An extrapolated iteratively reweighted $\ell_1$ method with complexity analysis

Hao Wang$^1$ · Hao Zeng$^1$ · Jiashan Wang$^2$

Received: 12 July 2020 / Accepted: 19 September 2022 / Published online: 5 October 2022
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

Abstract
The iteratively reweighted $\ell_1$ algorithm is a widely used method for solving various regularization problems, which generally minimize a differentiable loss function combined with a convex/nonconvex regularizer to induce sparsity in the solution. However, the convergence and the complexity of iteratively reweighted $\ell_1$ algorithms is generally difficult to analyze, especially for non-Lipschitz differentiable regularizers such as $\ell_p$ norm regularization with $0 < p < 1$. In this paper, we propose, analyze and test a reweighted $\ell_1$ algorithm combined with the extrapolation technique under the assumption of Kurdyka-Łojasiewicz (KL) property on the proximal function of the perturbed objective. Our method does not require the Lipschitz differentiability on the regularizers nor the smoothing parameters in the weights bounded away from 0. We show the proposed algorithm converges uniquely to a stationary point of the regularization problem and has local linear convergence for KL exponent at most 1/2 and local sublinear convergence for KL exponent greater than 1/2. We also provide results on calculating the KL exponents and discuss the cases when the KL exponent is at most 1/2. Numerical experiments show the efficiency of our proposed method.

Keywords $\ell_p$ regularization · Extrapolation techniques · Iteratively reweighted methods · Kurdyka-Łojasiewicz · Non-Lipschitz regularization

---

H Hao Wang
haw309@gmail.com

Hao Zeng
zenghao@shanghaitech.edu.cn

Jiashan Wang
jsw1119@gmail.com

$^1$ School of Information Science and Technology, ShanghaiTech University, Shanghai, China
$^2$ Department of Mathematics, University of Washington, Seattle, USA
1 Introduction

Recently, sparse regularization has received increasing contemporary attentions among researchers due to its various important applications, e.g., compressed sensing, machine learning, and image processing [14, 18, 26, 28–30, 43]. The goal of sparse regularization is to find sparse solutions of a mathematical model such that the model performance can be better generalized to future data. A common approach of this regularization technique is to add a regularizer term to the objective such that most of the components in the resulted solution are zero. In this paper, we focus on the $\ell_p$-norm regularization optimization problem of the following form

$$\min_{x \in \mathbb{R}^n} F(x) := f(x) + \lambda \|x\|_p^p,$$  \hspace{1cm} (P)

where $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable, $p \in (0, 1)$ and $\lambda > 0$ is the prescribed regularization parameter. The $\ell_p$ norm is defined as $\|x\|_p := (\sum_{i=1}^n |x_i|^p)^{1/p}$. Compared with the $\ell_1$ regularizer, $\ell_p$ regularizer is often believed to be a better approximation to the $\ell_0$ regularizer, i.e., the number of nonzeros of the involved vector. However, due to the nonsmooth and non-Lipschitz differentiable nature of the $\ell_p$-norm, this problem is difficult to handle and analyze. In fact, it has been proven in [16] to be strongly NP-hard.

The iteratively reweighted $\ell_1$ (IRL1) algorithm [10–12, 24, 37, 40] has been widely applied to solve regularization problems to induce sparsity in the solutions. It can easily handle various regularization terms including $\ell_p$-norm, log-sum [23], SCAD [13] and MCP [44] by approximating regularizer with a weighted $\ell_1$ norm in each iteration. For example, the technique proposed by Chen [12] and Lai [19] adds smoothing perturbation $\epsilon$ to each $|x_i|$ to formulate the $\epsilon$-approximation of the $\ell_p$ norm. In this case, the objective is replaced by

$$f(x) + \lambda \sum_{i=1}^n (|x_i| + \epsilon_i)^p,$$  \hspace{1cm} (1)

with prescribed $\epsilon > 0$. At each iteration $x^k$, the iteratively reweighted $\ell_1$ method solves the subproblem in which each $(|x_i| + \epsilon)^p$ is replaced by its linearization

$$p(|x_i^k| + \epsilon_i)^{p-1}|x_i|.$$  \hspace{1cm} (2)

In this case, large $\epsilon$ can smooth out many local minimizers, while small values make the subproblems difficult to solve and the algorithm easily get trapped into bad local minimizers. To obtain an accurate approximation of (P), Lu [25] proposed a dynamic updating strategy to drive $\epsilon$ from an initial relatively large value to 0 as $k \to \infty$. Recently, Wang et al. [37] show the property that the iterates generated by the IRL1 algorithm have local stable sign value. Based on this, they also present a novel updating strategy for $\epsilon_i$ that only drive $\epsilon_i$ associated with the nonzeros in the limit point to 0 while keeping others bounded away from 0.

Since Nesterov [34] first proposed the extrapolation techniques in gradient method, many works focus on analyzing and improving the convergence rate of the IRL1
algorithms. In this approach, a linear combination of previous two steps are used to update next step. Nesterov’s extrapolation techniques [31–34] have also been widely applied to accelerate the performance of the first-order methods and convex composite optimization problems, for example [4, 7, 20] and [35]. During the past decade, it is proven to be a successful accelerating approach when applied to various algorithms. For example, Beck and Teboulle [6] presented a fast iterative shrinkage-thresholding algorithm (FISTA) using this technique.

As for IRL1 methods, Yu and Pong [40] proposed several versions of IRL1 algorithms with extrapolation and analyzed the global convergence.

In view of the success of IRL1 combined with the extrapolation techniques, in this paper we propose and analyze the Extrapolated Proximal Iteratively Reweighted $\ell_1$ method (EPIRL1) to solve the $\ell_p$-norm regularization problem. We show the global convergence and the local complexity of the proposed methods under the Kurdyka-Łojasiewicz (KL) property [8, 9] on the proximal function of the perturbed objective. In particular, we show that our method converges to a first-order optimal solution of the $\ell_p$-regularization problem and local linear convergence rates can be guaranteed if the KL exponent is at most 1/2. Results on calculating the KL exponent are also provided, indicating a KL exponent no greater than 1/2 could happen in many cases.

1.1 Notation

We denote $\mathbb{R}$ and $\mathbb{Q}$ as the set of real numbers and rational numbers. The set $\mathbb{R}^n$ is the real $n$-dimensional Euclidean space with $\mathbb{R}_+^n$ being the positive orthant in $\mathbb{R}^n$ and $\mathbb{R}_{++}^n$ the interior of $\mathbb{R}_+^n$. In $\mathbb{R}^n$, denote $\| \cdot \|_p$ as the $\ell_p$-norm with $p \in (0, +\infty)$, i.e., $\| x \|_p = \left( \sum_{i=1}^n |x_i|^p \right)^{1/p}$. Note that for $p \in (0, 1)$, this does not define a proper norm due to its lack of subadditivity. For convex functions, the set of subgradients are defined as

$$\partial f(a) = \{ z \mid f(\tilde{x}) + \langle z, x - \tilde{x} \rangle \leq f(x), \forall x \in \mathbb{R}^n \}.$$ 

In particular, for $x \in \mathbb{R}^n$, $\partial \| x \|_1 = \{ \xi \in \mathbb{R}^n \mid \xi_i \in \partial |x_i|, i = 1, \ldots, n \}$. Given a lower semi-continuous function $f$, the Frechet subdifferential of $f$ at $a$ is defined as

$$\partial f(a) := \{ z \in \mathbb{R}^n : \liminf_{x \to a} \frac{f(x) - f(a) - \langle z, x - a \rangle}{\| x - a \|_2} \geq 0 \}$$

and the limiting subdifferential at $a \in \text{dom} f$ is defined as

$$\bar{\partial} f(a) := \{ z^* = \lim_{x^k \to a, f(x^k) \to f(a)} z^k, z^k \in \partial f(x^k) \}.$$ 

For any $x \in \mathbb{R}^n$, let $A(x)$ be the set of indices of zero components of $x$ and $I(x)$ be the set of indices of nonzero components of $x$; therefore, $A \cup I = \{ 1, \ldots, n \}$. For $f : \mathbb{R}^n \to \mathbb{R}$, let $f(x|_{I(x)})$ be the function in the reduced space $\mathbb{R}^{|I(x)|}$ by fixing $x_i = 0, i \in A(x)$. For $a, b \in \mathbb{R}^n$, $a \leq b$ means the inequality holds for each component, i.e., $a_i \leq b_i$ for $i = 1, \ldots, n$ and $a \odot b$ is the component-wise product of $a$ and $b$.
1.2 Kurdyka-Łojasiewicz property

Kurdyka-Łojasiewicz property is applicable to a wide range of problems such as nonsmooth semi-algebraic minimization problem [9], and serves as a basic assumption to guarantee the convergence of many algorithms. For example, a series of convergence results for gradient descent methods are proved in [3] under the assumption that the objective satisfies the KL property. The definition of Kurdyka-Łojasiewicz property is given below.

**Definition 1 (Kurdyka-Łojasiewicz property)** The function $f : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$ is said to have the Kurdyka-Łojasiewicz property at $x^* \in \text{dom} \bar{\partial}f$ if there exists $\eta \in (0, +\infty)$, a neighborhood $U$ of $x^*$ and a continuous concave function $\phi : [0, \eta) \rightarrow \mathbb{R}_+$ such that:

(i) $\phi(0) = 0$,
(ii) $\phi$ is $C^1$ on $(0, \eta)$,
(iii) for all $s \in (0, \eta)$, $\phi'(s) > 0$,
(iv) for all $x$ in $U \cap [f(x^*) < f < f(x^*) + \eta]$, the Kurdyka-Łojasiewicz inequality holds

$$\phi'(f(x) - f(x^*)) \text{dist}(0, \bar{\partial}f(x)) \geq 1.$$ 

The definition of the uniform KL property is given below.

**Definition 2 (Uniform KL property)** Suppose $f : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$ is a proper function and let $\Gamma$ be a compact set. If $f$ is a constant on $\Gamma$ and satisfies the KL property at each point of $\Gamma$ with $\phi$ as defined in the KL property, then there exists $\tau, \eta > 0$ such that

$$\phi'(f(x) - f(x^*)) \text{dist}(0, \bar{\partial}f(x)) \geq 1$$

for any $x^* \in \Gamma$ and any $x \in U = \{x \mid \text{dist}(x, \Gamma) < \tau\}$ and $f(x^*) < f(x) < f(x^*) + \eta$. We say $f$ has uniform KL property on $\Gamma$.

Recently, the KL and uniform KL property in Definition 1 have been used to establish the local convergence rates for many optimization methods [2, 15, 21, 39]. Li and Pong [22] summarized the relevant studies and present the following prototypical theorem on convergence rate of algorithms for minimizing functions satisfying the KL inequality:
“For a certain algorithm, consider a suitable objective function satisfying the KL property with an exponent of $\theta \in [0, 1)$, and the bounded sequence $\{x^k\}$ generated by the algorithm. Then the following results hold.

(i) If $\theta = 0$, then $\{x^k\}$ converges finitely.
(ii) If $\theta \in (0, \frac{1}{2}]$, then $\{x^k\}$ converges locally linearly.
(iii) If $\theta \in (\frac{1}{2}, 1)$, then $\{x^k\}$ converges locally sublinearly.”

2 Proximal Iteratively Reweighted $\ell_1$ Method with Extrapolation

In this section, we propose an extrapolated iteratively reweighted $\ell_1$ algorithm, hereinafter named as EIRL1. The framework of this algorithm is presented in Algorithm 1. The updating strategy (6) of $\epsilon$ was proposed in [37].

Algorithm 1 Extrapolated Proximal Iteratively Reweighted $\ell_1$ Algorithm

1: Input: $\mu \in (0, 1)$, $\beta \geq L_f$, $\epsilon^0 \in \mathbb{R}^n_{++}$, $0 \leq \alpha^0 \leq \bar{\alpha} < 1$ and $x^0$, where $\bar{\alpha}$ is defined in (7)
2: Initialize: set $k = 0$, $x^{-1} = x^0$.
3: repeat
4: Compute new iterate:
   \[ w^k_i = \mathbb{P}(|x^k_i| + \epsilon^k_i)^{p-1}, \]
   \[ y^k = x^k + \alpha^k(x^k - x^{k-1}), \]
   \[ x^{k+1} \leftarrow \arg\min_{x \in \mathbb{R}^n} \{ \nabla f(y^k)^T x + \frac{\beta}{2} \| x - y^k \|^2 + \lambda \sum_{i=1}^n w^k_i |x_i| \}, \]
5: Choose $0 \leq \alpha^k \leq \bar{\alpha}$ and $\epsilon^{k+1}$ according to
   \[
   \begin{cases}
   \epsilon^{k+1}_i = \epsilon^k_i & \text{if } x^{k+1}_i = 0, \\
   \epsilon^{k+1}_i \leq \mu \epsilon^k_i & \text{if } x^{k+1}_i \neq 0.
   \end{cases}
   \]
6: Set $k \leftarrow k + 1$.
7: until convergence

Define the smooth approximation $F(x, \epsilon)$ of $F(x)$ with smoothing parameter $\epsilon$ as

$F(x, \epsilon) := f(x) + \lambda \sum_{i=1}^n (|x_i| + \epsilon)^p$ with $F(x, 0) = F(x)$

and define the function of combining the objective with a proximal term as

$\psi(x, y, \epsilon) := F(x, \epsilon) + \frac{\beta}{2} \| x - y \|^2_2$.

We select $\bar{\alpha}$ according to the following
We make the following assumptions about $f$ and $F$.

**Assumption 1**

(i) $f$ is Lipschitz differentiable with constant $L_f \geq 0$.

(ii) The initial point $(x^0, \epsilon^0)$ and $\beta$ are chosen such that $\mathcal{L}(F^0) := \{ x \mid F(x) \leq F^0 \}$ is bounded where $F^0 := F(x^0, \epsilon^0)$ and $\beta \geq L_f$. Suppose $F$ is bounded below by $F$ on $\mathcal{L}(F^0)$.

The following properties hold true for Algorithm 1.

**Lemma 1** Suppose Assumption 1 holds true and $\{x^k\}$ is generated by Algorithm 1 for solving (P). Then the following statements hold.

(i) $\psi(x^k, x^{k-1}, \epsilon^k) - \psi(x^{k+1}, x^k, \epsilon^{k+1}) \geq \bar{\beta}\|x^k - x^{k-1}\|^2$, where

$$
\bar{\beta} = \left\{ \begin{array}{ll}
\bar{\alpha} \frac{1 - \bar{\alpha}^2}{2} & \text{if } f \text{ is convex and Lipschitz differentiable}, \\
\frac{\beta}{2} \left( 1 - \frac{\beta + 3L_f}{\beta} \bar{\alpha}^2 \right) & \text{if } f \text{ Lipschitz differentiable}.
\end{array} \right.
$$

(ii) The sequence $\{x^k\} \subset \mathcal{L}(F^0)$ and is bounded.

(iii) $\frac{\beta}{2} \|x^k - x^k\|^2 < +\infty$; therefore $\lim_{k \to \infty} \|x^{k+1} - x^k\|_2 = 0$ and the set of cluster points of $\{x^k\}$ is connected.

(iv) $\lim_{k \to \infty} \|y^k - x^k\|_2 = 0$ and $\lim_{k \to \infty} \|y^{k-1} - x^k\|_2 = 0$.

**Proof**

(i) Since $x^{k+1}$ is the optimal solution of subproblem (5), there exists $\xi^{k+1} \in \partial|x^{k+1}|$ such that

$$0 = \nabla f(y^k) + \beta(x^{k+1} - y^k) + \lambda w^k \xi^{k+1},
$$

which combined with the strong convexity of (5) yields

$$
\langle \nabla f(y^k), x^{k+1} \rangle + \frac{\beta}{2} \|x^{k+1} - y^k\|^2 + \lambda \sum_{i=1}^{n} w_i |x_i^{k+1}|
\leq \langle \nabla f(y^k), x^k \rangle + \frac{\beta}{2} \|x^k - y^k\|^2 + \lambda \sum_{i=1}^{n} w_i |x_i^k| - \frac{\beta}{2} \|x^{k+1} - x^k\|^2.
$$
An extrapolated iteratively reweighted $L_1$ method with complex…

From the concavity of $a^p$ on $\mathbb{R}_{++}$, we know for any $i \in \{1, \ldots, n\}$

\[
(|x_i^{k+1} + e_i^k|^p \leq (|x_i^k + e_i^k|^p + p(|x_i^k| + e_i^k)^{p-1} (|x_i^{k+1}| - |x_i^k|)) = (|x_i^k| + e_i^k)^p + w_i^k (|x_i^{k+1}| - |x_i^k|).
\]

Summing the above inequality over all $i$ yields

\[
\sum_{i=1}^n (|x_i^{k+1} + e_i^k|^p \leq \sum_{i=1}^n (|x_i^k| + e_i^k)^p + \sum_{i=1}^n w_i^k (|x_i^{k+1}| - |x_i^k|).
\tag{10}
\]

Combining (9) with (10),

\[
\langle \nabla f(y^k), x^{k+1} \rangle + \frac{\beta}{2} \|x^{k+1} - y^k\|^2 + \lambda \sum_{i=1}^n (|x_i^{k+1}| + e_i^k)^p
\leq \langle \nabla f(y^k), x^k \rangle + \frac{\beta}{2} \|x^k - y^k\|^2 + \lambda \sum_{i=1}^n (|x_i^k| + e_i^k)^p - \frac{\beta}{2} \|x^{k+1} - x^k\|^2.
\tag{11}
\]

It then follows that

\[
F(x^{k+1}, e^{k+1})
= f(x^{k+1}) + \lambda \sum_{i=1}^n (|x_i^{k+1}| + e_i^{k+1})^p
\leq f(y^k) + \langle \nabla f(y^k), x^{k+1} - y^k \rangle + \frac{L_f}{2} \|x^{k+1} - y^k\|^2 + \lambda \sum_{i=1}^n (|x_i^{k+1}| + e_i^k)^p
\leq f(y^k) + \langle \nabla f(y^k), x^{k+1} - y^k \rangle + \frac{\beta}{2} \|x^{k+1} - y^k\|^2 + \lambda \sum_{i=1}^n (|x_i^{k+1}| + e_i^k)^p - \frac{\beta}{2} \|x^{k+1} - x^k\|^2
\leq f(y^k) + \frac{3L_f}{2} \|x^k - y^k\|^2 + \lambda \sum_{i=1}^n (|x_i^k| + e_i^k)^p + \frac{\beta}{2} \|x^k - y^k\|^2 - \frac{\beta}{2} \|x^{k+1} - x^k\|^2
= F(x^k, e^k) + \frac{\beta + 3L_f}{2} \|x^k - y^k\|^2 - \frac{\beta}{2} \|x^{k+1} - x^k\|^2,
\tag{12}
\]

where the first inequality follows from the Lipschitz differentiability of $f$, the second inequality is by $\beta \geq L_f$, the third inequality follows from (11) and last inequality follows from Lipschitz differentiability of $f$ and Cauchy-Schwarz inequality. This means that

\[
F(x^{k+1}, e^{k+1}) \leq F(x^k, e^k) + \frac{\beta + 3L_f}{2} (\alpha^k)^2 \|x^k - x^{k-1}\|^2 - \frac{\beta}{2} \|x^{k+1} - x^k\|^2
\]

by the definition of $y^k$, which implies that
\[
\psi(x^k, x^{k-1}, e^k) - \psi(x^{k+1}, x^k, e^{k+1}) = F(x^k, e^k) + \frac{\beta}{2} \|x^k - x^{k-1}\|^2 - \left[F(x^{k+1}, e^{k+1}) + \frac{\beta}{2} \|x^{k+1} - x^k\|^2\right]
\]
\[
\geq \frac{\beta}{2} \left(1 - \frac{\beta + 3L_f}{\beta} (\alpha^k)^2\right) \|x^k - x^{k-1}\|^2
\]
\[
\geq \frac{\beta}{2} \left(1 - \frac{\beta + 3L_f}{\beta} \bar{\alpha}^2\right) \|x^k - x^{k-1}\|^2
\]

by \(\{\alpha^k\} \subset [0, \bar{\alpha}]\). We then deduce from (13) and \(0 < \bar{\alpha} < \sqrt{\frac{\beta}{\beta + 3L_f}}\) that the sequence \(\{F(x^k, e^k) + \frac{\beta}{2} \|x^k - x^{k-1}\|^2\}\) is monotonically decreasing.

Furthermore, if \(f\) is convex, in the fourth inequality (12), we can use convexity of \(f\) to have
\[
F(x^{k+1}, e^{k+1}) \leq f(x^k) + \lambda \sum_{i=1}^{n} (|x^k_i| + e^k_i) \beta + \frac{\beta}{2} \|x^k - y^k\|^2 - \frac{\beta}{2} \|x^{k+1} - x^k\|^2.
\]

Following the same arguments above, we have
\[
\psi(x^k, x^{k-1}, e^k) - \psi(x^{k+1}, x^k, e^{k+1}) \geq \frac{\beta}{2}(1 - \bar{\alpha}^2) \|x^k - x^{k-1}\|^2
\]

We then deduce from above and \(0 < \bar{\alpha} < 1\) that the sequence \(\{F(x^k, e^k) + \frac{\beta}{2} \|x^k - x^{k-1}\|^2\}\) is monotonically decreasing.

This proves part (i).

(ii) With \(x^0 = x^{-1}\), we know that for all \(k \geq 0\),
\[
F(x^k) \leq F(x^k, e^k) \leq F(x^k, e^k) + \frac{\beta}{2} \|x^k - x^{k-1}\|^2 \leq F(x^0, e^0)
\]

Under Assumption 1(ii), we know that \(\{x^k\} \subset L(F^0)\) and is bounded. This completes the proof of part (ii).

(iii) Summing both side of (13) from 0 to \(t\), we obtain that
\[
\bar{\beta} \sum_{k=0}^{t} \|x^k - x^{k-1}\|^2 \leq F(x^0, e^0) - F(x^t, e^t) - \frac{\beta}{2} \|x^{t+1} - x^t\|^2 \leq F(x^0, e^0) - F(x^0, e^0) - F(x^0, e^0) - F(x^0, e^0)
\]
\[
\leq F(x^0, e^0) - F(x^t, e^t) \leq F(x^0, e^0) - F(x^0, e^0) - F(x^0, e^0),
\]
yielding \(\sum_{k=0}^{\infty} \|x^k - x^{k-1}\|^2 < \infty\) and \(\lim_{k \to \infty} \|x^{k+1} - x^k\| = 0\). Based [5, Lemma 2.6], we know the set of cluster points is connected. Therefore, part (iii) is true.

(iv) This part is straightforward by noticing that
\[
y^k - x^k = a^k(x^k - x^{k-1}) \to 0
\]
\[
y^{k-1} - x^k = (x^{k-1} - x^k) + a^{k-1}(x^{k-1} - x^{k-2}) \to 0.
\]
from part (iii). \(\square\)
Remark 1 In case where \( f \) is twice differentiable, we can use Taylor’s theorem to have \( \frac{L}{2} \| x^k - y^k \|^2 \) instead of \( \frac{3L}{2} \| x^k - y^k \|^2 \) in the last inequality of (12), so that we can choose \( \bar{a} \in \left( 0, \sqrt{\frac{\beta}{\beta + t_r}} \right) \) if \( f \) is twice differentiable.

2.1 Locally stable support and sign

Using similar arguments from [37] and Lemma 1, we can also obtain results of local stable support and sign as shown in [37], which are listed below. It shows that after some iteration \( K, \{ x^k \}_{k \geq K} \) stays in the same orthant, and the nonzero components are bounded away from 0.

Theorem 1 Suppose Assumption 1 is true and \( \{ x^k \} \) is generated by Algorithm 1. There then exists \( C > 0 \) and \( K \in \mathbb{N} \) such that the following statements hold true.

(i) If \( w^K_i > C/\lambda \), then \( x^i_k = 0 \) and \( e^K_i = e^K_i \) for all \( k > K \).

(ii) The index sets \( I(x^k) \) and \( A(x^k) \) are fixed for all \( k > K \). Therefore, we can denote \( I^* = I(x^*) \) and \( A^* = A(x^k) \) for any \( k > K \).

(iii) For each \( i \in I^* \) and any \( k > K \),

\[
|x^*_i| \geq \left( \frac{C}{p\lambda} \right)^{\frac{1}{\rho-1}} - e^K_i > 0.
\]  

(iv) For any limit point \( x^* \) of \( \{ x^k \} \), it holds that \( I(x^*) = I^* \), \( A(x^*) = A^* \) and

\[
|x^*_i| \geq \left( \frac{C}{p\lambda} \right)^{\frac{1}{\rho-1}}, \quad i \in I^*.
\]  

(v) There exists \( s \in \{-1, 0, +1\}^n \) such that \( \text{sign}(x^k) \equiv s \) for any \( k > K \).

(vi) For all \( k > K \), \( e^K_i = e^K_i > 0, i \in A^* \) and \( \{ e^K_i \} \setminus 0, i \in I^* \).

Proof The bounded \( \{ x^k \} \) (Lemma 1 (ii)) implies that \( \{ y^k \} \) is also bounded. Then, there must exist \( C > 0 \) such that for any \( k \)

\[
\| \nabla f(y^k) + \beta(x^{k+1} - y^k) \|_{\infty} < C.
\]  

(i) If \( w^K_i > C/\lambda \) for some \( k \in \mathbb{N} \), then the optimality condition (8) implies \( x^{k+1}_i = 0 \). Otherwise we have \(| \nabla f(y^k) + \beta(x^{k+1} - y^k) | = \lambda w^K_i > C \), contradicting (16). By (6), \( e^{k+1}_i = e^k_i \). Monotonicity of \( (\cdot)^{\rho-1} \) and \( 0 + e^{k+1}_i \leq |x^k_i| + e^k_i \) yield

\[
w^k_i + 1 = p \left( 0 + e^{k+1}_i \right)^{\rho-1} \geq p \left( |x^k_i| + e^k_i \right)^{\rho-1} = w^K_i > C/\lambda.
\]
By induction we know that $\forall i \leq k \equiv 0$ and $\varepsilon_i^k = \varepsilon_i^K$ for any $k > K = \bar{k}$. This completes the proof of (i).

(ii) Suppose by contradiction this statement is not true. There exists $j \in \{1, \ldots, n\}$ such that $\{x_j^k\}$ takes zero and nonzero values both for infinite times. Hence, there exists a subsequence $S_1 \cup S_2 = \mathbb{N}$ such that $|S_1| = \infty, |S_2| = \infty$ and that

$$x_j^k = 0, \forall k \in S_1 \quad \text{and} \quad x_j^k \neq 0, \forall k \in S_2.$$ 

By (6), $\varepsilon_j^{k+1} < \mu \varepsilon_j^k$ for any $k \in S_2$. Therefore, $\lim_{k \to \infty} \varepsilon_j^{k+1} = 0$ due to $|S_2| = \infty$. Hence, there exists $\bar{k} \in S_1$ such that

$$w_j^k = p\left(\left|x_j^k + \varepsilon_j^k\right|^{p-1}\right) = p\left(\varepsilon_j^k\right)^{p-1} > C/\lambda.$$ 

It follows that $x_j^k \equiv 0$ for any $k > \bar{k}$ by (i) which implies $\{\bar{k} + 1, \bar{k} + 2, \ldots\} \subset S_1$ and $|S_2| < \infty$. This contradicts the assumption $|S_2| = \infty$. Hence, (ii) is true.

(iii) Combining (i) and (ii), we know for any $i \in \mathcal{I}^*$, $w_i^k \leq C/\lambda$, which is equivalent to (14). This proves (iii).

(iv) By (ii), $\mathcal{A}^* \subseteq \mathcal{A}(x^*)$ for any limit point $x^*$. It follows from (ii) and (iii) that $\mathcal{I}^* \subseteq \mathcal{I}(x^*)$ for any limit point $x^*$. Hence, $\mathcal{A}^* = \mathcal{A}(x^*)$ and $\mathcal{I}^* = \mathcal{I}(x^*)$ for any limit point $x^*$ since $\mathcal{A}^* \cup \mathcal{I}^* = \{1, \ldots, n\}$.

(v) By (iii) and Lemma 1(iii), there exists sufficiently large $\bar{k}$, such that for any $k > \bar{k}$

$$|x_i^k| > \epsilon := \frac{1}{2} \left(\frac{C}{p\lambda}\right)^{-\frac{1}{p-1}}, \quad \forall i \in \mathcal{I}^*.$$ 

(17)

and

$$\|x_i^{k+1} - x_i^k\|_2 < \epsilon$$

(18)

We prove (v) by contradiction. Assume there exists $j \in \mathcal{I}^*$ such that the sign of $x_j$ changes after $\bar{k}$. Hence there must be $\bar{k} \geq \bar{k}$ such that $x_j^\bar{k} x_j^{k+1} < 0$. It follows that

$$\|x^{k+1} - x^K\|_2 \geq |x_j^{k+1} - x_j^k| = \sqrt{\left(x_j^{k+1}\right)^2 + \left(x_j^k\right)^2 - 2x_j^k x_j^{k+1}} > \sqrt{\epsilon^2 + \epsilon^2} = \sqrt{2}\epsilon,$$

where the last inequality is by (17). This contradicts with (18); hence $\{x_j^k\}_{k \geq \bar{k}}$ have the same sign. Without loss of generality, we can reselect $K = \bar{k}$ and then (v) holds true.

(vi) This is trivially true by (ii) and the updating strategy (6). \qed
3 Global convergence

Defining $\chi$ as the set of all cluster points of $\{x^k\}$, we now show the global convergence of Algorithm 1, meaning every $x^* \in \chi$ satisfies the first-order stationary condition \[25, 37\]

\[
\nabla f(x^*) + \lambda p |x_i^*|^{p-1} \text{sign}(x_i^*) = 0, \quad i \in \mathcal{I}(x^*).
\]

**Theorem 2** (Global convergence) Suppose Assumption 1 holds true and $\{x^k\}$ is generated by Algorithm 1.

The following statements hold true

(i) $F$ attains the same value at every cluster point of $\{x^k\}$, i.e., there exists $\zeta \in \mathbb{R}$ such that $F(x^*) = \zeta$ for any $x^* \in \chi$.

(ii) Each point $x^* \in \chi$ is a stationary point of $F(x)$.

**Proof**

(i) For any $x^* \in \chi$ with subsequence $\{x^k\}_S \rightarrow x^*$, from Lemma 1 we know that

\[
\lim_{k \rightarrow \infty} F(x^k, e^k) = \lim_{k \rightarrow \infty} k \in S F(x^k, e^k) + \frac{\beta}{2} ||x^k - x^{k-1}||^2 = \lim_{k \rightarrow \infty} k \in S \psi(x^k, x^{k-1}, e^k),
\]

by the monotonicity of $\psi(x^k, x^{k-1}, e^k)$ from Lemma 1 (i), we know there is a unique limit value $\zeta$, i.e. $\zeta = \lim_{k \rightarrow \infty} k \in S F(x^k, e^k)$ and $F(x^*; 0) = \zeta$ for any $x^* \in \chi$.

(ii) Let $x^*$ be a limit point of $\{x^k\}$ with subsequence $\{x^k\}_S \rightarrow x^*$. We have for any $i \in \mathcal{I}(x^*)$,

\[
\nabla f(x^*) + \lambda p |x_i^*|^{p-1} \text{sign}(x_i^*) = \lim_{k \rightarrow \infty} k \in S \nabla f(x^k) + \lambda p |x_i^k|^{p-1} \text{sign}(x_i^k) = \lim_{k \rightarrow \infty} k \in S \nabla f(x^k) + \lambda p (|x_i^k| + e_i^k)^{p-1} \text{sign}(x_i^{k+1}) = \lim_{k \rightarrow \infty} k \in S -\beta (x_i^{k+1} - y_i^k) = 0,
\]
the second equality is by Theorem 1 (vi), the third equality is due to \( x^{k+1} \) satisfying the optimal condition of the subproblem for \( k > K \)
\[
\nabla f(y^i) + \beta (x^{k+1} - y^i) + \lambda \rho (|x^i_k| + \epsilon^k_i)^p - 1 \text{sign}(x^i_k) = 0, \quad i \in I(x^*)
\]
and last equality is by Lemma 1(iv).

Therefore, \( x^* \) is stationary, completing the proof. \( \square \)

To further analyze the property of \( \{ (x^k, \epsilon^k) \} \), denote \( \epsilon_i = \delta_i^2 \) since \( \epsilon_i \) is restricted to be nonnegative and write \( F \) and \( \psi \) as functions of \( (x, \delta) \) for simplicity, i.e.,
\[
F(x, \delta) = f(x) + \lambda \sum_{i=1}^n (|x_i| + \delta_i^2)^p,
\]
\[
\psi(x, y, \delta) = f(x) + \lambda \sum_{i=1}^n (|x_i| + \delta_i^2)^p + \frac{\lambda}{2} \| x - y \|_2^2.
\]

Next, we show the uniqueness of the limit points of \( \{ x^k \} \) under KL property. Notice that after the \( k \)th iteration, the iterates \( \{ x^k \} \) remains in the interior of the same orthant of \( \mathbb{R}^{|I^*|} \) and are bounded away from the axes by Theorem 1. Moreover, \( \delta_{A^*} \) remains as constants. Then \( \psi \) reverts to a function of \( (x_{I^*}, y_{I^*}, \delta_{I^*}) \) for sufficiently \( k \), i.e., \( (x_{I^*}, y_{I^*}, 0_{I^*}) \mapsto \psi(x_{I^*}, y_{I^*}, 0_{I^*}) \),
\[
\psi(x_{I^*}, y_{I^*}, \delta_{I^*}) = f(x_{I^*}) + \lambda \sum_{i \in I^*} (|x_i| + \delta_i^2)^p + \frac{\lambda}{2} \| x_{I^*} - y_{I^*} \|_2^2.
\]

Hence we can assume the reduced function \( \psi(x_{I^*}, y_{I^*}, \delta_{I^*}) \) has the KL property at \( (x^*, x^*, 0_{I^*}) \in \mathbb{R}^{|I^*|} \).

**Assumption 2** Suppose the reduced function \( (x_{I^*}, y_{I^*}, \delta_{I^*}) \mapsto \psi(x_{I^*}, \delta_{I^*}, \delta_{I^*}) \) has the uniform KL property 2 at every
\[
(x^*, x^*, 0_{I^*} \in \mathbb{R}^{|I^*|}, \quad \forall x^* \in \chi.
\]

**Lemma 2** Let \( \{ x^k \} \) be a sequence generated by Algorithm 1 and assumptions 1 and 2 are satisfied.

*For sufficiently large \( k \), the following statements hold.*

(i) \( \text{There exists } D_1 > 0 \text{ such that for all } k \)
\[
\| \nabla \psi(x^k_{I^*}, x^{k-1}_{I^*}, \delta_k^{I^*}) \|_2 \\
\leq D_1 (\| x^k_{I^*} - x^{k-1}_{I^*} \|_2 + \| x_{I^*}^{k-1} - x_{I^*}^{k-2} \|_2 + \| \delta_{I^*}^{k-1} \|_1 - \| \delta_{I^*}^{k} \|_1),
\]
so that \( \lim \| \nabla \psi(x^k_{I^*}, x^{k-1}_{I^*}, \delta_k^{I^*}) \| = 0. \)

(ii) \( \{ \psi(x^k_{I^*}, x^{k-1}_{I^*}, \delta_k^{I^*}) \} \) is monotonically decreasing. There exists \( D_2 > 0 \) such that
\[ ψ(x_k^T, x_{T-1}^k, δ_T^k) - ψ(x_{T}^{k+1}, x_{T}^k, δ_T^{k+1}) ≥ D_2 \| x_T^k - x_{T-1}^k \|^2. \]

(iii) \( ψ(x_*, x_*, 0^T) = \zeta := \lim_{k \to \infty} ψ(x^k_T, x_{T-1}^k, δ_T^k), \) where \( Γ \) is the set of the cluster points of \( \{x^k_T, x_{T-1}^k, δ_T^k\} \), i.e., \( Γ := \{(x_*^T, x_*^k, 0^T) \mid x_* \in \chi\}. \)

**Proof** By Theorem 1, for all sufficiently large \( k \), the components of \( \{x^k_T\} \) are all uniformly bounded away from 0, \( x_{A^c}^k \equiv 0 \) and that \( δ_{A^c}^k \) are fixed. Since all the remaining results in this subsection are properties for sufficiently large \( k \), without loss of generality and for simplicity, we can assume

\[ \mathcal{I}^* = \{1, \ldots, n\} \quad \text{and} \quad A^* = \emptyset, \]

so that all the components of \( \{x^k\} \) for sufficiently large \( k \) are bounded away from 0, i.e. \( (x^k, x_{k-1}^k, δ^k) = (x_*^T, x_*^{k-1}, δ_*^k). \)

(i) Notice for sufficiently large \( k \), the gradient of \( ψ \) at \( (x^k, x_{k-1}^k, δ^k) \) is

\[
\begin{align*}
\nabla_x ψ(x^k, x_{k-1}^k, δ^k) &= \nabla f(x^k) + β(x^k - x_{k-1}^k) + λ w^k \text{sign}(x^k), \\
\nabla_y ψ(x^k, x_{k-1}^k, δ^k) &= -β(x^k - x_{k-1}^k), \\
\nabla_δ ψ(x^k, x_{k-1}^k, δ^k) &= 2λ w^k \circ δ^k.
\end{align*}
\]

We first derive the upper bound for \( \|\nabla_x ψ(x^k, x_{k-1}^k, δ^k)\|_2 \). The first-order optimality condition of the \((k-1)\)th subproblem at \( x^k \) is

\[ \nabla f(y^{k-1}) + β(y^{k-1} - y^{k-1}) + λ w^{k-1} \circ \text{sign}(x^k) = 0. \]

Hence, we have

\[
\begin{align*}
\nabla_x ψ(x^k, x_{k-1}^k, δ^k) &= \nabla f(x^k) - \nabla f(y^{k-1}) \\
&+ β(y^{k-1} - y^{k-1}) + λ (w^k - w^{k-1}) \circ \text{sign}(x^k). \tag{23}
\end{align*}
\]

By Lemma 1(iv) and the Lipschitz differentiability of \( f \), we know

\[
\|\nabla f(x^k) - \nabla f(y^{k-1})\|_2 ≤ L_f \|x^k - y^{k-1}\|_2,
\]

\[
= L_f \|x^k - x_{k-1}^k - α^{k-1}(x_{k-1}^k - x_{k-2}^2)\|_2
\]

\[ = L_f \|x^k - x_{k-1}^k\|_2 + L_f \tilde{α} \|x_{k-1}^k - x_{k-2}^2\|_2, \tag{24}\]

and that

\[
\|β(y^{k-1} - y^{k-1})\|_2 = β \|y^{k-1} + α^{k-1}(x_{k-1}^k - x_{k-2}^2) - x_{k-1}^k\|_2 ≤ β \tilde{α} \|x_{k-1}^k - x_{k-2}^2\|_2. \tag{25}\]

On the other hand, by Lagrange’s mean value theorem, for each \( i \in \mathcal{I}^* \),
\[(w_i^k - w_i^{k-1}) \cdot \text{sign}(x_i^k) = |w_i^k - w_i^{k-1}|
\]

\[= |p|x_i^k| + (\delta_i^k)^2| - p|x_i^{k-1}| + (\delta_i^{k-1})^2| |
\]

\[= |p(1-p)(c_i^k)^{p-2}[|x_i^k| - |x_i^{k-1}| + (\delta_i^k)^2 - (\delta_i^{k-1})^2]| |
\]

\[\leq p(1-p)(c_i^k)^{p-2}[|x_i^k - x_i^{k-1}| + (\delta_i^k)^2 - (\delta_i^{k-1})^2]|
\]

\[\leq p(1-p)(c_i^k)^{p-2}[|x_i^k - x_i^{k-1}| + 2\delta_i^0(\delta_i^k - \delta_i^{k-1})],
\]

where the first equality is by the fact that \(x_i^k \neq 0\), and \(c_i^k\) is between \(|x_i^k| + (\delta_i^k)^2\) and \(|x_i^{k-1}| + (\delta_i^{k-1})^2\). It then follows that

\[
\|(w^* - w^{k-1}) \circ \text{sign}(x^k)\|_2
\]

\[\leq \|(w^* - w^{k-1}) \circ \text{sign}(x^k)\|_1
\]

\[\leq \overline{C}(\|x^k - x^{k-1}\|_1 + 2\|\delta^0\|_\infty(\|\delta^{k-1}\|_1 - \|\delta^k\|_1))
\]

\[\leq \overline{C}(\sqrt{n}\|x^k - x^{k-1}\|_2 + 2\|\delta^0\|_\infty(\|\delta^{k-1}\|_1 - \|\delta^k\|_1)),
\]

where the second inequality is by \((c_i^k)^{p-2} \leq \left(\frac{p\lambda}{C}\right)^\frac{p-2}{1-p}\) from Theorem 1(iii) with \(\overline{C} := p(1-p)\left(\frac{p\lambda}{C}\right)^\frac{p-2}{1-p}\). Combining (23), (24), (25) and (26), we know

\[
\|\nabla_x\psi(x^k, x^{k-1}, \delta^k)\|_2 \leq (L_f + \lambda \overline{C}\sqrt{n})\|x^k - x^{k-1}\|_2
\]

\[+ \tilde{a}(L_f + \beta)\|x^{k-1} - x^{k-2}\|_2
\]

\[+ 2\lambda \overline{C}\|\delta^0\|_\infty(\|\delta^{k-1}\|_1 - \|\delta^k\|_1).
\]

On the other hand, we have from (22) that

\[
\|\nabla_y\psi(x^k, x^{k-1}, \delta^k)\|_2 = \beta\|x^k - x^{k-1}\|_2,
\]

and that

\[
\|\nabla_\delta\psi(x^k, x^{k-1}, \delta^k)\|_2 \leq \|\nabla_\delta\psi(x^k, x^{k-1}, \delta^k)\|_1
\]

\[= \sum_{i \in I^*} 2\lambda w_i^k \delta_i^k
\]

\[\leq \sum_{i \in I^*} 2\lambda \overline{C}\frac{\sqrt{\mu}}{1 - \sqrt{\mu}}(\delta_i^{k-1} - \delta_i^k)
\]

\[\leq \frac{2\lambda \sqrt{\mu}}{1 - \sqrt{\mu}}(\|\delta^{k-1}\|_1 - \|\delta^k\|_1),
\]

where the second inequality is by Theorem 1(ii) and \(\delta^k \leq \sqrt{\mu}\delta^{k-1}\). Overall, we obtain from (27), (28) and (29) that Part (i) holds true by setting

\(\text{Springer}\)
\[ D_1 = \max \left( \beta + L_f + \lambda \bar{C} \sqrt{n}, \bar{a}(L_f + \beta), 2\lambda \bar{C} \| \delta^0 \|_\infty + \frac{2C \sqrt{n}}{1-\sqrt{n}} \right). \]

Part (ii) and (iii) follow directly from Lemma 1(i) with \( D_2 = \beta (1 - \bar{a}^2)/2 \) and Theorem 2(i), respectively. \( \square \)

Now we are ready to prove the uniqueness of limit points.

**Theorem 3** Suppose assumptions 1 and 2 are satisfied. Let \( \{x^k\} \) be a sequence generated by Algorithm 1.

Then \( \{x^k\} \) converges to a stationary point of \( F(x) \); moreover,

\[ \sum_{k=0}^{\infty} \|x^k - x^{k-1}\|_2 < \infty. \]

**Proof** Similar to the proof of Lemma 2 we can assume

\[ \mathcal{I}^* = \{1, \ldots, n\} \quad \text{and} \quad \mathcal{A}^* = \emptyset, \]

so that all the components of \( \{x^k\} \) for sufficiently large \( k \) are bounded away from 0, i.e. \( (x^k, x^{k-1}, \delta^k) = (x^k_1, x^{k-1}_2, \delta^k_2) \). By Theorem 2, it suffices to show that \( \{x^k\} \) has a unique cluster point.

By Lemma 2, \( \psi(x^k, x^{k-1}, \delta^k) \) is monotonically decreasing and converging to \( \zeta \). If \( \psi(x^k, x^{k-1}, \delta^k) = \zeta \) after some \( k_0 \), then from Lemma 2(ii), we know \( x^{k+1} = x^k \) for all \( k > k_0 \), meaning \( x^k \equiv x^{k_0} \in \mathcal{X} \), so that the proof is done.

We next consider the case that \( \psi(x^k, x^{k-1}, \delta^k) > \zeta \) for all \( k \). Since \( \psi \) has the uniform KL property at every \( (x^*, x^*, 0) \in \mathbb{R}^{3n} \), by the uniform KL property 2, there exists a continuous concave function \( \phi \) with \( \eta > 0, \tau > 0 \) and neighborhood

\[ U = \{(x, y, \delta) \in \mathbb{R}^{3n} \mid \text{dist}((x, y, \delta), \Gamma) < \tau \} \]

such that

\[ \phi'(\psi(x, y, \delta) - \zeta) \text{dist}((0, 0, 0), \partial \psi(x, y, \delta)) \geq 1 \quad (30) \]

for all \( (x, y, \delta) \in U \cap \{(x, y, \delta) \in \mathbb{R}^{3n} \mid \zeta < \psi(x, y, \delta) < \zeta + \eta \} \).

By the fact that \( \Gamma \) is the set of limit points of \( \{(x^k, x^{k-1}, \delta^k)\} \) and Lemma 1(ii), we have

\[ \lim_{k \to \infty} \text{dist}((x^k, x^{k-1}, \delta^k), \Gamma) = 0. \]

Hence, there exists \( k_1 \in \mathbb{N} \) such that \( \text{dist}((x^k, x^{k-1}, \delta^k), \Gamma) < \tau \) for any \( k > k_1 \). On the other hand, since \( \{\psi(x^k, x^{k-1}, \delta^k)\} \) is monotonically decreasing and converges to \( \zeta \), there exists \( k_2 \in \mathbb{N} \) such that \( \zeta < \psi(x^k, x^{k-1}, \delta^k) < \zeta + \eta \) for all \( k > k_2 \). Considering sufficiently large \( \bar{k} > \max\{k_1, k_2\} \) so that \( \mathcal{I}(x^\bar{k}) \) and \( \mathcal{A}(x^\bar{k}) \) are fixed for all \( k > \bar{k} \), and noticing that \( \psi \) is smooth at \( (x^\bar{k}, x^{\bar{k}-1}, \delta^\bar{k}) \) for all \( k > \bar{k} \), we know from (30) that
\( \phi'(\psi(x^k, x^{k-1}, \delta^k) - \zeta)\| \nabla \psi(x^k, x^{k-1}, \delta^k)\|_2 \geq 1, \) for all \( k \geq \bar{k}. \)

It follows that for any \( k \geq \bar{k}, \)

\[
[\phi(\psi(x^{k+1}, x^k, \delta^{k+1}) - \zeta) - \phi(\psi(x^{k+1}, x^k, \delta^{k+1}) - \zeta)] - D_1[||x^k - x^{k-1}||_2 + ||x^{k-1} - x^{k-2}||_2 + ||\delta^{k-1}||_1 - ||\delta^{k}||_1]
\]

\[
\geq [\phi(\psi(x^{k+1}, x^k, \delta^{k+1}) - \zeta) - \phi(\psi(x^{k+1}, x^k, \delta^{k+1}) - \zeta)]||\nabla \psi(x^{k+1}, x^k, \delta^{k+1})||_2
\]

\[
\geq \phi'(\psi(x^{k+1}, x^k, \delta^{k+1}) - \zeta) \cdot ||\nabla \psi(x^{k+1}, x^k, \delta^{k+1})||_2 \cdot [\nabla (x^{k+1}, x^k, \delta^{k+1}) - \psi(x^{k+1}, x^k, \delta^{k+1})]
\]

\[
\geq \psi(x^{k+1}, x^k, \delta^{k+1}) - \psi(x^{k+1}, x^k, \delta^{k+1}) \geq D_2||x^k - x^{k-1}||_2^2.
\]

where the first inequality is by Lemma 2(i), the second inequality is by the concavity of \( \phi, \) the fourth inequality is by Lemma 2(ii).

Rearranging and taking the square root of both sides, and using the inequality of arithmetic and geometric means, we have

\[
||x^k - x^{k-1}||_2 \leq \sqrt{\frac{2D_1}{D_2} [\phi(\psi(x^k, x^{k-1}, \delta^k) - \zeta) - \phi(\psi(x^{k+1}, x^k, \delta^{k+1}) - \zeta)]}
\]

\[
\times \sqrt{\frac{||x^k - x^{k-1}|| + ||x^{k-1} - x^{k-2}|| + (||\delta^{k-1}||_1 - ||\delta^{k}||_1)}{2}}
\]

\[
\leq \frac{D_1}{D_2} \left[ \phi(\psi(x^k, x^{k-1}, \delta^k) - \zeta) - \phi(\psi(x^{k+1}, x^k, \delta^{k+1}) - \zeta) \right]
\]

\[
+ \frac{1}{4} ||x^k - x^{k-1}||_2 + ||x^{k-1} - x^{k-2}||_2 + (||\delta^{k-1}||_1 - ||\delta^{k}||_1).}
\]

Subtracting \( \frac{1}{2} ||x^k - x^{k-1}||_2 \) from both sides, we have

\[
\frac{1}{2} ||x^k - x^{k-1}||_2 \leq \frac{D_1}{D_2} [\phi(\psi(x^k, x^{k-1}, \delta^k) - \zeta) - \phi(\psi(x^{k+1}, x^k, \delta^{k+1}) - \zeta)]
\]

\[
+ \frac{1}{4} (||x^{k-1} - x^{k-2}||_2 - ||x^k - x^{k-1}||_2) + \frac{1}{4} (||\delta^{k-1}||_1 - ||\delta^{k}||_1).
\]

Summing up both sides from \( \bar{k} \) to \( t, \) we have

\[
\frac{1}{2} \sum_{k=\bar{k}}^{t} ||x^k - x^{k-1}||_2 \leq \frac{D_1}{D_2} \left[ \phi(\psi(x^k, x^{k-1}, \delta^k) - \zeta) - \phi(\psi(x^{k+1}, x^k, \delta^{k+1}) - \zeta) \right]
\]

\[
+ \frac{1}{4} \left( ||x^{k-1} - x^{k-2}||_2 - ||x^k - x^{k-1}||_2 \right) + \frac{1}{4} (||\delta^{k-1}||_1 - ||\delta^{k}||_1).
\]

Now letting \( t \to \infty, \) we know \( ||\delta^t||_1 \to 0 \) and \( ||x^t - x^{t-1}||_2 \to 0 \) by Lemma 1(iii), and that \( \phi(\psi(x^{k+1}, x^k, \delta^{k+1}) - \zeta) \to \phi(\zeta - \zeta) = \phi(0) = 0. \) Therefore, we know

\[
\sum_{k=\bar{k}}^{\infty} ||x^k - x^{k-1}||_2 \leq \frac{2D_1}{D_2} \phi(\psi(x^\bar{k}, x^{\bar{k}-1}, \delta^\bar{k}) - \zeta) + \frac{1}{2} (||x^{\bar{k}-1} - x^{\bar{k}-2}||_2 + ||\delta^{\bar{k}-1}||_1) < \infty
\]
This implies that \( \{ x^k \} \) is a Cauchy sequence and consequently is convergent. \( \square \)

### 4 Local convergence rate

Now we investigate the local convergence rate of Algorithm 1 by assuming that \( \psi \) has the property at \((x_T^k, x_T^*, \delta_T^k) = (x_T^k, x_T^*, 0)\) with \( \phi \) in the KL definition taking the form \( \phi(s) = cs^{1-\theta} \) for some \( \theta \in (0, 1) \) and \( c > 0 \). This additional requirement will be discussed later.

We now show that Algorithm 1 has local linear convergence if \( \theta \) is at most \( \frac{1}{2} \) and local sublinear convergence if \( \theta \) is greater than \( \frac{1}{2} \). For simplicity, we still assume (21) in this section. The proof of this theorem essentially follows the convergence analysis in [1, Theorem 2] and [38, Theorem 4.3].

**Theorem 4** Suppose assumptions 1 and 2 are satisfied, and \( \{ x^k \} \) is generated by Algorithm 1 and converges to \( x^* \). Assume \( \phi \) in the uniform KL definition taking the form \( \phi(s) = cs^{1-\theta} \) for some \( \theta \in (0, 1) \) and \( c > 0 \). Then the following statements hold.

(i) If \( \theta = 0 \), then there exists \( k_0 \in \mathbb{N} \) so that \( x^k \equiv x^* \) for any \( k > k_0 \);

(ii) If \( \theta \in (0, \frac{1}{2}] \), then there exist \( \gamma \in (0, 1), c_1 > 0, c_2 > 0 \) such that

\[
\| x^k - x^* \|_2 < c_1 \gamma^k - c_2 \| \delta_T^k \|_1
\]

for sufficiently large \( k \);

(iii) If \( \theta \in (\frac{1}{2}, 1) \), then there exist \( c_3 > 0, c_4 > 0 \) such that

\[
\| x^k - x^* \|_2 < c_3 k^{-\frac{1-\theta}{2 \theta - 1}} - c_4 \| \delta_T^k \|_1
\]

for sufficiently large \( k \).

**Proof** Similar to the proof of Lemma 2 we can assume

\[
I^* = \{ 1, \ldots, n \} \quad \text{and} \quad A^* = \emptyset,
\]

so that all the components of \( \{ x^k \} \) for sufficiently large \( k \) are bounded away from 0, i.e. \((x^k, x^{k-1}, \delta^k) = (x_T^k, x_T^{k-1}, \delta_T^k)\).

(i) If \( \theta = 0 \), then \( \phi(s) = cs \) and \( \phi'(s) \equiv c \). We claim that there must exist \( k_0 > 0 \) such that \( \psi(x_T^{k_0}, x_T^{k_0-1}, \delta_T^{k_0}) = \zeta \). Suppose by contradiction this is not true so that \( \psi(x_T^k, x_T^{k-1}, \delta_T^k) > \zeta \) for all \( k \). Since \( \lim_{k \to \infty} x_T^k = x^* \) and the sequence \( \{ \psi(x_T^k, x_T^{k-1}, \delta_T^k) \} \) is monotonically decreasing to \( \zeta \) by Lemma 2. The KL inequality implies that all sufficiently large \( k \),

\[
c\| \nabla \psi(x_T^k, x_T^{k-1}, \delta_T^k) \|_2 \geq 1,
\]
contradicting $\|\nabla \psi(x^k, x^{k-1}, \delta^k)\|_2 \to 0$ by Lemma 2(i). Thus, there exists $k_0 \in \mathbb{N}$ such that $\psi(x^k, x^{k-1}, \delta^k) = \psi(x^{k_0}, x^{k_0-1}, \delta^{k_0}) = \zeta$ for all $k > k_0$. Hence, we conclude from Lemma 2(ii) that $x^{k+1} = x^k$ for all $k > k_0$, meaning $x^k \equiv x^* = x^{k_0}$ for all $k \geq k_0$. This proves (i).

(ii)-(iii) Now consider $\theta \in (0, 1)$. First of all, if there exists $k_0 \in \mathbb{N}$ such that $\psi(x^{k_0}, x^{k_0-1}, \delta^{k_0}) = \zeta$, then using the same argument in the second paragraph of the proof of Theorem 3, we can see that $\{x^k\}$ converges finitely.

Thus, we only need to consider the case that $\psi(x^k, x^{k-1}, \delta^k) > \zeta$ for all $k$.

Moreover, define $S^k = \sum_{l=k}^{\infty} \|x^{l+1} - x^l\|_2$. It holds that

$$\|x^k - x^*\|_2 = \|x^k - \lim_{l \to \infty} x'\|_2 = \lim_{l \to \infty} \sum_{l=k}^{\infty} \|x^{l+1} - x^l\|_2 \leq S^k = \sum_{l=k}^{\infty} \|x^{l+1} - x^l\|_2.$$ 

Therefore, we only have to prove $S^k$ also has the same upper bound as in (32) and (33).

To derive the upper bound for $S^k$, we continue with (31). For any $k > \bar{k}$ with $\bar{k}$ defined in the proof of Theorem 3, we can use the same argument of deriving (31) to have

$$S^k \leq \frac{2D_1}{D_2} \phi(\psi(x^k, x^{k-1}, \delta^k) - \zeta) + \frac{1}{2} \|x^{k-1} - x^{k-2}\|_2 + \frac{1}{2} \|\delta^{k-1}\|_1$$

$$= \frac{2D_1}{D_2} \phi(\psi(x^k, x^{k-1}, \delta^k) - \zeta) + \frac{1}{2} \|S^{k-2} - S^{k-1}\|_2 + \frac{1}{2} \|\delta^{k-1}\|_1. \quad (34)$$

By KL inequality with $\phi'(s) = c(1 - \theta)s^{-\theta}$, for $k > \bar{k}$,

$$c(1 - \theta)(\psi(x^k, x^{k-1}, \delta^k) - \zeta)^{-\theta} \|\nabla \psi(x^k, x^{k-1}, \delta^k)\|_2 \geq 1. \quad (35)$$

On the other hand, using Lemma 2(i) and the definition of $S^k$, we see that for all sufficiently large $k$,

$$\|\nabla \psi(x^k, x^{k-1}, \delta^k)\|_2 \leq D_1 (S^{k-2} - S^k + \|\delta^{k-1}\|_1 - \|\delta^k\|_1). \quad (36)$$

Combining (35) with (36), we have

$$(\psi(x^k, x^{k-1}, \delta^k) - \zeta)^{\theta} \leq D_1 c(1 - \theta)(S^{k-2} - S^k + \|\delta^{k-1}\|_1 - \|\delta^k\|_1)$$

Taking a power of $(1 - \theta)/\theta$ to both sides of the above inequality and scaling both sides by $c$, we obtain that for all $k > \bar{k}$
\[
\phi(\psi(x^k, x^{k-1}, \delta^k) - \zeta) = c[\psi(x^k, x^{k-1}) - \zeta]^{1-\theta}
\leq c \left[ D_1 c (1 - \theta) \left( S^{k-2} - S^k + \|\delta^{k-1}\|_1 - \|\delta^k\|_1 \right) \right]^{\frac{1-\theta}{\theta}}
\leq c \left[ D_1 c (1 - \theta) \left( S^{k-2} - S^k + \|\delta^{k-1}\|_1 \right) \right]^{\frac{1-\theta}{\theta}},
\]
which combined with (34) yields
\[
S^k \leq C_1 \left[ S^{k-2} - S^k + \|\delta^{k-1}\|_1 \right]^{\frac{1-\theta}{\theta}} + \frac{1}{2} \left[ S^{k-2} - S^k + \|\delta^{k-1}\|_1 \right]
\]
where \( C_1 = \frac{2D_1 c}{D_2} \left( D_1 \cdot c (1 - \theta) \right)^{\frac{1-\theta}{\theta}} \). It follows that
\[
S^k + \frac{\sqrt{\mu}}{1 - \mu} \|\delta^k\|_1
\leq C_1 \left[ S^{k-2} - S^k + \|\delta^{k-1}\|_1 \right]^{\frac{1-\theta}{\theta}} + \frac{1}{2} \left[ S^{k-2} - S^k + \|\delta^{k-1}\|_1 \right] + \frac{\sqrt{\mu}}{1 - \mu} \|\delta^k\|_1
\leq C_1 \left[ S^{k-2} - S^k + \|\delta^{k-1}\|_1 \right]^{\frac{1-\theta}{\theta}} + \frac{1}{2} \left[ S^{k-2} - S^k + \|\delta^{k-1}\|_1 \right] + \frac{\mu}{1 - \mu} \|\delta^{k-1}\|_1
\leq C_1 \left[ S^{k-2} - S^k + \|\delta^{k-1}\|_1 \right]^{\frac{1-\theta}{\theta}} + C_2 \left[ S^{k-2} - S^k + \|\delta^{k-1}\|_1 \right],
\]
with \( C_2 := \frac{1}{2} + \frac{\mu}{1 - \mu} \).

For part (ii), \( \theta \in (0, \frac{1}{2}] \). Notice that
\[
\frac{1 - \theta}{\theta} \geq 1 \quad \text{and} \quad S^{k-2} - S^k + \|\delta^{k-1}\|_1 \to 0.
\]
Hence, there exists \( k_1 \geq \bar{k} \) such that for any \( k \geq k_1 \)
\[
\left[ S^{k-2} - S^k + \|\delta^{k-1}\|_1 \right]^{\frac{1-\theta}{\theta}} \leq \left[ S^{k-2} - S^k + \|\delta^{k-1}\|_1 \right].
\]
This, combined with (39), yields
\[
S^k + \frac{\sqrt{\mu}}{1 - \mu} \|\delta^k\|_1 \leq (C_1 + C_2) \left[ S^{k-2} - S^k + \|\delta^{k-1}\|_1 \right]
\]
for any \( k \geq k_1 \). By \( \delta^k_i \leq \sqrt{\mu} \delta^{k-1}_i \), we know that for \( i \in I^a \),
\[
\delta^{k-1}_i \leq \frac{\sqrt{\mu}}{1 - \mu} (\delta^{k-2}_i - \delta^k_i).
\]
It follows that
So,
\[
S^k + \frac{\sqrt{\mu}}{1 - \mu} \| \delta^k \|_1 \leq (C_1 + C_2) \left[ (S^{k-2} + \frac{\sqrt{\mu}}{1 - \mu} \| \delta^{k-2} \|_1) - (S^k + \frac{\sqrt{\mu}}{1 - \mu} \| \delta^k \|_1) \right].
\]

Hence, we have
\[
S^k + \frac{\sqrt{\mu}}{1 - \mu} \| \delta^k \|_1 \leq \frac{C_1 + C_2}{C_1 + C_2 + 1} \left[ S^{k-2} + \frac{\sqrt{\mu}}{1 - \mu} \| \delta^{k-2} \|_1 \right]
\]
\[
\leq \left( \frac{C_1 + C_2}{C_1 + C_2 + 1} \right)^{\frac{1}{k-k_1}} \left[ S^{(k-k_1) \mod 2] + k_1} + \frac{\sqrt{\mu}}{1 - \mu} \| \delta^{(k-k_1) \mod 2] + k_1} \|_1 \right]
\]
\[
\leq \left( \frac{C_1 + C_2}{C_1 + C_2 + 1} \right)^{\frac{k-k_1}{k}} \left[ S^{k_1} + \frac{\sqrt{\mu}}{1 - \mu} \| \delta^{k_1} \|_1 \right].
\]

Hence, we have
\[
S^k + \frac{\sqrt{\mu}}{1 - \mu} \| \delta^k \|_1 \leq \left( \frac{C_1 + C_2}{C_1 + C_2 + 1} \right)^{k-k_1} \left( S^{k_1} + \frac{\sqrt{\mu}}{1 - \mu} \| \delta^{k_1} \|_1 \right)
\]

Therefore, for any \( k \geq k_1 \),
\[
\| x^k - x^* \|_2 \leq S^k \leq c_1 \gamma^k - c_2 \| \delta^k \|_1
\]

with
\[
\gamma = \sqrt{\frac{C_1 + C_2}{C_1 + C_2 + 1}}, c_1 = \frac{S^{k_1} + \frac{\sqrt{\mu}}{1 - \mu} \| \delta^{k_1} \|_1}{\gamma^{k_1}} \quad \text{and} \quad c_2 = \frac{\sqrt{\mu}}{1 - \mu},
\]

completing the proof of (ii).

For part (iii), \( \theta \in (\frac{1}{2}, 1) \). Notice that
\[
\frac{1 - \theta}{\theta} < 1 \quad \text{and} \quad S^{k-2} - S^k + \| \delta^{k-1} \|_1 \to 0.
\]

Hence, there exists \( k_2 \geq \bar{k} \) such that for any \( k \geq k_2 \)
\[
[S^{k-2} - S^k + \| \delta^{k-1} \|_1] \leq \left[ S^{k-2} - S^k + \| \delta^{k-1} \|_1 \right]^{\frac{1-\theta}{\theta}}.
\]

This, combined with (39), yields
\[
S^k + \frac{\sqrt{\mu}}{1 - \mu} \| \delta^k \|_1 \leq (C_1 + C_2) \left[ S^{k-2} - S^k + \| \delta^{k-1} \|_1 \right]^{\frac{1-\theta}{\theta}}
\]

for any \( k \geq k_1 \). Combining with (40) yields
An extrapolated iteratively reweighted $\ell_1$ method with complex…

Raising to a power of $\frac{\theta}{1-\theta}$ to both side of the above equation, we see further that for any $k > k_2$

$$(S^k + \frac{\sqrt{\mu}}{1-\mu} \|\delta^k\|_1)^{\frac{\theta}{1-\theta}} \leq C_3[S^{k-2} + \frac{\sqrt{\mu}}{1-\mu} \|\delta^{k-2}\|_1 - (S^k + \frac{\sqrt{\mu}}{1-\mu} \|\delta^k\|_1)]$$

(41)

with $C_3 := (C_1 + C_2)^{\frac{\theta}{1-\theta}}$.

Consider the “even” subsequence of $\{k_2, k_2 + 1, \ldots\}$ and define $\{\Delta_t\}_{t \geq k_2}$ with $\Delta_t := S^t + \frac{\sqrt{\mu}}{1-\mu} \|\delta^t\|_1$. Following from the proof for [1, Theorem 2], we have

$$\Delta_t \leq \left[\Delta_{N_1} + \hat{\mu}(t - N_1)\right]^{1/\nu} \leq C_4 t^{\nu - \frac{1}{2\nu - 1}},$$

(43)

for some $C_4 > 0$.

As for the “odd” subsequence of $\{k_2, k_2 + 1, \ldots\}$, we can define $\{\Delta_t\}_{t \geq k_2}$ with $\Delta_t := S^{t+1} + \frac{\sqrt{\mu}}{1-\mu} \|\delta^{t+1}\|_1$ and then can still show that (43) holds true.

Therefore, for all sufficiently large and even number $k$,

$$\|x^k - x^*\|_2 \leq S^k = \Delta_{\frac{k}{2}} - \frac{\sqrt{\mu}}{1-\mu} \|\delta^{k}\|_1 \leq 2^{\frac{1}{2\nu - 1}} C_4 k^{\nu - \frac{1}{2\nu - 1}} - \frac{\sqrt{\mu}}{1-\mu} \|\delta^k\|_1.$$

For all sufficiently large and odd number $k$,

$$\|x^k - x^*\|_2 \leq S^k = \Delta_{\frac{k-1}{2}} - \frac{\sqrt{\mu}}{1-\mu} \|\delta^{k}\|_1 \leq 2^{\frac{1}{2\nu - 1}} C_4 (k - 1)^{\nu - \frac{1}{2\nu - 1}} - \frac{\sqrt{\mu}}{1-\mu} \|\delta^k\|_1.$$

Overall, we have

$$\|x^k - x^*\| = \|x^k - x^*\|_2 \leq c_3 k^{\frac{1}{2\nu - 1}} - c_4 \|\delta^k\|_1$$

for sufficiently large $k$, where

$$c_3 := 2^{\frac{1}{2\nu - 1}} C_4 \max(1, 2^{\frac{1}{2\nu - 1}}) \quad \text{and} \quad c_4 := \frac{\sqrt{\mu}}{1-\mu}.$$ 

This completes the proof. □
5 Calculating the KL exponent of $\psi$

Though the convergence rate proved in the previous section gives the local behavior of our proposed algorithm, for the results to be informative, one has to be able to estimate the KL exponent of $\psi$ at $(x_T^*, x_T, \delta_T) = (x_T^*, x_T^*, 0)$. However, as noted in [27, Page 63, Section "Locally stable support and sign"], the KL exponent of a given function is often extremely hard to be estimated. There are only a few results available in the literature [22, 36, 41]. The most useful and related result is the following theorem given in [36] and [42] and its thorough proof is provided in [36]. Since the proof is in Chinese, we add its translation to the appendix.

**Theorem 5** ([36, Theorem 4.2]) Suppose that $h : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable in the local region $B(x^*, \epsilon_0)$ of $x^*$ and satisfies $\nabla h(x^*) = 0$ and $\nabla^2 h(x^*)$ is nonsingular. It holds that $h$ satisfies the KL inequality with the constant $\theta = \frac{1}{2}$ at $x^*$.

In our case, $\psi$ may not be differentiable at a local minimizer $(x, y, \delta)$ and Theorem 5 is not applicable to $\psi(x, y, \delta)$. However, we only need that $\psi(x_T^*, y_T, \delta_T)$ satisfies the KL inequality at $(x_T^*, y_T^*, \delta_T^*) = (x_T^*, x_T^*, 0)$. For simplicity, let $z = (x_T^*, y_T^*, \delta_T^*)$ and $z^* = (x_T^*, x_T^*, 0)$. We know $\psi(z)$ is twice continuously differentiable around $z^*$, so that Theorem 5 can be applied, resulting in the following theorem.

**Theorem 6** If $x^*$ is stationary for $F(x)$ and

$$\nabla^2 f(x_T^*) + \lambda p(p - 1)\text{diag}(|x_T^*|^{p-2}) + \beta I$$

(44)
is nonsingular, then $\psi(z)$ satisfies the uniform KL inequality at $z^*$ with $\theta = \frac{1}{2}$.

**Proof** First of all, it is trivial to see that for any local minimizer $z^*$, it holds that $y_T^* = x_T^*$ and $\delta_T^* = 0$.

By the first-order stationary condition (19), $\nabla_{x_T^*} \psi(z^*) = 0$. On the other hand, $\nabla_{y_T^*} \psi(z^*) = 0$ and $\nabla_{\delta_T^*} \psi(z^*) = 0$, meaning $\nabla \psi(z^*) = 0$. Moreover, we know $\psi$ is twice differentiable in a local region $B(z^*, \epsilon)$. It then suffices to show that

$$\nabla^2 \psi(z^*) = \begin{bmatrix} \nabla^2 f(x_T^*) + \lambda p(p - 1)\text{diag}(|x_T^*|^{p-2}) + \beta I & 0 & 0 \\ 0 & \beta I & 0 \\ 0 & 0 & 2\lambda p\text{diag}(|x_T^*|^{p-1}) \end{bmatrix}$$

is nonsingular, which is naturally true if $\nabla^2 f(x_T^*) + \lambda p(p - 1)\text{diag}(|x_T^*|^{p-2}) + \beta I$ is nonsingular.

The above result is applicable for any general $f$. Therefore, if (44) is satisfied at every stationary point for $F$, meaning we have a uniform KL exponent 1/2, then Theorem 4 implies that the proposed method for solving $\epsilon_p$ regularized least square problem has local linear convergence.
However, this condition is still not easy to verify. In the next theorem, we show that the KL exponent is $\frac{1}{2}$ when $f$ is the least-square error. We first show the equivalence among the local minimizers of $F(x)$, $F(x^*)$ and $\psi(x^*_T, y^*_T, \delta^*_T)$.

**Lemma 3** Suppose $f(x) = \frac{1}{2}\|Ax - b\|_2^2$ with $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$. Let $x^* \in \mathbb{R}^n \setminus \{0\}$ be a local minimizer of $F(x)$. Then the following assertions are equivalent:

(i) $x^*$ is a local minimizer of $F(x)$.
(ii) $x^*_T$ is a local minimizer of $F(x^*)$.
(iii) $z^* = (x^*_T, x^*_T, 0)$ is a local minimizer of $\psi(z)$.

**Proof** (i)$\iff$(ii) By [17, Theorem 1], assertion (i) is equivalent to that

$$A^T A + \lambda p(p - 1) \text{diag}(|x^*_T|^p - 2) > 0,$$

which, by [17, Theorem 1] again, is equivalent to assertion (ii).

(ii)$\iff$(iii). If $x^*$ is a local minimizer of $F(x)$, then there exists neighborhood $\mathbb{B}(x^*, \epsilon)$ with $\epsilon > 0$ such that $F(x) \geq F(x^*)$ for all $x \in \mathbb{B}(x^*, \epsilon)$. For any $z \in \mathbb{B}(z^*, \epsilon)$, it holds that $x^*_T \in \mathbb{B}(x^*_T, \epsilon)$. Hence

$$\psi(z) = F(x, \sigma^2) + \frac{\beta}{2}\|x^*_T - y^*_T\|_2^2 \geq F(x, 0) + \frac{\beta}{2}\|x^*_T - y^*_T\|_2^2 \geq F(x^*) = \psi(z^*),$$

meaning $z^*$ is a local minimizer of $\psi$.

(iii)$\iff$(ii). If $z^* = (x^*_T, x^*_T, 0)$ is a local minimizer of $\psi$, then there exists neighborhood $\mathbb{B}(z^*, \epsilon)$ with $\epsilon > 0$ such that for any $z \in \mathbb{B}(z^*, \epsilon)$, $\psi(z) \geq \psi(z^*)$. It holds that for any $x \in \mathbb{B}(x^*, \epsilon)$,

$$F(x) = \psi(x^*_T, x^*_T, 0) \geq \psi(x^*_T, x^*_T, 0) = F(x^*),$$

meaning $x^*_T$ is a local minimizer of $F(x^*)$. \hfill \Box

The next result provides more equivalent statements to Lemma 3 (iii).

**Lemma 4** Suppose $f(x) = \frac{1}{2}\|Ax - b\|_2^2$ with $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$. Then the following assertions are equivalent:

(i) $z^* = (x^*_T, x^*_T, 0)$ is a local minimizer of $\psi(z)$.
(ii) $\nabla \psi(z^*) = 0$ and

$$\nabla^2 \psi(z^*) = \begin{bmatrix} A^T A + \lambda p(p - 1) \text{diag}(|x^*_T|^p - 2) + \beta I & 0 & 0 \\ 0 & \beta I & 0 \\ 0 & 0 & 2\lambda p \text{diag}(|x^*_T|^p - 1) \end{bmatrix}$$

is positive definite.
(iii) \( \psi(z) \) satisfies the second-order growth property at \( z^* = (x^*_x, x^*_z, 0) \), i.e., there exist \( \epsilon_0, \epsilon_1 > 0 \) such that
\[
\psi(z) \geq \psi(z^*) + \epsilon_1 \|z - z^*\|^2_2 \quad \text{for any } z \in \mathbb{B}(z^*, \epsilon_0)
\]

**Proof** (i)\( \Rightarrow \) (ii). Suppose (i) holds. It suffices to show \( \nabla^2 \psi(z^*) > 0 \). To do this, suppose on the contrary that \( \nabla^2 \psi(z^*) \) is not positive definite. Then, there exists \( v \neq 0 \), such that \( v^T \nabla^2 \psi(z^*) v = 0 \). Let \( \varphi(t) = \psi(z^* + tv) = \psi(x^*_{x^*}, x^*_{x^*_z} + tv_x, x^*_{x^*_z} + tv_y, \delta^*_{x^*} + tv_\delta) \) for each \( t \in \mathbb{R} \). One can verify that
\[
\varphi'(0) = \nabla \psi(x^*, x^*_z, 0)^T v = 0, \quad \varphi''(0) = v^T \nabla^2 \psi(z^*) v = 0.
\]
Moreover, since \( t = 0 \) is a local minimum of \( \varphi \) (since \( z^* \) is a local minimum of \( \psi \)), it holds that \( \varphi^{(3)}(0) = 0 \). On the other hand, by the elementary calculus, one can check that
\[
\varphi^{(4)}(0) = \lambda \rho (p - 1)(p - 2)(p - 3)\|v_x\|_2^2 v_x^T \text{diag}(|x^*_{x^*_z}|^{p-4}) v_x + 4 \lambda \rho (p - 1)\|v_\delta\|_2^2 v_\delta^T \text{diag}(|x^*_{x^*_z}|^{p-2}) v_\delta 
\]
\[
\leq 0,
\]
where the inequality is an equality if and only if \( v_x = 0 \) and \( v_\delta = 0 \), implying \( v_y \neq 0 \) due to \( v \neq 0 \). However, this cannot happen since in this case \( v^T \nabla^2 \psi(z^*) v = \beta \|v_x\|_2^2 > 0 \) contradicting \( v^T \nabla^2 \psi(z^*) v = 0 \). Therefore, \( \varphi^{(4)}(0) < 0 \), implying that \( t = 0 \) is a local maximizer of \( \varphi \) — a contradiction. Hence, (ii) holds true.

(ii)\( \Rightarrow \) (iii). Suppose that (ii) holds. Then
\[
\nabla \psi(z^*) = 0 \quad \text{and} \quad \nabla^2 \psi(z^*) > 0.
\]
By Taylor’s formula, we have that
\[
\psi(z) = \psi(z^*) + \nabla \psi(z^*)^T(z - z^*) + \frac{1}{2}(z - z^*)^T \nabla^2 \psi(z^*)(z - z^*) + o(\|z - z^*\|^2)
\]
for each \((x, y, \delta)\). This, together with (45), implies that there exists \( \epsilon_0, \epsilon_1 > 0 \) such that
\[
\psi(z) \geq \psi(z^*) + \epsilon_1 \|z - z^*\|^2
\]
for any \( z \in \mathbb{B}(z^*, \epsilon_0) \).

(iii)\( \Rightarrow \) (i). It is trivial. The proof is done. \[ \square \]

We now can prove the main result when \( f \) is the least-square error.

**Theorem 7** Let \( x^* \in \mathbb{R}^n \setminus \{0\} \) and \( f(x) = \frac{1}{2} \|Ax - b\|^2_2 \) with \( A \in \mathbb{R}^{m \times n} \) and \( b \in \mathbb{R}^m \). If \( x^* \) is a local minimizer of \( F(x) \), then \( \psi \) satisfies the KL inequality at \((X^*_T, Y^*_T, \delta^*_T) = (x^*_x, x^*_z, 0)\) with \( \theta = \frac{1}{2} \).
Proof Since $x^*$ is a local minimizer of $F(x)$, by Lemma 3, $z^*$ is a local minimizer of $\psi$. It follows from Lemma 4 that $\nabla^2 \psi(z^*)$ is positive definite. By Theorem 5, $\psi$ satisfies the KL inequality at $(x^*_I, y^*_I, \delta^*_I)$ with $\theta = \frac{1}{2}$.

Now consider the sequence $\{x^k\}$ generated by the proposed algorithm in this case. If one of the cluster points is the local minimizer of $F$, then all the cluster points are local minimizers since they form a connected set by Lemma 1(iii). Therefore, $F$ has a uniform KL exponent $1/2$ on all the limit points, implying the algorithm has a unique limit point by Theorem (). Based on Theorem 7, we have the following result on the convergence rate of our proposed method for solving $\ell_p$ regularized least square problem.

**Corollary 1** Let $f(x) = \frac{1}{2} \|Ax - b\|_2^2$ with $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$. Suppose $\{x^k\}$ is generated by Algorithm 1 and converges to $x^* \in \mathbb{R}^n \setminus \{0\}$. If $x^*$ is a local minimizer of $F(x)$, then there exist $\gamma \in (0, 1), c_1 > 0, c_2 > 0$ such that

$$
\|x^k - x^*\|_2 < c_1 \gamma^{k-1} - c_2 \|\delta^k\|_1
$$

for sufficiently large $k$.

### 6 Numerical results

In this section, we perform sparse signal recovery experiments (similar to [14, 38, 40, 42]) to study the behaviors of Algorithm 1. The goal of the experiments is to reconstruct a length $n$ sparse signal $x$ from $m$ observations via its incomplete measurements with $m < n$. For this purpose, we first generate an $m \times n$ matrix $A$ with i.i.d. standard Gaussian entries and orthonormalizing the rows. We then set $y = Ax_{\text{true}} + \epsilon$, where the origin signal $x_{\text{true}}$ contains $K$ randomly placed $\pm 1$ spikes and $\epsilon \in \mathbb{R}^m$ has i.i.d. standard Gaussian entries with variance $\sigma^2 = 10^{-4}$. The objective takes the form $f(x) = \frac{1}{2} \|Ax - y\|_2^2$, where $A \in \mathbb{R}^{m \times n}, y \in \mathbb{R}^m$ and $x \in \mathbb{R}^n$.

In our experiments, we compare the performances of the proposed algorithm EIRL1 with IRL1, iteratively reweighted $\ell_2$ algorithm (IRL2), and the iterative jumping thresholding (IJT) algorithm [42] for solving $\ell_p$-norm optimization problem. Since the subproblems of IJT is trivial to solve for $p = 0.5$, we fix $p = 0.5$ in all experiments. All algorithms start from a random Gaussian vector $x^0$ with mean 0 and variance 1, and have the same termination criterion that either the number of iteration exceeds the preset limit or

$$
\frac{\|x^k - x^{k-1}\|}{\|x^k\|} \leq \text{opttol},
$$

where opttol is a small parameter, and the default is $10^{-4}$. 
6.1 Comparison with other types of algorithms

We compare the computational efficiency of algorithms for solving $\ell_{1/2}$ regularization problems with different scales $(m, n, K)$. We generate 50 random data sets $(A, x_{true}, y)$ for each scale $(m, n, K)$ and compute average CPU time, the number of correct nonzero components (sparsity) with respect to $x_{true}$, that is, $\text{MSE} = \frac{1}{n} \| x^k - x_{true} \|^2_2$. For EIRL1, we fix $\lambda = 0.05$, $\mu = 0.9$, $\beta = 1.1$, $e^0 = 1$, and Nesterov’s momentum coefficient for $a^k$. Table 1 shows that EIRL1 obtains the correct nonzero components and converges faster than other methods.

6.2 The role of $\{a^k\}$

EIRL1 introduced a new parameter $a^k$. In this session, we test the sensitivity of our algorithm performance to the selection of different $a^k$. We set $K = 200, 500, 800$ for $x_{true}$ which represent the situations with low, medium and high sparsity, respectively. For each $K$, we test EIRL1 with 5 different values of $a^k \equiv 0, 0.3, 0.5, 0.7, 0.9$. For each setting, we generate 50 random data sets $(A, x_{true}, y)$, compute the average MSE versus iterates, and record the number of nonzero components in the converged solution. We plot the evolution of MSE and the box-plots about the number of nonzero components with different $a^k$ in Fig. 1. Our experiment shows that larger $a^k$ converges faster to sparser solutions.

6.3 Comparison with other extrapolation parameters

In Fig. 2, we also compare the dynamic updating rules with constant coefficient $a^k = 0.9$. Here we use the same experimental setting as in the previous subsection and set $K = 800$. The dynamic updating strategies considered here are

| $(m, n, K)$ | Cpu time | Sparsity |
|-------------|---------|----------|
|             | EIRL1   | IRL1     | IRL2     | IJT      |
| (1000, 2000, 100) | 0.2234  | 0.3118   | 0.5406   | 0.5190   |
| (2000, 4000, 200) | 1.2918  | 1.8690   | 3.0128   | 3.0540   |
| (3000, 6000, 300) | 3.0904  | 4.6514   | 7.2166   | 7.1222   |
| (4000, 8000, 400) | 5.3846  | 8.0226   | 12.4336  | 12.2002  |

| $(m, n, K)$ | Sparsity |
|-------------|----------|
|             | EIRL1    | IRL1    | IRL2    | IJT     |
| (1000, 2000, 100) | 100      | 100     | 1159.6  | 100     |
| (2000, 4000, 200) | 200      | 200     | 2324.6  | 200     |
| (3000, 6000, 300) | 300      | 300     | 3490.8  | 300     |
| (4000, 8000, 400) | 400      | 400     | 4655.3  | 400     |
An extrapolated iteratively reweighted $\ell_1$ method with complex…

- Nesterov’s momentum coefficient \([34]\): \(a^k = \frac{k-1}{k+2} \) for \(k \geq 1\) and \(a^0 = 0\).
- The momentum coefficient in FISTA \([6]\): \(\theta^0 = 1, \theta^{k+1} = \frac{1+\sqrt{1+4(\theta^k)^2}}{2}, \alpha^k = \frac{\theta^{k-1}}{\theta^k}\)

**Fig. 1** The performance for different \(\alpha\) when \(K = 200, 500, 800\). It presents that larger \(\alpha\) has better performance.
The momentum coefficient in the accelerated proximal gradient (APG) method [35]:
\[ \theta^0 = 1, \theta^{k+1} = \frac{\sqrt{(\theta^k)^2 + 4(\theta^k)^2} - (\theta^k)^2}{2}, \alpha^k = \theta^k((\theta^{k-1})^{-1} - 1). \]

We can see that the dynamic strategies have nearly the same performance, outperformed by the constant \( \alpha = 0.9 \).

**Appendix**

The translation of the proof of Theorem 5.

**Proof** It follows from \( \nabla h(x^*) = 0 \) and the mean-value theorem that
\[
h(x) - h(x^*) = (x - x^*)^T \nabla^2 h(x^* + t_x(x - x^*))(x - x^*),
\]
with \( t_x \in (0, 1) \). Since \( h \) is twice continuously differentiable on \( \mathbb{B}(x^*, \epsilon_0) \), there exists \( L > 0 \) such that
\[
h(x) - h(x^*) \leq L\|x - x^*\|_2^2
\]
for any \( x \in \mathbb{B}(x^*, \epsilon_0) \). On the other hand,
\[
\nabla h(x) = \nabla h(x) - \nabla h(x^*)
\]
\[
= \int_0^1 \nabla^2 h(x + t(x - x^*))(x - x^*)dt
\]
\[
= \begin{bmatrix}
[\nabla^2 h(x^* + t_{1x}(x - x^*))]_1 \\
\vdots \\
[\nabla^2 h(x^* + t_{nx}(x - x^*))]_n
\end{bmatrix}(x - x^*)
\]
\[
:= A_x(x - x^*),
\]
An extrapolated iteratively reweighted $\ell_1$ method with complex…

where $t_{ix} \in (0, 1)$, $i = 1, 2, \ldots, n$.

Since $\nabla^2 h(x^*)$ is nonsingular, there exists $0 < \varepsilon_1 < \varepsilon_0$ such that for any $x \in B(x^*, \varepsilon_1)$, $A_x$ is nonsingular. Therefore, $A_x^T A_x$ is positive definite on $B(x^*, \varepsilon_1)$. Hence, there exists $0 < \varepsilon < \varepsilon_1$ and $\sigma > 0$ satisfying $\sigma = \min_{x \in B(x^*, \varepsilon)} \sigma(\min(A_x^T A_x))$. It then follows that

$$
\|\nabla h(x)\|^2 = (x - x^*)^T A_x^T A_x (x - x^*) \geq \sigma \|x - x^*\|^2 \geq \frac{\sigma}{L} (h(x) - h(x^*)).
$$

This implies that there exists $\theta = \frac{1}{2} \in (0, 1)$, $\varepsilon > 0$ and $c = (\frac{L}{\sigma})^{1/2}$ for any $x \in B(x^*, \varepsilon)$,

$$(h(x) - h(x^*))^{1-\theta} \leq c \|\nabla h(x)\|,
$$

or, equivalently, $h$ satisfies the KL inequality with $\theta = \frac{1}{2}$. □

**Acknowledgements** Hao Wang was supported by the Natural Science Foundation of Shanghai under Grant 21ZR1442800 and the National Natural Science Foundation of China under Grant 12001367. The authors would like to thank Professor Ting Kei Pong and Professor Kaiwen Meng for their advice on calculating the KL exponent.

**Data Availability** The data that support the findings of this study are available from the corresponding author upon request.

**Declarations**

**Conflict of interest** All authors disclosed no relevant relationships.

**References**

1. Attouch, Hedy, Bolte, Jérôme.: On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. Math. Program. **116**(1), 5–16 (2009)
2. Attouch, Hédy., Bolte, Jérôme., Redont, Patrick, Soubeyran, Antoine: Proximal alternating minimization and projection methods for nonconvex problems: an approach based on the kurdyka-
   lojasiewicz inequality. Math. Oper. Res. **35**(2), 438–457 (2010)
3. 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**(1–2), 91–129 (2013)
4. Auslender, Alfred, Teboulle, Marc: Interior gradient and proximal methods for convex and conic optimization. SIAM J. Optim. **16**(3), 697–725 (2006)
5. Bauschke,H.H., Dao, M.N., Moursi, W.M.: On fej\’er monotone sequences and nonexpansive mappings. arXiv preprint arXiv:1507.05585, 2015
6. Beck, Amir, Teboulle, Marc: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imag. Sci. **2**(1), 183–202 (2009)
7. Becker, S.R., Candès, E.J., Grant, M.C.: Templates for convex cone problems with applications to sparse signal recovery. Math. Program. comput. **3**(3), 165 (2011)
8. Bolte, J., Daniilidis, A., Lewis, A.: The lojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. SIAM J. Optim. **17**(4), 1205–1223 (2007)
9. Bolte, Jérôme., Sabach, Shoham, Teboulle, Marc: Proximal alternating linearized minimization for nonconvex and nonsmooth problems. Math. Program. 146(1), 459–494 (2014)
10. Candes, E.J., Wakin, M.B., Boyd, S.P.: Enhancing sparsity by reweighted $\ell_1$ minimization. J. Fourier Anal. Appl. 14(5–6), 877–905 (2008)
11. Chartrand, R., Yin, W.: Iteratively reweighted algorithms for compressive sensing. In 2008 IEEE International conference on acoustics, speech and signal processing, pp. 3869–3872. IEEE, 2008
12. Chen, Xiaojun, Zhou, Weijun: Convergence of reweighted $\ell_1$ minimization algorithms and unique solution of truncated lp minimization. The Hong Kong Polytechnic University, Department of Applied Mathematics (2010)
13. Fan, Jianqing, Li, Runze: Variable selection via nonconcave penalized likelihood and its oracle properties. J. Am. Stat. Assoc. 96(456), 1348–1360 (2001)
14. Frankel, Pierre, Garrigos, Guillaume, Peypouquet, Juan: Splitting methods with variable metric for kurdyka-łojasiewicz functions and general convergence rates. J. Optim. Theory Appl. 165(3), 874–900 (2015)
15. Ge, Dongdong, Jiang, Xiaoye, Ye, Yinyu: A note on the complexity of $\ell_p$ minimization. Math. Program. 129(2), 285–299 (2011)
16. Jaggi, M.: Sparse convex optimization methods for machine learning. PhD Thesis, ETH Zurich, 2011
17. Lai, Ming-Jun., Wang, Jingyue: An unconstrained $\ell_q$ minimization with $0 < q \leq 1$ for sparse solution of underdetermined linear systems. SIAM J. Optim. 21(1), 82–101 (2011)
18. Luo, Zhi-Quan., Pang, Jong-Shi., Ralph, Daniel: Mathematical programs with equilibrium constraints. Cambridge University Press, Cambridge (1996)
19. Lustig, M., Donoho, D., Pauly, J.M.: Sparse mri: the application of compressed sensing for rapid mr imaging. Magnetic Resonance Med.: Official J. Int. Soc. Magnetic Resonance Med. 58(6), 1182–1195 (2007)
20. Mairal, Julien, Elad, Michael, Sapiro, Guillermo: Sparse representation for color image restoration. IEEE Trans. Image Process. 17(1), 53–69 (2007)
21. Mairal, J., Bach, F., Ponce, J., Sapiro, G.: Online learning for matrix factorization and sparse coding. J. Mach. Learn. Res. 11(Jan), 19–60 (2010)
22. Nesterov, Yu.: Gradient methods for minimizing composite functions. Math. Program. 140(1), 125–161 (2013)
23. Nesterov, Yuriii: Introductory lectures on convex programming volume i: basic course. Lecture Notes 3(4), 5 (1998)
24. Nesterov, Yuriii: Primal-dual subgradient methods for convex problems. Math. Program. 120(1), 221–259 (2009)
34. Nesterov, Y.E.: A method for solving the convex programming problem with convergence rate $o(1/k^2)$. In Dokl. Akad. Nauk Sssr 269, 543–547 (1983)
35. Tseng, Paul: Approximation accuracy, gradient methods, and error bound for structured convex optimization. Math. Program. 125(2), 263–295 (2010)
36. Wang, Feng: Study on the Kurdyka-Łojasiewicz exponents of $\ell_p$ regularization functions (in Chinese). PhD thesis Southwest Jiaotong University, Chengdu (2021)
37. Wang, H., Zeng, H., Wang, J.: Relating $\ell_p$ regularization and reweighted $\ell_1$ regularization. arXiv preprint arXiv:1912.00723, 2019
38. Wen, B., Chen, X., Pong, T.K.: A proximal difference-of-convex algorithm with extrapolation. Comput. Optim. Appl. 69(2), 297–324 (2018)
39. Yangyang, Xu., Yin, Wotao: A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. SIAM J. Imag. Sci. 6(3), 1758–1789 (2013)
40. Peiran Yu and Ting Kei Pong: Iteratively reweighted $\ell_1$ algorithms with extrapolation. Comput. Optim. Appl. 73(2), 353–386 (2019)
41. Yu, P., Li, G., Pong, T.K.: Kurdyka-Łojasiewicz exponent via inf-projection. Found. Comput. Math. 22, 1–47 (2021)
42. Zeng, Jinshan, Lin, Shaobo, Zongben, Xu.: Sparse regularization: convergence of iterative jumping thresholding algorithm. IEEE Trans. Signal Process. 64(19), 5106–5118 (2016)
43. Roman Zeyde, Michael Elad, Matan Protter.: On single image scale-up using sparse-representations. In International conference on curves and surfaces, pp. 711–730. Springer, 2010
44. Zhang, Cun-Hui., et al.: Nearly unbiased variable selection under minimax concave penalty. Ann. Stat. 38(2), 894–942 (2010)

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.