More on second-order properties of the Moreau regularization-approximation of a convex function

J.-B. Hiriart-Urruty
Institut de Mathématiques
Université Paul Sabatier de Toulouse[1]

Abstract. We unify and improve existing results on the second-order differentiabilty of the so-called Moreau regularization of a convex function.

Keywords. Convex functions; Moreau regularization; generalized differentiation; proximal mapping; machine learning.

2020 Mathematics Subject Classification. 26B, 46N, 52A, 90C.

Introduction

Go to an Optimization congress, more precisely in sessions devoted to large scale problems such as those found in mathematical imagery, automatic learning or statistics (Machine Learning), and you will hear about Moreau's regularization, proximal (algorithmic) methods, etc. To understand them, you need a minimum of basic theoretical (*i.e.*, mathematical) knowledge. It is to this need that I had to respond by teaching in a Master 2R of Operation Research (course entitled "Contemporary Themes in (continuous) Optimization" during the last six years. The audience of students (all at the graduate level) mainly came from four engineering schools in Toulouse as well as the Paul Sabatier University[2].

For these purposes, we have chosen to start from a beginner level in the targeted field, hence the title of the pedagogical text [8], avoiding the temptation to take for granted things that seem simple to us (so much we are "in" by our own practices and work in Optimization).

The text referenced in [8] is divided into six parts of very unequal lengths. After the introductory paragraphs of Analysis (§1) and modern Convex Analysis (§2), we present in §3 the properties of Moreau's regularization (to first order); everything is distilled in the form of "facts" (= statements) without proofs (as we will do in section 2 of the current paper). It is, for the student-reader, the basis for understanding the so-called proximal algorithmic methods.

Section 4 is dedicated to the second-order properties of Moreau's regularization; in addition to summarizing the results available in terms of classical differential calculus, we

[1]Postal address:118 route de Narbonne 31062 Toulouse Cedex 09, France.
E-mail: jbhu@math.univ-toulouse.fr

[2]Oleg Burdakov was familiar with this ecosystem of engineering schools and universities in Toulouse. He was senior scientific adviser at CERFACS in 1995-1997. It was then that I had the opportunity to meet him for the first time.

improve some results from the literature. This is the subject of our paper here. The main results are displayed in section 3.

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

1. Moreau's construction

In works dating from 1963 – 1965, including a founding paper published in 1965 ([1]), the archetype, in my opinion, of an elegant and profound article of mathematics, the mechanic-mathematician J.-J. MOREAU[3] defined and studied the *approximate-regularized* (or envelope) of a convex function, which bears his name.

Consider a lower-semicontinuous (l.s.c.) convex function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ and $r > 0$ a parameter. Then, one defines the function $M_r f$ on $\mathbb{R}^n$ in the following way:

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

Here, $\|.\|$ stands for the usual Euclidean norm on $\mathbb{R}^n$. One remarks that the role of the parameter r is not essential in the construction of $M_r f$ since

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

Hence, if one is able to derive properties of the construction of MOREAU with the parameter $r = 1$, one will deduce similar conclusions for any parameter $r > 0$. That is what we will do for second-order differentiability properties. Indeed, the simplified notation Mf will be used for $M_1 f$.

The unique minimizer in the optimization problem (1) defining $M_r f(x)$ is the *proximal point* of x; it is denoted by $\operatorname{prox}_f^r(x)$. The mapping $\operatorname{prox}_f^r : \mathbb{R}^n \to \mathbb{R}^n$ is called *proximal mapping* or *proximal operator* attached with the function f and the parameter r.

Example 1. Let C be a (nonempty) closed convex set in $\mathbb{R}^n$ and f the l.s.c. convex function which takes the value 0 on C and $+\infty$ elsewhere (called the *indicator function* of C). Simple calculations lead to:

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

$$\operatorname{prox}_f^r(x) = \mathrm{p}_C(x) \text{ for all } x \in \mathbb{R}^n. \tag{3}$$

Here, $d_C(x)$ denotes the Euclidean distance from x to C, and $\mathrm{p}_C(x)$ is the projection of x onto C. This is this initial example that lead MOREAU to coin the qualifier *proximal*.

Example 2 (with Figure 1). Let $f : x \in \mathbb{R} \mapsto f(x) = |x|$. Then

$$M_r f(x) = \begin{cases} \frac{r}{2} x^2 \text{ if } x \in [-1/r, 1/r], \\ |x| - \frac{1}{2r} \text{if } |x| \geqslant 1/r, \end{cases} ; \tag{4}$$

[3] J.-J. MOREAU (1923 – 2014), who made his career at the University of Montpellier (France).

$$\text{prox}_f^r(x) = \begin{cases} 0 \text{ if } x \in [-1/r, 1/r], \\ x - 1/r \text{ if } x \geqslant 1/r, \\ x + 1/r \text{ if } x \leqslant -1/r \end{cases}. \tag{5}$$

If we stick to a condensed formula for $\text{prox}_f^r(x)$, one can write

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

where $\operatorname{sign}(x)$ equals 1 if $x > 0$, -1 if $x > 0$, 0 if $x = 0$.

The function $\frac{1}{r} M_r f$ is the so-called HUBER function, much used in Statistics. It is a compromise between quadratic behavior (near 0) and linear behavior (when the variable is large).

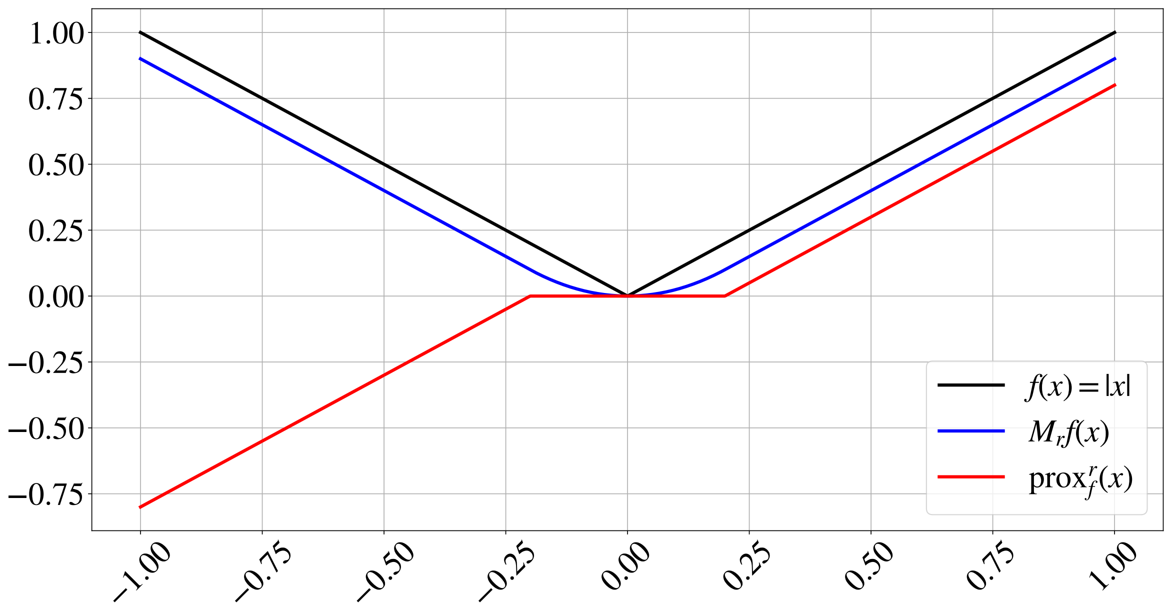

Figure 1

Example 3. A third example is with a convex quadratic function (of several variables) $f : x \in \mathbb{R}^n \mapsto f(x) = \frac{1}{2} \langle Ax, x \rangle$, where A is a symmetric positive semidefinite (n, n) matrix. Then, for all $x \in \mathbb{R}^n$,

$$\begin{cases} M_r f(x) = \frac{1}{2} \langle A_r x, x \rangle, \\ \text{with } A_r = A(I_n + \frac{1}{r}A)^{-1} = r \left[I_n - (I_n + \frac{1}{r}A)^{-1} \right]; \end{cases} \tag{6}$$

$$\text{prox}_f^r(x) = (I_n + \frac{1}{r}A)^{-1}(x). \tag{7}$$

Explicit calculations (I am not talking about numerical approximations by calculations) of $M_r f$ and of prox_f^r are sometimes possible; a repository is dedicated to them (see [2]), we will use it later (in section 3). From the point of view of numerical calculations, note the decomposable character of $\frac{1}{2}\|x-u\|^2 = \sum_{i=1}^n \frac{1}{2}(x_i - u_i)^2$. Thus, if f is itself decomposable, $f(x) = \sum_{i=1}^n f_i(x_i)$, the calculations of $M_r f(x)$ and $\mathrm{prox}_f^r(x)$ amount to n independent computations with functions f_i of a single real variable x_i. This is what happens with the (important) norm function $f(x) = \|x\|_1 = \sum_{i=1}^n |x_i|$.

In short, we have understood in view of these few examples that it is better to have to deal with convex functions f "prox friendly" (as I have seen it written by certain authors).

2. First-order properties of $M_r f$ and prox_f^r: *a digest*

The subject of modern Convex Analysis is widely covered in many books, whether teaching-research or corrected exercises ([3] for example). We only use here the rudiments on two essential objects: the subdifferential ∂f and the LEGENDRE-FENCHEL conjugate f^* of a convex function f.

An absolutely extraordinary result of MOREAU, concerning the regularization which bears his name, is that when we have regularized f, we have also regularized f^*, because:

$$Mf(x) + Mf^*(x) = \frac{1}{2}\|x\|^2, \tag{8}$$

$$\mathrm{prox}_f(x) + \mathrm{prox}_{f^*}(x) = x \text{ for all } x \in \mathbb{R}^n. \tag{9}$$

We understand that this will have consequences on the second-order differentiability of Mf and of Mf^*: they are twice differentiable or not at x *at the same time*. We will come back to this in section 3.

Here below are collected under the name of "Facts" the main results to know about $M_r f$ and prox_f^r. They are presented without proofs, knowing that they can be found in various books (Example: [3, Vol. 2, pages 317 − 330].

Fact 1. *$M_r f$ is a convex function, everywhere* finite *and* differentiable *on $\mathbb{R}^n$ (even with a Lipschitz gradient, but no more).*

Fact 2. *For all $y \in \mathbb{R}^n$,*

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

The prox mapping prox_f^r sends $\mathbb{R}^n$ onto $\mathcal{D} = \{x \in \mathrm{dom} f : \partial f(x)$ is nonempty$\}$ (exactly, no more, no less).

Fact 3. *For all $x \in \mathbb{R}^n$,*

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

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

Fact 4. $(r = 1)$ "*When you have one, you have the other one":*

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

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

Two remarks are in order here:

- The function Mf is not "too convex", in fact "less convex" than $(1/2)\|.\|^2$, since it is necessary to add another convex function, namely Mf^*, to get at $(1/2)\|.\|^2$. This assertion, a little vague at this point, will be clarified a little more during the study of the second-order differentiation of Mf in section 3.

- The mapping prox_f is a "gradient field" (or "derives from a potential function"), *i.e.*, it is the gradient of a function. This has an immediate consequence: at a point x where the mapping prox_f is differentiable, the Jacobian matrix $J(\mathrm{prox}_f)(x)$ is necessarily symmetric (a result in Differential Calculus).

Fact 5. The mapping prox_f is $r-$Lipschitz on $\mathbb{R}^n$, that is to say:

$$\left\|\mathrm{prox}_f(x) - \mathrm{prox}_f(y)\right\| \leqslant r \|x - y\| \text{ for all } x, y \text{ in } \mathbb{R}^n.$$

An example of consequence: Mf is a convex function with $1-$Lipschitz gradient mapping.

Fact(s) 6. Concerning lower bounds and minimizers of f and $M_r f$, we have:

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

$$(x \text{ minimizes } f \text{ on } \mathbb{R}^n) \Leftrightarrow (x \text{ minimizes } M_r f \text{ on } \mathbb{R}^n).$$

The next four statements are equivalent:

$$(i) \ \ x \text{ minimizes } f \text{ (or } M_r f \text{) on } \mathbb{R}^n \ ;$$

$$(ii) \ \ \mathrm{prox}_f^r(x) = x \ ;$$

$$(iii) \ \ f(\mathrm{prox}_f^r(x)) = f(x) \ ;$$

$$(iv) \ M_r f(x) = f(x).$$

3. Second-order properties of $M_r f$: what to expect, what can be proved

One is tempted to say - and I had the opportunity to read it - this: if the convex function f is twice differentiable (even of class $\mathcal{C}^\infty$) on $\mathrm{int}(\mathrm{dom} f)$, that is say on the largest set where it could be, then $M_r f$ is twice differentiable. This is clearly wrong, it suffices to see that to consider the indicator function f of $[-1, 1]$, which leads to a Moreau-regularized $M_r f$ which is not twice differentiable at the points -1 and 1 (see Example 1 or Example 4). Yet the result is true if f is assumed to have finite values (everywhere), that is to say if $\mathrm{dom} f = \mathbb{R}^n$. The studies on this subject of twice differentiability of $M_r f$ are old, they date from the years $1994 - 1997$; see for example the works [4], [5], [6]. We take up the essentials here in synthetic form, improving them in passing; Corollary 2 below is an example of covering and improving the existing results.

3.1 Preamble on the almost everywhere second-order differentiatibility of a convex function

Let us go back to basics, with functions $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$.

- We say that f admits at x_0 a TAYLOR-YOUNG second-order expansion when: f is differentiable at x_0 and there exists a symmetric matrix (denoted $A^2 f(x_0)$) such that

$$f(x_0 + h) = f(x_0) + \langle \nabla f(x_0), h \rangle + \frac{1}{2} \left\langle A^2 f(x_0) h, h \right\rangle + o(\|h\|^2). \tag{10}$$

The notation A is for A. D. ALEXANDROFF who, in 1939, published a paper proving that this takes place for almost any x_0 when the function f is convex, *i.e.* outside a set of null measure (in the LEBESGUE sense).

This is weaker than the (usual) second-order differentiability of f at x_0. But for a convex function it comes down to about the same: *If the convex function is (once) differentiable in a neighborhood of x_0, we have (*10*) if and only if f is twice differentiable at x_0, with $\nabla^2 f(x_0) = A^2 f(x_0)$.* This is far from being easy to prove ([7, Corollary 2.13]).

- According to R. T. ROCKAFELLAR and F. MIGNOT (in works published in 1976), one says that the set-valued mapping ∂f (for a convex function f) is differentiable at x_0 if, firstly f is differentiable at x_0, and then there exists a matrix $D^2 f(x_0)$ (that we could also denote as $J(\partial f)(x_0)$) such that

$$\left\{ \begin{array}{c} \|\partial f(x) - \nabla f(x_0) - D^2 f(x_0)(x - x_0)\| = o\left(\|x - x_0\|\right) \\ \text{(with } o(.) \text{ uniform for the } s \in \partial f(x)\text{).} \end{array} \right. \tag{11}$$

In a more detailed form, that means : For all $\varepsilon > 0$, there exists $\delta > 0$ such that

$$\left\{ \begin{array}{c} (\|x - x_0\| \leqslant \delta \text{ and } s \in \partial f(x)) \Rightarrow \\ (\|s - \nabla f(x_0) - D^2 f(x_0)(x - x_0)\| \leqslant \varepsilon \|x - x_0\|). \end{array} \right. \tag{11bis}$$

MIGNOT proved in a paper published in 1976 that ∂f is differentiable at almost all points x_0. In [7, Proposition 2.11], I proved that this matrix $D^2 f(x_0)$ is necessarily symmetric and positive semidefinite. In [7, Corollary 2.12], I also proved the following "logical and expected" result: *f admits at x_0 a* TAYLOR-YOUNG *second-order expansion if, and only if, ∂f is differentiable at x_0. In short, $A^2 f(x_0) = D^2 f(x_0)$.* By abuse of language, we therefore say that "f (convex) is twice A-differentiable at x_0" when we have (10) or (11), and we will keep the notation $A^2 f(x_0)$ (which - let us recall it - is $\nabla^2 f(x_0)$ when f is twice differentiable (in the usual sense) at x_0).

- A word about the counterpart of (10) for the conjugate function f^*: If we have the second-order expansion (10) at x_0, we have something similar for f^* at $s_0 = \nabla f(x_0)$, provided that $A^2 f(x_0)$ is invertible ([3, Vol. 2, page 89]):

$$\left\{ \begin{array}{c} f^*(s_0 + p) = f^*(s_0) + \langle x_0, p \rangle + \frac{1}{2} \left\langle \left[A^2 f(x_0)\right]^{-1} p, p \right\rangle + o(\|p\|^2), \\ \text{(with } x_0 = \nabla f^*(s_0)\text{, we recall it).} \end{array} \right. \tag{10*}$$

In what follows, one will choose, according to the case, the expansion (10) or (11); formulation (10) is more "palpable", while formulation (11) is more "powerful" (especially in proofs).

3.2 Getting twice differentiability of $M_r f$ from that of f

To lighten the notations, and without loss of generality, we now make $r = 1$ in the Moreau regularization process.

Let us recall (*cf.* Fact 3) that the twice differentiability of the function Mf at x_0, that is to say the differentiability of the mapping ∇Mf at x_0, is equivalent to the differentiability of the mapping prox_f at x_0, with

$$\nabla^2 Mf(x_0) = I_n - J(\mathrm{prox}_f)(x_0). \tag{12}$$

This relation confirms that $J(\mathrm{prox}_f)(x_0)$ is a symmetric matrix, as we announced it previously (at the end of Fact 4).

The key result linking the second-order differentiability of f and that of Mf is as follows.

Theorem 1. *If f is twice A-differentiable at* $\mathrm{prox}_f(x_0)$*, then Mf is twice differentiable (in the usual sense) at x_0, with*

$$\nabla^2 Mf(x_0) = I_n - \left[I_n + A^2 f(\mathrm{prox}_f(x_0))\right]^{-1}. \tag{13}$$

The proof is postponed in the Appendix.

Note immediately that formula (13) remains valid even if $A^2 f(\mathrm{prox}_f(x_0))$ is singular (*i.e.*, is not invertible). Since $A^2 f(\mathrm{prox}_f(x_0))$ is positive semidefinite, $I_n + A^2 f(\mathrm{prox}_f(x_0))$ is positive definite, hence invertible.

Even if f is twice differentiable (in the classical sense) wherever possible, *i.e.* at best on $\mathrm{int}(\mathrm{dom} f)$, the above result shows that this does not imply that Mf is twice differentiable everywhere: *it depends on the proximal points* $\mathrm{prox}_f(x)$, *if these fall in the twice differentiability zone of f or not.* The case where $\mathrm{prox}_f(x)$ falls on a boundary point of $\mathrm{dom} f$, this being however a point where the subdifferential of f is not empty, is particularly interesting; it will be considered below.

The "dual" version of Theorem 1 consists in writing the same result on the conjugate f^*, remembering that $\nabla^2 Mf(x_0) = I_n - \nabla^2 Mf^*(x_0)$.

Theorem 1*. *If f^* is twice A-differentiable at* $\mathrm{prox}_{f^*}(x_0)$ $(= x_0 - \mathrm{prox}_f(x_0))$*, then Mf is twice differentiable (in the usual sense) at x_0, with*

$$\nabla^2 Mf(x_0) = \left[I_n + A^2 f^*(\mathrm{prox}_{f^*}(x_0))\right]^{-1}. \tag{13*}$$

The two theorems above, Theorem 1 and Theorem 1*, do not lead to the twice differentiabilty of Mf at the same points x_0; the next example is an illustration of that.

Example 4. Let f be the indicator function of $[-1, 1]$. Then, f is trivially twice differentiable on $\mathrm{int}(\mathrm{dom} f) = (-1, 1)$, but

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

is not twice differentiable everywhere. Theorem 1 can be applied at $x_0 \in (-1,1)$ since, in that case, $\mathrm{prox}_f(x_0)$, which equals x_0, is in the twice differentiability zone of f. When $x_0 \notin (-1,1)$, $\mathrm{prox}_f(x_0) = \pm 1$ and everything can happen : Mf can be twice differentiable at x_0 just as Mf cannot be twice differentiable at x_0.

The dual version of this example is as follows. One has $f^* = |.|$. Then, f^* is clearly twice differentiable at all points except at 0, but

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

is not everywhere twice differentiable, exactly as (and at the same points ± 1 as) Mf. Theorem 1* can be applied (to f^*) at $x_0 \notin (-1,1)$ since, in that case, $\mathrm{prox}_{f^*}(x_0)$, which is different from 0, lies in the twice differentiability zone of f^*. When $x_0 \in [-1,1]$, $\mathrm{prox}_{f^*}(x_0) = 0$ and everything can happen: Mf^* can be twice differentiable at x_0 just as Mf^* cannot be twice differentiable at x_0.

By combining the two results, we arrived at the twice differentiability of Mf and Mf^* everywhere except perhaps in ± 1, and this is indeed the best we could do.

Let us draw some corollaries from the result of Theorem 1.

Corollary 2. *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be convex, l.s.c., satisfying the following assumption:*

$$(\mathcal{H}) \quad \begin{cases} f \textit{ is twice differentiable on } \mathrm{int}(\mathrm{dom} f), \\ \partial f(x) \textit{ is empty at any point of the boundary of } \mathrm{dom} f. \end{cases}$$

Then Mf is twice differentiable everywhere on $\mathbb{R}^n$.

The proof of Corollary 2 is fairly simple from the result of Theorem 1. Indeed, for all $x \in \mathbb{R}^n$, $\mathrm{prox}_f(x)$ is a point where the subdifferential of f is not empty (see Fact 2). However, by hypothesis $(\mathcal{H})$, such a point can only be inside $\mathrm{dom} f$, the zone where precisely f has been assumed to be twice differentiable. ⊡

Note that the second part of hypothesis $(\mathcal{H})$ only concerns points which are both on the boundary of $\mathrm{dom} f$ *and* in $\mathrm{dom} f$ (since, by definition, $\partial f(x)$ is empty when $x \notin \mathrm{dom} f$).

Corollary 3. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex twice differentiable on $\mathbb{R}^n$ (resp. of class $\mathcal{C}^2$ on $\mathbb{R}^n$). Then Mf is twice differentiable on $\mathbb{R}^n$ (resp. of class $\mathcal{C}^2$ on $\mathbb{R}^n$).*

For the twice differentiabilty of Mf, Corollary 2 trivially applies since the boundary of the domain of f is empty.

Let us see for the $\mathcal{C}^2$ property. We have $A^2 f(.) = \nabla^2 f(.)$ which is continuous by hypothesis, the mapping $\mathrm{prox}_f(.)$ which is continuous (*cf.* Fact 5), and the formula:

$$\nabla^2 Mf(x) = I_n - \left[I_n + \nabla^2 f(\mathrm{prox}_f(x))\right]^{-1}. \tag{14}$$

Then, it suffices to observe that $\nabla^2 Mf(.)$ results from the chaining (or composition) of continuous mappings. $\boxdot$

In the case of functions f of a single variable, formula (14) takes a simplified form:

$$(Mf)''(x) = \frac{f''(\mathrm{prox}_f(x))}{1 + f''(\mathrm{prox}_f(x))}. \tag{15}$$

We will have the opportunity to illustrate it several times.

Example 5 (with Figure 2)**.** Let $f : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ be the basic and familiar convex function defined by: $f(x) = -\ln(x)$ if $x > 0$, $+\infty$ otherwise. Thus assumption $(\mathcal{H})$ in Corollary 2 is satisfied, and thus Mf is twice differentiable on the whole of $\mathbb{R}$. This example is interesting because it shows that we could modify f by making it an affine function on a subinterval of $(0, +\infty)$ or by modifying its behavior when $x \to +\infty$, provided of course that we preserve its twice differentiability on $(0, +\infty)$, without destroying the twice differentiability of the resulting Mf.

If we want to have explicit calculations for the function f of this example, here they are:

$$\begin{cases} \mathrm{prox}_f(x) = \frac{x+\sqrt{x^2+4}}{2}, \\ Mf(x) = -\ln(\frac{x+\sqrt{x^2+4}}{2}) + \frac{1}{4}(x^2 + 2 - x\sqrt{x^2+4}), \\ (Mf)'(x) = \frac{x-\sqrt{x^2+4}}{2}, \\ (Mf)''(x) = \frac{1}{2}(\frac{1}{\sqrt{x^2+4}} - x). \end{cases} \tag{16}$$

One illustrates in this example, firstly $(Mf)'(x) = x - \mathrm{prox}_f(x)$, and secondly $(Mf)''(x) = \frac{f''(\mathrm{prox}_f(x))}{1+f''(\mathrm{prox}_f(x))}$ (formula (15)).

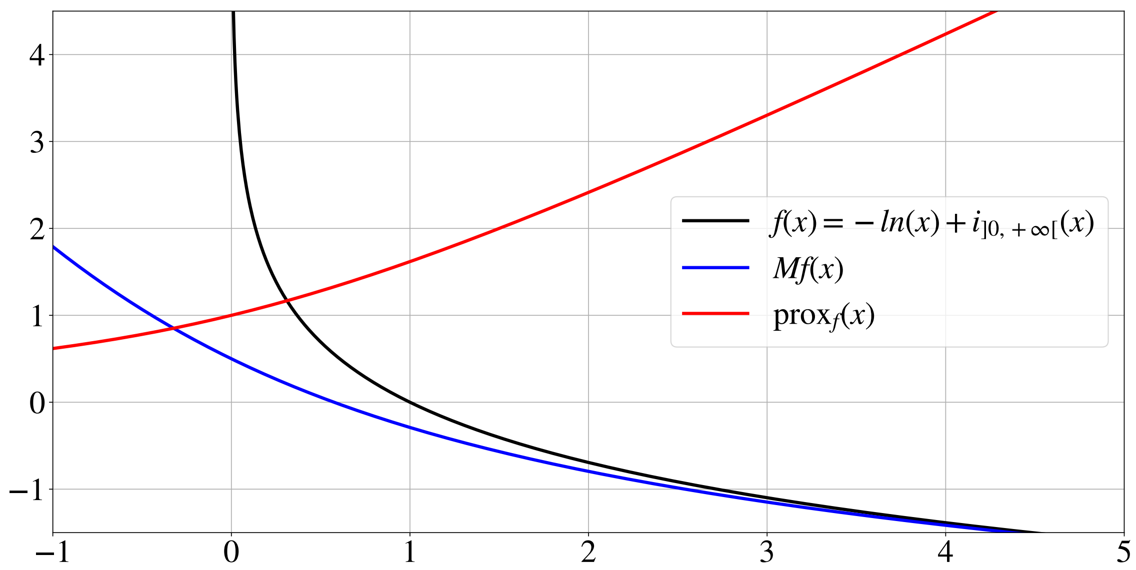

Figure 2

Example 6 (with Figure 3)**.** Let $f : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ be defined by

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

This example is interesting in the sense that at the boundary points ± 1 of $\operatorname{dom} f = [-1,1]$, the subdifferential of f is empty. Thus, the hypothesis $(\mathcal{H})$ in Corollary 2 is verified, and the function Mf therefore is twice differentiable everywhere on $\mathbb{R}$. Here again, we can carry out explicit calculations, here they are:

$$\begin{cases} \operatorname{prox}_f(x) = \frac{x}{\sqrt{1+x^2}}, \\ Mf(x) = \frac{1}{2}x^2 - \sqrt{1+x^2}, \\ (Mf)'(x) = x - \frac{x}{\sqrt{1+x^2}}, \\ (Mf)''(x) = 1 - \frac{1}{(1+x^2)^{3/2}}. \end{cases} \tag{17}$$

Again in this example, one illustrates that firstly $(Mf)'(x) = x - \operatorname{prox}_f(x)$, and secondly $(Mf)''(x) = \frac{f''(\operatorname{prox}_f(x))}{1+f''(\operatorname{prox}_f(x))}$ (formula (15)).

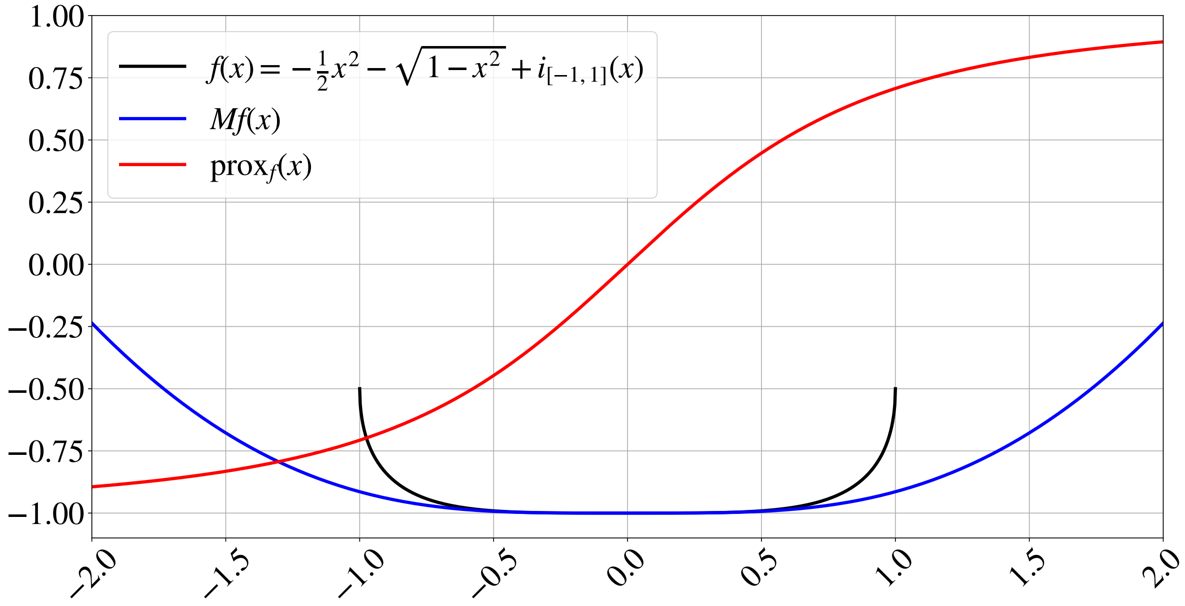

Figure 3

Example 7. Let $f : \mathbb{R} \to \mathbb{R}$ be convex and twice differentiable on $\mathbb{R}$. We know, from Corollary 3, that M_f is twice differentiable on $\mathbb{R}$, and even with its "curvature" bounded above by that of f and by 1:

$$(Mf)''(x) = \frac{f''(\mathrm{prox}_f(x))}{1 + f''(\mathrm{prox}_f(x))} \leqslant \min(1, f''(\mathrm{prox}_f(x)). \tag{18}$$

This type of upper bound will be taken up more generally in Corollary 4 below.

One can multiply illustrations with functions of a single variable, as in Examples $5 - 7$ above, as long as explicit calculations of $\mathrm{prox}_f(x)$ are available. For this, one can consult the repository [2].

Example 8 (Example 3 revisited). With the example of quadratic forms f, it is time to give variants of formula (14) and its cousins. Let $f : x \in \mathbb{R}^n \to \mathbb{R}$ be defined by $f(x) = \frac{1}{2}\langle Ax, x\rangle$, where A is a positive semidefinite (symmetric) matrix. Then, by defining $S = I_n - [I_n + A]^{-1}$, we have:

$$\begin{cases} \mathrm{prox}_f(x) = [I_n + A]^{-1}x = x - Sx, \\ Mf(x) = \frac{1}{2}\langle Sx, x\rangle, \\ \nabla Mf(x) = Sx, \\ \nabla^2 Mf(x) = S. \end{cases} \tag{19}$$

This is the prototype of the formula (14).

Since A is positive semidefinite, it turns out that

$$(S = \ ) \ I_n - [I_n + A]^{-1} = A\,[I_n + A]^{-1} \tag{20.1}$$

$$= [I_n + A]^{-1} A = A - A\,[I_n + A]^{-1} A \tag{20.2}$$

$$= \text{(if } A \text{ is invertible)} \ \left[I_n + A^{-1}\right]^{-1}. \tag{20.3}$$

These relations between matrices are a bit tricky to prove... You have to use $UU^{-1} = U^{-1}U = I_n$ with several different matrices U (formed with A and I_n). The matrix S shown in $(20.1) - (20.3)$ is sometimes called in the literature the *parallel sum* of A and I_n. Clearly, $\mathrm{Ker}\ S = \mathrm{Ker}\ A$, $\mathrm{Im}\ S = \mathrm{Im}\ A$.

For our use here, $(20.1) - (20.3)$ yield variants of the expression for $\nabla^2 Mf(x_0)$ in formula (14).

The different matrix forms seen in $(20.1) - (20.3)$ lead to specify a little more the relation between $\nabla^2 Mf(x)$ and $\nabla^2 f(\mathrm{prox}_f(x))$. In the next statement, the inequality $A \preccurlyeq B$ between (symmetric) positive semidefinite matrices means, as usual, that $B - A$ is positive semidefinite.

Corollary 4. *Let us place ourselves under the assumptions of Corollary 2. Then, for all* $x_0 \in \mathbb{R}^n$ *:*

$$\nabla^2 Mf(x_0) \preccurlyeq \nabla^2 f(\mathrm{prox}_f(x_0)) \ ; \tag{21.1}$$

$$\nabla^2 Mf(x_0) \preccurlyeq I_n \ ; \tag{21.2}$$

$$\left\{ \begin{array}{c} \textit{If } \lambda_1, ..., \lambda_n \textit{ denote the eigenvalues of } \nabla^2 f(\mathrm{prox}_f(x_0)), \\ \textit{then those of } \nabla^2 Mf(x_0) \textit{ are } \frac{\lambda_1}{1+\lambda_1}, ..., \frac{\lambda_n}{1+\lambda_n} \\ (\textit{thus, } \frac{\lambda_i}{1+\lambda_i} \leqslant \min(1, \lambda_i)). \end{array} \right. \tag{21.3}$$

The main result so far is that we have been able to prove the twice differentiability of f at x_0 whenever $\mathrm{prox}_f(x_0)$ "falls" in the zone of twice differentiability of f. Question: *What if otherwise?* Let us start with two very simple examples (the first one seen in Example 2). If f *is not twice twice differentiable at* $\mathrm{prox}_f(x_0)$ (for example, is not once differentiable), Theorem 1 cannot apply to x_0 nor to all x which have been "contaminated", those of $x_0 + \partial f(x_0)$ (since they give the same $\mathrm{pro}x_f(x_0)$!). Consider therefore the function $x \in \mathbb{R} \mapsto f(x) = |x|$. In cases where $\mathrm{prox}_f(x_0) = 0$, Theorem 1 does not apply; in fact all the "contaminated" x's are those of $[-1, 1]$; and indeed Mf is not twice differentiable at -1 and at 1, but nevertheless Mf is twice differentiable at $(-1, 1)$ ($Mf(x)$ equals $\frac{1}{2}x^2$ there).

The dual version of this example is the function f of Example 4. For $x_0 \geqslant 1$, we have $\mathrm{prox}_f(x_0) = 1$, and f is not differentiable there. Theorem 1 does not apply; in fact all the "contaminated" x's are those of $[1, +\infty)$; and indeed Mf is not twice differentiable at 1, yet Mf is twice differentiable at any $x \in (1, +\infty)$ ($(Mf)''(x)$ equals 1 there).

How to explain this phenomenon? The answer is in the following theorem (adapted from [6]).

Theorem 2. *Let u_0 be a point where f is not differentiable. Consider the closed convex set $C(u_0) = u_0 + \partial f(u_0)$, that is the set of points x_0 for which $\mathrm{prox}_f(x_0) = u_0$. Then Mf is twice differentiable on $\mathrm{int}C(u_0)$, with*

$$\nabla^2 Mf(x_0) = I_n \text{ for all } x_0 \in \mathrm{int}C(u_0). \tag{22}$$

There is, of course, a dual version of this theorem with f^*.

Proof. Let $x_0 \in \mathrm{int}C(u_0)$. There exists a neighborhood N de x_0 tel que $N \subset C(u_0)$. For all $x \in N$, $\mathrm{prox}_f(x)$ is constantly equal to u_0. Consequently, the mapping prox_f is differentiable x_0 and $J(\mathrm{prox}_f)(x_0) = 0$. Thus (see (12)),

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

Let us summarize what has been seen on the twice differentiability of Mf depending on the places "touched" by the proximal mapping prox_f:

$*$ If $\mathrm{prox}_f(x_0)$ is a point where f is twice differentiable, then Mf is twice differentiable at x_0;

$*$ If $\mathrm{prox}_f(x_0)$ is a point of "maximal" nondifferentiability of f, *i.e.* with $\partial f(\mathrm{prox}_f(x_0))$ of nonempty interior, then Mf is twice differentiable at x_0 (and we even know that $\nabla^2 Mf(x_0) = I_n$);

$*$ If $\mathrm{prox}_f(x_0)$ is a point of "partial" nondifferentiability of f, *i.e.* when $\partial f(\mathrm{prox}_f(x_0))$ has a nonempty interior, then we cannot conclude, with the knowledge developed here, whether Mf is twice differentiable at x_0 or not.

Example 9 (from [4, 5]).

This last example echoes the last point raised above, when $\partial f(\mathrm{prox}_f(x_0))$ is not reduced to a point (f is therefore not differentiable at $\mathrm{prox}_f(x_0)$), but $\partial f(\mathrm{prox}_f(x_0))$ has an empty interior. As surprising as it may seem, Mf could nevertheless be twice differentiable at x_0. To illustrate this possibility, it is necessary to consider functions of at least two variables.

Let $f : (x, y) \mapsto f(x, y) = |x| + \frac{1}{2}y^2$. In a neighborhood of $(0, 0)$, the function f has the "smooth" appearance of the letter U in the y direction, and the "kink" appearance of the letter V in the x direction. Simple calculations - moreover already done since f is separable in x and y - show that $Mf(x, y) = \frac{1}{2}x^2 + \frac{1}{4}y^2$ and $\mathrm{prox}_f(x, y) = (0, \frac{y}{2})$ in a neighborhood of $(0, 0)$. Thus Mf is twice differentiable at $(0, 0)$. This example is the root of the "U-V model" of nondifferentiable convex optimization developed over the past 25 years by several authors like C. Lemaréchal, F. Oustry, C. Sagastizabal, A. Lewis, and so on.

Brief conclusion

It has been almost 60 years since J.-J. Moreau introduced the approximation-regularization process that bears his name, as well as the name and properties of the *proximal mapping* that go with it. Since then, but especially in recent times where fields of application are very greedy for optimization algorithms (mathematical imaging, automatic or statistical learning (Machine Learning)), it is very common to call on these notions:

"Proximal methods are the natural algorithms for solving regularized learning problems" ([9]).

But to understand them, you need a minimum of basic theoretical knowledge, because:

"Nothing is more practical than a good theory" (O. von Helmholtz).

This was the aim of our presentation here, concentrated on the second-order smoothness properties of Moreau's approximation-regularization process.

Appendix

Proof of Theorem 1. The used techniques are rather classical in advanced Differential Calculus. The followed ideas are:

$*$ Since Mf is (once) differentiable (on $\mathbb{R}^n$) with $\nabla Mf(x) = x - \mathrm{prox}_f(x)$, we show that the mapping $\mathrm{prox}_f : \mathbb{R}^n \to \mathbb{R}^n$ is differentiable at x_0 with the following Jacobian matrix:

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

$*$ As we have recalled, $\mathrm{prox}_f(x)$ is $(I_n + \partial f)^{-1}(x)$ for all x; that means that $\mathrm{prox}_f(.)$ is the single-valued inverse of the set-valued mapping $(I_n + \partial f)(.)$.

To simplify the writing, we set $\mathrm{p}_f(.) = \mathrm{prox}_f(.)$ and $G = I_n + \partial f$.

According to the assumption made on f (differentiability of the set-valued mapping ∂f at $\mathrm{p}_f(x_0)$) and the definition just above of G, the set-valued mapping G satisfies $G(\mathrm{p}_f(x_0)) = x_0$ and is differentiable at x_0 with $JG(\mathrm{p}_f(x_0)) = I_n + A^2 f(\mathrm{p}_f(x_0))$. Thus, because $A^2 f(\mathrm{p}_f(x_0))$ is positive semidefinite, $JG(\mathrm{p}_f(x_0))$ is positive definite, hence invertible.

As a consequence of the assumption made on f, $\mathrm{p}_f(x_0)$ lies in the interior of the domain of f. Furthermore, since $\|\mathrm{p}_f(x_0+h)-\mathrm{p}_f(x_0)\| \leqslant \|h\|$ (because p_f is a 1-Lipschitz mapping), for $\|h\|$ small enough, $\mathrm{p}_f(x_0+h)$ also lies in the interior of f.

Consider now the expression

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

Our objective is to prove that this quantity is a $o(\|h\|)$, which will ensure that p_f is differentiable at x_0 with $J\mathrm{p}_f(x_0)=[JG(\mathrm{p}_f(x_0))]^{-1}$ $(=[I_n+A^2(\mathrm{p}_f(x_0))]^{-1})$.

Let us proceed. We have:

$$\left\{\begin{array}{c} \mathrm{p}_f(x_0+h)-\mathrm{p}_f(x_0)-[JG(\mathrm{p}_f(x_0))]^{-1}\,h \\ =\underbrace{-[JG(\mathrm{p}_f(x_0))]^{-1}}_{\text{a fixed term}}\ \underbrace{[h-JG(\mathrm{p}_f(x_0))(\mathrm{p}_f(x_0+h)-\mathrm{p}_f(x_0))]}_{\text{a quantity that we express otherwise.}} \end{array}\right. \tag{A2}$$

Now, recalling that $G=I_n+\partial f$, $\mathrm{p}_f=(I_n+\partial f)^{-1}=G^{-1}$, we have :

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

consequently,

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

and we almost are done.

Indeed, let us express that the (set-valued) mapping G is differentiable at $p_f(x_0)$:

$$\left\{\begin{array}{c} \text{Given } \varepsilon>0, \text{ there exists } \delta>0 \text{ such that } \|y-\mathrm{p}_f(x_0)\|\leqslant\delta \\ \text{implies } \|G(y)-G(\mathrm{p}_f(x_0))-JG(\mathrm{p}_f(x_0))(y-\mathrm{p}_f(x_0))\|\leqslant\varepsilon\,\|y-\mathrm{p}_f(x_0)\| \\ \text{(inequality uniform with respect to the elements of } G(y)). \end{array}\right. \tag{A4}$$

But, if $\|h\|\leqslant\delta$, on also has $\|\mathrm{p}_f(x_0+h)-\mathrm{p}_f(x_0)\|\leqslant\delta$ (this is the magic of the 1-Lipschitz property of p_f) ; so, following (A4):

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

We therefore have proved that $G(\mathrm{p}_f(x_0+h))-G(\mathrm{p}_f(x_0))-JG(\mathrm{p}_f(x_0))(\mathrm{p}_f(x_0+h)-\mathrm{p}_f(x_0))$ is a $o(\|h\|)$, which, with (A2) and (A3), allows us to conclude that the quantity in (A1) is also a $o(\|h\|)$.

Comments.

The proof above has the taste of the so-called theorem of inverse functions, it looks like the theorem of inverse functions, but it is not the theorem of inverse functions. Which made it possible to avoid recourse to the theorem of inverse functions is :

- to know from the beginning that $JG(\mathrm{p}_f(x_0))$ was invertible;
- the control of increments in $\mathrm{p}_f(u)$ by those in u ;
- knowing from the beginning that *there was* an inverse to G, that is to say p_f, whereas in usual Differential Calculus it is a *consequence* of the theorem of inverse functions.

References

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

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

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

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

5. C. Lemaréchal and C. Sagastizabal, *Practical aspects of the Moreau-Yosida regularization: theoretical preliminaries.* SIAM Journal on Optimization 7(2) (1997), $367-385$.

6. 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), $16-25$.

7. 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), $154-166$.

8. J.-B. Hiriart-Urruty, *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.* Pedagogical document (28 pages). Master 2R Operations Research, Toulouse $(2015-2021)$.

9. F. Iutzeler and J. Malick, *Nonsmoothness in Machine Learning: specific structure, proximal identification, and applications.* Set-Valued and Variational Analysis (2020) 28 (4), $661-678$.