Mathematical Programming manuscript No.
(will be inserted by the editor)

Warren Hare · Claudia Sagastizábal

Computing proximal points of nonconvex functions

the date of receipt and acceptance should be inserted later

Abstract The proximal point mapping is the basis of many optimization techniques for convex functions. By means of variational analysis, the concept of proximal mapping was recently extended to nonconvex functions that are prox-regular and prox-bounded. In such a setting, the proximal point mapping is locally Lipschitz continuous and its set of fixed points coincide with the critical points of the original function. This suggests that the many uses of proximal points, and their corresponding proximal envelopes (Moreau envelopes), will have a natural extension from Convex Optimization to Nonconvex Optimization. For example, the inexact proximal point methods for convex optimization might be redesigned to work for nonconvex functions. In order to begin the practical implementation of proximal points in a nonconvex setting, a first crucial step would be to design efficient methods of approximating nonconvex proximal points. This would provide a solid foundation on which future design and analysis for nonconvex proximal point methods could flourish.

In this paper we present a methodology based on the computation of proximal points of piecewise affine models of the nonconvex function. These models can be built with only the knowledge obtained from a black box providing, for each point, the function value and one subgradient. Convergence of the method is proved for the class of nonconvex functions that are prox-bounded and lower-$\mathcal{C}^2$ and encouraging preliminary numerical testing is reported.

Keywords Nonconvex Optimization, Nonsmooth Optimization, proximal point, prox-regular, lower-$\mathcal{C}^2$

AMS Subject Classification Primary: 90C26, 49J52; Secondary: 65K10, 49J53, 49M05.

1 Introduction and motivation

Envelope functions are useful for both theoretical and practical purposes in the analysis and manipulation of lower semicontinuous functions. This is the case in particular of the *Moreau envelope* [30], also known as *Moreau-Yosida regularization* [37] and *proximal envelope* [35].

For a convex function, for instance, the function and its Moreau envelope have the same set of minima. Since the Moreau envelope function is continuously differentiable, a gradient-like method can be applied to find a minimizer of the original function, possibly nondifferentiable. As such, the well known proximal point algorithm, [4], [23], [34], can be interpreted as a preconditioned gradient method for minimizing the Moreau envelope [7]. Many other examples can be found in convex programming,

Warren Hare
IMPA, Estrada Dona Castorina 110, Jardim Botânico, Rio de Janeiro RJ 22460-320, Brazil. Research supported by CNPq Grant No. 150234/2004-0, E-mail: whare@cecm.sfu.ca

Claudia Sagastizábal
IMPA, Estrada Dona Castorina 110, Jardim Botânico, Rio de Janeiro RJ 22460-320, Brazil. On leave from INRIA, France. Research supported by CNPq Grant No. 383066/2004-2, E-mail: sagastiz@impa.br

where several variants of proximal point methods have been extensively applied to develop different algorithms, such as [33], [2], [13], [11], [3], [8].

To make use of the Moreau envelope in algorithms requires the computation, at least approximately, of the proximal point mapping. The search direction used in an inexact proximal point algorithm is an approximation of the search direction which would be created if the exact proximal point were known. This direction is generally calculated by running a few iterations of a sub-algorithm which (if run to infinity) calculates the proximal point of a function. The proof of algorithmic convergence is then based on the relationships between optimality and proximal points, and the ability to calculate a proximal point for a convex function. For example, convex variants of bundle methods, [15, vol. II], define a new *serious step* when the approximate computation of certain proximal point (at the current serious step) is considered "good enough". Bundle *null steps* are nothing but iterations to better approximate a proximal point. In [1] and [9] it is shown that proximal points can be approximated with any desired accuracy via an adequate cutting-planes model of the convex function. This result is the basis for showing convergence of a bundle method when it generates a last serious step followed by an infinite number of null steps.

Recent research into proximal points of nonconvex functions has shown that much of the theoretical power of proximal points is not constrained to the convex case. In the nonconvex case, research on Moreau envelopes and the proximal point mapping has largely hinged upon the idea of *prox-regularity*; see [31], [32], [5], [6]. First introduced in [32], prox-regularity provides the necessary structure to extend many of the results on proximal points to a nonconvex setting. Specifically, if a function is *prox-regular* and *prox-bounded*, as defined in Section 2 below, the proximal point mapping is Lipschitz continuous near stationary points and its fixed points correspond to the critical points of the original function [31, Thm. 4.4]. The same relations are shown to hold in Hilbert and Banach spaces in [6] and [5], respectively.

The results obtained for prox-regular functions suggest that a stationary point of such a function could be found by applying a nonconvex variant of the proximal point method.

For both actual implementation and to create a firm foundation for the theory of nonconvex proximal point algorithms, a first crucial step would be a method of computing the proximal point for a nonconvex function. This is the important question we address in this paper. Along the lines of [9, Sec. 4], we focus on identifying a minimal set of conditions to be satisfied along iterations by simple model functions approximating the (unknown) nonconvex function. These conditions correspond to implementable models built on the knowledge of the information given by a black box, whose output for each point is the function value and one subgradient of the function at the given point.

This paper is organized as follows. In Subsection 1.1 we review some related previous work and highlight how our approach differs from past research. Section 2 reviews Variational Analysis definitions and results. Section 3 lays the background of our methodology, with some initial theoretical results ensuring convergence to the proximal point. A constructive approach to compute proximal points for nonconvex functions and a derived implementable black box algorithm are given, respectively, in Sections 4 and 5. The algorithm is shown to converge for prox-bounded *lower-*$\mathcal{C}^2$ functions in Theorem 14. Section 6 reports some encouraging preliminary numerical results, obtained by running the algorithm on a large collection of randomly generated lower-$\mathcal{C}^2$ functions. In particular, in Subsection 6.6, comparative runs with two well-known bundle methods show the potential advantages of our approach.

1.1 Relation with past research on Nonconvex Nonsmooth Optimization

Extending to the nonconvex setting a method designed for convex functions is not a new idea. In particular, this is the case of bundle methods, [24], [19], [25], [26]. As noted, when minimizing a convex function, null step-iterations following a serious step can be interpreted as the approximate calculation of the proximal point of the function at such serious step. For this reason, our work can be related to nonconvex bundle iterations between serious steps. We emphasize that this work focuses on how to approximate *one* proximal point, for a given point x^0 fixed. In a follow-up paper we shall consider how to use our work in a dynamic setting, letting x^0 vary at each serious step and, hence, developing a nonconvex bundle-like algorithm. Keeping these comments in mind, we now examine some features of nonconvex bundle methods.

Historically, convex bundle methods were developed in the early 80's in a dual form [15, Ch. XIV]. Likewise, the first nonconvex variants of bundle methods were also developed from a dual point of view.

That is, they are designed to eventually produce an approximate subgradient (playing the role of the ε-subgradient in the convex case) going to zero. In these methods, no attention is paid to the primal view of how to model the objective function by using a cutting-planes approximation (primal forms of convex bundle methods were mostly developed in the 90's; see [15, Ch. XV]). As a result, modifications introduced in the convex algorithm to adapt it to the nonconvex setting lead to "fixes" in the modelling function, that were not supported by a theoretical background from the primal point of view. Basically, such fixes consist in redefining the so-called *linearization errors*; see [15, Def. XIV.1.2.7], which can be negative in the nonconvex case, to make them nonnegative by force. Along these lines, various possibilities were considered, for example, using either the absolute value of (negative) linearization errors, or a "subgradient locality measure" depending on a penalty coefficient; we refer to [26] for a general study on this subject.

More recent nonconvex bundle methods, such as [17], [20], [21], [22], although developed from convex proximal bundle methods based on a primal view, still pursue the track of previous dual approaches, by using subgradient locality measures, possibly quadratic, to redefine negative linearization errors. The use of quadratic subgradient locality measures is similar to the idea we present here, but has the weakness that the penalty coefficient is fixed a priori (moreover, primal information, corresponding to function values, is again ignored). Instead, as explained in Remark 9, our penalty coefficient (η_n) will be calculated dynamically during each iteration in a manner which forces linearization errors (of a penalization of the objective function) to be positive. In this way, we gain both a primal and dual view of the underlying optimization problem; also see Remark 18.

Unlike past nonconvex approaches, in our research we begin by returning to the original building blocks of the proximal point algorithms of convex optimization: the proximal point itself. A strong theoretical base for nonconvex proximal points has already begun in works such as [31], [32], [5], [6]. The next logical step is therefore to create an algorithm capable of computing the proximal point of a nonconvex function. To the best of our knowledge, the present work is the first algorithm for computing the proximal point for a nonconvex function. In Subsection 6.6 numerical tests, comparing the computation of a proximal point via our algorithm and via running previously developed bundle algorithms on the proximal point subproblem, show favourable results which further support our methodology.

Finally, by approaching the problem from this direction we will, in the long run, not just create working algorithms, but also create a better understanding of why the algorithms work and how they can be improved. For example, the algorithm developed in this paper is not only capable of calculating the proximal point of a nonconvex function, but it is also capable of informing the user when the given proximal parameter is insufficient for the desired proximal point to be well defined (a concern relevant in the nonconvex setting only). This result will make it possible to create workable nonconvex algorithms based on the nonconvex $\mathcal{VU}$ theory in [27], [28], and the nonconvex partly smooth theory in [14]. Also, a full nonconvex nonsmooth algorithm based on the building blocks developed in this work will have the distinct advantage over the previous nonconvex bundle methods that it will provide both a primal and dual view of the nonconvex optimization problem, whereas previous methods only provide a dual view of the problem. This is because, in a method created from a primal-dual foundation, a modelling function directly related to the objective function is minimized at each iteration; while in previous approaches, only the subgradients of the minimized function were related to subgradients of the objective function. As a result, it is likely that future research in this direction will be more profitable than previous directions.

2 Basic definitions and past results

In this section we recall some important concepts and results related to Variational Analysis and the proximal point mapping.

2.1 Some notions from Variational Analysis

We begin by fixing the notation used throughout this paper.

Given a sequence of vectors $\{z_k\}$ converging to 0, we write $\zeta_k = o(|z_k|)$ if and only if for all $\varepsilon > 0$ there exists $k_\varepsilon > 0$ such that $|\zeta_k| \leq \varepsilon |z_k|$ for all $k \geq k_\varepsilon$.

For a set $S \subset \mathbb{R}^N$ and a point $x \in S$:

– A vector v is *normal* to S at x if there are sequences $x^\nu \to_S x$ and $v^\nu \to v$ such that $\langle v^\nu, z - x^\nu \rangle \leq o(|z - x^\nu|)$ for $z \in S$.
– A set S is said to be *Clarke regular* at $x \in S$ when S is locally closed at x and each normal vector v satisfies $\langle v, z - x \rangle \leq o(|z - x|)$ for all $z \in S$.

Let $f : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ be a proper lsc function, so that its epigraph, denoted and defined by $\operatorname{epi} f := \{(x, \beta) \in \mathbb{R}^N \times \mathbb{R} : \beta \geq f(x)\}$, is a closed set in $\mathbb{R}^{N+1}$.

For a point $\bar{x} \in \mathbb{R}^N$ where f is finite-valued:

– We use the Mordukhovich subdifferential [29] denoted by $\partial f(\bar{x})$ in [35, Def 8.3, p. 301].
– The function f is said to be *subdifferentially regular* at $\bar{x}$ if $\operatorname{epi} f$ is a Clarke regular set at the point $(\bar{x}, f(\bar{x}))$; see [35, Def. 7.25, p. 260]. For such functions, a relaxed subgradient inequality holds:

$$f(x) \geq f(\bar{x}) + \langle \bar{g}, x - \bar{x} \rangle + o(|x - \bar{x}|)$$

for each $\bar{g} \in \partial f(\bar{x})$ and all $x \in \mathbb{R}^N$.
– The function f is *prox-regular* at $\bar{x}$ for a subgradient $\bar{g} \in \partial f(\bar{x})$ (with parameter r_{pr}) if there exists $r_{pr} > 0$ such that

$$f(x') \geq f(x) + \langle g, x' - x \rangle - \frac{r_{pr}}{2}|x' - x|^2$$

whenever x and x' are near $\bar{x}$ with $f(x)$ near $f(\bar{x})$ and $g \in \partial f(x)$ near $\bar{g}$.
When this property holds for all subgradients in $\partial f(\bar{x})$, the function is said to be prox-regular at $\bar{x}$; see [35, Def. 13.27, p. 610].
– The function f is *prox-bounded* if there exists $R \geq 0$ such that the function $f(\cdot) + R\frac{1}{2}|\cdot|^2$ is bounded below. The corresponding *threshold* (of prox-boundedness) is the smallest $r_{pb} \geq 0$ such that $f(\cdot) + R\frac{1}{2}|\cdot|^2$ is bounded below for all $R > r_{pb}$. In this case, $f(\cdot) + R\frac{1}{2}|\cdot - \bar{x}|^2$ is bounded below for any $\bar{x}$ and $R > r_{pb}$ [35, Def. 1.23, p. 20 & Thm. 1.25, p. 21].
– The function f is *lower-$\mathcal{C}^2$* on an open set V if at any point x in V, f appended with a quadratic term is a convex function on an open neighbourhood V' of x; see [35, Thm. 10.33, p. 450].

The function f is said to be prox-regular if it is prox-regular at every $\bar{x} \in \mathbb{R}^N$. In this case, the prox-regularity parameter may either be *pointwise* (i.e., dependent on $x \in \mathbb{R}^N$) or *global* (independent of $x \in \mathbb{R}^N$). Likewise, f is said to be lower-$\mathcal{C}^2$ if the corresponding neighbourhood V is the entire space $\mathbb{R}^N$ (note that this does not imply the neighbourhood V' in the definition above is the entire space).

Remark 1 (lower-$\mathcal{C}^2$) The original definition of lower-$\mathcal{C}^2$ functions states:

> *The function f is lower-$\mathcal{C}^2$ on an open set V if for each $\bar{x} \in V$ there is a neighbourhood V' of $\bar{x}$ upon which a representation $f(x) = \max_{t \in T} f_t(x)$ holds, where T is a compact set and the functions f_t are twice continuously differentiable jointly in both x and t.*

The definition we cite above is proved equivalent to the original one in [35, Thm. 10.33, p. 450].

The next lemma, whose proof is straightforward, will be very useful in dealing with lower-$\mathcal{C}^2$ functions.

Lemma 2 *Suppose the function f is lower-$\mathcal{C}^2$ on V and $\bar{x} \in V$. Then there exists $\varepsilon > 0$, $K > 0$, and $\rho > 0$ such that*

(i) for any point x^0 and parameter $R \geq \rho$ the function $f + \frac{R}{2}|\cdot - x^0|^2$ is convex and finite valued on the closed ball $\bar{B}_\varepsilon(\bar{x})$, and
(ii) the function f is Lipschitz continuous with constant K on the closed ball $\bar{B}_\varepsilon(\bar{x})$.

Proof. We consider first the function $f + \frac{\rho}{2}|\cdot - x^0|^2$. Its convexity and the fact that ρ can be selected independently of x^0 is clear from the proof of [35, Thm. 10.33, p. 450]. Finite valuedness then follows by selecting $\varepsilon > 0$ such that the closed ball $\bar{B}_\varepsilon(\bar{x})$ is contained in the neighbourhood V', and applying the original definition of lower-$\mathcal{C}^2$ functions in Remark 1. Finally, since finite valued convex functions are Lipschitz continuous on compact domains [35, Ex. 9.14, p. 359], we have the existence of the Lipschitz constant K for the function f on the ball $\bar{B}_\varepsilon(\bar{x})$. □

Clearly, any function which is bounded below is prox-bounded with threshold $r_{pb} = 0$. Convex functions are prox-regular with global prox-regularity parameter zero ($r_{pr} = 0$) [35, Thm. 13.30, p. 611] while lower-$\mathcal{C}^2$ functions are prox-regular [35, Prop. 13.33, p. 613]. Finally, if a function is prox-regular at a point, then it is subdifferentially regular at that point [35, p. 610].

2.2 Proximal envelope results

Due to its key role in this work, we consider the proximal point mapping in more detail. Given a positive parameter R, the *proximal point mapping* of the function f at the point $x \in \mathbb{R}^N$ is defined by

$$p_R f(x) := \operatorname{argmin}_{w\in\mathbb{R}^N} \left\{ f(w) + R\frac{1}{2}|w - x|^2 \right\}, \tag{1}$$

where the minima above can form a nonsingleton set. Even for prox-bounded functions, some choices of R (below the prox-boundedness threshold) may result in $p_R f(x)$ being the empty set. Uniqueness of $p_R f(x)$ requires prox-boundedness and prox-regularity; see Theorem 4 below.

Usually, R and x are called the *prox-parameter* and *prox-center* of (1). The *proximal envelope* is the function that assigns to each x the optimal objective value in equation (1).

Intuitively, as the prox-parameter increases, the related proximal points should converge towards the given prox-center. We next show this result for Lipschitz continuous functions and give a simple bound for the distance between the set of proximal points and the prox-center.

Lemma 3 (Lipschitz continuous functions and proximal points) *Suppose the function f is Lipschitz continuous with constant K. Then for any $R > 0$ such that $p_R f(x^0)$ is nonempty the following relation holds:*

$$|p - x^0| < \frac{2K}{R} \quad \textit{for all} \quad p \in p_R f(x^0).$$

Proof. Given a prox-parameter $R > 0$ such that $p_R f(x^0)$ is nonempty, the inclusion $p_R f(x^0) \subset \left\{x : f(x) + \frac{R}{2}|x - x^0|^2 \le f(x^0)\right\}$ holds by (1). The result follows, using the Lipschitz continuity of f:

$$\begin{aligned} p_R f(x^0) &\subset \left\{x : f(x) + \tfrac{R}{2}|x - x^0|^2 \le f(x^0)\right\} \\ &\subseteq \left\{x : f(x^0) - K|x - x^0| + \tfrac{R}{2}|x - x^0|^2 \le f(x^0)\right\} \\ &= \left\{x : |x - x^0| \le \tfrac{2K}{R}\right\}. \end{aligned}$$

□

In [35, Prop. 13.37, p. 617], the (pointwise) assumptions of prox-regularity and prox-boundedness are used to show that the proximal point mapping is well defined near stationary points. The following proposition extends this result to points that need not be stationary and gives (global) conditions for the result to hold in the whole space.

Theorem 4 *Suppose that the lsc function f is prox-regular at $\bar{x}$ for $\bar{g} \in \partial f(\bar{x})$ with parameter r_{pr}, and that f is prox-bounded with threshold r_{pb}. Then for any $R > \max\{r_{pr}, r_{pb}\}$ and all x near $\bar{x} + \frac{1}{R}\bar{g}$ the following hold:*

- *the proximal point mapping $p_R f$ is single-valued and locally Lipschitz continuous,*
- *the gradient of the proximal envelope at x is given by $R(x - p_R f(x))$, and*
- *the proximal point is uniquely determined by the relation*

$$p = p_R f(x) \iff R\,(x - p) \in \partial f(p).$$

Suppose instead the lsc function f is prox-bounded with threshold r_{pb} and lower-$\mathcal{C}^2$ on V. Let $\bar{x} \in V$ and let $\varepsilon > 0$, $K > 0$ and $\rho > 0$ be given by Lemma 2. Then for any

$$R > R_{\bar{x}} := \max\{4K/\varepsilon, \rho, r_{pb}\}$$

and all $x \in B_{\varepsilon/2}(\bar{x})$ the three items above hold.

Thus, $x^\star$ is a stationary point of f if and only if $x^\star = p_R f(x^\star)$ for any $R > R_{x^\star}$.

Proof. Validity of the three items when the assumptions hold at $\bar{x}$ for $\bar{g} \in \partial f(\bar{x})$ follows from applying [31, Thm. 4.4] to the function $h(x) := f(x + \bar{x}) - \langle \bar{g}, x \rangle$, which is prox-regular at $x := 0$ for the subgradient $0 = \bar{g} - \bar{g} \in \partial h(0)$ and prox-bounded with the same threshold of prox-boundedness as f [35, Ex. 13.35, p. 615 & Ex. 1.24, p. 20].

Consider now the lower-$\mathcal{C}^2$ case. For a given $\bar{x}$ and the values $\varepsilon > 0$, $K > 0$ and $\rho > 0$ provided by Lemma 2 we have $f + \frac{\rho}{2}|\cdot - x|^2$ is convex on $\bar{B}_\varepsilon(\bar{x})$ for any $x \in \bar{B}_\varepsilon(\bar{x})$. As such, for any $x \in \bar{B}_\varepsilon(\bar{x})$, we have $f + \frac{\rho}{2}|\cdot - x|^2$ is prox-regular at x with parameter 0. Therefore, for any $x \in \bar{B}_\varepsilon(\bar{x})$, f is prox-regular at x with parameter ρ.

Since f is lower-$\mathcal{C}^2$ there also exists $K > 0$ such that f is Lipschitz continuous with constant K on $\bar{B}_\varepsilon(\bar{x})$, by Lemma 2. Letting $i_{B_\varepsilon(\bar{x})}$ denote the indicator function of the open ball $B_\varepsilon(\bar{x})$, our assumptions imply that the function $\tilde{f} := f + i_{B_\varepsilon(\bar{x})}$ is prox-regular with global parameter $\bar{r}_{pr}$. Moreover, if $R \geq 4K/\varepsilon$, then for any $x \in B_{\varepsilon/2}(\bar{x})$ by Lemma 3 we have that $p_R\tilde{f}(x) \in B_{\varepsilon/2}(x)$. Set

$$R_{\bar{x}} := \max\{4K/\varepsilon, \rho, r_{pb}\},$$

and notice that for any $R > R_{\bar{x}}$ the function $\tilde{f}$ is globally prox-regular with parameter $\rho \leq R$, prox-bounded with threshold $r_{pb} \leq R$, and satisfies

$$p_R f(x) = p_R\tilde{f}(x) \quad \text{for all } x \in B_{\varepsilon/2}(\bar{x}).$$

Now select any $x \in B_{\varepsilon/2}(\bar{x})$ and fix $R > R_{\bar{x}}$. Letting $p = p_R\tilde{f}(x)$, we have that $R(x - p) \in \partial\tilde{f}(p)$. The mapping $p_R\tilde{f}$ satisfies the three (pointwise) items in this proposition, with $\bar{x} = p \in B_{\varepsilon/2}(\bar{x})$, $\bar{g} = R(x - p) \in \partial\tilde{f}(p)$, and $r_{pr} = \bar{r}_{pr}$, near the point $p + \frac{1}{R}(R(x - p)) = x$. Since $p_R\tilde{f} = p_R f$ near x, the proof is complete. □

3 Main idea and results

In order to compute $p_R f(x^0)$ for a given $x^0 \in \mathbb{R}^N$, we proceed by splitting the prox-parameter R into two nonnegative terms η and μ satisfying $R = \eta + \mu$. These terms play two distinct roles derived from the relation

$$p_R f(x^0) = p_\mu(f + \eta\frac{1}{2}|\cdot - x^0|^2)(x^0).$$

In this expression, η defines a "convexified" function which we shall model by a simple function φ_η, while μ is used as a prox-parameter for this model. We refer to η and μ as the *convexification parameter* and *model prox-parameter*, respectively. Along the iterative process, η, μ and φ_η have to be suitably modified to guarantee convergence to the proximal point, or to detect failure if R is not sufficiently large for the proximal point to be well defined.

More formally, for $n = 0, 1, 2, \ldots$, given positive parameters η_n and μ_n and a model function

$$\varphi_{n,\eta_n} \approx f + \eta_n\frac{1}{2}|\cdot - x^0|^2,$$

we let $x^{n+1} \in \mathbb{R}^N$ be defined as follows:

$$x^{n+1} := \operatorname{argmin}_w \left\{ \varphi_{n,\eta_n}(w) + \mu_n\frac{1}{2}|w - x^0|^2 \right\}. \tag{2}$$

Since $x^{n+1} = p_{\mu_n}\varphi_{n,\eta_n}(x^0)$, we call it an *approximal point*.

To ensure that the sequence of approximal points is well defined and converges to $p_R f(x^0)$, the data in equation (2) must be suitably chosen. In our construction, the family of parameters and model functions satisfies:

$$\varphi_{n,\eta_n} \text{ is a convex function,} \tag{3a}$$

$$\varphi_{n,\eta_n}(x^0) \le f(x^0), \tag{3b}$$

$$\varphi_{n+1,\eta_{n+1}}(w) \ge \varphi_{n,\eta_n}(x^{n+1}) + \mu_n \left\langle x^0 - x^{n+1}, w - x^{n+1} \right\rangle \text{ for all } w \in \mathbb{R}^N, \tag{3c}$$

$$\varphi_{n,\eta_n}(w) \ge f(x^n) + \frac{1}{2}\eta_n |x^n - x^0|^2 + \langle g_{\mathbf{f}}^n + \eta_n(x^n - x^0), w - x^n \rangle \text{ for some } g_{\mathbf{f}}^n \in \partial f(x^n) \text{ and all } w \in \mathbb{R}^N, \tag{3d}$$

$$\underline{\mu} = \mu_n \text{ and } \eta_n = \overline{\eta} \text{ for some positive } \underline{\mu}, \text{ nonnegative } \overline{\eta}, \text{ and } n \text{ sufficiently large.} \tag{3e}$$

The parameter η_n will be defined to satisfy $\eta_n = R - \mu_n$, thus in condition (3e) the relation $\overline{\eta} = R - \underline{\mu}$ holds. Note that for model prox-parameters to be positive (a reasonable request), convexification parameters should always be smaller than R.

In Section 4 we give a black box implementable method ensuring satisfaction of conditions (3) above. In this method, the sequence of model prox-parameters $\{\mu_n\}$ is nonincreasing, thus the corresponding sequence of convexification parameters $\{\eta_n\}$ will be nondecreasing.

Global assumptions on the model functions, i.e., to be verified for all points in $\mathbb{R}^N$, are sound when f is a convex function; however, in our nonconvex setting, we should rather rely on local conditions. Some of the conditions listed in (3) are similar to the global conditions of [9, Sec. 4], but transformed into local counterparts that are more suitable for a nonconvex function f. In particular, condition (3b) can be thought of as a weakened form of [9, eq. (4.7)] which states that $\varphi_{n,\eta_n} \le f$ at all points (instead of at x^0 only). Condition (3c) is exactly equivalent to [9, eq. (4.8)], while condition (3d) is [9, eq. (4.9)], stated for the function $f + \frac{1}{2}\eta_n|\cdot - x^0|^2$; see Section 4 below. Although for our results we need satisfaction of (3c) at the point x^{n+1} only, it seems more natural to state the relation for any $w \in \mathbb{R}^N$. The next lemma uses conditions (3a), (3b), (3c), and (3e) to show convergence of the sequence of approximal points.

Lemma 5 *For $n = 0, 1, \ldots$, consider the sequence $\{x^n\}$ defined by (2), where the family of parameters and model functions satisfies conditions (3a), (3b), (3c), and (3e). The following hold:*
(i) the sequence $\left\{\varphi_{n,\eta_n}(x^{n+1}) + \mu_n \frac{1}{2}|x^{n+1} - x^0|^2\right\}$ is eventually strictly increasing and convergent.
(ii) The sequence $\{x^n\}$ is bounded and $|x^{n+1} - x^n| \to 0$ as $n \to \infty$.

Proof. Take n big enough for (3e) to hold. For notational simplicity, consider the function

$$L_n(w) := \varphi_{n,\overline{\eta}}(x^{n+1}) + \underline{\mu}\frac{1}{2}|x^{n+1} - x^0|^2 + \underline{\mu}\frac{1}{2}|w - x^{n+1}|^2$$

and let $s^{n+1} := \underline{\mu}(x^0 - x^{n+1}) \in \partial\varphi_{n,\overline{\eta}}(x^{n+1})$ be the subgradient given by the optimality condition corresponding to (2).
Sequence elements in (i) correspond both to $L_n(x^{n+1})$ and the optimal objective value of problem (2). Since x^{n+1} is the unique minimizer of this quadratic programming problem, $L_n(x^{n+1}) < \varphi_{n,\overline{\eta}}(x^0)$, and condition (3b) implies that

$$L_n(x^{n+1}) \le f(x^0) \text{ for all } n. \tag{4}$$

Condition (3c) written with $w = x^{n+2}$ gives the inequality

$$\varphi_{n+1,\overline{\eta}}(x^{n+2}) \ge \varphi_{n,\overline{\eta}}(x^{n+1}) + \left\langle s^{n+1}, x^{n+2} - x^{n+1} \right\rangle.$$

By definition, $L_{n+1}(x^{n+2}) = \varphi_{n+1,\overline{\eta}}(x^{n+2}) + \underline{\mu}\frac{1}{2}|x^{n+2} - x^0|^2$; hence,

$$L_{n+1}(x^{n+2}) \ge L_n(x^{n+1}) + \left\langle s^{n+1}, x^{n+2} - x^{n+1} \right\rangle + \underline{\mu}\frac{1}{2}|x^{n+2} - x^0|^2 - \underline{\mu}\frac{1}{2}|x^{n+1} - x^0|^2.$$

After expanding squares, the two rightmost terms above satisfy the relation

$$\begin{aligned}\underline{\mu}\tfrac{1}{2}|x^{n+2} - x^0|^2 - \underline{\mu}\tfrac{1}{2}|x^{n+1} - x^0|^2 &= -\underline{\mu}\left\langle x^0 - x^{n+1}, x^{n+2} - x^{n+1} \right\rangle + \underline{\mu}\tfrac{1}{2}|x^{n+2} - x^{n+1}|^2 \\ &= -\left\langle s^{n+1}, x^{n+2} - x^{n+1} \right\rangle + \underline{\mu}\tfrac{1}{2}|x^{n+2} - x^{n+1}|^2,\end{aligned}$$

so we obtain

$$L_{n+1}(x^{n+2}) \geq L_n(x^{n+1}) + \underline{\mu}\frac{1}{2}|x^{n+2} - x^{n+1}|^2. \tag{5}$$

Therefore, the sequence in (i) is strictly increasing for n big enough. It is also bounded above, by (4), so it must converge.
By conditions (3a) and (3b),

$$L_n(x^0) = \varphi_{n,\overline{\eta}}(x^{n+1}) + \left\langle s^{n+1}, x^0 - x^{n+1}\right\rangle \leq \varphi_{n,\overline{\eta}}(x^0) \leq f(x^0).$$

Since $L_n(x^{n+1}) + \underline{\mu}\frac{1}{2}|x^0 - x^{n+1}|^2 = L_n(x^0)$, we obtain, using condition (3e), that

$$\underline{\mu}\frac{1}{2}|x^0 - x^{n+1}|^2 + L_n(x^{n+1}) \leq f(x^0).$$

By (i), $L_n(x^{n+1})$ converges, so the sequence $\{x^n\}$ is bounded. The second part of item (ii) follows from equation (5) by noting

$$0 \leq \underline{\mu}\frac{1}{2}|x^{n+1} - x^n|^2 \leq L_n(x^{n+1}) - L_{n-1}(x^n) \to 0.$$

□

To show convergence of the sequence of approximal points to $p_R f(x^0)$, we apply a local nonconvex analog of the condition "$\varphi_{n,\eta_n} \leq f$". This is the purpose of condition (6) below, a relation we show holds for lower-$\mathcal{C}^2$ functions when applying the algorithm CompProx; see Section 5 and Theorem 14 below.

Theorem 6 *Suppose the lsc function f is prox-bounded and lower-$\mathcal{C}^2$ on V. Let $x^0 \in V$, and let $\varepsilon > 0$, $K > 0$ and $\rho > 0$ be given by Lemma 2 (written with $\bar{x} = x^0$). Suppose that $R > \max\{4K/\varepsilon, \rho, r_{pb}\}$ and consider the sequence $\{x^n\}$ defined by (2), where the family of parameters and model functions satisfies (3). Let p be an accumulation point of the sequence $\{x^n\}$, and let the subsequence $\{x^{n_i}\} \to p$ as $i \to \infty$. If f is continuous at p, and $\overline{\eta} \in [0, R)$ is such that*

$$\varphi_{n_i,\overline{\eta}}(w) \leq f(w) + \overline{\eta}\frac{1}{2}|w - x^0|^2 \quad \textit{for all } w \textit{ near } p \textit{ and } i \textit{ sufficiently large,} \tag{6}$$

then p is the proximal point of f at x^0: $p = p_R f(x^0)$. Furthermore,

$$\lim_{i\to\infty}\left(\varphi_{n_i,\eta_{n_i}}(x^{n_i+1}) + \mu_{n_i}\frac{1}{2}|x^{n_i+1} - x^0|^2\right) = f(p_R f(x^0)) + R\frac{1}{2}|p_R f(x^0) - x^0|^2. \tag{7}$$

If f is continuous and condition (6) holds near all accumulation points of $\{x^n\}$, then the entire sequence $\{x^n\}$ converges to the proximal point $p_R f(x^0)$ and (7) holds for the entire sequence.

Proof. Without loss of generality, we only consider the concluding portion of the sequence $\{x^n\}$, so assume $\eta_n = \overline{\eta}$ and $\mu_n = \underline{\mu}$ throughout.
Recall that in Lemma 5, for the bounded sequence $\{x^n\}$ we showed that $|x^{n+1} - x^n| \to 0$ as $n \to \infty$. Therefore $x^{n_i} \to p$ as $i \to \infty$ implies $x^{n_i+1} \to p$ as $i \to \infty$. For w near p, by equation (6) and the fact $\underline{\mu}(x^0 - x^{n+1})$ is a subgradient of the convex function $\varphi_{n,\overline{\eta}}$ at x^{n+1}, we have

$$\begin{aligned}
f(w) &\geq \varphi_{n_i,\overline{\eta}}(w) - \overline{\eta}\tfrac{1}{2}|w - x^0|^2 \\
&\geq \varphi_{n_i,\overline{\eta}}(x^{n_i+1}) - \overline{\eta}\tfrac{1}{2}|w - x^0|^2 + \underline{\mu}\langle x^0 - x^{n_i+1}, w - x^{n_i+1}\rangle \\
&= \varphi_{n_i,\overline{\eta}}(x^{n_i+1}) - \overline{\eta}\tfrac{1}{2}|x^{n_i+1} - x^0|^2 + \overline{\eta}\tfrac{1}{2}|x^{n_i+1} - x^0|^2 - \overline{\eta}\tfrac{1}{2}|w - x^0|^2 + \underline{\mu}\langle x^0 - x^{n_i+1}, w - x^{n_i+1}\rangle \\
&= \varphi_{n_i,\overline{\eta}}(x^{n_i+1}) - \overline{\eta}\tfrac{1}{2}|x^{n_i+1} - x^0|^2 + (\underline{\mu} + \overline{\eta})\langle x^0 - x^{n_i+1}, w - x^{n_i+1}\rangle - \overline{\eta}\tfrac{1}{2}|w - x^{n_i+1}|^2.
\end{aligned}$$

Condition (3d) written with $n = n_i$ for $w = x^{n_i+1}$ gives the following inequality,

$$\varphi_{n_i,\overline{\eta}}(x^{n_i+1}) - \frac{1}{2}\overline{\eta}|x^{n_i+1} - x^0|^2 \geq f(x^{n_i}) + \frac{1}{2}\overline{\eta}|x^{n_i} - x^0|^2 - \frac{1}{2}\overline{\eta}|x^{n_i+1} - x^0|^2 + \langle g_{\mathtt{f}}^{n_i} + \overline{\eta}(x^{n_i} - x^0), x^{n_i+1} - x^{n_i}\rangle,$$

for some $g_{\mathtt{f}}^{n_i} \in \partial f(x^{n_i})$. It follows that

$$\begin{aligned} f(w) &\geq \varphi_{n_i,\overline{\eta}}(x^{n_i+1}) - \overline{\eta}\tfrac{1}{2}|x^{n_i+1} - x^0|^2 + (\underline{\mu} + \overline{\eta})\langle x^0 - x^{n_i+1}, w - x^{n_i+1}\rangle - \overline{\eta}\tfrac{1}{2}|w - x^{n_i+1}|^2 \\ &\geq f(x^{n_i}) + \tfrac{1}{2}\overline{\eta}|x^{n_i} - x^0|^2 - \tfrac{1}{2}\overline{\eta}|x^{n_i+1} - x^0|^2 + \langle g_{\mathtt{f}}^{n_i} + \overline{\eta}(x^{n_i} - x^0), x^{n_i+1} - x^{n_i}\rangle \\ &\quad + R\langle x^0 - x^{n_i+1}, w - x^{n_i+1}\rangle - \tfrac{1}{2}\overline{\eta}|w - x^{n_i+1}|^2. \end{aligned}$$

Passing to the limit as $i \to \infty$, and using that f is continuous at p, yields the relations

$$\begin{aligned} f(w) &\geq \lim_{i\to\infty} \varphi_{n_i,\overline{\eta}}(x^{n_i+1}) - \overline{\eta}\frac{1}{2}|p - x^0|^2 + (\underline{\mu} + \overline{\eta})\langle x^0 - p, w - p\rangle - \overline{\eta}\frac{1}{2}|w - p|^2 \\ &\geq f(p) + R\left\langle x^0 - p, w - p\right\rangle - \overline{\eta}\tfrac{1}{2}|w - p|^2, \end{aligned}$$

for all w near p. Since $\overline{\eta}\frac{1}{2}|w - p|^2 = o(|w - p|)$, the last inequality means that $R(x^0 - p) \in \partial f(p)$, an inclusion equivalent to having $p = p_R f(x^0)$, by Theorem 4. Furthermore, evaluating the relations at $w = p$ shows that

$$\lim_{i\to\infty} \varphi_{n_i,\overline{\eta}}(x^{n_i+1}) = f(p) + \overline{\eta}\frac{1}{2}|p - x^0|^2,$$

which is equivalent to equation (7).
Convergence of the whole sequence $\{x^n\}$, when f is continuous and (6) holds at all accumulation points follows from the fact that, for prox-regular functions, the proximal point is unique. □

Theorem 6 shows convergence of the sequence of iterates to the proximal point, thereby extending to the nonconvex setting Proposition 4.3 of [9]. It further shows the corresponding optimal values in (2) to converge to the value of the proximal envelope at x^0.

We now proceed to construct adequate models with information provided by the black box.

4 Building the approximations

In this section we consider a constructive manner to define along iterations model functions, φ_{n,η_n}, as well as convexification parameters, η_n. Model prox-parameters, μ_n, will essentially be defined to ensure that $R = \eta_n + \mu_n$.

An important difference with respect to the convex case arises at this stage. Since the model function approximates $f + \eta\frac{1}{2}|\cdot - x^0|^2$ for varying η, we need not only the functional and subgradient values of f, but also those of the quadratic term $|\cdot - x^0|^2$. Accordingly, at iteration n, our bundle of information is composed of

$$\mathcal{B}\mathtt{f}_n := \left\{x^i,\, f_{\mathtt{f}}^i := f(x^i),\, g_{\mathtt{f}}^i \in \partial f(x^i)\right\}_{i\in I_n},\ \text{with } I_n \subset \{0, 1, \ldots, n\}. \tag{8}$$

The corresponding quadratic bundle elements, computable from $\mathcal{B}\mathtt{f}_n$, are denoted and defined by

$$\mathcal{B}\mathtt{q}_n := \left\{ f_{\mathtt{q}}^i := \frac{1}{2}|x^i - x^0|^2,\, g_{\mathtt{q}}^i := x^i - x^0 \right\}_{i\in I_n}. \tag{9}$$

The following useful relation holds for any pair of indices i, j:

$$f_{\mathtt{q}}^j - f_{\mathtt{q}}^i - \langle g_{\mathtt{q}}^i, x^j - x^i\rangle = \frac{1}{2}|x^i - x^j|^2. \tag{10}$$

With the notation introduced in (8) and (9), the model function used to compute x^{n+1} is given by

$$\varphi_{n,\eta_n}(w) := \max_{i\in I_n} \left\{ f_{\mathtt{f}}^i + \eta_n f_{\mathtt{q}}^i + \left\langle g_{\mathtt{f}}^i + \eta_n g_{\mathtt{q}}^i, w - x^i\right\rangle \right\}. \tag{11}$$

Since φ_{n,η_n} is a piecewise affine function, finding x^{n+1} via equation (2) amounts to solving a linear-quadratic programming problem on a simplex; see [16], [12] for specially designed solvers.

In order to solve (2), we must also define the convexification and model prox- parameters. For the convexification parameter, in particular, we desire to eventually satisfy conditions (3) and (6). Although it is not practical (nor possible with just black box information) to check an infinite number of points, we can update η so that condition (6) holds at least for all points in the bundle. The next lemma states a computable lower bound which ensures satisfaction of (6), written with $\overline{\eta}$ replaced by η_n, for all bundle elements; see (13) below.

Lemma 7 *Consider the bundle defined by* (8)-(9) *at iteration n with I_n containing at least two elements and the value*

$$\tilde{\eta}_n := \max_{i \neq j \in I_n} \left\{ \frac{f_{\mathbf{f}}^j - f_{\mathbf{f}}^i - \left\langle g_{\mathbf{f}}^j, x^j - x^i \right\rangle}{\frac{1}{2}|x^j - x^i|^2} \right\}. \tag{12}$$

The function φ_{n,η_n} defined by (11) *at iteration n satisfies*

$$\varphi_{n,\eta_n}(x^j) \leq f_{\mathbf{f}}^j + \eta_n f_{\mathbf{q}}^j \quad \text{for all } j \in I_n \tag{13}$$

whenever $\eta_n \geq \tilde{\eta}_n$.

Proof. We begin by noting that equation (13) holds if and only if for $\eta = \eta_n$

$$f_{\mathbf{f}}^i + \eta f_{\mathbf{q}}^i + \langle g_{\mathbf{f}}^i + \eta g_{\mathbf{q}}^i, x^j - x^i \rangle \leq f_{\mathbf{f}}^j + \eta f_{\mathbf{q}}^j \quad \text{for all } i, j \in I_{n+1}.$$

Using (10) we see satisfaction of (13) is equivalent to having

$$\frac{f_{\mathbf{f}}^i - f_{\mathbf{f}}^j - \langle g_{\mathbf{f}}^i, x^i - x^j \rangle}{\frac{1}{2}|x^j - x^i|^2} = \frac{f_{\mathbf{f}}^i - f_{\mathbf{f}}^j + \langle g_{\mathbf{f}}^i, x^j - x^i \rangle}{f_{\mathbf{q}}^j - f_{\mathbf{q}}^i - \langle g_{\mathbf{q}}^i, x^j - x^i \rangle} \leq \eta \quad \text{for all } i \neq j \in I_n$$

when $\eta = \eta_n$. Hence for any $\eta_n \geq \tilde{\eta}_n$ equation (13) holds. □

We now give a computable rule based on the lower bound in Lemma 7 that defines convexification and model prox- parameters in a monotone relation (nondecreasing for $\{\eta_n\}$ and nonincreasing for $\{\mu_n\}$). The rule makes use of a *convexification growth parameter* Γ, fixed at the beginning of the algorithm.

Proposition 8 *Fix a* convexification growth parameter $\Gamma > 1$. *For $n \geq 1$, consider the bundle sets defined by* (8)-(9) *at iteration $n-1$. Suppose the parameters satisfy the relation $\mu_{n-1} + \eta_{n-1} = R$ with μ_{n-1} positive, and that the index set I_n contains at least two elements. Having the lower bound $\tilde{\eta}_n$ from* (12)*, define new parameters according to the rule*

$$\begin{cases} \eta_n := \eta_{n-1} \text{ and } \mu := \mu_{n-1} & \text{if } \tilde{\eta}_n \leq \eta_{n-1} \\ \eta_n := \Gamma \tilde{\eta}_n \text{ and } \mu_n := R - \Gamma \tilde{\eta}_n & \text{if } \tilde{\eta}_n > \eta_{n-1}. \end{cases} \tag{14}$$

Then, either both parameters remain unchanged or they are strictly monotone:

$$\eta_{n-1} < \eta_n \quad \text{and} \quad \mu_{n-1} > \mu_n.$$

Moreover, $\eta_n \geq \max\{\eta_{n-1}, \tilde{\eta}_n\}$ and $\eta_n + \mu_n = R$. Lastly, if $\tilde{\eta}_n < R/\Gamma$, then $\eta_n < R$ and $\mu_n > 0$ positive.

Proof. All of the statements follow easily from rule (14). □

Remark 9 As mentioned in Subsection 1.1, nonnegativity of linearization errors is crucial for nonconvex bundle methods. The i^{th} linearization error for f (at x^0) is defined by

$$e_i^f = f(x^0) - \Big(f_{\mathbf{f}}^i + \langle g_{\mathbf{f}}^i, x^0 - x^i\rangle\Big).$$

For the "convexified" function $f_\eta := f + \eta\frac{1}{2}|\cdot - x^0|^2$, linearization errors have the expression

$$e_i^{f_\eta} = f(x^0) - \Big(f_{\mathbf{f}}^i + \eta f_{\mathbf{q}}^i + \langle g_{\mathbf{f}}^i + \eta g_{\mathbf{q}}^i, x^0 - x^i\rangle\Big).$$

From the relation

$$e_i^{f_\eta} = e_i^f + \eta\frac{1}{2}|x^i - x^0|^2,$$

we see that the value $\tilde\eta_n$ calculated in equation (12) is nothing but the smallest value of η ensuring that $e_i^{f_\eta} \geq 0$ for all $i \in I_n$. As a result, for any value of $\eta \geq \tilde\eta_n$, the corresponding function f_η will appear convex, as far as it can be seen from the available (bundle) information. Our rule (14) makes use of $\tilde\eta_n$ in a manner that can be viewed as a primal consistent method of forcing linearization errors to be nonnegative.

At this stage, it is interesting to refer to previous work exploring modifications of the cutting-planes model, although restricted to a convex function f. The modified model function introduced by [36] and further developed by [17], is the maximum over $i \in I_n$ of *tilted* affine functions, given by

$$\begin{cases} f_{\mathbf{f}}^i + (1-\theta)\langle g_{\mathbf{f}}^i, \cdot - x^i\rangle & \text{if } e_i^f > -\frac{\theta}{1-\kappa}\langle g_{\mathbf{f}}^i, x^0 - x^i\rangle \\ \kappa f_{\mathbf{f}}^i + (1-\kappa)f(x^0) + (1-\theta)\langle g_{\mathbf{f}}^i, \cdot - (\frac{\kappa}{1-\theta}x^i + (1-\frac{\kappa}{1-\theta})x^0\rangle\Big) & \text{otherwise,} \end{cases}$$

where $\kappa \in]0,1]$ and $\theta \in [0, 1-\kappa]$ are given parameters. In the expression above, note that the slope of the linearization is always tilted by a factor $1-\theta$. In addition, the linearization is shifted and (the slope further tilted) if the error e_i^f is not sufficiently positive, a consideration related to our definition of $\tilde\eta$. The original proposal in [36] corresponds to always defining the linearization by the first line above, for $\theta \in [0, 1/2]$. It is motivated by extending a modified Newton's method for finding multiple roots of a polynomial to the minimization of a strictly convex quadratic function. This is probably the reason why it is argued in [17] that *tilting errors avoids visiting successive "corners" of the epigraph of f and speeds up convergence, by employing second order information in the "ridge"* (i.e., in the $\mathcal{U}$-subspace in [27]). However, the resulting model can cut the graph of the (convex) function f, unlike the method we present here. More precisely, when the function f is convex, the numerator in (12) will always be nonpositive, yielding $\tilde\eta_n \leq 0$ for all n. By (14), η_n never increases and, when started with $\eta_0 = 0$, our model simply reduces to a standard cutting-planes model; see also the comments following Corollary 12 below.

An important effect of our parameter update is that, whenever the function $f + \eta_{n-1}|\cdot - x^0|^2$ is (locally) convex, the rule leaves the parameters unchanged, as shown in the next lemma.

Lemma 10 *For $n \geq 1$, consider the bundle sets defined by* (8)-(9) *at iteration $n-1$. Suppose the parameters satisfy the relation $\mu_{n-1} + \eta_{n-1} = R$ with μ_{n-1} positive, and that the index set I_n contains at least two elements. If the function $f + \hat\eta\frac{1}{2}|\cdot - x^0|^2$ is convex on an open ball containing $\{x^i : i \in I_n\}$, then the lower bound from* (12) *satisfies $\tilde\eta_n \leq \hat\eta$. Hence, if the function $f + \eta_{n-1}\frac{1}{2}|\cdot - x^0|^2$ is convex on an open ball containing $\{x^i : i \in I_n\}$, rule* (12)-(14) *sets $\eta_n = \eta_{n-1}$.*

Proof. Note that for arbitrary $i \in I_n$, $g_{\mathbf{f}}^i + \hat\eta g_{\mathbf{q}}^i \in \partial(f + \hat\eta\frac{1}{2}|\cdot - x^0|^2)(x^i)$. By the convexity assumption, the corresponding subgradient inequalities yield the relations

$$f_{\mathbf{f}}^i + \hat\eta f_{\mathbf{q}}^i + \langle g_{\mathbf{f}}^i + \hat\eta g_{\mathbf{q}}^i, x^j - x^i\rangle \leq f_{\mathbf{f}}^j + \hat\eta f_{\mathbf{q}}^j \text{ for any } i, j \in I_n.$$

Using (10) we see that

$$\frac{f_{\mathbf{f}}^i - f_{\mathbf{f}}^j + \langle g_{\mathbf{f}}^i, x^j - x^i\rangle}{0.5|x^j - x^i|^2} \leq \hat\eta \quad \text{for all } i \neq j \in I_n.$$

Thus, the lower bound from (12) satisfies the inequality $\tilde{\eta}_n \leq \hat{\eta}$, and the final result follows. □

Another important consequence of Proposition 8 is that when the lower bound $\tilde{\eta}_n$ is bigger than the prox-parameter R, the relation in (13) cannot hold with $\overline{\eta} \leq R$, and nothing can be said about the convergence of the approximal sequence. In the algorithm stated in Section 6, we use such lower bound to check if the given prox-parameter R is sufficiently large for our methodology to work.

The size of quadratic programming problems (2) is directly related to the bundle cardinality, i.e., to the number of indices in I_n. For the sake of efficiency, such cardinality should not grow beyond control as n increases. So far, in rule (12)-(14) the only requirement for index sets is to contain more than one element. In order to ensure satisfaction of (3b) and (3d), these elements should at least correspond to the prox-center and to the most recent approximal point ($\{0, n+1\} \subseteq I_{n+1}$). Satisfaction of condition (3c) is guaranteed if all active bundle points at x^{n+1} enter the index set I_{n+1}; this is the purpose of the index set J_n^{act} defined below.

Accordingly, for a given iteration n, we define a family of parameters and model functions $\mathcal{F} := \{\mu_n, \eta_n, \varphi_{n,\eta_n}\}_n$ via the following criteria:

($\mathcal{F}$-*i*) φ_{n,η_n} is given by (11) with $I_n \neq \emptyset$;

($\mathcal{F}$-*ii*) for $n \geq 1$, having parameters satisfying the relation $\mu_{n-1} + \eta_{n-1} = R$ with μ_{n-1} positive, η_n and μ_n are defined by (12)-(14);

($\mathcal{F}$-*iii*) each index set is chosen to satisfy $I_n \supseteq \{0, n\}$; and

($\mathcal{F}$-*iv*) $I_{n+1} \supseteq J_n^{act} := \left\{ i \in I_n : \varphi_{n,\eta_n}(x^{n+1}) = f_{\mathbf{f}}^i + \eta_n f_{\mathbf{q}}^i + \langle g_{\mathbf{f}}^i + g_{\mathbf{q}}^i, x^{n+1} - x^i \rangle \right\}$.

We now show how these properties combine to yield satisfaction of conditions (3a) to (3d).

Proposition 11 *Consider a family of parameters and model functions $\mathcal{F} := \{\mu_n, \eta_n, \varphi_{n,\eta_n}\}_n$ defining iterates x^{n+1} by* (2)*. At iteration n the following hold:*

(*i*) *If the family satisfies ($\mathcal{F}$-i), then $\{\varphi_{n,\eta_n}\}_n$ satisfies condition* (3a)*.*

(*ii*) *If for a fixed convexification growth parameter $\Gamma > 1$ the family satisfies ($\mathcal{F}$-ii) and ($\mathcal{F}$-iii), and the prox-parameter R is large enough for the lower bound defined in* (12) *to satisfy $\tilde{\eta}_n < R/\Gamma$, then $\eta_n < R$ and $\varphi_{n,\eta_n}(x^j) \leq f_{\mathbf{f}}^j + \eta_n f_{\mathbf{q}}^j$ for all $j \in I_n$. Therefore, $\{\varphi_{n,\eta_n}\}_n$ satisfies condition* (3b)*.*

(*iii*) *If $\eta_{n+1} = \eta_n$, and the family satisfies ($\mathcal{F}$-i) and ($\mathcal{F}$-iv), then $\{\varphi_{n,\eta_n}\}_n$ satisfies condition* (3c)*.*

(*iv*) *If the family satisfies ($\mathcal{F}$-i) and ($\mathcal{F}$-iii), then $\{\varphi_{n,\eta_n}\}_n$ satisfies condition* (3d)*.*

Proof. **[Item (i)]** By ($\mathcal{F}$-*i*), φ_{n,η_n} is defined in (11) as the pointwise maximum of a finite collection of affine functions, therefore convex, and item (*i*) holds.

[Item (ii)] If $\tilde{\eta}_n < R/\Gamma$, by Proposition 8, $R > \eta_n \geq \max\{\eta_{n-1}, \tilde{\eta}_n\}$. In particular, $\eta_n \geq \tilde{\eta}_n$, so $\varphi_{n,\eta_n}(x^j) \leq f_{\mathbf{f}}^j + \eta_n f_{\mathbf{q}}^j$ for all $j \in I_n$, by Lemma 7. Condition (3b) follows from applying this inequality at $j = 0$ (because $0 \in I_n$ by *($\mathcal{F}$-iii)*), recalling that $f_{\mathbf{q}}^0 = 0$: $\varphi_{n,\eta_n}(x^0) \leq f_{\mathbf{f}}^0 + \eta_n f_{\mathbf{q}}^0 = f_{\mathbf{f}}^0$.

[Item (iii)] Suppose that $\eta_{n+1} = \eta_n$ and the family satisfies *($\mathcal{F}$-i)* and *($\mathcal{F}$-iv)*. Writing the optimality conditions of equation (2), we see there exists an optimal nonnegative multiplier $\bar{\alpha} \in \mathbb{R}^{|I_n|}$ such that

$$\bar{\alpha}_j > 0 \text{ for all } j \in J_n^{act} \quad \text{and} \quad \sum_{i \in \mathbb{R}^{|I_n|}} \bar{\alpha}_i = 1$$

$$\varphi_{n,\eta_n}(x^{n+1}) = \sum_{i \in J_n^{act}} \bar{\alpha}_i \left(f_{\mathbf{f}}^i + \eta_n f_{\mathbf{q}}^i + \langle g_{\mathbf{f}}^i + \eta_n g_{\mathbf{q}}^i, x^{n+1} - x^i \rangle \right)$$

$$s^{n+1} = \mu_n(x^0 - x^{n+1}) = \sum_{i \in J_n^{act}} \bar{\alpha}_i \left(g_{\mathbf{f}}^i + \eta_n g_{\mathbf{q}}^i \right).$$

It follows that

$$\varphi_{n,\eta_n}(x^{n+1}) = \sum_{i \in J_n^{act}} \bar{\alpha}_i \left(f_{\mathbf{f}}^i + \eta_n f_{\mathbf{q}}^i \right) - \sum_{i \in J_n^{act}} \bar{\alpha}_i \langle g_{\mathbf{f}}^i + \eta_n g_{\mathbf{q}}^i, x^i \rangle + \langle s^{n+1}, x^{n+1} \rangle.$$

By ($\mathcal{F}$-*i*), using (11) with n replaced by $n+1$, for all $w \in \mathbb{R}^N$ and each $i \in I_{n+1}$ $\varphi_{n+1,\eta_{n+1}}(w) \geq f_{\mathbf{f}}^i + \eta_{n+1} f_{\mathbf{q}}^i + \langle g_{\mathbf{f}}^i + \eta_{n+1} g_{\mathbf{q}}^i, w - x^i \rangle$. As a result, writing the inequality for the convex sum of indices

$j \in J_n^{act}$ and using that $\eta_{n+1} = \eta_n$ we see that

$$\begin{aligned}
\varphi_{n+1,\eta_{n+1}}(w) &= \sum_{i\in J_n^{act}} \bar{\alpha}_i \varphi_{n+1,\eta_{n+1}}(w) \\
&\geq \sum_{i\in J_n^{act}} \bar{\alpha}_i \left(f_{\mathbf{f}}^i + \eta_{n+1} f_{\mathbf{q}}^i\right) + \sum_{i\in J_n^{act}} \bar{\alpha}_i \langle g_{\mathbf{f}}^i + \eta_{n+1} g_{\mathbf{q}}^i, w - x^i\rangle \\
&= \sum_{i\in J_n^{act}} \bar{\alpha}_i \left(f_{\mathbf{f}}^i + \eta_n f_{\mathbf{q}}^i\right) + \sum_{i\in J_n^{act}} \bar{\alpha}_i \langle g_{\mathbf{f}}^i + \eta_n g_{\mathbf{q}}^i, w - x^i\rangle \\
&= \varphi_{n,\eta_n}(x^{n+1}) - \langle s^{n+1}, x^{n+1}\rangle + \langle s^{n+1}, w\rangle,
\end{aligned}$$

which is condition (3c).
[Item (iv)] Setting $i = n$ in (11) ($n \in I_n$ because of *(F-iii)*) yields condition (3d). □

If the function $f + \eta_n \frac{1}{2}|\cdot - x^0|^2$ ever becomes (locally) convex, the application of rule (12)-(14) in *(F-ii)* leaves the subsequent parameters unchanged.

Corollary 12 *Consider a family of parameters and model functions $\mathcal{F} = \{\mu_n, \eta_n, \varphi_{n,\eta_n}\}_n$ chosen to satisfy (F-i)-(F-iv). If at some iteration n the convexification parameter $\eta_n < R$ is such that the function $f + \eta_n \frac{1}{2}|\cdot - x^0|^2$ is convex on a ball containing $\{x^i : i \in I_n\}$, then (F-ii) reduces to:*
(F-ii) $\eta_{n+1} = \eta_n$ and $\mu_{n+1} = R - \eta_n = \mu_n$.*

Proof. This follows easily from Lemma 10. □

When $f + \eta_n \frac{1}{2}|\cdot - x^0|^2$ is nonconvex, the model functions may not define a cutting-planes model for the function f. When f is a convex function, from Corollary 12 we see that ($\mathcal{F}$-*ii*) will always set $\eta_n = \eta_0 (= 0)$ and $\mu_n = \mu_0 (= R)$. Since in this case the quadratic bundle set $\mathcal{B}\mathbf{q}_n$ is just the zero vector, our method becomes a standard cutting-planes approximation for estimating the proximal point, as in [1], [9].

In the case when f is only locally convex, the application of Corollary 12 requires some knowledge of the region where the iterates lie. In the next lemma we see that for lower-$\mathcal{C}^2$ functions, by carefully selecting the prox-parameter R, it is possible to keep each new iterate in a ball around the prox-center.

Lemma 13 (lower-$\mathcal{C}^2$ and iteration distance) *Suppose the function f is lower-$\mathcal{C}^2$ on V and $x^0 \in V$. Let $\varepsilon > 0$, $K > 0$, and $\rho > 0$ be given by Lemma 2 written with $\bar{x} = x^0$.*

Consider a family of parameters and model functions $\mathcal{F} := \{\mu_n, \eta_n, \varphi_{n,\eta_n}\}_n$ defining iterates x^{n+1} by (2). If at iteration n

$$\{x^i : i \in I_n\} \subseteq B_\varepsilon(x^0) \quad \text{and} \quad R > \frac{2K}{\varepsilon} + 3\eta_n,$$

then the next iterate x^{n+1} generated by equation (2) also lies within $B_\varepsilon(x^0)$.

In particular, if $\Gamma > 1$ is a fixed convexification growth parameter, $\eta_0 = 0$, and $R > \frac{2K}{\varepsilon} + 3\Gamma\rho$, then all iterates lie within $B_\varepsilon(x^0)$.

Proof: Recall that, by definition (11),

$$\varphi_{n,\eta_n}(w) = \max_{i\in I_n}\{f_{\mathbf{f}}^i + \eta_n f_{\mathbf{q}}^i + \langle g_{\mathbf{f}}^i + \eta_n g_{\mathbf{q}}^i, w - x^i\rangle\},$$

with $x^{n+1} = p_{\mu_n}\varphi_{n,\eta_n}(x^0)$, by equation (2). In order to obtain a bound for $|x^{n+1} - x^0|$ via Lemma 3, we start by showing that φ_{n,η_n} is Lipschitz continuous on $\mathbb{R}^N$. Since, by assumption, $\{x^i : i \in I_n\} \subseteq B_\varepsilon(x^0)$, and by Lemma 2, f is Lipschitz continuous on $\bar{B}_\varepsilon(x^0)$, any subgradient $g_{\mathbf{f}}^i \in \partial f(x^i)$ satisfies $|g_{\mathbf{f}}^i| \leq K$. Together with the bound $|g_{\mathbf{q}}^i| \leq \varepsilon$, we obtain for φ_{n,η_n} the Lipschitz constant

$$L = \max_{i\in I_n}\{|g_{\mathbf{f}}^i + \eta_n g_{\mathbf{q}}^i|\} \leq \max_{i\in I_n}\{|g_{\mathbf{f}}^i|\} + \varepsilon\eta_n \leq K + \varepsilon\eta_n.$$

Lemma 3 written with $f = \varphi_{n,\eta_n}$ and our assumptions yield the desired result:

$$|x^{n+1} - x^0| \leq \frac{2L}{\mu_n} \leq \frac{2(K + \varepsilon\eta_n)}{R - \eta_n} < \varepsilon.$$

We prove the final statement of the lemma by induction on n. Given $R > 2K/\varepsilon + 3\Gamma\rho$, the statement clearly holds for $n = 0$. If $\{x^i : i \in I_n\} \subseteq B_\varepsilon(x^0)$ and $\eta_n \leq \Gamma\rho$, we have just shown that $x^{n+1} \in B_\varepsilon(x^0)$. Then, by Lemma 10, $\tilde{\eta}_n \leq \rho$ and hence, by rule (14), either $\eta_{n+1} = \eta_n \leq \Gamma\rho$, or $\eta_{n+1} = \Gamma\tilde{\eta}_n \leq \Gamma\rho$. In both cases, $\eta_{n+1} \leq \Gamma\rho$, and the result follows. □

Before continuing to the an actual implementation of our algorithm, we provide one final discussion as to its convergence for lower-$\mathcal{C}^2$ functions.

Theorem 14 (lower-$\mathcal{C}^2$) *Suppose the lsc function f is prox-bounded with threshold r_{pb}. Let the prox-center $x^0 \in \mathbb{R}^N$ and prox-parameter $R > r_{pb}$ be given. Suppose the function f is lower-$\mathcal{C}^2$ on V and $x^0 \in V$. Let $\varepsilon > 0$, $K > 0$, $\rho > 0$, and R_{x^0} be given by Lemma 2 and Theorem 4 written with $\bar{x} = x^0$. For $n = 0, 1, \ldots$, consider the sequence $\{x^n\}$ defined by (2), where the family of parameters and model functions $\mathcal{F} = \{\mu_n, \eta_n, \varphi_{n,\eta_n}\}_n$ is chosen according to ($\mathcal{F}$-i)-($\mathcal{F}$-iv) and the convexification growth parameter $\Gamma > 1$ is fixed. Then the following holds:*

(*i*) *If $R > 2K/\varepsilon + 3\Gamma\rho$, then (3) holds and the sequence of approximal points converges to some point $p \in \mathbb{R}^N$.*

(*ii*) *If at any iteration, $\eta_m \geq \rho$ and $R > 4K/\varepsilon + 3\eta_m$, then for all $n \geq m$ the family satisfies conditions (3) and (6). In this case, the sequence of approximal points converges to $p_R f(x^0)$, and*

$$\lim_{n\to\infty} \left(\varphi_{n,\eta_n}(x^{n+1}) + \mu_n \frac{1}{2}|x^{n+1} - x^0|^2 \right) = f(p_R f(x^0)) + R\frac{1}{2}|p_R f(x^0) - x^0|^2.$$

Proof. In item (*i*) we assume $R > 2K/\varepsilon + 3\Gamma\rho$ and use Lemma 13 to ensure that the sequence $\{x^n\}$ lies in the open ball $B_\varepsilon(x^0)$. To see item (*i*), we begin by recalling that, because the function $f + \rho\frac{1}{2}|\cdot - x^0|^2$ is convex on $B_\varepsilon(x^0)$, the lower bound from (12) satisfies $\tilde{\eta}_n \leq \rho < R$ (by Lemma 10 written with $\hat{\eta} = \rho$). Hence, by items (*i*), (*ii*), and (*iv*) in Proposition 11, conditions (3a), (3b), and (3d) hold. To show satisfaction of conditions (3c) and (3e), we need only to show that eventually η_n becomes constant (condition (3c) would then follow from Proposition 11, item (*iii*)). Suppose, for contradiction purposes, that $\{\eta_n\}$ never stabilizes. Then, from some iteration m_1 on, the convexification parameters must become positive: $\eta_n \geq \eta_{m_1} > 0 (= \eta_0)$ for $n \geq m_1$. For the sequence not to stabilize, there must be an infinite number of iterations $n \geq m_1$ during which $\eta_n = \Gamma\tilde{\eta}_{n-1} > \Gamma\eta_{n-1}$. But this process can only be repeated a finite number of times, because, since $\Gamma > 1$, the update will eventually result in $\eta_n \geq \rho$. Let m_2 denote the first of such iterations. Since $\eta_{m_2} \geq \rho$, and the function $f + \eta_m\frac{1}{2}|\cdot - x^0|^2$ is convex, by Corollary 12, we have that $\eta_n = \eta_{m_2} =: \overline{\eta}$ for all subsequent iterations. Therefore, convexification parameters stabilize, and (3) holds. Then, by item (*ii*) in Lemma 5, the sequence of approximal points converges to some point $p \in \mathbb{R}^N$.
In item (*ii*) we assume $R > 4K/\varepsilon + 3\eta_m$ and use again Lemma 13. Since $\eta_m \geq \rho$ and x^{m+1} is in $B_\varepsilon(x^0)$, by Corollary 12 we have that $\eta_{m+1} = \eta_m \geq \rho$. By mathematical induction we therefore have $x^n \in B_\varepsilon(x^0)$ and, hence, $\eta_n = \eta_m$ for all $n \geq m$. Moreover, $f + \eta_m\frac{1}{2}|\cdot - x^0|^2$ is convex on $B_\varepsilon(x^0)$. An argument similar to the one in item (*i*) shows satisfaction of conditions (3) and (6), with (6) holding for all $w \in B_\varepsilon(x^0)$. Since lower-$\mathcal{C}^2$ functions are continuous, and $R > \max\{4K/\varepsilon, \eta_m, r_{pb}\} \geq \max\{4K/\varepsilon, \rho, r_{pb}\}$, Theorem 6 completes the proof. □

5 Algorithmic Implementation

In this section we discuss effective implementation of the algorithm outlined in Section 3 and analyze its convergence. We start with the important question of defining stopping tests.

5.1 Developing a stopping criterion

We return once more to the algorithm's original inspiration, i.e., considering that for some $\overline{\eta} > 0$ the function $f + \overline{\eta}\frac{1}{2}|\cdot - x^0|^2$ is somehow "more convex" than the original function. With this relation in mind, we reexamine the stopping test for computing proximal points of convex functions.

For a convex function h, cutting-planes model functions satisfy $\varphi_n \leq h$, thus the relation $R(x^0 - x^{n+1}) \in \partial\varphi_n(x^{n+1})$ implies that for all $w \in \mathbb{R}^N$

$$h(w) \geq \varphi_n(w) \geq h(x^{n+1}) + R\langle x^0 - x^{n+1}, w - x^{n+1}\rangle - (h(x^{n+1}) - \varphi_n(x^{n+1})).$$

In particular, for $w = p_R h(x^0)$, we obtain

$$h(p_R h(x^0)) \geq h(x^{n+1}) + R\langle x^{n+1} - x^0, x^{n+1} - p_R h(x^0)\rangle - (h(x^{n+1}) - \varphi_n(x^{n+1})).$$

Likewise, since $p_R h(x^0)$ is characterized by the inclusion $R(x^0 - p_R h(x^0)) \in \partial h(p_R h(x^0))$, the associated subgradient inequality written at x^{n+1} gives

$$h(x^{n+1}) \geq h(p_R h(x^0)) + R\langle x^0 - p_R h(x^0), x^{n+1} - p_R h(x^0)\rangle.$$

The two inequalities above combined yield

$$0 \geq R\langle x^{n+1} - p_R h(x^0), x^{n+1} - p_R h(x^0)\rangle - (h(x^{n+1}) - \varphi_n(x^{n+1})),$$

i.e., $R|x^{n+1} - p_R h(x^0)|^2 \leq h(x^{n+1}) - \varphi_n(x^{n+1})$. Therefore, testing if the approximal estimate satisfies

$$\frac{h(x^{n+1}) - \varphi_n(x^{n+1})}{R} < \text{TOL}_{\text{stop}}^2,$$

where TOL_{stop} is a stopping tolerance, can be used as a stopping criterion in the convex case.

For our nonconvex function f, we seek $p_R f(x^0)$ via $p_{\mu_n}(f + \eta_n\frac{1}{2}|\cdot - x^0|^2)$. If $f + \eta_n\frac{1}{2}|\cdot - x^0|^2)$ is convex on a ball containing the points of interest, then the convex stopping criterion, written with the notation of (8)-(9), becomes

$$\text{if } \frac{f_{\mathbf{f}}^{n+1} + \eta_n f_{\mathbf{q}}^{n+1} - \varphi_{n,\eta_n}(x^{n+1})}{R - \eta_n} \leq \text{TOL}_{\text{stop}}^2 \text{ then } |x^{n+1} - p_R f(x^0)| \leq \text{TOL}_{\text{stop}}. \tag{15}$$

In general, we have no way of knowing when or where $f + \eta_n\frac{1}{2}|\cdot - x^0|^2$ is convex, so this stopping criterion may lead to false positive outcomes. However, if f is a lower-$\mathcal{C}^2$ function, there exists some positive $\overline{\eta}$ such that $f + \overline{\eta}\frac{1}{2}|\cdot - x^0|^2$ is convex locally near the points of interest. Thus, a reasonable stopping condition might be that

$$\frac{f(x^{n+1}) + \eta\frac{1}{2}|x^{n+1} - x^0|^2 - \varphi_{n,\eta}(x^{n+1})}{R - \eta} \leq \text{TOL}_{\text{stop}}^2,$$

for some "big enough" $\eta \in [\eta_n, R)$. The next proposition shows that this stopping criterion is indeed stronger than the one in equation (15). We also analyze a second test, where $\varphi_{n,\eta}(x^{n+1})$ (an unknown value) is replaced by $\varphi_{n,\eta_n}(x^{n+1})$ (a computable quantity).

Theorem 15 (Stopping Criterion) *Suppose x^{n+1} is computed as the solution to (2) with model function given by (11). For any positive η consider model functions $\varphi_{n,\eta}$ defined by (11) written with η_n replaced by η. For the test function*

$$\eta \mapsto T_n(\eta) := f(x^{n+1}) + \eta\frac{1}{2}|x^{n+1} - x^0|^2 - \varphi_{n,\eta}(x^{n+1}),$$

the following holds:
(i) T_n is a positive increasing function for $\eta \geq \tilde{\eta}_n$, where $\tilde{\eta}_n$ is defined in (12).
(ii) $T_n(\eta) \leq f(x^{n+1}) + \eta\frac{1}{2}|x^{n+1} - x^0|^2 - \varphi_{n,\eta_n}(x^{n+1})$ for any $\eta \geq \eta_n$.

Proof. Following the notation of (8)-(9), since $\varphi_{n,\eta}$ is defined as in (11), using (10) with $j = n+1$, we obtain that

$$\begin{aligned}
T_n(\eta) &= f_{\mathbf{f}}^{n+1} + \eta f_{\mathbf{q}}^{n+1} - \varphi_{n,\eta}(x^{n+1}) \\
&= f_{\mathbf{f}}^{n+1} + \eta f_{\mathbf{q}}^{n+1} - \max_{i\in I_n}\{f_{\mathbf{f}}^i + \eta f_{\mathbf{q}}^i + \langle g_{\mathbf{f}}^i + \eta g_{\mathbf{q}}^i, x^{n+1} - x^i\rangle\} \\
&= \min_{i\in I_n}\left\{f_{\mathbf{f}}^{n+1} - f_{\mathbf{f}}^i - \langle g_{\mathbf{f}}^i, x^{n+1} - x^i\rangle + \eta(f_{\mathbf{q}}^{n+1} - f_{\mathbf{q}}^i - \langle g_{\mathbf{q}}^i, x^{n+1} - x^0\rangle)\right\} \\
&= \min_{i\in I_n}\left\{f_{\mathbf{f}}^{n+1} - f_{\mathbf{f}}^i - \langle g_{\mathbf{f}}^i, x^{n+1} - x^i\rangle + \eta\frac{1}{2}|x^{n+1} - x^i|^2\right\}.
\end{aligned}$$

By Lemma 7 (equation (12) with (i,j) therein replaced by $(n+1,i)$), the fact that $\eta \geq \tilde{\eta}$ implies that

$$\eta\frac{1}{2}|x^{n+1} - x^i|^2 \geq f_{\mathbf{f}}^i - f_{\mathbf{f}}^{n+1} - \langle g_{\mathbf{f}}^i, x^i - x^{n+1}\rangle.$$

Thus, $T_n(\eta)$ is positive and increasing for all $\eta \geq \tilde{\eta}_n$, and item (i) follows.
If $\eta \geq \eta_n$, then clearly $f_{\mathbf{f}}^{n+1} + \eta f_{\mathbf{q}}^{n+1} - \varphi_{n,\eta_n}(x^{n+1}) \geq f_{\mathbf{f}}^{n+1} + \eta_n f_{\mathbf{q}}^{n+1} - \varphi_{n,\eta_n}(x^{n+1})$. Utilizing equations (11) and (10) we see that

$$\begin{aligned}
f_{\mathbf{f}}^{n+1} + \eta f_{\mathbf{q}}^{n+1} - \varphi_{n,\eta_n}(x^{n+1}) &\geq f_{\mathbf{f}}^{n+1} + \eta_n f_{\mathbf{q}}^{n+1} - \varphi_{n,\eta_n}(x^{n+1}) \\
&= \min_{i\in I_n}\left\{\begin{array}{l} f_{\mathbf{f}}^{n+1} + \eta f_{\mathbf{q}}^{n+1} - f_{\mathbf{f}}^i - \eta f_{\mathbf{q}}^i - \langle g_{\mathbf{f}}^i + \eta g_{\mathbf{q}}^i, x^{n+1} - x^i\rangle \\ \\ +(\eta_n - \eta)(f_{\mathbf{q}}^{n+1} - f_{\mathbf{q}}^i - \langle g_{\mathbf{q}}^i, x^{n+1} - x^0\rangle) \end{array}\right\} \\
&\geq f_{\mathbf{f}}^{n+1} + \eta f_{\mathbf{q}}^{n+1} - \varphi_{n,\eta}(x^{n+1}) + \min_{i\in I_n}\{-\frac{1}{2}(\eta_n - \eta)|x^{n+1} - x^i|^2\} \\
&= T_n(\eta) + \min_{i\in I_n}\{\tfrac{1}{2}(\eta - \eta_n)|x^{n+1} - x^i|^2\}.
\end{aligned}$$

Since $\eta \geq \eta_n$, the quadratic term above is nonnegative, and item (ii) follows. □

Our algorithm uses in the stopping test a quotient whose numerator is the right hand side expression in Theorem 15(ii), written with $\eta = R - \text{TOL}_\mu$, for $\text{TOL}_\mu \in (0, R]$ a given tolerance. For the denominator, along the lines of (15), we use $R - \eta = \text{TOL}_\mu$.

5.2 The algorithm

We are now in position to formally define our algorithm for computing proximal points of nonconvex functions. It requires the choice of some parameters used for stopping tests, that can be activated if no significant progress is observed (parameters MIN_{length} and $\text{MAX}_{\texttt{short}}$), if μ_n becomes too small (parameter TOL_μ), or if convergence is achieved (parameter TOL_{stop}).

Algorithm 1 (CompProx) A prox-center $x^0 \in \mathbb{R}^N$ and a prox-parameter $R > 0$ are given, as well as a black box procedure evaluating, for each $x \in \mathbb{R}^N$, the function value $f(x)$ and one subgradient in $\partial f(x)$.

Step 0 **(Initialization)** Choose a convexification growth parameter $\Gamma > 1$, a minimum step length $\text{MIN}_{length} \geq 0$, a maximum number of allowed short steps $\text{MAX}_{\texttt{short}} \geq 0$, a μ-tolerance level $\text{TOL}_\mu \in (0, R]$, and a stopping tolerance $\text{TOL}_{\text{stop}} \geq 0$.
Compute the black box information $f_{\mathbf{f}}^0 := f(x^0)$ and $g_{\mathbf{f}}^0 \in \partial f(x^0)$. Let $f_{\mathbf{q}}^0 := 0$, $g_{\mathbf{q}}^0 := 0$, $I_0 := \{0\}$, $\mu_0 := R$ and $\eta_0 := 0$. Set $n = 0$ and the short steps counter $\texttt{short} = 0$.

Step 1 **(Quadratic Subproblem)** Having the bundle information $\mathcal{B}\mathtt{f}_n$ from (8) and $\eta_n < R$, construct the bundle $\mathcal{B}\mathtt{q}_n$ as in (9) and the model function φ_{n,η_n} from (11). For $\mu_n = R - \eta_n$, find

$$(r^{n+1}, x^{n+1}) \in \begin{cases} \operatorname{argmin}\{r + \mu_n \frac{1}{2}|w - x^0|^2 : (r, w) \in \mathbb{R} \times \mathbb{R}^N\} \\ s.t.\ \ r \geq f^i_{\mathbf{f}} + \eta_n f^i_{\mathbf{q}} + \langle g^i_{\mathbf{f}} + \eta_n g^i_{\mathbf{q}}, w - x^i\rangle \text{ for } i \in I_n. \end{cases}$$

Step 2 **(Update Bundle)** Compute the black box information $f^{n+1}_{\mathbf{f}} := f(x^{n+1})$ and $g^{n+1}_{\mathbf{f}} \in \partial f(x^{n+1})$. Choose a new index set $I_{n+1} \supset \{0, n+1\} \cup J^{act}_n$, as defined in ($\mathcal{F}$-$iv$).

Step 3 **(Update μ and η)** If $|x^i - x^{n+1}| < \text{MIN}_{length}$ for some $i \in I_n$, then set $\mu_{n+1} := \max\{\mu_n/2, \text{TOL}_\mu\}$, $\eta_{n+1} := R - \mu_{n+1}$, and increase `short` by 1; if `short` $> \text{MAX}_{\texttt{short}}$, stop with the message

"Too many steps without significant progress."

Otherwise, apply rule (12)-(14) written with n therein replaced by $n+1$ to compute $\tilde{\eta}_{n+1}$ and to set μ_{n+1} and η_{n+1} accordingly.

Step 4 **(Check μ-tolerance)** If $\mu_{n+1} < \text{TOL}_\mu$, stop with the message

"R is insufficient, require R greater than $\text{TOL}_\mu + \Gamma(R - \mu_{n+1})$"

Step 5 **(Stopping Criterion)** If $\mu_{n+1} = \mu_n$, check the stopping condition

$$(0 \leq) \frac{f^{n+1}_{\mathbf{f}} + (R - \text{TOL}_\mu) f^{n+1}_{\mathbf{q}} - r^{n+1}}{\text{TOL}_\mu} \leq \text{TOL}^2_{\text{stop}}.$$

If this condition is true, stop with the message

"Convergence criterion activated"

and output x^{n+1} as the final approximal point.

Otherwise, increase n by 1 and loop to Step 1.

We already mentioned that Step 1 of CompProx is easily accomplished by a quadratic programming solver. In Step 3, we start with a test checking significant change with respect to past iterates, to avoid floating point errors in the calculation of the lower bound $\tilde{\eta}_{n+1}$. If too many "short steps" are taken, we stop the algorithm. In practice, we observed the beneficial effect of this test in the form of faster convergence. Step 4 eliminates instances for which the given prox-parameter fails to ensure existence of a well defined proximal point. In Step 5 we use the criterion in item (ii) of Theorem 15, for $\eta := R - \text{TOL}_\mu$. Note that if $\eta_n = R - \text{TOL}_\mu$, this test is just the stopping criterion for finding proximal points for convex functions (equation (15)). Finally, note that in Step 5 the stopping criterion is only checked if $\mu_{n+1} = \mu_n$. By examining Step 3, we then see that the stopping test is only performed under two circumstances:

either

$$\exists i \in I_n \text{ such that } |x^i - x^{n+1}| < \text{MIN}_{length}, \text{ with } \mu_{n+1} = \mu_n = \text{TOL}_\mu, \tag{16a}$$

or

$$\forall i \in I_n \quad |x^i - x^{n+1}| \geq \text{MIN}_{length}, \text{ with } \eta_{n+1} = \eta_n, \text{ because } \tilde{\eta}_{n+1} \leq \eta_n \text{ in (14)}. \tag{16b}$$

We now show that, for a suitable choice of parameters, algorithm CompProx either loops forever converging to the desired proximal point, or it terminates having found the proximal point.

Theorem 16 (Convergence under lower-$\mathcal{C}^2$) *Suppose the lsc function f is prox-bounded with threshold r_{pb} and lower-$\mathcal{C}^2$ on V. Let the prox-center $x^0 \in V$ and prox-parameter $R > r_{pb}$ be given and let $\varepsilon > 0$, $K > 0$, and $\rho > 0$ be given by Lemma 2 written with $\bar{x} = x^0$. Consider Algorithm 1, with parameters $\Gamma > 1$, $\text{MAX}_{\texttt{short}} = +\infty$, $\text{MIN}_{length} > 0$, $\text{TOL}_\mu \in (0, R]$, and $\text{TOL}_{\text{stop}} = 0$. If*

$$R > \frac{4K}{\varepsilon} + 3\Gamma\rho \quad \textit{and} \quad R \geq \text{TOL}_\mu + \Gamma\rho,$$

then the algorithm either terminates at $p_R f(x^0)$ or generates an infinite sequence $\{x^n\}$ converging to $p_R f(x^0)$ with

$$\lim_{n\to\infty} \left(\varphi_{n,\eta_n}(x^{n+1}) + \mu_n \frac{1}{2}|x^{n+1} - x^0|^2 \right) = f(p_R f(x^0)) + R\frac{1}{2}|p_R f(x^0) - x^0|^2.$$

Proof. We begin by noting that, as $R > \frac{2K}{\varepsilon}+3\Gamma\rho$, Lemma 13 implies all iterates created via CompProx will lie within $B_\varepsilon(x^0)$, an open ball in which $f+\rho\frac{1}{2}|\cdot - x^0|^2$ is convex.

[Part I: finitely many approximal points] Consider the case of finite termination of the algorithm. Algorithm CompProx can exit at Step 3 by $\text{MAX}_{\texttt{short}}$, at Step 4 by TOL_μ, or at Step 5 by TOL_{stop}. We start by showing that with our choice of stopping parameters, the only possible exit is at Step 5.

Since $\text{MAX}_{\texttt{short}} = +\infty$, the first stopping test is never active.

Considering the exit at Step 4, we note that μ_n can be decreased in two ways. First, if $|x^i - x^{n+1}| < \text{MIN}_{length}$ for some $i \in I_n$, which results in $\mu_{n+1} = \max\{\mu_n/2, \text{TOL}_\mu\}$; an update that will never result in exiting at Step 4. Second, if $|x^i - x^{n+1}| \geq \text{MIN}_{length}$ for all $i \in I_n$, and $\tilde{\eta}_{n+1} > \eta_n$, which results in $\mu_{n+1} = R - \Gamma\tilde{\eta}_n$. If $R - \Gamma\tilde{\eta}_n < \text{TOL}_\mu$ then $R < \text{TOL}_\mu + \Gamma\tilde{\eta}_n < \text{TOL}_\mu + \Gamma\rho$, by Lemma 10. This contradicts our assumption of the size of R, therefore the algorithm cannot exit at Step 4.

If CompProx exits at Step 5, then $\mu_n \geq \text{TOL}_\mu$, because the algorithm did not exit at Step 4. Hence, $R - \text{TOL}_\mu \geq \eta_n$, and by Theorem 15 (i), using the fact that $\text{TOL}_{\text{stop}} = 0$,

$$0 \leq f_{\mathbf{f}}^{n+1} + (R - \text{TOL}_\mu) f_{\mathbf{q}}^{n+1} - r^{n+1} \leq 0.$$

By (16), the stopping test is only checked if either (16a) or (16b) hold. Suppose first that case (b) holds. By using the fact that $R - \text{TOL}_\mu \geq \eta_n$, we obtain the relations $\tilde{\eta}_{n+1} \leq \eta_n \leq R - \text{TOL}_\mu$. By Theorem 15, this implies that

$$0 \leq f_{\mathbf{f}}^{n+1} + \eta_n f_{\mathbf{q}}^{n+1} - r^{n+1} \leq f_{\mathbf{f}}^{n+1} + (R - \text{TOL}_\mu) f_{\mathbf{q}}^{n+1} - r^{n+1} \leq 0.$$

Thus, all the inequalities are in fact equal to zero. Therefore

$$0 = \left(f_{\mathbf{f}}^{n+1} + \eta_n f_{\mathbf{q}}^{n+1} - r^{n+1}\right) - \left(f_{\mathbf{f}}^{n+1} + (R - \text{TOL}_\mu) f_{\mathbf{q}}^{n+1} - r^{n+1}\right) = (\eta_n - R + \text{TOL}_\mu)\frac{1}{2}|x^{n+1} - x^0|^2.$$

Since by (16b), $x^{n+1} \neq x^0$, the convexification parameter must be $\eta_n = R - \text{TOL}_\mu$. The remaining case for the stopping test, i.e. (16a), also gives $R - \text{TOL}_\mu = R - \mu_n = \eta_n$ so we have in both cases that $\eta_n = R - \text{TOL}_\mu \geq \Gamma\rho > \rho$. It follows that the function $f + \eta_n\frac{1}{2}|\cdot - x^0|^2$ is convex on $B_\varepsilon(x^0)$. By equation (15), if the stopping criterion with $\text{TOL}_{\text{stop}} = 0$ and $f + \eta_n\frac{1}{2}|\cdot - x^0|^2$ convex on an open ball containing all the points of interest, is successful, then $|x^{n+1} - p_R f(x^0)| = 0$. That is, the algorithm has terminated with x^{n+1} being the proximal point of x^0, $p_R f(x^0)$.

[Part II: infinitely many approximal points] Consider the case of the algorithm generating an infinite sequence of approximal points.

We begin by showing that condition (3e) holds. By construction, if $\mu_{n+1} \neq \mu_n$ then either $\mu_{n+1} = \mu_n/2$ or $\eta_{n+1} = \Gamma\tilde{\eta}_n$. The first of these cannot happen an infinite number of times, as then eventually $\mu_n/2$ will be less than TOL_μ. After this the line $\mu_{n+1} = \min\{\mu_n/2, \text{TOL}_\mu\}$ will return $\mu_{n+1} = \text{TOL}_\mu$. The second event, $\eta_{n+1} = \Gamma\tilde{\eta}_n$, also cannot happen an infinite number of times. Indeed, after the first time this occurs we must have $\eta_{m+1} = \Gamma\tilde{\eta}_n > \Gamma\eta_n \geq 0$. Each subsequent time therefore multiplies a increasing positive number by a factor of Γ. As $\Gamma > 1$, an infinite number of steps of this type would imply $\eta_n \to \infty$. But as soon as $\eta_n > \rho$, Lemma 10 tells us that rule (12)-(14) will leave η_n unchanged. Therefore, condition (3e) holds.

We now show that the family of parameters and model functions satisfies conditions *($\mathcal{F}$-i)-($\mathcal{F}$-iv)*. By construction in CompProx, the family satisfies *($\mathcal{F}$-i)*, *($\mathcal{F}$-iii)*, and *($\mathcal{F}$-iv)*. To show satisfaction of *($\mathcal{F}$-ii)*, we examine two possible stabilization values for the sequence of model prox-parameters: $\underline{\mu} = \text{TOL}_\mu$ and $\underline{\mu} > \text{TOL}_\mu$. If $\{\mu_n\}$ stabilizes at TOL_μ, then $\{\eta_n\}$ stabilizes at $R - \text{TOL}_\mu$. Since $R - \text{TOL}_\mu \geq \Gamma\rho > \rho$, at some iteration m the function $f + \eta_m|\cdot - x^0|^2$ is convex on $B_\varepsilon(x^0)$ and, by Corollary 12, condition *($\mathcal{F}$-ii)* reduces to $\eta_m = \eta_n$ for all $n \geq m$, a condition that clearly holds. If $\{\mu_n\}$ stabilizes at some value greater than TOL_μ, then Step 3 must never find $|x^i - x^{n+1}| < \text{MIN}_{length}$ for some $i \in I_n$, so *($\mathcal{F}$-ii)* is always used.

Hence, as soon as $\{\mu_n\}$ stabilizes, the family generated by algorithm CompProx satisfies *($\mathcal{F}$-i)-($\mathcal{F}$-iv)*. Proposition 11 yields the remaining conditions in (3) and, therefore, by Lemma 5 (ii), the sequence of approximal points converges. But such convergence implies in particular that Step 3 must eventually find some $i \in I_n$ for which $|x^i - x^{n+1}| < \text{MIN}_{length}$ (specifically, $|x^n - x^{n+1}|$ must eventually be less than MIN_{length}). Thus, the only possible value for the infinite sequence $\{\mu_n\}$ to stabilize is at $\underline{\mu} = \text{TOL}_\mu$.

Noting that $R - \text{TOL}_\mu \geq \rho$ and applying Theorem 14 (ii) completes the proof. □

Example 17 (Semi-convex and convex functions) Suppose the function f is semi-convex, in the sense that there exists $\rho > 0$ such that the function $f_\rho := f + \rho\frac{1}{2}|\cdot - x^0|^2$ is globally convex, then Theorem 16 holds whenever $R \geq \text{TOL}_\mu + \Gamma\rho$ (i.e. the conditions $R > r_{pb}$ and $R > 4K/\varepsilon + 3\Gamma\rho$ may be removed from the statement of the theorem).
To see this, note first that any semi-convex function f is lower-$\mathcal{C}^2$ on $V = \mathbb{R}^N$. Furthermore, ρ corresponds to the parameter ρ used in Lemma 2, while ε therein can be taken to be $\varepsilon = \infty$. The Lipschitz constant, K, is unfortunately undeterminable, but its value is not necessary for the case under consideration. More precisely, the proofs of Theorems 4, 6 and 14 only make use of K (and of ε) to ensure that all iterates remain in $B_\varepsilon(x^0)$. Since ρ makes the function f_ρ globally convex, the corresponding portions of the proofs may be skipped for semi-convex functions. That is, for semi-convex functions, the lower bound of $R > \max\{4K/\varepsilon, \rho, r_{pb}\}$ in Theorems 4 and 6 may be reduced to $R > \max\{\rho, r_{pb}\}$, while the lower bound of $R > 4K/\varepsilon + 3\Gamma\rho$ may be removed from the statement of Theorem 14. Finally, as f_ρ is convex, we know $r_{pb} \leq \rho$, hence the condition $R \geq \text{TOL}_\mu + \Gamma\rho$ is sufficient to allow the application of Theorems 4, 6 and 14 in the proof of Theorem 16.
If the function f is convex, it is semi-convex with parameter $\rho = 0$. Hence, the lower bound established above for R further reduces to $R \geq \text{TOL}_\mu$, which is also required at the initialization step of the algorithm.

Remark 18 As mentioned in Subsection 1.1, consider CompProx as a sub-algorithm generating serious steps of a bundle nonconvex method. Suppose x^0 is the last serious step generated by such bundle method:

$$f(x^{n+1}) > f(x^0) - m(f(x^0) + \overline{\eta}\frac{1}{2}|x^{n+1} - x^0|^2 - \varphi_{n,\eta_n}(x^{n+1})) \quad \text{for all } n$$

and a given parameter $m \in]0,1[$. The result in Theorem 16 shows that the sequence of null steps ($\{x^{n+1}\}$ in our setting) converges to $p_R f(x^0)$. Without entering into further details, the inequality above then implies that $x^0 = p_R f(x^0)$. Hence, not only x^0 is a stationary point, but the sequence of null steps converges to such stationary point. To the best of our knowledge, this last result is new in the nonconvex bundle literature.

6 Numerical Results

For our runs we work with a particular subclass of lower-$\mathcal{C}^2$ functions, that are piecewise quadratic and semi-convex, as defined in Example 17 and described below.

6.1 Test-functions and starting points

We consider functions defined as the pointwise maximum of a finite collection of quadratic functions:

$$f(x) := \max\{\langle x, A_i x\rangle + \langle B_i, x\rangle + C_i : i = 1, 2, ..., \texttt{nf}\}, \tag{17}$$

where A_i are $N \times N$ symmetrical matrices (no assumption on their positive definiteness is made), $B_i \in \mathbb{R}^N$, and $C_i \in \mathbb{R}$.

This class of functions has important practical advantages. First, many different examples are easily available, just by choosing values for N and $\texttt{nf}$, and then randomly generating $\texttt{nf}$ objects A_i, B_i and C_i. Second, black box oracles are easy to define, because for each given x any *active* index $j \leq \texttt{nf}$ (i.e., an index where the maximum in (17) is attained), yields a subgradient $A_j x + B_j \in \partial f(x)$. In particular, if there are $\texttt{nf}_{\texttt{act}}$ active indices at 0, the following relations hold (reordering indices if necessary):

$$f(0) = C_i = C \text{ for } i = 1, 2, ...\texttt{nf}_{\texttt{act}} \quad \text{and} \quad \partial f(0) = \text{conv}\{B_i : i = 1, \ldots, \texttt{nf}_{\texttt{act}}\}. \tag{18}$$

Furthermore, a function given by (17) is always lower-$\mathcal{C}^2$, prox-bounded, and prox-regular. Indeed such a function is actually semiconvex, prox-bounded, and prox-regular with identical thresholds and parameter:

$$\rho = r_{pb} = r_{pr} = \max\{|A_i| : i = 1, 2, ...\texttt{nf}\}.$$

Thus, a "large enough" prox-parameter R can be estimated a priori, by taking any value bigger than the norm of all matrices A_i.

To assess our method, in all our runs we fix the desired proximal point to be the zero vector, and choose appropriate starting points x^0 so that $p_R f(x^0) = 0$, where f is given by (17). We essentially proceed as follows:

– We choose a dimension N, the number of quadratic functions defining the maximum $\mathtt{nf}$, and the number of active subgradients $\mathtt{nf}_{\mathtt{act}} \leq \mathtt{nf}$ at $p_R f(x^0) = 0 \in \mathbb{R}^N$.
– Having selected a real number C, we set $C_1 = C_2 = \ldots = C_{\mathtt{nf}_{\mathtt{act}}} = C$ and take arbitrary $C_i < C$ for $i = \mathtt{nf}_{\mathtt{act}} + 1, \ldots, \mathtt{nf}$.
– We choose scalar upper and lower bounds and generate randomly numbers in this range to define A_i and B_i, for all $i \leq \mathtt{nf}$. We then take

$$R := 12\text{ceiling}\,[\,\max\{|A_i| : i = 1, \ldots, \mathtt{nf}\}\,] + 1.$$

– Finally, we use the characterization for proximal points given in Theorem 4 together with the expression for $\partial f(0)$ in (18):

$$p = p_R f(x^0) \iff R(x^0 - p) \in \partial f(p) \quad \text{i.e.,} \quad p = 0 \iff Rx^0 \in \partial f(0)\,,$$

to take any $x^0 \in \frac{1}{R}\text{conv}\{B_i : i = 1, \ldots, \mathtt{nf}_{\mathtt{act}}\}$ as starting point.

Table 1 shows relevant data for the 240 functions we randomly generated. Each table entry represents 20 functions with the same dimension (N), same number of quadratic functions defining the maximum ($\mathtt{nf}$), and same number of active indices at 0 ($\mathtt{nf}_{\mathtt{act}}$). For each dimension we generated two special instances, one grouping only convex functions (all matrices A_i are taken positive definite), and another with only nonconvex functions (all matrices are negative definite). These special instances are listed in the table with ♮ and †, respectively.

N	$\mathtt{nf}$	$\mathtt{nf}_{\mathtt{act}}$	numbers range	N	$\mathtt{nf}$	$\mathtt{nf}_{\mathtt{act}}$	numbers range
7	5	1	$[-10, 10]^{\dagger}$	11	9	1	$[-10, 0]$
7	5	3	[-10,10]	11	9	5	[-100,100]
7	5	5	[0,10]	11	9	9	$[-10, 10]^{\natural}$
7	10	1	$[-10, 10]^{\natural}$	11	18	1	[0,10]
7	10	5	[-100,100]	11	18	9	[-10,10]
7	10	10	[-10,0]	11	18	18	$[-10, 10]^{\dagger}$

Table 1 Test Problem Set († = nonconvex, ♮=convex)

6.2 Numerical testing

We proceed along three lines:

– Determination of "good" values for the user defined options in the algorithm. These options take two forms: the choice of the new index set in Step 2 of the algorithm (directly related to the bundle size), and the selection of values for the convexification growth parameter Γ and the stopping parameters MIN_{length}, $\text{MAX}_{\mathtt{short}}$, TOL_{μ}, and TOL_{stop}.
– Behaviour of the algorithm for various types of problems in various dimensions.
– Comparison with two bundle algorithms for nonsmooth optimization.

In order to ease the analysis of our results, we report them both in tables showing average values and in graphs with performance profiles. Performance profiles plot *the ratio of black box calls used by a given solver to number of black box calls used by the best solver* against *the portion of tests successfully completed*. The solvers which most rapidly reach the value of one are considered the most successful. Since our interest lies in solvers which quickly reach the value of one, we generally focus our performance profiles on the area where 30-35 percent of the tests have been successfully completed. We refer to [10] for more detail of performance profiles.

6.3 Choosing index sets

At Step 2 in the algorithm, the new index set must satisfy the inclusion $I_{n+1} \supset \{0, n+1\} \cup J_n^{act}$, but it is up to the user when to delete other elements in $\{0, 1, \ldots, n+1\}$, and which ones to discard. For example, the user might decide to always delete nonactive bundle elements, or to only delete bundle elements which have been nonactive for several iterations (say, 4, 8, or 16 iterations), or even to never delete any bundle element, regardless of their *inactivity*. We consider all these strategies and examine their effect with respect to convergence speed.

We set $\Gamma = 2$, $\text{MIN}_{length} = 10^{-8}$, $\text{MAX}_{\texttt{short}} = \infty$, and $\text{TOL}_\mu = 9/12R$, and ran the algorithm on the 240 test problems of Table 1 until the black box call returns a point x^n approximated the correct proximal point, $p_R f(x^0) = 0$, with six more digits of accuracy, (i.e. $|x^n| \leq 10^{-6}|x^0|$).

Figure 3 plots *the time in seconds elapsed* against *the ratio of tests successfully completed* for each of the four runs (all tests successfully completed). Table 2 provides statistical information from these tests in 7 and 11 dimensions, where the left (right) figure in each column corresponds to dimension 7 (11). Our results indicate the clear superiority of the "minimal" strategy, which sets $I_{n+1} := \{0, n+1\} \cup J_n^{act}$. Although a slight decrease in the number of black box calls is achieved by storing bundle information longer, this gain is outweighed by the increase in time each iteration takes to complete.

For this reason, in our subsequent tests we only use the minimal strategy.

Clean after # iterations	mean time in seconds	maximum time in seconds	mean black box calls
1	0.070, 0.131	0.20, 0.45	26.24, 34.51
4	0.085, 0.155	0.22, 0.58	26.48, 34.76
8	0.101, 0.190	0.28, 0.69	26.25, 33.44
16	0.133, 0.257	0.45, 1.05	26.27, 33.20
∞	0.190, 0.632	1.83, 10.47	26.24, 33.71

Table 2 Bundle cleaning test results *(dimension 7, dimension 11)*

6.4 Choosing CompProx parameters

Since "good" values for TOL_{stop} depend on the intended purpose of the algorithm's execution, we do not examine this parameter, instead setting it to $\text{TOL}_{\text{stop}} = 10^{-6}|x^0|$ (thus seeking an improvement of 10^{-6} in the accuracy of our solution). We consider different values for the remaining parameters, Γ, MIN_{length}, $\text{MAX}_{\texttt{short}}$, TOL_μ.

A run is considered successful if:

(i) The algorithm exited in either Step 3 or Step 5, and

(ii) the final point found was within 10^{-6} of the zero vector, the correct proximal point.

Since our choice of R should never be "insufficient", if the algorithm exits in Step 4, we consider the run a failure.

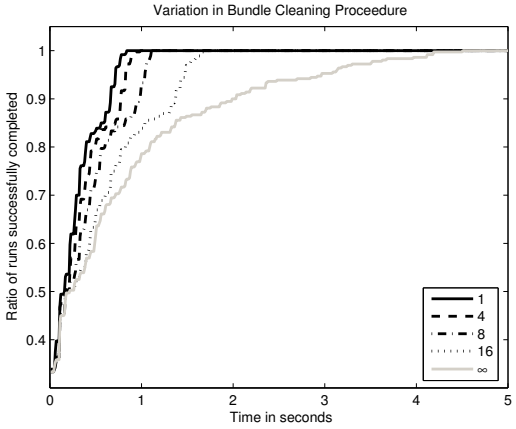

Fig. 3 Bundle cleaning and convergence speed

6.4.1 Convexification growth parameter

For these tests we set $\text{TOL}_{\text{stop}} = 10^{-6}|x^0|$, $\text{MIN}_{length} = 5$, $\text{MAX}_{\texttt{short}} = 10^{-8}$, and $\text{TOL}_\mu = 9/12R$, and make our runs for $\Gamma \in \{1.01, 2, 4, 8, 16\}$. In Table 4 we provide statistical results for these tests, while Figure 5 displays a performance profile for each of the options tested. According to the development in Example 17, any value of Γ in $(1, (R - \text{TOL}_\mu)/\rho]$ should guarantee convergence of the algorithm. In our case, $\rho \approx 1/12R$ and $\text{TOL}_\mu = 9/12R$, hence we expect $\Gamma \in (1, 3]$ to converge. Our numerical results support this expectation. It also appears that when Γ is set to 4, 8, or 16, the algorithm converges "most" of the times. In dimension 11, the higher values of Γ fail more often. This observation made us believe that the convergence for $\Gamma > 3$ is largely due to luck and small dimension size. When convergence is successful, it would appear that the convexification growth parameter has little effect on the rate of algorithmic convergence (notice in Figure 5 that, regardless of the value of Γ, approximately 80% of the runs were successfully concluded in the same amount of time, while the performance profiles for $\Gamma = 1.01$ and 2 are essentially identical).

It should also be noted that, when failures in convergence occur, they tend to occur rapidly. These failures are a result of the algorithm reporting that "R is insufficient," which suggests that the algorithm detects the minimal required value of R very efficiently.

Γ	successes	mean black box calls	failures	mean black box calls
1.01	120,120	30.00, 37.63	0, 0	NA, NA
2.00	120,120	29.52, 37.15	0, 0	NA, NA
4.00	120,117	29.33, 38.18	0, 3	NA, 1.00
8.00	114, 99	29.59, 42.41	6, 21	4.67, 1.57
16.0	103, 96	30.17, 42.88	17, 24	5.35, 2.75

Table 4 Convexification growth parameter test results, *(dimension 7, dimension 11)*

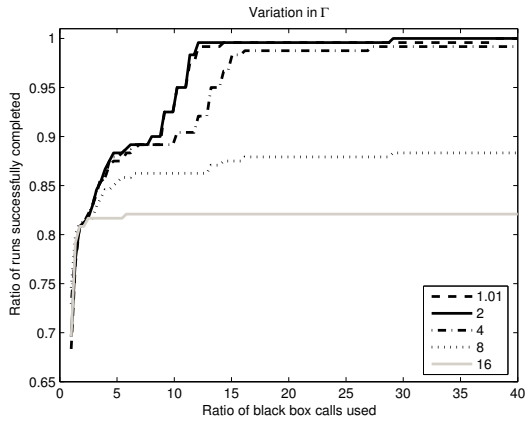

Fig. 5 Performance Profile for Γ parameter

6.4.2 Minimum step length

We set $\text{TOL}_{\text{stop}} = 10^{-6}|x^0|$, $\Gamma = 2$, $\text{MAX}_{\texttt{short}} = 5$, $\text{TOL}_{\mu} = 9/12R$, and make our runs for $\text{MIN}_{length} \in \{10^{-4}, 10^{-6}, 10^{-8}, 10^{-10}, 0\}$. In Table 6 we provide statistical results for these tests, while Figure 7 displays a performance profile for each of the options tested.

Our results indicate that a minimum step length of 10^{-8} is an appropriate value. Bigger values of MIN_{length} make the algorithm stop prematurely, exiting at Step 3, while smaller values result in erroneous exiting at Step 4.

MIN_{length}	successes	mean black box calls	failures	mean black box calls
10^{-4}	32, 40	11.25, 7.60	88, 80	15.28, 23.73
10^{-6}	86, 62	19.20, 20.32	34, 58	21.29, 35.81
10^{-8}	120,120	25.82, 36.02	0, 0	NA
10^{-10}	99,105	30.20, 40.23	21, 15	10.48, 11.47
0	78, 91	29.01, 40.86	42, 29	15.31, 19.03

Table 6 Minimum step length test results, *(dimension 7, dimension 11)*

6.4.3 Short steps

We set $\text{TOL}_{\text{stop}} = 10^{-6}|x^0|$, $\Gamma = 2$, $\text{TOL}_{\mu} = 9/12R$, $\text{MIN}_{length} = 10^{-8}$, and make our runs for $\text{MAX}_{\texttt{short}} \in \{1, 2, 5, 10, +\infty\}$.

The results reported in Table 8 and the Performance Profile in Figure 9 indicate that a maximum allowed of 5 short steps is appropriate both for speed and accuracy. Bigger values of $\text{MAX}_{\texttt{short}}$ make the algorithm stall, while smaller values make the algorithm stop prematurely.

6.4.4 Minimum μ

We set $\text{TOL}_{\text{stop}} = 10^{-6}|x^0|$, $\Gamma = 2$, $\text{MAX}_{\texttt{short}} = 5$, $\text{MIN}_{length} = 10^{-8}$, and make our runs for $\text{TOL}_{\mu} \in \{1/12R, 4/12R, 7/12R, 9/12R, 11/12R\}$.

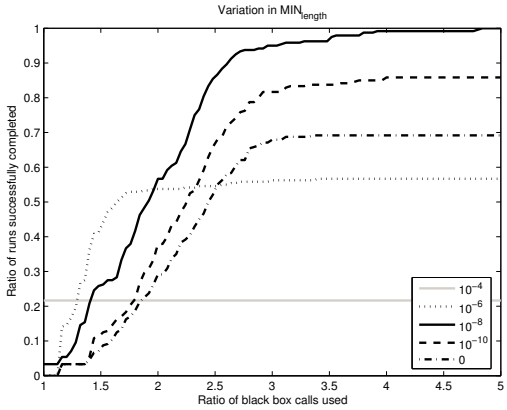

Fig. 7 Performance Profile for MIN_{length} parameter

$\text{MAX}_{\texttt{short}}$	successes	mean black box calls	failures	mean black box calls
1	58, 65	25.60, 27.46	62, 55	21.44, 22.69
2	115,114	27.93, 36.48	5, 6	134.40, 39.33
5	118,120	30.23, 39.99	2, 0	300.00, NA
10	118,120	32.99, 43.91	2, 0	300.00, NA
∞	118,120	158.05,234.20	2, 0	300.00, NA

Table 8 Maximum number of short steps test results, *(dimension 7, dimension 11)*

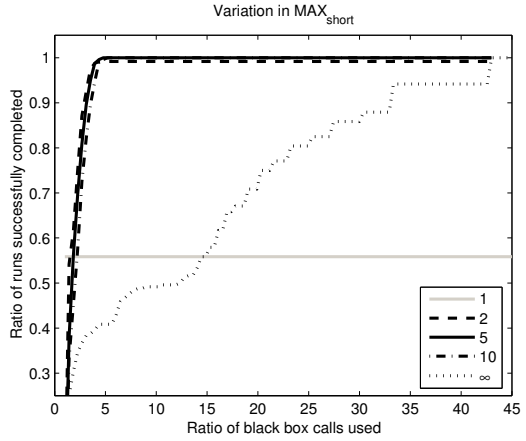

Fig. 9 Performance Profile for $\text{MAX}_{\texttt{short}}$ parameter

By the theory in Example 17, we expect convergence when $R \geq \text{TOL}_\mu + \Gamma\rho \approx \text{TOL}_\mu + 2/12R$, or $\text{TOL}_\mu \leq 10/12R$.

The results reported in Table 10 and the Performance Profile in Figure 11 support this for $\text{TOL}_\mu = 7/12R$ and $9/12R$. However, when TOL_μ becomes small convergence speeds appear to decrease, which result in premature violation of the $\text{MAX}_{\texttt{short}}$ parameter. This could be avoided by setting $\text{MAX}_{\texttt{short}}$ to a higher value. As before, (subsection 6.4.1, when TOL_μ is too large we see a very rapid detection that "R is insufficient."

TOL_μ	successes	mean black box calls	failures	mean black box calls
1/12R	93, 84	33.10, 40.60	27, 36	42.78, 62.14
4/12R	115,110	32.53, 44.01	5, 10	48.20, 59.60
7/12R	120,120	30.13, 42.99	0, 0	NA
9/12R	120,120	26.63, 40.18	0, 0	NA
11/12R	116,100	23.66, 42.46	4, 20	6.50, 1.05

Table 10 μ-tolerance test results, *(dimension 7, dimension 11)*

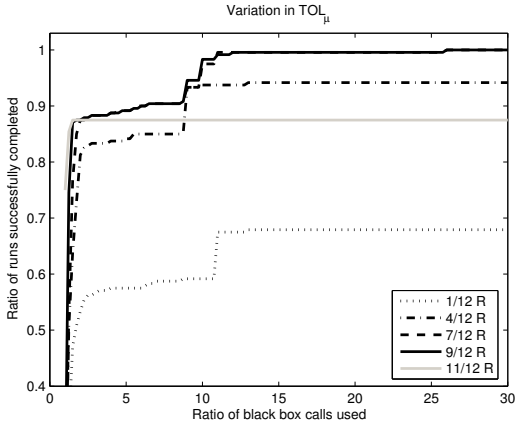

Fig. 11 Performance Profile for TOL_μ parameter

6.5 Increasing the problems' dimensions

We randomly generated 120 problems of dimension $N = 100$, again divided in 6 groups (with two special instances with only convex and nonconvex functions), as described in Table 12.

N	nf	nf_{act}	numbers range	N	nf	nf_{act}	numbers range
100	9	1	$[-10, 10]^\dagger$	100	121	1	$[-10, 10]^\natural$
100	9	5	[-10,10]	100	121	61	[-100,100]
100	9	9	[0,10]	100	121	121	[-10,0]

Table 12 Large Dimension Problem Set († = nonconvex, $\natural$=convex)

We set $\Gamma = 2$, $\text{MAX}_{\texttt{short}} = 5$, $\text{MIN}_{length} = 10^{-8}$, and $\text{TOL}_\mu = 9/12R$. Due to the time involved by the increase in dimension, we reduced TOL_{stop} to $10^{-4}|x^0|$. All 120 of the tests in 100 dimensions were successfully solved. The mean number of black box calls used was 125.08. By breaking down the tests into the various categories and examining the number of black box calls used in each case we observed the following:

– Tests with 9 or less active indices at 0 used on average 29.43 black box calls.

– Tests with 61 active indices used on average 201.09 black box calls.

– Tests with 121 active indices used on average 430.80 black box calls.
– Tests where the objective functions were strictly convex used on average 49.95 black box calls.
– Tests where the objective functions were nonconvex used on average 14.94 black box calls.

Overall these results are very positive. They show that, even in high dimensions, it is possible to accurately compute proximal points for nonconvex functions.

Contrary to our expectations, nonconvex runs seem to converge faster than convex ones, at least in average values. Nonetheless, it should be noted that in the nonconvex runs the number of quadratic functions used to define the problem was $\texttt{nf} = 9$, while for the convex runs 121 quadratic functions were used. This phenomenon suggests that the number of black box calls required to compute the proximal point is less dependent on the convexity of the problem than on overall smoothness of the problem near the solution (i.e., on the dimension of the $\mathcal{V}$-subspace in [27]). In our numerical experiments, tests with functions defined by less quadratic functions ($\texttt{nf}$) and, therefore, with less active indices at 0 ($\texttt{nf}_{\texttt{act}}$), appear to converge quicker than those with higher values for $\texttt{nf}$ and $\texttt{nf}_{\texttt{act}}$. Our results also suggest that the values of $\texttt{nf}$ and $\texttt{nf}_{\texttt{act}}$ have a greater effect on the speed of convergence than the problem dimension. Further research may provide enlightenment along these lines.

6.6 Comparison with bundle methods

Problem (1) written with $x = x^0$ gives the proximal point mapping of f at the point x^0. Since (1) is a nonsmooth unconstrained optimization problem, the computation of the proximal point could also be done by using some nonsmooth optimization method, applied to minimizing the function

$$f_R(\cdot) := f(\cdot) + R\frac{1}{2}|\cdot - x^0|^2.$$

In order to further assess our algorithm, we compare the behaviour of CompProx with two bundle methods minimizing f_R, for the functions f in the all-convex and all-nonconvex set problems from Table 1 (♮ and †, respectively). The first bundle method is n1cv2, the proximal bundle method for convex minimization from [18]; available upon request at

`http://www-rocq.inria.fr/estime/modulopt/optimization-routines/n1cv2.html`.

The second bundle method is n2fc1, an ε-steepest descent bundle method for nonconvex minimization with box constraints described in [19]. In our runs, we use a box constraint $|x_i| \leq 10^5$ for all $i \leq N$ for n2fc1.

To compare the three algorithms, we first adjusted their parameters to optimal performance ($\Gamma = 2$, $\text{MAX}_{\texttt{short}} = \infty$, $\text{MIN}_{length} = 10^{-8}$, $\text{TOL}_{\mu} = 9/12R$, and $\text{TOL}_{\text{stop}} = 0$ in CompProx). Then, for each of the 40 test problems, we ran each algorithm until 100 black box calls have been made, and saved the best point found.

With these settings, CompProx completed 100 black box calls on all 40 tests. By contrast, both n1cv2 and n2fc1 occasionally had an abnormal early termination (reporting problems in the quadratic programming solver computing the direction) for some functions having one active index ($\texttt{nf}_{\texttt{act}} = 1$). Table 13 provides statistical information for each code on both the convex and nonconvex test problem sets. The columns reporting accuracy use the formula

$$\text{relative accuracy} = \log_{10}\Big(\frac{|x^{\text{best}}|}{|x^0|}\Big)$$

where x^{best} is the best iterate found by each method.

Our results indicate the superiority of CompProx over the two bundle codes, especially in the worst case scenario. Similar conclusion follows from examining the performance profile in Figure 14, which plots *the accuracy achieved* against *the ratio of tests completed* for each of the cases in Table 13.

Code	Test Set	worst relative accuracy	mean relative accuracy	best relative accuracy	mean black box calls
N1CV2	convex	-3.4	-4.8	-6.5	100
N2FC1	convex	-2.0	-4.5	-7.9	74
CompProx	convex	-5.1	-6.3	-7.3	100
N1CV2	nonconvex	-1.1	-8.3	-13.0	98
N2FC1	nonconvex	-2.3	-4.7	-7.6	64
CompProx	nonconvex	-7.5	-9.9	-12.9	100

Table 13 Code comparison results on convex and nonconvex test sets

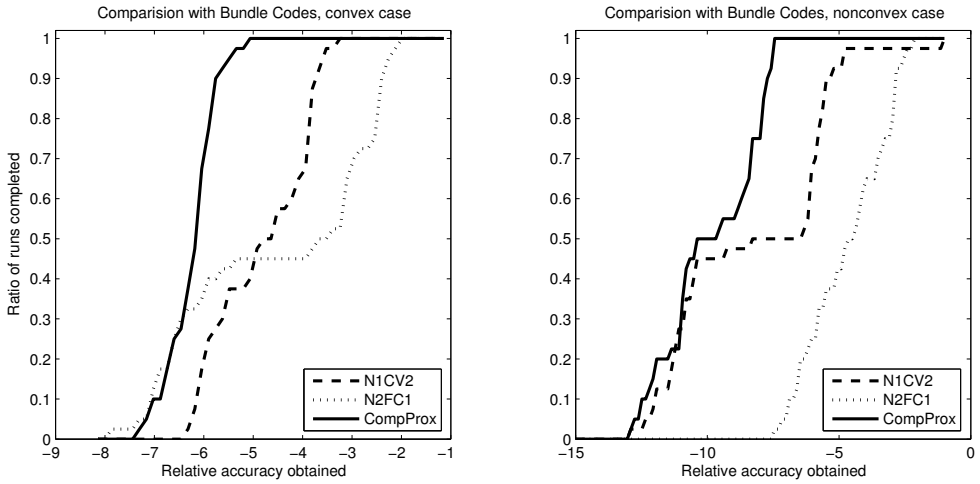

Fig. 14 Performance Profile for convex and nonconvex test sets

References

1. Auslender, A.: Numerical methods for nondifferentiable convex optimization. Math. Programming Stud. **30**, 102–126 (1987)
2. Auslender, A., Crouzeix, J.P., Fedit, P.: Penalty-proximal methods in convex programming. J. Optim. Theory Appl. **55**(1), 1–21 (1987)
3. Auslender, A., Haddou, M.: An interior-proximal method for convex linearly constrained problems and its extension to variational inequalities. Math. Programming **71**(1, Ser. A), 77–100 (1995)
4. Bellman, R., Kalaba, R., Lockett, J.: Numerical Inversion of the Laplace Transform. Elsevier (1966)
5. Bernard, F., Thibault, L.: Prox-regularity of functions and sets in Banach spaces. Set-Valued Anal. **12**(1-2), 25–47 (2004)
6. Bernard, F., Thibault, L.: Prox-regular functions in Hilbert spaces. J. Math. Anal. Appl. **303**(1), 1–14 (2005)
7. Bonnans, J.F., Gilbert, J., Lemaréchal, C., Sagastizábal, C.: A family of variable metric proximal methods. Math. Program., Ser. A **68**, 15–47 (1995)
8. Cominetti, R., Courdurier, M.: Coupling general penalty schemes for convex programming with the steepest descent and the proximal point algorithm. SIAM J. Optim. **13**(3), 745–765 (2002)
9. Correa, R., Lemaréchal, C.: Convergence of some algorithms for convex minimization. Math. Program. **62**(2), 261–275 (1993)
10. Dolan, E., Moré, J.: Benchmarking optimization software with performance profiles. Math. Program. **91**(2, Ser. A), 201–213 (2002)
11. Eckstein, J.: Nonlinear proximal point algorithms using Bregman functions, with applications to convex programming. Mathematics of Operations Research **18**, 202–226 (1993)
12. Frangioni, A.: Solving semidefinite quadratic problems within nonsmooth optimization algorithms. Comput. Oper. Res. **23**(11), 1099–1118 (1996)

13. Güler, O.: New proximal point algorithms for convex minimization. SIAM Journal on Optimization **2**, 649–664 (1992)
14. Hare, W. L., Lewis, A. S.: Identifying active constraints via partial smoothness and prox-regularity. J. Convex Anal. **11**, 251-266 (2004).
15. Hiriart-Urruty, J.-B., Lemaréchal, C. L.: Convex Analysis and Minimization Algorithms. Number 305-306 in Grundlehren der mathematischen Wissenschaften. Springer-Verlag, Berlin, (1993)
16. Kiwiel, K.: A method for solving certain quadratic programming problems arising in nonsmooth optimization. IMA Journal of Numerical Analysis **6**, 137–152 (1986)
17. Kiwiel, K.: A tilted cutting plane proximal bundle method for convex nondifferentiable optimization. Operations Research Letters, **10**, 75–81, (1991)
18. Lemaréchal, C. Sagastizábal, C. Variable metric bundle methods: From conceptual to implementable forms. *Mathematical Programming*, 76:393–410, 1997.
19. Lemaréchal, C., Strodiot, J.-J., Bihain, A.: On a bundle algorithm for nonsmooth optimization. Nonlinear programming, **4**, 245–282 (1980)
20. Lukšan, L., Vlček, J.: A bundle-Newton method for nonsmooth unconstrained minimization. Math. Programming, **83**, 373–391, (1998)
21. Lukšan, L., Vlček, J.: Globally convergent variable metric method for nonconvex nondifferentiable unconstrained minimization. J. Optim. Theory Appl., **2**, 407–430, (2001)
22. Makela, M. M., Neittaanmaki, P.: Nonsmooth Optimization: Analysis and Algorithms with Applications to Optimal Control. World Scientific: Singapore (1992)
23. Martinet, B.: Régularisation d'inéquations variationnelles par approximations successives. Rev. Française Informat. Recherche Opérationnelle **4**(Ser. R-3), 154–158 (1970)
24. Mifflin, R.: An algorithm for constrained optimization with semismooth functions. Math. Oper. Res. **2**, 191–207 (1977)
25. Mifflin, R.:Convergence of a modification of Lemaréchal's algorithm for nonsmooth optimization. Progress in nondifferentiable optimization, IIASA Collaborative Proc. **8**(Ser. CP-82), 85–95 (1982)
26. Mifflin, R.: A modification and extension of Lemarechal's algorithm for nonsmooth minimization. Math. Programming Stud. **17**, 77–90 (1982)
27. Mifflin, R., Sagastizábal, C.: Proximal points are on the fast track. J. Convex Anal. **9**, 563–579 (2002)
28. Mifflin, R., Sagastizábal, C.: *VU*-smoothness and proximal point results for some nonconvex functions. Optim. Methods Softw. **19**, 463–478 (2004)
29. Mordukhovich, B.S.: Maximum principle in the problem of time optimal response with nonsmooth constraints. Prikl. Mat. Meh. **40**(6), 1014–1023 (1976)
30. Moreau, J.: Proximité et dualité dans un espace Hilbertien. Bulletin de la Société Mathématique de France **93**, 273–299 (1965)
31. Poliquin, R.A., Rockafellar, R.T.: Generalized Hessian properties of regularized nonsmooth functions. SIAM J. Optim. **6**(4), 1121–1137 (1996)
32. Poliquin, R.A., Rockafellar, R.T.: Prox-regular functions in variational analysis. Trans. Amer. Math. Soc. **348**(5), 1805–1838 (1996)
33. Rockafellar, R.: Augmented Lagrangians and applications of the proximal point algorithm in convex programming. Mathematics of Operations Research **1**, 97–116 (1976)
34. Rockafellar, R.: Monotone operators and the proximal point algorithm. SIAM Journal on Control and Optimization **14**, 877–898 (1976)
35. Rockafellar, R., Wets, R.B.: Variational Analysis. No. 317 in Grund. der math. Wiss. Springer-Verlag (1998)
36. Tarasov, V.N., Popova, N.K.: A modification of the cutting-plane method with accelerated convergence. in *Nondifferentiaible Optimization: Motivations and application* (V.F. Demjanov, D. Pallaschke, eds.). Lecture Notes in Economics and Mathematical Systems, vol. 255, Springer-Verlag, 284–290 (1984)
37. Yosida, K.: Functional Analysis. Springer Verlag (1964)