Sparse Regularization: Convergence Of Iterative Jumping Thresholding Algorithm

Jinshan Zeng, Shaobo Lin, and Zongben Xu

***Abstract*—In recent studies on sparse modeling, nonconvex penalties have received considerable attentions due to their superiorities on sparsity-inducing over the convex counterparts. In this paper, we study the convergence of a nonconvex iterative thresholding algorithm for solving a class of sparse regularized optimization problems, where the corresponding thresholding functions of the penalties are discontinuous with jump discontinuities. Therefore, we call the algorithm the iterative jumping thresholding (IJT) algorithm. The finite support and sign convergence of IJT algorithm is first verified via taking advantage of such jump discontinuity. Together with the introduced restricted Kurdyka–Łojasiewicz property, then the global convergence[1] of the entire sequence can be further proved. Furthermore, we can show that the IJT algorithm converges to a strictly local minimizer at an eventual linear rate[2] under some additional conditions. Moreover, we derive *a posteriori* computable error estimate, which can be used to design an efficient terminate rule. It should be pointed out that the ℓ_q quasinorm $(0 < q < 1)$ is an important subclass of the nonconvex penalties studied in this paper. In particular, when applied to the ℓ_q regularization, IJT algorithm can converge to a local minimizer with an eventual linear rate under certain concentration conditions. We also apply the proposed algorithm to sparse signal recovery and synthetic aperture radar imaging problems. The experiment results show the effectiveness of the proposed algorithm.**

***Index Terms*—Sparse regularization, non-convex optimization, iterative thresholding algorithm, ℓ_q regularization $(0 < q < 1)$, Kurdyka-Łojasiewicz inequality.**

I. Introduction

THE sparse regularized optimization problems emerging in many areas of scientific research and engineering practice have attracted considerable attention in recent years. Typical applications include regression [37], visual coding [32], signal processing [20], compressed sensing [10], [23], and microwave imaging [40]. These problems can be intuitively modeled as the following ℓ_0 quasi-norm regularized optimization problem

$$\min_{x\in\mathbf{R}^N} \{F(x) + \lambda\|x\|_0\}, \tag{1}$$

where $F : \mathbf{R}^N \to [0, \infty)$ is a proper lower-semicontinuous function, $\|x\|_0$, commonly called the ℓ_0 quasi-norm, denotes the number of nonzero components of x and $\lambda > 0$ is a regularization parameter. Some efficient algorithms including the iterative *hard* thresholding algorithm ([3], [29]) were developed to solve (1).

Besides the ℓ_0 regularized optimization problem, a more general class of problems are considered in both practice and theory, that is,

$$\min_{x\in\mathbf{R}^N} \{F(x) + \lambda\Phi(x)\}, \tag{2}$$

where $\Phi(x)$ is a certain separable, continuous penalty with $\Phi(x) = \sum_{i=1}^N \phi(|x_i|)$, and $x = (x_1, \cdots, x_N)^T$. One of the most important cases is the ℓ_1-norm with $\Phi(x) = \|x\|_1 = \sum_{i=1}^N |x_i|$. The ℓ_1-norm is convex and thus, the corresponding ℓ_1-norm regularized optimization problem can be efficiently solved. Nevertheless, the ℓ_1-norm may not induce adequate sparsity when applied to certain applications [13]. Alternatively, many non-convex penalties were proposed as relaxations of the ℓ_0 quasi-norm. Some typical non-convex examples are the ℓ_q quasi-norm $(0 < q < 1)$ [13], [14], [39], smoothly clipped absolute deviation (SCAD) [24], and log-sum penalty [11]. Compared with the ℓ_1-norm, the non-convex penalties can usually induce better sparsity while the corresponding non-convex regularized optimization problems are generally more difficult to solve.

There are mainly four classes of algorithms to solve the nonconvex regularized optimization problem (2). The first one is the half-quadratic (HQ) algorithm [26], [27]. HQ algorithms can be efficient when both subproblems are easy to solve (particularly, when both subproblems have closed-form solutions). The second class is the iterative reweighted algorithm including iterative reweighted least squares (IRLS) minimization ([15], [21], [28], [30]) and iterative reweighted ℓ_1-minimization (IRL1) [11] algorithms. The basic idea of the iterative reweighted algorithm is to obtain an approximate sparse solution via solving a sequence of weighted least squares (or, ℓ_1-minimization) problems. The third class is the difference of convex functions algorithm (DC programming) [25]. DC programming method first converts the original problem into the difference of two convex problems (called primal and dual problems, respectively), then iteratively optimizes these two problems. The last class is the iterative thresholding algorithm, which fits the framework of the forward-backward splitting (FBS) algorithm [2] and the generalized gradient projection method [7] when applied to a separable non-convex penalty. Some typical iterative thresholding algorithms include iterative *hard* [3], *soft* [22] and *half* [39] thresholding algorithms. Compared to other types of non-convex

Manuscript received July 27, 2015; revised March 14, 2016 and May 26, 2016; accepted July 17, 2016. Date of publication July 28, 2016; date of current version August 08, 2016. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Ami Wiesel. The work of J. Zeng was supported in part by the National Science Foundation (NSF) under Grant 11501440. The work of S. Lin was supported in part by the NSF under Grants 61502342 and 11401462. *(Corresponding author: Shaobo Lin.)*

J. Zeng is with the College of Computer Information Engineering, Jiangxi Normal University, Nanchang 330022, China (e-mail: jsh.zeng@gmail.com).

S. Lin is with the College of Mathematics and Information Science, Wenzhou University, Wenzhou 325035, China (e-mail: sblin1983@gmail.com).

Z. Xu is with the School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: zbxu@mail.xjtu.edu.cn).

Color versions of one or more of the figures in this letter are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2016.2595499

[1]The global convergence in this paper is defined in the sense that the entire sequence converges regardless of the initial point.

[2]It is also known as asymptotic or local linear rate in other papers.

algorithms such as the HQ, IRLS, IRL1 and DC programming algorithms, the iterative thresholding algorithm is easy to implement and has almost the least computational complexity for large scale problems (see, [40] for instance).

Although the effectiveness of the iterative thresholding algorithms for the non-convex regularized optimization problems has been verified in many applications, except for the iterative *hard* [29] and *half* [41] thresholding algorithms, the convergence of most of these algorithms has not been thoroughly investigated. Basically, the three questions, i.e., *when, where, and how fast does the algorithm converge*, should be answered.

A. Main Contribution

In this paper, we give the convergence analysis for the iterative jumping thresholding algorithm (called IJT algorithm henceforth) for solving a certain class of non-convex regularized optimization problems. The main contributions can be summarized as follows:

a) We prove that the supports and signs of any sequence generated by IJT algorithm can converge with finitely many iterations.
b) Under a further assumption that there exists one limit point such that the objective function satisfies the so-called restricted Kurdyka-Łojasiewicz (rKL) property (see Definition 2) at this point, the whole sequence converges to this point (see Theorem 1).
c) Under certain second-order conditions, we demonstrate that IJT algorithm converges to a strictly local minimizer at an eventual linear rate (see Theorems 2 and 3).
d) When applied to the ℓ_q $(0 < q < 1)$ regularization, IJT algorithm converges to a local minimizer at an eventual linear rate as long as the matrix satisfies a certain concentration property (see Theorem 4).

B. Notations and Organization

We denote $\mathbf{R}$, $\mathbf{N}$ and $\mathbf{C}$ as the sets of real number, natural number and complex number, respectively. For any vector $x \in \mathbf{R}^N$, x_i is its ith component, and for a given index set $I \subset I_N \triangleq \{1, \ldots, N\}$, x_I represents its subvector containing all the components restricted to I. I^c represents the complementary set of I, i.e., $I^c = I_N \backslash I$. $\|x\|_2$ represents the Euclidean norm of a vector x. $\mathrm{Supp}(x)$ is the support of x, i.e., $\mathrm{Supp}(x) = \{i : |x_i| > 0, i = 1, \ldots, N\}$. For any matrix $A \in \mathbf{R}^{N \times N}$, $\sigma_i(A)$ and $\sigma_{\min}(A)$ ($\lambda_i(A)$ and $\lambda_{\min}(A)$) denote as the ith and minimal singular values (eigenvalues) of A, respectively. Similar to the vector case, for a given index set I, A_I represents the submatrix of A containing all the columns restricted to I. For any $z \in \mathbf{R}$, $\mathrm{sign}(z)$ denotes its sign function, i.e.,

$$\mathrm{sign}(z) = \begin{cases} 1, & \text{for } z > 0 \\ 0, & \text{for } z = 0 \\ -1, & \text{for } z < 0 \end{cases}.$$

The remainder of this paper is organized as follows. In Section II, we give the problem settings and then introduce IJT algorithm with some basic properties. In Section III, we give the convergence analysis of IJT algorithm. In Section IV, we apply the developed theoretical analysis to the ℓ_q $(0 < q < 1)$ regularization. In Section V, we give some related works and comparisons. In Section VI, we present some applications to show the effectiveness of the proposed algorithm. We conclude this paper in Section VII. The proofs are presented in Appendix.

II. IJT Algorithm

A. Problem Settings

We make several assumptions on the concerned problem

$$\min_{x \in \mathbf{R}^N} \{T_\lambda(x) = F(x) + \lambda\Phi(x)\}, \tag{3}$$

where $\Phi(x)$ is separable with $\Phi(x) = \sum_{i=1}^N \phi(|x_i|)$.

Assumption 1: $F : \mathbf{R}^N \to [0, \infty)$ is continuously differentiable with Lipschitz continuous gradient, i.e., it holds that

$$\|\nabla F(u) - \nabla F(v)\|_2 \le L\|u - v\|_2, \ \ \forall u, v \in \mathbf{R}^N,$$

where $L > 0$ is the Lipschitz constant.

Note that Assumption 1 is a general assumption for F. For example, the least squares and logistic loss functions used in machine learning are two typical cases.

Assumption 2: $\phi : [0, \infty) \to [0, \infty)$ is continuous and satisfies the following assumptions:

a) ϕ is non-decreasing with $\phi(0) = 0$ and $\phi(z) \to \infty$ when $z \to \infty$.
b) For each $b > 0$, there exists an $a > 0$ such that $\phi(z) \ge az^2$ for $z \in [0, b]$.
c) ϕ is differentiable on $(0, \infty)$, and its first derivative ϕ' is strictly convex with $\phi'(z) \to \infty$ for $z \to 0$ and $\lim_{z \to \infty} \phi'(z)/z = 0$.
d) ϕ has a continuous second derivative ϕ'' on $(0, \infty)$.

Most of the above assumptions were considered in [7]. It can be observed that Assumption 2(a) ensures the coercivity of ϕ, and thus the existence of the minimizer. Assumption 2(b) guarantees the weakly sequential lower semi-continuity of ϕ in l^2, and Assumption 2(c) is assumed to induce the sparsity. In practice, there are many non-convex functions satisfying Assumption 2. Two of the most typical subclasses are $\phi(z) = z^q$ and $\phi(z) = \log(1 + z^q)$ with $q \in (0, 1)$ as shown in Fig. 1.

B. IJT Algorithm

In order to describe IJT algorithm, we need to define the following proximity operator of Φ,

$$Prox_{\mu, \lambda\Phi}(x) = \arg\min_{u \in \mathbf{R}^N} \left\{ \frac{\|x - u\|_2^2}{2\mu} + \lambda\Phi(u) \right\}, \tag{4}$$

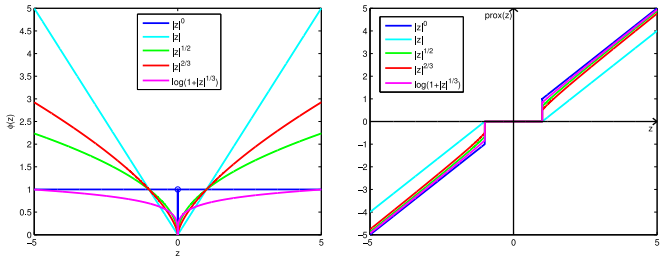
where $\mu > 0$ is a parameter. Since Φ is separable, computing $Prox_{\mu, \lambda\Phi}$ can be reduced to solve a one-dimensional minimization problem, that is,

$$prox_{\mu, \lambda\phi}(z) = \arg\min_{v \in \mathbf{R}} \left\{ \frac{|z - v|^2}{2\mu} + \lambda\phi(|v|) \right\}. \tag{5}$$

Therefore,

$$Prox_{\mu, \lambda\Phi}(x) = (prox_{\mu, \lambda\phi}(x_1), \cdots, prox_{\mu, \lambda\phi}(x_N))^T. \tag{6}$$

We list some useful results on $prox_{\mu, \lambda\phi}$ obtained in [7].



(a) (b)

Fig. 1. Typical ϕ satisfying Assumption 2 and the corresponding thresholding functions. We plot the figures of $\phi(|z|) = |z|^{1/2}, |z|^{2/3}, \log(1+|z|^{1/3})$, and their corresponding thresholding functions. For comparison, we also plot the figures of two well-known cases, i.e., ℓ_0-norm with $\phi(|z|) = 1_{|z|>0}$ as the indicator function of $|z| > 0$, ℓ_1-norm with $\phi(|z|) = |z|$, and their corresponding thresholding functions. (a) Typical penalty functions. (b) Thresholding functions.

Lemma 1. ([7, Lemmas 3.2 and 3.3]): Assume that ϕ satisfies Assumption 2, then

a) for each $\mu > 0$, the function $\rho_\mu : z \mapsto z + \lambda\mu\phi'(z)$ is well defined on $\mathbf{R}_+$;
b) the function $\psi : z \mapsto 2(\phi(z) - z\phi'(z))/z^2$ is strictly decreasing and one-to-one on $(0,\infty) \to (0,\infty)$;
c) for any $z > 0$, $\phi''(z)$ is negative and monotonically increasing;
d) $prox_{\mu,\lambda\phi}$ is well defined and can be specified as

$$prox_{\mu,\lambda\phi}(z) = \begin{cases} \text{sign}(z)\rho_\mu^{-1}(|z|), & \text{for } |z| \geq \tau_\mu \\ 0, & \text{for } |z| \leq \tau_\mu \end{cases}, \quad (7)$$

for any $z \in \mathbf{R}$ with

$$\tau_\mu = \rho_\mu(\eta_\mu) \text{ and } \eta_\mu = \psi^{-1}((\lambda\mu)^{-1}). \quad (8)$$

Moreover, the range of $prox_{\mu,\lambda\phi}$ is $\{0\} \cup [\eta_\mu, \infty)$.

It can be observed that the proximity operator is discontinuous with a jump discontinuity, which is one of the most significant features of such a class of non-convex penalties studied in this paper. Henceforth, we call $prox_{\mu,\lambda\phi}$ the *jumping* thresholding function. Moreover, it can be easily checked that the proximity operator is not nonexpansive in general. (Some specific proximity operators are shown in Fig. 1(b).)

Formally, the iterative form of IJT algorithm can be expressed as follows:

$$x^{n+1} \in Prox_{\mu,\lambda\Phi}(x^n - \mu\nabla F(x^n)), \quad (9)$$

where $\mu > 0$ is a step size parameter. For simplicity, we define

$$G_{\mu,\lambda\Phi}(x) = Prox_{\mu,\lambda\Phi}(x - \mu\nabla F(x)), \quad x \in \mathbf{R}^N,$$

and its fixed point set $\mathcal{F}_\mu \triangleq \{x : x = G_{\mu,\lambda\Phi}(x)\}$.

C. Some Basic Properties of IJT Algorithm

Property 1: Let x^* be a fixed point of $G_{\mu,\lambda\Phi}$ and $\{x^n\}$ be a sequence generated by IJT algorithm, then it holds

a) for any $i \in \text{Supp}(x^*)$, $|x_i^*| \geq \eta_\mu$ and $[\nabla F(x^*)]_i + \lambda\text{sign}(x_i^*)\phi'(|x_i^*|) = 0$; and for any $i \in \text{Supp}(x^*)^c$, $|x_i^*| = 0$ and $|[\nabla F(x^*)]_i| \leq \tau_\mu/\mu$;
b) for any $i \in \text{Supp}(x^{n+1})$, $|x_i^{n+1}| \geq \eta_\mu$ and $x_i^{n+1} + \lambda\mu\,\text{sign}(x_i^{n+1})\phi'(|x_i^{n+1}|) = x_i^n - \mu[\nabla F(x^n)]_i$; and for any $i \in \text{Supp}(x^{n+1})^c$, $|x_i^{n+1}| = 0$ and $|x_i^n - \mu[\nabla F(x^n)]_i| \leq \tau_\mu$, $n \in \mathbf{N}$,

where $[\nabla F(x^*)]_i$ and $[\nabla F(x^{n+1})]_i$ represent the ith component of $\nabla F(x^*)$ and $\nabla F(x^{n+1})$ respectively.

This property can be easily derived by the definition of proximity operator and Lemma 1(d). Actually, Property 1(a) is a certain type of optimality conditions of problem (3). We call x^* a *stationary point* of (3) if x^* satisfies Property 1(a), and we denote by Ω_μ the stationary point set for a given μ. Then according to Property 1(a), it holds $\mathcal{F}_\mu \subset \Omega_\mu$.

Property 2: Let $\{x^n\}$ be a sequence generated by IJT algorithm with a bounded initialization. Assume that $0 < \mu < \frac{1}{L}$, then it holds

a) $T_\lambda(x^{n+1}) \leq T_\lambda(x^n) - \frac{1}{2}(\frac{1}{\mu} - L)\|x^{n+1} - x^n\|_2^2$, and there exists a constant T_λ^* such that $\lim_{n\to\infty} T_\lambda(x^n) \to T_\lambda^*$;
b) $\|x^{n+1} - x^n\|_2 \to 0$ as $n \to \infty$;
c) each accumulation point of $\{x^n\}$ is a fixed point of $G_{\mu,\lambda\Phi}$;
d) if $\{x^n\}$ possess an isolated accumulation point, then the whole sequence converges to some $x^* \in \mathcal{F}_\mu$.

This property can be claimed from [7, Propositions 2.1, 2.3 and Corollary 2.1] with $\mu_n \equiv \mu$. Property 2(a) is commonly called the sufficient decrease property, which is a basic property desired for a descent method. Let $\mathcal{X}$ be the accumulation point set of $\{x^n\}$, then by Property 2(c), $\mathcal{X} \subset \mathcal{F}_\mu$, and further by Property 1(a), $\mathcal{X} \subset \Omega_\mu$.

Property 3: Suppose that $0 < \mu < \frac{1}{L}$, then each global minimizer of T_λ is a fixed point of $G_{\mu,\lambda\Phi}$. Let $\mathcal{M}$ be the set of global minimizers, then $\mathcal{M} \subset \mathcal{F}_\mu$.

Property 3 is a corollary of [7, Propositions 2.2] with a uniform step size. From Properties 2 and 3, the following relations hold

$$\mathcal{X} \subset \mathcal{F}_\mu, \mathcal{M} \subset \mathcal{F}_\mu \text{ and } \mathcal{F}_\mu \subset \Omega_\mu.$$

III. Convergence Analysis

In this section, we will answer the basic questions concerning IJT algorithm presented in introduction, i.e., when, where and how fast does the algorithm converge?

A. rKL Property

Kurdyka-Łojasiewicz (KL) property has been widely used to prove the convergence of the non-convex algorithms (see, [2] for instance).

Definition 1. (KL property): A function $f : \mathbf{R}^N \to \mathbf{R} \cup \{+\infty\}$ is said to have the KL property at $x^* \in \text{dom}(\partial f)$ if there exist $\eta \in (0, +\infty]$, a neighborhood U of x^* and a continuous concave function $\varphi : [0, \eta) \to \mathbf{R}_+$ such that:

i) $\varphi(0) = 0$ and φ is $\mathcal{C}^1$ on $(0, \eta)$;
ii) for all $s \in (0, \eta)$, $\varphi'(s) > 0$;
iii) for all x in $U \cap \{x : f(x^*) < f(x) < f(x^*) + \eta\}$, the KL inequality holds

$$\varphi'(f(x) - f(x^*))\text{dist}(0, \partial f(x)) \geq 1. \quad (10)$$

Proper lower semi-continuous functions which satisfy the KL inequality at each point of dom(∂f) are called KL functions.

The KL property of f at some point x^* means that "f is amenable to sharpness at x^*" [6], and the KL inequality (10) is equivalent to

$$\operatorname{dist}(0, \partial(\varphi \circ (f(x) - f(x^*)))) \geq 1, \tag{11}$$

for all $x \in U \cap \{x : f(x^*) < f(x) < f(x^*) + \eta\}$ (simply use the "one-sided" chain-rule [35, Theorem 10.6]). KL functions include real analytic functions, semialgebraic functions and locally strongly convex functions (more information can be referred to Sec. 2.2 in [38] and references therein). However, according to [4] (Sec. 1, page 1), some simple functions such as $f(x) = \exp(-\frac{1}{x^2}), \forall x \in \mathbf{R}$, are not KL function, and in the latter proof of Proposition 1 (see Appendix B), a class of simple functions are shown to be not KL functions.

Motivated by this, in this paper, we introduce another related but weaker property called the rKL property. Before describing the definition of rKL property formally, we define a projection mapping associated with an index set $I \subset I_N$,

$$P_I : \mathbf{R}^N \to \mathbf{R}^{|I|}, P_I x = x_I, \forall x \in \mathbf{R}^N.$$

We also denote P_I^T as the transpose of P_I,

$$P_I^T : \mathbf{R}^{|I|} \to \mathbf{R}^N, (P_I^T z)_I = z \text{ and } (P_I^T z)_{I^c} = 0, \forall z \in \mathbf{R}^{|I|},$$

where $|I|$ is the cardinality of I and $I^c = I_N \backslash I$.

Definition 2. (rKL property): A function $f : \mathbf{R}^N \to \mathbf{R} \cup \{+\infty\}$ is said to have the rKL property at $x^* \in \operatorname{dom}(\partial f)$, if $g : \mathbf{R}^{|I|} \to \mathbf{R} \cup \{+\infty\}, g(z) = f(P_I^T z)$ satisfies KL property at $z^* = x_I^*$ with $I = \operatorname{Supp}(x^*)$.

From Definition 2, rKL property only requires that the subgradient of f with respect to the nonzero variables can get sharp after certain a concave transform, while KL property requires such well property for all variables around some point. In the following, we give a sufficient condition of the rKL property.

Lemma 2: Given an index set $I \subset I_N$, consider the function $g(z) = f(P_I^T z)$. Assume that z^* is a stationary point of g (i.e., $\nabla g(z^*) = 0$), and g is twice continuously differentiable at a neighborhood of z^*, i.e., $B(z^*, \epsilon_0)$ for some $\epsilon_0 > 0$. Moreover, if $\nabla^2 g(z^*)$ is nonsingular, then f satisfies the rKL property at $P_I^T z^*$. Actually, it holds

$$|g(z) - g(z^*)| \leq C^* \|\nabla g(z)\|_2^2, \forall z \in B(z^*, \epsilon),$$

for some $0 < \epsilon < \epsilon_0$ and a positive constant $C^* > 0$.

The proof of this lemma is shown in Appendix A. Then we present a proposition to show that rKL property is an extension of KL property.

Proposition 1. (rKL is a generalization of KL): If f satisfies the KL property at x^*, then f satisfies the rKL property at x^*, but not vice versa.

The proof of this proposition is presented in Appendix B. According to the proof procedure of Proposition 1, the conditions listed in Lemma 2 are essential for the rKL property in the sense that there exists a function satisfying conditions in Lemma 2 but not KL property.

B. Convergence of Entire Sequence

Lemma 3. [Finite Support Convergence]: Let $\{x^n\}$ be a sequence generated by IJT algorithm and $I^n = \operatorname{Supp}(x^n)$. Assume that $0 < \mu < \frac{1}{L}$, then there exist a positive integer n^*, an index set I and a sign vector S^* such that when $n > n^*$, the following hold

a) $I^n = I$ and $\operatorname{Supp}(x^*) = I, \forall x^* \in \mathcal{X}$,

b) $\operatorname{sign}(x^n) = S^*$ and $\operatorname{sign}(x^*) = S^*, \forall x^* \in \mathcal{X}$.

The proof of this lemma is presented in Appendix C. According to Lemma 3, the support and sign freeze with finitely many iterations. Furthermore, by Lemma 3, we can claim that $\{x^n\}$ converges to x^* if the new sequence $\{x^{i+n^*}\}_{i \in \mathbf{N}}$ converges to x^*, which is also equivalent to the convergence of the sequence $\{z^{i+n^*}\}_{i \in \mathbf{N}}$,

$$z^{i+n^*} \to z^* \text{ as } i \to \infty \tag{12}$$

with $z^{i+n^*} = P_I x^{i+n^*}$ and $z^* = P_I x^*$. Let

$$\hat{z}^n = z^{n+n^*}, \tag{13}$$

then $\{\hat{z}^n\}$ has the same convergence behavior of $\{x^n\}$.

For any $\epsilon > 0$, we define a one-dimensional real set

$$\mathbf{R}_\epsilon \triangleq \mathbf{R} \backslash [-\epsilon, \epsilon].$$

Particularly, let $\mathbf{R}_0 = \mathbf{R} \backslash \{0\}$. We let

$$\mathcal{Z}^* \triangleq P_I \mathcal{X} = \{P_I x^* : x^* \in \mathcal{X}\},$$

then $\mathcal{Z}^*$ is the accumulation point set of the sequence $\{\hat{z}^n\}$. We define $T : \mathbf{R}_{\eta_\mu/2}^{|I|} \to \mathbf{R}$ and $f : \mathbf{R}_{\eta_\mu/2}^{|I|} \to \mathbf{R}$ as

$$T(z) = T_\lambda(P_I^T z) \text{ and } f(z) = F(P_I^T z), \forall z \in \mathbf{R}_{\eta_\mu/2}^{|I|}. \tag{14}$$

For any $z^* \in \mathcal{Z}^*$, it can be observed from Property 1(a) that $z^* \in \mathbf{R}_{\eta_\mu}^{|I|}$ and z^* is a stationary point of T. Moreover, we define a series of mappings $\phi_{1,m} : \mathbf{R}_0^m \to \mathbf{R}^m$ and $\phi_{2,m} : \mathbf{R}_0^m \to \mathbf{R}^{m \times m}$ as follows

$$\begin{aligned} \phi_{1,m}(z) &= (\operatorname{sign}(z_1)\phi'(|z_1|), \cdots, \operatorname{sign}(z_m)\phi'(|z_m|))^T, \\ \phi_{2,m}(z) &= \operatorname{diag}(\phi''(|z_1|), \cdots, \phi''(|z_m|)), m = 1, \ldots, N, \end{aligned} \tag{15}$$

where $\operatorname{diag}(z)$ represents the diagonal matrix generated by z. For brevity, we denote $\phi_{1,m}$ and $\phi_{2,m}$ as ϕ_1 and ϕ_2 respectively when m is fixed and there is no confusion.

By Properties 1 and 2, we can easily verify that $\{\hat{z}^n\}$ satisfies the following properties.

Lemma 4: $\{\hat{z}^n\}$ satisfies the following:

a) (Sufficient decrease condition). For each $n \in \mathbf{N}$,

$$T(\hat{z}^{n+1}) \leq T(\hat{z}^n) - \frac{1}{2}\left(\frac{1}{\mu} - L\right)\|\hat{z}^{n+1} - \hat{z}^n\|_2^2.$$

b) (Relative error condition). For each $n \in \mathbf{N}$,

$$\|\nabla T(\hat{z}^{n+1})\|_2 \leq (\frac{1}{\mu} + L)\|\hat{z}^{n+1} - \hat{z}^n\|_2.$$

c) (Continuity condition). There exists a subsequence $\{\hat{z}^{n_j}\}_{j \in \mathbf{N}}$ and z^* such that

$$\hat{z}^{n_j} \to z^* \text{ and } T(\hat{z}^{n_j}) \to T(z^*), \text{ as } j \to \infty.$$

Lemma 4(a) and (c) are obvious by Property 2, the specific form of T (14) and the construction of $\{\hat{z}^n\}$ (13). Lemma 4(b) holds mainly due to Property 1(b) and Assumptions 1–2. Specifically, by Property 1(b), it can be easily checked that

$$\hat{z}^{n+1} + \lambda\mu\phi_1(\hat{z}^{n+1}) = \hat{z}^n - \mu\nabla f(\hat{z}^n),$$

which implies

$$\mu(\nabla f(\hat{z}^{n+1}) + \lambda\phi_1(\hat{z}^{n+1})) = (\hat{z}^n - \hat{z}^{n+1}) + \mu(\nabla f(\hat{z}^{n+1}) - \nabla f(\hat{z}^n)).$$

Thus, $\|\nabla T(\hat{z}^{n+1})\|_2 =$

$$\frac{1}{\mu}\|(\hat{z}^n - \hat{z}^{n+1}) + \mu(\nabla f(\hat{z}^{n+1}) - \nabla f(\hat{z}^n))\|_2.$$

By Assumption 1, ∇F is Lipschitz continuous, then

$$\|\nabla f(\hat{z}^{n+1}) - \nabla f(\hat{z}^n)\|_2 \le \|\nabla F(P_I^T\hat{z}^{n+1}) - \nabla F(P_I^T\hat{z}^n)\|_2 \le L\|P_I^T\hat{z}^{n+1} - P_I^T\hat{z}^n\|_2 = L\|\hat{z}^{n+1} - \hat{z}^n\|_2.$$

Therefore, $\|\nabla T(\hat{z}^{n+1})\|_2 \le (\frac{1}{\mu} + L)\|\hat{z}^{n+1} - \hat{z}^n\|_2$.

From Lemma 4, if T further has the KL property at the limit point z^*, then according to Theorem 2.9 in [2], $\{\hat{z}^n\}$ converges to z^*. Note that the construction form of $\{\hat{z}^n\}$, we can obtain the following convergence result.

Theorem 1. [Global Convergence]: Assume that F and ϕ satisfy Assumptions 1 and 2, respectively. Let $\{x^n\}$ be a sequence generated by IJT algorithm. Suppose that $0 < \mu < \frac{1}{L}$, then $\{x^n\}$ converges subsequentially to a set $\mathcal{X}$. If further there is a limit point $x^* \in \mathcal{X}$ at which T_λ satisfies the rKL property, then the whole sequence converges to x^*.

Together with Lemma 2, the following corollary holds.

Corollary 1: Under Assumptions 1 and 2, suppose that $0 < \mu < \frac{1}{L}$, and that there exists a limit point x^* of $\{x^n\}$ such that F is twice continuously differentiable at x^* and $\nabla^2 T(P_I x^*)$ is nonsingular, then $\{x^n\}$ converges to x^*.

Remark 1: A similar condition is also used to guarantee the convergence of the steepest descent method in [34, Theorem 2, pp. 266]. Obviously, if z^* is a strictly local minimizer (or maximizer), or a strict saddle point of T, then the nonsingularity of $\nabla^2 T(z^*)$ holds naturally. Therefore, if T is locally strict convex or concave, then Corollary 1 holds.

C. Convergence to a Strictly Local Minimizer

As shown in Corollary 1, if $\nabla^2 T(P_I x^*)$ is nonsingular at some limit point x^*, then the sequence generated by IJT algorithm converges to x^*. In this subsection, we will justify that x^* is also a strictly local minimizer of the optimization problem if $\nabla^2 T(P_I x^*)$ is positive definite.

Theorem 2. [Convergence to a Strictly Local Minimizer]: Under assumptions of Corollary 1, if further $\nabla^2 T(P_I x^*)$ is positive definite, then x^* is a strictly local minimizer of T_λ.

The proof of this theorem is rather intuitive. By Property 1(a) we have

$$[\nabla F(x^*)]_I + \lambda\phi_1(x_I^*) = 0. \tag{16}$$

This together with the condition of the theorem

$$\nabla^2 T(P_I x^*) = \nabla_{II}^2 F(x^*) + \lambda\phi_2(x_I^*) \succ 0$$

imply that the second-order optimality conditions hold at $x^* = (x_I^*, 0)$, where $\nabla_{II}^2 F(x^*) = \frac{\partial^2 F(x)}{\partial x_I^2}\big|_{x=x^*}$. For sufficiently small vector h, we denote $x_h^* = (x_I^* + h_I, 0)$. It then follows

$$F(x_h^*) + \lambda\sum_{i\in I}\phi(|x_i^* + h_i|) \ge F(x^*) + \lambda\sum_{i\in I}\phi(|x_i^*|). \tag{17}$$

Furthermore, by Assumption 2(c), it obviously holds that

$$\phi(t) > (\|[\nabla F(x^*)]_{I^c}\|_\infty + 2)t/\lambda,$$

for sufficiently small $t > 0$. By this fact and the differentiability of F, for sufficiently small h, there hold

$$\begin{aligned} F(x^* + h) - F(x_h^*) + \lambda\sum_{i\in I^c}\phi(|h_i|) &= h_{I^c}^T[\nabla F(x^*)]_{I^c} + \lambda\sum_{i\in I^c}\phi(|h_i|) + o(h_{I^c}) \\ &\ge \sum_{i\in I^c}(\|[\nabla F(x^*)]_{I^c}\|_\infty - [\nabla F(x^*)]_i + 1)|h_i| \ge 0. \end{aligned} \tag{18}$$

Summing up the above two inequalities (17)–(18), one has that for all sufficiently small h,

$$T_\lambda(x^* + h) - T_\lambda(x^*) \ge 0, \tag{19}$$

and hence x^* is a local minimizer. Moreover, we can observe that when $h \neq 0$, then at least one of these two inequalities (17) and (18) will hold strictly, which implies that x^* is a strictly local minimizer.

D. Eventual Linear Convergence Rate

In order to derive the convergence rate of IJT algorithm, we first show some observations on ∇F and ϕ' in the neighborhood of x^*. For any $0 < \varepsilon < \eta_\mu$, we define a neighborhood of x^* as follows

$$\mathcal{N}(x^*, \varepsilon) = \{x \in \mathbf{R}^N : \|x_I - x_I^*\|_2 < \varepsilon, x_{I^c} = 0\}.$$

If F is twice continuously differentiable at x^* and also $\lambda_{\min}(\nabla_{II}^2 F(x^*)) > 0$, then for any $x \in \mathcal{N}(x^*, \varepsilon)$, there exist two sufficiently small positive constants c_F and c_ϕ (both c_F and c_ϕ depending on ε with $c_F \to 0$ and $c_\phi \to 0$ as $\varepsilon \to 0$) such that

$$\langle[\nabla F(x)]_I - [\nabla F(x^*)]_I, x_I - x_I^*\rangle \ge (\lambda_{\min}(\nabla_{II}^2 F(x^*)) - c_F)\|x_I - x_I^*\|_2^2, \tag{20}$$

$$\langle\phi_1(x_I) - \phi_1(x_I^*), x_I - x_I^*\rangle \ge (\phi''(e) - c_\phi)\|x_I - x_I^*\|_2^2, \tag{21}$$

where (21) holds for ϕ' being strictly convex on $(0, \infty)$, and thus ϕ'' being nondecreasing on $(0, \infty)$, consequently, $\min_{i\in I}\phi''(|x_i^*|) = \phi''(\min_{i\in I}|x_i^*|)$. With the observations (20) and (21), we obtain the following theorem.

Theorem 3. (Eventual Linear Rate): Under conditions of Corollary 1, if the following conditions also hold

a) $\lambda_{\min}(\nabla_{II}^2 F(x^*)) > 0$;
b) $0 < \lambda < -\frac{\lambda_{\min}(\nabla_{II}^2 F(x^*))}{\phi''(e)}$,
c) either $0 < \mu < \min\{\frac{2(\lambda_{\min}(\nabla_{II}^2 F(x^*)) + \lambda\phi''(e))}{L^2 - (\lambda\phi''(e))^2}, \frac{1}{L}\}$, or, for any sufficiently small $0 < \varepsilon < \eta_\mu$, the third derivative

ϕ''' is well-defined, bounded and nonzero on the set $\cup_{i\in I} B(x_i^*, \varepsilon)$, where $B(x_i^*, \varepsilon) := (x_i^* - \varepsilon, x_i^* + \varepsilon)$,

where $e = \min_{i\in I}|x_i^*|$, then there exists a positive integer n_0 and a constant $\rho \in (0,1)$ such that when $n > n_0$,

$$\|x^{n+1} - x^*\|_2 \le \rho \|x^n - x^*\|_2, \text{ and}$$

$$\|x^{n+1} - x^*\|_2 \le \frac{\rho}{1-\rho}\|x^{n+1} - x^n\|_2.$$

The proof of Theorem 3 is presented in Appendix D. As shown by this theorem, if we can fortunately obtain a good initial point, then IJT algorithm may converge fast with a linear rate. On the other hand, Theorem 3 also provides a posteriori computable error estimate of the algorithm, which can be used to design an efficient terminate rule of IJT algorithm. It can be observed that the conditions of Theorem 3 are slightly stricter than those of Theorem 2, and thus, x^* is also a strictly local minimizer under the conditions of Theorem 3.

IV. Application to ℓ_q $(0 < q < 1)$ Regularization

The ℓ_q $(0 < q < 1)$ regularization is formulated as follows:

$$\min_{x\in\mathbf{R}^N}\left\{T_\lambda(x) = \frac{1}{2}\|Ax - y\|_2^2 + \lambda\|x\|_q^q\right\}, \tag{22}$$

where $A \in \mathbf{R}^{M\times N}$ (commonly, $M < N$), $y \in \mathbf{R}^M$, and $\|x\|_q^q = \sum_{i=1}^N |x_i|^q$. The proximity operator $prox_{\mu,\lambda|\cdot|^q}$ can be expressed as (see [7])

$$prox_{\mu,\lambda|\cdot|^q}(z) = \begin{cases} (\cdot + \lambda\mu q \text{sign}(\cdot)|\cdot|^{q-1})^{-1}(z), & |z| \ge \tau_{\mu,q} \\ 0, & |z| \le \tau_{\mu,q} \end{cases} \tag{23}$$

for any $z \in \mathbf{R}$, where

$$\tau_{\mu,q} = \frac{2-q}{2-2q}(2\lambda\mu(1-q))^{\frac{1}{2-q}}, \tag{24}$$

$$\eta_{\mu,q} = (2\lambda\mu(1-q))^{\frac{1}{2-q}}, \tag{25}$$

and the range of $prox_{\mu,\lambda|\cdot|^q}$ is $\{0\} \cup [\eta_{\mu,q}, \infty)$. Specifically, for some special q (say, $q = 1/2, 2/3$), the corresponding proximity operators can be expressed analytically [39], [12].

According to [2] (See Example 5.4, page 122), the function $T_\lambda(x) = \frac{1}{2}\|Ax - y\|_2^2 + \lambda\|x\|_q^q$ is a KL function and obviously satisfies the rKL propety at any limit point. Then we can obtain the following corollary directly.

Corollary 2: Let $\{x^n\}$ be a sequence generated by IJT algorithm for ℓ_q regularization with $q \in (0,1)$. Assume that $0 < \mu < \frac{1}{\|A\|_2^2}$, then $\{x^n\}$ converges to a stationary point of ℓ_q regularization.

Moreover, it is easy to check that $\phi(z) = z^q$ satisfies the second part of condition (c) in Theorem 3. Therefore, the eventual linear convergence rate of IJT algorithm for ℓ_q regularization can be claimed as follows.

Corollary 3: Under conditions of Corollary 2, if the following conditions also hold:

a) $\lambda_{\min}(A_I^T A_I) > 0$,

b) $0 < \lambda < \frac{\lambda_{\min}(A_I^T A_I)e^{2-q}}{q(1-q)}$,

where $I = \text{Supp}(x^*)$ and $e = \min_{i\in I}|x_i^*|$, then IJT algorithm converges to a strictly local minimizer x^* with an eventual linear rate.

It can be observed that the minimal nonzero entry e of x^* is used in condition (b) of this corollary. A theoretical lower bound of e is estimated by Chen *et al.* [17]. In the following, we derive another sufficient conditions through the observation that the threshold value (25) is generally a tighter lower bound of e than that studied in [17]. Specifically, by (25), it holds

$$e \ge \eta_{\mu,q} = (2\lambda\mu(1-q))^{\frac{1}{2-q}}. \tag{26}$$

Then if $\frac{\lambda_{\min}(A_I^T A_I)}{\|A\|_2^2} > \frac{q}{2}$ and $\frac{q}{2\lambda_{\min}(A_I^T A_I)} < \mu < \frac{1}{\|A\|_2^2}$, the conditions in Corollary 3 hold naturally.

Theorem 4: Under conditions of Corollary 2, if the following conditions still hold:

a) $\dfrac{\lambda_{\min}(A_I^T A_I)}{\|A\|_2^2} > \dfrac{q}{2}$,

b) $\dfrac{q}{2\lambda_{\min}(A_I^T A_I)} < \mu < \dfrac{1}{\|A\|_2^2}$,

then IJT algorithm converges to a strictly local minimizer x^* with an eventual linear rate.

From Theorem 4, it means that if the matrix A satisfies a certain concentration property and the step size μ is chosen appropriately, then IJT algorithm can converge to a local minimizer with an eventual linear rate. Note that the condition (a) in Theorem 4 implies $\frac{q}{2\lambda_{\min}(A_I^T A_I)} < \frac{1}{\|A\|_2^2}$ naturally. Thus, the condition (b) of Theorem 4 is a natural and reachable condition and, furthermore, whenever this condition is satisfied, the sequence $\{x^n\}$ is indeed convergent by Corollary 2. This shows that only condition (a) is essential in Theorem 4. We notice that condition (a) is a concentration condition on eigenvalues of the submatrix $A_I^T A_I$, and, in particular, it implies

$$\lambda_{\min}(A_I^T A_I) > q\lambda_{\max}(A_I^T A_I)/2,$$

or equivalently

$$\text{Cond}(A_I^T A_I) := \frac{\lambda_{\max}(A_I^T A_I)}{\lambda_{\min}(A_I^T A_I)} < \frac{2}{q}, \tag{27}$$

where $\text{Cond}(A_I^T A_I)$ is the condition number of $A_I^T A_I$. (27) thus shows that the submatrix $A_I^T A_I$ is well-conditioned with the condition number lower than $2/q$.

In recent years, a property called the restricted isometry property (RIP) of a matrix A was introduced to characterize the concentration degree of the eigenvalues of its submatrix with k columns [9]. A matrix A is said to be of the k-order RIP (denoted then by δ_k-RIP) if there exists a $\delta_k \in (0,1)$ such that

$$(1-\delta_k)\|x\|_2^2 \le \|Ax\|_2^2 \le (1+\delta_k)\|x\|_2^2, \tilde{\forall}\|x\|_0 \le k. \tag{28}$$

In other words, the RIP ensures that all submatrices of A with k columns are close to an isometry, and therefore distance-preserving. Let $K = \|x^*\|_0$. It can be seen from (28) that if A possesses δ_K-RIP with $\delta_K < \frac{2-q}{2+q}$, then

$$\text{Cond}(A_I^T A_I) \le \frac{1+\delta_K}{1-\delta_K} < \frac{2}{q}.$$

Thus, we can claim that when A satisfies a certain RIP, the condition (a) in Theorem 4 can be satisfied. In particular, we have the following proposition.

Proposition 2: Assume that $K < N/2$ and A satisfies δ_K-RIP with $\delta_K < \frac{2-q}{2+2qN/K}$ or δ_{2K}-RIP with $\delta_{2K} < \frac{2-q}{2+qN/K}$, then the condition (a) in Theorem 4 holds.

This can be directly checked by the facts that $\lambda_{\min}(A_I^T A_I) \geq 1-\delta_K$, $\lambda_{\min}(A_I^T A_I) \geq 1-\delta_{2K}$, $\lambda_{\max}(A^T A) \leq 1+\delta_N$, $\delta_N \leq \frac{2N}{K}\delta_K$ and $\delta_N \leq \frac{N}{K}\delta_{2K}$ (c.f. [18, Proposition 1]).

From Proposition 2, we can see, for instance, when $q = 1/2, K/N = 1/3$ and A satisfies δ_K-RIP with $\delta_K < 3/10$ or δ_{2K}-RIP with $\delta_{2K} < 3/7$, the condition (a) in Theorem 4 is satisfied, and therefore, by Theorem 4, IJT algorithm converges to a local minimizer of the ℓ_q regularization with an eventual linear rate. It is noted that in the condition of Proposition 2, we always have $\delta_K < \frac{2-q}{2+4q}$ and $\delta_{2K} < \frac{2-q}{2+2q}$.

Remark 2: In [41], Zeng *et al.* have justified the convergence of a specific iterative thresholding algorithm called the iterative *half* thresholding algorithm for $\ell_{1/2}$ regularization. It can be observed that the convergence results of the iterative *half* thresholding algorithm obtained in [41] is just a special case of the results presented in this section.

Remark 3: Recently, Lu [29] proposed an iterative *hard* thresholding method and its variant for solving ℓ_0 regularization over a conic constraint, and established its convergence as well as the iteration complexity. Although the ℓ_0 quasi-norm does not satisfies Assumption 2, it can be observed that the finite support and sign convergence property (i.e., Lemma 3) holds naturally for *hard* algorithm due to the *hard* thresholding function possesses the same discontinuity of the jumping thresholding function. Furthermore, once the support of the sequence converges, the iterative form of *hard* algorithm reduces to a simple Landweber iteration, and thus the convergence and eventual linear convergence rate of *hard* algorithm can be directly claimed.

V. Related Works and Comparisons

There are many non-convex algorithms like HQ [1], FOCUSS [28], IRL1 [16] and DC programming [25] for solving the composite optimization problem. Compared with these algorithms, we derive a sufficient condition instead of the direct assumption that the accumulation points are isolated, for the convergence of IJT algorithm. Moreover, the convergence speed of IJT algorithm is also demonstrated in this paper.

Besides the aforementioned non-convex algorithms, there are some other tightly related non-convex algorithms mainly including two generic algorithms and some specific algorithms. The first generic algorithm very related to IJT algorithm is the generalized gradient projection method (called GGPM for short) [7], which is proposed to solve the following general non-convex optimization model in the infinite-dimensional Hilbert space

$$\min_{x\in\mathbf{X}} \{F(x) + \lambda\Phi(x)\}, \tag{29}$$

where $\mathbf{X}$ is an infinite-dimensional Hilbert space, $F : \mathbf{X} \to [0,\infty)$ is assumed to be a proper lower-semicontinuous function with Lipschitz continuous gradient $\nabla F(x)$ (the Lipschitz constant is L), and $\Phi : \mathbf{X} \to [0,\infty)$ is weakly lower-semicontinuous and satisfies Assumption 2. Specifically, the iteration of GGPM is

$$x^{n+1} \in Prox_{\mu_n,\lambda\Phi}(x^n - \mu_n \nabla F(x^n)). \tag{30}$$

The main convergence results of GGPM are stated as follows.

Theorem A. ([7, Theorem 4.1]): Consider $\{x^n\}$ generated by GGPM with step size sequence $\{\mu_n\}$ satisfying $0 < \underline{\mu} \leq \mu_n \leq \bar{\mu} < L^{-1}$. Then $x^n (n \geq 1)$ has a finite support, and the support only changes finitely many times. Moreover, every subsequence of $\{x^n\}$ has a strong accumulation point x^*, which is a fixed point in the sense $x^* \in G_{\mu^*,\lambda\Phi}(x^*)$ for some $\mu^* \in [\underline{\mu}, \bar{\mu}]$.

When applied to the ℓ_q $(0 < q < 1)$ regularization (22), the convergence results of GGP are strengthened as follows.

Theorem B. ([7, Theorem 5.1]): If $q \in (0,1)$ is rational and the following hold: (a) the set $\{x : A^T Ax = \|A^T A\|x\}$ is a finite-dimensional subspace, (b) the eigenvalues of $A^T A$ do not accumulate at $\|A^T A\|$, and (c) A satisfies the finite basis injective (FBI) property, i.e., A is injective whenever restricted to finitely many coefficients, then each sequence $\{x^n\}$ generated by GGPM with $\mu_n = \frac{n+1}{\|A^T A\|(n+1)+1}$ converges to a quasi-global minimizer of T_λ.

It can be observed from (30) that IJT algorithm fits the framework of GGPM in the finite-dimensional space with a uniform step size, i.e., $\mu_n \equiv \mu$. When applied to a general model (that is, under Assumptions 1 and 2), only subsequence convergence of GGPM is claimed as shown in Theorem A, while the global convergence of IJT algorithm is justified under the so-called rKL property as shown in Theorem 1. Moreover, the eventual linear convergence rate of IJT algorithm is also estimated in this paper. Furthermore, when applied to the ℓ_q $(0 < q < 1)$ regularization, the global convergence of GGPM is claimed only for ℓ_q regularization with rational q under certain conditions (see, Theorem B), while the global convergence of IJT algorithm for ℓ_q regularization with any q is claimed under the condition that $0 < \mu < 1/\|A^T A\|$ (see, Corollary 2). From Theorem B, the conditions (a)-(c) for the global convergence of GGPM generally mean that $A^T A$ is positive definite and possesses certain concentration property. While by Corollary 3 and Theorem 4, the eventual linear convergence rate of IJT algorithm can be estimated under the similar positive definite or concentration conditions.

Another tightly related generic algorithm is the inexact descent method studied in [2]. Such class of descent methods mainly include proximal algorithms, FBS, and regularized Gauss-Seidel methods. Let $f : \mathbf{R}^N \to \mathbf{R} \cup \{+\infty\}$ be a proper lower semicontinuous function, and $\{x^n\}$ be a sequence generated by the concerned descent method for solving the optimization problem

$$\underset{x\in\mathbf{R}^N}{\text{minimize}}\, f(x).$$

Moreover, the sequence $\{x^n\}$ is assumed to satisfy the following three abstract hypotheses.

- H1 *(Sufficient Decrease Condition)* For each $n \in \mathbf{N}$, $f(x^{n+1}) + a\|x^{n+1} - x^n\|^2 \leq f(x^n)$ for some $a > 0$.
- H2 *(Relative Error Condition)* For each $n \in \mathbf{N}$, there exists $w^{n+1} \in \partial f(x^{n+1})$ such that $\|w^{n+1}\| \leq b\|x^{n+1} - x^n\|$

for some $b > 0$, where ∂f denotes by the *limiting-subdifferential* of f ([31]).

H3 *(Continuity Condition)* There exists a subsequence $\{x^{n_j}\}_{j\in\mathbf{N}}$ and x^* such that $x^{n_j} \to x^*$ and $f(x^{n_j}) \to f(x^*)$ as $j \to \infty$.

The following is an abstract convergence result of a sequence $\{x^n\}$ satisfying H1–H3 [2].

Theorem C. ([2, Theorem 2.9]): Consider a sequence $\{x^n\}_{n\in\mathbf{N}}$ that satisfies H1–H3. If f has the KL property at the cluster point x^* specified in H3, then the sequence $\{x^n\}_{n\in\mathbf{N}}$ converges to x^*, and x^* is a critical point of f. Moreover, $\sum_{n=0}^{\infty} \|x^{n+1} - x^n\| < +\infty$.

It is shown in [2] that the FBS algorithm for non-convex function satisfies H1–H3. Specifically, FBS is generally proposed to solve the following composite optimization problem:

$$\text{minimize}_x\ f(x) := g(x) + h(x), \tag{31}$$

where g is smooth, and h is nonsmooth. Then the iteration of FBS can be described as follows:

$$x^{n+1} \in Prox_{\mu_n,h}(x^n - \mu_n \nabla g(x^n)), \tag{32}$$

and its convergence result is shown in the following.

Theorem D. ([2, Theorem 5.1]): Let $f = g + h$ be a proper lower semicontinuous KL function which is bounded from below. Assume further that $g : \mathbf{R}^N \to \mathbf{R}$ is finite valued, differentiable, has a L-Lipschitz continuous gradient, and that the restriction of h to its domain is continuous. If $\{x^n\}_{n\in\mathbf{N}}$ is a bounded sequence generated by FBS with $0 < \underline{\mu} < \mu_n < \bar{\mu} < 1/L$, then it converges to some critical point of f. Moreover, the sequence $\{x^n\}_{n\in\mathbf{N}}$ has a finite length.

It can be observed from (32) that IJT algorithm fits the framework of FBS with a fixed step size. As shown in Theorem D, if the step size $0 < \mu_n < 1/L$, the convergence of FBS is justified via taking advantage of the KL property of the objective function and under the boundedness assumption of the sequence. While the global convergence of IJT algorithm is justified under Assumptions 1–2, and the rKL property of the objective function as shown in Theorem 1. From Proposition 1, the rKL property is weaker than KL property, and can be viewed as a *generalization* of KL property. In addition, we give a sufficient condition to verify the rKL property as shown in Lemma 2. Besides the global convergence, we also give the estimate of the convergence rate of IJT algorithm under certain conditions (see, Theorem 3).

Moreover, there are some other specific iterative thresholding algorithms related to IJT. Among them, the *hard* algorithm and the *soft* algorithm are two representatives, which respectively solves the ℓ_0 regularization [3] and ℓ_1 regularization [22]. As shown in [3] and [22], when $\mu = 1$ both $hard$ and $soft$ algorithms can converge to a stationary point whenever $\|A\|_2 < 1$. These classical convergence results can be generalized when a step size parameter μ is incorporated with the iterative procedures, and in this case, the convergence condition becomes

$$0 < \mu < \|A\|_2^{-2}. \tag{33}$$

It can be seen from Corollary 2 that (33) is the same condition of the convergence of IJT when applied to the ℓ_q regularization with $0 < q < 1$, which then supports that the classical convergence results of iterative thresholding algorithms have been extended to the non-convex ℓ_q $(0 < q < 1)$ regularization case. Furthermore, it was shown in [8] that when the measurement matrix A satisfies the so-called FBI property and the stationary point possesses a strict sparsity pattern, the *soft* algorithm can converge to a global minimizer of ℓ_1 regularization with a linear convergence rate. Such result is not surprising because of the convexity of ℓ_1 regularization. As for convergence speed of the *hard* algorithm, it was demonstrated in [3] that under the condition $\mu = 1$ and $\|A\|_2 < 1$, *hard* algorithm converges to a local minimizer with an eventual linear convergence rate. However, as algorithms for solving non-convex models, Corollary 3 and Theorem 4 reveal that IJT shares the same eventual convergence speed with *hard* algorithm.

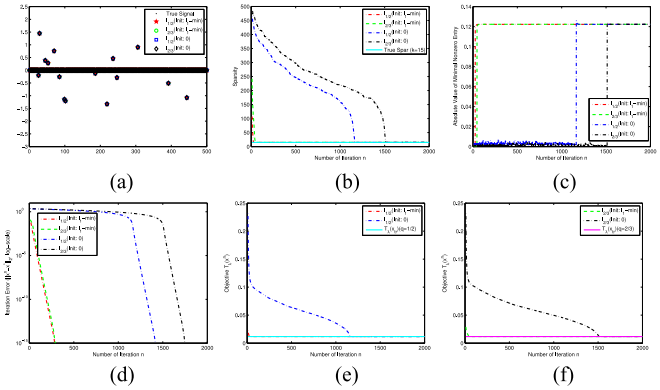


Fig. 2. Experiment for sparse signal recovery. (a) Recovery results. (b) Trends of sparsity level. (c) Trends of the absolute value of minimal nonzero entry. (d) Convergence rate, i.e., trends of $\|x^n - x^*\|_2$, where $x^* = x^{n_0}$ with $n_0 = 5000$ is taken as an approximation of the limit point. (e) Objective sequence for $\ell_{1/2}$ regularization. (f) Objective sequence for $\ell_{2/3}$ regularization. The labels "$l_{1/2}$ (Init: l_1-min)" and "$l_{2/3}$ (Init: l_1-min)" represent the cases of $\phi(|z|) = |z|^{1/2}$ and $\phi(|z|) = |z|^{2/3}$ with an inexact solution of the l_1-minimization problem as the initial guess, respectively. The labels "$l_{1/2}$ (Init: 0)" and "$l_{2/3}$ (Init: 0)" represent the cases of $\phi(|z|) = |z|^{1/2}$ and $\phi(|z|) = |z|^{2/3}$ with 0 as the initial guess, respectively. The Recovery RMSEs of the four cases, that is, $l_{1/2}$ (Init: l_1-min), $l_{2/3}$ (Init: l_1-min), $l_{1/2}$ (Init: 0) and $l_{2/3}$ (Init: 0) are $6.63e{-}5$, $7.35e{-}5$, $6.74e{-}5$ and $7.52e{-}5$, respectively.

VI. Applications

We first implement a synthesized sparse signal recovery experiment to demonstrate the effectiveness of IJT algorithm, and then show its performance when applied to a real application, i.e., sparse synthetic aperture radar (SAR) imaging. While the effectiveness of IJT algorithm for other applications like image deconvolution can be referred to [12]. (The corresponding matlab code of IJT algorithm can be got from https://github.com/JinshanZeng/IJT_Alg.)

A. Sparse Signal Recovery Problem

The sparse signal recovery problem aims to reconstruct a sparse signal from its incomplete measurements. For this purpose, given a true sparse signal x_{tr} with dimension $N = 500$ and sparsity $k = 15$, shown as in Fig. 2(a), its underdetermined observation is synthesized via $y = Ax_{tr}$, where the measurement matrix A is of dimension $M \times N = 250 \times 500$ with Gaussian

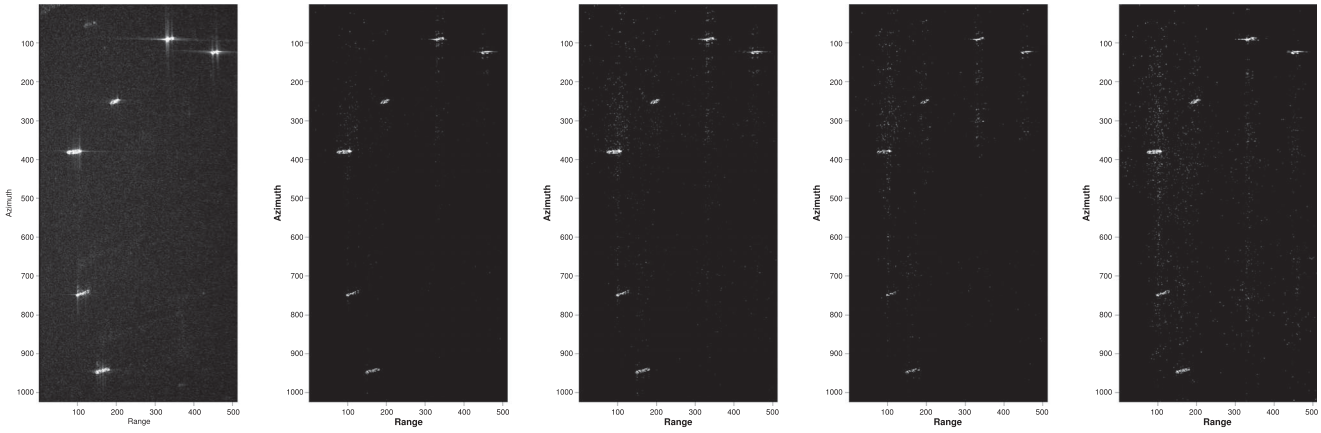


Fig. 3. RADARSAT-1 data imaging results via the traditional SAR imaging method (Range Doppler algorithm (RDA) [19]) using full sampling and IJT algorithm with $\phi(|z|) = |z|^{1/2}$ and $|z|^{2/3}$ using undersampling. (a) The traditional radar image under the full sampling data. (b)–(e) IJT for $\ell_{1/2}$ or $\ell_{2/3}$ regularization under the sampling rates 10% or 5%, respectively. The details are better seen by zooming on a computer scene.

$\mathcal{N}(0, 1/250)$ i.i.d. entries. Such measurement matrix is known to satisfy (with high probability) the RIP with optimal bounds [36]. We then applied IJT algorithm to the problem with two different non-convex penalties, that is, $\phi(|z|) = |z|^{1/2}, |z|^{2/3}$. In both cases, the jumping thresholding operators can be analytically expressed as shown in [39] and [12], respectively, and thus the corresponding IJT algorithms can be efficiently implemented. In both cases, $\mu = 0.99\|A\|_2^{-2}$, and λ were hand-tuned to be optimal ($\lambda = 0.0012$ and 0.0015 for $\ell_{1/2}$ and $\ell_{2/3}$, respectively). Moreover, we considered two different initial guesses including 0 and an inexact solution of the ℓ_1-minimization problem to justify the effect on the convergence speed. The experiment results are reported in Fig. 2.

It can be seen from Fig. 2(a) that the true sparse signal can be recovered with high accuracies in all cases. By Fig. 2(b), the sparsity levels of the iterates converge to the true sparsity level with finitely many iterations. Actually, the freezed support is exactly the true support of x_{tr}. Once the true support has been identified, the minimal nonzero entry of the iterate will be away from 0 with a deterministic distance as demonstrated by Fig. 2(c). The (eventual) linear convergence rate of IJT algorithm starting with a given initial guess can be also observed from Fig. 2(d). In both $\ell_{1/2}$ and $\ell_{2/3}$ cases, the objective sequences converge to the corresponding objective values at x_{tr}, which are generally the global minimum in both cases, as shown by Fig. 2(e) and (f).

B. Sparse SAR Imaging

SAR imaging is an inverse scattering problem that a spatial map of reflectivity is reconstructed from measurements of scattered electric fields, and it is normally modeled as an ill-posed linear inverse problem. Specifically, the SAR observation model is generally formulated as

$$y = A\xi + u, \tag{34}$$

where $A \in \mathbf{C}^{M\times N}$ is the SAR observation matrix, which is determined by SAR acquisition system and observation geometry, $\xi \in \mathbf{C}^N$ is the reflectivity of target, $u \in \mathbf{C}^M$ is the observation noise, and $y \in \mathbf{C}^M$ is the obtained observation. Assume that the reflectivity ξ is sparse under a basis $\Psi \in \mathbf{C}^{n\times n}$ with the sparse representation x. Then linear model (34) can be rewritten as

$$y = A\xi + u = A\Psi x + u = \bar{A}x + u,$$

where $\bar{A} = A\Psi$ and $\xi = \Psi x$. The sparse SAR imaging problem can be modeled as

$$\min_{x\in\mathbf{C}^N} \left\{ \frac{1}{2}\|y - \bar{A}x\|_2^2 + \lambda\Phi(x) \right\}, \tag{35}$$

where $\Phi(x)$ is a regularization term to induce the sparsity. We can then apply IJT algorithm to solve the sparse SAR imaging problem (35).

The used SAR dataset is from RADARSAT-1 in the fine mode-2 about Vancouver region. The detailed target and data descriptions are provided in [19]. We are interested in the region of the English Bay, where there are six sitting vessels. Such region is a typical sparse scene under the identity basis. The main radar parameters are listed as follows: the signal bandwidth is 30.111 MHz, the pulse repetition frequency is 1256.98 Hz. We applied IJT algorithm to solve the sparse SAR imaging problem with two different Φ's, i.e., $\Phi(x) = \sum_{i=1}^N |x_i|^{1/2}$ and $\sum_{i=1}^N |x_i|^{2/3}$ under different sampling rates. The experiment results are shown in Fig. 3.

Compared with the traditional SAR reconstruction result, IJT algorithm reconstructs higher quality images with increased resolution and reduced sidelobes at much lower sampling rate than the Nyquist rate, as shown in Fig. 3. More specifically, IJT algorithm implements SAR imaging effectively under 10% or even lower sampling rate, as demonstrated by Fig. 3(b)–(e).

VII. Conclusion

We have conducted a study of the convergence of IJT algorithm for a class of non-convex regularized optimization problems. One of the most significant features of such

class of iterative thresholding algorithms is that the associated thresholding functions are discontinuous with jump discontinuities. Among such class of non-convex optimization problems, the ℓ_q $(0 < q < 1)$ regularization problem is a typical subclass.

The main contribution of this paper is the establishment of the convergence and rate-of-convergence results of IJT algorithm. We first show that the support and sign of the sequence generated by IJT algorithm converge with finitely many iterations. Then we show the convergence of the entire sequence under the rKL property. Furthermore, we demonstrate that IJT algorithm converges to a strictly local minimizer with an eventual linear rate under some second-order conditions. When applied to the ℓ_q $(0 < q < 1)$ regularization, IJT algorithm can converge to a strictly local minimizer at an eventual linear rate as long as the matrix satisfies a certain concentration property. The obtained convergence results to a local minimizer generalize those known for the *soft* and *hard* algorithms. We also apply the proposed algorithm to some typical sparse applications, which demonstrate the effectiveness of IJT algorithm.

APPENDIX

A. Proof of Lemma 2

Proof: Note that $\nabla g(z^*) = 0$, then

$$|g(z) - g(z^*)| = |g(z) - g(z^*) - \nabla g(z^*)^T(z - z^*)| \leq \int_0^1 \|\nabla g(z^* + t(z - z^*)) - \nabla g(z^*)\|_2 \|z - z^*\|_2 dt. \tag{36}$$

Since g is twice continuously differentiable at $B(z^*, \epsilon_0)$, then it obviously exists constants $L_g > 0$ such that

$$\|\nabla g(z^* + t(z - z^*)) - \nabla g(z^*)\|_2 \leq L_g t \|z - z^*\|_2,$$

for any $z \in B(z^*, \epsilon_0)$ and $t \in (0, 1)$. Thus, it follows

$$|g(z) - g(z^*)| \leq \frac{L_g}{2}\|z - z^*\|_2^2, \forall z \in B(z^*, \epsilon_0). \tag{37}$$

On the other hand, for any $z \in B(z^*, \epsilon_0)$, there exists a $t_0 \in (0, 1)$ such that

$$\|\nabla g(z)\|_2 = \|\nabla g(z) - \nabla g(z^*)\|_2 = \|\nabla^2 g(z^* + t_0(z - z^*))(z - z^*)\|_2. \tag{38}$$

Since $\nabla^2 g(z^*)$ is nonsingular and by the continuity of $\nabla^2 g(z)$ at $B(z^*, \epsilon_0)$, then there exists $0 < \epsilon < \epsilon_0$ such that for any $z \in B(z^*, \epsilon)$,

$$\sigma_{\min}(\nabla^2 g(z^* + t_0(z - z^*))) \geq \min_{z \in B(z^*, \epsilon)} \sigma_{\min}(\nabla^2 g(z)) > 0.$$

Denote $\sigma_{\epsilon, z^*} \triangleq \min_{z \in B(z^*, \epsilon)} \sigma_{\min}(\nabla^2 g(z))$, then (38) becomes

$$\|\nabla g(z)\|_2 \geq \sigma_{\epsilon, z^*}\|z - z^*\|_2. \tag{39}$$

Let $C^* = \frac{L_g}{2\sigma^2_{\epsilon, z^*}}$. Combining (37) and (39), it implies

$$|g(z) - g(z^*)| \leq C^* \|\nabla g(z)\|_2^2.$$

■

B. Proof of Proposition 1

Proof: Let $I = \text{Supp}(x^*)$ and $z^* = x^*_I$. Define

$$g : \mathbf{R}^{|I|} \to \mathbf{R} \cup \{+\infty\}, g(z) = f(P_I^T z).$$

Assume that f satisfies the KL property at x^*, then by definition, there exist constants $\eta > 0$ and $\epsilon > 0$, a concave function φ and a neighborhood of x^*, $\mathcal{N}_{x^*} \triangleq \{x : \|x - x^*\|_2 \leq \epsilon\} \cap \{x : f(x^*) < f(x) < f(x^*) + \eta\}$, such that

$$\text{dist}(0, \partial(\varphi \circ (f(x) - f(x^*)))) \geq 1, \forall x \in \mathcal{N}_{x^*}.$$

Let $\mathcal{N}_{z^*}$ be a neighborhood of z^* with

$$\mathcal{N}_{z^*} = \{z : \|z - z^*\|_2 \leq \epsilon\} \cap \{z : g(z^*) < g(z) < g(z^*) + \eta\},$$

and $\mathcal{N}^*_{x^*} \triangleq \{P_I^T z : \forall z \in \mathcal{N}_{z^*}\} = \{x : x_I^c = 0, \|x_I - x^*_I\|_2 \leq \epsilon\} \cap \{x : f(x^*) < f(x) < f(x^*) + \eta\}$. Then $\mathcal{N}^*_{x^*} \subset \mathcal{N}_{x^*}$, and thus,

$$\text{dist}(0, \partial(\varphi \circ (f(x) - f(x^*)))) \geq 1, \forall x \in \mathcal{N}^*_{x^*}$$

which implies that

$$\text{dist}(0, \partial(\varphi \circ (g(z) - g(z^*)))) \geq 1, \forall z \in \mathcal{N}_{z^*}. \tag{40}$$

It follows that g satisfies the KL property at z^*. Therefore, f satisfies the rKL property at x^*.

In the following, we give a counterexample to show that f satisfies rKL but not KL property at its global minimizer. Specifically, let $f : \mathbf{R}^2 \to \mathbf{R}, f(x_1, x_2) = f_1(x_1) + f_2(x_2)$, with $f_1(x_1) = \frac{1}{2}|x_1 - \frac{3}{2}|^2 + |x_1|^{1/2}$, and $f_2(x_2) = \exp(-\frac{1}{x_2^2})$. It can be easily observed that f_1 reaches the global minima at $x_1^* = 1$, and f_2 reaches the global minima at $x_2^* = 0$. As a consequence, $x^* = (1, 0)$ is a global minimizer of f. According to [2] (See Example 5.4, page 122), f_1 is a KL function, and thus, satisfies the KL property at $x_1^* = 1$, while by [4] (Sec. 1, page 1), f_2 fails to satisfy the KL property at $x_2^* = 0$. By the definition of rKL property, we can show that f satisfies the rKL property at x^*.

We then prove that f does not satisfy KL property at x^* by contradiction. Assume this is not the case, that is, f satisfies the KL property at x^*, then there exist constants $\epsilon > 0$, $\eta > 0$ and a concave function φ such that

$$\text{dist}(0, \partial(\varphi \circ (f(x) - f(x^*)))) \geq 1, \tag{41}$$

for any $x \in \mathcal{N}_{x^*}$ with $\mathcal{N}_{x^*} \triangleq \{x : \|x - x^*\|_2 \leq \epsilon\} \cap \{x : f(x^*) < f(x) < f(x^*) + \eta\}$. Particularly, we consider another neighborhood of x^*, $\mathcal{N}^*$ defined as

$$\mathcal{N}^* = \{x : x_1 = 1, |x_2| \leq \epsilon\} \cap \{x : f(x^*) < f(x) < f(x^*) + \eta\}.$$

It is obvious that $\mathcal{N}^* \subset \mathcal{N}_{x^*}$, and thus (41) holds for any $x \in \mathcal{N}^*$. In this case, (41) reduces to

$$\text{dist}(0, \partial(\varphi \circ (f_2(x_2) - f_2(0)))) \geq 1, \tag{42}$$

for any $x_2 \in \{u : |u| \leq \epsilon\} \cap \{f_2(0) < f_2(u) < f_2(0) + \eta\}$. It follows that f_2 satisfies the KL property at 0, which contradicts with the fact that f_2 does not satisfy the KL property at 0 ([4], Sec. 1, page 1).

Such counterexample can be extended to much more general cases. Let f_1 be any one-dimensional KL function, and f_2 is still defined above. Then $f(x_1,x_2)=f_1(x_1)+f_2(x_2)$ satisfies the rKL but not KL property at any point $(x_1,0)$ with $x_1\neq 0$. ■

C. Proof of Lemma 3

Proof: We first show the convergence of $\{I^n\}$, then justify the convergence of $\{\text{sign}(x^n)\}$.

(i) Convergence of $\{I^n\}$: By Property 2(b), there exists a sufficiently large positive integer n_0 such that $\|x^n-x^{n+1}\|_2<\eta_\mu$ when $n>n_0$. We first show that

$$I^{n+1}=I^n,\forall n>n_0 \tag{43}$$

by contradiction. Assume this is not the case, that is, $I^{n_1+1}\neq I^{n_1}$ for some $n_1>n_0$. Then it is easy to derive a contradiction through distinguishing the following two possible cases:

Case 1: $I^{n_1+1}\neq I^{n_1}$ and $(I^{n_1+1}\cap I^{n_1})\subset I^{n_1+1}$. In this case, there exists an i_{n_1} such that $i_{n_1}\in I^{n_1+1}\backslash I^{n_1}$. By Property 1(b), it then implies

$$\|x^{n_1+1}-x^{n_1}\|_2\geq|x^{n_1+1}_{i_{n_1}}|\geq\eta_\mu,$$

which contradicts to $\|x^{n_1+1}-x^{n_1}\|_2<\eta_\mu$.

Case 2: $I^{n_1+1}\neq I^{n_1}$ and $(I^{n_1+1}\cap I^{n_1})=I^{n_1+1}$. Under this circumstance, it is obvious that $I^{n_1+1}\subset I^{n_1}$. Thus, there exists an k_{n_1} such that $k_{n_1}\in I^{n_1}\backslash I^{n_1+1}$. It then follows from Property 1 that

$$\|x^{n_1+1}-x^{n_1}\|_2\geq|x^{n_1}_{k_{n_1}}|\geq\eta_\mu,$$

and it contradicts to $\|x^{n_1+1}-x^{n_1}\|_2<\eta_\mu$. Thus, (43) holds true. It also means that the support sequence $\{I^n\}$ converges. We denote I the limit of I^n. Then for any $n>n_0$, $I^n=I$.

For any limit point $x^*\in\mathcal{X}$, there exists a subsequence $\{x^{n_j}\}$ converging to x^*, i.e.,

$$x^{n_j}\to x^*\ \text{ as }\ j\to\infty. \tag{44}$$

Thus, there exists a positive integer j_0 such that $n_{j_0}>n_0$ and $\|x^{n_j}-x^*\|_2<\eta_\mu$ when $j\geq j_0$. Similar to the previous proof, it can be also claimed that $I^{n_j}=\text{Supp}(x^*)$ for any $j\geq j_0$. On the other hand, by (43), $I^{n_j}=I$. Thus, for any limit point x^*, $\text{Supp}(x^*)=I$.

Taking $n^*=n_{j_0}$, then by the above analysis, it is obvious that the claim (a) in Lemma 3 holds.

(ii) Convergence of $\{sign(x^n)\}$: As $I^n=I=\text{Supp}(x^*)$ for any $n>n^*$ and $x^*\in\mathcal{X}$, it suffices to show that $\text{sign}(x_i^{n+1})=\text{sign}(x_i^n)$ and $\text{sign}(x_i^{n_j})=\text{sign}(x_i^*)$ for any $i\in I$, $j\geq j_0$, $n>n^*$. Similar to the first part (i) of proof, we will first check that $\text{sign}(x_i^{n+1})=\text{sign}(x_i^n)$, and then $\text{sign}(x_i^{n_j})=\text{sign}(x_i^*)$ for any $i\in I$ by contradiction. We now prove $\text{sign}(x_i^{n+1})=\text{sign}(x_i^n)$ for any $i\in I$ and $n>n^*$. Assume this is not the case. Then there exists an $i^*\in I$ such that $\text{sign}(x_{i^*}^{n+1})\neq\text{sign}(x_{i^*}^n)$, and hence,

$$\text{sign}(x_{i^*}^{n+1})\text{sign}(x_{i^*}^n)=-1.$$

From Property 1(b), it is easy to check

$$\|x^{n+1}-x^n\|_2\geq|x_{i^*}^{n+1}-x_{i^*}^n|=|x_{i^*}^{n+1}|+|x_{i^*}^n|\geq 2\eta_\mu,$$

which contradicts again to $\|x^{n+1}-x^n\|_2<\eta_\mu$. This contradiction shows $\text{sign}(x^{n+1})=\text{sign}(x^n)$ when $n>n^*$. It follows that $\{\text{sign}(x^n)\}$ converges. Let S^* be the limit of $\{\text{sign}(x^n)\}$. Similarly, we can also show that $\text{sign}(x^{n_j})=\text{sign}(x^*)$ whenever $j\geq j_0$. Therefore, $\text{sign}(x^n)=S^*=\text{sign}(x^*)$ when $n>n^*$ and for any $x^*\in\mathcal{X}$. ■

D. Proof of Theorem 3

Proof: We first prove Theorem 3 under conditions (a), (b) and the first part of (c), and then prove it under conditions (a), (b) and the second part of (c).

Part A: Let $C_2\triangleq\sqrt{1-2\mu\lambda_{\min}(\nabla^2_{II}F(x^*))+\mu^2L^2}$, $C_1\triangleq 1+\lambda\mu\phi''(e)$. By the conditions of Theorem 3, it is easy to check that $C_1>C_2>0$. Since both c_F and c_ϕ approach to zero as ε approaches zero, then we can take a sufficiently small $0<\varepsilon<\eta_\mu$ such that

$$0<c_F<\min\left\{\frac{(C_1-C_2)(C_1+3C_2)}{8\mu},\lambda_{\min}(\nabla^2_{II}F(x^*))\right\},$$

and $0<c_\phi<\frac{C_1-C_2}{2\lambda\mu}$. Let $\alpha_{F,\varepsilon}\triangleq\lambda_{\min}(\nabla^2_{II}F(x^*))-c_F$ and $\alpha_{\phi,\varepsilon}\triangleq-\phi''(e)+c_\phi$, then under conditions of Theorem 3, there hold $0<\alpha_{F,\varepsilon}<L$ and $\alpha_{\phi,\varepsilon}>0$, and further

$$1-\lambda\mu\alpha_{\phi,\varepsilon}=1+\lambda\mu\phi''(e)-\lambda\mu c_\phi>\frac{C_1+C_2}{2}>0, \tag{45}$$

$$1-2\mu\alpha_{F,\varepsilon}+\mu^2L^2\geq 1-2\mu\alpha_{F,\varepsilon}+\mu^2\alpha^2_{F,\varepsilon}\geq 0, \tag{46}$$

$$1-2\mu\alpha_{F,\varepsilon}+\mu^2L^2=C_2^2+2\mu c_F<\left(\frac{C_1+C_2}{2}\right)^2. \tag{47}$$

Since $\{x^n\}$ converges to x^*, then for any $0<\varepsilon<\eta_\mu$, there exists a sufficiently large integer $n_0>n^*$ (where n^* is specified as in Lemma 3) such that $\|x^n-x^*\|_2<\varepsilon$ when $n>n_0$. Let $I^n=\text{Supp}(x^n)$. By Lemma 3, it holds $I^n=I$ and $\text{sign}(x^n)=\text{sign}(x^*)$ when $n>n_0$. Furthermore, by Property 1, for any $i\in I$,

$$x_i^*+\lambda\mu\text{sign}(|x_i^*|)\phi'(|x_i^*|)=x_i^*-\mu[\nabla F(x^*)]_i,\ \text{and}$$
$$x_i^{n+1}+\lambda\mu\text{sign}(|x_i^{n+1}|)\phi'(|x_i^{n+1}|)=x_i^n-\mu[\nabla F(x^n)]_i,$$

when $n>n_0$. Consequently,

$$(x_I^{n+1}-x_I^*)+\lambda\mu(\phi_1(x_I^{n+1})-\phi_1(x_I^*))$$
$$=(x_I^n-x_I^*)-\mu([\nabla F(x^n)]_I-[\nabla F(x^*)]_I),$$

and then

$$\|x_I^{n+1}-x_I^*\|_2^2+\lambda\mu\langle\phi_1(x_I^{n+1})-\phi_1(x_I^*),x_I^{n+1}-x_I^*\rangle=$$
$$\langle x_I^{n+1}-x_I^*,(x_I^n-x_I^*)-\mu([\nabla F(x^n)]_I-[\nabla F(x^*)]_I)\rangle. \tag{48}$$

By (21), the left side of (48) satisfies

$$\|x_I^{n+1}-x_I^*\|_2^2+\lambda\mu\langle\phi_1(x_I^{n+1})-\phi_1(x_I^*),x_I^{n+1}-x_I^*\rangle$$
$$\geq(1-\lambda\mu\alpha_{\phi,\varepsilon})\|x_I^{n+1}-x_I^*\|_2^2,$$

and the right side of (48) satisfies

$$\langle x_I^{n+1}-x_I^*,(x_I^n-x_I^*)-\mu([\nabla F(x^n)]_I-[\nabla F(x^*)]_I)\rangle \le \|x_I^{n+1}-x_I^*\|_2\|(x_I^n-x_I^*)-\mu([\nabla F(x^n)]_I-[\nabla F(x^*)]_I)\|_2.$$

Without loss of generality, we assume that $\|x_I^{n+1}-x_I^*\|_2>0$, otherwise, it demonstrates that IJT algorithm converges to x^* with finitely many iterations. Thus, it becomes

$$(1-\lambda\mu\alpha_{\phi,\varepsilon})\|x_I^{n+1}-x_I^*\|_2 \le \|(x_I^n-x_I^*)-\mu([\nabla F(x^n)]_I-[\nabla F(x^*)]_I)\|_2. \quad (49)$$

Furthermore, by (20), it follows

$$\|(x_I^n-x_I^*)-\mu([\nabla F(x^n)]_I-[\nabla F(x^*)]_I)\|_2^2 \le (1-2\mu\alpha_{F,\varepsilon}+\mu^2L^2)\|x_I^n-x_I^*\|_2^2. \quad (50)$$

Combing (49) and (50), it implies

$$\|x_I^{n+1}-x_I^*\|_2 \le \frac{\sqrt{1-2\mu\alpha_{F,\varepsilon}+\mu^2L^2}}{1-\lambda\mu\alpha_{\phi,\varepsilon}}\|x_I^n-x_I^*\|_2.$$

Let $\rho \triangleq \frac{\sqrt{1-2\mu\alpha_{F,\varepsilon}+\mu^2L^2}}{1-\lambda\mu\alpha_{\phi,\varepsilon}}$. By (45)–(47), it is easy to check that $0<\rho<1$. Thus, when $n>n_0$

$$\|x^{n+1}-x^*\|_2=\|x_I^{n+1}-x_I^*\|_2 \le \rho\|x_I^n-x_I^*\|_2=\rho\|x^n-x^*\|_2. \quad (51)$$

Therefore, we have shown the eventual linear rate of IJT algorithm under conditions (a), (b) and the first part of (c).

Part B: Let $c_1 \triangleq \frac{1-\mu\lambda_{\min}(\nabla_{II}^2F(x^*))}{1+\lambda\mu\phi''(e)}$. By the conditions of Theorem 3, it holds $0<c_1<1$. For any $0<c<1$, let

$$g(c)=\max_{i\in I}\max_{\{x_i:|x_i-x_i^*|<c\eta_\mu\}}\left\{\frac{\lambda\mu|\phi'''(|x_i|)|}{2|1+\lambda\mu\phi''(|x_i^*|)|}\right\}, \quad (52)$$

and

$$c_\epsilon(c)=\frac{1-c_1-\epsilon}{g(c)\eta_\mu}, \quad (53)$$

for some $0<\epsilon<1-c_1$. Since $g(c)$ is non-decreasing with respective to c, and thus $c_\epsilon(c)$ is non-increasing with respect to c. Therefore, there exists a positive constant c^* such that

$$0<c^*<1 \text{ and } c^*<c_\epsilon(c^*). \quad (54)$$

Since $\{x^n\}$ converges to x^*, then there exists an $n^{**}>n^*$, such that for $n>n^{**}$, it holds $\|x^n-x^*\|_2<c^*\eta_\mu$. and by Lemma 3, one holds $I^n=I$ and $\text{sign}(x^n)=\text{sign}(x^*)$. Thus, $\|x^n-x^*\|_2=\|x_I^n-x_I^*\|_2$.

By Property 1, for any $i\in I$,

$$(x_i^n-x_i^*)-\mu([\nabla F(x^n)]_i-[\nabla F(x^*)]_i) = (x_i^{n+1}-x_i^*)+\text{sign}(x_i^*)\lambda\mu(\phi'(|x_i^{n+1}|)-\phi'(|x_i^*|)).$$

By Taylor expansion, for any $i\in I$, there exists an $\xi_i\in(0,1)$, such that

$$\phi'(|x_i^{n+1}|)-\phi'(|x_i^*|)=\text{sign}(x_i^*)\phi''(|x_i^*|)(x_i^{n+1}-x_i^*)+\frac{1}{2}\phi'''(|x_i^\xi|)(x_i^{n+1}-x_i^*)^2,$$

where $x_i^\xi=x_i^*+\xi_i(x_i^{n+1}-x_i^*)$. Let $h^n=x^n-x^*$, then by the above two inequalities, it follows

$$\Lambda_1h_I^{n+1}+\Lambda_2(h_I^{n+1}\odot h_I^{n+1}) = h_I^n-\mu([\nabla F(x^n)]_I-[\nabla F(x^*)]_I), \quad (55)$$

where $\odot$ denotes the Hadamard product or elementwise product, Λ_1 and Λ_2 are two different diagonal matrices with $\Lambda_1(i,i)=1+\lambda\mu\phi''(|x_i^*|), \Lambda_2(i,i)=\frac{1}{2}\text{sign}(x_i^*)\lambda\mu\phi'''(x_i^\xi)$. Moreover, by the twice differentiability of F at x^*, we have

$$[\nabla F(x^n)]_I-[\nabla F(x^*)]_I=\nabla_{II}^2F(x^*)h_I^n+o(\|h_I^n\|_2). \quad (56)$$

Plugging (56) into (55), it becomes

$$\Lambda_1h_I^{n+1}+\Lambda_2(h_I^{n+1}\odot h_I^{n+1}) = (\mathbf{I}-\mu\nabla_{II}^2F(x^*))h_I^n+o(\|h_I^n\|_2),$$

where $\mathbf{I}$ denotes as the identity matrix with the size $|I|\times|I|$ with $|I|$ being the cardinality of the set I. By the conditions of Theorem 3, for any $i\in I$,

$$\Lambda_1(i,i)=1+\lambda\mu\phi''(|x_i^*|)>1-\mu\lambda_{\min}(\nabla_{II}^2F(x^*))\ge 0,$$

thus, Λ_1 is invertible. Then it follows

$$h_I^{n+1}=\Lambda_1^{-1}(\mathbf{I}-\mu\nabla_{II}^2F(x^*))h_I^n-\Lambda_1^{-1}\Lambda_2(h_I^{n+1}\odot h_I^{n+1})+o(\|h_I^n\|_2). \quad (57)$$

By the definition of $o(\|h_I^n\|_2)$, there exists a constant c_ϵ^* (depending on ϵ) such that $|o(\|h_I^n\|_2)|\le\epsilon\|h_I^n\|_2$ when $\|h_I^n\|_2<c_\epsilon^*\eta_\mu$. Thus, we can take $c_0=\min\{c^*,c_\epsilon^*\}<1$ and $n_0>n^{**}$ such that when $n>n_0$, $\|x^n-x^*\|_2<c_0\eta_\mu$. Then (57) implies that

$$\begin{aligned}\|h_I^{n+1}\|_2 &\le \|\Lambda_1^{-1}(I-\mu\nabla_{II}^2F(x^*))h_I^n\|_2+\epsilon\|h_I^n\|_2 \\ &\quad+\|\Lambda_1^{-1}\Lambda_2(h_I^{n+1}\odot h_I^{n+1})\|_2 \\ &\le \left(\frac{1-\mu\lambda_{\min}(\nabla_{II}^2F(x^*))}{1+\lambda\mu\phi''(e)}+\epsilon\right)\|h_I^n\|_2+g(c^*)\|h_I^{n+1}\|_2^2 \\ &\le (c_1+\epsilon)\|h_I^n\|_2+g(c^*)c^*\eta_\mu\|h_I^{n+1}\|_2,\end{aligned}$$

where the second inequality holds by the definition of $g(c^*)$, the facts: $\lambda_{\max}(I-\mu\nabla_{II}^2F(x^*))\le 1-\mu\lambda_{\min}(\nabla_{II}^2F(x^*))$, $\min_{i\in I}|\Lambda_1(i,i)|\ge 1+\lambda\mu\phi''(e)>0$ and $c^*\ge c_0$, and the last inequality holds for $\|h_I^{n+1}\|_2<c^*\eta_\mu$ and the definition of c_1. Furthermore, by (53) and (54), it holds $1-c^*g(c^*)\eta_\mu>c_1+\varepsilon>0$. Therefore, it implies that $\|h_I^{n+1}\|_2\le\frac{c_1+\epsilon}{1-c^*g(c^*)\eta_\mu}\|h_I^n\|_2$, and then

$$\|x^{n+1}-x^*\|_2\le\frac{c_1+\epsilon}{1-c^*g(c^*)\eta_\mu}\|x^n-x^*\|_2.$$

Let $\rho\triangleq\frac{c_1+\epsilon}{1-c^*g(c^*)\eta_\mu}$, then $0<\rho<1$. Thus, the eventual convergence rate of IJT algorithm is linear.

Moreover, the posteriori error bound can be easily derived by the eventual convergence rate and triangle inequality. ■

ACKNOWLEDGMENT

The authors would like to thank two anonymous reviewers, and the associated editor for their constructive and helpful comments.

REFERENCES

[1] M. Allain, J. Idier, and Y. Goussard, "On global and local convergence of half-quadratic algorithms," *IEEE Trans. Image Process.*, vol. 15, no. 5. pp. 1130–1142, May 2006.

[2] H. Attouch, J. Bolte, and B. Svaiter, "Convergence of descent methods for semi-algebraic and tame problems: Proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods," *Math. Program. A*, vol. 137, pp. 91–129, 2013.

[3] T. Blumensath and M. Davies, "Iterative thresholding for sparse approximation," *J. Fourier Anal. Appl.*, vol. 14, no. 5, pp. 629–654, 2008.

[4] J. Bolte, A. Daniilidis, and A. Lewis, "The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems," *SIAM J. Optim.*, vol. 17, no. 4, pp. 1205–1223, 2006.

[5] J. Bolte, A. Daniilidis, A. Lewis, and M. Shiota, "Clarke subgradients of stratifiable functions," *SIAM J. Optim.*, vol. 18, no. 2, pp. 556–572, 2007.

[6] J. Bolte, S. Sabach, and M. Teboulle, "Proximal alternating linearized minimization for nonconvex and nonsmooth problems," *Math. Program. A*, vol. 146, pp. 459–494, 2014.

[7] K. Bredies, D. Lorenz, and S. Reiterer, "Minimization of non-smooth, non-convex functionals by iterative thresholding," *J. Optim. Theory Appl.*, vol. 165, pp. 78–122, 2015.

[8] K. Bredies and D. Lorenz, "Linear convergence of iterative soft-thresholding," *J. Fourier Anal. Appl.*, vol. 14, pp. 813–837, 2008.

[9] E. Candès and T. Tao, "Decoding by linear programming," *IEEE Trans. Inf. Theory*," vol. 51, no. 12, pp. 4203–4215, Dec. 2005.

[10] E. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

[11] E. Candès, M. Wakin, and S. Boyd, "Enhancing sparsity by reweighted ℓ_1 minimization," *J. Fourier Anal. Appl.*, vol. 14, no. 5, pp. 877–905, 2008.

[12] W. Cao, J. Sun, and Z. Xu, "Fast image deconvolution using closed-form thresholding formulas of L_q $(q = 1/2, 2/3)$ regularization," *J. Vis. Commun. Image Represent.*, vol. 24, no. 1, pp. 1529–1542, 2013.

[13] R. Chartrand, "Exact reconstruction of sparse signals via nonconvex minimization," *IEEE Signal Process. Lett.*, vol. 14, no. 10, pp. 707–710, Oct. 2007.

[14] R. Chartrand and V. Staneva, "Restricted isometry properties and nonconvex compressive sensing," *Inverse Problems*, vol. 24, no. 1–14, 2008.

[15] R. Chartrand and W. Yin, "Iterative reweighted algorithms for compressed sensing," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2008, pp. 3869–3872.

[16] X. Chen and W. Zhou, "Convergence of the reweighted ℓ_1 minimization algorithm for ℓ_2-ℓ_p minimization," *Comput. Optim. Appl.*, vol. 59, pp. 47–61, 2014.

[17] X. Chen, F. Xu, and Y. Ye, "Lower bound theory of nonzero entries in solutions of ℓ_2-ℓ_p minimization," *SIAM J. Sci. Comput*.," vol. 32, no. 5, pp. 2832–2852, 2010.

[18] S. Foucart, "Sparse recovery algorithms: Sufficient conditions in terms of restricted isometry constants," in *Approximation Theory XIII: San Antonio.* New York, NY, USA: Springer, 2010, pp. 65–77.

[19] I. Cumming and F. Wong, *Digital Processing of Synthetic Aperture Radar Data: Algorithms and Implementation*. Norwood, MA, USA: Artech House, 2004.

[20] P. Combettes and V. Wajs, "Signal recovery by proximal forward-backward splitting," *Multiscale Model. Simul.*, vol. 4, pp. 1168–1200, 2005.

[21] I. Daubechies, R. Devore, M. Fornasier, and C. Gunturk, "Iteratively reweighted least squares minimization for sparse recovery," *Commun. Pure Appl. Math.*," vol. 63, pp. 1–38, 2010.

[22] I. Duabechies, M. Defrise, and C. Mol, "An iterative thresholding algorithm for linear inverse problems with a sparse constraint," *Commun. Pure Appl. Math.*, vol. 57, pp. 1413–1457, 2004.

[23] D. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*," vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[24] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *J. Amer. Statist. Assoc.*," vol. 96, pp. 1348–1360, 2001.

[25] G. Gasso, A. Rakotomamonjy, and S. Canu, "Recovering sparse signals with a certain family of nonconvex penalties and dc programming," *IEEE Trans. Signal Process.*, vol. 57, no. 12, pp. 4686–4698, Dec. 2009.

[26] D. Geman and G. Reynolds, "Constrained restoration and the recovery of discontinuities," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, no. 3, pp. 367–383, Mar. 1992.

[27] D. Geman and C. Yang, "Nonlinear image recovery with half-quadratic regularization," *IEEE Trans. Image Process.*," vol. 4, no. 7, pp. 932–946, Jul. 1995.

[28] I. Gorodnitsky and B. Rao, "Sparse signal reconstruction from limited data using FOCUSS: A re-weighted minimum norm algorithm," *IEEE Trans. Signal Process.*, vol. 45, no. 3, pp. 600–616, Mar. 1997.

[29] Z. Lu, "Iterative hard thresholding methods for ℓ_0 regularized convex cone programming," *Math. Program.*, vol. 147, pp. 125–154, 2014.

[30] Z. Lu, "Iterative reweighted minimization methods for ℓ_p regularized unconstrained nonlinear programming," *Math. Program. A*," vol. 147, pp. 277–307, 2014.

[31] B. Mordukhovich, *Variational Analysis and Generalized Differentiation. I. Basic Theory*, Berlin, Germany: Springer, 2006, vol. 330.

[32] B. Olshausen and D. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, pp. 607–609, 1996.

[33] J. Ortega and W. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*. New York, NY, USA: Academic, 2000.

[34] A. Ostrowski, "Constributions to the theory of the method of steepest descent," *Archive Rational Mech. Anal.*, vol. 26, pp. 257–280, 1967.

[35] R. Rockafellar and R. Wets, *Variational Anal. Grundlehern der Mathematischen Wissenschaften*. Berlin, Germany: Springer, 1998, vol. 317.

[36] M. Rudelson and R. Vershynin, "On sparse reconstruction from Fourier and Gaussian measurements," *Commun. Pure Appl. Math.*, vol. 61, pp. 1025–1045, 2008.

[37] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Statist. Soc. B*, vol. 58, pp. 267–288, 1996.

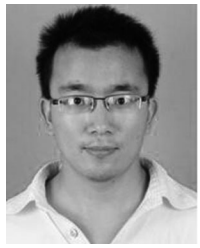
[38] Y. Xu and W. Yin, "A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion," *SIAM J. Imag. Sci.*, vol. 6, no. 3, pp. 1758–1789, 2013.

[39] Z. Xu, X. Chang, F. Xu, and H. Zhang, "$L_{1/2}$ regularization: A thresholding representation theory and a fast solver," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 7, pp. 1013–1027, Jul. 2012.

[40] J. Zeng, J. Fang, and Z. Xu, "Sparse SAR imaging based on $L_{1/2}$ regularization," *Sci. China Series F-Inf. Sci.*, vol. 55, pp. 1755–1775, 2012.

[41] J. Zeng, S. Lin, Y. Wang, and Z. Xu, $L_{1/2}$ Regularization: Convergence of iterative half thresholding algorithm," *IEEE Trans. Signal Process.*," vol. 62, no. 9, pp. 2317–2329, Jul. 2014.

Jinshan Zeng received the Ph.D. degree in applied mathematics from Xi'an Jiaotong University, Xi'an, China, in 2015. He is currently an Assistant Professor in the College of Computer Information Engineering, Jiangxi Normal University, Nanchang, China.



Shaobo Lin received the Ph.D. degree in applied mathematics from Xi'an Jiaotong University, Xi'an, China, in 2014. He is currently an Assistant Professor in the College of Mathematics and Information Science, Wenzhou University, Wenzhou, China.

Zongben Xu received the Ph.D. degree in mathematics from Xi'an Jiaotong University, Xi'an, China, in 1987. He was elected as a Member of Chinese Academy of Science in 2011. His current research interests include intelligent information processing and applied mathematics.