différentielle seconde, histoire de bien comprendre cet objet qui n'est pas évident au premier abord.

Soit $x \in \mathbb{R}^2$, et $h, k \in \mathbb{R}^2$, pour trouver $d^2f(x)(h,k)$, il faut écrire le DL de $df(x+k)(h)$ en x par rapport à k. On a $df(x+k)(h) = \left\langle 2A^2(x+k),\, h\right\rangle = df(x)(h) + \left\langle 2A^2k,\, h\right\rangle$. Par identification de la partie linéaire en k, on a donc $d(df)(x)(k) = \left\langle 2A^2k,\, \cdot\right\rangle$. On note plutôt cette quantité $d^2f(x)(k)$, elle appartient par définition à $\mathcal{L}(\mathbb{R}^2, \mathbb{R})$, puisque $d^2f(x) \in \mathcal{L}(\mathbb{R}^2, \mathcal{L}(\mathbb{R}^2, \mathbb{R}))$. $d^2f(x)$ est donc une forme bilinéaire, et on note pour cette raison pour tout $h, k \in \mathbb{R}^2$, $d^2f(x)(h,k)$ plutôt que $d^2f(x)(k)(h)$. Et pour rappel final, dès que f est deux fois différentiable en x, alors $d^2f(x)$ est une forme bilinéaire symétrique, donc pour tout $h, k \in \mathbb{R}^2$, $d^2f(x)(h,k) = d^2f(x)(k,h)$.

Rappel d'algèbre. Soit $u : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ une forme bilinéaire et $\mathcal{B} = (e_1, \dots, e_n)$ une base. Alors b est caractérisée par la donnée de la matrice $U = (u(e_i, e_j))_{1\leqslant i,j\leqslant n}$ que l'on appelle la matrice de la forme bilinéaire u. De plus si u est symétrique alors la matrice U est symétrique. Si $\mathcal{B}$ est orthonormée pour le produit scalaire $\langle\cdot,\cdot\rangle$, alors pour tout $x, y \in \mathbb{R}^n$, $u(x,y) = \langle Ux,\, y\rangle$.

Rappel de cours. Lorsque f est deux fois différentiable en $x \in \mathbb{R}^n$ et f à valeurs dans $\mathbb{R}$, alors $d^2f(x)$ est une forme bilinéaire symétrique. Sa matrice dans la base orthonormée canonique de $\mathbb{R}^n$ est appelée matrice hessienne de f et notée $\nabla^2 f(x)$. On a donc pour tout $h, k \in \mathbb{R}^n$, $d^2f(x)(h,k) = \left\langle \nabla^2 f(x)h,\, k\right\rangle = \left\langle \nabla^2 f(x)k,\, h\right\rangle$.

Comme pour tout $x \in \mathbb{R}^n$, pour tout $h, k \in \mathbb{R}^n$, $d^2f(x)(h,k) = \left\langle 2A^2k,\, h\right\rangle$, par identification, $\nabla^2 f(x) = 2A^2$.

4. a) On a pour tout $x = (x_1, x_2) \in \mathbb{R}^2$, $f(x) = \frac{x_1^2}{\left(\frac{1}{2}\right)^2} + x_2^2$. Ainsi L_1 est l'ellipse de centre $(0,0)$ et passant par les points $(\frac{1}{2}, 0)$ et $(0,1)$.

b) Voir Figure 1.

Exercice 3

On fixe $\varepsilon > 0$ et $n \in \mathbb{N}$ avec $n \geqslant 3$. On définit

$$\forall x \in \mathbb{R}^n, \quad J_\varepsilon(x) = \sum_{i=2}^{n-1} N_\varepsilon(x_{i+1} + x_{i-1} - 2x_i), \quad \text{où} \quad N_\varepsilon(t) = \sqrt{\varepsilon + t^2}.$$

1. Écrire le développement à l'ordre 1 de J_ε et en déduire que J_ε est différentiable, puis donner une expression de $\mathrm{d}J_\varepsilon(x)(h)$ pour tout $x, h \in \mathbb{R}^n$.

2. Montrer que J_ε peut s'écrire sous la forme

$$\forall x \in \mathbb{R}^n, \quad J_\varepsilon(x) = \sum_{i=2}^{n-1} N_\varepsilon(A_i x),$$

avec des matrices A_i, $i \in \{2, \dots, n-1\}$, que l'on explicitera. Réécrire la différentielle grâce aux A_i, et en déduire le gradient et la hessienne de J_ε en tout point $x \in \mathbb{R}^n$. *On pensera à justifier que J_ε est deux fois différentiable sur $\mathbb{R}^n$.*

3. Donner une expression de $\mathrm{d}^2J_\varepsilon(x)(h,k)$ pour tout $x \in \mathbb{R}^n$ et $(h,k) \in \mathbb{R}^n \times \mathbb{R}^n$ faisant intervenir les A_i.

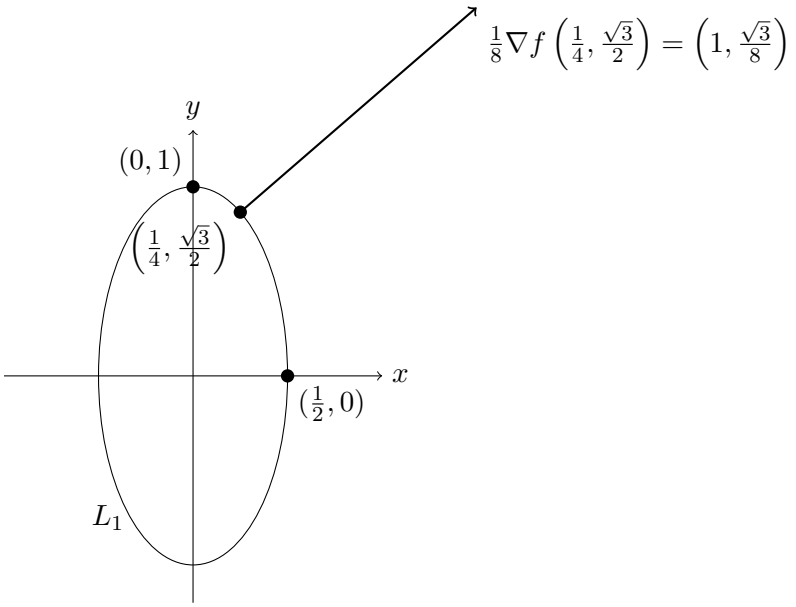

FIGURE 1. Ligne de niveau L_1 de f. On remarque, comme attendu, que le gradient au point de L_1 choisi est orthogonal à L_1 (i.e. orthogonal à la tangente à L_1 en ce point). Pour une démonstration générale de ce fait, voir l'Exercice 5.

Correction.

1. N_ε est $\mathcal{C}^\infty$ sur $\mathbb{R}$, en particulier admet un DL à l'ordre 1. On a pour tout $x, h \in \mathbb{R}^n$,

$$\begin{aligned} J_\varepsilon(x+h) &= \sum_{i=2}^{n-1} N_\varepsilon(x_{i+1}+x_{i-1}-2x_i+h_{i+1}+h_{i-1}-2h_i), \\ &= \underbrace{\sum_{i=2}^{n-1} N_\varepsilon(x_{i+1}+x_{i-1}-2x_i)}_{=J_\varepsilon(x)} + \underbrace{\sum_{i=2}^{n-1} N'_\varepsilon(x_{i+1}+x_{i-1}-2x_i)(h_{i+1}+h_{i-1}-2h_i)}_{\text{linéaire en } h} + o(\|h\|). \end{aligned}$$

Donc J_ε est différentiable en x et

$$\mathrm{d}J_\varepsilon(x)(h) = \sum_{i=2}^{n-1} N'_\varepsilon(x_{i+1}+x_{i-1}-2x_i)(h_{i+1}+h_{i-1}-2h_i).$$

2. Il suffit de poser pour tout $i \in \{2,\dots,n-1\}$, $A_i = (0\cdots 0\ 1\ -2\ 1\ 0\cdots 0) \in \mathcal{M}_{1,n}(\mathbb{R})$ où le -2 est en i-ème position.

> On vérifie par analyse dimensionnelle que $A_i x$ est un réel.

On a pour tout $x, h \in \mathbb{R}^n$

$$\mathrm{d}J_\varepsilon(x)(h) = \sum_{i=2}^{n-1} N'_\varepsilon(x_{i+1}+x_{i-1}-2x_i)A_i h = \left\langle \sum_{i=2}^{n-1} N'_\varepsilon(x_{i+1}+x_{i-1}-2x_i)A_i^T,\, h \right\rangle,$$

donc $\nabla J_\varepsilon(x) = \sum_{i=2}^{n-1} N'_\varepsilon(A_i x)A_i^T$.

> On vérifie par analyse dimensionnelle que $\nabla J_\varepsilon(x)$ est un vecteur (vecteur colonne).

L'application $x \in \mathbb{R}^n \mapsto \nabla J_\varepsilon(x)$ est $\mathcal{C}^1$ par composée d'applications $\mathcal{C}^1$, donc J_ε est deux fois continûment différentiable sur $\mathbb{R}^n$.

> Pour montrer que J_ε est deux fois différentiable en $x_0 \in \mathbb{R}^n$, sachant déjà que $\nabla J_\varepsilon(x)$ existe en tout $x \in \mathbb{R}^n$, il suffit de montrer que ∇J_ε est différentiable en x_0. En effet, comme pour tout $x \in \mathbb{R}^n$, on a $\mathrm{d}J_\varepsilon(x) = \langle \nabla J_\varepsilon(x), \cdot \rangle$, montrer que $\mathrm{d}J_\varepsilon : x \mapsto \mathrm{d}J_\varepsilon(x)$ est différentiable en x_0 est équivalent à montrer que ∇J_ε est différentiable en x_0.

On différencie ∇J_ε. Soit $x, h \in \mathbb{R}^n$, alors

$$\begin{aligned}\nabla J_\varepsilon(x+h) &= \sum_{i=2}^{n-1} N_\varepsilon'(A_i(x+h))A_i^T = \nabla J_\varepsilon(x) + \sum_{i=2}^{n-1} N_\varepsilon''(A_i x) \underbrace{A_i h}_{\in\mathbb{R}} \underbrace{A_i^T}_{\in\mathcal{M}_{n,1}(\mathbb{R})} + o(\|h\|),\\ &= \nabla J_\varepsilon(x) + \sum_{i=2}^{n-1} N_\varepsilon''(A_i x) \underbrace{A_i^T A_i}_{\in\mathcal{M}_{n,n}(\mathbb{R})} h + o(\|h\|),\\ &= \nabla J_\varepsilon(x) + \left(\sum_{i=2}^{n-1} N_\varepsilon''(A_i x) A_i^T A_i\right) h + o(\|h\|).\end{aligned}$$

> On a utilisé à la première ligne un développement de Taylor Young à l'ordre 1 de N_ε' en $A_i x$. C'est légitime comme N_ε' est $\mathcal{C}^\infty$ sur $\mathbb{R}$.

Donc $\nabla^2 J_\varepsilon(x) = \sum_{i=2}^{n-1} N_\varepsilon''(A_i x) A_i^T A_i$.

3. On a pour tout $x \in \mathbb{R}^n$ et $(h,k) \in \mathbb{R}^n \times \mathbb{R}^n$, $\mathrm{d}^2 J_\varepsilon(x)(h,k) = k^T \nabla^2 J_\varepsilon(x) h = \langle k, \nabla^2 J_\varepsilon(x) h\rangle = \langle \nabla^2 J_\varepsilon(x) k, h\rangle$, car $\nabla^2 J_\varepsilon(x)$ est la matrice dans la base canonique de $\mathbb{R}^n$ de la forme bilinéaire $\mathrm{d}^2 J_\varepsilon(x)$.

> Rappel : $\nabla^2 J_\varepsilon(x)$ est une matrice symétrique car $\mathrm{d}^2 J_\varepsilon(x)$ est une forme bilinéaire symétrique d'après le théorème de Schwarz.

On a donc $\mathrm{d}^2 J_\varepsilon(x)(h,k) = \sum_{i=2}^{n-1} N_\varepsilon''(A_i x) \underbrace{k^T A_i^T}_{\in\mathbb{R}} \underbrace{A_i h}_{\in\mathbb{R}} = \sum_{i=2}^{n-1} N_\varepsilon''(A_i x) \underbrace{A_i k}_{k_{i+1}+k_{i-1}-2k_i} \underbrace{A_i h}_{h_{i+1}+h_{i-1}-2h_i}$.

Exercice 4

Soit $N, M \in \mathbb{N}^*$. On considère la fonction suivante

$$J_\lambda : \alpha \in \mathbb{R}^N \mapsto \frac{1}{n}\sum_{i=1}^{n} \log(1 + e^{-y_i \langle \alpha, x_i\rangle}) + \lambda \|L\alpha\|^2,$$

où $\lambda > 0$, $L \in \mathcal{M}_{M,N}(\mathbb{R})$ et les x_i, y_i, pour tout $i \in \{1,\dots,n\}$, sont des éléments fixés.

1. Donner la nature des éléments x_i, y_i, $i \in \{1,\dots,n\}$. Sur quel espace est définie la norme euclidienne utilisée dans l'expression de J_λ ?

2. Donner une expression du gradient et de la hessienne de J_λ après avoir préalablement justifié qu'elle est deux fois continûment différentiable. *Indication : on pourra penser à introduire diverses notations comme dans l'Exercice 3 pour se faciliter la tâche.*

Correction.

1. Pour tout $i \in \{1,\dots,n\}$, y_i doit être un réel et $x_i \in \mathbb{R}^N$. Comme $L : \mathbb{R}^N \to \mathbb{R}^M$, la norme euclidienne est définie sur $\mathbb{R}^M$. Par contre le produit scalaire dans l'exponentielle est défini sur $\mathbb{R}^N$.

2. Posons pour tout $t \in \mathbb{R}$, $f(t) = \log(1+e^{-t})$. La fonction f est $\mathcal{C}^\infty$ sur $\mathbb{R}$ comme composée de fonctions $\mathcal{C}^\infty$ et de plus pour tout $t \in \mathbb{R}$, $f(t) > 0$. Pour tout $i \in \{1,\dots,n\}$, $\alpha \in \mathbb{R}^N \mapsto y_i \langle \alpha, x_i\rangle = y_i x_i^T \alpha$ est $\mathcal{C}^\infty$ sur $\mathbb{R}^n$ puisque c'est une fonction linéaire (c'est une forme linéaire). Enfin $\alpha \in \mathbb{R}^N \mapsto \lambda \|L\alpha\|^2$ est également $\mathcal{C}^\infty$ sur $\mathbb{R}^N$ puisque polynomiale en les coefficients de α. Par somme et composée de fonctions $\mathcal{C}^\infty$, on déduit que f est $\mathcal{C}^\infty$ sur $\mathbb{R}^N$.

Soit $\alpha, h \in \mathbb{R}^N$. On a

$$\begin{aligned}
J_\lambda(\alpha+h) &= \frac{1}{n}\sum_{i=1}^n f(y_i \langle x_i,\, \alpha+h\rangle) + \lambda \langle L(\alpha+h),\, L(\alpha+h)\rangle, \\
&= \frac{1}{n}\sum_{i=1}^n f(y_i \langle x_i,\, \alpha\rangle + y_i \langle x_i,\, h\rangle)) + \underbrace{\lambda \langle L\alpha,\, L\alpha\rangle}_{=\lambda\|L\alpha\|^2} + \underbrace{\lambda \langle Lh,\, L\alpha\rangle + \lambda \langle L\alpha,\, Lh\rangle}_{=2\lambda\langle L^TL\alpha,\, h\rangle} + \underbrace{\lambda \langle Lh,\, Lh\rangle}_{=o(\|h\|)}, \\
&= \frac{1}{n}\sum_{i=1}^n \left(\underbrace{f(y_i \langle x_i,\, \alpha\rangle) + f'(y_i \langle x_i,\, \alpha\rangle) y_i \langle x_i,\, h\rangle + o(\|h\|)}_{\text{DL1 de } f \text{ en } y_i \langle x_i,\, \alpha\rangle \in \mathbb{R}} \right) + \lambda \|L\alpha\|^2 + 2\lambda \left\langle L^TL\alpha,\, h\right\rangle + o(\|h\|), \\
&= J_\lambda(\alpha) + \underbrace{\frac{1}{n}\sum_{i=1}^n f'(y_i \langle x_i,\, \alpha\rangle) y_i \langle x_i,\, h\rangle + 2\lambda \left\langle L^TL\alpha,\, h\right\rangle}_{\text{linéaire en } h} + o(\|h\|).
\end{aligned}$$

> Dans le calcul $\langle Lh,\, L\alpha\rangle = \left\langle L^TL\alpha,\, h\right\rangle$, le premier produit scalaire est défini sur $\mathbb{R}^M$ alors que le second sur $\mathbb{R}^N$. On a utilisé les mêmes notations, par un léger abus de langage.

Donc

$$\mathrm{d}J_\lambda(\alpha)(h) = \frac{1}{n}\sum_{i=1}^n f'(y_i \langle x_i,\, \alpha\rangle) y_i \langle x_i,\, h\rangle + 2\lambda \left\langle L^TL\alpha,\, h\right\rangle = \left\langle \frac{1}{n}\sum_{i=1}^n f'(y_i \langle x_i,\, \alpha\rangle) y_i x_i + 2\lambda L^TL\alpha,\, h \right\rangle,$$

ainsi par identification $\nabla J_\lambda(\alpha) = \frac{1}{n}\sum_{i=1}^n \underbrace{f'(y_i \langle x_i,\, \alpha\rangle) y_i}_{\in\mathbb{R}} \underbrace{x_i}_{\in\mathbb{R}^N} + 2\lambda \underbrace{L^TL\alpha}_{\in\mathbb{R}^N} \in \mathbb{R}^N$.

Pour trouver $\nabla^2 J_\lambda(\alpha)$, on calcule

$$\begin{aligned}
\nabla J_\lambda(\alpha+h) &= \frac{1}{n}\sum_{i=1}^n f'(y_i \langle x_i,\, \alpha+h\rangle) y_i x_i + 2\lambda L^TL(\alpha+h), \\
&= \nabla J_\lambda(\alpha) + \frac{1}{n}\sum_{i=1}^n f''(y_i \langle x_i,\, \alpha\rangle)(y_i \langle x_i,\, h\rangle) y_i x_i + 2\lambda L^TLh + o(\|h\|),
\end{aligned}$$

où on a a utilisé, de manière similaire au précédent calcul, pour tout $i \in \{1, \dots, n\}$

$$f'(y_i \langle x_i,\, \alpha+h\rangle) = f'(y_i \langle x_i,\, \alpha\rangle) + f''(y_i \langle x_i,\, \alpha\rangle) y_i \langle x_i,\, h\rangle + o(\|h\|),$$

car la fonction $f' : \mathbb{R} \to \mathbb{R}$ est dérivable sur $\mathbb{R}$, donc admet un développement de Taylor Young à l'ordre 1 en tout point de $\mathbb{R}$.

Ainsi on déduit que

$$\mathrm{d}(\nabla J_\lambda)(\alpha)(h) = \frac{1}{n}\sum_{i=1}^n \underbrace{f''(y_i \langle x_i,\, \alpha\rangle) y_i^2}_{\in\mathbb{R}} \underbrace{\langle x_i,\, h\rangle}_{\in\mathbb{R}} \underbrace{x_i}_{\in\mathbb{R}^N} + 2\lambda L^TLh.$$

> Il faut essayer d'exprimer le membre de droite sous la forme Ah où A est une matrice de taille $N \times N$. Pour le terme $2\lambda L^TLh$ c'est déjà fait. Par contre il faut faire apparaître une matrice dans la quantité $f''(y_i \langle x_i,\, \alpha\rangle) y_i^2 \langle x_i,\, h\rangle x_i$, pout tout i.

On remarque que, pour tout $i \in \{1, \dots, n\}$, $\langle x_i,\, h\rangle x_i = \underbrace{(x_i^T h)}_{\in\mathbb{R}} x_i = x_i(x_i^T h) = \underbrace{(x_i x_i^T)}_{\in\mathcal{M}_{N,N}(\mathbb{R})} h$. Ainsi

$$\mathrm{d}(\nabla J_\lambda)(\alpha)(h) = \left(\frac{1}{n}\sum_{i=1}^n f''(y_i \langle x_i,\, \alpha\rangle) y_i^2 x_i x_i^T + 2\lambda L^TL \right) h,$$

et donc par identification $\nabla^2 J_\lambda(\alpha) = \frac{1}{n}\sum_{i=1}^n f''(y_i \langle x_i,\, \alpha\rangle) y_i^2 x_i x_i^T + 2\lambda L^TL$.

> Ce calcul de matrice hessienne ressemble pas mal à celui de l'Exercice 3. Notamment on a fait la manipulation $\langle x_i, h\rangle x_i = (x_i^T h)x_i = x_i(x_i^T h)$ qui mène à la matrice carré $x_i x_i^T$. C'est très proche du calcul de l'Exercice 3 : $(A_i h)A_i^T = A_i^T(A_i h)$, qui mène à la matrice carré $A_i^T A_i$. On pourrait se demander pourquoi dans le premier exercice la matrice est $A_i^T A_i$ et ici $x_i x_i^T$? La réponse : analyse dimensionnelle ! Les objets A_i et x_i ne sont pas de même nature. A_i a été défini comme une matrice ligne, c'est la matrice de la forme linéaire $x \in \mathbb{R}^n \mapsto \langle a_i, x\rangle = x_{i+1} + x_{i-1} - 2x_i$ avec a_i le vecteur $(0, \dots, 0, 1, -2, 1, 0, \dots, 0) \in \mathbb{R}^n$. En identifiant les vecteurs à des matrices colonnes, on a donc $A_i = a_i^T$. À l'inverse, x_i est un vecteur donc de même nature que a_i. D'où la différence.

Exercice 5 (Formalisation de l'interprétation géométrique du gradient)
Soit $f : \mathbb{R}^2 \to \mathbb{R}$ une fonction de classe $\mathcal{C}^1$. Pour $\lambda \in \mathbb{R}$, on note $L_\lambda = \{x \in \mathbb{R}^2 : f(x) = \lambda\}$ la ligne de niveau λ de f.

Soit $x_0 \in \mathbb{R}^2$ tel que $\nabla f(x_0) \neq 0$ et notons $\lambda_0 = f(x_0)$. On se donne $\gamma :]-\varepsilon, \varepsilon[\to \mathbb{R}^2$, pour un certain $\varepsilon > 0$, une fonction de classe $\mathcal{C}^1$ telle que $\gamma(0) = x_0$ et $\|\gamma'(0)\| = \|\nabla f(x_0)\|$.

1. Quelle condition doit-on avoir sur γ pour que $f \circ \gamma$ décroîsse le plus vite au voisinage de 0. En déduire que $-\nabla f(x_0)$ donne la direction de la plus forte pente de f en x_0.

2. On suppose maintenant que $f \circ \gamma$ est constante. Démontrer que $\nabla f(x_0)$ est orthogonale $\gamma'(0)$, i.e. à la tangente à la ligne de niveau L_{λ_0} en x_0.

3. On suppose que $\gamma'(0) = \nabla f(x_0)$. Montrer alors qu'il existe une fonction $\xi : \lambda \mapsto \xi(\lambda) \in \mathbb{R}^2$ définie sur un voisinage de λ_0, $\mathcal{C}^1$ et telle que $\xi(\lambda) \in L_\lambda$. En déduire un équivalent de $\|\xi(\lambda) - x_0\|$ lorsque $\lambda \to \lambda_0$. Interpréter.

Correction.

1. $f \circ \gamma :]-\varepsilon, \varepsilon[\to \mathbb{R}$ est $\mathcal{C}^1$ sur $]-\varepsilon, \varepsilon[$, par composée de fonctions $\mathcal{C}^1$, et pour tout $t \in]-\varepsilon, \varepsilon[$, on a

$$(f \circ \gamma)'(t) = \mathrm{d}f(\gamma(t))(\gamma'(t)) = \big\langle \nabla f(\gamma(t)), \gamma'(t)\big\rangle,$$

ainsi $|(f\circ\gamma)'(t)| \leqslant \|\nabla f(\gamma(t))\| \cdot \|\gamma'(t)\|$ par l'inégalité de Cauchy-Schwarz, et on a égalité dans l'inégalité si et seulement si les vecteurs $\nabla f(\gamma(t))$ et $\gamma'(t)$ sont liés.

Ainsi, parmi tous les choix possibles de γ, $f \circ \gamma$ décroit le plus vite au voisinage de 0 si et seulement si il y a égalité dans l'inégalité de Cauchy-Schwarz en $t = 0$ et avec $(f \circ \gamma)'(0) < 0$, i.e. si et seulement si $\gamma'(0) = -\nabla f(x_0)$.

Ainsi $-\nabla f(x_0)$ définie bien la direction de plus forte pente de f en x_0.

2. > Comme $f \circ \gamma$ est constante, on a donc $\gamma(]-\varepsilon, \varepsilon[) \subset L_{\lambda_0}$. La fonction γ est donc une courbe de $\mathbb{R}^2$ tracée, dans un voisinage de x_0, dans la ligne de niveau λ_0 de f.

On a pour tout $t \in]-\varepsilon, \varepsilon[$, $(f \circ \gamma)'(t) = 0$. Donc en particulier, on a $\langle \nabla f(x_0), \gamma'(0)\rangle = 0$, i.e. $\nabla f(x_0) \perp \gamma'(0)$.

3. Par hypothèse on a $(f \circ \gamma)'(0) > 0$. D'après la question 1., on déduit en particulier que $f \circ \gamma$ a la plus forte croissance, parmi tous les choix possibles de γ, dans un voisinage de 0.

Quitte à choisir ε plus petit, on peut supposer que pour tout $t \in]-\varepsilon, \varepsilon[$, on a $(f\circ\gamma)'(t) > 0$, puisque $(f\circ\gamma)'$ est continue. Ainsi $\varphi = f \circ \gamma$ est un $\mathcal{C}^1$-difféomorphisme de $]-\varepsilon, \varepsilon[$ dans $\Lambda = f \circ \gamma(]-\varepsilon, \varepsilon[) \subset \mathbb{R}$. L'ensemble Λ est un intervalle (image continue dans intervalle) ouvert de $\mathbb{R}$ contenant λ_0.

Posons $\xi = \gamma \circ \varphi^{-1} : \Lambda \to \mathbb{R}^2$. Alors ξ est $\mathcal{C}^1$ sur Λ, par composée de fonctions $\mathcal{C}^1$, et par définition de φ, pour tout $\lambda \in \Lambda$, $f(\xi(\lambda)) = f \circ \gamma \circ (f \circ \varphi)^{-1}(\lambda) = \lambda$, i.e. $\xi(\lambda) \in L_\lambda$. De plus

$$\forall \lambda \in \Lambda, \quad \xi'(\lambda) = \underbrace{\gamma'(\varphi^{-1}(\lambda))}_{\in\mathbb{R}^2}\underbrace{(\varphi^{-1})'(\lambda)}_{\in\mathbb{R}} = \frac{\gamma'(\varphi^{-1}(\lambda))}{\varphi'(\varphi^{-1}(\lambda))} = \frac{\gamma'(\varphi^{-1}(\lambda))}{\langle \nabla f(\gamma(\lambda)), \gamma'(\varphi^{-1}(\lambda))\rangle}.$$

En particulier, $\xi'(\lambda_0) = \frac{\nabla f(x_0)}{\|\nabla f(x_0)\|^2}$. Comme ξ est $\mathcal{C}^1$, d'après la formule de Taylor Young à l'ordre 1, on a

$$\forall \lambda \in \Lambda, \quad \xi(\lambda) = \underbrace{\xi(\lambda_0)}_{=x_0} + \underbrace{\xi'(\lambda_0)}_{=\frac{\nabla f(x_0)}{\|\nabla f(x_0)\|^2}}(\lambda - \lambda_0) + (\lambda - \lambda_0)^2\varepsilon(\lambda),$$

avec $\varepsilon : \Lambda \to \mathbb{R}^2$ tel que $\lim_{\lambda\to\lambda_0} \varepsilon(\lambda) = 0$. Ainsi

$$\forall \lambda \in \Lambda \setminus \{\lambda_0\}, \quad \left\| \frac{\xi(\lambda) - x_0}{\lambda - \lambda_0} \right\| = \left\| \xi'(\lambda_0) + (\lambda - \lambda_0)\varepsilon(\lambda) \right\|,$$

> On aimerait pouvoir conclure directement que $\|\xi'(\lambda_0) + (\lambda - \lambda_0)\varepsilon(\lambda)\|$ est équivalent à $\|\xi'(\lambda_0)\|$ quand $\lambda \to \lambda_0$. Mais ce n'est pas évident comme le $o_{\lambda\to\lambda_0}(\lambda - \lambda_0)$ est à l'intérieur de la norme. L'idée est de faire un DL1 de la norme $\|\cdot\|$.

Comme $\|\cdot\| = \sqrt{\|\cdot\|^2}$, $\|\cdot\|$ définie une application différentiable sur $\mathbb{R}^2 \setminus \{0\}$. On a pour tout $x \in \mathbb{R}^2 \setminus \{0\}$, $h \in \mathbb{R}^2$ suffisamment petit ($h \in B(0, r)$ avec $r > 0$ tel que $x + B(0, r) \subset \mathbb{R}^2 \setminus \{0\}$)

$$\|x + h\| = \sqrt{\|x\|^2 + 2\langle x,\, h\rangle + o(\|h\|)} = \|x\| + \frac{\langle x,\, h\rangle}{\|x\|} + o(\|h\|). \tag{1}$$

où on a utilisé le DL1 de $\sqrt{\cdot}$ en $\|x\|^2 \neq 0$. Ainsi, comme $\xi'(\lambda_0) \neq 0$, on a

$$\left\|\xi'(\lambda_0) + (\lambda - \lambda_0)\varepsilon(\lambda)\right\| = \left\|\xi'(\lambda_0)\right\| + o_{\lambda\to\lambda_0}(\lambda - \lambda_0).$$

Par conséquent, on déduit

$$\|\xi(\lambda) - x_0\| \underset{\lambda\to\lambda_0}{\sim} \left\|\xi'(\lambda_0)\right\| (\lambda - \lambda_0) = \frac{\lambda - \lambda_0}{\|\nabla f(x_0)\|} \cdot \frac{\nabla f(x_0)}{\|\nabla f(x_0)\|}.$$

Plus $\|\nabla f(x_0)\|$ est grand et plus les lignes de niveau sont resserrées.