Appl. Comput. Harmon. Anal. 56 (2022) 98–122

Contents lists available at ScienceDirect

Applied and Computational Harmonic Analysis

www.elsevier.com/locate/acha

A proximal algorithm with backtracked extrapolation for a class of structured fractional programming ☆

Qia Li [a], Lixin Shen [b], Na Zhang [c,*], Junpeng Zhou [d]

[a] *School of Computer Science and Engineering, Guangdong Province Key Laboratory of Computational Science, Sun Yat-sen University, Guangzhou 510275, PR China*
[b] *Department of Mathematics, Syracuse University, Syracuse, NY 13244, USA*
[c] *Department of Applied Mathematics, College of Mathematics and Informatics, South China Agricultural University, Guangzhou 510642, PR China*
[d] *School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510275, PR China*

ARTICLE INFO

Article history:
Received 23 February 2021
Received in revised form 15 July 2021
Accepted 11 August 2021
Available online 17 August 2021
Communicated by Qingtang Jiang

MSC:
90C26
90C30
65K05

Keywords:
Fractional programming
Proximal algorithm
Backtracked extrapolation
Sparse recovery

ABSTRACT

In this paper, we consider a class of structured fractional minimization problems where the numerator part of the objective is the sum of a convex function and a Lipschitz differentiable (possibly) nonconvex function, while the denominator part is a convex function. By exploiting the structure of the problem, we propose a first-order algorithm, namely, a proximal-gradient-subgradient algorithm with backtracked extrapolation (PGSA_BE) for solving this type of optimization problem. It is worth pointing out that there are a few differences between our backtracked extrapolation and other popular extrapolations used in convex and nonconvex optimization. One of such differences is as follows: if the new iterate obtained from the extrapolated iteration satisfies a backtracking condition, then this new iterate will be replaced by the one generated from the non-extrapolated iteration. We show that any accumulation point of the sequence generated by PGSA_BE is a critical point of the problem regarded. In addition, by assuming that some auxiliary functions satisfy the Kurdyka-Łojasiewicz property, we are able to establish global convergence of the entire sequence, in the case where the denominator is locally Lipschitz differentiable, or its conjugate satisfies the calmness condition. Finally, we present some preliminary numerical results to illustrate the efficiency of PGSA_BE.

☆ Qia Li's work was supported in part by the Natural Science Foundation of China under grant 11971499 and the Guangdong Province Key Laboratory of Computational Science at the Sun Yat-sen University (2020B1212060032). Na Zhang's work was supported in part by the Natural Science Foundation of China under grant 11701189 and the Opening Project of Guangdong Province Key Laboratory of Computational Science at the Sun Yat-sen University under grant 2021001. The work of Lixin Shen was supported in part by the National Science Foundation under grant DMS-1913039.

* Corresponding author.
E-mail address: nzhsysu@gmail.com (N. Zhang).

https://doi.org/10.1016/j.acha.2021.08.004
1063-5203/

1. Introduction

Fractional optimization, which refers to the problem of minimizing or maximizing an objective involving one or several rations of functions, has been investigated for several decades. It encompasses a large class of nonconvex optimization problems. In this paper, we consider a class of fractional minimization problems which takes the form of

$$\min\left\{\frac{f(x)+h(x)}{g(x)} : x\in\Omega\right\}, \tag{1.1}$$

where $f,g,h:\mathbb{R}^n\to\overline{\mathbb{R}}:=(-\infty,+\infty]$ are proper lower semicontinuous functions and the set $\Omega:=\{x\in\mathbb{R}^n : g(x)\neq 0\}$ is nonempty. Through this paper, we adopt the following blanket assumptions on problem (1.1).

Assumption 1.

(a) f is convex and continuous on $\mathrm{dom}(f)$.
(b) g is convex, real-valued and positive on $\Omega\cap\mathrm{dom}(f)$.
(c) h is Lipschitz differentiable with a Lipschitz constant $L>0$.
(d) $f+h$ is non-negative on $\mathrm{dom}(f)$ and $f(x)+h(x)\neq 0$ for $x\in\mathbb{R}^n\backslash\Omega$.

Many optimization problems arising in applications, such as sparse recovery and machine learning, can be cast into problem (1.1). Roughly speaking, the task of sparse signal recovery is to find a sparse solution to the linear system $Ax=b$ where $A\in\mathbb{R}^{m\times n}$ and $b\in\mathbb{R}^m$ are given (e.g., see [9,10,14,15,29,31]). Next, we provide two concrete examples of problem (1.1) in sparse signal recovery.

Example 1 *(L_1/L_2 sparse signal recovery [25]).* The model has received considerable attention very recently. Let $\|\cdot\|_2$ and $\|\cdot\|_1$ denote the Euclidean norm and ℓ_1-norm respectively. It is in the form of

$$\min\left\{\frac{\|x\|_1}{\|x\|_2} : Ax=b,\ \underline{x}\le x\le\overline{x},\ x\in\mathbb{R}^n\right\}, \tag{1.2}$$

where $\underline{x},\overline{x}\in\mathbb{R}^n$ denote lower and upper bounds of the underlying signal. To deal with the equality constraint $Ax=b$, a penalty problem of (1.2) is considered in [32] as follows,

$$\min\left\{\frac{\lambda\|x\|_1+\frac{1}{2}\|Ax-b\|_2^2}{\|x\|_2} :\ \underline{x}\le x\le\overline{x},\ x\in\mathbb{R}^n\right\}, \tag{1.3}$$

where $\lambda>0$ is a penalty parameter. Clearly, problem (1.1) reduces to (1.2) when f is the sum of $\lambda\|\cdot\|_1$ and the indicator function on $\{x\in\mathbb{R}^n : \underline{x}\le x\le\overline{x}\}$, $g=\|\cdot\|_2$, $h=\frac{1}{2}\|A\cdot-b\|_2^2$ and $\Omega=\{x\in\mathbb{R}^n : x\neq 0\}$.

Example 2 *(L_1/S_K sparse signal recovery).* For $x\in\mathbb{R}^n$ and a positive integer K, we use $\|x\|_{(K)}$, the largest-K norm of x, to denote the sum of the K largest absolute values of entries in x. Motivated by the truncated ℓ_1 function $\|\cdot\|_1-\|\cdot\|_{(K)}$ (see, for example, [16]) and the scale invariant property of $\|\cdot\|_1/\|\cdot\|_2$, we introduce a scale invariant function $\|\cdot\|_1/\|\cdot\|_{(K)}$, i.e., the ratio of ℓ_1-norm and the largest-K norm, which we name L_1/S_K function. The L_1/S_K function can serve as a sparsity-promoting function due to its nondifferentiability at a vector with at least one zero element. We refer interested readers to [28] for a rigorous definition of the sparsity promoting function. By applying the L_1/S_K function, we obtain the following two models for sparse signal recovery

$$\min\left\{\frac{\|x\|_1}{\|x\|_{(K)}} : Ax=b,\ \underline{x}\le x\le\overline{x},\ x\in\mathbb{R}^n\right\}, \tag{1.4}$$

and

$$\min\left\{\frac{\lambda\|x\|_1+\frac{1}{2}\|Ax-b\|_2^2}{\|x\|_{(K)}}:\ \underline{x}\le x\le\overline{x},\ x\in\mathbb{R}^n\right\}. \tag{1.5}$$

It is clear that problem (1.1) reduces to problem (1.5) when f is the sum of $\lambda\|\cdot\|_1$ and the indicator function on $\{x\in\mathbb{R}^n:\underline{x}\le x\le\overline{x}\}$, $g=\|\cdot\|_{(K)}$, $h=\frac{1}{2}\|A\cdot-b\|_2^2$ and $\Omega=\{x\in\mathbb{R}^n:x\neq 0\}$.

The parametric approach, that relates a fractional optimization problem to its associated parametric problem [13,18], is one of the classical approaches for the fractional programming. By the parametric approach, problem (1.1) has an optimal solution $x^\star\in\mathbb{R}^n$ if and only if $x^\star$ is an optimal solution to the following optimization problem:

$$\min\{f(x)+h(x)-c_\star g(x):x\in\Omega\}, \tag{1.6}$$

where $c_\star=\frac{f(x^\star)+h(x^\star)}{g(x^\star)}$. It is worth noting that the optimal objective value $c_\star$ is unknown in general. Therefore, iterative algorithms, which may date back to the Dinkelbach's method [12], were proposed to remedy this issue (e.g., see [17,24,27]). More precisely, beginning with x^0, an initial estimate of x, the x^{k+1} in the k-th iteration is the solution of the following subproblem:

$$x^{k+1}\in\arg\min\{f(x)+h(x)-c_k g(x):x\in\Omega\}. \tag{1.7}$$

Here, c_k is renewed via $c_k:=\frac{f(x^k)+h(x^k)}{g(x^k)}$. However, problem (1.7) is in fact a nonconvex programming, and it is very difficult to obtain its optimal solutions generally.

A proximal-gradient algorithm has been proposed for a class of fractional optimization problems in [7], where the numerator is convex and the denominator is a smooth convex function. It can be suitably applied to problem (1.1) in the case of smooth g and convex h. The resulting algorithm computes the new iterate by

$$x^{k+1}\in\arg\min\left\{f(x)+h(x)-c_k\langle\nabla g(x^k),x\rangle+\frac{1}{2\eta_k}\|x-x^k\|_2^2:x\in\Omega\right\}, \tag{1.8}$$

where $\eta_k>0$ and c_k is the objective value of problem (1.1) at x^k. Very recently, it was proposed that a proximity-gradient-subgradient algorithm (PGSA) in [32] for solving problem (1.1), where f is allowed to be nonconvex. Given an iterate x^k, PGSA generates the new iterate by

$$x^{k+1}\in\arg\min\left\{f(x)+\langle\nabla h(x^k)-c_k y^k,x-x^k\rangle+\frac{1}{2\alpha_k}\|x-x^k\|_2^2:x\in\Omega\right\} \tag{1.9}$$

for some $y^k\in\partial g(x^k)$, $0<\alpha_k<1/L$ and $c_k=(f(x^k)+h(x^k))/g(x^k)$. Additionally, PGSA with line search (PGSA_L) is also developed in [32] for possible acceleration, which solves almost the same subproblems as (1.9) except that the line search technique is applied to seeking for a potentially larger step size α_k. Almost as early as PGSA_L was proposed, an extrapolated proximal subgradient algorithm (ePSG) was presented in [8] for solving a similar class of fractional programs to (1.1), which allows g to be weekly convex but requires h to be convex. It is worth noting that when the infimum of g over $\Omega\cap\mathrm{dom}(f)$ is zero, then all the extrapolation parameters of ePSG are required to be zero too, which actually makes the iterations of ePSG coincide with (1.9), i.e., no extrapolations are performed in the method. It has been shown that any accumulation point of the sequence generated by the aforementioned algorithms is a critical point of problem (1.1). Convergence of the entire sequence generated by these algorithms is further established by assuming that a certain potential function satisfies the Kurdyka-Łojasiewicz property and g is differentiable with a

locally Lipschitz continuous gradient. However, this requirement on g is not fulfilled in some applications such as Example 2. Thus, the analysis of sequential convergence can not be applied to the above algorithms for these applications.

Inspired by extrapolation techniques in accelerating the proximal-gradient type algorithms for convex and nonconvex optimization (see, for example, [5,22,30]), we introduce in this paper so-called backtracked extrapolation to possibly accelerate PGSA for solving problem (1.1). The proposed algorithm is called PGSA with backtracked extrapolation (PGSA_BE). In each iteration of PGSA_BE, the new iterate is obtained by (1.9) with an extrapolation step when the backtracked condition evaluated at this new iterate is violated. Otherwise, the next iterate is simply computed by (1.9). We prove that, for a general choice of extrapolation parameters which is independent of the function g, any accumulation point of the sequence generated by PGSA_BE is a critical point of problem (1.1). Furthermore, we establish global sequential convergence of the sequence generated by PGSA_BE in two cases: (i) g is locally Lipschitz differentiable and (ii) the conjugate of g satisfies the calmness condition. It is easy to check that Example 1 falls in both cases while Example 2 only falls in the second case. In fact, there are many convex functions whose conjugates satisfy the calmness condition, e.g., positively homogeneous functions, whose conjugate functions are indicator functions of some closed convex sets [3, Proposition 14.11]. Finally, we conduct numerical experiments on sparse signal recovery problems to demonstrate the efficiency of PGSA_BE.

The rest of this paper is organized as follows. In Section 2, we present some preliminary materials. In Section 3, we propose our algorithm PGSA_BE and show subsequential convergence of the sequence generated by PGSA_BE. The convergence of the entire sequence generated by PGSA_BE is established in Section 4. Numerical results are presented in Section 5. Finally, we conclude this paper in Section 6.

2. Notation and preliminaries

We begin with our notation. Let $\mathbb{N}$ be the set of nonnegative integers. For $n \in \mathbb{N}$, we denote the n-dimensional Euclidean space by $\mathbb{R}^n$ and the standard inner product by $\langle \cdot, \cdot \rangle$. The Euclidean norm and ℓ_1-norm are denoted by $\|\cdot\|_2$ and $\|\cdot\|_1$ respectively. For a nonempty closed set $S \subseteq \mathbb{R}^n$, the indicator function on S is defined by

$$\iota_S(x) := \begin{cases} 0, & \text{if } x \in S, \\ +\infty, & \text{otherwise.} \end{cases}$$

Also, the distance from a point $x \in \mathbb{R}^n$ to S is denoted by $\operatorname{dist}(x, S) := \inf\{\|x - y\|_2 : y \in S\}$.

In the remaining part of this section, we introduce some technical preliminaries on subdifferential of nonconvex functions [20,26] and the Kurdyka-Łojasiewicz property [1].

2.1. Fréchet subdifferential

For an extended-real-valued function $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}}$, its domain is defined by $\operatorname{dom}(\varphi) := \{x \in \mathbb{R}^n : \varphi(x) < +\infty\}$. The Fréchet subdifferential of φ at $x \in \operatorname{dom}(\varphi)$, written as $\widehat{\partial}\varphi(x)$, is defined as follows:

$$\widehat{\partial}\varphi(x) := \left\{ y \in \mathbb{R}^n : \liminf_{\substack{z \to x \\ z \neq x}} \frac{\varphi(z) - \varphi(x) - \langle y, z - x \rangle}{\|z - x\|_2} \geq 0 \right\}.$$

The limiting (Fréchet) subdifferential, or simply the subdifferential for short, of φ at $x \in \operatorname{dom}(\varphi)$, is defined by

$$\partial\varphi(x) := \{y \in \mathbb{R}^n : \exists x^k \to x,\ \varphi(x^k) \to \varphi(x),\ y^k \in \widehat{\partial}\varphi(x^k) \to y\}.$$

It is obvious that $\widehat{\partial}\varphi(x) \subseteq \partial\varphi(x)$ for all $x \in \mathbb{R}^n$, where $\widehat{\partial}\varphi(x)$ is closed and convex, and $\partial\varphi(x)$ is closed. If φ is differentiable at x, then $\widehat{\partial}\varphi(x) = \{\nabla\varphi(x)\}$ with $\nabla\varphi(x)$ being the gradient of φ at x. If φ is continuously differentiable at x, then $\partial\varphi(x) = \{\nabla\varphi(x)\}$. For a convex function φ, the above subdifferentials reduce to the classical subdifferential [26, Proposition 8.12].

$$\widehat{\partial}\varphi(x) = \partial\varphi(x) = \{y \in \mathbb{R}^n : \varphi(z) - \varphi(x) - \langle y, z - x\rangle \geq 0, \forall z \in \mathbb{R}^n\}.$$

Moreover, for $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}}$, we use φ^* to denote the Fenchel conjugate function of φ, that is, for $y \in \mathbb{R}^n$

$$\varphi^*(y) := \sup\{\langle y, x\rangle - \varphi(x) : x \in \mathbb{R}^n\}.$$

If φ is a proper lower semicontinuous convex function, then $y \in \partial\varphi(x)$ if and only if $x \in \partial\varphi^*(y)$. We also need the notion of partial subdifferential. Let the variable x be decomposed into $p+1$ separated blocks $x_0, x_1, \ldots, x_p$ for $p \in \mathbb{N}$. For each x_i and fixing the other p blocks $x_0, x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_p$, we denote the Fréchet subdifferential of the function $\varphi(x_0, x_1, \ldots, x_{i-1}, \ \cdot \ , x_{i+1}, \ldots, x_p)$ at u by $\widehat{\partial}_{x_i}\varphi(x_0, x_1, \ldots, x_{i-1}, u, x_{i+1}, \ldots, x_p)$.

Next we recall some useful calculus results on Fréchet subdifferential. For any $\alpha > 0$ and $x \in \mathbb{R}^n$, $\widehat{\partial}(\alpha\varphi)(x) = \alpha\widehat{\partial}\varphi(x)$. Let $\varphi_1, \varphi_2 : \mathbb{R}^n \to (-\infty, +\infty]$ be proper lower semicontinuous. Then we have $\widehat{\partial}(\varphi_1 + \varphi_2)(x) \supseteq \widehat{\partial}\varphi_1(x) + \widehat{\partial}\varphi_2(x)$ for $x \in \mathrm{dom}(\varphi_1 + \varphi_2)$. Furthermore, if φ_2 is differentiable at x, then $\widehat{\partial}(\varphi_1 + \varphi_2)(x) = \widehat{\partial}\varphi_1(x) + \nabla\varphi_2(x)$. It was presented in [20, Corollary 1.111 and Proposition 3.45] some quotient rules for limiting subdifferential of φ_1/φ_2 at $\bar{x}$ with $\varphi_2(\bar{x}) \neq 0$ when φ_1 and φ_2 are assumed to be locally Lipschitz continuous around $\bar{x}$. Unfortunately, these quotient rules are not available at the border of $\mathrm{dom}(\varphi_1)$ if $\mathrm{dom}(\varphi_1) \neq \mathbb{R}^n$, since in this case the local Lipschitz continuity is not satisfied. Hence, we shall derive some rules for the Fréchet subdifferential $\widehat{\partial}(\varphi_1/\varphi_2)$ which can be used for $\bar{x}$ at the border of $\mathrm{dom}(\varphi_1)$. To this end, we first assume that $\mathrm{dom}(\varphi_2) = \mathbb{R}^n$ and introduce two functions defined by the quotient of φ_1 and φ_2.

We define $\psi : \mathbb{R}^n \to (-\infty, +\infty]$ at $x \in \mathbb{R}^n$ as

$$\psi(x) := \begin{cases} \frac{\varphi_1(x)}{\varphi_2(x)}, & \text{if } x \in \mathrm{dom}(\varphi_1) \text{ and } \varphi_2(x) \neq 0, \\ +\infty, & \text{else.} \end{cases}$$

Given $d > 0$, let $\rho : \mathbb{R}^n \times \mathbb{R}^n \to (-\infty, +\infty]$ be defined at $(x, y) \in \mathbb{R}^n \times \mathbb{R}^n$ as

$$\rho(x, y) = \begin{cases} \frac{\varphi_1(x)}{\langle x, y\rangle - \varphi_2^*(y)}, & (x, y) \in \mathrm{dom}(\varphi_1) \times \mathrm{dom}(\varphi_2^*) \text{ and } \langle x, y\rangle - \varphi_2^*(y) \geq d, \\ +\infty, & \text{else.} \end{cases}$$

We also need the concept of calmness condition.

Definition 2.1 *(Calmness condition [26]).* The function $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}}$ is said to satisfy the calmness condition at $x \in \mathrm{dom}(\varphi)$ (resp., relative to $S \subseteq \mathbb{R}^n$), if there exist $\kappa > 0$ and a neighborhood $\mathcal{O}$ of x, such that

$$|\varphi(u) - \varphi(x)| \leq \kappa\|u - x\|_2$$

for all $u \in \mathcal{O}$ (resp., $u \in \mathcal{O} \cap S$). We say φ satisfies the calmness condition on S if φ satisfies the calmness condition at any point in S relative to S.

The following two propositions concern the Fréchet subdifferentials of ψ and ρ respectively.

Proposition 2.2 *([32]). Let $x \in \mathrm{dom}(\psi)$ with $a_1 = \varphi_1(x)$ and $a_2 = \varphi_2(x) > 0$. Suppose that φ_1 is continuous at x relative to $\mathrm{dom}(\varphi_1)$ and φ_2 satisfies the calmness condition at x. Then*

$$\widehat{\partial}\psi(x) = \frac{\widehat{\partial}(a_2\varphi_1 - a_1\varphi_2)(x)}{a_2^2}.$$

Furthermore, if φ_2 is Fréchet differential at x, then

$$\widehat{\partial}\psi(x) = \frac{a_2\widehat{\partial}\varphi_1(x) - a_1\nabla\varphi_2(x)}{a_2^2}.$$

Proposition 2.3. *Let $(x,y) \in \mathrm{dom}(\rho)$ with $a_1 = \varphi_1(x) > 0$ and $a_2 = \langle x, y\rangle - \varphi_2^*(y) > d$. Suppose that φ_1 is continuous at x relative to $\mathrm{dom}(\varphi_1)$ and φ_2^* satisfies the calmness condition at y relative to $\mathrm{dom}(\varphi_2^*)$. Then*

$$\widehat{\partial}\rho(x,y) = \widehat{\partial}_x\rho(x,y) \times \widehat{\partial}_y\rho(x,y),$$

where

$$\widehat{\partial}_x\rho(x,y) = \frac{a_2\widehat{\partial}\varphi_1(x) - a_1 y}{a_2^2}, \quad \widehat{\partial}_y\rho(x,y) = \frac{\widehat{\partial}(a_1\varphi_2^*)(y) - a_1 x}{a_2^2}.$$

The proof is given in Appendix A.

2.2. Kurdyka-Łojasiewicz (KL) property

Definition 2.4 *(KL property [1]).* A proper function $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}}$ is said to satisfy the KL property at $\hat{x} \in \mathrm{dom}(\partial\varphi)$ if there exist $\eta \in (0, +\infty]$, a neighborhood O of $\hat{x}$ and a continuous concave function $\phi : [0,\eta) \to \mathbb{R}_+ := [0, +\infty)$, such that:

(i) $\phi(0) = 0$,
(ii) ϕ is continuously differentiable on $(0,\eta)$ with $\phi' > 0$,
(iii) For any $x \in O \cap \{x \in \mathbb{R}^n : \varphi(\hat{x}) < \varphi(x) < \varphi(\hat{x}) + \eta\}$, there holds $\phi'(\varphi(x) - \varphi(\hat{x}))\ \mathrm{dist}(0, \partial\varphi(x)) \geq 1$.

A proper lower semicontinuous function $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}}$ is called a KL function if φ satisfies the KL property at all points in $\mathrm{dom}(\partial\varphi)$. A wide range of functions is KL functions. Among those functions, the proper lower semicontinuous semialgebraic functions (see [2]) cover most frequently appeared functions in applications. Recall that a function $\varphi : \mathbb{R}^n \to \bar{\mathbb{R}}$ is semialgebraic if its graph $\mathrm{Graph}(\varphi) := \{(x,s) \in \mathbb{R}^n \times \mathbb{R} : s = \varphi(x)\}$ is a semialgebraic subset of $\mathbb{R}^{n+1}$, that is, there exist a finite number of real polynomial functions P_{ij}, $Q_{ij} : \mathbb{R}^{n+1} \to \mathbb{R}$ such that

$$\mathrm{Graph}(\varphi) = \bigcup_{j=1}^{p} \bigcap_{i=1}^{q} \{y \in \mathbb{R}^{n+1} : P_{ij}(y) = 0,\ Q_{ij}(y) < 0\}.$$

We also need the following result regrading the uniformed KL property in [1, Lemma 6].

Lemma 2.5 *(Uniformized KL property). Let $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper lower semi-continuous function and $\Gamma \subseteq \mathbb{R}^n$ be a compact set. Assume that φ is constant on Γ and satisfies the KL property at each point of Γ. Then, there exist $\delta > 0$, $\eta > 0$ and a continuous concave function $\phi : [0,\eta) \to \mathbb{R}_+$ satisfying Definition 2.4* (i) - (ii) *such that*

$$\phi'(\varphi(x) - \varphi(\hat{x}))\ \mathrm{dist}(0, \partial\varphi(x)) \geq 1$$

holds for any $\hat{x} \in \Gamma$ and $x \in \{x \in \mathbb{R}^n :\ \mathrm{dist}(x,\Gamma) < \delta,\ \varphi(\hat{x}) < \varphi(x) < \varphi(\hat{x}) + \eta\}$.

An abstract framework is provided in [2] for proving global sequential convergence based on the KL property. We review this result in the following proposition.

Proposition 2.6. *Let $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper lower semicontinuous function. Consider a sequence $\{x^k : k \in \mathbb{N}\}$ satisfying the following three conditions:*

(i) *(Sufficient decrease condition.) There exists $a > 0$ such that*

$$\varphi(x^{k+1}) + a\|x^{k+1} - x^k\|_2^2 \leq \varphi(x^k)$$

holds for any $k \in \mathbb{N}$;

(ii) *(Relative error condition.) There exist $b > 0$ and $\omega^{k+1} \in \partial\varphi(x^{k+1})$ such that*

$$\|\omega^{k+1}\|_2 \leq b\|x^{k+1} - x^k\|_2$$

holds for any $k \in \mathbb{N}$;

(iii) *(Continuity condition.) There exist a subsequence $\{x^{k_j} : j \in \mathbb{N}\}$ and $x^\star$ such that*

$$x^{k_j} \to x^\star \text{ and } \varphi(x^{k_j}) \to \varphi(x^\star), \text{ as } j \to \infty.$$

If φ satisfies the KL property at $x^\star$, then $\sum_{k=1}^{\infty} \|x^k - x^{k-1}\|_2 < +\infty$, $\lim_{k\to\infty} x^k = x^\star$ and $0 \in \partial\varphi(x^\star)$.

Following a similar line of arguments to Proposition 2.6, we generalize this framework in the next proposition.

Proposition 2.7. *Let $H : \mathbb{R}^n \times \mathbb{R}^m \to \overline{\mathbb{R}}$ be proper lower semicontinuous. Consider a bounded sequence $\{(u^k, v^k) \in \mathbb{R}^n \times \mathbb{R}^m : k \in \mathbb{N}\}$ satisfying the following three conditions:*

(i) *(Sufficient decrease condition.) There exist $a > 0$ and $K_1 > 0$ such that*

$$H(u^{k+1}, v^{k+1}) + a\|u^{k+1} - u^k\|_2^2 \leq H(u^k, v^k)$$

holds for any $k \geq K_1$;

(ii) *(Relative error condition.) There exist $b > 0$, $K_2 > 0$ and $\omega^{k+1} \in \partial H(u^{k+1}, v^{k+1})$ such that*

$$\|\omega^{k+1}\|_2 \leq b\|u^{k+1} - u^k\|_2$$

holds for any $k \geq K_2$.

(iii) *(Continuity condition.) $\xi := \lim_{k\to\infty} H(u^k, v^k)$ exists and $H \equiv \xi$ on Υ, where Υ denotes the set of accumulation points of $\{(u^k, v^k) : k \in \mathbb{N}\}$.*

If H satisfies the KL property at each point of Υ, then $\sum_{k=1}^{\infty} \|u^k - u^{k-1}\|_2 < +\infty$, $\lim_{k\to\infty} u^k = u^\star$ and $0 \in \partial H(u^\star, v^\star)$ for any $(u^\star, v^\star) \in \Upsilon$.

The proof is given in Appendix B.

3. The proximity-gradient-subgradient algorithm with backtracked extrapolation

In this section, we present our proximity-gradient-subgradient algorithm with backtracked extrapolation (PGSA_BE) for problem (1.1) and show its subsequential convergence.

Motivated by the success of extrapolation techniques used in convex and nonconvex optimization, we incorporate extrapolation to PGSA in (1.9) for possible acceleration. Moreover, the extrapolation used here is backtracked in each iteration, in the sense that an iteration without extrapolation ($\beta_k = 0$) will be performed instead if the backtracking condition is satisfied for the iterate x^{k+1} generated via a extrapolation step. We call the above extrapolation technique backtracked extrapolation. In particular, we present PGSA_BE for solving problem (1.1) in Algorithm 1.

Algorithm 1 PGSA with backtracked extrapolation (PGSA_BE) for solving (1.1).

Step 0. Input $x^{-1} = x^0 \in \Omega \cap \mathrm{dom}(f)$, $0 < \alpha \le 1/L$,
$l = 0$ if h is convex and $l = L$ else, $0 < \bar{\beta} < \sqrt{L/(L+l)}$,
$\{\beta_k : k \in \mathbb{N}\} \subseteq [0, \bar{\beta}]$, $0 < \epsilon < 1 - \bar{\beta}^2(1+\alpha l)$. Set $k \leftarrow 0$.

Step 1. Compute
$u^{k+1} = x^k + \beta_k(x^k - x^{k-1})$, // Extrapolation
$y^{k+1} \in \partial g(x^k)$,
$c_k = \dfrac{f(x^k) + h(x^k)}{g(x^k)}$,
$x^{k+1} = \mathrm{prox}_{\alpha f}(u^{k+1} - \alpha \nabla h(u^{k+1}) + \alpha c_k y^{k+1})$.

Step 2. If $\dfrac{g(x^{k+1})}{g(x^k)} < \dfrac{\beta_k^2(1+\alpha l)}{1-\epsilon}$,
set $x^{k+1} = \mathrm{prox}_{\alpha f}(x^k - \alpha \nabla h(x^k) + \alpha c_k y^{k+1})$. // Backtracking

Step 3. Set $k \leftarrow k+1$ and go to Step 1.

Before conducting the convergence analysis, we make some remarks on PGSA_BE. Since $\frac{\beta_k^2(1+\alpha l)}{1-\epsilon} \le \frac{\bar{\beta}^2(1+\alpha l)}{1-\epsilon} < 1$, the backtracking condition intuitively means that $g(x^{k+1})/g(x^k)$ is unexpectedly small, which is not preferred in the algorithm. In the very special case of $g \equiv 1$, the backtracked condition is never satisfied and PGSA_BE coincides with the extrapolated proximal gradient algorithm in [30].

Besides the backtracked extrapolation, PGSA_BE differs in several aspects from ePSG developed very recently in [8], which also uses some extrapolation technique for fractional optimization. For convenience, we denote the supremum and infimum of g over $\Omega \cap \mathrm{dom}(f)$ by M_1 and M_2. First, the extrapolation parameter $\{\beta_k : k \in \mathbb{N}\}$ in PGSA_BE is required to be in $[0, \bar{\beta}]$ with $0 < \bar{\beta} < \sqrt{L/(L+l)}$, where $l = 0$ if h is convex and $l = L$ otherwise. Specially, if h is convex, the requirement of $\{\beta_k : k \in \mathbb{N}\}$ reduces to $\{\beta_k : k \in \mathbb{N}\} \subseteq [0, 1)$ and $\sup\{\beta_k : k \in \mathbb{N}\} < 1$, which is general enough to cover the popular choice of the extrapolation parameters used in restart FISTA (see, for example, [4,23]). However, the choice of extrapolation parameters in ePSG relies on M_1 and M_2, and thus one has to estimate them before applying the algorithm. Second, in the case of $M_2 = 0$, ePSG reduces to PGSA which generates the new iterate by (1.9), i.e., no extrapolation is involved in the algorithm. In contrast to ePSG, after computing the new iterate by extrapolation, PGSA_BE incorporates a backtracked procedure that determines whether or not this new iterate will be used. Finally, when h is convex, to make the extrapolation parameters general enough, the step size for ∇h in ePSG should be in $(0, \frac{M_2}{(M_1+M_2)L})$ and thus may render a slow convergence if $\frac{M_2}{M_1}$ is small. Nevertheless, the step size α of PGSA_BE can be chosen in $(0, 1/L]$, which is independent of g and much larger than that of ePSG. Hence, PGSA_BE generally has a faster convergence than ePSG.

In what follows, we study the subsequential convergence of PGSA_BE. For convenience, we define $F : \mathbb{R}^n \to \overline{\mathbb{R}}$ at $x \in \mathbb{R}^n$ as

$$F(x) := \begin{cases} \frac{f(x)+h(x)}{g(x)}, & \text{if } x \in \Omega \cap \mathrm{dom}(f), \\ +\infty, & \text{else.} \end{cases}$$

Then problem (1.1) can be equivalently rewritten as

$$\min\{F(x) : x \in \mathbb{R}^n\}.$$

We recall the following definition of critical points in [32, Definition 3.4], where it is shown that any local minimizer of F is a critical point of F.

Definition 3.1. Let $x^\star \in \mathrm{dom}(F)$ and $c_\star = F(x^\star)$. We say that $x^\star$ is a critical point of F if

$$0 \in \partial f(x^\star) + \nabla h(x^\star) - c_\star \partial g(x^\star).$$

The definition of critical points (Definition 3.1) differs from the standard one $0 \in \widehat{\partial} F(x^\star)$. By Proposition 2.2 and Assumption 1, we have that

$$\begin{aligned} \widehat{\partial} F(x^\star) &= \frac{\widehat{\partial}\Big(g(x^\star)(f+h) - (f(x^\star)+h(x^\star))g\Big)(x^\star)}{(g(x^\star))^2} \\ &\subseteq \frac{1}{g(x^\star)}\Big(\partial f(x^\star) + \nabla h(x^\star) - c_\star \partial g(x^\star)\Big) \end{aligned}$$

where the last relation follows from the difference rule of Fréchet subdifferential [21, Theorem 3.1 (i)]. In view of Definition 3.1, $0 \in \widehat{\partial} F(x^\star)$ indicates that $x^\star$ is a critical point of F. However, the converse implication is generally not true. Specially, as pointed out in [32], in the special case that g is differentiable, Definition 3.1 coincides with $0 \in \widehat{\partial} F(x^\star)$. Below we present a lemma, which will be used later in establishing the subsequential convergence.

Lemma 3.2. *PGSA_BE generates a sequence* $\{x^k : k \in \mathbb{N}\} \subseteq \mathrm{dom}(F)$ *that satisfies*

$$f(x^{k+1}) + h(x^{k+1}) + \frac{1}{2\alpha}\|x^{k+1} - x^k\|_2^2 \le c_k g(x^{k+1}) + \frac{\beta_k^2(1/\alpha + l)}{2}\|x^k - x^{k-1}\|_2^2. \tag{3.1}$$

Proof. We prove this lemma by induction. First, the initial points $x^{-1} = x^0 \in \mathrm{dom}(F)$. Suppose $x^{-1}, x^0, \ldots, x^k \in \mathrm{dom}(F)$ for some $k \in \mathbb{N}$. By the definition of proximity operator and the convexity of f, we derive from PGSA_BE that

$$\frac{1}{\alpha}\Big(u^{k+1} - x^{k+1} - \alpha\nabla h(u^{k+1}) + \alpha c_k y^{k+1}\Big) \in \partial f(x^{k+1}), \tag{3.2}$$

which implies

$$f(x^{k+1}) + \frac{1}{\alpha}\langle u^{k+1} - x^{k+1} - \alpha\nabla h(u^{k+1}) + \alpha c_k y^{k+1}, x^k - x^{k+1}\rangle \le f(x^k). \tag{3.3}$$

Due to $u^{k+1} = x^k + \beta_k(x^k - x^{k-1})$ and the fact that $\langle a, b\rangle = \frac{1}{2}(\|a\|_2^2 + \|b\|_2^2 - \|a-b\|_2^2)$ for $a, b \in \mathbb{R}^n$, it follows from (3.3) that

$$\begin{aligned} &f(x^{k+1}) + \langle x^{k+1} - x^k, \nabla h(u^{k+1}) - c_k y^{k+1}\rangle + \frac{1}{2\alpha}\|u^{k+1} - x^{k+1}\|_2^2 + \frac{1}{2\alpha}\|x^{k+1} - x^k\|_2^2 \\ \le& f(x^k) + \frac{\beta_k^2}{2\alpha}\|x^k - x^{k-1}\|_2^2. \end{aligned} \tag{3.4}$$

Since ∇h is Lipschitz continuous with constant L, there hold

$$h(x^{k+1}) \leq h(u^{k+1}) + \langle \nabla h(u^{k+1}), x^{k+1} - u^{k+1} \rangle + \frac{L}{2}\|u^{k+1} - x^{k+1}\|_2^2, \tag{3.5}$$

$$h(u^{k+1}) + \langle \nabla h(u^{k+1}), x^k - u^{k+1} \rangle \leq h(x^k) + \frac{l}{2}\|x^k - u^{k+1}\|_2^2. \tag{3.6}$$

From the convexity of g and $c_k \geq 0$, we get

$$c_k g(x^k) + \langle c_k y^{k+1}, x^{k+1} - x^k \rangle \leq c_k g(x^{k+1}). \tag{3.7}$$

By summing (3.4)-(3.7), we obtain (3.1) from $\alpha \leq 1/L$ and $c_k g(x^k) = f(x^k) + h(x^k)$.

Finally, we prove $x^{k+1} \in \mathrm{dom}(F) = \mathrm{dom}(f) \cap \Omega$. It is obvious that $x^{k+1} \in \mathrm{dom}(f)$ and it suffices to show $x^{k+1} \in \Omega$, i.e., $g(x^{k+1}) \neq 0$. If the extrapolation step produces an iterate x^{k+1} such that $g(x^{k+1}) = 0$, then the backtracking condition is surely satisfied and thus a non-extrapolation step ($\beta_k = 0$) is applied instead. Next, we shall show that $g(x^{k+1}) \neq 0$ in the case of $\beta_k = 0$ by contradiction. Assume that $g(x^{k+1}) = 0$ and $\beta_k = 0$. Then we obtain from (3.1) that

$$f(x^{k+1}) + h(x^{k+1}) + \frac{1}{2\alpha}\|x^{k+1} - x^k\|_2^2 \leq 0.$$

Hence, we deduce that $x^{k+1} = x^k$ since $f + h \geq 0$. This contradicts to $x^k \in \Omega$ and we conclude that $\{x^k : k \in \mathbb{N}\} \subseteq \mathrm{dom}(F)$. □

Proposition 3.3. *Let $\{x^k : k \in \mathbb{N}\}$ be generated by PGSA_BE. Then, the following statements hold:*

(i) $F(x^{k+1}) + \dfrac{\|x^{k+1} - x^k\|_2^2}{2\alpha g(x^{k+1})} \leq F(x^k) + (1-\epsilon)\dfrac{\|x^k - x^{k-1}\|_2^2}{2\alpha g(x^k)}$ *for* $k \in \mathbb{N}$;

(ii) $\lim\limits_{k\to\infty} \dfrac{\|x^k - x^{k-1}\|_2^2}{g(x^k)} = 0$;

(iii) $\lim\limits_{k\to\infty} c_k = \lim\limits_{k\to\infty} F(x^k) = c_\star$ *exists;*

(iv) *Let $x^\star$ be any accumulation point of $\{x^k : k \in \mathbb{N}\}$. Then $x^\star \in \mathrm{dom}(F)$ and $F(x^\star) = c_\star$.*

Proof. We first prove Item (i). From Lemma 3.2, we know $g(x^k) \neq 0$ for $k \in \mathbb{N}$. By dividing $g(x^{k+1})$ on both sides of (3.1), we have

$$\begin{aligned} F(x^{k+1}) + \frac{\|x^{k+1} - x^k\|_2^2}{2\alpha g(x^{k+1})} &\leq F(x^k) + \frac{\beta_k^2(1/\alpha + l)}{2g(x^{k+1})}\|x^k - x^{k-1}\|_2^2 \\ &= F(x^k) + \frac{\beta_k^2(1+\alpha l)g(x^k)/g(x^{k+1})}{2\alpha g(x^k)}\|x^k - x^{k-1}\|_2^2 \\ &\leq F(x^k) + (1-\epsilon)\frac{\|x^k - x^{k-1}\|_2^2}{2\alpha g(x^k)}, \end{aligned}$$

where the last inequality follows from the backtracking step.

We next prove Item (ii). Summing the both sides of Item (i) from $k = 0$ to $K \in \mathbb{N}$, we obtain that

$$F(x^{K+1}) + \frac{\|x^{K+1} - x^K\|_2^2}{2\alpha g(x^{K+1})} + \frac{\epsilon}{2\alpha}\sum_{k=1}^{K}\frac{\|x^k - x^{k-1}\|_2^2}{g(x^k)} \leq F(x^0). \tag{3.8}$$

Then, Item (ii) follows immediately.

We next prove Item (iii). Item (i) implies that the sequence $\{F(x^k)+\frac{\|x^k-x^{k-1}\|_2^2}{2\alpha g(x^k)} : k \in \mathbb{N}\}$ is nonincreasing. Additionally, this sequence is also bounded below by 0. In view of Item (ii) and the aforementioned fact, we deduce that $\lim_{k\to\infty} F(x^k) = c_\star$ exists.

Finally, we prove Item (iv). Let $x^\star$ be an accumulation point of $\{x^k : k \in \mathbb{N}\}$ and $\{x^{k_j} : j \in \mathbb{N}\}$ be a subsequence such that $\lim_{j\to\infty} x^{k_j} = x^\star$. According to Lemma 3.2, it holds that

$$f(x^{k_j}) + h(x^{k_j}) + \frac{1}{2\alpha}\|x^{k_j} - x^{k_j-1}\|_2^2 \le c_{k_j-1} g(x^{k_j}) + \frac{\beta_{k_j-1}^2(1/\alpha + l)}{2}\|x^{k_j-1} - x^{k_j-2}\|_2^2. \tag{3.9}$$

By Proposition 3.3 (ii) and the continuity of g, we have

$$\lim_{j\to\infty} \|x^{k_j} - x^{k_j-1}\|_2^2 = \lim_{j\to\infty} g(x^{k_j})\frac{\|x^{k_j} - x^{k_j-1}\|_2^2}{g(x^{k_j})} = 0, \tag{3.10}$$

which implies that $\lim_{j\to\infty} x^{k_j-1} = \lim_{j\to\infty} x^{k_j} = x^\star$. Using this and the boundedness of $\{\beta_k : k \in \mathbb{N}\}$, we see that

$$\lim_{j\to\infty} \beta_{k_j-1}^2\|x^{k_j-1} - x^{k_j-2}\|_2^2 = \lim_{j\to\infty} \beta_{k_j-1}^2 g(x^{k_j-1})\frac{\|x^{k_j-1} - x^{k_j-2}\|_2^2}{g(x^{k_j-1})} = 0. \tag{3.11}$$

From Item (iii), (3.10) and (3.11), we have upon passing to the limit in (3.9) that $f(x^\star) + h(x^\star) \le c_\star g(x^\star)$. This together with the fact that $f + h > 0$ on $\mathbb{R}^n\backslash\Omega$ indicates that $x^\star \in \mathrm{dom}(F)$. Since F is continuous on $\mathrm{dom}(F)$, we conclude that $F(x^\star) = c_\star$. □

Now we are ready to show a subsequential convergence result of PGSA_BE for problem (1.1).

Theorem 3.4. *Let $\{x^k : k \in \mathbb{N}\}$ be generated by PGSA_BE. Then any accumulation point of $\{x^k : k \in \mathbb{N}\}$ is a critical point of F.*

Proof. Let $x^\star$ be an accumulation point of $\{x^k : k \in \mathbb{N}\}$ and $\{x^{k_j} : j \in \mathbb{N}\}$ be a subsequence such that $\lim_{j\to\infty} x^{k_j} = x^\star$. Since g is a real-valued convex function and $\{x^{k_j-1} : j \in \mathbb{N}\}$ is bounded, we know that $\{y^{k_j} : j \in \mathbb{N}\}$ is bounded. Without loss of generality, we may assume $\lim_{j\to\infty} y^{k_j}$ exists and $\lim_{j\to\infty} y^{k_j} = y^\star \in \partial g(x^\star)$ due to the closedness of operator ∂g. From the iteration of PGSA_BE, we have

$$x^{k_j} \in \mathrm{prox}_{\alpha f}\Big(u^{k_j} - \alpha\nabla h(u^{k_j}) + \alpha c_{k_j-1} y^{k_j}\Big). \tag{3.12}$$

As ∇h and F is continuous on $\mathrm{dom}(F)$, we obtain by Proposition 3.3 (iv) and passing to the limit in (3.12) that

$$x^\star \in \mathrm{prox}_{\alpha f}\Big(x^\star - \alpha\nabla h(x^\star) + \alpha F(x^\star) y^\star\Big).$$

By the definition of the proximity operator and the generalized *Fermat's Rule*, we deduce that $x^\star$ is a critical point of F. □

4. Global sequence convergence of PGSA_BE

We investigate in this subsection the global convergence of the entire sequence $\{x^k : k \in \mathbb{N}\}$ generated by PGSA_BE. We shall show $\{x^k : k \in \mathbb{N}\}$ converges to a critical point of F under suitable assumptions. To this end, we need to make two assumptions throughout this subsection as follows:

Assumption 2. Function F is level-bounded, i.e., for any $\gamma \in \mathbb{R}$, the level set $\{x \in \mathbb{R}^n : F(x) \leq \gamma\}$ is bounded.

Assumption 3. Function f is locally Lipschitz continuous on $\mathrm{dom}(f)$, i.e., for all $x \in \mathrm{dom}(f)$, there exist $L_x > 0$ and a neighborhood $\mathcal{O}$ of x, such that $|f(\hat{x}) - f(\tilde{x})| \leq L_x \|\hat{x} - \tilde{x}\|_2$ holds for all $\hat{x}, \tilde{x} \in \mathcal{O} \cap \mathrm{dom}(f)$.

Under Assumption 2, we have the following results regarding the sequence generated by PGSA_BE.

Proposition 4.1. *Let $\{(x^k, y^k) : k \in \mathbb{N}\}$ be generated by PGSA_BE. Suppose Assumption 2 holds. Then the following statements hold:*

(i) $\{(x^k, y^k) : k \in \mathbb{N}\}$ *is bounded;*
(ii) *There exist $0 < d_1 < d_2$ such that $d_1 \leq g(x^k) \leq d_2$ for all $k \in \mathbb{N}$;*
(iii) $\lim_{k\to\infty} \|x^{k+1} - x^k\|_2 = 0$;
(iv) $\lim_{k\to\infty} \dfrac{g(x^{k-1})}{g(x^k)} = 1$.

Proof. We first prove Item (i). Proposition 3.3 (i) indicates that $F(x^k) \leq F(x^0)$ for all $k \in \mathbb{N}$. This together with Assumption 2 leads to the boundedness of $\{x^k : k \in \mathbb{N}\}$. Since g is real-valued convex and $y^{k+1} \in \partial g(x^k)$, we deduce that $\{y^k : k \in \mathbb{N}\}$ is also bounded.

Next we show Items (ii) and (iii). According to Item (i) and the continuity of g, there exists $d_2 > 0$ such that $g(x^k) \leq d_2$ for all $k \in \mathbb{N}$. In addition, by Lemma 3.2 and Proposition 3.3 (iv), we note that $g(x^k) > 0$ for all $k \in \mathbb{N}$ and any accumulation point $x^\star$ of $\{x^k : k \in \mathbb{N}\}$ satisfies $g(x^\star) > 0$. Hence, we claim that $g(x^k) \geq d_1$ for some $d_1 > 0$, thanks to Item (i) and the continuity of g. Item (iii) follows from $g(x^k) \leq d_2$ and Proposition 3.3 (ii).

Finally we prove Item (iv). Let $S \subseteq \mathbb{R}^n$ be a bounded closed set satisfying $\{x^k : k \in \mathbb{N}\} \subseteq S \subseteq \mathrm{dom}(F)$. Then it is easy to verify that g is globally Lipschitz continuous on S since g is real-valued and convex. Hence, we have

$$\lim_{k\to\infty} \left| \frac{g(x^k)}{g(x^{k+1})} - 1 \right| = \lim_{k\to\infty} \left| \frac{g(x^k) - g(x^{k+1})}{g(x^{k+1})} \right| \leq \lim_{k\to\infty} \frac{|g(x^k) - g(x^{k+1})|}{d_1} = 0,$$

where the second inequality follows from Item (ii) and the last equality follows from Item (iii). □

Next, we assume either g is continuously differentiable on Ω with a locally Lipschitz continuous gradient or $g^\star$ satisfies the calmness condition on $\mathrm{dom}(g^*)$. It is worth noting that each of the two assumptions can not be deduced from the other one. The two examples below illustrate this point. Let $g(x) := \sum_{i=1}^n \sqrt{x_i^2 + 1}$ for $x \in \mathbb{R}^n$. Then we have that

$$g^*(y) = \begin{cases} -\sum_{i=1}^n \sqrt{1 - y_i^2}, & \|y\|_\infty \leq 1, \\ +\infty, & \text{else}, \end{cases}$$

where $\|\cdot\|_\infty$ denotes the ℓ_∞-norm.

In this case, g is continuously differentiable on $\mathbb{R}^n$ with a locally Lipschitz continuous gradient but g^* does not satisfy the calmness condition at any y with $\|y\|_\infty = 1$. On the other hand, we let $g(x) := \|x\|_1$ for $x \in \mathbb{R}^n$. Then we get that

$$g^*(y) = \begin{cases} 0, & \|y\|_\infty \leq 1, \\ +\infty, & \text{else}. \end{cases}$$

In this example, g is not differentiable at any x with a zero entry, but g^* satisfies the calmness condition on $\mathrm{dom}(g^*)$.

The sequential convergence of $\{x^k : k \in \mathbb{N}\}$ generated by PGSA_BE under each of these two assumptions will be analyzed in the next two subsections.

4.1. g is continuously differentiable on Ω with a locally Lipschitz continuous gradient

In this subsection, we derive the global sequential convergence result of the case where g is continuously differentiable on Ω with a locally Lipschitz continuous gradient. To this end, we first introduce an auxiliary function $G : \mathbb{R}^n \times \mathbb{R}^n \to \overline{\mathbb{R}}$, defined at $(x, z) \in \mathbb{R}^n \times \mathbb{R}^n$ as

$$G(x,z) := \begin{cases} \frac{f(x)+h(x)+\frac{1-\epsilon/2}{2\alpha}\|x-z\|_2^2}{g(x)}, & x \in \Omega \cap \mathrm{dom}(f), \\ +\infty, & \text{else.} \end{cases} \tag{4.1}$$

The next lemma concerns the lower semicontinuity of G.

Lemma 4.2. *The function G is lower semicontinuous.*

Proof. If $x \in \Omega$, there holds $0 < g(x) = \lim\limits_{y\to x} g(y)$. Then we obtain $G(x,z) \le \liminf\limits_{\substack{y\to x\\ u\to z}} G(y,u)$ since f is lower semicontinuous and h is continuous. If $x \notin \Omega$, we have $0 = g(x) = \lim\limits_{y\to x} g(y)$ and $G(x,z) = +\infty$. Then, invoking Assumption 1 (d), it holds that

$$0 < f(x) + h(x) + \frac{1-\epsilon/2}{2\alpha}\|x-z\|_2^2 \le \liminf_{\substack{y\to x\\ u\to z}} f(y) + h(y) + \frac{1-\epsilon/2}{2\alpha}\|y-u\|_2^2.$$

Thus, $\liminf\limits_{\substack{y\to x\\ u\to z}} G(y,u) = +\infty$ follows from $g \ge 0$. Therefore, we have $G(x,z) = \liminf\limits_{\substack{y\to x\\ u\to z}} G(y,u)$. This completes the proof. □

According to Proposition 2.6, if G satisfies the KL property, then global convergence of the whole sequence $\{x^k : k \in \mathbb{N}\}$ generated by PGSA_BE can be established by showing that Items (i)-(iii) of the proposition hold for G along the sequence $\{(x^{k+1}, x^k) : k \in \mathbb{N}\}$. In particular, we now prove that G and $\{(x^{k+1}, x^k) : k \in \mathbb{N}\}$ satisfy Item (ii) of Proposition 2.6.

Lemma 4.3. *Let $\{x^k : k \in \mathbb{N}\}$ be generated by PGSA_BE and suppose Assumptions 2-3 hold. If g is continuously differentiable on Ω with a locally Lipschitz continuous gradient, then there exist $b > 0$, $K \in \mathbb{N}$ and $\omega^{k+1} \in \partial G(x^{k+1}, x^k)$ such that for any $k \ge K$,*

$$\|\omega^{k+1}\|_2 \le b(\|x^{k+1} - x^k\|_2 + \|x^k - x^{k-1}\|_2).$$

Proof. Let S be the closure set of $\{x^k : k \in \mathbb{N}\}$. By Proposition 4.1 (i) and Proposition 3.3 (iv), S is bounded and $S \subseteq \mathrm{dom}(F)$. In view of our assumptions, ∇g and F are locally Lipschitz continuous on $\mathrm{dom}(F)$. Invoking Exercise 7.5(c) of [11], this together with the compactness of S implies that ∇g and F are globally Lipschitz continuous on S. In addition, by Proposition 4.1 (ii) and the boundedness of S, there exist $d_1, d_2, d_3 > 0$ such that $d_1 \le g(x) \le d_2$ and $\|\nabla g(x)\|_2 \le d_3$ for all $x \in S$.

Invoking Proposition 2.2 and the smoothness of g, we have

$$\widehat{\partial} G(x^{k+1}, x^k) = \widehat{\partial}_x G(x^{k+1}, x^k) \times \widehat{\partial}_z G(x^{k+1}, x^k),$$

where

$$\widehat{\partial}_x G(x^{k+1},x^k)=\frac{\partial f(x^{k+1})+\nabla h(x^{k+1})+\frac{1-\epsilon/2}{\alpha}(x^{k+1}-x^k)}{g(x^{k+1})}-\frac{G(x^{k+1},x^k)\nabla g(x^{k+1})}{g(x^{k+1})},\tag{4.2}$$
$$\widehat{\partial}_z G(x^{k+1},x^k)=\frac{(1-\epsilon/2)(x^k-x^{k+1})}{\alpha g(x^{k+1})}.$$

From the iteration of PGSA_BE, we obtain that

$$\frac{u^{k+1}-x^{k+1}}{\alpha}-\nabla h(u^{k+1})+c_k\nabla g(x^k)\in\partial f(x^{k+1}).$$

Substituting this into (4.2), we see that $\omega_x^{k+1}\in\widehat{\partial}_x G(x^{k+1},x^k)$, where

$$\begin{aligned}\omega_x^{k+1}:=&\frac{u^{k+1}-x^{k+1}+(1-\epsilon/2)(x^{k+1}-x^k)}{\alpha g(x^{k+1})}-\frac{\nabla h(u^{k+1})-\nabla h(x^{k+1})}{g(x^{k+1})}\\&+\frac{c_k\nabla g(x^k)-c_{k+1}\nabla g(x^{k+1})}{g(x^{k+1})}-\frac{(1-\epsilon/2)\nabla g(x^{k+1})\|x^{k+1}-x^k\|_2^2}{2\alpha g^2(x^{k+1})}.\end{aligned}\tag{4.3}$$

Using that $u^{k+1}=x^k+\beta_k(x^k-x^{k-1})$ and

$$\frac{c_k\nabla g(x^k)-c_{k+1}\nabla g(x^{k+1})}{g(x^{k+1})}=\frac{c_k\nabla g(x^k)-c_k\nabla g(x^{k+1})}{g(x^{k+1})}+\frac{c_k\nabla g(x^{k+1})-c_{k+1}\nabla g(x^{k+1})}{g(x^{k+1})},$$

we deduce from (4.3) that

$$\begin{aligned}\|\omega_x^{k+1}\|_2\le&\frac{(1/\alpha+L)\beta_k}{g(x^{k+1})}\|x^k-x^{k-1}\|_2\\&+\left(\frac{\epsilon/2}{\alpha g(x^{k+1})}+\frac{L}{g(x^{k+1})}+\frac{c_kL_g}{g(x^{k+1})}+\frac{L_F\|\nabla g(x^{k+1})\|_2}{g(x^{k+1})}\right)\|x^{k+1}-x^k\|_2\\&+\frac{(1-\epsilon/2)\|\nabla g(x^{k+1})\|_2}{2\alpha g^2(x^{k+1})}\|x^{k+1}-x^k\|_2^2,\end{aligned}\tag{4.4}$$

where L_g and L_F are the Lipschitz constants of ∇g and F on S respectively. Additionally, there exists $K\in\mathbb{N}$ such that $\|x^{k+1}-x^k\|_2^2\le\|x^{k+1}-x^k\|_2$ for $k\ge K$, thanks to Proposition 4.1 (iii). Using this and the facts that $0<\beta_k<1$, $g(x^k)\ge d_1$, $\|\nabla g(x^k)\|_2\le d_3$ and $c_k\le c_1$ for $k\in\mathbb{N}$, we obtain further from (4.4) that

$$\begin{aligned}\|\omega_x^{k+1}\|_2\le&\frac{1/\alpha+L}{d_1}\|x^k-x^{k-1}\|_2\\&+\left(\frac{\epsilon}{2\alpha}+L+c_1L_g+d_3L_F+\frac{(2-\epsilon)d_3}{4\alpha d_1}\right)\frac{\|x^{k+1}-x^k\|_2}{d_1}.\end{aligned}\tag{4.5}$$

On the other hand, a direct computation yields

$$\|\widehat{\partial}_z G(x^{k+1},x^k)\|_2=\frac{(1-\epsilon/2)}{\alpha g(x^{k+1})}\|x^{k+1}-x^k\|_2\le\frac{1-\epsilon/2}{\alpha d_1}\|x^{k+1}-x^k\|_2.\tag{4.6}$$

Combining (4.5) and (4.6), we finally obtain the desired result with

$$b:=\frac{1}{d_1}\max\left\{\frac{1}{\alpha}+L,\ \frac{\epsilon}{2\alpha}+L+c_1L_g+d_3L_F+\frac{(2-\epsilon)d_3}{4\alpha d_1}+\frac{1-\epsilon/2}{\alpha}\right\}.\quad\square$$

We are now ready to prove global convergence of the entire sequence $\{x^k : k \in \mathbb{N}\}$ generated by PGSA_BE.

Theorem 4.4. *Let $\{x^k : k \in \mathbb{N}\}$ be generated by PGSA_BE. Suppose that Assumptions 2-3 hold and G is a KL function. If g is continuously differentiable on Ω with a locally Lipschitz continuous gradient, then $\sum_{k=1}^{+\infty} \|x^k - x^{k-1}\|_2 < +\infty$ and $\{x^k : k \in \mathbb{N}\}$ converges to a critical point of F.*

Proof. In view of the definition of G, we have upon rearranging terms in Proposition 3.3 (i) that for $k \in \mathbb{N}$

$$G(x^{k+1}, x^k) + \frac{\epsilon}{4\alpha}\left(\frac{\|x^{k+1} - x^k\|_2^2}{g(x^{k+1})} + \frac{\|x^k - x^{k-1}\|_2^2}{g(x^k)}\right) \leq G(x^k, x^{k-1}),$$

which together with Proposition 4.1 (ii) leads to

$$G(x^{k+1}, x^k) + \frac{\epsilon}{4\alpha d_2}\left(\|x^{k+1} - x^k\|_2^2 + \|x^k - x^{k-1}\|_2^2\right) \leq G(x^k, x^{k-1}).$$

Using this and invoking Proposition 2.6, Theorem 3.4, Lemmas 4.2 and 4.3, we immediately obtain the desired result. □

We remark that in Theorem 4.4, the function G is required to satisfy the KL property. Since the sum or quotient of two semi-algebraic functions is also a semi-algebraic function and any semi-algebraic function is a KL function [1], this requirement is satisfied when f, h and g are all semi-algebraic functions. In particular, the associated G of problem (1.3) is a semi-algebraic function. Hence, we immediately obtain the global convergence of the sequence $\{x^k : k \in \mathbb{N}\}$ generated by PGSA_BE for problem (1.3).

4.2. g^* satisfies the calmness condition on $\mathrm{dom}(g^*)$

In this subsection, we establish the global sequential convergence of PGSA_BE when g^*, the Fenchel conjugate function of g, satisfies the calmness condition on $\mathrm{dom}(g^*)$. Our convergence analysis is motivated by [19], where the authors established the global sequential convergence of a proximal algorithm with extrapolation for a class of structure difference-of-convex optimization problems without assuming the smoothness of the second convex function involved. We begin with introducing an auxiliary function, which plays a crucial role in our analysis. Given $d > 0$, let $Q : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n \to (-\infty, +\infty]$ at $(x, y, z) \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n$ as

$$Q(x, y, z) := \begin{cases} \frac{f(x)+h(x)+\frac{1-\epsilon/2}{2\alpha}\|x-z\|_2^2}{\langle x,y\rangle - g^*(y)}, & (x, y) \in \mathrm{dom}(f) \times \mathrm{dom}(g^*) \text{ and} \\ & \langle x, y\rangle - g^*(y) \geq d, \\ +\infty, & \text{else.} \end{cases} \tag{4.7}$$

One can easily check by similar analysis in Lemma 4.2 that Q is proper and lower semicontinuous. Also, by the calculus for Fréchet subdifferential and Proposition 2.3, we have the following proposition concerning the Fréchet subdifferential of Q.

Proposition 4.5. *Let $Q : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n \to (-\infty, +\infty]$ be defined by (4.7) and $(x, y, z) \in \mathrm{dom}(Q)$ with $\langle x, y\rangle - g^*(y) > d$. Suppose that g^* satisfies the calmness condition on $\mathrm{dom}(g^*)$. Then it holds that*

$$\widehat{\partial} Q(x, y, z) = \widehat{\partial}_x Q(x, y, z) \times \widehat{\partial}_y Q(x, y, z) \times \widehat{\partial}_z Q(x, y, z),$$

where

$$\widehat{\partial}_x Q(x,y,z) := \frac{(\langle x,y\rangle - g^*(y))(\partial f(x) + \nabla h(x) + \frac{1-\epsilon/2}{\alpha}(x-z))}{(\langle x,y\rangle - g^*(y))^2} - \frac{(f(x)+h(x)+\frac{1-\epsilon/2}{2\alpha}\|x-z\|_2^2)y}{(\langle x,y\rangle - g^*(y))^2},$$

$$\widehat{\partial}_y Q(x,y,z) := \frac{(f(x)+h(x)+\frac{1-\epsilon/2}{2\alpha}\|x-z\|_2^2)(\partial g^*(y) - x)}{(\langle x,y\rangle - g^*(y))^2},$$

$$\widehat{\partial}_z Q(x,y,z) := \frac{\frac{1-\epsilon/2}{\alpha}(z-x)}{\langle x,y\rangle - g^*(y)}.$$

We also need to make extensive use of a sequence $\{\eta_k : k \in \mathbb{N}\}$ defined by $\eta_k := \langle x^k, y^k\rangle - g^*(y^k)$, where $\{x^k : k \in \mathbb{N}\}$ and $\{y^k : k \in \mathbb{N}\}$ are generated by PGSA_BE. With the help of d_1 and d_2, introduced in Proposition 4.1 (ii), we give some useful properties of η_k in the next lemma.

Lemma 4.6. *Let $\{(x^k, y^k) : k \in \mathbb{N}\}$ be generated by PGSA_BE. Suppose that Assumption 2 holds, then the following statements hold:*

(i) $\eta_{k+1} = g(x^k) + \langle x^{k+1} - x^k, y^{k+1}\rangle$ *and* $\eta_k \le g(x^k) \le d_2$ *for* $k \in \mathbb{N}$;
(ii) *there exists* $K_1 \in \mathbb{N}$ *such that* $\eta_k \ge d_1/2$ *for* $k \ge K_1$;
(iii) $\lim_{k\to\infty} \eta_k/\eta_{k+1} = 1$.

Proof. First we prove Item (i). By $y^{k+1} \in \partial g(x^k)$ and *Frechel-Young Inequality*, we immediately see that

$$\begin{aligned}\eta_{k+1} &= \langle x^{k+1}, y^{k+1}\rangle - g^*(y^{k+1}) = \langle x^{k+1} - x^k, y^{k+1}\rangle + \langle x^k, y^{k+1}\rangle - g^*(y^{k+1})\\ &= \langle x^{k+1} - x^k, y^{k+1}\rangle + g(x^k).\end{aligned}$$

Moreover, invoking the definition of g^* and Proposition 4.1 (ii), we have

$$\begin{aligned}\eta_k &= \langle x^k, y^k\rangle - \sup\{\langle x, y^k\rangle - g(x) : x \in \mathbb{R}^n\}\\ &\le \langle x^k, y^k\rangle - \langle x^k, y^k\rangle + g(x^k) = g(x^k) \le d_2.\end{aligned}$$

Next we show Item (ii). Item (i) yields $\eta_{k+1} - g(x^k) = \langle x^{k+1} - x^k, y^{k+1}\rangle$. In view of Proposition 4.1 (i) (iii), we have by passing to the limit that $\lim_{k\to\infty} \eta_{k+1} - g(x^k) = 0$. This together with Proposition 4.1 (ii) indicates Item (ii).

Finally, we prove Item (iii). A direct computation leads to

$$\frac{\eta_k}{\eta_{k+1}} - \frac{g(x^{k-1})}{g(x^k)} = \frac{\langle x^k - x^{k-1}, y^k\rangle g(x^k) - \langle x^{k+1} - x^k, y^{k+1}\rangle g(x^k)}{\eta_{k+1} g(x^k)}.$$

Invoking Item (ii) and Proposition 4.1 (i)-(iii), we have upon passing to the limit in the above relation that $\lim_{k\to\infty} \frac{\eta_k}{\eta_{k+1}} - \frac{g(x^{k-1})}{g(x^k)} = 0$. Combining this with Proposition 4.1 (iv), we immediately obtain $\lim_{k\to\infty} \eta_k/\eta_{k+1} = 1$. □

By Lemma 4.6 (ii), we note that $\{(x^{k+1}, y^{k+1}, x^k) : k \ge K_1\} \subseteq \mathrm{dom}(Q)$ with $0 < d \le d_1/2$, where $\{x^k : k \in \mathbb{N}\}$ and $\{y^k : k \in \mathbb{N}\}$ are generated by PGSA_BE. In the rest of this subsection, we always assume that $0 < d < d_1/2$ in the definition of Q. In view of Proposition 2.7, if Q satisfies the KL property, we can establish global convergence of the entire sequence generated by PGSA_BE by proving Q along

$\{(x^{k+1}, y^{k+1}, x^k) : k \geq K_1\}$ satisfies Items (i)-(iii) in the proposition. We shall show these results in the next two lemmas.

Lemma 4.7. *Let $\{(x^k, y^k) : k \in \mathbb{N}\}$ be generated by PGSA_BE. Suppose that Assumption 2 holds. Then there exist $a > 0$ and $K_2 \in \mathbb{N}$ such that for any $k \geq K_2$,*

$$Q(x^{k+1}, y^{k+1}, x^k) + a\left(\|x^{k+1} - x^k\|_2^2 + \|x^k - x^{k-1}\|_2^2\right) \leq Q(x^k, y^k, x^{k-1}).$$

Proof. By summing (3.4), (3.5), (3.6) and using the fact that $\alpha \leq 1/L$, we obtain that

$$\begin{aligned} f(x^{k+1}) + h(x^{k+1}) + \frac{1}{2\alpha}\|x^{k+1} - x^k\|_2^2 + \langle c_k y^{k+1}, x^k - x^{k+1}\rangle \\ \leq f(x^k) + h(x^k) + \frac{\beta_k^2(1/\alpha + l)}{2}\|x^k - x^{k-1}\|_2^2. \end{aligned} \tag{4.8}$$

Lemma 4.6 (i) yields that

$$\langle c_k y^{k+1}, x^k - x^{k+1}\rangle = c_k(g(x^k) - \eta_{k+1}) = f(x^k) + h(x^k) - c_k\eta_{k+1}.$$

Combining this with (4.8) and Lemma 4.6 (ii), we further obtain

$$\begin{aligned} \frac{f(x^{k+1}) + h(x^{k+1}) + \frac{1-\epsilon/2}{2\alpha}\|x^{k+1} - x^k\|_2^2}{\eta_{k+1}} + \frac{\epsilon}{4\alpha\eta_{k+1}}\|x^{k+1} - x^k\|_2^2 \\ \leq c_k + \frac{\beta_k^2(1/\alpha + l)}{2\eta_{k+1}}\|x^k - x^{k-1}\|_2^2 \end{aligned} \tag{4.9}$$

for $k \geq K_1$.

In addition, from Lemma 4.6 (iii), there exists $\widetilde{K} \in \mathbb{N}$ such that for $k \geq \widetilde{K}$,

$$\frac{\beta_k^2(1 + \alpha l)\eta_k}{\eta_{k+1}} \leq \frac{\bar{\beta}^2(1 + \alpha l)\eta_k}{\eta_{k+1}} \leq 1 - \frac{3}{4}\epsilon. \tag{4.10}$$

Therefore, by the definition of Q and Lemma 4.6 (i), we have for $k \geq K_2 := \max\{K_1, \widetilde{K}\}$ that

$$\begin{aligned} Q(x^k, y^k, x^{k-1}) &= \frac{f(x^k) + h(x^k) + \frac{1-\epsilon/2}{2\alpha}\|x^k - x^{k-1}\|_2^2}{\eta_k} \geq c_k + \frac{1 - \epsilon/2}{2\alpha\eta_k}\|x^k - x^{k-1}\|_2^2 \\ &\geq Q(x^{k+1}, y^{k+1}, x^k) + \frac{\epsilon}{4\alpha\eta_{k+1}}\|x^{k+1} - x^k\|_2^2 + \left(\frac{1 - \epsilon/2}{2\alpha} - \frac{\beta_k^2(1/\alpha + l)\eta_k}{2\eta_{k+1}}\right)\frac{\|x^k - x^{k-1}\|_2^2}{\eta_k} \\ &\geq Q(x^{k+1}, y^{k+1}, x^k) + \frac{\epsilon}{4\alpha\eta_{k+1}}\|x^{k+1} - x^k\|_2^2 + \frac{\epsilon}{8\alpha\eta_k}\|x^k - x^{k-1}\|_2^2 \\ &\geq Q(x^{k+1}, y^{k+1}, x^k) + \frac{\epsilon}{4\alpha d_2}\|x^{k+1} - x^k\|_2^2 + \frac{\epsilon}{8\alpha d_2}\|x^k - x^{k-1}\|_2^2, \end{aligned}$$

where the second and the third inequalities follow from (4.9) and (4.10) respectively. Let $a = \epsilon/(8\alpha d_2)$, then we get the desired result. □

Lemma 4.8. *Let $\{(x^k, y^k) : k \in \mathbb{N}\}$ be generated by PGSA_BE and suppose Assumptions 2-3 hold. If g^* satisfies the calmness condition on $\mathrm{dom}(g^*)$, then there exist $b > 0$, $K_3 \in \mathbb{N}$, and $\omega^{k+1} \in \partial Q(x^{k+1}, y^{k+1}, x^k)$ such that $\|\omega^{k+1}\|_2 \leq b(\|x^{k+1} - x^k\|_2 + \|x^k - x^{k-1}\|_2)$ holds for $k \geq K_3$.*

Proof. With the help of Proposition 4.5 and using the fact that $x^k \in \partial g^*(y^{k+1})$, one can verify that $\omega^{k+1} := (\omega_x^{k+1}, \omega_y^{k+1}, \omega_z^{k+1}) \in \widehat{\partial} Q(x^{k+1}, y^{k+1}, x^k)$, where

$$\begin{aligned}
\omega_x^{k+1} &= \frac{u^{k+1} - x^{k+1} + (1-\epsilon/2)(x^{k+1} - x^k)}{\alpha\eta^{k+1}} - \frac{\nabla h(u^{k+1}) - \nabla h(x^{k+1})}{\eta^{k+1}} \\
&\quad + \frac{(c_k - \frac{f(x^{k+1}) + h(x^{k+1})}{\eta_{k+1}})y^{k+1}}{\eta_{k+1}} - \frac{1-\epsilon/2}{2\alpha\eta_{k+1}^2}\|x^{k+1} - x^k\|_2^2 y^{k+1}, \\
\omega_y^{k+1} &= \frac{(f(x^{k+1}) + h(x^{k+1}) + \frac{1-\epsilon/2}{2\alpha}\|x^{k+1} - x^k\|_2^2)(x^k - x^{k+1})}{\eta_{k+1}^2}, \\
\omega_z^{k+1} &= \frac{(1-\epsilon/2)(x^k - x^{k+1})}{\alpha\eta_{k+1}}.
\end{aligned}$$

Next we shall bound $\|\omega^{k+1}\|_2$ by $b\|x^{k+1} - x^k\|_2$ for some $b > 0$. To this end, we first present some properties on $\{(x^k, y^k) : k \in \mathbb{N}\}$ which will be used in the estimation of $\|\omega^{k+1}\|_2$. Proposition 3.3 (i) indicates that $F(x^k) \le F(x^0)$ for all $k \in \mathbb{N}$, which together with Proposition 4.1 (ii) yields that

$$f(x^k) + h(x^k) \le F(x^0)d_2. \tag{4.11}$$

In addition, by Proposition 4.1 (iii), there exists $\widetilde{K} \in \mathbb{N}$ such that for $k \ge \widetilde{K}$,

$$\|x^{k+1} - x^k\|_2^2 \le \|x^{k+1} - x^k\|_2 \le \frac{d_1}{2}. \tag{4.12}$$

Furthermore, a direct computation yields

$$\begin{aligned}
c_k - \frac{f(x^{k+1}) + h(x^{k+1})}{\eta_{k+1}} &= \frac{f(x^k) + h(x^k) - f(x^{k+1}) - h(x^{k+1})}{g(x^k)} \\
&\quad + \frac{\langle x^k - x^{k+1}, y^{k+1}\rangle (f(x^k) + h(x^k))}{g(x^k)\eta_{k+1}}.
\end{aligned}$$

Also, we deduce that $f + h$ is Lipschitz continuous on the closure of $\{x^k : k \in \mathbb{N}\}$ from the boundedness of $\{x^k : k \in \mathbb{N}\}$ and Assumption 3. Let L_{f+h} denote the Lipschitz constant of $f + h$ on S.

Now we are ready to make an estimation of $\|\omega^{k+1}\|_2$. Let $M > 0$ denote the bound of $\{\|y^k\|_2 : k \in \mathbb{N}\}$. Using the aforementioned facts and invoking $u^{k+1} = x^{k+1} + \beta_k(x^{k+1} - x^k)$ with $\beta_k \le \bar{\beta}$ together with the Lipschitz continuity of ∇h and $\eta_k \ge d_1/2$ for $k \ge K_1$ from Lemma 4.6 (ii), we can verify that for $k \ge K_3 := \max\{K_1, \widetilde{K}\}$

$$\begin{aligned}
\|\omega_x^{k+1}\|_2 &\le \frac{2(1+\alpha L)\bar{\beta}}{\alpha d_1}\|x^k - x^{k-1}\|_2 \\
&\quad + \left(\frac{\epsilon + 2\alpha L + (1-\epsilon/2)M}{\alpha d_1} + \frac{2ML_{f+h}}{d_1^2} + \frac{4d_2M^2F(x^0)}{d_1^3}\right)\|x^{k+1} - x^k\|_2, \\
\|\omega_y^{k+1}\|_2 &\le \left(\frac{2-\epsilon}{4\alpha} + \frac{4d_2F(x^0)}{d_1^2}\right)\|x^{k+1} - x^k\|_2, \\
\|\omega_z^{k+1}\|_2 &\le \frac{2-\epsilon}{\alpha d_1}\|x^{k+1} - x^k\|_2.
\end{aligned}$$

Finally, we conclude that for $k \geq K_3$, $\|\omega^{k+1}\|_2 \leq b(\|x^{k+1}-x^k\|_2 + \|x^k - x^{k-1}\|_2)$ with

$$b := \frac{2-\epsilon}{4\alpha} + \frac{2+2\alpha L + (1-\epsilon/2)M}{\alpha d_1} + \frac{2ML_{f+h} + 4d_2F(x^0)}{d_1^2} + \frac{4d_2M^2F(x^0)}{d_1^3} \quad \square$$

Now we are ready to present the main result of this subsection.

Theorem 4.9. *Let $\{x^k : k \in \mathbb{N}\}$ be generated by PGSA_BE. Suppose that Assumptions 2-3 hold and Q is a KL function. If g^* satisfies the calmness condition on $\mathrm{dom}(g^*)$, then $\sum_{k=1}^{\infty}\|x^k - x^{k-1}\|_2 < +\infty$ and $\{x^k : k \in \mathbb{N}\}$ converges to a critical point of F.*

Proof. In view of Proposition 2.7, Theorem 3.4, Proposition 4.1, Lemma 4.7 and Lemma 4.8, it suffices to prove that $\lim_{k\to\infty} Q(x^{k+1}, y^{k+1}, x^k) = \xi$ exists and $Q(x^\star, y^\star, z^\star) = \xi$ for any accumulation point $(x^\star, y^\star, z^\star)$ of $\{(x^{k+1}, y^{k+1}, x^k) : k \in \mathbb{N}\}$. We see immediately from Lemma 4.7 that $\{Q(x^{k+1}, y^{k+1}, x^k) : k \in \mathbb{N}\}$ is nondecreasing. In addition, this sequence is bounded below by 0 thanks to Lemma 4.6 (ii). Hence, we deduce that $\lim_{k\to\infty} Q(x^{k+1}, y^{k+1}, x^k) = \xi$ exists.

Let $(x^\star, y^\star, z^\star)$ be an accumulation point of $\{(x^{k+1}, y^{k+1}, x^k) : k \in \mathbb{N}\}$. Then there exists a subsequence $\{(x^{k_j+1}, y^{k_j+1}, x^{k_j}) : j \in \mathbb{N}\}$ such that $\lim_{j\to\infty}(x^{k_j+1}, y^{k_j+1}, x^{k_j}) = (x^\star, y^\star, z^\star)$. Since Q is continuous on $\mathrm{dom}(Q)$, we have that $Q(x^\star, y^\star, z^\star) = \lim_{j\to\infty} Q(x^{k_j+1}, y^{k_j+1}, x^{k_j}) = \xi$. Since $(x^\star, y^\star, z^\star)$ is an arbitrary accumulation point, we complete the proof. $\square$

Before moving to the next section, we verify that the merit function Q for problem (1.5) satisfies the KL assumption needed in Theorem 4.9 and thus we can establish global convergence of the entire sequence $\{x^k : k \in \mathbb{N}\}$ generated by PGSA_BE for problem (1.5). Recall that $f = \lambda\|\cdot\|_1$, $g = \|\cdot\|_{(K)}$, and $h = \frac{1}{2}\|A\cdot - b\|_2^2$ in problem (1.5). It is shown in [6, Exercise IV 1.18 and Exercise IV 2.12] that $\|\cdot\|_{(K)}^* = \iota_{B_K}$, where B_K denotes the subset $\{y \in \mathbb{R}^n : \|y\|_\infty \leq 1, \|y\|_1 \leq K\}$. Hence, invoking (4.7), the merit function Q for problem (1.5) has the form of

$$Q(x,y,z) = \begin{cases} \frac{\lambda\|x\|_1 + \frac{1}{2}\|Ax-b\|_2^2 + \frac{1-\epsilon/2}{2\alpha}\|x-z\|_2^2}{\langle x, y\rangle}, & y \in B_K \text{ and } \langle x, y\rangle \geq d, \\ +\infty, & \text{else.} \end{cases}$$

By the above formulation of Q, it is clear that Q is a semi-algebraic function and thus satisfies the KL property.

5. Numerical experiments

In this section, we perform some preliminary numerical experiments to test the efficiency of our proposed PGSA_BE. All experiments are conducted in Matlab R2019b on a desktop with an Intel(R) Core(TM) i5-9500 CPU (3.00 GHz) and 16 GB of RAM.

In our numerical test, we focus on the two ratio regularized sparse recovery problems mentioned in Section 1. Specially, we consider sparse recovery with highly coherent matrices A, for which standard ℓ_1 regularization model usually fails. First, following [25], the matrix A is generated by oversampled discrete cosine transform (DCT), i.e., $A = [a_1, a_2, \cdots, a_n] \in \mathbb{R}^{m\times n}$ with

$$a_j = \frac{1}{\sqrt{m}}\cos\left(\frac{2\pi w j}{D}\right), \ j = 1, 2, \ldots, n.$$

Here $w \in \mathbb{R}^m$ is a random vector following the uniform distribution in $[0,1]^m$ and $D > 0$ is a parameter measuring how coherent the matrix is. Next we construct the ground truth $\widetilde{x} \in \mathbb{R}^n$ with sparsity $K \in \mathbb{N}$. We randomly choose a support subset of size K which has a minimum separation of at least $2D$ and generate a vector $\widetilde{v} \in \mathbb{R}^n$ supported on this set with i.i.d. standard normal entries. We set $\widetilde{x} = \mathrm{sgn}(\widetilde{v})$, where sgn denotes the standard signum function. Finally, we compute $b \in \mathbb{R}^m$ by $b = A\widetilde{x}$ and let $\underline{x} = -2 \times \mathbf{1}_n$ and $\overline{x} = 2 \times \mathbf{1}_n$, where $\mathbf{1}_n$ denotes the n-dimensional vector with all entries being 1.

Next we shall show that $\widetilde{x}$ is a critical point of problem (1.3). First, it can be checked directly that $\widetilde{c} = \lambda\sqrt{K}$ is the objective value at $\widetilde{x}$ of problem (1.3). Since $\widetilde{x} \in \partial\|\cdot\|_1(\widetilde{x})$ and $\nabla(\|\cdot\|_2)(\widetilde{x}) = \widetilde{x}/\sqrt{K}$, it holds that $0 \in \partial(\lambda\|\cdot\|_1)(\widetilde{x}) - \widetilde{c}\,\nabla(\|\cdot\|_2)(\widetilde{x})$. Using this and the facts that $\underline{x} < \widetilde{x} < \overline{x}$ and $A\widetilde{x} = b$, we deduce that

$$0 \in \partial(\lambda\|\cdot\|_1 + \iota_{\{x\in\mathbb{R}^n:\underline{x}\le x\le\overline{x}\}})(\widetilde{x}) + A^T(A\widetilde{x} - b) - \widetilde{c}\,\nabla(\|\cdot\|_2)(\widetilde{x}).$$

Hence, by Definition 3.1, we claim that $\widetilde{x}$ is a critical point of problem (1.3). Following a similar argument, we can also verify that $\widetilde{x}$ is a critical point of problem (1.5).

We shall compare the performance of PGSA_BE, PGSA_NL [32] for problem (1.3) and (1.5) as well as alternating direction method of multipliers (ADMM) [25] and ePSG [8] for problem (1.3). We set the parameter $\lambda \equiv 10^{-3}$ throughout the experiment. The implementation details for these algorithms are discussed below.

- PGSA_BE. First we set $\epsilon = 10^{-4}$ and $\alpha = 1/L$ with $L = \|A\|_2^2$. Inspired by the choice of extrapolation parameters used in FISTA [4], we calculate a recursive sequence $\theta_{k+1} = (1 + \sqrt{1 + 4\theta_k^2})/2$, where $\theta_{-1} = \theta_0 = 1$. We set $\beta_k = (\theta_{k-1} - 1)/\theta_k$ and reset $\theta_{k-1} = \theta_k = 1$ every 100 iterations ($\beta_{100} \approx 0.97$). Hence, we have $\{\beta_k\} \subseteq [0, \bar{\beta}]$ for some $0 < \bar{\beta} < 1$ and the requirements of the parameters in PGSA_BE is satisfied.
- PGSA_NL. We set $L = \|A\|_2^2$. Following the notation in [32, Algorithm 2], we set $\alpha = 10^{-3}$, $\underline{\alpha} = 1.99/L$, $\overline{\alpha} = 10^8$ and $N = 4$.
- ADMM. Following the way to using ADMM for problem (1.2) in [25], we first formulate problem (1.3) into

$$\min\left\{\frac{\lambda\|z\|_1 + \frac{1}{2}\|Ax - b\|_2^2}{\|y\|_2} : \ x = y,\ x = z,\ \underline{x} \le z \le \overline{x}\right\}$$

and introduce its augmented Lagrangian function

$$\begin{aligned}\mathcal{L}_{\mu_1,\mu_2}(x,y,z;v,w) =& \frac{\iota_{\{x\in\mathbb{R}^n:\underline{x}\le z\le\overline{x}\}}(z) + \lambda\|z\|_1 + \frac{1}{2}\|Ax - b\|_2^2}{\|y\|_2} \\ &+ \langle v, x - y\rangle + \frac{\mu_1}{2}\|x - y\|_2^2 + \langle w, x - z\rangle + \frac{\mu_2}{2}\|x - z\|_2^2.\end{aligned}$$

Then the ADMM for solving problem (1.3) consists of the following 5 steps:

$$\begin{cases} x^{k+1} := \arg\min\{\mathcal{L}_{\mu_1,\mu_2}(x, y^k, z^k; v^k, w^k) : x \in \mathbb{R}^n\}, \\ y^{k+1} := \arg\min\{\mathcal{L}_{\mu_1,\mu_2}(x^{k+1}, y, z^k; v^k, w^k) : y \in \mathbb{R}^n\}, \\ z^{k+1} := \arg\min\{\mathcal{L}_{\mu_1,\mu_2}(x^{k+1}, y^{k+1}, z; v^k, w^k) : z \in \mathbb{R}^n\}, \\ v^{k+1} := v^k + \mu_1(x^{k+1} - y^{k+1}), \\ w^{k+1} := w^k + \mu_2(x^{k+1} - z^{k+1}). \end{cases}$$

We set $\mu_1 = \mu_2 = 0.1$ in our experiments.

Table 1
Success rate (%).

		$D=1$	$D=5$	$D=10$	$D=15$	$D=20$
$K=12$	L_1/L_2-ADMM in [25]	100	100	87	60	48
	L_1/L_2-ePSG in [8]	100	100	83	58	41
	L_1/L_2-PGSA_NL in [32]	100	100	84	59	41
	L_1/L_2-PGSA_BE proposed	100	100	85	57	44
	L_1/S_k-PGSA_NL in [32]	100	100	100	100	100
	L_1/S_k-PGSA_BE proposed	100	100	100	100	100
$K=16$	L_1/L_2-ADMM in [25]	100	100	88	65	36
	L_1/L_2-ePSG in [8]	100	100	85	61	29
	L_1/L_2-PGSA_NL in [32]	100	100	85	63	32
	L_1/L_2-PGSA_BE proposed	100	100	86	63	33
	L_1/S_k-PGSA_NL in [32]	100	100	100	100	100
	L_1/S_k-PGSA_BE proposed	100	100	100	100	100
$K=20$	L_1/L_2-ADMM in [25]	95	98	81	60	41
	L_1/L_2-ePSG in [8]	95	98	78	53	37
	L_1/L_2-PGSA_NL in [32]	95	98	78	54	37
	L_1/L_2-PGSA_BE proposed	95	99	78	58	37
	L_1/S_k-PGSA_NL in [32]	100	100	100	100	100
	L_1/S_k-PGSA_BE proposed	100	100	100	100	100

- ePSG. Since $\inf\{\|x\|_2 : \underline{x} \le x \le \overline{x},\ x \neq 0\} = 0$, ePSG for problem (1.3) reduces to a non-extrapolation iterative algorithm (1.9). We set $\alpha_k \equiv 1.99/\|A\|_2^2$ in our experiments.

We remark that PGSA_BE, PGSA_NL and ePSG, involve the proximity operator of $f := \lambda\|\cdot\|_1 + \iota_{\{x\in\mathbb{R}^n:\underline{x}\le x\le\overline{x}\}}$ which can be easily and explicitly computed. Let $z \in \mathbb{R}^n$ and $\alpha > 0$, one can check that for $j = 1, 2, \cdots, n$,

$$(\mathrm{prox}_{\alpha f}(z))_j = \begin{cases} (\underline{x})_j, & \hat{z}_j < (\underline{x})_j, \\ \hat{z}_j, & (\underline{x})_j \le \hat{z}_j \le (\overline{x})_j, \\ (\overline{x})_j, & \hat{z}_j > (\overline{x})_j, \end{cases}$$

where $\hat{z}_j = \max\{0, |z_j| - \alpha\lambda\}\,\mathrm{sgn}(z_j)$.

Through the experiments, we fix $(m, n) = (64, 1024)$ and test on various kinds of sparse recovery problems with $D \in \{1, 5, 10, 15, 20\}$ and sparsity $K \in \{12, 16, 20\}$. In each setting (D, K), we first generate 100 instances randomly as described above and then perform all the computing algorithms. For each instance, we choose randomly the same initial point $x^0 = \widetilde{x} + 0.4\xi$ for all the algorithms, where the entries of $\xi \in \mathbb{R}^n$ are drawn randomly from the uniform distribution on $[-1, 1]$. Moreover, all the algorithms are terminated when

$$\frac{\|x^k - x^{k-1}\|_2}{\max\{1, \|x^k\|_2\}} < 10^{-8}.$$

Finally, the maximum iteration number is set to be $100n = 102400$ for ePSG and $20n = 20480$ for all other algorithms. The accuracy of the algorithms is evaluated in terms of success rate, defined as the number of successful trials over the total number of trials. A success is declared when the relative error of the output $x^\star$ to the ground truth $\widetilde{x}$ is less than 10^{-3}, that is, $\|x^\star - \widetilde{x}\|_2/\|\widetilde{x}\|_2 < 10^{-3}$. Tables 1 and 2 summarize the success rate and averaged CPU time of all the algorithms over 100 instances, respectively. To distinguish between algorithms for L_1/L_2 and L_1/S_K sparse recovery, in the tables we add the prefix "L_1/L_2" (resp., "L_1/S_K") to the algorithms for solving "L_1/L_2" (resp., "L_1/S_K") sparse recovery problems. One can observe from Table 1 that PGSA_NL and PGSA_BE for L_1/S_K sparse recovery achieve 100% success rate in every settings, while the success rates of all the algorithms for L_1/L_2 sparse recovery are comparable, and they decay as K or D increases. Finally, Table 2 shows in terms of CPU time, PGSA_BE slightly outperforms

Table 2
CPU time (in seconds).

		$D=1$	$D=5$	$D=10$	$D=15$	$D=20$
$K=12$	L_1/L_2-ADMM in [25]	0.719	0.738	0.981	1.122	1.230
	L_1/L_2-ePSG in [8]	0.457	0.567	0.968	1.290	1.544
	L_1/L_2-PGSA_NL in [32]	0.114	0.133	0.172	0.214	0.248
	L_1/L_2-PGSA_BE proposed	0.071	0.089	0.146	0.195	0.242
	L_1/S_k-PGSA_NL in [32]	0.125	0.144	0.171	0.197	0.220
	L_1/S_k-PGSA_BE proposed	0.078	0.091	0.106	0.120	0.133
$K=16$	L_1/L_2-ADMM in [25]	0.903	0.838	1.035	1.092	1.304
	L_1/L_2-ePSG in [8]	0.532	0.615	0.958	1.282	1.608
	L_1/L_2-PGSA_NL in [32]	0.121	0.136	0.173	0.215	0.255
	L_1/L_2-PGSA_BE proposed	0.085	0.097	0.152	0.195	0.241
	L_1/S_k-PGSA_NL in [32]	0.136	0.156	0.180	0.209	0.234
	L_1/S_k-PGSA_BE proposed	0.088	0.099	0.110	0.120	0.131
$K=20$	L_1/L_2-ADMM in [25]	1.259	1.112	1.163	1.254	1.354
	L_1/L_2-ePSG in [8]	0.733	0.817	1.038	1.293	1.623
	L_1/L_2-PGSA_NL in [32]	0.147	0.156	0.184	0.215	0.265
	L_1/L_2-PGSA_BE proposed	0.121	0.127	0.162	0.194	0.258
	L_1/S_k-PGSA_NL in [32]	0.143	0.163	0.188	0.213	0.242
	L_1/S_k-PGSA_BE proposed	0.092	0.104	0.111	0.120	0.129

PGSA_NL for the same ratio sparse recovery problem, while it substantially outperforms ADMM and ePSG for L_1/L_2 sparse recovery. This demonstrates the efficiency of PGSA_BE.

6. Conclusion

In this paper, we develop a proximal-gradient-subgradient algorithm with backtracked extrapolation (PGSA_BE) for solving problem (1.1). The proposed PGSA_BE allows a wide range of choices of extrapolation parameters. We prove that any accumulation point of the sequence $\{x^k : k \in \mathbb{N}\}$ generated by PGSA_BE is a critical point of problem (1.1). Moreover, under mild conditions and by assuming some merit functions are KL functions, we establish the global sequential convergence of $\{x^k : k \in \mathbb{N}\}$ in two cases: (i) g is locally Lipschitz differentiable and (ii) the conjugate of g satisfies the calmness condition. Finally, we conduct preliminary numerical experiments on sparse signal recovery problems to illustrate the efficiency of PGSA_BE.

Appendix A. Proof of Proposition 2.3

Proof. For $(u,v) \in \mathrm{dom}(\rho)$ and $(\omega_x, \omega_y) \in \mathbb{R}^n \times \mathbb{R}^n$, a direct computation yields that

$$\frac{\rho(u,v) - \rho(x,y) - \langle \omega_x, u-x\rangle - \langle \omega_y, v-y\rangle}{\|(u,v)-(x,y)\|_2} = T_1(x,y,u,v,\omega_x,\omega_y) + T_2(x,y,u,v),$$

where

$$T_1(x,y,u,v,\omega_x,\omega_y) = \frac{a_2\varphi_1(u) - a_1\langle u,v\rangle + a_1\varphi_2^*(v) - \langle a_2^2\omega_x, u-x\rangle - \langle a_2^2\omega_y, v-y\rangle}{a_2^2\|(u,v)-(x,y)\|_2},$$

$$T_2(x,y,u,v) = \frac{(a_2\varphi_1(u) - a_1\langle u,v\rangle + a_1\varphi_2^*(v))(a_2 - \langle u,v\rangle + \varphi_2^*(v))}{a_2^2(\langle u,v\rangle - \varphi_2^*(v))\|(u,v)-(x,y)\|_2}.$$

Since φ_1 is continuous at x relative to $\mathrm{dom}(\varphi_1)$ and φ_2^* satisfies the calmness condition at y relative to $\mathrm{dom}(\varphi_2^*)$, we get that

$$\lim_{\substack{(u,v)\to(x,y)\\(u,v)\in\mathrm{dom}(\rho)}} T_2(x,y,u,v) = 0. \tag{A.1}$$

By the definition of Fréchet subdifferential, we have

$$\begin{aligned}
&\widehat{\partial}\rho(x,y)\\
&=\left\{(\omega_x,\omega_y)\in\mathbb{R}^n\times\mathbb{R}^n:\ \liminf_{\substack{(u,v)\to(x,y)\\(u,v)\in\mathrm{dom}(\rho)}}\frac{\rho(u,v)-\rho(x,y)-\langle\omega_x,u-x\rangle-\langle\omega_y,v-y\rangle}{\|(u,v)-(x,y)\|_2}\geq 0\right\}\\
&=\left\{(\omega_x,\omega_y)\in\mathbb{R}^n\times\mathbb{R}^n:\ \liminf_{\substack{(u,v)\to(x,y)\\(u,v)\in\mathrm{dom}(\rho)}}T_1(x,y,u,v,\omega_x,\omega_y)\geq 0\right\}\\
&=\left\{(\omega_x,\omega_y)\in\mathbb{R}^n\times\mathbb{R}^n:\ \liminf_{\substack{(u,v)\to(x,y)\\(u,v)\in\mathrm{dom}(\eta)}}\frac{\eta(u,v)-\eta(x,y)-\langle a_2^2\omega_x,u-x\rangle-\langle a_2^2\omega_y,v-y\rangle}{a_2^2\|(u,v)-(x,y)\|_2}\geq 0\right\}\\
&=\frac{\widehat{\partial}\eta(x,y)}{a_2^2},
\end{aligned}$$

where $\eta:\mathbb{R}^n\times\mathbb{R}^n\to\overline{\mathbb{R}}$ defined at $(u,v)\in\mathbb{R}^n\times\mathbb{R}^n$ as $\eta(u,v)=a_2\varphi_1(u)-a_1\langle u,v\rangle+a_1\varphi_2^*(v)$. The second equality follows from (A.1) and the third equality holds due to $a_1>0$ and $a_2>0$. We then obtain the desired result from

$$\widehat{\partial}\eta(x,y)=\widehat{\partial}(a_2\varphi_1(x)+a_1\varphi_2^*(y))-a_1(y,x)=a_2\widehat{\partial}\varphi_1(x)\times a_1\widehat{\partial}\varphi_2^*(y)-a_1(y,x).\quad\square$$

Appendix B. Proof of Proposition 2.7

Proof. From Item (i), we get that $\{H(u^k,v^k):k\in\mathbb{N}\}$ is non-increasing. This together with Item (iii) implies that $H(u^k,v^k)\geq\xi$ for any $k\in\mathbb{N}$. If there exists $k_0>0$ such that $H(u^{k_0},v^{k_0})=\xi$, then $H(u^k,v^k)=\xi$ for all $k\geq k_0$ due to the non-increasing of $\{H(u^k,v^k):k\in\mathbb{N}\}$. Then, Item (i) yields $u^{k+1}=u^k$ for $k\geq k_0$. Thus, we obtain this proposition from Item (ii). Therefore, we only need to consider the case that for $k\in\mathbb{N}$

$$H(u^k,v^k)>\xi. \tag{B.1}$$

We first prove that there exist $K>\max\{K_1,K_2\}$, $\eta>0$ and a continuous concave function $\phi:[0,\eta)\to\mathbb{R}_+$ satisfying Item (i) - (ii) in Definition 2.4 such that there holds, for $k\geq K$,

$$\phi'(H(u^k,v^k)-\xi)\ \mathrm{dist}(0,\partial H(u^k,v^k))\geq 1. \tag{B.2}$$

Denote by Υ the set of accumulation points of $\{(u^k,v^k):k\in\mathbb{N}\}$. Since $\{(u^k,v^k):k\in\mathbb{N}\}$ is bounded, we have Υ is compact. From Item (iii) and (B.1), for any $\delta>0$ and $\eta>0$ there exists $K>0$ such that for $k\geq K$,

$$(u^k,v^k)\in\{(u,v)\in\mathbb{R}^n\times\mathbb{R}^m:\ \mathrm{dist}((u,v),\Upsilon)<\delta,\ \xi<H(u,v)<\xi+\eta\}.$$

Then, by Lemma 2.5, the fact that H satisfies the KL property at each point of Υ and Item (iii), we obtain (B.2) holds for $k\geq K$.

We next show the following key inequality, for $k\geq K$,

$$2\|u^{k+1}-u^k\|_2\leq\|u^k-u^{k-1}\|_2+\frac{b}{a}\Big(\phi(H(u^k,v^k)-\xi)-\phi(H(u^{k+1},v^{k+1})-\xi)\Big). \tag{B.3}$$

From the concavity of ϕ and $\phi' > 0$, Item (i) implies

$$\begin{aligned}\phi(H(u^k, v^k) - \xi) - \phi(H(u^{k+1}, v^{k+1}) - \xi) &\geq \phi'(H(u^k, v^k) - \xi)(H(u^k, v^k) - H(u^{k+1}, v^{k+1})) \\ &\geq a\phi'(H(u^k, v^k) - \xi)\|u^{k+1} - u^k\|_2^2. \end{aligned} \tag{B.4}$$

By Item (ii) and (B.2), it follows for $k \geq K$ that

$$\begin{aligned} b\,\phi'(H(u^k, v^k) - \xi)\|u^k - u^{k-1}\|_2 &\geq \phi'(H(u^k, v^k) - \xi)\|\omega^k\|_2 \\ &\geq \phi'(H(u^k, v^k) - \xi)\ \mathrm{dist}(0, \partial H(u^k, v^k)) \\ &\geq 1. \end{aligned} \tag{B.5}$$

Direct combination of (B.4) and (B.5) yields for $k \geq K$

$$\frac{b}{a}\Big(\phi(H(u^k, v^k) - \xi) - \phi(H(u^{k+1}, v^{k+1}) - \xi)\Big)\|u^k - u^{k-1}\|_2 \geq \|u^{k+1} - u^k\|_2^2.$$

We then obtain (B.3) for $k \geq K$, by utilizing $2\sqrt{\alpha\beta} \leq \alpha + \beta$ for $\alpha, \beta > 0$.

By summing (B.3) from $k = K$ to $k = J > K$, we have

$$\begin{aligned} &\sum_{k=K}^{J} \|u^{k+1} - u^k\|_2 + \|u^{J+1} - u^J\|_2 \\ &\leq \|u^K - u^{K-1}\|_2 + \frac{b}{a}\Big(\phi(H(u^K, v^K) - \xi) - \phi(H(u^{J+1}, v^{J+1}) - \xi)\Big) \\ &\leq \|u^K - u^{K-1}\|_2 + \frac{b}{a}\phi(H(u^K, v^K) - \xi). \end{aligned}$$

Let $J \to +\infty$, we obtain $\sum_{k=1}^{+\infty} \|u^k - u^{k-1}\|_2 < +\infty$ and $\lim_{k\to\infty} u^k = u^\star$. Finally, we prove this proposition by Item (ii) and the closeness of ∂H. □

References

[1] Hedy Attouch, Jerome Bolte, Patrick Redont, Antoine Soubeyran, Proximal alternating minimization and projection methods for nonconvex problems: an approach based on the Kurdyka-Łojasiewicz inequality, Math. Oper. Res. 35 (2010) 438–457.

[2] Hedy Attouch, Jerome Bolte, Benar Fux Svaiter, Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods, Math. Program. 137 (2013) 91–129.

[3] Heinz H. Bauschke, Patrick L. Combettes, et al., Convex Analysis and Monotone Operator Theory in Hilbert Spaces, vol. 408, Springer, 2011.

[4] Amir Beck, First-Order Methods in Optimization, SIAM, 2017.

[5] Amir Beck, Marc Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems, SIAM J. Imaging Sci. 2 (2009) 183–202.

[6] Rajendra Bhatia, Matrix Analysis, Springer, 1997.

[7] Radu Ioan Boţ, Ernö Robert Csetnek, Proximal-gradient algorithms for fractional programming, Optimization 66 (2017) 1383–1396.

[8] Radu Ioan Boţ, Minh N. Dao, Guoyin Li, Extrapolated proximal subgradient algorithms for nonconvex and nonsmooth fractional programs, arXiv preprint, arXiv:2003.04124, 2020.

[9] Emmanuel J. Candes, Justin K. Romberg, Terence Tao, Stable signal recovery from incomplete and inaccurate measurements, Commun. Pure Appl. Math. 59 (2006) 1207–1223.

[10] Rick Chartrand, Exact reconstruction of sparse signals via nonconvex minimization, IEEE Signal Process. Lett. 14 (2007) 707–710.

[11] Francis H. Clarke, Yuri S. Ledyaev, Ronald J. Stern, Peter R. Wolenski, Nonsmooth Analysis and Control Theory, vol. 178, Springer Science & Business Media, 2008.

[12] Line Clemmensen, Trevor Hastie, Daniela Witten, Bjarne Ersbøll, Sparse discriminant analysis, Technometrics 53 (2011) 406–413.
[13] Werner Dinkelbach, On nonlinear fractional programming, Manag. Sci. 13 (1967) 492–498.
[14] David L. Donoho, Compressed sensing, IEEE Trans. Inf. Theory 52 (2006) 1289–1306.
[15] Simon Foucart, Ming Jun Lai, Sparsest solutions of underdetermined linear systems via ℓ_q-minimization for $0 < q \leq 1$, Appl. Comput. Harmon. Anal. 26 (2009) 395–407.
[16] Jun-ya Gotoh, Akiko Takeda, Katsuya Tono, DC formulations and algorithms for sparse optimization problems, Math. Program. 169 (2018) 141–176.
[17] Toshihide Ibaraki, Parametric approaches to fractional programs, Math. Program. 26 (1983) 345–362.
[18] Raj Jagannathan, On some properties of programming problems in parametric form pertaining to fractional programming, Manag. Sci. 12 (1966) 609–615.
[19] Tianxiang Liu, Ting Kei Pong, Akiko Takeda, A refined convergence analysis of pdca$_e$ with applications to simultaneous sparse recovery and outlier detection, Comput. Optim. Appl. 73 (2019) 69–100.
[20] Boris S. Mordukhovich, Variational Analysis and Generalized Differentiation I: Basic Theory, vol. 330, Springer Science & Business Media, 2006.
[21] Boris S. Mordukhovich, Nguyen Mau Nam, N.D. Yen, Fréchet subdifferential calculus and optimality conditions in non-differentiable programming, Optimization 55 (2006) 685–708.
[22] Yu Nesterov, Gradient methods for minimizing composite functions, Math. Program. 140 (2013) 125–161.
[23] Brendan O'donoghue, Emmanuel Candes, Adaptive restart for accelerated gradient schemes, Found. Comput. Math. 15 (2015) 715–732.
[24] Jong-Shi Pang, A parametric linear complementarity technique for optimal portfolio selection with a risk-free asset, Oper. Res. 28 (1980) 927–941.
[25] Yaghoub Rahimi, Chao Wang, Hongbo Dong, Yifei Lou, A scale-invariant approach for sparse signal recovery, SIAM J. Sci. Comput. 41 (2019) A3649–A3672.
[26] R. Tyrrell Rockafellar, Roger J-B. Wets, Variational Analysis, Springer, 2004.
[27] Siegfried Schaible, Fractional programming. II, on Dinkelbach's algorithm, Manag. Sci. 22 (1976) 868–873.
[28] Lixin Shen, Bruce W. Suter, Erin E. Tripp, Structured sparsity promoting functions, J. Optim. Theory Appl. 183 (2019) 386–421.
[29] Christoph Studer, Richard G. Baraniuk, Stable restoration and separation of approximately sparse signals, Appl. Comput. Harmon. Anal. 37 (2014) 12–35.
[30] Bo Wen, Xiaojun Chen, Ting Kei Pong, Linear convergence of proximal gradient algorithm with extrapolation for a class of nonconvex nonsmooth minimization problems, SIAM J. Optim. 27 (2017) 124–145.
[31] Penghang Yin, Yifei Lou, Qi He, Jack Xin, Minimization of ℓ_{1-2} for compressed sensing, SIAM J. Sci. Comput. 37 (2015) A536–A563.
[32] Na Zhang, Qia Li, First-order algorithms for a class of fractional optimization problems, arXiv preprint, arXiv:2005.06207, 2020.