A Nonmonotone Descent Method for Optimization Problems Defined by Upper-$\mathcal{C}^2$ Functions over Submanifolds

Christian Kanzow* Leo Lehmann†

Tuesday 26th May, 2026

Abstract

We consider the optimization problem of minimizing a nonsmooth function characterized by a nonsmooth formulation of the descent lemma over a manifold. In the unconstrained case over a Euclidean space, this class of functions is called upper-$\mathcal{C}^2$. Using the recent notion of projectional subdifferentials, we show that their descent property carries over to submanifolds. We propose a nonmonotone subgradient method to solve these problems and prove stationarity of accumulation points of the generated sequence as well as convergence and rate-of-convergence results under the Kurdyka-Łojasiewicz property. We also perform numerical experiments and show how our approach can be applied to a certain type of difference of convex functions as well as clustering problems on manifolds.

1 Introduction

In this paper, we consider the optimization problem

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

where $\mathcal{M} \subseteq \mathcal{E}$ is an embedded Riemannian submanifold of some Euclidean space $\mathcal{E}$. The objective φ satisfies a nonsmooth (local) formulation of the descent lemma, which holds in particular if φ is a function from $\mathcal{E}$ to $\mathbb{R}$ and upper-$\mathcal{C}^2$. This class of functions has been recently recognized as being suitable for extensions of linesearch methods to the nonsmooth setting [2]. In particular, all functions satisfying the usual smooth descent lemma are upper-$\mathcal{C}^2$, but nonsmooth examples also arise directly in applications. In particular, problems of type (1.1) arise frequently in clustering tasks on manifolds with applications in natural sciences, text processing as well as image, video and time series analysis [10, 17, 21, 23, 25, 31, 40, 48, 50].

In the Euclidean setting, a descent method to find a (local) minimum of a continuously differentiable function φ is defined by the iteration

$$x^{k+1} := x^k + \tau_k d^k \tag{1.2}$$

*University of Würzburg, Institute of Mathematics, Emil-Fischer-Str. 30, 97074 Würzburg, Germany; e-mail: christian.kanzow@uni-wuerzburg.de

†University of Würzburg, Institute of Mathematics, Emil-Fischer-Str. 30, 97074 Würzburg, Germany; e-mail: leo.lehmann@uni-wuerzburg.de

where d^k is a descent direction and τ_k is a stepsize. For optimization problems on nonlinear spaces, i.e. manifolds, given a so-called retraction R, the iteration above can be generalized to

$$x^{k+1} = \mathrm{R}_{x^k}(\tau_k d^k).$$

If φ is smooth and d^k is the negative of the Riemannian gradient, the iteration is the so called Riemannian gradient descent method. In our case, besides the choice of d^k as a negative subgradient, our framework will be general enough to allow for d^k to be a Newton or quasi-Newton direction.

The choice of the stepsize τ_k plays an important role for the convergence of the algorithm. Using a backtracking linesearch, τ_k has to be chosen such that a suitable line search criterion holds like the Armijo rule [6]. The Armijo condition, however, enforces a descent in the objective function value in every iteration. In contrast, nonmonotone methods allow for controlled increases in the objective function value. Allowing nonmonotonicity often leads to fewer backtracking iterations and thus larger stepsizes, which is the reason why nonmonotone methods possibly perform better in practice [2, 11, 20].

There are essentially two prominent nonmonotone stepsize rules: The max-type rule by Grippo et al. [24] and the mean-type rule by Zhang and Hager [51]. The work [2], which motivated our study of upper-$\mathcal{C}^2$ optimization problems, considers the max-type rule. By employing the mean-rule in our subgradient method later, the global convergence results can be derived without much of the technical difficulties that arise due to the max-rule. Further, the different flavor of nonmonotonicity allows for stronger theoretical guarantees to be derived, namely convergence and convergence rates under the Kurdyka-Lojasiewicz property. To the best of our knowledge, these results are new even in the Euclidean case.

Linesearch methods are also popular in the context of manifold optimization. The nonmonotone mean-type rule in the smooth setting is studied in [42, 44]. In order to ensure convergence, these methods assume some Lipschitz-type conditions [15], which in the Euclidean setting are equivalent to Lipschitz continuity of the gradient or the Lipschitz smoothness property. We propose a more general nonsmooth condition. Similar to the Lipschitz-type conditions, using so-called projectional subdifferentials recently considered in [34], we show that the descent property characterizing the class of upper-$\mathcal{C}^2$ functions can be transferred from the ambient Euclidean space $\mathcal{E}$ to submanifolds.

The paper is organized in the following way. In Section 2, we introduce some concepts from variational analysis and optimization on manifolds. We further introduce the class of upper-$\mathcal{C}^2$ functions and show that these functions restricted to a submanifold satisfy a certain descent property. In Section 3, we present our nonmonotone descent method to solve (1.1) and discuss its (global) convergence properties. Subsequently, in Section 4, we obtain further convergence guarantees under the Kurdyka-Lojasiewicz condition. Next, in Section 5, we are concerned with some applications and conduct corresponding numerical experiments. We conclude with some final remarks in Section 6.

Notation: We will consider optimization problems on some submanifold $\mathcal{M} \subseteq \mathcal{E}$, where $\mathcal{E}$ is a Euclidean space (finite-dimensional Hilbert space). Throughout this manuscript, we identify the dual space $\mathcal{E}^*$ with $\mathcal{E}$ itself. We write $\langle x, y\rangle$ for the scalar product of two elements $x, y \in \mathcal{E}$ and $\|x\|$ denotes the induced norm of $x \in \mathcal{E}$. The distance of a point $x \in \mathcal{E}$ to a nonempty set $S \subseteq \mathcal{E}$ is denoted by $\operatorname{dist}(x, S) := \inf_{y\in S} \|x - y\|$. Further, for $x \in \mathcal{E}$ and a nonempty closed set $S \subseteq \mathcal{E}$, the projection of x onto S is denoted by $\operatorname{proj}_S(x) := \{y \in \mathcal{E} \mid \operatorname{dist}(x, S) = \|x - y\|\}$. If S is nonempty, closed and convex, $\operatorname{proj}_S(x)$ is single-valued for all $x \in X$. Also, we denote by δ_S the indicator function of S, which is 0 for arguments in S and ∞ else.

We write $B_r(x)$ for the open and $\overline{B}_r(x)$ for the closed ball of radius $r > 0$ around $x \in \mathcal{E}$. For arbitrary $S \subseteq \mathcal{E}$, we denote its interior by $\mathrm{int}(S)$, the closure by $\mathrm{cl}(S)$ and the convex hull by $\mathrm{conv}(S)$.

Finally, $\mathbb{R}$ denotes the set of real numbers, while $\overline{\mathbb{R}} := (-\infty, +\infty]$ is the set of extended reals except that we exclude the value $-\infty$. Given an extended-valued function $\theta : \mathcal{E} \to \overline{\mathbb{R}}$, we call $\mathrm{dom}(\theta) := \{x \in \mathcal{E} \mid \theta(x) < \infty\}$ the domain of θ. The function θ is said to be proper if $\mathrm{dom}(\theta)$ is nonempty.

2 Background Material

2.1 Basics from Variational Analysis

We first recall that a sequence $\{x^k\} \subseteq \mathcal{E}$ converges (locally) *Q-linearly* to some limit $x^* \in \mathcal{E}$ if there exists a constant $c \in (0, 1)$ such that $\|x^{k+1} - x^*\| \leq c\|x^k - x^*\|$ holds for all $k \in \mathbb{N}$ sufficiently large. Furthermore, we say that $\{x^k\}$ converges *R-linearly* to x^* if $\limsup_{k\to\infty} \|x^k - x^*\|^{1/k} < 1$ holds. Note that this property holds if there exist constants $\omega > 0$ and $\mu \in (0, 1)$ such that $\|x^k - x^*\| \leq \omega\mu^k$ for all $k \in \mathbb{N}$ sufficiently large, i.e., if the sequence $\|x^k - x^*\|$ is dominated by a Q-linearly convergent null sequence.

We next recall some results from variational analysis and refer the interested reader to the two monographs [39, 46] for more details.

Given a proper, lower semicontinuous function $g : \mathcal{E} \to \overline{\mathbb{R}}$ and any $x \in \mathrm{dom}(g)$, we call

$$\hat{\partial} g(x) := \Big\{ v \in \mathcal{E} \;\Big|\; \liminf_{y \to x, y \neq x} \frac{g(y) - g(x) - \langle v, y - x\rangle}{\|y - x\|} \geq 0 \Big\}$$

the *regular* or *Fréchet subdifferential* of f at x, whereas

$$\partial g(x) := \big\{ v \in \mathcal{E} \mid \exists x^k, v^k \in \mathcal{E} : x^k \to x, g(x^k) \to g(x), v^k \in \hat{\partial} g(x^k)\ \forall k \big\}$$

is called the *limiting, Mordukhovich*, or *basic subdifferential* of g at x.

For a locally Lipschitz continuous function g around $\bar{x} \in \mathbb{R}$, we define the *Clarke subdifferential* of g at $\bar{x}$ as

$$\partial_C g(\bar{x}) := \{ v \in \mathcal{E} \mid \langle v, h\rangle \leq g^\circ(\bar{x}; h) \text{ for all } h \in \mathbb{R}^n \},$$

where

$$g^\circ(\bar{x}; h) := \limsup_{x \to \bar{x}, t \to 0^+} \frac{g(x + th) - g(x)}{t}$$

is the *Clarke directional derivative* of g at $\bar{x}$ in the direction h. Note that $\partial_C g(\bar{x}) = \mathrm{conv}(\partial g(\bar{x}))$ holds for all locally Lipschitz continuous functions. Further, if g is locally Lipschitz continuous, then the Clarke subdifferentials of g on bounded subsets of $\mathcal{E}$ remain bounded, too, see e.g. [19]. By consequence, as $\partial g(\bar{x}) \subseteq \partial_C g(\bar{x})$ holds, the same is true for the limiting subdifferential.

Let us also recall some notation from variational geometry from [46]. For a set $C \subset \mathcal{E}$, we define the *tangent cone* of C at $\bar{x} \in C$ as

$$T_C(\bar{x}) := \Big\{ d \in \mathcal{E} \;\Big|\; \exists \{x^k\} \subset C, t_k \searrow 0, x^k \to \bar{x}, \frac{x^k - \bar{x}}{t_k} \to d \Big\}.$$

Further, the *regular normal cone* is given by

$$\hat{N}_C(\bar{x}) = T_C(\bar{x})^\circ := \{ n \in X \mid \langle n, d\rangle \leq 0 \forall d \in T_C(\bar{x}) \},$$

and the *limiting normal cone* is defined as

$$N_C(\bar{x}) = \{n \in \mathcal{E} \mid \exists \{x^k\} \subset C, \{n^k\} : x^k \to \bar{x}, n^k \in \hat{N}_C(\bar{x}), n^k \to n\}.$$

Before we proceed by stating some basic concepts and results about manifolds, we finally introduce the so-called Kurdyka–Łojasiewicz property. The following definition is a generalization of the classical one for nonsmooth functions, as introduced in [8, 12, 13]. The KL property plays a central role in the local convergence analysis and rate-of-convergence results of several algorithms for the solution of nonsmooth minimization problems, see e.g. [7, 44].

Definition 2.1. *Let $g : \mathcal{E} \to \overline{\mathbb{R}}$ be lower semicontinuous. We say that g satisfies the* Kurdyka–Łojasiewicz (KL) property *at $x^* \in dom(\partial g) := \{x \in \mathcal{E} \,|\, \partial g(x) \neq \emptyset\}$ if there exists a constant $\eta > 0$, a neighborhood $U \subset \mathcal{E}$ of x^*, and a continuous and concave function $\chi : [0, \eta] \to [0, \infty)$, called* desingularization function*, which is continuously differentiable on $(0, \eta)$ and satisfies $\chi(0) = 0$ and $\chi'(t) > 0$ for all $t \in (0, \eta)$, such that the so-called KL inequality*

$$\chi'\big(g(x) - g(x^*)\big)\mathrm{dist}\big(0, \partial g(x)\big) \geq 1 \tag{2.1}$$

holds for all $x \in U \cap \{x \in \mathcal{E} \,|\, g(x^*) < g(x) < g(x^*) + \eta\}$. Furthermore, we call g a KL function *if g satisfies the KL property at any point $x^* \in dom(\partial g)$.*

2.2 Essentials about Optimization on Submanifolds

As noted in [15], optimization on manifolds is not fundamentally different from optimization in a Euclidean space and many popular methods for unconstrained optimization have been generalized to optimization on manifolds. However, to do so, we require some preliminaries. For a broader overview, we refer to [1], [27] and [14].

Let us recall the following definition of a smooth (embedded) submanifold (see e.g. [46, Example 6.8], [14, Definition 3.10]):

Definition 2.2. *A set $\mathcal{M} \subseteq \mathcal{E}$ is called a* smooth embedded manifold *of dimension d around $\bar{x} \in \mathcal{M}$ if there exists an open neighborhood $U \subseteq \mathcal{E}$ of $\bar{x}$ such that $\mathcal{M}$ can be represented as the set of solutions to $F(x) = 0$, where $F : U \to \mathbb{R}^m$ is $\mathcal{C}^1$ with $\nabla F(\bar{x})$ of rank m, where $m = n - d$. The function F is called a* local defining function *for $\mathcal{M}$ at $\bar{x}$.*

Note that smooth embedded manifolds are manifolds by themselves, see [14]. As we will only be dealing with such manifolds, if we say that a set $\mathcal{M}$ is a manifold, we mean that it satisfies the definition above. Most of the main ideas presented here hold in a more general context and we refer the reader to the references mentioned above for more details. As we will see later, the reason is that the concept of projectional subdifferentials require the more special situation as introduced in Definition 2.2.

Let $\mathcal{M}$ be a smooth embedded submanifold and $x \in \mathcal{M}$. We follow [14] with the following definition and call

$$T_x\mathcal{M} = \{\dot{\gamma}(0) \mid \gamma : I \to \mathcal{M} \text{ is smooth and } \gamma(0) = x\},$$

the *tangent space* of $\mathcal{M}$ at x, where I is an open interval with $0 \in I$. Elements $\xi_x \in T_x\mathcal{M}$ are called *tangent vectors* to $\mathcal{M}$ at x. Note that if F is a local defining function of $\mathcal{M}$ at x, then by [14, Theorem 3.15], it holds that $T_x\mathcal{M} = \ker DF(x)$. On the other hand, by [46, Example 6.8], it also holds that $T_{\mathcal{M}}(x) = \ker DF(x)$. Hence, the tangent space and

the tangent cone agree and the notations $T_x\mathcal{M}$ and $T_{\mathcal{M}}(x)$ can be used interchangeably. Consequently, the normal cone to a submanifold as the orthogonal complement of the tangent space is a linear space itself, also called *normal space*, that is, we have

$$N_{\mathcal{M}}(x) = (T_x\mathcal{M})^{\perp}.$$

Note that $T_x\mathcal{M}$ is a vector space with the same dimension d as the manifold $\mathcal{M}$ and with its zero element denoted as 0_x.

The collection $T\mathcal{M} := \cup_{x\in\mathcal{M}}\{x\} \times T_x\mathcal{M}$ denotes the *tangent bundle* of the manifold $\mathcal{M}$.

If the tangent space $T_x\mathcal{M}$ is equipped with a smoothly varying inner product $\langle\cdot,\cdot\rangle_x$ called the *Riemannian metric*, the manifold $\mathcal{M}$ called a *Riemannian manifold*. Let us note that if $\mathcal{M} \subseteq \mathcal{E}$ is a submanifold embedded in some Euclidean space $\mathcal{E}$, then $T_x\mathcal{M}$ can be identified with a linear subspace of $\mathcal{E}$ and the inner product on $\mathcal{E}$ carries over to the tangent spaces in a natural way by means of restriction to the respective subspace. In that case, which we will assume in the following, $\mathcal{M}$ becomes a *Riemannian submanifold* of the Euclidean space $\mathcal{E}$.

For a smooth mapping $G : \mathcal{M} \to \mathcal{N}$ between two embedded submanifolds and $x \in \mathcal{M}$, we call the linear map $DG(x) : T_x\mathcal{M} \to T_{G(x)}\mathcal{N}$ defined by

$$DG(x)[\xi_x] = \frac{\mathrm{d}}{\mathrm{d}t}G(\gamma(t))|_{t=0} = (G \circ \gamma)'(0),$$

where γ is a smooth curve on $\mathcal{M}$ through x at $t = 0$ with velocity $\dot\gamma(0) = \xi_x$, the *differential* of G at x.

On a manifold $\mathcal{M}$, the *Riemannian gradient* of a smooth map $g : \mathcal{M} \to \mathbb{R}$ is the tangent vector $\mathrm{grad}g(x) \in T_x\mathcal{M}$ such that for all $\xi_x \in T_x\mathcal{M}$ it holds that

$$Dg(x)[\xi_x] = \langle \mathrm{grad}g(x), \xi_x\rangle_x.$$

Analogous to the Euclidean setting, the negative of the Riemannian gradient is the direction of steepest descent.

In manifold optimization, a key concept are so-called *retractions*. As noted in [1], these have two important purposes: They turn elements of $T_x\mathcal{M}$ into points of $\mathcal{M}$ and secondly, by means of the pullback through the retraction R, functions on $\mathcal{M}$ can be transformed to functions defined on the vector space $T_x\mathcal{M}$.

Definition 2.3. *Let $\mathcal{M}$ be a manifold. A smooth (i.e. at least $\mathcal{C}^2$) mapping $\mathrm{R} : T\mathcal{M} \to \mathcal{M}$ is called a* retraction *if for all $x \in \mathcal{M}$ the restriction $\mathrm{R}_x := \mathrm{R}(x,\cdot) : T_x\mathcal{M} \to \mathcal{M}$ satisfies the following properties:*

1. *$\mathrm{R}_x(0_x) = x$, where 0_x is the zero element of $T_x\mathcal{M}$.*
2. *With the identification $T_0T_x\mathcal{M} \simeq T_x\mathcal{M}$, it holds that*

$$D\mathrm{R}_x(0) = \mathrm{id}_{T_x\mathcal{M}},$$

where $\mathrm{id}_{T_x\mathcal{M}}$ is the identity mapping on $T_x\mathcal{M}$.

One possibility for R is the so-called *Riemannian exponential map*, which is, however, not always easy to compute. Finding a *good* retraction is crucial for the design of numerical optimization methods on manifolds. For example, in the Euclidean setting, the canonical

retraction is given by $\xi \mapsto x + \xi$ (note that this is well-defined as $T_x\mathcal{E} \simeq \mathcal{E}$) and for the sphere $\mathbb{S}^{n-1}$, we obtain a retraction by means the orthogonal projection mapping

$$\mathrm{R}_x(\xi) := \frac{x+\xi}{\|x+\xi\|}.$$

Remark 2.4. For simplicity, our subsequent theory will assume that we have a globally defined retraction on the manifold appearing in the constraints of our optimization problem, i.e. the retraction R is indeed a smooth mapping on $T\mathcal{M}$. However, if the stepsizes of our method are adjusted accordingly, our results still hold true under the weaker condition that the radius of definition of R_x remains uniformly bounded away from zero on compact subsets of $\mathcal{M}$.

Recently, there has been a growing interest in extending the classical concepts from nonsmooth and variational analysis to the context of manifolds. Notably, [34] consider so-called projectional subdifferentials on submanifolds. In particular, [34, Definition 3.1] and [34, Theorem 3.1] justify the following definition:

Definition 2.5. *Let $\mathcal{M} \subseteq \mathcal{E}$ be a smooth (embedded) manifold with $\bar{x} \in \mathcal{M}$ and $g : \mathcal{E} \to \overline{\mathbb{R}}$. Then we call*

$$\hat{\partial}_{\mathcal{M}} g(\bar{x}) := \mathrm{proj}_{T_{\bar{x}}\mathcal{M}} \hat{\partial}(g + \delta_{\mathcal{M}})(\bar{x})$$

the Fréchet projectional subdifferential *of g at $\bar{x}$ relative to $\mathcal{M}$, and*

$$\partial_{\mathcal{M}} g(\bar{x}) := \mathrm{proj}_{T_{\bar{x}}\mathcal{M}} \partial(g + \delta_{\mathcal{M}})(\bar{x})$$

the projectional (limiting/basic/Mordukhovich) subdifferential *of g at $\bar{x}$ relative to $\mathcal{M}$.*

If g is differentiable and $\mathcal{M} \subseteq \mathcal{E}$ an embedded manifold, it holds by [1, Equation 3.37] that $\mathrm{grad} g(x) = \mathrm{proj}_{T_x\mathcal{M}} \nabla g(x)$. Hence by [34, Proposition 3.2], we have $\partial_{\mathcal{M}} g(x) = \hat{\partial}_{\mathcal{M}} g(x) = \{\mathrm{grad} g(x)\}$.

Let us also state the following characterizations of the projectional subdifferentials from [34]:

$$\hat{\partial}_{\mathcal{M}} g(\bar{x}) = \hat{\partial}(g + \delta_{\mathcal{M}})(\bar{x}) \cap T_{\bar{x}}\mathcal{M} \quad \text{and} \quad \partial_{\mathcal{M}} g(\bar{x}) = \partial(g + \delta_{\mathcal{M}})(\bar{x}) \cap T_{\bar{x}}\mathcal{M}. \tag{2.2}$$

Any point $x^* \in \mathrm{dom} g \cap \mathcal{M}$ satisfying $0 \in \partial_{\mathcal{M}} g(x^*)$ is called a *stationary point* of g with respect to $\mathcal{M}$. This can be motivated by the fact that given a local minimizer x^* of $g + \delta_{\mathcal{M}}$, it holds that $0 \in \hat{\partial}(g+\delta_{\mathcal{M}})(x^*)$ and hence $0 \in \hat{\partial}(g+\delta_{\mathcal{M}})(x^*) \cap T_{\bar{x}}\mathcal{M} = \hat{\partial}_{\mathcal{M}} g(x^*) \subseteq \partial_{\mathcal{M}} g(x^*)$.

Let us also note that the projectional limiting subdifferential is *robust* in the following sense (see [34, Corollary 3.1]): Assume that $\{x^k\}$ is a sequence converging to some limit $\bar{x}$ such that $(g + \delta_{\mathcal{M}})(x^k) \to (g + \delta_{\mathcal{M}})(\bar{x})$ and $w^k \in \partial_{\mathcal{M}} g(x^k)$ for all $k \in \mathbb{N}$ converges to some $\bar{w}$. Then it holds that $\bar{w} \in \partial_{\mathcal{M}} g(\bar{x})$. This property will play an important role later on in our convergence theory.

The projectional subdifferential also has the following properties.

Lemma 2.6. *Assume that g is locally Lipschitz continuous on an open superset of an embedded manifold $\mathcal{M} \subseteq \mathcal{E}$. Then the following holds for all $\bar{x} \in \mathcal{M}$.*

1. *For all $w_{\mathcal{M}} \in \partial_{\mathcal{M}} g(\bar{x})$, there exists an element $w \in \partial g(\bar{x})$ such that $w_{\mathcal{M}} = \mathrm{proj}_{T_{\bar{x}}\mathcal{M}}(w)$. In particular, $\partial_{\mathcal{M}} g$ is locally bounded.*

2. *The projectional subdifferential is non-empty, i.e. we have $\partial_{\mathcal{M}} g(\bar{x}) \neq \emptyset$.*

Proof. As g is locally Lipschitz continuous, the sum-rule for the limiting subdifferential as in [39, Corallary 2.20] gives $\partial(g+\delta_{\mathcal{M}})(\bar{x}) \subseteq \partial g(\bar{x}) + N_{\mathcal{M}}(\bar{x})$. Hence, for every $w_{\mathcal{M}} \in \partial_{\mathcal{M}} g(\bar{x})$, there exists $w \in \partial g(\bar{x})$ and $n \in N_{\mathcal{M}}(\bar{x})$ with $w_{\mathcal{M}} = \mathrm{proj}_{T_{\bar{x}}\mathcal{M}}(w+n) = \mathrm{proj}_{T_{\bar{x}}\mathcal{M}}(w)$, as $N_{\mathcal{M}}(\bar{x}) = (T_{\bar{x}}\mathcal{M})^{\perp}$. The local boundedness of $\partial_{\mathcal{M}} g$ now follows directly from the corresponding property of the classical subdifferential for locally Lipschitz functions.

As in [38], the non-emptiness of $\partial_{\mathcal{M}} g(\bar{x})$ follows from [46, Corollary 8.10], which states that as $g + \delta_{\mathcal{M}}$ is finite and locally lower semicontinuous at $\bar{x}$, there exists a sequence $\{x^k\}$ with $x^k \to \bar{x}$, $(g+\delta_{\mathcal{M}})(x^k) \to (g+\delta_{\mathcal{M}})(\bar{x})$ and $\partial(g+\delta_M)(x^k) \neq \emptyset$. By consequence, we have $\partial_{\mathcal{M}} g(x^k) \neq \emptyset$ for all k large enough. Further, choosing $v^k \in \partial_{\mathcal{M}} g(x^k)$ gives a bounded sequence by the first part. Hence, on a subsequence, v^k converges to some $\bar{v}$ and by the robustness property we have $\bar{v} \in \partial_{\mathcal{M}} g(\bar{x})$, making $\partial_{\mathcal{M}} g(\bar{x})$ non-empty. □

Remark 2.7. Our algorithmic theory later is formulated in terms of the projectional limiting subdifferential. However, the conclusions remain valid for other subdifferentials. The key properties are the robustness property mentioned above and boundedness of the subdifferential for locally Lipschitz functions. In the Euclidean setting, these are shared by the classical limiting and the Clarke subdifferential, while the Fréchet subdifferential is not robust in this sense. Note that in some applications, Clarke subgradients can be computed (see [2]), however, the notion of stationarity in terms of the Clarke subdifferential is weaker. For our manifold setting, [26] also introduced a Clarke-type subdifferential on Riemannian manifolds for which also a Riemannian Kurdyka–Łojasiewicz property was developed in [29]. These subdifferentials can be defined for arbitrary manifolds, not only embedded Riemannian submanifolds. The advantage of projectional subdifferentials is that for submanifolds, properties on the ambient Euclidean space can be transferred to properties of the projectional subdifferential as in Section 2.4.

2.3 Upper-$\mathcal{C}^2$ functions

We next introduce the class of upper-$\mathcal{C}^2$ functions that will play a central role for the design and the convergence analysis of our descent method. For more details on this class of functions, see [46].

Definition 2.8. *Let $U \subseteq \mathcal{E}$ be an open set. We say that a function $\varphi : U \to \mathbb{R}$ is* upper-$\mathcal{C}^2$ *on U, if, on some neighborhood V of each $\bar{x} \in U$, there is a representation*

$$\varphi(x) = \min_{c \in C} \varphi_c(x),$$

where the functions φ_c are of class $\mathcal{C}^2$ on V, and C is a compact set (in some topological space) such that φ_c and its first- and second-order partial derivatives depend continuously on $(x, c) \in V \times C$.

Taking the discrete topology, it follows, for example, that functions of the form $\varphi(x) := \min\{f_1(x), \ldots, f_l(x)\}$ with twice continuously differentiable functions $f_i : U \to \mathbb{R}$ on some open set $U \subseteq \mathcal{E}$ are upper-$\mathcal{C}^2$ functions. Further examples can be derived from the subsequent characterization of upper-$\mathcal{C}^2$ functions from the recent report [2, Prop. 3.2] that will be particularly relevant for our setting.

Proposition 2.9. *Let $U \subseteq \mathcal{E}$ be an open set and $\varphi : \mathcal{E} \to \mathbb{R}$ be locally Lipschitz on U. Then the following statements are equivalent:*

(a) φ is upper-$\mathcal{C}^2$ on U.

(b) For each $\bar{x} \in U$, there exist a constant $\kappa \geq 0$ and some neighborhood V of $\bar{x}$ such that

$$\varphi(y) \leq \varphi(x) + \langle w, y - x \rangle + \kappa \|y - x\|^2 \tag{2.3}$$

for all $x, y \in V$ and all $w \in \partial\varphi(x)$.

(c) For each $\bar{x} \in U$, there exists some neighborhood V of $\bar{x}$ such that φ can be expressed as $\varphi = g - h$, where g is differentiable with Lipschitz gradient, and h is Lipschitz and prox-regular (see [46, Definition 3.27]). Indeed, one can take $g = \kappa\|\cdot\|^2$, for some $\kappa \geq 0$, and h to be convex.

The characterization from part (b) is particularly interesting in our case. The inequality (2.3) is the counterpart of the usual descent lemma for smooth functions with Lipschitz gradient, the constant κ in (2.3) plays the role of the corresponding Lipschitz constant. We stress, however, that κ in (2.3) is a local constant depending on the given point $\bar{x}$. Furthermore, it is worth noting that the inequality (2.3) holds for an arbitrary subgradient $w \in \partial\varphi(x)$, not just for a particular element from $\partial\varphi(x)$. This observation is highly important for our descent method to be well-defined. Finally, we note that part (c) shows that upper-$\mathcal{C}^2$ functions are closely related to the class of DC-functions (DC = difference-of-convex). In particular, this characterization provides several further examples of nonsmooth upper-$\mathcal{C}^2$ functions. In this respect, we refer the interested reader also to the corresponding discussion in [2].

2.4 Descent Property of Upper-$\mathcal{C}^2$ Functions on Submanifolds

Before presenting our algorithm in the next section, we establish the following result, which shows that upper-$\mathcal{C}^2$ functions defined on an open superset of an embedded Riemannian submanifold $\mathcal{M} \subseteq \mathcal{E}$ satisfy a certain descent property. This condition is inspired by the retraction smoothness condition in [15], where it was used for differentiable functions as a generalization to the well-known descent lemma to optimization problems on manifolds. Note again, however, that our Assumption 2.10 is both nonsmooth and merely local.

Assumption 2.10. *Let $\mathcal{M}$ be a manifold with retraction R and assume that for all $\bar{x} \in \mathcal{M}$ there exists a constant $\tilde{\kappa} > 0$, $r > 0$ and an open neighborhood $V \subseteq \mathcal{M}$ of $\bar{x}$, such that the objective function $\varphi : \mathcal{M} \to \mathbb{R}$ from (1.1) satisfies*

$$\varphi\big(\mathrm{R}_x(\xi_x)\big) - \big(\varphi(x) + \langle w_{\mathcal{M}}, \xi_x \rangle\big) \leq \tilde{\kappa}\|\xi_x\|^2, \tag{2.4}$$

for all $x \in V$, $\xi_x \in \overline{B}_r(0_x) \subset T_x\mathcal{M}$ and all $w_{\mathcal{M}} \in \partial_{\mathcal{M}}\varphi(x)$. Further, assume that $\partial_{\mathcal{M}}\varphi$ is locally bounded.

Let us show that it is sufficient for φ to be upper-$\mathcal{C}^2$ on an open superset of $\mathcal{M}$ for φ to satisfy (2.4). To do so, we first need the following properties of retractions. These are somewhat standard and similar arguments appear in the proof of [15, Lemma 2.7]. However, as they are also used later on, we present them as an independent lemma.

Lemma 2.11. *Let $\mathcal{M} \subseteq \mathcal{E}$ be a smooth embedded manifold and $\mathrm{R} : T\mathcal{M} \to \mathcal{M}$ a retraction on $\mathcal{M}$. Let $C \subseteq \mathcal{M}$ be compact. Then there exists a radius $r > 0$ and constants $\alpha, \beta \geq 0$ such that for all $x \in C$ the inequalities*

$$\|\mathrm{R}_x(\xi_x) - x\| \leq \alpha\|\xi_x\|, \tag{2.5}$$

$$\|\mathrm{R}_x(\xi_x) - x - \xi_x\| \leq \beta\|\xi_x\|^2 \tag{2.6}$$

hold for all $\xi_x \in \overline{B}_r(0_x) \subseteq T_x\mathcal{M}$.

Proof. Let $r > 0$. By assumption, R is defined and smooth on the set $K_0 := \{(x, \xi_x) \in T\mathcal{M} : x \in C, \|\xi_x\| \leq r\}$. Note that this set is compact by [36, Lemma C.8] and hence (by continuity), $\tilde{C} := \mathrm{R}(K_0)$ is compact, too. Now, denote

$$K := \{(x, \xi_x) \in T\mathcal{M} \mid x \in \tilde{C}, \|\xi_x\| \leq r\}, \tag{2.7}$$

which again is compact by the same argument.

For the remaining part, we follow the proof of [15, Lemma 2.7]. We first prove equation (2.5): Let $x \in C$, then for all $\xi_x \in \overline{B}_r(0_x) \subseteq T_x\mathcal{M}$, we have

$$\begin{aligned}\|\mathrm{R}_x(\xi_x) - x\| &\leq \int_0^1 \left\| \frac{\mathrm{d}}{\mathrm{d}t} \mathrm{R}_x(t\xi_x) \right\| \mathrm{d}t = \int_0^1 \|D\mathrm{R}_x(t\xi_x)[\xi_x]\| \,\mathrm{d}t \\ &\leq \int_0^1 \max_{(z,\zeta_z)\in K} \|D\mathrm{R}_z(\zeta_z)\| \|\xi_x\| \,\mathrm{d}t = \max_{(z,\zeta_z)\in K} \|D\mathrm{R}_z(\zeta_z)\| \|\xi_x\| =: \alpha \|\xi_x\|,\end{aligned}$$

where the maximum exists and is finite due to the smoothness of the retraction and the compactness of K.

The second claim (2.6) follows along similar lines: Let again $x \in C$, $\xi_x \in \overline{B}_r(0_x)$, then

$$\begin{aligned}\|\mathrm{R}_x(\xi_x) - x - \xi_x\| = \|\mathrm{R}_x(\xi_x) - \mathrm{R}_x(0_x) - \xi_x\| &\leq \int_0^1 \left\| \frac{\mathrm{d}}{\mathrm{d}t} \Big(\mathrm{R}_x(t\xi_x) - t\xi_x \Big) \right\| \mathrm{d}t \\ &\leq \int_0^1 \|D\mathrm{R}_x(t\xi_x)[\xi_x] - \xi_x\| \,\mathrm{d}t \leq \int_0^1 \|D\mathrm{R}_x(t\xi_x) - \mathrm{id}_{T_x\mathcal{M}}\| \|\xi_x\| \,\mathrm{d}t \\ &\leq \frac{1}{2} \max_{(z,\zeta_z)\in K} \|D^2\mathrm{R}_z(\zeta_z)\| \|\xi_x\|^2,\end{aligned}$$

where the last inequality follows from $\mathrm{id}_{T_x\mathcal{M}} = D\mathrm{R}_x(0_x)$ and the following calculation:

$$\|D\mathrm{R}_x(t\xi_x) - \mathrm{id}_{T_x\mathcal{M}}\| \leq \int_0^1 \left\| \frac{\mathrm{d}}{\mathrm{d}s} D\mathrm{R}_x(st\xi_x) \right\| \mathrm{d}s \leq \|t\xi_x\| \int_0^1 \|D^2\mathrm{R}_x(st\xi_x)\| \,\mathrm{d}s.$$

Again, as K is compact, the maximum above exists and hence (2.6) follows with $\beta := \frac{1}{2} \max_{(z,\zeta_z)\in K} \|D^2\mathrm{R}_z(\zeta_z)\|$. □

Finally, we prove a nonsmooth version of a corresponding result in [15]:

Lemma 2.12. *Let $\mathcal{M}$ be a (smooth embedded Riemannian) submanifold such that $\mathcal{M} \subseteq \mathcal{O} \subseteq \mathcal{E}$, where $\mathcal{O}$ is open. Further, assume that φ is upper-$\mathcal{C}^2$ on $\mathcal{O}$ and let* R *be a retraction on $\mathcal{M}$. Then Assumption 2.10 is satisfied.*

Proof. The local boundedness of $\partial_{\mathcal{M}}\varphi$ is a direct consequence of Lemma 2.6.

By Lemma 2.11, there exists some $r > 0$ such that (2.5) and (2.6) hold (also in an open neighborhood). Further, by Proposition 2.9, we have that for every point $\bar{x}$ in $\mathcal{O}$, the descent property (2.3) holds for all x, y in an open neighborhood V. As R is continuous, $\mathrm{R}^{-1}(V)$ is an open neighborhood of $(\bar{x}, 0_{\bar{x}}) \in T\mathcal{M}$. Hence (see e.g. [16]), there exists an open neighborhood $U \subseteq \mathcal{M}$ of $\bar{x}$ and $r > 0$ such that

$$\mathcal{U} := \{(x, \xi_x) \in T\mathcal{M} \mid x \in U, \|\xi_x\| < r\} \subseteq \mathrm{R}^{-1}(V),$$

and $\mathcal{U}$ is an open neighborhood of $(\bar{x}, 0_{\bar{x}})$ in $T\mathcal{M}$. Thus, by taking x from a possibly smaller open neighborhood, we can choose r small enough such that both properties above hold for all $\xi_x \in \overline{B}_r(x)$, with $y = \mathrm{R}_x(\xi_x)$ in (2.3), that is

$$\varphi(\mathrm{R}_x(\xi_x)) \leq \varphi(x) + \langle w, \mathrm{R}_x(\xi_x) - x \rangle + \kappa \|\mathrm{R}_x(\xi_x) - x\|^2 \tag{2.8}$$

holds for all $w \in \partial\varphi(x)$.

Now, let $w_{\mathcal{M}} \in \partial_{\mathcal{M}}\varphi(x)$. As in the proof of Lemma 2.6, we find $w \in \partial\varphi(x)$ and $n \in N_{\mathcal{M}}(x)$, such that $w_{\mathcal{M}} = w + n$. It now follows from the orthogonality $n \perp \xi_x$ that

$$\begin{aligned}\langle w, \mathrm{R}_x(\xi_x) - x\rangle &= \langle w, \xi_x + \mathrm{R}_x(\xi_x) - x - \xi_x\rangle \\ &= \langle w_{\mathcal{M}}, \xi_x\rangle + \langle w, \mathrm{R}_x(\xi_x) - x - \xi_x\rangle.\end{aligned}$$

Thus, by (2.8), we have

$$\begin{aligned}\varphi(\mathrm{R}_x(\xi_x)) &\leq \varphi(x) + \langle w_{\mathcal{M}}, \xi_x\rangle + \langle w, \mathrm{R}_x(\xi_x) - x - \xi_x\rangle + \kappa\|\mathrm{R}_x(\xi_x) - x\|^2 \\ &\leq \varphi(x) + \langle w_{\mathcal{M}}, \xi_x\rangle + \|w\|\|\mathrm{R}_x(\xi_x) - x - \xi_x\| + \kappa\|\mathrm{R}_x(\xi_x) - x\|^2.\end{aligned}$$

Now, the set $\partial\varphi(x)$ is bounded by assumption, thus there exists $G \geq 0$ (independent of the choice of $w_{\mathcal{M}}$), such that $\|w\| \leq G$ in the equation above.

Combining (2.5) and (2.6) with the calculation above, we deduce

$$\varphi(\mathrm{R}_x(\xi_x)) \leq \varphi(x) + \langle w_{\mathcal{M}}, \xi_x\rangle + (G\beta + \kappa\alpha^2)\|\xi_x\|^2.$$

As $w_{\mathcal{M}} \in \partial_{\mathcal{M}}\varphi(x)$ and $\xi_x \in \overline{B}_r(0_x)$ were arbitrary, the claim follows. □

3 Descent Method and Global Convergence

Before discussing our method and its convergence properties, let us first provide a motivation for the proposed algorithm. The standard method for the unconstrained minimization of a continuously differentiable function φ is based on the iteration

$$x^{k+1} := x^k + \tau_k d^k, \tag{3.1}$$

with a search direction $d^k \in \mathbb{R}^n$ satisfying the descent property $\langle \nabla\varphi(x^k), d^k\rangle < 0$ and a stepsize τ_k such that a suitable line search criterion holds like the Armijo rule

$$\varphi(x^k + \tau_k d^k) \leq \varphi(x^k) + \sigma\tau_k\langle \nabla\varphi(x^k), d^k\rangle \tag{3.2}$$

for some constant $\sigma \in (0, 1)$. The descent method considered in this section is a direct generalization of this approach to the class optimization problems on some manifold $\mathcal{M}$ with objective functions φ which are upper-$\mathcal{C}^2$ (on an open superset of $\mathcal{M}$ in the embedding space). In the Euclidean setting, which was considered in [2], the monotone version is based on the iteration (3.1) with some search direction d^k satisfying the descent-like property $\langle w^k, d^k\rangle < 0$ for an arbitrary element $w^k \in \partial\varphi(x^k)$. The counterpart of the (monotone) Armijo rule (3.2) reads

$$\varphi(x^k + \tau_k d^k) \leq \varphi(x^k) + \sigma\tau_k\langle w^k, d^k\rangle$$

for some $\sigma \in (0, 1)$.

Linesearch methods have also become popular for problems constrained to smooth manifolds, see e.g. [1, 15, 42]. The generalization from the Euclidean setting is straightforward, by simply replacing the canonical retraction $x \mapsto x + \xi$ in the Euclidean setting by a retraction defined for the respective manifold. Under suitable assumptions on d^k, the linesearch criterion is given by

$$\varphi(\mathrm{R}_{x^k}(d^k)) \leq \varphi(x^k) + \sigma\tau_k\langle \mathrm{grad}\varphi(x^k), d^k\rangle,$$

when φ is a differentiable function. In our nonsmooth case, we will consider the following condition, where we choose some $w_{\mathcal{M}}^k \in \partial_{\mathcal{M}}\varphi(x^k)$ and a direction d^k:

$$\varphi(\mathrm{R}_{x^k}(d^k)) \leq \varphi(x^k) + \sigma\tau_k\langle w_{\mathcal{M}}^k, d^k\rangle.$$

In order to prove global convergence results, we need to impose some conditions on the quality of the descent direction d^k and the underlying choice of the subgradients $w_{\mathcal{M}}^k$ to ensure that d^k remains related to the subgradients. The following conditions are similar to those already specified in [2] for upper-$\mathcal{C}^2$ functions in the Euclidean case and also appear already in [51]. Note that these conditions are also standard for linesearch methods in smooth manifold optimization, see [15, 42, 44], where a number of equivalent conditions are used.

Assumption 3.1. *Assume:*

(a) There exists a constant $a > 0$ such that

$$\langle w_{\mathcal{M}}^k, d^k\rangle \leq -a\|d^k\|^2, \text{ for all } k \in \mathbb{N}. \tag{3.3}$$

(b) There exists a constant $b > 0$ such that

$$\|w_{\mathcal{M}}^k\| \leq b\|d^k\|, \text{ for all } k \in \mathbb{N}. \tag{3.4}$$

Note that for $d^k \neq 0$, we obtain from Cauchy-Schwarz that

$$a\|d^k\| \leq \|w_{\mathcal{M}}^k\|, \tag{3.5}$$

and also

$$\langle -w_{\mathcal{M}}^k, d^k\rangle \geq a\|d^k\|^2 \geq \frac{a}{b}\|d^k\|\|w_{\mathcal{M}}^k\|. \tag{3.6}$$

The following result (see [2, Proposition 4.3] for a similar result for upper-$\mathcal{C}^2$ functions in the Euclidean setting and [15, Lemma 2.10] for the smooth case on a manifold) shows that the Armijo-type condition for our linesearch is satisfied for all $\tau_k > 0$ sufficiently small.

Proposition 3.2. *Assume that $\varphi : \mathcal{M} \to \mathbb{R}$ satisfies Assumption 2.10. Given any $\sigma \in (0,1)$, $x^k \in \mathcal{M}$, $d^k \in T_{x^k}\mathcal{M}$ and $w_{\mathcal{M}}^k \in \partial_{\mathcal{M}}\varphi(x^k)$ such that $w_{\mathcal{M}}^k$ and d^k satisfy Assumption 3.1, there exists a $t > 0$ such that*

$$\varphi\big(\mathrm{R}_{x^k}(\tau d^k)\big) \leq \varphi(x^k) + \sigma\tau\langle w_{\mathcal{M}}^k, d^k\rangle, \text{ for all } \tau \in (0,t). \tag{3.7}$$

Proof. By Assumption 2.10 there exists some $r > 0$ such that (2.4) holds, i.e.

$$\varphi\big(\mathrm{R}_{x^k}(d^k)\big) - \big(\varphi(x^k) + \langle w_{\mathcal{M}}^k, d^k\rangle\big) \leq \tilde{\kappa}\|d^k\|^2 \tag{3.8}$$

for all $d^k \in \overline{B}_r(0_{x^k}) \subset T_{x^k}\mathcal{M}$ and all $w_{\mathcal{M}}^k \in \partial_{\mathcal{M}}\varphi(x^k)$.

For $d^k = 0$, (3.7) is obviously true (for all $t > 0$). Otherwise, we now show that (3.7) holds for all $\tau \in (0,t)$ with $t = \min\left\{\frac{r}{\|d^k\|}, \frac{a^2(1-\sigma)}{b\tilde{\kappa}}\right\}$. By applying first (3.5) and then (3.6), we obtain

$$0 < \tau < t \leq \frac{a^2(1-\sigma)}{b\tilde{\kappa}} \leq \frac{a(1-\sigma)\|w_{\mathcal{M}}^k\|}{b\tilde{\kappa}\|d^k\|} \leq \frac{(1-\sigma)\langle -w_{\mathcal{M}}^k, d^k\rangle}{\tilde{\kappa}\|d^k\|^2}.$$

Consequently, one has for all $\tau \in (0, t)$

$$\tilde{\kappa}\tau^2\|d^k\|^2 \leq (1-\sigma)\tau\langle -w_{\mathcal{M}}^k, d^k\rangle$$

and therefore in combination with (3.8) with d^k replaced by τd^k it holds that

$$-\sigma\tau\langle w_{\mathcal{M}}^k, d^k\rangle \leq -\tau\langle w_{\mathcal{M}}^k, d^k\rangle - \tilde{\kappa}\tau^2\|d^k\|^2 \leq \varphi(x^k) - \varphi(\mathrm{R}_{x^k}(\tau d^k)).$$

Finally, by rearranging terms, we obtain the claimed inequality for all $\tau \in (0, t)$:

$$\varphi(\mathrm{R}_{x^k}(\tau d^k)) \leq \varphi(x^k) + \sigma\tau\langle w_{\mathcal{M}}^k, d^k\rangle.$$

This completes the proof. □

We next turn to a nonmonotone version of the previous iteration. It is clear from (3.7) and Assumption 3.1 that an algorithm with the iteration $x^{k+1} = \mathrm{R}_{x^k}(\tau_k d^k)$, where d^k is such the conditions (3.3) and (3.4) hold and τ_k is determined by a backtracking linesearch with termination criterion as in (3.7), will produce a monotonically decreasing sequence of function values $\{\varphi(x^k)\}_{k\in\mathbb{N}}$.

Of course, (3.7) will also hold if $\varphi(x^k)$ is replaced by an arbitrary upper bound. Therefore, suppose that we have a reference value $\mathcal{R}_k \geq \varphi(x^k)$. Then Proposition 3.2 guarantees that a backtracking linesearch with termination criterion

$$\varphi(\mathrm{R}_{x^k}(\tau_k d^k)) \leq \mathcal{R}_k + \sigma\tau_k\langle w_{\mathcal{M}}^k, d^k\rangle \tag{3.9}$$

terminates after a finite number of iterations. Hence, provided x^k is not already a stationary point, our linesearch method is well-defined.

Further, the condition (3.9) might already be satisfied for a larger choice of τ_k in comparison to the monotone version. This results in possibly larger steps, which is the reason why nonmonotone methods may outperform their monotone counterparts in practical applications.

In order to obtain suitable (global) convergence results, the reference value $\mathcal{R}_k$ has to be chosen in a careful way. One popular choice is due to Grippo et al. [24], where $\mathcal{R}_k := \max\{\varphi(x^j) \mid j = k, k-1, \ldots, k-m_k\}$ for some given $m_k \in \mathbb{N}$. We call this strategy the *max-rule* since $\mathcal{R}_k$ is defined as the maximum function value over the last few iterates, say, the last ten points. In the Euclidean setting, this choice of reference values was already studied in [2], where stationarity of accumulation points for arbitrary upper-$\mathcal{C}^2$ functions was derived.

In our Algorithm 3.3, however, we use the technique introduced by Zhang and Hager [51], where $\mathcal{R}_{k+1}$ is computed as a convex combination of the previous reference value $\mathcal{R}_k$ and the new function value $\varphi(x^{k+1})$. We therefore call this the *mean-rule*. Equation (3.10) in the upcoming result Lemma 3.4 shows that $\mathcal{R}_k$ chosen by the mean-rule is indeed an upper bound for $\varphi(x^k)$ in our setting. In the case where φ is a differentiable function and hence $w_{\mathcal{M}}^k = \mathrm{grad}\varphi(x^k)$, the nonmonotone linesearch with mean-rule is well-established in manifold optimization and was already studied in [42] and [44]. Let us note, however, that even in the differentiable case, the assumptions are usually more restrictive. Most notably, Assumption 2.10 is weaker than the assumption that φ has a globally Lipschitz continuous gradient on an open superset of the embedded manifold $\mathcal{M}$. Also, unlike previous results, in our case, $\mathcal{M}$ does not need to be compact. The algorithmic details are presented in Algorithm 3.3.

Assumption 3.1 allows a wide variety of options regarding the choice of the search direction d^k. For example, let $w_{\mathcal{M}}^k \in \partial_{\mathcal{M}}\varphi(x^k)$ be arbitrarily given. If $w_{\mathcal{M}}^k = 0$, then x^k is already a stationary point, and we terminate the iteration. Otherwise, we have $w_{\mathcal{M}}^k \neq 0$, in which case the gradient-type direction $d^k := -w_{\mathcal{M}}^k$ satisfies the two conditions (a) and (b) from Assumption 3.1 with $a := b := 1$. However, in the Euclidean setting, we can also take well-known quasi-Newton or limited-memory quasi-Newton directions like $d^k := -H_k w_{\mathcal{M}}^k$ for a bounded and uniformly positive definite sequence of (limited-memory) matrices $\{H_k\}$. Then both conditions from Assumption 3.1 are still satisfied. In the general setting, where $\mathcal{M}$ itself is not a Euclidean space, there exist some generalizations of quasi-Newton methods to Riemannian manifolds, see e.g. [43, 28]. However, they often involve vector transports between tangent spaces, which may be computationally expensive [22].

In the Euclidean setting, our main convergence result reduces to essentially the same as the one from [2]. However, as we use the mean rule as opposed to the max rule, the convergence analysis turns out to be much simpler.

Algorithm 3.3 (Nonmonotone Subgradient Method on Manifolds)**.**

Require: *$x^0 \in \mathcal{M}$, a retraction R on $T\mathcal{M}$, $0 < \tau_{\min} \leq \tau_{\max} < \infty$, $\sigma, \beta \in (0,1)$, $p_{\min} \in (0,1]$.*

Set $\mathcal{R}_0 := \varphi(x^0)$.
for *$k = 0, 1, 2, \ldots$* ***do***
  Choose $w_{\mathcal{M}}^k \in \partial_{\mathcal{M}}\varphi(x^k)$;
  if *$w^k = 0$* ***then***
    STOP and return x^k.
  end if
  Choose $d^k \in T_{x^k}\mathcal{M} \setminus \{0_{x^k}\}$ such that (3.3) and (3.4) hold.
  Choose $\tau_k \in [\tau_{\min}, \tau_{\max}]$.
  while *$\varphi(\mathrm{R}_{x^k}(\tau_k d^k)) > \mathcal{R}_k + \sigma\tau_k\langle w_{\mathcal{M}}^k, d^k\rangle$* ***do***
    $\tau_k = \beta\tau_k$
  end while
  Set $x^{k+1} := \mathrm{R}_{x^k}(\tau_k d^k)$.
  Choose $p_{k+1} \in [p_{\min}, 1]$ and set $\mathcal{R}_{k+1} := (1 - p_{k+1})\mathcal{R}_k + p_{k+1}\varphi(x^{k+1})$.
end for

Note that the sequence $\{x^k\}$ generated by Algorithm 3.3 satisfies $x^{k+1} \neq x^k$ for all k as otherwise, one would have $x^k = x^{k+1} = \mathrm{R}_{x^k}(\tau_k d^k)$ and hence by the property that R_{x^k} is locally invertible, we have $\tau_k d^k = 0_{x^k}$ contradicting our choice of d^k and $\tau_k > 0$. Now, we first collect some more properties of the sequence $\{x^k\}$, which are similar to those obtained for a nonmonotone proximal gradient method in [20].

Lemma 3.4. *Let Assumption 2.10 and Assumption 3.1 be satisfied. Assume that $\inf_{x\in\mathcal{M}} \varphi(x) > -\infty$. Then for all $x^0 \in \mathcal{M}$, Algorithm 3.3 either stops at some stationary point after a finite number of iterations, or the sequence $\{x^k\}_{k\in\mathbb{N}}$ satisfies the following properties:*

(a) For all $k \in \mathbb{N}$ it holds that

$$\varphi(x^{k+1}) + (1 - p_{k+1})\delta_k \leq \mathcal{R}_{k+1} \leq \mathcal{R}_k - p_{k+1}\delta_k, \tag{3.10}$$

where

$$\delta_k := -\sigma\tau_k\langle w_{\mathcal{M}}^k, d^k\rangle \geq \sigma a\tau_k\|d^k\|^2 \geq 0. \tag{3.11}$$

(b) The sequence $\{\mathcal{R}_k\}$ is monotonically decreasing.

(c) *Both* $\{\mathcal{R}_k\}$ *and* $\{\varphi(x^k)\}$ *converge to some value* $\varphi^* \in \mathbb{R}$*, i.e., both sequences converge and have the same limit.*

(d) $\tau_k d^k \to 0$ *for* $k \to \infty$*.*

Proof. The first inequality of part (a) follows from the definition of $\mathcal{R}_{k+1}$ and the linesearch termination criterion in (3.9):

$$\begin{aligned}\mathcal{R}_{k+1} &= (1-p_{k+1})\mathcal{R}_k + p_{k+1}\varphi(x^{k+1}) \\ &\geq (1-p_{k+1})\big(\varphi(x^{k+1}) - \sigma\tau_k\langle w_{\mathcal{M}}^k, d^k\rangle\big) + p_{k+1}\varphi(x^{k+1}) \\ &= \varphi(x^{k+1}) - (1-p_{k+1})\sigma\tau_k\langle w_{\mathcal{M}}^k, d^k\rangle.\end{aligned}$$

Similarly, we verify the second inequality:

$$\begin{aligned}\mathcal{R}_{k+1} &= (1-p_{k+1})\mathcal{R}_k + p_{k+1}\varphi(x^{k+1}) \\ &\leq (1-p_{k+1})\mathcal{R}_k + p_{k+1}(\mathcal{R}_k + \sigma\tau_k\langle w_{\mathcal{M}}^k, d^k\rangle) \\ &= \mathcal{R}_k + p_{k+1}\sigma\tau_k\langle w_{\mathcal{M}}^k, d^k\rangle.\end{aligned}$$

The second assertion directly follows from part (a), and we obtain for all k that

$$\varphi(x^k) \leq \mathcal{R}_k \leq \cdots \leq \mathcal{R}_0 = \varphi(x^0).$$

The last equation also shows that as φ is bounded from below, the sequence of reference values $\{\mathcal{R}_k\}$ is convergent to some limit φ^*. Now, as $p_k \geq p_{\min}$ and

$$\varphi(x^k) = \frac{1}{p_k}\big(\mathcal{R}_k - (1-p_k)\mathcal{R}_{k-1}\big) = \frac{1}{p_k}(\mathcal{R}_k - \mathcal{R}_{k-1}) + \mathcal{R}_{k-1}$$

due to the update of $\mathcal{R}_k$, the (usually nonmonotone) convergence of $\{\varphi(x^k)\}$ to the same limit φ^* follows.

Statement (d) is now a consequence of part (a) and the following telescoping argument:

$$\infty > \mathcal{R}_0 - \varphi^* \geq \sum_{j=1}^{k}\big(\mathcal{R}_{j-1} - \mathcal{R}_j\big) \geq \sum_{j=1}^{k} p_j\delta_{j-1} \geq \sum_{j=1}^{k} p_{\min}\sigma a\frac{\tau_{j-1}^2}{\tau_{\max}}\|d^{j-1}\|^2, \tag{3.12}$$

where the upper bound $\mathcal{R}_0 - \varphi^*$ is independent of k and holds for all $k \in \mathbb{N}$. □

Note that (3.12) directly implies the following estimate similar to [2, Proposition 4.4]:

Proposition 3.5. *There exists a constant* $c > 0$ *such that*

$$\min_{0\leq j\leq k}\tau_k\|d^k\| \leq \frac{c\sqrt{\mathcal{R}_0 - \varphi^*}}{\sqrt{k+1}}, \text{ for all } k \in \mathbb{N}. \tag{3.13}$$

Proof. By (3.12) we obtain

$$\min_{0\leq j\leq k}\tau_k^2\|d^k\|^2 \leq \frac{1}{k+1}\frac{\tau_{\max}}{p_{\min}\sigma a}\sum_{j=0}^{k} p_{\min}\frac{\sigma a}{\tau_{\max}}\tau_j^2\|d^j\|^2 \leq \frac{1}{k+1}\frac{\tau_{\max}}{p_{\min}\sigma a}\big(\mathcal{R}_0 - \varphi^*\big).$$

Taking the square root gives the claim with $c := \sqrt{\frac{\tau_{\max}}{p_{\min}\sigma a}}$. □

Next, we establish subsequential convergence to stationary points for the sequence $\{x^k\}$ generated by Algorithm 3.3.

Theorem 3.6. *Assume that Assumptions 2.10 and 3.1 hold and that φ satisfies $\inf_{x\in\mathcal{M}}\varphi(x) > -\infty$. Then, given any $x^0 \in \mathcal{M}$, either Algorithm 3.3 stops at a stationary point after a finite number of iterations, or the following assertions hold for the generated sequence $\{x^k\}$:*

1. *Suppose that $\{x^k\}_{k\in K}$ is a bounded subsequence of $\{x^k\}$, then the corresponding stepsizes are uniformly bounded away from zero, i.e. $\inf_{k\in K}\tau_k > 0$. Further, it holds that $\|w^k_{\mathcal{M}}\| \to_K 0$ and $\|d^k\| \to_K 0$.*
2. *Any accumulation point of $\{x^k\}_{k\in\mathbb{N}}$ is a stationary point.*
3. *Let x^* be an accumulation point of $\{x^k\}$. Then the entire sequence $\{\varphi(x^k)\}$ converges to $\varphi(x^*)$ and the sequence $\{\mathcal{R}_k\}$ converges monotonically to $\varphi(x^*)$.*

Proof. The crucial part is the claim that $\inf\tau_k > 0$ on bounded subsequences. Assume, by contradiction, that $\inf_{k\in K}\tau_k = 0$, where $\{x^k\}_{k\in K}$ is bounded. By taking subsequences, we may assume that $\tau_k \to_K 0$, and by boundedness of $\{x^k\}_K$, we assume that there exists some $\bar{x}$ such that $x^k \to_K \bar{x}$.

By Lemma 2.6, we have that for all $w^k_{\mathcal{M}}$ there exists $w^k \in \partial\varphi(x^k), k \in K$, such that $w^k_{\mathcal{M}} = \mathrm{proj}_{T_{x^k}\mathcal{M}}(w^k)$, it follows that as $\{w^k\}_{k\in K}$ is bounded, $\{w^k_{\mathcal{M}}\}_K$ is bounded, too. That is, by taking again a suitable subsequence, we have $w^k_{\mathcal{M}} \to_K \bar{w}_{\mathcal{M}}$ for some $\bar{w}_{\mathcal{M}}$. From (3.5), we have $\|w^k\| \geq a\|d^k\|$, which implies that $\{d^k\}_K$ is also bounded and hence, again on a suitable subsequence, we may assume $d^k \to_K \bar{d}$ for some $\bar{d} \in \mathcal{E}$. As $\tau_k \to_K 0$, we assume $\tau_k < \tau_{\min}$ for all $k \in K$ sufficiently large. Denote by $\hat{\tau}_k$ the previous trial stepsize, i.e., $\hat{\tau}_k = \beta^{-1}\tau_k$ and $\hat{x}^{k+1} := \mathrm{R}_{x^k}(\hat{\tau}_k d^k)$. That is, the inner loop does not terminate with the stepsize $\hat{\tau}_k$, therefore

$$\varphi(\hat{x}^{k+1}) > \mathcal{R}_k + \sigma\hat{\tau}_k\langle w^k_{\mathcal{M}}, d^k\rangle. \tag{3.14}$$

Since $\hat{\tau}_k\|d^k\| \to_K 0$, it follows by continuity of the retraction that $\hat{x}^{k+1} \to_K \bar{x}$. By our Assumption 2.10 there exists a constant $\hat{\kappa} > 0$ such that, for all $k \in K$ sufficiently large, we have

$$\varphi(\hat{x}^{k+1}) \leq \varphi(x^k) + \hat{\tau}_k\langle w^k_{\mathcal{M}}, d^k\rangle + \hat{\kappa}\hat{\tau}_k^2\|d^k\|^2. \tag{3.15}$$

A combination of first (3.14), the fact that $\mathcal{R}_k \geq \varphi(x^k)$, followed by (3.15) and (3.3) gives

$$\begin{aligned}\sigma\hat{\tau}_k\langle w^k_{\mathcal{M}}, d^k\rangle &< \varphi(\hat{x}^{k+1}) - \mathcal{R}_k \leq \varphi(\hat{x}^{k+1}) - \varphi(x^k)\\ &\leq \hat{\tau}_k\langle w^k_{\mathcal{M}}, d^k\rangle + \hat{\kappa}\hat{\tau}_k^2\|d^k\|^2 \leq \hat{\tau}_k\left(1 - \frac{\hat{\kappa}}{a}\hat{\tau}_k\right)\langle w^k_{\mathcal{M}}, d^k\rangle.\end{aligned}$$

Therefore, as $\langle w^k_{\mathcal{M}}, d^k\rangle < 0$, it follows that $\sigma > 1 - \frac{\hat{\kappa}}{a}\hat{\tau}_k$. As $\tau_k \to_K 0$ it follows that $\hat{\tau}_k \to_K 0$. Thus, the above inequality contradicts the fact that $\sigma \in (0, 1)$.

The claim concerning the convergence of $\{d^k\}_K$ is now a direct consequence of Lemma 3.4 (d) and since $\|w^k_{\mathcal{M}}\| \leq b\|d^k\|$ by Assumption 3.1, it also follows that $\|w^k\| \to_K 0$.

Now, we are ready to verify the second claim. Let x^* be an accumulation point of $\{x^k\}_{k\in\mathbb{N}}$, hence, there exists a subsequence $\{x^k\}_{k\in K}$ converging to x^*. By the first part, we have $w^k_{\mathcal{M}} \to_K 0$, which, by the so-called robustness property of the projectional subdifferential implies that $0 \in \partial_{\mathcal{M}}\varphi(x^*)$ as $w^k_{\mathcal{M}} \in \partial_{\mathcal{M}}\varphi(x^k)$ for all $k \in K$.

By Lemma 3.4 we already know that both $\{\varphi(x^k)\}$ and $\{\mathcal{R}_k\}$ converge (in the latter case monotonically) and their limits agree. However, on a subsequence, we have $\varphi(x^k) \to_K \varphi(x^*)$ by continuity of φ. Thus, the last claim concerning the convergence of $\{\varphi(x^k)\}$ and $\{\mathcal{R}_k\}$ follows. □

4 Convergence under the Kurdyka–Łojasiewicz Property

The key idea to obtain global and rate-of-convergence results for the nonmonotone subgradient method under the KL property is to exploit that the stepsizes along bounded subsequences remain bounded. This was shown in Theorem 3.6 and allows us to weaken the global assumptions in [44] (for the differentiable case) to merely local ones in the spirit of [32]. Therefore, we restrict the discussion here to highlight the main differences to the previous works [32] and [44]. We omit technical details and only discuss a sketch of proof. The full proofs can be found in the appendix.

For the remaining part, we assume that the Assumptions 2.10 and 3.1 hold. Further, assume that our objective function

$$\Phi := \varphi + \delta_{\mathcal{M}}$$

satisfies the KL property at a given accumulation point $x^* \in \mathcal{M}$. Let $\eta > 0$ be the corresponding constant and χ the associated desingularization function from Definition 2.1, and denote by $\{x^k\}_{k\in K}$ a subsequence converging to x^*.

Let $C \subseteq \mathcal{M}$ be a compact subset with nonempty interior relative to $\mathcal{M}$, such that $x^* \in \operatorname{int} C$ and $\varphi(x) \leq \mathcal{R}_0$ for all $x \in C$. Such a set always exists unless we are in the trivial case $x_0 = x^*$. By Theorem 3.6, there exists a constant $\underline{\tau}_C > 0$ such that

$$\tau_k \geq \underline{\tau}_C \quad \text{for all } k \text{ with } x^k \in C. \tag{4.1}$$

We denote the constants from Lemma 2.11 by α, β respectively and note that these can be chosen as global constants on the set C.

Following mostly [44], we also introduce the subsequent notation:

- $m := \min\left\{l \in \mathbb{N} \,\middle|\, (1-\sqrt{1-p_{\min}})\sqrt{l} \geq (1+\sqrt{1-p_{\min}})\right\}$,
- $l(k) := k + m - 1$,
- $\Xi_{k-1} := \sqrt{\mathcal{R}_{k-1} - \mathcal{R}_k}$ for $k \in \mathbb{N}$ and
- $\Delta_{i,j} := \chi\big(\mathcal{R}_i - \varphi(x^*)\big) - \chi\big(\mathcal{R}_j - \varphi(x^*)\big)$.

The index m is uniquely defined as the left-hand side of the inequality in its definition eventually becomes larger than the constant on the right-hand side. Moreover, the difference $l(k) - k = m - 1$ is a constant number for all $k \in \mathbb{N}$. It is thus guaranteed later that certain sums are always taken over a finite (fixed) number of terms only.

Since $\{\mathcal{R}_k\}$ is a monotone sequence, we have that $\Xi_{k-1} \geq 0$ holds for all $k \in \mathbb{N}$. In combination with the monotonicity of χ, it follows that $\Delta_{i,j} \geq 0$ for all $j \geq i$.

Let us also introduce the two index sets

$$K_1 := \left\{k \in \mathbb{N} \,\middle|\, \varphi(x^k) \leq \mathcal{R}_{k+m}\right\} \text{ and } K_2 := \left\{k \in \mathbb{N} \,\middle|\, \varphi(x^k) > \mathcal{R}_{k+m}\right\}$$

depending on the previously introduced number m.

From Lemma 3.4, we have $\tau_k d^k \to 0$ and

$$p_{\min}\sigma a \tau_k \|d^k\|^2 \leq \mathcal{R}_k - \mathcal{R}_{k+1}.$$

Assuming $x^k \in C$, we can deduce the following inequality

$$\|x^{k+1} - x^k\|^2 \leq \alpha^2 \tau_k^2 \|d^k\|^2 \leq \frac{\alpha^2 \tau_k}{p_{\min} \sigma a}(\mathcal{R}_k - \mathcal{R}_{k+1}) \leq \frac{\alpha^2 \tau_{\max}}{p_{\min} \sigma a} \Xi_k^2,$$

and hence

$$e\|x^{k+1} - x^k\| \leq \Xi_k, \tag{4.2}$$

where e denotes the constant $e = \frac{1}{\alpha}\sqrt{\frac{p_{\min}\sigma a}{\tau_{\max}}}$.

The following result is clear, as the individual summands can be made arbitrarily small (recall again that the difference $l(k) - k = m - 1$ is a constant).

Lemma 4.1. *Define the constant* $\hat{c} := \frac{b\alpha}{\underline{\tau}_C}\sqrt{\frac{\tau_{\max} p_{\min}}{\sigma a}}$, *with* $\underline{\tau}_C$ *as in* (4.1). *Then there exists a sufficiently large index* $k_0 - 1 \in K$ *such that*

$$\vartheta := \|x^{k_0 - 1} - x^*\| + \frac{1}{e}\sum_{j=k_0}^{l(k_0)} (3\Xi_{j-1} + \Xi_j) + \frac{2\hat{c}}{e}\sum_{j=k_0}^{l(k_0)} \chi\big(\mathcal{R}_j - \varphi(x^*)\big) \tag{4.3}$$

satisfies $\overline{B}_\vartheta(x^*) \subset U$, *where* U *is the neighborhood of* x^* *from the KL property in Definition 2.1 and* $\overline{B}_\vartheta(x^*) \cap \mathcal{M} \subset C$.

We next provide an upper bound for the distance of the current point x^k to stationarity, measured by the expression $\operatorname{dist}\big(0, \partial\Phi(x^k)\big)$, see also [44].

Lemma 4.2. *Under the conditions specified above, we have*

$$\operatorname{dist}\big(0, \partial\Phi(x^k)\big) \leq \frac{2b}{\underline{\tau}_C}\|x^{k+1} - x^k\|, \tag{4.4}$$

for all sufficiently large k *with* $x^k \in \overline{B}_\vartheta(x^*)$.

Proof. By Theorem 3.6 and noting that $\overline{B}_\vartheta(x^*) \cap \mathcal{M} \subseteq C$, we obtain that $\tau_k \geq \underline{\tau}_C$ for all k with $x^k \in \overline{B}_\vartheta(x^*)$. Therefore, by Lemma 2.11, we have

$$\|x^{k+1} - (x^k + \tau_k d^k)\| = \|\mathrm{R}_{x^k}(\tau_k d^k) - (x^k + \tau_k d^k)\| = o(\|\tau_k d^k\|).$$

Now the claim follows by Assumption 3.1 as for all k large enough

$$\begin{aligned}\|x^{k+1} - x^k\| &\geq \tau_k\|d^k\| - \|\mathrm{R}_{x^k}(\tau_k d^k) - (x^k + \tau_k d^k)\| \\ &\geq \frac{1}{2}\underline{\tau}_C\|d^k\| \geq \frac{1}{2b}\underline{\tau}_C\|w^k_{\mathcal{M}}\| \geq \frac{1}{2b}\underline{\tau}_C \operatorname{dist}\big(0, \partial\Phi(x^k)\big),\end{aligned}$$

where the last inequality is due to (2.2). □

An application of the corresponding result in [44] to the setting considered here, gives us the following technical result:

Lemma 4.3. *For all sufficiently large* $k \in \mathbb{N}$ *with* $\mathcal{R}_k < \varphi(x^*) + \eta$ *and* $x^k \in \overline{B}_\vartheta(x^*)$, *the following inequality holds:*

$$\frac{1 - \sqrt{1 - p_{\min}}}{\sqrt{m}} \sum_{i=k}^{l(k)} \Xi_i \leq \frac{1}{2}\Xi_k + \sqrt{1 - p_{\min}}\,\Xi_{k-1} + \hat{c}\Delta_{k,k+m}, \tag{4.5}$$

where $\hat{c}$ *denotes the constant from Lemma 4.1.*

Proof. As $x \mapsto \sqrt{x}$ is a concave function, the application of Jensen's inequality yields

$$\frac{1-\sqrt{1-p_{\min}}}{\sqrt{m}} \sum_{i=k}^{l(k)} \Xi_i \leq \big(1-\sqrt{1-p_{\min}}\big)\sqrt{\mathcal{R}_k - \mathcal{R}_{k+m}}. \tag{4.6}$$

We now distinguish two cases. The case $k \in K_1$ follows directly by calculation as in reference [44]. In the second case $k \in K_2$, the key difference to the proximal case in [32] is that from Lemma 4.2 together with the KL property, one obtains the bound

$$\chi'\big(\varphi(x^k) - \varphi(x^*)\big) \geq \frac{1}{\frac{2b}{\tau_C}\|x^{k+1} - x^k\|}.$$

By properties of the desingularization function χ, one arrives at the estimate

$$\varphi(x^k) - \mathcal{R}_{k+m} \leq \frac{2\hat{c}}{p_{\min}} \Xi_k \Delta_{k,k+m},$$

from which the claim can be deduced. The complete proof is in the appendix. □

Now, we obtain the global convergence of $\{x^k\}$ to a stationary point under the KL property of φ.

Theorem 4.4. *Let Assumptions 2.10 and 3.1 hold, let $\{x^k\}_K$ be a subsequence converging to some accumulation point x^*, and suppose that $\Phi := \varphi + \delta_{\mathcal{M}}$ satisfies the KL property at x^*. Then the entire sequence $\{x^k\}$ converges to x^*.*

Proof. Again, we refer to the appendix for the complete proof. Let k_0 be the index from the definition of ϑ, cf. Lemma 4.1. Without loss of generality, we may assume that k_0 is large enough, such that the results from Lemma 4.2 and Lemma 4.3 hold and that $\mathcal{R}_{k_0} < \varphi(x^*) + \eta$. Then, the following two statements hold:

(a) for all $k \geq k_0 - 1$: $x^k \in \overline{B}_\vartheta(x^*)$, and

(b) for all $k \geq l(k_0)$:

$$\big(1-\sqrt{1-p_{\min}}\big)\sqrt{m} \sum_{j=l(k_0)}^{k} \Xi_j \leq \sum_{j=k_0}^{k} \left(\frac{1}{2}\Xi_j + \sqrt{1-p_{\min}}\Xi_{j-1}\right) + \hat{c} \sum_{j=k_0}^{l(k_0)} \chi\big(\mathcal{R}_j - \varphi(x^*)\big), \tag{4.7}$$

where $\hat{c}$ denotes the constant from Lemma 4.1.

Both statements can be verified jointly by induction over k. Note that, in particular, due to the first statement, Lemma 4.3 can be used in the induction step for the second claim. The calculations can be done similar to those in the references mentioned above. In the end, one obtains

$$\|x^{k+1} - x^*\| \leq \|x^{k_0-1} - x^*\| + \sum_{j=k_0-1}^{k} \|x^{j+1} - x^j\| \leq \vartheta.$$

Hence, taking $k \to \infty$, shows that $\{x^k\}$ is a Cauchy sequence and thus convergent. □

Lastly, let us cite the following rate-of-convergence result from [44] for the case where the desingularization function is given by $\chi(t) = ct^\theta$ for some $c > 0$ and $\theta \in (0, 1)$. For completeness, we include a proof in the appendix.

Theorem 4.5. *Let Assumptions 2.10 and 3.1 hold, and suppose that $\{x^k\}$ converges on some subsequence $\{x^k\}_K$ to a limit point x^* such that φ has the KL property at x^*. Then the entire sequence $\{x^k\}$ converges to x^*. Further, if the corresponding desingularization function is given by $\chi(t) = ct^\theta$ (for some $c > 0$ and $\theta \in (0, 1)$), then the following statements hold:*

(a) If $\theta \in [1/2, 1)$, then $\{\mathcal{R}_k\}$ converges R-linearly to $\varphi(x^*)$ and $\{x^k\}$ converges R-linearly to x^*.

(b) If $\theta \in (0, 1/2)$, then there exist constants $\eta_1, \eta_2 > 0$ such that for all k large enough it holds that

$$\mathcal{R}_k - \varphi(x^*) \leq \eta_1 k^{-\frac{1}{1-2\theta}}, \tag{4.8}$$

$$\|x^k - x^*\| \leq \eta_2 k^{-\frac{\theta}{1-2\theta}}. \tag{4.9}$$

5 Numerical Experiments and Applications

The nonmonotone subgradient method comes in different flavors. The simplest choice of the descent direction is the negative of the subgradient, i.e. $d^k = -w^k$. Then, as noted before, if suitable positive definiteness conditions on the corresponding matrices are ensured, the limited memory BFGS search direction from [37] also satisfies our condition in Assumption 3.1. Lastly, problem specific descent direction can be designed (as for the Euclidean minimum sum-of-squares clustering problem in [2]).

Both, the choice of the stepsize strategy and the initial stepsize also have a significant influence on the numerical performance. For comparison, we will include the monotone version. We test the nonmonotone linesearch with parameters chosen manually ($p_k = 0.6$ and $m_k = 5$ for all $k \in \mathbb{N}$ for the mean- and max-rule respectively) in combination with the following initial stepsizes: For quasi-Newton methods, a constant initial guess, i.e. $\tau_k = 1$ works well. If we take the negative of the subgradient, we also test a stepsize inspired by the Barzilai-Borwein stepsize [9], that is:

$$\bar{\tau}_k := \frac{\langle \Delta x^k, \Delta x^k \rangle}{\langle \Delta x^k, \Delta g^k \rangle}, \tag{5.1}$$

with $\Delta x^k = x^k - x^{k-1}$ and $\Delta g^k = w^k - \operatorname{proj}_{T_{x^k}\mathcal{M}}(w^{k-1})$ (note that in a more general setting, we would need to employ parallel transport of w^{k-1} to the tangent space $T_{x^k}\mathcal{M}$). These stepsizes $\bar{\tau}_k$ are then projected onto the interval $[\tau_{\min}, \tau_{\max}]$. There are also so-called self-adaptive strategies, which were proposed in [2], where the nonmonotonicity parameter for the nonmonotone linesearch (i.e. the value p_k for our method or the number of function values over which we take the maximum for the max-rule) and the initial stepsize is chosen automatically. Essentially, if for the two last iterations the initial stepsize was accepted, it is increased and we aim for a monotone decrease of the objective in the next iteration. Conversely, if the initial stepsize is not accepted, we may decrease the initial stepsize and allow for more nonmonotone behavior (for details, we refer to [2]).

Defining

$$\Delta^k_{\text{rel}} x := \frac{\|x^k - x^{k-1}\|}{\max\{\|x^{k-1}\|, 1\}} \quad \text{and} \quad \Delta^k_{\text{rel}} \varphi := \frac{|\varphi(x^k) - \varphi(x^{k-1})|}{\max\{|\varphi(x^{k-1})|, 1\}},$$

for Euclidean problems, the stopping criterion is $\max\{\Delta^k_{\text{rel}} x, \Delta^k_{\text{rel}} \varphi\} \leq \varepsilon$ and without linear structure, we use $\Delta^k_{\text{rel}} \varphi \leq \varepsilon$ with a tolerance parameter $\varepsilon > 0$.

5.1 Euclidean Setting

In the following, we present two applications of our method that arise in data science and machine learning: First, we test our method on the minimum sum-of-squares clustering problem. As a second application, we also show its applicability to multidimensional scaling problems. Both problems also admit a decomposition as a difference of convex functions. Therefore, it seems reasonable to compare the performance of our method to suitable algorithms from DC-programming. In particular, we consider the classical difference of convex functions algorithm (DCA) from [35] and the boosted difference of convex function algorithm (BDCA) introduced in [5], which features an additional linesearch that helps improve convergence speed. In both methods, we solved the subproblems using the LBFGS algorithm.

Let us note that both test problems can be applied to similar datasets, where the data features a partition into a set of clusters. In the case of the minimum sum-of-squares clustering this is clear and the goal is to identify the centers of these clusters. On the other hand, the multidimensional scaling technique is usually employed to inspect the data and decide whether the data can be grouped into meaningful clusters. We test the methods on both artificial and real data:

1. *Artificial Data:* The minimum sum-of-squares clustering is tested on data which is generated as follows: First, l cluster centers are randomly generated from a normal distribution with a relatively high standard deviation. Secondly, the individual data points are again drawn from a normal distribution now centered around the cluster centers we obtained in the first step. The multidimensional scaling problem is tested on random normal data.

2. *Real Data:* We include experiments on the following datasets from the University of California Irvine Machine Learning Repository [33]: `https://archive.ics.uci.edu/dataset/53/iris`, `https://archive.ics.uci.edu/dataset/186/wine+quality`, `https://archive.ics.uci.edu/dataset/545/rice+cammeo+and+osmancik`.

5.1.1 Minimum Sum-of-Squares Clustering

Let $Y := \{y^1, \ldots, y^n\} \subset \mathbb{R}^s$ be a set of data points. The task is to split Y into l clusters given by centroids $\{x^1, \ldots, x^l\}$, i.e. we aim to solve the problem

$$\min \varphi(X) := \frac{1}{n} \sum_{j=1}^{n} \min_{t \in \{1,\ldots,l\}} \|x^t - y^j\|^2. \tag{5.2}$$

Note that this problem was already studied in great detail in [2] and previously in [4], where the authors derive formulas for an element of the (Clarke-) subdifferential and also proposed a search direction based on second-order information from a local approximation

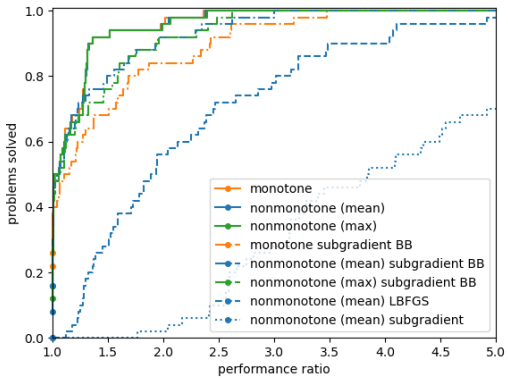

Figure 1: Performance profiles of CPU times for the Minimum Sum-of-Squares Clustering problem. The first 3 methods use the descent direction computed in [2], $\varepsilon = 10^{-4}$.

to the objective function φ. In [2], the nonmonotone subgradient method with the max-rule was shown to outperform several other optimization methods. In particular, they were also competitive in runtime compared to the k-means clustering algorithm. Therefore, the main objective of our numerical study here is to demonstrate that the mean-rule nonmonotone method performs equally well in this application.

We compute the elements $w^k \in \partial_C \varphi$ as in [2] as

$$w^k = \frac{1}{n} \sum_{j=1}^{n} w_j^k,$$

where $w_j^k \in \mathbb{R}^{s \times l}$ is zero except for one entry $2(x_t^k - y_j)$ at the position $t \in \arg\min_{t=1,\dots,l} \|x_t - y_j\|$. The descent directions can be taken as in [2].

Let us also note that from [41], φ admits the following DC formulation as $\varphi(X) = g(X) - h(X)$, where

$$g(X) := \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{l} \|x_j - y_i\|^2 + \frac{\rho}{2} \|X\|^2,$$

and

$$h(X) := \frac{1}{n} \sum_{i=1}^{n} \max_{j=1,\dots,l} \sum_{t=1, t \neq j}^{l} \|x_t - y_i\|^2 + \frac{\rho}{2} \|X\|^2.$$

In both cases, $\|X\|^2$ denotes the Frobenius norm of X and for $\rho > 0$ both functions are strongly convex. Note also that g is differentiable and the subgradients of h can be computed as in [41].

In Figure 1, we show performance profiles for the different methods applied to the minimum sum-of-squares clustering. Let us first note that both DCA as well as BDCA were also tested but are not included in the results as their results were significantly worse.

The methods with the descent direction from [2] performed better than those using only first order information from either the subgradient directly or the LBFGS quasi-Newton direction. We note that the method using the inverse subgradient together with the stepsize from (5.1) performs almost as good as the choice of descent direction from [2]. Further, in that case, both nonmonotone linesearches provide an additional speedup.

Dataset	Algorithm	Runtime (s)	Function Value
iris	nonmonotone (mean)	0.116	0.420
	nonmonotone (mean) LBFGS	0.142	0.420
	nonmonotone (mean) subgradient BB	0.066	0.515
	nonmonotone (mean) subgradient	0.215	0.420
wines	nonmonotone (mean)	4.316	1322.897
	nonmonotone (mean) LBFGS	2.483	1322.894
	nonmonotone (mean) subgradient BB	1.955	1322.892
	nonmonotone (mean) subgradient	3.165	1322.897
rice	nonmonotone (mean)	1.208	1864620.633
	nonmonotone (mean) LBFGS	0.790	1864620.169
	nonmonotone (mean) subgradient BB	0.826	1864620.528
	nonmonotone (mean) subgradient	1.023	1864620.668

Table 1: Numerical results for the minimum sum-of-squares clustering on real datasets (averaged over $N = 5$ runs with random initialization, $\varepsilon = 10^{-4}$).

In particular, both nonmonotone linesearches yield comparable results. This could also be observed for the other methods, i.e. the combination with LBFGS and the inverse subgradient with constant initial trial stepsize of 1. Both, however, have longer runtimes in our experiment. Table 1 shows additional experiments using real datasets. Note that the initialization plays an important role for clustering algorithms. In our experiments, we choose the initial data centroids as random points from the datasets.

5.1.2 Multidimensional Scaling

Denote by $D := (\delta_{ij})_{i,j}$ a given dissimilarity matrix such that the entry δ_{ij} at the position (i, j) measures the distance between two given data points $y_i \in \mathbb{R}^d$ and $y_j \in \mathbb{R}^d$, with d large ($i, j \in \{1, \dots, n\}$). We aim to find points $\{x_1, \dots, x_n\}$ in a lower dimensional space (often in $\mathbb{R}^2$ or $\mathbb{R}^3$) such that the dissimilarities between the original data points $\{y_1, \dots, y_n\}$ correspond to the respective Euclidean distances in the lower dimensional space. Denoting $d_{ij}(X) := \|x_i - x_j\|$, the multidimensional scaling problem is given by

$$\min_{X \in \mathbb{R}^{d \times n}} \varphi(X) := \sum_{i<j} (d_{ij}(X) - \delta_{ij})^2, \tag{5.3}$$

where w_{ij} are some weight parameters.

Again, the formulation of the multidimensional scaling problem in (5.3) can be written as a DC program $\varphi(X) = g(X) - h(X)$, where

$$g(X) := \frac{1}{2} \sum_{i<j} w_{ij} d_{ij}^2(X) + \frac{\rho}{2} \|X\|^2,$$

and

$$h(X) := \sum_{i<j} w_{ij} d_{ij}(X) + \frac{\rho}{2} \|X\|^2.$$

Note that g is differentiable and the subgradients of h can be directly computed.

In Figure 2, the performance profiles show that in this case, the method with the stepsize from (5.1) together with our nonmonotone linesearch method seems to perform

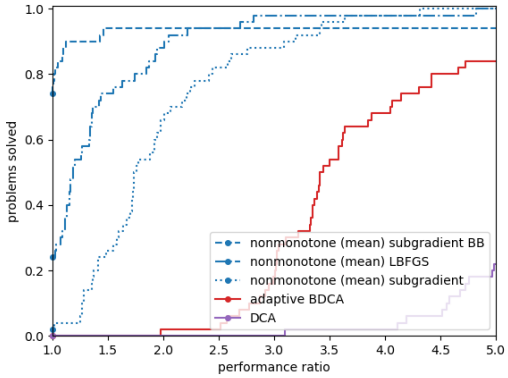

Figure 2: Performance profiles of CPU times for the Multidimensional Scaling problem, $\varepsilon = 10^{-4}$.

Method	Runtime (s)	Function Value
Nonmonotone (mean) subgradient BB	21.495	-50942.952
Nonmonotone (mean) LBFGS	30.386	-50961.759
Nonmonotone (mean) subgradient	52.662	-50969.116
Adaptive BDCA	112.953	-50963.378
DCA	312.823	-50968.881

Table 2: Numerical results for the multidimensional scaling problem on the iris dataset (averaged over $N = 5$ runs with random initialization, $\varepsilon = 10^{-4}$).

best, compared to the method in combination with the LBFGS descent direction and the inverse subgradient method with constant initial trial stepsize. Further, both DCA and BDCA have significantly longer runtimes. In Table 2, all algorithms are tested on the iris dataset.

5.2 Clustering on Manifolds

We will now propose an extension of the aforementioned minimum-sum-of-squares clustering to data on manifolds. Again, let $Y := \{y^1, \ldots, y^n\} \subset \mathcal{M}$ be a given set of data points on some manifold $\mathcal{M}$. The task is to split Y into l clusters given by centroids $\{x^1, \ldots, x^l\} \subset \mathcal{M}$ such that the following optimization problem is solved:

$$\min \varphi(X) := \frac{1}{n} \sum_{j=1}^{n} \min_{t \in \{1,\ldots,l\}} d(x^t, y^j). \tag{5.4}$$

Here, $d(x, y)$ measures the dissimilarity between two points $x, y \in \mathcal{M}$ (but d does not need to be a distance metric). In order to fit into our framework, we assume d to be smooth (i.e. $\mathcal{C}^2$). In the Euclidean setting, $d(x, y) := \|x - y\|^2$ is the case considered in (5.2), and then this smoothness assumption holds true.

Many of the clustering problems on manifolds have been separately analysed in a more statistical context. Our method is different as we take a geometric approach and cluster the data only based on their similarity and make no assumptions on e.g. the distribution.

We initialize the method by randomly sampling the initial estimates for the centroids as data points from the dataset. As the method might get stuck in local minima, at least for datasets of reasonable size, a good clustering can be obtained by comparing the results from several runs of the algorithm with different initializations.

Apart from the value of the objective function (5.4), we include the following evaluation metrics: homogeneity (a cluster is homogeneous if its members belong to a single class of the ground truth), completeness (members of a given class in the ground truth are assigned to the same cluster) and the V-measure (the harmonic mean of homogeneity and completeness) each take values between 0 and 1 [47]. The adjusted Rand index (measures cluster similarities through pairwise comparisons) takes values between -0.5 and 1 [45, 30]. Higher values indicate a better clustering.

Elements of the projectional subdifferentials are hard to compute. Thus, we take $w^k_{\mathcal{M}}$ from the upper bound $w^k_{\mathcal{M}} \in \mathrm{proj}_{T_{x^k}\mathcal{M}}\varphi(x^k)$ and assume that the sum-rule of the limiting subdifferential holds for the sum in problem (5.4). We take the initial stepsizes as in (5.1) and the termination criterion as before.

Finally, let us note that implementations of retractions and projections onto tangent spaces for many manifolds are available e.g. python in the package Pymanopt [49].

5.2.1 Spherical Clustering

In the case of spherical data, a common measure for the similarity between two vectors x and y is given by the so-called *cosine similarity* measure. This is simply the cosine of their angles, which corresponds to the inner product $\langle x, y\rangle$ as x and y have unit length. Our dissimilarity measure is hence defined as

$$d_{\cos}(x, y) = 1 - \langle x, y\rangle. \tag{5.5}$$

Data on the hypersphere arises frequently in data science if the data vectors are scaled to have unit length, which is the case if e.g. frequencies are counted or one is dealing with directional data. The formulation in (5.5), together with (5.4), was already used in [21] to cluster unlabeled document data based on word counts. On the other hand, directional data also arises naturally in many natural and physical sciences [23].

Let us note that the cosine dissimilarity is directly related to the Euclidean distance of two vectors of unit length. In that case, we have $\|x - y\|^2 = 2d_{\cos}(x, y)$.

On the sphere $\mathbb{S}^{n-1}$, by [1], we have the retraction $\mathrm{R}_x(\xi) := \frac{x+\xi}{\|x+\xi\|}$, and the projection onto tangent spaces are given by $\mathrm{proj}_{T_x\mathbb{S}^{n-1}}\xi = (I - xx^T)\xi$ for all $x \in \mathbb{S}^{n-1}$ and $\xi \in T_x\mathbb{S}^{n-1}$.

We show that our method can be applied to spherical clustering by comparing its results to those obtained from the sperical k-means clustering algorithm in [21]. We consider two test scenarios:

1. *Artificial data:* We generate k uniformly distributed random data points on the sphere as mean directions of von Mises-Fisher distributed data clusters with some concentration parameter κ. Each cluster is of the same size l.

2. *Real data:* We use the 20 newsgroups text dataset (available online e.g. here: `https://archive.ics.uci.edu/dataset/113/twenty+newsgroups`) of which we randomly select 5 categories in each run. The text data is then preprocessed by removing English stop words, vectorizing the data, counting word occurrences in each document and converting them to their so-called tf-idf measure. Subsequently, we use a singular value decomposition to reduce the dimension to $d = 100$ and

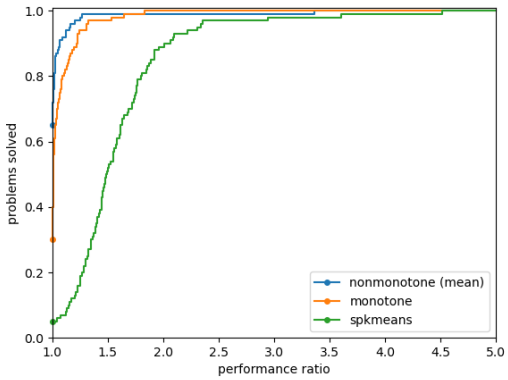

Figure 3: Performance profiles of CPU times for clustering on the sphere with artificial data (100 samples).

normalize each resulting vector. Hence, all data vectors lie on a high dimensional unit sphere.

	Monotone	Nonmonotone	Spherical k-means
Runtime (s)	0.3546 ± 0.1069	0.3480 ± 0.1279	0.5188 ± 0.1824
Function Value	0.840 ± 0.002	0.840 ± 0.002	0.840 ± 0.002
Homogeneity	0.499 ± 0.027	0.501 ± 0.022	0.498 ± 0.027
Completeness	0.499 ± 0.027	0.501 ± 0.022	0.498 ± 0.027
V-measure	0.499 ± 0.027	0.501 ± 0.022	0.498 ± 0.027
Adjusted Rand-Index	0.539 ± 0.035	0.542 ± 0.028	0.538 ± 0.035

Table 3: Results for spherical clustering on artificial data (averaged over 100 samples, with standard deviations, $\varepsilon = 10^{-6}$).

For an artificial dataset with 5 clusters, 1000 data points per cluster on the $\mathbb{S}^{200}$ and with a concentration parameter $\kappa = 10$, our numerical results are summarized in Table 3 with a performance profile for the runtime depicted in Figure 3. In this case, the solutions found by our method and the spherical k-means algorithm are of the same quality. However, our method is faster with the nonmonotone version providing an additional but only very slight advantage. In our experiments, we found that for large values of κ, i.e. more concentrated data, the spherical k-means algorithm performs better, whereas for small concentration parameters, our method is superior.

Results for the real dataset for document clustering are reported in Figure 4 and Table 4. Our method has a slightly worse performance compared to the spherical k-means algorithm.

5.2.2 Clustering on Matrix Manifolds

The approach for directional data, i.e. data on the hypersphere $\mathbb{S}^{n-1}$ can be directly extended to the so-called Stiefel manifold

$$\mathrm{St}(p,n) := \{X \in \mathbb{R}^{n\times p} \mid X^T X = I_p\},$$

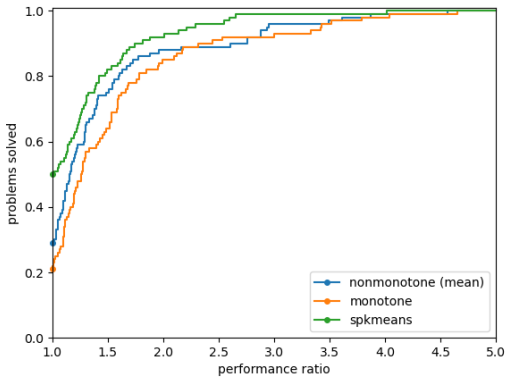

Figure 4: Performance profiles of CPU times for document clustering on the sphere (100 samples).

	Monotone	Nonmonotone	Spherical k-means
Runtime (s)	0.266 ± 0.135	0.243 ± 0.110	0.224 ± 0.107
Function Value	0.628 ± 0.010	0.628 ± 0.010	0.626 ± 0.009
Homogeneity	0.491 ± 0.073	0.488 ± 0.076	0.506 ± 0.072
Completeness	0.509 ± 0.070	0.507 ± 0.074	0.522 ± 0.071
V-measure	0.500 ± 0.071	0.497 ± 0.074	0.514 ± 0.071
Adjusted Rand-Index	0.471 ± 0.089	0.467 ± 0.093	0.490 ± 0.087

Table 4: Results for spherical clustering on document clustering data (averaged over 100 samples, with standard deviations, $\varepsilon = 10^{-6}$).

which inherits the scalar product $\langle X, Y \rangle = \operatorname{tr}(X^T Y)$ on its tangent space from $\mathbb{R}^{n \times p}$. Projections onto the tangent space are given by (see [1]):

$$\operatorname{proj}_{T_X \operatorname{St}(p,n)}(\xi) := (I - XX^T)\xi + \frac{1}{2} X(X^T \xi - \xi^T X).$$

There exist different retractions, for example one might chose

$$\mathrm{R}_X(\xi) := (X + \xi)(I_p - \xi^T \xi)^{-1/2},$$

or alternatively a QR decomposition of $X + \xi$ can be used (see [1] for details). Clustering problems on the Stiefel manifold have been considered, e.g. in [17, 48]. As suggested in [18, 17], we take the following dissimilarity measure

$$d(X, Y) := p - \operatorname{tr}(X^T Y)$$

in (5.4).

The Grassmann manifold $\operatorname{Gr}(p, n)$ is the set of all p dimensional subspaces of $\mathbb{R}^n$. We refer to [10] for an overview of the different representations and only mention two approaches: First, we can identify every subspace $\mathcal{U} \subseteq \mathbb{R}^n$ with its associated orthogonal projection matrix $P \in \mathbb{R}^{n \times n}$. With this representation, the Grassmannian can be written as the set of all $P \in \mathbb{R}^{n \times n}$ such that $P^T = P$, $P^2 = P$ and $\operatorname{rank} P = p$. On the other hand, the Grassmannian can also be described as the quotient manifold $\operatorname{St}(p, n)/\mathrm{O}(p)$ (where $\mathrm{O}(p)$ is the orthogonal group). These equivalent formulations are linked by the fact that

the orthogonal projector onto a subset $\mathcal{U} \subseteq \mathbb{R}^n$ spanned by the columns of $U \in \mathrm{St}(p,n)$ is given by $P = UU^T$.

Clustering on the Grassmann manifold was considered in [25, 17] and [40], the latter provides an application to the clustering of multivariate times series. Following [18, 17, 40], we can measure the dissimilarity of two points $P_1 = UU^T$ and $P_2 = VV^T$, where $U, V \in \mathrm{St}(p,n)$, on the Grassmannian with

$$d(P_1, P_2) = p - \mathrm{tr}(UU^T VV^T) = p - \mathrm{tr}(P_1 P_2).$$

Random matrices on these two matrix manifolds are generated with the method described in [18] and [17]: Let $O \in \mathbb{R}^{n\times n}$ be an orthogonal matrix. Then O can be represented as a product of (orthogonal) matrices $O_n^k(\theta)$ that take the form

$$O_n^k(\theta) := \begin{pmatrix} I_{k-1} & 0 & 0 & 0 \\ 0 & \cos(\theta) & -\sin(\theta) & 0 \\ 0 & \sin(\theta) & \cos(\theta) & 0 \\ 0 & 0 & 0 & I_{n-k-1} \end{pmatrix},$$

as $O = \prod_{k=1}^{n-1} \prod_{j=k}^{n-1} O_n^j(\theta_{k,j})$, where $\theta_{k,j}$ are a set of angles. By selecting the first p columns, we obtain an element $X \in \mathrm{St}(p,n)$ and $P := XX^T \in \mathrm{Gr}(p,n)$. For each cluster, such a set of angles is chosen and then (for the individual data points) we add a uniform random noise on the interval $[-\pi/9, \pi/9]$.

	Monotone	Nonmonotone
Runtime (s)	0.66 ± 0.44	0.54 ± 0.32
Function Value	0.847 ± 0.288	0.837 ± 0.290
Homogeneity	0.882 ± 0.108	0.885 ± 0.109
Completeness	0.940 ± 0.054	0.942 ± 0.055
V-measure	0.909 ± 0.084	0.911 ± 0.084
Adjusted Rand-Index	0.814 ± 0.164	0.820 ± 0.166

Table 5: Clustering results for clustering data on the Stiefel manifold (averaged over 100 examples, with standard deviations, $\varepsilon = 10^{-6}$).

	Monotone	Nonmonotone
Runtime (s)	0.85 ± 0.46	0.81 ± 0.40
Function Value	0.542 ± 0.091	0.538 ± 0.087
Homogeneity	0.862 ± 0.133	0.869 ± 0.130
Completeness	0.911 ± 0.081	0.917 ± 0.080
V-measure	0.885 ± 0.109	0.891 ± 0.107
Adjusted Rand-Index	0.807 ± 0.189	0.817 ± 0.186

Table 6: Clustering results for clustering on the Grassmann manifold (averaged over 100 examples, with standard deviations, $\varepsilon = 10^{-6}$).

Results are shown Table 5 for data on $\mathrm{St}(5,10)$, with 5 clusters, each of 100 data points and Table 6 for clustering on the Grassmann manifold $\mathrm{Gr}(5,10)$. In both cases, the V-measure and adjusted Rand-index are close to 1, indicating that the clustering obtained

by our method successfully reconstructs the structure of the artificial dataset. Further, we observe that the nonmonotone linesearch performs at least as good as its monotone counterpart.

6 Final Remarks

The motivation behind our work was mainly twofold: First, we noticed that by employing the mean-rule as nonmonotone linesearch procedures in our algorithm, the analysis turns out to be much simpler compared to the related max-rule used in [2]. By adapting the corresponding convergence theory for nonmonotone descent methods, we were further able to provide stronger theoretical guarantees in presence of the Kurdyka–Łojasiewicz property. Secondly, in the presence of specific constraints, namely if we consider minimization problems over submanifolds, these findings were further generalized through techniques from manifolds optimization.

It remains an open question whether other related methods, e.g. a projected (nonmonotone) subgradient method, can be realized under the same assumptions on the objective function (i.e. upper-$\mathcal{C}^2$) and with similar convergence properties.

References

[1] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2008.

[2] F. J. Aragón-Artacho, R. Campoy, P. Pérez-Aros, and D. Torregrosa-Belén. "Nonmonotone subgradient methods based on a local descent lemma". In: (2025). arXiv: 2510.19341 [math.OC].

[3] F. J. Aragón-Artacho, R. M. T. Fleming, and P. T. Vuong. "Accelerating the DC algorithm for smooth functions". In: *Mathematical Programming* 169.1 (2018), pp. 95–118. DOI: 10.1007/s10107-017-1180-1.

[4] F. J. Aragón-Artacho, P. Pérez-Aros, and D. Torregrosa-Belén. "The Boosted Double–proximal Subgradient Algorithm for nonconvex optimization". In: *Mathematical Programming* 214 (1 2025), pp. 491–537. DOI: 10.1007/s10107-024-02190-0.

[5] F. J. Aragón-Artacho and P. T. Vuong. "The Boosted Difference of Convex Functions Algorithm for nonsmooth functions". In: *SIAM Journal on Optimization* 30.1 (2020), pp. 980–1006. DOI: 10.1137/18M123339X.

[6] L. Armijo. "Minimization of functions having Lipschitz-continuous first partial derivatives". In: *Pacific Journal of Mathematics* 16 (1966), pp. 1–3.

[7] H. Attouch and J. Bolte. "On the convergence of the proximal algorithm for nonsmooth functions involving analytic features". In: *Mathematical Programming* 116.1 (2009), pp. 5–16. DOI: 10.1007/s10107-007-0133-5.

[8] H. Attouch, J. Bolte, P. Redont, and A. Soubeyran. "Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-Łojasiewicz inequality". In: *Mathematics of Operations Research* 35.2 (2010), pp. 438–457. DOI: 10.1287/moor.1100.0449.

[9] J. Barzilai and J. M. Borwein. "Two-Point step size gradient methods". In: *IMA Journal of Numerical Analysis* 8.1 (Jan. 1988), pp. 141–148. ISSN: 0272-4979. DOI: 10.1093/imanum/8.1.141.

[10] T. Bendokat, R. Zimmermann, and P.-A. Absil. "A Grassmann manifold handbook: basic geometry and computational aspects". In: *Advances in Computational Mathematics* 50 (1 2024). DOI: 10.1007/s10444-023-10090-8.

[11] E. G. Birgin, J. M. Martínez, and M. Raydan. "Nonmonotone Spectral Projected Gradient Methods on Convex Sets". In: *SIAM Journal on Optimization* 10.4 (2000), pp. 1196–1211. DOI: 10.1137/S1052623497330963.

[12] J. Bolte, A. Daniilidis, and A. Lewis. "The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems". In: *SIAM Journal on Optimization* 17.4 (2007), pp. 1205–1223. DOI: 10.1137/050644641.

[13] J. Bolte, A. Daniilidis, A. Lewis, and M. Shiota. "Clarke subgradients of stratifiable functions". In: *SIAM Journal on Optimization* 18.2 (2007), pp. 556–572. DOI: 10.1137/060670080.

[14] N. Boumal. *An Introduction to Optimization on Smooth Manifolds*. Cambridge University Press, 2023. DOI: 10.1017/9781009166164.

[15] N. Boumal, P.-A. Absil, and C. Cartis. "Global rates of convergence for nonconvex optimization on manifolds". In: *IMA Journal of Numerical Analysis* 39.1 (Feb. 2018), pp. 1–33. ISSN: 0272-4979. DOI: 10.1093/imanum/drx080.

[16] M. P. do Carmo. *Riemannian Geometry*. 1st ed. Birkhäuser Boston, MA, 1992. ISBN: 978-0-8176-3490-2. DOI: 10.1007/978-3-642-02431-3.

[17] H. E. Cetingul and R. Vidal. "Intrinsic mean shift for clustering on Stiefel and Grassmann manifolds". In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009, pp. 1896–1902. DOI: 10.1109/CVPR.2009.5206806.

[18] Y. Chikuse. *Statistics on Special Manifolds*. Springer New York, NY, 2003. ISBN: 978-0-387-00160-9. DOI: 10.1007/978-0-387-21540-2.

[19] F. H. Clarke, Y. S. Ledyaev, R. J. Stern, and R. R. Wolenski. *Nonsmooth Analysis and Control Theory*. Springer New York, NY, 1998. DOI: 10.1007/b97650.

[20] A. De Marchi. "Proximal gradient methods beyond monotony". In: *Journal of Nonsmooth Analysis and Optimization* 4 (2023).

[21] I. S. Dhillon and D. S. Modha. "Concept decompositions for large sparse text data using clustering". In: *Machine Learning* 42 (1 2001), pp. 143–175. DOI: 10.1023/A:1007612920971.

[22] R. Godaz, B. Ghojogh, R. Hosseini, R. Monsefi, F. Karray, and M. Crowley. *Vector transport free Riemannian LBFGS for optimization on symmetric positive definite matrix manifolds*. 2021. arXiv: 2108.11019 [math.OC].

[23] M. Golzy, M. Markatou, and A. Shivram. "Algorithms for clustering on the sphere: Advances & applications". In: *WCECS 2016 - World Congress on Engineering and Computer Science 2016*. Lecture Notes in Engineering and Computer Science. Newswood Limited, 2016, pp. 420–425.

[24] L. Grippo, F. Lampariello, and S. Lucidi. "A nonmonotone line search technique for Newton's method". In: *SIAM Journal on Numerical Analysis* 23.4 (1986), pp. 707–716.

[25] P. Gruber and F. J. Theis. "Grassmann clustering". In: *European Signal Processing Conference* (2006). 14th European Signal Processing Conference, EUSIPCO 2006 ; Conference date: 04-09-2006 Through 08-09-2006. ISSN: 2219-5491.

[26] S. Hosseini, W. Huang, and R. Yousefpour. "Line search algorithms for locally Lipschitz functions on Riemannian manifolds". In: *SIAM Journal on Optimization* 28.1 (2018), pp. 596–619. DOI: 10.1137/16M1108145.

[27] J. Hu, X. Liu, Z.-W. Wen, and Y.-X. Yuan. "A brief introduction to manifold optimization". In: *Journal of the Operations Research Society of China* 8 (2 20200), pp. 199–248. DOI: 10.1007/s40305-020-00295-9.

[28] W. Huang, P.-A. Absil, and K. A. Gallivan. "A Riemannian BFGS method for nonconvex optimization problems". In: *Numerical Mathematics and Advanced Applications ENUMATH 2015*. Cham: Springer International Publishing, 2016, pp. 627–634. ISBN: 978-3-319-39929-4.

[29] W. Huang and K. Wei. "Riemannian proximal gradient methods". In: *Mathematical Programming* 194 (1 2022), pp. 371–413. DOI: 10.1007/s10107-021-01632-3.

[30] L. Hubert and P. Arabie. "Comparing partitions". In: *Journal of Classification* 2 (1 1985), pp. 193–218. DOI: 10.1007/BF01908075.

[31] S. Jung, K. Park, and B. Kim. "Clustering on the torus by conformal prediction". In: *The Annals of Applied Statistics* 15.4 (2021), pp. 1583 –1603. DOI: 10.1214/21-AOAS1459.

[32] C. Kanzow and L. Lehmann. "Convergence of nonmonotone proximal gradient methods under the Kurdyka-Łojasiewicz property without a global Lipschitz assumption". In: *Journal of Optimization Theory and Applications* 207.4 (2025). DOI: 10.1007/s10957-025-02762-w.

[33] M. Kelly, R. Longjohn, and K. Nottingham. *The UCI Machine Learning Repository*. URL: https://archive.ics.uci.edu.

[34] P. Q. Khanh and T. T. Minh. "On projectional subdifferentials and projectional coderivatives with respect to smooth manifolds". In: *Journal of Optimization Theory and Applications* 208 (3 2026), p. 113. DOI: 10.1007/s10957-026-02941-3.

[35] H. A. Le Thi, T. Pham Dinh, and L. D. Muu. "Numerical solution for optimization over the efficient set by d.c. optimization algorithms". In: *Operations Research Letters* 19.3 (1996), pp. 117–128. ISSN: 0167-6377. DOI: 10.1016/0167-6377(96)00022-3.

[36] E. Levin, J. Kileel, and N Boumal. "Finding stationary points on bounded-rank matrices: a geometric hurdle and a smooth remedy". In: *Mathematical Programming* 199 (1 2023), pp. 831–864. DOI: 10.1007/s10107-022-01851-2.

[37] D. C. Liu and J. Nocedal. "On the limited memory BFGS method for large scale optimization". In: *Mathematical Programming* 45 (1 1989), pp. 503–528. DOI: 10.1007/BF01589116.

[38] K. W. Meng, M. H. Li, W. F. Yao, and X. Q. Yang. "Lipschitz-like property relative to a set and the generalized Mordukhovich criterion". In: *Mathematical Programming* 189 (1 2021), pp. 455–489. DOI: 10.1007/s10107-020-01568-0.

[39] B. S. Mordukhovich. *Variational Analysis and Applications*. Springer, 2018. DOI: 10.1007/978-3-319-92775-6.

[40] B.-S. Oh, A. B. J. Teoh, K.-A. Toh, and Z. Lin. "Grassmannian clustering for multivariate time sequences". In: *New Trends in Computer Technologies and Applications*. Springer Singapore, 2019, pp. 654–664. ISBN: 978-981-13-9190-3.

[41] B. Ordin and A. M. Bagirov. "A heuristic algorithm for solving the minimum sum-of-squares clustering problems". In: *Journal of Global Optimization* 61 (2 2015), pp. 341–361. DOI: 10.1007/s10898-014-0171-5.

[42] H. Oviedo. "Global convergence of Riemannian line search methods with a Zhang-Hager-type condition". In: *Numerical Algorithms* 91 (3 2022), pp. 1183–1203. DOI: 10.1007/s11075-022-01298-8.

[43] C. Qi, K. A. Gallivan, and P.-A. Absil. "Riemannian BFGS Algorithm with Applications". In: *Recent advances in optimization and its applications in engineering*. Springer Berlin Heidelberg, 2010, pp. 183–192. ISBN: 978-3-642-12598-0.

[44] Y. Qian, T. Tao, S. Pan, and H. Qi. "Convergence of ZH-Type nonmonotone descent method for Kurdyka–Łojasiewicz optimization problems". In: *SIAM Journal on Optimization* 35.2 (2025), pp. 1089–1109. DOI: 10.1137/24M1669153.

[45] W. M. Rand. "Objective criteria for the evaluation of clustering methods". In: *Journal of the American Statistical Association* 66.336 (1971), pp. 846–850. DOI: 10.1080/01621459.1971.10482356.

[46] R. T. Rockafellar and R. J.-B. Wets. *Variational Analysis*. Springer, 2009. DOI: 10.1007/978-3-642-02431-3.

[47] A. Rosenberg and J. Hirschberg. "V-Measure: A conditional entropy-based external cluster evaluation measure". In: *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*. Association for Computational Linguistics, 2007, pp. 410–420.

[48] S. Sengupta, S. Pal, R. Mitra, Y. Guo, A. Banerjee, and Y. Ji. *A Bayesian mixture model for clustering on the stiefel manifold*. 2017. arXiv: 1708.07196 [stat.ME].

[49] J. Townsend, N. Koep, and S. Weichwald. "Pymanopt: A Python Toolbox for Optimization on Manifolds using Automatic Differentiation". In: *Journal of Machine Learning Research* 17.137 (2016), 1–5. URL: http://jmlr.org/papers/v17/16-177.html.

[50] H. Wiechers, B. Eltzner, S. F. Huckemann, and K. V. Mardia. *Clustering schemes on the torus with application to RNA clashes*. 2021. arXiv: 2104.00094 [q-bio.BM].

[51] H. Zhang and W. W. Hager. "A nonmonotone line search technique and its application to unconstrained optimization". In: *SIAM Journal on Optimization* 14.4 (2004), pp. 1043–1056.

A Complete Proofs under the KL Property

Let us begin with a full proof of Lemma 4.3:

Lemma A.1. *For all sufficiently large $k \in \mathbb{N}$ with $\mathcal{R}_k < \varphi(x^*) + \eta$ and $x^k \in \overline{B}_\vartheta(x^*)$, the following inequality holds:*

$$\frac{1-\sqrt{1-p_{\min}}}{\sqrt{m}} \sum_{i=k}^{l(k)} \Xi_i \leq \frac{1}{2}\Xi_k + \sqrt{1-p_{\min}}\Xi_{k-1} + \hat{c}\Delta_{k,k+m}, \tag{A.1}$$

where $\hat{c}$ denotes the constant from Lemma 4.1.

Proof. As $x \mapsto \sqrt{x}$ is a concave function, the application of Jensen's inequality yields

$$\frac{1-\sqrt{1-p_{\min}}}{\sqrt{m}} \sum_{i=k}^{l(k)} \Xi_i \leq \big(1-\sqrt{1-p_{\min}}\big)\sqrt{\mathcal{R}_k - \mathcal{R}_{k+m}}. \tag{A.2}$$

We now distinguish two cases.

Case 1: $k \in K_1$. Then, we have $\varphi(x^k) \leq \mathcal{R}_{k+m}$, which implies

$$\begin{aligned}
\mathcal{R}_k - \mathcal{R}_{k+m} &= (1-p_k)\mathcal{R}_{k-1} + p_k\varphi(x^k) - \mathcal{R}_{k+m} \\
&\leq (1-p_k)\mathcal{R}_{k-1} + p_k\mathcal{R}_{k+m} - \mathcal{R}_{k+m} \\
&= (1-p_k)(\mathcal{R}_{k-1} - \mathcal{R}_{k+m}) \\
&\leq (1-p_{\min})(\mathcal{R}_{k-1} - \mathcal{R}_{k+m}) \qquad \text{(recall that } \mathcal{R}_{k-1} \geq \mathcal{R}_{k+m}) \\
&= (1-p_{\min})(\mathcal{R}_{k-1} - \mathcal{R}_k + \mathcal{R}_k - \mathcal{R}_{k+m}).
\end{aligned}$$

Using $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ for all $x, y \in \mathbb{R}_{\geq 0}$, we obtain

$$\big(1-\sqrt{1-p_{\min}}\big)\sqrt{\mathcal{R}_k - \mathcal{R}_{k+m}} \leq \sqrt{1-p_{\min}}\,\Xi_{k-1}.$$

The statement therefore follows from (A.2).

Case 2: $k \in K_2$. We then have $\varphi(x^*) \leq \mathcal{R}_{k+m} < \varphi(x^k) \leq \mathcal{R}_k < \varphi(x^*) + \eta$ by assumption. Using the KL property of Φ and the fact that Φ agrees with φ on $\mathcal{M}$, we get

$$\chi'\big(\varphi(x^k) - \varphi(x^*)\big)\operatorname{dist}\big(0, \partial\Phi(x^k)\big) = \chi'\big(\Phi(x^k) - \Phi(x^*)\big)\operatorname{dist}\big(0, \partial\Phi(x^k)\big) \geq 1.$$

As x^k was assumed to be in $\overline{B}_\vartheta(x^*)$, by application of (4.4), we have

$$\chi'(\varphi(x^k) - \varphi(x^*)) \geq \frac{1}{\frac{2b}{\tau_C}\|x^{k+1} - x^k\|}. \tag{A.3}$$

(recall that $x^{k+1} \neq x^k$ since the algorithm is assumed to generate an infinite sequence). Using the properties of χ, we now obtain

$$\begin{aligned}
\Delta_{k,k+m} &= \chi\big(\mathcal{R}_k - \varphi(x^*)\big) - \chi\big(\mathcal{R}_{k+m} - \varphi(x^*)\big) \\
&\geq \chi\big(\varphi(x^k) - \varphi(x^*)\big) - \chi\big(\mathcal{R}_{k+m} - \varphi(x^*)\big) \\
&\geq \chi'\big(\varphi(x^k) - \varphi(x^*)\big)\big(\varphi(x^k) - \mathcal{R}_{k+m}\big) \\
&\geq \frac{\varphi(x^k) - \mathcal{R}_{k+m}}{\frac{2b}{\tau_C}\|x^{k+1} - x^k\|},
\end{aligned}$$

where the first inequality results from the monotonicity of χ, the next one exploits the concavity of χ, and the final estimate exploits (A.3) together with the fact that $\varphi(x^k) - \mathcal{R}_{k+m} > 0$ in the case under consideration. Thus, with (4.2), we get

$$\varphi(x^k) - \mathcal{R}_{k+m} \leq \frac{2\hat{c}}{p_{\min}}\Xi_k\Delta_{k,k+m},$$

from the definition of $\hat{c}$. Similar to the first case, our aim is to bound the difference $\mathcal{R}_k - \mathcal{R}_{k+m}$. Using the fact that $\varphi(x^k) \leq \mathcal{R}_{k-1}$ by the acceptance criterion for our stepsize computation as well as $p_k \geq p_{\min}$, we have

$$p_k\varphi(x^k) + (1-p_k)\mathcal{R}_{k-1} \leq p_{\min}\varphi(x^k) + (1-p_{\min})\mathcal{R}_{k-1}.$$

Together with the definition of $\mathcal{R}_k := (1-p_k)\mathcal{R}_{k-1} + p_k\varphi(x^k)$, this yields

$$\begin{aligned}
\mathcal{R}_k - \mathcal{R}_{k+m} &= p_k\varphi(x^k) + (1-p_k)\mathcal{R}_{k-1} - \mathcal{R}_{k+m} \\
&\leq p_{\min}\varphi(x^k) + (1-p_{\min})\mathcal{R}_{k-1} - p_{\min}\mathcal{R}_{k+m} - (1-p_{\min})\mathcal{R}_{k+m} \\
&= p_{\min}\big(\varphi(x^k) - \mathcal{R}_{k+m}\big) + (1-p_{\min})(\mathcal{R}_{k-1} - \mathcal{R}_{k+m}) \\
&\leq 2\hat{c}\Xi_k\Delta_{k,k+m} + (1-p_{\min})(\mathcal{R}_{k-1} - \mathcal{R}_{k+m}) \\
&= 2\hat{c}\Xi_k\Delta_{k,k+m} + (1-p_{\min})(\mathcal{R}_{k-1} - \mathcal{R}_k + \mathcal{R}_k - \mathcal{R}_{k+m}).
\end{aligned}$$

Taking square roots on both sides and using $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ for all $x, y \in \mathbb{R}_{\geq 0}$, we obtain

$$\sqrt{\mathcal{R}_k - \mathcal{R}_{k+m}} \leq \sqrt{2\hat{c}\Xi_k\Delta_{k,k+m}} + \sqrt{1-p_{\min}}\big(\sqrt{\mathcal{R}_{k-1} - \mathcal{R}_k} + \sqrt{\mathcal{R}_k - \mathcal{R}_{k+m}}\big),$$

so we have

$$\big(1 - \sqrt{1-p_{\min}}\big)\sqrt{\mathcal{R}_k - \mathcal{R}_{k+m}} \leq \sqrt{2\hat{c}\Xi_k\Delta_{k,k+m}} + \sqrt{1-p_{\min}}\Xi_{k-1}.$$

Exploiting the inequality $2\sqrt{xy} \leq x + y$ for all $x, y \in \mathbb{R}_{\geq 0}$, this yields

$$\big(1 - \sqrt{1-p_{\min}}\big)\sqrt{\mathcal{R}_k - \mathcal{R}_{k+m}} \leq \frac{1}{2}\Xi_k + \sqrt{1-p_{\min}}\Xi_{k-1} + \hat{c}\Delta_{k,k+m}.$$

In view of (A.2), this completes the proof. □

Theorem A.2. *Let Assumptions 2.10 and 3.1 hold, let $\{x^k\}_K$ be a subsequence converging to some accumulation point x^*, and suppose that $\Phi := \varphi + \delta_{\mathcal{M}}$ satisfies the KL property at x^*. Then the entire sequence $\{x^k\}$ converges to x^*.*

Proof. Let k_0 be the index from the definition of ϑ, cf. Lemma 4.1. Without loss of generality, we may assume that k_0 is large enough such that the results from Lemma 4.2 and Lemma 4.3 hold and that $\mathcal{R}_{k_0} < \varphi(x^*) + \eta$. We now claim that the following statements hold:

(a) for all $k \geq k_0 - 1$: $x^k \in \overline{B}_\vartheta(x^*)$, and

(b) for all $k \geq l(k_0)$:

$$\big(1-\sqrt{1-p_{\min}}\big)\sqrt{m}\sum_{j=l(k_0)}^{k}\Xi_j \leq \sum_{j=k_0}^{k}\left(\frac{1}{2}\Xi_j + \sqrt{1-p_{\min}}\Xi_{j-1}\right) + \hat{c}\sum_{j=k_0}^{l(k_0)}\chi\big(\mathcal{R}_j - \varphi(x^*)\big), \tag{A.4}$$

where $\hat{c}$ denotes the constant from Lemma 4.1. We verify these two statements simultaneously by induction over k.

For $k = k_0 - 1$, the term $\|x^{k_0-1} - x^*\|$ is obviously smaller than the constant ϑ in Lemma 4.1. Hence $x^{k_0-1} \in \overline{B}_\vartheta(x^*)$. Now, assuming $x^j \in \overline{B}_\vartheta(x^*)$ for all j from $k_0 - 1$ to some $k \in \{k_0 - 1, \ldots, l(k_0)\}$, we can apply (4.2) to obtain

$$\begin{aligned}
\|x^k - x^*\| &\leq \|x^{k_0-1} - x^*\| + \sum_{j=k_0}^{k}\|x^j - x^{j-1}\| \\
&\leq \|x^{k_0-1} - x^*\| + \frac{1}{e}\sum_{j=k_0}^{k}\Xi_{j-1} \\
&\leq \|x^{k_0-1} - x^*\| + \frac{1}{e}\sum_{j=k_0}^{l(k_0)}\Xi_{j-1} \leq \vartheta.
\end{aligned}$$

This shows the first statement for $k = k_0 - 1, \dots, l(k_0)$. Now, we can apply Lemma 4.3 for indices $k = k_0, \dots, l(k_0)$ and obtain

$$\begin{aligned}\big(1 - \sqrt{1-p_{\min}}\big)\sqrt{m}\Xi_{l(k_0)} &\le \frac{1-\sqrt{1-p_{\min}}}{\sqrt{m}} \sum_{j=k_0}^{l(k_0)} \sum_{i=j}^{l(j)} \Xi_i \\ &\le \sum_{j=k_0}^{l(k_0)} \left(\frac{1}{2}\Xi_j + \sqrt{1-p_{\min}}\Xi_{j-1}\right) + \hat{c}\sum_{j=k_0}^{l(k_0)} \Delta_{j,j+m} \\ &\le \sum_{j=k_0}^{l(k_0)} \left(\frac{1}{2}\Xi_j + \sqrt{1-p_{\min}}\Xi_{j-1}\right) + \hat{c}\sum_{j=k_0}^{l(k_0)} \chi\big(\mathcal{R}_j - \varphi(x^*)\big),\end{aligned}$$

where the first inequality follows from the fact that the term $\Xi_{l(k_0)}$ occurs m times within the double sum on the right-hand side, whereas the other expressions Ξ_i are nonnegative, and the second inequality is obtained by summing (A.1) in Lemma 4.3 from k_0 to $l(k_0)$. In the final estimate, we simply omit some nonpositive terms. This shows that the second statement holds for $k = l(k_0)$.

Suppose that the first statement holds for all j from $k_0 - 1$ to some $k \ge l(k_0)$ and that the second statement is true for $k \ge l(k_0)$. We first show that the second statement for k implies the first statement for $k+1$. Using (A.4), we get

$$\begin{aligned}&\big(1 - \sqrt{1-p_{\min}}\big)\sqrt{m} \sum_{j=l(k_0)}^{k} \Xi_j \\ &\quad\le \sum_{j=k_0}^{k} \left(\frac{1}{2}\Xi_j + \sqrt{1-p_{\min}}\Xi_{j-1}\right) + \hat{c}\sum_{j=k_0}^{l(k_0)} \chi\big(\mathcal{R}_j - \varphi(x^*)\big) \\ &\quad\le \sum_{j=k_0}^{l(k_0)} \left(\frac{1}{2}\Xi_j + \sqrt{1-p_{\min}}\Xi_{j-1}\right) + \left(\frac{1}{2} + \sqrt{1-p_{\min}}\right) \sum_{j=l(k_0)}^{k} \Xi_j + \hat{c}\sum_{j=k_0}^{l(k_0)} \chi\big(\mathcal{R}_j - \varphi(x^*)\big).\end{aligned}$$

This implies

$$\begin{aligned}&\left(\left(1 - \sqrt{1-p_{\min}}\right)\sqrt{m} - \left(\frac{1}{2} + \sqrt{1-p_{\min}}\right)\right) \sum_{j=l(k_0)}^{k} \Xi_j \\ &\qquad\le \sum_{j=k_0}^{l(k_0)} \left(\frac{1}{2}\Xi_j + \sqrt{1-p_{\min}}\Xi_{j-1}\right) + \hat{c}\sum_{j=k_0}^{l(k_0)} \chi\big(\mathcal{R}_j - \varphi(x^*)\big).\end{aligned}$$

Noting that, by definition of m, it holds that $\big(1-\sqrt{1-p_{\min}}\big)\sqrt{m} - \big(1/2 + \sqrt{1-p_{\min}}\big) \ge 1/2$ and that we obviously have $\sqrt{1-p_{\min}} \le 1$, we get

$$\sum_{j=l(k_0)}^{k} \Xi_j \le \sum_{j=k_0}^{l(k_0)} (\Xi_j + 2\Xi_{j-1}) + 2\hat{c}\sum_{j=k_0}^{l(k_0)} \chi\big(\mathcal{R}_j - \varphi(x^*)\big).$$

This implies

$$\sum_{j=k_0-1}^{k} \Xi_j = \sum_{j=k_0}^{l(k_0)} \Xi_{j-1} + \sum_{j=l(k_0)}^{k} \Xi_j \le \sum_{j=k_0}^{l(k_0)} (3\Xi_{j-1} + \Xi_j) + 2\hat{c}\sum_{j=k_0}^{l(k_0)} \chi\big(\mathcal{R}_j - \varphi(x^*)\big). \qquad \text{(A.5)}$$

As we assumed that the first statement and by consequence $x^j \in \overline{B}_\vartheta(x^*) \cap \mathcal{M} \subset C$ is true for all $j \in \{k_0 - 1, \dots, k\}$, we obtain from (4.2) together with the equation above that

$$\begin{aligned} \|x^{k+1} - x^*\| &\leq \|x^{k_0-1} - x^*\| + \sum_{j=k_0-1}^{k} \|x^{j+1} - x^j\| \leq \|x^{k_0-1} - x^*\| + \frac{1}{e} \sum_{j=k_0-1}^{k} \Xi_j \\ &\leq \|x^{k_0-1} - x^*\| + \frac{1}{e} \sum_{j=k_0}^{l(k_0)} (3\Xi_{j-1} + \Xi_j) + \frac{2\hat{c}}{e} \sum_{j=k_0}^{l(k_0)} \chi\big(\mathcal{R}_j - \varphi(x^*)\big). \end{aligned} \tag{A.6}$$

The expression on the right-hand side is precisely the constant ϑ from Lemma 4.1. Thus, the first statement holds for $k+1$.

We next verify the second claim for $k+1$. Since we already know that $x^j \in \overline{B}_\vartheta(x^*)$ is true for all $j \in \{k_0 - 1, \dots, k+1\}$, we again apply Lemma 4.3 and sum over (A.1), now from k_0 to $k+1$. This yields

$$\begin{aligned} \big(1 - \sqrt{1 - p_{\min}}\big)\sqrt{m} \sum_{j=l(k_0)}^{k+1} \Xi_j &\leq \frac{1 - \sqrt{1 - p_{\min}}}{\sqrt{m}} \sum_{j=k_0}^{k+1} \sum_{i=j}^{l(j)} \Xi_i \\ &\leq \sum_{j=k_0}^{k+1} \left(\frac{1}{2}\Xi_j + \sqrt{1 - p_{\min}}\Xi_{j-1} \right) + \hat{c} \sum_{j=k_0}^{k+1} \Delta_{j,j+m} \\ &\leq \sum_{j=k_0}^{k+1} \left(\frac{1}{2}\Xi_j + \sqrt{1 - p_{\min}}\Xi_{j-1} \right) + \hat{c} \sum_{j=k_0}^{l(k_0)} \chi\big(\mathcal{R}_j - \varphi(x^*)\big), \end{aligned}$$

where the first inequality results from the fact that each term Ξ_j for $j = l(k_0), \dots, k+1$ from the left-hand side occurs m times within the double sum from the right-hand side (observe that the relation $l(j+1) = l(j) + 1$ holds for all j), whereas the remaining expressions Ξ_i are nonnegative, the next inequality exploits (A.1) from Lemma 4.3, and the final estimate uses a telescoping sum argument where we omit some nonpositive summands. This completes our induction step.

Hence, it follows that $x^k \in \overline{B}_\vartheta(x^*)$ for all $k \geq k_0 - 1$. Taking $k \to \infty$ in the resulting expression for $\sum_{j=k_0-1}^{k} \|x^{j+1} - x^j\|$ from (A.6) then shows that $\{x^k\}_{k\in\mathbb{N}}$ is a Cauchy sequence and, therefore, convergent. Thus, the accumulation point x^* is the limit of the entire sequence $\{x^k\}$. □

For completeness, we provide a proof for the following rate-of-convergence result in Theorem A.4 for the case where the desingularization function is given by $\chi(t) = ct^\theta$ for some $c > 0$ and $\theta \in (0, 1)$. Note, however, that it is also a direct consequence of the corresponding more general result found in [44]. We need the following technical result from [3, Lemma 1]:

Lemma A.3. *Let $\{s_k\} \subseteq [0, \infty)$ be any monotonically decreasing sequence satisfying $s_k \to 0$ and*

$$s_k^\alpha \leq \beta\big(s_k - s_{k+1}\big) \quad \textit{for all } k \textit{ sufficiently large,}$$

where $\alpha, \beta > 0$ are suitable constants. Then the following statements hold:

(a) If $\alpha \in (0, 1]$, the sequence $\{s_k\}$ converges linearly to zero with rate $1 - \frac{1}{\beta}$.

(b) If $\alpha > 1$, there exists a constant $\eta > 0$ such that

$$s_k \leq \eta k^{-\frac{1}{\alpha-1}} \quad \textit{for all } k \textit{ sufficiently large.}$$

Theorem A.4. *Let Assumptions 2.10 and 3.1 hold, and suppose that $\{x^k\}$ converges on some subsequence $\{x^k\}_K$ to a limit point x^* such that Φ has the KL property at x^*. Then the entire sequence $\{x^k\}$ converges to x^*. Further, if the corresponding desingularization function is given by $\chi(t) = ct^\theta$ (for some $c > 0$ and $\theta \in (0, 1)$), then the following statements hold:*

(a) If $\theta \in [1/2, 1)$, then $\{\mathcal{R}_k\}$ converges R-linearly to $\varphi(x^*)$ and $\{x^k\}$ converges R-linearly to x^*.

(b) If $\theta \in (0, 1/2)$, then there exist constants $\eta_1, \eta_2 > 0$ such that for all k large enough it holds that

$$\mathcal{R}_k - \varphi(x^*) \leq \eta_1 k^{-\frac{1}{1-2\theta}}, \tag{A.7}$$

$$\|x^k - x^*\| \leq \eta_2 k^{-\frac{\theta}{1-2\theta}}. \tag{A.8}$$

Proof. Taking Theorem 4.4 into account, we only need to verify statements (a) and (b). As a first step, let us prove the results for the sequence $\{\mathcal{R}_k\}$. We first claim that for $\theta \in (0, 1)$ and with

$$\omega := \Big(\frac{e\underline{\tau}_C}{2c\theta b}\Big)^2,$$

it holds that

$$\varphi(x^k) - \varphi(x^*) \leq \Big(\frac{1}{\omega}\Big)^{\frac{1}{2(1-\theta)}} \big(\mathcal{R}_k - \mathcal{R}_{k+1}\big)^{\frac{1}{2(1-\theta)}} \tag{A.9}$$

for all $k \in \mathbb{N}$ sufficiently large. In fact, if $\varphi(x^k) \leq \varphi(x^*)$ holds, then the left-hand side of (A.9) is nonpositive, hence, the claim holds trivially. By previous results, we may assume that k is large enough such that $x^k \in \overline{B}_\vartheta(x^*)$ and $\varphi(x^*) < \varphi(x^k) < \varphi(x^*) + \eta$ hold.

As Φ satisfies the KL property at x^* with $\chi(t) = ct^\theta$ and as $\Phi(x) = \varphi(x)$ for all $x \in \mathcal{M}$, we have

$$\begin{aligned} 1 &\leq \chi'\big(\varphi(x^k) - \varphi(x^*)\big)\operatorname{dist}\big(0, \partial\Phi(x^k)\big) \\ &= c\theta\big(\varphi(x^k) - \varphi(x^*)\big)^{\theta-1}\operatorname{dist}\big(0, \partial\Phi(x^k)\big). \end{aligned}$$

By (4.4), this yields

$$1 \leq c\theta\frac{2b}{\underline{\tau}_C}\big(\varphi(x^k) - \varphi(x^*)\big)^{\theta-1}\|x^{k+1} - x^k\|,$$

which gives the inequality

$$\|x^{k+1} - x^k\| \geq \frac{1}{c\theta\frac{2b}{\underline{\tau}_C}}\big(\varphi(x^k) - \varphi(x^*)\big)^{1-\theta}. \tag{A.10}$$

Further, from (4.2), we have

$$\mathcal{R}_{k+1} - \mathcal{R}_k \leq -e^2\|x^{k+1} - x^k\|^2. \tag{A.11}$$

Combination of (A.10) and (A.11) yields

$$\begin{aligned}\mathcal{R}_{k+1} - \mathcal{R}_k &\leq -e^2\|x^{k+1} - x^k\|^2 \\ &\leq -e^2\frac{1}{c^2\theta^2\frac{4b^2}{\underline{\tau}_C^2}}\big(\varphi(x^k) - \varphi(x^*)\big)^{2(1-\theta)} \\ &= -\omega\big(\varphi(x^k) - \varphi(x^*)\big)^{2(1-\theta)}.\end{aligned}$$

Rearranging these terms shows that the claim (A.9) holds.

Next recall that, by the acceptance criterion for the stepsize τ_k, we always have $\varphi(x^{k+1}) \leq \mathcal{R}_k$. Hence, it follows that

$$\mathcal{R}_{k+1} = (1 - p_{k+1})\mathcal{R}_k + p_{k+1}\varphi(x^{k+1}) \leq (1 - p_{\min})\mathcal{R}_k + p_{\min}\varphi(x^{k+1}). \tag{A.12}$$

Denote by $\{s_k\}$ the sequence defined by $s_k := \mathcal{R}_k - \varphi(x^*) \geq 0$. Then $s_k \to 0$ monotonically, and we obtain

$$\begin{aligned}s_{k+1} &\leq (1 - p_{\min})s_k + p_{\min}(\varphi(x^{k+1}) - \varphi(x^*)) \\ &\leq (1 - p_{\min})s_k + p_{\min}\Big(\frac{1}{\omega}\Big)^{\frac{1}{2(1-\theta)}}\big(s_{k+1} - s_{k+2}\big)^{\frac{1}{2(1-\theta)}},\end{aligned}$$

where the first inequality follows from (A.12) and the second one results from (A.9). By the monotonicity of $\{s_k\}$, this implies

$$s_{k+2} \leq (1 - p_{\min})s_k + p_{\min}\Big(\frac{1}{\omega}\Big)^{\frac{1}{2(1-\theta)}}\big(s_k - s_{k+2}\big)^{\frac{1}{2(1-\theta)}},$$

which, in turn, yields

$$\begin{aligned}s_k &\leq \frac{1}{p_{\min}}(s_k - s_{k+2}) + \Big(\frac{1}{\omega}\Big)^{\frac{1}{2(1-\theta)}}\big(s_k - s_{k+2}\big)^{\frac{1}{2(1-\theta)}} \\ &\leq \left(\frac{1}{p_{\min}} + \Big(\frac{1}{\omega}\Big)^{\frac{1}{2(1-\theta)}}\right)(s_k - s_{k+2})^{\min\{1,\frac{1}{2(1-\theta)}\}}\end{aligned}$$

for all k sufficiently large. As for all $a, b > 0$ it holds that $1/\min\{a, b\} = \max\{1/a, 1/b\}$, it follows that

$$s_k^{\max\{1,2(1-\theta)\}} \leq \gamma(s_k - s_{k+2}),$$

where

$$\gamma := \left(\frac{1}{p_{\min}} + \Big(\frac{1}{\omega}\Big)^{\frac{1}{2(1-\theta)}}\right)^{\max\{1,2(1-\theta)\}} > 0$$

is a constant.

By considering even and odd indices separately, we are now in the setting of Lemma A.3 and immediately obtain the corresponding rate-of-convergence results for the sequence $\{\mathcal{R}_k\}$ as $\theta \in (0, 1/2)$ implies that $2(1-\theta) > 1$ and $\theta \in [1/2, 1)$ implies that $2(1-\theta) \in (0, 1]$.

Let us now verify the statements for the sequence $\{x^k\}$. In view of Theorem 4.4, the equations in the proof of that result remain valid if $k_0 - 1$ is replaced by some k sufficiently large. Note that $\Xi_{j-1} \leq \sqrt{\mathcal{R}_{j-1} - \varphi(x^*)} = \sqrt{s_{j-1}}$. Taking the monotonicity

of the function χ and the sequences $\{\mathcal{R}_k\}$ and thus $\{s_k\}$ into account, it follows from (A.5) in combination with (4.2) and the definition of χ that, for $l > k$:

$$\begin{aligned}
\|x^k - x^l\| &\leq \sum_{j=k}^{l-1} \|x^{j+1} - x^j\| \leq \frac{1}{e} \sum_{j=k}^{l-1} \Xi_j \\
&\leq \frac{1}{e} \sum_{j=k+1}^{l(k+1)} (3\Xi_{j-1} + \Xi_j) + \frac{2\hat{c}}{e} \sum_{j=k+1}^{l(k+1)} \chi\big(\mathcal{R}_j - \varphi(x^*)\big) \\
&\leq \frac{4}{e} \sum_{j=k+1}^{l(k+1)} \sqrt{s_{j-1}} + \frac{2\hat{c}}{e} \sum_{j=k+1}^{l(k+1)} \chi(s_j) \\
&\leq \frac{4m}{e} \sqrt{s_k} + \frac{2\hat{c}m}{e} \chi(s_k) \\
&\leq \tilde{\eta} s_k^{\min\{1/2,\theta\}}
\end{aligned}$$

for all k sufficiently large, where

$$\tilde{\eta} := \frac{4 + 2\hat{c}c}{e} m.$$

Taking now $l \to \infty$, together with the corresponding rate-of-convergence results for $\{s_k\}$ from the first part, this completes the proof. □