Nonconvex Sparse Logistic Regression with Weakly Convex Regularization

Xinyue Shen, Student Member, IEEE, and Yuantao Gu, Senior Member, IEEE

Abstract—In this work we propose to fit a sparse logistic regression model by a weakly convex regularized nonconvex optimization problem. The idea is based on the finding that a weakly convex function as an approximation of the \( \ell_0 \) pseudo norm is able to better induce sparsity than the commonly used \( \ell_1 \) norm. For a class of weakly convex sparsity inducing functions, we prove the nonconvexity of the corresponding problem, and study its local optimality conditions and the choice of the regularizer parameter. Despite the nonconvexity, a method based on proximal gradient descent is used to solve the general weakly convex sparse logistic regression, and its convergence behavior is studied theoretically. Then the general framework is applied to a specific weakly convex function, and a local optimality condition and a bound on the logistic loss at a local optimum are provided. Its effectiveness is demonstrated in numerical experiments by both randomly generated and real datasets.

Index Terms—sparse logistic regression, weakly convex regularization, nonconvex optimization, proximal gradient descent

I. INTRODUCTION

Logistic regression is a widely used supervised machine learning method for classification. It learns a neutral hyperplane in the feature space of a learning problem according to a probabilistic model, and classifies test data points accordingly. The output of the classification result does not only give a class label, but also a natural probabilistic interpretation. It can be straightforwardly extended from two-class to multi-class problems, and it has been applied to text classification [1], gene selection and microarray analysis [2], [3], combinatorial chemistry [4], image analysis [5], [6], etc.

In a classification problem \( N \) pairs of training data \( \{(x^{(i)}, y^{(i)}) ; i = 1, \ldots, N \} \) are given, where every point \( x^{(i)} \in \mathbb{R}^d \) is a feature vector in the \( d \) dimensional feature space, and \( y^{(i)} \) is its corresponding class label. In a two-class logistic regression problem, \( y^{(i)} \in \{0, 1\} \), and it is assumed that the probability distribution of a class label \( y \) given a feature vector \( x \) is as follows

\[
p(y = 1|x; \theta) = \sigma(\theta^T x) = \frac{1}{1 + \exp(-\theta^T x)}
\]

\[
p(y = 0|x; \theta) = 1 - \sigma(\theta^T x) = \frac{1}{1 + \exp(\theta^T x)},
\]

where \( \sigma(\cdot) \) is the sigmoid function defined as above, and \( \theta \in \mathbb{R}^d \) is the model parameter to be learned. When \( \theta^T x = 0 \), the probability of having either label is 0.5, and thus the vector \( \theta \) gives the normal vector of a neutral hyperplane. Notice that if an affine hyperplane \( \theta^T x + b \) is to be considered, then we can simply add an additional dimension with value 1 to every feature vector, and then it will have the linear hyperplane form.

Suppose that the labels of the training samples are independently drawn from the probability distribution (1), then it has been proposed to learn \( \theta \) by minimizing the negative log-likelihood function, and the optimization problem is as follows

\[
\text{minimize} \quad l(\theta),
\]

where \( \theta \in \mathbb{R}^d \) is the variable, and \( l \) is the (empirical) logistic loss

\[
l(\theta) = \sum_{i=1}^{N} - \log p(y^{(i)}|x^{(i)}; \theta).
\]

Problem (2) is convex and differentiable, and can be readily solved [7]. Once we obtain a solution \( \hat{\theta} \), given a new feature vector \( x \), we can predict the probability of the two possible labels according to the logistic model, and take the one with larger probability by

\[
y = 1 \left( x^T \hat{\theta} \geq 0 \right) = \begin{cases} 
1, & x^T \hat{\theta} \geq 0; \\
0, & x^T \hat{\theta} < 0.
\end{cases}
\]

When the number of training samples \( N \) is relatively small compared to the feature space dimension \( d \), adding a regularization can avoid over-fitting and enhance classification accuracy on test data, and the \( \ell_2 \) norm has long been used as a regularization function [8]–[10]. Furthermore, a sparsity-inducing regularizer can select a subset of all available features that capture the relevant properties. Since \( \ell_1 \) norm is a convex function that induces sparsity, the \( \ell_1 \) norm regularized sparse logistic regression prevails [1], [11]–[15].

Despite that in general nonconvex optimization is hard to solve globally, nonconvex regularization has been extensively studied to induce sparsity in sparse logistic regression [16], [17], compressed sensing [18]–[21], and precision matrix estimation in Gaussian graphical models [22]. Inspired by results...
that tie binomial regression and one-bit compressed sensing [15], as well as results in [18], [19], [22], [23] indicating that weakly convex functions are able to better induce sparsity than the $\ell_1$ norm in these sparsity related problems, in this work we propose to use a weakly convex function in sparse logistic regression.

A. Contribution and outline

In this work, we consider a logistic regression problem in which the model parameter $\theta$ is sparse, i.e., the dimension $d$ can be large, and $\theta$ is assumed to have only $K$ non-zero elements, where $K$ is relatively small compared to $d$. We propose the following problem that uses a weakly convex (nonconvex) function $J$ in sparse logistic regression

$$\min_{\theta} \ l(\theta) + \beta J(\theta),$$

(4)

where the variable is $\theta \in \mathbb{R}^d$, $\beta > 0$ is a regularization parameter, and $l$ is the logistic loss (3). The contribution of this work can be summarized as follows.

- A class of weakly convex regularizations is proposed to be used in sparse logistic regression, in order to better induce sparsity compared with the classic $\ell_1$ regularization. The proposed problem is proved to be nonconvex, and a sufficient and a necessary local optimality conditions are provided, which can be used to derive further results in a specific case studied in the paper. Then we theoretically provide a tight bound on the choice of the regularization parameter $\beta$ to exclude 0 as a local optimum. These will be in section III.

- A solution method based on proximal gradient is proposed to solve the general weakly convex regularized problem (4). Despite its nonconvexity, we provide a theoretical convergence analysis, which includes the convergence of the objective value, the convergence of the increment of the iterates, and a performance bound on any limit point of the generated sequence. These will be in section IV.

- We apply the general framework to a specific weakly convex regularizer. As corollaries a necessary and sufficient condition on its local optimality and a bound on the logistic loss at a local optimum are obtained. The convergence analysis of the method for the general problem can also be applied, and a Nesterov acceleration is used with guaranteed convergence of the objective value. These will be in section V. In numerical experiments in section VI, we use this specific choice to verify the effectiveness of the model and the method on both randomly generated and real datasets.

B. Notations

In this work, for a vector $x$, we use $\|x\|_2$ to denote its $\ell_2$ norm, $\|x\|_1$ to denote its $\ell_1$ norm, and $\|x\|_\infty$ to denote its infinity norm. Its $i$th entry is denoted as $x_i$. For a matrix $X$, $\|X\|$ is its operator norm, i.e., its largest singular value. For a differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, its gradient is denoted as $\nabla f$, and if it is twice differentiable, then its Hessian is denoted as $\nabla^2 f$. If $f$ is a convex function, $\partial f(x)$ is its subgradient set at point $x$. For a function $F : \mathbb{R} \to \mathbb{R}$, we use $F^\prime_-$ and $F^\prime_+$ to denote its left and right derivatives, and $F^\prime_\pm$ and $F''_\pm$ to denote its left and right second derivatives, if they exist.

II. RELATED WORKS

A. $\ell_2$ and $\ell_1$ regularized logistic regression

The $\ell_2$ regularized logistic regression learns $\theta$ by solving the following problem

$$\min_{\theta} \ l(\theta) + \beta \|\theta\|^2_2,$$

where $\theta$ is the variable, $l$ is the logistic loss (3), and $\beta > 0$ is the regularization parameter. The solution can be interpreted as the maximum a posteriori probability (MAP) estimate of $\theta$, if $\theta$ has a Gaussian prior distribution with zero mean and covariance $\beta I$ [8]. The problem is strongly convex and differentiable, and can be solved by methods such as the Newton, quasi-Newton, coordinate descent, conjugate gradient descent, and iteratively reweighted least squares. For example, see [9], [10] and references therein.

It has been known that minimizing the $\ell_1$ norm of a variable induces sparsity to its solution, so the following $\ell_1$ norm regularized sparse logistic regression has been widely used to promote the sparsity of $\theta$

$$\min_{\theta} \ l(\theta) + \beta \|\theta\|_1,$$

(5)

where $\theta$ is the variable, $l$ is the logistic loss (3), and $\beta > 0$ is a parameter balancing the sparsity and the classification error on the training data. In logistic regression, $\theta_j = 0$ means that the $j$th feature does not have influence on the classification result. Thus, sparse logistic regression tries to find a few features that are relevant to the classification results from a large number of features. Its solution can also be interpreted as a MAP estimate, when $\theta$ has a Laplacian prior distribution $p(\theta) = (\beta/2)^d \exp(-\beta \|\theta\|_1)$.

The problem (5) is convex but non-differentiable, and several specialized solution methods have been proposed, such as an iteratively reweighted least squares (IRLS) method [13] in which every iteration solves a LASSO [24], a generalized LASSO method [11], a coordinate descent method [1], a Gauss-Seidel method [12], an interior point method that scales well to large problems [14], and some online algorithms such as [25].

B. Nonconvex sparse logistic regression and SVMs

The work [16] studies properties of local optima of a class of nonconvex regularized M-estimators including logistic regression and the convergence behavior of a proposed composite gradient descent solution method. The nonconvex regularizers considered in their work overlap with the ones in this work, but they have a convex constraint in addition.

Difference of convex (DC) functions are proposed in works such as [17], [26], [27] to approximate the $\ell_0$ pseudo norm and work as the regularization for feature selection in logistic
regression and support vector machines (SVMs). Their solution methods are based on the difference of convex functions algorithm (DCA), where each iteration involves solving a convex program (linear program for SVMs). In this work, our regularizer also belongs to the general class of DC functions, but we study a more specific class, i.e., the weakly convex functions, and there is no need to numerically solve a convex program in every iteration to solve the problem, given that the proximal operator of the weakly convex function has a closed form expression.

C. Nonconvex compressed sensing

From the perspective of reconstructing \( \theta \), one-bit compressed sensing [28] studies a similar problem, where a sparse vector \( \theta \) (or its normalization \( \theta/\|\theta\|_2 \)) is to be estimated from several one-bit measurements \( y^{(i)} = \mathbf{1}(\theta^T x^{(i)} \geq 0) \), and compressed sensing [29] studies a problem where a sparse \( \theta \) is to be estimated from several linear measurements \( y^{(i)} = \theta^T x^{(i)} \). In this setting \( x^{(i)} \) for \( i = 1, \ldots, N \) are known sensing vectors. Nonconvex regularizations have been used to promote sparsity in both compressed sensing [19], [23], [30]–[32] and one-bit compressed sensing [20]. These studies have shown that, despite that nonconvex optimization problems are usually hard to solve globally, with some proper choices of the nonconvex regularizers, using some local methods their recovery performances can be better than that of the \( \ell_1 \) regularization, both theoretically and numerically, in terms of required number of measurements and robustness against noise.

D. Weakly convex sparsity inducing function

In this work we use a class of weakly convex functions to induce sparsity, and the definition is as follows.

**Definition 1.** The weakly convex sparsity inducing function \( J \) is defined to be separable

\[
J(x) = \sum_{i=1}^{n} F(x_i),
\]

where the function \( F : \mathbb{R} \to \mathbb{R}^+ \) satisfies the following properties.

- Function \( F \) is even, not identically zero, and \( F(0) = 0 \);
- Function \( F \) is non-decreasing and concave on \([0, \infty)\);
- Function \( F \) is weakly convex [33] on \([0, \infty)\) with non-convexity parameter \( \zeta > 0 \), i.e., \( \zeta \) is the smallest positive scalar such that the function

\[
H(t) = F(t) + \zeta t^2
\]

is convex.

A similar definition of weakly convex sparsity inducing functions has been proposed in [19], and the difference is that in [19] the concavity condition is replaced with that the function \( t \mapsto F(t)/t \) is nonincreasing on \((0, \infty)\).

According to the definition function \( J \) is weakly convex, and

\[
G(x) = J(x) + \zeta \|x\|_2^2 = \sum_{i=1}^{d} H(x_i)
\]

is a convex function. Thus, \( J \) belongs to a wider class of DC functions [26], and both \( J \) and \( G \) are separable across all coordinates. Since \( \zeta > 0 \), the function \( J \) is nonconvex, and it can be nondifferentiable, which indicates that an optimization problem with \( J \) in the objective function can be hard to solve. Nevertheless, the fact that by adding a quadratic term the function becomes convex allows it to have some favorable properties, such as that its proximal operator is well defined by a convex problem with a unique solution.

The proximal operator \( \text{prox}_{\beta J}(\cdot) \) of function \( J \) with parameter \( \beta \) is defined as

\[
\text{prox}_{\beta J}(v) = \arg \min \beta J(x) + \frac{1}{2} \|x - v\|_2^2,
\]

where the minimization is with respect to \( x \). If \( \beta \) is small enough so that \( \beta \zeta < \frac{1}{2} \), then the objective function in (6) is strongly convex, and the minimizer is unique. For some weakly convex functions, their proximal operators have closed form expressions which are relatively easy to compute.

For instance a specific \( F \) satisfying Definition 1 is defined as follows

\[
F(t) = \begin{cases} 
|t| - \zeta t^2 & |t| \leq \frac{1}{\zeta} \\
\frac{1}{\zeta} t^2 & |t| > \frac{1}{\zeta}.
\end{cases}
\]

The function in (7) is also called minimax concave penalty (MCP) proposed in [34] for penalized variable selection in linear regression, and has been used in sparse logistic regression [16], compressed sensing [20], [23], and precision matrix estimation for high dimensional Gaussian graphical models [22]. Its proximal operator with \( \beta \zeta < \frac{1}{2} \) can be explicitly written as

\[
\text{prox}_{\beta F}(v) = \begin{cases} 
0 & |v| < \beta \\
\frac{v - \beta \text{sign}(v)}{1 - 2\beta \zeta} & \beta \leq |v| \leq \frac{1}{\zeta} \\
v & |v| > \frac{1}{\zeta}.
\end{cases}
\]

The proximal operator (8) is also called firm shrinkage operator [35], which generalizes the hard and soft shrinkage corresponding to the proximal operators of the \( \ell_1 \) norm and the pseudo \( \ell_0 \) norm, respectively.

III. SPARSE LOGISTIC REGRESSION WITH WEAKLY CONVEX REGULARIZATION

To fit a sparse logistic regression model, we propose to try to solve problem (4) with function \( J \) belonging to the class of weakly convex sparsity inducing functions in Definition 1. Note that when the nonconvexity parameter \( \zeta = 0 \), problem (4) becomes convex and the standard \( \ell_1 \) logistic regression is an instance of it.

A. Nonconvexity

An interesting observation is that at this point problem (4) with the nonconvexity parameter \( \zeta > 0 \) can either be convex or nonconvex, depending on the data matrix

\[
X = \begin{pmatrix} x^{(1)}, & \ldots, & x^{(N)} \end{pmatrix},
\]

the regularization parameter \( \beta \), as well as the nonconvexity parameter \( \zeta \). In the following, from a perspective we have a
conclusion that problem (4) is nonconvex with any $\zeta > 0$ (not necessarily sufficiently large).

**Theorem 1.** If the dimension of the column space of the matrix $X$ is less than $d$, i.e., matrix $X$ does not have full row rank, then problem (4) is nonconvex for any $\zeta > 0$.

**Proof.** The proof is in section VIII-A.

The dimension of the column space of $X$ is less than $d$, if the number of data points $N$ is less than the dimension $d$, which is a typical situation where regularization is needed, or the data points are on a low dimensional subspace in $\mathbb{R}^d$. In this work, we do not require that $X$ has full row rank, so in general the problem (4) that we try to solve is nonconvex.

### B. Local optimality

As revealed in the previous part, problem (4) can easily be nonconvex, so we study its local optimality conditions under some assumptions in this part.

What has already been known is that, for a DC function, its local minimum has to be a critical point [36] which is defined as follows.

**Definition 2.** [36] A point $x^*$ is said to be a critical point of a DC function $g(x)$ - $h(x)$, where $g(x)$ and $h(x)$ are convex, if $\partial g(x^*) \cap \partial h(x^*) = \emptyset$.

When a function is differentiable, the above definition is in consistent with the common definition that a critical point is a point where the derivative equals zero. Consequently, if $\theta^*$ is a local optimum of problem (4), then we at least know that $2\beta \zeta \theta^* \in \partial l(\theta^*)$, which is equivalent to

$$2\zeta \theta^* - \frac{1}{\beta} \nabla l(\theta^*) \in \partial G(\theta^*),$$

and boils down to

$$H'(\theta^*_i) \leq 2\zeta \theta^*_i - \frac{1}{\beta} \nabla l(\theta^*_i) \leq H'(+\theta^*_i)$$

for every $i$.

**Assumption 1.** Assume that the function $H$ in Definition 1 is smooth everywhere except for finite points, and $H''(t)$ and $H''(t)$ exist at every differentiable point.

In the following we have a conclusion on sufficient and necessary conditions on local optimality under Assumption 1.

**Theorem 2.** Suppose that Assumption 1 holds. If for every $\theta_i^*$, $i = 1, \ldots, d$, one of the following conditions holds, then $\theta^*$ is a local optimum of problem (4).

- If $\theta_i^* = 0$, then
  $$-\frac{1}{\beta} \nabla l(\theta^*) \in (F'_-(0), F'_+(0)).$$

- If $\theta_i^* \neq 0$, then
  $$2\zeta \theta_i^* - \frac{1}{\beta} \nabla l(\theta^*_i) = H'(\theta_i^*),$$

and there exists a neighborhood of $\theta_i^*$ such that $H''(\theta_i) = 2\zeta$ for any $\theta_i$ in the neighborhood.

If $\theta^*$ is a local optimum of problem (4), then for every $\theta_i^*$, $i = 1, \ldots, d$, one of the following conditions holds.

- If $\theta_i^* = 0$, then
  $$-\frac{1}{\beta} \nabla l(\theta^*) \in [F'_-(0), F'_+(0)].$$

- If $\theta_i^* \neq 0$, then (12) holds, and both $H'_+(\theta^*_i)$ and $H'_-(\theta^*_i)$ are no less than $2\zeta - 0.25|X|^2/\beta$.

**Proof.** The proof is in section VIII-B.

**Remark 1.** It is proved (in Lemma 1) that under Definition 1 (without Assumption 1) $F'_-(0), F'_+(0)$, and $H'(t)$ for $t \neq 0$ exist, and $F'_+(0) < F'_-(0)$.

The sufficient condition in Theorem 2 indicates that a critical point $\theta^*$ is a local minimum, if (10) is strict for any $\theta_i^* = 0$, and $H''(\theta_i^*) = 0$ for $\theta_i$, in a neighborhood of any $\theta_i^* \neq 0$. Since (12) and (13) imply (9), the necessary condition in Theorem 2 is stronger than the known condition that $\theta^*$ is a critical point. Theorem 2 can also be used to help the choice of the regularization parameter $\beta$ and unveil further properties for some specific choice of the regularizer $J$. These will be discussed in the next part and section V.

### C. Choice of the regularization parameter

In this part, we will show a condition on the choice of the regularization parameter $\beta$ to avoid $0$ to become a local optimum of problem (4). More specifically, in the following, we will show that if $\beta$ is larger than a certain value, then $0$ will be a local optimum of problem (4), and if $\beta$ is smaller than that value, then $0$ will not even be a critical point.

**Theorem 3.** Suppose that the mean of the data points is subtracted from them, i.e., $\sum_{i=0}^{N} x_i = 0$. If

$$\beta < \frac{\left\| \sum_{i=1}^{N} x_i^2 \right\|_\infty}{F'_+(0)},$$

then $0$ is not a critical point of problem (4). If

$$\beta > \frac{\left\| \sum_{i=1}^{N} x_i^2 \right\|_\infty}{F'_+(0)},$$

then $0$ is a local minimum of problem (4).

**Proof.** If Assumption 1 is imposed, the conclusion is a direct result of Theorem 2. Without Assumption 1, the conclusion still holds, and the proof is in section VIII-C.

**IV. A proximal gradient method**

In this section, we try to solve the weakly convex regularized sparse logistic regression problem (4) with any function $J$ satisfying Definition 1. Since the logistic loss $l$ is differentiable and the proximal operator of function $J$ can be well defined, the method that we use is proximal gradient descent, and the iterative update is as follows

$$\theta_{k+1} = \text{prox}_{\alpha_k \beta J}(\theta_k - \alpha_k \nabla l(\theta_k)),$$
where \( \alpha_k > 0 \) is a stepsize, and
\[
\nabla l(\theta_k) = \sum_{i=1}^{N} \left( \sigma \left( \theta_k^T x^{(i)} \right) - y^{(i)} \right) x^{(i)}.
\]

Note that the update (14) of the algorithm is equivalent to solving the following problem
\[
\text{minimize} \quad \alpha_k \beta J(\theta) + \frac{1}{2} \|\theta - \theta_k + \alpha_k \nabla l(\theta_k)\|^2_2,
\]
which is strongly convex for \( \alpha_k \beta \zeta < 1/2 \). The calculation of the proximal operator can be elementwise parallel, in that the function \( J \) is separable across the coordinates according to its definition.

The stepsize \( \alpha_k \) in the algorithm can be chosen as a constant \( \alpha \) or determined by backtracking. In the following we prove its convergence with the two stepsize rules.

**Theorem 4.** For stepsize \( \alpha_k \) chosen from one of the following ways,
- **constant stepsize** \( \alpha_k = \alpha \) and
  \[
  \frac{1}{\alpha} > \max \left( 2 \beta \zeta, \frac{1}{8} \|X\|^2 + \beta \zeta \right);
  \]
- **backtracking stepsize** \( \alpha_k = \eta n \alpha_{k-1} \), where \( \beta \zeta \alpha_0 < 1/2, 0 < \eta < 1 \), and \( n_k \) is the smallest nonnegative integer for the following to hold
  \[
  l(\theta_k) \leq l(\theta_{k-1}) + \langle \theta_k - \theta_{k-1}, \nabla l(\theta_{k-1}) \rangle
  + \frac{1}{2\alpha_k} \|\theta_{k-1} - \theta_k\|^2_2.
  \]

the sequence \( \{\theta_k\} \) generated by the algorithm satisfies the following.
- **Objective function** \( l(\theta_k) + \beta J(\theta_k) \) is monotonically non-increasing and convergent;
- **The update of the iterates converges to 0**, i.e.,
  \[
  \|\theta_k - \theta_{k-1}\|^2_2 \to 0;
  \]
- **The first order necessary local optimality condition will be approached**, i.e., there exists \( g_k \in \partial G(\theta_k) \) for every \( k \) such that
  \[
  \beta g_k - 2 \beta \zeta \theta_k + \nabla l(\theta_k) \to 0. \quad (16)
  \]

**Proof.** The proof is in section VIII-D.

**Remark 2.** If the sequence \( \{\theta_k\} \) has limit points, then the third conclusion means that every limit point of the sequence \( \{\theta_k\} \) is a critical point of the objective function.

A constant stepsize depending on the maximum eigenvalue of the data matrix \( X \) is able to guarantee convergence, but when \( X \) is of huge size or distributed and its eigenvalue is not attainable, a backtracking stepsize which does not depend on such information can be used. It should be noted that the proof is applicable to any convex and Lipchitz differentiable loss function \( l \), and because \( l \) is Lipchitz differentiable, \( n_k \) in the backtracking method always exists. According to Theorem 4, the objective function converges, so we can choose \( \epsilon_{tol} > 0 \) and set the following condition
\[
|l(\theta_{k+1}) + \beta J(\theta_{k+1}) - l(\theta_k) - \beta J(\theta_k)| \leq \epsilon_{tol} \quad (17)
\]
as a stopping criterion, and the algorithm is summarized in Table I.

Compared with the DCA method [27] in which every iteration involves numerically solving the following convex problem on \( \theta \) to obtain an update from \( \theta^k \) to \( \theta^{k+1} \)
\[
\text{minimize} \quad l(\theta) + \beta J(\theta) + \zeta \|\theta\|^2_2 - 2 \beta \zeta \theta^T \theta^k, \quad (18)
\]
the computational burden of the proposed method, which is dominated by the gradient calculation per iteration, is much lower, given that the proximal operator has a closed form expression.

Another way of applying the proximal gradient descent is to decompose the objective function into a convex function \( J(\theta) + \beta \zeta \|\theta\|^2_2 \) and a smooth function \( l(\theta) - \beta \zeta \|\theta\|^2_2 \), and use the results in [37] to ensure the convergence. However, because the proximal operator can be defined for a weakly convex function, there is no need to add the \( \zeta \|\theta\|^2_2 \) term to \( J \), and by using our decomposition and Theorem 4 the guaranteed constant stepsize is larger than that given by [37].

As a first-order method for nonconvex problems, Theorem 4 guarantees that any limit point of its generated sequence is a critical point. Next, we have the following bound on the logistic loss, in order to bound the performance of the proposed method on the training set.

**Proposition 1.** Suppose that \( \tilde{\theta} \) is a minimizer of the logistic loss on the training data, i.e., \( \nabla l(\tilde{\theta}) = 0 \), and \( \theta^\ast \) is a critical point of problem (4), then the following bound holds for the normalized \( \tilde{\theta} = \frac{\|\theta\|}{\|\tilde{\theta}\|} \theta^\ast \)
\[
l(\tilde{\theta}) - l(\theta) \leq \frac{1}{2M} \left( \sum_{j, \theta_j^\ast = 0} \beta F_j^\ast (0) + \sum_{j, \theta_j^\ast \neq 0} \beta F_j^\ast (|\theta_j^\ast|) + \frac{1}{2} \|X\|_{2,1} \right)^2, \quad (19)
\]
where \( M \) is a constant only depending on \( \|\theta\|_2 \).

**Proof.** The proof is in section VIII-E.

Since \( \tilde{\theta} \) has the same signs as \( \theta^\ast \) on all entries, and has the same amplitude as \( \theta \), Proposition 1 shows an upper bound on how well any critical point can fit the training data. From the concavity in Definition 1, we know that \( F^\ast (|\theta_j^\ast|) \leq F^\ast (0) \).

(For the specific \( F \) in the next section, \( F^\ast (|\theta_j^\ast|) < F^\ast (0) \) for \( \theta_j^\ast \neq 0 \).) It should be mentioned that with the \( \ell_1 \) norm regularization the same form of bound can also be derived, but \( F^\ast (|\theta_j^\ast|) = F^\ast (0) \) for any \( \theta_j^\ast \neq 0 \). Therefore, with the nonconvex regularization satisfying Definition 1, the bound (19) can be smaller than that with the \( \ell_1 \) norm.
V. A SPECIFIC WEAKLY CONVEX FUNCTION AND ITERATIVE FIRM-SHRINKAGE METHOD

In this section, we take the weakly convex function $J$ to be the specific one defined by $F$ in (7), in that its proximal operator has a closed form expression (8) that is easy to compute. We will first show a sufficient and necessary condition for local optimality, and then discuss the proximal gradient descent method studied in the previous section for this specific case.

A. Local optimality

With the specific function $J$ defined by $F$ in (7) we have the following conclusion on its local optimality.

Corollary 1. Suppose that function $J$ is defined by $F$ in (7), and $\beta \zeta > 0.125\|X\|^2$. Then $\theta^*$ is a local minimum of problem (4), if and only if one of the following conditions holds for every $\theta^*_j$, $j = 1, \ldots, d$. 

- $\theta^*_j = 0$ and $|\nabla l(\theta^*)_j| < \beta$;
- $|\theta^*_j| > \frac{1}{2\zeta}$ and $\nabla l(\theta^*)_j = 0$.

Proof. The function $J$ defined by $F$ in (7) satisfies Assumption 1, so part of the conclusion is a direct result of Theorem 2. The proof is in section VIII-F.

Remark 3. If the training data points are linearly separable, i.e., there exists $\|\theta\|_2 = 1$ such that $\theta^T x^{(i)} \neq 0$ and

$$y^{(i)} = \begin{cases} 1, & \theta^T x^{(i)} > 0; \\ 0, & \theta^T x^{(i)} < 0, \end{cases}$$

holds for all $i = 1, \ldots, N$, then

$$\nabla l(t\theta) \to 0, \quad t \to +\infty,$$

so we have that

$$\lim_{t \to +\infty} l(t\theta) + \beta J(t\theta) = l^* + \beta J^*$$

is a local optimal value. For such reason in [16] a constraint on a norm of $\theta$ is added in their optimization problem. However, here we note that for a given $t > 0$ the following bound holds

$$0 \leq l(t\theta) + \beta J(t\theta) - l^* - \beta J^* \leq \sum_{i=1}^{N} \exp \left(-t |\theta^T x^{(i)}| \right).$$

The above upper bound decreases exponentially in $t$ to 0, so with a sufficiently large $t$, a point $\theta$ can give an objective value numerically sufficiently close to $l^* + \beta J^*$.

Corollary 1 gives an effective way of checking if a point is a local optimum. Besides, it shows that for any $j$ belonging to the support set of $\theta^*$, the gradient $\nabla l(\theta^*)_j$ vanishes. Such property would be satisfied by the $\ell_0$ pseudo norm but not by the $\ell_1$ norm, and it indicates that the weakly convex function $J$ can induce sparsity and fit the training data well at the same time. The following bound on the logistic loss at any local optimum shows how well it can fit the training data.

Corollary 2. Suppose that function $J$ is defined by $F$ in (7) and $\beta \zeta > 0.125\|X\|^2$. If $\theta^*$ is a local optimum of problem (4), and $\bar{\theta}$ is a minimizer of the logistic loss on the training data, i.e., $\nabla l(\bar{\theta}) = 0$, then we have the following bound

$$l(\theta^*) - l(\bar{\theta}) \leq \beta\|\theta_T\|_1,$$

where $T$ is the support set of $\theta^*$ and $T^c$ is the complement set of $T$.

Proof. From the convexity of function $l$, we have

$$l(\theta^*) - l(\bar{\theta}) \leq \sum_{j \in T} \nabla l(\theta^*)_j(\theta^*_j - \bar{\theta}_j) + \sum_{j \in T^c} \nabla l(\theta^*)_j(-\bar{\theta}_j).$$

Since $T$ is the support set of $\theta^*$, according to Corollary 1, $\nabla l(\theta^*)_j = 0$ for any $j \in T^c$, and $|\nabla l(\theta^*)_j| < \beta$ for any $j \in T$, so we arrive at the result.

The right-hand-side of (20) is bounded by $\beta\|\theta\|_1$, and if the support set $T$ includes the support set of $\bar{\theta}$, then the bound becomes 0, meaning that $\theta^*$ also minimizes the logistic loss. Compared with Proposition 1 which studies the critical points of the general problem, the upper bound is smaller and the normalization of $\theta^*$ is not needed in Corollary 2, in that $\theta^*$ is under stronger assumption as a local optimum with a specific weakly convex $J$.

B. Iterative firm-shrinkage algorithm

When the function $J$ is defined by $F$ in (7), the proximal gradient method in Table I in section IV is instantiated and can be understood as a generalization of the iterative shrinkage-thresholding algorithm (ISTA) used to solve $\ell_1$ regularized least square problems [38], [39]. As the concrete proximal operator defined in (8) has been named as the firm-shrinkage operator, we call the method an iterative firm-shrinkage algorithm (IFSA). In every iteration, the computation complexity of IFSA is the same as that of ISTA.

According to Theorem 4, for IFSA if a constant or backtracking stepsize satisfying Theorem 4 is used, then we know that the objective function is non-increasing and convergent, that the sequence $\|\theta_k - \theta_{k-1}\|_2$ goes to 0, and that any limit point of $\{\theta_k\}$ (if there is any) is a critical point of the objective function.

To accelerate the convergence of a proximal gradient method, the Nesterov acceleration [37], [40] has been used in ISTA [38] and some other proximal gradient methods for nonconvex problems [41], [42]. In this work, we apply such technique into IFSA, and for a steady convergence behavior, in every iteration we check the objective values at both the accelerated and unaccelerated points, and only accept the acceleration when the accelerated point can give a lower objective value. The algorithm is summarized in Table II. According to Theorem 4, by using such acceleration, the objective value is still guaranteed to be nonincreasing and convergent.

For the function $J$ defined by $F$ in (7) or any function of which the proximal operator can be calculated in $O(1)$ time, without the acceleration, the proposed method and the ISTA have the same computational complexity in every iteration. Using the acceleration, compared with the accelerated ISTA called FISTA [37], the proposed method needs an additional evaluation of the objective function in every iteration.
TABLE II

<table>
<thead>
<tr>
<th>Iterative Firm Shrinkage Algorithm with Acceleration.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Input:</strong> initial point ( \theta_0, \alpha_0 &lt; 1/(2\beta\zeta) ) (or ( \alpha ) satisfying (15)).</td>
</tr>
<tr>
<td>( k = 1, ; \ell_1 = 1, ; \theta_1 = \theta_0; )</td>
</tr>
<tr>
<td><strong>Repeat:</strong></td>
</tr>
<tr>
<td>update ( \theta_k = \text{prox}_{\alpha_k\beta J}(\theta_k - \alpha_k\nabla l(\theta_k)) ) according to (8) by constant or backtracking stepsize;</td>
</tr>
<tr>
<td>update ( \theta_{k+1} = \frac{1}{\sqrt{1+4\alpha_k^2}} \left[ \frac{1}{1+\alpha_k^2} \right] \left( \theta_k - \alpha_k\nabla l(\theta_k) \right) + \frac{1}{\sqrt{1+4\alpha_k^2}} \left( \theta_k - \alpha_k\nabla l(\theta_k) \right); )</td>
</tr>
<tr>
<td>if ( l(\theta_k) + \beta J(\theta_k) &lt; l(\theta_{k+1}) + \beta J(\theta_{k+1}); )</td>
</tr>
<tr>
<td>( \theta_{k+1} = \theta_k; )</td>
</tr>
<tr>
<td>( k = k + 1; )</td>
</tr>
<tr>
<td><strong>Until</strong> stopping criterion (17) is satisfied.</td>
</tr>
</tbody>
</table>

VI. NUMERICAL EXPERIMENTS

In this section, we demonstrate numerical results of the weakly convex regularized sparse logistic regression (4) with function \( J \) specifically defined by \( F \) in (7). The solving method IFSA is implemented and tested both with and without the acceleration. As a comparison, we also show results of the \( \ell_1 \) logistic regression, of which there are more than one equivalent forms and we use the one in (5).

A. Randomly generated datasets

1) One example for convergence demonstration and comparison: To begin with, in one example we show the convergence curves of IFSA with different constant stepsizes \( \alpha \), and compare IFSA with the result of DCA [27] solving the same problem (4) and the results of the problem with \( \zeta = 0 \), i.e., the \( \ell_1 \) logistic regression (5), solved by FISTA [37] and by a generic solver SCs interfaced by CVXPY [43].

The dimensions are \( d = 50, \; N = 10000, \; K = 8. \) The data matrix is generated by \( X = AB/\|AB\| \), where \( A \in \mathbb{R}^{50 \times 45} \) and \( B \in \mathbb{R}^{45 \times 1000} \) are Gaussian matrices, so that the data points are in a latent 45-dimensional subspace. The positions of the non-zeros of the ground truth \( \theta^0 \) are uniformly randomly generated, and the amplitudes are uniformly distributed over \([5, 15]\). The label \( y \) is generated according to \( 1(x^T\theta^0 \geq 0) \) so that the data points are linearly separable.

We first sweep the space of the regularization parameter and the nonconvexity parameter for IFSA, and use the optimal choice \( \beta = 10^{-1.25} \) and \( \zeta = 10^{-2} \) to show the convergence curves with different stepsizes \( \alpha \) both with and without the Nesterov acceleration in Fig. 1. To satisfy the convergence condition in Theorem 4, we need \( \alpha < 7.9. \)

The result shows that with a larger stepsize (within the range covered by the theorem) the objective value decreases faster, and that with the acceleration the objective value decreases faster for all the tested stepsizes. All the curves are monotonically non-increasing, which verifies the conclusion in Theorem 4. After 2000 iterations, the results with acceleration obtained by different stepsizes are almost the same. According to such observation, for the accelerated IFSA, in the following subsections, we use an empirical and conventional fixed stepsize 0.1, a small \( \epsilon_{tol} = 10^{-6} \), and a large maximum number of iterations 2000 so that the iterative updates can be sufficient.

Next, we compare IFSA with the reference methods in this example. For the \( \ell_1 \) logistic regression solved by CVXPY, the regularization parameter is swept, and the optimal choice is \( \beta = 10^{-1.25} \), which is also used in FISTA solving the same problem (5). To compare the test error rate during iterations of IFSA with that of FISTA using a constant stepsize, we sweep the stepsize \( \alpha \) for FISTA, and find its optimal choice \( \alpha = 25 \), so the stepsize of IFSA is also taken as 25 in this comparison. The subproblems (18) of DCA are solved by the generic subgradient descent using diminishing stepsizes \( 3/k \), where the initial stepsize 3 is optimally chosen. The test error rates during 1000 iterations and the estimated \( \theta/\|\theta\|_2 \) after 1000 iterations are in Fig. 2.

Firstly, IFSA with the acceleration reaches a lower test error in fewer iterations than DCA and FISTA do, and when the test error of FISTA stops to decrease, that of IFSA continues to decrease further. The running time of 1000 iterations is 0.54s for IFSA (0.07s without the Nesterov acceleration) and 5.63s for DCA (on macos 3.2 GHz Intel Core i5 by Python 2.7.13.), so the proposed IFSA achieves lower test error using less computational time. Secondly, the result of FISTA coincides with the solution obtained by CVXPY for the \( \ell_1 \) logistic regression, and the normalized estimation obtained by the proposed method is better than those by FISTA and CVXPY.

2) Varying nonconvexity and regularization parameters: In the second experiment, we demonstrate the performance under various choices of the parameters \( \zeta \) and \( \beta \). The dimensions are \( d = 50, \; K = 5, \; N = 200. \) The training data \( X \) is randomly generated from i.i.d. normal distribution, and the ground truth \( \theta^0 \) is generated by uniformly randomly choosing
The test error is smaller than the test error with \( \zeta \), which is the logarithm of the test error averaged from 10 independently random experiments, each of which is tested by 1000 random test data points which are generated in the same way as the training data points.

The results show that with a fixed positive \( \beta \), as the value of \( \zeta \) increases from 0, the test error first decreases and then increases, and there is always a choice of \( \zeta > 0 \) under which the test error is smaller than the test error with \( \zeta = 0 \) which is the \( \ell_1 \) logistic regression. The results in Fig. 3 verify our motivation that weakly convex regularized logistic regression can better estimate the sparse model than the \( \ell_1 \) logistic regression and enhance test accuracy.

3) Non-separable datasets: In the above two settings the data points are linearly separable, while in this part we will show test errors when the training data points are not linearly separable. To be specific, the label \( y \) of a training data \( x \) is generated by \( y = 1(x^T\theta^0 + n \geq 0) \), where \( n \) is an additive noise generated from the Gaussian distribution \( \mathcal{N}(0, \epsilon^2) \). The training data matrix \( X \), the ground truth model vector \( \theta^0 \), and the test data points are randomly generated in the same way as the second experiment.

In the training process, under every noise level \( \epsilon \), we learned \( \theta \) under various \( \beta \) from \( 10^{-3} \) to 10 and \( \zeta \) from 0 to 10, and we repeated it 10 times with different random data to take the averaged test errors for every pair of \( \zeta \) and \( \beta \). For every noise level, we then took the lowest error rate obtained with \( \zeta = 0 \) as the error rate of \( \ell_1 \) logistic regression and the lowest error rate obtained with \( \zeta > 0 \) as the error rate of weakly convex logistic regression. The results are summarized in Table III.

From the results, we can see that, as the noise level increases, the error rates of both methods increase, but under every tested noise level the weakly convex logistic regression can achieve lower error rate than the \( \ell_1 \) logistic regression.

### TABLE III

<table>
<thead>
<tr>
<th>noise level</th>
<th>( \ell_1 ) logistic regression</th>
<th>weakly convex logistic regression</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>3.31%</td>
<td>0.92%</td>
</tr>
<tr>
<td>0.03</td>
<td>3.27%</td>
<td>1.48%</td>
</tr>
<tr>
<td>0.05</td>
<td>3.91%</td>
<td>1.85%</td>
</tr>
<tr>
<td>0.07</td>
<td>4.90%</td>
<td>3.39%</td>
</tr>
<tr>
<td>0.3</td>
<td>13.70%</td>
<td>12.37%</td>
</tr>
<tr>
<td>0.5</td>
<td>21.47%</td>
<td>19.70%</td>
</tr>
</tbody>
</table>

### B. Real datasets

In this part, we show experimental results of the weakly convex logistic regression on real datasets which have been commonly used in \( \ell_1 \) logistic regression, and compare the classification error rates between these two methods. The first dataset is a spam email database [44], where the number of features 57 is far smaller than the number of data points 4601, of which 20% are used for training. The classification result indicates whether or not an email is a spam. The second one is an arrhythmia dataset [44] which has 279 features and 452 data points, of which 80% are used for training. The two classes refer to the normal and arrhythmia, and missing values in the features are filled with zeros. The third one is a gene database from tumor and normal tissues [45], where the number of features 2000 is far larger than the number of data points 62, of which 40% are used for training. The classification result is whether or not it is a tumor tissue. The test and training data points are randomly separated.

In the experiments, we first run \( \ell_1 \) logistic regression with various \( \beta \) on the training data and use cross validation on the test data to find the best value of \( \beta \) and the corresponding error rate. Then we run the IFSAs for weakly convex logistic regression under various \( \beta \) and \( \zeta \), and still use cross validation to get the best \( \beta \), \( \zeta \) and its error rate.

Results in Table IV show that, for the first dataset, where the number of features is far smaller than the number of training data, the weakly convex logistic regression has a little improvement over the \( \ell_1 \) regularized logistic regression. For the second and the third datasets, where the number of training data points is inadequate compared to the number of features, the improvement achieved by weakly convex logistic regression is more significant.

### VII. Conclusion and Future Work

In this work we study weakly convex regularized sparse logistic regression. For a class of weakly convex sparsity inducing functions, we first prove that the optimization problem
with such functions as regularizers is in general nonconvex, and then we study its local optimality conditions, as well as the choice of the regularization parameter to exclude a trivial solution. Even though the general problem is nonconvex, a solution method based on the proximal gradient descent is devised with theoretical convergence analysis, which includes the convergence of the objective, the convergence of the increment of the generated sequence, and a bound on the logistic loss at any limit point of the sequence. Then the general framework is applied to a specific weakly convex function, and a necessary and sufficient local optimality condition is unveiled together with a bound on the logistic loss at any local optimum. The solution method for this specific case named iterative firm-shrinkage algorithm is implemented with the Nesterov acceleration. Its effectiveness is demonstrated in experiments by both randomly generated data and real datasets.

There can be several directions to extend this work, such as using only parts of the data in every iteration by applying stochastic proximal gradient method. More generally, weakly convex regularization could be used in other machine learning problems to fit sparse models.

VIII. APPENDIX

A. Proof of Theorem 1

Let us first prove a Lemma on some properties of the weakly convex function in Definition 1.

**Lemma 1.** If function $F$ (together with function $J$) satisfies Definition 1, then $F$ is differentiable on $(0, \infty)$, $F'_+(0) = -F'_-(0) < 0$, and $F(t)/t$ is non-increasing on $(0, \infty)$.

**Proof.** From Definition 1, $H(t) = F(t) + \zeta t^2$ is convex, so $H'_-(t), H'_+(t), F'_+(t)$, and $F'_-(t)$ exist and satisfy

$$F'_-(t) + 2\zeta t = H'_-(t) \leq H'_+(t) \leq F'_+(t) + 2\zeta t.$$  

Because $F$ is concave on $(0, \infty)$, we have $F'_-(t) \geq F'_+(t)$ on $(0, \infty)$, so $F$ is differentiable on $(0, \infty)$.

If we take the derivative of $F(t)/t$ on $(0, \infty)$, then

$$\left(\frac{F(t)}{t}\right)' = \frac{F'(t)t - F(t)}{t^2}.$$  

From the concavity in Definition 1, we know that for all $t > 0$,

$$0 \leq F(t) - F'(t)t,$$

so the derivative of $F(t)/t$ on $(0, \infty)$ is non-positive, and $F(t)/t$ is non-increasing on $(0, \infty)$.

Because of the above monotonicity, for all $t_1 > t_2 > 0$

$$\frac{F(t_1)}{t_1} \leq \frac{F(t_2)}{t_2} \leq F'_+(0).$$  

Because $F$ is not linear, there must exist $t_0 > 0$ such that for all $t > t_0$

$$0 \leq \frac{F(t)}{t} \leq \frac{F(t_0)}{t_0} < F'_+(0).$$  

(21)

Thus, together with that $F$ is even, we have $F'_+(0) = -F'_-(0) > 0$.

If problem (4) is convex, i.e., its objective is convex, then the subgradient set of the objective at any $\theta$ is

$$\nabla l(\theta) + \beta \partial G(\theta) - 2\beta\zeta \theta,$$

where the plus and minus signs operate on every element of the set $\partial G(\theta)$. We denote $\partial G(\theta) = -2\zeta \theta$ as $\partial J(\theta)$. If (4) is convex, the following must hold for any $\theta_0, \theta$ and any subgradient $g \in \nabla l(\theta_0) + \beta \partial J(\theta_0)$

$$l(\theta) + \beta J(\theta) \geq l(\theta_0) + \beta J(\theta_0) + g^T(\theta - \theta_0).$$  

(22)

In the following we will construct $\theta_0, \theta$, and $g$ such that (22) does not hold.

Because $X$ does not have full row rank, there exists $u \neq 0$ such that $u^TX = 0$. For such $u$, we have that

$$l(tu) = l(0)$$

holds for any $t$.

Next we will find $t > 0$ and $g \in \nabla l(0) + \beta \partial J(0)$, such that

$$\beta J(tu) < \beta J(0) + (tu - 0)^T(g - \nabla l(0)).$$

Note that $J(0) = 0$. Suppose that for any $t > 0$ and any $h = (g - \nabla l(0))/\beta \in \partial J(0)$, which is equivalent to $h_i \in [F'_-(0), F'_+(0)]$, the following holds

$$\sum_{i=1}^d F'(tu_i) = J(tu) \geq tu^T h = \sum_{i=1}^d h_iu_i.$$  

(23)

Because of (21) in Lemma 1, for every $u_i > 0$, there is a $t_i > 0$ such that for all $t > t_i$

$$F(tu_i) < F'_+(0)tu_i,$$

and for every $u_i < 0$ there is a $t_i > 0$ such that for all $t > t_i$

$$F(tu_i) < tF'_-(0)u_i.$$  

Thus, we have that the following holds for all $t > \max_i(t_i)$

$$\sum_{i=1}^d tF'(tu_i) < \sum_{u_i > 0} tF'_+(0)u_i + \sum_{u_i < 0} tF'_-(0)u_i.$$  

In (23), by taking $h_i = F'_+(0)$ when $u_i > 0$ and $h_i = F'_-(0)$ when $u_i < 0$, we have a contradiction. Now we have proved that

$$l(tu) + \beta J(tu) < 0$$

holds for some $t > 0$ and any $g \in \nabla l(0) + \beta \partial J(0)$, so (22) does not hold for all $\theta_0$ and $\theta$, and the objective in (4) is nonconvex.

B. Proof of Theorem 2

1) Proof of the sufficient condition: By definition of local optimality, $\theta^*$ is a local optimal point of problem (4), if and only if

$$\beta\zeta\|\theta\|^2 + l(\theta^*) + \beta G(\theta^*) - \beta\zeta\|\theta^*\|^2 \leq l(\theta) + \beta G(\theta).$$

$$\nabla l(\theta) + \beta \partial G(\theta) - 2\beta\zeta \theta,$$

where the plus and minus signs operate on every element of the set $\partial G(\theta)$. We denote $\partial G(\theta) = -2\zeta \theta$ as $\partial J(\theta)$. If (4) is convex, the following must hold for any $\theta_0, \theta$ and any subgradient $g \in \nabla l(\theta_0) + \beta \partial J(\theta_0)$

$$l(\theta) + \beta J(\theta) \geq l(\theta_0) + \beta J(\theta_0) + g^T(\theta - \theta_0).$$  

(22)
holds for any \( \theta \) in a small neighborhood of \( \theta^* \). It can be equivalently written as

\[
 l(\theta^*) + \beta G(\theta^*) \leq l(\theta) + \beta G(\theta) + 2\beta \zeta \langle \theta^* - \theta, \theta^* \rangle + \frac{\beta \zeta^2 \| \theta^* - \theta \|^2}{2}. \tag{24}
\]

We will prove that if every \( \theta^*_i \) for \( i = 1, \ldots, d \) satisfies either of the two conditions, then \( \theta^* \) is a local optimum, i.e., (24) holds in a small neighborhood of \( \theta^* \).

If the first condition (11) holds for \( \theta^*_i = 0 \), then together with the convexity inequalities that

\[
 H(0) \leq H(\theta_i) + (0 - \theta_i)H'_i(0)
\]

we know that

\[
 H(0) \leq H(\theta_i) + (0 - \theta_i)(0 - \nabla l(\theta^*)_{i}/\beta - \zeta(0 - \theta_i))
\]

holds for all \( \theta_i \) such that

\[
 0 \leq \zeta(0 - \theta_i) \leq 0 - \nabla l(\theta^*)_{i}/\beta - H'_i(0),
\]

and all \( \theta_i \) such that

\[
 0 \leq \zeta(\theta_i - 0) \leq H'_i(0) - 0 + \nabla l(\theta^*)_{i}/\beta.
\]

Therefore, there exists \( \delta_i > 0 \) such that for all \( (\theta_i - \theta^*_i)^2 \leq \delta_i \) we have

\[
 H(\theta^*_i) \leq H(\theta_i) + (\theta_i - \theta^*_i)(2\zeta \theta^*_i - \nabla l(\theta^*)_{i}/\beta - \zeta(\theta_i - \theta^*_i))^2.
\]

(25)

If the second condition in Theorem 2 holds for \( \theta^*_i \), then according to the second order Taylor expansion, there exists \( \delta_i > 0 \) such that the following holds in a small neighborhood \( (\theta_i - \theta^*_i)^2 \leq \delta_i \),

\[
 H(\theta^*_i) + (\theta_i - \theta^*_i)H'(\theta^*_i) + \zeta(\theta_i - \theta^*_i)^2 = H(\theta_i).
\]

Together with (12), we also arrive at (25). Thus, (25) holds for every coordinate with some \( \delta_i > 0 \), and we have that

\[
 G(\theta^*) \leq G(\theta) + (2\zeta \theta^* - \nabla l(\theta^*)/\beta)\langle \theta - \theta^* \rangle - \zeta \| \theta - \theta^* \|^2
\]

holds for \( \theta \) in a neighborhood \( \| \theta - \theta^* \|^2 \leq \delta \) with \( \delta = \min \delta_i > 0 \). Together with the fact that \( l \) is convex, we have that (24) holds in such neighborhood.

2) Proof of the necessary condition: Since \( \theta^* \) is a local optimum, it is a critical point, so we need only to prove that for a critical point \( \theta^* \), if there exists \( \theta^*_i \) that does not satisfy the two conditions, then such a critical point \( \theta^* \) cannot be a local optimum. Suppose that there is a \( \theta^*_i \) at which \( F \) (and also \( H \)) is differentiable, and one of \( H''_i(\theta^*_i) \) and \( H'_i(\theta^*_i) \) is less than \( 2\zeta - 0.25\|X\|^2/\beta \). Without loss of generality, we assume that

\[
 H''_i(\theta^*_i) < 2\zeta - 0.25\|X\|^2/\beta. \tag{26}
\]

Then we take

\[
 \theta = (\theta^*_1, \ldots, \theta^*_i - t, \theta^*_i + t, \ldots, \theta^*_d)
\]

for \( t > 0 \), so

\[
 \theta - \theta^* = (0, \ldots, 0, -t, 0, \ldots, 0),
\]

and

\[
 l(\theta) + \beta G(\theta) - l(\theta^*) - \beta G(\theta^*)
\]

\[
 = l(\theta) - l(\theta^*) + \beta H(\theta^*_i - t) - \beta H(\theta^*_i)
\]

\[
 \leq -\nabla l(\theta_i) + \beta H(\theta^*_i - t) - \beta H(\theta^*_i)
\]

\[
 \leq -\nabla l(\theta_i) + \beta H(\theta^*_i - t) + \beta H(\theta^*_i) \leq \beta H(\theta^*_i) t^2 + \beta \zeta(0 - \theta_i)^2
\]

\[
 = -\nabla l(\theta^*_i) t + \beta H(\theta^*_i) t + \beta \zeta t^2,
\]

where the first inequality holds in that the Lipchitz constant of \( \nabla l \) is 0.25\|X\|^2, and the second inequality holds for small positive \( t \) according to the second order Taylor expansion and (26). Consequently, together with (12) we have that in any small neighborhood there exists \( \theta \) such that

\[
 l(\theta^*) + \beta G(\theta^*) \geq l(\theta) + \beta G(\theta) + 2\beta \zeta (\theta^* - \theta)^T \theta^* - \beta \zeta \| \theta - \theta^* \|^2
\]

holds, so \( \theta^* \) cannot be a local optimum.

C. Proof of Theorem 3

If \( \nabla l(0) \notin \beta \partial G(0) \), then according to (9), \( 0 \) is not a critical point of the objective in problem (4), so \( 0 \) is not a local optimum of problem (4). Since \( \sum_{i=0}^{N} x^{(i)} = 0 \), the condition becomes

\[
 -\nabla l(0) = \sum_{y^{(i)} = 1} x^{(i)} \notin \beta \partial G(0),
\]

which is equivalent to that

\[
 \beta F'_+(0) < \left\| \sum_{y^{(i)} = 1} x^{(i)} \right\|_{\infty}.
\]

On the other side, if

\[
 \beta F'_+(0) > \left\| \sum_{y^{(i)} = 1} x^{(i)} \right\|_{\infty},
\]

then \( 0 \) is a local optimum of problem (4). To prove this, first notice that the following holds for any \( \theta \) and any \( g \in \partial G(0) \)

\[
 l(\theta) \geq l(0) + \theta^T \nabla l(0), \quad J(\theta) \geq G(0) + \theta^T \zeta - \zeta \| \theta \|^2.
\]

Thus, for any \( \theta \) and any \( g \in \partial G(0) \)

\[
 l(\theta) + \beta J(\theta) \geq l(0) + \beta J(0) + \beta \theta^T \nabla l(0) + \beta \theta^T g - \beta \zeta \| \theta \|^2.
\]

Then, for every \( \theta \) we take the following \( g \in \partial G(0) \)

\[
 g_i = \begin{cases} F'_+(0), & \theta_i \geq 0 \\ F'_-(0), & \theta_i < 0, \end{cases}
\]

so we have

\[
 \theta^T (\nabla l(0) + \beta g) - \beta \zeta \| \theta \|^2
\]

\[
 = \sum_{\theta_i > 0} \theta_i (\nabla l(0) + \beta F'_+(0)) - \beta \zeta \theta_i^2
\]

\[
 + \sum_{\theta_i < 0} \theta_i (\nabla l(0) + \beta F'_-(0)) - \beta \zeta \theta_i^2 > 0,
\]

where the last inequality holds for all \( \| \theta \|_2 \) small enough, in that \( \nabla l(0) + \beta F'_+(0) \) is strictly positive and \( \nabla l(0) + \beta F'_-(0) \) is strictly negative.
\( \beta F'_\beta(0) \) is strictly negative. Therefore, we have that in a small neighborhood of 0, the following holds
\[
l(\theta) + \beta J(\theta) > l(0) + \beta J(0),
\]
which means that 0 is a local minimum in this case. According to Lemma 1 we have that \( F'_\beta(0) > 0 \), so we reach the conclusions in Theorem 3.

\section*{D. Proof of Theorem 4}
The techniques used in this proof are similar to the ones in [38], [46]. To begin with, we are going to prove that the objective function in problem (4) is able to decrease monotonically during the iterations. First of all, notice the fact that the gradient of function \( l \) is Lipchitz continuous, which gives the following inequality according to the Lipchitz property
\[
l(\theta_k) + \beta J(\theta_k) \leq l(\theta_{k-1}) + \nabla l(\theta_{k-1})^T(\theta_k - \theta_{k-1}) + L\|\theta_k - \theta_{k-1}\|^2/2 + \beta J(\theta_k),
\]
where \( L \) is the Lipchitz constant. If the backtracking stepsize is used, then we have
\[
l(\theta_k) + \beta J(\theta_k) \leq l(\theta_{k-1}) + \nabla l(\theta_{k-1})^T(\theta_k - \theta_{k-1}) + \|\theta_k - \theta_{k-1}\|^2/(2\alpha_k) + \beta J(\theta_k). \tag{27}
\]
Secondly, according to our update rule, \( \theta_k \) minimizes the following function of \( u \)
\[
\beta J(u) + \frac{1}{2\alpha_k}\|u - \theta_{k-1}\|^2 + \nabla l^T(\theta_{k-1})(u - \theta_{k-1}),
\]
which is \( \mu_k \)-strongly convex given that
\[
\mu_k/2 = 1/(2\alpha_k) - \beta \zeta > 0.
\]
Thus, we have that
\[
\beta J(\theta_k) + \nabla l^T(\theta_{k-1})(\theta_k - \theta_{k-1}) + \frac{1}{2\alpha_k}\|\theta_k - \theta_{k-1}\|^2/2 
\leq \beta J(\theta_{k-1}) - \mu_k\|\theta_k - \theta_{k-1}\|^2/2. \tag{29}
\]
Combining (29) with (28) yields that
\[
l(\theta_k) + \beta J(\theta_k) \leq l(\theta_{k-1}) + \beta J(\theta_{k-1}) - \frac{\mu_k}{2}\|\theta_k - \theta_{k-1}\|^2/2,
\]
which means that the objective function is monotonically non-increasing with the backtracking stepsize. For the constant stepsize \( \alpha = \alpha_k \), combining (29) with (27), we have
\[
\beta J(\theta_k) + l(\theta_k) \leq l(\theta_{k-1}) + (L - 1/\alpha - \mu)\|\theta_k - \theta_{k-1}\|^2/2 + \beta J(\theta_{k-1}).
\]
For \( \alpha \) small enough such that
\[
L - 1/\alpha - \mu < 0,
\]
we have that the objective function is non-increasing. Because
\[
L \leq \sup_\theta \|H(\theta)\| = \|X\|^2/4,
\]
a sufficient condition for
\[
L - 1/\alpha - \mu = L - 2/\alpha + 2\beta \zeta < 0
\]
is that
\[
\|X\|^2/8 + \beta \zeta < 1/\alpha,
\]
which is the requirement of the constant stepsize. Together with the fact that the objective function is lower bounded, we have proved the first claim in the Theorem 4.

The second claim can be seen from that
\[
0 \leq (1/\alpha + \mu - L)\|\theta_k - \theta_{k-1}\|^2/2 
\leq l(\theta_{k-1}) + \beta J(\theta_{k-1}) - (l(\theta_k) + \beta J(\theta_k))
\]
holds for the constant stepsize, and
\[
0 \leq \mu_k\|\theta_k - \theta_{k-1}\|^2/2 
\leq l(\theta_{k-1}) + \beta J(\theta_{k-1}) - (l(\theta_k) + \beta J(\theta_k))
\]
holds for the backtracking stepsize. Note that \( \mu_k \) is nondecreasing during the iterations, so \( \|\theta_k - \theta_{k-1}\|_2 \) converges to 0 in both cases.

To see the third claim, remind that the update rule indicates that
\[
0 \in \beta \partial G(\theta)_k - 2\beta \zeta \theta_k + \nabla l(\theta_{k-1}) + \frac{1}{\alpha_k}(\theta_k - \theta_{k-1}),
\]
so there exists \( g_k \in G(\theta)_k \) for every \( k \) such that
\[
\beta g_k - 2\beta \zeta \theta_k + \nabla l(\theta_k) = \beta g_k - 2\beta \zeta \theta_k + \nabla l(\theta_{k-1}) + \nabla l(\theta_k) - \nabla l(\theta_{k-1}) 
\]
\[
= -\frac{\theta_k - \theta_{k-1}}{\alpha_k} + \nabla l(\theta_k) - \nabla l(\theta_{k-1}).
\]
Since \( \nabla l \) is continuous, as \( \theta_k - \theta_{k-1} \to 0 \), for a constant stepsize (16) holds. For the backtracking stepsize, \( 1/\alpha_k \) is nondecreasing, so (16) also holds.

\section*{E. Proof of Proposition 1}
Since \( \hat{\theta} \) is a minimizer of the logistic loss on the training data, according to [47], we have that
\[
l(\hat{\theta}) - l(\hat{\theta}) \leq \frac{1}{2M}\left\|\nabla l(\hat{\theta})\right\|_2^2,
\]
where \( M \) is a positive constant dependent on \( \|\hat{\theta}\|_2 \) and the data. Then we have the following bound on the \( \ell_2 \) norm of the gradient at \( \hat{\theta} \).
\[
\left\|\nabla l(\hat{\theta})\right\|_2 \leq \left\|\nabla l(\theta^*)\right\|_2 + \left\{\sum_{i=1}^N x^{(i)}(\sigma(\hat{\theta}^T x^{(i)}) - \sigma(\theta^*^T x^{(i)}))\right\}_2 
\leq \left(\sum_{i=1}^N \frac{\beta F'_\beta(0) + \sum_{j, \theta^*_j \neq 0} \beta F'_\beta(\theta^*_j)}{\|X\|_{2,1}/2}\right).
satisfies the condition that both $H''_1$ and $H''_2$ are no less than $\zeta$ when and only when $|t| > 1/(2\zeta)$, where $H'(t) = 2t$. From Theorem 2, together with the assumption that $\beta \zeta > 0.125\|X\|^2$, we can directly have a necessary condition that every $\theta^*_j$ satisfies one of the following

1. $\theta^*_j = 0$ and $|\nabla l(\theta^*_j)| \leq \beta$;
2. $|\theta^*_j| > \frac{1}{2\beta}$ and $\nabla l(\theta^*_j) = 0$.

Thus, to prove Corollary 1, we only need to show that if there is a coordinate $i$ such that $\theta^*_i = 0$ and $|\nabla l(\theta^*_i)| \leq \beta$, then $\theta^*$ is not a local optimum. To see this, we take

$$\theta = (\theta_1^*, \ldots, \theta_{i-1}^*, -t, \theta_{i+1}^*, \ldots, \theta_n^*),$$

and we will prove that for any $0 < \delta < 1/(2\zeta)$, there exists $t$ such that $0 < |t| \leq \delta$ and the local optimality inequality (24) does not hold. According to the Lipchitz condition, we have

$$l(\theta) - l(\theta^*) - l'\nabla l(\theta^*) + 0.125\|X\|^2|t|^2.$$ 

Since $\theta^*_i = 0$ and $|\theta_i| \leq 1/(2\zeta)$, we have

$$G(\theta^*_i) - G(\theta) = -|t|.$$ 

If $\nabla l(\theta^*_i) = \beta$, then for any $t > 0$

$$l(\theta) - l(\theta^*) + \beta G(\theta) - \beta G(\theta^*) + 2\beta \zeta (\theta^* - \theta)^T \theta^* - \beta \zeta \|\theta^* - \theta\|^2 = l(\theta) - l(\theta^*) + \beta|t| - \beta \zeta t^2 \leq 0.125\|X\|^2 t^2 - \beta \zeta t^2 < 0.$$ 

If $\nabla l(\theta^*_i) = -\beta$, then for any $t < 0$ the above inequality holds. Therefore, we prove that the local optimality inequality (24) cannot hold within any small neighborhood of $\theta^*$, and $\theta^*$ is not a local optimum.

### References


---

**Xinyue Shen** (S’14) is a PhD student in Electronic Engineering Department in Tsinghua University, Beijing, China, and her advisor is Professor Yuantao Gu. She received the B.E. degree with honor from Electronic Engineering Department in Tsinghua University in 2013. She was a visiting student researcher at Stanford University during 2015 to 2016. Her research interests include nonconvex optimization, sparse signal recovery, sparse learning, and subspace recovery. She received the Zhang Si-Ying (CCDC) Outstanding Youth Paper Award in 2017, and the national scholarship for graduate student in 2015.

---

**Yuantao Gu** (S’02–M’03–SM’17) received the B.E. degree from Xi’an Jiaotong University in 1998, and the Ph.D. degree with honor from Tsinghua University in 2003, both in Electronic Engineering. He joined the faculty of Tsinghua University in 2003 and is now a Tenured Associate Professor with Department of Electronic Engineering. He was a visiting scientist at Microsoft Research Asia during 2005 to 2006, Research Laboratory of Electronics at Massachusetts Institute of Technology during 2012 to 2013, and Department of Electrical Engineering and Computer Science at the University of Michigan in Ann Arbor during 2015. His research interests include high-dimensional statistics, sparse signal recovery, temporal-space and graph signal processing, related topics in wireless communications and information networks. He has been an Associate Editor of the IEEE Transactions on Signal Processing since 2015, a Handling Editor for EURASIP Digital Signal Processing from 2015 to 2017, and an Elected Member of the IEEE Signal Processing Theory and Methods (SPTM) Technical Committee since 2017. He is a senior member of IEEE.

He received the Best Paper Award of IEEE Global Conference on Signal and Information Processing (GlobalSIP) in 2015, the Award for Best Presentation of Journal Paper of IEEE International Conference on Signal and Information Processing (ChinaSIP) in 2015, and Zhang Si-Ying (CCDC) Outstanding Youth Paper Award (with his student) in 2017.