SIAM J. OPTIM.
Vol. 7, No. 2, pp. 367–385, May 1997

006

PRACTICAL ASPECTS OF THE MOREAU–YOSIDA REGULARIZATION: THEORETICAL PRELIMINARIES*

CLAUDE LEMARÉCHAL† AND CLAUDIA SAGASTIZÁBAL†

Abstract. When computing the infimal convolution of a convex function f with the squared norm, the so-called Moreau–Yosida regularization of f is obtained. Among other things, this function has a Lipschitzian gradient. We investigate some more of its properties, relevant for optimization. The most important part of our study concerns second-order differentiability: existence of a second-order development of f implies that its regularization has a Hessian. For the converse, we disclose the importance of the decomposition of $\mathbb{R}^N$ along $\mathcal{U}$ (the subspace where f is "smooth") and $\mathcal{V}$ (the subspace parallel to the subdifferential of f).

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

AMS subject classifications. Primary, 26B05; Secondary, 65K10

PII. S1052623494267127

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 (see [2], [13], [23]) 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 (see [15], [27]), whose minima coincide with those of f. The introduction of a preconditioner into this gradient method is thus natural; first steps in this direction were already made in [21], [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. We address this last, purely theoretical, question here; we also study some other properties relevant for optimization. Specifically, we relate the smoothness, behavior at infinity, and strong convexity of an objective function to the corresponding properties of its Moreau–Yosida regularization; for this, we use extensively the technical results of [11]. The companion paper [28] exploits the results obtained here to develop some related algorithms, emphasizing the implementable aspect; along these lines, we also mention the computational considerations contained in [5], [1], [9], [25], [14], [10].

Our notation follows closely that of [22] and [7]. 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. The conjugate of a closed (i.e., lower semicontinuous) convex function φ is

$$\varphi^*(g) := \sup_{x \in \mathbb{R}^N} \{\langle g, x \rangle - \varphi(x)\} . \tag{1}$$

Recall that the conjugacy operation is an involution; i.e., the conjugate of φ^* is φ itself. The indicator function of a closed convex set S (0 on S, $+\infty$ outside) is denoted by I_S. Given a symmetric positive definite linear operator M, we set $\langle \cdot, \cdot \rangle_M := \langle M \cdot, \cdot \rangle$;

* Received by the editors May 4, 1994; accepted for publication (in revised form) October 31, 1995.

http://www.siam.org/journals/siopt/7-2/26712.html

† INRIA, BP 105, 78153 Le Chesnay, France (claude.lemarechal@inria.fr, claudia.sagastizabal@inria.fr).

accordingly, we will shorten $\frac{1}{2}\|x\|_M^2 := \frac{1}{2}\langle x, x\rangle_M$, whose conjugate is $\frac{1}{2}\|g\|_{M^{-1}}^2$. 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 associated to the metric defined by M:

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

where $\overset{+}{\vee}$ stands for the infimal convolution. The dual relation

$$F^*(\cdot) = f^*(\cdot) + \frac{1}{2}\|\cdot\|_{M^{-1}}^2$$

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

$$\|\nabla F(x_1) - \nabla F(x_2)\|^2 \le \Lambda \left\langle \nabla F(x_1) - \nabla F(x_2), x_1 - x_2 \right\rangle . \tag{3}$$

(Note that the Lipschitz property comes with Cauchy–Schwarz.) If we denote by $p(x)$ the unique minimizer in (2), called the *proximal* point of x, $\nabla F(x)$ has the following expression:

$$G := \nabla F(x) = M(x - p(x)) \in \partial f(p(x))\,. \tag{4}$$

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

Our paper is organized as follows. First, we review a few elementary results on the Moreau–Yosida regularization F of (2), which are relevant when developing optimization algorithms. Some of them are easy and/or already known, at least for $M = \mathcal{I}$, the identity operator. Then in section 3 comes the main content of this paper: a study of second-order differentiability. We give a detailed answer to the question "when does F have a *Hessian*?" We also touch on second-order differentiability in the *epigraphical* sense, [24], which yields related but complementary results. This question is also addressed in [12] and [17]. In this last reference, the fairly general class of prox-regular functions, which contains lower-C^2, primal-lower-nice, and strongly amenable functions, is considered. Our present work is limited to a convex f; moreover, we will often consider the finite-valued case, $f : \mathbb{R}^N \to \mathbb{R}$. This avoids some technical difficulties and makes the reading lighter.

2. Properties of the Moreau–Yosida regularization.

We study here some properties which F of (2) inherits from f.

We first show that f and F have the same behavior 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 [22]: the recession function of an infimal convolution is the infimal convolution of the recession functions. Because the recession function of a squared norm is clearly $\mathrm{I}_{\{0\}}$, we obtain

$$F'_\infty(d) = \left(f'_\infty \mathbin{\overset{+}{\vee}} \mathrm{I}_{\{0\}} \right)(d) = \inf_{y=0} f'_\infty(d-y) = f'_\infty(d). \qquad \square$$

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 $1/\ell$;*
(ii) *f^* has a Lipschitzian gradient with Lipschitz constant ℓ;*
(iii) *F^* has a Lipschitzian gradient with Lipschitz constant L;*
(iv) *F is strongly convex with modulus $1/L$.*

Furthermore, we have the inequalities $\ell - 1/\lambda \le L \le \ell + 1/\lambda$.

Proof. Because f and F are finite valued, Theorems X.4.2.1 and X.4.2.2 in [7] 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 $\frac{1}{2}\|\cdot\|_M^2$, its conjugate is the sum of the respective conjugates: $F^*(\cdot) = f^*(\cdot) + \frac{1}{2}\|\cdot\|_{M^{-1}}^2$. Actually,

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

whenever one of the gradients exists. The equivalence between the Lipschitz properties is then clear; as for the relations between the constants, apply appropriate triangular inequalities. ∎

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 $\mathbb{R}^N$,*

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

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; $\langle G_1 - G_2, p_1 - p_2\rangle \ge 0$. Now take x_1 and x_2 in $\mathbb{R}^N$, and write the inequality for $p_i := p(x_i)$, G_i from (4):

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

which is (5). From this we can obtain

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

and the Lipschitz property follows immediately. ∎

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. The Hessian $\nabla^2 F(x)$ exists if and only if the Jacobian $\nabla p(x)$ exists:*

$$\nabla^2 F(x) = M(\mathcal{I} - \nabla p(x)) \quad \text{for all } x \in \mathbb{R}^N.$$

Proof. The proof is straightforward from (4). ∎

As observed in [14], a space decomposition may be important when combining quasi-Newton updates with proximal-point algorithms. Along these lines, we show that when $x \to x_0$, $p(x)$ is asymptotically close to the normal cone to $\partial f(p(x_0))$ at G. First we recall a well-known property of convex functions.

LEMMA 2.5. *Let f be a closed convex function and let $z_0 \in \operatorname{ri}\operatorname{dom} f$. Suppose $t \downarrow 0$ and $(z-z_0)/t$ has a cluster point ℓ; let $g \in \partial f(z)$ have a cluster point $g_0 \in \partial f(z_0)$. Then $\ell \in \mathcal{N} := \mathrm{N}_{\partial f(z_0)}(g_0)$, the normal cone to $\partial f(z_0)$ at g_0.*

Proof. Take any $\gamma \in \partial f(z_0)$; from convexity, $\langle g - \gamma, z - z_0\rangle \geq 0$. Dividing by $t > 0$ and passing to the limit, we obtain $\langle g_0 - \gamma, \ell\rangle \geq 0$. ∎

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

Proof. Set $g_0 = G = \nabla F(x_0) \in \partial f(p(x_0))$, $z_0 = p(x_0)$, $z = p(x)$, $g = \nabla F(x) \in \partial f(p(x))$, and $t = \|x - x_0\|$. Because F is continuously differentiable, $g \to G$. Then apply Lemma 2.5 and Proposition 2.3. ∎

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

COROLLARY 2.7. *For the closed convex function f, let G be defined by* (4), *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 (4), $\nabla F(x+td) - \nabla F(x) = tMd - M(p(x+td) - p(x))$; 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 left-hand 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 Corollary 2.6, the (bounded) right-hand side cannot have any positive cluster point; it must tend to 0 and the proof is complete. ∎

Of course, owing to the Lipschitz property of ∇F, a classical argument enables the following improvement of this directional result: if $x \to x_0$ in such a way that $(x - x_0)/\|x - x_0\| \to d$ with $Md \in \mathcal{T}$, then

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

To illustrate Corollary 2.7, take the bivariate function $f(\xi, \eta) = |\xi| + \frac{1}{2}\eta^2$ and $M = \mathcal{I}$. 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 $\left(\begin{smallmatrix} 0 & 0 \\ 0 & 1/2 \end{smallmatrix}\right)$. We see that the nondifferentiability 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 often forgotten observation: the proximal mapping has an explicit inverse. This may be very useful when designing algorithms; see [10].

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 (4). ∎

3. Second-order analysis. The aim of this section is to relate second-order derivatives of F and f. For the continuously differentiable F, there is no need of generalizing the classical notion of Hessian. For f, however, the multivalued ∂f calls for a special concept. We will say that the finite-valued convex function f admits at z_0 a generalized Hessian $\mathrm{H}f(z_0)$ when

(i) the gradient $\nabla f(z_0)$ exists,

(ii) there exists a symmetric positive semidefinite operator $\mathrm{H}f(z_0)$ such that

$$f(z_0 + h) = f(z_0) + \langle \nabla f(z_0), h\rangle + \frac{1}{2}\langle \mathrm{H}f(z_0)h, h\rangle + o(\|h\|^2)\,. \tag{6}$$

An important result of [6] is that (6) is equivalent to

$$\partial f(z_0 + h) \subset \nabla f(z_0) + \mathrm{H}f(z_0)h + o(\|h\|)B\,, \tag{7}$$

where B is the unit ball. Note also that when ∂f is single valued in a neighborhood of z_0, $\mathrm{H}f(z_0)$ is the classical Hessian $\nabla^2 f(z_0)$.

We present our study in several steps. First, we consider f strongly convex and differentiable; next, we eliminate the differentiability assumption. Finally, we take a general convex finite-valued function. We are interested in relating the existence of $\nabla^2 F(x_0)$ and $\mathrm{H}f(p_0)$, with $p_0 = p(x_0)$. The following growth condition plays a central role for most of our results:

$$f(p_0 + h) \le f(p_0) + f'(p_0; h) + \frac{1}{2}C\|h\|^2 \quad \text{for some } C > 0 \text{ and all } h \in B(0, \varepsilon). \tag{8}$$

3.1. Differentiable case. The following result has been proved independently by J.-B. Hiriart-Urruty and L. Q. Qi.

THEOREM 3.1. *Let the finite-valued convex function f have a generalized Hessian at $p(x_0)$. Then the Hessian of F exists at x_0; more precisely,*

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

Proof. In view of Proposition 2.4, we only need to exhibit $\nabla p(x_0)$. Write (7) with z_0 and $z_0 + h$ replaced by $p(x_0)$ and $p(x_0 + h)$, respectively:

$$\partial f(p(x_0+h)) \subset \nabla f(p(x_0)) + \mathrm{H}f(p(x_0))(p(x_0+h) - p(x_0)) + o(\|p(x_0+h) - p(x_0)\|)B\,.$$

Because $p(\cdot)$ is Lipschitzian (Proposition 2.3), $o(\|p(x_0 + h) - p(x_0)\|) = o(\|h\|)$. Multiply by M^{-1} and add $p(x_0 + h)$ to both sides to obtain the following:

$$\begin{aligned} M^{-1}\partial f(p(x_0 + h)) + p(x_0 + h) \quad &\subset \quad M^{-1}\nabla f(p(x_0)) + p(x_0 + h) \\ &\qquad + M^{-1}\mathrm{H}f(p(x_0))(p(x_0 + h) - p(x_0)) + o(\|h\|)B \\ &= \quad M^{-1}\nabla f(p(x_0)) + p(x_0) \\ &\qquad + [\mathcal{I} + M^{-1}\mathrm{H}f(p(x_0))](p(x_0 + h) - p(x_0)) \\ &\qquad + o(\|h\|)B\,. \end{aligned}$$

Now recall Theorem 2.8: the left-hand side contains x_0+h; likewise, $M^{-1}\nabla f(p(x_0))+ p(x_0)$ is the singleton x_0. Thus, we have proved

$$h \in [\mathcal{I} + M^{-1}\mathrm{H}f(p(x_0))](p(x_0 + h) - p(x_0)) + o(\|h\|)B\,.$$

Knowing that $\mathcal{I}+M^{-1}\mathrm{H}f(p(x_0))$ is invertible, this can also be written

$$p(x_0+h)-p(x_0)\in[\mathcal{I}+M^{-1}\mathrm{H}f(p(x_0))]^{-1}h+o(\|h\|)B\,,$$

which means that $[\mathcal{I}+M^{-1}\mathrm{H}f(p(x_0))]^{-1}$ is exactly the gradient of p at x_0. Therefore, $\nabla^2F(x_0)=M(\mathcal{I}-\nabla p(x_0))=M-M[\mathcal{I}+M^{-1}\mathrm{H}f(p(x_0))]^{-1}$. ◻

COROLLARY 3.2. *Let the finite-valued convex function f have a generalized Hessian at $p(x_0)$. Then $\nabla^2F(x_0)$ exists and*

$$\ker\nabla^2F(x_0)=\ker\mathrm{H}f(p(x_0)).$$

Proof. Use the notation $H:=\mathrm{H}f(p(x_0))$ and $H':=\nabla^2F(x_0)$. From Theorem 3.1, $M^{-1}H'=\mathcal{I}-[H+M]^{-1}M=\mathcal{I}-[M^{-1}H+\mathcal{I}]^{-1}$; hence,

$$\mathcal{I}-M^{-1}H'=(\mathcal{I}+M^{-1}H)^{-1}\,.$$

If $H'v=0$, then $(\mathcal{I}+M^{-1}H)v=v$ and $Hv=0$. Taking inverses, $(\mathcal{I}-M^{-1}H')^{-1}=\mathcal{I}+M^{-1}H$ and we show, likewise, that $Hv=0$ implies $H'v=0$. ◻

The converse part of Theorem 3.1 is not so simple; it will be stated in Theorems 3.14 and 3.15 below. First, the next geometrical result is crucial.

PROPOSITION 3.3. *Let f be a finite-valued strongly convex function satisfying* (8) *at a given $p_0=p(x_0)$. If $\nabla^2F(x_0)$ exists, then G of* (4) *lies in the relative interior of $\partial f(p_0)$.*

Proof. Let F have a Hessian at x_0; hence, by Proposition 2.4, $\nabla p(x_0)$ exists.

Assume for contradiction $G\in\operatorname{rbd}\partial f(p_0)$; by Proposition 2.2 in [11], the normal cone $\mathcal{N}=\mathrm{N}_{\partial f(p_0)}(G)$ is not a subspace. Now use Proposition 2.1 in [11] and the notation therein. We can take a unitary $\nu_0\in\mathcal{N}\cap\mathcal{M}^\perp$ (with $\mathcal{M}:=\mathcal{N}\cap-\mathcal{N}$) such that

$$\nu\in\mathcal{N}\text{ and }\langle\nu_0,\nu\rangle\neq0\quad\Longrightarrow\quad-\nu\notin\mathcal{N}\,. \tag{9}$$

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

$$\|p_t-p_0\|\leq\frac{1}{c}\|G_t-G\|=\frac{1}{c}t\,.$$

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

$$\|x_t-x_0\|\leq\|p_t-p_0\|+\frac{1}{\lambda}\|G_t-G\|\leq\left(\frac{1}{c}+\frac{1}{\lambda}\right)t. \tag{10}$$

Furthermore, since (8) holds, we can apply Corollary 3.3 in [11], with $\varphi=f$, $r=\frac{1}{2}C\|\cdot\|^2$, $z_0=p_0$, $x=p_t$, $g_0=G$, and $s=t\nu_0\in\mathcal{N}$: whenever $t\in[0,\varepsilon C/2]$,

$$\langle G_t-G,p_t-p_0\rangle\geq\frac{1}{2C}\|G_t-G\|^2=\frac{1}{2C}t^2\,.$$

Combine this equation with (10):

$$\frac{\lambda c}{2C(c+\lambda)}\leq\left\langle\nu_0,\frac{p_t-p_0}{\|x_t-x_0\|}\right\rangle\,.$$

Let $t \downarrow 0$. By (10), $x_t \to x_0$; extracting a subsequence if necessary, $[p_t - p_0]/\|x_t - x_0\| \to \nu \in \mathcal{N}$ (Corollary 2.6). Clearly, $\langle \nu_0, \nu \rangle > 0$; hence, by (9), $-\nu \notin \mathcal{N}$. This shows that $\mathrm{Im}\nabla p(x_0)$ is not a symmetric set; $\nabla p(x_0)$ cannot be a linear operator, which is the required contradiction. ∎

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

PROPOSITION 3.4. *Let f be a finite-valued strongly convex function satisfying* (8) *at a given $p_0 = p(x_0)$. If $\nabla^2 F(x_0)$ and $\nabla f(p_0)$ exist, then $\mathrm{H}f(p_0)$ exists.*

Proof. We have from (4) $p_0 = x_0 - M^{-1}G$ with $G = \nabla f(p_0)$. Apply Corollary 3.3 in [11] with $\varphi = f$, $r = \frac{1}{2}C\|\cdot\|^2$, $z_0 = p_0$, and $g_0 = G$; since $\partial f(p_0) = \{G\}$, $G + s$ is projected onto G for all s:

$$f^*(G+s) \geq f^*(G) + \langle s, p_0 \rangle + \frac{1}{2C}\|s\|^2 \quad \text{for } \|s\| \leq \varepsilon C/2\,. \tag{11}$$

By Corollary X.4.2.9 in [7], the existence of $\nabla^2 F(x_0)$ (positive definite, recall Theorem 2.2) implies the existence of $\nabla^2 F^*(G) = \nabla^2 f^*(G) + M^{-1}$. Therefore, $\nabla^2 f^*(G)$ exists and, by (11), is positive definite. Again, by Corollary X.4.2.9 in [7], f has a generalized Hessian at p_0. ∎

PROPOSITION 3.5. *Let f be a finite-valued strongly convex function satisfying* (8) *at $p_0 = p(x_0)$ for all $x_0 \in \mathbb{R}^N$. If $\nabla^2 F$ exists on the whole of $\mathbb{R}^N$, then f has a classical Hessian $\nabla^2 f$ on the whole of $\mathbb{R}^N$.*

Proof. We claim that f is differentiable at every $p \in \mathbb{R}^N$. Indeed, if $\partial f(p_0)$ is not a singleton, take a subgradient G in the relative boundary of $\partial f(p_0)$ (Proposition 2.3 in [11]). Because of Theorem 2.8, p_0 is the proximal point of $x_0 := p_0 + M^{-1}G$ and, by assumption, $\nabla^2 F(x_0)$ exists. From Proposition 3.3 we get the desired contradiction: G lies in the relative interior of $\partial f(p_0)$.

Then $\nabla^2 F$ and ∇f exist on the whole of $\mathbb{R}^N$ and Proposition 3.4 applies. ∎

3.2. Nondifferentiable case: The partial proximal operator. In this section, we consider a fixed x_0 such that f has no gradient at $p(x_0)$. In such a situation, can we relate the existence of $\nabla^2 F(x_0)$ with some smoothness of f at $p(x_0)$? To get an idea of what can happen, perturb the bivariate example $f(\xi, \eta) = |\xi| + \frac{1}{2}\eta^2$ at the end of section 2. Add to f an arbitrary convex univariate function $n(\xi)$, as nasty as desired, the extreme cases being $n \equiv 0$ and $n = \mathrm{I}_{\{0\}}$. It is easy to see that F remains the same near 0. In other words, F is totally blind to the behavior of f in the tangent cone $\mathcal{T}$.

We already know that the existence of $\nabla^2 F(x_0)$ implies $G \in \mathrm{ri}\,\partial f(p(x_0))$ (Proposition 3.3). As a result, the normal and tangent cones to $\partial f(p(x_0))$ at G are two complementary subspaces (Proposition 2.2 in [11]). Things are easier to visualize if the following notation is used: we set $\mathcal{U} = \mathcal{N} = \mathrm{N}_{\partial f(p(x_0))}(G)$ and $\mathcal{V} = \mathcal{T} = \mathrm{T}_{\partial f(p(x_0))}(G)$. The reason is that f is smooth or "U-shaped" in $p(x_0) + \mathcal{U}$ and kinky in $p(x_0) + \mathcal{V}$. Accordingly, we use the matrix-like decomposition $H = \left(\begin{smallmatrix} H_{\mathcal{UU}} & H_{\mathcal{UV}} \\ H_{\mathcal{VU}} & H_{\mathcal{VV}} \end{smallmatrix}\right)$ for linear operators. We denote indifferently by $\mathrm{Proj}_{\mathcal{U}}(s)$ or $s_{\mathcal{U}}$ the projection of s onto $\mathcal{U}$, similarly for $\mathcal{V}$. Note, incidentally, that the important subspace is really $\mathcal{U}$, which is completely defined by f alone; by contrast, $\mathcal{V}$ depends on the scalar product. Furthermore, these two subspaces do not depend on the particular $G \in \mathrm{ri}\,\partial f(p(x_0))$: $\mathcal{V}$ is the subspace parallel to $\mathrm{aff}\,\partial f(p(x_0))$.

LEMMA 3.6. *Let f be a finite-valued strongly convex function with modulus c, satisfying* (8) *at a given p_0. If $\nabla^2 f^*(G)$ exists, then it has the form $\left(\begin{smallmatrix} H^* & 0 \\ 0 & 0 \end{smallmatrix}\right)$, with*

$H^* : \mathcal{U} \to \mathcal{U}$ positive definite:

$$f^*(G+s) = f^*(G) + \langle s, p_0\rangle + \frac{1}{2}\Big\langle s, \begin{pmatrix} H^* & 0 \\ 0 & 0 \end{pmatrix} s\Big\rangle + o(\|s\|^2)\,. \tag{12}$$

Proof. The existence of $\nabla f^*(G)$ follows from Theorem 2.2. Then write the development (6) for f^* and apply Corollary 3.7 in [11] with $\varphi = f$, $r = \frac{1}{2}c\|\cdot\|^2$, $z_0 = p_0$, and $g_0 = G$. We obtain $\langle s, \nabla^2 f^*(G)s\rangle \le \frac{1}{2c}\|s_\mathcal{U}\|^2$, which implies $\operatorname{Im}\nabla^2 f^*(G) \subset \mathcal{U}$. The symmetric operator $\nabla^2 f^*(G)$ therefore has the stated form.

This establishes (12); let us now prove that H^* is positive definite. Because (8) holds, Corollary 3.4 in [11] applies with $\varphi = f$, $r = \frac{1}{2}C\|\cdot\|^2$, $z_0 = p_0$, and $g_0 = G$:

$$f^*(G+s) \ge f^*(G) + \langle s, p_0\rangle + \frac{1}{2C}\|s_\mathcal{U}\|^2$$

for all $s \in B(0, \varepsilon C/2)$. In particular, if $s \in \mathcal{U} \cap B(0, \varepsilon C/2)$, then $s_\mathcal{U} = s$ and we obtain, with (12),

$$f^*(G) + \langle s, p_0\rangle + \frac{1}{2}\Big\langle s, \begin{pmatrix} H^* & 0 \\ 0 & 0 \end{pmatrix} s\Big\rangle + o(\|s\|^2) \ge f^*(G) + \langle s, p_0\rangle + \frac{1}{2C}\|s\|^2\,.$$

This clearly implies $\langle s, (\begin{smallmatrix} H^* & 0 \\ 0 & 0 \end{smallmatrix})s\rangle \ge \frac{1}{C}\|s\|^2$ for all $s \in \mathcal{U}$. $\square$

While (12) describes the structure of $\nabla^2 f^*$, it also kills the proof technique used in Proposition 3.4: $\nabla^2 f^*$ is no longer invertible. At this stage we use the cure of section 4 in [11]: we consider the perturbation

$$\phi_\mathcal{V}(x) = \min_{y \in x+\mathcal{V}}\{f(y) - \langle G, y\rangle + \tfrac{1}{2}\|x-y\|^2\}\,, \quad \phi_\mathcal{V}^*(s) := f^*(G+s) + \frac{1}{2}\|s_\mathcal{V}\|^2\,. \tag{13}$$

LEMMA 3.7. *Let f be a finite-valued convex function. Take the Moreau–Yosida regularization of $\phi_\mathcal{V}$, defined in* (13)*: $\Phi_\mathcal{V}(x) := \min_{y\in\mathbb{R}^N}\{\phi_\mathcal{V}(y) + \frac{1}{2}\|y-x\|_M^2\}$, and denote by $\pi(x)$ the associated proximal point. Then the following holds:*

(i) *$\pi(p_0) = p_0$.*

(ii) *$\phi_\mathcal{V}$ is strongly convex if and only f is strongly convex.*

(iii) *If f is strongly convex and satisfies* (8) *at a given p_0, then $\mathrm{H}\phi_\mathcal{V}(p_0)$ exists if and only if $\nabla^2\Phi_\mathcal{V}(p_0)$ exists. In this case, $\mathrm{H}\phi_\mathcal{V}(p_0) = (\begin{smallmatrix} H^{*-1} & 0 \\ 0 & \mathcal{I}_\mathcal{V} \end{smallmatrix})$.*

Proof. (i): Use Theorem 2.8 at $p = p_0$, with f replaced by $\phi_\mathcal{V}$: $p_0 = \pi(p_0 + M^{-1}\gamma)$ for any $\gamma \in \partial\phi_\mathcal{V}(p_0)$. But Proposition 4.1 of [11] used with $z_0 = p_0$ gives $\partial\phi_\mathcal{V}(p_0) = \nabla\phi_V(p_0) = 0$. Thus, $\pi(p_0) = p_0$.

(ii): Theorem 2.2 yields the following chain of equivalences: f strongly convex $\iff$ ∇f^* Lipschitzian $\iff$ $\nabla(f^*(G+\cdot) + 1/2\|\operatorname{Proj}_\mathcal{V}(\cdot)\|^2)$ Lipschitzian $\iff$ $(f^*(G+\cdot) + 1/2\|\operatorname{Proj}_\mathcal{V}(\cdot)\|^2)^*$ strongly convex. The result follows from (13).

(iii): Because f is strongly convex, f^* is finite valued. Furthermore, due to Corollary 3.3 in [11], the assumptions of Proposition 4.2 in [11] hold and we have, for h small enough, $\phi_\mathcal{V}(p_0+h) \le \phi_\mathcal{V}(p_0) + 1/2C'\|h\|^2$. This is the growth condition (8) for $\phi_\mathcal{V}$ (recall $\nabla\phi_\mathcal{V}(p_0) = 0$). Write Theorem 3.1 and Proposition 3.4, with f, F, x_0, p_0 replaced by $\phi_\mathcal{V}, \Phi_\mathcal{V}, p_0, p_0$, to obtain the stated equivalence.

Finally, when $\mathrm{H}\phi_\mathcal{V}(p_0)$ exists, it is positive definite and its inverse is $\nabla^2\phi_\mathcal{V}^*(0)$ (Corollary X.4.2.9 of [7]). Because of (13), $\nabla^2\phi_\mathcal{V}^*(0) = \nabla^2 f^*(G) + \operatorname{Proj}_\mathcal{V}$, and because of Lemma 3.6, the diagonal form follows. $\square$

This enables us to state the key relation for the present nondifferentiable case.

PROPOSITION 3.8. *Let f be a finite-valued strongly convex function satisfying* (8) *at a given $p_0 = p(x_0)$. Then $\nabla^2 F(x_0)$ exists if and only if $\mathrm{H}\phi_{\mathcal{V}}(p_0)$ exists.*

Proof. In view of Lemma 3.7(iii), we have to prove the equivalence "$\nabla^2 F(x_0)$ exists $\iff$ $\nabla^2 \Phi_{\mathcal{V}}(p_0)$ exists." From Theorem 2.2, F is strongly convex; hence, by using Corollary X.4.2.9 in [7],

$$\exists \nabla^2 F(x_0) \iff \exists \nabla^2 F^*(G) \iff \exists \nabla^2 f^*(G) = \nabla^2 F^*(G) - M^{-1} .$$

Because of (13), this is further equivalent to (recall $\nabla \phi_{\mathcal{V}}(p_0) = \nabla \Phi_{\mathcal{V}}(p_0) = 0$)

$$\exists \nabla^2 \phi^*_{\mathcal{V}}(0) = \nabla^2 f^*(G) + \mathrm{Proj}_{\mathcal{V}} \iff \exists \nabla^2 \Phi^*_{\mathcal{V}}(0) = \nabla^2 \phi^*_{\mathcal{V}}(0) + M^{-1} .$$

Finally, $\nabla^2 \Phi^*_{\mathcal{V}}(0)$ is positive definite; by Corollary X.4.2.9 in [7], this last existence is equivalent to the existence of $\nabla^2 \Phi_{\mathcal{V}}(p_0)$. ∎

We devote the end of the section to interpretations of the above result in terms of the original function f. They crucially rely on the partial proximal operator associated to (13):

$$p_{\mathcal{V}}(x) := \operatorname*{argmin}_{y \in x + \mathcal{V}} \{ f(y) - \langle G, y \rangle + \tfrac{1}{2} \|x - y\|^2 \}.$$

Remember from Proposition 4.1 of [11] the useful characterization

$$\exists g \in \partial f(p_{\mathcal{V}}(x)) \quad \text{such that} \quad p_{\mathcal{V}}(x) - x = \mathrm{Proj}_{\mathcal{V}}(G - g), \tag{14}$$

as well as

$$\partial \phi_{\mathcal{V}}(x) = -G + \{ g \in \partial f(p_{\mathcal{V}}(x)) \, : \, \mathrm{Proj}_{\mathcal{V}}(G - g) = p_{\mathcal{V}}(x) - x \} . \tag{15}$$

First of all, the definition (7) of $\mathrm{H}\phi_{\mathcal{V}}(p_0)$ can be translated as follows.

COROLLARY 3.9. *Let f be a finite-valued strongly convex function satisfying* (8) *at a given $p_0 = p(x_0)$. Existence of $\nabla^2 F(x_0)$ is equivalent to the following property. Let $x \to p_0$ and let $g \in \partial f(p_{\mathcal{V}}(x))$ be such that $\mathrm{Proj}_{\mathcal{V}}(g - G) = x - p_{\mathcal{V}}(x)$. Then $g = G + \mathrm{H}\phi_{\mathcal{V}}(p_0)(x - p_0) + o(\|x - p_0\|)$.*

Proof. At each x, apply (15): with g as stated, $g - G$ describes $\partial \phi_{\mathcal{V}}(x)$. The result follows from Proposition 3.8, remembering that $\nabla \phi_{\mathcal{V}}(p_0) = 0$. ∎

This result concerns approximations of particular subgradients of f near p_0. Function values can also be approximated along the surface described by $p_{\mathcal{V}}(\cdot)$. In what follows, we study second-order developments of f with respect to the variable $p_{\mathcal{V}}(x)$ rather than x itself.

THEOREM 3.10. *Let f be a finite-valued strongly convex function satisfying* (8) *at a given p_0 such that $H := \mathrm{H}\phi_{\mathcal{V}}(p_0)$ exists. Taking $x = p_0 + h_{\mathcal{U}} + h_{\mathcal{V}}$ with $h_{\mathcal{U}} \to 0$ and $h_{\mathcal{V}} = O(\|h_{\mathcal{U}}\|)$, set $d(x) := p_{\mathcal{V}}(x) - p_0$. Then we have*

$$\begin{gathered} f(p_0 + d(x)) = \\ f(p_0) + \langle G, d(x) \rangle + \tfrac{1}{2} \langle d(x), H d(x) \rangle - \langle p_{\mathcal{V}}(x) - x, d(x) \rangle + o(\|d(x)\|^2) . \end{gathered} \tag{16}$$

Proof. By definition (13) and remembering that $\phi_{\mathcal{V}}(p_0) = f(p_0) - \langle G, p_0 \rangle$ (Proposition 4.1 of [11]), the second-order development of $\phi_{\mathcal{V}}$ gives

$$f(p_{\mathcal{V}}(x)) - \langle G, p_{\mathcal{V}}(x) \rangle + \frac{1}{2} \|p_{\mathcal{V}}(x) - x\|^2 = f(p_0) - \langle G, p_0 \rangle + \frac{1}{2} \langle h, Hh \rangle + o(\|h\|^2). \tag{17}$$

Now, H has the special form given in Lemma 3.7(iii), so the property $x - p_{\mathcal{V}}(x) \in \mathcal{V}$ implies $H(x - p_{\mathcal{V}}(x)) = x - p_{\mathcal{V}}(x)$; hence, by writing $h = (x - p_{\mathcal{V}}(x)) + (p_{\mathcal{V}}(x) - p_0) = (x - p_{\mathcal{V}}(x)) + d(x)$, we obtain $\langle h, Hh\rangle = \langle d(x), Hd(x)\rangle - 2\langle p_{\mathcal{V}}(x) - x, d(x)\rangle + \|p_{\mathcal{V}}(x) - x\|^2$. Plugging this equality into (17), we get

$$f(p_0 + d(x)) = f(p_0) + \langle G, d(x)\rangle + \frac{1}{2}\langle d(x), Hd(x)\rangle - \langle p_{\mathcal{V}}(x) - x, d(x)\rangle + o(\|h\|^2)\,.$$

Finally, $\|h\|^2 = \|h_{\mathcal{U}}\|^2 + \|h_{\mathcal{V}}\|^2 = O(\|h_{\mathcal{U}}\|^2)$, but $h_{\mathcal{U}}$ is just the component on $\mathcal{U}$ of $d(x)$; altogether, $\|h\|^2 = O(\|d(x)\|^2)$. ∎

Beware that (16) is not a regular development of f near p_0. First, it is valid only for special increments $d(\cdot)$ and, at this point, we have not even proved that they tend to 0. Second, what happens to the "extra" term $p_{\mathcal{V}}(x) - x \in \mathcal{V}$? To clarify this situation, we need to bound the difference $p_{\mathcal{V}}(x) - x$.

THEOREM 3.11. *Let f be a finite-valued strongly convex function, satisfying* (8) *at a given p_0 and such that $H := \mathrm{H}\phi_{\mathcal{V}}(p_0)$ exists. For $d = d_{\mathcal{U}} + d_{\mathcal{V}}$ tending to 0 in such a way that*

(i) $\|d_{\mathcal{V}}\| = o(\|d_{\mathcal{U}}\|)$,
(ii) $\exists g \in \partial f(p_0 + d)$ *such that* $\mathrm{Proj}_{\mathcal{V}}(g - G) = O(\|d\|)$,

we have the second-order development

$$f(p_0 + d) = f(p_0) + \langle G, d\rangle + \frac{1}{2}\langle d, Hd\rangle + o(\|d\|^2)\,. \tag{18}$$

Proof. With g as in (ii), define $x := p_0 + d + \mathrm{Proj}_{\mathcal{V}}(g - G)$; then, (14) shows that $p_{\mathcal{V}}(x) = p_0 + d$ and $h := x - p_0$ satisfies $h_{\mathcal{U}} = d_{\mathcal{U}}$ and $h_{\mathcal{V}} = d_{\mathcal{V}} + \mathrm{Proj}_{\mathcal{V}}(g - G) = O(\|h_{\mathcal{U}}\|)$. Write (16) and observe that $\langle p_{\mathcal{V}}(x) - x, d\rangle = \langle \mathrm{Proj}_{\mathcal{V}}(g - G), d_{\mathcal{V}}\rangle$. The assumptions (i), (ii) clearly imply that this is $o(\|d\|^2)$. ∎

This result describes points at which f behaves like a quadratic function. We now show a way of constructing such points.

COROLLARY 3.12. *Let f be a finite-valued strongly convex function satisfying* (8) *at a given p_0 such that $H := \mathrm{H}\phi_{\mathcal{V}}(p_0)$ exists. Let $h \to 0$ with $\|h_{\mathcal{V}}\| = O(\|h_{\mathcal{U}}\|)$ and set $x := p_0 + h$. Then $p_{\mathcal{V}}(x)$ from* (14) *tends to p_0 and* (18) *holds for $d := p_{\mathcal{V}}(x) - p_0 \to 0$.*

Proof. Proceeding as in the proof of Lemma 3.7(iii), we get the assumptions of Corollary 4.3 in [11] (with $z_0 = p_0$): $p_{\mathcal{V}}(\cdot)$ is radially Lipschitzian at p_0, hence $d := p_{\mathcal{V}}(x) - p_0 \to 0$. On the other hand, for some constant $\theta > 0$, $\|h\| \le \theta\|h_{\mathcal{U}}\| = \theta\|d_{\mathcal{U}}\| \le \theta\|d\|$. Because of (14), there is some $g \in \partial f(p_{\mathcal{V}}(x))$ such that $\|\mathrm{Proj}_{\mathcal{V}}(g - G)\| = \|p_{\mathcal{V}}(x) - x\| = O(\|h\|) = O(\|d\|)$. We are in the framework of Theorem 3.11: $d \to 0$, assumption (ii) holds; let us prove that assumption (i) is also satisfied. Any limit point of g lies in $\partial f(p_0)$ (graph closedness of ∂f). Since $\mathcal{V} = \mathrm{aff}\,\partial f(p_0) - G$, the property $\|\mathrm{Proj}_{\mathcal{V}}(g - G)\| = O(\|d\|)$ actually implies $\|g - G\| = O(\|d\|)$. Using Lemma 2.5 with $g_0 = G$, $z_0 = p_0$, $z = p_{\mathcal{V}}(x)$, and $t = \|p_{\mathcal{V}}(x) - p_0\|$, we deduce that any limit point of $\frac{p_{\mathcal{V}}(x) - p_0}{\|p_{\mathcal{V}}(x) - p_0\|}$ lies in $\mathcal{U}$: assumption (i) of Theorem 3.11 holds. ∎

Let us summarize our results: appropriate assumptions (strong convexity, growth condition, existence of $\nabla^2 F(x_0)$) provide the following second-order information:

(i) $\nabla^2 F(x_0)$ is positive definite (Theorem 2.2); $\nabla^2 f^*(G) = \nabla^{-2} F(x_0) - M^{-1}$ exists and is completely characterized by its $\mathcal{UU}$-block (Lemma 3.6).

(ii) $\nabla p(x_0)$ exists and, in view of Corollary 2.6, has the form $\nabla p(x_0) = \left(\begin{smallmatrix} P & 0 \\ T & 0 \end{smallmatrix}\right)$. From Proposition 2.4 we have $\nabla^{-2} F(x_0) = \nabla^2 F^*(G) = [\mathcal{I} - \nabla p(x_0)]^{-1} M^{-1}$.

(iii) A "partial" generalized Hessian of f at p_0, as described by (18), exists in $\mathcal{U}$. It is the $\mathcal{UU}$-block of the (diagonal) operator $[\nabla^{-2} F(x_0) - M^{-1} + \mathrm{Proj}_{\mathcal{V}}]^{-1}$ and we denote it by $\mathrm{H}_{\mathcal{U}} f(p_0)$.

This last block turns out to have the simple expression

$$\mathrm{H}_{\mathcal{U}} f(p_0) = M_{\mathcal{U}\mathcal{U}}(P^{-1} - \mathcal{I}_{\mathcal{U}})\,. \tag{19}$$

To see this, we need two results from linear algebra, stated without proof.

- For $M = \left(\begin{smallmatrix} M_{\mathcal{U}\mathcal{U}} & M_{\mathcal{U}\mathcal{V}} \\ M_{\mathcal{U}\mathcal{V}}^T & M_{\mathcal{V}\mathcal{V}} \end{smallmatrix}\right)$ and $M^{-1} = \left(\begin{smallmatrix} W_{\mathcal{U}\mathcal{U}} & W_{\mathcal{U}\mathcal{V}} \\ W_{\mathcal{U}\mathcal{V}}^T & W_{\mathcal{V}\mathcal{V}} \end{smallmatrix}\right)$, it holds that
$$M_{\mathcal{U}\mathcal{U}}^{-1} = W_{\mathcal{U}\mathcal{U}} - W_{\mathcal{U}\mathcal{V}} W_{\mathcal{V}\mathcal{V}}^{-1} W_{\mathcal{U}\mathcal{V}}^T\,.$$
- Let P be such that $\mathcal{I} - P$ and $(\mathcal{I} - P)^{-1} - \mathcal{I}$ are both invertible. Then P is invertible and

$$[(\mathcal{I} - P)^{-1} - \mathcal{I}]^{-1} = P^{-1} - \mathcal{I}\,. \tag{20}$$

Let us now compute H^* of (12): $\left(\begin{smallmatrix} H^* & 0 \\ 0 & 0 \end{smallmatrix}\right)$ is equal to

$$\nabla^2 f^*(G) = \nabla^2 F^*(G) - M^{-1} = \nabla^{-2} F(x_0) - M^{-1} = [(\mathcal{I} - \nabla p(x_0))^{-1} - \mathcal{I}]M^{-1}.$$

A straightforward computation gives

$$(\mathcal{I} - \nabla p(x_0))^{-1} = \begin{pmatrix} (\mathcal{I}_{\mathcal{U}} - P)^{-1} & (\mathcal{I}_{\mathcal{U}} - P)^{-1} T \\ 0 & \mathcal{I}_{\mathcal{V}} \end{pmatrix};$$

the $\mathcal{U}\mathcal{V}$-block of $\nabla^2 f^*(G)$ is therefore $[(\mathcal{I}_{\mathcal{U}} - P)^{-1} - \mathcal{I}_{\mathcal{U}}]W_{\mathcal{U}\mathcal{V}} + (\mathcal{I}_{\mathcal{U}} - P)^{-1} T W_{\mathcal{V}\mathcal{V}} = 0$. This serves to compute T and we obtain the $\mathcal{U}\mathcal{U}$-block

$$H^* = [(\mathcal{I}_{\mathcal{U}} - P)^{-1} - \mathcal{I}_{\mathcal{U}}](W_{\mathcal{U}\mathcal{U}} - W_{\mathcal{U}\mathcal{V}} W_{\mathcal{V}\mathcal{V}}^{-1} W_{\mathcal{U}\mathcal{V}}^T) = [(\mathcal{I}_{\mathcal{U}} - P)^{-1} - \mathcal{I}_{\mathcal{U}}]M_{\mathcal{U}\mathcal{U}}^{-1}\,.$$

This is precisely the inverse of $\mathrm{H}_{\mathcal{U}} f(p_0)$; then, (19) follows using (20).

3.3. Getting rid of strong convexity. Our final goal will be to eliminate the strong convexity assumption in the preceding second-order results. For this we perturb f to a strongly convex function f_τ, and we study the effect of this perturbation on the proximal point.

PROPOSITION 3.13. *Let f be a finite-valued convex function satisfying* (8) *at a given p_0. Take $z_0 \in \mathbb{R}^N$, $\tau \in]0,1[$ and define $f_\tau := f + \frac{1}{2}\tau\|\cdot - z_0\|_M^2$. Consider the Moreau–Yosida regularization of f_τ associated to the metric defined by $(1-\tau)M$:*

$$F_\tau(x) := \min_{y \in \mathbb{R}^n} \left\{ f_\tau(y) + \tfrac{1}{2}(1-\tau)\|y - x\|_M^2 \right\}. \tag{21}$$

Denote by $q_\tau(x)$ the unique minimizer of (21); *then, the following statements hold:*

(i) *the function f_τ is strongly convex and satisfies* (8) *at p_0 with C replaced by $C + \tau\Lambda$;*

(ii) *for all x, $q_\tau(x) = p(\tau z_0 + (1-\tau)x)$.*

Proof. The strong convexity of f_τ is clear. To prove that (8) holds for f_τ, add $\frac{1}{2}\tau\|p_0 + h - z_0\|_M^2$ to both sides of (8) written for f at p_0. Then use the properties $f'_\tau(p_0;h) = f'(p_0;h) + \frac{1}{2}\tau\,\langle M(p_0 - z_0), h\rangle$ and $\frac{1}{2}\tau\|\cdot\|_M^2 \le \frac{1}{2}\tau\Lambda\|\cdot\|^2$.

For proving (ii), write the optimality conditions for $p(\tau z_0 + 1 - \tau x)$ and $q_\tau(x)$:

$$\begin{array}{lll} p(\tau z_0 + (1-\tau)x) \text{ solves} & M(\tau z_0 + (1-\tau)x - p) \in \partial f(p) & \text{and} \\ q_\tau(x) \text{ solves} & (1-\tau)M(x - p) \in \partial f_\tau(p)\,. & \end{array}$$

Since $\partial f_\tau(p) = \partial f(p) + \{\tau M(p - z_0)\}$, they have the same solutions. $\square$

Thus, passing from f to f_τ can be absorbed by a perturbation of M in the Moreau–Yosida regularization of (2) and a (smooth) change of variables. Then the wording "strongly convex" can be removed in our second-order results. Note that our proof technique below will use f_τ and F_τ with arbitrary $\tau \in]0,1[$.

THEOREM 3.14. *Let f be a finite-valued convex function such that for a given $x_0 \in \mathbb{R}^n$,* (8) *holds at $p_0 = p(x_0)$. Assume $\nabla f(p_0)$ exists. Then $\nabla^2 F(x_0)$ exists if and only if* $\mathrm{H}f(p_0)$ *exists.*

Proof. The "only if" part is Theorem 3.1. As for the "if" part, suppose $\nabla^2 F(x_0)$ exists; hence, from Proposition 2.4, $\nabla p(x_0)$ exists. Then, consider f_τ as in Proposition 3.13, with $z_0 = x_0$: we have $q_\tau(x) = p(\tau x_0 + (1-\tau)x)$ for all x, therefore $\nabla q_\tau(x_0) = (1-\tau)\nabla p(x_0)$ exists. Again, using Proposition 2.4, F_τ has a Hessian at x_0. Since f_τ is strongly convex and satisfies (8) at p_0, Proposition 3.4 applies: $\mathrm{H}f_\tau(p_0)$ exists. Thus $\mathrm{H}f(p_0) = \mathrm{H}f_\tau(p_0) - \tau M$ exists as well. ∎

THEOREM 3.15. *Let f be a finite-valued convex function such that for all $x_0 \in \mathbb{R}^N$,* (8) *holds at $p_0 = p(x_0)$. 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. Recall that the existence of $\mathrm{H}f$ on the whole space implies the existence of $\nabla^2 f$ on the whole space. Then the "only if" part is Theorem 3.1. 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 3.13 with $z_0 = 0$; we have $\nabla q_\tau(x) = (1-\tau)\nabla p((1-\tau)x)$ for all x. Proceeding as in the proof of Theorem 3.14 but applying, this time, Proposition 3.5, we conclude that $\nabla^2 f_\tau$ (and hence $\nabla^2 f$) exists everywhere. ∎

For the nondifferentiable case, we again will use f_τ as in Proposition 3.13 with $z_0 = x_0$ and F_τ of (21). Then, because $G_\tau := \nabla F_\tau(x_0) = (1-\tau)G$, it follows that

$$\partial f_\tau(\cdot) - G_\tau = \partial f(\cdot) - G + \tau M(\cdot - p_0)\,. \tag{22}$$

We will also consider $\phi_{\mathcal{V},\tau}$, obtained by replacing f and G in (13) by f_τ and G_τ. The associated partial proximal operator $q_{\mathcal{V},\tau}$ is characterized by

$$\begin{gathered}\exists g \in \partial f(q_{\mathcal{V},\tau}(x)) \quad \text{such that} \\ q_{\mathcal{V},\tau}(x) - x = \mathrm{Proj}_{\mathcal{V}}(G-g) + \tau\,\mathrm{Proj}_{\mathcal{V}}(M(q_{\mathcal{V},\tau}(x) - p_0))\,.\end{gathered} \tag{23}$$

THEOREM 3.16. *Let f be a finite-valued convex function satisfying* (8) *at a given $p_0 = p(x_0)$. Assume $\nabla^2 F(x_0)$ exists. For $d = d_{\mathcal{U}} + d_{\mathcal{V}}$ tending to 0 in such a way that*

(i) $\|d_{\mathcal{V}}\| = o(\|d_{\mathcal{U}}\|)$,

(ii) $\exists g \in \partial f(p_0 + d)$ *such that* $\mathrm{Proj}_{\mathcal{V}}(g - G) = O(\|d\|)$,

we have the second-order development

$$f(p_0 + d) = f(p_0) + \langle G, d\rangle + \frac{1}{2}\langle d, H' d\rangle + o(\|d\|^2)\,, \tag{24}$$

where $H' = \begin{pmatrix} M_{\mathcal{U}\mathcal{U}}(P^{-1} - \mathcal{I}_{\mathcal{U}}) & -\tau M_{\mathcal{U}\mathcal{V}} \\ -\tau M_{\mathcal{U}\mathcal{V}}^T & \mathcal{I}_{\mathcal{V}} - \tau M_{\mathcal{V}\mathcal{V}} \end{pmatrix}$.

Proof. Consider again f_τ as in Proposition 3.13 with $z_0 = x_0$; F_τ of (21) has at x_0 a Hessian $\nabla^2 F_\tau(x_0) = (1-\tau)^2 \nabla^2 F(x_0) + \tau(1-\tau)M$. By Proposition 3.8, this is equivalent to the existence of $\mathrm{H}\phi_{\mathcal{V},\tau}(p_0)$. Take d satisfying (i) and apply (ii) together with (22): there is $g_\tau \in \partial f_\tau(p_0 + d)$ such that $\mathrm{Proj}_{\mathcal{V}}(g_\tau - G_\tau) = O(\|d\|) + \tau M d = O(\|d\|)$. Then Theorem 3.11 holds for the perturbed functions: mutatis mutandis, we write

$$f(p_0+d) + \frac{1}{2}\tau\|p_0+d-x_0\|_M^2 = f(p_0) + \frac{1}{2}\tau\|p_0-x_0\|_M^2 + \langle G_\tau, d\rangle + \frac{1}{2}\langle d, H_\tau d\rangle + o(\|d\|^2),$$

where H_τ is the corresponding generalized Hessian. By rearranging terms we obtain

$$f(p_0+d)=f(p_0)+\langle G,d\rangle+\frac{1}{2}\langle d,(H_\tau-\tau M)d\rangle+o(\|d\|^2)\,.$$

To finish the proof we give the expression of $H' := H_\tau - \tau M$. Indeed, by Lemma 3.7(iii), H_τ has the diagonal form $\left(\begin{smallmatrix} H_\tau^{*-1} & 0 \\ 0 & \mathcal{I}_\mathcal{V}\end{smallmatrix}\right)$. This, together with (19) and Proposition 3.13(ii), gives

$$H_\tau=\begin{pmatrix} \mathrm{H}_\mathcal{U} f_\tau(p_0) & 0 \\ 0 & \mathcal{I}_\mathcal{V}\end{pmatrix}=\begin{pmatrix} (1-\tau)M_{\mathcal{U}\mathcal{U}}(\frac{1}{1-\tau}P^{-1}-\mathcal{I}_\mathcal{U}) & 0 \\ 0 & \mathcal{I}_\mathcal{V}\end{pmatrix}.$$

Finally,

$$H' = H_\tau - \tau M = \begin{pmatrix} M_{\mathcal{U}\mathcal{U}}(P^{-1}-\mathcal{I}_\mathcal{U}) & -\tau M_{\mathcal{U}\mathcal{V}} \\ -\tau M_{\mathcal{U}\mathcal{V}}^T & \mathcal{I}_\mathcal{V}-\tau M_{\mathcal{V}\mathcal{V}}\end{pmatrix} \qquad \square$$

Of course, the $\mathcal{U}\mathcal{U}$-block of H' does not depend on τ: it *has to* be $\mathrm{H}_\mathcal{U} f(p_0)$ of (19).

COROLLARY 3.17. *Let f be a finite-valued convex function satisfying* (8) *at a given $p_0 = p(x_0)$. Assume $\nabla^2 F(x_0)$ exists. Let $h \to 0$ with $\|h_\mathcal{V}\| = O(\|h_\mathcal{U}\|)$ and set $x := p_0 + h$. For arbitrary $\tau \in]0,1[$, $q_{\mathcal{V},\tau}(x)$ from* (23) *tends to p_0 and* (24) *holds for $d := q_{\mathcal{V},\tau}(x) - p_0 \to 0$.*

Proof. Proceed as before; $q_{\mathcal{V},\tau}(\cdot)$ of (23) enjoys the same properties as $p_\mathcal{V}(\cdot)$. Also, the existence of $\nabla^2 F(x_0)$ implies the existence of $\nabla^2 F_\tau(x_0)$, which in turn is equivalent to the existence of $\mathrm{H}\phi_{\mathcal{V},\tau}(p_0)$. Thus Corollary 3.12 applies. $\square$

3.4. The epi-convergence approach. So far, our study has been dealing with rather classical derivatives: the Hessian for F and its (slight) generalization (6) for f. Another object appears as fairly handy when used in conjunction with Moreau–Yosida regularizations: the so-called epi-derivative. We proceed to explain in a few informal words what it is, referring to [26], [4], [24] for more detailed explanations.

- Let $\{E_t\}$ be a family of sets indexed by t. Form the set E of all possible clusterpoints of all possible sequences of elements $e_t \in E_t$ when $t \downarrow 0$. Under certain conditions which we do not specify here, we say that E is the *limit* of E_t and we write $E_t \to E$.
- Recall that the *epigraph* of a (convex) function φ is the set $\operatorname{epi}\varphi := \{(x,r) \in \mathbb{R}^n \times \mathbb{R} : r \geq \varphi(x)\} \subset \mathbb{R}^n \times \mathbb{R}$. The *graph* of its subdifferential is the set $\operatorname{gr}\partial\varphi := \{(z,g) : g \in \partial\varphi(z)\} \subset \mathbb{R}^n \times \mathbb{R}^n$.
- Several classical meanings can be given to a statement like "the function φ_t converges to the function φ" (pointwise convergence, uniform convergence,...). Here, we use the following concept: φ_t *epi-converges* to φ when $\operatorname{epi}\varphi_t \to \operatorname{epi}\varphi$ in the sense of the above set-convergence. We will then use the notation $\varphi = \operatorname{epi}\lim \varphi_t$.
- For closed convex functions, a fundamental property of epi-convergence is its stability under conjugacy and differentiation. More precisely, the statements $\varphi = \operatorname{epi}\lim\varphi_t$ and $\varphi^* = \operatorname{epi}\lim\varphi_t^*$ are equivalent. They are further equivalent to the statement $\operatorname{gr}\partial\varphi = \lim \operatorname{gr}\partial\varphi_t$, provided that pointwise convergence holds at least at one point (to fix the constant of integration).
- A (classical) second-order derivative can be viewed as a quadratic form. In the present theory, it is convenient to accept the value $+\infty$ for such an object. Accordingly, given a positive semidefinite operator H and a subspace S, we call *generalized quadratic form* characterized by H and S the closed convex function

$q(x) := 1/2\,\langle x, Hx\rangle + \mathrm{I}_S$. Note that $\partial q(x) = Hx + S^\perp$ for $x \in S$. This class of functions is invariant under conjugacy: the conjugate of a generalized quadratic form is another generalized quadratic form, characterized by K and T, say.

• Let φ be a closed convex function. For given z_0 and $g_0 \in \partial\varphi(z_0)$, we form the second-order difference quotient

$$\Delta_t(d) := \frac{\varphi(z_0 + td) - \varphi(z_0) - t\,\langle g_0, d\rangle}{t^2}\,.$$

This is a closed convex function of d, indexed by $t \downarrow 0$. Direct calculations give its subdifferential

$$\partial\Delta_t(d) = \frac{\partial\varphi(z_0 + td) - g_0}{t}$$

and its conjugate

$$\Delta_t^*(s) = \frac{\varphi^*(g_0 + ts) - \varphi^*(g_0) - t\,\langle s, x_0\rangle}{t^2}.$$

Observe that $\Delta_t(0) = \Delta_t^*(0) = 0$ and $\partial\Delta_t(0) \ni 0$.

• Then we say that φ has a second *epi-derivative* q at z_0, relative to g_0, when the function Δ_t epi-converges to q (a generalized quadratic form). Equivalently, φ^* has at g_0, relative to z_0, a second epi-derivative q^*. A further equivalence, probably the most useful, is $\operatorname{gr}\partial\Delta_t \to \operatorname{gr}\partial q$. In plain words, the clusterpoints of the difference quotients $[\partial\varphi(z_0 + td_t) - g_0]/t$, when $t \downarrow 0$ and $d_t \to d$, form some affine manifold $Hd + S$.

Now the stage is set and we can use these concepts in our Moreau–Yosida framework. Here again, we will not go into details, referring to [17] and also [12] for a more accurate analysis. Our aim here is to somehow "explain" our second-order results by heuristic observations rather than rigorous statements.

(a): We start with the following observation. Let the growth condition (8) hold at some p_0, and let f have a gradient at p_0. Then ∂f has at p_0 the radially Lipschitz behavior (see Corollary 3.5 in [11]); the two concepts of second epi-derivative and of generalized Hessian (7) coincide. Likewise, since ∇F is Lipschitzian, the two concepts of second epi-derivative and of classical Hessian coincide for F.

(b): Now suppose that f has at p_0 a second epi-derivative (H, S), relative to some subgradient $G \in \partial f(p_0)$. Equivalently, f^* has at G a second epi-derivative (K, T) relative to p_0. Clearly enough, $F^* = f^* + 1/2\,\|\cdot\|^2_{M-1}$ also has a second epi-derivative $(K + M^{-1}, T)$ at G, which is relative to $p_0 + M^{-1}G \in \partial F^*(G)$. Dualizing again, F has (at $p_0 + M^{-1}G =: x_0$ and relative to $G = \nabla F(x_0)$) a second epi-derivative (H', S'). Naturally, since $K + M^{-1}$ is positive definite, this last epi-derivative is an ordinary quadratic function: $S' = \mathbb{R}^n$. Indeed, as already mentioned, the difference quotients $[\nabla F(x_0 + td) - G]/t$ (which are bounded) converge uniformly to a linear function $H'd$. This explains and actually completes Theorem 3.1.

(c): Conversely, suppose that F has a second epi-derivative at some x_0 (relative to $G = M(x_0 - p(x_0)) = \nabla F(x_0)$; of course, it is actually a Hessian). Then F^* has a second epi-derivative (K', T') at G, relative to x_0. Here again, $f^* = F^* - 1/2\,\|\cdot\|^2_{M-1}$ has a second epi-derivative (K, T) at G, relative to $x_0 - M^{-1}G = p(x_0) \in \partial f^*(G)$. Finally, because f^* is convex, f itself has a second epi-derivative (H, S) at $p(x_0)$, relative to $G \in \partial f(p(x_0))$.

We conclude that as far as epi-derivatives are concerned, second differentiability of f and of F are always equivalent properties.

(c′): When F has a Hessian at x_0, suppose that the growth condition (8) holds and that $\nabla f(p(x_0))$ exists. As seen in (a) above, the second epi-derivative of f at $p(x_0)$ (which exists) is actually a generalized Hessian; this explains Proposition 3.4.

(d): Suppose f has a second epi-derivative at p_0, relative to $G \in \partial f(p_0)$ (for example, $\nabla^2 F(p_0 + M^{-1}G)$ exists), and consider curves $(g_t - G)/t$ with $g_t \in \partial f(p_0 + td_t)$ and $d_t \to d$. Their clusterpoints form the set $Hd + S^\perp$, where H is a symmetric positive semidefinite operator and S is a subspace.

• Such clusterpoints can exist only for $d \in \mathrm{N}_{\partial f(p_0)}(G)$ (otherwise, $g_t - G$ does not even tend to 0).

• To obtain these clusterpoints, one can in particular take $d_t \equiv 0$ and g_t arbitrary in $\partial f(p_0)$; this generates $\mathrm{T}_{\partial f(p_0)}(G)$, which is therefore contained in $S^\perp$.

• On the other hand, let the growth condition (8) hold. Then it takes some work (based on Corollary 3.3 in [11]) to realize that $S^\perp$ exactly reduces to $\mathrm{T}_{\partial f(p_0)}(G)$. As a result, $S^\perp = \mathrm{T}_{\partial f(p_0)}(G) = \mathcal{V}$. This explains Proposition 3.3; it also explains the extra term $p_{\mathcal{V}}(x) - x \in \mathcal{V} = S^\perp$ in (16).

• Finally, since these clusterpoints cover the whole of $Hd + S = Hd + \mathcal{U}$, some of them are exactly Hd; among the latter, we have those described by Theorem 3.11 and Corollary 3.12.

Let us summarize this section. Epi-derivatives are an elegant and powerful tool to relate second-order behaviors of f and F. They yield the essence of our results in sections 3.1–3.3 at practically no cost. This, however, is paid by the high degree of abstraction imposed by the concept. By contrast, our approach requires the heavy material developed in [11], but we deal with natural objects such as Taylor developments and ordinary (point-) convergence. We believe that the two approaches are in fact complementary and beneficial to each other: epi-derivatives give insightful guesses of the kind of result to be expected; standard convex analysis gives a more intuitive meaning to these "epi-results" and makes a closer description of $f(p_0 + h)$ for actual values of h. This last point becomes particularly useful when coming to numerical algorithms.

4. 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 H having extreme eigenvalues c and C. From Theorem 3.1, F is also quadratic with Hessian $M + M(H + M)^{-1}M$. Taking $M = \lambda\mathcal{I}$, a quick calculation shows that the condition number C/c of H is divided by $(\lambda + C)/(\lambda + c)$. This is a clear beneficial effect of the Moreau–Yosida regularization.

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 any good to minimize F (i.e., f). More specifically, what can be said about a superlinear algorithm minimizing F as compared to the ordinary proximal algorithm minimizing f?

When f is differentiable on the whole space, Theorem 3.15 and Corollary 3.2 tell us that such an approach brings exactly nothing. Either F still does not enjoy the necessary properties of smoothness and nondegeneracy or a superlinear algorithm could have been applied to f at the first place (ordinary Newton, quasi-Newton, or nonsmooth Newton as in [20]). 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 nonlinear program: 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.

When f is differentiable just at an optimum point $\bar{x}$, the situation is less clear. On the one hand, existence of a positive definite $\nabla^2 F(\bar{x})$ gives hope for an efficient nonsmooth Newton algorithm; the pending question is semismoothness of ∇F, a question which is investigated in [19]. On the other hand, the existence alone of a (positive definite) generalized Hessian $\mathrm{H}f(\bar{x})$ is probably not quite enough to obtain superlinearly convergent algorithms applied directly to f. Here we mention a technical question. As far as quasi-Newton methods are concerned, an important property is the *strict* differentiability of ∇F at an optimum point (see [3], [8]). It would be interesting to examine the consequences of such a property on the behavior of f; following [24], some useful insight might be provided by the epi-derivative approach.

The real issue is when f is not differentiable at $\bar{x}$, a situation which does not preclude the existence of a positive definite $\nabla^2 F(\bar{x})$. In this case, any kind of Newtonian method will minimize F rapidly but will not even be locally convergent when applied to f. Existence of $\nabla^2 F$ at an optimum point $\bar{x}$ implies some interesting properties for f. First of all, $0 \in \mathrm{ri}\, \partial f(\bar{x})$, a property which can be compared to the strict complementarity slackness in constrained optimization. Furthermore, f enjoys a partial second-order behavior, via the existence of $\mathrm{H}_{\mathcal{U}} f(\bar{x})$; see (19). In our analogy with constrained optimization, $\mathcal{U}$ is the subspace tangent to the active constraints. The $\mathcal{UV}$-decomposition appears as an important tool from the theoretical point of view; this observation assesses the algorithmic approach of [14].

Take for illustration the bivariate function $f(x) := \max\{\frac{1}{2}\|x\|^2 - \alpha \langle e, x\rangle , \langle e, x\rangle\}$; here, $e := (0, 1)^T$ and α is a nonnegative parameter.

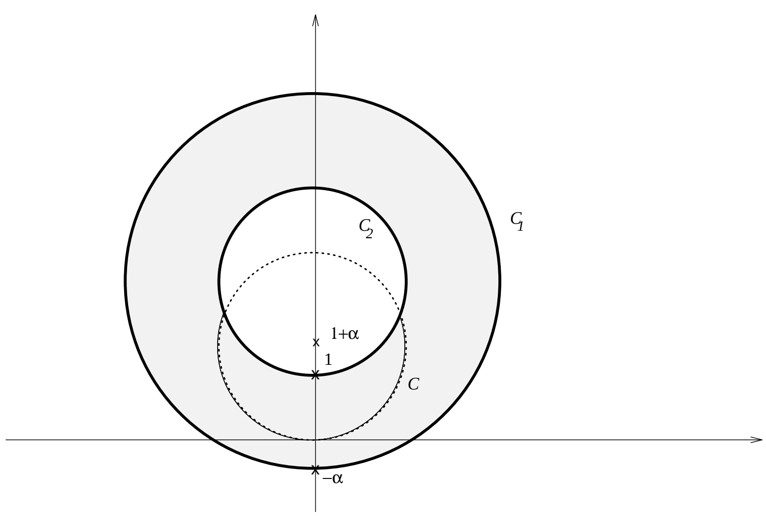

FIG. 1. *Moreau–Yosida regularization without Hessian.*

The kinks of f form a circle, denoted by C in Fig. 1. 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 $M = \mathcal{I}$, 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

$$\mu = \mu(x) := \frac{\|x - (\alpha+2)e\|}{\alpha+1} - 1\,. \tag{25}$$

In a condensed form,

$$p(x) = \frac{x - e + (\alpha+1)\nu(x)e}{1 + \nu(x)}\,, \tag{26}$$

where $\nu(x)$ is the projection of $\mu(x)$ in (25) onto $[0,1]$.

In Fig. 1, C_1 (respectively, C_2) is the boundary of the region where the first (respectively, 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 $\alpha = 0$, the origin is on C_1. Analytically, $\nu(x) = 1$ in (26) whenever $\|x - 2e\| > 2$. From this observation, the directional derivatives are easy to compute:

$$p'(0; d) = \begin{cases} \frac{1}{2}(d_1, 0)^T & \text{for } d_2 > 0, \\ \frac{1}{2}(d_1, d_2)^T & \text{for } d_2 \leq 0\,. \end{cases}$$

Here, the nonexistence of $\nabla p(0)$ illustrates Proposition 3.3: $G = 0$ is on the relative boundary of $\partial f(0) = [0,1]$.

(ii): When $\alpha > 0$, we have $\mu \in [0,1]$ in (25) 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

$$\nabla p(0) = \frac{\alpha+1}{\alpha+2}\left(\mathcal{I} - ee^T\right) = \frac{\alpha+1}{\alpha+2}\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}.$$

Let us turn now to the $\mathcal{UV}$-analysis developed in section 3.2. Here $\mathcal{V} = 0 \times \mathbb{R}$ is the vertical axis. The partial (generalized) Hessian (19) is $\mathrm{H}_{\mathcal{U}} f(0) = (\frac{\alpha+2}{\alpha+1} - 1)\mathcal{I}_{\mathcal{U}} = \frac{1}{\alpha+1}\mathcal{I}_{\mathcal{U}}$.

We claim that for small $\|x\|$, the partial proximal operator is

$$p_{\mathcal{V}}(x) = (x_1, \alpha + 1 - \sqrt{D})^T ,$$

where we have set $D := (\alpha+1)^2 - x_1^2$. Note: $\sqrt{D} = \alpha+1-\frac{x_1^2}{2(\alpha+1)}+o(x_1^2)$ and $[p_{\mathcal{V}}(x)]_2 = \frac{x_1^2}{2(\alpha+1)} + o(x_1^2)$. Geometrically, $p_{\mathcal{V}}(x)$ is obtained by intersecting $x + \mathcal{V}$ with C. To prove the formula analytically, plug $\partial f(p_{\mathcal{V}}(x)) = \{\lambda p_{\mathcal{V}}(x) + (1-(\alpha+1)\lambda)e : \lambda \in [0,1]\}$ and $G = 0$ in the characterization (14) to obtain

$$[p_{\mathcal{V}}(x) =]p(\lambda) := \left(x_1, \frac{x_2 - 1 + \lambda(\alpha+1)}{1+\lambda}\right)^T \tag{27}$$

for some $\lambda \in [0,1]$. With this change of variables, $p_{\mathcal{V}}(x)$ can be rewritten as

$$p_{\mathcal{V}}(x) = \underset{\lambda}{\operatorname{argmin}}\{f(p(\lambda)) + \tfrac{1}{2}\|x - p(\lambda)\|^2\} .$$

It takes some calculations to see that this minimum is attained at

$$\lambda := \frac{\alpha + 2 - \sqrt{D} - x_2}{\sqrt{D}} = \frac{1 - x_2}{\alpha+1} + o(x_1^2) \to \frac{1}{\alpha+1} \in]0,1[.$$

With this value of λ in (27), the claim follows.

This confirms that $p_{\mathcal{V}}(x)$ is a kink, as explained at the end of [11]. Then the function $\phi_{\mathcal{V}}$ of (13) has the expression

$$\phi_{\mathcal{V}}(x) = [p_{\mathcal{V}}(x)]_2 + \frac{1}{2}\|[p_{\mathcal{V}}(x)]_2 - x_2\|^2 = \frac{x_1^2}{2(\alpha+1)} + \frac{1}{2}x_2^2 + o(\|x\|^2) .$$

It can be differentiated directly, or (15) can be used: $\partial\phi_{\mathcal{V}}(x)$ is made up of those $g \in \partial f(p_{\mathcal{V}}(x))$ whose second coordinate is the same as $x - p_{\mathcal{V}}(x)$. We obtain just one vector: $\lambda p_{\mathcal{V}}(x) + (1-(\alpha+1)\lambda)e$, where λ takes the above value. Thus,

$$\nabla\phi_{\mathcal{V}}(x) = \begin{pmatrix} \frac{\alpha+2-\sqrt{D}x_2}{\sqrt{D}}x_1 \\ x_2 - [p_{\mathcal{V}}(x)]_2 \end{pmatrix} = \begin{pmatrix} \frac{x_1}{\alpha+1} \\ x_2 \end{pmatrix} + o(\|x\|) .$$

From there, (generalized) Hessians follow easily. Alternatively, the above expression of $\nabla p(0)$ can be plugged into the calculations made at the end of section 3.2: $P = \frac{\alpha+1}{\alpha+2}\mathcal{I}_{\mathcal{U}}$, $T = 0$, and we obtain

$$\mathrm{H}^* = (\mathcal{I}_{\mathcal{U}} - P)^{-1} - \mathcal{I}_{\mathcal{U}} = (\alpha+1)\mathcal{I}_{\mathcal{U}}, \quad \nabla^2\phi_{\mathcal{V}}(0) = \begin{pmatrix} \frac{1}{\alpha+1} & 0 \\ 0 & 1 \end{pmatrix}$$

and

$$\mathrm{H}_{\mathcal{U}}f(0) = P^{-1} - \mathcal{I}_{\mathcal{U}} = \left(\frac{\alpha+2}{\alpha+1} - 1\right)\mathcal{I}_{\mathcal{U}} = \frac{1}{\alpha+1}\mathcal{I}_{\mathcal{U}} .$$

We have a final comment. In this paper we focused our attention on the Fréchet point of view; as far as algorithms are concerned, this is well suited to the quasi-Newton pattern. For the Newton pattern (possibly approximate, see [18]), the directional point of view may be more relevant; see [20]. Likewise, the Moreau–Yosida regularization could be generalized to the resolvent of a maximal monotone operator in the framework of variational inequalities; for this, see [16].

We are indebted to J.-B. Hiriart-Urruty, L.Q. Qi, and R.T. Rockafellar, as well as three anonymous referees, for careful reading and many helpful suggestions.

REFERENCES

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

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

[3] J. Bonnans, J. Gilbert, C. Lemaréchal, and C. Sagastizábal, *A family of variable metric proximal methods*, Math. Programming, 68 (1995), pp. 15–48.

[4] S. Dolecki, G. Salinetti, and R. Wets, *Convergence of functions: Equi-semi-continuity*, Trans. Amer. Math. Soc., 276 (1983), pp. 409–429.

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

[6] J.-B. Hiriart-Urruty, *The approximate first-order and second-order directional derivatives for a convex function*, in Mathematical Theories of Optimization, J.-P. Cecconi and T. Zolezzi, eds., Lecture Notes in Mathematics 979, Springer-Verlag, Berlin, New York, 1983, pp. 154–166.

[7] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms*, Springer-Verlag, Berlin, New York, 1993.

[8] C.-M. Ip and J. Kyparisis, *Local convergence of quasi-Newton methods for B-differentiable equations*, Math. Programming, 56 (1992), pp. 71–89.

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

[10] C. Lemaréchal and C. Sagastizábal, *An approach to variable metric bundle methods*, in Systems Modelling and Optimization, J. Henry and J.-P. Yvon, eds., Lecture Notes in Control and Inform. Sci. 197, Springer-Verlag, Berlin, New York, 1994, pp. 144–162.

[11] C. Lemaréchal and C. Sagastizábal, *More than first-order developments of convex functions: Primal-dual relations*, J. Convex Anal., (1996), pp. 1–14.

[12] Y. Lucet, *Formule explicite du hessien de la régularisée de Moreau-Yosida d'une fonction convexe f en fonction de l'épi-différentielle seconde de f*, Report # LAO95-08, University of Toulouse, France, 1995.

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

[14] R. Mifflin, *A Quasi-Second-Order Proximal Bundle Algorithm*, Technical report 93-3, University of Washington, Pullman, WA, 1993.

[15] J. Moreau, *Proximité et dualité dans un espace hilbertien*, Bull. Soc. Math. France, 93 (1965), pp. 273–299.

[16] J. Pang and L. Qi, *Nonsmooth equations: Motivation and algorithms*, SIAM J. Control Optim., 3 (1993), pp. 443–465.

[17] R. Poliquin and R. Rockafellar, *Generalized hessian properties of regularized nonsmooth functions*, SIAM J. Optim., 6 (1996), pp. 1121–1137.

[18] L. Qi, *Superlinearly convergent approximate Newton methods for LC^1 optimization problems*, Math. Programming, 64 (1994), pp. 277–294.

[19] L. Qi, *Second-Order Analysis of the Moreau-Yosida Approximation of a Convex Function*, Applied Mathematics Report AMR94/20, School of Mathematics, The University of New South Wales, Sydney, Australia, 1994.

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

[21] M. Qian, *The Variable Metric Proximal Point Algorithm: Application to Optimization*, Department of Mathematics, Manuscript GN-50, University of Washington, Seattle, WA 98195, 1992.

[22] R. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[23] R. Rockafellar, *Monotone operators and the proximal point algorithm*, SIAM J. Control Optim., 14 (1976), pp. 877–898.

[24] R. Rockafellar, *Maximal monotone relations and the second derivatives of nonsmooth functions*, Annales de l'Institut Henri Poincaré, Analyse non linéaire, 2 (1985), pp. 167–186.

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

[26] R. Wets, *Convergence of convex functions, variational inequalities and convex optimization problems*, in Variational Inequalities and Complementarity Problems, F. G. R. Cottle and J.-L. Lions, eds., Wiley, New York, 1980, pp. 405–419.

[27] K. Yosida, *Functional Analysis*, Springer-Verlag, Berlin, New York, 1964.

[28] C. Lemaréchal and C. Sagastizábal, *Variable metric bundle methods: From conceptual to implementable forms*, Math. Programming, to appear.