Newton Method for $\ell_0$-Regularized Optimization

Shenglong Zhou · Lili Pan · Naihua Xiu

Received: 14 August 2020 / Accepted: 7 February 2021 / Published Online: 24 March 2021

Abstract As a tractable approach, regularization is frequently adopted in sparse optimization. This gives rise to regularized optimization, which aims to minimize the $\ell_0$ norm or its continuous surrogates that characterize the sparsity. From the continuity of surrogates to the discreteness of the $\ell_0$ norm, the most challenging model is the $\ell_0$-regularized optimization. There is an impressive body of work on the development of numerical algorithms to overcome this challenge. However, most of the developed methods only ensure that either the (sub)sequence converges to a stationary point from the deterministic optimization perspective or that the distance between each iteration and any given sparse reference point is bounded by an error bound in the sense of probability. In this paper, we develop a Newton-type method for the $\ell_0$-regularized optimization and prove that the generated sequence converges to a stationary point globally and quadratically under the standard assumptions, theoretically explaining that our method can perform surprisingly well.

Keywords $\ell_0$-regularized optimization · $\tau$-stationary point · Newton method · Global and quadratic convergence

Mathematics Subject Classification (2010) 65K05 · 90C46 · 90C27

1 Introduction

Over the last decade, sparsity has been thoroughly investigated due to its extensive applications ranging from compressed sensing [23, 15, 16], signal and image processing [25, 24, 17, 8], machine learning [48, 53] to neural networks [7, 33, 22] lately. Sparsity is frequently characterized by the $\ell_0$ norm, and its penalized version is expressed as the following $\ell_0$-regularized optimization,

\begin{equation}
\min_{x \in \mathbb{R}^n} f(x) + \lambda \|x\|_0,
\end{equation}

This work was funded by the National Science Foundation of China (11971052, 11801325, 11771255) and Young Innovation Teams of Shandong Province (2019KJI013).

Shenglong Zhou
School of Mathematics, University of Southampton, Southampton SO17 1BJ, UK
E-mail: shenglong.zhou@soton.ac.uk

Lili Pan
Department of Mathematics and Statistics, Shandong University of Technology, Zibo 255049, China
E-mail: panlili1979@163.com

Naihua Xiu
Department of Applied Mathematics, Beijing Jiaotong University, Beijing 100044, China
E-mail: nhxiu@bjtu.edu.cn
where $f : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable and bounded from below, \( \lambda > 0 \) is the penalty parameter and \( \|x\|_0 \) is the \( \ell_0 \) norm of \( x \), counting the number of non-zero elements of \( x \). Differing from the regularized optimization, another category of sparsity involved problems that have been well studied lately is the so-called sparsity constrained optimization,

\[
\min_{x \in \mathbb{R}^n} f(x), \quad \text{s.t. } \|x\|_0 \leq s,
\]

where \( s \ll n \) is a given positive integer. Based on the two optimizations, many state-of-the-art methods have been proposed in the last two decades. In particular, many of these were designed for the special case of compressed sensing (CS) where

\[
f(x) := f_{cs}(x) \equiv \frac{1}{2}\|Ax - y\|^2.
\]

Here, \( A \in \mathbb{R}^{m \times n} \) is the sensing matrix and \( y \in \mathbb{R}^m \) is the measurement.

### 1.1 Selective Literature Review

Since there is a vast body of work on the development of numerical methods to solve the problems expressed by (1.1) and (1.2), we only present a brief overview of the work that clarifies the motivation of this paper.

**Methods for (1.2)** are known as greedy methods. For the case of CS, the orthogonal matching [40,47, OMP], the gradient pursuit [12, GP], the compressive sample matching pursuit [38, CoSaMP], the subspace pursuit [20, SP], the normalized iterative hard-thresholding [14, NIHT], the hard-thresholding pursuit [28, HTP] and the accelerated iterative hard-thresholding [11, AIHT] methods have been developed. Methods for the general model (1.2) include the gradient support pursuit [2, GraSP], the iterative hard-thresholding [4, IHT], the Newton gradient pursuit [52, NTGP], the conjugate gradient iterative hard-thresholding [10, CGIHT], the gradient hard-thresholding pursuit [51, GraHTP], the improved iterative hard-thresholding [39, IIHT] and the Newton hard-thresholding pursuit [55, NHTP].

For the convergence results, most methods establish the theory that the distance between each iteration and any given reference (sparse) point is bounded by an error through statistical analysis. By contrast, methods such as IHT, IIHT and NHTP have been proven to converge to a stationary point globally in the deterministic sense. Moreover, if Newton directions are interpolated into some methods, for example, CoSaMP, SP, GraSP, NTGP and GraHTP, then their demonstrated empirical performances are extraordinary in terms of the super-fast computational speed and high accuracy, but without deterministic theoretical guarantees over a long time. Recently, [55] proved for the first time that their proposed NHTP has the global and quadratic convergence that justify the excellent performance of these methods.

**Methods for (1.1)** that aim to address the CS problems via the model (1.1) include the iterative hard-thresholding algorithm [13, IHT], the continuous exact \( \ell_0 \) penalty [44, CEL0], and the continuation single best replacement and the \( \ell_0 \)-regularization path descent in [45, CSBR, L0BD], the forward-backward splitting [1, FBS], the extrapolated proximal iterative hard-thresholding algorithm [3, EPIHT] and the mixed integer optimization method [6, MIO], to name just a few. For the general problem (1.1), plentiful methods are available. They are the penalty decomposition [36, PD] where equality and inequality constraints are also considered, the iterative hard-thresholding [35, see] where the box and convex cone are taken into account, the proximal gradient method and coordinate-wise support optimality method [5, PG, CowS] where sparse solutions are sought from a symmetric set, the random proximal alternating minimization method [41, RPA], the active set Barzilar-Borwein [18, ABB] and the very recently developed smoothing proximal gradient method [9, SPG].

We note that these methods are known as the first-order methods since they only exploit the first-order information such as gradients or function values. Recently, second-order methods have attracted much attention, including the primal dual active set [30, PDAS], the primal dual active set with continuation [31, PDASC] and the support detection and root finding [29, SDAR].
For convergence results, either error bounds are achieved for methods such as IHT, EPIHT, PDASC and SDAR, or a subsequence converges to a stationary point (which is a non-global convergence property) for methods such as PD, PG and ABB. It is important to mention that the authors in [1] prove that FBS converges to a critical point globally and authors [9, SPG] also show the global convergence to a relaxation problem of (1.1). Moreover, no better deterministic theoretical guarantees (such as quadratic convergence) have been established for algorithms used to solve (1.1). Therefore, a natural question arises: can we develop an algorithm based on the $\ell_0$-regularized optimization that enjoys the global and quadratic convergence?

1.2 Contributions

To answer the above question, we first introduce a $\tau$-stationary point, an optimality condition of (1.1), and then reveal its relationship with the local/global minimizers by Theorem 1. It is known that a $\tau$-stationary point is a necessary optimality condition by [5, Theorem 4.10]. However, we show that it is also a sufficient condition under the assumption of strong convexity.

The $\tau$-stationary point can be expressed as a stationary equation system (2.12), and allows us to employ the Newton-type method dubbed as NL0R, an abbreviation for Newton method for $\ell_0$-regularized optimization (1.1). Differing from the classical Newton methods that are usually employed on continuous equation systems, the stationery equation system turns out to be discontinuous. Nevertheless, we succeed in establishing the global and quadratic convergence properties for NL0R under standard assumptions, such as the strong smoothness and locally strong convexity of $f$, see Theorem 4. To the best of our knowledge, this work is the first paper that establishes both properties for an algorithm aiming to solve the $\ell_0$-regularized optimization problem.

Finally, extensive numerical experiments are conducted in this article to demonstrate that NL0R is highly competitive when benchmarked against a number of leading solvers for solving the compressed sensing and sparse complementarity problems. In a nutshell, NL0R can deliver relatively accurate sparse solutions with high computational speed.

To end this subsection, we would like to emphasize the differences between NL0R and some of the aforementioned methods. First, PDASC, SDAR and NHTP also adopt the concept of the $\tau$-stationary point. The former two methods always set $\tau = 1$, while similar to NHTP, NL0R benefits from more choices of $\tau$. Second, the gradient direction and Amijo-type rule of updating the step size are integrated. Those strategies are alternatives if the Newton direction does not guarantee sufficient descent of the objective function values during the process.

By contrast, PDASC and SDAR only take advantage of the Newton directions with unit step sizes. Therefore, it is difficult to establish the global convergence results for these methods. Finally, for the NHTP method for solving (1.2), the sparsity level $s$ is required, but is usually unknown and somehow determines the quality of the final solutions. In (1.1), the parameter $\lambda$ also plays an important role in pursuing sparse solutions. We will show that $\lambda$ can be set in a proper range and the proposed method NL0R can effectively tune it adaptively in numerical experiments.

1.3 Organization and Notation

The rest of the paper is organized as follows. The next section establishes the optimality conditions of (1.1) with the help of the $\tau$-stationary point. Its relationship with the local/global minimizers of (1.1) by Theorem 1 is also obtained. In Section 3, we design the Newton-type method for the $\ell_0$-regularized optimization (NL0R), followed by the main convergence results including the support set identification, and global and quadratic convergence properties. Extensive numerical experiments are presented in Section 4, where the implementation of NL0R and its comparisons with some excellent solvers for solving the compressed sensing and sparse complementarity problems are provided. Concluding remarks are made in the last section.
We end this section with some notation that is employed throughout the paper. Let \( N_n := \{1, 2, \cdots, n\} \) and \( \| \cdot \| \) be the Euclidean norm. Given a vector \( x \), let \( |x| := (|x_1|, |x_2|, \cdots, |x_n|)^\top \), and \( \text{supp}(x) \) be its support set consisting of the indices of the non-zero elements of \( x \). Given a set \( T \subseteq N_n, |T| \) and \( T \) are the cardinality and the complementary set. The sub-vector of \( x \) containing elements indexed on \( T \) is denoted by \( x_T \). To see this, let \( \text{supp}(x) \) be the Euclidean norm. Given a vector \( x \), let \( \|\cdot\| := \|\cdot\|_{L_2} \). Next, \( [a] \) represents the smallest integer that is no less than \( a \). Now, for a matrix \( A \in \mathbb{R}^{m \times n} \), let \( \|A\|_2 \) represent its spectral norm, i.e., its maximum singular value. \( A_{T,J} \) is the sub-matrix containing rows indexed by \( T \) and columns indexed by \( J \). In particular, we denote the sub-gradient and sub-Hessians by

\[
\nabla_T f(x) := (\nabla f(x))_T, \quad \nabla^2_T f(x) := (\nabla^2 f(x))_{T,T},
\]

\[
\nabla^2_{T,J} f(x) := (\nabla^2 f(x))_{T,J}, \quad \nabla^2_T f(x) := (\nabla^2 f(x))_{T,N_n}.
\]

## 2 Optimality

Some necessary optimality conditions of (1.1) have been established [36, Theorem 2.1] and [5, Theorem 4.10]. Here, inspired by the \( L \)-stationarity in [5], we introduce a \( \tau \)-stationary point.

### 2.1 \( \tau \)-stationary point

A point \( x \in \mathbb{R}^n \) is called a \( \tau \)-stationary point of the problem (1.1) if there is a \( \tau > 0 \) satisfying

\[
\begin{align*}
\nabla_T f(x) := (\nabla f(x))_T, \quad \nabla^2_T f(x) := (\nabla^2 f(x))_{T,T}, \\
\nabla^2_{T,J} f(x) := (\nabla^2 f(x))_{T,J}, \quad \nabla^2_T f(x) := (\nabla^2 f(x))_{T,N_n}.
\end{align*}
\]

\[
(2.1) \quad x \in \text{Prox}_{\tau \lambda \|\cdot\|_0} \left( x - \tau \nabla f(x) \right) := \arg\min_{z \in \mathbb{R}^n} \frac{1}{2} \| z - (x - \tau \nabla f(x)) \|^2 + \tau \lambda \| z \|_0.
\]

It follows from [1] that the operator \( \text{Prox}_{\tau \lambda \|\cdot\|_0} (z) \) takes a closed form as

\[
(2.2) \quad \left[ \text{Prox}_{\tau \lambda \|\cdot\|_0} (z) \right]_i = \begin{cases} 
\frac{z_i}{\lambda}, & \text{if } |z_i| > \frac{2\tau \lambda}{\lambda}, \\
0, & \text{if } |z_i| < \frac{2\tau \lambda}{\lambda}, \\
|z_i|, & \text{if } |z_i| = \frac{2\tau \lambda}{\lambda},
\end{cases}
\]

This allows us to equivalently characterize a \( \tau \)-stationary point by the conditions below or see [46, Theorem 24] and [13, Lemma 2].

**Lemma 1** A point \( x \) is a \( \tau \)-stationary point with \( \tau > 0 \) of (1.1) if and only if

\[
(2.3) \quad \begin{cases} 
\nabla_i f(x) = 0 \text{ and } |x_i| \geq \frac{2\tau \lambda}{\lambda}, & i \in \text{supp}(x), \\
\|\nabla f(x)\| \leq \frac{2\lambda}{\sqrt{\tau}}, & i \notin \text{supp}(x).
\end{cases}
\]

From Lemma 1, for any \( 0 < \tau_1 \leq \tau \), a \( \tau \)-stationary point \( x \) is also a \( \tau_1 \)-stationary point due to \( 2\tau \lambda \geq 2\tau_1 \lambda \) and \( 2\lambda/\tau \leq 2\lambda/\tau_1 \). Our next major result needs the strong smoothness and convexity of \( f \).

**Definition 1** A function \( f \) is strongly smooth with a constant \( L > 0 \) if

\[
(2.4) \quad f(z) \leq f(x) + \langle \nabla f(x), z - x \rangle + (L/2)\|z - x\|^2, \quad \forall \ x, z \in \mathbb{R}^n.
\]

A function \( f \) is strongly convex with a constant \( \ell > 0 \) if

\[
(2.5) \quad f(z) \geq f(x) + \langle \nabla f(x), z - x \rangle + (\ell/2)\|z - x\|^2, \quad \forall \ x, z \in \mathbb{R}^n.
\]

We say a function \( f \) is locally strongly convex with a constant \( \ell > 0 \) around \( x \) if (2.5) holds for any point \( z \) in the neighbourhood of \( x \).

Evidently, when the function is locally strongly convex, the constant \( \ell \) depends on the point \( x \). We drop the dependence for simplicity since it will not cause confusion in context. The strong convexity and smoothness respectively indicate that for any \( x, z \in \mathbb{R}^n \)

\[
(2.6) \quad \ell\|z - x\| \leq \|\nabla f(x) - \nabla f(z)\| \leq L\|z - x\|.
\]
2.2 First order optimality conditions

Our next major result is to establish the relationships between a \( \tau \)-stationary point and a local/global minimizer of (1.1).

**Theorem 1** For the problem (1.1), the following results hold.

1) **(Necessity)** A global minimizer \( x^\ast \) is also a \( \tau \)-stationary point for any \( 0 < \tau < 1/L \) if \( f \) is strongly smooth with \( L > 0 \). Moreover,

\[
x^\ast = \text{Prox}_{\tau \lambda \parallel \cdot \parallel_0} (x^\ast - \tau \nabla f(x^\ast)).
\]

2) **(Sufficiency)** A \( \tau \)-stationary point with \( \tau > 0 \) is a local minimizer if \( f \) is convex. Furthermore, a \( \tau \)-stationary point with \( \tau(> \geq 1/\ell) \) is also a (unique) global minimizer if \( f \) is strongly convex with \( \ell > 0 \).

**Proof** 1) Denote \( P := \text{Prox}_{\tau \lambda \parallel \cdot \parallel_0} (x^\ast - \tau \nabla f(x^\ast)) \) and \( \mu := L - 1/\tau < 0 \) due to \( 0 < \tau < 1/L \). Let \( x^\ast \) be a global minimizer and consider any point \( z \in P \). Then, we have

\[
2f(z) + 2\lambda \parallel z \parallel_0 \\
\leq 2f(x^\ast) + 2\langle \nabla f(x^\ast), z - x^\ast \rangle + L \parallel z - x^\ast \parallel^2 + 2\lambda \parallel z \parallel_0 \\
= 2f(x^\ast) + 2\langle \nabla f(x^\ast), z - x^\ast \rangle + (1/\tau) \parallel z - x^\ast \parallel^2 + \mu \parallel z - x^\ast \parallel^2 + 2\lambda \parallel z \parallel_0 \\
= 2f(x^\ast) + (1/\tau) \parallel z - (x^\ast - \tau \nabla f(x^\ast)) \parallel^2 - \tau \parallel \nabla f(x^\ast) \parallel^2 + 2\lambda \parallel z \parallel_0 + \mu \parallel z - x^\ast \parallel^2 \\
\leq 2f(x^\ast) + (1/\tau) \parallel z - (x^\ast - \tau \nabla f(x^\ast)) \parallel^2 + 2\lambda \parallel z \parallel_0 - \tau \parallel \nabla f(x^\ast) \parallel^2 + \mu \parallel z - x^\ast \parallel^2 \\
= 2f(x^\ast) + 2\lambda \parallel z \parallel_0 + \mu \parallel z - x^\ast \parallel^2.
\]

where the first, second and third inequalities hold from the facts that \( f \) is strongly smooth, and \( z \in P \) and \( x^\ast \) are the global minimizers of (1.1). This together with \( \mu < 0 \) leads to \( 0 \leq (\mu/2) \parallel z - x^\ast \parallel^2 \leq 0 \), yielding \( z = x^\ast \). Therefore, \( x^\ast \) is a \( \tau \)-stationary point of (1.1). Since \( z \) is arbitrary in \( P \) and \( z = x^\ast \), \( P \) is a singleton only containing \( x^\ast \).

2) Let \( x^\ast \) be a \( \tau \)-stationary point with \( \tau > 0 \) with \( T_\ast := \text{supp}(x^\ast) \). Consider a neighbor region of \( x^\ast \) as

\[
N(x^\ast) = \{ x \in \mathbb{R}^n : \parallel x - x^\ast \parallel < \epsilon_\ast \},
\]

where

\[
\epsilon_\ast := \left\{ \min_{t \in T_\ast} \left\{ |t^\ast_i|, \sqrt{\tau \lambda/(2n)} \right\}, \ x^\ast \neq 0, \left\{ \min_{t \in T_\ast} \left\{ |t^\ast_i|, \sqrt{\tau \lambda/(2n)} \right\}, \ x^\ast = 0. \right\}
\]

For any point \( x \in N(x^\ast) \), we conclude that \( T_\ast \subseteq \text{supp}(x) \). In fact, this is true when \( x^\ast = 0 \). When \( x^\ast \neq 0 \), if there is a \( j \) such that \( j \in T_\ast \) but \( j \notin \text{supp}(x) \), then we derive a contradiction:

\[
\epsilon_\ast \leq \min_{t \in T_\ast} \left\{ |t^\ast_i| \right\} \leq |x^\ast_j| = |x^\ast_j - x_j| \leq \parallel x - x^\ast \parallel < \epsilon_\ast.
\]

Therefore, we have \( T_\ast \subseteq \text{supp}(x) \). The convexity of \( f \) suffices to obtain

\[
f(x) - f(x^\ast) \geq \langle \nabla f(x^\ast), x - x^\ast \rangle \\
= \langle \nabla_{T_\ast} f(x^\ast), (x - x^\ast)_{T_\ast} \rangle + (\nabla_{T_\ast} f(x^\ast), (x - x^\ast)_{T_\ast}) \\
= (\frac{\lambda}{2 \tau}) \langle \nabla_{T_\ast} f(x^\ast), x_{T_\ast} \rangle := \phi.
\]

If \( T_\ast = \text{supp}(x) \), then \( \phi = 0 \) due to \( x_{T_\ast} = 0 \) and \( \parallel x^\ast \parallel_0 = \parallel x \parallel_0 \). These allow us to derive that

\[
f(x) + \lambda \parallel x \parallel_0 \geq f(x^\ast) + \phi + \lambda \parallel x \parallel_0 = f(x^\ast) + \lambda \parallel x^\ast \parallel_0.
\]
If $T_* \subset \text{supp}(x)$, then $\|x\|_0 - 1 \geq \|x^*\|_0$. In addition,

$$\phi = \langle \nabla_T f(x^*), x_T^* \rangle \geq -\|\nabla_T f(x^*)\|_T \|x_T^*\| \geq \frac{\|\tau\|}{|T_*| 2\lambda} \|x_T^* - x_T^*\| \geq -\sqrt{n2\lambda/\tau} > -\lambda.$$ (2.3)

These facts enable us to derive that

$$f(x) + \lambda\|x\|_0 \geq f(x^*) + \phi + \lambda\|x\|_0 > f(x^*) + \lambda\|x\|_0 - \lambda \geq f(x^*) + \lambda\|x^*\|_0.$$ (2.8)

Both cases show the local optimality of $x^*$ in the region $N(x^*)$. Again, it follows from $x^*$ being a $\tau$-stationary point with $\tau > 0$ that

$$(1/2)\|x - (x^* - \tau\nabla f(x^*))\|^2 + \tau\lambda\|x\|_0 \geq (1/2)\|x^* - (x^* - \tau\nabla f(x^*))\|^2 + \tau\lambda\|x^*\|_0,$$

for any $x \in \mathbb{R}^n$, which suffices to obtain that

$$(\nabla f(x^*), x - x^*) + \lambda\|x\|_0 \geq -(1/(2\tau))\|x - x^*\|^2 + \lambda\|x^*\|_0.$$ (2.9)

Since $f$ is strongly convex, for any $x \neq x^*$, we have

$$f(x) + \lambda\|x\|_0 \geq f(x^*) + \langle \nabla f(x^*), x - x^* \rangle + (\ell/2)\|x - x^*\|^2 + \lambda\|x\|_0 \geq (\ell - 1/\tau)/2\|x - x^*\|^2 + \lambda\|x^*\|_0 \geq f(x^*) + \lambda\|x^*\|_0.$$ (2.10)

where the last inequality is based on $\tau \geq 1/\ell$. Clearly, if $\tau > 1/\ell$, then the last inequality holds strictly, implying that $x^*$ is a unique global minimizer. \qed

Let us consider an example to illustrate the above theorem.

**Example 1** Let $a = (1 1 1)^T$, $\lambda > 8$ and $f$ be given by

$$f(x) := \frac{1}{2}(x - a)^T \begin{bmatrix} 2 & 0 & 0 \\ 0 & 3 & 1 \\ 0 & 1 & 3 \end{bmatrix} (x - a).$$ (2.10)

It is easy to verify that $f$ is strongly smooth with $L = 4$ and is also strongly convex with $\ell = 2$. Consider a point $x^* = (t 0 0)^T$ with $t \geq \lambda/2$. We can conclude that $x^*$ is a global minimizer of (1.1). In fact, $\nabla f(x^*) = (0 -4 -4)^T$ and $x^* - \tau\nabla f(x^*) = (t 4\tau 4\tau)^T$. This and (2.3) show that $x^*$ is a $\tau$-stationary point for some $\tau \in (1,\lambda/8]$ due to

$$\nabla_1 f(x^*) = 0$$ and $$|x_1^*| = t \geq \lambda/2 = \sqrt{2\lambda\lambda/8} \geq \sqrt{2\lambda\tau},$$

$$|\nabla_2 f(x^*)| = |\nabla_3 f(x^*)| = 4 = \sqrt{2 \times 8} \leq \sqrt{2\lambda/\tau}.$$ (2.10)

Then it follows from Theorem 1.2) and $\tau > 1/2 = 1/\ell$ that $x^*$ is a unique global minimizer of (1.1). Moreover, Theorem 1.1) concludes that a global minimizer (which is $x^*$) is also a $\tau_1$-stationary point with $\tau_1 \in (0, 1/L) = (0, 1/4)$. This does not conflict with $x^*$ being a $\tau$-stationary point with some $\tau \in (1, \lambda/8)$. 

Shenglong Zhou et al.
2.3 Stationary Equation

To express the solution of (2.1) more explicitly, we define

\[
T := T_r(x, \lambda) := \{ i \in \mathbb{N}_n : |x_i - \tau \nabla_i f(x)| \geq \sqrt{2\tau \lambda} \}.
\]

Based on the above set, we introduce the following stationary equation

\[
F_r(x; T) := \begin{bmatrix} \nabla_T f(x) \\ x_T \end{bmatrix} = 0.
\]

The relationship between (2.1) and (2.12) is revealed by the following theorem.

**Theorem 2** For any \( x \in \mathbb{R}^n \), by letting \( z := x - \tau \nabla f(x) \), we have

\[
x = \text{Prox}_{\tau \lambda} \|_{\ell_0} (z) \implies F_r(x; T) = 0 \implies x \in \text{Prox}_{\tau \lambda} \|_{\ell_0} (z).
\]

**Proof** If we have \( x = \text{Prox}_{\tau \lambda} \|_{\ell_0} (z) \), namely, \( \text{Prox}_{\tau \lambda} \|_{\ell_0} (z) \) is a singleton, then there is no index \( i \in T \) such that \( |z_i| = \sqrt{2\tau \lambda} \) by (2.2). This and (2.2) give rise to \((\text{Prox}_{\tau \lambda} \|_{\ell_0} (z))_T = z_T \). As a consequence,

\[
0 = x - \text{Prox}_{\tau \lambda} \|_{\ell_0} (z) \overset{(2.2)}{=} \begin{bmatrix} x_T \\ x_T \end{bmatrix} - \begin{bmatrix} z_T \\ 0 \end{bmatrix} = \begin{bmatrix} \tau \nabla_T f(x) \\ \tau \nabla_T f(x) \end{bmatrix},
\]

which suffices to obtain that \( F_r(x; T) = 0 \). We now prove the second claim. For any \( i \in T \), we have \( \nabla_i f(x) = 0 \) from (2.12) and thus \( |x_i| \geq \sqrt{2\tau \lambda} \) from (2.11). For any \( i \in T \), we have \( x_i = 0 \) from (2.12) and \( |\tau \nabla_i f(x)| = |x_i - \tau \nabla_i f(x)| < \sqrt{2\tau \lambda} \) from (2.11). Those together with Lemma 1 obtains the conclusion. \( \square \)

**Remark 1** We note that if \( \nabla f(0) = 0 \), then 0 is a \( \tau \)-stationary point of the problem (1.1), and even a global minimizer if \( f \) is convex. This case is trivial. However, we are more interested in the non-trivial case. Therefore from now on, we always suppose \( \nabla f(0) \neq 0 \) and denote

\[
\lambda := \min_i \left\{ \frac{\tau}{2} |\nabla_i f(0)|^2 : \nabla_i f(0) \neq 0 \right\}, \quad \overline{\lambda} := \max_i \frac{\tau}{2} |\nabla_i f(0)|^2.
\]

One can check that if \( \lambda \) is chosen to satisfy \( 0 < \lambda \leq \lambda \), then \( |0 - \tau \nabla_i f(0)| \geq \sqrt{2\tau \lambda} \) for any \( i \in J := \{ i \in \mathbb{N}_n : \nabla_i f(0) \neq 0 \} \), resulting in \( T_r(x, \lambda) = J \) in (2.11) and consequently, \( F_r(0; J) \neq 0 \) due to \( \nabla f(0) \neq 0 \). Namely, 0 is not a \( \tau \)-stationary point of (1.1). Hence, the trivial solution 0 is excluded.

On the other hand, if \( \lambda \) is chosen to satisfy \( \lambda > \overline{\lambda} \), then \( T_r(x, \lambda) = \emptyset \) in (2.11), leading to \( F_r(0; J) = 0 \), namely, 0 is a \( \tau \)-stationary point of (1.1). Therefore, for numerical experiments, \( \overline{\lambda} \) provides an upper bound for setting \( \lambda \) to an appropriate value.

3 Newton Method

Theorem 2 states that a point satisfying the stationary equation is a stronger condition than being a \( \tau \)-stationary point. The advantage of this equation allows us to cast a Newton-type algorithm based on its simple form.

3.1 Algorithm Design

To find a solution to the equation (2.12), we first need to locate the index set \( T \) that is unknown in general and then solve the equation. Therefore, we employ an adaptively updating rule as follows. For a computed point \( x^k \), we first calculate an approximation \( T_k \). Then, by fixing this set \( T_k \), we apply the Newton method on \( F_r(x; T_k) \) once into obtaining a direction \( d^k \). That is, \( d^k \) is a solution to the following equation system

\[
\nabla F_r(x^k; T_k)d = -F_r(x^k; T_k).
\]
The explicit formula of $F_k(x^k; T_k)$ from (2.12) implies that $d^k$ satisfies
\begin{align}
\nabla^2_{T_k} f(x^k) d^k_{T_k} = \nabla^2_{T_k} f(x^k) x^k_{T_k} - \nabla_{T_k} f(x^k),
\end{align}
\begin{align}
d^k_{T_k} = -x^k_{T_k}.
\end{align}

Now, let us consider the above formulas. The second part of $d^k$ can be derived directly without any difficulties. To find $d^k$, it is necessary to solve a linear equation with $|T_k|$ equations and $|T_k|$ variables. If a full Newton direction is taken, then the next iteration is $x^{k+1} = x^k + d^k = [(x^k_{T_k} + d^k_{T_k})^\top 0]^\top$. This means that the support set of $x^{k+1}$ will be located within $T_k$. Namely,
\begin{align}
supp(x^{k+1}) \subseteq T_k.
\end{align}

Based on this idea, we modify the standard rule associated with the Amijio line search $x^{k+1} = x^k + \alpha d^k$ as $x^{k+1} = x^k(\alpha)$, where
\begin{align}
x^k(\alpha) := \begin{bmatrix} x^k_{T_k} + \alpha d^k_{T_k} \\ x^k_{T_k} + d^k_{T_k} \\ 0 \end{bmatrix}.
\end{align}

For notational convenience, let
\begin{align}
J_k := T_{k-1} \setminus T_k, \quad S_k := \overline{T}_k \setminus T_{k-1},
\end{align}
\begin{align}
g^k := \nabla f(x^k), \quad H_k := \nabla^2_{T_k} f(x^k), \quad G_k := \nabla^2_{T_k \cup J_k} f(x^k).
\end{align}

We summarize the framework of the algorithm in Algorithm 1. From Algorithm 1, the following can be easily obtained:
\begin{align}
\begin{cases}
- \nabla^2_{T_k} f(x^k) = x^k_{T_k} & \text{supp}(x^{k+1}) \subseteq T_k, \\
\nabla^2_{T_k \cup J_k} f(x^k) = \begin{bmatrix} H_k & G_k \\
G_k^\top & \nabla^2_{J_k} f(x^k) \end{bmatrix} & \text{if } \nabla^2_{T_k + J_k} f(x^k) \text{ is solvable and its solution } d^k \text{ is Newton direction}.
\end{cases}
\end{align}

\begin{algorithm}

Algorithm 1 Newton method for the $\ell_0$-regularized optimization (NL0R)

If $\nabla f(0) = 0$, then return the solution 0 and terminate the algorithm. Otherwise, perform the following steps. Give parameters $\tau > 0, \delta > 0, \lambda \in (0, \Lambda), \sigma \in (0, 1/2), \beta \in (0, 1)$. Initialize $x^0$. Set $T_{-1} = \emptyset$ and $k \leftarrow 0$

while The halting conditions are violated do

Step 1. Set $T_k = \overline{T}_k$ if $S_k \neq \emptyset$, and $T_k = T_{k-1}$ otherwise, where $\overline{T}_k$ is computed by
\begin{align}
\overline{T}_k = \{ i \in N_n : |x_i^k - \tau g_i^k| \geq \sqrt{2\tau \lambda} \}.
\end{align}

Step 2. If (3.2) is solvable and its solution $d^k$ satisfies
\begin{align}
\langle g^k_{T_k}, d^k_{T_k} \rangle \leq -\delta \|d^k\|^2 + (1/4\tau)\|x^k_{T_k}\|^2,
\end{align}
then update $d^k$ by solving (3.2), namely by Newton direction,
\begin{align}
H_k d^k_{T_k} = G_k x^k_{T_k} - g^k_{T_k}, \quad d^k_{T_k} = -x^k_{T_k}.
\end{align}

Otherwise, update $d^k$ by Gradient direction
\begin{align}
\begin{cases}
- \nabla^2_{T_k} f(x^k) = x^k_{T_k} & \text{supp}(x^{k+1}) \subseteq T_k, \\
\nabla^2_{T_k \cup J_k} f(x^k) = \begin{bmatrix} H_k & G_k \\
G_k^\top & \nabla^2_{J_k} f(x^k) \end{bmatrix} & \text{if } \nabla^2_{T_k + J_k} f(x^k) \text{ is solvable and its solution } d^k \text{ is Newton direction}.
\end{cases}
\end{align}

Step 3. Find the smallest non-negative integer $m_k$ such that
\begin{align}
f(x^k(\beta^{m_k})) \leq f(x^k) + \sigma \beta^{m_k} \langle g^k, d^k \rangle.
\end{align}

Step 4. Set $\alpha_k = \beta^{m_k}, x^{k+1} = x^k(\alpha_k)$ and $k \leftarrow k + 1$.

end while

return $x^k$
\end{algorithm}
We emphasize that $J_k$ captures all nonzero elements in $x_k^{\hat{T}_k}$. This and (3.12) also allow us to explain that (3.2) is rewritten as (3.9). Therefore, we will see that $J_k$ instead of $T_k$ is used in the convergence analysis.

**Lemma 2** If $d^k$ is from (3.9), then we have

\begin{equation}
2\langle g^k_{\hat{T}_k}, d^k_{\hat{T}_k} \rangle + \langle d^k_{\hat{T}_k}, H_k d^k_{\hat{T}_k} \rangle = -\langle d^k_{T_k \cup J_k}, \nabla^2_{T_k \cup J_k} f(x^k) d^k_{T_k \cup J_k} \rangle + \langle d^k_{J_k}, \nabla^2_{J_k} f(x^k) d^k_{J_k} \rangle.
\end{equation}

**Proof** If $d^k$ is from (3.9), then we have the following chain of equations,

\begin{equation}
\langle d^k_{T_k \cup J_k}, \nabla^2_{T_k \cup J_k} f(x^k) d^k_{T_k \cup J_k} \rangle
= \begin{bmatrix}
d^k_{\hat{T}_k} \\
d^k_{J_k}
\end{bmatrix}^T
\begin{bmatrix}
H_k d^k_{\hat{T}_k} + G_k d^k_{J_k} \\
G_k^T d^k_{\hat{T}_k} + \nabla^2_{J_k} f(x^k) d^k_{J_k}
\end{bmatrix}
\end{equation}

\begin{equation}
= 2\langle d^k_{\hat{T}_k}, H_k d^k_{\hat{T}_k} - G_k x^k_{J_k} \rangle - \langle x^k_{J_k}, G_k^T d^k_{\hat{T}_k} \rangle + \langle d^k_{J_k}, \nabla^2_{J_k} f(x^k) d^k_{J_k} \rangle
\end{equation}

\begin{equation}
= -2\langle g^k_{\hat{T}_k}, d^k_{\hat{T}_k} \rangle - \langle d^k_{\hat{T}_k}, H_k d^k_{\hat{T}_k} \rangle + \langle d^k_{J_k}, \nabla^2_{J_k} f(x^k) d^k_{J_k} \rangle,
\end{equation}

which concludes our claim immediately.

\[ \square \]

Lemma 2 indicates that if $\nabla^2_{T_k \cup J_k} f(x^k)$ has a positive lower and upper bound, then $H_k$ is bounded from below and $\nabla^2_{J_k} f(x^k)$ is bounded from above. These bonds will enable (3.8) to be satisfied in each step under some properly chosen $\delta$ and $\tau$. This allows the Newton direction to be always imposed. Apparently, $\nabla^2_{T_k \cup J_k} f(x^k)$ being bounded from below can be guaranteed by some assumptions, such as the strong convexity of $f$; however, this is a strong assumption. To overcome this problem, the gradient direction compensates in the case when the condition (3.8) is violated.

**Remark 2** To end this subsection, we would like to show the computational complexity of NL0R. One can observe that the major computational cost arises from calculating the Hessian matrix $\nabla^2 f(x^k)$ and solving the equations (3.9). For the former, we only need its sub-matrices $H_k \in \mathbb{R}^{t_k \times t_k}$ and $G_k \in \mathbb{R}^{t_k \times J_k}$ with $t_k := |T_k|$ and $j_k := |J_k|$. This indicates that it is unnecessary to store/compute the entire Hessian matrix. For the latter the complexity is approximately $\mathcal{O}(t_k^4)$ since (3.9) has $t_k$ variables and $t_k$ equations. Overall, the whole computational complexity in each step depends on $t_k$ and $j_k$.

Note that if a problem admits a sparse solution $x^*$, namely, $||x^*||_0 \ll n$, then $t_k$ and $j_k$ can be quite small. For the CS examples (1.3), the Hessian matrix is $A^T A$, while $H_k = A_{T_k}^T A_{T_k}$ and $G_k = A_{J_k}^T A_{J_k}$. Then, the computational complexity is approximately $\mathcal{O}(\max\{m, t_k\} t_k \max\{t_k, j_k\})$, while $t_k$ and $j_k$ can be less than $0.05n$. Hence, NL0R can process relatively high-dimensional problems, as shown in the numerical experiments described in Section 4.2.

### 3.2 Global and quadratic convergence

As mentioned in Remark 1, if $\nabla f(0) = 0$, then 0 is a $\tau$-stationary point of the problem (1.1), and is even a global minimizer if $f$ is convex. However, this case is trivial. Therefore, we focus on the case of $\nabla f(0) \neq 0$ in Algorithm 1. Prior to presenting our main results, we define some parameters by

\begin{equation}
\pi := \min \left\{ \frac{1 - 2\alpha}{L/\delta - \alpha}, \frac{2(1 - \alpha)\delta}{L}, 1 \right\},
\end{equation}

\begin{equation}
\tau := \min \left\{ \frac{2\alpha \beta}{nL^2}, \frac{\alpha \beta}{4L}, \frac{1}{4L} \right\},
\end{equation}

\begin{equation}
\rho := \min \left\{ \frac{2\delta - n\tau L^2}{2}, \frac{2 - n\tau}{2} \right\}.
\end{equation}
Our first result shows that the direction in each step of NL0R is a descent with a decent declining rate, regardless of whether it is taken from the Newton or the gradient direction.

**Lemma 3 (Descent property)** Let $f$ be strongly smooth with $L > 0$ and $\tau, \rho$ be defined by (3.14). Then, for any $\tau \in (0, \overline{\tau})$, $\rho > 0$ and

\[
(3.15) \quad \langle g^k, d^k \rangle \leq -\rho \|d^k\|^2 - \frac{\tau}{2} \|g^k_{T_{k-1}}\|^2.
\]

**Proof** It follows from (3.14) that $\overline{\tau} \leq 1$ and thus $\overline{\tau} \beta < 1$ due to $\beta \in (0, 1)$. Hence $\overline{\tau} \leq \min \{2\delta / nL^2, 2 / n\}$, immediately showing that $\rho > 0$ if $\tau \in (0, \overline{\tau})$. In addition, if $d^k$ is updated by (3.9), then

\[
(3.16) \quad \|g^k_{T_k}\| \leq \|H_k d^k_{T_k} - G_k x^k_k\| = \|[H_k G_k]d^k_{T_k} \cup J_k\| \leq L\|d^k\|
\]

where the inequality holds because of $\|[H_k G_k]\|_2 \leq \|\nabla^2_{f(x_k)} f(x_k)\|_2 \leq L$ due to the strong smoothness of $f$ with the constant $L$. We now prove the conclusion for two cases.

**Case i** $S_k = \emptyset$. Step 1 in Algorithm 1 sets $T_k = T_{k-1}$. Consequently, $J_k = T_{k-1} \setminus T_k = \emptyset$ and $d^k_{T_k} = -x^k_{T_k} = 0$ from (3.12). If $d^k$ is updated by (3.9), then it follows that

\[
(3.17) \quad 2 \langle g^k, d^k \rangle = 2 \langle g^k_{T_k}, d^k_{T_k} \rangle = 2 \langle g^k_{T_k}, d^k_{T_k} \rangle = 2 \langle g^k_{T_k}, d^k_{T_k} \rangle \leq -2\delta \|d^k\|^2 + \|x^k_{T_k}\|^2 / (2\tau) = -2\delta \|d^k\|^2 \leq -2\delta \|d^k\|^2 + n\tau \|d^k_{T_k}\|^2 - \tau \|g^k_{T_k}\|^2
\]

where the last inequality holds due to $T_k = T_{k-1}$. If $d^k$ is updated by (3.10), then it follows from $d^k_{T_k} = -g^k_{T_k} = -g^k_{T_{k-1}}$ that

\[
(3.18) \quad 2 \langle g^k, d^k \rangle = 2 \langle g^k_{T_k}, d^k_{T_k} \rangle = 2 \langle g^k_{T_k}, d^k_{T_k} \rangle = 2 \langle g^k_{T_k}, d^k_{T_k} \rangle \leq -2\|d^k\|^2 + n\tau \|d^k_{T_k}\|^2 - \tau \|d^k_{T_k}\|^2 = -2\|d^k\|^2 - \tau \|g^k_{T_k}\|^2.
\]

**Case ii** $S_k \neq \emptyset$. For any $i \in S_k = T_k \setminus T_{k-1} = T_k \setminus T_{k-1}$, we have $x^k_i = 0$ because of $\text{supp}(x^k) \subseteq T_{k-1}$ by (3.3). Then, the definition of $T_k = \overline{T}_k$ in (3.7) gives rise to

\[
(3.19) \quad \forall i \in S_k, \quad \|g^k_i\|^2 = \|x^k_i - \tau g^k_i\|^2 \geq 2\tau \lambda \|x^k_j - \tau g^k_j\|^2, \quad \forall j \in J_k.
\]

This suffices to obtain the following chain of inequalities

\[
(3.19) \quad |J_k| / |S_k| \geq \|x^k_i - \tau g^k_i\|^2 = \|x^k_i - \tau g^k_i\|^2 \geq 2\tau \lambda \|x^k_j - \tau g^k_j\|^2, \quad \forall j \in J_k.
\]

Since $|J_k| / |S_k| \leq n$, the above inequalities establish our first fact

\[
(3.20) \quad -2 \langle x^k_j, g^k_j \rangle \leq n \tau \|g^k_j\|^2 - \tau \|g^k_{T_k}\|^2 - \|x^k_{T_k}\|^2 / \tau
\]
Lemma 4 (Existence and boundedness of $\alpha_k$) Let $f$ be strongly smooth with $L > 0$ and $\bar{\pi}, \bar{\tau}$ be defined by (3.14). Then,

\begin{equation}
(3.23) \quad f(x^k(\alpha)) \leq f(x^k) + \sigma \alpha \langle g^k, d^k \rangle
\end{equation}

holds for any $k \geq 0$ and any parameters

\[ 0 < \alpha \leq \bar{\pi}, \quad 0 < \delta \leq \min\{1, 2L\}, \quad 0 < \tau \leq \min\left\{ \frac{\alpha \delta}{nL^2}, \frac{\alpha}{n}, \frac{1}{4L} \right\}. \]

Moreover, for any $\tau \in (0, \bar{\tau})$, we have $\inf_{k \geq 0} \{ \alpha_k \} \geq \beta \tau > 0$.

Proof If $0 < \alpha \leq \bar{\pi}$ and $0 < \delta \leq \min\{1, 2L\}$, we have

\begin{equation}
(3.24) \quad \alpha \leq \frac{2(1 - \sigma)\delta}{L}, \quad \alpha \leq \frac{1 - 2\sigma}{L/\delta - \sigma} \leq \frac{1 - 2\sigma}{\max\{0, L - \sigma\}}.
\end{equation}

Since $f$ is strongly smooth, we obtain that

\[ 2f(x^k(\alpha)) - 2f(x^k) - 2\alpha \sigma \langle g^k, d^k \rangle \]

\begin{equation}
(2.4) \quad \leq 2(g^k, x^k(\alpha) - x^k) + L \|x^k(\alpha) - x^k\|^2 - 2\alpha \sigma \langle g^k, d^k \rangle
\end{equation}

\begin{equation}
(3.4) \quad = (1 - \sigma)2(g^k_{T_k}, d^k_{T_k}) - (1 - \alpha \sigma)2(g^k_{T_k}, x^k_{T_k}) + L[\alpha^2 \|d^k_{T_k}\|^2 + \|x^k_{T_k}\|^2]
\end{equation}

\begin{equation}
(3.12) \quad = (1 - \sigma)2(g^k_{T_k}, d^k_{T_k}) - (1 - \alpha \sigma)2(g^k_{T_k}, x^k_{T_k}) + L[\alpha^2 \|d^k_{T_k}\|^2 + \|x^k_{T_k}\|^2] =: \psi.
\end{equation}

To prove (3.23), one needs to show $\psi \leq 0$. Similar to the proof of Lemma 3, we consider two cases.

\textbf{Case i)} $S_k = \emptyset$. Step 1 in Algorithm 1 sets $T_k = T_{k-1}$, and thus $J_k = T_{k-1} \setminus T_k = \emptyset$. Then, we obtain

\[ \psi = \alpha(1 - \sigma)2(g^k_{T_k}, d^k_{T_k}) + L\alpha^2 \|d^k_{T_k}\|^2 \]
where the third inequality is due to $\delta \leq 1$, $\|d^k\|^2 = \|d^k_T\|^2$.

**Case ii) $S_k \neq \emptyset$.** If $d^k$ is from (3.9), then we have
\[
\psi \leq -(1 - \sigma)2\delta + (1 - \alpha\sigma)\delta\alpha + La^2 \quad \text{by } 1 - \alpha\sigma > 0, \tau \leq \alpha\delta/(nL^2).
\]
\[
= a [(L - \sigma\delta)\alpha - (1 - 2\sigma)\delta] \leq 0, \quad \text{by } L - \sigma\delta > 0, 1 - 2\sigma > 0, (3.24)
\]

If $d^k$ is updated by (3.10), namely $d^k_{T_k} = -g^k_{T_k}$, then
\[
\psi \leq -(1 - \sigma)2\delta + (1 - \alpha\sigma)\delta\alpha + La^2 \quad \text{by } 1 - \alpha\sigma > 0, \tau \leq \alpha\delta/(nL^2).
\]
\[
= a [(L - \sigma\delta)\alpha - (1 - 2\sigma)\delta] \leq 0, \quad \text{by } L - \sigma\delta > 0, 1 - 2\sigma > 0, (3.24)
\]

Thus, we verify (3.23). If $\tau \in (0, \bar{\tau})$, then for any $\alpha \in [\beta\bar{\tau}, \bar{\tau}]$, it follows that
\[
0 < \tau < \min \left\{ \bar{\tau}\delta/(nL^2), \bar{\tau}\beta/n, 1/(4L) \right\} 
\]
\[
\leq \min \left\{ \alpha\delta/(nL^2), \alpha/n, 1/(4L) \right\}.
\]
Therefore, (3.23) holds for any $\alpha \in [\beta\bar{\tau}, \bar{\tau}]$. Finally, the Armijo-type step size rule means that $\{\alpha_k\}$ must be bounded from below by $\beta\bar{\tau}$, that is,
\[
\inf_{k \geq 0} \{\alpha_k\} \geq \beta\bar{\tau} > 0.
\]
The whole proof is completed.
Lemma 4 allows us to conclude that the objective $f$ is strictly decreasing for each step, and the difference of two consecutive iterations and the entries of the stationary equation will vanish.

**Lemma 5** Let $f$ be strongly smooth with $L > 0$ and $\tau$ be defined by (3.14). Let $\{x^k\}$ be the sequence generated by NLOR with $\tau \in (0, \overline{\tau})$ and $\delta \in (0, \min\{1, 2L\})$. Then $\{f(x^k)\}$ is strictly nonincreasing sequence and

$$\lim_{k \to \infty} \max \left\{ \|F_r(x^k; T_k)\|, \|x^{k+1} - x^k\|, \|g_T^k\|, \|g_T^k\| \right\} = 0. \tag{3.27}$$

**Proof** By (3.23), (3.15) and denoting $c_0 := \sigma \rho \beta$, we have

$$f(x^{k+1}) - f(x^k) \leq \sigma \alpha_k \langle d^k, g^k \rangle \leq -\sigma \alpha_k \rho \|d^k\|^2 - \frac{\tau}{2} \|g_T^{k-1}\|^2 \leq -c_0 \|d^k\|^2 - \frac{\tau}{2} \|g_T^{k-1}\|^2. \tag{3.26}$$

Then, it follows from the above inequality that

$$\sum_{k=0}^{\infty} [c_0 \|d^k\|^2 + \frac{\tau}{2} \|g_T^{k-1}\|^2] \leq \sum_{k=0}^{\infty} \left[ f(x^k) - f(x^{k+1}) \right] = \left[ f(x^0) - \lim_{k \to \infty} f(x^k) \right] < +\infty,$$

where the last inequality is due to $f$ being bounded from below. Hence $\|d^k\| \to 0, \|g_T^{k-1}\| \to 0$, which suffices to obtain that $\|x^{k+1} - x^k\| \to 0$ because of

$$\|x^{k+1} - x^k\|^2 \leq \alpha_k^2 \|d_T^k\|^2 + \|x_T^k\|^2 \leq \|d_T^k\|^2 + \|x_T^k\|^2 = \|d^k\|^2. \tag{3.4}$$

The above relation also indicates that $\|x_T^k\|^2 \to 0$. In addition, if $d^k$ is taken from (3.9), then $\|g_T^k\| \leq L \|d^k\| \to 0$ by (3.16). If it is taken from (3.10) then $\|g_T^k\| = \|d_T^k\| \to 0$. These results together with (2.12) imply that $\|F_r(x^k; T_k)\|^2 = \|g_T^k\|^2 + \|x_T^k\|^2 \to 0$, finishing the whole proof. \qed

We are ready to conclude from Lemma 5 that the index set $T_k$ can be identified within finite steps and the sequence converges to a $\tau$-stationary point or a local minimizer globally that are presented by the following theorem.

**Theorem 3** (Convergence and identification of $T_k$) Let $f$ be strongly smooth with $L > 0$ and $\tau$ be defined by (3.14). Let $\{x^k\}$ be the sequence generated by NLOR with $\tau \in (0, \overline{\tau})$, $\delta \in (0, \min\{1, 2L\})$. The following results hold.

1) For any sufficiently large $k$, $T_k \equiv T_{k-1} \equiv T_\infty$.

2) Any accumulating point (say $x^*$) of the sequence satisfies

$$\nabla_{T_\infty} f(x^*) = 0, \quad x^*_{T_\infty} = 0, \quad \text{supp}(x^*) \subseteq T_\infty \tag{3.28}$$

and is non-trivial ($x^* \neq 0$), and it is a $\tau_*$-stationary point of (1.1) with

$$0 < \tau_* < \min \left\{ \overline{\tau}_*, \min_{i \in \text{supp}(x^*)} \left\{ |x^*_i|/(2\lambda) \right\} \right\}. \tag{3.29}$$

3) If $x^*$ is isolated, then the whole sequence converges to $x^*$.

**Proof** 1) For any sufficiently large $k$, $T_k \equiv T_{k-1}$ indicates $S_k = 0$ by Step 1 in Algorithm 1. Suppose there is a subsequence $K \subseteq \{0, 1, 2, \ldots \}$ such that $S_k \neq 0, k \in K$. Then, we have $S_k = T_k \setminus T_{k-1} = T_k \setminus T_{k-1} \neq \emptyset, k \in K$. Lemma 5 shows that $g_T^k \to 0$, yielding $g_S^k \to 0$. This contradicts $|\tau g_T^k| \geq \sqrt{2\tau} \lambda_i, i \in S_k$ by (3.19).

2) Let $\{x^{k_i}\}$ be the convergent subsequence of $\{x^k\}$ that converges to $x^*$. Since $x^{k_i} \to x^*$ and $\|x^{k_i+1} - x^k\| \to 0$ from Lemma 5, we have $x^{k_i+1} \to x^*$ and thus supp($x^*$) $\subseteq$ supp($x^{k_i+1}$) for sufficiently large $k_i$. Then, it follows from supp($x^{k_i+1}$) $\subseteq$ $T_{k_i}$ $\equiv$ $T_\infty$ by (3.3) and claim 1) that supp($x^*$) $\subseteq$ supp($x^{k_i+1}$) $\subseteq$ $T_\infty$. Moreover,\n
$$\nabla_{T_\infty} f(x^*) = \nabla_{T_{k_i}} f(x^{k_i}) = \lim_{k_i \to \infty} \nabla_{T_{k_i}} f(x^{k_i}) = \lim_{k_i \to \infty} g_T^{k_i} = 0. \tag{3.27}$$
Overall, (3.28) is true. Next, we claim that \( x^* \neq 0 \). Suppose \( x^* = 0 \). Algorithm 1 runs infinite steps only when \( \nabla f(0) \neq 0 \). By \( \lambda \in (0, \bar{\lambda}) \) and \( x_{k_i} \to x^* = 0 \), for sufficiently large \( k \), there is a sufficiently small \( \varepsilon > 0 \) satisfying

\[
|x_{k_i}^* - \tau g_{k_i}^*| \geq \tau|\nabla f(0)| - |x_{k_i}^*| - \tau|\nabla f(0) - \nabla f(x_{k_i}^*)| \geq \sqrt{2\tau \overline{\lambda}} - \varepsilon \geq \sqrt{2\tau \lambda}.
\]

(3.31)

Thus, \( \tilde{T}_{k_i} \neq \emptyset \). Recall that in claim 1), \( S_k = \emptyset \) for any sufficiently large \( k \). This implies \( \tilde{T}_{k_i} \subseteq T_{k_i} = T_{k_i} \equiv T_\infty \).

However, by (3.30), \( g_{k_i}^* \to 0 \) and \( x_{k_i}^* \to 0 \), contradicting with (3.31). Thus, \( x^* \neq 0 \). Now, by (3.29), it is easy to check that

\[
T_* := \text{supp}(x^*) = \{ i \in \mathbb{N}_n : x_i^* \neq 0 \} \equiv \{ i \in \mathbb{N}_n : |x_i^* - \tau_i \nabla_i f(x^*)| \geq \sqrt{2\tau \lambda} \}.
\]

This together with \( \nabla_{T_*} f(x^*) = 0 \), \( x_{T_*}^* = 0 \) from (3.28) suffices to obtain that \( F_{T_*}(x^*; T_*) = 0 \). Finally, Theorem 2 allows us to claim that \( x^* \) is a \( \tau \)-stationary point.

3) The whole sequence converges because of \( x^* \) being isolated, \([37, \text{Lemma 4.10}]\) and \( \|x^{k+1} - x^k\| \to 0 \) from Lemma 5.

Finally, we would like to see how fast our proposed method NL0R converges. Here, we need some assumptions. We say the Hessian of \( f \) is locally Lipschitz continuous around \( x^* \) with a constant \( M_* > 0 \) if

\[
\|\nabla^2 f(x) - \nabla^2 f(x')\|_2 \leq M_*\|x - x'\|
\]

for any points \( x, x' \) in the neighbourhood of \( x^* \). In addition, we also need that \( f \) is locally strongly convex with a constant \( \ell_* > 0 \) around \( x^* \). As we mentioned before, the constants \( M_* \) and \( \ell_* \) depend on the point \( x^* \).

Now, we are able to establish the following results.

**Theorem 4 (Global and quadratic convergence)** Let \( \{x^k\} \) be the sequence generated by NL0R with \( \tau \in (0, \tau) \) and \( \delta \in (0, \min\{1, \ell_*\}) \) and \( x^* \) be one of its accumulating points. Suppose \( f \) is strongly smooth with constant \( L > 0 \) and locally strongly convex with \( \ell_* > 0 \) around \( x^* \). Then, the following results hold.

1) The whole sequence converges to \( x^* \), a strictly local minimizer of (1.1).

2) The Newton direction is always accepted for sufficiently large \( k \).

3) Furthermore, if the Hessian of \( f \) is locally Lipschitz continuous around \( x^* \) with constant \( M_* > 0 \), then for sufficiently large \( k \),

\[
\|x^{k+1} - x^*\| \leq M_*/(2\ell_*)\|x^k - x^*\|^2.
\]

**Proof** 1) Denote \( T_* := \text{supp}(x^*) \). Theorem 3 shows that \( \nabla_{T_*} f(x^*) = 0 \) and \( x^* \neq 0 \). Consider a local region \( N(x^*) := \{ x \in \mathbb{R}^n : |x - x^*| < \epsilon_* \} \), where

\[
\epsilon_* := \min \left\{ \lambda/(2\|\nabla_{T_*} f(x^*)\|), \min_{i \in E_0} |x_i^*| \right\}
\]

For any \( x(\neq x^*) \in N(x^*) \), we have \( T_* \subseteq \text{supp}(x) \). In fact if there is a \( j \) such that \( j \in T_* \) but \( j \notin \text{supp}(x) \), then we derive a contradiction:

\[
\epsilon_* \leq \min_{i \in T_*} |x_i^*| \leq |x_j^*| = |x_j^* - x_j| \leq \|x - x^*\|_2 < \epsilon_*.\]

Since \( f \) is locally strongly convex with \( \ell_* > 0 \) around \( x^* \), for any \( x(\neq x^*) \in N(x^*) \), it is true that

\[
f(x) + \lambda|x||0 - f(x^*) - \lambda||x^*||0 \\
\geq (\nabla f(x^*), x - x^*) + (\ell_*/2)||x - x^*||^2 + \lambda||x||0 - \lambda||x^*||0 \\
= (\nabla_{T_*} f(x^*), x_{T_*}) + (\ell_*'/2)||x - x^*||^2 + \lambda||x||0 - \lambda||x^*||0 := \phi.
\]
where the first equality is due to $\nabla_T f(x^*) = 0$. Clearly, if $T_* = \text{supp}(x)$, then $x_{T_*} = 0$, $\|x\|_0 = \|x^*\|_0$ and hence $\phi = (\ell_/2)\|x - x^*\|^2 > 0$. If $T_* \neq (\subseteq)\text{supp}(x)$, then $\|x\|_0 \geq \|x^*\|_0 + 1$ and thus we obtain
\[
\phi \geq -\|\nabla_T f(x^*)\|\|x_{T_*}\| + (\ell_/2)\|x - x^*\|^2 + \lambda \geq -\|\nabla_T f(x^*)\|\|x - x^*\| + (\ell_/2)\|x - x^*\|^2 + \lambda \geq -\lambda/2 + (\ell_/2)\|x - x^*\|^2 + \lambda > 0.
\]
Both cases show that $x^*$ is a strictly local minimizer of (1.1) and is unique in $N(x^*)$, namely, $x^*$ is isolated local minimizer in $N(x^*)$. Therefore, the whole sequence tends to $x^*$ by Theorem 3.3).

2) We first verify that $H_k$ is nonsingular when $k$ is sufficiently large and
\[
\langle g_k^k, d_k^k \rangle \leq -\delta \|d_k\|^2 + \|x_k^k\|^2/(4\tau).
\]
Since $f$ is strongly smooth with $L$ and locally strongly convex with $\ell_*$ around $x^*$, it follows that
\[
\ell_* \leq \lambda_i(\nabla_{T_{\ldots \cup \ldots}}^2 f(x^k)), \lambda_i(H_k), \lambda_i(\nabla_{T_{\ldots \cup \ldots}}^2 f(x^k)) \leq L,
\]
where $\lambda_i(A)$ is the $i$th largest eigenvalue of $A$. Direct verification yields that
\[
2\langle g_k^k, d_k^k \rangle \overset{(3.13)}{=} -\langle d_k^k, \nabla_{T_{\ldots \cup \ldots}}^2 f(x^k) d_k^k \rangle - \langle H_k d_k^k, d_k^k \rangle + \langle d_k^k, \nabla_{J_k}^2 f(x^k) d_k^k \rangle \leq -\ell_* \|d_{T_k^\cup J_k}^k\|^2 + \ell \|d_k^k\|^2 + L\|x_k^k\|^2 \leq -2\ell_* \|d_{T_k^\cup J_k}^k\|^2 + \ell \|d_k^k\|^2 + L\|x_k^k\|^2 \leq -2\ell_* \|d_k^k\|^2 + 2L\|x_k^k\|^2 \leq -2\delta \|d_k^k\|^2 + \|x_k^k\|^2/(2\tau),
\]
where the last inequality is due to $\delta \leq \ell_*$ and $\tau < \tau \leq 1/(4L)$. This proves that $d_k^k$ from (3.9) is always admitted for sufficiently large $k$.

3) By Theorem 3.2), for sufficiently large $k$, we have (3.28), namely,
\[
x_{T_k}^k = 0, \quad \nabla_{T_k} f(x^*) = 0.
\]
For any $0 \leq t \leq 1$, by letting $x(t) := x^* + t(x_k - x^*)$, the Hessian of $f$ being locally Lipschitz continuous at $x^*$ derives
\[
\|\nabla_{T_k}^2 f(x^k) - \nabla_{T_k}^2 f(x(t))\|_2 \leq M_* \|x^k - x(t)\| = (1 - t)M_\tau \|x^k - x^*\|.
\]
Moreover, by Taylor expansion, we obtain
\[
\nabla f(x^k) - \nabla f(x^*) = \int_0^1 \nabla^2 f(x(t))(x^k - x^*)dt.
\]
Now, we have the following chain of inequalities
\[
\|x^{k+1} - x^*\|^2 = \|x_k^{k+1} - x_k^*\|^2 + \|x_k^{k+1} - x_k^*\|^2 \overset{(3.4, 3.34)}{=} \|x_k^{k+1} - x_k^*\|^2 = \|x_k^{k+1} - x_k^* + \alpha_k d_k^k\|^2 \leq (1 - \alpha_k)\|x_k^k - x_k^*\|^2 + \alpha_k \|x_k^k - x_k^* + d_k^k\|^2 \overset{(3.37)}{=} (1 - \alpha_k)\|x_k^k - x_k^*\|^2 + \alpha_k \|x_k^k - x_k^* + d_k^k\|^2 \leq (1 - \alpha_k)\|x^k - x^*\|^2 + \alpha_k \|x_k^k - x_k^* + d_k^k\|^2 \overset{(3.26)}{=} (1 - \alpha_k)\|x^k - x^*\|^2 + \alpha_k \|x_k^k - x_k^* + d_k^k\|^2.
\]
where (3.37) is due to \( \| \cdot \|^2 \) being a convex function. From 2), \( d^k \) is always updated by (3.9) for sufficiently large \( k \). Therefore, we have
\[
\ell_* \| x^k - x^* + d^k \| \leq \ell_* \| H^k_0^{-1}(\nabla^2 f(x^k)x^k - g_{T_k}) + x^k - x^*_{T_k} \|
\]
\[
\leq \| \nabla^2 f(x^k)x^k - g_{T_k} - H_k x^*_{T_k} \|
\]
\[
\leq \| \nabla^2 f(x^k)x^k - g_{T_k} - H_k x^* \|
\]
\[
\leq \| \nabla^2 f(x^k)(x^k - x^*) - \int_0^1 \nabla^2 f(x(t))(x^k - x^*)dt \|
\]
\[
\leq \| \int_0^1 [\nabla^2 f(x(t)) - \nabla^2 f(x(t))] (x^k - x^*) dt \|
\]
\[
\leq \| x^k - x^* \|^2 \int_0^1 (1-t) dt
\]
\[
\leq 0.5M_* \| x^k - x^* \|^2.
\]
It follows from \( d^k_{T_k} = -x^k_{T_k} \) and (3.34) that \( \| x^k + d^k - x^* \| = \| x^k_{T_k} + d^k_{T_k} - x^*_{T_k} \| \), leading to the following fact
\[
\| x^k + d^k - x^* \| \frac{\| x^k + d^k - x^* \|}{\| x^k - x^* \|} \leq \frac{\| x^k_{T_k} + d^k_{T_k} - x^*_{T_k} \|}{\| x^k - x^* \|} \leq \frac{M_* \| x^k - x^* \|^2}{2\ell_* \| x^k - x^* \|} \to 0.
\]
Now, we have three facts: (3.40), \( x^k \to x^* \) from 1), and \( \langle \nabla f(x^k), d^k \rangle \leq -\rho\| d^k \|^2 \) from Lemma 3, which together with [26, Theorem 3.3] allow us to claim that eventually the step size \( \alpha_k \) determined by the Armijo rule is 1, namely \( \alpha_k = 1 \). Then, it follows from (3.37) that
\[
\| x^{k+1} - x^* \|^2 \leq (1-\alpha_k)\| x^k_{T_k} - x^*_{T_k} \|^2 + \alpha_k \| x^k_{T_k} - x^*_{T_k} + d^k_{T_k} \|^2
\]
(3.41)
\[
= \| x^k_{T_k} - x^*_{T_k} + d^k_{T_k} \|^2 \leq (0.5M_*\ell_*)^2 \| x^k - x^* \|^4.
\]
Namely, the sequence converges quadratically, completing the proof. \( \square \)

4 Numerical Experiments

In this part, we will conduct extensive numerical experiments for our algorithm NL0R\(^1\) by using MATLAB (R2019a) on a laptop with 32GB memory and 2.3Ghz CPU for solving the CS problems and the sparse linear complementarity problems.

4.1 Implementation of NL0R

We initialize NL0R with \( x^0 = 0 \) so that \( T_0 \) in (3.7) is non-empty if \( \lambda \in (0, \Delta) \). The halting condition is set up as follows.

4.1.1 Halting conditions

If a point \( x^k \) satisfies \( \text{supp}(x^k) \subseteq T_k = T_{k-1}, \nabla T_k f(x^k) = 0 \) and \( x^k_{T_k} = 0 \), then similar reasoning to prove Theorem 3 2) allows us to show that it is also a \( \tau \)-stationary point of (1.1) with \( 0 < \tau < \min_i \{ |x_i^k|/(2\lambda), i \in \text{supp}(x^k) \} \). Therefore, it is reasonable to terminate NL0R at the \( k \)th step if either \( k \) reaches the maximum number of iterations (e.g., 2000) or \( x^k \) satisfies \( \text{supp}(x^k) \subseteq T_k = T_{k-1} \) and \( \| F_{T_k}(x;T_k) \| \leq 10^{-6} \), where \( \tau_k \) defined in (4.1).

\(^1\)Available at https://github.com/ShenglongZhou/NL0R
4.1.2 Selection of parameters

We fix $\sigma = 5 \times 10^{-5}$ and $\beta = 0.5$, while for $\lambda$, $\delta$ and $\tau$, the empirical numerical experience have indicated that a better strategy is to update them adaptively. Note that conditions in Theorem 4 are sufficient but unnecessary. Therefore, in practice there is no need to set the parameters to strictly meet these conditions.

More precisely, Theorem 4 states that any positive $\delta \in (0, \min\{1, \ell\})$ is acceptable. However, it is suggested that $\delta$ should be relatively small [21,27] so that more Newton directions will be taken, namely, there is a greater likelihood of the occurrence of condition (3.8). To demonstrate this, we tested NL0R under different choices of $\delta$. As shown in Figure 1, for each $\delta \in [10^{-8}, 10^{-1}]$, there were no violations of the condition (3.8), that is, the Newton directions were taken in all steps. Whereas when $\delta \in (10^{-1}, 10^0]$, an increasing number of violations occurred, resulting in increasing number of gradient directions that were taken during the process. It is observed that a greater value of $\delta$ corresponds to a greater number of iterations.

On the other hand, the condition $0 < \tau < \tau = 2\pi\delta/\beta^2$ from (3.14) suggests that $\tau$ should be small enough if $\delta$ is chosen to be small. However, $\hat{T}_k$ will not vary too much in (3.7) if a sufficiently small $\tau$ is selected at the beginning. This may push NL0R to fall in a local area rapidly, clearly degrading the performance of the algorithm. Therefore, we set

$$
\delta := \delta_k = \begin{cases} 
10^{-8}, & \text{if } S_k = \emptyset, \\
10^{-4}, & \text{if } S_k \neq \emptyset. 
\end{cases}
$$

Even though Theorem 4 provides a clue for choosing $0 < \tau < \tau$, it is still difficult to determine the proper value because $L$ is generally not easy to compute. An alternative is to update $\tau$ adaptively. Typically, we use the following rule: start $\tau$ with a fixed scalar $\tau_0$ (e.g., $\tau_0 = 1/2$ if no extra explanations are given) and then update it according to

$$
\tau_{k+1} = \begin{cases} 
4\tau_k/5, & \text{if } k/10 = [k/10] \text{ and } \|F_{\gamma_k}(x^k; T_k)\| > k^{-2}, \\
5\tau_k/4, & \text{if } k/10 = [k/10] \text{ and } \|F_{\gamma_k}(x^k; T_k)\| \leq k^{-2}, \\
\tau_k, & \text{otherwise.}
\end{cases}
$$

(4.1)
4.1.3 Tuning the penalty parameter $\lambda$

It is suggested to set $\lambda \in (0, \lambda)$ in Algorithm 1 to avoid a trivial solution 0, where $\lambda$ is given by (2.13). However, $\lambda$ may incur a very small $\lambda$ and thus a big size $|\tilde{T}_k|$ by (3.7). Recall that the complexity of deriving the Newton direction by (3.9) is at least approximately $O(|\tilde{T}_k|^3)$. Therefore, a small $\lambda$ not only increases the computational complexity but also results in a solution that is not sparse enough. On the other hand, as mentioned in Remark 1, a too large value of $\lambda$ (e.g. $\lambda > \lambda$ defined in (2.13)) will result in a trivial solution 0. To balance these two aspects, we start with a slightly bigger value of $\lambda$ as $\lambda_0 := \max\{\lambda, c\}$ and gradually reduce it through $\lambda_k = r\lambda_{k-1}$, where $r, c \in (0, 1]$. We pick $r = 0.5$ and $c = 0.5$ in our numerical experiments if no additional explanations are provided.

To determine the performance of NL0R under fixing $\lambda = \lambda_0$ or updating $\lambda = \lambda_k$, two instances of Example 2 are tested and the obtained results are shown in Figure 2. It is clearly observed that $\|F_k(x^k, T_k)\|$ declines dramatically for both fixing $\lambda = \lambda_0$ and updating $\lambda = \lambda_k$, indicating that NL0R enjoys a quadratic convergence property. The objective $f(x^k)$ produced by NL0R under fixing $\lambda = \lambda_0$ stabilizes at a certain level, implying that it achieves a local minimum. By contrast, NL0R under updating $\lambda = \lambda_k$ delivers the objective $f(x^k)$ that decreases sharply and approaches a globally optimal value. Therefore, the updating rule enables better performance of NL0R and is henceforth adopted for carrying out our numerical comparisons in the sequel.

![Fig. 2: Two strategies for setting $\lambda$ in NL0R for solving Example 2. The sub-figures in the top (bottom) row are produced under two scenarios.](image)

4.2 Compressed sensing

CS has seen revolutionary advances both in theory and algorithm development over the past decade. Ground-breaking papers that pioneered the advances are [23, 15, 16]. We will focus on two types of data: the randomly
generated data and the 2-dimensional image data. For the first data, we consider the exact recovery \( y = Ax \), where the sensing matrix \( A \) chosen as in [50,56], while for the image data, we consider the inexact recovery \( y = Ax + \xi \), where \( \xi \) is the noise and \( A \) will be described in Example 3.

**Example 2 (Random data)** Let \( A \in \mathbb{R}^{m \times n} \) be a random Gaussian matrix with each column being identically and independently distributed (i.i.d.) samples of the standard normal distribution. We then normalize each column to be a unit length. Next, the \( s_0 \) non-zero components of the ‘ground truth’ signal \( x^* \) are also iid samples of the standard normal distribution, and their locations are picked randomly. Finally, the measurement is given by \( y = Ax^* \).

**Example 3 (2-D image data)** Some images are naturally not sparse themselves but can be sparse under some wavelet transforms. Here, we take advantage of the Daubechies wavelet 1, denoted as \( W(\cdot) \). Then the images under this transform (i.e., \( x^* := W(\omega) \)) are sparse, and \( \omega \) is the vectorized intensity of an input image. Therefore, the explicit form of the sampling matrix may not be available. We consider a sampling matrix taking the form \( A = FW^{-1} \), where \( F \) is the partial fast Fourier transform, and \( W^{-1} \) is the inverse of \( W \). Finally, the added noise \( \xi \) has each element \( \xi_i \sim n_f \cdot N \) where \( N \) is the standard normal distribution and \( n_f \) is the noise factor. Three typical choices of \( n_f \) are considered, namely \( n_f \in \{0.01, 0.05, 0.1\} \). For this experiment, we compute a gray image (see the original image in Figure 4) with size \( 512 \times 512 \) (i.e. \( n = 512^2 = 262144 \)) and the sampling size \( m = 20033 \) and 29729, respectively.

### 4.2.1 Comparisons for random data

Since a large number of state-of-the-art methods have been proposed to solve the CS problems, it is far beyond the scope of our paper to compare all of them to our method. For a fair comparison, we only focus on the algorithms (often referred to as regularized methods) that aim to solve (1.1) or its relaxations, where \( \ell_0 \) norm is replaced by some approximations such as \( \ell_q(0 < q \leq 1) \) [32] or \( \ell_1 - \ell_2 \) [34]. We note that greedy methods mentioned in Subsection 1.1, for the model (1.2) with \( s \) being given, have been famous for the super-high computational speed and the high order of accuracy when \( s \) is small relative to \( n \). However, we will not compare them with NL0R since we would like to consider the scenario when \( s \) is unknown. We select MIRL1 [56], AWL1 [34, ADMM for weighted \( \ell_{1-2} \)] which is a faster approximation of the method proposed in [50], IRLS [32] (we choose \( q = 1/2 \)) and PDASC [31]. All parameters are set to their default values except for setting the maximum iteration number to 100 and removing the final refinement step for MIRL1 and \( \text{del}=1\text{e}^{-8} \) for PDASC. We note that PDASC and NL0R are second-order methods and the other three methods are first-order methods.

To examine the accuracy of the solutions and the speed of these five methods, we run 20 trials with medium dimensions \( n \) increasing from 10000 to 30000 and keeping \( m = \lceil 0.25n \rceil \), \( s_0 = \lceil 0.01n \rceil \) or \( s_0 = \lceil 0.05n \rceil \). Average results are reported in Figure 3, where \( s_0 = \lceil 0.01n \rceil \), and Table 1, where \( s_0 = \lceil 0.05n \rceil \). As shown in Figure 3, NL0R always generates the smallest \( \|x - x^*\| \), and the most accurate recovery, with accuracy order at least \( 10^{-14} \), followed by PDASC. By contrast, the other three methods obtain accuracy with the order above \( 10^{-5} \). These results clearly show that the second-order methods are advantageous for obtaining a higher order of accuracy. For the computational speed, it is clearly observed that NL0R always runs the fastest and requires only approximately 2 seconds when \( n = 30000 \). PDASC is the second-fastest method. This shows that for problems in higher dimensions, NL0R and PDASC can run faster than the first-order methods. Similar results are observed in Table 1. To summarize, NL0R delivers the most accurate recovery within the shortest time.

### 4.2.2 Comparisons for 2-D image data

In Example 3, the data size \( n \) is relatively large, possibly making most regularized methods extremely slow. Hence, we select three greedy methods CSMP (denoted for CoSaMP) [38], HTP [28] and AIHT [11] as well as PDSCA.
Fig. 3: Average recovery error and time of five methods for Example 2.

Table 1: Performance of five methods for Example 2.

| n     | \|x - x^*\| | Time (in seconds) |
|-------|---------------|------------------|
|       | \|x - x^*\|   | 10000 15000 20000 25000 30000 | 10000 15000 20000 25000 30000 |
| AWL12 | 8.38e-05      | 17.71            | 42.70            |
| RSLQ  | 4.32e-04      | 7.045            | 12.00            |
| MRIL1 | 1.57e-02      | 4.06e-04         | 23.21            |
| PDASC | 5.36e-14      | 0.972            | 2.290            |
| NL0R  | 1.16e-14      | 0.602            | 1.363            |

It is important to mention that despite the large dimensions of \(m\) and \(n\), \(A : \mathbb{R}^n \rightarrow \mathbb{R}^m\) and its transpose \(A^\top : \mathbb{R}^m \rightarrow \mathbb{R}^n\) are mappings that enable the acceleration of the calculations of \(A(x)\) and \(A^\top(y)\). This and the fact in Remark 2 that only some small sub-parts of the Hessian are needed by NL0R can lead to significantly savings in computer memory requirements.

As suggested in the PDSCA package, we set another rule to stop each method if at \(k\)th iteration it satisfies \(\|Ax^k - y\| \leq \|Ax^* - y\|\) to accelerate the termination. Moreover, for fair comparison, we first run PDSCA that can deliver a solution with a good sparsity level \(s\). Then, we set this sparsity level \(s\) for CSMP, HTP and AIHT since they need such prior information. Let \(x\) be a solution produced by a method. Apart from reporting the sparsity level \(\|x\|_0\) and the CPU time of a method, we also compute the peak signal to noise ratio (PSNR) defined by

\[
\text{PSNR} := 10 \log_{10} \left( \frac{\|x\| - \|x^*\|^2}{\|x^*\|^2} \right)
\]

to measure the performance of the method. Note that a larger PSNR, closer \(x\) approaches to the true image \(x^*\), namely a better performance. The results for Example 3 are presented in Figure 4 and Table 2, where it is observed that SPDSA offers the largest PSNR when \(nf=0.01\), whilst NL0R produces the largest PSNR for \(nf=0.05\) and \(nf=0.1\), implying that our method is more robust to noise. In addition, NL0R runs the fastest and renders the sparsest representations for most cases.

4.3 Sparse linear complementarity problem

Sparse linear complementarity problems have been applied in real-world applications such as bimatrix games and portfolio selection problems [19, 49, 42]. These problems aim to find a sparse vector \(x \in \mathbb{R}^n\) from \(\Omega := \{x \in \mathbb{R}^n : x \geq 0, \ Mx + q \geq 0, \ \langle x, Mx + q \rangle = 0\}\), where \(M \in \mathbb{R}^{n \times n}\) and \(q \in \mathbb{R}^n\). A point \(x \in \Omega\) is equivalent to

\[
f(x) := \sum_{i=1}^{n} \phi(x_i, M_i x + q_i) = 0,
\]
Fig. 4: Recovery results for Example 3 with $m = 20033$ and $n_f = 0.1$.

Table 2: Performance of five methods for Example 3.

| Method | $n_f = 0.01$ | | $n_f = 0.05$ | | $n_f = 0.1$ |
|--------|--------------|----------------|--------------|----------------|--------------|
|        | PSNR | Time | $\|x\|_0$ | PSNR | Time | $\|x\|_0$ | PSNR | Time | $\|x\|_0$ |
| SPDSA  | 21.62 | 15.53 | 9716 | 20.11 | 8.45 | 5982 | 19.60 | 5.72 | 2969 |
| AIHT   | 19.81 | 148.5 | 9716 | 20.15 | 2.23 | 5982 | 20.26 | 19.3 | 2969 |
| HTP    | 19.66 | 19.15 | 9716 | 20.27 | 3.40 | 5982 | 20.57 | 3.41 | 2969 |
| SCMP   | 12.49 | 51.54 | 9716 | 18.44 | 63.1 | 5982 | 16.35 | 14.8 | 2969 |
| NL0R   | 23.21 | 7.130 | 9690 | 21.91 | 4.43 | 4173 | 20.93 | 3.07 | 2803 |

where $M_i$ is the $i$th row of $M$ and $\phi$ is the so-called NCP function that is defined by $\phi(a, b) = 0$ if and only if $a \geq 0, b \geq 0, ab = 0$. We take advantage of an NCP function $\phi(a, b) = a_+^2 + b_+^2 + (-a)_+^2 + (-b)_+^2$, where $a_+ := \max\{a, 0\}$, and a testing example from [54].

Example 4 Let $M = ZZ^\top$ with $Z \in \mathbb{R}^{n \times m}$ and $m \leq n$ (e.g. $m = n/2$). Elements of $Z$ are iid samples from the standard normal distribution. Each column is then normalized to have a unit length. The ‘ground truth’ sparse solution $x^*$ with a sparsity level $s_*$ is produced in the same manner as in Example 2 and $q$ is obtained by $q_i = -(Mx^*)_i$ if $(Mx^*)_i > 0$ and $q_i = |(Mx^*)_i|$ otherwise.

Since there are very few methods that have been proposed to process the sparse LCP, we only select two solvers: the half-thresholding projection (HTP) method [43] and LEMKA’s method (LEMKE$^2$). We vary the sample size $n$ but fix $m = n/2, s_* = 0.01n$ and $s_* = 0.05n$. The average results over 20 trials are reported

*http://ftp.cs.wisc.edu/math-prog/matlab/lemke.m
in Figure 5 where $s_*=0.01n$ and Table 3 where $s_*=0.05n$. Comparing with HTP, LEMKE and NL0R produce much more accurate solutions because their obtained objective function values $f(x)$ and the recovered accuracy $||x-x^*||$ tend to be close to zero. With regard to the computational speed, the results are significantly different. As shown in Figure 5, NL0R runs super-fast, followed by LEMKE, and HTP is the slowest. Similar results are observed in Table 3, where for $n=20000$, NL0R only consumes approximately 8.826 seconds while LEMKE requires 531.1 seconds and HTP requires 207.9 seconds. Therefore, NL0R evidently outperforms the other methods in the high-dimensional settings.

Table 3: Performance of three methods for Example 4.

| $n$  | $f(x)$ | $||x-x^*||$ | Time (in seconds) |
|------|--------|-------------|-------------------|
|      | HTP    | LEMKE       | NL0R             |
|      | HTP    | LEMKE       | NL0R             |
|      | HTP    | LEMKE       | NL0R             |
| 5000 | 2.52e-06 | 1.15e-28   | 1.94e-27         | 5.55e-02 | 2.05e-14 | 6.46e-14 | 11.83 | 7.911 | 0.581 |
| 7500 | 4.20e-06 | 3.22e-28   | 3.84e-28         | 7.04e-02 | 3.15e-14 | 4.01e-14 | 27.69 | 27.14 | 1.240 |
| 10000| 5.38e-06| 7.21e-28   | 3.33e-27         | 8.36e-02 | 4.58e-14 | 9.88e-14 | 50.71 | 64.64 | 2.216 |
| 12500| 6.76e-06| 9.06e-28   | 4.00e-28         | 8.87e-02 | 5.10e-14 | 3.18e-14 | 79.96 | 127.7 | 3.434 |
| 15000| 7.99e-06| 1.18e-27   | 8.83e-28         | 9.86e-02 | 6.80e-14 | 6.07e-14 | 114.7 | 221.1 | 4.994 |
| 17500| 9.30e-06| 2.19e-27   | 8.37e-28         | 1.08e-01 | 7.69e-14 | 4.23e-14 | 158.4 | 354.0 | 6.862 |
| 20000| 1.12e-05| 3.22e-27   | 9.71e-27         | 1.18e-01 | 1.10e-13 | 2.72e-13 | 207.9 | 531.1 | 8.826 |

5 Conclusion

A vast body of work has developed numerical methods that only make use of the first-order information of the involved functions. Therefore, these methods can be executed rapidly in each step but may suffer from slow convergence. When Newton steps are integrated into some of these methods, the convergence can be accelerated significantly. However, there are few theoretical guarantees to thoroughly reveal the underlying reasons for this effect. In this paper, we developed a Newton-type method for the $\ell_0$-regularized optimization and proved that the generated sequence converged to a local minimizer globally and quadratically. Moreover, the numerical experiments displayed its strong ability to solve some high dimensional problems. Overall, the proposed method displays excellent performance theoretically and numerically.

Acknowledgements The authors sincerely thank the editor and an anonymous referee for their constructive comments, which have significantly improved the quality of the paper.
References

1. Attouch, H., Bolte, J., Svaiter, B.F.: Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized gauss-seidel methods. Mathematical Programming 137(1-2), 91–129 (2013)
2. Bahmani, S., Raj, B., Boufounos, P.T.: Greedy sparsity constrained optimization. Journal of Machine Learning Research 14(Mar), 807–841 (2013)
3. Bao, C., Dong, B., Hou, L., Shen, Z., Zhang, X., Zhang, X.: Image restoration by minimizing zero norm of wavelet frame coefficients. Inverse Problems 32(11), 115004 (2016)
4. Beck, A., Eldar, Y.C.: Sparsity constrained nonlinear optimization: Optimality conditions and algorithms. SIAM Journal on Optimization 23(3), 1480–1509 (2013)
5. Beck, A., Hallak, N.: Proximal mapping for symmetric penalty and sparsity. SIAM Journal on Optimization 28(1), 496–527 (2018)
6. Bertsimas, D., King, A., Mazumder, R.: Best subset selection via a modern optimization lens. The Annals of Statistics pp. 813–852 (2016)
7. Bian, W., Chen, X.: Smoothing neural network for constrained non-Lipschitz optimization with applications. IEEE Transactions on Neural Networks and Learning Systems 23(3), 399–411 (2012)
8. Bian, W., Chen, X.: Linearly constrained non-Lipschitz optimization for image restoration. SIAM Journal on Imaging Sciences 8(4), 2294–2322 (2015)
9. Bian, W., Chen, X.: A smoothing proximal gradient algorithm for nonsmooth convex regression with cardinality penalty. SIAM Journal on Numerical Analysis 58(1), 858–883 (2020)
10. Blanchard, J.D., Tanner, J., Wei, K.: CGIHT: conjugate gradient iterative hard thresholding for compressed sensing and matrix completion. Information and Inference: A Journal of the IMA 4(4), 289–327 (2015)
11. Blumensath, T.: Accelerated iterative hard thresholding. Signal Processing 92(3), 752–756 (2012)
12. Blumensath, T., Davies, M.E.: Gradient pursuits. IEEE Transactions on Signal Processing 56(6), 2370–2382 (2008)
13. Blumensath, T., Davies, M.E.: Iterative thresholding for sparse approximations. Journal of Fourier Analysis and Applications 14(5-6), 629–654 (2008)
14. Blumensath, T., Davies, M.E.: Normalized iterative hard thresholding: Guaranteed stability and performance. IEEE Journal of Selected Topics in Signal Processing 4(2), 298–309 (2010)
15. Candès, E.J., Romberg, J., Tao, T.: Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. IEEE Transactions on Information Theory 52(2), 489–509 (2006)
16. Candès, E.J., Tao, T.: Decoding by linear programming. IEEE Transactions on Information Theory 51(12), 4203–4215 (2005)
17. Chen, X., Ng, M.K., Zhang, C.: Non-Lipschitz \( \ell_p \)-regularization and box constrained model for image restoration. IEEE Transactions on Image Processing 21(12), 4709–4721 (2012)
18. Cheng, W., Chen, Z., Hu, Q.: An active set Barzilai-Borwein algorithm for \( \ell_0 \) regularized optimization. Journal of Global Optimization (2019)
19. Cottle, R.W.: Linear complementarity problem. Springer (2009)
20. Dai, W., Milenkovic, O.: Subspace pursuit for compressive sensing signal reconstruction. IEEE transactions on Information Theory 55(5), 2330–2349 (2009)
21. De Luca, T., Facchinei, F., Kanzow, C.: A semismooth equation approach to the solution of nonlinear complementarity problems. Mathematical Programming 75(3), 407–439 (1996)
22. Dinh, T., Wang, B., Bertozzi, A.L., Osher, S.J., Xin, J.: Sparsity meets robustness: channel pruning for the feynman-kac formalism principled robust deep neural nets. arXiv preprint arXiv:2003.08631 (2020)
23. Donoho, D.L.: Compressed sensing. IEEE Transactions on Information Theory 52(4), 1289–1306 (2006)
24. Elad, M.: Sparse and redundant representations: from theory to applications in signal and image processing. Springer Science & Business Media (2010)
25. Elad, M., Figueiredo, M.A., Ma, Y.: On the role of sparse and redundant representations in image processing. Proceedings of the IEEE 98(6), 972–982 (2010)
26. Facchinei, F.: Minimization of \( sl \) functions and the maratos effect. Operations Research Letters 17(3), 131–138 (1995)
27. Facchinei, F., Kanzow, C.: A nonsmooth inexact newton method for the solution of large-scale nonlinear complementarity problems. Mathematical Programming 76(3), 493–512 (1997)
28. Foucart, S.: Hard thresholding pursuit: an algorithm for compressive sensing. SIAM Journal on Numerical Analysis 48(3), 2543–2563 (2011)
29. Huang, J., Jiao, Y., Liu, Y., Lu, X.: A constructive approach to \( \ell_0 \) penalized regression. The Journal of Machine Learning Research 19(1), 403–439 (2018)
30. Ito, K., Kunisch, K.: A variational approach to sparsity optimization based on lagrange multiplier theory. Inverse Problems 30(1), 015001 (2013)
31. Jiao, Y., Jin, B., Lu, X.: A primal dual active set with continuation algorithm for the \( \ell_0 \)-regularized optimization problem. Applied and Computational Harmonic Analysis 39(3), 400–426 (2015)
32. Lai, M.J., Xu, Y., Yin, W.: Improved iteratively reweighted least squares for unconstrained smoothed $\ell_q$ minimization. SIAM Journal on Numerical Analysis 51(2), 927–957 (2013)
33. Lin, S., Ji, R., Li, Y., Deng, C., Li, X.: Toward compact convnets via structure-sparsity regularized filter pruning. IEEE Transactions on Neural Networks and Learning Systems (2019)
34. Lou, Y., Yan, M.: Fast $l_1 - l_2$ minimization via a proximal operator. Journal of Scientific Computing 74(2), 767–785 (2018)
35. Lu, Z.: Iterative hard thresholding methods for $\ell_0$ regularized convex cone programming. Mathematical Programming 147(1-2), 125–154 (2014)
36. Lu, Z., Zhang, Y.: Sparse approximation via penalty decomposition methods. SIAM Journal on Optimization 23(4), 2448–2478 (2013)
37. Moré, J.J., Sorensen, D.C.: Computing a trust region step. SIAM Journal on Scientific and Statistical Computing 4(3), 553–572 (1983)
38. Needell, D., Tropp, J.A.: CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. Applied and Computational Harmonic Analysis 26(3), 301–321 (2009)
39. Pan, L., Zhou, S., Xiu, N., Qi, H.D.: A convergent iterative hard thresholding for nonnegative sparsity optimization. Journal of Optimization Theory and Applications 163(3), 795–814 (2014)
40. Pati, Y.C., Rezaiifar, R., Krishnaprasad, P.S.: Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In: Signals, Systems and Computers, 1993. 1993 Conference Record of The Twenty-Seventh Asilomar Conference on, pp. 40–44. IEEE (1993)
41. Patrascu, A., Necoara, I., Patrinos, P.: A proximal alternating minimization method for $\ell_0$-regularized nonlinear optimization problems: Application to state estimation. Proceedings of the IEEE Conference on Decision and Control 2015, 4254–4259 (2015)
42. Shang, M., Zhang, C., Xiu, N.: Minimal zero norm solutions of linear complementarity problems. Journal of Optimization Theory and Applications 163(3), 795–814 (2014)
43. Shang, M., Zhou, S., Xiu, N.: Extragradient thresholding methods for sparse solutions of co-coercive ncps. Journal of Inequalities and Applications 2015(1), 34 (2015)
44. Soubies, E., Blanc-Féraud, L., Aubert, G.: A continuous exact $\ell_0$ penalty ($\text{cel}_0$) for least squares regularized problem. SIAM Journal on Imaging Sciences 8(3), 1607–1639 (2015)
45. Soussen, C., Idier, J., Brie, D.: Homotopy based algorithms for $\ell_0$-regularized least squares. IEEE Transactions on Signal Processing 63(13), 3301–3316 (2015)
46. Tropp, J.A.: Just relax: Convex programming methods for identifying sparse signals in noise. IEEE Transactions on Information Theory 52(3), 1030–1051 (2006)
47. Tropp, J.A., Gilbert, A.C.: Signal recovery from random measurements via orthogonal matching pursuit. IEEE Transactions on Information Theory 53(12), 4655–4666 (2007)
48. Wright, J., Ma, Y., Mairal, J., Sapiro, G., Huang, T.S., Yan, S.: Sparse representation for computer vision and pattern recognition. Proceedings of the IEEE 98(6), 1031–1044 (2010)
49. Xie, J., He, S., Zhang, S.: Randomized portfolio selection with constraints. Pacific Journal of Optimization 4(1), 89–112 (2008)
50. Yin, P., Lou, Y., He, Q., Xin, J.: Minimization of 1-2 for compressed sensing. SIAM Journal on Scientific Computing 37(1), A536–A563 (2015)
51. Yuan, X.T., Li, P., Zhang, T.: Gradient hard thresholding pursuit. The Journal of Machine Learning Research 18(1), 6027–6069 (2017)
52. Yuan, X.T., Liu, Q.: Newton greedy pursuit: A quadratic approximation method for sparsity-constrained optimization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4122–4129 (2014)
53. Yuan, X.T., Liu, X., Yan, S.: Visual classification with multitask joint sparse representation. IEEE Transactions on Image Processing 21(10), 4349–4360 (2012)
54. Zhou, S., Shang, M., Pan, L., Li, M.: Newton hard thresholding pursuit for sparse lcp via a new merit function. arXiv preprint arXiv:2004.02244 (2020)
55. Zhou, S., Xiu, N., Qi, H.D.: Global and quadratic convergence of Newton hard-thresholding pursuit. arXiv preprint arXiv:1901.02763 (2019)
56. Zhou, S., Xiu, N., Wang, Y., Kong, L., Qi, H.D.: A null-space-based weighted $\ell_1$ minimization approach to compressed sensing. Information and Inference: A Journal of the IMA 5(1), 76–102 (2016)