INEXACTNESS IN BUNDLE METHODS FOR LOCALLY LIPSCHITZ FUNCTIONS

W. HARE, C. SAGASTIZÁBAL, AND M. SOLODOV

ABSTRACT. We consider the problem of computing a critical point of a nonconvex locally Lipschitz function over a convex compact constraint set given an *inexact oracle* that provides an approximate function value and an approximate subgradient. We assume that the errors in function and subgradient evaluations are merely bounded, and in particular need not vanish in the limit. After some discussion on how to appropriately define an approximate subgradient in a nonconvex setting, the paper builds on bundle methods for convex functions defined with inexact information and for nonsmooth nonconvex functions defined through exact oracles. The algorithm herein incorporates a "noise attenuation" technique, activated when the inexactness of the oracle is excessive and causing difficulties in making progress. Convergence to approximately critical points is proven under the assumption that the objective function is regular, locally Lipschitz, and the bundle is appropriately managed.

Keywords: Nonconvex optimization, nonsmooth optimization, bundle methods, inexact oracles, locally Lipschitz functions, proximal point.
AMS subject classification: Primary, 90C26, 49J52; Secondary, 65K10, 49J53, 49M05

1. INTRODUCTION

In this paper we seek to approximately solve the problem

$$\min\{f(x) : x \in C\}, \tag{1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a nonconvex locally Lipschitz function given by an inexact oracle, and $C \subset \mathbb{R}^n$ is a convex compact set. As a practical matter, C must be simple enough, for example, defined by box or linear constraints, so that the resulting bundle method subproblems are quadratic programs. That said, modern computational tools also allow to solve efficiently somewhat more complex subproblems, such as consisting in minimizing quadratic functions subject to convex quadratic constraints, e.g., [AFGG11]. So, in practice, C could be defined by convex quadratics too. As a matter of the theory presented in the sequel, C can be any convex compact set.

Date: Manuscript draft: June 29, 2012.

W. Hare: University of British Columbia, Okanagan Campus, 3333 University Way, Kelowna, BC, V1Y 8C5, Canada. Research supported in part by NSERC DG program and UBC IRF. warren.hare@ubc.ca.

C. Sagastizábal: On leave from INRIA Rocquencourt, France.

Visiting researcher at IMPA, Estrada Dona Castorina 110, Jardim Botânico, Rio de Janeiro RJ 22460-320, Brazil. Research supported in part by CNPq 303840/2011-0, AFOSR FA9550-08-1-0370, NSF DMS 0707205, PRONEX-Optimization, and FAPERJ. sagastiz@impa.br.

M. Solodov: IMPA. Research supported in part by CNPq Grant 302637/2011-7, PRONEX-Optimization, and by FAPERJ. solodov@impa.br.

An *inexact oracle* is a procedure that, given a point x^i, returns estimations for both the function value and a subgradient. Accordingly, the available objection function information is $f^i \approx f(x^i)$ and $g^i \approx g(x^i) \in \partial f(x^i)$. Working with inexact oracles presents a natural challenge in a number of modern applications. For example, when solving large-scale problems by Lagrangian relaxation the objective function in (1) has the form $f(x) := \sup\{F_z(x) : z \in Z\}$ for a compact convex set Z. In many such applications, it may be impossible to evaluate f precisely, but controllable accuracy is easily obtained [ES10, Sag12]. Inexact oracles also arise in Stochastic Programming. For instance, when the expected value objective function is computed via Monte-Carlo simulation, by taking the mean over a large sample of random realizations, [RC04]. (Sub-)Gradient values for such functions could also be approximated numerically; for example, using Simplex Gradients, Centered Simplex Gradients, or Gupal estimators [Gup77, BKS08, CSV09, Kiw10, HM12]. Naturally, the accuracy of function and (sub-)gradient values can be increased by taking larger samples, but this involves increasing the computational effort. Another example refers to two-stage stochastic linear programming problems. When applying an L-shaped or Benders decomposition, the effort for evaluating the function f and a subgradient amounts to solving as many linear programs as scenarios compose the sample. To speed up calculations, one can compute the oracle information only approximately, skipping the exact linear programming solution for almost all scenarios and replacing the missing information by some sound value as in [OSS11]; see also [ZFEM12, OS12]. Alternatively, the linear programming solver can be stopped before optimality as in [ZPR00].

The minimization of nonsmooth *convex* functions that are given by *exact* oracles has been successfully approached in several manners. Amongst the most popular are the bundle and proximal-bundle methods [HUL93, Ch. XV]. Indeed, such methods are currently considered among the most efficient optimization methods for nonsmooth problems; see, e.g., [Lem01, SS05, HS10] for more detailed comments.

From the "primal" point of view, bundle methods can be thought of as replacing the true objective function by a piecewise-linear model function, constructed through a bundle of information gathering past evaluation points and their respective oracle output. In particular, proximal-bundle methods, [HUL93, Ch. XV], compute the *proximal point* of the model function to obtain new bundle elements and generate better minimizer estimates. This work extends such methods to handle both nonconvex objective functions and inexact oracles.

Not long after works on bundle methods were first developed, the problem of (locally) minimizing of a nonsmooth *nonconvex* function given by an *exact* oracle was considered by [Mif82b, Kiw85] and more recently in [MN92, Kiw96, LV98, HS10]. Most of these bundle methods were developed from a "dual" point of view. That is, they focus on driving certain convex combinations of subgradients towards satisfaction of first order optimality conditions [Lem75, Mif77, Lem78, LSB81, Mif82a, Mif82b]. In addition, except for [HS10], all of these methods handle nonconvexity by down-shitfing the so-called linearization errors if they are negative. Our algorithm, however, uses similar techniques to [HS10] to asymptotically ensure satisfaction of first order optimality without resorting to down-shifting; using instead the *Goldstein approximate subdifferential* (see Definition 1.2).

Inexact evaluations in *subgradient methods* had been studied in the nonconvex setting in [SZ98], and thoroughly in the convex case in [NB10]. Contrary to earlier work on inexact subgradient methods, both [SZ98] and [NB10] allow *non-vanishing* noise, i.e., evaluations of subgradients need not be asymptotically tightened. Inexact evaluations of function and subgradient values in convex bundle methods date back to [Kiw95]. However, the noise in [Kiw95] is asymptotically vanishing. The first work where non-vanishing perturbations in bundle methods had been considered appears to be [Hin01]; however, only subgradient values could be computed approximately, while function evaluations had to be exact. Non-vanishing inexactness (still in the convex case) in both functions and subgradient values was introduced in [Sol03], and thoroughly studied in [Kiw06]. Our goal in this work is to show that proximal bundle methods can also be adapted to work for *nonconvex* functions that are given by inexact oracles. However, it should be noted that inexactness of the subgradients in [Kiw06] was understood in terms of the ε-subdifferential for convex functions. This work employs a different error suited to the nonconvex setting, as discussed in Subsection 1.2.

In a manner similar to [Kiw06], our algorithm incorporates "noise attenuation" steps to deal with situations when the inexactness of the oracle is excessive and causes difficulty in making progress. Convergence to approximate critical points is proven under the assumption that the objective function is regular, locally Lipschitz, and the bundle is appropriately managed. To the best of our knowledge, the only other paper dealing with inexact oracles in bundle methods for nonconvex functions is [Nol12]. The method there is rather different, as [Nol12] does not employ the noise attenuation ideas and uses the "down-shift" mechanism to deal with nonconvexity. Convergence for the algorithm in our paper is proven for proper, regular, locally Lipschitz functions with full domain that are minimized over a compact convex set. Similar conditions are required in [Nol12]. In particular, [Nol12] examines the cases where the objective is either ε-convex (see [Nol12, eq. (1.14)]) or lower-$\mathscr{C}^1$ (see [RW98, Def 10.29]). When proper and lower semi-continuous, the class of ε-convex functions is locally Lipschitz [NLT00, Prop 3.2], while lower-$\mathscr{C}^1$ functions are proper, regular, and locally Lipschitz [RW98, Thm 10.31]. Unlike our work, which assumes a bounded constraint set, the work of [Nol12] assumes bounded lower level sets. However, as the constraint herein is only used to bound certain continuous variables, it is clear that our results could also be rephrased in this manner. Overall, our convergence results are quite similar to [Nol12]. However, the algorithms themselves are very different.

The remainder of this paper is organized as follows. This section continues by outlining some global conditions that we assume to hold throughout this work and then briefly explaining of what is meant by an inexact subgradient in a nonconvex setting. Section 2 summarizes the notation used in our algorithm. In Section 3 we discuss how linearization errors and criticality are to be understood in our setting. We further provide some discussion that explains how to interpret a null criticality measure for inexact oracles (equation (17)). Section 4 analyzes our noise attenuation step, including some useful consequences of the computation. In Section 5 we provide the algorithmic framework, called the Inexact Proximal Bundle Method, developed in this paper. Convergence results are given in Section 6, and the work ends with some concluding remarks.

1.1. **General notation and assumptions.**

Throughout this work we assume that, in problem (1),

$$\boxed{\begin{array}{c}\text{the objective function } f \text{ is } \textit{proper} \text{ [RW98, p. 5],}\\ \textit{regular} \text{ [RW98, Def 7.25], and locally Lipschitz with full domain.}\end{array}} \tag{2}$$

Note that, in the supremum function example in the introduction, that is when $f(x) := \sup\{F_z(x) : z \in Z\}$ and Z is a compact convex infinite set, if f_z is well-behaved in Z, then the function is a "lower-$\mathscr{C}^2$" function, so proper, regular, and locally Lipschitz [RW98, Def 10.29 & Thm 10.31]. Also note that, the assumption that f is proper with full domain means that f is finite-valued for all $x \in \mathbb{R}^n$. This is more than reasonable in a paper devoted to an algorithmic development.

In general we shall work with the definitions and notation laid out in [RW98]. The closed ball in $\mathbb{R}^n$ of radius $\varepsilon > 0$ is denoted by $B_\varepsilon(0)$. We shall use $\partial f(\bar{x})$ to denote the subdifferential of f at the point $\bar{x}$. Note that regularity implies that the subdifferential mapping is well-defined and can be computed by

$$\partial f(x) := \left\{ g \in \mathbb{R}^n : \liminf_{x \to \bar{x} x \neq \bar{x}} \frac{f(x) - f(\bar{x}) - \langle g, x - \bar{x}\rangle}{|x - \bar{x}|} \geq 0 \right\}. \tag{3}$$

Alternate definitions of the subdifferential mapping for regular functions can be found in [RW98, Chap. 8].

1.2. **Inexact Oracles and Nonconvexity.**

The idea of an inexact oracle for function values is easily defined. Given a point x^i, the statement "the value f^i approximates $f(x^i)$" mathematically translates to $|f^i - f(x^i)| \leq \sigma^i$ for some small error term σ^i. For (sub-)gradient values, the idea of an inexact oracle is less clear.

In [Kiw06], it is assumed that given a point x^i the inexact oracle for a *convex* function f returns a vector g^i in the Convex Analysis ε-subdifferential, $g^i \in \partial_{\varepsilon^i} f(x^i)$, where $\partial_\varepsilon f(\bar{x})$ is given by

$$\partial_\varepsilon f(\bar{x}) := \{ g : f(x) \geq f(\bar{x}) - \varepsilon + \langle g, x - \bar{x}\rangle, \text{ for all } x \in \mathbb{R}^n \}. \tag{4}$$

In this way, the inexact linearizations of f built with inexact oracle, that is hyperplanes of the form $f^i + \langle g, x - x^i\rangle$, stay below $f(x) + \varepsilon + \sigma^i$. For the particular case of an affine function $f(x) = Ax + b$, since $\partial_\varepsilon f(\bar{x}) = \partial f(\bar{x}) = A$, the "inexact' subgradients are in fact exact.

In a nonconvex setting, the definition of inexact subgradient is naturally more problematic. Considering the nonconvex function $f(x) = -\|x\|^2$, it is clear that $\partial_\varepsilon f(\bar{x}) = \emptyset$ at any point $\bar{x}$, if the definition (4) is employed. (This is not surprising, of course, as it considers a subdifferential designed for convex functions on a nonconvex function.) The next example shows that adapting the inequalities of Definition (3) to include inexactness is also unreasonable.

Example 1.1. *Consider the function f given by $f(x) = -x^4$ and fix an error level ε. One method to generate an inexact subdifferential at a point $\bar{x}$ might be to consider*

$$\left\{ \begin{array}{ll} g: & \textit{there exists } \delta > 0 \textit{ such that} \\ & f(x) \geq f(\bar{x}) - \varepsilon + \langle g, x - \bar{x}\rangle, \textit{for all } x \in B_\delta(\bar{x}) \end{array} \right\}.$$

But for the given function f, this set is $\mathbb{R}$, regardless of $\bar{x}$.
Another method to generate an inexact subdifferential at a point $\bar{x}$ might be to fix $\delta > 0$ a priori

and then consider

$$\{g : f(x) \geq f(\bar{x}) - \varepsilon + \langle g, x - \bar{x}\rangle, \textit{ for all } x \in B_\delta(\bar{x})\}$$

Simple calculations show that (for $f(x) = -x^4$ at $\bar{x}$) this set is

$$-4\bar{x}^3 + \left[-6\bar{x}^2\delta + 4\bar{x}\delta^2 - \delta^3 - \frac{\varepsilon}{\delta}, 6\bar{x}^2\delta + 4\bar{x}\delta^2 + \delta^3 + \frac{\varepsilon}{\delta}\right].$$

In particular, the error in the gradient approximations depends on $\bar{x}$ (and δ and ε).

Another method of generating inexact subdifferentials for nonconvex functions was proposed by Goldstein [Gol77].

Definition 1.2 (Goldstein subdifferential)**.** *The* Goldstein ε-approximate subdifferential *is denoted by ∂^G_ε and defined as*

$$\partial^G_\varepsilon f(x) = conv\{\partial f(y) : y \in B_\varepsilon(x)\}.$$

When ε is apparent, we shall refer to this object as the Goldstein subdifferential.

The Goldstein subdifferential is both outer and inner semicontinuous as a multifunction of (x, ε). When $\varepsilon = 0$, the enlargement reduces to the usual subdifferential. Like the Convex Analysis ε-subdifferential (employed in [Kiw06]) and the inexact approximations considered in Example 1.1, the error generated by the Goldstein subdifferential can be shown to be dependent on the local behaviour of the function. In particular, if the Goldstein subdifferntial is applied to a locally affine function, it will return the exact gradient for ε sufficiently small.

Finally, since for locally Lipschitz functions the subdifferential is locally bounded, for any given $y \in B_\varepsilon(x)$ the subgradients $g(y) \in \partial f(y)$ remain in a ball around $\partial f(x)$, with radius $L|y-x|$ where L is the local Lipschitz constant. So the inclusion

$$\partial^G_\varepsilon f(x) \subset \partial f(x) + B_{L\varepsilon}(0)$$

holds.

In this paper we consider inexact subgradients g for the function f at the point x^j of the form

$$g \in \partial f(x) + B_\varepsilon(0). \tag{5}$$

In view of the discussion above, this assumption provides a more consistent form of error on the subgradient values computed by the oracle. We conclude this section by noting that for any $\varepsilon_1, \varepsilon_2 > 0$ we have the following useful inclusion

$$\partial f(\bar{x}) + B_{\varepsilon_1}(0) \subseteq \partial^G_{\varepsilon_2} f(\bar{x}) + B_{\varepsilon_1}(0). \tag{6}$$

2. NOTATION FOR BUNDLE METHOD INGREDIENTS

Due to the technical nature of some of the developments in bundle methods, it is useful to provide a summary of notation upfront.

2.1. **Bundle Information.** The algorithm herein will hinge around a bundle of points. Basic elements include

k	an iteration counter,
$\{x^j\}_{j\in J^k}$	a set of points indexed by J^k,
$\hat{x}^k$	the algorithmic center at iteration k.

The algorithmic center will be one of the bundle points: $\hat{x}^k \in \{x^j\}_{j\in J^k}$. The algorithmic center is essentially the "best" known point up to the k-th iteration.

We shall use functional notation $f(x^j)$ to mean the exact value of the function f at the point x^j. However, the algorithm works with inexact oracle information. So we have inexact function values as follows:

$$\begin{array}{ll} f^j = f(x^j) - \sigma^j & \text{where } \sigma^j \text{ is an unknown error,} \\ \hat{f}^k = f(\hat{x}^k) - \hat{\sigma}^k & \text{where } \hat{\sigma}^k \text{ is an unknown error.} \end{array} \tag{7}$$

Note that the sign of σ^j is not specified, so that the true function value can be either overestimated or underestimated.

Similarly for the subgradient estimates, we have

$$\begin{array}{ll} g^j \in \partial f(x^j) + B_{\varepsilon^j}(0) & \text{where } \varepsilon^j \text{ is an unknown error,} \\ \hat{g}^k \in \partial f(\hat{x}^k) + B_{\hat{\varepsilon}^k}(0) & \text{where } \hat{\varepsilon}^k \text{ is an unknown error.} \end{array} \tag{8}$$

In both cases, the error term (σ^j or ε^j) is unknown, but a bound exists, i.e.,

$$|\sigma^j| \leq \bar{\sigma} \text{ and } 0 \leq \varepsilon^j \leq \bar{\varepsilon} \text{ for all } j.$$

The term *bundle information* will be used to denote

$$\mathscr{B}^k = \{(x^j, f^j, g^j) : j \in J^k\},$$

(recall that $\hat{x}^k = x^j$ for some $j \in J^k$, so the algorithmic center is always included in the bundle information).

2.2. **Model Functions.** In considering the nonconvex function f, we use bundle information to develop a piecewise-linear model. If the function f were convex, then given a point x, any subgradient $g(x) \in \partial f(x)$ would generate a linear lower bound for f: $f(x) + \langle g(x), y - x\rangle \leq f(y)$ for all y. This knowledge gives the classical cutting-plane model

$$\max_{j\in J^k} \left\{ f(x^j) + \left\langle g(x^j), y - x^j \right\rangle \right\} \quad \text{for some index set } J^k \subseteq \{1, \ldots, k\}.$$

However, in our setting, we are working with inexact oracle information. Therefore, we generate an *inexact* piecewise-linear model, defined by

$$\mathscr{M}^k(y) := \max_{j\in J^k} \left\{ f^j + \left\langle g^j, y - x^j \right\rangle \right\} \quad \text{for some index set } J^k. \tag{9}$$

Each new iterate in the algorithm is given by the *proximal point* of the inexact piecewise-linear model at the center $\hat{x}^k$. More precisely, given also a proximal-parameter $t^k > 0$,

$$\begin{array}{rcl} x^{k+1} & := & \arg\min_{y\in C}\{\mathscr{M}^k(y) + \frac{1}{2t^k}|y - \hat{x}^k|^2\} \\ & = & \arg\min_{y\in \mathbf{R}^n}\{\mathscr{M}^k(y) + \mathtt{i}_C(y) + \frac{1}{2t^k}|y - \hat{x}^k|^2\}, \end{array}$$

where the notation $\mathtt{i}_C$ stands for the indicator function of the set C:

$$\mathtt{i}_C(y) = \begin{cases} 0, & \text{if } y \in C, \\ +\infty, & \text{otherwise} . \end{cases}$$

From the optimality conditions of the problem above (which is a quadratic program if C is polyhedral; or assuming a constraint qualification if C is more general), there exists a simplicial multiplier

$$\alpha^k \in \mathbb{R}^{|J^k|}, \quad \alpha_j^k \geq 0, \quad \sum_{j=1}^{|J^k|} \alpha_j^k = 1$$

such that

$$x^{k+1} = \hat{x}^k - t^k(G^k + b^k) \quad \text{where} \quad \begin{cases} G^k := \sum_{j \in J^k} \alpha_j^k g^j \\ b^k \in \partial \mathtt{i}_C(x^{k+1}) . \end{cases} \tag{10}$$

Once the new iterate is known, we define the *inexact aggregate linearization*

$$\mathtt{M}^k(y) := \sum_{j \in J^k} \alpha_j^k \Big(f^j + \langle g^j, y - x^j \rangle \Big) . \tag{11}$$

Clearly,

$$\mathtt{M}^k(x^{k+1}) = \mathscr{M}^k(x^{k+1}) , \quad G^k \in \partial \mathscr{M}^k(x^{k+1}) , \text{ and } \quad G^k = \nabla \mathtt{M}^k(y) \text{ for all } y \in \mathbb{R}^n . \tag{12}$$

Moreover, because multipliers α^k are simplicial, it holds that

$$\mathtt{M}^k(y) \leq \mathscr{M}^k(y) \quad \text{for all } y \in \mathbb{R}^n . \tag{13}$$

3. Aggregate linearization error and criticality measure

When a bundle method is applied to information from an exact oracle for a convex function, an important fact is that the linearization errors are nonnegative. In other words, for an *exact oracle* and a *convex function,* the cutting-plane model underestimates the function at every point. This fact allows for a simple criticality measure for the setting of exact oracle of a convex function. In this section we examine how linearization errors can be adapted for inexact oracles of nonconvex functions, and the criticality measure that follows.

To begin we define the aggregate linearization error $\mathtt{E}^k$ (for the iteration k) by

$$\mathtt{E}^k := \hat{f}^k - \mathtt{M}^k(\hat{x}^k) - \left\langle b^k, \hat{x}^k - x^{k+1} \right\rangle , \tag{14}$$

where all the objects are as defined above. Our first result demonstrates that although these errors may be negative, they are nonetheless uniformly bounded below.

Proposition 3.1 (Error Bound)**.** *Suppose the points $\{x^j\}_{j \in J^k} \subset C$ are used to make a piecewise-linear model $\mathscr{M}^k$ given by* (9)*, where f^j and g^j are as described in* (7) *and* (8)*, respectively. Let the inexact aggregate linearization $\mathtt{M}^k$ be given by* (11) *and let b^k be the normal element in* (10)*.*

Then the error in (14) *has a lower bound, $\mathtt{E}^k \geq -c$, for a positive constant c depending only on the accuracies $\bar{\sigma}$ and $\bar{\varepsilon}$ and on the constraint set C.*

Proof. To bound the error in (14), we start by writing

$$\begin{aligned} -\mathtt{E}^k &= \mathtt{M}^k(\hat{x}) + \left\langle b^k, \hat{x}^k - x^{k+1} \right\rangle - \hat{f}^k \\ &\leq \mathtt{M}^k(\hat{x}) - \hat{f}^k \\ &= \sum_{j \in J^k} \alpha_j^k \left(f^j - \hat{f}^k + \left\langle g^j, \hat{x}^k - x^j \right\rangle \right), \end{aligned}$$

where we used that, by (10), $\hat{x}^k \in C$, $b^k \in \partial \mathtt{i}_C(x^{k+1})$, and (11) for the last relation. Denote

$$T^j := f^j - \hat{f}^k + \left\langle g^j, \hat{x}^k - x^j \right\rangle,$$

for each $j \in J^k$. Because the multiplier α^k is simplicial, it holds that

$$-\mathtt{E}^k \leq \sum_{j \in J^k} \alpha_j^k T^j \leq \max_{j \in J^k} |T^j| \sum_{j \in J^k} \alpha_j^k = \max_{j \in J^k} |T^j|.$$

By (8), for each x^j in the bundle there is an exact subgradient $g(x^j)$ and a vector residue r^j such that $g^j = g(x^j) - r^j$ with $|r^j| \leq \varepsilon^j \leq \bar{\varepsilon}$. Together with the functional relations in (7), we see that for each $j \in J^k$

$$\begin{aligned} |T^j| &= \left| f(x^j) - \sigma^j - f(\hat{x}^k) + \hat{\sigma} + \left\langle g(x^j), \hat{x}^k - x^j \right\rangle - \left\langle r^j, \hat{x}^k - x^j \right\rangle \right| \\ &\leq |f(x^j) - f(\hat{x}^k)| + 2|\bar{\sigma}| + (|g(x^j)| + \bar{\varepsilon})|\hat{x}^k - x^j|. \end{aligned}$$

The bound now follows, as all the iterates remain in the compact set C and locally the function is Lipschitzian so the subdifferential is locally bounded. □

To measure criticality, the algorithm examines the quantity

$$V^k := |G^k + b^k|. \tag{15}$$

To understand this measure, first note that for any approximate g^j subgradient satisfying (8) and any $\Delta \geq |x^j - \hat{x}^k|$, by relation (6) we have that

$$g^j \in \partial f(x^j) + B_{\varepsilon^j}(0) \subseteq \partial_\Delta^G f(\hat{x}^k) + B_{\varepsilon^j}(0) \subseteq \partial_\Delta^G f(\hat{x}^k) + B_{\bar{\varepsilon}}(0).$$

Accordingly, at iteration k define $\Delta^k := \max_{j \in J^k} |x^j - \hat{x}^k|$. This implies that $g^j \in \partial_{\Delta^k}^G f(\hat{x}^k) + B_{\bar{\varepsilon}}(0)$ for each $j \in J^k$, so

$$G^k = \sum_{j \in J^k} \alpha_j^k g^j \in \partial_{\Delta^k}^G f(\hat{x}^k) + B_{\bar{\varepsilon}}(0).$$

Since by (10) the normal element is a subgradient of the indicator function i_C, we see that

$$G^k + b^k \in \partial_{\Delta^k}^G f(\hat{x}^k) + \partial i_C(\hat{x}^k) + B_{\bar{\varepsilon}}(0) \subseteq \partial_{\Delta^k}^G (f + i_C)(\hat{x}^k) + B_{\bar{\varepsilon}}(0). \tag{16}$$

We shall see that the algorithm will generate a sequence such that $V^k = |G^k + b^k| \to 0$. Letting $\hat{x}^{acc}$ and Δ^{acc} denote accumulation points for the corresponding subsequences of $\hat{x}^k$ and $\{\Delta^k\}$ and passing to the limit in the inclusion this will imply that

$$0 \in \partial_{\Delta^{acc}}^G (f + i_b)(\hat{x}^{acc}) + B_{\bar{\varepsilon}}(0). \tag{17}$$

If $\Delta^{acc} = 0$, then x^{acc} is a $\bar{\varepsilon}$-critical point for f on C, while if $\Delta^{acc} > 0$ this states that x^{acc} is within a radius of Δ^{acc} of a $\bar{\varepsilon}$-critical point for f on C.

4. Explanation of noise attenuation

To measure progress towards a minimum and deciding when to change the algorithmic center, we use the *predicted decrease* (by the model), defined by

$$\delta^k := \hat{f}^k - \mathtt{M}^k(x^{k+1})\,. \tag{18}$$

Since the function $\mathtt{M}^k$ is affine and G^k is its gradient by (12), using the error definition in (14) yields that

$$\begin{aligned} \mathtt{E}^k + \delta^k &= \hat{f}^k - \mathtt{M}^k(\hat{x}^k) - \left\langle b^k, \hat{x}^k - x^{k+1}\right\rangle + \hat{f}^k - \mathtt{M}^k(x^{k+1}) \\ &= 2\Big(\hat{f}^k - \mathtt{M}^k(\hat{x}^k)\Big) - \langle G^k - b^k, x^{k+1} - \hat{x}^k\rangle\,. \end{aligned}$$

Noting that $\hat{x}^k - x^{k+1} = t^k(G^k + b^k)$ (equation (10)) and again using the error definition in (14) we have that

$$\mathtt{E}^k + \delta^k = 2\mathtt{E}^k + t^k|G^k + b^k|^2\,. \tag{19}$$

Recall that $\mathtt{E}^k$ represents the aggregate linearization error occurring at the point $\hat{x}^k$. In our algorithm, noise will be deemed too large when the sum of the predicted decrease with the error $\mathtt{E}^k$ is negative:

$$\mathtt{E}^k + \delta^k < 0. \tag{20}$$

In that case, we shall increase the proximal-parameter t_k to reduce the influence of noise, keeping the previous algorithmic center $\hat{x}^k$.

Note that if equation (20) holds, then equation (19) implies that $2\mathtt{E}^k + t^k|G^k + b^k|^2 < 0$, which in turn implies that $\mathtt{E}^k < 0$ and $t^k|G^k + b^k|^2 < -2\mathtt{E}^k$. Thus we have that

$$\delta^k + \mathtt{E}^k < 0 \quad \Rightarrow \quad \mathtt{E}^k < 0 \quad \text{and} \quad V^k \leq \sqrt{\frac{-2\mathtt{E}^k}{t^k}}\,. \tag{21}$$

Conversely, noise is deemed acceptable when

$$\delta^k + \mathtt{E}^k \geq 0. \tag{22}$$

In that case, $\delta^k \geq -\mathtt{E}^k$. Now, if $\mathtt{E}^k < 0$, this implies $\delta^k \geq |\mathtt{E}^k|$. On the other hand, if $\mathtt{E}^k \geq 0$, then equation (19) shows that $\delta^k \geq \mathtt{E}^k + t^k|G^k + b^k|^2 \geq |\mathtt{E}^k|$. Thus equation (22) implies that $\delta^k \geq |\mathtt{E}^k|$. In addition, in this case, equation (19) further implies that

$$\begin{aligned} 2\mathtt{E}^k + t^k|G^k + b^k|^2 &\geq 0 \\ \mathtt{E}^k &\geq -\tfrac{t^k}{2}|G^k + b^k|^2\,. \end{aligned}$$

Noting that $\delta^k = \mathtt{E}^k + t^k|G^k + b^k|^2$ yields that $\delta^k \geq \frac{t^k}{2}|G^k + b^k|^2$. In summary,

$$\delta^k + \mathtt{E}^k \geq 0 \quad \Rightarrow \quad \delta^k \geq \max\Big\{|\mathtt{E}^k|, \frac{t^k}{2}|G^k + b^k|^2\Big\} \geq 0\,. \tag{23}$$

We note that the descent test will be applied only when the noise is not too large, in which case the predicted descent δ^k is non-negative.

5. An inexact algorithm

The material presented above provides the necessary tools for dealing with both the nonconvexity of the objective function f and the inexactness of the oracle.

In the algorithm below, we first check if the noise is deemed acceptable. If equation (20) is satisfied, then the noise is deemed too large. In this case, we increase the proximal-parameter and compute the proximal point over the same cutting-plane model at the same algorithmic center with the new prox-parameter. If equation (20) is not satisfied, then the noise is deemed acceptable and we compare the descent for the (inexact) objective function values to the predicted δ^k (which will be non-negative). If the descent is better than the given fraction of δ^k, then a serious step is declared and the current iterate is accepted as the new algorithmic center. If descent was not sufficient, then a null step is declared.

We provide the resulting algorithmic framework now.

Inexact Proximal Bundle Method.
Oracle. An inexact oracle is given, providing for each x the values (f, g) approximating $(f(x), g(x))$ with $g(x) \in \partial f(x)$.

Step 0 (initialization).

Choose a starting point $x^1 \in \mathbb{R}^n$, compute f^1 and g^1, and set the initial index set $J^1 := \{1\}$.

Initialize the iteration counter $k = 1$, the noise-attenuation, serious-step, and null-steps counters $\mathtt{A}^1 := 0$, $\mathtt{S}^1 := 0$, and $\mathtt{N}^1 := 0$.

Select a stopping tolerance $\varepsilon_V \geq 0$, a descent parameter $m \in (0,1)$, a minimal prox-parameter $t_{\min} > 0$.

Select an initial prox-parameter $t^1 \geq t_{\min}$. Set $\hat{f}^1 = f^1$ and the initial prox-center $\hat{x}^1 := x^1$.

Step 1 (trial point finding).

Compute

$$x^{k+1} = \arg\min_{y \in C}\{\mathscr{M}^k(y) + \frac{1}{2t^k}|y - \hat{x}^k|^2\},$$

where $\mathscr{M}^k$ is given by (9).

Compute

$$\begin{aligned} \mathtt{E}^k &:= \hat{f}^k - \mathtt{M}^k(\hat{x}^k) - \langle b^k, \hat{x}^k - x^{k+1} \rangle, \\ V^k &:= |G^k + b^k|, \text{ and} \\ \delta^k &:= \hat{f}^k - \mathscr{M}^k(x^{k+1}) = \hat{f}^k - \mathtt{M}^k(x^{k+1}), \end{aligned}$$

where $\mathtt{M}^k$ is given by (11) and $G^k + b^k$ is given by (10).

Step 2 (noise attenuation and unboundedness detection).

If $\delta^k + \mathtt{E}^k < 0$, then noise needs to be attenuated by increasing t^k:

- Maintain $\hat{x}^{k+1} = \hat{x}^k$, $\hat{f}^{k+1} = \hat{f}^k$, the bundle $J^{k+1} = J^k$ (and, hence, the model $\mathscr{M}^{k+1} = \mathscr{M}^k$), increase $t^{k+1} = 10t^k$, set $\mathtt{A}^{k+1} = \mathtt{A}^k + 1$, $\mathtt{S}^{k+1} = \mathtt{S}^k$, $\mathtt{N}^{k+1} = \mathtt{N}^k$, increase k by 1, and go to Step 1.

Step 3 (stopping criterion).
If $V^k \leq \varepsilon_V$, then stop.
Step 4 (descent test).
Call the oracle at x^{k+1} to obtain (f^{k+1}, g^{k+1}).
If $f^{k+1} > \hat{f}^k - m\delta^k$, then declare the iteration a null-step and go to Step 5.
Otherwise, declare the iteration a serious-step and set $\hat{x}^{k+1} := x^{k+1}$, $\hat{f}^{k+1} := f^{k+1}$, select $t^{k+1} \geq t_{\min}$, set $\mathtt{S}^{k+1} = k+1$, $\mathtt{A}^{k+1} := 0$, $\mathtt{N}^{k+1} := 0$, and go to Step 6.
Step 5 (null-step).
Set $\hat{x}^{k+1} := \hat{x}^k$, $\hat{f}^{k+1} := \hat{f}^k$, $\mathtt{S}^{k+1} = \mathtt{S}^k$, and $\mathtt{A}^{k+1} = \mathtt{A}^k$. Increase $\mathtt{N}^{k+1} := \mathtt{N}^k + 1$. If $\mathtt{A}^k \neq 0$, set $t^{k+1} = t^k$; otherwise choose $t_{\min} \leq t^{k+1} \leq t^k$.
Step 6 (bundle update and loop).
Select the new bundle index set J^{k+1}. Increase k by 1 and go to Step 1.

□

6. Asymptotic convergence analysis

In this section we examine the asymptotic properties of the algorithm. We assume that the algorithm loops forever, so that $k \to \infty$ (i.e., the stopping test in Step 3 is removed). Noting that $\mathtt{A}^k$ is non-decreasing during null steps, this leads to three mutually exclusive cases:

Case 1: there are infinitely many serious iterates ($\mathtt{S}^k \to \infty$),
Case 2: the prox-center is unchanged after a finite number of iterations ($\mathtt{S}^k = \widehat{\mathtt{S}}$ for all k large), and there are an infinite number of noise attenuation steps ($\mathtt{A}^k \to \infty$), or
Case 3: the prox-center is unchanged after a finite number of iterations ($\mathtt{S}^k = \widehat{\mathtt{S}}$ for all k large), and there are a finite number of noise attenuation steps ($\mathtt{A}^k = \widehat{\mathtt{A}}$ for all k large).

We consider each of these cases now, showing that, in all the three cases the criticality measure V^k is driven to 0.

6.1. Infinite Serious Steps.

We begin with Case 1, i.e., an infinite number of serious steps.

Theorem 6.1 (Infinitely many serious iterates)**.** *Consider the Inexact Proximal Bundle Method given in Section 5 applied to solve* (1) *for a function satisfying* (2)*. Suppose $\mathtt{S}^k \to \infty$ and let K_s denote the indices of iterations that give a serious step (i.e., $\mathtt{S}^k = k$). Then $\delta^k \to 0$ as $K_s \ni k \to \infty$ and*

$$\lim_{K_s \ni k \to \infty} V^k = 0.$$

Proof. At each serious step $k \in K_s$ we have

$$\hat{f}^{k+1} \leq \hat{f}^k - m\delta^k \quad \text{and} \quad \delta^k \geq 0, \tag{24}$$

where the second inequality follows from the fact that a serious step can only take place when no noise attenuation occurs (thus we are in the case of (23)). It follows that the sequence $\{\hat{f}^k\}_{K_s}$ is nonincreasing.

Since the sequence $\{\hat{x}^k\} \subset C$ is bounded, by our assumption on f and σ^k the sequence $\{f(\hat{x}^k) - \hat{\sigma}^k\}_{K_s}$ is bounded below, i.e., $\{\hat{f}^k\}_{K_s}$ is bounded below. Since $\{\hat{f}^k\}_{K_s}$ is also nonincreasing, we conclude that it converges.

Equation (24) now implies that $\delta^k \to 0$ as $K_s \ni k \to \infty$. By equation (23), it then follows that

$$\mathtt{E}^k \to 0 \quad \text{and} \quad t_k|G^k + b^k|^2 \to 0 \quad \text{as} \quad K_s \ni k \to \infty.$$

Using the fact that $t^k \geq t_{\min} \geq 0$, we conclude that

$$V^k = |G^k + b^k| \to 0 \quad \text{as} \quad K_s \ni k \to \infty.$$

□

In the next two cases, a finite number of serious steps occurs. That is, after a finite number of iterations the algorithmic center remains unchanged, so that there exists $\hat{k}$ and $\hat{x}$ such that $\hat{x}^k = \hat{x}$ for all $k > \hat{k}$.

6.2. **Finite Serious Steps with Infinite Noise Attenuation.**

We now examine the Case 2, i.e., a finite number of serious steps with an infinite sequence of noise attenuation steps.

Theorem 6.2 (Finitely many serious iterates with infinite noise attenuation steps)**.** *Consider the Inexact Proximal Bundle Method given in Section 5 applied to solve* (1) *for a function satisfying* (2). *Suppose that* $\mathtt{S}^k = \widehat{\mathtt{S}}$ *for all k sufficiently large, and suppose that* $\mathtt{A}^k \to \infty$. *If* K_a *denote those iterations when inaccuracy is detected at Step 2, then*

$$\lim_{K_a \ni k \to \infty} V^k = 0.$$

Proof. First note that if $k \in K_a$, then $\delta^k + \mathtt{E}^k < 0$. By equation (21), we know that this implies that $\mathtt{E}^k < 0$ and

$$0 \leq V^k \leq \sqrt{-2\mathtt{E}^k/t^k}.$$

By Proposition 3.1, the quantities $(-\mathtt{E}^k)$ are uniformly bounded below; thus $\{|\mathtt{E}^k|\}$ is bounded above for $k \in K_a$. Moreover, once all serious steps have been done, the sequence $\{t^k\}$ is non-decreasing (according to Step 2, and to Step 5 where $\mathtt{A}^k \neq 0$). Moreover, as $\mathtt{A}^k \to \infty$, by Step 2 we have that $t^k \to \infty$. Thus, we conclude that $V^k \to 0$ as $K_a \ni k \to \infty$. □

6.3. **Finite Serious Steps with Finite Noise Attenuation.**

The remaining case is when the algorithm generates finitely many serious and noise attenuation iterates. Recall that the counter $\mathtt{A}^k$, of noise attenuation steps, is non-decreasing during null steps. Since the counter is reset to zero at serious iterations, if a finite number of serious steps occurs, then $\mathtt{A}^k$ is non-decreasing after a finite number of iterations. In this case, if $\mathtt{A}^k$ is bounded above, then $\mathtt{A}^k$ must reach its finite limit after a finite number of iterations. After this point, no noise attenuation occurs (so $\delta^k + \mathtt{E}^k \geq 0$) and $\mathtt{A}^k$ will remain unchanged, say at a value $\widehat{\mathtt{A}}$, in all future iterations. We start with some preliminary results, and state additional assumptions needed in this last case.

Proposition 6.3 (Properties of value function)**.** *In the setting of algorithm in Section 5, the following relations hold.*

(i) $2t^k\langle G^k + b^k, y - x^{k+1}\rangle = |x^{k+1} - \hat{x}|^2 + |y - x^{k+1}|^2 - |y - \hat{x}|^2$ *for all* $y \in \mathbb{R}^n$.

(ii) *For the optimal value of the subproblem solved in Step 1,*

$$\psi^k := \mathscr{M}^k(x^{k+1}) + \frac{1}{2t_k}|x^{k+1} - \hat{x}^k|^2,$$

it holds that $\mathtt{M}^k(y) + \left\langle b^k, y - x^{k+1}\right\rangle + \frac{1}{2t^k}|y - \hat{x}|^2 = \psi^k + \frac{1}{2t^k}|y - x^{k+1}|^2$ *for all* $y \in \mathbb{R}^n$.

Proof. The first item is by direct calculations, using equation (10). To show the second item, recall from equations (11) and (12) that $\mathtt{M}^k$ is linear and G^k is its gradient. Then, using item (i) and the definition of ψ^k, we see that

$$\begin{aligned}
\mathtt{M}^k(y) &= \mathtt{M}^k(x^{k+1}) + \langle G^k, y - x^{k+1}\rangle \\
&= \mathtt{M}^k(x^{k+1}) - \left\langle b^k, y - x^{k+1}\right\rangle + \langle G^k + b^k, y - x^{k+1}\rangle \\
&= \mathtt{M}^k(x^{k+1}) - \left\langle b^k, y - x^{k+1}\right\rangle + \tfrac{1}{2t^k}|x^{k+1} - \hat{x}|^2 + \tfrac{1}{2t^k}|y - x^{k+1}|^2 - \tfrac{1}{2t^k}|y - \hat{x}|^2 \\
&= \psi^k - \left\langle b^k, y - x^{k+1}\right\rangle + \tfrac{1}{2t^k}|y - x^{k+1}|^2 - \tfrac{1}{2t^k}|y - \hat{x}|^2,
\end{aligned}$$

where the final equality applies the fact that $\mathscr{M}^k(x^{k+1}) = \mathtt{M}^k(x^{k+1})$. □

The previous results did not make use of any specific condition on the bundle index set J^k. We now assume the bundle management subalgorithm in Step 6 is such that at two consecutive null steps, k and $k+1$, the choice of J^{k+1} ensures that, for all $y \in C$

$$\boxed{\begin{array}{ll}
(25) & \mathscr{M}^{k+1}(y) \geq f^{k+1} + \left\langle g^{k+1}, y - x^{k+1}\right\rangle, \text{ and} \\
(26) & \mathscr{M}^{k+1}(y) \geq \mathtt{M}^k(y).
\end{array}}$$

It is worth noting that conditions (25) and (26) are both easily achievable. The first one says that the newly computed information always enters the bundle; while the second means that if any past information is being removed, the aggregate linearization must replace it (if no information is removed, this condition holds automatically).

We are now in position to consider the last case.

Theorem 6.4 (Finite serious and noise attenuation steps with infinitely many null steps)**.** *Consider the Inexact Proximal Bundle Method given in Section 5 applied to solve* (1) *for a function satisfying* (2). *Suppose that a finite number of serious iterates and noise attenuation steps occur, so that for all* $k \geq \hat{k}$ *we have* $\mathtt{S}^k = \widehat{\mathtt{S}}$, $\mathtt{A}^k = \widehat{\mathtt{A}}$ *(and thus* $\delta^k + \mathtt{E}^k \geq 0$*). Let* $\hat{x}$ *be the point found by the last serious iteration (i.e.,* $\hat{x}^k = \hat{x}$ *for all* $k \geq \hat{k}$*), and let* $\hat{f}$ *denote the corresponding functional value. If the cutting-planes models satisfy conditions* (25) *and* (26) *for all* $k \geq \bar{k}$*, then*

$$\lim_{k\to\infty} x^k = \hat{x}, \quad \lim_{k\to\infty} \mathscr{M}^k(x^{k+1}) = \hat{f}, \quad \lim_{k\to\infty} V^k = 0.$$

Proof. We assume that $k \geq \hat{k}$ is large enough, so that $\hat{x}^k = \hat{x}$, $\hat{f}^k = \hat{f}$, and no noise attenuation occurs so $\{t^k\}$ is a non-increasing sequence bounded above by some $\hat{t}$ (according to Step 5).

We first show that the optimal value sequence ψ^k defined in Proposition 6.3 is bounded above. Writing Proposition 6.3 item (ii) for $y = \hat{x}$ and applying Proposition 3.1 we see that

$$\psi^k + \frac{1}{2t_k}|\hat{x} - x^{k+1}|^2 = \mathtt{M}^k(\hat{x}) + \left\langle b^k, \hat{x} - x^{k+1}\right\rangle = \hat{f} - \mathtt{E}^k \leq \hat{f} + c.$$

In particular, $\psi^k \leq \hat{f} + c$, so the sequence $\{\psi^k\}$ is bounded above.

We next show that ψ^k is increasing. Combining item (ii) in Proposition 6.3 with our assumed condition (26) yields the inequality

$$\mathscr{M}^{k+1}(y) + \left\langle b^k, y - x^{k+1} \right\rangle + \frac{1}{2t^k}|y - \hat{x}|^2 \geq \psi^k + \frac{1}{2t^k}|y - x^{k+1}|^2 \text{ for all } y \in C.$$

Since $b^k \in \partial \mathtt{i}_C(x^{k+1})$ (by equation (10)) and $t^k \geq t^{k+1}$, we obtain that

$$\forall y \in C \quad \mathscr{M}^{k+1}(y) + \frac{1}{2t^{k+1}}|y - \hat{x}|^2 \geq \psi^k + \frac{1}{2t^k}|y - x^{k+1}|^2 .$$

In particular, when $y = x^{k+2}$, this means that

$$\psi^{k+1} \geq \psi^k + \frac{1}{2t^k}|x^{k+2} - x^{k+1}|^2. \tag{27}$$

As the sequence $\{\psi^k\}$ is bounded and increasing, it must converge and consequently, by equation (27), it follows that $|x^{k+2} - x^{k+1}|^2 \to 0$.

We now show that $\delta^k \to 0$, by examining the limit of the auxiliary error sequence

$$e^k := f^{k+1} - \mathscr{M}^k(x^{k+1}).$$

By equation (25), written with $y = x^{k+2}$,

$$\begin{array}{rcl} e^k & \leq & \mathscr{M}^{k+1}(x^{k+2}) - \left\langle g^{k+1}, x^{k+2} - x^{k+1} \right\rangle - \mathscr{M}^k(x^{k+1}) \\ & = & \mathtt{M}^{k+1}(x^{k+2}) - \mathtt{M}^k(x^{k+1}) - \left\langle g^{k+1}, x^{k+2} - x^{k+1} \right\rangle , \end{array}$$

where in the last equality we used the first identity in (12). Writing item (ii) in Proposition 6.3, first with $y = x^{k+1}$ and then with k replaced by $k+1$ and $y = x^{k+2}$, we see that

$$\begin{array}{rcl} e^k & \leq & \psi^{k+1} - \frac{1}{2t^{k+1}}|x^{k+2} - \hat{x}|^2 - \psi^k + \frac{1}{2t^k}|x^{k+1} - \hat{x}|^2 - \left\langle g^{k+1}, x^{k+2} - x^{k+1} \right\rangle \\ & \leq & \psi^{k+1} - \psi^k + \frac{1}{2t^k}\left(|x^{k+1} - \hat{x}|^2 - |x^{k+2} - \hat{x}|^2 \right) - \left\langle g^{k+1}, x^{k+2} - x^{k+1} \right\rangle , \end{array}$$

because proximal parameters do not increase at null steps. By writing item (i) in Proposition 6.3 with $y = x^{k+2}$, the difference of squares above is equal to $2t^k \left\langle G^k + b^k, x^{k+2} - x^{k+1} \right\rangle - |x^{k+2} - x^{k+1}|^2$. Since, in addition, $\left\langle b^k, x^{k+2} - x^{k+1} \right\rangle \leq 0$, we obtain that

$$e^k \quad \leq \quad \psi^{k+1} - \psi^k + \left\langle G^k, x^{k+2} - x^{k+1} \right\rangle - \left\langle g^{k+1}, x^{k+2} - x^{k+1} \right\rangle .$$

As $t^k \leq \hat{t}$ and both G^k and g^{k+1} are bounded (recall that f is locally Lipschitzian), passing to the limit we see that

$$\limsup_{k \to \infty} e^k \leq 0 . \tag{28}$$

Since $\delta^k + \mathtt{E}^k \geq 0$, we have that $\delta^k \geq 0$ (see equation (23)), thus

$$\delta^k = \hat{f}^k - \mathtt{M}^k(x^{k+1}) = \hat{f} - \mathscr{M}^k(x^{k+1}) \geq 0.$$

As we are in a null step, we have $f^{k+1} > \hat{f}^k - m\delta^k$, so

$$\delta^k = \hat{f} - \mathscr{M}^k(x^{k+1}) < f^{k+1} + m\delta^k - \mathscr{M}^k(x^{k+1}),$$

implying that (using $m < 1$)

$$0 \leq (1-m)\delta^k < f^{k+1} - \mathscr{M}^k(x^{k+1}) = e^k.$$

By equation (28) we conclude that

$$\lim_{k\to\infty} \delta^k = 0\,.$$

By the definition of δ^k we have $\mathscr{M}^k(x^{k+1}) \to \hat{f}$, as stated. By equation (23), this implies that both $|\mathtt{E}^k| \to 0$ and $t^k|G^k + b^k|^2 \to 0$. Since $t^k \geq t_{\min}$ the later of these implies that $V^k = |G^k + b^k| \to 0$. Finally, by definition (10) we know that $x^{k+1} - \hat{x} = t^k(G^k + b^k)$, so that $x^k \to \hat{x}$ holds. □

6.4. **Bundle management.** Combining the convergence results above leads to the following corollary.

Corollary 6.5. *Consider the Inexact Proximal Bundle Method given in Section 5 applied to solve*

$$\min\{f(x) : x \in C\}$$

for a function satisfying (2). *Suppose that at any two consecutive null steps bundle management conditions* (25) *and* (26) *hold. If the algorithm loops forever, then there exists a cluster point x^{acc} of the subsequence of serious iterates satisfying the approximate criticality condition* (17) *with $\Delta^{acc} \geq 0$.*

Proof. Equations (16) and (15) tell us that $G^k + b^k \in \partial^G_{\Delta^k}(f + i_C)(\hat{x}^k) + B_{\bar{\varepsilon}}(0)$. Theorems 6.1, 6.2, and 6.4 ensure that $G^k + b^k \to 0$ as $K \ni k \to \infty$ for some infinite index set K. Since $\Delta^k = \max_{J^k} |x^j - \hat{x}^k|$ with $x^j, \hat{x}^k \in C$, the bounded subsequences $\{\hat{x}^k\}_K$ and $\{\Delta^k\}_K$ have accumulation points, say x^{acc} and Δ^{acc} (in the case of a finite number of serious iterates, x^{acc} is just $\hat{x}^k$ which stays fixed for all k large enough). The assertion follows. □

In principle, the above assertion may not be very compelling without some knowledge of Δ^{acc}. Therefore, in view of Corollary 6.5, the Inexact Proximal Bundle Method should choose the new index set J^{k+1} in a manner that makes Δ^k small enough asymptotically or even drives it to 0. At the same time, the bundle management strategy should ensure the satisfaction of conditions (25) and (26) at consecutive null steps.

For convex objective functions, a common practice ensuring the conditions above is to keep the so-called strongly active bundle elements. This is the set $J^k_{act} := \{j \in J^k : \alpha^k_j > 0\}$ where the simplicial multiplier α^k is computed when solving the subproblem (10) in Step 1. Together with (25), this means that $J^{k+1} = \{k+1\} \cup J^k_{act}$.

However, in our nonconvex inexact setting, bundle management is more involved. In particular, we must consider the three different situations in Theorems 6.1, 6.2 and 6.4, and reconcile the requirements of each of the situations. The following considerations do that.

If the algorithm generates infinitely many serious steps, no special conditions are needed for the bundle index set. Therefore, after each serious step we can select elements close enough to the algorithmic center, localizing the bundle set, for example as follows:

$$J^{k+1} = \{j \in \{k+1\} \cup J^k_{act} : |x^j - \hat{x}^{k+1}| \leq \theta V^{\mathtt{S}^{k+1}}\}$$

for some factor $\theta > 0$, (well) chosen at the initialization step. Since as $K_s \ni k \to \infty$ the criticality measure $V^{\mathtt{S}^{k+1}}$ goes to zero, the localization ensures that so does the corresponding diameter, as desired.

If there is a finite number of serious steps followed by infinitely many noise attenuation steps, as in the infinitely many serious steps case, no special conditions are needed for the bundle index set to ensure convergence. So, after each noise attentuation step, the same bundle management strategy as above can be used.

Finally, when there is a finite number of serious and noise steps, and infinitely many null steps, conditions (25) and (26) need to be satisfied from one null step to another. In that case, the following strategy can be used in order to drive Δ^k to zero: the bundle needs to be *restarted* after each first null step that follows a serious or noise attenuation step, and along the subsequent null iterations all "old" points in J^k_{act} such that the difference between the index $j \in J^k_{act}$ and the current iteration index $k+1$ is larger than a chosen fixed bound P need to be deleted. Specifically, when $\mathtt{N}^{k+1} = 1$ Step 6 takes $J^{k+1} = \{k+1\}$; and for $\mathtt{N}^{k+1} > 1$ Step 6 takes J^{k+1} as $\{k+1\} \cup \{j \in J^k_{act} : k+1-j \leq P\}$ plus the new aggregate function (and if the bundle contained a previously added aggregate function, it is deleted from it). This choice ensures that the two conditions (25) and (26) required for our convergence result are satisfied. In addition, with this choice, the diameter Δ^k of information at points employed tends to zero because, by Theorem 6.4, the whole tail of the sequence $\{x^k\}$ converges to $\hat{x}$ and the bundle given by J^k contains only points within fixed "index-distance" from each index k.

7. Concluding remarks

We presented a proximal bundle method for nonconvex functions capable to handle inexact oracles. The method applies a noise attenuation step that is similar to the 'stepsize correction' in [Kiw06, Step 3]. However, the situation examined in [Kiw06] relies heavily on convexity and does not easily carry over to nonconvex problems. Most notably, the algorithm in [Kiw06] is designed for the inexactness of the form of the convex ε-subdifferential of equation 4. Conversely, the algorithm herein is designed for the more general inexact subdifferential of equation (5).

Other recent work in this area includes [Nol12], which also develops a bundle method for nonconvex functions given by inexact oracles. Contrary to the work of [Nol12], the algorithm herein makes use of noise attenuation ideas without down-shifting linearization errors to deal with nonconvexity.

References

[AFGG11] A. Astorino, A. Frangioni, M. Gaudioso, and E. Gorgone. Piecewise-quadratic approximations in convex numerical optimization. *SIAM J. Optim.*, 21:1418–1438, 2011.

[BKS08] A. M. Bagirov, B. Karasözen, and M. Sezer. Discrete gradient method: derivative-free method for nonsmooth optimization. *J. Optim. Theory Appl.*, 137(2):317–334, 2008.

[CSV09] A. R. Conn, K. Scheinberg, and L. N. Vicente. *Introduction to Derivative-Free Optimization*, volume 8 of *MPS/SIAM Book Series on Optimization*. SIAM, 2009.

[ES10] G. Emiel and C. Sagastizábal. Incremental-like bundle methods with application to energy planning. *Computational Optimization and Applications*, 46:305–332, 2010.

[Gol77] A. A. Goldstein. Optimization of Lipschitz continuous functions. *Math. Programming*, 13(1):14–22, 1977.

[Gup77] A. M. Gupal. A method for the minimization of almost differentiable functions. *Kibernetika (Kiev)*, 1:114–116, 1977.

[Hin01] M. Hintermüller. A proximal bundle method based on approximate subgradients. *Comp. Optim. Appl.*, 20:245–266, 2001.

[HM12] W. Hare and M. Macklem. Derivative-free optimization methods for finite minimax problems. *Opt. Methods and Soft. (to appear)*, 2012.

[HS10] Warren Hare and Claudia Sagastizábal. A redistributed proximal bundle method for nonconvex optimization. *SIAM J. Optim.*, 20(5):2442–2473, 2010.

[HUL93] J.-B. Hiriart-Urruty and C. Lemaréchal. *Convex analysis and minimization algorithms. II*, volume 306 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 1993. Advanced theory and bundle methods.

[Kiw85] K. C. Kiwiel. A linearization algorithm for nonsmooth minimization. *Math. Oper. Res.*, 10(2):185–194, 1985.

[Kiw95] K. C. Kiwiel. Approximations in proximal bundle methods and decomposition of convex programs. *J. Optim. Theory Appl.*, 84:529–548, 1995.

[Kiw96] K. C. Kiwiel. Restricted step and Levenberg-Marquardt techniques in proximal bundle methods for nonconvex nondifferentiable optimization. *SIAM J. Optim.*, 6(1):227–249, 1996.

[Kiw06] K. C. Kiwiel. A proximal bundle method with approximate subgradient linearizations. *SIAM J. Optim.*, 16(4):1007–1023, 2006.

[Kiw10] Krzysztof C. Kiwiel. A nonderivative version of the gradient sampling algorithm for nonsmooth nonconvex optimization. *SIAM J. Optim.*, 20(4):1983–1994, 2010.

[Lem75] C. Lemaréchal. An extension of Davidon methods to non differentiable problems. *Math. Programming Stud.*, 3:95–109, 1975. Nondifferentiable optimization.

[Lem78] C. Lemaréchal. Bundle methods in nonsmooth optimization. In *Nonsmooth optimization (Proc. IIASA Workshop, Laxenburg, 1977)*, volume 3 of *IIASA Proc. Ser.*, pages 79–102. Pergamon, Oxford, 1978.

[Lem01] C. Lemaréchal. Lagrangian relaxation. In *Computational combinatorial optimization (Schloß Dagstuhl, 2000)*, volume 2241 of *Lecture Notes in Comput. Sci.*, pages 112–156. Springer, Berlin, 2001.

[LSB81] C. Lemaréchal, J.-J. Strodiot, and A. Bihain. On a bundle algorithm for nonsmooth optimization. In *Nonlinear programming, 4 (Madison, Wis., 1980)*, pages 245–282. Academic Press, New York, 1981.

[LV98] L. Lukšan and J. Vlček. A bundle-Newton method for nonsmooth unconstrained minimization. *Math. Programming*, 83(3, Ser. A):373–391, 1998.

[Mif77] R. Mifflin. Semismooth and semiconvex functions in constrained optimization. *SIAM J. Control Optimization*, 15(6):959–972, 1977.

[Mif82a] R. Mifflin. Convergence of a modification of Lemaréchal's algorithm for nonsmooth optimization. In *Progress in nondifferentiable optimization*, volume 8 of *IIASA Collaborative Proc. Ser. CP-82*, pages 85–95. Internat. Inst. Appl. Systems Anal., Laxenburg, 1982.

[Mif82b] R. Mifflin. A modification and extension of Lemarechal's algorithm for nonsmooth minimization. *Math. Programming Stud.*, 17:77–90, 1982. Nondifferential and variational techniques in optimization (Lexington, Ky., 1980).

[MN92] M. M. Mäkelä and P. Neittaanmäki. *Nonsmooth optimization*. World Scientific Publishing Co. Inc., River Edge, NJ, 1992. Analysis and algorithms with applications to optimal control.

[NB10] A. Nedić and D. P. Bertsekas. The effect of deterministic noise in subgradient methods. *Math. Program.*, 125:75–99, 2010.

[NLT00] H. Ngai, D. Luc, and M. Théra. Approximate convex functions. *J. Nonlinear Convex Anal.*, 1(2):155–176, 2000.

[Nol12] D. Noll. Bundle method for non-convex minimization with inexact subgradients and function values. In *Computational and Analytical Mathematics*, 2012. Springer Proceedings in Mathematics.

[OS12] W. L. Oliveira and C. Sagastizábal. Level bundle methods for oracles with on-demand accuracy. Technical report, 2012. Available at `http://www.optimization-online.org/DB_HTML/2012/03/3390.html`.

[OSS11] W. L. Oliveira, C. Sagastizábal, and S. Scheimberg. Inexact bundle methods for two-stage stochastic programming. *SIAM J. Optim.*, 21:517–544, 2011.

[RC04] C. P. Robert and G. Casella. *Monte Carlo statistical methods*. Springer Texts in Statistics. Springer-Verlag, New York, second edition, 2004.

[RW98] R. T. Rockafellar and J. J.-B. Wets. *Variational analysis*, volume 317 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 1998.

[Sag12] C. Sagastizábal. Divide to conquer: decomposition methods for energy optimization. *Math. Program. B*, 2012. Accepted for publication.

[Sol03] M. V. Solodov. On approximations with finite precision in bundle methods for nonsmooth optimization. *J. Optim. Theory Appl.*, 119:151–165, 2003.

[SS05] C. Sagastizábal and M. Solodov. An infeasible bundle method for nonsmooth convex constrained optimization without a penalty function or a filter. *SIAM J. Optim.*, 16:146–169, 2005.

[SZ98] M. V. Solodov and S. K. Zavriev. Error stabilty properties of generalized gradient-type algorithms. *J. Optim. Theory Appl.*, 98:663–680, 1998.

[ZFEM12] Victor Zverovich, Csaba Fábian, Eldon Ellison, and Gautam Mitra. A computational study of a solver system for processing two-stage stochastic lps with enhanced benders decomposition. *Mathematical Programming Computation*, pages 1–28, online first, 2012. 10.1007/s12532-012-0038-z.

[ZPR00] Golbon Zakeri, Andrew B. Philpott, and David M. Ryan. Inexact cuts in Benders decomposition. *SIAM J. Optim.*, 10(3):643–657 (electronic), 2000.