Nonconvex Sparse Logistic Regression via Proximal Gradient Descent

Xinyue Shen Yuantao Gu

Tsinghua University, Beijing, China

ICASSP
April 20, 2018

Outline

Outline

Logistic Regression

- training data $\{(\mathbf{x}^{(i)}, y^{(i)}), i = 1, \ldots, N\}$, feature $\mathbf{x}^{(i)} \in \mathbf{R}^d$, class label $y^{(i)}$
- two-class $y^{(i)} \in \{0, 1\}$ with assumption

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

- $\boldsymbol{\theta} \in \mathbf{R}^d$ is the model parameter to be learned
- $\boldsymbol{\theta}$ gives a neutral hyperplane
- minimize the negative log-likelihood

$$\text{minimize} \quad l(\boldsymbol{\theta}), \tag{1}$$

where l is the logistic loss

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{N} -\log p(y^{(i)}|\mathbf{x}^{(i)}; \boldsymbol{\theta}) \tag{2}$$

Sparse Logistic Regression

- $\boldsymbol{\theta}$ is assumed to be sparse
  - large dimension d, small number of non-zeros
  - $\theta_j = 0$ means that the jth feature is irrelevant
  - feature selection
- alleviate over-fitting and enhance test accuracy
- ℓ_1 norm regularized sparse logistic regression

$$\text{minimize} \quad l(\boldsymbol{\theta}) + \beta \|\boldsymbol{\theta}\|_1 \tag{3}$$

  - use ℓ_1 norm (convex relaxation of ℓ_0 pseudo norm) to induce sparsity
  - $\beta > 0$ is the regularization parameter
  - convex program

Nonconvex Regularization for Sparsity: Related Works

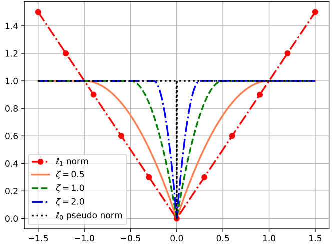
better approximation of the ℓ_0 pseudo norm

- a class of nonconvex regularized logistic regression with constraints on norms (Loh 2013)
- difference of convex (DC) functions regularized logistic regression (LeThi 2008, Cheng 2013, Yang 2016)
- other nonconvex regularizations in compressed sensing (Tropp 2006, Chartrand 2007, Candès 2008, Foucart 2009, Hyder 2010, Voronin, 2013, Chen 2014, Zhu 2015)

Weakly Convex Sparsity Inducing Functions

Definition 1 Weakly convex sparsity inducing function J is defined to be separable $J(\mathbf{x}) = \sum_{j=1}^{d} F(|x_j|)$, where $F : \mathbf{R} \to \mathbf{R}_+$

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



Outline

Weakly Convex Regularized Sparse Logistic Regression

$$\text{minimize} \quad l(\boldsymbol{\theta}) + \beta J(\boldsymbol{\theta}) \tag{4}$$

- nonconvex function J follows Definition 1
- difference of convex program
- problem (4) is nonconvex for any $\zeta > 0$ and $\beta > 0$, if

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

  does not have full row rank
- ℓ_1 logistic regression is an instance when $\zeta = 0$

Proximal Gradient Descent Solving Method

- J is nonconvex, but $J(\mathbf{x}) + \zeta\|\mathbf{x}\|_2^2$ is convex
- proximal operator well-defined when $\alpha\beta\zeta < \frac{1}{2}$

$$\operatorname{prox}_{\alpha\beta \mathrm{J}}(\mathbf{v}) = \operatorname{argmin}\ \alpha\beta \mathrm{J}(\mathbf{x}) + \frac{1}{2}\|\mathbf{x} - \mathbf{v}\|_2^2 \tag{5}$$

  - separable across the d coordinates
  - can have analytical solution via low-cost computation
- proximal gradient descent update with stepsize $\alpha_k > 0$

$$\boldsymbol{\theta}_{k+1} = \operatorname{prox}_{\alpha_{\mathrm{k}}\beta \mathrm{J}}(\boldsymbol{\theta}_{\mathrm{k}} - \alpha_{\mathrm{k}}\nabla \mathrm{l}(\boldsymbol{\theta}_{\mathrm{k}})) \tag{6}$$

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

Convergence

stepsize α_k satisfies one of the following

- constant stepsize $\alpha_k = \alpha$ and

$$\alpha < 1/\max\left(2\beta\zeta, \frac{1}{8}\|\mathbf{X}\|^2 + \beta\zeta\right) \tag{7}$$

- 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

$$l(\boldsymbol{\theta}_k) \leq l(\boldsymbol{\theta}_{k-1}) + \langle \boldsymbol{\theta}_k - \boldsymbol{\theta}_{k-1}, \nabla l(\boldsymbol{\theta}_{k-1})\rangle + \frac{\|\boldsymbol{\theta}_{k-1} - \boldsymbol{\theta}_k\|_2^2}{2\alpha_k}$$

then

- $l(\boldsymbol{\theta}_k) + \beta J(\boldsymbol{\theta}_k)$ monotonically nonincreasing and convergent
- $\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_{k-1}\|_2 \to 0$
- any limit point of $\{\boldsymbol{\theta}_k\}$ is a critical point of the objective function

Proximal Gradient Descent Solving Method

Table: Proximal gradient descent for problem (4)

Input: initial point $\boldsymbol{\theta}_0$, $\alpha_0 < 1/(2\beta\zeta)$ (or α satisfying (7)), $\epsilon_{\rm tol} > 0$.

$k := 0$;
Repeat:
  update $\boldsymbol{\theta}_{k+1}$ by (6) using constant or backtracking stepsize;
  $k := k + 1$;
Until $\lvert l(\boldsymbol{\theta}_{k+1}) + \beta J(\boldsymbol{\theta}_{k+1}) - l(\boldsymbol{\theta}_k) - \beta J(\boldsymbol{\theta}_k)\rvert \leq \epsilon_{\rm tol}$

- apply Nesterov acceleration
- apply stochastic gradient

Variation 1: Acceleration

Table: Accelerated proximal gradient descent for problem (4).

Input: initial point $\hat{\boldsymbol{\theta}}_0$, $\alpha_0 < 1/(2\beta\zeta)$ (or α satisfying (7)).
$k := 1$, $t_1 = 1$, $\boldsymbol{\theta}_1 = \hat{\boldsymbol{\theta}}_0$;
Repeat:
update $\hat{\boldsymbol{\theta}}_k = \mathrm{prox}_{\alpha_\mathrm{k}\beta\mathrm{J}}(\boldsymbol{\theta}_\mathrm{k} - \alpha_\mathrm{k}\nabla\mathrm{l}(\boldsymbol{\theta}_\mathrm{k}))$
according to (10) by constant or backtracking stepsize;
update $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$;
update $\boldsymbol{\theta}_{k+1} = \hat{\boldsymbol{\theta}}_k + \left(\frac{t_k-1}{t_{k+1}}\right)(\hat{\boldsymbol{\theta}}_k - \hat{\boldsymbol{\theta}}_{k-1})$;
if $l(\hat{\boldsymbol{\theta}}_k) + \beta J(\hat{\boldsymbol{\theta}}_k) < l(\boldsymbol{\theta}_{k+1}) + \beta J(\boldsymbol{\theta}_{k+1})$:
$\boldsymbol{\theta}_{k+1} = \hat{\boldsymbol{\theta}}_k$;
$k := k+1$;
Until $\lvert l(\boldsymbol{\theta}_{k+1}) + \beta J(\boldsymbol{\theta}_{k+1}) - l(\boldsymbol{\theta}_k) - \beta J(\boldsymbol{\theta}_k)\rvert \leq \epsilon_{\mathrm{tol}}$

Variation 2: Stochastic Gradient

use a stochastic gradient instead of the batch gradient $\nabla l(\boldsymbol{\theta}_k)$

$$\nabla \tilde{l}(\boldsymbol{\theta}_k) = N\left(\sigma\left(\boldsymbol{\theta}_k^{\mathrm{T}}\mathbf{x}^{(i)}\right) - y^{(i)}\right)\mathbf{x}^{(i)} \tag{8}$$

- $\mathbf{x}^{(i)}$ randomly chosen among all training samples
- one gradient calculation using only one data point
- a common choice of diminishing stepsize

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

where γ and α_0 are constant

A Specific Case

- F in J defined by minimax concave penalty (MCP)

$$F(t) = \begin{cases} |t| - \zeta t^2 & |t| \leq \frac{1}{2\zeta} \\ \frac{1}{4\zeta} & |t| > \frac{1}{2\zeta} \end{cases} \tag{9}$$

- proximal operator also known as firm-shrinkage operator

$$\mathrm{prox}_{\beta F}(v) = \begin{cases} 0 & |v| < \beta \\ \frac{v - \beta \mathrm{sign}(v)}{1 - 2\beta\zeta} & \beta \leq |v| \leq \frac{1}{2\zeta} \\ v & |v| > \frac{1}{2\zeta} \end{cases} \tag{10}$$

- method instantiated as Iterative Firm-shrinkage Algorithm (IFSA), above conclusions applicable
- a generalization of iterative shrinkage-thresholding algorithm (ISTA)

Outline

Sparse Logistic Regression

Weakly Convex Regularized Sparse Logistic Regression

Numerical Experiments

Conclusion

Convergence Demonstration and Comparison

- $d = 50$, $N = 1000$, $K = 8$ non-zeros in ground truth $\boldsymbol{\theta}^0$
- randomly generated $\mathbf{x}^{(i)}$ and $\boldsymbol{\theta}^0$, $y^{(i)} = \mathbf{1}((\mathbf{x}^{(i)})^{\mathrm{T}}\boldsymbol{\theta}^0 \geq 0)$
- choose $\beta = 10^{-1.25}$ and $\zeta = 10^{-2}$
- $\alpha < 7.9$ according to the convergence theorem

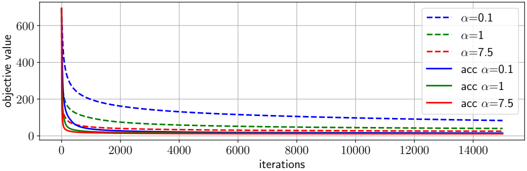
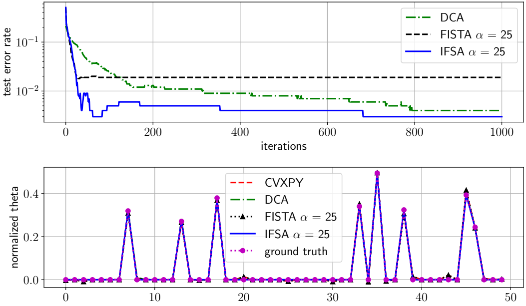


Figure: Convergence curves of IFSA in an example. Dashed lines: without acceleration. Solid lines: with acceleration.

Convergence Demonstration and Comparison

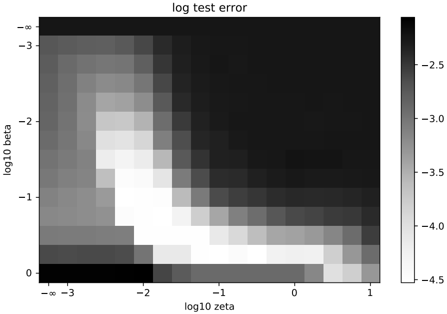
- weakly convex logistic regression by accelerated IFSA and DCA
- ℓ_1 logistic regression with optimal choice $\beta = 10^{-1.25}$
    - by CVXPY and FISTA (with optimal stepsize $\alpha = 25$)



- running time is 0.54s for IFSA and 5.63s for DCA

Varying Nonconvexity and Regularization Parameters

- $d = 50$, $K = 5$, and $N = 200$
- randomly generated $\mathbf{x}^{(i)}$ and $\boldsymbol{\theta}^0$, $y^{(i)} = \mathbf{1}((\mathbf{x}^{(i)})^{\mathrm{T}}\boldsymbol{\theta}^0 \geq 0)$
- accelerated IFSA stepsize $\alpha = 0.1$
- averaged from 20 experiments



Non-separable Noisy Dataset

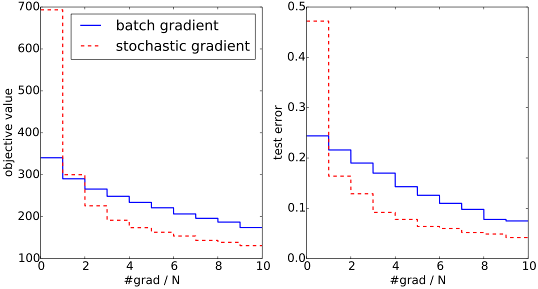
- $y = \mathbf{1}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\theta}^0 + n \geq 0)$, where $n \sim \mathcal{N}(0, \epsilon^2)$
- find optimal parameters among $\beta \in [10^{-3}, 10]$ and $\zeta \in [0, 10]$
- averaged from 10 experiments

Table: 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%

Stochastic Gradient versus Batch Gradient

- $\beta = 1.2$ and $\zeta = 0.1$
- batch gradient $\alpha = 15$, stochastic gradient $\alpha_0 = 0.0005$
- #grad per iteration: 1 for stochastic gradient, N for full gradient



Outline

Conclusion

- a class of weakly convex sparsity inducing functions as regularizer in sparse logistic regression
- solution method for this class of nonconvex problem
  - based on the proximal gradient descent
  - low computational complexity
  - convergence guarantee
- usage of Nesterov acceleration and stochastic gradient
- applied to a specific weakly convex function
  - iterative firm-shrinkage algorithm
- achieve lower test error within less running time in experiments
- code available at:
  `http://gu.ee.tsinghua.edu.cn/publications`
- extended version: "Nonconvex Sparse Logistic Regression with Weakly Convex Regularization". X. Shen, Y. Gu. IEEE Transactions on Signal Processing, 2018, accepted.

Thanks for listening!