A unified analysis of descent sequences in weakly convex optimization, including convergence rates for bundle methods

Felipe Atenas[†] , Claudia Sagastizábal[†] , Paulo J. S. Silva[†] , and Mikhail Solodov[‡]

December 15, 2021

ABSTRACT

We present a framework for analyzing convergence and local rates of convergence of a class of descent algorithms, assuming the objective function is weakly convex. The framework is general, in the sense that it combines the possibility of explicit iterations (based on the gradient or a subgradient at the current iterate), implicit iterations (using a subgradient at the next iteration, like in the proximal schemes), as well as iterations when the associated subgradient is specially constructed and does not correspond neither to the current nor the next point (this is the case of descent steps in bundle methods). Under the subdifferential-based error bound on the distance to critical points, linear rates of convergence are established. Our analysis applies, among other techniques, to prox-descent for decomposable functions, the proximal-gradient method for a sum of functions, redistributed bundle methods, and a class of algorithms that can be cast in the feasible descent framework for constrained optimization.

Key words: weak convexity, descent methods, bundle methods, model-based methods, proximal descent, proximal gradient method, error bound, linear convergence.

AMS subject classifications: 90C30, 90C33, 90C55, 65K05

[†] IMECC/UNICAMP, Campinas, São Paulo, Brazil.

Emails: f262900@dac.unicamp.br, sagastiz/pjssilva@unicamp.br

[‡] IMPA – Instituto de Matemática Pura e Aplicada, Estrada Dona Castorina 110, Jardim Botânico, Rio de Janeiro, RJ 22460-320, Brazil.

Email: solodov@impa.br

1 Introduction

We consider algorithmically generated descent sequences that aim at solving problems of the form

$$\min\ f(x), \quad x \in \mathbb{R}^n\,, \tag{1}$$

where

$$f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\} \text{ is weakly convex.}$$

The class of weakly convex functions is fairly broad and covers many problems of interest. It includes convex functions, differentiable functions with Lipschitzian gradient, certain compositions of convex functions with smooth functions, among others. We refer the readers to the discussion in [13], and §2 below. The case of constrained optimization will be handled by including into the objective function the indicator function of the feasible set.

We are interested in stating conditions that ensure global convergence and local linear convergence rates for algorithms whose sequence of iterates $\{x_k\}$ involves the Clarke's subgradient information about f, possibly collected along iterations. Together with the algorithmically generated sequence $\{x_k\} \subseteq \mathbb{R}^n$, we shall also consider a certain theoretical sequence $\{z_k\} \subseteq \mathbb{R}^n$, with associated perturbation parameters $\{\varepsilon_k\} \subseteq [0, +\infty)$. These objects are introduced to account for the fact that, to compute the iterate x_k, one often minimizes a model/approximation of f. This operation yields a subgradient of the model, which for some methods in general is not a subgradient of f itself at any point in the sequence $\{x_k\}$. We show that model subgradients can, however, be "transported" to a nearby point, where they are subgradients of f. For convex functions, this is the well-known transportation formula in [23, Ch.XI, § 4.2]. For weakly convex functions, a similar result requires a delicate construction, given in § 5 below. In particular, we think of $\{z_k\}$ as a (potential) perturbation, not necessarily computed by the algorithm, of the actual sequence $\{x_k\}$ which is computed indeed.

Formally, we shall consider frameworks with the following relations valid (again, recall that $\{x_k\}$ is the generated sequence, while $\{z_k\}$ is a theoretical one):

$$f(x_k) + a(\|x_k - x_{k-1}\|^2 + \varepsilon_{k-1}) \le f(x_{k-1}), \text{ for } a > 0\,; \tag{2a}$$

$$\exists g_k \in \partial f(z_{k-1}) \cup \partial f(z_k),\ \|g_k\| \le b\Big(\|x_k - x_{k-1}\| + \|x_k - z_k\|\Big), \text{ for } b > 0\,; \tag{2b}$$

$$\text{both } \|x_k - z_k\| \text{ and } \{\varepsilon_k\} \text{ tend to 0 as } k \to \infty\,. \tag{2c}$$

Some remarks are in order. To start with, notice that condition (2a) ensures that the sequence of functional values $\{f(x_k)\}$ is non-increasing. By contrast, the theoretical sequence $\{f(z_k)\}$ is not necessarily non-increasing.

To continue, consider first the simplest instance, with $z_k = x_k$ and $\varepsilon_k = 0$. Then the conditions in (2c) are automatic, while (2b) becomes

$$\|g_k\| \le b\|x_k - x_{k-1}\|\,,$$

for some subgradient g_k of f at either x_{k-1} or x_k. In the first case, it is natural to think of the scheme as being explicit (one obvious example is the gradient descent iteration, if f is differentiable: $x_k = x_{k-1} - t_k \nabla f(x_{k-1})$, with a suitable stepsize $t_k > 0$). In the second case, the scheme is in general implicit, and becomes essentially that of [2, § 2.3] if further $g_k \in \partial f(x_k)$ is taken. A prototypical instance is given by the proximal point iteration:

$$x_k \in \arg\min f(x) + \frac{1}{2t_k}\|x - x_{k-1}\|^2, \text{ for } t_k > 0\,, \tag{3}$$

which means that $x_k = x_{k-1} - t_k g_k$, for some $g_k \in \partial f(x_k)$.

Next, note that in the nonsmooth case, even the convex one, an explicit scheme with $g_k \in \partial f(z_{k-1})$ and $z_{k-1} = x_{k-1}$ in (2b) does not guarantee the descent condition (2a). Indeed, this

would be just the basic subgradient method, which is not of descent. General-purpose algorithms for nonsmooth optimization that build descent sequences are bundle methods [28, 23, 6]. Other nonsmooth methods can also be of descent, if they use more specific problem structure. Some examples are the prox-descent method for composite functions [30] and proximal-gradient methods for sums [3], considered together with the bundle method in § 5 below. It is precisely for treating those type of methods that the theoretical iterate z_k and associated perturbation ε_k were introduced in our framework (2). Essentially, such schemes compute the proximal point of a convex *model* of the function f. Thanks to our transportation formula for weakly convex functions, this amounts to performing an explicit step, using a subgradient of f at a perturbed point, that plays the role of z_k in (2). This relation holds as long as the model-functions satisfy general conditions stated in § 5. Therein, the process is developed in full details for model-based proximal methods, including bundle algorithms for weakly convex functions.

Our convergence analysis recovers, from a unified perspective, various (but not necessarily all) results in sources like [32], [2], and [13]. We also give new results, related to bundle methods for weakly convex functions. As stated in the concluding section of [38], developing a convergence theory along the lines of [2] for bundle methods based on practical oracles was an open question. We close this gap in § 5, most notably by stating the linear convergence of bundle methods with downshifted models that are typical in the nonconvex setting; we refer to § 5.1 for details. When the objective in (1) is convex, linear rates for bundle-like methods can be traced back to [26] and [42]; see also the efficiency estimates in [27]. The topic was revisited more recently in [16] and [10], respectively considering strongly convex functions and multi-cut models, and the classical proximal bundle method for convex optimization.

The rest of the paper is organized as follows. In § 2 we collect some facts on weakly convex functions, to be used in the sequel. Error bounds are briefly discussed in § 3. We proceed in § 4 with some general global convergence and local linear rate of convergence properties of the framework given by (2). In § 5 these results are applied to model-based algorithms, including prox-descent for composite functions, proximal-gradient methods for sums, Taylor-based models, and finally the bundle methods. In § 6 we show how our analysis applies to the class of feasible descent methods for constrained optimization considered in [32].

We conclude this section with some notation and definitions. By $\langle \cdot, \cdot \rangle$ we denote the Euclidean inner product (where the space is always clear from the context), with $\| \cdot \|$ being the associated norm. By $B(0, \epsilon)$ we denote the closed ball of radius $\epsilon \geq 0$, centered at the origin.

For a proper, lower semicontinuous, and locally Lipschitz-continuous function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, the Clarke subdifferential of f at $x \in \operatorname{dom}(f) = \{x \in \mathbb{R}^n \mid f(x) < +\infty\}$ is given by

$$\partial f(x) = \operatorname{conv} \{g \in \mathbb{R}^n \mid \exists \{y_k\} \subset \mathcal{D}_f : y_k \to x, \ \nabla f(y_k) \to g\},$$

where $\mathcal{D}_f$ is the set of points at which f is differentiable, and where $\operatorname{conv} D$ stands for the convex hull of a set D. For $x \in \operatorname{dom}(f)$, $\partial f(x)$ is a nonempty closed convex set, and ∂f is an upper semicontinuous mapping; see [7]. Note also that from [36, Proposition 3.1], [36, Theorem 3.6] and Proposition 2.2(iv) below, the so-called limiting subdifferential and the Clarke subdifferential coincide for weakly convex functions, the class of our interest. This is the reason why we define and use the Clarke subdifferential only. If f is a smooth function, $\partial f(x)$ reduces to the gradient $\nabla f(x)$, while in the case of f being convex, $\partial f(x)$ is the subdifferential of f at x in the usual convex analysis sense. For f convex and $\varepsilon \geq 0$, the ε-subdifferential of f at $x \in \operatorname{dom}(f)$ is given by

$$\partial_\varepsilon f(x) = \{g \in \mathbb{R}^n \mid f(y) \geq f(x) + \langle g, y - x \rangle - \varepsilon \ \forall \, y \in \mathbb{R}^n\}.$$

For (possibly nonconvex) closed set D, we denote the associated (possibly set-valued) projection mapping by

$$P_D(x) = \arg\min_{y \in D} \|y - x\| .$$

Then the distance from x to D is given by

$$d(x, D) = \|x - p\|, \quad \text{for any } p \in P_D(x) .$$

For a nonempty closed convex set $X \subseteq \mathbb{R}^n$, i_X denotes the indicator function of the set X, i.e., $i_X(x) = 0$ if $x \in X$ and $i_X(x) = +\infty$ otherwise. The normal cone to X at $x \in X$ is given by $N_X(x) = \partial i_X(x) = \{u \in \mathbb{R}^n \mid \langle u, y - x\rangle \leq 0 \text{ for all } y \in X\}$. If $x \notin X$ then $N_X(x) = \emptyset$.

Denote by $S := (\partial f)^{-1}(0)$ the set of critical points of f, i.e., points x such that $0 \in \partial f(x)$. The following property [32, Condition B] will play a role in our analysis.

Definition 1.1 (Proper separation of isocost surfaces)**.** *A closed function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ has properly separated isocost surfaces if there exists $\varepsilon > 0$ such that*

$$\bar{x} \in S,\ \bar{y} \in S,\ f(\bar{x}) \neq f(\bar{y}) \implies \|\bar{x} - \bar{y}\| \geq \varepsilon\,.$$

This property is very natural; we refer the readers to [32] for a discussion and sufficient conditions for it to hold.

2 Weakly convex functions

We start with the definition.

Definition 2.1 (Weakly convex functions)**.** *We say that $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is $\rho-$weakly convex, for $\rho > 0$, if $f(\cdot) + \frac{\rho}{2}\|\cdot\|^2$ is a convex function.*

The class of weakly convex functions is contained in some larger classes of nonsmooth functions, such as the generalized differentiable functions in the sense of Norkin [39], or the semismooth functions [35].

The following are some equivalent characterizations of weak convexity; see [9, Theorem 2.1], [8, Theorem 3.1].

Proposition 2.2 (Alternative characterizations of weak convexity)**.** *For a lower semicontinuous function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ and $\rho > 0$, the following statements are equivalent:*

(i) For any $z \in \mathbb{R}^n$, $f(\cdot) + \frac{\rho}{2}\|\cdot - z\|^2$ is a convex function.

(ii) For any $x, y \in \mathbb{R}^n$, such that $\partial f(y) \neq \emptyset$, any $g(y) \in \partial f(y)$ satisfies

$$f(y) + \langle g(y), x - y\rangle \leq f(x) + \frac{\rho}{2}\|x - y\|^2$$

or, equivalently,

$$\ell_{y,g(y)}(x) \leq f(x) + \frac{\rho}{2}\|x - y\|^2\,,$$

where $\ell_{y,g(y)}(\cdot) := f(y) + \langle g(y), \cdot - y\rangle$ is the linearization of f at the point y.

(iii) For all $x, y \in \mathbb{R}^n$, and $\lambda > 0$,

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y) + \frac{\rho\lambda(1-\lambda)}{2}\|x - y\|^2\,.$$

Note that in Proposition 2.2(i), by taking $z = 0$, we retrieve Definition 2.1, which means that f is convex up to a quadratic perturbation. Proposition 2.2(i) is completely equivalent to this way of defining weakly convex functions, since it states that f is convex up to a quadratic perturbation with a linear term. Regarding some other notions of nonconvexity in the literature, it is important to note that for a function to be weakly convex, Proposition 2.2(ii) must hold for all subgradients at all points. By contrast, for prox-regular functions ([43, Definition 13.27]), also known as lower-$\mathcal{C}^2$ functions, the inequality holds only locally for subgradients, points and functional values. As a result, weak convexity is equivalent to the function being prox-regular everywhere, and the parameter of prox-regularity being the same for all points, or simply uniformly prox-regular.

As already commented, the class of weakly convex functions is quite broad and includes many settings of interest, whose nonconvexity is *benign*, in the parlance of [49]. One example is the class of decomposable functions in [45], that contains max-functions, maximal eigenvalue functions, and norm-1 regularized functions; see also [30] and [44].

Definition 2.3 ($h \circ c$ decomposable functions)**.** *Given a continuously differentiable mapping $c : \mathbb{R}^n \to \mathbb{R}^m$ such that $c(\bar{x}) = 0$, and a finite-valued positively homogeneous convex function $h : \mathbb{R}^m \to \mathbb{R}$, the real-valued function f is $h \circ c$ decomposable at $\bar{x} \in \mathbb{R}^n$, if there exists a neighborhood U of $\bar{x}$ such that for all $x \in U$,*

$$f(x) = f(\bar{x}) + h(c(x)) \,.$$

If c is a $\mathcal{C}^1$ function with Lipschitz-continuous Jacobian, then such f is weakly convex. To see this, apply [15, Lemma 4.2]. Since h is finite-valued and sublinear, it is then convex and Lipschitz-continuous (see [23, V(1.2.6)]), while c is $\mathcal{C}^1$ with Lipschitz-continuous Jacobian from the assumptions. Therefore the composition $h \circ c$ and, hence, the function $f(\cdot) = f(\bar{x}) + h \circ c(\cdot)$, are weakly convex.

In association with other notions related to weak convexity, we further remark that all real-valued prox-regular functions (or, in our terminology, real-valued locally weakly convex functions) can also be locally decomposed as a sum of a convex continuous function and a concave quadratic function (in line with Definition 2.1), and can also be expressed as a composition of a convex continuous function with a differentiable function with locally Lipschitz gradient, see [8, Proposition 3.5, Remark 3.6].

We next give an example of weak convexity for extended real-valued functions, that will play a role in § 6 to include the class of feasible descent methods of [32] (for constrained optimization) into our framework.

Proposition 2.4. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function with $L-$Lipschitz continuous gradient on the nonempty closed convex set $X \subseteq \mathbb{R}^n$. Then, $f + i_X$ is a $L-$weakly convex function.*

Proof. Since f has Lipschitz-continuous gradient with constant L on X, then (e.g., from [25, Lemma A.11]), for all $x, y \in X$ it holds that

$$f(y) + \langle \nabla f(y), x - y \rangle - \frac{L}{2}\|x - y\|^2 \leq f(x) \,.$$

Furthermore, for $x \in X$, and $y \in \mathbb{R}^n$ such that $\partial(f + i_X)(y) \neq \emptyset$, that is, for $y \in X$, and for all $w \in N_X(y)$, we have that $\nabla f(y) + w \in \partial(f + i_X)(y)$, and

$$(f + i_X)(y) + \langle \nabla f(y) + w, x - y \rangle - \frac{L}{2}\|x - y\|^2 \leq (f + i_X)(x) \,.$$

If $x \notin X$, the above inequality holds trivially, because y needs to be an element of X to ensure that the subdifferential $\partial(f + i_X)(y)$ is nonempty (see Proposition 2.2(ii)). Therefore, $f + i_X$ is $L-$weakly convex. □

3 Error bounds

Error bounds are (upper) estimates of the distance to solutions (or critical points) of a given problem. Their role is paramount for various reasons, among which is convergence rate analyses; see, e.g., [41, 17, 25, 51].

In this work, we shall mostly employ the following subdifferential-based error bound. See, however, the end of this section for the so-called *natural residual* error bound [17] for constrained problems, and its relation with the subdifferential-based bound.

Definition 3.1 (Subdifferential error bound)**.** *We say that the subdifferential error bound holds for problem (1) where $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is bounded below, if for every $v \geq \inf_{x \in \mathbb{R}^n} f(x)$, there exist $\epsilon, \ell > 0$ such that whenever $x \in \mathbb{R}^n$, $f(x) \leq v$, and $w \in \partial f(x) \cap B(0, \epsilon)$, the following is true:*

$$d(x, S) \leq \ell \|w\| \,, \quad \text{where } S = (\partial f)^{-1}(0) \text{ is the set of critical points of } f.$$

The error bound above is related to various other notions that appear in the literature, such as the Kurdyka-Łojasiewicz inequality [29, 5], and quadratic growth of f around the set of its critical points [11, 12], or the set of minimizers when the function is convex [1, 50]. These conditions assure some regularity of the function near a critical point. Furthermore, the subdifferential error bound is related to metric subregularity of ∂f (see [24, Definition 3.17]).

Note that the error bound in Definition 3.1 uses the Clarke subdifferential, while in [2] the limiting subdifferential appears. As mentioned, for weakly convex functions these two subdifferentials coincide.

Another comment is that weak convexity, combined with the subdifferential error bound, entails the Łojasiewicz inequality with exponent $\theta = \frac{1}{2}$ [11, Proposition 3.8], and quadratic growth around the set of critical points [12, Theorem 3.1]. Furthermore, [11, Theorem 3.7] states that a variant of the Łojasiewicz inequality implies the error bound. In the convex setting, the quadratic growth condition is actually equivalent to the subdifferential error bound [1, Theorem 3.3].

We next turn our attention to constrained smooth optimization problems, the framework of [32], dealt with in § 6. Consider the problem

$$\min_{x \in X} \; f(x)\,, \tag{4}$$

where X is a closed convex set, and $f : \mathbb{R}^n \to \mathbb{R}$ is finite-valued and smooth. An equivalent problem is to handle constraints by adding to f the indicator function of the feasible set. It turns out that these two equivalent formulations are in fact different when it comes to error bounds, and some subtle issues arise.

Specifically, as is well known, criticality of a point x in the sense of

$$0 \in \partial(f + i_X)(x) = \nabla f(x) + N_X(x)$$

is equivalent to the condition

$$x - P_X(x - \nabla f(x)) = 0\,.$$

Hence, one can attempt to measure the distance to the set of critical points S by the violation of the projection equality above, or by the violation of the subdifferential inclusion above. It so happens that, at least in general, these are not the same. We next review the relations between the corresponding error bounds.

The subdifferential error bound would just read exactly the same as in Definition 3.1, redefining therein $f := f + i_X$ (then $w \in \nabla f(x) + N_X(x)$). The projection-based error bound states the following.

Definition 3.2 (Projection error bound)**.** *We say that the projection error bound holds for problem (4) where $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable and bounded below, if for every $v \geq \inf_{x \in X} f(x)$, there exist $\epsilon, \ell > 0$ such that whenever $x \in X$, $f(x) \leq v$, and $\|x - P_X(x - \nabla f(x))\| \leq \epsilon$, the following is true:*

$$d(x, S) \leq \ell \|x - P_X(x - \nabla f(x))\|\,.$$

The projection error bound is a natural way to measure violation of stationarity in convexly-constrained problems, used in many developments; see, e.g., [32, 48, 47].

Clearly, for problem (4) with smooth f, Definitions 3.1 and 3.2 amount to the same if $X = \mathbb{R}^n$ (or if S is in the interior of X). For constrained problems, there are two cases when these error bounds are equivalent. The first case is when the critical point is isolated, see [17, Proposition 6.2.4], [25, Proposition 1.31]. In that case, the projection error bound means the semistability property [25, Definition 1.29]. The second case when the two bounds are equivalent is when X is a generalized box in $\mathbb{R}^n$, i.e., X is defined by bound constraints on the variables (some bounds can be infinite), see [18, Theorem 2]. To the best of our knowledge, in other cases the relations between the subdifferential and projection error bounds are not known. However, the following simple argument shows that when the gradient of f is Lipschitz-continuous, the projection residual is bounded above by a multiple of $d(x, S)$, always. Then, if the subdifferential error bound holds, the right-hand side therein is of order no less than the projection residual. Hence, in principle,

the subdifferential error bound can hold when the projection one does not. Indeed, for each x let $p(x) \in P_S(x)$. Then,

$$\begin{aligned} \|x - P_X(x - \nabla f(x))\| &= \|x - P_X(x - \nabla f(x)) - (p(x) - P_X(p(x) - \nabla f(p(x)))\| \\ &\leq \|x - p(x)\| + \|P_X(x - \nabla f(x)) - P_X(p(x) - \nabla f(p(x)))\| \\ &\leq d(x, S) + \|x - \nabla f(x) - (p(x) - \nabla f(p(x)))\| \\ &\leq (2 + L)d(x, S)\,, \end{aligned}$$

where L is the Lipschitz constant of the gradient of f.

4 General asymptotic relations in the algorithmic pattern

In the sequel, we shall need the following technical result.

Lemma 4.1. *Let $\{a_k\} \subseteq \mathbb{R}^n$ and $\{b_k\} \subseteq [0, +\infty)$ be two sequences such that for all k it holds:*

$$\|a_k - a_{k-1}\| \leq \alpha_1 b_{k-1}$$

and

$$b_k \leq \alpha_2 b_{k-1}\,,$$

where $\alpha_1 > 0$ and $\alpha_2 \in (0, 1)$.

Then, there exists $a^* \in \mathbb{R}^n$ such that, for any $\bar{k}$, there exist $r \in (0, 1)$ and $c > 0$, such that for all $k \geq \bar{k}$,

$$\|a_k - a^*\| \leq c\alpha_2^k$$

with $c = \dfrac{\alpha_1 b_0}{1 - \alpha_2}$. In particular, $\{a_k\}$ converges to a^* R-linearly.

Proof. First, by direct induction, for all k it holds that $b_k \leq b_0(\alpha_2)^k$. By making a telescopic sum, for all $j \geq 1$,

$$\|a_{k+j} - a_k\| \leq \sum_{n=k+1}^{k+j} \|a_n - a_{n-1}\| \leq \frac{\alpha_1 b_0}{\alpha_2} \sum_{n=k+1}^{k+j} (\alpha_2)^n \leq \left(\frac{\alpha_1 b_0}{1 - \alpha_2}\right)(\alpha_2)^k\,, \tag{5}$$

where to obtain the last inequality we use that

$$\sum_{n=k+1}^{k+j} (\alpha_2)^n = (\alpha_2)^k \sum_{n=1}^{j} (\alpha_2)^n \leq (\alpha_2)^k \frac{\alpha_2}{1 - \alpha_2}\,,$$

since $\alpha_2 \in (0, 1)$. Therefore, $\{a_k\} \subseteq \mathbb{R}^n$ is a Cauchy sequence, and thus $\{a_k\}$ converges to some a^*. By taking the limit in (5) when $j \to \infty$, we obtain that $\|a_k - a^*\| \leq c\alpha_2^k$, as claimed. □

Regarding our problem of interest, if f in (1) is bounded below, the monotonically non-increasing sequence $\{f(x_k)\}$ from (2) converges, without any further assumptions (to some value, not necessarily a critical one). We next show that, for weakly convex functions satisfying the sub-differential error bound of Definition 3.1 and the isocost surfaces condition of Definition 1.1, the sequence of functional values of the projections of the theoretical sequence $\{z_k\}$ onto S stabilizes at a critical value (value of f at a critical point).

In the statements (iv) and (v) below, the index $\mathtt{i} \in \{0, 1\}$ is used to unify the analysis for explicit and implicit options in (2). Specifically, $\mathtt{i} = 1$ refers to explicit methods ($z_{k-\mathtt{i}} = z_{k-1}$, so that $g_k \in \partial f(z_{k-1})$), while $\mathtt{i} = 0$ refers to implicit methods ($z_{k-\mathtt{i}} = z_k$, so that $g_k \in \partial f(z_k)$).

Lemma 4.2 (Convergence to critical points and technical relations)**.** *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a $\rho-$weakly convex function, such that $\inf f > -\infty$. Then for any algorithmic scheme satisfying (2), the following holds:*

(i) $\{f(x_k)\}$ *monotonically converges to some value* $\tilde{f} \in \mathbb{R}$.

(ii) $x_k - x_{k-1} \to 0$, $z_k - z_{k-1} \to 0$ *and* $g_k \to 0$, *as* $k \to +\infty$.

Suppose, in addition, that f *satisfies the proper separation of isocost surfaces condition (Definition 1.1) and the subdifferential error bound (Definition 3.1). Then,*

(iii) $\{f(z_k)\}$ *converges to* f^*, *where* $f^* \in \mathbb{R}$ *is a critical value (i.e.,* $f^* = f(x)$ *for some* $x \in S$*).*

(iv) For $\mathtt{i} \in \{0,1\}$, *defining* $\tilde{p}_{k-i} \in P_S(z_{k-i})$, *for all* k *sufficiently large the distance from* $z_{k-\mathtt{i}}$ *to* S *can be estimated as*

$$\|z_{k-i} - \tilde{p}_{k-i}\|^2 \leq \frac{2\ell^2 b^2}{a}(f(x_{k-1}) - f(x_k)) + 2\ell^2 b^2 \|x_k - z_k\|^2 .$$

(v) For the functional value errors $v_k := f(x_k) - f^*$, *it holds that*

$$v_k \leq \frac{2\ell b^2}{a}(v_{k-1} - v_k) + 2\ell b^2 \|x_k - z_k\|^2 + \Theta_{k-i} ,$$

where

$$\Theta_{k-i} := f(x_{k-i}) - f(z_{k-i}) + \frac{\rho}{2}\|\tilde{p}_{k-i} - z_{k-i}\|^2 .$$

Proof. In view of (2a) and $\varepsilon_k \geq 0$, $\{f(x_k)\}$ is non-increasing. Since f is bounded below, item (i) follows immediately. Then also $f(x_{k-1}) - f(x_k) \to 0$.

As, by (2a),

$$\|x_k - x_{k-1}\|^2 \leq \frac{1}{a}(f(x_{k-1}) - f(x_k)) - \varepsilon_{k-1} , \tag{6}$$

it follows that $x_k - x_{k-1} \to 0$ in item (ii). Then (2b) and (2c) yield that $g_k \to 0$ and, $z_k - z_{k-1} = (z_k - x_k) + (x_k - x_{k-1}) + (x_{k-1} - z_{k-1}) \to 0$. Item (ii) is proven.

For the remaining items, we apply the subdifferential error bound at the tail of the auxiliary sequence $\{z_k\}$. The starting point is (2b), for which we use that f is a $\rho-$weakly convex function, considering the two possibilities $\mathtt{i} = 0$ and $\mathtt{i} = 1$ at the same time.

Since $g_k \in \partial f(z_k) \cup \partial f(z_{k-1})$, for $\mathtt{i} \in \{0,1\}$ it holds that

$$f(z_{k-\mathtt{i}}) + \langle g_k, x_k - z_{k-\mathtt{i}} \rangle \leq f(x_k) + \frac{\rho}{2}\|z_{k-\mathtt{i}} - x_k\|^2 .$$

In view of the fact that $f(x_k)$ decreases to $\tilde{f}$, $g_k \to 0$, $z_{k-\mathtt{i}} - x_k \to 0$, we have that for all $\epsilon > 0$, and all sufficiently large k, $f(z_{k-\mathtt{i}}) \leq \tilde{f} + \epsilon$ and $g_k \in \partial f(z_{k-\mathtt{i}}) \cap B(0,\epsilon)$. Thus, by the error bound,

$$\|z_{k-i} - \tilde{p}_{k-i}\| = d(z_{k-i}, S) \leq \ell \|g_k\| , \tag{7}$$

for $\mathtt{i} \in \{0,1\}$.

Since $g_k \to 0$, it follows from (7) that $z_{k-\mathtt{i}} - \tilde{p}_{k-\mathtt{i}} \to 0$, and then $z_k - \tilde{p}_k \to 0$ as $k \to +\infty$. Combining this with the fact that $z_k - z_{k-1} \to 0$, yields that $\tilde{p}_k - \tilde{p}_{k-1} \to 0$. Moreover, the property of separation of the isocost surfaces implies that $f(\tilde{p}_k) = f^*$ eventually, for a critical value f^* of f. To complete the proof of item (iii), we apply weak convexity of f for $0 \in \partial f(\tilde{p}_k)$, obtaining that for all sufficiently large k it holds that

$$f^* = f(\tilde{p}_k) + \langle 0, z_k - \tilde{p}_k \rangle \leq f(z_k) + \frac{\rho}{2}\|z_k - \tilde{p}_k\|^2 .$$

Hence,

$$-\frac{\rho}{2}\|z_k - \tilde{p}_k\|^2 \quad \leq \quad f(z_k) - f^* . \tag{8}$$

Notice that, in addition, $g_k \in \partial f(z_{k-\mathtt{i}})$ implies, for $\mathtt{i} \in \{0,1\}$, that

$$f(z_{k-\mathtt{i}}) + \langle g_k, \tilde{p}_{k-\mathtt{i}} - z_{k-\mathtt{i}} \rangle \leq f(\tilde{p}_{k-\mathtt{i}}) + \frac{\rho}{2}\|\tilde{p}_{k-\mathtt{i}} - z_{k-\mathtt{i}}\|^2 = f^* + \frac{\rho}{2}\|\tilde{p}_{k-\mathtt{i}} - z_{k-\mathtt{i}}\|^2 , \tag{9}$$

where the last equality holds for all k sufficiently large.

Next, combining (8) and (9), we obtain that

$$-\frac{\rho}{2}\|z_k - \tilde{p}_k\|^2 \leq f(z_k) - f^* \leq \langle g_{k+\mathtt{i}}, z_k - \tilde{p}_k\rangle + \frac{\rho}{2}\|\tilde{p}_k - z_k\|^2 .$$

Then, taking the limit as $k \to \infty$ yields that $f(z_k) \to f^*$.

Next, weak convexity implies that for $\mathtt{i} \in \{0,1\}$ and any $d \in \partial f(x_{k-\mathtt{i}})$,

$$f(x_{k-\mathtt{i}}) + \langle d, z_{k-\mathtt{i}} - x_{k-\mathtt{i}}\rangle \leq f(z_{k-\mathtt{i}}) + \frac{\rho}{2}\|z_{k-\mathtt{i}} - x_{k-\mathtt{i}}\|^2 .$$

Also, as $g_k \in \partial f(z_{k-\mathtt{i}})$,

$$f(z_{k-\mathtt{i}}) + \langle g_k, x_{k-\mathtt{i}} - z_{k-\mathtt{i}}\rangle \leq f(x_{k-\mathtt{i}}) + \frac{\rho}{2}\|x_{k-\mathtt{i}} - z_{k-\mathtt{i}}\|^2 .$$

Combining the two relations above, we obtain that

$$\langle g_k, x_{k-\mathtt{i}} - z_{k-\mathtt{i}}\rangle - \frac{\rho}{2}\|x_{k-\mathtt{i}} - z_{k-\mathtt{i}}\|^2 \leq f(x_{k-\mathtt{i}}) - f(z_{k-\mathtt{i}}) \leq \langle d, x_{k-\mathtt{i}} - z_{k-\mathtt{i}}\rangle + \frac{\rho}{2}\|z_{k-\mathtt{i}} - x_{k-\mathtt{i}}\|^2 .$$

Taking the limit in the last relation as $k \to +\infty$, Lemma 4.2(ii) and (2c) imply that $f(x_{k-\mathtt{i}}) - f(z_{k-\mathtt{i}}) \to 0$. Since $\{f(x_k)\}$ is a convergent sequence, and $f(z_k) \to f^*$, the sequences $\{f(x_k)\}$ and $\{f(z_k)\}$ both have the same limit. Thus, $\{f(x_k)\}$ is a non-increasing sequence converging to the critical value f^*, and $\{v_k\}$ is a nonnegative sequence.

To show statements (iv) and (v), recall that $(a+b)^2 \leq 2a^2 + 2b^2$, for all real numbers a, b. Then from (2b) we obtain that

$$\begin{aligned} \|g_k\|^2 &\leq b^2(\|x_k - x_{k-1}\| + \|x_k - z_k\|)^2 \\ &\leq 2b^2\|x_k - x_{k-1}\|^2 + 2b^2\|x_k - z_k\|^2 \\ &\leq \frac{2b^2}{a}(f(x_{k-1}) - f(x_k)) + 2b^2\|x_k - z_k\|^2 , \end{aligned} \tag{10}$$

where the last inequality follows from (6). In this manner, since $g_k \in \partial f(z_{k-\mathtt{i}})$ for $\mathtt{i} \in \{0,1\}$, from (7) and (10) it follows that

$$\|z_{k-\mathtt{i}} - \tilde{p}_{k-\mathtt{i}}\|^2 \leq \frac{2\ell^2 b^2}{a}(f(x_{k-1}) - f(x_k)) + 2\ell^2 b^2\|x_k - z_k\|^2 ,$$

which is statement (iv).

On the other hand, from (9), (7), and the fact that for all sufficiently large k it holds that $f(\tilde{p}_{k-\mathtt{i}}) = f^*$, we obtain that

$$\begin{aligned} f(z_{k-\mathtt{i}}) - f^* &\leq \|g_k\|\|z_{k-\mathtt{i}} - \tilde{p}_{k-\mathtt{i}}\| + \frac{\rho}{2}\|\tilde{p}_{k-\mathtt{i}} - z_{k-\mathtt{i}}\|^2 \\ &\leq \ell\|g_k\|^2 + \frac{\rho}{2}\|\tilde{p}_{k-\mathtt{i}} - z_{k-\mathtt{i}}\|^2 , \end{aligned}$$

for $\mathtt{i} \in \{0,1\}$. Therefore, combining this inequality with (10), yields

$$f(z_{k-\mathtt{i}}) - f^* \leq \frac{2\ell b^2}{a}(f(x_{k-1}) - f(x_k)) + 2\ell b^2\|x_k - z_k\|^2 + \frac{\rho}{2}\|\tilde{p}_{k-\mathtt{i}} - z_{k-\mathtt{i}}\|^2 .$$

Hence, as $\{v_k\}$ is non-increasing,

$$v_k \leq \frac{2\ell b^2}{a}(v_{k-1} - v_k) + f(x_{k-\mathtt{i}}) - f(z_{k-\mathtt{i}}) + 2\ell b^2\|x_k - z_k\|^2 + \frac{\rho}{2}\|\tilde{p}_{k-\mathtt{i}} - z_{k-\mathtt{i}}\|^2 .$$

This concludes the proof. □

The relations in Lemma 4.2 lead to the following result, on the convergence speed of both the sequence of functional values and of iterates. The respective rates are linear in the quotient (Q) and root (R) senses, as defined in [40].

Recall that the index $\mathtt{i} \in \{0, 1\}$ unifies the explicit and implicit options in (2).

Theorem 4.3 (Asymptotic results for weakly convex functions)**.** *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a $\rho-$weakly convex function such that $\inf f > -\infty$. Suppose, in addition, that f satisfies the proper separation of isocost surfaces condition (Definition 1.1) and the subdifferential error bound (Definition 3.1).*

Let $\{x_k\}$ and $\{z_k\}$ satisfy (2), and consider the sequence of functional errors $\{v_k\}$, defined in Lemma 4.2(v). If there exist $C_1\,, C_2 > 0$, such that, for all sufficiently large k it holds that

$$f(x_{k-i}) - f(z_{k-i}) \leq C_1(v_{k-1} - v_k) \tag{11}$$

and

$$\|x_k - z_k\|^2 \leq C_2(v_{k-1} - v_k)\,, \tag{12}$$

then there exist $q \in (0, 1)$ and $c > 0$ such that

(i) For all k sufficiently large,

$$v_k \leq q v_{k-1}\,,$$

where $q = M/(1 + M) \in (0, 1)$, and $M = C_1 + \ell b^2(2 + \rho\ell)(1/a + C_2)$.

(ii) The sequence of functional errors $\{v_k\}$ monotonically converges to 0 with $Q-$linear rate.

(iii) The sequence $\{x_k\}$ converges $R-$linearly to a critical point x^* of f, such that $f(x^*) = f^* = \lim_{k\to\infty} f(x_k)$. More specifically, for all sufficiently large k,

$$\|x_k - x^*\| \leq c\sqrt{q}^k\,,$$

where $c = \dfrac{\sqrt{v_0}}{\sqrt{a}\,(1 - \sqrt{q})}$.

Proof. First, convergence of $\{f(x_k)\}$ follows from Lemma 4.2(iii). The rate of convergence of $\{f(x_k)\}$ is derived from the technical estimates of Lemma 4.2. Indeed, combining the definition of $\Theta_{k-\mathtt{i}}$ with Lemma 4.2(iv) and (11), for all sufficiently large k it holds that

$$\begin{aligned}
\Theta_{k-\mathtt{i}} &\leq C_1(v_{k-1} - v_k) + \frac{\rho}{2}\left(\frac{2\ell^2 b^2}{a}(v_{k-1} - v_k) + 2\ell^2 b^2 \|x_k - z_k\|^2\right) \\
&= \left(C_1 + \frac{\rho\ell^2 b^2}{a}\right)(v_{k-1} - v_k) + \rho\ell^2 b^2 \|x_k - z_k\|^2\,.
\end{aligned}$$

Therefore, from Lemma 4.2(v), it further follows that

$$\begin{aligned}
v_k &\leq \left(C_1 + \frac{\ell b^2}{a}(2 + \rho\ell)\right)(v_{k-1} - v_k) + \ell b^2(2 + \rho\ell)\|x_k - z_k\|^2 \\
&\leq \left(C_1 + \frac{\ell b^2}{a}(2 + \rho\ell)\right)(v_{k-1} - v_k) + \ell b^2(2 + \rho\ell)C_2(v_{k-1} - v_k)\,,
\end{aligned}$$

where (12) is used to obtain the last inequality. Hence, $v_k \leq M(v_{k-1} - v_k)$, which gives item (i) with M specified therein.

Using inductively the inequality of item (i), we conclude that there exists $c > 0$ such that for $q = M/(1 + M)$ and all sufficiently large k,

$$v_k \leq cq^k.$$

To see item (iii), the estimate therein follows from Lemma 4.1. More specifically, there exists a point x^* such that $\{x_k\}$ converges to x^* R-linearly. In particular, from (2c), $\{z_{k-\mathtt{i}}\}$ also converges

to x^*, for $\mathtt{i} \in \{0,1\}$. Note that, since ∂f is an upper semicontinuous multifunction, Lemma 4.2(ii) and (2b) imply that $\partial f(z_{k-\mathtt{i}}) \ni g_k \to 0$, therefore $0 \in \partial f(x^*)$, that is, x^* is a critical point.

Finally, $z_k - \tilde{p}_k \to 0$ implies that $\tilde{p}_k \to x^*$, that is, x^* and $\tilde{p}_k$ are sufficiently close critical points. Therefore, in view of the proper separation of isocost surfaces property, $f(x^*) = f^*$. Hence, the limit of $\{x_k\}$ is a critical point $x^* \in f^{-1}(f^*)$. □

In the final two sections, Theorem 4.3 is applied to show the linear convergence rate of two different families of algorithms, proximal model-based ones akin to bundle methods, and the feasible descent framework of [32].

5 Bundle and proximal model-based methods

In nonsmooth optimization, satisfaction of (2a) is not straightforward. In addition to its role in Theorem 4.3, in this section weak convexity is an important ingredient in showing that iteratively minimizing appropriate approximating models of f indeed generates sequences that are of descent.

Suppose, for the moment, that f is a convex nonsmooth function. In this case, neither subgradient nor cutting-plane methods [6, Part II] fit the algorithmic pattern (2), because they do not guarantee the descent condition (2a). By contrast, as we shall show, serious steps within a bundle method do satisfy all the requirements. Bundle methods provide an implementable alternative for functions whose proximal point computation in (3) is difficult (or impossible). Before briefly reviewing the basic bundling mechanism, we mention that even for smooth functions, computing proximal points of some approximations of f has proven to be a useful technique to exploit decomposable structures. This is the basis of a plethora of approaches, including ADMM, as well as the prox-linear and prox-gradient methods considered below.

Having at hand a family of convex *model functions* for which computing proximal points is computationally implementable, in a bundle method [6, Part II] a candidate iterate is defined as the proximal point of the model function at x_{k-1}. If the candidate satisfies a condition of sufficient descent for f, it is labeled a *serious step* x_k, and (2a) holds; otherwise the candidate is declared a *null step*. At a new iteration, the bundling process improves the model function and/or adjusts the proximal parameter. By this token, at serious steps the approximation of the proximal point is sufficiently good to ensure that errors incurred when replacing f by its model satisfy (2c).

For a convex f, a key ingredient in the convergence analysis of bundle methods is to relate the model subgradient associated with the prox-computation to certain ε-subgradient of f. The nonconvex setting precludes the use of approximate subdifferentials in this part of the analysis. For this reason, different ad-hoc approaches have been proposed in the literature. Rather than singling out some specific approach, below we develop a general convergence theory that is applicable to weakly convex functions. The key is to complement the algorithmic pattern of (2) with a suitable condition on the model functions used to approximate the proximal point of f. Our proposal unifies the global convergence analysis of a wide variety of methods in the literature, and also provides their linear rate of convergence.

5.1 Model function assumptions

Approximating the proximal point scheme (3) involves defining a family of simpler (than f) model functions whose proximal point is computed at each iteration. Often, a trade-off must be found between simplicity (fast prox-computation) and accuracy (increased chances of accepting the candidate as a serious step, i.e., satisfying (2a)).

Given $x \in \mathbb{R}^n$, consider modelling the function $f - f(x)$ by a convex function $\varphi_x : \mathbb{R}^n \to \mathbb{R}$. Note that f might be extended real-valued, while its model is finite everywhere. The most synthetic model uses the linearization introduced in Proposition 2.2,

$$\varphi_x^{\mathtt{sg}}(\cdot) = \ell_{x,g(x)}(\cdot) - f(x)\,.$$

Incidentally, computing the proximal point of this model amounts to one subgradient iteration, with stepsize given by the inverse of the prox-parameter.

A cutting-plane model is richer, as it takes the maximum over several linearizations, generated with past iterates x_i for $i \in \mathbb{B}$, the *bundle*:

$$\varphi_x^{\mathtt{cp}}(\cdot) \;:=\; \max_{i\in\mathbb{B}} \left\{\ell_{x_i,g(x_i)}(\cdot) - f(x)\right\} = \max_{i\in\mathbb{B}} \left\{-e_i(x) + \langle g(x_i), \cdot - x\rangle\right\} , \\ \text{where we define } e_i(x) := f(x) - \ell_{x_i,g(x_i)}(x) .$$

The term $e_i(x)$, called *linearization error* in the bundle terminology, measures the quality of the linearization with respect to the reference point x. For convex f, the error is nonnegative and the cutting-plane model satisfies $\varphi_x^{\mathtt{cp}} \le f - f(x)$. But for nonconvex f this inequality cannot be ensured. To address this drawback, a common approach is to downshift negative linearization errors, making them nonnegative. This can be done in different ways; typically,

the term $e_i(x)$ is replaced by $e_i^q(x) := \max\{e_i(x), \frac{q}{2}\|x_i - x\|^2\}$ for $q > 0$ sufficiently large;

see [34, 28] and, more recently, [33, 31, 19, 37]. The approach in [21, 20, 22] differs from those works, as it handles nonconvexity using *redistributed* models that, in addition to downshifting, tilt the slopes, as in Proposition 5.2 below.

In order to account for many alternative models in the literature, we shall assume that the family of model functions satisfies the following property. In the sequel, we shall show that it holds for many methods of interest.

Definition 5.1 (Models 1QA)**.** *A convex proper function $\varphi_x : \mathbb{R}^n \to \mathbb{R}$ is said to model f at x with one-sided quadratical accuracy, if*

$$\exists q > 0 : \forall y \in \mathbb{R}^n \quad \varphi_x(y) \le f(y) - f(x) + \frac{q}{2}\|y - x\|^2 . \tag{13}$$

The property 1QA is a weakened form of the two-sided models considered in [14] and [11]. Making the condition unilateral is crucial for including bundle methods in the analysis (even when f is convex; see Figure 1 for an illustration).

The key role of convex 1QA models φ_x in convergence analyses is that they allow to *transport* subgradients, a mechanism that is not available for the nonconvex function f directly. Also, 1QA models are quite general, as the condition (13) can be satisfied both by cutting-plane-like models, where linearizations are oblivious to possible further information about f, and also by models that use structure. When a function has known structure, it is appealing to make the model inherit some of this feature. We next provide some examples.

5.1.1 Models defined using linearizations

For weakly convex functions, the simplest model $\varphi_x^{\mathtt{sg}}$ is clearly 1QA , taking $q = \rho$, the weak convexity parameter, but as already commented, the descent condition (2a) is not guaranteed for such a model, as it gives just a subgradient iteration. By contrast, the cutting-plane model with downshifted errors satisfies (13), as long as the iterates remain in a bounded set. The case of the more sophisticated model from [21, 20, 22] is analyzed below.

Proposition 5.2 (Redistributed models are 1QA)**.** *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a weakly convex function with parameter ρ and let $x \in \mathbb{R}^n$. Given bundle elements $x_i, f(x_i), g(x_i) \in \partial f(x_i)$ for $i \in \mathbb{B}$, consider the downshifted linearization errors and tilted subgradients, respectively defined by*

$$e_i^\rho(x) := f(x) - \ell_{x_i,g(x_i)}(x) + \frac{\rho}{2}\|x - x_i\|^2 \quad \text{and} \quad g_i^\rho(x) := g(x_i) - \rho(x - x_i) .$$

Then the associated model $\varphi_x^\rho(\cdot) := \max_{i\in\mathbb{B}} \{-e_i^\rho(x) + \langle g_i^\rho(x), \cdot - x\rangle\}$ is 1QA .

Proof. The model is convex, as the maximum of affine functions.

For any bundle element, weak convexity implies that, for all y,

$$f(y) + \frac{\rho}{2}\|y - x_i\|^2 \ge \ell_{x_i,g(x_i)}(y) = f(x_i) + \langle g(x_i), y - x_i\rangle .$$

Since $e_i(x) = f(x) - \ell_{x_i,g(x_i)}(x)$, rearranging terms, we obtain that

$$f(y) - f(x) \geq -e_i(x) + \langle g(x_i), y - x\rangle - \frac{\rho}{2}\|y - x_i\|^2 .$$

Adding $\frac{\rho}{2}\|y - x\|^2$ to both sides yields

$$f(y) - f(x) + \frac{\rho}{2}\|y - x\|^2 \geq -e_i(x) + \langle g(x_i), y - x\rangle + \frac{\rho}{2}\Big(\|y - x\|^2 - \|y - x_i\|^2\Big) .$$

As $\frac{\rho}{2}\Big(\|y - x\|^2 - \|y - x_i\|^2\Big) = -\frac{\rho}{2}\|x - x_i\|^2 - \langle \rho(x - x_i), y - x\rangle$, it follows that

$$\begin{array}{rcl} f(y) - f(x) + \frac{\rho}{2}\|y - x\|^2 & \geq & -\Big(e_i(x) + \frac{\rho}{2}\|x - x_i\|^2\Big) + \Big\langle \Big(g(x_i) - \rho(x - x_i)\Big), y - x\Big\rangle \\ & = & -e_i^\rho(x) + \langle g_i^\rho(x), y - x\rangle . \end{array}$$

Since each of the terms defining the model φ_x^ρ satisfies (13), so does the model. □

In the redistributed proximal bundle method [20] iterates are generated with a model $\varphi_x^{\rho_k}$ whose augmentation parameter ρ_k is updated along the process, without knowing ρ beforehand. It is shown in [21] that unless x_{k-1} is critical, the procedure generates a serious step after a finite number of null iterations for weakly convex functions (f is uniformly prox-bounded in the language of that work). In [20] the serious step sequence is shown to be globally convergent under the same assumptions. Thanks to the theory developed in § 5.2, based in Theorem 4.3, in addition to global convergence, we can now prove that serious steps converge at the linear rate. To the best of our knowledge, this is the first result on linear convergence rates for nonconvex bundle methods.

5.1.2 Decomposable functions, prox-descent and composite bundle methods

Recalling Definition 2.3, for decomposable functions $f = h \circ c$ the *ProxDescent* iterates [30, Algorithm 1] are defined by computing the proximal point of the model that is created by replacing the smooth mapping c with its Taylor expansion:

$$\varphi_x^{\texttt{LW}}(\cdot) := h(c(x) + \nabla c(x)^\top(\cdot - x)) - f(x) .$$

In [11], the associated method is called *prox-linear*. We next show that the model $\varphi_x^{\texttt{LW}}$ is 1QA under our assumptions (it should be noted that in [30] the outer function h can be more general, specifically extended-valued prox-regular).

Proposition 5.3 (Models for decomposable functions are 1QA)**.** *Let $h : \mathbb{R}^m \to \mathbb{R}$ be convex, finite-valued and positively homogeneous, and let $c : \mathbb{R}^n \to \mathbb{R}^m$ be continuously differentiable with its Jacobian being Lipschitz-continuous.*

Then the model $\varphi_x^{\texttt{LW}}$ is 1QA.

Proof. Under the stated assumptions, $\varphi_x^{\texttt{LW}}$ is convex.

As h is convex positive homogeneous and finite, it is the support function of a compact convex set D (that coincides with its subdifferential at 0), see [23, Chapter V] or [43, Theorem 8.24]. That is

$$h(d) = \max_{s\in D}\langle s, d\rangle.$$

Moreover, let L be the Lipschitz constant of the Jacobian of c. It follows that, for all $y, x \in \mathbb{R}^n$,

$$\|c(y) - c(x) - \nabla c(x)^\top(y - x)\| \leq \frac{L}{2}\|y - x\|^2.$$

Hence,

$$\begin{aligned} h(c(x)+\nabla c(x)^\top(y-x)) &= h(c(x)+\nabla c(x)^\top(y-x)-c(y)+c(y)) \\ &= \max_{s\in D}\langle s, c(x)+\nabla c(x)^\top(y-x)-c(y)+c(y)\rangle \\ &\le \max_{s\in D}\langle s, c(y)\rangle + \max_{s\in D}\langle s, c(x)+\nabla c(x)^\top(y-x)-c(y)\rangle \\ &\le \max_{s\in D}\langle s, c(y)\rangle + \max_{s\in D}\|s\|\|c(x)+\nabla c(x)^\top(y-x)-c(y)\| \\ &\le h(c(y)) + \frac{\max_{s\in D}\|s\|L}{2}\|y-x\|^2. \end{aligned}$$

After adding $-f(x)$ on both sides, this is (13) with $q = \max_{s\in D}\|s\|L$. □

When computing the proximal point of $\varphi_x^{\mathtt{LW}}$ is computationally expensive, an alternative is to employ the composite proximal bundle method of [44]. The proposal therein is to replace the outer function by its cutting-plane model, thereby computing the proximal point of the model

$$\varphi_x^{\mathtt{CS}}(\cdot) := h^{\mathtt{CP}}(c(x)+\nabla c(x)^\top(\cdot - x)) - f(x)\,.$$

By convexity of h, $\varphi_x^{\mathtt{CS}} \le \varphi_x^{\mathtt{LW}}$. This model is also 1QA , by Proposition 5.3.

5.1.3 Sum of functions and prox-gradient method

Given a C^2-function f_1 with Lipschitz-continuous gradient and a convex function f_2, the proximal gradient method [3] minimizes $f := f_1 + f_2$ computing the proximal point of f_2 at $x_k - t_k\nabla f_1(x_k)$, $t_k > 0$. This is equivalent to computing the proximal point of the model that makes a Taylor linearization of f_1 and keeps f_2:

$$\varphi_x^{\mathtt{pg}}(\cdot) := f_1(x) + \langle \nabla f_1(x), \cdot - x\rangle + f_2(\cdot) - f(x)\,.$$

If f_2 is convex, then so is $\varphi_x^{\mathtt{pg}}$. Also, the 1QA property for the model follows directly from the Lipschitz-continuity of the gradient of f_1.

5.1.4 Taylor-like models

The theory in [11] uses powerful tools in Variational Analysis, including Ekeland's variational principle, to prove convergence of a variety of algorithmic schemes. Like in this work, the iterates are generated by computing a proximal point of some model. An important difference, however, is [11, relation (1.4)], which requires the model to approximate f not only uniformly but also *bilaterally* (from above and from below). Specifically, with our notation, the theory presented in [11] requires that

$$\exists q > 0 : \forall y \in \mathbb{R}^n \quad f(y) - f(x) - \frac{q}{2}\|y-x\|^2 \le \varphi_x(y) \le f(y) - f(x) + \frac{q}{2}\|y-x\|^2\,.$$

While this condition holds in several situations described in [11] (related to Taylor-like models), the two-sided requirement excludes cutting-plane models from the analysis. The reason is that, even for a convex f, linearizations in the cutting-plane model $\varphi_x^{\mathtt{CP}}$, the key ingredient in a bundle algorithm, may deviate from below from f in a non-polynomial manner. Figure 1 illustrates this phenomenon.

5.2 Convergence theory for model-based methods

Using 1QA models φ_{x_k} approximating f, we shall consider the following algorithmic scheme, that will be shown to fit the framework of (2).

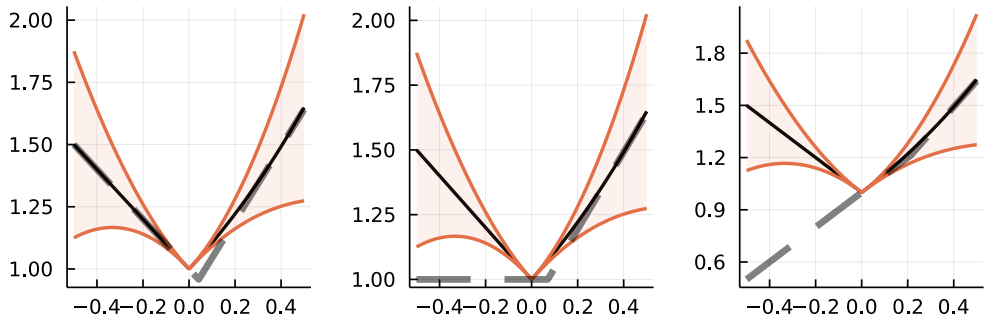

Figure 1: For the function $f(x) = \begin{cases} 1-x & x \le 0 \\ e^x & x > 0 \end{cases}$ plotted with a continuous dark line, three cutting-plane models are shown in dashed lines. These are all 1QA models, because they remain under the thick curved line in the top. By contrast, bilateral models considered in [11] must lie in the shaded region. Even for this simple convex function, none of the cutting-plane models satisfies the two-sided condition in [11].

Starting from some $x_0 \in \mathbb{R}^n$, for all $k \ge 1$,

$$x_k = x_{k-1} - t_{k-1} G_{k-1}, \qquad \text{for } G_{k-1} \in \partial\varphi_{x_{k-1}}(x_k)\,, \tag{14a}$$

$$f(x_k) - f(x_{k-1}) \le m\varphi_{x_{k-1}}(x_k), \qquad \text{for } m \in (0,1)\,. \tag{14b}$$

In particular, the new iterate is obtained computing the proximal point of the model, and the descent is measured using the value of the model at the new point. This is one of the characteristics of bundle methods. Other methods can also be recast in this manner. Below, we show that the sequences associated to the models described in § 5.1.2 and § 5.1.3 are of descent, both in the original sense of (2) and in the model-based sense of (14). Regarding the Taylor-like models in § 5.1.4, the proposal in [11] does not consider a specific type of problem to be tackled by a particular method. So, as long as we are able to generate a descent sequence in the sense of (14), the results in Proposition 5.4 below would hold, since Taylor-like models are bilateral, while 1QA models are one-sided (in this sense, more general).

5.2.1 Decomposable functions and prox-descent method

Let f be a decomposable function as in § 5.1.2, and consider the model $\varphi_x^{\mathtt{LW}}$ defined therein. Let $\{x_k\}$ be a prox-descent sequence as in [30, Algorithm 1].

First, (14a) is a direct consequence of the definition of the next iterate in [30, Algorithm 1] with stepsize $t_k := 1/\mu$. In order to see this, it suffices to recall that the step $d := x_k - x_{k-1}$ is characterized by the relations

$$\nabla c(x_{k-1})^\top v + \frac{1}{t_{k-1}} d = 0, \quad v \in \partial h(c(x_{k-1}) + \nabla c(x_{k-1})^\top d).$$

Setting $G_{k-1} := \nabla c(x_{k-1})^\top v$, it holds that

$$x_k - x_{k-1} = d = -t_{k-1} G_{k-1}, \quad G_{k-1} \in \partial\varphi_{x_{k-1}}(x_k),$$

which is (14a). As for (14b), it is the same as the acceptance criterion for the step in [30, Algorithm 1] with $m = \sigma$.

Note also that it is proven in [30, Theorem 5.4] that [30, Algorithm 1] generates stepsizes t_k that are bounded away from zero. Thus, the algorithm satisfies the assumptions in Proposition 5.4 below.

5.2.2 Sum of functions and prox-gradient method

Let $f = f_1 + f_2$ be as in § 5.1.3. The proximal gradient method conforms to the algorithmic pattern of (2) if $t_{\min} \le t_k \le 1/L_{f_1}$, where L_{f_1} is the Lipschitz constant of ∇f_1. Indeed, (2a) with $\varepsilon_{k-1} = 0$ and $a = L_{f_1}/2$ is a direct consequence of the decent properties of this algorithm; see, e.g., [4, Proposition 6.3.2]. As for (2b), we know that x_k minimizes

$$\varphi_x^{\mathtt{pg}}(\cdot) + \frac{1}{2t_k}\|\cdot - x_{k-1}\|^2.$$

Hence, there is $g_{2,k} \in \partial f_2(x_k)$ such that

$$0 = \nabla f_1(x_{k-1}) + g_{2,k} + \frac{1}{t_k}(x_k - x_{k-1}) = \nabla f_1(x_k) + g_{2,k} + \nabla f_1(x_{k-1}) - \nabla f_1(x_k) + \frac{1}{t_k}(x_k - x_{k-1}).$$

Defining $z_k := x_k$, we have $g_k := \nabla f_1(x_k) + g_{2,k} \in \partial f(x_k)$ and

$$\|g_k\| = \left\|\nabla f_1(x_{k-1}) - \nabla f_1(x_k) + \frac{1}{t_k}(x_k - x_{k-1})\right\| \le (L_{f_1} + 1/t_{\min})\|x_k - x_{k-1}\|.$$

This is (2b) with $b = L_{f_1} + 1/t_{\min}$. Finally, (2c) holds trivially.

5.2.3 Convergence of sequences generated by model-based methods

To continue with our analysis, we need to exhibit the errors ε_k and the theoretical sequence $\{z_k\}$ from (2) that are associated with the bundle-like scheme (14). We start by transporting subgradients of convex models of nonconvex functions to the convex function obtained from f, by weak convexity. This relation and Theorem 5.5 below yield z_k as a perturbation of the iterate x_k, as desired.

Proposition 5.4 (Transportation of subgradients and the validity of (2a))**.** *Consider the minimization of a proper $\rho-$weakly convex function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ applying the model-based proximal scheme in (14) with models φ_x that are 1QA with parameter $q \le \rho$ in Definition 5.1, and let $G_k \in \partial\varphi_{x_k}(x_{k+1})$ as in (14a). The following holds for all k.*

(i) The model aggregate error at x_k,

$$E_k := -t_k\|G_k\|^2 - \varphi_{x_k}(x_{k+1})\,,$$

satisfies $E_k \ge 0$.

(ii) If for all $x \in \mathbb{R}^n$, $F_x(\cdot)$ denotes the (convex) function $f(\cdot) + \frac{\rho}{2}\|\cdot - x\|^2$, then a subgradient G_k in (14a) can be transported to be the convex E_k-subgradient of F_{x_k} at x_k:

$$G_k \in \partial_{E_k} F_{x_k}(x_k).$$

Suppose, in addition, that $\inf f > -\infty$, and the proximal stepsizes are bounded away from zero: $t_k \ge t_{\min} > 0$. Then,

(iii) both $\{G_k\}, \{E_k\}$ converge to 0 as $k \to \infty$, and

(iv) condition (14b) is equivalent to (2a) written with $a = m/t_{\min}$ and $\varepsilon_k = t_{k-1}E_{k-1}$.

Proof. Since the models are 1QA, taking $x = y = x_k$ in (13) gives that $\varphi_{x_k}(x_k) \le 0$. By the convexity of the model and the iterate definition in (14a), it holds that

$$0 \ge \varphi_{x_{k-1}}(x_{k-1}) \ge \varphi_{x_{k-1}}(x_k) + \langle G_{k-1}, x_{k-1} - x_k\rangle = \varphi_{x_{k-1}}(x_k) + t_{k-1}\|G_{k-1}\|^2 = -E_{k-1}\,,$$

and $E_k \ge 0$ for all k, as stated in item (i).

To show item (ii), because the model is 1QA , we have that

$$\varphi_{x_{k-1}}(x) \leq f(x) - f(x_{k-1}) + \frac{\rho}{2}\|x - x_{k-1}\|^2 = F_{x_{k-1}}(x) - f(x_{k-1})\,.$$

Combining now the model convexity with (i) yields

$$\begin{aligned} F_{x_{k-1}}(x) \geq f(x_{k-1}) + \varphi_{x_{k-1}}(x) \quad &\geq \quad f(x_{k-1}) + \varphi_{x_{k-1}}(x_k) + \langle G_{k-1}, x - x_k\rangle \\ &= \quad f(x_{k-1}) + \langle G_{k-1}, x - x_{k-1}\rangle \\ &\quad\; +\varphi_{x_{k-1}}(x_k) + \langle G_{k-1}, x_{k-1} - x_k\rangle \\ &= \quad F_{x_{k-1}}(x_{k-1}) + \langle G_{k-1}, x - x_{k-1}\rangle - E_{k-1}\,. \end{aligned}$$

As the last relation is (ii) written with k replaced by $k-1$, the desired result follows.

To show item (iii), note that the descent condition (14b), written using the aggregate gradient and error definitions, gives

$$m(E_{k-1} + t_{k-1}\|G_{k-1}\|^2) \leq f(x_{k-1}) - f(x_k)\,. \tag{15}$$

As $\{f(x_k)\}$ is non-increasing and f is bounded below, this sequence is convergent. Hence, $f(x_{k-1}) - f(x_k) \to 0$ as $k \to \infty$. Then from (15) and $t_k \geq t_{\min} > 0$, it follows that $E_k \to 0$ and $G_k \to 0$.

Finally, rewriting the descent condition (14b) using the aggregate gradient and error definitions yields (iv), as

$$f(x_k) + \frac{m}{t_{k-1}}(\|x_k - x_{k-1}\|^2 + t_{k-1}E_{k-1}) \leq f(x_{k-1})\,.$$

□

To complete formulating (14) in the format of the algorithmic pattern in (2), we show the validity of (2b) and (2c). This is achieved applying the error bound inequality in Definition 3.1, noting that it involves the exact (Clarke) subgradients of f. We have just shown that the transported model subgradient is an E_k-subgradient of the auxiliary convex function F_{x_k} at x_k. The connection with the original function f is done by means of the following result, reproduced from [42].

Theorem 5.5 (Brøndsted-Rockafellar's like relation, Theorem 2 in [42])**.** *Let F be a proper lower semicontinuous convex function on $\mathbb{R}^n$. Suppose that $E \geq 0$ and that $G \in \partial_E F(x)$. Then, for each $\gamma > 0$, there is a unique $y = y(\gamma)$ such that*

$$G - \frac{1}{\gamma}y \in \partial F(x + \gamma y)\,, \quad \|y\| \leq \sqrt{E}\,.$$

By the above result, any $\varepsilon-$subgradient of a convex function can be perturbed to obtain an exact subgradient of the same function, at a perturbed point. Since weak convexity gives an explicit relation between f and the convex function F_x, we shall be able to relate the respective subgradients, and apply the subdifferential error bound for f using the perturbed points.

Lemma 5.6 (Casting (14) in the format of (2))**.** *Under the assumptions of Proposition 5.4, suppose f satisfies the subdifferential error bound of Definition 3.1 and the sequence of stepsizes $\{t_k\}$ in (14a) is bounded below by $t_{\min} > 0$. Then there exists a theoretical sequence $\{z_k\}$ such that all conditions in (2) hold, with $\|z_k - x_k\| \leq \sqrt{\ell E_k}$.*

Proof. The validity of (2a) was already shown in Proposition 5.4(iv).

To derive the expression for z_k, apply Theorem 5.5 written with $G := G_k \in \partial_{E_k} F_{x_k}(x_k)$ for the convex function $F := F_{x_k}$, $E := E_k$, taking $\gamma := \sqrt{\ell} > 0$, where $\ell > 0$ is the constant of the subdifferential error bound in Definition 3.1. It follows that there exists a unique y_k such that

$$\|y_k\| \leq \sqrt{E_k} \quad \text{and} \quad G_k - \frac{1}{\sqrt{\ell}}y_k \in \partial F_{x_k}\Big(x_k + \sqrt{\ell}y_k\Big) = \partial f\Big(x_k + \sqrt{\ell}y_k\Big) + \rho\sqrt{\ell}y_k\,,$$

by the definition of F_{x_k}. Therefore,

$$\text{letting } z_{k-1} := x_{k-1} + \sqrt{\ell} y_{k-1} \text{ it holds that } g_{k-1} := G_{k-1} - \left(\frac{1+\rho\ell}{\sqrt{\ell}}\right) y_{k-1} \in \partial f(z_{k-1})\,.$$

To show that condition (2c) holds, first notice that

$$\frac{1}{\sqrt{\ell}}\|z_{k-1} - x_{k-1}\| = \|y_{k-1}\| \leq \sqrt{E_{k-1}}\,.$$

Since $E_k \to 0$ by Proposition 5.4(iii), this means that $z_{k-1} - x_{k-1} \to 0$. The remaining condition $\varepsilon_k \to 0$ follows from the expression $\varepsilon_k = t_{k-1}E_{k-1}$ in Proposition 5.4(iv), combined with the boundedness assumption on t_k, using once more that $E_k \to 0$.

To show that the sequence $\{g_k \in \partial f(z_{k-1})\}$ satisfies condition (2b), notice that

$$\begin{aligned}
\|g_{k-1}\| &\leq \|G_{k-1}\| + \left(\tfrac{1+\rho\ell}{\sqrt{\ell}}\right)\|y_{k-1}\| \\
&= \tfrac{1}{t_k}\|x_k - x_{k-1}\| + \left(\tfrac{1+\rho\ell}{\sqrt{\ell}}\right)\tfrac{1}{\sqrt{\ell}}\|z_{k-1} - x_{k-1}\| \\
&\leq \tfrac{1}{t_{\min}}\|x_k - x_{k-1}\| + \left(\tfrac{1+\rho\ell}{\ell}\right)\|z_{k-1} - x_{k-1}\|\,.
\end{aligned}$$

Hence, (2b) holds with $b := \max\{1/t_{\min}, (1+\rho\ell)/\ell\}$. □

Thanks to Lemma 5.6, we are now in position of applying Theorem 4.3 to show that the general scheme based on models considered in this section converges, with a rate that is R-linear for the iterates and Q-linear for the functional values.

Theorem 5.7 (Global convergence of (14) and local linear rate)**.** *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a weakly convex function, such that $\inf f > -\infty$. Suppose, in addition, that f satisfies the proper separation of isocost surfaces (Definition 1.1) and the subdifferential error bound (Definition 3.1), and that the sequence of stepsizes $\{t_k\}$ in (14a) is bounded below by $t_{\min} > 0$. The following holds for the model-based proximal scheme in (14), as long as the models φ_x therein are 1QA (Definition 5.1 with parameter $q \leq \rho$).*

(i) $\{f(x_k)\}$ monotonically converges to some critical value f^*, such that the sequence of functional errors $\{v_k = f(x_k) - f^*\}$ converges to 0 with Q-linear rate:

$$\exists q \in (0,1) : v_k \leq q v_{k-1} \text{ for all sufficiently large } k.$$

(ii) The sequence of iterates $\{x_k\}$ converges to a critical point x_* of f with R-linear rate:

$$\exists q \in (0,1) \text{ and } c > 0 : \|x_k - x^*\| \leq c\sqrt{q}^k \text{ for all sufficiently large } k.$$

Proof. To see item (i), we apply Theorem 4.3. First, from the definition of the aggregate error E_k and (15), it follows that

$$E_{k-1} \leq \frac{1}{m}(v_{k-1} - v_k)\,,$$
$$\|G_k\|^2 \leq \frac{1}{m t_{k-1}}(v_{k-1} - v_k)\,.$$

The first inequality combined with the definition of z_k imply that

$$\|x_{k-1} - z_{k-1}\|^2 \leq \ell E_{k-1} \leq \frac{\ell}{m}(v_{k-1} - v_k)\,.$$

Moreover, combining the last inequalities with $G_{k-1} \in \partial_{E_{k-1}} F_{x_{k-1}}(x_{k-1})$, the definition of z_k, and the fact that t_k is bounded away from 0, we obtain that

$$
\begin{aligned}
f(x_{k-1}) - f(z_{k-1}) &\leq \frac{\rho}{2}\|z_{k-1} - x_{k-1}\|^2 - \langle G_{k-1}, z_{k-1} - x_{k-1}\rangle + E_{k-1} \\
&\leq \frac{\rho\ell}{2m}(v_{k-1} - v_k) + \|G_{k-1}\|\sqrt{\frac{\ell}{m}(v_{k-1} - v_k)} + \frac{1}{m}(v_{k-1} - v_k) \\
&\leq \frac{1}{m}\left(\frac{\rho\ell}{2} + \sqrt{\frac{\ell}{t_{\min}}} + 1\right)(v_{k-1} - v_k)\,.
\end{aligned}
$$

Since (11) and (12) in Theorem 4.3 hold for

$$
C_1 = \frac{1}{m}\left(\frac{\rho\ell}{2} + \sqrt{\frac{\ell}{t_{\min}}} + 1\right) \quad \text{and} \quad C_2 = \frac{\ell}{m}\,,
$$

items (i) and (ii) follow. □

6 The theory applied to constrained smooth optimization

Another application of our unified analysis is the feasible descent framework of [32] (see also [46]). Consider the constrained optimization problem (4), where $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable with Lipschitz-continuous gradient on the nonempty closed convex set $X \subseteq \mathbb{R}^n$.

The work [32] considers iterative sequences $\{x_k\}$ satisfying

$$
x_k = P_X(x_{k-1} - t_{k-1}\nabla f(x_{k-1}) + e_{k-1})\,, \quad t_{k-1} \geq t_{\min} > 0\,, \tag{16a}
$$

$$
\|e_{k-1}\| \leq \alpha\|x_k - x_{k-1}\|\,, \quad \alpha \in (0,1)\,. \tag{16b}
$$

This setting is quite broad. It includes, of course, the basic gradient projection method, taking $e_k = 0$ for all k. But, depending on the form of the mapping e that gives e_{k-1} in (2), it includes many other algorithms for solving problem (4). Some examples are the extragradient method, the proximal point method, coordinate descent, and several splitting techniques; see [32] and references therein.

We next show that our general analysis of (2) is applicable to methods given by (16) as well. We consider (2) for the function $(f + i_X)$ and take, for all $k \geq 1$, $\varepsilon_k = 0$ and $x_k = z_k$ (note that (2c) is then automatic). Under the stated assumptions, $f + i_X$ is weakly convex; see Proposition 2.4. We next need to show that (16) implies (2a) and (2b) for $f + i_X$. Once this is done, we apply Theorem 4.3 for the weakly convex function $f + i_X$.

The proof below that the sequence $\{x_k\}$ from (16) satisfies the descent condition (2a) for $f + i_X$ is essentially a similar argument as in [32] for f, because by (16a) it holds that $x_k \in X$ for all k (and so $(f + i_X)(x_k) = f(x_k)$). We include this part of the proof here mostly for completeness. Note, however, that the subgradients of f and of $(f + i_X)$ are not the same. Also, our rate of convergence analysis is different, as our results are based on the subdifferential error bound (Definition 3.1), while [32] uses the projection error bound (Definition 3.2).

Proposition 6.1 (The feasible descent framework (16) fits (2))**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function with $L-$Lipschitz continuous gradient on the nonempty closed convex set $X \subseteq \mathbb{R}^n$. Then any sequence $\{x_k\}$ satisfying (16) is a sequence of descent for the function $f + i_X$ in the sense of (2). More specifically,*

(i) For all k,

$$
f(x_k) + \left(\frac{1-\alpha}{t^*} - \frac{L}{2}\right)\|x_k - x_{k-1}\|^2 \leq f(x_{k-1})\,,
$$

whenever $t_k \leq t^* \leq 2(1-\alpha)/L$. I.e., (2a) holds for $f + i_X$ (recall that $x_k \in X$).

(ii) For all k, there exists $u_k \in N_X(x_k)$ such that

$$\|\nabla f(x_k) + u_k\| \leq \left(\frac{1+\alpha}{t_{\min}} + L\right) \|x_k - x_{k-1}\|\,,$$

i.e., (2b) holds for $f + i_X$.

Proof. From (16a) and the characterization of the projection operator, for all $y \in X$ it holds that

$$\langle x_{k-1} - t_{k-1}\nabla f(x_{k-1}) + e_{k-1} - x_k, y - x_k\rangle \leq 0\,.$$

Taking $y = x_{k-1}$ in this inequality and rearranging terms, we obtain that

$$\|x_{k-1} - x_k\|^2 - t_{k-1}\langle \nabla f(x_{k-1}), x_{k-1} - x_k\rangle \leq \langle e_{k-1}, x_k - x_{k-1}\rangle\,.$$

Using the Cauchy-Schwarz inequality and (16b) on the right-hand side, it holds that

$$\|x_{k-1} - x_k\|^2 - t_{k-1}\langle \nabla f(x_{k-1}), x_{k-1} - x_k\rangle \leq \alpha\|x_{k-1} - x_k\|^2\,.$$

It follows that

$$\langle \nabla f(x_{k-1}), x_k - x_{k-1}\rangle \leq \frac{\alpha - 1}{t_{k-1}}\|x_{k-1} - x_k\|^2\,.$$

Since the function is differentiable with Lipschitz-continuous gradient with constant L, by [25, Lemma A.11] we have that

$$f(x_k) - f(x_{k-1}) \leq \langle \nabla f(x_{k-1}), x_k - x_{k-1}\rangle + \frac{L}{2}\|x_k - x_{k-1}\|^2\,.$$

Combining the last two inequalities above gives

$$f(x_k) - f(x_{k-1}) \leq \left(\frac{\alpha - 1}{t_{k-1}} + \frac{L}{2}\right)\|x_{k-1} - x_k\|^2\,,$$

from which item (i) follows.

We next prove item (ii), i.e., condition (2b) for $f + i_X$. Again, from (16a) and the characterization of the projection operator, there exists $\nu_k \in N_X(x_k)$ such that

$$x_{k-1} - t_{k-1}\nabla f(x_{k-1}) + e_{k-1} - x_k = \nu_k\,.$$

Defining $u_k = \nu_k / t_{k-1} \in N_X(x_k)$, we have that

$$t_{k-1}u_k = x_{k-1} - x_k + e_{k-1} - t_{k-1}\nabla f(x_{k-1})\,,$$

and

$$t_{k-1}(\nabla f(x_k) + u_k) = x_{k-1} - x_k + e_{k-1} + t_{k-1}(\nabla f(x_k) - \nabla f(x_{k-1}))\,.$$

Define $w_k = \nabla f(x_k) + u_k \in \partial(f + i_X)(x_k)$. We then obtain that

$$\begin{aligned}\|w_k\| &\leq \frac{1}{t_{k-1}}\|x_{k-1} - x_k + e_{k-1}\| + \|\nabla f(x_k) - \nabla f(x_{k-1})\| \\ &\leq \left(\frac{1+\alpha}{t_{\min}} + L\right)\|x_k - x_{k-1}\|\,,\end{aligned}$$

where the triangle inequality, (16b), and the Lipschitz-continuity of the gradient of f were used. The proof is complete. □

Due to Propositions 6.1 and 2.4, we are now in position to apply our unified analysis for weakly convex functions to obtain estimates for the rate of convergence in (16).

Theorem 6.2 (Linear rate of convergence of (16))**.** *Under the assumptions of Proposition 6.1, if f is bounded from below, and the subdifferential error bound (Definition 3.1) and the proper separation of isocost surfaces condition (Definition 1.1) hold, then for the iterates $\{x_k\}$ satisfying (16) it holds that:*

(i) There exists some critical value $f^* \in \mathbb{R}$ of f such that $f(x_k) \to f^*$. For $v_k := f(x_k) - f^*$, there exists $q \in (0,1)$ such that for all sufficiently large k,

$$v_k \leq q v_{k-1}\,.$$

(ii) $\{x_k\}$ converges R-linearly to a critical point x^* of f with $f(x^*) = f^*$. More specifically, there exists $c > 0$ such that for all k sufficiently large,

$$\|x_k - x^*\| \leq c\sqrt{q}^k$$

Proof. By Proposition 2.4, $f + i_X$ is a weakly convex function. By Proposition 6.1, any sequence $\{x_k\}$ satisfying (16) conforms to (2) and all the conditions of Theorem 4.3, with $x_k = z_k$, $\varepsilon_k = 0$ for all k, and $g_k \in \partial(f + i_X)(x_k)$. Then the assertions follow from Theorem 4.3, with

$$q = \frac{M}{1+M}\,, \quad M = \frac{2\ell\left(\frac{1+\alpha}{t_{\min}+L}\right)^2}{\frac{1-\alpha}{t_{\min}} - \frac{L}{2}(1+L\ell)}\,, \quad c = \frac{\sqrt{v_0}}{\sqrt{\frac{1-\alpha}{t^*} - \frac{L}{2}}\left(1-\sqrt{q}\right)}\,.$$

□

Note that while the scheme (16) is explicit in our terminology, as it uses the gradient of f at x_{k-1}, it is cast in our framework (2) as being implicit, as the subgradient of $f + i_X$ is taken therein at x_k.

Acknowledgments The first author is supported by FAPESP grant 2019/20023-1. Research of the second author is partly supported by CNPq-Brazil Grant 306089/2019-0, FAPESP CEPID CeMEAI Grant 2013/07375-0, and PRONEX–Optimization. The third author is supported by FAPESP Grant 2018/24293-0, CNPq Grant 304301/2019-1, FAPESP CEPID CeMEAI Grant 2013/07375-0, and PRONEX–Optimization. The fourth author is supported in part by CNPq Grant 303913/2019-3, by FAPERJ Grant E-26/202.540/2019, and by PRONEX–Optimization.

References

[1] F. A. Artacho and M. H. Geoffroy. "Characterization of metric regularity of subdifferentials". In: *J. Convex Anal.* 15.2 (2008), p. 365.

[2] H. Attouch, J. Bolte, and B. F. Svaiter. "Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized Gauss–Seidel methods". In: *Math. Program.* 137.1-2 (2013), pp. 91–129.

[3] A. Beck. *First–Order Methods in Optimization*. SIAM, 2017.

[4] D. P. Bertsekas. *Convex Optimization Algorithms*. Athena scientific, 2015.

[5] J. Bolte, A. Daniilidis, A. Lewis, and M. Shiota. "Clarke subgradients of stratifiable functions". In: *SIAM J. Optim.* 18.2 (2007), pp. 556–572.

[6] J. Bonnans, J. Gilbert, C. Lemaréchal, and C. Sagastizábal. *Numerical Optimization: Theoretical and Practical Aspects*. Second Edition. Berlin, Germany: Springer, 2006.

[7] F. H. Clarke. *Optimization and nonsmooth analysis*. SIAM, 1990.

[8] A. Daniilidis and J. Malick. "Filling the gap between lower-C1 and lower-C2 functions". In: *J. Convex Anal.* 12.2 (2005), pp. 315–329.

[9] D. Davis and B. Grimmer. "Proximally guided stochastic subgradient method for nonsmooth, nonconvex problems". In: *SIAM J. Optim.* 29.3 (2019), pp. 1908–1930.

[10] M. Díaz and B. Grimmer. *Optimal Convergence Rates for the Proximal Bundle Method.* arXiv 2105.07874. 2021.

[11] D. Drusvyatskiy, A. Ioffe, and A. Lewis. "Nonsmooth optimization using Taylor-like models: error bounds, convergence, and termination criteria". In: *Math. Program.* 185.1-2 (2019), pp. 357–383.

[12] D. Drusvyatskiy, B. S. Mordukhovich, and T. T. A. Nghia. "Second-order growth, tilt stability, and metric regularity of the subdifferential". In: *J. Convex Anal. 21(4), 1165–1192* (2014).

[13] D. Drusvyatskiy and D. Davis. "Subgradient methods under weak convexity and tame geometry". In: *SIAG/OPT Views and News* 28.1 (2020), pp. 1–10.

[14] D. Drusvyatskiy and A. Lewis. "Error Bounds, Quadratic Growth, and Linear Convergence of Proximal Methods". In: *Math. Oper. Res.* 43.3 (2018), pp. 919–948.

[15] D. Drusvyatskiy and C. Paquette. "Efficiency of minimizing compositions of convex functions and smooth maps". In: *Math. Program.* 178.1 (2019), pp. 503–558.

[16] Y. Du and A. Ruszczyński. "Rate of convergence of the bundle method". In: *J. Optim. Theory Appl.* 173.3 (2017), pp. 908–922.

[17] F. Facchinei and J.-S. Pang. *Finite-dimensional variational inequalities and complementarity problems.* Springer Science & Business Media, 2007.

[18] A. Fischer. "Local behavior of an iterative framework for generalized equations with nonisolated solutions". In: *Math. Program.* 94.1 (2002), pp. 91–124.

[19] A. Fuduli, M. Gaudioso, and G. Giallombardo. "Minimizing nonconvex nonsmooth functions via cutting planes and proximity control". In: *SIAM J. Optim.* 14.3 (2003), pp. 743–756.

[20] W. Hare and C. Sagastizábal. "A redistributed proximal bundle method for nonconvex optimization". In: *SIAM J. Optim.* 20.5 (2010), pp. 2442–2473.

[21] W. Hare and C. Sagastizábal. "Computing proximal points of nonconvex functions". In: *Math. Program.* 116.1-2 (2009), pp. 221–258.

[22] W. Hare, C. Sagastizábal, and M. Solodov. "A proximal bundle method for nonsmooth nonconvex functions with inexact information". In: *Comput. Optim. Appl.* 63 (2016), pp. 1–28.

[23] J. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms I and II.* Grundlehren der mathematischen Wissenschaften 305 and 306. Springer-Verlag, 1996.

[24] A. Ioffe. "Metric regularity—a survey part 1. theory". In: *J. Aust. Math. Soc.* 101.2 (2016), pp. 188–243.

[25] A. Izmailov and M. Solodov. *Newton-type methods for optimization and variational problems.* Springer Series in Operations Research and Financial Engineering. Springer, Cham, 2014.

[26] K. Kiwiel. "A phase I–phase II method for inequality constrained minimax problems". In: *Control. Cybern.* 12.1-2 (1983), pp. 55–75.

[27] K. Kiwiel. "Efficiency of Proximal Bundle Methods". In: *J. Optim. Theory Appl.* 104.3 (2000), pp. 589–603.

[28] K. Kiwiel. *Methods of descent for nondifferentiable optimization.* Berlin: Springer-Verlag, 1985, pp. vi+362.

[29] K. Kurdyka. "On gradients of functions definable in o-minimal structures". In: *Annales de l'Institut Fourier.* Vol. 48. 3. 1998, pp. 769–783.

[30] A. Lewis and S. Wright. "A proximal method for composite minimization". In: *Math. Program.* 158.1-2 (2015), pp. 501–546.

[31] L. Lukšan and J. Vlček. "Globally convergent variable metric method for nonconvex nondifferentiable unconstrained minimization". In: *J. Optim. Theory Appl.* 2 (2001), pp. 407–430.

[32] Z.-Q. Luo and P. Tseng. "Error bounds and convergence analysis of feasible descent methods: a general approach". In: *Ann. Oper. Res.* 46.1 (1993), pp. 157–178.

[33] M. Makela and P. Neittaanmaki. *Nonsmooth Optimization: Analysis and Algorithms with Applications to Optimal Control*. World Scientific: Singapore, 1992.

[34] R. Mifflin. "A modification and extension of Lemarechal's algorithm for nonsmooth minimization". In: *Math. Programming Stud.* 17 (1982), pp. 77–90.

[35] R. Mifflin. "Semismooth and Semiconvex Functions in Constrained Optimization". In: *SIAM J. Control Optim.* 15.6 (1977), pp. 959–972.

[36] H. V. Ngai, D. T. Luc, and M. Théra. "Approximate convex functions". In: *J. Nonlinear Convex Anal.* 1.2 (2000), pp. 155–176.

[37] D. Noll, O. Prot, and A. Rondepierre. "A proximity control algorithm to minimize nonsmooth and nonconvex functions". In: *Pac. J. Optim.* 4.3 (2008), pp. 569–602.

[38] D. Noll. "Convergence of Non-smooth Descent Methods Using the Kurdyka–Łojasiewicz Inequality". In: *J. Optim. Theory Appl.* 160.2 (2013), pp. 553–572.

[39] V. Norkin. "Generalized-differentiable functions". In: *Cybernetics* 16.1 (1980), pp. 10–12.

[40] J. Ortega and W. Rheinboldt. *Iterative Solution of Nonlinear Equations in Several Variables*. Academic Press, New York, 1970.

[41] J.-S. Pang. "Error bounds in mathematical programming". In: *Math. Program.* 79 (1997), pp. 299–332.

[42] S. M. Robinson. "Linear convergence of epsilon-subgradient descent methods for a class of convex functions". In: *Math. Program.* 86.1 (1999), pp. 41–50.

[43] R. T. Rockafellar and R. J.-B. Wets. *Variational analysis*. Vol. 317. Springer Science & Business Media, 2009.

[44] C. Sagastizábal. "Composite Proximal Bundle Method". In: *Math. Program.* 140.1 (2013), pp. 189–233.

[45] A. Shapiro. "On a class of nonsmooth composite functions". In: *Math. Oper. Res.* 28.4 (2003), pp. 677–692.

[46] M. Solodov. "Convergence Analysis of Perturbed Feasible Descent Methods". In: *J. Optim. Theory Appl.* 93 (1997), pp. 337–353.

[47] M. Solodov. "Convergence rate analysis of iterative algorithms for solving variational inequality problems". In: *Math. Program.* 96 (2003), pp. 513–528.

[48] M. Solodov and P. Tseng. "Modified Projection-Type Methods for Monotone Variational Inequalities". In: *SIAM J. Control Optim.* 34 (1996), pp. 1814–1830.

[49] S. Wright. *Some Perspectives on Nonconvex Optimization*. IPAM workshop on Intersections between Control, Learning and Optimization. 2020.

[50] R. Zhang and J. Treiman. "Upper-Lipschitz multifunctions and inverse subdifferentials". In: *Nonlinear Anal., Theory Methods Appl.* 24.2 (1995), pp. 273–286.

[51] Z. Zhou and A. M.-C. So. "A unified approach to error bounds for structured convex optimization problems". In: *Math. Program.* 165.2 (2017), pp. 689–728.