Accepted Manuscript

GAITA: A Gauss-Seidel iterative thresholding algorithm for ℓ_q regularized least squares regression

Jinshan Zeng, Zhiming Peng, Shaobo Lin

PII: S0377-0427(17)30014-6
DOI: http://dx.doi.org/10.1016/j.cam.2017.01.010
Reference: CAM 10975

To appear in: *Journal of Computational and Applied Mathematics*

Received date: 27 June 2016
Revised date: 15 December 2016

Please cite this article as: J. Zeng, Z. Peng, S. Lin, GAITA: A Gauss-Seidel iterative thresholding algorithm for ℓ_q regularized least squares regression, *Journal of Computational and Applied Mathematics* (2017), http://dx.doi.org/10.1016/j.cam.2017.01.010

GAITA: A Gauss-Seidel Iterative Thresholding Algorithm for ℓ_q Regularized Least Squares Regression ☆

Jinshan Zeng[1], Zhiming Peng[2], Shaobo Lin[3] *

1. College of Computer Information Engineering, Jiangxi Normal University, Nanchang, 330022, P R China.
2. Department of Mathematics, University of California, Los Angeles (UCLA), Los Angeles, CA 90095, United States.
3.College of Mathematics and Information Science, Wenzhou University, Wezhou, 325035, P R China

Abstract

This paper studies the $\ell_q(0 < q < 1)$ regularized least squares regression (ℓ_qLS) problem, which arises in many applications of signal processing and machine learning. The iterative thresholding algorithm is an important algorithm for solving the ℓ_qLS problem, and can be viewed as a Jocabi-type iterative method. This paper proposes a Gauss-Seidel version of iterative thresholding algorithm called *GAITA* for solving the ℓ_qLS problem. Under certain conditions, we establish its global convergence[1], eventual linear rate, and the convergence to a local minimizer. Compared to the Jacobi counterpart, the proposed algorithm can allow larger step sizes and converge much faster. The effectiveness of the proposed algorithm is justified with numerical experiments on both synthetic data and real data.

Keywords: Gauss-Seidel, Jacobi, iterative thresholding algorithm, ℓ_q regularized least squares.

1. Introduction

In this paper, we consider the ℓ_q $(0 < q < 1)$ regularized least squares regression (ℓ_qLS) problem

$$\min_{x\in\mathbf{R}^N} f(x) = \frac{1}{2}\|Ax - y\|_2^2 + \lambda\|x\|_q^q, \tag{1.1}$$

where $\|x\|_q^q = \sum_{i=1}^N |x_i|^q$, N is the dimension of x and $\lambda > 0$ is a regularization parameter. The ℓ_qLS problem has attracted lots of attention for its stronger sparsity-promoting ability and

☆The work of J. Zeng is supported in part by the National Natural Science Foundation of China (Grants No. 61603162, 11401462). The work of S. Lin is supported in part by the National Natural Science Foundation of China (Grant Nos. 11501440, 61502342).

*Corresponding author: sblin1983@gmail.com

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

better bias-reduction property compared to the Lasso model [34]. Its applications include signal processing [13, 14], image processing [12, 25], and synthetic aperture radar imaging [41, 42].

One of the most popular method to solve (1.1) is the iterative thresholding algorithm (ITA) [9, 19, 40, 43],

$$x^{n+1} := \mathcal{T}(x^n) \in \mathrm{prox}_{\mu,\lambda\|\cdot\|_q^q}\left(x^n - \mu A^T(Ax^n - y)\right), \tag{1.2}$$

where $\mu > 0$ is a step-size parameter, and the proximity operator is defined by

$$\mathrm{prox}_{\mu,\lambda\|\cdot\|_q^q}(x) := \arg\min_{u\in\mathbf{R}^N} \lambda\|u\|_q^q + \frac{1}{2\mu}\|x-u\|_2^2. \tag{1.3}$$

For $q = \frac{1}{2}$ or $q = \frac{2}{3}$, $\mathrm{prox}_{\mu,\lambda|\cdot|^q}$ has a closed form solution [40]. While for other $q \in (0,1)$, we can use an iterative scheme proposed by [27] to get an approximation. One can view (1.2) as the Jacobi-type iteration, where all the components of x are updated simultaneously in each iteration. On the contrary, the Gauss-Seidel method updates one coordinate in a cyclic fashion. In general, the Gauss-Seidel iteration is faster than the corresponding Jacobi iteration [38], since it uses the latest updates at each iteration and it has more relaxed step size choice [29].

The goal of this paper is to develop a Gauss-Seidel based algorithm to solve the ℓ_qLS problem (1.1). Given the current iterate x^n and a step size μ, at the next iteration, the i-th coefficient is selected cyclically by

$$i = \begin{cases} N, & \text{if } 0 \equiv (n+1) \bmod N, \\ (n+1) \bmod N, & \text{otherwise.} \end{cases} \tag{1.4}$$

Then, we update

$$x_i^{n+1} = \mathcal{T}_i(x^n), \text{ and } x_j^{n+1} = x_j^n, \forall j \neq i, \tag{1.5}$$

where $\mathcal{T}_i(x^n) \in \mathrm{prox}_{\mu,\lambda|\cdot|^q}\left(x_i^n - \mu A_i^T(Ax^n - y)\right)$, x_i^n and A_i represent the i-th component of x^n and i-th column of A, respectively. The GAITA algorithm is summarized as follows.

Algorithm 1: Gauss-Seidel iterative thresholding algorithm (GAITA)

Input : $x^0 \in \mathbf{R}^N$, $N > 0$;

set global iteration counter $n = 0$;

while *stopping condition is not satisfied* **do**

- choose i according to (1.4);
- update $x_i^{n+1} = \mathcal{T}_i(x^n)$ and $x_j^{n+1} = x_j^n$ for $j \neq i$;
- update the global counter $n \leftarrow n+1$;

1.1. Contribution and novelty

Our main contribution is the establishment of the global convergence and convergence rate of the proposed algorithm. We first show the finite support and sign convergence (i.e., the support and sign of any sequence generated by the proposed algorithm converge in finitely many iterations), then justify the global convergence and convergence rate of GAITA. Moreover, we give several sufficient conditions to guarantee the convergence to a local minimizer. The novelties of this paper are summarized as follows.

(1) **Advantages of GAITA.** The proposed GAITA algorithm has several advantages over its Jacobi counterpart, i.e., the iterative jumping thresholding (IJT) algorithm studied in [44]. From Theorem 2.2, GAITA can allow a larger step size than IJT. While as shown by numerical experiments, GAITA converges much faster than IJT since it requires fewer epochs to reach the same precision, where one epoch is defined as updating all of the coordinates once. Moreover, note from Algorithm 1 that the computational cost of one epoch of GAITA equals to that of IJT. As a consequence, the total computational cost of the proposed algorithm is lower than that of IJT.

(2) **Novel techniques.** We introduced several new techniques in our analysis, although the analysis framework of this paper is similar to that in [44].

   (a) *Constructing an alternative sequence using a merging scheme.* The finite support and sign convergence property enable us to construct a new sequence that keeps the same convergence behavior of the original sequence. However, we can not directly take the new sequence as the nonzero part of the original sequence as done in [44], because many successive iterates of GAITA are the same, and thus, the desired sufficient descent and relative error properties of the new sequence will be violated. To overcome this, we introduce a merging scheme by leveraging the periodicity of the update order of GAITA. More detail can be referred to the proof of Lemma 5.4.

   (b) *Deriving the relative error property recursively.* The relative error property is a key property to prove the global convergence of a nonconvex algorithm as shown in [2]. Such desired property is generally derived from the optimality conditions. However, for the proposed algorithm, the optimality conditions only hold for the selected coordinate. To overcome such challenges, we first justify that the optimality conditions hold recursively for the concerned coordinates in a backward order. Summarizing

them by leveraging the periodicity of the iterates gives the desired condition. More details can be found in the proof of Lemma 5.6.

The remainder of this paper is organized as follows. Section 2 presents the main results of this paper. Related works and comparisons are provided in Section 3. In Section 4, a series of simulations are implemented to demonstrate the effectiveness of the proposed algorithm. Section 5 presents all the proofs. Finally, we conclude this paper in Section 6.

2. Main results

2.1. Finite support and sign convergence

Theorem 2.1. *Let $\{x^n\}$ be a sequence generated by GAITA, $I^n = \mathrm{supp}(x^n) \triangleq \{j : x_j^n \neq 0\}$ be the support of x^n, and $L_{\max} = \max_i \|A_i\|_2^2$. Assume that $0 < \mu < \frac{1}{L_{\max}}$, then*

(a) the objective sequence $\{f(x^n)\}$ converges;

(b) $\{x^n\}$ is bounded, and any limit point of $\{x^n\}$ is a stationary point;

(c) there exist a positive integer n_0, an index set I and a sign vector S^* such that $\forall n > n_0$, there hold:
(i) $I^n = I$ and $\mathrm{supp}(x^*) = I$,
(ii) $\mathrm{sign}(x^n) = S^*$ and $\mathrm{sign}(x^*) = S^*, \forall x^* \in \mathcal{L}$, where $\mathcal{L}$ is the set of limit points of $\{x^n\}$.

Theorem 2.1 implies that the supports and signs of the generated sequence $\{x^n\}$ converge in finitely many iterations. The finite support convergence property is crucial in many applications, including but not limited to the sparse signal recovery problem [45] and the feature screening problem [46]. In the sparse signal recovery problem, if the support converges, then the nonzero coefficients can be solved by a least squares problem. While for the feature screening problem, once the support gets freezing, then the main features can be further selected from these features retained in the support.

2.2. Global convergence and rate

Theorem 2.2. *Assume that $0 < \mu < \frac{1}{L_{\max}}$ and the initial iterate x^0 is bounded, then $\{x^n\}$ converges to a stationary point x^*.*

Moreover, let $I = \mathrm{supp}(x^*)$, and $K = \|x^*\|_0$. Assume further

$$A_I^T A_I + \lambda q(q-1)\Lambda(x_I^*) \succ 0, \tag{2.1}$$

where A_I represents the submatrix of A with column restricted to I, x_I^* is the subvector of x restricted to I, $\Lambda(x_I^*) \in \mathbf{R}^{K\times K}$ is a diagonal matrix with $(|x_i^*|^{q-2})_{i\in I}$ as the diagonal vector. Then there exist $n^* \in \mathbf{N}$, $C > 0$, and $\rho \in [0, 1)$ such that $\forall n > n^*$,

$$\|x^n - x^*\|_2 \leq C\rho^n.$$

Moreover, x^* is a strictly local minimizer.

The first part of Theorem 2.2 shows that GAITA converges to a stationary point when the step size is lower than a proper bound. Compared to its Jacobi counterpart [44], the proposed algorithm can adopt a larger range of the step-size parameter. More specifically, the restriction on the step size of IJT for the ℓ_qLS problem is $0 < \mu < \frac{1}{\|A\|_2^2}$, while that of GAITA is $0 < \mu < \frac{1}{\max_i \|A_i\|_2^2}$. The second part of Theorem 2.2 implies that GAITA converges to a strictly local minimizer at an eventual linear rate under certain second-order condition, i.e., the Hessian $\nabla^2 f(x^*)$ restricted to the support I is positive definite.

Furthermore, we can drive another two sufficient conditions for the eventual linear rate as well as convergence to a local minimizer via taking advantage of the specific form of the threshold value (5.1). Let $e = \min_{i\in I} |x_i^*|$. By the fact that $\lambda_{\min}(M_1 + M_2) \geq \lambda_{\min}(M_1) + \lambda_{\min}(M_2)$ for any two square matrices M_1, M_2 with the same sizes, then it is obvious that

$$\lambda_{\min}(A_I^T A_I + \lambda q(q-1)\Lambda(x_I^*)) \geq \lambda_{\min}(A_I^T A_I) + \lambda q(q-1)e^{q-2},$$

where $\lambda_{\min}(M)$ represents the minimal eigenvalue of a given matrix M. Thus, if

$$\lambda_{\min}(A_I^T A_I) > 0 \text{ and } 0 < \lambda < \frac{\lambda_{\min}(A_I^T A_I)e^{2-q}}{q(1-q)}, \tag{2.2}$$

then the condition (2.1) in Theorem 2.2 holds naturally.

Moreover, by Lemma 5.1, it holds

$$e = \min_{i\in I} |x_i^*| \geq \eta_{\lambda\mu,q} = (2\lambda\mu(1-q))^{\frac{1}{2-q}}. \tag{2.3}$$

Hence, if $\frac{\lambda_{\min}(A_I^T A_I)}{\max_i \|A_i\|_2^2} > \frac{q}{2}$ and $\frac{q}{2\lambda_{\min}(A_I^T A_I)} < \mu < \frac{1}{\max_i \|A_i\|_2^2}$, then the condition (2.2) holds and thus (2.1) also holds. According to the above analysis, we can easily obtain the following corollary.

Corollary 2.3. *Let $e = \min_{i\in I} |x_i^*|$. Assume either of the following conditions holds:*
(a) $\lambda_{\min}(A_I^T A_I) > 0, 0 < \lambda < \frac{\lambda_{\min}(A_I^T A_I)e^{2-q}}{q(1-q)}$;
(b) $\frac{\lambda_{\min}(A_I^T A_I)}{\max_i \|A_i\|_2^2} > \frac{q}{2}, \frac{q}{2\lambda_{\min}(A_I^T A_I)} < \mu < \frac{1}{\max_i \|A_i\|_2^2}$.
Then the rate of convergence of GAITA is eventually linear and x^* is a strictly local minimizer.

Intuitively, the condition (a) in Corollary 2.3 means that if the smooth part of the ℓ_qLS problem is strictly convex and the regularization parameter λ is sufficiently small, then the convexity of $f(x)$ at x^* can be guaranteed by the convexity of the smooth part. Suppose that A is column-normalized, i.e., $\|A_i\|_2 = 1$ for any i, then the condition (b) in Corollary 2.3 intuitively implies that if the smooth part of the ℓ_qLS problem is strongly convex, then the local convexity

of $f(x)$ at x^* can be guaranteed if the step size μ is chosen appropriately. Note that condition (a) in Corollary 2.3 is the same as that of IJT applied to the ℓ_qLS problem (see [44, Corollary 3]), while condition (b) in Corollary 2.3 is weaker than that of IJT (see [44, Theorem 4]).

3. Related works and comparisons

There are many methods for solving the ℓ_qLS problem. Some general methods can be found in [2, 3, 8, 10, 11, 15, 18, 21, 44] and references therein. However, those algorithms update the iterate by using the Jacobi but not Gauss-Seidel scheme. In [10], the subsequence convergence of the iterative thresholding algorithm for ℓ_qLS with an arbitrary $q \in (0, 1)$ and further the global convergence for ℓ_qLS with a rational q were verified under the condition $0 < \mu < \|A\|_2^{-2}$. In [2], the global convergence of the iterative thresholding algorithm for ℓ_qLS with an arbitrary q was justified under the same condition. In [44], the IJT algorithm is proposed for solving a class of sparse regularized problem, of which the ℓ_qLS problem is a special case. When applied to the ℓ_qLS problem, IJT converges to a stationary point if $0 < \mu < \|A\|_2^{-2}$. Besides these general methods, there are several specific iterative thresholding algorithms for solving ℓ_qLS with a specific q such as *hard* for ℓ_0 [9], *soft* for ℓ_1 [19] and *half* for $\ell_{1/2}$ [40]. Under the same condition (i.e., $0 < \mu < \|A\|_2^{-2}$), all these specific iterative thresholding algorithms were justified to converge to a stationary point. Compared with theses classical iterative thresholding algorithms, GAITA can adopt a larger step size (i.e., $0 < \mu < \frac{1}{\max_i \|A_i\|_2^2}$) with faster convergence in the sense of fewer epochs required for convergence as shown in the latter Fig. 2.

Note that GIATA is related to the class of block coordinate descent (BCD) algorithms, which have been applied in many applications. The original form of BCD, that is, block coordinate minimization (BCM) can be traced back to the 1950's [23]. The main idea of BCM is to update a block by minimizing the original objective with respect to that block. Its convergence was extensively studied under many different cases (cf. [22], [32], [36], [39] and the references therein). In [26], the convergence rate of BCM was developed under the strong convexity assumption for multi-block case, and in [6], its convergence rate was established under the general convexity assumption for two-block case. Besides BCM, the block coordinate gradient descent (BCGD) method was also numerously studied (cf. [7, 37]). Different from BCM, BCGD updates a block via taking a block gradient step, which is equivalent to minimizing a certain prox-linear approximation of the objective. Its global convergence was justified under the assumptions of the so-called local Lipschitzian error bound and the convexity of the non-differentiable part of

the objective.

One important subclass of BCD is the cyclic coordinate descent (CCD) algorithm. The CCD algorithm updates the iteration cyclically. The work [39] used cyclic update and supposed block-wise convexity. Specifically, the following optimization problem was considered in [39],

$$\min_{\mathbf{x}\in\mathbf{R}^N} f(\mathbf{x}_1,\ldots,\mathbf{x_s}) := g(\mathbf{x}_1,\ldots,\mathbf{x_s}) + \sum_{i=1}^{s} r_i(\mathbf{x}_i), \tag{3.1}$$

where variable $\mathbf{x}$ is decomposed into s blocks $\mathbf{x}_1,\ldots,\mathbf{x}_s$ with $\mathbf{x}_i \in \mathbf{R}^{N_i}, i = 1,\ldots,s$, and $\sum_{i=1}^{s} N_i = N$, function g is assumed to be a differentiable and *block multiconvex*[2] function, and r_i, $i = 1,\ldots,s$, are extended-value *convex* functions. In the perspective of iterative form, GAITA is similar as the BCD algorithm with prox-linear update studied in [39]. When applied to ℓ_qLS, the BCD with prox-linear update becomes:

$$x_i^{n+1} \in \operatorname*{argmin}_{v\in\mathbf{R}} \left\{ \langle A_i^T(Ax^n - y), v - \hat{x}_i^n\rangle + \frac{A_i^T A_i}{2}|v - \hat{x}_i^n|^2 + \lambda|v|^q \right\}, \tag{3.2}$$

where $\hat{x}_i^n = x_i^n + \omega_i^n(x_i^n - x_i^{n-1})$ denotes an extrapolated point, and $\omega_i^n \geq 0$ is the extrapolation weight. While GAITA performs the following form

$$x_i^{n+1} \in \operatorname*{argmin}_{v\in\mathbf{R}} \left\{ \langle A_i^T(Ax^n - y), v - x_i^n\rangle + \frac{1}{2\mu}|v - x_i^n|^2 + \lambda|v|^q \right\}, \tag{3.3}$$

where $\mu > 0$ is a step size. It can be seen from (3.2) and (3.3), if the extrapolation weight ω_i^n is set 0 in (3.2), and μ is taken as $\frac{1}{A_i^T A_i}$ for different coordinates in (3.3), then these two specific updates are the same. In [39], the global convergence and rate of the update (3.2) were justified under certain conditions. The main convergence results of BCD studied in [39] are stated as follows.

Theorem A. (Theorems 2.8 and 2.9 in [39]) *Let $\{x^n\}$ be a sequence generated by BCD algorithm with its parameter ω_i^n satisfying certain condition. Assume the following:*

(i) ∇f is Lipschitz continuous on any bounded set;

(ii) F satisfies the Kurdyka-Łojasiewicz (KL) property (see the latter Definition 3.1) at a finite limit point of $\{x^n\}$, x^*;

(iii) x^0 is sufficiently close to x^* and the objective sequence satisfies $F(x^n) > F(x^*)$ for $n > 0$.

[2] A function g is block multiconvex if for each i, g is a convex function of $\mathbf{x}_i$ while all the other blocks are fixed.

Then $\{x^n\}$ converges to x^*, which is a critical point. Moreover, if F satisfies the KL inequality with $\varphi(s) = cs^{1-\theta}$ for $c > 0$ and $\theta \in [0, 1)$. Then the following hold:

(a) If $\theta = 0$, then $\{x^n\}$ converges to x^* in finitely many iterations.

(b) If $\theta \in (0, 1/2]$, $\|x^n - x^*\| \leq C\rho^n$ for all $n \geq n_0$ for certain $n_0 > 0, C > 0, \rho \in [0, 1)$.

(c) If $\theta \in (1/2, 1)$, $\|x^n - x^*\| \leq Cn^{-(1-\theta)/(2\theta-1)}$ for all $n \geq n_0$ for certain $n_0 > 0, C > 0$.

Definition 3.1. (KL property [2]) *The function $f : \mathbf{R} \to \mathbf{R} \cup \{+\infty\}$ is said to have the Kurdyka-Łojasiewicz property at $x^* \in \mathrm{dom}(\partial f)$ 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) $\varphi(0) = 0$;

(ii) φ is $\mathcal{C}^1$ on $(0, \eta)$;

(iii) for all $s \in (0, \eta)$, $\varphi'(s) > 0$;

(iv) for all x in $U \cap \{x : f(x^*) < f(x) < f(x^*) + \eta\}$, the Kurdyka-Łojasiewicz inequality holds

$$\varphi'(f(x) - f(x^*))\mathrm{dist}(0, \partial f(x)) \geq 1. \tag{3.4}$$

Proper lower semi-continuous functions which satisfy the Kurdyka-Łojasiewicz inequality at each point of $\mathrm{dom}(\partial f)$ are called KL functions.

Intuitively, the main convergence results of BCD studied in [39] are similar as Theorem 2.2 of GAITA. However, there are still several differences between results in [39] and in this paper. The first one is that the regularized term r_i in (3.1) is assumed to be convex to guarantee each subproblem has a unique solution, while the function $|\cdot|^q$ with $q \in (0, 1)$ is nonconvex which leads to generally multi-solutions of the subproblem. The second one is besides the global convergence and rate, the finite support and sign convergence as well as the convergence to a local minimizer of GAITA are also justified in this paper. The last one is that the proof idea of this paper is significantly different with that of [39]. There are mainly two key stones of our proof of the global convergence, that is, the finite support convergence via utilizing the "jumping" property of the thresholding function, and then the global convergence via taking the support and sign convergence as a stepping stone, which will be shown in Section 5 specifically.

Besides [39], there are some specific CCD algorithms for nonconvex penalized least squares regression problems. In [28], a CCD algorithm was proposed for a class of nonconvex penalized least squares problems. However, both [28] and [36] did not consider the CCD algorithm for the ℓ_qLS problem. In [20], a CCD algorithm was implemented for solving the ℓ_1LS problem. Its convergence can be shown by referring to [36]. In [33], the ℓ_0LS-CD algorithm was proposed for the

ℓ_0LS problem, and its convergence to a local minimizer was also shown under certain conditions. Recently, Marjanovic and Solo [27] proposed a cyclic descent algorithm (called ℓ_qCD) for the ℓ_qLS problem with $0 < q < 1$ and A being column-normalized, i.e., $\|A_i\|_2 = 1$, $i = 1, 2, \ldots, N$, where A_i is the i-th column of A. They proved the subsequence convergence and further the convergence to a local minimizer under the so-called scalable restricted isometry property (SRIP) in [27]. In the perspective of the iterative form, ℓ_qCD is a special case of GAITA with A being column-normalized and $\mu = 1$. The main convergence results of ℓ_qCD obtained in [27] are shown as follows.

Theorem B. (Theorems 6 and 7 in [27]) *Let $\{x^n\}$ be a sequence generated by ℓ_qCD algorithm. Then x^n converges to a closed and connected set of coordinate-wise minimizers of $f(x)$. Furthermore, if A satisfies the so-called Scalable Restricted Isometry Property (SRIP, see Definition 3.2), i.e.,* SRIP(p, ϕ, α) *with some $p \geq 0$. Then for any $0 < q < \min\{1, 2/\alpha^2\}$, x^n converges to a local minimizer.*

From Theorem B, only subsequence convergence is claimed for ℓ_qCD. While the global convergence of GAITA is guaranteed as long as $0 < \mu < 1$ since in this case $L_{\max} = 1$ when A is normalized in column. The convergence to a local minimizer of ℓ_qCD algorithm is justified under certain SRIP condition of A, which is defined as follows.

Definition 3.2. (SRIP [27]). *We say A has the SRIP(p, ϕ, α) if there exist $\nu_\phi, \gamma_\phi > 0$ satisfying $\gamma_\phi/\nu_\phi < \alpha$ such that*

$$\nu_\phi \|x\|_2 \leq \|Ax\|_2 \leq \gamma_\phi \|x\|_2$$

holds for every $x \in B_p(\phi) := \{x : \|x\|_p^p \leq \phi\}$, and $\|\cdot\|_p^p := \|\cdot\|_0$ for $p = 0$.

Roughly speaking, ν_ϕ and γ_ϕ can be viewed as some type of the minimal and maximal singular values of A, respectively. Thus, SRIP essentially indicates that A possesses a good condition number. With the definition of SRIP, [27] demonstrates that if A has the SRIP(p, ϕ, α) with some $p \geq 0$, then for any $0 < q < q^*$ (where $q^* := \min\{1, 2/\alpha^2\}$), the ℓ_qCD algorithm converges to a local minimizer. Particularly, when $\alpha = \sqrt{2}$, that is, $\gamma_\phi/\nu_\phi < \sqrt{2}$, then the ℓ_qCD algorithm converges to a local minimizer for any $0 < q < 1$. In other words, if

$$0 < q < \min\left\{1, \frac{2\nu_\phi^2}{\gamma_\phi^2}\right\}, \tag{3.5}$$

then the ℓ_qCD algorithm converges to a local minimizer. It can be seen from Corollary 2.3 that the condition (b) is equivalent to

$$0 < q < \min\left\{1, \frac{2\lambda_{\min}(A_I^T A_I)}{\max_i \|A_i\|_2^2}\right\}. \tag{3.6}$$

It is generally hard to compare the conditions (3.5) and (3.6) directly. However, if $p = 0$, then SRIP may reduce to the standard restricted isometry property (RIP), and in this case, if further $\phi = K$ (where K is the cardinality of the support of x^*), then

$$\lambda_{\min}(A_I^T A_I) \geq \nu_K^2, \text{ and } \max_i \|A_i\|_2^2 \leq \gamma_K^2.$$

Therefore,

$$\frac{\lambda_{\min}(A_I^T A_I)}{\max_i \|A_i\|_2^2} \geq \frac{\nu_K^2}{\gamma_K^2},$$

which implies that the conditions of GAITA for convergence to a local minimizer are generally weaker than that of the ℓ_qCD algorithm in terms of the SRIP. Besides the global convergence and convergence to a local minimizer, we also show the finite support convergence and rate of GAITA in this paper.

In addition, there is another algorithm called proximal alternating linearized minimization (PALM) [7] for solving a class of nonconvex-nonsmooth problem of the form

$$\underset{x\in\mathbf{R}^N, y\in\mathbf{R}^M}{\text{minimize}} \ \Psi(x,y) := f(x) + g(y) + H(x,y) \tag{3.7}$$

where f, g are proper and lower semicontinuous functions with well-definedness of proximal maps, H is a continuously differentiable function and satisfies the following *block-wise Lipschitz differentiable*, that is,

$$\|\nabla_x H(x_1,y) - \nabla_x H(x_2,y)\|_2 \leq L_1(y)\|x_1 - x_2\|_2, \forall x_1, x_2 \in \mathbf{R}^N,$$
$$\|\nabla_y H(x,y_1) - \nabla_x H(x,y_2)\|_2 \leq L_1(x)\|y_1 - y_2\|_2, \forall y_1, y_2 \in \mathbf{R}^N,$$

for some $L_1(y) > 0$, $L_2(x) > 0$. The PALM for (3.7) can be described as follows:

$$x^{k+1} \in \text{prox}_{c_k,f}(x^k - c_k \nabla_x H(x^k, y^k)),$$
$$y^{k+1} \in \text{prox}_{d_k,g}(y^k - d_k \nabla_y H(x^{k+1}, y^k)),$$

where $c_k = \gamma_1 / L_1(y^k)$ and $d_k = \gamma_2 / L_2(x^{k+1})$ for some $\gamma_1, \gamma_2 \in (0,1)$. The main convergence results is listed as follows.

Theorem C. (Theorem 1 in [7]) *Suppose that Ψ satisfies certain continuous conditions described above and is a KL function. Let $\{z^k := (x^k, y^k)\}_{k\in\mathbf{N}}$ be a sequence generated by PALM which is assumed to be bounded. The following assertions hold.*

(i) *The sequence $\{z^k\}_{k\in\mathbf{N}}$ has finite length, that is, $\sum_{k=1}^{\infty} \|z^{k+1} - z^k\| < \infty$.*

(ii) *The sequence* $\{z^k\}_{k\in\mathbf{N}}$ *converges to a critical point* $z^* = (x^*, y^*)$ *of* Ψ.

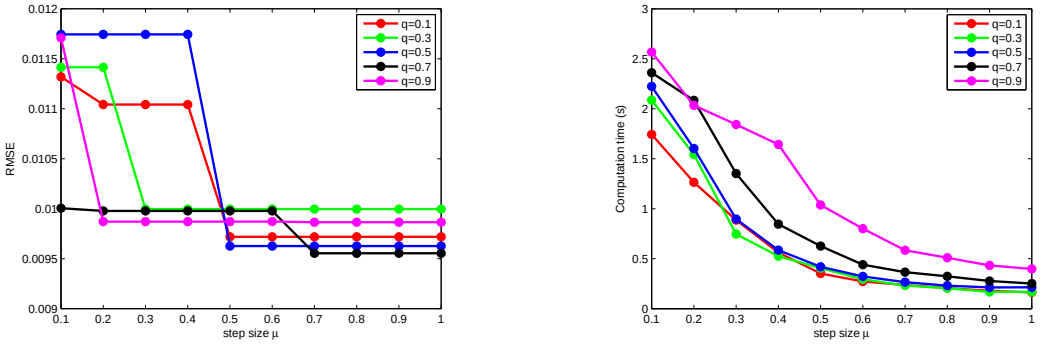
It was pointed out in [7] that PALM as well as its convergence result can be extended to p blocks ($p \geq 2$). When appled to ℓ_qLS problem with N blocks, i.e., each block contains one coordinate, PALM adopts different step sizes with respect to different coordinates while GAITA proposed in this paper adopts a uniform step size to all coordinates. Except the minor difference of the iterative form, the global convergence results of these two algorithms are almost the same. However, besides the global convergence, we also show the finite support and sign convergence and convergence to a local minimizer of GAITA in this paper. More importantly, the proof ideas of PALM and GAITA are significantly different. More specifically, we mainly use two key techniques to get the global convergence of GAITA, that is, support and sign convergence based on the "jumping" property of the proximity operator, and global convergence via taking the support and sign convergence as a step stone, as shown in Section 5. While the proof idea of the global convergence of PALM in [7] is firstly establishing the descent and bounded subgradient lemmas, and then utilizing the KL property of the objective function.

4. Numerical experiments

In this section, we conducted a series of sparse signal recovery experiments to show the effectiveness and efficiency of GAITA. We generate the measurement matrix $A \in \mathbf{R}^{250\times 500}$ with Gaussian $\mathcal{N}(0, 1/250)$ i.i.d. entries. The true sparse signal x_{tr} has 15 nonzeros, and the underdetermined observation is generated by $y = Ax_{tr} + \epsilon$, where ϵ is the noise. In all test cases, the regularization parameters λ were hand-tuned, and $x^0 = 0$ for all tests.

4.1. On effect of μ

In this section, we evaluate the performance of GAITA with different step size choices. The columns of A are normalized, i.e., $\|A_i\|_2 = 1$ for any i. The observation y was corrupted with 30 dB of noise. With these settings, the convergence condition of GAITA becomes $0 < \mu < 1$. We varied μ from 0 to 1, and considered different q, that is, $q = 0.1, 0.3, 0.5, 0.7, 0.9$. The stopping criteria is $\frac{\|x^n - x_{tr}\|_2}{\|x_{tr}\|_2} < 10^{-2}$. The experiment results are shown in Fig. 1. We can observe from Fig. 1 that the step size parameter μ has almost no influence on the recovery quality of the proposed algorithm (as shown in Fig. 1(a)) while it significantly affects the time efficiency of the proposed algorithm (as shown in Fig. 1(b)). Generally speaking, larger step size gives faster convergence.



(a) Recovery Error (b) Computational Time

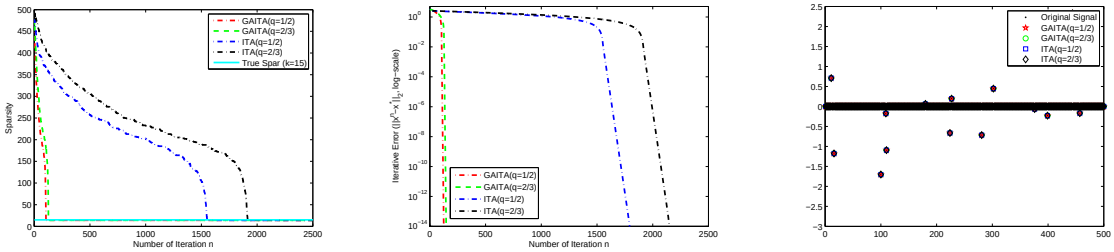
Figure 1: Experiment for the justification of the effect of the step size parameter μ on the performance of GAITA with different q. (a) The trends of recovery error of GAITA with different q. (b) The trends of the computational time of GAITA with different q.

4.2. Comparison with ITA

4.2.1. Faster convergence

In this section, we applied GAITA to the ℓ_qLS problems with $q = 1/2$, and $q = 2/3$. In both cases, the thresholding operators have closed form representations [40, 12], and thus the corresponding GAITA and ITA algorithms can be efficiently implemented. In all cases, the step size parameters μ were set as $\frac{0.95}{\max_i \|A_i\|_2^2}$ for GAITA and $0.99\|A\|_2^{-2}$ for ITA. We counted every N inner iterations of GAITA as one "completed" iteration because the computation cost of every N iterations of GAITA equals to that of one iteration of ITA [29]. The experiment results are reported in Fig. 2.

By Fig. 2(a), the sparsity levels of the iterates converge to the true sparsity level in finitely many iterations for both GAITA and ITA. The number of iterations for identifying the support of x by GAITA is significantly smaller than that of ITA. Fig. 2(b) shows the eventual linear convergence of GAITA and ITA. We can observe that GAITA converges much faster than ITA. Moreover, it can be observed from Fig. 2(a) and (b) that GAITA starts the linear decay until the support converges. Fig. 2(c) shows that the true sparse signal can be recovered with high accuracy in all cases. All these experiment results not only verify the theoretical assertions of this paper but also show the faster convergence of GAITA than ITA.

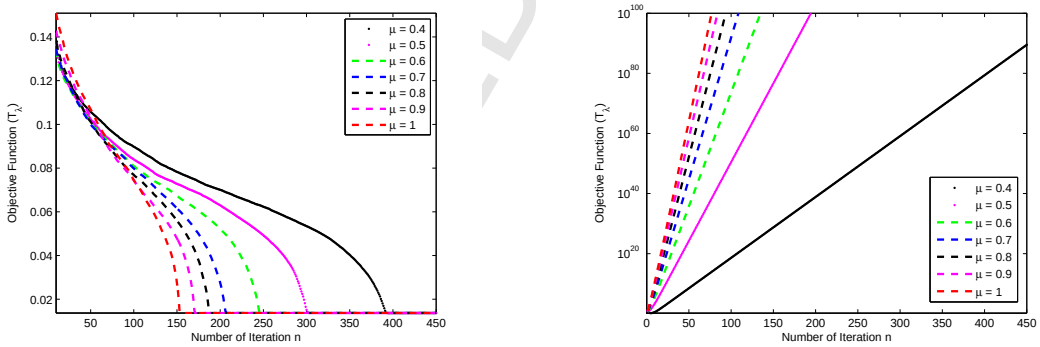


(a) Support convergence (b) Eventual linear Rate (c) Recovery signal

Figure 2: Experiment for faster convergence. (a) Trends of sparsity level. (b) 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. (c) Recovery results. The recovery MSEs of the four cases, that is, GAITA ($q = 1/2$), GAITA ($q = 2/3$), ITA ($q = 1/2$) and ITA ($q = 2/3$) are 1.63×10^{-5}, 2.35×10^{-5}, 6.23×10^{-5} and 7.52×10^{-5}, respectively.

4.2.2. Weaker convergence condition

In this section, we show that GAITA can adopt a larger step size than ITA. We applied GAITA and ITA to the ℓ_qLS problem with $q = 1/2$ and A being column-normalized. In this setting, the theoretical condition for convergence of ITA is $\mu \in (0, 0.1759)$ while the associated condition of GAITA is $\mu \in (0, 1)$. We used different μ (i.e., $\mu = 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1$) for both GAITA and ITA. The figures of the objective sequences are shown in Fig. 3.



(a) Convergence of GAITA (b) Divergence of ITA

Figure 3: An experiment that verifies the weaker convergence condition of GAITA compared with ITA. (a) The trends of the objective sequences, i.e., $\{f(x^n)\}$ of GAITA with different μ. (b) The trends of the objective sequences of ITA with different μ.

Fig. 3(a) shows that the objective sequences of GAITA converge for all selected μ, however, the objective sequences of ITA diverge for all selected μ as shown in Fig. 3(b). When $\mu = 1$ and A is column-normalized, GAITA is reduced to the ℓ_qCD method. Fig. 3(a) shows the

objective sequence of the ℓ_qCD method is convergent, which can be actually guaranteed by Lemma 5.2. However, different from GAITA, the global convergence of the ℓ_qCD method has not been justified if there is no additional condition.

4.2.3. Sparse recovery performance

We implement a series of numerical experiments to compare the sparse recovery performance of GAITA and ITA for two different ℓ_q regularization problems with $q = 1/2$ and $2/3$. For this purpose, we consider the signal recovery problem through the observation $y = Ax^* + \varepsilon$, where the measurement matrix $A \in \mathbf{R}^{250\times 500}$ is generated with i.i.d. Gaussian entries without column-normalization, the original sparse signal $x^* \in \mathbf{R}^{500}$ has k nonzero components, and ε is the Gaussian noise. The measurement y is added with 40 dB noise. We let k be the decile points of the interval between 0 and 100. For each k, we implement 50 experiments independently and record their recovery errors, i.e., $\frac{\|\hat{x}-x^*\|_\infty}{\|x^*\|_\infty}$, where $\hat{x}$ is the recovered signal. Then we take the average of recovery errors of these 50 independent experiments as the recovery error of this case. Moreover, we record the successful probability of recovery in the sense that if the recovery error $\frac{\|\hat{x}-x^*\|_\infty}{\|x^*\|_\infty} < 10^{-2}$, then we count it as a successful case. The recovery probability is taken as the ratio of the number of successful cases to the total number of experiments. The experiment results are illustrated in Fig. 4.

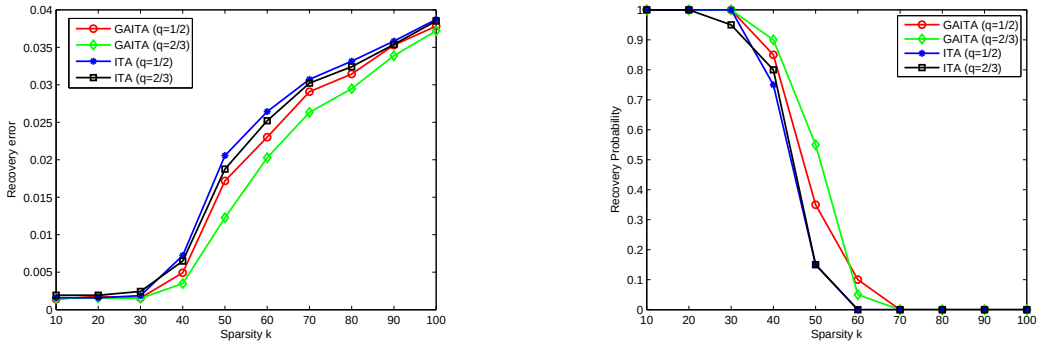
From Fig. 4, the performance of GAITA is slightly better than that of ITA in terms of both recovery error and probability. By Fig. 4, when the sparsity k is smaller than 40, both GAITA and ITA can recover the sparse signal with high probabilities. Once k increases to 50, the recovery probabilities of both GAITA and ITA will decay sharply to a small value. While when k is larger than 60, then both GAITA and ITA are failed to recover the sparse signal.

4.3. Sparse synthetic aperture radar imaging

Synthetic aperture radar (SAR) imaging is a typical inverse problem. According to [41], the sparse SAR imaging can be reformulated as the following ℓ_qLS model

$$\min_{x\in\mathbf{C}^N}\left\{\frac{1}{2}\|y-\Phi x\|_2^2+\lambda\|x\|_q^q\right\}, \tag{4.1}$$

where $\Phi = A\psi$, A is the SAR observation matrix determined by SAR acquisition system and observation geometry, Ψ is a sparse basis, $f = \Psi x$ is the reflectivity coefficient vector of the target scene.



(a) Trend of recovery error

(b) Trend of recovery probability

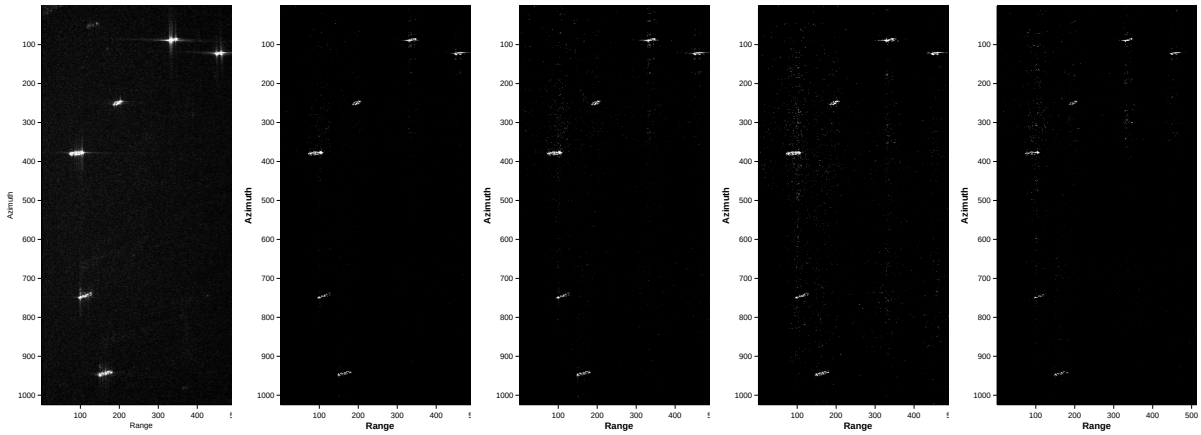
Figure 4: Experiments for comparing the recovery performance of GAITA with that of ITA. (a) The trends of recovery errors. (b) The trends of recovery probabilities.

We implement the experiments on RADARSAT-1 data. The detailed target and data descriptions can be found in [17]. We focus on the region of the English Bay, because this region is a typical sparse scene under the identity basis. There are six sitting vessels in this region. The main radar parameters are shown as follows: the signal bandwidth is 30.111 MHz, the pulse repetition frequency is 1256.98 Hz. We apply GAITA to solve the sparse SAR imaging problem with $q = 1/2$. Moreover, we compare the performance of GAITA with those of the traditional SAR imaging method, and some other state-of-the-art sparse SAR algorithms including ITA (i.e., *half* [40] when $q = 1/2$), fast iterative shrinkage-thresholding algorithm (FISTA) [4] and orthogonal matching pursuit (OMP) [35]. The experiment results are shown in Fig. 5.

Compared to the traditional SAR method, GAITA can reconstruct higher quality images (higher resolution and lower sidelobes) at much lower sampling rate than the Nyquist rate. While compared to the other sparse SAR imaging methods, GAITA can generally reconstruct relatively higher quality images (fewer artifacts) under the same sampling rates.

5. Proofs

In this section, we first present two propositions, and then give the proofs for the main theorems.



(a) RDA (100%) (b) GAITA (5%) (c) ITA (5%) (d) FISTA (5%) (e) OMP (5%)

Figure 5: RADARSAT-1 data imaging results via the traditional SAR imaging method (Range Doppler algorithm (RDA) [17]) using full sampling, and sparse SAR imaging methods using undersampling. (a) RDA under the full sampling data. (b)-(e) GAITA, ITA, FISTA and OMP under the sampling rate 5%, respectively. The details are better seen by zooming on a computer scene.

5.1. Support and sign convergence

According to [10], $\mathrm{prox}_{\mu,\lambda|\cdot|^q}$ can be expressed as

$$\mathrm{prox}_{\mu,\lambda|\cdot|^q}(z) = \begin{cases} (\cdot + \lambda\mu q \cdot \mathrm{sign}(\cdot)|\cdot|^{q-1})^{-1}(z), & \text{for } |z| \geq \tau_{\lambda\mu,q}, \\ 0, & \text{for } |z| < \tau_{\lambda\mu,q}, \end{cases} \tag{5.1}$$

where

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

and the range of $\mathrm{prox}_{\mu,\lambda|\cdot|^q}$ is $\{0\} \cup [\eta_{\lambda\mu,q}, \infty)$, $\mathrm{sign}(\cdot)$ represents the sign function henceforth. When $|z| \geq \tau_{\lambda\mu,q}$, the relation $\mathrm{prox}_{\mu,\lambda|\cdot|^q}(z) = (\cdot + \lambda\mu q \mathrm{sign}(\cdot)|\cdot|^{q-1})^{-1}(z)$ means that $\mathrm{prox}_{\mu,\lambda|\cdot|^q}(z)$ satisfies the equation $v + \lambda\mu q \cdot \mathrm{sign}(v)|v|^{q-1} = z$. It can be observed from (5.1) that $\mathrm{prox}_{\mu,\lambda|\cdot|^q}$ is discontinuous at $z = \pm\tau_{\lambda\mu,q}$. Such property is called the *jumping* property of the proximity operator in [44]. Based on the jumping property, the following *alternative* property (Lemma 5.1) holds for any sequence $\{x^n\}$ generated by GAITA. Besides the alterative property, the *sufficient descent* property (Lemma 5.2) also holds for $\{x^n\}$ by the definition of proximity operator.

Lemma 5.1. (Alternative property of $\{x^n\}$) *Given the current iterate x^n, let i be determined by (1.4), then x_i^{n+1} satisfies either (a) $x_i^{n+1} = 0$; or, (b) $|x_i^{n+1}| \geq \eta_{\lambda\mu,q}$ and also satisfies*

$$A_i^T(Ax^{n+1} - y) + \lambda q \mathrm{sign}(x_i^{n+1})|x_i^{n+1}|^{q-1} = (1/\mu - A_i^T A_i)(x_i^n - x_i^{n+1}). \tag{5.2}$$

That is, $\nabla_i f(x^{n+1}) = (1/\mu - A_i^T A_i)(x_i^n - x_i^{n+1})$, *where* $\nabla_i f(x^{n+1})$ *represents the* i*-th component of* $\nabla f(x^{n+1})$.

Proof. According to (5.1) and (1.5), we know that $x_i^{n+1} = 0$ or $|x_i^{n+1}| \geq \eta_{\lambda\mu,q}$. When $|x_i^{n+1}| \geq \eta_{\lambda\mu,q}$, x_i^{n+1} satisfies the following optimality condition

$$x_i^{n+1} - z_i^n + \lambda\mu q \mathrm{sign}(x_i^{n+1})|x_i^{n+1}|^{q-1} = 0, \tag{5.3}$$

where $z_i^n = x_i^n - \mu A_i^T(Ax^n - y)$. Then (5.3) gives $A_i^T(Ax^{n+1} - y) + \lambda q \mathrm{sign}(x_i^{n+1})|x_i^{n+1}|^{q-1} = \frac{1}{\mu}(x_i^n - x_i^{n+1}) - A_i^T A(x^n - x^{n+1})$. Note that $x_j^{n+1} = x_j^n$ for any $j \neq i$, hence, (5.2) holds. ■

Lemma 5.2. (Sufficient descent of $\{x^n\}$) *Let* $\{x^n\}$ *be a sequence generated by GAITA. If* $0 < \mu < L_{\max}^{-1}$, *then*

$$f(x^n) - f(x^{n+1}) \geq a\|x^n - x^{n+1}\|_2^2, \ \forall n \in \mathbf{N},$$

where $a \triangleq \frac{1}{2}(1/\mu - L_{\max})$ *with* $L_{\max} = \max_i \|A_i\|_2^2$.

Proof. Let i be determined via (1.4). According to (1.3) and (1.5),

$$x_i^{n+1} \in \arg\min_{v \in \mathbf{R}} \left\{ \frac{|z_i^n - v|^2}{2} + \lambda\mu|v|^q \right\},$$

where $z_i^n = x_i^n - \mu A_i^T(Ax^n - y)$. Then it implies

$$\frac{1}{2}|\mu A_i^T(Ax^n - y)|^2 + \lambda\mu|x_i^n|^q \geq \frac{1}{2}|(x_i^{n+1} - x_i^n) + \mu A_i^T(Ax^n - y)|^2 + \lambda\mu|x_i^{n+1}|^q.$$

Some simplifications give

$$\lambda|x_i^n|^q - \lambda|x_i^{n+1}|^q \geq \frac{|x_i^{n+1} - x_i^n|^2}{2\mu} + A_i^T(Ax^n - y)(x_i^{n+1} - x_i^n). \tag{5.4}$$

Moreover, since $x_j^{n+1} = x_j^n$ for any $j \neq i$, (5.4) becomes

$$\lambda\|x^n\|_q^q - \lambda\|x^{n+1}\|_q^q \geq \frac{\|x^{n+1} - x^n\|^2}{2\mu} + \langle Ax^n - y, A(x^{n+1} - x^n)\rangle. \tag{5.5}$$

Adding $\frac{1}{2}\|Ax^n - y\|_2^2 - \frac{1}{2}\|Ax^{n+1} - y\|_2^2$ to both sides of (5.5) gives

$$f(x^n) - f(x^{n+1}) \geq \frac{\|x^{n+1} - x^n\|^2}{2\mu} - \frac{1}{2}\|A(x^n - x^{n+1})\|_2^2 \geq \frac{1}{2}(1/\mu - L_{\max})\|x^n - x^{n+1}\|_2^2, \tag{5.6}$$

where the first equality holds for $\|A(x^n - x^{n+1})\|_2^2 = (A_i^T A_i)|x_i^n - x_i^{n+1}|^2 = (A_i^T A_i)\|x^n - x^{n+1}\|_2^2$, and the second inequality holds for $A_i^T A_i \leq L_{\max}$ and also $x_j^{n+1} = x_j^n$ for any $j \neq i$. ■

Based on these two properties, we can obtain the finite support and sign convergence.

Proposition 5.3. (Finite support and sign convergence) *Let* $\{x^n\}$ *be a sequence generated by GAITA with a finite initial point. Suppose that* $0 < \mu < 1/L_{\max}$. *Then,*

(a) $\{f(x^n)\}$ *converges to some value* f^**;*

(b) $\{x^n\}$ *is bounded;*

(c) there exist a positive integer n_0*, an index set* I *and a sign vector* S^* *such that when* $n > n_0$*, there holds*

$$I^n = I \text{ and } \operatorname{sign}(x^n) = S^*,$$

where $\operatorname{sign}(x^n)$ *represents the sign vector of* x^n*.*

Proof. By Lemma 5.2, $f(x^n)$ is monotonically decreasing. Then $\{f(x^n)\}$ converges since it is also lower bounded by 0. Denote $f^* = \lim_{n\to\infty} f(x^n)$.

From Lemma 5.2, we have $f(x^n) \leq f(x^0) < \infty$ for any $n \in \mathbf{N}$. Then $\{x^n\}$ is bounded by the coercivity of $f(x)$. Thus, $\{x^n\}$ has a convergent subsequence.

Combing Lemma 5.2 and the convergence of $\{f(x^n)\}$ gives the following

$$\sum_{n=0}^{\infty} \|x^{n+1} - x^n\|_2^2 \leq \frac{f(x^0)}{a}, \tag{5.7}$$

$$\|x^n - x^{n+1}\|_2 \to 0, \text{ as } n \to +\infty, \tag{5.8}$$

where (5.8) implies that there exists a sufficiently large positive integer n_0 such that $\|x^n - x^{n+1}\|_2 < \eta_{\lambda\mu,q}$ when $n > n_0$. Based on (5.7)-(5.8), and Lemma 5.2 and Lemma 5.1, we can prove the finite convergence of support and sign similarly to that of [44, Lemma 3]. ■

5.2. From support and sign convergence to global convergence

From Proposition 5.3, besides the finite convergence of support and sign, the subsequence convergence can be claimed by the boundedness of $\{x^n\}$. Upon the subsequence convergence, there are generally two ways to get the global convergence. One is by verifying that a limit point is an isolated point (e.g., [31]). The other one is by developing sufficient descent and relative error properties, and then assuming the KL property of the objective function at the limit point (e.g., [2]) to establish the convergence. However, for the first one, it is generally difficult to verify that a limit point is isolated. Actually, in many nonconvex cases, the limit point set may be possibly a continuum (a nonempty compact connected set). While for the second one, although the sufficient descent property and the KL property of the objective function have been shown by Lemma 5.2 and [2] (p. 122), respectively, the relative error property, i.e., there exists a constant $b > 0$ such that

$$\|g^{n+1}\|_2 \leq b\|x^{n+1} - x^n\|_2, \quad \text{where } g^{n+1} \in \partial f(x^{n+1}),$$

does not generally hold for GAITA, since x^{n+1} may equal to x^n for many n's, and in these cases, g^{n+1} are commonly not equal to zero. Specifically, let $i = N - ((n+1) \bmod N)$, then $x_j^{n+1} = x_j^n$ for any $j \neq i$, and also it is highly possible that $x_i^{n+1} = x_i^n = 0$. Thus, in this case, $x^{n+1} = x^n$. Fortunately, as shown in the following lemmas, it is possible to construct an auxiliary sequence $\{u^n\}$ which not only keeps the convergence behavior of $\{x^n\}$ but also possesses the sufficient descent and relative error properties.

Lemma 5.4. (Construction of $\{u^n\}$) *Under the conditions of Proposition 5.3, there exists a sequence $\{u^n\}$ which has the same convergence behavior of $\{x^n\}$.*

Proof. According to Proposition 5.3, for sufficiently large $n > n_0$, the supports and signs of $\{x^n\}$ are the same. Let $n^* = j_0 N > n_0$ for some positive integer j_0. Then we can define a new sequence $\{\hat{x}^n\}$ with $\hat{x}^n = x^{n^*+n}$ for $n \in \mathbf{N}$. By Proposition 5.3, all the supports and signs of $\{\hat{x}^n\}$ are the same. Let K be the cardinality of I. Without loss of generality, we assume $1 \leq I(1) < I(2) < \cdots < I(K) \leq N$. By the update rule (1.4)-(1.5) of GAITA, we can observe that many successive iterates of $\{\hat{x}^n\}$ are the same. Thus, we can merge these successive iterates into a single iterate. Moreover, the update of index is cyclic and thus periodic. As a consequence, the merging procedure can be repeated periodically. Formally, we consider such a periodic subsequence with N-length of $\{\hat{x}^n\}$, i.e.,

$$\{\hat{x}^{jN+I(1)}, \hat{x}^{jN+I(1)+1}, \cdots, \hat{x}^{jN+I(1)+N-1}\}$$

for $j \in \mathbf{N}$. Then for any $j \in \mathbf{N}$, we merge the N-length sequence $\{\hat{x}^{jN+I(1)}, \cdots, \hat{x}^{jN+I(1)+N-1}\}$ into a new K-length sequence $\{\bar{x}^{jK+1}, \bar{x}^{jK+2}, \cdots, \bar{x}^{jK+K}\}$ according to the rule

$$\{\hat{x}^{jN+I(i)}, \cdots, \hat{x}^{jN+I(i+1)-1}\} \mapsto \bar{x}^{jK+i},$$

with $\bar{x}^{jK+i} = \hat{x}^{jN+I(i)}$ for $i = 1, 2, \cdots, K$, since $\hat{x}^{jN+I(i)+k} = \hat{x}^{jN+I(i)}$ for $k = 1, \cdots, I(i+1) - I(i) - 1$. Moreover, we merge the first $I(1)$ iterates of $\{\hat{x}^n\}$ into $\bar{x}^0$, i.e.,

$$\{\hat{x}^0, \cdots, \hat{x}^{I(1)-1}\} \mapsto \bar{x}^0,$$

with $\bar{x}^0 = \hat{x}^0$, since these iterates keep invariant and are equal to $\hat{x}^0$. After this procedure, we obtain a new sequence $\{\bar{x}^n\}$ with $n = jK + i$, $i = 0, \cdots, K-1$ and $j \in \mathbf{N}$. It can be observed that such a merging procedure keeps the convergence behavior of $\{\bar{x}^n\}$ the same as that of $\{\hat{x}^n\}$ and thus, $\{x^n\}$. Furthermore, we define $\{u^n\}$ with $u^n = \bar{x}_I^n$, for $n \in \mathbf{N}$. As we can observe that

u^n keeps all the non-zero elements of $\bar{x}^n$ while getting rid of its zero elements. Therefore, $\{u^n\}$ keeps the same convergence behavior of $\{x^n\}$. ■

Define operators $P_I : \mathbf{R}^N \to \mathbf{R}^K, P_I x = x_I, \forall x \in \mathbf{R}^N$, and P_I^T as $P_I^T : \mathbf{R}^K \to \mathbf{R}^N, (P_I^T u)_I = u, (P_I^T u)_{I^c} = 0, \forall u \in \mathbf{R}^K$. Here I^c represents the complementary set of I, i.e., $I^c = \{1, 2, \cdots, N\} \backslash I$, $(P_I^T u)_I$ and $(P_I^T u)_{I^c}$ represent the subvectors of $P_I^T u$ restricted to I and I^c, respectively. Then we define a new function T as follows:

$$T : \mathbf{R}^K \to \mathbf{R}, T(u) = f(P_I^T u) = \frac{1}{2}\|Bu - y\|_2^2 + \lambda \|u\|_q^q, \forall u \in \mathbf{R}^K, \tag{5.9}$$

where $B = A_I$, and according to [2] (p. 122), T is a KL function with $\varphi(s) = cs^{1-\theta}$ for $c > 0$ and some $\theta \in [0, 1)$. Let $\{u^n\}$ be the sequence constructed in Lemma 5.4. Then $\{u^n\}$ possesses the following properties.

Lemma 5.5. (Properties of $\{u^n\}$) *The sequence $\{u^n\}$ possesses the following properties:*

(a) Given the current iterate u^n, u^{n+1} is performed as

$$u_i^{n+1} = \hat{\mathcal{T}}_i(u^n), \text{ and } u_j^{n+1} = u_j^n, \text{ for } j \neq i, \tag{5.10}$$

where $\hat{\mathcal{T}}_i(x^n) \in prox_{\mu,\lambda|\cdot|^q}\left(u_i^n - \mu B_i^T(Bu^n - y)\right)$ and

$$i = \begin{cases} K, & \text{if } 0 \equiv (n+1) \bmod K \\ (n+1) \bmod K, & \text{otherwise} \end{cases}. \tag{5.11}$$

(b) According to (5.10), for $n \geq K$, there exist two integers $1 \leq i_0 \leq K$ and $j_0 \geq 1$ such that $n = j_0 K + i_0$ and

$$u_j^n = \begin{cases} u_j^{n-(i_0-j)}, & \text{if } 1 \leq j \leq i_0 \\ u_j^{n-K-(i_0-j)}, & \text{if } i_0 + 1 \leq j \leq K \end{cases}. \tag{5.12}$$

(c) **(away from zero)** *$u^n \in \mathbf{R}^K_{\eta_{\lambda\mu,q^c}}$, where $\mathbf{R}_{\eta_{\lambda\mu,q^c}} \triangleq \mathbf{R} \setminus (-\eta_{\lambda\mu,q}, \eta_{\lambda\mu,q})$.*

(d) Given u^n, if i is determined by (5.11), then u_i^{n+1} satisfies the following equation

$$B_i^T(Bu^{n+1} - y) + \lambda q \text{sign}(u_i^{n+1})|u_i^{n+1}|^{q-1} = (\frac{1}{\mu} - B_i^T B_i)(u_i^n - u_i^{n+1}). \tag{5.13}$$

That is, $\nabla_i T(u^{n+1}) = (\frac{1}{\mu} - B_i^T B_i)(u_i^n - u_i^{n+1})$.

(e) **(sufficient descent)** *$T(u^n) - T(u^{n+1}) \geq a\|u^n - u^{n+1}\|_2^2$ for any $n \in \mathbf{N}$, where $a \triangleq \frac{1}{2}(\frac{1}{\mu} - L_{\max})$.*

Proof. It can be observed that the properties of $\{u^n\}$ listed in Lemma 5.5 are some direct extensions of those of $\{x^n\}$. More specifically, Lemma 5.5(a) can be derived by (1.4)-(1.5) and the construction procedure. Lemma 5.5(b) is obtained directly by the cyclic update rule.

Lemma 5.5(c) and (d) can be derived by Lemma 5.1(b) and Lemma 5.5(a). Lemma 5.5(e) can be obtained by Lemma 5.2 and the definition of T (5.9). ■

By Lemma 5.5(e), $\{u^n\}$ satisfies the sufficient descent property, and according to Lemma 5.5(c), $\{u^n\}$ lies in a special subspace, which does not contain the null point and all the points in axes. Thus, the function T is sufficiently smooth in such special subspace, and many proof techniques in smooth analysis can be adopted to analyse the convergence of $\{u^n\}$. As a consequence, the difficulty of the proof of global convergence of $\{x^n\}$ is significantly released via utilizing such technique. Besides Lemma 5.5, the following lemma shows that $\{u^n\}$ also satisfies the relative error property.

Lemma 5.6. (Relative error) *When $n \geq K-1$, $\nabla T(u^{n+1})$ satisfies*

$$\|\nabla T(u^{n+1})\|_2 \leq b\|u^{n+1} - u^n\|_2,$$

where $b = (\frac{1}{\mu} + K\delta)\sqrt{K}$, and $\delta = \max_{i,j=1,2,\cdots,K} |B_i^T B_j|$.

Proof. We assume that $n+1 = j^*K + i^*$ for some positive integers $j^* \geq 1$ and $1 \leq i^* \leq K$. For simplicity, let $i^* = K$. If not, we can renumber the indices of the coordinates such that $i^* = K$ holds while the sequence $\{u^n\}$ keeps invariant, since the update rule (5.11) is cyclic and thus periodic. Such an operation can be described as follows: for each $n \geq K$, by Lemma 5.5(b), we know that u^n are only related to the previous $K-1$ iterates. Thus, we consider the following a period of the original update order, i.e., $\{i^*+1, \cdots, K, 1, \cdots, i^*\}$, then we can renumber the above coordinate update order as $\{1', \cdots, (K-i^*)', (K-i^*+1)', \cdots, K'\}$, with

$$j' = \begin{cases} i^* + j, & \text{if } 1 \leq j \leq K - i^* \\ j - (K - i^*), & \text{if } K - i^* < j \leq K \end{cases}.$$

In the following, we calculate $\nabla_i T(u^{n+1})$ by a recursive way. For $i = K$, by Lemma 5.5(d), it holds

$$\nabla_K T(u^{n+1}) = (\frac{1}{\mu} - B_K^T B_K)(u_K^n - u_K^{n+1}). \tag{5.14}$$

For any $i = K-1, K-2, \ldots, 1$, $\nabla_i T(u^{n+1}) = B_i^T(Bu^{n+1} - y) + \lambda q \text{sign}(u_i^{n+1})|u_i^{n+1}|^{q-1}$ and $u_i^{n+1} = u_i^n$. Therefore, for $i = K-1, K-2, \ldots, 1$, $\nabla_i T(u^{n+1}) = \nabla_i T(u^n) + B_i^T B_K(u_K^{n+1} - u_K^n)$.

For any $j = 1, \ldots, K-1$, by a recursive way, we have

$$\begin{aligned}\nabla_{K-j} T(u^{n+1}) &= \nabla_{K-j} T(u^n) + B_{K-j}^T B_K (u_K^{n+1} - u_K^n) \\ &= \nabla_{K-j} T(u^{n-1}) + B_{K-j}^T \sum_{k=0}^{1} B_{K-k} (u_{K-k}^{n+1-k} - u_{K-k}^{n-k}) \\ &= \nabla_{K-j} T(u^{n-j+1}) + B_{K-j}^T \sum_{k=0}^{j-1} B_{K-k} (u_{K-k}^{n+1-k} - u_{K-k}^{n-k}). \end{aligned} \tag{5.15}$$

Moreover, Lemma 5.5(d) gives $\nabla_{K-j} T(u^{n-j+1}) = (\frac{1}{\mu} - B_{K-j}^T B_{K-j})(u_{K-j}^{n-j} - u_{K-j}^{n-j+1})$. Plugging this into (5.15), it holds $\nabla_{K-j} T(u^{n+1}) = \frac{1}{\mu}(u_{K-j}^{n-j} - u_{K-j}^{n-j+1}) + \sum_{k=0}^{j} B_{K-j}^T B_{K-k} (u_{K-k}^{n+1-k} - u_{K-k}^{n-k})$, which implies

$$\begin{aligned} |\nabla_{K-j} T(u^{n+1})| &\leq \frac{1}{\mu} |u_{K-j}^n - u_{K-j}^{n+1}| + \sum_{k=0}^{j} (|B_{K-j}^T B_{K-k}| \cdot |u_{K-k}^{n+1} - u_{K-k}^n|) \\ &\leq \frac{1}{\mu} |u_{K-j}^n - u_{K-j}^{n+1}| + \delta \|u^{n+1} - u^n\|_1, \end{aligned}$$

for $j = 0, 1, \cdots, K-1$, where the second inequality holds for $\delta = \max_{i,j=1,\cdots,K} |B_i^T B_j|$ and $\sum_{k=0}^{j} |u_{K-k}^{n+1} - u_{K-k}^n| \leq \|u^{n+1} - u^n\|_1$. Summing up $|\nabla_{K-j} T(u^{n+1})|$ with respect to j gives

$$\begin{aligned} \|\nabla T(u^{n+1})\|_1 &\leq \frac{1}{\mu} \|u^{n+1} - u^n\|_1 + K\delta \|u^{n+1} - u^n\|_1 \\ &\leq (\frac{1}{\mu} + K\delta) \sqrt{K} \|u^{n+1} - u^n\|_2, \end{aligned} \tag{5.16}$$

where the second inequality holds for the norm inequality between 1-norm and 2-norm, that is,

$$\|u\|_2 \leq \|u\|_1 \leq \sqrt{K} \|u\|_2, \tag{5.17}$$

for any $u \in \mathbf{R}^K$. Also, using the first inequality of (5.17), then (5.16) implies

$$\|\nabla T(u^{n+1})\|_2 \leq (\frac{1}{\mu} + K\delta) \sqrt{K} \|u^{n+1} - u^n\|_2.$$

■

By the boundedness of $\{x^n\}$ and the construction procedure of $\{u^n\}$, we know that $\{u^n\}$ is also bounded and has a convergent subsequence. Although we have known that T is a KL function with $\varphi(s) = cs^{1-\theta}$ for $c > 0$ and some $\theta \in [0, 1)$, the specific θ is still unknown. In the following, we present a lemma to show that T satisfies the KL inequality at the limit point of $\{u^n\}$ with $\theta = 1/2$ under certain conditions.

Lemma 5.7. (KL property of T) *Let u^* be a limit point of $\{u^n\}$. Suppose that the Hessian matrix $\nabla^2 T(u^*)$ is positive definite, then T satisfies the KL property at u^* with the function $\varphi(s) = cs^{1/2}$ for some $c > 0$.*

This Lemma is a corollary of [44, Lemma 2]. Actually, according to [44, Lemma 2], the condition of this lemma can be weakened that $\nabla^2 T(u^*)$ is nonsingular.

Proposition 5.8. (Global convergence and rate) *Let $\{u^n\}$ be a sequence constructed in Lemma 5.4. Then $\{u^n\}$ converges to a point u^*. If further $\nabla^2 T(u^*)$ is positive definite (actually, this can be weakened to be nonsingular), then there exist a positive integer n^*, constants $C > 0$ and $\rho \in [0, 1)$ such that $\|u^n - u^*\| \leq C\rho^n$ for any $n > n^*$.*

Proof. The convergence of $\{u^n\}$ is a direct result of Theorem 2.9 in [2] for an abstract algorithm, since $\{u^n\}$ satisfies the so-called sufficient decrease, relative error and continuity conditions, and T is also a KL function. The rate of $\{u^n\}$ can be estimated via the similar proof of [1, Theorem 5] for proximal algorithm. ■

From the above analysis, the support and sign convergence property plays a core role in the proof of global convergence of GAITA. Based on this property, an auxiliary sequence $\{u^n\}$ can be constructed from the original sequence $\{x^n\}$ generated by GAITA as shown in Lemma 5.4. While the convergence and rate of $\{u^n\}$ can be justified via a bit standard techniques as shown by Proposition 5.8.

5.3. Main Proofs

For any $x \in \mathbf{R}^N$, we define the operator $G_{\mu,\lambda\|\cdot\|_q^q}$ as

$$G_{\mu,\lambda\|\cdot\|_q^q}(x) = \text{prox}_{\mu,\lambda\|\cdot\|_q^q}(x - \mu A^T(Ax - y)). \tag{5.18}$$

Then we denote by $\mathcal{F}_q$ the fixed point set of the operator $G_{\mu,\lambda\|\cdot\|_q^q}$, i.e., $\mathcal{F}_q = \{x : x = G_{\mu,\lambda\|\cdot\|_q^q}(x)\}$.

Lemma 5.9. *(Theorem 3 in [27]). Given a point x^*, define the support of x^* as $\text{supp}(x^*) = \{i : x_i^* \neq 0\}$, then $x^* \in \mathcal{F}_q$ if and only if the following conditions hold:*

(a) for $i \in \text{supp}(x^*)$, $|x_i^*| \geq \eta_{\lambda\mu,q}$, and $A_i^T(Ax^* - y) + \lambda q \text{sign}(x_i^*)|x_i^*|^{q-1} = 0$;

(b) for $i \in \text{supp}(x^*)^c$, $|x_i^*| = 0$ and $|A_i^T(Ax^* - y)| \leq \tau_{\lambda\mu,q}/\mu$.

We call x^* a **stationary point** of the ℓ_qLS problem henceforth if it satisfies the optimality conditions in Lemma 5.9. Moreover, similar to Theorem 10 in [27], we can also claim that the mapping $\mathcal{T}$ is closed.

Lemma 5.10. (Any limit point is a stationary point) *Suppose that $0 < \mu < L_{\max}^{-1}$, and denote by $\mathcal{L}$ the limit point set of $\{x^n\}$, then $\mathcal{L} \subseteq \mathcal{F}_q$.*

The proof of this lemma is the same with that of [27, Theorem 5]. It only needs to note that $\text{prox}_{\mu,\lambda|\cdot|^q}$ is discontinuous at $\tau_{\lambda\mu,q}$ while $\text{prox}_{1,\lambda|\cdot|^q}$ is discontinuous at $\tau_{\lambda,q}$.

Proof. (for Theorem 2.1) By Proposition 5.3, the claims (a)-(b) and part of (c) in Theorem 2.1 hold. It only needs to show $\text{supp}(x^*) = I$ and $\text{sign}(x^*) = S^*$ for any $x^* \in \mathcal{L}$.

For any limit point $x^* \in \mathcal{L}$, there exits a subsequence $\{x^{n_j}\}$ converging to x^*, i.e., $x^{n_j} \to x^*$ as $j \to \infty$. Thus, there exists a sufficiently large positive integer j_0 such that $n_{j_0} > n_0$ and $\|x^{n_j} - x^*\|_2 < \eta_{\lambda\mu,q}$ when $j \geq j_0$. By Lemmas 5.10 and 5.9, any limit point x^* also has the similar alternative property of $\{x^n\}$. Thus, similar to the proof of Proposition 5.3, we have $\text{supp}(x^*) = I$ and $\text{sign}(x^*) = S^*$ for any $x^* \in \mathcal{L}$. ■

Proof. (for Theorem 2.2) According to the construction procedure of $\{u^n\}$ and by Proposition 5.8, we have that $\{x^n\}$ converges, and the rate of convergence is eventual linear under the conditions of this theorem. Let x^* be the limit point of $\{x^n\}$, then by Lemma 5.10, x^* is a stationary point.

In the following, we only need to justify that x^* is also a strictly local minimizer under the condition (2.1). Let $F(x) \triangleq \frac{1}{2}\|Ax - y\|_2^2$ and $\phi_1(x_I^*) \triangleq (q\text{sign}(x_{i_1}^*)|x_{i_1}^*|^{q-1}, \cdots, q\text{sign}(x_{i_K}^*)|x_{i_K}^*|^{q-1})^T$, where $i_j \in I, j = 1, \ldots, K$. By Lemma 5.9(a) we have

$$A_I^T(Ax^* - y) + \lambda\phi_1(x_I^*) = 0. \tag{5.19}$$

This together with the condition of the theorem, $A_I^T A_I + \lambda q(q-1)\Lambda(x_I^*) \succ 0$ imply that the second-order optimality conditions hold at $x^* = (x_I^*, 0)$. For sufficiently small vector h, we denote $x_h^* = (x_I^* + h_I, 0)$. It then follows

$$F(x_h^*) + \lambda \sum_{i\in I} |x_i^* + h_i|^q \geq F(x^*) + \lambda \sum_{i\in I} |x_i^*|^q. \tag{5.20}$$

Furthermore, for any $q \in (0, 1)$, it obviously holds that $t^q > (\|\nabla_{I^c} F(x^*)\|_\infty + 2)t/\lambda$, for sufficiently small $t > 0$. By this fact and the differentiability of F, one can observe that for sufficiently small h, there hold

$$\begin{aligned} &F(x^* + h) - F(x_h^*) + \lambda \sum_{i\in I^c} |h_i|^q = \nabla_{I^c}^T F(x^*) h_{I^c} + \lambda \sum_{i\in I^c} |h_i|^q + o(h_{I^c}) \\ &\geq \sum_{i\in I^c} (\|\nabla_{I^c} F(x^*)\|_\infty - [\nabla_{I^c} F(x^*)]_i + 1)|h_i| \geq 0. \end{aligned} \tag{5.21}$$

Summing up the above two inequalities (5.20)-(5.21), one has that for all sufficiently small h,

$$f(x^* + h) - f(x^*) \geq 0, \tag{5.22}$$

and hence x^* is a local minimizer. In addition, we can observe that when $h \neq 0$, then at least one of the two inequalities (5.20) and (5.21) will hold strictly, which implies that x^* is a strictly local minimizer. ■

6. Conclusion

In this paper, we focused on utilizing the Gauss-Seidel iteration rule to the iterative thresholding algorithm for solving the nonconvex ℓ_q regularized least squares regression problem and developed a new algorithm called GAITA. The convergence and convergence rate of the proposed algorithm are derived. The proposed algorithm can allow a larger fixed step size, and have faster rate of convergence as well as lower computational cost than its Jacobi counterpart.

Note that the proposed algorithm can be extended to solve a class of nonconvex composite (i.e., smooth+nonsmooth) optimization problems, where their proximity operators can be easily computed and have the so-called "*jumping*" property. Our analysis framework is also applicable to analyze the convergence of the extended algorithm. When it comes to parallel implementation, however, GAITA could have certain disadvantages because variables that depend on each other can only be updated sequentially.

References

[1] H. Attouch and J. Bolte, On the convergence of the proximal algorithm for nonsmooth functions involving analytic features, Mathematical Programming, 116: 5-16, 2009.

[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, Mathematical Programming, 137: 91-129, 2013.

[3] A. Bagirov, L. Jin, N. Karmitsa, A. Al Nuaimat and N. Sultanova, Subgradient method for nonconvex nonsmooth optimization, Journal of Optimization Theory and applications, 157: 416-435, 2013.

[4] A. Beck and M. Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM Journal on Imaging Sciences, 2: 183-202, 2009.

[5] A. Beck and M. Teboulle, A linearly convergent algorithm for solving a class of nonconvex/affine feasibility problems, in Fixed-Point Algorithms for Inverse Problems in Science

and Engineering, ser. Springer Optimization and Its Applications. New York, NY, USA: Springer, pp. 33-48, 2011.

[6] A. Beck and L. Tetruashvili, On the convergence of block coordinate descent type methods, SIAM Journal on Optimization, 23: 2037-2060, 2013.

[7] J. Bolte, S. Sabach and M. Teboulle, Proximal alternating linearized minimization for nonconvex and nonsmooth problems, Mathematical Programming, 146(1): 459-494, 2014.

[8] J. Burke, A. Lewis and M. Overton, A robust gradient sampling algorithm for nonsmooth, nonconvex optimization, SIAM Journal on Optimization, 15: 751-779, 2005.

[9] T. Blumensath and M. Davies, Iterative thresholding for sparse approximation, Journal of Fourier Analysis and Application, 14(5): 629-654, 2008.

[10] K. Bredies, D. Lorenz and S. Reiterer, Minimization of non-smooth, non-convex functionals by iterative thresholding, Journal of Optimization Theory and Applications, 165: 78-122, 2015.

[11] E. Candès, M. Wakin and S. Boyd, Enhancing sparsity by reweighted l_1 minimization, Journal of Fourier Analysis and Applications, 14(5): 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, Journal of Visual Communication and Image Representation, 24(1): 1529-1542, 2013.

[13] R. Chartrand, Exact reconstruction of sparse signals via nonconvex minimization, IEEE Signal Processing Letters, 14(10): 707-710, 2007.

[14] R. Chartrand and V. Staneva, Restricted isometry properties and nonconvex compressive sensing, Inverse Problems, 24: 1-14, 2008.

[15] X. Chen, Smoothing methods for nonsmooth, nonconvex minimization, Mathematical programming, 134: 71-99, 2012.

[16] P. Combettes and V. Wajs, Signal recovery by proximal forward-backward splitting. Multiscale Modeling and Simulation, 4: 1168-1200, 2005.

[17] I. Cumming and F. Wong. Digital processing of synthetic aperture radar data: algorithms and implementation, MA: Artech House, 2004.

[18] I. Daubechies, R. DeVore, M. Fornasier and C. Gunturk, Iteratively reweighted least squares minimization for sparse recovery, Communications on Pure and Applied Mathematics, 63: 1-38, 2010.

[19] I. Duabechies, M. Defrise and C. De Mol, An iterative thresholding algorithm for linear inverse problems with a sparse constraint, Communications on Pure and Applied Mathematics, 57: 1413-1457, 2004.

[20] J. Friedman, T. Hastie, H. Hofling and R. Tibshirani, Pathwise coordinate optimization, Annals of Applied Statistics, 1(2): 302-332, 2007.

[21] A. Fuduli, M. Gaudioso and G. Giallombardo, Minimizing nonconvex nonsmooth functions via cutting planes and proximity control, SIAM Journal on Optimization, 14: 743-756, 2004.

[22] L. Grippo and M. Sciandrone, Globally convergent block-coordinate techniques for unconstrained optimization, Optimization Methods and Software, 10: 587-637, 1999.

[23] C. Hildreth, A quadratic programming procedure, Naval Research Logistics Quarterly, 4: 79-85, 1957.

[24] J. Kivinen, Exponentiated gradient versus gradient descent for linear predictors, Information and Computation, 132(1): 1-63, 1997.

[25] D. Krishnan and R. Fergus, Fast image deconvolution using hyperLaplacian priors, in Proc. Adv. Neural Inf. Process. Syst. (NIPS), 2009.

[26] Z. Luo and P. Tseng, On the convergence of the coordinate descent method for convex differentiable minimization, Journal of Optimization Theory and Applications, 72: 7-35, 1992.

[27] G. Marjanovic and V. Solo, l_q sparsity penalized linear regression with cyclic descent, IEEE Transactions on Signal Processing, 62(6): 1464-1475, 2014.

[28] R. Mazumder, J. Friedman and T. Hastie, Sparsenet: Coordinate descent with nonconvex penalties, Journal of the American Statistical Association, 106: 1125-1138, 2007.

[29] Z. Peng, T. Wu, Y. Xu, M. Yan and W. Yin, Coordinate friendly structures, algorithms and applications, Annals of Mathematical Sciences and Applications, 1: 57-119, 2016.

[30] A. Ostrowski, Solutions of equaltions in Euclidean and Banach spaces, New York, NY, USA: Academic, 1973.

[31] J. Ortega and W. Rheinboldt, Iterative solution of nonlinear equations in several variables. SIAM, 2000.

[32] M. Razaviyayn, M. Hong and Z. Luo, A unified convergence analysis of block successive minimization methods for nonsmooth optimization, SIAM Journal on Optimization, 23: 1126-1153, 2013.

[33] A. Seneviratne and V. Solo, On exact denoising, School Elect. Eng. Telecommun., Univ. New South Wales, New South Wales, Australia, Tech. Rep., 2013.

[34] R. Tibshirani, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society. Series B (Methodological), 58(1): 267-288, 1996.

[35] J. Tropp and A. Gilbert, Signal recovery from random measurements via orthogonal matching pursuit, IEEE Transactions on Information Theory, 53: 4655-4666, 2007.

[36] P. Tseng, Convergence of a block coordinate descent method for nondifferentiable minimization, Journal of Optimization Theory and Applications, 109: 475-494, 2001.

[37] P. Tseng and S. Yun, A coordinate gradient descent method for nonsmooth separable minimization, Mathematical Programming, 117: 387-423, 2009.

[38] J. Tsitsiklis, A comparison of Jacobi and Gauss-Seidel parallel iterations, Applied Mathematics Letters, 2(2): 167-170, 1989.

[39] Y. Xu and W. Yin, A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion, SIAM Journal on Imaging Sciences, 6: 1758-1789, 2013.

[40] Z. Xu, X. Chang, F. Xu and H. Zhang, $L_{1/2}$ regularization: a thresholding representation theory and a fast solver, IEEE Transactions on Neural Networks and Learning Systems, 23: 1013-1027, 2012.

[41] J. Zeng, J. Fang and Z. Xu, Sparse SAR imaging based on $L_{1/2}$ regularization, Science China Series F-Information Science, 55: 1755-1775, 2012.

[42] J. Zeng, Z. Xu, B. Zhang, W. Hong and Y. Wu, Accelerated $L_{1/2}$ regularization based SAR imaging via BCR and reduced Newton skills, Signal Processing, 93(7): 1831-1844, 2013.

[43] J. Zeng, S. Lin, Y. Wang and Z. Xu, $L_{1/2}$ Regularization: convergence of iterative half thresholding algorithm, IEEE Transactions on Signal Processing, 62(9): 2317-2329, 2014.

[44] J. Zeng, S. Lin and Z. Xu, Sparse Regularization: Convergence of Iterative Jumping Thresholding Algorithm, IEEE Transactions on Signal Processing, 64(19): 5106-5117, 2016.

[45] Y. Wang, J. Zeng, Z. Peng, X. Chang and Z. Xu, Linear convergence of adaptively iterative thresholding algorithms for compressed sensing, IEEE Transactions on Signal Processing, 63(11): 2957-2971, 2015.

[46] L. Zhu, L. Li, R. Li and L. Zhu, Model-free feature screening for ultrahigh dimensional data. Journal of the American Statistical Association, 106: 1464-1475, 2011.