Newton acceleration on manifolds identified by proximal-gradient methods

Gilles Bareilles, Franck Iutzeler, Jérôme Malick

Noname manuscript No.
(will be inserted by the editor)

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 structure at no additional cost, and convergence is improved once it matches the structure of a minimizer. 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 superlinear convergence of the algorithm when solving some nondegenerated nonsmooth nonconvex optimization problems. We provide numerical illustrations on optimization problems regularized by ℓ_1-norm or 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 [18] or in sparsity-inducing regularization techniques in data analysis [4]. In these applications, the *nonsmooth* objective functions usually present a *smooth* substructure, which involves smooth submanifolds on which the functions are locally smooth. To fix ideas, consider the simple example of the ℓ_1 norm: though nonsmooth, it is obviously smooth around any point when restricted to the vector space of points with the same support.

* This work is partly funded by the ANR JCJC project *STROLL* (ANR-19-CE23-0008).

G. Bareilles · F. Iutzeler · J. Malick
Univ. Grenoble Alpes, CNRS, Grenoble INP, LJK, 38000 Grenoble, France
E-mail: firstname.lastname@univ-grenoble-alpes.fr

Exploiting the underlying smooth substructure of objective functions to develop second-order methods has been a subject of fruitful research in nonsmooth optimization, pioneered by the developments around $\mathcal{U}$-Newton algorithms [23] and the notion of partial smoothness [25]. Let us mention the $\mathcal{UV}$-Newton bundle method of [29], and the recent k-bundle Newton method of [24]. 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 [30].

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

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

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 ℓ_1-norm: its proximity operator puts exactly to 0 some coordinates of the input vector after 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 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 to minimize F are the (accelerated) proximal gradient algorithms. 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 general settings; we refer to [12], [33], [17], or [26]. For ℓ_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 [20] for further discussions.

In the ideal case where we know that the iterates are on the optimal manifold, one could switch to a more sophisticated method, e.g. updating parameters of first-order methods as in [28], considering Riemannian Newton methods as in [13], or other second-order schemes as in [27,21]. Unfortunately, even though we know the current structure of the iterates and we know that they will identify the optimal manifold in finite time, we *never* know if the current manifold is the optimal one.

We propose here a Newton acceleration of the proximal gradient algorithm solving the nonsmooth optimization problem ($\mathcal{P}$), that adaptatively uses identification. Our algorithm uses the same basic ingredients, that work behind the scenes, for the existing nonsmooth Newton algorithms recalled above (e.g. [29], [13], [24]). 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 superlinear convergence of the resulting algorithm under some qualification assumptions – but without prior knowledge on the final optimal submanifold. 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). Along the way, our study reveals

results that have some interest on their own, in particular: we refine the smoothness properties of the proximal gradient operator around structured critical points; we formalize complementary properties on line searches in Riemannian optimization; we also bring a careful attention to the technical details induced by nonconvexity.

The paper is organized as follows. First, in Section 2 we recall the useful notions of Riemannian optimization and variational analysis. Then, we introduce in Section 3 our template algorithm alternating a proximal gradient step with a Riemannian update on the identified manifold. 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 results are well-known and just recalled here, but several others seem to be less-known or not precisely treated in the 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 monographs [2] for Riemannian optimization and [32] 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 [11] for more extensive presentations. In the rest of the paper, $\mathcal{M}$ denotes a submanifold of $\mathbb{R}^n$ or $\mathbb{R}^{m\times n}$.

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 $\mathcal{C}^2$ function $\varphi : \mathbb{R}^p \to \mathbb{R}^n$ such that φ maps a neighborhood of $0 \in \mathbb{R}^p$ to a neighborhood of $\bar{x} \in \mathcal{M}$, that admits a smooth (local) inverse, and which derivative 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 $\Phi : \mathbb{R}^n \to \mathbb{R}^{n-p}$ with a surjective derivative at $\bar{x} \in \mathcal{M}$ that satisfies for all x close enough to $\bar{x}$: $x \in \mathcal{M} \Leftrightarrow \Phi(x) = 0$.

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 $\gamma : I \subset \mathbb{R} \to \mathcal{M} \subset \mathbb{R}^n$, 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:

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

The tangent space is a p-dimensional space containing *tangent vectors*. Each tangent space $T_x\mathcal{M}$ is equipped with a scalar product $\langle\cdot,\cdot\rangle_x : T_x\mathcal{M} \times T_x\mathcal{M} \to \mathbb{R}$, 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 φ or local equations Φ defining $\mathcal{M}$:

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

A *metric* on $\mathcal{M}$ can be defined as the minimal length over all curves joining two points $x, y \in \mathcal{M}$, ie. $\text{dist}_\mathcal{M}(x,y) = \inf_{c\in C_{x,y}} \int_0^1 \|c'(t)\|_{c(t)}\mathrm{d}t$, where $C_{x,y}$ is the set of $[0,1] \to \mathcal{M}$ 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. Let $F : \mathcal{M} \to \mathbb{R}$, the *Riemannian differential* of F at x is the linear operator $\mathrm{D}F(x) : T_x\mathcal{M} \to \mathbb{R}$ defined by $\mathrm{D}F(x)[\eta] \triangleq \frac{\mathrm{d}}{\mathrm{d}t} F \circ c(t)\big|_{t=0}$, where c is a smooth curve such that $c(0) = x$ and $c'(0) = \eta$. In turn, the *Riemannian gradient* $\operatorname{grad} F(x)$ is the unique vector of $T_x\mathcal{M}$ such that, for any tangent vector η, $\mathrm{D}F(x)[\eta] = \langle \operatorname{grad} F(x), \eta\rangle$. If $\operatorname{grad} F(x)$ 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 η at 0; then, for t near 0,

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

Notions of derivation for vector fields and of acceleration for curves are used to define second-order objects. Let a curve $c : I \to \mathcal{M}$ and a smooth vector field Z on c, ie. 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 $\frac{\mathrm{D}}{\mathrm{d}t} Z : I \to T\mathcal{B}$, is defined by $\frac{\mathrm{D}}{\mathrm{d}t} Z(t) \triangleq \text{proj}_{c(t)} Z'(t)$, where $Z'(t)$ denotes the derivative in the ambient space $\mathbb{R}^n$ and 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: $c''(t) \triangleq \frac{\mathrm{D}}{\mathrm{d}t} c'(0)$.

The *Riemannian Hessian* of F at x along η is the linear operator $\operatorname{Hess} F(x) : T_x\mathcal{M} \to T_x\mathcal{M}$ defined by the relation $\operatorname{Hess} F(x)[\eta] \triangleq \frac{\mathrm{D}}{\mathrm{d}t} \operatorname{grad} F(c(t))\big|_{t=0}$, where c is a smooth curve such that $c(0) = x$, $c'(0) = \eta$. Equivalently, we have $\langle \operatorname{Hess} F(x)[\eta], \eta\rangle = \frac{\mathrm{d}^2}{\mathrm{d}t^2} F \circ \gamma(t)\Big|_{t=0}$, where γ is a geodesic 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,

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

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

$$\operatorname{grad} F(x) = \text{proj}_x(\nabla \bar{F}(x)), \tag{2.1}$$

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

$$\operatorname{Hess} F(x)[\eta] = \text{proj}_x\left(\mathrm{D}\,\bar{G}(x)[\eta]\right). \tag{2.2}$$

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 η. A geodesic curve passing at (x, η), while attractive as the generalization of the straight line, has a prohibitive computational cost. We thus *retractions*, i.e. approximations of it, defined on a manifold $\mathcal{M}$ as a smooth map $\mathrm{R} : T\mathcal{B} \to \mathcal{M}$ such that

$$\mathrm{R}_x(0) = x \quad \text{and} \quad \mathrm{D}\,\mathrm{R}_x(0) : T_x\mathcal{M} \to T_x\mathcal{M} \text{ is the identity map: } D\,\mathrm{R}_x(0)[v] = v,$$

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

$$F \circ \mathrm{R}_x(t\eta) = F(x) + t\langle \operatorname{grad} F(x), \eta\rangle + \frac{t^2}{2}\langle \operatorname{Hess} F(x)[\eta], \eta\rangle + o(t^2\|\eta\|^2). \tag{2.3}$$

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

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

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

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

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

$$\mathcal{M}_I \triangleq \{x \in \mathbb{R}^n : x_i = 0 \text{ for } i \in I\}, \tag{2.4}$$

where $I \subset \{1, \ldots, n\}$. 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 $\operatorname{proj}_x(d)$, where $[\operatorname{proj}_x(d)]_i$ is d_i if $i \notin I$, and null otherwise. The map $\mathrm{R}_x(\eta) = x + \eta$ 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 $\mathcal{M}_I$ are obtained from their Euclidean counterparts by a simple projection: for $x, \eta \in T\mathcal{B}$,

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

Example 2.2 (Fixed rank matrices) We consider the manifold of fixed-rank matrices

$$\mathcal{M}_r \triangleq \{x \in \mathbb{R}^{m\times n} : \operatorname{rank}(x) = r\}, \tag{2.5}$$

for which we refer to [11, Sec. 7.5]. 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 $U^\top U = I_n$, $V^\top V = I_m$ and Σ 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 $\eta \in T_x\mathcal{M}_r$ writes

$$\eta = UMV^\top + U_pV^\top + UV_p^\top,$$

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 $\operatorname{proj}_x(d) = d - U^\top dV$. 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 $x, \eta \in T\mathcal{B}$, and with $P_U^\top = I_m - UU^\top$, $P_V^\top = I_n - VV^\top$.

$$\begin{aligned}\operatorname{grad} F(x) &= \operatorname{proj}_x(\nabla F(x))\\ \operatorname{Hess} F(x)[\eta] &= \operatorname{proj}_x(\nabla^2 F(x)[\eta]) + \left[P_U^\top \nabla F(x) V_p \Sigma^{-1}\right] V^\top + U\left[P_V^\top \nabla F(x)^\top U_p \Sigma^{-1}\right]^\top.\end{aligned}$$

2.2 Recalls on nonsmooth optimization

We review the basic notions of variational analysis used in this paper, following the monograph [32]. For this section, $g\colon \mathbb{R}^n \to \bar{\mathbb{R}} = \mathbb{R}\cup\{+\infty\}$ is a proper function.

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

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

is closed and convex, but the subdifferential mapping $\widehat{\partial} g(\cdot)$ may not be outer semi-continuous [32, Th. 8.6, Prop. 8.7]. To overcome this problem, the set of *(general or limiting) subgradients* is defined as

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

The limiting subdifferential is by design outer semi-continuous:

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

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 [32, Def. 7.25, Cor. 8.11]. This is notably the case for convex functions where the two above definitions coincide with the convex subdifferential [32, 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

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

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.

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

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

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[1], 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 as the ℓ_1-norm or the trace-norm; see Section 2.3.

Prox-regularity. 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, 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 $\bar{x}$ [32, Def. 13.27].

This property allows to have local Lipschitzness of the proximal operator as well as its characterization by first-order optimality conditions; see [19, Th. 4] and Lemma A.1. 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 Lipchitz 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, 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 (ℓ_1 norm) In the context of Example 2.1, we consider the ℓ_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 $r=0$. Its proximity operator admits a closed form:

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

which naturally gives sparse outputs. In other words, $x=\mathbf{prox}_{\gamma\|\cdot\|_1}(y)$ lies on $\mathcal{M}_I$ (see (2.4)) where I is the complementary of support of x. Observe also that the restriction of $\|\cdot\|_1$ to the manifold $\mathcal{M}_I$ is locally smooth. The ℓ_1 norm thus admits a Riemannian gradient and Hessian at point x:

$$\operatorname{grad}\|\cdot\|_1(x)=\operatorname{sign}(x) \qquad \text{and} \qquad \operatorname{Hess}\|\cdot\|_1(x)=0,$$

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

[1] The weaker assumption of prox-boundedness (ie. $g+r\|\cdot\|^2$ is bounded below for some r) implies that $\mathbf{prox}_{\gamma g}(y)$ is non-empty when γ is taken sufficiently small; see [32, Chap. 1.G].

Example 2.4 (nuclear norm) Following the notation of Example 2.2, we consider the nuclear norm, defined on $\mathbb{R}^{m\times n}$ as $\|x\|_* = \sum_{i=1}^{\mathrm{rank}(x)} \Sigma_{ii}$, where Σ 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: for matrix y $(= U\Sigma V^\top)$,

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

where the coefficient (i, j) of $(\Sigma - \gamma)_+$ is defined as $\max(\Sigma_{ij} - \gamma, 0)$. Thus, $x = \mathbf{prox}_{\gamma\|\cdot\|_*}(y)$ has low rank, by construction. Said otherwise, x lies on $\mathcal{M}_r$ (see (2.5)) where $r = \mathrm{rank}(\Sigma - \gamma)_+$. Observe also that the restriction of the nuclear norm to the manifold $\mathcal{M}_r$ is locally smooth, and thus admits a Riemannian gradient and Hessian at point x: denoting $\eta = UMV^\top + U_pV^\top + UV_p^\top \in T_x\mathcal{M}_r$ a tangent vector,

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

where $\circ$ denotes the Hadamard product and $\tilde{F} \in \mathbb{R}^{\bar{r}\times\bar{r}}$ 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.2.

3 General proximal algorithm with Riemannian acceleration

As mentioned in the introduction and in the previous examples, the output of a proximity operator often comes with the knowledge of the current manifold on which it lives. In this section, we leverage this ability to an algorithmic advantage by reducing our working space to the identified structure. "Smooth" structures (involving smooth submanifolds and smooth restrictions on it) are of special interest and open the way to Newton acceleration.

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

Assumption 1. *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) $\mathbf{prox}_{\gamma g}$ is non-empty on $\mathbb{R}^n$ for any $\gamma > 0$;
iv) $F(x) = f(x) + g(x)$ is bounded below.

In this setup, we propose a general algorithm (Algorithm 1) which consists in, first, performing a proximal gradient step $x_k \in \mathbf{prox}_{\gamma g}(y_{k-1} - \gamma\nabla f(y_{k-1}))$ 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 $\mathrm{ManAcc}_{\mathcal{M}_k}$ on the current manifold. This algorithm is general in the sense that we do not precise for now what is the Riemannian step ManAcc.

We start in Section 3.1 with a technical result about the local smoothness of the proximal gradient operator. In Section 3.2, we analyse the identification property of this algorithm. In Section 3.3, we study how Riemannian methods with local superlinear convergence propagate their rate to Algorithm 1. We will investigate later in Section 4 the Riemannian Newton acceleration falling into this scheme.

Algorithm 1 General structure exploiting algorithm

Require: Pick x_0 arbitrary, $\gamma < 1/L$.
1: **repeat**
2: Compute $x_k \in \mathbf{prox}_{\gamma g}(y_{k-1} - \gamma \nabla f(y_{k-1}))$ and get $\mathcal{M}_k \ni x_k$
3: Update $y_k = \mathrm{ManAcc}_{\mathcal{M}_k}(x_k)$ on the current manifold
4: **until** stopping criterion

3.1 Smoothness and localization of the proximal gradient

The results of this section are built on g being a partly smooth function; see [25].

Definition 3.1 (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 ∂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 [25, Sec. 5]. The proof extends and refines the rationale of [13, Th. 28] and [31, Th. 4.4] that deal with the proximity operator. We use this extension to allow for a full stepsize range of $(0, 1/r)$ in the proximal gradient around any point $\bar{x}$.

Theorem 3.1 (Proximal gradient points smoothly locate 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 γ, $\bar{\gamma}$ 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) $\frac{1}{\bar{\gamma}}(\bar{y} - \bar{x}) - \nabla f(\bar{y}) \in \mathrm{ri}\, \partial g(\bar{x})$ (the relative interior of the subdifferential at $\bar{x}$);
ii) either a) γ 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 [25, Sec. 5] and [13, Sec. 4.1], we consider the function

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

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 ρ_y.

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

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

Since g is $\mathcal{C}^2$-partly-smooth relative to $\mathcal{M}$ and f is $\mathcal{C}^2(\mathbb{R}^n)$, ρ_y is twice continuously differentiable on $\mathcal{M}$. Moreover, the r-prox-regularity gives easily (see Lemma A.2)

that $\rho_{\bar{y}}$ is lower-bounded by $(\frac{1}{\gamma} - r)\|\cdot - \bar{x}\|^2/2$ on a neighborhood of $\bar{x}$ in $\mathbb{R}^n$ and, a fortiori, in $\mathcal{M}$. From usual rationale (see e.g. [11, Chap 4.2, 6.1]), this implies

$$\operatorname{grad} \rho_{\bar{y}}(\bar{x}) = 0 \qquad \operatorname{Hess} \rho_{\bar{y}}(\bar{x}) \succeq \Big(\frac{1}{\gamma} - r\Big) I \succ 0,$$

which are the conditions to apply the implicit functions theorem, as follows.

We consider the equation $\Phi(x, y) = 0$, for x, y near $\bar{x}, \bar{y}$, where $\Phi : \mathcal{M} \times \mathbb{R}^n \to T\mathcal{B}$ is defined as $\Phi(x, y) = \operatorname{grad} \rho_y(x)$. This function is continuously differentiable on a neighborhood of $(\bar{x}, \bar{y})$, and its differential relative to $\bar{x}$ at that point, $\operatorname{Hess} \rho_{\bar{y}}(\bar{x})$, 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 $\hat{x} : \mathcal{N}_{\bar{y}} \to \mathcal{N}_{\bar{x}}$ such that, for any y in $\mathcal{N}_{\bar{y}}$, $\Phi(\hat{x}(y), y) = \operatorname{grad} \rho_y(\hat{x}(y)) = 0$. Actually, $\hat{x}(y)$ is a strong minimizer of ρ_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 $y \mapsto \operatorname{Hess} \rho_y(\hat{x}(y))$ is also continuous there and the property $\operatorname{Hess} \rho_{\bar{y}}(\hat{x}(\bar{y})) \succ 0$ extends locally around $\bar{y}$.

Step 2. As a second step, we turn to show that the minimizer $\hat{x}(y)$ of ρ_y on $\mathcal{M}$ is actually a strong critical point of ρ_y in $\mathbb{R}^n$ [25, 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 $0 \in \operatorname{ri} \partial \rho_y(x)$, that is

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

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 $\lim_{r\to\infty} h_r = \bar{h}$. The above property still holds at the limit $r \to \infty$. 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 $\lim_{r\to\infty} u_r = \bar{u}$. 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 ρ_y on $\mathcal{M}$, iii) $0 \in \operatorname{ri} \partial \rho_y(\hat{x}(y))$.

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 A.1) gives that $\hat{x}(y) = \mathbf{prox}_{\gamma g}(y - \gamma \nabla f(y))$ for y close enough to $\bar{y}$. □

3.2 Structure indentification

Theorem 3.1 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*, an illustration of which is depicted on Fig. 1.

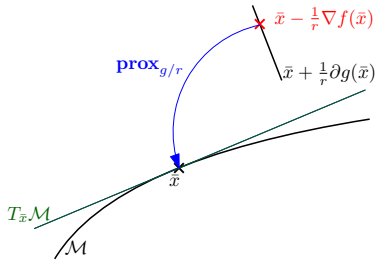


Fig. 1: Illustration of a r-structured critical point. Point i) is illustrated by the blue arrow, and point ii) implies that the red cross is in the interior of the black segment. Partial smoothness appears in the fact that the black segment is perpendicular to the tangent plane of $\mathcal{M}$ at $\bar{x}$.

Definition 3.2 A point $\bar{x}$ of a C^2 submanifold $\mathcal{M}$ is a *r-structured critical* point for (f, g) if we have:

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

While ii),iii),iv) are standard in the literature (see e.g. [13]), i) is not always explicited (an exception is for the notion of identifiability in [17]). It is directly verified when g is convex (for any $r > 0$), but this is not the case when g is nonconvex.[2]

Using this notion and Theorem 3.1, we get the precise identification result of the proximal gradient algorithm, that we need in the forthcoming analysis.

Corollary 3.1 (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 $y_k \to \bar{x}$, then $x_k \triangleq \mathbf{prox}_{\gamma g}(y_k - \gamma\nabla f(y_k)) \in \mathcal{M}$ for k large enough.*

Proof. The notion of r-structured critical point allows us to apply Theorem 3.1 with $\bar{x} = \bar{y} \in \mathcal{M}$ and $\bar{\gamma} = 1/r$. So we get that, 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 $\bar{x}$. Since its input (y_k) converges to $\bar{x}$, the proximal gradient mapping reaches the neighborhood in finite time, which guarantees that (y_k) are $\mathcal{M}$-valued. □

[2] The following example shows that in the nonconvex setting, ii) and iii) do not necessarily imply i). Take f null and g as follows, then the proximity operator of g at 0 writes:

$$g(x) = \begin{cases} x^2/2 & \text{if } |x| \le 1 \\ 1 - 3x/2 & \text{if } x \ge 1 \\ 1 + 3x/2 & \text{if } x \le -1 \end{cases}, \qquad \mathbf{prox}_{\gamma g}(0) = \begin{cases} 0 & \text{if } \gamma \in (0, 8/9) \\ \{-3\gamma/2, 0, 3\gamma/2\} & \text{if } \gamma = 8/9 \\ \{-3\gamma/2, 3\gamma/2\} & \text{if } \gamma > 8/9. \end{cases}$$

The function g is 1-prox-regular at 0, there holds $0 \in \mathrm{ri}\,\partial g(0) = \{0\}$, and yet 0 is not a fixed point of the proximal operator with stepsizes close to 1.

3.3 Superlinear convergence

Using the structure identification result 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 3.2 *Let Assumption 1 hold and take $\gamma \in (0, 1/L)$, where L is the Lipschitz constant for ∇f. 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) $\mathrm{ManAcc}_{\mathcal{M}}$ has superlinear convergence rate of order $1 + \theta \in (1, 2)$ near $\bar{x}$ 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 ManAcc*:*

$$\mathrm{dist}_{\mathcal{M}}(x_{k+1}, \bar{x}) \leq c\ \mathrm{dist}_{\mathcal{M}}(x_k, \bar{x})^{1+\theta} \qquad \text{for some } c > 0. \tag{3.1}$$

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, Theorem 3.1 implies that there exists $\delta_1 > 0$ and $C > 0$ such that,

$$\mathsf{T}(y) \in \mathcal{M} \quad \text{and} \quad \|\mathsf{T}(y) - \mathsf{T}(\bar{x})\| \leq C\|y - \bar{x}\| \qquad \text{for all } y \text{ in } \mathcal{B}(\bar{x}, \delta_1).$$

Now, if y belongs to $\mathcal{M}$, we get 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 Lipchitz continuity can be translated into a Riemannian one (see Lemma B.2) since for some $\delta > 0$,

$$\begin{aligned}(1-\delta)\mathrm{dist}_{\mathcal{M}}(\mathsf{T}(y), \bar{x}) = (1-\delta)\mathrm{dist}_{\mathcal{M}}(\mathsf{T}(y), \mathsf{T}(\bar{x})) &\leq \|\mathsf{T}(y) - \mathsf{T}(\bar{x})\| \\ &\leq C\|y - \bar{x}\| \leq C(1+\delta)\mathrm{dist}_{\mathcal{M}}(y, \bar{x})\end{aligned} \tag{3.2}$$

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

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

Then, the part ii) of the assumptions gives us the existence of ε_2, $q_2 > 0$ and $\theta \in (0, 1)$ such that, for any x in $\mathcal{B}_{\mathcal{M}}(\bar{x}, \varepsilon_2)$,

$$\mathrm{dist}_{\mathcal{M}}(\mathrm{ManAcc}_{\mathcal{M}}(x), \bar{x}) \leq q_2\ \mathrm{dist}_{\mathcal{M}}(x, \bar{x})^{1+\theta}. \tag{3.4}$$

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

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

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

$$\begin{aligned}\mathrm{dist}_{\mathcal{M}}(\mathsf{T}(\mathrm{ManAcc}_{\mathcal{M}}(x)), \bar{x}) &\leq q_1 \mathrm{dist}_{\mathcal{M}}(\mathrm{ManAcc}_{\mathcal{M}}(x), \bar{x}) \\ &\leq q_1 q_2\ \mathrm{dist}_{\mathcal{M}}(x, \bar{x})^{1+\theta} \leq q_1 q_2\ \varepsilon^{\theta} \mathrm{dist}_{\mathcal{M}}(x, \bar{x}).\end{aligned} \tag{3.5}$$

Since $q_2 q_1 \varepsilon^\theta \leq 1$ by construction, this gives

$$\operatorname{dist}_{\mathcal{M}}(\mathsf{T}(\operatorname{ManAcc}_{\mathcal{M}}(x)), \bar{x}) \leq \operatorname{dist}_{\mathcal{M}}(x, \bar{x}) \qquad \text{for any } x \in \mathcal{B}_{\mathcal{M}}(\bar{x}, \varepsilon). \tag{3.6}$$

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), there exists $K < \infty$ such that $y_K \in \mathcal{B}(\bar{x}, (1-\delta)\varepsilon/C)$. Besides, (3.2) tells us that $\operatorname{dist}_{\mathcal{M}}(\mathsf{T}(y_K), \bar{x}) \leq \varepsilon$ and thus x_k and y_k belong to $\mathcal{B}_{\mathcal{M}}(\bar{x}, \varepsilon)$ for all $k > K$ by (3.6). We conclude that $x_{k+1} = \mathsf{T}(y_k) \in \mathcal{M}$ for all $k \geq K$, and, using (3.5), that we have (3.1) with $c = q_1 q_2$, for all $k > K$. □

4 Newton acceleration

In this section, we investigate the possibilities of manifold acceleration within Algorithm 1. We show in Sections 4.2 and 4.3 how to use Riemannian (truncated) Newton accelerations within our framework and derive superlinear/quadratic convergence guarantees. A technical difficulty to ensure global convergence when interlacing proximal gradient updates with Riemannian Newton accelerations is to guarantee some functional decrease. Thus, we first study in Section 4.1 the use of line search for $\operatorname{ManAcc}_{\mathcal{M}}$ in our context.

4.1 Ensuring functional descent while preserving local rates: line search

We use in the following convergence proofs three properties of $\operatorname{ManAcc}_{\mathcal{M}}$: it should produce an update that lives on $\mathcal{M}$, enjoy a superlinear local convergence rate, and 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 monograph [15].

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

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

The line search employs a second-order retraction R_x (that could be the exponential map). The conditions under which stepsizes satisfying the Armijo rule exist are discussed in [15, Sec 6.3], the following lemma can then be derived.

Lemma 4.1 *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, $\langle \operatorname{grad} F(x), \eta \rangle < 0$, 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 [15, Th. 6.3.2] for the Armijo rule and the Riemannian setting. Since $m_1 < 1/2$, for any α sufficiently small there holds

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

Since F is bounded below, there exists a smallest $\hat{\alpha}$ such that $F(\mathrm{R}_x(\hat{\alpha}\eta)) = F(x) + m_1\hat{\alpha}\langle \operatorname{grad} F(x), \eta\rangle$. Thus all stepsizes in $(0, \hat{\alpha})$ are acceptable by (4.1). □

In addition, a line search performed near a minimizer with a Newton direction 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, which is a direct corollary of Theorem B.1.

Lemma 4.2 *Let Assumption 1 hold and consider a manifold $\mathcal{M}$ equipped with a retraction R, a point $x^\star \in \mathcal{M}$ and a pair $(x, \eta) \in T\mathcal{B}$. Assume that F is twice differentiable on $\mathcal{M}$ near a strong local minimizer $x^\star$ on $\mathcal{M}$, that is $\operatorname{Hess} F(x^\star)$ is positive definite. If the direction η brings a superlinear improvement towards $x^\star$, that is $\operatorname{dist}_{\mathcal{M}}(\mathrm{R}_x(\eta), x^\star) = o(\operatorname{dist}_{\mathcal{M}}(x, x^\star))$ as $x \to x^\star$, and $0 < m_1 < 1/2$, then η is acceptable by the Armijo rule (4.1) with unit stepsize $\alpha = 1$.*

In the following, we will consider a *backtracking* line search for finding an acceptable stepsize α: the unit stepsize is first tried, and then the search space is reduced geometrically. In practice, we use exactly [15, Alg. A6.3.1], which features polynomial interpolation of F in the search space.

4.2 Riemannian Newton & quadratic convergence

We construct a manifold update based on the Riemannian Newton method [2, Chap. 6], which is the simplest method with a local quadratic convergence. It consists in finding $d \in T_x\mathcal{M}$ that minimizes the second order model (2.3) of F at point $x \in \mathcal{M}$, or equivalently that solves Newton equation; see [11, Sec. 6.2].

Algorithm 2 ManAcc-Newton

Require: Manifold $\mathcal{M}$, point $x \in \mathcal{M}$
1: Find d in $T_x\mathcal{M}$ that solves

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

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

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 $\operatorname{Hess} F(x_k)$ 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 a limit point $x^\star$ such that

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

Then, after some finite time,

a) the sequence (x_k) lies on $\mathcal{M}$;
b) x_k converges to $x^\star$ quadratically: for large k, there exists $c > 0$ such that

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

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

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

The Riemannian Newton manifold step is therefore well-defined, and the line search terminates by Lemma 4.1, so that the manifold update is well-defined and provides descent ($F(y_k) \leq F(x_k)$).

Now, since the proximal gradient update provides a descent (see [6, Lem. 10.4]),

$$F(x_{k+1}) \leq F(y_k) - \frac{1-\gamma L}{2\gamma}\|x_{k+1} - y_k\|^2 \leq F(x_k) - \frac{1-\gamma L}{2\gamma}\|x_{k+1} - y_k\|^2. \quad (4.2)$$

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 (4.2) for $k = 1, \ldots, n$ yields:

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

Since $\operatorname{dist}(0, \partial F(x_{k+1})) \leq \frac{L\gamma+1}{\gamma}\|x_{k+1} - y_k\|$ (see e.g. the proof of [9, Prop. 13]), we have that $\operatorname{dist}(0, \partial F(x_{k+1}))$ converges to 0. The outer-semi continuity of the limiting subdifferential then yields criticiality of accumulation points.

Now we apply the local convergence of Riemannian Newton [2, Th. 6.3.2]: assumption ii) ensures that the Riemannian Newton direction d computed in step 1 of Algorithm 2 provides a quadratic improvement on a neighborhood of $x^\star$ on $\mathcal{M}$. Moreover, the line search returns the unit-stepsize after some finite time: $\alpha = 1$ is tried first, and is acceptable for directions providing superlinear improvement by Lemma 4.2. Thus the whole Riemannian Newton update provides quadratic improvement after some finite time. Using this and assumption i), Theorem 3.2 applies and yields the results. □

This theorem states that alternating proximal gradient steps and Riemannian Newton steps converges quadratically to structured points with virtually the same assumptions the Euclidean Newton method. However, the two standard issues of Newton's method still hold in our setting: at each iteration, a linear system has to be solved to produce the Newton direction; and this direction does not always provide descent (without positive definiteness of the Hessian). We show in the next section that truncated versions overcome these issues also in our framework.

4.3 Riemannian Truncated Newton & superlinear convergence

We consider a manifold update based on a truncated Newton procedure [14]. (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. 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:

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

Algorithm 3 ManAcc-Newton-CG

Require: Manifold $\mathcal{M}$, point $x \in \mathcal{M}$, convergence defining parameter $\theta \in (0, 1]$
1: Let $\eta = \|\operatorname{grad} F(x)\|^{\theta}$
2: Find d that solves (Inexact Newton eq.)
3: Find α 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 a limit point $x^\star$ such that

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

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

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

Proof. The direction provided by (Inexact Newton eq.) is a descent direction by Lemma B.3, the line search terminates by Lemma 4.1, so that the updates are well-defined and provide descent. Thus, as in the proof of Theorem 4.1 we get that every accumulation point of the iterate sequence is a critical point for F. We can apply now the local convergence of the Riemannian truncated Newton method [2, Th. 8.2.1]: assumptions ii) and iii) ensure that the direction d computed in step 1 of Algorithm 3 provides a local superlinear improvement towards $x^\star$. The end of the proof is the same as the one of the proof of Theorem 4.1. □

5 Numerical illustrations

In this section, we illustrate the effect of Newton acceleration. 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'. These methods are compared to the Proximal Gradient and the Accelerated Proximal Gradient, which serve as baseline. The algorithms and problems are implemented in Julia [8]; experiments may be reproduced using the code available at https://github.com/GillesBareilles/NewtonRiemannAccel-ProxGrad.

We report the numerical results in figures showing a) the suboptimality $F(x_k) - F(x^\star)$ 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:

$F(x_k) - F(x^\star)$	Suboptimality at current iteration.
#prox. grad. steps	Number of proximal gradient steps, each involve computing $\nabla f(\cdot)$ and $\mathbf{prox}_{\gamma g}(\cdot)$ once.
#ManAcc steps	Number of Riemannian steps, each involve computing grad $F(\cdot)$ once and 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 line search + some for the backtracking estimation of the Lipchitz constant.
$\#g$	Number of calls to $g(x)$, one per iteration + some for the line search.

The proximal gradient updates, present in all methods, include a backtracking procedure that maintains an estimate of the Lipschitz constant of ∇f, so that the proximal gradient step length is taken as the inverse of that estimate. The Conjugate Gradient used to solve (Newton equation) and (Inexact Newton eq.) follows [11, 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 $\langle \mathrm{Hess}\, F(x_k)[d], d\rangle / \|d\|^2$ small.[3] The manifold updates are completed by a backtracking line search started from unit stepsize, a direct implementation of [15, Alg. 6.3.1].

5.1 Two-dimensional nonsmooth example

We consider the piecewise quadratic problem of [24]:

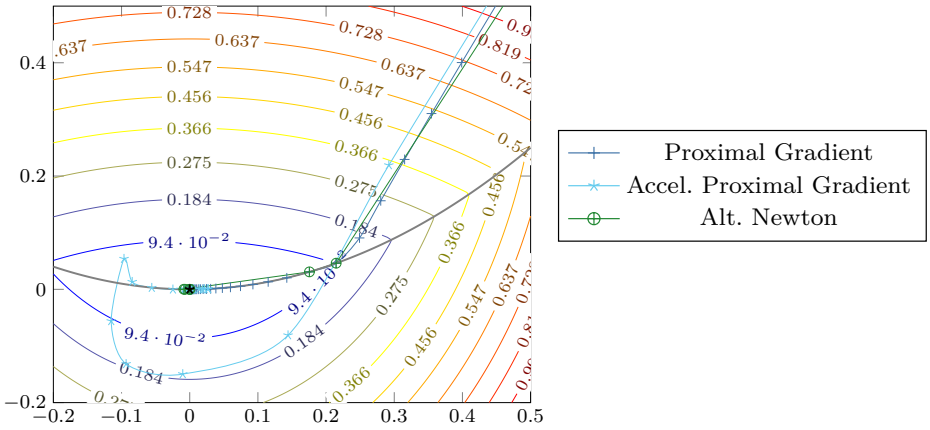
$$\min_{x\in\mathbb{R}^2} 2x_1^2 + x_2^2 + |x_1^2 - x_2|.$$

The objective function is partly-smooth relative to the parabola $\{x : x_2 = x_1^2\}$, 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 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 $(2, 3)$.

Observations The iterates are displayed in Fig. 2. The Proximal Gradient iterates reach the parabola in finite time, and then converge linearly on the parabola while the Accelerated Proximal Gradient iterates "overshoot" the optimal manifold (see [5]). The iterates of the Alt. Newton method stay on the parabola and the quadratic convergence behavior appears clearly since two Newton updates bring suboptimality below 10^{-3}, and one additional step gets it below 10^{-12}.

[3] Each CG iteration requires one Hessian-vector product, avoiding to form the Hessian matrix. A test on this ratio is used to detect a direction of quasi-negative curvature for the (Riemannian) Hessian, which is a stopping criterion of the Conjugate Gradient. In our implementation, we require this quantity to be smaller than 10^{-15} for the Newton method. For the truncated version, we reduce the threshold when getting close to the solution: intialized at 1, the threshold is decreased by a factor 10 each time the unit-step is accepted by the line search.



Algorithm	Tolerance	$F(x_k) - F(x^\star)$	#prox. grad. steps	#ManAcc 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 example

5.2 ℓ_1-regularized logistic problem

We now turn to the ℓ_1-regularized logistic problem:

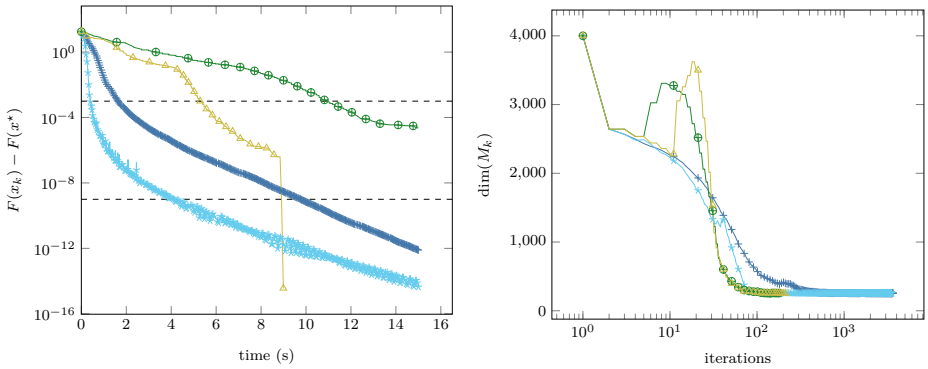
$$\min_{x \in \mathbb{R}^n} \frac{1}{m} \sum_{i=1}^{m} \log(1 + \exp(-y_i \langle A_i, x \rangle)) + \lambda \|x\|_1, \tag{5.1}$$

where $A \in \mathbb{R}^{m \times n}$, $y \in \{-1, 1\}^m$, and $\lambda > 0$. The nonsmooth part $g(x) = \lambda \|x\|_1$ is described in Section 2.3.

We consider an instance where $n = 4000$, $m = 400$, $\lambda = 10^{-2}$ and the final manifold has dimension 249. The coefficients of A are drawn independently following a normal law. From a sparse random vector s, y_i is set to 1 with probability $1/(1 + \exp(-\langle A_i, s \rangle))$, and -1 otherwise. All algorithms start from the same point which is the output of 35 iterations of the accelerated proximal gradient randomly initiated.

Observations The experiments are presented in Fig. 3. The optimal manifold is identified around iteration 200 for all methods except for Proximal Gradient, which needs 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.

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



(a) Suboptimality vs time (s) **(b)** Manifold dimension vs iteration

Prox. Gradient — Accel. Prox. Gradient — Alt. Newton — Alt. Truncated Newton

Algorithm	Tolerance	$F(x_k) - F(x^\star)$	#prox. grad. steps	#ManAcc 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-ℓ_1 problem

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 point having one non-null but very small coordinate. There, the second-order Taylor extension is valid on a small set however it may lead to a Newton step that lies outside that set, thus driving the iterate away. 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, Algorithm 1 retains a good rate overall.

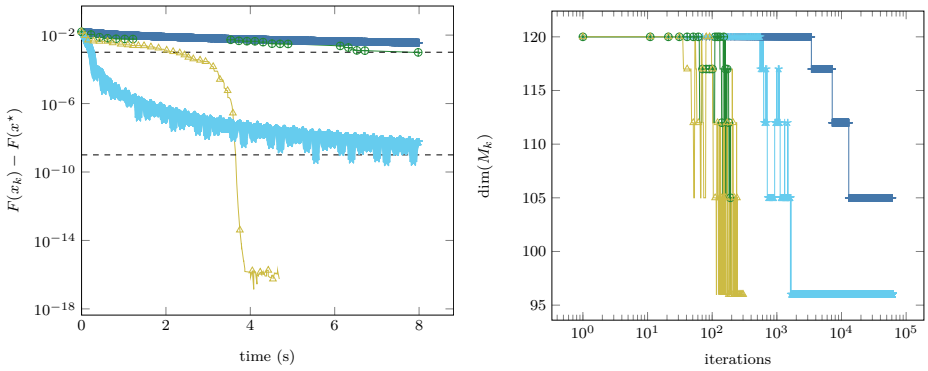
5.3 Trace-norm regularized problem

We consider the following matrix regression problem:

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

where $A_i \in \mathbb{R}^{n_1 \times n_2}$ for $i = 1, \ldots, m$, $y \in \mathbb{R}^m$ and λ 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 $n_1 = 10$, $n_2 = 12$, $m = 60$, $\lambda = 10^{-2}$ and the final manifold is that of matrices of rank 6. The coefficients of the A_i's are drawn independently from a normal law. From a sparse random vector s, y_i is taken as $\langle A_i, s \rangle + \xi_i$, where ξ_i follows a centered normal law with variance 0.01^2. All algorithms start from the same point which is the output of 10^3 iterations of the accelerated proximal gradient randomly initiated.



(a) Suboptimality vs time (s) **(b)** Manifold dimension vs iteration

Prox. Gradient — Accel. Prox. Gradient — Alt. Newton — Alt. Truncated Newton

Algorithm	Tolerance	$F(x_k) - F(x^\star)$	#prox. grad. steps	#ManAcc steps	#Hess $F(\cdot)[\cdot]$	$\#f$	$\#g$
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

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. The 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 of the final manifold, and (*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], cubic regularization [3]) as well as other proximal algorithms (e.g. prox-Newton [22], fast proximal gradient [7]). We focused here on the simplest Newton acceleration to highlight the ideas and the working horses of our approach.

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. Beck, A.: First-order methods in optimization, vol. 25. SIAM (2017)
7. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM journal on imaging sciences **2**(1), 183–202 (2009)
8. Bezanson, J., Edelman, A., Karpinski, S., Shah, V.B.: Julia: A fresh approach to numerical computing. SIAM review **59**(1), 65–98 (2017)
9. Bolte, J., Nguyen, T.P., Peypouquet, J., Suter, B.W.: From error bounds to the complexity of first-order descent methods for convex functions. Mathematical Programming (2015)
10. Bonnans, J.F., Gilbert, J.C., Lemaréchal, C., Sagastizábal, C.A.: Numerical optimization: theoretical and practical aspects. Springer Science & Business Media (2006)
11. Boumal, N.: An introduction to optimization on smooth manifolds. Available online (2020). URL http://www.nicolasboumal.net/book
12. Burke, J.V., Moré, J.J.: On the identification of active constraints. SIAM Journal on Numerical Analysis **25**(5), 1197–1211 (1988)
13. Daniilidis, A., Hare, W., Malick, J.: Geometrical interpretation of the predictor-corrector type algorithms in structured optimization problems. Optimization **55**(5-6) (2006)
14. Dembo, R.S., Steihaug, T.: Truncated-newtono algorithms for large-scale unconstrained optimization. Mathematical Programming **26**(2), 190–212 (1983)
15. Dennis Jr, J.E., Schnabel, R.B.: Numerical methods for unconstrained optimization and nonlinear equations. SIAM (1996)
16. Dolan, E.D., Moré, J.J.: Benchmarking optimization software with performance profiles. Mathematical programming **91**(2), 201–213 (2002)
17. Drusvyatskiy, D., Lewis, A.S.: Optimality, identifiability, and sensitivity. Mathematical Programming **147**(1-2), 467–498 (2014)
18. *et al*, O.B.: Comparison of bundle and classical column generation. Mathematical Programming **113**(2) (2008)
19. Hare, W., Sagastizábal, C.: Computing proximal points of nonconvex functions. Mathematical Programming **116**(1-2), 221–258 (2009)
20. Iutzeler, F., Malick, J.: Nonsmoothness in machine learning: specific structure, proximal identification, and applications. Set-Valued and Variational Analysis **28**(4), 661–678 (2020)
21. Lee, C.p.: Accelerating inexact successive quadratic approximation for regularized optimization through manifold identification. arXiv preprint arXiv:2012.02522 (2020)
22. 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)
23. Lemaréchal, C., Oustry, F., Sagastizábal, C.: The u-lagrangian of a convex function. Transactions of the American mathematical Society **352**(2), 711–729 (2000)
24. Lewis, A., Wylie, C.: A simple newton method for local nonsmooth optimization. arXiv preprint arXiv:1907.11742 (2019)
25. Lewis, A.S.: Active sets, nonsmoothness, and sensitivity. SIAM Journal on Optimization **13**(3), 702–725 (2002)
26. Lewis, A.S., Liang, J.: Partial smoothness and constant rank. arXiv preprint arXiv:1807.03134 (2018)
27. Lewis, A.S., Wright, S.J.: A proximal method for composite minimization. Mathematical Programming **158**(1), 501–546 (2016). DOI 10.1007/s10107-015-0943-9
28. 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)
29. Mifflin, R., Sagastizábal, C.: A vu-algorithm for convex minimization. Mathematical programming **104**(2-3), 583–608 (2005)

30. 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)
31. Poliquin, R., Rockafellar, R.: Prox-regular functions in variational analysis. Transactions of the American Mathematical Society **348**(5), 1805–1838 (1996)
32. Rockafellar, R.T., Wets, R.J.B.: Variational analysis, vol. 317. Springer (2009)
33. Wright, S.J.: Identifiable surfaces in constrained optimization. SIAM Journal on Control and Optimization **31**(4), 1063–1079 (1993)

A Preliminary Results on the Proximal Gradient

The fisrt result show the local Lipschitz continuity of the proximity operator. It can be proven using [31, Th. 4.4] using the arguments of [19, Th. 4] and the assumption that $\bar{x} = \mathbf{prox}_{\bar{\gamma} g}(\bar{y})$. We provide here a self-contained proof.

Lemma A.1 *Consider a function $g : \mathbb{R}^n \to \mathbb{R}$, a pair of points $\bar{x}, \bar{y}$ and a step length $\bar{\gamma} > 0$ 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 $\|(y - x)/\gamma - \bar{v}\| < \varepsilon$.

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

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

which we may write as

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

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

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

We may thus apply Theorem 4.4 from [31] 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 ε-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 $\partial g(x)$, 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 [31, Th. 4.4] includes a minor error relative to the Lipschitz constant computation, we report here a corrected value. □

Now, we show that critical points of prox-regular functions are strong local minimizers; this result appears more or less explicitly in some articles, including [13].

Lemma A.2 *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, the function $\rho_{\bar{y}}: x\mapsto g(x)+\frac{1}{2\gamma}\|\bar{y}-\gamma\nabla f(\bar{y})-x\|^2$ satisfies*

$$\rho_{\bar{y}}(x)\geq\rho_{\bar{y}}(\bar{x})+\frac{1}{2}\left(\frac{1}{\gamma}-r\right)\|x-\bar{x}\|^2,\qquad \textit{for all } x \textit{ near } \bar{x}.$$

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

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

which is the claimed inequality. □

B Technical results on Riemannian methods.

In this section, we provide basic results on Riemannian optimization that simplify our developements 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 Riemmannian and the Euclidian distances.

Lemma B.1 *Consider a point $\bar{x}$ of a Riemannian manifold $\mathcal{M}$, equipped with a retraction R such that $\mathrm{R}_{\bar{x}}$ is $\mathcal{C}^2$. For any $\varepsilon>0$, there exists a neighborhood $\mathcal{U}$ of $\bar{x}$ in $\mathcal{M}$ such that*

$$(1-\varepsilon)\mathrm{dist}_{\mathcal{M}}(x,\bar{x})\leq\|\mathrm{R}_{\bar{x}}^{-1}(x)\|\leq(1+\varepsilon)\mathrm{dist}_{\mathcal{M}}(x,\bar{x})\qquad \textit{for all } x\in\mathcal{U}.$$

where $\mathrm{R}_{\bar{x}}^{-1}:\mathcal{M}\to T_{\bar{x}}\mathcal{M}$ is the smooth inverse of $\mathrm{R}_{\bar{x}}$ defined locally around $\bar{x}$.

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

We consider the function $f:\mathcal{M}\to\mathbb{R}$ defined by $f(x)=\|\log_{\bar{x}}(x)\|-\|\mathrm{R}_{\bar{x}}^{-1}(x)\|$. Clearly $f(\bar{x})=0$, and $\mathrm{D}\,f(\bar{x})=0$ as the differentials of both $\mathrm{R}_{\bar{x}}^{-1}$ 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 $\hat{f}=f\circ\exp_{\bar{x}}:T_{\bar{x}}\mathcal{M}\to\mathbb{R}$. As $\hat{f}(\hat{\bar{x}})=0$, $\mathrm{D}\,\hat{f}(\hat{\bar{x}})=0$ and $\hat{f}$ is $\mathcal{C}^2$, there exists some $C>0$ such that

$$-C\|\hat{x}-\hat{\bar{x}}\|^2\leq\hat{f}(\hat{x})\leq C\|\hat{x}-\hat{\bar{x}}\|^2\qquad\text{in a neighborhood }\hat{\mathcal{U}}\text{ of }\hat{\bar{x}},.$$

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

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

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

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

as $\hat{x}=\log_{\bar{x}}(x)$, $\hat{\bar{x}}=0$. We conclude with $\mathrm{dist}_{\mathcal{M}}(x,\bar{x})=\|\hat{x}-\hat{\bar{x}}\|=\|\log_{\bar{x}}(x)\|$. □

We recall a slightly specialized version of [30, Th. 2.2], which is essentially the application of the implicit function theorem around a point of a manifold.

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

Proof. Let $\Psi : \mathbb{R}^n \to \mathbb{R}^{n-p}$ denote a $\mathcal{C}^k$ 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 $\mathrm{D}\,\Psi(x)$ is surjective. Consider the equation $\Phi(x, \eta_t, \eta_n) = 0$ around $(\bar{x}, 0, 0)$, with

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

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

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

Since $\bar{x} \in \mathcal{M}$, $\mathrm{D}_{\eta_n}\,\Phi(\bar{x}, 0, 0)$ 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 $\mathcal{C}^k$ function $\eta_n : \mathcal{N}_{\bar{x}}^1 \times \mathcal{N}_0^2 \to \mathcal{N}_0^3$ 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 η_n at $(x, 0)$ along η_t: for $\xi_t \in T_x\mathcal{M}$,

$$\mathrm{D}_{\eta_t}\,\eta_n(x, 0)[\xi_t] = -\left[\mathrm{D}_{\eta_n}\,\Phi(x, 0, 0)\right]^{-1} \mathrm{D}_{\eta_t}\,\Phi(x, 0, 0)[\xi_t].$$

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

Now, define a map $\mathrm{R} : \mathcal{N}_{\bar{x}}^1 \times \mathcal{N}_0^2 \to \mathcal{M}$ by $\mathrm{R}(x, \eta_t) = x + \eta_t + \eta_n(x, \eta_t)$. This map has degree of smoothness $\mathcal{C}^k$ since η_n is $\mathcal{C}^k$, satisfies $\mathrm{R}(x, 0) = x$ since $\eta_n(x, 0) = 0$ and satisfies $\mathrm{D}_{\eta_t}\,\eta_n(x, 0) = I + \mathrm{D}_{\eta_t}(x, 0) = I$. 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 $c(t) = \mathrm{R}(x, t\eta)$ 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

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

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. □

Lemma B.2 *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 $x \in \mathcal{U}$,*

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

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 B.1. As a retraction, it satisfies:

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

Taking $x = \mathrm{R}_{\bar{x}}(\eta)$ allows to write $x = \bar{x} + \mathcal{O}(\|\,\mathrm{R}_{\bar{x}}^{-1}(x)\|^2)$, 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

$$(1 - \varepsilon_1)\|\,\mathrm{R}_{\bar{x}}^{-1}(x)\| \leq \|x - \bar{x}\| \leq (1 + \varepsilon_1)\|\,\mathrm{R}_{\bar{x}}^{-1}(x)\|.$$

By Lemma B.1, for $\varepsilon_2 > 0$ small enough, there exists a neighborhood $\mathcal{U}_2 \subset \mathcal{U}$ of $\bar{x}$ such that,

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

With ε_1, ε_2 such that $1 - \varepsilon = (1 - \varepsilon_1)(1 - \varepsilon_2)$, we combine the two estimates to conclude. □

B.2 Two technical results on Riemannian descent algorithms

We provide here two technical results used in the proofs of Section 4. First, Theorem B.1 adapts [10, Th. 4.16] to the Riemannian setting. Second, Lemma B.3 adapts the proof of [14, Lemma A.2] to the Riemannian setting.

Theorem B.1 (Soundness of the Riemannian line search) *Consider a manifold $\mathcal{M}$ equipped with a retraction* R *and a twice differentiable function $F : \mathcal{M} \to \mathbb{R}$ that admits a strong local minimizer $x^\star$, that is, a point such that* $\operatorname{Hess} F(x^\star)$ *is positive definite. If x is close to $x^\star$, η brings a superlinear improvement towards $x^\star$, that is* $\operatorname{dist}_{\mathcal{M}}(\mathrm{R}_x(\eta), x^\star) = o(\operatorname{dist}_{\mathcal{M}}(x, x^\star))$ *as $x \to x^\star$, and $0 < m_1 < 1/2$, then η 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 $x^\star$ and $\operatorname{dist}_{\mathcal{M}}(\mathrm{R}_x(\eta), x^\star) = o(\operatorname{dist}_{\mathcal{M}}(x, x^\star))$. For convenience, let $x_+ = \mathrm{R}_x(\eta)$ denote the next point.

Following [2] (see e.g. the proof of Th. 6.3.2), we work in local coordinates around $x^\star$, representing any point $x \in \mathcal{M}$ by $\widehat{x} = \log_{x^\star}(x)$ and any tangent vector $\eta \in T_x\mathcal{M}$ by $\widehat{\eta}_x = \mathrm{D}\log_{x^\star}(x)[\eta]$. The function F is represented by $\widehat{F} = F \circ \exp_{x^\star} : T_{x^\star}\mathcal{M} \to \mathbb{R}$. 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: $\operatorname{dist}_{\mathcal{M}}(x, y) = \|\widehat{x} - \widehat{y}\|$. Besides, there holds

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

Indeed, $\mathrm{D}\, F(x)[\eta] = (F \circ \gamma)'(0)$ and $\operatorname{Hess} F(x)[\eta, \eta] = (F \circ \gamma)''(0)$, where γ 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 $\mathrm{D}\, F(x)[\eta] = \langle \operatorname{grad} F(x), \eta \rangle$ in terms of $\mathrm{D}^2\, \widehat{F}(\widehat{x^\star})[\widehat{x} - \widehat{x^\star}]^2$. To do so, we go through the intermediate quantity $\mathrm{D}\, \widehat{F}(\widehat{x})[\widehat{x_+} - \widehat{x}]$, and handle precisely the $o(\cdot)$ terms. By smoothness of $\widehat{F}$ and since $\mathrm{D}\, \widehat{F}(\widehat{x^\star}) = 0$, Taylor's formula for $\mathrm{D}\, \widehat{F}$ writes

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

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

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

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

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

Besides, the following estimate holds $\|\widehat{\eta} - (\widehat{x_+} - \widehat{x})\| = o(\|\widehat{x} - \widehat{x^\star}\|)$. Indeed, as the function $\log_{x^\star} \circ \mathrm{R}_x : T_x\mathcal{M} \to T_{x^\star}\mathcal{M}$ is differentiable, there holds for $\eta \in T_x\mathcal{M}$ small,

$$\log_{x^\star}(\mathrm{R}_x(\eta)) = \log_{x^\star}(\mathrm{R}_x(0)) + \mathrm{D}\log_{x^\star}(\mathrm{R}_x(0))[\mathrm{D}\, \mathrm{R}_x(0)[\eta]] + o(\|\eta\|),$$

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

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

so that the second estimate holds.

Combining the two above estimates allows to conclude that

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

so that overall,

$$\mathrm{D}\, F(x)[\eta] = \mathrm{D}\, \widehat{F}(\widehat{x})[\widehat{x_+} - \widehat{x}] + o(\|\widehat{x} - \widehat{x^\star}\|^2) = -\,\mathrm{D}^2\, \widehat{F}(\widehat{x^\star})[\widehat{x} - \widehat{x^\star}]^2 + o(\|\widehat{x} - \widehat{x^\star}\|^2).$$

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

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

Step 2. The function F admits a second-order development around $x^\star$: applying Eq. (2.3) with the exponential map $\exp_{x^\star}$ as a second-order retraction yields

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

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

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

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

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

Subtracting the two estimates yields

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

which ensures that the Armijo condition is satisfied.

□

Lemma B.3 (Riemannian Newton-CG a descent direction) *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:*

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

where d was obtained solving (Inexact Newton eq.) *with any forcing parameter η.*

Proof. The result is obtained by applying the analysis of [14, 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. □

C Complements to the experimental section

C.1 Oracles of Section 5.1

We detail here the 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

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

  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 $(x, \eta) \in T\mathcal{B}$, $\operatorname{grad} g(x) = 0$ and $\operatorname{Hess} g(x)[\eta] = 0$. Moreover, Euclidean gradient and Hessian-vector product are converted to Riemannian ones using equations (2.1) and (2.2):

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

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

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

C.2 Differentiating the singular-value decomposition

We establish the expressions of the derivative of the matrices involved in the singular value decomposition. These results may be seen as part of folkore, but, 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 C.1 *Consider the manifold of fixed rank matrices $\mathcal{M}_r$, a pair $x, \eta \in T\mathcal{B}$ and a smooth curve $c : I \to \mathcal{M}_r$ 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 $\eta = UMV^\top + U_pV^\top + UV_p^\top$, the above expressions simplify to

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

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

$$\eta = U'\Sigma V^\top + U\Sigma' V^\top + U\Sigma V'^\top \tag{C.1}$$

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

$$U' = U\Omega_U + U_\perp B_U, \tag{C.2}$$

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 $U^\top \times$ (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 η. 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

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

The off-diagonal elements of this equation write

$$\bar{I}_r \circ \left[U^\top \eta V\right] = \Omega_U \Sigma + \Sigma\Omega_V^\top, \tag{C.3}$$

where $\bar{I}_r$ has zeros on the diagonal and ones elsewhere. Adding (C.3)Σ and Σ(C.3)$^\top$ yields

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

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

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

hence $\Omega_U = F \circ \left[U^\top \eta V\Sigma + \Sigma V^\top \eta^\top U\right]$, where $F \in \mathbb{R}^{m-r\times r}$ 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 Ω_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 Σ(C.3) and (C.3)$^\top\Sigma$ yields

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

from which we get $\Omega_V = F \circ \left[\Sigma U^\top \eta V + V^\top \eta^\top U\Sigma\right]$. 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 $V^\top V_p = 0$. □

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

Proposition C.1 *The nuclear norm $g = \|\cdot\|_*$ restricted to $\mathcal{M}_r$ is $\mathcal{C}^2$ and admits a smooth second-order development of the form (2.3) near any point $x = U\Sigma V^\top \in \mathcal{M}_r$. Denoting $\eta = UMV^\top + U_pV^\top + UV_p^\top \in T_x\mathcal{M}_r$ a tangent vector, there holds:*

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

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 $c : I \to \mathcal{M}_r$ denote a smooth curve over $\mathcal{M}_r$ such that $\gamma(0) = x$ and $\gamma'(0) = \eta$, and consider $\varphi = \|c(\cdot)\|_* : I \to \mathbb{R}$. 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 $\varphi(t) = \operatorname{Tr}(\Sigma(t))$. Applying Lemma C.1 yields

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

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

In order to obtain the Riemannian Hessian, let $\bar{Z} : I \to \mathbb{R}^n$ denote a smooth extension of $\operatorname{grad} g(c(\cdot))$, defined by $\bar{Z}(t) = U(t)V(t)^\top$. The Riemannian Hessian is then obtained as $\operatorname{Hess} g(x)[\eta] = \operatorname{proj}_x \bar{Z}'(0)$. The derivative of $\bar{Z}$ at 0 is simply $\bar{Z}'(0) = U'V^\top + UV'^\top$ and thus writes, applying Lemma C.1

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

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

C.3 Additional numerical experiment

We illustrate in this appendix the robustness of the Newton acceleration on several instances of the same problem. More precisely, in the set-up of Section 5.3, we compare the 4 algorithms on 20 random instances of the tracenorm problem, in terms of wallclock time required to reach a suboptimality of 10^{-9}. We then provide in Fig. 5 a performance profile (i.e. the ordinate of a curve at abscissa $t \geq 1$ indicates the proportion of problems for which the corresponding algorithm was able to satisfy the criterion within t times the best algorithm time for each problem; see [16]).

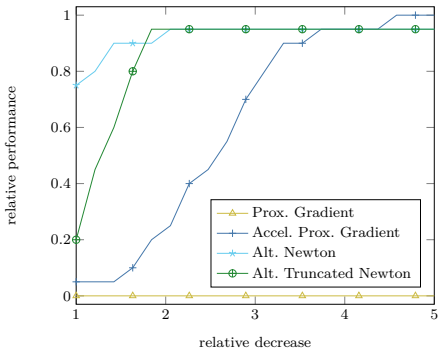


Fig. 5: Performance profile for the time to decrease suboptimality below 10^{-9}

We observe the following on Fig. 5. The ordinate at origin of a curve gives the proportion of problems for which the corresponding algorithm performed best: methods with Newton acceleration are the most efficient in 95%(= 75% + 20%) of the instances. Furthermore, they require about 2.5× less time to converge in half of the instances. Note also that the proximal gradient is completely outperformed by the others algorithms since it takes 5× more time than the best algorithm, for all instances.