SIAM J. OPTIM.
Vol. xx, No. x, pp. x–x

CONVERGENCE PROPERTIES OF PROXIMAL (SUB)GRADIENT METHODS WITHOUT CONVEXITY OR SMOOTHNESS OF ANY OF THE FUNCTIONS*

MIKHAIL SOLODOV†

Abstract. We establish convergence properties for a framework that includes a variety of proximal subgradient methods, where none of the involved functions needs to be convex or differentiable. The functions are assumed to be Clarke-regular. Our results cover the projected and conditional variants for the constrained case, the use of the inertial/momentum terms, and incremental methods when each of the functions is itself a sum, and the methods process the components in this sum separately.

Key words. proximal gradient methods, incremental methods, nonsmooth nonconvex optimization.

MSC codes. 90C30, 90C33, 90C55, 65K05

1. Introduction.

We consider constrained optimization problems of the form

$$\min_{x\in D} f(x) := \sum_{i=1}^{m} f_i(x), \quad f_i(x) = h_i(x) + g_i(x)\,,\ i = 1,\dots,m\,, \tag{1.1}$$

where $D \subset \mathbb{R}^n$ is a closed convex set and $m \geq 1$ is an integer. The functions $h_i : \mathbb{R}^n \to \mathbb{R}$ and $g_i : \mathbb{R}^n \to \mathbb{R}$, $i = 1,\dots,m$, are assumed to be locally Lipschitz-continuous and regular in the sense of Clarke [9, 28]. Note that the number of functions h_i and g_i in a given problem can be different. But one can always aggregate or split the components, or add trivial ones, to arrive to the format given by (1.1).

We emphasize that none of the functions need to be differentiable or convex. That said, while there are many situations where functions involved are Clarke-regular, this setting does not cover some important applications, like ReLU neural networks; see, e.g., the discussions in [13] and [33]. The purpose of this paper is to show that, when the functions are regular, a vareity of (incremental) proximal gradient type methods can be shown to converge (in a certain sense) in the *fully* nonsmooth and nonconvex settings.

When in (1.1) $m = 1$ and $D = \mathbb{R}^n$, the problem becomes

$$\min f(x) := (h(x) + g(x))\,.$$

If h is differentiable and its gradient is Lipschitz-continuous with modulus $L > 0$, and g is convex, the fundamental algorithm for solving the problem is that of proximal gradient; see, e.g., [2, Chapter 10] and [10, Chapter 2]. In its basic form, given the current iterate $x^k \in \mathbb{R}^n$, this method first makes the gradient step on h,

$$z^k = x^k - \frac{1}{L}h'(x^k),$$

*Submitted to the editors August 2023 (Revised February and September 2024).
Funding: The author is supported in part by CNPq Grant 306775/2023-9, by FAPERJ Grant E-26/200.347/2023, and by PRONEX–Optimization.
†IMPA – Instituto de Matemática Pura e Aplicada, Estrada Dona Castorina 110, Jardim Botânico, Rio de Janeiro, RJ 22460-320 Brazil (solodov@impa.br).

and then obtains the next iterate by computing the proximal point for g with respect to this z^k,

$$x^{k+1} = \arg\min \left(\frac{1}{L} g(x) + \frac{1}{2} \|x - z^k\|^2 \right) .$$

Or, equivalently,

$$x^{k+1} = \arg\min \left(\frac{1}{L} g(x) + \frac{1}{2} \|x - (x^k - \frac{1}{L} h'(x^k))\|^2 \right) . \tag{1.2}$$

The motivation for this scheme is that, in many important applications, the structure of g is such that computing proximal points in (1.2) can be easy or even explicit; see, e.g., [8]. Of course, (1.2) is the conceptual idea, and more sophisticated algorithms have been developed around it.

When m in (1.1) is large, incremental methods [4] process one function f_i at a time. For example, given the current iterate $x^k \in \mathbb{R}^n$, when $D = \mathbb{R}^n$ and the functions f_i are differentiable, the basic incremental gradient method proceeds as follows:

$$y^{k,0} = x^k, \ \alpha_k > 0, \quad y^{k,i} = y^{k,i-1} - \alpha_k f_i'(y^{k,i-1}) \, , i = 1, \ldots, m, \quad x^{k+1} = y^{k,m} .$$

Incremental gradient methods were originally motivated by machine learning applications, where they were known as *backpropagation*. Their convergence analysis dates back to [25, 23]. In this paper, instead of the incremental gradient steps for the components f_i, we shall consider incremental proximal (sub)gradient steps for $f_i = (h_i + g_i)$ of the form in (1.2), but without differentiability assumptions on h_i.

At its origins, in the proximal gradient methods (incremental or not) the function h, or the functions h_i comprising h, were always assumed to be differentiable, typically with Lipschitz-continuous gradient; see, e.g., [3, 15, 24, 31] for some examples (without attempting to be exhaustive in the list of relevant literature). In [20], Lipschitz-continuity of gradients is relaxed to a weaker property, but the functions h_i are still assumed to be differentiable. The function g was usually assumed to be convex; see, e.g., [2, Chapter 10], [31]. There are extentions allowing nonconvex g (but still with smooth h); see, e.g., [1] for the proximal gradient method (1.2), and more generally [6], and [20], [21] for some related incremental approaches. In [16] the proximal subgradient method is considered when both h and g are nonsmooth but convex (and $m = 1$).

We next discuss some literature where, like in this paper, none of the functions is assumed to be differentiable or convex. The fundamental work [13] proves convergence of the proximal subgradient method for the class of *tame* functions, with $m = 1$. On the other hand, apart from the different settings, the current paper covers the important case of incremental algorithms ($m > 1$), the use of momentum terms, allows possibly different subgradient and proximal parameters along the iterations, and larger stepsizes (here, the stepsizes are allowed to go to zero arbitrarily slowly, while [13] requires them to tend to zero fast enough – be square summable). Convergence rates for the proximal subgradient method for weakly convex functions [12] are obtained in [11]. In [33] some algorithms for more general than regular functions are considered. However, the nature of those algorithms is very different. First, the framework requires checking a descent test, which means knowing the value of the full function f. This cannot be applied in incremental methods, where knowledge of one f_i at a time is only available, while the full function f is never known. And more importantly, the framework in [33] is not of the "black-box" type when it comes to

computing subgradients. Even in the non-incremental case ($m = 1$) it requires to compute a special subgradient (not an arbitrary one), in particular the one which is associated to the directional derivative at the given point. While this is possible in some applications, it is not so in nonsmooth optimization in general (see, e.g., the discussion in [7, Part II]). The issue of using the traditional subgradient oracle has been addressed in [14] in the randomized framework. The work [19] derandomizes the algorithm, and [17] includes the constrained setting.

Some words about our notation. By $\mathbb{B}$ we denote the unit ball in $\mathbb{R}^n$ centered at the origin. By $P_D(z)$ we denote the Euclidean projection of the point $z \in \mathbb{R}^n$ onto D. The normal cone to D at $z \in \mathbb{R}^n$ is given by $N_D(z) = \{\nu \in \mathbb{R}^n \mid \langle \nu, y - z\rangle \leq 0 \; \forall\, y \in D\}$ for $z \in D$; $N_D(z) = \emptyset$ if $z \notin D$. Recall that for any $z \in \mathbb{R}^n$,

$$z - P_D(z) \in N_D(P_D(z))\,, \quad \langle z - P_D(z), y - P_D(z)\rangle \leq 0 \; \forall\, y \in D\,. \tag{1.3}$$

Let conv X stand for the convex hull of a set $X \subset \mathbb{R}^n$. Recall that a (locally) Lipschitz-continuous function $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable almost everywhere. Let $\mathcal{D}_f$ be the set of points where f is differentiable. Then the Clarke subdifferential of f at $x \in \mathbb{R}^n$ is the set

$$\partial f(x) = \mathrm{conv}\{v \in \mathbb{R}^n \mid \exists\, \{y^j\} \to x \text{ s.t. } \{y^j\} \subset \mathcal{D}_f,\ \{f'(y^j)\} \to v\,, j \to \infty\}\,.$$

When f is convex, ∂f is the same as the subdifferential in Convex Analysis.

Under our standing assumptions, it holds that for any bounded set $X \subset \mathbb{R}^n$ there exists some constant $L > 0$ such that

$$\|v_{h_i}\| \leq L,\ \|v_{g_i}\| \leq L, \quad \forall\, v_{h_i} \in \partial h_i(x),\ \forall\, v_{g_i} \in \partial g_i(x),\ \forall\, x \in X\,, \tag{1.4}$$

$i = 1, \ldots, m$.

The set of stationary points of problem (1.1) is given by

$$S := \{x \in D \mid 0 \in \partial f(x) + N_D(x)\}. \tag{1.5}$$

For a set-valued mapping F from $\mathbb{R}^n$ to the subsets of $\mathbb{R}^n$, its outer limit at $\bar{z} \in \mathbb{R}^n$ is the set

$$\limsup_{z \to \bar{z}} F(z) = \{v \in \mathbb{R}^n \mid \exists\, \{z^j\} \to \bar{z},\ \exists\, v^j \in F(z^j) \text{ s.t. } \{v^j\} \to v\,, j \to \infty\}.$$

The mapping F is outer semi-continuous at $\bar{z}$ if it holds that $\limsup_{z \to \bar{z}} F(z) \subset F(\bar{z})$. Under our standing assumptions, the Clarke subdifferentials of h_i, g_i, $i = 1, \ldots, m$, and of f, h, g, are outer semi-continuous at any $x \in D$. Also, the normal cone $N_D(x)$ is outer semi-continuous at all $x \in D$.

For a set-valued mapping Φ from $\mathbb{N} \times \mathbb{R}^n$ to the subsets of $\mathbb{R}^n$, its outer limit at $\bar{z} \in \mathbb{R}^n$ is

$$\limsup_{k \to \infty,\, z \to \bar{z}} \Phi(k, z) = \left\{ v \in \mathbb{R}^n \;\middle|\; \begin{array}{c} \exists \{j_k\} \to \infty \text{ as } k \to \infty, \exists \{z^k\} \to \bar{z}, \exists v^k \in \Phi(j_k, z^k) \\ \text{s.t. } \{v^k\} \to v\,, k \to \infty \end{array} \right\}.$$

2. Description of the algorithms. We proceed to formally state the algorithms in consideration. The relations that need to be imposed on the algorithmic parameters (the prox-paramenter, the subgradient stepsize, inexactness in solving subproblems, and the momentum term) are stated in the convergence analysis Section 3 below. One interesting new feature of all the algorithms is that the prox-paramenters

β_k and the subgradient stepsizes α_k are allowed to be unequal along iterations, as long as their ratio tends to one asymptotically (as will be stated in Section 3). Of course, they can also be taken equal, as is usual in the literature.

What is meant by an approximate stationary point of the subproblem (2.1) below is specified in (2.3), immediately after the algorithm. This algorithm takes the starting point $x^0 \in D$. Then, by (2.2), it generates feasible iterates $x^k \in D$ for all $k = 0, 1, 2, \ldots$.

ALGORITHM 2.1. **[Incremental Proximal Gradient Method with Momentum Terms]**

1. For the iteration $k \in \mathbb{N}$ and the corresponding iterate x^k, set $y^{k,0} = x^k$ and choose $\alpha_k > 0$, $\beta_k > 0$.
2. For $i = 1, \ldots, m$, choose the error-tolerance parameter $\varepsilon_{k,i} \geq 0$ and compute $y^{k,i}$ as an approximate, in the sense of (2.3), stationary point of the problem

   (2.1)
   $$\min_{x \in \mathbb{R}^n} \ \beta_k g_i(x) + \frac{1}{2}\|x - (y^{k,i-1} - \alpha_k v_h^{k,i-1})\|^2, \quad \text{where } v_h^{k,i-1} \in \partial h_i(y^{k,i-1}).$$

3. Choose $\gamma_k \geq 0$ and set

   (2.2)
   $$x^{k+1} = P_D(y^{k,m} + \gamma_k(x^k - x^{k-1})).$$

   Set $k := k + 1$ and go to Step 1.

By computing an approximate stationary point of (2.1) we mean the natural condition that $y^{k,i}$ satisfies

(2.3)
$$0 \in \beta_k \partial g_i(y^{k,i}) + y^{k,i} - y^{k,i-1} + \alpha_k v_h^{k,i-1} + \varepsilon_{k,i}\mathbb{B}\,.$$

Note that when g_i is not convex, for the step (2.1) in Algorithm 2.1 to be well-defined, the assumption that the proximal subproblems therein are solvable is required. This can be related, for example, to prox-boundedness [27] of g_i. However, we actually only need the existence of stationary points of proximal subproblems, which furthermore need not be unique. Also, as commented by a referee, prox-boundedness can be ensured by adding to g_i and subtracting from h_i a function with sufficient coercivity properties. For example, if g_i is weakly convex [12] then adding to it a certain multiple of $\|x\|^2$ would make the sum convex and thus prox-bounded.

The inertia term $\gamma_k(x^k - x^{k-1})$ in (2.2) is known as "momentum term" in the machine learning literature [18], and "heavy ball term" in optimization [26]. This, or some similar modification of the step playing the same role, is important for improving computational performance in many applications [18, 8]. Our momentum term parameters γ_k would require to tend to zero, which is different from the previous literature for smooth functions and non-incremental algorithms. However, it is known that it has to go to zero even in the simpler (than considered here) incremental gradient methods, and also in the smooth case [25, Theorem 3.1].

Remark 2.1. It is worth to mention that just like the tolerance parameter $\varepsilon_{k,i}$, we could allow the subgradient stepsize α_k and the prox-parameter β_k to vary over the components $i = 1, \ldots, m$. However, this would complicate considerably the notation and technical details in the convergence analysis in Section 3. For this reason, we limit this issue to a remark.

In the following *conditional* variant of the method, where in (2.4) below minimization is performed over the set D, no solvability assumptions are needed if the set D is compact.

Again, at $k = 0$, we take $x^0 \in D$.

ALGORITHM 2.2. **[Incremental Conditional Proximal Gradient Method with Momentum Terms]**

1. For the iteration $k \in \mathbb{N}$ and the corresponding iterate x^k, set $y^{k,0} = x^k$ and choose $\alpha_k > 0$, $\beta_k > 0$.
2. For $i = 1, \ldots, m$, choose the error-tolerance parameter $\varepsilon_{k,i} \geq 0$ and compute $y^{k,i}$ as an approximate, in the sense of (2.6), stationary point of the problem

$$\min_{x \in D} \ \beta_k g_i(x) + \frac{1}{2}\|x - (y^{k,i-1} - \alpha_k v_h^{k,i-1})\|^2, \quad \text{where } v_h^{k,i-1} \in \partial h_i(y^{k,i-1}). \tag{2.4}$$

3. Choose $\gamma_k \geq 0$ and set

$$x^{k+1} = P_D(y^{k,m} + \gamma_k(x^k - x^{k-1})). \tag{2.5}$$

Set $k := k + 1$ and go to Step 1.

By computing an approximate stationary point $y^{k,i}$ of (2.4) we mean the following natural condition for constrained problems:

$$\begin{array}{c} \exists\, w^{k,i} \in \beta_k \partial g_i(y^{k,i}) + y^{k,i} - y^{k,i-1} + \alpha_k v_h^{k,i-1} \text{ such that} \\ y^{k,i} - P_D(y^{k,i} - w^{k,i}) \in \varepsilon_{k,i}\mathbb{B}\,. \end{array} \tag{2.6}$$

In particular, for $\varepsilon_{k,i} = 0$, the relations in (2.6) mean computing an exact stationary point of (2.4).

The following algorithm differs from Algorithm 2.1 in that each proximal gradient step for $i = 1, \ldots, m$ is followed by projection onto D.

ALGORITHM 2.3. **[Incremental Projected Proximal Gradient Method with Momentum Terms]**

1. For the iteration $k \in \mathbb{N}$ and the corresponding iterate x^k, set $y^{k,0} = x^k$ and choose $\alpha_k > 0$, $\beta_k > 0$.
2. For $i = 1, \ldots, m$, choose the error-tolerance parameter $\varepsilon_{k,i} \geq 0$ and compute $z^{k,i}$ as an approximate, in the sense of (2.3) with $y^{k,i}$ therein substituted by $z^{k,i}$, stationary point of the problem (2.1), and set

$$y^{k,i} = P_D(z^{k,i}). \tag{2.7}$$

3. Choose $\gamma_k \geq 0$ and set

$$x^{k+1} = P_D(y^{k,m} + \gamma_k(x^k - x^{k-1})). \tag{2.8}$$

Set $k := k + 1$ and go to Step 1.

3. Convergence analysis. We first outline the Generalized Lyapunov Direct Method convergence analysis technique of [32] (see also [30]), adapted to our purposes. This technique is particularly useful when there does not exist a classical Lyapunov

function, which is guaranteed to behave monotonically along the iterative process. Note that, in general, such a function cannot be exhibited for incremental methods (even if the functions are smooth), as well as for nonsmooth subgradient methods (even without the incremental features).

Consider the general iterative process

$$x^{k+1} \in x^k - \alpha_k \Phi(k, x^k), \quad k = 0, 1, \ldots, \quad x^0 \in X, \tag{3.1}$$

where Φ is a set-valued mapping from $\mathbb{N} \times X$ to the subsets of X, with X being an open set in $\mathbb{R}^n$.

Choose a locally Lipschitz-continuous function $V : \mathbb{R}^n \to \mathbb{R}$, regular in the sense of Clarke. We call V pseudo-Lyapunov function ("pseudo", because it need not be monotone along the iterative sequence $\{x^k\}$ generated by (3.1)). The choice of V depends on the problem being solved, and the specific instance of (3.1), i.e., on the algorithm mapping Φ. In our applications below, $V = f$, the objective function in (1.1).

Let $\{x^k\}$ be a bounded sequence, with all its accumulation points belonging to some convex compact set $C \subset X$. Define

$$R(x) = \mathrm{conv}\{\partial V(x) \cup N_C(x)\}.$$

Denote the outer limit of Φ at x by

$$\bar{\Phi}(x) := \limsup_{k \to \infty,\, z \to x} \Phi(k, z)\,.$$

For the pseudo-Lyapunov function V, the set C, and the mapping $\bar{\Phi}$, define the following set:

$$A := \{x \in C \mid \max_{\rho \in R(x)} \min_{q \in \bar{\Phi}(x)} \langle \rho, q \rangle \leq 0\}\,. \tag{3.2}$$

This set serves as an attractor for the iterates generated by (3.1); it consists of all the points in C for which $-\bar{\Phi}(x)$ does not contain feasible directions (for C) that are of descent for the pseudo-Lyapunov function V. In our applications below, we will have $V = f$ (the objective function of (1.1)) and C being the intersection of D (the feasible set of (1.1)) with some compact convex set X that contains the (assumed bounded) iterative sequences in its interior. Furthermore, it will be proven that $A \subset S$, where S is given by (1.5) (i.e., the attractors are stationary points of problem (1.1)).

A set $\Omega \subset \mathbb{R}^n$ is said to be V-connected, if the set $V(\Omega) = \{t \in \mathbb{R} \mid \exists x \in \Omega, \quad t = V(x)\}$ is a connected set in $\mathbb{R}$. Denote by $\{A^t\}, \quad t \in T$, the (unique) decomposition of A into V-connected components, i.e.,

$$A = \cup_{t \in T} A^t, \quad A^{t'} \neq A^{t''} \quad \text{for } t' \neq t'', \quad t', t'' \in T.$$

THEOREM 3.1. *[32, 30, Adapted version] Let $\{x^k\}$ be any sequence generated by the process (3.1), where*

$$\sup_{x \in X} \limsup_{z \to x, k \to \infty} \sup_{u \in \Phi(k,z)} \|u\| < \infty\,, \tag{3.3}$$

$$\alpha_k > 0, \quad \lim_{k \to \infty} \alpha_k = 0, \quad \sum_{k=0}^{\infty} \alpha_k = \infty\,. \tag{3.4}$$

Suppose further that $\{x^k\}$ is bounded, with all its accumulation points belonging to some convex compact set $C \subset X$.

Then there exists $t \in T$ such that all accumulation points of the sequence $\{V(x^k)\}$ belong to the set $V\left(A^t \cap \bar{X}\right)$, where $\bar{X}$ is the set of all accumulation points of $\{x^k\}$.

If, in addition, the set $V(A)$ is nowhere dense in $\mathbb{R}$, then all the accumulation points of the sequence $\{x^k\}$ belong to a V-connected component of A defined by (3.2).

The following lemma will be used for relating the algorithms stated in Section 2 to the framework of (3.1) and Theorem 3.1.

LEMMA 3.2. *Let $D \subset \mathbb{R}^n$ be a closed convex set. Then for $x = P_D(y - \alpha d)$, where $y \in D$, $d \in \mathbb{R}^n$ and $\alpha > 0$, there exists $\nu \in N_D(x)$ such that $x = y - \alpha(d + \nu)$ and $\|\nu\| \leq \|d\|$.*

Proof. By the property of the projection operator (1.3), $(y - \alpha d) - x \in N_D(x)$. Hence, there exists $\hat{\nu} \in N_D(x)$ such that $y - \alpha d - x = \hat{\nu}$. Define $\nu = \frac{1}{\alpha}\hat{\nu} \in N_D(x)$. Then $d = \frac{1}{\alpha}(y - x) - \nu$. We obtain that

$$\begin{aligned}\|\nu\|^2 &\leq \|\nu\|^2 + \frac{1}{\alpha^2}\|y - x\|^2 \\ &\leq \|\nu\|^2 + \frac{1}{\alpha^2}\|y - x\|^2 - \frac{1}{\alpha}\langle \nu, y - x\rangle \\ &= \|d\|^2,\end{aligned}$$

where the second inequality holds because $\nu \in N_D(x)$ and $y \in D$. ∎

We start with convergence analysis of the Incremental Proximal Gradient Method with Momentum Terms (Algorithm 2.1).

It is well known that the condition of stepsize tending to zero (like that in (3.4)) is indispensable for convergence of subgradient methods in the nonsmooth case, even for convex problems and without any incremental features. This condition is also required for incremental methods, even when all the functions are smooth (the latter is demonstrated in [22, Section 2]; see also [25]). The exceptions are some special cases (like [5]) or when the theoretical convergence guarantees concern approximate solutions [29].

We shall assume implicitly that Algorithm 2.1 is well-defined, in the sense that the proximal subproblems therein are solvable (at least to approximate stationarity). This issue has been already discussed in Section 2. We shall also assume that all the generated iterates are bounded. In a sense, this is a limitation of the general nonconvex nonsmooth setting and of the adopted methodology.

THEOREM 3.3. *Let $\{x^k\}$ and $\{y^{k,i}\}$, $i = 1, \ldots, m$, be any bounded sequences generated by Algorithm 2.1, where the parameters satisfy (3.4) and*

$$\lim_{k\to\infty} \frac{\beta_k}{\alpha_k} = 1, \quad \lim_{k\to\infty} \frac{\gamma_k}{\alpha_k} = 0, \quad \lim_{k\to\infty} \frac{\varepsilon_{k,i}}{\alpha_k} = 0, \ i = 1, \ldots m. \tag{3.5}$$

Then there exists an f-connected component S^t of the set of stationary points S of problem (1.1) such that all accumulation points of the sequence $\{f(x^k)\}$ belong to the set $f(S^t \cap \bar{X})$, where $\bar{X}$ is the set of accumulation points of $\{x^k\}$.

If, in addition, the set $f(S)$ is nowhere dense in $\mathbb{R}$, then all accumulation points of $\{x^k\}$ belong to an f-connected component of the set of stationary points S.

Proof. By (2.3), we obtain that there exist $v_g^{k,i} \in \partial g_i(y^{k,i})$ and $r^{k,i} \in \mathbb{B}$ such that

$$\beta_k v_g^{k,i} + y^{k,i} - y^{k,i-1} + \alpha_k v_h^{k,i-1} + \varepsilon_{k,i} r^{k,i} = 0,\ i = 1, \ldots, m,$$

where $v_h^{k,i-1} \in \partial h_i(y^{k,i-1})$. Hence,

$$y^{k,i} = y^{k,i-1} - \alpha_k \left(v_h^{k,i-1} + v_g^{k,i} + \left(\frac{\beta_k}{\alpha_k} - 1 \right) v_g^{k,i} + \frac{\varepsilon_{k,i}}{\alpha_k} r^{k,i} \right),\ i = 1, \ldots, m\,,$$

where $y^{k,0} = x^k \in D$. Defining

$$d^{k,i} = v_h^{k,i-1} + v_g^{k,i} + \left(\frac{\beta_k}{\alpha_k} - 1 \right) v_g^{k,i} + \frac{\varepsilon_{k,i}}{\alpha_k} r^{k,i},\ i = 1, \ldots, m\,, \tag{3.6}$$

we obtain that

$$y^{k,m} = x^k - \alpha_k \sum_{i=1}^{m} d^{k,i}.$$

Therefore (see (2.2)),

$$\begin{aligned} x^{k+1} &= P_D(y^{k,m} + \gamma_k(x^k - x^{k-1})) \\ &= P_D\left(x^k - \alpha_k \left(\sum_{i=1}^{m} d^{k,i} - \frac{\gamma_k}{\alpha_k}(x^k - x^{k-1}) \right) \right) \\ &= x^k - \alpha_k \left(\sum_{i=1}^{m} d^{k,i} - \frac{\gamma_k}{\alpha_k}(x^k - x^{k-1}) + \nu^{k+1} \right), \end{aligned} \tag{3.7}$$

where the last equality is by Lemma 3.2, with

$$\nu^{k+1} \in N_D(x^{k+1}), \quad \|\nu^{k+1}\| \le \left\| \sum_{i=1}^{m} d^{k,i} - \frac{\gamma_k}{\alpha_k}(x^k - x^{k-1}) \right\|. \tag{3.8}$$

Let $Z \subset \mathbb{R}^n$ be some compact convex set containing $\{x^k\}$ and at least one stationary point of (1.1) (i.e, $Z \cap S \neq \emptyset$). Let further $X \subset \mathbb{R}^n$ be some open bounded set containing Z and $\{y^{k,i}\}$, $i = 1, \ldots, m$. By (1.4), by the definition of $d^{k,i}$ in (3.6), by the conditions on the algorithm parameters (3.5), and by (3.8), it follows that there exists some $c > 0$ such that

$$\left\| \sum_{i=1}^{m} d^{k,i} \right\| \le c, \quad \|x^k - x^{k-1}\| \le c, \quad \|\nu^{k+1}\| \le c \tag{3.9}$$

holds for all $k = 1, 2, \ldots$. We can further enlarge the sets Z and X, if necessary, so that still $Z \subset X$ and

$$\sum_{i=1}^{m} d^{k,i} - \frac{\gamma_k}{\alpha_k}(x^k - x^{k-1}) + \nu^{k+1} \in X,$$

for all $k = 1, 2, \ldots$.

We next write the iterates update (3.7) in the form of the iterative process (3.1) with the following mapping Φ from $\mathbb{N} \times X$ to the subsets of X:

(3.10)
$$\Phi(k,x) = \left\{ u \in \mathbb{R}^n \;\middle|\; \begin{array}{c} u = \sum_{i=1}^m d^i - \frac{\gamma_k}{\alpha_k} w + \nu, \\ d^i = v_h^{i-1} + v_g^i + \left(\frac{\beta_k}{\alpha_k} - 1\right) v_g^i + \frac{\varepsilon_{k,i}}{\alpha_k} r^i, \\ v_h^{i-1} \in \partial h_i(y^{i-1}), v_g^i \in \partial g_i(y^i), r^i \in \mathbb{B}, \\ y^0 = x,\ y^i = y^{i-1} - \alpha_k d^i,\ i = 1, \ldots, m, \\ \nu \in N_D(p),\ \|\nu\| \le c,\ p = x - \alpha_k \left(\sum_{i=1}^m d^i - \frac{\gamma_k}{\alpha_k} w + \nu \right), \\ w \in \mathbb{R}^n,\ \|w\| \le c. \end{array} \right\}$$

In the context of the iterative process (3.1) and Theorem 3.1, we choose $V = f$ and $C = D \cap Z$. We next show that the attraction set for (3.1),

$$A = \{ x \in C \mid \max_{\rho \in R(x)} \min_{q \in \bar{\Phi}(x)} \langle \rho, q \rangle \le 0 \},$$

where

$$R(x) = \operatorname{conv}\{\partial f(x) \cup N_C(x)\},$$

belongs to the set of stationary points S (1.5) of problem (1.1).

By (3.9) it holds that the mapping Φ given by (3.10) is bounded on X, and hence its outer limits are bounded, and the condition (3.3) of Theorem 3.1 holds.

We next estimate the outer limit $\bar{\Phi}(x) = \limsup_{k \to \infty,\, z \to x} \Phi(k,z)$. By (3.4), (3.5) and (3.9), we have that $y^i \to x$, $i = 1, \ldots, m$, and $p \to x$ as $z \to x$, $k \to \infty$. Then, by the other semicontinuity of the Clarke subdifferentials and of the normal cone to the closed convex set D, we have that the accumulation points of $v_h^{i-1} \in \partial h_i(y^{i-1})$ belong to $\partial h_i(x)$, the accumulation points of $v_g^i \in \partial g_i(y^i)$ belong to $\partial g_i(x)$, and the accumulation points of ν belong to $N_D(x)$, as $z \to x$, $k \to \infty$. It further follows, by (3.5), that the accumulation points of d^i belong to $\partial h_i(x) + \partial g_i(x)$. Taking into account again (3.5), (3.9), and putting things together, we conclude that

$$\begin{aligned} \bar{\Phi}(x) = \limsup_{k \to \infty,\, z \to x} \Phi(k,z) &\subset \sum_{i=1}^m (\partial h_i(x) + \partial g_i(x)) + N_D(x) \\ &= \partial f(x) + N_D(x)\,, \end{aligned} \tag{3.11}$$

where Clarke regularity of the functions involved was taken into account.

We prove that $A \subset Z \cap S \subset S$ by showing that if $x \notin Z \cap S$ then $x \notin A$. Suppose $x \notin Z \cap S$. If $x \notin C = D \cap Z$, then $x \notin A$ by the very definition of A. Let $x \in C = D \cap Z$. Then $x \notin Z \cap S$ implies that $x \notin S$. The latter means that the problem

$$\min \|s\| \text{ subject to } s \in \partial f(x) + N_D(x)\,,$$

has (unique) solution $\bar{s} \neq 0$. Since $x \in C = D \cap Z$, we have that $N_D(x) \subset N_C(x)$. Hence, $\bar{s} \in \partial f(x) + N_C(x)$.

Then,

$$\bar{s} = s^1 + s^2, \quad s^1 \in \partial f(x),\ s^2 \in N_C(x).$$

It follows that

$$\frac{1}{2}\bar{s} = \frac{1}{2}s^1 + \frac{1}{2}s^2 \in \operatorname{conv}\{\partial f(x) \cup N_C(x)\} = R(x).$$

We then obtain that

$$\begin{aligned}\max_{\rho\in R(x)}\min_{q\in\bar\Phi(x)}\langle\rho,q\rangle &\geq \frac{1}{2}\min_{q\in\bar\Phi(x)}\langle\bar s,q\rangle\\ &\geq \frac{1}{2}\min_{q\in\partial f(x)+N_D(x)}\langle\bar s,q\rangle\,,\end{aligned}\tag{3.12}$$

where the second inequality is by (3.11).

As $\bar s$ is the orthogonal projection of the origin onto $\partial f(x)+N_D(x)$, by (1.3) we have that

$$\|\bar s\|^2\leq\langle\bar s,q\rangle\quad\forall q\in\partial f(x)+N_D(x)\,.\tag{3.13}$$

Combining (3.12) and (3.13), we conclude that

$$\max_{\rho\in R(x)}\min_{q\in\bar\Phi(x)}\langle\rho,q\rangle\geq\frac{1}{2}\|\bar s\|^2>0\,,$$

because $\bar s\neq 0$ when $x\notin S$.

This proves that $A\subset S$, and the assertions now follow from Theorem 3.1. ∎

Remark 3.4. Note that according to (3.5), for convergence the proximal parameter β_k can be along iterations both smaller or larger than the subgradient stepsize α_k (or also equal to it). However, their ratio must tend to one eventually.

We next turn our attention to the conditional variant of the method, i.e., Algorithm 2.2. Note that as all the generated iterates in this case are feasible, their boundedness is automatic if the set D is compact.

THEOREM 3.5. *For sequences generated by Algorithm 2.2, under the same assumptions as those in Theorem 3.3, the same assertions hold.*

Proof. By (2.6), we have that there exist $v_g^{k,i}\in\partial g_i(y^{k,i})$ and $r^{k,i}\in\mathbb{B}$ such that, for $i=1,\ldots,m$,

$$\begin{aligned}y^{k,i}&=P_D(y^{k,i}-w^{k,i})+\varepsilon_{k,i}r^{k,i}\\ &=P_D(y^{k,i-1}-\beta_k v_g^{k,i}-\alpha_k v_h^{k,i-1})+\varepsilon_{k,i}r^{k,i}\\ &=P_D\left(y^{k,i-1}-\alpha_k\left(v_h^{k,i-1}+v_g^{k,i}+\left(\frac{\beta_k}{\alpha_k}-1\right)v_g^{k,i}\right)\right)+\varepsilon_{k,i}r^{k,i}\\ &=y^{k,i-1}-\alpha_k\left(v_h^{k,i-1}+v_g^{k,i}+\left(\frac{\beta_k}{\alpha_k}-1\right)v_g^{k,i}+\nu^{k,i}\right)+\varepsilon_{k,i}r^{k,i}\,,\end{aligned}$$

where the last equality is by Lemma 3.2, with

$$\nu^{k,i}\in N_D(y^{k,i}),\quad\|\nu^{k,i}\|\leq\left\|v_h^{k,i-1}+v_g^{k,i}+\left(\frac{\beta_k}{\alpha_k}-1\right)v_g^{k,i}\right\|\,.\tag{3.14}$$

Defining $d^{k,i}$ as in (3.6), and taking into account that $y^{k,0}=x^k$, we obtain that

$$y^{k,m}=x^k-\alpha_k\sum_{i=1}^m(d^{k,i}+\nu^{k,i}).$$

Therefore (see (2.5)),

$$\begin{aligned} x^{k+1} &= P_D(y^{k,m} + \gamma_k(x^k - x^{k-1})) \\ &= P_D\left(x^k - \alpha_k\left(\sum_{i=1}^m (d^{k,i} + \nu^{k,i}) - \frac{\gamma_k}{\alpha_k}(x^k - x^{k-1})\right)\right) \\ &= x^k - \alpha_k\left(\sum_{i=1}^m (d^{k,i} + \nu^{k,i}) - \frac{\gamma_k}{\alpha_k}(x^k - x^{k-1}) + \nu^{k+1}\right), \end{aligned} \tag{3.15}$$

where the last equality is by Lemma 3.2, with

$$\nu^{k+1} \in N_D(x^{k+1}), \quad \|\nu^{k+1}\| \le \left\|\sum_{i=1}^m (d^{k,i} + \nu^{k,i}) - \frac{\gamma_k}{\alpha_k}(x^k - x^{k-1})\right\|. \tag{3.16}$$

Taking again $Z \subset \mathbb{R}^n$ to be some compact convex set containing $\{x^k\}$ and at least one point in S, and taking $X \subset \mathbb{R}^n$ to be some open bounded set containing Z and $\{y^{k,i}\}$, $i = 1, \ldots, m$, by (1.4), by the definition of $d^{k,i}$ in (3.6), by the conditions on the algorithm parameters (3.5), and by (3.14) and (3.16), we can ensure that all the involved objects are bounded in norm by some $c > 0$. Also,

$$\sum_{i=1}^m (d^{k,i} + \nu^{k,i}) - \frac{\gamma_k}{\alpha_k}(x^k - x^{k-1}) \in X\,.$$

We next define the mapping Φ from $\mathbb{N} \times X$ to the subsets of X, associated to the iterates given by (3.15):

$$\Phi(k,x) = \left\{ u \in \mathbb{R}^n \;\middle|\; \begin{array}{c} u = \sum_{i=1}^m (d^i + \nu^i) - \frac{\gamma_k}{\alpha_k} w + \nu, \\ d^i = v_h^{i-1} + v_g^i + \left(\frac{\beta_k}{\alpha_k} - 1\right) v_g^i + \frac{\varepsilon_{k,i}}{\alpha_k} r^i, \\ v_h^{i-1} \in \partial h_i(y^{i-1}), v_g^i \in \partial g_i(y^i), r^i \in \mathbb{B}, \\ \nu^i \in N_D(y^i),\ \|\nu^i\| \le c, \\ y^0 = x,\ y^i = y^{i-1} - \alpha_k d^i,\ i = 1, \ldots, m, \\ \nu \in N_D(p),\ \|\nu\| \le c, \\ p = x - \alpha_k \left(\sum_{i=1}^m (d^i + \nu^i) - \frac{\gamma_k}{\alpha_k} w + \nu\right), \\ w \in \mathbb{R}^n,\ \|w\| \le c. \end{array} \right\} \tag{3.17}$$

Estimating the outer limit $\bar{\Phi}(x) = \limsup_{k\to\infty,\, z\to x} \Phi(k,z)$, we have that $y^i \to x$, $i = 1, \ldots, m$, and $p \to x$ as $z \to x$, $k \to \infty$. Then,

$$\begin{aligned} \bar{\Phi}(x) = \limsup_{k\to\infty,\, z\to x} \Phi(k,z) &\subset \sum_{i=1}^m (\partial h_i(x) + \partial g_i(x) + N_D(x)) + N_D(x) \\ &= \partial f(x) + N_D(x)\,, \end{aligned} \tag{3.18}$$

where Clarke regularity of the functions involved was taken into account, as well as the fact that for any convex cone K it holds that $K + K = K$.

The rest of the proof is the same as the corresponding part of Theorem 3.3. ∎

We finally consider Algorithm 2.3, where the projection is performed after each proximal-subgradient step.

THEOREM 3.6. *For sequences generated by Algorithm 2.3, under the same assumptions as those in Theorem 3.3, the same assertions hold.*

Proof. By (2.3) written for $z^{k,i}$, there exist $v_g^{k,i} \in \partial g_i(z^{k,i})$ and $r^{k,i} \in \mathbb{B}$ such that

$$\beta_k v_g^{k,i} + z^{k,i} - y^{k,i-1} + \alpha_k v_h^{k,i-1} + \varepsilon_{k,i} r^{k,i} = 0,\ i = 1, \ldots, m,$$

where $v_h^{k,i-1} \in \partial h_i(y^{k,i-1})$. We then obtain that

$$\begin{aligned}
y^{k,i} &= P_D(z^{k,i}) \\
&= P_D(y^{k,i-1} - \beta_k v_g^{k,i} - \alpha_k v_h^{k,i-1} - \varepsilon_{k,i} r^{k,i}) \\
&= P_D\left(y^{k,i-1} - \alpha_k \left(v_h^{k,i-1} + v_g^{k,i} + \left(\frac{\beta_k}{\alpha_k} - 1\right) v_g^{k,i} + \frac{\varepsilon_{k,i}}{\alpha_k} r^{k,i}\right)\right) \\
&= y^{k,i-1} - \alpha_k \left(v_h^{k,i-1} + v_g^{k,i} + \left(\frac{\beta_k}{\alpha_k} - 1\right) v_g^{k,i} + \frac{\varepsilon_{k,i}}{\alpha_k} r^{k,i} + \nu^{k,i}\right),
\end{aligned}$$

where the last equality is by Lemma 3.2, with

$$\nu^{k,i} \in N_D(y^{k,i}), \quad \|\nu^{k,i}\| \le \left\| v_h^{k,i-1} + v_g^{k,i} + \left(\frac{\beta_k}{\alpha_k} - 1\right) v_g^{k,i} + \frac{\varepsilon_{k,i}}{\alpha_k} r^{k,i} \right\|. \tag{3.19}$$

Defining again $d^{k,i}$ by (3.6), we have that the relations (3.15) and (3.16) hold, with the difference that now $d^{k,i}$ involves $v_g^{k,i} \in \partial g_i(z^{k,i})$ (instead of $v_g^{k,i} \in \partial g_i(y^{k,i})$ in Theorem 3.5).

Accordingly, the mapping characterizing Algorithm 2.3 is given by

$$\Phi(k,x) = \left\{ u \in \mathbb{R}^n \;\middle|\; \begin{array}{c} u = \sum_{i=1}^m (d^i + \nu^i) - \frac{\gamma_k}{\alpha_k} w + \nu, \\ d^i = v_h^{i-1} + v_g^i + \left(\frac{\beta_k}{\alpha_k} - 1\right) v_g^i + \frac{\varepsilon_{k,i}}{\alpha_k} r^i, \\ v_h^{i-1} \in \partial h_i(y^{i-1}), v_g^i \in \partial g_i(z^i), r^i \in \mathbb{B}, \\ \nu^i \in N_D(y^i),\ \|\nu^i\| \le c, \\ y^0 = x,\ y^i = y^{i-1} - \alpha_k d^i, \\ z^i = y^{i-1} - \beta_k v_g^i - \alpha_k v_h^{i-1} - \varepsilon_{k,i} r^i, \\ i = 1, \ldots, m, \\ \nu \in N_D(p),\ \|\nu\| \le c, \\ p = x - \alpha_k \left(\sum_{i=1}^m (d^i + \nu^i) - \frac{\gamma_k}{\alpha_k} w + \nu\right), \\ w \in \mathbb{R}^n,\ \|w\| \le c. \end{array} \right\}$$

Estimating the outer limit $\bar{\Phi}(x) = \limsup_{k\to\infty,\, z\to x} \Phi(k,z)$, we have that $y^i \to x$ and $z^i \to x$, $i = 1, \ldots, m$, and $p \to x$ as $z \to x$, $k \to \infty$. Then, (3.18) still holds, and the rest of the proof is the same as the corresponding part of Theorem 3.3. ∎

4. Concluding remarks. Convergence properties of proximal (sub)gradient methods had been shown for the case when none of the involved functions needs to be smooth or convex, but be regular in the sense of Clarke. The analysis covers methods with inertial (momentum) terms, as well as the incremental, projected, and conditional variants.

REFERENCES

[1] 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.
[2] A. Beck. *First–Order Methods in Optimization*. SIAM, 2017.
[3] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Science*, 2:183–202, 2009.
[4] D.P. Bertsekas. Incremental gradient, subgradient, and proximal methods for convex optimization: A survey. In Optimization for Machine Learning, S. Sra, S. Nowozin, and S. Wright, eds., MIT Press, Cambridge, MA, 2011, pp. 1–38.
[5] D. Blatt, A.O. Hero, and H. Gauchman. A convergent incremental gradient method with a constant step size. *SIAM J. Optimization*, 18:29–51, 2007.
[6] J. Bolte, S. Sabach, and M. Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146:459–494, 2014.
[7] J.F. Bonnans, J.C. Gilbert, C. Lemaréchal, and C. Sagastizábal. *Numerical Optimization: Theoretical and Practical Aspects*, 2nd ed., Springer, Berlin, 2006.
[8] L. Bottou, F.E. Curtis and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60:223–311, 2018.
[9] F.H. Clarke. *Optimization and Nonsmooth Analysis*. SIAM Publications, Philadelphia, 1990.
[10] Y. Cui and J.-S. Pang. *Modern Nonconvex Nondifferentiable Optimization*. SIAM–MOS, Philadelphia, 2022.
[11] D. Davis and D. Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29:207–239, 2019.
[12] D. Davis and D. Drusvyatskiy. Subgradient methods under weak convexity and tame geometry. *SIAG/OPT Views and News*, 28:1–10, 2020.
[13] D. Davis, D. Drusvyatskiy, S. Kakade, and J.D. Lee. Stochastic subgradient method converges on tame functions. *Foundations of Computational Mathematics*, 20:119–154, 2020.
[14] D. Davis, D. Drusvyatskiy, Y.T. lee, S. Padmanabhan, and G. Ye. A gradient sampling method with complexity guarantees for Lipschitz functions in high and low dimensions. In: *Advances in Neural Information Processing Systems*, NIPS 35: 6692–6703, 2022.
[15] A. Defazio, T. Caetano, and J. Domke. Finito: A faster, permutable incremental gradient method for big data problems. In: *Proceedings of the 31st International Conference on Machine Learning*, PMLR 32(2):1125–1133, 2014.
[16] D.H. Gutman and J.F. Peña. Convergence rates of proximal gradient methods via the convex conjugate. *SIAM J. Optimization*, 29:162–174, 2019.
[17] B. Grimmer and Z. Jia. Goldstein stationarity in Lipschitz constrained optimization. *Optimization Letters*, to appear. arXiv:2310.03690.
[18] T. Khanna. *Foundations of Neural Networks*. Addison-Wesley, Reading, Massachusetts, 1989.
[19] S. Kong and A.S. Lewis. The cost of nonconvexity in deterministic nonsmooth optimization. *Mathematics of Operations Research*, 2023. https://doi.org/10.1287/moor.2022.0289
[20] P. Latafat, A. Themelis, M. Ahookhosh, and P. Patrinos. Bregman Finito/MISO for nonconvex regularized finite sum minimization without Lipschitz gradient continuity. *SIAM J. Optimization*, 18:2230–2262, 2022.
[21] P. Latafat, A. Themelis, and P. Patrinos. Block-coordinate and incremental aggregated proximal gradient methods for nonsmooth nonconvex problems. *Mathematical Programming*, 193:195–224, 2022.
[22] Z.-Q. Luo. On the convergence of the LMS algorithm with adaptive learning rate for linear feedforward networks. *Neural Computation*, 3:226–245, 1991.
[23] Z.-Q. Luo and P. Tseng. Analysis of an approximate gradient projection method with applications to the backpropagation algorithm. *Optimization Methods and Software*, 4:85–101, 1994.
[24] J. Mairal. Incremental majorization-minimization optimization with application to large-scale machine learning. *SIAM J. Optimization*, 25:829–855, 2015.
[25] O.L. Mangasarian and M.V. Solodov. Serial and parallel backpropagation convergence via nonmonotone perturbed minimization. *Optimization Methods and Software*, 4:103–116, 1994.
[26] B.T. Polyak. *Introduction to Optimization*. Optimization Software, Inc., Publications Division, New York, 1987.
[27] R.A. Poliquin and R.T. Rockafellar. Prox-regular functions in Variational Analysis. *Transactions of the American Mathematical Society*, 348:1805–1838, 1996.
[28] R.T. Rockafellar and J.-B. Wets. *Variational Analysis*. Springer-Verlag, New York, 1997.
[29] M.V. Solodov. Incremental gradient algorithms with stepsizes bounded away from zero. *Computational Optimization and Applications*, 11:23–35, 1998.

[30] M.V. Solodov and S.K. Zavriev. Error stabilty properties of generalized gradient-type algorithms. *Journal of Optimization Theory and Applications*, 98:663–680, 1998.

[31] N.D. Vanli, M. Gürbüzbalaban, and A. Ozdaglar. Global convergence rate of proximal incremental aggregated gradient methods. *SIAM J. Optimization*, 28:1282–1300, 2018.

[32] S.K. Zavriev and A.G. Perevozchikov. Direct Lyapunov's method in attraction analysis of finite-difference inclusions. *USSR Computational Mathematics and Mathematical Physics*, 30(1):22–32, 1990.

[33] J. Zhang, H. Lin, S. Jegelka, S. Sra, and A. Jadbabaie. Complexity of finding stationary points of nonconvex nonsmooth functions. Proceedings of ICML 2020.