SIAM J. OPTIM.
Vol. 29, No. 4, pp. 2853–2880

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.

Key words. Lipschitz function, descent direction, Clarke subdifferential, algebraic varieties, Riemannian manifolds, robust low-rank matrix recovery

AMS subject classifications. 49J52, 65K05, 14P05, 15A99

DOI. 10.1137/17M1153571

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 solving an optimization problem 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, the projection of the negative gradient of the cost function at a point onto the tangent cone generally cannot be used to determine a direction along which the function is decreasing. Instead, one has to work with some replacements for the gradient, called subdifferentials.

In this paper we consider the general problem

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

*Received by the editors October 24, 2017; accepted for publication (in revised form) August 16, 2019; published electronically November 7, 2019.
https://doi.org/10.1137/17M1153571

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

‡Hausdorff Center for Mathematics & Institute for Numerical Simulation, University of Bonn, 53115 Bonn, Germany. Current address: Max Planck Institute for Mathematics in the Sciences, 04103 Leipzig, Germany (uschmajew@mis.mpg.de).

where $f : \mathbb{R}^n \to \mathbb{R}$ is a locally Lipschitz function and $\mathcal{M} \subseteq \mathbb{R}^n$ is a real algebraic variety which we assume to be closed in $\mathbb{R}^n$. Being a variety, $\mathcal{M}$ admits a so-called Whitney stratification [23], that is, a decomposition

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

where $\mathcal{M}_s$ are mutually disjoint smooth submanifolds of $\mathbb{R}^n$ of different dimensions. The manifolds $\mathcal{M}_s$ will be called strata of $\mathcal{M}$, and we assume $\dim(\mathcal{M}_s) < \dim(\mathcal{M}_r)$ for $s < r$. While we have real algebraic varieties in mind, in particular varieties of low-rank matrices, we note that many of the subsequent considerations only use that the set $\mathcal{M}$ is a closed union of disjoint smooth manifolds. However, an exception is Lemma 3.2, which makes use of the additional Whitney a-regularity condition regarding limits of tangent planes (see [22, section 8] and [23, section 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), then it holds that $T_{\mathcal{M}_s}(\bar{x}) \subseteq T$.

We will present an algorithm for solving problem (1.1) by generalizing the Riemannian gradient sampling (GS) method from [12]. In general, given a closed convex cone $T(x)$ at x and $\varepsilon > 0$, we can define the set

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

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 classic 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. Our GS algorithm then 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 will be shown to be a descent direction.

Since we are dealing with constrained optimization, we are of course interested in descent directions in the Bouligand tangent cone $T_{\mathcal{M}}^B(x)$. Consequently, we will 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 descent direction with a certain step-size rule, followed by a retraction to get back on $\mathcal{M}$. The resulting GS algorithm is given as Algorithm 1 in section 3.2.

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, section 19] or [13] and references therein),

(ii) the existence of linear subspaces in the Bouligand tangent cones (in particular, the tangent spaces $T_{\mathcal{M}_s}(x)$ of strata belong to the tangent cone), and

(iii) the existence of retractions in the sense of section 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 $T(x_\ell)$ at iterates $x_\ell \in \mathcal{M}_{s_\ell} \subseteq \mathcal{M}$ 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 at least a critical point of f on $\mathcal{M}_s$, in the sense that $0 \in \partial f(x) + (T_{\mathcal{M}_s}(x))^\perp$ (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 $(T_{\mathcal{M}_s}(x))^\perp$, obtained as a limit of the normal spaces $N(x_\ell) = (T(x_\ell))^\perp$. Hence, if the sequence converges to $x \in \mathcal{M}_s$ from strata of larger dimension, $N(x)$ will have smaller dimension than $(T_{\mathcal{M}_s}(x))^\perp$. Our motivation for this general setup are varieties of low-rank matrices, where in rank-deficient points the Bouligand tangent cones contain many reasonable subspaces $T(x) \supsetneq T_{\mathcal{M}_s}(x)$, as explained further below.

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

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

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

Specifically, in section 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 corrupted entries (outliers). Besides their practical relevance, the sets $\mathcal{M}_{\leq r}$ are interesting examples for our framework because the Euclidean metric projection is explicitly available via singular value decomposition (SVD), which is somewhat exceptional for such nontrivial sets. Furthermore, the variety $\mathcal{M}_{\leq r}$ naturally stratifies into smooth components

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

of fixed rank $s = 1, \ldots, r$. In practice, problem (1.5) could be addressed by the Riemannian optimization on the regular part $\mathcal{M}_r$ as in [9, 10, 12], but the theory in these works is developed for *complete* Riemannian manifolds and does technically hence not apply due to the nonclosedness of $\mathcal{M}_r$. This affects the existence of retractions, the free choice of step-sizes in the tangent space, and the convergence results (existence of cluster points). The newly proposed method in this paper can be seen as an extension of the GS method from [12] for the manifold $\mathcal{M}_r$ to its closure. Even though rank-deficient points are perhaps never encountered in practical computations, our approach has the advantage of allowing rigorous convergence statements without assuming limit points being regular points, that is, points of full rank r; cf. the similar remarks in [17].

As will be explained in section 4.1.3, when $s < r$, the Bouligand tangent cone $T_{\mathcal{M}_{\leq r}}(X)$ at $X \in \mathcal{M}_s$ contains linear spaces strictly larger than the tangent space $T_{\mathcal{M}_s}(X)$. Such linear spaces can be used for increasing the rank back to r as outlined in section 4.1.4. More generally, this provides a framework for the derivation of rank-increasing matrix optimization methods, which make it necessary to consider rank-deficient starting guesses for (1.5), obtained, say, as a "solution" for rank $r-1$. While

mainly heuristic, such rank-increasing strategies have shown superior performance, e.g., for matrix completion [20]. In section 4.2.3 we confirm this for robust low-rank approximation, and we test the method for reconstruction of scratched grayscale images in section 4.2.4.

Outline. The paper is organized as follows: Section 2 is concerned with some preliminaries and definitions of nonsmooth analysis. In section 3 the GS algorithm on real algebraic varieties together with a convergence result is presented. Finally, in section 4 some numerical results for low-rank optimization are presented for illustration.

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

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

Then 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, for every x, ξ it holds that

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

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 at 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 particular at 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 $g(x) := -\operatorname{argmin}\{\|v\| : v \in \partial f(x)\}$ 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

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

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

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

Obviously, it holds that $f^\circ(x; \xi) \leq f_\varepsilon^\circ(x; \xi)$. Let $g \in \mathbb{R}^n$ and $\|g\| \leq 1$; then 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, and for obtaining the largest descent guarantee one should solve

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

The solution to this problem can be computed by solving

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

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 $\mathcal{M} \subseteq \mathbb{R}^n$ be a closed set. 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

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

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 defined as the polar of the Clarke tangent cone:

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

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

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

In particular, every local minimum of f 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 a point $x \in \mathcal{M}$ and a closed convex cone $T(x)$ are given. Let $N(x) = (T(x))^{\circ}$ denote its polar cone. For example, the choice $T(x) = T_{\mathcal{M}}^{C}(x)$ is feasible here, but $T(x) = T_{\mathcal{M}}^{B}(x)$ 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

$$\partial_N f(x) \coloneqq \partial f(x) + N(x)$$

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

$$\partial_N^{\varepsilon} f(x) \coloneqq \partial_{\varepsilon} f(x) + N(x).$$

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

$$0 \in \partial_N^{\varepsilon} f(x),$$

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 $T(x)$.

As a first step we claim that if x is not critical with respect to $N(x)$, then there exists $\varepsilon > 0$ such that x is also not ε-critical with respect to $N(x)$.

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

Proof. Suppose to the contrary that $0 \in \partial_N^{1/i} f(x)$ 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 assumption made. ∎

The next lemma relates the minimum norm element in $\partial_N^\varepsilon f(x)$ to projections onto the cone $-T(x)$.

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 $P_{-T} + P_{-T^\circ}$ is the identity map. Therefore, it holds that

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

which is the assertion. ∎

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

$$\min_{w \in \partial_N^\varepsilon f(x)} \|w\| \tag{2.7}$$

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

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

Here, the first equality is obtained by the minimax theorem, and the last equality is obtained by [17, Equation (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

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

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

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

which implies

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

We conclude that

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

As stated in (2.8) it holds that

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

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

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

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

3. A GS algorithm. In this section, the proposed GS 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 we emphasize that these properties are in particular 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 GS algorithm in the next subsection.

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

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

When $\mathcal{M}$ is a real algebraic variety with stratification (1.2), then every $x \in \mathcal{M}$ belongs to one of the manifolds $\mathcal{M}_s$, and the tangent space $T(x) = T_{\mathcal{M}_s}(x)$ is a possible choice for a linear space in the Bouligand tangent cone.

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 $T(x) \subseteq T^B_{\mathcal{M}}(x)$.) Furthermore, (3.2) is satisfied with $\kappa = 2$.

Remark 3.1. Let R be a retraction on $\mathcal{M}$. In the setting of 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 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. A possible procedure for this step is described at the end of section 3.2.
- 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. Every step requires finding 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 $\partial_N^\varepsilon f(x_\ell)$ at iterate x_ℓ with a set G_ℓ that is the convex hull of a finite number of projected gradients sampled in the set $x_\ell + \varepsilon_\ell B_{T(x_\ell)}(0,1)$, where $T(x_\ell)$ is 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.

Algorithm 1. Gradient sampling algorithm.

Input: $x_0 \in \mathcal{M} \cap D$; $\delta_0, \varepsilon_0, \gamma, \varepsilon_{\text{opt}}, \delta_{\text{opt}} \in (0,1)$; $\beta \in (0,1)$; $\theta_\varepsilon, \theta_\delta \in (0,1]$.

1 **for** $\ell = 0, 1, 2, \ldots$ **do**

2 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`

3 **if** $\{x_\ell^i\}_{i=1}^{m_\ell} \not\subset D$ **then**

4 **return**

5 **end**

6 Let $G_\ell \coloneqq \operatorname{conv}\{P_{T(x_\ell)}(\nabla f(x_\ell)), P_{T(x_\ell)}(\nabla f(x_\ell^1)), \ldots, P_{T(x_\ell)}(\nabla f(x_\ell^{m_\ell}))\}$, and find

$$w_\ell = \operatorname{argmin}\{\|w\| : w \in G_\ell\}.$$

7 **if** $\|w_\ell\| \le \delta_{opt}$ *and* $\varepsilon_\ell \le \varepsilon_{opt}$ **then**

8 **return**

9 **if** $\|w_\ell\| \le \delta_\ell$ **then**

10 $\varepsilon_{\ell+1} \coloneqq \theta_\varepsilon \varepsilon_\ell$, $\delta_{\ell+1} \coloneqq \theta_\delta \delta_\ell$

11 $x_{\ell+1} \coloneqq x_\ell$

12 **else**

13 $\varepsilon_{\ell+1} = \varepsilon_\ell$, $\delta_{\ell+1} = \delta_\ell$, $g_\ell \coloneqq -\dfrac{w_\ell}{\|w_\ell\|}$ `// descent direction`

14 $t_\ell \coloneqq \max\{t : f(R_{x_\ell}(tg_\ell)) - f(x_l) < -\beta t \|w_\ell\|, t \in \{1, \gamma, \gamma^2, \ldots\}\}$ `// line search`

15 **if** $R_{x_\ell}(t_\ell g_\ell) \in D$ **then**

16 $x_{\ell+1} \coloneqq R_{x_\ell}(t_\ell g_\ell)$

17 **else**

18 Find $x_{\ell+1} \in \mathcal{M} \cap D$ s.t. $f(x_{\ell+1}) - f(x_\ell) < -\beta t_\ell \|w_\ell\|$ `// stay in D`

19 and $\|R_{x_\ell}(t_\ell g_\ell) - x_{\ell+1}\| \le \kappa t_\ell$. `// κ from (3.2)`

20 **end**

21 **end**

22 **end**

We remark that the line search in line 14 of the algorithm is well-defined 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

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

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

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

In practice, the back-tracking procedure for line 14 quickly reaches machine precision and will hence have a bounded execution time. In the numerical experiments in section 4.2 we rarely observed a step-size below 10^{-7} (most often much larger), partly due to additional stopping criteria. An exception was the experiment in section 4.2.1 where "zero" step-sizes occurred. We refer to the discussion there.

Let us comment on the adjustment step in line 18 which ensures that the iterates remain in $\mathcal{M} \cap D$. First of all, since $\mathcal{M} \cap D$ is assumed to have full relative measure, one may not expect it ever to be necessary to execute this step, since the if clause in line 15 should never fail. However, we have no rigorous argument that the failure probability for line 15 is zero. If it becomes necessary, the following procedure could be applied to execute line 18: 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 $f(R_{x_\ell}(tg_\ell)) - f(x_l) < -\beta t_\ell \, \|w_\ell\|$, and since $\mathcal{M} \cap D$ is assumed to have full measure, this process will terminate after finitely many steps with probability one. We do not have an estimate of how many steps would be required. Also note that this procedure assumes that one is able to construct random points on $\mathcal{M}$.

3.3. Convergence result. We begin with a lemma regarding possible limiting subspaces of the sequence $T(x_\ell)$ used in the algorithm. For the second part of this lemma 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). *Assume that the sequence $(x_\ell) \subseteq \mathcal{M}$ is infinite and 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 $(x_\ell)_{\ell \in \mathcal{L}}$ converging to $\bar{x}$ such that the following conditions hold:

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

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

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 $\ell \in \mathcal{L}$, 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 $\tilde{m}$, 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, section 8], [23, section 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 sequence generated by Algorithm* 1 *with 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 latter case, if $\bar{x} \in \mathcal{M}$ is a cluster point and S a subspace satisfying items* (i) *and* (ii) *in Lemma* 3.2 *for an infinite subsequence $(x_\ell)_{\ell\in\mathcal{L}}$ converging to $\bar{x}$, then $\bar{x}$ is a critical point of f 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 $\bar{x} \in \mathcal{M}_s$ implies

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

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

Remark 3.4. The meaning of "with probability one" here is similar to previous results on GS algorithms [4, 14, 12]; see in particular [4, page 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 isomorphic 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 $m = 1, \ldots, n+1$ *before* we run the algorithm and that we then simply use these sample points in the algorithm whenever this dimension 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 a 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 GS 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 and 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 that $\sum_{\ell\in\mathcal{L}} \operatorname{dist}(x_\ell, x_{\ell+1}) = \infty$.*

Next, following [4], we define the sets

$$G_\varepsilon^S(x) := \operatorname{cl}\operatorname{conv}\{P_S(\nabla f(y)) : y \in (x + \varepsilon \operatorname{cl} B_S(0,1)) \cap D\}, \tag{3.3}$$

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

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

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

and

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

where

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

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 $G_\varepsilon^S(\bar{x})$ is compact, there exists $w \in G_\varepsilon^S(\bar{x})$ 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 nonnegative $\lambda_1, \dots, \lambda_m$ with $\sum \lambda_i = 1$ such that $u \coloneqq \sum_{i=1}^m \lambda_i P_S(\nabla f(\tilde{y}^i))$ satisfies $\|u\| \le \|w\| + \nu/3 = \rho_\varepsilon(\bar{x}) + \nu/3$. Now choose $\bar{\varepsilon}$ such that

(3.4)
$$\tilde{V} \coloneqq \prod_{i=1}^m \left(\tilde{y}^i + \bar{\varepsilon} B_S(0,1)\right) \subseteq D_{\varepsilon - \bar{\varepsilon}}^m(\bar{x}), \quad \text{and} \quad \left\| \sum_{i=1}^m \lambda_i P_S\left(\nabla f\left(y^i\right)\right) \right\| \le \rho_\varepsilon(\bar{x}) + \nu$$

holds for all $y = (y^1, \dots, y^m) \in \tilde{V}$. Set $\tau \coloneqq \bar{\varepsilon}$. Then, by (3.4), 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 Algorithm 1 with fixed $\varepsilon_\ell = \varepsilon_0 \coloneqq \varepsilon$ 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 $C = G_\varepsilon^S(\bar{x})$ (and β from the algorithm), and let τ and $\hat{V}$ be obtained from Lemma 3.7 for this ν. Assume further that $(x_\ell^1, \dots, x_\ell^m) \in \hat{V}(\bar{x}, \varepsilon, \tau)$ for all ℓ. Then, if $0 \notin G_\varepsilon^S(\bar{x})$, it must hold that $\liminf_{\ell \to \infty} t_\ell > 0$.*

Proof. Let us denote $x_\ell^0 \coloneqq x_\ell$. By assumption (i) from Lemma 3.2, $m_\ell = m$ is fixed and

$$w_\ell \coloneqq \sum_{i=0}^m \lambda_\ell^i P_{T(x_\ell)}\left(\nabla f\left(x_\ell^i\right)\right)$$

has the minimum norm at the 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,

$$-\beta \gamma^{-1} t_\ell \|w_\ell\| \le f\left(R_{x_\ell}\left(\gamma^{-1} t_\ell g_\ell\right)\right) - f(x_\ell). \tag{3.5}$$

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

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

Multiplying by $-\|w_\ell\|\gamma/t_\ell$ and using that $g_\ell \in T(x_\ell)$, we get from (3.5) that

$$\left\langle P_{T(x_\ell)}(v_\ell), w_\ell \right\rangle - \frac{o\left(\gamma^{-1}t_\ell\right)\|w_\ell\|}{\gamma^{-1}t_\ell} \leq \beta \|w_\ell\|^2 . \tag{3.6}$$

Since the tuples $(x_\ell^1, \dots x_\ell^m) \in \hat{V} \subseteq V_\varepsilon(\bar{x}, \bar{x}, \nu)$ 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

$$\min\left\{\|w\| : w \in P_S\left(\operatorname{conv}\left\{\xi^i\right\}_{i=1}^m\right)\right\} \leq \rho_\varepsilon(\bar{x}) + \nu. \tag{3.7}$$

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,

$$\min\left\{\|w\| : w \in P_S\left(\operatorname{conv}\left\{\xi^i\right\}_{i=0}^m\right)\right\} \leq \rho_\varepsilon(\bar{x}) + \nu, \tag{3.8}$$

because the minimum is taken over a larger set compared to (3.7).

Assume that the minimum in (3.8) 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, $P_S(\xi^i) \in G_\varepsilon^S(\bar{x})$, $i = 1, \dots, m$. Also, it is easy to see that $P_S(\partial f(\bar{x})) \subseteq G_\varepsilon^S(\bar{x})$, and therefore $P_S(\xi^0) \in G_\varepsilon^S(\bar{x})$. We conclude that $\tilde{w} \in G_\varepsilon^S(\bar{x})$ 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 $P_S(v) \in G_\varepsilon^S(\bar{x})$. 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 (3.6) 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 $i = 0, 1, \dots, m$ the sequence λ_ℓ^i converges to some λ_*^i. Then, by assumption (ii) from Lemma 3.2, w_ℓ converges to

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

We need to show that $w_* = \tilde{w}$. Since $\tilde{w}$ is the unique minimizer of (3.8), 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 that

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

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

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

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 $\|x_{\ell+1} - x_\ell\| \le 2\kappa t_\ell$. Using telescopic sums, this implies

$$\sum_{\ell=1}^{\infty} t_\ell \|w_\ell\| < \infty \quad \text{and} \quad \sum_{\ell=1}^{\infty} \|x_{\ell+1} - x_\ell\| \|w_\ell\| < \infty. \tag{3.9}$$

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 is a projection of a convex combination of nearby gradients onto $T(x_\ell)$), 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 of $(w_\ell)_{\ell\in\mathcal{L}}$, 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 (3.9) yields $\liminf_{\ell\in\mathcal{L}} \|w_\ell\| = 0$. If on the other hand $x_\ell \to \bar{x}$, then the existence of a subsequence of w_ℓ converging to zero is actually 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 $\ell \ge \ell^*$. This only happens if $\|w_\ell\| > \delta$ for $\ell \ge \ell^*$ (see line 9). By (3.9), 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 $0 \notin G_\varepsilon^S(\bar{x})$. Let ν, τ and $\hat{V} = \hat{V}(\bar{x}, \varepsilon, \tau)$ be chosen as in Theorem 3.8. Since $(x_\ell^1, \dots, x_\ell^m)$ are sampled independently and uniformly from $D_\varepsilon^m(x_\ell)$, and $\hat{V}$ is a nonempty open subset of $D_\varepsilon^m(x_\ell)$, it will hold that $(x_\ell^1, \dots, x_\ell^m) \in \hat{V}$ 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. 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

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

This is a contradiction, because both sequences of minima 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. Application to low-rank optimization. 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 entrywise ℓ_1 distance. This is sometimes referred to as robust low-rank recovery and is explained in section 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

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

fit perfectly in the abstract setting considered in this paper for several reasons.

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

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

into smooth manifolds

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

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

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

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, the tangent space at X is efficiently parametrized as follows:

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

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

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

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, section 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 are 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

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

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 $\mathcal{U}_\perp$ and $\mathcal{V}_\perp$, 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 did not observe an advantage.

4.1.4. On 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. When embedding the result X of a fixed-rank optimization, say, of rank $s < r$, as a starting guess on a variety of higher rank, say, on $\mathcal{M}_{\leq s+r_{\mathrm{incr}}}$, one is faced with the scenario considered in this paper, namely, selecting a linear subspace $T(X)$ in the Bouligand tangent cone $T_{\mathcal{M}_{\leq s+r_{\mathrm{incr}}}}(X)$. 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 orthogonal to row and column space of X, respectively) 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 $U^TF = 0$ and $V^TG = 0$—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 $\mathbb{R}^4$. For simplicity, we implemented a slightly different sampling. 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, then replacing F and G with normalized versions of $F - UU^TF$ and $G - VV^TG$, respectively, and returning

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

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 $H \in \mathbb{R}^{r_{\mathrm{incr}} \times r_{\mathrm{incr}}}$ with the aid of `randn` and return

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

with a_4 uniformly random in $[0, 1]$. The choice of the matrices $U_\perp$, $V_\perp$ has been explained in section 4.1.3.

Finally, the quadratic program in line 6 of Algorithm 1 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 be reformulated as finding $\zeta \in \mathbb{R}^{m+1}$ that minimizes $\zeta^T\bar{G}\zeta$ subject to the constraints $\zeta \geq 0$ and $\sum_{i=1}^n \zeta_i = 1$, 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, except for the first experiments) 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 observed that setting up the matrix $\bar{G}$ dominates the computational cost when many sample points in a relatively high-dimensional tangent space are given. For instance, we can report that in the experiments of section 4.2.2, in which different sampling numbers are compared, more than one third of the overall execution time of our implementation of Algorithm 1 was spent on assembling the Gram matrices $\bar{G}$ (according to MATLAB timing function). We note that in principle the entries of the Gram matrix could be computed fully in parallel. While we did not consider such an implementation for our numerical experiments, this may be a promising approach for making the GS algorithm more efficient in high dimensions.

In all the subsequent numerical experiments, the sampling radius is initialized with $\varepsilon_0 = 10^{-3}$. The other parameters are fixed as follows: $\delta_0 = 10^{-3}$, $\gamma = 2^{-1}$, $\varepsilon_{\text{opt}} = 10^{-6}$, $\delta_{\text{opt}} = 10^{-12}$, $\beta = 10^{-4}$, $\theta_\varepsilon = 10^{-1}$, and $\theta_\delta = 10^{-1}$. A minimal step-size of 10^{-10} was used. In all but the first experiment (section 4.2.1) additional stopping criteria were used: optimization for a fixed-rank r is terminated after some prescribed number of iterations or if $|f(X_{\ell+1}) - f(X_\ell)| < 10^{-10}$ three times in a row.

We note that with these parameters we observed in almost all our experiments for the problem (4.7) below that the sampling radius remains fixed during the iterations, that is, line 9 is almost never activated. This can have several problem dependent reasons, and correspondingly we cannot state that we really find stationary points (typically $\|w_\ell\|$ stagnated 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, $m = 2 \cdot \dim(\mathcal{M})$), an observation also made in [11]. We will give more explanations on this issue in our first experiment in section 4.2.1, where the goal will to be to reproduce the theoretical convergence result numerically.

4.2. Numerical results for robust low-rank approximation. By robust low-rank approximation one means the approximation or 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, section 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 continuously differentiable in the set D of all matrices X for which $A - X$ contains no zero entries. The gradient is then given as

$$\nabla f(X) = \operatorname{sign}(A - X). \tag{4.8}$$

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. Note that the gradient has a Frobenius norm of at least one at any nonzero X and therefore does not serve as an indicator for optimality. This illustrates why the ε-subdifferential is needed in nonsmooth optimization.

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

$$A = A_{\text{ex}} + \lambda E_{\text{noise}} + \mu E_{\text{out}}, \quad \|A_{\text{ex}}\|_F = \|E_{\text{noise}}\|_F = \|E_{\text{out}}\|_F = 1, \tag{4.9}$$

where A_{ex} is the assumed ground truth, E_{noise} is a dense matrix with random entries (modeling general noise in measurements), and E_{out} is a sparse matrix with 1% random nonzero entries (modeling outliers). All three matrices have Frobenius norm one (denoted by $\|\cdot\|_F$). Thus the scalars $\lambda, \mu \ge 0$ in (4.9) 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 problem (4.7) to outliers and demonstrate that it can in principle be solved using the GS algorithm on $\mathcal{M}_{\leq r}$. However, it is not our aim to make a specific claim regarding the potential applications, where it can be important to take further variations of the above problem, for instance, 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 notebook with a 2.9 GHz dual core CPU and 16 GB of memory, using MATLAB R2017a with Optimization Toolbox and a modified version of `manopt`.

4.2.1. Illustration of convergence result. In order to reproduce the convergence statement of Theorem 3.3 numerically, several things have to be taken into account. The fact that the statement is only on limit points of convergent subsequences is less of an issue, since, if at all, usually only one accumulation point is observed in computations. More crucial is to show that the limit is a critical point. The way in which the proof of Theorem 3.3 works is highly nonconstructive. Recall that a basic idea is that the sampling radius ε_ℓ should converge to zero; otherwise one obtains a contradiction. Numerically, we can reproduce this behavior only to a certain accuracy: once the sample radius falls below machine precision, only the gradient $\nabla f(x_\ell)$ itself will be sampled, and it will typically not go to zero near nondifferentiable minimizers (the formula (4.8) gives an indication for this in the considered application).

More severe, however, is the following fact: even if we fix the sampling radius ε and assume that 0 is in the convex hull of *all* projected gradients in an ε-ball around some given point x (similar to the set $G_\varepsilon^S(x)$ in (3.3)) and that the dimension of this convex hull is d, the probability that the convex hull of only $d+1$ randomly sampled points from it does contain the zero vector will become quite small when d becomes large. One key argument in the theoretical result is that this eventually has to happen by repeatedly sampling such $d+1$ random points, but in practice one may have to wait very long. The probability increases when the number of sample points is increased.

Given these considerations, the first numerical experiment is designed as follows. We set λ and μ in (4.9) to zero (no noise) and let A_{ex} be a random rank r matrix. In this way the global minimum $\bar{X} = A_{\text{ex}}$ of $f(X) = \|A_{\text{ex}} - X\|_{\ell_1}$ lies on $\mathcal{M}_r$ and is a nonsmooth point also with respect to the manifold $\mathcal{M}_r$.[1] We then run our implementation of the GS algorithm for the problem (4.7) with the given rank r of A_{ex} for the two different sample sizes

$$m = \dim(\mathcal{M}_r) + 1 \quad \text{and} \quad m = 2 \cdot \dim(\mathcal{M}_r).$$

The starting points for the algorithm are random but the same for both choices of m. We use $\delta_0 = \varepsilon_0 = 10^{-3}$ but do not allow the sampling radius to drop below $\varepsilon_{\text{opt}} = 10^{-6}$ (which ultimately results in a GS algorithm with fixed sampling radius). Since we use $\theta_\varepsilon = 10^{-1}$, this means that line 9 in Algorithm 1 has to be activated three times (which due to $\theta_\delta = 10^{-1}$ corresponds to events $\|w_\ell\| \leq 10^{-1}$, $\|w_\ell\| \leq 10^{-2}$, and $\|w_\ell\| \leq 10^{-3}$) before the algorithm has a chance to terminate through line 7 as soon as

[1] This can be seen by noting that $\bar{X}$ is certainly a nonsmooth point of f on any linear subspace containing $\bar{X}$. Many such subspaces exist which are (locally) subsets of $\mathcal{M}_r$, for instance, the space of all matrices whose column space is contained in the column space of $\bar{X}$.

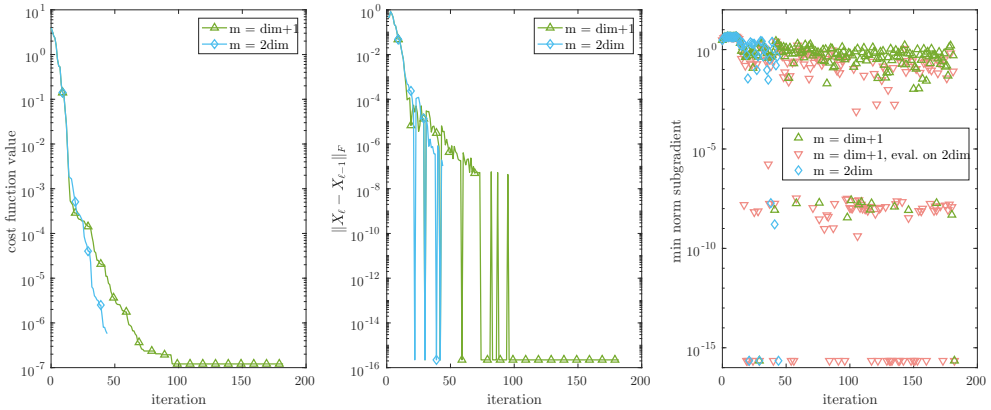

FIG. 1. *Typical outcome of GS algorithm for problem* (4.7) *with* $A \in \mathbb{R}^{5\times 5}$ *being a matrix of rank* $r = 1$. *Here* $\dim(\mathcal{M}) = 9$. *Two sampling sizes* $m = 10$ *and* $m = 18$ *have been tested. Left: cost function values* $\|A - X_\ell\|_{\ell_1}$. *Middle: Frobenius norms of differences between iterates. Right: Euclidean norm of the vector* $\|w_\ell\|$ *obtained in line* 6 *of Algorithm* 1. *The red values have been obtained by sampling eight more gradients in the algorithm for* $m = 10$, *but these have not been used for obtaining the descent direction. In the middle and right plot the values on the bottom indicate zero (machine precision has been artificially added).*

$\|w_\ell\| \le \delta_{\text{opt}} = 10^{-12}$ occurs. We plot the cost function values $f(X_\ell) = \|A_{\text{ex}} - X_\ell\|_{\ell_1}$, the difference in iterates $\|X_{\ell+1} - X_\ell\|_F$ (in Frobenius norm), and the lengths $\|w_\ell\|$ of shortest subgradients found in line 6 of the algorithm. For the sample sizes $m = \dim(\mathcal{M}_r) + 1$ we also plot the length of the shortest vectors that would have been found when sampling $2 \cdot \dim(\mathcal{M}_r)$ gradients, but these were not used as a descent direction. In this experiment we changed the default options of `quadprog` by setting `OptimalityTolerance` to 10^{-16} (default is 10^{-8}) and `StepTolerance` to zero (default is 10^{-12}).

Figure 1 shows the result of such an experiment with $M = N = 5$ and rank $r = 1$. Here, the dimensionality is relatively low, namely, $\dim(\mathcal{M}_1) = 9$. It can be seen that for both sample sizes, the GS method eventually terminated through line 7 of the algorithm,[2] that is, with an "optimal" point satisfying $\|w_\ell\| \le 10^{-12}$ and $\varepsilon_\ell \le 10^{-6}$. However, as the many triangles at the bottom of the right plot show, the method with $\dim(\mathcal{M}) + 1$ sample points actually found ε-critical points several times before, but without detecting it (since too few gradients were sampled). Further, comparing the left and the middle plot it can be seen that for both sample sizes the differences $X_{\ell+1} - X_\ell$ (and in consequence also $f(X_\ell) - f(X_{\ell+1})$) have been almost zero at some single iteration numbers, although decrease of the cost function was still possible in subsequent iterates. While in this experiment we ignored such occurrences and did not use them as a stopping condition (to ensure termination through the optimality criterion in line 7),[3] in the later experiments the algorithm will be terminated if function values stagnate a certain number of times in a row.

In higher dimension, things become more challenging. For instance, for $M = N = 10$ and rank $r = 2$ (then $\dim(\mathcal{M}_r) = 36$) we consistently observed that the method

[2] In most experiments with these parameters, the algorithm with $m = \dim(\mathcal{M}_1)+1$ sample points needed more than 200 iterates to terminate. For convenience we are presenting here a plot where this was not the case.

[3] However, we also did not check the optimality of these points (e.g., by repeatedly sampling gradients).

using $m = 2\dim(\mathcal{M}_r)$ sampled gradients would still succeed, while the method with $\dim(\mathcal{M}_r) + 1$ would find an optimal point but never detect this within 200 iterations. For even larger dimensions also $2\dim(\mathcal{M}_r)$ sample points would not suffice to technically detect optimal critical points within a reasonable number of iterations.

4.2.2. Influence of sample size. For obtaining the rigorous convergence results in the first part of the paper, it was crucial that at least $\dim(T(x))+1$ nearby gradients were sampled in addition to the one at the current iterate. At nonsingular 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 expensive. For optimization on low-rank matrix manifolds we observed that this issue happens already for medium sized matrices 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.

In a given application, however, it is not clear whether so many sample points are really necessary to achieve a desired goal. The opposite extreme is just taking $m = 0$ sample points, in which case the method reduces to the (Riemannian) steepest descent method. In the second experiment we aim to investigate the influence of the sample size on the performance for robust low-rank approximation. We run our implementation of the GS algorithm for problem (4.7) with $M = N = 30$ and target rank $r = 3$ for different sample sizes

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

Specifically, $\dim(\mathcal{M}_3) = 171$ here.

As a first scenario, we create the matrix A_{ex} as a random rank three matrix and only add the sparse outliers matrix E_{out} (it has nine nonzero entries) with factor $\mu = 0.9$ but no random noise ($\lambda = 0$). Then a possible outcome of the algorithm is given in Figure 2, where we plot the cost function values $\|A - X_\ell\|_{\ell_1}$, the Frobenius errors $\|A_{\text{ex}} - X_\ell\|_F$, as well as the execution times. The latter are given to make

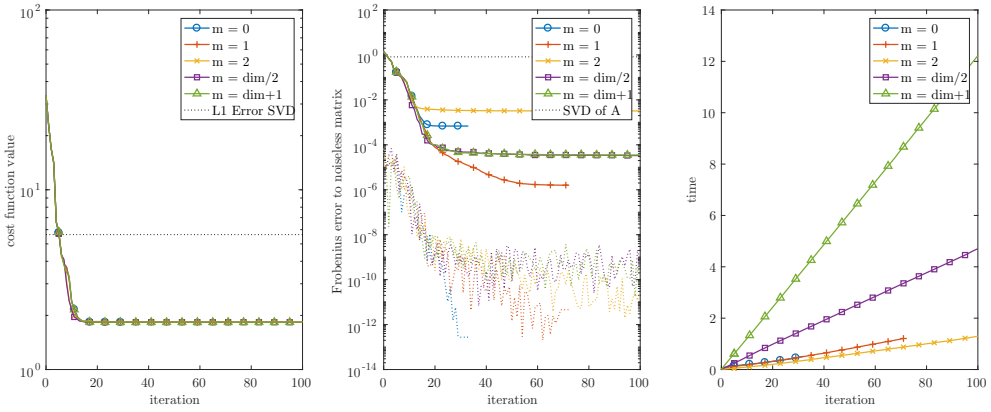

FIG. 2. *Results of GS algorithm for problem* (4.7) *with* $A = A_{\text{ex}} + 0.9E_{\text{out}} \in \mathbb{R}^{30\times 30}$ *and* $r = 3$*, tested for different sampling numbers* m*. Here* $\dim(\mathcal{M}_r) = 171$*. Matrix* A_{ex} *is of exact rank* r*. Left: cost function values* $\|A - X_\ell\|_{\ell_1}$*. Algorithms were terminated after* 100 *iterations or if* $|f(X_\ell) - f(X_{\ell+1})| \le 10^{-10}$ *three times in a row. The dotted line is the* ℓ_1 *error of the rank-*r *truncation of the SVD of* A*. Middle: Frobenius errors* $\|A_{\text{ex}} - X_\ell\|_F$ *and minimum absolute entry* $A_{\text{ex}} - X_\ell$*. Right: computing time in seconds.*

a relative comparison between sampling sizes; we did not aim for the most efficient implementation. In the left plot, we can see that all methods basically reach the same value for the ℓ_1 cost function and that this happened over all instances. Regarding the actual recovery of the original matrix A_{ex}, however, different outcomes were possible. Most often, all the methods were able to recover A_{ex} considerably below the level $\mu = 0.9$ of outliers. Specifically, a better recovery than using SVD of A was achieved. The methods with sampling sizes $m = \frac{1}{2}\dim(\mathcal{M}_r)$ and $m = \dim(\mathcal{M}_r) + 1$ were consistently observed to recover A_{ex} with an error of about 10^{-4} in Frobenius norm. The results for the methods with $m = 0, 1, 2$ did not seem to follow a particular pattern. Over several trials, each of them would sometimes yield the best, sometimes the worst approximation of A_{ex}. The selected plot in Figure 2 (middle) reflects this, in which the method with $m = 1$ was the best of all. Sometimes, all five methods only reached a recovery of 10^{-2}.

Since the norms $\|A - X_\ell\|_{\ell_1}$ are quite high in all cases, it might be that the minimizers are not really nonsmooth points of the problem. To check this we show in the dashed lines in the middle plot the minimum absolute value of the elements of $A - X_\ell$. They are close to zero, so X_ℓ are close to nondifferentiable points of f (which makes it likely that they are also close to nondifferentiable points on $\mathcal{M}_r$).

In a more realistic scenario, the matrix A_{ex} is not exactly of low rank. We provide a second result for this case by generating a matrix A_{ex} with exponentially decaying singular values (from 10^1 to 10^{-30} before normalization) and adding some noise of level $\lambda = 10^{-5}$ in (4.9). In this setup we consistently observed that a larger sampling size has basically no effect on the achievable result, while it considerably increases the computational cost. Most often, all five methods find an essentially optimal (in Frobenius norm) rank-r approximation of A_{ex}. In Figure 3 we selected a plot in which at least the methods with $m = 0$ and $m = 1$ took a bit longer, but this was not the rule. We also note that frequently some of the methods or even all of them stagnated at a suboptimal recovery of A_{ex} with an error in Frobenius norm of about 10^{-2}.

While these experiments demonstrate that ℓ_1-minimization is more robust to the influence of outliers than minimization of Frobenius norm, we draw a mixed conclusion regarding the required number of sample points for the GS algorithm: while it does not pay off using the $\dim(\mathcal{M}) + 1$ sample points as required by theory (and is simply

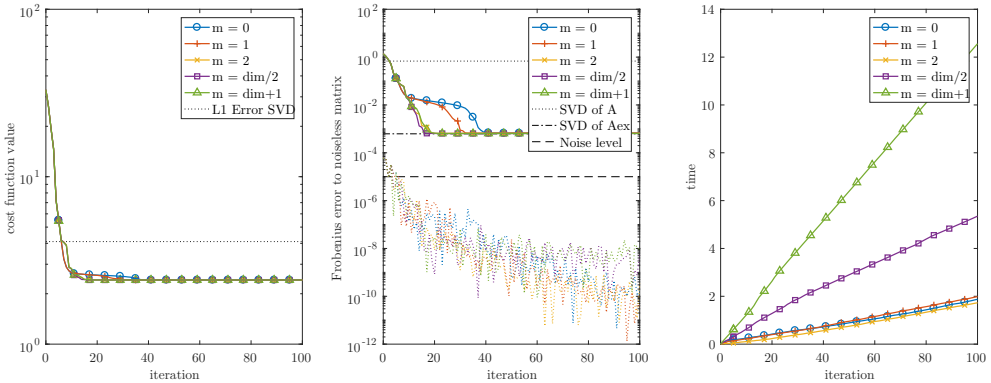

FIG. 3. *Same parameters as in Figure 2, but now* $A = A_{\text{ex}} + 10^{-5}E_{\text{noise}} + 0.9E_{\text{out}} \in \mathbb{R}^{30\times 30}$, *and* A_{ex} *is of full rank with exponentially decaying singular values. The middle plot also displays the Frobenius error of a best rank-r approximation of* A_{ex} *(dotted dashed line). As one can see, all algorithms find an essentially optimal rank-r approximation to* A_{ex}.

not possible when $\dim(\mathcal{M})$ is large), adding a few sampled gradients can be useful. It appears reasonable to choose the sample size 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 $2r$. This allows us to apply the GS algorithm to larger low-rank matrices.

4.2.3. Rank-increasing algorithm. In this experiment, the rank-increasing strategy described in section 4.1.4 is put to the test. We create a matrix A of the form (4.9). The matrix A_{ex} is generated as a dense full rank matrix of size $M \times N$, with singular 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 $\mu \in \{0, 0.5, 1\}$ for the outliers are tested.

Given A, we aim to compute a low-rank approximation of A_{ex} by solving a sequence of robust low-rank approximation problems (4.7) with increasing ranks, using the GS algorithm. We start with a random matrix of rank $s = 1$, iterate 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 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. section 4.1.4). Additionally we run the rank-increasing strategy using the GS algorithm with $m = 0$ on every manifold $\mathcal{M}_s$, that is, just using the negative projected gradient as descent direction without additional sampling (Riemannian steepest descent). For this algorithm the random subspace augmentation is used.

The rank-increasing strategy is compared with a fixed-rank approach by running the GS algorithm directly on the target manifold $\mathcal{M}_r$, using either a random point $X_0 \in \mathcal{M}_r$ or a best rank-r approximation of A in Frobenius norm (obtained from truncated SVD) as an initial guess. For the random point X_0, we additionally test the Riemannian steepest descent method on $\mathcal{M}_r$, that is, use our GS algorithm with $m = 0$ additional samples.

Figure 4 shows a result for $M = N = 100$, target rank $r = 21$, rank-increase by $r_{\text{incr}} = 2$, and at most ten iterations per rank. In this case, the best possible rank-21 approximation error to A_{ex} in Frobenius norm is around $4 \cdot 10^{-4}$. The matrix E_{out} has 100 nonzero entries. Running the same experiment over several instances, we consistently observed that for all three magnitudes μ of outliers none of the six methods ever terminated through the criterion $\|w_\ell\| \leq 10^{-12}$ in line 7 of Algorithm 1, yet the rank-increasing algorithms with sampling were able to essentially recover a best rank-21 approximation of A_{ex} in Frobenius norm. 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, say, by at least an order of magnitude, as is the case here (10^{-5} vs. 10^{-4}). When λ was larger, the methods would naturally only find an approximation of accuracy λ. We note that in very rare cases, the strategy using subspace augmentation from the projected gradient became stationary at a suboptimal level. For the rank-increasing algorithm without additional sampling ($m = 0$), an optimal recovery of A_{ex} was sometimes observed (as seen in the middle plot) but remained more of an exception. In many other cases (not displayed), a final error between 10^{-3} and 10^{-2} was reached for this method.

In comparison, the fixed-rank methods consistently stagnated at suboptimal values, an exception being of course the case $\mu = 0$ with optimal starting guess from the truncated SVD of A. In general, the algorithm starting from this starting point would stagnate around the initial error for all values of μ, which maybe indicates

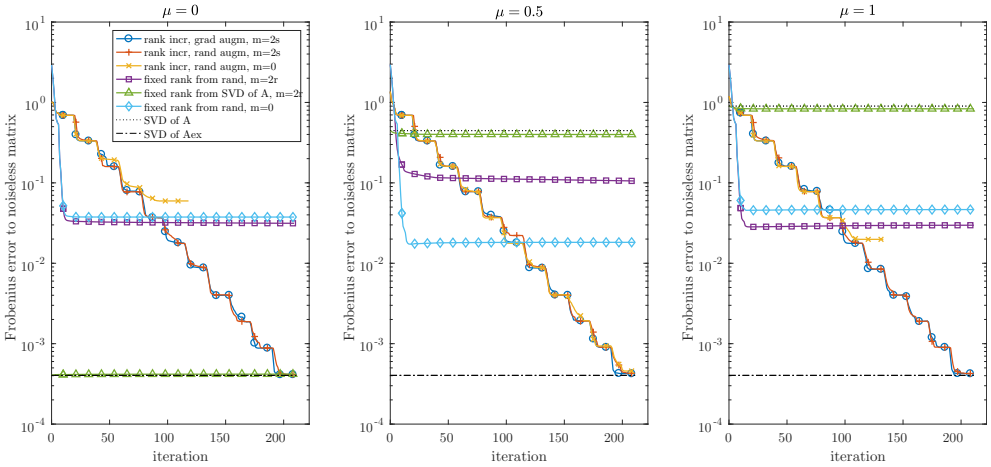

FIG. 4. *Rank-increasing strategy versus fixed-rank approach for problem* (4.7) *with* $r = 21$ *and* $A \in \mathbb{R}^{100\times 100}$ *of the form* (4.9) *with* $\lambda = 10^{-5}$*, tested for different magnitudes* $\mu \in \{0, 0.5, 1\}$ *of outliers. One can observe a staircase behavior where on every step the rank was increased by* $r_{incr} = 2$ *using two different subspace augmentation strategies. In the left plot, the dotted and dash-dotted line are (almost) on top of each other, since* A *and* A_{ex} *only vary by* $\lambda = 10^{-5}$*. The yellow and light blue curves were obtained without any GS (Riemannian steepest descent). The algorithm for a given rank was terminated if* $|f(X_\ell) - f(X_{\ell+1})| \leq 10^{-10}$ *three times in a row. This happened frequently for the yellow curve for all three values of* μ*.*

that this is also a critical point for the ℓ_1 distance function. When sparse outliers are present, a starting guess from the SVD of A is hence not recommended, as it takes too much false information from the outliers into account. The comparison of the two fixed-rank approaches (with or without sampling) with random starting guess was less predictable, and Figure 4 is an example for this. Frequently, both methods produced very close error curves for every sparse noise level μ.

It is important to note here that the rank-increasing strategy is particularly suited for fast decaying singular values. We observed that it is not competitive at all to fixed-rank optimization when the singular values of A_{ex} were not exponentially decaying but generated randomly, not even in case when the rank of A_{ex} was set to equal the target rank r. In such scenarios we observed that the error would only decrease significantly once the rank s was close to the target rank r. A possible explanation is that only in the case of fast decaying singular values an "optimal" rank-s approximation provides a significantly better starting guess for finding an optimal rank-$(s+r_{incr})$ approximation than a random point on $\mathcal{M}_{s+r_{incr}}$.

4.2.4. Inpainting. In the fourth 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` (Figure 5 left). As matrix A, we take a scratched version of this image (Figure 5 middle). With matrix A as an input, we conduct exactly the same experiment as in section 4.2.3. 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 limited to ten. The sampling size is again $2s$ for rank-s optimization and $2r$ for the fixed-rank methods. In Figure 5 (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.

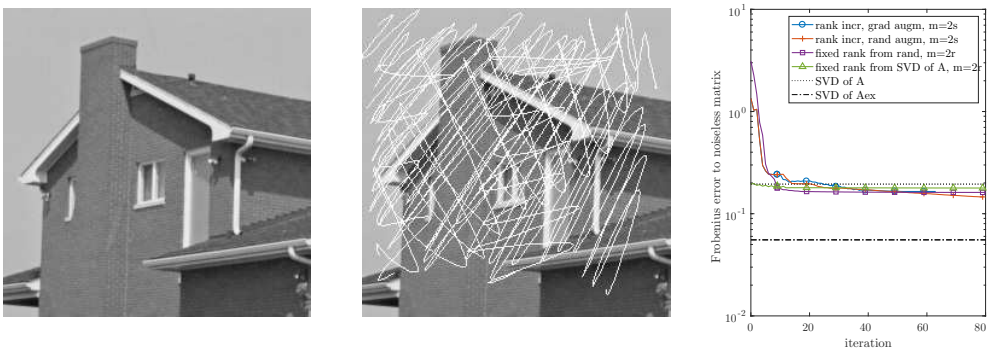

FIG. 5. *The original test image* `house.png` *is on the left. The scratched version in the middle is used as the input A for the GS algorithm (after normalization to Frobenius norm one). Right: error history for the four variants of GS algorithm, showing the Frobenius distance to (a normalized version) of the original image.*

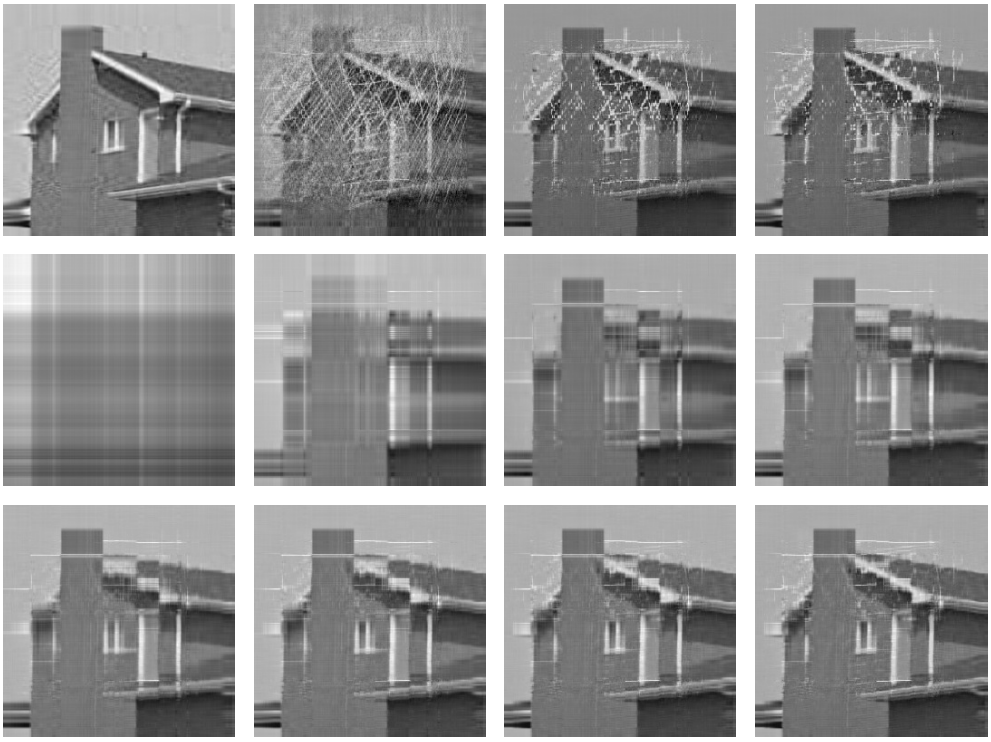

FIG. 6. *Low-rank approximations of* `house.png`*. Top line (from left to right): SVD truncation of the original image to rank $r = 22$, truncation of the scratched version, and the results of the two fixed-rank GS methods (random starting guess and SVD starting guess) for this rank r. Middle and top line: intermediate results of the rank-increasing algorithm (with random subspace augmentation) for the ranks* $1, 4, 7, 10, 13, 16, 19, 22$.

The outcome in terms of image reconstruction is nevertheless very different as can be seen in Figure 6. 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 image A, and the results of the two fixed-rank

GS methods (with random starting guess and SVD starting guess). The two other rows in Figure 6 show the intermediate results for the rank-increasing algorithm with the random subspace augmentation (red curve in Figure 5). 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 GS algorithm for minimization of locally Lipschitz functions on subvarieties of Euclidean space that admit stratifications into smooth submanifolds. The new method is considerably simpler than previous attempts since it only requires sampling in linear subspaces and no vector transport. Furthermore, the method can deal with singular points of the stratification using linear subspaces in the Bouligand tangent cone. We provided convergence results that are as strong as the analogous results for GS algorithm in linear space. The varieties of low-rank matrices constitute an important example for the considered setting, where the nontrivial 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, Cham, 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 (2014), pp. 1455–1459, http://www.manopt.org.

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

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

[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 (2013), pp. 1863–1882.

[7] F. H. Clarke, *Optimization and Nonsmooth Analysis*, 2nd ed., SIAM, Philadelphia, PA, 1990.

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

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

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

[11] S. Hosseini, W. Huang, and R. Yousefpour, *Line search algorithms for locally Lipschitz functions on Riemannian manifolds*, SIAM J. Optim., 28 (2018), pp. 596–619.

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

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

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

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

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

[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 (2015), pp. 622–646.

[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), 2014, pp. 1539–1547.

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

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

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

[22] H. Whitney, *Local properties of analytic varieties*, in Differential and Combinatorial Topology (A Symposium in Honor of Marston Morse), Princeton University Press, Princeton, NJ 1965, pp. 205–244.

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