SMOOTHING FAST ITERATIVE HARD THRESHOLDING ALGORITHM FOR ℓ_0 REGULARIZED NONSMOOTH CONVEX REGRESSION PROBLEM*

FAN WU†, WEI BIAN*‡, AND XIAOPING XUE‡

Abstract. We investigate a class of constrained sparse regression problem with cardinality penalty, where the feasible set is defined by box constraint, and the loss function is convex, but not necessarily smooth. First, we put forward a smoothing fast iterative hard thresholding (SFIHT) algorithm for solving such optimization problems, which combines smoothing approximations, extrapolation techniques and iterative hard thresholding methods. The extrapolation coefficients can be chosen to satisfy $\sup_k \beta_k = 1$ in the proposed algorithm. We discuss the convergence behavior of the algorithm with different extrapolation coefficients, and give sufficient conditions to ensure that any accumulation point of the iterates is a local minimizer of the original cardinality penalized problem. In particular, for a class of fixed extrapolation coefficients, we discuss several different update rules of the smoothing parameter and obtain the convergence rate of $O(\ln k/k)$ on the loss and objective function values. Second, we consider the case in which the loss function is Lipschitz continuously differentiable, and develop a fast iterative hard thresholding (FIHT) algorithm to solve it. We prove that the iterates of FIHT converge to a local minimizer of the problem that satisfies a desirable lower bound property. Moreover, we show that the convergence rate of loss and objective function values are $o(k^{-2})$. Finally, some numerical examples are presented to illustrate the theoretical results.

Key words. cardinality penalty, smoothing method, accelerated algorithm, extrapolation, convergence rate, local minimizer

AMS subject classifications. 90C30, 65K05, 49J52, 49M37

1. Introduction.

In this paper, we consider the following minimization problem:

$$\begin{array}{ll} \min & F(x) := f(x) + \lambda\|x\|_0 \\ \text{s.t.} & x \in \mathcal{X} := \{x \in \mathbb{R}^n : l \le x \le u\}, \end{array} \tag{1.1}$$

for some $\lambda > 0$, $l \in \overline{\mathbb{R}}^n_- := \{x \in \mathbb{R}^n : -\infty \le x_i \le 0, 1 \le i \le n\}$, $u \in \overline{\mathbb{R}}^n_+ := \{x \in \mathbb{R}^n : 0 \le x_i \le +\infty, 1 \le i \le n\}$ with $l < u$. In (1.1), we call that $f : \mathbb{R}^n \to \mathbb{R}$ is the loss function to characterize the data fitting, and $\|x\|_0$ is the penalty function to control the sparsity of solutions. Penalty parameter λ is to coordinate the trade-off between the data fitting and sparsity. Throughout this paper, we assume that f in (1.1) is convex but not necessarily smooth, and we focus on the case that f is nonsmooth.

Such nonsmooth convex regression problems with cardinality penalty arise from many important applications including compressed sensing [12, 19], variable selection [23], signal and image processing [32, 11], pattern recognition [8] and regression [34], etc. The purpose of these problems is to find the sparse solutions, most of whose elements are zeros. Owing to the existence of ℓ_0 function, optimization problem (1.1) is typical NP-hard in general. When f in (1.1) is a smooth convex function, a variety of first-order algorithms have been proposed. One important application is the least squares problem, i.e., $f(x) = \|Ax - b\|^2$, with sensing matrix $A \in \mathbb{R}^{m\times n}$ and observation vector $b \in \mathbb{R}^m$. Greedy methods have been proposed to seek the solutions of the

*Corresponding author.

Funding: This work is funded by the National Science Foundation of China (Nos: 11871178,61773136)

†School of Mathematics, Harbin Institute of Technology, Harbin, China (wufanmath@163.com).

‡School of Mathematics, Harbin Institute of Technology, Harbin, China; Institute of Advanced Study in Mathematics, Harbin Institute of Technology, Harbin 150001, China (bianweilvse520@163.com, xiaopingxue@hit.edu.cn).

ℓ_0 penalized least squares problems in the early stage, such as matching pursuit (MP) [26], orthogonal matching pursuit (OMP) [30], subspace pursuit (SP) [16], and so on. With the development of compressed sensing, Donoho [19] and Candès, Romberg, Tao [12] confirmed the equivalence of the ℓ_0 problem and ℓ_1 problem when A satisfies some proper conditions. Continuous convex relaxation methods is to replace the ℓ_0 function by a continuous convex function. Even though ℓ_1 penalty can be used to find the sparse solutions effectively, recent theoretical analysis shows that it often leads to an over-penalized problem or a biased estimator. Therefore, there occur various continuous nonconvex relaxation functions for ℓ_0 function, such as smoothly clipped absolute deviation (SCAD) function [20], hard thresholding function [37], capped-ℓ_1 function [31], transformed ℓ_1 function [29], etc. It is proved that these continuous nonconvex penalty functions can not only bring accurate sparse solutions, but also reduce the deviation of nonzero elements with respect to the true estimator. Nevertheless, in [6], it has been proved that finding global minimizers of these nonconvex relaxation problems are also NP-hard in general.

Blumensath and Davies [9] proposed an iterative hard thresholding (IHT) algorithm for solving the unconstrained and constrained ℓ_0 penalized problems, respectively. And they proved that the iterates converge to a local minimizer when $\|A\|_2 < 1$. In [10], they also verified that the IHT algorithm can obtain an approximated solution if A has restricted isometry property. Moreover, Lu and Zhang [25] presented a penalty decomposition method for general ℓ_0-penalized and ℓ_0-constrained minimization problem, and established that any accumulation point of the iterates satisfies the first-order optimality conditions. Lu [24] also studied an IHT algorithm and its variant for solving (1.1) when f is Lipschitz continuously differentiable.

The proximal forward-backward splitting algorithm [13, 15, 17, 21] is a classical first-order splitting method, which is also called the proximal gradient algorithm. When this method is used to solve the ℓ_1 penalized convex regression problem, it is often called the iterative shrinkage thresholding algorithm (ISTA). As we know, for the case that the loss function is Lipschitz continuously differentiable and convex, the convergence rate of the objective function values generated by ISTA is $O(k^{-1})$, where k is the iteration counter. Based on ISTA and Nesterov's acceleration scheme, Beck and Teboulle [4] proposed a fast iterative shrinkage thresholding algorithm (FISTA), which not only keeps the simplicity and computation cheapness of ISTA, but also improves the convergence rate of the objective function values to $O(k^{-2})$. Moreover, Nesterov [28] independently proposed an accelerated gradient algorithm for the same problem with the same convergence rate as FISTA. It's noteworthy that Su, Boyd and Candès [33] studied the relationships between a second-order ordinary differential equation and the Nesterov's accelerated gradient method. Inspired by the analysis in [33], Attouch and Peypouquet [3] investigated a proximal gradient method with extrapolation coefficients $\beta_k = \frac{k-1}{k+\alpha-1}$ for minimizing the sum of a Lipschitz continuously differentiable convex function and a proper closed convex function. In [3], it is proved that the convergence rate of the objective function values is $o(k^{-2})$, and the iterates $\{x^k\}$ converge to a minimizer of this problem as $\alpha > 3$. Recently, Doikov and Nesterov [18] presented a new accelerated algorithm for solving such problem, in which they used the high-order tensor methods to solve the inner subproblems and gave a complexity estimate under some assumptions. For the case with Lipschitz continuously differentiable but nonconvex loss function, Wen, Chen and Pong [35] proved that the iterates and objective function values generated by the proximal gradient algorithm with extrapolation are R-linearly convergent under the error bound condition. Later, Adly and Attouch [1] proposed an inertial proximal gradient algorithm

with Hessian damping and dry friction to obtain the finite convergence under some certain conditions.

A class of direct methods for solving nonsmooth convex minimization is the subgradient methods. For an $\epsilon > 0$, if x^ϵ satisfies $f(x^\epsilon) - \min f \leq \epsilon$, then x^ϵ is called an ϵ-approximation solution of $\min f$. It has been reported that the complexity of most subgradient methods for finding an ϵ-approximation solution is of the order $O(\epsilon^{-2})$. For a class of nonsmooth functions with max operator, Nesterov [27] gave a smooth convex function with Lipschitz continuous gradient of factor ϵ^{-1} to approximate it, where ϵ is a given and fixed positive paramter. Then, Nesterov [27] proved that the complexity for finding an ϵ-approximation solution of this nonsmooth problem can be improved to $O(\epsilon^{-1})$ when the accelerated gradient method is applied to solve the approximate smooth function. Later, the authors in [22] applied the Nesterov's smoothing technique to the Nash equilibria problem and gave a first-order method with the same complexity as [27]. Notably, Chen [14] presented a smoothing gradient method for solving the constrained nonsmooth nonconvex minimization problem and demonstrated how to update the smoothing parameter so that the algorithm converges to a stationary point of the problem. Recently, Bian and Chen [7] utilized an exact continuous relaxation problem to solve optimization problem (1.1) and presented a smoothing proximal gradient algorithm, whose iterates are globally convergent to a local minimizer of problem (1.1) and convergence rate on the objective function values is $o(k^{-\tau})$ with any $\tau \in (0, 1/2)$. In [2], the authors used the proximal regularized inertial Newton algorithm to solve the nonsmooth convex optimization problem, and proved that the convergence rate of the Moreau envelope values of the objective function is $o(k^{-2})$ when the index satisfies an updating rule.

Up to now, very few studies investigated accelerated algorithm for solving problem (1.1) in any systematic way. Inspired by the good performance of the accelerated algorithm with extrapolation and the smoothing method, we present a smoothing fast iterative hard thresholding (SFIHT) algorithm for solving problem (1.1). It is worth emphasizing that the SFIHT algorithm is used to minimize the sum of a nonsmooth convex function and a discontinuous nonconvex function. The strategy of acceleration is to adopt extrapolation on the iterates. As we know, the larger the range of the extrapolation coefficients, the better. It is worth noting that the extrapolation coefficients in our algorithm can satisfy $\sup_k \beta_k = 1$. A key technique in the SFIHT algorithm is that we split the range of the extrapolation coefficients into three cases, which are divided in line with the relationships among the ℓ_0 norms of the newest three adjacent iterates. Besides, though the subproblem in SFIHT algorithm is a nonconvex minimization problem, it has a closed-form solution and can be calculated exactly due to the special structure of ℓ_0 norm. So, the proposed SFIHT algorithm is well-defined. We show that the ℓ_0 norms of the iterates will not change after a finite number of iterations. Then, we discuss the convergence behavior of the SFIHT algorithm with different extrapolation coefficients. Also, we study the case that the loss function in problem (1.1) is smooth, and show a derived algorithm of SFIHT algorithm (called FIHT algorithm) for solving it. We prove that the iterates of FIHT algorithm converge to a local minimizer of this problem with an important lower bound property. Moreover, we also substantiate that the convergence rate of FIHT algorithm for the corresponding objective and loss function values is $o(k^{-2})$.

Contents. The rest of this paper is organized as follows. In Section 2, we first review some preliminary results on smoothing method, then we present the SFIHT algorithm for solving nonsmooth nonconvex problem (1.1). Next, we analyze the convergence properties of the proposed algorithm with different extrapolation coefficients

for solving (1.1). In Section 3, we focus on solving (1.1) with a smooth convex loss function. When the extrapolation coefficients in SFIHT algorithm are appropriately fixed, we give a better convergence behaviour of the proposed algorithm for solving this kind of problems. In Section 4, we apply the proposed algorithms to some practical instances, and show the value of acceleration by extrapolation in solving problem (1.1).

Notations. Throughout this paper, we denote $\mathbb{N} := \{1, 2, \cdots\}$. Let $\mathbb{R}^n$ be the Euclidean space with inner product $\langle\cdot,\cdot\rangle$ and corresponding Euclidean norm $\|\cdot\|$. For vectors $x, y \in \mathbb{R}^n$, $x \geq y$ means that $x_i \geq y_i$, $i = 1, 2, \cdots, n$. The ℓ_1 norm of vector $x \in \mathbb{R}^n$ is denoted by $\|x\|_1$, and let $I(x) := \{i : x_i = 0\}$. For a matrix $A \in \mathbb{R}^{m\times n}$, we use A^{T}, $\lambda_{\max}(A)$ and $\|A\| = \sqrt{\lambda_{\max}(A^{\mathrm{T}}A)}$ to denote its transpose, largest eigenvalue and spectral norm, respectively. Given a nonempty closed convex set $\Omega \subseteq \mathbb{R}^n$ and a vector $x \in \mathbb{R}^n$, $P_\Omega(x) := \arg\min\{\|x - z\| : z \in \Omega\}$, $N_\Omega(x)$ denotes the normal cone of Ω at x and $\Omega_J := \{x \in \Omega : x_j = 0, j \in J\}$ for a given index set $J \subseteq \{1, 2, \cdots, n\}$.

2. Numerical algorithm and its convergence analysis.

In this section, we focus on the case that f is a nonsmooth convex function. In what follows, we assume that f is level bounded on $\mathcal{X}$, i.e., set $\{x \in \mathcal{X} : f(x) \leq r\}$ is bounded for any $r \in \mathbb{R}$, which holds naturally if $\mathcal{X}$ is bounded. Note that function F in (1.1) is level bounded on $\mathcal{X}$ if and only if f is level bounded on $\mathcal{X}$.

2.1. Smoothing method and basic properties.

To overcome the nondifferentiability of loss function f in (1.1), we use a sequence of continuous differentiable functions to approximate f.

DEFINITION 2.1. *[7] We call $\tilde{f} : \mathbb{R}^n \times (0, \bar{\mu}] \to \mathbb{R}$ with $\bar{\mu} > 0$ a smoothing function of the convex function f on $\mathcal{X}$, if $\tilde{f}(x, \mu)$ satisfies the following conditions:*

(i) *for any fixed $\mu > 0$, $\tilde{f}(\cdot, \mu)$ is continuously differentiable in $\mathbb{R}^n$;*

(ii) $\lim_{z\to x, \mu\downarrow 0} \tilde{f}(z, \mu) = f(x), \quad \forall x \in \mathcal{X}$;

(iii) *$\tilde{f}(\cdot, \mu)$ is convex on $\mathcal{X}$ for any fixed $\mu > 0$;*

(iv) $\left\{\lim_{z\to x, \mu\downarrow 0} \nabla_z \tilde{f}(z, \mu)\right\} \subseteq \partial f(x), \quad \forall x \in \mathcal{X}$;

(v) *there exists a positive constant κ such that*

$$|\tilde{f}(x, \mu_2) - \tilde{f}(x, \mu_1)| \leq \kappa|\mu_1 - \mu_2|, \quad \forall x \in \mathcal{X},\ \mu_1, \mu_2 \in (0, \bar{\mu}];$$

(vi) *there exists a constant $L_{\tilde{f}} > 0$ such that for any $\mu \in (0, \bar{\mu}]$, $\nabla_x \tilde{f}(\cdot, \mu)$ is Lipschitz continuous on $\mathcal{X}$ with Lipschitz constant $L_{\tilde{f}}\mu^{-1}$.*

For the convenience of description, we provide $\tilde{f}$ a smoothing function of f with the definition in Definition 2.1 in the following analysis and denote $\nabla\tilde{f}(x, \mu)$ the gradient of $\tilde{f}(x, \mu)$ with respect to x. By virtue of Definition 2.1-(v), we see that

$$|\tilde{f}(x, \mu) - f(x)| \leq \kappa\mu, \quad \forall x \in \mathcal{X},\ 0 < \mu \leq \bar{\mu}. \tag{2.1}$$

Smooth approximations for nonsmooth optimization problems have been studied for decades. The fundamental of smoothing method we use in this paper is as follows. We first approximate loss function f by a smooth function with fixed smoothing parameter μ. Then, one find an approximate solution of the following problem

$$\min_{x\in\mathcal{X}} \tilde{F}(x, \mu) := \tilde{f}(x, \mu) + \lambda\|x\|_0. \tag{2.2}$$

Next, by updating the smoothing parameter μ, we can find a local minimizer of problem (1.1).

According to the definition of $\tilde{f}$, for any fixed $\mu \in (0, \bar{\mu}]$, $L \geq L_{\tilde{f}}$, it holds that

$$\tilde{f}(x,\mu) - \tilde{f}(y,\mu) \leq \langle \nabla \tilde{f}(y,\mu), x-y\rangle + \frac{L}{2\mu}\|x-y\|^2, \quad \forall x, y \in \mathcal{X}. \tag{2.3}$$

Using the convexity of function f, we can prove that $x^* \in \mathcal{X}$ is a local minimizer of problem (1.1) if and only if x^* satisfies $0 \in [\partial f(x^*)]_i + [N_{\mathcal{X}}(x^*)]_i, \forall i \notin I(x^*)$, which is equivalent to

$$x^* \in \arg\min\{f(x) : x \in \mathcal{X}_{I(x^*)}\}. \tag{2.4}$$

From (2.4), we can easily find that any local minimizer of problem (1.1) has the oracle property [20].

2.2. Smoothing fast iterative hard thresholding (SFIHT) algorithm. In this subsection, we combine the smoothing method, extrapolation technique and iterative hard thresholding algorithm to present a fast scheme for solving problem (1.1). We name it smoothing fast iterative hard thresholding algorithm and denote it by SFIHT algorithm for short.

In order to find an approximate solution of problem (2.2) with a fixed $\mu > 0$, we introduce an approximation of $\tilde{F}(x,\mu)$ around the given point y as follows

$$Q(x,y,\mu) := \tilde{f}(y,\mu) + \langle \nabla \tilde{f}(y,\mu), x-y\rangle + \frac{L}{2\mu}\|x-y\|^2 + \lambda\|x\|_0 \tag{2.5}$$

with a constant $L > L_{\tilde{f}}$. Further, we solve the following optimization problem

$$\min_{x\in\mathcal{X}} \; Q(x,y,\mu) \tag{2.6}$$

to find an approximate solution of problem (2.2). Although ℓ_0 function is nonsmooth and nonconvex, the objective function and the constraint set in (2.6) are both separable with respect to all elements of x. Using this fact, it has been proved in [24] that optimization problem (2.6) has the closed-form solution denoted by $\bar{x}$ and expressed by, for $i = 1, 2, \ldots, n$,

$$\bar{x}_i = \begin{cases} [P_{\mathcal{X}}\left(S_L(y,\mu)\right)]_i & \text{if } [S_L(y,\mu)]_i^2 - [q(y,\mu)]_i^2 > \dfrac{2\lambda\mu}{L}, \\ 0 & \text{if } [S_L(y,\mu)]_i^2 - [q(y,\mu)]_i^2 < \dfrac{2\lambda\mu}{L}, \\ [P_{\mathcal{X}}\left(S_L(y,\mu)\right)]_i \text{ or } 0 & \text{otherwise,} \end{cases} \tag{2.7}$$

where $S_L(y,\mu) := y - \frac{\mu}{L}\nabla\tilde{f}(y,\mu)$, $q(y,\mu) := P_{\mathcal{X}}\left(S_L(y,\mu)\right) - S_L(y,\mu)$. We take problem (2.6) as the unique subproblem of the proposed SFIHT algorithm. See Algorithm 2.1. Upon the above fact, we know that SFIHT algorithm is well-defined.

In each iteration of the SFIHT algorithm, the accelerated iterative hard thresholding method is used to find an approximate solution of problem (2.2). β_k in Step 1 satisfies $\beta_k \in \left[0, \sqrt{\frac{\mu_k}{\mu_{k-1}}}\right)$, which is a basic condition for the convergence analysis of the SFIHT algorithm. The new iterate depends on the two previous computed iterates. In order to improve the effect of extrapolation, we divide the extrapolation coefficients into three cases in the algorithm. We adjust the range of extrapolation coefficients according to the relationships among the ℓ_0 norms of the new computed iterate and two previous iterates. Step 4 is to update the smoothing parameter μ_k, which ensures that μ_k is decreasing and tends to zero.

Algorithm 2.1 Smoothing Fast Iterative Hard Thresholding (SFIHT) algorithm

Initialization: Take $x^1 = x^0 \in \mathcal{X}$, $L > L_{\tilde{f}}$, $\mu_1 = \mu_0 \in (0, \bar{\mu}]$ and $\sigma \in (0, 2)$. Set $k = 1$.

while a termination criterion is not met, **do**

Step 1. Choose $\beta_k \in \left[0, \sqrt{\frac{\mu_k}{\mu_{k-1}}}\right)$.

Step 2. Compute

$$y^k = x^k + \beta_k(x^k - x^{k-1}), \tag{2.8}$$

$$\bar{x}^{k+1} \in \arg\min\{Q(x, y^k, \mu_k) : x \in \mathcal{X}\}. \tag{2.9}$$

Step 3. **(3a)** If $I(x^{k-1}) = I(x^k) = I(\bar{x}^{k+1})$, let

$$x^{k+1} = \bar{x}^{k+1}$$

and go to **Step 4**.

(3b) Otherwise, choose $\beta_k \in \left[0, \sqrt{\frac{L - L_{\tilde{f}}}{4L} \frac{\mu_k}{\mu_{k-1}}}\right]$, compute **Step 2** to obtain $\bar{x}^{k+1}$.

(3b-1) If $I(x^k) = I(\bar{x}^{k+1})$, let

$$x^{k+1} = \bar{x}^{k+1}$$

and go to **Step 4**.

(3b-2) Otherwise, choose $\beta_k \in \left[0, \sqrt{\frac{L - L_{\tilde{f}}}{8L - 4L_{\tilde{f}}} \frac{\mu_k}{\mu_{k-1}}}\right]$, compute **Step 2** to obtain $\bar{x}^{k+1}$ and set $x^{k+1} = \bar{x}^{k+1}$.

Step 4. Set

$$\mu_{k+1} = \frac{\mu_0}{(k+2)^\sigma}.$$

Increment k by one and return to **Step 1**.

end while

Output x^k, μ_k and β_k.

2.3. Convergence analysis. Let $\{x^k\}$, $\{y^k\}$ and $\{\mu_k\}$ be the output iterates generated by the SFIHT algorithm. For $\kappa > 0$ in Definition 2.1-(v), we introduce the following sequence

$$H(x^k, \mu_k, \tau_k) := \tilde{F}(x^k, \mu_k) + \kappa\mu_k + \tau_k\|x^k - x^{k-1}\|^2,$$

where $\tau_k > 0$. Specially, we give a way of choosing τ_k as follows. For all $k \in \mathbb{N}$,

$$\tau_k = \begin{cases} \frac{L}{4}\mu_{k-1}^{-1} + \frac{L}{4}\beta_k^2\mu_k^{-1} & \text{if } I(x^{k-1}) = I(x^k) = I(x^{k+1}), \\ \frac{L - L_{\tilde{f}}}{8}\mu_{k-1}^{-1} & \text{otherwise.} \end{cases} \tag{2.10}$$

In the following, we will use the above τ_k to analyze the convergence of the SFIHT algorithm, which is the key point for all the following analysis.

First of all, we show that sequence $\{\tau_k\}$ guarantees that sequence $H(x^k, \mu_k, \tau_k)$ is nonincreasing and convergent, i.e., it can be used as an energy function of the SFIHT algorithm. To do this, we first state some basic properties and analyze the relationships between $H(x^k, \mu_k, \tau_k)$ and $H(x^{k+1}, \mu_{k+1}, \tau_{k+1})$.

LEMMA 2.2. *The following statements hold.*

(i) *For every $k \in \mathbb{N}$, $x^k \in \mathcal{X}$, $\{\mu_k\}$ is monotone decreasing and $\lim_{k\to\infty} \mu_k = 0$.*

(ii) *When $I(x^k) \neq I(x^{k+1})$, we have*

$$\begin{aligned} &H(x^{k+1}, \mu_{k+1}, \tau_{k+1}) - H(x^k, \mu_k, \tau_k) \\ &\leq \left[\tau_{k+1} - \frac{L - L_{\tilde{f}}}{4\mu_k}\right] \|x^{k+1} - x^k\|^2 + \left[\frac{2L - L_{\tilde{f}}}{2\mu_k}\beta_k^2 - \tau_k\right] \|x^k - x^{k-1}\|^2. \end{aligned} \tag{2.11}$$

(iii) *When $I(x^k) = I(x^{k+1})$, we have*

$$\begin{aligned} &H(x^{k+1}, \mu_{k+1}, \tau_{k+1}) - H(x^k, \mu_k, \tau_k) \\ &\leq \left[\tau_{k+1} - \frac{L}{2\mu_k}\right] \|x^{k+1} - x^k\|^2 + \left[\frac{L}{2\mu_k}\beta_k^2 - \tau_k\right] \|x^k - x^{k-1}\|^2. \end{aligned} \tag{2.12}$$

Proof. (i). By the proposed SFIHT algorithm, it's easy to verify this statement.

(ii). Using (2.3) with $x = x^{k+1}$, $y = y^k$ and $\mu = \mu_k$, we obtain

$$\tilde{f}(x^{k+1}, \mu_k) \leq \tilde{f}(y^k, \mu_k) + \langle \nabla \tilde{f}(y^k, \mu_k), x^{k+1} - y^k \rangle + \frac{L_{\tilde{f}}}{2\mu_k}\|x^{k+1} - y^k\|^2. \tag{2.13}$$

Since $\tilde{f}(x, \mu)$ is convex with respect to x for any fixed $\mu \in (0, \bar{\mu}]$, it holds that

$$\tilde{f}(y^k, \mu_k) + \langle \nabla \tilde{f}(y^k, \mu_k), x - y^k \rangle \leq \tilde{f}(x, \mu_k), \quad \forall x \in \mathcal{X}. \tag{2.14}$$

According to (2.13), for any $x \in \mathcal{X}$, $L > L_{\tilde{f}}$, we have

$$\begin{aligned} \tilde{F}(x^{k+1}, \mu_k) &= \tilde{f}(x^{k+1}, \mu_k) + \lambda\|x^{k+1}\|_0 \\ &\leq \tilde{f}(y^k, \mu_k) + \langle \nabla \tilde{f}(y^k, \mu_k), x^{k+1} - y^k \rangle + \frac{L}{2\mu_k}\|x^{k+1} - y^k\|^2 + \lambda\|x^{k+1}\|_0 \\ &\quad + \frac{L_{\tilde{f}} - L}{2\mu_k}\|x^{k+1} - y^k\|^2. \end{aligned}$$

This, combined with the definition of x^{k+1}, yields that for $x^k \in \mathcal{X}$,

$$\begin{aligned} \tilde{F}(x^{k+1}, \mu_k) &\leq \tilde{f}(y^k, \mu_k) + \langle \nabla \tilde{f}(y^k, \mu_k), x^k - y^k \rangle + \frac{L}{2\mu_k}\|x^k - y^k\|^2 + \lambda\|x^k\|_0 \\ &\quad + \frac{L_{\tilde{f}} - L}{2\mu_k}\|x^{k+1} - y^k\|^2 \\ &\leq \tilde{F}(x^k, \mu_k) + \frac{L}{2\mu_k}\|x^k - y^k\|^2 + \frac{L_{\tilde{f}} - L}{2\mu_k}\|x^{k+1} - y^k\|^2, \end{aligned}$$

where the last inequality holds by (2.14). The above inequality together with $y^k = x^k + \beta_k(x^k - x^{k-1})$, gives

$$\begin{aligned} \tilde{F}(x^{k+1}, \mu_k) \leq &\tilde{F}(x^k, \mu_k) + \frac{L_{\tilde{f}}}{2\mu_k}\beta_k^2\|x^k - x^{k-1}\|^2 + \frac{L_{\tilde{f}} - L}{2\mu_k}\|x^{k+1} - x^k\|^2 \\ &+ \frac{L - L_{\tilde{f}}}{\mu_k}\langle x^{k+1} - x^k, \beta_k(x^k - x^{k-1})\rangle. \end{aligned} \tag{2.15}$$

Using the algebratic inequality, it holds that

$$\frac{L-L_{\tilde{f}}}{\mu_k}\langle x^{k+1}-x^k, \beta_k(x^k-x^{k-1})\rangle \le \frac{L-L_{\tilde{f}}}{4\mu_k}\|x^{k+1}-x^k\|^2 + \frac{L-L_{\tilde{f}}}{\mu_k}\beta_k^2\|x^k-x^{k-1}\|^2.$$

Combining this with (2.15), we obtain

$$\tilde{F}(x^{k+1},\mu_k) \le \tilde{F}(x^k,\mu_k) - \frac{L-L_{\tilde{f}}}{4\mu_k}\|x^{k+1}-x^k\|^2 + \frac{2L-L_{\tilde{f}}}{2\mu_k}\beta_k^2\|x^k-x^{k-1}\|^2. \tag{2.16}$$

By Definition 2.1-(v) and the monotone decreasing of $\{\mu_k\}$, we have

$$\tilde{F}(x^{k+1},\mu_{k+1}) + \kappa\mu_{k+1} - \kappa\mu_k \le \tilde{F}(x^{k+1},\mu_k), \tag{2.17}$$

plugging (2.17) in (2.16), then we get

$$\begin{aligned}&\tilde{F}(x^{k+1},\mu_{k+1}) + \kappa\mu_{k+1}\\ \le&\tilde{F}(x^k,\mu_k) + \kappa\mu_k - \frac{L-L_{\tilde{f}}}{4\mu_k}\|x^{k+1}-x^k\|^2 + \frac{2L-L_{\tilde{f}}}{2\mu_k}\beta_k^2\|x^k-x^{k-1}\|^2.\end{aligned} \tag{2.18}$$

Using (2.18) and the definition of $H(x^k,\mu_k,\tau_k)$, we obtain (2.11) immediately.

(iii). Let

$$G(x,y,\mu) := \tilde{f}(y,\mu) + \langle\nabla\tilde{f}(y,\mu), x-y\rangle + \frac{L}{2\mu}\|x-y\|^2.$$

For each fixed y and μ, $G(x,y,\mu)$ is differentiable and strongly convex with respect to x with modulus $L\mu^{-1}$. Hence, for arbitrary $x\in\mathcal{X}$, we have

$$\begin{aligned}&\langle\nabla_x G(x^{k+1},y^k,\mu_k), x-x^{k+1}\rangle\\ =&\langle\nabla\tilde{f}(y^k,\mu_k), x-x^{k+1}\rangle + \frac{L}{2\mu_k}\|x-y^k\|^2 - \frac{L}{2\mu_k}\|x^{k+1}-x\|^2 - \frac{L}{2\mu_k}\|x^{k+1}-y^k\|^2.\end{aligned}$$

When $I(x^k) = I(x^{k+1})$, in view of the construction of the SFIHT algorithm, we find

$$x^{k+1} = \arg\min_{x\in\mathcal{X}}\{G(x,y^k,\mu_k) + \lambda\|x^k\|_0\}. \tag{2.19}$$

By the above fact, we obtain that $\langle\nabla_x G(x^{k+1},y^k,\mu_k), x-x^{k+1}\rangle \ge 0$, $\forall x\in\mathcal{X}$, which indicates that, for all $x\in\mathcal{X}$,

$$0 \le \langle\nabla\tilde{f}(y^k,\mu_k), x-x^{k+1}\rangle - \frac{L}{2\mu_k}\|x^{k+1}-y^k\|^2 + \frac{L}{2\mu_k}\|x-y^k\|^2 - \frac{L}{2\mu_k}\|x^{k+1}-x\|^2. \tag{2.20}$$

Summing up (2.13) and (2.20), and by $L > L_{\tilde{f}}$, we have

$$\begin{aligned}&\tilde{f}(x^{k+1},\mu_k)\\ \le&\tilde{f}(y^k,\mu_k) + \langle\nabla\tilde{f}(y^k,\mu_k), x-y^k\rangle + \frac{L}{2\mu_k}\|x-y^k\|^2 - \frac{L}{2\mu_k}\|x^{k+1}-x\|^2.\end{aligned} \tag{2.21}$$

Substituting (2.14) into the right side of (2.21), one has

$$\tilde{f}(x^{k+1},\mu_k) \le \tilde{f}(x,\mu_k) + \frac{L}{2\mu_k}\|x-y^k\|^2 - \frac{L}{2\mu_k}\|x^{k+1}-x\|^2. \tag{2.22}$$

Letting $x = x^k$ in (2.22), and using the definition of y^k, $\|x^k\|_0 = \|x^{k+1}\|_0$ and (2.17), we obtain

$$\begin{aligned}&\tilde{F}(x^{k+1},\mu_{k+1}) + \kappa\mu_{k+1}\\ \le&\tilde{F}(x^k,\mu_k) + \kappa\mu_k + \frac{L}{2\mu_k}\beta_k^2\|x^k - x^{k-1}\|^2 - \frac{L}{2\mu_k}\|x^{k+1} - x^k\|^2.\end{aligned} \tag{2.23}$$

By the definition of $H(x^k,\mu_k,\tau_k)$ and (2.23), we see the statement in item (iii). ∎

For simplicity, we define the notations

$$\gamma := \min\left\{\frac{L}{4}, \frac{L - L_{\tilde{f}}}{8}\right\}, \quad \mathcal{K} := \{k : I(x^k) \neq I(x^{k+1})\} \tag{2.24}$$

and

$$\nu := \min\left\{l_i^2\mu_0^{-1}, u_j^2\mu_0^{-1}, 2\lambda L^{-1} : l_i \neq 0, u_j \neq 0,\ i,j \in \{1,2,\ldots,n\}\right\}. \tag{2.25}$$

LEMMA 2.3. *The following statements hold.*

(i) *When* $k \in \mathcal{K}$, $\frac{2L-L_{\tilde{f}}}{2\mu_k}\beta_k^2 \le \tau_k$; *otherwise,* $\frac{L}{2\mu_k}\beta_k^2 \le \tau_k$.

(ii) $\{H(x^k,\mu_k,\tau_k)\}$ *is nonincreasing and convergent, i.e.,* $\lim_{k\to\infty} H(x^k,\mu_k,\tau_k) = H_\infty < \infty$.

Proof. (i). For $k \in \mathcal{K}$, there must be $\beta_k^2 \le \frac{L-L_{\tilde{f}}}{8L-4L_{\tilde{f}}}\frac{\mu_k}{\mu_{k-1}}$, which means that $\tau_k = \frac{L-L_{\tilde{f}}}{8}\mu_{k-1}^{-1} \ge \frac{2L-L_{\tilde{f}}}{2}\beta_k^2\mu_k^{-1}$.

$k \notin \mathcal{K}$ implies $k \in \mathcal{N}_1 := \{k : I(x^{k-1}) = I(x^k) = I(x^{k+1})\}$ or $k \in \mathcal{N}_2 := \{k : I(x^{k-1}) \neq I(x^k) = I(x^{k+1})\}$. From the SFIHT algorithm, we know $\beta_k^2 < \frac{\mu_k}{\mu_{k-1}}$ for $k \in \mathcal{N}_1$ and $\beta_k^2 \le \frac{L-L_{\tilde{f}}}{4L}\frac{\mu_k}{\mu_{k-1}}$ for $k \in \mathcal{N}_2$. Then, for $k \in \mathcal{N}_1$, $\mu_{k-1}^{-1} > \beta_k^2\mu_k^{-1}$, and by (2.10), we have

$$\tau_k = \frac{L}{4}\mu_{k-1}^{-1} + \frac{L}{4}\beta_k^2\mu_k^{-1} > \frac{L}{4}\beta_k^2\mu_k^{-1} + \frac{L}{4}\beta_k^2\mu_k^{-1} = \frac{L}{2}\beta_k^2\mu_k^{-1};$$

for $k \in \mathcal{N}_2$, $\mu_{k-1}^{-1} \ge \frac{4L}{L-L_{\tilde{f}}}\beta_k^2\mu_k^{-1}$, and by (2.10), we have $\tau_k = \frac{L-L_{\tilde{f}}}{8}\mu_{k-1}^{-1} \ge \frac{L}{2}\beta_k^2\mu_k^{-1}$. Then, we establish result (i).

(ii). We first analyze the nonincreasing of $\{H(x^k,\mu_k,\tau_k)\}$, and divide the proof into two cases as follows.

Case 1. If $k \in \mathcal{K}$, by (2.10), we have $\tau_{k+1} = \frac{L-L_{\tilde{f}}}{8}\mu_k^{-1}$. This together with (2.11) and (i) of this lemma, it holds that

$$H(x^{k+1},\mu_{k+1},\tau_{k+1}) - H(x^k,\mu_k,\tau_k) \le -\frac{L - L_{\tilde{f}}}{8}\mu_k^{-1}\|x^{k+1} - x^k\|^2, \quad \forall k \in \mathcal{K}. \tag{2.26}$$

Case 2. If $k \notin \mathcal{K}$, we know that $k \in \{I(x^k) = I(x^{k+1}) = I(x^{k+2})\}$ or $k \in \{I(x^k) = I(x^{k+1}) \neq I(x^{k+2})\}$.

When $k \in \{I(x^k) = I(x^{k+1}) = I(x^{k+2})\}$, $\tau_{k+1} = \frac{L}{4}\mu_k^{-1} + \frac{L}{4}\beta_{k+1}^2\mu_{k+1}^{-1}$. By (2.12) and statement (i) in this lemma, we obtain

$$\begin{aligned}&H(x^{k+1},\mu_{k+1},\tau_{k+1}) - H(x^k,\mu_k,\tau_k)\\ \le&\left[\tau_{k+1} - \frac{L}{2\mu_k}\right]\|x^{k+1} - x^k\|^2 = -\frac{L}{4}\left[1 - \beta_{k+1}^2\frac{\mu_k}{\mu_{k+1}}\right]\mu_k^{-1}\|x^{k+1} - x^k\|^2.\end{aligned} \tag{2.27}$$

By $\beta_{k+1} \in \left[0, \sqrt{\frac{\mu_{k+1}}{\mu_k}}\right)$ in this case, we obtain $H(x^{k+1}, \mu_{k+1}, \tau_{k+1}) \leq H(x^k, \mu_k, \tau_k)$.

When $k \in \{I(x^k) = I(x^{k+1}) \neq I(x^{k+2})\}$, by (2.10), we have $\tau_{k+1} = \frac{L-L_{\tilde{f}}}{8}\mu_k^{-1}$. According to (2.12) and result (i) of this lemma, it yields that

$$H(x^{k+1}, \mu_{k+1}, \tau_{k+1}) - H(x^k, \mu_k, \tau_k) \leq -\frac{3L + L_{\tilde{f}}}{8}\mu_k^{-1}\|x^{k+1} - x^k\|^2. \tag{2.28}$$

Hence, $\{H(x^k, \mu_k, \tau_k)\}$ is nonincreasing. Since f is bounded from below on $\mathcal{X}$ and $\{x^k\} \subseteq \mathcal{X}$, $\{H(x^k, \mu_k, \tau_k)\}$ is also bounded from below on $\mathcal{X}$. This together with the nonincreasing of $\{H(x^k, \mu_k, \tau_k)\}$ implies that $\{H(x^k, \mu_k, \tau_k)\}$ is convergent. ∎

The next lemma explores the boundedness of sequence $\{x^k\}$, and gives an estimate on $\{x^k\}$ and $\{\mu_k\}$, which lays a foundation for the analysis of $I(x^k)$.

LEMMA 2.4. *The following statements hold:*

(i) *sequence $\{x^k\}$ is bounded;*

(ii) *for any $k \in \mathcal{K}$, $\|x^{k+1} - x^k\|^2 \geq \nu\mu_k$, where ν is defined as in (2.25).*

Proof. (i). By (2.1) and the definition of $\{H(x^k, \mu_k, \tau_k)\}$, we have

$$F(x^k) \leq \tilde{F}(x^k, \mu_k) + \kappa\mu_k \leq H(x^k, \mu_k, \tau_k) \leq H(x^1, \mu_1, \tau_1) = \tilde{F}(x^1, \mu_1) + \kappa\mu_1,$$

which together with the level boundedness of F on $\mathcal{X}$ gives result (i).

(ii). For any $k \in \mathcal{K}$, from (2.7), there exists some $i \in \{1, 2, \ldots, n\}$ such that

$$x_i^k = [P_{\mathcal{X}}(S_L(y^{k-1}, \mu_{k-1}))]_i \neq 0,\ x_i^{k+1} = 0$$

or

$$x_i^k = 0,\ x_i^{k+1} = [P_{\mathcal{X}}(S_L(y^k, \mu_k))]_i \neq 0.$$

When $x_i^k = [P_{\mathcal{X}}(S_L(y^{k-1}, \mu_{k-1}))]_i \neq 0$ and $x_i^{k+1} = 0$, by the definition of $P_{\mathcal{X}}(\cdot)$, there exist three cases: (a) $x_i^k = l_i < 0$; (b) $x_i^k = u_i > 0$; (c) $x_i^k = [S_L(y^{k-1}, \mu_{k-1})]_i$. For case (a), we have

$$\|x^{k+1} - x^k\|^2 \geq |x_i^{k+1} - x_i^k|^2 = |x_i^k|^2 = |l_i|^2 \geq \nu\mu_0. \tag{2.29}$$

For case (b), similar to the analysis in case (a), we see that (2.29) also holds. For case (c), according to (2.7), we have $\left[S_L(y^{k-1}, \mu_{k-1})\right]_i^2 \geq 2\lambda L^{-1}\mu_{k-1}$. Hence,

$$\|x^{k+1} - x^k\|^2 \geq |x_i^{k+1} - x_i^k|^2 = |x_i^k|^2 = \left[S_L(y^{k-1}, \mu_{k-1})\right]_i^2 \geq \nu\mu_{k-1}.$$

These together with the decreasing of $\{\mu_k\}$, yields that

$$\|x^{k+1} - x^k\|^2 \geq \nu\mu_k. \tag{2.30}$$

When $x_i^k = 0$ and $x_i^{k+1} = [P_{\mathcal{X}}(S_L(y^k, \mu_k))]_i \neq 0$, by the similar arguments, we can get the same result as above. Then, we have thus proved this statement. ∎

Now we show that the support set of $\{x^k\}$ generated by the SFIHT algorithm will no longer change after finitely many iterations, and give some other useful estimates.

LEMMA 2.5. *The sequences generated by the SFIHT algorithm own the following properties:*

(i) *$I(x^k) = \{i : x_i^k = 0\}$ changes finite times at most;*

(ii) *the output extrapolation coefficients can be chosen to satisfy $\sup_k \beta_k = 1$;*

(iii) $\sum_{k=1}^{\infty}\left[1-\beta_{k+1}^2\frac{\mu_k}{\mu_{k+1}}\right]\mu_k^{-1}\|x^{k+1}-x^k\|^2<\infty$;

(iv) $\lim_{k\to\infty}\left[f(x^k)+\tau_k\|x^k-x^{k-1}\|^2\right]$ *exists.*

Proof. (i). Recalling $\mathcal{K}=\{k: I(x^k)\neq I(x^{k+1})\}$, then, we only need to show that set $\mathcal{K}$ has at most finite elements. We argue it by contradiction and suppose there are infinite elements in $\mathcal{K}$. This, together with (2.26) and (ii) of Lemma 2.3, we have

$$\begin{aligned}0&\le\sum_{k\in\mathcal{K}}\frac{1}{8}(L-L_{\tilde{f}})\mu_k^{-1}\|x^{k+1}-x^k\|^2\\&\le\sum_{k=1}^{\infty}\left[H(x^k,\mu_k,\tau_k)-H(x^{k+1},\mu_{k+1},\tau_{k+1})\right]=H(x^1,\mu_1,\tau_1)-H_\infty<\infty.\end{aligned}\tag{2.31}$$

On the basis of Lemma 2.4-(ii), we obtain

$$\sum_{k\in\mathcal{K}}\frac{1}{8}(L-L_{\tilde{f}})\mu_k^{-1}\|x^{k+1}-x^k\|^2\ge\sum_{k\in\mathcal{K}}\frac{1}{8}(L-L_{\tilde{f}})\nu=\infty.$$

This leads to a contradiction to (2.31). Hence, set $\mathcal{K}$ has at most finite elements, we see further that $I(x^k)=\{i: x_i^k=0\}$ changes finite times at most.

(ii). In view of result (i) of this lemma, we know that the SFIHT algorithm will continue to run (3a) in Step 3 after finite iterations. Then, by the update rule of μ_k in Step 4, the statement in (ii) holds.

(iii). According to (2.26), (2.27) and (2.28), and by (2.24), for $\forall k\in\mathbb{N}$, we find

$$H(x^{k+1},\mu_{k+1},\tau_{k+1})-H(x^k,\mu_k,\tau_k)\le-\gamma\left(1-\beta_{k+1}^2\frac{\mu_k}{\mu_{k+1}}\right)\mu_k^{-1}\|x^{k+1}-x^k\|^2.$$

Summing up the above inequality from 1 to ∞ and by Lemma 2.3-(ii), we obtain

$$\begin{aligned}0&\le\sum_{k=1}^{\infty}\gamma\left(1-\beta_{k+1}^2\frac{\mu_k}{\mu_{k+1}}\right)\mu_k^{-1}\|x^{k+1}-x^k\|^2\\&\le\sum_{k=1}^{\infty}\left[H(x^k,\mu_k,\tau_k)-H(x^{k+1},\mu_{k+1},\tau_{k+1})\right]=H(x^1,\mu_1,\tau_1)-H_\infty<\infty.\end{aligned}$$

In view of the definition of γ in (2.24), we get the desired result (iii).

(iv). Taking $x=x^k$ and $\mu=\mu_k$ in (2.1), and by direct computation, it yields that

$$H(x^k,\mu_k,\tau_k)-2\kappa\mu_k\le f(x^k)+\lambda\|x^k\|_0+\tau_k\|x^k-x^{k-1}\|^2\le H(x^k,\mu_k,\tau_k).$$

Letting k tend to infinity in the above inequality, along with $\lim_{k\to\infty}\mu_k=0$, we get

$$\lim_{k\to\infty}f(x^k)+\lambda\|x^k\|_0+\tau_k\|x^k-x^{k-1}\|^2=\lim_{k\to\infty}H(x^k,\mu_k,\tau_k).$$

This, combined with the fact that $\lim_{k\to\infty}\|x^k\|_0$ exists, we deduce the existence of

$$\lim_{k\to\infty}\left[f(x^k)+\tau_k\|x^k-x^{k-1}\|^2\right].$$

∎

Refs. [24] and [36] also study the IHT algorithm for solving the constrained ℓ_0 penalized convex regression problem modeled by (1.1). In terms of problem, the main difference is that the loss functions studied by them are smooth, while it can be nonsmooth in this paper. In terms of algorithm, [24] considers the IHT algorithm, while both [36] and this paper focus on the IHT algorithm with extrapolation. It's worth stressing that the extrapolation coefficients in [36] need satisfy $\sup_k \beta_k \le \frac{\sqrt{2}}{2}$, but the SFIHT algorithm proposed in this paper expands the range of extrapolation coefficients in a significant way, which can be seen clearly by the following results on convergence rate.

Remark 2.6. If $\beta_k \in \left[0, \sqrt{(1-a\mu_{k-1})\frac{\mu_k}{\mu_{k-1}}}\right]$ with $a > 0$ in Step 1, by Lemma 2.5-(iii), we obtain $\sum_{k=1}^{\infty}\|x^{k+1}-x^k\|^2 < \infty$.

Combining Lemma 2.5-(i) and the framework of the SFIHT algorithm, we know that the algorithm only runs (3a) in Step 3 after a finite number of iterations, which means that the SFIHT algorithm solves subproblem (2.9) only one time and is used to solve a nonsmooth convex optimization after finite itarations. Hence, in order to improve the convergence behavior of the iterates generated by the SFIHT algorithm, we will consider two different choices of β_k in Step 1 of the SFIHT algorithm.

2.3.1. A sufficient condition for the convergence of $\{x^k\}$. In this subsubsection, we will analyze the convergence of the iterates generated by the SFIHT algorithm for solving (1.1) when

$$\sigma \in \left[\frac{1}{2}, 1\right] \tag{2.32}$$

and

$$\beta_k \in \left[0, \sqrt{\left(1-\frac{1}{2k^{1-\sigma}}\right)\frac{\mu_k}{\mu_{k-1}}}\right] \tag{2.33}$$

in Step 1. It's easy to find that $\beta_k < \sqrt{\frac{\mu_k}{\mu_{k-1}}}$, hence, the previous results are still hold in this case. Based on these results, we first give the following estimations.

LEMMA 2.7. *When σ and β_k in Step 1 are chosen as in (2.32) and (2.33), it holds that*

$$\sum_{k=1}^{\infty}\frac{1}{k+1}\mu_k^{-2}\|x^{k+1}-x^k\|^2 < \infty \quad \textit{and} \quad \lim_{k\to\infty}\|x^{k+1}-x^k\| = 0.$$

Proof. From Lemma 2.5-(i) and (2.33), there exists $K \in \mathbb{N}$, such that

$$\beta_{k+1}^2 \le \left(1-\frac{1}{2(k+1)^{1-\sigma}}\right)\frac{\mu_{k+1}}{\mu_k}, \quad \forall k \ge K.$$

This together with Lemma 2.5-(iii), we obtain $\sum_{k=1}^{\infty}\frac{1}{2(k+1)^{1-\sigma}}\mu_k^{-1}\|x^{k+1}-x^k\|^2 < \infty$. According to $\mu_k = \frac{\mu_0}{(k+1)^\sigma}$, we see that

$$\sum_{k=1}^{\infty}\frac{1}{k+1}\mu_k^{-2}\|x^{k+1}-x^k\|^2 < \infty \quad \text{and} \quad \sum_{k=1}^{\infty}(k+1)^{2\sigma-1}\|x^{k+1}-x^k\|^2 < \infty.$$

Thanks to $\sigma \in [\frac{1}{2}, 1]$, we get $\sum_{k=1}^{\infty}\|x^{k+1}-x^k\|^2 < \infty$. Then, we have got all the estimates in this lemma. ∎

Next, we give another preliminary result.

PROPOSITION 2.8. *Let $\{a_k\}$ and $\{b_k\}$ be nonnegative sequences with $\sum_{k=1}^{\infty} a_k b_k < \infty$. If $\{a_k\}$ is a nonincreasing sequence and satisfies $\sum_{k=1}^{\infty} a_k = \infty$, then there exists a subsequence of $\{b_k\}$, denoted by $\{b_{k_i}\}$, satisfying $\lim_{i\to\infty} b_{k_i-1} = 0$ and $\lim_{i\to\infty} b_{k_i} = 0$.*

Proof. Since sequence $\{a_k\}$ is nonincreasing, we have

$$\sum_{k=2}^{\infty} a_k(b_{k-1} + b_k) \le \sum_{k=2}^{\infty} a_{k-1}b_{k-1} + \sum_{k=2}^{\infty} a_k b_k < \infty.$$

This together with $\sum_{k=2}^{\infty} a_k = \infty$ implies that there exists a subsequence of $\{b_{k-1} + b_k\}$, denoted by$\{b_{k_i-1} + b_{k_i}\}$, such that $\lim_{i\to\infty}(b_{k_i-1} + b_{k_i}) = 0$. Due to the nonnegativity of $\{b_k\}$, it follows that $\lim_{i\to\infty} b_{k_i-1} = 0$ and $\lim_{i\to\infty} b_{k_i} = 0$. ∎

THEOREM 2.9. *Any accumulation point of sequence $\{x^k\}$ generated by the SFIHT algorithm is a local minimizer of problem (1.1).*

Proof. From Lemma 2.5-(i), we know that there exist some $K \in \mathbb{N}$ and $I \subseteq \{1, 2, \ldots, n\}$ such that $I(x^k) = I$ for all $k \ge K$. This, combined with the SFIHT algorithm, we have

(2.34)
$$x^{k+1} = \arg\min_{x\in\mathcal{X}_I} \left\{ \tilde{f}(y^k, \mu_k) + \langle \nabla \tilde{f}(y^k, \mu_k), x - y^k\rangle + \frac{L}{2\mu_k}\|x - y^k\|^2 \right\}, \quad \forall k \ge K.$$

By Lemma 2.7, we know

$$\sum_{k=1}^{\infty} \frac{1}{k+1}\left[\mu_k^{-2}\|x^{k+1} - x^k\|^2\right] < \infty.$$

Using the above result, Proposition 2.8 with $a_k = \frac{1}{k+1}$ and $b_k = \mu_k^{-2}\|x^{k+1} - x^k\|^2$, there exists a subsequence of $\{\mu_k^{-2}\|x^{k+1} - x^k\|^2\}$ satisfying

$$\lim_{i\to\infty} \mu_{k_i-1}^{-2}\|x^{k_i} - x^{k_i-1}\|^2 = 0 \quad \text{and} \quad \lim_{i\to\infty} \mu_{k_i}^{-2}\|x^{k_i+1} - x^{k_i}\|^2 = 0. \tag{2.35}$$

We see further that $\lim_{i\to\infty} \mu_{k_i-1}^{-1}(x^{k_i} - x^{k_i-1}) = 0$ and $\lim_{i\to\infty} \mu_{k_i}^{-1}(x^{k_i+1} - x^{k_i}) = 0$, which implies

$$\lim_{i\to\infty} \frac{x^{k_i+1} - y^{k_i}}{\mu_{k_i}} = \lim_{i\to\infty} \left[\frac{x^{k_i+1} - x^{k_i}}{\mu_{k_i}} - \beta_{k_i}\frac{\mu_{k_i-1}}{\mu_{k_i}}\frac{x^{k_i} - x^{k_i-1}}{\mu_{k_i-1}}\right] = 0. \tag{2.36}$$

From Lemma 2.4-(i) and Lemma 2.5-(i), we know that there exists a subsequence of $\{x^{k_i}\}$ (also denoted by $\{x^{k_i}\}$ for simplicity) and $\bar{x} \in \mathcal{X}_I$ such that $\lim_{i\to\infty} x^{k_i} = \bar{x}$. By the triangle inequality, it holds that

$$\|x^{k_i+1} - \bar{x}\| \le \|x^{k_i+1} - x^{k_i}\| + \|x^{k_i} - \bar{x}\|,$$

then we immediately deduce that $\lim_{i\to\infty} x^{k_i+1} = \bar{x}$ by $\lim_{i\to\infty} \|x^{k+1} - x^k\| = 0$ proved in Lemma 2.7. According to the definitions of y^{k_i} and β_{k_i}, we can also obtain $\lim_{i\to\infty} y^{k_i} = \bar{x}$, and by Definition 2.1-(iv), we know that

$$\left\{\lim_{i\to\infty} \nabla \tilde{f}(y^{k_i}, \mu_{k_i})\right\} \subseteq \partial f(\bar{x}). \tag{2.37}$$

For any $k_i \geq K$, recalling the definition of x^{k_i+1} in (2.34), we have

$$\left\langle \nabla \tilde{f}(y^{k_i}, \mu_{k_i}) + L\mu_{k_i}^{-1}(x^{k_i+1} - y^{k_i}), x - x^{k_i+1} \right\rangle \geq 0, \quad \forall x \in \mathcal{X}_I,\ k_i \geq K. \tag{2.38}$$

Letting $i \to \infty$ in (2.38), by (2.36), (2.37) and $\lim_{i\to\infty} x^{k_i+1} = \bar{x}$, we notice that there exists $\bar{\xi} \in \partial f(\bar{x})$ satisfying

$$\langle \bar{\xi}, x - \bar{x} \rangle \geq 0, \quad \forall x \in \mathcal{X}_I. \tag{2.39}$$

Since f is a convex function and $\mathcal{X}_I$ is a nonempty closed convex set, (2.39) implies that $\bar{x}$ is a global minimizer of f on $\mathcal{X}_I$. By (2.35) and $\lim_{k\to\infty} \mu_k = 0$, we have $\lim_{i\to\infty} \mu_{k_i-1}^{-1} \|x^{k_i} - x^{k_i-1}\|^2 = 0$, which together with the choice of τ_k in (2.10), we deduce

$$\lim_{i\to\infty} \left[f(x^{k_i}) + \tau_{k_i} \|x^{k_i} - x^{k_i-1}\|^2 \right] = \lim_{i\to\infty} f(x^{k_i}) = f(\bar{x}).$$

This, combined with Lemma 2.5-(iv), we have

$$\lim_{k\to\infty} \left[f(x^k) + \tau_k \|x^k - x^{k-1}\|^2 \right] = f(\bar{x}).$$

Assume $\hat{x}$ is an accumulation point of $\{x^k\}$ with the convergence of subsequence $\{x^{l_j}\}$. Then, it holds that

$$f(\hat{x}) = \lim_{j\to\infty} f(x^{l_j}) \leq \lim_{j\to\infty} \left[f(x^{l_j}) + \tau_{l_j} \|x^{l_j} - x^{l_j-1}\|^2 \right] = f(\bar{x}). \tag{2.40}$$

Since $\hat{x} \in \mathcal{X}_I$ and $\bar{x} \in \arg\min_{x\in\mathcal{X}_I} f(x)$, by (2.40), we obtain $\hat{x} \in \arg\min_{x\in\mathcal{X}_I} f(x)$. Hence, any accumulation point of $\{x^k\}$ is a global minimizer of f on $\mathcal{X}_I$. Equivalently, any accumulation point of $\{x^k\}$ is a local minimizer of problem (1.1). ∎

Combining the construction and convergence analysis of the SFIHT algorithm, we know that the extrapolation coefficients can be considered into two cases for simplicity, i.e.,

$$\begin{cases} \beta_k \in \left[0, \sqrt{\left(1 - \dfrac{1}{2k^{1-\sigma}}\right) \dfrac{\mu_k}{\mu_{k-1}}}\ \right] & \text{if } I(x^{k-1}) = I(x^k) = I(\bar{x}^{k+1}), \\ \beta_k \in \left[0, \sqrt{\dfrac{L - L_{\tilde{f}}}{8L - 4L_{\tilde{f}}} \dfrac{\mu_k}{\mu_{k-1}}}\ \right] & \text{otherwise.} \end{cases}$$

for which the theoretical results in this paper are still valid. It's well known that the condition of the extrapolation parameters is weaker is better and $\sup_k \beta_k = 1$ is the most important. Thus, we divide the extrapolation coefficients into three cases, mainly to expand the range of extrapolation coefficients so as to greatly improve the convergence performance of the SFIHT algorithm. Though this behavior will increase the computation of the algorithm, it is surprising by Lemma 2.5-(i) that the computational increase is finite.

2.3.2. Convergence rate on objection function values. In this subsubsection, we discuss a specific β_k in Step 1 of the SFIHT algorithm. It not only keeps the validity for finding the local minimizers of problem (1.1), but also obtains the convergence rate of the loss and objective function values. Based on the FISTA algorithm in [4] and the smoothing method, Bian in [5] proposed a smoothing fast iterative shrinkage thresholding algorithm (FISTA$_S$) for solving the constrained nonsmooth convex

optimization problem. We set extrapolation coefficient β_k in Step 1 to be the same as that in [5], namely, the following recurrence relation:

$$\begin{cases} t_k = \dfrac{1+\sqrt{1+4\left(\frac{\mu_{k-1}}{\mu_k}\right)t_{k-1}^2}}{2}, \\ \beta_k = \dfrac{t_{k-1}-1}{t_k}, \end{cases} \tag{2.41}$$

with $t_0 = 1$. Then, $\beta_k < \sqrt{\frac{\mu_k}{\mu_{k-1}}}$, which satisfies the condition in Step 1. Since the SFIHT algorithm always runs (3a) in Step 3 after finite iterations, by [5, Theorem 1], we can directly get the following corollary.

COROLLARY 2.10. *Let $\{x^k\}$ be the sequence generated by the SFIHT algorithm, in which the β_k in Step 1 is defined by (2.41). Then, the following statements hold:*
(i) *any accumulation point of $\{x^k\}$ is a local minimizer of problem (1.1);*
(ii) $\lim_{k\to\infty} F(x^k) = F_\infty < \infty$ *exists and*

$$F(x^{k+1}) - F_\infty = \begin{cases} O(k^{-\sigma}) & \text{if } 0 < \sigma < 1, \\ O(k^{-1}\ln k) & \text{if } \sigma = 1, \\ O(k^{\sigma-2}) & \text{if } 1 < \sigma < 2. \end{cases}$$

Since the ℓ_0 norms of the iterates change only a finite number of times, the above convergence rate also holds for loss function f. From the estimation in Corollary 2.10, we can see that the convergence rate on the objective function values generated by the SFIHT algorithm greatly improves the rate given by the algorithm in [7].

3. A case on smooth loss function. In this section, we discuss the case that loss function f in problem (1.1) is Lipschitz continuously differentiable. In summary, throughout this section, we require f in (1.1) to satisfy the following assumptions

$$\begin{cases} \bullet\ f \text{ is a smooth convex function on } \mathcal{X}; \\ \bullet\ f \text{ is level bounded on } \mathcal{X}; \\ \bullet\ \nabla f \text{ is Lipschitz continuous with Lipschitz constant } L_f. \end{cases}$$

In the context of this problem, we select a special extrapolation coefficient β_k in the Step 1 of the SFIHT algorithm. Since the smoothing method is not needed in this case, we call it fast iterative hard thresholding (FIHT) algorithm. Please see Algorithm 3.1. For the sake of brevity, we also define an approximation of F at a given point y as follows:

$$Q(x,y) := f(y) + \langle \nabla f(y), x-y\rangle + \frac{L}{2}\|x-y\|^2 + \lambda\|x\|_0,$$

where $L > L_f$.

Let $\{x^k\}$ and $\{y^k\}$ be the iterates generated by the FIHT algorithm. Similarly, the subproblem in Step 2 of FIHT algorithm has a closed-form solution. Hence, the FIHT algorithm is also well-defined. The following lemma shows a lower bound property of $\{x^k\}$, which can be easily obtained by [36, Lemma 3.2], so we omit its proof here.

LEMMA 3.1. *The following statements hold.*
(i) *When $x_j^k \neq 0$ for some $j \in \{1,2,\ldots,n\}$, it holds that*

$$|x_j^k| \geq \delta := \min_{i=1,2,\ldots,n} \delta_i > 0, \tag{3.1}$$

Algorithm 3.1 Fast Iterative Hard Thresholding (FIHT) algorithm

Initialization: Take $x^1 = x^0 \in \mathcal{X}$, $\alpha > 0$ and $L > L_f$. Set $k = 1$.
while a termination criterion is not met, **do**
Step 1. Take $\beta_k = \frac{k-1}{k+\alpha-1}$.
Step 2. Compute

$$y^k = x^k + \beta_k(x^k - x^{k-1}),$$
$$\bar{x}^{k+1} \in \arg\min\{Q(x, y^k) : x \in \mathcal{X}\}.$$

Step 3. **(3a)** If $I(x^{k-1}) = I(x^k) = I(\bar{x}^{k+1})$, let

$$x^{k+1} = \bar{x}^{k+1},$$

increment k by one and return to **Step 1**.
(3b) Otherwise, choose $\beta_k \in \left[0, \sqrt{\frac{L-L_f}{4L}}\right]$, compute **Step 2** to obtain $\bar{x}^{k+1}$.
(3b-1) If $I(x^k) = I(\bar{x}^{k+1})$, let

$$x^{k+1} = \bar{x}^{k+1},$$

increment k by one and return to **Step 1**.
(3b-2) Otherwise, choose $\beta_k \in \left[0, \sqrt{\frac{L-L_f}{8L-4L_f}}\right]$, compute **Step 2** to obtain $\bar{x}^{k+1}$ and let $x^{k+1} = \bar{x}^{k+1}$.
Increment k by one and return to **Step 1**.
end while
Output x^k, μ_k and β_k.

where

$$\delta_i = \begin{cases} \min\left(u_i, \sqrt{2\lambda/L}\right) & \text{if } l_i = 0, \\ \min\left(-l_i, \sqrt{2\lambda/L}\right) & \text{if } u_i = 0, \\ \min\left(-l_i, u_i, \sqrt{2\lambda/L}\right) & \text{otherwise.} \end{cases}$$

(ii) *For every $k \in \mathbb{N}$, if $I(x^k) \neq I(x^{k+1})$, then $\|x^{k+1} - x^k\| \geq \delta$.*

We begin the convergence analysis of the FIHT algorithm by defining the following important auxiliary sequence

$$W(x^k, \zeta_k) := F(x^k) + \zeta_k\|x^k - x^{k-1}\|^2,$$

where $\zeta_k > 0$. Again, for all $k \in \mathbb{N}$, we give a choice of ζ_k as follows,

$$\zeta_k = \begin{cases} \frac{L}{4}(1+\beta_k^2) & \text{if } I(x^{k-1}) = I(x^k) = I(x^{k+1}), \\ \frac{L-L_f}{8} & \text{otherwise.} \end{cases} \tag{3.2}$$

By a similar analysis to sequence $H(x^k, \mu_k, \tau_k)$ in Section 2, we can get some basic

results on the sequence $W(x^k, \zeta_k)$ for the FIHT algorithm. For easy of reading, we only list them in the following lemma, but omit their proofs.

LEMMA 3.2. *The following properties are satisfied:*

(i) *for every* $k \in \mathbb{N}$, $x^k \in \mathcal{X}$;

(ii) *when* $I(x^k) \neq I(x^{k+1})$, *we have*

$$\begin{aligned} &W(x^{k+1}, \zeta_{k+1}) - W(x^k, \zeta_k) \\ &\leq \left[\zeta_{k+1} - \frac{L - L_f}{4}\right] \|x^{k+1} - x^k\|^2 + \left[\frac{2L - L_f}{2}\beta_k^2 - \zeta_k\right] \|x^k - x^{k-1}\|^2; \end{aligned}$$

(iii) *when* $I(x^k) = I(x^{k+1})$, *we have*

$$\begin{aligned} &W(x^{k+1}, \zeta_{k+1}) - W(x^k, \zeta_k) \\ &\leq \left[\zeta_{k+1} - \frac{L}{2}\right] \|x^{k+1} - x^k\|^2 + \left[\frac{L}{2}\beta_k^2 - \zeta_k\right] \|x^k - x^{k-1}\|^2; \end{aligned}$$

(iv) *when* $I(x^k) = I(x^{k+1})$, $\frac{L}{2}\beta_k^2 \leq \zeta_k$; *otherwise,* $\frac{2L - L_f}{2}\beta_k^2 \leq \zeta_k$;

(v) $\{W(x^k, \zeta_k)\}$ *is nonincreasing and* $\lim_{k\to\infty} W(x^k, \zeta_k) = W_\infty < \infty$ *exists.*

Based on the above results, the next lemma shows that the iterate sequence $\{x^k\}$ generated by the FIHT algorithm is bounded and $I(x^k)$ only changes finite times.

LEMMA 3.3. *The following statements hold:*

(i) *there exists a positive constant* R *such that* $\|x^k\| \leq R$, $\forall k \in \mathbb{N}$;

(ii) *the* ℓ_0 *norm of sequence* $\{x^k\}$ *changes only finitely often.*

Proof. (i). By virtue of the nonincreasing of sequence $\{W(x^k, \zeta_k)\}$, we have

$$F(x^k) \leq W(x^k, \zeta_k) \leq W(x^1, \zeta_1) = F(x^1).$$

Combining this and the level boundedness of f on $\mathcal{X}$, it is easy to obtain the boundedness of $\{x^k\}$.

(ii). Also denote $\mathcal{K} = \{k : I(x^k) \neq I(x^{k+1})\}$, and we prove the finiteness of set $\mathcal{K}$ by contradiction. Let's assume that $\mathcal{K}$ contains infinite elements. By (ii) and (iv) of Lemma 3.2 and $\zeta_{k+1} = \frac{L - L_f}{8}$ for $k \in \mathcal{K}$, we obtain

$$W(x^{k+1}, \zeta_{k+1}) - W(x^k, \zeta_k) \leq -\frac{L - L_f}{8}\|x^{k+1} - x^k\|^2, \quad \forall k \in \mathcal{K}.$$

Summing up the above inequality over $k \in \mathcal{K}$ and using Lemma 3.2-(v), we find

$$\begin{aligned} \sum_{k\in\mathcal{K}} \frac{L - L_f}{8}\|x^{k+1} - x^k\|^2 &\leq \sum_{k\in\mathcal{K}} \left[W(x^k, \zeta_k) - W(x^{k+1}, \zeta_{k+1})\right] \\ &\leq \sum_{k=1}^{\infty} \left[W(x^k, \zeta_k) - W(x^{k+1}, \zeta_{k+1})\right] \\ &= W(x^1, \zeta_1) - W_\infty < \infty. \end{aligned} \tag{3.3}$$

In addition, from Lemma 3.1-(ii), we have

$$\sum_{k\in\mathcal{K}} \|x^{k+1} - x^k\|^2 \geq \sum_{k\in\mathcal{K}} \delta^2 = \infty,$$

which is inconsistent with (3.3). Hence, $I(x^k)$ only changes finite times, as claimed. □

From the above lemma, we can easily find that the FIHT algorithm is reduced to the proximal gradient algorithm with extrapolation $\beta_k = \frac{k-1}{k+\alpha-1}$ for solving $\min_{\mathcal{X}_J} f$ after a finite number of iterations, where J is a fixed index set by Lemma 3.3. Literature [3] studied the forward-backward method with extrapolation coefficient $\frac{k-1}{k+\alpha-1}$ for solving the sum of a convex function with Lipschitz continuous gradient and a proper closed convex function. In the context of this section, the algorithm can be written as follows:

$$\begin{cases} y^k = x^k + \dfrac{k-1}{k+\alpha-1}(x^k - x^{k-1}), \\ x^{k+1} = \arg\min\limits_{x\in\Omega} \left\{ \langle \nabla f(y^k), x - y^k\rangle + \dfrac{L}{2}\|x - y^k\|^2 \right\}, \end{cases} \tag{3.4}$$

where Ω is a nonempty closed convex set of $\mathbb{R}^n$. The convergence results of this algorithm in [3] are as follows.

LEMMA 3.4. *[3] Let $\{x^k\}$ be a sequence generated by (3.4). When $\alpha > 3$, the following statements hold:*

(i) *the iterates $\{x^k\}$ converge to a global minimizer of $\min_\Omega f$;*
(ii) $\lim_{k\to\infty} k^2(f(x^k) - \min f) = 0$ *and* $\lim_{k\to\infty} k\|x^{k+1} - x^k\| = 0$.

Using the above lemma, we obtain the following significant conclusions for the proposed FIHT algorithm.

THEOREM 3.5. *When $\alpha > 3$, sequence $\{x^k\}$ generated by the FIHT algorithm satisfies*

(i) $\lim_{k\to\infty} x^k = x^*$, *where x^* is a local minimizer of problem (1.1) satisfying the lower bound property:*

$$|x_i^*| \geq \delta \quad or \quad x_i^* = 0, \quad \forall i = 1, 2, \ldots, n, \tag{3.5}$$

where δ is defined as in (3.1). Moreover, $\lim_{k\to\infty} f(x^k) = f_\infty < \infty$ and $\lim_{k\to\infty} F(x^k) = F_\infty < \infty$ exist;
(ii) $\lim_{k\to\infty} k\|x^{k+1} - x^k\| = 0$;
(iii) $\lim_{k\to\infty} k^2(f(x^k) - f_\infty) = 0$ *and* $\lim_{k\to\infty} k^2(F(x^k) - F_\infty) = 0$.

Proof. (i). Statement (ii) of Lemma 3.3 implies that there exist $\bar{K} \in \mathbb{N}$ and $J \subset \{1, 2, \ldots, n\}$ such that $I(x^k) = J$, $\forall k \geq \bar{K}$. Therefore, we have

$$x^{k+1} = \arg\min_{x\in\mathcal{X}_J} \left\{ f(y^k) + \langle \nabla f(y^k), x - y^k\rangle + \frac{L}{2}\|x - y^k\|^2 \right\}, \quad \forall k \geq \bar{K}. \tag{3.6}$$

Observing the above relation and Lemma 3.4, we know $\lim_{k\to\infty} x^k = x^*$, where x^* is a global minimizer of $\min_{\mathcal{X}_J} f$, which is a local minimizer of problem (1.1). Combining this with Lemma 3.1-(i), we deduce that x^* has the lower bound property in (3.5). This together with the continuity of f, we obtain all the results in this statement.

(ii). In view of (3.6) and Lemma 3.4-(ii), we have the estimation in item (ii).

(iii). From Lemma 3.3-(ii) and (3.5), there exists $\bar{K} \in \mathbb{N}$ such that

$$\|x^k\|_0 = \|x^*\|_0, \quad \forall k \geq \bar{K}.$$

Further, it holds that

$$f(x^k) - f(x^*) = F(x^k) - F(x^*), \quad \forall k \geq \bar{K}. \tag{3.7}$$

By (3.6) and Lemma 3.4-(ii), we obtain the estimation in (iii). ∎

If the extrapolation coefficients are chosen below some threshold, the authors in [35] proved that the iterates and the function value sequence of the proximal gradient algorithm with extrapolation are R-linear convergent when the objective function satisfies the error bound condition. It's worth emphasizing that when loss function f satisfies the error bound condition and β_k in the FIHT algorithm satisfies $\sup_k \beta_k < 1$, the R-linear convergence of sequence $\{x^k\}$ and objective function values can also be obtained for the FIHT algorithm. For easy reading, the convergence results obtained in this paper are summarized in Table 1.

TABLE 1
Summary of convergence of the SFIHT algorithm

f	β_k in Step 1	σ	local minimizer	lower bound	$\{F(x^k) - F_\infty\}$
nonsmooth	$\left[0, \sqrt{\frac{\mu_k}{\mu_{k-1}}}\right)$	$(0, 2)$	—	—	—
	(2.33)	$[\frac{1}{2}, 1]$	✓	—	—
	(2.41)	$(0, 1)$	✓	—	$O(k^{-\sigma})$
		1	✓	—	$O(k^{-1}\ln k)$
		$(1, 2)$	✓	—	$O(k^{\sigma-2})$
smooth	$\frac{k-1}{k+\alpha-1}$	—	✓	✓	$o(k^{-2})$

4. Experimental results. The aim of this section is to verify the theoretical results and performance of the proposed two algorithms by some numerical experiments. The SFIHT algorithm without extrapolation is called the smoothing iterative hard thresholding (SIHT) algorithm in this paper. Example 4.1 and Example 4.2 are an under-determined linear regression problem and an over-determined censored regression problem, respectively. The purpose of Example 4.1 and Example 4.2 is to illustrate the ability of the SFIHT algorithm for solving the problem, and to compare the good performance of the SFIHT algorithm with respect to the SIHT algorithm. In Example 4.3, we use the FIHT algorithm to solve the under-determined ℓ_0 regularized least squares problem. At the same time, we compare the performance of FIHT algorithm and IHT algorithm in Example 4.3. For different problems, we choose appropriate equilibrium parameter λ to adjust the data fitting and sparsity. The CPU time (in seconds) reported here doesn't include the time of data initialization.

The numerical experiments are performed in Python 3.7.0 on a 64-bit Lenovo PC with an Intel(R) Core(TM) i7-10710U CPU @1.10GHz 1.61GHz and 16GB RAM.

For any given $\epsilon > 0$, we call x^ϵ an ϵ local minimizer of problem (1.1), if it holds

$$\|[\nabla \tilde{f}(x^\epsilon, \mu)]_{I(x^\epsilon)^c}\|_\infty \le \epsilon \quad \text{and} \quad \mu \le \epsilon,$$

where $I(x^\epsilon)^c = \{i : x_i^\epsilon \neq 0\}$. We stop the proposed algorithm if x^k is an ϵ local minimizer of problem (1.1) or the number of iteration k exceeds 15000. Set some fixed parameters $\mu_0 = 0.7$, $L = 2L_{\tilde{f}}$ and $\alpha = 4$ throughout the numerical experiments.

Example 4.1 We consider the following ℓ_0 regularized linear regression problem:

$$\min_{-\mathbf{1} \le x \le \mathbf{1}} F(x) := \|Ax - b\|_1 + \lambda \|x\|_0, \tag{4.1}$$

where $A \in \mathbb{R}^{m\times n}$ with $m < n$, $b \in \mathbb{R}^m$. We choose a smoothing function of the ℓ_1 loss

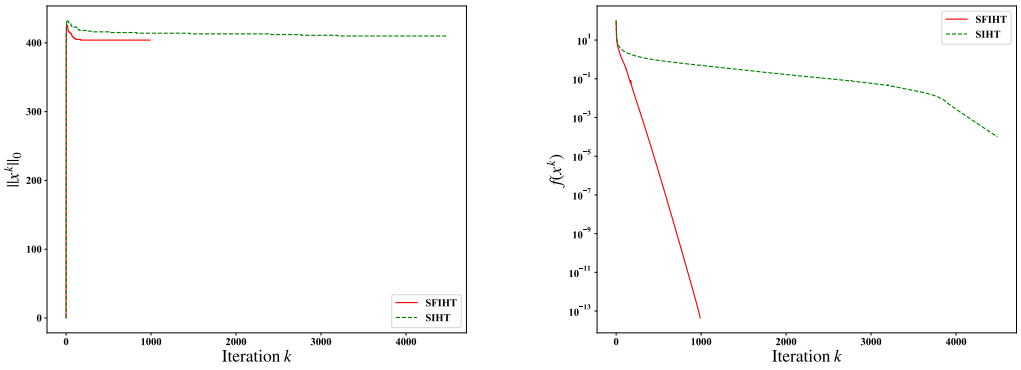
function as below, and it satisfies the conditions in Definition 2.1,

$$\tilde{f}(x,\mu)=\sum_{i=1}^{m}\tilde{\theta}(A_ix-b_i,\mu)\quad\text{with}\quad\tilde{\theta}(z,\mu)=\begin{cases}|z| & \text{if } |z|>\mu,\\ \dfrac{z^2}{2\mu}+\dfrac{\mu}{2} & \text{if } |z|\le\mu.\end{cases}\tag{4.2}$$

Three choices of β_k in Step 1 and Step 3 of the SFIHT algorithm are set as follows: $\beta_k=\frac{k-1}{k+\alpha-1}\sqrt{\left(1-\frac{1}{2k^{1-\sigma}}\right)\frac{\mu_k}{\mu_{k-1}}}$ (Step 1), $\beta_k=\sqrt{\frac{L-L_{\tilde{f}}}{4L}\frac{\mu_k}{\mu_{k-1}}}$ (Step 3b) and $\beta_k=\sqrt{\frac{L-L_{\tilde{f}}}{8L-4L_{\tilde{f}}}\frac{\mu_k}{\mu_{k-1}}}$ (Step 3b-2). Denote s the ℓ_0 norm of true solution $x^\star$, i.e., $\|x^\star\|_0=s$. For positive integers m, n and s, the data is generated as follows:

$$\bar{x}=\texttt{zeros}(n,1);\ \ \bar{n}=\texttt{randperm}(n);\ \ \bar{x}(\bar{n}(1:s))=\texttt{randn}(s,1);\ \ a=0.005;$$
$$x^\star=(\texttt{median}([\bar{x}';l';u']))';\ \ A=\texttt{orth}(\texttt{randn}(m,n)')';\ \ b=A*x^\star+a*\texttt{randn}(m,1).$$

Set $\epsilon=10^{-3}$, $L_{\tilde{f}}=\|A^{\mathrm{T}}A\|$, $\sigma=0.95$ and $x^0=\text{zeros}(n,1)$. We randomly generate A, b and $x^\star$ with $(m,n)=(300,1000)$ and $(m,n)=(500,5000)$. Fig. 1(a) shows that the support set of sequences $\{x^k\}$ generated by the SFIHT and SIHT algorithm only change finite times and are convergent. Observing Fig. 1(b), we find that compared with the SIHT algorithm, the SFIHT algorithm can find a better solution with fewer iterations. From Fig. 1, we see that the convergence rate of the SFIHT algorithm is faster than the SIHT algorithm, and the sparsity of the solution obtained by the SFIHT algorithm is also closer to the true solution $x^\star$. For three different stopping criterions, we record the CPU time and iterations of the two algorithms in Table 2. It's clear that the computational cost of the SFIHT algorithm is much less than that of the SIHT algorithm. The two algorithms find ϵ local minimizer with $\epsilon=10^{-4}$ with the same iterations, because μ_k doesn't meet the termination condition when $\|[\nabla\tilde{f}(x^k,\mu_k)]_{I(x^k)^c}\|_\infty\le\epsilon$ holds.



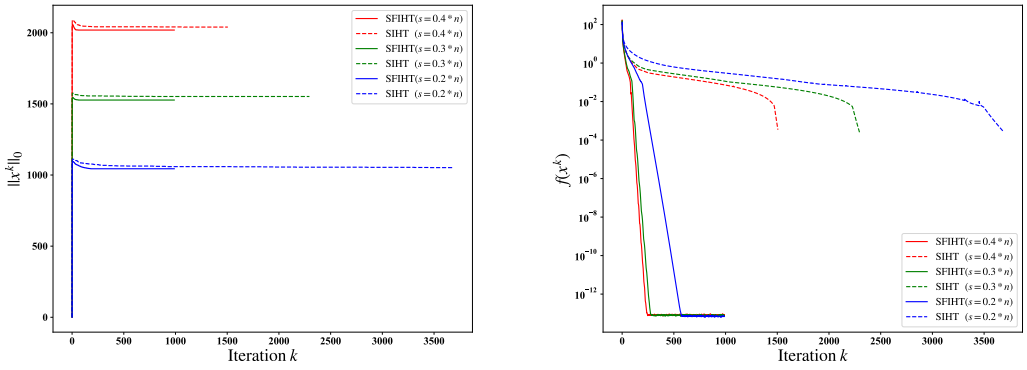
(a) (b)

FIG. 1. *Convergence of cardinality and loss function values for Example 4.1 with* $m=300$, $n=1000$ *and* $s=400$.

When $m=500$ and $n=5000$, we generate the problem data with three different sparsity levels, which are 20%, 30% and 40%. The results drawn in Fig. 2 show that the SFIHT algorithm substantially outperforms the SIHT algorithm in terms of the solution quality both on the loss function value and the cardinality.

TABLE 2
Computational cost for Example 4.1 with different stopping criterions when $m = 300$, $n = 1000$ and $s = 400$.

ϵ	10^{-2}		10^{-3}		10^{-4}	
Algorithm	SFIHT	SIHT	SFIHT	SIHT	SFIHT	SIHT
Time	**0.168**	23.388	**1.644**	27.108	**192.258**	209.628
Iterations	**217**	4150	**988**	4487	11155	11155



(a) (b)

FIG. 2. *Convergence of cardinality and loss function values for Example 4.1 with $m = 500$, $n = 5000$ and different sparsity levels.*

Example 4.2 We consider the following ℓ_0 regularized censored regression problem:

$$\min_{\mathbf{0}\le x\le \mathbf{1}} F(x) := \frac{1}{m}\|\max\{Ax, 0\} - b\|_1 + \lambda\|x\|_0,$$

where $A \in \mathbb{R}^{m\times n}$ with $m > n$ and $b \in \mathbb{R}^m$. A smoothing function satisfying Definition 2.1 of the above loss function is defined by

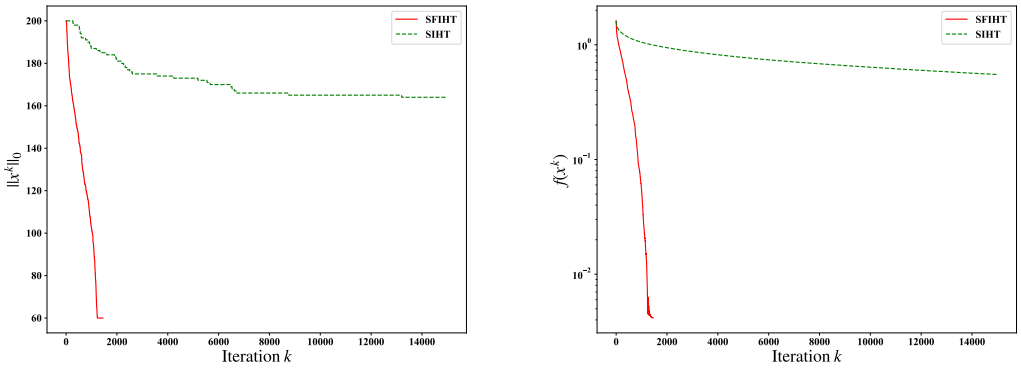
$$\tilde{f}(x,\mu) = \frac{1}{m}\sum_{i=1}^{m}\tilde{\theta}(\tilde{\phi}(A_i x,\mu) - b_i,\mu) \quad \text{with} \quad \tilde{\phi}(s,\mu) = \begin{cases} \max\{s,0\} & \text{if } |s| > \mu, \\ \dfrac{(s+\mu)^2}{4\mu} & \text{if } |s| \le \mu. \end{cases}$$

Set $\epsilon = 10^{-2}$, $L_{\tilde{f}} = \frac{3}{2m}\|A^{\mathrm{T}}A\|$, $\sigma = 0.7$ and $x^0 = 0.1 * \mathrm{ones}(n,1)$. β_k in Step 1 of the SFIHT algorithm is the same as that in (2.41), and the others are the same as in Example 4.1. We randomly generate the problem data as follows:

$$A = \texttt{randn}(m,n);\ \ \bar{n} = \texttt{randperm}(n);\ \ x^\star = \texttt{zeros}(n,1);$$
$$x^\star(\bar{n}(1:s)) = \texttt{unifrnd}(0.1,1,[s,1]);\ \ b = \max(A * x^\star + 0.01 * \texttt{randn}(m,1), 0).$$

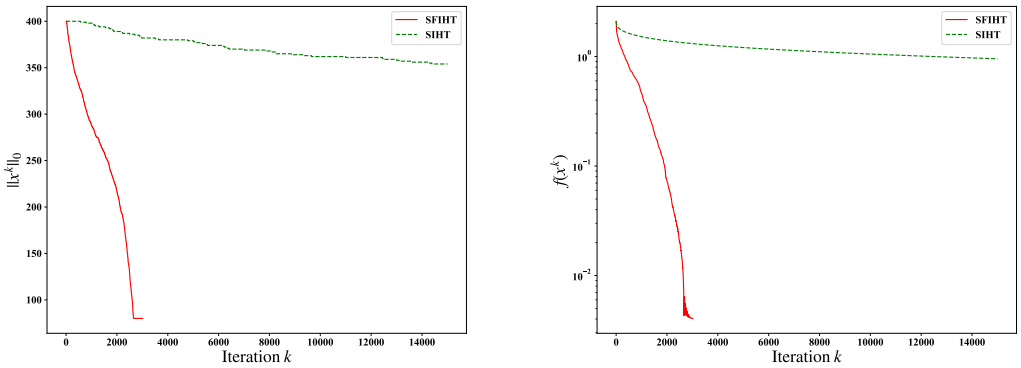
In this example, we run numerical experiments with $(m,n,s) = (1000,200,60)$ and $(m,n,s) = (2000,400,80)$. Results recorded in Fig. 3 and Fig. 4 show that the SFIHT algorithm performs much better than the SIHT algorithm in terms of both

the ℓ_0 norms and loss function values. Similarly, the SFIHT algorithm needs much less iterations to get an ϵ local minimizer of problem (1.1) than the SIHT algorithm.



(a) (b)

FIG. 3. *Convergence of cardinality and loss function values for Example 4.2 with* $m = 1000$, $n = 200$ *and* $s = 60$.



(a) (b)

FIG. 4. *Convergence of cardinality and loss function values for Example 4.2 with* $m = 2000$, $n = 400$ *and* $s = 80$.

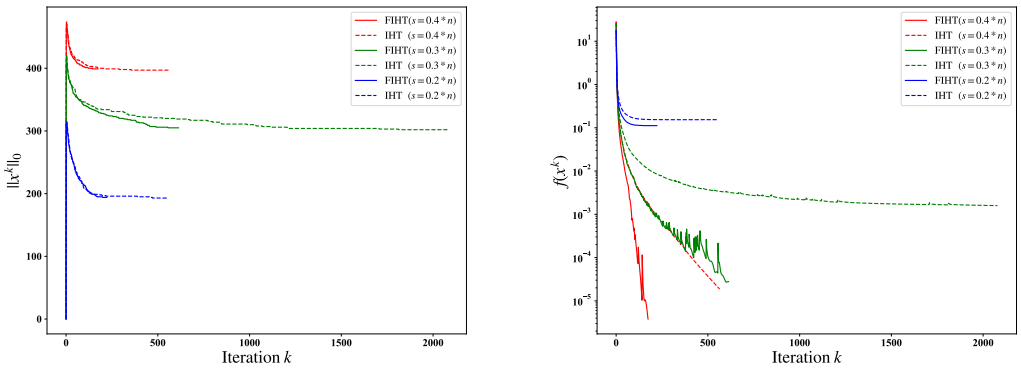
Example 4.3 We consider the following ℓ_0 regularized least squares problem:

$$\min_{\mathbf{0}\le x\le \mathbf{5}} F(x) := \frac{1}{2}\|Ax - b\|^2 + \lambda\|x\|_0 \tag{4.3}$$

where $A \in \mathbb{R}^{m\times n}$ with $m < n$ and $b \in \mathbb{R}^m$.

We set $\epsilon = 10^{-4}$, $L_f = \|A^{\mathrm{T}}A\|$ and $x^0 = \mathrm{zeros}(n, 1)$. We take the extrapolation coefficients $\beta_k = \sqrt{\frac{k}{k+1}\frac{L-L_f}{4L}}$ (Step 3b) and $\beta_k = \sqrt{\frac{k}{k+1}\frac{L-L_f}{8L-4L_f}}$ (Step 3b-2) in the FIHT algorithm. For two cases of (m, n) and three cases of s, we randomly generate

the problem data as them in Example 4.1. When $(m, n) = (300, 1000)$, for the different choices of s, the cardinalities and loss function values versus iteration k are plotted in Fig. 5. From this figure, we can see that the total iterations of the FIHT algorithm is much smaller than that of the IHT algorithm, and both the cardinalities and loss function values of the final output iterate obtained by the FIHT algorithm are more accurate. Fig. 6 illustrates the convergence rate of the FIHT algorithm is also faster than that of the IHT algorithm for solving problem (4.3) when the dimension of the problems is larger. It's also worth emphasizing that the loss function value at iterative point generated by SFIHT algorithm is smaller for all iterations. For different values of ϵ, the CPU time and iterations for obtaining an ϵ local minimizer by the FIHT algorithm and the IHT algorithm are recorded in Table 3. From the above results, We can observe that the FIHT algorithm performs much better than the IHT algorithm.



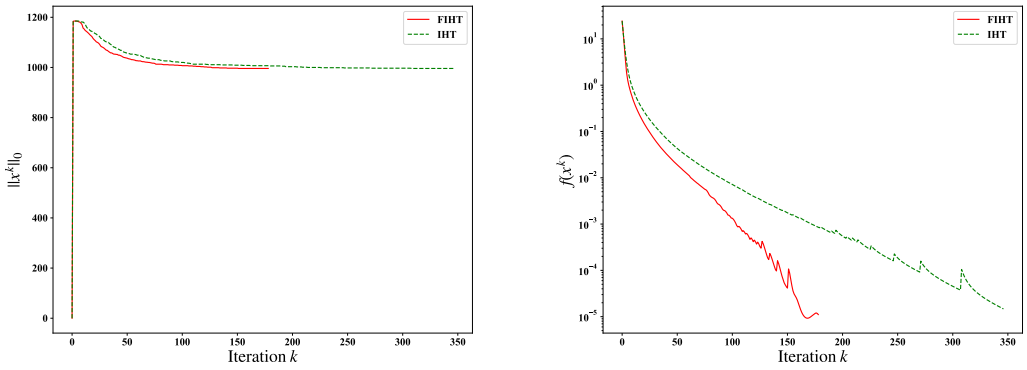
(a) (b)

FIG. 5. *Convergence of cardinality and loss function values for Example 4.3 with* $m = 300$, $n = 1000$ *and different sparsity levels.*

TABLE 3
Computational cost for Example 4.3 with different stopping criterions when $m = 500$, $n = 5000$ *and* $s = 1000$.

ϵ	10^{-2}		10^{-3}		10^{-4}		10^{-5}	
Algorithm	FIHT	IHT	FIHT	IHT	FIHT	IHT	FIHT	IHT
Time	0.180	**0.144**	**0.696**	0.738	**1.302**	2.322	**2.280**	4.41
Iterations	**26**	37	**92**	144	**178**	346	**284**	542

5. Conclusions. The main contribution of this paper is to propose an effective and fast algorithm for solving the constrained cardinality penalty problem with a continuous convex loss function, and analyze its convergence properties. We first use a parametric smoothing approximation of the loss function to generate a cardinality penalty problem with smooth loss function. Then, the iterative hard thresholding algorithm with extrapolation is used, in which the smoothing parameter is updated step by step. The only one subproblem in the proposed algorithm has a closed-form



(a) (b)

FIG. 6. *Convergence of cardinality and loss function values for Example 4.3 with* $m = 500$, $n = 5000$ *and* $s = 1000$.

solution, and the extrapolation coefficients can be chosen to satisfy $\sup_k \beta_k = 1$. After finitely many iterations, the support set of the iterate does not change. Further, we give sufficient conditions to guarantee that any accumulation point of $\{x^k\}$ generated by the proposed algorithm is a local minimizer of the considered problem. For a class of extrapolation coefficients, we obtain not only the convergence of the sequence, but also the convergence rate of $O(\ln k/k)$ on the loss and objective function values. Moreover, we consider in particular the case that the loss function is a Lipschitz continuous convex function. To solve it, we provide an algorithm, which can be viewed as a specific form of the above proposed algorithm. This algorithm owns both sequence convergence on the iterates and convergence rate of $o(k^{-2})$ on the loss and objective function values. Additionally, the limit point not only is a local minimizer of the considered problem but also possesses a desirable lower bound.

REFERENCES

[1] S. ADLY AND H. ATTOUCH, *Finite convergence of proximal-gradient inertial algorithms combining dry friction with Hessian-driven damping*, SIAM J. Optim., 30 (2020), pp. 2134–2162.
[2] H. ATTOUCH AND S. LÁSZLÓ, *Newton-like inertial dynamics and proximal algorithms governed by maximally monotone operators*, SIAM J. Optim., 30 (2020), pp. 3252–3283.
[3] H. ATTOUCH AND J. PEYPOUQUET, *The rate of convergence of Nesterov's accelerated forward-backward method is actually faster than* $1/k^2$, SIAM J. Optim., 26 (2016), pp. 1824–1834.
[4] A. BECK AND M. TEBOULLE, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM J. Imaging Sci., 2 (2009), pp. 183–202.
[5] W. BIAN, *Smoothing accelerated algorithm for constrained nonsmooth convex optimization problems (in chinese)*, Sci. Sin. Math., 50 (2020), pp. 1651–1666.
[6] W. BIAN AND X. CHEN, *Optimality and complexity for constrained optimization problems with nonconvex regularization*, Math. Oper. Res., 42 (2017), pp. 1063–1084.
[7] W. BIAN AND X. CHEN, *A smoothing proximal gradient algorithm for nonsmooth convex regression with cardinality penalty*, SIAM J. Numer. Anal., 58 (2020), pp. 858–883.
[8] T. BLUMENSATH AND M. DAVIES, *Sparse and shift-invariant representations of music*, IEEE Trans. Audio Speech Lang. Process., 14 (2006), pp. 50–57.
[9] T. BLUMENSATH AND M. DAVIES, *Iterative thresholding for sparse approximations*, J. Fourier Anal. Appl., 14 (2008), pp. 629–654.
[10] T. BLUMENSATH AND M. DAVIES, *Iterative hard thresholding for compressed sensing*, Appl.

Comput. Harmon. Anal., 27 (2009), pp. 265–274.
[11] A. Bruckstein, D. Donoho, and M. Elad, *From sparse solutions of systems of equations to sparse modeling of signals and images*, SIAM Rev., 51 (2009), pp. 34–81.
[12] E. Candes, J. Romberg, and T. Tao, *Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information*, IEEE Trans. Inf. Theory, 52 (2006), pp. 489–509.
[13] A. Chambolle, R. DeVore, N. Lee, and B. Lucier, *Nonlinear wavelet image processing: variational problems, compression, and noise removal through wavelet shrinkage*, IEEE Trans. Image Process., 7 (1998), pp. 319–335.
[14] X. Chen, *Smoothing methods for nonsmooth, nonconvex minimization*, Math. Program., 134 (2012), pp. 71–99.
[15] P. Combettes and V. Wajs, *Signal recovery by proximal forward-backward splitting*, Multiscale Model. Simul., 4 (2005), pp. 1168–1200.
[16] W. Dai and O. Milenkovic, *Subspace pursuit for compressive sensing signal reconstruction*, IEEE Trans. Inf. Theory, 55 (2009), pp. 2230–2249.
[17] I. Daubechies, M. Defrise, and C. De Mol, *An iterative thresholding algorithm for linear inverse problems with a sparsity constraint*, Commun. Pure Appl. Math., 57 (2004), pp. 1413–1457.
[18] N. Doikov and Y. Nesterov, *Contracting proximal methods for smooth convex optimization*, SIAM J. Optim., 30 (2020), pp. 3146–3169.
[19] D. Donoho, *Compressed sensing*, IEEE Trans. Inf. Theory, 52 (2006), pp. 1289–1306.
[20] J. Fan and R. Li, *Variable selection via nonconcave penalized likelihood and its oracle properties*, J. Am. Stat. Assoc., 96 (2001), pp. 1348–1360.
[21] E. Hale, W. Yin, and Y. Zhang, *Fixed-point continuation for ℓ_1-minimization: methodology and convergence*, SIAM J. Optim., 19 (2008), pp. 1107–1130.
[22] S. Hoda, A. Gilpin, J. Pena, and T. Sandholm, *Smoothing techniques for computing Nash equilibria of sequential games*, Math. Oper. Res., 35 (2010), pp. 494–512.
[23] Y. Liu and Y. Wu, *Variable selection via a combination of the ℓ_0 and ℓ_1 penalties*, J. Comput. Graph. Stat., 16 (2007), pp. 782–798.
[24] Z. Lu, *Iterative hard thresholding methods for ℓ_0 regularized convex cone programming*, Math. Program., 147 (2014), pp. 125–154.
[25] Z. Lu and Y. Zhang, *Sparse approximation via penalty decomposition methods*, SIAM J. Optim., 23 (2013), pp. 2448–2478.
[26] S. Mallat and Z. Zhang, *Matching pursuits with time-frequency dictionaries*, IEEE Trans. Signal Process., 41 (1993), pp. 3397–3415.
[27] Y. Nesterov, *Smooth minimization of non-smooth functions*, Math. Program., 103 (2005), pp. 127–152.
[28] Y. Nesterov, *Gradient methods for minimizing composite functions*, Math. Program., 140 (2013), pp. 125–161.
[29] M. Nikolova, *Local strong homogeneity of a regularized estimator*, SIAM J. Appl. Math., 61 (2000), pp. 633–658.
[30] Y. Pati, R. Rezaiifar, and P. Krishnaprasad, *Orthogonal matching pursuit-recursive function approximation with applications to wavelet decomposition*, in Conference Record of the Twenty-Seventh Asilomar Conference on Signal, Systems and Computers, vol. 1-2, 1993, pp. 40–44.
[31] D. Peleg and R. Meir, *A bilinear formulation for vector sparsity optimization*, Signal Process., 88 (2008), pp. 375–389.
[32] E. Soubies, L. Blanc-Feraud, and G. Aubert, *A continuous exact ℓ_0 penalty (CEL0) for least squares regularized problem*, SIAM J. Imaging Sci., 8 (2015), pp. 1607–1639.
[33] W. Su, S. Boyd, and E. Candès, *A differential equation for modeling Nesterov's accelerated gradient method: theory and insights*, J. Mach. Learn. Res., 17 (2016), pp. 1–43.
[34] R. Tibshirani, *Regression shrinkage and selection via the Lasso*, J. R. Stat. Soc. Ser. B-Methodol., 58 (1996), pp. 267–288.
[35] B. Wen, X. Chen, and T. Pong, *Linear convergence of proximal gradient algorithm with extrapolation for a class of nonconvex nonsmooth minimization problems*, SIAM J. Optim., 27 (2017), pp. 124–145.
[36] F. Wu and W. Bian, *Accelerated iterative hard thresholding algorithm for ℓ_0 regularized regression problem*, J. Glob. Optim., 76 (2020), pp. 819–840.
[37] Z. Zheng, Y. Fan, and J. Lv, *High dimensional thresholded regression and shrinkage effect*, J. R. Stat. Soc. Ser. B-Stat. Methodol., 76 (2014), pp. 627–649.