Practical aspects of the Moreau-Yosida regularization I: theoretical properties

Claude Lemaréchal, Claudia Sagastizábal

INRIA
INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Practical aspects of the Moreau-Yosida regularization I: theoretical properties

Claude Lemaréchal , Claudia Sagastizábal

N° 2250

May 1994

PROGRAMME 5

Traitement du signal,

automatique

et productique

ISSN 0249-6399

INRIA
ROCQUENCOURT

Practical aspects of the Moreau-Yosida regularization I: theoretical properties

Claude Lemaréchal , Claudia Sagastizábal

Programme 5 — Traitement du signal, automatique et productique
Projet Promath

Rapport de recherche n ° 2250 — May 1994 — 19 pages

Abstract: When computing the infimal convolution of a convex function f with the squared norm, one obtains the so-called Moreau-Yosida regularization of f. Among other things, this function has a Lipschitzian gradient. We investigate some more of its properties, relevant for optimization. Our main result concerns second-order differentiability and is as follows. Under assumptions that are quite reasonable in optimization, the Moreau-Yosida is twice diffferentiable if and only if f is twice differentiable as well. In the course of our development, we give

some results of general interest in convex analysis. In particular, we establish primal-dual relationship between the remainder terms in the first-order development of a convex function and its conjugate.

Key-words: Convex optimization, mathematical programming, proximal point, second-order differentiability.

(Résumé : tsvp)

Unité de recherche INRIA Rocquencourt
Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex (France)
Téléphone : (33 1) 39 63 55 11 – Télécopie : (33 1) 39 63 53 30

Aspects pratiques de la régularisée de Moreau-Yosida I: propriétés théoriques

Résumé : La régularésée de Moreau-Yosida d'une fonction convexe f est l'inf-convolution de f avec le carré de la norme. Cette régularisée a un gradient lipschitzien, nous avons étudié d'autres propriétés, d'intéirêt en optimisation. Notre résultat principal concerne les dérivées secondes. Sous des hypothèses courantes en optimisation, la Moreau-Yosida est deux fois différentiable si et seulement si f l'est aussi. Au cours de notre étude, nous prouvons quelques résultats généraux d'analyse convexe. Nous établissons en particulier des relations primales-duales entre les développements au premier ordre d'une fonction convexe et de sa conjuguée.

1. Introduction. The motivation for this paper is to explore the possibility of introducing efficient preconditioners into the proximal-point algorithm to minimize a convex function f. This algorithm ([2], [8], [14]) is essentially an implicit (sub)gradient method. However it is much more fruitful to see it as the ordinary gradient method applied to a certain perturbation of f: the Moreau-Yosida regularization ([10], [16]), whose minima coincide with those of f. The idea of a preconditioner is thus natural, first steps into this direction were already made in [12], [3]. Naturally, such a preconditioner has to exploit the second-order properties of the perturbed objective function; a study of these properties is therefore a prerequisite to the development of any reasonable algorithm. Here, we address this last, purely theoretical, question; we study also some other properties relevant for optimization; specifically, we relate the smoothness, behaviour at infinity, strong convexity of an objective function to the corresponding properties of its Moreau-Yosida. A subsequent paper will exploit the results obtained here to develop some related algorithms, emphasizing the implementable aspect. Along these lines, we also mention the computational considerations contained in [4], [1], [6], [15], [9], [7].

Our notation follows closely that of [13] and [5]. In the space $\mathbb{R}^N$, the Euclidean product is denoted by $\langle \cdot, \cdot \rangle$, and $\| \cdot \|$ is the associated norm; $B(x, \rho)$ is the ball centered at x with radius ρ. Given a symmetric positive definite linear operator M, we set $\langle \cdot, \cdot \rangle_M := \langle M \cdot, \cdot \rangle$; accordingly, we will shorten $\|x\|_M^2 := \langle x, x \rangle_M$. The smallest and largest eigenvalues of M will be denoted by λ and Λ respectively.

We denote by F the Moreau-Yosida regularization of a given closed convex function f, asssociated to the metric defined by M:

$$F(x) := \min_{y \in \mathbb{R}^n} \left\{ f(y) + \tfrac{1}{2}\|y - x\|_M^2 \right\} =: \left(f \mathbin{\dot{+}} \tfrac{1}{2}\|\cdot\|_M^2 \right)(x), \tag{1}$$

where $\dot{+}$ stands for the infimal convolution.

First-order regularity of F is well known: without any further assumption, F has a Lipschitzian gradient. More precisely, for all x_1, $x_2 \in \mathbb{R}^N$:

$$\|\nabla F(x_1) - \nabla F(x_2)\|^2 \leq \Lambda \langle \nabla F(x_1) - \nabla F(x_2), x_1 - x_2 \rangle . \tag{2}$$

If we denote by $p(x)$ the unique minimizer in (1), called the *proximal* point of x, $\nabla F(x)$ has the following expression

$$G := \nabla F(x) = M(x - p(x)) \in \partial f(p(x)) . \tag{3}$$

Note in particular that f has a nonempty subdifferential at any point p of the form $p(x)$.

The main aim of this paper is to suggest the following: even though the Moreau-Yosida is a powerful tool to bring first-order differentiability, it can hardly go beyond. Indeed, the existence of a Hessian of F normally implies that f itself has a Hessian. This sort of rough statement will be made precise in §4, using a rather sophisticated machinery which we develop in §3. In §2, we review a few elementary results on the Moreau-Yosida F of (1), which are relevant when developing optimization algorithms. Some of them are easy and/or already known, at least for $M = Id$. For the sake of simplicity, we often consider finite-valued objective functions $f : \mathbb{R}^N \to \mathbb{R}$. This avoids some technical difficulties and makes the reading lighter.

2. Properties of the Moreau-Yosida. We study here some properties which F of (1) inherits from f. For the reader's convenience, we start by recalling a few classical results in convex analysis, which will be frequently used in the sequel. They can be found in [13] or Chapter X in [5].

(i) The conjugate of a closed (i.e. lsc) convex function φ is

$$\varphi^*(g) := \sup_{x\in\mathbb{R}^N} \{\langle g,x\rangle - \varphi(x)\} \,. \tag{4}$$

(ii) The conjugacy operation is an involution, i.e. the conjugate of φ^* is φ itself.

(iii) The conjugate of a sum is the infimal convolution of the conjugates.

(iv) If C is a closed convex set, the conjugate of its indicator function I_C (0 on C, $+\infty$ outside) is

$$(\mathrm{I}_C)^*(g) = \sigma_C(g) := \sup\{\langle g,x\rangle \,:\, x\in\mathbb{R}^N\}\,,$$

the support function of C.

(v) The subdifferential of φ^* is the argmax in (4). Said otherwise,

$$g\in\partial\varphi(x) \quad\Longleftrightarrow\quad \langle g,x\rangle - \varphi^*(g) \geq \langle \gamma,x\rangle - \varphi^*(\gamma) \quad \text{for all } \gamma.$$

(vi) The directional derivative of a finite-valued convex function φ is the support function of its subdifferential: $\varphi'(x;\cdot) = \sigma_{\partial\varphi(x)}(\cdot)$.

(vii) Naturally, all these properties imply for example $[\mathrm{I}_{B(0,\varepsilon)}]^*(\cdot) = \varepsilon\|\cdot\|$, and also $[\varphi'(x;\cdot)]^* = \mathrm{I}_{\partial\varphi(x)}$.

We first show that f and F have the same behaviour at infinity. Recall that the recession (or asymptotic) function of a closed convex function φ is defined by

$$\varphi'_\infty(d) = \lim_{t\to+\infty} [\varphi(x+td)-\varphi(x)]/t$$

(a limit which does not depend on $x\in\operatorname{dom}\varphi$). This function is useful because φ has a nonempty bounded set of minima if and only if $\varphi'_\infty(d)>0$ for all $d\neq 0$.

THEOREM 2.1. *The recession functions of f and F are identical.*

Proof. Apply Corollary 9.2.1 in [13]: the recession function of an infimal convolution is the infimal convolution of the recession functions. Then we obtain

$$F'_\infty(d) = \left(f'_\infty \mathbin{\dot\vee} \mathrm{I}_{\{0\}}\right)(d) = \inf_{y=0} f'_\infty(d-y) = f'_\infty(d)\,,$$

because the recession function of a squared norm is clearly $\mathrm{I}_{\{0\}}$. □

Recall that a function φ is said to be *strongly convex* with modulus $c>0$ if and only if $\varphi(\cdot) - \frac{1}{2}c\|\cdot\|^2$ is a convex function. This property plays the role of nondegenerate Hessians in smooth optimization; as such, it is fairly relevant for optimization algorithms. We show that strong convexity is transmitted between f and F. Dually, smoothness is likewise transmitted between f^* and F^*.

THEOREM 2.2. *For a finite-valued convex function f, the following statements are equivalent:*

(*i*) *f is strongly convex with modulus c;*
(*ii*) *f^* has a Lipschitzian gradient with Lipschitz constant $1/c$;*
(*iii*) *F^* has a Lipschitz continuous gradient with constant $1/c + 1/\lambda$;*
(*iv*) *F is strongly convex with modulus $c\frac{\lambda}{\lambda+c}$.*

Proof. Because f and F are finite-valued, Theorems X.4.2.1 and X.4.2.2 in [5] can be applied to yield the equivalences $(i) \iff (ii)$ and $(iii) \iff (iv)$.

Let us prove $(ii) \iff (iii)$. Since F is the infimal convolution of f and $\frac{1}{2}\|\cdot\|^2_M$, its conjugate is the sum of the respective conjugates: $F^*(\cdot) = f^*(\cdot) + \frac{1}{2}\|\cdot\|^2_{M^{-1}}$. Actually

$$\nabla F^*(\cdot) = \nabla f^*(\cdot) + M^{-1}(\cdot)\,,$$

whenever one of the gradients exists. □

We now turn our attention to properties involving the proximal operator more directly. They will be useful for the study of second-order smoothness.

PROPOSITION 2.3. *For any x_1 and x_2 in $\mathbb{R}^N$,*

$$\|p(x_1) - p(x_2)\|^2_M \leq \langle x_1 - x_2, p(x_1) - p(x_2)\rangle_M \ . \tag{5}$$

It follows that the mapping $x \mapsto p(x)$ is Lipschitzian with constant Λ/λ.

Proof. For arbitrary $p_1, p_2 \in \mathbb{R}^N$ and $G_i \in \partial f(p_i)$, the convexity of f gives the monotonicity of the subgradients: $\langle G_1 - G_2, p_1 - p_2\rangle \geq 0$. Take now x_1 and x_2 in $\mathbb{R}^N$, and write the inequality for $p_i := p(x_i)$, G_i from (3):

$$\langle M(x_1 - p(x_1)) - M(x_2 - p(x_2)), p(x_1) - p(x_2)\rangle \geq 0\,,$$

which is (5). From this we can obtain

$$\lambda\|p(x_1) - p(x_2)\|^2 \leq \Lambda\|x_1 - x_2\|\ \|p(x_1) - p(x_2)\|$$

and the Lipschitz property follows inmediately. □

PROPOSITION 2.4. *Assume f is a closed convex function. Then $\nabla F(\cdot)$ has directional derivatives if and only if $p(\cdot)$ has directional derivatives. Moreover, the Hessian $\nabla^2 F(x)$ exists if and only if p has a Jacobian $P'(x)$:*

$$\nabla^2 F(x) = M(Id - P'(x)) \quad \textit{for all } x \in \mathbb{R}^N\,.$$

Proof. Straightforward from (3). □

As observed in [9], a space decomposition reveals important when combining quasi-Newton updates with proximal-point algorithms. Along these lines, we show that the directional derivatives of $p(\cdot)$ lie in $\mathrm{N}_{\partial f(p(x))}(G)$, the normal cone to $\partial f(p(x))$ at G, whenever they exist.

LEMMA 2.5. *For the closed convex function f, let G be defined by (3), and denote by $\mathcal{N} := \mathrm{N}_{\partial f(p(x))}(G)$ the normal cone to $\partial f(p(x))$ at G. Then any cluster point of $\{[p(x+td) - p(x)]/t\}_{t\downarrow 0}$ lies in $\mathcal{N}$.*

Proof. By definition, $p(x+td)$ minimizes (1) with x replaced by $x+td$, therefore

$$f(p(x+td)) + \tfrac{1}{2}\|p(x+td) - x - td\|^2_M \leq f(p(x)) + \tfrac{1}{2}\|p(x) - x - td\|^2_M\,,$$

that is,

$$f(p(x+td)) - f(p(x)) \leq \tfrac{1}{2}\|p(x) - x - td\|_M^2 - \tfrac{1}{2}\|p(x+td) - x - td\|_M^2 .$$

Apply to the lefthand side the subgradient inequality for any $g \in \partial f(p(x))$ and develop the righthand side to obtain

$$\langle p(x+td) - p(x), g\rangle \leq \tfrac{1}{2}\langle p(x) - p(x+td), p(x) + p(x+td) - 2x - 2td\rangle_M .$$

Due to (3), $\langle p(x) - p(x+td), G\rangle = \frac{1}{2}\langle p(x) - p(x+td), 2x - 2p(x)\rangle_M$ which, added to both sides, gives:

$$\langle p(x+td) - p(x), g - G\rangle \leq -\tfrac{1}{2}\|p(x+td) - p(x)\|_M^2 + t\langle p(x+td) - p(x), d\rangle_M .$$

Finally, divide by t:

$$\begin{aligned}\left\langle \frac{p(x+td) - p(x)}{t}, g - G \right\rangle &\leq -\frac{\|p(x+td) - p(x)\|_M^2}{2t} + \langle p(x+td) - p(x), d\rangle_M \\ &\leq \langle p(x+td) - p(x), d\rangle_M ,\end{aligned}$$

for all $g \in \partial f(p(x))$. Passing to the limit we obtain the desired result, because of the continuity of $p(\cdot)$. □

COROLLARY 2.6. *If $p(\cdot)$ has a Jacobian $P'(x)$, then* $\operatorname{Im} P'(x) \subset \mathcal{N}$. *When $x \to x_0$, all the cluster points of $\frac{p(x)-p(x_0)}{\|x-x_0\|}$ lie in $\mathcal{N} \cap B(0, \Lambda/\lambda)$.*

Proof. The first statement is clear from Lemma 2.5. The proof of the second statement is classical for Lipschitzian functions: with x tending to x_0, extract a subsequence such that $\frac{x-x_0}{\|x-x_0\|} \to d$; setting $t := \|x - x_0\| \downarrow 0$, we have

$$x = x_0 + td + q(t) , \qquad \text{with } \frac{q(t)}{t} \to 0 \text{ when } t \downarrow 0 .$$

Then write

$$p(x) - p(x_0) = p(x_0 + td) - p(x_0) + p(x_0 + td + q(t)) - p(x_0 + td)$$

to obtain with Proposition 2.3

$$\left\| \frac{p(x) - p(x_0)}{\|x - x_0\|} - \frac{p(x_0 + td) - p(x_0)}{t} \right\| = \frac{\|p(x_0 + td + q(t)) - p(x_0 + td)\|}{t} \leq \frac{\Lambda}{\lambda}\frac{\|q(t)\|}{t}$$

and the result follows. □

A direct consequence of Lemma 2.5 is that ∇F enjoys automatically some directional differentiability.

PROPOSITION 2.7. *For the closed convex function f, let G be defined by* (3), *and denote by $\mathcal{T}$ the tangent cone to $\partial f(p(x))$ at G. Then, for any d such that $Md \in \mathcal{T}$,*

$$\frac{\nabla F(x+td) - \nabla F(x)}{t} \longrightarrow Md \quad \textit{when } t \downarrow 0 .$$

Proof. From (3), $\nabla F(x+td) - \nabla F(x) = tMd - M(p(x+td) - p(x))$, we only need to show that $[p(x+td) - p(x)]/t$ tends to 0 when $t \downarrow 0$. For this, use (5):

$$\langle M(p(x+td) - p(x)), p(x+td) - p(x)\rangle \leq t \langle p(x+td) - p(x), Md\rangle \ .$$

Observing that the lefthand side is minorized by $\lambda\|p(x+td)-p(x)\|^2$, divide by t^2 to obtain

$$0 \leq \lambda \frac{\|p(x+td) - p(x)\|^2}{t^2} \leq \left\langle \frac{p(x+td) - p(x)}{t}, Md \right\rangle \ .$$

In view of Lemma 2.5, the (bounded) righthand side cannot have any positive cluster point, it must tend to 0 and the proof is complete. □ Of course, because ∇F is Lipschitzian, we can proceed as in Corollary 2.6 to see that, if $x \to x_0$ in such a way that $(x-x_0)/\|x-x_0\| \to d$, with $Md \in \mathcal{T}$, then

$$\frac{\nabla F(x) - \nabla F(x_0)}{\|x - x_0\|} \longrightarrow Md \, .$$

As an illustration, take the bivariate function $f(\xi,\eta) = |\xi| + \frac{1}{2}\eta^2$ and $M = Id$. The optimality condition for the proximal point (π,ρ) of (ξ,η) close to 0 results in

$$\pi = 0 \quad \text{if } |\xi| \leq 1 \, , \qquad \text{and} \qquad \rho = \eta/2 \, .$$

Thus, at $x = 0$, $\partial f(x) = [-1,1] \times \{0\}$, $p(x) = 0$, $G = 0$ and $p(\cdot)$ has the Jacobian $\begin{pmatrix} 0 & 0 \\ 0 & 1/2 \end{pmatrix}$. We see that the non-differentiability of f at 0 in the subspace $\mathcal{T} = \mathbb{R} \times \{0\}$ does not affect the second-order differentiability of F.

We conclude this section with a trivial but not so well-known observation: the proximal mapping has an explicit inverse. This may be very useful when designing algorithms, see [7].

THEOREM 2.8. *Let p be such that $\partial f(p) \neq \emptyset$ and take $G \in \partial f(p)$. Then p is the proximal point of $x := p + M^{-1}G$.*

Proof. We have $M(x-p) \in \partial f(p)$ and this characterizes the proximal point in a unique way, see (3). □

3. Some results in convex analysis. In this section we gather some results of general interest concerning convex sets and functions. They will be instrumental for our second-order analysis.

3.1. On the geometry of convex sets. In the following Propositions we characterize those convex cones that are subspaces. For the subspace $\mathcal{M}$ in Proposition 3.1, see Theorem 2.7 in [13].

PROPOSITION 3.1. *Let $\mathcal{N}$ be a closed convex cone and call $\mathcal{M} := \mathcal{N} \cap \{-\mathcal{N}\}$ the largest subspace contained in $\mathcal{N}$. Then*

$$\mathcal{N} \ \textit{is a subspace if and only if} \ \mathcal{N} \cap \mathcal{M}^{\perp} = \{0\} \, .$$

Moreover, for any $\nu_0 \in \mathcal{M}^\perp$ *and* $\nu \in \mathcal{N}$,

$$\langle \nu_0, \nu \rangle \neq 0 \quad \Longrightarrow \quad -\nu \notin \mathcal{N}\,. \tag{6}$$

Proof. When the convex cone $\mathcal{N}$ is a subspace, it is symmetric: $\mathcal{N} = -\mathcal{N} = \mathcal{M}$. In this case, $\mathcal{N} \cap \mathcal{M}^\perp = \mathcal{M} \cap \mathcal{M}^\perp = \{0\}$.

Conversely, when $\mathcal{N} \cap \mathcal{M}^\perp = \{0\}$, suppose for contradiction that $\mathcal{N}$ is not a subspace; we can take $\nu \in \mathcal{N} \backslash \mathcal{M}$, which can be expressed as a direct sum:

$$\nu = \nu_m + \nu_0 \quad \text{with } \nu_m \in \mathcal{M} \text{ and } 0 \neq \nu_0 \in \mathcal{M}^\perp\,.$$

Since $\mathcal{M}$ is a symmetric set, $-\nu_m \in \mathcal{M} \subset \mathcal{N}$ and, since $\mathcal{N}$ is a convex cone,

$$\nu_0 = \nu + (-\nu_m) \in \mathcal{N}\,;$$

thus we have exhibited a nonzero $\nu_0 \in \mathcal{N} \cap \mathcal{M}^\perp$. This is the required contradiction.

Finally, we have to prove (6). Take $\nu_0 \in \mathcal{M}^\perp$ and $\nu \in \mathcal{N}$. If $-\nu$ were in $\mathcal{N}$, it would be also in $\mathcal{M}$ and this would contradict $\langle \nu_0, \nu \rangle \neq 0$. □

The next result deals with $\mathrm{N}_C(g_0)$, the normal cone to a closed convex set C at $g_0 \in C$. It will be used in Proposition 4.3 with $C := \partial f(p(x))$ and $g_0 := G$ of (3).

PROPOSITION 3.2. *Assume* $C \subset \mathbb{R}^N$ *is a closed convex set and let* $g_0 \in C$. *Then*

$$\mathcal{N} := \mathrm{N}_C(g_0) \textit{ is a subspace} \quad \Longleftrightarrow \quad g_0 \in \mathrm{ri}\, C\,,$$

where ri *denotes the relative interior.*

Proof. Call σ_C the support function of C. By definition of normal cones, $s \in \mathcal{N}$ exactly when $\langle s, g_0 \rangle = \sigma_C(s)$. The convex cone $\mathcal{N}$ is a subspace if and only if it is symmetric. Thus, $\mathcal{N}$ is a subspace if and only if the following property holds:

$$s \in \mathcal{N} \implies \sigma_C(s) + \sigma_C(-s) = 0\,, \qquad [= \langle s, g_0 \rangle + \langle -s, g_0 \rangle]$$

or equivalently

$$\sigma_C(d) + \sigma_C(-d) > 0 \quad \Longrightarrow \quad d \notin \mathcal{N}\,,$$

i.e., applying the definition of $\mathcal{N}$:

$$\sigma_C(d) + \sigma_C(-d) > 0 \quad \Longrightarrow \quad \exists g_d \in C \,:\, \langle d, g_d - g_0 \rangle > 0\,,$$

which in turn can be written

$$\sigma_C(d) + \sigma_C(-d) > 0 \quad \Longrightarrow \quad \sigma_C(d) > \langle d, g_0 \rangle\,.$$

By Theorem 13.1 in [13] or Theorem V.2.2.3(ii) in [5], this last property just expresses $g_0 \in \mathrm{ri}\, C$. □

We now characterize those convex sets that are simultaneously closed and relatively open.

PROPOSITION 3.3. *Assume $C \subset \mathbb{R}^N$ is a closed convex set. Then*

$$C = \operatorname{ri} C \quad \Longleftrightarrow \quad C = \operatorname{aff} C\,,$$

where aff *denotes the affine hull. If, in addition, C is bounded, then $C = \operatorname{ri} C$ if and only if C is a singleton.*

Proof. Since an affine set is relatively open, the "$\Leftarrow$" part is straightforward. Our proof of the "$\Rightarrow$" part is geometric. Assume $C = \operatorname{ri} C$; from Theorem 6.4 in [13], this means

$$\forall x_1, x_2 \in C\,, \quad \sup\{t : tx_1 + (1-t)x_2 \in C\} > 1\,. \tag{7}$$

Fix x_1, x_2 and call T the above supremum; we claim that $T = +\infty$. If T were finite, the vector $x_3 := Tx_1 + (1-T)x_2$ would be in C (C is closed). But then the optimality of T would imply

$$\sup\{t : tx_1 + (1-t)x_3 \in C\} = 1\,,$$

which is impossible: replace x_2 by x_3 in (7).

Exchanging x_1 and x_2, we see that C contains the affine hull of $\{x_1, x_2\}$. We conclude $C = \operatorname{aff} C$, since x_1 and x_2 were arbitrary.

The rest of the statement follows because the only affine bounded sets are the singletons. □

3.2. First-order developments of convex functions. We now present a theory analogous to that of § X.4.2(b) in [5]. Instead of finding quadratic bounds depending on a particular subgradient, we consider here the first-order expansion of a convex function φ:

$$\varphi(x_0 + h) = \varphi(x_0) + \varphi'(x_0; h) + o(\|h\|)\,.$$

Clearly, the remainder term is nonnegative, we want to estimate it more accurately. To bound it from above means to find $\varepsilon > 0$ and some function r satisfying

$$r \text{ is convex, nonnegative, differentiable at } 0,\ r(0) = 0, \nabla r(0) = 0\,,$$

such that

$$\varphi(x_0 + h) \le \varphi(x_0) + \varphi'(x_0; h) + r(h) + \mathrm{I}_{B_\varepsilon}(h) \quad \text{for all } h\,. \tag{8}$$

For a lower bound we need likewise ε and r such that

$$\varphi(x_0 + h) + \mathrm{I}_{B_\varepsilon}(h) \ge \varphi(x_0) + \varphi'(x_0; h) + r(h) \quad \text{for all } h\,. \tag{9}$$

Here and throughout, $\mathrm{I}_{B_\varepsilon}$ denotes the indicator function of $B(0, \varepsilon)$. Note that, despite the appearances, the existence of $\nabla r(0)$ is automatic from (8); by contrast, convexity of r is a more restrictive assumption.

Our aim is to study the duality between (8) and (9): if φ satisfies one of these inequalties, does φ^* satisfy the other? The answer is given in Theorems 3.5 and 3.7 below, which play

the role of respectively Theorems X.4.2.7 and X.4.2.6 in [5]. To establish this, we will always need the following assumption on r:

$$\text{there exists } c>0 \text{ such that} \quad r(h) \geq \frac{1}{2}c\|h\|^2 \quad \text{for all } h\,. \tag{10}$$

We will frequently compute the conjugate of $\psi + \mathrm{I}_{B_\varepsilon}$, i.e. the infimal convolution

$$\Psi_\varepsilon^*(g) := (\psi + \mathrm{I}_{B_\varepsilon})^*(g) = \min_{s\in\mathbb{R}^N}\{\psi^*(g-s)+\varepsilon\|s\|\}\,. \tag{11}$$

This is the so-called Lipschitzian regularization of ψ^*, and we start by showing some properties of this operation, illustrated by Fig. 1.

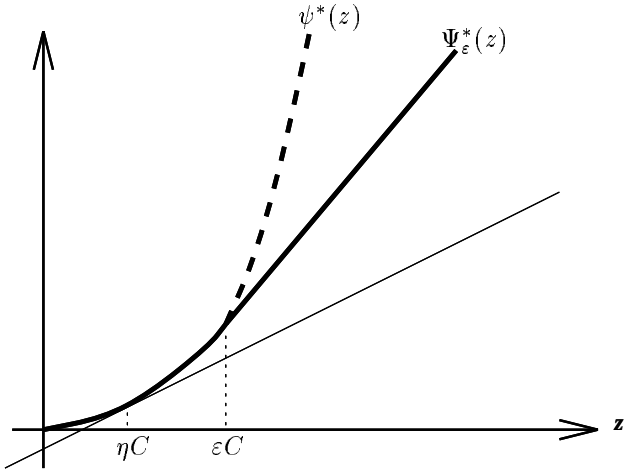

FIG. 1. Lipschitzian Regularization

LEMMA 3.4. *Let ψ be a finite-valued convex function and denote by Ψ_ε^* the Lipschitzian regularization of its conjugate ψ^*.*

(i) If there are $c>0$ and $\delta>0$ such that

$$\psi(h) \geq \psi(0) + \psi'(0;h) + \frac{1}{2}c\|h\|^2 \quad \textit{for } \|h\| \leq \delta\,, \tag{12}$$

then $\Psi_\varepsilon^*(s) = \psi^*(s)$ for all $\varepsilon \leq \delta$ and $s \in \partial\psi(0) + B(0,\varepsilon c/2)$.

(ii) If ψ has the form $\psi(h) = \frac{1}{2}C\|h\|^2$ (i.e. $\psi^* = \frac{1}{2C}\|\cdot\|^2$) for some $C>0$, then

$$\Psi_\varepsilon^*(z) = \begin{cases} \frac{1}{2C}\|z\|^2 & \textit{if } \|z\| \leq \varepsilon C \\ -\frac{\varepsilon^2 C}{2} + \varepsilon\|z\| & \textit{if not.} \end{cases}$$

(iii) With ψ as in (ii) and $\eta \in]0,\varepsilon]$, there holds

$$\Psi_\varepsilon^*(z) \geq -\frac{\eta^2 C}{2} + \eta\|z\| \qquad \textit{for all } z.$$

Proof. [(i)] Apply Proposition XI.3.4.5 of [5]: ψ^* and its regularization Ψ_ε^* coincide on the set

$$\{s \in \mathbb{R}^N : \psi^*(s) = \Psi_\varepsilon^*(s)\} = \{s \in \mathbb{R}^N : \partial\psi^*(s) \cap B(0,\varepsilon) \neq \emptyset\}\,.$$

Since $s \in \partial\psi(0)$ means $0 \in \partial\psi^*(s)$, this coincidence set contains in particular $\partial\psi(0)$.

To say that $\partial\psi^*(s) \cap B(0,\varepsilon) \neq \emptyset$ is to say that $\langle s, \cdot\rangle - \psi(\cdot)$ attains its maximum $\psi^*(s)$ on $B(0,\varepsilon)$. Accordingly, let us find an upper bound for

$$A := \sup_{\|h\|>\varepsilon} \{\langle s, h\rangle - \psi(h)\}\,.$$

For this, use the convexity of ψ on $[0,h]$: if $h \notin B(0,\varepsilon)$,

$$\psi(\varepsilon h/\|h\|) \leq \psi(0) + \frac{\varepsilon}{\|h\|}[\psi(h) - \psi(0)]\,,$$

which, after some algebraic manipulations using (12), gives when $\varepsilon \leq \delta$:

$$\psi(h) \geq \psi(0) + \psi'(0;h) + \frac{1}{2}c\varepsilon\|h\|\,.$$

Therefore

$$\begin{array}{rcl} A & \leq & \sup_{\|h\|>\varepsilon}\{\langle s,h\rangle - \psi(0) - \psi'(0;h) - \frac{1}{2}c\varepsilon\|h\|\} \\ & \leq & -\psi(0) + \sup_{\|h\|\in\mathbb{R}^N}\{\langle s,h\rangle - \psi'(0;h) - \frac{1}{2}c\varepsilon\|h\|\} \\ & = & -\psi(0) + \left(\psi'(0;\cdot) + \frac{1}{2}c\varepsilon\|\cdot\|\right)^*(s) \\ & = & -\psi(0) + \left(\mathrm{I}_{\partial\psi(0)} \,\dot{\nabla}\, \mathrm{I}_{B(0,\varepsilon c/2)}\right)(s)\,. \end{array}$$

The last infimal convolution is zero when $s \in \partial\psi(0) + B(0,\varepsilon c/2)$. For every such s which is not in $\partial\psi(0)$, we thus have

$$A \leq -\psi(0) = \inf\psi^* < \psi^*(s)\,.$$

Then $\langle s,h\rangle - \psi(h)$ attains its maximum on $B(0,\varepsilon)$; (i) is proved.

[(ii)] For our quadratic function ψ, apply the dual definition (11):

$$\begin{array}{rcl} \Psi_\varepsilon^*(z) = \left(\frac{1}{2}C\|\cdot\|^2 + \mathrm{I}_{B_\varepsilon}\right)^*(z) & = & \sup_{x\in B(0,\varepsilon)}\{\langle z,x\rangle - \frac{1}{2}C\|x\|^2\} \\ & = & C\sup_{x\in B(0,\varepsilon)}\{\langle z/C,x\rangle - \frac{1}{2}\|x\|^2\} \\ & = & C\sup_{x\in B(0,\varepsilon)}\{\frac{1}{2}\|z/C\|^2 - \frac{1}{2}\|x - z/C\|^2\} \\ & = & \frac{1}{2C}\|z\|^2 - \frac{C}{2}\inf_{x\in B(0,\varepsilon)}\|x - z/C\|^2 \\ & = & \frac{1}{2C}\|z\|^2 - \frac{C}{2}\mathrm{d}^2_{B(0,\varepsilon)}(z/C)\,. \end{array}$$

To finish the proof observe that the distance function $\mathrm{d}_{B(0,\varepsilon)}(z/C)$ is equal to $\varepsilon - \|z\|/C$ whenever $z/C \notin B(0,\varepsilon)$.

[(iii)] We consider two cases. If $\|z\| \leq \varepsilon C$, the result follows from

$$\Psi_\varepsilon^*(z) - \left(-\eta^2 C/2 + \eta\|z\|\right) = \frac{1}{2C}\|z\|^2 + \frac{1}{2}\eta^2 C - \eta\|z\| = \frac{1}{2C}(\|z\| - \eta C)^2\,.$$

In the other case, set $q(\eta) := -\eta^2 C/2 + \eta\|z\|$ (so $q = \Psi_\varepsilon^*$ when $\eta = \varepsilon$). Compute $q'(\eta) = -\eta C + \|z\|$. We see that q is an increasing function when $-\eta C + \|z\| \geq 0$, in particular when $\eta \leq \varepsilon < \|z\|/C$. □

We are now in a position to establish the "duality" between (8) and (9).

THEOREM 3.5. *Let φ be a finite-valued convex function satisfying* (8). *Assume also that r satisfies* (10) *as well as*

$$r(h) \leq \frac{1}{2} C\|h\|^2 \quad \textit{for all } h\,. \tag{13}$$

Then, for all $g_0 \in \partial\varphi(x_0)$ and $s \in B(0, \dfrac{\varepsilon c^2}{2C})$, we have

$$\varphi^*(g_0+s) \geq \varphi^*(g_0) + \langle s, x_0\rangle + \min_{\gamma\in\partial\varphi(x_0)} r^*(g_0+s-\gamma)\,. \tag{14}$$

Proof. We conjugate both sides in (8): for all $g \in \mathbb{R}^N$

$$\begin{array}{rcl} \varphi^*(g) - \langle g, x_0\rangle & \geq & [\varphi(x_0) + \varphi'(x_0;\cdot) + r + \mathrm{I}_{B_\varepsilon}]^*(g) \\ & = & -\varphi(x_0) + \left(\mathrm{I}_{\partial\varphi(x_0)} \mathbin{\overset{+}{\vee}} (r + \mathrm{I}_{B_\varepsilon})^*\right)(g) \\ & = & -\varphi(x_0) + \left(\mathrm{I}_{\partial\varphi(x_0)} \mathbin{\overset{+}{\vee}} R_\varepsilon^*\right)(g) \\ & = & -\varphi(x_0) + \min_{\gamma\in\partial\varphi(x_0)} R_\varepsilon^*(g-\gamma)\,, \end{array} \tag{15}$$

where R_ε^* denotes the Lipschitzian regularization of r^*, see (11). We will prove that, for g close enough to $\partial\varphi(x_0)$, R_ε^* can be replaced by r^* in (15).

First, observe that (10) and (13) are transformed in the dual space to

$$\frac{1}{2C}\|\cdot\|^2 \leq r^* \leq \frac{1}{2c}\|\cdot\|^2\,. \tag{16}$$

These inequalities are transmitted to the Lipschitzian regularizations; by Lemma 3.4(*iii*) we therefore have, for all $\eta \in]0, \varepsilon]$,

$$R_\varepsilon^*(z) \geq \frac{1}{2}\left(-\eta^2 C + 2\eta\|z\|\right) \qquad \text{for all } z.$$

Apply (16) to obtain by division

$$\frac{r^*(z)}{R_\varepsilon^*(z)} \leq \frac{\|z\|^2}{-\eta^2 cC + 2\eta c\|z\|} \qquad \text{for all } z. \tag{17}$$

Let γ_g denote an optimal γ in (15). From (16) and (17),

$$\frac{1}{2C}\|g-\gamma_g\|^2 \leq r^*(g-\gamma_g) \leq \frac{\|g-\gamma_g\|^2}{-\eta^2 cC + 2\eta c\|g-\gamma_g\|} R_\varepsilon^*(g-\gamma_g)\,.$$

We append to this chain additional upper bounds, using the optimality of γ_g, $R_\varepsilon^* \leq r^*$ and (16): for all $\gamma \in \partial\varphi(x_0)$,

$$\frac{1}{2C} \leq \frac{1}{-\eta^2 cC + 2\eta c\|g-\gamma_g\|}\,\frac{\|g-\gamma\|^2}{2c}\,.$$

After some algebra this results in

$$\|g-\gamma_g\| \le \frac{C}{2\eta c^2}\|g-\gamma\|^2 + \frac{\eta C}{2} \qquad \begin{array}{l}\text{for } \gamma_g \text{ optimal in (15), } \eta \le \varepsilon \\ \text{and } \gamma \text{ arbitrary in } \partial\varphi(x_0).\end{array} \tag{18}$$

To obtain (14), take $\eta = \frac{\varepsilon c}{2C}$, $g_0 \in \partial\varphi(x_0)$ and $s \in B(0, \frac{\varepsilon c^2}{2C})$. For $g := g_0 + s$, let γ in (18) be the projection of g onto $\partial\varphi(x_0)$: we do have $\|g-\gamma_g\| \le \varepsilon c/2$. Then Lemma 3.4($i$) allows us to replace R_ε^* by r^* in (15). The result follows, since $\varphi(x_0)+\varphi^*(g_0) = \langle g_0, x_0\rangle$. □

The computation of the remainder term in (14) may be deemed abstract. However it can be made explicit when using a quadratic r in (8).

COROLLARY 3.6. *Let φ be a finite-valued convex function satisfying (8), with $r = \frac{1}{2}C\|\cdot\|^2$. Then, for all $g_0 \in \partial\varphi(x_0)$ and $s \in B(0, \varepsilon C/2)$, we have*

$$\varphi^*(g_0+s) \ge \varphi^*(g_0) + \langle s, x_0\rangle + \frac{1}{2C}\|g_0+s-\mathcal{P}(g_0+s)\|^2\,,$$

where $\mathcal{P}$ is the projection onto $\partial\varphi(x_0)$.

As a result

$$\langle s, x-x_0\rangle \ge \frac{\|s\|^2}{2C} \quad \textit{for all } s \in \mathrm{N}_{\partial\varphi(x_0)}(g_0) \cap B(0,\varepsilon C/2) \textit{ and } x \in \partial\varphi^*(g_0+s)\,. \tag{19}$$

Proof. Use Theorem 3.5 with $r^* = \frac{1}{2C}\|\cdot\|^2 = \frac{1}{2c}\|\cdot\|^2$. To prove (19), use the subgradient inequality $\varphi^*(g_0) \ge \varphi^*(g_0+s) - \langle s, x\rangle$ for $x \in \partial\varphi^*(g_0+s)$, and observe that g_0+s is projected onto g_0. □

This result reveals a sort of continuity of *particular* subgradients of a convex function. More precisely, take x close to x_0; because φ is finite-valued, any $g \in \partial\varphi(x)$ is close to $\partial\varphi(x_0)$ (closedness of the subdifferential mapping). Then use the Cauchy-Schwarz inequality in (19) to see that g has actually a Lipschitzian behaviour with respect to its projection g_0 onto $\partial\varphi(x_0)$.

It is not clear whether our assumptions (10) and (13) are really essential for the above results. At least, (13) is natural: useful lower bounds for φ^* need nontrivial upper bounds for φ; but the role of (10) is more obscure. By contrast, the assumptions in the following dual counterpart to Theorem 3.5 seem rather minimal.

THEOREM 3.7. *Let φ be a finite-valued convex function satisfying (9). Assume also that r satisfies (10). Then, for all $g_0 \in \partial\varphi(x_0)$, φ^* is differentiable at g_0 ($\nabla\varphi^*(g_0) = x_0$) and, for all $s \in B(0, \varepsilon c/2)$, we have*

$$\varphi^*(g_0+s) \le \varphi^*(g_0) + \langle s, x_0\rangle + \min_{\gamma\in\partial\varphi(x_0)} r^*(g_0+s-\gamma)\,. \tag{20}$$

Proof. Proceed as in the proof of Theorem 3.5. Clearly (12) holds with $\psi = \varphi$ and $\delta = \varepsilon$, so Lemma 3.4(i) can be applied to conjugate the lefthand side of (9):

$$\varphi^*(g) - \langle g, x_0\rangle \le -\varphi(x_0) + \min_{\gamma\in\partial\varphi(x_0)} r^*(g-\gamma) \quad \text{for } g \in \partial\varphi(x_0) + B(0,\varepsilon c/2)\,.$$

Set $g = g_0 + s$, with s as stated, and observe that $\varphi(x_0) + \varphi^*(g_0) = \langle g_0, x_0 \rangle$ to obtain (20).

Now, in view of (10), $r^* \leq \frac{1}{2c}\|\cdot\|^2$, so (20) gives

$$\varphi^*(g_0+s) - \varphi^*(g_0) - \langle s, x_0\rangle \leq \frac{1}{2c} \min_{\gamma\in\partial\varphi(x_0)} \|g_0 + s - \gamma\|^2 \leq \frac{1}{2c}\|s\|^2 .$$

Because the lefthand side is nonnegative, this clearly implies $\nabla\varphi^*(g_0) = x_0$. □

This result explains why full duality cannot hold between (8) and (9): indeed, if (8) implied (9) for φ^*, then Theorem 3.7 applied to φ would imply the existence of $\nabla\varphi(x_0)$.

4. Second-order analysis. The aim of this section is to study the statement "F has second-order derivatives if and only if f has second-order derivatives". We start with the easy part.

PROPOSITION 4.1. *Let f be a finite-valued convex function and x_0 such that $\nabla^2 f(p(x_0))$ exists. Then the Hessian of F exists at x_0, more precisely*

$$\nabla^2 F(x_0) = M - M[\nabla^2 f(p(x_0)) + M]^{-1} M$$

(here ∇^2 means the Hessian in the classical sense).

Proof. The assumptions imply the existence of ∇f in a neighbourhood of $p(x_0)$. Then (3) shows that $p(x)$ is implicitly defined by

$$\Phi(x,p) := \nabla f(p) + M(p-x) = 0 .$$

Since Φ has Jacobians $\nabla_x\Phi = -M$ and $\nabla_p\Phi(x_0, p(x_0)) = \nabla^2 f(p(x_0)) + M$, the Implicit Function Theorem applies: $P'(x_0) = [\nabla^2 f(p(x_0)) + M]^{-1} M$. The result follows from Proposition 2.4. □

COROLLARY 4.2. *Let f be a finite-valued convex function and x_0 such that $\nabla^2 f(p(x_0))$ exists. Then*

$$\ker \nabla^2 F(x_0) = \ker \nabla^2 f(p(x_0)) .$$

Proof. Use the notation $A := \nabla^2 f(p(x_0))$ and $A' := \nabla^2 F(x_0)$. From Proposition 4.1, $M^{-1}A' = Id - [A+M]^{-1}M = Id - [M^{-1}A + Id]^{-1}$, hence

$$Id - M^{-1}A' = (Id + M^{-1}A)^{-1} .$$

If $A'v = 0$, then $(Id + M^{-1}A)v = v$ and $Av = 0$. Taking inverses, $(Id - M^{-1}A')^{-1} = Id + M^{-1}A$ and we show likewise that $Av = 0$ implies $A'v = 0$. □

The converse part of Theorem 4.1 is not so simple; it will be stated in Theorems 4.7 and 4.8 below. First the next geometrical result is crucial.

PROPOSITION 4.3. *Let f be a finite-valued strongly convex function such that, given x, (8) holds with $x_0 = p(x)$ and $r = \frac{1}{2}C\|\cdot\|^2$. More precisely,*

$$(21) f(p(x)+h) \leq f(p(x)) + f'(p(x);h) + \frac{1}{2}C\|h\|^2 + \mathrm{I}_{B_\varepsilon}(h) \quad \textit{for some } C>0 \textit{ and all } h.$$

If $\nabla^2 F(x)$ exists, then G of (3) lies in the relative interior of $\partial f(p(x))$.

Proof. Let F have a Hessian at x, hence, by Proposition 2.4, $P'(x)$ exists.

Assume for contradiction $G \in \mathrm{rbd}\,\partial f(p(x))$; by Proposition 3.2, the normal cone $\mathcal{N} = \mathrm{N}_{\partial f(p(x))}(G)$ is not a subspace. In view of Proposition 3.1, we can take a unitary $\nu_0 \in \mathcal{N} \cap \mathcal{M}^\perp$; then take $G_t := G + t\nu_0$, with $t > 0$. Calling c the modulus of strong convexity of f, Theorem 2.2 guarantees that $p_t := \nabla f^*(G_t)$ satisfies the Lipschitz condition

$$\|p_t - p(x)\| \le \frac{1}{c}\|G_t - G\| = \frac{1}{c}t\,. \tag{22}$$

By Theorem 2.8 this p_t is the proximal point of $x_t := p_t + M^{-1}G_t$ which, because of (22), satisfies

$$\|x_t - x\| \le \|p_t - p(x)\| + \frac{1}{\lambda}\|G_t - G\| \le (\frac{1}{c} + \frac{1}{\lambda})t\,. \tag{23}$$

Furthermore, G_t is projected onto G: $\mathcal{P}(G_t) = G$ and, whenever $t \in]0, \varepsilon C/2]$, (19) applied to f with $x_0 = p(x)$, $x = p_t$ and $g_0 = G$ gives

$$\langle G_t - G, p_t - p(x)\rangle \ge \frac{1}{2C}\|G_t - G\|^2 = \frac{1}{2C}t^2\,.$$

Combine this with (23):

$$\frac{\lambda c}{2C(c+\lambda)} \le \left\langle \nu_0, \frac{p_t - p(x)}{\|x_t - x\|}\right\rangle\,.$$

Let $t \downarrow 0$. By (23), $x_t \to x$; by Corollary 2.6 (extract a subsequence if necessary) $[p_t - p(x)]/\|x_t - x\| \to \nu$; clearly, $\langle \nu_0, \nu\rangle > 0$. By Lemma 2.5, $\nu \in \mathcal{N}$ and, by (6), $-\nu \notin \mathcal{N}$. This shows that $\mathrm{Im}P'(x)$ is not a symmetric set; $P'(x)$ cannot be a linear operator, and this is the required contradiction. □

We can now establish a local and a global second-order result, valid for strongly convex functions.

PROPOSITION 4.4. *Let f be a finite-valued strongly convex function such that (21) holds for a given x. If $\nabla^2 F(x)$ and $\nabla f(p(x))$ exist, then $\nabla^2 f(p(x))$ exists.*

Proof. We have from (3) $p(x) = x - M^{-1}G$ with $G = \nabla f(p(x))$. Apply Corollary 3.6 with $\varphi = f$, $x_0 = p(x)$ and $g_0 = G$; since $\partial f(p(x)) = \{G\}$, $G + s$ is projected onto G for all s:

$$f^*(G+s) \ge f^*(G) + \langle s, p(x)\rangle + \frac{1}{2C}\|s\|^2 \quad \text{for } \|s\| \le \varepsilon C/2\,. \tag{24}$$

By Corollary X.4.2.9 in [5], the existence of $\nabla^2 F(x)$ (positive definite, recall Theorem 2.2) implies the existence of $\nabla^2 F^*(G) = \nabla^2 f^*(G) + M$. Therefore $\nabla^2 f^*(G)$ exists and, by (24), is positive definite. Again by Corollary X.4.2.9 in [5], f is twice differentiable at $p(x)$. □

PROPOSITION 4.5. *Let f be a finite-valued strongly convex function such that (21) holds for all x. If $\nabla^2 F$ exists on the whole of $\mathbb{R}^N$, then $\nabla^2 f$ exists on the whole of $\mathbb{R}^N$.*

Proof. We claim that f is differentiable at every $p \in \mathbb{R}^N$. Indeed, if $\partial f(p)$ is not a singleton, take a subgradient G in the relative boundary of $\partial f(p)$ (Proposition 3.3). Because

of Theorem 2.8, p is the proximal point of $x := p + M^{-1}G$ and, by assumption, $\nabla^2 F(x)$ exists. From Proposition 4.3 we get the desired contradiction: G lies in the relative interior of $\partial f(p)$.

Then $\nabla^2 F$ and ∇f exist on the whole of $\mathbb{R}^N$ and Proposition 4.4 applies. □

Our final goal will be to eliminate the strong convexity assumption in these last two results. For this we perturb f to a strongly convex function $f + \frac{1}{2}\tau\|\cdot\|_M^2$ and we study the effect of this perturbation on the proximal point.

PROPOSITION 4.6. *Let f be a finite-valued convex function satisfying* (21) *for a given x, and define $f_\tau := f + \frac{1}{2}\tau\|\cdot\|_M^2$, for $\tau \in [0,1[$. Consider the Moreau-Yosida of f_τ associated to the metric defined by $(1-\tau)M$:*

$$F_\tau(x) := \min_{y\in\mathbb{R}^n} \left\{ f_\tau(y) + \tfrac{1}{2}(1-\tau)\|y-x\|_M^2 \right\} \tag{25}$$

and denote by $q_\tau(x)$ the unique minimizer of (25).

Then the following statements hold

(i) The function f_τ is strongly convex and satisfies (21) *with C replaced by $C + \tau\Lambda$;*

(ii) for all x, $p(x) = q_\tau(\frac{x}{1-\tau})$.

Proof. The strong convexity of f_τ is clear. To prove that (21) holds for f_τ, add $\frac{1}{2}\tau\|x_0+h\|_M^2$ to both sides of (21) applied to f. Then use the properties $f'_\tau(x_0;h) = f'(x_0;h) + \frac{1}{2}\tau\langle Mx_0, h\rangle$ and $\frac{1}{2}\tau\|\cdot\|_M^2 \le \frac{1}{2}\tau\Lambda\|\cdot\|^2$.

For proving (ii), write the optimality condition for $p(x)$ and $q_\tau(\frac{x}{1-\tau})$:

$$\begin{array}{ll} p(x) \text{ solves} & M(x-p) \in \partial f(p) \quad \text{and} \\ q_\tau(\frac{x}{1-\tau}) \text{ solves} & (1-\tau)M(\frac{x}{1-\tau} - p) \in \partial f_\tau(p)\,. \end{array}$$

Since $\partial f_\tau(p) = \partial f(p) + \{\tau Mp\}$, we are done. □

Thus, passing from f to f_τ can be absorbed by perturbing M to $(1-\tau)M$ in the Moreau-Yosida of (1). Then Proposition 4.5 can be applied to the perturbed data. Our global and local results become as follows.

THEOREM 4.7. *Let f be a finite-valued convex function such that* (21) *holds for all x. Then $\nabla^2 F$ exists on the whole of $\mathbb{R}^N$ if and only if $\nabla^2 f$ exists on the whole of $\mathbb{R}^N$.*

Proof. The "only if" part is Theorem 4.1. As for the "if" part, suppose $\nabla^2 F$ exists everywhere; hence, from Proposition 2.4, $p(\cdot)$ has a Jacobian everywhere. Then consider f_τ as in Proposition 4.6; by virtue of Proposition 4.6(ii), q_τ also has a Jacobian $Q'_\tau(\frac{x}{1-\tau}) = (1-\tau)P'(x)$ for all x. Apply now Proposition 2.4 to F_τ: $\nabla^2 F_\tau = (1-\tau)M(Id - Q'_\tau)$, so F_τ has a Hessian everywhere. Since f_τ is strongly convex and satisfies (21), Proposition 4.5 holds: $\nabla^2 f_\tau$ exists everywhere. Then $\nabla^2 f = \nabla^2 f_\tau - \tau M$ exists everywhere. □

THEOREM 4.8. *Let f be a finite-valued convex function such that* (21) *holds for an optimal $x = \bar{x}$. Assume $\nabla f(\bar{x})$ exists. Then $\nabla^2 F(\bar{x})$ exists if and only if $\nabla^2 f(\bar{x})$ exists.*

Proof. Without loss of generality assume that $\bar{x} = 0$; hence $p(\bar{x}) = \bar{x} = 0$. Copy the proof of Theorem 4.7, replacing "everywhere" by "at $0 = \bar{x} = p(\bar{x}) = q_\tau(\bar{x})$" and using the relation $Q'(0) = (1-\tau)P'(0)$. □

5. Concluding Remarks. The very first motivation for the Moreau-Yosida regularization was to solve ill-conditioned systems of linear equations ([2], Chap.V). In fact, suppose f is quadratic, its Hessian A having extreme eigenvalues c and C. From Proposition 4.1, F is also quadratic with Hessian $M + M(A+M)^{-1}M$. Taking $M = \lambda Id$, a quick calculation shows that the condition number C/c of A is divided by $(\lambda + C)/(\lambda + c)$. This is a clear beneficial effect of the Moreau-Yosida.

Consider now a general objective function. Barring all implementation considerations, assume that the proximal point $p(x)$ can be computed for each x (perhaps approximately, but for a negligible computation cost). Then the question arises whether such a computation is of any use for a superlinearly convergent algorithm minimizing F (i.e. f).

When f is differentiable on the whole space, Theorems 4.7 and 4.2 tell us that a combination Moreau-Yosida & Newton brings essentially nothing. Either F still does not enjoy the necessary properties of smoothness and non-degeneracy, or a superlinear algorithm could have been applied to f at the first place (ordinary Newton, quasi-Newton, or nonsmooth Newton as in [11]). For example, take an augmented Lagrangian:

$$f(x) := f_0(x) + \frac{\pi}{2}\left[\max\left(0, f_1(x) + \frac{\mu}{\pi}\right)\right]^2 ,$$

which is associated to the NLP: minimize f_0 subject to $f_1 \leq 0$. The minimization of the corresponding F (for given μ, π and M) is not easier than the minimization of f: $\nabla^2 F$ exists and is positive definite only when $\nabla^2 f$ enjoys the same properties.

On the other hand, suppose f is a non-differentiable function. Then our results are no longer so conclusive: as far as superlinear convergence is concerned, the important property is the second-order differentiability *at a minimum point*. Theorem 4.7 does not single out such an optimal point, while Theorem 4.8 does not apply at an optimal (kinky) point. To illustrate this deficiency, take the bivariate function $f(x) := \max\{\frac{1}{2}\|x\|^2 - \alpha\langle e, x\rangle, \langle e, x\rangle\}$, for $e := (0,1)^T$ and α a nonnegative parameter. 4 The kinks of f form a circle, denoted by C in Fig. 2. The subdifferential of f at 0 is the segment $[-\alpha e, e]$ hence f is minimized at $0 = p(0)$, for all $\alpha \geq 0$. For $M = Id$, let us compute the proximal point of $x \neq 0$:

$$x - p = \begin{cases} p - \alpha e & \text{if } \frac{1}{2}\|p\|^2 - \alpha\langle e, p\rangle > \langle e, p\rangle \\ \mu(p - \alpha e) + (1-\mu)e & \text{for some } \mu \in [0,1] \text{ if } \frac{1}{2}\|p\|^2 - \alpha\langle e, p\rangle = \langle e, p\rangle \\ e & \text{if } \frac{1}{2}\|p\|^2 - \alpha\langle e, p\rangle < \langle e, p\rangle \ . \end{cases}$$

Working out the calculations, we find that

$$p(x) = \begin{cases} \dfrac{x + \alpha e}{2} & \text{if } \|x - (\alpha+2)e\| > 2(\alpha+1) \\ \dfrac{x + (\alpha\mu + \mu - 1)e}{1+\mu} & \text{if } \alpha + 1 \leq \|x - (\alpha+2)e\| \leq 2(\alpha+1) \\ x - e & \text{if } \|x - (\alpha+2)e\| < \alpha + 1 \, , \end{cases}$$

where

$$\mu = \mu(x) := \frac{\|x - (\alpha+2)e\|}{\alpha+1} - 1 \, . \tag{26}$$

In a condensed form,

$$p(x) = \frac{x - e + (\alpha+1)\nu(x)e}{1+\nu(x)}\,, \tag{27}$$

where $\nu(x)$ is the projection of $\mu(x)$ in (26) onto $[0,1]$.

In Fig. 2, C_1 (resp. C_2) is the boundary of the region where the first (resp. second) function prevails at $p(x)$. The dashed crown is the locus of those x such that $p(x)$ is a kink. The point is that C_2 is always far from 0, while C_1 does contain 0 when $\alpha = 0$. As a result, $\nabla^2 F(0)$ exists if $\alpha > 0$ but not if $\alpha = 0$. To show this, we consider two cases.

(i) When $\alpha > 0$, we have $\mu \in [0,1]$ in (26) for small $\|x\|$. This comes from $\mu(0) = 1/(\alpha+1)$, together with the continuity of $\mu(\cdot)$. In this region, which includes the origin in its interior,

$$p(x) = (\alpha+1)\frac{x-(\alpha+2)e}{\|x-(\alpha+2)e\|} + (\alpha+1)e\,.$$

A mere differentiation gives

$$P'(0) = \frac{\alpha+1}{\alpha+2}\left(Id - ee^T\right) = \frac{\alpha+1}{\alpha+2}\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}\,.$$

(ii) When $\alpha = 0$, the origin is on C_1. Analytically, $\nu(x) = 1$ in (27) whenever $\|x-2e\| > 2$. From this observation, the directional derivatives are easy to compute:

$$P'(0;d) = \begin{cases} \frac{1}{2}d_1 & \text{for } d_2 > 0 \\ \frac{1}{2}d_1 + \frac{1}{2}d_2 & \text{for } d_2 \le 0\,. \end{cases}$$

Here, the nonexistence of $P'(0)$ illustrates Theorem 4.3: $G = 0$ is on the relative boundary of $\partial f(0) = [0,1]$.

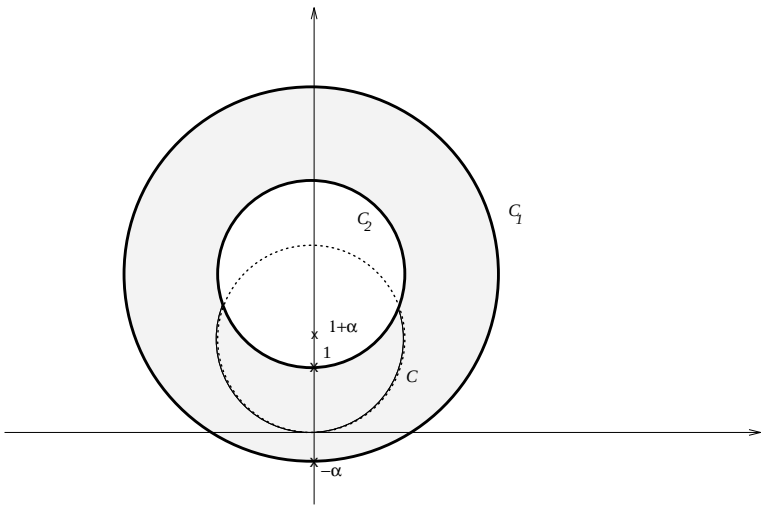

FIG. 2. Moreau-Yosida without Hessian

Let us sum up our results: as far as minimization algorithms are concerned, efficient combinations of Moreau-Yosida with Newtonian schemes might not be so straightforward. For instance, the algorithmic pattern (AP1) in [3] seems of little relevance. At present we can foresee two ways round this difficulty:

(i) A metric M in (1) that varies along the minimizing algorithm (see (AP2) and (AP3) in [3] or Algorithm 13 in [7]).

(ii) A space decomposition. From Proposition 2.7, F has a nice second-order behaviour in the tangent cone $\mathcal{T}$; as for the normal cone $\mathcal{N}$, an approximation scheme is proposed in [9]. Thanks to the property $\mathbb{R}^n = \mathcal{T} \oplus \mathcal{N}$ (Theorem III.3.2.5 in [5]), second-order information can thus be collected in the whole space.

REFERENCES

[1] A. Auslender, *Numerical methods for nondifferentiable convex optimization*, Mathematical Programming Study, 30 (1987), pp. 102–126.

[2] R. Bellman, R. Kalaba, and J. Lockett, *Numerical Inversion of the Laplace Transform*, Elsevier, 1966.

[3] J. Bonnans, J. Gilbert, C. Lemaréchal, and C. Sagastizábal, *A family of Variable Metric Proximal methods*, Mathematical Programming, To appear. Also as Rapport de Recherche INRIA #1851 (1993).

[4] M. Fukushima, *A descent algorithm for nonsmooth convex programming*, Mathematical Programming, 30 (1984), pp. 163–175.

[5] J. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms*, Springer-Verlag, 1993. (two volumes).

[6] K. Kiwiel, *Proximity control in bundle methods for convex nondifferentiable minimization*, Mathematical Programming, 46 (1990), pp. 105–122.

[7] C. Lemaréchal and C. Sagastizábal, *An approach to variable metric bundle methods*, in Systems Modelling and Optimization, J. Henry and J.-P. Yvon, eds., no. 197 in Lecture Notes in Control and Information Sciences, Springer-Verlag, to appear. Also as Rapport de Recherche INRIA #2128, 1994.

[8] B. Martinet, *Régularisation d'inéquations variationnelles par approximations successives*, Revue Française d'Informatique et Recherche Opérationnelle, R-3 (1970), pp. 154–179.

[9] R. Mifflin, *A quasi-second-order proximal bundle algorithm*, Technical Report 93-3, University of Washington, Pullman, (1993).

[10] J. Moreau, *Proximité et dualité dans un espace hilbertien*, Bulletin de la Société Mathématique de France, 93 (1965), pp. 273–299.

[11] L. Qi and J. Sun, *A nonsmooth version of Newton's method*, Mathematical Programming, 58 (1993), pp. 353–367.

[12] M. Qian, *The variable metric proximal point algorithm: application to optimization*, manuscript, Department of Mathematics, GN-50, University of Washington, Seattle, WA 98195, 1992.

[13] R. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, New Jersey, 1970.

[14] ———, *Monotone operators and the proximal point algorithm*, SIAM Journal on Control and Optimization, 14 (1976), pp. 877–898.

[15] H. Schramm and J. Zowe, *A version of the bundle idea for minimizing a nonsmooth function: conceptual idea, convergence analysis, numerical results*, SIAM Journal on Optimization, 2 (1992), pp. 121–152.

[16] K. Yosida, *Functional Analysis*, Springer Verlag, 1964.

Unité de recherche INRIA Lorraine, Technopôle de Nancy-Brabois, Campus scientifique,
615 rue du Jardin Botanique, BP 101, 54600 VILLERS LÈS NANCY
Unité de recherche INRIA Rennes, Irisa, Campus universitaire de Beaulieu, 35042 RENNES Cedex
Unité de recherche INRIA Rhône-Alpes, 46 avenue Félix Viallet, 38031 GRENOBLE Cedex 1
Unité de recherche INRIA Rocquencourt, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex
Unité de recherche INRIA Sophia-Antipolis, 2004 route des Lucioles, BP 93, 06902 SOPHIA-ANTIPOLIS Cedex

Éditeur
INRIA, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex (France)
ISSN 0249-6399