Geometrical interpretation of the predictor-corrector type algorithms in structured optimization problems

Aris Daniilidis, Warren Hare, Jérôme Malick

Geometrical interpretation of the predictor-corrector type algorithms in structured optimization problems

Aris DANIILIDIS Warren HARE Jérôme MALICK

Dedicated to Professor D. Pallaschke for the occasion of his 65th birthday.

Key words: Proximal algorithm, $\mathcal{U}$-Lagrangian, partly-smooth function, Newton-type methods, Riemannian gradient.

AMS 2000 Subject Classification: *Primary*: 49J52 49K30 53B21 ; *Secondary*: 90C26 90C53 90C55

Abstract

It has been observed that in many optimization problems, nonsmooth objective functions often appear smooth on naturally arising manifolds. This has led to the development of optimization algorithms which attempt to exploit this smoothness. Many of these algorithms follow the same two step pattern: first predict a direction of decrease, and second make a correction step to return to the manifold. In this paper we examine some of the theoretical components used in such predictor-corrector methods. We begin our examination under the minimal assumption that the restriction of the function to the manifold is smooth. At the second stage, we add the condition of "partial smoothness" relative to the manifold. Finally, we examine the case when the function is both "prox-regular" and partly smooth. In this final setting we show that the proximal point mapping can be used to return to the manifold, and argue that returning in this manner is preferable to returning via the projection mapping. We finish by developing sufficient conditions for quadratic convergence of predictor-corrector methods using a proximal point correction step.

1 Introduction

In considering the minimization of a nonsmooth function it has often been noted that, in general, the minimum will occur at a point of nondifferentiability. It has also been noted however, that nonsmoothness seldom occurs in a random manner, but instead often has an underlining structure which can be exploited in optimization [BM88], [Bur90], [Wri93], [MS99], [MS02], [Sha03], [Lew03], [HL04], [Har04b]. This underlining structure often appears to take the form of a manifold along which the function appears smooth, but away from which the function appears nonsmooth. If the minima lie on such a manifold we refer to this manifold as the *active manifold*. Many researchers have developed algorithms which force iterates onto the active manifold to ensure rapid convergence [BM88], [Bur90], [AKK91], [MM05], [MS05].

These algorithms, which can be called *predictor-corrector methods*, follow a common two-step form. Supposing that we have an iterate which lies on the active manifold, we use the smoothness of the function along with manifold to take a prediction step in a direction tangent to the active manifold. Since, in general, this operation results in a point outside the manifold, it is followed by a corrector step with returns the iterate to the active manifold. Algorithms of this form can be found in [LOS00], [Ous99], [MS05]. As explained in [MM05], these methods can be seen as concrete versions of the intrinsic Riemannian Newton method (see [Gab82] and [Smi94] among others).

For predictor-corrector algorithms to work efficiently, it is essential to understand the tools which the algorithm uses. During the predictor step, these algorithms use the smoothness of the function along the active manifold to predict directions of decrease. This leads to the need for gradient and Hessian like structures for the function restricted to the active manifold. The primary focus of this work is to explore an extrinsic method for defining the so-called Riemann gradient. This extrinsic definition can be found in Section 2, which considers the Riemann gradient under the most basic assumption for its existence: that the function be smooth along a manifold. In this setting the Riemann gradient is shown to be a well defined object, and several formulas for its computation are developed.

In order to provide more structure to the function, in Section 3, we introduce the notion of "partly smooth" function. Originally developed in [Lew03], partly smooth functions are functions which are not just smooth along a manifold, but also satisfy some regularity and sharpness conditions (see Definition 14). These conditions create a stronger relationship between the function and its Riemann gradient. Of particular interest is how this relationship behaves under the assumption that the Riemann gradient lies within the relative interior of the subdifferential. In Subsection 3.3 we show that if this assumption holds at a point, then it must hold locally, and the direction of steepest descent along the manifold is the direction of steepest descent in the entire space.

As a secondary focus for this work, we consider the correction step of predictor-corrector methods. In examining the correction step, it is clear that returning to the manifold can be accomplished in many different ways. Theoretically one could simply project the prediction step onto the manifold [MM05]. This approach however has problems in practice, for either the active manifold or the projection onto it is not known. Recent work by Mifflin and Sagastizábal has suggested that another manner of returning to the active manifold is by the use of proximal points [MS02]. In [MS02] they show that under certain conditions the proximal point mapping identifies the active manifold for a convex function. This result has been extended to *prox-regular* functions in [MS04].

Prox-regularity, originally studied in [PR96a], is a generalization of convexity which provides the necessary structure for the proximal point mapping to be single valued [PR96b]. In Section 4 we consider the Riemann gradient and the proximal point mapping for prox-regular partly smooth functions. We begin by providing an alternate proof for the theorems of [MS04], which yields a slightly stronger result. Our approach not only shows that proximal points identify active manifolds, but describes the smoothness of the proximal point mapping in this setting. Subsection 4.2 then provides a theoretical comparison of the proximal point method and the projection method for returning to the active manifold. We show that, in general, the proximal point method takes a larger step and causes greater function decrease than the projection method, thus reinforcing the idea that the proximal point method is a more effective manner to return to the active manifold. The study finishes with the statement of the quadratic convergence of the conceptual form of the proximal algorithm of [MS05].

1.1 Notation and preliminaries

We begin with outlining the notation used throughout this work.

(a) Notions from differential and Riemannian geometry A subset $\mathcal{M}$ of $\mathbb{R}^n$ is said to be a *p-dimensional C^k submanifold* of $\mathbb{R}^n$ around $x \in \mathcal{M}$ ($1 \leq k \leq +\infty$) if there exists *local parameterization* of $\mathcal{M}$ around x, that is, a C^k function $\varphi : \mathbb{R}^p \to \mathbb{R}^n$ such that φ realizes a local homeomorphism between a neighborhood of $0 \in \mathbb{R}^p$ and a neighborhood of $x \in \mathcal{M}$ and the derivative of φ at $\varphi^{-1}(x) = 0$ is injective. A p-dimensional C^k submanifold $\mathcal{M}$ of $\mathbb{R}^n$ can alternatively be defined via a *local equation*, that is, a C^k function $\Phi : \mathbb{R}^n \to \mathbb{R}^{n-p}$ with a surjective derivative at $\bar{x} \in \mathcal{M}$, that satisfies for all x close enough to $\bar{x}$

$$x \in \mathcal{M} \quad \Longleftrightarrow \quad \Phi(x) = 0.$$

To lighten notation, henceforth we shall write "C^k manifold" instead of "p-dimensional C^k submanifold of $\mathbb{R}^n$". We shall also omit the "C^k" whenever the level of smoothness of a manifold or a function is

irrelevant. In this case, we shall say "smooth" to express that a function is of class C^k where k will be sufficiently large for our purposes.

Given a point $x \in \mathcal{M}$, we denote respectively by $\mathrm{T}_{\mathcal{M}}(x)$ and $\mathrm{N}_{\mathcal{M}}(x)$ the *tangent space* (of dimension p) and the *normal space* (of dimension $n-p$) of $\mathcal{M}$ at x, defined through either a local parameterization φ or a local equation Φ as follows:

$$\mathrm{T}_{\mathcal{M}}(x) = \operatorname{Im} \mathrm{D}\varphi(0) = \ker \mathrm{D}\Phi(x) \quad \text{and} \quad \mathrm{N}_{\mathcal{M}}(x) = \ker \mathrm{D}\varphi(0)^* = \operatorname{Im} \mathrm{D}\Phi(x)^* \tag{1}$$

(where A^* denotes the adjoint operator of A). For a C^k manifold $\mathcal{M}$, the *tangent bundle* $\mathrm{T}\mathcal{M}$ is the C^{k-1} manifold of $\mathbb{R}^{2n}$ of dimension $2p$ defined by

$$\mathrm{T}\mathcal{M} = \bigcup_{x \in \mathcal{M}} (x, \mathrm{T}_{\mathcal{M}}(x)).$$

Let us point out two local parameterizations that we will consider in this paper. Given a C^k manifold $\mathcal{M}$ ($k \geq 2$) and $x \in \mathcal{M}$, the function defined for $u \in \mathrm{T}_{\mathcal{M}}(x)$ (sufficiently small) by

$$\varphi_x^{\mathrm{proj}}(u) = \mathrm{P}_{\mathcal{M}}(x+u) = \operatorname{argmin} \{\delta_{\mathcal{M}}(y) + \|x+u-y\|^2\}$$

is the *projection parameterization* of $\mathcal{M}$ at x [MM05, Lemma 4.8], where $\mathrm{P}_{\mathcal{M}}$ denotes the projection onto $\mathcal{M}$ and $\delta_{\mathcal{M}}$ denotes the indicator function for the manifold $\mathcal{M}$. Similarly, the function defined for small $u \in \mathrm{T}_{\mathcal{M}}(x)$ by

$$\varphi_x^{\tan}(u) = \operatorname{argmin} \{\delta_{\mathcal{M}}(y) + \delta_{N_{\mathcal{M}}(x)}(x+u-y)\}$$

is *the tangential parameterization* of $\mathcal{M}$ at x (see [Ous99, Theorem 3.4] or [MM05, Corollary 2.3]). Both parameterizations are locally well defined (as single-valued functions). Let us also note that the local inverse of $\varphi_x^{\tan}$ is the projection onto $\mathrm{T}_{\mathcal{M}}(x)$, that is,

$$z = \varphi_x^{\tan}(u) \Longleftrightarrow u = \mathrm{P}_{\mathrm{T}_{\mathcal{M}}(x)}(z-x). \tag{2}$$

The natural embedding of a submanifold $\mathcal{M}$ into $\mathbb{R}^n$ permits to define a Riemannian structure and to introduce geodesics on $\mathcal{M}$ (see [dC92] for instance). Roughly speaking, a geodesic is locally the shortest path between two points on $\mathcal{M}$. We denote by $\gamma(x,u,t)$ the value at $t \in \mathbb{R}$ of the geodesic starting at $x \in \mathcal{M}$ with velocity $u \in \mathrm{T}_{\mathcal{M}}(x)$ (it is uniquely defined – see [dC92]). For instance, if the manifold is the whole Euclidean space $\mathcal{M} = \mathbb{R}^n$, then the geodesics are the straight lines traversed with a constant speed: $\gamma(x,u,t) = x + tu$ in this case.

(b) Notions from variational analysis We mainly follow the notation of [RW98]. We define the *regular* (or *Fréchet*) *subdifferential* of f at $\bar{x}$ as

$$\hat{\partial} f(\bar{x}) = \{p \in \mathbb{R}^n : \quad f(x) \geq f(\bar{x}) + \langle p, x-\bar{x}\rangle + o(\|x-\bar{x}\|)\},$$

and the *limiting subdifferential* as

$$\partial f(\bar{x}) = \limsup_{x \to \bar{x}, f(x) \to f(\bar{x})} \hat{\partial} f(x), \tag{3}$$

where lim sup is the set upper limit in the sense of Kuratowski.

If the function f is *(Clarke) regular* (or *subdifferential regular*) at $\bar{x}$, then the regular subdifferential at $\bar{x}$ coincides with the limiting subdifferential (and with the Clarke subdifferential, see [RW98]). For instance, the smooth functions, the convex functions and the indicator functions of manifolds are regular functions [RW98, Chapter 6.8].

For any set S we define the *affine span*, $\operatorname{aff} S$, to be the smallest affine space which contains S. The *relative interior* $\operatorname{ri} S$ of S, is then the interior of S relative to the space described by its affine span.

A lower semicontinuous function $f : \mathbb{R}^n \to \mathbb{R}$, is called *prox-bounded* if for some point x and some parameter $\lambda > 0$ the function $f(y) + \frac{1}{2\lambda}\|x - y\|^2$ is bounded below. In this case we define the *proximal mapping* $x \mapsto P_\lambda(x)$ by

$$P_\lambda(x) = \operatorname*{argmin}_{y \in \mathbb{R}^n} \left\{ f(y) + \frac{1}{2\lambda}\|x - y\|^2 \right\}. \tag{4}$$

The points of the set $P_\lambda(x)$ are called *proximal points* of f at x, and the parameter λ the *prox-parameter*. It is known (see [RW98], for example) that for any convex function (with any $\lambda > 0$), or for any "prox-regular" function (with λ sufficiently small, cf. Definition 26), one has $P_\lambda(x) = x$ if and only if x is a critical point of f. Henceforth when dealing with prox-bounded functions we shall always assume that the prox-parameter is sufficiently small for the proximal mapping to be well defined.

2 Riemannian gradient

Throughout this paper we will repeatedly make the assumption that a function f has a *smooth restriction* on a submanifold $\mathcal{M}$ near a point $\bar{x} \in \mathcal{M}$ (or, simply worded, that f is *smooth along* $\mathcal{M}$ near $\bar{x}$). This smoothness can be defined intrinsically by expressing $\mathcal{M}$ using its local parameterizations ([dC92]). Taking advantage of the natural embedding of $\mathcal{M}$ into $\mathbb{R}^n$, it can also be characterized by the existence of a *smooth representation* $\widetilde{f} : \mathbb{R}^n \to \mathbb{R}$, that is a smooth function $\widetilde{f}$ such that $\widetilde{f}(x) = f(x)$ for all points on $\mathcal{M}$. From the outlook of this work, we find it advantageous to use the extrinsic approach.

2.1 Examples

To provide some insight into the abundance and interest of functions smooth along a manifold, we begin by providing some simple examples of nonsmooth functions which have smooth restrictions on a given manifold.

Example 1 (Finite Max Functions). Suppose the function f is defined as the maximum of a finite number of C^k functions:

$$f(x) = \max_{i=1,2,\ldots,N} f_i(x), \qquad f_i \in C^k \quad \text{for } i = 1, 2, \ldots, N.$$

For each point $\bar{x}$ we define its *active set* by $A(\bar{x}) = \{i : f(\bar{x}) = f_i(\bar{x})\}$. Then assuming that the *active gradients* $\{\nabla f_i(\bar{x}) : i \in A(\bar{x})\}$ are linearly independent, it follows that f is C^k along the manifold

$$\mathcal{M} = \{x : A(x) = A(\bar{x})\},$$

that is, the restriction of f to $\mathcal{M}$ is C^k. Moreover, for any $i \in A(\bar{x})$, the function f_i is a C^k representation of f on $\mathcal{M}$. ∎

Example 1 shows us that functions which have smooth representations exist in abundance. Many researchers have examined methods of extending the idea of finite max functions to an even broader class of functions. For example, [Roc82] consider the class of "lower-C^2" functions, in which an infinite index set is permitted provided a certain constraint qualification condition is preserved. In [MS00a], [MS00b], the idea of a "primal-dual gradient" structure (`pdg` structures) is developed. Originally these structures were a method of defining functions along a manifold via finite max functions in a manner which ensures the possibility to reconstruct subgradient information. However, the definition has expanded over research to include the ability to define the functions along a manifold in other manners.

Our second example explains some of the interest in functions with smooth representations. Indeed, it shows that the highly studied maximum eigenvalue function has a smooth representation. Moreover, compositions of the maximum eigenvalue function with a smooth function, also have a smooth representation provided a constraint qualification is satisfied. The example is well-known, and details can be found in [Arn71], [Ous99] amongst others.

Example 2 ($\lambda_{\max} \circ \mathcal{F}$). Let S_m the space of symmetric $m \times m$ matrices, and r a positive integer. We denote the eigenvalues of a symmetric matrix $X \in S_m$ by $\lambda_1(X) \geq \lambda_2(X) \geq \ldots \geq \lambda_m(X)$. Then the set of symmetric matrices whose maximum eigenvalue has multiplicity r,

$$\mathcal{M}_r = \{X \in S_m,\ \lambda_1(X) = \cdots = \lambda_r(X) > \lambda_{r+1}(X)\},$$

is a C^∞ submanifold of S_m. Moreover, for any matrix $\bar{X} \in \mathcal{M}_r$ the maximum eigenvalue function $\lambda_{\max}$ has a C^∞ representation on $\mathcal{M}_r$ near $\bar{X}$ which is

$$\widetilde{\lambda}(X) = \frac{1}{r}(\lambda_1(X) + \cdots + \lambda_r(X)).$$

Furthermore, consider the composition $f = \lambda_{\max} \circ \mathcal{F}$, where the function $\mathcal{F} : \mathbb{R}^n \to S_m$ is C^k. At any point $\bar{x}$ where the *transversality condition*

$$\mathrm{N}_{\mathcal{M}r}(\mathcal{F}(x)) \bigcap \ker \mathrm{D}\mathcal{F}(x)^* = \{0\} \tag{5}$$

holds, the set $\mathcal{N}_r = \mathcal{F}^{-1}(\mathcal{M}_r)$ is a submanifold of $\mathbb{R}^n$ around $\bar{x}$, and $\widetilde{f} = \widetilde{\lambda} \circ \mathcal{F}$ is a C^k representation of $\lambda_{\max} \circ \mathcal{F}$ on $\mathcal{N}_r$ near x. ∎

In our final example we explore the idea of "$\mathcal{VU}$-decompositions" and "fast tracks" [LOS00], [MS99], [MS00b], [MS02], (amongst others). The example demonstrates a method of determining potential manifolds along which a convex function might have a smooth representation. Although somewhat theoretical in appearance, the ideas have lead to a new "$\mathcal{VU}$-proximal point algorithm" which has shown some success in practice. This algorithm is discussed in more detail in Section 4.1 of this work.

Example 3 ($\mathcal{VU}$-theory and fast tracks). Consider a convex function f and a point $\bar{x}$. In examining the subdifferential $\partial f(\bar{x})$, one might notice that the "nonsmoothness" of f at $\bar{x}$ is essentially contained in directions parallel to the subdifferential. This led to the idea of *$\mathcal{VU}$-decomposition* at $\bar{x}$, in which the space $\mathbb{R}^n$ is decomposed into two orthogonal subspaces

$$\mathcal{V}(\bar{x}) = \operatorname{par} \partial f(\bar{x}) \qquad \text{and} \qquad \mathcal{U}(\bar{x}) = \mathcal{V}(\bar{x})^\perp,$$

where $\operatorname{par} \partial f(\bar{x})$ denotes the linear space parallel to the affine space generated by the subdifferential $\partial f(\bar{x})$. These spaces represent the directions from $\bar{x}$ for which f behaves nonsmoothly ($\mathcal{V}$) and smoothly ($\mathcal{U}$) (see [LOS00]). The goal is then to find a smooth function which describes f in the directions of $\mathcal{U}$.

Following the notation of [MS00b], [MS05], we consider $\bar{V}$ and $\bar{U}$ to be basis of $\mathcal{V}(\bar{x})$ and $\mathcal{U}(\bar{x})$. Suppose $\dim \mathcal{V}(\bar{x}) = n - p$, thus $\dim \mathcal{U}(\bar{x}) = p$. For any $g \in \operatorname{ri} \partial f(\bar{x})$ the *$\mathcal{U}$-Lagrangian* is defined by

$$L_{\mathcal{U}}(\,\cdot\,, g)\colon \begin{cases} \mathbb{R}^p \longrightarrow \mathbb{R} \\ u \longmapsto \min\limits_{v \in \mathbb{R}^{n-p}} \{f(\bar{x} + \bar{U}u + \bar{V}v) - g^\top \bar{V} v\} \end{cases} \tag{6}$$

We denote by $W(u, g)$ the points where the minimum is attained. If for all $g \in \operatorname{ri} \partial f(\bar{x})$ there exists a C^2 selection $v : \mathbb{R}^p \to \mathbb{R}^{n-p}$ such that $v(u) \in W(u, g)$ and if the $\mathcal{U}$-Lagrangian $L_{\mathcal{U}}(u, g)$ is C^2 with respect to u, then $\bar{\mathcal{M}} = \{\bar{x} + (\bar{U}\, u + \bar{V}\, v(u)) : u \in \mathbb{R}^p\}$ is a manifold and f is C^2 on $\bar{\mathcal{M}}$ [MS00b]. Such a function $(u \mapsto v(u))$ is called a *fast track* of f and studied extensively in [MS00b] and [MS02].

Note that $\operatorname{ri} \partial f(\bar{x})$ is nonempty whenever $\partial f(\bar{x}) \neq \emptyset$. Moreover, if f is nonsmooth, then $\operatorname{ri} \partial f(\bar{x})$ cannot be a singleton (in fact $\operatorname{ri} \partial f(\bar{x})$ is a singleton if, and only if, $\operatorname{ri} \partial f(\bar{x}) = \{g\} = \partial f(\bar{x})$, which implies $\mathcal{V}(\bar{x}) = 0$ and $p = n$), containing thus an infinity of points. ∎

It is worth noting here that the research on `pdg` structures (mentioned earlier) sprang largely from the search for concrete examples of functions which contain fast tracks.

2.2 The Riemannian gradient of f

Given a function f which has a smooth representation on the manifold $\mathcal{M}$ (near the point $\bar{x}$), it is natural to ask how the gradient of the representation $\widetilde{f}$ relates to the original function f. To do so, we require the following proposition from [Lew03, Proposition 2.2].

Proposition 4 (Normal space and subdifferential). *Let $\mathcal{M}$ be a submanifold of $\mathbb{R}^n$ and $f : \mathbb{R}^n \to \mathbb{R}$ be a function with a smooth restriction on $\mathcal{M}$, and x be a point of $\mathcal{M}$. Then*

$$\mathrm{N}_{\mathcal{M}}(x) \supset \hat{\partial} f(x) - \nabla \widetilde{f}(x). \tag{7}$$

It is easily seen that (7) yields in particular

$$\mathrm{N}_{\mathcal{M}}(x) \supset \operatorname{par} \hat{\partial} f(x). \tag{8}$$

Let us now provide two examples related to Proposition 4. The first example shows that the inclusions in Equation (7) can be strict, while the second shows that the regular subdifferential $\hat{\partial} f(x)$ cannot be replaced by the limiting subdifferential $\partial f(x)$. In Section 3, we will introduce partial smoothness, which provides enough structure to avoid these complications.

Example 5 (Strict inclusion). Let f be a smooth function on $\mathbb{R}^n$ (thus $f = \widetilde{f}$) and $\mathcal{M}$ be a strict vector subspace of $\mathbb{R}^n$. The subspace parallel to $\hat{\partial} f(x) = \{\nabla f(x)\}$ is the singleton $\{0\}$, while $\mathrm{N}_{\mathcal{M}}(x)$ is the nontrivial subspace normal to $\mathcal{M}$. Thus the inclusion in (7) can be strict. ∎

Example 6 (Necessity of regular subdifferential). Consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ defined by

$$f(x,y) = \begin{cases} -|y|, & \text{if } y \leq 0 \\ \min\left\{(4/\pi)\,\arctan(y/x), \sqrt{x^2+y^2}\right\}, & \text{if } x \geq y > 0 \\ \min\left\{\sqrt{(\max\{x,0\})^2 + y^2}, \sqrt{2}\right\}, & \text{if } y \geq \max\{x,0\} \end{cases}$$

Notice f is continuous and that it is constant (equal to 0) along the manifold $\mathcal{M} = \mathbb{R} \times \{0\}$. However, for the limiting subdifferential $\partial f(0,0)$ (cf. (3)) we have

$$\operatorname{par} \partial f(0,0) \not\subset \{0\} \times \mathbb{R} = \mathrm{N}_{\mathcal{M}}(0,0).$$

To see this, observe first that, for any $k > 0$ the function f is differentiable at $(0, -1/k)$, with derivative $\nabla f(0, -1/k) = (0, -1)$. Thus $(0,-1) \in \partial f(0,0)$. Consider now the sequence $(x_k, y_k) = \frac{1}{2k}(\cos\frac{1}{k}, \sin\frac{1}{k})$, for $k \in \mathbb{N}$. The function f is also differentiable at (x_k, y_k), and as $k \to +\infty$ we get $(x_k, y_k) \to (0,0)$ and

$$\nabla f(x_k, y_k) = \left(\frac{x_k}{\sqrt{{x_k}^2 + {y_k}^2}}, \frac{y_k}{\sqrt{{x_k}^2 + {y_k}^2}}\right) = \left(\cos\frac{1}{k}, \sin\frac{1}{k}\right) \longrightarrow (1,0).$$

Thus $\{(0,-1),(1,0)\} \in \partial f(0,0)$ and therefore $\operatorname{par} \partial f(0,0) \not\subset \{0\} \times \mathbb{R}$. ∎

For a function f which is smooth along a manifold $\mathcal{M}$, there is no reason to expect the representation function $\widetilde{f}$ to be unique. In the next proposition we see that for any vector $g \in \hat{\partial} f(x) + \mathrm{N}_{\mathcal{M}}(x)$ a smooth representation $\widetilde{f}$ of f can be found satisfying $\nabla \widetilde{f}(x) = g$. Thus if either $\hat{\partial} f$ or $\mathrm{N}_{\mathcal{M}}$ is nontrivial, there are an infinite number of distinct representation functions.

Proposition 7 (Abundance of $\nabla\widetilde{f}(x)$). *Let $\mathcal{M}$ be a smooth manifold, $\bar{x} \in \mathcal{M}$ and $p \in \mathrm{N}_{\mathcal{M}}(\bar{x})$. Then there exists a smooth representation h of the null function on $\mathcal{M}$ near $\bar{x}$ satisfying $\nabla h(\bar{x}) = p$. Thus, if $f : \mathbb{R}^n \to \mathbb{R}$ is a function having a smooth representation on $\mathcal{M}$ near $\bar{x} \in \mathcal{M}$, then for every $g \in \hat{\partial} f(\bar{x}) + \mathrm{N}_{\mathcal{M}}(\bar{x})$, there exists a smooth representative $\widetilde{f}$ of f on $\mathcal{M}$ near $\bar{x}$ with $\nabla\widetilde{f}(\bar{x}) = g$.*

Proof. Let $\Phi : \mathbb{R}^n \to \mathbb{R}^{n-p}$ be a local equation of the manifold $\mathcal{M}$ (of dimension p) around the point $\bar{x}$. Since $\mathrm{D}\Phi(\bar{x})$ is surjective, the gradients of the components φ_i of Φ, $\{\nabla\varphi_i(\bar{x})\}_i$, form a basis of $\mathrm{N}_{\mathcal{M}}(\bar{x})$. Thus for every $p \in \mathrm{N}_{\mathcal{M}}(\bar{x})$, there exist $\{a_i\}_i \subset \mathbb{R}$ such that $p = \sum_{i=1}^{n-p} a_i \nabla\varphi_i(\bar{x})$. The function $h = \sum_{i=1}^{n-p} a_i\varphi_i$ is a smooth representation of the null function on $\mathcal{M}$ having the prescribed derivative.

To see the second part of the proof, consider any smooth representation $\widetilde{f}$ of f on $\mathcal{M}$ near $\bar{x}$. For a given $g \in \hat{\partial} f(\bar{x}) + \mathrm{N}_{\mathcal{M}}(\bar{x})$, we have $p = g - \nabla\widetilde{f}(\bar{x}) \in \mathrm{N}_{\mathcal{M}}(\bar{x})$, by Proposition 4. Applying the first part, we obtain a smooth representation h of the null function on $\mathcal{M}$ with $\nabla h(\bar{x}) = p$. Thus $\widetilde{f} + h$ is a smooth representation of f on $\mathcal{M}$ with the desired derivative. ∎

Even if the gradient of the representation function $\widetilde{f}$ can take any value, it is still somewhat controlled by the fact that $\widetilde{f}$ agrees with the original function on the manifold. This means that in directions tangent to the manifold $\nabla\widetilde{f}$ must behave like a subgradient of f. Indeed Proposition 4 yields that, given any subgradient $g \in \hat{\partial} f(\bar{x})$, we have

$$\langle \nabla\widetilde{f}(\bar{x}), u\rangle = \langle g, u\rangle, \quad \text{for all } u \in \mathrm{T}_{\mathcal{M}}(x).$$

This leads to the key definition of this work.

Definition 8 (Riemannian gradient). *Suppose f is smooth along the manifold $\mathcal{M}$ at the point $x \in \mathcal{M}$. Let $\widetilde{f}$ be any smooth representation of f on $\mathcal{M}$ near x. Then the* Riemannian gradient *of f at x relative to $\mathcal{M}$ is defined by*

$$\nabla_{\mathcal{M}} f(x) = \mathrm{P}_{\mathrm{T}_{\mathcal{M}}(x)}(\nabla\widetilde{f}(x)).$$

Proposition 9. *The Riemannian gradient of f at x does not depend on the smooth representation $\widetilde{f}$ of f at x.*

Proof. Let f be smooth along $\mathcal{M}$ at the point $\bar{x} \in \mathcal{M}$. Let $\widetilde{f}_1$ and $\widetilde{f}_2$ be two smooth representations of f. Then $\widetilde{f}_1 + \delta_{\mathcal{M}} = \widetilde{f}_2 + \delta_{\mathcal{M}}$, where $\delta_{\mathcal{M}}$ is the indicator function of $\mathcal{M}$. By [RW98, Corollary 1.9] we have for $i = 1, 2$

$$\partial(\widetilde{f}_i + \delta_{\mathcal{M}})(x) = \nabla\widetilde{f}_i(x) + \mathrm{N}_{\mathcal{M}}(x).$$

It follows

$$\nabla\widetilde{f}_1(x) - \nabla\widetilde{f}_2(x) \in \mathrm{N}_{\mathcal{M}}(x),$$

that is, $\mathrm{P}_{\mathrm{T}_{\mathcal{M}}(x)}(\nabla\widetilde{f}_1(x) - \nabla\widetilde{f}_2(x)) = 0$. ∎

Remark 10. (i) intrinsic definition. Similar to the smoothness of f, the gradient of f along the manifold $\mathcal{M}$ can either be defined intrinsicly, via a local parameterization of $\mathcal{M}$, or extrinsicly, via a local smooth representative $\widetilde{f}$. Despite the nonuniqueness of the representative function $\widetilde{f}$, this latter definition will be more helpful for our purposes.

(ii) gradients in differential geometry. In differential geometry there is a conceptual difference between tangent vectors and gradients of functions. Tangent vectors are (classes of equivalences of) derivatives of curves passing through x (identify to $\mathrm{T}_{\mathcal{M}}(x)$), while gradients of functions are one-forms and belong to the co-tangent (or dual) space $\mathrm{T}_{\mathcal{M}}(x)^*$. Using the natural identification between $\mathrm{T}_{\mathcal{M}}(\bar{x})$ and $\mathrm{T}_{\mathcal{M}}(\bar{x})^*$ allowed by the Riemannian structure, we indistinguishly mix up gradients and tangent vectors.

(iii) projected gradient. The Riemann gradient bares a strong relationship to the object commonly referred to as the "projected gradient" (see [CM87] for example). The major difference is that the

projected gradient is constructed when examining a smooth function over a constraint set, while the Riemann gradient considers a function f with smooth restriction on some manifold $\mathcal{M}$ and a smooth representation $\widetilde{f}$ of f in the ambient space. By viewing the manifold as a constraint set, and by replacing the objective function f by $\widetilde{f}$, the Riemannian gradient could be viewed as the projected gradient of the representation function $\widetilde{f}$. ■

Let us give a simple argument showing the smoothness of $\nabla_{\mathcal{M}} f$ in relation to the smoothness of f and $\mathcal{M}$.

Lemma 11 (Smoothness of $\nabla_{\mathcal{M}} f$). *If the function f is C^k along the C^k-manifold $\mathcal{M}$ ($k \geq 1$), then the function $x \mapsto \nabla_{\mathcal{M}} f(x)$ is of class C^{k-1} on $\mathcal{M}$.*

Proof. Since $\nabla \widetilde{f}(\cdot)$ is of class C^{k-1}, we just have to justify that $x \mapsto \mathrm{P}_{\mathrm{T}_{\mathcal{M}}(x)}$ is of class C^{k-1}. Let $\bar{x} \in \mathcal{M}$, and φ a C^k local parameterization of $\mathcal{M}$ around $\bar{x}$. For x close to $\bar{x}$, the columns of a matrix representing $\mathrm{D}\varphi(\varphi^{-1}(x))$ form a basis of $\mathrm{T}_{\mathcal{M}}(x)$ which has a C^{k-1} dependence on x. Using these basis to express $\mathrm{P}_{\mathrm{T}_{\mathcal{M}}(x)}$, we get its desired smoothness, and the one of $\nabla_{\mathcal{M}} f(\cdot)$ follows. ■

Although we used an explicit representation to define the Riemannian gradient, the Riemannian gradient is unique and intrinsically defined (Proposition 9). Let us also show that $\widetilde{f}$ can be omitted when defining $\nabla_{\mathcal{M}} f$ in case that $\hat{\partial} f(x) \neq \emptyset$.

Proposition 12 (Riemannian gradient and subdifferential). *Suppose the function f is smooth along the manifold $\mathcal{M}$ near $x \in \mathcal{M}$. If the regular subdifferential $\hat{\partial} f(x)$ is non-empty, then*

$$\nabla_{\mathcal{M}} f(x) = \mathrm{P}_{\mathrm{T}_{\mathcal{M}}(x)}(\hat{\partial} f(x)).$$

Proof. Let $\widetilde{f}$ be a representation of f so that

$$\nabla_{\mathcal{M}} f(\bar{x}) = \mathrm{P}_{\mathrm{T}_{\mathcal{M}}(x)}(\nabla \widetilde{f}(\bar{x})).$$

Inclusion (7) implies

$$\mathrm{P}_{\mathrm{T}_{\mathcal{M}}(x)}(\hat{\partial} f(x)) \subset \mathrm{P}_{\mathrm{T}_{\mathcal{M}}(x)}(\nabla \widetilde{f}(x)) = \{\nabla_{\mathcal{M}} f(x)\}.$$

Thus $\mathrm{P}_{\mathrm{T}_{\mathcal{M}}(x)}(\hat{\partial} f(x))$ can be either empty or singleton. Since $\hat{\partial} f(x) \neq \emptyset$ and the projection exists, the proof is complete. ■

2.3 Algorithms using the smoothness along $\mathcal{M}$

The smoothness of f along a manifold can be exploited to design optimization methods to solve

$$\begin{cases} \min & f(x) \\ x \in \mathcal{M} \end{cases}$$

These methods follow a two-step process which resembles to a predictor-corrector process: the next iterate is computed by

1. (Predictor) a step in the tangent space,
2. (Corrector) a step to regain the manifold $\mathcal{M}$.

A very important method following this pattern is the Riemannian Newton method (see [Gab82], [Smi94], [AES99], [DMP03] among others). To compute the predictor step, this method uses the Riemannian gradient and also the so-called Riemannian Hessian that can be defined via geodesics as follows. For $(x, u) \in \mathrm{T}\mathcal{M}$, the value of the Riemannian Hessian at x along (u, u) is set as

$$\nabla^2_{\mathcal{M}} f(x)(u, u) = \frac{\mathrm{d}^2}{\mathrm{d}t^2} f(\gamma(x, u, t))\Big|_{t=0}.$$

One iteration of the Riemannian Newton method is then

1. compute the direction $u = -[\nabla^2_{\mathcal{M}} f(x)]^{-1} \nabla_{\mathcal{M}} f(x)$ in the tangent space,

2. compute the step $x_+ = \gamma(x, u, 1)$ on $\mathcal{M}$.

If $\mathcal{M} = \mathbb{R}^n$, this appears to be exactly the classical Newton method: the next iterate is $x_+ = \gamma(x, u, 1) = x + u$ with Newton direction $u = -[\nabla f(x)]^{-1} \nabla f(x)$ and step $t = 1$. The Riemannian Newton method is shown to be quadratically convergent under classical assumptions (see references above). In [MM05], this method is generalized in using the local parameterizations $\varphi^{\tan}$ and φ^{proj} to stay on the manifold (corrector step). For instance, with the projection parameterization, the next iterate is computed in processing

1. make a Newton step in the tangent space

$$\widetilde{x}_+ = x - [\nabla^2_{\mathcal{M}} f(x)]^{-1} \nabla_{\mathcal{M}} f(x);$$

2. correct it in projecting onto $\mathcal{M}$

$$x_+ = \mathrm{P}_{\mathcal{M}}(\widetilde{x}_+).$$

Introducing in an appropriate way the parameterization permits to emphasize that the $\mathcal{U}$-Newton methods developed in [LOS00] and [Ous99] follow the two-step process too. Since the two parameterizations coincide with geodesics up to second-order, the quadratic convergence is also maintained for the corresponding algorithms (see [MM05] for details).

We will see in Subsection 4.2 that the $\mathcal{VU}$-proximal algorithm of [MS05] follow the same pattern too, and that it is also quadratically convergent. For the moment, we just note that a step of the basic proximal algorithm also follow this two-step process: usually expressed as "an implicit subgradient step", it can be interpreted here as an "implicit Riemannian gradient step".

Proposition 13 (An interpretation of the proximal step). *Let f be Clarke regular and smooth along $\mathcal{M}$ near the point $x \in \mathcal{M}$. If for a given prox-parameter $\lambda > 0$ there exists a proximal point $y \in P_\lambda(x)$ which belongs to $\mathcal{M}$, then*

(i) $\lambda \nabla_{\mathcal{M}} f(y) = \mathrm{P}_{\mathrm{T}_{\mathcal{M}}(y)}(x - y)$, *or equivalently*

(ii) $x = \varphi^{\tan}_y(\lambda \nabla_{\mathcal{M}} f(y))$, *which means that we recover x by computing the gradient step in the tangent space followed by the correction step provided by $\varphi^{\tan}$ to stay on the manifold.*

Proof. By definition, $y \in P_\lambda(x)$ satisfies

$$0 \in \hat{\partial} \left(f + \frac{1}{2\lambda} \| \cdot - x \|^2 \right)(y) = \hat{\partial} f(y) + \frac{1}{\lambda}(y - x).$$

This equation follows from [RW98, Corollary 10.9] and the smoothness of $y \mapsto \frac{1}{2\lambda}\|y - x\|^2$. Proposition 4 yields $(x - y) \in \lambda \nabla \widetilde{f}(y) + \mathrm{N}_{\mathcal{M}}(y)$. Taking the projection onto $\mathrm{T}_{\mathcal{M}}(y)$ completes the proof of (i).

To get (ii), recall that $\mathrm{P}_{\mathrm{T}_{\mathcal{M}}(y)}$ is the inverse of the parameterization $\varphi^{\tan}_y$ (see (2)): we can write

$$x = \varphi^{\tan}_y(\mathrm{P}_{\mathrm{T}_{\mathcal{M}}(y)}(x - y))$$

and we use (i) to conclude. ∎

In Section 4 we will see that under the additional assumptions of partial smoothness and prox-regularity (Definition 14 and Definition 26 in this work) even stronger relationships between proximal points and the Riemann gradient exist. For example, under those conditions, the converse of Proposition 13 also holds true (see Theorem 29).

3 Riemann gradient and partial smoothness

3.1 Definition and examples

The notion of partly smooth functions is introduced in [Lew03]. This concept expresses a certain regularity on the underlying smooth structure of a nonsmooth function.

Definition 14 (partial smoothness). *A function $f : \mathbb{R}^n \to \mathbb{R}$ is called* C^k-partly smooth *relative to a nonempty subset $\mathcal{M} \subset \mathbb{R}^n$ ($k \in \mathbb{N} \cup \{\infty\}$) at a point $\bar{x} \in \mathcal{M}$, if $\mathcal{M}$ is a C^k-manifold and the following properties hold:*

(i) smoothness: *the restriction of f to $\mathcal{M}$ is a C^k-function;*

(ii) regularity: *f is Clarke regular with $\partial f(y) \neq \emptyset$ at all $y \in \mathcal{M}$ near $\bar{x}$;*

(iii) sharpness: *the affine span of $\partial f(\bar{x})$ is a translate of $\mathrm{N}_{\mathcal{M}}(\bar{x})$;*

(iv) subdifferential continuity: *the set-valued map ∂f restricted to $\mathcal{M}$ is continuous near $\bar{x}$.*

If these conditions hold for all $\bar{x} \in \mathcal{M}$ we say that f is partly smooth *relative to $\mathcal{M}$.*

Examples of partly smooth functions are abundant in optimization. For example, finite max functions of the style in Example 1 are partly smooth along the manifold defined therein [Lew03, Corollary 4.8]. In [Lew03, Example 3.6] it was shown that the maximal eigenvalue function (Example 2 of this work) is partly smooth. Furthermore, [Har04b] and [MM05, Theorem 2.9] have independently shown that in the convex case, the ideas of partial smoothness and "fast tracks" coincide exactly. The next example clarifies this statement.

Example 15 (Fast tracks). Consider a convex function f and a point $\bar{x}$. Suppose f admits a fast track $u \mapsto v(u)$ (see definition and notation in Example 3). Then $\bar{\mathcal{M}} = \{\bar{x} + (\bar{U}u + \bar{V}v(u)) : u \in \mathbb{R}^p\}$ is a manifold and f is partly smooth at $\bar{x}$ relative to $\bar{\mathcal{M}}$. Conversely, if a convex function f is C^2-partly smooth at $\bar{x}$ relative to a C^2-manifold $\mathcal{M}$ and $0 \in \partial f(\bar{x})$, then f admits a fast track. Therefore the $\mathcal{U}$-Lagrangian $L_{\mathcal{U}}(u, g)$ is C^2 with respect to u (for any $g \in \operatorname{ri} \partial f(\bar{x})$). ■

As mentioned, in Lewis's original work on partial smoothness it is shown that the maximum eigenvalue is partly smooth. By applying [Lew03, Theorem 4.2] to this fact, we see that the composition of the maximum eigenvalue function with a smooth function is partly smooth.

Example 16 (Eigenvalue functions). Consider the composition of the maximum eigenvalue function $\lambda_{\max}$ with a function $\mathcal{F}$ smooth at any point $\bar{x}$. It is remarkable that condition (5), ensuring that $\mathcal{N}_r = \mathcal{F}^{-1}(\mathcal{M}_r)$ is a manifold around $\bar{x}$, also guarantees that $\lambda_{\max} \circ \mathcal{F}$ is partly smooth at $\bar{x}$ relative to $\mathcal{N}_r$. ■

Examples of functions which are not partly smooth at a given point are also easily constructed (see [Har04a] or [Har04b] for example).

3.2 Expressions of the Riemannian gradient

We now turn our attention to determining the effect of partial smoothness on descriptions of the Riemann gradient.

Proposition 17 (The gradient as projection of 0). *Let the function f be partly smooth relative to the manifold $\mathcal{M}$ at the point $\bar{x}$. Then for any $x \in \mathcal{M}$ near $\bar{x}$*

$$\nabla_{\mathcal{M}} f(x) = \mathrm{P}_{\mathrm{T}_{\mathcal{M}}(x)}(\partial f(x)) = \mathrm{P}_{\operatorname{aff} \partial f(x)}(0).$$

Proof. The first equality follows from Proposition 12 and condition (ii) of partial smoothness (which yields $\partial f(x) = \hat{\partial} f(x) \neq \emptyset$, for all $x \in \mathcal{M}$). To prove the second equality, consider $\bar{g} = \mathrm{P}_{\mathrm{aff}\,\partial f(x)}(0)$ and any $g \in \partial f(x)$. Notice that,

$$\mathrm{P}_{\mathrm{T}_{\mathcal{M}}(x)}(g) = \mathrm{P}_{\mathrm{T}_{\mathcal{M}}(x)}(g - \bar{g}) + \mathrm{P}_{\mathrm{T}_{\mathcal{M}}(x)}(\bar{g}) = \mathrm{P}_{\mathrm{T}_{\mathcal{M}}(x)}(\bar{g}),$$

since $g - \bar{g} \in \mathrm{par}\,\hat{\partial} f(x) \subset \mathrm{N}_{\mathcal{M}}(x)$ (Proposition 4). Thus it is sufficient to show $\bar{g} \in \mathrm{T}_{\mathcal{M}}(x)$ in order to deduce $\mathrm{P}_{\mathrm{T}_{\mathcal{M}}(x)}(g) = \bar{g} = \mathrm{P}_{\mathrm{aff}\,\partial f(x)}(0)$ for all $g \in \partial f(x)$. Since $0 - \bar{g} \in N_{\mathrm{aff}\,\partial f(x)}(\bar{g})$ and $N_{\mathrm{aff}\,\partial f(x)}(\bar{g})$ is a subspace, we obtain $\bar{g} \in (\mathrm{aff}\,\partial f(x))^{\perp}$. Applying [Lew03, Proposition 2.10] (which states that the sharpness condition of partial smoothness holds locally) we obtain $\bar{g} \in \mathrm{T}_{\mathcal{M}}(x)$ as required. ∎

The proof of Proposition 17 makes use of the smoothness of the function along the manifold, regularity, normal sharpness, and the fact that normal sharpness holds locally. The necessity of smoothness along the manifold and regularity are obvious, while the necessity of normal sharpness is illustrated by the following simple example. Finally, in order for normal sharpness to hold locally subdifferential continuity is required (see the proof of [Lew03, Proposition 2.10]). Thus all four conditions of partial smoothness are used in Proposition 17.

Example 18 (Necessity of normal sharpness). Consider in $\mathbb{R}^2$, the affine space $\mathcal{M} = \{1\} \times \mathbb{R}$ and the function $f(x, y) = \frac{1}{2} x^2$. It is easy to show that f satisfies conditions (i), (ii) and (iv) of Definition 14 at $\bar{x} = (1, 0)$. To calculate the Riemann gradient we note $\partial f(1, 0) = \{(1, 0)\}$ and thus

$$\nabla_{\mathcal{M}} f(1, 0) = \mathrm{P}_{\mathrm{T}_{\mathcal{M}}(\bar{x})}(\partial f(1, 0)) = (0, 0).$$

However since $\mathrm{aff}\,\partial f(1, 0) = \{(1, 0)\}$, we find

$$\mathrm{P}_{\mathrm{aff}\,\partial f(1,0)}(0, 0) = (1, 0) \neq (0, 0) = \nabla_{\mathcal{M}} f(1, 0).$$

Thus Proposition 17 fails without the normal sharpness. ∎

The next example illustrates Proposition 17 in the language of the $\mathcal{U}$-Lagrangian.

Example 19 ($\mathcal{U}$-Lagrangian). We consider a convex function f, a point $\bar{x}$ and a subgradient $g \in \partial f(\bar{x})$. Using the notation of Example 3, we express explicitly the dependence on x of the $\mathcal{U}$-Lagrangian (see (6)). The function $u \mapsto L_{\mathcal{U}(x)}(u, g)$ is shown in [LOS00, Theorem 3.3] to be differentiable at $u = 0$ with

$$\nabla L_{\mathcal{U}(x)}(0, g) = \mathrm{P}_{\mathcal{U}(x)}(g).$$

The gradient $\nabla L_{\mathcal{U}(x)}(0, g)$ is called the $\mathcal{U}$-gradient of f at x. If f is C^k-partly smooth at $\bar{x}$ relative to the C^k-manifold $\mathcal{M} = \{\bar{x} + \bar{U}u + \bar{V}v(u), u \in \mathbb{R}^p\}$, then Proposition 12 implies that

$$\nabla L_{\mathcal{U}(x)}(0, g) = \mathrm{P}_{\mathrm{T}_{\mathcal{M}}(x)}(g) = \nabla_{\mathcal{M}} f(x).$$

Since $\nabla_{\mathcal{M}} f$ is C^{k-1} as a function over $\mathcal{M}$ (Lemma 11), the $\mathcal{U}$-gradient is also C^{k-1} with respect to $x \in \mathcal{M}$. This completes a previous result in [MM05] which has asserted the continuity of the $\mathcal{U}$-gradient with respect to $x \in \mathcal{M}$. ∎

3.3 Persistence and consequences

In this subsection we consider the situation of the Riemann gradient lying in the relative interior of the subdifferential. We begin by showing that this situation is a persistent one. That is, given a partly smooth function f, if the Riemann gradient at the point $\bar{x}$ lies in the relative interior of the subdifferential at $\bar{x}$, then the Riemann gradient of a point $x \in \mathcal{M}$ near $\bar{x}$ also lie in the relative interior of the subdifferential $\partial f(x)$. For this we need the following lemma (which mainly follows the spirit of [MM05, Theorem 2.12]).

Lemma 20 (Persistence inside the subdifferential). *Suppose the function f is partly smooth relative to the manifold $\mathcal{M}$ at the point $\bar{x} \in \mathcal{M}$. Let g be a continuous selection of* aff $\partial f(\cdot)$ *on $\mathcal{M}$, that is, a continuous function $g : \mathcal{M} \to \mathbb{R}^n$ such that $g(x) \in \operatorname{aff} \partial f(x)$. If $g(\bar{x}) \in \operatorname{ri} \partial f(\bar{x})$, then for any $x \in \mathcal{M}$ near $\bar{x}$, we have $g(x) \in \operatorname{ri} \partial f(x)$.*

Proof. Observe first that the sharpness of f on $\mathcal{M}$ at $\bar{x}$ (partial smoothness assumption (iii)) yields by [Lew03, Proposition 2.10] that $\mathrm{N}_{\mathcal{M}}(x) = \operatorname{span}(\partial f(x) - g(x))$, for all $x \in \mathcal{M}$ close to $\bar{x}$. With the help of a basis of $\mathrm{N}_{\mathcal{M}}(x)$ depending continuously on $x \in \mathcal{M}$, we construct a continuous function $x \mapsto \psi_x$ such that

$$\psi_x : \mathrm{N}_{\mathcal{M}}(x) \longrightarrow \mathbb{R}^{n-p}$$

is a linear bijection between $\mathrm{N}_{\mathcal{M}}(x)$ and $\mathbb{R}^{n-p}$. Consider then the convex-valued multi-function $F : \mathcal{M} \rightrightarrows \mathbb{R}^{n-p}$ defined by

$$F(x) = \psi_x(\partial f(x) - g(x)).$$

Continuity of ∂f (by partial smoothness assumption (iv)), of g (by assumption) and of ψ_x (by construction) yield the continuity of F as a multifunction on $\mathcal{M}$ around $\bar{x}$. Furthermore, observe that

$$g(x) \in \operatorname{ri} \partial f(x) \quad \Longleftrightarrow \quad 0 \in \operatorname{int} F(x).$$

Now, suppose for contradiction that there exists a sequence $\{x_k\}$ of points in $\mathcal{M}$ such that x_k tends to x and $g(x_k) \notin \operatorname{ri} \partial f(x_k)$. Set $F_k = F(x_k)$ so that $0 \notin \operatorname{int} F_k$. We separate now 0 from $\operatorname{int} F_k$: there exist $s_k \in \mathbb{R}^{n-p}$ with $\|s_k\| = 1$ such that

$$\forall k \in \mathbb{N}, \quad \forall y \in F_k, \qquad s_k^\top y \leq 0. \tag{9}$$

Extracting a subsequence if necessary, we can suppose that $s_k \to s$ with $\|s\| = 1$. Since $0 \in \operatorname{int} F(\bar{x})$, let $r > 0$ be such that $B(0, r) \subset F(\bar{x})$. Let $v \in B(0, r)$; the continuity of F implies that there are $v_k \in F_k$ such that $v_k \to v$. With (9), we can write $s_k^\top v_k \leq 0$, for all $k \in \mathbb{N}$. Passing to the limit, this gives $s^\top v \leq 0$. This can be done for any $v \in B(0, r)$, so we have $s^\top v = 0$ for all $v \in B(0, r)$. We conclude that $s = 0$, which contradicts $\|s\| = 1$. ∎

As an immediate corollary to Lemma 20, we obtain a more precise version of Proposition 17.

Corollary 21 (Steepest Descent). *Suppose the function f is partly smooth relative to the manifold $\mathcal{M}$ at the point $\bar{x} \in \mathcal{M}$ such that the Riemann gradient $\nabla_{\mathcal{M}} f(\bar{x}) \in \operatorname{ri} \partial f(\bar{x})$. Then for all $x \in \mathcal{M}$ near $\bar{x}$*

$$\nabla_{\mathcal{M}} f(x) = \mathrm{P}_{\partial f(x)}(0).$$

Proof. Proposition 17 enables us to say that $\mathrm{P}_{\operatorname{aff} \partial f(x)}(0) = \nabla_{\mathcal{M}} f(x)$. Applying Lemma 20 with $g(x) = \nabla_{\mathcal{M}} f(x)$, we obtain $\nabla_{\mathcal{M}} f(x) \in \operatorname{ri} \partial f(x)$ for $x \in \mathcal{M}$ near $\bar{x}$. Thus $\mathrm{P}_{\partial f(x)}(0) = \mathrm{P}_{\operatorname{aff} \partial f(x)}(0)$. ∎

Remark 22 (Steepest descent and Riemannian gradient). The direction of the steepest descent plays an important role in optimization problems. In the nonsmooth case, instead of normalizing the opposite direction of the gradient, we take the direction that minimizes the support function of $\partial f(x)$

$$\min_{\|d\| \leq 1} \max_{g \in \partial f(x)} \langle g, d \rangle,$$

see for instance [Wol75]. It is well known (by the min-max theorem), that this consists in taking

$$\max_{g \in \partial f(x)} \min_{\|d\| \leq 1} \langle g, d \rangle = \max_{g \in \partial f(x)} -\langle g, g/\|g\| \rangle = \max_{g \in \partial f(x)} -\|g\|,$$

so that the direction of the steepest descent is obtained by calculating

$$-\operatorname{argmin}\{\|g\|,\ g \in \partial f(x)\} = -\mathrm{P}_{\partial f(x)}(0).$$

Corollary 21 thus guarantees that if $x \in \mathcal{M}$ is close to a point $\bar{x} \in \mathcal{M}$ where $\nabla_{\mathcal{M}} f(\bar{x}) \in \operatorname{ri} \partial f(\bar{x})$, then the direction of the steepest descent in the "smooth world" of $\mathcal{M}$ and the direction of the steepest descent of nonsmooth analysis are the same. ■

In our next proposition, we demonstrate how the Riemann gradient can be constructed using a fixed subgradient.

Proposition 23 (Expression of $\nabla_{\mathcal{M}} f(x)$ with a fixed subgradient). *Suppose the function f is partly smooth relative to the manifold $\mathcal{M}$ at the point $\bar{x} \in \mathcal{M}$. Let $\bar{g}$ be any element of $\operatorname{ri} \partial f(\bar{x})$. Then for all $x \in \mathcal{M}$ close to $\bar{x}$,*

$$\nabla_{\mathcal{M}} f(x) = \mathrm{P}_{\partial f(x)}(\bar{g}) - \mathrm{P}_{\mathrm{N}_{\mathcal{M}}(x)}(\bar{g}).$$

Proof. Consider the function $h(x) = f(x) - \langle \bar{g}, x \rangle$. It is easy to check that h also is partly smooth relative to $\mathcal{M}$. Moreover, since

$$\partial h(x) = \partial f(x) - \bar{g}, \tag{10}$$

we have $0 = \nabla_{\mathcal{M}} h(\bar{x}) \in \operatorname{ri} \partial h(\bar{x})$. Applying Corollary 21 to h at $\bar{x}$ implies that for $x \in \mathcal{M}$ near $\bar{x}$,

$$\nabla_{\mathcal{M}} h(x) = \mathrm{P}_{\partial h(x)}(0). \tag{11}$$

Observe that the left-hand side of this equation can be expressed as

$$\mathrm{P}_{\partial h(x)}(0) = \mathrm{P}_{\partial f(x) - \bar{g}}(0) = \mathrm{P}_{\partial f(x)}(\bar{g}) - \bar{g}.$$

By using equation (10) and Proposition 17 (and the linearity of the projection mapping), the left-hand side of equation (11) becomes

$$\nabla_{\mathcal{M}} h(x) = \mathrm{P}_{\mathrm{T}_{\mathcal{M}}(x)}(\partial f(x) - \bar{g}) = \nabla_{\mathcal{M}} f(x) - \mathrm{P}_{\mathrm{T}_{\mathcal{M}}(x)}(\bar{g}).$$

Thus we have

$$\nabla_{\mathcal{M}} f(x) - \mathrm{P}_{\mathrm{T}_{\mathcal{M}}(x)}(\bar{g}) = \mathrm{P}_{\partial f(x)}(\bar{g}) - \bar{g},$$

which is equivalent to the desired equality. ■

Example 24 (Primal-dual track). In [MS05], an improved version of fast-track is considered: this consists in adding to the fast-track (or primal track) the so-called dual track defined as

$$\Gamma(u) = \operatorname{argmin}\{\|g\|^2, \quad g \in \partial f(\bar{x} + \bar{U}u + \bar{V}v(u))\}.$$

We saw in Example 15 that a convex function admitting a fast-track is partly smooth. Assuming in addition that $\Gamma(0) \in \operatorname{ri} \partial f(\bar{x})$, thus $\Gamma(0) = \nabla_{\bar{\mathcal{M}}} f(\bar{x}))$ and applying Corollary 21 we obtain that

$$\Gamma(u) = \nabla_{\bar{\mathcal{M}}} f(\bar{x} + \bar{U}u + \bar{V}v(u))$$

for $u \in \mathbb{R}^p$ small. The primal-dual track is thus a tangent vector field on $\bar{\mathcal{M}}$, namely the Riemann gradient vector field. ■

4 Prox-regularity and proximal points

4.1 Identification and characterization

In [Lew03] an example of a function which is partly smooth relative to two distinct manifolds is provided. This naturally provides some concern, as many of our equivalent definitions for $\nabla_{\mathcal{M}} f$ from Section 3 make no reference to the active manifold itself. The next proposition explains why these equivalences can exist even if the active manifold is not unique.

Proposition 25 (Multiple Manifolds). *If the function f is partly smooth at the point $\bar{x}$ relative to two distinct manifolds $\mathcal{M}_1$ and $\mathcal{M}_2$, then the Riemann gradient is independent of the active manifold examined, that is,*

$$\nabla_{\mathcal{M}_1} f(\bar{x}) = \nabla_{\mathcal{M}_2} f(\bar{x}).$$

Proof. Since f is partly smooth at $\bar{x}$ relative to $\mathcal{M}_1$ we know that $N_{\mathcal{M}_1}(\bar{x}) = \operatorname{aff} \partial f(\bar{x})$. Similarly, $N_{\mathcal{M}_2}(\bar{x}) = \operatorname{aff} \partial f(\bar{x})$. Thus $N_{\mathcal{M}_1}(\bar{x}) = N_{\mathcal{M}_2}(\bar{x})$ and by regularity we have $T_{\mathcal{M}_1}(\bar{x}) = T_{\mathcal{M}_2}(\bar{x})$. Therefore, in view of Definition 8 we have $\nabla_{\mathcal{M}_1} f(\bar{x}) = \nabla_{\mathcal{M}_2} f(\bar{x})$ as desired. ∎

Despite this, it might happen that the uniqueness of the active manifold is paramount. In [HL04] it is shown that by the addition of prox-regularity, the active manifold of partial smoothness becomes unique. In this section we investigate further effects that prox-regularity has on partly smooth functions. We begin with the pertinent definition.

Definition 26 (Prox-Regularity). *A function f is* prox-regular *at a point $\bar{x}$ for a subgradient $\bar{w} \in \partial f(\bar{x})$ if f is finite and locally lower semicontinuous at $\bar{x}$ and there exist $\varepsilon > 0$ and $R > 0$ such that*

$$f(x') \geq f(x) + \langle w, x' - x \rangle - \frac{R}{2}\|x' - x\|^2$$

whenever $\|x' - \bar{x}\| < \varepsilon$, $\|x - \bar{x}\| < \varepsilon$, $|f(x) - f(\bar{x})| < \varepsilon$, and $\|w - \bar{w}\| < \varepsilon$ with $w \in \partial f(x)$. We call a function prox-regular *at $\bar{x}$ if it is prox-regular at $\bar{x}$ for all $\bar{w} \in \partial f(\bar{x})$.*

It is worth noting that prox-regularity is a stronger condition than Clarke regularity in the sense that, if a function is prox-regular at $\bar{x}$ for a subgradient $\bar{w}$, then $\bar{w} \in \hat{\partial} f(\bar{x})$. Hence, if a function is prox-regular at $\bar{x}$ then it is Clarke regular at $\bar{x}$ [RW98, p. 610].

In [MS02] it was shown that for convex functions which admit a fast track, the proximal point mapping was attracted to the fast track. By recalling the relationship between fast tracks and partial smoothness (see Example 15), one can easily obtain that the proximal point map for a convex partly smooth function is attracted to the active manifold of the function. In our next theorem we see that this holds true for prox-regular partly smooth functions as well, a result which has also been shown in [MS04]. Theorem 28 further contains the previously unknown result that the proximal point mapping for a prox-regular C^k-partly smooth function, is C^{k-1}. In order to prove this result, we require the following lemma from [HL04, Theorem 3.2].

Lemma 27. *Suppose the function $\rho : \mathbb{R}^k \times \mathbb{R}^m \to \mathbb{R}$ is C^k-partly smooth at the point $(\bar{y}, \bar{z})$ relative to the manifold $\mathbb{R}^k \times \mathcal{M}$. Consider the family of parameterization functions $\rho_{\bar{y}}(\cdot) = \rho(\bar{y}, \cdot)$. If $\bar{z}$ is a nondegenerate critical point of $\rho_{\bar{y}}$ (i.e. $0 \in \operatorname{ri} \partial \rho_{\bar{y}}(\bar{z})$) and there exists $\varepsilon > 0$ such that*

$$\rho_{\bar{y}}(z) \geq \rho_{\bar{y}}(\bar{z}) + \varepsilon ||z - \bar{z}||^2$$

for all $z \in \mathcal{M}$ near $\bar{z}$, then there exist neighborhoods $\mathcal{N}_{\bar{z}}$ of $\bar{z}$ and $\mathcal{N}_{\bar{y}}$ of $\bar{y}$ and a function $\Phi \in C^{k-1}$ such that for all parameters $y \in \mathcal{N}_{\bar{y}}$, $\Phi(y) \in \mathcal{M}$ is a critical point of ρ_y restricted to $\mathcal{N}_{\bar{z}}$.

Theorem 28 (Proximal points locate active manifolds). *Suppose the function f is prox-bounded, and prox-regular at the point $\bar{x}$ and that f is C^k-partly smooth at $\bar{x}$ relative to the C^k-manifold $\mathcal{M}$ ($k \geq 1$) with $0 \in \operatorname{ri} \partial f(\bar{x})$. Then for $\lambda > 0$ sufficiently small, the proximal point mapping $P_\lambda f$ is C^{k-1} near $\bar{x}$ and*

$$P_\lambda f(x) = \operatorname*{argmin}_{y \in \mathcal{M}} \left\{ f(y) + \frac{1}{2\lambda} ||y - x||^2 \right\}. \tag{12}$$

In particular the proximal point $P_\lambda f(x)$ belongs to the manifold $\mathcal{M}$.

Proof. We begin by showing the function

$$\rho(y,z) = \frac{1}{2\lambda}||z-y||^2 + f(z),$$

satisfies the conditions of Lemma 27 at $(\bar{x},\bar{x})$ with $\varepsilon = 1$. Since f is partly smooth relative to $\mathcal{M}$ at $\bar{x}$, the function $(y,z) \mapsto f(z)$ is partly smooth relative to $\mathbb{R}^n \times \mathcal{M}$ [Lew03, Proposition 4.5]. As the addition of a smooth function does not alter partial smoothness [Lew03, Corollary 4.6], we conclude that ρ is partly smooth relative to $\mathbb{R}^m \times \mathcal{M}$ at $(\bar{x},\bar{x})$. Examining the parameterization functions $\rho_y(\cdot) = \rho(y,\cdot)$ we immediately see that $\partial\rho_{\bar{x}}(\bar{x}) = \partial f(\bar{x})$. Therefore, as $\bar{x}$ is a nondegenerate critical point of f, it is a nondegenerate critical point of $\rho_{\bar{x}}$. The inequality $\rho_{\bar{x}}(x) \geq \rho_{\bar{x}}(\bar{x}) + ||x-\bar{x}||^2$ is equivalent to

$$f(x) \geq f(\bar{x}) + \langle 0, x-\bar{x}\rangle - \left(\frac{1}{2\lambda} - 1\right)||x-\bar{x}||^2. \tag{13}$$

Since f is prox-regular at $\bar{x}$ for 0, there exists $R > 0$ such that

$$f(x) \geq f(\bar{x}) + \langle 0, x-\bar{x}\rangle - \frac{R}{2}||x-\bar{x}||^2$$

for x near $\bar{x}$. Then selecting λ sufficiently small to ensure that $\frac{1}{2\lambda} - 1 > R/2$, we see inequality (13) must hold. Thus ρ satisfies the conditions of Lemma 27 (with $\varepsilon = 1$). Therefore, there exists a function $\Phi \in C^{k-1}$ such that for all parameters y near $\bar{x}$, $\Phi(y) \in \mathcal{M}$ is a critical point of ρ_y near $\bar{x}$. That is

$$0 \in \partial\rho_y(\Phi(y)) = \partial f(\Phi(y)) + \frac{1}{\lambda}(\Phi(y) - y),$$

for all y near $\bar{x}$. By [PR96b, Theorem 4.4] this implies $\Phi(y) = P_\lambda(y)$. Since $\Phi(y) \in \mathcal{M}$ and $\Phi(y) \in C^{k-1}$ the proof is complete. ∎

Theorem 28 tells us that for prox-regular partly smooth functions, the proximal points are extremely well behaved. Our next theorem characterizes proximal points in terms of the Riemann gradient.

Theorem 29 (Characterization of proximal points). *Suppose the function f is prox-bounded, and prox-regular at the point $\bar{x}$ and that f is partly smooth at $\bar{x}$ relative to a manifold $\mathcal{M}$ with $0 \in \operatorname{ri}\partial f(\bar{x})$. Then for $\lambda > 0$ sufficiently small and x sufficiently close to $\bar{x} \in \mathcal{M}$ the proximal point $P_\lambda(x)$ belongs to $\mathcal{M}$ and is characterized by*

$$\begin{aligned} y = P_\lambda(x) \quad &\iff \quad \lambda\nabla_{\mathcal{M}} f(y) = \mathrm{P}_{\mathrm{T}_{\mathcal{M}}(y)}(x-y) \\ &\iff \quad \mathrm{P}_{\mathrm{T}_{\mathcal{M}}(y)}(\nabla\widetilde{f}(y) + \tfrac{1}{\lambda}(y-x)) = 0 \\ &\iff \quad \tfrac{1}{\lambda}(x-y) \in \nabla\widetilde{f}(y) + \mathrm{N}_{\mathcal{M}}(y), \end{aligned}$$

where $\widetilde{f}$ is a smooth representation of f along $\mathcal{M}$ such that $\nabla\widetilde{f}(\bar{x}) = 0$.

Proof. By Proposition 7 we can find a smooth representation of f along $\mathcal{M}$, $\widetilde{f}$ such that $\nabla\widetilde{f}(\bar{x}) = 0$. For λ sufficiently small and x sufficiently close to $\bar{x}$ we know by Theorem 28 that $P_\lambda f(x) \in \mathcal{M}$, and therefore

$$\begin{aligned} P_\lambda(x) &= \operatorname{argmin}_{y\in\mathcal{M}}\{f(y) + \tfrac{1}{2\lambda}\|x-y\|^2\} \\ &= \operatorname{argmin}_{y\in\mathbb{R}^n}\{\widetilde{f}(y) + \delta_{\mathcal{M}}(y) + \tfrac{1}{2\lambda}\|x-y\|^2\} \\ &= P_\lambda(\widetilde{f} + \delta_{\mathcal{M}})(x). \end{aligned}$$

Observe now that $\widetilde{f} + \delta_{\mathcal{M}}$ is prox-regular at $\bar{x}$ and that $0 \in \partial(\widetilde{f} + \delta_{\mathcal{M}})(\bar{x})$. Then [PR96b, Theorem 4.4] yields

$$y = P_\lambda(\widetilde{f} + \delta_{\mathcal{M}})(x) \iff \frac{1}{\lambda}(x-y) \in \nabla\widetilde{f}(y) + \mathrm{N}_{\mathcal{M}}(y),$$

for x close to $\bar{x}$. Projecting this equality on $\mathrm{T}_{\mathcal{M}}(y)$ gives the remaining two equivalences, by using linearity of the projection and Definition 8. ∎

4.2 Interpretation of the $\mathcal{VU}$-proximal algorithm

The $\mathcal{VU}$ proximal point algorithm developed by Mifflin and Sagastizábal [MS05] follow the predictor-corrector pattern of Subsection 2.3. Specifically, the conceptual form of this algorithm computes the next iterate by the following process for $x \in \mathcal{M}$:

1. (Predictor) make a Newton step in the tangent space

$$\widetilde{x}_+ = x - [\nabla^2_{\mathcal{M}} f(x)]^{-1} \nabla_{\mathcal{M}} f(x).$$

2. (Corrector) make a proximal step (towards $\mathcal{M}$)

$$x_+ = P_\lambda(\widetilde{x}_+).$$

The algorithm presented in [MS05, Algorithm 6] is an implementable form of this predictor-corrector scheme: the matrix H_k approximates $[\nabla^2_{\mathcal{M}} f(x)]^{-1}$, s_k approximates the "dual track" (that is $\nabla_{\mathcal{M}} f(x)$ by Example 24) and p_{k+1} approximates x_+ as well.

In this subsection we compare the $\mathcal{VU}$-proximal algorithm and the projected Newton method (see Subsection 2.3). To enlighten notation, we introduce, for $(x, u) \in \mathrm{T}\mathcal{M}$,

$$\begin{aligned} (y_\delta =)\, y^\lambda_\delta &= \mathrm{P}_{\mathcal{M}}(x+u) = \operatorname{argmin}_y \{\delta_{\mathcal{M}}(y) + \tfrac{1}{2\lambda}\|x+u-y\|^2\} \\ y^\lambda_f &= P_\lambda(x+u) = \operatorname{argmin}_y \{f(y) + \tfrac{1}{2\lambda}\|x+u-y\|^2\}. \end{aligned}$$

It follows directly from its definition that $y^\lambda_\delta \in \mathcal{M}$. Note also that since y^λ_δ is independent of $\lambda > 0$, the abbreviation of $y^\lambda_\delta = y_\delta$ is acceptable. On the other hand, Theorem 28 guarantees that $y^\lambda_f \in \mathcal{M}$ (provided $x + u$ is sufficiently close to $\bar{x}$ and λ is sufficiently small). Thus we obtain that

$$\|x + u - y_\delta\|^2 \leq \|x + u - y^\lambda_f\|^2. \tag{14}$$

Applying this to the definition of y^λ_f we find

$$\begin{aligned} f(y^\lambda_f) + \tfrac{1}{2\lambda}\|x+u-y^\lambda_f\|^2 &\leq f(y_\delta) + \tfrac{1}{2\lambda}\|x+u-y_\delta\|^2 \\ &\leq f(y_\delta) + \tfrac{1}{2\lambda}\|x+u-y^\lambda_f\|^2, \end{aligned}$$

therefore

$$f(y^\lambda_f) \leq f(y_\delta). \tag{15}$$

Thus, comparing with the projection step, the proximal point step is at least as large (in norm) and causes at least as great of a decrease in function value.

Proposition 30 shows that in general the proximal point method actually takes a larger step and causes greater decrease in the function value. In fact the only time the proximal point and projection coincide is when the projection method successfully finds a critical point.

Proposition 30 (Prox vs. Projection). *Suppose the function f is prox-regular at the point $\bar{x}$, and partly smooth there relative to the manifold $\mathcal{M}$ with $0 \in \operatorname{ri} \partial f(\bar{x})$. For a point x, let us consider*

$$(x_\delta =)\, x^\lambda_\delta = \mathrm{P}_{\mathcal{M}}(x) \quad \text{(proximal step)} \quad \textit{and} \quad x^\lambda_f = P_\lambda f(x) \quad \text{(projection step)}.$$

Suppose x is sufficiently close to $\bar{x}$ and $\rho > 0$ is sufficiently small that for any $\lambda < \rho$ Theorem 28 and Theorem 29 apply. Then for a smooth representation $\widetilde{f}$ of f along $\mathcal{M}$ such that $\nabla \widetilde{f}(\bar{x}) = 0$ the following properties are equivalent:

(i) $x_\delta = x^\lambda_f$ *for some* $\lambda < \rho$;

(ii) $\nabla_{\mathcal{M}} f(x_\delta) = 0$;

(iii) $0 \in \partial f(x_\delta)$;

(iv) $\nabla \widetilde{f}(x_\delta) \in \mathrm{N}_{\mathcal{M}}(x_\delta)$;

(v) $x_f^\lambda = x_\delta$, *for all* $\lambda < \rho$.

Proof. Note again that x_δ^λ is independent of $\lambda > 0$, so we can make use of the abbreviation $x_\delta^\lambda = x_\delta$. On the other hand, Proposition 7 shows that $\widetilde{f}$ exists as required. Since x_δ is independent of λ we have $-\frac{1}{\lambda}(x_\delta - x) \in \mathrm{N}_{\mathcal{M}}(x_\delta)$, for all $\lambda > 0$. In fact, as $\mathrm{N}_{\mathcal{M}}(x_\delta)$ is a subspace this optimality condition can be strengthened to:

$$y = x_\delta \quad \Longleftrightarrow \quad -R(y - x) \in \mathrm{N}_{\mathcal{M}}(y), \text{ for all } R. \tag{16}$$

(i) ⇒ (ii). If $x_f^\lambda = x_\delta$ we have $(x - x_f^\lambda) \in \mathrm{N}_{\mathcal{M}}(x_f^\lambda)$. Therefore $\mathrm{P}_{\mathrm{T}_{\mathcal{M}}(x_f^\lambda)}(x - x_f^\lambda) = 0$, which by Theorem 29 implies $0 = \nabla_{\mathcal{M}} f(x_f^\lambda) = \nabla_{\mathcal{M}} f(x_\delta)$.

(ii) ⇔ (iii) ⇔ (iv). It follows from Definition 8 and Corollary 21.

(iv) ⇒ (v). Suppose $\nabla \widetilde{f}(x_\delta) \in \mathrm{N}_{\mathcal{M}}(x_\delta)$, thus $-\nabla \widetilde{f}(x_\delta) \in \mathrm{N}_{\mathcal{M}}(x_\delta)$ too, since $\mathrm{N}_{\mathcal{M}}(x_\delta)$ is a subspace. For any $\lambda > 0$, by equation (16) we have $\frac{1}{\lambda}(x - x_\delta) \in \mathrm{N}_{\mathcal{M}}(x_\delta)$. Summing up these two inclusions, we obtain

$$\frac{1}{\lambda}(x - x_\delta) \in \nabla \widetilde{f}(x_\delta) + \mathrm{N}_{\mathcal{M}}(x_\delta).$$

Thus, for $\lambda \leq \rho$ such that Theorem 29 applies, we obtain $x_\delta = P_\lambda(x)$, thus (v) holds.

(v) ⇒ (i). It is obvious. ∎

The $\mathcal{VU}$-proximal algorithm has thus a good behavior. In addition, the forthcoming theorem shows that the quadratic convergence is not lost.

Theorem 31 (Quadratic convergence of the $\mathcal{VU}$-proximal algorithm). *Let f be prox-bounded, prox-regular at the point $\bar{x}$ and $\mathcal{C}^2$-partly smooth at $\bar{x}$ relative to the manifold $\mathcal{M}$. Suppose that $0 \in \mathrm{ri}\,\partial f(\bar{x})$ and that $\nabla^2_{\mathcal{M}} f(\bar{x})$ is nonsingular. Then the conceptual proximal $\mathcal{VU}$-proximal algorithm converges quadratically when started sufficiently close to $\bar{x}$.*

Proof. For $x \in \mathcal{M}$, let us set

$$h(x) = x - [\nabla^2_{\mathcal{M}} f(x)]^{-1} \nabla_{\mathcal{M}} f(x) \qquad \text{and} \qquad N(x) = P_\lambda(h(x)),$$

such that one iteration of the algorithm is $x_+ = N(x)$. Note that $\nabla^2_{\mathcal{M}} f(x)$ is also nonsingular at $x \in \mathcal{M}$ around $\bar{x}$, hence h is well-defined and smooth on $\mathcal{M}$ near $\bar{x}$. By Theorem 28, P_λ is smooth near $\bar{x}$ with values on $\mathcal{M}$. This yields that $N : \mathcal{M} \to \mathcal{M}$ is smooth too. Taking the first order development of this smooth function around $\bar{x}$ (relative to $\mathcal{M}$), we have

$$N(x) = N(\bar{x}) + \mathrm{D}N(\bar{x})(x - \bar{x}) + \mathrm{O}(\|x - \bar{x}\|^2) \tag{17}$$

Since $0 \in \partial f(\bar{x})$ and f is prox-regular at x, we have

$$N(\bar{x}) = P_\lambda(h(\bar{x})) = P_\lambda(\bar{x}) = \bar{x}.$$

On the other hand,

$$\mathrm{D}N(\bar{x}) = \mathrm{D}P_\lambda(h(\bar{x}))\mathrm{D}h(\bar{x}).$$

Since $\nabla_{\mathcal{M}} f(\bar{x}) = 0$, for any $u \in \mathrm{T}_{\mathcal{M}}(\bar{x})$ we have

$$\mathrm{D}h(\bar{x})(u) = u - \mathrm{D}([\nabla^2_{\mathcal{M}} f(\bar{x})]^{-1})\nabla_{\mathcal{M}} f(\bar{x})u + [\nabla^2_{\mathcal{M}} f(\bar{x})]^{-1}\nabla^2_{\mathcal{M}} f(\bar{x})u = u - u = 0.$$

It follows from (17) that $N(x) - \bar{x} = \mathrm{O}(\|x - \bar{x}\|^2)$ which proves the quadratic convergence. ∎

Remark 32. Theorem 31 presents assumptions that ensure the efficiency of the conceptual $\mathcal{VU}$-algorithm. In [MS05], such a theorem is stated, but it requires more technical assumptions that seem unavoidable when dealing with concrete algorithms that do not require explicit knowledge of $\mathcal{M}$.

Acknowledgement

Part of this work was carried out during a visit of the third author in the CRM and the Autonomous University of Barcelona (January 2005) and a research stay of the first author to the University of Savoie (March-May 2005). These two authors wish to thank their hosts for the hospitality.

References

[AES99] T. Arias A. Edelman and S. T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl.*, 20(2):303–353, 1999.

[AKK91] F. Al-Khayyal and J. Kyparisis. Finite convergence of algorithms for nonlinear programs and variational inequalities. *J. Optim. Theory Appl.*, 70(2):319–332, 1991.

[Arn71] V. Arnold. On matrices depending on parameters. *Russian Math. Surveys*, 26:29–43, 1971.

[BM88] J. V. Burke and J. J. Moré. On the identification of active constraints. *SIAM J. Numer. Anal.*, 25(5):1197–1211, 1988.

[Bur90] J. Burke. On the identification of active constraints. II. The nonconvex case. *SIAM J. Numer. Anal.*, 27(4):1081–1103, 1990.

[CM87] P. H. Calamai and J.J. Moré. Projected gradient methods for linearly constrained problems. *Math. Programming*, 39(1):93–116, 1987.

[dC92] M. P. do Carmo. *Riemannian Geometry*. Mathematics: Theory and Applications. Birkhäuser, 1992.

[DMP03] J.-P. Dedieu, G. Malajovich, and P. Priouret. Newton's method on Riemannian manifolds: covariant alpha-theory. *IMA J. Num. Anal.*, 23(3):395–419, 2003.

[Gab82] D. Gabay. Minimizing a differentiable function over a differentiable manifold. *J. Optim. Theory Appl.*, 37(2):177–219, June 1982.

[Har04a] W.L Hare. *Nonsmooth Optimization with Smooth Substructure*. PhD thesis, Simon Fraser University, 2004.

[Har04b] W.L Hare. Recent functions and sets of smooth substructure: Relationships and examples. to appear *J. Comput. Optim. Appl.*, 33(2), April 2006.

[HL04] W.L Hare and A. S. Lewis. Identifying active constraints via partial smoothness and prox-regularity. *J. Convex Anal.*, 11(2):251–266, 2004.

[Lew03] A. S. Lewis. Active sets, nonsmoothness and sensitivity. *SIAM J. Optim.*, 13:702–725, 2003.

[LOS00] C. Lemaréchal, F. Oustry, and C. Sagastizábal. The $\mathcal{U}$-Lagrangian of a convex function. *Trans. Amer. Math. Soc.*, 352(2):711–729, 2000.

[MM05] J. Malick and S. Miller. Newton methods for nonsmooth convex minimization: connection among $\mathcal{U}$-Lagrangian, Riemannian Newton and SQP methods. *Math. Programming*, 104(3), 2005.

[MS99] R. Mifflin and C. Sagastizábal. $\mathcal{VU}$-decomposition derivatives for convex max-functions. In *Ill-Posed Variational Problems and Regularization Techniques (Trier, 1998)*, volume 477 of *Lecture Notes in Econom. and Math. Systems*, pages 167–186. Springer, Berlin, 1999.

[MS00a] R. Mifflin and C. Sagastizábal. Functions with primal-dual gradient structure and $\mathcal{U}$-hessians. In *Nonlinear Optimization and Related Topics (Erice, 1998)*, volume 36 of *Appl. Optim.*, pages 219–233. Kluwer Acad. Publ., Dordrecht, 2000.

[MS00b] R. Mifflin and C. Sagastizábal. On $\mathcal{VU}$-theory for functions with primal-dual gradient structure. *SIAM J. Optim.*, 11(2):547–571, 2000.

[MS02] R. Mifflin and C. Sagastizábal. Proximal points are on the fast track. *J. Convex Anal.*, 9(2):563–579, 2002.

[MS04] R. Mifflin and C. Sagastizábal. $\mathcal{VU}$-smoothness and proximal point results for some non-convex functions. *Optim. Meth. Soft.*, 19(5), 2004.

[MS05] R. Mifflin and C. Sagastizábal. A $\mathcal{VU}$-proximal point algorithm for convex minimization. *Math. Programming*, 104(3), 2005.

[Ous99] F. Oustry. The $\mathcal{U}$-Lagrangian of the maximum eigenvalue function. *SIAM J. Optim.*, 9(2):526–549, 1999.

[PR96a] R. A. Poliquin and R. T. Rockafellar. Generalized hessian properties of regularized nonsmooth functions. *SIAM J. Optim.*, 6(4):1121–1137, 1996.

[PR96b] R. A. Poliquin and R. T. Rockafellar. Prox-regular functions in variational analysis. *Trans. Amer. Math. Soc*, 348(5):1805–1838, 1996.

[Roc82] R. T. Rockafellar. Favorable classes of Lipschitz-continuous functions in subgradient optimization. In *Progress in Nondifferentiable Optimization*, volume 8 of *IIASA Collaborative Proc. Ser. CP-82*, pages 125–143. Internat. Inst. Appl. Systems Anal., Laxenburg, 1982.

[RW98] R. Tyrrell Rockafellar and R. J.-B. Wets. *Variational Analysis.* Number 317 in Grundlehren der mathematischen Wissenschaften. Springer-Verlag, Berlin, 1998.

[Sha03] A. Shapiro. On a class of nonsmooth composite functions. *Math. Oper. Res.*, 28:677–692, 2003.

[Smi94] S. T. Smith. Optimization techniques on Riemannian manifolds. *Fields Inst. Comm.*, 3:113–136, 1994.

[Wol75] P. Wolfe. A method of conjuguate subgradients. *Math. Program. Study*, 3:145–173, 1975.

[Wri93] S. J. Wright. Identifiable surfaces in constrained optimization. *SIAM J. Control Optim.*, 31(4):1063–1079, 1993.

Aris DANIILIDIS

Departament de Matemàtiques
Universitat Autònoma de Barcelona
E-08193 Bellaterra (Cerdanyola del Vallès), Spain.

e-mail: arisd@mat.uab.es
http://mat.uab.es/~arisd

Research supported by the MEC Grant No. MTM2005-08572-C03-03 (Spain).

Warren HARE

IMPA - Instituto Nacional de Matemática Pura e Aplicada,
Estrada Dona Castorina 110
Rio de Janeiro, Brasil
22460-320

email: whare@cecm.sfu.ca
http://www.cecm.sfu.ca/~whare

Research supported by CNPq Grant No. 150234/2004-0

Jérôme MALICK

INRIA, Rhône-Alpes
655 avenue de l'Europe
Montbonnot, St Martin
F-38334 Saint Ismier, France

email: jerome.malick@inria.fr
http://www.inrialpes.fr/bipop/people/malick/