Module A4
Fonctions régulières

Sauf mention contraire, $\mathcal{X}$ (respectivement $\mathcal{Y}$...) est un espace de Hilbert, muni du produit scalaire noté $\langle \cdot, \cdot \rangle$ et on note $\| \cdot \|$ la norme euclidienne associée.

Dans ce module, on s'intéresse à la classe des fonctions dites *régulières* (*smooth* en anglais), qui interviennent dans de nombreux schémas numériques d'optimisation.

1 Fonctions régulières

1.1 Définition

Introduisons la définition suivante :

Définition 1 (Fonctions régulières)

Soit $J : \mathcal{X} \to \mathbb{R}$ une fonction différentiable. Soit $L \geq 0$. On dit que J est L-*régulière* si son gradient $\nabla J : \mathcal{X} \to \mathcal{X}$ est lipschitzien de constante de Lipschitz L. Autrement dit, J est L-régulière si

$$\forall\, (x,z) \in \mathcal{X}^2, \qquad \|\nabla J(x) - \nabla J(z)\| \leq L\, \|x - z\|$$

Remarque : Attention : dans d'autres domaines des mathématiques, l'adjectif *régulière* peut désigner d'autres propriétés (par exemple le caractère $\mathcal{C}^\infty$). Il est donc essentiel d'être attentif à la définition considéré par l'auteur.

On peut déjà faire les observations suivantes :

- si J est L-régulière, alors J est L'-régulière pour $L' \geq L$;
- si J est L-régulière, alors J est de classe $\mathcal{C}^1$. La réciproque est évidemment fausse.

Lorsque la fonction est deux fois différentiable, on a la proposition suivante :

Proposition 1

Soit $J : \mathcal{X} \to \mathbb{R}$ une fonction deux fois différentiable. On suppose qu'il existe $M \geq 0$ tel que

$$\forall\, x \in \mathcal{X}, \qquad |||\operatorname{Hess} J(x)||| \leq M$$

Alors J est M-régulière.

Démonstration : Il s'agit d'une conséquence de l'inégalité des accroissements finis dans le cas vectoriel. ■

Pour les rappels sur les normes d'opérateurs, on pourra consulter le module **Compléments C1 : Éléments d'algèbre linéaire**.

1.2 Composition avec un opérateur affine

L'ensemble des fonctions régulières est stable par combinaison linéaire.

Proposition 2

Soit $J_1 : \mathcal{X} \to \mathbb{R}$ une fonction L_1-régulière et $J_2 : \mathcal{X} \to \mathbb{R}$ une fonction L_2-régulière. Soit $(\lambda, \mu) \in \mathbb{R}^2$. Alors la fonction $\lambda\, J_1 + \mu\, J_2$ est $(|\lambda|\, L_1 + |\mu|\, L_2)$-régulière.

DÉMONSTRATION : On applique la linéarité du gradient et l'inégalité triangulaire. ■

Proposition 3

Soit $J : \mathcal{X} \to \mathbb{R}$ une fonction L-régulière. Soit $A : \mathcal{Y} \to \mathcal{X}$ un opérateur linéaire borné et $b \in \mathcal{X}$. Alors la fonction $x \mapsto J(A\,x + b)$ est $(|||A|||^2 L)$-régulière.

REMARQUE : Dans ce cours, on écrit $A\,x$ au lieu de $A(x)$ lorsque A est linéaire.

DÉMONSTRATION : La fonction considérée est différentiable de gradient

$$x \mapsto A^* \nabla J(A\,x + b)$$

puis on utilise la régularité de J pour écrire que, pour tout $(x, z) \in \mathcal{X}^2$,

$$\|A^* \nabla J(A\,x + b) - A^* \nabla J(A\,z + b)\| \leq |||A^*||| \times L \times \|(A\,x + b) - (A\,z + b)\|$$

Puisque

$$\|(A\,x + b) - (A\,z + b)\| = \|A\,x - A\,z\| = \|A\,(x - z)\| \leq |||A|||\, \|x - z\|$$

on obtient le résultat désiré en combinant les deux majorations. ■

> En pratique, de nombreux algorithmes d'optimisation du premier ordre nécessitent d'être capables de déterminer **une** constante de LIPSCHITZ du gradient d'une fonction régulière. Plus celle-ci est proche de la plus petite constante de LIPSCHITZ du gradient, meilleur peut être le comportement de l'algorithme considéré.

2 Lemmes de descente et de BAILLON–HADDAD

Dans cette section, on va établir deux résultats qui vont nous être utiles de nombreuses preuves de convergence d'algorithmes d'optimisation du premier ordre.

2.1 Lemme de descente

Proposition 4 (Lemme de descente)

Soit $J : \mathcal{X} \to \mathbb{R}$ une fonction L-régulière. Alors on a

$$\forall\,(x, z) \in \mathcal{X}^2, \qquad J(z) \leq J(x) + \langle \nabla J(x), z - x \rangle + \frac{L}{2}\, \|x - z\|^2$$

DÉMONSTRATION : Soit $(x,z) \in \mathcal{X}^2$. Considérons la fonction définie sur $[\,0\,;1\,]$ par

$$\forall\, t \in [\,0\,;1\,], \qquad f(t) = J\big(x + t\,(z-x)\big) - \big\langle \nabla J(x), x + t\,(z-x) \big\rangle$$

Cette fonction est dérivable ; il est alors aisé de vérifier que

$$\begin{aligned} J(z) - J(x) - \langle \nabla J(x), z - x\rangle = \Big[f(t)\Big]_0^1 &= \int_0^1 f'(t)\,\mathrm{d}t \\ &= \int_0^1 \big\langle \nabla J\big(x + t\,(z-x)\big) - \nabla J(x), z - x \big\rangle\,\mathrm{d}t \end{aligned}$$

Or, on a

$$\begin{aligned} &\left| \int_0^1 \big\langle \nabla J\big(x + t\,(z-x)\big) - \nabla J(x), z - x \big\rangle\,\mathrm{d}t \right| \\ &\qquad \le \int_0^1 \left|\big\langle \nabla J\big(x + t\,(z-x)\big) - \nabla J(x), z - x \big\rangle\right|\,\mathrm{d}t \end{aligned}$$

En utilisant l'inégalité de CAUCHY–SCHWARZ et la régularité de J, on obtient alors

$$\int_0^1 \left|\big\langle \nabla J\big(x + t\,(z-x)\big) - \nabla J(x), z - x \big\rangle\right|\,\mathrm{d}t \le L \int_0^1 t\,\|z-x\|^2\,\mathrm{d}t$$

et on en déduit le résultat souhaité après calcul de l'intégrale. ■

2.2 Cas d'une fonction convexe

Dans le cas d'une fonction convexe, on peut démontrer le résultat suivant :

Lemme 1 (BAILLON–HADDAD)

Soit $J : \mathcal{X} \to \mathbb{R}$ une fonction différentiable. Soit $L > 0$. Alors les deux énoncés suivants sont équivalents :

(i) J est **convexe** et L-régulière.

(ii) $\forall\,(x,z) \in \mathcal{X}^2, \qquad \langle \nabla J(x) - \nabla J(z), x - z\rangle \ge \dfrac{1}{L}\,\|\nabla J(x) - \nabla J(z)\|^2$

Dans ce cas, on dit que $(1/L)\nabla J$ est un opérateur *fermement non-expansif* ou que ∇J est un opérateur $(1/L)$-*co-coercif*.

DÉMONSTRATION : On démontre séparément les deux sens de l'équivalence.

- **(i) $\Longrightarrow$ (ii) :** Commençons par remarquer que, pour tout $(x,z) \in \mathcal{X}^2$ et pour tout $p \in \mathcal{X}$, on a

$$\begin{aligned} \langle \nabla J(z) - \nabla J(x), p - x\rangle - \frac{L}{2}\,\|p-x\|^2 = {} & \frac{1}{2\,L}\,\|\nabla J(z) - \nabla J(x)\|^2 \\ & - \frac{L}{2}\,\left\| \frac{\nabla J(z) - \nabla J(x)}{L} - (p-x)\right\|^2 \end{aligned}$$

On passe à la borne supérieure sur $p \in \mathcal{X}$; on obtient que

$$\sup_{p\in\mathcal{X}} \left\{ \langle \nabla J(z) - \nabla J(x), p - x\rangle - \frac{L}{2}\,\|p-x\|^2 \right\} = \frac{1}{2\,L}\,\|\nabla J(z) - \nabla J(x)\|^2$$

de sorte que

$$\begin{aligned} &\sup_{p\in\mathcal{X}} \left\{ \langle \nabla J(z) - \nabla J(x), p\rangle - \frac{L}{2}\,\|p-x\|^2 \right\} \\ &\qquad = \langle \nabla J(z) - \nabla J(x), x\rangle + \frac{1}{2\,L}\,\|\nabla J(z) - \nabla J(x)\|^2 \end{aligned}$$

Puisque J est L-régulière, le lemme de descente (proposition 4) assure que :

$$\langle \nabla J(z) - \nabla J(x), p \rangle - \frac{L}{2} \|p - x\|^2 \leq J(x) - J(p) + \langle \nabla J(z), p \rangle - \langle \nabla J(x), x \rangle$$

On passe à nouveau à la borne supérieure sur $p \in \mathcal{X}$:

$$\begin{aligned}\langle \nabla J(z) - \nabla J(x), x \rangle + \frac{1}{2\,L} \|\nabla J(z) - \nabla J(x)\|^2 \leq & \; -\langle \nabla J(x), x \rangle + J(x) \\ & + \sup_{p \in \mathcal{X}} \Big\{ \langle \nabla J(z), p \rangle - J(p) \Big\}\end{aligned}$$

Après réarrangement, on obtient que

$$\begin{aligned}\langle \nabla J(x), x \rangle - J(x) \leq & \; \langle \nabla J(x) - \nabla J(z), x \rangle - \frac{1}{2\,L} \|\nabla J(z) - \nabla J(x)\|^2 \\ & + \sup_{p \in \mathcal{X}} \Big\{ \langle \nabla J(z), p \rangle - J(p) \Big\}\end{aligned}$$

La fonction J est convexe et différentiable ; on a donc pour tout $p \in \mathcal{X}$

$$J(p) \geq J(z) + \langle \nabla J(z), p - z \rangle \qquad \text{soit} \qquad \langle \nabla J(z), z \rangle - J(z) \geq \langle \nabla J(z), p \rangle - J(p)$$

Autrement dit,

$$\langle \nabla J(z), z \rangle - J(z) = \sup_{p \in \mathcal{X}} \Big\{ \langle \nabla J(z), p \rangle - J(p) \Big\}$$

On a donc démontré que, pour tout $(x, z) \in \mathcal{X}^2$,

$$\begin{aligned}\langle \nabla J(x), x \rangle - J(x) \leq & \; \langle \nabla J(x) - \nabla J(z), x \rangle - \frac{1}{2\,L} \|\nabla J(z) - \nabla J(x)\|^2 \\ & + \langle \nabla J(z), z \rangle - J(z)\end{aligned}$$

En réécrivant cette inégalité en inversant le rôle de x et de z, et en sommant les deux inégalités, on démontre finalement le résultat annoncé.

- **(ii) $\Longrightarrow$ (i) :** Il suffit d'utiliser l'inégalité de Cauchy–Schwarz pour majorer le produit scalaire et ainsi obtenir la définition de la régularité. La convexité découle alors de la monotonie du gradient. ■

2.3 Cas fortement convexe

Enfin, dans le cas d'une fonction J différentiable et fortement convexe de module α, on rappelle qu'on a la relation suivante :

$$\forall\,(x, z) \in \mathcal{X}^2, \qquad J(z) \geq J(x) + \langle \nabla J(x), z - x \rangle + \frac{\alpha}{2} \|x - z\|^2$$

On en déduit qu'une fonction régulière et fortement convexe est encadrée par deux fonctions quadratiques. La "distance" entre ces deux fonctions dépend du rapport entre la constante de Lipschitz du gradient et le module de forte convexité de la fonction.

Définition 2 (Conditionnement d'une fonction régulière fortement convexe)

Soit $J : \mathcal{X} \to \mathbb{R}$ une fonction L-régulière et fortement convexe de module α. On définit le *conditionnement* de J par

$$\kappa = \frac{L}{\alpha}$$

REMARQUE : D'après ce qui précède cette définition, le conditionnement est nécessairement un nombre positif supérieur à 1.

> Plus le conditionnement de J est petit (c'est-à-dire, proche de 1), plus les deux fonctions quadratiques encadrant J sont proches. *A contrario*, plus il est grand, et plus ces deux fonctions quadratiques sont éloignées. Or, si ces deux fonctions sont proches, le comportement de J est *a priori* mieux contrôlé (l'espace dont elle dispose est plus réduit). C'est pourquoi un faible conditionnement est intuitivement préférable à un grand conditionnement.

Dans l'exemple suivant, on va montrer qu'il existe un lien entre le conditionnement d'une fonction et le conditionnement d'une matrice :

EXEMPLE

Conditionnement d'une forme quadratique strictement convexe.
Soit $n \in \mathbb{N}$. Considérons la forme quadratique suivante :

$$J : \left\{ \begin{array}{ccc} \mathbb{R}^n & \to & \mathbb{R} \\ x & \mapsto & \dfrac{1}{2} \langle A\,x, x \rangle \end{array} \right.$$

avec $A \in \mathcal{M}_{m,n}(\mathbb{R})$ une matrice symétrique définie positive. Comme on le verra dans la section suivante, la fonction J est fortement convexe, de module $\lambda_{\min}$, et $|\lambda_{\max}|$-régulière. Il s'ensuit que le conditionnement de J vaut

$$\kappa = \frac{\lambda_{\max}}{\lambda_{\min}}$$

On reconnaît là la définition usuelle du conditionnement de la matrice A.

3 Exemples

3.1 Fonctions quadratiques généralisées

Définition 3 (Fonction quadratique généralisée)

Soit $A \in \mathcal{M}_{n,n}(\mathbb{R})$ une matrice réelle symétrique, $b \in \mathbb{R}^n$ un vecteur et $c \in \mathbb{R}$ un scalaire. Alors on appelle *fonction quadratique généralisée* la fonction définie par

$$\left\{ \begin{array}{ccc} \mathbb{R}^n & \to & \mathbb{R} \\ x & \mapsto & \dfrac{1}{2} \langle A\,x, x \rangle - \langle b, x \rangle + c \end{array} \right. \tag{$\star$}$$

Autrement dit, toute fonction quadratique généralisée est la somme d'une forme quadratique, d'une forme linéaire et d'une constante.

REMARQUE : Si A n'est pas symétrique, il suffit de la remplacer par sa symétrisée

$$\tilde{A} = \frac{A + A^\top}{2}$$

Les fonctions quadratiques généralisées sont régulières, comme le montre la proposition suivante :

Proposition 5

Soit $A \in \mathcal{M}_{n,n}(\mathbb{R})$ une matrice réelle **symétrique**, $b \in \mathbb{R}^n$ un vecteur et $c \in \mathbb{R}$ un scalaire, et soit f la fonction quadratique généralisée définie dans $(\star)$. Alors f est différentiable et son gradient est donné pour tout $x \in \mathbb{R}^n$ par

$$\nabla f(x) = A\,x - b$$

De plus, f est $|\lambda_{\max}|$-régulière, où $\lambda_{\max}$ est la plus grande valeur propre de A.

On peut également signaler les résultats suivants :

Proposition 6

Soit $A \in \mathcal{M}_{n,n}(\mathbb{R})$ une matrice réelle **symétrique**, $b \in \mathbb{R}^n$ un vecteur et $c \in \mathbb{R}$ un scalaire, et soit f la fonction quadratique généralisée définie dans $(\star)$. Alors les énoncés suivants sont équivalents :

(i) f est convexe (resp. $\lambda_{\min}$-fortement convexe)

(ii) A est semi-définie positive (resp. définie positive)

(iii) les valeurs propres de A sont positives (resp. strictement positives), avec $\lambda_{\min}$ la plus petite valeur propre de A.

Ainsi, le lemme de descente assure que pour tout $x \in \mathbb{R}^n$,

$$\frac{\lambda_{\min}}{2}\,\|x\|_2^2 - \langle b, x\rangle + c \leq \frac{1}{2}\,\langle A\,x, x\rangle - \langle b, x\rangle + c \leq \frac{\lambda_{\max}}{2}\,\|x\|_2^2 - \langle b, x\rangle + c$$

L'inégalité de Cauchy–Schwarz permet alors d'écrire

$$\frac{\lambda_{\min}}{2}\,\|x\|_2^2 - \|b\|_2 \cdot \|x\|_2 + c \leq \frac{1}{2}\,\langle A\,x, x\rangle - \langle b, x\rangle + c \leq \frac{\lambda_{\max}}{2}\,\|x\|_2^2 - \|b\|_2 \cdot \|x\|_2 + c$$

ce qui donne un encadrement de la fonction quadratique f qui ne dépend que de la norme de la variable x.

> Pour plus de détails à propos des fonctions quadratiques généralisées et leur rôle en optimisation, on se reportera au module **B2 : Introduction à la méthode des moindres carrés** (3MA261).

3.2 Fonction de Huber

Soit $\alpha > 0$. On considère la fonction $J : \mathbb{R} \to \mathbb{R}$ définie de la manière suivante :

$$\forall\, x \in \mathbb{R}, \qquad J(x) = \begin{cases} \dfrac{x^2}{2} & \text{si } |x| \leq \alpha \\[2ex] \alpha\,|x| - \dfrac{\alpha^2}{2} & \text{sinon} \end{cases}$$

Elle est visiblement dérivable sur $]-\infty\,;-\alpha\,[$, $]-\alpha\,;\alpha\,[$ et $]\,\alpha\,;\infty\,[$, de dérivée

$$J'(x) = \begin{cases} x & \text{si } |x| < \alpha \\ \alpha & \text{si } x > \alpha \\ -\alpha & \text{si } x < -\alpha \end{cases}$$

Il est aisé de vérifier que J' est continue en α et $-\alpha$, de sorte que J est différentiable sur tout $\mathbb{R}$. Soit $(x,z) \in \mathbb{R}^2$. En calculant $|J'(x) - J'(z)|$ pour tous les cas de figure possibles, on démontre que

$$\forall\,(x,z) \in \mathbb{R}^2, \qquad |J'(x) - J'(z)| \leq |x - z|$$

ce qui assure que J est 1-régulière. L'allure de la courbe représentative de quelques unes de ces fonctions est présentée en figure 1.

> Dans le module **A5 : Opérateur proximal de Moreau**, on donnera un sens plus précis à la notion de régularisation pour les fonctions de Huber.

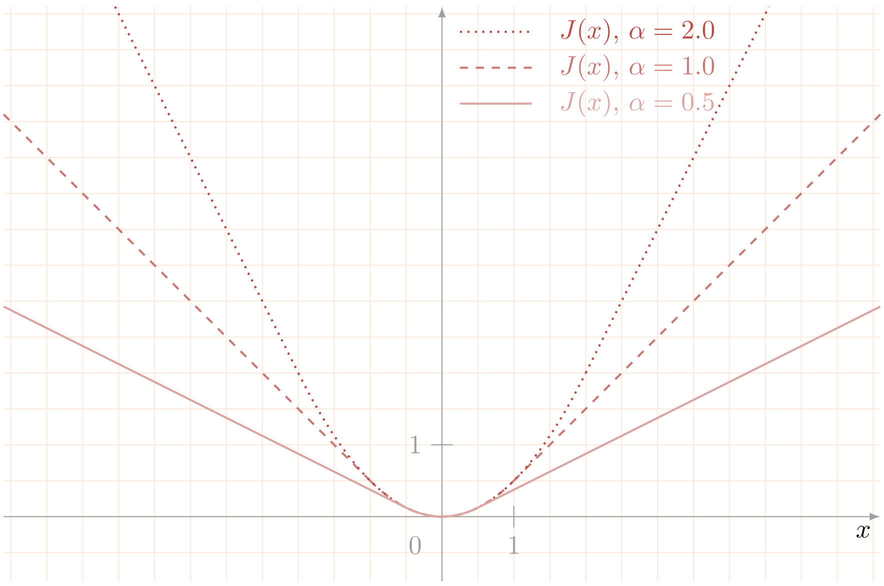

Figure 1 – Exemples de fonctions de Huber.

> D'autres régularisation de la valeur absolue existent. Parmi elles, mentionnons les fonctions suivantes :
>
> - $t \mapsto \sqrt{t^2 + \varepsilon}$ pour $\varepsilon > 0$
> - $t \mapsto |t| - \alpha\,\ln\left(1 + \dfrac{|t|}{\alpha}\right)$ pour $\alpha > 0$
>
> qui apparaissent parfois en traitement du signal.

Pour aller plus loin

Descentes de gradient explicite. L'hypothèse de régularité apparaît dans de nombreux algorithmes d'optimisation du premier ordre, dès lors que les itérations sont définies à l'aide du gradient de tout ou partie de la fonction objectif (dans le second cas, seule la partie concernée est supposée régulière). Parmi ces algorithmes figurent les méthodes de gradient explicite (module **B2 : Méthodes du gradient explicite**), les algorithmes utilisant l'éclatement explicite-implicite (*Forward-Backward Splitting*, module **B4 : Éclatement primal d'opérateurs**) tels que l'algorithme PALM (module **B6 : Éclatement de variables**). Dans ces cas, les preuves de convergence reposent fortement sur le lemme de descente que nous avons établi dans ce module.

Conjuguée convexe. On verra dans le module **A8 : Dualité de Fenchel et conjuguée convexe**que la régularité d'une fonction convexe se traduit par la forte convexité de sa conjuguée convexe. La démonstration de ce résultat repose en partie sur le lemme de Baillon–Haddad. La conjuguée convexe apparaît d'ailleurs furtivement dans la preuve du lemme de Baillon–Haddad.

Fonctions quadratiques généralisées. Ces fonctions apparaissent naturellement dans la méthode des moindres carrés, qui constitue une classe de problèmes classiques en analyse numérique (cf. module **B2 : Introduction à la méthode des moindres carrés** (3MA261)). On les retrouve aussi souvent dans l'analyse de la convergence des algorithmes de ce cours, en tant qu'approximation quadratique (locale) des fonctions objectifs. Enfin, signalons qu'elles jouent un rôle important en recherche linéaire (module **B2 : Méthodes du gradient explicite**).

Fonctions de Huber. Ces fonctions apparaissent naturellement comme régularisations de Moreau–Yosida (cf. module **A5 : Opérateur proximal de Moreau**).