# Harnessing structure in composite nonsmooth minimization

Gilles Bareilles, Franck Iutzeler, Jérôme Malick

# HARNESSING STRUCTURE IN COMPOSITE NONSMOOTH MINIMIZATION*

GILLES BAREILLES†, FRANCK IUTZELER†, AND JÉRÔME MALICK‡

**Abstract.** We consider the problem of minimizing the composition of a nonsmooth function with a smooth mapping in the case where the proximity operator of the nonsmooth function can be explicitly computed. We first show that this proximity operator can provide the exact smooth substructure of minimizers, not only of the nonsmooth function, but also of the full composite function. We then exploit this proximal identification by proposing an algorithm which combines proximal steps with sequential quadratic programming steps. We show that our method identifies the optimal smooth substructure and converges locally quadratically. We illustrate its behavior on two problems: the minimization of a maximum of quadratic functions and the minimization of the maximal eigenvalue of a parametrized matrix.

**Key words.** Nonsmooth optimization, proximal operator, partial smoothness, manifold identification, maximum eigenvalue minimization, sequential quadratic programming.

**AMS subject classifications.** 65K10, 90C26, 49Q12, 90C55.

## 1. Introduction.

### 1.1. Context: structured nonsmooth optimization.

In this paper, we consider nonsmooth optimization problems of the form

$$\min_{x\in\mathbb{R}^n} \; F(x) \;\triangleq\; g(c(x)), \tag{1.1}$$

where the inner mapping $c : \mathbb{R}^n \to \mathbb{R}^m$ is smooth and the outer function $g : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ is nonsmooth and may be nonconvex, but admits an explicit proximity operator. Such composite nonsmooth optimization problems appear in a variety of applications in signal processing, machine learning, and control, such as robust nonlinear regression, phase synchronization, nonsmooth penalty functions; see *e.g.* [20, 27] and the references therein.

Throughout the paper, we illustrate our developments on two classes of functions: the pointwise maximum of $m$ smooth real-valued functions $c_i$

$$F(x) = \max_{i=1,\dots,m} (c_i(x)) \tag{1.2}$$

and the maximum eigenvalue of a parametrized symmetric real matrix $c$

$$F(x) = \lambda_{\max}(c(x)). \tag{1.3}$$

In these two examples and many others, subgradients of $F$ can be computed and thus the composite function can be minimized using standard nonsmooth optimization algorithms (*e.g.* subgradient methods, gradient sampling [7], nonsmooth BFGS [19], or bundle methods [13]). Nevertheless, these methods do not exploit the fact that $F$ is a composition of a smooth mapping $c$, which can hinder their performance. In contrast, the so-called prox-linear methods leverage this composite expression by introducing an extension of the proximity operator where the nonlinear mapping $c$

*Submitted to the editors June 27, 2022.

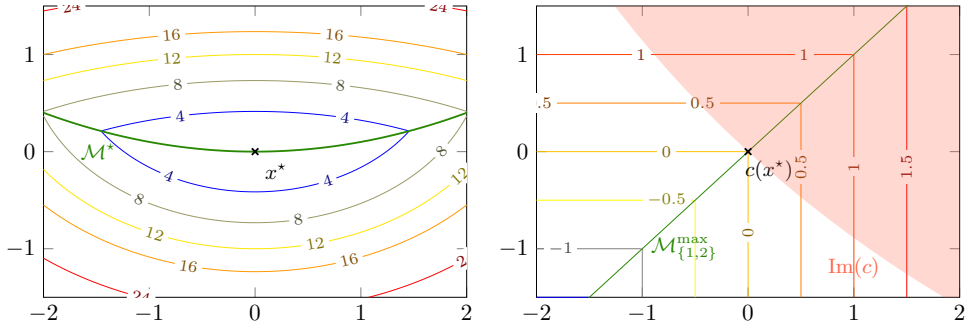
†Univ. Grenoble Alpes, LJK, France (gilles.bareilles@univ-grenoble-alpes.fr, franck.iutzeler@univ-grenoble-alpes.fr).

‡Univ. Grenoble Alpes, CNRS, Grenoble INP, LJK (jerome.malick@univ-grenoble-alpes.fr).

is iteratively replaced by a first-order Taylor approximation [20]. These methods benefit from theoretical convergence guarantees, and nicely generalize to Taylor-like approximations [9, 3]. However these methods are not always directly implementable because the prox-linear step may be hard to compute, as in (1.3).

In this paper, we propose an optimization algorithm for solving (1.1) exploiting that the nonsmooth objective function $F = g \circ c$ writes as a composition with a simple nonsmooth function $g$, which displays some smooth substructure, as discussed below.

**1.2. Smooth substructure, identification, and existing algorithms.** For many composite functions, including (1.2) and (1.3), the nondifferentiability points locally organize into *smooth manifolds over which $F$ evolves smoothly.* We illustrate in Figure 1 such a smooth substructure for a maximum of two functions.




(a) Level curves of $F$ (b) Level curves of $g$ and image of $c$

Fig. 1: Smooth substructure on a simple example ($n = m = 2$). The figures show the level curves of $g(y) = \max(y_1, y_2)$ (on the right, in the intermediate space) and of $F = g \circ c$, with two quadratic functions $c_1(x)$, $c_2(x)$ (on the left, in the input space). The manifolds of non-differentiability are in green; the image of $c$ is the red area.

The smooth substructure of $F$ can help in solving (1.1). Indeed, if the optimal solution $x^\star$ belongs to a manifold $\mathcal{M}^\star$ that is known beforehand, then minimizing the nonsmooth function $F$ over $\mathbb{R}^n$ boils down to minimizing the smooth restriction $F|_{\mathcal{M}^\star}$ over this smooth *optimal manifold* $\mathcal{M}^\star$. This would enable to solve (1.1) by smooth constrained optimization algorithms, such as Sequential Quadratic Programming (SQP) methods (see *e.g.* [23, 5]). The main difficulty in practice is that *we do not know $\mathcal{M}^\star$ in advance.*

Thus, the algorithms exploiting this smooth substructure require two ingredients:

i) a mechanism to identify the optimal manifold;

ii) an efficient method to minimize $F$ restricted to this manifold.

For general convex functions, the algorithm of [21] mixes a proximal bundle iteration (as a heuristic for identification) and a so-called $\mathcal{U}$-Newton iteration (which interprets as an SQP step; see [22, Sec. 5]). The obtained superlinear rate hinges on the identification of the optimal manifold.

For max-of-smooth functions (1.2), the paper [29] pioneered the idea of seeking the optimal manifold and using it to make second-order steps. Their identification heuristic uses the indices of the maximal function along a descent direction. Recently, [17, 10] investigate a related setting and propose bundle-like algorithms incorporating high-order information that converge (super)linearly on max-of-smooth functions when the optimal manifold is known.

For the maximum eigenvalue of a parametrized matrix (1.3), a specific version of the $\mathcal{U}$-Newton method discussed above is studied by [24]. Again, the identification mechanism is a heuristic determining the multiplicity of the maximal eigenvalue and the optimization step is an SQP iteration.

None of these methods guarantee identification of the optimal manifold: they either assume that the optimal manifold is known in advance, or rely on heuristics for identification. Here, we aim at further harnessing the smooth substructure of $F = g \circ c$ to have *guaranteed local identification* of the optimal manifold and then *guaranteed quadratic convergence* when using SQP iterations.

**1.3. Contributions and outline.** We propose a second-order algorithm for solving the nonsmooth composite problem (1.1) that identifies the optimal manifold of non-differentiability. The two main ingredients of our algorithm are the following:

i) we use the explicit proximal operator of $g$ with chosen stepsizes to provide a guaranteed identification procedure;
ii) for a candidate manifold $\mathcal{M}$, we make an SQP iteration minimizing a smooth extension of $F|_{\mathcal{M}}$ subject to the constraint of belonging to $\mathcal{M}$.

The fact proximal-based operators have identification properties around minimizers is well-known: the proximal operator [11], the proximal gradient operator [1], approximate variable-metric proximal gradient operators [14], and prox-linear operators [20] locally identify the optimal manifold under some natural geometrical assumptions. Here, we only have access to the proximity operator of $g$, and in order to exploit the structure it provides, we face the double challenge of, first, identifying the smooth structure around a point which is not a minimizer for $g$, and, second, deducing the corresponding structure of $F = g \circ c$. Thus, our main technical contribution is to establish that $\mathbf{prox}_{\gamma g}$ maps a point $y$ close to $c(x^\star)$ to $c(\mathcal{M}^\star)$. The step $\gamma$ should be carefully chosen, in particular larger than the distance of $y$ to $c(x^\star)$. Mathematically, we study the range of steps for which the curve $\gamma \mapsto \mathbf{prox}_{\gamma g}(y)$ belongs to $c(\mathcal{M}^\star)$. This analysis shows connections with recent works in nonsmooth analysis, such as the modulus of identifiability appearing in [16].

We combine this new identification result with standard SQP-steps to propose an algorithm for minimizing the composite function $F$. We pay a special attention to prevent the quadratic convergence of SQP from jeopardizing identification: we prove that, for a well-chosen stepsize policy, the method identifies the optimal structure and locally converges quadratically. We illustrate numerically these properties on problems of the form (1.2) and (1.3).

The outline of the remainder of the paper is as follows. First, in Section 2, we introduce the technical tools to describe the manifold identification brought by proximity operators (including prox-regularity and partial smoothness). Furthermore, we lay out two technical properties needed for proximal identification in the composite setting. In Section 3, we show our main result consisting in a description of a stepsize range for which the proximity operator of $g$ identifies the optimal manifold locally around a minimizer. In Section 4 we detail the proposed method combining SQP-steps and proximal identification steps. Finally, we present in Section 5 numerical illustrations of our method and of the identification result.

**2. Setting and assumptions.** Let us start by representing schematically the type of functions we consider:

$$\mathbb{R}^n \xrightarrow[\text{smooth mapping}]{c} \operatorname{Im}(c) \subset \mathbb{R}^m \xrightarrow[\text{nonsmooth function}]{g} \mathbb{R} \cup \{+\infty\}.$$

Throughout the paper, we denote by $x$ points in the *input space* $\mathbb{R}^n$ and by $y$ points in the *intermediate space* $\mathbb{R}^m$.

In all the results presented in this paper, we make the following assumption that describes the minimal global properties on $g$ and $c$ to conduct our reasoning.

ASSUMPTION 2.1. *The mapping $c : \mathbb{R}^n \to \mathbb{R}^m$ is $\mathcal{C}^2$, the function $g : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ is proper and lower semi-continuous.*

We work with the set of *(general) subgradients* (see [26, Def. 8.3]), defined at a point $\bar{y}$ where $g(\bar{y})$ is finite as:

$$\partial g(\bar{y}) \triangleq \left\{ \lim_r v_r : v_r \in \widehat{\partial} g(y_r), y_r \to \bar{y}, g(y_r) \to g(\bar{y}) \right\},$$

where $\widehat{\partial} g(\bar{y})$ denotes the set of *regular (or Fréchet) subgradients*, defined as

$$\widehat{\partial} g(\bar{y}) \triangleq \{ v : g(y) \geq g(\bar{y}) + \langle v, y - \bar{y} \rangle + o(\|y - \bar{y}\|) \text{ for all } y \in \mathbb{R}^m \}.$$

These two subdifferentials match if (and only if) $g$ is (Clarke) regular at $\bar{y}$. Closed convex functions are regular everywhere and these subdifferentials match the usual convex subdifferential (see [26, Chap. 8.11-12] for details).

In the remainder of this section, we provide quick recalls and definitions about the two important objects of our analysis: the proximity operator in Subsection 2.1 and the structure manifolds in Subsection 2.2. We illustrate them on our running examples (1.2) and (1.3).

**2.1. Proximity operator.** The proximity operator of a function $g$ with step $\gamma > 0$ at $y \in \mathbb{R}^m$ is defined as the set-valued mapping

$$\mathbf{prox}_{\gamma g}(y) \triangleq \operatorname{argmin}_{u \in \mathbb{R}^m} \left\{ g(u) + \frac{1}{2\gamma} \|u - y\|^2 \right\}.$$

This operator is well-defined when $g$ is prox-regular and prox-bounded; see *e.g.* [26, 13.37]. We quickly introduce these two notions and recall a result on the uniqueness and characterization of the prox operator, which is important in our developments.

A function $g$ is *prox-regular* at a point $\bar{y}$ for a subgradient $\bar{v} \in \partial g(\bar{y})$ if $g$ is finite, locally lower semi-continuous at $\bar{y}$, and there exists $r > 0$ and $\epsilon > 0$ such that

$$g(y') \geq g(y) + \langle v, y' - y \rangle - \frac{r}{2} \|y' - y\|^2$$

whenever $v \in \partial g(y)$, $\|y - \bar{y}\| < \epsilon$, $\|y' - \bar{y}\| < \epsilon$, $\|v - \bar{v}\| < \epsilon$, and $g(y) < g(\bar{y}) + \epsilon$. When this holds for all $\bar{v} \in \partial g(\bar{y})$, we say that $g$ is prox-regular at $\bar{y}$ [26, Def. 13.27].

A function $g$ is *prox-bounded* if there exists $R \geq 0$ such that the function $g + \frac{R}{2}\|\cdot\|^2$ is bounded below. The corresponding *threshold* (of prox-boundedness) is the smallest $r_{pb} \geq 0$ such that $g + \frac{R}{2}\|\cdot\|^2$ is bounded below for all $R > r_{pb}$. In this case, $g + \frac{R}{2}\|\cdot - \bar{y}\|^2$ is bounded below for any $\bar{y}$ and $R > r_{pb}$ [26, Def. 1.23, Th. 1.25].

We can now recall a relevant result on the characterization of proximal points.

PROPOSITION 2.2 ([12, Th. 1]). *Suppose that the function $g$ is prox-regular at $\bar{y}$ for $\bar{v} \in \partial g(\bar{y})$ with parameter $r_{pr}$, and prox-bounded with threshold $r_{pb}$. Then, for any $\gamma < \min(r_{pr}^{-1}, r_{pb}^{-1})$ and all $y$ near $\bar{y} + \gamma \bar{v}$, the proximal operator is:*

- *single-valued and locally Lipschitz continuous;*
- *uniquely determined by the relation*

$$p = \mathbf{prox}_{\gamma g}(y) \Leftrightarrow y - p \in \gamma \partial g(p).$$

In addition to its existence and characterization provided by the result above, the proximity operator has a closed-form expression in our running examples.

EXAMPLE 2.3 (Maximum). *The subdifferential of* $g(y) = \max(y_1, \dots, y_m)$ *is*

$$\partial \max(y) = \operatorname{Conv}\{e_i : y_i = \max(y)\},$$

*where* $e_i$ *is the* $i$*-th element of the Cartesian basis of* $\mathbb{R}^m$*. This function is convex, thus globally prox-regular and prox-bounded everywhere (with parameters* 0*). Its proximity operator is given (coordinate-wise) by*

$$\left[\mathbf{prox}_{\gamma \max}(y)\right]_i = \begin{cases} s & \text{if } y_i > s \\ y_i & \text{else} \end{cases}$$

*where* $s$ *is the unique real number such that* $\sum_{\{i:y_i>s\}}(y_i - s) = \gamma$*.*

EXAMPLE 2.4 (Maximum eigenvalue). *Denote the eigenvalue decomposition of a point* $y \in \mathbb{S}_m$ *as* $y = E \operatorname{Diag}(\lambda) E^\top$*, where* $\lambda \in \mathbb{R}^m$ *is a vector with decreasing entries and* $E \in \mathbb{R}^{m\times m}$ *an orthogonal matrix. The subdifferential of the maximum eigenvalue at* $y$ *writes [18, Ex. 3.6]*

$$\partial \lambda_{\max}(y) = \{E_{1:r} Z E_{1:r}^\top, Z \in \mathbb{S}_r, Z \succeq 0, \operatorname{trace} Z = 1\}$$

*where* $r$ *is the multiplicity of the maximum eigenvalue of* $y$*. This function is convex, thus prox-regular and prox-bounded (with parameters* 0*). Its proximity operator can be expressed using the one of the* max *function as*

$$\mathbf{prox}_{\gamma \lambda_{\max}}(y) = E \operatorname{Diag}(\mathbf{prox}_{\gamma \max}(\lambda)) E^\top.$$

**2.2. Structure manifolds.** We now specify the notion of structure manifold in relation with a nonsmooth function $g$.

A subset $\mathcal{M}$ of $\mathbb{R}^n$ is said to be a *$p$-dimensional $\mathcal{C}^2$-submanifold* of $\mathbb{R}^n$ around $\bar{x} \in \mathcal{M}$ if there exists a $\mathcal{C}^2$ *manifold-defining map* $h : \mathbb{R}^n \to \mathbb{R}^{n-p}$ with a surjective derivative at $\bar{x} \in \mathcal{M}$ that satisfies for all $x$ close enough to $\bar{x}$: $x \in \mathcal{M} \Leftrightarrow h(x) = 0$. We define the tangent and normal spaces at a point $x \in \mathcal{M}$ as follows:

$$T_x\mathcal{M} = \ker \mathrm{D}\, h(x) \qquad N_x\mathcal{M} = \operatorname{Im}\ \mathrm{D}\, h(x)^*$$

The important notion of *structure manifolds of* $g$ can be defined as a manifold $\mathcal{M}^g$ where $g$ is nondifferentiable. More precisely, at a point $\bar{y} \in \mathcal{M}^g$, we require $g$ to be prox-regular and *partly smooth*. This property of ($\mathcal{C}^2$-)partial smoothness is verified at a point $\bar{y}$ for a function $g$ relatively to a set $\mathcal{M}^g$ containing $\bar{y}$ if $\mathcal{M}^g$ is a $\mathcal{C}^2$ manifold around $\bar{y}$ and if

- (smoothness) the restriction of $g$ to $\mathcal{M}^g$ is a $\mathcal{C}^2$ function near $\bar{y}$;
- (regularity) $g$ is (Clarke) regular at all points $y \in \mathcal{M}^g$ near $\bar{y}$, with $\partial g(y) \neq \emptyset$;
- (sharpness) the affine span of $\partial g(\bar{y})$ is a translate of $N_{\bar{y}}\mathcal{M}^g$;
- (sub-continuity) the mapping $\partial g$ restricted to $\mathcal{M}^g$ is continuous at $\bar{y}$.

The concept of partial smoothness, introduced in [18], captures (locally) well-behaved nonsmoothness by requiring $g$ to be smooth along a manifold and non-smooth across it. In addition, the prox-regularity of $g$ ensures unicity of the structure manifold near $\bar{y}$ [11, Corollary 4.2, Example 7.1]. To highlight the relation between the manifold and the function $g$, we use the notation $\mathcal{M}^g$ for the structure manifold related to $g$.

EXAMPLE 2.5. *The structure manifolds of* max *are*

$$\mathcal{M}_I^{\max} = \{y \in \mathbb{R}^m : y_i = \max(y) \text{ for } i \in I\},$$

*where* $I \subset \{1, \dots, m\}$. *A smooth manifold-defining map for* $\mathcal{M}_I^{\max}$ *is* $h : \mathbb{R}^m \to \mathbb{R}^{|I|-1}$ *such that* $h(y)_l = y_{i_l} - y_{i_{|I|}}$, *where* $|I|$ *denotes the size of* $I$ *and* $i_l$ *the* $l$*-th element of* $I$ *(with some ordering). As required, this map is surjective. At any point* $y \in \mathbb{R}^m$, *the maximum is partly smooth relative to* $\mathcal{M}_I^{\max}$, *where* $I = \{i : y_i = \max(y)\}$.

EXAMPLE 2.6. *The structure manifolds of* $\lambda_{\max}$ *in* $\mathbb{S}_m$ *consist of all matrices having a largest eigenvalue with fixed multiplicity* $r$*:*

$$\mathcal{M}_r^{\lambda_{\max}} = \{y \in \mathbb{S}_m : \lambda_1(y) = \cdots = \lambda_r(y)\}.$$

*A manifold-defining map of* $\mathcal{M}_r^{\lambda_{\max}}$ *is described in [28] and* $\lambda_{\max}$ *is partly smooth relative to* $\mathcal{M}_r^{\lambda_{\max}}$ *at any point* $y \in \mathcal{M}_r^{\lambda_{\max}}$.

In view of the expression of the proximity operators in our examples, their output naturally lie on the structure manifolds described above. More precisely, $\mathbf{prox}_{\gamma \max}(y)$ belongs to the structure manifold $\mathcal{M}_I^{\max}$, where $I$ collects the indices of the $k$ largest entries of $y$ and $k$ grows as $\gamma$ increases. Similarly, $\mathbf{prox}_{\gamma\lambda_{\max}}(y)$ belongs to the structure manifold $\mathcal{M}_r^{\lambda_{\max}}$, where $r$ increases as $\gamma$ does. This observation is at the core of the ability of proximal operators to identify neighboring structure manifolds.

**2.3. Structure Identification.** It is well-known that the proximity operator identifies structure locally around critical points (see *e.g.* [11, Th. 4]): all points near a minimizer are mapped to the manifold containing the minimizer. Furthermore, this structure is revealed during the computation of the operator.[1]

In the situation we consider, the proximity operator of $F$ cannot be explicitly computed. However, $\mathbf{prox}_{\gamma g}$ is available and can provide some structure in the intermediate space $\mathbb{R}^m$ that we would like to exploit. To do so, we introduce two properties (holding on our two running examples), that will allow us to retrieve the structural information in the intermediate space near points that are not minimizers of $g$.

The first property holds at point $\bar{y} \in \mathcal{M}^g$ if the nonsmooth function $g$ strictly increases on all directions on which it is nonsmooth.

PROPERTY 2.7 (Normal ascent). *A function* $g$ *satisfies the* normal ascent *property at point* $\bar{y}$ *if* $0$ *lies in the relative interior of the projection of* $\partial g(\bar{y})$ *on the normal space at* $\bar{y}$, *that is:* $0 \in \operatorname{ri} \operatorname{proj}_{N_{\bar{y}}\mathcal{M}^g} \partial g(\bar{y})$.

*Remark* 2.8 (Positive directional derivative). In a "nice" setting where $g$ is Lipschitz continuous and regular at $\bar{y}$, Property 2.7 implies that the directional derivative of $g$ along any normal direction $d \in N_{\bar{y}}\mathcal{M}^g$ is positive. Indeed, in that case one-sided directional derivatives are well-defined [26, p. 358, Th. 9.16], and the derivative along direction $w$ equals $\max_{v \in \partial g(x)} \langle v, w \rangle$. Along a normal direction $d \in N_y\mathcal{M}^g$, by partial smoothness the directional derivative writes $\max_{v_n \in \operatorname{proj}_{N_y\mathcal{M}^g}(\partial g(x))} \langle v_n, d \rangle$. Property 2.7 ensures the existence of $\alpha > 0$ such that $\alpha d \in \operatorname{proj}_{N_y\mathcal{M}^g}(\partial g(x))$, making the derivative positive.

The second property is more technical and controls the velocity of a curve on the manifold $\mathcal{M}^g$.

[1] Computing exactly the *structure* of the output point of the operator, as can be done for the prox, is opposed to merely observing the structure of the output after its computation. This last option is not desirable in our opinion as it entails delicate numerical questions such as testing equality between reals for the maximum, or computing the multiplicity of the maximal eigenvalue of a matrix.

PROPERTY 2.9 (Curve property). *A function $g$ partly smooth at $\bar{y}$ relative to $\mathcal{M}^g$ satisfies the* curve property *at $\bar{y}$ when there exists a neighborhood $\mathcal{N}_{\bar{y}}$ of $\bar{y}$ and $T > 0$ such that any smooth application $e : \mathcal{N}_{\bar{y}} \times [0, T] \to \mathcal{M}^g$ such that $e(y, 0) = \operatorname{proj}_{\mathcal{M}^g}(y)$, $\frac{\mathrm{d}}{\mathrm{d}t} e(y, t)|_{t=0} = -\operatorname{grad} g(\operatorname{proj}_{\mathcal{M}^g}(y))$ satisfies*

$$\|\operatorname{proj}_{N_{e(y,t)}\mathcal{M}^g}(e(y, t) - y)\| \leq \operatorname{dist}_{\mathcal{M}^g}(y) + \tilde{L}\, t^2 \quad \text{for all } y \in \mathcal{N}_{\bar{y}}, t \in [0, T],$$

*where $\operatorname{dist}_{\mathcal{M}^g}(y) \triangleq \|y - \operatorname{proj}_{\mathcal{M}^g}(y)\|$ is the distance between $\mathcal{M}^g$ and $y$, and $\operatorname{grad} g(p) \in T_p\mathcal{M}^g$ denotes the Riemannian gradient of $g$ obtained as $\operatorname{proj}_{T_p\mathcal{M}^g}(\partial g(p))$.*

The idea behind this property is to ensure that the differential of the projection of the (time dependent) normal space is (uniformly) negligible at time 0. Note that for affine spaces, we trivially have $\|\operatorname{proj}_{N_{e(y,t)}\mathcal{M}^g}(y - e(y, t))\| = \operatorname{dist}_{\mathcal{M}^g}(y)$ for all $t$ near 0: the normal spaces are equal at all points of the manifold.

These two properties are satisfied at any structured point for the two nonsmooth functions max and $\lambda_{\max}$ of our running examples as detailed in the following lemma. The proofs for the two functions are rather direct but require precise technical descriptions; we defer them to Appendix A.

LEMMA 2.10. *Consider either:*

- *$g = \max$, $\bar{y} \in \mathbb{R}^m$, and the structure manifold $\mathcal{M}_I^{\max}$ (of Example 2.5);*
- *$g = \lambda_{\max}$, $\bar{y} \in \mathbb{S}_m$, and the structure manifold $\mathcal{M}_r^{\lambda_{\max}}$ (of Example 2.6).*

*Then, Properties 2.7 and 2.9 hold at $\bar{y}$.*

Finally, the structure provided by $\mathbf{prox}_{\gamma g}$ lies in the intermediate space $\mathbb{R}^m$, while the optimization variable lives in $\mathbb{R}^n$. In order to transfer the structure information to the input space, we will also require the smooth map $c : \mathbb{R}^n \to \mathbb{R}^m$ to be transversal to $\mathcal{M}^g \subset \mathbb{R}^m$ at some point $\bar{x} \in \mathbb{R}^n$, which holds when $\mathcal{M}^g$ is a manifold around $c(\bar{x})$ and the following (equivalent) conditions hold:

$$\ker(\mathrm{D}\, c(\bar{x})^*) \cap N_{c(\bar{x})}\mathcal{M}^g = \{0\} \quad \text{or} \quad T_{c(\bar{x})}\mathcal{M}^g + \operatorname{Im}\, \mathrm{D}\, c(\bar{x}) = \mathbb{R}^m.$$

In that case, the set $c^{-1}(\mathcal{M}^g)$ is a submanifold of $\mathbb{R}^n$ [15, Th. 6.30], whose normal space has the same dimension as the one of $\mathcal{M}^g$. Furthermore, we have [15, Ex. 6-10]

$$N_x c^{-1}(\mathcal{M}^g) = \mathrm{D}\, c(x)^* N_{c(x)}\mathcal{M}^g \quad \text{and} \quad T_x c^{-1}(\mathcal{M}^g) = \mathrm{D}\, c(x)^{-1} T_{c(x)}\mathcal{M}^g.$$

**3. Collecting structure with the proximity operator.** We show in this section how to exactly detect the optimal structure manifold of the composite function $F = g \circ c$ around a point $\bar{x}$ using the proximity operator of $g$.

In our nonconvex and nonsmooth setting, we seek only structured points which satisfy certain assumptions summarized in our definition of a *qualified point*.

DEFINITION 3.1 (Qualified points). *A point $\bar{x} \in \mathbb{R}^n$ is* qualified *relative to a decomposition $(g, c)$ of $F$ and manifold $\mathcal{M}^g$ if*

*i) $g$ is prox-bounded and prox-regular at $c(\bar{x})$;*
*ii) $g$ is partly smooth at $c(\bar{x})$ relative to $\mathcal{M}^g$;*
*iii) $c$ is transversal to $\mathcal{M}^g$ at $\bar{x}$;*
*iv) $g$ satisfies Properties 2.7 and 2.9 at point $c(\bar{x})$.*

Three of these assumptions constrain only the nonsmooth function $g$ and are easily verifiable in practice. Only the transversality condition limits the range of acceptable smooth mappings; see *e.g.* [18, Sec. 4]. For such *qualified* points, we get two useful properties: first, $F$ is partly smooth at $\bar{x}$ relative to $\mathcal{M} = c^{-1}(\mathcal{M}^g) \ni \bar{x}$ (by the chain rule of [18, Th. 4.2]) and second, the operator $\mathbf{prox}_{\gamma g}$ is single-valued, locally Lipschitz, and defined by its optimality condition near $c(\bar{x})$.

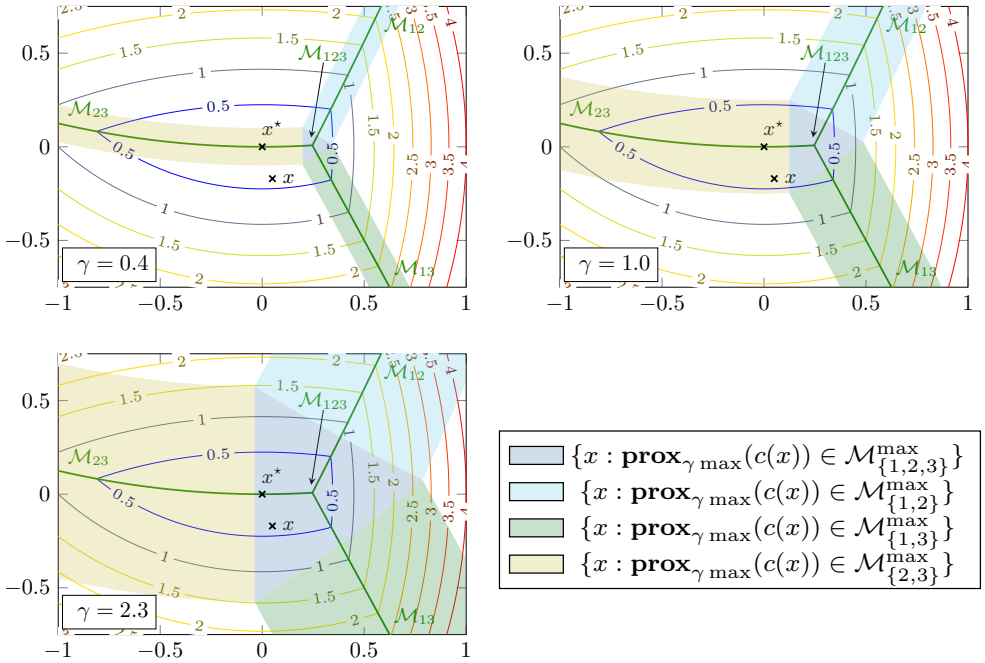



Fig. 2: Illustration of the main result on a maximum of three quadratic functions, with $\bar{x} \in \mathcal{M}^{\max}_{\{1,2\}}$ and a point $\tilde{x}$ near $\bar{x}$. The three figures show the areas where $\mathbf{prox}_{\gamma g} \circ c$ detects manifolds for three stepsizes: $\gamma = 0.4$ (upper left), $\gamma = 1$ (upper right) and $\gamma = 2.3$ (lower left). We see on the upper left fig. that $\mathbf{prox}_{\gamma g} \circ c$ detects no structure from $\tilde{x}$ because $\gamma$ is too small, and in constrast, on the lower fig., that it wrongly detects too much structure ($\mathcal{M}^{\max}_{\{1,2,3\}}$) because $\gamma$ is too large. On the upper right fig., the optimal manifold is detected with $\gamma$ chosen in the right interval.

**3.1. Main result.** We show in the following theorem that if $x$ is near a qualified point of $F$ with structure $\mathcal{M}$, then $\mathbf{prox}_{\gamma g}(c(x))$ will output a point on $\mathcal{M}^g = c(\mathcal{M})$, the structure manifold of $g$ corresponding to $\mathcal{M}$ (in the intermediate space). Our theorem provides precise conditions on $x$ and $\gamma$ that guarantee this structure identification and forms the main theoretical contribution of the paper. We illustrate this behavior in Figure 2.

The position of this result with respect to the literature is discussed right after in Remark 3.3, and the proof is given in the following Subsection 3.2, in a succession of technical lemmas. We stress that we give guarantees on the *structure* to which the point $\mathbf{prox}_{\gamma g}(c(x))$ belongs, rather than on the point itself.

THEOREM 3.2. *Consider a function $F = g \circ c$ and a point $\bar{x}$. Assume that $\bar{x}$ is* qualified *relative to a manifold $\mathcal{M}^g \subset \mathbb{R}^m$. Then, there exists a neighborhood $\mathcal{N}_{\bar{x}}$ of $\bar{x}$ and a constant $\Gamma$ such that, for all $x \in \mathcal{N}_{\bar{x}}$,*

$$\mathbf{prox}_{\gamma g}(c(x)) \in \mathcal{M}^g \text{ for all } \gamma \in [\varphi(\mathrm{dist}_{\mathcal{M}}(x)), \Gamma],$$

*where* $\operatorname{dist}_{\mathcal{M}}(x)$ *denotes the distance from* $x$ *to the manifold* $\mathcal{M}$ *and* $\varphi$ *is defined as*

$$\varphi(t) = \frac{c_{ri}}{2\tilde{L}}\left(1 - \sqrt{1 - \frac{4\tilde{L}c_{map}t}{c_{ri}^2}}\right) = \frac{c_{map}}{c_{ri}}t + \frac{\tilde{L}c_{map}^2}{c_{ri}^3}t^2 + o(t^2),$$

*with* $c_{ri}$, $c_{map}$, *and* $\tilde{L}$ *(of Property 2.9) positive constants.*

*In particular, there exists* $L > 0$, $\epsilon > 0$ *such that*

$$\|x - x^\star\| \leq \epsilon \text{ and } L\|x - x^\star\| \leq \gamma \leq \Gamma \Longrightarrow \mathbf{prox}_{\gamma g}(c(x)) \in \mathcal{M}^g.$$

Note that Property 2.9 is only used to compute explicitly an interval of $\gamma$ guaranteed to provide the correct structure; the existence of that interval holds independently.

*Remark* 3.3 (Relation with existing results). The difference between Theorem 3.2 and existing results lies in two aspects. First, the identification properties of the proximal operator [8, Th. 28], the proximal-gradient operator [1, Th. 3.1], or even approximate prox-gradient operators [14] give structure information directly in the input space (even in abstract algorithmic frameworks [11, Th. 4]). In the composite case, the proximity operator reveals structure in the intermediate space only, and extra work is required to bring it back to the input space.

Second, most existing results investigate identification properties near minimizers, and not just arbitrary points (a notable exception is [1] in a different context). Here, we evaluate $\mathbf{prox}_{\gamma g}$ near $c(\bar{x})$, a point without any specific properties even if $\bar{x}$ is a local minimizer. This is why we need Property 2.7 to guarantee identification in the intermediate space, and bring the structure information to the input space.

**3.2. Proof of Theorem 3.2.** The proof consists in giving conditions on $y$ and $\gamma$ so that $p = \mathbf{prox}_{\gamma g}(y)$ lies on the manifold. We characterize this relation by the first-order optimality condition:

$$y \in p + \gamma \partial g(p).$$

We first show that, for $y$ near $\bar{y}$ and $\gamma$ small, there is a unique point $p$ *on the manifold* $\mathcal{M}^g$ that satisfies the tangent component of this optimality condition:

$$\operatorname{proj}_{T_p\mathcal{M}^g}(y - p) = \gamma \operatorname{grad} g(p), \tag{3.1}$$

where $\operatorname{grad} g(p) \triangleq \operatorname{proj}_{T_p\mathcal{M}^g} \partial(g(p))$ is unique by the sharpness property of partial smoothness, and matches the Riemannian gradient of $g$ on $\mathcal{M}^g$ (see [6, Sec. 7.7]). Such points $p$ are given by a smooth manifold-valued application $e(y, \gamma)$, the existence of which is guaranteed by the following lemma.

LEMMA 3.4. *Consider a function* $g : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$, *a point* $\bar{y} \in \mathbb{R}^m$, *and a manifold* $\mathcal{M}^g$ *with* $g$ *partly smooth at* $\bar{y}$ *relative to* $\mathcal{M}^g$. *Then, there exists a smooth curve* $e : \mathcal{N}_{\bar{y}} \times \mathcal{N}_0 \to \mathcal{M}$ *defined on a neighborhood of* $(\bar{y}, 0)$ *in* $\mathbb{R}^m \times \mathbb{R}_+$ *such that*

- *for all* $y \in \mathcal{N}_{\bar{y}}$, $e(y, 0) = \operatorname{proj}_{\mathcal{M}^g}(y)$ *and* $\frac{\mathrm{d}}{\mathrm{d}\gamma} e(y, \gamma)|_{\gamma=0} = -\operatorname{grad} g(\operatorname{proj}_{\mathcal{M}^g}(y))$;
- *for all* $y \in \mathcal{N}_{\bar{y}}$, $\gamma \in \mathcal{N}_0$, *Eq. (3.1) is satisfied for* $p = e(y, \gamma)$.

*Proof.* We define the mapping $\Phi : \mathbb{R}^m \times \mathbb{R} \times \mathcal{M}^g \to \cup_{x \in \mathcal{M}^g} T_x\mathcal{M}^g$ as

$$\Phi(y, \gamma, p) = \gamma \operatorname{grad} g(p) - \operatorname{proj}_{T_p\mathcal{M}^g}(y - p)$$

and consider the equation $\Phi(y, \gamma, p) = 0$ near the point $(\bar{y}, 0, \bar{y})$. Using the smoothness of $g$ on $\mathcal{M}^g$ given by partial smoothness, we have that this mapping is continuously

differentiable on a neighborhood of $(\bar{y}, 0, \bar{y})$. We see that its differential with respect to $p$ is $\mathrm{D}_p\,\Phi(\bar{y}, 0, \bar{y}) = I$. Indeed, for $\eta \in T_p\mathcal{M}^g$,

$$\mathrm{D}_p\,\Phi(y, \gamma, p)[\eta] = \gamma\,\mathrm{Hess}\,g(p)[\eta] + \eta - \mathrm{D}_{p'}\left(p' \mapsto \mathrm{proj}_{T_{p'}\mathcal{M}^g}(y - p)\right)(p)[\eta].$$

At point $(\bar{y}, 0, \bar{y})$, the first term vanishes, and the third term writes

$$\mathrm{D}_{p'}\left(p' \mapsto \mathrm{proj}_{T_{p'}\mathcal{M}^g}(0)\right)(\bar{y})[\eta]$$

and vanishes as well as the differential of the null function $p' \mapsto \mathrm{proj}_{T_{p'}\mathcal{M}^g}(0)$. Thus $\mathrm{D}_p\,\Phi(\bar{y}, 0, \bar{y}) = I$ is invertible. The implicit functions theorem thus grants the existence of neighborhoods $\mathcal{N}^1_{\bar{y}}$, $\mathcal{N}^2_0$, $\mathcal{N}^3_{\bar{y}}$ of $\bar{y}$, $0$, $\bar{y}$ in $\mathbb{R}^m$, $\mathbb{R}$, $\mathcal{M}^g$ and a continuously differentiable function $c : \mathcal{N}^1_{\bar{y}} \times \mathcal{N}^2_0 \to \mathcal{N}^3_{\bar{y}}$ such that, for any $(y, \gamma) \in \mathcal{N}^1_{\bar{y}} \times \mathcal{N}^2_0$, Equation (3.1) is satisfied with $p = e(y, \gamma)$. For $y \in \mathcal{N}^1_{\bar{y}}$, $e(y, 0)$ satisfies $y - e(y, 0) \in N_{e(y,0)}\mathcal{M}^g$, which is the first-order optimality condition of $e(y, 0) = \mathrm{proj}_{\mathcal{M}^g}(y)$. Possibly reducing $\mathcal{N}^1_{\bar{y}}$ so that, for all $y \in \mathcal{N}_{\bar{y}}$ $\mathrm{proj}_{\mathcal{M}^g}(y)$ is well-defined and unique, the previous optimality condition is equivalent to $e(y, 0) = \mathrm{proj}_{\mathcal{M}^g}(y)$. Besides, differentiating $\Phi(y, \gamma, e(y, \gamma)) = 0$ relative to $\gamma$ at $\gamma = 0$ yields

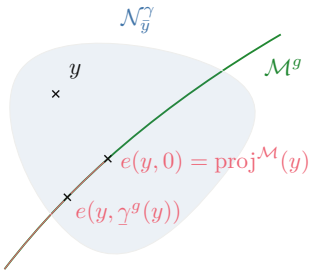
$$\mathrm{D}_\gamma\,e(y, 0) = -[\mathrm{D}_p\,\Phi(y, 0, \mathrm{proj}_{\mathcal{M}^g}(y))]^{-1}\,\mathrm{D}_\gamma\,\Phi(y, 0, \mathrm{proj}_{\mathcal{M}^g}(y)) = -\,\mathrm{grad}\,g(\mathrm{proj}_{\mathcal{M}^g}(y)),$$

which concludes the proof. ∎

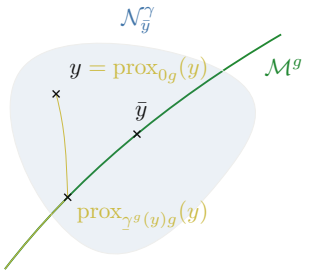
The previous lemma shows that for every $(y, \gamma)$ one can find a point $e(y, \gamma)$ *on the manifold* $\mathcal{M}^g$ that solves the tangent part of the optimality condition (3.1). The next lemma determines the values of $y$ and $\gamma$ for which the whole optimality condition

$$y \in e(y, \gamma) + \gamma\,\mathrm{ri}\,\partial g(e(y, \gamma)) \tag{3.2}$$

holds, as illustrated in Figure 3a.




(a) the curve $\gamma \mapsto e(y, \gamma)$ on $\mathcal{M}$.




(b) the curve $\gamma \mapsto \mathbf{prox}_{\gamma g}(y)$ on $\mathcal{M}$ for $\gamma \geq \varphi(\mathrm{dist}_{\mathcal{M}^g}(y))$.

Fig. 3: Illustration of Lemma 3.5 and its consequences.

LEMMA 3.5. *Consider a function $g$, a point $\bar{y} \in \mathbb{R}^m$ and a manifold $\mathcal{M}^g$ such that $g$ is partly smooth at $\bar{y}$ relative to $\mathcal{M}^g$ and that $g$ satisfies Property 2.7 at $\bar{y}$. Let $e$ denote a smooth $\mathcal{M}$-valued application defined on a neighborhood of $(\bar{y}, 0)$ provided by Lemma 3.4. Then, there exists $C > 0$ such that:*

*i) for all $\gamma \in [0, C]$, $e(\bar{y}, \gamma)$ verifies (3.2) with $y = \bar{y}$,*

*ii) for all $\gamma \in [0, C]$, there exists a neighborhood $\mathcal{N}^\gamma_{\bar{y}}$ of $\bar{y}$ such that, for all $y \in \mathcal{N}^\gamma_{\bar{y}}$, $e(y, \gamma)$ verifies (3.2),*

*Further assume that $g$ satisfies Property 2.9 at $\bar{y}$ with constant $\tilde{L}$, then*

*iii) there exist $\Gamma^g > 0$ and a neighborhood $\mathcal{N}_{\bar{y}}$ of $\bar{y}$ such that for all $y \in \mathcal{N}_{\bar{y}}$*

$$e(y,\gamma) \text{ verifies (3.2) for all } \gamma \in [\varphi^g(\operatorname{dist}_{\mathcal{M}^g}(y)), \Gamma^g],$$

$$\text{where } c_{ri} \geq 0 \text{ and } \varphi^g(t) = \tfrac{c_{ri}}{2\tilde{L}}\left(1 - \sqrt{1 - \tfrac{4\tilde{L}t}{c_{ri}^2}}\right) = \tfrac{1}{c_{ri}}t + \tfrac{\tilde{L}}{c_{ri}^3}t^2 + o(t^2).$$

The proof consists in finding the points $y, \gamma$ such that $0 \in \operatorname{ri}\Psi(y,\gamma)$, where the mapping $\Psi : \mathbb{R}^m \times \mathbb{R} \to \cup_{x\in\mathcal{M}^g} N_x\mathcal{M}^g$ is defined as

$$\Psi(y,\gamma) = \operatorname{proj}_{N_{e(y,\gamma)}\mathcal{M}^g}\left(\frac{1}{\gamma}\left(e(y,\gamma) - y\right) + \partial g\left(e(y,\gamma)\right)\right).$$

Items *i)* and *ii)* are shown by extending the property $0 \in \Psi(\bar{y}, 0)$ to a neighborhood of $(\bar{y}, 0)$, using the inner-semicontinuity properties of $\Psi$. We then derive explicit bounds on the interval of steps such that $0 \in \operatorname{ri}\Psi(y,\gamma)$: for a fixed $y \in \mathcal{N}_{\bar{y}}$, when $\gamma$ decreases past some value, say $\underline{\gamma}(y)$, the condition $0 \in \operatorname{ri}\Psi(y,\gamma)$ no longer holds. Precisely at $\underline{\gamma}(y)$, 0 lies on the (relative) boundary of $\Psi(y,\underline{\gamma}(y))$: denoting $\operatorname{rbd} S \triangleq S \setminus \operatorname{ri} S$ the relative boundary of set $S$,

$$0 \in \operatorname{rbd}\operatorname{proj}_{N_{e(y,\underline{\gamma}(y))}\mathcal{M}^g}\left(\frac{1}{\underline{\gamma}(y)}\left(e(y,\underline{\gamma}(y)) - y\right) + \partial g\left(e(y,\underline{\gamma}(y))\right)\right).$$

Denoting $\partial^N g(p) \triangleq \operatorname{proj}_{N_p\mathcal{M}^g}(\partial g(p))$ the projection of the subdifferential on the normal space of its structure manifold and taking norms yields:

$$\begin{aligned}\|\operatorname{proj}_{N_{e(y,\underline{\gamma}(y))}\mathcal{M}^g}(y - e(y,\underline{\gamma}(y)))\| &\geq \underline{\gamma}(y) \inf_{v_n \in \operatorname{rbd}\partial^N g(e(y,\underline{\gamma}(y)))} \|v_n\| \\ &\geq \underline{\gamma}(y) \underbrace{\inf_{p\in\mathcal{N}_{\bar{y}}} \inf_{v_n\in\operatorname{rbd}\partial^N g(p)} \|v_n\|}_{\triangleq c_{\mathrm{ri}}}.\end{aligned}$$

Since $0 \in \operatorname{ri}\operatorname{proj}_{N_{\bar{y}}\mathcal{M}^g}\partial g(\bar{y})$ and $\partial g$ is inner-semicontinuous, the former property actually holds on a neighborhood of $\bar{y}$ in $\mathcal{M}^g$, thus making the constant $c_{\mathrm{ri}}$ positive. We note that this kind of quantity also appears as the *modulus of identifiability* in the recent [16, Def. 2.3] where it has the same property: its positivity enables the identification of the associated structure manifold.

Using Property 2.9, the left-hand side is upper bounded by a simpler expression:

$$\tilde{L}\underline{\gamma}(y)^2 + \operatorname{dist}_{\mathcal{M}^g}(y) \geq c_{\mathrm{ri}}\underline{\gamma}(y), \quad \text{that is} \quad \underline{\gamma}(y) \leq \frac{c_{\mathrm{ri}}}{2\tilde{L}}\left(1 - \sqrt{1 - \frac{4\tilde{L}\operatorname{dist}_{\mathcal{M}^g}(y)}{c_{\mathrm{ri}}^2}}\right),$$

which provides the expression for $\varphi^g$ used in the lemma.

*Proof. Item i)* We first consider $\Psi_{\bar{y}}(\cdot) = \Psi(\bar{y}, \cdot)$. Since $\bar{y} \in \mathcal{M}^g$, Lemma 3.4 tells us that $e(\bar{y},\gamma) = \bar{y} - \gamma\operatorname{grad} g(\bar{y}) + o(\gamma)$, and thus

$$\Psi_{\bar{y}}(0) = \operatorname{proj}_{N_{\bar{y}}\mathcal{M}^g}\left(-\operatorname{grad} g(\bar{y}) + \partial g(\bar{y})\right) = \operatorname{proj}_{N_{\bar{y}}\mathcal{M}^g}(\partial g(\bar{y}))$$

where we used that $\operatorname{grad} g(\bar{y}) \in T_{\bar{y}}\mathcal{M}^g$ is orthogonal to $N_{\bar{y}}\mathcal{M}^g$.

Property 2.7 provides that $0 \in \operatorname{ri} \Psi_{\bar{y}}(0)$. We now turn to showing that there exists $C'$ such that, for all $\gamma \in [0, C']$, $0 \in \operatorname{ri} \Psi_{\bar{y}}(\gamma)$.

By contradiction, assume there exist a sequence $\gamma_k \to 0$ such that $0 \notin \operatorname{ri} \Psi_{\bar{y}}(\gamma_k)$. This means that there exists a sequence of unit norm vectors $(s_k)$ such that for all $k$,

$$\langle s_k, z \rangle \leq 0 \text{ for all } z \in \Psi_{\bar{y}}(\gamma_k). \tag{3.3}$$

As a bounded sequence, $s_k$ admits at least one limit point, say $\bar{s}$. Take $\bar{z} \in \Psi_{\bar{y}}(0)$. The continuity of $\partial g$ (by partial smoothness, item iv), of $\gamma \mapsto (e(\bar{y}, \gamma) - \bar{y})/\gamma$ (by smoothness of $e$), and of $\gamma \mapsto \operatorname{proj}_{N_{e(\bar{y},\gamma)}\mathcal{M}^g}$ (by smoothness of $\mathcal{M}^g$) yield the continuity of $\Psi_{\bar{y}}$ as a set-valued map. This mapping is thus inner-semicontinuous [26, Def. 5.4], so there exists a sequence $z_k \in \Psi_{\bar{y}}(\gamma_k)$ such that $z_k$ converges to $\bar{z}$. Taking the correct subsequence and renaming iterates, we can write $s_k \to \bar{s}$ and $z_k \to \bar{z}$. Equation (3.3) provides $\langle s_k, z_k \rangle \leq 0$ for all $k$, which gives at the limit $\langle \bar{s}, \bar{z} \rangle \leq 0$. This actually holds for all $\bar{z} \in \Psi_{\bar{y}}(0)$: $\bar{s}$ separates 0 and $\Psi(0)$, which contradicts $0 \in \operatorname{ri} \Psi_{\bar{y}}(0)$.

Finally, let us take the constant $C$ such that $[0, C]$ is included in $[0, C']$ and the neighborhood of 0 provided by Lemma 3.4. Then, for any $\gamma \in [0, C]$, adding the two orthogonal inclusions $0 \in \operatorname{ri} \Psi_{\bar{y}}(\gamma)$ and $0 = \Phi(y, \gamma, c(y, \gamma))$, we obtain that $e(\bar{y}, \gamma)$ verifies (3.2) with $y = \bar{y}$.

*Item ii)* Let $\gamma \in [0, C]$. We turn to show the existence of a neighborhood $\mathcal{N}_{\bar{y}}^{\gamma}$ of $\bar{y}$ such that, for all $y \in \mathcal{N}_{\bar{y}}^{\gamma}$, $e(y, \gamma)$ verifies (3.2). By contradiction, assume that there exists a sequence $(y_k)$ that converges to $\bar{y}$ such that (3.2) fails for $(y_k, \gamma)$. Since the tangent component of (3.2) does hold, necessarily $0 \notin \operatorname{ri} \Psi(y_k, \gamma)$. However, the mapping $y \mapsto \Psi(y, \gamma)$ is inner-semicontinuous (from the same arguments as in the proof of *item i)* and there holds $0 \in \operatorname{ri} \Psi(\bar{y}, \gamma)$. A reasoning similar to that of *item i)* reveals the contradiction.

*Item iii)* Define $\mathcal{N}_{\bar{y}}$ a neighborhood of $\bar{y}$ and $\Gamma^g$ a positive constant such that Property 2.9 applies over $\mathcal{N}_{\bar{y}} \times [0, \Gamma^g]$, $\mathcal{N}_{\bar{y}}$ is contained in $\cup_{\gamma \in [0,C]} \mathcal{N}_{\bar{y}}^{\gamma} \cap \mathcal{N}_{\bar{y}}^{C}$ and $0 \in \operatorname{ri} \Psi(y, \gamma)$ holds for all $(y, \gamma) \in \mathcal{N}_{\bar{y}} \times [0, \Gamma^g]$. This last condition can be met on a nontrivial neighborhood of $(\bar{y}, 0)$ since it holds at that point, and $\Psi$ is inner-semicontinuous ($e(y, \gamma)$ lies on $\mathcal{M}^g$ and $\partial g$ is inner-semicontinuous by partial smoothness of $g$).

Let $y \in \mathcal{N}_{\bar{y}}$ and $\gamma > 0$ such that $\varphi^g(\operatorname{dist}_{\mathcal{M}^g}(y)) \leq \gamma \leq \Gamma^g$. The lower bound on $\gamma$ implies that $\tilde{L}\gamma^2 + \operatorname{dist}_{\mathcal{M}}(y) < c_{\mathrm{ri}}\gamma$. We have successively by Property 2.9 and the above bound that

$$\begin{aligned} \| \operatorname{proj}_{N_{e(y,\gamma)}\mathcal{M}^g}(y - e(y, \gamma)) \| &\leq \operatorname{dist}_{\mathcal{M}}(y) + \tilde{L}\gamma^2 < c_{\mathrm{ri}}\gamma \\ &\leq \gamma \inf\{\|n\|, n \in \operatorname{rbd} \partial^N g(e(y, \gamma))\}. \end{aligned}$$

This means that $\operatorname{proj}_{N_{e(y,\gamma)}\mathcal{M}^g}(y - e(y, \gamma))$ belongs to the ball of center 0 and radius $\gamma \inf\{\|n\|, n \in \operatorname{rbd} \partial^N g(e(y, \gamma))\}$ in $N_{e(y,\gamma)}\mathcal{M}^g$. In addition, this ball is included in $\partial^N g(e(y, \gamma)$ since $0 \in \partial^N g(e(y, \gamma)$ by definition of $\mathcal{N}_{\bar{y}}$. Therefore, $0 \in \operatorname{ri} \Psi(y, \gamma)$ for all $y \in \mathcal{N}_{\bar{y}}$ and $\gamma \in [\varphi^g(\operatorname{dist}_{\mathcal{M}^g}(y)), \Gamma^g]$. ∎

We can now proceed to the proof of Theorem 3.2. It consists in first identifying the curve $e(y, \gamma)$ to $\mathbf{prox}_{\gamma g}(y)$ and thus prove that it belongs to the sought manifold, as illustrated in Figure 3b. Then, this intermediate identification result is brought back to the input space using transversality.

*Proof.* The standing assumptions allow to call Lemma 3.5 at point $c(\bar{x})$ with manifold $\mathcal{M}^g$. This yields the neighborhood $\mathcal{N}_{c(\bar{x})}$, constants $\Gamma^g$ and $C$, a function

$\varphi^g$, and a smooth mapping $e : \mathcal{N}_{c(\bar{x})} \times [0, C] \to \mathcal{M}^g$ such that, for $y \in \mathcal{N}_{c(\bar{x})}$ and $\gamma \in [\varphi^g(\mathrm{dist}_{\mathcal{M}^g}(y)), \Gamma^g]$, $e(y, \gamma)$ verifies the optimality condition (3.2) of $e(y, \gamma) = \mathbf{prox}_{\gamma g}(y)$. Besides, since $g$ is prox-regular and prox-bounded at point $c(\bar{x})$, these properties also hold on a neighborhood of that point. Under these conditions, Proposition 2.2 allows to recover the equality $e(y, \gamma) = \mathbf{prox}_{\gamma g}(y)$. Take $\mathcal{N}_{\bar{x}} = c^{-1}(\mathcal{N}_{c(\bar{x})})$, a neighborhood of $\bar{x}$ as the preimage of a neighborhood of $c(\bar{x})$ by the continuous $c$. For all $x \in \mathcal{N}_{\bar{x}}$,

$$\mathbf{prox}_{\gamma g}(c(x)) \in \mathcal{M}^g \text{ for all } \gamma \in [\varphi^g(\mathrm{dist}_{\mathcal{M}^g}(c(x))), \Gamma^g].$$

We turn to show that, for some constant $c_{\mathrm{map}} > 0$, there holds $\mathrm{dist}_{\mathcal{M}^g}(c(x)) \leq c_{\mathrm{map}} \mathrm{dist}_{\mathcal{M}}(x)$ for all $x \in \mathcal{N}_{\bar{x}}$. Let $x \in \mathcal{N}_{\bar{x}}$ and $x^{\mathcal{M}} = \mathrm{proj}_{\mathcal{M}}(x)$, so that $\mathrm{dist}_{\mathcal{M}}(x) = \|x^{\mathcal{M}} - x\|$. Using successively that $c(x^{\mathcal{M}}) \in \mathcal{M}^g$ and smoothness of $c$, there holds for $x$ near $\bar{x}$

$$\begin{aligned}
\mathrm{dist}_{\mathcal{M}^g}(c(x)) &\leq \|c(x) - c(x^{\mathcal{M}})\| \\
&\leq \|\mathrm{Jac}_c(x^{\mathcal{M}}) \cdot (x - x^{\mathcal{M}})\| + \mathcal{O}(\|x - x^{\mathcal{M}}\|^2) \\
&\leq \left( \sup_{v_n \in N_{x^{\mathcal{M}}}\mathcal{M}, \|v_n\|=1} \|\mathrm{Jac}_c(x^{\mathcal{M}}) \cdot v_n\| \right) \|x - x^{\mathcal{M}}\| + \mathcal{O}(\|x - x^{\mathcal{M}}\|^2) \\
&\leq \underbrace{\left( \sup_{u \in \mathcal{N}_{\bar{x}}} \sup_{v_n \in N_u\mathcal{M}, \|v_n\|=1} \|\mathrm{Jac}_c(u) \cdot v_n\| \right)}_{C''} \|x - x^{\mathcal{M}}\| + \mathcal{O}(\|x - x^{\mathcal{M}}\|^2).
\end{aligned}$$

Since $c$ is transversal to $\mathcal{M}^g$ at $\bar{x}$, its Jacobian restricted to the normal space is invertible: $C''$ is positive. Therefore, for all $x \in \mathcal{N}_{\bar{x}}$ and a constant $c_{\mathrm{map}} > C''$, there holds $\mathrm{dist}_{\mathcal{M}^g}(c(x)) \leq c_{\mathrm{map}} \mathrm{dist}_{\mathcal{M}}(x)$. Monotony of $\varphi^g$ implies that $\varphi^g(\mathrm{dist}_{\mathcal{M}^g}(c(x))) \leq \varphi^g(c_{\mathrm{map}} \mathrm{dist}_{\mathcal{M}}(x))$. Hence the claimed bounds with

$$\varphi(t) = \frac{c_{\mathrm{ri}}}{2\tilde{L}} \left( 1 - \sqrt{1 - \frac{4\tilde{L} c_{\mathrm{map}} t}{c_{\mathrm{ri}}^2}} \right) \quad \text{and} \quad \Gamma = \Gamma^g.$$

Finally, we show the existence of positive constants $\epsilon$, $L$ such that

$$\|x - \bar{x}\| \leq \epsilon \text{ and } L\|x - \bar{x}\| \leq \gamma \leq \Gamma \Longrightarrow \mathbf{prox}_{\gamma g}(c(x)) \in \mathcal{M}^g.$$

Since $\bar{x} \in \mathcal{M}$, $\mathrm{dist}_{\mathcal{M}}(\cdot) \leq \|\cdot - \bar{x}\|$. By monotony and smoothness of $\varphi$, there exists $L > 0$ such that $\varphi(\mathrm{dist}_{\mathcal{M}^\star}(\cdot)) \leq L\|\cdot - x^\star\|$ over $\mathcal{B}(x^\star, \epsilon)$. Reducing $\epsilon$ if necessary so that $L\epsilon < \Gamma$ yields the result. ∎

**4. Proposed method.** In this section, we use the results of Section 3 to propose an optimization method that locally identifies the structure of a minimizer and converges quadratically to this point.

Recall the basic idea: if the optimal manifold $\mathcal{M}^\star$ corresponding to a minimizer $x^\star$ is known, the *nonsmooth* optimization problem turns into a *smooth constrained* optimization problem. In turn, this problem can be solved using algorithms from smooth constrained optimization such as Sequential Quadratic Programming.

Using this idea and the structure identification mechanism developed in the previous section, we propose a method which: i) uses the proximity operator of $g$ to gather structure in the intermediate space, ii) brings back this structure to the input space, and iii) optimizes smoothly along the identified manifold. The resulting algorithm is precisely described in Subsection 4.1 and then analyzed in Subsection 4.2.

**4.1. Description of the algorithm.** We proceed to describe the three steps exposed above. The full algorithm is depicted in Algorithm 4.1.

*Gathering structure.* We showed in Theorem 3.2 that near a qualified point in $\mathbb{R}^n$, the operator $\mathbf{prox}_{\gamma g}(c(\cdot))$ provides the optimal structure $\mathcal{M}^{g\star}$ (in the intermediate space $\mathbb{R}^m$) for an explicit range of steps. We thus define from the current iterate $x_k \in \mathbb{R}^n$ and stepsize $\gamma_k$ the working manifold $\mathcal{M}_k^g$ (in the intermediate space) as the structure of $\mathbf{prox}_{\gamma_k g}(c(x_k))$. One technical point is to guarantee that, after some time, $\gamma_k \in [L\|x_k - x^\star\|, \Gamma]$ so that the optimal manifold is identified; this is done by decreasing $\gamma_k$ linearly at each iteration.

*From the intermediate to the input space.* We now have a structure manifold $\mathcal{M}_k^g$ in the intermediate space, and can define $\tilde{g}_k$, a smooth extension of $g$ on $\mathcal{M}_k^g$ to $\mathbb{R}^m$. Using a local equation $h_k^g$ of $\mathcal{M}_k^g$, we define the smooth map $h_k = h_k^g \circ c : \mathbb{R}^n \to \mathbb{R}^{p_k}$, which locally defines $\mathcal{M}_k = c^{-1}(\mathcal{M}_k^g)$. Similarly, a smooth extension of $F$ on $\mathcal{M}_k$ is defined by $\tilde{F}_k = \tilde{g}_k \circ c$.

*Optimizing in the input space.* We can now take steps to minimize the smooth extension $\tilde{F}_k$ on the smooth set $\mathcal{M}_k$ characterized by $h_k(x) = 0$:

$$\min_{x\in\mathbb{R}^n} \tilde{F}_k(x) \quad \text{s.t.} \quad h_k(x) = 0.$$

We turn to an elementary version of the traditional second-order Sequential Quadratic Programming methodology; see *e.g.* [5, Chap. 14]. At iteration $k$, the SQP direction $d_k^{\mathrm{SQP}}(x_k)$ at point $x_k$ is defined as the solution of the following quadratic problem:

$$\begin{aligned} d_k^{\mathrm{SQP}}(x_k) = \operatorname{argmin}_{d\in\mathbb{R}^n} \quad & \langle \nabla \tilde{F}_k(x_k), d\rangle + \frac{1}{2}\langle \nabla^2_{xx} L_k(x_k, \lambda_k(x_k))d, d\rangle \\ \text{s.t.} \quad & h_k(x_k) + \mathrm{D}\, h_k(x_k)d = 0 \end{aligned} \tag{4.1}$$

where $\nabla^2_{xx} L_k$ denotes the Hessian of the Lagrangian $L_k(x,\lambda) = \tilde{F}_k(x) + \langle \lambda, h_k(x)\rangle$, and the multiplier $\lambda_k(x_k)$ defined from the following least-squares problem:

$$\lambda_k(x_k) = \operatorname{argmin}_{\lambda\in\mathbb{R}^{p_k}} \left\| \nabla \tilde{F}_k(x_k) + \sum_{i=1}^{p_k} \lambda_i \nabla h_{k,i}(x_k) \right\|^2 . \tag{4.2}$$

Finally, we check that $x_k + d_k^{\mathrm{SQP}}(x_k)$ provides a functional decrease in order to avoid degrading the iterate when the current structure is suboptimal. If the test is not verified, $x_k$ is not updated and $\gamma_k$ is decreased until a satisfying structure is detected.

**4.2. Convergence of Algorithm 4.1.** We proceed to give the result guaranteeing identification and local quadratic convergence of Algorithm 4.1.

In order to benefit from the quadratic rate of SQP, the elements of (4.1) should have the minimal regularity typically required by smooth constrained Newton methods (see *e.g.* [5, Th. 14.5]); we thus make the following assumption.

ASSUMPTION 4.1 (Regularity of functions). *The smooth extension and the manifold defining map are $\mathcal{C}^2$ with Lipschitz second derivatives, and the Jacobian of the constraints is full rank near the solution.*

In order to focus on the algorithmic originality of the method, we slightly simplify the situation and make the two following algorithmic assumptions.

ASSUMPTION 4.2 (Nonconvex stability). *The iterates of Algorithm 4.1 remain in the connex component of the sublevel set $\{x : F(x) \leq F(x_0)\}$ that contains $x^\star$.*

**Algorithm 4.1** General structure exploiting algorithm

**Require:** Pick $x_0$ near a minimizer, $\gamma_0$ large enough.

**repeat**
  $\gamma_k = \frac{\gamma_{k-1}}{2}$
  Compute $\mathbf{prox}_{\gamma_k g}(c(x_k))$ and obtain $\mathcal{M}_k^g$ locally defined by $h_k^g$
  $h_k = h_k^g \circ c$ (local equation of $\mathcal{M}_k$), $\tilde{F}_k = \tilde{g}_k \circ c$ (smooth extension)
  Compute $d_k^{\mathrm{SQP}}(x_k)$ by solving (4.1)
  **if** $F(x_k + d_k^{\mathrm{SQP}}(x_k)) \leq F(x_k)$ **then**
    $x_{k+1} = x_k + d_k^{\mathrm{SQP}}(x_k)$
  **else**
    $x_{k+1} = x_k$
**until** stopping criterion

This assumption ensures that an update that decreases the functional value remains in the neighborhood of the minimizer $x^\star$. It is naturally satisfied when $F$ is convex, or when $x^\star$ is a global minimizer of $F$ and $x_0$ is close enough to $x^\star$.

ASSUMPTION 4.3 (No Maratos effect). *The iterates of Algorithm 4.1 are such that a step $d$ that makes $x+d$ quadratically closer to $x$ yields descent: $F(x+d) \leq F(x)$.*

In smooth constrained optimization, getting closer (even at quadratic rate) to a minimizer does not imply decrease of objective value and constraint violation (measured by a merit function). This so-called Maratos effect (see *e.g.* [5]) is one of the main difficulties in globalizing SQP schemes, which is out of the scope of the current paper. We thus assume this effect does not affect our algorithm in theory, and use in practice one of the successful refinements, as discussed in Subsection 5.2.

We are now ready for the main convergence result of Algorithm 4.1, which establish that, after some finite time, the iterates identify exactly the optimal manifold and converge to the minimizer at a quadratic rate.

THEOREM 4.4 (Exact identification and quadratic convergence). *Consider a function $F = g \circ c$ and $x^\star$ a strong minimizer,[2] qualified relative to the optimal manifold $\mathcal{M}^\star$. Assume that the smooth extension $\tilde{F}$ of $F$ relative to $\mathcal{M}^\star$ and the corresponding manifold defining map $h$ satisfy Assumption 4.1.*

*If $x_0$ and $F(x_0)$ are close enough to $x^\star$ and $F(x^\star)$, $\gamma_0$ is large enough and the simplifying algorithmic Assumptions 4.2 and 4.3 hold, then there exists $C > 0$ such that the iterates $(x_k, \mathcal{M}_k)$ generated by Algorithm 4.1 verify:*

$$\mathcal{M}_k = \mathcal{M}^\star \quad \text{and} \quad \|x_{k+1} - x^\star\| \leq C\|x_k - x^\star\|^2 \quad \text{for all } k \text{ large enough.}$$

The proof of this result consists in two steps. We first show the existence of a neighborhood of initialization on which the proximity operator will eventually identify the optimal manifold, once the stepsize has been sufficiently decreased. From this point onward, we prove that the SQP step provides a quadratic improvement and that the stepsize policy makes the manifold identification stable.

*Proof. Local identification of the optimal structure.* By Theorem 3.2, there exists a ball centered around $x^\star$ of radius $\epsilon_1 > 0$ and two positive constants $L, \Gamma$ such that, for all $x \in \mathcal{B}(x^\star, \epsilon_1)$ and $\gamma \in [L\|x - x^\star\|, \Gamma]$, $\mathbf{prox}_{\gamma g}(c(x))$ belongs to the optimal manifold $\mathcal{M}^{g\star} = c(\mathcal{M}^\star)$.

[2] There exists $\eta > 0$, $\epsilon > 0$ such that $F(x) \geq F(x^\star) + \eta\|x - x^\star\|^2$ for all $x \in \mathcal{B}(x^\star, \epsilon)$.

*Local quadratic convergence of SQP on the optimal structure.* Let us assume that the optimal manifold has been identified. The least square multiplier $\lambda$ is defined by the optimality condition of (4.2):

$$\lambda(x) = -[\mathrm{Jac}_h(x)\,\mathrm{Jac}_h(x)^\top]^{-1}\,\mathrm{Jac}_h(x)\nabla\tilde{F}(x).$$

and since $h$ is smooth and its Jacobian is full-rank near $x^\star$, $\lambda$ is a Lipschitz continuous function near $x^\star$.

Since $x^\star$ is a strong minimizer of $F$, the Hessian of the Lagrangian restricted to the tangent space is positive definite. Indeed, since $x^\star$ is a strong minimizer of $F$ on $\mathcal{M}^\star$, the Riemannian Hessian relative to the optimal manifold is positive definite. With the choice of multiplier (4.2), the Riemannian Hessian is exactly the Hessian of the Lagrangian restricted to the tangent space to $\mathcal{M}^\star$ at $x^\star$ (see [6, Sec. 7.7]), which is thus itself positive definite.

Thus, using the local quadratic convergence of SQP [5, Th. 14.5], we get that there exists a ball centered around $x^\star$ of radius $\epsilon_2 > 0$ such that the SQP step computed at a point $x$ in that neighborhood relative to the optimal manifold provides a quadratic improvement towards $x^\star$. Reducing $\epsilon_2$ if necessary, we can in addition have that the convergence is at least linear with rate $1/2$.

*Initialization, identification, and quadratic convergence.* Let $\epsilon = \min(\epsilon_1, \epsilon_2, \Gamma/(2L))$. We will now show that initializing with $x_0 \in \{x : F(x) \leq F(x^\star) + \eta\epsilon^2\}$ and $\gamma_0 \geq \Gamma$ provides the claimed behavior.

First, the functional decrease test of the algorithm and Assumption 4.3 guarantee that all iterates satisfy $F(x_k) \leq F(x_0)$. Using that $x^\star$ is a strong minimizer, we get that $\eta\|x_k - x^\star\|^2 \leq F(x_k) - F(x^\star) \leq F(x_0) - F(x^\star) \leq \eta\epsilon^2$, and thus that the iterates remain in $\mathcal{B}(x^\star, \epsilon)$.

Second, as $L\|x - x^\star\| \leq \Gamma/2$ for all $x \in \mathcal{B}(x^\star, \epsilon)$ by construction, the fact that $\gamma_0 > \Gamma$ and $(\gamma_k)$ decreases with geometric rate $1/2$ implies that there exists $K$ such that $L\|x_K - x^\star\| \leq \gamma_K \leq \Gamma$.

Now, assume that at iteration $k \geq K$, $L\|x_k - x^\star\| \leq \gamma_k \leq \Gamma$. Since $x_k \in \mathcal{B}(x^\star, \epsilon_1)$, we have from above that $\mathcal{M}^\star$ is identified. Thus, the SQP step is performed relative to the optimal manifold and $x_k + d_k^{\mathrm{SQP}}(x_k)$ brings a linear improvement of factor $1/2$ at least. Assumption 4.2 ensures that $F(x_k + d_k^{\mathrm{SQP}}(x_k)) \leq F(x_k)$ so that $x_{k+1} = x_k + d_k^{\mathrm{SQP}}(x_k)$ and thus

$$L\|x_{k+1} - x^\star\| \leq \frac{L}{2}\|x_k - x^\star\| \leq \frac{\gamma_k}{2} = \gamma_{k+1}.$$

This shows that $L\|x_{k+1} - x^\star\| \leq \gamma_{k+1} \leq \Gamma$, which completes the induction. We get that $\gamma_k \in [L\|x_k - x^\star\|, \Gamma]$ for all $k \geq K$. Finally, we have that for all $k \geq K$, $\mathcal{M}_k = \mathcal{M}^\star$ and $x_{k+1}$ is quadratically closer to $x^\star$ than $x_k$. ∎

*Direct generalizations.* Theorem 4.4 actually holds for any decrease factor of $\gamma_k$ in $(0, 1)$ with the presented SQP update, or actually any superlinearly convergent update (*e.g.* a quasi-Newton type update). The above result is also readily adapted to an update that converges merely linearly, as long as its rate of convergence is faster than that of $\gamma_k$. This opens the possibility of using SQP methods that rely only on first-order information (see *e.g.* [4]).

## 5. Numerical experiments.

In this section, we provide numerical illustrations for our results. Our goal here is twofold:

i) to illustrate the identification of the optimal manifold by the proximity operator near a minimizer as provided by Theorem 3.2;
ii) to demonstrate the applicability of Algorithm 4.1 and observe the quadratic rates predicted by Theorem 4.4 on our running examples.

**5.1. Test problems.** We first consider the minimization of a pointwise maximum of smooth functions (1.2):

$$\min_{x\in\mathbb{R}^n} \max_{i=1,\dots,m} (c_i(x)).$$

We take the celebrated `MaxQuad` instance, where $n = 10$, $m = 5$ and each $c_i$ is quadratic convex, making the whole function $F$ convex [5, p. 153]. In this instance, the optimal manifold is $\mathcal{M}_I^{\max}$ with $I = \{2, 3, 4, 5\}$.

Second, we consider the minimization of the maximum eigenvalue of an affine mapping (1.3):

$$\min_{x\in\mathbb{R}^n} \lambda_{\max}\left(A_0 + \sum_{i=1}^{n} x_i A_i\right).$$

We take $n = 25$ and we generate randomly $n + 1$ symmetric matrices of size 50. In this instance, the multiplicity of the maximum eigenvalue at the minimizer is $r = 3$.

**5.2. Numerical setup.** All the algorithms are implemented in Julia [2]; experiments may be reproduced using the code available online[3].

*Algorithm.* For the initialization of Algorithm 4.1, we set $\gamma_0$ as the smallest $\gamma$ such that $\mathbf{prox}_{\gamma g}(c(x_0))$ has the most structure (*e.g.* if $g = \max$, we increase $\gamma$ until the output of the proximity operator sets all coordinates to the same value). We solve the quadratic subproblem (4.1) providing the SQP step by the reduced system approach presented in [5, p. 133]. Tangent vectors are expressed in an orthonormal basis of the nullspace of the Jacobian of the constraints at the current iterate. At iterate $x_k$, a second-order correction step $d^{\mathrm{corr}}(x_k)$ is added to the SQP step $d^{\mathrm{SQP}}(x_k)$. It is obtained as $d^{\mathrm{corr}}(x_k) = \mathrm{argmin}_{d\in\mathbb{R}^n}\{\|h(x_k) + \mathrm{Jac}_h(x_k)\, d\|,\ \text{s.t. } d \in \mathrm{Im}\ \mathrm{Jac}_h(x_k)^\top\}$. The full-step is thus $x_k + d^{\mathrm{SQP}}(x_k) + d^{\mathrm{corr}}(x_k)$.

*Baselines.* For the two nonsmooth problems, we compare with the nonsmooth BFGS algorithm of [19] (nsBFGS) and the gradient sampling algorithm [7]. The nsBFGS method is not covered by any theoretical guarantees; it is known to perform relatively well in practice, often displaying a linear rate of convergence. In contrast, the Gradient Sampling algorithm generates with probability one a sequence of iterates for which all cluster points are Clarke stationary for $F$ [7, Th. 3.1].[4]

*Oracles.* Traditional methods for nonsmooth optimization, and notably bundle methods, require a first-order oracle:

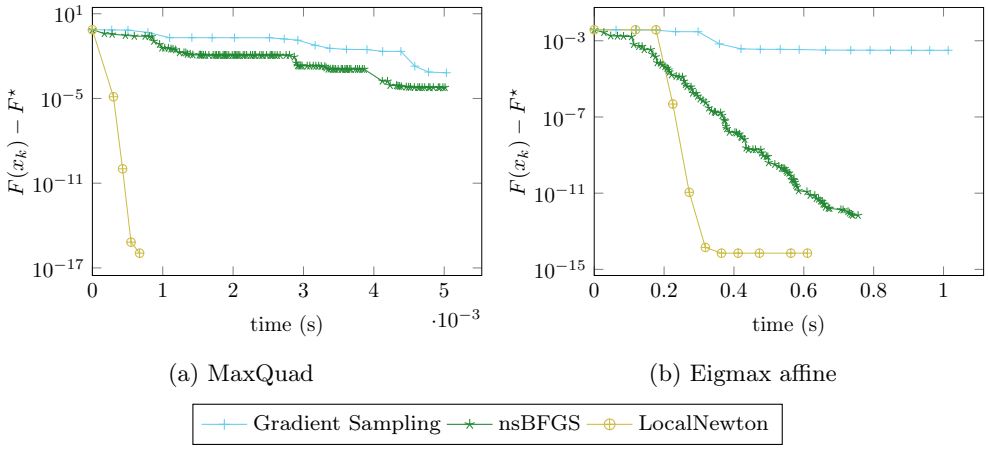
$$x \mapsto (F(x), v) \qquad \text{where } v \in \partial F(x)$$

while Gradient Sampling and nsBFGS require additionally to know if $F$ is differentiable at point $x$. Algorithm 4.1 requires rather different information oracles:

$$x \mapsto F(x)$$
$$x \mapsto \mathcal{M}^g \ni \mathbf{prox}_{\gamma g}(c(x))$$
$$\mathcal{M}, x \mapsto h(x), \mathrm{Jac}_h(x), \nabla\tilde{F}(x), \nabla^2 L(x, \lambda).$$

[3] https://github.com/GillesBareilles/LocalCompositeNewton.jl

[4] This holds when $F$ is locally Lipschitz over $\mathbb{R}^n$ and lower bounded, the algorithm iterates indefinitely and the sampling radius decreases to 0.




(a) MaxQuad (b) Eigmax affine

Fig. 4: Suboptimality vs time (s)

The second part of the oracle provides the candidate structure at point $x$. The last part of the oracle, which requires a point *and a candidate structure*, provides the second-order information of $F$ required by the SQP step.

**5.3. Experiments.** Figure 4 reports the suboptimality of the considered methods in terms of CPU time and each marker corresponds to one iteration. All algorithms are initialized at a point $x_0$ obtained by running nsBFGS for several iterations.

Our algorithm compares favorably to nsBFGS and Gradient Sampling: it converges in a handful of iterations and less time. Note that this happens even though the iteration cost of our algorithm is higher than that of the other methods. Indeed, the oracles of our method are more complex and a quadratic problem needs to be solved, while the iteration cost of nsBFGS and Gradient Sampling is dominated by the computation of function values and subgradients at each trials of the linesearch.

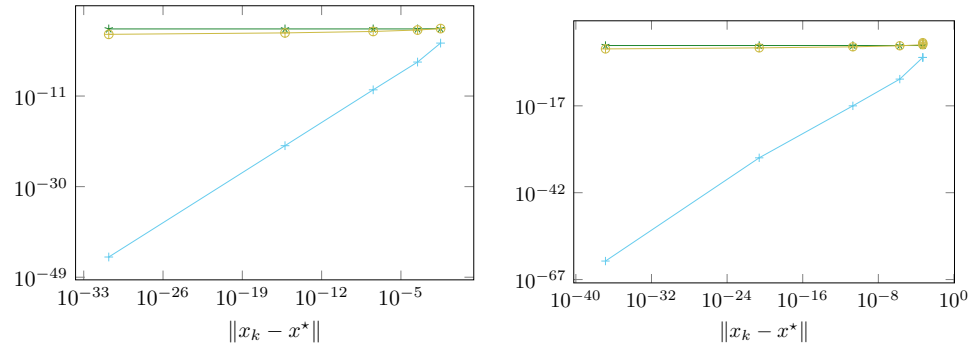
In terms of identification, our method finds the correct manifold at the first iteration for MaxQuad, and at the third iteration for Eigmax. From that point, the iterates of Algorithm 4.1 reach machine precision in 3 iterations. This illustrates the quadratic convergence, and supports the idea that, for nondifferentiable problems as well, it is worth computing higher-order information to get fast local methods.

Figure 5 allows to observe the identification of the algorithm and the quality of the bounds of Theorem 3.2. For each iterate $x_k$ of Algorithm 4.1, we report the current step $\gamma_k$ along with the minimal and maximal steps $\underline{\gamma}(x_k), \bar{\gamma}(x_k)$ such that $\mathbf{prox}_{\gamma g}(c(x_k))$ belongs to the optimal manifold.[5] A first remark is that, as predicted by Theorem 4.4, the pair $x_k, \gamma_k$ satisfies the identification condition $\gamma_k \in [L\|x_k - x^\star, \Gamma]$ after a few iterations. We also observe that $\bar{\gamma}(x_k)$ is near constant and that $\underline{\gamma}(x_k)$ converges to zero linearly with $\|x_k - x^\star\|$, as predicted by our result. Finally, we note that even though the initial point is not structured and away from the minimizer ($\|x_0 - x^\star\| \approx 10^{-2}$), the initialization of $\gamma_0$ ensures a quick identification.

**6. Conclusions.** This paper studies the local structure of functions that write as a composition of a nonsmooth function with a smooth mapping. When the proximity operator of the nonsmooth function is explicitly available, we show that the structure

[5] To better illustrate the local behavior of our method, we also ran the algorithms with a high precision floating type. Details and corresponding experiments can be found in Appendix B.

of the minimizer can be detected. We further use this information to propose a local Newton method to minimize the objective harnessing the detected structure. This method is guaranteed to identify the structure of the minimizer and to converge quadratically. We illustrate this behavior on two standard nonsmooth problems.




(a) MaxQuad (b) Eigmax affine

Fig. 5: Stepsize $\gamma_k$ vs iteration

**Appendix A. The maximum and maximum eigenvalue satisfy the normal ascent and curve properties.** We show here that the maximum and the maximum eigenvalue meet the normal ascent Property 2.7 and curve properties Property 2.9. We begin with a lemma that simplifies verification of Property 2.9.

LEMMA A.1. *Consider a function $g$, partly smooth at a point $\bar{y}$ relative to a manifold $\mathcal{M}^g$, and a smooth application $e : \mathcal{N}_{\bar{y}} \times [0,T] \to \mathcal{M}^g$ defined for a neighborhood $\mathcal{N}_{\bar{y}}$ of $\bar{y}$ and $T > 0$ such that $e(y,0) = \operatorname{proj}_{\mathcal{M}^g}(y)$, $\frac{\mathrm{d}}{\mathrm{d}t} e(y,t)|_{t=0} = -\operatorname{grad} g(\operatorname{proj}_{\mathcal{M}^g}(y))$. If $\mathrm{D}\left(t \mapsto \operatorname{proj}_{N_{e(y,t)}\mathcal{M}^g}(\operatorname{proj}_{\mathcal{M}}(y) - y)\right) = 0$ for all $y \in \mathcal{N}_{\bar{y}}$, then $g$ satisfies Property 2.9 at point $\bar{y}$.*

*Proof.* We denote $\theta(y,t) = \operatorname{proj}_{N_{e(y,t)}\mathcal{M}^g}(e(y,t) - y)$. First,

$$\frac{\mathrm{d}}{\mathrm{d}t}\theta(y,t)|_{t=0} = \mathrm{D}\left(t \mapsto \operatorname{proj}_{N_{e(y,t)}\mathcal{M}^g}(\operatorname{proj}_{\mathcal{M}^g}(y) - y)\right) + \operatorname{proj}_{N_{\operatorname{proj}_{\mathcal{M}}(y)}\mathcal{M}^g}\left(\mathrm{D}(t \mapsto (e(y,t) - y))(0)\right),$$

where the first term is null by assumption and the second is also null since it is the normal projection of the tangent vector $\operatorname{grad} g(\operatorname{proj}_{\mathcal{M}^g}(y))$. Thus, $\frac{\mathrm{d}}{\mathrm{d}t}\theta(y,t)|_{t=0} = 0$. Using this fact and smoothness of $\theta$, Taylor's theorem with Lagrange remainder yields, for all $y \in \mathcal{N}_{\bar{y}}$, the existence of $\bar{t} \in [0,T]$ such that, for all $t \in [0,T]$,

$$\theta(y,t) = \theta(y,0) + \frac{t^2}{2}\frac{\mathrm{d}^2}{\mathrm{d}t^2}\theta(y,\bar{t}).$$

Therefore, for all $y \in \mathcal{N}_{\bar{y}}$ and $t \in [0,T]$,

$$\|\theta(y,t)\| \leq \|\theta(y,0)\| + \frac{t^2}{2}\sup_{\bar{t}\in[0,T]}\frac{\mathrm{d}^2}{\mathrm{d}t^2}\theta(y,\bar{t}) \leq \|\theta(y,0)\| + t^2\tilde{L},$$

where $\tilde{L} = \sup_{y \in \mathcal{N}_{\bar{y}}} \sup_{\bar{t} \in [0,T]} \frac{\mathrm{d}^2}{\mathrm{d}t^2} \theta(y, \bar{t})$. ∎

We can now proceed with the proof of Lemma 2.10, divided into two parts corresponding to the two cases of the result. The case $g = \max$ comes easily, due to the polyhedral nature of the function.

LEMMA A.2. *Consider $g = \max$, a point $\bar{y} \in \mathbb{R}^m$ and the corresponding structure manifold $\mathcal{M}_I^{\max}$ (of Example 2.5). Then Properties 2.7 and 2.9 hold at $\bar{y}$.*

*Proof. Normal ascent* Take $y \in \mathcal{M}_I^{\max}$ for some active indices $I \subset \{1, \dots, m\}$. A normal direction $d \in N_y \mathcal{M}_I^{\max}$ is such that $d_i = 0$ for $i \notin I$ and $\sum_{i \in I} d_i = 0$. Thus $\max(y + td) = y_i + td_i$ with $i = \operatorname{argmax}_i d_i$, and $\mathrm{D} \max(y)[d] = \lim_{t \searrow 0} (\max(y + td) - \max(y))/t = d_i > 0$ for all $d \neq 0$.

*Curve assumption* Since the structure manifold of max are affine subspaces, the normal spaces are equal at all points of the manifold. Therefore the derivative of the projection at a parametrized point is null and Lemma A.1 provides the result. ∎

The case $g = \lambda_{\max}$ is not difficult *per se*, but requires a precise description of the geometry of the maximum eigenvalue function and its structure manifolds; we refer to [28, 25] for the derivation of these tools.

LEMMA A.3. *Consider $g = \lambda_{\max}$, a point $\bar{y} \in \mathbb{S}_m$ and the corresponding structure manifold $\mathcal{M}_r^{\lambda_{\max}}$ (of Example 2.6). Then Properties 2.7 and 2.9 hold at $\bar{y}$.*

*Proof. Normal ascent* Take $y \in \mathcal{M}_r^{\lambda_{\max}}$, let $U \in \mathbb{R}^{m \times r}$ denote a basis of the first eigenspace of matrix $y$ and $d \in N_y \mathcal{M}_r^{\lambda_{\max}}$. The normal space at $y \in \mathcal{M}_r^{\lambda_{\max}}$ writes ([25, Th. 4.3, Cor. 4.8])

$$N_y \mathcal{M}_r^{\lambda_{\max}} = \{U(y) Z U(y)^\top, Z \in \mathbb{S}_r, \operatorname{trace}(Z) = 0\}.$$

Therefore, $d = UZU^\top$ for some $Z \in \mathbb{S}_r$ such that $\operatorname{trace}(Z) = 0$. Let $s = U(I/r + \alpha Z)U^\top$ where $\alpha > 0$ is small enough so that $s$ is positive definite. Since $s$ has also unit trace, it is a subgradient of $\lambda_{\max}$ at $y$ [25, Th. 4.1]. Thus $\lambda'_{\max}(y; d) = \sup_{v \in \partial \lambda_{\max}(y)} \langle v, d \rangle \geq \langle s, d \rangle = \langle I/r + \alpha Z, Z \rangle = \alpha \|Z\|^2$. Hence $\lambda'_{\max}(y; d) > 0$ for any $d \in N_y \mathcal{M}_r^{\lambda_{\max}} \setminus \{0\}$.

*Curve assumption* Let $\bar{y} \in \mathcal{M}_r^{\lambda_{\max}}$. For any $y \in \mathbb{S}_m$, we denote by $P(y)$ the orthogonal projection on the eigenspace corresponding to the $r$ largest eigenvalues of $y$ (counting multiplicities). This operator is smooth. We can define a mapping $U : \mathbb{S}_m \to \mathbb{R}^{m \times r}$ such that: $U(y)^\top U(y) = I_r$, $P(y) = U(y)U(y)^\top$, $U$ is smooth near our reference point $\bar{y}$ and its derivative at $\bar{y}$ satisfies $\mathrm{D}\, U(\bar{y})^\top U(\bar{y}) = 0$. The mapping $U$ defines a smooth orthonormal basis of the eigenspace corresponding to the $r$ largest eigenvalues [28, p. 557]. Finally, for a point $y' \in \mathcal{M}_r^{\lambda_{\max}}$, the projection of $d \in \mathbb{S}_m$ on $N_{y'} \mathcal{M}_r^{\lambda_{\max}}$ writes

$$\operatorname{proj}_{N_{y'} \mathcal{M}_r^{\lambda_{\max}}}(d) = U(y') \left\{ U(y')^\top dU(y') - \frac{1}{r} \operatorname{trace}(U(y')^\top dU(y')) I_r \right\} U(y')^\top.$$

Now, fix $y$ near $\bar{y}$, consider the eigenbasis $U$ with reference point $e(y, 0) = \operatorname{proj}_{\mathcal{M}_r^{\lambda_{\max}}}(y)$. Following Lemma A.1, let $\nu : t \mapsto \operatorname{proj}_{N_{e(y,t)} \mathcal{M}_r^{\lambda_{\max}}}(d)$ with $d = \operatorname{proj}_{\mathcal{M}_r^{\lambda_{\max}}}(y) - y$. We can now give an explicit expression of $\nu(t)$ and show that

$\frac{\mathrm{d}}{\mathrm{d}t}\nu(0)$ is null. Denoting $U(t) = U(e(y,t))$, we have

$$\nu(t) = U(t)\underbrace{\left\{U(t)^\top dU(t) - \frac{1}{r}\operatorname{trace}(U(t)^\top dU(t))I_r\right\}}_{\triangleq \chi(t)}U(t)^\top.$$

First, as $d$ is a normal vector to $\mathcal{M}_r^{\lambda_{\max}}$ at point $\operatorname{proj}_{\mathcal{M}_r^{\lambda_{\max}}}(y)$, there exists $Z \in \mathbb{S}_r$ such that $d = U(0)ZU(0)^\top$. Using that $\mathrm{D}\,U(0)^\top U(0) = 0$ yields

$$\mathrm{D}\,U(0)^\top dU(0) = \mathrm{D}\,U(0)^\top U(0)ZU(0)^\top U(0) = 0.$$

Then, one readily checks that $U(0)\,\mathrm{D}\,\chi(0)U(0) = 0$.

We turn to the term $\mathrm{D}\,U(0)\chi(0)U(0)^\top$. A quick computation from the eigen decomposition of $y$ shows that $d$ writes $U(0)ZU(0)^\top$, where $Z$ is actually diagonal. Therefore, $\chi(0) = Z - (1/r)\operatorname{trace}(Z)I_r$ is a diagonal matrix, so that

$$\mathrm{D}\,U(0)\chi(0)U(0)^\top = \sum_{i=1}^{r}\chi(0)_{ii}\,\mathrm{D}\,U_i(0)U_i(0)^\top.$$

Following [28], the differential of $t \mapsto U(e(y,t))$ at $t = 0$ writes

$$\mathrm{D}\,U_i(0) = \sum_{k=r+1}^{m}\frac{1}{\lambda_1 - \lambda_k}U_k(0)U_k(0)^\top\eta U_i(0),$$

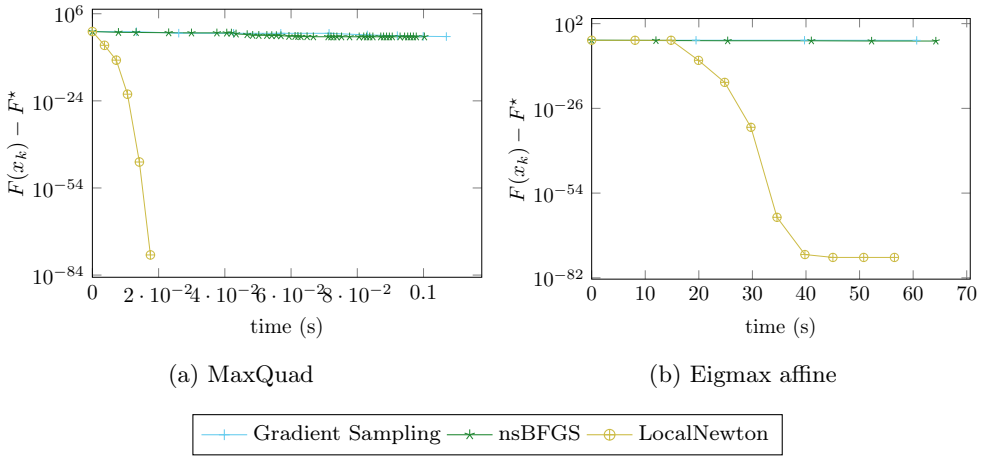
with $\eta = \operatorname{grad}\lambda_{\max}(\operatorname{proj}_{\mathcal{M}_r^{\lambda_{\max}}}(y))$. Using that $\lambda_{\max}(y) = (1/r)\sum_{i=1}^{r}U_i(y)^\top yU_i(y)$, we compute the Riemannian gradient: $\operatorname{grad}\lambda_{\max}(y) = (1/r)\sum_{i=1}^{r}U_i(y)^\top U_i(y)$ (see [6, Sec. 7.7]). By orthogonality of the smooth basis of eigenvectors, the terms $U_k(0)^\top U_i(0)$ vanish for all $i \in \{1,\dots,r\}$ and $k \in \{r+1,\dots,m\}$. We get that $\mathrm{D}\,U(0)\chi(0)U(0)^\top = 0$, and thus that $\mathrm{D}\,\nu(0) = 0$. Thus, Lemma A.1 applies and yields the result. ∎

**Appendix B. Numerical experiments in high precision.** We report in Figure 6 the evolution of suboptimality versus computing time, for the same problems and algorithms as in section 5, but with a high precision floating type. Indeed, the flexibility of the Julia language allows to use the same implementation with the high precision `BigFloat` type, which precision is $1.73 \cdot 10^{-72}$, or the usual `Float64` type, which precision is $2.22 \cdot 10^{-16}$.

**Acknowledgments.** This work is funded by the ANR JCJC project STROLL (ANR-19-CE23-0008) and MIAI@Grenoble Alpes (ANR-19-P3IA-0003).

## REFERENCES

[1] G. Bareilles, F. Iutzeler, and J. Malick, *Newton acceleration on manifolds identified by proximal-gradient methods*, arXiv preprint arXiv:2012.12936, (2020).
[2] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah, *Julia: A fresh approach to numerical computing*, SIAM review, 59 (2017), pp. 65–98.
[3] J. Bolte, Z. Chen, and E. Pauwels, *The multiproximal linearization method for convex composite problems*, Mathematical Programming, 182 (2020), pp. 1–36, https://doi.org/10.1007/s10107-019-01382-3.
[4] J. Bolte and E. Pauwels, *Majorization-Minimization Procedures and Convergence of SQP Methods for Semi-Algebraic and Tame Programs*, Mathematics of Operations Research, 41 (2016), pp. 442–465, https://doi.org/10.1287/moor.2015.0735.




(a) MaxQuad (b) Eigmax affine

Fig. 6: Suboptimality vs time (s)

[5] J.-F. Bonnans, J. C. Gilbert, C. Lemaréchal, and C. A. Sagastizábal, *Numerical optimization: theoretical and practical aspects*, Springer Science & Business Media, 2006.
[6] N. Boumal, *An introduction to optimization on smooth manifolds*. To appear with Cambridge University Press, Jan 2022, http://www.nicolasboumal.net/book.
[7] J. V. Burke, F. E. Curtis, A. S. Lewis, M. L. Overton, and L. E. Simões, *Gradient sampling methods for nonsmooth optimization*, in Numerical Nonsmooth Optimization, Springer, 2020, pp. 201–225.
[8] A. Daniilidis, W. Hare, and J. Malick, *Geometrical interpretation of the predictor-corrector type algorithms in structured optimization problems*, Optimization, 55 (2006), pp. 481–503.
[9] D. Drusvyatskiy, A. D. Ioffe, and A. S. Lewis, *Nonsmooth optimization using Taylor-like models: Error bounds, convergence, and termination criteria*, Mathematical Programming, 185 (2021), pp. 357–383, https://doi.org/10.1007/s10107-019-01432-w.
[10] X. Y. Han and A. S. Lewis, *Survey Descent: A Multipoint Generalization of Gradient Descent for Nonsmooth Optimization*, (2021), p. 29.
[11] W. Hare and A. S. Lewis, *Identifying active constraints via partial smoothness and prox-regularity*, Journal of Convex Analysis, 11 (2004), pp. 251–266.
[12] W. Hare and C. Sagastizábal, *Computing proximal points of nonconvex functions*, Mathematical Programming, 116 (2009), pp. 221–258.
[13] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms*, Springer Verlag, Heidelberg, 1993. Two volumes.
[14] C.-p. Lee, *Accelerating Inexact Successive Quadratic Approximation for Regularized Optimization Through Manifold Identification*, arXiv:2012.02522 [math], (2021), https://arxiv.org/abs/2012.02522.
[15] J. M. Lee, *Introduction to Smooth Manifolds*, Graduate Texts in Mathematics, Springer-Verlag, New York, 2003, https://doi.org/10.1007/978-0-387-21752-9.
[16] A. Lewis and T. Tian, *Identifiability, the kl property in metric spaces, and subgradient curves*, arXiv preprint arXiv:2205.02868, (2022).
[17] A. Lewis and C. Wylie, *A simple Newton method for local nonsmooth optimization*, arXiv:1907.11742 [cs, math], (2019), https://arxiv.org/abs/1907.11742.
[18] A. S. Lewis, *Active sets, nonsmoothness, and sensitivity*, SIAM Journal on Optimization, 13 (2002), pp. 702–725.
[19] A. S. Lewis and M. L. Overton, *Nonsmooth optimization via quasi-Newton methods*, Mathematical Programming, 141 (2013), pp. 135–163, https://doi.org/10.1007/s10107-012-0514-2.
[20] A. S. Lewis and S. J. Wright, *A proximal method for composite minimization*, Mathematical Programming, 158 (2016), pp. 501–546.
[21] R. Mifflin and C. Sagastizábal, *A $\mathcal{VU}$-algorithm for convex minimization*, Mathematical programming, 104 (2005), pp. 583–608.
[22] S. A. Miller and J. Malick, *Newton methods for nonsmooth convex minimization: con-*

*nections among-lagrangian, riemannian newton and sqp methods*, Mathematical programming, 104 (2005), pp. 609–633.

[23] J. Nocedal and S. Wright, *Numerical optimization*, Springer Science & Business Media, 2006.

[24] D. Noll and P. Apkarian, *Spectral bundle methods for non-convex maximum eigenvalue functions: second-order methods*, Mathematical Programming, 104 (2005), pp. 729–747.

[25] F. Oustry, *The $\mathcal{U}$-Lagrangian of the Maximum Eigenvalue Function*, SIAM Journal on Optimization, 9 (1999), pp. 526–549, https://doi.org/10.1137/S1052623496311776.

[26] R. T. Rockafellar and R. J.-B. Wets, *Variational analysis*, vol. 317, Springer Science & Business Media, 2009.

[27] A. Shapiro, *On a Class of Nonsmooth Composite Functions*, Mathematics of Operations Research, 28 (2003), pp. 677–692, https://doi.org/10.1287/moor.28.4.677.20512.

[28] A. Shapiro and M. K. H. Fan, *On Eigenvalue Optimization*, SIAM Journal on Optimization, 5 (1995), pp. 552–569, https://doi.org/10.1137/0805028.

[29] R. S. Womersley and R. Fletcher, *An algorithm for composite nonsmooth optimization problems*, Journal of Optimization Theory and Applications, 48 (1986), pp. 493–523, https://doi.org/10.1007/BF00940574.