Journal of Global Optimization (2021) 79:853–883
https://doi.org/10.1007/s10898-020-00955-3

Linear convergence of inexact descent method and inexact proximal gradient algorithms for lower-order regularization problems

Yaohua Hu[1] · Chong Li[2] · Kaiwen Meng[3] · Xiaoqi Yang[4]

Received: 22 October 2019 / Accepted: 27 September 2020 / Published online: 5 October 2020

Abstract
The ℓ_p regularization problem with $0 < p < 1$ has been widely studied for finding sparse solutions of linear inverse problems and gained successful applications in various mathematics and applied science fields. The proximal gradient algorithm is one of the most popular algorithms for solving the ℓ_p regularisation problem. In the present paper, we investigate the linear convergence issue of one inexact descent method and two inexact proximal gradient algorithms (PGA). For this purpose, an optimality condition theorem is explored to provide the equivalences among a local minimum, second-order optimality condition and second-order growth property of the ℓ_p regularization problem. By virtue of the second-order optimality condition and second-order growth property, we establish the linear convergence properties of the inexact descent method and inexact PGAs under some simple assumptions. Both linear convergence to a local minimal value and linear convergence to a local minimum are provided. Finally, the linear convergence results of these methods are extended to the infinite-dimensional Hilbert spaces. Our results cannot be established under the framework of Kurdyka–Łojasiewicz theory.

✉ Kaiwen Meng
mengkw@swufe.edu.cn

Yaohua Hu
mayhhu@szu.edu.cn

Chong Li
cli@zju.edu.cn

Xiaoqi Yang
mayangxq@polyu.edu.hk

[1] Shenzhen Key Laboratory of Advanced Machine Learning and Applications, College of Mathematics and Statistics, Shenzhen University, Shenzhen 518060, People's Republic of China

[2] School of Mathematical Sciences, Zhejiang University, Hangzhou 310027, People's Republic of China

[3] School of Economics and Mathematics, Southwestern University of Finance and Economics, Chengdu 611130, People's Republic of China

[4] Department of Applied Mathematics, The Hong Kong Polytechnic University, Kowloon, Hong Kong

Keywords Sparse optimization · Nonconvex regularization · Inexact approach · Descent methods · Proximal gradient algorithms · Linear convergence

1 Introduction

The following linear inverse problem is at the core of many problems in various areas of mathematics and applied sciences: finding $x \in \mathbb{R}^n$ such that

$$Ax = b,$$

where $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$ are known, and an unknown noise is included in b. If $m \ll n$, the above linear inverse problem is seriously ill-conditioned and has infinitely many solutions, and researchers are interested in finding solutions with certain structures, e.g., the sparsity structure. A popular technique for approaching a sparse solution of the linear inverse problem is to solve the ℓ_1 regularization problem

$$\min_{x\in\mathbb{R}^n} \ \|Ax - b\|^2 + \lambda\|x\|_1,$$

where $\|\cdot\|$ denotes the Euclidean norm, $\|x\|_1 := \sum_{i=1}^n |x_i|$ is a sparsity promoting norm, and $\lambda > 0$ is a regularization parameter providing a tradeoff between accuracy and sparsity. In the past decade, the ℓ_1 regularization problem has been extensively investigated (see, e.g., [4,18,19,37,53,56]) and gained successful applications in a wide range of fields, such as compressive sensing [13,20], image science [4,21], systems biology [46,50] and machine learning [3,35].

However, in recent years, it has been revealed by extensive empirical studies that the solutions obtained from the ℓ_1 regularization may be much less sparse than the true sparse solution, and that the ℓ_1 regularization cannot recover a signal or an image with the least measurements when applied to compressive sensing; see, e.g., [15,55,61]. To overcome these drawbacks, the following ℓ_p regularization problem ($0 < p < 1$) was introduced in [15,55] to improve the performance of sparsity recovery:

$$\min_{x\in\mathbb{R}^n} \ \|Ax - b\|^2 + \lambda\|x\|_p^p, \tag{1}$$

where $\|x\|_p := \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}$ is the ℓ_p quasi-norm. It was shown in [15] that the ℓ_p regularization requires a weaker restricted isometry property to guarantee perfect sparsity recovery and allows to obtain a more sparse solution from fewer linear measurements than that required by the ℓ_1 regularization; and it was illustrated in [24,55] that the ℓ_p regularization has a significantly stronger capability in obtaining a sparse solution than the ℓ_1 regularization. Benefitting from these advantages, the ℓ_p regularization technique has been applied in many fields; see [24,33,36,40,41] and references therein. It is worth noting that the ℓ_p regularization problem (1) is a variant of lower-order penalty problems, investigated in [11,26,34], for a constrained optimization problem. The main advantage of the lower-order penalty functions over the classical ℓ_1 penalty function in the context of constrained optimization is that they require weaker conditions to guarantee an exact penalization property and that their least exact penalty parameter is smaller.

Motivated by these significant advantages and successful applications of the ℓ_p regularization, tremendous efforts have been devoted to the study of optimization algorithms for the ℓ_p regularization problem. Many practical algorithms have been investigated for solving problem (1), such as an interior-point potential reduction algorithm [23], smoothing methods

[16,17], splitting methods [29,30] and iterative reweighted minimization methods [27,31]. In particular, Xu et al. [55] proposed an iterative half thresholding algorithm, which is efficient in signal recovery and image deconvolution. In the present paper, we are particularly interested in the proximal gradient algorithm (in short, PGA) for solving problem (1).

ALGORITHM PGA. Given an initial point $x^0 \in \mathbb{R}^n$ and a sequence of stepsizes $\{v_k\} \subseteq \mathbb{R}_+$. For each $k \in \mathbb{N}$, having x^k, we determine x^{k+1} as follows:

$$\begin{aligned} z^k &:= x^k - 2v_k A^\top (Ax^k - b), \\ x^{k+1} &\in \arg\min_{x\in\mathbb{R}^n} \left\{ \lambda \|x\|_p^p + \frac{1}{2v_k}\|x - z^k\|^2 \right\}. \end{aligned} \tag{2}$$

The practicability of the PGA is an important issue when solving the nonconvex ℓ_p regularization problem (1). It is worth noting that the main computation of the PGA is the calculation of the proximity operator of the ℓ_p regularizer in (2). In particular, the analytical solutions of the proximity operator of the ℓ_p regularizer (2) when $p = 1$ (resp. 0, $\frac{1}{2}$, $\frac{2}{3}$) were provided in [19] (resp. [6,14,55]); see also [24, Proposition 18] for the group-wised ℓ_p regularizer. In such cases, the PGA is reduced to the iterative soft [19] (resp., hard [6], half [55]) thresholding algorithm when $p = 1$ (resp, 0, $\frac{1}{2}$) and the algorithm proposed in [14] when $p = \frac{2}{3}$ for solving the associated ℓ_p regularization problems. However, in the scenario of general $p \in (0, 1)$, the proximity operator of the ℓ_p regularizer may not have an analytic solution (see [24, Remark 21]), and it could be computationally expensive to solve subproblem (2) exactly at each iteration. Moreover, recall that the sparsity of x is defined by $\|x\|_0$ (i.e., the number of its nonzero components) and a well-known fact that $\lim_{p\to 0} \|x\|_p^p = \|x\|_0$ for each $x \in \mathbb{R}^n$. From this theoretical perspective, the ℓ_p regularization problem is close to the sparse recovery when p is close to 0; see [15,17,55]. It was also illustrated by the numerical studies in [24,55,57] that the PGA with $p \in (0, 1)$ outperforms the numerical algorithms for the ℓ_1 regularization problem on both accuracy and robustness. Hence, the excellent numerical performance of the PGA with $p \in (0, 1)$ but lacking the analytic formula with general $p \in (0, 1)$ motivates us to concern the convergence theory of the inexact PGA for the ℓ_p regularization problem (1).

The PGA is one of the most widely studied first-order iterative algorithms for solving regularization problems, and a special case of several iterative methods (see [1,2,8,42,49]) for solving the composite minimization problem

$$\min_{x\in\mathbb{R}^n} F(x) := H(x) + \Phi(x), \tag{3}$$

where $H : \mathbb{R}^n \to \overline{\mathbb{R}} := \mathbb{R} \cup \{+\infty\}$ is smooth and convex, and $\Phi : \mathbb{R}^n \to \overline{\mathbb{R}}$ is nonsmooth and possibly nonconvex. The convergence properties of these iterative methods have been explored under the framework of so-call Kurdyka–Łojasiewicz (in short, KL) theory. In particular, Attouch et al. [2] established the global convergence of abstract descent methods for minimizing a KL function $F : \mathbb{R}^n \to \overline{\mathbb{R}}$, in which the sequence $\{x_k\}$ satisfies the following hypotheses for two positive constants α and β:

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

$$F(x^{k+1}) - F(x^k) \le -\alpha \|x^{k+1} - x^k\|^2;$$

(H2) (*Relative error condition*). For each $k \in \mathbb{N}$, there exists $w^{k+1} \in \partial F(x^{k+1})$ such that

$$\|w^{k+1}\| \le \beta \|x^{k+1} - x^k\|;$$

(H3) (*Continuity condition*).[1] There exist a subsequence $\{x^{k_j}\}$ and a point x^* such that

$$\lim_{j\to\infty} x^{k_j} \to x^* \quad \text{and} \quad \lim_{j\to\infty} F(x^{k_j}) \to F(x^*).$$

The global convergence of Algorithm PGA follows from the convergence results of [2].

The study of convergence rates of optimization algorithms is an important issue of numerical optimization, and much attention has been paid to establish the convergence rates of relevant iterative algorithms for solving the structured optimization problem (3); see [1,7,25,29,38,48,49,52,54] and references therein. For example, the linear convergence of the PGA for solving the classical ℓ_1 (convex) regularization problem has been well investigated; see, e.g., [9,47,58–60] and references therein. Under the general framework of the KL (possibly nonconvex) functions, the linear convergence of several iterative algorithms for solving problem (3), including the PGA as a special case, have been established in [1,8,49,54] under the assumption that the KL exponent of the objective function is $\frac{1}{2}$. On the other hand, Zeng et al. [57] obtained the linear convergence of the PGA for problem (1) with an upper bound on p, which may be less than 1, and a lower bound on the stepsizes $\{v_k\}$ (Example 1 shows its restriction on parameters); Hu et al. [24] established the linear convergence of the PGA for the group-wised ℓ_p regularization problem under the assumption that the limiting point is a local minimum.

The inexact numerical algorithms have been widely used and extensively applied in engineering and application fields due to practical considerations of computational error and noise or the difficulty in solving the subproblems. Although some recent works studied the convergence properties of inexact numerical methods (see, e.g., [12,25,28]) for the convex composite optimization problem, to the best of our knowledge, there is few theoretical analysis on how the error in the calculation of the proximity operator affects the convergence rate of the inexact PGA for solving the ℓ_p regularization problem (1). Two relevant papers on the linear convergence study of the inexact PGA should be mentioned: (a) Schmidt et al. [45] proved the linear convergence of the inexact PGA for solving the convex composite optimization problem (3), in which H is strongly convex and Φ is convex; (b) Frankel et al. [22] provided a framework of establishing the linear convergence for descent methods satisfying (H1)–(H3), where (H2) is replaced by inexact form (H2°) (see Sect. 4). However, the convergence analysis in [22] was based on the descent property of the sequence of function values (H1) and the inexact version would be not convenient to implement for applications; see the explanation in Remark 6 below. Therefore, neither of the convergence analysis in [22,45] can be applied to establish the linear convergence of the inexact PGA for solving the ℓ_p regularization problem, in which the sequence of function values is not necessarily descent. Thus, a clear analysis of the convergence rate of the inexact PGA is required to advance our understanding of its strength for solving the ℓ_p regularization problem (1).

The aim of the present paper is to investigate the linear convergence issue of an inexact descent method and inexact PGAs for solving the ℓ_p regularization problem (1). For this purpose, we first investigate an optimality condition theorem for the local minima of the ℓ_p regularization problem (1), in which we establish the equivalences among a local minimum, second-order optimality condition and second-order growth property of the ℓ_p regularization problem (1). The established optimality conditions are not only of independent interest (which, in particular, improve the result in [17] and ensure the KL property of the ℓ_p regularized function with exponent $\frac{1}{2}$ at the local minimum; see Remark 1) in investigating the structure of local minima, but also provide a crucial tool for establishing the linear conver-

[1] This condition is satisfied automatically for the ℓ_p regularization problem (1).

gence of the inexact descent method and inexact PGAs for solving the ℓ_p regularization problem in Sects. 4 and 5.

We then consider a general framework of an inexact descent method, in which both (H1) and (H2) are relaxed to inexact forms (see (H1°) and (H2°) in Sect. 4), for solving the ℓ_p regularization problem. Correspondingly, the solution sequence does not satisfy the descent property. This is an essential difference from the extensive studies in descent methods and the work of Frankel et al. [22], which obstructs the application of the methodology of KL theory for exact descent methods proposed in [1,2] to the convergence analysis of the inexact descent method; see Remark 3 for the detail explanations. Instead, under some mild assumptions on the limiting points and inexact terms, we establish the linear convergence of the inexact descent method by virtue of both second-order optimality condition and second-order growth property (see Theorem 2). Our convergence analysis deviates significantly from that of [22] and relevant works in descent methods [1,2], where the KL inequality is used as a standard technique.

The convergence theorem for the inexact descent method further provides a useful tool for establishing the linear convergence of the inexact PGAs in Sect. 5. In particular, we investigate the inexact versions of the PGA for solving the ℓ_p regularization problem (1), in which the proximity operator of the ℓ_p regularizer (2) is approximately solved at each iteration (with progressively better accuracy). Inspired by the ideas in the seminal work of Rockafellar [43], we consider two types of inexact PGAs: one measures the inexact term by the approximation of proximal regularized function value, and the other is measured by the distance of the iterate to the exact proximal operator (see Algorithms IPGA-I and IPGA-II). Under some suitable assumptions on the inexact terms, we establish the linear convergence of these two inexact PGAs to a local minimum of problem (1); see Theorems 5 and 6. It is worth noting that neither of these inexact PGAs satisfies the conditions of the inexact descent method mentioned earlier; see the explanation in Remark 5(ii). In our analysis in this part, Theorem 2 plays an important role in such a way that we are able to show that the components sequence on the support of the limiting point satisfies the conditions of Theorem 2. We further propose two implementable inexact PGAs that satisfy the assumptions made in the convergence theorems and thus share the linear convergence property. It is worth noting that our results cannot be established under the framework of the KL theory.

As an interesting byproduct, the results obtained above are extended to the infinite-dimensional Hilbert spaces. Bredies et al. [10] investigated the PGA for solving the ℓ_p regularization problem in infinite-dimensional Hilbert spaces and proved its global convergence to a critical point under some technical assumptions and using dedicated tools from algebraic geometry; see the explanation before Theorem 9. Dropping these technical assumptions, we prove the global convergence of the PGA under the only assumption on stepsizes (as in [10]), which significantly improves [10, Theorem 5.1], and, under a simple additional assumption, further establish the linear convergence of the descent method and PGA, as well as their inexact versions, for solving the ℓ_p regularization problem in infinite-dimensional Hilbert spaces.

The paper is organized as follows. In Sect. 2, we present the notations and preliminary results to be used in the present paper. In Sect. 3, we establish the equivalences among a local minimum, second-order optimality condition and second-order growth property of the ℓ_p regularization problem (1), as well as some interesting corollaries. By virtue of the second-order optimality condition and second-order growth property, the linear convergence of an inexact descent method and inexact PGAs for solving problem (1) are established in Sects. 4 and 5, respectively. Finally, the convergence properties of relevant algorithms are extended to the infinite-dimensional Hilbert spaces in Sect. 6.

2 Notation and preliminary results

We consider the n-dimensional Euclidean space $\mathbb{R}^n$ with inner product $\langle \cdot\,, \cdot \rangle$ and Euclidean norm $\|\cdot\|$. For $0 < p < 1$ and $x \in \mathbb{R}^n$, the ℓ_p "norm" on $\mathbb{R}^n$ is denoted by $\|\cdot\|_p$ and defined as follows:

$$\|x\|_p := \left(\sum_{i=1}^{n} |x_i|^p\right)^{\frac{1}{p}} \quad \text{for each } x \in \mathbb{R}^n;$$

while $\|x\|_0$ denotes the number of nonzero components of x. It is well-known (see, e.g., [24, Eq. (7)]) that

$$\|x\|_p \geq \|x\|_q \quad \text{for each } x \in \mathbb{R}^n \text{ and } 0 < p \leq q. \tag{4}$$

We write $\mathrm{supp} : \mathbb{R}^n \to \mathbb{R}$ and $\mathrm{sign} : \mathbb{R} \to \mathbb{R}$ to denote the support function and signum function, respectively. For an integer $l \leq n$, fixing $x \in \mathbb{R}^l$ and $\delta \in \mathbb{R}_+$, we use $\mathbf{B}(x, \delta)$ to denote the open ball of radius δ centered at x (in the Euclidean norm). Moreover, we write

$$\mathbb{R}^l_{\neq} := \{x \in \mathbb{R}^l : x_i \neq 0 \text{ for each } i = 1, \ldots, l\}.$$

Let $\mathbb{R}^{l\times l}$ denote the space of all $l \times l$ matrices. We endow $\mathbb{R}^{l\times l}$ with the partial orders $\succ$ and $\succeq$, which are defined for any $Y,\ Z \in \mathbb{R}^{l\times l}$ by

$$Y \succ (\text{resp., } \succeq)\ Z \quad \Longleftrightarrow \quad Y - Z \text{ is positive definite (resp., positive semi-definite).}$$

Thus, for $Z \in \mathbb{R}^{l\times l}$, $Z \succ 0$ (resp., $Z \succeq 0$, $Z \prec 0$) means that Z is positive definite (resp., positive semi-definite, negative definite). In particular, we use $\mathrm{diag}(x)$ to denote a square diagonal matrix with the components of vector x on its main diagonal.

For simplicity, associated with problem (1), we use $F : \mathbb{R}^n \to \mathbb{R}$ to denote the ℓ_p regularized function, and $H : \mathbb{R}^n \to \mathbb{R}$ and $\Phi : \mathbb{R}^n \to \mathbb{R}$ are the functions defined by

$$F(\cdot) := H(\cdot) + \Phi(\cdot), \quad H(\cdot) := \|A \cdot -b\|^2 \quad \text{and} \quad \Phi(\cdot) := \lambda\|\cdot\|_p^p. \tag{5}$$

Letting $x^* \in \mathbb{R}^n \backslash \{0\}$, we write

$$s := \|x^*\|_0 \quad \text{and} \quad I := \mathrm{supp}(x^*), \tag{6}$$

We write A_i to denote the i-th column of A, $A_I := (A_i)_{i\in I}$ and $x_I := (x_i)_{i\in I}$. Let $f : \mathbb{R}^s \to \mathbb{R}$, $h : \mathbb{R}^s \to \mathbb{R}$ and $\varphi : \mathbb{R}^s \to \mathbb{R}$ be the functions defined by

$$f(\cdot) := h(\cdot) + \varphi(\cdot), \quad h(\cdot) := \|A_I \cdot -b\|^2 \quad \text{and} \quad \varphi(\cdot) := \lambda\|\cdot\|_p^p \tag{7}$$

Obviously, φ is smooth (of arbitrary order) on $\mathbb{R}^s_{\neq}$, and so is f. The first- and second-order derivatives of φ at each $y \in \mathbb{R}^s_{\neq}$ are respectively given by

$$\nabla\varphi(y) = \lambda p\left(\left(|y_i|^{p-1}\mathrm{sign}(y_i)\right)_{i\in I}\right) \quad \text{and} \quad \nabla^2\varphi(y) = \lambda p(p-1)\mathrm{diag}\left(\left(|y_i|^{p-2}\right)_{i\in I}\right). \tag{8}$$

Since $0 < p < 1$, it is clear that $\nabla^2\varphi(y) \prec 0$ for any $y \in \mathbb{R}^s_{\neq}$. By (5) and (7), one sees that

$$\Phi(x) = \varphi(x_I) \text{ and } F(x) = f(x_I) \text{ for each } x \text{ satisfying } \mathrm{supp}(x) = I. \tag{9}$$

The point x^* is called a critical point of problem (1) if it satisfies that $\nabla f(x^*_I) = 0$. The following elementary equality is repeatedly used in our convergence analysis:

$$\|Ay - b\|^2 - \|Ax - b\|^2 = \left\langle y - x, 2A^\top(Ax - b)\right\rangle + \|A(y - x)\|^2 \tag{10}$$

(by Taylor's formula applied to the function $\|A\cdot -b\|^2$). We end this section by providing the following lemma, which is useful to establish the linear convergence of inexact decent methods.

Lemma 1 *Let $\eta\in(0,1)$, and let $\{a_k\}$ and $\{\delta_k\}$ be two sequences of nonnegative scalars such that*

$$a_{k+1}\le a_k\eta+\delta_k \text{ for each } k\in\mathbb{N} \quad \text{and} \quad \limsup_{k\to\infty}\frac{\delta_{k+1}}{\delta_k}<1. \tag{11}$$

Then there exist $\theta\in(0,1)$ and $K>0$ such that

$$a_k\le K\theta^k \quad \text{for each } k\in\mathbb{N}. \tag{12}$$

Proof By the second inequality of (11), there exist $\theta\in(\eta,1)$ and $N\in\mathbb{N}$ such that

$$\delta_N<1 \quad \text{and} \quad \delta_{k+1}\le\theta\delta_k \text{ for each } k\ge N. \tag{13}$$

Then we show by mathematical induction that

$$a_{N+k}\le a_N\eta^k+\frac{\theta^{k-1}}{\theta-\eta} \text{ for each } k\in\mathbb{N}. \tag{14}$$

Clearly, (14) holds for $k=1$ (by (11) and $\delta_N<1$). Suppose that (14) holds when $k=i$. Then it follows from (11) and (13) that

$$a_{N+i+1}\le\left(a_N\eta^i+\frac{\theta^{i-1}}{\theta-\eta}\right)\eta+\theta^i\delta_N\le a_N\eta^{i+1}+\frac{\theta^i}{\theta-\eta}.$$

That is, (14) holds when $k=i+1$, and so, it holds for each $k\in\mathbb{N}$ by mathematical induction. Then (14) is applicable to deriving (12) by letting $K:=\max\{2a_N,\frac{2}{\theta(\theta-\eta)},\max_{k\le N}\frac{a_k}{\theta^k}\}$, and the proof is complete. □

3 Characterizations of local minima

Optimality condition is a crucial tool for optimization problems, either providing the useful characterizations of (local) minima or designing effective optimization algorithms. Some sufficient or necessary optimality conditions for the ℓ_p regularization problem (1) have been developed in the literature; see [17,24,32,39] and references therein. In particular, Chen et al. [17] established the following first- and second-order necessary optimality conditions for a local minimum x^* of problem (1), i.e.,

$$2A_I^\top(A_Ix_I^*-b)+\lambda p\left(\left(|x_i^*|^{p-1}\mathrm{sign}(x_i^*)\right)_{i\in I}\right)=0, \tag{15}$$

and

$$2A_I^\top A_I+\lambda p(p-1)\mathrm{diag}\left(\left(|x_i^*|^{p-2}\right)_{i\in I}\right)\succeq 0, \tag{16}$$

where $I=\mathrm{supp}(x^*)$ is defined by (6). These necessary conditions were used to estimate the (lower/upper) bounds for the absolute values and the number of nonzero components of local minima. However, it seems that a complete optimality condition that is both necessary and sufficient for the local minima of the ℓ_p regularization problem has not been established yet in the literature. To remedy this gap, this section is devoted to providing some necessary and sufficient characterizations for the local minima of problem (1).

To begin with, the following lemma (i.e., [24, Lemma 10]) illustrates that the ℓ_p regularized function satisfies a first-order growth property at 0, which is useful for proving the equivalent characterizations of its local minima. This property also indicates a significant advantage of the ℓ_p regularization over the ℓ_1 regularization that the ℓ_p regularization has a strong sparsity promoting capability.

Lemma 2 *Let $h : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function. Then there exist $\epsilon > 0$ and $\delta > 0$ such that*

$$h(x) + \lambda \|x\|_p^p \geq h(0) + \epsilon \|x\| \quad \text{for any } x \in \mathbf{B}(0, \delta).$$

The main result of this section is presented in the following theorem, in which we establish the equivalences among a local minimum, second-order optimality condition and second-order growth property of the ℓ_p regularization problem (1). Note that the latter two conditions were provided in [24] as necessary conditions for the group-wised ℓ_p regularization problem, while the second-order optimality condition is an improvement of the result in [17] in that the matrix in the left-hand side of (16) is indeed positive definite. Recall that $F : \mathbb{R}^n \to \mathbb{R}$ is the ℓ_p regularized function defined by (5) and $I = \text{supp}(x^*)$ is defined by (6).

Theorem 1 *Let $x^* \in \mathbb{R}^n \backslash \{0\}$. Then the following assertions are equivalent:*

(i) *x^* is a local minimum of problem (1).*

(ii) *(15) and the following condition hold:*

$$2A_I^\top A_I + \lambda p(p-1)\text{diag}\left(\left(|x_i^*|^{p-2}\right)_{i\in I}\right) \succ 0. \tag{17}$$

(iii) *Problem (1) satisfies the second-order growth property at x^*, i.e., there exist $\epsilon > 0$ and $\delta > 0$ such that*

$$F(x) \geq F(x^*) + \epsilon \|x - x^*\|^2 \quad \text{for any } x \in \mathbf{B}(x^*, \delta). \tag{18}$$

Proof Without loss of generality, we assume that $I = \{1, \ldots, s\}$.

(i) $\Rightarrow$ (ii). Suppose that (i) holds. Then x_I^* is a local minimum of f [by (9)], and (15) and (16) hold by [17, p. 76] (they can also be checked directly by the optimality condition for smooth optimization in [5, Proposition 1.1.1]): $\nabla f(x_I^*) = 0$ and $\nabla^2 f(x_I^*) \succeq 0$. Thus, it remains to prove (17), i.e., $\nabla^2 f(x_I^*) \succ 0$. To do this, suppose on the contrary that (17) does not hold. Then, by (16), there exists $w \neq 0$ such that $\langle w, \nabla^2 f(x_I^*)w\rangle = 0$. Let $\psi : \mathbb{R} \to \mathbb{R}$ be defined by

$$\psi(t) := f(x_I^* + tw) \quad \text{for each } t \in \mathbb{R}.$$

Then one sees that $\psi'(0) = \langle w, \nabla f(x_I^*)\rangle = 0$ and $\psi''(0) = \langle w, \nabla^2 f(x_I^*)w\rangle = 0$, and 0 is a local minimum of ψ (as x_I^* is a local minimum of f). Therefore, $\psi^{(3)}(0) = 0$ and $\psi^{(4)}(0) \geq 0$. However, by the elementary calculus, one can check that

$$\psi^{(4)}(0) = \lambda p(p-1)(p-2)(p-3)\sum_{i\in I}\left(w_i^4 |x_i^*|^{p-4}\right) < 0,$$

which yields a contradiction. Hence, assertion (ii) holds.

(ii) $\Rightarrow$ (iii). Suppose that assertion (ii) of this theorem holds. Then

$$\nabla f(x_I^*) = 0 \quad \text{and} \quad \nabla^2 f(x_I^*) \succ 0. \tag{19}$$

By Taylor's formula, we have that

$$f(y) = f(x_I^*) + \nabla f(x_I^*)(y - x_I^*) + \frac{1}{2}\langle y - x_I^*, \nabla^2 f(x_I^*)(y - x_I^*)\rangle$$

$$+o(\|y - x_I^*\|^2) \quad \text{for each } y \in \mathbb{R}^s.$$

This, together with (19), implies that there exist $\epsilon_1 > 0$ and $\delta_1 > 0$ such that

$$f(y) \geq f(x_I^*) + 2\epsilon_1 \|y - x_I^*\|^2 \quad \text{for any } y \in \mathbf{B}(x_I^*, \delta_1). \tag{20}$$

Let $\tau > 0$ be such that $\sqrt{\epsilon_1 \tau} \geq \|A_I\| \|A_{I^c}\|$, and define $g : \mathbb{R}^{n-s} \to \mathbb{R}$ by

$$g(z) := \|A_{I^c} z\|^2 + 2\langle A_I x_I^* - b, A_{I^c} z\rangle - 2\tau \|z\|^2 \quad \text{for each } z \in \mathbb{R}^{n-s}. \tag{21}$$

Clearly, g is continuously differentiable on $\mathbb{R}^{n-s}$ with $g(0) = 0$. Then, by Lemma 2, there exist $\epsilon_2 > 0$ and $\delta_2 > 0$ such that

$$g(z) + \lambda \|z\|_p^p \geq g(0) + \epsilon_2 \|z\| = \epsilon_2 \|z\| \geq 0 \quad \text{for any } z \in \mathbf{B}(0, \delta_2). \tag{22}$$

Fix $x := \begin{pmatrix} x_I \\ x_{I^c} \end{pmatrix}$ with $x_I \in \mathbf{B}(x_I^*, \delta_1)$ and $x_{I^c} \in \mathbf{B}(0, \delta_2)$. Then it follows from the definitions of the functions F, f and g [see (5), (7) and (21)] that

$$\begin{aligned} F(x) &= \|A_I x_I + A_{I^c} x_{I^c} - b\|^2 + \lambda \|x_I\|_p^p + \lambda \|x_{I^c}\|_p^p \\ &= \|A_I x_I - b\|^2 + \|A_{I^c} x_{I^c}\|^2 + 2\langle A_I x_I - b, A_{I^c} x_{I^c}\rangle + \lambda \|x_I\|_p^p + \lambda \|x_{I^c}\|_p^p \\ &= f(x_I) + g(x_{I^c}) + 2\tau \|x_{I^c}\|^2 + \lambda \|x_{I^c}\|_p^p + 2\langle A_I(x_I - x_I^*), A_{I^c} x_{I^c}\rangle. \end{aligned}$$

Applying (20) (to x_I in place of y) and (22) (to x_{I^c} in place of z), we have that

$$F(x) \geq f(x_I^*) + 2\epsilon_1 \|x_I - x_I^*\|^2 + 2\tau \|x_{I^c}\|^2 + 2\langle A_I(x_I - x_I^*), A_{I^c} x_{I^c}\rangle.$$

By the definition of τ, we have that

$$2|\langle A_I(x_I - x_I^*), A_{I^c} x_{I^c}\rangle| \leq 2\sqrt{\epsilon_1 \tau} \|x_I - x_I^*\| \|x_{I^c}\| \leq \epsilon_1 \|x_I - x_I^*\|^2 + \tau \|x_{I^c}\|^2,$$

and then, it follows that

$$F(x) \geq f(x_I^*) + \epsilon_1 \|x_I - x_I^*\|^2 + \tau \|x_{I^c}\|^2 \geq f(x_I^*) + \min\{\epsilon_1, \tau\} \|x - x^*\|^2$$

(noting that $x_{I^c} = x_{I^c} - x_{I^c}^*$). Hence $F(x) \geq F(x^*) + \min\{\epsilon_1, \tau\} \|x - x^*\|^2$, as $f(x_I^*) = F(x^*)$ by (9). This means that (18) holds with $\epsilon := \min\{\epsilon_1, \tau\}$ and $\delta := \min\{\delta_1, \delta_2\}$, and so (iii) is verified.

(iii) $\Rightarrow$ (i). It is trivial. The proof is complete. □

Remark 1 (i) A function $F : \mathbb{R}^n \to \mathbb{R}$ is said to satisfy the Kurdyka–Łojasiewicz (in short, KL) property at $x^* \in \mathbb{R}^n$ if there exist $\eta > 0$, a neighborhood U of x^* and a continuous concave function $\psi : [0, \eta) \to \mathbb{R}_+$ such that

(a) $\psi(0) = 0$,
(b) ψ is continuously differentiable on $(0, \eta)$ with $\psi'(s) > 0$ for each $s \in (0, \eta)$, and
(c) for each $x \in U \cap \{x \in \mathbb{R}^n : F(x^*) < F(x) < F(x^*) + \eta\}$, the KL inequality holds

$$\psi'(F(x) - F(x^*)) \operatorname{dist}(0, \partial F(x)) \geq 1. \tag{23}$$

The function F is said to be a KL function at x^* with the exponent being $\frac{1}{2}$ if it satisfies the KL property with ψ chosen as $\psi(s) := c\sqrt{s}$ for each $s \in [0, \eta)$. The KL property of F is crucial in the establishment of the global convergence of exact descent methods (i.e., the sequence satisfying (H1)–(H3) as stated in Sect. 1; see [2]), while the KL property with exponent $\frac{1}{2}$ can guarantee the linear convergence rate of exact descent methods (see, e.g., [1,7]). It is well-known that the ℓ_p regularized function ($0 < p < 1$) is a KL function (see, e.g., [1,2]). For convex function F, it was shown by [7, Theorem 5] that

second-order growth property (18) can ensure that it is a KL function with exponent $\frac{1}{2}$ at x^*. Note that the ℓ_p regularized function F is nonconvex. As communicated with Dr. Ting Kei Pong, the positive definiteness of $\nabla^2 f(x_I^*)$ [i.e., (17)] implies that f satisfies the KL property at x_I^* with exponent $\frac{1}{2}$ (by [57, Lemma 2]). One can then obtain that F is a KL function with exponent $\frac{1}{2}$ at x^* by considering the restriction of F on I and I^c respectively.

(ii) As shown in Lemma 2, for the case when $x^* = 0$, the equivalence between assertions (i) and (iii) in Theorem 1 is true, while assertion (ii) is not well defined (as $I = \emptyset$).

The structure of local minima is a useful property for the numerical study of the ℓ_p regularization problem; see, e.g., [17,55]. As a byproduct of Theorem 1, we will prove that the number of local minima of problem (1) is finite, which was claimed in [17, Corollary 2.2] but with an incomplete proof (because their proof is based on the fact that f has at most one local minimum whenever $A_I^\top A_I$ is of full rank, which is unclear).

Corollary 1 *The ℓ_p regularization problem (1) has only a finite number of local minima.*

Proof Let $I \subseteq \{1, \ldots, n\}$. We use $\mathrm{LM}(F, \mathbb{R}^n; I)$ to denote the set of local minima x^* of problem (1) with $\mathrm{supp}(x^*) = I$, and set

$$\Theta(I) := \left\{x_I : x \in \mathrm{LM}(F, \mathbb{R}^n; I)\right\}. \tag{24}$$

Then the set of local minima of problem (1) can be expressed as the union of $\mathrm{LM}(F, \mathbb{R}^n; I)$ over all subsets $I \subseteq \{1, \ldots, n\}$. Clearly, $\mathrm{LM}(F, \mathbb{R}^n; I)$ and $\Theta(I)$ have the same cardinality. Thus, to complete the proof, it suffices to show that $\Theta(I)$ is finite. To do this, we may assume that, without loss of generality, $I = \{1, \ldots, s\}$, and write

$$O := \{y \in \mathbb{R}^s_{\neq} : \nabla^2 f(y) \succ 0\}, \tag{25}$$

where $f : \mathbb{R}^s \to \mathbb{R}$ is defined by (7). Clearly, O is open in $\mathbb{R}^s$, and $\Theta(I) \subseteq O$ by Theorem 1. Thus, it follows from (24) that

$$\Theta(I) \subseteq \mathrm{LM}(f, \mathbb{R}^s) \cap O \tag{26}$$

(we indeed can show an equality), where, for an open subset U of $\mathbb{R}^s$, $\mathrm{LM}(f, U)$ stands for the set of local minima of f over U. For simplicity, we set

$$\mathbb{R}^s_J := \{y \in \mathbb{R}^s : y_j > 0 \text{ for } j \in J,\ y_j < 0 \text{ for } j \in I \backslash J\}$$

and $O_J := O \cap \mathbb{R}^s_J$ for any $J \subseteq I$. Then each O_J is open in $\mathbb{R}^s$ (as so are O and $\mathbb{R}^s_J$). This particularly implies that

$$\mathrm{LM}(f, \mathbb{R}^s) \cap O_J = \mathrm{LM}(f, O_J) \quad \text{for each } J \subseteq I. \tag{27}$$

Moreover, it is clear that $O = \cup_{J \subseteq I} O_J$. Hence

$$\Theta(I) \subseteq \mathrm{LM}(f, \mathbb{R}^s) \cap O = \cup_{J \subseteq I} \left(\mathrm{LM}(f, \mathbb{R}^s) \cap O_J\right) = \cup_{J \subseteq I} \mathrm{LM}(f, O_J) \tag{28}$$

[thanks to (26) and (27)]. Below we show that

$$O_J \text{ is convex for each } J \subseteq I. \tag{29}$$

Granting this, one concludes that each $\mathrm{LM}(f, O_J)$ is at most a singleton, because $\nabla^2 f \succ 0$ on O_J by (25) and then f is strictly convex on O_J by the higher-dimensional derivative tests for convexity (see, e.g., [44, Theorem 2.14]); hence $\Theta(I)$ is finite by (28), completing the proof.

To show (29), fix $J \subseteq I$, and let $y, z \in O_J$. Then, by definition, one has that

$$\nabla^2 f(y) \succ 0 \quad \text{and} \quad \nabla^2 f(z) \succ 0. \tag{30}$$

By elementary calculus, the map $t \mapsto t^{p-2}$ is convex on $(0, +\infty)$, and so

$$\frac{|y_i|^{p-2} + |z_i|^{p-2}}{2} \geq \left(\frac{|y_i| + |z_i|}{2}\right)^{p-2} \quad \text{for each } i \in I.$$

Consequently, we have

$$\operatorname{diag}\left(\left(\frac{|y_i|^{p-2} + |z_i|^{p-2}}{2}\right)_{i\in I}\right) \succeq \operatorname{diag}\left(\left(\left(\frac{|y_i| + |z_i|}{2}\right)^{p-2}\right)_{i\in I}\right).$$

This, together with (8) and (30), implies that

$$\nabla^2 f\left(\frac{y+z}{2}\right) \succeq \frac{\nabla^2 f(y) + \nabla^2 f(z)}{2} \succ 0.$$

Since $\frac{y+z}{2} \in \mathbb{R}^s_J \subseteq \mathbb{R}^s_{\neq}$, it follows that $\frac{y+z}{2} \in O \cap \mathbb{R}^s_J = O_J$ and (29) is proved. □

Another byproduct of Theorem 1 is the following corollary, in which we show the isolation of a local minimum of problem (1) in the sense of critical points. This property is useful for establishing the global convergence of the inexact descent method and inexact PGA. For simplicity, we use S to denote the set of critical points of problem (1).

Corollary 2 *Let x^* be a local minimum of the ℓ_p regularization problem (1). Then x^* is an isolated critical point of problem (1), that is, there exists $\tau > 0$ such that $S \cap \mathbf{B}(x^*, \tau) = \{x^*\}$.*

Proof Recall that $I = \operatorname{supp}(x^*)$ and f are defined by (6) and (7), respectively. Since x^* is a local minimum of problem (1), it follows from (9) that x^*_I is a local minimum of f and from Theorem 1 [cf. (17)] that $\nabla^2 f(x^*_I) \succ 0$. By the fact that $x^*_I \in \mathbb{R}^s_{\neq}$ and by the smoothness of f at x^*_I, we can find a constant τ with

$$0 < \tau < \left(\frac{4}{\lambda p}\|A^\top(Ax^* - b)\|_\infty\right)^{\frac{1}{p-1}} \tag{31}$$

such that

$$\mathbf{B}(x^*_I, \tau) \subseteq \mathbb{R}^s_{\neq} \cap \{y \in \mathbb{R}^s : \nabla^2 f(y) \succ 0\}. \tag{32}$$

We aim to show that $S \cap \mathbf{B}(x^*, \tau) = \{x^*\}$. To do this, let $x \in S \cap \mathbf{B}(x^*, \tau)$. We first claim that $\operatorname{supp}(x) = I$. It is clear by (32) that

$$x_i \neq 0 \quad \text{when } i \in I, \text{ and } |x_i| < \tau \text{ otherwise.} \tag{33}$$

If $i \in \operatorname{supp}(x)$, by the definition of critical point, it follows that $2A_i^\top(Ax - b) + \lambda p|x_i|^{p-1}\operatorname{sign}(x_i) = 0$; consequently, by the fact that x is closed to x^*, we obtain that

$$|x_i| = \left(\frac{2|A_i^\top(Ax - b)|}{\lambda p}\right)^{\frac{1}{p-1}} > \left(\frac{4|A_i^\top(Ax^* - b)|}{\lambda p}\right)^{\frac{1}{p-1}} > \left(\frac{4\|A^\top(Ax^* - b)\|_\infty}{\lambda p}\right)^{\frac{1}{p-1}} > \tau$$

[due to (31)]. This, together with (33), shows that $\operatorname{supp}(x) = I$, as desired.

Finally, we show that $x = x^*$. By (32), one has that f is strongly convex on $\mathbf{B}(x^*_I, \tau)$. Since x is a critical point of problem (1), one has by the definition of critical point that

$\nabla f(x_I) = 0$, and so x_I is a minimum of f on $\mathbf{B}(x_I^*, \tau)$. By the strongly convexity of f on $\mathbf{B}(x_I^*, \tau)$, we obtain $x_I = x_I^*$, and hence that $x = x^*$ (since $\mathrm{supp}(x) = I$). The proof is complete. □

4 Linear convergence of inexact descent method

This section aims to establish the linear convergence of an inexact version of descent methods in a general framework. In our analysis, we will employ both second-order optimality condition and second-order growth property, established in Theorem 1.

Let α and β be fixed positive constants and $\{\epsilon_k\} \subseteq \mathbb{R}_+$ be a sequence of nonnegative scalars, and recall that $F : \mathbb{R}^n \to \mathbb{R}$ is the ℓ_p regularized function defined by (5). We consider a sequence $\{x^k\}$ that satisfies the following relaxed conditions of (H1) and (H2).

(H1°) For each $k \in \mathbb{N}$,

$$F(x^{k+1}) - F(x^k) \leq -\alpha \|x^{k+1} - x^k\|^2 + \epsilon_k^2; \tag{34}$$

(H2°) For each $k \in \mathbb{N}$, there exists $w^{k+1} \in \partial F(x^{k+1})$ such that

$$\|w^{k+1}\| \leq \beta \|x^{k+1} - x^k\| + \epsilon_k.$$

Remark 2 Frankel et al. [22] proposed an inexact version of descent methods, in which only (H2) is relaxed to the inexact form (H2°) while the exact form (H1) is maintained. Consequently, the corresponding sequence $\{x^k\}$ satisfies a descent property. However, in our framework, note by (34) that the sequence $\{x^k\}$ does not necessarily satisfy a descent property. This is an essential difference between [22], as well as extensive studies in descent methods [1,2,8,54], and our study in this paper.

Remark 3 (i) Due to Remark 2, the methodology of KL theory for exact descent methods proposed in [2] cannot be directly applied (with some minor modification) to establish the global convergence of the inexact descent method with the solution sequence $\{x^k\}$ satisfying (H1°) and (H2°). Indeed, the descent property of objective function values (H1) is crucial in the framework of convergence analysis in [2], which guarantees that the sequence of function values $\{F(x^k)\}$ converges decreasingly to $F(x^*)$. Then the KL property of F at a cluster point x^* is applicable to establishing the global convergence of $\{x^k\}$ by deriving the following key relation that, for each $k \in \mathbb{N}$,

$$2\|x^{k+1} - x^k\| \leq \|x^k - x^{k-1}\| + \frac{\beta}{\alpha}\left(\psi(F(x^k) - F(x^*)) - \psi(F(x^{k+1}) - F(x^*))\right). \tag{35}$$

However, for the inexact descent method (i.e., satisfying (H1°) and (H2°)), the decreasing property of $\{F(x^k)\}$ to $F(x^*)$ is not satisfied; hence, the KL property (23) is not available at x^k because $\psi'(F(x^k) - F(x^*))$ is not well defined in the case when $F(x^k) - F(x^*)$ is negative. Even though we assume that the cluster point x^* is a local minimum of F (in which $\psi'(F(x^k) - F(x^*))$ is well defined), the line of convergence analysis in [2] does not follow for the inexact descent method yet. In particular, a relation similar to (35) can be obtained only if $\alpha\|x^{k+1} - x^k\|^2 \geq \epsilon_k^2$ for each $k \in \mathbb{N}$, which is essentially the case of the descent method; otherwise, we cannot derive any relation similar to (35) via the framework of KL theory. Consequently, the methodology of KL in [2] theory cannot be directly applied to achieve the global convergence of the inexact descent method, even if a cluster point of $\{x^k\}$ is a local minimum of F.

(ii) The methodology of KL theory for exact descent methods proposed in [1] cannot be directly applied to establish the linear convergence rate of the inexact descent method, even in the case where $\{x^k\}$ converges to a local minimum x^* and F is a KL function with exponent $\frac{1}{2}$ at x^*. In fact, relation (35) is crucial in the framework of convergence rate analysis in [1], which cannot be obtained for the inexact descent method (as explained in statement (i) of this remark). In this paper, we provide convergence analysis of the inexact descent method and inexact proximal gradient algorithms for the ℓ_p regularization problem (1) beyond the framework of KL theory.

We begin with the following useful properties of the inexact descent method; in particular, a consistent property that x^k has the same support as x^* when k is large [assertion (ii)] is useful for providing a uniform decomposition of $\{x^k\}$ in convergence analysis.

Proposition 1 (i) *Let $\{x^k\}$ be a sequence satisfying* (H1°) *with*

$$\sum_{k=0}^{\infty} \epsilon_k^2 < +\infty. \tag{36}$$

Then $\sum_{k=0}^{\infty} \|x^{k+1} - x^k\|^2 < +\infty$.

(ii) *Let $\{x^k\}$ be a sequence satisfying* (H2°) *with $\lim_{k\to\infty} \epsilon_k = 0$. Suppose that $\{x^k\}$ converges to x^*. Then there exists $N \in \mathbb{N}$ such that*

$$\mathrm{supp}(x^k) = \mathrm{supp}(x^*) \quad \textit{for each } k \geq N. \tag{37}$$

Proof Assertion (i) of this theorem is trivial by the assumption and the fact that $F \geq 0$. Below, we prove assertion (ii). Write

$$\gamma := \left(\frac{\lambda p}{\beta + 1 + 4\|A^\top(Ax^* - b)\|_\infty}\right)^{\frac{1}{1-p}}. \tag{38}$$

By the assumption that $\{x^k\}$ converges to x^*, there exists $N \in \mathbb{N}$ such that for each $k \geq N$

$$x_i^k \neq 0 \quad \text{when } i \in \mathrm{supp}(x^*), \text{ and } |x_i^k| < \gamma \text{ otherwise.} \tag{39}$$

Fix $k \geq N$ and $i \in \mathrm{supp}(x^k)$. By the assumption (H2°), there exists $w^k \in \partial F(x^k)$ such that

$$\|w^k\| \leq \beta\|x^k - x^{k-1}\| + \epsilon_k < \beta + 1 \tag{40}$$

(by the assumptions that $\lim_{k\to\infty} \epsilon_k = 0$ and $\lim_{k\to\infty} x^k = x^*$). Noting that $i \in \mathrm{supp}(x^k)$, we obtain by (8) that

$$|w_i^k| = |2A_i^\top(Ax^k - b) + \lambda p|x_i^k|^{p-1}\mathrm{sign}(x_i^k)| \geq \lambda p|x_i^k|^{p-1} - 4\|A^\top(Ax^* - b)\|_\infty.$$

This, together with (40) and (38), shows that $|x_i^k| > \gamma$ when $i \in \mathrm{supp}(x^k)$. This, together with (39), shows that $\mathrm{supp}(x^k) = \mathrm{supp}(x^*)$ for each $k \geq N$. The proof is complete. □

The main theorem of this section is as follows. The convergence theorem is not only of independent interest in establishing the linear convergence of inexact descent method, but also provides a useful approach for the linear convergence study of the inexact PGA in the next section. Recall that functions F and f are defined by (5) and (7), respectively.

Theorem 2 *Let $\{x^k\}$ be a sequence satisfying* (H1°) *and $\{\epsilon^k\}$ satisfy* (36). *Suppose one of limiting points of $\{x^k\}$, denoted by x^*, is a local minimum of problem* (1). *Then the following assertions are true.*

(i) $\{x^k\}$ *converges to* x^*.
(ii) *Suppose further that* $\{x^k\}$ *satisfies* (H2°) *and*

$$\limsup_{k\to\infty} \frac{\epsilon_{k+1}}{\epsilon_k} < 1. \tag{41}$$

Then $\{x^k\}$ *converges linearly to* x^*, *that is, there exist* $C > 0$ *and* $\eta \in (0, 1)$ *such that*

$$F(x^k) - F(x^*) \le C\eta^k \quad \text{and} \quad \|x^k - x^*\| \le C\eta^k \quad \textit{for each } k \in \mathbb{N}. \tag{42}$$

Proof (i) It follows from Proposition 1(i) that $\lim_{k\to\infty} \|x^{k+1} - x^k\| = 0$. By the assumption that x^* is a local minimum of problem (1), it follows from Lemma 2 that x^* is an isolated critical point of problem (1). Then, we can prove that $\{x^k\}$ converges to x^* (the proof is standard; see, e.g., the proof of [10, Proposition 2.3]).

(ii) If $x^* = 0$, it follows from Proposition 1(ii) that there exists $N \in \mathbb{N}$ such that $x^k = 0$ for each $k \ge N$, and so the conclusion holds. Then it remains to prove (42) for the case when $x^* \ne 0$.

Suppose that $x^* \ne 0$. Recall that $I = \mathrm{supp}(x^*)$ is defined by (6). By the assumption that x^* is a local minimum of problem (1), assertions (ii) and (iii) of Theorem 1 are satisfied; hence, it follows from (17) and (8) that $2A_I^\top A_I + \nabla^2\varphi(x_I^*) = \nabla^2 f(x_I^*) \succ 0$. This, together with $x_I^* \in \mathbb{R}^s_{\ne}$ [cf. (6)] and the smoothness of φ at x_I^*, implies that there exist $\epsilon > 0$, $\delta > 0$ and $L_\varphi > 0$ such that (18) holds and

$$\begin{aligned} &\mathbf{B}(x_I^*, \delta) \subseteq \mathbb{R}^s_{\ne} \cap \{y \in \mathbb{R}^s : \nabla^2\varphi(y) \succ -2A_I^\top A_I\}, \\ &\|\nabla\varphi(y) - \nabla\varphi(z)\| \le L_\varphi \|y - z\| \quad \text{for any } y, z \in \mathbf{B}(x_I^*, \delta). \end{aligned} \tag{43}$$

By assertion (i) of this theorem that $\{x^k\}$ converges to x^*, there exists $N \in \mathbb{N}$ such that (37) holds [by Proposition 1(ii)] and $x_I^k \in \mathbf{B}(x_I^*, \delta)$ for each $k \ge N$. In particular, the following relations hold for each $k \ge N$:

$$F(x^{k+1}) \ge F(x^*) + \epsilon\|x^{k+1} - x^*\|^2, \tag{44}$$

and

$$\|\nabla\varphi(x_I^k) - \nabla\varphi\left(x_I^{k+1}\right)\| \le L_\varphi \|x_I^k - x_I^{k+1}\|. \tag{45}$$

Noting by (8) and (43) that

$$\nabla^2\varphi(w) \prec 0 \quad \text{and} \quad \nabla^2 f(w) \succ 0 \quad \text{for any } w \in \mathbf{B}(x_I^*, \delta),$$

it follows that φ is concave and f is convex on $\mathbf{B}(x_I^*, \delta)$. Fix $k \ge N$. Then one has that

$$\left\langle \nabla\varphi(x_I^k), x_I^k - x_I^{k+1} \right\rangle \le \varphi(x_I^k) - \varphi\left(x_I^{k+1}\right) \tag{46}$$

and

$$f(x_I^k) - f(x_I^*) \le \left\langle \nabla f(x_I^k), x_I^k - x_I^* \right\rangle \tag{47}$$

(as $x_I^k, x_I^{k+1} \in \mathbf{B}(x_I^*, \delta)$). To proceed, we define

$$r_k := F(x^k) - F(x^*) \quad \text{for each } k \in \mathbb{N}, \tag{48}$$

and then it follows from (37) and (9) that

$$r_k = f(x_I^k) - f(x_I^*). \tag{49}$$

Hence, using (47), we obtain that

$$r_k \le \left\langle \nabla f(x_I^k), x_I^k - x_I^* \right\rangle = \left\langle \nabla f(x_I^k), x_I^k - x_I^{k+1} \right\rangle + \left\langle \nabla f(x_I^k), x_I^{k+1} - x_I^* \right\rangle. \tag{50}$$

By (7) and (46), it follows that

$$\begin{aligned} \left\langle \nabla f(x_I^k), x_I^k - x_I^{k+1} \right\rangle &= \left\langle \nabla h(x_I^k), x_I^k - x_I^{k+1} \right\rangle + \left\langle \nabla \varphi(x_I^k), x_I^k - x_I^{k+1} \right\rangle \\ &\le \left\langle \nabla h(x_I^k), x_I^k - x_I^{k+1} \right\rangle + \varphi(x_I^k) - \varphi\left(x_I^{k+1}\right). \end{aligned}$$

Recall from (7) that $\nabla h(x_I^k) = 2A_I^\top (A_I x_I^k - b)$. Then, by (10) (with $A_I, x_I^{k+1}, x_I^{k+1}$ in place of A, y, x), we have that

$$\begin{aligned} \left\langle \nabla f(x_I^k), x_I^k - x_I^{k+1} \right\rangle &\le f(x_I^k) - f\left(x_I^{k+1}\right) + \|A_I(x_I^{k+1} - x_I^k)\|^2 \\ &\le r_k - r_{k+1} + \|A\|^2 \|x^{k+1} - x^k\|^2 \end{aligned} \tag{51}$$

[due to (49)]. On the other hand, one has that

$$\left\langle \nabla f(x_I^k), x_I^{k+1} - x_I^* \right\rangle = \left\langle \nabla f\left(x_I^{k+1}\right), x_I^{k+1} - x_I^* \right\rangle + \left\langle \nabla f(x_I^k) - \nabla f\left(x_I^{k+1}\right), x_I^{k+1} - x_I^* \right\rangle. \tag{52}$$

By the assumption (H2°), we obtain that

$$\begin{aligned} \left\langle \nabla f\left(x_I^{k+1}\right), x_I^{k+1} - x_I^* \right\rangle &\le \|\nabla f\left(x_I^{k+1}\right)\| \|x_I^{k+1} - x_I^*\| \\ &\le \|w^{k+1}\| \|x_I^{k+1} - x_I^*\| \\ &\le \beta \|x^{k+1} - x^k\| \|x^{k+1} - x^*\| + \epsilon_k \|x^{k+1} - x^*\|; \end{aligned}$$

while by (7) and (45), we conclude that

$$\begin{aligned} &\left\langle \nabla f(x_I^k) - \nabla f\left(x_I^{k+1}\right), x_I^{k+1} - x_I^* \right\rangle \\ &= \left\langle \nabla h(x_I^k) - \nabla h\left(x_I^{k+1}\right) + \nabla \varphi(x_I^k) - \nabla \varphi\left(x_I^{k+1}\right), x_I^{k+1} - x_I^* \right\rangle \\ &\le (2\|A\|^2 + L_\varphi) \|x_I^{k+1} - x_I^k\| \|x_I^{k+1} - x_I^*\| \\ &\le (2\|A\|^2 + L_\varphi) \|x^{k+1} - x^k\| \|x^{k+1} - x^*\|. \end{aligned}$$

Combining the above two inequalities, it follows from (52) that

$$\left\langle \nabla f(x_I^k), x_I^{k+1} - x_I^* \right\rangle \le \left(\beta + 2\|A\|^2 + L_\varphi\right) \|x^{k+1} - x^k\| \|x^{k+1} - x^*\| + \epsilon_k \|x^{k+1} - x^*\|.$$

Let

$$\sigma := \beta + 2\|A\|^2 + L_\varphi \quad \text{and} \quad \tau \in (0, \epsilon). \tag{53}$$

Then one has that

$$\begin{aligned} \left\langle \nabla f(x_I^k), x_I^{k+1} - x_I^* \right\rangle &\le \frac{\sigma^2}{2\tau} \|x^{k+1} - x^k\|^2 + \frac{\tau}{2} \|x^{k+1} - x^*\|^2 + \frac{1}{2\tau} \epsilon_k^2 + \frac{\tau}{2} \|x^{k+1} - x^*\|^2 \\ &= \frac{\sigma^2}{2\tau} \|x^{k+1} - x^k\|^2 + \tau \|x^{k+1} - x^*\|^2 + \frac{1}{2\tau} \epsilon_k^2. \end{aligned}$$

This, together with (50) and (51), shows that

$$r_k \le r_k - r_{k+1} + \left(\|A\|^2 + \frac{\sigma^2}{2\tau}\right) \|x^{k+1} - x^k\|^2 + \tau \|x^{k+1} - x^*\|^2 + \frac{1}{2\tau} \epsilon_k^2. \tag{54}$$

Recalling (48), we obtain by the assumption (H1°) that

$$\|x^{k+1}-x^k\|^2 \le \frac{1}{\alpha}\left(F(x^k)-F(x^{k+1})\right)+\frac{1}{\alpha}\epsilon_k^2 = \frac{1}{\alpha}(r_k-r_{k+1})+\frac{1}{\alpha}\epsilon_k^2,$$

and by (44) that

$$\|x^{k+1}-x^*\|^2 \le \frac{1}{\epsilon}\left(F(x^{k+1})-F(x^*)\right) = \frac{1}{\epsilon}r_{k+1}.$$

Hence, (54) reduces to

$$r_k \le r_k - r_{k+1} + \frac{2\tau\|A\|^2+\sigma^2}{2\tau\alpha}(r_k-r_{k+1}) + \frac{2\tau\|A\|^2+\sigma^2}{2\tau\alpha}\epsilon_k^2 + \frac{\tau}{\epsilon}r_{k+1} + \frac{1}{2\tau}\epsilon_k^2,$$

that is,

$$r_{k+1} \le \left(1-\frac{1-\frac{\tau}{\epsilon}}{1+\frac{2\tau\|A\|^2+\sigma^2}{2\tau\alpha}-\frac{\tau}{\epsilon}}\right)r_k + \left(\frac{2\tau\|A\|^2+\sigma^2+\alpha}{2\tau\alpha+2\tau\|A\|^2+\sigma^2-2\tau^2\alpha\frac{1}{\epsilon}}\right)\epsilon_k^2. \quad (55)$$

Let

$$\bar{\eta} := 1-\frac{1-\frac{\tau}{\epsilon}}{1+\frac{2\tau\|A\|^2+\sigma^2}{2\tau\alpha}-\frac{\tau}{\epsilon}} \quad \text{and} \quad \bar{c} := \frac{2\tau\|A\|^2+\sigma^2+\alpha}{2\tau\alpha+2\tau\|A\|^2+\sigma^2-2\tau^2\alpha\frac{1}{\epsilon}}.$$

Then (55) reduces to

$$r_{k+1} \le \bar{\eta}r_k + \bar{c}\epsilon_k^2 \quad \text{for each } k \ge N.$$

One can check that $0<\bar{\eta}<1$ and $\bar{c}>0$ by (53), and note (41). Applying Lemma 1 (with r_k, $\bar{\eta}$ and $\bar{c}\epsilon_k^2$ in place of a_k, η and δ_k), there exist $\theta\in(0,1)$ and $K>0$ such that

$$F(x^k)-F(x^*) = r_k \le K\theta^k \quad \text{for each } k \ge N$$

[by (48)]. Furthermore, using (44), we have that

$$\|x^k-x^*\| \le \left(\frac{F(x^k)-F(x^*)}{\epsilon}\right)^{\frac{1}{2}} \le \left(\frac{K}{\epsilon}\right)^{\frac{1}{2}}\left(\sqrt{\theta}\right)^k \quad \text{for each } k \ge N.$$

This shows that (42) holds with $C := \max\left\{K, \left(\frac{K}{\epsilon}\right)^{\frac{1}{2}}\right\}$ and $\eta := \sqrt{\theta}$. The proof is complete. □

Remark 4 It is worth noting in (42) that the linear convergence of $\{F(x^k)\}$ to $F(x^*)$ is a direct consequence of that of $\{x^k\}$ to x^*. Indeed, recalling from [24, Lemma 2] that $\|x\|_p^p-\|y\|_p^p \le \|x-y\|_p^p$ for any $x, y\in\mathbb{R}^n$, we obtain by (5) that

$$F(x^k)-F(x^*) \le \|A\|^2\|x^k-x^*\|^2+\lambda\|x^k-x^*\|_p^p.$$

As an application of Theorem 2 for the case when $\epsilon_k \equiv 0$, the linear convergence of the descent methods investigated in [1,2] for solving the ℓ_p regularization problem (1) is presented in the following theorem.

Theorem 3 *Let $\{x^k\}$ be a sequence satisfying* (H1) *and* (H2). *Then $\{x^k\}$ converges to a critical point x^* of problem* (1). *Suppose that x^* is a local minimum of problem* (1). *Then $\{x^k\}$ converges linearly to x^*.*

5 Linear convergence of inexact proximal gradient algorithms

The main purpose of this section is to investigate the linear convergence rate of two inexact PGAs for solving the ℓ_p regularization problem (1). Associated to problem (2), we denote the (inexact) proximal operator of the ℓ_p regularizer by

$$\mathcal{P}_{v,\epsilon}(x) := \epsilon\text{-}\arg\min_{y\in\mathbb{R}^n}\left\{\lambda\|y\|_p^p + \frac{1}{2v}\|y-x\|^2\right\}. \tag{56}$$

In the special case when $\epsilon = 0$, we write $\mathcal{P}_v(x)$ for $\mathcal{P}_{v,0}(x)$ for simplicity. Recall that functions F and H are defined by (5). It is clear that the iterative formula of Algorithm PGA is

$$x^{k+1} \in \mathcal{P}_{v_k}\left(x^k - v_k\nabla H\left(x^k\right)\right).$$

Some useful properties of the proximal operator of the ℓ_p regularizer are presented as follows.

Proposition 2 *Let $v > 0$, $\epsilon > 0$, $x \in \mathbb{R}^n$, $\xi \in \mathbb{R}^n$, $y \in \mathcal{P}_v(x - v\nabla H(x))$ and $z \in \mathcal{P}_{v,\epsilon}(x - v(\nabla H(x) + \xi))$. Then the following assertions are true.*

(i) $F(z) - F(x) \le -\left(\frac{1}{2v} - \|A\|^2\right)\|z-x\|^2 - \langle z-x, \xi\rangle + \epsilon$.
(ii) *For each $i \in \mathbb{N}$, the following implication holds*

$$y_i \neq 0 \quad\Rightarrow\quad |y_i| \ge (v\lambda p(1-p))^{\frac{1}{2-p}}.$$

Proof (i) Recall that H and Φ are defined by (5), that is, $H(\cdot) = \|A\cdot - b\|^2$ and $\Phi(\cdot) = \lambda\|\cdot\|_p^p$. It follows from (56) that

$$\Phi(z) + \frac{1}{2v}\|z - (x - v(\nabla H(x) + \xi))\|^2 \le \Phi(x) + \frac{1}{2v}\|v(\nabla H(x) + \xi)\|^2 + \epsilon,$$

that is,

$$\Phi(z) - \Phi(x) \le -\frac{1}{2v}\|z-x\|^2 - \left\langle z-x, 2A^\top(Ax-b)\right\rangle - \langle z-x, \xi\rangle + \epsilon.$$

Combining this with (10), we prove assertion (i) of this theorem.

(ii) Let $i \in \mathbb{N}$ be such that $y_i \neq 0$. Then, by (56) (with $\epsilon = 0$), one has that

$$y_i \in \arg\min_{t\in\mathbb{R}}\left\{\lambda|t|^p + \frac{1}{2v}(t - (x - v\nabla H(x))_i)^2\right\}.$$

Thus, using its second-order necessary condition, we obtain that $\lambda p(p-1)|y_i|^{p-2} + \frac{1}{v} \ge 0$; consequently, $|y_i| \ge (v\lambda p(1-p))^{\frac{1}{2-p}}$. The proof is complete. □

Inspired by the ideas in the seminal work of Rockafellar [43], we propose the following two types of inexact PGAs.

ALGORITHM IPGA- I. Given an initial point $x^0 \in \mathbb{R}^n$, a sequence of stepsizes $\{v_k\} \subseteq \mathbb{R}_+$ and a sequence of inexact terms $\{\epsilon_k\} \subseteq \mathbb{R}_+$. For each $k \in \mathbb{N}$, having x^k, we determine x^{k+1} by

$$x^{k+1} \in \mathcal{P}_{v_k,\epsilon_k}\left(x^k - v_k\nabla H\left(x^k\right)\right). \tag{57}$$

ALGORITHM IPGA- II. Given an initial point $x^0 \in \mathbb{R}^n$, a sequence of stepsizes $\{v_k\} \subseteq \mathbb{R}_+$ and a sequence of inexact terms $\{\epsilon_k\} \subseteq \mathbb{R}_+$. For each $k \in \mathbb{N}$, having x^k, we determine x^{k+1} satisfying

$$\operatorname{dist}\left(x^{k+1}, \mathcal{P}_{v_k}\left(x^k - v_k \nabla H\left(x^k\right)\right)\right) \leq \epsilon_k. \tag{58}$$

Remark 5 (i) Algorithms IPGA-I and IPGA-II adopts two popular inexact schemes in the calculation of proximal operators, respectively: Algorithm IPGA-I (resp., Algorithm IPGA-II) measures the inexact term by the approximation of proximal regularized function value (resp., by the distance of the iterate to the exact proximal operator). The latter type of inexact scheme is commonly considered in theoretical analysis, while the former one is more attractive to implement in practical applications. Recently, Frankel et al. [22] proposed an inexact PGA (based on a similar inexact scheme to Algorithm IPGA-II) for solving the general problem (3).

(ii) Neither Algorithms IPGA-I nor IPGA-II satisfies both conditions (H1°) and (H2°) of the inexact descent method mentioned in Sect. 4. Indeed, if both conditions (H1°) and (H2°) are satisfied, then Lemma 1 ensures a consistent property of the support of $\{x^k\}$ to x^* [cf. (37)], which is impossible for either Algorithms IPGA-I or IPGA-II. In particular, Algorithms IPGA-I only satisfies condition (H1°) (shown in the proof of Theorem 4), while neither (H1°) nor (H2°) can be shown for Algorithms IPGA-II.

Using Theorem 2, the global convergence result of Algorithm IPGA-I is presented in the following theorem. However, we are not able to prove the global convergence of Algorithm IPGA-II at this moment.

Theorem 4 *Let $\{x^k\}$ be a sequence generated by Algorithm IPGA-I with $\{v_k\}$ satisfying*

$$0 < \underline{v} \leq v_k \leq \bar{v} < \frac{1}{2}\|A\|^{-2} \quad \text{for each } k \in \mathbb{N}. \tag{59}$$

and $\{\epsilon_k\}$ satisfying (36). Suppose that one of limiting points of $\{x^k\}$, denoted by x^*, is a local minimum of problem (1). Then $\{x^k\}$ converges to x^*.

Proof In view of Algorithm IPGA-I [cf. (57)] and by Proposition 2(i) (with x^{k+1}, x^k, v_k, 0, ϵ_k in place of z, x, v, ξ, ϵ), we obtain that

$$\begin{aligned} F(x^{k+1}) - F(x^k) &\leq -\left(\frac{1}{2v_k} - \|A\|^2\right)\|x^{k+1} - x^k\|^2 + \epsilon_k \\ &\leq -\left(\frac{1}{2\bar{v}} - \|A\|^2\right)\|x^{k+1} - x^k\|^2 + \epsilon_k \end{aligned}$$

[by (59)]. Note also by (59) that $\frac{1}{2\bar{v}} - \|A\|^2 > 0$. This shows that $\{x^k\}$ satisfies (H1°) with $\frac{1}{2\bar{v}} - \|A\|^2$ and $\sqrt{\epsilon_k}$ in place of α and ϵ_k, respectively. Then the conclusion directly follows from Theorem 2(i). The proof is complete. □

Recall that, for the inexact proximal point algorithm (see, e.g., [43,51]), the inexact term is assumed to have progressively better accuracy to investigate its convergence rate; specifically, it is assumed that $x^{k+1} \in \mathcal{P}_{v_k,\epsilon_k}(x^k)$ with $\epsilon_k = o(\|x^{k+1} - x^k\|^2)$ or that $\operatorname{dist}\left(x^{k+1}, \mathcal{P}_{v_k}(x^k)\right) \leq o(\|x^{k+1} - x^k\|)$. However, we are not able to prove the linear convergence of the inexact PGAs under this assumption of inexact term yet (due to the nonconvexity of the ℓ_p regularized function), and we need some additional assumptions to ensure the linear

convergence. Recall that $I = \mathrm{supp}(x^*)$ is defined by (6). Let $\{t_k\} \subseteq \mathbb{R}_+$ and $\{\tau_k\} \subseteq \mathbb{R}_+$. For Algorithms IPGA-I and IPGA-II, we assume

$$x_I^{k+1} \in \mathcal{P}_{v_k,\epsilon_k}\left(\left(x^k - v_k \nabla H(x^k)\right)_I\right) \quad \text{with} \quad \epsilon_k \le \tau_k \|x_I^{k+1} - x_I^k\|^2, \tag{60}$$

$$x_{I^c}^{k+1} \in \mathcal{P}_{v_k,\epsilon_k}\left(\left(x^k - v_k \nabla H(x^k)\right)_{I^c}\right) \quad \text{with} \quad \epsilon_k \le \tau_k \|x_{I^c}^{k+1} - x_{I^c}^k\|^2, \tag{61}$$

and

$$\mathrm{dist}\left(x_I^{k+1}, \left(\mathcal{P}_{v_k}\left(x^k - v_k \nabla H\left(x^k\right)\right)\right)_I\right) \le t_k \|x_I^{k+1} - x_I^k\|, \tag{62}$$

$$\mathrm{dist}\left(x_{I^c}^{k+1}, \left(\mathcal{P}_{v_k}\left(x^k - v_k \nabla H\left(x^k\right)\right)\right)_{I^c}\right) \le t_k \|x_{I^c}^{k+1} - x_{I^c}^k\|, \tag{63}$$

respectively. Note that (60)–(61) and (62)–(63) are sufficient conditions for guaranteeing (57) with $\epsilon_k = t_k\|x^{k+1} - x^k\|$ and (58) with $\epsilon_k = t_k\|x^{k+1} - x^k\|$, respectively. (The implementable strategy of inexact PGAs that conditions (60)–(61) or (62)–(63) are satisfied will be proposed at the end of this section.) Now, we establish the linear convergence of the above two inexact PGAs for solving the ℓ_p regularization problem under the additional assumptions, respectively. Recall that f, h and φ are defined by (7). In the special case when $t_k = 0$, Theorem 5 is reduced to [24, Corollary 17] and shows the linear convergence of the exact PGA for the ℓ_p regularization problem (1).

Theorem 5 *Let $\{x^k\}$ be a sequence generated by Algorithm IPGA-II with $\{v_k\}$ satisfying (59). Suppose that $\{x^k\}$ converges to a local minimum x^* of problem (1) and that (62) and (63) are satisfied for each $k \in \mathbb{N}$ with $\lim_{k\to\infty} t_k = 0$. Then $\{x^k\}$ converges linearly to x^*.*

Proof Note that $\mathcal{P}_{v_k}\left(x^k - v_k \nabla H\left(x^k\right)\right)$ is closed for each $k \in \mathbb{N}$. Then, by (62) and (63), one can choose

$$y^k \in \mathcal{P}_{v_k}\left(x^k - v_k \nabla H\left(x^k\right)\right) \tag{64}$$

such that

$$\|x_I^{k+1} - y_I^k\| \le t_k \|x_I^{k+1} - x_I^k\| \quad \text{and} \quad \|x_{I^c}^{k+1} - y_{I^c}^k\| \le t_k \|x_{I^c}^{k+1} - x_{I^c}^k\| \quad \text{for each } k \in \mathbb{N} \tag{65}$$

Noting that $x_I^* \in \mathbb{R}^s_{\neq}$ [cf. (6)] and recalling that f, h and φ are defined by (7), there exists $0 < \delta < \left(\underline{v}\lambda p(1-p)\right)^{\frac{1}{2-p}}$ such that $\mathbf{B}(x_I^*, \delta) \subseteq \mathbb{R}^s_{\neq}$ and

$$\|\nabla\varphi(y) - \nabla\varphi(z)\| \le L_\varphi \|y - z\| \quad \text{for any } y, z \in \mathbf{B}(x_I^*, \delta). \tag{66}$$

By the assumption that $\lim_{k\to\infty} x^k = x^*$ and $I = \mathrm{supp}(x^*)$ [cf. (6)], we have by (65) that $\lim_{k\to\infty} y_I^k = x_I^*$ and $\lim_{k\to\infty} y_{I^c}^k = x_{I^c}^* = 0$. Then there exists $N \in \mathbb{N}$ such that

$$\|x_I^k - x_I^*\| \le \delta, \quad \|y_I^k - x_I^*\| \le \delta \quad \text{and} \quad \|y_{I^c}^k\| \le \delta \quad \text{for each } k \ge N.$$

Consequently, one sees that

$$x_I^k,\ y_I^k \in \mathbf{B}(x_I^*, \delta) \subseteq \mathbb{R}^s_{\neq} \quad \text{and} \quad y_{I^c}^k = 0 \quad \text{for each } k \ge N \tag{67}$$

[by Proposition 2(ii)], and by (66) that

$$\|\nabla\varphi\left(x_I^{k+1}\right) - \nabla\varphi(y_I^k)\| \le L_\varphi \|x_I^{k+1} - y_I^k\| \quad \text{for each } k \ge N. \tag{68}$$

We first provide an estimate on $\{x^k_{I^c}\}_{k\geq N}$. By the assumption that $\lim_{k\to\infty} t_k = 0$, we can assume, without loss of generality, that $t_k < \frac{1}{2}$ for each $k \geq N$. By (67), we obtain from the second inequality of (65) that

$$\|x^{k+1}_{I^c}\| \leq t_k\|x^{k+1}_{I^c} - x^k_{I^c}\| \leq t_k\|x^{k+1}_{I^c}\| + t_k\|x^k_{I^c}\|,$$

and so,

$$\|x^{k+1}_{I^c}\| \leq \frac{t_k}{1-t_k}\|x^k_{I^c}\| < 2t_k\|x^k_{I^c}\| \quad \text{for each } k \geq N. \tag{69}$$

Below, we estimate $\{x^k_I\}_{k\geq N}$. To do this, we fix $k \geq N$ and let τ be a constant such that $0 < \tau < \frac{1}{4\bar{v}} - \frac{1}{2}\|A\|^2$ [recalling (59)]. By (65) and using the triangle inequality, one has that

$$\begin{aligned}\frac{1}{2}\|x^{k+1}_I - x^k_I\| < (1-t_k)\|x^{k+1}_I - x^k_I\| \leq \|y^k_I - x^k_I\| &\leq (1+t_k)\|x^{k+1}_I - x^k_I\| \\ &< \frac{3}{2}\|x^{k+1}_I - x^k_I\|\end{aligned} \tag{70}$$

(by $t_k < \frac{1}{2}$). By (64), (5) and (7), we check that $y^k_I \in \mathcal{P}_{v_k}\left(x^k_I - v_k\left(\nabla h(x^k_I) + 2A_I A_{I^c}x^k_{I^c}\right)\right)$, and so, we obtain from Proposition 2(i) (with $f, h, A_I, y^k_I, x^k_I, v_k, 2A^\top_I A_{I^c}x^k_{I^c}, 0$ in place of $F, H, A, z, x, v, \xi, \epsilon$) that

$$\begin{aligned}f(y^k_I) - f(x^k_I) &\leq -\left(\frac{1}{2v_k} - \|A_I\|^2\right)\|y^k_I - x^k_I\|^2 - \left\langle y^k_I - x^k_I, 2A^\top_I A_{I^c}x^k_{I^c}\right\rangle \\ &\leq -\left(\frac{1}{2v_k} - \|A\|^2\right)\|y^k_I - x^k_I\|^2 + \tau\|y^k_I - x^k_I\|^2 + \frac{1}{\tau}\|A\|^4\|x^k_{I^c}\|^2 \\ &\leq -\frac{1}{4}\left(\frac{1}{2\bar{v}} - \|A\|^2 - \tau\right)\|x^{k+1}_I - x^k_I\|^2 + \frac{1}{\tau}\|A\|^4\|x^k_{I^c}\|^2\end{aligned} \tag{71}$$

[by (59) and (70)]. By the smoothness of f on $\mathbf{B}(x^*_I, \delta)(\subseteq \mathbb{R}^s_{\neq})$ and (67), there exists $L > 0$ such that

$$f(x^{k+1}_I) - f(y^k_I) \leq \|\nabla f(y^k_I)\|\|x^{k+1}_I - y^k_I\| + L\|x^{k+1}_I - y^k_I\|^2. \tag{72}$$

(by TaylorØs formula). The first-order optimality condition of (64) says that

$$\nabla\varphi(y^k_I) + \frac{1}{v_k}\left(y^k_I - x^k_I + 2v_k A^\top_I(Ax^k - b)\right) = 0. \tag{73}$$

Then we obtain by (7) that

$$\nabla f(y^k_I) = 2A^\top_I(A_I y^k_I - b) + \nabla\varphi(y^k_I) = -\left(\frac{1}{v_k} - 2A^\top_I A_I\right)(y^k_I - x^k_I) - 2A^\top_I A_{I^c}x^k_{I^c};$$

consequently,

$$\begin{aligned}\|\nabla f(y^k_I)\| &\leq \left(\frac{1}{v_k} - 2\|A\|^2\right)\|y^k_I - x^k_I\| + 2\|A\|^2\|x^k_{I^c}\| \\ &\leq \frac{3}{2}\left(\frac{1}{\bar{v}} - 2\|A\|^2\right)\|x^{k+1}_I - x^k_I\| + 2\|A\|^2\|x^k_{I^c}\|\end{aligned}$$

[due to (59) and (70)]. Combing this with (72), we conclude by the first inequality of (65) that

$$
\begin{aligned}
&f(x_I^{k+1}) - f(y_I^k) \\
&\le \frac{3}{2}\left(\frac{1}{\bar{v}} - 2\|A\|^2\right) t_k \|x_I^{k+1} - x_I^k\|^2 + 2\|A\|^2 t_k \|x_{I^c}^k\| \|x_I^{k+1} - x_I^k\| + L t_k^2 \|x_I^{k+1} - x_I^k\|^2 \\
&\le \left(\frac{3}{2}\left(\frac{1}{\bar{v}} - 2\|A\|^2\right) t_k + t_k^2(L+\tau)\right) \|x_I^{k+1} - x_I^k\|^2 + \frac{1}{\tau}\|A\|^4 \|x_{I^c}^k\|^2. \qquad (74)
\end{aligned}
$$

Recalling that $\lim_{k\to\infty} t_k = 0$, we can assume, without loss of generality, that

$$
\frac{3}{2}\left(\frac{1}{\bar{v}} - 2\|A\|^2\right) t_k + t_k^2(L+\tau) \le \frac{1}{4}\tau \quad \text{for each } k \ge N.
$$

This, together with (71) and (74), yields that

$$
f(x_I^{k+1}) - f(x_I^k) \le -\frac{1}{4}\left(\frac{1}{2\bar{v}} - \|A\|^2 - 2\tau\right) \|x_I^{k+1} - x_I^k\|^2 + \frac{2}{\tau}\|A\|^4 \|x_{I^c}^k\|^2. \qquad (75)
$$

On the other hand, by the smoothness of f on $\mathbf{B}(x_I^*, \delta)$, we obtain by (67) and (7) that

$$
\begin{aligned}
\|\nabla f\left(x_I^{k+1}\right)\| \le{}& \|\nabla h(x_I^k) + \nabla\varphi(y_I^k)\| + \|\nabla h(x_I^{k+1}) - \nabla h(x_I^k)\| \\
&+ \|\nabla\varphi(x_I^{k+1}) - \nabla\varphi(y_I^k)\|. \qquad (76)
\end{aligned}
$$

Note by (73), (70) and (59) that

$$
\|\nabla h(x_I^k) + \nabla\varphi(y_I^k)\| = \|\frac{1}{v_k}(x_I^k - y_I^k) - 2A_I^\top A_{I^c} x_{I^c}^k\| \le \frac{3}{2\underline{v}}\|x_I^{k+1} - x_I^k\| + 2\|A\|^2\|x_{I^c}^k\|,
$$

$$
\|\nabla h(x_I^{k+1}) - \nabla h(x_I^k)\| \le 2\|A\|^2\|x_I^{k+1} - x_I^k\|,
$$

and by (68) and (65) that

$$
\|\nabla\varphi(x_I^{k+1}) - \nabla\varphi(y_I^k)\| \le L_\varphi\|x_I^{k+1} - y_I^k\| \le L_\varphi t_k\|x_I^{k+1} - x_I^k\|.
$$

Hence, (76) implies that

$$
\|\nabla f\left(x_I^{k+1}\right)\| \le \left(\frac{3}{2\underline{v}} + 2\|A\|^2 + L_\varphi t_k\right)\|x_I^{k+1} - x_I^k\| + 2\|A\|^2\|x_{I^c}^k\|.
$$

This and (75) show that $\{x_I^k\}_{k\ge N}$ satisfies (H1°) and (H2°) with f, x_I^k, $\frac{1}{4}\left(\frac{1}{2\bar{v}} - \|A\|^2 - 2\tau\right)$, $\left(\frac{3}{2\underline{v}} + 2\|A\|^2 + L_\varphi t_k\right)$ and $\max\left\{\sqrt{\frac{2}{\tau}}, 2\right\}\|A\|^2\|x_{I^c}^k\|$ in place of F, x^k α, β and ϵ_k, respectively. Furthermore, it follows from (69) that $\lim_{k\to\infty}\frac{\|x_{I^c}^{k+1}\|}{\|x_{I^c}^k\|} \le \lim_{k\to\infty} 2t_k = 0$. This verifies (41) assumed in Theorem 2(ii). Therefore, the assumptions of Theorem 2(ii) are satisfied, and so it follows that $\{x_I^k\}$ converges linearly to x_I^*. Recall from (69) that $\{x_{I^c}^k\}$ converges linearly to $x_{I^c}^*$ (=0). Therefore, $\{x^k\}$ converges linearly to x^*. The proof is complete. □

Remark 6 Frankel et al. [22] considered an inexact PGA similar to Algorithm IPGA-II with the inexact control being given by

$$
\epsilon_k = t_k \operatorname{dist}\left(\mathcal{P}_{v_k}\left(x^k - v_k\nabla H\left(x^k\right)\right), \mathcal{P}_{v_k}\left(x^{k-1} - v_{k-1}\nabla H\left(x^{k-1}\right)\right)\right).
$$

However, this inexact control would be not convenient to implement for applications because ϵ_k is expressed in terms of $\mathcal{P}_v(\cdot)$ that is usually expensive to calculate exactly. In Theorem 5, we established the linear convergence of Algorithm IPGA-II with the inexact control being

given by (62) and (63). Our convergence analysis deviates significantly from that of [22], in which the KL inequality is used as a standard technique.

Theorem 6 *Let $\{x^k\}$ be a sequence generated by Algorithm IPGA-I with $\{v_k\}$ satisfying (59). Suppose that $\{x^k\}$ converges to a global minimum x^* of problem (1) and that (60) and (61) are satisfied for each $k \in \mathbb{N}$ with $\lim_{k\to\infty} \tau_k = 0$. Then $\{x^k\}$ converges linearly to x^*.*

Proof For simplicity, we write $y^k \in \mathcal{P}_{v_k}(x^k - v_k \nabla H(x^k))$ for each $k \in \mathbb{N}$. By Proposition 2(i) (with y^k, x^k, v_k, 0, 0 in place of z, x, v, ξ, ϵ) and by (59), one has that

$$\left(\frac{1}{2\bar{v}} - \|A\|^2\right)\|y^k - x^k\|^2 \le F(x^k) - F(y^k) \le F(x^k) - \min_{x\in\mathbb{R}^n} F(x).$$

Then, by the assumption that $\{x^k\}$ converges to a global minimum x^* of F, we have that $\{y^k\}$ also converges to this x^*. By Theorem 1, it follows from (17) that $2A_I^\top A_I + \nabla^2\varphi(x_I^*) = \nabla^2 f(x_I^*) \succ 0$. This, together with $x_I^* \in \mathbb{R}^s_{\neq}$ [cf. (6)] and the smoothness of φ at x_I^*, implies that there exists $0 < \delta < \left(\underline{v}\lambda p(1-p)\right)^{\frac{1}{2-p}}$ such that

$$\mathbf{B}(x_I^*, \delta) \subseteq \mathbb{R}^s_{\neq} \cap \{y \in \mathbb{R}^s : \nabla^2\varphi(y) \succ -2A_I^\top A_I\}. \tag{77}$$

By the convergence of $\{x^k\}$ and $\{y^k\}$ to x^*, there exists $N \in \mathbb{N}$ such that

$$x_I^k, y_I^k \in \mathbf{B}(x_I^*, \delta), \quad x_{I^c}^k \in \mathbf{B}(0, \delta) \quad \text{and} \quad y_{I^c}^k = 0 \quad \text{for each } k \ge N \tag{78}$$

[by Proposition 2(ii)]. Fix $k \ge N$. Then, by (61) and (56), we have that

$$\varphi(x_{I^c}^{k+1}) + \frac{1}{2v_k}\|x_{I^c}^{k+1} - x_{I^c}^k + 2v_k A_{I^c}^\top(Ax^k - b)\|^2 \le \epsilon_k + \frac{1}{2v_k}\| - x_{I^c}^k + 2v_k A_{I^c}^\top(Ax^k - b)\|^2.$$

This implies that

$$\varphi(x_{I^c}^{k+1}) \le \epsilon_k + \frac{1}{2v_k}\left(\|x_{I^c}^k\|^2 - \|x_{I^c}^k - x_{I^c}^{k+1}\|^2\right) - \left\langle x_{I^c}^{k+1}, 2A_{I^c}(Ax^k - b)\right\rangle. \tag{79}$$

Note that $\lim_{k\to\infty} x_{I^c}^k = 0$ and $\lim_{k\to\infty} \tau_k = 0$. By (79) and (61), there exists $K > 0$ such that

$$\|x_{I^c}^{k+1}\|_p^p \le K(\|x_{I^c}^{k+1}\| + \|x_{I^c}^k\|).$$

Then it follows from (4) (as $p < 1$) that

$$\left(1 - K\|x_{I^c}^{k+1}\|^{1-p}\right)\|x_{I^c}^{k+1}\|^p \le \|x_{I^c}^{k+1}\|_p^p - K\|x_{I^c}^{k+1}\| \le K\|x_{I^c}^k\|.$$

Since $\lim_{k\to\infty} x_{I^c}^k = 0$, we assume, without loss of generality, that $\|x_{I^c}^{k+1}\| \le (2K)^{-\frac{1}{1-p}}$. Hence,

$$\|x_{I^c}^{k+1}\|^p \le 2K\|x_{I^c}^k\| = \left(2K\|x_{I^c}^k\|^{1-p}\right)\|x_{I^c}^k\|^p.$$

Let $\alpha_k := \left(2K\|x_{I^c}^k\|^{1-p}\right)^{\frac{1}{p}}$. Then it follows that

$$\|x_{I^c}^{k+1} - x_{I^c}^k\| \ge \|x_{I^c}^k\| - \|x_{I^c}^{k+1}\| \ge \frac{1-\alpha_k}{\alpha_k}\|x_{I^c}^{k+1}\|. \tag{80}$$

On the other hand, let $f_k : \mathbb{R}^s \to \mathbb{R}$ be an auxiliary function defined by

$$f_k(y) := \varphi(y) + \frac{1}{2v_k}\|y - \left(x_I^k - 2v_k A_I^\top(Ax^k - b)\right)\|^2 \quad \text{for each } y \in \mathbb{R}^s. \tag{81}$$

Obviously, f_k is smooth on $\mathbb{R}^s_{\neq}$ and note by Taylor's formula of f_k at y_I^k that

$$f_k(y) = f_k(y_I^k) + \nabla f_k(y_I^k)(y - y_I^k) + \frac{1}{2}\left\langle y - y_I^k, \nabla^2 f_k(y_I^k)(y - y_I^k)\right\rangle + o(\|y - y_I^k\|^2), \forall y \in \mathbb{R}^s. \tag{82}$$

By (81), it is clear that $y_I^k \in \arg\min_{y\in\mathbb{R}^s} f_k(y)$. Its first-order necessary optimality condition says that $\nabla f_k(y_I^k) = 0$, and its second-order derivative is $\nabla^2 f_k(y_I^k) = \nabla^2\varphi(y_I^k) + \frac{1}{v_k}\mathbf{I}_s$, where $\mathbf{I}_s$ denotes the identical matrix in $\mathbb{R}^{s\times s}$. Note by (77) and (78) that $\nabla^2\varphi(y_I^k) \succ -2A_I^\top A_I$. Then

$$\nabla^2 f_k(y_I^k) \succ \frac{1}{v_k}\mathbf{I}_s - 2A_I^\top A_I \succ \frac{1}{\bar{v}}\mathbf{I}_s - 2A_I^\top A_I \succ 0$$

[by (59)]. Hence, letting σ be the smallest eigenvalue of $\frac{1}{\bar{v}}\mathbf{I}_s - 2A_I^\top A_I$, we obtain by (82) that

$$f_k(y) \geq f_k(y_I^k) + \frac{\sigma}{2}\|y - y_I^k\|^2 \quad \text{for any } y \in \mathbf{B}(y_I^k, 2\delta) \tag{83}$$

(otherwise we can select a smaller δ). By (78), one observes that

$$\|x_I^{k+1} - y_I^k\| \leq \|x_I^{k+1} - x_I^*\| + \|y_I^k - x_I^*\| \leq 2\delta,$$

and so, (83) and (60) imply that

$$\|x_I^{k+1} - y_I^k\|^2 \leq \frac{2}{\sigma}\left(f_k(x_I^{k+1}) - f_k(y_I^k)\right) \leq \frac{2}{\sigma}\tau_k\|x_I^{k+1} - x_I^k\|^2.$$

Note that $y^k \in \mathcal{P}_{v_k}(x^k)$ is arbitrary. This, together with (80), shows that $\{x^k\}$ can be seen as a special sequence generated by Algorithm IPGA-II that satisfies (62) and (63) with $\max\{\frac{\alpha_k}{1-\alpha_k}, \frac{2}{\sigma}\tau_k\}$ in place of t_k. Since $\lim_{k\to\infty}\tau_k = 0$ and $\lim_{k\to\infty}\alpha_k = 0$ (by the definition of α_k), one has that $\lim_{k\to\infty}\max\{\frac{\alpha_k}{1-\alpha_k}, \frac{2}{\sigma}\tau_k\} = 0$, and so, the conclusion directly follows from Theorem 5. □

It is a natural question how to design the inexact PGA that satisfies (60)–(61) or (62)–(63). Note that both functions $\|\cdot\|_p^p$ and $\|\cdot - x\|^2$ in the proximal operator are separable [see (56)]. We can propose two implementable inexact PGAs, Algorithms IPGA-Ip and IPGA-IIp, which are the parallel versions of Algorithms IPGA-I and IPGA-II, respectively.

ALGORITHM IPGA- IP. Given an initial point $x^0 \in \mathbb{R}^n$, a sequence of stepsizes $\{v_k\} \subseteq \mathbb{R}_+$ and a sequence of nonnegative scalars $\{\epsilon_k\} \subseteq \mathbb{R}_+$. For each $k \in \mathbb{N}$, having x^k, we determine x^{k+1} by

$$x_i^{k+1} \in \mathcal{P}_{v_k,\epsilon_k}\left(\left(x^k - v_k\nabla H(x^k)\right)_i\right) \text{ with } \epsilon_k = \tau_k\|x_i^{k+1} - x_i^k\|^2 \quad \text{for each } i = 1, \ldots, n.$$

ALGORITHM IPGA- IIP. Given an initial point $x^0 \in \mathbb{R}^n$, a sequence of stepsizes $\{v_k\} \subseteq \mathbb{R}_+$ and a sequence of nonnegative scalars $\{t_k\} \subseteq \mathbb{R}_+$. For each $k \in \mathbb{N}$, having x^k, we determine x^{k+1} satisfying

$$\operatorname{dist}\left(x_i^{k+1}, \left(\mathcal{P}_{v_k}\left(x^k - v_k\nabla H\left(x^k\right)\right)\right)_i\right) \leq t_k\|x_i^{k+1} - x_i^k\| \quad \text{for each } i = 1, \ldots, n.$$

It is easy to verify that Algorithms IPGA-Ip and IPGA-IIp satisfy conditions (60)–(61) and (62)–(63) respectively, and so, their linear convergence properties follow directly from Theorems 5 and 6.

The linear convergence of the exact PGA for the ℓ_p regularization problem (1) has been established by Zeng et al. [57] and Hu et al. [24] under different assumptions. In particular, [57, Theorem 4] proved the linear convergence of the PGA under a restrictive assumption on the stepsize v and the regularization order p that

$$\frac{p}{2} < \frac{\lambda_{\min}(A_I^\top A_I)}{\|A\|^2} \quad \text{and} \quad \frac{p}{4\lambda_{\min}(A_I^\top A_I)} < v < \frac{1}{2\|A\|^2}, \tag{84}$$

where I denotes the support of the limiting point x^* and $\lambda_{\min}(\cdot)$ denotes the smallest eigenvalue of a matrix. While, [24, Corollary 17] obtained the linear convergence of the exact PGA for each stepsize $0 < v < \frac{1}{2}\|A\|^{-2}$ and regularization order $0 < p < 1$ under the assumption that the limiting point is a local minimum. This work extends the linear convergence theory to the inexact PGA for the ℓ_p regularization problem (1), and particularly, Theorem 5 in this paper is the same as [24, Corollary 17] in the case when $t_k = 0$.

At the end of this section, two numerical examples are provided to compare the linear convergence results of the PGA. Example 1 provides a small-dimensional toy where the convergence theory in this paper is available but not the one in [57]; and Example 2 compares the numerical performance of the exact and inexact PGAs in high-dimensional sparse recovery. The numerical experiments are implemented in Matlab R2014a and executed on a personal laptop (Intel Core i7-8550U, 1.80 GHz, 16.00 GB of RAM).

Example 1 Consider the ℓ_p regularization problem (1) with $A \in \mathbb{R}^{5\times 10}$ being an i.i.d. Gaussian ensemble. In a random trial, the linear matrix is

$$A := \begin{pmatrix} -0.44 & 0.31 & 0.55 & -0.095 & -0.18 & 0.36 & -0.026 & -0.17 & 0.41 & -0.22 \\ 0.12 & -0.036 & 0.018 & 0.032 & 0.16 & 0.60 & 0.51 & 0.44 & 0.097 & 0.36 \\ -0.34 & -0.26 & -0.051 & 0.24 & 0.64 & 0.36 & -0.31 & -0.091 & -0.26 & -0.21 \\ 0.46 & -0.19 & 0.26 & 0.29 & -0.44 & 0.37 & -0.074 & -0.092 & -0.37 & -0.33 \\ -0.41 & -0.79 & 0.086 & 0.036 & -0.35 & -0.076 & 0.064 & -0.030 & 0.090 & 0.26 \end{pmatrix}$$

with $\|A\| = 1$, and the ground-true sparse solution and the observation are

$$\bar{x} := (0.82, 0.64, 0, 0, 0, 0, 0, 0, 0, 0)^\top \quad \text{and} \quad b := (-0.17, 0.078, -0.44, 0.26, -0.84)^\top,$$

respectively. Algorithm PGA is conducted to solve the problem (1) to approach the sparse solution. In the implementation of the Algorithm PGA, we set the regularization parameter $\lambda = 0.01$, the regularization order $p = 0.7$, the initial point $x^0 = 0$ and the constant stepsize $v_k \equiv v$. Three criteria are used to measure the numerical performance of the PGA: the violation of the first-order optimality condition (15):

$$\text{FOC} = \left\| 2A_I^\top (A_I x_I - b) + \lambda p \left(\left(|x_i|^{p-1} \text{sign}(x_i) \right)_{i\in I} \right) \right\|,$$

the obedience of the second-order optimality condition (17):

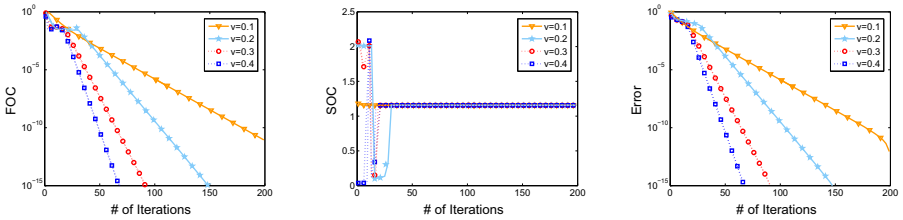
$$\text{SOC} := \lambda_{\min} \left(2A_I^\top A_I + \lambda p(p-1) \text{diag}\left(\left(|x_i|^{p-2} \right)_{i\in I} \right) \right),$$

and the relative error to the limiting point x^*:

$$\text{RE} := \frac{\|x - x^*\|}{\|x^*\|}.$$

The stopping criterion of the PGA is set as $\text{RE} \le 1e{-}16$ or the number of iterations is greater than 200.

The numerical results of the PGA with $v \in [0.1, 0.4]$ for this example are plotted in Fig. 1, including the FOC, SOC and RE along the number of iterations. It is demonstrated from



(a) First-order optimality condition. **(b)** Second-order optimality condition. **(c)** Linear convergence rate.

Fig. 1 Numerical results of the exact PGA with different stepsizes for a random trial

Fig. 1c that the PGA converges linearly for this example for all the stepsizes $v \in [0.1, 0.4]$. From Fig. 1a, b, one observes that the optimality conditions (15) and (17) are satisfied, and thus, the limiting point of the sequence generated by the PGA is a local minimum of problem (1). Hence, Theorem 5 in this paper (at a special case when $t_k = 0$, also [24, Corollary 17]) is available to guarantee the linear convergence of the PGA for all the stepsizes $v \in [0.1, 0.4]$, which is consistent with Fig. 1c.

For this sequence of the PGA, one has that $I = \{1, 2\}$ and $\lambda_{\min}(A_I^\top A_I) = 0.5763$, and thus, (84) is reduced to $v \in (0.3036, 0.5)$. Hence, [57, Theorem 4] is able to ensure the linear convergence of the PGA only for the stepsize $v = 0.4$ in this example; while the linear convergence behavior in Fig. 1c when $v = 0.1, 0.2$ and 0.3 cannot be ensured by [57, Theorem 4].

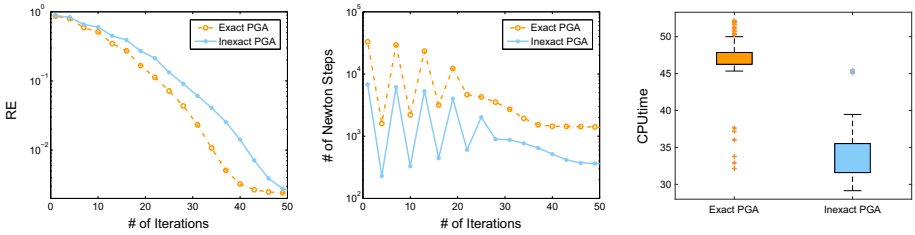
Example 2 In this example, we compare the numerical performance of the exact and inexact PGAs in high-dimensional sparse recovery. The simulation data are generated via the standard process of compressive sensing. Particularly, the matrix $A \in \mathbb{R}^{m\times n}$ is randomly generated with each entry being an i.i.d. Gaussian ensemble and $AA^\top = \mathbf{I}_m$, the ground-true sparse solution $\bar{x} \in \mathbb{R}^n$ is a random s-sparse vector with each nonzero entry drawn from the standard uniform distribution on (0, 1), and the observation $b := A\bar{x} + \varepsilon$, where ε is an additive Gaussian noise with its standard deviation being $1e$-3.

In this experiment, the numbers of samples and variables $(m, n) = (2500, 10{,}000)$ and the sparsity $s = 400$. The exact and inexact PGAs (i.e., Algorithm PGA and Algorithm IPGA-Ip) are conducted to solve the problem (1) to approach the sparse solution. In the implementation of the PGAs, we set the regularization parameter $\lambda = 0.01$, the regularization order $p = 0.7$, the initial point $x^0 = 0$ and the constant stepsize $v_k \equiv 0.4$. As reported in [36, Theorem 1], the solution of the proximity operator of the ℓ_p regularizer has a threshold value

$$\varphi(\lambda, p) := (2-p)\,(2(1-p))^{-\frac{1-p}{2-p}}\,\lambda^{\frac{1}{2-p}}.$$

For the component where $|z_i^k|$ is larger than the threshold value, the proximity subproblem (2) is approached via applying the Newton method to solve the nonlinear equation of its first-order optimality condition:

$$\lambda p x_i + \frac{1}{v}(x_i - z_i^k)|x_i|^{2-p} = 0.$$



(a) Convergence behavior. **(b)** Number of Newton steps for subproblems. **(c)** CPUtime.

Fig. 2 Numerical comparison of the exact and inexact PGAs

The calculation accuracy for the subproblems of the exact and inexact PGAs are $\tau_k = 1e\text{-}16$ and $\tau_k = \frac{1}{\sqrt{k}}$, respectively. The criterion to measure the numerical performance of the PGAs is the relative error to the ground-true solution $\bar{x}$:

$$\text{RE} := \frac{\|x - \bar{x}\|}{\|\bar{x}\|}.$$

The stopping criterion of the PGAs is set as $\text{RE} \leq 1e{-}3$ or the number of iterations is greater than 100.

We conduct 500 random simulations of sparse recovery and the numerical results of averaging these 500 simulations are illustrated in Fig. 2, including the averaged RE, the averaged number of Newton steps for subproblems and the averaged CPUtime along the number of iterations. Two observations are illustrated from Fig. 2: (i) Both the exact and inexact PGAs converge linearly to the ground-true solution, which are consistent with [24, Corollary 17] and Theorem 5, respectively. (ii) Although the inexact PGA requires a little more outer iterations than the exact PGA, however, it costs much less Newton steps (only one-tenth) in solving the proximity subproblems and thus spends less CPUtime (about half) than that of the exact PGA. Hence, the inexact PGA is more effective than the exact PGA in high-dimensional sparse recovery.

6 Extension to infinite dimensional cases

This section extends the results in preceding sections to the infinite-dimensional Hilbert spaces. In this section, we adopt the following notations. Let $\mathcal{H}$ be a Hilbert space, and let ℓ^2 denote the Hilbert space consisting of all square-summable sequences. We consider the following ℓ_p regularized least squares problem in infinite-dimensional Hilbert spaces

$$\min_{x \in l^2} \; F(x) := \|Ax - b\|^2 + \sum_{i=1}^{\infty} \lambda_i |x_i|^p, \tag{85}$$

where $A : \ell^2 \to \mathcal{H}$ is a bounded linear operator, and $\lambda := (\lambda_i)$ is a sequence of weights satisfying

$$\lambda_i \geq \underline{\lambda} > 0 \quad \text{for each } i \in \mathbb{N}. \tag{86}$$

We start from some useful properties of the (inexact) descent methods and then present the linear convergence of (inexact) descent methods and PGA for solving problem (85).

Proposition 3 *Let* $\{x^k\} \subseteq \ell^2$ *be a sequence satisfying* (H1°) *and* (H2°), *and* $\{\epsilon^k\}$ *satisfy* (36). *Then there exist* $N \in \mathbb{N}$ *and a finite index set* $J \subseteq \mathbb{N}$ *such that*

$$\mathrm{supp}(x^k) = J \quad \text{for each } k \geq N. \tag{87}$$

Proof Fix $k \in \mathbb{N}$. By (H1°), one has that

$$F(x^k) \leq F(x^{k-1}) - \alpha\|x^k - x^{k-1}\|^2 + \epsilon_{k-1}^2 \leq F(x^{k-1}) + \epsilon_{k-1}^2 \leq F(x^0) + \sum_{i=0}^{\infty} \epsilon_i^2 < +\infty$$

[due to (36)]. Then, it follows from (4) and (86) that

$$\|x^k\|^p \leq \|x^k\|_p^p \leq \frac{1}{\underline{\lambda}} \sum_{i=1}^{\infty} \lambda_i |x_i^k|^p \leq \frac{1}{\underline{\lambda}} F(x^k) < +\infty.$$

Then $\{x^k\}$ is bounded, denoting the upper bound of their norms by M. Let

$$\tau := \min\left\{\frac{1}{\beta}, \left(\frac{\underline{\lambda} p}{2 + 2\|A\|^2 M + 2\|A\|\|b\|}\right)^{1-p}\right\} \ (> 0). \tag{88}$$

Note by Proposition 1(i) that $\lim_{k\to\infty} \|x^{k+1} - x^k\| = 0$, which, together with (36), shows that there exists $N \in \mathbb{N}$ such that

$$\|x^{k+1} - x^k\| \leq \tau \quad \text{and} \quad \epsilon_k < 1 \quad \text{for each } k \geq N. \tag{89}$$

We claim that the following implication is true for each $k \geq N$ and $i \in \mathbb{N}$

$$x_i^k \neq 0 \quad \Rightarrow \quad |x_i^k| > \tau; \tag{90}$$

hence, this, together with (89), implies (87), as desired.

Finally, we complete the proof by showing (90). Fix $k > N$ and $i \in \mathbb{N}$, and suppose that $x_i^k \neq 0$. Then, it follows from (86) and (H2°) that

$$\underline{\lambda} p |x_i^k|^{p-1} + 2A_i^\top (Ax^k - b) \leq \|w^k\| \leq \beta\|x^k - x^{k-1}\| + \epsilon_k < 2$$

[due to (89) and $\tau \leq \frac{1}{\beta}$ by (88)]. Noting that $\|x^k\| \leq M$, we obtain from the above relation that

$$|x_i^k| > \left(\frac{\underline{\lambda} p}{2 + 2\|A\|^2 M + 2\|A\|\|b\|}\right)^{1-p} \geq \tau$$

[by (88)], which verifies (90), as desired. □

Remark 7 (i) Problem (85) for the n-dimensional Euclidean space has an equivalent formula to that of problem (1). Indeed, let $u_i := \left(\frac{\lambda_i}{\underline{\lambda}}\right)^{\frac{1}{p}} x_i$ and $K_i := \left(\frac{\underline{\lambda}}{\lambda_i}\right)^{\frac{1}{p}} A_i$ for $i = 1, \ldots, n$. Then, problem (85) is reformulated to $\min_{u\in\mathbb{R}^n} \ \|Ku - b\|^2 + \underline{\lambda}\|u\|_p^p$ that is (1) with K and u in place of A and x.

(ii) It is easy to verify by the similar proofs that Theorem 1 and Corollary 2 are also true for problem (85) in the infinite-dimensional Hilbert spaces.

Theorem 7 *Let* $\{x^k\} \subseteq \ell^2$ *be a sequence satisfying* (H1) *and* (H2). *Then* $\{x^k\}$ *converges to a critical point* x^* *of problem* (85). *Suppose that* x^* *is a local minimum of problem* (85). *Then* $\{x^k\}$ *converges linearly to* x^*.

Proof By the assumptions, it follows from Proposition 3 that there exist $N \in \mathbb{N}$ and a finite index set J such that (87) is satisfied. Let $f_J : \mathbb{R}^{|J|} \to \mathbb{R}$ be a function denoted by

$$f_J(y) := \|A_J y - b\|^2 + \sum_{i \in J} \lambda_i |y_i|^p \quad \text{for any } y \in \mathbb{R}^{|J|}.$$

By the assumptions and (87), we can check that $\{x_J^k\}_{k \geq N}$ satisfies (H1) and (H2) with x_J^k and f_J in place of x^k and F. Hence, the convergence of $\{x_J^k\}$ to a critical point x_J^* of f_J directly follows Theorem 3. Let $x_{J_c}^* = 0$. Then, by (87), it follows that $\{x^k\}$ converges to this x^*, which is a critical point of problem (85). Furthermore, suppose that x^* is a local minimum of problem (85). Then x_J^* is also a local minimum of f_J, and so, the linear convergence of $\{x_J^k\}$ to x_J^* also follows from Theorem 3. Then, by (87), we conclude that $\{x^k\}$ converges linearly to this x^*. □

Theorem 8 *Let $\{x^k\} \subseteq \ell^2$ be a sequence satisfying* (H1°) *and $\{\epsilon^k\}$ satisfy* (36). *Suppose one of limiting points of $\{x^k\}$, denoted by x^*, is a local minimum of problem* (85). *Then the following assertions are true.*

(i) *$\{x^k\}$ converges to x^*.*
(ii) *Suppose further that $\{x^k\}$ satisfies* (H2°) *and $\{\epsilon^k\}$ satisfies* (41). *Then $\{x^k\}$ converges linearly to x^*.*

Proof The proofs of assertions (i) and (ii) of this theorem use the lines of analysis similar to that of assertion (i) of Theorem 2 (recalling from Remark 7(ii) that Corollary 2 is true for the infinite-dimensional cases) and that of Theorem 7, respectively. The details are omitted. □

Bredies et al. [10] investigated the PGA for solving problem (85) in infinite-dimensional Hilbert spaces and proved that the generated sequence converges to a critical point under the following additional assumptions: (a) $\{x \in \ell^2 : A^\top Ax = \|A^\top A\|x\}$ is finite dimensional, (b) $\|A^\top A\|$ is not an accumulation point of the eigenvalues of $A^\top A$, (c) A satisfies a finite basis injectivity property, and (d) p is a rational. Dropping these technical assumptions, we prove the global convergence of the PGA only under the common made assumption on stepsizes, which significantly improves [10, Theorem 5.1], and further establish its linear convergence under a simple additional assumption in the following theorem. Recall from [2, Theorem 5.1] that the sequence $\{x^k\}$ generated by Algorithm PGA satisfies conditions (H1) and (H2) under the assumption (59). Hence, as an application of Theorem 7, the results in the following theorem directly follow.

Theorem 9 *Let $\{x^k\} \subseteq \ell^2$ be a sequence generated by Algorithm PGA with $\{v_k\}$ satisfying* (59). *Then $\{x^k\}$ converges to a critical point x^* of problem* (85). *Furthermore, suppose that x^* is a local minimum of problem* (85). *Then $\{x^k\}$ converges linearly to x^*.*

Let x^* be a local minimum of problem (85). It was reported in [17, Theorem 2.1(i)] that

$$|x_i^*| \geq \left(\frac{\underline{\lambda} p(1-p)}{2\|A_i\|^2} \right)^{\frac{1}{2-p}} \quad \text{for each } i \in \operatorname{supp}(x^*).$$

This indicates that $\operatorname{supp}(x^*)$ is a finite index set. Then, following the proof lines of Theorems 4–6, we can obtain the linear convergence of inexact PGAs for infinite-dimensional Hilbert spaces, which are provided as follows.

Theorem 10 *Let $\{x^k\} \subseteq \ell^2$ be a sequence generated by Algorithm IPGA-I with $\{v_k\}$ satisfying* (59). *Then the following assertions are true.*

(i) *Suppose that* (36) *is satisfied, and that one of limiting points of* $\{x^k\}$*, denoted by* x^**, is a local minimum of problem* (85)*. Then* $\{x^k\}$ *converges to* x^*.
(ii) *Suppose that* $\{x^k\}$ *converges to a global minimum* x^* *of problem* (85) *and that* (60) *and* (61) *are satisfied for each* $k \in \mathbb{N}$ *with* $\lim_{k\to\infty} \tau_k = 0$*. Then* $\{x^k\}$ *converges linearly to* x^*.

Theorem 11 *Let* $\{x^k\} \subseteq \ell^2$ *be a sequence generated by Algorithm IPGA-II with* $\{v_k\}$ *satisfying* (59)*. Suppose that* $\{x^k\}$ *converges to a local minimum* x^* *of problem* (85) *and that* (62) *and* (63) *are satisfied for each* $k \in \mathbb{N}$ *with* $\lim_{k\to\infty} t_k = 0$*. Then* $\{x^k\}$ *converges linearly to* x^*.

Remark 8 Algorithms IPGA-Ip and IPGA-IIp, the parallel versions of Algorithms IPGA-I and IPGA-II, are implementable for solving problem (85) in the infinite-dimensional Hilbert spaces, and the generated sequences share the same linear convergence properties as shown in Theorems 10 and 11, respectively.

Acknowledgements The authors are grateful to the editor and the anonymous reviewer for their valuable comments and suggestions toward the improvement of this paper. Yaohua Hu's work was supported in part by the National Natural Science Foundation of China (12071306, 11871347), Natural Science Foundation of Guangdong Province of China (2019A1515011917, 2020B1515310008), Project of Educational Commission of Guangdong Province of China (2019KZDZX1007), Natural Science Foundation of Shenzhen (JCYJ20190808173603590, JCYJ20170817100950436) and Interdisciplinary Innovation Team of Shenzhen University. Chong Li's work was supported in part by the National Natural Science Foundation of China (11971429) and Zhejiang Provincial Natural Science Foundation of China (LY18A010004). Kaiwen Meng's work was supported in part by the National Natural Science Foundation of China (11671329) and the Fundamental Research Funds for the Central Universities (JBK1805001). Xiaoqi Yang's work was supported in part by the Research Grants Council of Hong Kong (PolyU 152342/16E).

References

1. Attouch, H., Bolte, J., Redont, P., Soubeyran, A.: Proximal alternating minimization and projection methods for nonconvex problems: an approach based on the Kurdyka–Łojasiewicz inequality. Math. Oper. Res. **35**, 438–457 (2010)
2. Attouch, H., Bolte, J., Svaiter, B.F.: Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized Gauss–Seidel methods. Math. Program. **137**, 91–129 (2013)
3. Bach, F., Jenatton, R., Mairal, J., Obozinski, G.: Structured sparsity through convex optimization. Stat. Sci. **27**, 450–468 (2012)
4. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imaging Sci. **2**, 183–202 (2009)
5. Bertsekas, D.P.: Nonlinear Programming. Athena Scientific, Cambridge (1999)
6. Blumensath, T., Davies, M.E.: Iterative thresholding for sparse approximations. J. Fourier Anal. Appl. **14**, 629–654 (2008)
7. Bolte, J., Nguyen, T.P., Peypouquet, J., Suter, B.W.: From error bounds to the complexity of first-order descent methods for convex functions. Math. Program. **2016**, 1–37 (2016)
8. Bolte, J., Sabach, S., Teboulle, M.: Proximal alternating linearized minimization for nonconvex and nonsmooth problems. Math. Program. **146**, 459–494 (2013)
9. Bredies, K., Lorenz, D.A.: Linear convergence of iterative soft-thresholding. J. Fourier Anal. Appl. **14**, 813–837 (2008)
10. Bredies, K., Lorenz, D.A., Reiterer, S.: Minimization of non-smooth, non-convex functionals by iterative thresholding. J. Optim. Theory App. **165**, 78–112 (2015)
11. Burachik, R.S., Rubinov, A.: Abstract convexity and augmented Lagrangians. SIAM J. Optim. **18**, 413–436 (2007)
12. Byrd, R.H., Nocedal, J., Oztoprak, F.: An inexact successive quadratic approximation method for $L-1$ regularized optimization. Math. Program. **157**, 375–396 (2016)
13. Candès, E., Tao, T.: Decoding by linear programming. IEEE Trans. Inform. Theory **51**, 4203–4215 (2005)

14. Cao, W., Sun, J., Xu, Z.: Fast image deconvolution using closed-form thresholding formulas of L_q ($q = \frac{1}{2}, \frac{2}{3}$) regularization. J. Vis. Commun. Image R. **24**, 31–41 (2013)
15. Chartrand, R., Staneva, V.: Restricted isometry properties and nonconvex compressive sensing. Inverse Probl. **24**, 1–14 (2008)
16. Chen, X.: Smoothing methods for nonsmooth, nonconvex minimization. Math. Program. **134**, 71–99 (2012)
17. Chen, X., Xu, F., Ye, Y.: Lower bound theory of nonzero entries in solutions of ℓ_2–ℓ_p minimization. SIAM J. Sci. Comput. **32**, 2832–2852 (2010)
18. Combettes, P.L., Wajs, V.R.: Signal recovery by proximal forward–backward splitting. Multiscale Model. Sim. **4**, 1168–1200 (2005)
19. Daubechies, I., Defrise, M., De Mol, C.: An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. Commun. Pur. Appl. Math. **57**, 1413–1457 (2004)
20. Donoho, D.L.: Compressed sensing. IEEE Trans. Inform. Theory **52**, 1289–1306 (2006)
21. Elad, M.: Sparse and Redundant Representations. Springer, New York (2010)
22. Frankel, P., Garrigos, G., Peypouquet, J.: Splitting methods with variable metric for Kurdyka–Lojasiewicz functions and general convergence rates. J. Optim. Theory Appl. **165**, 874–900 (2015)
23. Ge, D., Jiang, X., Ye, Y.: A note on complexity of L_p minimization. Mathm. Program. **129**, 285–299 (2011)
24. Hu, Y., Li, C., Meng, K., Qin, J., Yang, X.: Group sparse optimizatin via $\ell_{p,q}$ regularization. J. Mach. Learn. Res **18**, 1–52 (2017)
25. Hu, Y., Li, C., Yang, X.: On convergence rates of linearized proximal algorithms for convex composite optimization with applications. SIAM J. Optim. **26**, 1207–1235 (2016)
26. Huang, X., Yang, X.: A unified augmented Lagrangian approach to duality and exact penalization. Math. Oper. Res. **28**, 533–552 (2003)
27. Lai, M., Wang, J.: An unconstrained ℓ_q minimization with $0<q \leq 1$ for sparse solution of underdetermined linear systems. SIAM J. Optim. **21**, 82–101 (2011)
28. Lee, J.D., Sun, Y., Saunders, M.A.: Proximal Newton-type methods for minimizing composite functions. SIAM J. Optim. **24**, 1420–1443 (2014)
29. Li, G., Pong, T.K.: Douglas–Rachford splitting for nonconvex optimization with application to nonconvex feasibility problems. Math. Program. **159**, 1–31 (2015)
30. Li, G., Pong, T.K.: Global convergence of splitting methods for nonconvex composite optimization. SIAM J. Optim. **25**, 2434–2460 (2015)
31. Lu, Z., Zhang, Y., Lu, J.: ℓ_p Regularized low-rank approximation via iterative reweighted singular value minimization. Comput. Optim. Appl. **68**, 619–642 (2017)
32. Lu, Z., Zhang, Y.: Sparse approximation via penalty decomposition methods. SIAM J. Optim. **23**, 2448–2478 (2013)
33. Lu, J., Qiao, K., Li, X., Zou, Y., Lu, Z.: ℓ_0-minimization methods for image restoration problems based on wavelet frames. Inverse Probl. **35**, 064001 (2019)
34. Luo, Z., Pang, J., Ralph, D.: Mathematical Programs with Equilibrium Constraints. Cambridge University Press, Cambridge (1996)
35. Mairal, J.: Incremental majorization-minimization optimization with application to large-scale machine learning. SIAM J. Optim. **25**, 829–855 (2015)
36. Marjanovic, G., Solo, V.: On l_q optimization and sparse inverse covariance selection. IEEE Trans. Sig. Proc. **62**, 1644–1654 (2014)
37. Nesterov, Y.: Gradient methods for minimizing composite functions. Math. Program. **140**, 125–161 (2013)
38. Ochs, P., Chen, Y., Brox, T., Pock, T.: iPiano: inertial proximal algorithm for nonconvex optimization. SIAM J. Imaging Sci. **7**, 1388–1419 (2014)
39. Nikolova, M.: Description of the minimizers of least squares regularized with ℓ_0-norm. Uniqueness of the global minimizer. SIAM J. Imaging Sci. **6**, 904–937 (2013)
40. Pant, J.K., Lu, W.S., Antoniou, A.: New improved algorithms for compressive sensing based on ℓ_p norm. IEEE Trans. Circuits II **61**, 198–202 (2014)
41. Qin, J., Hu, Y.H., Xu, F., Yalamanchili, H.K., Wang, J.: Inferring gene regulatory networks by integrating ChIP-seq/chip and transcriptome data via LASSO-type regularization methods. Methods **67**, 294–303 (2014)
42. Razaviyayn, M., Hong, M., Luo, Z.: A unified convergence analysis of block successive minimization methods for nonsmooth optimization. SIAM J. Optim. **23**, 1126–1153 (2013)
43. Rockafellar, R.T.: Monotone operators and the proximal point algorithm. SIAM J. Control Optim. **14**, 877–898 (1976)
44. Rockafellar, R.T., Wets, R.J.-B.: Variational Analysis. Springer, Berlin (1998)

45. Schmidt, M., Roux, N.L., Bach, F.: Convergence rates of inexact proximal-gradient methods for convex optimization. Adv. Neural Inf. Process. Syst. **24**, 1458–1466 (2011)
46. Simon, N., Friedman, J., Hastie, T., Tibshirani, R.: A sparse-group Lasso. J. Comput. Graph. Stat. **22**, 231–245 (2013)
47. Tao, S., Boley, D., Zhang, S.: Local linear convergence of ISTA and FISTA on the LASSO problem. SIAM J. Optim. **26**, 313–336 (2016)
48. Tseng, P.: Approximation accuracy, gradient methods, and error bound for structured convex optimization. Math. Program. **125**, 263–295 (2010)
49. Tseng, P., Yun, S.: A coordinate gradient descent method for nonsmooth separable minimization. Math. Program. **117**, 387–423 (2009)
50. Wang, J., Hu, Y., Li, C., Yao, J.-C.: Linear convergence of CQ algorithms and applications in gene regulatory network inference. Inverse Probl. **33**, 055017 (2017)
51. Wang, J., Li, C., Lopez, G., Yao, J.-C.: Proximal point algorithms on Hadamard manifolds: linear convergence and finite termination. SIAM J. Optim. **26**, 2696–2729 (2017)
52. Wen, B., Chen, X., Pong, T.K.: Linear convergence of proximal gradient algorithm with extrapolation for a class of nonconvex nonsmooth minimization problems. SIAM J. Optim. **27**, 124–145 (2017)
53. Xiao, L., Zhang, T.: A proximal-gradient homotopy method for the sparse least-squares problem. SIAM J. Optim. **23**, 1062–1091 (2013)
54. Xu, Y., Yin, W.: A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. SIAM J. Imaging Sci. **6**, 1758–1789 (2013)
55. Xu, Z., Chang, X., Xu, F., Zhang, H.: $L_{1/2}$ regularization: a thresholding representation theory and a fast solver. IEEE Trans. Neur. Net. Lear. **23**, 1013–1027 (2012)
56. Yang, J., Zhang, Y.: Alternating direction algorithms for ℓ_1-problems in compressive sensing. SIAM J. Sci. Comput. **33**, 250–278 (2011)
57. Zeng, J., Lin, S., Xu, Z.: Sparse regularization: convergence of iterative jumping thresholding algorithm. IEEE Trans. Sig. Proc. **64**, 5106–5118 (2016)
58. Zhang, H., Jiang, J., Luo, Z.-Q.: On the linear convergence of a proximal gradient method for a class of nonsmooth convex minimization problems. J. Oper. Res. Soc. China **1**, 163–186 (2013)
59. Zhang, L., Hu, Y., Li, C., Yao, J.-C.: A new linear convergence result for the iterative soft thresholding algorithm. Optimization **66**, 1177–1189 (2017)
60. Zhang, L., Hu, Y., Yu, C.K.W., Wang, J.: Iterative positive thresholding algorithm for nonnegative sparse optimization. Optimization **67**, 1345–1363 (2018)
61. Zhang, T.: Analysis of multi-stage convex relaxation for sparse regularization. J. Mach. Learn. Res. **11**, 1081–1107 (2010)