



**HAL**  
open science

## Newton acceleration on manifolds identified by proximal-gradient methods

Gilles Bareilles, Franck Iutzeler, Jérôme Malick

► **To cite this version:**

Gilles Bareilles, Franck Iutzeler, Jérôme Malick. Newton acceleration on manifolds identified by proximal-gradient methods. 2021. hal-03197686v1

**HAL Id: hal-03197686**

**<https://hal.science/hal-03197686v1>**

Preprint submitted on 30 Sep 2021 (v1), last revised 25 May 2022 (v3)

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Newton acceleration on manifolds identified by proximal-gradient methods

Gilles Bareilles, Franck Iutzeler, Jérôme Malick

Received: date / Accepted: date

**Abstract** Proximal methods are known to identify the underlying substructure of nonsmooth optimization problems. Even more, in many interesting situations, the output of a proximity operator comes with its local structure to no additional cost, and convergence is improved once some final neighborhood of interest is attained. However, it is impossible in general to know whether the current structure is final or not; such highly valuable information has to be exploited adaptively. To do so, we place ourselves in the case where a proximal gradient method can identify manifolds of differentiability of the nonsmooth objective. Leveraging this manifold identification, we show that Riemannian Newton-like methods can be intertwined with the proximal gradient steps to drastically boost the convergence. We prove the super-linear convergence of the algorithm when solving some nondegenerated nonsmooth nonconvex optimization problems. We provide numerical illustrations on optimization problems regularized by  $\ell_1$ -norm or the trace-norm.

**Keywords** Nonsmooth optimization · Riemannian optimization · Proximal Gradient · Identification · Partial Smoothness · Sparsity-inducing regularization

## 1 Introduction

Nonsmoothness naturally appears in various applications of optimization, e.g. in decomposition methods in operations research [13], [35], or in sparsity-inducing regularization techniques in data analysis [4], [39]. In these applications, the nonsmoothness is not completely arbitrary but strongly structured. The *nonsmooth* objective functions usually present a *smooth* substructure, which involves smooth manifolds

---

G. Bareilles · F. Iutzeler  
Univ. Grenoble Alpes  
Laboratoire Jean Kuntzmann  
E-mail: firstname.lastname@univ-grenoble-alpes.fr  
J. Malick  
Univ. Grenoble Alpes  
CNRS, Laboratoire Jean Kuntzmann  
E-mail: firstname.lastname@univ-grenoble-alpes.fr

on which the functions are locally smooth. To fix ideas, consider the simple example of the  $\ell_1$  norm: though nonsmooth, it is obviously smooth around any point when restricted to the vector space of points with the same support.

Exploiting the underlying smooth substructure of objective functions to develop second-order methods has been a subject of fruitful and creative research in nonsmooth optimization, pioneered by the developments around  $\mathcal{U}$ -Newton algorithms [26] and the notion of partial smoothness [28]. Let us mention the  $\mathcal{UV}$ -Newton bundle method of [31] approximating the substructure from first-order information, and the recent  $k$ -bundle Newton method of [27] refining the approximation from a partial second-order oracle. Interestingly, these Newton-type methods for nonsmooth optimization are connected to the standard Newton methods of nonlinear programming (SQP) and to the Newton methods of Riemannian optimization; see [32]. We refer to [38] for a recent review of these nonsmooth Newton methods relying on an implicit smooth substructure.

In this paper, we focus on a special situation where the smooth substructure can be exploited numerically. We consider the nonsmooth optimization problem

$$\min_{x \in \mathbb{R}^n} F(x) \triangleq f(x) + g(x) \quad (\mathcal{P})$$

where  $f$  is a smooth differentiable function, and  $g$  is not everywhere differentiable – but admits a *simple* proximal operator. More precisely, we assume that the proximal operator of  $g$  outputs an explicit expression of the proximal point together with a representation of the current active submanifold. Coming back to the example of the  $\ell_1$ -norm: its proximity operator, called soft-thresholding, puts exactly to 0 some coordinates of the input vector, using on a comparison test; hence, the output has some sparsity structure, which is known as a byproduct of the computation. More generally, this situation covers a large class of signal processing and machine learning applications, where  $g$  is used to enforce some prior structure such as sparsity of vectors (when  $g$  is one of the  $\ell_1, \ell_{0.5}, \ell_0$ -norms) or low rank of matrices (when  $g$  is the nuclear norm); see e.g. [4].

Since  $g$  has a simple proximal operator, first-order methods of choice to minimize  $F$  are the (accelerated) proximal-gradient algorithms, interlacing “forward” steps with the gradient of the smooth part  $f$  and “backward” steps with the proximal operator of the nonsmooth part  $g$ . Interestingly, in nondegenerate cases, the iterates produced by these algorithms eventually reach the optimal submanifold (ie. the manifold which contains the minimizer): it is the so-called *identification* property of proximal algorithms, extensively studied in more general settings; we refer to [14], [40], [19], or [29]. For  $\ell_1$ -norm regularization, this means that after a finite but unknown number of iterations the algorithm “identifies” the final set of non-zero variables; see the pedagogical paper [23] for further discussions.

In the ideal case where we know that the iterates have reached the optimal manifold, one could switch to a more sophisticated method, e.g. updating parameters of first-order methods as in [30] or considering Riemannian Newton methods as in [15]. Unfortunately, even though we know the current structure of the iterates generated by a proximity operator and we know that they will identify the optimal manifold in finite time, we *never* know if the current manifold is the optimal one. Adaptivity is thus key in the exploitation of proximally discovered substructure.

We propose here a Newton acceleration of the proximal-gradient algorithm solving the nonsmooth optimization problem  $(\mathcal{P})$ . Our algorithm uses the same basic

ingredients that work behind the scenes for the existing nonsmooth Newton algorithms recalled above (e.g. [31], [15], [27]). In our context, they become apparent, and we can then heavily rely on recent developments of Riemannian optimization [2], [12]. Specifically our algorithm relies on (i) explicit proximal operations for structure identification and (ii) the efficiency of Riemannian Newton-type methods to finally benefit from faster convergence. We present a convergence analysis showing super-linear convergence of the resulting algorithm under some qualification assumptions – but without prior knowledge on the final manifold.

The simpler algorithm combining these features consists in alternating between a proximal gradient step and a Riemannian step on the identified manifold. We present a detailed convergence analysis of this algorithm, with a specific attention to the technical details induced by nonconvexity. We then propose to use Newton-like methods for the Riemannian step, leading to a Newton acceleration of the proximal gradient. Along the way, our study reveals results that have some interest on their own, in particular: we establish the smoothness of the proximal gradient operator around structured critical points; we also formalize complementary properties on line-searches in Riemannian optimization. Finally, we provide numerical illustrations showing the interests of the proposed Newton acceleration on typical structure-inducing regularized problems (sparse logistic regression and low-rank least-squares).

The paper is organized as follows. First, in Section 2 we recall the notions of Riemannian optimization and variational analysis that will be central in the paper. Then, we introduce in Section 3 our template algorithm alternating a proximal gradient step with a Riemannian update on the identified manifold. We show its convergence and give local superlinear convergence results towards qualified points. In Section 4, we specify the implementation of efficient Riemannian Newton-type methods and illustrate their performances in Section 5. The paper also contains three appendices with material used in our proofs; some of these complementary results are well-known and just recalled here, but several others seem to be less-known or not precisely treated in the existing literature.

## 2 Preliminaries: definitions, recalls, and examples

In this section, we introduce the notions which will be central in our developments. Our notation and terminology follow closely those of the textbooks [2] for Riemannian optimization and [37] for nonsmooth optimization. This section can be skipped by readers familiar with these topics.

### 2.1 Recalls on Riemannian optimization

We briefly introduce below the tools of Riemannian optimization used in this paper. We refer the reader to [2] and [12] for more extensive presentations.

**Submanifolds.** 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 local parameterization of  $\mathcal{M}$  around  $\bar{x}$ , that is, a  $\mathcal{C}^2$  function  $\varphi : \mathbb{R}^p \rightarrow \mathbb{R}^n$  such that  $\varphi$  realizes a local homeomorphism between a neighborhood of  $0 \in \mathbb{R}^p$  and a neighborhood of  $\bar{x} \in \mathcal{M}$  and the derivative of  $\varphi$  at  $\varphi^{-1}(\bar{x}) = 0$  is injective. A *p-dimensional  $\mathcal{C}^2$ -submanifold* of  $\mathbb{R}^n$  can alternatively be

defined via a local equation, that is a  $\mathcal{C}^2$  function  $\Phi : \mathbb{R}^n \rightarrow \mathbb{R}^{n-p}$  with a surjective derivative at  $\bar{x} \in \mathcal{M}$  that satisfies for all  $x$  close enough to  $\bar{x}$ :  $x \in \mathcal{M} \Leftrightarrow \Phi(x) = 0$ .

A basic tool to investigate approximations on manifolds is notion of the *smooth curves*. A smooth curve on  $\mathcal{M}$  is a  $\mathcal{C}^2$  application  $\gamma : I \subset \mathbb{R} \rightarrow \mathcal{M} \subset \mathbb{R}^n$ , where  $I$  is an open interval containing 0. At each point  $x \in \mathcal{M}$ , the *tangent space*, noted  $T_x\mathcal{M}$ , can be defined as the velocities of all smooth curves passing by  $x$  at 0:

$$T_x\mathcal{M} \triangleq \{c'(0) \mid c : I \rightarrow \mathcal{M} \text{ is a smooth curve around } 0 \text{ and } c(0) = x\}.$$

The tangent space is a  $p$ -dimensional Euclidean space containing *tangent vectors*. Each tangent space  $T_x\mathcal{M}$  is equipped with a scalar product  $\langle \cdot, \cdot \rangle_x : T_x\mathcal{M} \times T_x\mathcal{M} \rightarrow \mathbb{R}$ , and the associated norm  $\|\cdot\|_x$ . In many cases, the tangent metric varies smoothly with  $x$ , making the manifold *Riemannian*. In this paper, we use the ambient space scalar product to define the scalar product on tangent spaces; we will thus drop the subscript in the tangent scalar product and norm notations when there is no confusion possible. Related to the tangent space, we will also consider the *normal space*  $N_x\mathcal{M}$  at  $x \in \mathcal{M}$ , defined as the orthogonal space to  $T_x\mathcal{M}$  in  $\mathbb{R}^n$ , and the *tangent bundle manifold* defined by:

$$T\mathcal{B} \triangleq \bigcup_{x \in \mathcal{M}} (x, T_x\mathcal{M}).$$

Note also that both tangent and normal spaces at  $x \in \mathcal{M}$  admit explicit expressions from derivatives of local parametrization  $\varphi$  or local equations  $\Phi$  defining  $\mathcal{M}$ :

$$T_x\mathcal{M} = \text{Im } D\varphi(0) = \text{Ker } D\Phi(x) \quad N_x\mathcal{M} = \text{Ker } D\varphi(0)^* = \text{Im } D\Phi(x)^*$$

A *metric* on  $\mathcal{M}$  can be defined as the minimal length over all curves joining two points  $x, y \in \mathcal{M}$ , ie.  $\text{dist}_{\mathcal{M}}(x, y) = \inf_{c \in C_{x,y}} \int_0^1 \|c'(t)\|_{c(t)} dt$ , where  $C_{x,y}$  is the set of  $[0, 1] \rightarrow \mathcal{M}$  smooth curves  $c$  such that  $c(0) = x$ ,  $c(1) = y$ . The minimizing curves generalize the notion of straight line between two points to manifolds. The constant speed parametrization of any minimizing curve is called a *geodesic*.

**Riemannian Gradients and Hessian.** The notion of derivative and gradient are extended to functions defined on Riemannian manifolds, as follows. Let  $F : \mathcal{M} \rightarrow \mathbb{R}$ , the *Riemannian differential* of  $F$  at  $x$  is the linear operator  $DF(x) : T_x\mathcal{M} \rightarrow \mathbb{R}$  defined by  $DF(x)[\eta] \triangleq \frac{d}{dt} F \circ c(t) \Big|_{t=0}$ , where  $c$  is a smooth curve such that  $c(0) = x$  and  $c'(0) = \eta$ . In turn, the *Riemannian gradient*  $\text{grad } F(x)$  is the unique vector of  $T_x\mathcal{M}$  such that, for any tangent vector  $\eta$ ,  $DF(x)[\eta] = \langle \text{grad } F(x), \eta \rangle$ . If  $\text{grad } F(x)$  exists, a first order Taylor development can be formulated. Let  $x \in \mathcal{M}$ ,  $\eta \in T_x\mathcal{M}$  and  $c$  denote a smooth curve passing by  $x$ , with velocity  $\eta$  at 0; then, for  $t$  near 0,

$$F \circ c(t) = F(x) + t \langle \text{grad } F(x), \eta \rangle + o(t).$$

Before defining second order objects, we need a notion of derivation for vector fields and of acceleration for curves. Consider a curve  $c : I \rightarrow \mathcal{M}$  and a smooth vector field  $Z$  on that curve, that is a smooth map such that  $Z(t) \in T_{c(t)}\mathcal{M}$  for  $t \in I$ . The *covariant derivative* of  $Z$  on the curve  $c$ , denoted  $\frac{D}{dt} Z : I \rightarrow T\mathcal{B}$ , is defined by  $\frac{D}{dt} Z(t) \triangleq \text{proj}_{c(t)} Z'(t)$ , where  $Z'(t)$  denotes the derivative in the ambient space  $\mathbb{R}^n$  and  $\text{proj}_x$  corresponds to the orthogonal projector from  $\mathbb{R}^n$  to  $T_x\mathcal{M}$ . The *acceleration* of a curve  $c$  is defined as the covariant derivative of its velocity:  $c''(t) \triangleq \frac{D}{dt} c'(0)$ . It amounts to the tangential component of the second derivative of the curve in  $\mathbb{R}^n$ .

The *Riemannian hessian* of  $F$  at  $x$  along  $\eta$  is the linear operator  $\text{Hess} : T_x\mathcal{M} \rightarrow T_x\mathcal{M}$  defined by the relation  $\text{Hess } F(x)[\eta] \triangleq \frac{D}{dt} \text{grad } F(c(t))|_{t=0}$ , where  $c$  is a smooth curve such that  $c(0) = x$ ,  $c'(0) = \eta$ . The Riemannian hessian may be defined equivalently by the relation  $\langle \text{Hess } F(x)[\eta], \eta \rangle = \frac{d^2}{dt^2} F \circ \gamma(t)|_{t=0}$ , where  $\gamma$  is a geodesic curve such that  $\gamma(0) = x$ ,  $\gamma'(0) = \eta$ . A second order Taylor development can now be formulated. Let  $x \in \mathcal{M}$ ,  $\eta \in T_x\mathcal{M}$ , and  $c$  be a smooth curve such that  $c(0) = x$ ,  $c'(0) = \eta$ . Then, for  $t$  near 0,

$$F \circ c(t) = F(x) + t \langle \text{grad } F(x), \eta \rangle + \frac{t^2}{2} \langle \text{Hess } F(x)[\eta], \eta \rangle + \frac{t^2}{2} \langle \text{grad } F(x), c''(0) \rangle + o(t^2). \quad (2.1)$$

If  $F : \mathcal{M} \rightarrow \mathbb{R}$  has a smooth extension on  $\mathbb{R}^n$ , the Riemannian gradient and hessian can be computed from their Euclidean counterparts: for a smooth function  $\bar{F} : \mathbb{R}^n \rightarrow \mathbb{R}$  that coincides with  $F$  on  $\mathcal{M}$ , we have

$$\text{grad } F(x) = \text{proj}_x(\nabla \bar{F}(x)), \quad (2.2)$$

and, for  $\bar{G} : \mathbb{R}^n \rightarrow \mathbb{R}^n$  a smooth mapping that coincides with  $\text{grad } F$  on  $\mathcal{M}$ , we have

$$\text{Hess } F(x)[\eta] = \text{proj}_x(D\bar{G}(x)[\eta]). \quad (2.3)$$

**Algorithms on manifolds: retractions and convergence rates.** Iterative Riemannian methods require a way to produce curves on  $\mathcal{M}$  given a point  $x$  and a tangent vector  $\eta$ . A geodesic curve passing at  $(x, \eta)$ , while attractive as the generalization of the straight line, has a prohibitive computational cost. We thus use first and second-order approximations of it. A *retraction* on a manifold  $\mathcal{M}$  is a smooth map  $R : T\mathcal{B} \rightarrow \mathcal{M}$  such that

$$R_x(0) = x \quad \text{and} \quad DR_x(0) : T_x\mathcal{M} \rightarrow T_x\mathcal{M} \text{ is the identity map: } DR_x(0)[v] = v,$$

where, for each  $x \in \mathcal{M}$ ,  $R_x : T_x\mathcal{M} \rightarrow \mathcal{M}$  is defined as the restriction of  $R$  at  $x$ , so that  $R_x(v) = R(x, v)$ . A *second-order retraction* is a retraction  $R$  such that, for all  $(x, \eta) \in T\mathcal{B}$ , the curve  $c(t) = R_x(t\eta)$  has zero acceleration at 0:  $c''(0) = 0$ . Thus  $t \mapsto R_x(t\eta)$  is a practical curve passing by  $(x, \eta)$  at 0, and provides a similar development as (2.1): for  $t$  near 0,

$$F \circ R_x(t\eta) = F(x) + t \langle \text{grad } F(x), \eta \rangle + \frac{t^2}{2} \langle \text{Hess } F(x)[\eta], \eta \rangle + o(t^2 \|\eta\|^2). \quad (2.4)$$

Finally, the convergence rates on manifolds are defined as follows. A sequence of points  $(x_k)$  *converges (Q-)linearly* to some point  $\bar{x} \in \mathcal{M}$  if there exist an integer  $K > 0$  and a constant  $q \in (0, 1)$  such that, for all  $k \geq K$ , there holds

$$\text{dist}_{\mathcal{M}}(x_{k+1}, \bar{x}) \leq q \text{dist}_{\mathcal{M}}(x_k, \bar{x}).$$

The sequence *converges with order at least  $p$*  if there exists an integer  $K > 0$  and a constant  $q \in (0, 1)$  such that, for all  $k \geq K$ , there holds

$$\text{dist}_{\mathcal{M}}(x_{k+1}, \bar{x}) \leq q \text{dist}_{\mathcal{M}}(x_k, \bar{x})^p.$$

The convergence is *superlinear* when  $p > 1$  and *quadratic* when  $p = 2$ .

**Examples of submanifolds and related objects.** In this paper, we will illustrate our developments with two sparsity-inducing norms (see Section 2.3) involving respectively the two following manifolds.

*Example 2.1 (Fixed coordinate-sparsity subspaces)* We consider the submanifold of  $\mathbb{R}^n$  defined as

$$\mathcal{M}_I \triangleq \{x \in \mathbb{R}^n : x_i = 0 \text{ for } i \in I\}, \quad (2.5)$$

where  $I \subset \{1, \dots, n\}$  denotes the set indices of the zero coordinates. This manifold is actually a vector space and all related notions have simple expressions, as follows.

The tangent space at any point identifies with the manifold itself:  $T_x \mathcal{M}_I = \mathcal{M}_I$ . The orthogonal projection of a vector  $d \in \mathbb{R}^n$  on the tangent space writes  $\text{proj}_x(d)$ , where  $[\text{proj}_x(d)]_i$  is  $d_i$  if  $i \notin I$ , and null otherwise. The map  $\mathbb{R}_x(\eta) = x + \eta$  defines a second-order retraction. Given a function  $F$  defined on the ambient space, the Riemannian gradient and hessian-vector product of the restriction of  $F$  to  $\mathcal{M}_I$  can be obtained from their Euclidean counterparts by a simple projection: for  $x, \eta \in T\mathcal{B}$ ,

$$\text{grad } F(x) = \text{proj}_x(\nabla F(x)) \quad \text{Hess } F(x)[\eta] = \text{proj}_x(\nabla^2 F(x)[\eta]).$$

*Example 2.2 (Fixed rank matrices)* The second running example is the manifold of fixed-rank matrices

$$\mathcal{M}_r \triangleq \{x \in \mathbb{R}^{m \times n} : \text{rank}(x) = r\}, \quad (2.6)$$

for which we refer to [12, Sec. 7.5] for more details. A rank- $r$  matrix  $x \in \mathcal{M}_r$  is represented as  $x = U\Sigma V^\top$ , where  $U \in \mathbb{R}^{m \times r}$ ,  $V \in \mathbb{R}^{n \times r}$ ,  $\Sigma \in \mathbb{R}^{r \times r}$  such that  $U^\top U = I_r$ ,  $V^\top V = I_r$  and  $\Sigma$  is a diagonal matrix with positive entries. Such a decomposition can be obtained by computing the singular value decomposition of the matrix  $x$ . Using this representation, a tangent vector  $\eta \in T_x \mathcal{M}_r$  can be written as

$$\eta = UMV^\top + U_p V_p^\top + UV_p^\top,$$

where  $M \in \mathbb{R}^{r \times r}$ ,  $U_p \in \mathbb{R}^{m \times r}$ ,  $V_p \in \mathbb{R}^{n \times r}$  such that  $U^\top U_p = 0$ ,  $V^\top V_p = 0$ . The orthogonal projection of a vector  $d \in \mathbb{R}^{m \times n}$  onto  $T_x \mathcal{M}_r$  writes  $\text{proj}_x(d) = d - U^\top d V$ . Given a function  $F$  defined on the ambient space, a Riemannian gradient and hessian-vector product of  $F$  restricted to  $\mathcal{M}_r$  can be obtained from their Euclidean counterparts: for  $x, \eta \in T\mathcal{B}$ ,

$$\text{grad } F(x) = \text{proj}_x(\nabla F(x))$$

$$\text{Hess } F(x)[\eta] = \text{proj}_x(\nabla^2 F(x)[\eta]) + \left[ P_U^\top \nabla F(x) V_p \Sigma^{-1} \right] V^\top + U \left[ P_V^\top \nabla F(x)^\top U_p \Sigma^{-1} \right]^\top,$$

where  $P_U^\top = I_m - UU^\top$ ,  $P_V^\top = I_n - VV^\top$ .

## 2.2 Recalls on nonsmooth optimization

We review the basic notions of variational analysis used in this paper; our main reference is the textbook [37]. For this section,  $g: \mathbb{R}^n \rightarrow \bar{\mathbb{R}} = \mathbb{R} \cup \{+\infty\}$  is a proper function.

**Subgradients.** Consider a point  $\bar{x}$  with  $g(\bar{x})$  finite. The set of *regular subgradients* is defined as

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

This set is closed and convex, but the subdifferential mapping  $\widehat{\partial}g(\cdot)$  may not be outer semi-continuous [37, Th. 8.6, Prop. 8.7]. To overcome this problem, the set of (*general or limiting*) *subgradients* is defined as

$$\partial g(\bar{x}) \triangleq \left\{ \lim_r v_r : v_r \in \widehat{\partial}g(x_r), x_r \rightarrow \bar{x}, g(x_r) \rightarrow g(\bar{x}) \right\}.$$

The limiting subdifferential is by design outer semi-continuous

$$\limsup_{x \rightarrow \bar{x}} \partial g(x) = \{u : \exists x_r \rightarrow \bar{x}, \exists u_r \rightarrow u \text{ with } u_r \in \partial g(x_r)\} \subset \partial g(\bar{x}),$$

which is an attractive property to study the properties of sequences of points whose subgradients converge. We say that a function is (*Clarke*) *regular* at  $\bar{x}$  if the regular and limiting subdifferentials at  $\bar{x}$  coincide [37, Def. 7.25, Cor. 8.11]. This is notably the case for convex functions where the two above definitions coincide with the convex subdifferential [37, Prop. 8.12].

**Optimality conditions and critical points.** The subdifferential allows to derive optimality conditions: for a local minimizer  $\bar{x}$  of  $F$ , we have  $0 \in \partial F(\bar{x})$ . For the objective function of  $(\mathcal{P})$ , this writes

$$0 \in \nabla f(\bar{x}) + \partial g(\bar{x}) \quad \text{or equivalently} \quad -\nabla f(\bar{x}) \in \partial g(\bar{x}).$$

A point satisfying these conditions is called a critical point. The analysis of the algorithms of this paper will provide convergence guarantees towards critical points. When  $F$  is convex, critical points coincide with (global) minimizers.

**Proximity operator.** A central tool to tackle non-differentiable functions is the *proximity operator*, introduced by [33]. For a step-size  $\gamma > 0$  and a function  $g$ ; it is defined as the set-valued mapping

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

This operator enables to achieve some descent on  $g$  while, as we will investigate next, preserving the structure of non-differentiability points of  $g$ . Since this operator will be at the core of our future developments, *we will assume that it is non-empty for all  $y$* . Note that this is a reasonable assumption since it is satisfied as soon as  $g$  is lower-bounded<sup>1</sup>, which is trivially verified by our functions of interest (see Section 2.3). Though computing proximal points is in general difficult, it is easy for some relevant cases, in particular for the sparsity-inducing regularizers often used in machine learning and signal processing, as the  $\ell_1$ -norm or the trace-norm; see Section 2.3 below.

An important property in our developments is that the output of the proximal operator are usually attracted to the non-differentiability points of  $g$ . We can use this additional valuable information to reduce the dimension of the space of the associated minimization problem; see e.g. the pedagogical paper [23].

**Prox-regularity.** Although we consider nonconvex functions in general, we will need some local stability, and we introduce the notion of prox-regularity [37, Def. 13.27].

<sup>1</sup> The weaker assumption that  $g$  is prox-bounded (ie.  $g + r\|\cdot\|^2$  is bounded below for some  $r$ ) implies that  $\mathbf{prox}_{\gamma g}(y)$  is non-empty for all  $y$  as long as  $\gamma$  is taken sufficiently small; see [37, Chap. 1.G] for additional details.



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

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

whenever  $v \in \partial g(x)$ ,  $\|x - \bar{x}\| < \varepsilon$ ,  $\|x' - \bar{x}\| < \varepsilon$ ,  $\|v - \bar{v}\| < \varepsilon$  and  $g(x) < g(\bar{x}) + \varepsilon$ . When this holds for all  $\bar{v} \in \partial g(\bar{x})$ , we say that  $g$  is prox-regular at  $\bar{x}$ .

This property allows to have local Lipschitzness of the proximal operator as well as its characterization by first-order optimality conditions; see [20, Th. 4] and Lemma 5. Specifically, we will use that if  $g$  is  $r$ -prox-regular at  $\bar{x}$ , then, for any  $\gamma < 1/r$ ,  $\mathbf{prox}_{\gamma g}(y)$  is single-valued and Lipschitz continuous for any  $y$  near  $\bar{x} + \gamma \bar{v}$  where  $\bar{v} \in \partial g(\bar{x})$  and  $\bar{x} = \mathbf{prox}_{g/r}(\bar{x} + \bar{v}/r)$ . Furthermore, in this neighborhood for  $y$ , it is uniquely determined by the relation

$$x = \mathbf{prox}_{\gamma g}(y) \Leftrightarrow \frac{y - x}{\gamma} \in \partial g(x),$$

which characterizes proximal maps using first-order optimality conditions.

### 2.3 Running examples

*Example 2.3 ( $\ell_1$  norm)* We consider the  $\ell_1$  norm defined on  $\mathbb{R}^n$  as  $\|x\|_1 = \sum_{i=1}^n |x_i|$ . This function is convex, thus prox-regular at every point with  $r = 0$ . Besides, its proximity operator admits a closed form expression: for  $y \in \mathbb{R}^n$  and  $\gamma > 0$ , coordinate  $i$  writes

$$[\mathbf{prox}_{\gamma \|\cdot\|_1}(y)]_i = \begin{cases} y_i + \gamma & \text{if } y_i < -\gamma \\ 0 & \text{if } -\gamma \leq y_i \leq \gamma \\ y_i - \gamma & \text{if } y_i > \gamma \end{cases}$$

For a given point  $\bar{x} \in \mathbb{R}^n$ , observe that the restriction of  $\|\cdot\|_1$  to the manifold  $\mathcal{M}_{\bar{I}} = \{x \in \mathbb{R}^n : x_i = 0 \text{ if } \bar{x}_i = 0\}$  (see (2.5)) is smooth. The  $\ell_1$  norm thus admits a Riemannian gradient and hessian at point  $\bar{x}$ :

$$\text{grad } \|\cdot\|_1(\bar{x}) = \text{proj}_{\bar{x}}(\text{sign}(\bar{x})) \quad \text{and} \quad \text{Hess } \|\cdot\|_1(\bar{x}) = 0,$$

where  $\text{sign}(x) \in \{-1, 0, 1\}$  denotes the sign of  $x$ , null when  $x = 0$ .

*Example 2.4 (nuclear norm)* We consider the nuclear norm, defined on  $\mathbb{R}^{m \times n}$  as  $\|x\|_* = \sum_{i=1}^{\min(m,n)} \Sigma_{ii}$ , where  $\Sigma$  denotes the diagonal term of the singular value decomposition of  $x$ . This function is convex, and thus prox-regular at every point with  $r = 0$ . Its proximity operator admits a closed form expression: for matrix  $y (= U \Sigma V^\top)$  and step  $\gamma > 0$ ,

$$\mathbf{prox}_{\gamma \|\cdot\|_*}(y) = U(\Sigma - \gamma)_+ V^\top,$$

where coefficient  $(i, j)$  of  $(\Sigma - \gamma)_+$  is defined as  $\max(\Sigma_{ij} - \gamma, 0)$ .

The nuclear norm restricted to the manifold  $\mathcal{M}_{\bar{r}}$  of the matrices of rank  $\bar{r} = \text{rank}(\Sigma - \gamma)_+$  (see (2.6)) is locally  $\mathcal{C}^2$  around  $\bar{x}$  and admits a smooth second-order

development of the form (2.4) at any point  $x = U\Sigma V^\top \in \mathcal{M}_{\tilde{r}}$  close to  $\bar{x}$ . Denoting  $\eta = U_M V^\top + U_p V^\top + U V_p^\top \in T_x \mathcal{M}_{\tilde{r}}$  a tangent vector, there holds indeed:

$$\begin{aligned} \text{grad } \|\cdot\|_*(x) &= UV^\top \\ \text{Hess } \|\cdot\|_*(x)[\eta] &= U \left[ \tilde{F} \circ (M - M^\top) \right] V^\top + U_p \Sigma^{-1} V^\top + U \Sigma^{-1} V_p^\top, \end{aligned}$$

where  $\circ$  denotes the Hadamard product and  $\tilde{F} \in \mathbb{R}^{\tilde{r} \times \tilde{r}}$  is such that  $\tilde{F}_{ij} = 1/(\Sigma_{jj} + \Sigma_{ii})$  if  $\Sigma_{jj} \neq \Sigma_{ii}$ , and  $\tilde{F}_{ij} = 0$  otherwise. This statement is proved in [Appendix C](#).

### 3 A general algorithm for explicit nonsmooth problems

Let us start by specifying our blanket assumptions on the composite problem ( $\mathcal{P}$ ).

**Assumption 1** (On the functions). *The functions  $f$  and  $g$  are proper and*

- i)  $f$  is  $\mathcal{C}^2(\mathbb{R}^n)$  with an  $L$ -Lipschitz continuous gradient;*
- ii)  $g$  is lower semi-continuous;*
- iii)  $\text{prox}_{\gamma g}$  is non-empty on  $\mathbb{R}^n$  for any  $\gamma > 0$ ;*
- iv)  $F(x) = f(x) + g(x)$  is bounded below.*

These assumptions are mostly common except the third point which directly comes from our idea of using the proximal operator both for the optimization itself and as an oracle for the current structure of the iterates. In this section, we lay down the general algorithm based on this idea (in [Section 3.1](#)) and study its global convergence under the above assumption (in [Section 3.2](#)) and its local convergence under some mild additional structure assumptions (in [Section 3.3](#))

#### 3.1 General proximal algorithm with Riemannian acceleration

We describe here the core ideas that guide the methods presented in this paper. Building on the well-established identification properties of proximal methods, we leverage this ability to an algorithmic advantage by reducing our working space to the identified structure. ‘‘Smooth’’ structures (involving smooth manifolds and smooth restrictions on it) are of special interest and open the way to Newton acceleration. As mentioned in the introduction and in [Section 2.3](#), the output of a proximity operator often comes with the knowledge of the current manifold on which it lives. Three ingredients thus come into play: (i) identification by a proximal method, (ii) adaptive scrutiny of the identified structure, and (iii) exploitation of the current structure by an optimization method.

---

#### Algorithm 1 General structure exploiting algorithm

---

- 1: **repeat**
  - 2:   Compute  $x_k \in \text{prox}_{\gamma g}(y_{k-1} - \gamma \nabla f(y_{k-1}))$  and get  $\mathcal{M}_k \ni x_k$
  - 3:   Update  $y_k = \text{ManUp}_{\mathcal{M}_k}(x_k)$  on the current manifold
  - 4: **until** stopping criterion
- 

We propose the general algorithm ([Algorithm 1](#)) which consists in, first, performing a proximal gradient step  $x_k \in \text{prox}_{\gamma g}(y_{k-1} - \gamma \nabla f(y_{k-1}))$  that provides both

the current point  $x_k$  and the manifold  $\mathcal{M}_k$  where it lies, and, second, carrying out a Riemannian optimization update,  $\text{ManUp}_{\mathcal{M}_k}$ , on the current manifold. This algorithm is general in the sense that we do not precise for now what is the Riemannian step  $\text{ManUp}$ . Our plan of action is to lay out the generic assumptions on the manifold updates that provide global and local convergence, respectively in Sections 3.2 and 3.3. Then we will investigate in Section 4 the Riemannian Newton acceleration falling into this scheme.

### 3.2 Global convergence

The following result show that [Algorithm 1](#) converges to a critical value of  $F$ , and all accumulation points of its iterates are critical points. In order for this to hold, we only need the mild assumption that the manifold update does not increase the functional value (this offers a broad choice of methods since this kind of descent is easily obtained by line-search as discussed in [Section 4.1](#)).

**Theorem 3.1 (Global convergence)** *Let [Assumption 1](#) hold and take  $\gamma \in (0, 1/L)$ . Suppose that the manifold update  $\text{ManUp}_{\mathcal{M}}$  provides descent, that is for any  $x$  in  $\mathcal{M}$*

$$F(\text{ManUp}_{\mathcal{M}}(x)) \leq F(x).$$

*Then, [Algorithm 1](#) generates non-increasing functional values ( $F(x_{k+1}) \leq F(y_k) \leq F(x_k)$  for all  $k$ ) and all limit points of  $(x_k)$  and  $(y_k)$  are critical points of  $F$  that share the same functional value.*

*Proof.* It is well-known that the proximal-gradient update provides a descent (see the result and reference in [Lemma 3](#) in appendix). Choosing  $y_k$  such that  $F(y_k) \leq F(x_k)$  (by assumption on the manifold update) yields:

$$F(x_{k+1}) \stackrel{\text{Lemma 3}}{\leq} F(y_k) - \frac{1 - \gamma L}{2\gamma} \|x_{k+1} - y_k\|^2 \leq F(x_k) - \frac{1 - \gamma L}{2\gamma} \|x_{k+1} - y_k\|^2. \quad (3.1)$$

The sequence  $(F(x_k))$  is thus non-increasing and lower-bounded, therefore it converges. Besides, any accumulation point of  $(x_k)$  is a critical point of  $F$ . Indeed, summing equation (3.1) for  $k = 1, \dots, n$  yields:

$$\frac{1 - \gamma L}{2\gamma} \sum_{k=1}^n \|x_{k+1} - y_k\|^2 \leq F(x_1) - F(x_{n+1}) \leq F(x_1) - \inf F < +\infty.$$

Therefore the general term of the above series  $\|x_{k+1} - y_k\|^2$  converges to 0, which implies, by [Lemma 4](#), that the distance from  $\partial F(x_k)$  to 0 converges to 0. The outer-semi continuity property of the limiting subdifferential allows to conclude that every accumulation point of  $(x_k)$  is a critical point of  $F$ . Finally, all limit points share the same functional value as  $F$  is lower semi-continuous.  $\square$

### 3.3 Local identification and superlinear rate

We study here how Riemannian methods with local superlinear convergence (such as Riemannian Newton's method) propagate their superior rate to our generic [Algorithm 1](#). We start with a result about the local smoothness of the proximal-gradient operator.

### 3.3.1 Smoothness and localization of the proximal gradient

The results of this section are built on  $g$  being a partly smooth function; see [28].

**Definition 1 (partial smoothness)** A function  $g$  is  $(\mathcal{C}^2)$ -partly smooth at a point  $\bar{x}$  relative to a set  $\mathcal{M}$  containing  $\bar{x}$  if  $\mathcal{M}$  is a  $\mathcal{C}^2$  manifold around  $\bar{x}$  and:

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

Under this assumption, we show in the next theorem that the proximal gradient smoothly locates active manifolds: if some input  $\bar{y}$  is mapped onto  $\mathcal{M}$ , then the proximal gradient is  $\mathcal{M}$ -valued and  $\mathcal{C}^1$  around  $\bar{y}$ . This result is based on the sensitivity analysis of partly smooth functions [28, Sec. 5]. The proof extends and refines the rationale of [15, Th. 28] about the proximity operator.

**Theorem 3.2 (Proximal gradient points locate active manifolds)** *Let  $f$  be a  $\mathcal{C}^2$  function on  $\mathbb{R}^n$  and  $g$  a lower semi-continuous function on  $\mathbb{R}^n$ . Suppose that  $g$  is both  $r$ -prox-regular at  $\bar{x}$  and partly-smooth relative to  $\mathcal{M}$  at  $\bar{x}$ .*

*Take  $\gamma, \bar{\gamma}$  such that  $0 < \gamma < \bar{\gamma} \leq 1/r$  and  $\bar{x} = \mathbf{prox}_{\bar{\gamma}g}(\bar{y} - \bar{\gamma}\nabla f(\bar{y}))$ . If*

- i)  $\frac{1}{\bar{\gamma}}(\bar{y} - \bar{x}) - \nabla f(\bar{y}) \in \text{ri } \partial g(\bar{x})$ ;*
- ii) either a)  $\gamma$  is sufficiently close to  $\bar{\gamma}$ , or b)  $\bar{y}$  is sufficiently close to  $\bar{x}$ ;*

*then, the proximal gradient  $y \mapsto \mathbf{prox}_{\gamma g}(y - \gamma\nabla f(y))$  is  $\mathcal{C}^1$  and  $\mathcal{M}$ -valued near  $\bar{y}$ .*

*Proof.* Adopting the same reasoning as in [28, Sec. 5] and [15, Sec. 4.1], we consider the function

$$\begin{aligned} \rho : \mathbb{R}^n \times \mathbb{R}^n &\rightarrow \mathbb{R}^n \\ (x, y) &\mapsto g(x) + \frac{1}{2\gamma}\|x - y + \gamma\nabla f(y)\|^2, \end{aligned}$$

and denote by  $\rho_y = \rho(\cdot, y)$ . Computing the proximal gradient  $\mathbf{prox}_{\gamma g}(y - \gamma\nabla f(y))$  can then be seen as minimizing the parametrized function  $\rho_y$ .

**Step 1.** As a first step, we study the minimizers of  $\rho_y$  restricted to  $\mathcal{M}$ , for  $y$  near  $\bar{y}$ . We consider the parametric manifold optimization problem, for  $y$  near  $\bar{y}$ :

$$\min_{x \in \mathcal{M}} \rho_y(x). \quad (P_{\mathcal{M}}(y))$$

Since  $g$  is  $\mathcal{C}^2$ -partly-smooth relative to  $\mathcal{M}$  and  $f$  is  $\mathcal{C}^2(\mathbb{R}^n)$ ,  $\rho_y$  is twice continuously differentiable on  $\mathcal{M}$ . By Lemma 6,  $\bar{x}$  is a strong local minimizer of  $\rho_{\bar{y}}$ . Thus, the second-order sufficient optimality condition on manifolds Lemma 7 hold:

$$\text{grad } \rho_{\bar{y}}(\bar{x}) = 0 \quad \text{Hess } \rho_{\bar{y}}(\bar{x}) \succeq \left(\frac{1}{\bar{\gamma}} - r\right)I \succ 0.$$

We consider the equation  $\Phi(x, y) = 0$ , for  $x, y$  near  $\bar{x}, \bar{y}$ , where  $\Phi : \mathcal{M} \times \mathbb{R}^n \rightarrow T\mathcal{B}$  is defined as  $\Phi(x, y) = \text{grad } \rho_y(x)$ . This function is continuously differentiable on a neighborhood of  $(\bar{x}, \bar{y})$ , and its differential relative to  $\bar{x}$  at that point,  $\text{Hess } \rho_{\bar{y}}(\bar{x})$ , is invertible. The implicit function theorem thus grants the existence of neighborhoods  $\mathcal{N}_{\bar{x}}, \mathcal{N}_{\bar{y}}$  of  $\bar{x}, \bar{y}$  in  $\mathcal{M}, \mathbb{R}^n$ , and a continuously differentiable function  $\hat{x} : \mathcal{N}_{\bar{y}} \rightarrow \mathcal{N}_{\bar{x}}$  such that, for any  $y$  in  $\mathcal{N}_{\bar{y}}$ ,  $\Phi(\hat{x}(y), y) = \text{grad } \rho_y(\hat{x}(y)) = 0$ . Actually,  $\hat{x}(y)$  is

a strong minimizer of  $\rho_y$  on  $\mathcal{M}$  for  $y$  close enough to  $\bar{y}$ . Indeed, the mapping  $\hat{x}$  is continuous on  $\mathcal{N}_{\bar{y}}$ , so that  $y \mapsto \text{Hess } \rho_y(\hat{x}(y))$  is also continuous there and the property  $\text{Hess } \rho_{\bar{y}}(\hat{x}(\bar{y})) \succ 0$  extends locally around  $\bar{y}$ .

**Step 2.** As a second step, we turn to show that the minimizer  $\hat{x}(y)$  of  $\rho_y$  on  $\mathcal{M}$  is actually a strong critical point of  $\rho_y$  in  $\mathbb{R}^n$  [28, Def. 5.3], and thus the proximal gradient of point  $y$ . More precisely, we claim that, for  $y$  near  $\bar{y}$  and  $x = \hat{x}(y)$ , there holds  $0 \in \text{ri } \partial \rho_y(x)$ , that is

$$\frac{1}{\gamma}(y - x) - \nabla f(y) \in \text{ri } \partial g(x).$$

This property holds at  $(\bar{x}, \bar{y})$  by assumption. By contradiction, assume there exist sequences of points  $(y_r)$  with limit  $\bar{y}$ ,  $(x_r) = (\hat{x}(y_r))$  with limit  $\bar{x} = \hat{x}(\bar{y})$  and  $(h_r)$  of unit norm  $\|h_r\| = 1$  such that for all  $r$ ,  $h_r$  separates 0 from  $\partial \rho_{y_r}(x_r)$ :

$$\inf_{h \in \partial \rho_{y_r}(x_r)} \langle h_r, h \rangle \geq 0.$$

Since  $(h_r)$  is bounded, a converging subsequence can be extracted from it, let  $\bar{h}$  denote its limit. At the cost of renaming iterates, we assume that  $\lim_{r \rightarrow \infty} h_r = \bar{h}$ . The above property still holds at the limit  $r \rightarrow \infty$ . Indeed, let  $\bar{u} \in \partial \rho_{\bar{y}}(\bar{x})$ . Since  $g$  is partly smooth, the mapping  $(x, y) \in \mathcal{N}_{\bar{x}} \times \mathcal{N}_{\bar{y}} \mapsto \partial \rho_y(x) = \partial g(x) + \frac{1}{\gamma}(x - y)$  is continuous. Therefore, there exists a sequence  $(u_r)$  such that  $u_r \in \partial \rho_{y_r}(x_r)$  and  $\lim_{r \rightarrow \infty} u_r = \bar{u}$ . We have for all  $r$ :  $\langle u_r, h_r \rangle \geq 0$ , which yields at the limit  $\langle \bar{u}, \bar{h} \rangle \geq 0$ . Thus  $\bar{h}$  separates 0 from  $\partial \rho_{\bar{y}}(\bar{x})$ , which contradicts our assumption.

**Conclusion.** We thus have a continuously differentiable function  $\hat{x}$  defined on a neighborhood of  $\bar{y}$  such that i)  $\hat{x}(\bar{y}) = \bar{x}$ , ii)  $\hat{x}(y)$  is a strong minimizer of  $\rho_y$  on  $\mathcal{M}$ , iii)  $0 \in \text{ri } \partial \rho_y(\hat{x}(y))$ .

This last point tells us that  $(y - \hat{x}(y))/\gamma - \nabla f(y) \in \partial g(\hat{x}(y))$ . The characterization of proximity by the optimality condition (Lemma 5) gives that  $\hat{x}(y) = \mathbf{prox}_{\gamma g}(y - \gamma \nabla f(y))$  for  $y$  close enough to  $\bar{y}$ .  $\square$

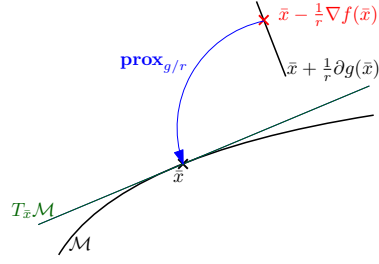
This theorem captures the localization properties of the proximal gradient operator. It also enables us to precisely define a condition under which a point can be localized. We formalize it in the definition of *r-structured critical points*.

**Definition 2** A point  $\bar{x}$  of a  $C^2$  manifold  $\mathcal{M}$  is a *r-structured critical point* for  $(f, g)$  if:

- i) proximal gradient stability:  $\bar{x} = \mathbf{prox}_{g/r}(\bar{x} - 1/r \nabla f(\bar{x}))$  ;
- ii) qualification condition:  $0 \in \text{ri}(\nabla f + \partial g)(\bar{x})$ ;
- iii) prox-regularity:  $g$  is  $r$ -prox-regular at  $\bar{x}$ ;
- iv) partial smoothness:  $g$  is partly-smooth at  $\bar{x}$  with respect to  $\mathcal{M}$ .

Definition 2 gives a precise characterization of the conditions allowing a critical point to benefit from a local structure stability by the proximal gradient operator. An illustration of this structure is shown in Fig. 1. While points ii,iii,iv are rather standard in the literature (see e.g. [15]), point i is less standard. First, this point is directly verified when  $g$  is convex, in which case any positive  $r$  is valid. In the nonconvex case, this condition is not often laid out precisely (still, it appears in the notion of identifiability in [19]) but it is necessary to take a stepsize range of  $(0, 1/r)$  in the proximal gradient around  $\bar{x}$ .

With this additional assumption, Theorem 3.2 applied to  $\bar{x} = \bar{y} \in \mathcal{M}$  a structured critical point for  $(f, g)$  gives us the following corollary.



**Fig. 1:** Illustration of a  $r$ -structured critical point. Point  $i$  is illustrated by the blue arrow, while point  $ii$  implies that the red cross is in the interior of the black segment. Partial smoothness appears in the fact that the black segment is perpendicular to the tangent plane of  $\mathcal{M}$  at  $\bar{x}$ .

**Corollary 1** *Let  $f$  be a  $\mathcal{C}^2$  function on  $\mathbb{R}^n$  and  $g$  a lower semi-continuous function on  $\mathbb{R}^n$ . Take  $\bar{x} \in \mathcal{M}$  a  $r$ -structured critical point for  $(f, g)$ . Then, for any  $\gamma \in (0, 1/r)$ , the proximal gradient map  $y \mapsto \mathbf{prox}_{\gamma g}(y - \gamma \nabla f(y))$  is  $\mathcal{C}^1$  and  $\mathcal{M}$ -valued near  $\bar{x}$ .*

### 3.3.2 Structure identification

Corollary 1 tells us that the proximal gradient mapping is  $\mathcal{M}$ -valued near  $\bar{x}$ . Thus, if the input  $(y_k)$  of the proximal gradient mapping converges to  $\bar{x}$ , it will reach this neighborhood in finite time, after which the proximal gradient of  $(y_k)$  will be  $\mathcal{M}$ -valued, as formalized in the following result.

**Corollary 2 (Identification)** *Let  $f$  be a  $\mathcal{C}^2$  function on  $\mathbb{R}^n$  and  $g$  a lower semi-continuous function on  $\mathbb{R}^n$ . Take  $\bar{x} \in \mathcal{M}$  a  $r$ -structured critical point for  $(f, g)$ . Then, for any  $\gamma \in (0, 1/r)$ , if the sequence  $(y_k)$  satisfies  $y_k \rightarrow \bar{x}$ , then  $x_k \triangleq \mathbf{prox}_{\gamma g}(y_k - \gamma \nabla f(y_k)) \in \mathcal{M}$  for  $k$  large enough.*

Since our algorithm is based on the use of the manifold in which the output of the proximity operator lies, it is natural to wonder if two manifolds could be identified at the same time. Fortunately, this is impossible with our assumptions as formalized in the next result, whose proof follows the rationale of [21, Cor. 3.2].

**Proposition 1 (Uniqueness of manifold)** *Consider a function  $g$ , two manifolds  $\mathcal{M}_1, \mathcal{M}_2$  and a point  $\bar{x} \in \mathcal{M}_1 \cap \mathcal{M}_2$  such that  $g$  is  $r$ -prox-regular at  $\bar{x}$  and partly-smooth relative to both manifolds  $\mathcal{M}_1$  and  $\mathcal{M}_2$ . Then, near  $\bar{x}$ ,  $\mathcal{M}_1 = \mathcal{M}_2$ .*

*Proof.* For the sake of eventual contradiction, let  $(x_k)$  denote any sequence converging to  $\bar{x}$  such that  $x_k \in \mathcal{M}_1 \setminus \mathcal{M}_2$  for all  $k$ . Since  $g$  is prox-regular, [37, Prop. 13.37] tells us that there is  $\bar{\gamma} > 0$  such that  $\bar{x} = \mathbf{prox}_{\bar{\gamma} g}(\bar{y})$  for some  $\bar{y} \in \bar{x} + \bar{\gamma} \text{ri } \partial g(\bar{x}) \in \mathbb{R}^n$  (since  $g$  is partly smooth,  $\partial g(\bar{x})$  has non-empty relative interior, and  $\bar{y}$  can be taken as  $\bar{x} + \bar{\gamma} \bar{v}$  for any  $\bar{v} \in \text{ri } \partial g(\bar{x})$  by reasoning as in the proof of [20, Th. 4]).

We can thus select a sequence  $v_k \in \partial g(x_k)$  converging to  $\bar{v} = (\bar{y} - \bar{x})/\bar{\gamma} \in \text{ri } \partial g(\bar{x})$  and define  $y_k = x_k + \gamma v_k$  for some  $\gamma \in (0, \bar{\gamma})$ . It is immediate to see that the sequence  $(y_k)$  converges to  $y^\gamma = (1 - (\gamma/\bar{\gamma}))\bar{x} + (\gamma/\bar{\gamma})\bar{y}$  and that  $y^\gamma$  can be made arbitrarily close to  $\bar{y}$  by taking  $\gamma$  close to  $\bar{\gamma}$ . Thus, we can consider that, properly choosing  $\gamma$ ,  $y_k$  reaches any neighborhood of  $\bar{y}$  in a finite number of iterations.

Lemma 5 then indicates that for  $k$  large enough, we have  $x_k = \mathbf{prox}_{\gamma g}(y_k)$ . Furthermore, Theorem 3.2 applied with  $f = 0$ ,  $\mathcal{M} = \mathcal{M}_2$  shows that  $\mathbf{prox}_{\gamma g}$  is  $\mathcal{M}_2$ -valued near  $\bar{y}$  which implies that  $x_k = \mathbf{prox}_{\gamma g}(y_k) \in \mathcal{M}_2$  for  $k$  large enough which contradicts  $x_k$  being in  $\mathcal{M}_1 \setminus \mathcal{M}_2$ .  $\square$

### 3.3.3 Super-linear convergence of Algorithm 1

Using the structure identification results above, we can guarantee that our method benefits from superlinear convergence, provided that the considered Riemannian method is superlinearly convergent locally around a limit point.

**Theorem 3.3** *Let Assumption 1 hold and take  $\gamma \in (0, 1/L)$ . Assume that Algorithm 1 generates a sequence  $(y_k)$  which admits at least one limit point  $\bar{x}$  such that:*

- i)  $\bar{x} \in \mathcal{M}$  is a  $r$ -structured critical point for  $(f, g)$  with  $r < 1/\gamma$ ;
- ii)  $\text{ManUp}_{\mathcal{M}}$  has superlinear convergence rate of order  $1 + \theta \in (1, 2)$  on a neighborhood of  $\bar{x}$  in  $\mathcal{M}$ .

Then, after some finite time:

- a) the full sequence  $(x_k)$  lies on  $\mathcal{M}$ ;
- b)  $x_k$  converges to  $\bar{x}$  superlinearly with the same order as  $\text{ManUp}$ :

$$\text{dist}_{\mathcal{M}}(x_{k+1}, \bar{x}) \leq c \text{dist}_{\mathcal{M}}(x_k, \bar{x})^{1+\theta} \quad \text{for some } c > 0.$$

*Proof.* Let us note  $\mathsf{T}(y) = \mathbf{prox}_{\gamma g}(y - \gamma \nabla f(y))$  for  $y \in \mathbb{R}^n$ .

The part i) of the assumptions enables us to show the existence of some neighborhood of  $\bar{x}$  on which the proximal gradient operation is  $\mathcal{M}$ -valued and Lipschitz continuous. More precisely, Corollary 1 tells us that there exists  $\delta_1 > 0$  and  $C > 0$  such that, for any  $y$  in  $\mathcal{B}(\bar{x}, \delta_1)$ , we have

$$\mathsf{T}(y) \in \mathcal{M} \quad \text{and} \quad \|\mathsf{T}(y) - \mathsf{T}(\bar{x})\| \leq C\|y - \bar{x}\|.$$

Now, if  $y$  additionally belongs to  $\mathcal{M}$ , we immediately deduce that there exists  $\varepsilon_1 > 0$  such that for any  $y$  in  $\mathcal{B}_{\mathcal{M}}(\bar{x}, \varepsilon_1)$ ,  $\mathsf{T}(y) \in \mathcal{M}$ ; but in addition, the Euclidean Lipschitz continuity can be translated into a Riemannian one (see Lemma 9) since for some  $\delta > 0$ ,

$$(1 - \delta)\text{dist}_{\mathcal{M}}(\mathsf{T}(y), \bar{x}) = (1 - \delta)\text{dist}_{\mathcal{M}}(\mathsf{T}(y), \mathsf{T}(\bar{x})) \leq \|\mathsf{T}(y) - \mathsf{T}(\bar{x})\| \leq C\|y - \bar{x}\| \leq C(1 + \delta)\text{dist}_{\mathcal{M}}(y, \bar{x}) \quad (3.2)$$

Hence, there is  $q_1 > 0$  such that for any  $y$  in  $\mathcal{B}_{\mathcal{M}}(\bar{x}, \varepsilon_1)$

$$\text{dist}_{\mathcal{M}}(\mathsf{T}(y), \bar{x}) = \text{dist}_{\mathcal{M}}(\mathsf{T}(y), \mathsf{T}(\bar{x})) \leq q_1 \text{dist}_{\mathcal{M}}(y, \bar{x}). \quad (3.3)$$

Then, the part ii) of the assumptions gives us the existence of  $\varepsilon_2, q_2 > 0$  and  $\theta \in (0, 1)$  such that, for any  $x$  in  $\mathcal{B}_{\mathcal{M}}(\bar{x}, \varepsilon_2)$ ,

$$\text{dist}_{\mathcal{M}}(\text{ManUp}_{\mathcal{M}}(x), \bar{x}) \leq q_2 \text{dist}_{\mathcal{M}}(x, \bar{x})^{1+\theta}. \quad (3.4)$$

Let us now take any  $x \in \mathcal{B}_{\mathcal{M}}(\bar{x}, \varepsilon)$  where  $\varepsilon = \min(\varepsilon_1, \varepsilon_2, (\varepsilon_1/q_2)^{\frac{1}{1+\theta}}, (q_2 q_1)^{-\frac{1}{\theta}})$ :

(i) Since  $x \in \mathcal{B}_{\mathcal{M}}(\bar{x}, \varepsilon_2)$ , the manifold update (3.4) yields

$$\text{dist}_{\mathcal{M}}(\text{ManUp}_{\mathcal{M}}(x), \bar{x}) \leq q_2 \text{dist}_{\mathcal{M}}(x, \bar{x})^{1+\theta} \leq q_2 \varepsilon^{1+\theta} \leq \varepsilon_1.$$

(ii) As  $\text{ManUp}_{\mathcal{M}}(x)$  lies in  $\mathcal{B}_{\mathcal{M}}(\bar{x}, \varepsilon_1)$ , the proximal gradient update (3.3) applied to  $y = \text{ManUp}_{\mathcal{M}}(x)$  gives

$$\begin{aligned} \text{dist}_{\mathcal{M}}(\mathbb{T}(\text{ManUp}_{\mathcal{M}}(x)), \bar{x}) &\leq q_1 \text{dist}_{\mathcal{M}}(\text{ManUp}_{\mathcal{M}}(x), \bar{x}) \\ &\leq q_1 q_2 \text{dist}_{\mathcal{M}}(x, \bar{x})^{1+\theta} + \leq q_1 q_2 \varepsilon^\theta \text{dist}_{\mathcal{M}}(x, \bar{x}). \end{aligned} \quad (3.5)$$

Since  $q_2 q_1 \varepsilon^\theta \leq 1$  by construction, this allows us to conclude that for any  $x \in \mathcal{B}_{\mathcal{M}}(\bar{x}, \varepsilon)$ , we have

$$\text{dist}_{\mathcal{M}}(\mathbb{T}(\text{ManUp}_{\mathcal{M}}(x)), \bar{x}) \leq \text{dist}_{\mathcal{M}}(x, \bar{x}). \quad (3.6)$$

We have thus proved the existence of a neighborhood  $\mathcal{B}_{\mathcal{M}}(\bar{x}, \varepsilon)$  of  $\bar{x}$  in  $\mathcal{M}$  which is stable for an iteration of Algorithm 1 and over which one iteration has a superlinear improvement of order  $1 + \theta$  (by (3.5)).

Finally, since  $\bar{x}$  is a limit point of  $(y_k)$ :

- (i) an iterate will reach any (Euclidean) ball around  $\bar{x}$  in finite time, notably there exists  $K < \infty$  such that  $y_K \in \mathcal{B}(\bar{x}, (1 - \delta)\varepsilon/C)$ ;
- (ii) equation (3.2) then tells us that  $\text{dist}_{\mathcal{M}}(\mathbb{T}(y_K), \bar{x}) \leq \varepsilon$  and thus  $x_k$  and  $y_k$  belong to  $\mathcal{B}_{\mathcal{M}}(\bar{x}, \varepsilon)$  for all  $k > K$  by (3.6) and the definition of  $\varepsilon$ .

We may thus conclude that  $x_{k+1} = \mathbb{T}(y_k) \in \mathcal{M}$  for all  $k \geq K$ , and, using (3.5),

$$\text{dist}_{\mathcal{M}}(x_{k+1}, \bar{x}) \leq q \text{dist}_{\mathcal{M}}(x_k, \bar{x})^{1+\theta}$$

for all  $k > K$  and  $q = q_1 q_2$ . □

## 4 Newton acceleration

In this section, we investigate the possibilities of manifold updates for our method. Since the procedure  $\text{ManUp}_{\mathcal{M}}$  is required not to degrade function value, we study in Section 4.1 the use of line-search in our context. Then, we show in Sections 4.2 and 4.3 how to use Riemannian (truncated) Newton method within our framework and derive superlinear/quadratic convergence guarantees.

### 4.1 Ensuring functional descent while preserving local rates: Linesearches

As developed in the previous section,  $\text{ManUp}_{\mathcal{M}}$  should produce an update that (i) lives on  $\mathcal{M}$ , (ii) enjoys a superlinear local convergence rate, and (iii) does not degrade function value. For this last point, we consider a simple line-search and we prove that, under mild assumptions, it helps to find a point which decreases function value, and retains the favorable local properties. Surprisingly, this result does not appear in the standard references on Riemannian optimization. We provide here the necessary developments inspired from the classical textbook [18].

Standing at point  $x \in \mathcal{M}$  with a proposed direction  $\eta \in T_x \mathcal{M}$ , a stepsize  $\alpha > 0$  is *acceptable* if it satisfies the following *Armijo* condition, aimed at ensuring sufficient decrease in function value:

$$F(\mathbb{R}_x(\alpha\eta)) \leq F(x) + m_1 \alpha \langle \text{grad } F(x), \eta \rangle, \quad \text{for } 0 < m_1 < 1/2. \quad (4.1)$$

The conditions under which stepsizes satisfying the Armijo rule exist are discussed in [18, Sec 6.3], the following lemma can then be derived.



**Lemma 1** *Let Assumption 1 hold and consider a manifold  $\mathcal{M}$  equipped with a retraction  $R$  and a pair  $(x, \eta) \in T\mathcal{B}$ . If  $F$  is differentiable on  $\mathcal{M}$  at  $x$ ,  $\langle \text{grad } F(x), \eta \rangle < 0$  and  $m_1 < 1$ , then there exists  $\hat{\alpha} > 0$  such that any step size  $\alpha \in (0, \hat{\alpha})$  is acceptable by the Armijo rule (4.1).*

*Proof.* We adapt a part of the proof of [18, Th. 6.3.2] for the Armijo rule and the Riemannian setting. Since  $m_1 < 1$ , for any  $\alpha$  sufficiently small there holds

$$F \circ R_x(\alpha\eta) \leq F \circ R_x(0) + m_1 D(F \circ R_x)(0)[\alpha\eta] = F(x) + m_1\alpha \langle \text{grad } F(x), \eta \rangle.$$

Since  $F$  is bounded below, there exists some smallest positive  $\hat{\alpha}$  such that  $F(R_x(\hat{\alpha}\eta)) = F(x) + m_1\hat{\alpha} \langle \text{grad } F(x), \eta \rangle$ . Thus all stepsizes in  $(0, \hat{\alpha})$  are acceptable by the Armijo rule (4.1).  $\square$

In addition, a linesearch performed near a minimizer with a Newton direction (or in general a superlinearly converging method) should accept the unit stepsize, so that a full step may be taken. This is the case when the Riemannian Hessian around this minimizer is positive definite as stated by the next lemma.

**Lemma 2** *Let Assumption 1 hold and consider a manifold  $\mathcal{M}$  equipped with a retraction  $R$ , a point  $x^* \in \mathcal{M}$  and a pair  $(x, \eta) \in T\mathcal{B}$ . Assume that  $F$  is twice differentiable on  $\mathcal{M}$  near  $x^*$  and  $x^*$  is a local minimizer on  $\mathcal{M}$ , that is  $\text{Hess } F(x^*)$  is positive definite. If the direction  $\eta$  brings a superlinear improvement towards  $x^*$ , that is  $\text{dist}_{\mathcal{M}}(R_x(\eta), x^*) = o(\text{dist}_{\mathcal{M}}(x, x^*))$  as  $x \rightarrow x^*$ , and  $0 < m_1 < 1/2$ , then  $\eta$  is acceptable by the Armijo rule (4.1) with unit stepsize  $\alpha = 1$ .*

*Proof.* The result is exactly Theorem B.1 applied to the composite function  $g+f$ .  $\square$

We consider a *backtracking* linesearch for finding an acceptable stepsize  $\alpha$ . The unit stepsize is first tried, and then the search space is reduced geometrically so that superlinear steps are taken whenever possible. The procedure terminates with a finite number of iterations. Besides, the backtracking nature of the linesearch ensures that step sizes are not overly small. In practice, we use exactly [18, Alg. A6.3.1], which features polynomial interpolation of  $F$  in the search space.

## 4.2 Riemannian Newton & quadratic convergence

We now construct a manifold update based on the Riemannian Newton method [2, Chap. 6], this is the simplest method enjoying a local quadratic convergence rate. It consists in finding a direction  $d \in T_x\mathcal{M}$  that minimizes the second order model (2.4) of  $F$  at point  $x \in \mathcal{M}$ , or equivalently that solves Newton equation; see [12, Sec. 6.2].

In order to ensure that the Newton direction is well-defined and provides descent, we assume that the Riemannian Hessian of  $F$  is positive definite at each iterate (relative to the working manifold).

**Theorem 4.1** *Let Assumption 1 hold and take  $\gamma \in (0, 1/L)$ . Consider the sequence of iterates  $(x_k)$  generated by Algorithm 1 equipped with the Riemannian Newton manifold update (Algorithm 2). If  $\text{Hess } F(x_k)$  is positive definite at each step, then all limit points of  $(x_k)$  are critical points of  $F$  and share the same functional value.*

*Furthermore, assume that the sequence  $(y_k)$  admits at least one limit point  $x^*$  such that*

**Algorithm 2** ManUp-Newton**Require:** Manifold  $\mathcal{M}$ , point  $x \in \mathcal{M}$ .1: Find  $d$  in  $T_x\mathcal{M}$  that solves

$$\text{grad } F(x) + \text{Hess } F(x)[d] = 0. \quad (\text{Newton equation})$$

2: Find  $\alpha$  satisfying the Armijo condition (4.1) with direction  $d$ .3: **return**  $y = R_x(\alpha d)$ 

- i)*  $x^* \in \mathcal{M}$  is a  $r$ -structured critical point for  $(f, g)$  with  $r < 1/\gamma$ ;  
*ii)*  $\text{Hess}_{\mathcal{M}} F(x^*) \succ 0$  and  $\text{Hess}_{\mathcal{M}} F$  is locally Lipschitz around  $x^*$ .

Then, after some finite time,

- a)* the sequence  $(x_k)$  lies on  $\mathcal{M}$ ;  
*b)*  $x_k$  converges to  $x^*$  quadratically, ie.

$$\text{dist}_{\mathcal{M}}(x_{k+1}, x^*) \leq \text{dist}_{\mathcal{M}}(x_k, x^*)^2.$$

*Proof.* As the Riemannian hessian is assumed to be positive definite, Newton's direction is a descent direction:

$$\langle \text{grad } F(x_k), d_k \rangle = -\langle \text{grad } F(x_k), \text{Hess } F(x_k)^{-1} \text{grad } F(x_k) \rangle < 0.$$

The Riemannian Newton manifold step is therefore well-defined, and stepsizes acceptable by the linesearch exist by Lemma 1, so that the manifold update is well defined and provides descent. Thus, Theorem 3.1 ensures that every accumulation point of the iterate sequence is a critical point for  $F$ .

Furthermore, assumption ii) ensures that Lemma 10 applies: the Riemannian Newton direction  $d$  computed in step 1 of Algorithm 2 provides a quadratic improvement towards  $x^*$  on a neighborhood of  $x^*$  on  $\mathcal{M}$ . Besides, the linesearch returns the unit-stepsize after some finite time:  $\alpha = 1$  is tried first, and is acceptable for direction providing superlinear improvement by Lemma 2. Thus the whole Riemannian Newton update provides quadratic improvement after some finite time. Using this and assumption i), Theorem 3.3 applies and yields the results.  $\square$

This theorem states that alternating proximal gradient steps and Riemannian Newton steps on the identified manifold converges quadratically to structured points with virtually the same assumptions the Euclidean Newton method. Notably, a point  $x^* \in \mathcal{M}$  that is both  $r$ -structured critical for  $(f, g)$  and such that  $\text{Hess}_{\mathcal{M}} F(x^*) \succ 0$  is a strong local minimizer of  $F$  on  $\mathbb{R}^n$ , that is there exists some  $\varepsilon > 0$  such that, for  $x$  near  $x^*$ ,

$$F(x) \geq F(x^*) + \frac{\varepsilon}{2} \|x - x^*\|^2.$$

However, we can notice two issues of Newton's method (present in both the Euclidean and Riemannian settings): (i) at each iteration a linear system has to be solved to produce the Newton direction  $d$ ; and (ii) the direction  $d$  does not always provide descent without the strong assumption that the Riemannian Hessian is positive definite at each step. Truncated versions of Newton's method overcome these issues, as we see for our context, in the next section.

### 4.3 Riemannian Truncated Newton & superlinear convergence

We now consider a manifold update based on a truncated Newton procedure. (Riemannian) Truncated Newton consists in solving (Newton equation) partially by using a (Riemannian) conjugate gradient procedure so that whenever the resolution of (Newton equation) is stopped, the resulting direction provides descent on the function. This method was proposed first by Dembo and Steinhaug [17] as a specific case of inexact Newton method [16]. [34] discusses it as *line search Newton-CG*, and a review can be found in [24].

The quality of the truncated Newton direction is controlled by a parameter  $\eta \in [0, 1)$  which bounds the ratio of residual and gradient norms:

$$\|\text{grad } F(x) + \text{Hess } F(x)[d]\| \leq \eta \|\text{grad } F(x)\|. \quad (\text{Inexact Newton eq.})$$

$\eta = 0$  allows only the exact Newton step, while  $\eta = 1$  allows a broad set of directions, including  $d = 0$ .

---

#### Algorithm 3 ManUp-Newton-CG

---

**Require:** Manifold  $\mathcal{M}$ , point  $x \in \mathcal{M}$ , convergence defining parameter  $\theta \in (0, 1]$ .

- 1: Let  $\eta = \|\text{grad } F(x)\|^\theta$
- 2: Find  $d$  that solves

$$\|\text{grad } F(x) + \text{Hess } F(x)[d]\| \leq \eta \|\text{grad } F(x)\|. \quad (\text{Inexact Newton eq.})$$

- 3: Find  $\alpha$  satisfying the Armijo condition (4.1) with direction  $d$ .
  - 4: **return**  $y = R_x(\alpha d)$
- 

**Theorem 4.2** *Let Assumption 1 hold and take  $\gamma \in (0, 1/L)$ . Consider the sequence of iterates  $(x_k)$  generated by Algorithm 1 equipped with the Riemannian Truncated Newton manifold update (Algorithm 3). Then all limit points of  $(x_k)$  are critical points of  $F$  and share the same function value.*

*Furthermore, assume that sequence  $(y_k)$  admits at least one limit point  $x^*$  such that*

- i)  $x^* \in \mathcal{M}$  is a  $r$ -structured critical point for  $(f, g)$  with  $r < 1/\gamma$ ;*
- ii)  $\text{Hess}_{\mathcal{M}} F(x^*) \succ 0$  and  $\text{Hess}_{\mathcal{M}} F$  is locally Lipschitz around  $x^*$ .*
- iii) we take  $\eta_k = \mathcal{O}(\|\text{grad } F(x_k)\|^\theta)$ , for some  $\theta \in (0, 1]$ .*

*Then,*

- a) for  $k$  large enough, the full sequence  $(x_k)$  lies on  $\mathcal{M}$ ;*
- b)  $x_k$  converges to  $x^*$  superlinearly with order  $1 + \theta$ : for large  $k$ , there exist  $c > 0$ ,*

$$\text{dist}_{\mathcal{M}}(x_{k+1}, x^*) \leq c \text{dist}_{\mathcal{M}}(x_k, x^*)^{1+\theta}.$$

*Proof.* The direction provided by the approximate resolution of (Inexact Newton eq.) is a descent direction by Lemma 11. Stepsizes acceptable by the linesearch thus exist by Lemma 1, so that the Riemannian Newton-CG manifold update is well defined and provides descent. Thus, Theorem 3.1 ensures that every accumulation point of the iterate sequence is a critical point for  $F$ .

Furthermore, assumptions ii) and iii) ensure that Lemma 12 applies: the Riemannian Newton CG direction  $d$  computed in step 1 of Algorithm 3 provides a

superlinear improvement towards  $x^*$  on a neighborhood of  $x^*$  on  $\mathcal{M}$ . Besides, the linesearch returns the unit-stepsizes after some finite time:  $\alpha = 1$  is tried first, and is acceptable for direction providing superlinear improvement by Lemma 2. Thus the whole Riemannian Newton-CG update eventually provides fast improvement. Using this and assumption i), Theorem 3.3 applies and yields the results.  $\square$

## 5 Numerical illustrations

In this section, we illustrate the effect of Newton acceleration, in terms of identification of the final manifold and local convergence. The algorithms and problems are implemented in Julia [10]; experiments may be reproduced using the code available at <https://github.com/GillesBareilles/NewtonRiemannAccel-ProxGrad>.

We consider Algorithm 1 equipped with either the Newton update of Algorithm 2, denoted ‘Alt. Newton’ or the truncated Newton update of Algorithm 3, denoted ‘Alt. Truncated Newton’. Both algorithms use a Conjugate Gradient procedure to solve the Newton equation, either exactly or not. Each CG iteration requires one (Riemannian) Hessian-vector product, avoiding to form a Hessian matrix. These methods are compared to the Proximal Gradient and the Accelerated Proximal Gradient, which serve as baseline.

We report the numerical results in figures showing a) the suboptimality  $F(x_k) - F(x^*)$  of the current iterate  $x_k$  versus time, and b) the dimension of the current manifold  $\mathcal{M}_k \ni x_k$  versus iteration. We also report a table comparing the algorithms at the first iteration that makes suboptimality lower than tolerances  $10^{-3}$  and  $10^{-9}$  for various measures summarized in the following table:

$F(x_k) - F(x^*)$	Suboptimality at current iteration.
#prox. grad. steps	Number of proximal gradient steps, each involve computing $\nabla f(\cdot)$ and $\text{prox}_{\gamma g}(\cdot)$ once.
#ManUp steps	Number of Riemannian steps, each involve computing $\text{grad } F(\cdot)$ once and $\text{Hess } F(\cdot)[\cdot]$ multiple times (one per Conjugate Gradient iteration).
#Hess $F(\cdot)[\cdot]$	Number of Riemannian Hessian-vector products, approximates the effort spent in manifold updates since algorithm started.
# $f$	Number of calls to $f(x)$ , one per iteration, some for the Riemannian linesearch, some for the backtracking estimation of the Lipschitz constant of $\nabla f$ .
# $g$	Number of calls to $g(x)$ , one per iteration, some for the Riemannian linesearch.

The proximal gradient updates, present in all methods, include a backtracking procedure that maintains an estimate of the Lipschitz constant of  $\nabla f$ , so that the proximal gradient steplength is taken as the inverse of that estimate. The Conjugate Gradient used to solve (Newton equation) and (Inexact Newton eq.) follows [12, Alg. 6.2]; it is stopped when the (in)exactness criterion is met, or after 50 iterations for the logistic problem and 150 for the trace-norm one, or when the inner direction  $d$  makes the ratio  $\langle \text{Hess } F(x_k)[d], d \rangle / \|d\|$  small. The manifold updates are completed by a backtracking linesearch started from unit stepsizes, a direct implementation of [18, Alg. 6.3.1].

### 5.1 Two-dimensional nonsmooth nonconvex example

We consider the piecewise quadratic problem of [27]:

$$\min_{x \in \mathbb{R}^2} 2x_1^2 + x_2^2 + |x_1^2 - x_2|.$$

The objective function is partly-smooth relative to the parabola  $\mathcal{M} \triangleq \{x : x_2 = x_1^2\}$ , for which an expression for the tangent space, the orthogonal projection on tangent space, a second-order retraction and conversion from Euclidean gradients and hessian-vector products to Riemannian ones are readily available. We detail here the different oracles of  $f(x) \triangleq 2x_1^2 + x_2^2$  and  $g(x) \triangleq |x_1^2 - x_2|$ :

- *proximity operator*: For  $\gamma < 1/2$ , there holds

$$\mathbf{prox}_{\gamma g}(x) = \begin{cases} \left(\frac{x_1}{1+2\gamma}, x_2 + \gamma\right) & \text{if } x_2 \leq \frac{x_1^2}{(1+2\gamma)^2} - \gamma \\ \left(\frac{x_1}{1+4\gamma t - 2\gamma}, x_2 + 2\gamma t - \gamma\right) & \text{if } \frac{x_1^2}{(1+2\gamma)^2} - \gamma \leq x_2 \leq \frac{x_1^2}{(1-2\gamma)^2} + \gamma \\ \left(\frac{x_1}{1-2\gamma}, x_2 - \gamma\right) & \text{if } \frac{x_1^2}{(1-2\gamma)^2} + \gamma \leq x_2 \end{cases}$$

where  $t$  solves  $x_2^2 + (-2\gamma t + \gamma - x_2)(1 + 4\gamma t - 2\gamma)^2 = 0$ .

- *Riemannian gradient and hessian*: Since  $g$  is identically null on  $\mathcal{M}$ , for any point  $(x, \eta) \in T\mathcal{B}$ ,

$$\mathbf{grad} g(x) = 0 \quad \text{and} \quad \mathbf{Hess} g(x)[\eta] = 0.$$

Besides, euclidean gradient and hessian-vector product are converted to Riemannian ones using equations (2.2) and (2.3):

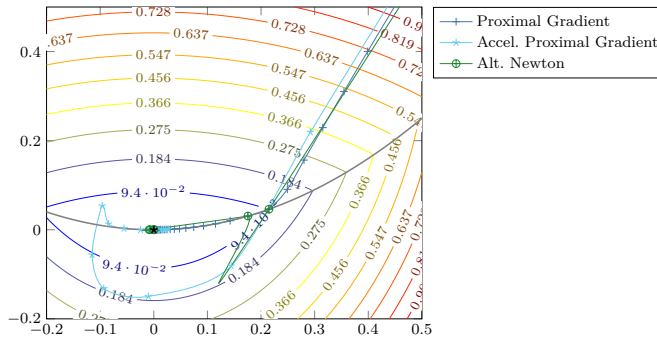
$$\begin{aligned} \mathbf{grad} f(x) &= \mathbf{proj}_x(\nabla f(x)) \\ \mathbf{Hess} f(x)[\eta] &= \mathbf{proj}_x \left( \nabla^2 f(x)[\eta] - \begin{pmatrix} 2\eta_1 \\ 0 \end{pmatrix} \left\langle \nabla f(x), \begin{pmatrix} 2x_1 \\ -1 \end{pmatrix} \right\rangle \frac{1}{1 + 4x_1^2} \right), \end{aligned}$$

and the orthogonal projection onto  $T_x\mathcal{M}$  writes

$$\mathbf{proj}_x(d) = d - \left\langle d, \begin{pmatrix} 2x_1 \\ -1 \end{pmatrix} \right\rangle \frac{1}{1 + 4x_1^2} \begin{pmatrix} 2x_1 \\ -1 \end{pmatrix}$$

We run the proximal gradient, its accelerated counterpart and [Algorithm 1](#) with the Newton update [Algorithm 2](#). The proximal gradient steps of all algorithms have a constant step-size  $\gamma = 0.05$ , all algorithms are started from point  $(2, 3)$ , and converge to the minimizer of the function  $(0, 0)$ .

**Observations** The iterates are displayed in [Fig. 2](#). The first Proximal Gradient step (not visible on the figure) yields an iterate lying on the parabola. While the (accelerated) proximal gradient iterates leave the parabola to reach it later, the Alt. Newton iterates remains on the parabola until convergence. The quadratic convergence behavior appears clearly as two Newton manifold updates bring suboptimality below  $10^{-3}$ , and one additional step gets below  $10^{-12}$ . The Proximal Gradient iterates reach the parabola in finite time, and then converge linearly towards  $x^*$  on the parabola. The Accelerated Proximal Gradient iterates feature two known negative properties: they “overshoot” the manifold to be identified twice before reaching it, and they oscillate around  $x^*$  in the parabola (see [5]).



Algorithm	Tolerance	$F(x_k) - F(x^*)$	#prox. grad. steps	#ManUp steps	#Hess $F(\cdot)[\cdot]$	#f	#g
Prox. Gradient	$1 \cdot 10^{-3}$	$7.74 \cdot 10^{-4}$	29	–	–	30	30
Prox. Gradient	$1 \cdot 10^{-9}$	$7.59 \cdot 10^{-10}$	60	–	–	61	61
Accel. Prox. Gradient	$1 \cdot 10^{-3}$	$9.63 \cdot 10^{-4}$	16	–	–	17	17
Accel. Prox. Gradient	$1 \cdot 10^{-9}$	$5.18 \cdot 10^{-10}$	63	–	–	64	64
Alt. Newton	$1 \cdot 10^{-3}$	$1.49 \cdot 10^{-4}$	2	2	10	7	7
Alt. Newton	$1 \cdot 10^{-9}$	$8.75 \cdot 10^{-13}$	3	3	15	10	10

Fig. 2: nonsmooth, nonconvex example

## 5.2 $\ell_1$ -regularized logistic problem

We now turn to the  $\ell_1$ -regularized logistic problem:

$$\min_{x \in \mathbb{R}^n} \frac{1}{m} \sum_{i=1}^m \log(y_i \sigma(\langle A_i, x \rangle)) + \lambda \|x\|_1, \quad (5.1)$$

where  $A \in \mathbb{R}^{m \times n}$ ,  $y \in \{-1, 1\}^m$ ,  $\lambda$  denotes a positive scalar and  $\sigma$  the sigmoid map  $x \mapsto 1/(1 + e^{-x})$ . The nonsmooth part  $g(x) = \lambda \|x\|_1$  is described in Section 2.3.

We consider an instance of (5.1) where  $n = 4 \cdot 10^3$ ,  $m = 8 \cdot 10^3$ ,  $\lambda = 10^{-2}$  and the final manifold has dimension 249. The coefficients of  $A$  are drawn independently following a normal law. The measurements  $y$  are generated as follows: a sparse random source signal  $s$  is generated, where each coordinate either is drawn following a normal law or is set to 0, with probability 1/2. Measurement  $y_i$  is 1 with probability  $(1 + \sigma(\langle A_i, s \rangle))/2$ , and  $-1$  otherwise.

**Observations** The experiments are presented in Fig. 3. The optimal manifold is identified roughly around iteration 200 for all methods except for Proximal Gradient, which needs about 1000 iterations. The two baselines Proximal Gradient and its accelerated version show linear convergence, with a better rate for the non accelerated version once the final manifold is reached. Alt. Truncated Newton shows superlinear acceleration, while Alt. Newton fails to converge in the given time budget. This fact, along with the counts of Hessian-vector products, shows the interest of the inexact solve of Newton equation performed by the Truncated Newton procedure Algorithm 3 as opposed to the exact solve performed by the Newton procedure Algorithm 2.

As iterations grow, the (Accelerated) Proximal Gradient identifies manifolds of decreasing dimension in a roughly monotonical way. Alt. Truncated Newton behaves differently: after identifying monotonically manifolds of dimension lower than 2000,

the dimension of the current manifold jumps to about 3000 for about 10 iterations, to finally reach quickly the final manifold. We believe that this partial loss of identified structure is caused by iterates getting close to a non differentiable point of  $F$  on the current manifold, e.g. having one non-null but very small coordinate. There, the second-order Taylor extension is valid on a small set, broadly speaking only up to the non-differentiability point; however it may lead to a Newton step that lies outside that set, thus driving the iterate away from more structured points instead of identifying them. The same behavior occurs for Alt. Newton. This difficulty can be related to the well-known problem of constraint activation in nonlinear programming. Despite this behavior, [Algorithm 1](#) retains a good rate overall.

### 5.3 Trace-norm regularized problem

We consider the following matrix regression problem:

$$\min_{x \in \mathbb{R}^{n_1 \times n_2}} \frac{1}{2} \sum_{i=1}^m (\langle A_i, x \rangle - y_i)^2 + \lambda \|x\|_*, \quad (5.2)$$

where  $A_i \in \mathbb{R}^{n_1 \times n_2}$  for  $i = 1, \dots, m$ ,  $y \in \mathbb{R}^m$  and  $\lambda$  denotes a positive scalar. The nonsmooth part  $g(x) = \lambda \|x\|_*$  is described in [Section 2.3](#).

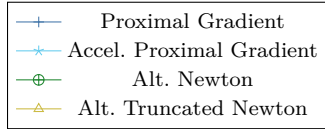
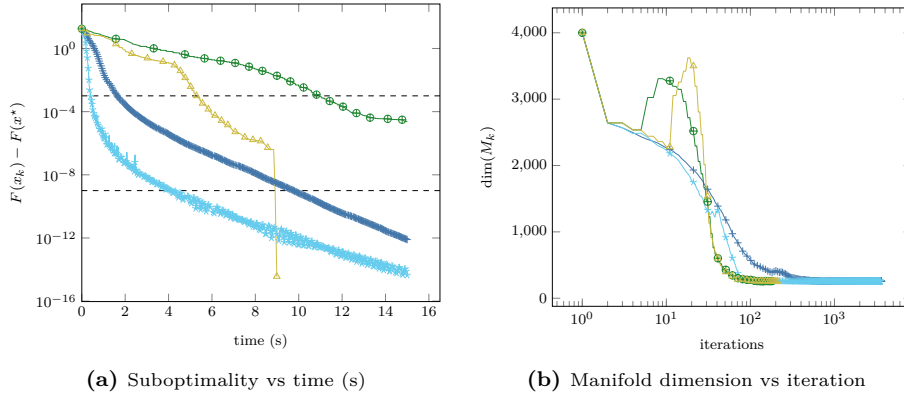
We consider an instance of (5.2) where  $n_1 = 10$ ,  $n_2 = 12$ ,  $m = 60$ ,  $\lambda = 1 \cdot 10^{-2}$  and the final manifold is that of matrices of rank 6. The coefficients of each matrix  $A_i$  are drawn independently following a normal law. The measurements  $y$  are generated as follows: a sparse random source signal  $s$  is generated as the projection of a matrix which coordinates are drawn independently following a normal law; the measurement  $y_i$  is then  $\langle A_i, s \rangle + \xi_i$ , where  $\xi_i$  follows a centered gaussian law with variance  $0.01^2$ .

**Observations** The experiments are presented in [Fig. 4](#). We see on [Fig. 4a](#) that the Proximal Gradient algorithm and its accelerated version converge sublinearly, which is to be related to the lack of strong convexity of the objective problem. Alt. Truncated Newton converges superlinearly, and shows the interest of the Newtonian acceleration. [Figure 4b](#) shows that the Proximal Gradient does not reach the final optimal manifold within the budget of iterations; similarly for the Newton method, within the budget of time.

## 6 Concluding remarks

This paper proposes and studies a nonsmooth optimization algorithm exploiting the underlying smooth geometry revealed by the proximal operator. This simple method alternates between a proximal gradient step providing identification and a Riemann Newton acceleration providing superlinear convergence. This algorithm has two special features: (i) it does not rely on prior knowledge on the final manifold, (ii) its convergence is guaranteed in the (structured) nonconvex case.

Several extensions of this algorithm are possible; specifically, both building blocks can be refined: other Newton accelerations could be considered (e.g. trust-region [1], BFGS [22], cubic regularization [3]) as well as other proximal algorithms (e.g. prox-(quasi-)Newton [9, 25], fast proximal gradient [8]). We focused here on the simplest Newton acceleration to highlight the ideas and the working horses of our approach.



Algorithm	Tolerance	$F(x_k) - F(x^*)$	#prox. grad. steps	#ManUp steps	#Hess $F(\cdot)[\cdot]$	#f	#g
Prox. Gradient	$1 \cdot 10^{-3}$	$9.96 \cdot 10^{-4}$	357	—	—	779	358
Prox. Gradient	$1 \cdot 10^{-9}$	$9.97 \cdot 10^{-10}$	2,306	—	—	4,677	2,307
Accel. Prox. Gradient	$1 \cdot 10^{-3}$	$9.26 \cdot 10^{-4}$	90	—	—	246	91
Accel. Prox. Gradient	$1 \cdot 10^{-9}$	$9.9 \cdot 10^{-10}$	953	—	—	1,972	954
Alt. Newton	$1 \cdot 10^{-3}$	$9.76 \cdot 10^{-4}$	62	61	6,303	556	427
Alt. Newton	$1 \cdot 10^{-9}$	—	—	—	—	—	—
Alt. Truncated Newton	$1 \cdot 10^{-3}$	$9.56 \cdot 10^{-4}$	51	50	2,616	437	321
Alt. Truncated Newton	$1 \cdot 10^{-9}$	$3.77 \cdot 10^{-15}$	105	105	5,091	742	572

Fig. 3: Logistic- $\ell_1$  problem

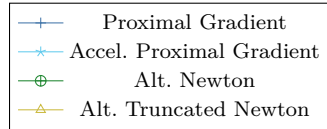
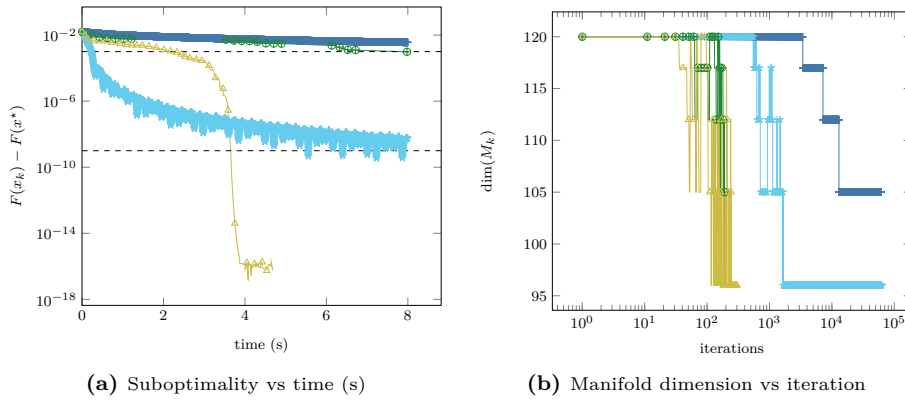
Acknowledgements

GB and FI benefited from the support of the ANR JCJC project *STROLL* (ANR-19-CE23-0008).

References

1. Absil, P.A., Baker, C.G., Gallivan, K.A.: Trust-region methods on riemannian manifolds. *Foundations of Computational Mathematics* **7**(3), 303–330 (2007)
2. Absil, P.A., Mahony, R., Sepulchre, R.: *Optimization algorithms on matrix manifolds*. Princeton University Press (2009)
3. Agarwal, N., Boumal, N., Bullins, B., Cartis, C.: Adaptive regularization with cubics on manifolds. *Mathematical Programming* (2020)
4. Bach, F., Jenatton, R., Mairal, J., Obozinski, G.: Optimization with sparsity-inducing penalties. *Foundations and Trends® in Machine Learning* **4**(1), 1–106 (2012)
5. Bareilles, G., Iutzeler, F.: On the interplay between acceleration and identification for the proximal gradient algorithm. *Computational Optimization and Applications* **77**, 351–378 (2020)
6. Bauschke, H.H., Combettes, P.L.: *Convex analysis and monotone operator theory in Hilbert spaces*. Springer Science & Business Media (2011)
7. Beck, A.: *First-order methods in optimization*, vol. 25. SIAM (2017)





Algorithm	Tolerance	$F(x_k) - F(x^*)$	#prox. grad. steps	#ManUp steps	#Hess $F(\cdot)$ [·]	#f	#g
Prox. Gradient	$1 \cdot 10^{-3}$	—	—	—	—	—	—
Prox. Gradient	$1 \cdot 10^{-9}$	—	—	—	—	—	—
Accel. Prox. Gradient	$1 \cdot 10^{-3}$	$9.99 \cdot 10^{-4}$	1,489	—	—	3,073	1,490
Accel. Prox. Gradient	$1 \cdot 10^{-9}$	$9.86 \cdot 10^{-10}$	43,283	—	—	86,661	43,284
Alt. Newton	$1 \cdot 10^{-3}$	$9.83 \cdot 10^{-4}$	93	93	28,063	873	687
Alt. Newton	$1 \cdot 10^{-9}$	—	—	—	—	—	—
Alt. Truncated Newton	$1 \cdot 10^{-3}$	$9.7 \cdot 10^{-4}$	76	76	16,342	738	568
Alt. Truncated Newton	$1 \cdot 10^{-9}$	$2.27 \cdot 10^{-11}$	128	128	27,786	1,101	879

Fig. 4: Trace-norm problem

8. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences* **2**(1), 183–202 (2009)
9. Becker, S., Fadili, J.: A quasi-newton proximal splitting method. *Advances in neural information processing systems* **25**, 2618–2626 (2012)
10. Bezanson, J., Edelman, A., Karpinski, S., Shah, V.B.: Julia: A fresh approach to numerical computing. *SIAM review* **59**(1), 65–98 (2017)
11. Bonnans, J.F., Gilbert, J.C., Lemaréchal, C., Sagastizábal, C.A.: *Numerical optimization: theoretical and practical aspects*. Springer Science & Business Media (2006)
12. Boumal, N.: *An introduction to optimization on smooth manifolds*. Available online (2020). URL <http://www.nicolasboumal.net/book>
13. Briant, O., Lemaréchal, C., Meurdesoif, P., Michel, S., Perrot, N., Vanderbeck, F.: Comparison of bundle and classical column generation. *Mathematical Programming* **113**(2), 299–344 (2008)
14. Burke, J.V., Moré, J.J.: On the identification of active constraints. *SIAM Journal on Numerical Analysis* **25**(5), 1197–1211 (1988)
15. Daniilidis, A., Hare, W., Malick, J.: Geometrical interpretation of the predictor-corrector type algorithms in structured optimization problems. *Optimization* **55**(5-6), 481–503 (2006)
16. Dembo, R.S., Eisenstat, S.C., Steihaug, T.: Inexact newton methods. *SIAM Journal on Numerical analysis* **19**(2), 400–408 (1982)
17. Dembo, R.S., Steihaug, T.: Truncated-newtono algorithms for large-scale unconstrained optimization. *Mathematical Programming* **26**(2), 190–212 (1983)
18. Dennis Jr, J.E., Schnabel, R.B.: *Numerical methods for unconstrained optimization and nonlinear equations*. SIAM (1996)

19. Drusvyatskiy, D., Lewis, A.S.: Optimality, identifiability, and sensitivity. *Mathematical Programming* **147**(1-2), 467–498 (2014)
20. Hare, W., Sagastizábal, C.: Computing proximal points of nonconvex functions. *Mathematical Programming* **116**(1-2), 221–258 (2009)
21. Hare, W.L., Lewis, A.S.: Identifying active constraints via partial smoothness and prox-regularity. *Journal of Convex Analysis* **11**(2), 251–266 (2004)
22. Huang, W., Absil, P.A., Gallivan, K.A.: A riemannian bfgs method without differentiated retraction for nonconvex optimization problems. *SIAM Journal on Optimization* **28**(1), 470–495 (2018)
23. Iutzeler, F., Malick, J.: Nonsmoothness in machine learning: specific structure, proximal identification, and applications. *Set-Valued and Variational Analysis* **28**(4), 661–678 (2020)
24. Knoll, D.A., Keyes, D.E.: Jacobian-free newton–krylov methods: a survey of approaches and applications. *Journal of Computational Physics* **193**(2), 357–397 (2004)
25. Lee, J.D., Sun, Y., Saunders, M.A.: Proximal newton-type methods for minimizing composite functions. *SIAM Journal on Optimization* **24**(3), 1420–1443 (2014)
26. Lemaréchal, C., Oustry, F., Sagastizábal, C.: The u-lagrangian of a convex function. *Transactions of the American mathematical Society* **352**(2), 711–729 (2000)
27. Lewis, A., Wylie, C.: A simple newton method for local nonsmooth optimization. *arXiv preprint arXiv:1907.11742* (2019)
28. Lewis, A.S.: Active sets, nonsmoothness, and sensitivity. *SIAM Journal on Optimization* **13**(3), 702–725 (2002)
29. Lewis, A.S., Liang, J.: Partial smoothness and constant rank. *arXiv preprint arXiv:1807.03134* (2018)
30. Liang, J., Fadili, J., Peyré, G.: Activity identification and local linear convergence of forward–backward-type methods. *SIAM Journal on Optimization* **27**(1), 408–437 (2017)
31. Mifflin, R., Sagastizábal, C.: A vu-algorithm for convex minimization. *Mathematical programming* **104**(2-3), 583–608 (2005)
32. Miller, S.A., Malick, J.: Newton methods for nonsmooth convex minimization: connections among u-lagrangian, riemannian newton and sqp methods. *Mathematical programming* **104**(2-3), 609–633 (2005)
33. Moreau, J.J.: Proximité et dualité dans un espace hilbertien. *Bulletin de la Société mathématique de France* **93**, 273–299 (1965)
34. Nocedal, J., Wright, S.: *Numerical optimization*. Springer Science & Business Media (2006)
35. Oliveira, W., Sagastizábal, C., Scheimberg, S.: Inexact bundle methods for two-stage stochastic programming. *SIAM Journal on Optimization* **21**(2), 517–544 (2011). DOI <http://dx.doi.org/10.1137/100808289>
36. Poliquin, R., Rockafellar, R.: Prox-regular functions in variational analysis. *Transactions of the American Mathematical Society* **348**(5), 1805–1838 (1996)
37. Rockafellar, R.T., Wets, R.J.B.: *Variational analysis*, vol. 317. Springer Science & Business Media (2009)
38. Sagastizábal, C.: A VU-point of view of nonsmooth optimization. *Proceedings of the International Congress of Mathematicians* **4**, 3815–3836 (2018)
39. Vaïter, S., Peyré, G., Fadili, J.: Low complexity regularization of linear inverse problems. In: *Sampling Theory, a Renaissance*, pp. 103–153. Springer (2015)
40. Wright, S.J.: Identifiable surfaces in constrained optimization. *SIAM Journal on Control and Optimization* **31**(4), 1063–1079 (1993)

## A Results on the Proximal Gradient

### A.1 Preliminary lemmas

Out of completeness, we present below basic results about the proximal gradient along with their proofs. We refer to [7, Chap. 10] for a general reference on the topic.

**Lemma 3 (Functional descent)** *Let Assumption 1 hold. For any  $y \in \mathbb{R}^n, \gamma \in \mathbb{R}_+$ , and  $x \in \text{prox}_{\gamma g}(y - \gamma \nabla f(y))$ , we have*

$$F(x) + \frac{1 - \gamma L}{2\gamma} \|x - y\|^2 \leq F(y).$$

*Proof.* By definition,

$$\begin{aligned} x &\in \operatorname{argmin}_{u \in \mathbb{R}^n} \left( g(u) + \frac{1}{2\gamma} \|u - (y - \gamma \nabla f(y))\|^2 \right) \\ &= \operatorname{argmin}_{u \in \mathbb{R}^n} \underbrace{\left( f(y) + \langle \nabla f(y), u - y \rangle + g(u) + \frac{1}{2\gamma} \|u - y\|^2 \right)}_{\triangleq s_y(u)} \end{aligned}$$

and the optimality of  $x$  implies that  $s_y(x) \leq s_y(y)$ , i.e.:

$$f(y) + \langle \nabla f(y), x - y \rangle + g(x) + \frac{1}{2\gamma} \|x - y\|^2 \leq f(y) + g(y).$$

Finally, the  $L$ -Lipschitz continuity the gradient of  $f$  implies that  $f(x) \leq f(y) + \langle \nabla f(y), x - y \rangle + L/2 \|y - x\|^2$  [6, Th. 18.15]. Combined with the previous equation, this yields the result.  $\square$

The following lemma links the output of the proximal gradient operator and the subdifferential of  $F$ .

**Lemma 4 (Bound on distance to subdifferential)** *Let Assumption 1 hold. For any  $y \in \mathbb{R}^n$ ,  $\gamma \in \mathbb{R}_+$ , and  $x = \mathbf{prox}_{\gamma g}(y - \gamma \nabla f(y))$ , we have*

$$\operatorname{dist}(0, \partial F(x)) \leq \frac{L\gamma + 1}{\gamma} \|y - x\|.$$

*Proof.* The first order optimality condition defining  $x$  are

$$0 \in \partial g(x) + \frac{1}{\gamma} (x - (y - \gamma \nabla f(y))) \Leftrightarrow 0 \in \partial g(x) + \nabla f(x) + \frac{x - y}{\gamma} + \nabla f(y) - \nabla f(x),$$

which can be rewritten as  $\frac{y-x}{\gamma} + \nabla f(x) - \nabla f(y) \in \partial F(x)$ . Using the  $L$ -Lipschitz continuity of  $\nabla f$  yields the following bound:

$$\operatorname{dist}(0, \partial F(x)) \leq \left\| \frac{y-x}{\gamma} + \nabla f(x) - \nabla f(y) \right\| \leq \left( \frac{1}{\gamma} + L \right) \|y - x\|.$$

$\square$

The following result is a less standard, it can be found in [20, Th. 4].

**Lemma 5** *Consider a function  $g : \mathbb{R}^n \rightarrow \mathbb{R}$ , a pair of points  $\bar{x}, \bar{y}$  and a steplength  $\bar{\gamma} > 0$  such that  $\bar{x} = \mathbf{prox}_{\bar{\gamma} g}(\bar{y})$  and  $g$  is  $r$  prox-regular at  $\bar{x}$  for subgradient  $\bar{v} \triangleq (\bar{y} - \bar{x})/\bar{\gamma}$ .*

*Then, for any  $\gamma \in (0, \min(1/r, \bar{\gamma}))$ , there exists a neighborhood  $\mathcal{N}_{\bar{y}}$  of  $\bar{y}$  over which  $\mathbf{prox}_{\gamma g}$  is single-valued and  $(1 - \gamma r)^{-1}$ -Lipschitz continuous. Furthermore, there holds*

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

*for  $y \in \mathcal{N}_{\bar{y}}$  and  $x$  near  $\bar{x}$  in the sense  $\|x - \bar{x}\| < \varepsilon$ ,  $|g(x) - g(\bar{x})| < \varepsilon$  and  $\|(y - x)/\gamma - \bar{v}\| < \varepsilon$  for some  $\varepsilon > 0$ .*

*Proof.* One can easily check that prox-regularity of  $g$  at  $\bar{x}$  for subgradient  $\bar{v}$  is equivalent to prox-regularity of function  $\tilde{g}$  around 0 for subgradient 0, with  $\tilde{g} = g(\cdot + \bar{x}) - \langle \bar{v}, \cdot \rangle - g(\bar{x})$  and a change of variable  $\tilde{x} = x - \bar{x}$ . Similarly,  $\bar{x} = \mathbf{prox}_{\bar{\gamma} g}(\bar{y})$  is characterized by its global optimality condition

$$g(x) + \frac{1}{2\bar{\gamma}} \|x - \bar{y}\| > g(\bar{x}) + \frac{1}{2\bar{\gamma}} \|\bar{x} - \bar{y}\|^2 \quad \text{for all } x \neq \bar{x},$$

which we may write as

$$g(x) > g(\bar{x}) + \langle \bar{v}, x - \bar{x} \rangle - \frac{1}{2\bar{\gamma}} \|x - \bar{x}\|^2 \quad \text{for all } x \neq \bar{x}.$$

Under that same change of variables, since  $\tilde{g}(0) = 0$ , this optimality condition rewrites as

$$\tilde{g}(\tilde{x}) > -\frac{1}{2\bar{\gamma}}\|\tilde{x}\|^2 \quad \text{for all } \tilde{x} \neq 0.$$

We may thus apply Theorem 4.4 from [36] to get the claimed result on  $\tilde{g}$ , which transfer back to  $g$  as our change of function and variable is bijective. We thus obtain that for  $\gamma \in (0, \min(1/r, \bar{\gamma}))$ , on a neighborhood  $\mathcal{N}_{\bar{y}}$  of  $\bar{y}$ ,  $\mathbf{prox}_{\gamma g}$  is single-valued,  $(1 - \gamma r)^{-1}$ -Lipschitz continuous and  $\mathbf{prox}_{\gamma g}(y) = [I + \gamma T]^{-1}(y)$ , where  $T$  denotes the  $g$ -attentive  $\varepsilon$ -localization of  $\partial g(\bar{x})$ . Taking  $y$  near  $\bar{y}$  and  $x$  near  $\bar{x}$  such that  $\|x - \bar{x}\| < \varepsilon$ ,  $|g(x) - g(\bar{x})| < \varepsilon$  and  $\|(y - x)/\gamma - \bar{v}\| < \varepsilon$  allows to identify the localization of  $\partial g(x)$  with  $\partial g(\bar{x})$ , so that

$$\frac{y - x}{\gamma} \in \partial g(x) \Leftrightarrow \frac{y - x}{\gamma} \in T(x) \Leftrightarrow (I + \gamma T)(x) = y \Leftrightarrow x = \mathbf{prox}_{\gamma g}(y).$$

Note that the proof of [36, Th. 4.4] includes a minor error relative to the Lipschitz constant computation, we report here a corrected value.  $\square$

Note here that the condition  $\gamma \in (0, \min(1/r, \bar{\gamma}))$  may be misleading since for the second part of the result, the conditions i)  $\|(y - x)/\gamma - \bar{v}\| < \varepsilon$  and ii)  $y \in \mathcal{N}_{\bar{y}}$  also have to be fulfilled. This means that the quantity

$$\|\bar{y} - \bar{x}\| \left( \frac{1}{\gamma} - \frac{1}{\bar{\gamma}} \right)$$

also has to be small. This can be done either a) by taking  $\gamma$  sufficiently close to  $\bar{\gamma}$  (which may not be possible since  $\gamma$  has to be smaller than  $1/r$ ); or b) when  $\|\bar{y} - \bar{x}\|$  are sufficiently small, ie. around fixed points of the proximal operator.

## A.2 Strong local minimizers

In this section, we show that critical points of prox-regular functions are strong minimizers. This property and the associated optimality conditions on manifolds are important for proving the localization properties of proximal gradient steps. These results can be considered as existing results, and they appear more or less explicitly in some articles, including [15].

**Lemma 6** *Let  $f$  and  $g$  denote two functions and  $\bar{x}, \bar{y}$  two points such that  $f$  is differentiable at  $\bar{y}$  and  $g$  is  $r$ -prox-regular at  $\bar{x}$  for subgradient  $\frac{1}{\gamma}(\bar{y} - \bar{x}) - \nabla f(\bar{y}) \in \partial g(\bar{x})$  with  $\gamma \in (0, 1/r)$ .*

*Then,  $\bar{x}$  is a strong local minimizer of  $\rho_{\bar{y}} : x \mapsto g(x) + \frac{1}{2\gamma}\|\bar{y} - \gamma\nabla f(\bar{y}) - x\|^2$ , ie. for all  $x$  near  $\bar{x}$ ,*

$$\rho_{\bar{y}}(x) \geq \rho_{\bar{y}}(\bar{x}) + \frac{1}{2} \left( \frac{1}{\gamma} - r \right) \|x - \bar{x}\|^2.$$

*Proof.* Prox-regularity of  $g$  at  $\bar{x}$  with subgradient  $\frac{1}{\gamma}(\bar{y} - \gamma\nabla f(\bar{y}) - \bar{x}) \in \partial g(\bar{x})$  writes

$$g(x) \geq g(\bar{x}) + \frac{1}{\gamma} \langle \bar{y} - \gamma\nabla f(\bar{y}) - \bar{x}, x - \bar{x} \rangle - \frac{r}{2} \|x - \bar{x}\|^2.$$

The identity  $2\langle b - a, c - a \rangle = \|b - a\|^2 + \|c - a\|^2 - \|b - c\|^2$  applied to the previous scalar product yields:

$$g(x) \geq g(\bar{x}) + \frac{1}{2\gamma} \|\bar{y} - \gamma\nabla f(\bar{y}) - \bar{x}\|^2 + \frac{1}{2\gamma} \|x - \bar{x}\|^2 - \frac{1}{2\gamma} \|\bar{y} - \gamma\nabla f(\bar{y}) - x\|^2 - \frac{r}{2} \|x - \bar{x}\|^2,$$

which rewrites

$$\underbrace{g(x) + \frac{1}{2\gamma} \|\bar{y} - \gamma\nabla f(\bar{y}) - x\|^2}_{=\rho_{\bar{y}}(x)} \geq \underbrace{g(\bar{x}) + \frac{1}{2\gamma} \|\bar{y} - \gamma\nabla f(\bar{y}) - \bar{x}\|^2}_{=\rho_{\bar{y}}(\bar{x})} + \frac{1}{2} \left( \frac{1}{\gamma} - r \right) \|x - \bar{x}\|^2.$$

$\square$

**Lemma 7 (Sufficient optimality conditions)** *Consider a manifold  $\mathcal{M}$  embedded in  $\mathbb{R}^n$ , a point  $\bar{x}$  of  $\mathcal{M}$  and a function  $F$  defined on a neighborhood of  $\bar{x}$  in  $\mathcal{M}$  for which  $\bar{x}$  is a strong minimizer ie. for all  $x$  near  $\bar{x}$  in  $\mathcal{M}$ ,*

$$F(x) \geq F(\bar{x}) + \frac{c}{2} \|x - \bar{x}\|^2.$$

Then, there holds

$$\text{grad } F(\bar{x}) = 0 \quad \text{and} \quad \text{Hess } F(\bar{x}) \succeq cI.$$

*Proof.* Let  $\eta \in T_{\bar{x}}\mathcal{M}$  and denote  $\gamma$  a geodesic going through  $\bar{x}$  with velocity  $\eta$  at  $t = 0$ . The quadratic growth assumption can be applied at  $x = \gamma(t)$ , which allows to write

$$\frac{1}{t}(F \circ \gamma(t) - F \circ \gamma(0)) \geq \frac{c}{2} \left\| \frac{\gamma(t) - \gamma(0)}{\sqrt{t}} \right\|^2.$$

Taking the limit  $t \rightarrow 0$  yields  $\langle \text{grad } F(\bar{x}), \eta \rangle \geq 0$ . The same reasoning holds with  $x = \gamma(-t)$  and yields the converse inequality, so that  $\langle \text{grad } F(\bar{x}), \eta \rangle = 0$ .

Besides, summing the quadratic growth conditions applied at  $\gamma(t)$  and  $\gamma(-t)$  provides

$$\frac{1}{t^2}(F \circ \gamma(t) - 2F \circ \gamma(0) + F \circ \gamma(-t)) \geq \frac{c}{2} \left\| \frac{\gamma(t) - \gamma(0)}{t} \right\|^2 + \frac{c}{2} \left\| \frac{\gamma(-t) - \gamma(0)}{t} \right\|^2.$$

Taking the limit as  $t \rightarrow 0$  yields  $\langle \text{Hess } F(\bar{x})[\eta], \eta \rangle \geq c\|\eta\|^2$ . As  $\eta$  is picked arbitrarily in  $T_{\bar{x}}\mathcal{M}$ , the results are obtained.  $\square$

## B Technical results on Riemannian methods.

In this section, we provide basic results on Riemannian optimization that simplify our developments and that we have not been able to find in the existing literature.

### B.1 Euclidean spaces and manifolds, back and forth

We establish here a connection between the Riemannian and the Euclidian distances.

**Lemma 8** *Consider a point  $\bar{x}$  of a Riemannian manifold  $\mathcal{M}$ , equipped with a retraction  $R$  such that  $R_{\bar{x}}$  is  $\mathcal{C}^2$ . For any  $\varepsilon > 0$ , there exists a neighborhood  $\mathcal{U}$  of  $\bar{x}$  in  $\mathcal{M}$  such that, for all  $x \in \mathcal{U}$ ,*

$$(1 - \varepsilon)\text{dist}_{\mathcal{M}}(x, \bar{x}) \leq \|R_{\bar{x}}^{-1}(x)\| \leq (1 + \varepsilon)\text{dist}_{\mathcal{M}}(x, \bar{x}).$$

where  $R_{\bar{x}}^{-1} : \mathcal{M} \rightarrow T_{\bar{x}}\mathcal{M}$  is the smooth inverse of  $R_{\bar{x}}$  defined locally around  $\bar{x}$ .

*Proof.* The retraction at  $\bar{x}$  can be inverted locally around 0. Indeed, as  $D R_{\bar{x}}(0_{T_{\bar{x}}\mathcal{M}}) = I$  is invertible and  $R_{\bar{x}}$  is  $\mathcal{C}^2$ , the implicit function theorem provides the existence of a  $\mathcal{C}^2$  inverse function  $R_{\bar{x}}^{-1} : \mathcal{M} \rightarrow T_{\bar{x}}\mathcal{M}$  defined locally around  $\bar{x}$ . Furthermore, one shows by differentiating the relation  $R_{\bar{x}} \circ R_{\bar{x}}^{-1}$  that the differential of  $R_{\bar{x}}^{-1}$  at  $\bar{x}$  is the identity.

We consider the function  $f : \mathcal{M} \rightarrow \mathbb{R}$  defined by  $f(x) = \|\log_{\bar{x}}(x)\| - \|R_{\bar{x}}^{-1}(x)\|$ . Clearly  $f(\bar{x}) = 0$ , and  $D f(\bar{x}) = 0$  as the differentials of both  $R_{\bar{x}}^{-1}$  and logarithm at  $\bar{x}$  are the identity. In local coordinates  $\hat{x} = \log_{\bar{x}} x$  around  $\bar{x}$ ,  $f$  is represented by the function  $\hat{f} = f \circ \exp_{\bar{x}} : T_{\bar{x}}\mathcal{M} \rightarrow \mathbb{R}$ . As  $\hat{f}(\hat{x}) = 0$ ,  $D \hat{f}(\hat{x}) = 0$  and  $\hat{f}$  is  $\mathcal{C}^2$ , there exists some  $C > 0$  such that, over a neighborhood  $\hat{\mathcal{U}}$  of  $\hat{x}$ ,

$$-C\|\hat{x} - \hat{x}\|^2 \leq \hat{f}(\hat{x}) \leq C\|\hat{x} - \hat{x}\|^2.$$

For any  $\varepsilon > 0$ , by taking a small enough neighborhood  $\hat{\mathcal{U}}' \subset \hat{\mathcal{U}}$ , there holds

$$-\varepsilon\|\hat{x} - \hat{x}\| \leq \hat{f}(\hat{x}) \leq \varepsilon\|\hat{x} - \hat{x}\|.$$

Thus for all  $x$  in  $\mathcal{U} = R_{\bar{x}}(\hat{\mathcal{U}}')$ ,

$$-\varepsilon\|\log_{\bar{x}}(x)\| \leq \|\log_{\bar{x}}(x)\| - \|R_{\bar{x}}^{-1}(x)\| \leq \varepsilon\|\log_{\bar{x}}(x)\|,$$

as  $\hat{x} = \log_{\bar{x}}(x)$ ,  $\hat{x} = 0$ . The result's equivalent form is obtained using that  $\text{dist}_{\mathcal{M}}(x, \bar{x}) = \|\hat{x} - \hat{x}\| = \|\log_{\bar{x}}(x)\|$ .  $\square$

We recall a slightly specialized version of [32, Th. 2.2], which is essentially the application of the implicit function theorem around a point of a manifold.

**Proposition 2 (Tangential retraction)** *Consider a  $p$ -dimensional  $C^k$ -submanifold  $\mathcal{M}$  of  $\mathbb{R}^n$  around a point  $\bar{x} \in \mathcal{M}$ . The mapping  $R : T\mathcal{B} \rightarrow \mathcal{M}$ , defined for  $(x, \eta) \in T\mathcal{B}$  near  $(\bar{x}, 0)$  by  $\text{proj}_{\bar{x}}(R(x, \eta)) = \eta$  defines a second-order retraction near  $(\bar{x}, 0)$ . The point-wise retraction, defined as  $R_x = R(x, \cdot)$ , is locally invertible with inverse  $R_x^{-1} = \text{proj}_x$ .*

*Proof.* Let  $\Psi : \mathbb{R}^n \rightarrow \mathbb{R}^{n-p}$  denote a  $C^k$  function defining  $\mathcal{M}$  around  $\bar{x}$ : for all  $x$  close enough to  $\bar{x}$ , there holds  $x \in \mathcal{M} \Leftrightarrow \Psi(x) = 0$ , and  $D\Psi(x)$  is surjective. Consider the equation  $\Phi(x, \eta_t, \eta_n) = 0$  around  $(\bar{x}, 0, 0)$ , with

$$\begin{aligned} \Phi : \{x, \eta_t, \eta_n : x \in \mathcal{M}, \eta_t \in T_x\mathcal{M}, \eta_n \in N_x\mathcal{M}\} &\rightarrow \mathbb{R} \\ x, \eta_t, \eta_n &\mapsto \Psi(x + \eta_t + \eta_n). \end{aligned}$$

The partial differential  $D_{\eta_n} \Phi(\bar{x}, 0, 0)$  is, for  $\xi_n \in N_{\bar{x}}\mathcal{M}$ ,

$$D_{\eta_n} \Phi(\bar{x}, 0, 0)[\xi_n] = D\Psi(\bar{x})[\xi_n].$$

Since  $\bar{x} \in \mathcal{M}$ ,  $D_{\eta_n} \Phi(\bar{x}, 0, 0)$  is surjective from  $N_{\bar{x}}\mathcal{M}$  to  $\mathbb{R}^{n-p}$  so its a bijection. The implicit function theorem provides the existence of neighborhoods  $\mathcal{N}_{\bar{x}}^1 \subset \mathcal{M}$ ,  $\mathcal{N}_0^2 \subset \cup_{x \in \mathcal{M}} T_x\mathcal{M}$ ,  $\mathcal{N}_0^3 \subset \cup_{x \in \mathcal{M}} N_x\mathcal{M}$  and a unique  $C^k$  function  $\eta_n : \mathcal{N}_{\bar{x}}^1 \times \mathcal{N}_0^2 \rightarrow \mathcal{N}_0^3$  such that, for all  $x \in \mathcal{N}_{\bar{x}}^1$ ,  $\eta_t \in \mathcal{N}_0^2$  and  $\eta_n \in \mathcal{N}_0^3$ ,  $\eta_n(\bar{x}, 0) = 0$  and

$$\Phi(x, \eta_t, \eta_n(x, \eta_t)) = 0 \Leftrightarrow x + \eta_t + \eta_n(x, \eta_t) \in \mathcal{M}.$$

It also provides an expression for the partial derivative of  $\eta_n$  at  $(x, 0)$  along  $\eta_t$ : for  $\xi_t \in T_x\mathcal{M}$ ,

$$D_{\eta_t} \eta_n(x, 0)[\xi_t] = -[D_{\eta_n} \Phi(x, 0, 0)]^{-1} D_{\eta_t} \Phi(x, 0, 0)[\xi_t].$$

As noted before,  $D_{\eta_n} \Phi(x, 0, 0)$  is bijective since  $x \in \mathcal{M}$ . Besides,  $D_{\eta_t} \Phi(x, 0, 0) = D\Phi(x)[\xi_t] = 0$  since  $T_x\mathcal{M}$  identifies as the kernel of  $D\Phi(x)$ . Thus  $D_{\eta_t} \eta_n(x, 0) = 0$ .

Now, define a map  $R : \mathcal{N}_{\bar{x}}^1 \times \mathcal{N}_0^2 \rightarrow \mathcal{M}$  by  $R(x, \eta_t) = x + \eta_t + \eta_n(x, \eta_t)$ . This map has degree of smoothness  $C^k$  since  $\eta_n$  is  $C^k$ , satisfies  $R(x, 0) = x$  since  $\eta_n(x, 0) = 0$  and satisfies  $D_{\eta_t} \eta_n(x, 0) = I + D_{\eta_t} \eta_n(x, 0) = I$ . Thus  $R$  defines a retraction on a neighborhood of  $(\bar{x}, 0)$ .

We turn to show the second-order property of  $R$ . Consider the smooth curve  $c$  defined as  $c(t) = R(x, t\eta)$  for some  $x \in \mathcal{N}_{\bar{x}}^1$ ,  $\eta_t \in T_x\mathcal{M} \cap \mathcal{N}_0^2$ . It's first derivative writes

$$c'(t) = \eta + D_{\eta_t} \eta_n(x, t\eta)[\eta] = \eta.$$

As  $c'(t) \in T_x\mathcal{M}$ , and we consider a Riemannian submanifold of  $\mathbb{R}^n$ , the acceleration of the curve  $c$  is obtained by computing the derivative of  $c'(\cdot)$  in the ambient space and then projecting onto  $T_x\mathcal{M}$ . Thus  $c''(t) = 0$  and in particular,  $c''(0) = 0$  which makes  $R$  a second-order retraction.  $\square$

**Lemma 9** *Consider a point  $\bar{x}$  of a Riemannian manifold  $\mathcal{M}$ . For any  $\varepsilon > 0$ , there exists a neighborhood  $\mathcal{U}$  of  $\bar{x}$  in  $\mathcal{M}$  such that, for all  $x \in \mathcal{U}$ ,*

$$(1 - \varepsilon)\text{dist}_{\mathcal{M}}(x, \bar{x}) \leq \|x - \bar{x}\| \leq (1 + \varepsilon)\text{dist}_{\mathcal{M}}(x, \bar{x}),$$

where  $\|x - \bar{x}\|$  is the Euclidean distance in the ambient space.

*Proof.* Let  $\bar{x}, x$  denote two close points on  $\mathcal{M}$ . Consider the tangential retraction introduced in Proposition 2. As a retraction, it satisfies:

$$R_{\bar{x}}(\eta) = R_{\bar{x}}(0) + D R_{\bar{x}}(0)[\eta] + \mathcal{O}(\|\eta\|^2) = \bar{x} + \mathcal{O}(\|\eta\|^2).$$

Taking  $x = R_{\bar{x}}(\eta)$  allows to write  $x = \bar{x} + \mathcal{O}(\|R_{\bar{x}}^{-1}(x)\|^2)$ , so that for any small  $\varepsilon_1 > 0$ , there exists a small enough neighborhood  $\mathcal{U}_1 \subset \mathcal{U}$  of  $\bar{x}$  in  $\mathcal{M}$  such that

$$(1 - \varepsilon_1)\|R_{\bar{x}}^{-1}(x)\| \leq \|x - \bar{x}\| \leq (1 + \varepsilon_1)\|R_{\bar{x}}^{-1}(x)\|.$$

By Lemma 8, for  $\varepsilon_2 > 0$  small enough, there exists a neighborhood  $\mathcal{U}_2 \subset \mathcal{U}$  of  $\bar{x}$  such that,

$$(1 - \varepsilon_2)\text{dist}_{\mathcal{M}}(x, \bar{x}) \leq \|R_{\bar{x}}^{-1}(x)\| \leq (1 + \varepsilon_2)\text{dist}_{\mathcal{M}}(x, \bar{x}).$$

By choosing  $\varepsilon_1, \varepsilon_2$  such that  $1 - \varepsilon = (1 - \varepsilon_1)(1 - \varepsilon_2)$ , these two estimates can be combined to get the result.  $\square$

## B.2 Soundness of the Riemannian line-search

Theorem B.1 and its proof are adapted from [11, Th. 4.16] to the Riemannian setting.

**Theorem B.1** *Consider a manifold  $\mathcal{M}$  equipped with a retraction  $R$  and a twice differentiable function  $F : \mathcal{M} \rightarrow \mathbb{R}$  that admits a strong local minimizer  $x^*$ , that is a point such that  $\text{Hess } F(x^*)$  is positive definite. If  $x$  is close to  $x^*$ ,  $\eta$  brings a superlinear improvement towards  $x^*$ , that is  $\text{dist}_{\mathcal{M}}(R_x(\eta), x^*) = o(\text{dist}_{\mathcal{M}}(x, x^*))$  as  $x \rightarrow x^*$ , and  $0 < m_1 < 1/2$ , then  $\eta$  is acceptable by the Armijo rule (4.1) with unit stepsize  $\alpha = 1$ .*

*Proof.* Let  $x, \eta \in T\mathcal{B}$  denote a pair such that  $x$  is close to  $x^*$  and  $\text{dist}_{\mathcal{M}}(R_x(\eta), x^*) = o(\text{dist}_{\mathcal{M}}(x, x^*))$ . For convenience, let  $x_+ = R_x(\eta)$  denote the next point.

Following [2] (see e.g. the proof of Th. 6.3.2), we work in local coordinates around  $x^*$ , representing any point  $x \in \mathcal{M}$  by  $\widehat{x} = \log_{x^*}(x)$  and any tangent vector  $\eta \in T_x\mathcal{M}$  by  $\widehat{\eta}_x = D\log_{x^*}(x)[\eta]$ . The function  $F$  is represented by  $\widehat{F} = F \circ \exp_{x^*} : T_{x^*}\mathcal{M} \rightarrow \mathbb{R}$ . Defining the coordinates via the logarithm grants the useful property that the riemannian distance of any two points  $x, y \in \mathcal{M}$  matches the euclidean distance between their representatives:  $\text{dist}_{\mathcal{M}}(x, y) = \|\widehat{x} - \widehat{y}\|$ . Besides, there holds

$$D F(x)[\eta] = D \widehat{F}(\widehat{x})[\widehat{\eta}] \quad \text{and} \quad \text{Hess } F(x)[\eta, \eta] = D^2 \widehat{F}(\widehat{x})[\widehat{\eta}, \widehat{\eta}]. \quad (\text{B.1})$$

Indeed,  $D F(x)[\eta] = (F \circ \gamma)'(0)$  and  $\text{Hess } F(x)[\eta, \eta] = (F \circ \gamma)''(0)$ , where  $\gamma$  denotes the geodesic curve defined by  $\widehat{\gamma}(t) = \widehat{x} + t\widehat{\eta}$ . Using  $F \circ \gamma = \widehat{F} \circ \widehat{\gamma}$ , one obtains the result.

**Step 1.** We derive an approximation of  $D F(x)[\eta] = \langle \text{grad } F(x), \eta \rangle$  in terms of  $D^2 \widehat{F}(\widehat{x}^*)[\widehat{x} - \widehat{x}^*]^2$ . To do so, we go through the intermediate quantity  $D \widehat{F}(\widehat{x})[\widehat{x}_+ - \widehat{x}]$ , and handle precisely the  $o(\cdot)$  terms. By smoothness of  $\widehat{F}$  and since  $D \widehat{F}(\widehat{x}^*) = 0$ , Taylor's formula for  $D \widehat{F}$  writes

$$\begin{aligned} D \widehat{F}(\widehat{x})[\widehat{x}_+ - \widehat{x}] &= D^2 \widehat{F}(\widehat{x}^*)[\widehat{x}_+ - \widehat{x}, \widehat{x} - \widehat{x}^*] + o(\|\widehat{x} - \widehat{x}^*\|^2) \\ &= -D^2 \widehat{F}(\widehat{x}^*)[\widehat{x} - \widehat{x}^*]^2 + D^2 \widehat{F}(\widehat{x}^*)[\widehat{x}_+ - \widehat{x}^*, \widehat{x} - \widehat{x}^*] + o(\|\widehat{x} - \widehat{x}^*\|^2) \\ &= -D^2 \widehat{F}(\widehat{x}^*)[\widehat{x} - \widehat{x}^*]^2 + o(\|\widehat{x} - \widehat{x}^*\|^2), \end{aligned}$$

where, in the last step, we used that  $\|\widehat{x}_+ - \widehat{x}^*\| = o(\|\widehat{x} - \widehat{x}^*\|)$  to get that  $\|D^2 \widehat{F}(\widehat{x}^*)[\widehat{x}_+ - \widehat{x}^*, \widehat{x} - \widehat{x}^*]\| = \|D^2 \widehat{F}(\widehat{x}^*)\| \|\widehat{x}_+ - \widehat{x}^*\| \|\widehat{x} - \widehat{x}^*\| = o(\|\widehat{x} - \widehat{x}^*\|^2)$ . We now turn to show that  $D \widehat{F}(\widehat{x})[\widehat{x}_+ - \widehat{x}]$  behaves as  $D F(x)[\eta]$  up to  $o(\|\widehat{x}_+ - \widehat{x}\|^2)$ . Since  $D F(x)[\eta] = D \widehat{F}(\widehat{x})[\widehat{\eta}]$  by (B.1), there holds:

$$\|D F(x)[\eta] - D \widehat{F}(\widehat{x})[\widehat{x}_+ - \widehat{x}]\| = \|D \widehat{F}(\widehat{x})[\widehat{\eta} - (\widehat{x}_+ - \widehat{x})]\| \leq \|D \widehat{F}(\widehat{x})\| \|\widehat{\eta} - (\widehat{x}_+ - \widehat{x})\|.$$

As  $F$  is twice differentiable and  $\exp$  is  $C^\infty$ ,  $\widehat{F}$  is twice differentiable as well. In particular its derivative is locally Lipschitz continuous, so that for  $\widehat{x}$  near  $\widehat{x}^*$ , we obtain a first estimate:

$$\|D \widehat{F}(\widehat{x})\| = \|D \widehat{F}(\widehat{x}) - D \widehat{F}(\widehat{x}^*)\| = \mathcal{O}(\|\widehat{x} - \widehat{x}^*\|).$$

Besides, the following estimate holds  $\|\widehat{\eta} - (\widehat{x}_+ - \widehat{x})\| = o(\|\widehat{x} - \widehat{x}^*\|)$ . Indeed, as the function  $\log_{x^*} \circ R_x : T_x\mathcal{M} \rightarrow T_{x^*}\mathcal{M}$  is differentiable, there holds for  $\eta \in T_x\mathcal{M}$  small,

$$\log_{x^*}(R_x(\eta)) = \log_{x^*}(R_x(0)) + D \log_{x^*}(R_x(0))[D R_x(0)[\eta]] + o(\|\eta\|),$$

which simplifies to  $\widehat{x}_+ = \widehat{x} + \widehat{\eta} + o(\|\eta\|)$ . Lemma 8 allows to write  $\|\eta\| = \|R_x^{-1}(x_+)\| = \mathcal{O}(\text{dist}_{\mathcal{M}}(x, x_+))$ . Using the triangular inequality and the assumption that  $\text{dist}_{\mathcal{M}}(x_+, x^*) = o(\text{dist}_{\mathcal{M}}(x, x^*))$  we get

$$\text{dist}_{\mathcal{M}}(x, x_+) \leq \text{dist}_{\mathcal{M}}(x, x^*) + \text{dist}_{\mathcal{M}}(x^*, x_+) = \mathcal{O}(\text{dist}_{\mathcal{M}}(x, x^*)) = \mathcal{O}(\|\widehat{x} - \widehat{x}^*\|),$$

so that the second estimate holds.

Combining the two above estimates allows to conclude that

$$\|D F(x)[\eta] - D \widehat{F}(\widehat{x})[\widehat{x}_+ - \widehat{x}]\| = o(\|\widehat{x} - \widehat{x}^*\|^2),$$

so that overall,

$$D F(x)[\eta] = D \widehat{F}(\widehat{x})[\widehat{x}_+ - \widehat{x}] + o(\|\widehat{x} - \widehat{x}^*\|^2) = -D^2 \widehat{F}(\widehat{x}^*)[\widehat{x} - \widehat{x}^*]^2 + o(\|\widehat{x} - \widehat{x}^*\|^2).$$

Using that  $\|\widehat{x} - \widehat{x}^*\| = \text{dist}_{\mathcal{M}}(x, x^*)$  and  $D^2 \widehat{F}(\widehat{x}^*) = \text{Hess } F(x^*)$  (B.1), we obtain

$$D F(x)[\eta] = -\text{Hess } F(x^*)[\widehat{x} - \widehat{x}^*]^2 + o(\text{dist}_{\mathcal{M}}(x, x^*)^2). \quad (\text{B.2})$$

**Step 2.** The function  $F$  admits a second-order development around  $x^*$ : applying Eq. (2.4) with the exponential map  $\exp_{x^*}$  as a second-order retraction yields

$$F(x) = F(x^*) + \frac{1}{2} \text{Hess } F(x^*)[\widehat{x} - \widehat{x}^*]^2 + o(\text{dist}_{\mathcal{M}}(x, x^*)^2), \quad (\text{B.3})$$

where we used that  $\text{dist}_{\mathcal{M}}(x, x^*) = \|\log_{x^*}(x) - \log_{x^*}(x^*)\|$ . Denote  $0 < l \leq L$  the lower and upper eigenvalues of  $\text{Hess } F(x^*)$ . The combination (B.3) +  $m_1$ (B.2) writes

$$\begin{aligned} F(x) + m_1 D F(x)[\eta] &= F(x^*) + \left(\frac{1}{2} - m_1\right) \text{Hess } F(x^*)[\widehat{x} - \widehat{x}^*]^2 + o(\text{dist}_{\mathcal{M}}(x, x^*)^2) \\ &\geq F(x^*) + \left(\frac{1}{2} - m_1\right) l \text{dist}_{\mathcal{M}}(x, x^*)^2 + o(\text{dist}_{\mathcal{M}}(x, x^*)^2), \end{aligned}$$

Let  $\varepsilon > 0$  such that  $\frac{1}{2}L\varepsilon^2 < (\frac{1}{2} - m_1)l$ . As  $\text{dist}_{\mathcal{M}}(x_+, x^*) = o(\text{dist}_{\mathcal{M}}(x, x^*))$ , for  $x$  close enough to  $x^*$  there holds  $\text{dist}_{\mathcal{M}}(x_+, x^*) \leq \varepsilon \text{dist}_{\mathcal{M}}(x, x^*)$ . Combining this with the second-order development of  $f$  at  $x_+$ , there holds:

$$\begin{aligned} F(x_+) &= F(x^*) + \frac{1}{2} \text{Hess } F(x^*)[\widehat{x}_+ - \widehat{x}^*]^2 + o(\text{dist}_{\mathcal{M}}(x_+, x^*)^2) \\ &\leq F(x^*) + \frac{1}{2} L \text{dist}_{\mathcal{M}}(x_+, x^*)^2 + o(\text{dist}_{\mathcal{M}}(x_+, x^*)^2) \\ &\leq F(x^*) + \frac{1}{2} L \varepsilon^2 \text{dist}_{\mathcal{M}}(x, x^*)^2 + o(\text{dist}_{\mathcal{M}}(x, x^*)^2). \end{aligned}$$

Subtracting the two estimates yields

$$F(x_+) - (F(x) + m_1 D F(x)[\eta]) \leq \left(\frac{1}{2}L\varepsilon^2 - \left(\frac{1}{2} - m_1\right)l\right) \text{dist}_{\mathcal{M}}(x, x^*)^2 + o(\text{dist}_{\mathcal{M}}(x, x^*)^2),$$

which ensures that the Armijo condition is satisfied.  $\square$

### B.3 Riemannian Newton method

The Riemannian Newton method is shown to converge quadratically by Absil et al [2, Th. 6.3.2], which we recall in our setting.

**Lemma 10** *Let Assumption 1 hold and consider a manifold  $\mathcal{M}$  and point  $x^* \in \mathcal{M}$  such that*

- $F$  is twice differentiable on  $\mathcal{M}$  near  $x^*$ ;
- $\text{Hess}_{\mathcal{M}} F(x^*) \succ 0$ ;
- $\text{Hess}_{\mathcal{M}} F$  is locally Lipschitz near  $x^*$ .

*Then there exists a neighborhood  $\mathcal{U} \subset \mathcal{M}$  of  $x^*$  such that the Riemannian Newton algorithm, defined as*

$$x_{k+1} = \text{R}_{x_k}(-\text{Hess } F(x_k)^{-1} \text{grad } F(x_k)),$$

*converges quadratically to  $x^*$  for any initial point  $x_0 \in \mathcal{U}$ .*

*Proof.* The newton update, analysed in [2, Th. 6.3.2], provides the existence of a neighborhood  $\mathcal{U}$  over which the following estimate holds:

$$\text{dist}_{\mathcal{M}}(\text{ManUp}(x), x^*) \leq (\beta\gamma_J + \gamma_R) \text{dist}_{\mathcal{M}}(x, x^*)^2,$$

where  $\gamma_J$  and  $\gamma_R$  quantify smoothness of  $\text{Hess}$  and  $\text{R}_x$  as defined in the proof. We specialize the result to normal coordinates, that is  $\varphi = \exp_{x^*}$ , which allows to write estimates in terms of Riemannian distances and make the Christoffel terms becomes negligible.  $\square$



## B.4 Riemannian Newton-CG method

In order to prove the global and local behavior of the Newton-CG manifold update, we first show that  $d_k$  is indeed a descent direction for  $F \circ R_{x_k}$ , adapting the proof of [17, Lemma A.2] to the Riemannian setting.

**Lemma 11** *Let Assumption 1 hold and consider a manifold  $\mathcal{M}$  and a point  $x \in \mathcal{M}$ . If  $F$  is twice differentiable on  $\mathcal{M}$  at  $x$  and  $x$  is not a stationary point of  $F$ , then there holds:*

$$\langle \text{grad } F(x), d \rangle \leq -\min(1, \|\text{Hess } F(x)\|^{-1}) \|\text{grad } F(x)\|^2,$$

where  $d$  was obtained solving Eq. (Inexact Newton eq.) with any forcing parameter  $\eta$ .

*Proof.* The result is obtained by applying the analysis of [17, Lemma A.2] to the approximate resolution of (Inexact Newton eq.) on the euclidean space  $T_x\mathcal{M}$ , with constant specified according to the proof.  $\square$

We also show that the Riemannian Newton-CG direction provides superlinear improvement.

**Lemma 12** *Let Assumption 1 hold and consider a manifold  $\mathcal{M}$  and point  $\bar{x} \in \mathcal{M}$  such that*

- $F$  is twice differentiable on  $\mathcal{M}$  near  $\bar{x}$ ;
- $\bar{x}$  is a critical point on  $\mathcal{M}$ :  $\text{grad } F(\bar{x}) = 0$ ;
- $\text{Hess}_{\mathcal{M}}$  is invertible and locally Lipschitz at  $\bar{x}$ ;
- the forcing sequence behaves as  $\eta_k = \mathcal{O}(\|\text{grad } F(x_k)\|^\theta)$ , for some  $\theta \in (0, 1]$ .

*Then there exists a neighborhood  $\mathcal{U} \subset \mathcal{M}$  of  $\bar{x}$  such that the Riemannian Truncated Newton algorithm, defined as*

$$x_{k+1} = R_{x_k}(d_k) \quad \text{where } d_k \text{ solves } \quad \|\text{grad } F(x_k) + \text{Hess } F(x_k)[d_k]\| \leq \eta_k \|\text{grad } F(x_k)\|,$$

*converges superlinearly to  $\bar{x}$  with order  $1 + \theta$  for any initial point  $x_0 \in \mathcal{U}$ .*

*Proof.* The inexact newton update, analysed in [2, Th. 8.2.1], provides the existence of a neighborhood  $\mathcal{U}$  over which the following estimate holds:

$$\text{dist}_{\mathcal{M}}(\text{ManUp}(x), \bar{x}) \leq 2\beta\gamma_\xi^{1+\theta} \text{dist}_{\mathcal{M}}(x, \bar{x})^{1+\theta} + C \text{dist}_{\mathcal{M}}(x, \bar{x})^2,$$

where  $\gamma_\xi$  is a Lipschitz constant for  $\text{grad } F$  near  $\bar{x}$  and  $C, \beta$  denote two positive constants. Again, we have specialized the result to normal coordinates, that is  $\varphi = \exp_{\bar{x}}$ , which allows to write estimates in terms of Riemannian distances.  $\square$

## C Differentiating the singular-value decomposition

In this section, we establish the expressions of the derivative of the matrices involved in the singular value decomposition. These results may be seen as part of folklore, and, up to our knowledge, there are not explicitly written in the literature. We need them for the computations related to trace-norm regularized problems.

**Lemma 13** *Consider the manifold of fixed rank matrices  $\mathcal{M}_r$ , a pair  $x, \eta \in T\mathcal{B}$  and a smooth curve  $c : I \rightarrow \mathcal{M}_r$  such that  $c(0) = x$ ,  $c'(0) = \eta$ . Besides, let  $U(t), \Sigma(t), V(t)$  denote smooth curves of  $St(m, r)$ ,  $\mathbb{R}^{r \times r}$ ,  $St(n, r)$  such that  $\gamma(t) = U(t)\Sigma(t)V(t)^\top$ . The derivatives of the decomposition factors at  $t = 0$  write*

$$\begin{aligned} U' &= U \left( F \circ \left[ U^\top \eta V \Sigma + \Sigma V^\top \eta^\top U \right] \right) + (I_m - UU^\top) \eta V \Sigma^{-1} \\ V' &= V \left( F \circ \left[ \Sigma U^\top \eta V + V^\top \eta^\top U \Sigma \right] \right) + (I_n - VV^\top) \eta^\top U \Sigma^{-1} \\ \Sigma' &= I_k \circ \left[ U^\top \eta V \right], \end{aligned}$$

where  $I_k$  is the identity of  $\mathbb{R}^{k \times k}$ ,  $\circ$  denotes the Hadamard product and  $F \in \mathbb{R}^{r \times r}$  is such that  $F_{ij} = 1/(\Sigma_{jj}^2 - \Sigma_{ii}^2)$  if  $\Sigma_{jj} \neq \Sigma_{ii}$ , and  $F_{ij} = 0$  otherwise. Equivalently, when the tangent vector is represented as  $\eta = UMV^\top + U_pV_p^\top + UV_p^\top$ , the above expressions simplify to

$$\begin{aligned} U' &= U \left( F \circ [M\Sigma + \Sigma M^\top] \right) + U_p \Sigma^{-1} \\ V' &= V \left( F \circ [\Sigma M + M^\top \Sigma] \right) + V_p \Sigma^{-1} \\ \Sigma' &= I_k \circ M, \end{aligned}$$

*Proof.* We consider the curve  $\gamma$  and all components and derivatives at  $t = 0$ , therefore we don't mention evaluation time. Differentiating  $\gamma = U\Sigma V^\top$  yields

$$\eta = U' \Sigma V^\top + U \Sigma' V^\top + U \Sigma V'^\top \quad (\text{C.1})$$

As a tangent vector to the Stiefel manifold at point  $U$ ,  $U'$  can be expressed as [2, Ex. 3.5.2]

$$U' = U \Omega_U + U_\perp B_U, \quad (\text{C.2})$$

where  $\Omega_U \in \mathbb{R}^{r \times r}$  is a skew-symmetric matrix,  $B_U \in \mathbb{R}^{m-r \times m-r}$ , and  $U_\perp$  is any matrix such that  $U^\top U_\perp = 0$  and  $U_\perp^\top U_\perp = I_{m-r}$ . Similarly,  $V' = V \Omega_V + V_\perp B_V$ , where  $\Omega_V \in \mathbb{R}^{r \times r}$  is skew-symmetric,  $B_V \in \mathbb{R}^{n-r \times n-r}$ , and  $V_\perp$  is any matrix such that  $V^\top V_\perp = 0$  and  $V_\perp^\top V_\perp = I_{n-r}$ .

Computing  $U^\top \times (\text{C.1}) \times V$  yields

$$U^\top \eta V = \Omega_U \Sigma + \Sigma' + \Sigma \Omega_V^\top.$$

Looking at the diagonal elements of this equation yields the derivative of the diagonal component of  $\eta$ . This is done by taking the Hadamard product of both sides of previous equation with the identity matrix of  $\mathbb{R}^{r \times r}$ , and writes

$$\Sigma' = I_r \circ [U^\top \eta V].$$

The off-diagonal elements of this equation write

$$\bar{I}_r \circ [U^\top \eta V] = \Omega_U \Sigma + \Sigma \Omega_V^\top, \quad (\text{C.3})$$

where  $\bar{I}_r \in \mathbb{R}^{r \times r}$  has zeros on the diagonal and ones everywhere else. Adding (C.3)  $\Sigma$  and  $\Sigma(\text{C.3})^\top$  yields

$$\bar{I}_r \circ [U^\top \eta V \Sigma + \Sigma V^\top \eta^\top U] = \Omega_U \Sigma^2 - \Sigma^2 \Omega_U,$$

which decouples coefficient-wise. At coefficient  $(ij)$ , with  $i \neq j$ ,

$$[U^\top \eta V \Sigma + \Sigma V^\top \eta^\top U]_{ij} = [\Omega_U]_{ij} (\Sigma_{jj}^2 - \Sigma_{ii}^2),$$

hence  $\Omega_U = F \circ [U^\top \eta V \Sigma + \Sigma V^\top \eta^\top U]$ , where  $F \in \mathbb{R}^{m-r \times m-r}$  has zeros on the diagonal and for  $i \neq j$ ,  $F_{ij} = 1/(\Sigma_{jj}^2 - \Sigma_{ii}^2)$  if  $\Sigma_{jj}^2 \neq \Sigma_{ii}^2$ , 0 otherwise. Besides, left-multiplying (C.1) by  $U_\perp^\top$  yields  $U_\perp^\top \eta = U_\perp^\top U' \Sigma V^\top$ , which rewrites, using the decomposition (C.2) of  $U'$ , as  $U_\perp^\top \eta = B_U \Sigma V^\top$ . Hence  $B_U = U_\perp^\top \eta V \Sigma^{-1}$  and we get the complete expression for  $U'$  by assembling the expressions of  $\Omega_U$  and  $B_U$  with the decomposition (C.2). The term  $U_\perp^\top U_\perp$  is eliminated using that  $U^\top U + U_\perp^\top U_\perp = I_m$ .

Let's follow the same steps to get expressions for  $V'$ . Adding  $\Sigma(\text{C.3})$  and  $(\text{C.3})^\top \Sigma$  yields

$$\bar{I}_r \circ [\Sigma U^\top \eta V + V^\top \eta^\top U \Sigma] = \Omega_V \Sigma^2 - \Sigma^2 \Omega_V,$$

from which we get  $\Omega_V = F \circ [\Sigma U^\top \eta V + V^\top \eta^\top U \Sigma]$ . Besides, right-multiplying (C.1) by  $V_\perp$  yields  $\eta V_\perp = U \Sigma V'^\top V_\perp$ , which rewrites using the decomposition  $V' = V \Omega_V + V_\perp B_V$  as  $\eta V_\perp = U \Sigma B_V^\top$ . Hence  $B_V = V_\perp^\top \eta^\top U \Sigma^{-1}$ , and we get the claimed formula by eliminating the  $V_\perp$  terms with  $V^\top V + V_\perp^\top V_\perp = I_n$ .

The simplified expressions are obtained using that  $U^\top U = I_m$ ,  $U^\top U_p = 0$ ,  $V^\top V = I_n$  and  $V^\top V_p = 0$ .  $\square$

We are now ready to give the expression of the Riemannian gradient and Hessian of the nuclear norm.

**Proposition 3** *The nuclear norm  $g = \|\cdot\|_*$  restricted to  $\mathcal{M}_r$  is  $\mathcal{C}^2$  and admits a smooth second-order development of the form (2.4) near any point  $x = U\Sigma V^\top \in \mathcal{M}_r$ . Denoting  $\eta = U\eta V^\top + U_p V_p^\top + U V_p^\top \in T_x \mathcal{M}_r$  a tangent vector, there holds:*

$$\begin{aligned} \text{grad } g(x) &= UV^\top \\ \text{Hess } g(x)[\eta] &= U \left[ \tilde{F} \circ (M - M^\top) \right] V^\top + U_p \Sigma^{-1} V^\top + U \Sigma^{-1} V_p^\top, \end{aligned}$$

where  $\circ$  denotes the Hadamard product and  $\tilde{F} \in \mathbb{R}^{r \times r}$  is such that  $\tilde{F}_{ij} = 1/(\Sigma_{jj} + \Sigma_{ii})$  if  $\Sigma_{jj} \neq \Sigma_{ii}$ , and  $\tilde{F}_{ij} = 0$  otherwise.

*Proof.* Let  $c : I \rightarrow \mathcal{M}_r$  denote a smooth curve over  $\mathcal{M}_r$  such that  $\gamma(0) = x$  and  $\gamma'(0) = \eta$ , and consider  $\varphi = \|c(\cdot)\|_* : I \rightarrow \mathbb{R}$ . Writing the decomposition  $c(t) = U(t)\Sigma(t)V(t)^\top$ , for  $U(t)$ ,  $\Sigma(t)$ ,  $V(t)$  smooth curves of  $St(m, r)$ ,  $\mathbb{R}^{r \times r}$ ,  $St(n, r)$  allows to write  $\varphi(t) = \text{Tr}(\Sigma(t))$ . Applying Lemma 13 yields

$$\varphi'(0) = \text{Tr}(\Sigma'(0)) = \text{Tr}(U^\top \eta V) = \text{Tr}(\eta V U^\top) = \langle \eta, UV^\top \rangle,$$

so that  $\text{grad } g(x) = UV^\top \in T_x \mathcal{M}$ .

In order to obtain the Riemannian hessian, let  $\bar{Z} : I \rightarrow \mathbb{R}^n$  denote a smooth extension of  $\text{grad } g(c(\cdot))$ , defined by  $\bar{Z}(t) = U(t)V(t)^\top$ . The Riemannian hessian is then obtained as  $\text{Hess } g(x)[\eta] = \text{proj}_x \bar{Z}'(0)$ . The derivative of  $\bar{Z}$  at 0 is simply  $\bar{Z}'(0) = U'V^\top + UV'^\top$  and thus writes, applying Lemma 13

$$\bar{Z}'(0) = U \left( F \circ \left[ M\Sigma + \Sigma M^\top \right] \right) V^\top + U_p \Sigma^{-1} V^\top + U \left( F \circ \left[ \Sigma M + M^\top \Sigma \right] \right)^\top V^\top + U \Sigma^{-1} V_p^\top$$

This expression simplifies to the statement by using the fact that  $F$  is antisymmetric and applying the identity  $(A \circ B)^\top = A^\top \circ B^\top$ .  $\square$