THE $\mathcal{U}$-LAGRANGIAN OF A CONVEX FUNCTION

CLAUDIUS LÉMARÉCHAL, FRANÇOIS OUSTRY, AND CLAUDIA SAGASTIZÁBAL

ABSTRACT. At a given point $\mathbf{p}$, a convex function $f$ is differentiable in a certain subspace $\mathcal{U}$ (the subspace along which $\partial f(\mathbf{p})$ has 0-breadth). This property opens the way to defining a suitably restricted second derivative of $f$ at $\mathbf{p}$. We do this via an intermediate function, convex on $\mathcal{U}$. We call this function the $\mathcal{U}$-Lagrangian; it coincides with the ordinary Lagrangian in composite cases: exact penalty, semidefinite programming. Also, we use this new theory to design a conceptual pattern for superlinearly convergent minimization algorithms. Finally, we establish a connection with the Moreau-Yosida regularization.

1. Introduction

This paper deals with higher-order expansions of a nonsmooth function, a problem addressed in [4], [5], [7], [9], [13], [25], and [31] among others.

The initial motivation for our present work lies in the following facts. When trying to generalize the classical second-order Taylor expansion of a function $f$ at a nondifferentiability point $\mathbf{p}$, the major difficulty is by far the nonlinearity of the first-order approximation. Said otherwise, the gradient vector $\nabla f(\mathbf{p})$ is now a set $\partial f(\mathbf{p})$ and we have to consider difference quotients between sets, say

$$\frac{\partial f(\mathbf{p} + h) - \partial f(\mathbf{p})}{\|h\|}. \tag{1.1}$$

Giving a sensible meaning to the minus-sign in this expression is a difficult problem, to say the least; it has received only abstract answers so far; see [1], [3], [10], [12], [16], [18], [23], [24], [30]. However, here are two crucial observations (already mentioned in [22]):

- There is a subspace $\mathcal{U}$ (the “ridge”) in which the first-order approximation $f'(\mathbf{p}; \cdot)$ (the directional derivative) is linear.

- Defining a second-order expansion of $f$ is unnecessary along directions not in $\mathcal{U}$. Consider for example the case where $f = \max_i f_i$ with smooth $f_i$'s; then a minimization algorithm of the SQP-type will converge superlinearly, even if the second-order behaviour of $f$ is identified in the ridge only ([26], [6]).

Here, starting from results presented in [14] and [15], we take advantage of these observations. After some preliminary theory in §2, we define our key-objects in §3: the $\mathcal{U}$-Lagrangian and its derivatives. In §4 we give some specific examples (further studied in [17], [20]): how the $\mathcal{U}$-Lagrangian specializes in an NLP and an SDP.

Received by the editors July 18, 1996 and, in revised form, August 1, 1997.

1991 Mathematics Subject Classification. Primary 49J52, 58C20; Secondary 49Q12, 65K10.

Key words and phrases. Nonsmooth analysis, generalized derivative, second-order derivative, composite optimization.
framework, and how it could help designing superlinearly convergent algorithms for general convex functions. Finally, we show in §5 a connection between our objects thus defined and the Moreau-Yosida regularization. Indeed, the present paper clarifies and formalizes the theory sketched in §3.2 of [15]; for a related subject see also [29], [25].

Our notation follows closely that of [28] and [11]. The space $\mathbb{R}^n$ is equipped with a scalar product $(\cdot, \cdot)$, and $\|\cdot\|$ is the associated norm; in a subspace $S$, we will write $(\cdot, \cdot)_S$ and $\|\cdot\|_S$ for the induced scalar product and norm. The open ball of $\mathbb{R}^n$ centered at $x$ with radius $r$ is $B(x, r)$; and once again, we use the notation $B_S(x, r)$ in a subspace $S$. We denote by $x_S$ the projection of a vector $x \in \mathbb{R}^n$ onto the subspace $S$. Throughout this paper, we consider the following situation:

$$(1.2) \quad f \text{ is a finite-valued convex function, } \overline{p} \text{ and } \overline{g} \in \partial f(\overline{p}) \text{ are fixed.}$$

We will also often assume that $\overline{g}$ lies in the relative interior of $\partial f(\overline{p})$.

2. The $\mathcal{U}\mathcal{V}$ decomposition

We start by defining a decomposition of the space $\mathbb{R}^n = \mathcal{U} \oplus \mathcal{V}$, associated with a given $\overline{p} \in \mathbb{R}^n$. We give three equivalent definitions for the subspaces $\mathcal{U}$ and $\mathcal{V}$; each has its own merit to help the intuition.

**Definition 2.1.**

(i) Define $\mathcal{U}_1$ as the subspace where $f'(\overline{p}; \cdot)$ is linear and take $\mathcal{V}_1 := \mathcal{U}_1^\perp$. Because $f'(\overline{p}; \cdot)$ is sublinear, we have

$$\mathcal{U}_1 := \{ d \in \mathbb{R}^n : f'(\overline{p}; d) = -f'(\overline{p}; -d) \};$$

if necessary, see for instance Proposition V.1.1.6 in [11]. In other words, $\mathcal{U}_1$ is the subspace where $f(\overline{p} + \cdot)$ appears to be “differentiable” at 0. Note that this definition of $\mathcal{U}_1$ does not rely on a particular scalar product.

(ii) Define $\mathcal{V}_2$ as the subspace parallel to the affine hull of $\partial f(\overline{p})$ and take $\mathcal{U}_2 := \mathcal{V}_2^\perp$. In other words, $\mathcal{V}_2 := \text{lin}(\partial f(\overline{p}) - \overline{g})$ for an arbitrary $\overline{g} \in \partial f(\overline{p})$, and $d \in \mathcal{U}_2$ means $(\overline{g} + v, d) = (\overline{g}, d)$ for all $v \in \mathcal{V}_2$.

(iii) Define $\mathcal{U}_3$ and $\mathcal{V}_3$ respectively as the normal and tangent cones to $\partial f(\overline{p})$ at an arbitrary $g^\circ$ in the relative interior of $\partial f(\overline{p})$. It is known (see, for example, Proposition 2.2 in [14]) that the property $g^\circ \in \text{ri } \partial f(\overline{p})$ is equivalent to these cones being subspaces.

To visualize these definitions, the reader may look at Figure 1 in §3.2 (where $\overline{g} = g^\circ \in \text{ri } \partial f(\overline{p})$). We recall the definition of the relative interior: $g^\circ \in \text{ri } \partial f(\overline{p})$ means

$$(2.1) \quad g^\circ + (B(0, \eta) \cap \mathcal{V}_2) \subset \partial f(\overline{p}) \quad \text{for some } \eta > 0.$$  

We start with a preliminary result, showing in particular that Definition 2.1 does define the same pair $\mathcal{U}\mathcal{V}$ three times.

**Proposition 2.2.** In Definition 2.1,

(i) the subspace $\mathcal{U}_3$ is actually given by

$$(2.2) \quad \{ d \in \mathbb{R}^n : (g - g^\circ, d) = 0 \text{ for all } g \in \partial f(\overline{p}) \} = N_{\partial f(\overline{p})}(g^\circ)$$

and is independent of the particular $g^\circ \in \text{ri } \partial f(\overline{p})$;

(ii) $\mathcal{U}_1 = \mathcal{U}_2 = \mathcal{U}_3 =: \mathcal{U}$;

(iii) $\mathcal{U} \subset N_{\partial f(\overline{p})}(\overline{g})$ for all $\overline{g} \in \partial f(\overline{p})$.  

Proof. (i) To prove (2.2), take $g^o \in ri \partial\tilde{f}(\overline{p})$ and set $N := N_{\partial\tilde{f}(\overline{p})}(g^o)$. By definition of a normal cone, $N$ contains the left-hand side in (2.2); we only need to establish the converse inclusion. Let $d \in N$ and $g \in \partial\tilde{f}(\overline{p})$; it suffices to prove $\langle g - g^o, d \rangle \geq 0$. Indeed, (assuming $g - g^o \neq 0$), $v := -\frac{g - g^o}{\|g - g^o\|} \in \mathcal{V}_2$, hence (2.1) and $d \in N$ imply that

$$0 \geq \langle g^o + \eta v - g^o, d \rangle = -\frac{\eta}{\|g - g^o\|}(g - g^o, d) \quad \text{for some } \eta > 0$$

and we are done.

To see the independence on the particular $g^o$, replace $g^o$ in (2.2) by some other $\gamma^o \in ri \partial\tilde{f}(\overline{p})$:

$$N_{\partial\tilde{f}(\overline{p})}(\gamma^o) = \{d \in \mathbb{R}^n : \langle g, d \rangle = \langle \gamma^o, d \rangle = \langle g^o, d \rangle, \text{ for all } g \in \partial\tilde{f}(\overline{p})\} = \mathcal{U}_3.$$

(ii) Write

$$(2.3) \quad \mathcal{U}_1 = \{d \in \mathbb{R}^n : \max_{g \in \partial\tilde{f}(\overline{p})} \langle g, d \rangle = \min_{g \in \partial\tilde{f}(\overline{p})} \langle g, d \rangle\}$$

to see from (i) that $\mathcal{U}_1 = \mathcal{U}_3$. Then we only need to prove $\mathcal{U}_1 \subset \mathcal{U}_2 \subset \mathcal{U}_3$.

Let $d \in \mathcal{U}_1$. For an arbitrary $v = \sum_j \lambda_j (g_j - \overline{g}) \in \mathcal{V}_2$ with $g_j \in \partial\tilde{f}(\overline{p})$, we have from (2.3)

$$\langle v, d \rangle = \sum_j \lambda_j \langle (g_j, d) - (\overline{g}, d) \rangle = 0;$$

hence $d \in \mathcal{V}_2 = \mathcal{U}_2$.

Let $d \in \mathcal{U}_2$. We have $\langle g, d \rangle = \langle \overline{g}, d \rangle$ for all $g \in \partial\tilde{f}(\overline{p})$. It follows that $\langle g, d \rangle = \langle g^o, d \rangle$ and this, together with (i), implies $d \in \mathcal{U}_3$.

(iii) Let $d \in \mathcal{U} = \mathcal{U}_3$. Given $\overline{g} \in \partial\tilde{f}(\overline{p})$, we have $\langle g^o, d \rangle = \langle g, d \rangle = \langle \overline{g}, d \rangle$ for all $g \in \partial\tilde{f}(\overline{p})$; hence $d \in N_{\partial\tilde{f}(\overline{p})}(\overline{g})$. $\square$

Using projections, every $x \in \mathbb{R}^n$ can be decomposed as $x = (x_u, x_v)^T$. Throughout this paper we use the notation $x_u \oplus x_v$ for the vector with components $x_u$ and $x_v$. In other words, $\oplus$ stands for the linear mapping from $\mathcal{U} \times \mathcal{V}$ onto $\mathbb{R}^n$ defined by

$$(2.4) \quad \mathcal{U} \times \mathcal{V} \ni (u, v) \mapsto u \oplus v := \begin{pmatrix} u \\ v \end{pmatrix} \in \mathbb{R}^n.$$

With this convention, $\mathcal{U}$ and $\mathcal{V}$ are themselves considered as vector spaces. We equip them with the scalar product induced by $\mathbb{R}^n$, so that

$$\langle g, x \rangle = \langle g_u \oplus g_v, x_u \oplus x_v \rangle = \langle g_u, x_u \rangle_u + \langle g_v, x_v \rangle_v,$$

with similar expressions for norms.

Remark 2.3. The projection $x \mapsto x_u$, as well as the operation $(u, v) \mapsto \overline{p} + u \oplus v$, will appear recurrently in all our development. Consider the three convex functions $h_1, h_2$ and $h$ defined by

$$\mathcal{U} \ni u \mapsto h_1(u) := f(\overline{p} + u \oplus v), \quad \text{with } v \in \mathcal{V} \text{ arbitrary};$$

$$\mathcal{V} \ni v \mapsto h_2(v) := f(\overline{p} + u \oplus v), \quad \text{with } u \in \mathcal{U} \text{ arbitrary};$$

$$\mathcal{U} \times \mathcal{V} \ni (u, v) \mapsto h(u, v) := f(\overline{p} + u \oplus v).$$

Their subdifferentials have the expressions

$$\partial h_1(u) = \{g_u : g \in \partial \tilde{f}(\overline{p} + u \oplus v)\},$$

$$\partial h_2(v) = \{g_v : g \in \partial \tilde{f}(\overline{p} + u \oplus v)\},$$

$$\partial h(x_u, x_v) = \{g_u \oplus g_v : g \in \partial \tilde{f}(\overline{p} + x)\}.$$
Proving these formulae is a good exercise to become familiar with the operation \( \oplus \) of (2.4) and with our \( \mathcal{VU} \) notation. Just consider the adjoint of \( \oplus \) and of the projections onto the various subspaces involved.

In the \( \mathcal{VU} \) language, (2.1) gives the following elementary result.

**Proposition 2.4.** Suppose in (1.2) that \( \overline{g} \in \text{ri} \partial f(p) \). Then there exists \( \eta > 0 \) small enough such that

\[
\overline{g} + 0 \oplus \frac{\eta v}{\|v\|_V} \in \partial f(p)
\]

for any \( 0 \neq v \in V \). In particular,

(2.5) \[ f(p + u \oplus v) \geq f(p) + \langle \overline{g}_U, u \rangle_U + \langle \overline{g}_V, v \rangle_V + \eta\|v\|_V, \]

for any \( (u, v) \in \mathcal{U} \times V \).

**Proof.** Just translate (2.1): with \( v \) as stated, \( u \oplus v \overline{g}_U + \frac{\eta v}{\|v\|_V} \in \partial f(p) \) and the rest follows easily. 

---

### 3. The \( \mathcal{U} \)-Lagrangian

In this section we formalize the theory outlined in §3.2 of [15]. Along with the \( \mathcal{VU} \) decomposition, we introduced there the “tangential” regularization \( \phi_V \). Here, we find it convenient to consider \( \phi_V \) as a function defined on \( U \) only; in addition, we drop the quadratic term appearing in (13) of [15]. As will be seen in §4, these modifications result in some sort of Lagrangian, which we denote by \( L_U \) instead of \( \phi_V \).

#### 3.1. Definition and basic properties.

Following the above introduction, we define the function \( L_U \) as follows:

(3.1) \[
U \ni u \mapsto L_U(u) := \inf_{v \in V} \{ f(p + u \oplus v) - \langle \overline{g}_V, v \rangle_V \}.
\]

Associated with (3.1) we have the set of minimizers

(3.2) \[
W(u) := \text{Argmin}_{v \in V} \{ f(p + u \oplus v) - \langle \overline{g}_V, v \rangle_V \}.
\]

It will be seen below that an important question is whether \( W(u) \) is nonempty.

**Remark 3.1.** The function \( L_U \) of (3.1) will be called the \( \mathcal{U} \)-Lagrangian. Note that it depends on the particular \( \overline{g} \), a notation \( L_U(u, \overline{g}) \) is also possible. In fact, since \( \overline{g} \) lies in the dual of \( \mathbb{R}^n \), it connotes a dual variable; this will become even more visible in §4.1 (just observe here that \( \overline{g} \mapsto -L_U \) is a conjugate function).

At this point, the idea behind (3.1) can be roughly explained. As is commonly known, smoothness of a convex function is related to strong convexity of its conjugate. In our context, a useful property is the “radial” strong convexity of \( f^* \) at \( \overline{g} \), say,

\[
f^*(\overline{g} + s) \geq f^*(\overline{g}) + \langle s, \overline{p} \rangle + \frac{1}{2} c\|s\|^2 + o(\|s\|^2)
\]

for some \( c > 0 \). However, the above inequality is hopeless for an \( s \) of the form \( s = 0 \oplus v \) (see §4 in [14]; see also [2] for related developments). To obtain radial strong convexity on \( V \), we introduce the function

(3.3) \[
f^*(\overline{g} + s) + \frac{1}{2} c\|s_V\|^2_V.
\]
Its conjugate (restricted to $\mathcal{U}$) is precisely $L_\mathcal{U}$ when $c = +\infty$ (a value which yields the “strongest” possible convexity); Theorem 3.3 will confirm the smoothness of $L_\mathcal{U}$.

The value $c = 1$ in (3.3) may be deemed more natural – and indeed, it will be useful in §5; in fact, Lemma 5.1 will show that the choice of $c$ has minor importance for second order. \hfill $\square$

**Theorem 3.2.** Assume (1.2).

1. The function $L_\mathcal{U}$ defined in (3.1) is convex and finite everywhere.
2. A minimum point $w \in W(u)$ in (3.2) is characterized by the existence of some $g \in \partial f(\bar{p} + u \oplus w)$ such that $g_v = \bar{g}_v$.
3. In particular, $0 \in W(0)$ and $L_\mathcal{U}(0) = f(\bar{p})$.
4. If $\bar{p} \in \text{ri} \partial f(p)$, then $W(u)$ is nonempty for each $u \in \mathcal{U}$ and $W(0) = \{0\}$.

**Proof.** (i) The infimand in (3.1) is $h(u, v) - \langle \bar{g}_v, v \rangle \nu$, where the function $h$ was defined in Remark 2.3. It is clearly finite-valued and convex on $\mathcal{U} \times \mathcal{V}$, and the subgradient inequality at $(u, v) = (0, 0)$ gives

$$h(u, v) - \langle \bar{g}_v, v \rangle \nu \geq f(\bar{p}) + \langle \bar{g}_\mathcal{U}, u \rangle_{\mathcal{U}} \text{ for any } v \in \mathcal{V}.$$ 

It follows that $L_\mathcal{U}$ is nowhere $-\infty$ and, being a partial infimum of a jointly convex function, it is convex as well, see for example §IV.2.4 in [11].

(ii) The optimality condition for $w \in W(u)$ is $0 \in \partial h_2(w) - \bar{g}_v$, with $h_2$ as in Remark 2.3. Knowing the expression of $\partial h_2$, we obtain $0 = g_v - \bar{g}_v$, for some $g \in \partial f(\bar{p} + u \oplus w)$.

(iii) In particular, for $u = 0$, we can take $w = 0$ and $g = \bar{g} \in \partial f(\bar{p} + 0 \oplus 0)$ in (ii). This proves that $v = 0$ satisfies the optimality condition for (3.1); then $L_\mathcal{U}(0) = f(\bar{p})$.

(iv) Apply (2.5): there exists $\eta > 0$ such that, for any $v \neq 0$,

$$h(u, v) - \langle \bar{g}_v, v \rangle \nu \geq f(\bar{p}) + \langle \bar{g}_\mathcal{U}, u \rangle_{\mathcal{U}} + \eta\|v\|\nu.$$ 

Thus, the infimand in (3.1) is inf-compact on $\mathcal{V}$ and the set $W(u)$ is nonempty. At $u = 0$, we have

$$h(0, v) - \langle \bar{g}_v, v \rangle \nu \geq f(\bar{p}) + \eta\|v\|\nu,$$

which shows that $v = 0$ is the unique minimizer. \hfill $\square$
Corollary VI.4.5.3 in [11] gives the calculus rule

\[ s \in \partial_u L_U(u) \iff s \oplus 0 \in \partial_u (h - \langle 0 \oplus g_V, \cdot \rangle)(u, w) \iff s \oplus 0 \in \partial_u h(u, w) - 0g_V \iff s \oplus g_V \in \partial_{u,v} h(u, w), \]

where \( w \in W(u) \) is arbitrary. From the expression of \( \partial_{u,v} h = \partial h \) in Remark 2.3, this is (3.4).

(ii) Because of Theorem 3.2(iii), (3.4) holds at \( u = 0 \) and becomes \( \partial L_U(0) = \{ g_U : g_U \oplus g_V \in \partial f(p) \} \). This latter set clearly contains \( \overline{g}_U \). Actually, it does not contain any other point, due to Definition 2.1(ii): \( \partial f(p) \subset g + V \), i.e., all subgradients at \( p \) have the same \( U \)-component, namely \( g_U \).

This result is illustrated in Figure 1. We stress the fact that the set in the right-hand-side of (3.4) does not depend on the particular \( w \in W(u) \). In other words, (3.4) expresses the following: to obtain the subgradients of \( L_U \) at \( u \), take those subgradients \( g \) of \( f \) at \( p + u \oplus W(u) \) that have the same \( V \)-component as \( g \) (namely \( g_V \)); then take their \( U \)-component. Remembering that \( U \) is in effect a subset of \( \mathbb{R}^n \), we can also write more informally

\[ \partial L_U(u) = [\partial f(p + u \oplus W(u)) \cap (g + U)]_U. \]

This operation somewhat simplifies when \( g_V = 0 \):

(3.5) if \( g_V = 0 \), then \( \partial L_U(u) = \partial f(p + u \oplus W(u)) \cap U \).

See the end of §3.2 below for additional comments on the “trajectories” \( p + u \oplus W(u) \).

Another observation is that, for all \( u \in U \),

\[ f'(p; u \oplus 0) = \langle g, u \oplus 0 \rangle = \langle \overline{g}_U, u \rangle_U = \langle \nabla L_U(0), u \rangle_U. \]
In other words, $L_U$ agrees, up to first order, with the restriction of $f$ to $\overline{p} + U$. Continuing with our $U$-terminology, we will say that $\overline{g}_U$ is the $U$-gradient of $f$ at $\overline{p}$, and note that $\overline{g}_U$ is actually independent of the particular $\overline{g} \in \partial f(\overline{p})$ (recall Proposition 2.2(i)).

**Remark 3.4.** We add that, because $f$ is locally Lipschitzian, this $U$-differentiability property holds also tangentially to $U$:

$$f(\overline{p} + h) = f(\overline{p}) + \langle \overline{g}, h \rangle + o(\|h\|) \text{ whenever } \|h\| = o(\|h_U\|_U).$$

This remark will be instrumental when coming to higher order; then we will have to *select* $h$ appropriately, to allow a specification of the remainder term in (3.6); see Theorem 3.9.

As already mentioned, the existence of $\nabla L_U(0)$ is of paramount importance, since it suppresses the difficulty pointed out in the introduction of this paper; now the difference quotient in (1.1) takes the form

$$\frac{\partial L_U(u) - \overline{g}_U}{\|u\|_U},$$

which does make sense. Here is a useful first consequence: $W(u) = o(\|u\|_U)$.

**Corollary 3.5.** Assume (1.2). If $\overline{g} \in \ri \partial f(\overline{p})$, then

$$\forall \varepsilon > 0 \exists \delta > 0 : \|u\|_U \leq \delta \Rightarrow \|w\|_V \leq \varepsilon \|u\|_U \text{ for any } w \in W(u).$$

**Proof.** Use Theorem 3.3(ii) to write the first-order expansion of $L_U$:

$$L_U(u) = L_U(0) + \langle \nabla L_U(0), u \rangle_U + o(\|u\|_U) = f(\overline{p}) + \langle \overline{g}_U, u \rangle_U + o(\|u\|_U).$$

For any $w \in W(u)$ we have $L_U(u) = f(\overline{p} + u + w) - \langle \overline{g}_V, w \rangle_V$; therefore, (2.5) written for $v = w$, gives $L_U(u) \geq f(\overline{p}) + \langle \overline{g}_U, u \rangle_U + \eta \|w\|_V$. Altogether, we obtain

$$o(\|u\|_U) = L_U(u) - f(\overline{p}) - \langle \overline{g}_U, u \rangle_U \geq \eta \|w\|_V. \quad \Box$$

Let us sum up our results so far.

- Given $\overline{g} \in \partial f(\overline{p})$, we define via (3.1) a convex function $L_U$ (Theorem 3.2(i)), which is differentiable at $0$ and coincides up to first order with the restriction of $f$ to $\overline{p} + U$ (Theorem 3.3(ii)).
- When $W(\cdot) \neq 0$, this $U$-Lagrangian is indeed the restriction of $f$ to a “thick surface” $\{ \overline{p} + \cdot \cap W(\cdot) \}$, parametrized by $u \in U$.
- We also define, via Theorem 3.2(ii), a “thick selection” of $\partial f$ on this thick surface, made up of those subgradients that have the same $V$-component as $\overline{g}$.
- As a function of the parameter $u$, this thick selection behaves like a subdifferential, namely $\partial L_U$ (Theorem 3.3(i)).
- When $\overline{g} \in \ri \partial f(\overline{p})$, our thick surface has $U$ as “tangent space” at $\overline{p}$ (Corollary 3.5; we use quotation marks because $W$ is multivalued).

**Remark 3.6.** We note in passing two extreme cases in which our theory becomes trivial:

- when $f$ is differentiable at $\overline{p}$, then $U = \mathbb{R}^n$, $V = \{0\}$ and $L_U \equiv f$;
- when $\partial f(\overline{p})$ has full dimension, then $U = \{0\}$ and there is no $U$-Lagrangian. \( \Box \)
3.3. Higher-order behaviour. Proceeding further in our differential analysis of \( L_\mathcal{U} \), we now study the behaviour of \( \partial L_\mathcal{U} \) near 0. A very basic property of this set is its radial Lipschitz continuity. We say that \( f \) has a radially Lipschitz subdifferential at \( p \) when there is a \( D > 0 \) and a \( \delta > 0 \) such that
\[
\partial f(p + d) \subset \partial f(p) + B(0, D\|d\|), \quad \text{for all } d \in B(0, \delta).
\]
This is equivalent to an upper quadratic growth condition on the function itself (recall Corollary 3.5 in [14]): there is a \( C > 0 \) and an \( \varepsilon > 0 \) such that
\[
f(p + d) \leq f(p) + f'(p; d) + \frac{1}{2}C\|d\|^2, \quad \text{for all } d \in B(0, \varepsilon).
\]
This property is transmitted from \( f \) to \( L_\mathcal{U} \):

**Proposition 3.7.** Assume (1.2). Assume also that \( W(u) \) is nonempty for \( u \) small enough, and that (3.7) is satisfied. Then
\[
\text{(i) } \partial L_\mathcal{U}(u) \subset \mathcal{F}_\mathcal{U} + B_\mathcal{U}(0, 2C\|u\|_\mathcal{U}), \text{ for some } \delta > 0 \text{ and all } u \in B_\mathcal{U}(0, \delta);
\]
\[
\text{(ii) } L_\mathcal{U}(u) \leq L_\mathcal{U}(0) + \langle \mathcal{F}_\mathcal{U}, u \rangle + \frac{1}{2}R\|u\|_\mathcal{U}^2, \text{ for some } \rho > 0, R > 0 \text{ and all } u \in B_\mathcal{U}(0, \rho).
\]

**Proof.** Remember that \( \nabla L_\mathcal{U}(0) = \mathcal{F}_\mathcal{U} \). Because the subdifferential is an outer-semicontinuous mapping, we can choose \( \delta > 0 \) such that for all \( u \in B_\mathcal{U}(0, \delta) \) and \( g_\mathcal{U} \in \partial L_\mathcal{U}(u) \), \( \|g_\mathcal{U} - \mathcal{F}_\mathcal{U}\|_\mathcal{U} \leq \frac{\varepsilon}{2} \) (see §VI.6.2 of [11] for example). On the other hand, assume \( \delta \) so small that \( \mathcal{W}(u) \) contains some \( w; \) from Theorem 3.2(ii), \( g_\mathcal{U} + \mathcal{F}_\mathcal{U} \in \partial f(p + u + w) \).

Now \( \mathcal{U} \subset N_{\partial f(p)}(\mathcal{F}) \) (Proposition 2.2(iii)). Using the notation \( s := (g_\mathcal{U} - \mathcal{F}_\mathcal{U}) \oplus 0 \), so that \( g_\mathcal{U} + \mathcal{F}_\mathcal{U} = \mathcal{F} + s \in \partial f(p + u + w) \), we are in the conditions of Corollary 3.3 in [14] written with \( \varphi = f \), \( z_0 = p \), \( g_0 = \mathcal{F} \), \( x = p + u + w \). Inequality (14) therein becomes
\[
\|g_\mathcal{U} - \mathcal{F}_\mathcal{U}\|_\mathcal{U}^2 = \|s\|^2 \leq 2C(s, u + w) = 2C(g_\mathcal{U} - \mathcal{F}_\mathcal{U}, u)_\mathcal{U} \leq 2C\|g_\mathcal{U} - \mathcal{F}_\mathcal{U}\|_\mathcal{U}\|u\|_\mathcal{U},
\]
which is \( (i) \). As for \( (ii) \), it is equivalent to \( (i) \) (Corollary 3.5 in [14]).

Back to the \( f \)-context, Proposition 3.7 says: for small \( u \in \mathcal{U} \) and all \( w \in W(u) \), there holds
\[
\{g_\mathcal{U} : g_\mathcal{U} \oplus \mathcal{F}_\mathcal{U} \in \partial f(p + u + w)\} \subset \mathcal{F}_\mathcal{U} + B_\mathcal{U}(0, 2C\|u\|_\mathcal{U})
\]
as well as
\[
f(p + u + w) \leq f(p) + \langle \mathcal{F}, u + w \rangle + \frac{1}{2}R\|u\|_\mathcal{U}^2.
\]

Now, we have a function \( L_\mathcal{U} \), which is differentiable at 0, and whose second-order difference quotients inherit the qualitative properties of those of \( f \). The stage is therefore set to consider the case where \( L_\mathcal{U} \) has a generalized Hessian at 0, in the sense of [9] (see also [15], §3). Generally speaking, we say that a convex function \( \varphi \) has at \( z_0 \) a generalized Hessian \( H\varphi(z_0) \) when
\[
\text{(i) the gradient } \nabla \varphi(z_0) \text{ exists;}
\]
\[
\text{(ii) there exists a symmetric positive semidefinite operator } H\varphi(z_0) \text{ such that }
\]
\[
\varphi(z_0 + d) = \varphi(z_0) + \langle \nabla \varphi(z_0), d \rangle + \frac{1}{2}(H\varphi(z_0)d, d) + o(\|d\|^2);
\]
\[
\text{(iii) or equivalently,}
\]
\[
\partial \varphi(z_0 + d) \subset \nabla \varphi(z_0) + H\varphi(z_0)d + B(0, 0, o(\|d\|)).
\]
**Definition 3.8.** Assume (1.2). We say that $f$ has at $\bar{p}$ a $\mathcal{U}$-Hessian $H_{\mathcal{U}} f(\bar{p})$ (associated with $\bar{g}$) if $L_{\mathcal{U}}$ has a generalized Hessian at 0; then we set

$$H_{\mathcal{U}} f(\bar{p}) := H L_{\mathcal{U}} (0).$$

When it exists, the $\mathcal{U}$-Hessian $H_{\mathcal{U}} f(\bar{p})$ is therefore a symmetric positive semi-definite operator from $\mathcal{U}$ to $\mathcal{U}$. Its existence means the possibility of expanding $f$ along the thick surface $\bar{p} + \cdot \oplus W(\cdot)$ introduced at the end of §3.2.

**Theorem 3.9.** Take $\bar{p} \in \text{ri} \partial f(\bar{p})$ and let the $\mathcal{U}$-Hessian $H_{\mathcal{U}} f(\bar{p})$ exist. For $u \in \mathcal{U}$ and $h \in u \oplus W(u)$, there holds

$$f(\bar{p} + h) = f(\bar{p}) + \langle \bar{g}, h \rangle + \frac{1}{2} \langle H_{\mathcal{U}} f(\bar{p}) u, u \rangle_{\mathcal{U}} + o(\|h\|^2). \quad (3.10)$$

**Proof.** We know from Theorem 3.2(iv) that $W(u) \neq \emptyset$. Then apply the definition of $L_{\mathcal{U}}$ and expand $L_{\mathcal{U}}$ to obtain for all $u$ and $w \in W(u)$:

$$L_{\mathcal{U}}(u) = f(\bar{p} + u \oplus w) - \langle \bar{g}_V, w \rangle_V = L_{\mathcal{U}}(0) + \langle \nabla L_{\mathcal{U}}(0), u \rangle_{\mathcal{U}} + \frac{1}{2} \langle H_{\mathcal{U}} f(\bar{p}) u, u \rangle_{\mathcal{U}} + o(\|u\|^2_{\mathcal{U}}) = f(\bar{p}) + \langle \bar{g}_U, u \rangle_{\mathcal{U}} + \frac{1}{2} \langle H_{\mathcal{U}} f(\bar{p}) u, u \rangle_{\mathcal{U}} + o(\|u\|^2_{\mathcal{U}}).$$

In view of Corollary 3.5, $o(\|u\|^2_{\mathcal{U}}) = o(\|h\|^2)$; (3.10) follows, adding $\langle \bar{g}_V, w \rangle_V$ to both sides.

To the second-order expansion (3.10), there corresponds a first-order expansion of selected subgradients along the thick surface $\bar{p} + \cdot \oplus W(\cdot)$: with the notation and assumptions of Theorem 3.9,

$$\{g_{\mathcal{U}} : g_{\mathcal{U}} \oplus \bar{g}_V \in \partial f(\bar{p} + h)\} \subset \bar{g}_U + H_{\mathcal{U}} f(\bar{p}) u + B_{\mathcal{U}}(0, o(\|h\|)).$$

With reference to Remark 3.4, the expansion (3.10) makes (3.6) more explicit, for increments $h = h_{\mathcal{U}} \oplus h_V$ such that $h_V \in W(h_{\mathcal{U}})$. The aim of the next section is to disclose some intrinsic interest of these particular $h$’s.

**4. Examples of application**

This section shows how the $\mathcal{U}$-concepts developed in §3 generalize well-known objects. We will first consider special situations: max-functions (§4.1) and semi-definite programming (§4.2). Then in §4.3 we outline a conceptual minimization algorithm.

**4.1. Exact penalty.** Consider an ordinary nonlinear programming problem

$$\begin{align*}
\min \psi(p), \\
f_i(p) \leq 0, & \quad i = 1, \ldots, m,
\end{align*} \quad (4.1)$$

with convex $C^2$ data $\psi$ and $f_i$. Take an optimal $\bar{p}$ and suppose that the KKT conditions hold: with $L(p, \lambda) := \psi(p) + \sum \lambda_i f_i(p)$, defined for $(p, \lambda) \in \mathbb{R}^n \times \mathbb{R}^m$, there exist Lagrange multipliers $\lambda_i$ such that

$$\begin{align*}
[\nabla_p L(p, \lambda) =] \quad \nabla \psi(\bar{p}) + \sum \lambda_i \nabla f_i(\bar{p}) = 0, \\
\lambda_i \geq 0 \quad \lambda_i f_i(\bar{p}) = 0, & \quad \text{for } i = 1, \ldots, m. \quad (4.2)
\end{align*}$$

We will use the notation $\gamma := \nabla \psi$, $g_i := \nabla f_i$, $\gamma := \nabla \psi(\bar{p})$, $\bar{g}_i := \nabla f_i(\bar{p})$. 

---

License or copyright restrictions may apply to redistribution; see https://www.ams.org/journal-terms-of-use
Consider now an exact penalty function associated with (4.1): with \( f_0(p) \equiv 0 \) (and \( g_0(p) := \nabla f_0(p) \equiv 0 \)), set
\[
(4.3) \quad f(p) := \psi(p) + \pi \max\{f_0(p), \ldots, f_m(p)\},
\]
where \( \pi > 0 \) is a penalty parameter. Call
\[
J(p) := \{ j \in \{0, \ldots, m\} : \psi(p) + \pi f_j(p) = f(p) \}
\]
the set of indices realizing the max at \( p \). Standard subdifferential calculus gives
\[
\partial f(p) = \gamma(p) + \pi \text{conv}\{g_j(p) : j \in J(p)\}.
\]
In NLP language, instead of maximal functions, one speaks of active constraints. We therefore set
\[
\mathcal{T} := \{ i \in \{1, \ldots, m\} : f_i(\overline{p}) = 0 \}
\]
(naturally, we assume \( \mathcal{T} \neq \emptyset \); otherwise, the problem lacks interest). It is easy to see that \( J(\overline{p}) = \mathcal{T} \cup \{0\} \); correspondingly, we associate with \( J(\overline{p}) \) the “multipliers”
\[
(4.4) \quad \pi_i := \overline{\lambda}_i \text{ for } i \in \mathcal{T} \quad \text{and} \quad \pi_0 := \pi - \sum_{i \in \mathcal{T}} \overline{\lambda}_i.
\]

For \( \pi \) large enough, it is well known that \( \overline{p} \) solving (4.1) also minimizes \( f \) of (4.3).

We proceed to apply the theory of §3 to the present situation: \( f \) is the exact penalty function of (4.3), \( \overline{p} \) is optimal and \( \overline{\gamma} = 0 \). We will show that the \( \mathcal{U} \)-Lagrangian \( L_{\mathcal{U}} \)
coincides up to second order with the restriction to \( \mathcal{U} \) of the ordinary Lagrangian \( L(\overline{p} + \cdot, \overline{\lambda}) \). All along this subsection, we make the following assumptions:

- the active gradients \( \{\overline{g}_i\}_{i \in \mathcal{T}} \) are linearly independent (hence \( \overline{\lambda} \) is unique in the KKT conditions (4.2)),
- \( \overline{\lambda}_i > 0 \) for \( i \in \mathcal{T} \) (strict complementarity),
- and \( \pi > \sum_{i \in \mathcal{T}} \overline{\lambda}_i \), i.e., \( \pi_0 > 0 \) in (4.4).

The following development should be considered as a mere illustration of the \( \mathcal{U} \)-theory. This is why we content ourselves with the above simplifying assumptions, which are relaxed in the more complete work of [17].

We start with a basic result, stating in particular that \( \mathcal{U} \) is the space tangent to the surface defined by the active constraints (well-defined thanks to our simplifying assumptions).

**Proposition 4.1.** With the above notation and assumptions, we have the following relations for \( p = \overline{p} \):

(i) \( \partial f(\overline{p}) = \overline{\gamma} + \{ \sum_{i \in \mathcal{T}} \mu_i \overline{g}_i : \mu_i \geq 0, \sum_{i \in \mathcal{T}} \mu_i \leq \pi \} \);

(ii) the subspaces \( \mathcal{U} \) and \( \mathcal{V} \) of Definition 2.1 are
\[
\mathcal{V} = \text{lin}\{\overline{g}_i\}_{i \in \mathcal{T}}, \quad \mathcal{U} = \{ d \in \mathbb{R}^n : \langle \overline{g}_i, d \rangle = 0, i \in \mathcal{T} \};
\]

(iii) \( \overline{\gamma} := 0 \in \text{ri} \partial f(\overline{p}) \).

**Proof.** (i) We have
\[
\partial f(\overline{p}) = \overline{\gamma} + \pi \text{conv}\{\overline{g}_i : i \in \mathcal{T} \cup \{0\}\}
= \overline{\gamma} + \{ \pi \alpha_0 + \sum_{i \in \mathcal{T}} \pi \alpha_i \overline{g}_i : \alpha_i \geq 0, \alpha_0 + \sum_{i \in \mathcal{T}} \alpha_i = 1 \}.
\]
The formula is then straightforward, setting \( \mu_i := \pi \alpha_i \) and eliminating the unnecessary vector 0.
(ii) Apply Definition 2.1(ii): \( V = \text{lin}\{\partial f(\mathbf{p}) - \mathbf{v}\} \) because \( \mathbf{v} \in \partial f(\mathbf{p}) \). Together with (i), the results clearly follow.

(iii) Consider the set \( B := \{\sum_{i} \mu_i \mathbf{g}_i : \mu_i \geq -\mathbf{n}_i, \sum_{i} \mu_i \leq \mathbf{n}_0\} \), where \( \mathbf{p} \) was defined in (4.4). Because of (ii), \( B \subset V \). Because of strict complementarity and \( \mathbf{n}_0 > 0 \), \( B \) is a relative neighborhood of \( 0 = \mathbf{g} \in V \). Finally, because of (4.2) and (4.4),

\[
B = \mathbf{g} + B + \sum_{i} \lambda_i \mathbf{g}_i
\]

\[
= \mathbf{g} + \left\{ \sum_{i} (\mu_i + \mathbf{n}_i) \mathbf{g}_i : \mu_i + \mathbf{n}_i \geq 0, \sum_{i} (\mu_i + \mathbf{n}_i) \leq \pi \right\}.
\]

In view of (i), \( B \subset \partial f(\mathbf{p}) \) and we are done. \( \Box \)

**Lemma 4.2.** With the notation and assumptions of this subsection, let \( p \) be close to \( \mathbf{p} \). Then \( J(\mathbf{p}) \subset J(\mathbf{p}) = T \cup \{0\} \) and the system in \( \{\mu_i\}_{J(\mathbf{p})} \)

\[
\begin{align*}
\langle \mathbf{g}_i, \gamma(p) \rangle + \sum_{j \in J(\mathbf{p})} \mu_j \langle \mathbf{g}_i, g_j(p) \rangle &= 0 \quad \text{for all } i \in T, \\
\sum_{j \in J(\mathbf{p})} \mu_j &= \pi
\end{align*}
\]

(4.5)

has a solution, which is unique, if and only if \( J(p) = J(\mathbf{p}) = T \cup \{0\} \). The solution \( \mu(p) \) satisfies \( \mu_j(p) > 0 \) for all \( j \in J(p) = J(\mathbf{p}) \). Moreover, \( \mu(\mathbf{p}) = \mathbf{p} \) of (4.4) and \( p \mapsto \mu(p) \) is differentiable at \( p = \mathbf{p} \).

**Proof.** Let \( j \notin J(\mathbf{p}) \). By continuity, \( f_j(p) < f_i(p) \) for all \( i \in J(\mathbf{p}) \), hence \( J(p) \subset J(\mathbf{p}) \).

Now consider (4.5). First, observe that, because of (4.2), \( \mathbf{p} \) of (4.4) is a solution at \( p = \mathbf{p} \).

(a) Assume first that \( J(p) = J(\mathbf{p}) = T \cup \{0\} \). Since \( g_0(p) \equiv 0 \), the variable \( \mu_0 \) is again directly given by \( \mu_0(p) = \pi - \sum_{j \in T} \mu_j(p) \). As for the \( \mu_j \)'s, \( j \in T \), they are given by an \( T \times T \) linear system, whose matrix is \( (\mathbf{g}_i, g_j(p))_{ij} \). Because the \( \mathbf{g}_i \)'s are linearly independent, this matrix is positive definite. The solution \( \mu(p) \) is unique; it is also close to \( \mathbf{p} \), is therefore positive and sums up to less than \( \pi \): \( \mu_0(p) > 0 \). In particular, \( \mu(\mathbf{p}) = \mathbf{p} \) is the unique solution at \( p = \mathbf{p} \). The differentiability property then follows from the Implicit Function Theorem.

(b) On the other hand, assume the set \( I_0 := J(\mathbf{p}) \setminus J(p) \) is nonempty and suppose (4.5) has a solution \( \{\mu^*_j\}_{j \in J(p)} \). Set \( \mu^*_j := 0 \) for \( j \in I_0 \); then \( \mu^* \) also solves (4.5) with \( J(p) \) replaced by \( J(\mathbf{p}) \). This contradicts part (a) of the proof. \( \Box \)

The next result reveals a nice interpretation of \( W(\cdot) \) in (3.2): it makes a local description of the surface defined by the active constraints.

**Theorem 4.3.** Use the notation and assumptions of this subsection. For \( u \in U \) small enough, \( W(u) \) defined in (3.2) is a singleton \( w(u) \), which is the unique solution of the system with unknown \( v \in V \)

\[
f_i(p + u \oplus v) = 0, \quad \text{for all } i \in T.
\]

(4.6)

**Proof.** According to Theorem 3.2(ii) and (3.5), an arbitrary \( p \in \mathbf{p} + u \oplus W(u) \) is characterized by \( \partial f(p) \cap U \neq \emptyset \); there are convex multipliers \( \{\alpha_j\}_{j \in J(p)} \) such that \( \gamma(p) + \pi \sum_{j \in J(p)} \alpha_j g_j(p) \in U \). Setting \( \mu_j := \pi \alpha_j \), this means that the system (4.5)
has a nonnegative solution. Now, in view of Proposition 4.1(iii) and Corollary 3.5, \( p - \overline{p} \) is small; we can apply Lemma 4.2, \( J(p) = I \cup \{0\} \), and this is just (4.6).

Uniqueness of such a \( p \) is then easy to prove. Substituting \( f_i \) for \( h_2 \) in Remark 2.3, the gradients of the functions \( v \mapsto f_i(\overline{p} + u \oplus v) \) are \( g_i(\overline{p} + u \oplus v) \), which are linearly independent for \((u, v) = (0, 0)\). By the Implicit Function Theorem, (4.6) has a unique solution \( w(u) \) for small \( u \).

Now we are in a position to give specific expressions for the derivatives of the \( \mathcal{U} \)-Lagrangian.

**Theorem 4.4.** Use the notation and assumptions of this subsection.

(i) The \( \mathcal{U} \)-Lagrangian is differentiable in a neighborhood of 0. With \( \mu(\cdot) \) and \( w(\cdot) \) defined in Lemma 4.2 and Theorem 4.3 respectively, and with \( p(u) := \overline{p} + u \oplus w(u) \),

we have for \( u \in \mathcal{U} \) small enough

\[
\nabla L_\mathcal{U}(u) \oplus 0 = \gamma(p(u)) + \sum_{j \in I} \mu_j(p(u))g_j(p(u)).
\]

(ii) The Hessian \( \nabla^2 L_\mathcal{U}(0) \) exists. Using the matrix-like decomposition

\[
\nabla^2_{pp} L(\overline{p}, \overline{\lambda}) = \begin{pmatrix}
H_{\mathcal{U}\mathcal{U}} & H_{\mathcal{U}\mathcal{V}} \\
H_{\mathcal{V}\mathcal{U}} & H_{\mathcal{V}\mathcal{V}}
\end{pmatrix}
\]

for the Hessian of the Lagrangian, we have \( \nabla^2 L_\mathcal{U}(0) = H_{\mathcal{U}\mathcal{U}} \).

**Proof.** (i) Put together Lemma 4.2 and Theorem 4.3. Observe, in particular, that the right-hand side of (4.7) lies in \( \mathcal{U} \). Then invoke (3.5).

(ii) In view of Lemma 4.1(iii) and Corollary 3.5, \( w(u) = o(||u||_U) \), hence \( p(\cdot) \) has a Jacobian at 0; in fact, \( J(p(0))u = u \oplus 0 \) for all \( u \in \mathcal{U} \). Then, using Lemma 4.2, (4.7) clearly shows that \( \nabla L_\mathcal{U} \) is differentiable at 0. Compute from (4.7) the differential \( \nabla^2 L_\mathcal{U}(0)u \) for \( u \in \mathcal{U} \):

\[
\nabla^2 L_\mathcal{U}(0)u \oplus 0 = \nabla^2 \psi(\overline{p})Jp(0)u + \sum_{j \in I} \lambda_j \nabla^2 f_j(\overline{p})Jp(0)u \\
+ \sum_{j \in I} \langle \nabla \mu_j(\overline{p}), Jp(0)u \rangle \overline{g}_j \\
= \nabla^2_{pp} L(\overline{p}, \overline{\lambda})(u \oplus 0) + \sum_{j \in I} \langle \nabla \mu_j(\overline{p}), Jp(0)u \rangle \overline{g}_j.
\]

Thus, \( \nabla^2 L_\mathcal{U}(0)u \) is the \( \mathcal{U} \)-part of the right-hand side. The second term is a sum of vectors in \( \mathcal{V} \), which does not count; we do obtain (ii).

In Remark 3.1 we have said that \( \overline{g} \) in §3 plays the role of a dual variable. This is suggested by the relation \( 0 = \overline{g} + \sum_{i \in I} \lambda_i \overline{g}_i \in \partial f(\overline{p}) \), which, in the present NLP context, establishes a correspondence between \( \overline{g} = 0 \) and the multipliers \( \lambda_i \) or \( \overline{p}_i \). Taking some nonzero \( \overline{g}' \in ri \partial f(\overline{p}) \) does not change the situation much; this just amounts to applying the theory to \( f - \langle \overline{g}', \cdot \rangle \), which is still minimal at \( \overline{p} \) — but of course the multipliers are changed, say, to \( \overline{\lambda}_i \) or \( \overline{p}'_i \). Denoting by \( g(p(u)) \) the right-hand side in (4.7), the correspondence \( \overline{g} \leftrightarrow \overline{\lambda} \leftrightarrow p \) can even be extended to \( g(p(u)) \leftrightarrow \overline{\lambda}(u) \leftrightarrow \mu(u) \).
4.2. Eigenvalue optimization. Consider the problem of minimizing with respect to \( x \in \mathbb{R}^m \) the largest eigenvalue \( \lambda_1 \) of a real symmetric \( n \times n \) matrix \( A \), depending affinely on \( x \). Most of the relevant information for the function \( \lambda_1 \circ A \) can be obtained by analyzing the maximum eigenvalue function \( \lambda_1(A) \), which is convex (and finite-valued). We briefly describe here how the \( \mathcal{U} \)-theory applies to this context. For a detailed study, we refer to [20] where an interesting connection is established with the geometrical approach of [21].

For the sake of consistency, we keep the notation \( p := A(\pi) \) for the reference matrix where the analysis is performed. If \( r \) denotes the multiplicity of \( \lambda_1(p) \), then \( \mathcal{W}_r := \{ p : p \text{ is a symmetric matrix and } \lambda_1(p) \text{ has multiplicity } r \} \) is the smooth manifold \( \Omega \) of [21].

First, the subspaces \( \mathcal{U} \) and \( \mathcal{V} \) in Definition 2.1 are just the tangent and normal spaces to \( \mathcal{W}_r \) at \( p \) (Corollary 4.8 in [20]). Similarly to Theorem 4.3, Theorem 4.11 in [20] shows that the set \( \mathcal{W}(u) \) of (3.2) is a singleton \( w(u) \), characterized by \( p + u \oplus w(u) \in \mathcal{W}_r \).

As for second order, the \( \mathcal{U} \)-Lagrangian (3.1) is twice continuously differentiable in a neighbourhood of \( 0 \in \mathcal{U} \). Finally, use again the matrix-like decomposition

\[
\begin{pmatrix}
H_{\mathcal{U}U} & H_{\mathcal{U}V} \\
H_{\mathcal{V}U} & H_{\mathcal{V}V}
\end{pmatrix}
\]

for the Hessian of the Lagrangian introduced in Theorem 5 of [21]. Then Theorem 4.12 in [20] shows that \( \nabla^2 L_{\mathcal{U}}(0) = H_{\mathcal{U}U} \) is the reduced Hessian matrix (5.31) in [21].

4.3. A conceptual superlinear scheme. The previous subsections have shown that our \( \mathcal{U} \)-objects become classical when \( f \) has some special form. It is also demonstrated in [17] and [20] how these \( \mathcal{U} \)-objects can provide interpretations of known minimization algorithms. Here we go back to a general \( f \) and we design a superlinearly convergent conceptual algorithm for minimizing \( f \). Again, we obtain a general formalization of known techniques from classical optimization.

Given \( p \) close to a minimum point \( \bar{p} \), the problem is to compute some \( p_+ \), superlinearly closer to \( \bar{p} \). We propose a conceptual scheme, in which we compute first the \( \mathcal{V} \)-component of the increment \( p_+ - p \), and then its \( \mathcal{U} \)-component. This idea of decomposing the move from \( p \) to \( p_+ \) in a “vertical” and a “horizontal” step can be traced back to [8].

**Algorithm 4.5.** \( \mathcal{V} \)-Step. Compute a solution \( \delta v \in \mathcal{V} \) of

\[
\min \{ f(p + 0 \oplus \delta v) : \delta v \in \mathcal{V} \}
\]

and set \( p' := p + 0 \oplus \delta v \).

\( \mathcal{U} \)-Step. Make a Newton step in \( p' + \mathcal{U} \): compute the solution \( \delta u \in \mathcal{U} \) of

\[
g'_{\mathcal{U}} + H_{\mathcal{U}f}(\bar{p})\delta u = 0,
\]

where \( g' \in \partial f(p') \) is such that \( g'_{\mathcal{V}} = 0 \), so that \( g'_{\mathcal{U}} \in \partial L_{\mathcal{U}}((p' - \bar{p})_{\mathcal{U}}) \).

Update. Set \( p := p' + \delta u \oplus 0 = p + \delta u \oplus \delta v \).

**Remark 4.6.** This algorithm needs the subspace \( \mathcal{U} \) associated with \( \bar{p} \), as well as the \( \mathcal{U} \)-Hessian \( H_{\mathcal{U}f}(\bar{p}) \), which must exist and be positive definite. The knowledge of \( \mathcal{U} \) may be considered as a bold requirement; constructing appropriate approximations of it is for sure a key to obtain implementable forms. As for existence and positive
definiteness of $H_uf(\overline{p})$, it is a natural assumption. Quasi-Newton approximations of it might be suitable, as well as other approaches in the lines of [27].

The next result supports our scheme.

Theorem 4.7. Using the notation of §3, assume that $g := 0 \in \text{ri} \partial f(\overline{p})$, and that $f$ has at $\overline{p}$ a positive definite $U$-Hessian. Then the point $p_+$ constructed by Algorithm 4.5 satisfies

$$\|p_+ - p\| = o(\|p - \overline{p}\|).$$

Proof. We denote by $u := (p - \overline{p})_U$ the $U$-component of $p - \overline{p}$ (see Figure 2). For $\delta v \in V$, make the change of variables $v := (p - \overline{p})_V + \delta v$, so that (4.8) can be written

$$\min_{v \in V} f(p + u \oplus v).$$

Denoting by $v_+$ a solution, we have

$$v_+ = (p - \overline{p})_V + \delta v = (p_+ - \overline{p})_V \in W(u)$$

and Corollary 3.5 implies that

$$\|v_+\|_V = o(\|u\|_U) = o(\|p - \overline{p}\|).$$

From the definition (3.9) of $H_uf(\overline{p})$ and observing that $\nabla L_{U}(0) = 0$, we have

$$\partial L_{U}(u) \ni g_u = 0 + H_uf(\overline{p})u + o(\|u\|_U).$$

Subtracting from (4.9), $H_uf(\overline{p})(u + \delta u) = o(\|u\|_U)$ and, since $H_uf(\overline{p})$ is invertible, $\|u + \delta u\|_U = o(\|u\|_U)$. Then, writing

$$x = (p_+ - \overline{p})_U = o(\|u\|_U) = o(\|p - \overline{p}\|).$$

we do have

$$\|(p_+ - \overline{p})_U\| = o(\|u\|_U) = o(\|p - \overline{p}\|).$$

With (4.10), the conclusion follows. □

5. $U$-Hessian and Moreau-Yosida regularizations

The whole business of §3 was to develop a theory ending up with the definition of a $U$-Hessian (Definition 3.8). Our aim now is to assess this concept: we give a necessary and sufficient condition for the existence of $H_uf$, in terms of Moreau-Yosida regularization ([32], [19]).
We denote by $F$ the Moreau-Yosida regularization of $f$, associated with the Euclidean metric,

\begin{equation}
F(x) := \min_{y \in \mathbb{R}^n} \{ f(y) + \frac{1}{2} \| x - y \|^2 \}.
\end{equation}

The unique minimizer in (5.1), called the proximal point of $x$, is denoted by

\begin{equation}
p(x) := \arg\min_{y \in \mathbb{R}^n} \{ f(y) + \frac{1}{2} \| x - y \|^2 \}.
\end{equation}

It is well known that $F$ has a (globally) Lipschitzian gradient, satisfying

\begin{equation}
\nabla F(x) = x - p(x) \in \partial f(p(x)).
\end{equation}

Given $\overline{p}$ and $\overline{g}$ satisfying (1.2), we are interested in the behaviour of $F$ near

\begin{equation}
\overline{x} := \overline{p} + \overline{g}
\end{equation}
(recall, for example, Theorem 2.8 of [15]: $\overline{g} = \nabla F(\overline{x})$ and $\overline{x}$ is such that $p(\overline{x}) = \overline{p}$).

More precisely, restricting our attention to $\overline{x} + \mathcal{U}$, we will give an equivalence result and a formula linking the so restricted Hessian of $F$, with the $\mathcal{U}$-Hessian of $f$ at $\overline{p}$. To prove our results, we introduce an intermediate function, similar to $\phi_V$ in §3.2 of [15], but adapted to our $\mathcal{U}$-context:

\begin{equation}
\mathcal{U} \ni u \mapsto \phi_V(u) := \min_{v \in V} \{ f(\overline{p} + u + v) - \langle \overline{g}_V, v \rangle_V + \frac{1}{2} \| v \|_V^2 \}.
\end{equation}

We start by showing that this function agrees up to second order with $L_{\mathcal{U}}$.

**Lemma 5.1.** With the notation above, assume that the conclusion of Corollary 3.5 holds for at least one $w \in W(u)$ – for example, let $\overline{g}$ be in $\text{ri} \partial f(\overline{p})$. Then

$$\forall \varepsilon > 0 \exists \delta > 0 : \| u \|_{\mathcal{U}} \leq \delta \Rightarrow |\phi_V(u) - L_{\mathcal{U}}(u)| \leq \varepsilon \| u \|_{\mathcal{U}}^2.$$

In particular,

\begin{equation}
\nabla \phi_V(0) = \overline{g}_{\mathcal{U}} \quad \text{and} \quad \exists \text{HL}_{\mathcal{U}}(0) \iff \exists \text{H} \phi_V(0) = \text{HL}_{\mathcal{U}}(0).
\end{equation}

**Proof.** Clearly $\phi_V(u) \geq L_{\mathcal{U}}(u)$. To obtain an opposite inequality, write the minimand in (5.5) for $v = w \in W(u)$:

$$\phi_V(u) \leq f(\overline{p} + u + w) - \langle \overline{g}_V, w \rangle_V + \frac{1}{2} \| w \|_V^2$$

$$= L_{\mathcal{U}}(u) + \frac{1}{2} \| w \|_V^2.$$

Taking, in particular, $w$ such that $\| w \|_V = o(\| u \|_{\mathcal{U}})$ (or applying Corollary 3.5), the results follow. \hfill \Box

The reason for introducing $\phi_V$ is that its Moreau-Yosida regularization $\Phi_V$ is obtained from the restriction $F_{\mathcal{U}}$ of $F$ to $\overline{x} + \mathcal{U}$ by a mere translation.

**Proposition 5.2.** Assume (1.2). The two functions

$$\mathcal{U} \ni d_\mathcal{U} \mapsto \begin{cases} 
\Phi_V(d_\mathcal{U}) := \min_{u \in \mathcal{U}} \{ \phi_V(u) + \frac{1}{2} \| d_\mathcal{U} - u \|_{\mathcal{U}}^2 \}, \\
F_{\mathcal{U}}(d_\mathcal{U}) := F(\overline{x} + d_\mathcal{U} + 0),
\end{cases}$$

satisfy

$$F_{\mathcal{U}}(d_\mathcal{U}) = \Phi_V(\overline{g}_{\mathcal{U}} + d_\mathcal{U}) + \frac{1}{2} \| \overline{g}_V \|_V^2 \quad \text{for all } d_\mathcal{U} \in \mathcal{U}.$$
Proof. Take $d_U \in U$. Recalling (5.4), compute $F_U(d_U) = F(\varphi + (\overline{g}_U + d_U) \circ \overline{g}_V)$ in the following tricky way:

$$F_U(d_U) = \min_{(u,v) \in U \times V} \{ f(\varphi + u \oplus v) + \frac{1}{2}\| (\overline{g}_U + d_U - u) \circ \overline{g}_V - v \|^2 \} = \min_{u \in U} \left\{ \min_{v \in V} \{ f(\varphi + u \oplus v) + \frac{1}{2}\| \overline{g}_V - v \|^2 \} + \frac{1}{2}\| \overline{g}_U + d_U - u \|^2 \right\} = \min_{u \in U} \{ \phi_V(u) + \frac{1}{2}\| \overline{g}_U \|^2 + \frac{1}{2}\| \overline{g}_U + d_U - u \|^2 \} = \Phi_V(g_U + d_U) + \frac{1}{2}\| \overline{g}_V \|^2. \quad \square$$

Since $L_U$ is so close to $\phi_V$ (Lemma 5.1), its Moreau-Yosida regularization is close to $\Phi_V$, i.e., to $F_U$, up to a translation. This explains the next result, which is the core of this section.

**Theorem 5.3.** Make the assumptions of Lemma 5.1.

(i) If $H_U f(\varphi)$ exists, then $\nabla^2 F_U(0)$ exists and is given by

$$\nabla^2 F_U(0) = I_U - (I_U + H_U f(\varphi))^{-1};$$

here $I_U$ denotes the identity in $U$.

(ii) Conversely, assume that $\nabla^2 F_U(0)$ exists. If (3.7) \(\equiv\) (3.8) holds, then $H_U f(\varphi)$ exists and is given by

$$H_U f(\varphi) = (I_U - \nabla^2 F_U(0))^{-1} - I_U.$$

If, in addition, $H_U f(\varphi)$ is positive definite – for example, if $f$ is strongly convex, we also have

$$H_U f(\varphi) = (\nabla^2 F_U(0))^{-1} - I_U.$$ 

Proof. (i) When $H_U f(\varphi)$ exists, use (5.6) to see that

$$H_U f(\varphi) = HL_U(0) = H\phi_V(0).$$

Then we can apply Theorem 3.1 of [15] to $\phi_V$. We see from (5.6) that the proximal point giving $\Phi_V(g_U)$ is $0 \in U$, so we have

$$\nabla^2 \Phi_V(g_U) = I_U - (I_U + H\phi_V(0))^{-1}.$$ 

In view of Proposition 5.2 and (5.9), this is just (5.7).

(ii) Combine Proposition 3.7(i) with Lemma 5.1 to see that (3.7) \(\equiv\) (3.8) also holds for $\phi_V$ at $0 \in U$; furthermore, $\nabla \phi_V(0)$ exists. Then we can apply Theorem 3.14 of [15] to $\phi_V$: when $\nabla^2 \Phi_V(g_U) = \nabla^2 F_U(0)$ exists, then $H\phi_V(0) = H_U f(\varphi)$ exists. We can write (5.7) and invert it to obtain (5.8).

Finally, suppose that $f$ is strongly convex: for some $c > 0$ and all $(u, w) \in U \times V$,

$$f(\varphi + u \oplus w) \geq f(\varphi) + (\overline{g}, u \oplus w) + \frac{c}{2}\| u \oplus w \|^2 \geq f(\varphi) + (\overline{g}_U, u) + (\overline{g}_V, w) + \frac{c}{2}\| u \|^2_U.$$ 

Take $w \in W(u)$ and subtract $(\overline{g}_V, w)\nu$ from both sides

$$L_U(u) \geq L_U(0) + \langle \nabla L_U(0), u \rangle_U + \frac{c}{2}\| u \|^2_U,$$

hence $H_U f(\varphi) = HL_U(0)$ is certainly positive definite. Computing its inverse from (5.8) and applying (20) from [15], we obtain the last relation. \(\square\)
A consequence of this result is that, when $\nabla^2 F(x)$ exists, then $H_{\mathcal{U}t} f(\overline{p})$ exists; $\nabla^2 F_U(0)$ is just the $\mathcal{U}t$-block of $\nabla^2 F(x)$. Furthermore, $x \mapsto p(x)$ has at $\overline{p}$ a Jacobian of the form

$$Jp(\overline{p}) = I - \nabla^2 F(\overline{p}) = \begin{pmatrix} P & 0 \\ 0 & 0 \end{pmatrix}$$

(recall Corollary 2.6 in [15]). If $f$ satisfies (3.8) at $\overline{p}$, then

$$P = (I - \nabla^2 F(\overline{p}))_{\mathcal{U}t} = I_{\mathcal{U}t} - \nabla^2 F_U(0) = (H_{\mathcal{U}t} f(\overline{p}) + I_{\mathcal{U}t})^{-1}$$

is positive definite.

6. Conclusion

The distinctive difficulty of nonsmooth optimization is that the graph of $f$ near a minimum point $\overline{p}$ behaves like an elongated, gully-shaped valley. Such a valley is relatively easy to describe in the composite case (max-functions, maximal eigenvalues): it consists of those points where the non-differentiability of $f$ stays qualitatively the same as at $\overline{p}$; see the considerations developed in [22]. In the general case, however, even an appropriate definition of this valley is already not clear. We believe that the main contribution of this paper lies precisely here: we have generalized the concept of the gully-shaped valley to arbitrary (finite-valued) convex functions. To this aim, we have adopted the following process:

- First, we have used the tangent space to the active constraints, familiar in the NLP world; this was $\mathcal{U}$ of Definition 2.1.
- Then we have defined the gully-shaped valley, together with its parametrization by $u \in \mathcal{U}$, namely the mapping $W(\cdot)$ of (3.2).
- At the same time, we have singled out in (3.5) a selection of subgradients of $f$, together with a potential function $L_{\mathcal{U}}$. A nice feature is that our definitions are constructive via (3.1).
- This has allowed us to reduce the second-order study of $f$, restricted to the valley, to that of $L_{\mathcal{U}}$ (in $\mathcal{U}$).
- We have shown how our generalizations reduce to known objects in composite optimization, and how they can be used for the design of superlinearly convergent algorithms.
- Finally, we have related our new objects with the Moreau-Yosida regularization of $f$.

Acknowledgment

We are deeply indebted to R. Mifflin, for his careful reading and numerous helpful suggestions. The $\mathcal{U}$-terminology is due to him.

References

15. M. Kawasaki, 
14. C. Lemaréchal and C. Sagastizábal,
Nonsmooth analysis and the theory of fans
12. A.D. Ioffe,
10. J.-B. Hiriart-Urruty and C. Lemaréchal,
Nonsmooth optimization and the theory of fans
9. J.-B. Hiriart-Urruty, 
8. A.R. Conn,
Constrained optimization using a nondifferentiable penalty function
7. R. Cominetti and R. Correa,
On second derivatives for nonsmooth functions
5. R.W. Chaney,
6. T.F. Coleman and A.R. Conn,
Nonlinear optimization via an exact penalty function: asymptotic analysis, Mathematical Programming 24 (1982), 123–136. MR 84e:90087a


INRIA, 655 avenue de l’Europe, 38330 Montbonnot, France
E-mail address: Claude.Lemarechal@inria.fr

INRIA, 655 avenue de l’Europe, 38330 Montbonnot, France
E-mail address: Francois.Oustry@inria.fr

INRIA, BP 105, 78153 Le Chesnay, France
E-mail address: Claudia.Sagastizabal@inria.fr