Convergence of Nonmonotone Proximal Gradient Methods under the Kurdyka–Łojasiewicz Property without a Global Lipschitz Assumption

Christian Kanzow[*] Leo Lehmann[†]

November 18, 2024

Abstract

We consider the composite minimization problem with the objective function being the sum of a continuously differentiable and a merely lower semicontinuous and extended-valued function. The proximal gradient method is probably the most popular solver for this class of problems. Its convergence theory typically requires that either the gradient of the smooth part of the objective function is globally Lipschitz continuous or the (implicit or explicit) a priori assumption that the iterates generated by this method are bounded. Some recent results show that, without these assumptions, the proximal gradient method, combined with a monotone stepsize strategy, is still globally convergent with a suitable rate-of-convergence under the Kurdyka–Łojasiewicz property. For a nonmonotone stepsize strategy, there exist some attempts to verify similar convergence results, but, so far, they need stronger assumptions. This paper is the first which shows that nonmonotone proximal gradient methods for composite optimization problems share essentially the same nice global and rate-of-convergence properties as its monotone counterparts, still without assuming a global Lipschitz assumption and without an a priori knowledge of the boundedness of the iterates.

Keywords. Composite Optimization, Nonsmooth Optimization, Proximal Gradient Method, Kurdyka–Łojasiewicz Property, Nonmonotone Line Search, Global Convergence, Linear Convergence

AMS subject classifications. 49J52, 90C30

1 Introduction

We consider the composite optimization problem

$$\min_x \psi(x) := f(x) + \phi(x), \quad x \in \mathbb{X}, \tag{1.1}$$

[*]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@stud-mail.uni-wuerzburg.de

where $f : \mathbb{X} \to \mathbb{R}$ is continuously differentiable, $\phi : \mathbb{X} \to \overline{\mathbb{R}}$ is proper and lower semicontinuous, and $\mathbb{X}$ denotes a Euclidean space (finite-dimensional Hilbert space). Note that neither f nor ϕ has to be convex. Composite optimization problems of this kind occur frequently in many applications like machine learning, data compression, matrix completion, image processing, low-rank approximation, or dictionary learning [12, 13, 16, 17, 23, 24, 25].

In many applications, the nonsmooth term ϕ plays the role of a sparsity, regularization, or penalty term. For example, the standard choice $\phi(x) := \lambda \|x\|_1$ for $x \in \mathbb{R}^n$ and a constant $\lambda > 0$ is known to impose some sparsity in different applications. On the other, improved sparsity can often be obtained by using terms like $\phi(x) := \lambda \|x\|_p$ with $p \in (0,1)$ or $\phi(x) := \lambda \|x\|_0$, with $\|x\|_0$ denoting the number of nonzero components of the vector x. Note that these two latter functions ϕ are nonconvex, the second one is even discontinuous, but lower semicontinuous. The same is true if ϕ represents the indicator function of a nonempty and closed feasible set. Furthermore, we stress that the standard global Lipschitz assumption on the derivative of f is satisfied for quadratic objective functions, but usually violated for more general nonlinear functions f. In particular, this non-Lipschitz behaviour occurs if f stands for an augmented Lagrangian function or for the dual of a (not necessarily uniformly convex) primal problem, cf. [15, 19]. This indicates that it is reasonable to consider composite optimization problems in this general setting.

The standard solver for composite optimization problems is the proximal gradient method given by the iteration

$$x^{k+1} := \operatorname{argmin}_{x \in \mathbb{X}} \Big\{ f(x^k) + \big\langle f'(x^k), x - x^k \big\rangle + \frac{1}{2\gamma_k} \|x - x^k\|^2 + \phi(x) \Big\} \tag{1.2}$$

for some given $\gamma_k > 0$, i.e., the new iterate x^{k+1} is obtained by using a simple quadratic approximation of the nonlinear and smooth function f, whereas the nonsmooth function ϕ is moved into the subproblem without any modifications. Using some simple algebraic manipulations, it is well-known and easy to see that this procedure can be rewritten as

$$x^{k+1} := \operatorname{Prox}_{\gamma_k \phi}\big(x^k - \gamma_k f'(x^k)\big) \tag{1.3}$$

with the prox operator defined by

$$\operatorname{Prox}_{\gamma\phi}(x) := \operatorname{argmin}_{z \in \mathbb{X}} \Big\{ \phi(z) + \frac{1}{2\gamma} \|z - x\|^2 \Big\}.$$

Observe that (1.2) reduces to the steepest descent method $x^{k+1} := x^k - \gamma_k f'(x^k)$ for $\phi \equiv 0$, which indicates that the parameter $\gamma_k > 0$ may be viewed as a stepsize.

Note that the proximal gradient method allows an efficient implementation only if the prox operator can be evaluated in a simple way. Fortunately, there exist several practically relevant scenarios where this prox operator can be computed either analytically or with a very small computational overhead, cf. the excellent monograph [5] by Beck for a list of examples. The same monograph is also a perfect reference for the existing convergence theory of proximal point methods, at least for the case of convex functions ϕ.

Extensions of the convergence theory to nonconvex and only lower semicontinuous functions ϕ occur in the seminal papers [4, 9]. Both papers cover global and rate-of-convergence results, and their technique is based on a global Lipschitz assumption for the gradient of f and the *Kurdyka–Łojasiewicz (KL) property.* While the latter seems indispensable, the former assumption is very strong especially for applications with non-quadratic functions f. In some other papers, this kind of global Lipschitz assumption is hidden by the a priori condition that the iterates remain bounded or the assumption that the level sets of ψ are bounded, which itself implies the boundedness of the iterates.

In the recent paper [21], it was shown that global convergence can also be obtained without this global Lipschitz assumption, only the much weaker condition of f' begin locally Lipschitz continuous is required (depending on the smoothness properties of ϕ, it might also be enough to have f only continuously differentiable). Moreover, the subsequent work [19] also shows that one can obtain the standard global and rate-of-convergence results under the KL property, again without the assumption that the derivative of f is globally Lipschitz continuous. These results were obtained for a proximal gradient method with a suitably updated stepsize parameter γ_k in such a way that the sequence of function values $\{\psi(x^k)\}$ is monotonically decreasing.

The situation is more delicate when we allow a nonmonotone stepsize rule, i.e., a choice of γ_k such that the sequence $\{\psi(x^k)\}$ is no longer monotonically decreasing. The two most prominent monotonicity strategies are the max-type rule by Grippo et al. [18] as well as the mean-type rule [33] by Zhang and Hager. Both were originally introduced for smooth unconstrained optimization problems and shown to provide global convergence under the same set of assumptions. A closer inspection indicates, however, that the mean-type rule requires slightly weaker conditions than the max-type rule. This is also reflected by the two recent papers [21, 14]. While [21] considers a nonmonotone proximal gradient method with the max-type rule, the subsequent work [14] investigates a nonmonotone proximal gradient method with the mean-type rule. Both papers show that each accumulation point is a suitable stationary point of the objective function ψ, but [21] requires a stronger (uniform continuity) assumption, whereas [14] shows that the convergence theory for the monotone proximal gradient method can be adapted to the mean-type nonmonotone version without any extra condition.

The aim of this paper is to present a convergence and a rate-of-convergence result for the mean-type nonmonotone proximal gradient method under the KL property, and without assuming a global Lipschitz assumption regarding f' or an a priori condition like the boundedness of the iterates x^k. To the best of our knowledge, this is the first time that such results are shown without any of these extra conditions. Our results are build on a combination of ideas from the convergence theory in the monotone setting as in [19], recent contributions to mean-type nonmonotone proximal gradient methods in [14] and a refinement of the (global) convergence theory in [30] for our setting with weaker assumptions. Unfortunately, the usual analysis for monotone methods based on the KL property is already technical, and the introduction of the nonmonotonicity complicates things even further. Nevertheless, we feel it is worth going through this analysis taking into account that nonmonotone methods often outperform their monotone versions in practical applications, see [14, 29] for numerical comparisons in the context of proximal gradient methods. Taking this into account, we note that the current paper is a purely theoretical work providing an improved insight into the behaviour of a well-established method for the solution of composite optimization problems.

The paper is organized in the following way. We first recall some background material in Section 2. We then state our nonmonotone proximal gradient method in Section 3 and recall some of its basic properties. The corresponding convergence analysis is then presented in Section 4. We close with some final remarks in Section 5.

Notation: We write $\langle x, y\rangle$ for the scalar product of two elements $x, y \in \mathbb{X}$, and $\|x\|$ for the corresponding norm of $x \in \mathbb{X}$. The induced distance of a point $x \in \mathbb{X}$ to a nonempty set $S \subseteq \mathbb{X}$ is denoted by $\operatorname{dist}(x, S) := \inf_{y\in S} \|x - y\|$. The closed ball around a given point $x \in \mathbb{X}$ with radius $r > 0$ is denoted by $B_r(x)$. Given a differentiable mapping $f : \mathbb{X} \to \mathbb{R}$, we write $f'(x)$ for its derivative at $x \in \mathbb{X}$. 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 $\theta : \mathbb{X} \to \overline{\mathbb{R}}$, we call $\operatorname{dom}(\theta) := \{x \in \mathbb{X} \mid \theta(x) < \infty\}$ the *domain* of θ. The function θ is said to be *proper* if $\operatorname{dom}(\theta)$ is nonempty. Throughout this manuscript, we identify the dual space $\mathbb{X}^*$ with $\mathbb{X}$ itself.

2 Background Material

We first recall that a sequence $\{x^k\} \subseteq \mathbb{X}$ converges (locally) *Q-linearly* to some limit $x^* \in \mathbb{X}$ 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 [26, 31] for more details.

Given a proper, lower semicontinuous function $\theta : \mathbb{X} \to \overline{\mathbb{R}}$ and any $x \in \operatorname{dom}(\theta)$, we call

$$\hat{\partial}\theta(x) := \Big\{ v \in \mathbb{X} \,\Big|\, \liminf_{y\to x, y\neq x} \frac{\theta(y) - \theta(x) - \langle v, y - x\rangle}{\|y - x\|} \geq 0 \Big\}$$

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

$$\partial\theta(x) := \big\{ v \in \mathbb{X} \mid \exists x^k, v^k \in \mathbb{X} : x^k \to x, \theta(x^k) \to \theta(x), v^k \in \hat{\partial}\theta(x^k)\ \forall k \big\}$$

is called the *limiting, Mordukhovich*, or *basic subdifferential* of f at x. Recall that both subdifferentials coincide with the convex subdifferential for convex functions θ, i.e., in this case, we have

$$\hat{\partial}\theta(x) = \partial\theta(x) = \big\{ v \in \mathbb{X} \,\big|\, \theta(y) \geq \theta(x) + \langle v, y - x\rangle\ \forall y \in \operatorname{dom}(\theta) \big\},$$

whereas for general nonconvex functions, only the inclusion $\hat{\partial}\theta(x) \subseteq \partial\theta(x)$ holds. These two subdifferentials have different properties, e.g., $\hat{\partial}\theta(x)$ might be empty even for locally Lipschitz continuous functions, whereas $\partial\theta(x)$ is known to be nonempty in this situation. Moreover, the limiting subdifferential is *robust* in the following sense: Given a sequence $\{x^k\}$ converging to some limit x and a corresponding sequence $\{v^k\}$ with $v^k \in \partial\theta(x^k)$ for all $k \in \mathbb{N}$ such that $v^k \to v$ for some $v \in \mathbb{X}$, then the inclusion $v \in \partial\theta(x)$ holds. This robustness property is highly important in the convergence theory for many algorithms in the context of nonsmooth optimization, and simple examples show that, in general, it is not shared by the Fréchet subdifferential.

Given a lower semicontinuous function $\theta : \mathbb{X} \to \overline{\mathbb{R}}$ and a local minimum x^* of θ, it follows immediately from the definition of the Fréchet subdifferential that $0 \in \hat{\partial}\theta(x^*)$ holds. In particular, we therefore get $0 \in \partial\theta(x^*)$. Any point x^* satisfying this relation is called an *M-stationary point* (M = Mordukhovich) or simply a *stationary point* of θ.

Now, coming back to our composite optimization problem where $\psi = f + \phi$ is the sum of a continuously differentiable and a lower semicontinuous function, the sum rule

$$\partial\psi(x) = f'(x) + \partial\phi(x) \quad \forall x \in \operatorname{dom}(\phi) \tag{2.1}$$

holds for the limiting subdifferential, cf. [26]. In particular, we therefore have a stationary point x^* of problem (1.1) if and only if

$$0 \in f'(x^*) + \partial\phi(x^*)$$

holds.

We finally introduce the Kurdyka–Łojasiewicz property which will play a central role for our subsequent rate-of-convergence result. The following definition is a generalization of the classical one for nonsmooth functions, as introduced in [3, 7, 8]. Note that this KL property plays a central role in the local convergence analysis of several algorithms for the solution of nonsmooth minimization problems, see [2, 4, 9, 10, 11, 19, 27, 28] for a couple of examples.

Definition 2.1. *Let $g : \mathbb{X} \to \overline{\mathbb{R}}$ be lower semicontinuous. We say that g satisfies the* Kurdyka–Łojasiewicz (KL) property *at $x^* \in \{x \in \mathbb{X} \,|\, \partial g(x) \neq \emptyset\}$ if there exists a constant $\eta > 0$, a neighborhood $U \subset \mathbb{X}$ 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*

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

holds for all $x \in U \cap \{x \in \mathbb{X} \,|\, g(x^*) < g(x) < g(x^*) + \eta\}$.

Note that there exist whole classes of functions where the KL property is known to hold with the corresponding desingularization function $\chi(t) := ct^\kappa$ for some $\kappa \in (0, 1]$ and a constant $c > 0$, where κ is called the *KL exponent*, see [8, 22]. This KL property looks somewhat artificial at a first glance, but turns out to be a very useful and general tool for proving global convergence of the entire sequence as well as local rate-of-convergence results. Moreover, for g being a convex function, the KL property is known to be equivalent to several other concepts like a quadratic growth condition, a proximal error bound, or the metric subregularity condition, see [32] for more details and corresponding references.

We finally restate a technical result from [1, Lemma 1] which will be used in order to simplify our final rate-of-convergence result.

Lemma 2.2. *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

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

3 Nonmonotone Proximal Gradient Method

This section gives a precise presentation of our nonmonotone proximal gradient method and provides some of its basic properties. We first state the assumptions that we suppose to hold throughout our theoretical investigation of the method.

Assumption 3.1. *Assume:*

(a) ψ is bounded from below on $\operatorname{dom}(\phi)$,

(b) ϕ is bounded from below by an affine function,

(c) f' is locally Lipschitz continuous.

Note that the first condition is very reasonable since otherwise the given composite optimization problem (1.1) would be unbounded from below. The second condition essentially guarantees that the proximal gradient subproblems (1.2) have a solution (not necessarily a unique one) since this implies that eventually the quadratic term dominates the behaviour of the corresponding function. Finally, we stress that the local Lipschitz condition is equivalent to f' being globally Lipschitz continuous on *compact* sets, and that this local Lipschitz property is a much weaker condition than the usual global Lipschitz assumption, e.g., the exponential function, the natural logarithm, and all polynomials of degree higher than two are locally Lipschitz, but not globally Lipschitz on their respective domains.

Before we present our algorithm, recall that the basic iteration of the proximal gradient method is given by (1.2) or, equivalently, by (1.3). Observe that the next iterate x^{k+1} depends on the choice of the stepsize parameter γ_k, though, for simplicity, this is not made explicit in our notation. A central observation is that, for sufficiently small $\gamma_k > 0$, the corresponding solution x^{k+1} satisfies the property

$$\psi(x^{k+1}) \leq \psi(x^k) - \omega \frac{1}{2\gamma_k} \|x^{k+1} - x^k\|^2 \tag{3.1}$$

for any given parameter $\omega \in (0, 1)$, provided that x^k is not already an M-stationary point of ψ, see, e.g., [14, 21] for a formal proof. This implies that the sequence $\{\psi(x^k)\}$ is monotonically decreasing. Now, suppose that we have a reference value $\mathcal{R}_k \geq \psi(x^k)$. Then, of course, any x^{k+1} satisfying the monotone criterion (3.1) also satisfies the inequality

$$\psi(x^{k+1}) \leq \mathcal{R}_k - \omega \frac{1}{2\gamma_k} \|x^{k+1} - x^k\|^2, \tag{3.2}$$

but this condition might already be satisfied for a larger choice of γ_k, hence, resulting in a larger step, which is the reason why nonmonotone methods often 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. [18], where $\mathcal{R}_k := \max\{\psi(x^l) \mid l = k, k-1, \ldots, k-l_k\}$ for some given $l_k \in \mathbb{N}$. 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 our Algorithm 3.2, however, we use the technique introduced by Zhang and Hager [33], where $\mathcal{R}_{k+1}$ is computed as a convex combination of the previous reference value $\mathcal{R}_k$ and the new function value $\psi(x^{k+1})$. We therefore call this the *mean-rule*. The details are given in Algorithm 3.2. Note that, to be closer to the original version of this nonmonotonicity criterion, we replace the constant $\omega \in (0, 1)$ used in (3.1) and (3.2) by a sequence of some values $1 - \alpha_k \in (0, 1)$.

Algorithm 3.2 (Nonmonotone Proximal Gradient Method).

Require: *$x^0 \in \mathrm{dom}(\phi)$, $\varepsilon > 0$, $0 < \gamma_{\min} \leq \gamma_{\max} < \infty$, $0 < \alpha_{\min} \leq \alpha_{\max} < 1$, $0 < \beta_{min} \leq \beta_{\max} < 1$, $p_{\min} \in (0, 1]$.*

1: *set $\mathcal{R}_0 := \psi(x_0)$.*
2: ***for*** $k = 0, 1, 2, \ldots$ ***do***
3: *choose $\gamma_k \in [\gamma_{\min}, \gamma_{\max}]$.*
4: *compute $x^{k+1} \in Prox_{\gamma_k \phi}(x^k - \gamma_k f'(x^k))$.*
5: ***if*** $\|\frac{1}{\gamma_k}(x^{k+1} - x^k) - f'(x^{k+1}) + f'(x^k)\| \leq \varepsilon$ ***then***
6: ***return*** $x^* := x^{k+1}$
7: ***end if***
8: *choose $\alpha_k \in [\alpha_{\min}, \alpha_{\max}]$ and $\beta_k \in [\beta_{\min}, \beta_{\max}]$*
9: ***if*** $\psi(x^{k+1}) > \mathcal{R}_k - \frac{1-\alpha_k}{2\gamma_k}\|x^{k+1} - x^k\|^2$ ***then***
10: *set $\gamma_k := \beta_k \gamma_k$ and go back to step 4.*
11: ***end if***
12: *choose $p_{k+1} \in [p_{\min}, 1]$ and set $\mathcal{R}_{k+1} := (1 - p_{k+1})\mathcal{R}_k + p_{k+1}\psi(x^{k+1})$.*
13: ***end for***

We note that our convergence theory implicitly assumes that Algorithm 3.2 generates an infinite sequence. In particular, we assume that the practical termination criterion included into line 5 of Algorithm 3.2 never holds. This test can be interpreted as a measure for x^{k+1} being close to an M-stationary point, see [21] for further details. Note that this also implies that $x^{k+1} \neq x^k$

holds for all k since otherwise the current iterate x^k is both an M-stationary point of ψ and also a point satisfying the termination criterion from line 5 of Algorithm 3.2.

The following properties can be verified for Algorithm 3.2, see [14, 21] for the details.

Lemma 3.3. *Let Assumption 3.1 be satisfied. Then the following statements hold for each sequence $\{x^k\}$ generated by Algorithm 3.2:*

(a) $\mathcal{R}_k \geq \psi(x^k)$ for all $k \in \mathbb{N}$.

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

(c) The inner loop for the stepsize calculation γ_k is finite for each k (provided that x^k is not already an M-stationary point).

(d) $\|x^{k+1} - x^k\| \to 0$ for $k \to \infty$.

The formal proof of this result can be found in [14, 21]. We note, however, that statements (a), (b), and (c) are relatively simple to verify or standard observations. Part (d) is then a direct consequence of these statements, in fact, from the computation of $\mathcal{R}_k$, the acceptance criterion for the stepsize γ_k, and the updates of the corresponding parameters, we obtain

$$\begin{aligned} \mathcal{R}_{k+1} &= (1 - p_{k+1})\mathcal{R}_k + p_{k+1}\psi(x^{k+1}) \\ &\leq (1 - p_{k+1})\mathcal{R}_k + p_{k+1}\Big(\mathcal{R}_k - \frac{1-\alpha_k}{2\gamma_k}\|x^{k+1} - x^k\|^2\Big) \\ &\leq \mathcal{R}_k - p_{\min}\frac{1-\alpha_{\max}}{2\gamma_{\max}}\|x^{k+1} - x^k\|^2. \end{aligned} \tag{3.3}$$

Rearranging these terms yields

$$\mathcal{R}_{k+1} - \mathcal{R}_k \leq -p_{\min}\frac{1-\alpha_{\max}}{2\gamma_{\max}}\|x^{k+1} - x^k\|^2 \leq 0 \tag{3.4}$$

for all $k \in \mathbb{N}$. Now, since ψ is bounded from below by Assumption 3.1, we obtain from Lemma 3.3 (a) that the sequence $\{\mathcal{R}_k\}$ is also bounded from below. In view of Lemma 3.3 (b), it follows that this sequence is convergent. Consequently, the left-hand side from (3.4) converges to zero. Hence, statement (d) of Lemma 3.3 is a consequence of (3.4) and the sandwich theorem.

We next summarize the main global convergence properties of Algorithm 3.2, the corresponding proofs can be found in [14].

Theorem 3.4. *Let Assumption 3.1 be satisfied and x^* be an accumulation point of a sequence $\{x^k\}$ generated by Algorithm 3.2. Then the following statements hold:*

(a) x^* is an M-stationary point of ψ.

(b) The sequence $\{\psi(x^k)\}$ converges to $\psi(x^*)$.

(c) The sequence $\{\mathcal{R}_k\}$ converges monotonically to $\psi(x^*)$.

Note that the central statement of Theorem 3.4 is assertion (a), the corresponding technical proof in [14] follows the ideas from [21]. The other two statements are easier to verify. In fact, statement (c) is mainly a consequence of the observation from Lemma 3.3 (b) that the sequence $\{\mathcal{R}_k\}$ is monotonically decreasing, and the (usually nonmonotone) convergence of the sequence $\{\psi(x^k)\}$ can then be derived from this observation together with the update of $\mathcal{R}_{k+1}$ in Algorithm 3.2.

4 Convergence Theory

Theorem 3.4 shows very satisfactory global (subsequential) convergence properties under fairly mild assumptions regarding the two functions f and ϕ. The aim of this section (and of this paper) is to show that, given some accumulation point of a sequence generated by Algorithm 3.2 such that the KL property holds at this point, then the entire sequence converges to this limit point and, in addition, has very favourable rate-of-convergence properties. In other words, we obtain essentially the same convergence properties for our nonmonotone proximal gradient method as those known for its monotone counterpart from [19]. Once again, we stress that this result holds without any convexity of f or ϕ, without a global Lipschitz assumption regarding f', without any explicit knowledge of a local Lipschitz constant, without an a priori assumption that the sequence $\{x^k\}$ remains bounded, and with ϕ being merely lower semicontinuous.

The corresponding convergence theory requires some technical results which are inspired by the corresponding ones in the papers [14, 19, 30] and modified in a suitable way to deal with the above general setting.

To this end, we begin with the following result which is the nonmonotone counterpart of [19, Lemma 4.1], see also [14, Corollary 4.5].

Lemma 4.1. *Let Assumption 3.1 hold, $\{x^k\}$ be any sequence generated by Algorithm 3.2, and let x^* be an accumulation point of $\{x^k\}$. Then, for each $\rho > 0$, there exists a constant $\underline{\gamma}_\rho > 0$ such that $\gamma_k \geq \underline{\gamma}_\rho$ holds for all $k \in \mathbb{N}$ satisfying $x^k \in B_\rho(x^*)$.*

Proof. Recall that the stepsizes $\{\gamma_k\}$ are well-defined in view of Lemma 3.3 (c). Let $\rho > 0$ be fixed. Since f' is locally Lipschitz continuous, it is globally Lipschitz continuous on $B_{2\rho}(x^*)$ with Lipschitz constant denoted by $L_{2\rho}$. Since x^* is an accumulation point, there are infinitely many iterates belonging to $B_\rho(x^*)$.

Assume, by contradiction, that there exists a subsequence $\{x^k\}_K$ with $x^k \in B_\rho(x^*)$ for all $k \in K$ and such that $\{\gamma_k\}_K$ is not bounded away from 0. Without loss of generality, we may assume $\gamma_k \to_K 0$ and $x^k \to_K \bar{x}$ for some $\bar{x} \in B_\rho(x^*)$, and that the acceptance criterion for the computation of the stepsize γ_k is violated in the previous iteration of the inner loop. For the corresponding trial stepsize $\hat{\gamma}_k := \gamma_k/\beta_k$, we then have $\gamma_k/\beta_{\max} \leq \hat{\gamma}_k \leq \gamma_k/\beta_{\min}$. This shows that we also have $\hat{\gamma}_k \to_K 0$.

Then the corresponding trial vector $\hat{x}^{k+1}$, i.e., the solution of the subproblem

$$\min_x \; f(x^k) + \langle f'(x^k), x - x^k \rangle + \frac{1}{2\hat{\gamma}_k}\|x - x^k\|^2 + \phi(x), \quad x \in \mathbb{X}, \tag{4.1}$$

does not satisfy the stepsize condition with associated parameter $\hat{\alpha}_k \in [\alpha_{\min}, \alpha_{\max}]$, i.e.,

$$\psi(\hat{x}^{k+1}) > \mathcal{R}_k - \frac{1-\hat{\alpha}_k}{2\hat{\gamma}_k}\|\hat{x}^{k+1} - x^k\|^2 \geq \mathcal{R}_k - \frac{1-\alpha_{\min}}{2\hat{\gamma}_k}\|\hat{x}^{k+1} - x^k\|^2. \tag{4.2}$$

Note that (4.2) in combination with $\mathcal{R}_k \geq \psi(x^k)$, cf. Lemma 3.3 (a), implies, in particular, that we have $x^k \neq \hat{x}^{k+1}$. Moreover, since $\hat{x}^{k+1}$ is a solution of (4.1), we have

$$\langle f'(x^k), \hat{x}^{k+1} - x^k \rangle + \frac{1}{2\hat{\gamma}_k}\|\hat{x}^{k+1} - x^k\|^2 + \phi(\hat{x}^{k+1}) \leq \phi(x^k). \tag{4.3}$$

Using the Cauchy-Schwarz inequality and the fact that $\psi(x^k) \leq \mathcal{R}_k \leq \mathcal{R}_0$ for all k by Lemma 3.3 (a), (b), we obtain

$$\begin{aligned}\frac{1}{2\hat{\gamma}_k}\|\hat{x}^{k+1} - x^k\|^2 &\leq \|f'(x^k)\|\|\hat{x}^{k+1} - x^k\| + \phi(x^k) - \phi(\hat{x}^{k+1}) \\ &= \|f'(x^k)\|\|\hat{x}^{k+1} - x^k\| + \psi(x^k) - f(x^k) - \phi(\hat{x}^{k+1}) \\ &\leq \|f'(x^k)\|\|\hat{x}^{k+1} - x^k\| + \mathcal{R}_0 - f(x^k) - \phi(\hat{x}^{k+1}).\end{aligned}$$

Using the continuous differentiability of f and the boundedness condition from Assumption 3.1 for ϕ, we claim that the above inequality implies $\hat{x}^{k+1} - x^k \to_K 0$: Assume, by contradiction, that $\{\|\hat{x}^{k+1} - x^k\|\}_{k\in K}$ would be unbounded, then the left-hand side would grow more rapidly than the right-hand side. If $\{\|\hat{x}^{k+1} - x^k\|\}_{k\in K}$ remains bounded but (at least on a subsequence) staying away from zero, then the right-hand side is bounded but the left-hand side is unbounded as we have $\hat{\gamma}_k \to_K 0$. Hence, we have $\hat{x}^{k+1} - x^k \to_K 0$.

Taking this into account and using the fact that $\bar{x} \in B_\rho(x^*)$, it follows that

$$x^k \in B_\rho(x^*) \subset B_{2\rho}(x^*) \quad \text{and} \quad \hat{x}^{k+1} \in B_{2\rho}(x^*) \quad \text{for sufficiently large } k \in K. \tag{4.4}$$

We will exploit this observation later.

Using the differential mean-value theorem, there exists a point ξ^k on the line segment between x^k and $\hat{x}^{k+1}$ such that

$$\begin{aligned}\psi(\hat{x}^{k+1}) - \psi(x^k) &= f(\hat{x}^{k+1}) + \phi(\hat{x}^{k+1}) - f(x^k) - \phi(x^k) \\ &= \langle f'(\xi^k), \hat{x}^{k+1} - x^k \rangle + \phi(\hat{x}^{k+1}) - \phi(x^k).\end{aligned}$$

Substituting $\phi(\hat{x}^{k+1}) - \phi(x^k)$ into (4.3) yields

$$\langle f'(x^k) - f'(\xi^k), \hat{x}^{k+1} - x^k \rangle + \frac{1}{2\hat{\gamma}_k}\|\hat{x}^{k+1} - x^k\|^2 + \psi(\hat{x}^{k+1}) - \psi(x^k) \le 0.$$

Now, we obtain

$$\begin{aligned}\frac{1}{2\hat{\gamma}_k}\|\hat{x}^{k+1} - x^k\|^2 &\le -\langle f'(x^k) - f'(\xi^k), \hat{x}^{k+1} - x^k \rangle + \psi(x^k) - \psi(\hat{x}^{k+1}) \\ &\le -\langle f'(x^k) - f'(\xi^k), \hat{x}^{k+1} - x^k \rangle + \psi(x^k) - \mathcal{R}_k + \frac{1-\alpha_{\min}}{2\hat{\gamma}_k}\|\hat{x}^{k+1} - x^k\|^2 \\ &\le \|f'(x^k) - f'(\xi^k)\|\|\hat{x}^{k+1} - x^k\| + \frac{1-\alpha_{\min}}{2\hat{\gamma}_k}\|\hat{x}^{k+1} - x^k\|^2,\end{aligned}$$

where the second inequality is due to (4.2) and the final estimate follows from the Cauchy-Schwarz inequality together with $\mathcal{R}_k \ge \psi(x^k)$, cf. Lemma 3.3 (a). As $\hat{x}^{k+1} \neq x^k$ in view of our previous discussion, the resulting expression can be simplified to

$$\frac{\alpha_{\min}}{2\hat{\gamma}_k}\|\hat{x}^{k+1} - x^k\| \le \|f'(x^k) - f'(\xi^k)\|.$$

Since ξ^k is on the line connecting x^k and $\hat{x}^{k+1}$, it follows from (4.4) that also $\xi^k \in B_{2\rho}(x^*)$ holds for all $k \in K$ sufficiently large. Thus, using the Lipschitz continuity of f' on $B_{2\rho}(x^*)$, we conclude

$$\frac{\alpha_{\min}}{2\hat{\gamma}_k}\|\hat{x}^{k+1} - x^k\| \le L_{2\rho}\|x^k - \xi^k\| \le L_{2\rho}\|x^k - \hat{x}^{k+1}\|.$$

As $\hat{x}^{k+1} \neq x^k$, we get $\hat{\gamma}_k \ge \frac{\alpha_{\min}}{2L_{2\rho}}$ for all $k \in K$ sufficiently large. This, in turn, implies that the corresponding subsequence $\{\gamma_k\}_K$ is also bounded from below by a positive constant, but this contradicts our assumption. Altogether, this completes the proof. □

For the remaining part, assume that our objective function ψ satisfies the KL property at a given accumulation point x^*. Let $\eta > 0$ be the corresponding constant and χ the associated desingularization function from Definition 2.1. Furthermore, we denote by $\{x^k\}_{k\in K}$ a subsequence converging to x^*.

The subsequent theory requires some further constants and indices which will be introduced here and which will be used throughout the remaining part of this section. To this end, we first note that, in view of Lemma 3.3 (d), there exists an index $\hat{k} \in \mathbb{N}$ such that

$$\sup_{k\ge\hat{k}} \|x^{k+1} - x^k\| \le \eta. \tag{4.5}$$

Define

$$\rho := \eta + \frac{1}{2} \quad \text{and} \quad C_\rho := B_\rho(x^*) \cap \{x \in \mathbb{X} \,|\, \psi(x) \le \mathcal{R}_0\}.$$

Let L_ρ be a global Lipschitz constant of f' on C_ρ. By Lemma 4.1, there exists a constant $\underline{\gamma}_\rho > 0$ such that

$$\gamma_k \ge \underline{\gamma}_\rho \quad \text{for all } k \text{ with } x^k \in C_\rho.$$

Following mostly [30], we also introduce the following notation:

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

Note that the index m is obviously uniquely defined since the left-hand side of the inequality eventually becomes larger than the constant on the right-hand side. Furthermore, note that the difference $l(k) - k = m - 1$ is a constant number for all $k \in \mathbb{N}$, in particular, this difference does not increase to infinity for $k \to \infty$. This simple observation plays some role in the subsequent convergence analysis since it guarantees that certain sums are always taken over a finite (fixed) number of terms only. Moreover, we note that $\Xi_{k-1} \ge 0$ holds for all $k \in \mathbb{N}$ by the monotonicity property of the sequence $\{\mathcal{R}_k\}$ from Lemma 3.3 (b). Finally, we also have $\Delta_{i,j} \ge 0$ for all $j \ge i$ by monotonicity of χ in combination with Lemma 3.3 (b) once again.

We further introduce the two index sets

$$K_1 := \{k \in \mathbb{N} \,|\, \psi(x^k) \le \mathcal{R}_{k+m}\},$$

and

$$K_2 := \{k \in \mathbb{N} \,|\, \psi(x^k) > \mathcal{R}_{k+m}\}$$

depending on the previously introduced number m. To simplify the notation, we define the constant

$$a := \frac{1 - \alpha_{\max}}{2\gamma_{\max}} > 0.$$

Using the update rule for $\mathcal{R}_k$ together with the acceptance criterion for the step size γ_k, we obtain

$$\mathcal{R}_k \le \mathcal{R}_{k-1} - \frac{1 - \alpha_{k-1}}{2\gamma_{k-1}} p_k \|x^k - x^{k-1}\|^2 \le \mathcal{R}_{k-1} - a p_{\min} \|x^k - x^{k-1}\|^2,$$

cf. (3.3). Therefore,

$$\sqrt{a p_{\min}} \|x^k - x^{k-1}\| \le \Xi_{k-1}. \tag{4.6}$$

The following results and proofs are motivated by the corresponding analysis in [30]. However, to avoid the a priori assumption that the sequence generated by Algorithm 3.2 is bounded, we need to modify the arguments to some extend, using ideas from [19, 21].

Lemma 4.2. *Define the constant* $\hat{c} := \frac{\sqrt{p_{\min}}}{2\sqrt{a}}\Big(\frac{1}{\underline{\gamma}_\rho} + L_\rho\Big)$. *Then there exists a sufficiently large index* $k_0 \in K$ *such that*

$$\alpha := \|x^{k_0 - 1} - x^*\| + \frac{4}{\sqrt{a p_{\min}}} \sum_{j=k_0}^{l(k_0)} \Xi_{j-1} + \frac{2\hat{c}}{\sqrt{a p_{\min}}} \sum_{j=k_0}^{l(k_0)} \chi\big(\mathcal{R}_j - \psi(x^*)\big) \tag{4.7}$$

satisfies $\alpha < \frac{1}{2}$ *and* $B_\alpha(x^*) \subset U$, *where* U *is the neighborhood of* x^* *from the KL property in Definition 2.1.*

Proof. First recall that $l(k) - k = m - 1$ is a fixed constant, hence, the number of terms within each of the summations is fixed, independent of k. Therefore, it is easy to see that each term on the right-hand side can be made arbitrarily small: For the first term, this follows from $x^k \to_K x^*$ together with the fact that $\|x^k - x^{k-1}\| \to 0$ in view of Lemma 3.3 (d). For the second term, recall that $\mathcal{R}_k \to \psi(x^*)$ monotonically by Theorem 3.4 (c). For the final term, note again that $\mathcal{R}_k \to \psi(x^*)$ monotonically and that, by definition, the desingularization function χ is continuous at the origin. $\square$

Note that, in principle, Lemma 4.2 holds for an arbitrary constant $\hat{c} > 0$. However, the index k_0 depends on this constant $\hat{c}$. Since the previous result will later be applied to the particular choice of $\hat{c}$ from Lemma 4.2 with the corresponding index k_0, the previous result is formulated for this particular value of $\hat{c}$.

The following result is the nonmonotone counterpart of [19, Lemma 4.4]. Regarding its assumption, we recall from Theorem 3.4 that the sequence $\{\mathcal{R}_k\}$ converges monotonically to the function value $\psi(x^*)$ at the accumulation point x^*, hence, we eventually have $\mathcal{R}_k < \psi(x^*) + \eta$, where $\eta > 0$ denotes the constant from Definition 2.1.

Lemma 4.3. *Under the conditions specified above, we have*

$$\operatorname{dist}\big(0, \partial\psi(x^{k+1})\big) \le \Big(\frac{1}{\underline{\gamma}_\rho} + L_\rho\Big)\|x^{k+1} - x^k\| \tag{4.8}$$

for all sufficiently large $k \in \mathbb{N}$ such that $x^k \in B_\alpha(x^*)$ holds.

Proof. For all k, since x^{k+1} solves the proximal gradient subproblem (1.2), we obtain

$$0 \in f'(x^k) + \frac{1}{\gamma_k}(x^{k+1} - x^k) + \partial\phi(x^{k+1}) \tag{4.9}$$

from the M-stationary condition together with the sum rule (2.1) for the Mordukhovich subdifferential. This implies

$$\frac{1}{\gamma_k}(x^k - x^{k+1}) + f'(x^{k+1}) - f'(x^k) \in f'(x^{k+1}) + \partial\phi(x^{k+1}) = \partial\psi(x^{k+1}) \tag{4.10}$$

using once again the sum rule from (2.1). Take $k \in \mathbb{N}$ sufficiently large such that $x^k \in B_\alpha(x^*)$ and $k \ge \hat{k}$, where $\hat{k}$ is the index from (4.5) and α is the constant defined in Lemma 4.2. Since $\alpha \le \rho$, we have $\gamma_k \ge \underline{\gamma}_\rho$ by Lemma 4.1. Now, the estimate

$$\|x^{k+1} - x^*\| \le \|x^{k+1} - x^k\| + \|x^k - x^*\| \le \eta + \alpha \le \rho$$

shows that $x^k, x^{k+1} \in C_\rho$. Therefore, we get

$$\|f'(x^{k+1}) - f'(x^k)\| \le L_\rho\|x^{k+1} - x^k\|.$$

Using the bound on γ_k and (4.10) gives

$$\begin{aligned}
\operatorname{dist}\big(0, \partial\psi(x^{k+1})\big) &\le \Big\|\frac{1}{\gamma_k}(x^k - x^{k+1}) + f'(x^{k+1}) - f'(x^k)\Big\| \\
&\le \frac{1}{\gamma_k}\|x^k - x^{k+1}\| + L_\rho\|x^{k+1} - x^k\| \\
&\le \Big(\frac{1}{\underline{\gamma}_\rho} + L_\rho\Big)\|x^{k+1} - x^k\|
\end{aligned}$$

for all k with $k \ge \hat{k}$ and $x^k \in B_\alpha(x^*)$. $\square$

We now present our final technical result, motivated by [30].

Lemma 4.4. *For all sufficiently large $k \in \mathbb{N}$ with $\mathcal{R}_k < \psi(x^*) + \eta$ and $x^{k-1} \in B_\alpha(x^*)$, the following inequality holds:*

$$\frac{1-\sqrt{1-p_{\min}}}{\sqrt{m}} \sum_{i=k}^{l(k)} \Xi_i \leq \left(1/2 + \sqrt{1-p_{\min}}\right) \Xi_{k-1} + \hat{c}\Delta_{k,k+m}, \tag{4.11}$$

where $\hat{c}$ denotes the constant from Lemma 4.2.

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

$$\frac{1-\sqrt{1-p_{\min}}}{\sqrt{m}} \sum_{i=k}^{l(k)} \Xi_i \leq \left(1-\sqrt{1-p_{\min}}\right)\sqrt{\mathcal{R}_k - \mathcal{R}_{k+m}}. \tag{4.12}$$

We now distinguish two cases.

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

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

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

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

The statement therefore follows from (4.12).

Case 2: $k \in K_2$. We then have $\psi(x^*) \leq \mathcal{R}_{k+m} < \psi(x^k) \leq \mathcal{R}_k < \psi(x^*) + \eta$ by assumption. Using the KL property of χ, we get

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

As x^{k-1} was assumed to be in $B_\alpha(x^*)$, by application of Lemma 4.3, we have

$$\chi'\big(\psi(x^k) - \psi(x^*)\big) \geq \frac{1}{\left(\frac{1}{\underline{\gamma}_\rho} + L_\rho\right)\|x^k - x^{k-1}\|} \tag{4.13}$$

(recall that $x^k \neq x^{k-1}$ since Algorithm 3.2 is assumed to generate an infinite sequence). Using the properties of χ, we now obtain

$$\begin{aligned}
\Delta_{k,k+m} &= \chi\big(\mathcal{R}_k - \psi(x^*)\big) - \chi\big(\mathcal{R}_{k+m} - \psi(x^*)\big) \\
&\geq \chi\big(\psi(x^k) - \psi(x^*)\big) - \chi\big(\mathcal{R}_{k+m} - \psi(x^*)\big) \\
&\geq \chi'\big(\psi(x^k) - \psi(x^*)\big)\big(\psi(x^k) - \mathcal{R}_{k+m}\big) \\
&\geq \frac{\psi(x^k) - \mathcal{R}_{k+m}}{\left(\frac{1}{\underline{\gamma}_\rho} + L_\rho\right)\|x^k - x^{k-1}\|},
\end{aligned}$$

where the first inequality results from the monotonicity of χ, the next one exploits the concavity of χ, and the final estimate exploits (4.13) together with the fact that $\psi(x^k) - \mathcal{R}_{k+m} > 0$ in the case under consideration. Thus, with (4.6), we get

$$\psi(x^k) - \mathcal{R}_{k+m} \leq \frac{2\hat{c}}{p_{\min}}\Xi_{k-1}\Delta_{k,k+m}$$

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 $\psi(x^k) \leq \mathcal{R}_{k-1}$ by the acceptance criterion for our stepsize computation as well as $p_k \geq p_{\min}$, we have

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

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

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

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

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

so we have

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

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

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

In view of (4.12), this completes the proof. □

The following result shows global convergence of the entire sequence $\{x^k\}$ generated by the nonmonotone proximal gradient method to one of its accumulation points x^*, given that ψ satisfies the KL property at this point. The proof follows the technique of the global convergence result in [30]. However, by using our previous results, we neither assume the a priori boundedness of the iterates $\{x^k\}$ nor do we require f to satisfy a global Lipschitz condition.

Theorem 4.5. *Let Assumption 3.1 hold, let $\{x^k\}_K$ be a subsequence converging to some limit point x^*, and suppose that the KL property for ψ holds at x^*. Then the entire sequence $\{x^k\}$ converges to x^*.*

Proof. Let k_0 be the index from the definition of α, cf. Lemma 4.2. Without loss of generality, we may assume $k_0 \geq \hat{k}$, where $\hat{k}$ is the index from (4.5) and that $\mathcal{R}_{k_0} < \psi(x^*) + \eta$. We now claim that the following statements hold:

(a) for all $k \geq k_0 - 1$: $x^k \in B_\alpha(x^*)$, and

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

$$\big(1 - \sqrt{1 - p_{\min}}\big)\sqrt{m}\sum_{j=l(k_0)}^{k}\Xi_j \leq \Big(\frac{1}{2} + \sqrt{1 - p_{\min}}\Big)\sum_{j=k_0}^{k}\Xi_{j-1} + \hat{c}\sum_{j=k_0}^{l(k_0)}\chi\big(\mathcal{R}_j - \psi(x^*)\big), \quad (4.14)$$

where $\hat{c}$ denotes the constant from Lemma 4.2. We verify these two statements jointly by induction over k. For all $k \in \{k_0 - 1, \ldots, l(k_0)\}$, we obtain from (4.6) and the definition of the constant α in Lemma 4.2 that

$$\begin{aligned}\|x^k - x^*\| &\le \|x^{k_0-1} - x^*\| + \sum_{j=k_0}^{k} \|x^j - x^{j-1}\| \\ &\le \|x^{k_0-1} - x^*\| + \sum_{j=k_0}^{l(k_0)} \|x^j - x^{j-1}\| \\ &\le \|x^{k_0-1} - x^*\| + \frac{1}{\sqrt{a p_{\min}}} \sum_{j=k_0}^{l(k_0)} \Xi_{j-1} \le \alpha,\end{aligned}$$

which shows the first statement for $k = k_0 - 1, \ldots, l(k_0)$. Now, we can apply Lemma 4.4 for indices $k = k_0, \ldots, l(k_0)$ and obtain

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

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 (4.11) in Lemma 4.4 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 (4.14), we get

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

This implies

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

Noting that, by definition of m, it holds that $\big(1-\sqrt{1-p_{\min}}\big)\sqrt{m}-\big(1/2+\sqrt{1-p_{\min}}\big)\geq 1/2$ and that we obviously have $1/2+\sqrt{1-p_{\min}}\leq 3/2$, we get

$$\sum_{j=l(k_0)}^{k}\Xi_j\leq 3\sum_{j=k_0}^{l(k_0)}\Xi_{j-1}+2\hat{c}\sum_{j=k_0}^{l(k_0)}\chi\big(\mathcal{R}_j-\psi(x^*)\big).$$

This implies

$$\sum_{j=k_0-1}^{k}\Xi_j=\sum_{j=k_0}^{l(k_0)}\Xi_{j-1}+\sum_{j=l(k_0)}^{k}\Xi_j\leq 4\sum_{j=k_0}^{l(k_0)}\Xi_{j-1}+2\hat{c}\sum_{j=k_0}^{l(k_0)}\chi\big(\mathcal{R}_j-\psi(x^*)\big),$$

hence, we obtain from (4.6) that

$$\begin{aligned}\|x^{k+1}-x^*\|&\leq\|x^{k_0-1}-x^*\|+\sum_{j=k_0-1}^{k}\|x^{j+1}-x^j\|\leq\|x^{k_0-1}-x^*\|+\frac{1}{\sqrt{ap_{\min}}}\sum_{j=k_0-1}^{k}\Xi_j\\&\leq\|x^{k_0-1}-x^*\|+\frac{4}{\sqrt{ap_{\min}}}\sum_{j=k_0}^{l(k_0)}\Xi_{j-1}+\frac{2\hat{c}}{\sqrt{ap_{\min}}}\sum_{j=k_0}^{l(k_0)}\chi\big(\mathcal{R}_j-\psi(x^*)\big).\end{aligned}\tag{4.15}$$

The expression on the right-hand side is precisely the constant α from Lemma 4.2. Thus, the first statement holds for $k+1$.

We next verify the second part for $k+1$. Since we know that $x^j\in B_\alpha(x^*)$ is true for all $j\in\{k_0-1,\dots,k+1\}$ by our induction hypothesis, we again apply Lemma 4.4 and sum over (4.11), now from k_0 to $k+1$. This yields

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

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 (4.11) from Lemma 4.4, and the final estimate uses a telescoping sum argument where we omit some nonpositive summands. This completes our induction step.

Hence, it follows that $x^k\in B_\alpha(x^*)$ 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 (4.15) 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 $\{x^k\}$. □

Next, we present a rate-of-convergence result for the case where the desingularization function is given by $\chi(t)=ct^\kappa$ for some $c>0$ and $\kappa\in(0,1)$. We recover the same rate-of-convergence as for the monotone proximal gradient method. The proof is based on [19] with suitable adaptations for the nonmonotone case.

Theorem 4.6. *Let Assumption 3.1 hold, and suppose that $\{x^k\}$ converges on some subsequence $\{x^k\}_K$ to a limit point x^* such that ψ has the KL property at x^*. Then the entire sequence $\{x^k\}$ converges to x^*. Further, if the corresponding desingularization function is given by $\chi(t)=ct^\kappa$ (for some $c>0$ and $\kappa\in(0,1)$), then the following statements hold:*

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

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

$$\mathcal{R}_k - \psi(x^*) \le \eta_1 k^{-\frac{1}{1-2\kappa}}, \tag{4.16}$$

$$\|x^k - x^*\| \le \eta_2 k^{-\frac{\kappa}{1-2\kappa}}. \tag{4.17}$$

Proof. Taking Theorem 4.5 into account, we only need to verify the statements (a) and (b). As a first step, let us prove the results for the $\{\mathcal{R}_k\}$. We first claim that for $\kappa \in (0, 1)$ and with

$$\sigma := \frac{1 - \alpha_{\max}}{2\gamma_{\max}} \frac{p_{\min}}{c^2\kappa^2\Big(\frac{1}{\underline{\gamma}_\rho} + L_\rho\Big)^2},$$

it holds that

$$\psi(x^{k+1}) - \psi(x^*) \le \Big(\frac{1}{\sigma}\Big)^{\frac{1}{2(1-\kappa)}} \big(\mathcal{R}_k - \mathcal{R}_{k+1}\big)^{\frac{1}{2(1-\kappa)}} \tag{4.18}$$

for all $k \in \mathbb{N}$ sufficiently large. In fact, if $\psi(x^{k+1}) \le \psi(x^*)$ holds, then the left-hand side of (4.18) is nonpositive, hence, the claim holds trivially, Thus, it remains to consider the case $\psi(x^{k+1}) > \psi(x^*)$. In view of our previous results, we may assume that k is large enough such that $x^k \in B_\alpha(x^*)$ and $\psi(x^*) < \psi(x^{k+1}) < \psi(x^*) + \eta$ hold.

As ψ satisfies the KL property at x^* with $\chi(t) = ct^\kappa$, we have

$$\begin{aligned} 1 &\le \chi'\big(\psi(x^{k+1}) - \psi(x^*)\big)\operatorname{dist}\big(0, \partial\psi(x^{k+1})\big) \\ &= c\kappa\big(\psi(x^{k+1}) - \psi(x^*)\big)^{\kappa-1}\operatorname{dist}\big(0, \partial\psi(x^{k+1})\big). \end{aligned}$$

By Lemma 4.3, this yields

$$1 \le c\kappa\Big(\frac{1}{\underline{\gamma}_\rho} + L_\rho\Big)\big(\psi(x^{k+1}) - \psi(x^*)\big)^{\kappa-1}\|x^{k+1} - x^k\|,$$

which gives the inequality

$$\|x^{k+1} - x^k\| \ge \frac{1}{c\kappa\Big(\frac{1}{\underline{\gamma}_\rho} + L_\rho\Big)}\big(\psi(x^{k+1}) - \psi(x^*)\big)^{1-\kappa}. \tag{4.19}$$

By (3.3), we also have

$$\mathcal{R}_{k+1} - \mathcal{R}_k \le -\frac{1 - \alpha_{\max}}{2\gamma_{\max}} p_{\min}\|x^{k+1} - x^k\|^2. \tag{4.20}$$

Combination of (4.19) and (4.20) yields

$$\begin{aligned} \mathcal{R}_{k+1} - \mathcal{R}_k &\le -\frac{1 - \alpha_{\max}}{2\gamma_{\max}} p_{\min}\|x^{k+1} - x^k\|^2 \\ &\le -\frac{1 - \alpha_{\max}}{2\gamma_{\max}} p_{\min} \frac{1}{c^2\kappa^2\Big(\frac{1}{\underline{\gamma}_\rho} + L_\rho\Big)^2}\big(\psi(x^{k+1}) - \psi(x^*)\big)^{2(1-\kappa)} \\ &= -\sigma\big(\psi(x^{k+1}) - \psi(x^*)\big)^{2(1-\kappa)}. \end{aligned}$$

Rearranging these terms shows that the claim (4.18) holds.

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

$$\mathcal{R}_{k+1} = (1 - p_{k+1})\mathcal{R}_k + p_{k+1}\psi(x^{k+1}) \leq (1 - p_{\min})\mathcal{R}_k + p_{\min}\psi(x^{k+1}). \tag{4.21}$$

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

$$\begin{aligned} s_{k+1} &\leq (1 - p_{\min})s_k + p_{\min}(\psi(x^{k+1}) - \psi(x^*)) \\ &\leq (1 - p_{\min})s_k + p_{\min}\left(\frac{1}{\sigma}\right)^{\frac{1}{2(1-\kappa)}} \left(s_k - s_{k+1}\right)^{\frac{1}{2(1-\kappa)}}, \end{aligned}$$

where the first inequality follows from (4.21) and the second one results from (4.18). This implies

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

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

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

where

$$\beta := \left(\frac{1}{p_{\min}} + \left(\frac{1}{\sigma}\right)^{\frac{1}{2(1-\kappa)}}\right)^{\max\{1,2(1-\kappa)\}} > 0$$

is a constant.

We are now in the setting of Lemma 2.2 and immediately obtain the corresponding rate-of-convergence results for the sequence $\{\mathcal{R}_k\}$ as $\kappa \in (0, 1/2)$ implies that $2(1 - \kappa) > 1$ and $\kappa \in [1/2, 1)$ implies that $2(1 - \kappa) \in (0, 1]$.

Let us now verify the statements for the sequence $\{x^k\}$. In view of Theorem 4.5, the equations in the proof of that result remain valid if $k_0 - 1$ is replaced by some k sufficiently large. Note that $\Xi_{j-1} \leq \sqrt{\mathcal{R}_{j-1} - \psi(x^*)} = \sqrt{s_{j-1}}$. Taking the monotonicity of the function χ and the sequences $\{\mathcal{R}_k\}$ and thus $\{s_k\}$ into account, it follows from (4.15) that, for $l > k$:

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

for all k sufficiently large, where

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

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

We finally consider a generalized projected gradient method as a simple application of our theory.

Example 4.7. (Generalized Nonmonotone Projected Gradient Method)
Consider the constrained optimization problem

$$\min\ f(x) \quad \text{subject to} \quad x \in S$$

for some given set $S \subseteq \mathbb{X}$ which is assumed to be nonempty and closed (not necessarily convex). This problem can be reformulated as the unconstrained composite optimization problem

$$\min\ f(x) + \phi(x), \qquad x \in \mathbb{X},$$

where $\phi(x) := \delta_S(x)$ denotes the indicator function of S, i.e., $\delta_S(x) = 0$ for $x \in S$ and $\delta_S(x) = +\infty$ otherwise. Since S is nonempty and closed, this indicator function is proper and lower semicontinuous. Moreover, the proximal subproblem (1.3) reduces to compute a projection of a given point onto the set S. Recall that this projection always exists, albeit it is not necessarily unique since S is not assumed to be convex. Consequently, if f satisfies Assumption 3.1 (c), and in particular if f further has the respective KL property, our convergence results apply to this generalized projected gradient method. Note that this may be viewed as a generalization of the famous spectral gradient method from [6] where the feasible set S is assumed to be convex (and the globalization uses the max-type nonmonotone stepsize rule, see also the extension in [20]).

5 Final Remarks

The current paper presents a global and rate-of-convergence result for nonmonotone proximal gradient methods applied to composite optimization problems under fairly mild assumptions. To the best of our knowledge, this is the first time that results of this kind are shown for a nonmonotone method without a global Lipschitz assumption or the a priori knowledge that the iterates generated by the given method are bounded. Though the technique of proof is quite technical and it is currently not clear whether these results can be extended to other classes of first-order methods, we plan to have a closer look at this topic as part of our future research.

Acknowledgement. The authors thank Xiaoxi Jia for her comments on a previous version of the manuscript which helped to improve Theorem 4.6.

References

[1] 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.

[2] 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.

[3] 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.

[4] H. Attouch, J. Bolte, and B. F. Svaiter. "Convergence of descent methods for semi-algebraic and tame problems, proximal algorithms, forward-backward splitting, and regularized Gauss–Seidel methods". In: *Mathematical Programming* 137 (2013), pp. 91 –129. DOI: 10.1007/s10107-011-0484-9.

[5] A. Beck. *First-Order Methods in Optimization.* SIAM, 2017. DOI: 10.1137/1.9781611974997.

[6] 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.

[7] 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.

[8] 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.

[9] J. Bolte, S. Sabach, and M. Teboulle. “Proximal alternating linearized minimization for nonconvex and nonsmooth problems”. In: *Mathematical Programming* 146 (2014), pp. 459 –494. DOI: 10.1007/s10107-013-0701-9.

[10] R. I. Boţ and E. R. Csetnek. “An inertial Tseng’s type proximal algorithm for nonsmooth and nonconvex optimization problems”. In: *Journal of Optimization Theory and Applications* 171.2 (2016), pp. 600–616. DOI: 10.1007/s10957-015-0730-z.

[11] R. I. Boţ, E. R. Csetnek, and S. C. László. “An inertial forward–backward algorithm for the minimization of the sum of two nonconvex functions”. In: *EURO Journal on Computational Optimization* 4.1 (2016), pp. 3–25. DOI: 10.1007/s13675-015-0045-8.

[12] A. M. Bruckstein, D. L. Donoho, and M. Elad. “From sparse solutions of systems of equations to sparse modeling of signals and images”. In: *SIAM Review* 51.1 (2009), pp. 34–81. DOI: 10.1137/060657704.

[13] R. Chartrand. “Exact reconstruction of sparse signals via nonconvex minimization”. In: *IEEE Signal Processing Letters* 14.10 (2007), pp. 707–710. DOI: 10.1109/LSP.2007.898300.

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

[15] A. De Marchi, X. Jia, C. Kanzow, and P. Mehlitz. “Constrained composite optimization and augmented Lagrangian methods”. In: *Mathematical Programming* 201.1 (2023), pp. 863–896.

[16] D. Di Lorenzo, G. Liuzzi, F. Rinaldi, F. Schoen, and M. Sciandrone. “A concave optimization-based approach for sparse portfolio selection”. In: *Optimization Methods and Software* 27.6 (2012), pp. 983–1000. DOI: 10.1080/10556788.2011.577773.

[17] B. Dumitrescu and P. Irofti. *Dictionary Learning Algorithms and Applications.* Springer, 2018.

[18] 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.

[19] X. Jia, C. Kanzow, and P. Mehlitz. “Convergence Analysis of the Proximal Gradient Method in the Presence of the Kurdyka–Łojasiewicz Property Without Global Lipschitz Assumptions”. In: *SIAM Journal on Optimization* 33.4 (2023), pp. 3038–3056. DOI: 10.1137/23M1548293. URL: https://doi.org/10.1137/23M1548293.

[20] X. Jia, C. Kanzow, P. Mehlitz, and G. Wachsmuth. “An augmented Lagrangian method for optimization problems with structured geometric constraints”. In: *Mathematical Programming* (2022). DOI: 10.1007/s10107-022-01870-z.

[21] C. Kanzow and P. Mehlitz. “Convergence properties of monotone and nonmonotone proximal gradient methods revisited”. In: *Journal of Optimization Theory and Applications* 195.2 (2022), pp. 624–646. DOI: 10.1007/s10957-022-02101-3.

[22] K. Kurdyka. “On gradients of functions definable in o-minimal structures”. In: *Annales de l’institut Fourier* 48.3 (1998), pp. 769–783. DOI: 10.5802/aif.1638.

[23] Y.-F. Liu, Y.-H. Dai, and S. Ma. “Joint power and admission control: non-convex ℓ_q approximation and an effective polynomial time deflation approach”. In: *IEEE Transactions on Signal Processing* 63.14 (2015), pp. 3641–3656. DOI: 10.1109/TSP.2015.2428224.

[24] G. Marjanovic and V. Solo. “On ℓ_q optimization and matrix completion”. In: *IEEE Transactions on Signal Processing* 60.11 (2012), pp. 5714–5724. DOI: 10.1109/TSP.2012.2212015.

[25] I. Markovsky. *Low Rank Approximation*. Springer, 2018.

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

[27] P. Ochs. “Local convergence of the heavy-ball method and iPiano for non-convex optimization”. In: *Journal of Optimization Theory and Applications* 177.1 (2018), pp. 153–180. DOI: 10.1007/s10957-018-1272-y.

[28] P. Ochs, Y. Chen, T. Brox, and T. Pock. “iPiano: Inertial proximal algorithm for nonconvex optimization”. In: *SIAM Journal on Imaging Sciences* 7.2 (2014), pp. 1388–1419. DOI: 10.1137/130942954.

[29] Y. Qian and S. Pan. “Convergence of a class of nonmonotone descent methods for KL optimization problems”. In: *arXiv preprint arXiv:2201.00639* (2022).

[30] Y. Qian, T. Tao, S. Pan, and H. Qi. “Convergence of ZH-type nonmonotone descent method for Kurdyka-Łojasiewicz optimization problems”. In: *arXiv:2406.05740* (2024).

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

[32] J. J. Ye, X. Yuan, S. Zeng, and J. Zhang. “Variational analysis perspective on linear convergence of some first order methods for nonsmooth convex optimization problems”. In: *Set-Valued and Variational Analysis* (2021), pp. 1–35.

[33] 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.