Convergence Analysis of the Proximal Gradient Method in the Presence of the Kurdyka–Łojasiewicz Property without Global Lipschitz Assumptions

Xiaoxi Jia[*] Christian Kanzow[†] Patrick Mehlitz[‡]

January 12, 2023

Abstract. We consider a composite optimization problem where the sum of a continuously differentiable and a merely lower semicontinuous function has to be minimized. The proximal gradient algorithm is the classical method for solving such a problem numerically. The corresponding global convergence and local rate-of-convergence theory typically assumes, besides some technical conditions, that the smooth function has a globally Lipschitz continuous gradient and that the objective function satisfies the Kurdyka–Łojasiewicz property. Though this global Lipschitz assumption is satisfied in several applications where the objective function is, e.g., quadratic, this requirement is very restrictive in the non-quadratic case. Some recent contributions therefore try to overcome this global Lipschitz condition by replacing it with a local one, but, to the best of our knowledge, they still require some extra condition in order to obtain the desired global and rate-of-convergence results. The aim of this paper is to show that the local Lipschitz assumption together with the Kurdyka–Łojasiewicz property is sufficient to recover these convergence results.

Keywords. Non-Lipschitz Optimization, Nonsmooth Optimization, Proximal Gradient Method, Kurdyka–Łojasiewicz Property, Rate-of-Convergence

AMS subject classifications. 49J52, 90C30

1 Introduction

In this paper, we are concerned with problems from *composite optimization* where the sum of a continuously differentiable function f and a merely lower semicontinuous function ϕ has to be minimized. Problems of this type appear quite frequently in

[*] University of Würzburg, Institute of Mathematics, 97074 Würzburg, Germany, xiaoxi.jia@mathematik.uni-wuerzburg.de

[†] University of Würzburg, Institute of Mathematics, 97074 Würzburg, Germany, kanzow@mathematik.uni-wuerzburg.de, ORCID: 0000-0003-2897-2509

[‡] Brandenburgische Technische Universität Cottbus-Senftenberg, Institute of Mathematics, 03046 Cottbus, Germany, mehlitz@b-tu.de, ORCID: 0000-0002-9355-850X

many practically relevant areas like, e.g., machine learning, data compression, matrix completion, and image processing, see [10, 18, 19, 24, 30, 33], where, typically, f models a tracking-type term while ϕ is used to promote sparse structures in the solutions.

For an algorithmic treatment of such problems, it seems a nearby idea to exploit the composite form, i.e., differentiability of f on the one hand and additional structural properties of the function ϕ on the other hand (typically, the nonsmoothness encapsulated within ϕ is of specific type in all aforementioned applications). More precisely, the so-called proximal mapping of the function ϕ has to be available, which is typically the case in the aforementioned practically relevant scenarios. The idea behind the definition of proximal mappings is to interrelate the search for minimizers (or at least stationary points) with a fixed-point problem, and to apply a fixed-point iteration to the proximal mapping in order to tackle the minimization of the underlying function. Combining the available oracles for f and ϕ in order to construct an algorithm to minimize $f + \phi$ led to the development of so-called proximal gradient methods which date back to [25]. It is worth to note that proximal gradient algorithms can be interpreted as so-called forward-backward splitting methods which are far older, see [17, 37] for their origins and [6] for a modern view. Popular instances of proximal gradient methods are the iterative shrinkage/threshold algorithm (ISTA) and its accelerated version (FISTA = fast ISTA), see [8], where ϕ has to be convex. The monograph [7] presents a nice overview of existing results addressing proximal gradient methods where the nonsmooth part enjoys convexity.

It has been pointed out in the seminal works [4, 13] that the convergence theory for proximal gradient methods can be extended to situations where the nonsmooth part ϕ is merely lower semicontinuous and not necessarily convex. In both aforementioned papers, the analysis, which covers both (global) convergence and rate-of-convergence results, requires a so-called *descent lemma* as well as the celebrated *Kurdyka–Łojasiewicz property*, originating from [29, 31, 32]. The majority of available convergence results regarding proximal gradient methods seems to indicate that the price we have to pay for allowing ϕ to be nonsmooth is that the gradient ∇f of the smooth part has to be globally Lipschitz continuous. This requirement, which holds naturally when f is a (convex) quadratic function (as indicated above, this happens to be the case in many standard applications from image processing and data science), turns out to be rather restrictive in the non-quadratic situation which also is of practical interest, see Examples 3.6 and 3.7 below.

Let us review some contributions where the authors try to get rid of this global Lipschitz assumption. First, we would like to mention [5] where composite optimization problems with convex functions f and ϕ are considered without postulating global Lipschitzness of ∇f. It is shown that local Lipschitz continuity of ∇f is enough to obtain rate-of-convergence results for the iterates generated by a Bregman-type proximal gradient method. However, the authors of [5] require the additional assumption that there is a constant $L > 0$ such that $Lh - f$ is convex, where h is a convex function which defines the Bregman distance (let us mention that h equals the squared Euclidean norm in our setting). This convexity-type condition is satisfied in a couple of practically relevant situations. The approach of [5] was generalized to the nonconvex setting in [14] using, once again, a local Lipschitz assumption on ∇f, as well as the slightly stronger assumption (in order to deal with the nonconvexity) that there exist a constant $L > 0$ and a convex function h such that both $Lh - f$ and $Lh + f$ are convex. Let us emphasize that this constant L plays a central role in the design of the corresponding proximal-type methods. More precisely, it is used explicitly for the determination of the stepsizes. In the recent paper [21], global convergence results are proven under a local Lipschitz assumption on ∇f (without postulating any of the convexity-type conditions from above), but the authors assume (a priori) boundedness of iterates and stepsizes.

The present paper is based on [28] where the authors show global convergence results for proximal gradient methods in the sense that every accumulation point is shown to be a suitable stationary point of the composite optimization problem. The analysis in [28] is based on the local Lipschitz continuity of ∇f, and does not require the iterates to be bounded. An extension of this work, using a nonmonotone line search, is given in [22]. In contrast to most existing papers on proximal gradient methods, however, convergence of the entire sequence is not addressed in [22, 28]. Hence, no associated rate-of-convergence results could be given ([22] presents some standard worst-case rate-of-convergence results addressing the difference of two subsequent iterates along convergent subsequences). The aim of this paper is to fill this gap. More precisely, we show that the entire sequence generated by the proximal gradient method converges to a limit with a suitable rate, provided that this point satisfies the aforementioned Kurdyka–Łojasiewicz property. The underlying convergence theory is still based on a merely local Lipschitz assumption on ∇f, neither its global Lipschitzness nor the (a priori) boundedness of the iterates and stepsizes is presumed. To this end, we stress that our analysis is not based on any kind of descent lemma, which is in contrast to the contributions [5, 14] mentioned above.

The paper is organized as follows: In Section 2, we formally introduce the model problem of interest and provide some necessary notation as well as background material from generalized differentiation. The proximal gradient method together with the global convergence properties known from [28] are stated in Section 3. The convergence and rate-of-convergence analysis is then given in Section 4. We close with some final remarks in Section 5.

2 Problem Setting and Preliminaries

2.1 Problem Setting

Throughout the paper, we investigate the numerical treatment of the composite optimization problem

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

where $f\colon \mathbb{X} \to \mathbb{R}$ is continuously differentiable, $\phi\colon \mathbb{X} \to \overline{\mathbb{R}} := \mathbb{R} \cup \{\infty\}$ is lower semicontinuous (possibly infinite-valued and nondifferentiable), and $\mathbb{X}$ denotes a Euclidean space, i.e., a real and finite-dimensional Hilbert space. Since we do not want to deal with trivial situations, we assume that there exist points in $\mathbb{X}$ where the value of ϕ is finite. Let us underline that $\mathbb{X}$ is chosen to be Euclidean because this allows to cover applications from matrix analysis like low-rank optimization or matrix completion.

In order to minimize the function $\psi\colon \mathbb{X} \to \overline{\mathbb{R}}$ in (P), we will exploit its composite structure which allows for gradient steps with respect to the continuously differentiable function f on the one hand and so-called proximal steps with respect to ϕ on the other hand, i.e., we rely on a splitting approach. Throughout the last decades, experiments on numerous practically relevant optimization problems have shown that splitting methods are superior to the direct applications of standard methods from nonsmooth optimization to the function ψ.

2.2 Basic Notation

Throughout the paper, the Euclidean space $\mathbb{X}$ will be equipped with the inner product $\langle\cdot,\cdot\rangle\colon \mathbb{X}\times\mathbb{X}\to\mathbb{R}$ and the associated norm $\|\cdot\|$. Given a set $A\subset\mathbb{X}$ and an element $x\in\mathbb{X}$, we use $A+x := x+A := \{x\}+A := \{x+a \mid a\in A\}$ for brevity. Furthermore,

$$\operatorname{dist}(x,A) := \inf\{\|y-x\| \mid y\in A\}$$

denotes the distance of the point x to the set A with $\operatorname{dist}(x,\emptyset) := \infty$. For given $\varepsilon>0$, $B_\varepsilon(x) := \{y\in\mathbb{X} \mid \|y-x\|\le\varepsilon\}$ denotes the closed ε-ball around x.

The continuous linear operator $f'(x)\colon \mathbb{X}\to\mathbb{R}$ denotes the derivative of the continuously differentiable function $f\colon\mathbb{X}\to\mathbb{R}$ at $x\in\mathbb{X}$, and we will make use of $\nabla f(x) := f'(x)^*1$ where $f'(x)^*\colon\mathbb{R}\to\mathbb{X}$ is the adjoint of $f'(x)$. This way, ∇f is a mapping from $\mathbb{X}$ to $\mathbb{X}$.

We further say that a sequence $\{x^k\}\subset\mathbb{X}$ converges *Q-linearly* to $x^*\in\mathbb{X}$ if there is a constant $c\in(0,1)$ such that the inequality

$$\|x^{k+1}-x^*\|\le c\|x^k-x^*\|$$

holds for all sufficiently large $k\in\mathbb{N}$. Furthermore, $\{x^k\}$ is said to converge *R-linearly* to x^* if we have

$$\limsup_{k\to\infty}\|x^k-x^*\|^{1/k}<1.$$

Note that this R-linear convergence holds if there exist constants $\omega>0$ and $\mu\in(0,1)$ such that $\|x^k-x^*\|\le\omega\mu^k$ holds for all sufficiently large $k\in\mathbb{N}$, i.e., if the expression $\|x^k-x^*\|$ is dominated by a Q-linearly convergent null sequence.

2.3 Generalized Differentiation

The following concepts are standard in variational analysis, and we refer the interested reader to the monographs [34, 39] for more details.

Let us fix a merely lower semicontinuous function $\vartheta\colon\mathbb{X}\to\overline{\mathbb{R}}$ and pick $x\in\operatorname{dom}\vartheta$ where $\operatorname{dom}\vartheta := \{x\in\mathbb{X}\mid\vartheta(x)<\infty\}$ denotes the domain of ϑ. Then the set

$$\widehat{\partial}\vartheta(x) := \left\{\eta\in\mathbb{X}\,\middle|\,\liminf_{y\to x,\,y\ne x}\frac{\vartheta(y)-\vartheta(x)-\langle\eta,y-x\rangle}{\|y-x\|}\ge 0\right\}$$

is called the *regular* (or *Fréchet*) *subdifferential* of ϑ at x. Furthermore, the set

$$\partial\vartheta(x) := \left\{\eta \in \mathbb{X} \,\middle|\, \begin{array}{l} \exists \{x^k\}, \{\eta^k\} \subset \mathbb{X}\colon \\ \qquad x^k \to x,\, \vartheta(x^k) \to \vartheta(x),\, \eta^k \to \eta,\, \eta^k \in \widehat{\partial}\vartheta(x^k)\,\forall k \in \mathbb{N} \end{array}\right\}$$

is well known as the *limiting* (or *Mordukhovich*) *subdifferential* of ϑ at x. Clearly, we always have $\widehat{\partial}\vartheta(x) \subset \partial\vartheta(x)$ by construction of these sets. Whenever ϑ is a convex function, equality holds, and both subdifferentials coincide with the subdifferential of convex analysis, i.e.,

$$\widehat{\partial}\vartheta(x) = \partial\vartheta(x) = \{\eta \in \mathbb{X} \,|\, \forall y \in \operatorname{dom}\vartheta\colon\; \vartheta(y) \geq \vartheta(x) + \langle \eta, y - x\rangle\}$$

is valid in this situation. By definition of the regular subdifferential, it is clear that whenever $x^* \in \operatorname{dom}\vartheta$ is a local minimizer of ϑ, then $0 \in \widehat{\partial}\vartheta(x^*)$ hold. The latter fact is known as Fermat's rule, see [34, Proposition 1.30(i)]. Thus, the inclusion $0 \in \partial\vartheta(x^*)$ is a necessary optimality condition for x^* being a local minimizer of ϑ as well. Note that, for ϑ being convex, this necessary optimality condition is also sufficient for (global) minimality of x^* for ϑ.

Let us now apply this to the special case where $\vartheta := \psi$ is the sum of the continuously differentiable function f and a merely lower semicontinuous function ϕ, as it happens to be the case when investigating (P). Whenever $x \in \operatorname{dom}\phi$ is fixed, the sum rule

$$\partial(f + \phi)(x) = \nabla f(x) + \partial\phi(x) \tag{2.1}$$

holds due to the assumed continuous differentiability of f, see [34, Proposition 1.30(ii)]. Application of Fermat's rule therefore shows that the optimality condition

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

holds at any local minimizer $x^* \in \operatorname{dom}\phi$ of the composite optimization problem (P). Any point $x^* \in \operatorname{dom}\phi$ satisfying this necessary optimality condition will be called an *M-stationary point* of (P) due to the appearance of the limiting (or Mordukhovich) subdifferential.

We next introduce the famous Kurdyka–Łojasiewicz property that was already mentioned in Section 1 and which plays a central role in our subsequent convergence analysis. The version of this property stated below is a generalization of the classical Kurdyka–Łojasiewicz inequality for nonsmooth functions as introduced in [3, 11, 12] and afterwards used in the local convergence analysis of several nonsmooth optimization methods, see [2, 4, 13, 15, 16, 35, 36] for a couple of examples.

Definition 2.1. Let $g\colon \mathbb{X} \to \overline{\mathbb{R}}$ be lower semicontinuous. We say that g has the *KL property*, where KL abbreviates *Kurdyka–Łojasiewicz*, at $x^* \in \{x \in \mathbb{X} \,|\, \partial g(x) \neq \emptyset\}$ if there exist a constant $\eta > 0$, a neighborhood $U \subset \mathbb{X}$ of x^*, and a continuous concave function $\chi\colon [0, \eta] \to [0, \infty)$ which is continuously differentiable on $(0, \eta)$ and satisfies $\chi(0) = 0$ as well as $\chi'(t) > 0$ 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 $x \in U \cap \big\{x \in \mathbb{X} \,|\, g(x^*) < g(x) < g(x^*) + \eta\big\}$. The function χ from above is referred to as the *desingularization function*.

We note that there exist classes of functions where the KL property holds with the corresponding desingularization function given by $\chi(t) := ct^{\kappa}$ for $\kappa \in (0,1]$ and some constant $c > 0$, where the parameter κ is called the *KL exponent*, see [12, 29].

3 A Proximal Gradient Method and its Global Convergence Properties

This section begins with a formal description of a proximal gradient method for the composite optimization problem (P), and then summarizes the associated global convergence properties established in [28]. Note that our proximal gradient method uses a line search which is important to get global convergence properties without a global Lipschitz assumption. We start with a precise statement of the algorithm.

Algorithm 3.1 (Proximal Gradient Method).

```
Require: $\tau > 1$, $0 < \gamma_{\min} \le \gamma_{\max} < \infty$, $\delta \in (0,1)$, $x^0 \in \operatorname{dom}\phi$
Set $k := 0$.
while A suitable termination criterion is violated at iteration $k$ do
    Choose $\gamma_k^0 \in [\gamma_{\min}, \gamma_{\max}]$.
    For $i = 0, 1, 2, \ldots$, compute a solution $x^{k,i}$ of
        $\min_x \; f(x^k) + \langle \nabla f(x^k), x - x^k \rangle + \frac{\gamma_{k,i}}{2}\|x - x^k\|^2 + \phi(x), \quad x \in \mathbb{X} \qquad (3.1)$
    with $\gamma_{k,i} := \tau^i \gamma_k^0$, until the acceptance criterion
        $\psi(x^{k,i}) \le \psi(x^k) - \delta \frac{\gamma_{k,i}}{2}\|x^{k,i} - x^k\|^2 \qquad (3.2)$
    holds.
    Denote by $i_k := i$ the terminal value, and set $\gamma_k := \gamma_{k,i_k}$ and $x^{k+1} := x^{k,i_k}$.
    Set $k \leftarrow k + 1$.
end while
return $x^k$
```

Our convergence analysis requires some technical assumptions as well as a local Lipschitz condition on the gradient of the continuously differentiable function f.

Assumption 3.2.

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

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

(c) The function $\nabla f \colon \mathbb{X} \to \mathbb{X}$ is locally Lipschitz continuous.

Keeping in mind that our goal is to minimize the function ψ in (P), Assumption 3.2 (a) is reasonable. Furthermore, Assumption 3.2 (b) is employed to guarantee existence of solutions for the appearing subproblems (3.1). To be precise, Assumption 3.2 (b) implies that the objective function of the subproblem (3.1) is, for fixed

$k, i \in \mathbb{N}$, coercive, and therefore always attains a global minimizer $x^{k,i}$ (which does not need to be unique). Finally, the local Lipschitz condition for ∇f from Assumption 3.2 (c) will play a crucial role especially in Section 4 where we consider situations where a sequence generated by Algorithm 3.1 converges as a whole and give associated rate-of-convergence results.

In the following, we recall the central global convergence properties of Algorithm 3.1 whose proofs can be found in [28, Section 3]. Note that, throughout our analysis of Algorithm 3.1, we implicitly assume that this method generates an infinite sequence. For a discussion of a practical termination criterion, we refer to [28, Remark 3.1] for more details.

First, we recall that the stepsize rule in Step 4 of Algorithm 3.1 is always finite if the current iterate is not already stationary. Hence, the overall method is well-defined.

Lemma 3.3. *Consider a fixed iteration $k \in \mathbb{N}$ of Algorithm 3.1, assume that x^k is not an M-stationary point of (P), and suppose that Assumption 3.2 (b) holds. Then the inner loop in Step 4 of Algorithm 3.1 is finite, i.e., we have $\gamma_k = \gamma_{k,i_k}$ for some finite index $i_k \in \{0, 1, 2, \dots\}$.*

The following result summarizes some of the properties of Algorithm 3.1 that will later be used in Section 4.

Proposition 3.4. *Let Assumption 3.2 (a) and (b) hold, and let $\{x^k\}$ be a sequence generated by Algorithm 3.1. Then the following statements hold:*

(a) $\|x^{k+1} - x^k\| \to 0$ *as* $k \to \infty$,

(b) *for any convergent subsequence* $\{x^k\}_K$, $\gamma_k \|x^{k+1} - x^k\| \to_K 0$ *holds as* $k \to_K \infty$,

(c) *if, additionally, Assumption 3.2 (c) is valid, then for any convergent subsequence* $\{x^k\}_K$, $\{\gamma_k\}_K$ *is bounded.*

Finally, we restate the main global convergence result for Algorithm 3.1, see again [28, Section 3] for the corresponding details.

Theorem 3.5. *Let Assumption 3.2 be satisfied. Then each accumulation point of a sequence $\{x^k\}$ generated by Algorithm 3.1 is an M-stationary point of (P).*

Note that [28, Theorem 3.1] shows that a result like Theorem 3.5 also holds without any Lipschitz condition regarding ∇f, but it then requires a slightly stronger condition for the nonsmooth function ϕ, namely the continuity of ϕ on its domain (this condition holds, e.g., if ϕ is the indicator function of a constraint set). Our analysis in Section 4, however, requires the local Lipschitz condition for the gradient ∇f, so we decided to treat it as a standing assumption.

We close this section by mentioning two classes of examples where the standard global Lipschitz assumption on the gradient of f is typically violated, whereas a local Lipschitz condition is often satisfied.

Example 3.6. (Augmented Lagrangian Methods)
Consider the constrained optimization problem

$$\min_x \; f(x) + \phi(x) \quad \text{s.t.} \quad c(x) \in C,$$

where $f\colon \mathbb{X} \to \mathbb{R}$ and $\phi\colon \mathbb{X} \to \overline{\mathbb{R}}$ are as in (P). In addition, we have some constraints defined by a continuously differentiable function $c\colon \mathbb{X} \to \mathbb{Y}$, where $\mathbb{Y}$ is another Euclidean space, and a nonempty, closed, and convex set $C \subset \mathbb{Y}$.

Given a current iterate $x^k \in \mathbb{X}$ and a corresponding Lagrange multiplier estimate $\lambda^k \in \mathbb{Y}$, augmented Lagrangian techniques then compute the next iterate x^{k+1} by solving (approximately) the subproblem

$$\min_x f(x) + \phi(x) + \frac{\rho_k}{2} \operatorname{dist}^2 \left(c(x) + \frac{\lambda^k}{\rho_k}, C \right), \qquad x \in \mathbb{X}$$

for some penalty parameter $\rho_k > 0$. Since the squared distance function $y \mapsto \operatorname{dist}^2(y, C)$ is continuously differentiable by convexity of C, see [6, Corollary 12.30], this subproblem has exactly the structure of the composite optimization problem (P) and can therefore, in principle, be solved by a proximal gradient method, see [20,23,26,27] for suitable realizations of this approach.

Assuming that the gradient of the smooth part of this objective function (with respect to the variable x) is globally Lipschitz continuous, however, is pretty strong is this setting and, basically, requires the constraint function c to be linear and the set C to be polyhedral, whereas local Lipschitzness of this gradient holds under mild conditions on the smoothness of f and c.

The following example makes use of conjugate functions, see [6, Definition 3.1]. Since, within this paper, they only occur in this particular application, we refrain from stating their precise definitions and properties, and refer the interested reader to the excellent monographs [6,7,39] for more details.

Example 3.7. (Dual Proximal Gradient Methods)
Consider the (primal) optimization problem

$$\min_x \; g(x) + h(Ax), \qquad x \in \mathbb{X} \tag{3.3}$$

where both functions $g\colon \mathbb{X} \to \overline{\mathbb{R}}$ and $h\colon \mathbb{Y} \to \overline{\mathbb{R}}$ are lower semicontinuous and convex while possessing nonempty domains, and $A\colon \mathbb{X} \to \mathbb{Y}$ is a linear operator. Above, $\mathbb{Y}$ is another Euclidean space. Note that none of the functions g or h is assumed to be (continuously) differentiable.

The (Fenchel) dual problem of (3.3) is given by

$$\min_y \; g^*(A^*y) + h^*(-y), \qquad y \in \mathbb{Y} \tag{3.4}$$

with the two conjugate functions $g^*\colon \mathbb{X} \to \overline{\mathbb{R}}$ and $h^*\colon \mathbb{Y} \to \overline{\mathbb{R}}$ being lower semicontinuous and convex, and $A^*\colon \mathbb{Y} \to \mathbb{X}$ being the adjoint of A. Under suitable assumptions, the pair (3.3), (3.4) enjoys strong duality, i.e., the optimal objective function values of these problems coincide, see [38], which motivates to solve (3.4) instead of (3.3) in some applications where the conjugate functions are explicitly available.

Assuming, in addition, that g is uniformly convex, it is known that g^* is real-valued everywhere and continuously differentiable with a globally Lipschitz continuous gradient, see [39, Proposition 12.60]. Consequently, as promoted in [9], a standard proximal gradient algorithm can be applied to the dual problem (3.4). On the other hand, if g is only strictly convex, then the domain of g^* is, in general, no longer the entire space, but g^* can still be shown to be continuously differentiable on the interior of its domain. Its gradient, however, is no longer guaranteed to be globally Lipschitz continuous on the domain.

4 Convergence Analysis in the Presence of the KL Property

The aim of this section is to show convergence of the entire sequence $\{x^k\}$ generated by Algorithm 3.1 provided that there exists an accumulation point x^* which, in addition, satisfies the KL property, and to present associated rate-of-convergence results. The proofs of these results are based on a local Lipschitz assumption on ∇f only, without the a priori assumption that the whole sequence $\{x^k\}$ is bounded. Based on some recent contributions in the area of proximal gradient and related first-order methods, it seems reasonable to expect such a result to hold. For example, [13, 35] consider a whole class of first-order methods and investigate their (essentially local) convergence showing, in particular, that the entire sequence $\{x^k\}$ generated by their methods stays within a certain neighborhood of a solution provided that the KL property holds at this solution.

Their approach is not directly applicable to our situation since, on the one hand, we do not use the a priori assumption that our iterates are bounded, and, on the other hand, because the adaption of the methods considered in [13, 35] to the proximal gradient setting would result in an algorithm with a constant stepsize. However, having an accumulation point of Algorithm 3.1 satisfying the KL property, we know from the local Lipschitz assumption on ∇f that a respective global Lipschitz condition holds in a suitable neighborhood of this point, which then can be used to verify that the stepsizes computed by Algorithm 3.1 remain bounded. This – more or less heuristic – idea fortifies us to believe that one can also get convergence and rate-of-convergence results under the KL property in the presence of Assumption 3.2 (c). The following analysis is a careful mathematical realization of this somewhat vague idea.

We begin with a result which shows that, locally around an accumulation point of the sequence $\{x^k\}$, the associated stepsizes γ_k remain bounded. This observation and its proof are related to [28, Corollary 3.1]. Note that this statement is essentially different from the boundedness of stepsizes along convergent subsequences of iterates which is inherent in the presence of Assumption 3.2, see Proposition 3.4 (c).

Lemma 4.1. *Let Assumption 3.2 hold, let $\{x^k\}$ be any sequence generated by Algorithm 3.1, and let x^* be an accumulation point of this sequence. Then, for any $\rho > 0$, there is a constant $\bar{\gamma}_\rho > 0$ (usually depending on ρ) such that $\gamma_k \leq \bar{\gamma}_\rho$ holds for all $k \in \mathbb{N}$ such that $x^k \in B_\rho(x^*)$.*

Proof. First, recall from Lemma 3.3 that the stepsize γ_k is well-defined for each $k \in \mathbb{N}$. Let $\rho > 0$ be fixed, and recall that the assumed local Lipschitz continuity of ∇f implies that this gradient mapping is (globally) Lipschitz continuous on the compact set $B_{2\rho}(x^*)$ (note that we took 2ρ as the radius of this ball here). Let us denote the corresponding Lipschitz constant by $L_{2\rho}$. Since x^* is an accumulation point of the sequence $\{x^k\}$, there are infinitely many iterates of this sequence belonging to $B_\rho(x^*)$.

Now, assume, by contradiction, that there is a subsequence $\{\gamma_k\}_K$ with $x^k \in B_\rho(x^*)$ for all $k \in K$ such that $\{\gamma_k\}_K$ is unbounded. Without loss of generality, we may assume that $\gamma_k \to_K \infty$, that the subsequence of iterates $\{x^k\}_K$ converges to some point $\bar{x}$ (not necessarily equal to x^*), and that, for each $k \in K$, the acceptance criterion (3.2) is violated in the first iteration of the inner loop. Then, for the trial stepsize $\hat{\gamma}_k := \gamma_k/\tau = \tau^{i_k-1}\gamma_k^0$, we also have $\hat{\gamma}_k \to_K \infty$, whereas the corresponding trial vector $\hat{x}^k := x^{k,i_k-1}$ does not satisfy the acceptance criterion from (3.2), i.e., we have

$$\psi(\hat{x}^k) > \psi(x^k) - \delta\frac{\hat{\gamma}_k}{2}\|\hat{x}^k - x^k\|^2 \quad \forall k \in K. \tag{4.1}$$

On the other hand, since $\hat{x}^k$ solves the corresponding subproblem (3.1) with $\hat{\gamma}_k$ in place of $\gamma_{k,i}$, we have

$$\langle \nabla f(x^k), \hat{x}^k - x^k \rangle + \frac{\hat{\gamma}_k}{2}\|\hat{x}^k - x^k\|^2 + \phi(\hat{x}^k) - \phi(x^k) \leq 0. \tag{4.2}$$

We claim that this, in particular, implies $\hat{x}^k \to_K \bar{x}$. In fact, using (4.2), the Cauchy-Schwarz inequality, and the fact that $\{\psi(x^k)\}$ is monotonically decreasing by construction of Algorithm 3.1, we obtain

$$\begin{aligned}
\frac{\hat{\gamma}_k}{2}\|\hat{x}^k - x^k\|^2 &\leq \|\nabla f(x^k)\|\|\hat{x}^k - x^k\| + \phi(x^k) - \phi(\hat{x}^k) \\
&= \|\nabla f(x^k)\|\|\hat{x}^k - x^k\| + \psi(x^k) - f(x^k) - \phi(\hat{x}^k) \\
&\leq \|\nabla f(x^k)\|\|\hat{x}^k - x^k\| + \psi(x^0) - f(x^k) - \phi(\hat{x}^k).
\end{aligned}$$

Since f is continuously differentiable and $-\phi$ is bounded from above by an affine function in view of Assumption 3.2 (b), the above estimate implies $\|\hat{x}^k - x^k\| \to_K 0$. In fact, if $\{\|\hat{x}^k - x^k\|\}_K$ would be unbounded, then the left-hand side would grow more rapidly than the right-hand side, and if $\{\|\hat{x}^k - x^k\|\}_K$ would be bounded, but staying away, at least on a subsequence, from zero by a positive number, the right-hand side would be bounded, whereas the left-hand side would be unbounded on the corresponding subsequence. Consequently, we have $\|\hat{x}^k - x^k\| \to_K 0$, and since $x^k \to_K \bar{x}$, this implies $\hat{x}^k \to_K \bar{x}$. In particular, since $\bar{x} \in B_\rho(x^*)$, this implies that, for all sufficiently large $k \in K$, we have both $x^k \in B_{2\rho}(x^*)$ and $\hat{x}^k \in B_{2\rho}(x^*)$.

Let us fix some $k \in K$. Using the mean-value theorem yields the existence of a point ξ^k on the line segment connecting x^k with $\hat{x}^k$ such that

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

Substituting the resulting expression for $\phi(\hat{x}^k) - \phi(x^k)$ into (4.2) yields

$$\langle \nabla f(x^k) - \nabla f(\xi^k), \hat{x}^k - x^k \rangle + \frac{\hat{\gamma}_k}{2} \|\hat{x}^k - x^k\|^2 + \psi(\hat{x}^k) - \psi(x^k) \leq 0. \tag{4.3}$$

Exploiting (4.1), we therefore obtain

$$\begin{aligned} \frac{\hat{\gamma}_k}{2} \|\hat{x}^k - x^k\|^2 &\leq -\langle \nabla f(x^k) - \nabla f(\xi^k), \hat{x}^k - x^k \rangle + \psi(x^k) - \psi(\hat{x}^k) \\ &\leq \|\nabla f(x^k) - \nabla f(\xi^k)\| \|\hat{x}^k - x^k\| + \delta \frac{\hat{\gamma}_k}{2} \|\hat{x}^k - x^k\|^2 \end{aligned}$$

which can be rewritten as

$$(1 - \delta) \frac{\hat{\gamma}_k}{2} \|\hat{x}^k - x^k\| \leq \|\nabla f(x^k) - \nabla f(\xi^k)\|.$$

Since ξ^k in an element from the line connecting x^k and $\hat{x}^k$, it follows that $\xi^k \in B_{2\rho}(x^*)$ for all $k \in K$ sufficiently large. Hence, the Lipschitz continuity of ∇f on this ball yields

$$(1 - \delta) \frac{\hat{\gamma}_k}{2} \|\hat{x}^k - x^k\| \leq L_{2\rho} \|x^k - \xi^k\| \leq L_{2\rho} \|x^k - \hat{x}^k\|$$

for all sufficiently large $k \in K$. Since $\hat{x}^k \neq x^k$ in view of (4.1), this implies that $\{\hat{\gamma}_k\}_K$ is bounded which, in turn, yields the boundedness of the subsequence $\{\gamma_k\}_K$, contradicting our assumption. This completes the proof. □

We next show that the entire sequence $\{\psi(x^k)\}$ converges to $\psi(x^*)$, where x^* is an arbitrary accumulation point of a sequence $\{x^k\}$ generated by Algorithm 3.1. Note that this result is not completely obvious since ψ is only lower semicontinuous but not continuous in general. Indeed, this property results from the construction of the iterates x^{k+1} of Algorithm 3.1.

Lemma 4.2. *Let Assumption 3.2 be satisfied, and let x^* be an accumulation point of a sequence $\{x^k\}$ generated by Algorithm 3.1. Then the entire sequence $\{\psi(x^k)\}$ converges to $\psi(x^*)$.*

Proof. Let $\{x^k\}_K$ be a subsequence converging to x^*. By means of Proposition 3.4 (a), we also have $x^{k+1} \to_K x^*$. Since ψ is lower semicontinuous, we then obtain

$$\psi(x^*) \leq \liminf_{k \to_K \infty} \psi(x^{k+1}). \tag{4.4}$$

On the other hand, by construction, the entire sequence $\{\psi(x^k)\}$ is monotonically decreasing. Since it is also bounded from below by $\psi(x^*)$ as a consequence of (4.4), it follows that the whole sequence $\{\psi(x^k)\}$ converges. It remains to show that its limit is equal to (the lower bound) $\psi(x^*)$.

To this end, we first note that x^{k+1} solves the subproblem (3.1) with stepsize γ_k in place of $\gamma_{k,i}$. Hence, we have

$$\begin{aligned} &\langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{\gamma_k}{2} \|x^{k+1} - x^k\|^2 + \phi(x^{k+1}) \\ &\qquad \leq \langle \nabla f(x^k), x^* - x^k \rangle + \frac{\gamma_k}{2} \|x^* - x^k\|^2 + \phi(x^*) \end{aligned}$$

for each $k \in \mathbb{N}$. Taking the upper limit as $k \to_K \infty$, and using the continuity of ∇f as well as Proposition 3.4, we obtain

$$\limsup_{k \to_K \infty} \phi(x^{k+1}) \leq \phi(x^*).$$

Combining this with (4.4) and using the continuity of f yields $\psi(x^{k+1}) \to_K \psi(x^*)$. Since $\{\psi(x^k)\}$ converges, the assertion follows. □

All results stated so far are independent of the KL property. The remaining part of our analysis, however, is heavily based on the assumption that our objective function ψ satisfies the KL property at a given accumulation point x^* of a sequence $\{x^k\}$ generated by Algorithm 3.1. In particular, let $\eta > 0$ be the corresponding constant from the definition of the associated desingularization function χ. Furthermore, we will assume that Assumption 3.2 is valid. In view of Proposition 3.4, we can find a sufficiently large index $\hat{k} \in \mathbb{N}$ such that

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

We then define

$$\rho := \eta + \frac{1}{2} \tag{4.6}$$

as well as the compact set

$$C_\rho := B_\rho(x^*) \cap \mathcal{L}_\psi(x^0), \tag{4.7}$$

where $\mathcal{L}_\psi(x^0) := \{x \in \mathbb{X} \,|\, \psi(x) \leq \psi(x^0)\}$ is the sublevel set of ψ with respect to x^0, the starting point exploited in Algorithm 3.1. By monotonicity of $\{\psi(x^k)\}$, we have $\{x^k\} \subset \mathcal{L}_\psi(x^0)$. Finally, throughout the section, let $L_\rho > 0$ be a (global) Lipschitz constant of ∇f on C_ρ from (4.7). Finally, in view of Lemma 4.1, we have

$$\gamma_k \leq \bar{\gamma}_\rho \quad \forall x^k \in C_\rho \tag{4.8}$$

with some suitable upper bound $\bar{\gamma}_\rho > 0$ (depending on our choice of ρ from (4.6)). Using this notation, we can formulate the following result.

Lemma 4.3. *Let Assumption 3.2 hold, and let $\{x^k\}$ be any sequence generated by Algorithm 3.1. Suppose that $\{x^k\}_K$ is a subsequence converging to some limit point x^*, and that ψ has the KL property at x^* with desingularization function χ. Then there is a sufficiently large constant $k_0 \in K$ such that the corresponding constant*

$$\alpha := \|x^{k_0} - x^*\| + \sqrt{\frac{8\big(\psi(x^{k_0}) - \psi(x^*)\big)}{\delta \gamma_{\min}}} + \frac{2\big(\bar{\gamma}_\rho + L_\rho\big)}{\delta \gamma_{\min}} \chi\big(\psi(x^{k_0}) - \psi(x^*)\big) \tag{4.9}$$

satisfies $\alpha < \frac{1}{2}$, where $\rho > 0$ and $\bar{\gamma}_\rho > 0$ are the constants defined in (4.6) and (4.8), respectively, while $L_\rho > 0$ is a Lipschitz constant of ∇f on C_ρ from (4.7), and $\delta > 0$ as well as $\gamma_{\min} > 0$ are the parameters from Algorithm 3.1.

Proof. The statement follows from the fact that each summand on the right-hand side of (4.9) can be made arbitrarily small. This is clear for the first one since the subsequence $\{x^k\}_K$ converges to x^*. This is also true for the second summand as a consequence of Lemma 4.2. Finally, the third one can be made arbitrarily small since we have $\psi(x^k) \to \psi(x^*)$ by Lemma 4.2, taking into account that the desingularization function χ is continuous at the origin. Hence, the statement follows by taking an index $k_0 \in K$ sufficiently large. □

We next state another technical result.

Lemma 4.4. *Let Assumption 3.2 hold, and let $\{x^k\}$ be any sequence generated by Algorithm 3.1. Suppose that $\{x^k\}_K$ is a subsequence converging to some limit point x^*, and that ψ has the KL property at x^* with desingularization function χ. Then*

$$\operatorname{dist}\big(0, \partial\psi(x^{k+1})\big) \leq \big(\bar{\gamma}_\rho + L_\rho\big)\|x^{k+1} - x^k\|$$

holds for all sufficiently large $k \in \mathbb{N}$ such that $x^k \in B_\alpha(x^*)$, where $\alpha < \frac{1}{2}$ denotes the constant from (4.9), $\bar{\gamma}_\rho > 0$ is the constant from (4.8), and $L_\rho > 0$ is the Lipschitz constant of ∇f on C_ρ from (4.7).

Proof. For any $k \in \mathbb{N}$, since x^{k+1} is a solution of (3.1), we obtain

$$0 \in \nabla f(x^k) + \gamma_k(x^{k+1} - x^k) + \partial\phi(x^{k+1})$$

from the corresponding M-stationary condition. This implies

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

for all $k \in \mathbb{N}$, where we used the sum rule (2.1) for the limiting subdifferential.

Now, take an arbitrary index $k \in \mathbb{N}$ sufficiently large such that $x^k \in B_\alpha(x^*)$ and $k \geq \hat{k}$, where $\hat{k}$ is the index from (4.5). In view of (4.6) and Lemma 4.3, we have $\alpha \leq \rho$. Therefore, Lemma 4.1 shows that

$$\gamma_k \leq \bar{\gamma}_\rho. \tag{4.11}$$

Moreover, using (4.5), (4.6), and Lemma 4.3, we get

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

Hence, $x^k, x^{k+1} \in C_\rho$ holds with the compact set C_ρ from (4.7). Therefore, we have

$$\big\|\nabla f(x^{k+1}) - \nabla f(x^k)\big\| \leq L_\rho\|x^{k+1} - x^k\|$$

by definition of L_ρ. Together with (4.10) and (4.11), we thus obtain

$$\begin{aligned}\operatorname{dist}\big(0, \partial\psi(x^{k+1})\big) &\leq \big\|\gamma_k(x^k - x^{k+1}) + \nabla f(x^{k+1}) - \nabla f(x^k)\big\| \\ &\leq \gamma_k\|x^{k+1} - x^k\| + L_\rho\|x^{k+1} - x^k\| \\ &\leq \big(\bar{\gamma}_\rho + L_\rho\big)\|x^{k+1} - x^k\|\end{aligned}$$

for all $k \in \mathbb{N}$ satisfying $k \geq \hat{k}$ and $x^k \in B_\alpha(x^*)$. □

The following result shows that the entire sequence $\{x^k\}$, generated by Algorithm 3.1, already converges to one of its accumulation points x^* provided that the objective function ψ satisfies the KL property at this point. The proof combines our previous results with a technique used in [13].

Theorem 4.5. *Let Assumption 3.2 hold, and let $\{x^k\}$ be any sequence generated by Algorithm 3.1. Suppose that $\{x^k\}_K$ is a subsequence converging to some limit point x^*, and that ψ has the KL property at x^*. Then the entire sequence $\{x^k\}$ converges to x^*.*

Proof. In view of Lemma 4.2, we know that the whole sequence $\{\psi(x^k)\}$ is monotonically decreasing and converging to $\psi(x^*)$. This implies that $\psi(x^k) \geq \psi(x^*)$ holds for all $k \in \mathbb{N}$.

Now, suppose we have $\psi(x^k) = \psi(x^*)$ for some index $k \in \mathbb{N}$. Then, by monotonicity, we also get $\psi(x^{k+1}) = \psi(x^*)$. Consequently, we obtain from (3.2) that

$$0 \leq \frac{\delta \gamma_{\min}}{2} \|x^{k+1} - x^k\|^2 \leq \psi(x^k) - \psi(x^{k+1}) = 0$$

and, thus, $x^{k+1} = x^k$. Since, by assumption, the subsequence $\{x^k\}_K$ converges to x^*, this implies that $x^k = x^*$ for all $k \in \mathbb{N}$ sufficiently large. In particular, we have convergence of the entire (eventually constant) sequence $\{x^k\}$ to x^* in this situation.

For the remainder of this proof, we can therefore assume that $\psi(x^k) > \psi(x^*)$ holds for all $k \in \mathbb{N}$. We then let $\alpha \in (0, 1/2)$ be the constant from (4.9), and $k_0 \in K$ be the corresponding iteration index which is used in the definition of α, see Lemma 4.3. We then have $0 < \psi(x^k) - \psi(x^*) \leq \psi(x^{k_0}) - \psi(x^*)$ for all $k \geq k_0$. Without loss of generality, we may also assume that $k_0 \geq \hat{k}$ (the latter being the index defined by (4.5)) and that k_0 is sufficiently large to satisfy

$$\psi(x^{k_0}) < \psi(x^*) + \eta. \tag{4.12}$$

Let $\chi \colon [0, \eta] \to [0, \infty)$ be the desingularization function which comes along with the validity of the KL property at x^*. Due to $\chi(0) = 0$ and $\chi'(t) > 0$ for all $t \in (0, \eta)$, we obtain

$$\chi\big(\psi(x^k) - \psi(x^*)\big) \geq 0 \quad \forall k \geq k_0. \tag{4.13}$$

We now claim that the following two statements hold for all $k \geq k_0$:

(a) $x^k \in B_\alpha(x^*)$,

(b) $\|x^{k_0} - x^*\| + \sum_{i=k_0}^{k} \|x^{i+1} - x^i\| \leq \alpha$, which is equivalent to

$$\sum_{i=k_0}^{k} \|x^{i+1} - x^i\| \leq \sqrt{\frac{8\big(\psi(x^{k_0}) - \psi(x^*)\big)}{\delta \gamma_{\min}}} + \frac{2\big(\bar{\gamma}_\rho + L_\rho\big)}{\delta \gamma_{\min}} \chi\big(\psi(x^{k_0}) - \psi(x^*)\big). \tag{4.14}$$

We verify these two statements jointly by induction. For $k = k_0$, statement (a) holds simply by the definition of α in (4.9). Furthermore, the acceptance criterion (3.2) together with the monotonicity of $\{\psi(x^k)\}$ implies

$$\|x^{k_0+1} - x^{k_0}\| \leq \sqrt{\frac{2\big(\psi(x^{k_0}) - \psi(x^{k_0+1})\big)}{\delta \gamma_{\min}}} \leq \sqrt{\frac{2\big(\psi(x^{k_0}) - \psi(x^*)\big)}{\delta \gamma_{\min}}}. \tag{4.15}$$

In particular, this shows that (4.14) holds for $k = k_0$. Suppose that both statements hold for some $k \geq k_0$. Using the triangle inequality, the induction hypothesis, and the definition of α, we obtain

$$\|x^{k+1} - x^*\| \leq \sum_{i=k_0}^{k} \|x^{i+1} - x^i\| + \|x^{k_0} - x^*\| \leq \alpha,$$

i.e., statement (a) holds for $k+1$ in place of k. The verification of the induction step for (b) is more involved.

To this end, first note that (4.12) implies

$$\psi(x^*) < \psi(x^i) < \psi(x^*) + \eta \quad \forall i \geq k_0. \tag{4.16}$$

Since ψ has the KL property at x^*, we have

$$\chi'\big(\psi(x^i) - \psi(x^*)\big) \operatorname{dist}\big(0, \partial\psi(x^i)\big) \geq 1 \quad \forall i \geq k_0. \tag{4.17}$$

Since $x^i \in B_\alpha(x^*)$ for all $i \in \{k_0, k_0+1, \ldots, k\}$ by our induction hypothesis, we can apply Lemma 4.4 and obtain (after a simple index shift)

$$\operatorname{dist}\big(0, \partial\psi(x^i)\big) \leq \big(\bar{\gamma}_\rho + L_\rho\big)\|x^i - x^{i-1}\| \quad \forall i \in \{k_0+1, k_0+2, \ldots, k+1\}.$$

In view of (4.17), we therefore obtain

$$\chi'\big(\psi(x^i) - \psi(x^*)\big) \geq \frac{1}{\big(\bar{\gamma}_\rho + L_\rho\big)\|x^i - x^{i-1}\|} \quad \forall i \in \{k_0+1, k_0+2, \ldots, k+1\}. \tag{4.18}$$

To simplify some of the subsequent formulas, we follow [13] and introduce the shorthand notation

$$\Delta_{i,j} := \chi\big(\psi(x^i) - \psi(x^*)\big) - \chi\big(\psi(x^j) - \psi(x^*)\big)$$

for $i, j \in \mathbb{N}$. The assumed concavity of χ then implies

$$\Delta_{i,i+1} \geq \chi'\big(\psi(x^i) - \psi(x^*)\big)\big(\psi(x^i) - \psi(x^{i+1})\big). \tag{4.19}$$

Using (4.18), (4.19), and the acceptance criterion (3.2), we therefore get

$$\begin{aligned} \Delta_{i,i+1} &\geq \chi'\big(\psi(x^i) - \psi(x^*)\big)\big(\psi(x^i) - \psi(x^{i+1})\big) \\ &\geq \frac{\psi(x^i) - \psi(x^{i+1})}{(\bar{\gamma}_\rho + L_\rho)\|x^i - x^{i-1}\|} \geq \frac{\delta\gamma_{\min}}{2(\bar{\gamma}_\rho + L_\rho)} \frac{\|x^{i+1} - x^i\|^2}{\|x^i - x^{i-1}\|} = \beta\frac{\|x^{i+1} - x^i\|^2}{\|x^i - x^{i-1}\|} \end{aligned}$$

for all $i \in \{k_0+1, k_0+2, \ldots, k+1\}$, where we used the constant $\beta := \frac{\delta\gamma_{\min}}{2(\bar{\gamma}_\rho + L_\rho)}$. Noting that $a + b \geq 2\sqrt{ab}$ holds for all real numbers $a, b \geq 0$, we therefore obtain

$$\frac{1}{\beta}\Delta_{i,i+1} + \|x^i - x^{i-1}\| \geq 2\sqrt{\frac{1}{\beta}\Delta_{i,i+1}\|x^i - x^{i-1}\|} \geq 2\|x^{i+1} - x^i\|$$

for all $i \in \{k_0+1, k_0+2, \ldots, k+1\}$. Summation yields

$$\begin{aligned}
2\sum_{i=k_0+1}^{k+1} \|x^{i+1}-x^i\| &\leq \sum_{i=k_0+1}^{k+1} \|x^i - x^{i-1}\| + \frac{1}{\beta}\sum_{i=k_0+1}^{k+1} \Delta_{i,i+1} \\
&= \sum_{i=k_0+1}^{k} \|x^{i+1}-x^i\| + \|x^{k_0+1}-x^{k_0}\| + \frac{1}{\beta}\Delta_{k_0+1,k+2} \\
&\leq \sum_{i=k_0+1}^{k+1} \|x^{i+1}-x^i\| + \|x^{k_0+1}-x^{k_0}\| + \frac{1}{\beta}\Delta_{k_0+1,k+2}.
\end{aligned}$$

Subtracting the first summand from the right-hand side, exploiting the estimate (4.15), and using the nonnegativity as well as monotonicity of the desingularization function χ, we obtain

$$\sum_{i=k_0+1}^{k+1} \|x^{i+1}-x^i\| \leq \sqrt{\frac{2\big(\psi(x^{k_0})-\psi(x^*)\big)}{\delta\gamma_{\min}}} + \frac{1}{\beta}\chi\big(\psi(x^{k_0})-\psi(x^*)\big).$$

Adding the term $\|x^{k_0+1}-x^{k_0}\|$ to both sides and using (4.15) once again, we get

$$\sum_{i=k_0}^{k+1} \|x^{i+1}-x^i\| \leq \sqrt{\frac{8\big(\psi(x^{k_0})-\psi(x^*)\big)}{\delta\gamma_{\min}}} + \frac{1}{\beta}\chi\big(\psi(x^{k_0})-\psi(x^*)\big).$$

Hence, statement (b) holds for $k+1$ in place of k, and this completes the induction.

In particular, it follows from (a) that $x^k \in B_\alpha(x^*)$ for all $k \geq k_0$. Taking $k \to \infty$ in (4.14) therefore shows that $\{x^k\}$ is a Cauchy sequence and, thus, convergent. Since we already know that x^* is an accumulation point, it follows that the entire sequence $\{x^k\}$ converges to x^*. □

We finally state our rate-of-convergence result for one particular class of desingularization functions. The result holds for a more general class of such functions, and we comment on this after the proof. To keep the notation simple and since this result, having in mind the previous ones, is more or less a standard observation, we decided to state this rate-of-convergence result in the following way.

Theorem 4.6. *Let Assumption 3.2 hold, and let $\{x^k\}$ be any sequence generated by Algorithm 3.1. Suppose that $\{x^k\}_K$ is a subsequence converging to some limit point x^*, and that ψ has the KL property at x^*. Then the entire sequence $\{x^k\}$ converges to x^*, and if the corresponding desingularization function has the form $\chi(t) = ct^{1/2}$ for some $c > 0$, the following statements hold:*

(a) *the sequence $\{\psi(x^k)\}$ converges Q-linearly to $\psi(x^*)$,*

(b) *the sequence $\{x^k\}$ converges R-linearly to x^*.*

Proof. In view of Theorem 4.5, we only need to verify the quantitative statements (a) and (b) of the theorem.

As noted at the beginning of the proof of Theorem 4.5, we may assume, without loss of generality, that $\psi(x^k) > \psi(x^*)$ holds for all $k \in \mathbb{N}$. In view of Lemma 4.2, we then have

$$x^k \in B_\alpha(x^*) \cap \big\{x \in \operatorname{dom}\phi \mid \psi(x^*) < \psi(x) < \psi(x^*) + \eta\big\}$$

for all $k \in \mathbb{N}$ sufficiently large, where $\alpha > 0$ is the constant from (4.9) and $\eta > 0$ denotes the constant from the definition of the desingularization function χ. Since ψ satisfies the KL property at x^* with $\chi(t) = ct^{1/2}$, we have

$$\begin{aligned} 1 &\leq \chi'\big(\psi(x^{k+1}) - \psi(x^*)\big) \operatorname{dist}\big(0, \partial\psi(x^{k+1})\big) \\ &= \frac{c}{2}\big(\psi(x^{k+1}) - \psi(x^*)\big)^{-1/2} \operatorname{dist}\big(0, \partial\psi(x^{k+1})\big) \end{aligned}$$

for all sufficiently large $k \in \mathbb{N}$. Taking into account Lemma 4.4, this yields

$$1 \leq \frac{c(\bar{\gamma}_\rho + L_\rho)}{2}\big(\psi(x^{k+1}) - \psi(x^*)\big)^{-1/2}\|x^{k+1} - x^k\|$$

for all $k \in \mathbb{N}$ sufficiently large, where $\bar{\gamma}_\rho > 0$ is the constant from (4.8) and $L_\rho > 0$ is the global Lipschitz constant of ∇f on C_ρ from (4.7). Rearranging this expression yields

$$\|x^{k+1} - x^k\| \geq \frac{2}{c(\bar{\gamma}_\rho + L_\rho)}\big(\psi(x^{k+1}) - \psi(x^*)\big)^{1/2}. \tag{4.20}$$

On the other hand, by the acceptance criterion (3.2) and $\gamma_k \geq \gamma_{\min}$, we have

$$\psi(x^{k+1}) - \psi(x^k) \leq -\delta\frac{\gamma_{\min}}{2}\|x^{k+1} - x^k\|^2. \tag{4.21}$$

Combining (4.20) and (4.21), we obtain

$$\begin{aligned} \big(\psi(x^{k+1}) - \psi(x^*)\big) - \big(\psi(x^k) - \psi(x^*)\big) &= \psi(x^{k+1}) - \psi(x^k) \\ &\leq -\delta\frac{\gamma_{\min}}{2}\|x^{k+1} - x^k\|^2 \\ &\leq -\frac{2\delta\gamma_{\min}}{c^2(\bar{\gamma}_\rho + L_\rho)^2}\big(\psi(x^{k+1}) - \psi(x^*)\big) \\ &= -\sigma\big(\psi(x^{k+1}) - \psi(x^*)\big) \end{aligned}$$

for all $k \in \mathbb{N}$ sufficiently large, where we used the constant $\sigma := \frac{2\delta\gamma_{\min}}{c^2(\bar{\gamma}_\rho + L_\rho)^2}$ for brevity. Rearranging these terms yields

$$\psi(x^{k+1}) - \psi(x^*) \leq \frac{1}{1+\sigma}\big(\psi(x^k) - \psi(x^*)\big) \tag{4.22}$$

for all $k \in \mathbb{N}$ large enough, which shows that the sequence $\{\psi(x^k)\}$ converges Q-linearly to $\psi(x^*)$.

To verify statement (b), observe that the descent test (3.2) and the monotonicity of the sequence $\{\psi(x^k)\}$ yield

$$\frac{\delta\gamma_{\min}}{2}\|x^{k+1} - x^k\|^2 \leq \psi(x^k) - \psi(x^{k+1}) \leq \psi(x^k) - \psi(x^*) =: \psi_k,$$

and that the sequence $\{\psi_k\}$ is Q-linearly convergent in view of part (a). Taking this into account, it is not difficult to see that there exist constants $\omega > 0$ and $\mu \in (0,1)$ such that

$$\|x^{k+1} - x^k\| \leq \omega\mu^k$$

holds for all sufficiently large $k \in \mathbb{N}$. Hence, for given integers $\ell > k > 0$ large enough, we therefore obtain

$$\|x^{\ell+1} - x^k\| \leq \sum_{j=k}^{\ell} \|x^{j+1} - x^j\| \leq \omega \sum_{j=k}^{\ell} \mu^j \leq \omega\mu^k \sum_{j=0}^{\infty} \mu^j = \frac{\omega}{1-\mu}\mu^k.$$

Taking the limit $\ell \to \infty$ yields

$$\|x^k - x^*\| \leq \frac{\omega}{1-\mu}\mu^k$$

for all large enough $k \in \mathbb{N}$. This completes the proof of the (local) R-linear convergence of $\{x^k\}$ to its limit x^*. □

We note that similar rate-of-convergence results can be obtained for the more general case where the desingularization function is given by $\chi(t) = ct^\kappa$ for some $\kappa \in (0,1]$. The easiest way to see that is to modify the previous proof and to apply, for example, [1, Lemma 1].

5 Conclusions

In this paper, we have shown that convergence of the whole sequence generated by proximal gradient methods applied to the composite optimization problem (P) can be achieved whenever the gradient of the smooth function f is locally Lipschitz continuous while the objective function ψ possesses the KL property at all points of its domain. For our analysis, we neither needed a priori boundedness of iterates and stepsizes nor any additional convexity assumptions. Our findings also gave rise to the statement of associated rate-or-convergence results.

Several generalizations of the proximal gradient method involving, e.g., inertial terms or Bregman distances, see [5, 14–16] and the references therein, have been investigated in the presence of global Lipschitzness of the gradient associated with the smooth term, as well as the KL property. Keeping our findings in mind, it might be promising to check whether our technique of proof can be applied in these settings to weaken the appearing Lipschitz assumptions.

References

[1] F. J. Aragón Artacho, R. M. T. Fleming, and P. T. Vuong. Accelerating the DC algorithm for smooth functions. *Mathematical Programming*, 169(1):95–118, 2018. 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. *Mathematical Programming*, 116(1):5–16, 2009. 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. *Mathematics of Operations Research*, 35(2):438–457, 2010. 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. *Mathematical Programming*, 137:91 – 129, 2013. doi:10.1007/s10107-011-0484-9.

[5] H. H. Bauschke, J. Bolte, and M. Teboulle. A descent lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications. *Mathematics of Operations Research*, 42(2):330–348, 2017. doi:10.1287/moor.2016.0817.

[6] H. H. Bauschke and P. L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, 2011. doi:10.1007/978-1-4419-9467-7.

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

[8] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009. doi:10.1137/080716542.

[9] A. Beck and M. Teboulle. A fast dual proximal gradient algorithm for convex minimization and applications. *Operations Research Letters*, 42(1):1–6, 2014. doi:10.1016/j.orl.2013.10.007.

[10] W. Bian and X. Chen. Linearly constrained non-Lipschitz optimization for image restoration. *SIAM Journal on Imaging Sciences*, 8(4):2294–2322, 2015. doi:10.1137/140985639.

[11] J. Bolte, A. Daniilidis, and A. Lewis. The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal on Optimization*, 17(4):1205–1223, 2007. doi:10.1137/050644641.

[12] J. Bolte, A. Daniilidis, A. Lewis, and M. Shiota. Clarke subgradients of stratifiable functions. *SIAM Journal on Optimization*, 18(2):556–572, 2007. doi:10.1137/060670080.

[13] J. Bolte, S. Sabach, and M. Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146:459 – 494, 2014. doi:10.1007/s10107-013-0701-9.

[14] J. Bolte, S. Sabach, M. Teboulle, and Y. Vaisbourd. First order methods beyond convexity and Lipschitz gradient continuity with applications to quadratic inverse problems. *SIAM Journal on Optimization*, 28(3):2131–2151, 2018. doi:10.1137/17M1138558.

[15] R. I. Boţ and E. R. Csetnek. An inertial Tseng's type proximal algorithm for nonsmooth and nonconvex optimization problems. *Journal of Optimization Theory and Applications*, 171(2):600–616, 2016. doi:10.1007/s10957-015-0730-z.

[16] 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. *EURO Journal on Computational Optimization*, 4(1):3–25, 2016. doi:10.1007/s13675-015-0045-8.

[17] R. E. Bruck. On the weak convergence of an ergodic iteration for the solution of variational inequalities for monotone operators in Hilbert space. *Journal of Mathematical Analysis and Applications*, 61(1):159–164, 1977. doi:10.1016/0022-247X(77)90152-4.

[18] A. M. Bruckstein, D. L. Donoho, and M. Elad. From sparse solutions of systems of equations to sparse modeling of signals and images. *SIAM Review*, 51(1):34–81, 2009. doi:10.1137/060657704.

[19] R. Chartrand. Exact reconstruction of sparse signals via nonconvex minimization. *IEEE Signal Processing Letters*, 14(10):707–710, 2007. doi:10.1109/LSP.2007.898300.

[20] X. Chen, L. Guo, Z. Lu, and J. J. Ye. An augmented Lagrangian method for non-Lipschitz nonconvex programming. *SIAM Journal on Numerical Analysis*, 55(1):168–193, 2017. doi:10.1137/15M1052834.

[21] E. Cohen, N. Hallak, and M. Teboulle. Dynamic alternating direction of multipliers for nonconvex minimization with nonlinear functional equality constraints. *Journal of Optimization Theory and Applications*, 193:324–353, 2022. doi:10.1007/s10957-021-01929-5.

[22] A. De Marchi. Proximal gradient methods beyond monotony. Technical report, preprint arXiv, 2022. URL `https://arxiv.org/abs/2211.04827`.

[23] A. De Marchi, X. Jia, C. Kanzow, and P. Mehlitz. Constrained composite optimization and augmented Lagrangian methods. Technical report, preprint arXiv, 2022. URL `https://arxiv.org/abs/2203.05276`. accepted for publication in Mathematical Programming.

[24] D. Di Lorenzo, G. Liuzzi, F. Rinaldi, F. Schoen, and M. Sciandrone. A concave optimization-based approach for sparse portfolio selection. *Optimization Methods and Software*, 27(6):983–1000, 2012. doi:10.1080/10556788.2011.577773.

[25] M. Fukushima and H. Mine. A generalized proximal point algorithm for certain non-convex minimization problems. *International Journal of Systems Science*, 12(8):989–1000, 1981. doi:10.1080/00207728108963798.

[26] L. Guo and Z. Deng. A new augmented Lagrangian method for MPCCs - theoretical and numerical comparison with existing augmented Lagrangian methods. *Mathematics of Operations Research*, 47(2):1229–1246, 2022. doi:10.1287/moor.2021.1165.

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

[28] C. Kanzow and P. Mehlitz. Convergence properties of monotone and nonmonotone proximal gradient methods revisited. *Journal of Optimization Theory and Applications*, 195(2):624–646, 2022. doi:10.1007/s10957-022-02101-3.

[29] K. Kurdyka. On gradients of functions definable in o-minimal structures. *Annales de l'institut Fourier*, 48(3):769–783, 1998. doi:10.5802/aif.1638.

[30] 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. *IEEE Transactions on Signal Processing*, 63(14):3641–3656, 2015. doi:10.1109/TSP.2015.2428224.

[31] S. Łojasiewicz. Une propriété topologique des sous-ensembles analytiques réels. les Équations aux dérivées partielles. *Éditions du Centre National de la Recherche Scientifique Paris*, pages 87–89, 1963.

[32] S. Łojasiewicz. *Ensembles semi-analytiques*. Centre De Physique Theorique De L'Ecole Polytechnique, 1965.

[33] G. Marjanovic and V. Solo. On ℓ_q optimization and matrix completion. *IEEE Transactions on Signal Processing*, 60(11):5714–5724, 2012. doi:10.1109/TSP.2012.2212015.

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

[35] P. Ochs. Local convergence of the heavy-ball method and iPiano for non-convex optimization. *Journal of Optimization Theory and Applications*, 177(1):153–180, 2018. doi:10.1007/s10957-018-1272-y.

[36] P. Ochs, Y. Chen, T. Brox, and T. Pock. iPiano: Inertial proximal algorithm for nonconvex optimization. *SIAM Journal on Imaging Sciences*, 7(2):1388–1419, 2014. doi:10.1137/130942954.

[37] G. B. Passty. Ergodic convergence to a zero of the sum of monotone operators in Hilbert space. *Journal of Mathematical Analysis and Applications*, 72(2):383–390, 1979. doi:10.1016/0022-247X(79)90234-8.

[38] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970. doi:10.1515/9781400873173.

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