Journal of Optimization Theory and Applications (2022) 195:624–646
https://doi.org/10.1007/s10957-022-02101-3

Convergence Properties of Monotone and Nonmonotone Proximal Gradient Methods Revisited

Christian Kanzow[1] · **Patrick Mehlitz**[2,3]

Received: 3 December 2021 / Accepted: 22 August 2022 / Published online: 29 September 2022

Abstract
Composite optimization problems, where the sum of a smooth and a merely lower semicontinuous function has to be minimized, are often tackled numerically by means of proximal gradient methods as soon as the lower semicontinuous part of the objective function is of simple enough structure. The available convergence theory associated with these methods (mostly) requires the derivative of the smooth part of the objective function to be (globally) Lipschitz continuous, and this might be a restrictive assumption in some practically relevant scenarios. In this paper, we readdress this classical topic and provide convergence results for the classical (monotone) proximal gradient method and one of its nonmonotone extensions which are applicable in the absence of (strong) Lipschitz assumptions. This is possible since, for the price of forgoing convergence rates, we omit the use of descent-type lemmas in our analysis.

Keywords Non-Lipschitz optimization · Nonsmooth optimization · Proximal gradient method

Mathematics Subject Classification 49J52 · 90C30

Communicated by Paulo J. S. Silva.

✉ Christian Kanzow
kanzow@mathematik.uni-wuerzburg.de

Patrick Mehlitz
mehlitz@b-tu.de

1 Institute of Mathematics, University of Würzburg, 97074 Würzburg, Germany

2 Institute of Mathematics, Brandenburg University of Technology Cottbus-Senftenberg, 03046 Cottbus, Germany

3 School of Business Informatics and Mathematics, University of Mannheim, 68159 Mannheim, Germany

1 Introduction

In this paper, we address the classical problem of minimizing the sum of a smooth function f and a nonsmooth function ϕ, also known under the name *composite optimization*. This setting received much attention throughout the last years due to its inherent practical relevance in, e.g., machine learning, data compression, matrix completion, and image processing, see, e.g., [6, 13, 14, 20, 27, 28].

A standard technique for the solution of composite optimization problems is the proximal gradient method, introduced by Fukushima and Mine [21] and popularized, e.g., by Combettes and Wajs in [18]. A particular instance of this method is the celebrated iterative shrinkage/threshold algorithm (ISTA), see, e.g., [5]. A summary of existing results for the case where the nonsmooth term ϕ is defined by a convex function is given in the monograph by Beck [4].

The proximal gradient method can also be interpreted as a forward-backward splitting method, see [12, 31] for its origins and [3] for a modern view, and is able to handle problems where the nonsmooth term ϕ is given by a merely lower semicontinuous function, see, e.g., the seminal works [1, 8]. These references also provide convergence and rate-of-convergence results by using the popular *descent lemma* together with the celebrated *Kurdyka–Łojasiewicz property*.

To the best of our knowledge, however, the majority of available convergence results for proximal gradient methods assume that the smooth term f is continuously differentiable with a globally Lipschitz continuous gradient (or they require local Lipschitzness together with a bounded level set which, again, implies the global Lipschitz continuity on this level set). This requirement, which is the essential ingredient for the classical descent lemma, is often satisfied for standard applications of the proximal gradient method in data science and image processing, where f appears to be a quadratic function.

In this paper, we aim to get rid of this global Lipschitz condition. This is motivated by the fact that the algorithmic application we have in mind does not satisfy this Lipschitz property since the smooth term f corresponds to the augmented Lagrangian function of a general nonlinear constrained optimization problem, which rarely has a globally Lipschitz continuous gradient or a bounded level set. The proximal gradient method will be used to solve the resulting subproblems which forces us to generalize the convergence theory up to reasonable assumptions which are likely to hold in our framework. We refer the interested reader to [15, 19, 23, 25] where such augmented Lagrangian *proximal* methods are investigated.

Numerically, a nonmonotone version of the proximal gradient method is often preferred. Based on ideas by Grippo et al. [22] in the context of smooth unconstrained optimization problems, Birgin et al. [7] developed a nonmonotone projected gradient method for the minimization of a differentiable function over a convex set. Later, this theory was further refined in [34] where the authors present a nonmonotone proximal gradient method, known under the name SpaRSA, for composite optimization problems where the nonsmooth part ϕ of the objective function is convex (and not just an indicator function of a convex set as in [7]). The ideas from [7, 34] were subsequently generalized in the papers [15, 16] where the proximal gradient method is used as a subproblem solver within an augmented Lagrangian and penalization scheme,

respectively. However, the authors did not address the aforementioned problematic lack of Lipschitzness in these papers which causes their convergence theory to be barely applicable in their algorithmic framework. In [26, 33], the authors present nonmonotone extensions of ISTA which can handle merely lower semicontinuous terms in the objective function. Again, for the convergence analysis, global Lipschitzness of the smooth term's derivative is assumed. Due to its practical importance, we therefore aim to provide a convergence theory for the nonmonotone proximal gradient method without using any Lipschitz assumption.

In the seminal paper [2], the authors consider the composite optimization problem with both terms being convex, but without a global Lipschitz assumption for the gradient of the smooth part f. They get suitable rate-of-convergence results for the iterates generated by a Bregman-type proximal gradient method using only a local Lipschitz condition. In addition, however, they require that there is a constant $L > 0$ such that $Lh - f$ is convex, where h is a convex function which defines the Bregman distance (in our setting, h equals the squared norm). Some examples indicate that this convexity-type condition is satisfied in many practically relevant situations. Subsequently, this approach was generalized to the nonconvex setting in [9] using, once again, a local Lipschitz assumption only, as well as the slightly stronger assumption (in order to deal with the nonconvexity) that there exist $L > 0$ and a convex function h such that both $Lh - f$ and $Lh + f$ are convex. Note that the constant L plays a central role in the design of the corresponding proximal-type methods. Particularly, it is used explicitly for the choice of stepsizes. Finally, the very recent paper [17] proves global convergence results under a local Lipschitz assumption (without the additional convexity-type condition), but assumes that the iterates and stepsizes of the underlying proximal gradient method remain bounded.

To the best of our knowledge, this is the current state-of-the-art regarding the convergence properties of proximal gradient methods. The aim of this paper is slightly different, since we do not provide rate-of-convergence results, but conditions which guarantee accumulation points to be suitable stationary points of the composite optimization problem. This is the essential feature of the proximal gradient method which, for example, is exploited in [15, 19, 25] to develop augmented Lagrangian proximal methods. We also stress that, in this particular situation, the above assumption that $Lh \pm f$ is convex for some $L > 0$ is often violated unless we are dealing with linear constraints only.

Our analysis does not require a global Lipschitz assumption and is not based on the crucial descent lemma, contrasting [2, 9] mentioned above. The results show that we can get stationary accumulation points only under a local Lipschitz assumption and, depending on the properties of ϕ, sometimes even without any Lipschitz condition. In any case, a convexity-type condition like $Lh \pm f$ being convex for some constant L is not required at all. Moreover, the implementation of our proximal gradient method does not need any knowledge of the size of any Lipschitz-type constant.

Since the aim of this paper is to get a better understanding of the theoretical convergence properties of both monotone and nonmonotone proximal gradient methods, and since these methods have already been applied numerically to a large variety of problems, we do not include any numerical results in this paper.

Let us recall that we are mainly interested in conditions ensuring that accumulation points of sequences produced by the proximal gradient method are stationary. The main contributions of this paper show that this property holds (neglecting a few technical conditions) for the monotone proximal gradient method if either the smooth function f is continuously differentiable and the nonsmooth function ϕ is continuous on its domain (e.g., this assumption holds for a constrained optimization problem where ϕ corresponds to the indicator function of a nonempty and closed set), or if f is differentiable with a locally Lipschitz continuous derivative and ϕ is an arbitrary lower semicontinuous function. Corresponding statements for the nonmonotone proximal gradient method require stronger assumptions, basically the uniform continuity of the objective function on a level set. That, however, is a standard assumption in the literature dealing with nonmonotone stepsize rules.

The paper is organized as follows: In Sect. 2, we give a detailed statement of the composite optimization problem and provide some necessary background material from variational analysis. The convergence properties of the monotone and nonmonotone proximal gradient method are then discussed in Sects. 3 and 4, respectively. We close with some final remarks in Sect. 5.

2 Problem Setting and Preliminaries

We consider the *composite* optimization problem

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

where $f : \mathbb{X} \to \mathbb{R}$ is continuously differentiable, $\phi : \mathbb{X} \to \overline{\mathbb{R}} := \mathbb{R} \cup \{\infty\}$ is lower semicontinuous (possibly infinite-valued and nondifferentiable), and $\mathbb{X}$ denotes a Euclidean space, i.e., a real and finite-dimensional Hilbert space. We assume that the domain $\operatorname{dom}\phi := \{x \in \mathbb{X} \,|\, \phi(x) < \infty\}$ of ϕ is nonempty to rule out trivial situations. In order to minimize the function $\psi : \mathbb{X} \to \overline{\mathbb{R}}$ in (P), it seems reasonable to exploit the composite structure of ψ, i.e., to rely on the differentiability of f on the one hand, and on some beneficial structural properties of ϕ on the other one. This is the idea behind splitting methods.

Throughout the paper, the Euclidean space $\mathbb{X}$ will be equipped with the inner product $\langle \cdot, \cdot \rangle : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ and the associated norm $\|\cdot\|$. For some set $A \subset \mathbb{X}$ and some point $x \in \mathbb{X}$, we make use of $A + x = x + A := \{x + a \,|\, a \in A\}$ for the purpose of simplicity. For some sequence $\{x^k\} \subset \mathbb{X}$ and $x \in \mathbb{X}$, $x^k \to_\phi x$ means that $x^k \to x$ and $\phi(x^k) \to \phi(x)$. The continuous linear operator $f'(x) : \mathbb{X} \to \mathbb{R}$ denotes the derivative of f at $x \in \mathbb{X}$, and we will make use of $\nabla f(x) := f'(x)^* 1$ where $f'(x)^* : \mathbb{R} \to \mathbb{X}$ is the adjoint of $f'(x)$. This way, ∇f is a mapping from $\mathbb{X}$ to $\mathbb{X}$. Furthermore, we find $f'(x)d = \langle \nabla f(x), d \rangle$ for each $d \in \mathbb{X}$.

The following concepts are standard in variational analysis, see, e.g., [29, 32]. Let us fix some point $x \in \operatorname{dom}\phi$. Then,

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

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

$$\partial\phi(x) := \left\{\eta \in \mathbb{X} \,\middle|\, \exists\{x^k\}, \{\eta^k\} \subset \mathbb{X}\colon\ x^k \to_\phi x,\ \eta^k \to \eta,\ \eta^k \in \widehat{\partial}\phi(x^k)\ \forall k \in \mathbb{N}\right\}$$

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

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

holds in this situation. It can be seen right from the definition that whenever $x^* \in \operatorname{dom}\phi$ is a local minimizer of ϕ, then $0 \in \widehat{\partial}\phi(x^*)$, which is referred to as Fermat's rule, see [29, Proposition 1.30(i)].

Given $x \in \operatorname{dom}\phi$, the limiting subdifferential has the important robustness property

$$\left\{\eta \in \mathbb{X} \,\middle|\, \exists\{x^k\}, \{\eta^k\} \subset \mathbb{X}\colon\ x^k \to_\phi x,\ \eta^k \to \eta,\ \eta^k \in \partial\phi(x^k)\ \forall k \in \mathbb{N}\right\} \subset \partial\phi(x), \tag{1}$$

see [29, Proposition 1.20]. Clearly, the converse inclusion $\supset$ is also valid by definition of the limiting subdifferential. Note that in situations where ϕ is discontinuous at x, the requirement $x^k \to_\phi x$ in the definition of the set on the left-hand side in (1) is strictly necessary. In fact, the usual outer semicontinuity in the sense of set-valued mappings, given by

$$\left\{\eta \in \mathbb{X} \,\middle|\, \exists\{x^k\}, \{\eta^k\} \subset \mathbb{X}\colon\ x^k \to x,\ \eta^k \to \eta,\ \eta^k \in \partial\phi(x^k)\ \forall k \in \mathbb{N}\right\} \subset \partial\phi(x), \tag{2}$$

would be a much stronger condition in this situation and does not hold in general.

Whenever $x \in \operatorname{dom}\phi$ is fixed, the sum rule

$$\widehat{\partial}(f + \phi)(x) = \nabla f(x) + \widehat{\partial}\phi(x) \tag{3}$$

holds, see [29, Proposition 1.30(ii)]. Thus, due to Fermat's rule, whenever $x^* \in \operatorname{dom}\phi$ is a local minimizer of $f + \phi$, we have $0 \in \nabla f(x^*) + \widehat{\partial}\phi(x^*)$. This condition is potentially more restrictive than $0 \in \nabla f(x^*) + \partial\phi(x^*)$ which, naturally, also serves as a necessary optimality condition for (P). However, the latter is more interesting from an algorithmic point of view as it is well known from the literature on splitting methods comprising nonconvex functions ϕ. If ϕ is convex, there is no difference between those stationarity conditions.

Throughout the paper, a point $x^* \in \operatorname{dom}\phi$ satisfying $0 \in \nabla f(x^*) + \partial\phi(x^*)$ will be called a *Mordukhovich-stationary* (*M-stationary* for short) point of (P) due to the appearance of the limiting subdifferential. In the literature, the name *limiting critical point* is used as well. We close this section with two special instances of problem (P) and comment on the corresponding M-stationary conditions.

Remark 2.1 Consider the constrained optimization problem

$$\min_x \ f(x) \quad \text{subject to} \quad x \in C$$

for a continuously differentiable function $f\colon \mathbb{X} \to \mathbb{R}$ and a nonempty and closed (not necessarily convex) set $C \subset \mathbb{X}$. This problem is equivalent to the unconstrained problem (P) by setting $\phi := \delta_C$, where $\delta_C\colon \mathbb{X} \to \overline{\mathbb{R}}$ denotes the indicator function of the set C, vanishing on C and taking the value ∞ on $\mathbb{X} \setminus C$, which is lower semicontinuous due to the assumptions regarding C. The corresponding M-stationarity condition is given by

$$0 \in \nabla f(x^*) + \partial \delta_C(x^*) = \nabla f(x^*) + \mathcal{N}_C(x^*),$$

where $\mathcal{N}_C(x^*)$ denotes the *limiting* (or *Mordukhovich*) *normal cone*, see [29, Proposition 1.19].

Remark 2.2 Consider the more general constrained optimization problem

$$\min_x \ f(x) + \varphi(x) \quad \text{subject to} \quad x \in C$$

with $f\colon \mathbb{X} \to \mathbb{R}$ and $C \subset \mathbb{X}$ as in Remark 2.1, and $\varphi\colon \mathbb{X} \to \overline{\mathbb{R}}$ being another lower semicontinuous function (which might represent a regularization, penalty, or sparsity-promoting term, for example). Setting $\phi := \varphi + \delta_C$, we obtain once again an optimization problem of the form (P). The corresponding M-stationarity condition is given by

$$0 \in \nabla f(x^*) + \partial \phi(x^*) = \nabla f(x^*) + \partial(\varphi + \delta_C)(x^*).$$

Unfortunately, the sum rule

$$\partial(\varphi + \delta_C)(x^*) \subset \partial \varphi(x^*) + \partial \delta_C(x^*) = \partial \varphi(x^*) + \mathcal{N}_C(x^*)$$

does not hold in general. However, for locally Lipschitz functions φ, for example, it applies, see [29, Theorems 1.22, 2.19]. Note that the resulting stationarity condition

$$0 \in \nabla f(x^*) + \partial \varphi(x^*) + \mathcal{N}_C(x^*)$$

might be slightly weaker than M-stationarity as introduced above. Related discussions can be found in [24, Section 3].

3 Monotone Proximal Gradient Method

We first investigate a monotone version of the proximal gradient method applied to the composite optimization problem (P) with f being continuously differentiable and

ϕ being lower semicontinuous. Recall that the corresponding M-stationarity condition is given by

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

Our aim is to find, at least approximately, an M-stationary point of (P). The following algorithm is the classical proximal gradient method for this class of problems. Since we will also consider a nonmonotone variant of this algorithm in the following section, we call this the monotone proximal gradient method.

Algorithm 3.1 (*Monotone proximal gradient method*)

Require: $\tau > 1, 0 < \gamma_{\min} \leq \gamma_{\max} < \infty, \delta \in (0, 1), x^0 \in \operatorname{dom} \phi$

1: Set $k := 0$.
2: **while** A suitable termination criterion is violated at iteration k **do**
3: Choose $\gamma_k^0 \in [\gamma_{\min}, \gamma_{\max}]$.
4: For $i = 0, 1, 2, \ldots$, compute a solution $x^{k,i}$ of

$$\min_x \; f(x^k) + \langle \nabla f(x^k), x - x^k \rangle + \frac{\gamma_{k,i}}{2} \|x - x^k\|^2 + \phi(x), \quad x \in \mathbb{X} \tag{4}$$

with $\gamma_{k,i} := \tau^i \gamma_k^0$, until the acceptance criterion

$$\psi(x^{k,i}) \leq \psi(x^k) - \delta \frac{\gamma_{k,i}}{2} \|x^{k,i} - x^k\|^2 \tag{5}$$

holds.
5: Denote by $i_k := i$ the terminal value, and set $\gamma_k := \gamma_{k,i_k}$ and $x^{k+1} := x^{k,i_k}$.
6: Set $k \leftarrow k + 1$.
7: **end while**
8: **return** x^k

The convergence theory requires some technical assumptions.

Assumption 3.1 (a) The function ψ is bounded from below on $\operatorname{dom} \phi$.
(b) The function ϕ is bounded from below by an affine function.

Assumption 3.1 (a) is a reasonable condition regarding the given composite optimization problem, whereas Assumption 3.1 (b) is essentially a statement relevant for the subproblems from (4). In particular, Assumption 3.1 (b) implies that the quadratic objective function of the subproblems (4) are, for fixed $k, i \in \mathbb{N}$, coercive, and therefore always attain a solution $x^{k,i}$ (which, however, may not be unique).

The subsequent convergence theory assumes implicitly that Algorithm 3.1 generates an infinite sequence.

We first establish that the stepsize rule in Step 4 of Algorithm 3.1 is always finite.

Lemma 3.1 *Consider a fixed iteration k of Algorithm 3.1, assume that x^k is not an M-stationary point of (P), and suppose that Assumption 3.1 (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 $i_k \in \{0, 1, 2, \ldots\}$.*

Proof Suppose that the inner loop of Algorithm 3.1 does not terminate after a finite number of steps in iteration k. Recall that $x^{k,i}$ is a solution of (4). Therefore, we get

$$\langle \nabla f(x^k), x^{k,i} - x^k \rangle + \frac{\gamma_{k,i}}{2}\|x^{k,i} - x^k\|^2 + \phi(x^{k,i}) \leq \phi(x^k). \tag{6}$$

Noting that $\gamma_{k,i} \to \infty$ for $i \to \infty$ and using Assumption 3.1 (b), we obtain $x^{k,i} \to x^k$ for $i \to \infty$. Taking the limit $i \to \infty$ therefore yields

$$\phi(x^k) \leq \liminf_{i\to\infty} \phi(x^{k,i}) \leq \limsup_{i\to\infty} \phi(x^{k,i}) \leq \phi(x^k),$$

where the first estimate follows from the lower semicontinuity of ϕ and the final inequality is a consequence of (6). Therefore, we have

$$\phi(x^{k,i}) \to \phi(x^k) \quad \text{for } i \to \infty. \tag{7}$$

We claim that

$$\liminf_{i\to\infty} \gamma_{k,i}\|x^{k,i} - x^k\| > 0. \tag{8}$$

Assume, by contradiction, that there is a subsequence $i_l \to \infty$ such that

$$\liminf_{l\to\infty} \gamma_{k,i_l}\|x^{k,i_l} - x^k\| = 0. \tag{9}$$

Since x^{k,i_l} is optimal for (4), Fermat's rule and the sum rule (3) yield

$$0 \in \nabla f(x^k) + \gamma_{k,i_l}(x^{k,i_l} - x^k) + \widehat{\partial}\phi(x^{k,i_l}) \tag{10}$$

for all $l \in \mathbb{N}$. Taking the limit $l \to \infty$ while using (7) and (9), we obtain

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

which means that x^k is already an M-stationary point of (P). This contradiction shows that (8) holds. Hence, there is a constant $c > 0$ such that

$$\gamma_{k,i}\|x^{k,i} - x^k\| \geq c$$

holds for all large enough $i \in \mathbb{N}$. In particular, this implies

$$(1-\delta)\frac{\gamma_{k,i}}{2}\|x^{k,i} - x^k\|^2 \geq \frac{1-\delta}{2}c\|x^{k,i} - x^k\| \geq o\big(\|x^{k,i} - x^k\|\big) \tag{11}$$

for all sufficiently large $i \in \mathbb{N}$. Furthermore, (6) shows that

$$\langle \nabla f(x^k), x^{k,i} - x^k \rangle + \phi\big(x^{k,i}\big) - \phi(x^k) \leq -\frac{\gamma_{k,i}}{2}\|x^{k,i} - x^k\|^2. \tag{12}$$

Using a Taylor expansion of the function f and exploiting (11), (12), we obtain

$$\begin{aligned}\psi(x^{k,i}) - \psi(x^k) &= f(x^{k,i}) + \phi(x^{k,i}) - f(x^k) - \phi(x^k)\\ &= \langle \nabla f(x^k), x^{k,i} - x^k\rangle + \phi(x^{k,i}) - \phi(x^k) + o\big(\|x^{k,i} - x^k\|\big)\\ &\leq -\frac{\gamma_{k,i}}{2}\|x^{k,i} - x^k\|^2 + o\big(\|x^{k,i} - x^k\|\big)\\ &\leq -\delta\frac{\gamma_{k,i}}{2}\|x^{k,i} - x^k\|^2\end{aligned}$$

for all $i \in \mathbb{N}$ sufficiently large. This, however, means that the acceptance criterion (5) is valid for sufficiently large $i \in \mathbb{N}$, contradicting our assumption. This completes the proof. □

Let us note that the above proof actually shows that the inner loop from Step 4 of Algorithm 3.1 is either finite, or we have $\gamma_{k,i_l}\|x^{k,i_l} - x^k\| \to 0$ along a subsequence $i_l \to \infty$. Rewriting (10) by means of

$$\nabla f(x^{k,i_l}) - \nabla f(x^k) + \gamma_{k,i_l}(x^k - x^{k,i_l}) \in \nabla f(x^{k,i_l}) + \widehat{\partial}\phi(x^{k,i_l}) \tag{13}$$

and recalling that $\nabla f : \mathbb{X} \to \mathbb{X}$ is continuous motivates to also use

$$\|\nabla f(x^{k,i}) - \nabla f(x^k) + \gamma_{k,i}(x^k - x^{k,i})\| \leq \tau_{\text{abs}}$$

for some $\tau_{\text{abs}} > 0$ as a termination criterion of the inner loop since this encodes, in some sense, approximate M-stationarity of $x^{k,i}$ for (P) (note that taking the limit $l \to \infty$ in (13) would recover the limiting subdifferential of ϕ at x^k since we have $x^{k,i_l} \to_\phi x^k$ by (7)).

A critical step for the convergence theory of Algorithm 3.1 is provided by the following result.

Proposition 3.1 *Let Assumption 3.1 hold. Then, each sequence $\{x^k\}$ generated by Algorithm 3.1 satisfies $\|x^{k+1} - x^k\| \to 0$.*

Proof First recall that the sequence $\{x^k\}$ is well defined by Lemma 3.1. Using the acceptance criterion (5), we get

$$\psi(x^{k+1}) \leq \psi(x^k) - \delta\frac{\gamma_k}{2}\|x^{k+1} - x^k\|^2 \leq \psi(x^k) \tag{14}$$

for all $k \in \mathbb{N}$. Hence, the sequence $\{\psi(x^k)\}$ is monotonically decreasing. Since ψ is bounded from below on $\operatorname{dom}\phi$ by Assumption 3.1 (a) and $\{x^k\} \subset \operatorname{dom}\phi$, it follows that this sequence is convergent. Therefore, (14) implies

$$\gamma_k\|x^{k+1} - x^k\|^2 \to 0 \quad \text{for } k \to \infty.$$

Hence, the assertion follows from the fact that, by construction, we have $\gamma_k \geq \gamma_{\min} > 0$ for all $k \in \mathbb{N}$. □

A refined analysis gives the following result.

Proposition 3.2 *Let Assumption 3.1 hold, let $\{x^k\}$ be a sequence generated by Algorithm 3.1, and let $\{x^k\}_K$ be a subsequence converging to some point x^*. Then, $\gamma_k \|x^{k+1} - x^k\| \to_K 0$ holds.*

Proof If the subsequence $\{\gamma_k\}_K$ is bounded, the statement follows immediately from Proposition 3.1. The remaining part of this proof therefore assumes that this subsequence is unbounded. Without loss of generality, we may assume that $\gamma_k \to_K \infty$ and that the acceptance criterion (5) is violated in the first iteration of the inner loop for each $k \in K$. Then, for $\hat{\gamma}_k := \gamma_k/\tau$, $k \in K$, we also have $\hat{\gamma}_k \to_K \infty$, but the corresponding vector $\hat{x}^k := x^{k,i_k-1}$ does not satisfy the stepsize condition from (5), i.e., we have

$$\psi(\hat{x}^k) > \psi(x^k) - \delta \frac{\hat{\gamma}_k}{2} \|\hat{x}^k - x^k\|^2 \quad \forall k \in K. \tag{15}$$

On the other hand, since $\hat{x}^k$ solves the corresponding subproblem (4) with $\hat{\gamma}_k = \gamma_{k,i_k-1}$, we have

$$\langle \nabla f(x^k), \hat{x}^k - x^k \rangle + \frac{\hat{\gamma}_k}{2} \|\hat{x}^k - x^k\|^2 + \phi(\hat{x}^k) - \phi(x^k) \leq 0. \tag{16}$$

We claim that this, in particular, implies $\hat{x}^k \to_K x^*$. In fact, using (16), the Cauchy-Schwarz inequality, and the monotonicity of $\{\psi(x^k)\}$, 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 3.1 (b), this 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.

Now, by the mean-value theorem, there exists ξ^k on the line segment connecting x^k with $\hat{x}^k$ such that

$$\begin{aligned} \psi(\hat{x}^k) - \psi(x^k) &= f(\hat{x}^k) + \phi(\hat{x}^k) - f(x^k) - \phi(x^k) \\ &= \langle \nabla f(\xi^k), \hat{x}^k - x^k \rangle + \phi(\hat{x}^k) - \phi(x^k). \end{aligned} \tag{17}$$

Substituting the expression $\phi(\hat{x}^k) - \phi(x^k)$ from (17) into (16) yields

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

Exploiting (15), 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

$$(1 - \delta) \frac{\hat{\gamma}_k}{2} \|\hat{x}^k - x^k\| \leq \|\nabla f(x^k) - \nabla f(\xi^k)\| \tag{18}$$

(note that $\hat{x}^k \neq x^k$ in view of (15)). Since $x^k \to_K x^*$ (by assumption) and $\hat{x}^k \to_K x^*$ (by the previous part of this proof), we also get $\xi^k \to_K x^*$. Using $\delta \in (0, 1)$ and the continuous differentiability of f, it follows from (18) that $\hat{\gamma}_k \|\hat{x}^k - x^k\| \to_K 0$.

Finally, exploiting the fact that x^{k+1} and $\hat{x}^k$ are solutions of the subproblems (4) with parameters γ_k and $\hat{\gamma}_k$, respectively, we find

$$\begin{aligned} &\langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{\gamma_k}{2} \|x^{k+1} - x^k\|^2 + \phi(x^{k+1}) \\ &\quad \leq \langle \nabla f(x^k), \hat{x}^k - x^k \rangle + \frac{\gamma_k}{2} \|\hat{x}^k - x^k\|^2 + \phi(\hat{x}^k), \\ &\langle \nabla f(x^k), \hat{x}^k - x^k \rangle + \frac{\hat{\gamma}_k}{2} \|\hat{x}^k - x^k\|^2 + \phi(\hat{x}^k) \\ &\quad \leq \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{\hat{\gamma}_k}{2} \|x^{k+1} - x^k\|^2 + \phi(x^{k+1}). \end{aligned}$$

Adding these two inequalities and noting that $\gamma_k = \tau \hat{\gamma}_k > \hat{\gamma}_k$ yields $\|x^{k+1} - x^k\| \leq \|\hat{x}^k - x^k\|$ and, therefore,

$$\gamma_k \|x^{k+1} - x^k\| = \tau \hat{\gamma}_k \|x^{k+1} - x^k\| \leq \tau \hat{\gamma}_k \|\hat{x}^k - x^k\| \to_K 0.$$

This completes the proof. □

The above technique of proof implies a boundedness result for the sequence $\{\gamma_k\}_K$ if ∇f satisfies a local Lipschitz property around the associated accumulation point of iterates. This observation is stated explicitly in the following result.

Corollary 3.1 *Let Assumption 3.1 hold, let $\{x^k\}$ be a sequence generated by Algorithm 3.1, let $\{x^k\}_K$ be a subsequence converging to some point x^*, and assume that $\nabla f : \mathbb{X} \to \mathbb{X}$ is locally Lipschitz continuous around x^*. Then, the corresponding subsequence $\{\gamma_k\}_K$ is bounded.*

Proof We may argue as in the proof of Proposition 3.2. Hence, on the contrary, assume that $\gamma_k \to_K \infty$. For each $k \in K$, define $\hat{\gamma}_k$ and $\hat{x}^k$ as in that proof, and let $L > 0$ denote the local Lipschitz constant of ∇f around x^*. Recall that $x^k \to_K x^*$ (by assumption) and $\hat{x}^k \to_K x^*$ (from the proof of Proposition 3.2). Exploiting (18), we therefore obtain

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

for all $k \in K$ sufficiently large, using the fact that ξ^k is on the line segment between x^k and $\hat{x}^k$. Since $\hat{\gamma}_k \to_K \infty$ and $\hat{x}^k \neq x^k$, see once again (15), this gives a contradiction. Hence, $\{\gamma_k\}_K$ stays bounded. □

The following is the main convergence result for Algorithm 3.1 which requires a slightly stronger smoothness assumption on either f or ϕ.

Theorem 3.1 *Assume that Assumption 3.1 holds, while either ϕ is continuous on* dom ϕ *or $\nabla f : \mathbb{X} \to \mathbb{X}$ is locally Lipschitz continuous. Then, each accumulation point x^* of a sequence $\{x^k\}$ generated by Algorithm 3.1 is an M-stationary point of (P).*

Proof Let $\{x^k\}_K$ be a subsequence converging to x^*. In view of Proposition 3.1, it follows that also the subsequence $\{x^{k+1}\}_K$ converges to x^*. Furthermore, Proposition 3.2 yields $\gamma_k\|x^{k+1} - x^k\| \to_K 0$. The minimizing property of x^{k+1}, Fermat's rule, and the sum rule (3) imply that

$$0 \in \nabla f(x^k) + \gamma_k(x^{k+1} - x^k) + \widehat{\partial}\phi(x^{k+1}) \tag{19}$$

holds for each $k \in K$. Hence, if we can show $\phi(x^{k+1}) \to_K \phi(x^*)$, we can take the limit $k \to_K \infty$ in (19) to obtain the desired statement $0 \in \nabla f(x^*) + \partial\phi(x^*)$.

Due to (14), we find $\psi(x^{k+1}) \leq \psi(x^0)$ for each $k \in K$. Taking the limit $k \to_K \infty$ while respecting the lower semicontinuity of ϕ gives $\psi(x^*) \leq \psi(x^0)$, and due to $x^0 \in \text{dom}\,\phi$, we find $x^* \in \text{dom}\,\phi$. Thus, the condition $\phi(x^{k+1}) \to_K \phi(x^*)$ obviously holds if ϕ is continuous on its domain since all iterates x^k generated by Algorithm 3.1 as well as x^* belong to dom ϕ.

Hence, it remains to consider the situation where ϕ is only lower semicontinuous, but ∇f is locally Lipschitz continuous. From $x^{k+1} \to_K x^*$ and the lower semicontinuity of ϕ, we find

$$\phi(x^*) \leq \liminf_{k\in K} \phi(x^{k+1}) \leq \limsup_{k\in K} \phi(x^{k+1}).$$

It therefore suffices to show that $\limsup_{k\in K} \phi(x^{k+1}) \leq \phi(x^*)$ holds. Since x^{k+1} solves the subproblem (4) with parameter γ_k, we obtain

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

for each $k \in K$. We now take the upper limit over K on both sides. Using the continuity of ∇f, the convergences $x^{k+1} - x^k \to_K 0$ as well as $\gamma_k \|x^{k+1} - x^k\|^2 \to_K 0$ (see Propositions 3.1 and 3.2), and taking into account that $\gamma_k \|x^k - x^*\|^2 \to_K 0$ due to the boundedness of the subsequence $\{\gamma_k\}_K$ in this situation, see Corollary 3.1, we obtain $\limsup_{k \in K} \phi(x^{k+1}) \leq \phi(x^*)$. Altogether, we therefore get $\phi(x^{k+1}) \to_K \phi(x^*)$, and this completes the proof. □

Note that ϕ being continuous on $\operatorname{dom} \phi$ is an assumption which holds, e.g., if ϕ is the indicator function of a closed set, see Remark 2.1. Therefore, Theorem 3.1 provides a global convergence result for constrained optimization problems with an arbitrary continuously differentiable objective function over any closed (not necessarily convex) feasible set. Moreover, the previous convergence result also holds for a general lower semicontinuous function ϕ provided that ∇f is locally Lipschitz continuous. This includes, for example, sparse optimization problems in $\mathbb{X} \in \{\mathbb{R}^n, \mathbb{R}^{n \times m}\}$ involving the so-called ℓ_0-quasi-norm, which counts the number of nonzero entries of the input vector, as a penalty term or optimization problems in $\mathbb{X} := \mathbb{R}^{n \times m}$ comprising rank penalties. Note that we still do not require the global Lipschitz continuity of ∇f. However, it is an open question whether the previous convergence result also holds for the general setting where f is only continuously differentiable and ϕ is just lower semicontinuous.

Remark 3.1 Let $\{x^k\}$ be a sequence generated by Algorithm 3.1. In iteration $k \in \mathbb{N}$, x^{k+1} satisfies the necessary optimality condition (19) of the subproblem (4). Hence, from the next iteration's point of view, we obtain

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

for each $k \in \mathbb{N}$ with $k \geq 1$. This justifies evaluation of the termination criterion

$$\left\| \gamma_{k-1}(x^{k-1} - x^k) + \nabla f(x^k) - \nabla f(x^{k-1}) \right\| \leq \tau_{\text{abs}} \tag{20}$$

for some $\tau_{\text{abs}} > 0$ since this means that x^k is, in some sense, approximately M-stationary for (P). Observe that, along a subsequence $\{x^k\}_K$ satisfying $x^{k-1} \to_K x^*$ for some x^*, Propositions 3.1 and 3.2 yield $x^k \to_K x^*$ and $\gamma_{k-1}(x^k - x^{k-1}) \to_K 0$ under appropriate assumptions, which means that (20) is satisfied for large enough $k \in K$ due to continuity of $\nabla f : \mathbb{X} \to \mathbb{X}$, see the discussion after Lemma 3.1 as well.

Recall that the existence of accumulation points is guaranteed by the coercivity of the function ψ. A simple criterion for the convergence of the entire sequence $\{x^k\}$ is provided by the following comment.

Remark 3.2 Let $\{x^k\}$ be any sequence generated by Algorithm 3.1 such that x^* is an isolated accumulation point of this sequence. Then, the entire sequence converges to x^*. This follows immediately from [30, Lemma 4.10] and the property of the proximal gradient method stated in Proposition 3.1. The accumulation point x^* is isolated, in particular, if f is twice continuously differentiable with $\nabla^2 f(x^*)$ being positive

definite and ϕ is convex. In this situation, x^* is a strict local minimum of ψ and therefore the only stationary point of ψ is a neighborhood of x^*. Since, by Theorem 3.1, every accumulation point is stationary, it follows that x^* is necessarily an isolated stationary point in this situation and, thus, convergence of the whole sequence $\{x^k\}$ to x^* follows.

4 Nonmonotone Proximal Gradient Method

The method to be presented here is a nonmonotone version of the proximal gradient method from the previous section. The kind of nonmonotonicity used here was introduced by Grippo et al. [22] for a class of smooth unconstrained optimization problems and then discussed, in the framework of composite optimization problems, by Wright et al. [34] as well as in some subsequent papers. We first state the precise algorithm and investigate its convergence properties. The relation to the existing convergence results is postponed until the end of this section.

Algorithm 4.1 (*Nonmonotone proximal gradient method*)

Require: $\tau > 0$, $0 < \gamma_{\min} \leq \gamma_{\max} < \infty$, $m \in \mathbb{N}$, $\delta \in (0, 1)$, $x^0 \in \operatorname{dom} \phi$

1: Set $k := 0$.
2: **while** A suitable termination criterion is violated at iteration k **do**
3: Set $m_k := \min\{k, m\}$ and choose $\gamma_k^0 \in [\gamma_{\min}, \gamma_{\max}]$.
4: For $i = 0, 1, 2, \ldots$, compute a solution $x^{k,i}$ of

$$\min_x \; f(x^k) + \langle \nabla f(x^k), x - x^k \rangle + \frac{\gamma_{k,i}}{2} \|x - x^k\|^2 + \phi(x), \quad x \in \mathbb{X} \tag{21}$$

with $\gamma_{k,i} := \tau^i \gamma_k^0$, until the acceptance criterion

$$\psi(x^{k,i}) \leq \max_{j=0,1,\ldots,m_k} \psi(x^{k-j}) - \delta \frac{\gamma_{k,i}}{2} \|x^{k,i} - x^k\|^2 \tag{22}$$

holds.
5: Denote by $i_k := i$ the terminal value, and set $\gamma_k := \gamma_{k,i_k}$ and $x^{k+1} := x^{k,i_k}$.
6: Set $k \leftarrow k + 1$.
7: **end while**
8: **return** x^k

The only difference between Algorithms 3.1 and 4.1 is in the stepsize rule. More precisely, Algorithm 4.1 may be viewed as a generalization of Algorithm 3.1 since the particular choice $m = 0$ recovers Algorithm 3.1. Numerically, in many examples, the choice $m > 0$ leads to better results and is therefore preferred in practice. On the other hand, for $m > 0$, we usually get a nonmonotone behavior of the function values $\{\psi(x^k)\}$ which complicates the theory significantly. In addition, the nonmontone proximal gradient method also requires stronger assumptions in order to prove a suitable convergence result.

In particular, in addition to the requirements from Assumption 3.1, we need the following additional conditions on the data functions in order to proceed.

Assumption 4.1 (a) The function ψ is uniformly continuous on the sublevel set $\mathcal{L}_\psi(x^0) := \{x \in \mathbb{X} \mid \psi(x) \leq \psi(x^0)\}$.
(b) The function ϕ is continuous on $\operatorname{dom}\phi$.

Note that we always have $\mathcal{L}_\psi(x^0) \subset \operatorname{dom}\phi$ by the continuity of f. Furthermore, whenever ψ is coercive, Assumption 4.1 (b) already implies Assumption 4.1 (a) since $\mathcal{L}_\psi(x^0)$ would be a compact subset of $\operatorname{dom}\phi$ in this situation, and continuous functions are uniformly continuous on compact sets. Observe that coercivity of ψ is an inherent property in many practically relevant settings. We further note that, in general, Assumption 4.1 (a) does not imply Assumption 4.1 (b), and the latter is a necessary requirement since, in our convergence theory, we will also evaluate the function ϕ in some points resulting from an auxiliary sequence $\{\hat{x}^k\}$ which may not belong to the level set $\mathcal{L}_\psi(x^0)$.

For the convergence theory, we assume implicitly that Algorithm 4.1 generates an infinite sequence $\{x^k\}$. We first note that the stepsize rule in the inner loop of Algorithm 4.1 is always finite. Since

$$\psi(x^k) \leq \max_{j=0,1,\ldots,m_k} \psi(x^{k-j})$$

this observation follows immediately from Lemma 3.1.

Throughout the section, for each $k \in \mathbb{N}$, let $l(k) \in \{k - m_k, \ldots, k\}$ be an index such that

$$\psi(x^{l(k)}) = \max_{j=0,1,\ldots,m_k} \psi(x^{k-j})$$

is valid. We already mentioned that $\{\psi(x^k)\}$ may possess a nonmonotone behavior. However, as the following lemma shows, $\{\psi(x^{l(k)})\}$ is monotonically decreasing.

Lemma 4.1 *Let Assumption 3.1 (b) hold and let $\{x^k\}$ be a sequence generated by Algorithm 4.1. Then $\{\psi(x^{l(k)})\}$ is monotonically decreasing.*

Proof The nonmonotone stepsize rule from (22) can be rewritten as

$$\psi(x^{k+1}) \leq \psi(x^{l(k)}) - \delta\frac{\gamma_k}{2}\|x^{k+1} - x^k\|^2. \tag{23}$$

Using $m_{k+1} \leq m_k + 1$, we find

$$\begin{aligned}
\psi(x^{l(k+1)}) &= \max_{j=0,1,\ldots,m_{k+1}} \psi(x^{k+1-j}) \\
&\leq \max_{j=0,1,\ldots,m_k+1} \psi(x^{k+1-j}) \\
&= \max\left\{\max_{j=0,1,\ldots,m_k} \psi(x^{k-j}), \psi(x^{k+1})\right\} \\
&= \max\left\{\psi(x^{l(k)}), \psi(x^{k+1})\right\}
\end{aligned}$$

$$= \psi(x^{l(k)}),$$

where the last equality follows from (23). This shows the claim. □

As a corollary of the above result, we obtain that the iterates of Algorithm 4.1 belong to the level set $\mathcal{L}_\psi(x^0)$.

Corollary 4.1 *Let Assumption 3.1 (b) hold and let $\{x^k\}$ be a sequence generated by Algorithm 4.1. Then $\{x^k\}, \{x^{l(k)}\} \subset \mathcal{L}_\psi(x^0)$ holds.*

Proof Noting that $l(0) = 0$ holds by construction, Lemma 4.1 and (23) yield the estimate $\psi(x^{k+1}) \leq \psi(x^{l(k)}) \leq \psi(x^{l(0)}) = \psi(x^0)$ for each $k \in \mathbb{N}$ which shows the claim. □

The counterpart of Proposition 3.1 is significantly more difficult to prove in the nonmonotone setting. In fact, it is this central result which requires the uniform continuity of the objective function ψ from Assumption 4.1 (a). Though its proof is essentially the one from [34], we present all details since they turn out to be of some importance for the discussion at the end of this section.

Proposition 4.1 *Let Assumption 3.1 and Assumption 4.1 (a) hold. Then, each sequence $\{x^k\}$ generated by Algorithm 4.1 satisfies $\|x^{k+1} - x^k\| \to 0$.*

Proof Since ψ is bounded from below due to Assumption 3.1 (a), Lemma 4.1 implies

$$\lim_{k\to\infty} \psi(x^{l(k)}) = \psi^* \tag{24}$$

for some finite $\psi^* \in \mathbb{R}$. From Corollary 4.1, we find $\{x^{l(k)}\} \subset \mathcal{L}_\psi(x^0)$. Applying (23) with k replaced by $l(k) - n - 1$ for some $n \in \mathbb{N}$ gives $\psi(x^{l(k)-n}) \leq \psi(x^{l(l(k)-n-1)}) \leq \psi(x^0)$, i.e., $\{x^{l(k)-n}\} \subset \mathcal{L}_\psi(x^0)$ (here, we assume implicitly that k is large enough such that no negative indices $l(k) - n - 1$ occur). More precisely, for $n = 0$, we have

$$\psi(x^{l(k)}) - \psi(x^{l(l(k)-1)}) \leq -\delta \frac{\gamma_{l(k)-1}}{2} \|x^{l(k)} - x^{l(k)-1}\|^2 \leq 0.$$

Taking the limit $k \to \infty$ in the previous inequality and using (24), we therefore obtain

$$\lim_{k\to\infty} \gamma_{l(k)-1} \|x^{l(k)} - x^{l(k)-1}\|^2 = 0.$$

Since $\gamma_k \geq \gamma_{\min} > 0$ for all $k \in \mathbb{N}$, we get

$$\lim_{k\to\infty} d^{l(k)-1} = 0, \tag{25}$$

where $d^k := x^{k+1} - x^k$ for all $k \in \mathbb{N}$. Using (24) and (25), it follows that

$$\psi^* = \lim_{k\to\infty} \psi(x^{l(k)}) = \lim_{k\to\infty} \psi\left(x^{l(k)-1} + d^{l(k)-1}\right) = \lim_{k\to\infty} \psi(x^{l(k)-1}), \tag{26}$$

where the last equality takes into account the uniform continuity of ψ from Assumption 4.1 (a) and (25).

We will now prove, by induction, that the limits

$$\lim_{k\to\infty} d^{l(k)-j} = 0, \qquad \lim_{k\to\infty} \psi(x^{l(k)-j}) = \psi^* \tag{27}$$

hold for all $j \in \mathbb{N}$ with $j \geq 1$. We already know from (25) and (26) that (27) holds for $j = 1$. Suppose that (27) holds for some $j \geq 1$. We need to show that it holds for $j + 1$. Using (23) with k replaced by $l(k) - j - 1$, we have

$$\psi(x^{l(k)-j}) \leq \psi(x^{l(l(k)-j-1)}) - \delta \frac{\gamma_{l(k)-j-1}}{2} \|d^{l(k)-j-1}\|^2$$

(again, we assume implicitly that k is large enough such that $l(k)-j-1$ is nonnegative). Rearranging this expression and using $\gamma_k \geq \gamma_{\min}$ for all k yields

$$\|d^{l(k)-j-1}\|^2 \leq \frac{2}{\gamma_{\min}\delta}\big(\psi(x^{l(l(k)-j-1)}) - \psi(x^{l(k)-j})\big).$$

Taking $k \to \infty$, using (24), as well as the induction hypothesis, it follows that

$$\lim_{k\to\infty} d^{l(k)-j-1} = 0, \tag{28}$$

which proves the induction step for the first limit in (27). The second limit then follows from

$$\lim_{k\to\infty} \psi\big(x^{l(k)-(j+1)}\big) = \lim_{k\to\infty} \psi\big(x^{l(k)-(j+1)} + d^{l(k)-j-1}\big) = \lim_{k\to\infty} \psi\big(x^{l(k)-j}\big) = \psi^*,$$

where the first equation exploits (28) together with the uniform continuity of ψ from Assumption 4.1 (a) and $\{x^{l(k)-j}\}, \{x^{l(k)-(j+1)}\} \subset \mathcal{L}_\psi(x^0)$, whereas the final equation is the induction hypothesis.

In the last step of our proof, we now show that $\lim_{k\to\infty} d^k = 0$ holds. Suppose that this is not true. Then there is a (suitably shifted, for notational simplicity) subsequence $\{d^{k-m-1}\}_{k\in K}$ and a constant $c > 0$ such that

$$\|d^{k-m-1}\| \geq c \quad \forall k \in K. \tag{29}$$

Now, for each $k \in K$, the corresponding index $l(k)$ is one of the indices $k-m, k-m+1, \ldots, k$. Hence, we can write $k-m-1 = l(k)-j_k$ for some index $j_k \in \{1, 2, \ldots, m+1\}$. Since there are only finitely many possible indices j_k, we may assume without loss of generality that $j_k = j$ holds for some fixed index $j \in \{1, \ldots, m+1\}$. Then (27) implies

$$\lim_{k\to_K\infty} d^{k-m-1} = \lim_{k\to_K\infty} d^{l(k)-j} = 0.$$

This contradicts (29) and therefore completes the proof. □

Theorem 4.1 *Assume that Assumptions 3.1 and 4.1 hold and let $\{x^k\}$ be a sequence generated by Algorithm 4.1. Suppose that x^* is an accumulation point of $\{x^k\}$ such that $x^k \to_K x^*$ holds along a subsequence $k \to_K \infty$. Then, x^* is an M-stationary point of (P), and $\gamma_k(x^{k+1} - x^k) \to_K 0$ is valid.*

Proof Since $\{x^k\}_K$ is a subsequence converging to x^*, it follows from Proposition 4.1 that also the subsequence $\{x^{k+1}\}_K$ converges to x^*. We note that $x^* \in \operatorname{dom} \phi$ follows from Corollary 4.1 by closedness of $\mathcal{L}_\psi(x^0)$. The minimizing property of x^{k+1} for (21) together with Fermat's rule and the sum rule from (3) imply that the necessary optimality condition (19) holds for each $k \in K$. We claim that the subsequence $\{\gamma_k\}_K$ is bounded. Assume, by contradiction, that this is not true. Without loss of generality, let us assume that $\gamma_k \to_K \infty$ and that the acceptance criterion (22) is violated in the first iteration of the inner loop for each $k \in K$. Setting $\hat{\gamma}_k := \gamma_k/\tau$ for each $k \in K$, $\{\hat{\gamma}^k\}_K$ also tends to infinity, but the corresponding vectors $\hat{x}^k := x^{k,i_k-1}$, $k \in K$, do not satisfy the stepsize condition from (22), i.e., we have

$$\psi(\hat{x}^k) > \max_{j=0,1,\ldots,m_k} \psi(x^{k-j}) - \delta \frac{\hat{\gamma}_k}{2} \|\hat{x}^k - x^k\|^2 \qquad \forall k \in K. \tag{30}$$

On the other hand, since $\hat{x}^k = x^{k,i_k-1}$ solves the corresponding subproblem (4) with $\hat{\gamma}_k = \gamma_{k,i_k-1}$, we have

$$\langle \nabla f(x^k), \hat{x}^k - x^k \rangle + \frac{\hat{\gamma}_k}{2} \|\hat{x}^k - x^k\|^2 + \phi(\hat{x}^k) \le \phi(x^k) \tag{31}$$

for each $k \in K$. Due to $\hat{\gamma}_k \to_K \infty$ and since ϕ is bounded from below by an affine function due to Assumption 3.1 (b) while ϕ is continuous on its domain by Assumption 4.1 (b) (which yields boundedness of the right-hand side of (31)), this implies $\hat{x}^k - x^k \to_K 0$. Consequently, we have $\hat{x}^k \to_K x^*$ as well.

Now, if $\hat{\gamma}_k \|\hat{x}^k - x^k\| \to_{K'} 0$ holds along a subsequence $k \to_{K'} \infty$ such that $K' \subset K$, then, due to

$$0 \in \nabla f(x^k) + \hat{\gamma}_k(\hat{x}^k - x^k) + \widehat{\partial}\phi(\hat{x}^k),$$

which holds for each $k \in K'$ by means of Fermat's rule and the sum rule (3), we immediately see that x^* is an M-stationary point of (P) by taking the limit $k \to_{K'} \infty$ and exploiting the continuity of ϕ on $\operatorname{dom} \phi$ from Assumption 4.1 (b). Thus, for the remainder of the proof, we may assume that there is a constant $c > 0$ such that

$$\hat{\gamma}_k \|\hat{x}^k - x^k\| \ge c$$

holds for each $k \in K$. Further, we then also get

$$(1-\delta)\frac{\hat{\gamma}_k}{2}\|\hat{x}^k - x^k\|^2 \ge \frac{1-\delta}{2} c \|\hat{x}^k - x^k\| \ge o\big(\|\hat{x}^k - x^k\|\big)$$

for all $k \in K$ sufficiently large. Rearranging (31) gives us

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

for each $k \in K$. From the mean-value theorem, we obtain some ξ^k on the line segment between $\hat{x}^k$ and x^k such that

$$\begin{aligned}
&\psi(\hat{x}^k) - \max_{j=0,1,\ldots,m_k} \psi(x^{k-j}) \\
&\quad \leq \psi(\hat{x}^k) - \psi(x^k) \\
&\quad = \langle \nabla f(\xi^k), \hat{x}^k - x^k \rangle + \phi(\hat{x}^k) - \phi(x^k) \\
&\quad = \langle \nabla f(x^k), \hat{x}^k - x^k \rangle + \phi(\hat{x}^k) - \phi(x^k) + \langle \nabla f(\xi^k) - \nabla f(x^k), \hat{x}^k - x^k \rangle \\
&\quad \leq -\frac{\hat{\gamma}^k}{2} \|\hat{x}^k - x^k\|^2 + o(\|\hat{x}^k - x^k\|) \\
&\quad \leq -\delta \frac{\hat{\gamma}^k}{2} \|\hat{x}^k - x^k\|^2
\end{aligned}$$

for all $k \in K$ sufficiently large. This contradiction to (30) shows that the sequence $\{\gamma_k\}_K$ is bounded.

Finally, the continuity of ϕ from Assumption 4.1 (b) gives $\phi(x^{k+1}) \to_K \phi(x^*)$ due to $x^{k+1} \to_K x^*$. Thus, recalling $x^k \to_K x^*$ and the boundedness of $\{\gamma_k\}_K$, we find $\gamma_k(x^{k+1} - x^k) \to_K 0$, and taking the limit $k \to_K \infty$ in (19) gives us M-stationarity of x^* for (P). □

Remark 4.1 (a) Note that Assumptions 3.1 and 4.1 do not comprise any Lipschitz conditions on ∇f.

(b) The results in this section recover the findings from [23, Section 4] and [25, Section 3] which were obtained in the special situation where ϕ is the indicator function associated with a closed set, see Remark 2.1 as well.

(c) Based on Theorem 4.1, (20) also provides a reasonable termination criterion for Algorithm 4.1, see Remark 3.1 as well.

(d) In view of Proposition 4.1, it follows in the same way as in Remark 3.2 that the entire sequence $\{x^k\}$ generated by Algorithm 4.1 converges if there exists an isolated accumulation point.

The uniform continuity of ψ which is demanded in Assumption 4.1 (a) is obviously a much stronger assumption than the one used in the previous section for the monotone proximal gradient method. In particular, this assumption rules out applications where ϕ is given by the ℓ_0-quasi-norm. Nevertheless, the theory still covers the situation where the role of ϕ is played by an ℓ_p-type penalty function for $p \in (0, 1)$ over $\mathbb{X} \in \{\mathbb{R}^n, \mathbb{R}^{n \times m}\}$ which is known to promote sparse solutions. More precisely, this choice is popular in sparse optimization if the more common ℓ_1-norm does not provide satisfactory sparsity results, and the application of the ℓ_0-quasi-norm seems too difficult, see [6, 14, 15, 19, 27, 28] for some applications and numerical results

based on the ℓ_p-quasi-norm or closely related expressions. We would like to note that uniform continuity is a standard assumption in the context of nonmonotone stepsize rules involving acceptance criteria of type (22), see [22, page 710].

We close this section with a discussion on existing convergence results for nonmonotone proximal gradient methods. To the best of our knowledge, the first one can be found in [34]. The authors prove convergence under the assumptions that f is differentiable with a globally Lipschitz continuous gradient and ϕ being real-valued and convex, see [34, Section II.G]. Implicitly, however, they also exploit the uniform continuity of $\psi = f + \phi$ in their proof of [34, Lemma 4], a result like Proposition 4.1, without stating this assumption explicitly. Taking this into account, our Assumption 4.1 (a) is actually weaker than the requirements used in [34], so that the results of this section can be viewed as a generalization of the convergence theory from [34].

Furthermore, [15, Section 3.1] and [16, Appendix A] consider a nonmonotone proximal gradient method which is slightly different from Algorithm 4.1 since the acceptance criterion (22) is replaced by the slightly simpler condition

$$\psi(x^{k,i}) \leq \max_{j=0,1,\ldots,m_k} \psi(x^{k-j}) - \frac{\delta}{2}\|x^{k,i} - x^k\|^2.$$

In [16, Theorem 4.1], the authors obtain convergence to M-stationary points whenever ψ is bounded from below as well as uniformly continuous on the level set $\mathcal{L}_\psi(x^0)$, f possesses a Lipschitzian derivative on some enlargement of $\mathcal{L}_\psi(x^0)$, and ϕ is continuous. Clearly, our convergence analysis of Algorithm 4.1 does not exploit any Lipschitzianity of ∇f, so our assumptions are weaker than those ones used in [16]. In [15, Theorem 3.3], the authors claim that the results from [16] even hold when the continuity assumption on ϕ is dropped. The proof of [15, Theorem 3.3], however, relies on the outer semicontinuity property (2) of the limiting subdifferential, which does not hold for general discontinuous functions ϕ, so this result is not reliable.

Finally, let us mention that the two references [26, 33] also consider nonmonotone (and accelerated) proximal gradient methods. These methods are not directly comparable to our algorithm since they are based on a different kind of nonmonotonicity. In any case, although the analysis in both papers works for merely lower semicontinuous functions ϕ, the provided convergence theory requires ∇f to be globally Lipschitz continuous.

5 Conclusions

In this paper, we demonstrated how the convergence analysis for monotone and nonmonotone proximal gradient methods can be carried out in the absence of (global) Lipschitz continuity of the derivative associated with the smooth function. Our results, thus, open up these algorithms to be reasonable candidates for subproblem solvers within an augmented Lagrangian framework for the numerical treatment of constrained optimization problems with lower semicontinuous objective functions, see, e.g., [15]

where this approach has been suggested but suffers from an incomplete analysis, and [19, 23, 25] where this approach has been corrected and extended.

Let us mention some remaining open problems regarding the investigated proximal gradient methods. First, it might be interesting to find minimum requirements which ensure global convergence of Algorithms 3.1 and 4.1. We already mentioned in Sect. 3 that it is an open question whether the convergence analysis for Algorithm 3.1 can be generalized to the setting where f is only continuously differentiable while ϕ is just lower semicontinuous. Second, we did not investigate if the *Kurdyka–Łojasiewicz* property could be efficiently incorporated into the convergence analysis in order to get stronger results even in the absence of strong Lipschitz assumptions on the derivative of f. Third, our analysis has shown that Algorithms 3.1 and 4.1 compute M-stationary points of (P) in general. In the setting of Remark 2.2, i.e., where constrained programs with a merely lower semicontinuous objective function are considered, the introduced concept of M-stationarity is, to some extent, *implicit* since it comprises an unknown subdifferential. In general, the latter can be approximated from above in terms of initial problem data only in situations where a qualification condition is valid. The resulting stationarity condition may be referred to as *explicit* M-stationarity. It seems to be a relevant topic of future research to investigate whether Algorithms 3.1 and 4.1 can be modified such that they compute explicitly M-stationary points in this rather general setting. Fourth, it might be interesting to investigate whether other types of nonmonotonicity, different from the one used in Algorithm 4.1, can be exploited in order to get rid of the uniform continuity requirement from Assumption 4.1 (a).

Finally, we note that there exist several generalizations of proximal gradient methods using, e.g., inertial terms and Bregman distances, see, e.g., [2, 9–11] and the references therein. The corresponding convergence theory is also based on a global Lipschitz assumption for the gradient of the smooth term or additional convexity assumptions which allow the application of a descent-type lemma. It might be interesting to see whether our technique of proof can be adapted to these generalized proximal gradient methods in order to weaken the postulated assumptions.

Acknowledgements We thank the two anonymous reviewers for the suggestion of several interesting, subject-related references.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data availability Data sharing is not applicable to this article as no data sets were generated or analyzed during the current study.

Declarations

Conflict of interest No conflict of interest has been reported by the authors.

References

1. Attouch, H., Bolte, J., Svaiter, B.F.: Convergence of descent methods for semi-algebraic and tame problems, proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods. Math. Program. **137**, 91–129 (2013). https://doi.org/10.1007/s10107-011-0484-9
2. Bauschke, H.H., Bolte, J., Teboulle, M.: A descent lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications. Math. Oper. Res. **42**(2), 330–348 (2017). https://doi.org/10.1287/moor.2016.0817
3. Bauschke, H.H., Combettes, P.L.: Convex Analysis and Monotone Operator Theory in Hilbert Spaces. Springer, Berlin (2011). https://doi.org/10.1007/978-1-4419-9467-7
4. Beck, A.: First-Order Methods in Optimization. SIAM (2017). https://doi.org/10.1137/1.9781611974997
5. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imag. Sci. **2**(1), 183–202 (2009). https://doi.org/10.1137/080716542
6. Bian, W., Chen, X.: Linearly constrained non-Lipschitz optimization for image restoration. SIAM J. Imag. Sci. **8**(4), 2294–2322 (2015). https://doi.org/10.1137/140985639
7. Birgin, E.G., Martínez, J.M., Raydan, M.: Nonmonotone spectral projected gradient methods on convex sets. SIAM J. Optim. **10**(4), 1196–1211 (2000). https://doi.org/10.1137/s1052623497330963
8. Bolte, J., Sabach, S., Teboulle, M.: Proximal alternating linearized minimization for nonconvex and nonsmooth problems. Math. Program. **146**, 459–494 (2014). https://doi.org/10.1007/s10107-013-0701-9
9. Bolte, J., Sabach, S., Teboulle, M., Vaisbourd, Y.: First order methods beyond convexity and Lipschitz gradient continuity with applications to quadratic inverse problems. SIAM J. Optim. **28**(3), 2131–2151 (2018). https://doi.org/10.1137/17M1138558
10. Boţ, R.I., Csetnek, E.R.: An inertial Tseng's type proximal algorithm for nonsmooth and nonconvex optimization problems. J. Optim. Theory Appl. **171**(2), 600–616 (2016). https://doi.org/10.1007/s10957-015-0730-z
11. Boţ, R.I., Csetnek, E.R., László, S.C.: An inertial forward-backward algorithm for the minimization of the sum of two nonconvex functions. EURO J. Comput. Optim. **4**(1), 3–25 (2016). https://doi.org/10.1007/s13675-015-0045-8
12. Bruck, R.E.: On the weak convergence of an ergodic iteration for the solution of variational inequalities for monotone operators in Hilbert space. J. Math. Anal. Appl. **61**(1), 159–164 (1977). https://doi.org/10.1016/0022-247X(77)90152-4
13. Bruckstein, A.M., Donoho, D.L., Elad, M.: From sparse solutions of systems of equations to sparse modeling of signals and images. SIAM Rev. **51**(1), 34–81 (2009). https://doi.org/10.1137/060657704
14. Chartrand, R.: Exact reconstruction of sparse signals via nonconvex minimization. IEEE Signal Process. Lett. **14**(10), 707–710 (2007). https://doi.org/10.1109/LSP.2007.898300
15. Chen, X., Guo, L., Lu, Z., Ye, J.J.: An augmented Lagrangian method for non-Lipschitz nonconvex programming. SIAM J. Numer. Anal. **55**(1), 168–193 (2017). https://doi.org/10.1137/15M1052834
16. Chen, X., Lu, Z., Pong, T.-K.: Penalty methods for a class of non-Lipschitz optimization problems. SIAM J. Optim. **26**(3), 1465–1492 (2016). https://doi.org/10.1137/15M1028054
17. Cohen, E., Hallak, N., Teboulle, M.: Dynamic alternating direction of multipliers for nonconvex minimization with nonlinear functional equality constraints. J. Optim. Theory Appl. **193**, 324–353 (2022). https://doi.org/10.1007/s10957-021-01929-5
18. Combettes, P.L., Wajs, V.R.: Signal recovery by proximal forward-backward splitting. Multiscale Model. Simul. **4**(4), 1168–1200 (2005). https://doi.org/10.1137/050626090
19. De Marchi, A., Jia, X., Kanzow, C., Mehlitz, P.: Constrained structured optimization and augmented Lagrangian proximal methods. Technical report, preprint arXiv (2022). arXiv:2203.05276
20. Di Lorenzo, D., Liuzzi, G., Rinaldi, F., Schoen, F., Sciandrone, M.: A concave optimization-based approach for sparse portfolio selection. Optim. Methods Softw. **27**(6), 983–1000 (2012). https://doi.org/10.1080/10556788.2011.577773

21. Fukushima, M., Mine, H.: A generalized proximal point algorithm for certain non-convex minimization problems. Int. J. Syst. Sci. **12**(8), 989–1000 (1981). https://doi.org/10.1080/00207728108963798
22. Grippo, L., Lampariello, F., Lucidi, S.: A nonmonotone line search technique for Newton's method. SIAM J. Numer. Anal. **23**(4), 707–716 (1986). https://doi.org/10.1137/0723046
23. Guo, L., Deng, Z.: A new augmented Lagrangian method for MPCCs - theoretical and numerical comparison with existing augmented Lagrangian methods. Math. Oper. Res. **47**(2), 1229–1246 (2022). https://doi.org/10.1287/moor.2021.1165
24. Guo, L., Ye, J.J.: Necessary optimality conditions and exact penalization for non-Lipschitz nonlinear programs. Math. Program. **168**, 571–598 (2018). https://doi.org/10.1007/s10107-017-1112-0
25. Jia, X., Kanzow, C., Mehlitz, P., Wachsmuth, G.: An augmented Lagrangian method for optimization problems with structured geometric constraints. Math. Programm. (2021). https://doi.org/10.1007/s10107-022-01870-z. to appear
26. Li, H., Lin, Z.: Accelerated proximal gradient methods for nonconvex programming. In: Proceedings of the 28th International Conference on Neural Information Processing Systems, vol. 1, pp. 379–387 (2015). https://doi.org/10.5555/2969239.2969282
27. Liu, Y.-F., Dai, Y.-H., Ma, S.: Joint power and admission control: non-convex ℓ_q approximation and an effective polynomial time deflation approach. IEEE Trans. Signal Process. **63**(14), 3641–3656 (2015). https://doi.org/10.1109/TSP.2015.2428224
28. Marjanovic, G., Solo, V.: On ℓ_q optimization and matrix completion. IEEE Trans. Signal Process. **60**(11), 5714–5724 (2012). https://doi.org/10.1109/TSP.2012.2212015
29. Mordukhovich, B.S.: Variational Analysis and Applications. Springer, Berlin (2018). https://doi.org/10.1007/978-3-319-92775-6
30. Moré, J.J., Sorensen, D.C.: Computing a trust region step. SIAM J. Sci. Stat. Comput. **4**(3), 553–572 (1983). https://doi.org/10.1137/0904038
31. Passty, G.B.: Ergodic convergence to a zero of the sum of monotone operators in Hilbert space. J. Math. Anal. Appl. **72**(2), 383–390 (1979). https://doi.org/10.1016/0022-247X(79)90234-8
32. Rockafellar, R.T., Wets, R.J.-B.: Variational Analysis. Springer, Berlin (2009). https://doi.org/10.1007/978-3-642-02431-3
33. Wang, T., Liu, H.: A nonmonotone accelerated proximal gradient method with variable stepsize strategy for nonsmooth and nonconvex minimization problems. Technical report, preprint Optimization-Online (2021). http://www.optimization-online.org/DB_HTML/2021/04/8365.html
34. Wright, S.J., Nowak, R.D., Figueiredo, M.A.T.: Sparse reconstruction by separable approximation. IEEE Trans. Signal Process. **57**(7), 2479–2493 (2009). https://doi.org/10.1109/tsp.2009.2016892