MATHEMATICS OF OPERATIONS RESEARCH
Vol. 00, No. 0, Xxxxx 0000, pp. 000–000
ISSN 0364-765X | EISSN 1526-5471 | 00 | 0000 | 0001

INFORMS
DOI 10.1287/xxxx.0000.0000

Extrapolated Proximal Subgradient Algorithms for Nonconvex and Nonsmooth Fractional Programs

Radu Ioan Boţ
Faculty of Mathematics, University of Vienna, A-1090 Vienna, Austria, radu.bot@univie.ac.at,

Minh N. Dao
School of Engineering, Information Technology and Physical Sciences, Federation University Australia, Ballarat 3353, Australia, m.dao@federation.edu.au,

Guoyin Li
Department of Applied Mathematics, University of New South Wales, Sydney 2052, Australia, g.li@unsw.edu.au,

In this paper, we consider a broad class of nonsmooth and nonconvex fractional programs, which encompass many important modern optimization problems arising from diverse areas such as the recently proposed scale invariant sparse signal reconstruction problem in signal processing. We propose a proximal subgradient algorithm with extrapolations for solving this optimization model and show that the iterated sequence generated by the algorithm is bounded and any of its limit points is a stationary point of the model problem. The choice of our extrapolation parameter is flexible and includes the popular extrapolation parameter adopted in the restarted Fast Iterative Shrinking-Threshold Algorithm (FISTA). By providing a unified analysis framework of descent methods, we establish the convergence of the full sequence under the assumption that a suitable merit function satisfies the Kurdyka–Łojasiewicz property. Our algorithm exhibits *linear convergence* for the scale invariant sparse signal reconstruction problem and the Rayleigh quotient problem over spherical constraint. When the denominator is the maximum of finitely many continuously differentiable weakly convex functions, we also propose another extrapolated proximal subgradient algorithm with guaranteed convergence to a stronger notion of stationary points of the model problem. Finally, we illustrate the proposed methods by both analytical and simulated numerical examples.

Key words: descent method; extrapolation; fractional program; Kurdyka–Łojasiewicz property; linear convergence; proximal subgradient algorithm
MSC2000 subject classification: Primary: 90C26, 90C32; secondary: 49M27, 65K05
OR/MS subject classification: Primary: programming: fractional; secondary: programming: nondifferentiable; programming: nonlinear: algorithms

---

1. Introduction In this paper, we consider the following class of nonsmooth and nonconvex fractional program which takes the form

$$\min_{x\in S} \frac{f(x)}{g(x)}, \tag{P}$$

where $\mathcal{H}$ is a Euclidean space (or a finite-dimensional real Hilbert space), S is a nonempty closed convex subset of $\mathcal{H}$, and $f, g\colon \mathcal{H} \to (-\infty, +\infty]$ are proper lower semicontinuous functions which are not necessarily convex. Unless stated otherwise, the numerator f can be written as the sum of $f^{\mathfrak{s}}$ and $f^{\mathfrak{n}}$, where $f^{\mathfrak{s}}$ is a differentiable convex function whose gradient is Lipschitz continuous and $f^{\mathfrak{n}}$

is a nonconvex function, and the denominator g is a continuous weakly convex function on an open convex set containing S, and always takes positive values on S. We note that weakly convex functions form a broad class of functions which covers convex functions, nonconvex quadratic functions and differentiable functions whose gradient is Lipschitz continuous.

This class of nonsmooth and nonconvex fractional program is a broad optimization model which encompasses many important modern optimization problems arising from diverse areas. This includes, for example, the recently proposed scale invariant sparse signal reconstruction problem in signal processing [28] and the robust Sharpe ratio optimization problems in finance [11]. Moreover, in the special case where the denominator $g(x) \equiv 1$ and $S = \mathcal{H}$, problem (P) reduces to the well-studied nonsmooth composite optimization with the form

$$\min_{x\in\mathcal{H}} f(x) = f^{\mathfrak{s}}(x) + f^{\mathfrak{n}}(x),$$

which covers a lot of modern optimization problems in machine learning (for example, the Lasso problem in computer science). Below we provide a few motivating examples illustrating the model problem (P).

(i) **Scale invariant sparse signal recovery problem:** In signal processing, to reconstruct a sparse signal from its observation, one considers the following scale invariant minimization problem [28]

$$\min_{x\in\mathbb{R}^N} \frac{\|x\|_1}{\|x\|_2} \quad \text{s.t.} \quad Ax \le b,\ Cx = d,$$

where $\|\cdot\|_1$ and $\|\cdot\|_2$ are the ℓ_1-norm and Euclidean norm respectively, $A \in \mathbb{R}^{M\times N}$, $b \in \mathbb{R}^M$, $C \in \mathbb{R}^{P\times N}$, $d \in \mathbb{R}^P$, and the constraint set is bounded and does not contain the origin. Here, the objective function relates to the restricted isometry constant and serves as a surrogate of the cardinality of x. It was shown in [28] that this model can outperform the celebrated Lasso model in recovering a sparse solution. This model problem is indeed a special case of problem (P) with $f(x) = \|x\|_1$ (that is, $f^{\mathfrak{s}} = 0$, $f^{\mathfrak{n}} = f$), $g(x) = \|x\|_2$ and S being the polytope $\{x \in \mathbb{R}^N : Ax \le b, Cx = d\}$.

(ii) **Rayleigh quotient optimization with spherical constraint:** The Rayleigh quotient optimization problem with spherical constraint can be formulated as

$$\min_{x\in\mathbb{R}^N} \frac{x^\top Ax}{x^\top Bx} \quad \text{s.t.} \quad \|x\|_2 = 1,$$

where A and B are symmetric positive definite matrices. This is a special case of problem (P) with $S = \mathbb{R}^N$, $f(x) = x^\top Ax + \iota_C(x)$ (that is, $f^{\mathfrak{s}}(x) = x^\top Ax$, $f^{\mathfrak{n}}(x) = \iota_C(x)$), where C is the unit sphere $\{x \in \mathbb{R}^N : \|x\|_2 = 1\}$ and ι_C is the indicator function of the set C (see (4) later for the definition of indicator function), and $g(x) = x^\top Bx$.

(iii) **Robust Sharpe ratio minimization problem:** The standard Sharpe ratio optimization problem (see, e.g., [11]) can be formulated as

$$\max_{x\in\mathbb{R}^N} \frac{a^\top x - r}{\sqrt{x^\top Ax}} \quad \text{s.t.} \quad e^\top x = 1,\ x \ge 0,$$

where the numerator is the expected return and the denominator measures the risk. In practice, the data associated with the model is often uncertain due to prediction or estimation errors. Following robust optimization approach, we assume that the data (A, a, r) are uncertain and belong to the polyhedral uncertainty set $\mathcal{U} = \mathcal{U}_1 \times \mathcal{U}_2$, where $\mathcal{U}_1 = \mathrm{conv}\{(a_1, r_1), \ldots, (a_{m_1}, r_{m_1})\}$ and $\mathcal{U}_2 = \mathrm{conv}\{A_1, \ldots, A_{m_2}\}$. Here, $(a_i, r_i) \in \mathbb{R}^N \times \mathbb{R}$, $i = 1, \ldots, m_1$, are such that $a_i^\top x - r_i \le 0$ for

all $x \in S = \{x \in \mathbb{R}^N : e^\top x = 1, x \geq 0\}$, and A_j are symmetric positive definite matrix, $j = 1, \ldots, m_2$. The robust Sharpe ratio optimization problem can be written as

$$\max_{x \in \mathbb{R}^N} \frac{\min_{(a,r)\in \mathcal{U}_1}\{a^\top x - r\}}{\max_{A \in \mathcal{U}_2} \sqrt{x^\top A x}} \quad \text{s.t.} \quad e^\top x = 1,\ x \geq 0,$$

which can be written as

$$\min_{x \in \mathbb{R}^N} -\frac{\min_{1\leq i\leq m_1}\{a_i^\top x - r_i\}}{\max_{1\leq i\leq m_2} \sqrt{x^\top A_i x}} \quad \text{s.t.} \quad e^\top x = 1,\ x \geq 0.$$

This is a special case of problem (P) with $f(x) = -\min_{1\leq i\leq m_1}\{a_i^\top x - r_i\} = \max_{1\leq i\leq m_1}\{r_i - a_i^\top x\}$ (that is, $f^{\mathfrak{s}} = 0$, $f^{\mathfrak{n}} = f$), $g(x) = \max_{1\leq i\leq m_2} \sqrt{x^\top A_i x}$ and S as mentioned above.

Before we proceed, we also note that the model problem (P) can be simplified as $\min_{x\in\mathcal{H}} \frac{\widehat{f}(x)}{g(x)}$ with $\widehat{f} = f + \iota_S$ and some appropriate modifications of the assumptions. On the other hand, we stick to the model problem (P), as this simplification complicates the statement of the assumptions needed for ensuring the convergence of the algorithms later on.

The fractional programming problem has a long history, and a classical and popular approach for solving the fractional programming problem is the Dinkelbach's method (see, for example, [12, 13]) which relates it to the following optimization problem

$$\min_{x\in S} f(x) - \overline{\theta} g(x). \tag{1}$$

In particular, (P) has an optimal solution $\overline{x} \in S$ if and only if $\overline{x}$ is an optimal solution to (1) and the optimal objective value of (1) is equal to zero with $\overline{\theta} = \frac{f(\overline{x})}{g(\overline{x})}$. However, one drawback of this procedure is that this can only be done in the very restrictive case when the optimal objective value of (P) is known. To overcome this drawback, in the literature (see [12, 13, 15, 16, 30]) an iterative scheme was proposed which requires solving in each iteration n of the optimization problem

$$\min_{x\in S}\{f(x) - \theta_n g(x)\} \tag{2}$$

while θ_n is updated by $\theta_{n+1} := \frac{f(x_{n+1})}{g(x_{n+1})}$, where x_{n+1} is an optimal solution of (2). However, solving in each iteration an optimization problem of type (2) can be as expensive and difficult as solving the fractional programming problem (P) in general.

Recently, in view of the success of the proximal algorithms in solving composite optimization problems (that is, when the denominator $g(x) \equiv 1$), [7] proposed proximal gradient type algorithms for fractional programming problems, where the numerator f is a proper, convex and lower semicontinuous function and the denominator g is a smooth function, either concave or convex. The approach of [7] is appealing because the proposed iterative methods there perform a gradient step with respect to g and a proximal step with respect to f. In this way, the functions f and g are processed separately in each iteration.

Although the approach in [7] is of interest, still many research questions need to be answered. For example,

- firstly, how to extend the approach in [7] to the case where the numerator and denominator are both nonconvex and nonsmooth? Such an extension would allow us to cover, for example, robust Sharpe ratio optimization problems where both the numerator f and the denominator g are nonsmooth, and the Rayleigh quotient optimization problem with spherical constraints where the numerator f is a nonconvex function.

- secondly, it is known that the performance of the proximal gradient method can be largely improved (see [23]) if one can incorporate extrapolation steps in solving composite optimization problems (that is, when the denominator $g(x) \equiv 1$ in (P)), as for example for the restarted Fast Iterative Shrinking-Threshold Algorithm (FISTA) [5, Chapter 10]. Therefore, it is of great interest to develop proximal algorithms with extrapolations for solving fractional programs.
- thirdly, in the case where f and g are convex, and g is continuously differentiable, it was shown in [7] that the proximal gradient method generates a sequence of iterates which converges to a stationary point of problem (P). Recently, algorithms were proposed for computing a stronger version of stationary points called d(irectional)-stationary points for a class of difference-of-convex optimization problems (for example see [1, 27]). Taking this into account, developing algorithms which converge to sharper notions of stationary points of problem (P) is also highly desirable.

The purpose of this paper is to provide answers to the above questions. Specifically, the contributions of this paper are as follows.

(1) In Section 4, we propose a proximal subgradient algorithm with extrapolations for solving the model problem (P). We then establish that the sequence of iterates generated by the algorithm is bounded and any of its limit points is a stationary point of the model problem (P). Interestingly, the convergence of our algorithm does not require the numerator and denominator to be convex or smooth. Moreover, our extrapolation parameter is broad enough to accommodate the popular extrapolation parameter used for restarted FISTA.
(2) In Section 5, we establish a general framework for analyzing descent methods which is amenable for optimization methods with multi-steps and inexact subproblems. Our conditions are weaker than those in the literature and complement the existing results. With the help of this framework, we establish the convergence of the full sequence under the assumption that a suitable merit function satisfies the KL property. In particular, by identifying the explicit KL exponent, we establish linear convergence of the proposed algorithm for scale invariant sparse signal recovery problem and Rayleigh quotient optimization with spherical constraint.
(3) In the case where the denominator is the maximum of finitely many continuously differentiable weakly convex functions, in Section 6, we also propose an enhanced proximal subgradient algorithm with extrapolations, and show that this enhanced algorithm converges to a stronger notion of stationary points of the model problem.
(4) Finally, we illustrate the proposed methods via analytical and simulated numerical examples in Section 7.

2. Preliminaries Throughout this work, we assume that $\mathcal{H}$ is a Euclidean space (or a finite-dimensional real Hilbert space) with inner product $\langle \cdot, \cdot \rangle$ and the induced norm $\|\cdot\|$. The set of nonnegative integers is denoted by $\mathbb{N}$, the set of real numbers by $\mathbb{R}$, the set of nonnegative real numbers by $\mathbb{R}_+ := \{x \in \mathbb{R} : x \geq 0\}$, and the set of the positive real numbers by $\mathbb{R}_{++} := \{x \in \mathbb{R} : x > 0\}$.

Let $h \colon \mathcal{H} \to [-\infty, +\infty]$ be an extended real-valued function. The *domain* of h is $\operatorname{dom} h := \{x \in \mathcal{H} : h(x) < +\infty\}$. We say that h is *proper* if $\operatorname{dom} h \neq \varnothing$ and it never takes the value $-\infty$. The function h is *lower semicontinuous* if, for all $x \in \operatorname{dom} h$, $h(x) \leq \liminf_{z \to x} h(z)$. We use the symbol $z \xrightarrow{h} x$ to indicate $z \to x$ and $h(z) \to h(x)$. Given $x \in \mathcal{H}$ with $|h(x)| < +\infty$, the *Fréchet subdifferential* of h at x is defined by

$$\widehat{\partial} h(x) := \left\{ u \in \mathcal{H} : \liminf_{z \to x} \frac{h(z) - h(x) - \langle u, z - x \rangle}{\|z - x\|} \geq 0 \right\}$$

and the *limiting subdifferential* of h at x is defined by

$$\partial_L h(x) := \left\{ u \in \mathcal{H} : \exists x_n \xrightarrow{h} x,\ u_n \to u \ \text{ with } \ u_n \in \widehat{\partial} h(x_n) \right\}.$$

We set $\widehat{\partial}h(x) = \partial_L h(x) := \varnothing$ when $|h(x)| = +\infty$ and define $\operatorname{dom}\partial_L h := \{x \in \mathcal{H} : \partial_L h(x) \neq \varnothing\}$. It follows from the definition that the limiting subdifferential has the *robustness property*

$$\partial_L h(x) = \left\{u \in \mathcal{H} : \exists x_n \xrightarrow{h} x,\ u_n \to u \ \text{ with } \ u_n \in \partial_L h(x_n)\right\}. \tag{3}$$

For a convex function h, both Fréchet and limiting subdifferentials reduce to the classical subdifferential in convex analysis (see, for example, [21, Theorem 1.93])

$$\partial h(x) := \{u \in \mathcal{H} : \forall z \in \mathcal{H},\ \langle u, z - x\rangle \leq h(z) - h(x)\}.$$

Moreover, for a strictly differentiable[1] function h, both Fréchet and limiting subdifferentials reduce to the derivative of h denoted by ∇h.

Let S be a nonempty subset of $\mathcal{H}$. Its convex hull is denoted by $\operatorname{conv} S$. The indicator function of S is given by

$$\iota_S(x) := \begin{cases} 0 & \text{if } x \in S, \\ +\infty & \text{if } x \notin S. \end{cases} \tag{4}$$

Given $x \in \mathcal{H}$, the *Fréchet normal cone* of S at x is given by $\widehat{N}_S(x) := \widehat{\partial}\iota_S(x)$ and the *limiting normal cone* of S at x is $N_S(x) := \partial_L \iota_S(x)$. The set S is *regular* at $x \in S$ if $N_S(x) = \widehat{N}_S(x)$. We say that S is regular if it is regular at all of its points. It is known, e.g., from [21, Proposition 1.5] that if S is locally convex around x, i.e., $S \cap U$ is convex for some neighborhood U of x, then S is regular at x.

For a function $h\colon \mathcal{H} \to [-\infty, +\infty]$ finite at x, we say that h is *regular*[2] at x if $\widehat{\partial}h(x) = \partial_L h(x)$. The function h is said to be regular on a set C if it is regular at any $x \in C$. For a proper lower semicontinuous function h, it is clear that if h is convex around x or strictly differentiable at x, then it is regular at x. In the case where h is an indicator function of a closed set or is a Lipschitz continuous function around x, according to [21, Proposition 1.92], h is regular at x if and only if $\operatorname{epi} h := \{(x, r) \in \mathcal{H} \times \mathbb{R} : r \geq h(x)\}$ is regular at $(x, h(x))$.

In general, the limiting subdifferential set can be nonconvex (e.g., for $h(x) = -|x|$ at $0 \in \mathbb{R}$), while $\partial_L h$ enjoys comprehensive calculus rules based on *variational/extremal principles* of variational analysis [21, 29]. In particular, the following sum rule and quotient rule and for limiting subdifferential will be useful for us later.

Lemma 2.1 (Sum and quotient rules). *Let $h_1, h_2\colon \mathcal{H} \to (-\infty, +\infty]$ be proper lower semicontinuous functions, and let $x \in \mathcal{H}$. Then the following hold:*

(i) *Suppose that h_1 is finite at $\overline{x}$ and h_2 is locally Lipschitz around $\overline{x}$. Then $\partial_L(h_1 + h_2)(\overline{x}) \subseteq \partial_L h_1(\overline{x}) + \partial_L h_2(\overline{x})$, where the equality holds if both h_1 and h_2 are regular at $\overline{x}$, in which case $h_1 + h_2$ is also regular at $\overline{x}$. Moreover, if h_2 is strictly differentiable at $\overline{x}$, then $\partial_L(h_1 + h_2)(\overline{x}) = \partial_L h_1(\overline{x}) + \nabla h_2(\overline{x})$.*

(ii) *Suppose that h_1 and h_2 are Lipschitz continuous around $\overline{x}$, and $h_2(\overline{x}) \neq 0$. Then, if $\widehat{\partial}h_2$ is nonempty-valued around $\overline{x}$, one has*

$$\partial_L\left(\frac{h_1}{h_2}\right)(\overline{x}) \subseteq \frac{\partial_L(h_2(\overline{x})h_1)(\overline{x}) - h_1(\overline{x})\partial_L h_2(\overline{x})}{h_2(\overline{x})^2}. \tag{5}$$

If h_2 is strictly differentiable at $\overline{x}$, one has

$$\partial_L\left(\frac{h_1}{h_2}\right)(\overline{x}) = \frac{\partial_L(h_2(\overline{x})h_1)(\overline{x}) - h_1(\overline{x})\nabla h_2(\overline{x})}{h_2(\overline{x})^2}, \tag{6}$$

and consequently h_1/h_2 is regular at $\overline{x}$ if and only if the function $x \mapsto h_2(\overline{x})h_1(x)$ is regular at $\overline{x}$.

[1] A function h is strictly differentiable at x if there exists $u \in \mathcal{H}$ such that $\lim\limits_{y,z \to x} \frac{h(y) - h(z) - \langle u, y - z\rangle}{\|y - z\|} = 0$. Clearly, if h is continuously differentiable at x, then it is strictly differentiable at x.

[2] This is also referred as *lower regular* in [21, 22].

Proof. (i): We first derive from [21, Theorem 3.36] and its following remark that $\partial_L(h_1+h_2)(\overline{x}) \subseteq \partial_L h_1(\overline{x}) + \partial_L h_2(\overline{x})$ and that if both h_1 and h_2 are regular at $\overline{x}$, then so is h_1+h_2 and $\partial_L(h_1+h_2)(\overline{x}) = \partial_L h_1(\overline{x}) + \partial_L h_2(\overline{x})$. By [21, Proposition 1.107(ii)], this equality also holds if h_2 is strictly differentiable at $\overline{x}$.

(ii): As h_1 and h_2 are Lipschitz continuous around $\overline{x}$ and $h_2(\overline{x}) \neq 0$, [21, Proposition 1.111(ii)] implies that

$$\partial_L\left(\frac{h_1}{h_2}\right)(\overline{x}) = \frac{\partial_L(h_2(\overline{x})h_1 - h_1(\overline{x})h_2)(\overline{x})}{h_2(\overline{x})^2}. \tag{7}$$

Thus, to see (5), it suffices to show that $\partial_L(h_2(\overline{x})h_1 - h_1(\overline{x})h_2)(\overline{x}) \subseteq \partial_L(h_2(\overline{x})h_1)(\overline{x}) - h_1(\overline{x})\partial_L h_2(\overline{x})$. This is obvious if $h_1(\overline{x}) = 0$. If $h_1(\overline{x}) < 0$, then $-h_1(\overline{x}) > 0$ and, by (i),

$$\partial_L(h_2(\overline{x})h_1 - h_1(\overline{x})h_2)(\overline{x}) \subseteq \partial_L(h_2(\overline{x})h_1)(\overline{x}) + \partial_L(-h_1(\overline{x})h_2)(\overline{x}) = \partial_L(h_2(\overline{x})h_1)(\overline{x}) - h_1(\overline{x})\partial_L h_2(\overline{x}).$$

If $h_1(\overline{x}) > 0$, then $\widehat{\partial}(h_1(\overline{x})h_2) = h_1(\overline{x})\widehat{\partial}h_2$ is nonempty-valued around $\overline{x}$ and, by [22, Corollary 3.4],

$$\partial_L(h_2(\overline{x})h_1 - h_1(\overline{x})h_2)(\overline{x}) \subseteq \partial_L(h_2(\overline{x})h_1)(\overline{x}) - \partial_L(h_1(\overline{x})h_2)(\overline{x}) = \partial_L(h_2(\overline{x})h_1)(\overline{x}) - h_1(\overline{x})\partial_L h_2(\overline{x}),$$

from which we get the claimed inclusion.

To see (6), we assume in addition that h_2 is strictly differentiable. Then, $-h_1(\overline{x})h_2$ is also strictly differentiable. Applying the last assertion of (i) by replacing h_1 and h_2 with $h_2(\overline{x})h_1$ and $-h_1(\overline{x})h_2$ respectively, it follows from (7) that (6) holds. Finally, the conclusion for the regularity of h_1/h_2 follows from (6) and [17, Corollaries 1.12.2 and 1.14.2]. □

We say that a function h is *weakly convex (on $\mathcal{H}$)* if there exists $\rho \geq 0$ such that $h + \frac{\rho}{2}\|\cdot\|^2$ is a convex function. Moreover, the smallest constant ρ such that $h + \frac{\rho}{2}\|\cdot\|^2$ is convex is called the *modulus* for a weakly convex function h. More generally, a function h is said to be *weakly convex on $S \subseteq \mathcal{H}$ with modulus ρ* if $h + \iota_S$ is weakly convex with modulus ρ. Weakly convex functions form a broad class of functions which covers convex functions and differentiable functions whose gradient is Lipschitz continuous, in particular, any (possibly nonconvex) quadratic functions. Recall that the (one-sided) *directional derivative* of a proper function h at $x \in \operatorname{dom} h$ in the direction d is defined by

$$h'(x;d) = \lim_{t \to 0^+} \frac{h(x+td) - h(x)}{t},$$

provided the limit exists in $[-\infty, +\infty]$; see [4, Definition 17.1]. We end this section with the following lemma.

Lemma 2.2. *Let S be a nonempty closed convex subset of $\mathcal{H}$, let $\overline{x} \in S$, and let $h\colon \mathcal{H} \to (-\infty, +\infty]$ be a proper lower semicontinuous function which is weakly convex on S. Then the following hold:*

(i) *For all $x \in \mathcal{H}$, $\partial_L(h+\iota_S)(x)$ is a (possibly empty) closed convex set.*

(ii) *If $\overline{x} \in \operatorname{int} S$ and h is continuous at $\overline{x}$, then $\partial_L h(\overline{x}) = \partial_L(h+\iota_S)(\overline{x}) \neq \varnothing$ and, for all $x \in S$, $h'(\overline{x}; x-\overline{x}) = \max\{\langle v, x-\overline{x}\rangle : v \in \partial_L(h+\iota_S)(\overline{x})\}$. In particular, if h is a weakly convex function on $\mathcal{H}$ which is continuous at $\overline{x}$, then, for all $d \in \mathcal{H}$, $h'(\overline{x}; d) = \max\{\langle v, d\rangle : v \in \partial_L h(\overline{x})\}$.*

(iii) *$0 \in \partial_L(h+\iota_S)(\overline{x})$ if and only if, for all $x \in S$, $h'(\overline{x}; x-\overline{x}) \geq 0$.*

Proof. By assumption, there exists $\rho \geq 0$ such that $H := h + \iota_S + \frac{\rho}{2}\|\cdot\|^2$ is a convex function. Using Lemma 2.1(i), we have that, for all $x \in \mathcal{H}$, $\partial H(x) = \partial_L H(x) = \partial_L(h+\iota_S)(x) + \rho x$, and so $\partial_L(h+\iota_S)(x) = \partial H(x) - \rho x$. Since S is convex, it follows from the definition of directional derivative that, for all $x \in \mathcal{H}$,

$$h'(\overline{x}; x-\overline{x}) \leq (h+\iota_S)'(\overline{x}; x-\overline{x}) = H'(\overline{x}; x-\overline{x}) - \langle \rho\overline{x}, x-\overline{x}\rangle, \tag{8}$$

where the first inequality is an equality if $x \in S$.

(i): Since $\partial H(x)$ is a closed convex set, so is $\partial_L(h+\iota_S)(x)$.

(ii): Assume that $\overline{x} \in \operatorname{int} S$ and h is continuous at $\overline{x}$, then $h(x) = (h+\iota_S)(x)$ for x near $\overline{x}$ and H is a convex function which is continuous at $\overline{x} \in \operatorname{int} S$ and h is continuous at $\overline{x}$, and hence $\partial_L h(\overline{x}) = \partial_L(h+\iota_S)(\overline{x}) = \partial_L H(\overline{x}) - \rho\overline{x} \neq \varnothing$.

Now, let $x \in S$. By [4, Theorem 17.18], $H'(\overline{x}; x-\overline{x}) = \max\{\langle u, x-\overline{x}\rangle : u \in \partial H(\overline{x})\} = \max\{\langle v+\rho\overline{x}, x-\overline{x}\rangle : v \in \partial(h+\iota_S)(\overline{x})\}$, which combined with (8) implies the desired claim.

(iii): Set $H_1 := h + \iota_S + \frac{\rho}{2}\|\cdot - \overline{x}\|^2$. Then H_1 is also a convex function. We derive from Lemma 2.1(i) that $\partial_L(h+\iota_S)(\overline{x}) = \partial H_1(\overline{x})$ and from (8) that $h'(\overline{x}; x-\overline{x}) \geq 0$ for all $x \in S$ if and only if $(h+\iota_S)'(\overline{x}; x-\overline{x}) = H_1'(\overline{x}; x-\overline{x}) \geq 0$ for all $x \in \mathcal{H}$. The conclusion then follows from [4, Theorem 16.3 and Proposition 17.3]. □

Kurdyka–Łojasiewicz property Next, we recall the celebrated Kurdyka–Łojasiewicz (KL) property [18, 20] which plays an important role in our convergence analysis later on. For each $\eta \in (0, +\infty]$, we denote by Φ_η the class of all continuous concave functions $\varphi\colon [0,\eta) \to \mathbb{R}_+$ such that $\varphi(0) = 0$ and φ is continuously differentiable on $(0,\eta)$ with $\varphi' > 0$.

Let $h\colon \mathcal{H} \to (-\infty, +\infty]$ be a proper lower semicontinuous function. We say that h satisfies the *KL property* [18, 20] at $\overline{x} \in \operatorname{dom} \partial_L h$ if there exist a neighborhood U of $\overline{x}$, $\eta \in (0, +\infty]$, and a function $\varphi \in \Phi_\eta$ such that, for all $x \in U$ with $h(\overline{x}) < h(x) < h(\overline{x}) + \eta$, one has

$$\varphi'(h(x) - h(\overline{x}))\operatorname{dist}(0, \partial_L h(x)) \geq 1.$$

If h satisfies the KL property at each point in $\operatorname{dom} \partial_L h$, then h is called a *KL function*. For a function h satisfying the KL property at $\overline{x} \in \operatorname{dom} \partial_L h$, if the corresponding function φ can be chosen as $\varphi(s) = \gamma s^{1-\alpha}$ for some $\gamma \in \mathbb{R}_{++}$ and $\alpha \in [0,1)$, then we say that h has the *KL property at $\overline{x}$ with an exponent of α*. If h is a KL function and has the same exponent α at any $\overline{x} \in \operatorname{dom} \partial_L h$, then h is called a *KL function with an exponent of α*.

This definition encompasses broad classes of functions that arise in practical optimization problems. For example, it is known that if h is a proper lower semicontinuous semi-algebraic function, then h is a KL function with a suitable exponent of $\alpha \in [0,1)$. The semi-algebraic function covers many common nonsmooth functions that appear in modern optimization problems such as functions which can be written as maximum or minimum of finitely many polynomials, Euclidean norms and the eigenvalues and rank of a matrix. Also, sums, products, and quotients of semi-algebraic functions are still semi-algebraic. For some recent development of KL property, see [2, 8, 19].

Next, we state a uniform version of KL property which will be used later on. The proof is essentially based on [10, Lemma 6]. On the other hand, as our conclusion here slightly differs from [10, Lemma 6], we provide a short proof for completeness.

Lemma 2.3. *Let $(x_n)_{n\in\mathbb{N}}$ be a bounded sequence in $\mathcal{H}$, let Ω be the set of cluster points of $(x_n)_{n\in\mathbb{N}}$, and let $h\colon \mathcal{H} \to (-\infty, +\infty]$ be a proper lower semicontinuous function that is constant on Ω and satisfies the KL property at each point of Ω. Set $\Omega_0 := \{\overline{x} \in \Omega : h(x_n) \to h(\overline{x}) \text{ as } n \to +\infty\}$ and suppose that $\Omega_0 \neq \varnothing$. Then there exist $\eta \in (0, +\infty]$, $\varphi \in \Phi_\eta$, and $n_0 \in \mathbb{N}$ such that, for all $\overline{x} \in \Omega_0$,*

$$\varphi'(h(x_n) - h(\overline{x}))\operatorname{dist}(0, \partial_L h(x_n)) \geq 1 \tag{9}$$

whenever $n \geq n_0$ and $h(x_n) > h(\overline{x})$. Moreover, if h satisfies the KL property at every point of Ω with an exponent of α, then the function φ can be chosen as $\varphi(s) = \gamma s^{1-\alpha}$ for some $\gamma \in \mathbb{R}_{++}$.

Proof. Since $(x_n)_{n\in\mathbb{N}}$ is bounded, Ω is nonempty and compact. According to [10, Lemma 6], there exists $\varepsilon > 0$, $\eta > 0$, and $\varphi \in \Phi_\eta$ such that

$$\varphi'(h(x) - h(\overline{x}))\operatorname{dist}(0, \partial_L h(x)) \geq 1 \tag{10}$$

whenever $\operatorname{dist}(x,\Omega)<\varepsilon$ and $h(\overline{x})<h(x)<h(\overline{x})+\eta$. From the proof of [10, Lemma 6], we also see that, if h satisfies the KL property at every point of Ω with an exponent of α, then the function φ can be chosen as $\varphi(s)=\gamma s^{1-\alpha}$ for some $\gamma\in\mathbb{R}_{++}$.

We note that $\operatorname{dist}(x_n,\Omega)\to 0$ as $n\to+\infty$. Indeed, suppose otherwise. Then there exist $\overline{\varepsilon}>0$ and a subsequence $(x_{k_n})_{n\in\mathbb{N}}$ of $(x_n)_{n\in\mathbb{N}}$ such that, for all $n\in\mathbb{N}$, $\operatorname{dist}(x_{k_n},\Omega)\geq\overline{\varepsilon}$. Since $(x_{k_n})_{n\in\mathbb{N}}$ is also bounded, there exists a subsequence $(x_{l_{k_n}})_{n\in\mathbb{N}}$ such that $x_{l_{k_n}}\to x^*$. We have that $x^*\in\Omega$ and that, for all $n\in\mathbb{N}$, $\operatorname{dist}(x_{l_{k_n}},\Omega)\geq\overline{\varepsilon}$. By the continuity of the distance function (see, e.g., [4, Example 1.48]), $\operatorname{dist}(x^*,\Omega)\geq\overline{\varepsilon}$, which contradicts the fact that $x^*\in\Omega$.

Now, let $\overline{x}\in\Omega_0$. Since $\operatorname{dist}(x_n,\Omega)\to 0$ and $h(x_n)\to h(\overline{x})$ as $n\to+\infty$, one can find $n_0\in\mathbb{N}$ such that, for all $n\geq n_0$,

$$\operatorname{dist}(x_n,\Omega)<\varepsilon \quad\text{and}\quad h(x_n)<h(\overline{x})+\eta.$$

Here, we note that n_0 does not depend on $\overline{x}\in\Omega_0\subseteq\Omega$ because h is constant on Ω. The conclusion follows from (10) and its following remark. □

3. Stationary points of fractional programs In this section, we introduce various versions of stationary points for fractional programs and examine their relationships.

Definition 3.1 (Stationary points, lifted stationary points & strong lifted stationary points). For problem (P), we say that $\overline{x}\in S$ is

(i) a *(limiting) stationary point* if $0\in\partial_L(\frac{f}{g}+\iota_S)(\overline{x})$;
(ii) a *(limiting) lifted stationary point* if $0\in g(\overline{x})\partial_L(f+\iota_S)(\overline{x})-f(\overline{x})\partial_L g(\overline{x})$;
(iii) a *(limiting) strong lifted stationary point* if $f(\overline{x})\partial_L g(\overline{x})\subseteq g(\overline{x})\partial_L(f+\iota_S)(\overline{x})$.

It is well known that a necessary condition for $\overline{x}\in S$ to be a local minimizer of $\frac{f}{g}$ on S is $0\in\partial_L(\frac{f}{g}+\iota_S)(\overline{x})$. Thus, any local minimizer must be a stationary point. Next, we examine the relationships between the above three versions of stationary points.

Lemma 3.2 (Stationary points vs. lifted stationary points). *Consider problem (P) in which $f,g\colon\mathcal{H}\to(-\infty,+\infty]$ are proper lower semicontinuous functions and S is a nonempty closed subset of $\mathcal{H}$. Let C be a nonempty closed subset of $\mathcal{H}$ such that $C\cap S\neq\varnothing$ and let $\overline{x}\in C\cap S$. Suppose that g is Lipschitz continuous around $\overline{x}$ with $g(\overline{x})>0$, and that $f=f_1+\iota_C$, where f_1 is Lipschitz continuous around $\overline{x}$ and one of the following is satisfied:*

(a) $\overline{x}\in\operatorname{int}(C\cap S)$;
(b) f_1 and $C\cap S$ are regular at $\overline{x}$;
(c) f_1 is strictly differentiable at $\overline{x}$.

Then the following statements hold:

(i) *If $\widehat{\partial}g$ is nonempty-valued around $\overline{x}$, then*

$$\partial_L\left(\frac{f}{g}+\iota_S\right)(\overline{x})\subseteq\frac{g(\overline{x})\partial_L(f+\iota_S)(\overline{x})-f(\overline{x})\partial_L g(\overline{x})}{g(\overline{x})^2}, \tag{11}$$

in which case, if $\overline{x}$ is a stationary point of (P), then it is a lifted stationary point of (P).

(ii) *If g is strictly differentiable at $\overline{x}$, then*

$$\partial_L\left(\frac{f}{g}+\iota_S\right)(\overline{x})=\frac{g(\overline{x})\partial_L(f+\iota_S)(\overline{x})-f(\overline{x})\nabla g(\overline{x})}{g(\overline{x})^2}, \tag{12}$$

in which case, $\overline{x}$ is a stationary point of (P) if and only if it is a lifted stationary point of (P).

Proof. (i): By assumption, g is positive around $\overline{x}$ and f_1/g is Lipschitz continuous around $\overline{x}$. Using this fact and applying Lemma 2.1(i) and then Lemma 2.1(ii), we have

$$\partial_L\left(\frac{f}{g}+\iota_S\right)(\overline{x})=\partial_L\left(\frac{f_1}{g}+\iota_{C\cap S}\right)(\overline{x})\subseteq\partial_L\left(\frac{f_1}{g}\right)(\overline{x})+\partial_L\iota_{C\cap S}(\overline{x})\subseteq\frac{g(\overline{x})\partial_L f_1(\overline{x})-f_1(\overline{x})\partial_L g(\overline{x})}{g(\overline{x})^2}+\partial_L\iota_{C\cap S}(\overline{x}).$$

Now, if (a) holds, then $\partial_L f_1(\overline{x}) + \partial_L \iota_{C\cap S}(\overline{x}) = \partial_L(f_1 + \iota_{C\cap S})(\overline{x}) = \partial_L(f+\iota_S)(\overline{x})$. By Lemma 2.1(i), this is also valid if (b) or (c) holds. Noting that $f_1(\overline{x}) = f(\overline{x})$ and that $\partial_L \iota_{C\cap S}(\overline{x}) = \frac{1}{g(\overline{x})}\partial_L \iota_{C\cap S}(\overline{x})$ since $g(\overline{x}) > 0$, we get (11).

(ii): As g is strictly differentiable at $\overline{x}$ with $g(\overline{x}) \neq 0$, we note that f_1/g is regular at $\overline{x}$ if f_1 is regular at $\overline{x}$ (by Lemma 2.1(ii)), and that f_1/g is strictly differentiable at $\overline{x}$ if f_1 is strictly differentiable at $\overline{x}$. Now, (12) is obtained by using the same argument as in (i) and noting that the inclusions become equalities due to the strict differentiability of g (in all of three cases) and either the fact that $\overline{x} \in \operatorname{int}(C\cap S)$ (in the case of (a)), or the regularity of f_1/g and $C\cap S$ (in the case of (b)), or the strict differentiability of f_1/g (in the case of (c)). □

From the definition, any strong lifted stationary point $\overline{x}$ with $\partial_L g(\overline{x}) \neq \varnothing$ is also a lifted stationary point. Moreover, if g is strictly differentiable, then strong lifted stationary points and lifted stationary points are the same. However, if g is not strictly differentiable, then a lifted stationary point need not to be a strong lifted stationary point in general, as in the following example.

Example 3.3. Consider the following one-dimensional fractional program

$$\min_{x\in[-1,1]} \frac{x^2+1}{|x|+1}. \tag{13}$$

Let $\overline{x} = 0$, $f(x) = x^2+1$, $g(x) = |x|+1$ and $S = [-1,1]$. Clearly, $\partial_L(f+\iota_S)(\overline{x}) = \{0\}$ and $\partial_L g(\overline{x}) = [-1,1]$. Then, $\overline{x}$ is a lifted stationary point because $0 \in g(\overline{x})\partial_L(f+\iota_S)(\overline{x}) - f(\overline{x})\partial_L g(\overline{x}) = [-1,1]$. On the other hand, $\overline{x}$ is not a strong lifted stationary point as

$$[-1,1] = f(\overline{x})\partial_L g(\overline{x}) \nsubseteq g(\overline{x})\partial_L(f+\iota_S)(\overline{x}) = \{0\}.$$

Indeed, a direct verification shows that the lifted stationary points of (13) are $-\sqrt{2}+1$, 0, and $\sqrt{2}-1$; while the set of strong lifted stationary points of (13) is $\{-\sqrt{2}+1, \sqrt{2}-1\}$, which coincides with the set of local/global minimizers of problem (13).

Finally, we establish the relationship between the strong lifted stationary points and the recently studied d(irectional)-stationary points in the difference-of-convex (DC) optimization literature [27, 6]. Recall that $\overline{x} \in S$ is a *d-stationary point* of a function h on S if, for all $x \in S$, $h'(\overline{x}; x-\overline{x}) \geq 0$.

Lemma 3.4 (Strong lifted stationary point vs. d-stationary points). *Consider problem (P) in which S is a nonempty closed convex subset of $\mathcal{H}$ and both $f+\iota_S$ and g are proper lower semicontinuous weakly convex functions. Let $\overline{x} \in S$. Suppose that g is continuous on an open set containing S and that $g(\overline{x}) > 0$. Then $\overline{x}$ is a strong lifted stationary point of (P) if and only if it is a d-stationary point of $f - \frac{f(\overline{x})}{g(\overline{x})}g$ on S.*

Proof. Set $\overline{\theta} := f(\overline{x})/g(\overline{x})$. First, $\overline{x}$ is a strong lifted stationary point of (P) if and only if, for all $v \in \partial_L g(\overline{x})$, $0 \in \partial_L(f+\iota_S)(\overline{x}) - \overline{\theta}v = \partial_L(f - \overline{\theta}\langle v, \cdot\rangle + \iota_S)(\overline{x})$, which is equivalent to, for all $v \in \partial_L g(\overline{x})$ and all $x \in S$, $f'(\overline{x}; x-\overline{x}) - \overline{\theta}\langle v, x-\overline{x}\rangle = (f - \langle \overline{\theta}v, \cdot\rangle)'(\overline{x}; x-\overline{x}) \geq 0$ due to Lemma 2.2(iii). Now, as g is weakly convex on $\mathcal{H}$ and continuous at $\overline{x}$, applying the last conclusion of Lemma 2.2(ii) with $h = g$, we have, for all $x \in \mathcal{H}$, $g'(\overline{x}; x-\overline{x}) = \max\{\langle v, x-\overline{x}\rangle : v \in \partial_L g(\overline{x})\}$, which completes the proof. □

4. Extrapolated proximal subgradient (e-PSG) algorithm In this section, we consider problem (P) under the following assumptions.

Assumption A1. $f = f^{\mathfrak{s}} + f^{\mathfrak{n}}$, where $f^{\mathfrak{s}}$ is a differentiable convex function whose gradient $\nabla f^{\mathfrak{s}}$ is Lipschitz continuous with modulus ℓ on $\mathcal{H}$, and $f^{\mathfrak{n}}$ is a proper lower semicontinuous function, $S \cap \operatorname{dom} f \neq \varnothing$ and, for all $x \in S \cap \operatorname{dom} f$, $f(x) \geq 0$.

Assumption A2. g is a continuous weakly convex function with modulus β on an open convex set containing S, and always takes positive values on S.

We note that the nonnegative assumption of the numerator f and the positivity assumption of the denominator g are standard in the literature of fractional programs [7, 12, 13]. Although, at a first glance, these assumptions might be restrictive, they are indeed easily satisfied for many practical optimization models in diverse areas, in particular, for all the motivating examples we mentioned in the introduction. Moreover, consider the case where the constraint set S is compact and the denominator g takes positive values on S. Note that problem (P) is equivalent to $\min_{x\in S}\frac{f(x)+\alpha g(x)}{g(x)}$ for any $\alpha>0$. By replacing f with $f+\alpha g$ for large enough α if necessary, we may assume without loss of generality that, in this case, the numerator of the objective function always takes nonnegative values.

We now propose the following proximal subgradient algorithm with extrapolation for solving the nonsmooth and nonconvex fractional programming problem (P). To do this, we define the following boundedness condition (BC): There exist $m, M\in\mathbb{R}_{++}$ such that, for all $x\in S\cap\operatorname{dom} f$,

$$m\le g(x)\le M. \tag{BC}$$

Algorithm 4.1 (Extrapolated proximal subgradient (e-PSG) algorithm).

▷ **Step 1.** Choose $x_{-1}=x_0\in S\cap\operatorname{dom} f$ and set $n=0$. Let $\delta\in\mathbb{R}_{++}$, let $\zeta\in\mathbb{R}_{++}$ be such that $1-\sqrt{\beta}\zeta>0$, and let

$$\overline{\mu}\in\left[0,\frac{\delta(1-\sqrt{\beta}\zeta)\sqrt{mM}}{2M}\right)\quad\text{and}\quad\overline{\kappa}\in\left[0,\sqrt{\frac{m\delta(1-\sqrt{\beta}\zeta)}{\ell M}-\frac{2m\overline{\mu}}{\ell\sqrt{mM}}}\right),$$

where ℓ is defined in Assumption A1, β is defined in Assumption A2, while m and M are given in (BC). In the absence of (BC), we let $\overline{\mu}=0$ and $\overline{\kappa}=0$.

▷ **Step 2.** Set $\theta_n=\frac{f(x_n)}{g(x_n)}$, let $g_n\in\partial_L g(x_n)$, choose $\tau_n\in\mathbb{R}$ such that $0<\tau_n\le 1/\max\{\sqrt{\beta}\theta_n/\zeta,\delta\}$. Let $u_n=x_n+\kappa_n(x_n-x_{n-1})$ with $\kappa_n\in[0,\overline{\kappa}]$, $v_n=x_n+\mu_n(x_n-x_{n-1})$ with $\mu_n\in[0,\overline{\mu}\tau_n]$, and find

$$x_{n+1}\in\operatorname*{arg\,min}_{x\in S}\left(f^{\mathfrak{n}}(x)+f^{\mathfrak{s}}(u_n)+\langle\nabla f^{\mathfrak{s}}(u_n),x-u_n\rangle+\frac{1}{2\tau_n}\|x-v_n-\tau_n\theta_n g_n\|^2+\frac{\ell}{2}\|x-u_n\|^2\right).$$

▷ **Step 3.** If a termination criterion is not met, let $n=n+1$ and go to Step 2.

Before proceeding, we first make a few observations. Firstly, in the special case where $f^{\mathfrak{s}}\equiv 0$, $f^{\mathfrak{n}}$ is convex, $\kappa_n=0$, $\mu_n=0$, $\ell=0$ and g is continuously differentiable (and so, $g_n=\nabla g(x_n)$), Algorithm 4.1 reduces to the proximal gradient algorithm proposed in [7]. Secondly, in Step 2, the part "$f^{\mathfrak{s}}(u_n)+\langle\nabla f^{\mathfrak{s}}(u_n),x-u_n\rangle$" serves as the linear approximation of $f^{\mathfrak{s}}$ at u_n. Although the term "$f^{\mathfrak{s}}(u_n)$" can be removed as it does not contribute to the minimization problem, we prefer to leave it here for understanding the algorithm intuitively. Finally, it is worth noting that when $\overline{\mu}<\frac{\delta(1-\sqrt{\beta}\zeta)\sqrt{mM}}{2M}$, then $\frac{m\delta(1-\sqrt{\beta}\zeta)}{\ell M}>\frac{2m\overline{\mu}}{\ell\sqrt{mM}}$, and so, the choice of $\overline{\kappa}$ in Step 1 makes sense.

Remark 4.2 (Computing the subproblems). In the above algorithm, the major computational cost lies in solving the subproblem in Step 2. In Step 2, finding x_{n+1} is indeed equivalent to computing the proximal operator[3] of $\frac{\tau_n}{1+\ell\tau_n}(f^{\mathfrak{n}}+\iota_S)$ at the point $\frac{v_n+\tau_n\theta_n g_n+\ell\tau_n u_n-\tau_n\nabla f^{\mathfrak{s}}(u_n)}{1+\ell\tau_n}$, where $f^{\mathfrak{n}}$ is the nonsmooth part of the numerator. This can be done efficiently for functions f and sets S with specific structures. For example,

(i) In the case where S is a polyhedral and $f^{\mathfrak{n}}$ is the maximum of finitely many affine functions, the optimization problem in Step 2 can be reformulated as a convex quadratic optimization problem with linear constraints, and so, can be solved by calling a QP solver. This, in particular, covers the motivating examples (i) and (iii) in the introduction.

[3] The proximal operator of a function h is denoted by Prox_h and is defined as $\operatorname{Prox}_h(x)=\operatorname{argmin}_y\{h(y)+\frac{1}{2}\|y-x\|^2\}$.

(ii) In the case of $S=\mathbb{R}^N$, $f^{\mathfrak{s}}$ is a convex quadratic function, $f^{\mathfrak{n}}=\iota_C$ with C being the unit sphere (as in the motivating example (ii) in the introduction), the optimization problem in Step 2 reduces to computing the projection onto the unit sphere C which has a closed form solution.

(iii) In the case of $f^{\mathfrak{n}}$ is the minimum of finitely many (nonconvex) quadratic function, that is, $f^{\mathfrak{n}}(x)=\min_{1\leq i\leq m}\{x^\top A_i x+a_i^\top x+\alpha_i\}$ and $S=\{x:\|x\|^2\leq\rho\}$, the optimization problem in Step 2 can be computed by solving m many (nonconvex) quadratic optimization problem with a ball constraint. As each quadratic optimization problem with a ball constraint is a trust-region problem, it can be equivalently reformulated as either a semi-definite program (SDP) or an eigenvalue problem. So, the subproblem can be solved by calling an SDP solver or an eigenvalue problem solver.

(iv) In the case of $S=\{x:q_i(x)\leq 0, i=1,\ldots,m_1\}$ where q_i are convex quadratic functions, and $f^{\mathfrak{n}}(x)=\max_{1\leq i\leq m_2}h_i(x)$ where each h_i is a convex quadratic function, the optimization problem in Step 2 can be reformulated as a convex quadratic optimization problem with convex quadratic constraints, and so, can be further rewritten as a semidefinite programming problem (SDP) and solved by calling an SDP solver.

We also note that, when solving the subproblem in Step 2, we require an exact minimizer. On the other hand, one can suitably modify the algorithm to solve the subproblem inexactly and establish convergence to approximate stationary points in suitable sense. For brevity, we will not discuss this in this paper, and leave it as a future work.

Remark 4.3 (Extrapolation parameters). Our choice of the extrapolation parameters covers the popular extrapolation parameter used for restarted FISTA in the case where g is convex and satisfies (BC) (see, for example, [5, Chapter 10] and [23]). To see this, as g is convex, one has $\beta=0$. Choose $\overline{\mu}=0$, $\delta=\frac{\ell M}{m}$, and $\overline{\kappa}\in(0,1)$. Let $\kappa_n=\overline{\kappa}\frac{\nu_{n-1}-1}{\nu_n}$, where

$$\nu_{-1}=\nu_0=1 \quad\text{and}\quad \nu_{n+1}=\frac{1+\sqrt{1+4\nu_n^2}}{2},$$

and reset $\nu_{n-1}=\nu_n=1$ when $n=n_0,2n_0,3n_0,\ldots$ for some integer n_0. In this case, it can be directly verified that $0\leq\kappa_n\leq\overline{\kappa}<1$, and so, the requirement of our extrapolation parameter is satisfied. Also, it is worth noting that our proposed algorithm (Algorithm 4.1) allows one to perform extrapolation even when the smooth part of the numerator $f^{\mathfrak{s}}\equiv 0$ (as in the the motivating examples (i) and (iii) in the introduction).

Remark 4.4 (Choices of the parameters). Firstly, in Algorithm 4.1, δ is any positive real number and ζ is any positive real number such that $1-\sqrt{\beta}\zeta>0$. If the modulus of weak convexity $\beta>0$, then we require $\zeta\in(0,1/\sqrt{\beta})$. This can be easily satisfied by setting, for example, $\zeta=1/(2\sqrt{\beta})$. If $\beta=0$ (that is, g is convex), then ζ can be chosen as any positive number (and actually, in this case, ζ does not involve in Algorithm 4.1). In our numerical experiments later, when choosing the extrapolation parameter in the form of restarted FISTA, we follow the choice in the previous remark and set $\delta=\frac{\ell M}{m}$.

Regarding the parameter τ_n in Algorithm 4.1, we require that $0<\tau_n\leq 1/\max\{\sqrt{\beta}\theta_n/\zeta,\delta\}$. In our numerical experiment, we observe that if τ_n are chosen small, then the step size tends to be very small, and so, the progress of the algorithm can be slow. Therefore, to avoid small step sizes, we choose $\tau_n=1/\max\{\sqrt{\beta}\theta_n/\zeta,\delta\}$.

For the choices of $\overline{\kappa}$ and $\overline{\mu}$, we divide the discussions into two cases. If g does not satisfy (BC), then $\overline{\kappa}=\overline{\mu}=0$. We now consider the case that g satisfies (BC). If the modulus of weak convexity $\beta>0$, by choosing $\zeta\in\big(0,1/(2\sqrt{\beta})\big]$, one can set $\overline{\mu}=\frac{\delta\sqrt{mM}}{8M}$ and choose $\overline{\kappa}\in\Big[0,\sqrt{\frac{m\delta}{4\ell M}}\Big)$. If $\beta=0$ (that is, g is convex), the conditions on $\overline{\kappa}$ and $\overline{\mu}$ in Algorithm 4.1 read as

$$\overline{\mu}\in\left[0,\frac{\delta\sqrt{mM}}{2M}\right) \quad\text{and}\quad \overline{\kappa}\in\left[0,\sqrt{\frac{m\delta}{\ell M}-\frac{2m\overline{\mu}}{\ell\sqrt{mM}}}\right).$$

For the choices κ_n and μ_n, motivated by the restarted FISTA, a plausible choice for extrapolation parameters κ_n and μ_n is to let

$$\kappa_n = \overline{\kappa}\frac{\nu_{n-1}-1}{\nu_n} \quad \text{and} \quad \mu_n = \overline{\mu}\tau_n\frac{\nu_{n-1}-1}{\nu_n},$$

where

$$\nu_{-1} = \nu_0 = 1 \quad \text{and} \quad \nu_{n+1} = \frac{1+\sqrt{1+4\nu_n^2}}{2},$$

and reset $\nu_{n-1} = \nu_n = 1$ when $n = n_0, 2n_0, 3n_0, \ldots$ for some positive integer n_0. This strategy in choosing the extrapolation parameters κ_n and μ_n indeed will be utilized in our numerical experiments later.

Next, we establish the subsequential convergence of Algorithm 4.1. To do this, we will need the following lemmas which will be used later on. The first lemma shows that our Assumption A2 on weak convexity implies an important subgradient inequality. The second lemma is known as the decent lemma for differentiable functions whose gradient is Lipschitz continuous.

Lemma 4.5 (Subgradient inequality for weakly convex functions). *Let S be a nonempty closed convex subset of $\mathcal{H}$. Suppose that either g is regular and weakly convex with modulus β on S, or g is weakly convex with modulus β on an open convex set containing S. Then, for all $x, y \in S$ and $u \in \partial_L g(x)$,*

$$\langle u, y-x\rangle \le g(y) - g(x) + \frac{\beta}{2}\|y-x\|^2.$$

Proof. Let $x, y \in S$. By assumption, $G := g + \iota_C + \frac{\beta}{2}\|\cdot\|^2$ is a convex function for $C = S$ or $C = O$, where O is some open convex set containing S. This implies that $\partial G(x) = \partial_L G(x) = \partial_L(g+\iota_C)(x) + \beta x$, where the second equality is from Lemma 2.1(i). If $C = O$, then $\partial_L(g+\iota_C)(x) = \partial_L g(x)$. In the case where $C = S$, since S is convex and g is regular on S, Lemma 2.1(i) also implies that $\partial_L(g+\iota_C)(x) = \partial_L g(x) + \partial_L \iota_S(x)$. Noting that $0 \in \partial_L \iota_S(x)$, we deduce that, in both cases, $\partial_L g(x) + \beta x \subseteq \partial_L(g+\iota_C)(x) + \beta x = \partial G(x)$.

Now, let any $u \in \partial_L g(x)$. Then $u + \beta x \in \partial G(x)$, and so

$$\begin{aligned}\langle u, y-x\rangle = \langle u+\beta x, y-x\rangle + \langle -\beta x, y-x\rangle &\le G(y) - G(x) - \beta\langle x, y-x\rangle \\ &= \left(g(y) + \iota_C(y) + \frac{\beta}{2}\|y\|^2\right) - \left(g(x) + \iota_C(x) + \frac{\beta}{2}\|x\|^2\right) - \beta\langle x, y-x\rangle \\ &= g(y) - g(x) + \frac{\beta}{2}\|y-x\|^2,\end{aligned}$$

which completes the proof. □

Lemma 4.6 (Descent lemma). *Let $h\colon \mathcal{H} \to \mathbb{R}$ be a differentiable function whose gradient is Lipschitz continuous with modulus ℓ. Then, for all $x, y \in \mathcal{H}$,*

$$h(y) \le h(x) + \langle \nabla h(x), y-x\rangle + \frac{\ell}{2}\|y-x\|^2.$$

Proof. This follows from [23, Lemma 1.2.3], see also [4, Lemma 2.64(i)]. □

We are now ready to state the subsequential convergence of Algorithm 4.1.

Theorem 4.7 (Subsequential convergence). *Let $(x_n)_{n\in\mathbb{N}}$ be the sequence generated by Algorithm 4.1. Suppose that Assumptions A1 and A2 hold, and that the set $S_0 := \{x \in S : \frac{f(x)}{g(x)} \le \frac{f(x_0)}{g(x_0)}\}$ is bounded. Then the following hold:*

(i) *For all* $n \in \mathbb{N}$, $x_n \in S \cap \operatorname{dom} f$ *and*

$$\left(\frac{f(x_{n+1})}{g(x_{n+1})} + c\|x_{n+1} - x_n\|^2\right) + \alpha\|x_{n+1} - x_n\|^2 \leq \frac{f(x_n)}{g(x_n)} + c\|x_n - x_{n-1}\|^2,$$

where

$$\begin{cases} c := \frac{\ell\overline{\kappa}^2}{2m} + \frac{\overline{\mu}}{2\sqrt{mM}},\ \alpha := \frac{\delta(1-\sqrt{\beta}\zeta)}{2M} - \frac{\overline{\mu}}{\sqrt{mM}} - \frac{\ell\overline{\kappa}^2}{2m} & \text{if (BC) holds,} \\ c := 0, \quad \alpha := \frac{\delta(1-\sqrt{\beta}\zeta)}{2M'} \text{ with } M' := \sup_{x\in S_0} g(x) & \text{otherwise.} \end{cases} \tag{14}$$

Consequently, the sequence $\left(\frac{f(x_n)}{g(x_n)}\right)_{n\in\mathbb{N}}$ *is convergent.*

(ii) *The sequence* $(x_n)_{n\in\mathbb{N}}$ *is bounded and* $\sum_{n=0}^{+\infty}\|x_{n+1} - x_n\|^2 < +\infty$. *Consequently,* $\lim_{n\to+\infty}\|x_{n+1} - x_n\| = 0$.

(iii) *If* $\liminf_{n\to+\infty}\tau_n = \overline{\tau} > 0$, *then, for every cluster point* $\overline{x}$ *of* $(x_n)_{n\in\mathbb{N}}$, *it holds that* $\overline{x} \in S \cap \operatorname{dom} f$, $\lim_{n\to+\infty}\frac{f(x_n)}{g(x_n)} = \frac{f(\overline{x})}{g(\overline{x})}$, *and* $\overline{x}$ *is a lifted stationary point of* (P).

Proof. (i)&(ii): First, it is clear that, for all $n \in \mathbb{N}$, $x_n \in S \cap \operatorname{dom} f$, and so

$$g(x_n) > 0 \ \text{ and } \ \theta_n = \frac{f(x_n)}{g(x_n)} \geq 0. \tag{15}$$

We see that, for all $n \in \mathbb{N}$ and $x \in S$,

$$\begin{aligned}
&f(x) + \frac{1}{2\tau_n}\|x - v_n - \tau_n\theta_n g_n\|^2 + \frac{\ell}{2}\|x - u_n\|^2 \\
&= f^{\mathfrak{n}}(x) + f^{\mathfrak{s}}(x) + \frac{1}{2\tau_n}\|x - v_n - \tau_n\theta_n g_n\|^2 + \frac{\ell}{2}\|x - u_n\|^2 \\
&\geq f^{\mathfrak{n}}(x) + f^{\mathfrak{s}}(u_n) + \langle \nabla f^{\mathfrak{s}}(u_n), x - u_n\rangle + \frac{1}{2\tau_n}\|x - v_n - \tau_n\theta_n g_n\|^2 + \frac{\ell}{2}\|x - u_n\|^2 \\
&\geq f^{\mathfrak{n}}(x_{n+1}) + f^{\mathfrak{s}}(u_n) + \langle \nabla f^{\mathfrak{s}}(u_n), x_{n+1} - u_n\rangle + \frac{1}{2\tau_n}\|x_{n+1} - v_n - \tau_n\theta_n g_n\|^2 + \frac{\ell}{2}\|x_{n+1} - u_n\|^2 \\
&\geq f^{\mathfrak{n}}(x_{n+1}) + f^{\mathfrak{s}}(x_{n+1}) - \frac{\ell}{2}\|x_{n+1} - u_n\|^2 + \frac{1}{2\tau_n}\|x_{n+1} - v_n - \tau_n\theta_n g_n\|^2 + \frac{\ell}{2}\|x_{n+1} - u_n\|^2 \\
&= f(x_{n+1}) + \frac{1}{2\tau_n}\|x_{n+1} - v_n - \tau_n\theta_n g_n\|^2,
\end{aligned}$$

where the first inequality follows from the convexity of $f^{\mathfrak{s}}$, the second inequality is from the definition of x_{n+1} in Step 2 of the algorithm, and the last inequality follows from the fact that $f^{\mathfrak{s}}$ is a differentiable function whose gradient is Lipschitz continuous with modulus ℓ (Lemma 4.6 with $h = f^{\mathfrak{s}}$, $y = x_{n+1}$ and $x = u_n$). Therefore, for all $n \in \mathbb{N}$ and $x \in S$,

$$f(x) \geq f(x_{n+1}) + \frac{1}{2\tau_n}(\|x_{n+1} - v_n\|^2 - \|x - v_n\|^2) - \theta_n\langle g_n, x_{n+1} - x\rangle - \frac{\ell}{2}\|x - u_n\|^2. \tag{16}$$

Letting $x = x_n$ and noting that $x_{n+1} - v_n = (x_{n+1} - x_n) - \mu_n(x_n - x_{n-1})$, $x_n - v_n = -\mu_n(x_n - x_{n-1})$, and $x_n - u_n = -\kappa_n(x_n - x_{n-1})$, we have

$$f(x_n) \geq f(x_{n+1}) + \frac{1}{2\tau_n}\left(\|x_{n+1} - x_n\|^2 - 2\mu_n\langle x_{n+1} - x_n, x_n - x_{n-1}\rangle\right) - \theta_n\langle g_n, x_{n+1} - x_n\rangle - \frac{\ell\kappa_n^2}{2}\|x_n - x_{n-1}\|^2.$$

Next, let $\omega \in \mathbb{R}_{++}$. By Young's inequality,

$$\langle x_{n+1} - x_n, x_n - x_{n-1}\rangle \leq \frac{1}{2\omega}\|x_{n+1} - x_n\|^2 + \frac{\omega}{2}\|x_n - x_{n-1}\|^2.$$

Since $x_n, x_{n+1} \in S$ and $g_n \in \partial_L g(x_n)$, Lemma 4.5 implies that

$$\langle g_n, x_{n+1} - x_n \rangle \leq g(x_{n+1}) - g(x_n) + \frac{\beta}{2}\|x_{n+1} - x_n\|^2.$$

Combining the three above inequalities yields

$$f(x_n) \geq f(x_{n+1}) + \frac{1}{2}\left(\frac{1}{\tau_n} - \beta\theta_n - \frac{\mu_n}{\omega\tau_n}\right)\|x_{n+1} - x_n\|^2 + \theta_n(g(x_n) - g(x_{n+1})) - \frac{1}{2}\left(\ell\kappa_n^2 + \frac{\omega\mu_n}{\tau_n}\right)\|x_n - x_{n-1}\|^2.$$

Since $1/\tau_n \geq \max\{\sqrt{\beta}\theta_n/\zeta, \delta\} \geq \sqrt{\beta}\theta_n/\zeta$ (and so, $\frac{1}{\tau_n} - \beta\theta_n \geq \frac{1-\sqrt{\beta}\zeta}{\tau_n}$) and $\theta_n = f(x_n)/g(x_n)$, dividing by $g(x_{n+1}) > 0$ on both sides, it follows that

$$\frac{f(x_n)}{g(x_n)} + \frac{1}{2g(x_{n+1})}\left(\ell\kappa_n^2 + \frac{\omega\mu_n}{\tau_n}\right)\|x_n - x_{n-1}\|^2 \geq \frac{f(x_{n+1})}{g(x_{n+1})} + \frac{1}{2g(x_{n+1})}\left(\frac{1-\sqrt{\beta}\zeta}{\tau_n} - \frac{\mu_n}{\omega\tau_n}\right)\|x_{n+1} - x_n\|^2. \tag{17}$$

We now distinguish two following cases.

Case 1: (BC) holds. Combining with $\kappa_n \leq \overline{\kappa}$, $\mu_n \leq \overline{\mu}\tau_n$, $1/\tau_n \geq \delta$, and choosing $\omega := \sqrt{m/M}$, we derive from (17) that

$$\frac{f(x_n)}{g(x_n)} + \frac{\ell\overline{\kappa}^2 + \omega\overline{\mu}}{2m}\|x_n - x_{n-1}\|^2 \geq \frac{f(x_{n+1})}{g(x_{n+1})} + \left(\frac{\delta(1-\sqrt{\beta}\zeta)}{2M} - \frac{\overline{\mu}}{2M\omega}\right)\|x_{n+1} - x_n\|^2,$$

which means

$$\frac{f(x_n)}{g(x_n)} + \left(\frac{\ell\overline{\kappa}^2}{2m} + \frac{\overline{\mu}}{2\sqrt{mM}}\right)\|x_n - x_{n-1}\|^2 \geq \frac{f(x_{n+1})}{g(x_{n+1})} + \left(\frac{\delta(1-\sqrt{\beta}\zeta)}{2M} - \frac{\overline{\mu}}{2\sqrt{mM}}\right)\|x_{n+1} - x_n\|^2.$$

Setting $F_n := \frac{f(x_n)}{g(x_n)} + c\|x_n - x_{n-1}\|^2$, we deduce that

$$F_{n+1} + \alpha\|x_{n+1} - x_n\|^2 \leq F_n. \tag{18}$$

From the choice of $\overline{\kappa}$, we have $\frac{\ell\overline{\kappa}^2}{2m} < \frac{\delta(1-\sqrt{\beta}\zeta)}{2M} - \frac{\overline{\mu}}{\sqrt{mM}}$. Thus, $\alpha > 0$ and the sequence $(F_n)_{n\in\mathbb{N}}$ is nonincreasing. As F_n is nonnegative, $(F_n)_{n\in\mathbb{N}}$ is a convergent sequence, say $F_n \to \overline{F}$. Furthermore, one also has from (18) that, for any positive integer m,

$$\sum_{n=0}^{m} \alpha\|x_{n+1} - x_n\|^2 \leq \sum_{n=0}^{m}(F_n - F_{n+1}) = F_0 - F_{m+1} \leq F_0.$$

It follows that

$$\sum_{n=0}^{+\infty} \|x_{n+1} - x_n\|^2 < +\infty.$$

In particular, $x_{n+1} - x_n \to 0$ as $n \to +\infty$, and so

$$\frac{f(x_n)}{g(x_n)} = F_n - c\|x_n - x_{n-1}\|^2 \to \overline{F} \quad \text{as } n \to +\infty.$$

Next, to see the boundedness of $(x_n)_{n\in\mathbb{N}}$, observe that

$$\frac{f(x_n)}{g(x_n)} \leq F_n \leq F_0 = \frac{f(x_0)}{g(x_0)} + c\|x_0 - x_{-1}\|^2 = \frac{f(x_0)}{g(x_0)}.$$

So, $x_n \in S_0 = \{x \in S : \frac{f(x)}{g(x)} \leq \frac{f(x_0)}{g(x_0)}\}$, and hence $(x_n)_{n\in\mathbb{N}}$ is bounded by the assumption that S_0 is bounded.

Case 2: (BC) does not hold. Then, by the construction of the algorithm, $\overline{\mu} = \overline{\kappa} = 0$, so $\mu_n = \kappa_n = 0$ for all $n \in \mathbb{N}$ and (17) becomes

$$\frac{f(x_n)}{g(x_n)} \geq \frac{f(x_{n+1})}{g(x_{n+1})} + \frac{1-\sqrt{\beta}\zeta}{2\tau_n g(x_{n+1})}\|x_{n+1}-x_n\|^2,$$

which implies that $(\theta_n)_{n\in\mathbb{N}} = (\frac{f(x_n)}{g(x_n)})_{n\in\mathbb{N}}$ is nonincreasing. As $(\theta_n)_{n\in\mathbb{N}}$ is bounded below, it is convergent. Therefore, for all $n \in \mathbb{N}$, $x_n \in S_0 = \{x \in S : \frac{f(x)}{g(x)} \leq \frac{f(x_0)}{g(x_0)}\}$, and the sequence $(x_n)_{n\in\mathbb{N}}$ is thus bounded. Combining with the continuity of g on S and the boundedness of S_0, one has $\sup_{n\in\mathbb{N}} g(x_n) \leq M' = \sup_{x\in S_0} g(x) < +\infty$. Since $1/\tau_n \geq \delta$, it follows that

$$\frac{f(x_{n+1})}{g(x_{n+1})} + \frac{\delta(1-\sqrt{\beta}\zeta)}{2M'}\|x_{n+1}-x_n\|^2 \leq \frac{f(x_n)}{g(x_n)}. \tag{19}$$

Since $\left(\frac{f(x_n)}{g(x_n)}\right)_{n\in\mathbb{N}}$ is convergent, telescoping (19) yields

$$\sum_{n=0}^{+\infty}\|x_{n+1}-x_n\|^2 < +\infty.$$

(iii): Let $\overline{x}$ be any cluster point of $(x_n)_{n\in\mathbb{N}}$ and let $(x_{k_n})_{n\in\mathbb{N}}$ be a subsequence of $(x_n)_{n\in\mathbb{N}}$ such that $x_{k_n} \to \overline{x}$. Then $\overline{x} \in S$ and, by (ii), $x_{k_n-1} \to \overline{x}$ and also $u_{k_n-1} \to \overline{x}$ and $v_{k_n-1} \to \overline{x}$. We have from (16) that, for all $n \in \mathbb{N}$ and $x \in S$,

$$f(x) \geq f(x_{k_n}) - \frac{1}{2\tau_{k_n-1}}\|x-v_{k_n-1}\|^2 - \theta_{k_n-1}\langle g_{k_n-1}, x_{k_n}-x\rangle - \frac{\ell}{2}\|x-u_{k_n-1}\|^2. \tag{20}$$

Since g is locally Lipschitz continuous on an open set containing S (due to Assumption A2 and [29, Example 9.14]), we have from $x_{k_n} \to \overline{x} \in S$ that $g(x_{k_n}) \to g(\overline{x}) > 0$. Moreover, as $x_{k_n-1} \to \overline{x}$ and [21, Corollary 1.81], we have that $(g_{k_n-1})_{n\in\mathbb{N}}$ is bounded. By (3) and passing to a subsequence if necessary, we may assume that $g_{k_n-1} \to \overline{g} \in \partial_L g(\overline{x})$. Letting $x = \overline{x}$ and $n \to +\infty$ in (20) and noting that $\liminf_{n\to+\infty} \tau_n = \overline{\tau} > 0$, we get $\limsup_{n\to+\infty} f(x_{k_n}) \leq f(\overline{x})$. This together with the lower semicontinuity of f implies that $\lim_{n\to+\infty} f(x_{k_n}) = f(\overline{x})$. It then follows that

$$\lim_{n\to+\infty}\frac{f(x_n)}{g(x_n)} = \lim_{n\to+\infty}\frac{f(x_{k_n})}{g(x_{k_n})} = \frac{f(\overline{x})}{g(\overline{x})}.$$

Now, letting $n \to +\infty$ in (20), one has, for all $x \in S$,

$$f(x) - f(\overline{x}) \geq -\frac{1}{2\overline{\tau}}\|x-\overline{x}\|^2 - \frac{f(\overline{x})}{g(\overline{x})}\langle \overline{g}, \overline{x}-x\rangle - \frac{\ell}{2}\|x-\overline{x}\|^2,$$

or equivalently, for all $x \in S$,

$$\varphi(x) \geq \varphi(\overline{x}), \quad \text{where } \varphi(x) := f(x) + \left(\frac{1}{2\overline{\tau}} + \frac{\ell}{2}\right)\|x-\overline{x}\|^2 - \frac{f(\overline{x})}{g(\overline{x})}\langle \overline{g}, x\rangle.$$

We must have $0 \in \partial_L(\varphi + \iota_S)(\overline{x})$, and so $\frac{f(\overline{x})}{g(\overline{x})}\overline{g} \in \partial_L(f+\iota_S)(\overline{x})$. In particular, $\overline{x} \in S \cap \operatorname{dom} f$. Since $\overline{g} \in \partial_L g(\overline{x})$, we obtain that

$$0 \in g(\overline{x})\partial_L(f+\iota_S)(\overline{x}) - f(\overline{x})\partial_L g(\overline{x}),$$

and the proof is complete. □

Remark 4.8 (Discussions on Theorem 4.7). Firstly, in Theorem 4.7(iii), we require that $\liminf_{n\to+\infty}\tau_n = \overline{\tau} > 0$. This can be ensured easily. Indeed, we note that, as shown in the proof of Theorem 4.7(i)&(ii), for all $n \in \mathbb{N}$,

$$\theta_n \leq \frac{f(x_n)}{g(x_n)} + c\|x_n - x_{n-1}\| \leq \frac{f(x_0)}{g(x_0)} + c\|x_0 - x_{-1}\| = \theta_0,$$

and so $1/\max\{\sqrt{\beta}\theta_n/\zeta, \delta\} \geq \overline{\varepsilon} := 1/\max\{\sqrt{\beta}\theta_0/\zeta, \delta\} > 0$. Now, if we fix an $\varepsilon \in (0, \overline{\varepsilon})$ and, for each $n \in \mathbb{N}$, choose τ_n such that $\varepsilon \leq \tau_n \leq 1/\max\{\sqrt{\beta}\theta_n/\zeta, \delta\}$, then $\liminf_{n\to+\infty}\tau_n \geq \varepsilon > 0$. Thus, the required condition holds.

Secondly, as we have seen in the construction of Algorithm 4.1, the choices of parameters κ_n and μ_n depend on whether the condition (BC) holds or not. These choices play an important role when deriving the subsequential convergence of the algorithm. A natural question is to see whether condition (BC) can be weaken to a form so that the choices of the parameters can be unified. This would be an interesting future research topic to examine.

Next, we consider the following assumption.

Assumption A3. $f^{\mathfrak{n}} = f^{\mathfrak{l}} + \iota_C$, where C is a nonempty closed subset of $\mathcal{H}$ such that $C \cap S \neq \varnothing$ and one of the following is satisfied:

(a) $f^{\mathfrak{l}}$ is locally Lipschitz continuous on $\mathcal{H}$ and $C = S = \mathcal{H}$;

(b) $f^{\mathfrak{l}}$ is locally Lipschitz continuous on an open set containing $S \cap \operatorname{dom} f$ and both $f^{\mathfrak{l}}$ and $C \cap S$ are regular at any $x \in S \cap \operatorname{dom} f$;

(c) $f^{\mathfrak{l}}$ is strictly differentiable on an open set containing $S \cap \operatorname{dom} f$.

All of our motivating examples in the introduction satisfy this assumption. Indeed, we note that convex sets and the unit sphere $C = \{x \in \mathbb{R}^N : \|x\| = 1\}$ are regular, a continuous convex function on $\mathbb{R}^N$ is regular at any $x \in \mathbb{R}^N$, and ι_C is regular at any $x \in C$. It follows that examples (i) and (iii) both satisfy Assumption A3(a)&(b), while example (ii) satisfies Assumption A3(b)&(c).

Corollary 4.9. *Under the hypotheses of Theorem 4.7, suppose further that* $\liminf_{n\to+\infty}\tau_n = \overline{\tau} > 0$, *Assumption A3 holds, and g is strictly differentiable on an open set containing $S \cap \operatorname{dom} f$. Then every cluster point $\overline{x}$ of $(x_n)_{n\in\mathbb{N}}$ is a stationary point of* (P).

Proof. This follows from Theorem 4.7(iii) and Lemma 3.2(ii). □

5. A unified analysis framework and global convergence of e-PSG In this section, we will prove that the global convergence of the whole sequence of $(x_n)_{n\in\mathbb{N}}$ generated by Algorithm 4.1, under the assumption that a suitable merit function satisfies the KL property. To do this, we first establish a general framework for analyzing descent methods which is amenable for optimization method with multi-steps and inexact subproblems. As we will see later on, the proximal subgradient method with extrapolation which we proposed fits to this framework, and so, our desired global convergence result follows consequently.

Firstly, we fix some notation which will be used later on. Let $\mathcal{H}, \mathcal{K}$ be two finite-dimensional real Hilbert spaces. Let $h\colon \mathcal{K} \to (-\infty, +\infty]$ be a proper lower semicontinuous function, let $(x_n)_{n\in\mathbb{N}}$ and $(z_n)_{n\in\mathbb{N}}$ be respectively sequences in $\mathcal{H}$ and $\mathcal{K}$, $(\alpha_n)_{n\in\mathbb{N}}$ and $(\beta_n)_{n\in\mathbb{N}}$ sequences in $\mathbb{R}_{++}$, $(\Delta_n)_{n\in\mathbb{N}}$ and $(\varepsilon_n)_{n\in\mathbb{N}}$ sequences in $\mathbb{R}_+$, and let $\underline{\imath} \leq \overline{\imath}$ be two (not necessarily positive) integers and $\lambda_i \in \mathbb{R}_+$, $i \in I := \{\underline{\imath}, \underline{\imath}+1, \ldots, \overline{\imath}\}$, with $\sum_{i\in I}\lambda_i = 1$. We set $\Delta_k = 0$ for $k < 0$ and consider the following conditions:

H1 (*Sufficient decrease condition*). For each $n \in \mathbb{N}$,

$$h(z_{n+1}) + \alpha_n\Delta_n^2 \leq h(z_n);$$

H2 (*Relative error condition*). For each $n \in \mathbb{N}$,

$$\beta_n \operatorname{dist}(0, \partial_L h(z_n)) \leq \sum_{i\in I}\lambda_i\Delta_{n-i} + \varepsilon_n;$$

H3 (*Continuity condition*). There exist a subsequence $(z_{k_n})_{n\in\mathbb{N}}$ and $\widetilde{z}$ such that

$$z_{k_n} \to \widetilde{z} \quad \text{and} \quad h(z_{k_n}) \to h(\widetilde{z}) \quad \text{as } n \to +\infty;$$

H4 (*Parameter condition*). It holds that

$$\underline{\alpha} := \inf_{n\in\mathbb{N}} \alpha_n > 0, \quad \underline{\gamma} := \inf_{n\in\mathbb{N}} \alpha_n\beta_n > 0, \quad \text{and} \quad \sum_{n=1}^{+\infty} \varepsilon_n < +\infty;$$

H5 (*Distance condition*). There exist $j \in \mathbb{Z}$ and $c \in \mathbb{R}$ such that, for all $n \in \mathbb{N}$,

$$\|x_{n+1} - x_n\| \le c\Delta_{n+j}.$$

Next, we present a lemma which serves as a preparation for our abstract convergence result later on.

Lemma 5.1. *Suppose that (H1) and (H3) hold. Let Ω be the set of cluster points of $(z_n)_{n\in\mathbb{N}}$ and set $\Omega_0 := \{\overline{z} \in \Omega : h(z_n) \to h(\overline{z}) \text{ as } n \to +\infty\}$. Then the following hold:*

(i) *$\Omega_0 = \{\overline{z} \in \mathcal{K} : \exists z_{k_n} \to \overline{z}$ with $h(z_{k_n}) \to h(\overline{z})$ as $n \to +\infty\}$, $\Omega_0 \neq \varnothing$, and, for all $\overline{z} \in \Omega_0$,*

$$h(z_n) \downarrow h(\overline{z}) \quad \text{as } n \to +\infty.$$

(ii) *If $\underline{\alpha} := \inf_{n\in\mathbb{N}} \alpha_n > 0$, then, for all $\overline{z} \in \Omega_0$,*

$$\sum_{n=0}^{+\infty} \Delta_n^2 \le \frac{h(z_0) - h(\overline{z})}{\underline{\alpha}} < +\infty$$

and, consequently, $\Delta_n \to 0$ as $n \to +\infty$.

(iii) *If (H2) holds and $\underline{\delta} := \inf_{n\in\mathbb{N}, i\in I} \alpha_{n-i}\beta_n^2 > 0$, then, for all $n \ge \max\{0, \bar{\imath}\}$,*

$$\operatorname{dist}(0, \partial_L h(z_n)) \le \sqrt{\frac{h(z_{n-\bar{\imath}}) - h(z_{n+1-\underline{\imath}})}{\underline{\delta}}} + \frac{\varepsilon_n}{\beta_n}.$$

If additionally $\lim_{n\to+\infty} \varepsilon_n/\beta_n = 0$, then, for all $\overline{z} \in \Omega_0$,

$$0 \in \partial_L h(\overline{z}).$$

Proof. (i): We first have from (H1) that $(h(z_n))_{n\in\mathbb{N}}$ is nonincreasing. Therefore, $(h(z_n))_{n\in\mathbb{N}}$ is convergent if and only if it has a converging subsequence. It follows that

$$\Omega_0 = \{\overline{z} \in \mathcal{K} : \exists z_{k_n} \to \overline{z} \text{ with } h(z_{k_n}) \to h(\overline{z}) \text{ as } n \to +\infty\}$$

and by (H3), $\Omega_0 \neq \varnothing$. The remaining statement follows from the definition of Ω_0 and the monotonicity of $(h(z_n))_{n\in\mathbb{N}}$.

(ii): Let $\overline{z} \in \Omega_0$. By (H1) and (i),

$$\sum_{n=0}^{+\infty} \alpha_n \Delta_n^2 \le \sum_{n=0}^{+\infty} (h(z_n) - h(z_{n+1})) = h(z_0) - h(\overline{z}).$$

Since $\underline{\alpha} = \inf_{n\in\mathbb{N}} \alpha_n > 0$, it follows that

$$\sum_{n=0}^{+\infty} \Delta_n^2 \le \frac{h(z_0) - h(\overline{z})}{\underline{\alpha}} < +\infty,$$

and hence, $\Delta_n \to 0$ as $n \to +\infty$.

(iii): Assume that (H2) holds and $\underline{\delta} := \inf_{n\in\mathbb{N},i\in I} \alpha_{n-i}\beta_n^2 > 0$. Let $n \geq \max\{0, \bar{\imath}\}$. Applying Cauchy–Schwarz inequality and using the fact that $\sum_{i\in I} \lambda_i^2 \leq \sum_{i\in I} \lambda_i = 1$, we have

$$\left(\sum_{i\in I} \lambda_i \Delta_{n-i}\right)^2 \leq \left(\sum_{i\in I} \lambda_i^2\right)\left(\sum_{i\in I} \Delta_{n-i}^2\right) \leq \sum_{i\in I} \Delta_{n-i}^2.$$

Combining with (H2) and then with (H1) yields

$$\beta_n \operatorname{dist}(0, \partial_L h(z_n)) \leq \sqrt{\sum_{i\in I} \Delta_{n-i}^2} + \varepsilon_n \leq \sqrt{\sum_{i\in I} \frac{h(z_{n-i}) - h(z_{n+1-i})}{\alpha_{n-i}}} + \varepsilon_n.$$

Since $\underline{\delta} = \inf_{n\in\mathbb{N},i\in I} \alpha_{n-i}\beta_n^2 > 0$, we derive that

$$\begin{aligned}
\operatorname{dist}(0, \partial_L h(z_n)) &\leq \sqrt{\sum_{i\in I} \frac{h(z_{n-i}) - h(z_{n+1-i})}{\alpha_{n-i}\beta_n^2}} + \frac{\varepsilon_n}{\beta_n} \\
&\leq \sqrt{\sum_{i\in I} \frac{h(z_{n-i}) - h(z_{n+1-i})}{\underline{\delta}}} + \frac{\varepsilon_n}{\beta_n} \\
&= \sqrt{\frac{h(z_{n-\bar{\imath}}) - h(z_{n+1-\underline{\imath}})}{\underline{\delta}}} + \frac{\varepsilon_n}{\beta_n}.
\end{aligned}$$

Finally, if $\lim_{n\to+\infty} \varepsilon_n/\beta_n = 0$, then, noting from (i) that $(h(z_n))_{n\in\mathbb{N}}$ is convergent, we get $\lim_{n\to+\infty} \operatorname{dist}(0, \partial_L h(z_n)) = 0$. This shows that $0 \in \partial_L h(\overline{z})$ for all $\overline{z} \in \Omega_0$, which completes the proof. □

Theorem 5.2 (Abstract convergence). *Suppose that (H1), (H2), (H3), and (H4) hold and that the sequence $(z_n)_{n\in\mathbb{N}}$ is bounded. Let Ω be the set of cluster points of $(z_n)_{n\in\mathbb{N}}$ and suppose that h is constant on Ω and satisfies the KL property at each point of Ω. Set $\Omega_0 := \{\overline{z} \in \Omega : h(z_n) \to h(\overline{z}) \text{ as } n \to +\infty\}$ and $\overline{h} := h(z)$ for $z \in \Omega_0$. Then the following hold:*

(i) *The sequence $(\Delta_n)_{n\in\mathbb{N}}$ satisfies*

$$\sum_{n=0}^{+\infty} \Delta_n < +\infty.$$

(ii) *If (H5) holds, then $\sum_{n=0}^{+\infty} \|x_{n+1} - x_n\| < +\infty$, and the sequence $(x_n)_{n\in\mathbb{N}}$ is convergent.*

(iii) *If $\inf_{n\in\mathbb{N}} \beta_n > 0$, then, for all $\overline{z} \in \Omega_0$,*

$$0 \in \partial_L h(\overline{z}).$$

(iv) *Suppose further that h satisfies the KL property at every point of Ω with an exponent of $\alpha \leq 1/2$, that $\underline{\imath} \leq 1$, and that*

$$\delta := \inf_{n\in\mathbb{N},i\in I} \alpha_{n-i}\beta_n^2 > 0 \quad \text{and} \quad \frac{\varepsilon_n}{\beta_n} = O\left(\sqrt{h(z_{n-\bar{\imath}}) - h(z_{n+1-\underline{\imath}})}\right) \quad \text{as } n \to +\infty. \tag{21}$$

Then there exist $\gamma_1 \in \mathbb{R}_{++}$ and $\rho \in (0,1)$ such that, for all $n \in \mathbb{N}$,

$$h(z_n) - \overline{h} \leq \gamma_1 \rho^n.$$

Moreover, if additionally (H5) holds and $\sum_{k=n}^{+\infty} \varepsilon_k = O\left(\sqrt{h(z_{n-\bar{\imath}}) - \overline{h}}\right)$ as $n \to +\infty$, then there exist $\overline{x} \in \mathcal{H}$ and $\gamma_2 \in \mathbb{R}_{++}$ such that, for all $n \in \mathbb{N}$,

$$\|x_n - \overline{x}\| \leq \gamma_2 \rho^{\frac{n}{2}}.$$

Proof. First, $\Omega_0 \neq \varnothing$ due to Lemma 5.1(i). Let $\overline{z} \in \Omega_0$. Again by Lemma 5.1(i),

$$h(z_n) \downarrow \overline{h} = h(\overline{z}) \quad \text{as } n \to +\infty.$$

(i): Noting that, for all $n \in \mathbb{N}$, $h(z_n) \geq h(\overline{z})$, we distinguish the following two cases.

Case 1: There exists $n_1 \in \mathbb{N}$ such that $h(z_{n_1}) = h(\overline{z})$. Then, since $(h(z_n))_{n\in\mathbb{N}}$ is nondecreasing, $h(z_n) = h(\overline{z})$ for all $n \geq n_1$. It follows from (H1) that $\Delta_n = 0$ for all $n \geq n_1$, so $\sum_{n=0}^{+\infty} \Delta_n < +\infty$.

Case 2: For all $n \in \mathbb{N}$, $h(z_n) > h(\overline{z})$. We derive from Lemma 2.3 that there exist $\eta \in (0, +\infty]$, $\varphi \in \Phi_\eta$, and $n_0 \in \mathbb{N}$ such that, for all $n \geq n_0$,

$$\varphi'(h(z_n) - h(\overline{z}))\operatorname{dist}(0, \partial_L h(z_n)) \geq 1. \tag{22}$$

Setting $r_n := h(z_n) - h(\overline{z}) \downarrow 0$, by combining with (H1), (H2), (H4), and the concavity of φ, it follows that, for all $n \geq n_0$,

$$\begin{aligned}
\Delta_n^2 &\leq \frac{1}{\alpha_n}(h(z_n) - h(z_{n+1}))\varphi'(h(z_n) - h(\overline{z}))\operatorname{dist}(0, \partial_L h(z_n)) \\
&\leq \frac{1}{\alpha_n \beta_n}\Big(\varphi(r_n) - \varphi(r_{n+1})\Big)\Big(\sum_{i\in I} \lambda_i \Delta_{n-i} + \varepsilon_n\Big) \\
&\leq \frac{1}{\underline{\gamma}}\Big(\varphi(r_n) - \varphi(r_{n+1})\Big)\Big(\sum_{i\in I} \lambda_i \Delta_{n-i} + \varepsilon_n\Big).
\end{aligned}$$

Using the inequality of arithmetic and geometric means (AM-GM) gives us that, for all $n \geq n_0$,

$$2\Delta_n \leq \frac{1}{\underline{\gamma}}\Big(\varphi(r_n) - \varphi(r_{n+1})\Big) + \Big(\sum_{i\in I} \lambda_i \Delta_{n-i} + \varepsilon_n\Big).$$

Since this inequality holds for all $n \geq n_0$, we derive that, for all $m \geq n \geq \max\{n_0, \bar{\imath}\}$,

$$2\sum_{k=n}^{m} \Delta_k \leq \frac{1}{\underline{\gamma}}\Big(\varphi(r_n) - \varphi(r_{m+1})\Big) + \sum_{k=n}^{m}\sum_{i\in I} \lambda_i \Delta_{k-i} + \sum_{k=n}^{m} \varepsilon_k. \tag{23}$$

We have that

$$\sum_{k=n}^{m}\sum_{i\in I} \lambda_i \Delta_{k-i} = \sum_{i\in I} \lambda_i \sum_{k=n}^{m} \Delta_{k-i} = \sum_{i\in I} \lambda_i \sum_{k=n-i}^{m-i} \Delta_k \leq \sum_{i\in I} \lambda_i \sum_{k=n-\bar{\imath}}^{m-\underline{\imath}} \Delta_k = \sum_{k=n-\bar{\imath}}^{m-\underline{\imath}} \Delta_k,$$

using the fact that $\Delta_k \geq 0$ for all $k \in \mathbb{Z}$ and that $\sum_{i\in I} \lambda_i = 1$. Now, by adopting the convention that a summation is zero when the starting index is larger than the ending index,

$$\sum_{k=n-\bar{\imath}}^{m-\underline{\imath}} \Delta_k \leq \sum_{k=n}^{m} \Delta_k + \sum_{k=n-\bar{\imath}}^{n-1} \Delta_k + \sum_{k=m+1}^{m-\underline{\imath}} \Delta_k = \sum_{k=n}^{m} \Delta_k + \sum_{i=1}^{\bar{\imath}} \Delta_{n-i} + \sum_{i=\underline{\imath}}^{-1} \Delta_{m-i}.$$

We continue (23) as

$$\sum_{k=n}^{m} \Delta_k \leq \frac{1}{\underline{\gamma}}\Big(\varphi(r_n) - \varphi(r_{m+1})\Big) + \sum_{i=1}^{\bar{\imath}} \Delta_{n-i} + \sum_{i=\underline{\imath}}^{-1} \Delta_{m-i} + \sum_{k=n}^{m} \varepsilon_k.$$

Letting $m \to +\infty$ and noting from Lemma 5.1(ii) that $\Delta_m \to 0$, we obtain

$$\sum_{k=n}^{+\infty} \Delta_k \leq \frac{1}{\underline{\gamma}}\varphi(r_n) + \sum_{i=1}^{\bar{\imath}} \Delta_{n-i} + \sum_{k=n}^{+\infty} \varepsilon_k < +\infty, \tag{24}$$

which yields $\sum_{n=0}^{+\infty}\Delta_n<+\infty$.

(ii): This follows from (i) and (H5).

(iii): As $\inf_{n\in\mathbb{N}}\beta_n>0$, noting that $\inf_{n\in\mathbb{N}}\alpha_n>0$ and $\lim_{n\to+\infty}\varepsilon_n=0$, we have $\inf_{n\in\mathbb{N},i\in I}\alpha_{n-i}\beta_n^2>0$ and $\lim_{n\to+\infty}\varepsilon_n/\beta_n=0$. Therefore, the conclusion of this part follows from Lemma 5.1(iii).

(iv): Using Lemma 5.1(iii) and (21), and by increasing n_0 if necessary, we find $c_1\in\mathbb{R}_{++}$ such that, for all $n\geq n_0$,

$$\operatorname{dist}(0,\partial_L h(z_n))\leq\sqrt{\frac{h(z_{n-\bar{\imath}})-h(z_{n+1-\underline{\imath}})}{\delta}}+\frac{\varepsilon_n}{\beta_n}\leq c_1\sqrt{h(z_{n-\bar{\imath}})-h(z_{n+1-\underline{\imath}})}.$$

Combining with (22) yields

$$1\leq c_1\varphi'(h(z_n)-h(\overline{z}))\sqrt{h(z_{n-\bar{\imath}})-h(z_{n+1-\underline{\imath}})}.$$

Since $r_n=h(z_n)-h(\overline{z})$, it follows that

$$1\leq c_1^2\left[\varphi'(r_n)\right]^2(r_{n-\bar{\imath}}-r_{n+1-\underline{\imath}}).$$

As φ can be chosen as $\varphi(s)=\gamma s^{1-\alpha}$ for some $\gamma\in\mathbb{R}_{++}$, there exists $c_2\in\mathbb{R}_{++}$ such that, for all $n\geq n_0$,

$$c_2 r_n^{2\alpha}\leq r_{n-\bar{\imath}}-r_{n+1-\underline{\imath}}.$$

Since $r_n\downarrow 0$, $2\alpha\leq 1$, and $1-\underline{\imath}\geq 0$, by increasing n_0 if necessary, it holds that, for all $n\geq n_0$, $r_n^{2\alpha}\geq r_n\geq r_{n+1-\underline{\imath}}$. We deduce that, for all $n\geq n_0$,

$$r_{n+1-\underline{\imath}}\leq\frac{1}{c_2+1}r_{n-\bar{\imath}},$$

and hence, there exist $\gamma_1\in\mathbb{R}_{++}$ and $\rho\in(0,1)$ such that, for all $n\in\mathbb{N}$, $0\leq r_n=h(z_n)-h(\overline{z})\leq\gamma_1\rho^n$.

Now, recalling the convention that a summation is zero when the starting index is larger than the ending index, it follows from Cauchy–Schwarz inequality, (H1), and (H4) that

$$\begin{aligned}\left(\sum_{i=1}^{\bar{\imath}}\Delta_{n-i}\right)^2\leq\max\{0,\bar{\imath}\}\sum_{i=1}^{\bar{\imath}}\Delta_{n-i}^2&\leq\max\{0,\bar{\imath}\}\sum_{i=1}^{\bar{\imath}}\frac{h(z_{n-i})-h(z_{n+1-i})}{\underline{\alpha}}\\&\leq\frac{\max\{0,\bar{\imath}\}}{\underline{\alpha}}\max\{0,h(z_{n-\bar{\imath}})-h(z_n)\}\\&\leq\frac{\max\{0,\bar{\imath}\}}{\underline{\alpha}}r_{n-\bar{\imath}}.\end{aligned}$$

Combining with (24) gives, for all $n\geq n_0$,

$$\sum_{k=n}^{+\infty}\Delta_k\leq\frac{1}{\underline{\gamma}}\varphi(r_n)+\sqrt{\frac{\max\{0,\bar{\imath}\}}{\underline{\alpha}}r_{n-\bar{\imath}}}+\sum_{k=n}^{+\infty}\varepsilon_k.$$

As $\sum_{k=n}^{+\infty}\varepsilon_k=O\left(\sqrt{h(z_{n-\bar{\imath}})-h(\overline{z})}\right)=O(\sqrt{r_{n-\bar{\imath}}})$ as $n\to+\infty$ and $\varphi(r_n)=\gamma r_n^{1-\alpha}\leq\gamma r_{n-\bar{\imath}}^{1-\alpha}\leq\gamma\sqrt{r_{n-\bar{\imath}}}$ for all n large enough, by increasing n_0 if necessary, there exists $c_3\in\mathbb{R}_{++}$ such that, for all $n\geq n_0$,

$$\sum_{k=n}^{+\infty}\Delta_k\leq c_3\sqrt{r_{n-\bar{\imath}}}\leq c_3\sqrt{\gamma_1}\rho^{\frac{n-\bar{\imath}}{2}}.$$

Since (H5) holds, (ii) implies that $(x_n)_{n\in\mathbb{N}}$ is convergent to some $\overline{x}\in\mathcal{H}$. Then, for all $n\in\mathbb{N}$,

$$\|x_n-\overline{x}\|\leq\sum_{k=n}^{+\infty}\|x_{k+1}-x_k\|\leq c\sum_{k=n}^{+\infty}\Delta_k$$

and the conclusion follows. □

Remark 5.3 (Parameter conditions). In view of (H4) and as shown in the proof of Theorem 5.2(iii), if $\inf_{n\in\mathbb{N}}\beta_n>0$, then the conditions $\inf_{n\in\mathbb{N},i\in I}\alpha_{n-i}\beta_n^2>0$ and $\lim_{n\to+\infty}\varepsilon_n/\beta_n=0$ in Lemma 5.1(iii) are guaranteed. If additionally $\varepsilon_n=O(h(z_{n-\bar{\imath}})-h(z_{n+1-\underline{\imath}}))$ as $n\to+\infty$, then the parameter conditions

$$\frac{\varepsilon_n}{\beta_n}=O\left(\sqrt{h(z_{n-\bar{\imath}})-h(z_{n+1-\underline{\imath}})}\right)\quad\text{and}\quad\sum_{k=n}^{+\infty}\varepsilon_k=O\left(\sqrt{h(z_{n-\bar{\imath}})-h(\overline{z})}\right)\quad\text{as } n\to+\infty.$$

in Theorem 5.2(iv) are also satisfied. Indeed, since $h(z_n)\downarrow h(\overline{z})$, we have $h(z_{n-\bar{\imath}})-h(z_{n+1-\underline{\imath}})\to 0$ as $n\to+\infty$, and so, for all n large enough, $h(z_{n-\bar{\imath}})-h(z_{n+1-\underline{\imath}})\le\sqrt{h(z_{n-\bar{\imath}})-h(z_{n+1-\underline{\imath}})}$. It follows that $\varepsilon_n=O(\sqrt{h(z_{n-\bar{\imath}})-h(z_{n+1-\underline{\imath}})})$ and, since $\inf_{n\in\mathbb{N}}\beta_n>0$, $\varepsilon_n/\beta_n=O(\sqrt{h(z_{n-\bar{\imath}})-h(z_{n+1-\underline{\imath}})})$ as $n\to+\infty$. Now, we note that

$$\sum_{k=n}^{+\infty}(h(z_{k-\bar{\imath}})-h(z_{k+1-\underline{\imath}}))=\sum_{i=\underline{\imath}}^{\bar{\imath}}(h(z_{n-i})-h(\overline{z}))\le(\bar{\imath}-\underline{\imath}+1)(h(z_{n-\bar{\imath}})-h(\overline{z})),$$

which implies that $\sum_{k=n}^{+\infty}\varepsilon_k=O(h(z_{n-\bar{\imath}})-h(\overline{z}))$, and so $\sum_{k=n}^{+\infty}\varepsilon_k=O\left(\sqrt{h(z_{n-\bar{\imath}})-h(\overline{z})}\right)$ as $n\to+\infty$.

Remark 5.4 (Comparison to the existing literature). The general framework (H1)–(H5) extends various convergence conditions for exact and inexact descent methods in the literature. Specifically, in [3, 10], the authors proposed conditions that satisfied (H1)–(H5) with $\mathcal{K}=\mathcal{H}=\mathbb{R}^N$, $z_n=x_n$, $\Delta_n=\|x_{n+1}-x_n\|$, $\alpha_n\equiv a$, $\beta_n\equiv 1/b$, $\varepsilon_n\equiv 0$, $I=\{1\}$, and $\lambda_1=1$. These conditions were then generalized in [14] to flexible parameters and real Hilbert spaces. In the finite-dimensional setting, the conditions in [14] fulfill (H1)–(H5) with $\mathcal{K}=\mathcal{H}$, $z_n=x_n$, $\Delta_n=\|x_{n+1}-x_n\|$, $I=\{1\}$, and $\lambda_1=1$.

The framework (H1)–(H5) also holds in the case of [9, Proposition 4] with $\mathcal{K}=\mathcal{H}=\mathbb{R}^N$, $z_n=x_n$, $\Delta_n=\|x_{n+2}-x_{n+1}\|$, $\alpha_n\equiv a$, $\beta_n\equiv 1/b$, $\varepsilon_n\equiv 0$, $I=\{1\}$, and $\lambda_1=1$. Here, Δ_n is shifted one step forward comparing to the two aforementioned studies. This difference makes the relative error condition explicit; see [24, Section 2.4] for a discussion.

In [26], the authors provided a framework for convergence analysis of iPiano, a proximal gradient algorithm with extrapolation. In turn, their conditions satisfied (H1)–(H5) with $\mathcal{K}=\mathcal{H}^2$, $z_n=(x_n,x_{n-1})$, $\Delta_n=\|x_n-x_{n-1}\|$, $\alpha_n\equiv a$, $\beta_n\equiv 1/b$, $\varepsilon_n\equiv 0$, $I=\{0,1\}$, and $\lambda_0=\lambda_1=1/2$. Recently, these conditions have been extended in [25] with $\mathcal{H}=\mathbb{R}^N$, $\mathcal{K}=\mathbb{R}^{N+P}$ and $z_n=(x_n,u_n)$. It is worth noting that the finite index set I of integers in [25] can always be written as $I=\{\underline{\imath},\underline{\imath}+1,\ldots,\bar{\imath}\}$ for $\underline{\imath}\le\bar{\imath}$. To get the global convergence of $(x_n)_{n\in\mathbb{N}}$, [25, Theorem 10] not only needs (H5) as our Theorem 5.2 but also requires that h is bounded from below and that, for any converging subsequence $(z_{k_n})_{n\in\mathbb{N}}$ of $(z_n)_{n\in\mathbb{N}}$,

$$z_{k_n}\to\widetilde{z}\quad\text{and}\quad h(z_{k_n})\to h(\widetilde{z})\quad\text{as } n\to+\infty,$$

which implies that h is constant on Ω. We also note that linear convergence of $(x_n)_{n\in\mathbb{N}}$ has been not investigated in the framework of [25, 26].

Next, we show that the full sequence generated by Algorithm 4.1 is globally convergent by further assuming that a suitable merit function is a KL function. We note that, as we will see later in Remark 5.6, this assumption is automatically fulfilled if f and g are both semi-algebraic functions and S is a semi-algebraic set, which, in particular, holds for all the motivating examples mentioned before.

Theorem 5.5 (Global convergence). *Let* $\liminf_{n\to\infty}\tau_n = \overline{\tau} > 0$ *and let* $(x_n)_{n\in\mathbb{N}}$ *be the sequence generated by Algorithm 4.1. Suppose that Assumptions A1, A2, and A3 hold, that* g *is differentiable on an open set containing* $S\cap\operatorname{dom} f$ *whose gradient* ∇g *is Lipschitz continuous*[4] *with modulus* ℓ_g *on* $S\cap\operatorname{dom} f$*, that, for* c *given in (14),*

$$h(x,y) := \frac{f(x)}{g(x)} + \iota_S(x) + c\|x-y\|^2$$

satisfies the KL property at $(\overline{x},\overline{x})$ *for all* $\overline{x}\in S\cap\operatorname{dom} f$*, and that the set* $\{x\in S : \frac{f(x)}{g(x)} \le \frac{f(x_0)}{g(x_0)}\}$ *is bounded. Then* $\sum_{n=0}^{+\infty}\|x_{n+1}-x_n\| < +\infty$*, and the sequence* $(x_n)_{n\in\mathbb{N}}$ *converges to a stationary point of (P). Moreover, if* h *satisfies the KL property with an exponent of* $\alpha\le 1/2$ *at* $(\overline{x},\overline{x})$ *for all* $\overline{x}\in S\cap\operatorname{dom} f$*, then the convergence rate of* $(x_n)_{n\in\mathbb{N}}$ *and* $(h(x_{n+1},x_n))_{n\in\mathbb{N}}$ *is linear in the sense that there exist* $\gamma_1,\gamma_2\in\mathbb{R}_{++}$ *and* $\rho\in(0,1)$ *such that, for all* $n\in\mathbb{N}$*,*

$$|h(x_{n+1},x_n) - h(x_\infty,x_\infty)| \le \gamma_1\rho^n \quad \text{and} \quad \|x_n - x_\infty\| \le \gamma_2\rho^{\frac{n}{2}}.$$

Proof. Let $z_n = (x_{n+1},x_n)$. Let Ω be the set of cluster points of $(z_n)_{n\in\mathbb{N}}$. Theorem 4.7 asserts that the sequence $(z_n)_{n\in\mathbb{N}}$ is bounded and that, for all $n\in\mathbb{N}$, $x_n\in S\cap\operatorname{dom} f$ and

$$h(z_{n+1}) + \alpha\|x_{n+1}-x_n\|^2 \le h(z_n) \quad \text{with } \alpha>0 \text{ given in (14).} \tag{25}$$

By combining with Corollary 4.9, for every $\overline{z}\in\Omega$, one has $\overline{z} = (\overline{x},\overline{x})$ with $\overline{x}\in S\cap\operatorname{dom} f$ a stationary point of (P) and

$$\theta_n = \frac{f(x_n)}{g(x_n)} \to \frac{f(\overline{x})}{g(\overline{x})} \quad \text{as } n\to+\infty.$$

In particular, $h(z_n) = h(x_{n+1},x_n) = \frac{f(x_{n+1})}{g(x_{n+1})} + c\|x_{n+1}-x_n\|^2 \to \frac{f(\overline{x})}{g(\overline{x})}$ as $n\to+\infty$.

From Step 2 of Algorithm 4.1 and noting that $g_n = \nabla g(x_n)$, we have for all $n\in\mathbb{N}$,

$$0\in\partial_L(f^{\mathfrak{n}}+\iota_S)(x_{n+1}) + \nabla f^{\mathfrak{s}}(u_n) + \frac{1}{\tau_n}(x_{n+1}-v_n-\tau_n\theta_n\nabla g(x_n)) + \ell(x_{n+1}-u_n),$$

which combined with $\partial_L(f+\iota_S) = \nabla f^{\mathfrak{s}} + \partial_L(f^{\mathfrak{n}}+\iota_S)$ yields

$$\begin{aligned}\hat{x}_{n+1} &:= \nabla f^{\mathfrak{s}}(x_{n+1}) - \nabla f^{\mathfrak{s}}(u_n) - \ell(x_{n+1}-u_n) - \frac{1}{\tau_n}(x_{n+1}-v_n) + \theta_n\nabla g(x_n)\\ &\in \partial_L(f+\iota_S)(x_{n+1}).\end{aligned}$$

Since g is continuously differentiable at x_n and $g(x_n)>0$, it follows from Lemma 3.2(ii) that

$$\begin{aligned}\partial_L\left(\frac{f}{g}+\iota_S\right)(x_{n+1}) &= \frac{g(x_{n+1})\partial_L(f+\iota_S)(x_{n+1}) - f(x_{n+1})\nabla g(x_{n+1})}{g(x_{n+1})^2}\\ &= \frac{\partial_L(f+\iota_S)(x_{n+1}) - \theta_{n+1}\nabla g(x_{n+1})}{g(x_{n+1})},\end{aligned}$$

and so

$$x_n^* := \frac{\hat{x}_{n+1}-\theta_{n+1}\nabla g(x_{n+1})}{g(x_{n+1})} \in \partial_L\left(\frac{f}{g}+\iota_S\right)(x_{n+1}).$$

Therefore, $(x_n^* + 2c(x_{n+1}-x_n), 2c(x_n-x_{n+1})) \in \partial_L h(z_n)$.

[4] Assumption on Lipschitz continuity of ∇g can be relaxed to a weaker assumption that there exist $\varepsilon,\ell_g\in\mathbb{R}_{++}$ such that $\|\nabla g(x)-\nabla g(y))\| \le \ell_g\|x-y\|$ for all $x,y\in S\cap\operatorname{dom} f$ with $\|x-y\|\le\varepsilon$.

Note that $\tau_n \leq 1/\max\{\sqrt{\beta}\theta_n/\zeta, \delta\} \leq \frac{1}{\delta}$, so $\mu_n \leq \overline{\mu}\tau_n \leq \frac{\overline{\mu}}{\delta}$. Next, we see that, for all $n \in \mathbb{N}$,

$$\|x_{n+1} - v_n\| \leq \|x_{n+1} - x_n\| + \mu_n\|x_n - x_{n-1}\| \leq \|x_{n+1} - x_n\| + \frac{\overline{\mu}}{\delta}\|x_n - x_{n-1}\|,$$
$$\|x_{n+1} - u_n\| \leq \|x_{n+1} - x_n\| + \kappa_n\|x_n - x_{n-1}\| \leq \|x_{n+1} - x_n\| + \overline{\kappa}\|x_n - x_{n-1}\|,$$

and by the Lipschitz continuity of $\nabla f^{\mathfrak{s}}$,

$$\|\nabla f^{\mathfrak{s}}(x_{n+1}) - \nabla f^{\mathfrak{s}}(u_n)\| \leq \ell\|x_{n+1} - u_n\| \leq \ell\|x_{n+1} - x_n\| + \ell\overline{\kappa}\|x_n - x_{n-1}\|.$$

Since $(x_n)_{n\in\mathbb{N}}$ is bounded, the continuity of ∇g implies that $(\nabla g(x_n))_{n\in\mathbb{N}}$ is also bounded. There thus exists $\mu \in \mathbb{R}_{++}$ such that, for all $n \in \mathbb{N}$, $\|\nabla g(x_n)\| \leq \mu$. Since $\liminf_{n\to+\infty} \tau_n = \overline{\tau} > 0$ and $\lim_{n\to+\infty} \|x_{n+1} - x_n\| = 0$ (see Theorem 4.7(ii)), there exists $n_0 \in \mathbb{N}$ such that, for all $n \geq n_0$,

$$\tau_n \geq \overline{\tau}/2 \quad \text{and} \quad \|x_n - x_{n-1}\| \leq \varepsilon.$$

Now, from the definition of $h(z_n)$, we see that

$$\begin{aligned}\theta_n\nabla g(x_n) - \theta_{n+1}\nabla g(x_{n+1}) = {} & \theta_n(\nabla g(x_n) - \nabla g(x_{n+1})) - c\|x_n - x_{n-1}\|^2\nabla g(x_{n+1}) \\ & + c\|x_{n+1} - x_n\|^2\nabla g(x_{n+1}) + (h(z_{n-1}) - h(z_n))\nabla g(x_{n+1})\end{aligned}$$

and by the Lipschitz continuity of ∇g and the boundedness of $(\nabla g(x_n))$, for all $n \geq n_0$,

$$\begin{aligned}\|\theta_n\nabla g(x_n) - \theta_{n+1}\nabla g(x_{n+1})\| \leq {} & \ell_g\theta_n\|x_{n+1} - x_n\| + c\varepsilon\mu\|x_n - x_{n-1}\| \\ & + c\varepsilon\mu\|x_{n+1} - x_n\| + \mu(h(z_{n-1}) - h(z_n)).\end{aligned}$$

Altogether, it follows from the definition of x_n^* that, for all $n \geq n_0$,

$$\begin{aligned}\|\hat{x}_{n+1} - \theta_{n+1}\nabla g(x_{n+1})\| &\leq \|\nabla f^{\mathfrak{s}}(x_{n+1}) - \nabla f^{\mathfrak{s}}(u_n)\| + \ell\|x_{n+1} - u_n\| + \frac{1}{\tau_n}\|x_{n+1} - v_n\| \\ &\quad + \|\theta_n\nabla g(x_n) - \theta_{n+1}\nabla g(x_{n+1})\| \\ &\leq 2\ell\|x_{n+1} - x_n\| + 2\ell\overline{\kappa}\|x_n - x_{n-1}\| + \frac{2}{\overline{\tau}}(\|x_{n+1} - x_n\| + \frac{\overline{\mu}}{\delta}\|x_n - x_{n-1}\|) \\ &\quad + (\ell_g\theta_n + c\varepsilon\mu)\|x_{n+1} - x_n\| + c\varepsilon\mu\|x_n - x_{n-1}\| + \mu(h(z_{n-1}) - h(z_n)).\end{aligned}$$

Recalling that $(\theta_n)_{n\in\mathbb{N}}$ is convergent and hence bounded and noting that $\inf_{n\in\mathbb{N}} g(x_n) > 0$ due to the continuity and positivity of g, the boundedness of $(x_n)_{n\in\mathbb{N}}$ and the closedness of S, we find $K \in \mathbb{R}_{++}$ such that, for all $n \geq n_0$,

$$\begin{aligned}\|x_n^*\| &= \frac{\|\hat{x}_{n+1} - \theta_{n+1}\nabla g(x_{n+1})\|}{|g(x_{n+1})|} \\ &\leq K\left(\|x_{n+1} - x_n\| + \|x_n - x_{n-1}\| + (h(z_{n-1}) - h(z_n))\right).\end{aligned}$$

We deduce that there exists $K_1 \in \mathbb{R}_{++}$ such that, for all $n \geq n_0$,

$$\begin{aligned}\operatorname{dist}(0, \partial_L h(z_n)) &\leq \sqrt{\|x_n^* + 2c(x_{n+1} - x_n)\|^2 + 4c^2\|x_n - x_{n+1}\|^2} \\ &\leq \sqrt{2\|x_n^*\|^2 + 8c^2\|x_{n+1} - x_n\|^2 + 4c^2\|x_n - x_{n+1}\|^2} \\ &\leq K_1\left(\|x_{n+1} - x_n\| + \|x_n - x_{n-1}\| + h(z_{n-1}) - h(z_n)\right),\end{aligned}$$

where the second inequality is from the elementary inequality that $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$. Now, by applying Theorem 5.2 and Remark 5.3 with $I = \{0,1\}$, $\lambda_0 = \lambda_1 = 1/2$, $\Delta_n = 2K_1\|x_{n+1} - x_n\|$, $\alpha_n \equiv \frac{\alpha}{4K_1^2} > 0$, $\beta_n \equiv 1$, and $\varepsilon_n = K_1(h(z_{n-1}) - h(z_n)) \leq K_1(h(z_{n-1}) - h(z_{n+1}))$, we get the conclusion. □

Remark 5.6 (KL property of the merit function). In Theorem 5.5, we impose the assumption that the merit function $h(x,y)=\frac{f(x)}{g(x)}+\iota_S(x)+c\|x-y\|^2$ is a KL function with c given in (14). Note that sum or quotient of two semi-algebraic functions is a semi-algebraic function, and indicator function of a semi-algebraic set (sets described as union or intersections of finitely many sets which can be expressed as lower level sets of polynomials) is also a semi-algebraic function. We note that this assumption is automatically satisfied when f and g are semi-algebraic functions, and S is a semi-algebraic set. This, in particular, covers all the motivating examples we mentioned in the introduction.

It is also known from [19, Theorem 3.6] that the merit function h has the KL property with an exponent of $\alpha\in[1/2,1)$ at $(\overline{x},\overline{x})$ as long as $f/g+\iota_S$ has the KL property with an exponent of α at $\overline{x}$.

Next, we see that Algorithm 4.1 converges at a linear rate when applied to the scale invariant sparse signal recovery problem and Rayleigh quotient optimization with spherical constraint, if the parameters τ_n satisfy $\liminf_{n\to\infty}\tau_n=\overline{\tau}>0$.

Proposition 5.7 (KL exponent 1/2 & linear convergence). *Let $\mathcal{H}=\mathbb{R}^N$, and suppose that one of the following holds:*

(i) *$f(x)=x^\top Ax+\iota_C(x)$, $g(x)=x^\top Bx$, and $S=\mathbb{R}^N$, where A and B are symmetric positive definite matrices and $C:=\{x\in\mathbb{R}^N:\|x\|=1\}$.*

(ii) *$f(x)=\|x\|_1$, $g(x)=\|x\|_2$, and $S=\{x\in\mathbb{R}^N:Ax\le b,Cx=d\}$, where $A\in\mathbb{R}^{M\times N}$, $b\in\mathbb{R}^M$, $C\in\mathbb{R}^{P\times N}$, and $d\in\mathbb{R}^P$.*

Then, for all $c\in\mathbb{R}_+$, $h(x,y)=\frac{f(x)}{g(x)}+\iota_S(x)+c\|x-y\|^2$ satisfies the KL property with an exponent of 1/2 at $(\overline{x},\overline{x})$ for all $\overline{x}\in\operatorname{dom}f$. Consequently, if $\liminf_{n\to\infty}\tau_n=\overline{\tau}>0$, then Algorithm 4.1 exhibits linear convergence when applied to the above cases.

Proof. In view of [19, Theorem 3.6] and Theorem 5.5, it suffices to show that $F:=f/g+\iota_S$ is a KL function with an exponent of 1/2.

(i): We see that $F(x)=\frac{x^\top Ax}{x^\top Bx}+\iota_C(x)$. For all $x\notin C$, $\partial_L F(x)=\varnothing$. For all $x\in C$, since $\partial_L\iota_C(x)=N_C(x)=\{\xi x:\xi\in\mathbb{R}\}$, it holds that

$$\partial_L F(x)=\left\{\frac{2(x^\top Bx)Ax-2(x^\top Ax)Bx}{(x^\top Bx)^2}+\xi x:\xi\in\mathbb{R}\right\}. \tag{26}$$

Let $\overline{x}\in\operatorname{dom}\partial_L F$. We must have $\overline{x}\in C$. Let $\varepsilon,\eta\in(0,1)$ and let x be such that $\|x-\overline{x}\|\le\varepsilon$ and $F(\overline{x})<F(x)<F(\overline{x})+\eta$. Then $F(x)<+\infty$, and so $x\in C$. It follows from (26) that

$$\begin{aligned}
\operatorname{dist}(0,\partial_L F(x))&=\inf_{\xi\in\mathbb{R}}\left\|\frac{2(x^\top Bx)Ax-2(x^\top Ax)Bx}{(x^\top Bx)^2}+\xi x\right\|\\
&=\inf_{\xi\in\mathbb{R}}\left(\left\|\frac{2(x^\top Bx)Ax-2(x^\top Ax)Bx}{(x^\top Bx)^2}\right\|^2+\xi^2\right)^{1/2}\\
&=\left\|\frac{2(x^\top Bx)Ax-2(x^\top Ax)Bx}{(x^\top Bx)^2}\right\|\\
&=\frac{2}{x^\top Bx}\left\|(A-F(\overline{x})B)x-(F(x)-F(\overline{x}))Bx\right\|,
\end{aligned}$$

where the second equality follows from the fact that $x^\top(2(x^\top Bx)Ax-2(x^\top Ax)Bx)=0$ and $\|x\|=1$. Now, since $A-F(\overline{x})B$ is a symmetric matrix, there exists $c>0$ such that, for all $z\in\mathbb{R}^N$,

$$\|(A-F(\overline{x})B)z\|^2\ge c(z^\top(A-F(\overline{x})B)z)=c(z^\top Bz)(F(z)-F(\overline{x})).$$

Therefore,

$$\operatorname{dist}(0,\partial_L F(x)) \geq 2(F(x)-F(\overline{x}))^{1/2}\left(\frac{\sqrt{c}}{\sqrt{x^\top Bx}} - (F(x)-F(\overline{x}))^{1/2}\frac{\|Bx\|}{x^\top Bx}\right).$$

Let $\lambda_{\max}$ and $\lambda_{\min}$ are the maximum and minimum eigenvalues of B, respectively. Then $\lambda_{\min} \leq x^\top Bx \leq \lambda_{\max}$ since $\|x\| = 1$. By shrinking η if necessary, we have

$$(F(x)-F(\overline{x}))^{1/2}\frac{\|Bx\|}{x^\top Bx} \leq \eta^{1/2}\frac{\|Bx\|}{\lambda_{\min}} \leq \frac{\sqrt{c}}{2\sqrt{\lambda_{\max}}}.$$

We deduce that $\operatorname{dist}(0,\partial_L F(x)) \geq \frac{\sqrt{c}}{\sqrt{\lambda_{\max}}}(F(x)-F(\overline{x}))^{1/2}$, and F is thus a KL function with an exponent of $1/2$.

(ii): By a similar argument as in [31, Theorem 4.4], F is a KL function with an exponent of $1/2$. □

6. Convergence to strong stationary points In this section, we propose another algorithm which converges to a strong lifted stationary points of the fractional programming problem (P). To do this, we now consider the case where Assumption A2 is replaced by the following stronger assumption.

Assumption A2'. $g(x) = \max\{g_i(x) : 1 \leq i \leq p\}$, where each g_i is continuously differentiable on an open set containing S and weakly convex on S with modulus $\beta \in \mathbb{R}_+$, and (BC) holds.

Recall that the *ε-active set* for $g(x) = \max\{g_i(x) : 1 \leq i \leq p\}$ is defined by

$$I_\varepsilon(x) = \{i \in \{1,\ldots,p\} : g_i(x) \geq g(x) - \varepsilon\}.$$

We then propose an extrapolated proximal subgradient algorithm as follows.

Algorithm 6.1 (e-PSG for strong stationary points).

▷ **Step 1.** Choose $x_{-1} = x_0 \in S \cap \operatorname{dom} f$ and set $n = 0$. Let $\varepsilon, \delta \in \mathbb{R}_{++}$, let $\zeta \in \mathbb{R}_{++}$ be such that $1 - \sqrt{\beta}\zeta > 0$, and let

$$\overline{\mu} \in \left[0, \frac{\delta(1-\sqrt{\beta}\zeta)\sqrt{mM}}{2M}\right) \quad \text{and} \quad \overline{\kappa} \in \left[0, \sqrt{\frac{m\delta(1-\sqrt{\beta}\zeta)}{\ell M} - \frac{2m\overline{\mu}}{\ell\sqrt{mM}}}\right),$$

where ℓ is defined in Assumption A1, β is defined in Assumption A2', while m and M are given in (BC).

▷ **Step 2.** Set $\theta_n = \frac{f(x_n)}{g(x_n)}$ and choose $\tau_n \in \mathbb{R}$ such that $0 < \tau_n \leq 1/\max\{\sqrt{\beta}\theta_n/\zeta, \delta\}$. Let $u_n = x_n + \kappa_n(x_n - x_{n-1})$ with $\kappa_n \in [0,\overline{\kappa}]$ and $v_n = x_n + \mu_n(x_n - x_{n-1})$ with $\mu_n \in [0, \overline{\mu}\tau_n]$. For each $i_n \in I_\varepsilon(x_n)$, find

$$w_n^{i_n} \in \operatorname*{arg\,min}_{x\in S}\left(f^{\mathfrak{n}}(x) + f^{\mathfrak{s}}(u_n) + \langle \nabla f^{\mathfrak{s}}(u_n), x-u_n\rangle + \frac{1}{2\tau_n}\|x - v_n - \tau_n\theta_n\nabla g_{i_n}(x_n)\|^2 + \frac{\ell}{2}\|x-u_n\|^2\right).$$

▷ **Step 3.** Set $x_{n+1} := w_n^{\hat{i}_n}$, where

$$\hat{i}_n \in \operatorname*{arg\,min}_{i_n \in I_\varepsilon(x_n)}\left(f(w_n^{i_n}) - \theta_n g(w_n^{i_n}) + \frac{1}{2}\left(\frac{1-\sqrt{\beta}\zeta}{\tau_n} - \frac{M\mu_n}{\sqrt{mM}\tau_n}\right)\|w_n^{i_n} - x_n\|^2\right).$$

▷ **Step 4.** If a termination criterion is not met, let $n = n+1$ and go to Step 2.

Before we proceed, we note that Step 3 in Algorithm 6.1 is motivated by the recent work of Pang et al. [27] which proposes an enhanced version of the DC algorithm for solving DC programs that converges to a stronger notion of stationary points, namely, to d-stationary points. Similar to the work of Pang et al., in Step 2, we need to compute the proximal mapping of $f^{\mathfrak{n}}+\iota_S$ for $|I_\varepsilon(x_n)|$ times (which is at most p). Although comparing to Algorithm 4.1, the computation cost in solving each subproblem may be higher, as we will see later, the algorithm converges to a strong lifted stationary point of (P).

Theorem 6.2. *Let $(x_n)_{n\in\mathbb{N}}$ be the sequence generated by Algorithm 6.1. Suppose that Assumptions A1 and A2' hold, and that the set $\{x\in S:\frac{f(x)}{g(x)}\le\frac{f(x_0)}{g(x_0)}\}$ is bounded. Then the following hold:*

(i) *For all $n\in\mathbb{N}$, $x_n\in S\cap\operatorname{dom} f$ and*

$$F_n:=\frac{f(x_n)}{g(x_n)}+\left(\frac{\ell\overline{\kappa}^2}{2m}+\frac{\overline{\mu}}{2\sqrt{mM}}\right)\|x_n-x_{n-1}\|^2 \tag{27}$$

is nonincreasing and convergent. Consequently, the sequence $\left(\frac{f(x_n)}{g(x_n)}\right)_{n\in\mathbb{N}}$ is convergent.

(ii) *The sequence $(x_n)_{n\in\mathbb{N}}$ is bounded and*

$$\sum_{n=0}^{+\infty}\|x_{n+1}-x_n\|^2<+\infty.$$

Consequently, $\lim_{n\to+\infty}\|x_{n+1}-x_n\|=0$.

(iii) *If $\liminf_{n\to+\infty}\tau_n=\overline{\tau}>0$, then, for every cluster point $\overline{x}$ of $(x_n)_{n\in\mathbb{N}}$, it holds that $\overline{x}\in S\cap\operatorname{dom} f$, $\lim_{n\to+\infty}\frac{f(x_n)}{g(x_n)}=\frac{f(\overline{x})}{g(\overline{x})}$, and*

$$\frac{f(\overline{x})}{g(\overline{x})}\bigcup_{i\in I_0(\overline{x})}\nabla g_i(\overline{x})\subseteq\partial_L(f+\iota_S)(\overline{x}). \tag{28}$$

In addition, if f is weakly convex on S, then $\overline{x}$ is a strong lifted stationary point of (P).

Proof. (i)&(ii): We first see that, for all $n\in\mathbb{N}$, $x_n\in S\cap\operatorname{dom} f$, and so $g(x_n)>0$ and $\theta_n=\frac{f(x_n)}{g(x_n)}\ge 0$.

Next, for all $n\in\mathbb{N}$, $i_n\in I_\varepsilon(x_n)$, and $x\in S$,

$$\begin{aligned}
f(w_n^{i_n})&=f^{\mathfrak{n}}(w_n^{i_n})+f^{\mathfrak{s}}(w_n^{i_n})\\
&\le f^{\mathfrak{n}}(w_n^{i_n})+f^{\mathfrak{s}}(u_n)+\langle\nabla f^{\mathfrak{s}}(u_n),w_n^{i_n}-u_n\rangle+\frac{\ell}{2}\|w_n^{i_n}-u_n\|^2\\
&\le f^{\mathfrak{n}}(x)+f^{\mathfrak{s}}(u_n)+\langle\nabla f^{\mathfrak{s}}(u_n),x-u_n\rangle+\frac{1}{2\tau_n}\|x-v_n-\tau_n\theta_n\nabla g_{i_n}(x_n)\|^2+\frac{\ell}{2}\|x-u_n\|^2\\
&\quad-\frac{1}{2\tau_n}\|w_n^{i_n}-v_n-\tau_n\theta_n\nabla g_{i_n}(x_n)\|^2\\
&\le f^{\mathfrak{n}}(x)+f^{\mathfrak{s}}(x)+\frac{1}{2\tau_n}\|x-v_n-\tau_n\theta_n\nabla g_{i_n}(x_n)\|^2+\frac{\ell}{2}\|x-u_n\|^2\\
&\quad-\frac{1}{2\tau_n}\|w_n^{i_n}-v_n-\tau_n\theta_n\nabla g_{i_n}(x_n)\|^2\\
&=f(x)+\frac{1}{2\tau_n}\|x-v_n\|^2-\frac{1}{2\tau_n}\|w_n^{i_n}-v_n\|^2+\theta_n\langle\nabla g_{i_n}(x_n),w_n^{i_n}-x\rangle+\frac{\ell}{2}\|x-u_n\|^2\\
&=f(x)+\frac{1}{2\tau_n}\|x-v_n\|^2-\frac{1}{2\tau_n}\|x_n-v_n\|^2-\frac{1}{2\tau_n}\|w_n^{i_n}-x_n\|^2+\frac{\mu_n}{\tau_n}\langle w_n^{i_n}-x_n,x_n-x_{n-1}\rangle\\
&\quad+\theta_n\langle\nabla g_{i_n}(x_n),w_n^{i_n}-x_n\rangle-\theta_n\langle\nabla g_{i_n}(x_n),x-x_n\rangle+\frac{\ell}{2}\|x-u_n\|^2,
\end{aligned}\tag{29}$$

where the first inequality is from the fact that $\nabla f^{\mathfrak{s}}$ is Lipschitz continuous with modulus ℓ (Lemma 4.6), the second inequality is from Step 2 of Algorithm 6.1, the third inequality follows

from the convexity of f^s, and the last equality uses the fact that $x_n - v_n = -\mu_n(x_n - x_{n-1})$. For $\omega = \sqrt{m/M} > 0$, one has from Young's inequality that

$$\begin{aligned}\langle w_n^{i_n} - x_n, x_n - x_{n-1}\rangle &\leq \frac{1}{2\omega}\|w_n^{i_n} - x_n\|^2 + \frac{\omega}{2}\|x_n - x_{n-1}\|^2 \\ &= \frac{M}{2\sqrt{mM}}\|w_n^{i_n} - x_n\|^2 + \frac{m}{2\sqrt{mM}}\|x_n - x_{n-1}\|^2. \end{aligned} \tag{30}$$

It follows from Assumption A2' that g is regular and weakly convex with modulus β on S. By Lemma 4.5,

$$\langle \nabla g_{i_n}(x_n), w_n^{i_n} - x_n\rangle \leq g_{i_n}(w_n^{i_n}) - g_{i_n}(x_n) + \frac{\beta}{2}\|w_n^{i_n} - x_n\|^2. \tag{31}$$

Combining inequalities (29), (30) and (31), and noting that $g_{i_n}(w_n^{i_n}) \leq g(w_n^{i_n})$ by the definition of g and that $\beta\theta_n \leq \sqrt{\beta}\zeta/\tau_n$ by the choice of τ_n, one has

$$\begin{aligned} f(w_n^{i_n}) \leq f(x) &+ \frac{1}{2\tau_n}\|x - v_n\|^2 - \frac{1}{2\tau_n}\|x_n - v_n\|^2 - \frac{1}{2}\left(\frac{1-\sqrt{\beta}\zeta}{\tau_n} - \frac{M\mu_n}{\sqrt{mM}\tau_n}\right)\|w_n^{i_n} - x_n\|^2 \\ &+ \theta_n(g(w_n^{i_n}) - g_{i_n}(x_n)) - \theta_n\langle \nabla g_{i_n}(x_n), x - x_n\rangle + \frac{\ell}{2}\|x - u_n\|^2 + \frac{m\mu_n}{2\sqrt{mM}\tau_n}\|x_n - x_{n-1}\|^2. \end{aligned}$$

Now, using the definition of x_{n+1}, we derive that, for all $n \in \mathbb{N}$, $i_n \in I_\varepsilon(x_n)$, and $x \in S$,

$$\begin{aligned} &f(x_{n+1}) - \theta_n g(x_{n+1}) + \frac{1}{2}\left(\frac{1-\sqrt{\beta}\zeta}{\tau_n} - \frac{M\mu_n}{\sqrt{mM}\tau_n}\right)\|x_{n+1} - x_n\|^2 \\ &\leq f(w_n^{i_n}) - \theta_n g(w_n^{i_n}) + \frac{1}{2}\left(\frac{1-\sqrt{\beta}\zeta}{\tau_n} - \frac{M\mu_n}{\sqrt{mM}\tau_n}\right)\|w_n^{i_n} - x_n\|^2 \\ &\leq f(x) - \theta_n g_{i_n}(x_n) + \frac{1}{2\tau_n}\|x - v_n\|^2 - \frac{1}{2\tau_n}\|x_n - v_n\|^2 \\ &\quad - \theta_n\langle \nabla g_{i_n}(x_n), x - x_n\rangle + \frac{\ell}{2}\|x - u_n\|^2 + \frac{m\mu_n}{2\sqrt{mM}\tau_n}\|x_n - x_{n-1}\|^2. \end{aligned} \tag{32}$$

Let $i_n \in I_0(x_n) \subseteq I_\varepsilon(x_n)$. Then $g_{i_n}(x_n) = g(x_n)$. Since $f(x_n) = \theta_n g(x_n)$ and $x_n - u_n = -\kappa_n(x_n - x_{n-1})$, letting $x = x_n$ in (32) yields

$$f(x_{n+1}) - \theta_n g(x_{n+1}) + \frac{1}{2}\left(\frac{1-\sqrt{\beta}\zeta}{\tau_n} - \frac{M\mu_n}{\sqrt{mM}\tau_n}\right)\|x_{n+1} - x_n\|^2 \leq \frac{1}{2}\left(\ell\kappa_n^2 + \frac{m\mu_n}{\sqrt{mM}\tau_n}\right)\|x_n - x_{n-1}\|^2.$$

Dividing by $g(x_{n+1}) > 0$ on both sides and recalling that $m \leq g(x_{n+1}) \leq M$, $\kappa_n \leq \overline{\kappa}$, $\mu_n \leq \overline{\mu}\tau_n$, and $1/\tau_n \geq \delta$, we have that

$$\frac{f(x_{n+1})}{g(x_{n+1})} + \left(\frac{\delta(1-\sqrt{\beta}\zeta)}{2M} - \frac{\overline{\mu}}{2\sqrt{mM}}\right)\|x_{n+1} - x_n\|^2 \leq \frac{f(x_n)}{g(x_n)} + \left(\frac{\ell\overline{\kappa}^2}{2m} + \frac{\overline{\mu}}{2\sqrt{mM}}\right)\|x_n - x_{n-1}\|^2.$$

Proceeding as in the proof of Theorem 4.7(i)&(ii), we obtain conclusions (i) and (ii) of this theorem.

(iii): In view of (i), we set

$$\overline{\theta} := \lim_{n\to+\infty}\theta_n = \lim_{n\to+\infty}\frac{f(x_n)}{g(x_n)}.$$

Let $\overline{x}$ be a cluster point of $(x_n)_{n\in\mathbb{N}}$ and let $(x_{k_n})_{n\in\mathbb{N}}$ be a subsequence convergent to $\overline{x}$. Then $\overline{x} \in S$ as well as $x_{k_n+1} \to \overline{x}$, $u_{k_n} \to \overline{x}$, and $v_{k_n} \to \overline{x}$ due to (ii). By the continuity of each g_i, there exists $n_0 \in \mathbb{N}$ such that, for all $i \in \{1,\ldots,p\}$ and all $n \geq n_0$, $g_i(x_{k_n}) \geq g_i(\overline{x}) - \varepsilon/2$ and $g(\overline{x}) \geq g(x_{k_n}) - \varepsilon/2$. It follows that, for all $n \geq n_0$, $I_0(\overline{x}) \subseteq I_\varepsilon(x_{k_n})$.

Let $n \geq n_0$ and let $i \in I_0(\overline{x}) \subseteq I_\varepsilon(x_{k_n})$. We have from (32) that, for all $x \in S$,

$$\begin{aligned} &f(x_{k_n+1}) - \theta_{k_n} g(x_{k_n+1}) + \frac{1}{2}\left(\frac{1-\sqrt{\beta}\zeta}{\tau_{k_n}} - \frac{M\mu_{k_n}}{\sqrt{mM}\tau_{k_n}}\right)\|x_{k_n+1} - x_{k_n}\|^2 \\ &\leq f(x) - \theta_{k_n} g_i(x_{k_n}) + \frac{1}{2\tau_{k_n}}\|x - v_{k_n}\|^2 - \frac{1}{2\tau_{k_n}}\|x_{k_n} - v_{k_n}\|^2 \\ &\quad - \theta_{k_n}\langle \nabla g_i(x_{k_n}), x - x_{k_n}\rangle + \frac{\ell}{2}\|x - u_{k_n}\|^2 + \frac{m\mu_{k_n}}{2\sqrt{mM}\tau_{k_n}}\|x_{k_n} - x_{k_n-1}\|^2. \end{aligned} \tag{33}$$

It follows from the continuity of g, g_i, and ∇g_i that $g(x_{k_n+1}) \to g(\overline{x})$, $g_i(x_{k_n}) \to g_i(\overline{x}) = g(\overline{x})$ (as $i \in I_0(\overline{x})$), and $\nabla g_i(x_{k_n}) \to \nabla g_i(\overline{x})$. Letting $x = \overline{x}$ and $n \to +\infty$ in (33) and noting that $\overline{\tau} = \liminf_{k\to\infty} \tau_n > 0$, we have $\limsup_{n\to+\infty} f(x_{k_n+1}) \leq f(\overline{x})$. Combining with the lower semicontinuity of f gives $f(x_{k_n+1}) \to f(\overline{x})$ as $n \to +\infty$. Thus, $\theta_{k_n} \to \overline{\theta} = \frac{f(\overline{x})}{g(\overline{x})}$ as $n \to +\infty$.

Now, letting $n \to +\infty$ in (33), we obtain that, for all $x \in S$,

$$f(\overline{x}) \leq f(x) + \left(\frac{1}{2\overline{\tau}} + \frac{\ell}{2}\right)\|x - \overline{x}\|^2 + \frac{f(\overline{x})}{g(\overline{x})}\langle \nabla g_i(\overline{x}), \overline{x} - x\rangle.$$

This shows that $\overline{x}$ minimizes the function φ over S, where

$$\varphi(x) := f(x) + \left(\frac{1}{2\overline{\tau}} + \frac{\ell}{2}\right)\|x - \overline{x}\|^2 - \frac{f(\overline{x})}{g(\overline{x})}\langle \nabla g_i(\overline{x}), x\rangle$$

In particular, one sees that, for all $i \in I_0(\overline{x})$, $\frac{f(\overline{x})}{g(\overline{x})}\nabla g_i(\overline{x}) \in \partial_L(f + \iota_S)(\overline{x})$. So, $\overline{x} \in S \cap \operatorname{dom} f$ and

$$\bigcup_{i\in I_0(\overline{x})} \frac{f(\overline{x})}{g(\overline{x})}\nabla g_i(\overline{x}) \subseteq \partial_L(f + \iota_S)(\overline{x}). \tag{34}$$

By taking convex hull on both sides, we see that

$$\frac{f(\overline{x})}{g(\overline{x})}\partial_L g(\overline{x}) = \operatorname{conv}\bigcup_{i\in I_0(\overline{x})} \frac{f(\overline{x})}{g(\overline{x})}\nabla g_i(\overline{x}) \subseteq \operatorname{conv}\partial_L(f + \iota_S)(\overline{x}).$$

If f is weakly convex on S, Lemma 2.2(i) implies that $\partial(f + \iota_S)(\overline{x})$ is convex. Thus, the conclusion follows. □

Remark 6.3 (Absence of the boundedness condition). As with Algorithm 4.1 and Theorem 4.7, in the case where (BC) fails, if we set $\overline{\mu} = \overline{\kappa} = 0$ in Step 1 and let

$$\hat{i}_n \in \operatorname*{arg\,min}_{i_n \in I_\varepsilon(x_n)} \left(f(w_n^{i_n}) - \theta_n g(w_n^{i_n}) + \frac{1-\sqrt{\beta}\zeta}{2\tau_n}\|w_n^{i_n} - x_n\|^2\right)$$

in Step 3 of Algorithm 6.1, then Theorem 6.2 still holds with $F_n = \frac{f(x_n)}{g(x_n)}$.

Remark 6.4 (Discussion of the results). (i) Firstly, a close inspection of the proof and noting that, for all $\eta < \varepsilon$, one has for all large n, $I_\eta(\overline{x}) \subseteq I_\varepsilon(x_{k_n})$. So, (28) in the conclusion of Theorem 6.2(iii) indeed can be strengthened as: for all $\eta < \varepsilon$,

$$\frac{f(\overline{x})}{g(\overline{x})}\bigcup_{i\in I_\eta(\overline{x})} \nabla g_i(\overline{x}) \subseteq \partial_L(f + \iota_S)(\overline{x}).$$

(ii) Secondly, following the same method of proof used in Theorem 5.5, one can establish the global convergence of Algorithm 6.1 under the KL assumptions in Theorem 5.5 and also the additional assumption that $I_0(\overline{x}) = \{i \in \{1, \ldots, p\} : g_i(\overline{x}) = g(\overline{x})\}$ is a singleton for all $\overline{x} \in \Omega$, where Ω is the set of cluster points of $(x_n)_{n\in\mathbb{N}}$. Another sufficient condition ensuring the global convergence would be that any point $\overline{x} \in \Omega$ is isolated. For brevity purpose, we omit the proof here. Unfortunately, these conditions are rather restrictive for the setting of Algorithm 6.1. It would be interesting to see how one can obtain further weaker conditions ensuring the global convergence of Algorithm 6.1. This would be an interesting open question and will be examined later.

7. Numerical examples In the section, we illustrate our proposed algorithms via numerical examples. We first start with an explicit analytic example and use it to demonstrate the different behavior of Algorithm 4.1 and Algorithm 6.1 as well as the effect of the extrapolations. Then, we examine the performance of the algorithm for the scale invariant sparse signal reconstruction model. All the numerical tests were conducted on a computer with a 2.8 GHz Intel Core i7 and 8 GB RAM, equipped with MATLAB R2015a.

7.1. An analytical example Consider the analytical example discussed in Example 3.3

$$\min_{x\in[-1,1]} \frac{x^2+1}{|x|+1}. \tag{EP$_1$}$$

In this case, $g(x)=|x|+1$ is convex, and so, $\beta=0$. Also, for all $x\in[-1,1]$, $m\le g(x)\le M$, where $m=1$ and $M=2$. The numerator $f(x)=f^{\mathfrak{s}}(x)=x^2+1$ is a convex and differentiable function whose gradient is Lipschitz continuous with modulus $\ell=2$.

Algorithm 4.1 vs. Algorithm 6.1. Let $\delta=\frac{\ell M}{m}=4$ and $\tau_n=\frac{1}{\delta}=\frac{1}{4}$ for all n. Set $\overline{\mu}=0$ and let $\overline{\kappa}\in(0,1)$ and $\kappa_n\in[0,\overline{\kappa}]$. We now compare the behavior of Algorithm 4.1 and Algorithm 6.1 for (EP$_1$):

Firstly, it can be directly verified that $g_n=\operatorname{sign}(x_n)\in\partial g(x_n)$ and that $f^{\mathfrak{s}}(u_n)+\langle\nabla f^{\mathfrak{s}}(u_n),x-u_n\rangle+\frac{\ell}{2}\|x-u_n\|^2=x^2+1$. In this case, Algorithm 4.1 reduces to

$$x_{n+1}=\mathrm{P}_{[-1,1]}\left(\frac{2}{3}\left[x_n+\frac{1}{4}\frac{x_n^2+1}{|x_n|+1}\operatorname{sign}(x_n)\right]\right).$$

Here, $P_{[-1,1]}$ denotes the Euclidean projection onto the set $[-1,1]$. If one chooses as initial point $x_0=0$, then $x_n=0$ for all n, and so, $(x_n)_{n\in\mathbb{N}}$ converges to a lifted stationary point (but not a strong lifted stationary point).

If one chooses as initial point $x_0>0$, then, by induction, it is easy to see that $x_n>0$ and so, $x_n\in(0,1]$. This implies that

$$x_{n+1}=\mathrm{P}_{[-1,1]}\left(\frac{2}{3}\left[x_n+\frac{x_n^2+1}{4(x_n+1)}\right]\right)=\frac{2}{3}\left[x_n+\frac{x_n^2+1}{4(x_n+1)}\right],$$

where the last equality is from the fact that $x_n+\frac{x_n^2+1}{4(x_n+1)}\in[0,\frac{3}{2}]$ for all $x_n\in(0,1]$. Thus, $x_n\to\sqrt{2}-1$ which is a lifted stationary point.

Similarly, if one chooses as initial point $x_0<0$, then, $x_n\to1-\sqrt{2}$ which is also a lifted stationary point.

Next, we analyze the behavior of Algorithm 6.1. Recall that $\delta=\frac{\ell M}{m}=4$, $\tau_n=\frac{1}{\delta}=\frac{1}{4}$, $\overline{\mu}=0$, $\kappa_n\in[0,\overline{\kappa}]$ with $\overline{\kappa}\in(0,1)$. Let $\varepsilon=2$. Note that $g(x)=\max\{x+1,-x+1\}$. Then $I_\varepsilon(x_n)=\{1,2\}$, and so,

$$w_n^1=\mathrm{P}_{[-1,1]}\left(\frac{2}{3}\left[x_n+\frac{1}{4}\frac{x_n^2+1}{|x_n|+1}\right]\right)\ \text{and}\ w_n^2=\mathrm{P}_{[-1,1]}\left(\frac{2}{3}\left[x_n-\frac{1}{4}\frac{x_n^2+1}{|x_n|+1}\right]\right).$$

In Algorithm 6.1, we set $x_{n+1}:=w_n^{\hat{i}_n}$, where

$$\hat{i}_n\in\operatorname*{arg\,min}_{i\in\{1,2\}}\left((w_n^i)^2+1-\frac{x_n^2+1}{|x_n|+1}(|w_n^i|+1)+2(w_n^i-x_n)^2\right).$$

For the proceeding step for updating x_{n+1}, if the values happens to be the same in the above argmin operations, we choose $\hat{i}_n$ to be the smallest index. By randomly generating the initial guess x_0, we observe that Algorithm 6.1 generates a sequence $(x_n)_{n\in\mathbb{N}}$ such that $x_n\to\sqrt{2}-1$ if $x_0\ge0$ and $x_n\to1-\sqrt{2}$ if $x_0<0$. Figure 1 depicts the trajectory x_n of Algorithm 6.1 with three initial points:

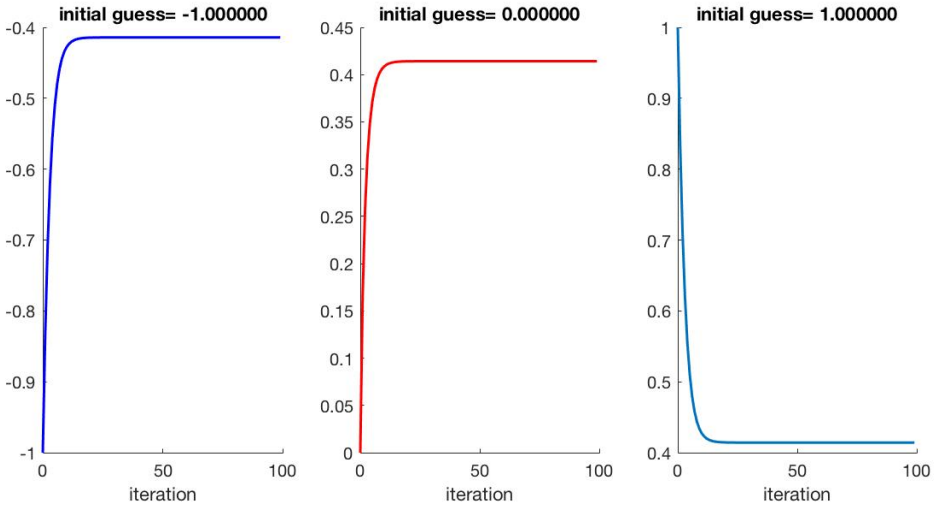

FIGURE 1. Trajectory of Algorithm 6.1 with different initial guess x_0 for (EP_1)

$x_0 = 0, -1, 1$. Interestingly, we note that, in the case where $x_0 = 0$, Algorithm 6.1 converges to a strong lifted stationary point $\sqrt{2}-1$ while Algorithm 4.1 converges to a lifted stationary point 0, which is not a strong lifted stationary point.

Effect of the extrapolation parameter. We now illustrate the behavior of Algorithm 4.1 by varying the extrapolation parameters. To do this, again let $\delta = \frac{\ell M}{m} = 4$, $\tau_n = \frac{1}{\delta} = \frac{1}{4}$, and $g_n = \text{sign}(x_n) \in \partial_L g(x_n)$ for all n. Set $\alpha \in [0,1)$, $\overline{\mu} = \frac{\alpha\delta\sqrt{mM}}{2M} = \sqrt{2}\alpha$, and $\overline{\kappa} \in \left[0, \sqrt{1-\alpha}\right)$. Let any $\kappa_n \in [0, \overline{\kappa}]$ and $\mu_n = \overline{\mu}\tau_n \frac{\nu_{n-1}-1}{\nu_n} = \frac{\sqrt{2}}{4}\alpha\frac{\nu_{n-1}-1}{\nu_n}$, where

$$\nu_{-1} = \nu_0 = 1 \quad \text{and} \quad \nu_{n+1} = \frac{1+\sqrt{1+4\nu_n^2}}{2},$$

and reset $\nu_{n-1} = \nu_n = 1$ when $n = n_0, 2n_0, 3n_0, \ldots$ for the integer $n_0 = 50$. In this case, direct verification shows that $\sup_n \nu_n \le 1$, and hence $\mu_n \le \frac{\sqrt{2}}{4}\alpha = \overline{\mu}\tau_n$. Starting with the initialization $x_0 = 1$, we then run Algorithm 4.1 with different $\alpha \in [0,1)$. Figure 2 depicts the distance, in the log scale, between the iterates x_n and the solution $x^* = \sqrt{2}-1$ for $\alpha \in \{0, 0.5, 0.7, 0.99\}$, where the case $\alpha = 0$ indeed corresponds to the un-extrapolated cases. As one can see from Figure 2, as α increases and approaches 1, the algorithm tends to converge faster. Moreover, we note that, from our derivation, we require $\alpha < 1$ to ensure the convergence of the algorithm. On the other hand, when α increases and approaches one, the algorithm exhibits some oscillation phenomenon.

7.2. Scale invariant sparse signal recovery problem As another illustration, we examine the following scale invariant sparse signal recovery problem discussed in the motivating example

$$\min_{x\in\mathbb{R}^N} \frac{\|x\|_1}{\|x\|_2} \quad \text{s.t.} \quad Ax = b,\ \text{lb}_i \le x_i \le \text{ub}_i,\ i = 1, \ldots, N, \tag{EP$_2$}$$

where lb_i and ub_i are the lower bound and upper bound for the variables x_i, $i = 1, \ldots, N$. We follow [28] and generate the matrix A via the so-called oversampled discrete cosine transform (DCT), that is, $A = [a_1, a_2, \ldots, a_N] \in \mathbb{R}^{P\times N}$ where

$$a_j = \frac{1}{\sqrt{P}} \cos\left(\frac{2\pi w\, j}{F}\right),\ j = 1, \ldots, N.$$

where w is a random vector uniformly distributed in $[0,1]^P$ and F is a positive number which gives a measure on how coherent the matrix is. The ground truth $x^g \in \mathbb{R}^N$ is simulated as an s-sparse

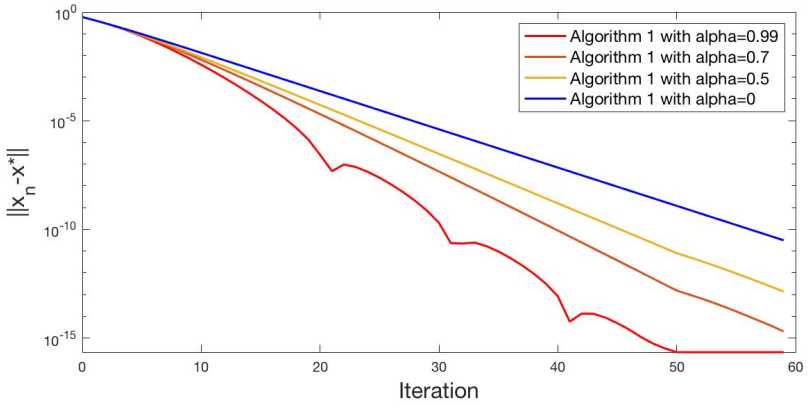

FIGURE 2. Distance to the solution vs iterations in solving (EP_1)

signal where s is the total number of nonzero entries. The support of x^g is a random index set, and the values of nonzero elements follow a Gaussian normal distribution. Then the ground-truth is normalized to have maximum magnitude as 1 so that we can examine the performance within the $[-1,1]^N$ box constraint. Then, we generate $b = Ax^g$, and set $\mathrm{lb}_i = -1$ and $\mathrm{ub}_i = 1$. Specifically, in our experiment, following [28], we consider the above matrix A of size $(P,N) = (64,1024)$, $F = 10$ and the ground-truth sparse vector has 12 nonzero elements.

We use two methods for solving this scale invariant sparse signal recovery problem: our proposed extrapolated proximal subgradient method (e-PSG) and the alternating direction of method of multipliers (ADMM) proposed in [28]. It was shown in [28] that the ADMM method works very efficiently although the theoretical justification of the convergence of this method is still lacking.

- ADMM method: We first solve the L_1-optimization problem which results when replacing the objective of (EP_2) by $\|x\|_1 := \sum_{i=1}^N |x_i|$. This is done by using the commercial software Gurobi and produces a solution x_0 for the L_1-optimization problem. Following [28], we use x_0 as an initialization and use the ADMM method proposed therein. We terminate the algorithm when the relative error $\frac{\|x_{n+1}-x_n\|}{\max\{\|x_n\|,1\}}$ is smaller than 10^{-9}.
- Algorithm 4.1 (e-PSG method): Similar to the ADMM method, we also use the solution of the L_1-optimization problem as the initial point. We choose $f^s \equiv 0$ (and so, $\ell = 0$), $\kappa_n = 0$. As $g(x) = \|x\|_2$ is convex, $\beta = 0$. Moreover, for all x feasible for (EP_2), $m \le g(x) \le M$ where $M = \sqrt{N}$ and m is a positive number computed as the Euclidean norm of the least norm solution of $Ax = b$ via the Matlab code `m = norm(pinv(A)*b)`. Let $\alpha = 0.99$ and set $\mu_n = \frac{\alpha\sqrt{mM}}{2M}\frac{\nu_{n-1}-1}{\nu_n}$, where

$$\nu_{-1} = \nu_0 = 1 \quad \text{and} \quad \nu_{n+1} = \frac{1+\sqrt{1+4\nu_n^2}}{2},$$

  and reset $\nu_{n-1} = \nu_n = 1$ when $n = n_0, 2n_0, 3n_0, \ldots$ for the integer $n_0 = 50$. For $\delta = \frac{\ell M}{m} > 0$, let $\tau_n = \frac{1}{\delta}$ and $\overline{\mu} = \frac{\alpha\delta\sqrt{mM}}{2M} < \frac{\delta\sqrt{mM}}{2M}$. It can be verified that $\mu_n \le \frac{\alpha\sqrt{mM}}{2M} = \overline{\mu}\tau_n$, and so, the requirements of the parameters in Algorithm 4.1 are satisfied. We use the same termination criterion as for the ADMM method. For the subproblem arising in Step 2 of Algorithm 4.1, we reformulate the problem as an equivalent quadratic program with linear constraints, and solve it using the software Gurobi.

We run the ADMM and the e-PSG method (Algorithm 4.1) for 50 trials. The following table summarizes the output of the two methods by listing the average number of

- sparsity level of the initial guess: the number of entries of the initialization (the solution for L_1-optimization problem) with value larger than 10^{-6};
- sparsity level of the solution: the number of entries of the computed solution with value larger than 10^{-6};

- error with respect to the ground truth: the Euclidean norm of the difference of the computed solution and the ground truth x^g;
- the objective value of the computed solution;
- CPU time measured in seconds.

From Table 1, one can see that e-PSG method is competitive with the ADMM method in terms of sparsity level and the CPU time used, and produces a solution with slightly better quality in terms of the final objective value and the error with respect to the ground truth. As plotted in Figure 3, one can see that ADMM uses around 2000 iterations to reach the desired relative error tolerance, and has sharp oscillating phenomenon in terms of the objective value (this has also been observed in [28], and the authors of [28] believed that this is one of the major obstacles in establishing the convergence of the ADMM method); while the proposed e-PSG method quickly approaches the desired error tolerance. On the other hand, it should be noted that the subproblems in the ADMM method have closed form solutions while the subproblems in the e-PSG method are reformulated as quadratic programming problems with linear constraints and solved via the software Gurobi[5].

TABLE 1. Computation results for (EP_2)

	sparsity level		error w.r.t the ground truth	objective value of the computed solution	CPU time
	initial guess	computed solution			
ADMM	64	12	6.948329e-06	2.724348	1.970365
e-PSG	64	12	4.539185e-10	2.724326	2.375557

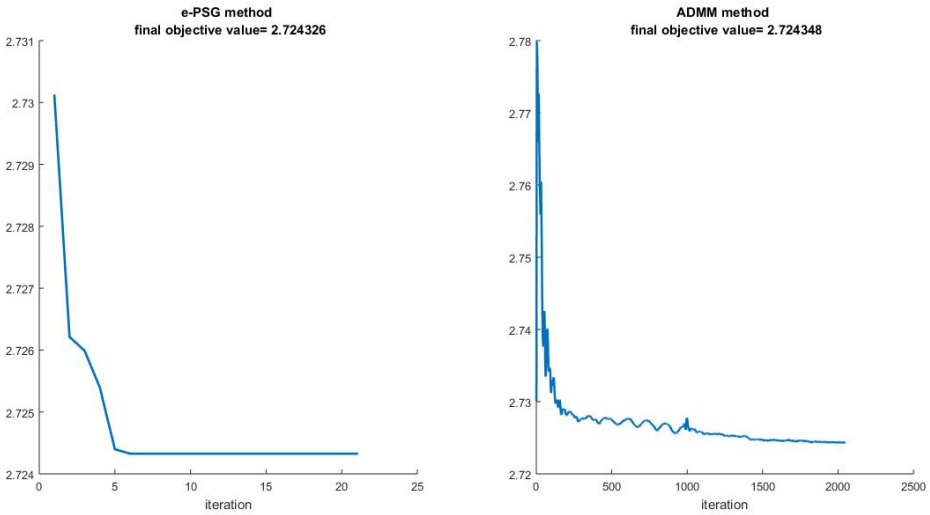

FIGURE 3. Objective values vs. iterations in solving (EP_2)

[5] One possible way to improve the CPU time in using e-PSG is to solve the subproblem via alternating direction method of multiplier method directly. We leave this as a future study.

8. Conclusions We have proposed proximal subgradient algorithms with extrapolations for solving fractional optimization model where both the numerator and denominator can be nonsmooth and nonconvex. We have shown that the sequence of iterates generated by the algorithm is bounded and any of its limit points is a stationary point of the model problem. We have also established the global convergence of the sequence by further assuming the KL property for a suitable merit function by providing a unified analysis framework of descent methods. Finally, in the case where the denominator is the maximum of finitely many continuously differentiable weakly convex functions, we have also proposed an enhanced proximal subgradient algorithm with extrapolations, and showed that this enhanced algorithm converges to a stronger notion of stationary points of the model problem.

Our results in this paper point out the following interesting open questions and future work: (1) For the enhanced proximal subgradient algorithm with extrapolations (Algorithm 6.1), is it possible to extend the case from $g(x) = \max_{1\leq i\leq p}\{g_i(x)\}$ to $g(x) = \max_{t\in T}\{g_t(x)\}$ where T is a (possibly) infinite set? (2) In Algorithm 6.1, as one needs to solve the subproblem $|I_\varepsilon(x_n)|$ times, this can be time consuming when the dimension is high. Is it possible to incorporate any randomize technique to save the computational cost and establish the convergence in probability? (3) How to obtain the global convergence of the full sequence of Algorithm 6.1 under weaker and reasonable assumptions is also an important topic to be examined. (4) In our model problem (P), we assume that the numerator f can be written as the sum of $f^{\mathfrak{s}}$ and $f^{\mathfrak{n}}$, where $f^{\mathfrak{s}}$ is a differentiable convex function whose gradient is Lipschitz continuous and $f^{\mathfrak{n}}$ is a nonconvex function. It would be interesting to see how one could develop algorithms which allow the smooth part $f^{\mathfrak{s}}$ being possibly nonconvex as well. (5) Finally, further numerical implementations of our algorithms and comparisons with other competitive methods are left as future research.

Acknowledgments The authors would like to thank Dr. Yifei Lou for kindly sharing the MATLAB code for the ADMM method used in [28]. The authors are also grateful to the Associate Editor and the referees for their constructive comments and suggestions. R.I. Boţ was partially supported by the Austrian Science Fund (FWF) under project I 2419-N32. M.N. Dao and G. Li were partially supported by the Australian Research Council (ARC) under project DP190100555.

References

[1] Aragón Artacho FJ, Campoy R, Vuong PT (2020) Using positive spanning sets to achieve d-stationarity with the boosted DC algorithm. *Vietnam J. Math.* 48(2):363–376.

[2] Attouch H, Bolte J (2009) On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Math. Program. Ser. B* 116(1–2):5–16.

[3] Attouch H, Bolte J, Svaiter BF (2013) Convergence of descent methods for semi-algebraic and tame problems: Proximal algorithms, forward-backward splitting, and regularized Gauss–Seidel methods. *Math. Program. Ser. A* 137(1–2):91–129.

[4] Bauschke HH, Combettes PL (2017) *Convex Analysis and Monotone Operator Theory in Hilbert Spaces* (Cham: Springer), 2nd edition.

[5] Beck A (2017) *First-Order Methods in Optimization* (Philadelphia: Society for Industrial and Applied Mathematics (SIAM)).

[6] Beck A, Hallak N (2020) On the convergence to stationary points of deterministic and randomized feasible descent directions methods. *SIAM J. Optim.* 30(1):56–79.

[7] Boţ RI, Csetnek ER (2017) Proximal-gradient algorithms for fractional programming. *Optimization* 66(8):1383–1396.

[8] Bolte J, Daniilidis A, Lewis A, Shiota M (2007) Clarke subgradients of stratifiable functions. *SIAM J. Optim.* 18(2):556–572.

[9] Bolte J, Pauwels E (2016) Majorization-minimization procedures and convergence of sqp methods for semi-algebraic and tame programs. *Math. Oper. Res.* 41(2):442–465.

[10] Bolte J, Sabach S, Teboulle M (2014) Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Math. Program. Ser. A* 146(1–2):459–494.

[11] Chen L, He S, Zhang SZ (2011) When all risk-adjusted performance measures are the same: in praise of the Sharpe ratio. *Quant. Finance* 11(10):1439–1447.

[12] Crouzeix JP, Ferland JA, Schaible S (1985) An algorithm for generalized fractional programs. *J. Optim. Theory Appl.* 47(1):35–49.

[13] Dinkelbach W (1967) On nonlinear fractional programming. *Management Sci.* 13:492–498.

[14] Frankel P, Garrigos G, Peypouquet J (2014) Splitting methods with variable metric for Kurdyka–łojasiewicz functions and general convergence rates. *J. Optim. Theory Appl.* 165(3):874–900.

[15] Ibaraki T (1981) Solving mathematical programming problems with fractional objective functions. Schaible S, Ziemba W, eds., *Generalized Concavity in Optimization and Economics*, 441–472 (New York-London: Academic Press).

[16] Ibaraki T (1983) Parametric approaches to fractional programs. *Math. Program.* 26(3):345–362.

[17] Kruger AY (2003) On Fréchet subdifferentials. *J. Math. Sci.* 116:3325–3358.

[18] Kurdyka K (1998) On gradients of functions definable in o-minimal structures. *Ann. Inst. Fourier (Grenoble)* 48(3):769–783.

[19] Li G, Pong TK (2018) Calculus of the exponent of Kurdyka–łojasiewicz inequality and its applications to linear convergence of first-order methods. *Found. Comput. Math.* 18(5):1199–1232.

[20] Lojasiewicz S (1963) Une propriété topologique des sous-ensembles analytiques réels. *Les Équations aux Dérivées Partielles*, 87–89 (Paris: Éditions du Centre National de la Recherche Scientifique (CNRS)).

[21] Mordukhovich BS (2006) *Variational Analysis and Generalized Differentiation I. Basic Theory* (Berlin: Springer).

[22] Mordukhovich BS, Nam NM, Yen ND (2006) Fréchet subdifferential calculus and optimality conditions in nondifferentiable programming. *Optimization* 55(5–6):685–708.

[23] Nesterov Y (2004) *Introductory Lectures on Convex Optimization: A Basic Course* (Boston: Kluwer Academic).

[24] Noll D (2013) Convergence of non-smooth descent methods using the Kurdyka–łojasiewicz inequality. *J. Optim. Theory Appl.* 160(2):553–572.

[25] Ochs P (2019) Unifying abstract inexact convergence theorems and block coordinate variable metric iPiano. *SIAM J. Optim.* 29(1):541–570.

[26] Ochs P, Chen Y, Brox T, Pock T (2014) iPiano: Inertial proximal algorithm for nonconvex optimization. *SIAM J. Imaging Sci.* 7(2):1388–1419.

[27] Pang JS, Razaviyayn M, Alvarado A (2016) Computing B-stationary points of nonsmooth DC programs. *Math. Oper. Res.* 42(1):95–118.

[28] Rahimi Y, Wang C, Dong H, Lou Y (2019) A scale-invariant approach for sparse signal recovery. *SIAM J. Sci. Comput.* 41(6):3649–3672.

[29] Rockafellar RT, Wets RJB (1998) *Variational Analysis* (Berlin: Springer).

[30] Schaible S (1976) Fractional programming II. On Dinkelbach's algorithm. *Management Sci.* 22(8):868–873.

[31] Zeng L, Yu P, Pong TK (2020) Analysis and algorithms for some compressed sensing models based on L1/L2 minimization, arXiv:2007.12821, to appear in SIAM J. Optim., 2021.