A Convergent Iterative Hard Thresholding for Nonnegative Sparsity Optimization

Lili Pan[1,2], Shenglong Zhou[3], Naihua Xiu[1], Houduo Qi[3]

Abstract: The iterative hard thresholding (IHT) algorithm is a popular greedy-type method in (linear and nonlinear) compressed sensing and sparse optimization problems. In this paper, we give an improved iterative hard thresholding algorithm for solving the nonnegative sparsity optimization (NSO) by employing the Armijo-type stepsize rule, which automatically adjusts the stepsize and support set and leads to a sufficient decrease of the objective function each iteration. Consequently, the improved IHT algorithm enjoys several convergence properties under standard assumptions. Those include the convergence to α-stationary point (also known as L-stationary point in literature if the objective function has Lipschitz gradient) and the finite identification of the true support set. We also characterize the conditions that the full sequence converges to a local minimizer of NSO and establish its linear convergence rate. Extensive numerical experiments are included to demonstrate the good performance of the proposed algorithm.

Keywords: sparsity constrained optimization, improved iterative hard thresholding, convergence, convergence rate, numerical experiment
Mathematics Subject Classification: 90C26, 90C30, 90C90

---

1 Introduction

In this paper, we are concerned with efficient numerical methods for the nonnegative sparsity optimization (NSO for short):

$$\min \quad f(x), \qquad \text{s.t.} \qquad x \in S \cap \mathbb{R}^n_+, \tag{1}$$

where $f(x) : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable and is bounded from below, $s < n$ is a positive integer and defines the sparse set $S := \{x \in \mathbb{R}^n : \|x\|_0 \le s\}$ with $\|x\|_0$ being the l_0-norm of x (the number of nonzero elements in x) and $\mathbb{R}^n_+$ is the nonnegative orthant in $\mathbb{R}^n$. Important examples covered by (1) include the linear compressed sensing problem of $f(x) = f_A(x) := \|Ax - b\|^2$ with $A \in \mathbb{R}^{m\times n}$ being a linear measurement matrix and $b \in \mathbb{R}^m$ is the observation vector (see, e.g., [19] for an extensive treatment of this problem) and the nonlinear compressed sensing problem of $f(x) = f_\Phi(x) := \|\Phi(x) - b\|^2$ with $\Phi : \mathbb{R}^n \mapsto \mathbb{R}^m$ being a nonlinear measurement function [8]. Another important example is $f(x)$ being a regularized logistic regression cost function [2, Sect. 4]. Those examples of (1) may or may not have the nonnegativity constraint $x \ge 0$. We include it mainly because of the two reasons. One is that in many real-world problems the underlying parameters represent quantities that can only take on nonnegative values, e.g., amounts of materials, chemical concentrations, pixel intensities, to name a few [16]. Another reason is that it is one of the prototype examples of the symmetric set considered by Beck and Hallak [6] in their nonlinear sparse optimization. We hope that by including the nonnegativity constraint some of our results may have their counterparts when a more general symmetric set is used instead of $\mathbb{R}^n_+$.

It is usually expected that numerical methods for the linear (nonlinear) compressed sensing should naturally extend to solve (1). A class of such methods are of the greedy methods. One advantage of these methods is that they are generally faster than the relaxation approaches, which often lead to separable convex programming problems that can be solved, for example, by methods of alternating directions or splitting methods [21]. Another advantage is that many of them have stable recovery properties under some conditions [15]. A variety of greedy methods have been proposed in compressed sensing, such as matching pursuit (MP) [29], orthogonal MP (OMP) [18], compressive sampling matching pursuit (CoSaMP) [31], subspace pursuit (SP) [17], hard thresholding pursuit (HTP) [20], conjugate gradient iterative hard

The work was supported in part by the National Natural Science Foundation of China (11431002, 71611130218) and Shandong Province Natural Science Foundation (ZR2016AM07). 1. Department of Applied Mathematics, Beijing Jiaotong University, Beijing 100044, P. R. China; 2. Department of Mathematics, Shandong University of Technology, Zibo 255049, P. R. China; 3. School of Mathematics, University of Southampton, Southampton SO17 1BJ, United Kingdom. L. Pan (panlili1979@163.com), S. Zhou (longnan_zsl@163.com), N. Xiu (nhxiu@bjtu.edu.cn), H. Qi (hdqi@soton.ac.uk). Date: February 13, 2017.

thresholding (CGIHT) [7], to name just a few. Some of those methods have been extended to the sparsity constrained nonlinear optimization. For example, Bahmani et al. [2] proposed a gradient hard-thresholding method which generalizes CoSaMP. Yuan et al. [39] generalized HTP to the sparsity constrained convex optimization. Yuan and Liu [40] proposed a Newton greedy pursuit (NTGP) method to approximately minimize a twice differentiable function over the sparsity constraint.

In particular, the iterative hard thresholding (IHT) algorithm, a popular greedy method which was proposed for the linear compressed sensing problem by Blumensath and Davies in [9, 10] (and later extended to the nonlinear case by Blumensath [8]), has attracted much attention due to its nice recovery properties. For example, when the matrix A in defining f_A is of full row-rank and its spectral norm satisfies $\|A\|_2 < 1$, IHT converges to a local minimum [9]. Furthermore, it was observed in [11] that the algorithm may fail to converge if the spectral norm condition is violated. They then proposed a normalised IHT (NIHT) with an adaptive stepsize by the line search strategy and proved its convergence to a local minimum if A is of full row-rank and is s-regular (i.e., any s columns of A are linearly independent). A latest result of Cartis and Thompson [15] showed that NIHT converges to a local minimum if the matrix A is $2s$-regular.

There recently emerges a new line of research on those problems mainly attempted from the numerical optimization community [1, 5, 6, 34, 35, 27], which tend to ask the following fundamental questions:

(Q1) Towards what *stationary* points that a generated sequence converges?

(Q2) Under what conditions that such a stationary point may become a local/global minimizer?

(Q3) What is the convergence rate to a local/global minimizer if the convergence is taking place?

There are two key elements that seem to be indispensable in the delicate analysis among the existing literature in answering those questions. One is on introducing a well characterized stationarity appropriate to the data at hand and the other is on a well defined stepsize rule that is to force certain sufficient decrease in the merit function used in the respective algorithms. For example, assuming that the function f has Lipschitz gradient, Beck and Eldar [5] introduced L-stationarity (among others) and by using a fixed or the accurate minimization stepsize rule, they established the convergence to an L-stationary point of various algorithms including IHT. See [6] for further results along this line on the sparse optimization problem with a symmetric constraint set. The results in [6] were further significantly enhanced by Lu [27] by employing a nonmonotone line search stepsize rule. When f is nondifferentiable, Attouch et al. [1] introduced the concept of critical point and showed that a few classes of algorithms actually converge to such a critical point. In particular, a variant of IHT with a fixed (or varying) stepsize on the linear compressed sensing problem is proved to converge to a critical point [1, Example 5.4], which in this special case is also the L-stationary point of Beck and Eldar [5]. We note that (Q3) is hardly addressed in the literature.

In contrast to the research reviewed above, the stationarities studied by Pan et al. [34, 35] followed the classical derivation of optimality conditions for nonconvex programming and are based on Bouligand or Clarke tangent cones for nonconvex sets (see [12, Section 6.3] for the definitions of those two cones). This leads to B-, C- and α-stationarites. Their relationships to L-stationarity (and others) have been briefly discussed in [6, Remark 5.3]. The blanket assumption used in [34, 35] is that f is continuously differentiable (its gradient is not necessarily Lipschitzian).

In this paper, we continue the research of [35, 34] by applying their stationarities to the algorithm of IHT with the Armijo stepsize rule to solve (1). In answering the questions (Q1)-(Q3), we asked whether our obtained results have been as general as they can be. This effort has led to the important relationships among the global/local minimizer and the three stationary points (α-, B-, and C-stationarities) in Theorem 2.1 and Figure 1, which clearly show what extra conditions are required for one to imply another. This theorem is fundamental to our algorithmic analysis later on. It turns out that the extra conditions needed are satisfied by the restricted strong convexity and restricted strong smoothness of f. Both of the concepts are introduced and popularized in [32]. The resulting IHT enjoys a number of very nice convergence properties. We single out a few that partially answered the questions (Q1)-(Q3):

(i) (for Q1) Any accumulation point of iterative sequence is an α-stationary point of NSO if the objective function f is $2s$-restricted strongly smooth (Theorem 3.1).

(ii) (for Q2) The full iterative sequence converges to a local minimizer of NSO if f is $2s$-restricted strongly smooth and $2s$-restricted strongly convex (Theorem 3.2).

(iii) (for Q3) The sequence of functional values converges at a sublinear rate if f is $2s$-restricted strongly smooth and $2s$-restricted strongly convex (Theorem 3.3). Furthermore, the sequence of iterates converges at a Q-linear rate under the condition that the sparsity constraint is tight at the solution (Theorem 3.4).

In addition, the numerical performance of our improved IHT is also very satisfactory for a large number of commonly tested problems. Finally, we would like to emphasize that one of our convergence results, namely Thm. 3.2(ii), is similar to what have been reported in [6, 27], but under different assumptions and

on different algorithms. Our basic assumption is on the continuity of the gradient of f. When the gradient is also Lipschitzian, our α-startionarity becomes the L-stationarity. We will make more comments on the similarity right after Thm. 3.2.

This paper is organized as follows. Section 2 presents some technical results on the optimality conditions of (1). Section 3 contains the IIHT algorithm for (1) and proves its convergence properties. Numerical results are given in Section 4. The last section makes some concluding remarks. For the sake of easy reading, we introduce some notations to end this section.

Table 1: Notations used in the paper.

Notation	**Description**
S	The sparse set $\{x \in \mathbb{R}^n : \ \|x\|_0 \leq s\}$;
S_+	The feasible region of (1), i.e. $S \cap \mathbb{R}^n_+$;
$\mathrm{supp}(x)$	The support set of $x \in \mathbb{R}^n$, i.e., $\{i \in \{1, \cdots, n\} : \ x_i \neq 0\}$;
Γ^* (or Γ^k)	The support set of $x^* \in \mathbb{R}^n$ (or $x^k \in \mathbb{R}^n$);
Γ_{xy}	The union of support sets between $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^n$, i.e., $\mathrm{supp}(x) \cup \mathrm{supp}(y)$;
$\|\Gamma\|$	The cardinality of Γ;
x_Γ	The subvector consisting of entries of $x \in \mathbb{R}^n$ indexed by Γ;
A_Γ	The submatrix consisting of columns of $A \in \mathbb{R}^{m\times n}$ indexed by Γ;
∇f	The gradient of $f(x)$ on $\mathbb{R}^n$, i.e., $\nabla f(x)$;
$\nabla_\Gamma f(x)$	The subvector of $\nabla f(x)$ indexed on Γ, i.e., $(\nabla f(x))_\Gamma$;
$N(x, \delta)$	The neighbor region of $x \in \mathbb{R}^n$ with radius $\delta > 0$, i.e., $\{y \in \mathbb{R}^n : \ \|y - x\| < \delta\}$;
e_i	The vector in $\mathbb{R}^n$ whose ith component is one and others are zeros;
$x_i^\downarrow$	The ith largest (in absolute value) element of $x \in \mathbb{R}^n$.
$\mathbb{R}^n_\Gamma$	The subspace of $\mathbb{R}^n$ spanned by $\{\mathrm{e}_i : i \in \Gamma\}$, i.e., $\mathrm{span}\{\mathrm{e}_i : i \in \Gamma\}$

2 Characterizations of Various Stationarities

In this section, we will give detailed characterizations of the relationships among the three stationary points (namely, α-, B-, and C-stationary point) and the local/global minimizers of (1). We will also report some consequences of those characterizations under some additional conditions such as the restricted strong convexity/smoothness of f. Those results will be used in the convergence analysis of the improved IHT algorithm in later sections.

2.1 On the three stationarities

In this part, we assume that f is continuously differentiable. We will use the orthogonal projection onto a closed set $\Omega \subseteq \mathbb{R}^n$ defined as follows:

$$\mathrm{P}_\Omega(x) := \arg\min \left\{ \|y - x\|^2 : \ \text{ s.t. } \ y \in \Omega \right\},$$

where $\|\cdot\|$ is the Euclidean norm in $\mathbb{R}^n$. Since Ω is not convex, there may be multiple optimal solutions. In this case, $\mathrm{P}_\Omega(x)$ can be any one of them. In particular, when $\Omega = S$, $\mathrm{P}_\Omega(x)$ can be obtained by setting all but s largest absolute components of x to zero ($\mathrm{P}_\Omega(x)$ is also known as the support project of x onto S). Furthermore, it was proved in [35, Prop. 3.1] that

$$\mathrm{P}_{S_+}(x) = \mathrm{P}_S\Big(\mathrm{P}_{\mathbb{R}^n_+}(x)\Big). \tag{2}$$

Moreover,

$$x_{\Gamma_x} = y_{\Gamma_x} \qquad \text{for } \ x = \mathrm{P}_{S_+}(y). \tag{3}$$

The B- and C-stationary points are respectively defined through the orthogonal projection onto the Bouligand and Clarke tangent cones of S_+. In our analysis, we will only use an equivalent characterization of the each cone and will not need their original definitions, which are described in [12, Sect. 6.3] and [35, Sect. 2.1]. We note that the Bouligand tangent cone below can also be derived following [4].

Proposition 2.1 *[35, Thms. 2.1 and 2.2] (Characterizations of Bouligand and Clarke tangent cones). Recall from Table 1, Γ^* is the support set of $x^* \in \mathbb{R}^n$. If $x^* \in S$, the Bouligand and Clarke tangent cones of S at x^*, respectively denoted by $\mathrm{T}_S^B(x^*)$ and $\mathrm{T}_S^C(x^*)$ are given by*

$$\begin{aligned} \mathrm{T}_S^B(x^*) &= \begin{cases} \mathbb{R}^n_{\Gamma^*}, & \text{if } |\Gamma^*| = s, \\ \bigcup\limits_{\Upsilon \supseteq \Gamma^*, |\Upsilon| = s} \mathbb{R}^n_{\Upsilon}, & \text{if } |\Gamma^*| < s, \end{cases} \\ \mathrm{T}_S^C(x^*) &= \mathbb{R}^n_{\Gamma^*}. \end{aligned}$$

Furthermore, if $x^* \in S_+$ we have

$$\mathrm{T}_{S_+}^B(x^*) = \mathrm{T}_S^B(x^*) \cap \mathrm{T}_{\mathbb{R}^n_+}(x^*), \qquad \mathrm{T}_{S_+}^C(x^*) = \mathrm{T}_S^C(x^*),$$

where $\mathrm{T}_{\mathbb{R}^n_+}(x^*) := \{d \in \mathbb{R}^n : \ d_i \geq 0, i \notin \Gamma^*\}$ is the usual tangent cone of $\mathbb{R}^n_+$ at x^*.

The α-stationary point defined below is actually the L-stationary point [5] when f has Lipschitz gradient with the Lipschitz constant L_f. The difference lies in that α in our definition is allowed to take any positive value, while it is restricted within $0 < \alpha \leq 1/L_f$ in [5].

Definition 2.1 *Let $x^* \in S_+$ be a given feasible point of (1).*

(i) We say that x^* is an α-stationary point if there exists $\alpha > 0$ such that

$$x^* \in \mathrm{P}_{S_+}\left(x^* - \alpha \nabla f(x^*)\right).$$

(ii) We say the x^* is a B-stationary point if

$$0 \in \mathrm{P}_{\mathrm{T}^B_{S_+}(x^*)}(-\nabla f(x^*)).$$

(iii) We say the x^* is a C-stationary point if

$$0 = \mathrm{P}_{\mathrm{T}^C_{S_+}(x^*)}(-\nabla f(x^*)).$$

The following table is extracted from [35, Table 3], which is very useful in helping us understand the subtle differences among the definitions. We will frequently use those characterizations in our analysis below.

Table 2: Gradient characterizations of the three stationary points.

	$\|x^*\|_0 = s,\ x^* \geq 0$	$\|x^*\|_0 < s,\ x^* \geq 0$
α-stationary point	$\nabla_i f(x^*) \begin{cases} = 0, & i \in \Gamma^* \\ \geq -\alpha (x^*)^{\downarrow}_s, & i \notin \Gamma^* \end{cases}$	$\nabla_i f(x^*) \begin{cases} = 0, & i \in \Gamma^* \\ \in \mathbb{R}_+, & i \notin \Gamma^* \end{cases}$
B-stationary point	$\nabla_i f(x^*) \begin{cases} = 0, & i \in \Gamma^* \\ \in \mathbb{R}, & i \notin \Gamma^* \end{cases}$	$\nabla_i f(x^*) \begin{cases} = 0, & i \in \Gamma^* \\ \in \mathbb{R}_+, & i \notin \Gamma^* \end{cases}$
C-stationary point	$\nabla_i f(x^*) \begin{cases} = 0, & i \in \Gamma^* \\ \in \mathbb{R}, & i \notin \Gamma^* \end{cases}$	$\nabla_i f(x^*) \begin{cases} = 0, & i \in \Gamma^* \\ \in \mathbb{R}, & i \notin \Gamma^* \end{cases}$

It is well known that sparse optimization in general fundamentally differs from classical optimization. One way to appreciate such difference, as well demonstrated in [5], is that the classical variational inequality

$$\langle \nabla f(x^*),\ x - x^* \rangle \geq 0, \quad \forall x \in S$$

is not a necessary optimality condition. Interestingly, as proved below, within sufficiently small a neighbourhood of x^*, the variational inequality is equivalent to the B-stationary point.

Proposition 2.2 *Let $x^* \in S_+$. Then the following results hold.*

(i) x^* is a B-stationary point of (1) if and only if there exists δ satisfying $0 < \delta < \min\{x_i^* : \ i \in \Gamma^*\}$ such that

$$\langle \nabla f(x^*),\ x - x^* \rangle \begin{cases} = 0, & \text{if } \|x^*\|_0 = s \\ \geq 0, & \text{if } \|x^*\|_0 < s \end{cases} \tag{4}$$

holds for any $x \in N(x^*, \delta) \cap S_+$.

(ii) In particular, if $\|x^*\|_0 < s$ (the sparse constraint is not tight), then x^* is a B-stationary point of (1) if and only if

$$\langle \nabla f(x^*),\ x - x^* \rangle \geq 0,\ \forall\ x \in S_+.$$

Proof (i) (Only if part) Suppose first that $x^* \in S_+$ is a B-stationary point of (1). We prove (4) for any $x \in N(x^*, \delta) \cap S_+$ with some $\delta > 0$ by considering two cases.

Case 1. $\|x^*\|_0 = s$ and $x^* \geq 0$. Take δ satisfying $0 < \delta < \min\{x_i^* : i \in \Gamma^*\}$, for any $x \in N(x^*, \delta) \cap S_+$ and $i \in \Gamma^*$, we have

$$x_i = x_i^* - (x_i^* - x_i) \geq x_i^* - |x_i^* - x_i| > x_i^* - \delta > 0,$$

which yields that $\Gamma^* \subseteq \mathrm{supp}(x)$. By $\|x\|_0 \leq s$ and $|\Gamma^*| = \|x^*\|_0 = s$, we can obtain

$$\mathrm{supp}(x) \equiv \Gamma^*, \quad \forall x \in N(x^*, \delta) \cap S_+. \tag{5}$$

Since x^* is a B-stationary point of (1), by Table 2, we have

$$\nabla_i f(x^*) \begin{cases} = 0, & \text{for } \ i \in \Gamma^*, \\ \in \mathbb{R}, & \text{for } \ i \notin \Gamma^*, \end{cases}$$

which together with (5) yields that

$$\langle \nabla f(x^*),\ x - x^* \rangle = \sum_{i \in \Gamma^*} \nabla_i f(x^*)(x_i - x_i^*) + \sum_{i \notin \Gamma^*} \nabla_i f(x^*)(x_i - x_i^*) = 0.$$

Case 2. $\|x^*\|_0 < s$ and $x^* \geq 0$. Since x^* is a B-stationary point of (1), we have

$$\nabla_i f(x^*) \begin{cases} = 0, & i \in \Gamma^*, \\ \in \mathbb{R}_+, & i \notin \Gamma^*. \end{cases}$$

It follows that for any $\delta > 0$ and $x \in N(x^*, \delta) \cap S_+$,

$$\langle \nabla f(x^*),\ x - x^* \rangle = \sum_{i \in \Gamma^*} \nabla_i f(x^*)(x_i - x_i^*) + \sum_{i \notin \Gamma^*} \nabla_i f(x^*)(x_i - x_i^*) \geq 0, \tag{6}$$

where the last inequality follows from the facts that (a) $\nabla_i f(x^*) = 0$, for $i \in \Gamma^*$, and (b) $x_i \geq 0$, $x_i^* = 0, \nabla_i f(x^*) \geq 0$ for $i \notin \Gamma^*$. We note that this part of the proof also applies to all $x \in S_+$ without having to be restricted in a neighbourhood of x^*.

(i) (The if part) Conversely, suppose that $x^* \in S_+$ satisfies (4) for any $x \in N(x^*, \delta) \cap S_+$ and $0 < \delta < \min\{x_i^* : i \in \Gamma^*\}$. We show x^* is a B-stationary point of (1) also by two cases.

Case 1. $\|x^*\|_0 = s$ and $x^* \geq 0$. For any $i \in \Gamma^*$ and δ satisfying $0 < \delta < \min\{x_i^* : i \in \Gamma^*\}$, by letting $x = x^* + \delta \mathrm{e}_i/2$, we have $x \in N(x^*, \delta) \cap S_+$. It follows from (4) that

$$0 = \langle \nabla f(x^*),\ x - x^* \rangle = \langle \nabla f(x^*),\ \delta \mathrm{e}_i/2 \rangle = \delta \nabla_i f(x^*)/2.$$

Hence, $\nabla_i f(x^*) = 0$ for $i \in \Gamma^*$ and $\nabla_i f(x^*)$ for $i \notin \Gamma^*$ is not restricted.
Case 2. $\|x^*\|_0 < s$ and $x^* \geq 0$. If $i \in \Gamma^*$, using the same proof as Case 1 above, we obtain $\nabla_i f(x^*) = 0$. If $i \notin \Gamma^*$, let $x = x^* + \delta \mathrm{e}_i/2$. Then $x \in N(x^*, \delta) \cap S_+$. It follows from (4) that

$$0 \leq \langle \nabla f(x^*),\ x - x^* \rangle = \langle \nabla f(x^*),\ \delta \mathrm{e}_i/2 \rangle = \delta \nabla_i f(x^*)/2.$$

Hence, $\nabla_i f(x^*) \geq 0$ for $i \notin \Gamma^*$. It follows from Table 2 that x^* is a B-stationary point of (1).

(ii) The only-if part follows from Case 2 of the only-if part of (i), where it was noted that the proof does not rely on the neighbourhood of x^* used. The if-part proof follows from Case 2 of the if-part of (i), where the current condition in (ii) necessarily implies the condition within a neighborhood used in (i). □

Our next major result is to establish the relationships among the three stationary points and the global/local minimizers of (1). Some of the relationships need a certain kind of convexity. We choose to use the one of the restricted strong convexity introduced in [32, 22]. Slightly different forms of this concept were also presented in [2, 8, 39]. Note that these properties are all analogous to the restricted isometry property (RIP) [14] in the standard (linear) compressed sensing. For easy reference, we include a definition.

Definition 2.2 *A function f is called s-restricted strongly smooth (s-RSS) with parameter $L_s > 0$, if for any $x, y \in \mathbb{R}^n$ satisfying $|\Gamma_{xy}| \leq s$, it holds that*

$$f(y) - f(x) - \langle \nabla f(x),\ y - x \rangle \leq \frac{L_s}{2} \|y - x\|^2. \tag{7}$$

We say that the function f is s-restricted strongly convex (s-RSC) with parameter $l_s > 0$, if for any $x, y \in \mathbb{R}^n$ satisfying $|\Gamma_{xy}| \leq s$, it holds that

$$f(y) - f(x) - \langle \nabla f(x),\ y - x \rangle \geq \frac{l_s}{2}\|y - x\|^2. \tag{8}$$

In particular, if $l_s = 0$, the function f is said to be s-restricted convex (s-RC).

We note that f being s-RSS is a weaker condition than that f having a Lipschitz gradient, and s-RSC may not imply the convexity of f on $\mathbb{R}^n$. We are ready to report our main result below.

Theorem 2.1 *For (1) and $x^* \in S_+$, consider three conditions: (a) $\|x^*\|_0 = s$; (b) $\|x^*\|_0 < s$; (c) f is $2s$-RC. Then we have the following $(1) - (14)$ relationships shown in Figure 1 among the α-, B-, C-stationary points and global/local minimizers. For example, for the relationship (3), an α-stationary point will be a global minimizer of (1) under the conditions (b) and (c).*

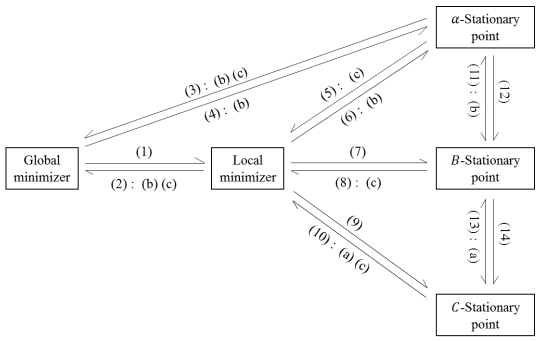

Figure 1: Relationships among α-, B-, C-stationary points and global/local minimizers.

Proof Clearly, (1) holds. By using Table 2, (11)-(14) can be verified directly. We actually only need to prove (3), (7) and (8). In fact, if (3), (7) and (8) hold, then (12) and (8) $\Rightarrow$ (5); (7) and (11) $\Rightarrow$ (6); (6) and (3) $\Rightarrow$ (2); (1) and (6) $\Rightarrow$ (4); (7) and (14) $\Rightarrow$ (9); (13) and (8) $\Rightarrow$ (10).

For (3), if f is $2s$-restricted convex, then for any $x \in S_+$ which implies $|\Gamma_{xx^*}| = |\mathrm{supp}(x) \cup \Gamma^*| \leq 2s$, we have

$$\begin{aligned} f(x) &\geq f(x^*) + \langle \nabla f(x^*),\ x - x^* \rangle \\ &= f(x^*) + \sum_{i \in \Gamma^*} \nabla_i f(x^*)(x_i - x_i^*) + \sum_{i \notin \Gamma^*} \nabla_i f(x^*) x_i \\ &\geq f(x^*), \end{aligned}$$

where the last inequality is from expression of α-stationary point in Table 2 for the case $\|x^*\|_0 < s$. This proves that x^* is a global minimizer of (1).

For (7), if $x^* \in S_+$ is a local minimizer of (1), then there is a constant $\delta > 0$ such that

$$f(x^*) \leq f(x),\ \forall\ x \in N(x^*, \delta) \cap S_+.$$

If $\|x^*\|_0 < s$, then for any $i \in \Gamma^*$, we have $x^* + te_i \in N(x^*, \delta) \cap S_+$ with sufficiently small $t > 0$ or $t < 0$ such that

$$f(x^*) \leq f(x^* + te_i) = f(x^*) + t\nabla_i f(x^*) + o(t),$$

thus $\nabla_i f(x^*) = 0$ for $i \in \Gamma^*$. For any $i \notin \Gamma^*$, the above inequality holds for sufficiently small $t > 0$, which yields $\nabla_i f(x^*) \geq 0$. If $\|x^*\|_0 = s$, the same argument leads to $\nabla_i f(x^*) = 0$ for any $i \in \Gamma^*$. Therefore, x^* is a B-stationary point of (1) by Table 2.

For (8), if x^* is B-stationary point and f is $2s$-restricted convex, then for any $x \in N(x^*, \delta) \cap S_+$ that implies $|\Gamma_{xx^*}| = |\mathrm{supp}(x) \cup \Gamma^*| = |\Gamma^*| = s \leq 2s$, we have

$$f(x) \geq f(x^*) + \langle \nabla f(x^*), x - x^* \rangle \geq f(x^*),$$

where the last inequality is from Pro. 2.2(i), which means x^* is a local minimizer of (1). □

We would like to make some comments regarding the above theorem.

(R1) It follows from the relations (7) and (9) in Fig. 1 that a local minimizer must be a B- or C-stationary point. This means that the B- or C-stationarity forms a necessary condition for the sparse optimization (1). For the converse to be true, one must need some condition such as (c). In particular, the condition (a): $\|x^*\|_0 = s$ (i.e., the sparsity constraint is tight) is also part of the sufficient condition for a C-stationary point to be a local minimizer. Without this condition, a C-stationary point may fail to be a local minimizer even f is assumed to be convex, as shown by the following example:

$$\begin{array}{ll} \min & f(x) = (x_1+1)^2 + (x_2-1)^2 + (x_3-1)^2 \\ \text{s.t.} & \|x\|_0 \le 2, \quad x \ge 0. \end{array}$$

The objective function f is convex on $\mathbb{R}^3_+$ and its gradient $\nabla f(x) = 2(x_1+1, x_2-1, x_3-1)^\top$. It is obvious that $x^* = (0,0,1)^\top$ with $\nabla f(x^*) = (2,-2,0)^\top$ is C-stationary point, but not a local minimizer because $f((0,\epsilon,1)^\top) < f(x^*)$, $0 < \epsilon \le 1$.

(R2) If one further assumes that f has Lipschitz gradient (not just being continuously differentiable), α-stationarity becomes the L-stationaritiy introduced in [5]. Moreover, α-stationarity is also a necessary condition of x^* being a local minimizer [5, Thm. 2.2]. Without the Lipschitz property of the gradient function, relation (6) shows that it is also a necessary condition provided that the sparse constraint is not tight.

2.2 Global properties

In this subsection, we collect several useful global properties of B- and C-stationary points under the restricted (strong) convexity. Our first result is a simple consequence of the results reported above. Recalling the variational inequality characterization of the B-stationary point in Prop. 2.2(ii), the relationships (11) and (3) in Fig. 1 establish the following important characterization of a global minimizer of (1).

Corollary 2.1 *Suppose f is $2s$-RC and $x^* \in S_+$ with $\|x^*\|_0 < s$. The following are equivalent.*

(i) x^* is a global minimizer of (1).

(ii) x^* is an α-stationary point.

(iii) x^* is a B-stationary point.

(iv) It holds that $\langle \nabla f(x^*),\ x - x^* \rangle \ge 0,\ \forall\ x \in S_+$.

The next result shows that a B-stationary point or a C-stationary point can be a global minimizer when restricted to certain subspace.

Theorem 2.2 *Suppose f is s-RC. Let $x^* \in S_+$. Then the following hold.*

(i) If x^* is a B-stationary point, then it is a global minimizer on the subspace $\mathbb{R}^n_\Upsilon$ for any $\Upsilon \subseteq \{1,\cdots,n\}$ that satisfies $\Gamma^* \subseteq \Upsilon$ and $|\Upsilon| = s$.

(ii) If x^* is a C-stationary point, then it is a global minimizer on the subspace $\mathbb{R}^n_{\Gamma^*}$.

(iii) If f is s-RSC, then the local minimizer of problem (1) on any s-dimensional subspace is unique. Furthermore, the number of the local minimizers is finite.

Proof (i) For any $x \in \mathbb{R}^n_\Upsilon \cap \mathbb{R}^n_+$, if $\|x^*\|_0 = s = |\Gamma^*|$ (and hence $\Upsilon = \Gamma^*$), we have $(x - x^*)_i = 0, \forall i \notin \Gamma^*$; if $\|x^*\|_0 < s$, $(x - x^*)_i = x_i \ge 0, \forall i \notin \Gamma^*$, which together with the fact of f being s-restricted convex and the expression of B-stationary point in Table 2 yields that

$$\begin{array}{rcl} f(x) & \ge & f(x^*) + \langle \nabla f(x^*),\ x - x^* \rangle \\ & = & f(x^*) + \displaystyle\sum_{i \in \Gamma^*} \nabla_i f(x^*)(x_i - x_i^*) + \sum_{i \notin \Gamma^*} \nabla_i f(x^*)(x_i - x_i^*) \\ & \ge & f(x^*). \end{array}$$

Thus the conclusion is derived. The proof of (ii) is similar to (i), and thus its proof is omitted.

(iii) We note that under the assumption of s-restricted strong convexity of f, the inequality in the proof of (i) becomes strict. Therefore, there exists only one local minimizer on any s-dimensional subspace. We also note that from Table 2 (Relation (9)), any local minimizer of (1) is also a C-stationary point. However, according to (ii), any C-stationary point must be a unique minimizer on an s-dimensional subspace. Since the number of the subspaces whose dimension is no larger than s is finite, we conclude that the number of the local minimizers of (1) is finite. □

The following example shows that there may exist multiple minimizers of (1) under the s-RSC. That is, one cannot establish the uniqueness of the global minimizer in Thm. 2.2(iii), unless stronger assumptions are in place.

$$\min \ \|x-\mathbf{1}\|, \ \ \text{s.t.} \ \ \|x\|_0 \le 1, \ x \ge 0,$$

where $\mathbf{1} = (1, 1, \cdots, 1)^\top$. The objective function is strongly convex on $\mathbb{R}^n$, and every e_i, $i = 1, \ldots, n$, is a global minimizer.

3 A Convergent IHT and Its Theoretical Analysis

In this section, we will present an improved iterative hard thresholding (IIHT) algorithm for (1) and then analyze its convergence properties utilizing the results reported above.

3.1 IIHT Algorithm

As reviewed in the Introduction, in order for the generated iterates by the IHT algorithm to converge to a point satisfying certain optimality conditions (stationarities), a proper selection of stepsize seems necessary at each iteration. For example, for classical linear compressed sensing, Blumensath and Davies introduced an adaptive stepsize rule based on the RIP to ensure a sufficient decrease in the objective per iteration. Recently, for the nonlinear sparse optimization problem, Lu [27] introduced a nonmonotone line search to a projection algorithm to ensure its convergence. In this paper, we choose to use the classical Armijo stepsize rule in IHT, leading to what we call an Improved IHT (IIHT) algorithm. Another new element that we introduce in IIHT is a new stopping criterion that is motived by C-stationarity. The remaining part of IIHT just follows the original IHT and hence the framework of IIHT is very simple and is described as follows.

Table 3: The framework of IIHT algorithm for (1).

Step 0 Initialize $x^0 = 0$, $0 < \alpha_0 < \frac{1}{L_{2s}}$, $\sigma > 0$, $0 < \beta < 1, \epsilon > 0$. Set $k \Leftarrow 0$.

Step 1 Compute $x^{k+1} = \mathrm{P}_{S_+}\big(x^k - \alpha_k \nabla f(x^k)\big)$, where $\alpha_k = \alpha_0 \beta^{q_k}$ and q_k is the smallest nonnegative integer q such that

$$f(x^k(\alpha_k^0 \beta^q)) \le f(x^k) - \tfrac{\sigma}{2}\|x^k(\alpha_k^0\beta^q) - x^k\|^2,$$

and $x^k(\alpha) := \mathrm{P}_{S_+}(x^k - \alpha \nabla f(x^k))$.

Step 2 If $\|\nabla_{\Gamma^k} f(x^k)\| \le \epsilon$, then **Stop**; Otherwise, let $k \Leftarrow k+1$ and go to **Step 1**.

The stopping criterion used will be justified by Thm. 3.2(iii). We emphasize that the major computaion $\mathrm{P}_{S_+}(\cdot)$ is very easy to obtain via (2). The following result shows that the Armijo stepsize is well defined under some condition.

Lemma 3.1 *Let f be $2s$-RSS and $x^k \in S_+$ be given. Then it holds*

$$f(x^k(\alpha)) \le f(x^k) - \frac{\sigma}{2}\|x^k(\alpha) - x^k\|^2 \qquad \textit{for} \ \ 0 < \alpha \le \frac{1}{L_{2s}+\sigma}. \tag{9}$$

Therefore α_k in the algorithm is well defined.

Proof According to the computation of $x^k(\alpha)$ in Step 1, we have

$$x^k(\alpha) \in \mathrm{argmin}\big\{\|x - x^k + \alpha\nabla f(x^k)\|^2 : \ \ x \in S_+\big\},$$

which implies that $\|x^k(\alpha) - x^k + \alpha\nabla f(x^k)\|^2 \le \|\alpha \nabla f(x^k)\|^2$ by $x^k \in S_+$. This leads to

$$\|x^k(\alpha) - x^k\|^2 \le -2\alpha\langle \nabla f(x^k), \ x^k(\alpha) - x^k\rangle. \tag{10}$$

It follows from the property of $2s$-RSS of f and (10) that

$$\begin{aligned} f(x^k(\alpha)) &\le f(x^k) + \langle \nabla f(x^k), \ x^k(\alpha) - x^k\rangle + \frac{L_{2s}}{2}\|(x^k(\alpha) - x^k)\|^2 \\ &\le f(x^k) - \frac{1}{2\alpha}\|x^k(\alpha) - x^k\|^2 + \frac{L_{2s}}{2}\|(x^k(\alpha) - x^k)\|^2 \\ &= f(x^k) - \frac{1}{2}(1/\alpha - L_{2s})\|x^k(\alpha) - x^k\|^2. \end{aligned}$$

By restricting $\alpha \in (0, \frac{1}{L_{2s}+\sigma}]$, we obtain the desired result. The proof is completed. □

3.2 Convergence Analysis

Combining the restricted strong convexity and smoothness of f, the convergence of IIHT can be established in this subsection. We first present a technical result.

Lemma 3.2 *Suppose that the function f is s-RC and s-RSS with parameter L_s. Then for any $x, y \in \mathbb{R}^n$ satisfying $|\Gamma_{xy}| \leq s$, we have*

$$\|(\nabla f(y) - \nabla f(x))_{\Gamma_{xy}}\| \leq L_s \|y - x\|.$$

Proof Let us fix $x \in S$ and define the following function of variable y at point x:

$$\phi_x(y) := f(y) - \langle \nabla f(x),\ y - x \rangle.$$

Due to the s-RC of $f(\cdot)$, the point x is a minimizer of $\phi_x(y)$ over all y satisfying $|\Gamma_{xy}| \leq s$. This is because

$$\phi_x(y) - \phi_x(x) = f(y) - \langle \nabla f(x),\ y - x \rangle - f(x) \geq 0 \qquad \forall\ y \text{ such that } |\Gamma_{xy}| \leq s. \tag{11}$$

We note that function $\phi_x(\cdot)$ has the same properties of s-restricted strong smoothness as $f(\cdot)$. Define $d \in \mathbb{R}^n$ by

$$d_i := \begin{cases} \frac{1}{L_s}(\nabla \phi_x(y))_i, & \text{if } i \in \Gamma_{xy} \\ 0, & \text{otherwise.} \end{cases}$$

We have

$$\big\|y - d\big\|_0 \leq |\Gamma_{xy}| \leq s \quad \text{and} \quad \langle \nabla \phi_x(y),\ d \rangle = \frac{1}{L_s} \|(\nabla \phi_x(y))_{\Gamma_{xy}}\|^2,$$

which, together with (11) and the s-RSS of $\phi_x(\cdot)$, imply

$$\begin{aligned} \phi_x(x) &\leq \phi_x\big(y - d\big) \\ &\leq \phi_x(y) + \big\langle \nabla \phi_x(y),\ -d \big\rangle + \frac{L_s}{2} \big\| \frac{1}{L_s} (\nabla \phi_x(y))_{\Gamma_{xy}} \big\|^2 \\ &= \phi_x(y) - \frac{1}{2L_s} \big\| (\nabla \phi_x(y))_{\Gamma_{xy}} \big\|^2. \end{aligned} \tag{12}$$

Rewrite (12) as

$$f(x) \leq f(y) - \langle \nabla f(x), y - x \rangle - \frac{1}{2L_s} \big\| (\nabla f(y) - \nabla f(x))_{\Gamma_{xy}} \big\|^2. \tag{13}$$

By interchanging x and y in (13) and adding the resulting inequality to (13), we get

$$\big\| (\nabla f(y) - \nabla f(x))_{\Gamma_{xy}} \big\|^2 \leq L_s \langle \nabla f(y) - \nabla f(x), y - x \rangle. \tag{14}$$

The desired results then follows from applying the Cauchy-Schwarz inequality to (14). □

We report our first convergence result below.

Theorem 3.1 *Let the sequence $\{x^k\}$ be generated by IIHT. Suppose f is $2s$-RSS. Then the following hold.*

(i) $\lim_{k\to\infty} \|x^{k+1} - x^k\| = 0$ *and* $\inf_{k\geq 0} \{\alpha_k\} > 0$*;*

(ii) Any accumulation point of $\{x^k\}$ is an α-stationary point of (1).

Moreover, if f is $2s$-RC, then the following hold.

(iii) The sequence of projected gradients converges to zero, i.e.,

$$\lim_{k\to\infty} \|\nabla_{\Gamma^k} f(x^k)\| = 0.$$

(iv) Any accumulation point of $\{x^k\}$ is a local minimizer of (1).

Proof (i) As required in IIHT, we have $f(x^k) - f(x^{k+1}) \geq \frac{\sigma}{2}\|x^{k+1} - x^k\|^2$. Then

$$\sum_{k=0}^{\infty} \|x^{k+1} - x^k\|^2 \leq \frac{2}{\sigma} \sum_{k=0}^{\infty} \big(f(x^k) - f(x^{k+1})\big) < \frac{2}{\sigma} \left(f(x^0) - \lim_{k\to+\infty} f(x^k) \right) < +\infty,$$

where the last inequality is due to f being bounded from below. Hence $\lim_{k\to\infty} \|x^{k+1} - x^k\| = 0$.

Armijo-type stepsize rule and Lemma 3.1 imply that $\{\alpha_k\}$ is bounded from below by a positive constant. In fact,

$$\inf_{k\geq 0} \{\alpha_k\} \geq \frac{\beta}{\sigma + L_{2s}} := \underline{\alpha} > 0. \tag{15}$$

(ii) Suppose that x^* is an accumulation point of the sequence $\{x^k\}$. There exists a subsequence $\{x^{k_j}\}$ converging to x^*. It follows from (i) that

$$\lim_{j\to\infty} x^{k_j+1} = \lim_{j\to\infty} x^{k_j} = x^*. \tag{16}$$

Based on the update

$$x^{k_j+1} = \mathrm{P}_{S_+}\big(x^{k_j} - \alpha_{k_j}\nabla f(x^{k_j})\big) \tag{17}$$

in Step 1 of the IIHT algorithm, we consider two cases.

Case 1. For $i \in \Gamma^*$. There must exist a sufficiently large index n_1 and a positive constant c_0 such that

$$\min\{x_i^{k_j},\ x_i^{k_j+1}\} \geq c_0 \qquad \forall\ j \geq n_1.$$

This together with the projection formula of $\mathrm{P}_{S_+}(\cdot)$ in (2) and (17) implies

$$x_i^{k_j+1} = x_i^{k_j} - \alpha_{k_j}\nabla_i f(x^{k_j}).$$

Therefore, the positive lower bound in (15) and the limit in (16) yield

$$\nabla_i f(x^*) = 0, \qquad \forall\ i \in \Gamma^*.$$

Case 2. For $i \notin \Gamma^*$. Without loss of any generality, we may assume $\lim_{j\to\infty}\alpha_{k_j} = c_1 > 0$ on the subsequence $\{k_j\}$. We consider two subcases. Subcase 2.1: $\|x^*\|_0 = s$. Then we have

$$0 = \lim_{j\to\infty} x_i^{k_j+1} = \lim_{j\to\infty}\Big(\mathrm{P}_S\Big(\mathrm{P}_{\mathbb{R}^n_+}(x^{k_j} - \alpha_{k_j}\nabla f(x^{k_j}))\Big)\Big)_i$$

Due to the property of the projections $\mathrm{P}_S(\cdot)$ and $\mathrm{P}_{\mathbb{R}^n_+}(\cdot)$, we must have

$$\max\Big\{x_i^{k_j} - \alpha_{k_j}\nabla_i f(x^{k_j}),\ 0\Big\} \leq \Big(x^{k_j+1}\Big)_s^{\downarrow}.$$

Taking limits on both sides, we obtain

$$\nabla_i f(x^*) \geq -\frac{1}{c_1}(x^*)_s^{\downarrow}.$$

Subcase 2.2: $\|x^*\|_0 < s$. Suppose $\nabla_i f(x^*) < 0$. We then have

$$\lim_{j\to\infty}\Big(x_i^{k_j} - \alpha_{k_j}\nabla_i f(x^{k_j})\Big) = -c_1\nabla_i f(x^*) > 0,$$

leading to

$$\Big(\mathrm{P}_{\mathbb{R}^n_+}(x^{k_j} - \alpha_{k_j}\nabla f(x^{k_j}))\Big)_i \geq -\frac{1}{2}c_1\nabla_i f(x^*)$$

for all sufficiently large j. Since $\|x^*\|_0 < s$, we must have for j sufficiently large

$$x_i^{k_j+1} = \Big(\mathrm{P}_S\Big(\mathrm{P}_{\mathbb{R}^n_+}(x^{k_j} - \alpha_{k_j}\nabla f(x^{k_j}))\Big)\Big)_i = \Big(\mathrm{P}_{\mathbb{R}^n_+}(x^{k_j} - \alpha_{k_j}\nabla f(x^{k_j}))\Big)_i \geq -\frac{1}{2}c_1\nabla_i f(x^*) > 0.$$

This contradicts the assumption $i \notin \Gamma^*$ (which in turn implies $\lim_{j\to\infty} x_i^{k_j+1} = 0$). Therefore, we must have $\nabla_i f(x^*) \geq 0$ for Subcase 2.2.

Summarizing the above two cases, we obtained

$$\nabla_i f(x^*)\begin{cases} = 0, & \text{if } \ i \in \Gamma^*, \\ \in [-\frac{1}{c_1}(x^*)_s^{\downarrow},\ \infty), & \text{if } \ i \notin \Gamma^*, \end{cases} \tag{18}$$

which means that x^* is an α-stationary point of (1) by Table 2.

(iii) Notice that $\mathrm{T}^C_{S_+}(x^k) = \mathbb{R}^n_{\Gamma^k}$ is a subspace. The projection of the negative gradient $(-\nabla f(x^k))$ to this subspace has the following property due to [13, Lemma 3.1]:

$$\|\mathrm{P}_{\mathbb{R}^n_{\Gamma^k}}(-\nabla f(x^k))\| = \max\big\{\langle -\nabla f(x^k),\ v\rangle :\ v \in \mathbb{R}^n_{\Gamma^k},\ \ \|v\| \leq 1\big\} = \big\|\nabla_{\Gamma^k} f(x^k)\big\|.$$

Moreover, the maximum takes place at the boundary of $\|v\| = 1$. Therefore, for any given $\epsilon > 0$, there exists $v^k \in \mathbb{R}^n_{\Gamma^k}$ with $\|v^k\| = 1$ such that

$$\big\|\nabla_{\Gamma^k} f(x^k)\big\| \leq -\langle \nabla f(x^k),\ v^k\rangle + \epsilon. \tag{19}$$

It follows from $x^{k+1} = \mathrm{P}_{S_+}\big(x^k - \alpha_k \nabla f(x^k)\big)$ and the property in (3) that

$$x^{k+1}_{\Gamma^{k+1}} = \Big(x^k - \alpha_k \nabla f(x^k)\Big)_{\Gamma^{k+1}}.$$

In other words, the vector $(x^{k+1} - (x^k - \alpha_k \nabla f(x^k)))$ is orthogonal to $\mathbb{R}^n_{\Gamma^{k+1}}$. This yields that

$$\big\langle x^{k+1} - (x^k - \alpha_k \nabla f(x^k)),\ w^{k+1} - x^{k+1}\big\rangle = 0, \qquad \forall\ w^{k+1} \in \mathbb{R}^n_{\Gamma^{k+1}}.$$

Choose a particular w^{k+1} by $w^{k+1} := x^{k+1} + v^{k+1}$. The Cauchy-Schwartz inequality implies

$$-\langle \nabla f(x^k), v^{k+1}\rangle \le \frac{\|x^{k+1} - x^k\|}{\alpha_k}. \tag{20}$$

Since f is $2s$-RSS, we have from Lemma 3.2 and (20) that

$$\begin{aligned} -\langle \nabla f(x^{k+1}), v^{k+1}\rangle &= -\langle \nabla f(x^{k+1}) - \nabla f(x^k),\ v^{k+1}\rangle - \langle \nabla f(x^k),\ v^{k+1}\rangle \\ &\le L_{2s}\|x^{k+1} - x^k\| + \frac{\|x^{k+1} - x^k\|}{\alpha_k}. \end{aligned}$$

Taking limits on both sides and using the facts established in (i), we have

$$\limsup_{k\to\infty} -\langle \nabla f(x^{k+1}), v^{k+1}\rangle \le 0.$$

From (19) and the arbitrariness of ϵ, we proved $\lim\limits_{k\to\infty} \big\|\nabla_{\Gamma^k} f(x^k)\big\| = 0$.

(iv) The convergence to a local minimizer of (1) provided that f is $2s$-restricted convex follows directly from Theorem 2.1 (Relation (5)). □

The following result further characterizes when the whole sequence converges to a local minimizer and when the local minimizer becomes a global one.

Theorem 3.2 *Assume f is both $2s$-RSS and $2s$-RSC. Then the whole sequence $\{x^k\}$ converges to a local minimizer x^* of (1). Furthermore, depending on whether the sparse constraint is tight or not at x^*, we have the following detailed characterization of x^*.*

(i) If the sparse constraint is tight at x^* (i.e., $\|x^*\|_0 = s$), then

$$\Gamma^k \equiv \Gamma^* \qquad \textit{for all sufficiently large } k.$$

(ii) If the sparse constraint is not tight at x^* (i.e., $\|x^*\|_0 < s$), then x^* is a global minimizer of (1).

Proof From Thm. 2.2(iii), the number of the local minimizers of (1) is finite and from Theorem 3.1(iv), every accumulation point of $\{x^k\}$ is a local minimizer of (1). Hence, the number of accumulation points of sequence $\{x^k\}$ is finite and every accumulation point, which is also a local minimizer, is isolated. Since f is $2s$-RSC, the sequence $\{x^k\}$ is bounded. Theorem 3.1(i) has established that the whole sequence $\{x^k\}$ satisfies $\|x^{k+1} - x^k\| \to 0$. It follows from [30, Lemma 4.10] (which is restated as [23, Prop. 7], which is more relevant to our current setting) that the whole sequence must converge to a local minimizer. We now prove the remaining two claims.

(i) If $\|x^*\|_0 = s$, since $x^k \to x^*$, we have $\|x^k - x^*\| < \delta$ where $0 < \delta < \min\{x^*_i : i \in \Gamma^*\}$ for all sufficiently large k. Then following the same reasoning as proving (5), we have $\Gamma^k \equiv \Gamma^*$ for all sufficiently large k.

(ii) If $\|x^*\|_0 < s$, the conclusion can be derived immediately due to f being $2s$-RSC and Theorem 2.1 (Relation (2)). □

What we have proved in the above theorem is that when the sparse constraint $\|x\|_0 \le s$ is tight at x^*, we can only claim that the whole sequence converges to a local minimizer, whereas when it is not tight, the whole sequence converges to the global minimizer. We note that a similar result has also been recently proved by Lu [27] though under different assumptions. If the sparse constraint is tight, then the sequence generated in [27] only converges to a local minimizer under the assumptions that f has Lipschitz gradient and is convex. If the sparse constraint is not tight, then the sequence generated in [27] converges to the global minimizer provided that f further satisfies Assumption (3) in [27]. Here we assumed $2s$-RSS and $2s$-RSC. Therefore, our basic assumptions as well as the proof techniques are fundamentally different from those in [27]. Moreover, our result in (i) states that the active index set can be correctly identified in the case of tight constraint. This is the crucial property that allows us to establish the Q-linear convergence rate (Thm. 3.4) below.

3.3 Sub-linear and Q-linear convergence rate

In this subsection, we will show the linear convergence rate both in terms of functional value sequence $\{f(x^k)\}$ and the sequence itself $\{x^k\}$. From the view of point in Theorem 3.2, we need the assumptions of both $2s$-RSC and $2s$-RSS. Consequently, the whole sequence $\{x^k\}$ converges to a local minimizer x^*.

First we make an easy observation. For x^k, denote

$$F_k(x) := \|x - x^k + \alpha_k \nabla f(x^k)\|^2.$$

Then it is obvious that

$$x^{k+1} = \mathrm{P}_{S_+}(x^k - \alpha_k \nabla f(x^k)) = \arg\min_{x \in S_+} \; F_k(x).$$

We claim that it holds

$$\langle \nabla F_k(x^{k+1}),\ x^* - x^{k+1} \rangle \geq 0 \qquad \text{for any } k \text{ such that } \ \Gamma^* \subseteq \Gamma^{k+1}. \tag{21}$$

We prove above inequality by considering two cases. Case 1: $\|x^{k+1}\|_0 < s$ and Case 2: $\|x^{k+1}\|_0 = s$. For Case 1, apply Cor. 2.1(iv) to F_k (instead of f therein) to get (21) because $F_k(\cdot)$ is $2s$-RC due to $F_k(\cdot)$ being strongly convex. For Case 2, x^{k+1} is the global minimizer of $F_k(x)$ and thus a B-stationary point, which implies $\nabla_{\Gamma^{k+1}} F_k(x^{k+1}) = 0$. Then by $\Gamma^* \subseteq \Gamma^{k+1}$ when k is sufficiently large, we must have

$$\langle \nabla F_k(x^{k+1}),\ x^* - x^{k+1} \rangle = \sum_{i \in \Gamma^{k+1}} \nabla_i F_k(x^{k+1})(x_i^* - x_i^{k+1}) = 0.$$

Hence (21) holds and its leads to the following linear rate convergence.

Theorem 3.3 *Assume f is $2s$-RSS and $2s$-RSC. Let $\{x^k\}$ be generated by IIHT and be convergent to a local minimizer x^* of (1) (the convergence is guaranteed by Thm. 3.2). Then for any $k > k_0$, the following inequality holds:*

$$f(x^k) - f(x^*) \leq \frac{1}{(k-k_0)\underline{\alpha} l_{2s}} \big(f(x^{k_0}) - f(x^*)\big), \tag{22}$$

where $\underline{\alpha} := \frac{\beta}{L_{2s}+\sigma}$ and k_0 is the smallest positive integer such that $\Gamma^* \subseteq \Gamma^k$ for any $k > k_0$.

Proof Since $x^k \to x^*$, there exists k_0 such that $\Gamma^* \subseteq \Gamma^k,\ \forall\ k > k_0$. Therefore, (21) holds for any $k > k_0$. Since f is convex on any $2s$-dimensional subspace and $2s$-RSS with $0 < \alpha_k < \frac{1}{L_{2s}}$, it follows that

$$f(x^*) \quad \geq \quad f(x^k) + \langle \nabla f(x^k),\ x^* - x^k \rangle, \tag{23}$$

$$f(x^{k+1}) \quad \leq \quad f(x^k) + \langle \nabla f(x^k),\ x^{k+1} - x^k \rangle + \frac{1}{2\alpha_k}\|x^{k+1} - x^k\|^2. \tag{24}$$

By the function $F_k(x)$ being strongly convex with modulus 2 and (21), we have

$$F_k(x^*) = F_k(x^{k+1}) + \langle \nabla F_k(x^{k+1}),\ x^* - x^{k+1} \rangle + \|x^* - x^{k+1}\|^2 \geq F_k(x^{k+1}) + \|x^* - x^{k+1}\|^2.$$

Substituting the definition of $F_k(x)$ into the above inequality and simplifying lead to

$$\begin{aligned} & \langle \nabla f(x^k),\ x^{k+1} - x^k \rangle + \frac{1}{2\alpha_k}\|x^{k+1} - x^k\|^2 \\ \leq \quad & \langle \nabla f(x^k),\ x^* - x^k \rangle + \frac{1}{2\alpha_k}(\|x^* - x^k\|^2 - \|x^* - x^{k+1}\|^2). \end{aligned} \tag{25}$$

Combining (24), (25) and (23), it holds that

$$\begin{aligned} & f(x^{k+1}) + \frac{1}{2\alpha_k}\|x^* - x^{k+1}\|^2 \\ \leq \quad & f(x^k) + \langle \nabla f(x^k),\ x^{k+1} - x^k \rangle + \frac{1}{2\alpha_k}\|x^{k+1} - x^k\|^2 + \frac{1}{2\alpha_k}\|x^* - x^{k+1}\|^2 \quad \text{(by (24))} \\ \leq \quad & f(x^k) + \langle \nabla f(x^k),\ x^* - x^k \rangle + \frac{1}{2\alpha_k}\|x^* - x^k\|^2 \qquad \text{(by (25))} \\ \leq \quad & f(x^*) + \frac{1}{2\alpha_k}\|x^* - x^k\|^2, \qquad \text{(by (23))} \end{aligned}$$

which amounts to

$$\begin{aligned} f(x^{k+1}) - f(x^*) \quad & \leq \quad \frac{1}{2\alpha_k}\left(\|x^k - x^*\|^2 - \|x^{k+1} - x^*\|^2\right) \\ & \leq \quad \frac{1}{2\underline{\alpha}}\left(\|x^k - x^*\|^2 - \|x^{k+1} - x^*\|^2\right), \end{aligned}$$

where we have used the fact $\alpha_k \geq \frac{\beta}{\sigma+L_{2s}} = \underline{\alpha}$ proved in Theorem 3.1(i) (15). For any positive integer j, using this inequality and the monotonically decreasing property of $\{f(x^k)\}$, we have

$$j(f(x^{k+j}) - f(x^*)) \leq \sum_{i=k}^{k+j-1} \big(f(x^{i+1}) - f(x^*)\big) \leq \frac{1}{2\underline{\alpha}}(\|x^k - x^*\|^2 - \|x^{k+j} - x^*\|^2).$$

We thus have

$$f(x^{k+j}) - f(x^*) \leq \frac{1}{2j\underline{\alpha}}\|x^k - x^*\|^2. \tag{26}$$

In addition, since x^* is a local minimizer and thus a B-stationary point. By Prop. 2.2(i), it holds $\langle \nabla f(x^*), x^k - x^* \rangle \geq 0$ because x^k is in a neighborhood of x^*. This and f being $2s$-RSC yield

$$f(x^k) - f(x^*) \geq \langle \nabla f(x^*),\ x^k - x^* \rangle + \frac{l_{2s}}{2}\|x^k - x^*\|^2 \geq \frac{l_{2s}}{2}\|x^k - x^*\|^2,$$

which together with (26) contributes to

$$f(x^{k+j}) - f(x^*) \leq \frac{1}{j\underline{\alpha}l_{2s}}\big(f(x^k) - f(x^*)\big).$$

Therefore, for any $k > k_0$, it holds that

$$f(x^k) - f(x^*) \leq \frac{1}{(k-k_0)\underline{\alpha}l_{2s}}\big(f(x^{k_0}) - f(x^*)\big),$$

which completes the proof. □

We now show the Q-linear convergence rate of the iterative points sequence of IIHT under assumption $\|x^*\|_0 = s$.

Theorem 3.4 *Assume f is $2s$-RSS and $2s$-RSC. Let x^* be the limit of the sequence $\{x^k\}$ generated by IIHT that satisfies $\|x^*\|_0 = s$. Then for any sufficiently large k, it holds,*

$$\|x^{k+1} - x^*\|^2 \leq \rho\|x^k - x^*\|^2, \quad 0 < \rho < 1, \tag{27}$$

where $\rho := 1 - 2l_{2s}^2\underline{\alpha}/L_{2s} + l_{2s}^2\underline{\alpha}^2$ with $\underline{\alpha}$ being defined in Theorem 3.3.

Proof As already used, the convergence of $\{x^k\}$ to x^* is guaranteed by Thm. 3.2. Since f is $2s$-restricted strongly convex with parameters l_{2s} in (8), we can easily obtain that

$$\|(\nabla f(x) - \nabla f(y))_{\Gamma_{xy}}\| \geq l_{2s}\|x - y\| \qquad \forall\ |\Gamma_{xy}| \leq 2s.$$

This together with Lemma 3.2 and Thm. 3.2(i) (proving $\Gamma^k \equiv \Gamma^*$ for all sufficiently large k) yields that for any sufficiently large k,

$$\begin{aligned}
\|x^{k+1} - x^*\|^2 &= \|x^k_{\Gamma^*} - \alpha_k \nabla_{\Gamma^*} f(x^k) - x^*_{\Gamma^*} + \alpha_k \nabla_{\Gamma^*} f(x^*)\|^2 \\
&= \|x^k - x^*\|^2 - 2\alpha_k\langle x^k - x^*, \nabla f(x^k) - \nabla f(x^*)\rangle + \alpha_k^2\|(\nabla f(x^k) - \nabla f(x^*))_{\Gamma^*}\|^2 \\
&\leq \|x^k - x^*\|^2 - (2\alpha_k/L_{2s} - \alpha_k^2)\|(\nabla f(x^k) - \nabla f(x^*))_{\Gamma^*}\|^2 \\
&\leq \big(1 - 2l_{2s}^2\alpha_k/L_{2s} + l_{2s}^2\alpha_k^2\big)\|x^k - x^*\|^2,
\end{aligned}$$

where $\nabla_{\Gamma^*} f(x^*) = 0$ in the first equality holds due to Thm. 3.1(iii), namely, $\nabla_{\Gamma^*} f(x^*) = \lim_{k\to\infty} \nabla_{\Gamma^k} f(x^k) = 0$. It follows from $\underline{\alpha} \leq \alpha_k < 1/L_{2s}$ that

$$\begin{aligned}
1 - 2l_{2s}^2\alpha_k/L_{2s} + l_{2s}^2\alpha_k^2 &= 1 + l_{2s}^2\big(\alpha_k - 1/L_{2s}\big)^2 - l_{2s}^2/L_{2s}^2 \\
&\leq 1 + l_{2s}^2\big[\big(\underline{\alpha} - 1/L_{2s}\big)^2 - 1/L_{2s}^2\big] \\
&= 1 - 2l_{2s}^2\underline{\alpha}/L_{2s} + l_{2s}^2\underline{\alpha}^2 = \rho.
\end{aligned}$$

Moreover, $\rho = l_{2s}^2\big(\underline{\alpha} - 1/L_{2s}\big)^2 + 1 - l_{2s}^2/L_{2s}^2 > 0$ and $\rho = 1 - l_{2s}^2\underline{\alpha}\Big(\frac{2}{L_{2s}} - \frac{\beta}{L_{2s}+\sigma}\Big) < 1$. The proof is completed. □

We note that convergence result of the type (27) is known to be Q linear rate in optimization. We are only able to establish this result for the special case when the sparse constraint is tight. The key reason is that we were able to correctly identify the active index set for this case.

4 Numerical Experiments

In this section, we report our numerical experiments of `IIHT` on three classes of problems: Linear compressed sensing under nonnegativity constraints, Sparse logistic regression and Phase retrieval. Our stopping criterion is set as

$$\text{number of iterations} \leq \texttt{Maxiter} \qquad \text{or} \qquad \|(\nabla f(x^k))_{\Gamma^k}\| \leq \epsilon,$$

where we stop our algorithm whenever the number of iterations exceeds `Maxiter` or the projected gradient becomes less than ϵ. We will set a different level for ϵ and `Maxiter` for each class of test problems. The CPU time reported here does not include the time for data initialization. All those simulations are carried out on a CPU 3.2GHz, RAM 4.0GB desktop.

4.1 Compressed Sensing

We first test the classical linear CS problem under the nonnegativity constraint with $f(x) = f_A(x) := \|Ax - b\|^2$, where $A \in \mathbb{R}^{m\times n}$ is a linear measurement matrix satisfying

$$b = Ax + \xi.$$

We will test two scenarios. One is the exact recovery where $\xi \equiv 0$ and the other is the so-called stable recovery where ξ follows the normal distribution. More specifically, two types of sensing matrices of A will be generated, namely, random Gaussian matrix, and random partial Discrete Cosine Transform (pDCT) matrix:

$$\begin{array}{lll} \text{Gaussian:} & A_{\cdot j} \overset{i.i.d.}{\sim} \mathcal{N}(0, I/m), & j = 1, 2, \cdots, n, \\ \text{pDCT:} & A_{ij} = m^{-1/2}\cos(2\pi(j-1)\psi_i), & i = 1, 2, \cdots, m, \quad j = 1, 2, \cdots, n, \end{array}$$

where $A_{\cdot j}$ denotes the jth column of A, ψ_i, $i = 1, \ldots, m$ are uniformly and independently sampled from $[0, 1]$. After generating them, we orthogonalize them to satisfy $AA^\top = I$.

We generate the 'true' original signal x_{orig} with nonnegative elements as follow: first produce an index set T with s indices randomly selected from $\{1, \cdots, n\}$; then for each element of x_{orig} with index in T, uniformly choose them from $[0, 10]$. The data are generated as follows (in Matlab format):

$$\begin{aligned} & x_{\text{orig}} = \texttt{zeros}(n, 1);\ T = \texttt{randperm}(n); \\ & x_{\text{orig}}(T(1:s)) = 10 * \texttt{rand}(s, 1); \\ & b = A * x_{\text{orig}} + \sigma_0 * \texttt{randn}(m, 1). \end{aligned} \tag{28}$$

Clearly, the case $\sigma_0 = 0$ is the exact recovery. For stable recovery, we take $\sigma_0 = 0.01$.

(a) Parameter setting. In our implementation, we set `Maxiter` $= 1000, \epsilon = 10^{-5}$, $\beta = 0.8$ and $\sigma = 10^{-5}$ for simplicity. Instead of fixing α_0 for each step in `IIHT`, we update it according to [11] to accelerate the computational speed, namely,

$$\alpha_0^k = \frac{\|A_{\Gamma^k}^\top (b - Ax^k)\|^2}{\|A_{\Gamma^k} A_{\Gamma^k}^\top (b - Ax^k)\|^2}.$$

We run 40 trials for Gaussian and pDCT matrices with $n = 5000, m = n/4$ and $s = 0.01n$ or $s = 0.05n$ for exact and stable recovery to see the decreasing of objective function at each iteration. Results recorded in Figure 2 show that only 5 (15) iterations are needed to get the desirable solutions when $s = 0.01n$ ($s = 0.05n$) for both exact and stable recovery, which shows that the gain in decreasing the objective function per iteration is sufficient.

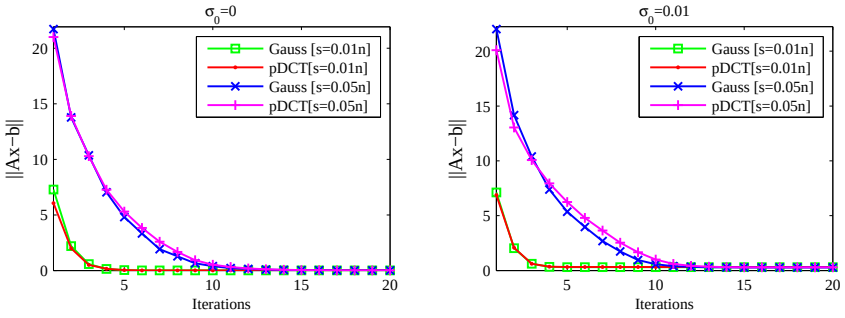

Figure 2: Objective function value at each iteration.

(b) Comparison of different methods. The reason for us to consider linear mapping is that we can compare our algorithm with other state-of-the-art greedy methods which are used to address liner compressed sensing. For example, Normalized Iterative Hard Thresholding (`NIHT`) proposed by Blumensath in [11], Compressive Sampling Matching Pursuit (`CoSaMP`) established by Thomas et al. in [31], and Subspace Pursuit(`SP`) in [17] [1].

We begin with running 100 independent trials for each type of matrix under $m = 64, n = 256$ and recording the corresponding success rate at sparsity levels from 5 to 30. The success rate is defined as the percentage of successful recovery of 100 trials. If the relative error is smaller than 10^{-2}, i.e.,

$$\texttt{Relative Error} := \frac{\|x - x_{\text{orig}}\|}{\|x\|} < 10^{-2},$$

the recovery is regarded as a successful one. Here x denotes computed solutions by four methods. Corresponding results are seen in Figure 3. Obviously, for these two types of matrices, `IIHT` basically runs the best results, followed by `SP` which outperforms `NIHT` and `CoSaMP`.

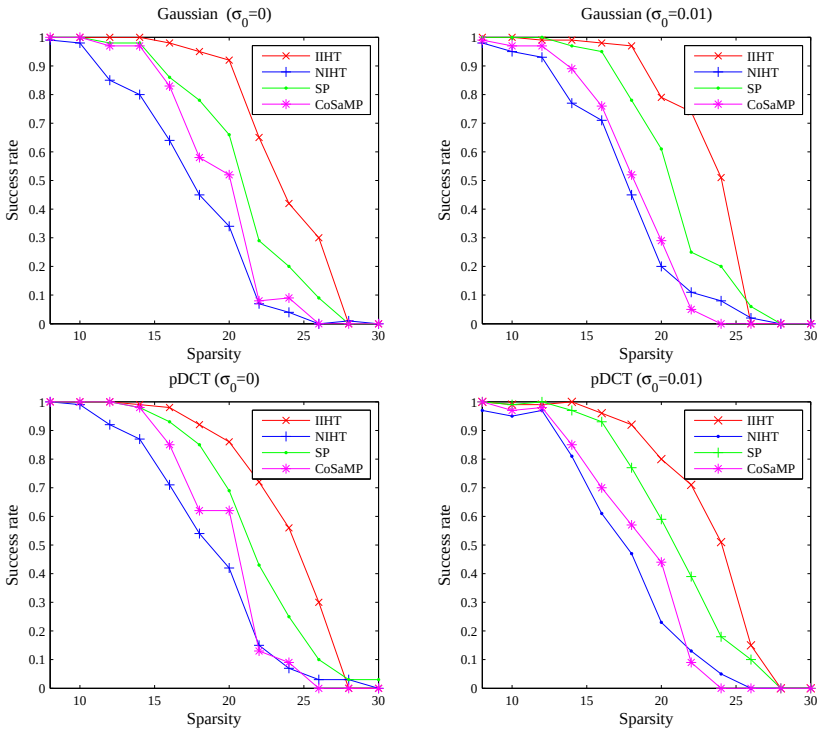

Figure 3: Success rates with two types of matrices with $m = 64, n = 256$.

To see the accuracy of the solutions and the speed of these four methods, we now run 40 trials for each kind of matrices with higher dimensions n increasing from 1000 to 9000 and keeping $m = n/4, s = 0.01n, 0.05n$. We also fix $\sigma_0 = 0.01$. Specific results produced by these four methods are recorded in Tables 4 and 5. The most obvious property of the data in the table is that the relative error of those four methods almost are identical. However, in terms of computational time, `SP` performs the best when $s = 0.01n$, followed by `IIHT`, `CoSaMP` and `NIHT`; When $s = 0.05n$, our proposed `IIHT` behaves better than `SP` and `NIHT`, and `CoSaMP` comes the last.

Table 4: Average results for Gaussian matrix.

1. CoSaMP and SP are available at: *http://media.aau.dk/null_space_pursuits/2011/07/a-few-corrections-to-cosamp-and-sp-matlab.html.*

s	n	Relative Error				Time			
		IIHT	NIHT	SP	CoSaMP	IIHT	NIHT	SP	CoSaMP
	1000	0.0040	0.0046	0.0046	0.0051	0.0049	0.0070	0.0052	0.0059
	3000	0.0035	0.0037	0.0037	0.0041	0.0124	0.0914	0.0081	0.0125
$0.01n$	5000	0.0036	0.0035	0.0036	0.0041	0.0254	0.2817	0.0116	0.0282
	7000	0.0042	0.0041	0.0043	0.0051	0.0541	1.6674	0.0292	0.0862
	9000	0.0038	0.0039	0.0041	0.0047	0.0905	2.8940	0.0719	0.1070
	1000	0.0043	0.0046	0.0045	0.0067	0.0113	0.0787	0.0145	0.0428
	3000	0.0038	0.0038	0.0039	0.0067	0.0651	2.8045	0.1194	37.491
$0.05n$	5000	0.0044	0.0042	0.0044	0.0065	0.2946	11.075	0.5455	150.69
	7000	0.0038	0.0038	0.0038	0.0061	0.4264	23.916	1.2049	559.55
	9000	0.0040	0.0041	0.0041	0.0065	0.8517	58.758	1.9868	1492.2

Table 5: Average results for pDCT matrix.

s	n	Relative Error				Time			
		IIHT	NIHT	SP	CoSaMP	IIHT	NIHT	SP	CoSaMP
	1000	0.0038	0.0039	0.0039	0.0039	0.0047	0.0060	0.0051	0.0049
	3000	0.0034	0.0034	0.0034	0.0042	0.0123	0.0877	0.0083	0.0127
$0.01n$	5000	0.0035	0.0035	0.0035	0.0044	0.0321	0.3904	0.0143	0.0350
	7000	0.0039	0.0039	0.0037	0.0044	0.0530	1.4925	0.0318	0.0728
	9000	0.0037	0.0037	0.0037	0.0045	0.0903	3.2969	0.0532	0.2942
	1000	0.0041	0.0040	0.0040	0.0068	0.0104	0.0754	0.0167	3.0316
	3000	0.0038	0.0038	0.0038	0.0061	0.0654	2.3984	0.1477	32.955
$0.051n$	5000	0.0041	0.0040	0.0041	0.0069	0.2266	11.858	0.3876	141.56
	7000	0.0042	0.0040	0.0042	0.0068	0.4541	17.088	1.5226	562.10
	9000	0.0040	0.0040	0.0041	0.0063	0.9101	56.779	8.73688	1485.2

4.2 Sparse Logistic Regression Problem

The logistic regression model plays an important role in two-class classification method that has been used widely in many applications ranging from data mining, machine learning, computer vision, to bioinformatics. Specifically, given data $z \in \mathbb{R}^n$ and weights (v, w), it assumes the following probability model

$$\mathbb{P}(b = \pm 1 | v, w) = \frac{1}{1 + \exp(-b(v + w^\top z))},$$

where b is the class label. If $z^i \in \mathbb{R}^n, i = 1, \cdots, m$ are m given samples with n features and $b_i \in \{1, -1\}, i = 1, \cdots, m$ are given m binary outcomes or labels, one estimates (v, w) by minimizing the negative log-likelihood:

$$\min_{v,w} \quad L(w, v) := \sum_{i=1}^{m} \log \left(1 + \exp(-b_i(v + w^\top z^i))\right)$$

Recently, sparse logistic regression is attractive in many applications involving high-dimensional data, seen [24, 28] and references therein. The corresponding optimization model is

$$\min_{v,w} \quad L(v, w), \quad \text{s.t. } \|w\|_0 \leq s. \tag{29}$$

Letting $x = (v; w) \in \mathbb{R}^{n+1}$ and $p^i = (1; z^i) \in \mathbb{R}^{n+1}$, denote the so-called *logistic loss* as

$$f(x) := \frac{1}{m} \sum_{i=1}^{m} \log \left(1 + \exp(-b_i \cdot x^\top p^i)\right),$$

We select two popular methods for numerical comparison. One is the penalty method of Lu and Zhang [28] proposed a penalty decomposition (PD) method. The other is the first-order method SLEP of [26]. To

compare the solution quality of the three methods, we adopt the criterion, *error rate*, from [28], which is defined by

$$\texttt{Error Rate} := \frac{1}{m}\sum_{i=1}^{m} |\text{sign}(x^\top p^i) - b_i|, \tag{30}$$

where x is the solution obtained by methods and sign(a) is the sign function, i.e., sign$(a) = 1$ if $a > 0$; sign$(a) = -1$ if $a < 0$; sign$(a) = 0$, otherwise.

(a) Parameter setting. We will test two kinds of data sets: real data sets and random data sets to be described below. For the `PD` method, we set $\texttt{eps} = 10^{-3}, \texttt{maxit} = 1000$, and the rest of its parameters are set by default. For `SLEP` method, we set $\texttt{opts.mFlag} = 1, \texttt{opts.lFlag} = 1, \texttt{opts.tFlag} = 2$, and fix $\texttt{rho} = 0.05$ for the random data sets, where `rho` corresponds to l_1 norm penalty parameter λ. However, `rho` is appropriately adjusted for the real data sets. The rest of its parameters are set by default. For our `IIHT`, we use

Table 6: Parameters for `IIHT`.

Real data	$\alpha_0 = 0.01, \quad \beta = 0.2, \quad \sigma = 10^{-5}, \quad \texttt{Maxiter} = 1000, \quad \epsilon = rm/\lambda_{\max}(A^\top A)$
Random data	$\alpha_0 = 0.2, \quad \beta = 0.5, \quad \sigma = 10^{-3}, \quad \texttt{Maxiter} = 1000, \quad \epsilon = m \max\left\{10^{-4}, 10^{\frac{m+n}{1000}-13}\right\}$

Here, $A := [p^1, \cdots, p^m]$ and $r := \max\{m,n\}/\min\{m,n\}, x^0 = (v^0; w^0)$. We always start with w^0 as a zero vector, and initialize $v^0 = 10$ for real data sets but $v^0 = 1$ for random data sets.

(b) Comparison on real data. In our first experiment, we test three real data sets. The first data set is the colon tumor gene expression data[1] with 62 samples and 2000 features. The second data is the ionosphere[2] data with 351 samples and 34 features. The third one is the German Credit data[3] with 1000 samples and 24 features. The first and third data sets are from the UCI machine learning bench market repository [33]. We standardize each data set so that the sample mean is zero and the sample variance is one. We first apply `SLEP` to (29) with a sequence of suitably chosen `rho` to obtain solutions $\widehat{w}$ with an increasing sparsity sequence such as $\|\widehat{w}\|_0 = 1, 2, \cdots, 20$. We then set s being same as $\|\widehat{w}\|_0$ for `PD` and `IIHT`, so that the solutions of these three method are of the same sparsity.

Results for the first two data sets are recorded in Figure 4. In terms of CPU time, `PD` performs poorly for both data sets, while `SLEP` and `IIHT` run very fast. For Colon data, `SLEP` basically gets lowest logistic loss and error rate, and `PD` produces the highest ones. For Ionosphere data, there is no big difference for logistic loss between `PD` and `IIHT`. Both are better than `SLEP`. In terms of error rate, `IIHT` behaves the best, followed by `PD` and `SLEP`.

In fact, the error rate is often used to evaluate the quality of a model vector, which is taken the sum over the testing samples instead of the training samples in (30). It is well known that when the ratio between the number of training samples and the number of features is small, namely, m/n, the error rate is usually high for most of models. Thus, Colon and Ionosphere data sets may not be appropriate for evaluating the error rate. Based on this, we chose German Credit data to estimate it. Specifically, we simply divide this data into two parts: the first 900 samples being training data and the rest 100 samples being testing data. Then we apply three methods in the way as above. The results are shown in Figure 5. For training data, `PD` returns the best results in terms of logistic loss and error rate, followed by `IIHT` and `SLEP`. However, `IIHT` basically outperforms `SLEP` and `PD` for testing data, as it generates lowest logistic loss for most cases.

(c) Comparison on random data. Now we compare the three methods on the random data sets, where the samples $\{z^1, \cdots, z^m\}$ and the corresponding outcomes $\{b_1, \cdots, b_m\}$ are generated in the same manner as [28]. In detail, for each instance we choose equal number of positive and negative samples, that is, $m_+ = m_- = m/2$, where m_+ (resp., m_-) is the number of samples with outcome +1 (resp., -1). The features of positive (resp., negative) samples are independent and identically distributed, drawn from a normal distribution $\mathcal{N}(\phi, 1)$, where ϕ is in turn drawn from a uniform distribution on $[0, 1]$ (resp., $[-1, 0]$). Corresponding pseudo MTALAB codes are:

$$T = \texttt{randperm}(m); \quad b = \texttt{ones}(m, 1); \quad b(T(1 : m/2)) = -1;$$
$$z^i = b_i * \texttt{rand} + \texttt{randn}(n, 1), \quad i = 1, \cdots, m.$$

Data of different sizes are generated. For each size, we randomly generate the data set consisting of 40 trials. For each trial, let $\widehat{w}$ be the approximate optimal solution obtained by `SLEP`. We then apply our `PD` and `IIHT` methods to solve problem (29) with $s = \|\widehat{w}\|_0$ so that the resulting approximate optimal solutions

1. Colon tumor gene expression data: *http://genomics-pubs.princeton.edu/oncology/affydata/index.html.*
2. Ionosphere data: *http://archive.ics.uci.edu/ml/datasets/Ionosphere.*
3. German Credit data: *https://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data).*

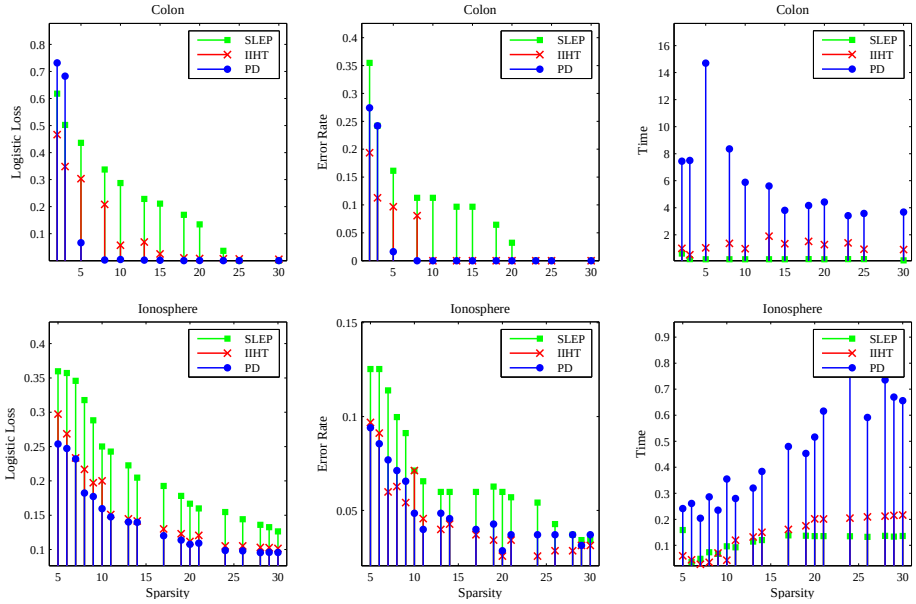

Figure 4: Results for Colon tumor gene expression data and Ionosphere data.

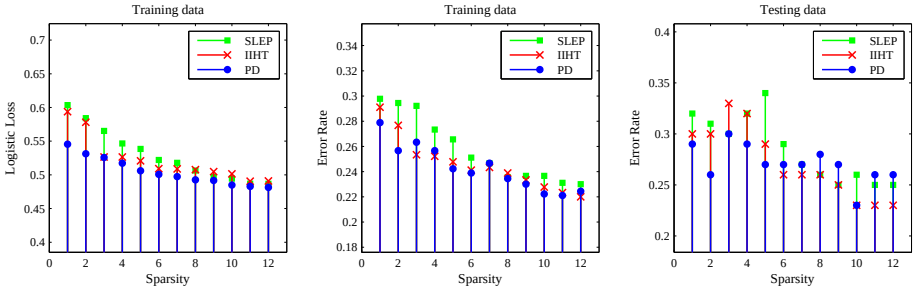

Figure 5: Results for German Credit data.

are at least as sparse as $\widehat{w}$. The results of the three methods for the these randomly generated instances are presented in Table 7. Clearly, `IIHT` obtains the best results with lowest logistic loss and least CPU time. `PD` outperforms `SLEP` in terms of logistic loss but takes the most time.

Table 7: Average results generated by three methods.

			Logistic Loss			Time		
m	n	s	IIHT	PD	SLEP	IIHT	PD	SLEP
	1000	127.2	2.42e-04	2.46e-04	1.65e-01	0.51	13.32	0.75
1000	3000	145.3	4.90e-05	2.26e-04	1.55e-01	1.33	65.41	4.23
	5000	165.2	3.40e-05	1.89e-04	1.50e-01	2.87	100.68	6.83
	1000	195.4	9.94e-05	4.10e-04	1.91e-01	2.22	81.49	3.56
3000	3000	233.3	5.31e-05	4.10e-04	1.83e-01	4.14	201.88	11.33
	5000	246.5	4.21e-05	2.97e-04	1.79e-01	6.08	360.44	20.75
	1000	239.7	3.06e-05	5.72e-04	1.94e-01	4.06	139.51	6.41
5000	3000	304.5	1.58e-05	3.84e-04	1.87e-01	10.07	362.73	18.76
	5000	326.3	2.14e-05	2.96e-04	1.86e-01	16.07	549.65	33.12

4.3 Phase Retrieval Problem

Phase retrieval is the problem that aims at recovering a signal from the magnitude of its Fourier transform. Namely, it is to find a real-valued discrete time signal $x \in \mathbb{R}^N$ from its magnitude-squared of an N point discrete Fourier transform (DFT):

$$b_j = \Big| \sum_{k=1}^{n} x_k e^{-2\pi \mathrm{i}(j-1)(k-1)/N} \Big|^2, \quad j = 1, \cdots, N.$$

Here x is constructed as $x = (x_1, \cdots, x_n, 0, \cdots, 0)^\top \in \mathbb{R}^N$. If we denote F the DFT matrix, then each elements $F_{jk} = e^{-2\pi \mathrm{i}(j-1)(k-1)/N}$ and $b = |Fx|^2$, where $|\cdot|^2$ denotes the element-wise absolute-squared value. Therefore, phase retrieval of sparse signals actually can be reformulated as the following model (see [36] for details).

$$\begin{array}{ll} \min_{x \in \mathbb{R}^N} & \sum_{i=1}^{N} (|F_i x|^2 - b_i)^2, \\ \text{s.t.} & \|x\|_0 \le s, \\ & \mathrm{supp}(x) \subseteq \{1, \cdots, n\}. \end{array}$$

where F_i is the i-th row of F. Actually, phase retrieval of sparse signals is a special case of the more general quadratic compressed sensing (QCS) problem [5, 37]. For our IIHT, we set $\texttt{Maxiter} = 2000, \alpha_0 = 0.001, \epsilon = 10^{-2}$ and $\beta = 0.1$, $\sigma = 10^{-4}$. We generate $y \in \mathbb{R}^n$ with sparsity s as in (28), and then get x_{orig} and b by the pseudo MATLAB codes:

$$\begin{array}{l} x_{\mathrm{orig}} = [y; \texttt{zeros}(N-n, 1)]; \\ b = \texttt{abs}(\texttt{fft}(x_{\mathrm{orig}})).\hat{}2 + \sigma_0 * \texttt{randn}(N, 1); \end{array}$$

In order to evaluate the performance of IIHT, we compare it with GESPAR proposed in [36]. Its parameters are set by default.

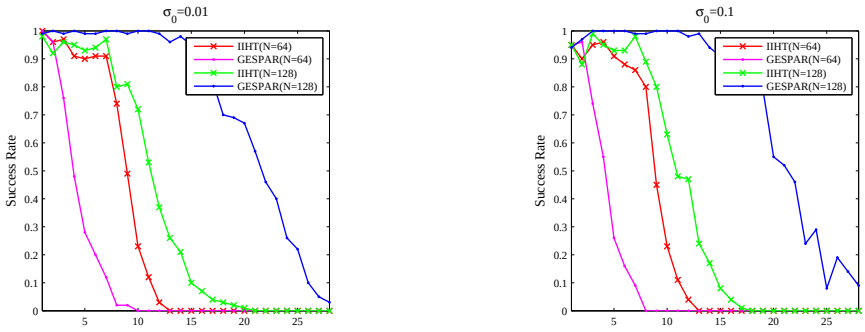

Figure 6: Success rates with three types of matrices.

By fixing $n = 64$ and $N = 64, 128$ under $\sigma_0 = 0.01$ and $\sigma_0 = 0.1$, we test 100 trials for these two methods with different sparsity level s. The corresponding success rate which is defined as before and CPU time are taken into consideration in illustrating their performance. Results shown in Figure 6 demonstrate that IIHT

outperforms `GESPAR` when $N = 64$, while performs worse than `GESPAR` when $N = 128$, regardless of noise level.

To see the accuracy of the solutions and the speed of these two methods, we now run 40 trials with slightly higher dimensions n increasing from 500 to 3000 and keeping $N = 2n, s = 1\%n$. We also test them under two noise levels $\sigma_0 = 0.01$ and $\sigma_0 = 0.1$. We only report results associated with successful recovery, i.e., `Relative Error` < 0.01. Such results are recorded in Table 8, in which `IIHT` outperforms `GESPAR` in terms of both average CPU time and average relative error when $n \geq 1500$.

Table 8: Average results with $N = 2n, s = 1\%n$.

n	$\sigma_0 = 0.01$				$\sigma_0 = 0.1$			
	Time		Relative Error		Time		Relative Error	
	`IIHT`	`GESPAR`	`IIHT`	`GESPAR`	`IIHT`	`GESPAR`	`IIHT`	`GESPAR`
500	1.82	0.82	8.73e-06	4.97e-06	1.58	0.65	7.71e-05	5.72e-05
1000	3.38	8.47	2.73e-06	1.93e-06	4.70	8.82	2.59e-05	1.79e-05
1500	4.22	45.64	1.65e-06	2.07e-04	4.07	54.40	1.65e-05	5.46e-05
2000	7.14	98.53	1.08e-06	4.93e-04	6.33	115.02	1.03e-05	3.46e-04
2500	9.07	284.07	9.57e-07	5.89e-04	9.48	360.91	9.22e-06	1.05e-03
3000	12.45	490.90	7.29e-07	2.69e-04	9.88	754.86	7.73e-06	1.21e-03

5 Conclusion

In this paper, we studied an improved version of the popular Iterated Hard-Thresholding (IHT) algorithm, for the sparsity and nonnegativity constrained optimization, from the veiwpoint of optimization. We try to answer the questions that are common in optimization. Those questions include towards what stationary point that the IHT would converge to and at what speed. In order to answer those questions, we studied the relationships among the three stationary points (α-, B- and C-stationary points) and local (global) minimizers of (1). Moreover, we established some results on convergence and linear convergence rates of IHT by including the Armijo line search in IHT. Numerical experiments demonstrated the efficiency of the improved IHT on three widely tested problems.

Two immediate questions arise from this research. One is to assess whether the nonmonotone line search strategy used by Lu [27] would lead to more efficient performance of IHT and lead to stronger convergence results. The second question is whether we can establish convergence to the global minimizer for the case that the sparse constraint is not tight.

References

[1] H. Attouch, J. Bolte and B. F. Svaiter, Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting and regularized Gauss-Seidel methods, *Math. Program.* 137 (2013) 91-129.

[2] S. Bahmani, B. Raj and P. Boufounos, Greedy sparsity-constrained optimization, *J. Mach. Learn. Res.* 14 (2013) 807-841.

[3] R. Baraniuk, V. Cevher, M. Duarte and C. Hegde, Model-based compressive sensing, *IEEE Trans. Inform. Theorey.* 56 (2010) 1982-2001.

[4] H. H. Bauschke, D. R. Luke, H.M. Phan and X. Wang, Restricted normal cones and sparsity optimization with affine constraints, *Found Comput Math.* 14 (2014) 63-83.

[5] A. Beck and and Y. Eldar, Sparsity constrained nonlinear optimization: optimality conditions and algorithms, *SIAM J Optim.* 23 (2013) 1480-1509.

[6] A. Beck and N. Hallak, On the minimization over sparse symmetric sets. *Math. Oper. Res.* 41 (2015) 196-223.

[7] J. D. Blanchard, J. Tanner and K. Wei, CGIHT: conjugate gradient iterative hard thresholding for compressed sensing and matrix completion, *Inf. Inference: J. IMA*, 4 (2015) 1-36.

[8] T. Blumensath, Compressed sensing with nonlinear observations and related nonlinear optimisation problems, *IEEE Trans. Inform. Theorey.* 59 (2013) 3466-3474.

[9] T. Blumensath and M. E. Davies, Iterative thresholding for sparse approximations, *J Fourier. Anal. Appl.* 14 (2008) 626-654.

[10] T. Blumensath and M. E. Davies, Iterative hard thresholding for compressed sensing, *Appl. Comp. Harmon. Anal.* 27 (2009) 265-274.

[11] T. Blumensath and M. E. Davies, Normalized iterative hard thresholding: Guaranteed stability and performance, *IEEE J. Selected Topics in Signal Processing.* 4 (2010) 298-309.

[12] J. M. Borwein and A. S. Lewis, Convex Analysis and Nonlinear Optimization, Springer, 2010.

[13] P. H. Calamai and J. J. Moré, Projection gradient methods for linearly constrained problems, *Math. Program.* 39 (1987) 93-116.

[14] E. J. Candès and T. Tao, Decoding by linear programming, *IEEE Trans. Inform. Theorey.* 51 (2005) 4203-4215.

[15] C. Cartis and A. Thompson, A new and improved quantitative recovery analysis for iterative hard thresholding algorithms in compressed sensing, *IEEE Trans. Inform. Theorey.* 61 (2015) 2019-2042.

[16] D. Chen and R.,J. Plemmons, Nonnegativity constraints in numerical analysis *In: Proceedings of the Symposium on the Birth of Numerical Analysis, Leuven Belgium*, (2009) 109-140.

[17] W. Dai and O. Milenkovic, Subspace pursuit for compressive sensing signal reconstruction *IEEE Trans. Inform. Theorey.* 55 (2009) 2230-2249.

[18] G. Davis, S. Mallat and M. Avellaneda, Adaptive greedy approximations, *Constr Approx*, 13 (1997) 57-98.

[19] M. Elad, Sparse and redundant representations: From theory to applications in signal and image processing, Springer, 2010.

[20] S. Foucart, Hard thresholding pursuit: an algorithm for compressive sensing, *SIAM J. Numer. Anal*, 49 (2011) 2543-2563.

[21] B. He, M. Tao and X.M. Yuan, A splitting method for separable convex programming, *IMA J. Numer. Anal.* 35 (2015) 394-426.

[22] A. Jalali, C. C. Johnson and P. K. Ravikumar, On learning discrete graphical models using greedy methods *Advances in Neural Information Processing Systems.* 24 (2011) 1935-1943.

[23] C. Kanzow and H.-D. Qi, A QP-free constrained Newton-type method for variational inequality problems, *Math. Program.* 85 (1999) 81-106.

[24] K. Koh, S. J. Kim and S. Boyd, An interior-point method for large-scale l_1-regularized logistic regression, *J. Mach. Learn. Res.* 8 (2007) 1519-1555.

[25] B. Krishnapuram, L. Carin, M. A. Figueiredo and A. J. Hartemink, Sparse multinomial logistic regression: Fast algorithms and generalization bounds, *IEEE Trans. Pattern Anal. Mach. Intell.* 27 (2005), 957-968.

[26] J. Liu, S. Ji and J. Ye, SLEP: Sparse learning with efficient projections. Arizona State University, 2009. Available at http://www.public.asu.edu/∼ jye02/Software/SLEP.

[27] Z. Lu, Optimization over Sparse Symmetric Sets via a Nonmonotone Projected Gradient Method, 2015. Available at http://people.math.sfu.ca/ zhaosong/ResearchPapers/NPG-sparse.pdf

[28] Z. Lu and Y. Zhang, Sparse approximation via penalty decomposition methods,*SIAM. J. Optim.* 23 (2013), pp. 2448-2478.

[29] S. Mallat and Z. Zhang, Matching pursuits with time-frequency dictionaries, *IEEE Trans. Signal Process.* 41 (1993) 3397-3415.

[30] J. J. Moré and D. C. Sorensen, Comuting a trust region step, *SIAM J. Sci. Stat. Comput.*, 4 (1983) 553-572.

[31] D. Needell and J. A. Tropp, CoSaMP: Iterative signal recovery from incomplete and inaccurate samples, *Appl. Comp. Harmon. Anal.* 26 (2009) 301-332.

[32] S. Negahban, P. Ravikumar, M. Wainwright and B. Yu, A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers, *in Advances in Neural Information Processing Systems (NIPS)* 2009.

[33] D. Newman, S. Hettich, C. Blake and C. Merz, UCI repository of machine learning databases, 1998. Available at www.ics.uci.edu/∼ mlearn/MLRepository.html.

[34] L. Pan, N. Xiu and S. Zhou, Gradient support projection algorithm for the affine feasibility problem with sparsity and nonnegativity, 2014. Avaiable at http://arxiv.org/abs/1406.7178

[35] L. Pan, N. Xiu and S. Zhou, On solutions of sparsity constrained optimization, *J. Oper. Res. Soc. China*. 3 (2015) 421-439.

[36] Y. Shechtman, A. Beck and Y. C. Eldar, GESPAR: Efficient phase retrieval of sparse signals, *IEEE Trans. Signal Process.* 62 (2014) 928-938.

[37] Y. Shechtman, Y. C. Eldar, A. Szameit and M. Segev, Sparsity based sub-wavelength imaging with partially incoherent light via quadratic compressed sensing, *Optics express.* 19 (2011) 14807-14822.

[38] L. Wang, The L_1 penalized LAD estimator for high dimensional linear regression, *J. Multivariate Anal.* 120 (2013) 135-151.

[39] X. Yuan, P. Li and T. Zhang, Gradient hard thresholding pursuit for sparsity-constrained optimization, *ICML*, (2014) 127-135.

[40] X. Yuan and Q. Liu, Newton greedy pursuit: A quadratic approximation method for sparsity-constrained optimization, *CVPR*, (2014) 4122-4129