NONCONVEX SPARSE LOGISTIC REGRESSION VIA PROXIMAL GRADIENT DESCENT

Xinyue Shen and Yuantao Gu

Electronic Engineering Department, Tsinghua University, Beijing 100084, China

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 ℓ_0 pseudo norm is able to better induce sparsity than the commonly used ℓ_1 norm. For a class of weakly convex sparsity inducing functions, despite the nonconvexity, the algorithm proposed to solve the problem is based on proximal gradient descent, which allows the use of convergence acceleration techniques and stochastic gradient. Then the general framework is applied to a specific weakly convex function, and the solution method is instantiated as an iterative firm-shrinkage algorithm, of which the effectiveness is demonstrated in numerical experiments.

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

1. INTRODUCTION

1.1. Background

Logistic regression is a widely used supervised machine learning method for classification. It learns a neutral hyperplane in the feature space according to a probabilistic model, and classifies test data correspondingly. The output 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 has been applied to text classification [2], gene selection and microarray analysis [3,4], combinatorial chemistry [5], image analysis [6,7], etc.

In a classification problem training data $\{(\mathbf{x}^{(i)}, y^{(i)}), i = 1, \dots, N\}$ are given, where every point $\mathbf{x}^{(i)} \in \mathbb{R}^d$ is a feature vector, 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 $\mathbf{x}$ is as the following

$$\begin{aligned} p(y = 1|\mathbf{x}; \boldsymbol{\theta}) &= \sigma(\boldsymbol{\theta}^{\mathrm{T}}\mathbf{x}) = \frac{1}{1 + \exp(-\boldsymbol{\theta}^{\mathrm{T}}\mathbf{x})} \\ p(y = 0|\mathbf{x}; \boldsymbol{\theta}) &= 1 - \sigma(\boldsymbol{\theta}^{\mathrm{T}}\mathbf{x}), \end{aligned} \tag{1}$$

where $\boldsymbol{\theta} \in \mathbb{R}^d$ is the model parameter to be learned, $\sigma(\cdot)$ is the sigmoid function. When $\mathbf{x}^{\mathrm{T}}\boldsymbol{\theta} = 0$, the probability of having either label is 0.5, so $\boldsymbol{\theta}$ is a normal vector of a neutral hyperplane.

In sparse logistic regression, the model parameter $\boldsymbol{\theta}$ is assumed to be sparse, i.e., the dimension d can be large, and $\boldsymbol{\theta}$ is assumed to have only a few non-zero elements. An element $\boldsymbol{\theta}_j = 0$ means that the jth feature does not have influence on the classification result, so sparse logistic regression tries to find a few features that are relevant to the classification results from a large number of features. It is also a way of alleviating over-fitting and enhancing classification accuracy on test data.

As a convex function that induces sparsity, ℓ_1 norm has been widely used as the regularization in sparse logistic regression, and the optimization problem is as the following

$$\text{minimize} \quad l(\boldsymbol{\theta}) + \beta\|\boldsymbol{\theta}\|_1, \tag{2}$$

where $\boldsymbol{\theta}$ is the variable, $\beta > 0$ is a parameter balancing the sparsity and the error on the training data, and l is the logistic loss

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{N} -\log p(y^{(i)}|\mathbf{x}^{(i)}; \boldsymbol{\theta}). \tag{3}$$

Problem (2) is convex but nondifferentiable, and several specialized solution methods have been proposed [2, 8–12]. Once we obtain a solution $\hat{\boldsymbol{\theta}}$, given a new feature vector $\mathbf{x}$, we can predict the probability of the two labels by (1) and take the one with higher probability.

1.2. Contribution

In this work we propose to use a weakly convex function as the regularization in sparse logistic regression. The idea is inspired by the relation between logistic regression and one-bit compressed sensing [13] and results indicating that weakly convex functions are able to better induce sparsity than the ℓ_1 norm [14–16]. We formulate the problem as a weakly convex sparsity inducing function regularized nonconvex program, and a solution method based on proximal gradient descent is devised, where the usage of Nesterov acceleration and stochastic gradient is also considered. Then we apply the framework to a specific weakly convex regularizer, and the effectiveness of the model and the method is verified in numerical experiments.

1.3. Related Works

Despite that in general nonconvex optimization is hard to solve, nonconvex regularization has been extensively studied to induce sparsity in machine learning for feature selection and other sparsity related topics such as compressed sensing.

The work [17] 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 used as an approximation of the ℓ_0 pseudo norm for feature selection in logistic regression and support vector machines (SVMs) in [18–20]. Their solution

This work was partially supported by National Natural Science Foundation of China (NSFC 61531166005, 61571263), the National Key Research and Development Program of China (Project No. 2016YFE0201900, 2017YFC0403600), and Tsinghua University Initiative Scientific Research Program (Grant 2014Z01005). The corresponding author of this work is Y. Gu (E-mail: gyt@tsinghua.edu.cn). An extended manuscript of this work is available at [1].

978-1-5386-4658-8/18/$31.00 ©2018 IEEE

methods are based on the difference of convex functions algorithm (DCA), where each iteration involves solving a convex program. In this work, our regularizer also belongs to the 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 given that the proximal operator of the weakly convex function admits an easy computation.

From the perspective of reconstructing $\boldsymbol{\theta}$, one-bit compressed sensing [21] studies a similar problem, where a sparse vector $\boldsymbol{\theta}$ (or its normalization $\boldsymbol{\theta}/\|\boldsymbol{\theta}\|_2$) is to be estimated from several one-bit measurements $y^{(i)} = \mathbf{1}(\boldsymbol{\theta}^{\mathrm{T}}\mathbf{x}^{(i)} \geq 0)$, where $\mathbf{1}(\cdot)$ takes the value 1 when the condition holds and 0 if not, and compressed sensing [22] studies the recovery of a sparse $\boldsymbol{\theta}$ from several linear measurements $y^{(i)} = \boldsymbol{\theta}^{\mathrm{T}}\mathbf{x}^{(i)}$. Nonconvex regularizations have been used to promote sparsity in both compressed sensing [15,16,23,24] and one-bit compressed sensing [25]. These studies have shown that, despite that nonconvex optimization problems are 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 ℓ_1 regularization, both theoretically and numerically, in terms of required number of measurements and robustness against noise.

2. PRELIMINARIES

A class of weakly convex functions has been proposed to induce sparsity [15]. The definition is as the following.

Definition 1. *[15] The weakly convex sparsity inducing function J is defined to be separable $J(\mathbf{x}) = \sum_{i=1}^{n} F(|\mathbf{x}_i|)$, where the function $F : \mathbb{R} \to \mathbb{R}_+$ satisfies the following.*

- *Function F is even and not identically zero, and $F(0) = 0$;*
- *Function F is non-decreasing on $[0, \infty)$;*
- *Function $t \mapsto F(t)/t$ is nonincreasing on $(0, \infty)$;*
- *Function F is weakly convex [26] on $[0, \infty)$ with nonconvexity parameter $\zeta > 0$, i.e., ζ is the smallest positive scalar such that the function $F(t) + \zeta t^2$ is convex.*

According to the definition J is weakly convex, and $J(\mathbf{x}) + \zeta\|\mathbf{x}\|_2^2$ is a convex function, so J belongs to the class of DC functions [18]. 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.

The proximal operator of function J with parameter β is defined as

$$\operatorname{prox}_{\beta J}(\mathbf{v}) = \operatorname{argmin} \beta J(\mathbf{x}) + \frac{1}{2}\|\mathbf{x} - \mathbf{v}\|_2^2, \tag{4}$$

where the minimization is with respect to $\mathbf{x}$. If β is small enough so that $\beta\zeta < \frac{1}{2}$, then the objective function in (4) is strongly convex, and the minimizer is unique. For some weakly convex functions, their proximal operators allow easy computation. For instance, the following F known as minimax concave penalty (MCP) proposed in [27] satisfies Definition 1

$$F(t) = \begin{cases} |t| - \zeta t^2 & |t| \leq \frac{1}{2\zeta} \\ \frac{1}{4\zeta} & |t| > \frac{1}{2\zeta} \end{cases}. \tag{5}$$

Its proximal operator with $\beta\zeta < 1/2$ can be explicitly written as

$$\operatorname{prox}_{\beta F}(v) = \begin{cases} 0 & |v| < \beta \\ \frac{v - \beta \operatorname{sign}(v)}{1 - 2\beta\zeta} & \beta \leq |v| \leq \frac{1}{2\zeta} \\ v & |v| > \frac{1}{2\zeta} \end{cases}, \tag{6}$$

and is also known as the firm shrinkage operator [28].

3. SPARSE LOGISTIC REGRESSION WITH WEAKLY CONVEX REGULARIZATION

We propose to use the following problem, in which function J belongs to the class of weakly convex sparsity inducing functions in Definition 1, to learn the parameter $\boldsymbol{\theta}$ in sparse logistic regression

$$\text{minimize} \quad l(\boldsymbol{\theta}) + \beta J(\boldsymbol{\theta}), \tag{7}$$

where the variable is $\boldsymbol{\theta} \in \mathbb{R}^d$, $\beta > 0$ is a regularization parameter, and l is the logistic loss (3). Note that when the nonconvexity parameter $\zeta = 0$, the problem becomes convex and the standard ℓ_1 logistic regression is an instance of it. When $\zeta > 0$, it is not straight forward to see if problem (7) is convex for any training data and any choice of β and the nonconvexity parameter ζ. In the extended version of this work [1], we proved that when the data matrix

$$\mathbf{X} = \left(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}\right)$$

does not have full row rank, then problem (7) is nonconvex for any ζ and β.

As for the solving method, since the logistic loss l is differentiable and the proximal operator of function J can be well defined, we use the proximal gradient descent method, and the iterative update is as the following

$$\boldsymbol{\theta}_{k+1} = \operatorname{prox}_{\alpha_k \beta J}(\boldsymbol{\theta}_k - \alpha_k \nabla l(\boldsymbol{\theta}_k)), \tag{8}$$

where $\alpha_k > 0$ is a stepsize, and the gradient is calculated as follows

$$\nabla l(\boldsymbol{\theta}_k) = \sum_{i=1}^{N} \left(\sigma\left(\boldsymbol{\theta}_k^{\mathrm{T}}\mathbf{x}^{(i)}\right) - y^{(i)}\right)\mathbf{x}^{(i)}. \tag{9}$$

Note that the update (8) of the algorithm is equivalent to solving the following problem

$$\text{minimize} \quad \alpha_k \beta J(\boldsymbol{\theta}) + \frac{1}{2}\left\|\boldsymbol{\theta} - \boldsymbol{\theta}_k + \alpha_k \nabla l(\boldsymbol{\theta}_k)\right\|_2^2,$$

which is strongly convex for $\alpha_k\beta\zeta < 1/2$ and separable across the d coordinates.

We have proved in the extended version [1] that if the stepsize α_k satisfies one of the following

- constant stepsize $\alpha_k = \alpha$ and
$$\frac{1}{\alpha} > \max\left(2\beta\zeta, \frac{1}{8}\|\mathbf{X}\|^2 + \beta\zeta\right); \tag{10}$$
- backtracking stepsize $\alpha_k = \eta^{n_k}\alpha_{k-1}$, where $\beta\zeta\alpha_0 < 1/2$, $0 < \eta < 1$, and n_k is the smallest nonnegative integer for the following to hold
$$l(\boldsymbol{\theta}_k) \leq l(\boldsymbol{\theta}_{k-1}) + \langle \boldsymbol{\theta}_k - \boldsymbol{\theta}_{k-1}, \nabla l(\boldsymbol{\theta}_{k-1})\rangle + \frac{\|\boldsymbol{\theta}_{k-1} - \boldsymbol{\theta}_k\|_2^2}{2\alpha_k};$$

then the sequence $\{\boldsymbol{\theta}_k\}$ generated by the algorithm satisfies that the objective function is non-increasing and convergent, and that $\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_{k-1}\|_2 \to 0$, so we can have the following stopping criterion

$$|l(\boldsymbol{\theta}_{k+1}) + \beta J(\boldsymbol{\theta}_{k+1}) - l(\boldsymbol{\theta}_k) - \beta J(\boldsymbol{\theta}_k)| \leq \epsilon_{\text{tol}}. \tag{11}$$

The algorithm can be summarized in Table 1.

Though algorithm 1 is convergent, proximal gradient methods have been known to suffer from slow convergence, and the Nesterov

Table 1. Proximal gradient descent for weakly convex regularized logistic regression.

```
Input: initial point $\boldsymbol{\theta}_0$, $\alpha_0 < 1/(2\beta\zeta)$ (or $\alpha$ satisfying (10)), $\epsilon_{\mathrm{tol}} > 0$.
$k = 0$;
Repeat:
    update $\boldsymbol{\theta}_{k+1}$ by (8) using constant or backtracking stepsize;
    $k = k + 1$;
Until stopping criterion (11) is satisfied.
```

Table 2. Accelerated proximal gradient descent for weakly convex regularized logistic regression.

```
Input: initial point $\hat{\boldsymbol{\theta}}_0$, $\alpha_0 < 1/(2\beta\zeta)$ (or $\alpha$ satisfying (10)).
$k = 1, t_1 = 1, \boldsymbol{\theta}_1 = \hat{\boldsymbol{\theta}}_0$;
Repeat:
    update $\hat{\boldsymbol{\theta}}_k = \mathrm{prox}_{\alpha_k \beta J}(\boldsymbol{\theta}_k - \alpha_k \nabla l(\boldsymbol{\theta}_k))$
        by constant or backtracking stepsize;
    update $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$;
    update $\boldsymbol{\theta}_{k+1} = \hat{\boldsymbol{\theta}}_k + \left(\frac{t_k - 1}{t_{k+1}}\right)(\hat{\boldsymbol{\theta}}_k - \hat{\boldsymbol{\theta}}_{k-1})$;
    $k = k + 1$;
Until maximum number of iterations is reached.
```

acceleration [29] has been used in proximal gradient based algorithms such as ISTA [30]. Such technique is also applicable in our case, so we have the accelerated algorithm summarized in Table 2.

Another variation in implementation is that the batch gradient $\nabla l(\boldsymbol{\theta}_k)$ used in every iteration can be replaced with a stochastic gradient

$$\nabla \tilde{l}(\boldsymbol{\theta}_k) = N\left(\sigma\left(\boldsymbol{\theta}_k^{\mathrm{T}} \mathbf{x}^{(i)}\right) - y^{(i)}\right)\mathbf{x}^{(i)}, \tag{12}$$

where $\mathbf{x}^{(i)}$ is randomly chosen among all training samples. We say that (12) is one gradient calculation, which only uses one data point, and correspondingly the batch gradient (9) which uses all training samples needs N gradient calculations. Diminishing stepsize is required due to the gradient noise caused by the randomness, and a common choice of the stepsize is as follows [31]

$$\alpha_k = \alpha_0/(1 + k\gamma\alpha_0),$$

where γ and α_0 are constant parameters. Stochastic gradient is widely used in learning [31, 32], in that with only one data sample and one gradient calculation per iteration, methods using stochastic gradient can reach a lower error rate with fewer gradient calculations compared with methods using batch gradient.

3.1. A Specific Case: Iterative Firm-shrinkage Method

In this section, we take the weakly convex function J to be the specific one defined by F in (5), in that its proximal operator has a closed form expression (6) that only needs parallel scalar multiplications to compute.

When the function J is defined by F in (5), the proximal gradient method in Table 1 is instantiated and can be understood as a generalization of the iterative shrinkage-thresholding algorithm (ISTA) used to solve ℓ_1 regularized least square problems [30, 33]. As the concrete proximal operator defined in (6) has been named as the firm-shrinkage operator, we call the method an *iterative firm-shrinkage algorithm* (IFSA). According to the theorems proved in the extended paper [1], for IFSA, if a constant or backtracking stepsize is used, then we know that the objective function is non-increasing and convergent, that the update $\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_{k-1}\|_2$ goes to 0, and that any limit point of $\{\boldsymbol{\theta}_k\}$ (if there is any) is a critical point of the objective function.

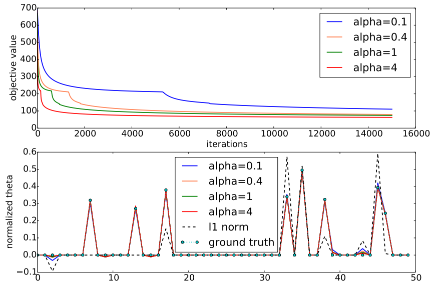

Fig. 1. An example without acceleration. *Upper:* objective value during iterations. *Lower:* estimated $\boldsymbol{\theta}/\|\boldsymbol{\theta}\|_2$ and the ground truth.

4. NUMERICAL EXPERIMENTS

In this section, we demonstrate numerical results of the weakly convex regularized sparse logistic regression (7) with function J specifically defined by F in (5). The solving method IFSA is implemented and tested both with and without acceleration. As a comparison, we also show results of the ℓ_1 logistic regression (2), for which there are many algorithms, and we simply use a generic solver SCS interfaced by CVXPY [34], in that in such comparison we focus on replacing the ℓ_1 norm with a weakly convex function.

4.1. One Example for Convergence Demonstration

To begin with, for one example we show the convergence curves and the estimated $\boldsymbol{\theta}/\|\boldsymbol{\theta}\|_2$ of our algorithm with different constant stepsizes. The dimensions are $d = 50$, $N = 1000$, and $K = 8$. The data matrix is generated by $\mathbf{X} = \mathbf{AB}/\|\mathbf{AB}\|$, where $\mathbf{A} \in \mathbb{R}^{50\times 45}$ and $\mathbf{B} \in \mathbb{R}^{45\times 1000}$ are Gaussian matrices, so that the data points are in a latent 45-dimensional subspace. The positions of the nonzeros of the ground truth $\boldsymbol{\theta}_0$ are uniformly randomly generated, and the amplitudes are uniformly distributed over $[5, 15]$. The label y is generated according to $\mathbf{1}(\boldsymbol{\theta}_0^{\mathrm{T}}\mathbf{x} \geq 0)$, so that the data points are linearly separable. We set the regularization parameter $\beta = 1.2$ and the nonconvexity parameter $\zeta = 0.1$.

The results without and with the Nesterov acceleration are shown in Fig. 1 and Fig. 2, respectively. Fig. 1 shows that, with larger stepsize (within the range), the objective function decreases faster, and when terminated at the given number of iterations the estimated $\boldsymbol{\theta}/\|\boldsymbol{\theta}\|_2$ becomes closer to the ground truth. Fig. 2 shows that with the acceleration the objective function decreases faster for all the tested stepsizes, and that the estimations of $\boldsymbol{\theta}/\|\boldsymbol{\theta}\|_2$ are better than the estimation obtained from the ℓ_1 logistic regression.

4.2. Varying Nonconvexity and Regularization Parameters

In the second experiment, we demonstrate the performance under various choices of the parameters ζ and β. The dimensions are $d = 50$, $K = 5$, and $N = 200$. The training data $\mathbf{X}$ is randomly generated from i.i.d. normal distribution, and the ground truth $\boldsymbol{\theta}_0$

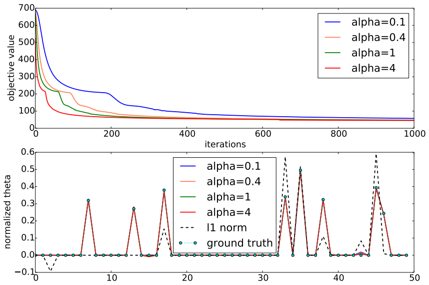

Fig. 2. An example with acceleration. *Upper:* objective value during iterations. *Lower:* estimated $\boldsymbol{\theta}/\|\boldsymbol{\theta}\|_2$ and the ground truth.

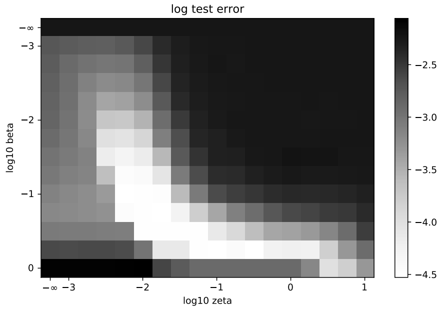

Fig. 3. Logarithm of test error under various values of ζ and β.

is generated by uniformly randomly choosing K nonzero elements with i.i.d. normal distribution. The stepsize of IFSA is chosen as 0.1. The labels are generated so that the data points are linearly separable. The results are in Fig. 3, where the horizontal axis is the logarithm of ζ, the vertical axis is the logarithm of β, and the gray scale represents the logarithm of the test error averaged from 10 independent experiments, each of which is tested by 1000 random test samples generated in the same way as the training data.

The results show that with a fixed value of β no less than $10^{-2.8}$, as the value of ζ 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 $\zeta = 0$. The results in Fig. 3 verify our motivation that weakly convex regularized logistic regression can better estimate the sparse model than the ℓ_1 logistic regression and enhance test accuracy.

4.3. Non-separable Noisy Dataset

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 $\mathbf{x}$ is generated by $y = \mathbf{1}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\theta} + n \geq 0)$, where n is an additive noise generated from the Gaussian distribution $\mathcal{N}(0, \epsilon^2)$. The training data matrix $\mathbf{X}$, the ground truth model vector $\boldsymbol{\theta}_0$, and the test data points are randomly generated in the same way as the second experiment.

In the training process, under every noise level ϵ, we learned $\boldsymbol{\theta}$ under various β from 10^{-3} to 10 and ζ from 0 to 10, and we repeat it 10 times with different random data to take the averaged test errors for every pair of ζ and β. For every noise level, we then took the lowest error rate obtained with $\zeta = 0$ as the error rate of ℓ_1 logistic regression and the lowest error rate obtained with $\zeta > 0$ as the error rate of weakly convex logistic regression. The results are in Table 3. From the results we can see that, under every tested noise level the weakly convex logistic regression can achieve lower error rate than the ℓ_1 logistic regression.

Table 3. Test error for non-separable data.

noise level	ℓ_1 logistic regression	weakly convex logistic regression
0.01	3.31%	0.92%
0.03	3.27%	1.48%
0.05	3.91%	1.85%
0.07	4.90%	3.39%
0.3	13.70%	12.37%
0.5	21.47%	19.70%

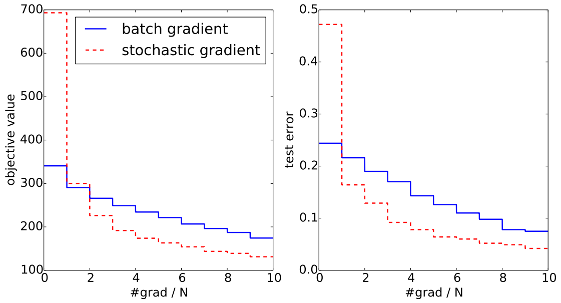

Fig. 4. An example with stochastic gradient.

4.4. Stochastic Gradient versus Batch Gradient

Before concluding this work, we show a numerical example in which stochastic gradient is used in the accelerated IFSA. The data generation is the same as section 4.1, and we set $\beta = 1.2$ and $\zeta = 0.1$. For batch gradient we run 10 iterations with $\alpha = 15$ for fast convergence, and for stochastic gradient we run $10N$ iterations with $\alpha_0 = 0.0005$. The objective value and the test error rate are calculated every N iterations for stochastic gradient and every iteration for the full gradient, and the curves are in Fig. 4. Please notice that the horizontal axis is proportional to the number of gradient calculations, which is 1 for stochastic gradient and N for full gradient per iteration. The result in Fig. 4 shows that using stochastic gradient in the accelerated IFSA is able to achieve lower objective value and error rate, when the number of gradient calculations is limited.

5. CONCLUSION AND FUTURE WORK

In this work we study weakly convex regularized sparse logistic regression. For a class of weakly convex sparsity inducing functions, even though the problem is nonconvex, a solution method based on the proximal gradient descent is devised with possible usage of Nesterov acceleration and stochastic gradient. Then the general framework is applied to a specific weakly convex function, and the solution method for this specific case named as iterative firm-shrinkage algorithm is implemented. Its effectiveness is demonstrated in numerical experiments. There can be future works on the theoretical analysis of stochastic proximal gradient descent for this weakly convex regularized nonconvex program. More generally, weakly convex regularization could be used in other machine learning problems to fit sparse models.

6. REFERENCES

[1] X. Shen and Y. Gu, "Nonconvex sparse logistic regression with weakly convex regularization," *arXiv preprint arXiv:1708.02059*, 2017.

[2] A. Genkin, D. D. Lewis, and David Madigan, "Large-scale bayesian logistic regression for text categorization," *Technometrics*, 2006.

[3] J. Zhu and T. Hastie, "Classification of gene microarrays by penalized logistic regression," *Biostatistics*, vol. 5, no. 3, pp. 427–43, 2004.

[4] G. C. Cawley and N. L.C. Talbot, "Gene selection in cancer classification using sparse logistic regression with bayesian regularization," *Bioinformatics*, vol. 22, no. 19, pp. 2348–2355, 9 2006.

[5] M. T. D. Cronin, A. O. Aptula, J. C. Dearden, J. C. Duffy, T. I. Netzeva, H. Patel, P. H. Rowe, T. W. Schultz, A. P. Worth, and K. Voutzoulidis, "Structure-based classification of antibacterial activity," *Journal of Chemical Information and Computer Sciences*, vol. 42, no. 4, pp. 869, 2002.

[6] R. F. Murray, "Classification images: A review.," *Journal of Vision*, vol. 11, no. 5, pp. 74–76, 2011.

[7] G. Ciocca, C. Cusano, and R. Schettini, "Image orientation detection using LBP-based features and logistic regression," *Multimedia Tools and Applications*, vol. 74, no. 9, pp. 3013–3034, 2015.

[8] S. Lee, H. Lee, P. Abbeel, and A. Y. Ng, "Efficient l1 regularized logistic regression," in *AAAI*, 2006.

[9] V. Roth, "The generalized lasso: a wrapper approach to," *IEEE Trans Neural Netw*, vol. 15, no. 1, pp. 16 – 28, 2002.

[10] S. K. Shevade and S. S. Keerthi, "A simple and efficient algorithm for gene selection using sparse logistic regression," *Bioinformatics*, vol. 19, no. 17, pp. 2246–2253, 2003.

[11] K. Koh, S. Kim, and S. Boyd, "An interior-point method for large-scale l1-regularized logistic regression," *Journal of Machine Learning Research*, vol. 8, no. 4, pp. 1519–1555, 2007.

[12] S. Perkins and J. Theiler, "Online feature selection using grafting," in *In International Conference on Machine Learning*. 2003, pp. 592–599, ACM Press.

[13] Y. Plan and R. Vershynin, "Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach," *IEEE Transactions on Information Theory*, vol. 59, no. 1, pp. 482–494, 2013.

[14] R. Chartrand, "Exact reconstruction of sparse signals via nonconvex minimization," *IEEE Signal Processing Letters*, vol. 14, no. 10, pp. 707–710, 2007.

[15] L. Chen and Y. Gu, "The convergence guarantees of a nonconvex approach for sparse recovery," *IEEE Transactions on Signal Processing*, vol. 62, no. 15, pp. 3754–3767, 2014.

[16] X. Shen, L. Chen, Y. Gu, and H. C. So, "Square-root lasso with nonconvex regularization: An admm approach," *IEEE Signal Processing Letters*, vol. 23, no. 7, pp. 934–938, 2016.

[17] P. Loh and M. J. Wainwright, "Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima," in *Advances in Neural Information Processing Systems 26*, pp. 476–484. 2013.

[18] H. A. Le Thi, H. M. Le, V. V. Nguyen, and T. Pham Dinh, "A DC programming approach for feature selection in support vector machines learning," *Advances in Data Analysis and Classification*, vol. 2, no. 3, pp. 259–278, Dec 2008.

[19] S. O. Cheng and H. A. Le Thi, "Learning sparse classifiers with difference of convex functions algorithms," *Optimization Methods and Software*, vol. 28, no. 4, pp. 830–854, 2013.

[20] L. Yang and Y. Qian, "A sparse logistic regression framework by difference of convex functions programming," *Applied Intelligence*, vol. 45, no. 2, pp. 241–254, Sep 2016.

[21] P. T. Boufounos and R. G. Baraniuk, "1-bit compressive sensing," in *2008 42nd Annual Conference on Information Sciences and Systems*, March 2008, pp. 16–21.

[22] D.L. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[23] Laming Chen and Yuantao Gu, "The convergence guarantees of a non-convex approach for sparse recovery using regularized least squares," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 3350–3354.

[24] Laming Chen and Yuantao Gu, "Fast sparse recovery via nonconvex optimization," in *2015 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 2015.

[25] R. Zhu and Q. Gu, "Towards a lower sample complexity for robust one-bit compressed sensing," in *Proceedings of the 32nd International Conference on Machine Learning (ICML15)*, David Blei and Francis Bach, Eds., 2015, pp. 739–747.

[26] J. Vial, "Strong and weak convexity of sets and functions," *Mathematics of Operations Research*, vol. 8, no. 2, pp. 231–259, 1983.

[27] C. Zhang, "Nearly unbiased variable selection under minimax concave penalty," *The Annals of Statistics*, vol. 38, no. 2, pp. 894–942, 2010.

[28] H. Gao and A. G. Bruce, "Waveshrink with firm shrinkage," *Statistica Sinica*, vol. 7, no. 4, pp. 855–874, 1997.

[29] Y. E. Nesterov, "A method for solving the convex programming problem with convergence rate $O(1/k^2)$," pp. 372–376, 1983.

[30] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[31] A. Nitanda, "Stochastic proximal gradient descent with acceleration techniques," in *Advances in Neural Information Processing Systems 27*, pp. 1574–1582. 2014.

[32] L. Rosasco, S. Villa, and Bang Công Vũ, "Convergence of stochastic proximal gradient algorithm," *arXiv preprint arXiv:1403.5074*, 2014.

[33] E. T. Hale, W. Yin, and Y. Zhang, "A fixed-point continuation method for l1-regularized minimization with applications to compressed sensing," *CAAM Technical report TR07-07*, 2007.

[34] S. Diamond and S. Boyd, "CVXPY: A Python-embedded modeling language for convex optimization," *Journal of Machine Learning Research*, vol. 17, no. 83, pp. 1–5, 2016.