Pacific Journal of Optimization, vol. 4, pp. 319-333, 2008

A proximal-like algorithm for a class of nonconvex programming

Jein-Shan Chen [1]
Department of Mathematics
National Taiwan Normal University
Taipei, Taiwan 11677

Shaohua Pan [2]
School of Mathematical Sciences
South China University of Technology
Guangzhou 510640, China

January 5, 2006
(first revised, July 14, 2006)
(second revised, April 12, 2007)
(final version, February 16, 2008)

Abstract. In this paper, we study a proximal-like algorithm for minimizing a closed proper function $f(x)$ subject to $x \geq 0$, based on the iterative scheme: $x^k \in \operatorname{argmin}\{f(x)+\mu_k d(x, x^{k-1})\}$, where $d(\cdot,\cdot)$ is an entropy-like distance function. The algorithm is well-defined under the assumption that the problem has a nonempty and bounded solution set. If, in addition, f is a differentiable quasi-convex function (or f is a differentiable function which is homogeneous with respect to a solution), we show that the sequence generated by the algorithm is convergent (or bounded), and furthermore, it converges to a solution of the problem (or every accumulation point is a solution of the problem) when the parameter μ_k approaches to zero. Preliminary numerical results are also reported, which further verify the theoretical results obtained.

Key words. Proximal algorithm, entropy-like distance, quasi-convex, homogeneous.

[1]Member of Mathematics Division, National Center for Theoretical Sciences, Taipei Office , E-mail: jschen@math.ntnu.edu.tw. The author's work is partially supported by National Taiwan Normal University.

[2]E-mail: shhpan@scut.edu.cn.

1 Introduction

The proximal point algorithm for minimizing a convex function $f(x)$ on $\mathbb{R}^n$ generates a sequence $\{x^k\}_{k\in N}$ by the iterative scheme as below:

$$x^k = \operatorname*{argmin}_{x\in\mathbb{R}^n} \left\{ f(x) + \mu_k \|x - x^{k-1}\|^2 \right\}, \tag{1}$$

where μ_k is a sequence of positive numbers and $\|\cdot\|$ denotes the Euclidean norm in $\mathbb{R}^n$. This method, which was originally introduced by Martinet [13], is based on the Moreau proximal approximation [14] of f, defined by

$$f_\lambda(x) = \inf_u \left\{ f(u) + \frac{1}{2\lambda}\|x - u\|^2 \right\}, \quad \lambda > 0. \tag{2}$$

This proximal algorithm was then further developed and studied by Rockafellar [18, 19]. In 1992, Teboulle [16] introduced the so-called *entropic proximal map* based on imitating the proximal map of Moreau, which replaces the quadratic distance in (1)–(2) with the following entropy-like distance, also called φ-divergence:

$$d_\varphi(x, y) = \sum_{i=1}^{n} y_i \varphi(x_i/y_i), \tag{3}$$

where $\varphi\colon \mathbb{R}_+ \to \mathbb{R}$ is a closed proper strictly convex function satisfying certain conditions (see [2, 9, 10, 16, 17]). An important choice of φ is $\varphi(t) = t\ln t - t + 1$, for which the corresponding d_φ is the well known Kullback-Leibler entropy function [5, 6, 7, 16] from statistics and that is the "entropy" terminology stems from.

The algorithm associated with the φ-divergence for minimizing a convex function f subject to nonnegative constraints $x \geq 0$ is given as follows

$$\begin{array}{rcl} x^0 & > & 0 \\ x^k & = & \operatorname*{argmin}_{x\geq 0} \left\{ f(x) + \mu_k d_\varphi(x, x^{k-1}) \right\}, \end{array} \tag{4}$$

where μ_k is same as in (1). The algorithm in (4) is a proximal-like one and has been studied extensively for convex programming; see [9, 10, 16, 17] and references therein. In fact, the algorithm (4) with $\varphi(t) = -\ln t + t - 1$ was first proposed by Eggermont in [6]. It is worthwhile to point out that the fundamental difference between (1) and (4) is that the term d_φ is used to force the iterates $\{x^k\}_{k\in N}$ to stay in the interior of the nonnegative orthant $\mathbb{R}^n_+$, namely the algorithm in (4) will automatically generate a positive sequence $\{x^k\}_{k\in N}$. Similar extensions and convergence results for the proximal-like methods using a Bregman distance have also been studied (see, for example, [3, 11, 12]). However, the analysis of the proximal-like method based on a Bregman distance does not carry over to the algorithm defined as (4) except for the case $\varphi(t) = t\ln t - t + 1$ where the two distances coincide. As explained in [17], this is due to the fact that one nice property

[3, Lemma 3.1], which holds for Bregman distances, does not hold in general for d_φ. In addition, we also observe that the algorithm in (4) was adopted by [4] to solve the problem of minimizing a closed function f over $\mathbb{R}^n$ without assuming convexity of f.

In this paper, we wish to employ the φ-divergence algorithm defined as in (4) with

$$\varphi(t) := -\ln t + t - 1 \quad (t > 0) \tag{5}$$

to solve the nonconvex optimization problem of the following form

$$\begin{array}{ll} \min & f(x) \\ \text{s.t.} & x \geq 0, \end{array} \tag{6}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a closed proper function with $\mathrm{dom} f \supseteq \mathbb{R}^n_{++}$. This particular choice of φ, used for convex minimization, was also discussed in [9, 17]. Since we do not require the convexity of f, the algorithm to be studied in this paper is as follows:

$$\begin{array}{rcl} x^0 & > & 0 \\ x^k & \in & \displaystyle\operatorname*{argmin}_{x \geq 0} \left\{ f(x) + \mu_k d(x, x^{k-1}) \right\}, \end{array} \tag{7}$$

where μ_k is a sequence of positive numbers and $d(x, y)$ is specified as follows:

$$d(x, y) := \sum_{i=1}^{n} y_i \varphi(x_i/y_i) = \sum_{i=1}^{n} \Big[y_i \ln(y_i/x_i) + (x_i - y_i) \Big]. \tag{8}$$

The main purpose of this paper is to establish convergence results (see Propositions 3.3 and 3.4) of the algorithm (7)–(8) under some mild assumptions for the problem (6).

Throughout this paper, $\mathbb{R}^n$ denotes the space of n-dimensional real column vectors, $\mathbb{R}^n_+$ represents the nonnegative orthant in $\mathbb{R}^n$ with its interior being $\mathbb{R}^n_{++}$, $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$ denote the Euclidean inner product and the Euclidean norm, respectively. For a given function f defined on $\mathbb{R}^n$, if f is differentiable at x, the notation $\nabla f(x)$ denotes the gradient of f at x while $(\nabla f(x))_i$ means the ith partial derivative of f with respect to x.

2 Preliminaries

In this section, we recall some preliminary results that will be used in the next section. We start with the definition of Fejér convergence to a nonempty set with respect to $d(\cdot, \cdot)$.

Definition 2.1 *A sequence $\{x^k\}_{k \in N} \subset \mathbb{R}^n_{++}$ is Fejér convergent to a nonempty set $U \subseteq \mathbb{R}^n_+$ with respect to the divergence $d(\cdot, \cdot)$ if $d(x^k, u) \leq d(x^{k-1}, u)$ for each k and any $u \in U$.*

Given an extended real-valued function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, denote its domain by

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

The function f is said to be proper if $\mathrm{dom} f \neq \emptyset$ and $f(x) > -\infty$ for any $x \in \mathrm{dom} f$, and f is called closed, which is equivalent to f being lower semicontinuous, if its epigraph is a closed set. Next we present some properties of quasi-convex and homogeneous functions, and recall the definition of stationary point for a constrained optimization problem.

Definition 2.2 *[1] Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a proper function. If*

$$f(\alpha x + (1-\alpha)y) \leq \max\{f(x), f(y)\}$$

for any $x, y \in \mathrm{dom} f$ and $\alpha \in (0,1)$, then f is called quasi-convex.

It is easy to verify that any convex function is quasi-convex as well as strictly quasi-convex, but the converse is not true. For quasi-convex functions, we have the following results.

Lemma 2.3 *[1] Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a proper function. Then,*

(a) *f is quasi-convex if and only if the level sets $L_f(\gamma) := \{x \in \mathrm{dom} f \mid f(x) \leq \gamma\}$ are convex for all $\gamma \in \mathbb{R}$.*

(b) *If f is differentiable on $\mathrm{dom} f$, then f is quasi-convex if and only if $\langle \nabla f(y), x - y\rangle \leq 0$ whenever $f(x) \leq f(y)$ for any $x, y \in \mathrm{dom} f$.*

Definition 2.4 *[15] A proper function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is called homogeneous with respect to $\bar{x} \in \mathrm{dom} f$ with exponential $\kappa > 0$ if for any $x \in \mathrm{dom} f$ and $\lambda \geq 0$,*

$$f(\bar{x} + \lambda(x - \bar{x})) - f(\bar{x}) = \lambda^{\kappa}(f(x) - f(\bar{x})).$$

Lemma 2.5 *[15] Assume that the proper function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is differentiable on $\mathrm{dom} f$. If f is homogeneous with respect to $\bar{x} \in \mathrm{dom} f$ with exponential $\kappa > 0$, then*

$$f(x) - f(\bar{x}) = \kappa^{-1}\langle \nabla f(x), x - \bar{x}\rangle \quad \forall x \in \mathrm{dom} f.$$

Definition 2.6 *For a constrained optimization problem $\min_{x \in C} f(x)$, where $C \subseteq \mathbb{R}^n$ is a nonempty convex set, x^* is called a stationary point if $\nabla f(x^*)^T(x - x^*) \geq 0$ for all $x \in C$.*

In what follows, we focus on the properties of φ given as (5) and the induced function $d(\cdot,\cdot)$, which will be used in the subsequent analysis. First, we summarize some special properties of φ. Since their verifications are direct by computations, we omit the details.

Property 2.7 *Let $\varphi : \mathbb{R}_{++} \to \mathbb{R}$ be defined as in (5). Then, the following results hold.*

(a) $\varphi(t) \geq 0$ *and* $\varphi(t) = 0$ *if and only if* $t = 1$.

(b) $\varphi(t)$ *is decreasing in* $(0,1)$ *with* $\lim_{t\to 0^+} \varphi(t) = +\infty$, *and increasing in* $(1,\infty)$ *with* $\lim_{t\to+\infty} \varphi(t) = +\infty$.

(c) $\varphi(1) = 0$, $\varphi'(1) = 0$, *and* $\varphi''(1) > 0$.

(d) $\varphi'(t)$ *is nondecreasing on* $(0,+\infty)$ *and* $\lim_{t\to+\infty} \varphi'(t) = 1$, $\lim_{t\to 0^+} \varphi'(t) = -\infty$.

(e) $\varphi'(t) \leq \ln t$ *for all* $t > 0$ *and* $\varphi'(t) > 0$ *when* $t \in [1,\infty)$.

From Property 2.7 (a) and the definition of $d(x,y)$, it is not hard to see that

$$d(x,y) \geq 0 \ \text{ and } \ d(x,y) = 0 \iff x = y, \ \forall x, y \in \mathbb{R}^n_{++}.$$

This means that $d(\cdot,\cdot)$ can be viewed as a distance-like function, though $d(\cdot,\cdot)$ itself can not be a distance since the triangle inequality does not hold in general. In fact, d is a *divergence measure* in $\mathbb{R}^n_{++} \times \mathbb{R}^n_{++}$ (see [10]) that enjoys some favorable properties, for example, the ones given by Lemmas 2.8 and 2.9. In addition, we notice that $d : \mathbb{R}^n_{++} \times \mathbb{R}^n_{++} \to \mathbb{R}$ is a continuous function and can be continuously extended to $\mathbb{R}^n_{++} \times \mathbb{R}^n_{+}$ by adopting the convention $0 \ln 0 = 0$, i.e., $d(\cdot,\cdot)$ admits points with 0 component in its second argument. The following two lemmas characterize some crucial properties of the divergence measure.

Lemma 2.8 *Let φ and d be defined as in (5) and (8), respectively. Then,*

(a) $d(x,z) - d(y,z) \geq \sum_{i=1}^{n} (z_i - y_i)\varphi'(y_i/x_i)$ *for any* $x, y \in \mathbb{R}^n_{++}$ *and* $z \in \mathbb{R}^n_{+}$;

(b) *for any fixed* $y \in \mathbb{R}^n_{+}$, $L_x(y,\gamma) := \left\{x \in \mathbb{R}^n_{++} \mid d(x,y) \leq \gamma\right\}$ *are bounded for all* $\gamma \geq 0$;

(c) *for any fixed* $x \in \mathbb{R}^n_{++}$, $L_y(x,\gamma) := \left\{y \in \mathbb{R}^n_{+} \mid d(x,y) \leq \gamma\right\}$ *are bounded for all* $\gamma \geq 0$.

Proof. (a) By the definition of the function d, we can compute that

$$\begin{aligned} d(x,z) - d(y,z) &= \sum_{i=1}^{n} \Big[z_i \ln(z_i/x_i) + x_i - z_i\Big] - \sum_{i=1}^{n} \Big[z_i \ln(z_i/y_i) + y_i - z_i\Big] \\ &= \sum_{i=1}^{n} \Big[z_i \ln(y_i/x_i) + x_i - y_i\Big]. \end{aligned} \tag{9}$$

Since $\varphi'(y_i/x_i) = 1 - x_i/y_i$, we have $y_i\varphi'(y_i/x_i) = y_i - x_i$, i.e.,

$$x_i - y_i = -y_i\varphi'(y_i/x_i). \tag{10}$$

In addition, using Property 2.7 (e) and noting that $z_i \geq 0$ for all i, we readily have

$$z_i \ln(x_i/y_i) \geq z_i\varphi'(y_i/x_i). \tag{11}$$

From equations (9)–(11), we immediately obtain that

$$d(x,z) - d(y,z) \geq \sum_{i=1}^{n} \Big[z_i\varphi'(y_i/x_i) - y_i\varphi'(y_i/x_i)\Big] = \sum_{i=1}^{n}(z_i - y_i)\varphi'(y_i/x_i).$$

(b) The boundedness of the level sets $L_x(y,\gamma)$ for all $\gamma \geq 0$ is direct by Property 2.7 (b).

(c) Let $\psi(t) := t\ln t - t + 1$ $(t \geq 0)$. By the definition of $d(\cdot,\cdot)$, we can verify that

$$d(x,y) = \sum_{i=1}^{n} x_i\psi(y_i/x_i) \quad \forall x \in \mathbb{R}^n_{++} \ \text{ and } \ y \in \mathbb{R}^n_{+}.$$

Thus, for any fixed $x \in \mathbb{R}^n_{++}$, to show that $L_y(x,\gamma)$ are bounded for all $\gamma \geq 0$, it suffices to prove that $\psi(t)$ has bounded level sets, which is clear since $\lim\limits_{t\to+\infty} \psi(t) = +\infty$. □

Lemma 2.9 *Given any two sequences $\{y^k\}_{k\in N} \subset \mathbb{R}^n_{++}$ and $\{x^k\}_{k\in N} \subseteq \mathbb{R}^n_{+}$,*

(a) *if $\{y^k\}_{k\in N}$ converges to $\bar{y} \in \mathbb{R}^n_{+}$, then $\lim_{k\to+\infty} d(y^k,\bar{y}) = 0$;*

(b) *if $\{y^k\}_{k\in N}$ is bounded and $\{x^k\}_{k\in N}$ is such that $\lim_{k\to+\infty} d(y^k,x^k) = 0$, then we have $\lim_{k\to+\infty} \|y^k - x^k\| = 0$.*

Proof. (a) From the definition of $d(\cdot,\cdot)$, it follows that

$$d(y^k,\bar{y}) = \sum_{i=1}^{n} \Big[\bar{y}_i\ln\bar{y}_i - \bar{y}_i\ln y_i^k + (y_i^k - \bar{y}_i)\Big].$$

For any $i \in \{1,2,\ldots,n\}$, if $\bar{y}_i = 0$, clearly $\bar{y}_i\ln\bar{y}_i - \bar{y}_i\ln y_i^k + (y_i^k - \bar{y}_i) \to 0$ as $k \to +\infty$; if $\bar{y}_i > 0$, then $\ln(\bar{y}_i/y_i^k) \to 0$ and $(y_i^k - \bar{y}_i) \to 0$ since $\{y_i^k\} \to \bar{y}_i$, which means that $\bar{y}_i\ln(\bar{y}_i/y_i^k) + (y_i^k - \bar{y}_i) \to 0$ as $k \to +\infty$. The two sides yield that $\lim_{k\to+\infty} d(y^k,\bar{y}) = 0$.

(b) First, by Lemma 2.8 (b) and the fact that $\lim_{k\to+\infty} d(y^k,x^k) = 0$, we may verify that $\{x^k\}_{k\in N}$ is bounded. Now suppose $\lim\limits_{k\to+\infty} \|y^k - x^k\| \neq 0$. Then, there exists a subsequence $\{y^{\sigma(k)}\}_{k\in N}$ such that $\|y^{\sigma(k)} - x^{\sigma(k)}\| \geq 3\varepsilon$ for some $\varepsilon > 0$ and for all sufficiently large k. Since $\{y^{\sigma(k)}\}_{k\in N}$ is bounded, we can extract a convergent subsequence. Without loss of

generality, we still use $\{y^{\sigma(k)}\}_{k\in N} \to y^*$ to represent the convergent subsequence. From the triangle inequality, it then follows that

$$\|y^{\sigma(k)} - y^*\| + \|y^* - x^{\sigma(k)}\| \geq \|y^{\sigma(k)} - x^{\sigma(k)}\| \geq 3\varepsilon.$$

Since $\{y^{\sigma(k)}\}_{k\in N} \to y^*$, there exists a positive integer K such that $\|y^{\sigma(k)} - y^*\| \leq \varepsilon$ for $k \geq K$. Thus, we have $\|y^* - x^{\sigma(k)}\| \geq 3\varepsilon - \|y^{\sigma(k)} - y^*\| \geq 2\varepsilon$ for $k \geq K$. On the other hand, the boundedness of $\{x^{\sigma(k)}\}_{k\in N}$ implies that there is a convergent subsequence $\{x^{\sigma(\gamma(k))}\} \to x^*$ for $k \geq K$ and $k \to +\infty$. Then by the same arguments as above, we obtain $\|y^* - x^*\| \geq \varepsilon$. However, $\lim\limits_{k\to+\infty} d(y^k, x^k) = 0$ yields $\lim\limits_{k\to+\infty} d(y^{\sigma(\gamma(k))}, x^{\sigma(\gamma(k))}) = 0$. Therefore, from the continuity of d and boundedness of the sequences, we have $d(y^*, x^*) = 0$ which implies $y^* = x^*$. Thus, we obtain a contradiction. The proof is complete. □

3 Main results

In this section, we establish the convergence results of the proximal-like algorithm (7)–(8). First, we show that the algorithm is well-defined under the following assumption:

(A1) The solution set of problem (6), denoted by $\mathcal{X}^*$, is nonempty and bounded.

Lemma 3.1 *Let d be defined as in (8). Then, under assumption (A1),*

(a) *the sequence $\{x^k\}_{k\in N}$ generated by (7)–(8) is well-defined;*

(b) *$\{f(x^k)\}_{k\in N}$ is a decreasing and convergent sequence.*

Proof. (a) The proof proceeds by induction. Clearly, when $k = 0$, the conclusion holds since $x^0 > 0$. Suppose that x^{k-1} is well-defined. Let f^* be the optimal value of (6). Then, from the iterative scheme (7) and the nonnegativity of d, it follows that

$$f(x) + \mu_k d(x, x^{k-1}) \geq f^* + \mu_k d(x, x^{k-1}) \quad \text{for all } x \in \mathbb{R}^n_{++}. \tag{12}$$

Let $f_k(x) := f(x) + \mu_k d(x, x^{k-1})$ and denote its level sets by

$$L_{f_k}(\gamma) := \{x \in \mathbb{R}^n_{++} : f_k(x) \leq \gamma\} \quad \text{for any } \gamma \in \mathbb{R}.$$

Using the inequality in (12), we have $L_{f_k}(\gamma) \subseteq L_x(x^{k-1}, \mu_k^{-1}(\gamma - f^*))$. This, together with Lemma 2.8 (b), implies that $L_{f_k}(\gamma)$ is bounded for any $\gamma \geq f^*$. Notice that $L_{f_k}(\gamma) = \mathcal{X}^*$ for any $\gamma \leq f^*$ since $\mu_k d(x, x^{k-1}) \geq 0$, and consequently $L_{f_k}(\gamma)$ is also bounded for this case by assumption (A1). This shows that the level sets of $f_k(x)$ are bounded. Also, $f_k(x)$ is lower semicontinuity on $\mathbb{R}^n$. Therefore, the level sets of $f_k(x)$ are compact. Using the

lower semicontinuity of $f_k(x)$ again, we have that $f_k(x)$ has a global minimum which may not be unique due to the nonconvexity of f. In such case, x^k can be arbitrarily chosen among the set of minimizers of $f_k(x)$. The sequence $\{x^k\}_{k\in N}$ is thus well-defined.

(b) From the iterative scheme in (7), it readily follows that

$$f(x^k) + \mu_k d(x^k, x^{k-1}) \leq f(x) + \mu_k d(x, x^{k-1}), \quad \forall x \in \mathbb{R}^n_{++}. \tag{13}$$

Setting $x = x^{k-1}$ in the last inequality, we obtain that

$$f(x^k) + \mu_k d(x^k, x^{k-1}) \leq f(x^{k-1}) + \mu_k d(x^{k-1}, x^{k-1}) = f(x^{k-1}),$$

which, by the nonnegativity of d and μ_k, implies that

$$0 \leq \mu_k d(x^k, x^{k-1}) \leq f(x^{k-1}) - f(x^k).$$

This show that $\{f(x^k)\}_{k\in N}$ is a decreasing sequence, and furthermore, it is convergent by assumption (A1). The proof is thus completed. □

By Lemma 3.1 (b), let $\beta := \lim_{k\to+\infty} f(x^k)$ and define the following set

$$U := \{x \in \mathbb{R}^n_+ \mid f(x) \leq \beta\}. \tag{14}$$

Clearly, $\mathcal{X}^* \subseteq U$, and consequently U is nonempty by assumption (A1). In what follows, we show that the sequence $\{x^k\}_{k\in N}$ generated by (7)–(8) is Fejér convergent to U with respect to d under the following additional assumption for f:

(A2) f is a quasi-convex function which is differentiable on $\mathrm{dom} f$.

Lemma 3.2 *Let $\{\mu_k\}$ be an arbitrary sequence of positive numbers and $\{x^k\}_{k\in N}$ be the sequence generated by (7)–(8). Then, under assumptions (A1) and (A2),*

(a) $d(x^k, x) \leq d(x^{k-1}, x)$ for any $x \in \mathbb{R}^n_+$ such that $f(x) \leq f(x^k)$;

(b) $\{x^k\}_{k\in N}$ *is Fejér convergent to the set U with respect to d;*

(c) *for any $x \in U$, the sequence $\{d(x^k, x)\}_{k\in N}$ is convergent.*

Proof. (a) For any $x \in \mathbb{R}^n_+$ satisfying $f(x) \leq f(x^k)$, from Lemma 2.3 (b) it follows that

$$\nabla f(x^k)^T (x - x^k) \leq 0.$$

In addition, since x^k is a minimizer of the function $f(x) + \mu_k d(x, x^{k-1})$, we have

$$\nabla f(x^k) + \mu_k \nabla_x d(x^k, x^{k-1}) = 0. \tag{15}$$

Combining the last two equations, we obtain that

$$\mu_k\langle\nabla_x d(x^k, x^{k-1}), x - x^k\rangle \geq 0 \quad \forall x \in \mathbb{R}^n_+ \text{ such that } f(x) \leq f(x^k).$$

Noting that

$$\nabla_x d(x^k, x^{k-1}) = (\varphi'(x^k_1/x^{k-1}_1), \ldots, \varphi'(x^k_n/x^{k-1}_n))^T,$$

and using Lemma 2.8 (a) with $x = x^{k-1}, y = x^k$ and $z = x$, it follows that

$$d(x^{k-1}, x) - d(x^k, x) \geq \mu_k\langle\nabla_x d(x^k, x^{k-1}), x - x^k\rangle \geq 0$$

for any $x \in \mathbb{R}^n_+$ satisfying $f(x) \leq f(x^k)$. The desired result follows.

(b) By the definition of U, clearly, $x \in U$ means $f(x) \leq f(x^k)$ for all k since $\{f(x^k)\}_{k\in N}$ is decreasing. The proof is direct by part (a) and Definition 2.1.

(c) The proof follows directly from part (b) and the nonnegativity of d. □

Now we are in a position to establish the convergence results of the algorithm.

Proposition 3.3 *Let $\{\mu_k\}_{k\in N}$ be an arbitrary sequence of positive numbers and $\{x^k\}_{k\in N}$ be generated by (7)–(8). If assumptions (A1) and (A2) hold, then $\{x^k\}_{k\in N}$ converges, and*

(a) *if there exist $\hat{\mu}$ and $\bar{\mu}$ such that $0 < \hat{\mu} < \mu_k \leq \bar{\mu}$ for each k, then*

$$\lim_{k\to+\infty}(\nabla f(x^k))_i \geq 0, \quad \lim_{k\to+\infty}(\nabla f(x^k))_i(x_i - x^k_i) \geq 0 \text{ for any } x \in \mathbb{R}^n_+, \quad i = 1, 2, \ldots, n;$$

(b) *if $\lim_{k\to+\infty}\mu_k = 0$, then $\{x^k\}_{k\in N}$ converges to a solution of (6).*

Proof. We first prove that the sequence $\{x^k\}_{k\in N}$ is convergent. By Lemma 3.2 (b), $\{x^k\}_{k\in N}$ is Fejér convergent to the set U with respect to d, which in turn implies that

$$\{x^k\}_{k\in N} \subseteq \left\{y \in \mathbb{R}^n_{++} \mid d(y, x) \leq d(x^0, x)\right\} \quad \forall x \in U.$$

From Lemma 2.8 (b), the set on the right hand side of the last equation is bounded, and consequently, $\{x^k\}_{k\in N}$ is bounded. Let $\bar{x}$ be an accumulation point of $\{x^k\}_{k\in N}$ and $\{x^{k_j}\}$ be a subsequence converging to $\bar{x}$. From the continuity of f, it then follows that

$$\lim_{j\to+\infty} f(x^{k_j}) = f(\bar{x}),$$

which, by the definition of U, implies that $\bar{x} \in U$. Using Lemma 3.2 (c), we have that the sequence $\{d(x^k, \bar{x})\}_{k\in N}$ is convergent. Notice that $\lim_{j\to+\infty} d(x^{k_j}, \bar{x}) = 0$ by Lemma 2.9 (a), and therefore, we conclude that $\lim_{k\to+\infty} d(x^k, \bar{x}) = 0$. Using Lemma 2.9 (b) with $y^k = x^k$ and $x^k = \bar{x}$, we prove that $\{x^k\}_{k\in N}$ converges to $\bar{x}$.

(a) By the iterative formula in (7), x^k is the minimizer of $f(x)+\mu_k d(x,x^{k-1})$, and hence

$$\nabla f(x^k) = -\mu_k \nabla_x d(x^k, x^{k-1}),$$

which means that

$$(\nabla f(x^k))_i = -\mu_k \varphi'(x_i^k/x_i^{k-1}), \quad i=1,2,\ldots,n. \tag{16}$$

Let $\bar{x}$ be the limit of the sequence $\{x^k\}_{k\in N}$. Define the index sets

$$I(\bar{x}) := \Big\{ i \in \{1,2,\ldots,n\} \mid \bar{x}_i > 0 \Big\} \ \text{ and } \ J(\bar{x}) := \Big\{ i \in \{1,2,\ldots,n\} \mid \bar{x}_i = 0 \Big\}.$$

Clearly, the two disjoint sets form a division of $\{1,2,\ldots,n\}$. We proceed the arguments by the cases where $i \in I(\bar{x})$ or $i \in J(\bar{x})$ for any $i \in \{1,2,\ldots,n\}$.

Case (1): If $i \in I(\bar{x})$, then $\lim_{k\to+\infty} x_i^k/x_i^{k-1} = 1$ by the convergence of $\{x^k\}_{k\in N}$. Notice that φ is continuous on $(0,+\infty)$ and $\varphi'(1)=0$, and hence,

$$\lim_{k\to+\infty} \varphi'(x_i^k/x_i^{k-1}) = 0.$$

This, together with (16) and the boundedness of $\{\mu_k\}_{k\in N}$, immediately yields

$$\lim_{k\to+\infty} (\nabla f(x^k))_i = 0.$$

Consequently, by the convergence of $\{x^k\}_{k\in N}$, we have $\lim_{k\to+\infty} (\nabla f(x^k))_i (x_i - x_i^k) = 0$.

Case (2): If $i \in J(\bar{x})$, we will show that $\lim_{k\to+\infty} (\nabla f(x^k))_i \geq 0$. Suppose that it does not hold. Then, $(\nabla f(x^k))_i < 0$ for large enough k. From (16) and the assumption for $\{\mu_k\}_{k\in N}$, it then follows that $\varphi'(x_i^k/x_i^{k-1}) > 0$ for sufficiently large $k \in N$. Thus, by Property 2.7 (c)–(d), $x_i^k > x_i^{k-1}$ for large enough $k \in N$, which contradicts the fact that

$$\{x^k\}_{k\in N} \to \bar{x} \quad \text{and} \quad \bar{x}_i = 0 \quad \forall i \in J(\bar{x}).$$

Consequently, $\lim_{k\to+\infty} (\nabla f(x^k))_i \geq 0$. Noting that $\lim_{k\to+\infty} (x_i - x_i^k) \geq 0$ since $x_i \geq 0$ and $\lim_{k\to+\infty} x_i^k = \bar{x}_i = 0$, we readily have $\lim_{k\to+\infty} (\nabla f(x^k))_i (x_i - x_i^k) \geq 0$.

(b) Since x^k is the minimizer of $f(x)+\mu_k d(x,x^{k-1})$, we have

$$f(x^k) + \mu_k d(x^k, x^{k-1}) \leq f(x) + \mu_k d(x, x^{k-1}) \quad \forall x \in \mathbb{R}^n_{++},$$

which, by the nonnegativity of d, implies that

$$f(x^k) \leq f(x) + \mu_k d(x, x^{k-1}) \quad \forall x \in \mathbb{R}^n_{++}. \tag{17}$$

Taking the limit $k \to +\infty$ on the inequality and using the continuity of f yields that

$$f(\bar{x}) \leq f(x), \quad \forall x \in \mathbb{R}^n_{++}$$

since $\lim_{k\to+\infty} x^k = \bar{x}$, $\lim_{k\to+\infty} \mu_k = 0$ and the sequence $\{d(x, x^{k-1})\}_{k\in N}$ is bounded by Lemma 2.8 (c). This, together with the continuity of f, means that $f(\bar{x}) \le f(x)$ for any $x \in \mathbb{R}^n_+$, and consequently, $\bar{x} \in \mathcal{X}^*$. We thus complete the proof. $\square$

From Proposition 3.3 (a), we see that the sequence $\{x^k\}_{k\in N}$ converges to a stationary point of (6) without requiring $\mu_k \to 0$ if f is continuously differentiable quasi-convex. When f is not quasi-convex, the following proposition states that the sequence $\{x^k\}_{k\in N}$ generated by (7)–(8) is bounded and every limit point is a solution of problem (6) under (A1) and the following additional assumption:

(A3) f is differentiable on dom$f$ and homogeneous with respect to a solution of the problem (6) with exponential $\kappa > 0$.

Proposition 3.4 *Let $\{\mu_k\}_{k\in N}$ be any sequence of positive numbers and $\{x^k\}_{k\in N}$ be generated by (7)–(8). If assumptions (A1) and (A3) hold, then $\{x^k\}_{k\in N}$ is bounded, and*

(a) *if there exist $\hat{\mu}$ and $\bar{\mu}$ such that $0 < \hat{\mu} < \mu_k \le \bar{\mu}$ for each k, then*

$$\lim_{k\to+\infty} (\nabla f(x^k))_i \ge 0, \quad \lim_{k\to+\infty} (\nabla f(x^k))_i (x_i - x_i^k) \ge 0 \text{ for any } x \in \mathbb{R}^n_+, \ \ i = 1, 2, \ldots, n;$$

(b) *if $\lim_{k\to+\infty} \mu_k = 0$, every limit point of $\{x^k\}_{k\in N}$ is an optimal solution of (6).*

Proof. Let $x^* \in \mathcal{X}^*$ be such that f is homogeneous with respect to x^* with exponential $\kappa > 0$. Then, using Lemma 2.5 it follows that

$$0 \le f(x^k) - f(x^*) = \kappa^{-1} \langle \nabla f(x^k), x^k - x^* \rangle \quad \forall k.$$

This, together with (15), means that

$$\langle \mu_k \nabla_x d(x^k, x^{k-1}), x^* - x^k \rangle \ge 0 \quad \forall k.$$

Using Lemma 2.8 (a) with $x = x^{k-1}, y = x^k$ and $z = x^*$ then yields that

$$d(x^{k-1}, x^*) - d(x^k, x^*) \ge \mu_k \langle \nabla_x d(x^k, x^{k-1}), x^* - x^k \rangle \ge 0 \quad \forall k.$$

This shows that the sequence $\{d(x^k, x^*)\}_{k\in N}$ is monotonically decreasing, and hence,

$$\{x^k\}_{k\in N} \subseteq \left\{ y \in \mathbb{R}^n_{++} \mid d(y, x^*) \le d(x^0, x^*) \right\}.$$

Using Lemma 2.8 (b), we have that $\{x^k\}_{k\in N}$ is bounded. Let $\bar{x}$ be an arbitrary limit point of $\{x^k\}_{k\in N}$ and $\{x^{k_j}\}$ be a subsequence such that $\lim_{j\to+\infty} x^{k_j} = \bar{x}$. Then, it is easy to prove part (a) by following the same arguments as Proposition 3.3 (a). Notice that the inequality (17) still holds, and consequently, we have

$$f(x^{k_j}) \le f(x) + \mu_k d(x, x^{k_j - 1}) \quad \forall x \in \mathbb{R}^n_{++}.$$

Taking the limit $j \to +\infty$ and using the same arguments as Proposition 3.3 (b), we have

$$f(\bar{x}) \le f(x), \quad \forall x \in \mathbb{R}^n_+.$$

This shows that $\bar{x}$ is a solution of the problem (6). The proof is thus completed. $\square$

4 Numerical experiments

In this section, we verify the theoretical results obtained by applying the algorithm (7)–(8) for some differentiable quasi-convex optimization problems of the form (6). Since f is quasi-convex if and only if the level sets of f are convex, we generate the quasi-convex function $f(x)$ by compounding a quadratic convex function $g(x) := (1/2)x^T Mx$ with a monotone increasing but nonconvex function $h : \mathbb{R} \to \mathbb{R}$, i.e., $f(x) = h(g(x))$, where $M \in \mathbb{R}^{n\times n}$ is a given symmetric positive semidefinite matrix.

In our experiments, the matrix M was obtained by setting $M = NN^T$, where N was a square matrix whose nonzero elements were chosen randomly from a normal distribution with mean -1, variance one and standard deviation one. In this procedure, the number of nonzero elements of N is determined so that the nonzero density of M can be approximately estimated. The function h is given as below.

Experiment A. We took $h(t) = -\dfrac{1}{1+t}$ $(t \neq -1)$, and generated 10 matrices M of dimension $n = 100$ with approximate nonzero density 0.1% and 10%, respectively. Then, we solved the quasi-convex programming problem (6) with $f(x) = -\dfrac{1}{1+(x^T Mx)/2}$.

Experiment B. Set $h(t) = \sqrt{t} + 1$ $(t \geq 0)$. We adopted this h and the matrix M in Experiment A to yield the quasi-convex function $f(x) = \sqrt{(x^T Mx)/2} + 1$.

Experiment C. Set $h(t) = \ln(1 + t)$ $(t > -1)$. We employed h and the matrix M in Experiment A to generate the quasi-convex function $f(x) = \ln[1 + (x^T Mx)/2]$.

Experiment D. We took $h(t) = \arctan(t) + t + 2$ and used such h and the matrix M in Experiment A to generate the function $f(x) = \arctan[(x^T Mx)/2] + (x^T Mx)/2 + 2$.

It is not difficult to verify that each h in Experiments A-D is pseudo-convex, and hence the corresponding f is also pseudo-convex by [1, Excercie 3.43]. Thus, every stationary point of (6) is a globally optimal solution by [1, Section 3.5]. Moreover, we notice that all test problems have a globally optimal solution $x^* = 0$. In view of this, throughout the experiments, we employed an approximate version of the algorithm (7)–(8), described as follows, to solve the test problems generated randomly.

Approximate entropy-like proximal algorithm

Given $\epsilon > 0$ and $\tau > 0$. Select a starting point $x^0 \in \mathbb{R}^n_{++}$, and set $k := 0$.

For $k = 1, 2, \cdots$ **until** $|\nabla f(x^k)^T x^k| < \epsilon$ **do**

1. Use an unconstrained minimization method to solve approximately the problem

$$\min_{x\in\mathbb{R}^n}\left\{f(x)+\mu_k d(x,x^{k-1})\right\}, \tag{18}$$

and obtain an x^k such that $\|\nabla f(x^k)+\mu_k\nabla_x d(x^k,x^{k-1})\|\le\tau$.

2. Let $k:=k+1$, and then go back to Step 1.

End

We implemented the approximate entropy-like proximal algorithm with our code in MATLAB 6.5. All numerical experiments were done at a PC with 2.8GHz CPU and 512MB memory. We chose a BFGS algorithm with Armijo line search to solve (18). To improve numerical behavior, we replaced the standard Armijo line search by the nonmonotone line search technique described as in [8] to seek a suitable steplength, i.e., we computed the smallest nonnegative integer l such that

$$f_k(x^k+\beta^l d^k)\le\mathcal{W}_k+\sigma\beta^l\nabla f_k(x^k)^T d^k$$

where $f_k(x)=f(x)+\mu_k d(x,x^{k-1})$ and $\mathcal{W}_k$ is given by

$$\mathcal{W}_k:=\max\left\{f_k(x^j)\mid j=k-m_k,\ldots,k\right\},$$

and, for a given nonnegative integer $\hat{m}$ and s, we set

$$m_k=\begin{cases}0 & \text{if } k\le s\\ \min\{m_{k-1}+1,\hat{m}\} & \text{otherwise}\end{cases}.$$

Throughout the experiments, the following parameters were used for the line search:

$$\beta=0.5,\quad \sigma=10^{-4},\quad \hat{m}=5\ \text{ and }\ s=5.$$

The parameters ϵ and τ in the algorithm were chosen as $\epsilon=10^{-5}$ and $\tau=10^{-5}$, respectively. In addition, we updated the proximal parameter μ_k by the following formula:

$$\mu_{k+1}=0.1\mu_k\quad\text{with}\quad\mu_0=1,$$

and used $x^0=w$ as the starting point, where ω was chosen randomly from $[1,2]$.

Numerical results for Experiments A-D are summarized in Tables 1-4 of the appendix. In these tables, **Obj**. represents the value of $f(x)$ at the final iteration, **Nf** denotes the total number of function evaluations for the objective function of subproblem (18) for solving each quasi-convex programming problem, **Den** denotes the approximate nonzero density of M, and **Time** represents the CPU time in second for solving each test problem.

From Tables 1–4, we see that the approximate entropy-like proximal algorithm can find the stationary point successfully for all test problems in Experiments A–D from the given starting point $x^0 = \omega$ except those test problems with nonzero density 10% in Experiment A, for which we used the starting point $x^0 = 0.5\omega$ instead. This verifies the theoretical results obtained in last section. In addition, from these tables, it seems that the proximal-like algorithm needs more function evaluations for those problems involved in a matrix M with a higher nonzero density.

5 Conclusions

We have considered the proximal-like method defined by (7)–(8) for a class of nonconvex problems of the form (6) and established the convergence results of the algorithm under some suitable assumptions. Specifically, we have shown that, under assumptions (A1) and (A2), the sequence $\{x^k\}_{k\in N}$ generated by the algorithm is convergent, and furthermore, converges to a solution of (6) when $\mu_k \to 0$; while under assumptions (A1) and (A3), the sequence $\{x^k\}_{k\in N}$ is only bounded but every accumulation point is a solution if $\mu_k \to 0$. Preliminary numerical experiments are also reported to verify the theoretical results. In our future research work, we will consider applying the proximal-like algorithm for more extensive classes of nonconvex optimization problems.

Acknowledgement. The authors are grateful to Prof. Fukushima and Prof. Tseng for inspiring them to generate test examples. They also thank two referees for valuable suggestions.

References

[1] M. S. Bazaraa and C. M. Shetty, *Nonlinear Programming Theory and Algorithms*, New York: John Wiley & Sons, 1979.

[2] Y. Censor and S. A. Zenios, *The proximal minimization algorithm with D-functions*, Journal of Optimization Theory and Applications, vol. 73 (1992), pp. 451-464.

[3] G. Chen and M. Teboulle, *Convergence analysis of a proximal-like minimization algorithm using Bregman functions*, SIAM Journal on Optimization, vol. 3 (1993), pp. 538-543.

[4] S. Chretien and O. Hero, *Generalized proximal point algorithms*, Technical Report, The University of Michigan, 1998.

[5] J. Eckstein, *Nonlinear proximal point algorithms using Bregman functions, with applications to convex programming*, Mathematics of Operation Research, vol. 18 (1993), pp. 206-226.

[6] P. B. Eggermont, *Multilicative iterative algorithms for convex programming*, Linear Algebra Appl., vol. 130 (1990), pp. 25-42.

[7] O. Guler, *On the convergence of the proximal point algorithm for convex minimization*, SIAM Journal on Control Optimization, vol. 29 (1991), pp. 403-419.

[8] L. Grippo, F. Lampariello and S. Lucidi, *A nonmonotone line search technique for Newton's method*, SIAM Journal on Numerical Analysis, vol. 23 (1986), pp. 707-716.

[9] A. Iusem and M. Teboulle, *Convergence rate analysis of nonquadratic proximal method for convex and linear programming*, Mathematics of Operation Research, vol. 20 (1995), pp. 657-677.

[10] A. Iusem, B. Svaiter, and M. Teboulle, *Entropy-Like proximal methods in convex programming*, Mathematics of Operation Research, vol. 19 (1994), pp. 790-814.

[11] K. C. Kiwiel, *Proximal minimization methods with generalized Bregman functions*, SIAM Journal on Control and Optimization, vol. 35 (1997), pp. 1142-1168.

[12] S. Kabbadj, *Méthodes proximales entropiques*, Thése de Doctorat, Uinversité Montpellier II, 1994.

[13] B. Martinet, *Perturbation des methodes d'Optimisation*, Application, R.A.I.R.O. Numer. Anal., 1978.

[14] J. J. Moreau, *Promimité et Dualité dans un Espace Hilbertien*, Bull. Soc. Math. France, vol. 93 (1965), pp. 273-299.

[15] B. T. Polyak, *Introduction to Optimization*. Optimization Software, INC., 1987.

[16] M. Teboulle, *Entropic proximal mappings with applications to nonlinear programming*, Mathematics of Operation Research , vol. 17 (1992), pp. 670-690.

[17] M. Teboulle, *Convergence of proximal-like algorithms*, SIAM Journal on Optimization, vol. 7 (1997), pp. 1069-1083.

[18] R.T. Rockafellar, *Augmented lagrangians and applications of proximal point algorithm in convex programming*, Mathematics of Operation Research, vol. 1 (1976), pp. 97-116.

Table 1: **Numerical results for Experiment A**

	Den=0.1%			**Den**=10%		
N0.	**Obj.**	**Nf**	**Time(s)**	**Obj.**	**Nf**	**Time(s)**
1	-0.9999958e-0	628	0.31	-0.99999530e-0	52984	21.71
2	-0.9999974e-0	582	0.31	-0.99999504e-0	32168	16.65
3	-0.9999950e-0	997	0.39	-0.99999500e-0	45363	27.04
4	-0.9999955e-0	728	0.37	-0.99999521e-0	45944	18.11
5	-0.9999964e-0	571	0.30	-0.99999519e-0	49001	26.40
6	-0.9999960e-0	869	0.52	-0.99999501e-0	61891	29.70
7	-0.9999995e-0	631	0.33	-0.99999513e-0	40297	20.50
8	-0.9999958e-0	1048	0.44	-0.99999511e-0	64684	37.60
9	-0.9999953e-0	461	0.31	-0.99999503e-0	33232	15.03
10	-0.9999988e-0	808	0.37	-0.99999510e-0	31660	21.09

Note: the starting point $x^0 = 0.5w$ was used for all test problems with density 10%.

[19] R.T. Rockafellar, *Monotone operators and the proximal point algorithm*, SIAM Journal on Control and Optimization, vol. 14 (1976), pp. 877-898.

Appendix

Table 2: **Numerical results for Experiment B**

	Den=0.1%			**Den**=10%		
N0.	**Obj.**	**Nf**	**Time(s)**	**Obj.**	**Nf**	**Time(s)**
1	2.0000005e-0	619	0.36	2.0000099e-0	11412	7.67
2	2.0000099e-0	1599	0.48	2.0000095e-0	11705	7.40
3	2.0000094e-0	652	0.33	2.0000096e-0	9636	7.32
4	2.0000093e-0	563	0.28	2.0000098e-0	11041	7.65
5	2.0000072e-0	791	0.41	2.0000095e-0	10211	7.00
6	2.0000082e-0	996	0.58	2.0000095e-0	12963	7.51
7	2.0000054e-0	514	0.22	2.0000100e-0	11631	8.68
8	2.0000045e-0	622	0.31	2.0000097e-0	10702	6.09
9	2.0000052e-0	573	0.31	2.0000099e-0	10927	8.43
10	2.0000028e-0	540	0.26	2.0000096e-0	11354	8.18

Table 3: **Numerical results for Experiment C**

	Den=0.1%			**Den**=10%		
N0.	**Obj.**	**Nf**	**Time(s)**	**Obj.**	**Nf**	**Time(s)**
1	4.0504722e-6	434	0.25	4.2542862e-6	4473	3.51
2	1.5335386e-6	506	0.31	4.8823994e-6	7448	5.47
3	4.7463833e-6	731	0.45	4.9623792e-6	5539	3.51
4	4.5687416e-6	525	0.31	4.6959189e-6	5664	3.72
5	3.0131432e-6	529	0.30	4.9137922e-6	5199	3.40
6	4.8877190e-6	910	0.44	4.9240652e-6	4991	3.90
7	4.8123156e-6	693	0.44	3.6413794e-6	5228	3.51
8	4.4356008e-6	843	0.44	4.9984977e-6	5178	4.55
9	4.3297744e-6	423	0.23	4.4443700e-6	5277	4.47
10	4.0444890e-7	605	0.26	4.9962998e-6	5650	4.31

Table 4: **Numerical results for Experiment D**

	Den=0.1%			**Den**=10%		
N0.	**Obj.**	**Nf**	**Time(s)**	**Obj.**	**Nf**	**Time(s)**
1	2.0000050e-0	506	0.30	2.0000048e-0	10384	8.20
2	2.0000033e-0	538	0.31	2.0000048e-0	11434	7.86
3	2.0000048e-0	831	0.44	2.0000043e-0	11252	8.84
4	2.0000042e-0	432	0.25	2.0000049e-0	12107	8.65
5	2.0000035e-0	396	0.25	2.0000049e-0	10428	7.75
6	2.0000048e-0	853	0.50	2.0000046e-0	10424	8.05
7	2.0000050e-0	546	0.33	2.0000045e-0	11610	7.92
8	2.0000032e-0	693	0.39	2.0000042e-0	12283	8.91
9	2.0000037e-0	488	0.28	2.0000047e-0	11758	8.06
10	2.0000048e-0	556	0.36	2.0000049e-0	12344	8.55