SUBSPACE NEWTON METHOD FOR ℓ_0-REGULARIZED OPTIMIZATION *

SHENGLONG ZHOU†, LILI PAN‡, AND NAIHUA XIU§

Abstract. Sparse optimization has seen an evolutionary advance in the past decade with extensive applications ranging from image and signal processing, statistics to machine learning. As a tractable approach, regularization is frequently used, leading to a regularized optimization where ℓ_0 norm or its continuous approximations that characterize the sparsity are punished in its objective. From the continuity of approximations to the discreteness of ℓ_0 norm, the most challenging model is the ℓ_0-regularized optimization. To conquer its hardness, numerous numerically effective methods have been proposed. However, most of them only enjoy that the (sub)sequence converges to a stationary point from the deterministic optimization perspective or the distance between each iterate and any given sparse reference point is bounded by an error bound in the sense of probability. We design a method SNL0: subspace Newton method for the ℓ_0-regularized optimization, and prove that its generated sequence converges to a stationary point globally under the strong smoothness condition. In addition, it is also quadratic convergent with the help of locally strong convexity, well explaining that our method, as a second order method, is able to converge very fast. Moreover, a novel mechanism to effectively update the penalty parameter is created, which allows us to get rid of the tedious parameter tuning task that is suffered by most regularized optimization methods.

Key words. Sparse optimization, τ-stationary point, Subspace Newton method, Global and quadratic convergence

AMS subject classifications. 65K05, 90C46, 90C06, 90C27

1. Introduction.

Over the last decade, sparsity has been thoroughly investigated due to its extensive applications ranging from compressed sensing [24, 15, 16], signal and image processing [26, 25, 17, 8], machine learning [51, 56] to neural networks [7, 36, 22] lately. Sparsity is frequently characterized by ℓ_0 norm and its penalized problem is commonly phrased as ℓ_0-regularized optimization, taking the form of

$$\min_{\mathrm{x}\in\mathbb{R}^n} \quad f(\mathrm{x}) + \lambda\|\mathrm{x}\|_0, \tag{1.1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable and bounded from below, $\lambda > 0$ is the penalty parameter and $\|\mathrm{x}\|_0$ is ℓ_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 is the so-called sparsity constrained optimization:

$$\min_{\mathrm{x}\in\mathbb{R}^n} \ f(\mathrm{x}), \quad \text{s.t. } \|\mathrm{x}\|_0 \le s, \tag{1.2}$$

where $s \le n$ is a given positive integer. Based on the two optimizations, large numbers of state-of-the-art methods have been proposed in the last decade. In particular, many of them are designed for a special application, compressed sensing (CS), where the

*Submitted to the editors DATE.

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

†School of Mathematics, University of Southampton, Southampton SO17 1BJ, United Kingdom. (shenglong.zhou@soton.ac.uk)

‡Department of Mathematics, Shandong University of Technology, Zibo 255049, P.R. China. (panlili1979@126.com).

§Department of Applied Mathematics, Beijing Jiaotong University, Beijing 10044, P.R. China. (nhxiu@bjtu.edu.cn).

least squares are taken into account, namely

$$f(\mathrm{x}) := f_{cs}(\mathrm{x}) \equiv \|A\mathrm{x} - \mathrm{y}\|^2. \tag{1.3}$$

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

1.1. Selective Literature Review. Since the varieties of kinds of methods are beyond our scope of reviewing, we only cite a small portion which should be enough to clarify our motivations of this paper.

(a) Methods for (1.2) are known as greedy ones. For the case of CS, one can refer to orthogonal matching [44, 50, OMP], gradient pursuit [12, GP], compressive sample matching pursuit [41, CoSaMP], subspace pursuit [20, SP], normalized iterative hard-thresholding [14, NIHT], hard-thresholding pursuit [31, HTP] and accelerated iterative hard-thresholding [11, AIHT]. Methods for the general model (1.2) include the gradient support pursuit [2, GraSP], iterative hard-thresholding [4, IHT], Newton gradient pursuit [55, NTGP], conjugate gradient iterative hard-thresholding [10, CGIHT], gradient hard-thresholding pursuit [54, GraHTP], improved iterate hard-thresholding [43, IIHT] and Newton hard-thresholding pursuit [59, NHTP].

To derive the convergence results, most methods enjoy the theory that the distance between each iterate to any given reference (sparse) point is bounded by an error through statistic analysis. By contrast, methods like IHT, IIHT and NHTP have been proved to converge to a stationary point globally in the sense of the deterministic way. 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 super fast computational speed and high order of accuracy, but without deterministic theoretical guarantees for a long time. Until recently, authors in [59] first proved that their proposed NHTP has global and quadratic convergence properties, which unravel the reason why these methods behave exceptionally well.

(b) Methods for (1.1) aiming at addressing CS problem via the model (1.1) include iterative hard-thresholding algorithm [13, IHT], continuous exact ℓ_0 penalty [47, CEL0], two methods: continuation single best replacement and ℓ_0-regularization path descent in [48, CSBR, L0BD], forward-backward splitting [1, FBS], extrapolated proximal iterative hard-thresholding algorithm [3, EPIHT] and mixed integer optimization method [6, MIO], to name just a few. While for the general problem (1.1), one can see penalty decomposition [39, PD] where equality and inequality constraints are also considered, iterative hard-thresholding [38, see] where the box and convex cone are taken into account, proximal gradient method and coordinate-wise support optimality method [5, PG, CowS] where sparse solutions are sought from a symmetric set, random proximal alternating minimization method [45, RPA], active set Barzilar-Borwein [18, ABB] and a very recently smoothing proximal gradient method [9, SPG]. Note that these methods can be regarded as the first order methods since they only benefit from the first order information such as gradients or function values. Then second order methods have attracted much attention lately, including primal dual active set [33, PDAS], primal dual active set with continuation [34, PDASC] and support detection and root finding [32, SDAR].

As 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 local convergence property) for methods like PD, PG and ABB. It is worth mentioning that 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). Apart from that, no better deterministic theoretical guarantees (like quadratic convergence) have been established on algorithms for solving (1.1). Therefore, a natural question is: can we develop an algorithm based on ℓ_0-regularized optimization that enjoys the global and quadratic convergence?

1.2. Contributions. In order to answer the above question, we begin with introducing a τ-stationary point, an optimality condition of (1.1), and then reveal its relationship with local/global minimizers by Theorem 2.3. It is known that a τ-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.

Following the idea of the τ-stationary point, we perform Newton step only on chosen subspaces decided by the support sets. The proposed method is dubbed as SNL0, an abbreviation for subspace Newton method for ℓ_0-regularized optimization (1.1). Differing from methods PDAS or SDAR, where similar algorithmic schemes are used, the Armijo line search is adopted to ensure sufficient descent for each step. This successfully extends the classical convergence result of Newton's method to ℓ_0-regularization. Namely, the proposed method enjoys the global and quadratic convergence properties under standard assumptions (see Theorem 3.9). As far as we know, it is the first paper that establishes both for an algorithm aiming at solving (1.1).

Parameter tuning is always a tedious but crucial task for penalty problems with (1.1) as a special case. We design a novel mechanism (see `PSS` in Algorithm 3.1) interpolated into SNL0 to update the parameter λ adaptively, which enables the method to identify the support set eventually. Numerical experiments demonstrate that this mechanism is very effective.

Finally, extensive numerical experiments illustrate that SNL0 is very competitive when benchmarked against other methods for solving problems such as compressed sensing and sparse complementarity problems. In a nutshell, it is capable of generating desirably sparse solutions by consuming significantly short time.

It is worth mentioning that to find the subspace, PDASC and SDAR also make use of the idea of the τ-stationary point but always with $\tau = 1$, while SNL0 benefits from more choices of $\tau > 0$. In addition, gradient direction is exploited as an alternative of Newton direction if the latter does not guarantee sufficient decline of f in some steps. However, PDASC and SDAR always take advantage of full Newton directions. Comparing with the method NHTP in [59] for the model (1.2), where the sparsity level s is required and decides the quality of its final solutions, SNL0 is able to find a sparse solution only with a rough input instead of a rigorous sparsity level s.

1.3. Organization and Notation. The rest of the paper is organized as follows. Next section establishes the optimality conditions of (1.1) with the help of the τ-stationary point whose relationship with the local/global minimizers of (1.1) by Theorem 2.3 is also given. In Section 3, we design the subspace Newton method for the ℓ_0-regularized optimization (SNL0). Particularly, a novel strategy (`PSS` in Algorithm 3.1) is proposed to update the penalty parameter λ and the support set in each step of the algorithm, followed by the main convergence results including the support set identification, global and quadratic convergence properties under some standard assumptions. Extensive numerical experiments are presented in Section 4, where the implementation of SNL0 as well as its comparisons with some other excellent solvers for solving problems, such as compressed sensing and sparse complementarity problems, are provided. Concluding remarks are made in the last section.

We end this section with some notation to be employed throughout the paper. Let $\mathbb{N}_n := \{1, 2, \cdots, n\}$. Given a vector x, let $|\mathrm{x}| := (|x_1, |x_2|, \cdots, |x_n|)^\top$,

$\|\mathrm{x}\| := \sqrt{\sum_i |x_i|^2}$ and $\|\mathrm{x}\|_1 := \sum_i |x_i|$ be its ℓ_2 and ℓ_1 norm, respectively. Moreover, denote $\|\mathrm{x}\|_{[i]}$ as the ith largest element $|\mathrm{x}|$. For example, $\|\mathrm{x}\|_{[2]} = \|\mathrm{x}\|_{[3]} = 2$ if $\mathrm{x} = (3, 2, 1, -2)^\top$. The support set of x is supp(x) consisting of indices of its non-zero elements. Given a set $T \subseteq \mathbb{N}_n$, $|T|$ and $\overline{T}$ are the cardinality and the complementary set. The sub vector of x containing elements indexed on T is denoted by $\mathrm{x}_T \in \mathbb{R}^{|T|}$. Now, for a matrix $A \in \mathbb{R}^{m\times n}$, let $\|A\|_2$ represent its spectral norm, i.e., its maximum singular value. Write $A_{T,J}$ is the submatrix containing rows indexed on T and columns indexed on J. In particular, denote $A_{:J} = A_{\mathbb{N}_m,J}, A_{T:} = A_{T,\mathbb{N}_n}$, the sub-gradient and the sub-Hessians by

$$\nabla_T f(\mathrm{x}) := (\nabla f(\mathrm{x}))_T, \quad \nabla^2_{T,J} f(\mathrm{x}) := (\nabla^2 f(\mathrm{x}))_{T,J}, \quad \nabla^2_T f(\mathrm{x}) := (\nabla^2 f(\mathrm{x}))_{T,T}.$$

Finally, $\lceil a \rceil$ stands for the smallest integer that is no less than a.

2. Optimality.

Some necessary optimality conditions of (1.1) have been studied. These include ones in [39, Theorem 2.1] and [5, Theorem 4.10]. Here, inspired by the latter, we introduce a τ-stationary point (this is the same as the L-stationarity in [5]).

2.1. τ-stationary point. A vector $\mathrm{x} \in \mathbb{R}^n$ is called a τ-stationary point of (1.1) if there is a $\tau > 0$ such that

$$\mathrm{x} \in \mathrm{Prox}_{\tau\lambda\|\cdot\|_0}(\mathrm{x} - \tau\nabla f(\mathrm{x})) := \underset{\mathrm{z}\in\mathbb{R}^n}{\mathrm{argmin}}\ \frac{1}{2}\|\mathrm{z} - (\mathrm{x} - \tau\nabla f(\mathrm{x}))\|^2 + \tau\lambda\|\mathrm{z}\|_0. \tag{2.1}$$

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

$$\Big[\mathrm{Prox}_{\tau\lambda\|\cdot\|_0}(\mathrm{z})\Big]_i = \begin{cases} z_i, & |z_i| > \sqrt{2\tau\lambda}, \\ \{z_i, 0\}, & |z_i| = \sqrt{2\tau\lambda}, \\ 0, & |z_i| < \sqrt{2\tau\lambda}. \end{cases} \tag{2.2}$$

This allows us to characterize a τ-stationary point by conditions below equivalently, see [49, Theorem 24] and [13, Lemma 2].

LEMMA 2.1. *A point* x *is a τ-stationary point with $\tau > 0$ of* (1.1) *if and only if*

$$\begin{cases} \nabla_i f(\mathrm{x}) = 0 \text{ and } |x_i| \geq \sqrt{2\tau\lambda}, & i \in \mathrm{supp}(\mathrm{x}), \\ |\nabla_i f(\mathrm{x})| \leq \sqrt{2\lambda/\tau}, & i \notin \mathrm{supp}(\mathrm{x}). \end{cases} \tag{2.3}$$

From Lemma 2.1, for any $0 < \tau_1 \leq \tau$, a τ-stationary point x is also a τ_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 2.2. *A function f is strongly smooth with a constant $L > 0$ if*

$$f(\mathrm{z}) \leq f(\mathrm{x}) + \langle \nabla f(\mathrm{x}), \mathrm{z} - \mathrm{x}\rangle + \frac{L}{2}\|\mathrm{z} - \mathrm{x}\|^2,\ \forall\ \mathrm{x}, \mathrm{z} \in \mathbb{R}^n. \tag{2.4}$$

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

$$f(\mathrm{z}) \geq f(\mathrm{x}) + \langle \nabla f(\mathrm{x}), \mathrm{z} - \mathrm{x}\rangle + \frac{\ell}{2}\|\mathrm{z} - \mathrm{x}\|^2,\ \forall\ \mathrm{x}, \mathrm{z} \in \mathbb{R}^n. \tag{2.5}$$

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

Something needs emphasize here is that when the function is locally strongly convex, the constant ℓ depends on the point x. We drop the dependence for simplicity since it would not cause confusion in the context. The strong convexity and smoothness respectively indicate that, for any $\mathrm{x}, \mathrm{z} \in \mathbb{R}^n$

$$\ell\|\mathrm{z} - \mathrm{x}\| \leq \|\nabla f(\mathrm{z}) - \nabla f(\mathrm{x})\| \leq L\|\mathrm{z} - \mathrm{x}\|. \tag{2.6}$$

2.2. First order optimality conditions. Our next major result is to establish the relationships between a τ-stationary point and a local/global minimizer of (1.1).

THEOREM 2.3. *For problem* (1.1)*, the following results hold.*

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

$$\mathrm{x}^* = \mathrm{Prox}_{\tau\lambda\|\cdot\|_0}\left(\mathrm{x}^* - \tau\nabla f(\mathrm{x}^*)\right). \tag{2.7}$$

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

Proof. 1) Denote $\mathbb{P} := \mathrm{Prox}_{\tau\lambda\|\cdot\|_0}\left(\mathrm{x}^* - \tau\nabla f(\mathrm{x}^*)\right)$ and $\mu := L - 1/\tau < 0$ due to $0 < \tau < 1/L$. Let x^* be a global minimizer and consider any point $\mathrm{z} \in \mathbb{P}$. Then we have the following chain of inequality

$$\begin{aligned}
& f(\mathrm{z}) + \lambda\|\mathrm{z}\|_0 \\
\overset{(a)}{\leq}\ & f(\mathrm{x}^*) + \langle\nabla f(\mathrm{x}^*), \mathrm{z}-\mathrm{x}^*\rangle + \frac{L}{2}\|\mathrm{z}-\mathrm{x}^*\|^2 + \lambda\|\mathrm{z}\|_0 \\
=\ & f(\mathrm{x}^*) + \langle\nabla f(\mathrm{x}^*), \mathrm{z}-\mathrm{x}^*\rangle + \frac{1}{2\tau}\|\mathrm{z}-\mathrm{x}^*\|^2 + \frac{\mu}{2}\|\mathrm{z}-\mathrm{x}^*\|^2 + \lambda\|\mathrm{z}\|_0 \\
=\ & f(\mathrm{x}^*) + \frac{1}{2\tau}\|\mathrm{z}-(\mathrm{x}^*-\tau\nabla f(\mathrm{x}^*))\|^2 - \frac{\tau}{2}\|\nabla f(\mathrm{x}^*)\|^2 + \lambda\|\mathrm{z}\|_0 + \frac{\mu}{2}\|\mathrm{z}-\mathrm{x}^*\|^2 \\
\overset{(b)}{\leq}\ & f(\mathrm{x}^*) + \frac{1}{2\tau}\|\mathrm{x}^*-(\mathrm{x}^*-\tau\nabla f(\mathrm{x}^*))\|^2 + \lambda\|\mathrm{x}^*\|_0 - \frac{\tau}{2}\|\nabla f(\mathrm{x}^*)\|^2 + \frac{\mu}{2}\|\mathrm{z}-\mathrm{x}^*\|^2 \\
=\ & f(\mathrm{x}^*) + \lambda\|\mathrm{x}^*\|_0 + \frac{\mu}{2}\|\mathrm{z}-\mathrm{x}^*\|^2 \overset{(c)}{\leq} f(\mathrm{z}) + \lambda\|\mathrm{z}\|_0 + \frac{\mu}{2}\|\mathrm{z}-\mathrm{x}^*\|^2,
\end{aligned}$$

where (a), (b) and (c) hold respectively from facts that f being strongly smooth, $\mathrm{z} \in \mathbb{P}$ and x^* being the global minimizer of (1.1). This together with $\mu < 0$ leads to $0 \leq (\mu/2)\|\mathrm{z}-\mathrm{x}^*\|^2 < 0$, which yields $\mathrm{z} = \mathrm{x}^*$. Therefore, x^* is a τ-stationary point of (1.1). Since z is arbitrary in $\mathbb{P}$ and $\mathrm{z} = \mathrm{x}^*$, $\mathbb{P}$ is a singleton only containing x^*.

2) Let x^* be a τ-stationary point with $\tau > 0$ with $T_* := \mathrm{supp}(\mathrm{x}^*)$ and $\epsilon := \min_{i\in T_*}|\mathrm{x}^*|$. Consider a neighbour region of x^* as

$$N(\mathrm{x}^*) = \left\{\mathrm{x} \in \mathbb{R}^n : \|\mathrm{x}-\mathrm{x}^*\|_1 < \left\{\begin{array}{ll} \min\{\epsilon, \sqrt{0.5\tau\lambda}\}, & \mathrm{x}^* \neq 0 \\ \sqrt{0.5\tau\lambda}, & \mathrm{x}^* = 0 \end{array}\right\}\right\}.$$

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

$$\epsilon \leq |x_j^*| = |x_j^* - x_j| \leq \|\mathrm{x}-\mathrm{x}^*\|_1 < \min\{\epsilon, \sqrt{0.5\tau\lambda}\} \leq \epsilon.$$

Therefore, we have $T_* \subseteq \mathrm{supp}(\mathrm{x})$. The convexity of f suffices to

$$\begin{aligned}
f(\mathrm{x}) &\geq f(\mathrm{x}^*) + \langle\nabla f(\mathrm{x}^*), \mathrm{x}-\mathrm{x}^*\rangle \\
&= f(\mathrm{x}^*) + \langle\nabla_{T_*} f(\mathrm{x}^*), (\mathrm{x}-\mathrm{x}^*)_{T_*}\rangle + \langle\nabla_{\overline{T}_*} f(\mathrm{x}^*), (\mathrm{x}-\mathrm{x}^*)_{\overline{T}_*}\rangle \\
&\overset{(2.3)}{=} f(\mathrm{x}^*) + \langle\nabla_{\overline{T}_*} f(\mathrm{x}^*), \mathrm{x}_{\overline{T}_*}\rangle =: f(\mathrm{x}^*) + \phi.
\end{aligned} \tag{2.8}$$

If $T_* = \mathrm{supp}(\mathrm{x})$, then we have $\phi = 0$ due to $\mathrm{x}_{\overline{T}_*} = 0$ and $\|\mathrm{x}^*\|_0 = \|\mathrm{x}\|_0$. These allow us to derive that

$$f(\mathrm{x}) + \lambda\|\mathrm{x}\|_0 \overset{(2.8)}{\geq} f(\mathrm{x}^*) + \phi + \lambda\|\mathrm{x}\|_0 = f(\mathrm{x}^*) + \lambda\|\mathrm{x}^*\|_0.$$

If $T_* \subseteq (\neq)\mathrm{supp}(\mathrm{x})$, then $\|\mathrm{x}\|_0 - 1 \geq \|\mathrm{x}^*\|_0$. In addition,

$$\begin{aligned} \phi = \langle \nabla_{\overline{T}_*} f(\mathrm{x}^*), \mathrm{x}_{\overline{T}_*}\rangle &\overset{(2.3)}{\geq} -\sqrt{2\lambda/\tau}\sum_{i\in\overline{T}_*}|x_i| = -\sqrt{2\lambda/\tau}\sum_{i\in\overline{T}_*}|x_i - x_i^*| \\ &\geq\ -\sqrt{2\lambda/\tau}\|\mathrm{x}-\mathrm{x}^*\|_1 > -\sqrt{2\lambda/\tau}\sqrt{0.5\tau\lambda} = -\lambda. \end{aligned}$$

These facts enable us to derive that

$$f(\mathrm{x}) + \lambda\|\mathrm{x}\|_0 \overset{(2.8)}{\geq} f(\mathrm{x}^*) + \phi + \lambda\|\mathrm{x}\|_0 > f(\mathrm{x}^*) + \lambda(\|\mathrm{x}\|_0 - 1) \geq f(\mathrm{x}^*) + \lambda\|\mathrm{x}^*\|_0.$$

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

$$\frac{1}{2}\|\mathrm{x} - (\mathrm{x}^* - \tau\nabla f(\mathrm{x}^*))\|^2 + \tau\lambda\|\mathrm{x}\|_0 \geq \frac{1}{2}\|\mathrm{x}^* - (\mathrm{x}^* - \tau\nabla f(\mathrm{x}^*))\|^2 + \tau\lambda\|\mathrm{x}^*\|_0,$$

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

$$\langle \nabla f(\mathrm{x}^*),\ \mathrm{x} - \mathrm{x}^*\rangle + \lambda\|\mathrm{x}\|_0 \geq -\frac{1}{2\tau}\|\mathrm{x}-\mathrm{x}^*\|^2 + \lambda\|\mathrm{x}^*\|_0. \tag{2.9}$$

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

$$\begin{aligned} f(\mathrm{x}) + \lambda\|\mathrm{x}\|_0 &\overset{(2.5)}{\geq} f(\mathrm{x}^*) + \langle \nabla f(\mathrm{x}^*), \mathrm{x} - \mathrm{x}^*\rangle + \frac{\ell}{2}\|\mathrm{x}-\mathrm{x}^*\|^2 + \lambda\|\mathrm{x}\|_0 \\ &\overset{(2.9)}{\geq} f(\mathrm{x}^*) + \frac{\ell - 1/\tau}{2}\|\mathrm{x}-\mathrm{x}^*\|^2 + \lambda\|\mathrm{x}^*\|_0 \geq f(\mathrm{x}^*) + \lambda\|\mathrm{x}^*\|_0, \end{aligned}$$

where the last inequality is from $\tau \geq 1/\ell$. Clearly, if $\tau > 1/\ell$, then the last inequality holds strictly, which means x^* is a unique global minimizer. □

Let us consider an example to illustrate the above theorem.

EXAMPLE 2.1. *Let* $\mathrm{a} = (t\ 1\ 1)^\top$, $\lambda > 8$ *and* f *be given by*

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

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

$$\begin{aligned} \nabla_1 f(\mathrm{x}^*) = 0 \text{ and } |x_1| = t \geq \lambda/2 = \sqrt{2\lambda\lambda/8} \geq \sqrt{2\lambda\tau}, \\ |\nabla_2 f(\mathrm{x}^*)| = |\nabla_3 f(\mathrm{x}^*)| = 4 = \sqrt{2\times 8} \leq \sqrt{2\lambda/\tau}. \end{aligned}$$

Then it follows from Theorem 2.3 1) and $\tau > 1 = 1/\ell$ that x^* is a unique global minimizer of the problem (1.1). Moreover, Theorem 2.3 2) concludes that a global minimizer (which is x^*) is also a τ_1-stationary point with $\tau_1 \in (0, 1/L) = (0, 1/2)$. This is not conflicted with x^* being a τ-stationary point with some $\tau \in (1, \lambda/8]$.

2.3. Stationary Equation. To well express the solution of (2.1), define

$$T := T_\tau(\mathrm{x}, \lambda) := \left\{ i \in \mathbb{N}_n : |x_i - \tau \nabla_i f(\mathrm{x})| \geq \sqrt{2\tau\lambda} \right\}. \tag{2.11}$$

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

$$F_\tau(\mathrm{x}; T) := \left[\begin{array}{c} \nabla_T f(\mathrm{x}) \\ \mathrm{x}_{\overline{T}} \end{array} \right] = 0. \tag{2.12}$$

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

THEOREM 2.4. *For any* $\mathrm{x} \in \mathbb{R}^n$*, by letting* $\mathrm{z} := \mathrm{x} - \tau \nabla f(\mathrm{x})$*, we have*

$$\mathrm{x} = \mathrm{Prox}_{\tau\lambda\|\cdot\|_0}(\mathrm{z}) \quad \Longrightarrow \quad F_\tau(\mathrm{x}; T) = 0 \quad \Longrightarrow \quad \mathrm{x} \in \mathrm{Prox}_{\tau\lambda\|\cdot\|_0}(\mathrm{z}).$$

Proof. If we have $\mathrm{x} = \mathrm{Prox}_{\tau\lambda\|\cdot\|_0}(\mathrm{z})$, namely, $\mathrm{Prox}_{\tau\lambda\|\cdot\|_0}(\mathrm{z})$ 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 $(\mathrm{Prox}_{\tau\lambda\|\cdot\|_0}(\mathrm{z}))_T = \mathrm{z}_T$. As a consequence,

$$0 = \mathrm{x} - \mathrm{Prox}_{\tau\lambda\|\cdot\|_0}(\mathrm{z}) \overset{(2.2)}{=} \left[\begin{array}{c} \mathrm{x}_T \\ \mathrm{x}_{\overline{T}} \end{array} \right] - \left[\begin{array}{c} \mathrm{z}_T \\ 0 \end{array} \right] = \left[\begin{array}{c} \tau \nabla_T f(\mathrm{x}) \\ \mathrm{x}_{\overline{T}} \end{array} \right],$$

which suffices to $F_\tau(\mathrm{x}; T) = 0$. We now prove the second claim. For any $i \in T$, we have $\nabla_i f(\mathrm{x}) = 0$ from (2.12) and thus $|x_i| \geq \sqrt{2\tau\lambda}$ from (2.11). For any $i \in \overline{T}$, we have $x_i = 0$ from (2.12) and $|\tau \nabla_i f(\mathrm{x})| = |x_i - \tau \nabla_i f(\mathrm{x})| < \sqrt{2\tau\lambda}$ from (2.11). Those together with Lemma 2.1 claim the conclusion immediately. ∎

Above theorem states that a point satisfying the stationary equation is a stronger condition than being a τ-stationary point. The advantage of this equation is that it allows us to design an efficient Newton-type algorithm based on its simple form.

3. Subspace Newton Method. This section casts a Newton-type method to solve the stationary equation (2.12). Hereafter, for notational simplicity, we denote

$$g^k := \nabla f(\mathrm{x}^k).$$

3.1. Algorithmic Design. If the current x^k is computed, we need to find T_k by (2.11). The first issue confronted us is how to guarantee a non-empty T_k. To do that, we introduce the following scheme to update λ and set a proper T_k adaptively. Given parameters $\epsilon > 0, c \geq 1$ and an integer $K \geq 1$, let $\mathrm{x}^k \in \mathbb{R}^n$, $T_{k-1} \subseteq \mathbb{N}_n$ and $s_{k-1} \in \mathbb{N}_n$ be computed for the current step. Then T_k is updated by

$$T_k = \mathtt{PSS}(\mathrm{x}^k, T_{k-1}, s_{k-1}), \tag{3.1}$$

which is presented in Algorithm 3.1.

Note that the s_kth largest element of $|\mathrm{x}^k - \tau g^k|$ might be multiple, so the difference between T_k from (3.4) and $T_\tau(\mathrm{x}^k, \lambda_k)$ from (2.11) is that $|T_k| = s_k \leq |T_\tau(\mathrm{x}^k, \lambda_k)|$. We choose to update T_k by (3.4) rather than by (2.11) since it is easier to control the cardinality of T_k and consequently, easier to control the support set of the iterate in each step. Moreover, setting T_k by (3.4) still preserves the property of (2.11), which together with other several useful properties are given in the following lemma.

LEMMA 3.1. *If* $s_{-1} > 0$*, then for any* $k \geq 0$*,* T_k *is non-empty and* $|T_k| = s_k \leq s_{k+1} = |T_{k+1}|$*. Furthermore,*

$$|\mathrm{x}_i^k - \tau g_i^k| \geq \sqrt{2\tau\lambda_k} \geq |\mathrm{x}_j^k - \tau g_j^k|, \quad \forall\ i \in T_k,\ \forall\ j \in \overline{T}_k. \tag{3.5}$$

Algorithm 3.1 Penalty and subspace selection ($\mathtt{PSS}(\mathrm{x}^k, T_{k-1}, s_{k-1})$)

Tuning the sparsity level s_k by

$$s_k = \begin{cases} \max\{s_{k-1}, \lceil c|T_{k-1}|\rceil\}, & \text{if } k/K = \lceil k/K \rceil \text{ and } \|g^k_{T_{k-1}}\| \geq \epsilon, \\ s_{k-1}, & \text{otherwise.} \end{cases} \tag{3.2}$$

Setting the penalty λ_k by

$$\lambda_k = \frac{1}{2\tau}\|\mathrm{x}^k - \tau g^k\|^2_{[s_k]}. \tag{3.3}$$

Selecting the subspace T_k by

$$\begin{aligned} & T_k = \Gamma_k \cup \Upsilon_k, \quad \text{where} \\ & \Gamma_k = \{i \in \mathbb{N}_n : |\mathrm{x}^k_i - \tau g^k_i| > \sqrt{2\tau\lambda_k}\}, \\ & \Upsilon_k \subseteq \{i \in \mathbb{N}_n : |\mathrm{x}^k_i - \tau g^k_i| = \sqrt{2\tau\lambda_k}\}, \quad \text{with} \quad |\Upsilon_k| = s_k - |\Gamma_k|. \end{aligned} \tag{3.4}$$

return T

The proof is quite straightforward by the fact $0 < s_{-1} \leq s_1 \leq s_2 \leq \cdots$, so we omit it here. One can discern that T_k actually coincides with the indices of a point which is one of the solutions of the problem $\Pi_{s_k}(\mathrm{z}^k) \in \mathrm{argmin}_{\|\mathrm{z}\|_0 \leq s_k}\|\mathrm{z} - \mathrm{z}^k\|$, where $\mathrm{z}^k := \mathrm{x}^k - \tau g^k$. Here, Π_s is also known as the Hard-Thresholding operator that keeps s largest (in modulus) components of a vector and sets the other ones to zeros [31]. Now T_k is well defined. For the equation (2.12) with such fixed set T_k, the classical Newton direction d^k is a solution of the following equation:

$$\nabla F_\tau(\mathrm{x}^k; T_k)\mathrm{d} = -F_\tau(\mathrm{x}^k; T_k). \tag{3.6}$$

The explicit formula of $F_\tau(\mathrm{x}^k; T_k)$ from (2.12) implies that d^k satisfies

$$\begin{aligned} \nabla^2_{T_k} f(\mathrm{x}^k)\mathrm{d}^k_{T_k} &= \nabla^2_{T_k, \overline{T}_k} f(\mathrm{x}^k)\mathrm{x}^k_{\overline{T}_k} - g^k_{T_k}, \\ \mathrm{d}^k_{\overline{T}_k} &= -\mathrm{x}^k_{\overline{T}_k}. \end{aligned} \tag{3.7}$$

Now let us take a look at the above formulas. The second part of d^k can be derived directly without any difficult computations. To find d^k, one needs to solve a linear equation with s_k equations and s_k variables since $|T_k| = s_k$ by Lemma 3.1. If a full Newton direction is taken, then next iterate $\mathrm{x}^{k+1} = \mathrm{x}^k + \mathrm{d}^k = [(\mathrm{x}^k_{T_k} + \mathrm{d}^k_{T_k})^\top\ 0]^\top$. This means the support set of the next iterate will be located within T_k. Based on this idea, we modify the standard rule associated with Amijio line search $\mathrm{x}^{k+1} = \mathrm{x}^k + \alpha\mathrm{d}^k$ as $\mathrm{x}^{k+1} = \mathrm{x}^k(\alpha)$, where

$$\mathrm{x}^k(\alpha) := \begin{bmatrix} \mathrm{x}^k_{T_k} + \alpha\mathrm{d}^k_{T_k} \\ \mathrm{x}^k_{\overline{T}_k} + \mathrm{d}^k_{\overline{T}_k} \end{bmatrix} = \begin{bmatrix} \mathrm{x}^k_{T_k} + \alpha\mathrm{d}^k_{T_k} \\ 0 \end{bmatrix}. \tag{3.8}$$

In this way, the updated iterate is sparser than s_k. We summarize the framework of the algorithm in Algorithm 3.2 entitled with subspace Newton method for the ℓ_0-regularized optimization, where some notation are defined by

$$J_k := T_{k-1} \backslash T_k, \qquad H_k := \nabla^2_{T_k} f(\mathrm{x}^k), \qquad G_k := \nabla^2_{T_k, J_k} f(\mathrm{x}^k). \tag{3.9}$$

Algorithm 3.2 Subspace Newton method for the ℓ_0-regularized optimization (SNL0)

Initialize x^0 and choose $\tau > 0, \delta > 0, \sigma \in (0, 1/2), c \geq 1, K \geq 1, \epsilon > 0, \beta \in (0, 1)$
Set $s_{-1} > 0, T_{-1} = \emptyset$ and $k \Leftarrow 0$
while The halting conditions do not meet **do**

Step 1. Select $T_k = \mathtt{PSS}(\mathrm{x}^k, T_{k-1}, s_{k-1})$ in Algorithm 3.1

Step 2. If (3.7) is solvable and its solution d^k satisfies

$$\langle g^k_{T_k}, \mathrm{d}^k_{T_k}\rangle \leq -\delta\|\mathrm{d}^k\|^2 + \|\mathrm{x}^k_{\overline{T}_k}\|^2/(4\tau), \tag{3.10}$$

then update d^k by solving (3.7), namely by Newton direction,

$$H_k \mathrm{d}^k_{T_k} = G_k \mathrm{x}^k_{J_k} - g^k_{T_k}, \qquad \mathrm{d}^k_{\overline{T}_k} = -\mathrm{x}^k_{\overline{T}_k}. \tag{3.11}$$

Otherwise, update d^k by Gradient direction

$$\mathrm{d}^k_{T_k} = -g^k_{T_k}, \qquad \mathrm{d}^k_{\overline{T}_k} = -\mathrm{x}^k_{\overline{T}_k}. \tag{3.12}$$

Step 3. Find the smallest non-negative integer m_k such that

$$f(\mathrm{x}^k(\beta^{m_k})) \leq f(\mathrm{x}^k) + \sigma\beta^{m_k}\langle g^k, \mathrm{d}^k\rangle. \tag{3.13}$$

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

end while
return x^k

From our proposed algorithm in Algorithm 3.2, we have the following facts:

$$\begin{cases} \mathrm{supp}(\mathrm{x}^{k+1}) \subseteq T_k, \\ -\mathrm{d}^k_{\overline{T}_k} = \mathrm{x}^k_{\overline{T}_k} = \begin{bmatrix} \mathrm{x}^k_{T_{k-1}\cap\overline{T}_k} \\ 0 \end{bmatrix} = \begin{bmatrix} \mathrm{x}^k_{T_{k-1}\setminus T_k} \\ 0 \end{bmatrix} \overset{(3.9)}{=} \begin{bmatrix} \mathrm{x}^k_{J_k} \\ 0 \end{bmatrix}, \\ \nabla^2 f(\mathrm{x}^k) = \begin{bmatrix} \nabla^2_{T_k\cup J_k} f(\mathrm{x}^k) & 0 \\ 0 & 0 \end{bmatrix}, \ \ \nabla^2_{T_k\cup J_k} f(\mathrm{x}^k) = \begin{bmatrix} H_k & G_k \\ G_k^\top & \nabla^2_{J_k} f(\mathrm{x}^k) \end{bmatrix}. \end{cases} \tag{3.14}$$

We emphasize that J_k captures all nonzero elements in $\mathrm{x}^k_{\overline{T}_k}$. This and (3.14) also allow us to explain that (3.7) is rewritten as (3.11). Therefore, we will see more J_k instead of $\overline{T}_k$ being used in convergence analysis.

Remark 3.2. With regard to the proposed algorithm, we have some comments.

i) The purpose of initializing $s_{-1} > 0$ is to guarantee a non-empty T_k in each step and $|T_k| = s_k$. To update $\mathrm{x}^{k+1}_{T_k}$, one needs to solve a linear equation (3.11) to derive $\mathrm{d}^k_{T_k}$ with s_k equations and s_k variables. The complexity of solving this equation is at most $\mathcal{O}(|s_k|^3)$.

ii) As stated in [57], a general subspace approach requires

$$\mathrm{x}^{k+1} - \mathrm{x}^k \in \mathcal{S}^k,$$

where $\mathcal{S}^k$ is a subspace in $\mathbb{R}^n$ with the good feature that the dimension of $\mathcal{S}^k$ being much less than n, which allows large scale computation possible [42, 23]. Let $\mathbb{R}^n_T := \mathrm{span}\{\mathrm{e}_i, i \in T\}$ be a subspace of $\mathbb{R}^n$ spanned by $\mathrm{e}_i, i \in T$, where

e_i has i-th element one and zero elements for the rest. The updating rule (3.8) leads to $\mathrm{supp}(\mathrm{x}^{k+1}) \subseteq T_k$ and thus yields $\mathrm{x}^{k+1} - \mathrm{x}^k \in \mathbb{R}^n_{T_k \cup T_{k-1}}$. This is the reason why our method in Algorithm 3.2 is entitled as subspace Newton method. We will show that when k is sufficiently large, $T_k \cup T_{k-1} = T_k$ has a small cardinality comparing with n.

LEMMA 3.3. *If* d^k *is from* (3.11)*, then we have*

$$\begin{aligned}(3.15)\quad & 2\langle g^k_{T_k}, \mathrm{d}^k_{T_k}\rangle + \langle \mathrm{d}^k_{T_k}, H_k \mathrm{d}^k_{T_k}\rangle \\ &\quad = -\langle \mathrm{d}^k_{T_k\cup J_k}, \nabla^2_{T_k\cup J_k} f(\mathrm{x}^k)\mathrm{d}^k_{T_k\cup J_k}\rangle + \langle \mathrm{d}^k_{J_k}, \nabla^2_{J_k} f(\mathrm{x}^k)\mathrm{d}^k_{J_k}\rangle.\end{aligned}$$

Proof. If d^k is from (3.11), then we have the following chain of equations,

$$\begin{aligned}
&\langle \mathrm{d}^k_{T_k\cup J_k}, \nabla^2_{T_k\cup J_k} f(\mathrm{x}^k)\mathrm{d}^k_{T_k\cup J_k}\rangle \\
\overset{(3.14)}{=}\ & \begin{bmatrix} \mathrm{d}^k_{T_k} \\ \mathrm{d}^k_{J_k} \end{bmatrix}^\top \begin{bmatrix} H_k & G_k \\ G_k^\top & \nabla^2_{J_k} f(\mathrm{x}^k) \end{bmatrix} \begin{bmatrix} \mathrm{d}^k_{T_k} \\ \mathrm{d}^k_{J_k} \end{bmatrix} = \begin{bmatrix} \mathrm{d}^k_{T_k} \\ \mathrm{d}^k_{J_k} \end{bmatrix}^\top \begin{bmatrix} H_k \mathrm{d}^k_{T_k} + G_k \mathrm{d}^k_{J_k} \\ G_k^\top \mathrm{d}^k_{T_k} + \nabla^2_{J_k} f(\mathrm{x}^k)\mathrm{d}^k_{J_k} \end{bmatrix} \\
\overset{(3.14)}{=}\ & \langle \mathrm{d}^k_{T_k}, H_k \mathrm{d}^k_{T_k} - G_k \mathrm{x}^k_{J_k}\rangle - \langle \mathrm{x}^k_{J_k}, G_k^\top \mathrm{d}^k_{T_k}\rangle + \langle \mathrm{d}^k_{J_k}, \nabla^2_{J_k} f(\mathrm{x}^k)\mathrm{d}^k_{J_k}\rangle \\
=\ & 2\langle \mathrm{d}^k_{T_k}, H_k \mathrm{d}^k_{T_k} - G_k \mathrm{x}^k_{J_k}\rangle - \langle H_k \mathrm{d}^k_{T_k}, \mathrm{d}^k_{T_k}\rangle + \langle \mathrm{d}^k_{J_k}, \nabla^2_{J_k} f(\mathrm{x}^k)\mathrm{d}^k_{J_k}\rangle \\
\overset{(3.11)}{=}\ & -2\langle g^k_{T_k}, \mathrm{d}^k_{T_k}\rangle - \langle \mathrm{d}^k_{T_k}, H_k \mathrm{d}^k_{T_k}\rangle + \langle \mathrm{d}^k_{J_k}, \nabla^2_{J_k} f(\mathrm{x}^k)\mathrm{d}^k_{J_k}\rangle,
\end{aligned}$$

which conclude our claim immediately. ∎

Lemma 3.3 indicates that if $\nabla^2_{T_k\cup J_k} f(\mathrm{x}^k)$ has a positive lower and upper bound, so is H_k bounded from below and $\nabla^2_{J_k} f(\mathrm{x}^k)$ bounded from above, then (3.10) is satisfied in each step under some properly chosen δ and τ. This allows the Newton direction to be always imposed. Apparently, $\nabla^2_{T_k\cup J_k} f(\mathrm{x}^k)$ being bounded from below can be guaranteed by some assumptions, such as the strong convexity of f, which, however, is a strong assumption. To overcome this, the gradient direction compensates the case when the condition (3.10) is violated.

3.2. Convergence Analysis. Before our main results, we would like to define some parameters. Denote

$$\begin{aligned}
\overline{\alpha} &:= \min\left\{\frac{1-2\sigma}{L/\delta - \sigma}, \frac{2(1-\sigma)\delta}{L}, \ \ 1\right\}, \\
\overline{\tau} &:= \min\left\{\frac{2\overline{\alpha}\delta\beta}{L^2}, \ \ \overline{\alpha}\beta, \ \ \frac{1}{4L}\right\}, \qquad \rho := \min\left\{\frac{2\delta - \tau L^2}{2}, \ \ \frac{2-\tau}{2}\right\}.
\end{aligned}$$

Our first result shows that the direction in each step of SNL0 is a descent one with a decent declining rate, no matter it is taken from the Newton or the gradient direction.

LEMMA 3.4 (Descent property). *Let f be strongly smooth with $L > 0$ and $\overline{\tau}, \rho$ be defined as* (3.16)*. Then for any $\tau \in (0, \overline{\tau})$, it holds $\rho > 0$ and*

$$(3.16)\qquad \langle g^k, \mathrm{d}^k\rangle \le -\rho\|\mathrm{d}^k\|^2 - \frac{\tau}{2}\|g^k_{T_{k-1}}\|^2.$$

Proof. It follows from (3.16) that $\overline{\alpha}\beta < 1$ and hence $\overline{\tau} \le \min\left\{2\delta/L^2, 2\right\}$, which immediately shows $\rho > 0$ if $\tau \in (0, \overline{\tau})$. In addition, if d^k is updated by (3.11), then

$$(3.17)\qquad \|g^k_{T_k}\| \overset{(3.11)}{=} \|H_k \mathrm{d}^k_{T_k} - G_k \mathrm{x}^k_{J_k}\| \overset{(3.14)}{=} \|[H_k\ G_k]\mathrm{d}^k_{T_k\cup J_k}\| \overset{(3.14)}{\le} L\|\mathrm{d}^k\|,$$

where the inequality holds because of $\|[H_k\ G_k]\|_2 \leq \|\nabla^2_{T_k\cup J_k} f(\mathrm{x}^k)\|_2 \leq L$ due to strong smoothness of f with the constant L. We now take two cases: **Case i)** $S_k := T_k \setminus T_{k-1} = \emptyset$ and **Case ii)** $S_k \neq \emptyset$ into consideration.

Case i) $S_k = \emptyset$. Lemma 3.1 1) that $|T_k| \geq |T_{k-1}|$ results in $T_k = T_{k-1}$. Consequently, $J_k = T_{k-1} \setminus T_k = \emptyset$ and $\mathrm{d}^k_{\overline{T}_k} = -\mathrm{x}^k_{\overline{T}_k} = 0$ from (3.14). If d^k is updated by (3.11), the following chain of inequalities holds

$$\begin{aligned}
2\langle g^k, \mathrm{d}^k\rangle &= 2\langle g^k_{T_k}, \mathrm{d}^k_{T_k}\rangle - 2\langle g^k_{\overline{T}_k}, \mathrm{x}^k_{\overline{T}_k}\rangle \qquad = 2\langle g^k_{T_k}, \mathrm{d}^k_{T_k}\rangle \\
&\overset{(3.10)}{\leq} -2\delta\|\mathrm{d}^k\|^2 + \|\mathrm{x}^k_{\overline{T}_k}\|^2/(2\tau) \qquad = -2\delta\|\mathrm{d}^k\|^2 + \tau\|g^k_{T_k}\|^2 - \tau\|g^k_{T_k}\|^2 \\
&\overset{(3.17)}{\leq} -[2\delta - \tau L^2]\|\mathrm{d}^k\|^2 - \tau\|g^k_{T_k}\|^2 \overset{(3.16)}{\leq} -2\rho\|\mathrm{d}^k\|^2 - \tau\|g^k_{T_{k-1}}\|^2,
\end{aligned} \tag{3.18}$$

where the last inequality holds due to $T_k = T_{k-1}$. If d^k is updated by (3.12), then it follows from $\mathrm{d}^k_{T_k} = -g^k_{T_k} = -g^k_{T_{k-1}}$ that

$$\begin{aligned}
2\langle g^k, \mathrm{d}^k\rangle &\overset{(3.18)}{=} 2\langle g^k_{T_k}, \mathrm{d}^k_{T_k}\rangle \quad = -2\|\mathrm{d}^k_{T_k}\|^2 \quad = -(2-\tau)\|\mathrm{d}^k_{T_k}\|^2 - \tau\|\mathrm{d}^k_{T_k}\|^2 \\
&= -(2-\tau)\|\mathrm{d}^k\|^2 - \tau\|g^k_{T_{k-1}}\|^2 \overset{(3.16)}{\leq} -2\rho\|\mathrm{d}^k\|^2 - \tau\|g^k_{T_{k-1}}\|^2.
\end{aligned} \tag{3.19}$$

Case ii) $S_k \neq \emptyset$. For any $i \in S_k = T_k \setminus T_{k-1}$, we have $x^k_i = 0$ because of $\mathrm{supp}(\mathrm{x}^k) \subseteq T_{k-1}$ by (3.14). Then (3.5) in Lemma 3.1 gives rise to

$$\forall i \in S_k, \quad |\tau g^k_i|^2 = |x^k_i - \tau g^k_i|^2 \geq 2\tau\lambda_k \geq |x^k_j - \tau g^k_j|^2, \quad \forall j \in J_k. \tag{3.20}$$

Again $|T_k| \geq |T_{k-1}|$ drives that $|S_k| = |T_k| - |T_k \cap T_{k-1}| \geq |T_{k-1}| - |T_k \cap T_{k-1}| = |J_k|$. This suffices to

$$\begin{aligned}
&\tau^2\left[\|g^k_{T_k}\|^2 - \|g^k_{T_k\cap T_{k-1}}\|^2\right] = \tau^2\|g^k_{S_k}\|^2 \\
\overset{(3.20)}{\geq}\ & |S_k|2\tau\lambda_k \geq |J_k|2\tau\lambda_k = \sum_{j\in J_k} 2\tau\lambda_k \overset{(3.20)}{\geq} \sum_{j\in J_k} |x^k_j - \tau g^k_j|^2 = \|\mathrm{x}^k_{J_k} - \tau g^k_{J_k}\|^2 \\
=\ & \|\mathrm{x}^k_{J_k}\|^2 - 2\tau\langle \mathrm{x}^k_{J_k}, g^k_{J_k}\rangle + \tau^2\|g^k_{J_k}\|^2 \overset{(3.14)}{=} \|\mathrm{x}^k_{\overline{T}_k}\|^2 - 2\tau\langle \mathrm{x}^k_{J_k}, g^k_{J_k}\rangle + \tau^2\|g^k_{J_k}\|^2 \\
=\ & \|\mathrm{x}^k_{\overline{T}_k}\|^2 - 2\tau\langle \mathrm{x}^k_{J_k}, g^k_{J_k}\rangle + \tau^2\left[\|g^k_{T_{k-1}}\|^2 - \|g^k_{T_k\cap T_{k-1}}\|^2\right]
\end{aligned}$$

and thus results in our first fact

$$-2\langle \mathrm{x}^k_{J_k}, g^k_{J_k}\rangle \leq \tau\|g^k_{T_k}\|^2 - \tau\|g^k_{T_{k-1}}\|^2 - \|\mathrm{x}^k_{\overline{T}_k}\|^2/\tau \tag{3.21}$$

$$\overset{(3.17)}{\leq} \tau L^2\|\mathrm{d}^k\|^2 - \tau\|g^k_{T_{k-1}}\|^2 - \|\mathrm{x}^k_{\overline{T}_k}\|^2/\tau. \tag{3.22}$$

Now we are ready to establish our claim. If d^k is updated by (3.11), then

$$2\langle g^k_{T_k}, \mathrm{d}^k_{T_k}\rangle \overset{(3.10)}{\leq} -2\delta\|\mathrm{d}^k\|^2 + \|\mathrm{x}^k_{\overline{T}_k}\|^2/(2\tau). \tag{3.23}$$

The direct calculation yields the following chain of inequalities,

$$\begin{aligned}
2\langle g^k, \mathrm{d}^k\rangle &= 2\langle g^k_{T_k}, \mathrm{d}^k_{T_k}\rangle - 2\langle g^k_{\overline{T}_k}, \mathrm{x}^k_{\overline{T}_k}\rangle \overset{(3.14)}{=} 2\langle g^k_{T_k}, \mathrm{d}^k_{T_k}\rangle - 2\langle g^k_{J_k}, \mathrm{x}^k_{J_k}\rangle \\
&\overset{(3.23),(3.22)}{\leq} -[2\delta - \tau L^2]\|\mathrm{d}^k\|^2 - \|\mathrm{x}^k_{\overline{T}_k}\|^2/(2\tau) - \tau\|g^k_{T_{k-1}}\|^2 \\
&\overset{(3.16)}{\leq} -2\rho\|\mathrm{d}^k\|^2 - \tau\|g^k_{T_{k-1}}\|^2.
\end{aligned}$$

If d^k is updated by (3.12), then $\mathrm{d}^k_{T_k} = -g^k_{T_k}$ yields that

$$\begin{aligned}
2\langle g^k, \mathrm{d}^k\rangle &= 2\langle g^k_{T_k}, \mathrm{d}^k_{T_k}\rangle - 2\langle g^k_{\overline{T}_k}, \mathrm{x}^k_{\overline{T}_k}\rangle = -2\|\mathrm{d}^k_{T_k}\|^2 - 2\langle g^k_{J_k}, \mathrm{x}^k_{J_k}\rangle \\
&\overset{(3.21)}{\leq} -2\|\mathrm{d}^k_{T_k}\|^2 + \tau\|g^k_{T_k}\|^2 - \|\mathrm{x}^k_{\overline{T}_k}\|^2/\tau - \tau\|g^k_{T_{k-1}}\|^2 \\
&\overset{(3.14)}{=} -(2-\tau)\|\mathrm{d}^k_{T_k}\|^2 - \|\mathrm{d}^k_{\overline{T}_k}\|^2/\tau - \tau\|g^k_{T_{k-1}}\|^2 \\
&\leq -(2-\tau)(\|\mathrm{d}^k_{T_k}\|^2 + \|\mathrm{d}^k_{\overline{T}_k}\|^2) - \tau\|g^k_{T_{k-1}}\|^2 \overset{(3.16)}{\leq} -2\rho\|\mathrm{d}^k\|^2 - \tau\|g^k_{T_{k-1}}\|^2,
\end{aligned}$$

where the second inequality is from $-1/\tau \leq \tau - 2$ for any $\tau > 0$. ◻

Our next result shows that α_k exists and is bound away from zero. This means the step length to update next point is well defined and would not be too small, which is expected to speed up the convergence.

Lemma 3.5 (Existence and boundedness of α_k). *Let f be strongly smooth with $L > 0$ and $\overline{\alpha}, \overline{\tau}$ be defined as (3.16). Then*

$$f(\mathrm{x}^k(\alpha)) \leq f(\mathrm{x}^k) + \sigma\alpha\langle g^k, \mathrm{d}^k\rangle \tag{3.24}$$

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

$$0 < \alpha \leq \overline{\alpha}, \quad 0 < \delta \leq \min\{1, 2L\}, \quad 0 < \tau \leq \min\left\{\alpha\delta/L^2,\ \alpha,\ 1/(4L)\right\}.$$

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

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

$$\alpha \leq \frac{2(1-\sigma)\delta}{L}, \quad \alpha \leq \frac{1-2\sigma}{L/\delta - \sigma} \quad \text{and} \quad \alpha \leq \frac{1-2\sigma}{L-\sigma}.$$

Since f is strongly smooth, we obtain that

$$\begin{aligned}
&2f(\mathrm{x}^k(\alpha)) - 2f(\mathrm{x}^k) - 2\alpha\sigma\langle g^k, \mathrm{d}^k\rangle \\
\overset{(2.4)}{\leq}\ & 2\langle g^k, \mathrm{x}^k(\alpha) - \mathrm{x}^k\rangle + L\|\mathrm{x}^k(\alpha) - \mathrm{x}^k\|^2 - 2\alpha\sigma\langle g^k, \mathrm{d}^k\rangle \\
\overset{(3.8)}{=}\ & \alpha(1-\sigma)2\langle g^k_{T_k}, \mathrm{d}^k_{T_k}\rangle - (1-\alpha\sigma)2\langle g^k_{\overline{T}_k}, \mathrm{x}^k_{\overline{T}_k}\rangle + L\left[\alpha^2\|\mathrm{d}^k_{T_k}\|^2 + \|\mathrm{x}^k_{\overline{T}_k}\|^2\right] \\
\overset{(3.14)}{=}\ & \alpha(1-\sigma)2\langle g^k_{T_k}, \mathrm{d}^k_{T_k}\rangle - (1-\alpha\sigma)2\langle g^k_{J_k}, \mathrm{x}^k_{J_k}\rangle + L\left[\alpha^2\|\mathrm{d}^k_{T_k}\|^2 + \|\mathrm{x}^k_{\overline{T}_k}\|^2\right] =: \psi.
\end{aligned}$$

To conclude the conclusion, one needs to show $\psi \leq 0$. Similar to the proof of Lemma (3.4), we consider two cases: **Case i)** $S_k = T_k \setminus T_{k-1} = \emptyset$ and **Case ii)** $S_k \neq \emptyset$.

Case i) $S_k = \emptyset$. This case indicates $J_k = \emptyset$ from the proof of Lemma (3.4). If d^k is updated by (3.11), then we obtain

$$\begin{aligned}
\psi &= \alpha(1-\sigma)2\langle g^k_{T_k}, \mathrm{d}^k_{T_k}\rangle + L\alpha^2\|\mathrm{d}^k_{T_k}\|^2 \\
&\begin{cases} \overset{(3.10)}{\leq} -2\alpha(1-\sigma)\delta\|\mathrm{d}^k\|^2 + L\alpha^2\|\mathrm{d}^k_{T_k}\|^2, & \text{if } \mathrm{d}^k \text{ is from (3.11)} \\ \overset{(3.12)}{=} -2\alpha(1-\sigma)\|\mathrm{d}^k_{T_k}\|^2 + L\alpha^2\|\mathrm{d}^k_{T_k}\|^2, & \text{if } \mathrm{d}^k \text{ is from (3.12)} \end{cases}
\end{aligned}$$

$$\leq -2\alpha(1-\sigma)\delta\|\mathrm{d}^k\|^2 + L\alpha^2\|\mathrm{d}^k\|^2 = \alpha[L\alpha - 2(1-\sigma)\delta]\|\mathrm{d}^k\|^2, \tag{3.25}$$

where the third inequality is due to $-1 \leq -\delta$, $\|\mathrm{d}^k\|^2 = \|\mathrm{d}^k_{T_k}\|^2$. This suffices to $\psi \leq 0$ since $\alpha \leq 2(1-\sigma)\delta/L$.

Case ii) $S_k \neq \emptyset$. If d^k is from (3.11), then we have

$$\begin{aligned}
\psi &\overset{(3.23),(3.22)}{\leq} && \alpha(1-\sigma)[-2\delta\|\mathrm{d}^k\|^2 + \frac{1}{2\tau}\|\mathrm{x}^k_{\overline{T}_k}\|^2] + L\alpha^2\|\mathrm{d}^k_{T_k}\|^2 \\
&+ && (1-\alpha\sigma)[\tau L^2\|\mathrm{d}^k\|^2 - \tau\|g^k_{T_{k-1}}\|^2 - \frac{1}{\tau}\|\mathrm{x}^k_{\overline{T}_k}\|^2] + L\|\mathrm{x}^k_{\overline{T}_k}\|^2 \\
&\leq && c_1\|\mathrm{d}^k\|^2 + c_2\|\mathrm{x}^k_{\overline{T}_k}\|^2 - (1-\alpha\sigma)\tau\|g^k_{T_{k-1}}\|^2 \leq c_1\|\mathrm{d}^k\|^2 + c_2\|\mathrm{x}^k_{\overline{T}_k}\|^2,
\end{aligned}$$

where $1-\alpha\sigma > 0$ due to $0<\alpha<1, 0<\sigma\leq 1/2$ and c_1 and c_2 are given by

$$\begin{aligned}
c_1 &:= -\alpha(1-\sigma)2\delta + (1-\alpha\sigma)\tau L^2 + L\alpha^2, \\
&\leq -\alpha(1-\sigma)2\delta + (1-\alpha\sigma)\delta\alpha + L\alpha^2 && \text{because of } \alpha\leq 1, \sigma\leq\frac{1}{2}, \tau\leq\frac{\alpha\delta}{L^2} \\
&= \alpha\left[(L-\sigma\delta)\alpha - (1-2\sigma)\delta\right] \leq 0, && \text{because of } \sigma\delta\leq L, \alpha\leq\frac{1-2\sigma}{L/\delta-\sigma} \\
c_2 &:= \alpha(1-\sigma)/(2\tau) - (1-\alpha\sigma)/\tau + L \\
&\leq (1-\alpha\sigma)/(2\tau) - (1-\alpha\sigma)/\tau + L && \text{because of } \alpha\leq 1 \\
&\leq -(1-\alpha\sigma)/(2\tau) + L \leq 0. && \text{because of } \alpha\leq 1, \sigma\leq\frac{1}{2}, \tau\leq\frac{1}{4L}
\end{aligned}$$

If d^k is updated by (3.12), namely $\mathrm{d}^k_{T_k} = -g^k_{T_k}$, then

$$\begin{aligned}
\psi &\overset{(3.12),(3.21)}{\leq} && -2\alpha(1-\sigma)\|\mathrm{d}^k_{T_k}\|^2 + L\alpha^2\|\mathrm{d}^k_{T_k}\|^2 \\
&+ && (1-\alpha\sigma)[\tau\|g^k_{T_k}\|^2 - \tau\|g^k_{T_{k-1}}\|^2 - \frac{1}{\tau}\|\mathrm{x}^k_{\overline{T}_k}\|^2] + L\|\mathrm{x}^k_{\overline{T}_k}\|^2 \\
&\overset{(3.12)}{\leq} && c_3\|\mathrm{d}^k_{T_k}\|^2 + c_4\|\mathrm{x}^k_{\overline{T}_k}\|^2 - (1-\alpha\sigma)\tau\|g^k_{T_{k-1}}\|^2 \leq c_3\|\mathrm{d}^k_{T_k}\|^2 + c_4\|\mathrm{x}^k_{\overline{T}_k}\|^2,
\end{aligned}$$

where c_3 and c_4 are given by

$$\begin{aligned}
c_3 &:= -2\alpha(1-\sigma) + (1-\alpha\sigma)\tau + L\alpha^2 \\
&\leq \alpha\left[(L-\sigma)\alpha - (1-2\sigma)\right] \leq 0 && \text{because of } \alpha\leq 1, \sigma\leq\frac{1}{2}, \tau\leq\alpha, \alpha\leq\frac{1-2\sigma}{L-\sigma} \\
c_4 &:= -(1-\alpha\sigma)/\tau + L \\
&\leq -1/(2\tau) + L \leq 0, && \text{because of } \alpha\leq 1, \sigma\leq\frac{1}{2}, \tau\leq\frac{1}{4L}
\end{aligned}$$

which finishes the proof of the first claim. If further $\tau\in(0,\overline{\tau})$ where $\overline{\tau}$ is defined as (3.16), then for any $\beta\overline{\alpha}\leq\alpha\leq\overline{\alpha}$, we have

$$0 < \tau < \min\left\{\overline{\alpha}\delta\beta/L^2,\ \overline{\alpha}\beta,\ 1/(4L)\right\} \leq \min\left\{\alpha\delta/L^2,\ \alpha,\ 1/(4L)\right\}.$$

This together with (3.24) shows

$$f(\mathrm{x}^k(\alpha)) - f(\mathrm{x}^k) \leq \sigma\alpha\langle g^k, \mathrm{d}^k\rangle. \tag{3.26}$$

Finally, the Armijo-type step size rule means that $\{\alpha_k\}$ is bounded from below by a positive constant, that is,

$$\inf_{k\geq 0}\{\alpha_k\} \geq \beta\overline{\alpha} > 0. \tag{3.27}$$

The whole proof is completed. ∎

Remark 3.6. Regarding to the proof of above lemma, we can easily see that if for some iterations k such that $J_k = \emptyset$, namely $T_k = T_{k-1}$, then (3.25) shows that if $0 < \alpha \leq 2(1-\sigma)\delta/L$, then (3.24) holds. This means there is no restriction on $\tau > 0$ by an upper bound $\overline{\tau}$ which is related to δ.

Lemma 3.5 allows us to conclude that the objective f is strictly decreasing for each step, and the difference of two consecutive iterates and the entries of the stationary equation will vanish.

LEMMA 3.7. *Let f be strongly smooth with $L > 0$ and $\overline{\tau}$ be defined as* (3.16). *Let* $\{\mathrm{x}^k\}$ *be the sequence generated by* SNL0 *with $\tau \in (0, \overline{\tau})$ and $\delta \in (0, \min\{1, 2L\})$. Then* $\{f(\mathrm{x}^k)\}$ *is a strictly nonincreasing sequence and*

$$\lim_{k\to\infty} \max\left\{ \ \|F_\tau(\mathrm{x}^k; T_k)\|, \ \|\mathrm{x}^{k+1} - \mathrm{x}^k\|, \ \|g^k_{T_{k-1}}\|, \ \|g^k_{T_k}\| \ \right\} = 0.$$

Proof. By (3.26), (3.16) and denoting $c_0 := \sigma\overline{\alpha}\beta\rho$, we have

$$\begin{aligned} f(x^{k+1}) - f(x^k) \leq \sigma\alpha_k\langle g^k, \mathrm{d}^k\rangle &\overset{(3.16)}{\leq} -\sigma\alpha_k\rho\|\mathrm{d}^k\|^2 - \frac{\tau}{2}\|g^k_{T_{k-1}}\|^2 \\ &\overset{(3.27)}{\leq} -c_0\|\mathrm{d}^k\|^2 - \frac{\tau}{2}\|g^k_{T_{k-1}}\|^2. \end{aligned}$$

Then it follows from the above inequality that

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

where the last inequality is due to f being bounded from below. Hence $\lim_{k\to\infty}\|\mathrm{d}^k\| = \lim_{k\to\infty}\|g^k_{T_{k-1}}\| = 0$, which suffices to $\lim_{k\to\infty}\|\mathrm{x}^{k+1} - \mathrm{x}^k\| = 0$ because of

$$\|\mathrm{x}^{k+1} - \mathrm{x}^k\|^2 \overset{(3.8)}{=} \alpha_k^2\|\mathrm{d}^k_{T_k}\|^2 + \|\mathrm{x}^k_{\overline{T}_k}\|^2 \leq \|\mathrm{d}^k_{T_k}\|^2 + \|\mathrm{d}^k_{\overline{T}_k}\|^2 = \|\mathrm{d}^k\|^2.$$

In addition, if d^k is taken from (3.11), then $\|g^k_{T_k}\| \leq L\|\mathrm{d}^k\|$ by (3.17). If it is taken from (3.12) then $\|g^k_{T_k}\| = \|\mathrm{d}^k_{T_k}\|$. Those together with (2.12) that $\|F_\tau(\mathrm{x}^k; T_k)\|^2 = \|g^k_{T_k}\|^2 + \|\mathrm{x}^k_{\overline{T}_k}\|^2 \leq (L^2+1)\|\mathrm{d}^k\|^2 \to 0$. The whole proof is completed. ∎

We are ready to conclude from Lemma 3.7 that the support set of $\{\mathrm{x}^k\}$ can be identified within finite steps and the sequence converges to a τ-stationary point or a local minimizer globally, which are presented by the following theorem.

THEOREM 3.8 (Convergence and support sets identification). *Let f be strongly smooth with $L > 0$ and $\overline{\tau}$ be defined as* (3.16). *Let* $\{\mathrm{x}^k\}$ *be the sequence generated by* SNL0 *with $\tau \in (0, \overline{\tau})$ and $\delta \in (0, \min\{1, 2L\})$. Then the following results hold.*

1) *The whole sequence $\{s_k\}$ converges (to s_*).*
2) *Any accumulating point (say x^*) of the sequence $\{\mathrm{x}^k\}$ is a τ-stationary point of the following problem with $\lambda_* := \|\mathrm{x}^*\|^2_{[s_*]}/(2\tau)$,*

$$\min_{\mathrm{x}} \ f(\mathrm{x}) + \lambda_*\|\mathrm{x}\|_0. \tag{3.28}$$

Furthermore, x^* is a local minimizer of the above model if f is convex.

3) *If x^* is isolated, then the whole sequence converges to x^*. Moreover, for any sufficiently large k,*

a) if $\|x^*\|_0 = s_*$*, then* $\nabla_{\mathrm{supp}(x^*)} f(x^*) = 0$ *and* $\mathrm{supp}(x^k) \equiv T_k \equiv \mathrm{supp}(x^*)$.
b) if $\|x^*\|_0 < s_*$*, then* $\nabla f(x^*) = 0$ *and* $\mathrm{supp}(x^*) \subseteq (\mathrm{supp}(x^k) \cap T_k)$.

Proof. 1) We already proved in Lemma 3.7 that

$$\lim_{k\to\infty} g^k_{T_{k-1}} = \lim_{k\to\infty} g^k_{T_k} = \lim_{k\to\infty} g^{k+1}_{T_k} = 0. \tag{3.29}$$

This yields $\|g^k_{T_{k-1}}\| < \epsilon$ for sufficiently large k, and thus $s_k = s_{k-1}$ from (3.2). Namely, $\{s_k\}$ converges to s_*, which and Lemma 3.1 indicate that, for sufficiently large k,

$$|T_k| = s_k = s_{k+1} = |T_{k+1}| = \cdots = s_*. \tag{3.30}$$

2) Lemma 3.1 states that T_k is non-empty and

$$|x^k_i - \tau g^k_i| \geq \sqrt{2\tau\lambda_k} \geq |x^k_j - \tau g^k_j|, \quad \forall\, i \in T_k, \forall\, j \in \overline{T}_k, \tag{3.31}$$

Let $\{x^{k_t}\}$ be the convergent subsequence of $\{x^k\}$ that converges to x^*. Since $\{T_{k_t}\} \subseteq \mathbb{N}_n$ and $\mathbb{N}_n$ has finite elements, it must have a subsequence $\{T_{k_t}\}_{k_t \in \mathcal{I}}$, where $\mathcal{I} \subseteq \{k_1, k_2, \cdots\}$ satisfying that $T_{k_t} \equiv T_\infty, \forall k_t \in \mathcal{I}$ for sufficiently large k_t. Then we consider the subsequence $\{x^{k_t}\}_{k_t \in \mathcal{I}}$. Therefore, for notational convenience, without loss of any generality, we focus on $\{x^{k_t}\}$ itself and assume

$$T_{k_t} \equiv T_{k_{t+1}} \equiv \cdots \equiv: T_* \quad \text{with} \quad s_* \overset{(3.30)}{=} |T_*| \tag{3.32}$$

for sufficiently large k_t. Since $x^{k_t} \to x^*$ and $\|x^{k+1} - x^k\| \to 0$ from Lemma 3.7, we have $x^{k_t+1} \to x^*$ and thus $\mathrm{supp}(x^*) \subseteq \mathrm{supp}(x^{k_t+1})$ (see the proof of [59, Theorem 8]). Then it follows from $\mathrm{supp}(x^{k_t+1}) \subseteq T_{k_t} \equiv T_*$ by (3.14) that

$$\mathrm{supp}(x^*) \subseteq \mathrm{supp}(x^{k_t+1}) \subseteq T_*, \tag{3.33}$$

which indicates

$$\nabla_{T_*} f(x^*) \overset{(3.32)}{=} \nabla_{T_{k_t}} f(x^*) = \lim_{k_t\to\infty} \nabla_{T_{k_t}} f(x^{k_t}) = \lim_{k_t\to\infty} g^{k_t}_{T_{k_t}} \overset{(3.29)}{=} 0. \tag{3.34}$$

In addition,

$$\begin{aligned} \lim_{k_t\to\infty} \lambda_{k_t} &\overset{(3.3)}{=} \lim_{k_t\to\infty} \frac{1}{2\tau}\|x^{k_t} - \tau g^{k_t}\|^2_{[s_{k_t}]} \overset{(3.4)}{=} \lim_{k_t\to\infty} \frac{1}{2\tau}\|x^{k_t}_{T_{k_t}} - \tau g^{k_t}_{T_{k_t}}\|^2_{[s_{k_t}]} \\ &\overset{(3.29)}{=} \frac{1}{2\tau}\|x^*_{T_*}\|^2_{[s_*]} \overset{(3.33)}{=} \frac{1}{2\tau}\|x^*\|^2_{[s_*]} = \lambda_*. \end{aligned} \tag{3.35}$$

Now for any $i \in T_{k_t} \overset{(3.32)}{=} T_*,\ j \in \overline{T}_{k_t} \overset{(3.32)}{=} \overline{T}_*$, we have

$$\begin{aligned} |x^*_i| &\overset{(3.34)}{=} |x^*_i - \tau\nabla_i f(x^*)| = \lim_{k_t\to\infty} |x^{k_t}_i - \tau g^{k_t}_i| \overset{(3.31)}{\geq} \lim_{k_t\to\infty} \sqrt{2\tau\lambda_{k_t}} \overset{(3.35)}{=} \sqrt{2\tau\lambda^*} \\ &\overset{(3.31)}{\geq} \lim_{k_t\to\infty} |x^{k_t}_j - \tau g^{k_t}_j| = |x^*_j - \tau\nabla_j f(x^*)| \overset{(3.33)}{=} \tau|\nabla_j f(x^*)|, \end{aligned}$$

which leads to

$$|x^*_i| \geq \sqrt{2\tau\lambda^*} \text{ and } |\nabla_j f(x^*)| \leq \sqrt{2\lambda^*/\tau}. \tag{3.36}$$

If $\|x^*\|_0 = s_*$, then we must have $\mathrm{supp}(x^*) = T_*$ from (3.33) and (3.32). This together with (3.36), (3.34) that $\nabla_{T_*} f(x^*) = 0$ and Lemma (2.1) show x^* is a τ-stationary point of the problem (3.28). If $\|x^*\|_0 < s_*$, then $\lambda_* = \frac{1}{2\tau}\|x^*\|^2_{[s_*]} = 0$ from (3.35) and hence $\nabla_j f(x^*) = 0, \forall j \in \overline{T}_*$. Recall (3.34) that $\nabla_{T_*} f(x^*) = 0$, giving $\nabla f(x^*) = 0$ and thus showing (2.3), a τ-stationary point. Finally, a τ-stationary point is a local minimizer if f is convex by Theorem 2.3.

3) The whole sequence converges because of x^* being isolated, [40, Lemma 4.10] and $\|x^{k+1} - x^k\| \to 0$ from Lemma 3.7. Since $x^k \to x^*$, we must have $T_* := \mathrm{supp}(x^*) \subseteq \mathrm{supp}(x^k) \subseteq T_{k-1}$(see the proof of [59, Theorem 8]). If $\|x^*\|_0 = s_*$, then

$$s_* = |T_*| \leq |\mathrm{supp}(x^k)| \leq |T_{k-1}| \overset{(3.30)}{=} s_*,$$

which yields $|\mathrm{supp}(x^k)| = s_*$. Therefore we have $\mathrm{supp}(x^k) = T_{k-1} = T_*$ by $T_* \subseteq \mathrm{supp}(x^k) \subseteq T_{k-1}$. Similar reasons allow us to show $\mathrm{supp}(x^{k+1}) = T_k = T_*$. Overall, $\mathrm{supp}(x^k) = T_k = T_*$. If $\|x^*\|_0 < s_*$, then $T_* \subseteq \mathrm{supp}(x^{k+1}) \subseteq T_k$. This together with $T_* \subseteq \mathrm{supp}(x^k)$ brings out $T_* \subseteq (\mathrm{supp}(x^k) \cap T_k)$. The whole proof is finished. ∎

Finally, we would like to see how fast our proposed method SNL0 converges. To proceed that, we need the locally Lipschitz continuity. We say the Hessian of f is locally Lipschitz continuous around x with constant $M > 0$ if for any points z in the neighbourhood of x, it has

$$\|\nabla^2 f(z) - \nabla^2 f(x)\|_2 \leq M\|z - x\|.$$

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 ℓ depend on the point x. Now we are able to establish the following results.

THEOREM 3.9 (Global and quadratic convergence). *Let* $\{x^k\}$ *be the sequence generated by* SNL0 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^*. Choose $\tau \in (0, \overline{\tau})$ and $\delta \in (0, \min\{1, \ell\})$. Then the following results hold.

1) The whole sequence converges to x^*, namely, x^* is the limit point.
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,

$$\|x^{k+1} - x^*\| \leq M/(2\ell)\|x^k - x^*\|^2, \tag{3.37}$$

$$\|F_\tau(x^{k+1}; T_{k+1})\| \leq \frac{M\sqrt{L^2+1}}{\min\{\ell^3, \ell\}}\|F_\tau(x^k; T_k)\|^2. \tag{3.38}$$

Proof. 1) Denote $T_* := \mathrm{supp}(x^*)$. Theorem 3.8 shows that $s_k \to s_*$, $\nabla_{T_*} f(x^*) = 0$ if $\|x^*\|_0 = s_*$ and $\nabla f(x^*) = 0$ if $\|x^*\|_0 < s_*$. Consider a local region $N(x^*) := \{x \in \mathbb{R}^n : \mathrm{supp}(x) \subseteq T_*\}$ if $\|x^*\|_0 = s_*$ and $N(x^*) = \mathbb{R}^n$ if $\|x^*\|_0 < s_*$. Then these facts together with f being locally strongly convex with $\ell > 0$ around x^* derive that, for any $x(\neq x^*) \in N(x^*)$,

$$\begin{aligned} f(x) - f(x^*) &\geq \langle \nabla f(x^*), x - x^* \rangle + (\ell/2)\|x - x^*\|^2 \\ &> \langle \nabla_{T_*} f(x^*), (x - x^*)_{T_*} \rangle + \langle \nabla_{\overline{T}_*} f(x^*), (x - x^*)_{\overline{T}_*} \rangle \\ &= \langle \nabla_{\overline{T}_*} f(x^*), x_{\overline{T}_*} \rangle = \begin{cases} \langle \nabla_{\overline{T}_*} f(x^*), 0 \rangle, & \|x^*\|_0 = s_* \\ \langle 0, x_{\overline{T}_*} \rangle, & \|x^*\|_0 < s_*. \end{cases} = 0. \end{aligned}$$

This means x^* is a strictly local minimizer of (3.28) and also shows x^* is isolated. Therefore, the whole sequence tends to x^* by Theorem 3.8 2).

2) We first verify H_k is nonsingular when k is sufficiently large and

$$\langle g^k_{T_k}, \mathrm{d}^k_{T_k}\rangle \le -\delta\|\mathrm{d}^k\|^2 + \|\mathrm{x}^k_{\overline{T}_k}\|^2/(4\tau).$$

Let I_n be the identity matrix with order n. Since f is strongly smooth with L and locally strongly convex with ℓ in a neighbourhood of x^*, we have

$$\ell \le \lambda_i(\nabla^2_{T_k\cup J_k} f(\mathrm{x}^k)), \lambda_i(H_k), \lambda_i(\nabla^2_{J_k} f(\mathrm{x}^k)) \le L, \tag{3.39}$$

where $\lambda_i(A)$ is the ith largest eigenvalue of A. Now by (3.15), we have

$$\begin{aligned} 2\langle g^k_{T_k}, \mathrm{d}^k_{T_k}\rangle &= -\langle \mathrm{d}^k_{T_k\cup J_k}, \nabla^2_{T_k\cup J_k} f(\mathrm{x}^k)\mathrm{d}^k_{T_k\cup J_k}\rangle \\ &\quad -\langle H_k \mathrm{d}^k_{T_k}, \mathrm{d}^k_{T_k}\rangle + \langle \mathrm{d}^k_{J_k}, \nabla^2_{J_k} f(\mathrm{x}^k)\mathrm{d}^k_{J_k}\rangle \\ &\overset{(3.39)}{\le} -\ell\left[\|\mathrm{d}^k_{T_k\cup J_k}\|^2 + \|\mathrm{d}^k_{T_k}\|^2\right] + L\|\mathrm{x}^k_{\overline{T}_k}\|^2 \\ &= -\ell\left[\|\mathrm{d}^k_{T_k\cup J_k}\|^2 + \|\mathrm{d}^k_{T_k}\|^2 + \|\mathrm{d}^k_{J_k}\|^2 - \|\mathrm{d}^k_{J_k}\|^2\right] + L\|\mathrm{x}^k_{\overline{T}_k}\|^2 \\ &= -2\ell\|\mathrm{d}^k_{T_k\cup J_k}\|^2 + \ell\|\mathrm{d}^k_{J_k}\|^2 + L\|\mathrm{x}^k_{\overline{T}_k}\|^2 \overset{(3.14)}{=} -2\ell\|\mathrm{d}^k\|^2 + (\ell+L)\|\mathrm{x}^k_{\overline{T}_k}\|^2 \\ &\le -2\ell\|\mathrm{d}^k\|^2 + 2L\|\mathrm{x}^k_{\overline{T}_k}\|^2 \le -2\delta\|\mathrm{d}^k\|^2 + \|\mathrm{x}^k_{\overline{T}_k}\|^2/(2\tau), \end{aligned}$$

where the last inequality is owing to $\delta \le \ell$ and $\tau < \overline{\tau} \le 1/(4L)$. This proves that d^k from (3.11) is always admitted for sufficiently large k.

3) By Theorem 3.8 3), for sufficiently large k, $\mathrm{supp}(\mathrm{x}^*) \subseteq T_k$ and $\nabla_{T_k} f(\mathrm{x}^*) = \nabla_{T_*} f(\mathrm{x}^*) = 0$ if $\|\mathrm{x}^*\|_0 = s_*$ and $\nabla f(\mathrm{x}^*) = 0$ if $\|\mathrm{x}^*\|_0 < s_*$. These suffice to

$$\mathrm{x}^*_{\overline{T}_k} = 0, \quad \nabla_{T_k} f(\mathrm{x}^*) = 0. \tag{3.40}$$

For any $0 \le t \le 1$, by letting $\mathrm{x}(t) := \mathrm{x}^* + t(\mathrm{x}^k - \mathrm{x}^*)$. the Hessian of f being locally Lipschitz continuous at x^* derives

$$\|\nabla^2_{T_k:} f(\mathrm{x}^k) - \nabla^2_{T_k:} f(\mathrm{x}(t))\|_2 \le M\|\mathrm{x}^k - \mathrm{x}(t)\| = (1-t)M\|\mathrm{x}^k - \mathrm{x}^*\|. \tag{3.41}$$

Moreover, by Taylor expansion, one has

$$\nabla f(\mathrm{x}^k) - \nabla f(\mathrm{x}^*) = \int_0^1 \nabla^2 f(\mathrm{x}(t))(\mathrm{x}^k - \mathrm{x}^*)dt. \tag{3.42}$$

Now, we have the following chain of inequalities

$$\begin{aligned} \|\mathrm{x}^{k+1} - \mathrm{x}^*\|^2 &= \|\mathrm{x}^{k+1}_{T_k} - \mathrm{x}^*_{T_k}\|^2 + \|\mathrm{x}^{k+1}_{\overline{T}_k} - \mathrm{x}^*_{\overline{T}_k}\|^2 \\ &\overset{(3.8,3.40)}{=} \|\mathrm{x}^{k+1}_{T_k} - \mathrm{x}^*_{T_k}\|^2 \overset{(3.8)}{=} \|\mathrm{x}^k_{T_k} - \mathrm{x}^*_{T_k} + \alpha_k \mathrm{d}^k_{T_k}\|^2 \\ &= \|(1-\alpha_k)(\mathrm{x}^k_{T_k} - \mathrm{x}^*_{T_k}) + \alpha_k(\mathrm{x}^k_{T_k} - \mathrm{x}^*_{T_k} + \mathrm{d}^k_{T_k})\|^2 \\ &\le (1-\alpha_k)\|\mathrm{x}^k_{T_k} - \mathrm{x}^*_{T_k}\|^2 + \alpha_k\|\mathrm{x}^k_{T_k} - \mathrm{x}^*_{T_k} + \mathrm{d}^k_{T_k}\|^2 \qquad (3.43) \\ &\overset{(3.27)}{\le} (1-\overline{\alpha}\beta)\|\mathrm{x}^k - \mathrm{x}^*\|^2 + \overline{\alpha}\|\mathrm{x}^k_{T_k} - \mathrm{x}^*_{T_k} + \mathrm{d}^k_{T_k}\|^2, \qquad (3.44) \end{aligned}$$

where (3.43) is due to $\|\cdot\|^2$ is a convex function. From 2), d^k is always updated by (3.11) for sufficiently large k. Therefore, we have

$$\ell\|\mathrm{x}^k_{T_k} - \mathrm{x}^*_{T_k} + \mathrm{d}^k_{T_k}\| \overset{(3.7)}{=} \ell\|H_k^{-1}(\nabla^2_{T_k\overline{T}_k} f(\mathrm{x}^k)\mathrm{x}^k_{\overline{T}_k} - g^k_{T_k}) + \mathrm{x}^k_{T_k} - \mathrm{x}^*_{T_k}\|$$

$$
\begin{aligned}
&\leq \ \|\nabla^2_{T_k:}f(\mathrm{x}^k)\mathrm{x}^k - g^k_{T_k} - H_k\mathrm{x}^*_{T_k}\| \\
&\overset{(3.40)}{=} \|\nabla^2_{T_k:}f(\mathrm{x}^k)\mathrm{x}^k - g^k_{T_k} - \nabla^2_{T_k:}f(\mathrm{x}^k)\mathrm{x}^* + \nabla_{T_k}f(\mathrm{x}^*)\| \\
&\overset{(3.42)}{=} \|\nabla^2_{T_k:}f(\mathrm{x}^k)(\mathrm{x}^k - \mathrm{x}^*) - \int_0^1 \nabla^2_{T_k:}f(\mathrm{x}(t))(\mathrm{x}^k - \mathrm{x})dt\| \\
&= \ \|\int_0^1 [\nabla^2_{T_k:}f(\mathrm{x}^k) - \nabla^2_{T_k:}f(\mathrm{x}(t))](\mathrm{x}^k - \mathrm{x}^*)dt\| \\
&\leq \ \int_0^1 \|\nabla^2_{T_k:}f(\mathrm{x}^k) - \nabla^2_{T_k:}f(\mathrm{x}(t))\|_2\|\mathrm{x}^k - \mathrm{x}^*\|dt \\
&\overset{(3.41)}{\leq} M\|\mathrm{x}^k - \mathrm{x}^*\|^2 \int_0^1 (1-t)dt \leq 0.5M\|\mathrm{x}^k - \mathrm{x}^*\|^2. \qquad (3.45)
\end{aligned}
$$

In addition, it follows from $\mathrm{d}^k_{\overline{T}_k} = -\mathrm{x}^k_{\overline{T}_k}$ and (3.40) that $\|\mathrm{x}^k + \mathrm{d}^k - \mathrm{x}^*\| = \|\mathrm{x}^k_{T_k} + \mathrm{d}^k_{T_k} - \mathrm{x}^*_{T_k}\|$ and thus

$$
\frac{\|\mathrm{x}^k + \mathrm{d}^k - \mathrm{x}^*\|}{\|\mathrm{x}^k - \mathrm{x}^*\|} = \frac{\|\mathrm{x}^k_{T_k} + \mathrm{d}^k_{T_k} - \mathrm{x}^*_{T_k}\|}{\|\mathrm{x}^k - \mathrm{x}^*\|} \overset{(3.45)}{\leq} \frac{(0.5M/\ell)\|\mathrm{x}^k - \mathrm{x}^*\|^2}{\|\mathrm{x}^k - \mathrm{x}^*\|} \to 0. \qquad (3.46)
$$

Now we have three facts: (3.46), $\mathrm{x}^k \to \mathrm{x}^*$ from 1), and $\langle \nabla f(\mathrm{x}^k), \mathrm{d}^k\rangle \leq -\rho\|\mathrm{d}^k\|^2$ from Lemma 3.4, which together with [27, Theorem 3.3] allow us to claim that eventually the step size α_k determined by the Armijo rule is 1, namely $\alpha_k = 1$. Then it follows from (3.43) that

$$
\begin{aligned}
\|\mathrm{x}^{k+1} - \mathrm{x}^*\|^2 &\overset{(3.43)}{\leq} (1-\alpha_k)\|\mathrm{x}^k_{T_k} - \mathrm{x}^*_{T_k}\|^2 + \alpha_k\|\mathrm{x}^k_{T_k} - \mathrm{x}^*_{T_k} + \mathrm{d}^k_{T_k}\|^2 \\
&= \ \|\mathrm{x}^k_{T_k} - \mathrm{x}^*_{T_k} + \mathrm{d}^k_{T_k}\|^2 \overset{(3.45)}{\leq} (0.5M/\ell)^2\|\mathrm{x}^k - \mathrm{x}^*\|^4. \qquad (3.47)
\end{aligned}
$$

Namely, the sequence converges quadratically. Finally, for sufficiently large k,

$$
\begin{aligned}
\|F_\tau(\mathrm{x}^{k+1}; T_{k+1})\|^2 &\overset{(2.12)}{=} \|g^{k+1}_{T_{k+1}}\|^2 + \|\mathrm{x}^{k+1}_{\overline{T}_{k+1}}\|^2 \\
&\overset{(3.40)}{=} \|g^{k+1}_{T_{k+1}} - \nabla_{T_{k+1}}f(\mathrm{x}^*)\|^2 + \|\mathrm{x}^{k+1}_{\overline{T}_{k+1}} - \mathrm{x}^*_{\overline{T}_{k+1}}\|^2 \\
&\overset{(2.6)}{\leq} (L^2+1)\|\mathrm{x}^{k+1} - \mathrm{x}^*\|^2 \overset{(3.47)}{\leq} (L^2+1)(0.5M/\ell)^2\|\mathrm{x}^k - \mathrm{x}^*\|^4. \qquad (3.48)
\end{aligned}
$$

Since f is strongly convex in a neighbour of x^*, then $f(\mathrm{x})$ is also strongly convex in a neighborhood of x^k due to x^k being in neighborhood of x^* for sufficiently large k. Because of this, we have $\nabla^2_{T_k}f(\mathrm{x}^*) \succeq \ell I$, which yields

$$
\sigma_{\min}(\nabla F_\tau(\mathrm{x}^*; T_k)) = \sigma_{\min}\left(\begin{bmatrix} \nabla^2_{T_k}f(\mathrm{x}^*) & \nabla^2_{T_k,T_k^c}f(\mathrm{x}^*) \\ 0 & I \end{bmatrix}\right) \succeq \min\{\ell, 1\}.
$$

where $\sigma_{\min}(A)$ in the minimum singular value of A. Now we have the following Taylor expansion for a fixed T_k,

$$
\begin{aligned}
\|F_\tau(\mathrm{x}^k; T_k)\| &\geq \|F_\tau(\mathrm{x}^*; T_k) + \nabla F_\tau(\mathrm{x}^*; T_k)(\mathrm{x}^k - \mathrm{x}^*)\| - o(\|\mathrm{x}^k - \mathrm{x}^*\|) \\
&= \|\nabla F_\tau(\mathrm{x}^*; T_k)(\mathrm{x}^k - \mathrm{x}^*)\| - o(\|\mathrm{x}^k - \mathrm{x}^*\|), \\
&\geq (1/\sqrt{2})\|\nabla F_\tau(\mathrm{x}^*; T_k)(\mathrm{x}^k - \mathrm{x}^*)\| \geq (\min\{\ell, 1\}/\sqrt{2})\|\mathrm{x}^k - \mathrm{x}^*\|,
\end{aligned}
$$

where the first equation holds due to $F_\eta(\mathrm{x}^*;T_k)=0$ by (3.40). Finally,

$$\|F_\tau(\mathrm{x}^k;T_k)\|^2 \geq \frac{\min\{\ell^2,1\}}{2}\|\mathrm{x}^k-\mathrm{x}^*\|^2 \overset{(3.48)}{\geq} \frac{\min\{\ell^3,\ell\}}{M\sqrt{L^2+1}}\|F_\tau(\mathrm{x}^{k+1};T_{k+1})\|,$$

which completes the whole proof. □

4. Numerical Experiments. In this part, we will conduct extensive numerical experiments of our algorithm SNL0 by using MATLAB (R2019a) on a laptop of 32GB memory and Inter(R) Core(TM) i9-9880H 2.3Ghz CPU.

4.1. Implementation of SNL0. We initialize SNL0 with $\mathrm{x}^0=0$. Parameters are set as $\sigma=5\times10^{-5}$ and $\beta=0.5$, $c=1.05$, $\epsilon=10^{-5}$ and $K=50$. As for s_{-1}, an upper bound can be taken as $\mathcal{O}(n/\ln(n))$ suggested by [29] in Compressed sensing problems. Following this idea, we set $s_{-1}=\lceil rn/\ln(n)\rceil$ with $r=\max\{0.05,|\log_n(\|\nabla f(0)\|)|\}$ for all experiments if there is no extra explanations.

4.1.1. Tuning δ and τ. Note that conditions in Theorem 3.8 and Theorem 3.9 are sufficient but not necessary. Therefore, there is no need to set parameters strictly meeting them in practice. More precisely, Theorem 3.9 states any positive $\delta\in(0,\min\{1,\ell\})$ is acceptable, but in practice to guarantee more steps with Newton directions, it is suggested to be relatively small [21, 28]. While Theorem 3.8 requires $0<\tau<\overline{\tau}\leq 2\overline{\alpha}\delta\beta/L^2$ from (3.16), which means τ should be small enough if δ is chosen to be small. However, T_k would not vary too much in **Step 1** if a sufficiently small τ is set at the beginning. This potentially causes SNL0 to fall in a local area all the time, which clearly degrades the performance of the algorithm. Therefore, we set

$$\delta:=\delta_k=\begin{cases}10^{-10}, & \text{if} \quad T_k=T_{k-1},\\ 10^{-4}, & \text{if} \quad T_k\neq T_{k-1}.\end{cases}$$

This can be explained by Remark 3.6. If $T_k=T_{k-1}$, there is no restriction on τ by δ, which means τ is not necessary to be sufficiently small. Otherwise, $\delta_k=10^{-4}$ would not result in a small τ, which benefits for finding the support set T_k.

In spite of that Theorem 3.8 has given us a clue to set $0<\tau<\overline{\tau}$, it is still difficult to fix a proper one since L is not easy to compute in general. Overall, we choose to update τ adaptively. Typically, we use the following rule: starting τ with a fixed scalar τ_0 (e.g., 5 if no extra explanations are given) and then update it as,

$$\tau_{k+1}=\begin{cases}\tau_k/1.05, & \text{if } k/10=\lceil k/10\rceil \text{ and } \|F_{\tau_k}(\mathrm{x}^k;T_k)\|>k^{-2},\\ \tau_k 1.25, & \text{if } k/10=\lceil k/10\rceil \text{ and } \|F_{\tau_k}(\mathrm{x}^k;T_k)\|\leq k^{-2},\\ \tau_k, & \text{otherwise.}\end{cases}$$

4.1.2. Halting conditions. We terminate SNL0 at kth step if it meets one of following conditions: 1) k reaches the maximum number (e.g., 2000) of iterations or 2) $\mathrm{Tol}_{\tau_k}(\mathrm{x}^k;\ T_k)\leq10^{-6}$, where

$$\mathrm{Tol}_{\tau_k}(\mathrm{x}^k;\ T_k):=\|F_{\tau_k}(\mathrm{x}^k;T_k)\|+\max_{i\in\overline{T}_k}\{|\nabla_i f(\mathrm{x}^k)|-\|\mathrm{x}^k\|_{[s]}/\tau_k,\ 0\}. \tag{4.1}$$

In fact, if a point x^k satisfies that $\mathrm{Tol}_{\tau_k}(\mathrm{x}^k;\ T_k)=0$, then both terms on the right-hand side of (4.1) are zeros, which imply that $\nabla_{T_k}f_r(\mathrm{x}^k)=0, \mathrm{x}^k_{T_k^c}=0$ and $|\nabla_i f(\mathrm{x}^k)|\leq\|\mathrm{x}^k\|_{[s]}/\tau_k$. Hence $\mathrm{supp}(x^k)\subseteq T_k$. If $\|x^k\|_0=s_k$, then $\nabla_{T_k}f_r(\mathrm{x}^k)=0$, $|x^k_i|\geq\sqrt{2\lambda_k\tau_k}, i\in T_k$ and $|\nabla_i f(\mathrm{x}^k)|\leq\sqrt{2\lambda_k/\tau_k}, i\in\overline{T}_k$ with $\lambda_k=\|\mathrm{x}^k\|^2_{[s]}/(2\tau_k)$. If $\|x^k\|_0<s_k$, $\nabla f_r(\mathrm{x}^k)=0$. Namely, x^k is kind of a stationary point.

4.2. Compressed sensing. CS has seen revolutionary advances both in theory and algorithm over the past decade. Ground-breaking papers that pioneered the advances are [24, 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 $\mathrm{y} = A\mathrm{x}$, where the sensing matrix A chosen as in [53, 60]. While for the image data, we consider the inexact recovery $\mathrm{y} = A\mathrm{x} + \xi$, where ξ is the noise and A will be described in Example 4.2.

EXAMPLE 4.1 (Random data). *Let $A \in \mathbb{R}^{m\times n}$ be a random Gaussian matrix with each column $A_j, j \in \mathbb{N}_n$ being identically and independently generated from the standard normal distribution. We then normalize each column such that $\|A_j\| = 1$. Finally, the 'ground truth' signal x^* and the measurement y are produced by the following pseudo Matlab codes:*

$$\begin{aligned} \mathrm{x}^* &= \mathtt{zeros}(n,1), \quad \Gamma = \mathtt{randperm}(n), \\ \mathrm{x}^*(\Gamma(1:s_*)) &= \mathtt{randn}(s_*,1), \quad \mathrm{y} = A\mathrm{x}^*. \end{aligned} \tag{4.2}$$

where s_* is the sparsity level of the signal x^*. Let x be the solution produced by a method. We say the recovery of this method is successful if

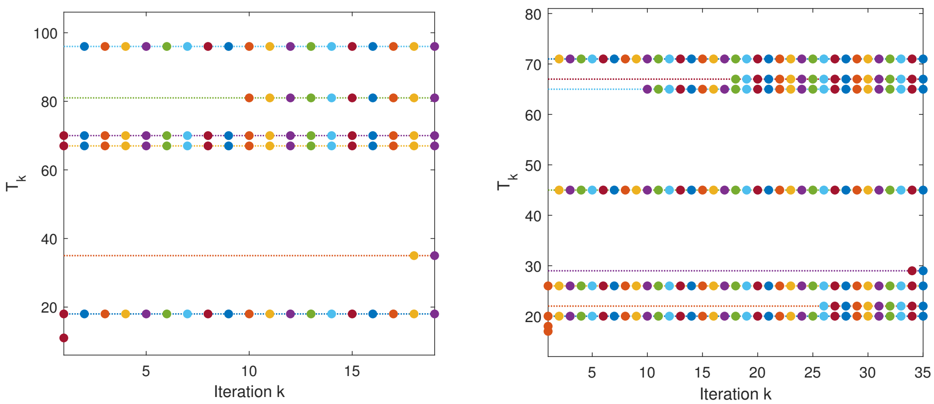
$$\|\mathrm{x} - \mathrm{x}^*\| < 0.01\|\mathrm{x}^*\|.$$

EXAMPLE 4.2 (2-D image data). *Some images are naturally not sparse themselves but could be sparse under some wavelet transforms. Here, we take advantage of the Daubechies wavelet 1, denoted as $W(\cdot)$. Then images under this transform (i.e., $x^* := W(\omega)$) is sparse, ω be the vectorized intensity of an input image. Because of this, 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 (FFT) and W^{-1} is the inverse of W. Finally, the added noise ξ has each element $\xi_i \sim \mathtt{nf}\cdot\mathcal{N}$ with $\mathcal{N}$ being the standard normal distribution and* `nf` *being the noise factor. Three typical choices of* `nf` *are taken into account, namely $\mathtt{nf} \in \{0.01, 0.05, 0.1\}$. For this experiment, we compute two images.*

- `Img1`*: A gray image (see the original image in Figure 4) with size 512×512 (i.e. $n = 512^2 = 262144$). The sampling size is $m = 20033$ and 29729.*
- `Img2`*: A color image (see the original image in Figure 5) with size $256 \times 256 \times 3$ (i.e. $n = 256^2 = 65536$). The sampling size is $m = 3767$ and 6213.*

4.2.1. Effectiveness of PSS. First of all, we would like to see the effectiveness of `PSS` in Algorithm 3.1. We compute two instances from Example 4.1 with $n = 100, m = 25$. One has the solution being 6-sparse and another is 8-sparse. As shown in Figure 1, where points on lines are correct indices, SNL0 is able to identify the true support within a few iterations. Actually, as long as the true support set is found, it stops very quickly, usually with a couple of steps (2 steps for both instances). Note that, for example in Fig 1a, at step 1, index 11 is a false one since 11 does not belong to the true support set $\{18, 35, 67, 70, 81, 96\}$. This means SNL0 is able to correct T_k.

4.2.2. 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 of our scope to compare all of them. To make comparisons fair, we only focus on those algorithms (often referred as regularized methods) which aim at solving (1.1) or its relaxations, where ℓ_0 norm is replaced by some approximations such as $\ell_q(0 < q \leq 1)$ [35] or $\ell_1 - \ell_2$



(a) True sparsity level 6

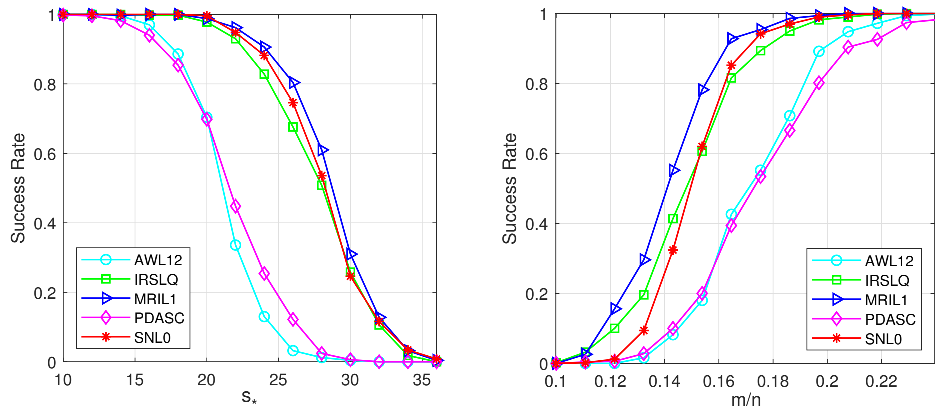
(b) True sparsity level 8

Fig. 1: T_k at kth iteration for Example 4.1 with $n = 100, m = 25$.

[37]. Note that greedy methods mentioned in Subsection 1.1, for the model (1.2) with s being given, have been famous for the super fast computational speed and the high order of accuracy when s is relatively small to n. However, we will not compare them with SNL0 since we would like to consider the scenario when s is unknown. We select MIRL1 [60], AWL1 [37, ADMM for weighted ℓ_{1-2}] which is a faster approximation of the method proposed in [53], IRSLQ [35] (we choose $q = 1/2$) and PDASC [34]. All parameters are set as default except for setting the maximum iteration number as 100 and removing the final refinement step for MRIL1 and `del=1e-8` for PDASC. Note that PDASC and SNL0 are the second order methods and the other three belong to the category of the first order methods.

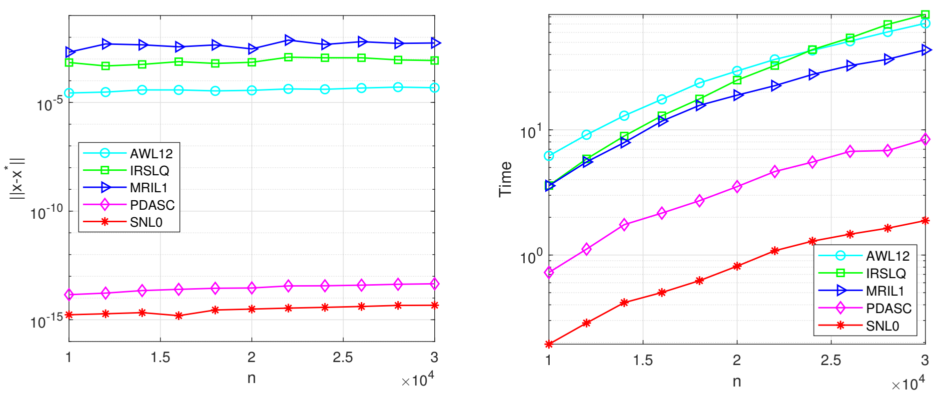
We begin with running 500 independent trials with fixind $n = 256, m = \lceil n/4 \rceil$ and then report the corresponding success rates (which is defined by the percentage of the number of successful recoveries over all trails) at sparsity levels s_* from 10 to 36. From Figure 2a, one can observe that MIRL1, SNL0 and IRSLQ basically generate similar results for each sparsity level, but all outperform AWL12 and PDASC. Moreover, all lines decline along with the rising of s_*, which indicates that the CS problems become more difficult to get recovered successfully. Our next experiment is to see how sample size m affects the performance of those methods. We run 500 independent trials with fixing $n = 256, s = 13$ but varying $m = \lceil rn \rceil$ where $r \in \{0.1, 0.12, \cdots, 0.3\}$. Obviously, the larger m is, the easier the problem is to be solved, which is illustrated by Figure 2b. Again MIRL1, SNL0 and IRSLQ behave better than the other two.

To see the accuracy of the solutions and the speed of these five methods, we now run 50 trials with higher dimensions n increasing from 10000 to 30000 and keeping $m = \lceil 0.25n \rceil, s_* = \lceil 0.01n \rceil$, and record average results in Figure 3. As shown in Figure 3a, SNL0 always generates the smallest $\|\mathrm{x} - \mathrm{x}^*\|$, the most accurate recovery, with accuracy order at least 10^{-14}, followed by PDSAC. By contrast, the other three methods get accuracy with the order being above 10^{-5}. This phenomenon well testifies that the second order methods have their advantages in producing a higher order of accuracy. When it comes to the computational speed, it can be clearly seen from Figure 3b that SNL0 always runs the fastest, with only consuming about 2 seconds when $n = 30000$. PDSAC is the runner up. This shows that, for problems in higher dimensions, SNL0 and PDSAC are able to run faster than the first order methods.



(a) Success rates v.s. s_* with $m = 64$ (b) Success rates v.s. m/n with $s_* = 13$

Fig. 2: Success rates of five methods for solving Example 4.1 with $n = 256$.



(a) $\|x - x^*\|$ v.s. n (b) CPU time v.s. n

Fig. 3: Average recovery error and time of five methods for solving Example 4.1.

4.2.3. Comparisons for 2-D image data. In Example 4.2, data size n is relatively large, which possibly makes most regularized methods suffer extremely slow computation. Hence, we select three greedy methods CSMP (denoted for CoSaMP) [41], HTP [31] and AIHT [11] as well as PDSCA. As suggested in package PDSCA, we set another rule to stop each method if at kth iteration it satisfies $\|Ax^k - y\| \leq \|Ax^* - y\|$ to speed up the termination. Moreover, to make comparisons fair, we fist run PDSCA, which has a strong ability to obtain a solution with good sparsity level s_P. Then we set s for CSMP, HTP and AIHT since they need such prior information, but only set $s_{-1} = \lceil 0.75 s_P \rceil$ for our method to accelerate it. 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

$$\text{PSNR} := 10 \log_{10}(n \|x - x^*\|^{-2})$$

to measure the performance of the method. Note that the larger PSNR is, the much closer x approaches to the true image x*, namely the better performance of a method yields. Results for `Img1` are presented in Table 1 and Figure 4, where SPDSA offers the biggest PSNR when `nf`$= 0.01$, whilst SNL0 produces the biggest ones when `nf`$= 0.05$

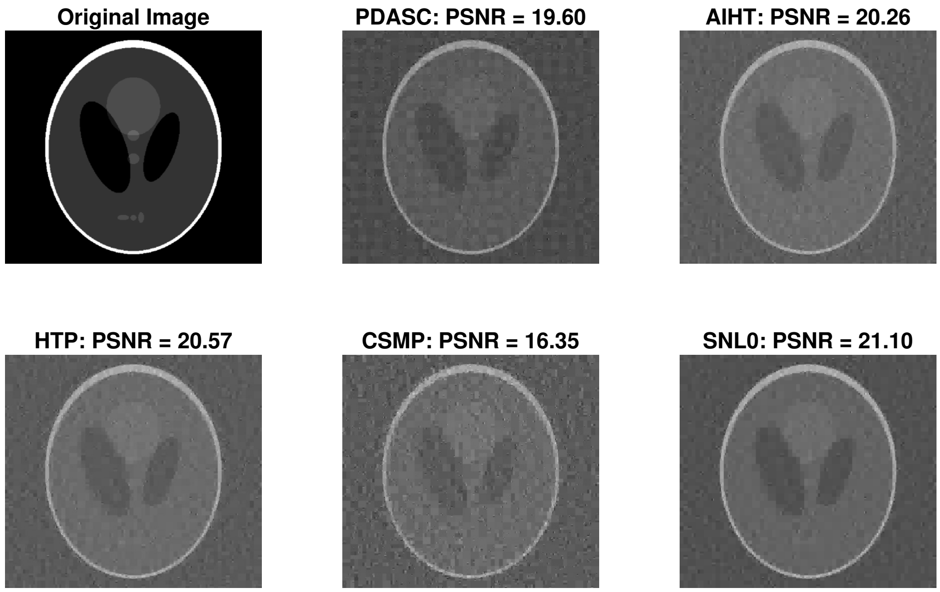


Fig. 4: Recovery results for `Img1` in Example 4.2 with $m = 20033$ and `nf`$= 0.1$.

Table 1: Performance of five methods for `Img1` in Example 4.2.

	`nf`$= 0.01$			`nf`$= 0.05$			`nf`$= 0.1$		
Algs.	PSNR	Time	$\|\|x\|\|_0$	PSNR	Time	$\|\|x\|\|_0$	PSNR	Time	$\|\|x\|\|_0$
				$m = 20033, n = 262144, m/n = 0.076$					
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	19.25	2969
HTP	19.66	19.15	9716	20.27	3.40	5982	20.57	3.41	2969
SCMP	12.49	51.54	9716	8.44	63.09	5982	16.35	14.83	2969
SNL0	20.74	10.17	9639	21.01	2.62	4487	21.10	2.58	2227
				$m = 29729, n = 262144, m/n = 0.113$					
SPDSA	35.37	11.54	9902	25.07	6.58	5002	22.61	5.44	3513
AIHT	32.21	71.42	9902	24.78	9.52	5002	23.07	9.16	3513
HTP	34.89	14.38	9902	25.14	4.57	5002	23.19	2.02	3513
SCMP	21.48	39.79	9902	23.00	9.94	5002	20.73	2.26	3513
SNL0	33.80	9.68	9824	25.35	3.06	3752	23.44	1.30	2635

and `nf`$= 0.1$, which means our method is more robust to the noise. In addition, SNL0 runs the fastest and renders the sparsest representations for all cases. Finally, for color image `Img2`, we repeat the recovery for three (rgb) channels and report the total results of three channels. For example, we report the PSNR$= \sum_{i=1}^{3}$PSNR$_i$, where PSNR$_i$ is the peak signal to noise ratio obtained by one method for solving channel i. As demonstrated in Figure 5, the quality of the image recovered by SNL0 clearly is better than others. More detailed results are reported in Table 2, where similar observations to Table 1 can be seen.

4.3. Sparse Linear Complementarity Problem. Sparse linear complementarity problems (SLCP) have been applied to deal with real-world applications such

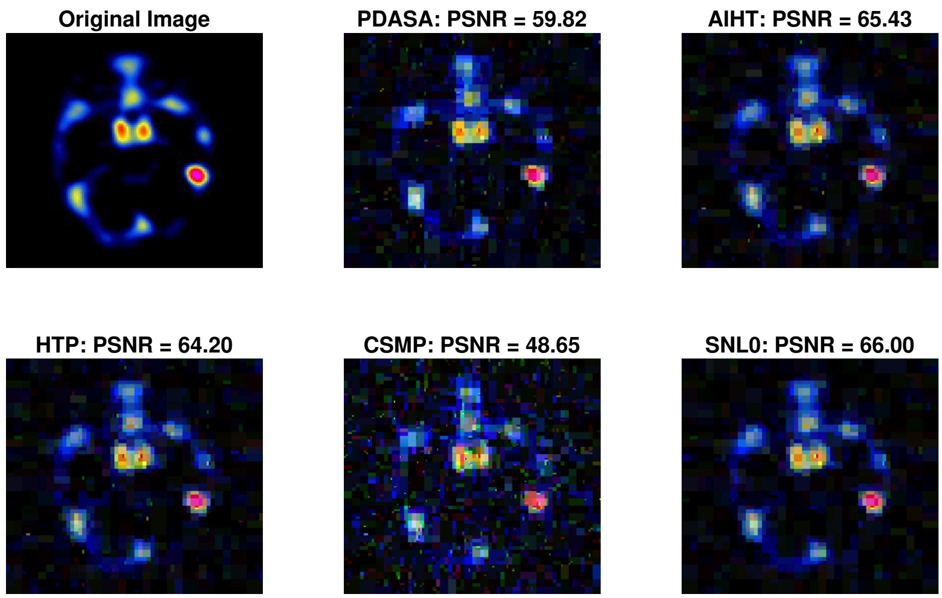


Fig. 5: Recovery results for `Img2` in Example 4.2 with $m = 3767$ and `nf`$= 0.1$.

Table 2: Performance of five methods for `Img2` in Example 4.2.

Algs.	`nf`$= 0.01$			`nf`$= 0.05$			`nf`$= 0.1$		
	PSNR	Time	$\|\|\mathrm{x}\|\|_0$	PSNR	Time	$\|\|\mathrm{x}\|\|_0$	PSNR	Time	$\|\|\mathrm{x}\|\|_0$
				$m = 3767, n = 65536, m/n = 0.057$					
SPDSA	64.26	9.78	4885	62.44	5.52	2474	59.82	4.08	1585
AIHT	66.35	21.94	4885	67.08	5.69	2474	65.43	1.57	1585
HTP	64.15	12.65	4885	66.32	2.75	2474	64.20	1.40	1585
SCMP	52.74	42.47	4885	43.86	30.58	2474	48.65	13.9	1585
SNL0	67.42	5.97	4458	67.56	2.41	1857	66.00	1.40	1190
				$m = 6213, n = 65536, m/n = 0.095$					
SPDSA	79.10	8.68	6851	73.49	4.74	3487	69.85	3.54	2203
AIHT	86.25	19.47	6851	81.30	1.12	3487	77.18	2.44	2203
HTP	86.45	6.14	6851	81.06	1.30	3487	76.88	1.24	2203
SCMP	62.62	33.89	6851	65.01	10.80	3487	65.59	4.85	2203
SNL0	87.41	5.92	6563	81.44	1.63	2616	76.16	0.84	1653

as bimatrix games and portfolio selection problems [19, 52, 46]. The problem aims at finding a sparse vector $\mathrm{x} \in \mathbb{R}^n$ from

$$\{\mathrm{x} \in \mathbb{R}^n : \ \mathrm{x} \geq 0, \ Q\mathrm{x} + \mathrm{q} \geq 0, \ \langle \mathrm{x}, Q\mathrm{x} + \mathrm{q} \rangle = 0\}, \tag{4.3}$$

where $Q \in \mathbb{R}^{n\times n}$ and $q \in \mathbb{R}^n$. A point x from (4.3) is equivalent to

$$f(x) := \sum_{i=1}^{n} \phi(x_i, Q_i\mathrm{x} + q_i) = 0, \tag{4.4}$$

where ϕ is the so-called NCP function, which is defined by $\phi(a, b) = 0 \Leftrightarrow a \geq 0, b \geq 0, ab = 0$. We benefit from two NCP functions: the Fischer-Burmeister (FB) function

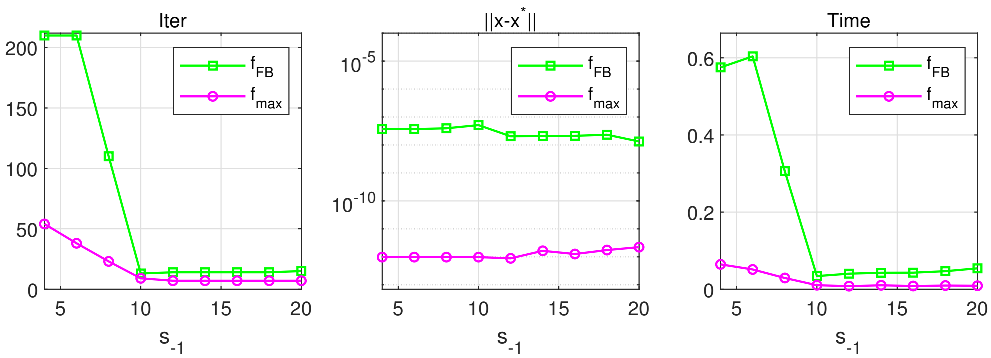


Fig. 6: Performance of SNL0 effected by s_{-1} for solving Example 4.3.

$\phi_{FB}(a,b) = \sqrt{a^2+b^2} - a - b$ [30] and the one $\phi_{\max}(a,b) = a_+^2 b_+^2 + (-a)_+^2 + (-b)_+^2$ introduced in [58], where $a_+ := \max\{a, 0\}$, with corresponding merit functions f:

$$f_{FB}(\mathrm{x}) = 0.5\left[\|\mathrm{x}\|^2 + \|\mathrm{y}\|^2 + \|\mathrm{x}+\mathrm{y}\|^2 - 2\langle\sqrt{\mathrm{x}\circ\mathrm{x}+\mathrm{y}\circ\mathrm{y}}, \mathrm{x}+\mathrm{y}\rangle\right],$$
$$f_{\max}(\mathrm{x}) = 0.5\left[\|(\mathrm{x}_+)\circ(\mathrm{y}_+)\|^2 + \|(-\mathrm{x})_+\|^2 + \|(-\mathrm{y})_+\|^2\right],$$

where $\mathrm{y} := Q\mathrm{x} + \mathrm{q}$, $\sqrt{\mathrm{z}} := (\sqrt{z_1}, \cdots, \sqrt{z_n})^\top$, $\mathrm{z}_+ := ((z_1)_+, \cdots, (z_n)_+)^\top$ and $\mathrm{x}\circ\mathrm{z} := (x_1 z_1, \cdots, x_n z_n)^\top$. Note that f_{FB} has unbounded Hessian at $(0,0)$. Therefore, to make use of SNL0, we add a small scalar ε (e.g. 10^{-10}) to smooth $\sqrt{z}$, namely, replacing $\sqrt{z}$ by $\sqrt{z+\varepsilon}$ in f_{FB}. The following testing example is from [58].

EXAMPLE 4.3. *Let $Q = ZZ^\top$ with $Z \in \mathbb{R}^{n\times m}$ whose elements are generated from the standard normal distribution, where $m \le n$ (e.g. $m = n/2$). Then, the 'ground truth' sparse solution x^* is produced same as (4.2) and q is obtained by $q_i = -(Q\mathrm{x}^*)_i$ if $x_i^* > 0$ and $q_i = |(Q\mathrm{x}^*)_i|$ otherwise.*

We first run an experiment with $n = 500$ and $s_* = 10$ to see how initial guess s_{-1} would affect the performance of SNL0. Results are presented in Figure 6. It seems that if s_{-1} is chosen to be higher than s_*, then SNL0 gets better results for both models. More precisely, the number of iterations declines when s_{-1} increases from 4 to 10 and then stabilizes when $s_{-1} \ge 10 = s_*$. A similar phenomenon can be seen for CPU time. Note that $\|\mathrm{x}-\mathrm{x}^*\|$ is in order of 10^{-7} and 10^{-11} for the model (1.1) with $f = f_{FB}$ and $f = f_{\max}$ respectively, which means SNL0 achieves global solutions (i.e. x^*) for both models. In addition, this experiment illustrates that (1.1) with $f = f_{\max}$ allows SNL0 to perform better results than (1.1) with $f = f_{FB}$, since it solves the former model faster and more accurate.

Next we run 50 trials with higher dimensions n rising from 10000 to 30000 and keeping $s_* = \lceil 0.01n\rceil$. To make fair comparisons of SNL0 for solving two models: (1.1) with $f = f_{\max}$ and (1.1) with $f = f_{FB}$, we set $s_{-1} = \lceil 0.1n/\ln(n)\rceil$. Average results are displayed in Figure 7, where a clear conclusion which can be made is that SNL0 addresses SLCP more effectively through the model with $f = f_{\max}$, with fewer iterations, much shorter CPU time and more accurate recovery.

5. Conclusion. Many methods only make use of the first order information of the involved functions. Because of this, they are able to run fast but suffer from slow convergence. When Newton steps only performed on chosen subspaces are integrated into some of these methods, then much more rapid convergence can be achieved. To the best of our knowledge, theoretic guarantees include two groups: either the

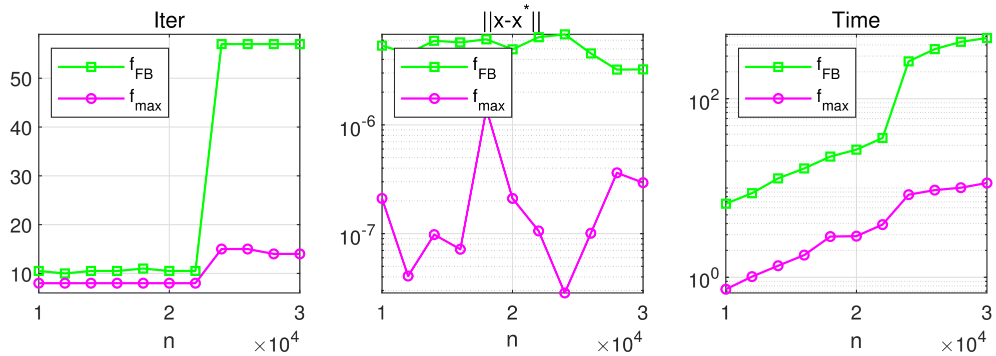


Fig. 7: Performance of SNL0 effected by n for solving Example 4.3.

(sub)sequence converges to a stationary point of ℓ_0-regularized optimization or the distance between each iterate and any given sparse reference point is bounded by an error bound in the sense of probability. However, those are still not enough to unravel the reasons why those methods with Newton steps perform very well. In this paper, we designed a subspace Newton method equipped with an effective mechanism adaptively updating the penalty parameter for the ℓ_0-regularized optimization. Theoretically, we proved that the sequence generated by our proposed method converges to a stationary point globally and quadratically, well explaining our method possessing an extraordinary performance. Numerically, the method is capable of running very fast and rendering extremely high order of accuracy, being competitive to against some other excellent solvers.

REFERENCES

[1] H. Attouch, J. Bolte, and B. F. Svaiter, *Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized gauss–seidel methods*, Mathematical Programming, 137 (2013), pp. 91–129.

[2] S. Bahmani, B. Raj, and P. T. Boufounos, *Greedy sparsity constrained optimization*, Journal of Machine Learning Research, 14 (2013), pp. 807–841.

[3] C. Bao, B. Dong, L. Hou, Z. Shen, X. Zhang, and X. Zhang, *Image restoration by minimizing zero norm of wavelet frame coefficients*, Inverse Problems, 32 (2016), pp. 115004.

[4] A. Beck and Y. C. Eldar, *Sparsity constrained nonlinear optimization: Optimality conditions and algorithms*, SIAM Journal on Optimization, 23 (2013), pp. 1480–1509.

[5] A. Beck and N. Hallak, *Proximal mapping for symmetric penalty and sparsity*, SIAM Journal on Optimization, 28 (2018), pp. 496–527.

[6] D. Bertsimas, A. King, and R. Mazumder, *Best subset selection via a modern optimization lens*, The Annals of Statistics, (2016), pp. 813–852.

[7] W. Bian and X. Chen, *Smoothing neural network for constrained non-lipschitz optimization with applications*, IEEE Transactions on Neural Networks and Learning Systems, 23 (2012), pp. 399–411.

[8] W. Bian and X. Chen, *Linearly constrained non-Lipschitz optimization for image restoration*, SIAM Journal on Imaging Sciences, 8 (2015), pp. 2294–2322.

[9] W. Bian and X. Chen, *A smoothing proximal gradient algorithm for nonsmooth convex regression with cardinality penalty*, SIAM Journal on Numerical Analysis, 58 (2020), pp. 858–883.

[10] J. D. Blanchard, J. Tanner, and K. Wei, *CGIHT: conjugate gradient iterative hard thresholding for compressed sensing and matrix completion*, Information and Inference: A Journal of the IMA, 4 (2015), pp. 289–327.

[11] T. Blumensath, *Accelerated iterative hard thresholding*, Signal Processing, 92 (2012), pp. 752–756.

[12] T. Blumensath and M. E. Davies, *Gradient pursuits*, IEEE Transactions on Signal Processing, 56 (2008), pp. 2370–2382.

[13] T. Blumensath and M. E. Davies, *Iterative thresholding for sparse approximations*, Journal of Fourier analysis and Applications, 14 (2008), pp. 629–654.
[14] T. Blumensath and M. E. Davies, *Normalized iterative hard thresholding: Guaranteed stability and performance*, IEEE Journal of Selected Topics in Signal Processing, 4 (2010), pp. 298–309.
[15] E. J. Candès, J. Romberg, and T. Tao, *Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information*, IEEE Transactions on Information Theory, 52 (2006), pp. 489–509.
[16] E. J. Candès and T. Tao, *Decoding by linear programming*, IEEE Transactions on Information Theory, 51 (2005), pp. 4203–4215.
[17] X. Chen, M. K. Ng, and C. Zhang, *Non-Lipschitz ℓ_p-regularization and box constrained model for image restoration*, IEEE transactions on image processing, 21 (2012), pp. 4709–4721.
[18] W. Cheng, Z. Chen, and Q. Hu, *An active set Barzilar-Borwein algorithm for ℓ_0 regularized optimization*, Journal of Global Optimization, (2019).
[19] R. W. Cottle, *Linear complementarity problem*, Springer, 2009.
[20] W. Dai and O. Milenkovic, *Subspace pursuit for compressive sensing signal reconstruction*, IEEE transactions on Information Theory, 55 (2009), pp. 2230–2249.
[21] T. De Luca, F. Facchinei, and C. Kanzow, *A semismooth equation approach to the solution of nonlinear complementarity problems*, Mathematical Programming, 75 (1996), pp. 407–439.
[22] T. Dinh, B. Wang, A. L. Bertozzi, S. J. Osher, and J. Xin, *Sparsity meets robustness: channel pruning for the feynman-kac formalism principled robust deep neural nets*, arXiv preprint arXiv:2003.00631, (2020).
[23] Q. Dong, X. Liu, Z.-W. Wen, and Y.-X. Yuan, *A parallel line search subspace correction method for composite convex optimization*, Journal of the Operations Research Society of China, 3 (2015), pp. 163–187.
[24] D. L. Donoho, *Compressed sensing*, IEEE Transactions on Information Theory, 52 (2006), pp. 1289–1306.
[25] M. Elad, *Sparse and redundant representations: from theory to applications in signal and image processing*, Springer Science & Business Media, 2010.
[26] M. Elad, M. A. Figueiredo, and Y. Ma, *On the role of sparse and redundant representations in image processing*, Proceedings of the IEEE, 98 (2010), pp. 972–982.
[27] F. Facchinei, *Minimization of sc1 functions and the maratos effect*, Operations Research Letters, 17 (1995), pp. 131–138.
[28] F. Facchinei and C. Kanzow, *A nonsmooth inexact newton method for the solution of large-scale nonlinear complementarity problems*, Mathematical Programming, 76 (1997), pp. 493–512.
[29] J. Fan and J. Lv, *Sure independence screening for ultrahigh dimensional feature space*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 70 (2008), pp. 849–911.
[30] A. Fischer, *A special newton-type optimization method*, Optimization, 24 (1992), pp. 269–284.
[31] S. Foucart, *Hard thresholding pursuit: an algorithm for compressive sensing*, SIAM Journal on Numerical Analysis, 49 (2011), pp. 2543–2563.
[32] J. Huang, Y. Jiao, Y. Liu, and X. Lu, *A constructive approach to l0 penalized regression*, The Journal of Machine Learning Research, 19 (2018), pp. 403–439.
[33] K. Ito and K. Kunisch, *A variational approach to sparsity optimization based on lagrange multiplier theory*, Inverse Problems, 30 (2013), p. 015001.
[34] Y. Jiao, B. Jin, and X. Lu, *A primal dual active set with continuation algorithm for the ℓ_0-regularized optimization problem*, Applied and Computational Harmonic Analysis, 39 (2015), pp. 400–426.
[35] M.-J. Lai, Y. Xu, and W. Yin, *Improved iteratively reweighted least squares for unconstrained smoothed ℓ_q minimization*, SIAM Journal on Numerical Analysis, 51 (2013), pp. 927–957.
[36] S. Lin, R. Ji, Y. Li, C. Deng, and X. Li, *Toward compact convnets via structure-sparsity regularized filter pruning*, IEEE transactions on neural networks and learning systems, (2019).
[37] Y. Lou and M. Yan, *Fast $l1-l2$ minimization via a proximal operator*, Journal of Scientific Computing, 74 (2018), pp. 767–785.
[38] Z. Lu, *Iterative hard thresholding methods for ℓ_0 regularized convex cone programming*, Mathematical Programming, 147 (2014), pp. 125–154.
[39] Z. Lu and Y. Zhang, *Sparse approximation via penalty decomposition methods*, SIAM Journal on Optimization, 23 (2013), pp. 2448–2478.
[40] J. J. Moré and D. C. Sorensen, *Computing a trust region step*, SIAM Journal on Scientific and Statistical Computing, 4 (1983), pp. 553–572.

[41] D. Needell and J. A. Tropp, *Cosamp: Iterative signal recovery from incomplete and inaccurate samples*, Applied and cComputational Harmonic Analysis, 26 (2009), pp. 301–321.
[42] Q. Ni and Y. Yuan, *A subspace limited memory quasi-newton algorithm for large-scale nonlinear bound constrained optimization*, Mathematics of Computation, 66 (1997), pp. 1509–1520.
[43] L. Pan, S. Zhou, N. Xiu, and H.-D. Qi, *A convergent iterative hard thresholding for nonnegative sparsity optimization*, Pacific Journal of Optimization, 13 (2017), pp. 325–353.
[44] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, *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, IEEE, 1993, pp. 40–44.
[45] A. Patrascu, I. Necoara, and P. Patrinos, *A proximal alternating minimization method for ℓ_0-regularized nonlinear optimization problems: Application to state estimation*, Proceedings of the IEEE Conference on Decision and Control, 2015 (2015), pp. 4254–4259.
[46] M. Shang, C. Zhang, and N. Xiu, *Minimal zero norm solutions of linear complementarity problems*, Journal of Optimization Theory and Applications, 163 (2014), pp. 795–814.
[47] E. Soubies, L. Blanc-Féraud, and G. Aubert, *A continuous exact ℓ_0 penalty (cel0) for least squares regularized problem*, SIAM Journal on Imaging Sciences, 8 (2015), pp. 1607–1639.
[48] C. Soussen, J. Idier, J. Duan, and D. Brie, *Homotopy based algorithms for ℓ_0-regularized least squares*, IEEE Transactions on Signal Processing, 63 (2015), pp. 3301–3316.
[49] J. A. Tropp, *Just relax: Convex programming methods for identifying sparse signals in noise*, IEEE Transactions on Information Theory, 52 (2006), pp. 1030–1051.
[50] J. A. Tropp and A. C. Gilbert, *Signal recovery from random measurements via orthogonal matching pursuit*, IEEE Transactions on Information Theory, 53 (2007), pp. 4655–4666.
[51] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. S. Huang, and S. Yan, *Sparse representation for computer vision and pattern recognition*, Proceedings of the IEEE, 98 (2010), pp. 1031–1044.
[52] J. Xie, S. He, and S. Zhang, *Randomized portfolio selection with constraints*, Pacific Journal of Optimization, 4 (2008), pp. 89–112.
[53] P. Yin, Y. Lou, Q. He, and J. Xin, *Minimization of 1-2 for compressed sensing*, SIAM Journal on Scientific Computing, 37 (2015), pp. A536–A563.
[54] X.-T. Yuan, P. Li, and T. Zhang, *Gradient hard thresholding pursuit*, The Journal of Machine Learning Research, 18 (2017), pp. 6027–6069.
[55] X.-T. Yuan and Q. Liu, *Newton greedy pursuit: A quadratic approximation method for sparsity-constrained optimization*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 4122–4129.
[56] X.-T. Yuan, X. Liu, and S. Yan, *Visual classification with multitask joint sparse representation*, IEEE Transactions on Image Processing, 21 (2012), pp. 4349–4360.
[57] Y.-X. Yuan, *A review on subspace methods for nonlinear optimization*, in Proceedings of the International Congress of Mathematics, 2014, pp. 807–827.
[58] S. Zhou, L. Pan, L. Mu, and M. Shang, *Newton hard thresholding pursuit for sparse lcp via a new merit function*, (2019), https://www.researchgate.net/publication/337948990.
[59] S. Zhou, N. Xiu, and H.-D. Qi, *Global and quadratic convergence of Newton hard-thresholding pursuit*, arXiv preprint arXiv:1901.02763, (2019).
[60] S. Zhou, N. Xiu, Y. Wang, L. Kong, and H.-D. Qi, *A null-space-based weighted ℓ_1 minimization approach to compressed sensing*, Information and Inference: A Journal of the IMA, 5 (2016), pp. 76–102.