Institut für Numerische Simulation
Rheinische Friedrich-Wilhelms-Universität Bonn

Wegelerstraße 6 • 53115 Bonn • Germany
phone +49 228 73-3427 • fax +49 228 73-7527
www.ins.uni-bonn.de

S. Hosseini and A. Uschmajew

A gradient sampling method on algebraic varieties and application to nonsmooth low-rank optimization

INS Preprint No. 1624

October 2016
(Extended and revised version, March 2017)

A GRADIENT SAMPLING METHOD ON ALGEBRAIC VARIETIES AND APPLICATION TO NONSMOOTH LOW-RANK OPTIMIZATION

SEYEDEHSOMAYEH HOSSEINI* AND ANDRÉ USCHMAJEW*

Abstract. In this paper, a nonsmooth optimization method for locally Lipschitz functions on real algebraic varieties is developed. To this end, the set-valued map ε-conditional subdifferential $x \to \partial_\varepsilon^N f(x) := \partial_\varepsilon f(x) + N(x)$ is introduced, where $\partial_\varepsilon f(x)$ is the Goldstein-ε-subdifferential and $N(x)$ is a closed convex cone at x. It is proved that negative of the shortest ε-conditional subgradient provides a descent direction in $T(x)$, which denotes the polar of $N(x)$. The ε-conditional subdifferential at an iterate x_ℓ can be approximated by a convex hull of a finite set of projected gradients at sampling points in $x_\ell + \varepsilon_\ell B_{T(x_\ell)}(0,1)$ to $T(x_\ell)$, where $T(x_\ell)$ is a linear space in the Bouligand tangent cone and $B_{T(x_\ell)}(0,1)$ denotes the unit ball in $T(x_\ell)$. The negative of the shortest vector in this convex hull is shown to be a descent direction in the Bouligand tangent cone at x_ℓ. The proposed algorithm makes a step along this descent direction with a certain step-size rule, followed by a retraction to lift back to points on the algebraic variety $\mathcal{M}$. The convergence of the resulting algorithm to a critical point is proved. For numerical illustration, the considered method is applied to some nonsmooth problems on varieties of low-rank matrices $\mathcal{M}_{\leq r}$ of real $M \times N$ matrices of rank at most r, specifically robust low-rank matrix approximation and recovery in the presence of outliers.

1. Introduction

This paper is concerned with the numerical solution of nonsmooth optimization problems on real algebraic varieties. The method proposed in this work generalizes the gradient sampling method for Riemannian manifolds to problems on such sets. Our motivation comes from applications in low-rank matrix and tensor optimization, where one is faced with the fact that smooth manifolds of fixed rank, say, manifolds of rank-r matrices, are not closed, and hence convergence of Riemannian algorithms is difficult to establish even for smooth functions [1, 15, 17, 20, 21].

As most iterative methods for finding an optimizer are based on the idea of a sequential descent of the cost function based on local information, the development of nonlinear optimization algorithms has always been intimately related to the understanding of the geometric properties of the constraints and the objective function. In a nondifferentiable problem on a constraint set, projection of the negative gradient of the cost function in a point on the tangent cone generally cannot be used to determine a direction along which the function is decreasing. Therefore, one has to work with some replacements for the gradient, called subdifferentials.

In this paper, we consider the general problem

$$\min_{x \in \mathcal{M}} f(x), \tag{1.1}$$

*Hausdorff Center for Mathematics & Institute for Numerical Simulation, University of Bonn, 53115 Bonn, Germany (hosseini@ins.uni-bonn.de, uschmajew@ins.uni-bonn.de).

2010 *Mathematics Subject Classification.* 49J52, 65K05, 14P05,15A99.

Key words and phrases. Lipschitz function, Descent direction, Clarke subdifferential, Algebraic varieties, Riemannian manifolds, Robust low-rank matrix recovery.

where $f : \mathbb{R}^n \to \mathbb{R}$ is a *locally Lipschitz function* and $\mathcal{M} \subseteq \mathbb{R}^n$ is a *closed* real algebraic variety. As such, $\mathcal{M}$ admits a so-called Whitney stratification [23], that is,

$$\mathcal{M} = \bigcup_{s=0}^{r} \mathcal{M}_s, \tag{1.2}$$

where $\mathcal{M}_s$ are mutually disjoint smooth submanifolds of $\mathbb{R}^n$ (not necessarily connected) of different dimensions. The manifolds $\mathcal{M}_s$ will be called strata of $\mathcal{M}$. We assume $\dim(\mathcal{M}_s) < \dim(\mathcal{M}_r)$ for $s < r$. While we will consider real algebraic varieties in the sequel, we note that many subsequent considerations only use that $\mathcal{M}$ is closed and a disjoint union of smooth (C^∞) manifolds. However, a potential exception is Lemma 3.2, which makes use of the additional Whitney a-regularity condition regarding limits of tangent planes (see [22, Sec. 8] and [23, Sec. 19]) and enters crucially into the main convergence result (Theorem 3.3) via Theorem 3.8. Specifically, a-regularity means that whenever a sequence $(x_\ell) \subset \mathcal{M}_r$ converges to $\bar{x} \in \mathcal{M}_s$, where $\dim(\mathcal{M}_r) > \dim(\mathcal{M}_s)$, and the tangent spaces $T_{\mathcal{M}_r}(x_\ell)$ converge to some subspace T (in the usual sense, say, in the sense of orthogonal projections), then it should hold $T_{\mathcal{M}_s}(\bar{x}) \subseteq T$.

We aim to present an algorithm to solve problem (1.1) by generalizing the Riemannian gradient sampling method from [12]. In general, given a closed convex cone $T(x)$ at x, we can define

$$\partial_\varepsilon^N f(x) := \partial_\varepsilon f(x) + N(x), \tag{1.3}$$

where $\partial_\varepsilon f(x)$ is the Goldstein-ε-subdifferential, and $N(x) = (T(x))^\circ$ is the polar cone of $T(x)$. Such a set $\partial_\varepsilon^N f(x)$ in (1.3) is called an ε-conditional subdifferential, and every $\xi \in \partial_\varepsilon^N f(x)$ is called an ε-conditional subgradient. We shall prove that the shortest ε-conditional subgradient proposes a descent direction in the cone $T(x)$ (Theorem 2.3), which assigns an essential role to the ε-conditional subdifferential for deriving optimization algorithms for (1.1). In practice, however, $\partial_\varepsilon^N f(x)$ might be unavailable in closed form and has to be approximated using some of its elements. One possibility is based on random gradient sampling as in the GS algorithm [4]. In order to design such a method for the constraint setting at hand, we assume that $T(x)$ is actually a *linear* space. The gradient sampling algorithm as outlined in Sec. 3 approximates the ε-conditional subdifferential corresponding to $N(x) = (T(x))^\circ$ by a convex hull of vectors, which are obtained from projecting gradients at sample points in $x + \varepsilon B_{T(x)}(0, 1)$ to $T(x)$, where $B_{T(x)}(0, 1)$ denotes the unit ball in $T(x)$. The negative of the shortest vector in this convex hull is shown to be a descent direction.

Since we are dealing with constraint optimization, we are of course interested in descent directions in the Bouligand tangent cone $T_{\mathcal{M}}^B(x)$. Therefore we have to choose the linear space $T(x)$ as a subset of the Bouligand cone:

$$T(x) \subseteq T_{\mathcal{M}}^B(x). \tag{1.4}$$

A new iterate is then obtained by making a step along this direction with a certain step-size rule, followed by a retraction to get back on $\mathcal{M}$.

Our restriction to a closed real algebraic variety $\mathcal{M}$ generally ensures

(i) the existence of an a-regular stratification (1.2), see the original proof [23, Sec. 19], or [13] and references therein;
(ii) the existence of linear subspaces in the Bouligand cones (in particular, tangent spaces $T_{\mathcal{M}_s}(x)$ of strata), and
(iii) the existence of retractions in the sense of Sec. 3.1.2. In particular, the metric projection onto $\mathcal{M}$ will have the desired properties.

Of course, instead of restricting to real algebraic varieties, one may include these three properties into a list of assumptions for general closed sets $\mathcal{M} \subseteq \mathbb{R}^n$.

As a main result, we prove that if the subspaces $T(x_\ell)$ are chosen such that they contain the tangent spaces $T_{\mathcal{M}_{s_\ell}}(x_\ell)$ of the current strata, then a cluster point $x \in \mathcal{M}_s$ will be critical point of f on $\mathcal{M}_s$ in the sense $0 \in \partial f(x) + (T_{\mathcal{M}_s}(x))^\perp$ (see Theorem 3.3). Formally, the result of the theorem is in fact a little stronger, namely $0 \in \partial f(x) + N(x)$, where $N(x)$ is in general only a subspace of $(T_{\mathcal{M}_s}(x))^\perp$, which is obtained as a limit of normal spaces $N(x_\ell)$. Hence, if we converge to x from strata with $\dim(\mathcal{M}_r) > \dim(\mathcal{M}_s)$, the dimension of $N(x)$ will be smaller than dimension of $(T_{\mathcal{M}_s}(x))^\perp$. Our motivation for this general setup are varieties of low-rank matrices. The Bouligand tangent cone to varieties of low-rank matrices contains many reasonable subspaces $T(x) \supsetneq T_{\mathcal{M}_s}(x)$ obeying (1.4) in rank-deficient points.

We note that even when applied to submanifolds of Euclidean spaces, the new method is conceptually and technically considerably simpler than the algorithm in [12], as it uses only gradients from nearby points within the tangent plane at the current iterate, and projects them to the tangent space of that point. Therefore, no vector transport is required. It is also worth mentioning that in [12] we worked on manifolds whose injectivity radius is bounded below, while we relax also this assumption in this paper.

As applications of our method, we consider minimizing nonsmooth functions on real algebraic varieties of low-rank matrices

$$\min_{X \in \mathcal{M}_{\leq r}} f(X), \quad \mathcal{M}_{\leq r} \coloneqq \{X \in \mathbb{R}^{M \times N} : \operatorname{rank}(X) \leq r\}, \quad r \leq \min(M, N). \tag{1.5}$$

Specifically, in Sec. 4 we conduct experiments in which we use the GS method for some problems of robust recovery of low-rank matrices in the presence of outliers. The variety $\mathcal{M}_{\leq r}$ naturally stratifies by rank into smooth components

$$\mathcal{M}_s \coloneqq \{X \in \mathbb{R}^{M \times N} : \operatorname{rank}(X) = s\}$$

of fixed rank $s \leq r$. The Riemannian optimization algorithms in [9, 10, 12] are applicable to that manifold in practice, but the theory are developed for *complete* Riemannian manifolds and does not apply due to the nonclosedness of $\mathcal{M}_s$. This affects the existence of retractions, the free choice of step-sizes in the tangent space, and the convergence results (existence of cluster points).

When applied to problems on $\mathcal{M}_{\leq r}$, the newly proposed method can be seen as an extension of the gradient sampling method from [12] for the manifold $\mathcal{M}_r$ to its closure. It has the advantage that it provides rigorous convergence results even in the case that a rank-deficient point is never encountered; cf. the similar remarks in [17] for smooth functions. Also it provides a sound framework for the derivation of rank-increasing methods, which make it necessary to consider rank-deficient starting guesses, obtained, say, as a "solution" of (1.5) for rank $r-1$, for (1.5). Such rank-increasing strategies have shown superior performance for solving (1.5), e.g., for matrix completion [20]. In Sec. 4, we use them for reconstruction of scratched grayscale images.

Besides its practical relevance, the set $\mathcal{M}_{\leq r}$ is an interesting example for our framework because the Euclidean metric projection is explicitly available via singular value decomposition, which is somewhat exceptional for such a nontrivial set. When $s < r$, the Bouligand cone $T_{\mathcal{M}_{\leq r}}(X)$ also contains the matrices in the orthogonal complement of $T_{\mathcal{M}_s}(X)$ which are of rank at most $r - s$; see [6, 17]. In such points, it is reasonable to run the minimization algorithm on $\mathcal{M}_s$ to get a local minimizer on $\mathcal{M}_s$ and afterward to use a linear space of tangent vectors orthogonal to $T_{\mathcal{M}_s}(X)$ to increase the rank and move to a manifold with a higher dimension. In the rank-increasing step, we may consider different linear subspaces of the Bouligand tangent cone. In our experiments, we consider random subspace augmentation and the subspaces spanned by the dominant $r - s$ left and right singular vectors of the orthogonal projection of $\nabla f(X)$ on $(T_{\mathcal{M}_s}(X))^\perp$; (cf. Sec. 4.1.4). Unfortunately, this remains an heuristic as there is no guarantee that this spaces will contain a descent direction in the case that the Bouligand cone contains a descent direction (which we do not know how to compute due to the nonconvexity of the Bouligand cone). Nevertheless, the strategy turned out to be useful for the rank-increasing strategy.

Outline. The paper is organized as follows: Section 2 is concerned with some preliminaries and definitions of nonsmooth analysis. Section 3 is devoted to the gradient sampling algorithm on real algebraic varieties and convergence results of the algorithm. Finally, in Section 4 some numerical experiments are illustrated.

2. Prerequisites

We consider the space $\mathbb{R}^n$ equipped with a fixed Euclidean norm $\|\cdot\|$, generated by an inner product $\langle\cdot,\cdot\rangle$. By $B(x,\varepsilon)$ we denote the open set $\{y \in \mathbb{R}^n : \|y-x\| < \varepsilon\}$. The super-script $^\perp$ indicates orthogonal complements with respect to this inner product. For a closed set $\mathcal{M} \in \mathbb{R}^n$, let $P_{\mathcal{M}}(y) = \operatorname{argmin}_{x\in\mathcal{M}} \|x-y\|$ denote the metric projection on $\mathcal{M}$. If $\mathcal{M}$ is convex, then $y \mapsto P_{\mathcal{M}}(y)$ is single-valued and continuous. We denote by $\operatorname{cl} N$ and $\operatorname{conv} N$ the closure and the convex hull of a set N.

2.1. Unconstrained optimization.

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a locally Lipschitz function and $L = L(x)$ be its Lipschitz constant around x. We first recall basic concepts of unconstrained optimization

$$\min_{x\in\mathbb{R}^n} f(x), \tag{2.1}$$

for such a function. The Clarke generalized directional derivative of f at x in the direction ξ is defined as

$$f^\circ(x;\xi) \coloneqq \limsup_{\substack{y\to x\\ t\downarrow 0}} \frac{f(y+t\xi)-f(y)}{t}.$$

Moreover, the Clarke subdifferential is defined as follows:

$$\partial f(x) \coloneqq \{v \in \mathbb{R}^n : f^\circ(x;\xi) \geq \langle v,\xi\rangle \text{ for all } \xi \in \mathbb{R}^n\}.$$

This set is closed and convex. It is also bounded since we have

$$\|v\| \leq L \quad \text{for all } v \in \partial f(x). \tag{2.2}$$

Moreover,

$$f^\circ(x;\xi) = \sup_{v\in\partial f(x)} \langle v,\xi\rangle.$$

If f is differentiable at x, then $\nabla f(x) \in \partial f(x)$. Furthermore, if f is continuously differentiable at x, then it holds that

$$\partial f(x) = \{\nabla f(x)\}.$$

In general, letting Ω_f denote the set of points on which f is differentiable (which is dense in $\mathbb{R}^n$, see [7]), we have the characterization

$$\partial f(x) \coloneqq \operatorname{conv}\{v \in \mathbb{R}^n : \text{there exists } (x_i) \subset \Omega_f \text{ s.t. } x_i \to x \text{ and } \nabla f(x_i) \to v\}.$$

The unconstrained necessary optimality condition in the sense of Clarke is $0 \in \partial f(x)$, and holds in local minima and maxima of f. We refer to [7] for proofs of all these properties.

A vector $g = g(x) \in \mathbb{R}^n$ is called a descent direction for f at x, if there exists $\alpha > 0$ such that

$$f(x+tg) - f(x) < 0 \quad \text{for all } t \in (0,\alpha).$$

The extension of the steepest descent method for smooth optimization to (2.1) uses in every step the search direction $g(x) \coloneqq -\operatorname{argmin}\{\|v\| : v \in \partial f(x)\}$ in combination with a step-size rule. But since $\partial f(\cdot)$ is not continuous, such extension can fail to be convergent to critical points for locally Lipschitz functions. To obtain a powerful convergence property, it is necessary to enlarge the set

$\partial f(x)$; see [2]. An adequate replacement is the ε-subdifferential $\partial_\varepsilon f(x)$; see [8], which for $\varepsilon > 0$ is defined by

$$\partial_\varepsilon f(x) \coloneqq \operatorname{conv}\{v \in \mathbb{R}^n : v \in \partial f(y), y \in \operatorname{cl} B(x, \varepsilon)\}.$$

If $0 \in \partial_\varepsilon f(x)$, then x is said to be an ε-critical point. Note that $\partial_\varepsilon f(x)$ is closed. Correspondingly, one defines

$$f_\varepsilon^\circ(x; \xi) \coloneqq \sup_{v \in \partial_\varepsilon f(x)} \langle v, \xi \rangle.$$

Obviously, it holds that $f^\circ(x; \xi) \leq f_\varepsilon^\circ(x; \xi)$. Let $g \in \mathbb{R}^n$ and $\|g\| \leq 1$, by Lebourg's Mean Value Theorem [7], there exist $\theta \in (0, 1)$ and $v \in \partial f(x + t\theta g)$ for all $t \in (0, \varepsilon]$, such that

$$f(x + tg) - f(x) = t\langle v, g \rangle \leq t f_\varepsilon^\circ(x; g).$$

According to this inequality, a descent direction g is found when $f_\varepsilon^\circ(x; g)$ is negative. For the largest descent guarantee one has to solve

$$\min_{\|g\| \leq 1} f_\varepsilon^\circ(x; g) = \min_{\|g\| \leq 1} \max_{v \in \partial_\varepsilon f(x)} \langle v, g \rangle. \tag{2.3}$$

This problem has a solution, which can be computed by solving the problem,

$$- \min_{v \in \partial_\varepsilon f(x)} \|v\|. \tag{2.4}$$

If v^* is the solution of (2.4), then $g = -\frac{v^*}{\|v^*\|}$ is the solution of (2.3), and we have

$$f_\varepsilon^\circ(x; g) = - \|v^*\|,$$

see, e.g., [2].

2.2. **Constrained optimization.** In this paper we are concerned with constrained optimization problems. Here and in the following, let $f : \mathbb{R}^n \to \mathbb{R}$ be a locally Lipschitz function, and $\mathcal{M}$ be closed. We consider the minimization problem

$$\min_{x \in \mathcal{M}} f(x), \tag{2.5}$$

and assume that it has at least one solution.

The Bouligand tangent cone, also called contingent cone, to $\mathcal{M}$ at x is defined as

$$T_{\mathcal{M}}^B(x) = \{\xi \in \mathbb{R}^n : \text{there exist } (x_i) \subset \mathcal{M} \text{ and } (t_i) \subset \mathbb{R} \text{ such that } x_i \to x,\ t_i \downarrow 0 \text{ and } \frac{x_i - x}{t_i} \to \xi\}.$$

This cone is closed, but in general not convex, which makes it difficult to use for nonsmooth optimization. In contrast, the Clarke tangent cone to $\mathcal{M}$ at x, defined by

$$T_{\mathcal{M}}^C(x) \coloneqq \{\xi \in \mathbb{R}^n : \text{for all } (x_i) \subset \mathcal{M} \text{ with } x_i \to x, \text{ and all } t_i \downarrow 0 \text{ there exist } (\xi_i) \subset \mathbb{R}^n \text{ such that } x_i + t_i\xi_i \in \mathcal{M} \text{ for all } i \text{ and } \xi_i \to \xi\},$$

is closed and convex [7]. It holds that $T_{\mathcal{M}}^C(x) \subseteq T_{\mathcal{M}}^B(x)$; see [7]. When $\mathcal{M}$ is a smooth submanifold in a neighborhood of x, then both cones coincide with the tangent space to $\mathcal{M}$ at x.

The Clarke normal cone is the polar of the Clarke tangent cone:

$$N_{\mathcal{M}}^C(x) \coloneqq \left(T_{\mathcal{M}}^C(x)\right)^\circ = \{y \in \mathbb{R}^n : \langle y, \xi \rangle \leq 0 \text{ for all } \xi \in T_{\mathcal{M}}^C(x)\}.$$

It is also closed and convex. A point $x \in \mathcal{M}$ is critical point for (2.5) in the sense of Clarke, if

$$0 \in \partial f(x) + N_{\mathcal{M}}^C(x).$$

Every local minimum on $\mathcal{M}$ satisfies this.

2.2.1. *Existence of descent directions.* We shall prove here a rather general result on the existence of descent directions.

Assume that $x \in \mathcal{M}$ and a closed convex cone $T(x)$ are given. Let $N(x) = (T(x))^\circ$ denote its polar cone. For example, the choice $T(x) = T^C_{\mathcal{M}}(x)$ is feasible, but $T(x) = T^B_{\mathcal{M}}(x)$ may be not feasible (due to nonconvexity). Let us say that x is *critical* with respect to $N(x)$, if

$$0 \in \partial f(x) + N(x). \tag{2.6}$$

We denote

$$\partial_N f(x) := \partial f(x) + N(x),$$

and call $\partial_N f(x)$ the *conditional subdifferential.*

Proceeding now as in the unconstrained case, for each $\varepsilon \geq 0$, the *conditional ε-subdifferential* is defined as

$$\partial^\varepsilon_N f(x) := \partial_\varepsilon f(x) + N(x).$$

If $x \in \mathcal{M}$ satisfies the weaker condition

$$0 \in \partial^\varepsilon_N f(x),$$

then x is said to be an *ε-critical point* with respect to $N(x)$. Note that for $\varepsilon = 0$ we recover (2.6).

We aim to show that if x is not critical with respect to N, that is,

$$0 \notin \partial f(x) + N(x),$$

then there exists a descent direction in $T(x)$.

The first statement is that if x is not critical, then there exists $\varepsilon > 0$ such that x is not ε-critical.

Proposition 2.1. *Let $x \in \mathcal{M}$ such that $0 \notin \partial_N f(x)$. Then there exists $\varepsilon > 0$ such that $0 \notin \partial^\varepsilon_N f(x)$.*

Proof. Suppose to the contrary that $0 \in \partial^{1/i}_N f(x)$ for all i, that is, there exists $w_i \in \partial_{1/i} f(x) \cap -N(x)$. Since w_i is a bounded sequence by (2.2), it has a convergent subsequence to some point w. Note that $w_i \in \partial_{1/j} f(x)$ for $i \geq j$. Since these sets are closed, it follows that $w \in \bigcap_{j=1}^\infty \partial_{1/j} f(x) = \partial f(x)$. As the normal cone $N(x)$ is also closed, we obtain $w \in \partial f(x) \cap -N(x)$, that is, $0 \in \partial_N f(x)$ in contradiction to the made assumption. □

The next lemma relates the minimum norm element in $\partial^\varepsilon_N f(x)$ to projections on $-T(x)$.

Lemma 2.2. *Let $T \subseteq \mathbb{R}^n$ be a closed convex cone and $v^* \in \mathbb{R}^n$. Then*

$$\operatorname{argmin}\{\|w\| : w \in v^* + T^\circ\} = \operatorname{argmin}\{\|\xi - v^*\| : \xi \in -T\} = P_{-T}(v^*).$$

Proof. For a closed convex cone, it is known (and easy to see) that $P_{-T} + P_{-T^\circ} = \mathrm{Id}$. Therefore, it holds

$$\begin{aligned}\operatorname{argmin}\{\|w\|^2 : w \in v^* + T^\circ\} &= v^* + \operatorname{argmin}\{\|v^* + \eta\|^2 : \eta \in T^\circ\} \\ &= v^* - \operatorname{argmin}\{\|v^* - \eta\|^2 : \eta \in -T^\circ\} \\ &= v^* - P_{-T^\circ}(v^*) = P_{-T}(v^*),\end{aligned}$$

which is the assertion. □

Consider now the situation $0 \notin \partial^\varepsilon_N f(x)$. Then the minimizer w^* of the problem

$$\min_{w \in \partial^\varepsilon_N f(x)} \|w\| \tag{2.7}$$

is nonzero. It is also unique, since the norm $\|\cdot\|$ is assumed strictly convex. From Lemma 2.2 with $T = T(x)$ it follows that actually $w^* \in -T(x)$, and specifically

$$w^* = P_{-T(x)}(v^*), \quad \text{where} \quad v^* = \operatorname{argmin}\{\|P_{-T(x)}(v)\| : v \in \partial_\varepsilon f(x)\}. \tag{2.8}$$

The main result of this section is that $-w^*$ provides a descent direction on $T(x)$.

Indeed, similar to (2.3), a natural approach to seek a descent direction in $T(x)$ is to consider the problem

$$\begin{aligned} \min_{\|g\|\le 1, g\in T(x)} f_\varepsilon^\circ(x;g) &= \min_{\|g\|\le 1, g\in T(x)} \max_{v\in\partial_\varepsilon f(x)} \langle v, g\rangle \\ &= \max_{v\in\partial_\varepsilon f(x)} \min_{\|g\|\le 1, g\in T(x)} \langle v, g\rangle \\ &= \max_{v\in\partial_\varepsilon f(x)} (- \max_{\|g\|\le 1, g\in T(x)} \langle -v, g\rangle) \\ &= - \min_{v\in\partial_\varepsilon f(x)} \|P_{-T(x)}(v)\|. \end{aligned} \tag{2.9}$$

Here, the first equality is obtained by the minimax theorem and the last equality is obtained by [17, Eq. (2.4)]. It is clear that the common value of (2.9) is negative. The theorem below generalizes the equivalence of (2.3) and (2.4) to the constrained case.

Theorem 2.3. *Consider $x \in \mathcal{M}$ such that $0 \notin \partial_N^\varepsilon f(x)$. Let w^* be the solution of (2.7). Then $w^* \in -T(x)$, $w^* \neq 0$, and for*

$$g = -\frac{w^*}{\|w^*\|} \in T(x)$$

it holds that

$$f_\varepsilon^\circ(x;g) = -\|w^*\| < 0,$$

that is, g is a descent direction.

Proof. From $0 \notin \partial_N^\varepsilon f(x)$ it follows $w^* \neq 0$. We have the variational inequality

$$\langle w^*, w^*\rangle \le \langle w^*, w\rangle \quad \text{for all } w \in \partial_N^\varepsilon f(x).$$

In particular, for every $v \in \partial_\varepsilon f(x)$ it holds

$$\langle w^*, w^*\rangle \le \langle w^*, v\rangle,$$

which implies

$$\max_{v\in\partial_\varepsilon f(x)} \langle -w^*, v\rangle \le \langle -w^*, w^*\rangle.$$

We conclude that

$$f_\varepsilon^\circ(x;g) \le -\|w^*\| < 0.$$

As stated in (2.8) it holds

$$w^* = P_{-T(x)}(v^*),$$

where v^* solves the last problem in (2.9). From this we get the reverse relation

$$f_\varepsilon^\circ(x;g) = \sup_{v\in\partial_\varepsilon f(x)} \langle v, g\rangle \ge \left\langle v^*, -\frac{P_{-T(x)}(v^*)}{\|P_{-T(x)}(v^*)\|}\right\rangle = -\left\|P_{-T(x)}(v^*)\right\| = -\|w^*\|,$$

where the second last equality holds because $-T(x)$ is a cone. □

3. A gradient sampling algorithm

In this section, the proposed gradient sampling algorithm is presented together with a suitable convergence result. We first list the properties that a general closed set $\mathcal{M} \subseteq \mathbb{R}^n$ must have in order to define the algorithm, and emphasize that they are satisfied for closed real algebraic varieties.

3.1. **Assumptions for the minimization algorithm.**

Recall that we are considering the minimization problem

$$\min_{x\in\mathcal{M}} f(x).$$

The following assumptions on f and $\mathcal{M}$ are required to formulate the gradient sampling algorithm in the next subsection. We highlight that the assumptions on $\mathcal{M}$ are satisfied if $\mathcal{M}$ is a closed real algebraic variety.

3.1.1. *Existence of linear spaces in Bouligand tangent cones.* In what follows, we assume, for every $x \in \mathcal{M}$, a linear space $T(x)$ contained in the Bouligand tangent cone exists, that is (repeating (1.4)),

$$T(x) \subseteq T^B_{\mathcal{M}}(x).$$

When $\mathcal{M}$ is a real algebraic variety and $x \in \mathcal{M}_s$, then due to (1.2) there is at least one possible choice, namely $T(x) = T_{\mathcal{M}_s}(x)$.

3.1.2. *Existence of retractions.* Following [17], a map $R : \bigcup_{x\in\mathcal{M}}\{x\} \times T(x) \to \mathcal{M}$ will be called a *retraction* if for any fixed $x \in \mathcal{M}$ and $\xi \in T(x)$, we have

$$\lim_{t\downarrow 0} \frac{R_x(t\xi) - (x + t\xi)}{t} = 0. \tag{3.1}$$

When talking about retractions, we silently assume that there exists a constant $\kappa > 0$ such that

$$\|R_x(\xi) - x\| \le \kappa\|\xi\| \tag{3.2}$$

for all $x \in \mathcal{M}$ and $\xi \in T(x)$.

For closed real algebraic varieties, any metric projection $P_{\mathcal{M}} : \mathbb{R}^n \to \mathcal{M}$, $P_{\mathcal{M}}(y) \in \operatorname{argmin}_{x\in\mathcal{M}} \|x-y\|$ defines the retraction

$$R_x(\xi) = P_{\mathcal{M}}(x + \xi).$$

This can be seen from the fact that every tangent vector to a real algebraic variety is tangent to some analytic arc $\gamma_{x,\xi}(t) = x + t^p\xi + O(t^{p+1})$ with $p > 0$ and $\gamma_{x,\xi}(t) \in \mathcal{M}$ for small t; see [16, Proposition 2]. Hence

$$\frac{\|R_x(t\xi) - (x + t\xi)\|}{t} \le \frac{\|\gamma_{x,\xi}(t^{1/p}) - (x + t\xi)\|}{t} \to 0$$

for $t \downarrow 0$. (Recall that $T(x) \subseteq T^B_{\mathcal{M}}(x)$.) Further, (3.2) is satisfied with $\kappa = 2$.

Remark 3.1. Let R be a retraction on $\mathcal{M}$. Using Theorem 2.3 with $T(x) \subseteq T^B_{\mathcal{M}}(x)$, we can prove that there exists $\alpha > 0$ such that

$$f(R_x(tg)) - f(x) \le -t\frac{\|w^*\|}{2} \quad \text{for all } t \in (0, \alpha).$$

Indeed, we have

$$\begin{aligned} f(R_x(tg)) - f(x) &\le f(x + tg) - f(x) + L\|R_x(tg) - (x + tg)\| \\ &\le f^\circ(x; g) \cdot t + o(t) + L\|R_x(tg) - (x + tg)\| \\ &\le f^\circ_\varepsilon(x; g) \cdot t + o(t) + L\|R_x(tg) - (x + tg)\| \\ &= f^\circ_\varepsilon(x; g) \cdot t + o(t). \end{aligned}$$

We obtain

$$\frac{f(R_x(tg)) - f(x)}{t} \le f^\circ_\varepsilon(x; g) + \frac{1}{2}|f^\circ_\varepsilon(x; g)| = -\frac{\|w^*\|}{2}$$

for t small enough.

3.1.3. *Continuously differentiability on a set of full measure.* Algorithm 1 below will feature a subset $D \subseteq \Omega_f$ with the following properties: $D \subseteq \Omega_f$ is an *open* set in $\mathbb{R}^n$ of full measure on which f is *continuously differentiable.* Furthermore, $\mathcal{M} \cap D$ is an open set of full measure in $\mathcal{M}$ (w.r.t. to the induced topology).

Let us comment on these assumptions already here.

- The assumption that D is an open set of full measure is made to ensure that the termination in line 3 has zero probability.
- The assumption that $\mathcal{M} \cap D$ is an open set of full measure in $\mathcal{M}$ ensures that the adjustment step in line 18 is possible. The following procedure could be applied (in theory; it is not expected to be ever necessary in practice): if $R_{x_\ell}(t_\ell g_\ell) \notin D$, one continues choosing $x_{\ell+1}$ uniformly at random from $\mathcal{M} \cap B(R_{x_\ell}(t_\ell g_\ell), \kappa t_\ell/k)$, $k = 1, 2 \ldots$, until $x_{\ell+1} \in D$ and $f(x_{\ell+1}) - f(x_\ell) < -\beta t_\ell \|w_\ell\|$ as desired. By continuity of f and the inequality $f(R_{x_\ell}(tg_\ell)) - f(x_l) < -\beta t_\ell \|w_\ell\|$, this process will terminate after finitely many steps with probability one. Of course, this requires that one is able to construct random points on $\mathcal{M}$.
- Finally, the assumption that f is continuously differentiable on D will be crucial for the convergence proof.

In many cases of interest, one can reasonably expect that $D = \Omega_f$ satisfies these assumptions.

3.2. **A minimization algorithm.** Theorem 2.3 and Remark 3.1 suggest a descent algorithm using descent directions obtained from (2.7) combined with a line-search. In every step, it requires to find the shortest element in $\partial_N^\varepsilon f(x)$. However, since in many applications an explicit description of $\partial_N^\varepsilon f(x)$ will not be available, an approximation has to be used. Our algorithm adopts the reasoning in [4] to the constrained optimization problem at hand by replacing $\partial_N^\varepsilon f(x_\ell)$ with G_ℓ at iteration ℓ, where G_ℓ is convex hull of a finite set of projected gradients at sampling points in $x_\ell + \varepsilon_\ell B_{T(x_\ell)}(0,1)$ to $T(x_\ell)$, where $T(x_\ell)$ denotes a linear space in the Bouligand tangent cone and $B_{T(x_\ell)}(0,1)$ is the unit ball in $T(x_\ell)$.

The resulting minimization algorithm is given as Algorithm 1.

We remark that the line search in line 14 of the algorithm is well-define and t_ℓ can be found using a finite process. To see this, observe that for $w_\ell = \operatorname{argmin}\{\|w\| : w \in G_\ell\}$ we have

$$\langle P_{T(x_\ell)}(\nabla f(x_\ell)), g_\ell\rangle \leq \sup_{w \in G_\ell} \langle w, g_\ell\rangle \leq -\|w_\ell\|,$$

where $g_\ell = \frac{-w_\ell}{\|w_\ell\|}$. By (3.1), $t \mapsto R_{x_\ell}(tg_\ell)$ has the right derivative g_ℓ at zero. Then, since $x_\ell \in D$, the function $\varphi(t) = f(R_{x_\ell}(tg_\ell))$ has the right derivative $\varphi'_+(0) = \langle \nabla f(x_\ell), g_\ell\rangle = \langle P_{T(x_\ell)}(\nabla f(x_\ell)), g_\ell\rangle < 0$. Therefore, since $\beta < 1$, there exists $\alpha > 0$ such that for all $t \in (0, \alpha)$ we have

$$f(R_{x_\ell}(tg_\ell)) - f(x_\ell) = \varphi(t) - \varphi(0) < t\beta\langle \nabla f(x_\ell), g_\ell\rangle \leq -t\beta \|w_\ell\|.$$

3.3. **Convergence result.** We begin with a lemma regarding possible limiting subspaces of $T(x_\ell)$. For the second part it is essential that the stratification (1.2) of $\mathcal{M}$ is a-regular.

Lemma 3.2. *Let $\mathcal{M}$ be a real algebraic variety with an a-regular stratification* (1.2) *by dimension. Assume that the sequence x_ℓ has a cluster point $\bar{x}$. Let*

$$\tilde{m} = \limsup_{\rho \to 0} \left(\max\{\dim(T(x_\ell)) : \|x_\ell - \bar{x}\| \leq \rho\}\right).$$

Then there exists a linear space S and an infinite subsequence $(x_\ell)_{\ell \in L}$ such that the following conditions hold:

(i) $\dim T(x_\ell) = \dim S = \tilde{m}$ *for all* $\ell \in L$,

(ii) $P_{T(x_\ell)} \to P_S$ *for* $\ell \in L$, $\ell \to \infty$.

Algorithm 1: Gradient sampling algorithm

```
Input: $x_0 \in \mathcal{M} \cap D$; $\delta_0, \varepsilon_0, \gamma, \varepsilon_{\mathrm{opt}}, \delta_{\mathrm{opt}} \in (0,1)$; $\beta \in (0,1)$; $\theta_\varepsilon, \theta_\delta \in (0,1]$.
for $\ell = 0, 1, 2, \dots$ do
    Choose $m_\ell = \dim(T(x_\ell)) + 1$ points $\{x_\ell^i\}_{i=1}^{m_\ell}$ independently and uniformly from
    $x_\ell + \varepsilon_\ell B_{T(x_\ell)}(0,1)$.                                // gradient sampling
    if $\{x_\ell^i\}_{i=1}^{m_\ell} \not\subset D$ then
        return
    end
    Let $G_\ell \coloneqq \mathrm{conv}\{P_{T(x_\ell)}(\nabla f(x_\ell)), P_{T(x_\ell)}(\nabla f(x_\ell^1)), \dots, P_{T(x_\ell)}(\nabla f(x_\ell^{m_\ell}))\}$, and find
        $w_\ell = \mathrm{argmin}\{\|w\| : w \in G_\ell\}$
    if $\|w_\ell\| \leq \delta_{opt}$ and $\varepsilon_\ell \leq \varepsilon_{opt}$ then
        return
    end
    if $\|w_\ell\| \leq \delta_\ell$ then
        $\varepsilon_{\ell+1} \coloneqq \theta_\varepsilon \varepsilon_\ell$, $\delta_{\ell+1} \coloneqq \theta_\delta \delta_\ell$
        $x_{\ell+1} \coloneqq x_\ell$
    else
        $\varepsilon_{\ell+1} = \varepsilon_\ell$, $\delta_{\ell+1} = \delta_\ell$, $g_\ell \coloneqq -\frac{w_\ell}{\|w_\ell\|}$      // descent direction
        $t_\ell \coloneqq \max\{t : f(R_{x_\ell}(tg_\ell)) - f(x_l) < -\beta t \|w_\ell\|, t \in \{1, \gamma, \gamma^2, \dots\}\}$   // line search
        if $R_{x_\ell}(t_\ell g_\ell) \in D$ then
            $x_{\ell+1} \coloneqq R_{x_\ell}(t_\ell g_\ell)$
        else
            Find $x_{\ell+1} \in \mathcal{M} \cap D$ such that $f(x_{\ell+1}) - f(x_\ell) < -\beta t_\ell \|w_\ell\|$      // stay in D
            and $\|R_{x_\ell}(t_\ell g_\ell) - x_{\ell+1}\| \leq \kappa t_\ell$.                  // $\kappa$ from (3.2)
        end
    end
end
```

Furthermore, assume $x_\ell \in \mathcal{M}_{s_\ell}$ and $\bar{x} \in \mathcal{M}_s$. Then if $T(x_\ell)$ contains the tangent space $T_{\mathcal{M}_{s_\ell}}(x_\ell)$ for almost all $\ell \in L$, then S contains $T_{\mathcal{M}_s}(\bar{x})$.

Proof. Without loss of generality, $x_\ell \to \bar{x}$ and $\dim T(x_\ell) = \tilde{m}$ for all ℓ. The sequence of orthogonal projections $P_{T(x_\ell)}$ lies on the spectral unit sphere $\|P_{T(x_\ell)}\| = 1$, i.e., is bounded. Therefore, after eventually switching to a subsequence, we may assume that $P_{T(x_\ell)}$ converges to some P. It is easy to show that P is an orthogonal projection of same rank as well, so we set S to be the range of P. To prove the second part, we assume without loss of generality that s_ℓ is constant (but not necessarily equal to s), and that $T_{\mathcal{M}_{s_\ell}}(x_\ell)$ converges to a subspace Q (in the sense of projections; $\dim Q = \dim T_{\mathcal{M}_{s_\ell}}(x_\ell)$). By the Whitney condition (a) [22, Sec. 8], [23, Sec. 19], the range of Q contains $T_{\mathcal{M}_s}(\bar{x})$. On the other hand,

$$P_S P_Q = \lim_{\ell \to \infty} P_{T(x_\ell)} P_{T_{\mathcal{M}_{s_\ell}}(x_\ell)} = \lim_{\ell \to \infty} P_{T_{\mathcal{M}_{s_\ell}}(x_\ell)} = P_Q,$$

which proves $T_{\mathcal{M}_s}(\bar{x}) \subseteq Q \subseteq S$. □

We now turn to the convergence result for Algorithm 1 given our main assumptions.

Theorem 3.3. *Let (x_ℓ) be a generated sequence for parameters $\delta_{opt} = \varepsilon_{opt} = 0$ and $\theta_\varepsilon, \theta_\delta \in (0,1)$. With probability one the algorithm does not stop and we either have $f(x_\ell) \downarrow -\infty$, or $\delta_\ell \downarrow 0$, $\varepsilon_\ell \downarrow 0$. In the later case, if $\bar{x}$ is a cluster point and S a corresponding subspace as in Lemma 3.2, then*

$\bar{x} \in \mathcal{M}_s$ is a critical point of f on $\mathcal{M}_s$ in the sense that

$$0 \in \partial f(\bar{x}) + S^{\perp}.$$

In particular, when $x_\ell \in \mathcal{M}_{s_\ell}$ and $T_{\mathcal{M}_{s_\ell}}(x_\ell) \subseteq T(x_\ell)$ for all ℓ, then

$$0 \in \partial f(\bar{x}) + (T_{\mathcal{M}_s}(\bar{x}))^{\perp},$$

that is, $\bar{x}$ is critical on the submanifold $\mathcal{M}_s$.

Remark 3.4. The meaning of "with probability one" here is similar to previous results on GS algorithm [4, 14, 12]; see in particular [4, p. 757]. The random nature of the algorithm is in the selection of sampling points in every iteration. These points are sampled in line 2 from the unit ball in $T(x_\ell)$, which is isomoprhic to the unit ball in $\mathbb{R}^{m_\ell - 1}$. Let B_{m-1} denote the unit ball in $\mathbb{R}^{m-1}$. Then we can regard the tuple of sample points in iteration ℓ as an element of $B_{m_\ell - 1}^{m_\ell}$. Only finitely many values for $m_\ell = \dim T(x_\ell) + 1 \leq n+1$ are possible. We may imagine that an infinite sequence $\mathbf{x}^m \in (B_{m-1}^m)^\infty$ of sample point tuples has been generated for every possible dimension $m = 1, \dots, n+1$ *before* we run the algorithm, and that we then simply use these sample points in the algorithm whenever this subdimension occurs, for instance, $\{x_\ell^i\}_{i=1}^{m_\ell} = \mathbf{x}_\ell^{m_\ell}$. In this interpretation, the randomness gets "outside" of the algorithm. Now "with probability one" refers to the fact, that for every m and almost every realization $\mathbf{x}^m \in (B_{m-1}^m)^\infty$ (with respect to suitable measure on $(B_{m-1}^m)^\infty$), any infinite subsequence of $\mathbf{x}^m$ hits every positive measure subset of B_{m-1}^m infinitely often. This will be a crucial argument in the proof of Theorem 3.3.

The logic of the proof follows the arguments for unconstrained gradient sampling by Burke, Lewis and Overton [4] and [14] in general, and corresponding arguments for a recent generalization to Riemannian manifolds [12] in particular. Thus, we will refer to proofs in [12] for some similar steps. However, since the algorithm at hand requires no Riemannian gradients, no vector transports, and works for general real algebraic varieties, some nontrivial modifications of the arguments will be needed. We first state two observations originally used by Kiwiel [14]; see also [12] for a proof of the second one.

Lemma 3.5. *Assume that a nonempty compact convex set C in an Euclidean space does not contain zero. Then for every $\beta \in (0,1)$ there exists $\nu > 0$ such that if $u, v \in C$ and $\|u\| \leq \min\{\|w\| : w \in C\} + \nu$, we deduce that $\langle v, u\rangle > \beta \|u\|^2$.*

Lemma 3.6. *Let $(x_\ell)_{\ell \in \mathbb{N}}$ be a divergent sequence in a metric space, and let* dist *denote the metric. Then for every infinite convergent subsequence $(x_\ell)_{\ell \in \mathcal{L}}$, $\mathcal{L} \subset \mathbb{N}$, it holds $\sum_{\ell \in \mathcal{L}} \operatorname{dist}(x_\ell, x_{\ell+1}) = \infty$.*

Next, following [4], we define the sets

$$G_\varepsilon^S(x) := \operatorname{cl}\operatorname{conv}\{P_S(\nabla f(y)) : y \in (x + \varepsilon \operatorname{cl} B_S(0,1)) \cap D\},$$

where S is a linear subspace of $\mathbb{R}^n$. For every $\varepsilon, \nu > 0$ and $\bar{x} \in \mathcal{M}$, let further $m = \dim S + 1$ and

$$\rho_\varepsilon(\bar{x}) := \min\{\|w\| : w \in G_\varepsilon^S(\bar{x})\},$$

$$D_\varepsilon(x) := (x + \varepsilon \operatorname{cl} B_S(0,1)) \cap D, \ D_\varepsilon^m(x) := \prod_1^m D_\varepsilon(x),$$

and

$$V_\varepsilon(\bar{x}, x, \nu) := \{y = (y^1, \dots, y^m) \in D_\varepsilon^m(x) : \tilde{\rho}_\varepsilon(y) \leq \rho_\varepsilon(\bar{x}) + \nu\},$$

where

$$\tilde{\rho}_\varepsilon(y) := \min\{\|w\| : w \in \operatorname{conv}\{P_S(\nabla f(y^i))\}_{i=1}^m\}.$$

Lemma 3.7. *Let $\varepsilon > 0$, $\bar{x} \in \mathcal{M}$. For any $\nu > 0$, there exist $\tau > 0$ and a nonempty open set $\hat{V} = \hat{V}(\bar{x}, \varepsilon, \tau)$ such that $\operatorname{cl} \hat{V} \subseteq V_\varepsilon(\bar{x}, x, \nu)$ for all $x \in B(\bar{x}, \tau)$.*

Proof. Since $G^S_\varepsilon(\bar{x})$ is compact, there exists $w \in G^S_\varepsilon(\bar{x})$, such that $\rho_\varepsilon(\bar{x}) = \|w\|$. The argumentation now follows along similar lines as [12, Lemma 4.2]: using Carathéodory's Theorem and the continuity of $y \mapsto P_S(\nabla f(y))$ on D, we can find $\tilde{y} = (\tilde{y}^1, \dots, \tilde{y}^m) \in \prod_1^m (\bar{x} + \varepsilon B_S(0,1)) \cap D$ and non-negative $\lambda_1, \dots, \lambda_m$ with $\sum \lambda_i = 1$ such that $u := \sum_{i=1}^m \lambda_i P_S(\nabla f(\tilde{y}^i))$ satisfies $\|u\| \leq \|w\| + \nu/3 = \rho_\varepsilon(\bar{x}) + \nu/3$. Now choose $\bar{\varepsilon}$ such that

$$\tilde{V} := \prod_{i=1}^m (\tilde{y}^i + \bar{\varepsilon} B_S(0,1)) \subseteq D^m_{\varepsilon - \bar{\varepsilon}}(\bar{x}), \quad \text{and} \quad \left\| \sum_{i=1}^m \lambda_i P_S(\nabla f(y^i)) \right\| \leq \rho_\varepsilon(\bar{x}) + \nu \tag{3.3}$$

holds for all $y = (y^1, \dots, y^m) \in \tilde{V}$. Set $\tau := \bar{\varepsilon}$. Then, by (3.3), for all $x \in B(\bar{x}, \tau)$ we have $\tilde{V} \subseteq D_\varepsilon(x)$, and $\tilde{V} \subseteq V_\varepsilon(\bar{x}, x, \nu)$. Then we can choose any nonempty open subset $\hat{V}$ of $\tilde{V}$ such that $\operatorname{cl} \hat{V} \subset \tilde{V}$. □

Theorem 3.8. *Suppose (in a slight abuse of notation) that (x_ℓ) is a subsequence of iterates constructed by Alg. 1 with fixed $\varepsilon_\ell = \varepsilon_0 := \varepsilon$ such that x_ℓ converges to $\bar{x} \in \mathcal{M}$ and, furthermore, satisfies properties* (i) *and* (ii) *of Lemma 3.2 for some subspace S. Let $\nu > 0$ be taken from Lemma 3.5 for $C = G^S_\varepsilon(\bar{x})$ (and β from the algorithm), and τ and $\hat{V}$ be obtained from Lemma 3.7 for this ν. Assume further that $(x^1_\ell, \dots, x^m_\ell) \in \hat{V}(\bar{x}, \varepsilon, \tau)$ for all ℓ. Then, if $0 \notin G^S_\varepsilon(\bar{x})$, it must hold* $\liminf_{\ell \to \infty} t_\ell > 0$.

Proof. Let's denote $x^0_\ell := x_\ell$. By assumption (i) from Lemma 3.2, $m_\ell = m$ is fixed and

$$w_\ell := \sum_{i=0}^m \lambda^i_\ell P_{T(x_\ell)}(\nabla f(x^i_\ell))$$

has the minimum norm at ℓth iteration of the algorithm. By switching to another subsequence, we may assume to the contrary that $t_\ell \to 0$. By construction, $\gamma^{-1} t_\ell$ does not satisfy the Armijo condition, that is,

$$-\beta \gamma^{-1} t_\ell \|w_\ell\| \leq f(R_{x_\ell}(\gamma^{-1} t_\ell g_\ell)) - f(x_\ell). \tag{3.4}$$

By Lebourg's mean value Theorem, there exists $y_\ell \in [x_\ell, R_{x_\ell}(\gamma^{-1} t_\ell g_\ell)]$ and $v_\ell \in \partial f(y_\ell)$, such that

$$f(R_{x_\ell}(\gamma^{-1} t_\ell g_\ell)) - f(x_\ell) = \langle v_\ell, \gamma^{-1} t_\ell g_\ell \rangle + o(\gamma^{-1} t_\ell).$$

Multiplying by $-\|w_\ell\|\gamma/t_\ell$ and using that $g_\ell \in T(x_\ell)$, we get from (3.4) that

$$\langle P_{T(x_\ell)}(v_\ell), w_\ell \rangle - \frac{o(\gamma^{-1} t_\ell) \|w_\ell\|}{\gamma^{-1} t_\ell} \leq \beta \|w_\ell\|^2. \tag{3.5}$$

Since the tuples $(x^1_\ell, \dots x^m_\ell) \in \hat{V} \subseteq V_\varepsilon(\bar{x}, \bar{x}, \nu)$ are bounded, we may assume they converge to some $(z^1, \dots, z^m) \in \operatorname{cl} \hat{V}$. By Lemma 3.7, $(z^1, \dots, z^m) \in V_\varepsilon(\bar{x}, \bar{x}, \nu)$. Hence, denoting $\xi^i = \nabla f(z^i)$, we have

$$\min\{\|w\| : w \in P_S(\operatorname{conv}\{\xi^i\}_{i=1}^m)\} \leq \rho_\varepsilon(\bar{x}) + \nu. \tag{3.6}$$

Restricting the subsequence even further, we can assume $\nabla f(x_\ell)$ to be convergent to some $\xi^0 \in \partial f(\bar{x})$. Then,

$$\min\{\|w\| : w \in P_S(\operatorname{conv}\{\xi^i\}_{i=0}^m)\} \leq \rho_\varepsilon(\bar{x}) + \nu, \tag{3.7}$$

because the minimum is taken over a larger set compared to (3.6).

Assume that the minimum in (3.7) is attained at

$$\tilde{w} = P_S\left(\sum_{i=0}^m \tilde{\lambda}^i \xi^i\right).$$

Note that $\tilde{w}$ is unique. Obviously, $P_S(\xi^i) \in G^S_\varepsilon(\bar{x})$, $i = 1, \ldots, m$. Also, it is easy to see, that $P_S(\partial f(\bar{x})) \subseteq G^S_\varepsilon(\bar{x})$, and therefore $P_S(\xi^0) \in G^S_\varepsilon(\bar{x})$. We conclude that $\tilde{w} \in G^S_\varepsilon(\bar{x})$ and $\|\tilde{w}\| \leq \rho_\varepsilon(\bar{x}) + \nu$. By Lemma 3.5,

$$\langle P_S(v), \tilde{w}\rangle > \beta \|\tilde{w}\|^2$$

for every $P_S(v) \in G^S_\varepsilon(\bar{x})$. The aim is now to show that w_ℓ has a subsequence converging to $\tilde{w}$. Then, since $v_\ell \in \partial f(y_\ell)$ has a convergent subsequence to some $v \in \partial f(\bar{x})$, a limitation of (3.5) in subsequences using $t_\ell \to 0$, yields the contradiction $\langle P_S(v), \tilde{w}\rangle \leq \beta \|\tilde{w}\|^2$.

Restricting to further subsequences, we can assume that for $i = 0, 1, \ldots, m$ the sequence λ^i_ℓ converges to some λ^i_*. Then, by assumption (ii) from Lemma 3.2, w_ℓ converges to

$$w_* = P_S\left(\sum_{i=0}^{m} \lambda^i_* \xi^i\right) \in P_S(\operatorname{conv}\{\xi^i\}_{i=0}^m).$$

We need to show that $w_* = \tilde{w}$. Since $\tilde{w}$ is the unique minimizer of (3.7), it is enough to prove that $\|w_*\| \leq \|\tilde{w}\|$ in order to make this conclusion. Let $\eta > 0$. For large enough ℓ it holds

$$\|w_*\| \leq \|w_\ell\| + \eta \leq \left\| P_{T(x_\ell)}\left(\sum_{i=0}^{m} \tilde{\lambda}^i \nabla f(x^i_\ell)\right)\right\| + \eta.$$

The second inequality holds by the choice of w_ℓ. The expression in the norm converges to $\tilde{w}$, so we may also assume

$$\left\| \tilde{w} - P_{T(x_\ell)}\left(\sum_{i=0}^{m} \tilde{\lambda}^i \nabla f(x^i_\ell)\right)\right\| \leq \eta.$$

In conclusion, $\|w_*\| \leq \|\tilde{w}\| + 2\eta$ for any $\eta > 0$. □

Proof of Theorem 3.3. We assume the case $\liminf_{\ell\to\infty} f(x_\ell) > -\infty$. By construction, $f(x_{\ell+1}) - f(x_\ell) < -\beta t_\ell \|w_\ell\|$ and $\|x_{\ell+1} - x_\ell\| \leq 2\kappa t_\ell$. Using telescopic sums, this implies

$$\sum_{\ell=1}^{\infty} t_\ell \|w_\ell\| < \infty \quad \text{and} \quad \sum_{\ell=1}^{\infty} \|x_{\ell+1} - x_\ell\| \|w_\ell\| < \infty. \tag{3.8}$$

Let $(x_\ell)_{\ell\in\mathcal{L}}$ be a convergent subsequence with limit $\bar{x}$. To show that $\bar{x}$ is a critical point, we aim to prove that $(w_\ell)_{\ell\in\mathcal{L}}$ has a subsequence that converges to zero. Then, since $w_\ell \in \partial_{\varepsilon_\ell} f(x_\ell) + N(x_\ell)$, it follows that $0 \in \partial f(\bar{x}) + S^\perp$. In particular, when $x_\ell \in \mathcal{M}_{s_\ell}$ and $T_{\mathcal{M}_{s_\ell}}(x_\ell) \subseteq T(x_\ell)$, since $w_\ell \in \partial_{\varepsilon_\ell} f(x_\ell) + (T_{\mathcal{M}_{s_\ell}}(x_\ell))^\perp$, we conclude by Lemma 3.2 that $0 \in \partial f(\bar{x}) + (T_{\mathcal{M}_s}(\bar{x}))^\perp$.

To prove the existence of such a subsequence, we use two different arguments, depending on whether (x_ℓ) itself converges or not. If (x_ℓ) diverges, we argue as Kiwiel [14], namely that combining Lemma 3.6 and (3.8) yields $\liminf_{\ell\in\mathcal{L}} \|w_\ell\| = 0$. If $x_\ell \to \bar{x}$, then the existence of a subsequence of w_ℓ converging to zero is equivalent to the statement $\delta_\ell \downarrow 0$, $\varepsilon_\ell \downarrow 0$, which is shown below.

By construction of the algorithm, the contrary would mean that there exists ℓ^* such that $\delta_\ell = \delta$ and $\varepsilon_\ell = \varepsilon$ remain fixed for all $\ell \geq \ell^*$. This only happens if $\|w_\ell\| > \delta$ for $\ell \geq \ell^*$ (see line 9). By (3.8), this implies $t_\ell \to 0$ and $\sum_{\ell=1}^{\infty} \|x_{\ell+1} - x_\ell\| < \infty$. In particular, x_ℓ is then a Cauchy sequence and has a limit $\bar{x} \in \mathcal{M}$. We then consider a subsequence of (x_ℓ) that satisfies the properties (i) and (ii) of Lemma 3.2, but for notational convenience, we assume that this is the whole sequence (x_ℓ) itself. To derive a contradiction, we distinguish between two possible cases.

First, assume $0 \notin G^S_\varepsilon(\bar{x})$. Let ν, τ and $\hat{V} = \hat{V}(\bar{x}, \varepsilon, \tau)$ be chosen as in Theorem 3.8. Since $(x^1_\ell, \ldots, x^m_\ell)$ are sampled independently and uniformly from $D^m_\varepsilon(x_\ell)$, and $\hat{V}$ is a nonempty open subset of $D_\varepsilon^m(x_\ell)$, it will hold $(x_\ell^1,\dots,x_\ell^m)\in\hat V$ infinitely often. By Theorem 3.8, this contradicts $t_\ell\to 0$.

In the second case, assume $0\in G_\varepsilon^S(\bar x)$. Then $\rho_\varepsilon(\bar x)=0$. Let $\nu=\delta/2$ and choose τ and $\hat V=\hat V(\bar x,\tau,\nu)$ according to Lemma 3.7. Similar as before, we will have $(x_\ell^1,\dots,x_\ell^m)\in\hat V$ infinitely often. Also, $x_\ell\in B(\bar x,\tau)$ for ℓ large enough. Then

$$\begin{aligned}&\min\{\|w\|: w\in P_S(\operatorname{conv}\{\nabla f(x_\ell^i)\}_{i=1}^m)\}\\ \le{}& \rho_\varepsilon(\bar x)+\nu=\delta/2\le\|w_\ell\|-\delta/2\\ \le{}&\min\{\|w\|: w\in P_{T(x_\ell)}(\operatorname{conv}\{\nabla f(x_\ell^i)\}_{i=1}^m)\}-\delta/2.\end{aligned}$$

This is a contradiction, because both sequences of minimal have the same limit inferior. This can be shown using similar arguments as in the proof of Theorem 3.8 by taking a convergent subsequence $(x_\ell^1,\dots,x_\ell^m)\to(z^1,\dots,z^m)$. In summary, we have shown that $\delta_\ell\downarrow 0$, $\varepsilon_\ell\downarrow 0$. □

4. Some numerical experiments for robust low-rank matrix recovery

As an application, we have implemented our algorithm for solving problems of the form

$$\min_{\operatorname{rank}(X)\le r} f(X)$$

on the space $\mathbb{R}^{M\times N}$ of $M\times N$ matrices (equipped with the Frobenius inner product). Specifically, we conducted numerical experiments for low-rank recovery of noisy matrices via minimization of entry-wise ℓ_1 distance. This is sometimes referred to as *robust* low-rank recovery and is explained in Sec. 4.2 below. But first, we shall give some background on the low-rank matrix varieties that are the main geometric object in this optimization task.

4.1. **Low-rank matrix varieties.**

The real algebraic varieties

$$\mathcal{M}_{\le r}=\{X\in\mathbb{R}^{M\times N}:\operatorname{rank}(X)\le r\}$$

fit perfectly in the abstract setting considered above for several reasons.

4.1.1. *Stratification into fixed-rank manifolds.* First, as in (1.2), they admit a stratification by dimension

$$\mathcal{M}_{\le r}=\bigcup_{s=0}^{r}\mathcal{M}_s \tag{4.1}$$

into smooth manifolds

$$\mathcal{M}_s=\{X\in\mathbb{R}^{M\times N}:\operatorname{rank}(X)=s\}$$

of fixed-rank matrices. The geometry of these manifolds is well-understood. In particular, we have

$$\dim(\mathcal{M}_s)=(M+N-s)s, \tag{4.2}$$

and

$$T_{\mathcal{M}_s}(X)=\mathcal{U}\otimes\mathbb{R}^N+\mathbb{R}^M\otimes\mathcal{V}=(\mathcal{U}\otimes\mathcal{V})\oplus(\mathcal{U}^\perp\otimes\mathcal{V})\oplus(\mathcal{U}\otimes\mathcal{V}^\perp), \tag{4.3}$$

where $\mathcal{U}\subseteq\mathbb{R}^M$ is the column space of the matrix X (its image), and $\mathcal{V}\subseteq\mathbb{R}^N$ is its row space (the image of X^T). Here we have identified $\mathbb{R}^{M\times N}$ as a tensor product $\mathbb{R}^M\otimes\mathbb{R}^N$. The symbol $\oplus$ indicates that the splitting into subspaces is orthogonal (with respect to Frobenius inner product).

If we are given a decomposition $X=USV^T\in\mathcal{M}_s$ with $U\in\mathbb{R}^{M\times s}$ and $V\in R^{N\times s}$ having orthonormal columns, then tangent space is efficiently parametrized as follows:

$$T_{\mathcal{M}_s}(X)=\{UEV^T+FV^T+UG^T: E\in\mathbb{R}^{s\times s},\ U^TF=0,\ V^TG=0\}. \tag{4.4}$$

The orthogonal projection (with respect to the Frobenius inner product) of a matrix Z on the subspace $T_{\mathcal{M}_s}(X)$ is given as

$$P_{T_{\mathcal{M}_s}(X)}(Z) = U \underbrace{U^T Z V}_{E} V^T + \underbrace{(Z - UU^T Z)V}_{F} V^T + U \underbrace{U^T(Z - ZVV^T)}_{G^T},$$

yielding the parameters E, F and G as indicated. When s is small compared to M and N, it is important that these parameters can be computed by performing sequential matrix products with the "tall" matrices U and V (or their transposes) only. The full projectors UU^T and VV^T should never be computed.

4.1.2. *Regularity of the stratification.* It follows from Whitney's abstract construction [23, Sec. 19] that the stratification (4.1) is a-regular. However, thanks to the simple structure of the tangent spaces $T_{\mathcal{M}_s}(X)$ it is very easy to verify this directly. Let (X_ℓ) be a sequence of rank-r matrices converging to $\bar{X}$ having rank $s < r$. Assume that $T_{\mathcal{M}_r}(X_\ell)$ converges to T in the sense of subspaces. In light of (4.3), after passing to subsequences, we can assume that the column and row spaces of X_ℓ converge to subspaces $\mathcal{U}$ and $\mathcal{V}$, respectively, so that $T = \mathcal{U} \otimes \mathbb{R}^N + \mathbb{R}^M \otimes \mathcal{V}$. Then, in order to show $T_{\mathcal{M}_s}(\bar{X}) \subseteq T$ (which means a-regularity), it is enough to argue that $\mathcal{U}$ contains the column space of $\bar{X}$, while $\mathcal{V}$ contains the row space of $\bar{X}$. Both is obviously true, since $X_\ell v \to \bar{X} v$ for all $v \in \mathbb{R}^N$ and $X_\ell^T u \to \bar{X}^T u$ for all $u \in \mathbb{R}^M$.

4.1.3. *Linear subspaces in the Bouligand tangent cone.* A simple description of the Bouligand tangent cone to $\mathcal{M}_{\leq r}$ in singular points is available [6, 17]. Let $X \in \mathcal{M}_{\leq r}$ have rank $s \leq r$, then the tangent cone is given as

$$T^B_{\mathcal{M}_{\leq r}}(X) = T_{\mathcal{M}_s}(X) \oplus \{Y \in (T_{\mathcal{M}_s}(X))^\perp : \operatorname{rank}(Y) \leq r - s\}.$$

Hence, when $s < r$, $T^B_{\mathcal{M}_{\leq r}}(X)$ contains many linear subspaces $T(X)$ satisfying

$$T_{\mathcal{M}_s}(X) \subseteq T(X) \subseteq T^B_{\mathcal{M}_{\leq r}}(X).$$

Possible choices include subspaces of the form

$$T(X) = T_{\mathcal{M}_s}(X) \oplus (\mathcal{U}_\perp \otimes \mathcal{V}_\perp), \tag{4.5}$$

where $\mathcal{U}_\perp \subseteq \mathbb{R}^M$ and $\mathcal{V}_\perp \subseteq \mathbb{R}^N$ are subspaces of dimension $r - s$ that are orthogonal to the column and row spaces of X respectively. Using the parametrization (4.4) of $T_{\mathcal{M}_s}(X)$, and letting $U_\perp \in \mathbb{R}^{M \times (r-s)}$, $V_\perp \in \mathbb{R}^{N \times (r-s)}$ be basis representations of $\mathcal{U}_\perp$, $\mathcal{V}_\perp$, respectively, elements in such a space $T(X)$ are then represented as

$$Z = UEV^T + FV^T + UG^T + U_\perp H V_\perp^T \tag{4.6}$$

subject to $U^T F = 0$ and $V^T G = 0$. Here $H \in \mathbb{R}^{(r-s)\times(r-s)}$.

In the optimization algorithm, the choice of subspaces $T(X)$ determines which row and column spaces can be reached from the current singular point X. In our experiments we used spaces of the form (4.5), taking as $\mathcal{U}_\perp$ and $\mathcal{V}_\perp$ either random subspaces orthogonal to $\mathcal{U}$ and $\mathcal{V}$, or, alternatively, the subspaces spanned by the dominant $r - s$ left and right singular vectors of the orthogonal projection of $\nabla f(X)$ on $(T_{\mathcal{M}_s}(X))^\perp$. Compared to random subspaces, this second choice based on the gradient appears very reasonable and has been observed to be beneficial in smooth low-rank matrix completion [19]. However, in our experiments on robust low-rank approximation we could not confirm this.

4.1.4. *On our implementation of the GS algorithm on $\mathcal{M}_{\leq r}$.* The manifold $\mathcal{M}_r$ of matrices with full possible rank r is dense and open in $\mathcal{M}_{\leq r}$. Hence in a practical computation on $\mathcal{M}_{\leq r}$ with initial guess on $\mathcal{M}_r$, an iterate of rank less than r is never encountered. Also a nonsmooth point of f will never occur in practice. This makes it possible to deal with the algorithm as a Riemannian optimization algorithm on fixed rank matrix manifolds, in the same way as in [5, 20]. In this viewpoint, the GS algorithms is just a specific way to select a search direction in the tangent space. For our implementation, we used the `manopt` toolbox [3] for MATLAB, which provides a convenient framework for defining Riemannian solvers on manifolds of fixed-rank matrices.

A different situation occurs when one wishes to sequentially increase the rank during the optimization of the cost function. A rank-increasing strategy is useful when the target rank of a satisfying solution is not known in advance. Also it has been observed to be computationally beneficial [18, 19]: starting with small ranks is not only computationally cheaper, but also provides starting guesses for a higher rank which are potentially better than starting at random. Every time one embeds the result X of a fixed-rank optimization, say of rank $s < r$, as a starting guess for a variety of higher rank, say, $\mathcal{M}_{\leq s+r_{\text{incr}}}$, one is faced with the scenario considered in this paper of selecting a linear subspace $T(X)$ in the Bouligand tangent cone $T_{\mathcal{M}_{\leq s+r_{\text{incr}}}}(X)$. We choose subspaces of the form (4.5). The subspaces $\mathcal{U}_\perp$, $\mathcal{V}_\perp$ are represented by orthonormal matrices $U_\perp$ and $V_\perp$ with r_{incr} columns. In the experiments these matrices are either randomly chosen (but respectively orthogonal to row and column space of X), or obtained from the dominant singular vectors of $\nabla f(X) - P_{\mathcal{M}_s}(\nabla f(X))$ (the orthogonal projection of $\nabla f(X)$ on $(T_{\mathcal{M}_s}(X))^\perp$).

In the algorithm, one has to draw random elements from the unit ball in the space $T(X)$. Since the decomposition (4.6) is orthogonal, this is achieved by randomly drawing E, F, G and H (the latter only at the rank increasing steps) – each of Frobenius norm one and subject to the constraints on F and G – and then forming a linear combination $a_1E + a_2F + a_3G + a_4H$ where (a_1, a_2, a_3, a_4) is a random vector in the unit sphere of $\mathbb{R}^4$. In our implementation, we do it a bit differently. Given $X = USV^T$ of rank s, the random sampling in the unit ball of $T_{\mathcal{M}_s}(X)$ is realized by constructing E, F and G using `randn` in MATLAB, replacing F and G with normalized versions of $F - UU^TF$ and $G - VV^TG$, respectively, and returning

$$Z_1 = (a_1UEV^T + a_2FV^T + a_3UG^T)/\sqrt{3},$$

where (a_1, a_2, a_3) is uniformly random in $[0, 1]^3$. In iterations when the rank is increased by r_{incr}, we construct $Z_1 \in T_{\mathcal{M}_s}(X)$ as just described. Then we construct normalized $H \in \mathbb{R}^{r_{\text{incr}} \times r_{\text{incr}}}$ with the aid of `randn` and return

$$Z = (Z_1 + a_4U_\perp HV_\perp)/\sqrt{2},$$

with a_4 uniformly random in $[0, 1]$. The choice of the matrices $U_\perp$, $V_\perp$ has been explained further above.

Finally, the quadratic program in line 6 of the algorithm needs to be solved. Assembling and solving this problem becomes the computationally most expensive part of the GS algorithm when the number m of sample points is large. The problem can reformulated as finding $\zeta \in \mathbb{R}^{m+1}$ that minimizes $\zeta^T\bar{G}\zeta$ subject to the constraints $\zeta \geq 0$ and $\sum_{i=1}^n \zeta_i = 0$, where $\bar{G}$ is the Gram matrix of the $m + 1$ tangent vectors obtained from projecting the gradients. To solve this problem we used the function `quadprog` (with default values) which is part of the MATLAB optimization Toolbox. For assembling the matrix $\bar{G}$, it is useful to note that the inner product of tangent vectors represented in a form as in (4.4) can be rather efficiently computed when the rank is small. Such a functionality is provided by `manopt`. Still we observe that setting up the matrix $\bar{G}$ dominates the computational cost when many sample points in a relatively high-dimensional tangent space are given.

In all the subsequent numerical experiments, the sampling radius is initialized with $\varepsilon_0 = 10^{-4}$. The other parameters are fixed as follows: $\delta_0 = 10^{-3}$, $\gamma = 2^{-1}$, $\varepsilon_{\text{opt}} = 10^{-16}$, $\delta_{\text{opt}} = 10^{-16}$, $\beta = 10^{-4}$, $\theta_\varepsilon = 10^{-1}$, and $\theta_\delta = 10^{-1}$. The optimization for a fixed rank r is terminated either after some prescribed number of iterations, or if $|f(X_{\ell+1}) - f(X_\ell)| < 10^{-10}$. We note that with these parameters we observed in all our experiments that the sampling radius remains fixed during the iterations, that is, line 9 is never activated. This can have several problem dependent reasons, and correspondingly we cannot state that we really find stationary points (typically $\|w_\ell\|$ stagnates in the order of 10^{-1}). However, shrinkage of sampling radius can be encountered when the sampling size is significantly larger than $\dim(\mathcal{M}) + 1$, for instance, $3(\dim(\mathcal{M}) + 1)$. This confirms an observation also made in [11].

4.2. **Numerical results for robust low-rank approximation.** By robust low-rank approximation one means the approximation and recovery of low-rank matrices based on some or all given entries, of which some are corrupted by large error, so-called outliers. For such a task, minimization of different combinations of Frobenius, ℓ_1, nuclear norm and other error measures have been proposed; cf. [5, Sec. 1.1] for references. Here, we consider the very basic and prototypical problem

$$\min_{\operatorname{rank}(X)\le r} \|A - X\|_{\ell_1} = \min_{\operatorname{rank}(X)\le r} \sum_{ij} |a_{ij} - x_{ij}| \tag{4.7}$$

for a given matrix $A \in \mathbb{R}^{M\times N}$. The cost function $f(X) = \|A - X\|_{\ell_1}$ is locally Lipschitz, and is continuously differentiable in the set D of all matrices X for which $A - X$ contains no zero entries. The gradient is then given as $\nabla f(X) = \operatorname{sign}(A - X)$. It is likely the case, but we did not attempt to prove it, that for any r the set $\mathcal{M}_{\le r} \cap D$ is of relative full measure in $\mathcal{M}_{\le r}$, which was crucial for the convergence proof.

In practice, the matrix A may not be exactly available, but is measured subject to Gaussian noise with some extreme outliers. In comparison to low-rank approximation in the Frobenius norm (which in the case that all entries are given can be solved using SVD), it is expected that minimization in ℓ_1-norm is more robust to sparse noise and extreme outliers. In the first two experiments below, A will be generated as

$$A = A_{\text{ex}} + \lambda E_{\text{noise}} + \mu E_{\text{out}}, \quad \|A_{\text{ex}}\|_F = \|E_{\text{noise}}\|_F = \|E_{\text{out}}\|_F = 1, \tag{4.8}$$

where A_{ex} is the assumed ground truth, E_{noise} is a dense matrix with random entries (modelling general noise in measurements), and E_{out} is a sparse matrix with 1% random nonzero entries (modelling outliers). All three matrices have Frobenius norm one (denote by $\|\cdot\|_F$). Thus the scalars $\lambda, \mu \ge 0$ in (4.8) determine the noise level. The goal in solving the robust low-rank approximation problem (4.7) is then to recover a good rank-r approximation to A_{ex}, which is, say, optimal in Frobenius norm up to the noise level λ.

With our experiments below we are able to confirm this robustness of the problem (4.7) to outliers and demonstrate that it can be in principle solved using the GS algorithm on $\mathcal{M}_{\le r}$. However, it is not our aim to make a specific claim regarding the potential applications, where it can be important to further take variations of the above problem including smooth or nonsmooth penalty terms into account. In first place, we consider the problem (4.7) as an interesting, nontrivial instance of the abstract scenario considered in this paper, for which the GS algorithm might be useful.

All experiments have been conducted on a Linux workstation with 3.2 GHz CPU cores and 6 GB of memory, using MATLAB R2015b with Optimization Toolbox and a modified version of `manopt`.

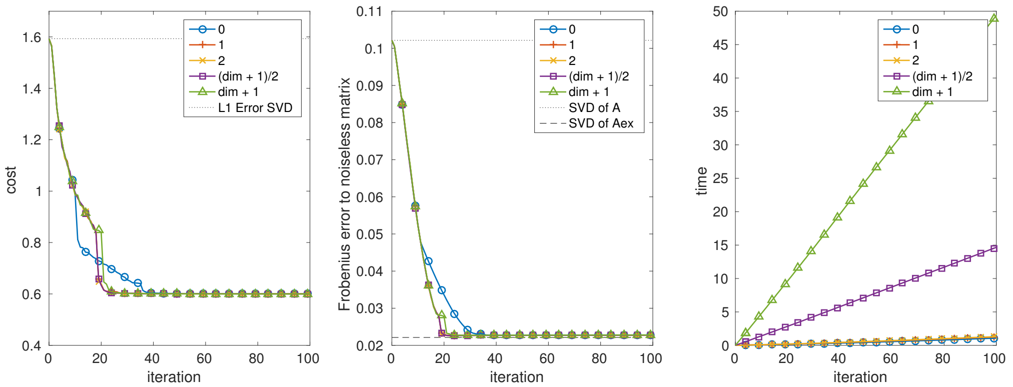

FIGURE 1. Results of GS algorithm for problem (4.7) with $A = A_{\mathrm{ex}} + E_{\mathrm{out}} \in \mathbb{R}^{30\times 30}$ and $r = 3$, tested for different sampling sizes. Here dim = 171. Left: cost function values $\|A - X_\ell\|_{\ell_1}$. The initial value (dotted line) is obtained from a best rank-three approximation (in Frobenius norm) of A via SVD. Middle: Frobenius errors $\|A_{\mathrm{ex}} - X_\ell\|_F$. The best possible is the best rank-three approximation error of A_{ex}, plotted as dashed line. Right: computational time (in seconds) vs. GS iterations.

4.2.1. *Influence of sample size.* For obtaining the convergence results in the first part of the paper, it was crucial that at least $\dim(T(x)) + 1$ nearby gradients are sampled in addition to the one at the current iterate. At non-singular points $T(x)$ is just the tangent space to $\mathcal{M}$. If the dimension of $\mathcal{M}$ is large, the solution of the quadratic program in line 6 of the algorithm for finding the search direction becomes computationally very expensive. For optimization on low-rank matrix manifolds we observed that this issue happens already for medium sized matrix of moderate rank due to (4.2). For instance, already when dealing with 100×100 matrices of rank one with $m = \dim(\mathcal{M}_1) + 1 = 200$ sample points, the algorithm is quite slow.

It is not clear whether in practice so many sample points are really necessary. The opposite extreme is taking $m = 0$ sample points, in which case the method reduces to the (Riemannian) steepest descent method. In this subsection we aim to investigate the influence of the sample size.

In the first experiment, we run our implementation of the GS algorithm for the problem (4.7) with $M = N = 30$ and target rank $r = 3$ (robust rank-three approximation) for different sample sizes

$$m \in \left\{0, 1, 2, \frac{\dim(\mathcal{M}_r + 1)}{2}, \dim(\mathcal{M}_r) + 1\right\}.$$

Specifically, $\dim(\mathcal{M}_3) = 171$ here. The matrix A is given as in (4.8), with $\lambda = 0$ (no background noise) and $\mu = 0.1$. The ground truth A_{ex} is a matrix with exponentially decaying singular values. It is generated by replacing the singular values of a random matrix with ones that are logarithmically distributed between 1 and 10^{-16}, and normalizing to Frobenius norm one. The matrix E_{out} has nine nonzero entries representing outliers in the measurement of A_{ex}. As a starting guess, we choose the best rank three approximation (in Frobenius norm) of A, obtained from an SVD. A typical outcome is given in Fig. 1, where we plot the cost function values $\|A - X_\ell\|_{\ell_1}$, the Frobenius errors $\|A_{\mathrm{ex}} - X_\ell\|_F$ as well as the execution times. The latter is given to make a relative comparison between sampling sizes. We did not aim for the most efficient implementation.

It can be seen that the ℓ_1-minimization is more robust to the influence of the outliers: for all sampling sizes a rank-three approximation is obtained, whose Frobenius error to the initial matrix A_{ex} is comparable to the best possible one, which is quite remarkable. In particular, the

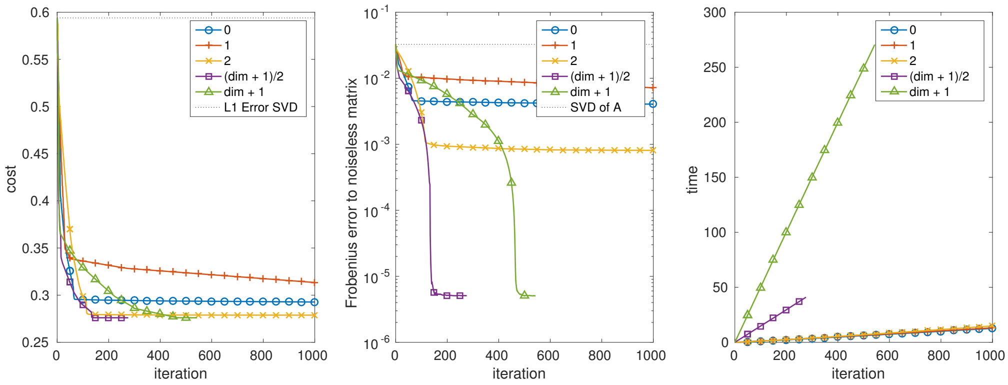

FIGURE 2. Results of GS algorithm for problem (4.7) with $A = A_{\text{ex}} + E_{\text{out}} + E_{\text{noise}} \in \mathbb{R}^{30\times 30}$ and $r = 3$, tested for different sampling sizes. Here $\dim = 171$. The matrix A_{ex} is of rank r. Left: cost function values $\|A - X_\ell\|_{\ell_1}$. The initial value (dotted line) is obtained from a best rank-three approximation (in Frobenius norm) of A via SVD. The curves for $\frac{\dim(\mathcal{M}_r+1)}{2}$, $\dim(\mathcal{M}_r)+1$ terminated because the stopping condition $|f(X_{\ell+1}) - f(X_\ell)| < 10^{-10}$ was satisfied. Middle: Frobenius errors $\|A_{\text{ex}} - X_\ell\|_F$. The best possible would be zero, but given the background noise $\|E_{\text{noise}}\|_F = 10^{-5}$ a recovery in this order of magnitude can be considered optimal. Right: computational time (in seconds) vs. GS iterations.

initial error obtained from the rank-three truncation of the SVD of the perturbed matrix A is significantly improved.

However, as one can also see, a larger sampling size has basically no effect on the achievable result, while considerably increasing the computational cost. We believe that one possible reason for this behavior is that the error $A - X$ at local minima of (4.7) cannot be expected to contain zero entries in the considered setup. This means that (most likely) the cost function is smooth at critical points. Nonsmooth optimization methods like the GS algorithm aim at situations where a minimum is achieved at non-differentiable points. To mimic such a situation we set up a second test case.

In this second case, A_{ex} is a matrix of rank r, whose upper $r \times N$ block is generated with `randn`, while the lower $(M - r) \times N$ block contains only zeros. Then A is generated via (4.8) with outliers as before ($\mu = 0.1$), but also this time we add Gaussian noise of level $\lambda = 10^{-5}$ in (4.8). A typical convergence history for the GS algorithm in this case is shown in Fig. 2. In this scenario larger sample sizes yield better results. In fact, only for $m = \frac{\dim(\mathcal{M}_r+1)}{2}$ and $m = \dim(\mathcal{M}_r) + 1$ the exact matrix A_{ex} was reliably recovered up to the background noise level 10^{-5}, which is the best to hope for. The ordering of the curves for zero, one or two sample points was not very predictable.

From both experiments, we draw a mixed conclusion: while it does not pay off using the $\dim(\mathcal{M}) + 1$ sample points as required by theory (and is simply not possible when $\dim(\mathcal{M})$ is large), it does not influence the computational cost to use at least a few gradient samplings. In some situations it appears important that the sample size is at least proportional to the dimension of the variety. In the following experiments, we set the sample size on the manifold $\mathcal{M}_r$ to be $2r$. In this way we are able to deal with larger low-rank matrices.

4.2.2. *Rank-increasing algorithm.* In this experiment, we put the rank-increasing strategy described in Sec. 4.1.4 to the test. We create a matrix A of the form (4.8). As in the first experiment of Sec. 4.2.1, the matrix A_{ex} is generated as a dense full rank matrix of size $M \times N$, with singular

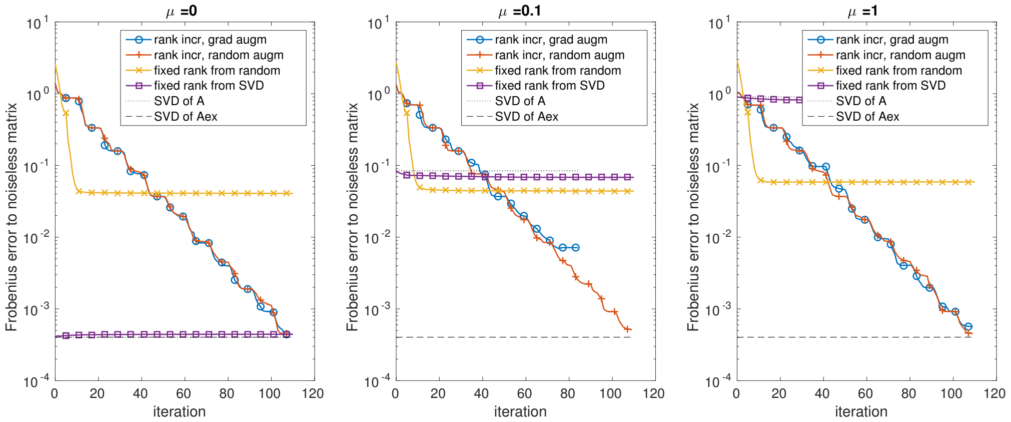

FIGURE 3. Rank-increasing strategy for problem (4.7) with $A \in \mathbb{R}^{100\times 100}$ of the form (4.8) and $r = 21$ with $\lambda = 10^{-5}$, tested for different magnitudes $\mu \in \{0, 0.1, 1\}$ of outliers. One can nicely see the staircase behavior where on every step the rank was increased by $r_{\text{incr}} = 2$ using two different subspace augmentation strategies. In the left plot, the dashed and dotted line are (almost) on top of each other, since A and A_{ex} only vary by $\lambda = 10^{-5}$. In the right plot, the dotted line is beneath the line with the square markers.

values logarithmically distributed between 1 and 10^{-16}, and then scaled to Frobenius norm one. For the Gaussian noise we choose again $\lambda = 10^{-5}$. Three different magnitudes $\mu \in \{0, 0.1, 1\}$ for the outliers will be tested.

Given A, we aim to compute a low-rank approximation of A_{ex}. We use the rank-increasing strategy, starting with a random matrix of rank $s = 1$, iterating on the manifold $\mathcal{M}_s$, increasing the rank by r_{incr}, iterating on $\mathcal{M}_{s+r_{\text{incr}}}$ and so forth, until a certain target rank $s = r$ is reached. The sampling size on the rank-s manifold is set to $2s$, the number of iterations per rank is limited. In the rank-increasing step we distinguish between a random subspace augmentation and an augmentation by subspaces obtained from rank-r_{incr} truncation of the projection of $\nabla f(X)$ to $(T_{\mathcal{M}_r}(X))^{\perp}$ (cf. Sec. 4.1.4). For comparison, we also run the GS algorithm on the fixed rank manifold $\mathcal{M}_r$, both with random starting guess and with starting guess obtained from an SVD of A.

In Fig. 3, we see results for $M = N = 100$, target rank $r = 21$, rank-increase by $r_{\text{incr}} = 2$, and at most nine iterations per rank. In this case, the best possible rank-21 approximation error to A_{ex} in Frobenius norm is around $4 \cdot 10^{-4}$. For all three magnitudes μ of outliers, the rank-increasing algorithm with random subspace augmentation is able to essentially recover a best rank-21 approximation of A_{ex}. Note that for this to be possible it is necessary that the norm λ of the background noise is lower than the best rank-r approximation error in Frobenius norm to A_{ex}, say at least by an order of magnitude, as it is the case here (10^{-5} vs. 10^{-4}). For the algorithm using low-rank approximation of the projected gradient for rank increase we sometimes (but not often) observed stagnation at an earlier point as can be seen in the middle plot, and therefore decided to report such a case. It is interesting that, even in the case $\mu = 0$ (no outliers), the minimization of the function (4.7) finds an almost optimal low-rank approximation in Frobenius norm, but we do not have an explanation for this.

In comparison, the fixed-rank methods stagnate at suboptimal values, except in the case $\mu = 0$. It is worthwhile to remark that for large magnitude of outliers ($\mu = 1$) a starting guess from the

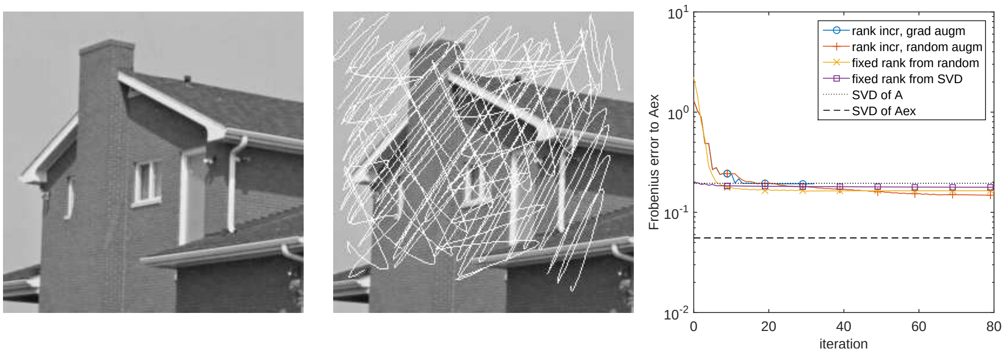

FIGURE 4. The original test image `'house.png'` is at the left. The scratched version in the middle is used as the input A for the GS algorithm (after normalization to Frobenius norm one). Right: error history for the four variants of GS algorithm, showing the Frobenius distance to (a normalized version) of the original image.

SVD of A is not recommended, presumably because it takes too much false information from the outliers into account.

4.2.3. *Inpainting.* In the third experiment, we try to use our algorithm for reconstruction of scratched grayscale images. The ground truth A_{ex} is now the 512×512 matrix obtained from scaling the grayscale test image `'house.png'` (Fig. 4 left). As matrix A, we take a scratched version of this image (Fig. 4 middle). With matrix A as an input, we conduct exactly the same experiment as in Sec. 4.2.2. For the rank-increasing algorithm we start with rank $s = 1$ and then increase seven times by $r_{\text{incr}} = 3$, leading to a final rank $r = 22$. The number of iterations per rank is now set to ten. The sampling size is again $2s$ for rank-s optimization, and $2r$ for the fixed-rank methods. In Fig. 4 right one can see a corresponding error history. It is interesting that in this case, the rank-increasing algorithm does not exhibit the staircase behavior as for the previous experiment.

Further, all four methods produce results in the order of the best rank-r approximation of the corrupted matrix A. The outcome in terms of image reconstruction is nevertheless very different as can be seen in Fig. 5. The first row in this picture shows, from left to right, the truncation of the original image A_{ex} to rank $r = 22$ (which would be the ideal goal), the SVD truncation of the corrupted matrix A, and the results of the two fixed-rank GS methods (with random starting guess and SVD starting guess). The two other rows in Fig. 5 show the intermediate results for the rank-increasing algorithm with the random subspace augmentation (red curve in Fig 4). It produces arguably a better reconstruction. It is interesting that while all 'diagonal' scratches have been successfully removed, some axis aligned scratches remained, perhaps because they do not violate the low-rank constraint.

5. CONCLUSION

In this paper we have developed a gradient sampling algorithm for minimization of locally Lipschitz functions on subvarieties of Euclidean spaces that admit stratifications into manifolds. The new method is considerably simpler than previous attempts since it only requires sampling in linear subspaces and no vector transport. We are able to deal with singular points of the stratification using linear subspaces in the Bouligand tangent cone, and provide convergence results that are as strong as the analogous results for GS algorithm in linear space. The varieties of low-rank matrices provide an important example for the considered setting, where the non-trivial

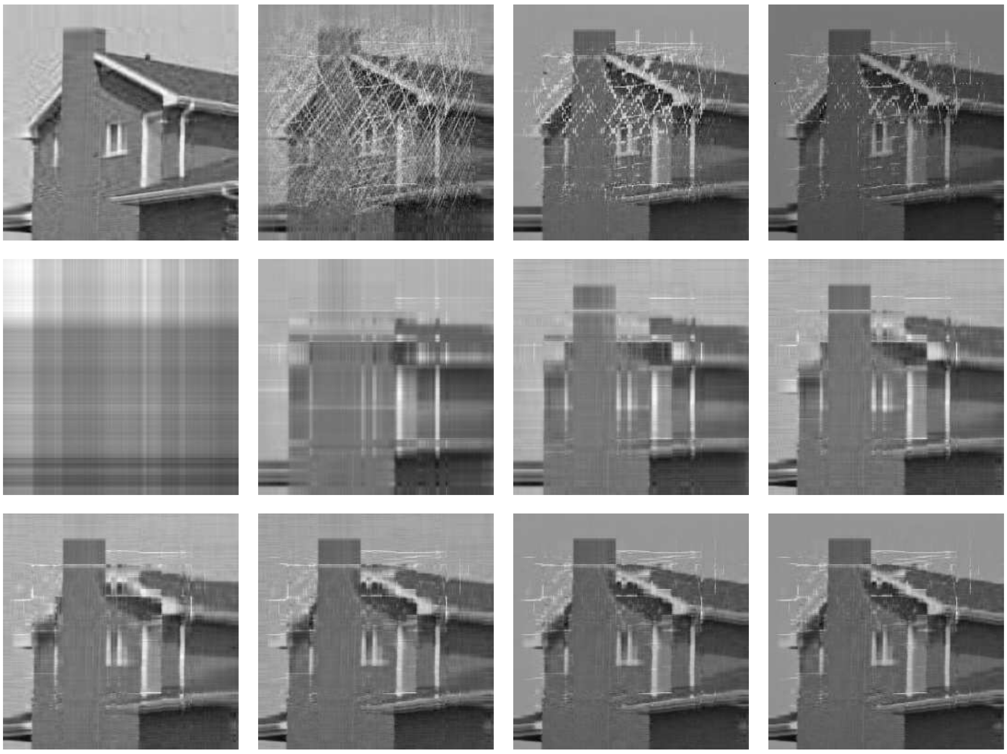

FIGURE 5. Low-rank approximations of `'house.png'`. Top line (from left to right): truncation of the original image to rank $r = 22$, the SVD truncation of the scratched version, and the results of the two fixed-rank GS methods (random starting guess and SVD starting guess). The other eight pictures show the intermediate results of the rank-increasing algorithm (with random subspace augmentation) for the ranks $1, 4, 7, 10, 13, 16, 19, 22$.

linear subspaces in the tangent cone at rank-deficient matrices correspond to subspace enrichment of the corresponding column and row space. In this way, rank-increasing algorithms can be easily incorporated into the considered framework. Our numerical experiments on robust low-rank recovery indicate that the GS method can be successfully used on such problems.

References

[1] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization algorithms on matrix manifolds.* Princeton University Press, Princeton, NJ, 2008.

[2] A. Bagirov, N. Karmitsa, and M. M. Mäkelä. *Introduction to nonsmooth optimization.* Springer, 2014.

[3] N. Boumal, B. Mishra, P.-A. Absil, and R. Sepulchre. Manopt, a MATLAB toolbox for optimization on manifolds. *J. Mach. Learn. Res.*, 15:1455–1459, 2014.

[4] J. V. Burke, A. S. Lewis, and M. L. Overton. A robust gradient sampling algorithm for nonsmooth, nonconvex optimization. *SIAM J. Optim.*, 15(3):751–779, 2005.

[5] L. Cambier and P.-A. Absil. Robust low-rank matrix completion by Riemannian optimization. *SIAM J. Sci. Comput.*, 38(5):S440–S460, 2016.

[6] T. P. Cason, P.-A. Absil, and P. Van Dooren. Iterative methods for low rank approximation of graph similarity matrices. *Linear Algebra Appl.*, 438(4):1863–1882, 2013.

[7] F. H. Clarke. *Optimization and nonsmooth analysis.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, second edition, 1990.

[8] A. A. Goldstein. Optimization of Lipschitz continuous functions. *Math. Programming*, 13(1):14–22, 1977.

[9] P. Grohs and S. Hosseini. Nonsmooth trust region algorithms for locally Lipschitz functions on Riemannian manifolds. *IMA J. Numer. Anal.*, 36(3):1167–1192, 2016.

[10] P. Grohs and S. Hosseini. ε-subgradient algorithms for locally Lipschitz functions on Riemannian manifolds. *Adv. Comput. Math.*, 42(2):333–360, 2016.

[11] S. Hosseini, W. Huang, and R. Yousefpour. Line search algorithms for locally Lipschitz functions on Riemannian manifolds. *Bonn INS Preprint No. 1626*, 2016.

[12] S. Hosseini and A. Uschmajew. A Riemannian gradient sampling algorithm for nonsmooth optimization on manifolds. *SIAM J. Optim.*, 27(1):173–189, 2017.

[13] V. Yu. Kaloshin. A geometric proof of the existence of Whitney stratifications. *Mosc. Math. J.*, 5(1):125–133, 2005.

[14] K. C. Kiwiel. Convergence of the gradient sampling algorithm for nonsmooth nonconvex optimization. *SIAM J. Optim.*, 18(2):379–388, 2007.

[15] D. Kressner, M. Steinlechner, and B. Vandereycken. Low-rank tensor completion by Riemannian optimization. *BIT*, 54(2):447–468, 2014.

[16] D. B. O'Shea and L. C. Wilson. Limits of tangent spaces to real surfaces. *Amer. J. Math.*, 126(5):951–980, 2004.

[17] R. Schneider and A. Uschmajew. Convergence results for projected line-search methods on varieties of low-rank matrices via Łojasiewicz inequality. *SIAM J. Optim.*, 25(1):622–646, 2015.

[18] M. Tan, I. W. Tsang, L. Wang, B. Vandereycken, and S. J. Pan. Riemannian pursuit for big matrix recovery. In *Proceedings of the 31st International Conference on Machine Learning (ICML 2014)*, pages 1539–1547, 2014.

[19] A. Uschmajew and B. Vandereycken. Greedy rank updates combined with Riemannian descent methods for low-rank optimization. In *2015 International Conference on Sampling Theory and Applications (SampTA)*, pages 420–424, 2015.

[20] B. Vandereycken. Low-rank matrix completion by Riemannian optimization. *SIAM J. Optim.*, 23(2):1214–1236, 2013.

[21] B. Vandereycken and S. Vandewalle. A Riemannian optimization approach for computing low-rank solutions of Lyapunov equations. *SIAM J. Matrix Anal. Appl.*, 31(5):2553–2579, 2010.

[22] H. Whitney. Local properties of analytic varieties. In *Differential and Combinatorial Topology*, pages 205–244. Princeton Univeristy Press, 1965.

[23] H. Whitney. Tangents to an analytic variety. *Ann. of Math. (2)*, 81:496–549, 1965.