La régularisation-approximation de Moreau d'une fonction convexe, l'opérateur proximal, les méthodes de gradient proximal, etc. : une présentation synthétique pour ceux qui n'en ont jamais entendu parler

J.-B. Hiriart-Urruty
Institut de Mathématiques
Université Paul Sabatier de Toulouse
(automne 2021)

"Optimization is revolutionized by its interactions with Machine Learning and Data Analysis", S. Wright (2016)

"Proximal methods are the natural algorithms for solving regularized learning problems", F. Iutzeler and J. Malick (2020)

Introduction

Allez dans un colloque d'Optimisation, plus précisément dans des sessions consacrées à des problèmes de grande taille comme il s'en trouve en imagerie mathématique, apprentissage automatique ou statistique (Machine Learning), et vous entendrez parler de la *régularisée de Moreau,* des *méthodes (algorithmiques) proximales*, etc. Pour les comprendre, il faut un minimum de connaissances de base théoriques (c'est-à-dire mathématiques). C'est à cette nécessité que j'ai dû répondre en enseignant dans un Master 2R de Recherche Opérationnelle (cours intitulé *"Thèmes contemporains en Optimisation (continue)"* lors des six dernières années. L'auditoire d'étudiants (à Bac + 5) venait essentiellement de quatre écoles d'ingénieurs de Toulouse ainsi que de l'université Paul Sabatier [1].

Nous avons fait le choix de partir d'un niveau débutant dans le domaine visé, d'où le titre de ce texte, en évitant la tentation de considérer comme acquis des choses qui nous paraissent simples (tant on est "dedans" par nos propres pratiques et travaux en Optimisation).

L'exposé qui suit est divisé 6 en parties de longueurs très inégales. Après des paragraphes liminaires d'Analyse (§1) et d'Analyse convexe moderne (§2), nous présentons au §3 les propriétés de la régularisée de Moreau (au premier ordre) ; le tout est distillé sous forme de "faits" (= des énoncés) sans démonstrations. C'est, pour l'étudiant-lecteur, **le socle pour comprendre les méthodes algorithmiques dites proximales**. Le paragraphe 4 est dédié aux propriétés de la régularisée de Moreau (au deuxième ordre) ; en plus de faire la synthèse des résultats disponibles en termes de calcul différentiel classique, nous améliorons quelques résultats de la littérature. Au §5, un coup d'oeil est jeté sur les modèles généraux d'algorithmes dits de type proximal en optimisation convexe ; des références indiquées permettront à l'étudiant-lecteur d'aller plus loin et de se préparer à l'utilisation, voire l'amélioration, de ces méthodes dans des domaines d'applications. Le §6 ne consiste qu'en une inflexion vers le monde de l'optimisation non convexe.

1. Je remercie Marcel Mongeau (Professeur à l'ENAC, Toulouse) de cette initiative, et d'avoir mis en place ce Master 2R "de site" en Recherche Opérationnelle à Toulouse, domaine auquel il a toujours été attaché.

1. Préliminaires d'Analyse

1.1. Ce qu'est une fonction convexe s.c.i.

Nous allons considérer des fonctions à valeurs étendues, c'est-à-dire dans $\mathbb{R} \cup \{+\infty\}$. Ce n'est pas par un caprice de mathématicien ou un désir exagéré de généralisation, mais bien une approche utile et même nécessaire pour aborder de manière à la fois générale et synthétique ce que nous avons à présenter. Pour une telle fonction $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, son *domaine de finitude* est noté $\mathrm{dom} f$; dit autrement, $\mathrm{dom} f = \{x : f(x) \in \mathbb{R}\}$. Nous ne considérerons, évidemment, que des fonctions f pour lesquelles $\mathrm{dom} f$ n'est pas vide.

Une première notion, plus "vectorielle" que "d'Analyse" est celle de *convexité*. Nous ne nous contentons que d'une définition géométrique globale : une fonction f est convexe lorsque son *épigraphe* $\mathrm{epi} f$, littéralement ce qui est au-dessus du graphe de f, c'est-à-dire $\{(x, y) \in \mathbb{R}^n \times \mathbb{R} : f(x) \leqslant y\}$, est une partie convexe (de $\mathbb{R}^n \times \mathbb{R}$, donc).

Une deuxième notion, d'Analyse cette fois, est celle de *semi-continuité inférieure* (s.c.i. en abrégé) d'une fonction f. C'est, en termes imagés, "la moitié de la continuité qui vient par dessous". Nous en donnons deux définitions rapides, équivalentes bien sûr.

- En tout point x de $\mathbb{R}^n$, chaque fois qu'une suite (x_k) tend vers x et que $f(x_k)$ tend vers ℓ (quand $k \to \infty$) on doit avoir $\ell \geqslant f(x)$. Dit en termes à peine dégrossis, "on chute à la limite". Un exemple est la fonction rang d'une matrice : quand une suite (M_k) de matrices tend vers une matrice M, si la suite $(\mathrm{rang} M_k)_k$ a une limite r, alors $r \geqslant \mathrm{rang} M$ (le rang ne peut que chuter à la limite).

- Une fonction $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ est s.c.i. (sur $\mathbb{R}^n$) lorsque son épigraphe $\mathrm{epi} f$ est une partie fermée (de $\mathbb{R}^n \times \mathbb{R}$). Cela explique la terminologie *closed functions* utilisée parfois dans la littérature anglo-saxonne pour de telles fonctions.

C'est le moment de mettre en garde contre un piège. *"Une fonction continue est s.c.i."* ... vrai, bien sûr... sauf qu'il y a une chausse-trappe ici : l'inégalité sur les limites de suites évoquée dans la première définition au-dessus doit être vérifiée pour tous les x, y compris ceux qui ne sont pas dans le domaine de finitude $\mathrm{dom} f$ de f, c'est-à-dire en fait ceux qui sont au bord (ou à la frontière) de $\mathrm{dom} f$. Bref, l'inégalité $\ell \geqslant f(x)$ doit être comprise dans $\mathbb{R} \cup \{+\infty\}$. Prenons par exemple la fonction $f : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ qui vaut 0 si $x \in]-1, 1[$, $+\infty$ sinon. Bien sûr, cette fonction est continue là où elle est finie, sur $]-1, 1[$ donc, ... mais elle n'est pas s.c.i. sur $\mathbb{R}$.

Les effets de bord ennuyeux sont facilement corrigés de la manière suivante : si une fonction $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ n'est pas s.c.i., c'est-à-dire si son épigraphe $\mathrm{epi} f$ n'est pas fermé, on a quand même de la chance car $\mathrm{cl}(\mathrm{epi} f)$ (l'adhérence ou la fermeture de $\mathrm{epi} f$) est à son tour l'épigraphe d'une fonction, s.c.i. celle-là, notée $\overline{f}$ ou $\mathrm{cl} f$. Dans l'exemple simple juste au-dessus, $\mathrm{cl} f(x)$ vaut 0 si $x \in [-1, 1]$, $+\infty$ sinon.

Il se trouve que pour une fonction $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ qui se trouve être déjà convexe, la construction et la définition analytique de $\mathrm{cl} f$ se trouvent simplifiées. Voici comment on peut procéder. Soit a fixé à l'intérieur du domaine de f ; alors

$$\mathrm{cl} f(x) = \lim_{t \downarrow 0} f(x + t(a - x)) \text{ pour tout } x \in \mathrm{cl}(\mathrm{dom} f). \tag{1}$$

Bref, cette construction "par limites le long des rayons" ne coûte pas cher.

Désormais, et pour toute la suite, dès qu'on évoquera "une fonction convexe $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$", il s'agira toujours *d'une fonction convexe s.c.i. sur $\mathbb{R}^n$, finie au moins en un point.*

1.2 La construction de MOREAU

Dans des travaux datant de $1963 - 1965$, dont un article fondateur publié en 1965 ([1]), l'archétype, selon moi, d'un article élégant et profond de mathématiques, le mécanicien-mathématicien J.-J. MOREAU [2] définit et étudie l'approximée-régularisée (ou enveloppe) qui porte son nom (on dit aussi régularisée de MOREAU-YOSIDA). Voici sa construction.

Soit une fonction convexe $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ et $r > 0$ un paramètre. Alors, on définit la fonction $M_r f$ sur $\mathbb{R}^n$ de la manière suivante :

$$M_r f(x) = \inf_{u \in \mathbb{R}^n} \left\{ f(u) + \frac{r}{2} \|x - u\|^2 \right\}. \tag{2}$$

Ici, $\|.\|$ est la norme euclidienne usuelle sur $\mathbb{R}^n$. Un cadre plus général, et bien adapté à la construction de $M_r f$, est celui d'un espace de Hilbert mais nous n'avons pas besoin de cette extension pour les visées algorithmiques qui sont les nôtres. On remarquera que le rôle du paramètre r n'est pas essentiel dans la construction de $M_r f$ puisque

$$M_r f(x) = r \inf_{u \in \mathbb{R}^n} \left\{ f(u)/r + \frac{1}{2} \|x - u\|^2 \right\}. \tag{3}$$

Ainsi, si on est capable d'étudier la construction de MOREAU avec le parmètre $r = 1$, on en déduira les conclusions utiles pour n'importe quel paramètre $r > 0$. C'est ce qui se passera au paragraphe 4.

Deux choses à remarquer et retenir dans la définition (2) :

- $\frac{1}{2} \|x - u\|^2$ joue le rôle d'un élastique de rappel quadratique. Une conséquence : la borne inférieure dans la définition de $M_r f(x)$ est (finie et) *atteinte* ; on aurait pu écrire min dans (2).

- La stricte convexité de la fonction $\frac{1}{2} \|.\|^2$, alliée à la convexité de f, font qu'il n'y a qu'*un seul* point de minimisation dans le problème d'optimisation définissant $M_r f(x)$.

- La construction (2) fait apparaître un mélange de contributions de f et de $\frac{1}{2} \|.\|^2$; en effet

$$M_r f(x) = \inf_{\substack{u,\, v \in \mathbb{R}^n \\ u+v=x}} \left\{ f(u) + \frac{r}{2} \|v\|^2 \right\}. \tag{4}$$

En termes plus savants, on dit que $M_r f$ est le résultat de l'*inf-convolution* (opération notée par le symbole $\diamond$) de f et de $\frac{r}{2} \|.\|^2$, il nous arrivera d'écrire $M_r f = f \diamond \frac{r}{2} \|.\|^2$. Cette "contamination" de f par la fonction très régulière $\frac{r}{2} \|.\|^2$ aura des effets régularisants sur le résultat $M_r f$ (voir les paragraphes 3 et 4 plus loin). Noter d'ores et déjà les propriétés "différentielles" (vecteur gradient et matrice hessienne) de la fonction "noyau de régularisation" $N_r = \frac{r}{2} \|.\|^2$ utilisée : Pour tout $x \in \mathbb{R}^n$,

$$\nabla N_r(x) = rx \,;\, \nabla^2 N_r(x) = r I_n.$$

2. J.-J. MOREAU (1923 – 2014), qui a fait sa carrière à l'université de MONTPELLIER.

L'unique point de minimisation dans le problème d'optimisation (2) définissant $M_r f(x)$ porte un nom : c'est le *point proximal* de x (relativement à la fonction f (et de paramètre r)) ; il se note $\mathrm{prox}_f^r(x)$. Encore une fois, il n'y a pas vraiment de perte de généralité à faire $r = 1$, auquel cas on utilisera les notations simplifiées Mf et prox_f. D'ailleurs, en raison de la propriété énoncée en (3), certains auteurs utilisent la notation $\mathrm{prox}_{\varepsilon f}$ pour $\mathrm{prox}_f^{1/\varepsilon}$, ce qui revient au même.

L'application $\mathrm{prox}_f^r : \mathbb{R}^n \to \mathbb{R}^n$ s'appelle l'*application proximale* ou l'*opérateur proximal* (relative(-if) à la fonction f, de paramètre r).

Certains auteurs utilisent le paramètre $\frac{1}{\varepsilon}$ au lieu de r. Bien entendu, comme tout ε en mathématiques, il est destiné à être petit et à tendre vers zéro ; ce qui se lit sur r en disant que celui-ci va tendre vers l'infini.

Exemple 1. Soit C une partie convexe fermée (non vide) de $\mathbb{R}^n$ et f la fonction convexe qui vaut 0 dans C et $+\infty$ ailleurs. Cette fonction est appelée *indicatrice de* C, elle est notée parfois i_C dans la littérature spécialisée, mais il ne faut pas la confondre avec la fonction 1_C utilisée en Probabilités. De simples calculs conduisent à :

$$M_r f(x) = \frac{r}{2}(d_C(x))^2, \tag{5}$$

$$\mathrm{prox}_f^r(x) = \mathrm{p}_C(x) \text{ pour tout } x \in \mathbb{R}^n. \tag{6}$$

Ici, $d_C(x)$ désigne la distance de x à C, et $\mathrm{p}_C(x)$ la projection de x sur C. C'est cet exemple initial qui a conduit Moreau à l'appellation point proximal.

Il suffit d'observer attentivement deux ou trois exemples typiques, même avec des fonctions d'une seule variable, pour voir comment fonctionne la régularisation de Moreau. L'exemple ci-dessus, avec $C = [-1, 1]$ en est un. Un deuxième exemple est comme suit.

Exemple 2. Soit $f : x \in \mathbb{R} \mapsto f(x) = |x|$. Alors

$$M_r f(x) = \left\{ \begin{array}{l} \frac{r}{2}x^2 \text{ si } x \in [-1/r, 1/r]\,, \\ |x| - \frac{1}{2r} \text{si } |x| \geqslant 1/r, \end{array} \right. ; \tag{7}$$

$$\mathrm{prox}_f^r(x) = \left\{ \begin{array}{l} 0 \text{ si } x \in [-1/r, 1/r]\,, \\ x - 1/r \text{ si } x \geqslant 1/r, \\ x + 1/r \text{ si } x \leqslant -1/r \end{array} \right. . \tag{8}$$

Si on tient à une formule ramassée pour $\mathrm{prox}_f^r(x)$, on peut écrire

$$\mathrm{prox}_f^r(x) = [|x| - 1/r]^+ \,\mathrm{sign}(x),$$

où $\mathrm{sign}(x)$ vaut 1 si $x > 0$, -1 si $x > 0$, 0 si $x = 0$.

La fonction $\frac{1}{r} M_r f$ est la fonction de Huber, utilisée en Statistique. C'est un compromis entre le comportement quadratique (au voisinage de 0) et le comportement linéaire (quand la variable est grande).

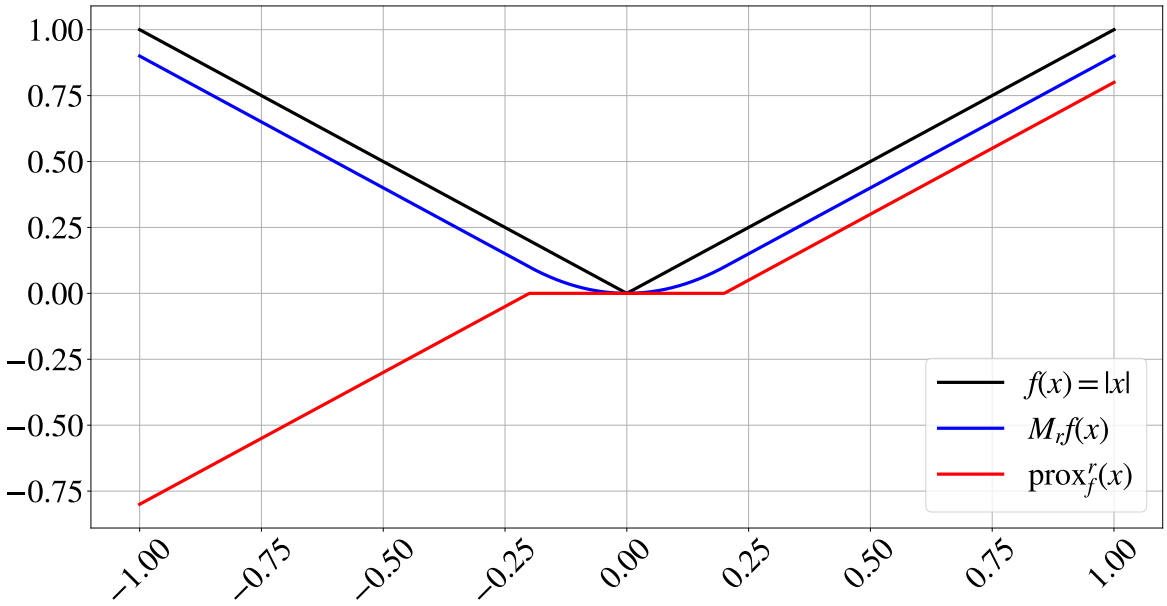

Figure 1 [3]

Exemple 3. Un troisième exemple est avec la fonction quadratique convexe (de plusieurs variables) $f : x \in \mathbb{R}^n \mapsto f(x) = \frac{1}{2}\langle Ax, x\rangle$, où A est une matrice (n, n) symétrique semidéfinie positive. Alors, pour tout $x \in \mathbb{R}^n$,

$$\begin{cases} M_r f(x) = \frac{1}{2}\langle A_r x, x\rangle, \\ \text{avec } A_r = A(I_n + \frac{1}{r}A)^{-1} = r\left[I_n - (I_n + \frac{1}{r}A)^{-1}\right]; \end{cases} \tag{9}$$

$$\text{prox}_f^r(x) = (I_n + \frac{1}{r}A)^{-1}(x). \tag{10}$$

Les calculs explicites (je ne parle pas d'approximations numériques par calculs) de $M_r f$ et de prox_f^r sont parfois possibles ; un repositoire leur est consacré ([2]), nous nous en servirons plus loin (au paragraphe 4). Du point de vue des calculs numériques, notons le caractère décomposable de $\frac{1}{2}\|x - u\|^2 = \sum_{i=1}^n \frac{1}{2}(x_i - u_i)^2$. Ainsi, si f est elle-même décomposable, $f(x) = \sum_{i=1}^n f_i(x_i)$, les calculs de $M_r f(x)$ et de $\text{prox}_f^r(x)$ reviennent à n calculs indépendants avec des fonctions f_i d'une seule variable réelle x_i. C'est ce qui se passe avec la fonction norme (importante) $f(x) = \|x\| = \sum_{i=1}^n |x_i|$.

Bref, on a compris au vu de ces quelques exemples qu'il vaut mieux avoir à traiter des fonctions convexes f "prox friendly" (comme je l'ai vu écrire par certains auteurs).

3. Cette figure, comme les suivantes, a été confectionnée par Felipe Atenas, que je remercie.

2. Rudiments d'Analyse convexe moderne

Le sujet est largement couvert dans de nombreux livres, qu'ils soient d'enseignement-recherche ou d'exercices corrigés ([3],[4], par exemple). Nous n'en prenons ici que des rudiments sur deux objets essentiels : le sous-différentiel et la conjuguée de LEGENDRE-FENCHEL.

2.1 Le sous-différentiel

Pour une fonction convexe $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, en un point x où f est finie, le *sous-différentiel* de f en x est l'ensemble, noté $\partial f(x)$, des pentes s des fonctions minorantes affines de f qui coïncident avec elle en x,

$$\partial f(x) = \{s : f(y) \geqslant f(x) + \langle s, y - x\rangle \text{ pour tout } y \in \mathbb{R}^n\}\,. \tag{11}$$

La notation ∂f peut paraître bizarre au premier abord, et elle l'est... Le concept a peu de choses à voir avec une dérivée partielle de même symbole. Un élément s de $\partial f(x)$ est appelé *sous-gradient* de f en x (la lettre s est pour rappeler *slope*).

Note historique. L'appellation "sous-gradient" est due à J.-J. MOREAU (Note aux Comptes Rendus de l'Académie des Sciences de Paris, 1963), ainsi que la traduction anglaise "subgradient" (lettre de MOREAU à l'auteur, 1993). La même année (1963), R. T. ROCKAFELLAR dans sa thèse disait plutôt *"s is a differential of f at x"*. Il est arrivé à MOREAU de parler à ses débuts de "la sous-différentielle de f en x" (comme on dit "la différentielle de f en x"), avant de basculer vers l'appellation "le sous-différentiel de f en x" qui s'est imposée depuis.

Comme on s'y attend, $\partial f(x)$ ne contient qu'un seul élément lorsque f est différentiable en x, auquel cas cet élément est $\nabla f(x)$. D'une manière générale, $\partial f(x)$ est un "paquet" convexe fermé de $\mathbb{R}^n$. Lorsque x est à l'intérieur du domaine de f, ce paquet est borné (et non vide !) ... mais il est intéressant (et primordial) d'aller voir ce qui se passe au bord du domaine de f, lorsque f y est finie. A titre d'illustration, dans l'Exemple 1, simplifié au cas unidimensionnel $C = [-1, 1]$, $\partial f(1) = [0, +\infty[$. Mais il se peut que le sous-différentiel soit vide en un point du bord du domaine.

Parmi les multiples règles de calcul relatives au sous-différentiel d'une fonction convexe, retenons simplement celle-ci : Si $g : \mathbb{R}^n \to \mathbb{R}$ est une fonction convexe différentiable, en tout x où la fonction (convexe) f est finie [4],

$$\partial(f + g)(x) = \partial f(x) + \nabla g(x).$$

En particulier,

$$\partial(f + \frac{1}{2}\,\|.\|^2)(x) = \partial f(x) + x. \tag{12}$$

Enfin, de manière évidente d'après la définition (11), un point x est un minimiseur de la fonction (convexe) f sur $\mathbb{R}^n$ si, et seulement si, $0 \in \partial f(x)$ (c'est-à-dire : 0 est *un* sous-gradient de f en x ; c'est ce qui remplace la condition $\nabla f(x) = 0$).

4. Dans la formule qui suit, $\partial f(x) + \nabla g(x)$ doit être compris comme $\{s + \nabla g(x) : s \in \partial f(x)\}$.

2.2. La conjuguée de Legendre-Fenchel

Dès qu'on a affaire à une fonction convexe f, on doit se préparer à rencontrer sa cousine *la conjuguée* f^* au sens de Legendre-Fenchel. Alors que l'Analyse traditionnelle parle de transformées de Fourier, de Laplace, ... définies par des intégrales, celle-ci, fondamentale en Optimisation, est construite via des opérations du style "prise d'un infimum, ou d'un supremum". Voici la définition de base :

$$f^* : s \in \mathbb{R}^n \mapsto f^*(s) = \sup_{x \in \mathbb{R}^n} \left(\langle s, x \rangle - f(x) \right), \tag{13}$$

le supremum en question pouvant être $+\infty$. Nous avons ainsi une nouvelle fonction convexe (comme nous les avons considérées dès le début, à savoir une fonction convexe s.c.i. sur $\mathbb{R}^n$, finie au moins en un point). Une des propriétés fondamentales de l'Analyse convexe moderne est l'effet-miroir de la conjugaison de Legendre-Fenchel : en conjuguant deux fois de suite, on retombe sur nos pieds, $(f^*)^* = f$. Moralité (à retenir) : *dès qu'on évoque une propriéte de* f, *surgit de façon claire ou demeure cachée, une propriété de* f^*.

Il se trouve que la seule fonction convexe vérifiant $f = f^*$ est cette fonction pivot $\frac{1}{2} \|.\|^2$ dont on a déjà parlé.

Note historique. Dans la littérature, il y a eu plusieurs appellations pour la transformation $f \rightsquigarrow f^*$. "De Legendre-Fenchel" est celle qui est communément admise et comprise par le plus grand nombre.

Un résultat absolument extraordinaire de Moreau, concernant la régularisation qui porte son nom, est que quand on a régularisé f, on a aussi régularisé f^*, car :

$$Mf(x) + Mf^*(x) = \frac{1}{2} \|x\|^2, \tag{14}$$

$$\mathrm{prox}_f(x) + \mathrm{prox}_{f^*}(x) = x \text{ pour tout } x \in \mathbb{R}^n. \tag{15}$$

On comprend que cela aura des conséquences sur la différentiabilité seconde de Mf et de Mf^* : elles sont deux fois différentiables ou pas en x *en même temps*. On y reviendra au §4.

La seule relation liant ∂f (paragraphe précédent) et f^* (paragraphe présent) dont nous aurons un besoin ponctuel est la suivante :

$$s \in \partial f(x) \text{ si, et seulement si, } f^*(s) + f(x) = \langle s, x \rangle. \tag{16}$$

3. Les propriété de la régularisée $M_r f$ (au premier ordre) : *a digest*

Dans ce paragraphe, sont collectés sous la dénomination de "faits" les résultats principaux à connaître sur $M_r f$ et prox_f^r. Ils sont présentés sans démonstrations, sachant qu'on peut les trouver dans divers livres (Exemple [3, Vol. 2, pages $317-330$] et même sous forme d'exercices (par exemple dans [4, Problème 7.15]).

Fait 1. *$M_r f$ est une fonction convexe, partout finie et différentiable sur $\mathbb{R}^n$.*

Donc, même pour une fonction convexe f un peu chahutée au départ (elle pourrait prendre la valeur $+\infty$!), la nouvelle fonction $M_r f$ est automatiquement partout finie et, surtout, différentiable sur $\mathbb{R}^n$.

Inutile de faire un coup de régularisation de plus, cela ne change pas fondamentalement la donne ; plus précisément :

$$M_\sigma(M_\rho f) = M_\tau f,$$

où $\frac{1}{\tau} = \frac{1}{\rho} + \frac{1}{\sigma}$ (cette vieille règle sur la résistance équivalente à deux résistances placées en parallèle).

Fait 2. *Le point proximal* $\mathrm{prox}_f^r(x)$ *est caractérisé de la manière suivante :* $\mathrm{prox}_f^r(x)$ *est le seul point* $y \in \mathbb{R}^n$ *satisfaisant la relation*

$$x \in \left(I + \frac{1}{r}\partial f\right)(y), \tag{17}$$

où I désigne l'application identité ($y \mapsto y$) de $\mathbb{R}^n$. C'est cette caractérisation qui fait écrire parfois $\mathrm{prox}_f^r(x)$ sous la forme $\mathrm{prox}_f^r(x) = \left(I + \frac{1}{r}\partial f\right)^{-1}(x)$. Cette écriture peut paraître abusive, mais pas tant que cela : bien que ∂f soit une multiapplication (c'est-à-dire : $x \mapsto$ (un sous-ensemble de $\mathbb{R}^n$)), $x \mapsto \mathrm{prox}_f^r(x)$ est bien une application (c'est-à-dire : pour un x, il n'y a qu'un seul point $\mathrm{prox}_f^r(x)$, défini sans ambiguïté).

On peut déduire - et c'est utile - des caractérisations précédentes : Pour tout $y \in \mathbb{R}^n$,

$$\left(\mathrm{prox}_f^r\right)^{-1}(y) = y + \frac{1}{r}\partial f(y). \tag{18}$$

C'est-à-dire : *tous les points qui, par l'application* prox_f^r, *sont envoyés sur* y *sont exactement ceux appartenant à l'ensemble convexe fermé* $y + \frac{1}{r}\partial f(y)$.

Conséquence de ce qui vient d'être explicité au-dessus :

Fait 3. *L'application* $prox_f^r$ *envoie* $\mathbb{R}^n$ *sur* $\mathcal{D} = \{x \in \mathrm{dom} f : \partial f(x)$ *n'est pas vide*$\}$ *(exactement, ni plus ni moins).*

Cet ensemble $\mathcal{D}$ est coincé entre $\mathrm{int}(\mathrm{dom} f)$ et $\mathrm{cl}(\mathrm{dom} f)$, il n'est pas nécessairement convexe... mais n'en est pas loin. On sent déjà que le comportement de f au bord de $\mathrm{dom} f$, savoir si ∂f y est vide ou pas, sera une information cruciale pour les propriétés de $M_r f$.

Fait 4. $M_r f$ *est différentiable sur* $\mathbb{R}^n$, *avec, pour tout* $x \in \mathbb{R}^n$,

$$\nabla M_r f(x) = r(x - \mathrm{prox}_f^r(x)), \tag{19}$$

$$\nabla M_r f(x) \in \partial f(\mathrm{prox}_f^r(x)). \tag{20}$$

Ainsi, le vecteur $\nabla M_r f(x)$ est *un* sous-gradient de f au point $\mathrm{prox}_f^r(x)$.

Autre manière d'écrire (19) :

$$\mathrm{prox}_f^r(x) = x - \frac{1}{r}\nabla M_r f(x). \tag{19bis}$$

On peut donc voir $\mathrm{prox}_f^r(x)$ comme le résultat d'une itération de la méthode du gradient appliquée à la fonction $M_r f$ en x. Cet aspect des choses sera revu au paragraphe 5.

Mentionnons juste qu'il y a d'autres caractérisations de $\text{prox}_f^r(x)$ sous forme d'inéquations, comme dans le cas de la projection d'un point sur un convexe fermé (sujet qui fut l'inspirateur de prox_f^r).

Note d'extension (pour ceux qui veulent travailler dans le contexte d'un espace de Hilbert). La différentiabilité obtenue au-dessus ne souffre d'aucune ambiguïté car nous travaillons avec une fonction convexe définie sur un espace de dimension finie. Dans le contexte d'un espace de Hilbert, les résultats d'Analyse convexe conduisent facilement à la GATEAUX-différentiabilité de $M_r f$, mais il faudrait travailler un peu plus sur les propriétés de $M_r f$ et de prox_f^r pour accéder à la FRÉCHET-différentiabilité de $M_r f$ (celle, usuelle, apprise en cours de Calcul différentiel) ... Encore une fois, dans notre contexte de dimension finie et pour des fonctions convexes, pas de souci : GATEAUX et FRÉCHET, c'est la même chose.

Fait 5. (Faisons $r = 1$ pour simplifier). *La conjuguée* $(Mf)^*$ *(de* LEGENDRE-FENCHEL*) de* Mf *n'est autre que* $f^* + \frac{1}{2}\|.\|^2$.

Ceci sert surtout à expliquer un certain nombre de choses. Une conséquence de cette simple règle : Dès qu'une propriété de f se caractérise via f^*, il est probable qu'elle se transmet à Mf.

Avec ce qui a été vu en Fait 4 et Fait 5, on peut facilement démontrer les résultats suivants :

$$\begin{cases} (Mf = Mg) \Leftrightarrow (f = g)\ ; \\ (\text{prox}_f = \text{prox}_f) \Leftrightarrow (f = g + \text{constante}). \end{cases}$$

Fait 6. (Faisons $r = 1$ pour simplifier). *Quand on a l'un, on a l'autre... , c'est-à-dire toute propriété écrite pour* f *a automatiquement et immédiatement son pendant pour* f^*. Ainsi, à partir de (14) et (15) : Pour tout $x \in \mathbb{R}^n$,

$$Mf^*(x) = \frac{1}{2}\|x\|^2 - Mf(x), \tag{21}$$

$$Mf(x) = \frac{1}{2}\|x\|^2 - Mf^*(x); \tag{21*}$$

$$\text{prox}_{f^*}(x) = x - \text{prox}_f(x)\ (= \nabla Mf(x)), \tag{22}$$

$$\text{prox}_f(x) = x - \text{prox}_{f^*}(x)\ (= \nabla Mf^*(x)). \tag{22*}$$

Retenons de ceci deux choses :

- La fonction Mf n'est pas "trop convexe", en fait "moins convexe" que $\frac{1}{2}\|.\|^2$, puisqu'il faut ajouter une autre fonction convexe, Mf^*, pour arriver à $\frac{1}{2}\|.\|^2$. Cette assertion, au demeurant un peu vague, sera explicitée un peu plus lors de l'étude de la différentiation seconde de Mf au paragraphe 4.

- L'application prox_f est un "champ de gradient" (ou "dérive d'un potentiel"), c'est-à-dire qu'elle est le gradient d'une fonction. Cela a une conséquence immédiate : en un point x où l'application $\text{prox}_f : \mathbb{R}^n \to \mathbb{R}^n$ est différentiable, la matrice jacobienne $J(\text{prox}_f)(x)$ est nécessairement *symétrique* (résultat de Calcul différentiel).

Fait 7. *L'application* $\text{prox}_f^r : \mathbb{R}^n \to \mathbb{R}^n$ *est monotone (croissante) et lipschitzienne de constante* r.

La monotonie (croissante) de prox_f^r, c'est-à-dire la propriété

$$\left\langle \mathrm{prox}_f^r(x) - \mathrm{prox}_f^r(y), x - y \right\rangle \geqslant 0 \text{ pour tout } x, y \text{ dans } \mathbb{R}^n, \tag{23}$$

n'est pas surprenante puisque prox_f^r est le gradient d'une fonction convexe. En fait, on a un peu mieux que l'inégalité (23), à savoir

$$\left\langle \mathrm{prox}_f^r(x) - \mathrm{prox}_f^r(y), x - y \right\rangle \geqslant \frac{1}{r} \left\| \mathrm{prox}_f^r(x) - \mathrm{prox}_f^r(y) \right\|^2. \tag{23 bis}$$

Le caractère lipschitzien de l'application prox_f^r,

$$\left\| \mathrm{prox}_f^r(x) - \mathrm{prox}_f^r(y) \right\| \leqslant r \left\| x - y \right\| \text{ pour tout } x, y \text{ dans } \mathbb{R}^n, \tag{24}$$

est intéressant car il induit, par exemple, que prox_f^r est différentiable presque partout sur $\mathbb{R}^n$, c'est-à-dire sauf sur un ensemble de mesure de Lebesgue nulle (cela résulte d'un théorème de Rademacher).

A retenir : $M_r f$ (comme $M_r f^*$) est une fonction convexe différentiable et à gradient lipschitzien sur $\mathbb{R}^n$.

Fait 8. *Les fonctions f et Mf ont le même comportement à l'infini.*

Expliquons ce que cela veut dire. Pour une fonction g telle que celles que nous considérons (c'est-à-dire une fonction convexe s.c.i. sur $\mathbb{R}^n$, finie au moins en un point), le comportement à l'infini est décrit par la fonction suivante :

$$g'_\infty : d \in \mathbb{R}^n \mapsto g'_\infty(d) = \lim_{t \to +\infty} \frac{g(a + td) - g(a)}{t} \quad (\in \mathbb{R} \cup \{+\infty\}),$$

où a est un point du domaine de g, fixé une fois pour toutes. Il est étonnant - et intéressant - que la limite $g'_\infty(d)$ ne dépende pas du point a choisi. La notation $g'_\infty(d)$, introduite pour la première fois dans [3], suggère bien que $g'_\infty(d)$ est "la pente à l'infini de g dans la direction d".

Pour la situation qui nous concerne, nous avons :

$$(M_r f)'_\infty(d) = f'_\infty(d) \text{ pour tout } d \text{ dans } \mathbb{R}^n. \tag{25}$$

Ainsi, dans l'Exemple 1, avec comme fonction f l'indicatrice de $C = [-1, 1]$, nous avons

$$(M_r f)'_\infty(d) = f'_\infty(d) = +\infty \text{ partout sauf en } d = 0 \text{ où cela vaut } 0.$$

Notons que l'ensemble des minimiseurs de g sur $\mathbb{R}^n$ est un ensemble (convexe) fermé borné lorsque $g'_\infty(d) > 0$ pour toute direction non nulle d.

Fait 9. *$M_r f$ est bien une approximation (par dessous) de f.* Plus précisément : Pour tout $x \in \mathbb{R}^n$,

$$M_r f(x) \text{ tend en croissant vers } f(x) \text{ quand } r \to +\infty. \tag{26}$$

Il faut bien saisir la généralité du résultat (26) : Si $x \in \mathrm{dom} f$, $M_r f(x)$ tend vers la valeur (finie) $f(x)$; mais également, $M_r f(x)$ tend vers $+\infty$ quand $x \notin \mathrm{dom} f$.

Concernant prox_f^r, nous avons le résultat suivant : Si $x \in \text{dom} f$ (et uniquement dans ce cas),

$$\text{prox}_f^r(x) \to x \text{ quand } r \to +\infty. \tag{27}$$

Ainsi, la fonction $M_r f$ est à la fois une régularisation (*cf.* Fait **1**) et une approximation (*cf.* Fait **9**) de la fonction f. Il n'est donc pas inapproprié d'appeler $M_r f$ "*la régularisée-approximée de* MOREAU *de* f ".

Fait(s) 10. - *Concernant les bornes inférieures et les solutions des problèmes de minimisation de f et de $M_r f$. Nous avons* :

$$\inf_{x\in\mathbb{R}^n} f(x) = \inf_{x\in\mathbb{R}^n} M_r f(x) \; ; \tag{28}$$

$$(x \text{ minimise } f \text{ sur } \mathbb{R}^n) \Leftrightarrow (x \text{ minimise } M_r f \text{ sur } \mathbb{R}^n)\,. \tag{29}$$

L'égalité (28) doit être comprise dans $\mathbb{R} \cup \{-\infty\}$, il se peut que les fonctions f et $M_r f$ ne soient pas bornées inférieurement sur $\mathbb{R}^n$. L'équivalence (29) ne stipule pas qu'il y a des solutions aux problèmes de minimisation de f et de $M_r f$, il se peut très bien qu'il n'y ait aucun point minimiseur.

- *Caractérisations diverses des minimiseurs de f ou de $M_r f$*. Les assertions suivantes sont équivalentes :

$$(i) \;\; x \text{ minimise } f \text{ (ou } M_r f \text{) sur } \mathbb{R}^n \; ; \tag{30}$$

$$(ii) \;\; \text{prox}_f^r(x) = x \; ; \tag{31}$$

$$(iii) \quad f(\text{prox}_f^r(x)) = f(x) \; ; \tag{32}$$

$$(iv) \;\; M_r f \; (x) = f(x). \tag{33}$$

L'équivalence entre (i) et (iv) est intéressante à noter : $M_r f \leqslant f$ de manière générale, et $M_r f$ et f ne se "touchent" qu'aux points (communs) de minimisation.

L'équivalence entre (i) et (ii) (ou sa voisine (iii)) explique qu'on ait pu penser à un algorithme de "proximations successives" du style (on fait $r = 1$ pour simplifier) :

$$\left\{ \begin{array}{c} x_0 \text{ point d'initialisation quelconque}\,; \\ x_{k+1} = \text{prox}_f(x_k). \end{array} \right.$$

De fait, si les ensembles de sous-niveau $\{x : f(x) \leqslant \ell\}$ sont bornés, l'algorithme décrit au-dessus produit une suite $(x_k)_k$ de points avec les propriétés suivantes :

$$\left\{ \begin{array}{c} (i) \text{ La suite } (f(x_k))_k \text{ est décroissante avec } k \text{ et } \lim_{k\to+\infty} f(x_k) = \min_{x\in\mathbb{R}^n} f(x)\,; \\ (ii) \text{ La suite } (x_k)_k \text{ converge vers } \textit{un} \text{ point minimiseur de } f \text{ sur } \mathbb{R}^n. \end{array} \right. \tag{34}$$

L'algorithme décrit au-dessus est rustre, mais on peut en garder l'esprit pour expliquer un certain nombre d'algorithmes de type proximal. Pour des premières propriétés de cet algorithme et des commentaires historiques, voir [3, Vol. 2, pages $318-330$].

4. Les propriétés de la régularisée $M_r f$ au deuxième ordre : ce qu'on peut espérer et obtenir

On est tenté de dire - et j'ai eu l'occasion de le lire - ceci : si la fonction convexe f est deux fois différentiable (voire de classe $\mathcal{C}^\infty$) sur $\mathrm{int}(\mathrm{dom} f)$, c'est-à-dire sur le plus grand ensemble où elle peut l'être, alors $M_r f$ est deux fois différentiable... Ceci est clairement faux, il suffit pour s'en assurer de considérer la fonction f indicatrice de $[-1,1]$, laquelle conduit à une MOREAU-régularisée $M_r f$ qui n'est pas deux fois différentiable aux points -1 et 1 (voir Exemple 1). Pourtant le résultat est vrai si f est supposée à valeurs finies (partout), c'est-à-dire si $\mathrm{dom} f = \mathbb{R}^n$. Les études à ce sujet sont anciennes, elles datent des années $1994-1997$; *cf.* par exemple les travaux [5], [6], [7]. Nous en reprenons ici l'essentiel sous forme synthétique, en les améliorant légèrement au passage ; le Corollaire 2 est un exemple d'amélioration de l'existant.

4.1 Ce que nous dit le Calcul différentiel classique

Nous avons appris, quand nous étions petits, que la différentiabilité (ou dérivabilité) d'une fonction numérique de la variable réelle f en x_0 équivaut au développement du 1^{er} ordre suivant : $f(x_0+h) = f(x_0) + a.h + o(h)$; dans ce cas $a = f'(x_0)$. Il n'en est plus de même avec la différentiation d'ordre 2 (ou d'ordre supérieur en général) : Si f, fonction dérivable, admet en x_0 le développement du $2^{ème}$ ordre suivant

$$f(x_0+h) = f(x_0) + f'(x_0).h + \ell.\frac{h^2}{2} + o(h^2), \tag{35}$$

il n'est pas sûr que f soit deux fois différentiable en x_0 ; on ne peut conclure que $\ell = f''(x_0)$. Un contre-exemple classique est la fonction f qui à x associe $f(x) = x^3 \sin(1/x)$ si $x \neq 0$, et 0 si $x = 0$. On a bien une fonction partout différentiable, avec un développement du type (35) en $x_0 = 0$ clair, $f(0+h) = h^3 \sin(1/h) = o(h^2)$, mais f n'est pas 2 fois différentiable en 0.

Cette difficulté ne se produit pas lorsque f est convexe. Voici le résultat dans ce cas, qui fait l'objet d'un excellent exercice d'Analyse, même si sa démonstration n'est pas immédiate (il faut utiliser la monotonie de f') : Soit f une fonction convexe (ou concave) différentiable sur un intervalle I, soit $x_0 \in I$ en lequel on a un développement du type (35) ; alors, $f''(x_0)$ existe et vaut ℓ. La fonction convexe et différentiable $f : x \in \mathbb{R} \mapsto f(x) = |x|^{3/2}$ est une illustration de notre propos, sous sa forme négative : elle n'est pas 2 fois différentiable en 0 parce que, justement, un développement à l'ordre 2 comme en (35) n'est pas possible.

4.2 Un peu plus sur la différentiation d'ordre 1 et 2 d'une fonction convexe

Reprenons les choses à la base, avec des fonctions $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$.

Définitions.

- On dit que f admet en x_0 un développement de TAYLOR-YOUNG à l'ordre 2 lorsque : f est différentiable en x_0 et il existe une matrice symétrique (notée $A^2 f(x_0)$) telle que

$$f(x_0+h) = f(x_0) + \langle \nabla f(x_0), h\rangle + \frac{1}{2}\left\langle A^2 f(x_0)h, h\right\rangle + o(\|h\|^2). \tag{36}$$

La notation A est pour A. D. Alexandroff qui, en 1939, publie un article démontrant que ceci a lieu pour presque tout x_0 lorsque la fonction f est convexe, c'est-à-dire en dehors d'un ensemble de mesure (de Lebesgue) nulle.

Ceci est plus faible que la différentiabilité (usuelle) d'ordre 2 de f en x_0. Mais pour une fonction convexe cela revient à peu près au même : *Si la fonction convexe est (une fois) différentiable dans un voisinage de* x_0, *on a* (36) *si et seulement si* f *est deux fois différentiable en* x_0, *avec* $\nabla^2 f(x_0) = A^2 f(x_0)$. Ceci est loin d'être évident à démontrer ([8, Corollary 2.13]).

- Suivant R. T. Rockafellar et F. Mignot (dans des travaux publiés en 1976), on dit que la multiapplication ∂f (pour une fonction convexe f) est différentiable en x_0 si, d'abord f est différentiable en x_0, et ensuite il existe une matrice $D^2 f(x_0)$ (que l'on pourrait noter aussi $J(\partial f)(x_0)$) telle que

$$\left\{ \begin{array}{c} \|\partial f(x) - \nabla f(x_0) - D^2 f(x_0)(x - x_0)\| = o\left(\|x - x_0\|\right) \\ \text{(un } o(.) \text{ uniforme en les } s \in \partial f(x)). \end{array} \right. \tag{37}$$

Détaillons ce que cela signifie : Pour tout $\varepsilon > 0$, il existe $\delta > 0$ tel que

$$\left\{ \begin{array}{c} (\|x - x_0\| \leqslant \delta \text{ et } s \in \partial f(x)) \Rightarrow \\ (\|s - \nabla f(x_0) - D^2 f(x_0)(x - x_0)\| \leqslant \varepsilon \|x - x_0\|). \end{array} \right. \tag{37bis}$$

Mignot démontre dans un article publié en 1976 que ∂f est différentiable en presque tout x_0. Dans [8, Proposition 2.11], je démontre que cette matrice $D^2 f(x_0)$ est nécessairement symétrique et semidéfinie positive (il y a un peu de travail pour cela, mais la démonstration fait appel à des techniques usuelles de Calcul différentiel). Dans [8, Corollary 2.12], je démontre aussi le résultat "logique et attendu" que voici : f *admet en* x_0 *un développement de* Taylor-Young *à l'ordre* 2 *si, et seulement si,* ∂f *est différentiable en* x_0. *Bref,* $A^2 f(x_0) = D^2 f(x_0)$. Par abus de langage, on dira donc que "f (convexe) est deux fois A-différentiable en x_0" lorsqu'on a (36) ou (37), et on gardera la notation $A^2 f(x_0)$ (qui, rappelons-le, est $\nabla^2 f(x_0)$ si f est deux fois différentiable (au sens usuel) en x_0).

Un mot sur le pendant de (36) pour la fonction conjuguée f^* : Si on a le développement (36) en x_0, on a quelque chose de similaire pour f^* en $s_0 = \nabla f(x_0)$, à condition de supposer $A^2 f(x_0)$ inversible ([3, Vol. 2, page 89]) :

$$\left\{ \begin{array}{c} f^*(s_0 + p) = f^*(s_0) + \langle x_0, p\rangle + \frac{1}{2}\left\langle \left[A^2 f(x_0)\right]^{-1} p, p \right\rangle + o(\|p\|^2), \\ \text{(avec } x_0 = \nabla f^*(s_0), \text{ rappelons-le).} \end{array} \right. \tag{36*}$$

Dans la suite, on choisira, suivant le cas, le développement (36) ou (37) ; la formulation (36) est plus "palpable", la formulation (37) est plus "puissante" (notamment dans les démonstrations).

4.3 La différentiabilité seconde de $M_r f$ à partir de celle de f

Pour alléger les notations, et sans perte de généralité, nous faisons désormais $r = 1$ dans le processus de régularisation de Moreau.

Rappelons (*cf.* la relation (19)), que la 2-différentiabilité (ou différentiabilité d'ordre 2) de la fonction Mf en x_0, c'est-à-dire la différentiabilité de l'application ∇Mf en x_0, équivaut à la différentiabilité de l'application prox_f en x_0, avec

$$\nabla^2 Mf(x_0) = I_n - J(\mathrm{prox}_f)(x_0). \tag{38}$$

Cette relation confirme bien que $J(\mathrm{prox}_f)(x_0)$ est une matrice symétrique, comme nous l'annoncions précédemment (à la fin de Fait 6).

Le résultat-clé liant la différentiabilité d'ordre 2 de f et celle de Mf est comme suit.

Théorème 1. *Si f est deux fois A-différentiable en $\mathrm{prox}_f(x_0)$, alors Mf est deux fois différentiable (au sens usuel) en x_0, avec*

$$\nabla^2 Mf(x_0) = I_n - \left[I_n + A^2 f(\mathrm{prox}_f(x_0))\right]^{-1}. \tag{39}$$

La démonstration figure en Annexe à la fin de notre présentation.

Notons tout de suite que la formule (39) reste valable même si $A^2 f(\mathrm{prox}_f(x_0))$ est singulière (c'est-à-dire, n'est pas inversible). Comme $A^2 f(\mathrm{prox}_f(x_0))$ est semidéfinie positive, $I_n + A^2 f(\mathrm{prox}_f(x_0))$ est définie positive, donc inversible.

Même si f est 2 fois différentiable (au sens classique) partout où cela est possible, c'est-à-dire au mieux sur $\mathrm{int}(\mathrm{dom} f)$, le résultat au-dessus montre que cela n'induit pas que Mf est partout 2 fois différentiable : *cela dépend des points* $\mathrm{prox}_f(x)$, *si ceux-ci tombent dans la zone de 2-différentiabilité de f ou pas*. Le cas où $\mathrm{prox}_f(x)$ tombe sur un point-frontière de $\mathrm{dom} f$, celui-ci étant toutefois un point où le sous-différentiel de f n'est pas vide, est particulièrement intéressant ; il sera considéré plus bas.

La version "duale" du Théorème 1 consiste à écrire le même résultat sur la conjuguée f^*, en se souvenant que $\nabla^2 Mf(x_0) = I_n - \nabla^2 Mf^*(x_0)$.

Théorème 1*. *Si f^*est deux fois A-différentiable en $\mathrm{prox}_{f^*}(x_0)$ $(= x_0 - \mathrm{prox}_f(x_0))$, alors Mf est deux fois différentiable (au sens usuel) en x_0, avec*

$$\nabla^2 Mf(x_0) = \left[I_n + A^2 f^*(\mathrm{prox}_{f^*}(x_0))\right]^{-1}. \tag{39*}$$

Les deux théorèmes au-dessus, le Théorème 1 et le Théorème 1*, ne conduisent pas à la 2-différentiabilité de Mf aux mêmes points x_0 ; l'exemple ci-dessous en est une illustration.

Exemple 4. Soit f la fonction indicatrice de $[-1, 1]$. Alors, f est trivialement 2 fois différentiable sur $\mathrm{int}(\mathrm{dom} f) =]-1, 1[$, mais

$$x \mapsto Mf(x) = \begin{cases} 0 \text{ si } x \in [-1, 1]\,, \\ \frac{1}{2}(x-1)^2 \text{ si } x \geqslant 1, \\ \frac{1}{2}(x+1)^2 \text{ si } x \leqslant -1, \end{cases}$$

n'est pas partout 2 fois différentiable. Le Théorème 1 est applicable en $x_0 \in]-1, 1[$ puisque, dans ce cas, $\mathrm{prox}_f(x_0)$, qui vaut x_0, est dans la zone de 2-différentiabilité de f. Lorsque $x_0 \notin]-1, 1[$, $\mathrm{prox}_f(x_0) = \pm 1$ et tout peut se passer : Mf peut être 2 fois différentiable en x_0 comme Mf peut ne pas être 2 fois différentiable en x_0.

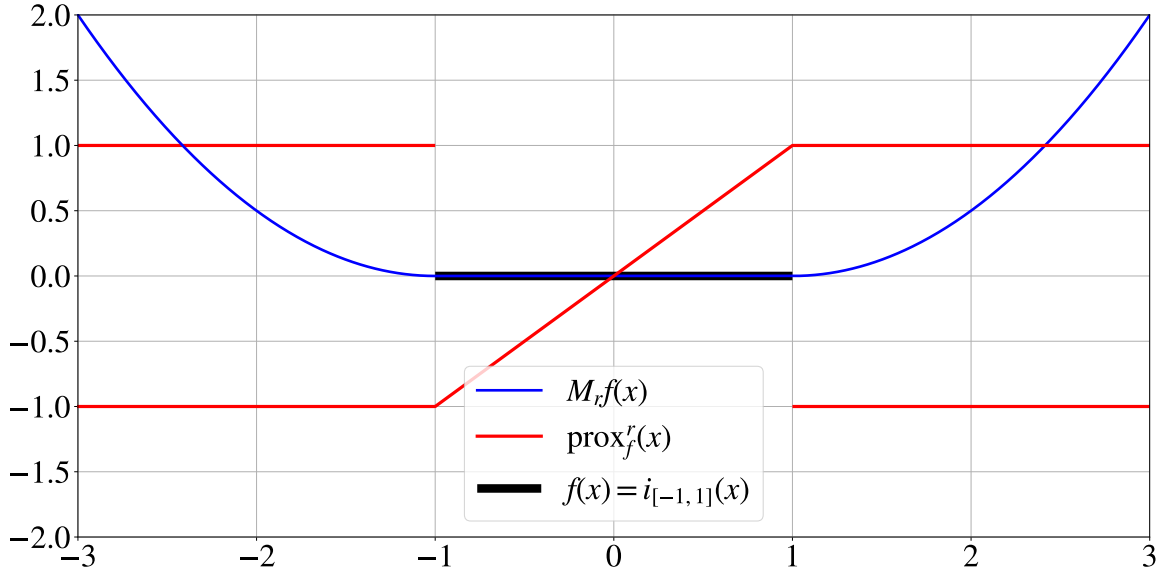

Figure 2

La version duale de cet exemple est comme suit. On a $f^* = |.|$. Alors, f^* est clairement 2 fois différentiable sur $\mathbb{R}^*$ (c'est-à-dire $\mathbb{R}$ privé du point 0), mais

$$x \mapsto Mf^*(x) = \begin{cases} \frac{1}{2}x^2 \text{ si } x \in [-1,1]\,, \\ x - \frac{1}{2} \text{ si } x \geqslant 1, \\ -x - \frac{1}{2} \text{ si } x \leqslant -1, \end{cases}$$

n'est pas partout 2 fois différentiable, exactement comme (et aux mêmes points ± 1 que) Mf. Le Théorème 1^* est applicable (à f^*) en $x_0 \notin]-1,1[$ puisque, dans ce cas, $\text{prox}_{f^*}(x_0)$, qui est différent de 0, est dans la zone de 2-différentiabilité de f^*. Lorsque $x_0 \in [-1,1]$, $\text{prox}_{f^*}(x_0) = 0$ et tout peut se passer : Mf^* peut être 2 fois différentiable en x_0 comme Mf^* peut ne pas être 2 fois différentiable en x_0.

En cumulant les deux résultats, on est arrivé à la 2-différentiabilité de Mf et Mf^* partout sauf peut-être en ± 1, et c'est effectivement ce qu'on pouvait faire de mieux.

Tirons quelques corollaires du résultat du Théorème 1.

Corollaire 2. *Soit $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ une fonction convexe vérifiant l'hypothèse suivante :*

$$(\mathcal{H}) \quad \begin{cases} * \ f \text{ est 2 fois différentiable sur } \text{int}(\text{dom} f), \\ * \ \partial f(x) \text{ est vide en tout point frontière de } \text{dom} f. \end{cases}$$

Alors Mf est 2 fois différentiable partout sur $\mathbb{R}^n$.

Notons que la deuxième partie de l'hypothèse ($\mathcal{H}$) ne concerne que les points qui sont à la fois sur la frontière de dom$f$ et dans dom$f$ (puisque, par définition, $\partial f(x)$ est vide lorsque $x \notin \text{dom}f$). La démonstration du Corollaire 2 est très simple à partir du résultat du Théorème 1. En effet, pour tout $x \in \mathbb{R}^n$, $\text{prox}_f(x)$ est un point où le sous-différentiel de f n'est pas vide. Or, de par l'hypothèse ($\mathcal{H}$), un tel point ne peut être qu'à l'intérieur de dom$f$, zone où précisément f a été supposée 2 fois différentiable.

Corollaire 3. *Soit $f : \mathbb{R}^n \to \mathbb{R}$ une fonction convexe 2 fois différentiable sur $\mathbb{R}^n$ (resp. de classe $\mathcal{C}^2$ sur $\mathbb{R}^n$). Alors Mf est 2 fois différentiable sur $\mathbb{R}^n$ (resp. de classe $\mathcal{C}^2$ sur $\mathbb{R}^n$).*

Pour la 2-différentiabilité de Mf, le Corollaire 2 s'applique trivialement puisque la frontière du domaine de f est vide.

Voyons pour le caractère $\mathcal{C}^2$. On a $A^2f(.) = \nabla^2 f(.)$ qui est continue par hypothèse, l'application $\text{prox}_f(.)$ qui est continue (*cf.* (24)), et la formule :

$$\nabla^2 Mf(x) = I_n - \left[I_n + \nabla^2 f(\text{prox}_f(x))\right]^{-1}. \tag{40}$$

Ensuite, il suffit d'observer que $\nabla^2 M_f(.)$ résulte de l'enchaînement (ou composition) d'applications continues.

Dans le cas de fonctions f d'une seule variable, la formule (40) prend une forme simplifiée :

$$(Mf)''(x) = \frac{f''(\text{prox}_f(x))}{1 + f''(\text{prox}_f(x))}. \tag{41}$$

Nous aurons l'occasion de l'illustrer à plusieurs reprises.

Exemple 5. Soit $f : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ définie par $f(x) = -\ln(x)$ si $x > 0$, $+\infty$ sinon. Alors l'hypothèse ($\mathcal{H}$) du Corollaire 2 est satisfaite, et donc Mf est 2 fois différentiable sur $\mathbb{R}$. Cet exemple est intéressant car il montre que l'on pourrait modifier f en en faisant une fonction affine sur un sous-intervalle de $]0, +\infty[$ ou en modifiant son comportement quand $x \to +\infty$, à condition bien sûr de préserver la 2-différentiabilité sur $]0, +\infty[$, sans que cela détruise la 2-différentiabilité de Mf.

Si l'on tient à avoir des calculs explicites pour la fonction f de cet exemple, les voici :

$$\left\{\begin{array}{c} \text{prox}_f(x) = \frac{x+\sqrt{x^2+4}}{2}, \\ Mf(x) = -\ln(\frac{x+\sqrt{x^2+4}}{2}) + \frac{1}{4}(x^2 + 2 - x\sqrt{x^2+4}), \\ (Mf)'(x) = \frac{x-\sqrt{x^2+4}}{2}, \\ (Mf)''(x) = \frac{1}{2}(\frac{1}{\sqrt{x^2+4}} - x). \end{array}\right. \tag{42}$$

On vérifie sur cet exemple que, d'une part $(Mf)'(x) = x - \text{prox}_f(x)$, et que, d'autre part, $(Mf)''(x) = \frac{f''(\text{prox}_f(x))}{1+f''(\text{prox}_f(x))}$ (formule (41)).

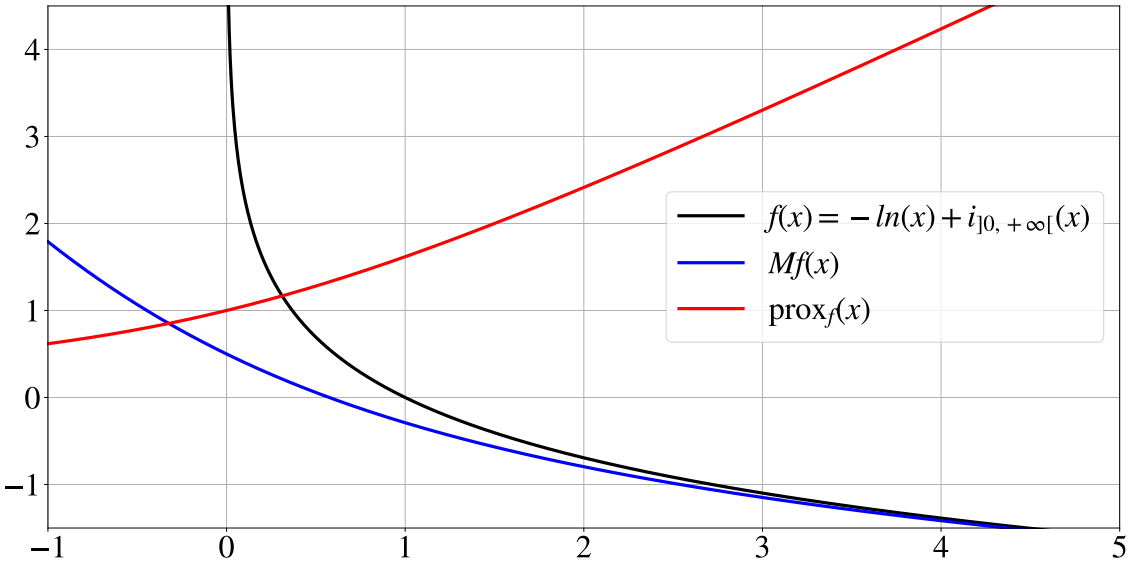

Figure 3

Exemple 6. Soit $f : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ définie par

$$f(x) = \begin{cases} -\frac{1}{2}x^2 - \sqrt{1-x^2} \text{ si } x \in [-1,1]\,, \\ +\infty \text{ sinon.} \end{cases}$$

Cet exemple est intéressant au sens où aux points-frontières ± 1 de $\text{dom} f = [-1,1]$, le sous-différentiel de f est vide. Ainsi, l'hypothèse $(\mathcal{H})$ du Corollaire 2 est vérifiée, et la fonction Mf est 2 fois différentiable partout sur $\mathbb{R}$. Ici encore, on peut mener jusqu'au bout des calculs explicites, les voici :

$$\begin{cases} \text{prox}_f(x) = \frac{x}{\sqrt{1+x^2}}, \\ Mf(x) = \frac{1}{2}x^2 - \sqrt{1+x^2}, \\ (Mf)'(x) = x - \frac{x}{\sqrt{1+x^2}}, \\ (Mf)''(x) = 1 - \frac{1}{(1+x^2)^{3/2}}. \end{cases} \tag{43}$$

On vérifie sur cet exemple que, d'une part $(Mf)'(x) = x - \text{prox}_f(x)$ et que, d'autre part, $(Mf)''(x) = \frac{f''(\text{prox}_f(x))}{1+f''(\text{prox}_f(x))}$ (formule (41)).

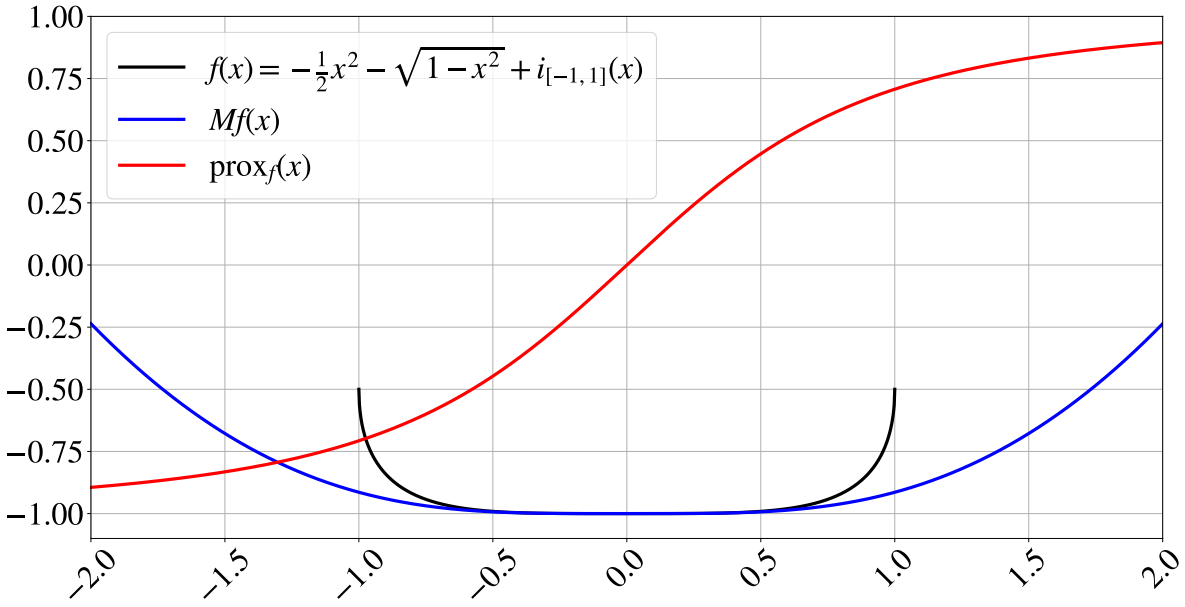

Figure 4

Exemple 7. Soit $f : \mathbb{R} \to \mathbb{R}$ définie par $f(x) = \cosh(x) - \frac{x^2}{2}$. On sait, d'après le Corollaire 3, que M_f sera 2 fois différentiable sur $\mathbb{R}$, et même avec une "courbure" majorée par celle de f et par 1 :

$$(Mf)''(x) = \frac{f''(\mathrm{prox}_f(x))}{1 + f''(\mathrm{prox}_f(x))} \leqslant \min(1, f''(\mathrm{prox}_f(x)). \tag{44}$$

On reprendra ce type de majoration de manière plus générale dans le Corollaire 4 plus loin.

Ici aussi, des calculs explicites peuvent être menées à leur terme, se souvenant que $\cosh(\mathrm{Arg\,sinh}(x)) = \sqrt{1+x^2}$ et $\mathrm{Arg\,sinh}'(x) = \frac{1}{\sqrt{1+x^2}}$. Voici ce que cela donne :

$$\begin{cases} \mathrm{prox}_f(x) = \mathrm{Arg\,sinh}(x), \\ Mf(x) = \sqrt{1+x^2} + \frac{x^2}{2} - x\mathrm{Arg\,sinh}(x), \\ (Mf)'(x) = x - \mathrm{Arg\,sinh}(x), \\ (Mf)''(x) = 1 - \frac{1}{\sqrt{1+x^2}}. \end{cases} \tag{45}$$

On vérifie également sur cet exemple que, d'une part $(Mf)'(x) = x - \mathrm{prox}_f(x)$ et que, d'autre part, $(Mf)''(x) = \frac{f''(\mathrm{prox}_f(x))}{1+f''(\mathrm{prox}_f(x))}$ (formule (41)).

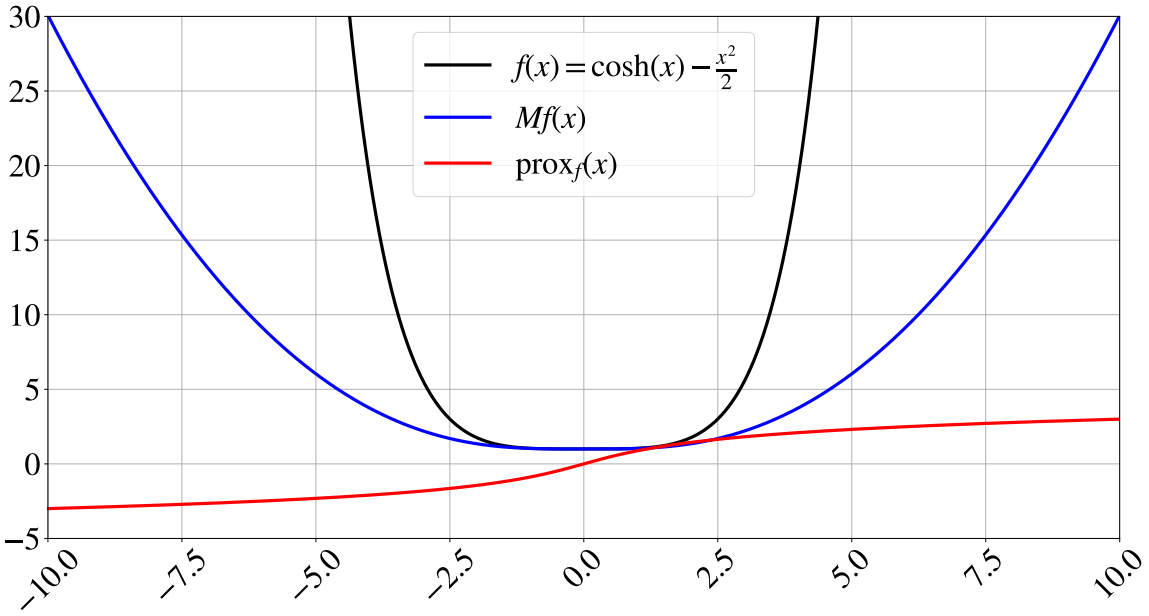

Figure 5

Exemple 8. Soit $f : \mathbb{R} \to \mathbb{R}$ définie par $f(x) = \exp(x)$. Pour une fonction aussi basique que celle-là, il est intéressant d'avoir une forme explicite de Mf. Cela est possible grâce à la fonction particulière dite de Lambert. La fonction ℓ de Lambert est définie sur $[-\frac{1}{e}, +\infty[$ comme suit : $\ell(y)$ est l'unique solution de l'équation (en x) : $x \exp(x) = y$. Par exemple, $\ell(e^x)e^{\ell(x)} = e^x$, soit $\ell(e^x) = e^{x-\ell(x)}$. Nous n'utiliserons $\ell(y)$ que pour les $y > 0$. Dans ce contexte, nous avons ([9, page 41]) :

$$\ell'(y) = \frac{e^{-\ell(y)}}{1+\ell(y)} = \frac{\ell(y)}{[1+\ell(y)]y}.$$

Les résultats suivants sont alors faciles à obtenir, ils étaient d'ailleurs déjà disponibles dans ([4, Exercice 7.9]) :

$$\begin{cases} \mathrm{prox}_f(x) = x - \ell(e^x), \\ Mf(x) = \frac{1}{2}\left[\ell(e^x)\right]^2 + \ell(e^x), \\ (Mf)'(x) = \ell(e^x), \\ (Mf)''(x) = \frac{\ell(e^x)}{1+\ell(e^x)}. \end{cases} \tag{46}$$

Ici encore, nous avons l'illustration du fait que $(Mf)'(x) = x - \mathrm{prox}_f(x)$ et que $(Mf)''(x) = \frac{f''(\mathrm{prox}_f(x))}{1+f''(\mathrm{prox}_f(x))}$ (formule (41)).

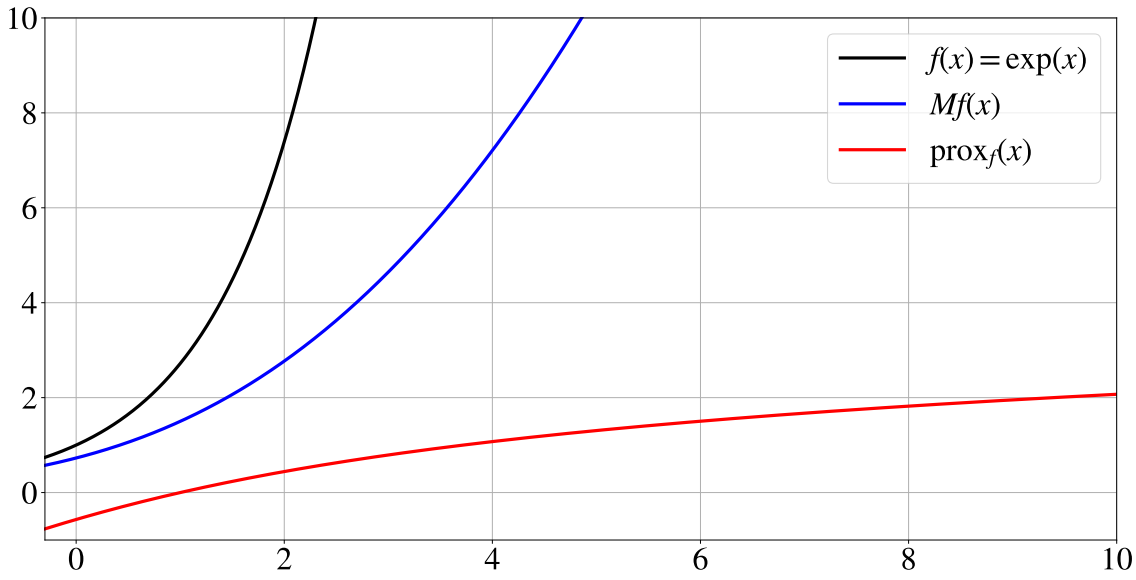

Figure 6

On peut multiplier des illustrations avec des fonctions d'une seule variable, comme dans les Exemples $5-8$ au-dessus, dans la mesure où les calculs explicites de $\mathrm{prox}_f(x)$ sont disponibles. Pour cela, on pourra consulter le repositoire [2].

Exemple 9 (Exemple 3 revisité). Avec l'exemple des formes quadratiques f, c'est le moment de donner des variantes de la formule (39) et ses cousines. Soit $f : x \in \mathbb{R}^n \to \mathbb{R}$ définie par $f(x) = \frac{1}{2}\langle Ax, x\rangle$ où A est une matrice (symétrique) semidéfinie positive. Alors, en définissant $S = I_n - [I_n + A]^{-1}$, on a :

$$\begin{cases} \mathrm{prox}_f(x) = [I_n + A]^{-1}x = x - Sx, \\ Mf(x) = \frac{1}{2}\langle Sx, x\rangle, \\ \nabla Mf(x) = Sx, \\ \nabla^2 Mf(x) = S. \end{cases} \tag{47}$$

C'est le prototype de la formule (39).

Comme A est semidéfinie positive, il se trouve que

$$(S =)\ I_n - [I_n + A]^{-1} = A\,[I_n + A]^{-1} \tag{48.1}$$

$$= [I_n + A]^{-1}\,A = A - A\,[I_n + A]^{-1}\,A. \tag{48.2}$$

$$= (\text{si } A \text{ est inversible})\ \left[I_n + A^{-1}\right]^{-1}. \tag{48.3}$$

C'est un peu astucieux à démontrer... Il faut utiliser $UU^{-1} = U^{-1}U = I_n$ avec plusieurs matrices différentes U (formées avec A et I_n). La matrice S explicitée en (48.1)-(48.3) est

parfois appelée dans la littérature la *somme parallèle* de A et de I_n. Clairement, Ker $S =$ Ker A, Im $S =$ Im A.

Pour notre utilisation ici, (48.1)-(48.3) donnent quatre ou cinq variantes de l'expression de $\nabla^2 Mf(x_0)$ dans la formule (39).

On peut donc réinterpréter le résultat du Corollaire 2 (ou 3) en disant ceci : *La forme quadratique associée à la matrice hessienne de la* MOREAU*-régularisée de f en $x_0 \in \mathbb{R}^n$ (soit $\nabla^2(f \diamond \frac{1}{2}\|.\|^2)(x_0)$) n'est autre que la* MOREAU*-régularisée de la forme quadratique associée à matrice hessienne de f en* $\text{prox}_f(x_0) \in \text{int}(\text{dom} f)$ (*soit* $\nabla^2 f(\text{prox}_f(x_0))$; ce qui est fort élégant.

Les différentes formes matricielles vues en (48.1)-(48.3) conduisent à préciser un peu plus la relation entre $\nabla^2 Mf(x)$ et $\nabla^2 f(\text{prox}_f(x))$.

Corollaire 4. *Plaçons-nous sous les hypothèses du Corollaire 2. Alors, pour tout* $x_0 \in \mathbb{R}^n$ *:*

$$* \quad \nabla^2 Mf(x_0) \preccurlyeq \nabla^2 f(\text{prox}_f(x_0)) \ ; \tag{49.1}$$

$$* \quad \nabla^2 Mf(x_0) \preccurlyeq I_n. \tag{49.2}$$

$$* \begin{cases} \text{Si } \lambda_1, ..., \lambda_n \text{ désignent les valeurs propres de } \nabla^2 f(\text{prox}_f(x_0)), \\ \text{celles de } \nabla^2 Mf(x_0) \text{ sont } \frac{\lambda_1}{1+\lambda_1}, ..., \frac{\lambda_n}{1+\lambda_n}. \\ (\text{ainsi, } \frac{\lambda_i}{1+\lambda_i} \leqslant \min(1, \lambda_i)). \end{cases} \tag{49.3}$$

Le résultat principal jusqu'à présent est que l'on a pu démontrer la 2-différentiabilité de f en x_0 dès lors que $\text{prox}_f(x_0)$ "tombe" dans la zone de 2-différentiablité de f. Question : *Qu'en est-il sinon ?* Commençons avec deux exemples très simples (le premier vu en Exemple 2). Si f *n'est pas deux fois* 2*-différentiable en* $\text{prox}_f(x_0)$ (par exemple, n'est pas simplement différentiable), le Théorème 1 ne peut s'appliquer à x_0 ni à tous les x qui ont été "contaminés", ceux de $x_0 + \partial f(x_0)$ (puisqu'ils donnent le même $\text{pro}x_f(x_0)$!). Considérons par exemple la fonction $x \in \mathbb{R} \mapsto f(x) = |x|$. Dans les cas où $\text{prox}_f(x_0) = 0$, le Théorème 1 ne s'applique pas ; de fait tous les x "contaminés" sont ceux de $[-1, 1]$; et en effet Mf n'est pas 2-différentiable en -1 et en 1..., mais pourtant Mf est 2-différentiable sur $]-1, 1[$ ($Mf(x)$ y vaut $\frac{1}{2}x^2$).

La version duale de cet exemple est la fonction f de l'Exemple 4. Pour $x_0 \geqslant 1$, on a $\text{prox}_f(x_0) = 1$, f n'y est pas différentiable. Le Théorème 1 ne s'applique pas ; de fait tous les x "contaminés" sont ceux de $[1, +\infty[$; et en effet Mf n'est pas 2-différentiable en 1..., mais pourtant Mf est 2-différentiable sur $]1, +\infty[$ ($(Mf)''(x)$ y vaut 1).

Comment expliquer ce phénomène ? La réponse est dans le théorème qui suit (adapté de [7]).

Théorème 2. *Soit u_0 un point de non différentiabilité de f. Considérons le convexe fermé $C(u_0) = u_0 + \partial f(u_0)$, c'est-à-dire l'ensemble des points x_0 tels que* $\text{prox}_f(x_0) = u_0$. *Alors Mf est 2-différentiable sur* $\text{int}C(u_0)$, *avec*

$$\nabla^2 Mf(x_0) = I_n \text{ pour tout } x_0 \in \text{int}C(u_0). \tag{50}$$

Il y a, bien sûr, une version duale de ce théorème avec f^*.

Démonstration. Rappelons au préalable que $C(u_0) = \partial(\frac{1}{2}\|.\|^2 + f)(u_0)$.

Soit donc $x_0 \in \text{int}C(u_0)$. Il existe donc un voisinage V de x_0 tel que $V \subset C(u_0)$. Observant à la fois que $C(u_0) = \partial(\frac{1}{2}\|.\|^2 + f)(u_0)$ et que $(\frac{1}{2}\|.\|^2 + f)^* = f^* \diamond \frac{1}{2}\|.\|^2 = Mf^*$, on a alors (*cf.* formule (16)) :

$$Mf^*(x) + (\frac{1}{2}\|.\|^2 + f)(u_0) = \langle x, u_0\rangle \text{ pour tout } x \in V.$$

Ainsi, Mf^* est une fonction affine (de x) dans le voisinage V de x_0. En conséquence, $\nabla^2 Mf^*(x_0) = 0$. Comme $\nabla^2 Mf(x_0) = I_n - \nabla^2 Mf^*(x_0)$, le résultat annoncé suit.

Une deuxième démonstration. On démarre comme dans la première au-dessus, puis la conclusion est plus simple et rapide.

Pour tout $x \in V$, $\text{prox}_f(x)$ est constante et vaut u_0. En conséquence, prox_f est différentiable en x_0 et $J(\text{prox}_f)(x_0) = 0$. D'où (revoir (38)),

$$\nabla^2 Mf(x_0) = I_n - J(\text{prox}_f)(x_0) = I_n.$$

Résumons ce qui a été vu sur la 2-différentiabilité de Mf en fonction des endroits "touchés" par l'application proximale prox_f :

$*$ Si $\text{prox}_f(x_0)$ est un point de 2-différentiabilité de f, alors Mf est 2 fois différentiable en x_0 ;

$*$ Si $\text{prox}_f(x_0)$ est un point de non différentiabilité "maximale" de f, c'est-à-dire avec $\partial f(\text{prox}_f(x_0))$ d'intérieur non vide, alors Mf est 2 fois différentiable en x_0 (et on connaît même $\nabla^2 Mf(x_0) = I_n$) ;

$*$ Si $\text{prox}_f(x_0)$ est un point de non différentiabilité "partielle" de f, c'est-à-dire avec $\partial f(\text{prox}_f(x_0))$ d'intérieur vide, alors on ne sait conclure, avec les connaissances développées ici, si Mf est 2 fois différentiable en x_0 ou pas.

Exemple 10 (tiré de [5, 6]). Ce dernier exemple fait écho au dernier point soulevé au-dessus, lorsque $\partial f(\text{prox}_f(x_0))$ n'est pas réduit à un point (f n'est donc pas différentiable en $\text{prox}_f(x_0)$), mais $\partial f(\text{prox}_f(x_0))$ est d'intérieur vide. Aussi étonnant que cela puisse paraître, Mf peut néanmoins être 2 fois différentiable en x_0. Il faut, pour illustrer cette possibilité, prendre des fonctions de deux variables au moins.

Soit $f : (x, y) \mapsto f(x, y) = |x| + \frac{1}{2}y^2$. Au voisinage de $(0, 0)$, la fonction f a l'allure "lisse" de la lettre U dans la direction des y, et l'allure "avec un coin" de la lettre V dans la direction des x. Des calculs simples - d'ailleurs déjà faits puisque f est séparable en x et y - montrent que $Mf(x, y) = \frac{1}{2}x^2 + \frac{1}{4}y^2$ et $\text{prox}_f(x, y) = (0, \frac{y}{2})$ au voisinage de $(0, 0)$. Ainsi Mf est 2 fois différentiable en $(0, 0)$. Cet exemple est la racine du "modèle U-V" d'optimisation convexe non différentiable développé depuis 25 ans par plusieurs auteurs (C. Lemaréchal, F. Oustry, C. Sagastizabal, A. Lewis, ...).

5. Un coup d'oeil aux méthodes de gradient proximal

"Proximal methods are the natural algorithms for solving regularized learning problems" (extrait du récent article [10]).

A côté de la méthode conceptuelle et primitive de "proximations successives" (voir fin de Fait 10 au §3), jettons un coup d'oeil à ce qu'est un algorithme général de type proximal.

La fonction f à minimiser sur $\mathbb{R}^n$ a la forme suivante : $f = g + h$. Les contributions et les propriétés de g et h sont différentes : g est la partie "sympathique", elle est par exemple convexe et différentiable sur $\mathbb{R}^n$ (on peut donc invoquer et utiliser son gradient) ; h est la partie "moins sympathique", convexe mais non différentiable, éventuellement prenant la valeur $+\infty$ (on ne peut l'appréhender directement). Deux exemples :

- La fonction $h(x)$ est $\lambda \|x\|_1$, la fonction $g(x)$ est $\frac{1}{2}\|Ax - b\|^2$. Minimiser $f(x) = g(x) + h(x)$ est un célèbre problème d'optimisation répertorié sous l'acronyme de LASSO (*Least Absolute Shrinkage and Selection Operator*).

- La fonction f est l'indicatrice d'un convexe fermé C (elle vaut donc 0 si $x \in C$, $+\infty$ si $x \notin C$). La minimisation de $f(x) = g(x) + h(x)$ sur $\mathbb{R}^n$ est donc la minimisation de $g(x)$ sur C.

Un algorithme général de type gradient proximal est conçu comme suit : un coup de gradient sur la fonction g,

$$x_k - \alpha_k \nabla g(x_k), \tag{51}$$

suivi d'un coup de proximation sur la fonction h,

$$\mathrm{prox}_h^{r_k}(x_k - \alpha_k \nabla g(x_k)). \tag{52}$$

Ainsi, ce qui figure en (5.2) est le nouvel itéré x_{k+1}. Les paramètres $\alpha_k > 0$ et $r_k > 0$ sont à gérer, bien sûr, et leurs comportements sont à la source des résultats de convergence de (x_k) vers un minimiseur de $f = g + h$. Une autre manière de comprendre le passage (5.2) de x_k à x_{k+1} est comme suit (faisons $r_k = 1/\alpha_k$) : x minimise $f = g + h$ si et seulement si $0 \in \nabla g(x) + \partial h(x)$, soit $-\nabla g(x) \in \partial h(x)$; or cette dernière relation peut encore s'écrire

$$x - \alpha_k \nabla g(x) \in (I + \frac{1}{r_k}\partial h)(x), \tag{53}$$

ce qui est exactement la caractérisation de $x = \mathrm{prox}_h^{r_k}(x - \alpha_k \nabla g(x))$ (revoir (17) si nécessaire).

A titre d'exemples couverts par ce modèle d'algorithme :

- Si $h = 0$, c'est l'algorithme de gradient standard sur g ;

- Si $g = 0$, c'est l'algorithme de proximations successives sur h ;

- Si h est la fonction indicatrice du convexe fermé C, $x_{k+1} = \mathrm{P}_C(x_k - \alpha_k \nabla g(x_k))$, c'est donc l'algorithme de gradient projeté ;

- Dans le cas du problème LASSO, $\nabla g(x_k)$ est connu, c'est $A^T A x_k - A^T b$, et $\mathrm{prox}_{\|.\|_1}^{r_k}(.)$ est donné par une formule explicite (*cf.* Exemple 2).

Nous entrons ici dans un nouveau monde, celui des méthodes de gradient proximal (descriptions, théorèmes de convergence, complexités, accélérations possibles, comparaisons

numériques, ...). Pour en avoir une idée, au stade d'apprentissage où il est, nous conseillons à l'étudiant-lecteur les deux exposés *Proximal gradient methods* (24 diapositives) et *Fast gradient proximal methods* (20 diapositives) d'E. Candès (via le site web référencé en [11]).

6. Un mot vers la minimisation de fonctions non convexes

Après tout, rien n'empêche de tenter de régulariser une fonction non convexe par le procédé de Moreau. Disons que les choses se passent alors moins bien : $\text{prox}_f(x)$ peut être un ensemble (non réduit à un point) par exemple ; les résultats d'Analyse convexe ne s'appliquent pas (forcément... , f n'étant pas convexe). Toutefois l'aspect de régularisation-approximation n'est pas complètement évacué. Voyons rapidement cela.

Soit $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ *une fonction* (seulement) *s.c.i. et bornée inférieurement sur* $\mathbb{R}^n$. Alors :

Fait(s) 11.

(*i*) $M_r f$ *est une fonction à valeurs finies et continue sur* $\mathbb{R}^n$;

(*ii*) *La suite de fonctions* $(M_r f)_{r>0}$ *est croissante avec* r, *et, pour tout* x, $M_r f(x) \to f(x)$ *quand* $r \to +\infty$;

(*iii*) $\text{prox}_f^r(x)$ *est un ensemble non vide compact* ;

(*iv*) *Les bornes inférieures de* f *et de* $M_r f$ *sur* $\mathbb{R}^n$ *sont égales.*

On a donc déjà perdu une propriété essentielle de $M_r f$ dans le cas où f est convexe, à savoir sa différentiabilité (*cf.* Fait 4).

Illustrons les Faits 11 avec une fonction très utilisée en optimisation (dite) parcimonieuse,

$$x = (x_1, x_2, ..., x_n) \mapsto \|x\|_0 = \text{Card}\,\{i : x_i \neq 0\}\,. \tag{54}$$

La fonction $\|.\|_0$ est "séparable" en les variables x_i, au sens suivant :

$$\|(x_1, x_2, ..., x_n)\|_0 = \sum_{i=1}^{n} \gamma(x_i),$$

où γ est la fonction de comptage basique définie par : $\gamma(u) = 1$ si $u \neq 0$, $\gamma(u) = 0$ si $u = 0$. Propriété partagée avec la fonction-norme $\|(x_1, x_2, ..., x_n)\|_1 = \sum_{i=1}^n |x_i|$. Cette séparabilité en les variables aide dans le calcul de $M_r f(x)$:

$$x = (x_1, x_2, ..., x_n) \in \mathbb{R}^n \mapsto M_r f(x) = \inf_{u \in \mathbb{R}^n} \left\{ \|u\|_0 + \frac{r}{2} \|x - u\|^2 \right\}$$

$$= \inf_{u \in \mathbb{R}^n} \left\{ \sum_{i=1}^{n} \left[\gamma(u_i) + \frac{r}{2}(x_i - u_i)^2 \right] \right\}$$

$$= \sum_{i=1}^{n} \inf_{v \in \mathbb{R}} \left[\gamma(v) + \frac{r}{2}(x_i - v)^2 \right]. \tag{55}$$

Le graphe de $M_r\gamma$ est une parabole écrêtée par une droite, dont la courbure en 0, $(M_rf)''(0) = r$, croît vers l'infini avec r.

Voici des résultats de calculs explicites, avec $r = 1$ pour simplifier :

$$Mf(x) = \begin{cases} \frac{1}{2}x^2 \text{ si } |x| \leqslant \sqrt{2}, \\ 1 \text{ si } |x| \geqslant \sqrt{2}. \end{cases} \tag{56}$$

$$\text{prox}_f(x) = \begin{cases} x \text{ si } |x| > \sqrt{2}, \\ \{0, x\} \text{ si } |x| = \sqrt{2}, \\ 0 \text{ si } |x| \leqslant \sqrt{2}. \end{cases} \tag{57}$$

En conséquence, $x \mapsto M_rf(x) = \sum_{i=1}^n M_r\gamma(x_i)$ est une fonction *continue* jouant le rôle d'approximation de la fonction $x \mapsto \|x\|_0$.

La "cousine matricielle" de la fonction $\|.\|_0$ étant la fonction rang (via la décomposition en valeurs singulières de matrices), cette manière de régulariser s'applique aussi à cette fonction de matrices pourtant bien chahutée (voir [12]), mais à chaque jour suffit sa peine.

Conclusion succincte

Voilà bientôt 60 ans que J.-J. Moreau introduisit le procédé d'*approximation-régularisation* qui porte son nom, ainsi que l'appellation et les propriétes du *point proximal* qui vont avec. Depuis, mais surtout dans une époque récente où des domaines d'applications sont très gourmands d'algorithmes d'optimisation (imagerie mathématique, apprentissage automatique ou statistique (Machine Learning)), il est très fréquent qu'on fasse appel à ces notions. Mais pour les comprendre, il faut un minimum de connaissances de base théoriques, car :

"Rien n'est plus pratique qu'une bonne théorie" (O. von Helmholtz) ;

"Theory is the first term in the Taylor series of practice" (Th. M. Cover, 1990 Shannon Lecture).

C'était l'objet de notre présentation ici.

Annexe

Démonstration du Théorème 1.

- Les techniques utilisées sont celles classiques du Calcul différentiel.
- Idées suivies :

$*$ Puisque Mf est différentiable (sur $\mathbb{R}^n$) avec $\nabla Mf(x) = x - \text{prox}_f(x)$, on va démontrer que l'application $\text{prox}_f : \mathbb{R}^n \to \mathbb{R}^n$ est différentiable en x_0 avec comme matrice jacobienne

$$J(\text{prox}_f)(x_0) = \left[I_n + A^2 f(\text{prox}_f(x_0))\right]^{-1}.$$

$*$ Comme on l'a vu, $\text{prox}_f(x)$ est $(I_n + \partial f)^{-1}(x)$ pour tout x, $\text{prox}_f(.)$ est l'inverse (univoque) de l'application (éventuellement multivoque) $(I_n + \partial f)(.)$.

Pour alléger l'ecriture, posons $\text{p}_f(.) = \text{prox}_f(.)$ et $G = I_n + \partial f$.

D'après l'hypothèse faite sur f (différentiabilité de ∂f en $\text{p}_f(x_0)$) et la définition juste au-dessus de G, la multiapplication G vérifie $G(\text{p}_f(x_0)) = x_0$ et est différentiable en x_0 avec

$JG(\mathrm{p}_f(x_0)) = I_n + A^2 f(\mathrm{p}_f(x_0))$. Donc, en raison de la propriété de semidéfinie positivité de $A^2 f(\mathrm{p}_f(x_0))$, $JG(\mathrm{p}_f(x_0))$ est définie positive et donc inversible.

Comme conséquence de l'hypothèse faite sur f, $\mathrm{p}_f(x_0)$ se trouve à l'intérieur du domaine de f. Puisque $\|\mathrm{p}_f(x_0+h) - \mathrm{p}_f(x_0)\| \leqslant \|h\|$, pour $\|h\|$ assez petit, $\mathrm{p}_f(x_0+h)$ est aussi à l'intérieur du domaine de f.

Considérons l'expression

$$\mathrm{p}_f(x_0+h) - \mathrm{p}_f(x_0) - [JG(\mathrm{p}_f(x_0))]^{-1} h. \tag{A1}$$

On va démontrer que cette expression est un $o(\|h\|)$, ce qui assurera que p_f est différentiable en x_0 avec $J\mathrm{p}_f(x_0) = [JG(\mathrm{p}_f(x_0))]^{-1}$ $(= [I_n + A^2(\mathrm{p}_f(x_0))]^{-1})$.

Allons-y. On a :

$$\left\{ \begin{array}{c} \mathrm{p}_f(x_0+h) - \mathrm{p}_f(x_0) - [JG(\mathrm{p}_f(x_0))]^{-1} h \\ = \underbrace{-\,[JG(\mathrm{p}_f(x_0))]^{-1}}_{\textit{terme fixe}} \ \underbrace{[h - JG(\mathrm{p}_f(x_0))(\mathrm{p}_f(x_0+h) - \mathrm{p}_f(x_0))]}_{\textit{quantité qu'on va exprimer autrement.}} \end{array} \right. \tag{A2}$$

A présent, se rappelant que $G = I_n + \partial f$, $\mathrm{p}_f = (I_n + \partial f)^{-1} = G^{-1}$, on a :

$$x_0 \in G(\mathrm{p}_f(x_0)), \text{ en fait } x_0 = G(\mathrm{p}_f(x_0))\,;\ x_0 + h \in G(\mathrm{p}_f(x_0+h))\,;$$

donc

$$\left\{ \begin{array}{c} h - JG(\mathrm{p}_f(x_0))(\mathrm{p}_f(x_0+h) - \mathrm{p}_f(x_0)) \\ = (x_0+h) - x_0 - JG(\mathrm{p}_f(x_0))(\mathrm{p}_f(x_0+h) - \mathrm{p}_f(x_0)) \\ \in G(\mathrm{p}_f(x_0+h)) - G(\mathrm{p}_f(x_0)) - JG(\mathrm{p}_f(x_0))(\mathrm{p}_f(x_0+h) - \mathrm{p}_f(x_0)), \end{array} \right. \tag{A3}$$

et c'est quasiment terminé.

En effet, exprimons que (la multiapplication) G est différentiable en $p_f(x_0)$:

$$\left\{ \begin{array}{c} \varepsilon > 0 \text{ étant donné, il existe } \delta > 0 \text{ tel que } \|y - \mathrm{p}_f(x_0)\| \leqslant \delta \\ \text{implique } \|G(y) - G(\mathrm{p}_f(x_0)) - JG(\mathrm{p}_f(x_0))(y - \mathrm{p}_f(x_0))\| \leqslant \varepsilon \|y - \mathrm{p}_f(x_0)\| \\ \text{(inégalité uniforme en les éléments de } G(y)). \end{array} \right. \tag{A4}$$

Mais, si $\|h\| \leqslant \delta$, on a aussi $\|\mathrm{p}_f(x_0+h) - \mathrm{p}_f(x_0)\| \leqslant \delta$ (magie de la propriété de Lipschitz de l'application p_f) ; donc d'après (A4) :

$$\left\{ \begin{array}{c} \|h\| \leqslant \delta \Rightarrow \|G(\mathrm{p}_f(x_0+h)) - G(\mathrm{p}_f(x_0)) - JG(\mathrm{p}_f(x_0))(\mathrm{p}_f(x_0+h) - \mathrm{p}_f(x_0))\| \\ \leqslant \varepsilon \|\mathrm{p}_f(x_0+h) - \mathrm{p}_f(x_0)\| \leqslant \varepsilon \|h\|\,. \end{array} \right.$$

On a donc démontré que $G(\mathrm{p}_f(x_0+h)) - G(\mathrm{p}_f(x_0)) - JG(\mathrm{p}_f(x_0))(\mathrm{p}_f(x_0+h) - \mathrm{p}_f(x_0))$ est un $o(\|h\|)$, ce qui, avec (A2) et (A3), permet d'obtenir que la quantité qui est en (A1) est bien un $o(\|h\|)$.

Commentaires.

La démonstration au-dessus a le goût du théorème des fonctions inverses, cela ressemble au théorème des fonctions inverses, mais ce n'est pas le théorème des fonctions inverses... Ce qui a permis d'éviter le recours au théorème des fonctions inverses est :

- savoir dès le départ que $JG(\mathrm{p}_f(x_0))$ est inversible ;

- le contrôle des accroissements en $\mathrm{p}_f(u)$ par ceux en u ;

- savoir dès le début qu'*il y avait* un inverse à G, c'est-à-dire p_f ici, alors qu'en Calcul différentiel usuel c'est une *conséquence* du théorème des fonctions inverses.

Références (par ordre de $1^{ère}$ citation dans le texte)

1. J.-J. Moreau, *Proximité et dualité dans un espace hilbertien.* Bull. Soc. Math. France 93 (1965), 273 − 279.

C'est l'article fondateur du domaine. Près de 60 ans après sa publication, il a gardé toute sa modernité.

2. *The proximity operator repository.* Website proximity-operator.net (rubrique Examples & Programs).

3. J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms* (2 volumes), Springer-Verlag (1993).

4. J.-B. Hiriart-Urruty, *Optimisation et Analyse convexe. Exercices et problèmes corrigés, avec rappels de cours.* Presses Universitaires de France (1998).

5. C. Lemaréchal and C. Sagastizabal, *Practical aspects of the Moreau-Yosida regularization I : theoretical properties.* Research Report-2250, INRIA (1994).

6. C. Lemaréchal and C. Sagastizabal, *Practical aspects of the Moreau-Yosida regularization : theoretical preliminaries.* SIAM Journal on Optimization, Vol. 7, No 2 (1997), 367 − 385.

7. L. Qi, *Second-order analysis of the Moreau-Yosida regularization.* Proceedings of the International Conference on Nonlinear analysis and Convex analysis. World Sci. Publ. (1999), 16 − 25.

8. J.-B. Hiriart-Urruty, *The approximate first-order and second-order directional derivatives for a convex function.* Proceedings of the conference Mathematical Theories of Optimization in Santa Margherita Ligure, Italy (1981), Lecture Notes in Mathematics 979, J. P. Cecconi and T. Zolezzi, eds., Springer-Verlag (1983), 154 − 166.

Ceci est le type de longs papiers de revue, 35 pages tapées à la machine, qu'on écrivait à l'époque, qui demandent beaucoup de travail, faisant état de travaux en cours (et des "personal communications", par exemple par Mignot, Lemaréchal, ...), et qui ont disparu des radars des bases de données car la publication n'est pas dans des journaux bien répértoriés... On ne fait plus trop cela de nos jours.

9. J.-B. Hiriart-Urruty, *Des fonctions... pas si particulières que ça : celles de Lambert, Gudermann et Airy.* Revue Quadrature, No 101, (2016) 40 − 45.

10. F. Iutzeler and J. Malick, *Nonsmoothness in Machine Learning : specific structure, proximal identification, and applications.* Journal *S*et-Valued and Variational Analysis (2020) 28 (4), 661 − 678.

11. E. Candès, *Advanced Topics in Convex Optimization.* Website candes.su.domains, University of Stanford (Teaching > Handouts > Lectures).

12. J.-B. Hiriart-Urruty and Hai Yen Le, *A variational approach of the rank function.* TOP (Journal of the Spanish Society of Statistics and Operations Research), Vol. 21, No 2 (2013), 207 − 240.