

# Fast Newton Method for Sparse Logistic Regression

**Rui Wang**

*Department of Applied Mathematics  
Beijing Jiaotong University  
Beijing, P. R. China*

WANGRUIBJTU@BJTU.EDU.COM

**Naihua Xiu**

*Department of Applied Mathematics  
Beijing Jiaotong University  
Beijing, P. R. China*

NHXIU@BJTU.EDU.CN

**Shenglong Zhou**

*School of Mathematics  
University of Southampton  
Southampton, UK*

SZ3G14@SOTON.AC.UK

## Abstract

Sparse logistic regression has been developed tremendously in recent two decades, from its origination the  $\ell_1$ -regularized version by Tibshirani (1996) to the sparsity constrained models by Bahmani, Raj, and Boufounos (2013); Plan and Vershynin (2013). This paper is carried out on the sparsity constrained logistic regression through the classical Newton method. We begin with analysing its first optimality condition to acquire a strong  $\tau$ -stationary point for some  $\tau > 0$ . This point enables us to equivalently derive a stationary equation system which is able to be efficiently solved by Newton method. The proposed method NSLR an abbreviation for Newton method for sparse logistic regression, enjoys a very low computational complexity, local quadratic convergence rate and termination within finite steps. Numerical experiments on random data and real data demonstrate its superior performance when against with seven state-of-the-art solvers.

**Keywords:** Sparsity constrained logistic regression, Strong  $\tau$ -stationary point, Stationary equation, Newton method, Quadratic convergence rate

## 1. Introduction

Logistic regression is a well-known effective tool of classification with extensive applications ranging from machine learning, data mining, pattern recognition, medical science to statistics. It describes the relation between a sample data  $\mathbf{x}$  and its associated binary response/label  $y \in \{0, 1\}$  through the conditional probability

$$\Pr\{y|\mathbf{x}, \mathbf{z}\} = \frac{e^{y\langle \mathbf{x}, \mathbf{z} \rangle}}{1 + e^{\langle \mathbf{x}, \mathbf{z} \rangle}},$$

where  $\Pr\{y|\mathbf{x}, \mathbf{z}\}$  is the conditional probability of the label  $y$ , given the sample  $\mathbf{x}$  and a parameter vector  $\mathbf{z}$ . To find the maximum likelihood estimate of the parameter  $\mathbf{z}$ , a set of  $n$  independent and identically distributed samples  $(\mathbf{x}_i, y_i), i = 1, 2, \dots, n$  are first drawn, where  $\mathbf{x}_i \in \mathbb{R}^p$  and  $y_i \in \{0, 1\}$ , yielding a joint likelihood of the interested parameter/classifier  $\mathbf{z}$ . Then the maximum likelihood estimate is obtained by addressing the classical logistic

regression model,

$$\min_{\mathbf{z} \in \mathbb{R}^p} \ell(\mathbf{z}) := \frac{1}{n} \sum_{i=1}^n \left\{ \ln(1 + e^{\langle \mathbf{x}_i, \mathbf{z} \rangle}) - y_i \langle \mathbf{x}_i, \mathbf{z} \rangle \right\}. \quad (1)$$

This model usually performs well in scenario when the number  $n$  of samples is larger than the number  $p$  of features because it enjoys a strictly convexity and thus admits a unique optimal solution provided that the matrix  $[\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n]$  has a full row rank. However, when it comes to the case  $n < (\ll) p$ , it usually suffers from the over-fitting. That is, the solved classifier through (1) well fits the model (making the loss sufficiently small) on training data but may behave poorly on unseen data.

Unfortunately, the case  $n < (\ll) p$  occurs often in many real applications. For instance, each gene expression data sample comprises of thousands of genes whilst common medical equipments are only able to obtain very limited samples. In image processing, each image contains large amounts of pixels, which is far more than the number of observed images. Fortunately, despite numerous features in those data, there is only a small portion that is of importance. For example, apart from the classification task, the micro-array data experiments also attempt to identify a small set of informative genes (to distinguish the tumour and the normal tissues) in each gene expression data. The purpose is to remove the irrelevant genes so as to simplify the inference. This naturally gives rise to the so-called sparse logistic regression.

### 1.1 Sparse logistic regression (SLR)

Sparse logistic regression was originated from the  $\ell_1$ -regularized logistic regression proposed by Tibshirani (1996),

$$\min_{\mathbf{z} \in \mathbb{R}^p} \ell(\mathbf{z}) + \nu \|\mathbf{z}\|_1, \quad (2)$$

where  $\|\mathbf{z}\|_1$  is the  $\ell_1$ -norm and  $\nu > 0$ . Under the help of  $\ell_1$ -regularization, this model is capable of rendering a sparse solution, which allows us to capture important features among others. A vector is called sparse if only a few entries are non-zero and the rest are zeros. With the advance in sparse optimization in recent decade, (2) has been extensively extended to the following general model,

$$\min_{\mathbf{z} \in \mathbb{R}^p} f_\phi(\mathbf{z}) := \ell(\mathbf{z}) + \phi_\nu(\mathbf{z}), \quad (3)$$

where the regularized function  $\phi_\nu(\mathbf{z}) : \mathbb{R}^p \rightarrow \mathbb{R}$  is designed to pursue a sparse solution. Methods associated with model (3) are treated as the relaxation/regularization methods from the perspective of optimization. They share the advantage that (3) is unconstrained, making most generic optimization methods for non-differentiable problems tractable.

An alternative is to consider logistic regression with a sparsity constraint, which was first studied by Bahmani, Raj, and Boufounos (2013); Plan and Vershynin (2013) separately and then well investigated by Wang et al. (2017). The model they considered is

$$\min_{\mathbf{z} \in \mathbb{R}^p} \ell(\mathbf{z}), \quad \text{s.t. } \mathbf{z} \in S := \{\mathbf{z} \in \mathbb{R}^p : \|\mathbf{z}\|_0 \leq s\}, \quad (4)$$

where  $S$  is the sparse set with  $s \ll p$ , and  $\|\mathbf{z}\|_0$  is the  $\ell_0$  pseudo norm of  $\mathbf{z}$ , counting the number of nonzero elements of  $\mathbf{z}$ . Apparently, this model suffers from the discreteness of

the constraint set, which causes the NP-hardness of solving it. However, the general case of this model, where  $\ell(\mathbf{z})$  is replaced by a general function, has been extensively explored (see Beck and Hallak, 2015; Pan et al., 2015) since it was first introduced by Bahmani, Raj, and Boufounos (2013) and Beck and Eldar (2013). In statistics, this model (4) where the logistic loss  $\ell$  is replaced by a square norm loss of linear regression, is the so-called best subspace selection (see Hastie et al., 2017; Mazumder et al., 2017; Hazimeh and Mazumder, 2018; Xie and Deng, 2018). Those researches have presented fruitful results, which not only show that sparsity constrained model possesses many advantages over the regularized model, such as parameter-free, ease of sparsity controlling and low computational complexity in terms of algorithmic design, but also provide a series of tractable tools to address this NP-hard problem. Therefore, the work in this paper is carried out along with this model.

## 1.2 Methods of solving SLR

Since there are large numbers of methods that have been proposed to deal with the sparse optimization problems containing the SLR as a special example, which is far beyond our scope of review, we only focus on those directly processing (3) and (4).

For regularization model (3), different penalty functions  $\phi_\nu$  yield different methods, which enable to be summarized as two categories based on the convexity of  $\phi_\nu$ .

**Convex** regularizations are mainly associated with the usage of  $\ell_1$  norm.

- $\phi_\nu(\mathbf{z}) = \nu\|\mathbf{z}\|_1$ . Expectation maximization method (EM, see Figueiredo, 2003; Krishnapuram et al., 2005), at each iteration, managed to find a majorization/upper-bound function  $F_\phi$  of  $f_\phi$  and then generate a Newton step to minimize  $F_\phi$ ; Andrew and Gao (2007) proposed an Orthant-Wise Limited-memory Quasi-Newton method (OWL-QN) where BFGS Quasi-Newton steps were carried out to minimize a quadratic approximation of  $f_\phi$  at each iteration; Koh et al. (2007) took interior-point method into account to form a truncated Newton interior-point method (TNIP); Shi et al. (2010) created a hybrid iterative shrinkage method (HIS) which was comprised of two phases: iterative shrinkage (fixed-point method) and interior-point scheme; Yu et al. (2010) adopted the OWL-QN but with their own direction finding procedure to derive an improved OWL-QN\*; Yuan et al. (2010) exploited the coordinate descent method with each coordinate being updated by using one dimensional Newton direction (CDN).
- $\phi_\nu(\mathbf{z}) = \nu_1\|\mathbf{z}\|_2^2 + \nu_2\|\mathbf{z}\|_1$ , where  $\nu = [\nu_1, \nu_2] > 0$  and  $\|\mathbf{z}\|_2$  is the  $\ell_2$  norm. For this penalty, Liu et al. (2009b) created a well known package SLEP. The information they mainly used were the function value and gradient; Friedman et al. (2010) found a quadratic approximation  $L$  of  $\ell$  and applied the coordinate descent method into solving  $L + \phi_\nu$ . They packed the method into a solver GLMNET which was then improved by Yuan et al. (2012) to get the improved GLMNET.
- $\phi_\nu(\mathbf{z}) = \nu\|\mathbf{z}\|_2^2 + \delta_{\|\mathbf{z}\|_1 \leq t}(\mathbf{z})$ , where  $t > 0$  is given,  $\delta_{\|\mathbf{z}\|_1 \leq t}(\mathbf{z}) = 0$  if  $\|\mathbf{z}\|_1 \leq t$  and  $+\infty$  otherwise. Model (3) with this penalty can be solved by a first order method Lassplore (Liu et al., 2009a) or SLEP (Liu et al., 2009b); When  $\nu = 0$ , namely, the  $\ell_1$  constrained logistic regression,

$$\min_{\mathbf{z} \in \mathbb{R}^p} \ell(\mathbf{z}), \quad \text{s.t. } \|\mathbf{z}\|_1 \leq t, \quad (5)$$

Lee et al. (2006) developed the IRLS-LARS scheme where LARS (see Efron et al., 2004) was first introduced to address an  $\ell_1$  constrained least square problem and a direction that was analogous to Newton direction was then chosen to update next iteration.

**Nonconvex** regularizations differ slightly.

- **SCAD**: Fan and Li (2001) first took advantage of the smoothly clipped absolute deviation (SCAD) as the penalty.
- **One-step LLA**: Zou and Li (2008) developed a one step local linear approximation (LLA) algorithm in which at each step, the LLA estimator was applied to adopt a sparse representation.
- **Group Bridge**: Huang et al. (2009) benefited from a group bridge for multiple regression problems when covariates are grouped.
- **GIST**: Gong et al. (2013) designed a general iterative shrinkage and thresholding algorithm (GIST), which has the ability to precess a group of nonconvex regularizations such as SCAD, Log-sum penalty (LSP, Candes et al., 2008), Capped- $\ell_1$  penalty (Zhang, 2010), Minimax Concave Penalty (MCP, Zhang et al., 2010) and to name a few.
- **APG**: Li and Lin (2015) proposed the accelerated proximal gradient method (APG) to deal with (3) with the capped- $\ell_1$  penalty.
- **HONOR**: Gong and Ye (2015) proposed an efficient hybrid optimization algorithm for non-convex regularized problems (HONOR), where they effectively integrated a Quasi-Newton step and a gradient descent step.
- **DC-PN**: Rakotomamonjy et al. (2016) combined the idea of the difference of two convex functions and the proximal Newton optimization scheme (DC-PN) to solve (3). The regularization used in the model was a difference of two convex functions that equalled to the capped- $\ell_1$  penalty.

For sparsity constrained model (4), since it was introduced by Bahmani et al. (2013) and Beck and Eldar (2013), various approaches have been proposed.

- **GraSP**: Bahmani et al. (2013) generalized the compressive sampling matching pursuit (CoSaMP, see Needell and Tropp, 2009) to get the gradient support pursuit (GraSP). The latter is able to be reduced to the former when the logistic regression is replaced by the linear regression. Notice that apart from solving (4), GraSP was also able to address the model

$$\min \ell(\mathbf{z}) + (\nu/2)\|\mathbf{z}\|_2^2, \quad \text{s.t. } \mathbf{z} \in S. \quad (6)$$

- **Logit-GOMP**: Lozano et al. (2011) applied orthogonal matching pursuit (OMP, see Mallat and Zhang, 1993) into model (4), where the data is assumed to be classified into several groups, to develop a group OMP method (Logit-GOMP).
- **PD**: Lu and Zhang (2013) designed a penalty decomposition (PD) method in which a sequence of penalty subproblems were solved by a block coordinate descent method.

- **GraHTP**: Inspired by some greedy algorithms for compressing sensing (Donoho, 2006) such as iterative hard thresholding (IHT, see Blumensath and Davies, 2009) and hard thresholding pursuit (HTP, see Foucart, 2011), Yuan et al. (2014, 2018) designed gradient HTP methods GraHTP and FGraHTP.
- **NTGP**: Yuan and Liu (2014, 2017) benefited from Newton method to establish Newton greedy pursuit methods NTGP and QNTGP to solve the model (6).
- **IIHT**: Pan et al. (2017) followed the IHT method by (Blumensath and Davies, 2009) and built a convergent improved IHT method (IIHT).
- **GPGN**: Wang et al. (2017) combined the projected gradient method and the Newton method to derive a greedy projected gradient-Newton method (GPGN). The Newton steps were only applied into a series of chosen subspaces.
- **FNHTP**: Chen and Gu (2017) proposed a fast Newton hard thresholding pursuit (FNHTP) by using an unbiased stochastic Hessian estimator for the inverse Hessian matrix.

Notice that many of those above mentioned algorithms involve the Newton method, a second order method, with being able to be summarized into three groups. The first group comprises of EM, OWL-QN, OWL-QN\*, CDN, TNIP, HIS and HONOR. Those algorithms adopted Newton method to solve an unconstrained optimization. Most of them derived Newton directions through solving a linear equation system with at least  $p$  variables and  $p$  equations. The second group contains NTGP and GPGN, with performing Newton steps on chosen subspaces. The former in each step updated an iterate by processing a subproblem, that is  $\mathbf{z}^{k+1} = \operatorname{argmin}_{\mathbf{z} \in S} Q(\mathbf{z}; \mathbf{z}^k)$ , where  $Q(\mathbf{z}; \mathbf{z}^k)$  a local quadratic approximation of  $\ell(\mathbf{z})$  at current iterate  $\mathbf{z}^k$ . This subproblem was then solved by Newton method. While GPGN benefited from IHT to update an iterate and only imposed Newton step on a subspace when two consecutive iterates had same support sets. The third group seeks an approximation to estimate the inverse of the whole Hessian matrix, such as an unbiased stochastic Hessian estimator used in FNHTP, with hence low computational complexity compared with methods from group one. Inspired by those well established researches, we make use of Newton method to solve (4) as well. However, the way for us to use this method is different with that of any of above algorithms.

### 1.3 Our contributions

As what we mentioned above, we focus on the model (4) via the classic Newton method in this paper. The main contributions are summarized as follows.

- i) We start with considering the optimality condition of model (4) by introducing a strong  $\tau$ -stationary point (see Definition 3 for more details) which turns out to be a local (even a global under some circumstances) minimizer of problem (4) by Theorem 6. More importantly, a strong  $\tau$ -stationary point draws forth the stationary equation (20), a key concept in this paper that makes the classic Newton method applicable.
- ii) Differing with any of above mentioned algorithms, we perform Newton method directly on solving the stationary equation, one kind of optimality conditions of problem (4). Thanks to nice properties of the stationary equation, our proposed Newton

method dubbed as **NSLR**, an abbreviation for Newton method for the SLR, has a simple framework (see Table 1) making its implementation easy and has a low computational complexity  $\mathcal{O}(s^3 + s^2n + np)$  per each iteration. One of reasons for such low computational complexity is that we only calculate a small linear equation system with  $s$  variables and  $s$  equations to update the Newton direction.

- iii) In spite of many well established convergence results of Newton method to process continuous optimization, it is not trivial to do convergence analysis of Newton method to tackle (4), a combinatorial optimization. We show that the generated sequence has a local quadratic convergence and **NSLR** terminates within finite steps which can be estimated as  $\mathcal{O}(\log_2(c/\sqrt{\epsilon}))$  (see Theorem 12 iv) for details) if the chosen initial point is sufficiently close to a strong  $\tau$ -stationary point, where  $c$  is a constant being associated with samples and the initial point,  $\epsilon$  is the the stopping tolerance of **NSLR**.
- iv) Finally, the efficiency of **NSLR** is demonstrated against seven state-of-the-art methods on a number of randomly generated and real datasets. The fitting accuracy and computational speed are very competitive. Especially, in high dimensional data setting, **NSLR** outperforms others in terms of the computational time.

Now we would like to highlight the difference between above mentioned second order methods and our proposed method. Firstly, **NSLR** exploits Newton method to solve the stationary equation of the original problem (4) itself, which clearly differs with those in group one that aimed at solving unconstrained optimizations. Most of those methods were typical computationally slow due to addressing a large linear equation system in each step with complexity roughly about  $\mathcal{O}(p^3)$ . By contrast, **NSLR** only tackles a relatively small linear equation system with complexity roughly about  $\mathcal{O}(s^3)$ . Secondly, unlike **NTGP** using Newton method to solve quadratic subproblems, no subproblems in **NSLR** are considered. Moreover, our proposed method enjoys the local quadratic convergence rate and terminates with finite steps. Compared with **GPGN** who combined **IHT** procedures and Newton directions together, **NSLR** always benefits from Newton direction. Finally, differing with **FNHTP** to seek an unbiased stochastic Hessian to approximate the inverse of the whole Hessian matrix, **NSLR** only concentrates on a small sub-matrix of the whole Hessian, which results in a small scale linear equation to be calculated.

#### 1.4 Organization and notation

This paper is organized as follows. To explore the optimality conditions of (4), the next section introduces a strong  $\tau$ -stationary point (see Definition 3) and establishes the its relationship with a local/global minimizer (see Theorem 6). This relation allows us to solve a stationary equation system (20) to acquire a local/global optimal solution of (4). Based on the stationary equation, Section 3 proposes our method **NSLR**, an abbreviation for Newton method for SLR, which turns out to have simple algorithmic framework and low computational complexity. In Section 4, we establish the convergence results, including the local quadratic convergence and termination with finite steps, of the sequence generated by **NSLR**. In Section 5, the superior performance of **NSLR** is demonstrated against some of the state-of-the-art solvers on randomly generated and real datasets in high dimensional scenarios. Concluding remarks are made in the last section.

To end this section, we would like to define some notation. Let  $[\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n] =: X^\top \in \mathbb{R}^{p \times n}$  be the sample matrix and  $\mathbf{y} = (y_1, y_2, \dots, y_n)^\top \in \mathbb{R}^n$  be the response vector. Denote  $N_p := \{1, 2, \dots, p\}$ . Let  $|T|$  be the number of elements of  $T \subseteq N_p$  and  $\bar{T} := N_p \setminus T$  be the complementary set  $T$ . Write  $\mathbf{z}_T \in \mathbb{R}^{|T|}$  (resp.  $X_T$ ) as the sub vector (resp. matrix) of  $\mathbf{z}$  (resp.  $X$ ) containing elements (resp. columns) indexed on  $T$ . Similarly,  $X_{I,T}$  is the sub matrix containing rows indexed on  $I$  and columns indexed on  $T$ . Let  $0_{ab}$  stand for an order  $a \times b$  zero matrix and  $\mathcal{I}_a$  be an order  $a \times a$  identity matrix. For notational convenience, we write  $0$  to denote all  $0_{ab}$  if there is no ambiguity in the context and combine two vectors as  $(\mathbf{x}; \mathbf{y}) := (\mathbf{x}^\top \ \mathbf{y}^\top)^\top$ . We denote  $[\mathbf{z}]_i^\downarrow$  the  $i$ th largest (in absolute) elements of  $\mathbf{z}$ . Let  $\|\cdot\|$  denote the Frobenius norm for a matrix and Euclidean norm for a vector respectively, and  $\|\cdot\|_2$  denote the Spectral norm for a matrix. Furthermore,  $\lambda_{\min}(A)$  and  $\lambda_{\max}(A)$  are respectively the minimal and maximal eigenvalues of the symmetric matrix  $A$ . Denote  $\mathbf{1}$  the vector with all entries being ones.

## 2. Optimality and Stationary Equation

This section is devoted to investigate the optimality conditions and a stationary equation of (4). We summarize some properties of its objective function  $\ell(\mathbf{z})$  based on Wang et al. (2017), which will benefit for this paper.

**Property 1 (Lemma 2.2-2.4, Lemma A.3 (Wang et al., 2017))** *The objective function  $\ell(\mathbf{z})$  of (4) is twice continuously differentiable and has the following basic properties:*

- i) *The logistic loss function  $\ell(\mathbf{z})$  is nonnegative, convex and strongly smooth on  $\mathbb{R}^p$  with a parameter  $\lambda_x := \lambda_{\max}(X^\top X)/(4n)$ , namely, for any  $\mathbf{z}, \mathbf{z}^* \in \mathbb{R}^p$*

$$\ell(\mathbf{z}) \leq \ell(\mathbf{z}^*) + \langle \nabla \ell(\mathbf{z}^*), \mathbf{z} - \mathbf{z}^* \rangle + (\lambda_x/2) \|\mathbf{z} - \mathbf{z}^*\|^2. \quad (7)$$

- ii) *The gradient  $\nabla \ell(\mathbf{z})$  is given by*

$$\nabla \ell(\mathbf{z}) = X^\top (h(\mathbf{z}) - \mathbf{y})/n, \quad (8)$$

where  $h(\mathbf{z})$  is a vector with  $(h(\mathbf{z}))_i = 1/(1 + e^{-\langle \mathbf{x}_i, \mathbf{z} \rangle})$ ,  $i = 1, \dots, n$ . And it enjoys

$$\|\nabla \ell(\mathbf{z}) - \nabla \ell(\mathbf{z}^*)\| \leq \lambda_x \|\mathbf{z} - \mathbf{z}^*\|. \quad (9)$$

- iii) *The Hessian matrix  $\nabla^2 \ell(\mathbf{z})$  is given by*

$$\nabla^2 \ell(\mathbf{z}) = X^\top D(\mathbf{z}) X/n, \quad (10)$$

where  $D(\mathbf{z})$  is a diagonal matrix with with

$$(D(\mathbf{z}))_{ii} = \frac{e^{\langle \mathbf{x}_i, \mathbf{z} \rangle}}{(1 + e^{\langle \mathbf{x}_i, \mathbf{z} \rangle})^2} \in (0, 1/4], \quad i = 1, \dots, n.$$

- iv) *For any  $\mathbf{z}_1, \mathbf{z}_2 \in \mathbb{R}^p$ ,*

$$\|\nabla^2 \ell(\mathbf{z}_1) - \nabla^2 \ell(\mathbf{z}_2)\| \leq \gamma_x \|\mathbf{z}_1 - \mathbf{z}_2\|, \quad (11)$$

where  $\gamma_x := 12\lambda_x \max_{i=1, \dots, n} \|\mathbf{x}_i\|_1$ .

Our first result reveals the relation between the logistic regression and linear regression.

**Property 2** For (4), denoting  $\mathbf{c} := 2\mathbf{y} - \mathbf{1}$ , we have results below.

- i)  $0 \leq \min_{\mathbf{z} \in S} \ell(\mathbf{z}) \leq (\ln(2) - 1/4) + \min_{\mathbf{z} \in S} \|X\mathbf{z} - \mathbf{c}\|^2 / (4n)$ .
- ii) If  $X^\top \mathbf{c} \neq 0$ , then there always exists a subset  $T \subset N_p$  with  $r := |T| = \text{rank}(X_T) \leq s$  that satisfies  $X_T^\top \mathbf{c} \neq 0$ . For such  $T$ , let the singular value decomposition of  $X_T$  be

$$X_T = U\Lambda V^\top,$$

where  $U \in \mathbb{R}^{n \times r}$  and  $V \in \mathbb{R}^{r \times r}$  are column orthogonal matrices and  $\Lambda \in \mathbb{R}^{r \times r}$  is a diagonal matrix, then there is a  $\bar{\mathbf{z}} \in \mathbb{R}^p$  with  $\bar{\mathbf{z}}_T = V\Lambda^{-1}U^\top \mathbf{c}$  and  $\bar{\mathbf{z}}_{\bar{T}} = 0$  such that

$$\bar{\mathbf{z}} \in S \quad \text{and} \quad \ell(\bar{\mathbf{z}}) < \ell(0).$$

If  $X^\top \mathbf{c} = 0$ , then 0 is a global minimizer of (4) and  $0 \in \text{argmin}_{\mathbf{z} \in S} \|X\mathbf{z} - \mathbf{c}\|^2$ .

The relation between the logistic regression and linear regression in Property 2 i) provides a hint to initialize a starting point (i.e.,  $\mathbf{z}^0 \in \text{argmin}_{\mathbf{z} \in S} \|X\mathbf{z} - \mathbf{c}\|^2$ ) in terms of algorithmic design. Property 2 ii) states that there are many non-trivial feasible solutions such that their loss function values are strictly less than  $\ell(0)$  if  $X^\top \mathbf{c} \neq 0$ . Otherwise 0 is the global solution of (4) and meantime  $0 \in \text{argmin}_{\mathbf{z} \in S} \|X\mathbf{z} - \mathbf{c}\|^2$  due to  $X^\top(X0 - \mathbf{c}) = 0$ , which means the equation of the second inequality of i) holds.

When it comes to characterize the solutions of (4), we introduce the definition of a strong  $\tau$ -stationary point. To proceed, we first give the conception of projection of a vector  $\beta \in \mathbb{R}^p$  onto the sparse set  $S$ :

$$P_S(\mathbf{z}) := \text{argmin}\{\|\mathbf{z} - \mathbf{x}\| : \mathbf{x} \in S\},$$

which sets all but  $s$  largest absolute value components of  $\mathbf{z}$  to zero. Since the right hand side may have multiple solutions,  $P_S(\mathbf{z})$  is a set. But for the sake of notational convenience, we write  $P_S(\mathbf{z}) = \mathbf{x}$  when  $P_S(\mathbf{z}) = \{\mathbf{x}\}$  is a singleton.

**Definition 3** (Lu, 2015, Definition 3.2) A vector  $\mathbf{z} \in S$  is called a strong  $\tau$ -stationary point of (4) if there is a  $\tau > 0$  such that

$$\mathbf{z} = P_S(\mathbf{z} - \tau \nabla \ell(\mathbf{z})). \quad (12)$$

Recall that a  $\tau$ -stationary point is defined by Beck and Eldar (2013),

$$\mathbf{z} \in P_S(\mathbf{z} - \tau \nabla \ell(\mathbf{z})).$$

Lemma 2.2 in (Beck and Eldar, 2013) tells us that  $\mathbf{z}$  is a  $\tau$ -stationary point if and only if

$$\|\mathbf{z}\|_0 \leq s \quad \text{and} \quad |\nabla_i \ell(\mathbf{z})| \begin{cases} = 0, & i \in \text{supp}(\mathbf{z}), \\ \leq \frac{1}{\tau} [\mathbf{z}]_s^\dagger, & i \notin \text{supp}(\mathbf{z}). \end{cases} \quad (13)$$

Similarly, we have following properties regarding to a strong  $\tau$ -stationary point.



**Property 4** For a given  $\tau > 0$ ,  $\mathbf{z} \in \mathbb{R}^p$  is a strong  $\tau$ -stationary point if and only if

- i)  $\|\mathbf{z}\|_0 < s$  and  $\nabla\ell(\mathbf{z}) = 0$ , or
- ii)  $\|\mathbf{z}\|_0 = s$  and

$$|\nabla_i\ell(\mathbf{z})| \begin{cases} = 0, & i \in \text{supp}(\mathbf{z}), \\ < \frac{1}{\tau}[\mathbf{z}]_s^\downarrow, & i \notin \text{supp}(\mathbf{z}). \end{cases} \quad (14)$$

The proof is similar to that of Lemma 2.2 by Beck and Eldar (2013), and thus is omitted here. Now we would like to emphasize the relationship between these two kinds of points.

**Remark 5** The relationship between a strong  $\tau$ -stationary point and  $\tau$ -stationary point can be described as follows:

- i) If  $\|\mathbf{z}\|_0 < s$ , then  $\mathbf{z}$  is a  $\tau$ -stationary point if and only if it is a strong  $\tau$ -stationary point because of  $\nabla\ell(\mathbf{z}) = 0$ .
- ii) If  $\|\mathbf{z}\|_0 = s$ , then a strong  $\tau$ -stationary point  $\mathbf{z}$  is also a  $\tau$ -stationary point. In the meantime, a  $\tau$ -stationary point  $\mathbf{z}$  is also a strong  $\tau_1$ -stationary point with any  $0 < \tau_1 < \tau$  since  $|\nabla_i\ell(\mathbf{z})| \leq [\mathbf{z}]_s^\downarrow/\tau < [\mathbf{z}]_s^\downarrow/\tau_1$ .

We use one simple example to illustrate ii) of above Remark. Consider  $s = 2$  and

$$X = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad y = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}, \quad \mathbf{z}^* = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}.$$

Direct calculation yields  $\nabla\ell(\mathbf{z}^*) = (0 \ 0 \ 0.5)^\top$ . One can verify that  $\mathbf{z}^* \in P_S(\mathbf{z}^* - 2\nabla\ell(\mathbf{z}^*))$  and  $\mathbf{z}^* \in P_S(\mathbf{z}^* - \tau_1\nabla\ell(\mathbf{z}^*))$  for any given  $0 < \tau_1 < 2$ . This means  $\mathbf{z}^*$  is a  $\tau$ -stationary point for any given  $0 < \tau \leq 2$  and a strong  $\tau_1$ -stationary point for any given  $0 < \tau_1 < 2$ . Based on the definition of strong  $\tau$ -stationary point, our first main result is to establish the relationship between a local/global optimal solution and a strong  $\tau$ -stationary point of (4).

**Theorem 6** Considering (4), following results hold.

- i) A global minimizer is a strong  $\tau$ -stationary point  $\mathbf{z}^*$  for any given  $\tau \in (0, 1/\lambda_x)$ .
- ii) If  $\|\mathbf{z}^*\|_0 = s$ , a strong  $\tau$ -stationary point  $\mathbf{z}^*$  for a given  $\tau > 0$  is a local minimizer. If it further satisfies  $\nabla\ell(\mathbf{z}^*) = 0$  then it is also a global minimizer.
- iii) If  $\|\mathbf{z}^*\|_0 < s$ , a strong  $\tau$ -stationary point  $\mathbf{z}^*$  for a given  $\tau > 0$  is a global minimizer.

The above established relationship allows us to focus on a strong  $\tau$ -stationary point itself to pursuit a ‘good’ solution of (4). Rewrite (12) as

$$\begin{cases} \mathbf{z} = P_S(\mathbf{z} - \tau\mathbf{d}), \\ \mathbf{d} = \nabla\ell(\mathbf{z}), \end{cases} \quad (15)$$

which is equivalent to

$$\begin{bmatrix} \mathbf{z} - P_S(\mathbf{z} - \tau \mathbf{d}) \\ \mathbf{d} - \nabla \ell(\mathbf{z}) \end{bmatrix} = 0. \quad (16)$$

Regarding to the projected property of  $P_S(\mathbf{z})$ , it sets all but  $s$  largest absolute value components of  $\mathbf{z}$  to zero. To well express the solution of  $P_S(\mathbf{z} - \tau \mathbf{d})$ , we define two sets

$$T(\mathbf{u}; \tau) := \{i \in N_p : |\mathbf{z}_i - \tau \mathbf{d}_i| \geq [\mathbf{z} - \tau \mathbf{d}]_s^\downarrow\} \quad \text{with} \quad |T(\mathbf{u}; \tau)| = s \quad (17)$$

$$\bar{T}(\mathbf{u}; \tau) := \{i \in N_p : i \notin T(\mathbf{u}; \tau)\} \quad (18)$$

where  $\mathbf{u} = (\mathbf{z}; \mathbf{d}) \in \mathbb{R}^{2p}$ . Clearly, it has  $T(\mathbf{u}, \tau) \cap \bar{T}(\mathbf{u}, \tau) = \emptyset$ ,  $T(\mathbf{u}, \tau) \cup \bar{T}(\mathbf{u}, \tau) = N_p$ . One may discern that  $T(\mathbf{u}; \tau)$  may not be unique. So we denote  $T^B(\mathbf{u}; \tau)$  the set that covers all  $T(\mathbf{u}; \tau)$  in (17). In other words,  $T(\mathbf{u}; \tau)$  is a particular element of  $T^B(\mathbf{u}; \tau)$ , i.e.,

$$T(\mathbf{u}; \tau) \in T^B(\mathbf{u}; \tau).$$

Hereafter, if no confusion arises, we use the simple notation

$$T := T(\mathbf{u}; \tau), \quad \bar{T} := \bar{T}(\mathbf{u}; \tau).$$

One should keep in mind  $T$  is associated with  $\mathbf{u}$  and  $\tau$ , but for notational convenience, we drop the dependence. Based on those notation, for any  $T \in T^B(\mathbf{u}; \tau)$ , the projection is able to be expressed as

$$P_S(\mathbf{z} - \tau \mathbf{d}) = \begin{bmatrix} (\mathbf{z} - \tau \mathbf{d})_T \\ 0 \end{bmatrix}, \quad (19)$$

and thus (16) implies that the following *stationary equation*

$$F_\tau(\mathbf{u}; T) := \begin{bmatrix} \mathbf{d}_T \\ \mathbf{z}_{\bar{T}} \\ \mathbf{d}_T - \nabla_T \ell(\mathbf{z}) \\ \mathbf{d}_{\bar{T}} - \nabla_{\bar{T}} \ell(\mathbf{z}) \end{bmatrix} = 0 \quad (20)$$

holds for any  $T \in T^B(\mathbf{u}; \tau)$ . It is worth mentioning that if  $T \notin T^B(\mathbf{u}; \tau)$ , then (19) is not correct and hence (20) may not hold.

The idea of the *stationary equation* refers to the results in Huang et al. (2018), in which authors concentrated on  $\ell_0$ -regularization linear regression. In this paper, we aim to solve the logistic regression with sparsity constraint, and we provide different theoretical analysis as well as some novel convergence results (refer to Section 4) from the perspective of optimization. The following theorem confirms the equivalence between (16) and (20).

**Theorem 7**  $\mathbf{z}$  is a strong  $\tau$ -stationary point for a given  $\tau > 0$  if and only if

$$F_\tau(\mathbf{u}; T) = 0 \quad \text{for all} \quad T \in T^B(\mathbf{u}; \tau).$$

where  $\mathbf{u} = (\mathbf{z}; \mathbf{d})$  is denoted by the above.

For a given  $\mathbf{u} = (\mathbf{z}; \mathbf{d})$  and  $\tau > 0$ , if a  $T \in T^B(\mathbf{u}; \tau)$  is given, then the Jacobian matrix of  $F_\tau(\mathbf{u}; T)$  enjoys the form

$$\nabla F_\tau(\mathbf{u}; T) = \begin{bmatrix} 0_{ss} & 0_{sr} & \mathcal{I}_s & 0_{sr} \\ 0_{rs} & \mathcal{I}_r & 0_{rs} & 0_{rr} \\ -\nabla_{T,T}^2 \ell(\mathbf{z}) & -\nabla_{T,\bar{T}}^2 \ell(\mathbf{z}) & \mathcal{I}_s & 0_{sr} \\ -\nabla_{\bar{T},T}^2 \ell(\mathbf{z}) & -\nabla_{\bar{T},\bar{T}}^2 \ell(\mathbf{z}) & 0_{rs} & \mathcal{I}_r \end{bmatrix}, \quad (21)$$

where  $r := p - s$  and  $\nabla_{T,\bar{T}}^2 \ell(\mathbf{z}) := (\nabla^2 \ell(\mathbf{z}))_{T,\bar{T}}$ . It is worth mention that since  $T$  is related to  $\mathbf{u}$ , the Jacobian of  $F_\tau(\mathbf{u}; T)$  may differ with (21) if we treat  $T$  to be unknown. However, we do not concern about the latter case and only focus on the case  $T \in T^B(\mathbf{u}; \tau)$  being given when it comes to the algorithmic design. We next show that for any given  $T$ , it does not affect the overall structure and non-singularity of the matrix  $\nabla F_\tau(\mathbf{u}; T)$ . To see this, we would like to introduce the so-called  $s$ -regularity of  $X$ .

**Definition 8** (Definition 2.2, Beck and Eldar, 2013) *A matrix is  $s$ -regular if its any  $s$  columns are linearly independent.*

If  $X$  is  $s$ -regular, then it enables us to well define

$$\underline{\lambda} := \min_{\mathbf{z} \in S} \frac{\mathbf{z}^\top X^\top X \mathbf{z}}{\mathbf{z}^\top \mathbf{z}} = \min_{|T| \leq s} \lambda_{\min}(X_T^\top X_T) > 0. \quad (22)$$

**Theorem 9** *Suppose the matrix  $X$  is  $s$ -regular. Then for any given  $T \in T^B(\mathbf{u}; \tau)$  with a given  $\tau > 0$ ,  $\nabla F_\tau(\mathbf{u}; T)$  is nonsingular on  $\mathbb{R}^{2p}$  and its inverse matrix has the form*

$$(\nabla F_\tau(\mathbf{u}; T))^{-1} = \begin{bmatrix} (\nabla_{T,T}^2 \ell)^{-1} & -(\nabla_{T,T}^2 \ell)^{-1} \nabla_{T,\bar{T}}^2 \ell & -(\nabla_{T,T}^2 \ell)^{-1} & 0 \\ 0 & \mathcal{I}_r & 0 & 0 \\ \mathcal{I}_s & 0 & 0 & 0 \\ (\nabla_{\bar{T},T}^2 \ell)(\nabla_{T,T}^2 \ell)^{-1} & R(\mathbf{z}) & -(\nabla_{\bar{T},T}^2 \ell)(\nabla_{T,T}^2 \ell)^{-1} & \mathcal{I}_r \end{bmatrix}, \quad (23)$$

where  $\ell := \ell(\mathbf{z})$  and  $R(\mathbf{z}) := -\nabla_{T,T}^2 \ell (\nabla_{T,T}^2 \ell)^{-1} \nabla_{T,\bar{T}}^2 \ell + \nabla_{\bar{T},\bar{T}}^2 \ell$ . Moreover, for any  $\mathbf{z}^* \in \mathbb{R}^p$  and a given  $\delta > 0$ ,

$$\|(\nabla_{T,T}^2 \ell(\mathbf{z}))^{-1}\|_2 \leq \mu(X, \mathbf{z}^*, \delta) := \frac{e^{2(\sqrt{\lambda_x} \|\mathbf{z}^*\| + \delta)}}{\underline{\lambda}/(4n)}, \quad \forall \mathbf{z} \in N(\mathbf{z}^*, \delta), \quad (24)$$

which means  $\|(\nabla F_\tau(\mathbf{u}; T))^{-1}\|$  is bounded on  $N(\mathbf{u}^*, \delta)$ .

### 3. Fast Newton Method

In this section, we aim to design an efficient algorithm to solve the stationary equation (20). Let  $\mathbf{u}^k = (\mathbf{z}^k; \mathbf{d}^k)$  be the  $k$ th iterative point and fix  $\tau > 0$ . Similar to (17) and (18), the  $k$ th iterative index sets are defined as

$$T_k := T(\mathbf{u}^k; \tau), \quad \bar{T}_k := \bar{T}(\mathbf{u}^k; \tau).$$

Theorem 9 tells us that  $\nabla F_\tau(\mathbf{u}^k; T_k)$  is nonsingular under the condition of the  $s$ -regularity of the matrix  $X$ . Therefore we naturally think of the Newton method to solve the system of stationary equation (20). The classical Newton algorithm reads

$$\nabla F_\tau(\mathbf{u}^k; T_k)(\mathbf{u}^{k+1} - \mathbf{u}^k) = -F_\tau(\mathbf{u}^k; T_k). \quad (25)$$

From the explicit formula of  $F_\tau(\mathbf{u}^k; T_k)$  in (20), we have

$$\begin{aligned} & \begin{aligned} \mathbf{d}_{T_k}^{k+1} &= 0, \\ \mathbf{z}_{\bar{T}_k}^{k+1} &= 0, \end{aligned} \\ \left[ \begin{array}{cc} \nabla_{T_k, T_k}^2 \ell(\mathbf{z}^k) & \nabla_{T_k, \bar{T}_k}^2 \ell(\mathbf{z}^k) \\ \nabla_{\bar{T}_k, T_k}^2 \ell(\mathbf{z}^k) & \nabla_{\bar{T}_k, \bar{T}_k}^2 \ell(\mathbf{z}^k) \end{array} \right] \begin{bmatrix} \mathbf{z}_{T_k}^{k+1} - \mathbf{z}_{T_k}^k \\ -\mathbf{z}_{\bar{T}_k}^k \end{bmatrix} - \begin{bmatrix} 0 \\ \mathbf{d}_{\bar{T}_k}^{k+1} \end{bmatrix} &= - \begin{bmatrix} \nabla_{T_k} \ell(\mathbf{z}^k) \\ \nabla_{\bar{T}_k} \ell(\mathbf{z}^k) \end{bmatrix}. \end{aligned} \quad (26)$$

The third equation in (26) yields that

$$\nabla_{T_k, T_k}^2 \ell(\mathbf{z}^k) \mathbf{z}_{T_k}^{k+1} - \nabla_{T_k}^2 \ell(\mathbf{z}^k) \mathbf{z}^k = -\nabla_{T_k} \ell(\mathbf{z}^k),$$

where  $\nabla_{T_k}^2 \ell(\mathbf{z})$  is the sub-matrix of  $\nabla^2 \ell(\mathbf{z})$  containing rows indexed on  $T$ , which implies that  $\mathbf{z}_{T_k}^{k+1}$  is a solution of the following linear equation

$$\begin{aligned} \nabla_{T_k, T_k}^2 \ell(\mathbf{z}^k) \mathbf{v} &= \nabla_{T_k}^2 \ell(\mathbf{z}^k) \mathbf{z}^k - \nabla_{T_k} \ell(\mathbf{z}^k) \\ &= \nabla_{T_k, T_{k-1}}^2 \ell(\mathbf{z}^k) \mathbf{z}_{T_{k-1}}^k - \nabla_{T_k} \ell(\mathbf{z}^k), \end{aligned} \quad (27)$$

where the second equation holds due to  $\text{supp}(\mathbf{z}^k) \subseteq T_{k-1}$  from (26). Overall, we update  $\mathbf{u}^{k+1}$  through (26) by

$$\begin{cases} \mathbf{z}_{T_k}^{k+1} &= \mathbf{v}^k, \\ \mathbf{z}_{\bar{T}_k}^{k+1} &= 0, \\ \mathbf{d}_{T_k}^{k+1} &= 0, \\ \mathbf{d}_{\bar{T}_k}^{k+1} &= \nabla_{\bar{T}_k} \ell(\mathbf{z}^k) + \nabla_{\bar{T}_k}^2 \ell(\mathbf{z}^k) (\mathbf{z}^{k+1} - \mathbf{z}^k), \end{cases} \quad (28)$$

where  $\mathbf{v}^k$  is a solution of (27). If  $\nabla_{T_k, T_k}^2 \ell(\mathbf{z}^k)$  is nonsingular, (28) means we do not need to calculate the inverse of the whole matrix  $\nabla F_\tau(\mathbf{u}^k; T_k)$  to update  $\mathbf{u}^{k+1}$ , because it can be done through computing the inverse of  $\nabla_{T_k, T_k}^2 \ell(\mathbf{z}^k)$  to update  $\mathbf{z}_{T_k}^{k+1}$  as

$$\mathbf{z}_{T_k}^{k+1} = \mathbf{v}^k = \mathbf{z}_{T_k}^k - \left[ \nabla_{T_k, T_k}^2 \ell(\mathbf{z}^k) \right]^{-1} \left[ \nabla_{T_k} \ell(\mathbf{z}^k) - \nabla_{T_k, \bar{T}_k}^2 \ell(\mathbf{z}^k) \mathbf{z}_{\bar{T}_k}^k \right]. \quad (29)$$

Hence we propose the Newton method for (4) to iteratively process the stationary equation (20). The algorithmic framework is summarized as in following table.

Table 1: Framework of the algorithm.

---

NSLR: Newton method for the SLR (4)	
<b>Step 0</b>	Initialize $\mathbf{z}^0$ , $\mathbf{d}^0 = \nabla \ell(\mathbf{z}^0)$ . Choose $\epsilon, \tau > 0$ . Set $k \leftarrow 0$ .
<b>Step 1</b>	(Support set Selection) Choose $T_k := T(\mathbf{u}^k; \tau) \in T^B(\mathbf{u}^k; \tau)$ by (17).
<b>Step 2</b>	(Convergence Check) If $\ F_\tau(\mathbf{u}^k; T_k)\  \leq \epsilon$ , then stop. Otherwise, go to <b>Step 3</b> .
<b>Step 3</b>	(Full Newton) Update $\mathbf{u}^{k+1} = (\mathbf{z}^{k+1}; \mathbf{d}^{k+1})$ by (28), set $k \leftarrow k + 1$ and go to <b>Step 1</b> .

---

Regarding to the proposed algorithm NSLR, we have some comments.

- i) For **Step 0**, Property 2 gives us a clue to find a starting point  $\mathbf{z}^0 \in \operatorname{argmin}_{\mathbf{z} \in S} \|X\mathbf{z} - \mathbf{c}\|^2$ , which, however, would consume much time. Therefore, we will simply use  $\mathbf{z}^0 = 0$ . In addition, we will show that if the starting point  $\mathbf{z}^0$  is chosen sufficiently closely to a strong  $\tau$ -stationary point  $\mathbf{z}^*$  with  $\|\mathbf{z}^*\|_0 = s$ , all  $T_k$  will be identical and, furthermore, NSLR will terminate within finite steps (see Theorem 12). Interestingly, in terms of numerical experiments, to fasten the convergence of the proposed method, the choice of starting point  $\mathbf{z}^0$  is not necessary to be close enough to a strong  $\tau$ -stationary point. For simplicity, one could just start from the origin.
- ii) For **Step 1**, we only pick  $s$  largest elements (in absolute) to form  $T_k$ , which allows us to use `min` function in MATLAB (2017b or later version) whose computational complexity is  $\mathcal{O}(p + s \log s)$ .
- iii) For **Step 3**, to update  $\mathbf{z}_{T_k}^{k+1}$ , one need to solve a linear equation (27). The complexity of calculating  $\nabla_{T_1, T_2}^2 \ell(\mathbf{z}) = X_{T_1}^\top D(\mathbf{z}) X_{T_2} / n$  with  $|T_1| = |T_2| = s$  is  $\mathcal{O}(sn + s^2n)$  since  $D(\mathbf{z})$  is the diagonal matrix. The complexity of solving a linear equation with  $s$  equations and  $s$  variables is at most  $\mathcal{O}(s^3)$ . Therefore, the whole complexity of updating  $\mathbf{z}_{T_k}^{k+1}$  is  $\mathcal{O}(s^3 + s^2n)$ . For updating  $\mathbf{d}_{T_k}^{k+1}$ , since  $X(\mathbf{z}^{k+1} - \mathbf{z}^k) = X_{T_k \cup T_{k-1}}(\mathbf{z}^{k+1} - \mathbf{z}^k)_{T_k \cup T_{k-1}}$ , its computational complexity is  $\mathcal{O}(2sn)$ . So combining  $\nabla_{T_k}^2 \ell(\mathbf{z}^k)(\mathbf{z}^{k+1} - \mathbf{z}^k)$  in (28), the complexity of updating  $\mathbf{d}_{T_k}^{k+1}$  is  $\mathcal{O}(np + ns)$ . Overall, the computational complexity of **Step 3** is

$$\mathcal{O}(s^3 + s^2n + np),$$

which means the computation is quite fast if  $\max\{s, n\} \ll p$ .

- iv) It is worth mentioning that theoretically under the assumption  $X$  being  $s$ -regular (see Theorem 9), the linear equation (27) always admits a unique solution. Numerically, to avoid a non-singular case, one could calculate  $[\nabla_{T_k, T_k}^2 \ell(\mathbf{z}^k) + \mu \mathcal{I}]^{-1}$  instead with a small positive  $\mu$ . Actually, if we consider the model (6), namely  $\ell$  is replace by  $\ell + \mu \|\cdot\|_2^2$ , then (27) always has a unique solution without any assumptions.

## 4. Quadratic Convergence

Our first result states that if the sequence generated by NSLR is convergent, then it must converge to a  $\tau$ -stationary point of (4).

**Theorem 10** *For any given  $\tau > 0$ , let  $\mathbf{u}^\infty := (\mathbf{z}^\infty; \mathbf{d}^\infty)$  be any limit point of the sequence  $\{\mathbf{u}^k\}$  generated by NSLR, then  $\mathbf{z}^\infty$  is a  $\tau$ -stationary point of (4).*

Our next main result shows that if the initial point  $\mathbf{u}^0$  of the sequence  $\{\mathbf{u}^k\}$  is sufficiently close to a point  $\mathbf{u}^* = (\mathbf{z}^*; \nabla\ell(\mathbf{z}^*))$  where  $\mathbf{z}^*$  is a strong  $\tau^*$ -stationary point of (4) for a given  $\tau^* > 0$ , then the sequence converges to  $\mathbf{u}^*$  quadratically. Before this, we would like to define some notation to ease the reading. Denote  $\Gamma_* := \text{supp}(\mathbf{z}^*)$  and

$$\tau \in \begin{cases} (0, \tau^*], & \text{if } \|\mathbf{z}^*\|_0 = s, \\ (0, \infty), & \text{if } \|\mathbf{z}^*\|_0 < s, \end{cases} \quad (30)$$

$$\delta^* := \frac{\min_{i \in \Gamma_*} |z_i^*| - \tau^* \max_{i \notin \Gamma_*} |d_i^*|}{2 \max\{1, \tau^*\}}, \quad (31)$$

$$N(\mathbf{u}^*, \delta^*) := \begin{cases} \{\mathbf{u} = (\mathbf{z}; \mathbf{d}) \in \mathbb{R}^{2p} \mid \mathbf{z} \in S, \|\mathbf{u} - \mathbf{u}^*\| < \delta^*\}, & \text{if } \delta^* \neq 0, \\ \{\mathbf{u} = (\mathbf{z}; \mathbf{d}) \in \mathbb{R}^{2p} \mid \mathbf{z} \in S\}, & \text{if } \delta^* = 0. \end{cases} \quad (32)$$

Theorem 7 says that a strong  $\tau^*$ -stationary point  $\mathbf{z}^*$  is equivalent to  $F_{\tau^*}(\mathbf{u}^*; T_*) = 0$  for any  $T_* \in T^B(\mathbf{u}^*; \tau^*)$ . This implies  $\mathbf{d}^* = \nabla\ell(\mathbf{z}^*)$ . So one can easily verify that if  $\mathbf{z}^*$  is a strong  $\tau^*$ -stationary point, then  $\delta^* > 0$  if  $\mathbf{z}^* \neq 0$  and  $\delta^* = 0$  only if  $\mathbf{z}^* = 0$  by Property 4. The following lemma presents that for any  $\mathbf{z}$  being sufficiently close to  $\mathbf{z}^*$ , they share the same support set  $\text{supp}(\mathbf{z}^*)$ .

**Lemma 11** *Let  $\mathbf{z}^*$  be a strong  $\tau^*$ -stationary point of (4) for a given  $\tau^* > 0$ . Then for any  $\mathbf{u} \in N_S(\mathbf{u}^*, \delta^*)$ , following results hold.*

i) *If  $\|\mathbf{z}^*\|_0 = s$ , then for any given  $0 < \tau \leq \tau^*$  there is*

$$\{\text{supp}(\mathbf{z})\} = T^B(\mathbf{u}; \tau) = T^B(\mathbf{u}^*; \tau^*) = \{\text{supp}(\mathbf{z}^*)\}.$$

ii) *If  $\|\mathbf{z}^*\|_0 < s$ , then for any given  $\tau > 0$  there is  $T^B(\mathbf{u}; \tau) \subseteq T^B(\mathbf{u}^*; \tau^*)$  and*

$$\text{supp}(\mathbf{z}^*) \subseteq (\text{supp}(\mathbf{z}) \cap T), \quad \forall T \in T^B(\mathbf{u}; \tau).$$

Now we are ready to claim our main result.

**Theorem 12** *For (4), assume the matrix  $X$  is  $s$ -regular. Let  $\mathbf{z}^*$  be a strong  $\tau^*$ -stationary point for a given  $\tau^* > 0$  and  $\mathbf{u}^* = (\mathbf{z}^*; \mathbf{d}^*)$  with  $\mathbf{d}^* = \nabla\ell(\mathbf{z}^*)$ . Let  $\tau$ ,  $\delta^*$ ,  $N_S(\cdot, \cdot)$  and  $\mu(X, \mathbf{u}^*, \delta^*) =: \mu^*$  be defined as (30), (31), (32) and (24) respectively. Suppose that the initial point of sequence  $\{\mathbf{u}^k\}$  generated by NSLR satisfies*

$$\mathbf{u}^0 \in N_S(\mathbf{u}^*, \min\{\delta^*, \delta_x^*\}), \quad \text{with} \quad \delta_x^* := (\sqrt{1 + 4\lambda_x^2} \gamma_x \mu^*)^{-1}. \quad (33)$$

*Then for any  $k \geq 0$ , following results hold.*

i)  $\lim_{k \rightarrow \infty} \mathbf{u}^k = \mathbf{u}^*$  with quadratic convergence rate, namely,

$$\|\mathbf{u}^{k+1} - \mathbf{u}^*\| \leq (0.5/\delta_x^*) \|\mathbf{u}^k - \mathbf{u}^*\|^2.$$

ii) If  $\|\mathbf{z}^*\|_0 = s$  it holds

$$T^B(\mathbf{u}^k; \tau) = \{\text{supp}(\mathbf{z}^k)\} = \{\text{supp}(\mathbf{z}^*)\} = T^B(\mathbf{u}^*; \tau^*).$$

If  $\|\mathbf{z}^*\|_0 < s$  it holds

$$\text{supp}(\mathbf{z}^*) \subseteq (\text{supp}(\mathbf{z}^k) \cap T_k), \quad \forall T_k \in T^B(\mathbf{u}^k; \tau).$$

iii)  $\|F_\tau(\mathbf{u}^{k+1}; T_{k+1})\| \leq c_x \|\mathbf{u}^k - \mathbf{u}^*\|^2$  and NSLR will terminate when

$$k \geq \lceil \log_2(\sqrt{c_x} \|\mathbf{u}^0 - \mathbf{u}^*\|) + \log_2(1/\sqrt{\epsilon}) \rceil,$$

where  $\lceil a \rceil$  is the smallest integer being no less than  $a$  and  $c_x := [(0.5/\delta_x^*)^2 + (1.25\gamma_x)^2]^{\frac{1}{2}}$ .

Compared with the classical convergence results of Newton method, Theorem 12 gives the explicit value of the neighborhood and proves that NSLR will terminate within finite steps.

## 5. Numerical Experiments

In this part, we will conduct extensive numerical experiments of our algorithm NSLR by using MATLAB (R2017b) on a desktop of 8GB of memory and Inter Core i5 2.7Ghz CPU, against seven leading solvers both on synthetic data and real data.

### 5.1 Test examples

We first do numerical experiments on synthetic data. In such simulations, we adopt two types of data generation. Lu and Zhang (2013); Pan et al. (2017) have used the model with identically independently generated features  $[\mathbf{x}_1 \cdots \mathbf{x}_n] \in \mathbb{R}^{p \times n}$ , whilst Agarwal et al. (2010); Bahmani et al. (2013) have considered independent features with each of which  $\mathbf{x}_i$  being generated by an autoregressive process (Hamilton, 1994).

**Example 1 (Independent Data (Lu and Zhang, 2013; Pan et al., 2017))** To generate data labels  $\mathbf{y} \in \{0, 1\}^n$ , we first randomly separate  $\{1, \dots, n\}$  into two parts  $I$  and  $\bar{I}$  and set  $y_i = 0$  for  $i \in I$  and  $y_i = 1$  for  $i \in \bar{I}$ . Then the feature data is produced by

$$\mathbf{x}_i = y_i v_i \mathbf{1} + \mathbf{w}_i, \quad i = 1, \dots, n$$

with  $\mathbb{R} \ni v_i \sim \mathcal{N}(0, 1)$ ,  $\mathbb{R}^p \ni \mathbf{w}_i \sim \mathcal{N}(0, \mathcal{I}_p)$ , where  $\mathcal{N}(0, \mathcal{I})$  is the normal distribution with mean zero and variance identity. Since the sparse parameter  $\mathbf{z}^* \in \mathbb{R}^p$  is unknown, different  $s (< n)$  will be tested to pursuit a sparse solution.

**Example 2 (Correlated Data (Agarwal et al., 2010; Bahmani et al., 2013))** The sparse parameter  $\mathbf{z}^* \in \mathbb{R}^p$  has  $s$  nonzero entries drawn independently from the standard Gaussian distribution. Each data sample  $\mathbf{x}_i = [x_{i1} \cdots x_{ip}]^\top$ ,  $i = 1, \dots, n$  is an independent

instance of the random vector generated by an autoregressive process (see Hamilton, 1994) determined by

$$x_{i(j+1)} = \rho x_{ij} + \sqrt{1 - \rho^2} v_{ij}, j = 1, \dots, p - 1$$

with  $x_{i1} \sim \mathcal{N}(0, 1)$ ,  $v_{ij} \sim \mathcal{N}(0, 1)$  and  $\rho \in [0, 1]$  being the correlation parameter. The data labels  $\mathbf{y} \in \{0, 1\}^n$  are then drawn randomly according to the Bernoulli distribution with

$$\Pr\{y_i = 0 | \mathbf{x}_i\} = \frac{1}{1 + e^{\langle \mathbf{x}_i, \mathbf{z}^* \rangle}}, \quad i = 1, \dots, n.$$

**Example 3 (Real data)** Eight real data sets are taken into consideration: *arcene*<sup>1</sup>, *colon-cancer*<sup>2</sup>, *news20.binary*<sup>2</sup>, *newsgroup*<sup>3</sup>, *duke breast-cancer*<sup>2</sup>, *leukemia*<sup>2</sup>, *gisette*<sup>2</sup> and *rcv1.binary*<sup>2</sup>, which are summarized in following table, where only last four data sets have testing data. Moreover, for *arcene*, *colon-cancer*, *duke breast-cancer* and *leukemia*, sample-wise normalization has been conducted so that each sample has mean zero and variance one, and then feature-wise normalization has been conducted so that each feature has mean zero and variance one. For the rest three data sets, they are feature-wisely scaled to  $[-1, 1]$ . For *rcv1.binary*, we only use 20000 samples for the testing data. All  $-1$ s in the label classes  $\mathbf{y}$  are replaced by 0.

Table 2: Details of eight real datasets.

Data name	$n$ samples	$p$ features	training size $m_1$	testing size $m_2$
<i>arcene</i>	100	10,000	100	0
<i>colon-cancer</i>	62	2,000	62	0
<i>news20.binary</i>	19,996	1,355,191	19,996	0
<i>newsgroup</i>	11,314	777,811	11,314	0
<i>duke breast-cancer</i>	42	7,129	38	4
<i>leukemia</i>	72	7,129	38	34
<i>gisette</i>	7,000	5,000	6,000	1,000
<i>rcv1.binary</i>	697,641	47,236	20,242	20,000

## 5.2 Implementation

For parameter  $\tau$ , despite that Theorem 12 has given us a clue, it is still difficult to fix a proper one since  $\mathbf{z}^*$  is unknown. Alternative is to update  $\tau$  adaptively. By (14), it has

$$\tau^* < \frac{[\mathbf{z}^*]_s^\downarrow}{\max_{\{i: d_i^* \neq 0\}} |d_i^*|}, \quad \text{if } \|\mathbf{z}^*\|_0 = s.$$

<sup>1</sup><http://archive.ics.uci.edu/ml/index.php>

<sup>2</sup><https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/>

<sup>3</sup>[https://web.stanford.edu/~hastie/glmnet\\_matlab/](https://web.stanford.edu/~hastie/glmnet_matlab/)



Based on this, we would like to start  $\tau$  with a fixed one  $\tau_0 = 1$  and then decreasingly update it as following rule,

$$\tau_{k+1} = \begin{cases} 0.1\tau_k, & \text{if } \tau_k \geq \frac{[\mathbf{z}^k]_s^\downarrow}{\max_{j \in \bar{T}_k} |d_j^k|} \text{ and } \|F_{\tau_k}(\mathbf{u}^k; T_k)\| > 1/k, \\ \tau_k, & \text{otherwise,} \end{cases} \quad (34)$$

where  $T_k \in T^B(\mathbf{u}^k, \tau^k)$ . The reason of keeping  $\tau_{k+1} = \tau_k$  if  $\|F_{\tau_k}(\mathbf{u}^k; T_k)\| < 1/k$  is that this error tends to zero with order  $1/k$ , a desirable decreasing rate.

As for the initialization of NSLR, we set  $\mathbf{z}^0 = 0$  as what we described before. Due to updating  $\tau$  by  $\tau_k$ , the stop criteria is set as

$$\|F_{\tau_k}(\mathbf{u}^k; T_k)\| < \epsilon = 10^{-6}.$$

Notice that if the final  $\mathbf{u}^k$  satisfying  $\|F_{\tau_k}(\mathbf{u}^k; T_k)\| = 0$  and  $\nabla \ell(\mathbf{z}^k) = 0$ , then it is a global minimizer of (4) owing to  $\mathbf{z}^k \in S$  and Theorem 6 iii).

### 5.3 Benchmark methods

Since there are too many solvers being able to address the sparse logistic regression, we only focus on those programmed by Matlab. Solvers with codes being online unavailable or being written by other languages, such as R and C, are not selected for comparisons. We thus choose 7 solvers mentioned in Subsection 1.2, which should be enough to make comprehensive comparisons. We summarize them into following table.

Table 3: Benchmark Methods

Models	(3) with convex $\phi_\nu$	(3) with non-convex $\phi_\nu$	(4)
First order method	SLEP	APG, GIST	GraSP
Second order method	IRLS-LARS	--	NTGP, GPGN

For SLEP, we use it to solve (3) with  $\phi_\nu(\mathbf{z}) = \nu_1 \|\mathbf{z}\|_2^2 + \nu_2 \|\mathbf{z}\|_1$ , whilst IRLS-LARS aims to solve the  $\ell_1$  constrained logistic regression, namely, (5). APG and GIST are taken to solve the capped  $\ell_1$  logistic regression with  $\phi_\nu(\mathbf{z}) = \nu_3 \min(|z_i|, \nu_4)$ . We only use non-monotonous version of APG since its numerical performance was better than that of the monotonous version (see Li and Lin, 2015). For GraSP, the version chosen here is to solve (4) directly instead of (6). Notice that methods that aim at solving model (3) involve a penalty parameter  $\nu$ , whilst those tackling (4) need the sparsity level  $s$ . To make results comparable, we adjust their default parameters  $\nu$  for each method to guarantee the generated solution  $\mathbf{z}$  satisfying  $\|\mathbf{z}\|_0 \leq p/2$ . We will report the four indicators:

$$(\ell(\mathbf{z}), \|\nabla \ell(\mathbf{z})\|, \text{SER}, \text{Time})$$

to illustrate the performance of methods, where Time (in seconds) is the CPU time,  $\mathbf{z}$  is the solution obtained by each method and SER is the sign error rate defined by

$$\text{SER} := \frac{1}{n} \sum_{i=1}^m |y - \text{sign}(\langle \mathbf{x}_i, \mathbf{z} \rangle_+)|.$$

Here  $\text{sign}(a_+)$  is the sign of the projection of  $a$  onto a non-negative space, namely, it returns 1 if  $a > 0$  and 0 otherwise.

## 5.4 Numerical comparison

We now report the performance of eight methods on above three examples. To avoid randomness, we report average results over 10-time independent trails for the first two example since they are involving in randomly generated data.

**(a) Comparison on Example 1.** To observe the influence of the sparsity level  $s$  on four greedy methods: NSLR, GPGN, GraSP and NTGP, we fix  $p = 10000, n = p/5$  and alter  $s \in \{400, 600, \dots, 1600\}$ . As demonstrated in Figure 1, NSLR outperforms others in terms of the lowest  $\ell(\mathbf{z})$ ,  $\|\nabla \ell(\mathbf{z})\|$  and SER and the shortest time, followed by GPGN. By contrast, GraSP always performs the worst results, which means this first order method is not competitive when against with the other three methods, three second order methods.

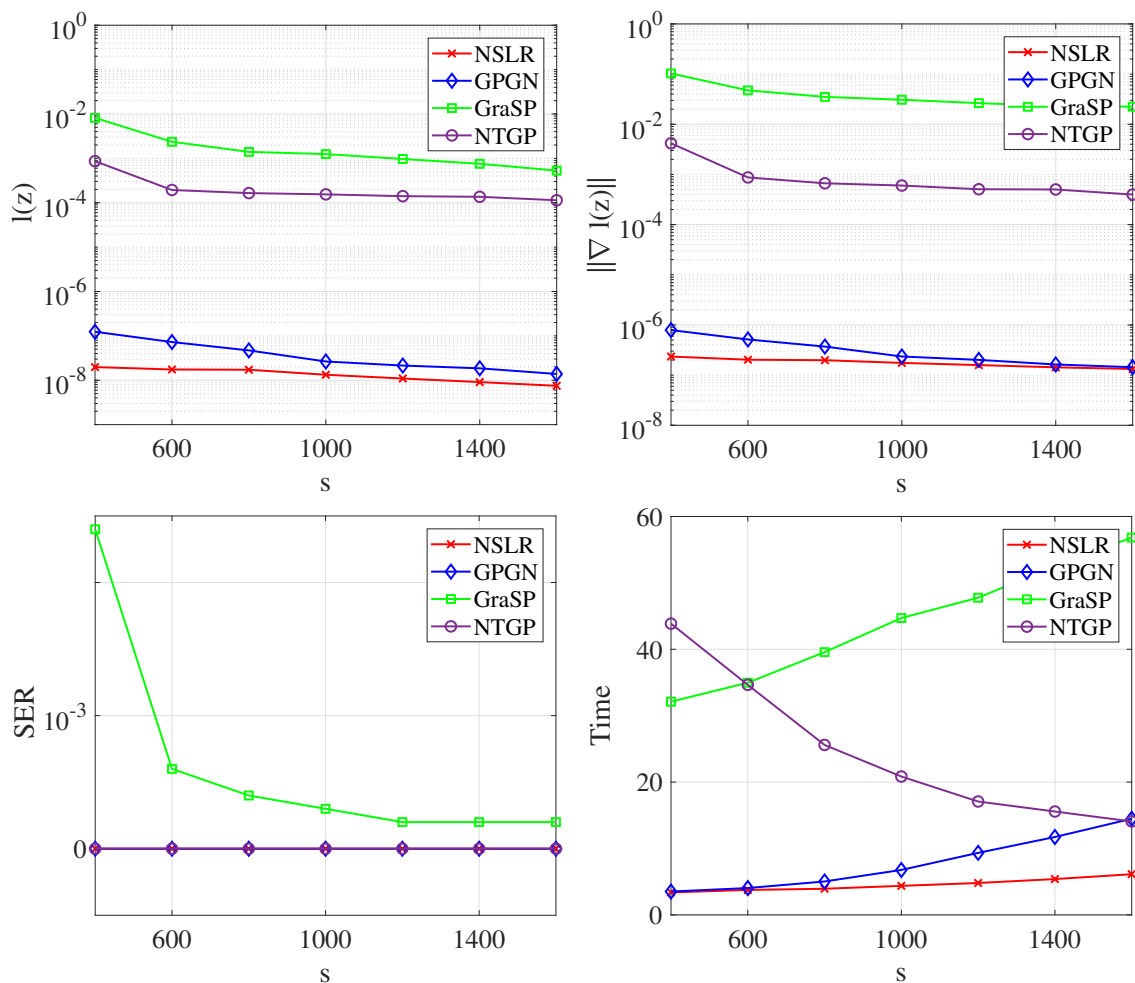


Figure 1: Comparison of four methods for Example 1 with  $p = 10000, n = 0.2p$

To observe the influence of the ratio of the sample size  $n$  and the number of features  $p$  on all eight methods, we fix  $p = 20000$ ,  $s = 0.1p$  and vary  $n/p \in \{0.1, 0.2, \dots, 0.7\}$ . Apart from recording the four indicators, we also report the number of non-zeros of the solution  $\mathbf{z}$  generated by each method. Here, for the LARS, we stop it when  $s$  variables are selected and their default stopping conditions are met since LARS only adds one variable at each iteration (see Huang et al. (2018)). We set  $\nu_1 = 10^{-1}, \nu_2 = 10^{-2}$  for SLEP,  $\nu_3 = 10^{-2}, \nu_4 = 10^{-4}$  for APG and  $\nu_3 = 10^{-3}\text{abs}(\text{randn}), \nu_4 = 10^{-5}\text{abs}(\text{randn})$  for GIST. As presented in Figure 2, in terms of Obj, Grad and SER, again NSLR performs the best results, followed by GPGN and GIST. It is obviously that LARS and SLEP produce undesirable results compared with other methods. For the computational time, NSLR runs the fastest, while GraSP and APG run relatively slow with over 1000 seconds when  $n/p \geq 0.6$ . Table 4 shows the sparsity levels  $\|\mathbf{z}\|_0$  only in LARS is lower than our NSLR. This is because LARS fails to recover the support and vanishes when  $s = 500$  in this numerical experiment (this phenomenon had also been observed in Huang et al. (2018) and Garg and Khandekar (2009).)

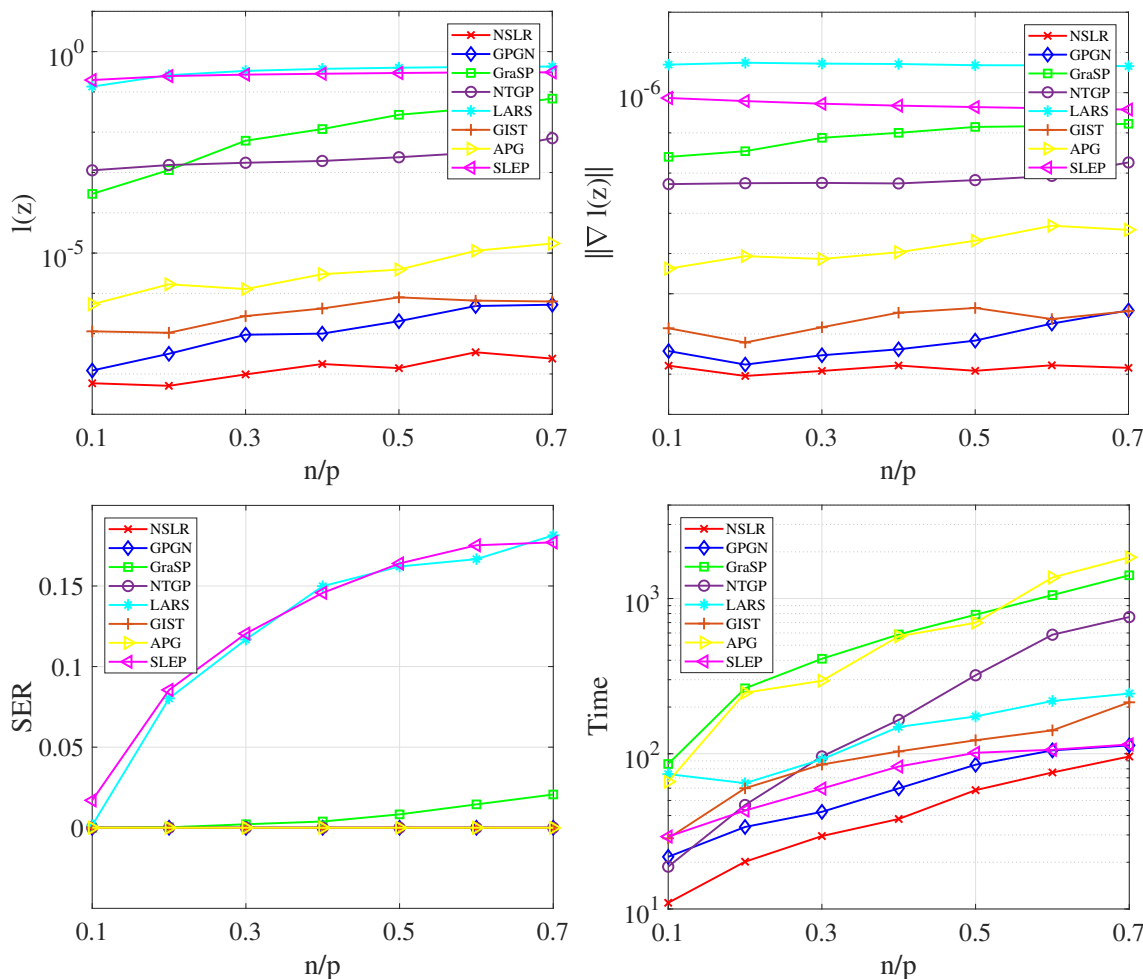


Figure 2: Comparison of eight methods for Example 1 with  $p = 20000$ ,  $s = 0.1p$ .

Table 4: Sparsity levels  $\|\mathbf{z}\|_0$  of eight methods for Example 1 with  $p = 20000, s = 0.1p$ .

$n/p$	0.1	0.2	0.3	0.4	0.5	0.6	0.7
NSLR, GPGN, GraSP, NTGP	2000	2000	2000	2000	2000	2000	2000
LARS	500	500	500	500	500	500	500
GIST	4403	4309	5274	7832	7913	8614	8904
APG	6138	5857	5720	6170	5574	5048	5043
SLEP	2076	2534	2980	3235	3498	3596	3873

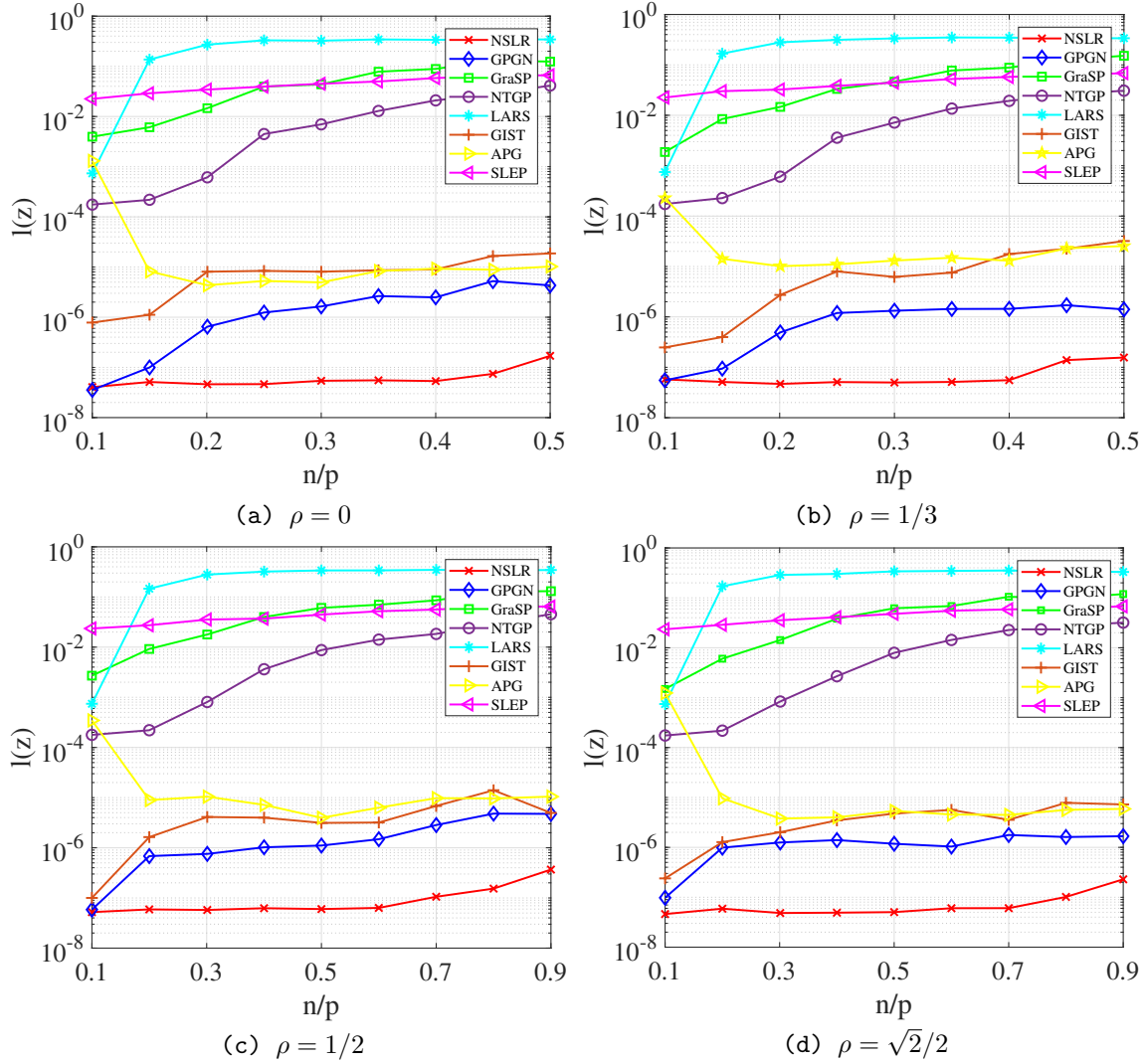


Figure 3: Comparison of eight methods for Example 2 with  $p = 1000, s = 0.1p$ .

**(b) Comparison on Example 2.** To observe the influence of the correlation parameter  $\rho$  on eight methods, we set  $p = 1000, s = 0.1p$  but choose  $\rho \in \{0, 1/3, 1/2, \sqrt{2}/2\}$ . Figure

3 shows the average  $\ell(\mathbf{z})$  gotten by eight methods for a wide range of the ratio  $n/p \in \{0.1, 0.2, \dots, 0.9\}$ . Apparently, at four different value of  $\rho$ , NSLR always performs stably best results. Moreover, the trends in these eight methods perform generally consistent, which indicates the correlation parameter has little influence on these methods. Therefore, we further fix  $\rho = 1/2$  and observe the performance of eight methods under higher dimensions.

Now we alter  $p \in \{10000, 20000, 30000\}$  with  $n = 0.2p, s = 0.05p$  or  $s = 0.1p$ . Here, we set  $\nu_1 = 10^{-2}, \nu_2 = 10^{-1}$  for SLEP  $\nu_3 = 10^{-2}, \nu_4 = 5 \times 10^{-4}$  for APG and  $\nu_3 = 5 \times 10^{-3} \text{abs}(\text{randn}), \nu_4 = 5 \times 10^{-5} \text{abs}(\text{randn})$  for GIST. As reported in Table 5, for cases of  $p = 20000, 30000$ , LARS basically fails to render desirable solutions due to the highest  $\ell(\mathbf{z})$ . It can be clearly seen that NSLR always provides the best accuracies with consuming shortest time. More importantly, its generated  $\|\nabla\ell(\mathbf{z})\|$  has accuracies with order of  $10^{-7}$ , which means global optimal solutions are achieved. By contrast, SLEP and LARS behave the worst due to highest  $\ell(\mathbf{z})$  and  $\|\nabla\ell(\mathbf{z})\|$  with order of  $10^{-1}$ . As for computational speed, NTGP and LARS seem to be the slowest, while NSLR and GPGN run the fastest.

Table 5: Average results of eight methods for Example 2.

	$s = 0.05p, n = 0.2p$					$s = 0.1p, n = 0.2p$				
	$\ell(\mathbf{z})$	$\ \nabla\ell(\mathbf{z})\ $	SER	Time(s)	$\ \mathbf{z}\ _0$	$\ell(\mathbf{z})$	$\ \nabla\ell(\mathbf{z})\ $	SER	Time(s)	$\ \mathbf{z}\ _0$
$p = 10000$										
NSLR	7.29e-8	2.39e-7	0.00e+0	3.22	500	2.43e-8	1.97e-7	0.00e+0	4.12	1000
GPGN	1.09e-7	5.49e-7	0.00e+0	4.19	500	6.84e-8	2.41e-7	0.00e+0	6.96	1000
GraSP	1.20e-2	1.32e-1	5.00e-3	17.10	500	7.41e-3	8.91e-2	1.70e-3	17.01	1000
NTGP	4.37e-3	1.70e-2	0.00e+0	33.54	500	2.18e-3	6.18e-3	0.00e+0	13.53	1000
LARS	1.43e-1	3.52e-1	1.25e-2	27.69	500	2.85e-1	5.74e-1	4.90e-3	71.80	1000
GIST	4.81e-6	2.99e-5	0.00e+0	12.54	2381	8.08e-6	6.27e-5	0.00e+0	11.40	2974
APG	1.29e-4	1.82e-3	0.00e+0	30.00	1407	1.25e-4	4.43e-3	0.00e+0	27.82	5211
SLEP	1.75e-1	3.89e-1	2.00e-3	6.56	811	1.32e-1	2.89e-1	0.00e+0	10.84	1019
$p = 20000$										
NSLR	7.33e-8	2.54e-7	0.00e+0	13.32	1000	2.66e-8	1.45e-7	0.00e+0	15.11	2000
GPGN	8.48e-8	4.15e-7	0.00e+0	16.47	1000	5.91e-8	2.55e-7	0.00e+0	15.83	2000
GraSP	1.29e-2	1.35e-1	5.25e-3	62.22	1000	5.00e-3	7.44e-2	1.50e-3	92.69	2000
NTGP	5.43e-3	2.15e-2	0.00e+0	134.65	1000	2.30e-3	6.68e-3	0.00e+0	54.50	2000
LARS	3.96e-1	7.75e-1	5.43e-2	107.45	1000	4.21e-1	8.09e-1	6.18e-2	117.1	1000
GIST	7.20e-7	4.74e-6	0.00e+0	33.52	1542	9.98e-7	4.60e-6	0.00e+0	54.55	4137
APG	5.06e-5	2.09e-3	0.00e+0	24.94	1849	6.04e-5	2.99e-3	0.00e+0	69.46	4323
SLEP	1.88e-1	4.08e-1	3.25e-3	22.67	1511	1.45e-1	3.18e-1	0.00e+0	34.59	2005
$p = 30000$										
NSLR	7.12e-8	2.27e-7	0.00e+0	33.98	1500	3.14e-8	1.66e-7	0.00e+0	49.11	3000
GPGN	8.58e-8	4.65e-7	0.00e+0	35.06	1500	4.09e-8	2.44e-7	0.00e+0	51.49	3000
GraSP	2.28e-2	1.76e-1	7.50e-3	181.95	1500	8.96e-3	1.07e-1	2.83e-3	208.12	3000
NTGP	5.36e-3	2.09e-2	0.00e+0	307.17	1500	2.32e-3	6.76e-3	0.00e+0	125.14	3000
LARS	4.59e-1	8.70e-1	1.03e-1	205.79	1000	4.99e-1	9.25e-1	1.21e-1	206.76	1000
GIST	2.76e-6	1.12e-5	0.00e+0	121.16	6124	7.97e-7	4.25e-6	0.00e+0	150.95	6278
APG	9.40e-4	2.82e-2	1.67e-4	206.78	7460	3.11e-4	7.03e-3	0.00e+0	247.16	6450
SLEP	1.96e-1	4.21e-1	2.33e-3	54.20	2392	1.49e-1	3.23e-1	1.67e-4	80.19	3009

**(c) Comparison on Example 3.** To observe the performance for above all eight methods on real data sets, we select eight real data sets with different dimensions. The highest dimension is up to millions (see `news20.binary`). Table 6 reports results for eight methods on four datasets without testing data ( $m_2 = 0$ ): `arcene`, `colon-cancer`, `news20.binary` and `newsgroup`. For the last two datasets, LARS makes our desktop run out of memory, thus its results are omitted here. NaN obtained by GraSP on data `newsgroup` means  $\|\nabla\ell(\mathbf{z})\|$  tends to infinity. Clearly, NSLR is more efficient than others for all test instances. For example, NSLR only uses 9 seconds for data `newsgroup` with  $p = 777811$  features and achieves the smallest logistic loss with the sparsest solution.

Table 6: Results of eight methods for Example 3 with  $m_2 = 0$ .

Data	$\ell(\mathbf{z})$	$\ \nabla\ell(\mathbf{z})\ $	SER	Time(s)	$\ \mathbf{z}\ _0$	$\ell(\mathbf{z})$	$\ \nabla\ell(\mathbf{z})\ $	SER	Time(s)	$\ \mathbf{z}\ _0$
	Arcene: $m_1 = 100, p = 10000$					colon-cancer: $m_1 = 62, p = 2000$				
NSLR	6.58e-8	1.31e-6	0.00e+0	0.63	60	1.50e-7	1.29e-6	0.00e+0	0.13	20
GPGN	1.84e-5	2.72e-4	0.00e+0	0.58	60	3.35e-5	3.51e-4	0.00e+0	0.17	20
GraSP	1.88e-3	1.78e-1	0.00e+0	1.23	60	3.03e-1	7.00e-1	1.61e-2	0.26	20
NTGP	1.09e-1	1.34e+0	1.00e-2	2.81	60	4.51e-3	3.60e-2	0.00e+0	0.19	20
LARS	7.98e-2	7.68e-1	0.00e+0	0.97	60	2.84e-2	1.60e-1	0.00e+0	0.23	30
GIST	3.99e-7	6.12e-6	0.00e+0	5.64	134	6.22e-7	4.09e-6	0.00e+0	0.13	41
APG	1.84e-5	9.53e-4	0.00e+0	3.34	430	1.79e-7	3.66e-6	0.00e+0	0.23	20
SLEP	1.05e-1	9.58e-1	0.00e+0	3.66	66	1.68e-1	6.80e-1	4.84e-2	0.14	23
Data	news20.binary: $m_1 = 19996, p = 1355191$					newsgroup: $m_1 = 11314, p = 777811$				
NSLR	1.82e-2	4.24e-3	4.80e-3	20.43	2500	2.20e-2	6.52e-3	9.90e-3	9.55	3000
GPGN	2.94e-2	6.10e-3	8.00e-3	25.44	2500	5.17e-2	7.90e-3	1.20e-2	30.71	3000
GraSP	2.46e-2	1.00e-2	9.25e-3	205.7	2500	5.08e-1	NaN	4.75e-2	33.06	3000
NTGP	7.94e-2	1.38e-2	6.55e-3	93.20	2500	3.11e-1	4.77e-2	5.80e-2	13.07	3000
LARS	—	—	—	—	—	—	—	—	—	—
GIST	3.18e-2	9.98e-3	6.60e-3	27.57	4091	6.84e-2	9.70e-3	1.52e-2	10.55	3017
APG	4.18e-2	8.10e-3	1.24e-2	37.15	3869	5.12e-2	7.70e-3	1.77e-2	10.62	3257
SLEP	1.49e-1	3.19e-2	3.34e-2	43.70	5299	2.60e-1	4.15e-2	4.60e-2	23.89	5138

When all methods solve datasets with testing data ( $m_2 > 0$ ): `duke breast-cancer`, `leukemia`, `gisette` and `rcv1.binary`, the table is a little different. Since the testing data is taken into consideration, we add two indicators to illustrate the performance of each method:  $\ell(\mathbf{z})$ -test and SER-test. Results are reported in Table 7, where  $\ell(\mathbf{z})$  and  $\ell(\mathbf{z})$ -test denote the objective function value on training data and testing data respectively, and similar to SER-train and SER-test. For cases of `duke breast-cancer` and `leukemia`, LARS stops when the maximum number of iterations reaches the  $\min\{m_1, p\}$ . Hence its produced  $\|\mathbf{z}\|_0$ s are less than other methods. We can see that NSLR could guarantee a good performance on the testing data as well as the training data. It is capable of rendering sparsest solutions with smallest logistic loss and consuming least time.

Table 7: Average results of eight methods for Example 3 with  $m_2 > 0$ .

	$\ \nabla\ell(\mathbf{z})\ $	$\ell(\mathbf{z})$	$\ell(\mathbf{z})$ -test	SER-train	SER-test	Time(s)	$\ \mathbf{z}\ _0$
Data	<b>duke breast-cancer:</b> $m_1 = 38, m_2 = 4, p = 7129$						
NSLR	1.53e-10	3.16e-11	1.72e-7	0.00e+0	0.00e+0	0.39	100
GPGN	9.73e-5	1.31e-5	2.55e-3	0.00e+0	0.00e+0	0.44	100
GraSP	8.76e-2	3.57e-3	1.82e-3	0.00e+0	0.00e+0	0.50	100
NTGP	8.76e-2	1.21e-5	1.20e-4	0.00e+0	0.00e+0	0.64	100
LARS	5.12e-4	1.01e-4	1.21e-4	0.00e+0	0.00e+0	0.61	37
GIST	5.78e-8	6.64e-9	1.60e-1	0.00e+0	0.00e+0	0.54	614
APG	1.12e-5	6.35e-7	2.68e-7	0.00e+0	0.00e+0	0.74	136
SLEP	9.14e-3	1.93e-3	1.04e-2	0.00e+0	0.00e+0	0.43	203
Data	<b>leukemia:</b> $m_1 = 38, m_2 = 34, p = 7129$						
NSLR	2.81e-5	1.54e-6	2.66e-1	0.00e+0	2.94e-2	0.30	150
GPGN	1.11e-3	1.29e-5	2.71e+0	0.00e+0	1.47e-1	0.35	150
GraSP	2.68e-1	3.40e-3	5.08e-1	0.00e+0	1.47e-1	0.56	150
NTGP	2.68e-1	4.22e-4	2.11e-1	0.00e+0	5.88e-2	0.75	150
LARS	8.09e-3	6.07e-4	1.11e-1	0.00e+0	8.82e-2	1.05	37
GIST	2.65e-1	5.68e-3	1.82e-1	0.00e+0	1.18e-1	0.42	295
APG	6.99e-3	1.05e-4	3.63e-1	0.00e+0	8.82e-2	0.54	1066
SLEP	2.00e+0	1.71e-1	2.85e-1	0.00e+0	2.94e-2	0.58	269
Data	<b>gisette:</b> $m_1 = 6000, m_2 = 1000, p = 5000$						
NSLR	2.43e-4	1.55e-4	1.91e-1	0.00e+0	4.18e-2	1.40	500
GPGN	3.25e-4	2.25e-4	2.63e-1	0.00e+0	4.60e-2	1.41	500
GraSP	1.23e-3	1.51e-4	2.30e-1	0.00e+0	4.82e-2	2.43	500
NTGP	1.23e-3	1.17e-3	9.58e-1	0.00e+0	4.18e-2	2.55	500
LARS	3.27e-1	1.63e-1	1.39e+0	1.00e-3	4.18e-2	8.26	500
GIST	4.02e-4	2.40e-4	1.02e+0	0.00e+0	4.30e-2	1.78	1303
APG	1.12e-3	3.72e-4	1.31e+0	0.00e+0	4.90e-2	1.64	907
SLEP	1.49e-1	1.92e-2	8.30e-1	0.00e+0	4.97e-2	2.04	1569
Data	<b>rcv1.binary:</b> $m_1 = 20242, m_2 = 20000, p = 47236$						
NSLR	2.99e-3	6.62e-2	1.21e-1	2.10e-2	4.33e-2	3.71	1000
GPGN	1.73e-3	2.90e-2	2.35e-1	8.05e-3	5.58e-2	6.81	1000
GraSP	9.07e-3	3.09e-1	1.98e+0	4.15e-2	9.51e-2	21.7	1000
NTGP	5.30e-3	7.48e-2	1.37e-1	1.06e-2	4.64e-2	4.47	1000
LARS	2.96e-2	2.27e-1	2.61e-1	5.13e-1	5.46e-1	33.5	1000
GIST	4.43e-3	3.44e-2	1.41e-1	8.20e-3	4.85e-2	4.57	1545
APG	1.54e-3	2.51e-2	2.78e-1	7.31e-3	6.00e-2	7.57	1537
SLEP	1.52e-2	1.24e-1	1.65e-1	3.02e-2	5.23e-2	8.45	3527

## 6. Conclusion

In this paper, we considered the sparsity constrained logistic regression (4), which is a non-linear combinatorial problem. Despite the NP-hardness, we benefited from its nice properties of objective function and introduced the strong  $\tau$ -stationary point as a optimality condition of (4). This contributed to a stationary equation, which can be efficiently solved by Newton

method. It turned out that our proposed method NSLR has a relatively low computational complexity in each step because only a small linear equation system with  $s$  variables and  $s$  equations need to be solved to update the Newton direction. We further established its quadratic convergence to a strong  $\tau$ -stationary point and showed termination within finite steps. It is worth mentioning that we reasonably extended the classical Newton method for solving unconstrained problems to sparsity constrained logistic regression. The numerical performance against several state-of-the-art methods demonstrated that NSLR is remarkably efficient and competitive in solving (4), especially in large scale settings.

A drawback of NSLR is that the starting point is required to be close enough to its limit to establish the quadratic convergence results. A potential way to compensate this is to adopt a line search scheme (see Yuan and Liu (2017) for example), which would guarantee that the generated sequence by NSLR converges to its limit, because of this, the closeness to the limit of the starting point is no more demanded. We leave this as a future research.

## Acknowledgments

We would like to acknowledge support for this project from the National Natural Science Foundation of China (11431002) and the 111 project of China (B16002). We are also very grateful to Prof. Xiao-Tong Yuan for sharing with us their excellent package NTGP.

## Appendix A.

In this appendix we prove all associated theorems from Section 2:

### Proof of Property 2

**Proof** i) Denote  $\ell_i(\mathbf{z}) = \log(1 + e^{t_i}) - y_i t_i$  with  $t_i = \langle \mathbf{x}_i, \mathbf{z} \rangle$ . By Taylor expansion of  $\ell_i(\mathbf{z})$  at  $t_i = 0$ , for any  $\min\{0, t_i\} \leq \varsigma_i \leq \max\{0, t_i\}$ , it holds

$$\begin{aligned}
\ell(\mathbf{z}) &= \frac{1}{n} \sum_{i=1}^n \ell_i(\mathbf{z}) = \frac{1}{n} \sum_{i=1}^n [\ln(2) + t_i/2 + t_i^2/(2(2 + e^{\varsigma_i} + e^{-\varsigma_i})) - y_i t_i] \\
&\leq \frac{1}{n} \sum_{i=1}^n [\ln(2) + t_i/2 + t_i^2/(2 + e^{\varsigma_i} + e^{-\varsigma_i}) - y_i t_i] \\
&\leq \frac{1}{n} \sum_{i=1}^n [\ln(2) + (1/2 - y_i)t_i + t_i^2/4] \\
&= \frac{1}{n} \sum_{i=1}^n [\ln(2) + [t_i - (2y_i - 1)]^2/4 - (2y_i - 1)^2/4] \\
&= \ln(2) + (\|X\mathbf{z} - \mathbf{c}\|^2 - \|\mathbf{c}\|^2)/(4n) \\
&= \ln(2) - 1/4 + \|X\mathbf{z} - \mathbf{c}\|^2/(4n), \tag{35}
\end{aligned}$$

where the last equality is due to  $y_i \in \{0, 1\}$  and thus  $2y_i - 1 \in \{-1, 1\}$ . Then for any  $\mathbf{z} \in S$ , we must have  $\min_{\mathbf{z} \in S} \ell(\mathbf{z}) \leq \ell(\mathbf{z})$ , which results in the conclusion.



ii) If  $X^\top \mathbf{c} \neq 0$ , the first claim holds clearly (In fact, a typical case is when  $r = 1$ ). This indicates  $U^\top \mathbf{c} \neq 0$ . Direct calculation yields that

$$\begin{aligned} \|X\bar{\mathbf{z}} - \mathbf{c}\|^2 &= \|X_T V \Lambda^{-1} U^\top \mathbf{c} - \mathbf{c}\|^2 = \|U \Lambda V^\top V \Lambda^{-1} U^\top \mathbf{c} - \mathbf{c}\|^2 \\ &= \|U U^\top \mathbf{c} - \mathbf{c}\|^2 = \|U U^\top \mathbf{c}\|^2 - 2\langle U U^\top \mathbf{c}, \mathbf{c} \rangle + \|\mathbf{c}\|^2 \\ &= \|\mathbf{c}\|^2 - \|U^\top \mathbf{c}\|^2 \quad (\text{because of } U^\top U = \mathcal{I}) \\ &< \|\mathbf{c}\|^2 = n, \end{aligned}$$

which together with (35) concludes the the second claim. If  $X^\top \mathbf{c} = 0$ , then  $\nabla \ell(0) = X^\top (h(0) - \mathbf{y})/n = -2X^\top \mathbf{c}/n = 0$ . This together with Property 1 i) suffices to

$$\ell(\mathbf{z}) \geq \ell(0) + \langle \nabla \ell(0), \mathbf{z} - 0 \rangle = \ell(0)$$

for any  $\mathbf{z} \in S$ , which means 0 is a global optimal solution. And one can easily check that  $0 \in \operatorname{argmin}_{\mathbf{z} \in S} \|X\mathbf{z} - \mathbf{c}\|^2$  due to  $X^\top (X0 - \mathbf{c}) = 0$ .  $\blacksquare$

### Proof of Theorem 6

**Proof** i) We argue by contradiction and suppose that for any given  $\tau \in (0, 1/\lambda_x)$ ,  $\mathbf{z}^*$  is not a strong  $\tau$ -stationary point. Then there exists a  $\mathbf{z} \neq \mathbf{z}^* \in \mathbb{R}^p$  satisfies

$$\mathbf{z} = P_S(\mathbf{z}^* - \tau \nabla \ell(\mathbf{z}^*)),$$

which together with the definition of projection on  $S$  imply that

$$\|\mathbf{z} - (\mathbf{z}^* - \tau \nabla \ell(\mathbf{z}^*))\|^2 \leq \|\mathbf{z}^* - (\mathbf{z}^* - \tau \nabla \ell(\mathbf{z}^*))\|^2.$$

This leads to

$$\langle \nabla \ell(\mathbf{z}^*), \mathbf{z} - \mathbf{z}^* \rangle \leq -1/(2\tau) \|\mathbf{z} - \mathbf{z}^*\|^2. \quad (36)$$

It follows from Property 1 (i) and (36) that for any  $\tau \in (0, 1/\lambda_x)$ ,

$$\begin{aligned} \ell(\mathbf{z}) &\leq \ell(\mathbf{z}^*) + \langle \nabla \ell(\mathbf{z}^*), \mathbf{z} - \mathbf{z}^* \rangle + (\lambda_x/2) \|\mathbf{z} - \mathbf{z}^*\|^2 \\ &\leq \ell(\mathbf{z}^*) - 1/(2\tau) \|\mathbf{z} - \mathbf{z}^*\|^2 + (\lambda_x/2) \|\mathbf{z} - \mathbf{z}^*\|^2 < \ell(\mathbf{z}^*) \end{aligned}$$

which contradicts the assumption that  $\mathbf{z}^*$  is a global minimizer of SLR (4).

ii) For a given  $\tau > 0$ , let  $\mathbf{z}^*$  be a strong  $\tau$ -stationary point with  $\|\mathbf{z}^*\|_0 = s$ . For any  $\mathbf{z} \in N(\mathbf{z}^*, \delta) \cap S$  with  $0 < \delta < [\mathbf{z}]_s^\downarrow$ , we have

$$|\mathbf{z}_i| = |\mathbf{z}_i^* - (\mathbf{z}_i^* - \mathbf{z}_i)| \geq |\mathbf{z}_i^*| - \delta > [\mathbf{z}]_s^\downarrow - [\mathbf{z}]_s^\downarrow = 0, \quad \forall i \in \operatorname{supp}(\mathbf{z}^*),$$

which means  $\operatorname{supp}(\mathbf{z}^*) \subseteq \operatorname{supp}(\mathbf{z})$ . By  $\|\mathbf{z}\|_0 \leq s$  and  $|\operatorname{supp}(\mathbf{z}^*)| = \|\mathbf{z}^*\|_0 = s$ , we obtain

$$\operatorname{supp}(\mathbf{z}^*) = \operatorname{supp}(\mathbf{z}), \quad \forall \mathbf{z} \in N(\mathbf{z}^*, \delta) \cap S. \quad (37)$$

This implies that  $\mathbf{z}_i = \mathbf{z}_i^* = 0$  for  $i \notin \text{supp}(\mathbf{z}^*)$ . It follows that the convexity of  $\ell$  and Property 4 ii) ( $\nabla_i \ell(\mathbf{z}^*) = 0, i \in \text{supp}(\mathbf{z}^*)$ )

$$\begin{aligned} \ell(\mathbf{z}) &\geq \ell(\mathbf{z}^*) + \langle \nabla \ell(\mathbf{z}^*), \mathbf{z} - \mathbf{z}^* \rangle \\ &= \ell(\mathbf{z}^*) + \sum_{i \in \text{supp}(\mathbf{z}^*)} \nabla_i \ell(\mathbf{z}^*)(\mathbf{z}_i - \mathbf{z}_i^*) + \sum_{i \notin \text{supp}(\mathbf{z}^*)} \nabla_i \ell(\mathbf{z}^*)(\mathbf{z}_i - \mathbf{z}_i^*) = \ell(\mathbf{z}^*). \end{aligned}$$

Thus  $\mathbf{z}^*$  is a local minimizer of SLR (4). If  $\mathbf{z}^*$  further satisfies  $\ell(\mathbf{z}^*) = 0$ , then for any  $\mathbf{z} \in S$ , the convexity of  $\ell$  suffices to

$$\ell(\mathbf{z}) \geq \ell(\mathbf{z}^*) + \langle \nabla \ell(\mathbf{z}^*), \mathbf{z} - \mathbf{z}^* \rangle = \ell(\mathbf{z}^*), \quad (38)$$

which proves the global optimality of  $\mathbf{z}^*$ .

iii) When  $\|\mathbf{z}^*\|_0 < s$ , Property 4 i) guarantees  $\nabla \ell(\mathbf{z}^*) = 0$ , which together with (38) yields the conclusion.  $\blacksquare$

### Proof of Theorem 7

**Proof** We only prove the ‘if’ part since the proof of the ‘only if’ part is quite straightforward. Suppose we have  $F_\tau(\mathbf{u}; T) = 0$  for all  $T \in T^B(\mathbf{u}; \tau)$ , namely,

$$\mathbf{z}_{\overline{T}} = 0, \quad \mathbf{d}_T = 0, \quad \mathbf{d} = \nabla \ell(\mathbf{z}) \quad (39)$$

If  $T^B(\mathbf{u}; \tau)$  is a singleton, by letting  $T$  be the only entry of  $T^B(\mathbf{u}; \tau)$ , then

$$\mathbf{z} - \text{P}_S(\mathbf{z} - \tau \nabla \ell(\mathbf{z})) \stackrel{(39)}{=} \mathbf{z} - \text{P}_S(\mathbf{z} - \tau \mathbf{d}) = \begin{bmatrix} \mathbf{z}_T \\ \mathbf{z}_{\overline{T}} \end{bmatrix} - \begin{bmatrix} (\mathbf{z} - \tau \mathbf{d})_T \\ 0 \end{bmatrix} \stackrel{(39)}{=} \begin{bmatrix} \mathbf{z}_T - \mathbf{z}_T \\ 0 - 0 \end{bmatrix} = 0,$$

which means  $\mathbf{z}$  is a strong  $\tau$ -stationary point. If  $T^B(\mathbf{u}; \tau)$  has multiple elements, then by the definition (17) of  $T(\mathbf{u}; \tau)$  we have two claims:  $[\mathbf{z} - \tau \mathbf{d}]_s^\downarrow = [\mathbf{z} - \tau \mathbf{d}]_{s+1}^\downarrow > 0$  or  $[\mathbf{z} - \tau \mathbf{d}]_s^\downarrow = 0$ . Now we exclude the first claim. Without loss of any generality, we assume  $|z_1 - \tau d_1| \geq \dots \geq |z_s - \tau d_s| = |z_{s+1} - \tau d_{s+1}| = [\mathbf{z} - \tau \mathbf{d}]_s^\downarrow$ . Let  $T_1 = \{1, 2, \dots, s\}$  and  $T_2 = \{1, 2, \dots, s-1, s+1\}$ . Then  $F_\tau(\mathbf{u}; T_1) = F_\tau(\mathbf{u}; T_2) = 0$  imply that  $\mathbf{d}_{T_1} = \mathbf{d}_{T_2} = 0$  and  $\mathbf{z}_{\overline{T_1}} = \mathbf{z}_{\overline{T_2}} = 0$ , which lead to

$$|z_1| = |z_1 - \tau d_1| \geq \dots \geq |z_s| = |z_s - \tau d_s| = |z_{s+1}| = |z_{s+1} - \tau d_{s+1}| = [\mathbf{z} - \tau \mathbf{d}]_s^\downarrow > 0.$$

This is contradicted with  $\mathbf{z}_{\overline{T_1}} = 0$  because of  $s+1 \in \overline{T_1}$ . Therefore, we have  $[\mathbf{z} - \tau \mathbf{d}]_s^\downarrow = 0$ . This together with the definition (17) of  $T(\mathbf{z}; \tau)$  derives  $0 = [\mathbf{z} - \tau \mathbf{d}]_s^\downarrow \geq |z_i - \tau d_i| = |\tau d_i|$  for any  $i \in \overline{T_1}$ , which combining  $\mathbf{d}_{T_1} = 0$  renders  $\mathbf{d} = 0$  and hence  $\ell(\mathbf{z}) = 0$  from (39). Moreover,  $[\mathbf{z}]_s^\downarrow = [\mathbf{z} - \tau \mathbf{d}]_s^\downarrow = 0$ , yielding  $\|\mathbf{z}\|_0 < s$ . Consequently,  $\mathbf{z}$  is a strong  $\tau$ -stationary point from Property 4.  $\blacksquare$

## Proof of Theorem 9

**Proof** By using the elementary transformation of the matrix  $\nabla F_\tau(\mathbf{u})$ , we have

$$\nabla F_\tau(\mathbf{u}; T) \xrightarrow[\text{operation}]{\text{row}} \begin{bmatrix} 0 & 0 & \mathcal{I}_s & 0 \\ 0 & \mathcal{I}_r & 0 & 0 \\ -\nabla_{T,T}^2 \ell(\mathbf{z}) & 0 & 0 & 0 \\ -\nabla_{T,T}^2 \ell(\mathbf{z}) & 0 & 0 & \mathcal{I}_r \end{bmatrix} \xrightarrow[\text{operation}]{\text{column}} \begin{bmatrix} 0 & 0 & \mathcal{I}_s & 0 \\ 0 & \mathcal{I}_r & 0 & 0 \\ -\nabla_{T,T}^2 \ell(\mathbf{z}) & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathcal{I}_r \end{bmatrix} \xrightarrow[\text{operation}]{\text{column}} \begin{bmatrix} \mathcal{I}_s & 0 & 0 & 0 \\ 0 & \mathcal{I}_r & 0 & 0 \\ 0 & 0 & -\nabla_{T,T}^2 \ell(\mathbf{z}) & 0 \\ 0 & 0 & 0 & \mathcal{I}_r \end{bmatrix}$$

Clearly, we have

$$\text{rank}(\nabla F_\tau(\mathbf{u}; T)) = \text{rank}(\nabla_{T,T}^2 \ell(\mathbf{z})) + 2p - s = \text{rank}(X_T^\top D(\mathbf{z})X_T) + 2p - s,$$

If the matrix  $X$  is  $s$ -regular, then for  $\forall \mathbf{h} \neq 0$  we have  $X_T \mathbf{h} \neq 0$  and  $(X_T \mathbf{h})^\top D(\mathbf{z})(X_T \mathbf{h}) > 0$  by  $D(\mathbf{z}) \succ 0$ . This shows  $X_T^\top D(\mathbf{z})X_T \succ 0$ , yielding  $\text{rank}(X_T^\top D(\mathbf{z})X_T) = s$ .

Direct calculation yields the form of  $(\nabla F_\tau(\mathbf{u}; T))^{-1}$  as (23). It remains to verify the last statement of the theorem. For any given  $\mathbf{z}^* \in \mathbb{R}^p$  and  $\delta > 0$ ,  $\mathbf{z} \in N(\mathbf{z}^*, \delta)$  implies that  $\|\mathbf{z}\| < \|\mathbf{z}^*\| + \delta$  and hence

$$\|X\mathbf{z}\| \leq (\lambda_{\max}(X^\top X))^{1/2} \|\mathbf{z}\| = 2\lambda_x^{1/2} \|\mathbf{z}\| < 2\lambda_x^{1/2} (\|\mathbf{z}^*\| + \delta) =: \sigma,$$

which leads to

$$\langle \mathbf{x}_i, \mathbf{z} \rangle \leq \|X\mathbf{z}\| < \sigma, \quad i = 1, \dots, n.$$

This together with Property 1 (iii) suffices to that for any  $\mathbf{z} \in N(\mathbf{z}^*, \delta)$ ,

$$1/(4e^\sigma) \leq e^\sigma / (1 + e^\sigma)^2 \leq \lambda_{\min}(D(\mathbf{z})) \leq \lambda_{\max}(D(\mathbf{z})) \leq 1/4. \quad (40)$$

We are ready to see that

$$\begin{aligned} \|(\nabla_{T,T}^2 \ell(\mathbf{z}))^{-1}\|_2 &= \lambda_{\max}((\nabla_{T,T}^2 \ell(\mathbf{z}))^{-1}) \\ &= [\lambda_{\min}(\nabla_{T,T}^2 \ell(\mathbf{z}))]^{-1} = [\lambda_{\min}(X_T^\top D(\mathbf{z})X_T/n)]^{-1} \\ &\leq n [\lambda_{\min}(X_T^\top X_T) \lambda_{\min}(D(\mathbf{z}))]^{-1} \\ &\stackrel{(22,40)}{\leq} 4ne^\sigma/\underline{\lambda}, \end{aligned}$$

which completes the whole proof. ■

## Appendix B.

In this appendix we prove results from Section 4:

### Proof of Theorem 10

**Proof** Since  $\mathbf{u}^k \rightarrow \mathbf{u}^\infty$ , it holds  $\mathbf{z}^k \rightarrow \mathbf{z}^\infty$  and  $\mathbf{d}^k \rightarrow \mathbf{d}^\infty$  as  $k \rightarrow \infty$ , which suffices to  $\mathbf{z}^\infty \in S$  due to  $\|\mathbf{z}^k\|_0 \leq s$ . Clearly,  $\{T_k\}$  is bounded owing to  $T_k \subseteq N_p$  and  $|T_k| = s$ , and thus has a subsequence  $\{T_{k_t}\}$  such that

$$T_{k_t} = T_{k_{t+1}} = \dots =: T_\infty. \quad (41)$$

Since  $\mathbf{z}^{k_{t+1}} \rightarrow \mathbf{z}^\infty$ ,  $\text{supp}(\mathbf{z}^{k_{t+1}}) \subseteq T_{k_t} = T_\infty$ , we must have

$$T_\infty \begin{cases} = \text{supp}(\mathbf{z}^\infty), & \text{if } \|\mathbf{z}^\infty\|_0 = s, \\ \supset \text{supp}(\mathbf{z}^\infty), & \text{if } \|\mathbf{z}^\infty\|_0 < s. \end{cases} \quad \text{and hence } \mathbf{z}_{\overline{T}_\infty}^\infty = 0. \quad (42)$$

From (27) and (28), we obtain

$$\begin{aligned} \nabla_{T_\infty}^2 \ell(\mathbf{z}^\infty) \mathbf{z}^\infty &= \lim_{k_t \rightarrow \infty} \nabla_{T_{k_t}}^2 \ell(\mathbf{z}^{k_t}) \mathbf{z}^{k_{t+1}} = \lim_{k_t \rightarrow \infty} \left[ \nabla_{T_{k_t}}^2 \ell(\mathbf{z}^{k_t}) \mathbf{z}^{k_t} - \nabla_{T_{k_t}} \ell(\mathbf{z}^{k_t}) \right] \\ &= \nabla_{T_\infty}^2 \ell(\mathbf{z}^\infty) \mathbf{z}^\infty - \nabla_{T_\infty} \ell(\mathbf{z}^\infty), \end{aligned}$$

which yields that

$$\nabla_{T_\infty} \ell(\mathbf{z}^\infty) = 0. \quad (43)$$

Again by (28), we have

$$\mathbf{d}_{T_\infty}^\infty = \lim_{k_t \rightarrow \infty} \mathbf{d}_{T_{k_t}}^{k_{t+1}} = 0, \quad (44)$$

$$\begin{aligned} \mathbf{d}_{\overline{T}_\infty}^\infty &= \lim_{k_t \rightarrow \infty} \mathbf{d}_{\overline{T}_{k_t}}^{k_{t+1}} = \lim_{k_t \rightarrow \infty} \left[ \nabla_{\overline{T}_{k_t}} \ell(\mathbf{z}^{k_t}) + \nabla_{\overline{T}_{k_t}}^2 \ell(\mathbf{z}^{k_t}) (\mathbf{z}^{k_{t+1}} - \mathbf{z}^{k_t}) \right] \\ &= \nabla_{\overline{T}_\infty} \ell(\mathbf{z}^\infty). \end{aligned} \quad (45)$$

Now for any  $i \in T_{k_t} \stackrel{(41)}{=} T_\infty$ ,  $j \in \overline{T}_{k_t} \stackrel{(41)}{=} \overline{T}_\infty$ , we have

$$\begin{aligned} |z_i^\infty| &\stackrel{(44)}{=} |z_i^\infty - \tau d_i^\infty| = \lim_{k_t \rightarrow \infty} |z_i^{k_t} - \tau d_i^{k_t}| \\ &\stackrel{(17)}{\geq} \lim_{k_t \rightarrow \infty} |z_j^{k_t} - \tau d_j^{k_t}| = |z_j^\infty - \tau d_j^\infty| \stackrel{(42)}{=} \tau |d_j^\infty|, \end{aligned}$$

which leads to

$$[\mathbf{z}^\infty]_s^\downarrow = \min_{i \in T_\infty} |z_i^\infty| \geq \tau |d_j^\infty| \stackrel{(45)}{=} \tau |\nabla_j \ell(\mathbf{z}^\infty)|, \quad \forall j \in \overline{T}_\infty, \quad (46)$$

If  $\|\mathbf{z}^\infty\|_0 = s$ , then  $T_\infty \stackrel{(42)}{=} \text{supp}(\mathbf{z}^\infty)$ . Consequently,  $[\mathbf{z}^\infty]_s^\downarrow \geq \tau |\nabla_j \ell(\mathbf{z}^\infty)|, \forall j \notin \text{supp}(\mathbf{z}^\infty)$ . If  $\|\mathbf{z}^\infty\|_0 < s$ , then  $[\mathbf{z}^\infty]_s^\downarrow = 0$  and  $\nabla \ell(\mathbf{z}^\infty) = 0$  from (46) and (43). Those together with (13) show  $\mathbf{z}^\infty$  is a  $\tau$ -stationary point.  $\blacksquare$

Before we prove Lemma 11, we need the following property.

**Property 13** *Let  $\mathbf{z}$  be a strong  $\tau$ -stationary point of (4). We have following results.*

i) If  $\|\mathbf{z}\|_0 = s$ , then  $T^B(\mathbf{z}; \tau)$  is a singleton and for any  $0 < \tau_1 \leq \tau$

$$T^B(\mathbf{z}; \tau_1) = T^B(\mathbf{z}; \tau) = \{\text{supp}(\mathbf{z})\}.$$

ii) If  $\|\mathbf{z}\|_0 < s$ , then  $T^B(\mathbf{z}; \tau)$  has finite many entries, namely,

$$T^B(\mathbf{z}; \tau) = \{T \subseteq N_p : T \supset \text{supp}(\mathbf{z}), |T| = s\}.$$

Moreover, for any given  $\tau_1 > 0$ ,  $T^B(\mathbf{z}; \tau_1) = T^B(\mathbf{z}; \tau)$ .

**Proof** Since  $\mathbf{z}$  is a strong  $\tau$ -stationary point of (4), it follows from Theorem (21) that  $\mathbf{d} = \nabla \ell(\mathbf{z})$ . i) When  $\|\mathbf{z}\|_0 = s$ , Property 4 ii) yields

$$|d_i| = |\nabla_i \ell(\mathbf{z})| \begin{cases} = 0, & i \in \text{supp}(\mathbf{z}), \\ < \frac{1}{\tau} [\mathbf{z}]_s^\downarrow, & i \notin \text{supp}(\mathbf{z}), \end{cases} \quad (47)$$

which suffices to

$$|z_i - \tau d_i| = |z_i| \geq [\mathbf{z}]_s^\downarrow > \tau |d_j| = |z_j - \tau d_j|$$

for any  $i \in \text{supp}(\mathbf{z}), j \notin \text{supp}(\mathbf{z})$ . The above inequalities together with  $\|\mathbf{z}\|_0 = s$  imply  $\text{supp}(\mathbf{z}) \in T^B(\mathbf{z}; \tau)$ . If there is another  $T \in T^B(\mathbf{z}; \tau)$  such that  $T \neq \text{supp}(\mathbf{z})$ , then it follows that there exists a  $j_0 \in T \setminus \text{supp}(\mathbf{z})$  and  $i_0 \in \text{supp}(\mathbf{z}) \setminus T$ . This means

$$[\mathbf{z}]_s^\downarrow \stackrel{(47)}{>} \tau |d_{j_0}| = |z_{j_0} - \tau d_{j_0}| \stackrel{(17)}{\geq} |z_{i_0} - \tau d_{i_0}| \stackrel{(47)}{=} |z_{i_0}| \geq [\mathbf{z}]_s^\downarrow,$$

which is a contradiction. Therefore,  $T^B(\mathbf{z}; \tau) = \{T(\mathbf{z}; \tau)\} = \{\text{supp}(\mathbf{z})\}$  is a singleton. Again by Property 4, a strong  $\tau$ -stationary point  $\mathbf{z}$  is also a strong  $\tau_1$ -stationary point  $\mathbf{z}$  due to  $\frac{1}{\tau} [\mathbf{z}]_s^\downarrow \leq \frac{1}{\tau_1} [\mathbf{z}]_s^\downarrow$  for any  $0 < \tau_1 \leq \tau$ , thus  $T^B(\mathbf{z}; \tau) = \{T(\mathbf{z}; \tau_1)\} = \{\text{supp}(\mathbf{z})\}$ .

ii) When  $\|\mathbf{z}\|_0 < s$ , Property 4 i) yields  $\mathbf{d} = \nabla \ell(\mathbf{z}) = 0$ , which implies that  $[\mathbf{z} - \tau \mathbf{d}]_s^\downarrow = [\mathbf{z}]_s^\downarrow = 0$  and  $T(\mathbf{z}; \tau) = \{i \in N_p : |x_i| \geq 0\}$  with  $|T(\mathbf{z}; \tau)| = s$ . As  $\|\mathbf{z}\|_0 < s$ , we must have  $\text{supp}(\mathbf{z}) \subset T(\mathbf{z}; \tau)$  due to  $|T(\mathbf{z}; \tau)| = s$ . To form  $T(\mathbf{z}; \tau)$ , we need pick  $s - \|\mathbf{z}\|_0$  indices from  $N_p \setminus \text{supp}(\mathbf{z})$ , with  $C_{p-\|\mathbf{z}\|_0}^{s-\|\mathbf{z}\|_0}$  choices, yielding  $C_{p-\|\mathbf{z}\|_0}^{s-\|\mathbf{z}\|_0} T(\mathbf{z}; \tau)$ s to form  $T^B(\mathbf{z}; \tau)$ . Finally,  $\nabla f(\mathbf{z}) = 0$  suffices to  $\mathbf{z}$  being also a  $\tau_1$ -stationary point for any  $\tau_1 > 0$ . Similar choices to form  $T^B(\mathbf{z}; \tau_1)$  lead to  $T^B(\mathbf{z}; \tau) = T^B(\mathbf{z}; \tau_1)$ .  $\blacksquare$

### Proof of Lemma 11

**Proof** i) If  $\mathbf{z}^*$  is a strong  $\tau^*$ -stationary point of (4), then Theorem 7 shows  $F_{\tau^*}(\mathbf{u}^*; T_*) = 0$  for any  $T_* \in T^B(\mathbf{u}^*; \tau^*)$ . This means

$$\mathbf{z}_{T_*}^* = 0, \quad \mathbf{d}_{T_*}^* = 0, \quad \mathbf{d}^* = \nabla \ell(\mathbf{z}^*). \quad (48)$$

When  $\|\mathbf{z}^*\|_0 = s$ , Property 13 i) already shows  $T^B(\mathbf{u}^*; \tau^*) = T^B(\mathbf{u}^*; \tau) = \{\text{supp}(\mathbf{z}^*)\}$ . This means  $T_* = \text{supp}(\mathbf{z}^*)$ . Next, we show that

$$\text{supp}(\mathbf{z}) = \text{supp}(\mathbf{z}^*). \quad (49)$$

To prove above equation, we only prove  $\text{supp}(\mathbf{z}) \supseteq \text{supp}(\mathbf{z}^*)$  since  $\mathbf{z} \in S$  and  $\|\mathbf{z}^*\|_0 = s$ . If it is not true, i.e., there is an  $i \in \text{supp}(\mathbf{z}^*)$  but  $i \notin \text{supp}(\mathbf{z})$ , then it follows from the definition of  $N_S(\mathbf{u}^*, \delta^*)$  that

$$[\mathbf{z}^*]_s^\downarrow \stackrel{(31)}{\geq} \delta^* \stackrel{(32)}{>} \|\mathbf{u} - \mathbf{u}^*\| \geq \|\mathbf{z} - \mathbf{z}^*\| \geq |0 - z_i^*| \geq [\mathbf{z}^*]_s^\downarrow,$$

which is a contradiction. Again,  $\mathbf{u} \in N_S(\mathbf{u}^*, \delta^*)$  suffices to

$$\begin{aligned} (|z_i - z_i^*| + |d_i - d_i^*| + |d_j - d_j^*|)^2 &\leq 3(|z_i - z_i^*|^2 + |d_i - d_i^*|^2 + |d_j - d_j^*|^2) \\ &\leq 3\|\mathbf{u} - \mathbf{u}^*\|^2 \stackrel{(32)}{<} 4(\delta^*)^2. \end{aligned} \quad (50)$$

For any  $T \in T^B(\mathbf{u}; \tau)$ , we now prove  $\text{supp}(\mathbf{z}^*) = T$ . Considering any  $i \in \text{supp}(\mathbf{z}^*), j \notin \text{supp}(\mathbf{z}^*)$ , direct calculation derives following chain of inequalities

$$\begin{aligned} |z_i - \tau d_i| - |z_j - \tau d_j| &\stackrel{(49)}{=} |z_i - \tau d_i| - |\tau d_j| \geq |z_i| - \tau|d_i| - \tau|d_j| \\ &\stackrel{(48)}{=} |z_i - z_i^* + z_i^*| - \tau|d_i - d_i^*| - \tau|d_j - d_j^* + d_j^*| \\ &\geq |z_i^*| - |z_i - z_i^*| - \tau^*|d_i - d_i^*| - \tau^*|d_j - d_j^*| - \tau^*|d_j^*| \\ &\geq |z_i^*| - \max\{1, \tau^*\}(|z_i - z_i^*| + |d_i - d_i^*| + |d_j - d_j^*|) - \tau^*|d_j^*| \\ &\stackrel{(50)}{>} [\mathbf{z}^*]_s^\downarrow - 2\max\{1, \tau^*\}\delta^* - \tau^* \max_{j \notin \text{supp}(\mathbf{z}^*)} |d_j^*| \\ &\stackrel{(31)}{\geq} 0. \end{aligned}$$

This means for any  $i \in \text{supp}(\mathbf{z}^*)$  it has  $i \in T$ , namely,  $\text{supp}(\mathbf{z}^*) \subseteq T$ . Finally  $\|\mathbf{z}^*\|_0 = s$  and  $|T| = s$  suffice to  $\text{supp}(\mathbf{z}^*) = T$ . This combining the arbitrariness of  $T \in T^B(\mathbf{u}; \tau)$ , indicates  $T^B(\mathbf{u}; \tau) = \{\text{supp}(\mathbf{z}^*)\}$ .

ii) For  $\|\mathbf{z}^*\|_0 < s$ , if  $\mathbf{z}^* = 0$  the conclusion holds obviously due to  $\text{supp}(\mathbf{z}^*) = \emptyset$ . We now only focus on  $\mathbf{z}^* \neq 0$ , which means  $\delta^* > 0$  from (31). We first show that  $\Gamma_* := \text{supp}(\mathbf{z}^*) \subseteq \text{supp}(\mathbf{z})$ . If it is not true, i.e., there is an  $i \in \Gamma_*$  but  $i \notin \text{supp}(\mathbf{z})$ , then by the definition of  $N_S(\mathbf{u}^*, \delta^*)$  as (32), it follows

$$\min_{i \in \Gamma_*} |z_i^*| \stackrel{(31)}{\geq} \delta^* > \|\mathbf{u} - \mathbf{u}^*\| \geq \|\mathbf{z} - \mathbf{z}^*\| \geq |0 - z_i^*| \geq \min_{i \in \Gamma_*} |z_i^*|,$$

which is a contradiction. We then prove  $\Gamma_* \subseteq T$  for any  $T \in T^B(\mathbf{u}; \tau)$ , namely,  $|z_i - \tau d_i| > |z_j - \tau d_j|$  for any  $i \in \Gamma_*, j \notin \Gamma_*$ . For any given  $\tau > 0$  and any  $\mathbf{u} \in N_S(\mathbf{u}^*, \delta^*)$ , we have

$$\begin{aligned} &(|z_i - z_i^*| + |z_j - z_j^*| + |d_i - d_i^*| + |d_j - d_j^*|)^2 \\ &\leq 4(|z_i - z_i^*|^2 + |z_j - z_j^*|^2 + |d_i - d_i^*|^2 + |d_j - d_j^*|^2) < 4(\delta^*)^2. \end{aligned} \quad (51)$$

Notice that  $\mathbf{z}^*$  is a strong  $\tau^*$ -stationary point with  $\|\mathbf{z}^*\|_0 < s$ , then it must satisfy  $\mathbf{d}^* = \nabla\ell(\mathbf{z}^*) = 0$  by Property 4 i). Direct calculation derives following chain of inequalities

$$\begin{aligned}
& |z_i - \tau d_i| - |z_j - \tau d_j| \\
\geq & |z_i| - \tau|d_i| - |z_j| - \tau|d_j| \\
= & |z_i - z_i^* + z_i^*| - |z_j - z_j^*| - \tau|d_i| - \tau|d_j| && \text{(because of } z_j^* = 0, \forall j \notin \Gamma_*\text{)} \\
\geq & |z_i^*| - |z_i - z_i^*| - |z_j - z_j^*| - \tau^*|d_i - d_i^*| - \tau^*|d_j - d_j^*| && \text{(because of } \mathbf{d}^* = 0\text{)} \\
\geq & |z_i^*| - \max\{1, \tau^*\}(|z_i - z_i^*| + |z_j - z_j^*| + |d_i - d_i^*| + |d_j - d_j^*|) \\
> & \min_{i \in \Gamma_*} |z_i^*| - 2 \max\{1, \tau^*\} \delta^* && \text{(because of (51))} \\
\geq & 0. && \text{(because of (31))}
\end{aligned}$$

This means for any  $i \in \Gamma_*$  it has  $i \in T$ , namely,  $\Gamma_* \subseteq T$ . Overall we prove  $\Gamma_* \subseteq \text{supp}(\mathbf{x}) \cap T$ ,  $\forall T \in T^B(\mathbf{z}, \tau)$ . Finally, the proof of Property 13 ii) says that  $T^B(\mathbf{z}^*; \tau^*)$  contains every  $T_*$  which satisfies  $T_* \supseteq \Gamma_*$  and  $|T_*| = s$ . Particularly, it covers all  $T \in T^B(\mathbf{z}; \tau)$  because of  $T \supseteq \Gamma_*$  and  $|T| = s$ , namely  $T^B(\mathbf{z}; \tau) \subseteq T^B(\mathbf{z}^*; \tau^*)$ .  $\blacksquare$

## Proof of Theorem 12

**Proof** i) Since  $\mathbf{z}^*$  is a strong  $\tau^*$ -stationary point, then  $F_{\tau^*}(\mathbf{u}^*; T_*) = 0$  for any  $T_* \in T^B(\mathbf{u}^*; \tau^*)$  from Theorem 7, namely,

$$\mathbf{z}_{T_*}^* = 0, \quad \mathbf{d}_{T_*}^* = 0, \quad \mathbf{d}^* = \nabla\ell(\mathbf{z}^*). \quad (52)$$

Choose  $T_0 \in T^B(\mathbf{u}^0; \tau)$ . Lemma 11 and  $\mathbf{u}^0 \in N_S(\mathbf{u}^*, \delta^*)$  suffice to  $T_0 \in T^B(\mathbf{u}^0; \tau) \subseteq T^B(\mathbf{u}^*; \tau^*)$ . This together with (52) derives

$$\mathbf{z}_{T_0}^* = 0, \quad \mathbf{d}_{T_0}^* = \nabla_{T_0}\ell(\mathbf{z}^*) = 0. \quad (53)$$

For any  $0 \leq t \leq 1$ , denote

$$\begin{bmatrix} \mathbf{z}(t) \\ \mathbf{d}(t) \end{bmatrix} = \mathbf{u}(t) := \mathbf{u}^* + t(\mathbf{u}^0 - \mathbf{u}^*) = \begin{bmatrix} \mathbf{z}^* + t(\mathbf{z}^0 - \mathbf{z}^*) \\ \mathbf{d}^* + t(\mathbf{d}^0 - \mathbf{d}^*) \end{bmatrix}.$$

It is easy to verify that  $\mathbf{u}(t) \in N_S(\mathbf{u}^*, \delta^*)$ . This together with Property 1 iv) generates

$$\begin{aligned}
& \max \{ \|\nabla_T^2 \ell(\mathbf{z}^0) - \nabla_T^2 \ell(\mathbf{z}(t))\|, \|\nabla_T^2 \ell(\mathbf{z}^0) - \nabla_T^2 \ell(\mathbf{z}(t))\| \} \\
\leq & \|\nabla^2 \ell(\mathbf{z}^0) - \nabla^2 \ell(\mathbf{z}(t))\| \leq \gamma_x \|\mathbf{z}^0 - \bar{\mathbf{z}}\| = (1-t)\gamma_x \|\mathbf{z}^0 - \mathbf{z}^*\|, \quad (54)
\end{aligned}$$

where  $\nabla_T^2 \ell(\mathbf{z}^0)$  is the sub-matrix of  $\nabla^2 \ell(\mathbf{z}^0)$  containing rows indexed on  $T$ . Moreover, by Taylor expansion, one has

$$\nabla\ell(\mathbf{z}^0) - \nabla\ell(\mathbf{z}^*) = \int_0^1 \nabla^2 \ell(\mathbf{z}(t))(\mathbf{z}^0 - \mathbf{z}^*) dt. \quad (55)$$

It follows from  $\mathbf{u}^0 \in N_S(\mathbf{u}^*, \min\{\delta^*, \delta_x^*\})$  that  $\|\mathbf{z}^0 - \mathbf{z}^*\| \leq \|\mathbf{u}^0 - \mathbf{u}^*\| < \delta^*$ , which combining with (24) in Theorem 9 leads to

$$\|(\nabla_{T_0, T_0}^2 \ell(\mathbf{z}^0))^{-1}\|_2 \leq \mu^* = \frac{e^{2(\sqrt{\lambda_x} \|\mathbf{z}^*\| + \delta^*)}}{\underline{\lambda}/(4n)}. \quad (56)$$

Now, we have following chain of inequalities

$$\begin{aligned} & \|\mathbf{z}^1 - \mathbf{z}^*\| \\ &= (\|\mathbf{z}_{T_0}^1 - \mathbf{z}_{T_0}^*\|^2 + \|\mathbf{z}_{\bar{T}_0}^1 - \mathbf{z}_{\bar{T}_0}^*\|^2)^{1/2} \stackrel{(28,53)}{=} \|\mathbf{z}_{T_0}^1 - \mathbf{z}_{T_0}^*\| \\ &\stackrel{(28)}{=} \|\mathbf{z}_{T_0}^0 - \mathbf{z}_{T_0}^* - (\nabla_{T_0, T_0}^2 \ell(\mathbf{z}^0))^{-1} (\nabla_{T_0} \ell(\mathbf{z}^0) - \nabla_{T_0, \bar{T}_0}^2 \ell(\mathbf{z}^0) \mathbf{z}_{\bar{T}_0}^0)\| \\ &\stackrel{(56)}{\leq} \mu^* \|\nabla_{T_0, T_0}^2 \ell(\mathbf{z}^0) (\mathbf{z}_{T_0}^0 - \mathbf{z}_{T_0}^*) + \nabla_{T_0} \ell(\mathbf{z}^0) - \nabla_{T_0, \bar{T}_0}^2 \ell(\mathbf{z}^0) \mathbf{z}_{\bar{T}_0}^0\| \\ &\stackrel{(53)}{=} \mu^* \|\nabla_{T_0, T_0}^2 \ell(\mathbf{z}^0) (\mathbf{z}_{T_0}^0 - \mathbf{z}_{T_0}^*) - \nabla_{T_0} \ell(\mathbf{z}^0) + \nabla_{T_0} \ell(\mathbf{z}^*) + \nabla_{T_0, \bar{T}_0}^2 \ell(\mathbf{z}^0) (\mathbf{z}_{\bar{T}_0}^0 - \mathbf{z}_{\bar{T}_0}^*)\| \\ &\stackrel{(55)}{=} \mu^* \|\nabla_{T_0}^2 \ell(\mathbf{z}^0) (\mathbf{z}^0 - \mathbf{z}^*) - \int_0^1 \nabla_{T_0}^2 \ell(\mathbf{z}(t)) (\mathbf{z}^0 - \mathbf{z}^*) dt\| \\ &= \mu^* \left\| \int_0^1 [\nabla_{T_0}^2 \ell(\mathbf{z}^0) - \nabla_{T_0}^2 \ell(\mathbf{z}(t))] (\mathbf{z}^0 - \mathbf{z}^*) dt \right\| \\ &\leq \mu^* \int_0^1 \|\nabla_{T_0}^2 \ell(\mathbf{z}^0) - \nabla_{T_0}^2 \ell(\mathbf{z}(t))\| \|\mathbf{z}^0 - \mathbf{z}^*\| dt \\ &\stackrel{(54)}{\leq} \mu^* \gamma_x \|\mathbf{z}^0 - \mathbf{z}^*\|^2 \int_0^1 (1-t) dt = 0.5 \mu^* \gamma_x \|\mathbf{z}^0 - \mathbf{z}^*\|^2. \end{aligned} \quad (57)$$

Next we estimate  $\|\mathbf{d} - \mathbf{d}^*\|$ . One can verify that

$$\mu^* \lambda_x = \frac{e^{2(\sqrt{\lambda_x} \|\mathbf{z}^*\| + \delta^*)}}{\underline{\lambda}/(4n)} \lambda_x > \frac{\lambda_x}{\underline{\lambda}/(4n)} = \frac{\lambda_{\max}(X^\top X)}{\underline{\lambda}} \stackrel{(22)}{=} \frac{\lambda_{\max}(X^\top X)}{\min_{|T| \leq s} \lambda_{\min}(X_T^\top X_T)} \geq 1. \quad (58)$$

Similarly, we have following chain of inequalities

$$\begin{aligned} \|\mathbf{d}^1 - \mathbf{d}^*\| &= (\|\mathbf{d}_{T_0}^1 - \mathbf{d}_{T_0}^*\|^2 + \|\mathbf{d}_{\bar{T}_0}^1 - \mathbf{d}_{\bar{T}_0}^*\|^2)^{1/2} \stackrel{(28,53)}{=} \|\mathbf{d}_{T_0}^1 - \mathbf{d}_{T_0}^*\| \\ &\stackrel{(28,53)}{=} \|\nabla_{\bar{T}_0} \ell(\mathbf{z}^0) + \nabla_{\bar{T}_0}^2 \ell(\mathbf{z}^0) (\mathbf{z}^1 - \mathbf{z}^0) - \nabla_{\bar{T}_0} \ell(\mathbf{z}^*)\| \\ &= \|\nabla_{\bar{T}_0} \ell(\mathbf{z}^0) - \nabla_{\bar{T}_0} \ell(\mathbf{z}^*) - \nabla_{\bar{T}_0}^2 \ell(\mathbf{z}^0) (\mathbf{z}^0 - \mathbf{z}^*) + \nabla_{\bar{T}_0}^2 \ell(\mathbf{z}^0) (\mathbf{z}^1 - \mathbf{z}^*)\| \\ &\stackrel{(55)}{=} \left\| \int_0^1 [\nabla_{\bar{T}_0}^2 \ell(\mathbf{z}^0) - \nabla_{\bar{T}_0}^2 \ell(\mathbf{z}(t))] (\mathbf{z}^0 - \mathbf{z}^*) dt + \nabla_{\bar{T}_0}^2 \ell(\mathbf{z}^0) (\mathbf{z}^1 - \mathbf{z}^*) \right\| \\ &\stackrel{(54)}{\leq} \gamma_x \|\mathbf{z}^0 - \mathbf{z}^*\|^2 \int_0^1 (1-t) dt + \|\nabla_{\bar{T}_0}^2 \ell(\mathbf{z}^0)\|_2 \|\mathbf{z}^1 - \mathbf{z}^*\| \\ &\leq 0.5 \gamma_x \|\mathbf{z}^0 - \mathbf{z}^*\|^2 + \|\nabla_{\bar{T}_0}^2 \ell(\mathbf{z}^0)\|_2 \|\mathbf{z}^1 - \mathbf{z}^*\| \\ &\leq 0.5 \gamma_x \|\mathbf{z}^0 - \mathbf{z}^*\|^2 + \lambda_x \|\mathbf{z}^1 - \mathbf{z}^*\| \\ &\stackrel{(57)}{\leq} 0.5 \gamma_x (1 + \mu^* \lambda_x) \|\mathbf{z}^0 - \mathbf{z}^*\|^2 \stackrel{(58)}{\leq} \gamma_x \mu^* \lambda_x \|\mathbf{z}^0 - \mathbf{z}^*\|^2. \end{aligned} \quad (59)$$



Based on (57) and (59), we have

$$\begin{aligned}
\|\mathbf{u}^1 - \mathbf{u}^*\|^2 &= \|\mathbf{z}^1 - \mathbf{z}^*\|^2 + \|\mathbf{d}^1 - \mathbf{d}^*\|^2 \\
&\leq (1/2\gamma_x\mu^*)^2\|\mathbf{z}^0 - \mathbf{z}^*\|^4 + (\gamma_x\mu^*\lambda_x)^2\|\mathbf{z}^0 - \mathbf{z}^*\|^4 \\
&\leq (1/4 + \lambda_x^2)(\gamma_x\mu^*)^2\|\mathbf{z}^0 - \mathbf{z}^*\|^4 \\
&= (0.5/\delta_x^*)^2\|\mathbf{z}^0 - \mathbf{z}^*\|^4 \leq (0.5/\delta_x^*)^2\|\mathbf{u}^0 - \mathbf{u}^*\|^4,
\end{aligned} \tag{60}$$

which gives rise to

$$\|\mathbf{u}^1 - \mathbf{u}^*\| \leq (0.5/\delta_x^*)\|\mathbf{u}^0 - \mathbf{u}^*\|^2 \leq 0.5\|\mathbf{u}^0 - \mathbf{u}^*\|. \tag{61}$$

The above inequality suffices to  $\|\mathbf{u}^1 - \mathbf{u}^*\| \leq \|\mathbf{u}^0 - \mathbf{u}^*\| < \min\{\delta^*, \delta_x^*\}$ . In addition,  $\mathbf{z}_{T_0}^1 = 0$  from (28) means  $\mathbf{z}^1 \in S$ . Then  $\mathbf{u}^1 \in N_S(\mathbf{u}^*, \min\{\delta^*, \delta_x^*\})$ . Similar reasons to derive (60) allow us to get

$$\|\mathbf{u}^2 - \mathbf{u}^*\| \leq (0.5/\delta_x^*)\|\mathbf{u}^1 - \mathbf{u}^*\|^2.$$

By induction, one could conclude that  $\mathbf{u}^k \in N_S(\mathbf{u}^*, \min\{\delta^*, \delta_x^*\})$  and

$$\|\mathbf{u}^{k+1} - \mathbf{u}^*\| \leq (0.5/\delta_x^*)\|\mathbf{u}^k - \mathbf{u}^*\|^2. \tag{62}$$

Hence  $\lim_{k \rightarrow \infty} \mathbf{u}^k = \mathbf{u}^*$ , and the sequence  $\{\mathbf{z}^k\}$  has quadratic convergence rate.

ii) The above proof shows that  $\mathbf{u}^k \in N_S(\mathbf{u}^*, \min\{\delta^*, \delta_x^*\})$  for any  $k \geq 0$ , which indicates  $\mathbf{u}^k \in N_S(\mathbf{u}^*, \delta^*)$ . This combining Lemma 11 directly derives the claim.

iii) Since  $\mathbf{u}^k \in N_S(\mathbf{u}^*, \min\{\delta^*, \delta_x^*\})$  and (62), we get

$$\|\mathbf{u}^{k+1} - \mathbf{u}^*\| \leq 0.5\|\mathbf{u}^k - \mathbf{u}^*\|. \tag{63}$$

By the last equation of (26), we have

$$\mathbf{d}^{k+1} = \nabla\ell(\mathbf{z}^k) + \nabla^2\ell(\mathbf{z}^k)(\mathbf{z}^{k+1} - \mathbf{z}^k). \tag{64}$$

In addition, it follows from ii)  $T_{k+1} \in T^B(\mathbf{z}^*; \tau^*)$ , Property 4 and (52) that

$$\begin{aligned}
\|\mathbf{z}^*\|_0 = s &\implies \text{supp}(\mathbf{z}^*) = T_{k+1} \\
&\implies \mathbf{z}_{T_{k+1}}^* = 0, \quad \mathbf{d}_{T_{k+1}}^* \stackrel{(52)}{=} 0, \\
\|\mathbf{z}^*\|_0 < s &\implies \text{supp}(\mathbf{z}^*) \subset T_{k+1}, \quad \mathbf{d}^* \stackrel{(52)}{=} \nabla\ell(\mathbf{z}^*) \stackrel{\text{Property 4 i)}}{=} 0 \\
&\implies \mathbf{z}_{T_{k+1}}^* = 0, \quad \mathbf{d}_{T_{k+1}}^* = 0.
\end{aligned} \tag{65}$$

These give rise to

$$\begin{aligned}
F_\tau(\mathbf{u}^{k+1}; T_{k+1}) &\stackrel{(20)}{=} \begin{bmatrix} \mathbf{d}_{T_{k+1}}^{k+1} \\ \mathbf{z}_{T_{k+1}}^{k+1} \\ \mathbf{d}^{k+1} - \nabla\ell(\mathbf{z}^{k+1}) \end{bmatrix} \stackrel{(65)}{=} \begin{bmatrix} \mathbf{d}_{T_{k+1}}^{k+1} - \mathbf{d}_{T_{k+1}}^* \\ \mathbf{z}_{T_{k+1}}^{k+1} - \mathbf{z}_{T_{k+1}}^* \\ \mathbf{d}^{k+1} - \nabla\ell(\mathbf{z}^{k+1}) \end{bmatrix} \\
&\stackrel{(64)}{=} \begin{bmatrix} \mathbf{d}_{T_{k+1}}^{k+1} - \mathbf{d}_{T_{k+1}}^* \\ \mathbf{z}_{T_{k+1}}^{k+1} - \mathbf{z}_{T_{k+1}}^* \\ \nabla\ell(\mathbf{z}^k) + \nabla^2\ell(\mathbf{z}^k)(\mathbf{z}^{k+1} - \mathbf{z}^k) - \nabla\ell(\mathbf{z}^{k+1}) \end{bmatrix}.
\end{aligned}$$

Then Taylor Expansion,

$$\nabla\ell(\mathbf{z}^{k+1}) - \nabla\ell(\mathbf{z}^k) = \int_0^1 \nabla^2\ell(\mathbf{z}^k + t(\mathbf{z}^{k+1} - \mathbf{z}^k))(\mathbf{z}^{k+1} - \mathbf{z}^k)dt$$

enables us to derive that

$$\begin{aligned} I &:= \|\nabla\ell(\mathbf{z}^k) + \nabla^2\ell(\mathbf{z}^k)(\mathbf{z}^{k+1} - \mathbf{z}^k) - \nabla\ell(\mathbf{z}^{k+1})\| \\ &= \left\| \int_0^1 (\nabla^2\ell(\mathbf{z}^k + t(\mathbf{z}^{k+1} - \mathbf{z}^k)) - \nabla^2\ell(\mathbf{z}^k))(\mathbf{z}^{k+1} - \mathbf{z}^k)dt \right\| \\ &\leq \int_0^1 \|\nabla^2\ell(\mathbf{z}^k + t(\mathbf{z}^{k+1} - \mathbf{z}^k)) - \nabla^2\ell(\mathbf{z}^k)\| \|\mathbf{z}^{k+1} - \mathbf{z}^k\| dt \\ &\stackrel{(11)}{\leq} \int_0^1 t\gamma_x \|\mathbf{z}^{k+1} - \mathbf{z}^k\|^2 dt = 0.5\gamma_x \|\mathbf{z}^{k+1} - \mathbf{z}^k\|^2 \\ &\leq 0.5\gamma_x \|\mathbf{u}^{k+1} - \mathbf{u}^k\|^2 \leq \gamma_x (\|\mathbf{u}^{k+1} - \mathbf{u}^*\|^2 + \|\mathbf{u}^k - \mathbf{u}^*\|^2) \\ &\stackrel{(63)}{\leq} 1.25\gamma_x \|\mathbf{u}^k - \mathbf{u}^*\|^2. \end{aligned}$$

Next we estimate

$$\begin{aligned} II &:= \|\mathbf{d}_{T_{k+1}}^{k+1} - \mathbf{d}_{T_{k+1}}^*\|^2 + \|\mathbf{z}_{T_{k+1}}^{k+1} - \mathbf{z}_{T_{k+1}}^*\|^2 \\ &\leq \|\mathbf{u}^{k+1} - \mathbf{u}^*\|^2 \stackrel{(62)}{\leq} (0.5/\delta_x^*)^2 \|\mathbf{u}^k - \mathbf{u}^*\|^4. \end{aligned}$$

Based on those facts, we obtain

$$\begin{aligned} \|F_\tau(\mathbf{u}^{k+1}; T_{k+1})\| &= \sqrt{I^2 + II} \leq [(0.5/\delta_x^*)^2 + (1.25\gamma_x)^2]^{1/2} \|\mathbf{u}^k - \mathbf{u}^*\|^2, \\ &= c_x \|\mathbf{u}^k - \mathbf{u}^*\|^2 \stackrel{(63)}{\leq} c_x 2^{-2k} \|\mathbf{u}^0 - \mathbf{u}^*\|^2, \end{aligned}$$

which suffices to the conclusion. ■

## References

- Alekh Agarwal, Sahand Negahban, and Martin J Wainwright. Fast global convergence rates of gradient methods for high-dimensional statistical recovery. In *Advances in Neural Information Processing Systems*, pages 37–45, 2010.
- Galen Andrew and Jianfeng Gao. Scalable training of  $\ell_1$ -regularized log-linear models. In *Proceedings of the 24th international conference on Machine learning*, pages 33–40. ACM, 2007.
- Sohail Bahmani, Bhiksha Raj, and Petros T Boufounos. Greedy sparsity-constrained optimization. *Journal of Machine Learning Research*, 14(Mar):807–841, 2013.
- Amir Beck and Yonina C Eldar. Sparsity constrained nonlinear optimization: Optimality conditions and algorithms. *SIAM Journal on Optimization*, 23(3):1480–1509, 2013.

- Amir Beck and Nadav Hallak. On the minimization over sparse symmetric sets: projections, optimality conditions, and algorithms. *Mathematics of Operations Research*, 41(1):196–223, 2015.
- Thomas Blumensath and Mike E Davies. Iterative hard thresholding for compressed sensing. *Applied and computational harmonic analysis*, 27(3):265–274, 2009.
- Emmanuel J Candes, Michael B Wakin, and Stephen P Boyd. Enhancing sparsity by reweighted  $\ell_1$  minimization. *Journal of Fourier analysis and applications*, 14(5-6):877–905, 2008.
- Jinghui Chen and Quanquan Gu. Fast newton hard thresholding pursuit for sparsity constrained nonconvex optimization. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 757–766. ACM, 2017.
- David L Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006.
- Bradley Efron, Trevor Hastie, Iain Johnstone, Robert Tibshirani, et al. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.
- Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.
- Mário AT Figueiredo. Adaptive sparseness for supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 25(9):1150–1159, 2003.
- Simon Foucart. Hard thresholding pursuit: an algorithm for compressive sensing. *SIAM Journal on Numerical Analysis*, 49(6):2543–2563, 2011.
- Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.
- Rahul Garg and Rohit Khandekar. Gradient descent with sparsification: an iterative algorithm for sparse recovery with restricted isometry property. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 337–344. ACM, 2009.
- Pinghua Gong and Jieping Ye. Honor: Hybrid optimization for non-convex regularized problems. In *Advances in Neural Information Processing Systems*, pages 415–423, 2015.
- Pinghua Gong, Changshui Zhang, Zhaosong Lu, Jianhua Huang, and Jieping Ye. A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems. In *International Conference on Machine Learning*, pages 37–45, 2013.
- James Douglas Hamilton. *Time series analysis*, volume 2. Princeton university press Princeton, NJ, 1994.
- Trevor Hastie, Robert Tibshirani, and Ryan J Tibshirani. Extended comparisons of best subset selection, forward stepwise selection, and the lasso. *arXiv preprint arXiv:1707.08692*, 2017.

- Hussein Hazimeh and Rahul Mazumder. Fast best subset selection: Coordinate descent and local combinatorial optimization algorithms. *arXiv preprint arXiv:1803.01454*, 2018.
- Jian Huang, Shuangge Ma, Huiliang Xie, and Cun-Hui Zhang. A group bridge approach for variable selection. *Biometrika*, 96(2):339–355, 2009.
- Jian Huang, Yuling Jiao, Yanyan Liu, and Xiliang Lu. A constructive approach to  $l_0$  penalized regression. *Journal of Machine Learning Research*, 19(10), 2018.
- Kwangmoo Koh, Seung-Jean Kim, and Stephen Boyd. An interior-point method for large-scale  $\ell_1$ -regularized logistic regression. *Journal of Machine learning research*, 8(Jul): 1519–1555, 2007.
- Balaji Krishnapuram, Lawrence Carin, Mario AT Figueiredo, and Alexander J Hartemink. Sparse multinomial logistic regression: Fast algorithms and generalization bounds. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (6):957–968, 2005.
- Su-In Lee, Honglak Lee, Pieter Abbeel, and Andrew Y Ng. Efficient  $l_1$  regularized logistic regression. In *AAAI*, volume 6, pages 401–408, 2006.
- Huan Li and Zhouchen Lin. Accelerated proximal gradient methods for nonconvex programming. In *Advances in neural information processing systems*, pages 379–387, 2015.
- Jun Liu, Jianhui Chen, and Jieping Ye. Large-scale sparse logistic regression. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 547–556. ACM, 2009a.
- Jun Liu, Shuiwang Ji, Jieping Ye, et al. Slep: Sparse learning with efficient projections. *Arizona State University*, 6(491):7, 2009b.
- Aurelie Lozano, Grzegorz Swirszcz, and Naoki Abe. Group orthogonal matching pursuit for logistic regression. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 452–460, 2011.
- Zhaosong Lu. Optimization over sparse symmetric sets via a nonmonotone projected gradient method. *arXiv preprint arXiv:1509.08581*, 2015.
- Zhaosong Lu and Yong Zhang. Sparse approximation via penalty decomposition methods. *SIAM Journal on Optimization*, 23(4):2448–2478, 2013.
- Stéphane Mallat and Zhifeng Zhang. Matching pursuit with time-frequency dictionaries. Technical report, Courant Institute of Mathematical Sciences New York United States, 1993.
- Rahul Mazumder, Peter Radchenko, and Antoine Dedieu. Subset selection with shrinkage: Sparse linear modeling when the snr is low. *arXiv preprint arXiv:1708.03288*, 2017.
- Deanna Needell and Joel A Tropp. Cosamp: Iterative signal recovery from incomplete and inaccurate samples. *Applied and computational harmonic analysis*, 26(3):301–321, 2009.

- Li-Li Pan, Nai-Hua Xiu, and Sheng-Long Zhou. On solutions of sparsity constrained optimization. *Journal of the Operations Research Society of China*, 3(4):421–439, 2015.
- Lili Pan, Shenglong Zhou, Naihua Xiu, and Hou-Duo Qi. A convergent iterative hard thresholding for nonnegative sparsity optimization. *Pacific Journal of Optimization*, 13(2):325–353, 2017.
- Yaniv Plan and Roman Vershynin. Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach. *IEEE Transactions on Information Theory*, 59(1):482–494, 2013.
- Alain Rakotomamonjy, Remi Flamary, and Gilles Gasso. Dc proximal newton for nonconvex optimization problems. *IEEE transactions on neural networks and learning systems*, 27(3):636–647, 2016.
- Jianing Shi, Wotao Yin, Stanley Osher, and Paul Sajda. A fast hybrid algorithm for large-scale  $\ell_1$ -regularized logistic regression. *Journal of Machine Learning Research*, 11(Feb):713–741, 2010.
- Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.
- Rui Wang, Naihua Xiu, and Chao Zhang. Greedy projected gradient-newton method for large-scale sparse logistic regression. *technical report*, 2017.
- Weijun Xie and Xinwei Deng. The ccp selector: Scalable algorithms for sparse ridge regression from chance-constrained programming. *arXiv preprint arXiv:1806.03756*, 2018.
- Jin Yu, SVN Vishwanathan, Simon Günter, and Nicol N Schraudolph. A quasi-newton approach to nonsmooth convex optimization problems in machine learning. *Journal of Machine Learning Research*, 11(Mar):1145–1200, 2010.
- Guo-Xun Yuan, Kai-Wei Chang, Cho-Jui Hsieh, and Chih-Jen Lin. A comparison of optimization methods and software for large-scale  $\ell_1$ -regularized linear classification. *Journal of Machine Learning Research*, 11(Nov):3183–3234, 2010.
- Guo-Xun Yuan, Chia-Hua Ho, and Chih-Jen Lin. An improved glmnet for  $\ell_1$ -regularized logistic regression. *Journal of Machine Learning Research*, 13(Jun):1999–2030, 2012.
- Xiao-Tong Yuan and Qingshan Liu. Newton greedy pursuit: A quadratic approximation method for sparsity-constrained optimization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4122–4129, 2014.
- Xiao-Tong Yuan and Qingshan Liu. Newton-type greedy selection methods for  $\ell_0$ -constrained minimization. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2437–2450, 2017.
- Xiao-Tong Yuan, Ping Li, and Tong Zhang. Gradient hard thresholding pursuit for sparsity-constrained optimization. In *International Conference on Machine Learning*, pages 127–135, 2014.

- Xiao-Tong Yuan, Ping Li, and Tong Zhang. Gradient hard thresholding pursuit. *Journal of Machine Learning Research*, 18(166):1–43, 2018.
- Cun-Hui Zhang et al. Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, 38(2):894–942, 2010.
- Tong Zhang. Analysis of multi-stage convex relaxation for sparse regularization. *Journal of Machine Learning Research*, 11(Mar):1081–1107, 2010.
- Hui Zou and Runze Li. One-step sparse estimates in nonconcave penalized likelihood models. *Annals of statistics*, 36(4):1509, 2008.