Comput Optim Appl
DOI 10.1007/s10589-015-9762-4

A proximal bundle method for nonsmooth nonconvex functions with inexact information

W. Hare[1] · **C. Sagastizábal**[2] · **M. Solodov**[2]

Received: 2 July 2012

Abstract For a class of nonconvex nonsmooth functions, we consider the problem of computing an approximate critical point, in the case when only *inexact* information about the function and subgradient values is available. We assume that the errors in function and subgradient evaluations are merely bounded, and in principle need not vanish in the limit. We examine the redistributed proximal bundle approach in this setting, and show that reasonable convergence properties are obtained. We further consider a battery of difficult nonsmooth nonconvex problems, made even more difficult by introducing inexactness in the available information. We verify that very satisfactory outcomes are obtained in our computational implementation of the inexact algorithm.

Keywords Nonsmooth optimization · Nonconvex optimization · Bundle method · Inexact information · Locally Lipschitz functions · Proximal point

Mathematics Subject Classification 90C25 · 49J52 · 65K10 · 49M05

C. Sagastizábal—Visiting Researcher.

✉ M. Solodov
solodov@impa.br

W. Hare
warren.hare@ubc.ca

C. Sagastizábal
sagastiz@impa.br

[1] University of British Columbia, Okanagan Campus, 3333 University Way, Kelowna, BC V1Y 8C5, Canada

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

Published online: 23 May 2015

1 Introduction

In this paper we seek to approximately solve the problem

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

where $f : \mathbb{R}^n \to \mathbb{R}$ is a locally Lipschitz function and $D \subset \mathbb{R}^n$ is a convex compact set.

The information at disposal is provided by an *inexact* procedure that, given a point x^i, returns some estimations for the function value at x^i and for one subgradient at x^i. Accordingly, the available information is $f^i \approx f(x^i)$ and $g^i \approx g(x^i) \in \partial f(x^i)$. Working with inexact information presents a natural challenge in a number of modern applications. In this paper, we shall assume that the inexact information is provided in a manner such that the errors in the function and subgradient values are bounded by universal constants (see Sect. 2.1 for mathematical details). While the algorithm and analysis do not require these constants to be known, they do require them to exist across the entire (compact) constraint set. This assumption is not restrictive; it encompasses several useful situations. Clearly, if the information is exact, then the assumption holds trivially (with all errors bounded by 0). Three, more interesting, examples include derivative-free optimization, Large-scale Lagrangian or Semidefinite relaxations, and stochastic simulations. We discuss these next.

Example 1 (*Derivative-free optimization*) Suppose $f \in \mathcal{C}^2$, and a procedure is provided that returns exact function values, but does not return any gradient information. This is the framework for the large research area of derivative-free optimization (DFO) [5]. One common technique in DFO is to approximate the gradients using finite differences, linear interpolation, or some other approximation technique. Numerical analysis and DFO contain a ripe literature on how to approximate gradients, and more importantly error bounds for various approximation techniques (see [5, § 2–5] for a few examples). Similar error bounds exist for a variety of (sub-)gradient approximation techniques [3,5,14,16,17,26]. In general, error bounds are based on the Lipschitz constant of the true gradient, the geometry of the sample set (the set of points used to create the approximation), and the diameter of the sample set. As the sample set is created by the user, its geometry and diameter are assumed to be controlled. The compactness of D can be used to assume a universal bound on the Lipschitz constant, and thereby create a universal error bound for the approximated gradients. In this case the exact value of the universal constant would be unknown, as it would depend on the bound for the Lipschitz constant of the true gradient, but the bounding constant itself is known to exist.

Example 2 (Large-scale Lagrangian or Semidefinite relaxations) Another example arises when solving large-scale or difficult problems by Lagrangian or Semidefinite relaxations, which amount to minimizing a function of the form $f(x) := \sup\{F_z(x) : z \in Z\}$ for functions $F_z(\cdot)$ that are usually smooth but sometimes may be nonconvex, [29,33,40]. In some applications it may be impossible to evaluate f precisely but controllable accuracy is easily obtained; in particular when the set Z is bounded, such is the case in [10,44]. A similar situation arises in H_∞-control, as presented in [1,39].

In [1] it is argued that certain nonconvex functions can be locally approximated by use of the support function of a compact set. A detailed explanation of this approximation technique is given in [39, §1.9]. An error bound of the form required in this paper is provided in [1, Lem2.1] and proof that the function is lower-$\mathcal{C}^2$ (and therefore locally Lipschitz) is given in [39, Lem9].

Another example where errors arise in function and gradient evaluation is when the objective function is provided through a stochastic simulation.

Example 3 (Stochastic simulations) If the objective function is provided through a stochastic simulation, then the errors in the function and subgradient values are understood through probability distribution functions. (This would encompass, for example, the situation where the objective function is provided by an expected value estimated via Monte-Carlo simulation [42].) Errors can be controlled and reduced by running the simulation multiple times and applying the central limit theorem. However, it should be noted that, an error bound of the form used in this paper is not truly accessible in this situation, as there will always be some nonzero probability of the error being surprisingly large.

The minimization of nonsmooth *convex* functions that are given by *exact* information has been successfully approached in several manners. Amongst the most popular are the bundle and proximal-bundle methods [19, Ch.XV]. Indeed, such methods are currently considered the most efficient optimization methods for nonsmooth problems; see, e.g., [29,44,45] for more detailed comments.

From the "primal" point of view, bundle methods can be thought of as replacing the true objective function by a model, constructed through a bundle of information gathering past evaluation points and their respective f, g-values. In particular, proximal-bundle methods, [19, Ch.XV], compute the *proximal point* of the model function to obtain new bundle elements and generate better minimizer estimates. This work is in the direction of adapting one such method to handle both nonconvex objective functions and inexact information.

Not long after works on bundle methods for the convex case were first developed, the problem of (locally) minimizing a nonsmooth *nonconvex* function using *exact* information was considered in [20,36] and more recently in [1,18,23,31,37]. Many 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 [27,28,30,34–36]. Except for [18], all of these methods handle nonconvexity by downshitfing the so-called linearization errors if they are negative. The method of [18] tilts the slopes in addition to downshifting. Our algorithm here is along the lines of [18].

Inexact evaluations in *subgradient methods* had been studied in the nonconvex setting in [48], and in the convex case, for a variety of algorithms, in [8,24,38]. Contrary to earlier work on inexact subgradient methods, both [48] and [38] allow *nonvanishing* 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 [22]. However, the noise in [22] is asymptotically vanishing. The first work where nonvanishing perturbations in bundle methods had been considered appears to be [15];

but 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 [46], and thoroughly studied in [25]. For the latest unified theory of convex inexact bundle methods, see [9]. In this work, we consider behavior of the redistributed bundle method of [18] for *nonconvex* functions that are given by inexact information. To the best of our knowledge, the only other work dealing with inexact information in bundle methods for nonconvex functions is [39]. The method of [39] employs the "downshift" mechanism that modifies linearization errors if they are negative. In addition to downshifting the cutting-planes, our method also tilts its slopes (and of course, there are also some other differences in the algorithms). In [39], the cases where the objective function is either ε-convex ([39, eq. (1.14)]) or lower-$\mathcal{C}^1$ ([43, Def.10.29]) are examined. Unlike our work, which assumes a bounded constraint set, the work of [39] assumes bounded lower level sets. Overall, our convergence results are quite similar to [39] (see Sect. 5.2 for a thorough description of some details of this comparison). The algorithms themselves are quite different, however.

The remainder of this paper is organized as follows. This section continues with outlining some general terminology and notation for our nonconvex setting. Section 2 summarizes the notation used in our algorithm. In Sects. 3 and 4 we formally state our Inexact Proximal Bundle Method and analyze its convergence. Section 6 presents numerical results.

1.1 General notation and assumptions

Throughout this work we assume that, in problem (1), the objective function f is *proper* [43, p. 5], *regular* [43, Def 7.25], and locally Lipschitz with full domain. 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-$\mathcal{C}^2$" function, so proper, regular, and locally Lipschitz [43, 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 $x \in \mathbb{R}^n$.

In general we shall work with the definitions and notation laid out in [43]. The closed ball in $\mathbb{R}^n$ with the center in $x \in \mathbb{R}^n$ and radius $\rho > 0$ is denoted by $B_\rho(x)$. 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 is given by

$$\partial f(x) := \left\{ g \in \mathbb{R}^n : \lim_{x \to \bar{x}} \inf_{x \neq \bar{x}} \frac{f(x) - f(\bar{x}) - \langle g, x - \bar{x} \rangle}{|x - \bar{x}|} \geq 0 \right\}. \tag{2}$$

Alternative equivalent definitions of the subdifferential mapping for regular functions can be found in [43, Chap. 8].

The family of lower-$\mathcal{C}^1$ functions, defined below, was introduced by [49]. It constitutes a broad class of locally Lipschitz functions that contains lower-$\mathcal{C}^2$ functions as a subfamily. Given an open set Ω containing D, combining [6, Thm.2,Cor.3] with [49, Prop.2.4], the following statements are equivalent:

$$\begin{cases} \text{(i)} & f \text{ is lower}-\mathcal{C}^1 \text{ on } \Omega \\ \text{(ii)} & \left.\begin{array}{l} \forall \bar{x} \in \Omega\,, \forall \varepsilon > 0\, \exists \rho > 0\colon \\ \forall x \in B_\rho(\bar{x}) \text{ and } g \in \partial f(x) \end{array}\right\} & f(x+u) \geq f(x) + \langle g, u\rangle - \varepsilon|u| \\ & & \text{whenever } |u| \leq \rho \text{ and } x+u \in B_\rho(\bar{x}) \\ \text{(iii)} & \left.\begin{array}{l} \forall \bar{x} \in \Omega\,, \forall \varepsilon > 0\, \exists \rho > 0\colon \\ \forall y^1\,, y^2 \in B_\rho(\bar{x}) \text{ and } g^1 \in \partial f(y^1)\,, g^2 \in \partial f(y^2) \end{array}\right\} & \left\langle g^1 - g^2, y^1 - y^2\right\rangle \geq -\varepsilon|y^1 - y^2| \\ \text{(iv)} & f \text{ is semismooth ([34]) and regular on } \Omega\,. \end{cases} \tag{3}$$

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 Available information

Defining the concept of inexact information for function values is straigthforward. Given a point x and some error tolerance $\sigma \geq 0$, the statement "$\phi \in \mathbb{R}$ approximates the value $f(x)$ within σ", means that $|\phi - f(x)| \leq \sigma$. By contrast, for subgradient values, the notion of inexact information allows more interpretations. We shall consider the following estimates, which make good sense especially in the nonconvex case. At a point x, an element $g \in \mathbb{R}^n$ approximates within tolerance $\theta \geq 0$ some subgradient of f at x if $g \in \partial f(x) + B_\theta(0)$.

The algorithm herein will hinge around previously generated information. Basic elements include

k	an iteration counter,
J^k	an index set for the information used in iteration k,
$\{x^j\}_{j\in J^k}$	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.

The algorithm works with inexact information. So we have inexact function and subgradient values as follows:

$$\begin{array}{rll} f^j = f(x^j) - \sigma^j & \text{where} & \sigma^j \text{ is an unknown error,} \\ \hat{f}^k = f(\hat{x}^k) - \hat{\sigma}^k & \text{where} & \hat{\sigma}^k \text{ is an unknown error.} \\ g^j \in \partial f(x^j) + B_{\theta^j}(0) & \text{where} & \theta^j \text{ is an unknown error.} \end{array} \tag{4}$$

Note that the sign of errors σ^j is not specified, so that the true function value can be either overestimated or underestimated. Both error terms σ^j and θ^j are assumed bounded:

$$|\sigma^j| \leq \bar{\sigma} \quad \text{and} \quad 0 \leq \theta^j \leq \bar{\theta} \quad \text{for all} \quad j. \tag{5}$$

But the error terms themselves, and their bounds $\bar{\sigma}$, $\bar{\theta}$, are generally unknown.

2.2 Model functions and some basic relations

As usual in bundle methods, we use the available information to define a piecewise-linear model of f. If the function f were convex and the data exact, then given a point x^j, any subgradient $g^j \in \partial f(x^j)$ would generate a linear lower bound for the function: $f(x^j) + \langle g^j, y - x^j \rangle \leq f(y)$ for all y. This knowledge gives the classical cutting-plane model for f:

$$\max_{j \in J^k} \left\{ f(x^j) + \left\langle g^j, y - x^j \right\rangle \right\} = f(\hat{x}^k) + \max_{j \in J^k} \left\{ -e_j^k + \left\langle g^j, y - \hat{x}^k \right\rangle \right\}$$

for some index set $J^k \subseteq \{1, \ldots, k\}$ referring to some previous iterates, where

$$e_j^k = f(\hat{x}^k) - f(x^j) - \left\langle g^j, \hat{x}^k - x^j \right\rangle$$

are the linearization errors (nonnegative in the convex case).

In our setting, we are working with inexact information. Furthermore, we have to deal with possible nonconvexity of f. Following the redistributed proximal approach of [18], we generate a convex piecewise-linear model defined by

$$M^k(\hat{x}^k + d) := \hat{f}^k + \max_{j \in J^k} \left\{ -c_j^k + \left\langle s_j^k, d \right\rangle \right\}. \tag{6}$$

In each affine piece, both the intercept and the slope correspond, respectively, to the linearization error and subgradient of the "locally convexified" function of the form $f(\cdot) + \frac{\eta^k}{2} |\cdot - \hat{x}^k|^2$, for certain convexification parameter η^k adjusted dynamically, along iterations. Similarly to [18], such parameter is taken sufficiently large to make the intercept c_j^k nonnegative (in the nonconvex case, the linearization errors may be negative even if the exact data is used).

Accordingly, each affine piece has a shifted nonnegative intercept

$$0 \leq c_j^k := e_j^k + b_j^k, \quad \text{for} \quad \begin{cases} e_j^k := \hat{f}^k - f^j - \langle g^j, \hat{x}^k - x^j \rangle, \\ b_j^k := \dfrac{\eta^k}{2} \left| x^j - \hat{x}^k \right|^2; \end{cases} \tag{7}$$

and a modified slope,

$$s_j^k := g^j + \eta^k \left(x^j - \hat{x}^k \right), \tag{8}$$

which results from tilting the given approximate subgradient g^j at x^j by means of η^k.

Any choice for the convexification parameter that keeps c_j^k in (7) nonnegative is acceptable. In our proximal redistributed method we take

$$\eta^k \geq \max\left\{\max_{j\in J^k,\, x^j\neq\hat{x}^k} \frac{-2e_j^k}{\left|x^j-\hat{x}^k\right|^2}, 0\right\} + \gamma, \tag{9}$$

for a (small) positive parameter γ, whose role is explained in Remark 1 below.

The term *bundle information* will be used to denote all the data needed to define the model M^k in (6); the relevant objects will be indexed by the set 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 (this is not strictly necessary, but simplifies some issues; and keeping the last best iterate in the bundle makes some general sense anyway).

Notice that taking in (7) the index $J \in J^k$ for which $\hat{x}^k = x^J$ gives $b_J^k = 0$ and $e_J^k = \hat{f}^k - f^J = 0$, so $c_J^k = 0$ and, hence,

$$M^k(\hat{x}^k) = \hat{f}^k + \max_{j\in J^k}\{-c_j^k\} = \hat{f}^k. \tag{10}$$

Each new iterate in the algorithm is given by solving the *proximal point* subproblem for the model M^k. Specifically,

$$x^{k+1} = \hat{x}^k + d^k,$$

for the (uniquely defined) direction

$$\begin{aligned} d^k &:= \arg\min_{\hat{x}^k+d\in D}\left\{M^k(\hat{x}^k+d) + \frac{1}{2t^k}|d|^2\right\} \\ &= \arg\min_{d\in\mathbb{R}^n}\left\{M^k(\hat{x}^k+d) + \mathtt{i}_D(\hat{x}^k+d) + \frac{1}{2t^k}|d|^2\right\}, \end{aligned} \tag{11}$$

where $t^k > 0$ is an inverse proximal-parameter, and the notation $\mathtt{i}_D$ stands for the indicator function of the set D:

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

As a practical matter, D 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., [2]. So, in practice, D could be defined by convex quadratics too. As a matter of the theory presented in the sequel, D can be any convex compact set (subject to the comments in Remark 2 below).

From the optimality conditions of the subproblem above (which is linearly constrained if D is polyhedral; or assuming a constraint qualification [47] if D is more general),

$$0 \in \partial M^k\left(x^{k+1}\right) + \partial\,\mathtt{i}_D\left(x^{k+1}\right) + \frac{1}{t^k}d^k.$$

Since the model (6) is piecewise-linear, this means that 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

$$d^k = -t^k\left(G^k + v^k\right), \quad \text{where} \quad G^k := \sum_{j \in J^k} \alpha_j^k s_j^k, \quad v^k \in \partial \mathtt{i}_D\left(x^{k+1}\right). \tag{12}$$

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

$$A^k\left(\hat{x}^k + d\right) := M^k\left(x^{k+1}\right) + \left\langle G^k, d - d^k\right\rangle. \tag{13}$$

Thus we have,

$$\begin{array}{l} A^k\left(x^{k+1}\right) = M^k\left(x^{k+1}\right), G^k \in \partial M^k\left(x^{k+1}\right), \text{ and} \\ G^k = \nabla A^k\left(\hat{x}^k + d\right) \quad \text{for all} \quad d \in \mathbb{R}^n. \end{array} \tag{14}$$

By the subgradient inequality, it holds that

$$A^k\left(\hat{x}^k + d\right) \leq M^k\left(\hat{x}^k + d\right) \quad \text{for all} \quad d \in \mathbb{R}^n. \tag{15}$$

The *aggregate error* is defined by

$$E^k := M^k\left(\hat{x}^k\right) - M^k\left(x^{k+1}\right) + \left\langle G^k, d^k\right\rangle \geq 0, \tag{16}$$

where the inequality follows from $G^k \in \partial M^k(x^{k+1})$ and $d^k = x^{k+1} - \hat{x}^k$. Using that $\hat{f}^k = M^k(\hat{x}^k)$ (see (10)) and the optimal multipliers from (12), gives the following alternative aggregate error expressions:

$$E^k = \sum_{j \in J^k} \alpha_j^k c_j^k, \tag{17}$$

and

$$E^k = \hat{f}^k - A^k\left(x^{k+1}\right) + \left\langle G^k, d^k\right\rangle.$$

Similarly, for the aggregate linearization it holds that

$$\begin{aligned} A^k(\hat{x}^k + d) &= \hat{f}^k + \sum_{j \in J^k} \alpha_j^k \left(-c_j^k + \left\langle s_j^k, d\right\rangle\right) \\ &= \hat{f}^k - E^k + \langle G^k, d\rangle, \quad d \in \mathbb{R}^n, \end{aligned} \tag{18}$$

where we have used (12).

3 Algorithm statement

After the new iterate is computed, we first check whether it provides sufficient decrease of the objective function as compared to the previous stability center (naturally, both are inexact values in our setting). Specifically, the quality of decrease is measured as a fraction of the quantity

$$\delta^k := \left(\hat{f}^k - M^k(\hat{x}^k)\right) + E^k + t_k\left|G^k + \nu^k\right|^2 = E^k + t_k\left|G^k + \nu^k\right|^2, \tag{19}$$

where E^k is defined in (17) and G^k and ν^k are given by (12); the right-most equality is by (10). Note that since $E^k \geq 0$ by (16), it follows from (19) that

$$\delta^k \geq 0.$$

(Here, we note that our definition of the predicted decrease δ^k differs from [18].) If the descent is sufficient, then the corresponding point is declared the new stability center (a so-called serious iteration). Otherwise, the stability center $\hat{x}^k$ remains unchanged, and the model M^k is refined (a so-called null step).

Our assumptions on defining the next model M^{k+1} are standard:

$$\begin{aligned} M^{k+1}(\hat{x}^k + d) &\geq \hat{f}^{k+1} - c_{k+1}^{k+1} + \left\langle s_{k+1}^{k+1}, d\right\rangle, \\ M^{k+1}(\hat{x}^k + d) &\geq A^k(\hat{x}^k + d). \end{aligned} \tag{20}$$

The conditions in (20) are required to hold on consecutive null steps only; they need not be required after a serious step is performed. The first relation in (20) just means that the newly computed information always enters the bundle. The second condition holds automatically if no information is removed (due to (15)), or if only inactive pieces (corresponding to $\alpha_j^k = 0$) are removed.

We next describe the different steps of the proposed algorithm.

Algorithm 4 (*Nonconvex Proximal Bundle Method with Inexact Information*) A procedure is given, providing for each x a value f approximating $f(x)$ and a vector g approximating some element in $\partial f(x)$, as in (4), (5).

Step 0 (initialization)

Select parameters $m \in (0, 1)$ and $\gamma > 0$ and a stopping tolerance $\texttt{tol} \geq 0$.

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 to $k = 1$. Select an initial inverse prox-parameter $t^1 > 0$. Set $\hat{f}^1 = f^1$ and the initial prox-center $\hat{x}^1 := x^1$.

Step 1 (trial point finding and stopping test)

Given the model M^k defined by (6), compute the direction d^k by solving the subproblem (11). Define the associated G^k and ν^k by (12), E^k by (17), and δ^k by (19).

Set $x^{k+1} = \hat{x}^k + d^k$. If $\delta^k \leq \texttt{tol}$, stop.

Step 2 (descent test)
Compute (f^{k+1}, g^{k+1}), the information at x^{k+1}. If

$$f^{k+1} > \hat{f}^k - m\delta^k, \tag{21}$$

then declare the iteration a null-step and go to Step 3.

Otherwise, declare the iteration a serious-step and set $\hat{x}^{k+1} := x^{k+1}$, $\hat{f}^{k+1} := f^{k+1}$, select $t^{k+1} > 0$, and go to Step 4.

Step 3 (null-step)
Set $\hat{x}^{k+1} := \hat{x}^k$, $\hat{f}^{k+1} := \hat{f}^k$; choose $0 < t^{k+1} \leq t^k$.

Step 4 (bundle update and loop)
Select the new bundle index set J^{k+1}, keeping the active elements. Select η^k as in (9) and update the model M^{k+1} as needed. Increase k by 1 and go to Step 1. □

The use of δ^k as a stationarity measure to stop the algorithm will be clarified by the relations in Lemma 5; see also Theorems 6 and 7.

As mentioned, Algorithm 4 follows the framework and ideas laid out in [18]. However, in order to ensure convergence of the algorithm in the presence of inexact information, some adaptations are made. To begin with, the algorithm now assumes a convex compact constraint set D. As a result, the normal (to the set D) elements v^k are introduced and carried throughout. Next, the computation for the predicted decrease looks somewhat different. However, applying in (19) the relations (12) and (16), we see that

$$\begin{aligned}
\delta^k &= \left(\hat{f}^k - M^k(\hat{x}^k)\right) + E^k + t_k\left|G^k + v^k\right|^2 \\
&= \hat{f}^k - M^k\left(x^{k+1}\right) + \left\langle G^k, d^k\right\rangle + t^k\left|G^k + v^k\right|^2 \\
&= \hat{f}^k - \left(M^k\left(\hat{x}^k + d^k\right) + \left\langle v^k, d^k\right\rangle\right).
\end{aligned}$$

Then the predicted decrease from [18] is recovered if there is no constraint set D (or if $\hat{x} + d^k \in \text{int}(D)$, so that $v^k = 0$).

Another change from [18] is in the computation of η^k, explained below.

Remark 1 (The choice of the convexificaton parameter η^k) As explained in Sect. 2.2, the term

$$\max_{j \in J^k,\, x^j \neq \hat{x}^k} \frac{-2e_j^k}{|x^j - \hat{x}^k|^2}$$

represents the minimal value of η to imply that for all $j \in J^k$ the linearization errors of the "locally convexified" function remain nonnegative:

$$e_j^k + \frac{\eta}{2}\left|x^j - \hat{x}^k\right|^2 \geq 0.$$

Taking the maximum of this term with 0 yields nonnegativity of η^k, and adding the "safeguarding" small positive parameter γ makes η^k strictly larger that the minimal value. This differs from the update in [18] where instead the minimal term was multiplied by a constant $\Gamma > 1$. As illustrated by Fig. 1 in our numerical experiments, the update (9) works somewhat better than the original version of [18]. The reason appears to be that it deals better with situations where the minimal term in question is null or too small. □

Finally, it is worth remarking that, contrary to many nonconvex bundle methods endowed with a linesearch, e.g., [21,23,32,34,37], our method does not employ a linesearch sub-routine. In Sect. 5 we give some indications as to why a linesearch is not required in our approach.

Remark 2 (Uniformly bounded number of active indices in subproblems) In our analysis below, we shall make the following assumption: "The number of active indices, i.e., of $j \in J^k$ such that $\alpha_j^k > 0$, is uniformly bounded in k". As a practical matter, this can be readily achieved if D is polyhedral (the typical case). This is because most (if not all) active-set QP solvers choose linearly independent bases, i.e., work with "minimal" representations. In the expression of G^k in (12), this means that QP solver gives a solution with no more than $n+1$ positive simplicial multipliers (such a solution always exists by the Carathéodory Theorem). A similar assumption/property for a QP solver had been used for a different QP-based method in [13, Sec.5], and specifically for a bundle procedure in [7].

That said, it is should be noted that if a non-active-set method (for example, an interior point method) is used, then this assumption need not hold.

We also use below the assumption that $\{\eta^k\}$ is bounded. Boundedness of $\{\eta^k\}$ has been established in [18] for the lower-$\mathcal{C}^2$ case when the function information is exact. However, in our setting it is theoretically possible that inexactness results in an unbounded η^k, even if the objective function is convex. The experiments in Sect. 6 show, however, that behavior of the sequence $\{\eta^k\}$ is adequate under various kinds of perturbations, and the overall performance of the inexact algorithm is satisfactory indeed.

4 Convergence properties

We proceed with the convergence analysis of Algorithm 4, considering two cases: either there is an infinite sequence of serious/descent iterations, or from some index on the stability center $\hat{x}^k$ remains fixed and all the subsequent iterations are of the null type.

4.1 General asymptotic relations

We start with some relations that are relevant for all the cases.

Lemma 5 *Suppose the cardinality of the set $\{j \in J^k \mid \alpha_j^k > 0\}$ is uniformly bounded in k (recall Remark 2).*

If $E^k \to 0$ as $k \to \infty$, then

(i) $\sum_{j \in J^k} \alpha_j^k |x^j - \hat{x}^k| \to 0$ *as* $k \to \infty$.

If, in addition, for some subset $K \subset \{1, 2, \ldots\}$,

$$\hat{x}^k \to \bar{x}, \ G^k \to \bar{G} \ \text{ as } K \ni k \to \infty, \ \text{with } \{\eta^k \mid k \in K\} \text{ bounded},$$

then we also have

(ii) $\bar{G} \in \partial f(\bar{x}) + B_{\bar{\theta}}(0)$.

If, in addition, $G^k + v^k \to 0$ *as* $K \ni k \to \infty$, *then*

(iii) $\bar{x}$ *satisfies the following approximate stationarity condition:*

$$0 \in \left(\partial f(\bar{x}) + \partial i_D(\bar{x})\right) + B_{\bar{\theta}}(0). \tag{22}$$

Finally, if in addition, f *is lower-*$\mathcal{C}^1$, *then*

(iv) for each $\varepsilon > 0$ *there exists* $\rho > 0$ *such that*

$$f(y) \geq f(\bar{x}) - (\bar{\theta} + \varepsilon)|y - \bar{x}| - 2\bar{\sigma}, \quad \text{for all } y \in D \cap B_\rho(\bar{x}). \tag{23}$$

Proof Recall that the first term in the right-hand side of (9) is the minimal value of $\eta \geq 0$ to imply that

$$e_j^k + \frac{\eta}{2}\left|x^j - \hat{x}^k\right|^2 \geq 0$$

for all $i \in J^k$. It is then easily seen that, for such η and for $\eta^k \geq \eta + \gamma$, we have that

$$c_j^k = e_j^k + \frac{\eta^k}{2}\left|x^j - \hat{x}^k\right|^2 \geq \frac{\gamma}{2}\left|x^j - \hat{x}^k\right|^2.$$

Taking into account that α_j^k and c_j^k are nonnegative, if $E^k \to 0$ then it follows from (17) that $\alpha_j^k c_j^k \to 0$ for all $j \in J^k$. Hence,

$$\alpha_j^k c_j^k \geq \left(\alpha_j^k\right)^2 c_j^k \geq \frac{\gamma}{2}\left(\alpha_j^k |x^j - \hat{x}^k|\right)^2 \to 0.$$

Thus, $\alpha_j^k |x^j - \hat{x}^k| \to 0$ for all $j \in J^k$. As, by the assumption, the sum in the item (i) is over a finite set of indices and each element in the sum tends to zero, the assertion (i) follows.

For each j, let p^j be the orthogonal projection of g^j onto the (convex, closed) set $\partial f(x^j)$. It holds that $|g^j - p^j| \leq \theta^j \leq \bar{\theta}$. By (12) and (8), we have that

$$G^k = \sum_{j\in J^k} \alpha_j^k g^j + \eta^k \sum_{j\in J^k} \alpha_j^k \big(x^j - \hat{x}^k\big)$$
$$= \sum_{j\in J^k} \alpha_j^k p^j + \sum_{j\in J^k} \alpha_j^k \big(g^j - p^j\big) + \eta^k \sum_{j\in J^k} \alpha_j^k \big(x^j - \hat{x}^k\big). \tag{24}$$

As the number of active indices is uniformly bounded in k, by re-numbering the indices and filling unused indices with $\alpha_j^k = 0$, we can consider that J^k is some fixed index set (say, $\{1, \ldots, N\}$). Let J be the set of all $j \in J^k$ such that $\liminf \alpha_j^k > 0$. Then item (i) imples that $|x^j - \hat{x}^k| \to 0$. Thus, $|x^j - \bar{x}| \le |x^j - \hat{x}^k| + |\hat{x}^k - \bar{x}| \to 0$. As $p^j \in \partial f(x^j)$ and $x^j \to \bar{x}$ for $j \in J$, and $\{\alpha_j^k\} \to 0$ for $j \notin J$, passing onto a further subsequence in the set K, if necessary, outer semicontinuity of the Clarke subdifferential [43, Thm 6.6] implies that

$$\lim_{k\to\infty} \sum_{j\in J^k} \alpha_j^k p^j \in \partial f(\bar{x}).$$

As the second term in (24) is clearly in $B_{\bar{\theta}}(0)$, while the last term tends to zero by item (i), this shows the assertion (ii).

Item (iii) follows from noting that $(G^k + v^k) \to 0$ as $K \ni k \to \infty$ implies that $\{v^k\} \to -\bar{G}$. As $v^k \in \partial \mathtt{i}_D(\hat{x}^k)$ for each k, we conclude that $-\bar{G} \in \partial \mathtt{i}_D(\bar{x})$ (by [43, Thm6.6]). Adding the latter inclusion and result (ii) gives (22).

We finally consider item (iv). Fix any $\varepsilon > 0$. Let $\rho > 0$ be such that (3.ii) holds for $\bar{x}$. Let $y \in D \cap B_\rho(\bar{x})$ be arbitrary but fixed. Again, we can consider that J^k is a fixed index set. Let J be the set of $j \in J^k$ for which $|x^j - \hat{x}^k| \to 0$. In particular, it then holds that $x^j \in B_\rho(\bar{x})$. By item (i), we have that $\{\alpha_j^k\} \to 0$ for $j \notin J$.

Using (3) together with (4), for $j \in J$ we obtain that

$$f(y) \ge f^j + \left\langle g^j, y - x^j \right\rangle + \sigma^j + \left\langle p^j - g^j, y - x^j \right\rangle - \varepsilon |y - x^j|$$
$$\ge f^j + \left\langle g^j, y - x^j \right\rangle + \sigma^j - (\theta^j + \varepsilon)|y - x^j|.$$

By (7) and the linearization error definition,

$$f^j + \left\langle g^j, -x^j \right\rangle = \hat{f}^k - \left\langle g^j, \hat{x}^k \right\rangle + b_j^k - c_j^k.$$

As a result, it holds that

$$f(y) \ge \hat{f}^k - c_j^k + b_j^k + \left\langle g^j, y - \hat{x}^k \right\rangle + \sigma^j - (\theta^j + \varepsilon)|y - x^j|.$$

Since $b_j^k \ge 0$ and $g^j = s_j^k - \eta^k(x^j - \hat{x}^k)$, we obtain that

$$f(y) \ge f(\hat{x}^k) - c_j^k + \left\langle s_j^k, y - \hat{x}^k \right\rangle - \eta^k \left\langle x^j - \hat{x}^k, y - \hat{x}^k \right\rangle + \sigma^j + \hat{\sigma}^k$$
$$- (\theta^j + \varepsilon)|y - x^j|.$$

Taking the convex combination in the latter relation using the simplicial multipliers in (12), and using (17), gives

$$\begin{aligned} f(y)\sum_{j\in J}\alpha_j^k &\geq \sum_{j\in J}\alpha_j^k\left(f(\hat{x}^k)-c_j^k+\left\langle s_j^k, y-\hat{x}^k\right\rangle\right)-\eta^k\left\langle\sum_{j\in J}\alpha_j^k(x^j-\hat{x}^k), y-\hat{x}^k\right\rangle \\ &\quad+\sum_{j\in J}\alpha_j^k\left(\sigma^j+\hat{\sigma}^k\right)-(\theta^j+\varepsilon)\sum_{j\in J}\alpha_j^k|y-x^j| \\ &\geq f(\hat{x}^k)\sum_{j\in J}\alpha_j^k-E^k+\left\langle G^k, y-\hat{x}^k\right\rangle-\sum_{j\notin J}\alpha_j^k\left\langle s_j^k, y-\hat{x}^k\right\rangle \\ &\quad-\eta^k\left\langle\sum_{j\in J}\alpha_j^k\left(x^j-\hat{x}^k\right), y-\hat{x}^k\right\rangle \\ &\quad-2\bar{\sigma}-(\bar{\theta}+\varepsilon)\sum_{j\in J}\alpha_j^k\left(\left|y-\hat{x}^k\right|+\left|x^j-\hat{x}^k\right|\right). \end{aligned} \tag{25}$$

Passing onto the limit in (25) as $K\ni k\to\infty$, using item (i) and also that $\{\alpha_j^k\}\to 0$ for $j\notin J$ (so that, in particular, $\sum_{j\in J}\alpha_j^k\to 1$), we obtain that

$$f(y)\geq f(\bar{x})+\left\langle\bar{G}, y-\bar{x}\right\rangle-2\bar{\sigma}-(\bar{\theta}+\varepsilon)|y-\bar{x}|. \tag{26}$$

As already seen, $(G^k+v^k)\to 0$ implies that $-\bar{G}\in\partial\mathtt{i}_D(\bar{x})$, so that $\langle-\bar{G}, y-\bar{x}\rangle\leq 0$ for all $y\in D$. Adding this inequality to (26) gives the assertion (23). □

If f is convex and $D=\mathbb{R}^n$, then condition (23) can be seen to be equivalent to $0\in\partial_{2\bar{\sigma}}f(\bar{x})+B_{\bar{\theta}}(0)$, where $\partial_{2\sigma}f(\bar{x})$ denotes the usual 2σ-subdifferential of f at $\bar{x}$. This supports that an approximate optimality condition of this order ("linear" in the errors levels) is what is reasonable to strive to achieve in the setting of perturbed data. That said, we note that for the convex case (and for so-called "lower models" [9]), our result is weaker than what can be obtained by other means (basically, in the convex case a $\bar{\sigma}$-approximate solution can be achieved). This is quite natural, however, as the convex case analysis takes advantage of the corresponding tools (like the subgradient inequality), which are not available in our more general setting.

4.2 Null and serious steps

Dependent on the assumptions about t_k, we shall prove that the approximate optimality condition holds for: some accumulation point $\bar{x}$ of $\{\hat{x}^k\}$; all accumulation points of $\{\hat{x}^k\}$; or for the last serious iterate $\hat{x}^k=\bar{x}$.

Consider first the case of the infinite number of serious steps.

Theorem 6 (Infinitely many serious iterates) *Let the algorithm generate an infinite number of serious steps. Then $\delta^k\to 0$ as $k\to\infty$.*

Let the sequence $\{\eta^k\}$ be bounded.

(*i*) *If* $\sum_{k=1}^{\infty} t_k = +\infty$, *then as* $k \to \infty$ *we have* $E^k \to 0$, *and there exist* $K \subset \{1, 2, \ldots\}$ *and* $\bar{x}$, $\bar{G}$ *such that* $\hat{x}^k \to \bar{x}$, $G^k \to \bar{G}$, *and* $G^k + v^k \to 0$ *as* $K \ni k \to \infty$.
In particular, if the cardinality of the set $\{j \in J^k \mid \alpha_j^k > 0\}$ *is uniformly bounded in* k *(recall Remark 2), then the conclusions of Lemma 5 hold.*

(*ii*) *If* $\liminf_{k\to\infty} t^k > 0$, *then these assertions hold for all accumulation points* $\bar{x}$ *of* $\{\hat{x}^k\}$.

Proof At each serious step k, the opposite of (21) holds. Thus, we have that

$$\hat{f}^{k+1} \le \hat{f}^k - m\delta^k, \tag{27}$$

where $\delta^k \ge 0$. It follows that the sequence $\{\hat{f}^k\}$ is nonincreasing.

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

Using (27), we obtain that

$$0 \le m \sum_{k=1}^{l} \delta^k \le \sum_{k=1}^{l-1} \left(\hat{f}^k - \hat{f}^{k+1}\right),$$

so that, letting $l \to \infty$,

$$0 \le m \sum_{k=1}^{\infty} \delta^k \le \hat{f}^1 - \lim_{k\to\infty} \hat{f}^k.$$

As a result,

$$\sum_{k=1}^{\infty} \delta^k = \sum_{k=1}^{\infty} \left(E^k + t_k \left|G^k + v^k\right|^2\right) < +\infty. \tag{28}$$

Hence, $\delta^k \to 0$ as $k \to \infty$. As all the quantities above are nonnegative, it also holds that

$$E^k \to 0 \quad \text{and} \quad t_k |G^k + v^k|^2 \to 0 \quad \text{as} \quad k \to \infty. \tag{29}$$

If $\sum_{k=1}^{\infty} t_k = +\infty$, but $|G^k + v^k| \ge \beta$ for some $\beta > 0$ and all k, then (28) results in a contradiction. The fact that no such β exists, means precisely that there exists an index set $K \subset \{1, 2, \ldots\}$ such that

$$G^k + v^k \to 0, \quad K \ni k \to \infty. \tag{30}$$

Passing onto a further subsequence, if necessary, we can assume that $\{x^k\} \to \bar{x}$ and $G^k \to \bar{G}$ as $K \ni k \to \infty$. Item (i) is now proven.

If $\liminf_{k\to\infty} t_k > 0$, then the second relation in (29) readily implies (30) for $K = \{1, 2, \ldots\}$, and thus the same assertions can be seen to hold for all accumulation points of $\{\hat{x}^k\}$. □

In the remaining case, a finite number of serious steps occurs. That is, after a finite number of iterations the algorithmic center is no more changed: from some $\bar{k}$ on, the center is $\hat{x}^k = \hat{x}$ for all $k > \bar{k}$, and only null steps follow. The proof makes use of a simple, yet crucial, relation that we show next. Specifically, by the intercept and slope definitions in (7) and (8), we see that

$$\begin{aligned}
&-c_{k+1}^{k+1} + \left\langle s_{k+1}^{k+1}, x^{k+1} - \hat{x}^k \right\rangle \\
&\quad = -e_{k+1}^{k+1} - b_{k+1}^{k+1} + \left\langle g^{k+1} + \eta^{k+1}\left(x^{k+1} - \hat{x}^k\right), x^{k+1} - \hat{x}^k \right\rangle \\
&\quad = -\left(\hat{f}^k - f^{k+1} - \left\langle g^{k+1}, \hat{x}^k - x^{k+1} \right\rangle\right) - \frac{\eta^{k+1}}{2}\left|x^{k+1} - \hat{x}^k\right|^2 \\
&\qquad + \left\langle g^{k+1}, x^{k+1} - \hat{x}^k \right\rangle + \eta^{k+1}\left|x^{k+1} - \hat{x}^k\right|^2 \\
&\quad = f^{k+1} - \hat{f}^k + \frac{\eta^{k+1}}{2}\left|x^{k+1} - \hat{x}^k\right|^2.
\end{aligned}$$

As a result, whenever x^{k+1} is declared a null step, (21) implies that

$$-c_{k+1}^{k+1} + \left\langle s_{k+1}^{k+1}, x^{k+1} - \hat{x}^k \right\rangle \geq -m\delta^k. \tag{31}$$

We also note that this crucial relation eliminates the need of performing linesearch at null steps; see Sect. 5.1 below.

Theorem 7 (Finite serious steps followed by infinitely many null steps)

Let a finite number of serious iterates be followed by infinite null steps. Let the sequence $\{\eta^k\}$ *be bounded and* $\liminf_{k\to\infty} t^k > 0$.

Then $\{x^k\} \to \hat{x}$, $\delta^k \to 0$, $E^k \to 0$, $G^k + v^k \to 0$, *and there exist* $K \subset \{1, 2, \ldots\}$ *and* $\bar{G}$ *such that* $G^k \to \bar{G}$ *as* $K \ni k \to \infty$.

In particular, if the cardinality of the set $\{j \in J^k \mid \alpha_j^k > 0\}$ *is uniformly bounded in k (recall Remark 2), the conclusions of Lemma 5 hold for* $\bar{x} = \hat{x}$.

Proof Let k be large enough, so that $k \geq \bar{k}$ and $\hat{x}^k = \hat{x}$, $\hat{f}^k = \hat{f}$ are fixed.

Define the optimal value of the subproblem (11) by

$$\psi^k := M^k\left(x^{k+1}\right) + \frac{1}{2t^k}\left|d^k\right|^2. \tag{32}$$

We first show that the sequence $\{\psi^k\}$ is bounded above. Recall that, by (13),

$$A^k(\hat{x}) = M^k\left(x^{k+1}\right) - \left\langle G^k, d^k \right\rangle.$$

We then obtain that

$$\begin{aligned}\psi^k + \frac{1}{2t^k}\left|d^k\right|^2 &= A^k(\hat{x}) + \left\langle G^k, d^k\right\rangle + \frac{1}{t^k}\left|d^k\right|^2\\ &= A^k(\hat{x}) - \left\langle v^k, d^k\right\rangle\\ &\leq A^k(\hat{x})\\ &\leq M^k(\hat{x})\\ &= \hat{f},\end{aligned}$$

where the second equality follows from $G^k + v^k = -d^k/t^k$, the first inequality is by $v^k \in \partial \mathtt{i}_D(x^{k+1})$ and $d^k = x^{k+1} - \hat{x}$, the second inequality is by (15), and the last is by (10). In particular, $\psi^k \leq \hat{f}$, so the sequence $\{\psi^k\}$ is bounded above.

We next show that $\{\psi^k\}$ is increasing. To that end, we obtain that

$$\begin{aligned}\psi^{k+1} &= M^{k+1}\left(x^{k+2}\right) + \frac{1}{2t^{k+1}}\left|d^{k+1}\right|^2\\ &\geq A^k\left(x^{k+2}\right) + \frac{1}{2t^k}\left|d^{k+1}\right|^2\\ &= M^k\left(x^{k+1}\right) + \left\langle G^k, x^{k+2} - x^{k+1}\right\rangle + \frac{1}{2t^k}\left|d^{k+1}\right|^2\\ &= \psi^k - \frac{1}{2t^k}\left|d^k\right|^2 - \left\langle v^k, x^{k+2} - x^{k+1}\right\rangle - \frac{1}{t^k}\left\langle d^k, d^{k+1} - d^k\right\rangle + \frac{1}{2t^k}\left|d^{k+1}\right|^2\\ &\geq \psi^k + \frac{1}{2t^k}\left|d^{k+1} - d^k\right|^2,\end{aligned}$$

where the first inequality is by the second assumption in (20) and the fact that $t^{k+1} \leq t^k$, the second equality is by (13), the third equality is by (12) and (32), and the last is by $v^k \in \partial \mathtt{i}_D(x^{k+1})$.

As the sequence $\{\psi^k\}$ is bounded above and increasing, it converges. Consequently, taking also into account that $1/t^k \geq 1/t^{\bar{k}}$, it follows that

$$\left|d^{k+1} - d^k\right| \to 0, \quad k \to \infty. \tag{33}$$

Next, by the definition (19) of δ^k and the characterization (16) of E^k, we have that

$$\begin{aligned}\hat{f} &= \delta^k + M^k\left(\hat{x}\right) - E^k - t^k\left|G^k + v^k\right|^2\\ &= \delta^k + M^k\left(x^{k+1}\right) - \left\langle G^k, d^k\right\rangle - t^k\left|G^k + v^k\right|^2\\ &= \delta^k + M^k\left(\hat{x} + d^k\right) + \left\langle v^k, d^k\right\rangle\\ &\geq \delta^k + M^k\left(\hat{x} + d^k\right),\end{aligned}$$

where the inequality is by $v^k \in \partial \mathtt{i}_D(x^{k+1})$. Therefore,

$$\delta^{k+1} \leq \hat{f} - M^{k+1}\left(\hat{x} + d^{k+1}\right). \tag{34}$$

By the first inequality in the assumption (20) on the model, written for $d = d^{k+1}$,

$$-\hat{f}^{k+1} + c_{k+1}^{k+1} - \left\langle s_{k+1}^{k+1}, d^{k+1} \right\rangle \geq -M^{k+1}(\hat{x} + d^{k+1}).$$

As $\hat{f}^{k+1} = \hat{f}$, adding condition (31) to the inequality above, we obtain that

$$m\delta^k + \left\langle s_{k+1}^{k+1}, d^k - d^{k+1} \right\rangle \geq \hat{f} - M^{k+1}(\hat{x} + d^{k+1}).$$

Combining this relation with (34) yields

$$0 \leq \delta^{k+1} \leq m\delta^k + \left\langle s_{k+1}^{k+1}, d^k - d^{k+1} \right\rangle. \tag{35}$$

Since $m \in (0, 1)$ and $\left\langle s_{k+1}^{k+1}, d^k - d^{k+1} \right\rangle \to 0$ as $k \to \infty$ (recall that $\{d^k - d^{k+1}\} \to 0$ by (33) and $\{\eta^k\}$ is bounded), using [41, Lemma 3, p. 45] it follows from (35) that

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

Since $\delta^k = E^k + t_k|G^k + v^k|^2$, and $\liminf_{k\to\infty} t^k > 0$, we have $\lim_{k\to\infty} E^k = 0$ and $\lim_{k\to\infty} |G^k + v^k| = 0$. Also $\lim_{k\to\infty} d^k = 0$, so that $\lim_{k\to\infty} x^k = \hat{x}$. Passing onto a subsequence if necessary, we may also conclude that G^k converges to some $\bar{G}$. Finally, as $\hat{x}^k = \bar{x}$ for all k, we clearly have all of the requirements in Lemma 5 fulfilled. The conclusions follow. □

5 Putting the algorithm in perspective

We next comment on how our approach relates to other methods in the nonsmooth nonconvex literature. As linesearch is common in methods that tackle nonconvex problems, we first explain why our method does not need such a procedure. After that, we describe some differences with the nonconvex bundle method of [39], which also deals with inexact information on both the function and subgradient values.

5.1 Lack of linesearch

In nonsmooth nonconvex optimization methods, linesearch is a subalgorithm that usually uses three parameters $m, m_S, m_N \in (0, 1)$, and is invoked at each iteration k. Let its inner iterations be labeled by $\ell = 1, 2, \ldots$.

By using a linesearch, instead of just taking $\hat{x}^k + d^k$ as the iterate for which the f, g information is computed, a trial stepsize $\tau_\ell > 0$ is chosen, to define the trial point $y^\ell := \hat{x}^k + \tau_\ell d^k$, for which the values f_{y^ℓ}, g_{y^ℓ} are computed.

Analogously to our algorithm, the inner linesearch iterations define trial intercepts and slopes

$$c_\ell^k := \hat{f}^k - f_{y^\ell} - \langle g_{y^\ell}, \hat{x}^k - y^\ell \rangle + \frac{\eta^k}{2}|y^\ell - \hat{x}^k|^2, \quad s_\ell^k := g_{y^\ell} + \eta^k(y^\ell - \hat{x}^k).$$

Then, before incorporating the corresponding affine piece in the next model function, the trial point is classified as follows:

- y^ℓ is declared a serious iterate and the linesearch ends when the opposite of (21) holds (written with the function estimate f_{y^ℓ} instead of f^{k+1}), *and* if the step is "not too short". The latter, in view of (7), means that

$$\text{either} \quad \tau_\ell \geq 1 \quad \text{or} \quad c_\ell^k \geq m_S \delta^k. \tag{36}$$

  The alternative above (i.e., the first condition in (36)) was introduced in [21] to prevent insignificant descent.
- y^ℓ is declared a null iterate and the linesearch ends when there is no sufficient decrease ((21) holds for f_{y^ℓ}) *and*

$$-c_\ell^k + \left\langle s_\ell^k, y^\ell - \hat{x}^k \right\rangle \geq -m_N \delta^k \tag{37}$$

  holds. The latter condition can be interpreted as a nonsmooth extension of the Wolfe condition, see also [36].
- If y^ℓ could not be declared serious or null, the inner iteration continues: the counter ℓ is increased by 1, a new stepsize $\tau^{\ell+1}$ is chosen, and the process is repeated.

For the inner loop with the linesearch to be well-defined, it must (of course) have finite termination. When the information is exact, this can be shown taking $0 < m + m_S < m_N < 1$ when f is upper semidifferentiable [4], a weaker property than semismoothness. With inexact information, as in (4), it is not clear that linesearch terminates finitely (unless, perhaps, the information becomes asymptotically exact).

In our proximal redistributed method, there is no need for linesearch, because one of the two situations above (i.e., satisfaction of either (36) or (37)) always holds for $\tau_\ell = 1$ and $\ell = 1$. To see this, take $m_N = m$ and m_S arbitrary and recall the descent test in Step 2 of the algorithm. If a serious step is declared, i.e., the opposite of (21) holds, then (36) is obviously automatic, as the method always employs $\tau_1 = 1$. If, instead, a null step is declared, (21) holds and, again, no linesearch is necessary, because (31) is just (37), written with $m_N = m$, $y^1 = \hat{x}^k + d^k = x^{k+1}$, $c_1^k = c_{k+1}^{k+1}$, and $s_1^k = s_{k+1}^{k+1}$.

5.2 Relation with Noll's proximity control algorithm

The *proximity control algorithm* of [39] uses certain *second-order model*, denoted by Φ_k, which adds to the cutting-plane model M^k a quadratic term of the form $\frac{1}{2}\langle d, Q(\hat{x}^k)d\rangle$ for a matrix varying with $\hat{x}^k$. Here, without impairing convergence properties of [39], we take the zero matrix, so that

$$\Phi_k(\hat{x}^k + d, \hat{x}^k) \text{ in [39] corresponds to } M^k(d) \text{ in (6)},$$

and, in the parlance of [39], the first and second order models coincide. We emphasize the approach in [39] is more general, as the matrices $Q(\cdot)$ can be indefinite, as long as $Q(\hat{x}^k) + \frac{1}{t_k} I$ remains positive definite. We mention in passing that the proximity control parameter τ_k in [39] corresponds to $1/t_k$ in our method.

The considered problem is unconstrained, but in Sect. 1.5 of the work all null iterates are assumed to remain in some compact set (this is the ball $B(0, M)$ in (1.13) in [39]).

The cutting-plane model in the proximity control method ensures positivity of the intercept c_j^k in (6) by downshifting only, without tilting of the gradients:

$$\text{In [39], the model (6) takes } \begin{cases} c_j^k := e_j^k + \max\{-e_j^k, 0\} + \gamma |x^j - \hat{x}^k|^2 \\ s_j^k := g^j. \end{cases}$$

Downshifting preserves all the important properties in Lemma 5 (which depend on having $c_j^k \geq \gamma |x^j - \hat{x}^k|^2$). But without tilting the slopes, the relation (31) is no longer valid at null steps. To address this issue, the proximity control method distinguishes two cases to update the parameter t_k when (27) does not hold. Instead of introducing a linesearch, as in most of nonconvex bundle algorithms, changing t_k results in a *curve search*.

More precisely, in [39], an iterate is declared a null step and $t_{k+1} = t_k$ when, for parameters $0 < m < \tilde{m} < 1$,

$$f^{k+1} > \hat{f}^k - m\delta^k \quad \text{and} \quad f^{k+1} + \left\langle g^{k+1}, \hat{x}^k - x^{k+1} \right\rangle \leq \hat{f}^k - \tilde{m}\delta^k,$$

or in other words, (21) and $e_{k+1}^k \geq \tilde{m}\delta^k$ hold. Otherwise, the stepsize is deemed "too bad" and it is updated by $t_{k+1} = t_k/2$. Combining both conditions above for null steps gives the inequality

$$\left\langle g^{k+1}, \hat{x}^k - x^{k+1} \right\rangle \leq \hat{f}^k - f^{k+1} - \tilde{m}\delta^k \leq (m - \tilde{m})\delta^k < 0,$$

because $m < \tilde{m}$. Since in addition the downshifting procedure ensures that $c_{k+1}^k \geq e_{k+1}^k$ always, at null steps the inequality in (31) is satisfied with m replaced by $\tilde{m}$:

$$-c_{k+1}^k + \left\langle s_{k+1}^k, x^{k+1} - \hat{x}^k \right\rangle < -c_{k+1}^k < -e_{k+1}^k \leq -\tilde{m}\delta^k.$$

Since the parameter t_k remains fixed for the subsequence of infinite null steps, the proximity control update of this parameter satisfies the conditions in Theorem 7: $t_{k+1} \geq t_k$ with $\liminf t_k > 0$. In this sense, Lemma 7 in [39] corresponds to our Theorem 7 and reaches the same conclusion on approximate stationarity (22), involving only the gradient errors $\bar{\theta}$. The case of an infinite tail of "too bad" steps (which cannot happen in our approach), drives $t_k \to 0$ and is much more involved:

(i) When $\bar{\sigma} = 0$ (no error in the function evaluation), [39, Lemma3] shows that for lower-$\mathcal{C}^1$ functions, whenever (21) holds with $t_k \to 0$, the last center is approximately stationary. Specifically, changing the ball $B_{\bar{\theta}}(0)$ in (22) to the larger ball

$$B_{\Theta}(0) \quad \text{where} \quad \Theta = \bar{\theta} + \frac{\bar{\theta} + \varepsilon}{\tilde{m} - m}$$

and ε is given in (3).

(ii) When there is noise in the function too, [39, Lemma6] gives a similar result for lower-$\mathcal{C}^1$ functions, under the following additional assumption on the evaluation errors:

$$[39, \textbf{ axiom(1.42)}] : \ \exists \varepsilon'' > 0 \text{ and } \Delta^k \to 0^+ \text{ such that}$$
$$\hat{\sigma}^k \leq \sigma^{k+1} + (\varepsilon'' + \Delta_k)|\hat{x}^k - x^{k+1}|.$$

This condition imposes some kind of "upper-semicontinuity of the noise" at the centers. Under this assumption, when there are infinitely many "too bad" steps, the last center is approximately stationary in the ball $B_{\Theta'}(0)$, where $\Theta' = \bar{\theta} + \frac{\bar{\theta}+\varepsilon+\varepsilon''}{\tilde{m}-m}$.

The agressive management of t_k, halving the parameter when steps are "too bad", has also an impact on the convergence analysis for the serious step sequence. When t_k remains bounded away from zero, part ii in [39, Theorems 1 and 2] corresponds to our result in Theorem 6(ii), proving stationarity on a ball depending only on the gradient error bound, $\bar{\theta}$.

As before, the analysis becomes more involved when $t_k \to 0$ (parts (iii) to (ix) in the theorems). Once again, but now for the accumulation points of the serious step sequence, axiom (1.42) yields stationarity on the ball above. This result is not in contradiction with our statement in Theorem 6(i), as halving the parameter t_k results in a (convergent) geometric series with ratio 1/2, which does not satisfy our divergence assumption.

To finish this discussion, we mention the following useful feature of [39]. Section 1.9 therein describes an H_∞ control problem for which the axiom (1.42) on "upper-semicontinuity of the noise" can be effectively ensured in practice.

6 Numerical illustration

In this section we first check the behaviour of Algorithm 4 when the information is exact, by comparing it with the exact nonconvex bundle method of [18]. We also numerically test Algorithm 4 for various kinds of inexactness. While there is no intention to make strong general claims, at least on the given test examples, the approach appears to be satisfactory. Finally, we explore how the convexification parameter behaves in the computational tests. Indeed, Theorems 6 and 7 assume that the parameter η^k remains bounded. In the case of exact information (i.e., $\bar{\sigma} = \bar{\theta} = 0$), under reasonable conditions, [18, Lem. 3] shows boundedness of the convexification parameter sequence. However, for inexact information showing a similar result would be difficult (if not

impossible) without imposing additional assumptions on the behavior of the errors (perhaps of the nature of axiom (1.42) in [39]). This difficulty is illustrated by the following simple example. Consider a constant function $f(x) = 0$ and 3 arbitrarily close points: x^0, $\hat{x} = x^1$, and x^2. Suppose that the function error at x^0 and x^2 both shift function values down slightly, but the function error at x^1 is zero. As the linearization errors indicate that f is an arbitrarily concave quadratic function, the update (9) would force η^k to ∞. Nevertheless, our numerical experience in Sect. 6.4 indicates that one might expect not to encounter such pathological/artificial situations in computation.

6.1 Test functions and types of inexactness

Algorithm 4 was implemented in MATLAB, version 8.1.0.604 (R2013a). Default values for the parameters were set as follows: $m = 0.05$, $\gamma = 2$, and $t^1 = 0.1$. To select η^k (in step 5) we use (9) with equality, i.e.,

$$\eta^k = \max\left\{\max_{j\in J^k,\, x^j\neq\hat{x}^k} \frac{-2e_j^k}{|x^j-\hat{x}^k|^2}, 0\right\} + \gamma.$$

(We also considered $\eta^k = \max\left\{\max_{j\in J^k,\, x^j\neq\hat{x}^k} \frac{-2e_j^k}{|x^j-\hat{x}^k|^2}, 0, \eta^{k-1}\right\} + \gamma$, but the formula above provided slightly better results.) No parameter tuning was performed for any of these values here. Although the values for m and t^1 correspond to the parameters tuned in [18], they are not necessarily optimal in an inexact setting.

In Step 4 of Algorithm 4 the bundle of information keeps only active elements in J^{k+1}. In addition to the stopping test in Step 1, there are two emergency exits: when the iteration count passes $\max(300, 250n)$, and when the QP solver computing the direction d^k in (11) fails.

Like [18], we use the Ferrier polynomials as a collection of nonconvex test problems (see [11], [12]). The Ferrier polynomials are constructed as follows. For each $i = 1, 2, \ldots, n$, we define

$$\begin{aligned} h_i : \mathbb{R}^n &\mapsto \mathbb{R}, \\ x &\mapsto (ix_i^2 - 2x_i) + \textstyle\sum_{j=1}^n x_j. \end{aligned}$$

Using the functions h_i, we define

$$\begin{aligned} f_1(x) &:= \sum_{i=1}^n |h_i(x)|, \\ f_2(x) &:= \sum_{i=1}^n (h_i(x))^2, \\ f_3(x) &:= \max_{i\in\{1,2,\ldots,n\}} |h_i(x)|, \end{aligned}$$

$$f_4(x) := \sum_{i=1}^{n} |h_i(x)| + \frac{1}{2}|x|^2,$$
$$f_5(x) := \sum_{i=1}^{n} |h_i(x)| + \frac{1}{2}|x|.$$

These functions have 0 as a global minimizer, are known to be nonconvex, nonsmooth (except for f_2), lower-$\mathcal{C}^2$, and generally challenging to minimize [18]. As our closed compact feasible set, we use $D = B_{10}(0)$. We consider 75 test problems

$$\min_{x\in D} f_k(x) \quad \text{for} \quad \begin{array}{l} k \in \{1, 2, 3, 4, 5\} \\ n \in \{2, 3, 4, \ldots, 16\}. \end{array}$$

We set $x^1 = [1, 1/4, 1/9, \ldots, 1/n^2]$ for each test problem.

To introduce errors in the available information, at each evaluation we add a randomly generated element to the exact values $f(x^{k+1})$ and $g(x^{k+1})$, with norm less or equal to σ^k and θ^k respectively.

We test 5 different forms of noise:

- $\mathtt{N}_0$: No noise, $\bar{\sigma} = \sigma^k = 0$ and $\theta^k = 0$ for all k,
- $\mathtt{N}_c^{f,g}$: Constant noise, $\bar{\sigma} = \sigma^k = 0.01$ and $\theta^k = 0.01$ for all k,
- $\mathtt{N}_v^{f,g}$: Vanishing noise, $\bar{\sigma} = 0.01$, $\sigma^k = \min\{0.01, |x^k|/100\}$, $\theta^k = \min\{0.01, |x^k|^2/100\}$ for all k,
- $\mathtt{N}_c^{g}$: Constant Gradient noise, $\bar{\sigma} = \sigma^k = 0$ and $\theta^k = 0.01$ for all k, and
- $\mathtt{N}_v^{g}$: Vanishing Gradient noise, $\bar{\sigma} = \sigma^k = 0$ and $\theta^k = \min\{0.01, |x^k|/100\}$ for all k.

The first noise form, $\mathtt{N}_0$, is used as a benchmark for comparison. Noise form $\mathtt{N}_c^{f,g}$ is representative of a noisy function where the noise is outside of the optimizer's control. The third, $\mathtt{N}_v^{f,g}$, is representative of a noisy simulation where the optimization can use some technique to reduce noise. The technique is assumed to be expensive, so the optimizer only applies the technique as a solution is approached. The fourth and fifth, $\mathtt{N}_c^{g}$ and $\mathtt{N}_v^{g}$, represent exact functions where subgradient information is approximated numerically. Like $\mathtt{N}_v^{f,g}$, in $\mathtt{N}_v^{g}$ as a solution is approached, we decrease the amount of noise.

To address the random nature of the problem, for noise forms $\mathtt{N}_c^{f,g}$, $\mathtt{N}_v^{f,g}$, $\mathtt{N}_c^{g}$, and $\mathtt{N}_v^{g}$, we repeat each test 10 times. (Noise form $\mathtt{N}_0$ is deterministic, so no repeating is required).

As for all the functions the global minimum is zero, we use the formula

$$\text{Accuracy} = \left| \log_{10}(\hat{f}^{\bar{k}}) \right|$$

to check the performance of the different methods. In all the figures that follow we plot the resulting average achieved accuracy, when running the corresponding algorithms until satisfaction of its stopping test, taking $\mathtt{tol} = 10^{-3}$ and $\mathtt{tol} = 10^{-6}$ (left and right graphs, respectively).

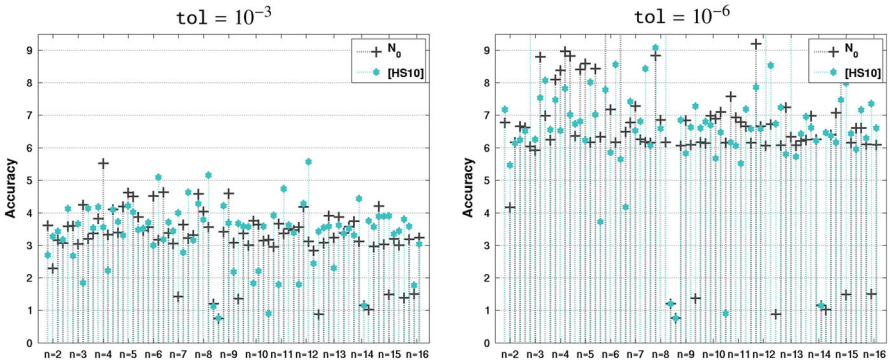

Fig. 1 Accuracy at termination for Algorithm 4 and [18]

6.2 Comparison with RedistProx algorithm in the exact case

We start by benchmarking Algorithm 4 with the exact variant $\mathtt{N}_0$ and the (exact) RedistProx Algorithm from [18]. Both methods use the relative stopping criterion $\delta^k \leq \mathtt{tol}\left(1 + |\hat{f}^k|\right)$.

Examining Fig. 1, we see that Algorithm 4 exhibits a performance comparable to the exact RedisProx method of [18]. We also notice that the relative stopping criterion is fairly successful in reaching the desired accuracy (of 3 or 6 digits). When the tolerance is 10^{-6}, Algorithm 4 seems to behave somewhat better than RedistProx, possibly because the version of the η^k-update employed in [18] is more likely to cause QP instabilities (recall Remark 1).

6.3 Impact of noise on solution accuracy

Next, we explore convergence over the variety of error forms $\mathtt{N}_c^{f,g}$,$\mathtt{N}_v^{f,g}$,$\mathtt{N}_c^{g}$,$\mathtt{N}_v^{g}$, taking as relative stopping criterion $\delta^k \leq \max\left(\mathtt{tol}, \bar{\sigma}\right)\left(1 + |\hat{f}^k|\right)$. (It is clearly unreasonable/not-meaningful to aim for accuracy higher than the error bound).

In Figs. 2 and 3, we present the algorithm's average performance when noise is present (the results are averaged across all 10 runs of the algorithm). To ease the interpretation of the graphs, we replot the results with no noise (Algorithm 4 with $\mathtt{N}_0$).

Figure 2 reports the result for constant noise (variants $\mathtt{N}_c^{f,g}$ and $\mathtt{N}_c^{g}$).

Examining Fig. 2, we see that errors in the evaluations result in poorer final accuracy, as expected. When the function values and gradients have constant noise, we achieve an accuracy roughly equal to the magnitude of that error. When function values are exact, but gradients contain constant error, the results are better, but still notably worse than when exact calculations are available. Nevertheless, we notice a general increase in accuracy as `tol` is decreased.

Figure 3 reports the results for vanishing noise (variants $\mathtt{N}_v^{f,g}$ and $\mathtt{N}_v^{g}$). In this case, when $\mathtt{tol} = 10^{-3}$, both noise forms $\mathtt{N}_v^{f,g}$ and $\mathtt{N}_v^{g}$ have an accuracy similar to the

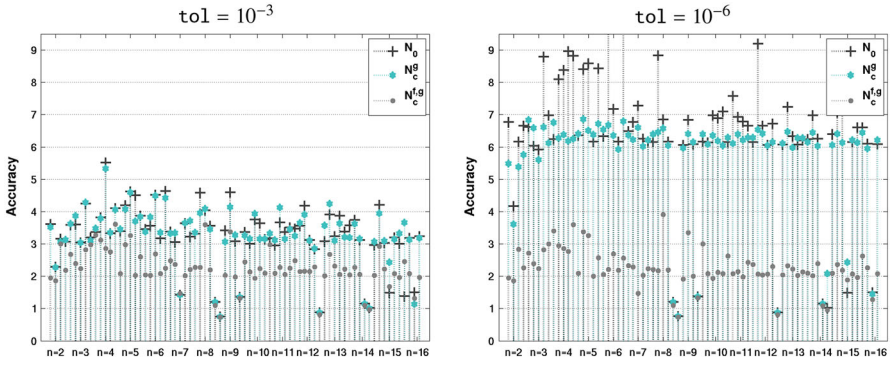

Fig. 2 Accuracy at termination for noise forms $\mathtt{N}_0$, $\mathtt{N}_c^{f,g}$, and $\mathtt{N}_c^g$

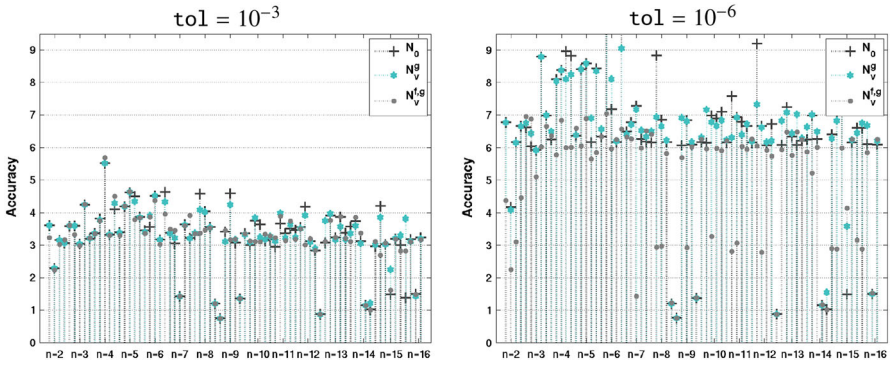

Fig. 3 Accuracy at termination for noise forms $\mathtt{N}_0$, $\mathtt{N}_v^{f,g}$, and $\mathtt{N}_v^g$

exact variant $\mathtt{N}_0$. For $\mathtt{tol} = 10^{-6}$, the noisy variants achieve reduced accuracy, but generally better than in the constant noise case.

6.4 Impact of noise on the convexification parameter η^k

We are interested in exploring the assumption that the parameter η^k remains bounded. To discuss this, define

$$\eta_{\min} := \max_{j \in J^k,\, x^j \neq \hat{x}^k} \frac{-2e_j^k}{|x^j - \hat{x}^k|^2}.$$

In [18, Rem. 2], it is shown that if $\eta > 2n$, then the convexified Ferrier polynomial $f_i + \eta/2|\cdot|^2$ is convex ($i \in \{1, 2, 3, 4, 5\}$). Thus, if there is no noise present, then we would have $\eta_{\min} \leq 2n$ at each iteration, and consequently (if there is no noise) $\eta^k \leq \eta_{\min} + \gamma \leq 2n + 2$ at each iteration (as $\gamma = 2$). Of course, in the presence of noise, we cannot expect $\eta^k \leq 2n + 2$ at all iterations. However, we would hope that η^k does not grow greatly out of proportion to this value.

Table 1 Termination value of η^k

Noise form	Problems with		
	$\eta^k \leq 2n+2$	$2n+2 < \eta^k \leq 25n$	$25n < \eta^k$
N_0	73	1	1
$\mathrm{N}_c^{f,g}$	582	94	74
$\mathrm{N}_v^{f,g}$	703	21	26
N_c^{g}	729	13	8
N_v^{g}	731	10	9

In this set of tests, we set $\texttt{tol} = 0$ and allow the algorithm to run until $25n$ function/subgradient evaluations are used (effectively forcing a limiting state to the algorithm). In Table 1, we report the number of times η^k is below $2n+2$, between $2n+2$ and $25n$, or exceeds $25n$ by the termination of the algorithm.

Examining Table 1, we see that the only situation where η^k seems somewhat uncontrolled is the noise form $\mathrm{N}_c^{f,g}$. Recalling that in that case the noise is constant on both f and g values, this is clearly the hardest noise form to deal with. Overall, the experiments support that the assumption of η^k remaining bounded is quite reasonable in general, particularly if noise asymptotically vanishes or if the f values are exact. It is interesting to note that, for all noise forms (including "no-noise") and on all tests, Ferrier polynomial f_4 in dimension 14 results in $\eta^k > 25n$.

Acknowledgments The authors thank the referees for many useful and insightful comments. In fact, this version looks very much different from (and is much better than) the original, thanks to the input received. Research of the first author is supported in part by NSERC DG program and UBC IRF. The second author is supported by CNPq 303840/2011-0, AFOSR FA9550-08-1-0370, NSF DMS 0707205, PRONEX-Optimization, and FAPERJ. The third author is supported in part by CNPq Grant 302637/2011-7, PRONEX-Optimization, and by FAPERJ.

References

1. Apkarian, P., Noll, D., Prot, O.: A proximity control algorithm to minimize nonsmooth and nonconvex semi-infinite maximum eigenvalue functions. J. Convex Anal. **16**(3–4), 641–666 (2009)
2. Astorino, A., Frangioni, A., Gaudioso, M., Gorgone, E.: Piecewise-quadratic approximations in convex numerical optimization. SIAM J. Optim. **21**, 1418–1438 (2011)
3. Bagirov, A.M., Karasözen, B., Sezer, M.: Discrete gradient method: derivative-free method for nonsmooth optimization. J. Optim. Theory Appl. **137**(2), 317–334 (2008)
4. Bihain: Optimization of upper semidifferentiable functions. J. Optim. Theory Appl., **44**, (1985)
5. Conn, A.R., Scheinberg, K., Vicente, L.N.: Introduction to Derivative-Free Optimization, volume 8 of MPS/SIAM Book Series on Optimization. SIAM, New Delhi (2009)
6. Daniilidis, A., Georgiev, P.: Approximate convexity and submonotonicity. J. Math. Anal. Appl. **291**(1), 292–301 (2004)
7. Daniilidis, A., Sagastizábal, C., Solodov, M.: Identifying structure of nonsmooth convex functions by the bundle technique. SIAM J. Optim. **20**(2), 820–840 (2009)
8. d'Antonio, G., Frangioni, A.: Convergence analysis of deflected conditional approximate subgradient methods SIAM. J. Optim. **20**(1), 357–386 (2009)
9. de Oliveira, W., Sagastizábal, C., Lemaréchal, C.: Convex proximal bundle methods in depth: a unified analysis for inexact oracles. Math. Program. **148**(1–2), 241–277 (2014)

10. Emiel, G., Sagastizábal, C.: Incremental-like bundle methods with application to energy planning. Comp. Optim. Appl. **46**, 305–332 (2010)
11. Ferrier, C.: Bornes Duales de Problémes d'Optimisation Polynomiaux, Ph.D. thesis, Laboratoire Approximation et Optimisation, Université Paul Sabatier, Toulouse, France (1997)
12. Ferrier, C.: Computation of the distance to semi-algebraic sets. ESAIM Control Optim. Calc. Var. **5**, 139–156 (2000)
13. Fletcher, R., Leyffer, S., Ralph, D., Scholtes, S.: Local convergence of SQP methods for mathematical programs with equilibrium constraints. SIAM J. Optim. **17**, 259–286 (2006)
14. Gupal, A.M.: A method for the minimization of almost differentiable functions. Kibernetika (Kiev) **1**, 114–116 (1977)
15. Hintermüller, M.: A proximal bundle method based on approximate subgradients. Comp. Optim. Appl. **20**, 245–266 (2001)
16. Hare, W., Macklem, M.: Derivative-free optimization methods for finite minimax problems. Opt. Methods Soft. **28**(2), 300–312 (2013)
17. Hare, W., Nutini, J.: A derivative-free approximate gradient sampling algorithm for finite minimax problems. Comput. Optim. Appl. **56**(1), 1–38 (2013)
18. Hare, W., Sagastizábal, C.: A redistributed proximal bundle method for nonconvex optimization. SIAM J. Optim. **20**(5), 2442–2473 (2010)
19. Hiriart-Urruty, J.-B., Lemaréchal, C.: Convex analysis and minimization algorithms. II, Volume 306 of Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Springer, Berlin (1993). Advanced theory and bundle methods
20. Kiwiel, K.C.: A linearization algorithm for nonsmooth minimization. Math. Oper. Res. **10**(2), 185–194 (1985)
21. Kiwiel, K.C.: Methods of Descent for Nondifferentiable Optimization. Springer, Berlin (1985)
22. Kiwiel, K.C.: Approximations in proximal bundle methods and decomposition of convex programs. J. Optim. Theory Appl. **84**, 529–548 (1995)
23. Kiwiel, K.C.: Restricted step and Levenberg-Marquardt techniques in proximal bundle methods for nonconvex nondifferentiable optimization. SIAM J. Optim. **6**(1), 227–249 (1996)
24. Kiwiel, K.C.: Convergence of approximate and incremental subgradient methods for convex optimization SIAM. J. Optim. **14**(3), 807–840 (2004)
25. Kiwiel, K.C.: A proximal bundle method with approximate subgradient linearizations. SIAM J. Optim. **16**(4), 1007–1023 (2006)
26. Kiwiel, K.C.: A nonderivative version of the gradient sampling algorithm for nonsmooth nonconvex optimization. SIAM J. Optim. **20**(4), 1983–1994 (2010)
27. Lemaréchal, C.: An extension of Davidon methods to non differentiable problems. Math. Progr. Study **3**, 95–109 (1975). Nondifferentiable optimization
28. Lemaréchal, C.: Bundle Methods in Nonsmooth Optimization. In: Nonsmooth Optimization (Proceedings of IIASA Workshop, Laxenburg, 1977), volume 3 of IIASA Proceeding Series, pp. 79–102. Pergamon, Oxford (1978)
29. Lemaréchal, C.: Lagrangian relaxation. In: Computational Combinatorial Optimization (Schloß Dagstuhl, 2000), volume 2241 of Lecture Notes in Computer Science, pp. 112–156. Springer, Berlin (2001)
30. Lemaréchal, C., Strodiot, J.-J., Bihain, A.: On a bundle algorithm for nonsmooth optimization. In: Nonlinear Programming, 4 (Madison, Wis., 1980), pp. 245–282. Academic Press, New York (1981)
31. Lukšan, L., Vlček, J.: A bundle-Newton method for nonsmooth unconstrained minimization. Math. Program. **83**(3, Ser. A), 373–391 (1998)
32. Lukšan, L., Vlček, J.: Globally convergent variable metric method for nonconvex nondifferentiable unconstrained minimization. J. Optim. Theory Appl. **151**(3), 425–454 (2011)
33. Luz, C.J., Schrijver, A.: A convex quadratic characterization of the Lovász theta number. SIAM J. Discret. Math. **19**(2), 382–387 (2005)
34. Mifflin, R.: Semismooth and semiconvex functions in constrained optimization. SIAM J. Control Optim. **15**(6), 959–972 (1977)
35. Mifflin, R.: Convergence of a modification of Lemaréchal's algorithm for nonsmooth optimization. In: Progress in Nondifferentiable Optimization, volume 8 of IIASA Collaborative Proceeding Series CP-82, pp. 85–95. Int. Inst. Appl. Systems Anal., Laxenburg (1982)

36. Mifflin, R.: A modification and extension of Lemarechal's algorithm for nonsmooth minimization. Math. Program. Study **17**, 77–90 (1982). Nondifferential and variational techniques in optimization (Lexington, Ky., 1980)
37. Mäkelä, M.M., Neittaanmäki, P.: Nonsmooth Optimization. Analysis and algorithms with applications to optimal control. World Scientific Publishing Co., Inc, River Edge (1992)
38. Nedić, A., Bertsekas, D.P.: The effect of deterministic noise in subgradient methods. Math. Program. **125**, 75–99 (2010)
39. Noll, D.: Bundle method for non-convex minimization with inexact subgradients and function values. In: Computational and Analytical Mathematics, vol. 50, pp. 555–592. Springer Proceedings in Mathematics (2013)
40. Pinar, M., Teboulle, M.: On semidefinite bounds for maximization of a non-convex quadratic objective over the L1-unit ball. RAIRO Oper. Res. **40**(3), 253–265 (2006)
41. Polyak, B.T.: Introduction to Optimization. Optimization Software, Inc., Publications Division, New York (1987)
42. Robert, C.P., Casella, G.: Monte Carlo Statistical Methods. Springer Texts in Statistics, 2nd edn. Springer, New York (2004)
43. Rockafellar, R.T., Wets, J.J.-B.: Variational Analysis, Volume 317 of Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Springer, Berlin (1998)
44. Sagastizábal, C.: Divide to conquer: decomposition methods for energy optimization. Math. Program. B **134**(1), 187–222 (2012)
45. Sagastizábal, C., Solodov, M.: An infeasible bundle method for nonsmooth convex constrained optimization without a penalty function or a filter. SIAM J. Optim. **16**, 146–169 (2005)
46. Solodov, M.V.: On approximations with finite precision in bundle methods for nonsmooth optimization. J. Optim. Theory Appl. **119**, 151–165 (2003)
47. Solodov, M.V.: Constraint qualifications. In: Cochran, James J., et al. (eds.) Wiley Encyclopedia of Operations Research and Management Science. Wiley, New York (2010)
48. Solodov, M.V., Zavriev, S.K.: Error stabilty properties of generalized gradient-type algorithms. J. Optim. Theory Appl. **98**, 663–680 (1998)
49. Spingarn, J.E.: Submonotone subdifferentials of Lipschitz functions. Trans. Am. Math. Soc. **264**, 77–89 (1981)