A generalized proximal linearized algorithm for DC functions with application to the optimal size of the firm problem

J. Cruz Neto, P. Oliveira, Antoine Soubeyran, J. Souza

A generalized proximal linearized algorithm for DC functions with application to the optimal size of the firm problem

J. X. Cruz Neto[1] · P. R. Oliveira[2] · A. Soubeyran[3] · J. C. O. Souza[1]

Abstract
The purpose of this paper is twofold. First, we examine convergence properties of an inexact proximal point method with a quasi distance as a regularization term in order to find a critical point (in the sense of Toland) of a DC function (difference of two convex functions). Global convergence of the sequence and some convergence rates are obtained with additional assumptions. Second, as an application and its inspiration, we study in a dynamic setting, the very important and difficult problem of the limit of the firm and the time it takes to reach it (maturation time), when increasing returns matter in the short run. Both the formalization of the critical size of the firm in term of a recent variational rationality approach of human dynamics and the speed of convergence results are new in Behavioral Sciences.

Keywords Proximal point method · DC function · Kurdyka–Łojasiewicz inequality · Limit of the firm · Variational rationality

1 Introduction

In this paper, we show how a generalized proximal point method applied to DC functions (a function which can be written as difference of two convex functions) can be a nice tool to

J. X. Cruz Neto was supported in part by CNPq Grant 305462/2014-8, P. R. Oliveira was supported in part by CNPq and J. C. O. Souza was supported in part by CNPq-Ciências sem Fronteiras Grant 203360/2014-1.

✉ J. C. O. Souza
joaocos.mat@ufpi.edu.br

J. X. Cruz Neto
jxavier@ufpi.edu.br

P. R. Oliveira
poliveir@cos.ufrj.br

A. Soubeyran
antoine.soubeyran@gmail.com

1 DM, Universidade Federal do Piauí, Teresina, Brazil

2 PESC-COPPE, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil

3 Aix-Marseille University (Aix-Marseille School of Economics), CNRS & EHESS, Marseille, France

solve a dynamic version of the very important and difficult problem of the limit of the firm and the time it takes to reach it by using the recent variational rationality (VR) approach of human dynamics; see Soubeyran (2009, 2010, 2016). To this end, we consider a proximal point method which finds at each iteration an approximated solution of a minimization problem involving a convex approximation of the objective DC function (possible non-convex) and a generalized regularization (possible non-symmetric) and we consider a hierarchical firm including an entrepreneur, a profile of workers, and a succession of periods where the entrepreneur can hire, fire or keep working workers in a changing environment. Each period, the entrepreneur chooses how much to produce of the same final good (of a given quality) and sells each unit of this good at the same and fixed price. Therefore, we show that a firm can achieve an optimal size (and the time it takes to reach it) by means of a generalized proximal method for DC functions.

It is well known that the proximal point method is one of the most studied method for finding zeros of maximal monotone operators, and in particular, it is used to solve convex optimization problems. The proximal point method was introduced into optimization literature by Martinet (1970) and popularized by Rockafellar (1976), who showed that the algorithm can be used for finding zeros of monotone operators, even if each subproblem is performed inexactly, which is an important consideration in practice; see e.g. Bento and Soubeyran (2015a, b), Burachik and Svaiter (2001), Zaslavski (2010). In particular, the following algorithm has been used for finding a minimizer of a convex function

$$x^{k+1} = \arg\min_{x\in\mathbb{R}^n}\{f(x) + \frac{1}{2\lambda_k}||x - x^k||^2\}. \tag{1}$$

Even the idea underlying the convergence results of (1) being based on the monotonicity of subdifferential operators of convex functions this procedure has been adapted to deal with possible non-convex functions; see Hare and Sagastizábal (2009), Kaplan and Tichatschke (1998), Pan and Chen (2007), Papa Quiroz and Oliveira (2012) and references therein. To the best of our knowledge, a proximal point method for DC function was first introduced by Sun et al. (2003). There is a huge literature on DC theory both from a theoretical point of view and for algorithmic purposes; see e.g. Bačák and Borwein (2011), Fernández Cara and Moreno (1988), Hartman (1959), Hiriart-Urruty (1986), Moudafi and Maingé (2006), Muu and Quoc (2010), Pham and Souad (1986), Pham and An (1997), Pham et al. (2005), Souza and Oliveira (2015), Sun et al. (2003). DC optimization algorithms have been proved to be particularly successful for analyzing and solving a variety of highly structured and practical problems; see, for instance, Pham et al. (2005) and references therein. Applications of DC theory in game theory can be found in Muu and Quoc (2010), plasma physics and fluid mechanics can be found in Fernández Cara and Moreno (1988), and examples of DC functions from various parts of analysis can be found in Bačák and Borwein (2011).

On the other hand, our regularization term in (1) is no longer an Euclidian norm but a quasi distance, where the quasi distance from a point x to an other point y can be different from the reverse. Soubeyran (2009, 2010, 2016) showed how such a quasi distance modelizes in Behavioral Sciences costs of being able to move from on position to an other one where costs of being able to move from x to y are usually different from costs of being able to move from y to x. Hence the symmetric assumption of distances must be dropped. Nevertheless, a quasi distance can preserve nice properties useful for the convergent analysis, such as continuity and coercivity. Extensions of proximal point methods by using nonlinear or generalized regularizations were considered, for instance, in Bento and Soubeyran (2015a), Bento and Soubeyran (2015b), Burachik and Svaiter (2001), Chen and Teboulle (1993), Eckstein (1993), Kiwiel (1997), Moreno et al. (2012), Pan and Chen (2007). The works (Bento and Soubeyran

2015a, b; Moreno et al. 2012) are devoted to study convergence properties of generalized proximal point methods where the regularization term is a quasi distance (or quasi metric). Bento and Soubeyran (2015a) discussed how such generalized proximal point method can be a nice tool to modelize the dynamics of human behaviors in the context of the (VR) variational rationality approach. Applications of quasi metric spaces to Behavioral Sciences (Psychology, Economics, Management, Game theory, etc.) and theoretical computer science can be found for instance in Bao et al. (2016a), Brattka (2003), Künzi et al. (2006), Romaguera and Sanchis (2003).

The goal of this paper is two-fold. Firstly, we propose an inexact generalized proximal linearized algorithm for solving optimization problems involving a DC function which is neither necessarily convex nor smooth. We also study global convergence of the sequence and some convergence rate results under reasonable assumptions. As a second contribution, we provide an application, in a dynamic context, to the very important problem of the limit of the firm, when increasing returns (concave costs of production) prevail in the short run, using the recent (VR) variationality approach of a lot of stay/stability and change dynamics in Behavioral Sciences (see Soubeyran 2009, 2010, 2016). This is a difficult problem, both for conceptual and technical reasons. Different variants of this example can be found in Soubeyran (2009, 2010), Bento and Soubeyran (2015a, b), and Bao et al. (2015). But none of them examines the very important case of increasing returns, which is the realistic case for production costs, as we do here, as an application of the proximal point method for DC optimization and its related speed of convergence.

The organization of this paper is as follows. In Sect. 2 some concepts in subdifferential theory, DC optimization and quasi distance are presented. In Sect. 3 an inexact generalized proximal linearized algorithm is discussed. In Sect. 4, the convergence analysis of the method is studied and we compare our results with the existing DC literature. Finally, Sect. 5 is devoted to determine the limit of the firm, when increasing returns matter in the short run. Future works are mentioned in the conclusions.

2 Preliminaries

Let $\Gamma_0(\mathbb{R}^n)$ denote the convex cone of all the proper (i.e. not identically equal to $+\infty$) lower semicontinuous convex functions from $\mathbb{R}^n$ to $\mathbb{R} \cup \{+\infty\}$, let $\langle \cdot, \cdot \rangle$ be the canonical inner product, and $||\cdot||$ the corresponding Euclidean norm on $\mathbb{R}^n$. We denote by $\mathcal{D} = \text{int}(\text{dom}(h))$, where the term "int X" refers to the interior of the set X. The effective domain of a function f, denoted by $\text{dom}(f)$, is defined as

$$\text{dom}(f) = \{x \in \mathbb{R}^n \ : \ f(x) < +\infty\}.$$

Let $\{x^k\}$ be a sequence in $\mathbb{R}^n$. We call cluster point (or accumulation point) of the sequence $\{x^k\}$ every point $x \in \mathbb{R}^n$ such that there exists a subsequence $\{x^{k_j}\}$ of $\{x^k\}$ converging to x.

2.1 Subdifferential theory

Let us recall some definitions and properties of the subdifferential theory which can be found, for instance, in Mordukhovich and Shao (1996), Rockafellar and Wets (1998).

Definition 1 Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a proper lower semicontinuous function.

1. The Fréchet subdifferential of f at x, denoted by $\hat{\partial} f(x)$, is defined as follows

$$\hat{\partial} f(x) = \begin{cases} \{v \in \mathbb{R}^n \ : \ \liminf\limits_{\substack{y \to x \\ y \neq x}} \frac{f(y)-f(x)-\langle v, y-x\rangle}{||x-y||} \geq 0\}, & \text{if } x \in \text{dom}(f); \\ \emptyset, \quad \text{if } x \notin \text{dom}(f). \end{cases}$$

2. The limiting-subdifferential of f at x, denoted by $\partial f(x)$, is defined as follows

$$\partial f(x) = \begin{cases} \{v \in \mathbb{R}^n \ : \ \exists x^k \to x, \ f(x^k) \to f(x), \ v^k \in \hat{\partial} f(x^k) \to v\}, & \text{if } x \in \text{dom}(f); \\ \emptyset, \quad \text{if } x \notin \text{dom}(f). \end{cases}$$

We denote by $\text{dom}\, \partial f = \{x \in \mathbb{R}^n \ : \ \partial f(x) \neq \emptyset\}$. Recall that the limiting-subdifferential is closed and $\hat{\partial} f(x) \subset \partial f(x)$. If f is a proper, lower semicontinuous and convex function, and $x \in \text{dom}(f)$, then $\hat{\partial} f(x)$ coincides with the classical subdifferential in the sense of convex analysis and it is nonempty, closed and convex set.

2.2 Difference of convex functions

Let $\mathcal{DC}(\mathbb{R}^n)$ denote the class of DC functions defined on $\mathbb{R}^n$. A general DC program is of the form

$$f^* = \inf\{f(x) = g(x) - h(x) \quad : \quad x \in \mathbb{R}^n\},$$

with $g, h \in \Gamma_0(\mathbb{R}^n)$. Such a function f is called a DC function while the convex functions g and h are DC components of f. In DC programming, the convention

$$(+\infty) - (+\infty) = +\infty$$

has been adopted to avoid the ambiguity $(+\infty) - (+\infty)$ that does not present any interest here. Note that the finiteness of f^* implies that $\text{dom}(g) \subseteq \text{dom}(h)$. Such inclusion will be assumed throughout the paper.

Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a proper lower semicontinuous DC function (resp. bounded from below). Then, f has lower semicontinuous DC components g and h, with $\inf_{x \in \mathbb{R}^n} h(x) = 0$ (resp. g bounded from below); see Yassine et al. (2001, Proposition 2.4) and Lhilali Alaoui (1996, Proposition 3.2). It is well known that a necessary condition for $x \in \text{dom}(f)$ to be a local minimizer of f is $\partial h(x) \subset \partial g(x)$. The same condition holds true when x is a local maximum of f; see Hiriart-Urruty (1986). We will focus our attention on finding points such that $\partial h(x) \cap \partial g(x) \neq \emptyset$ called critical points of f. This concept was introduced in the DC literature by Toland (1979) and has been used widely; see for instance An et al. (2009), Moudafi and Maingé (2006), Pham and Souad (1986), Pham and An (1997), Pham et al. (2005), Souza and Oliveira (2015), Sun et al. (2003) and references therein. Clearly, local minimizer and local maximum are critical points.

It is worth to mention that the class of DC functions is the vector space generated by the cone of convex functions which contains for instance the class of lower-$\mathcal{C}^2$ functions (f is said to be lower-$\mathcal{C}^2$ if f is locally a supremum of a family of $\mathcal{C}^2$ functions). In particular, $\mathcal{DC}(\mathbb{R}^n)$ contains the space $\mathcal{C}^{1,1}$ which is the class of continuously differentiable functions whose its gradient is locally Lipschitz. $\mathcal{DC}(\mathbb{R}^n)$ is closed under the operations usually considered in optimization. For instance, a linear combination, a finite supremum or the product of two DC functions remains DC. Locally DC functions on $\mathbb{R}^n$ are DC functions on $\mathbb{R}^n$; see Hiriart-Urruty (1986) and references therein for the details.

2.3 Quasi distance

Definition 2 A quasi metric space is a pair (X, q) such that X is a nonempty set, and $q : X \times X \to \mathbb{R}_+$, called a quasi metric or quasi distance, is a mapping satisfying:

1. For all $x, y \in X$, $q(x, y) = q(y, x) = 0 \Leftrightarrow x = y$;
2. For all $x, y, z \in X$, $q(x, z) \leq q(x, y) + q(y, z)$.

Therefore, metric spaces are quasi metric spaces satisfying the symmetric property $q(x, y) = q(y, x)$. Quasi distances are not necessarily convex or differentiable but it preserves nice properties such as continuity and coercivity; see assumption A3 in Sect. 4. Examples of quasi distances can be found in Moreno et al. (2012) and references therein. The full justifications for using quasi distances in order to modelize costs of moving in Behavioral Sciences are given in the presentation of the Variational Rationality approach, see Soubeyran (2009, 2010, 2016), and more recently in Soubeyran (2018a, b). A quick justification is that costs of moving are non symmetric, because costs of moving from x to y are not the same as costs of moving from y to x. Furthermore, the triangle inequality comes from the fact that detours are costly, because of some kind of fixed starting and stopping costs for each detour. A complete answer for this justification would be too long. It requires to have in mind the framework of the Variational Rationality approach, where resistance to change plays a major role. For other papers where other generalized distances are used to modelize costs of moving, including w-distances, proximal distances, Bregman distances, and partial quasi distances as regularization terms, see Moreno et al. (2012) for quasi distances, Bao et al. (2016b) for w-distances and partial quasi distances, Bento et al. (2016) for proximal distances and Cruz Neto et al. (2010) for Bregman distances.

3 A generalized proximal linearized algorithm

Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a proper DC function bounded from below with DC components g and h, i.e., $f(x) = g(x) - h(x)$ with $g, h \in \Gamma_0(\mathbb{R}^n)$. In this section, we consider a generalized proximal linearized algorithm for finding a critical point of the DC function f. At each iteration our algorithm linearizes the function $f(x)$ but never directly minimize it, while it minimizes the function $g(x)$ in conjunction with the linearization of $h(x)$ and a generalized regularization. As mentioned before, this kind of linearized method has proved to be efficient for solving a large number of problems, for instance, the convex problem of minimizing a sum of two convex functions; see Bolte et al. (2014), Goldfarb et al. (2013), Kiwiel et al. (1999). Our method minimizes upper-bounds of the objective function as we will see in Remark 4. This kind of method is often called *majorization-minimization* (see Lange et al. 2000) or *successive upper-bound minimization* (see Razaviyayn et al. 2016). Majorizing surrogates have been used successfully in large scale problems (Mairal 2015), signal processing literature about sparse estimation (Daubechies et al. 2004; Gasso et al. 2009), linear inverse problems in image processing (Ahn et al. 2006; Erdogan and Fessler 1999), and matrix factorization (Lee and Seung 2001; Mairal et al. 2010).

The algorithms analyzed in this paper are based on the computation of the proximity operator popularized by Rockafellar (1976). Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a proper, convex and lower semicontinuous function, $\lambda > 0$ and $y \in \mathbb{R}^n$, the proximal operator at y with respect to λf $\text{prox}_{\lambda f} : \mathbb{R}^n \to \mathbb{R}^n$ is defined by

$$\text{prox}_{\lambda f}(y) = \arg\min_{x\in\mathbb{R}^n}\left\{f(x) + \frac{1}{2\lambda}||x-y||^2\right\}.$$

As remarked by Rockafellar (1976), from a practical point of view, it is essential to replace the proximal point with an approximate version of it. Therefore, for a given $\epsilon \geq 0$, we say that $z \in \mathbb{R}^n$ is an approximation of $\text{prox}_{\lambda f}(y)$ with ϵ-precision, denote by $z \approx_\epsilon \text{prox}_{\lambda f}(y)$, if

$$f(z) + \frac{1}{2\lambda}||z-y||^2 \leq f(x) + \frac{1}{2\lambda}||x-y||^2 + \epsilon, \quad \forall x \in \mathbb{R}^n,$$

i.e.,

$$z \approx_\epsilon \arg\min_{x\in\mathbb{R}^n}\left\{f(x) + \frac{1}{2\lambda}||x-y||^2\right\}.$$

This kind of inexact proximal algorithm is analyzed in Zaslavski (2010) for convex functions.

Next, we propose an inexact generalized proximal linearized algorithm for solving optimization problems involving a DC function, where the term generalized stands to replace the Euclidean distance by a quasi distance in the proximal operator.

Algorithm 1—generalized inexact method

Step 1 Given an initial point $x^0 \in \mathcal{D}$, $\mu > 0$ and two auxiliar sequences $\{\epsilon_k\}$, with $\epsilon_k \geq 0$ for all $k \in \mathbb{N}$, and $\{\lambda_k\}$ a bounded sequence of positive numbers such that $\liminf_k \lambda_k > 0$. Set $k = 0$.

Step 2 Calculate

$$w^k \in \partial h(x^k). \tag{2}$$

Step 3 Compute $(x^{k+1}, \xi^{k+1}) \in \mathbb{R}^n \times \mathbb{R}^n$ such that

$$x^{k+1} \approx_{\epsilon_k} \arg\min_{x\in\mathbb{R}^n}\left\{g(x) - \langle w^k, x - x^k\rangle + \frac{1}{2\lambda_k}q^2(x^k, x)\right\}, \tag{3}$$

with

$$\xi^{k+1} \in \partial g(x^{k+1}) \tag{4}$$

where

$$||\xi^{k+1} - w^k|| \leq \mu q(x^k, x^{k+1}). \tag{5}$$

If $x^{k+1} = x^k$, stop. Otherwise, set $k := k + 1$ and return to Step 2.

Remark 1 The well definition of $\{w^k\}$ and $\{x^k\}$ is guaranteed if h is a convex function, and g is a convex and bounded from below function, respectively. It is straightforward to chech that, if f is a DC function bounded from below, then f admits a DC decomposition with g bounded from below. The conditions on the parameters $\{\lambda_k\}$ are that it is a bounded sequence of positive numbers such that $\liminf_k \lambda_k > 0$. These conditions are equivalent to the following: there exist constants $c_1, c_2 \in \mathbb{R}$ such that

$$0 < c_1 \leq \lambda_k \leq c_2, \quad \forall k \in \mathbb{N}. \tag{6}$$

Remark 2 If $\epsilon_k = 0$, for all $k \in \mathbb{N}$ (called exact version of Algorithm 1), then (3) becomes the following:

$$x^{k+1} \in \arg\min_{x\in\mathbb{R}^n} \left\{ g(x) - \langle w^k, x - x^k\rangle + \frac{1}{2\lambda_k} q^2(x^k, x) \right\}, \tag{7}$$

which implies that

$$0 \in \partial g(x^{k+1}) - w^k + \frac{1}{2\lambda_k} \partial q^2(x^k, \cdot)(x^{k+1}). \tag{8}$$

Since, for each $k \geq 0$, the map $y \mapsto q(x^k, y)$ is locally Lipschitz, i.e., q is locally Lipschitz in the second variable (see Moreno et al. 2012, Proposition 3.6) it follows from (Mordukhovich and Shao 1996, Theorem 7.1) that $\partial q^2(x^k, \cdot)(x^{k+1}) \subset 2q(x^k, x^{k+1})\partial q(x^k, \cdot)(x^{k+1})$. Thus, from (8), we obtain that there exist $\xi^{k+1} \in \partial g(x^{k+1})$ and $\eta^{k+1} \in \partial(q(x^k, \cdot))(x^{k+1})$ such that

$$w^k - \xi^{k+1} = \frac{\eta^{k+1}}{\lambda_k} q(x^k, x^{k+1}). \tag{9}$$

Therefore, (9) implies that any solution x^{k+1} of (7) is also a solution of (3) satisfying (5) as long as $\{x^k\}$ is bounded (which implies that $\{\eta^k\}$ is bounded; see remark below). In this context, the constant μ in (5) can be taken as any upper bound of the (bounded) sequence $\{\frac{\eta^{k+1}}{\lambda_k}\}$.

Remark 3 Note that we are not assuming the boundedness of the sequences $\{w^k\}$, $\{\xi^k\}$ and $\{\eta^k\}$. It comes from the fact the subdifferential of a locally Lipschitz function is locally bounded (see Rockafellar and Wets 1998, Theorem 9.13) together with assumption that $\{x^k\}$ is bounded. Recall that $w^k \in \partial h(x^k)$, $\xi^k \in \partial g(x^k)$ and $\eta^k \in \partial(q(x^{k-1}, \cdot))(x^k)$ taking in account that h, g are convex functions (and hence locally Lipschitz) and $q(x^{k-1}, \cdot)$ is locally Lipchitz (see Moreno et al. 2012, Proposition 3.6).

Remark 4 Note that in (7), instead of minimize $f(\cdot) = g(\cdot) - h(\cdot)$ directly, we minimize the linear approximation $g(\cdot) - \langle w^k, \cdot - x^k\rangle$ in addition with the proximal regularization. Without loss of generality such an approximation can be taken as $g(\cdot) - h(x^k) - \langle w^k, \cdot - x^k\rangle$ because the term $h(x^k)$ is constant. Then (7) can be viewed as

$$x^{k+1} \in \arg\min_{x\in\mathbb{R}^n} \varphi_k(x), \tag{10}$$

where $\varphi_k(x) = g(x) - h(x^k) - \langle w^k, x - x^k\rangle + \frac{1}{2\lambda_k} q^2(x^k, x)$. In this case, for each $k \geq 0$, from the convexity of h, we have $f(x) \leq \varphi_k(x)$, for all $x \in \mathrm{dom}(f)$. Therefore, in the problem (10) one minimizes upper-bounds of the objective function f and each minimization step decreases the value of the objective function, i.e., $f(x^{k+1}) < f(x^k)$, for all $k \in \mathbb{N}$. This will be prove in Theorem 1.

We emphasize that our algorithm is different of the DCA algorithm considered by Pham and Souad (1986). Our algorithm shares the same idea of DCA algorithm, namely, linearizing some component $g(\cdot)$ or $h(\cdot)$; or both of the DC objective function $f(x) = g(x) - h(x)$. However, our algorithm is simpler, because linearization is done directly, and not on the dual components, besides the fact that it is well known that subgradient method may not converge or decrease monotonically even for (non-differentiable) convex functions; see Polyak (1978). An example comparing the performance of a proximal point method and DCA algorithm can be found in Moudafi and Maingé (2006). In the recent "variational rationality approach" (Soubeyran 2009, 2010, 2016), the perturbation term of a proximal algorithm can be seen as

a crude formulation of the complex concept of resistance to change, while the utility generated by a change in the objective function can represent a crude formulation of the motivation to change concept; see Bento and Soubeyran (2015a, b). Algorithm 1 is closely related to the proximal method proposed by Sun et al. (2003), but our modeling approach seems to be more appropriate for applications in behavioral science using the "variational rationality approach" where costs of being able to change from a current position to other one, and costs of being able to stay in the current position are not necessarily symmetric and equal to zero, respectively; see Bento and Soubeyran (2015a, b), Moreno et al. (2012). The fact that the inexact proximal point algorithm is well adapted to applications in Behavioral Sciences is shown extensively in Bento and Soubeyran (2015a, b), using the variational rationality approach of human behaviors (Soubeyran 2009, 2010, 2016). On the contrary, the DCA method proposed by Pham and Souad (1986) is inappropriate for direct applications in Behavioral Sciences, because it uses, at the very beginning, a dual formulation for each component $g(\cdot)$ and $h(\cdot)$, who have no clear behavioral meaning and would be very expensive (if not impossible) to calculate, even for a very clever agent.

4 Convergence analysis

DCA algorithm was first introduced by Pham and Souad (1986) and the proximal point method for DC functions was first analyzed by Sun et al. (2003). These methods have been extensively developed since then and have the following convergence properties:

(i) The objective function f decreases through the sequence $\{x^k\}$, i.e., $f(x^{k+1}) < f(x^k)$, for all $k \in \mathbb{N}$;
(ii) The sequence is asymptotically regular, i.e., $\sum_{k=0}^{+\infty} ||x^{k+1} - x^k||^2 < \infty$;
(iii) If the sequence $\{x^k\}$ is bounded, then its cluster points are critical points of f;

see Moudafi and Maingé (2006), Pham and Souad (1986), Pham and An (1997), Pham et al. (2005), Souza and Oliveira (2015), Sun et al. (2003). Next, we prove that our method has the above convergence behavior and further we also prove global convergence of the method (convergence of the whole sequence) and establish some convergence rates under additional assumptions in Sects. 4.1 and 4.2, respectively.

Theorem 1 *The sequence $\{x^k\}$ generated by Algorithm 1 satisfies:*

1. *either the algorithm stops at a critical point;*
2. *or f ϵ_k-decreases, i.e., $f(x^{k+1}) < f(x^k) + \epsilon_k$, $\forall k \geq 0$.*

Proof If $x^{k+1} = x^k$, it follows from (5) that $\xi^{k+1} = w^k$, and so, $w^k \in \partial h(x^k) \cap \partial g(x^k)$ which means that x^k is a critical point of f. Now, suppose that $x^{k+1} \neq x^k$. From (2) and (3), we have

$$h(x^k) + \langle w^k, x^{k+1} - x^k \rangle \leq h(x^{k+1}),$$

and

$$g(x^{k+1}) - \langle w^k, x^{k+1} - x^k \rangle + \frac{1}{2\lambda_k} q^2(x^k, x^{k+1}) \leq g(x^k) + \epsilon_k,$$

respectively. Adding last two inequalities, we obtain

$$f(x^{k+1}) + \frac{1}{2\lambda_k} q^2(x^k, x^{k+1}) \leq f(x^k) + \epsilon_k, \tag{11}$$

which leads to $f(x^{k+1}) < f(x^k) + \epsilon_k$. □

Proposition 1 *Consider* $\{x^k\}$ *generated by Algorithm 1. If* $\sum_{k=0}^{+\infty} \epsilon_k < \infty$, *then* $\sum_{k=0}^{\infty} q^2(x^k, x^{k+1}) < \infty$ *and, in particular,* $\lim_{k\to+\infty} q(x^k, x^{k+1}) = 0$.

Proof From (11), we obtain

$$\frac{1}{2\lambda_k} q^2(x^k, x^{k+1}) \leq f(x^k) - f(x^{k+1}) + \epsilon_k,$$

and therefore

$$\frac{1}{2}\sum_{k=0}^{n-1} \frac{1}{\lambda_k} q^2(x^k, x^{k+1}) \leq f(x^0) - f(x^n) + \sum_{k=0}^{n-1} \epsilon_k, \forall n \in \mathbb{N}. \tag{12}$$

Since f is bounded from below and (6) holds, taking n goes to $+\infty$ in (12), we obtain $\sum_{k=0}^{\infty} q^2(x^k, x^{k+1}) < \infty$ because $\sum_{k=0}^{+\infty} \epsilon_k < \infty$. In particular, we have $\lim_{k\to+\infty} q(x^k, x^{k+1}) = 0$. □

Theorem 2 *Consider* $\{x^k\}$ *generated by Algorithm 1. Then, every cluster point of* $\{x^k\}$, *if any, is a critical point of* f.

Proof Let $\hat{x}$ be a cluster point of $\{x^k\}$, and let $\{x^{k_j}\}$ be a subsequence of $\{x^k\}$ converging to $\hat{x}$. Since $w^{k_j} \in \partial h(x^{k_j})$ and $\{x^{k_j}\}$ is bounded, it follows from (Rockafellar and Wets 1998, Theorem 9.13) that $\{w^{k_j}\}$ is also bounded. So, we can suppose that $\{w^{k_j}\}$ converges to a point $\hat{w}$ (one can extract other subsequences if necessary). Hence, combining (5) with Proposition 1, we have

$$\lim_{j\to+\infty} \xi^{k_j+1} = \lim_{j\to+\infty} w^{k_j} = \hat{w}. \tag{13}$$

On the other hand, it follows from definition of the algorithm that $w^{k_j} \in \partial h(x^{k_j})$ and $\xi^{k_j+1} \in \partial g(x^{k_j+1})$. Thus, letting j goes to $+\infty$ in last two inclusions, from (13), we obtain $\hat{w} \in \partial h(\hat{x}) \cap \partial g(\hat{x})$. This means that $\hat{x}$ is a critical point of f, and the proof is completed.□

4.1 Global convergence

Dealing with descent methods for convex functions, we can expect that the algorithm provides globally convergent sequences, i.e., convergence of the whole sequence. When the functions under consideration are neither convex nor quasiconvex, the method may provide sequences that exhibit highly oscillatory behaviors, and partial convergence results are obtained. The Kurdyka–Łojasiewicz property has been successfully applied to analyze various types of asymptotic behavior, in particular, proximal point methods; see for instance Attouch et al. (2013), Bento and Soubeyran (2015b), Frankel et al. (2015).

Global convergence of DCA algorithm (a subgradient-type algorithm) was considered by An et al. (2009) for subanalytic DC functions. A function $f : \mathbb{R}^n \to \mathbb{R}$ is called subanalytic if its graph is a subanalytic subset of $\mathbb{R}^n \times \mathbb{R}$. The set A is called subanalytic if each point of $\mathbb{R}^n$ admits a neighborhood V such that

$$A \cap V = \{x \in \mathbb{R}^n \ : \ (x, y) \in B\},$$

where B is a bounded semianalytic subset of $\mathbb{R}^n \times \mathbb{R}^m$ for some $m \geq 1$. A subset A is called semianalytic if each point of $\mathbb{R}^n$ admits a neighborhood V for which $A \cap V$ assumes the form

as follows

$$\bigcup_{i=1}^{p}\bigcap_{j=1}^{q}\{x \in V \ : \ f_{ij}(x) = 0, \ g_{ij}(x) > 0\},$$

where the functions $f_{ij}, g_{ij} : V \to \mathbb{R}$ are real-analytic for all $i = 1, \ldots, p$ and $j = 1, \ldots, q$. Readers who are unfamiliar with "subanalytic" might in a first reading replace it by "semialgebraic". A set $A \subset \mathbb{R}^n$ is called semialgebric if it assumes the following form:

$$A = \bigcup_{i=1}^{p}\bigcap_{j=1}^{q}\{x \in V \ : \ f_{ij}(x) = 0, \ g_{ij}(x) > 0\},$$

where $f_{ij}, g_{ij} : \mathbb{R}^n \to \mathbb{R}$ are polynomial functions for all $i = 1, \ldots, p$ and $j = 1, \ldots, q$. The class of semialgebraic sets provides an important subclass of subanalytic sets; see for instance An et al. (2009), Bolte et al. (2007).

A function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is said to have the Kurdyka–Łojasiewicz property at $x^* \in \operatorname{dom} \partial f$ if there exist $\eta \in (0, +\infty]$, a neighborhood U of x^* and a continuous concave function $\varphi : [0, \eta) \to \mathbb{R}_+$ (called desingularizing function) such that:

$$\varphi(0) = 0, \quad \varphi \text{ is } C^1 \text{ on } (0, \eta), \quad \varphi'(s) > 0, \ \forall s \in (0, \eta); \tag{14}$$

$$\varphi'(f(x) - f(x^*))\operatorname{dist}(0, \partial f(x)) \geq 1, \quad \forall x \in U \cap [f(x^*) < f < f(x^*) + \eta], \tag{15}$$

where $[\eta_1 < f < \eta_2] = \{x \in \mathbb{R}^n : \eta_1 < f(x) < \eta_2\}$ and C^1 means differentiable with continuous derivative; see the definition of the Kurdyka–Łojasiewicz property and other references about this subject in Attouch et al. (2013).

Remark 5 One can easily check that the Kurdyka–Łojasiewicz property is satisfied at any non-critical point $\widehat{x} \in \operatorname{dom} \partial f$. It follows from the Kurdyka–Łojasiewicz property that the critical points of f lying in $U \cap [f(x^*) < f < f(x^*) + \eta]$ have the same critical value $f(x^*)$. If f is differentiable and $f(x^*) = 0$, then (15) can be rewritten as

$$\nabla(\varphi \circ f)(x) \geq 1,$$

for each convenient $x \in \mathbb{R}^n$. This property basically expresses the fact that a function can be made sharp by a reparameterization of its values; see Attouch et al. (2013).

From Bolte et al. (2007, Theorem 3.1), a subanalytic function f which is continuous when restricted to its closed domain satisfies the Kurdyka–Łojasiewicz property with desingularising function $\varphi(t) = \frac{C}{\theta}t^{\theta}$ with $C > 0$ and $\theta \in (0, 1]$. We prove in the sequel global convergence and rate of convergence of the proximal point method as in An et al. (2009) which deals with a subgradient-type method involving subanalytic DC functions. However, we go further proving the global convergence of the proximal sequence for DC functions that satisfy the Kurdyka–Łojasiewicz property which include the subanalytic DC functions.

From now on in this section, in order to obtain global convergence results, we consider the exact version of Algorithm 1 which means to set $\epsilon_k = 0$, for all $k \geq 0$, under the following assumptions:

A1. f is continuous;
A2. h is continuously differentiable with its gradient Lipschitz with constant $L > 0$;
A3. There exist real numbers $\alpha > 0$ and $\beta > 0$ such that

$$\alpha||x - y|| \leq q(x, y) \leq \beta||x - y||, \quad \forall x, y \in \mathbb{R}^n. \tag{16}$$

Last condition was used to prove convergence of the proximal point algorithm for non-convex and non-smooth functions that verify the Kurdyka–Łojasiewicz property, see Bento and Soubeyran (2015a, b), Moreno et al. (2012). Moreno et al. (2012) present several examples of quasi distances highlighting two that satisfy (16).

Theorem 3 *Let $\{x^k\}$ be the sequence generated by the exact version of Algorithm 1. Assume that conditions A1–A3 hold, $\{x^k\}$ is bounded and f satisfies the Kurdyka–Łojasiewicz property at a cluster point $\hat{x}$ of $\{x^k\}$. Then $\{x^k\}$ converges to $\hat{x}$ which is a critical point of f.*

Proof It is straightforward to check the assertion just applying (Attouch et al. 2013, Theorem 2.9) together with the following facts:

1. It follows from (6) that $0 < c_1 \leq \lambda_k \leq c_2$, for all $k \geq 0$. Hence, for each k, combining (11) with (16), we obtain
$$f(x^{k+1}) + a||x^{k+1} - x^k||^2 \leq f(x^k), \tag{17}$$
where $a := \dfrac{\alpha^2}{2c_2} > 0$;
2. From definition of Algorithm 1 there exists $\xi^{k+1} \in \partial g(x^{k+1})$ such that (5) holds. Note that
$$\xi^{k+1} - \nabla h(x^{k+1}) = \xi^{k+1} - \nabla h(x^k) + \nabla h(x^k) - \nabla h(x^{k+1}),$$
where $\xi^{k+1} - \nabla h(x^{k+1}) \in \partial f(x^{k+1})$. Thus, from triangular inequality, we have
$$\begin{aligned} ||\xi^{k+1} - \nabla h(x^{k+1})|| &\leq ||\xi^{k+1} - \nabla h(x^k)|| + ||\nabla h(x^{k+1}) - \nabla h(x^k)|| \\ &\leq \mu q(x^k, x^{k+1}) + L||x^{k+1} - x^k|| \\ &\leq \mu\beta||x^{k+1} - x^k|| + L||x^{k+1} - x^k|| \\ &= b||x^{k+1} - x^k||, \end{aligned} \tag{18}$$
for $b = (\mu\beta + L)$, where the second inequality is due to (5) and assumption A2, the third inequality comes from (16);
3. Let $\{x^{k_j}\}$ be a subsequence of $\{x^k\}$ converging to $\hat{x}$. Since f is bounded from below it follows from Theorem 1 that $\{f(x^k)\}$ is convergent. Thus, from A1, we have that $\{f(x^{k_j})\}$ converges to $f(\hat{x})$, and so $\{f(x^k)\}$ converges to $f(\hat{x})$.

This complete the proof. □

4.2 Convergence rates

Next, we study the convergence rate of the sequence $\{f(x^k)\}$ depending on the nature of the desingularizing function φ. We prove a similar rates of convergence of the sequence $\{f(x^k)\}$ generated by our proximal method as the ones obtained by An et al. (2009, Theorem 3.3) for the sequence $\{x^k\}$ generated by DCA algorithm. Denote by $r_k := f(x^k) - f(x^*)$, where x^* is the limit point of $\{x^k\}$.

Theorem 4 *Let $\{x^k\}$ be the sequence generated by the exact version of the Algorithm 1. Assume that conditions A1–A3 hold, $\{x^k\}$ is bounded and f satisfies the Kurdyka–Łojasiewicz property at the limit point x^* of $\{x^k\}$ with desingularizing function $\varphi(t) = \frac{C}{\theta}t^\theta$ with $C > 0$ and $\theta \in (0, 1]$. Then, the following estimations hold:*

1. *If $\theta = 1$, the sequence $\{x^k\}$ converges in a finite number of steps;*

2. *If* $\theta \in [\frac{1}{2}, 1)$, *then there exist constants* $c > 0$ *and* $k_0 \in \mathbb{N}$ *such that*

$$f(x^k) - f(x^*) = O\left[exp(-c(k - k_0))\right];$$

3. *If* $\theta \in (0, \frac{1}{2})$, *then there exists* $c > 0$ *such that*

$$f(x^k) - f(x^*) = O\left[c(k-1)^{\frac{-1}{1-2\theta}}\right].$$

Proof Since f is bounded from below from Theorem 1 (with $\epsilon_k = 0$), we have that $\{f(x^k)\}$ converges and, from A1, it converges to $f(x^*)$ because x^k converges to x^*. Hence, we obtain that $\{r_k\}$ is convergent sequence such that $r_k \geq 0$, for all $k \in \mathbb{N}$, and there exist $k_0 \in \mathbb{N}$ and a neighborhood $\mathcal{N}(x^*)$ such that $x^k \in \mathcal{N}(x^*)$, for all $k \geq k_0$, where the Kurdyka–Łojasiewicz property holds. Combining (15) with (17) and (18), we obtain

$$(\varphi')^2(r_{k+1})(r_k - r_{k+1}) \geq \frac{a}{b^2}(\varphi')^2(r_{k+1})||\xi^{k+1} - \nabla h(x^{k+1})|| \geq \frac{a}{b^2}, \tag{19}$$

taking in account that $\xi^{k+1} - \nabla h(x^{k+1}) \in \partial f(x^{k+1})$. If $\theta = 1$, from (19), we have $C^2(r_k - r_{k+1}) \geq \frac{a}{b^2} > 0$ which contradicts the fact that $\{r_k\}$ converges. Therefore, there exists some $k \in \mathbb{N}$ such that $r_k = 0$, and the algorithm terminates in a finite number of steps. Now, assume that $r_k > 0$, for all $k \in \mathbb{N}$. For $\theta \in (0, 1)$, (19) gives

$$C^2 r_{k+1}^{2\theta-2}(r_k - r_{k+1}) \geq \frac{a}{b^2}. \tag{20}$$

Since $r_k \to 0$, if $\theta \in [1/2, 1)$, then $0 < 2 - 2\theta \leq 1$ and we have (enlarging k_0 if necessary) that $r_{k+1}^{2-2\theta} \geq r_{k+1}$, for all $k \geq k_0$. Thus, (20) implies that $r_{k+1} \leq (\frac{1}{1+d})r_k$, where $d = \frac{a}{b^2}$. Using last inequality successively one has

$$r_{k+1} \leq r_{k_0}\left(\frac{1}{1+d}\right)^{k-k_0+1} = r_{k_0}\exp\left[-\log(1+d)(k-k_0+1)\right],$$

and the second statement is proved. Now, assume that $\theta \in (0, 1/2)$. Set $\gamma(t) := \frac{C}{1-2\theta}t^{2\theta-1}$. Then,

$$\gamma(r_{k+1}) - \gamma(r_k) = \int_{r_k}^{r_{k+1}} \gamma'(t)dt = C\int_{r_{k+1}}^{r_k} t^{2\theta-2}dt \geq C(r_k - r_{k+1})r_k^{2\theta-2}. \tag{21}$$

We consider two cases: if $r_{k+1}^{2\theta-2} \leq 2r_k^{2\theta-2}$, combining (21) with (20), we obtain that $\gamma(r_{k+1}) - \gamma(r_k) \geq \frac{a}{2b^2}$. On the other hand, if $r_{k+1}^{2\theta-2} > 2r_k^{2\theta-2}$, since $2\theta - 2 < 2\theta - 1 < 0$, we obtain $\frac{2\theta-1}{2\theta-2} > 0$, and thus, $r_{k+1}^{2\theta-1} > qr_k^{2\theta-1}$, where $q = 2^{\frac{2\theta-1}{2\theta-2}}$. Therefore,

$$\gamma(r_{k+1}) - \gamma(r_k) = C\int_{r_{k+1}}^{r_k} t^{2\theta-2}dt > \frac{C}{1-2\theta}(q-1)r_k^{2\theta-1} \geq \frac{C}{1-2\theta}(q-1)r_{k_0}^{2\theta-1}. \tag{22}$$

Set $\hat{c} := \min\{\frac{a}{2b^2}, \frac{C}{1-2\theta}(q-1)r_{k_0}^{2\theta-1}\}$. Then

$$\gamma(r_{k+1}) - \gamma(r_k) \geq \hat{c}, \quad \forall k \geq k_0.$$

This implies

$$\gamma(r_{k+1}) \geq \gamma(r_{k+1}) - \gamma(r_{k_0}) \geq \sum_{n=k_0}^{k} \gamma(r_{n+1}) - \gamma(r_n) \geq \hat{c}(k - k_0).$$

Then, $r_{k+1} \leq \left[\frac{(1-2\theta)}{C}\hat{c}(k-k_0)\right]^{\frac{1}{2\theta-1}}$ and the desired result is proved. □

Next, we study the rate of convergence of the sequence $\{x^k\}$. Besides the fact that our method is different to the one considered in An et al. (2009) our next result differs to Theorem 3.3 by An et al. (2009) mainly because it does not depend on the nature of the desingularizing function φ.

Theorem 5 *Let $\{x^k\}$ be the sequence generated by the exact version of Algorithm 1. Assume that conditions A1–A3 hold, $\{x^k\}$ is bounded and f satisfies the Kurdyka–Łojasiewicz property at the limit point x^* of $\{x^k\}$ with desingularising function φ. Set $\tilde{\varphi}(t) = \max\{\varphi(t), \sqrt{t}\}$. Then, $||x^k - x^*|| = O\left[\tilde{\varphi}(f(x^{k-1}) - f(x^*))\right]$.*

Proof By assumption x^k converges to x^*, then there exist $K \in \mathbb{N}$ and a neighborhood $\mathcal{N}(x^*)$ such that $x^k \in \mathcal{N}(x^*)$, for all $k \geq K$, where the Kurdyka–Łojasiewicz property holds. We may suppose that $r_k > 0$, for all $k \in \mathbb{N}$, because otherwise the algorithm terminates in a finite number of steps. Taking in account fact on the proof of Theorem 3, it is straightforward to adapt the ideas in the proof of Attouch et al. (2013, Lemma 2.6) to obtain the following estimation:

$$2||x^{k+1} - x^k|| \leq ||x^k - x^{k-1}|| + M[\varphi(f(x^k) - f(x^*)) - \varphi(f(x^{k+1}) - f(x^*))],$$

for some constant $M > 0$. Summing up this inequality for $k = K, \ldots, n$, we have

$$\sum_{k=K}^{n} ||x^{k+1} - x^k|| \leq ||x^K - x^{K-1}|| + M\varphi(r_K).$$

Using the triangular inequality and letting $n \to +\infty$, we obtain

$$\begin{aligned} ||x^K - x^*|| &\leq \sum_{k=K}^{+\infty} ||x^{k+1} - x^k|| \leq ||x^K - x^{K-1}|| + M\varphi(r_K) \\ &\leq \sqrt{\frac{f(x^{K-1}) - f(x^K))}{a}} + M\varphi(r_K) \end{aligned}$$

by (17). Then, using that $\{r_k\}$ is decreasing together with the fact that $r_K \geq 0$, we obtain

$$||x^K - x^*|| \leq \frac{1}{\sqrt{a}}\sqrt{r_{K-1}} + M\varphi(r_{K-1})$$

which gives the desired result. □

Remark 6 Note that combining last two theorems, we obtain to the proximal point method the convergence rate result proved by An et al. (2009, Theorem 3.3) for DCA algorithm.

4.3 Numerical illustration

In this section we present some numerical results to verify the practical efficiency of the proposed algorithm. We especially analyze the exact case, i.e., $\epsilon_k = 0$ for all $k \in \mathbb{N}$. It is well known that the proximal point method is indeed a conceptual scheme for optimization which has been the starting point for other methods. The performance of the method depends essentially on the algorithm used to solve the subproblems. In this situation, it makes little sense to compare the proximal method with other methods in terms of computational efficiency and hence, we skip a discussion of comparison with other algorithms. In this section, algorithm 1 is coded in SCILAB 5.5.2 on a machine with a Intel(R) Core(TM) i5, 1.0 GHz CPU and 6 GB memory. The subproblems are solved using the subroutine "fminsearch".

Example 1 Let $f : \mathbb{R} \to \mathbb{R}$ be a (non-convex) DC function given by

$$f(x) = \frac{x^4}{4} - \frac{x^2}{2} + 1$$

which satisfies the Kurdyka–Łojasiewicz property at $x = 1$ with desingularising function $\varphi(t) = t^{1/2}$. The critical points of f are $\widehat{x}_1 = -1$ and $\widehat{x}_2 = 1$ (global minimum), and $\widehat{x}_3 = 0$ (local maximum). Set $\lambda_k = 1/2$ for all $k \in \mathbb{N}$. Consider the (non-symmetric) quasi distance $q(x, y) = \begin{cases} y - x, & \text{if } x < y, \\ 2(x - y), & \text{if } x \geq y. \end{cases}$ Thus, Algorithm 1 becomes the following: find $x^{k+1} \in \mathbb{R}$ such that

$$0 \in x_{k+1}^3 - x^k + \partial q^2(\cdot, x^k)(x^{k+1}).$$

Since the algorithm stops if $x^{k+1} = x^k$, we can assume that $x^{k+1} \neq x^k$, for all $k \in \mathbb{N}$. Then,

$$\partial q^2(\cdot, x^k)(x^{k+1}) = \begin{cases} -2(x^k - x^{k+1}), & \text{if } x^{k+1} < x^k, \\ 8(x^{k+1} - x^k), & \text{if } x^{k+1} > x^k. \end{cases}$$

Thus, given $x_k \in \mathbb{R}$ the iterative step of Algorithm 1 is the real number solution of one of the following equations:

$$x_{k+1}^3 + 8x_{k+1} - 9x_k = 0 \quad \text{and} \quad x_{k+1}^3 + 2x_{k+1} - 3x_k = 0.$$

If $x_0 \in (0, 1)$, then we have that $\{x_k\}$ is increasing and $x_k \in (0, 1)$, for all $k \in \mathbb{N}$. Hence x_k converges to 1 as $k \to +\infty$. From Theorems 4 and 5, there exists $k_0 \in \mathbb{N}$ such that $|x_k - 1| = O\,[exp(-c(k - k_0 - 1))]$ and $f(x^k) - f(1) = O\,[exp(-c(k - k_0))]$, where $c = \log\left(\frac{10}{9}\right)$ comes from the constants $c_1 = c_2 = 1/2$, $a = \alpha = \mu = L = 1$, $\beta = 2$ and $b = 3$ taken in this example. A similar analysis can be done for different initial and critical points.

Example 2 In this example we consider the 2-dimensional case of last example, i.e., let $f : \mathbb{R}^2 \to \mathbb{R}$ be a (non-convex) DC function given by

$$f(x, y) = \frac{1}{4}(x^4 + y^4) - \frac{1}{2}(x^2 + y^2).$$

The critical points of f are $\widehat{x}_1 = (-1, -1)$, $\widehat{x}_2 = (-1, 1)$, $\widehat{x}_3 = (1, -1)$ and $\widehat{x}_4 = (1, 1)$ (global minimum), and $\widehat{x}_5 = (0, 0)$ (local maximum).

For this problem, we consider 10 randomly generated starting points with coordinates belonging to the box $[-10, 10] \times [-10, 10]$ and $\lambda_k = 5$ for all $k \in \mathbb{N}$. Moreover, we adopt the stopping criteria $||x^{k+1} - x^k|| < 10^{-8}$.

Table 1 summurizes the results. Column x^0 provides the starting point generated, column iter.(k) refers to iteration which the algorithm stopped, column L-time gives the longest time (in seconds) to solve the subproblems of the algorithm, column x^k represents the limit point found by the algorithm and $||x^k - \widehat{x}||$ is the distance of the end-point of the algorithm to the corresponding (shortest) critical point.

In Fig. 1 are plotted the starting points and the iterations of the algorithm. Figure 2 shows the time to obtain the limit point in each of 10 times that the algorithm was performed.

Table 1 Results with randomly generated starting points

$x^0 = (x_1^0, x_2^0)$	Iter. (k)	L-time (s)	$x^k = (x_1^k, x_2^k)$	$\|\|x^k - \widehat{x}_i\|\|$
(− 4.387004, − 7.4398831)	12	0.51	(− 1.0000012, − 1.0000024)	0.0000027
(5.5662572, − 5.7619391)	12	0.82	(1.0000071, − 1.0000417)	0.0000423
(− 7.7572907, 3.7137919)	13	0.85	(− 1.0000078, 1.0000285)	0.0000295
(− 6.9375666, 3.9417012)	12	0.78	(− 1.0000416, 0.9999977)	0.0000417
(6.8310369, − 1.8759505)	13	0.91	(1.000017, − 0.9999906)	0.0000194
(− 1.810349, 7.5682516)	12	0.88	(− 0.9999647, 1.0000496)	0.0000609
(− 7.7232806, − 6.0033245)	13	0.61	(− 1.0000218, − 0.9999945)	0.0000225
(1.2373215, 1.7923547)	11	0.68	(0.9999910, 1.000021)	0.0000228
(3.7079593, 7.8124495)	13	0.94	(1.0000161, 0.9999911)	0.0000184
(0.0844256, −3.0127692)	13	0.92	(1.0000032, − 1.000005)	0.0000059

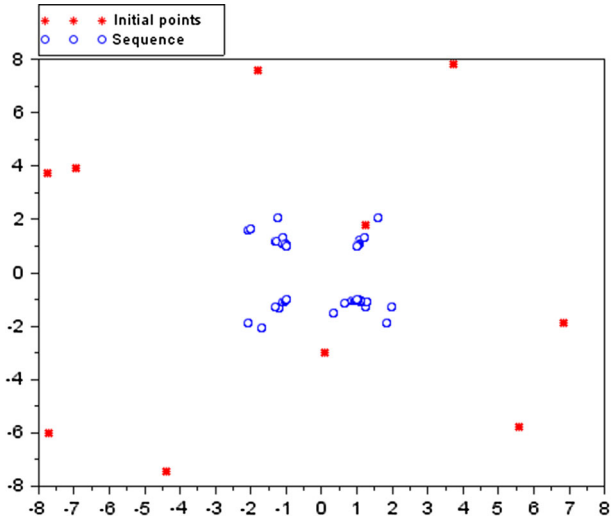

Fig. 1 Algorithm running with 10 randomly starting points

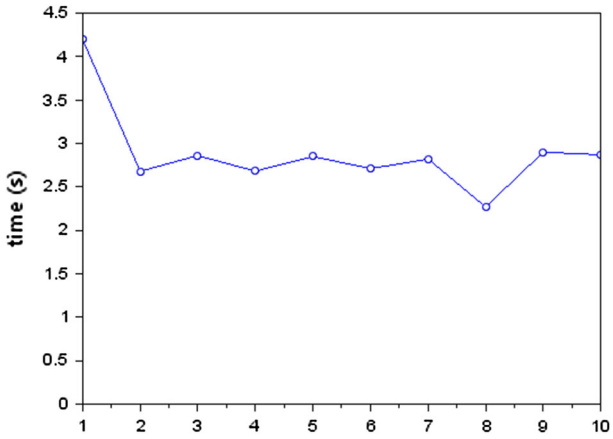

Fig. 2 Time of convergence

5 Application: the optimal size of the firm problem

One of the main topic in Economic and Management Sciences is to determine the optimal size of an organization. This is a difficult problem, both for conceptual and technical reasons. This optimal size can refer to the quantity of the final good produced, the range of different final goods that the multi-product firm produces, the number and quality of workers of different types employed, the amount of means used in the production process, as well as the number of intermediate stages in the production process and their different locations in different countries in the globalization process. A huge literature exists and a lot of aspects must be examined. This literature separates two cases:

(i) the easy and rather irrealistic case of decreasing returns, when costs of production of the firm are convex;

(ii) the much more difficult but realistic case of increasing returns, when costs of production are concave.

Because the reader can be either not fully trained in Behavioral Sciences or in Variational Analysis and Optimizing Algorithms, we start to make clear in the easy case of convex costs of production what is the problem of the optimal size of a firm and build a simple connection between this application and a standard proximal algorithm. A next example will consider the much more interesting case of concave costs of production which is fully adapted to the present DC proximal algorithm. Consider the simplest linear-quadratic static formulation of this problem (decreasing returns). In this case the revenue function $r = px$ of the firm grows linearly with the quantity produced of a final good x (sold at the given price $p > 0$) while the cost of production function $c(x) = wx + \mu x^2$ grows quadratically in the quantity produced x. The first term wx refer to the sum of each given individual wage $w > 0$ paid to each employed worker, where each employed worker produces one unit of the final good. Hence the number of employed workers is equal to the number of units of the final good produced x. The second term μx^2, $\mu > 0$, of the cost function refer to costs of using means. In this very simple case the optimal size of employed workers and the optimal quantity produced of the final good exist trivially. The neoclassical theory of the firm supposes that the entrepreneur, knowing his linear/quadratic revenue and cost functions, can calculate this optimal size $x^* = (p-w)/2\mu > 0$, via the maximization of his profit function $\pi(x) = px - c(x)$. When this is not the case (in a bounded rationality context, Simon 1955), the firm can be initially at $x \neq x^*$, out of its optimal size. Then, its initial size x (number of hired workers, or units of the final good produced) can be too big $x > x^*$ (or too low, $x < x^*$), the entrepreneur hiring too much (not enough) workers in the initial period. The (VR) variational rationality approach (Soubeyran 2009, 2010), considers this out of equilibrium starting case as the starting point of a possibly convergent process $x = x^k \curvearrowright y = x^{k+1}$ towards the optimal size, where, each current period $k+1$, the entrepreneur hires $y - x > 0$, fires $x - y < 0$ workers, or continues using $y = x$ workers, with hiring/firing costs $C(x, y) \geq 0$ of being able to change the number of employed workers (the size of the firm). Hence the proximal payoff of the entrepreneur becomes $P_\xi(y/x) = \pi(y) - \xi C(x, y)$, where $\xi > 0$ weights the relative importance of per period costs of being able to change $C(x, y)$ relative to the per period static profit. This crude example shows very succinctly how the (VR) approach makes the connection between the problem of convergence to the optimal size of the firm and the exact proximal point algorithm, which solves, at each period the subproblem $\sup\left\{P_{\xi_{k+1}}(y/x),\, y \in X = \mathbb{R}_+\right\}$, where $x = x^k$, $y = x^{k+1}$ and $\xi = \xi_{k+1}$.

In this last section we will consider, in a dynamic setting, the most realistic but difficult case where the firm is supposed to exhibit increasing returns, when (execution) costs of production are concave. Its size refers to the production level, i.e, the number of units of the final good the firm produces. Then, using the recent variational rationality approach (Soubeyran 2009, 2010, 2016), we will determine the long run optimal size of this firm, when it can, each period, hire, fire and keep again workers. This offers an original and dynamic theory of the limit of the firm.

5.1 A simple model of the firm with increasing returns in the short run

5.1.1 An example of "to be increased" and "to be decreased" payoffs

To better see how DC optimization works in applications, let us examine the simplest case we can imagine, which can be generalized to the multidimensional setting. Different variants

of this example can be found in Soubeyran (2009), Soubeyran (2010), Bento and Soubeyran (2015a, b) and Bao et al. (2015). But none of them examine the very important case of increasing returns, which is the realistic case for production costs, as we do here, as an application of the proximal point method for DC optimization. Consider a hierarchical firm including an entrepreneur, a profile of workers, and a succession of periods where the entrepreneur can hire, fire or keep working workers in a changing environment. Each period, the entrepreneur chooses how much to produce of the same final good (of a given quality) and sells each unit of this good at the same and fixed price $p > 0$. In the current period, the firm produces $x \in \mathbb{R}_+$ units of a final good and employs $l(x) \in \mathbb{R}_+$ workers. In this simple model the size of the firm refers to x. For simplification, each worker is asked to produce one unit of the final good. Then, $l(x) = x$. The current profit of the entrepreneur, $\pi(x) = r(x) - c(x)$, is the difference between the revenue of the firm $r(x) = px \geq 0$ and the cost of production $c(x) \geq 0$. To produce one unit of the final good, each employed worker must use a given bundle of individual means (tools and ingredients) and a fixed collective mean (say, some given piece of land, a given infrastructure). The entrepreneur rents the durable tools and buys the non-durable ingredients. Let $\overline{\pi} = \sup\{\pi(y), y \in X\} < +\infty$ be the highest profit the entrepreneur can hope to achieve. Then, $f(x) = \overline{\pi} - \pi(x) \geq 0$ is the current unrealized profit he can hope to carry out in the current period, or later. The profit function $\pi(\cdot)$ is a "to be increased" payoff, while the unrealized profit function $f(\cdot)$ is a "to be decreased" payoff.

In the mathematical part of the paper, the objective function $f(x) = g(x) - h(x)$ is the difference between two convex functions, $g(x)$ and $h(x)$. In our behavioral example, $f(x)$ represents the unrealized profit the entrepreneur can hope to achieve, i.e, $f(x) = \overline{\pi} - \pi(x) = \overline{\pi} + c(x) - r(x)$, where $g(x) = \overline{\pi} + c(x)$ and $h(x) = r(x)$. Then, the cost and the revenue functions $r(\cdot)$ and $c(\cdot)$ must be concave to fit with the mathematical part of the paper.

Clearly, in a perfectly competitive market where the price p of the final good is a given, the revenue function $r(\cdot) : x \in \mathbb{R}_+ \longmapsto r(x) = px$ is linear, hence concave with respect to the production level x. What we have to make clear is why a cost function $c(\cdot)$ is usually concave in the short run. But, to escape to mathematical difficulties, textbooks in Economics focus the attention on the less usual case of convex costs of production in the short run. Costs of production are concave when the technology of the firm exhibits increasing returns, coming from economies of scale, economies of specialization, learning by doing several times the same thing, limited capacities, lack of time to be able to change fixed costs in the short run, which become variable costs in the long run. In our standard model of the firm, costs of production $c(x) = wx + hx + K$ are the sum of three different costs, where, (i) $w > 0$ is the given wage paid to each employed worker, (ii) $h > 0$ is the price paid to suppliers to acquire each bundle of means used by each employed worker to produce one unit of the final good, and, (iii) $K > 0$ is the cost to rent a durable, fixed, collective and indivisible mean.

This cost of production exhibits increasing returns to scale because, in the current period, before production takes place, the fixed costs $K > 0$ must be paid even if, later, no worker are required to work, i.e, $c(0) = K > 0$. This implies that the unit cost of production $c(x)/x = w + h + K/x$ decreases when the production level x increases. The cost of production will be strictly concave if, for example, the price $h = h(x)$ of each bundle of means used by each employed worker decreases with the number x of bundles of means the entrepreneur must buy to produce x units of the final good (when suppliers offer discounts).

5.2 The variational rationality approach: the simplest formulation

5.2.1 Stay/stability and change dynamics

The (VR) variational rationality approach (Soubeyran 2009, 2010) modelizes and unifies a lot of different models of stay and change dynamics which appeared in Behavioral Sciences (Economics, Management Sciences, Psychology, Sociology, Political Sciences, Decision theory, Game theory, Artificial Intelligence, etc.); see Georgeff et al. (1998), Wooldridge (2000). Stays refer to exploitation phases, temporary repetitions of the same action, temporary habits, routines, rules and norms, etc. while changes refer to exploration phases, learning and innovations processes, forming and breaking habits and routines, changing doings (actions), havings and beings, etc. This dynamical approach considers entities (an agent, an organization or several interacting agents), which are, at the beginning of the story, in an undesirable initial position, and are unable to reach immediately a final desired position. The goal of this approach is to examine the transition problem: how such entities can find, build and use an acceptable and feasible transition which is able to overcome a lot of intermediate obstacles, difficulties and resistance to change, with not too much intermediate sacrifices and enough intermediate satisfactions to sustain motivation to change and persevere until reaching the final desired position. This (VR) approach admits a lot of variants, based on the same short list of general principles and concepts. The four main concepts refer to changes and stays, worthwhile changes and stays, worthwhile transitions and variational traps, worthwhile to approach and reach but not worthwhile to leave. A stay and change dynamic refers to a succession of periods, where $k+1$ is the current period and k is the past period, where $x = x^k \in X$ can be a past action (doing), having or being and $y = x^{k+1} \in X$ can be a current action (doing), having or being. A single change from $x = x^k \in X$ to $y = x^{k+1} \in X$ is $x \curvearrowright y$, $y \neq x$. A single stay at x is $x \curvearrowright y$, $y = x$.

Let us give, starting from our previous example, the simplest prototype of the (VR) variational rationality approach, to finally show how, at the end of a worthwhile transition, a firm can achieve an optimal size.

5.2.2 Worthwhile changes

The (VR) approach starts with the following broad definition of a worthwhile change: a change is worthwhile if motivation to change rather than to stay is "high enough" with respect to resistance to change rather than to stay. This definition allows a lot of variants, as much variants as the definitions of motivation (more than one hundred theories of motivations exist in Psychology), resistance (which includes a lot of different aspects) and "high enough" (see Soubeyran 2009, 2010). Let us give a very simple formulation of the worthwhile to change concept.

In the previous example a change refers to a move from having produced $x \in X = \mathbb{R}_+$ units of a final good in the previous period to produce $y \in \mathbb{R}_+$ units of this final good in the current period. A stay is a particular move, from having produced a given quantity $x = x^k$ of the final good in the previous period to produce again the same quantity $y = x^{k+1} = x^k$ of this final good in the current period. The previous and current "to be increased" payoffs of the entrepreneur are the profit $\pi(x)$ and $\pi(y)$. His previous and current "to be decreased" payoffs are his unrealized profits $f(x) = \overline{\pi} - \pi(x) \geq 0$ and $f(y) = \overline{\pi} - \pi(y) \geq 0$.

Advantages to change from x to y, if they exist, represent the difference in profits or unrealized profits, $A(x, y) = \pi(y) - \pi(x) = f(x) - f(y) \geq 0$.

Inconveniences to change from x to y refer to the difference $I(x, y) = C(x, y) - C(x, x) \geq 0$.

$C(x, y) \geq 0$ modelizes the costs of being able to change from x to y. In the present model $C(x, y)$ modelizes costs of hiring, firing and keeping working workers to be able to move from producing x units of the final good, to produce y units of the final good, where y can be higher, lower or the same than x. Costs of hiring $y - x > 0$ workers are $C(x, y) = \rho^+(y - x)$, where $\rho^+ > 0$ is the cost of hiring one worker. Costs of firing $x - y > 0$ workers are $C(x, y) = \rho^-(x - y)$, where $\rho^- > 0$ is the cost of firing one worker. Costs of keeping working $y = x$ workers are $C(x, x) = \rho^= x$, where $\rho^= \geq 0$ is the cost of keeping working one period more one worker. For simplification (this will require a too long discussion), we will suppose that $\rho^= = 0$. Then, $C(x, x) = 0$, and inconveniences to change are

$$I(x, y) = C(x, y) = \left\{ \begin{array}{l} \rho^+(y - x) \text{ if } y \geq x \\ \rho^-(x - y) \text{ if } y \leq x \end{array} \right\} \geq 0.$$

Motivation to change $M(x, y) = U[A(x, y)]$ is the utility $U[A]$ of advantages to change $A = A(x, y) \geq 0$.

Resistance to change $R(x, y) = D[I(x, y)]$ is the disutility $D[I]$ of inconveniences to change $I = I(x, y) \geq 0$, where the utility function $U[\cdot] : A \in \mathbb{R}_+ \longmapsto U[A] \in \mathbb{R}_+$ and the disutility function $D[\cdot] : I \in R_+ \longmapsto D[I] \in \mathbb{R}_+$ are strictly increasing and zero at zero.

A worthwhile change from x to y is such that motivation to change $M(x, y) \in \mathbb{R}_+$ from x to y is higher than resistance to change $R(x, y)$ from x to y, up to a chosen worthwhile to change satisfaction ratio $\xi > 0$, i.e., such that $M(x, y) \geq \xi R(x, y)$.

In the example, the utility $U[A]$ of advantages to change and the disutility $D[I]$ of inconveniences to change are linear-quadratic, i.e., $M = U[A] = A$, $R = D[I] = I^2$ (see Soubeyran 2009, 2010 for more general cases). In this context, a change $x \curvearrowright y$ from producing again the quantity x of the final good to produce a different quantity y of this final good is worthwhile if advantages to change are high enough with respect to resistances to change, i.e., $A(x, y) = \pi(y) - \pi(x) = f(x) - f(y) \geq \xi R(x, y) = \xi C(x, y)^2$, where $C(x, x) = 0$. What is "high enough" is defined by the size of $\xi > 0$.

5.2.3 Worthwhile transitions

A transition is a succession of single stays and changes $x^0 \curvearrowright x^1 \curvearrowright ....x^k \curvearrowright x^{k+1} \curvearrowright ....$ where $x^{k+1} \neq x^k$ or $x^{k+1} = x^k$ for each $k \in \mathbb{N}$.

A worthwhile transition is a transition such that each stay or change is worthwhile, i.e., $x^{k+1} \in W_{\xi_{k+1}}(x^k)$, $k \in \mathbb{N}$, that is,

$$\begin{aligned} A(x^k, x^{k+1}) &= \pi(x^{k+1}) - \pi(x^k) \\ &= f(x^k) - f(x^{k+1}) \\ &\geq \xi_{k+1} R(x^k, x^{k+1}) \\ &= \xi_{k+1} C(x^k, x^{k+1})^2, \ k \in \mathbb{N}. \end{aligned}$$

5.2.4 Ends as variational traps

A (strong) variational trap $x^* \in X$ is both, (i) an aspiration point $x^* \in W_{\xi_{k+1}}(x^k)$, $k \in \mathbb{N}$, worthwhile to reach from any position of the transition, (ii) a stationary trap $W_{\xi_*}(x^*) = \{x^*\}$,

where it is not worthwhile to move to any other position $y \neq x^*$, given that the worthwhile to change ratio tends to a limit, $\xi_{k+1} \to \xi_* > 0, k \to +\infty$, and finally (iii) worthwhile to approach, i.e., which converges to the aspiration point. More explicitly, x^* is a variational trap if,

(i) $A(x^k, x^*) = \pi(x^*) - \pi(x^k) = f(x^k) - f(x^*) \geq \xi_{k+1} R(x^k, x^*) = \xi_{k+1} C(x^k, x^*)^2$, $k \in \mathbb{N}$;
(ii) $A(x^*, y) = \pi(y) - \pi(x^*) = f(x^*) - f(y) < \xi_* R(x^*, y) = \xi_* C(x^*, y)^2$, for all $y \neq x^*$;
(iii) It is a limit point of the worthwhile transition, i.e., $x^k \to x^*, k \to +\infty$.

A weak variational trap does not require to be an aspiration point.

5.3 Proximal algorithms as worthwhile transitions

To show how an exact or inexact proximal algorithm can be seen as a leading example of a worthwhile transition, the present paper uses a specific formulation of the (VR) approach, where the utility of advantages to change and the disutility of inconveniences to change are linear quadratic, i.e., $M = U[A] = A$ and $R = D[I] = I^2 = C^2$, where $C(x, y) = q(x, y) \geq 0$ is a quasi distance (see Moreno et al. (2012) for this linear quadratic case, and Bento and Soubeyran (2015a, b) for more general cases).

5.3.1 The proximal formulation of a worthwhile change

In a linear quadratic setting, motivation and resistance to change are $M(x, y) = A(x, y) = \pi(y) - \pi(x) = f(x) - f(y)$ and $R(x, y) = q(x, y)^2$. These simplifications allow to define,

(1) a proximal "to be increased" payoff $P_\xi(y/x) = \pi(y) - \xi R(x, y)$, which is the difference between the current "to be increased" payoff $\pi(y)$ and the weighted current resistance to change $R(x, y)$, where the weight $\xi > 0$ balances the importance of the current 'to be increased" payoff and the current resistance to change.
(2) a proximal "to be decreased" payoff $Q_\xi(y/x) = f(y) + \xi R(x, y)$, which is the sum of the current 'to be decreased" payoff $f(y)$ and the weighted current resistance to change $R(x, y)$.

Then, a change $x \curvearrowright y \in W_\xi(x)$ is worthwhile if moving from x to y, the proximal 'to be increased" payoff increases, $P_\xi(y/x) \geq P_\xi(x/x)$, and the proximal "to be decreased" payoff decreases, $Q_\xi(y/x) \leq Q_\xi(x/x)$. This comes from the following equivalences

$$\begin{aligned} y \in W_\xi(x) &\iff M(x, y) \geq \xi R(x, y) \\ &\iff \pi(y) - \pi(x) = f(x) - f(y) \geq \xi R(x, y) \\ &\iff P_\xi(y/x) \geq P_\xi(x/x) \\ &\iff Q_\xi(y/x) \leq Q_\xi(x/x). \end{aligned}$$

5.3.2 Exact and inexact proximal algorithms as examples of worthwhile transitions

A transition is a succession of single stays and changes $x^0 \curvearrowright x^1 \curvearrowright ....x^k \curvearrowright x^{k+1} \curvearrowright ....$ where $x^{k+1} \neq x^k$ or $x^{k+1} = x^k$ for each $k \in \mathbb{N}$.

A worthwhile transition is a transition such that each stay or change is worthwhile, i.e., in term of proximal payoffs to change,

$$x^{k+1} \in W_{\xi_{k+1}}(x^k) = \left\{ \begin{array}{c} y \in X, \text{ such that} \\ P_{\xi_{k+1}}(y/x^k) \geq P_{\xi_{k+1}}(x^k/x^k), \text{ i.e.,} \\ \pi(y) - \xi_{k+1} R(x_n, y) \geq \pi(x^k), \text{ or,} \\ Q_{\xi_{k+1}}(y/x^k) \leq Q_{\xi_{k+1}}(x^k/x^k), \text{ i.e.,} \\ f(y) + \xi_{k+1} R(x^k, y) \leq f(x^k) \end{array} \right\},$$

where each $\xi_{k+1} > 0, k \in \mathbb{N}$ can be chosen and $R(x^k, y) = q(x^k, y)^2$.

In the context of this paper,

$$x^{k+1} \in W_{\xi_{k+1}}(x^k) \Longleftrightarrow f(x^{k+1}) + \xi_{k+1} q(x^k, x^{k+1})^2 \leq f(x^k),$$

where $\xi_{k+1} = 1/2\lambda_k > 0$.

A worthwhile change is exact if $x^{k+1} \in \arg\max \left\{ P_{\xi_{k+1}}(y/x_n), \ y \in X \right\}$.

An inexact worthwhile change is any worthwhile change "close enough" to an exact worthwhile change, where the term "close enough" can have several different interpretations, depending of chosen reference points and frames. In this paper "close enough" is given by condition (3) and (5) given (2) and (4). The explicit justifications follow Bento and Soubeyran (2015a).

5.4 Surrogate proximal algorithms as worthwhile transitions

Usually the entrepreneur does not know the whole profit function $\pi(\cdot)$. Then, he must perform, each current period $k+1$, an approximate evaluation of this function $\widetilde{\pi}(\cdot/x)$, where $x = x^k$. This requires to consider a more complex formulation of the (VR) approach, where past experience and current evaluations $\widetilde{\pi}(\cdot/x)$ of the payoff functions $\pi(\cdot)$ are included in the worthwhile to change process (see Soubeyran 2016). In the present paper, we will discard the role of past experience to focus our attention on the current evaluation process, when the entrepreneur knows from the very beginning the whole revenue function $r(\cdot) = -g(\cdot)$, but does not know very well the execution cost function $c(\cdot)$. Then, he needs, each period, to make an approximate evaluation of the execution cost function $c(\cdot)$, in term of a simple function $\widetilde{c}(\cdot/x^k)$, which over-estimates globally this cost function $c(\cdot) = -h(\cdot),$, i.e., $\widetilde{c}(y/x^k) \geq c(y)$, for all $y \in X$ with $\widetilde{c}(x^k/x^k) = c(x^k)$. Then, the surrogate evaluation function $\widetilde{\pi}(./x) : y \in X \longmapsto \widetilde{\pi}(y/x) = r(x) - \widetilde{c}(y/x^k)$ under-estimates the "to be increased" profit function $\pi(\cdot) = r(\cdot) - c(\cdot)$, because $\widetilde{\pi}(y/x) \leq \pi(y)$ for all $y \in X$ and $\widetilde{\pi}(x/x) = \pi(x)$.

Similarly, the surrogate evaluation function $\widetilde{f}(\cdot/x) : y \in X \longmapsto \widetilde{f}(y/x) = g(y) - \widetilde{h}(y/x)$ over-estimates the "to be decreased" profit function $f(\cdot) = g(\cdot) - h(\cdot)$, because $\widetilde{f}(y/x) \geq f(y)$, for all $y \in X$, with $\widetilde{f}(x/x) = f(x)$, where $\widetilde{h}(y/x) \leq h(y)$, for all $y \in X$, with $\widetilde{h}(x/x) = h(x)$.

To fit with the mathematical part of the paper, we will suppose that, in the current period $k+1$, the entrepreneur knows the resistance to change function $R(x, y)$. Then, given this knowledge structure, where, each period, the entrepreneur is allowed to make an under-estimation of his profit function $\widetilde{\pi}(\cdot/x)$, a change $x = x^k \curvearrowright y$ is worthwhile if $\widetilde{A}(x, y) = \widetilde{\pi}(y/x) - \pi(x) = f(x) - \widetilde{f}(y/x) \geq \xi R(x, y)$.

The proximal version of this worthwhile to change condition is

$$\widetilde{P}_\xi(y/x) = \widetilde{\pi}(y/x) - \xi R(x, y) \geq \pi(x) = \widetilde{P}_\xi(x/x),$$

or

$$\widetilde{Q}_\xi(y/x) = \widetilde{f}(y/x) + \xi R(x, y) \leq f(x) = \widetilde{Q}_\xi(x/x).$$

This behavioral evaluation process fits with the mathematical part of the paper, which uses a convex-concave procedure (see Yuille and Rangarajan 2003) in the context of DC programming (see also Horst and Thoai 1999).

5.5 Ends

5.5.1 When critical points are variational traps

A weak variational trap is both a limit point of a worthwhile transition, and a stationary trap not worthwhile to leave. This modelizes the approach, and the end of a worthwhile stay and change process. Usually, a critical point is not a stationary trap. Then, the last question of this paper is: when critical points of the exact and inexact proximal algorithm are variational traps?

Definition 3 A function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is said to be weakly convex if there exists $\rho > 0$ such that for all $x, y \in \mathbb{R}^n$ and $t \in [0, 1]$

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y) + \rho t(1-t)||x - y||^2. \tag{23}$$

The function f is said to be locally weakly convex at x if there exists $\epsilon > 0$ such that f is weakly convex on $B(x, \epsilon)$. It is locally weakly convex if it is locally weakly convex at every point of its domain.

Proposition 2 *Let f be a weakly convex function. If x^* is a critical point of f, then*

$$f(x^*) \leq f(y) + \frac{\rho}{\alpha^2} q^2(x^*, y) \quad \forall y \in \mathbb{R}^n, \tag{24}$$

where $\alpha > 0$ satisfies (16).

Proof It follows from Vial (1983, Proposition 4.8) together with (16). □

Proposition 3 *Let f be a weakly convex function. If x^* is a critical point of f and $\lambda > \frac{\rho}{\alpha^2}$, then $W_\lambda(x^*) = \{x^*\}$.*

Proof From (24) and $\lambda > \frac{\rho}{\alpha^2}$, we have

$$f(x^*) \leq f(y) + \frac{\rho}{\alpha^2} q^2(x^*, y) < f(y) + \lambda q^2(x^*, y) \quad \forall y \neq x^*.$$

The result follows from last inequality and definition of $W_\lambda(x)$. □

Remark 7 It is obvious that convex functions are weakly convex. Moreover, $C^{1,1}$ functions and lower-C^2 functions are locally weakly convex and locally Lipschitz weakly convex, respectively. It is known that $C^{1,1}$ functions and lower-C^2 functions are DC functions.

5.5.2 The optimal size of the firm

In the example a variational trap $x^* \in X$ defines an optimal size of the firm where the entrepreneur hires and fires less and less workers to finally stops to hire and fire workers, when resistance to change wins motivation to change. This offers an original theory of the limit of the firm in term of the VR approach, where the entrepreneur optimizes at the end, and satisfies with not too much sacrifices during the transition.

5.6 Speed of convergence

Let $\widehat{x}$ be a Fréchet critical point of f, i.e., $0 \in \hat{\partial} f(\widehat{x})$. Assume that f has a closed domain and restricted to its domain f is continuous. The Kurdyka–Łojasiewicz inequality for a subanalytic function f at $\widehat{x}$ becomes

$$|f(x) - f(\widehat{x})|^{1-\theta} \leq L \|s\|, \tag{25}$$

for all $x \in V$, and $s \in \hat{\partial} f(x)$; see Bolte et al. (2007, Theorem 3.1). It shows that convergence of the norm of the marginal profit $\|s\|$ to zero, $s \in \hat{\partial} f(x)$, implies convergence of the profit $f(x)$ to the critical one $f(\widehat{x})$. Then, Theorems 4 and 5 are very important for applications because they give us informations on the speed of convergence. In our specific example, they tell us how quickly the firm approach its optimal size $\widehat{x}$. Indeed, (25) means that, for a given $x \in V$, the larger the distance from $f(x)$ to $f(\widehat{x})$, the larger is the norm of the subgradient $\|s\|$, for $s \in \hat{\partial} f(x)$. This is a curvature hypothesis. This means that function f is sharp enough close to the critical point. Even more, the higher θ is (moving from $\theta \in (0, 1/2)$ to $\theta \in [1/2, 1)$ to $\theta = 1$), the sharper is this function. Intuition shows that, the higher θ is, the higher is the speed of convergence. Theorem 4 confirms very precisely this intuition. The sharper is the profit function close to its critical size, the speedy is convergence to its critical size: when $\theta = 1$, convergence in a finite number of periods is a very nice and realistic result. When $\theta \in [1/2, 1)$, convergence is of the exponential form. Hence speed of convergence is high. The higher the constant $d = a/b^2$, the speedy is the convergence. This is the case when $a = \alpha^2/(2c_2)$ is high. That is, when resistance to change $R(x, y) = q(x, y)$) is high enough (α high in (16), namely, $\alpha \|y - x\| \leq q(x, y) \leq \beta \|y - x\|$) and, for a given advantage to change $f(x) - f(y) \geq 0$, motivation to change $M_k(x, y) = \lambda_k [f(x) - f(y)]$ is low enough (c_2 low requires λ_k low, from $c_1 \leq \lambda_k \leq c_2$). This is also the case when $b = \mu\beta + L$ is low. That is, for example, when β is low (a not too high resistance to change). The last case $\theta \in (0, 1/2)$ works as well.

6 Conclusions

We presented a generalized proximal linearized algorithm for finding critical points of a DC function (difference of two convex functions) and some convergence rate results. We also provided an application, in a dynamic setting, to determine the limit of the firm, when increasing returns matter in the short run. Future research will examine the case where the concave revenue function is not well known, while the concave execution cost function is perfectly known to the entrepreneur. The case of a more changing environment can also be considered. Finally multi-objective DC programming must be examined to consider the limit of firms which produce different final products. This is the main realistic case.

Acknowledgements We would like to thank the referee for his/her constructive remarks which allow us to improve our work.

References

Ahn, S., Fessler, J. A., Blatt, D., & Hero, A. O. (2006). Convergent incremental optimization transfer algorithms: Application to tomography. *IEEE Transactions on Medical Imaging*, *25*, 283–296.

An, L. T. H., Ngai, H. V., & Pham, D. T. (2009). Convergence analysis of DC algorithm for DC programming with subanalytic data. Ann. Oper. Res., Technical Report, LMI, INSA-Rouen.

Attouch, H., Bolte, J., & Svaiter, B. F. (2013). Convergence of descent methods for semi-algebraic and tame problems: Proximal algorithms, forward–backward splitting, and regularized Gauss–Seidel methods. *Mathematical Programming*, *137*(1–2), 91–129.

Bačák, M., & Borwein, J. M. (2011). On difference convexity of locally Lipschitz functions. *Optimization*, *60*, 961–978.

Bao, T., Mordukhovich, B. S., & Soubeyran, A. (2015). Variational analysis in psychological modelling. *Journal of Optimization Theory and Applications*, *164*, 290–315.

Bao, T. Q., Cobzaş, S., & Soubeyran, A. (2016a). Variational principles, completeness and the existence of traps in behavioral sciences. *Annals of Operations Research*, *269*(1–2), 53–79.

Bao, T. Q., Khanh, P. Q., & Soubeyran, A. (2016b). Variational principles with generalized distances and the modelization of organizational change. *Optimization*, *65*(12), 2049–2066.

Bento, G. C., Cruz Neto, J. X., Lopes, J., Soares, Jr P., & Soubeyran, A. (2016). Generalized proximal distances for bilevel equilibrium problems. *SIAM Journal on Optimization*, *26*(1), 810–830.

Bento, G. C., & Soubeyran, A. (2015a). Generalized inexact proximal algorithms: Routines formation with resistance to change, following worthwhile changes. *Journal of Optimization Theory and Applications*, *166*, 172–187.

Bento, G. C., & Soubeyran, A. (2015b). A generalized inexact proximal point method for nonsmooth functions that satisfies Kurdyka-Łojasiewicz inequality. *Set-Valued and Variational Analysis*, *23*, 501–517.

Bolte, J., Daniliidis, A., & Lewis, A. (2007). Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamic systems. *SIAM Optimization*, *17*(4), 1205–1223.

Bolte, J., Sabach, S., & Teboulle, M. (2014). Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, *146*, 459–494.

Brattka, V. (2003). Recursive quasi-metric spaces. *Theoretical Computer Science*, *305*, 17–42.

Burachik, R. S., & Svaiter, B. F. (2001). A relative error tolerance for a family of generalized proximal point methods. *Mathematics of Operations Research*, *26*(4), 816–831.

Chen, G., & Teboulle, M. (1993). Convergence analysis of proximal-like optimization algorithm using Bregman functions. *SIAM Journal on Optimization*, *3*, 538–543.

Cruz Neto, J. X., Oliveira, P. R., Souza, S. D. S., & Soubeyran, A. (2010). A proximal algorithm with separable Bregman distances for quasiconvex optimization over the nonnegative orthant. *European Journal of Operation Research*, *201*(2), 365–376.

Daubechies, I., Defrise, M., & De Mol, C. (2004). An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics*, *57*, 1413–1457.

Eckstein, J. (1993). Nonlinear proximal point algorithms using Bregman functions, with applications to convex programming. *Mathematics of Operations Research*, *18*, 202–226.

Erdogan, H., & Fessler, J. A. (1999). Ordered subsets algorithms for transmission tomography. *Physics in Medicine and Biology*, *44*, 2835–2851.

Fernández Cara, E., & Moreno, C. (1988). Critical point approximation through exact regularization. *Mathematics of Computation*, *50*, 139–153.

Frankel, P., Garrigos, G., & Peypouquet, J. (2015). Splitting methods with variable metric for Kurdyka–Łojasiewicz functions and general convergence rates. *Journal of Optimization Theory and Applications*, *165*, 874–900.

Gasso, G., Rakotomamonjy, A., & Canu, S. (2009). Recovering sparse signals with non-convex penalties and DC programming. *IEEE Transactions on Signal Processing*, *57*, 4686–4698.

Georgeff, M., Pell, B., Pollack, M., Tambe, M., & Wooldridge, M. (1998). The belief–desire–intention model of agency. In *International workshop on agent theories, architectures, and languages* (pp. 1–10). Berlin, Heidelberg: Springer.

Goldfarb, D., Ma, S., & Scheinberg, K. (2013). Fast alternating linearization methods for minimizing the sum of two convex functions. *Mathematical Programming*, *141*, 349–382.

Hare, W., & Sagastizábal, C. (2009). Computing proximal points of nonconvex functions. *Mathematical Programming*, *116*, 221–258.

Hartman, P. (1959). On functions representable as a difference of convex functions. *Pacific Journal of Mathematics*, *9*, 707–713.

Hiriart-Urruty, J. B. (1986). Generalized differentiability, duality and optimization for problems dealing with diference of convex functions, convexity and duality in optimization. *Lecture Notes in Economics and Mathematical Systems*, *256*, 37–70.

Horst, R., & Thoai, N. V. (1999). DC programming: Overview. *Journal of Optimization Theory and Applications*, *103*(1), 1–43.

Kaplan, A., & Tichatschke, R. (1998). Proximal point methods and nonconvex optimization. *Journal of Global Optimization*, *13*, 389–406.
Kiwiel, K. C. (1997). Proximal minimization methods with generalized Bregman functions. *SIAM Journal on Control and Optimization*, *35*, 1142–1168.
Kiwiel, K. C., Rosa, C. H., & Ruszczyski, A. (1999). Proximal decomposition via alternating linearization. *SIAM Journal on Optimization*, *9*(3), 668–689.
Künzi, H. P. A., Pajoohesh, H., & Schellekens, M. P. (2006). Partial quasi-metrics. *Theoretical Computer Science*, *365*, 237–246.
Lange, K., Hunter, D. R., & Yang, I. (2000). Optimization transfer using surrogate objective functions. *Journal of Computational and Graphical Statistics*, *9*, 1–20.
Lee, D. D. & Seung, H. S. (2001). Algorithms for non-negative matrix factorization. In *Advances in NIPS*.
Lhilali Alaoui, A. (1996). Caractérisation des fonctions D.C. *Annales des sciences Mathématiques du Québec*, *20*(1), 1–13.
Mairal, J. (2015). Incremental majorization–minimization optimization with application to large-scale machine learning. *SIAM Journal on Optimization*, *25*(2), 829–855.
Mairal, J., Bach, F., Ponce, J., & Sapiro, G. (2010). Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, *11*, 19–60.
Martinet, B. (1970). Régularisation d'inéquations variationelles par approximations successives. *Revue Française d'informatique Recherche Opérationnelle*, *4*, 154–159.
Mordukhovich, B. S., & Shao, Y. (1996). Nonsmooth sequential analysis in asplund spaces. *Transactions of the American Mathematical Society*, *348*, 1235–1280.
Moreno, F. G., Oliveira, P. R., & Soubeyran, A. (2012). A proximal point algorithm with quasi distance. Application to habit's formation. *Optimization*, *61*, 1383–1403.
Moudafi, A., & Maingé, P.-E. (2006). On the convergence of an approximate proximal method for d.c. functions. *Journal of Computational Mathematics*, *24*, 475–480.
Muu, L. D., & Quoc, T. D. (2010). One step from DC optimization to DC mixed variational inequalities. *Optimization*, *59*, 63–76.
Pan, S., & Chen, J. S. (2007). Entropy-like proximal algorithms based on a second-order homogeneous distance function for quasi-convex programming. *Journal of Global Optimization*, *39*, 555–575.
Papa Quiroz, E. A., & Oliveira, P. R. (2012). An extension of proximal methods for quasiconvex minimization on the nonnegative orthant. *European Journal of Operational Research*, *216*, 26–32.
Pham, D. T., & An, L. T. H. (1997). Convex analysis approach to DC programming: Theory, algorithms and applications. *ACTA Mathematica Vietnamica*, *22*, 289–355.
Pham, D. T., An, L. T. H., & Akoa, F. (2005). The DC (difference of convex functions) programming and DCA revisited with DC models of real world nonconvex optimization problems. *Annals of Operations Research*, *133*, 23–46.
Pham, D. T., & Souad, E. B. (1986). Algorithms for solving a class of nonconvex optimization problems: Methods of subgradient. *Fermat Days 85: Mathematics for Optimization*, *129*, 249–271.
Polyak, B. T. (1978). Subgradient methods: A survey of Soviet research. In *Nonsmooth optimization: Proceedings of the IIASA workshop March* (pp. 5–30).
Razaviyayn, M., Sanjabi, M., & Luo, Z.-Q. (2016). A stochastic successive minimization method for nonsmooth nonconvex optimization with applications to transceiver design in wireless communication networks. *Mathematical Programming*, *157*(2), 515–545.
Rockafellar, R. T. (1976). Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, *14*, 877–898.
Rockafellar, R. T., & Wets, R. J.-B. (1998). *Variational analysis*. Berlin: Springer.
Romaguera, S., & Sanchis, M. (2003). Applications of utility functions defined on quasi-metric spaces. *Journal of Mathematical Analysis and Applications*, *283*, 219–235.
Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics*, *69*(1), 99–118.
Soubeyran, A. (2009). *Variational rationality, a theory of individual stability and change: Worthwhile and ambidextry behaviors*. Marseille: GREQAM, Aix Marseillle University. (**Preprint**).
Soubeyran, A. (2010). *Variational rationality and the "unsatisfied man": Routines and the course pursuit between aspirations, capabilities and beliefs*. Marseille: GREQAM, Aix Marseillle University. (**Preprint**).
Soubeyran, A. (2016). *Variational rationality. A theory of worthwhile stay and change approach-avoidance transitions ending in traps*. Marseille: GREQAM, Aix Marseillle University. (**Preprint**).
Soubeyran, A. (2018a). *Variational rationality 1.a. Proximal dynamics and stationary traps: When is it worthwhile to move?*. Marseille: GREQAM-AMSE, Aix-Marseille University. (**Preprint**).

Soubeyran, A. (2018b). *Variational rationality 1.b. The formation of preferences and intentions*. Marseille: GREQAM-AMSE, Aix-Marseille University. (**Preprint**).

Souza, J. C. O., & Oliveira, P. R. (2015). A proximal point algorithm for DC functions on Hadamard manifolds. *Journal of Global Optimization*, *63*(4), 797–810.

Sun, W., Sampaio, R. J. B., & Candido, M. A. B. (2003). Proximal point algorithm for minimization of DC Functions. *Journal of Computational Mathematics*, *21*, 451–462.

Toland, J. F. (1979). On subdifferential calculus and duality in nonconvex optimization. *Bulletin Société Mathématique de France, Mémoire*, *60*, 177–183.

Vial, J.-P. (1983). Strong and weak convexity of sets and functions. *Mathematics of Operations Research*, *8*, 231–259.

Yassine, A., Alaa, N., & Lhilali Alaoui, A. (2001). Convergence of Toland's critical points for sequences of D.C. functions and application to the resolution of semilinear elliptic problems. *Control and Cybernetics*, *30*(4), 405–417.

Yuille, A., & Rangarajan, A. (2003). The concave–convex procedure. *Neural Computation*, *15*(4), 915–936.

Wooldridge, M. (2000). *Reasoning about rational agents*. Cambridge: The MIT Press.

Zaslavski, A. (2010). Convergence of a proximal point method in the presence of computational errors in Hilbert spaces. *SIAM Journal on Optimization*, *20*(5), 2413–2421.