Accelerated Gradient Algorithms for Variable Selection with Nonconvex Penalties

Kai Yang, Masoud Asgharian, Sahir Bhatnagar

McGill University

Research supported by NSERC and MITACS

Summary

- **What:** Nonconvex penalties are popular in high-dimensional variable selection, largely due to their oracle property. Nonconvexity proposes a problem here as most optimization methods are designed for convex problems only. Meanwhile, high statistical dimensionality of the data suggests that only first-order methods are adequate for the problem.
- **Current approach:** Currently, first-order methods used in statistical computing literature are mainly ISTA. Coordinate descent has also been used, but it lacks proof of convergence and rate of convergence.
- **Problem:** Developing first-order methods for statistical learning objectives, while attaining global convergence.
- **Our Solution:** Adapt a generalization of Nesterov's accelerated gradient method to nonconvex problems and derive optimal parameter settings.

Problem

- Variable selection has a sigificant application in bioinformatics
- Nonconvex penalties such as SCAD, MCP process *Oracle property*, which makes them a better choice in general for variable selection comparing to LASSO
- However, the nonconvexity and nonsmoothness proposes a challenge for statistical computing, *particularly for high-dimensional data*
- When the statistical dimensionality of the data goes high (such as the number of SNPs), second-order methods are usually not efficient due to the need to evaluate secant conditions per step and the lack of global convergence when not performing line search
- Coordinate descent usually lacks of proof for global convergence. Furthermore, rate of convergence for such methods is usually not feasible to establish.
- ISTA is a "smoothing" version of gradient descent. However, for ill-conditioned problems, ISTA will not be efficient.
- FISTA was proposed to solve this issue: Nesterov's accelerated gradient (AG) was used instead of gradient descent
- However, Nesterov's AG does not achieve global convergence for nonconvex problems

Generalization of Nesterov's AG

- A recent paper by Ghadimi and Lan [1] generalized Nesterov's AG to nonconvex composite settings
- The optimization problem is denoted by:

$$\min_{x\in\mathbb{R}^n} \Psi(x) + \chi(x), \quad \Psi(x) := f(x) + h(x),$$

- $f \in \mathcal{C}^{1,1}_{L_f}(\mathbb{R}^n)$ is possibly nonconvex, $h \in \mathcal{C}^{1,1}_{L_h}(\mathbb{R}^n)$ is convex
- χ: $B_M(0) \mapsto \mathbb{R}$ is a bounded simple convex function for some $M > 0$
- $\mathcal{C}^{1,1}_L$ denotes the class of first-order $L-$smooth functions

Algorithm 1 Accelerated Gradient Method for Nonconvex Composite Problems

Require: starting point $x_0 \in \mathbb{R}^n$, $\{\alpha_k\}$ s.t. $\alpha_1 = 1$ and $\forall k \geq 2, 0 < \alpha_k < 1$, $\{\beta_k > 0\}$, and $\{\lambda_k > 0\}$

0. Set $x_0^{ag} = x_0$ and $k = 1$
1. Set
$$x_k^{md} = \alpha_k x_{k-1}^{ag} + (1-\alpha_k)\, x_{k-1}$$
2. Compute $\nabla\Psi\left(x_k^{md}\right)$ and set
$$\begin{aligned} x_k^{ag} &= \mathcal{P}\left(x_{k-1}, \nabla\Psi\left(x_k^{md}\right), \lambda_k\right)\\ x_k &= \mathcal{P}\left(x_k^{md}, \nabla\Psi\left(x_k^{md}\right), \beta_k\right)\end{aligned}$$
3. Set $k = k+1$ and go to step 1

Ensure: x_N^{md}

- $\mathcal{P}$ is the proximal operator defined as:

$$\mathcal{P}(x, y, c) := \arg\min_{u\in\mathbb{R}^n}\left\{\langle y, u\rangle + \frac{1}{2c}\|u-x\|^2 + \chi(u)\right\}.$$

- The algorithm attains global convergence while:

$$\alpha_k\lambda_k \leq \beta_k < \frac{1}{L_\Psi}, \ \forall k = 1, 2, \ldots N-1 \text{ and}$$
$$\frac{\alpha_1}{\lambda_1\Gamma_1} \geq \frac{\alpha_2}{\lambda_2\Gamma_2} \geq \cdots \geq \frac{\alpha_N}{\lambda_N\Gamma_N},$$

- Our interpretation showed that the algorithm is a *damped Nesterov's AG*, with the optimal parameter settings set to be [2]:

$$\alpha_1 = 1, \qquad \alpha_{k+1} = \frac{2}{1+\sqrt{1+\frac{4}{\alpha_k^2}}},$$
$$\lambda_1 = \beta, \qquad \lambda_{k+1} = \beta/\alpha_{k+1}, \qquad \beta = \frac{1-\delta}{L_\Psi}$$

- $0 < \delta \leq \frac{1}{2}$ depends on the nonconvex smooth component f in the objective function; specifically, L_f and L_Ψ

Simulation Study

- Data simulated with $\mathbf{x}_i \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\varepsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$
- $N = 1000$ for linear models and $N = 3000$ for logistic models
- $\boldsymbol{\tau}_{\text{generate}} \in \mathbb{R}^{10006}$ is a sparse constant vector with 6 "true" values of $1.23, 3, 4, 5, 6, 59$ as the true effect coefficients and 10000 values of 0 as coefficients to be eliminated by the penalty
- *AG improves convergence rate and is less likely to be stuck in local minimizers*

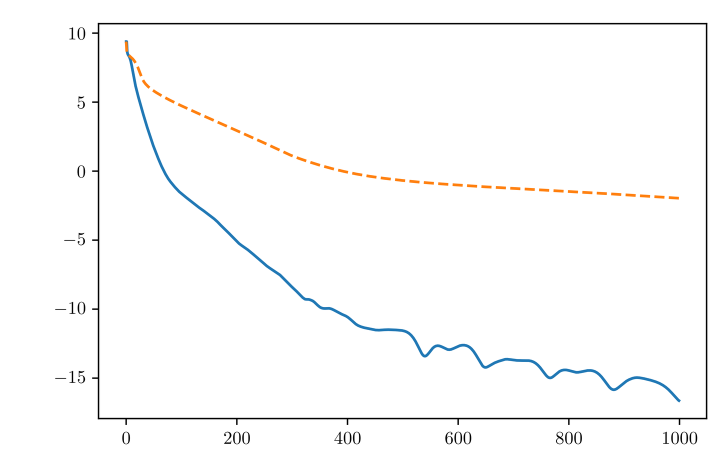

Figure: (A). LM penalized by SCAD

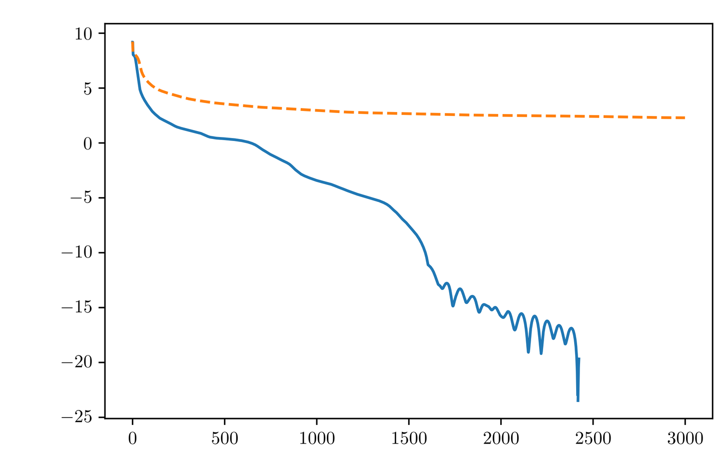

Figure: (B). LM penalized by MCP

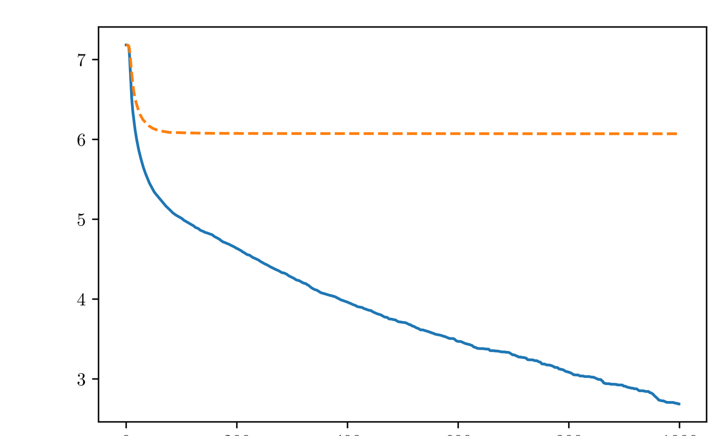

Figure: (C). logistic penalized by SCAD

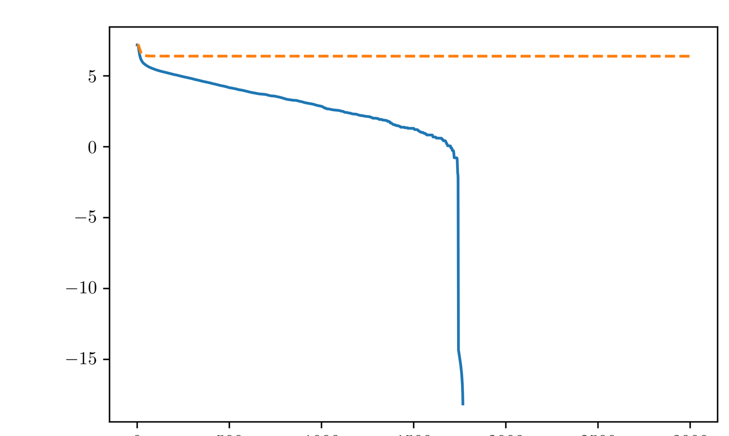

Figure: (D). logistic penalized by MCP

Figure: AG vs ISTA for Linear Models (LM) and Logistic Models penalized by SCAD and MCP. Horizontal axis represents the number of iterations k, vertical axis represents $\log(\text{Objective Value}_k - \text{Optimal Value})$. Orange dotted line for ISTA, blue line for AG (our proposed method).

References

[1] Saeed Ghadimi and Guanghui Lan. "Accelerated gradient methods for nonconvex nonlinear and stochastic programming". In: *Mathematical Programming* 156.1-2 (Feb. 2015), pp. 59–99. DOI: 10.1007/s10107-015-0871-8.

[2] Kai Yang, Masoud Asgharian, and Sahir Bhatnagar. "Improving Convergence for Nonconvex Composite Programming". In: (Sept. 22, 2020). arXiv: 2009.10629 [math.OC].