Pathwise Coordinate Optimization for Nonconvex Sparse Learning

Tuo Zhao

http://www.princeton.edu/~tuoz

Department of Computer Science
Johns Hopkins University

Mar. 25. 2015

Collaborators

This is joint work with

- Prof. Han Liu at Princeton University,
- Prof. Tong Zhang at Rutgers University and Baidu,
- Xingguo Li at University of Minnesota.

Manuscript:

http://arxiv.org/abs/1412.7477

Software Package:

http://cran.r-project.org/web/packages/picasso/

Outline

- Background
- Pathwise Coordinate Optimization
- Computational and Statistical Theories
- Numerical Simulations
- Conclusions

Background

Regularized M-Estimation

- Let $\boldsymbol{\beta}^*$ denote the parameter to be estimated. We solve the following regularized M-estimation problem

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^d} \underbrace{\mathcal{L}(\boldsymbol{\beta})+\mathcal{R}_\lambda(\boldsymbol{\beta})}_{\mathcal{F}_\lambda(\boldsymbol{\beta})},$$

  where $\mathcal{L}(\boldsymbol{\beta})$ is a smooth loss function, and $\mathcal{R}_\lambda(\boldsymbol{\beta})$ is a regularization function with a tuning parameter λ.

- Examples: Lasso, Logistic Lasso (Tibshirani, 1996), Group Lasso (Yuan and Lin, 2006), Graphical Lasso (Yuan and Lin, 2007; Banerjee et al., 2008; Friedman et al. 2008), ...

Regularization Functions

$\mathcal{R}_\lambda(\boldsymbol{\beta})$ is coordinate separable,

$$\mathcal{R}_\lambda(\boldsymbol{\beta}) = \sum_{j=1}^{d} r_\lambda(\beta_j).$$

$\mathcal{R}_\lambda(\boldsymbol{\beta})$ is decomposable,

$$\mathcal{R}_\lambda(\boldsymbol{\beta}) = \lambda \|\boldsymbol{\beta}\|_1 + \mathcal{H}_\lambda(\boldsymbol{\beta}) = \lambda \sum_{j=1}^{d} [|\beta_j| + h_\lambda(\beta_j)].$$

Examples: Smooth Clipped Absolute Deviation (SCAD, Fan and Li, 2001) and Minimax Concavity Penalty (MCP, Zhang, 2010)

Regularization Functions

For any $\gamma > 2$, SCAD is defined as

- $$r_\lambda(\beta_j) = \begin{cases} \lambda|\beta_j| & \text{if } |\beta_j| \leqslant \lambda, \\ -\dfrac{|\beta_j|^2 - 2\lambda\gamma|\beta_j| + \lambda^2}{2(\gamma-1)} & \text{if } \lambda < |\beta_j| \leqslant \lambda\gamma, \\ \dfrac{(\gamma+1)\lambda^2}{2} & \text{if } |\beta_j| > \lambda\gamma. \end{cases}$$
- $$h_\lambda(\beta_j) = \begin{cases} 0 & \text{if } |\beta_j| \leqslant \lambda, \\ \dfrac{2\lambda|\beta_j| - |\beta_j|^2 - \lambda^2}{2(\gamma-1)} & \text{if } \lambda < |\beta_j| \leqslant \lambda\gamma, \\ \dfrac{(\gamma+1)\lambda^2 - 2\lambda|\beta_j|}{2} & \text{if } |\beta_j| > \lambda\gamma. \end{cases}$$

Regularization Functions

For any $\gamma > 1$, MCP is defined as

- $r_\lambda(\beta_j) = \begin{cases} \lambda\left(|\beta_j| - \dfrac{|\beta_j|^2}{2\lambda\gamma}\right) & \text{if } |\beta_j| \leqslant \lambda\gamma, \\ \dfrac{\lambda^2\gamma}{2} & \text{if } |\beta_j| > \lambda\gamma. \end{cases}$
- $h_\lambda(\beta_j) = \begin{cases} -\dfrac{|\beta_j|^2}{2\gamma} & \text{if } |\beta_j| \leqslant \lambda\gamma, \\ \dfrac{\lambda^2\gamma - 2\lambda|\beta_j|}{2} & \text{if } |\beta_j| > \lambda\gamma. \end{cases}$

Regularization Functions

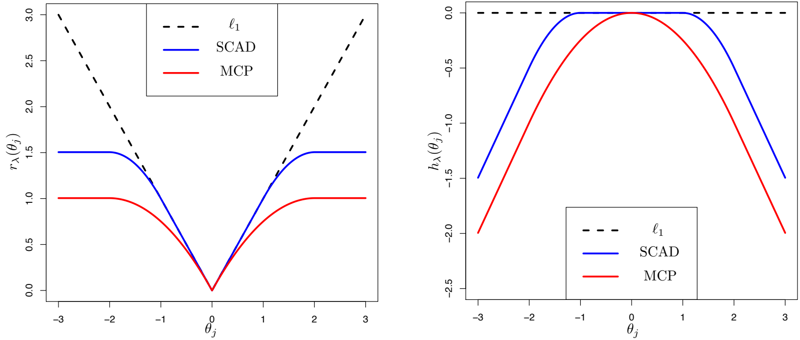

Figure: 1. $\lambda = 1$ and $\gamma = 2.01$.

Loss Functions

$\mathbf{X} \in \mathbb{R}^{n \times d}$ – design matrix, $\boldsymbol{y} \in \mathbb{R}^n$ – response vector.

- Least Square Loss:

$$\mathcal{L}(\boldsymbol{\beta}) = \frac{1}{2n} \|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2.$$

- Logistic Loss:

$$\mathcal{L}(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \left(\log\left[1 + \exp(\mathbf{X}_{i*}^T \boldsymbol{\beta})\right] - y_i \mathbf{X}_{i*}^T \boldsymbol{\beta} \right).$$

- Others: Huber Loss, Multi-category Logistic Loss,...

Reformulation

We rewrite the regularized M-estimation problem as

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^d} \underbrace{\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1}_{\mathcal{F}_\lambda(\boldsymbol{\beta})}.$$

- $\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta})$ is smooth but nonconvex,

$$\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) = \mathcal{L}(\boldsymbol{\beta}) + \mathcal{H}_\lambda(\boldsymbol{\beta}).$$

- $\lambda\|\boldsymbol{\beta}\|_1$ is nonsmooth but convex.

Remark: Amenable to theoretical analysis.

Randomized Coordinate Descent Algorithm

At the t-th iteration, we randomly select a coordinate j from d coordinates. We then take $\boldsymbol{\beta}_{\setminus j}^{(t+1)} \leftarrow \boldsymbol{\beta}_{\setminus j}^{(t)}$, and

- Exact Coordinate Minimization (Fu, 1998)

$$\beta_j^{(t+1)} \leftarrow \mathop{\arg\min}_{\beta_j} \ \widetilde{\mathcal{L}}_{\lambda}(\beta_j; \boldsymbol{\beta}_{\setminus j}^{(t)}) + \lambda|\beta_j|.$$

- Inexact Coordinate Minimization (Shalev-Shwartz, 2011)

$$\beta_j^{(t+1)} \leftarrow \mathop{\arg\min}_{\beta_j} \ (\beta_j - \beta^{(t)})\nabla_j \widetilde{\mathcal{L}}_{\lambda}(\boldsymbol{\beta}^{(t)}) + \frac{L}{2}(\beta_j - \beta^{(t)})^2 + \lambda|\beta_j|,$$

where L is the step size parameter.

Examples

- Sparse Linear Regression + MCP:

$$\mathcal{T}_{j,\lambda}(\boldsymbol{\beta}^{(t)}) = \begin{cases} \widetilde{\beta}_j^{(t+1)} & \text{if } |\widetilde{\beta}_j^{(t+1)}| \geqslant \gamma\lambda, \\ \dfrac{\mathcal{S}_\lambda(\widetilde{\beta}_j^{(t+1)})}{1 - 1/\gamma} & \text{if } |\widetilde{\beta}_j^{(t+1)}| < \gamma\lambda. \end{cases}$$

  where $\widetilde{\beta}_j^{(t+1)} = \mathbf{X}_{*j}^T(\mathbf{y} - \mathbf{X}_{*\setminus j}\boldsymbol{\beta}_{\setminus j}^{(t)})/n$.

- Sparse Logistic Regression + MCP:

$$\mathcal{T}_{j,\lambda}(\boldsymbol{\beta}^{(t)}) = \mathcal{S}_\lambda(\boldsymbol{\beta}^{(t)} - \nabla_j \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}^{(t)})/L)$$

Remark: Sublinear Convergence to Local Optima without Statistical Guarantees (Shalev-Shwartz, 2011).

Pathwise Coordinate Optimization

Pathwise Coordinate Optimization

- Much faster than other competing algorithms.
- Very simple implementation.
- Easily scale to large problems.
- NO computational analysis in existing literature
- NO statistical guranratee on the obtained estimator.

Our Contribution:

- The FIRST pathwise coordinate optimization algorithm with both computational and statistical guarantees.
- The FIRST two-step estimator with both computational and statistical guarantees.

Pathwise Coordinate Optimization

Friedman et al. 2007, Mazumder et al. 2011

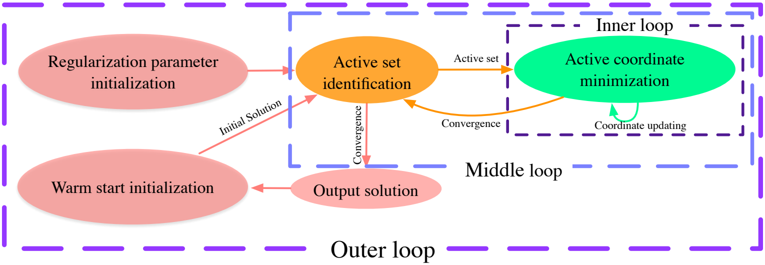

Figure: 2. The pathwise coordinate optimization framework contains 3 nested loops : (I) Warm start initialization; (II) Active set identification; (III) Active coordinate minimization.

Restricted Strong Convexity and Smoothness

Motivation: For any $\boldsymbol{\beta},\ \boldsymbol{\beta}' \in \mathbb{R}^d$ such that $|\{j \mid \beta_j \neq 0 \text{ or } \beta_j' \neq 0\}| \leqslant s$, we have

- $\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}') - \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) - (\boldsymbol{\beta}' - \beta)^T \nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) \geqslant \dfrac{C_-(s)}{2} \|\boldsymbol{\beta}' - \beta\|_2^2,$
- $\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}') - \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) - (\boldsymbol{\beta}' - \beta)^T \nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) \leqslant \dfrac{C_+(s)}{2} \|\boldsymbol{\beta}' - \beta\|_2^2,$

where $C_-(s),\ C_+(s) > 0$ are two constants depending on s.

Remark: An algorithm, which can maintain SPARSE solutions throughout all iterations, behaves like minimizing a STRONGLY CONVEX function. Therefore a linear convergence can be expected.

Warm Start Initialization (Outer Loop)

- We choose a sequence of DECREASING regularization parameters $\{\lambda_K\}_{K=1}^{N}$:

$$\lambda_0 \geqslant \lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_{N-1} \geqslant \lambda_N.$$

- The algorithm yields a sequence of output solutions $\{\widehat{\beta}^{\{K\}}\}_{K=0}^{N}$ from sparse to dense,

$$\widehat{\beta}^{\{K\}} \leftarrow \min_{\beta} \widetilde{\mathcal{L}}_{\lambda_K}(\beta) + \lambda_K \|\beta\|_1 .$$

Warm Start Initialization (Outer Loop)

- We choose $\lambda_0 = \|\nabla\mathcal{L}(\mathbf{0})\|_\infty$, then have

$$\min_{\boldsymbol{\xi}\in\partial\|\mathbf{0}\|_1} \|\nabla\mathcal{L}(\mathbf{0}) + \nabla\mathcal{H}_\lambda(\mathbf{0}) + \lambda_0\boldsymbol{\xi}\|_\infty = 0 \text{ and } \widehat{\boldsymbol{\beta}}^{\{0\}} = \mathbf{0}.$$

- The regularization sequence $\{\lambda_K\}_{K=0}^N$ is geometrically decreasing

$$\lambda_K = \eta\lambda_{K-1} \text{ with } \eta \in (0,1).$$

- When solving the optimization problem with λ_K, we use $\widehat{\boldsymbol{\beta}}^{\{K-1\}}$ as INITIALIZATION.

Geometric Interpretation

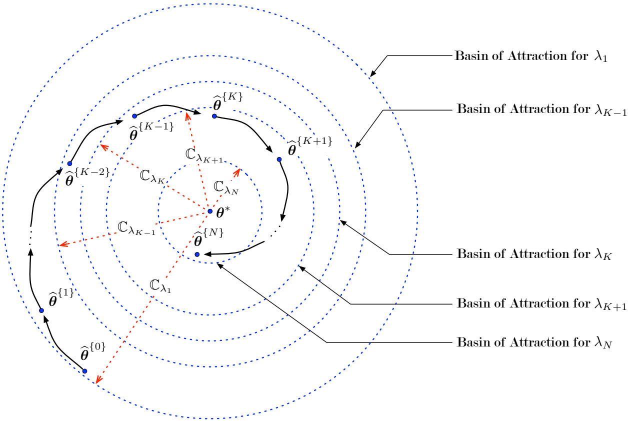

Figure: 3. Large regularization parameters suppress the overselection of irrelevant variables $\{j \mid \beta_j^* = 0\}$ and yields highly sparse solutions.

Active Set Strategy (Friedman et al. 2007)

Define $\mathcal{A} = \{j \mid \beta_j \neq 0\}$ be the set of indices of nonzero coordinates, and $\overline{\mathcal{A}} = \{j \mid \beta_j = 0\}$ be the set of indices of zero coordinates. A naive updating scheme is:

- (1) **Active coordinate minimization**: Cyclically update β_j's in $\mathcal{A}$ until convergence.
- (2) **Sweeping coordinates**: Check all β_j in $\mathcal{A}$, and if any coordinate becomes zero, move it to $\overline{\mathcal{A}}$.
- (3) **Adding coordinates**: Update β_j's over $\overline{\mathcal{A}}$ for only once, and if any coordinate becomes nonzero, move it to $\mathcal{A}$. Then we go back to (1).

Remark: Heuristic tricks without theoretical guarantees.

Active Set Identification (Middle Loop)

For notational simplicity, the outer loop index K is omitted.

Greedy Selection:

At the m-th iteration, we have $\boldsymbol{\beta}^{[m]}$ and define
$\mathcal{A}_m = \{j \mid \beta_j^{[m]} \neq 0\}$ and $\overline{\mathcal{A}}_m = \{j \mid \beta_j^{[m]} = 0\}$.

- $\boldsymbol{\beta}^{[m+0.5]} \leftarrow$ Active Coordinate Minimization over $\mathcal{A}_m$.
- $k_m \leftarrow \arg\max_{k \in \overline{\mathcal{A}}_m} |\nabla_k \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}^{[m+0.5]})|$.
- $\beta_{k_m}^{[m+1]} \leftarrow \mathcal{T}_{k_m,\lambda}(\boldsymbol{\beta}^{[m+0.5]})$ and $\boldsymbol{\beta}_{\setminus k_m}^{[m+1]} = \boldsymbol{\beta}_{\setminus k_m}^{[m+0.5]}$.

Remark: Conservative coordinate selection.

Active Set Identification (Middle Loop)

For notational simplicity, the outer loop index K is omitted.

Randomized Selection:

At the m-th iteration, we have $\boldsymbol{\beta}^{[m]}$ and define $\mathcal{A}_m = \{j \mid \beta_j^{[m]} \neq 0\}$ and $\overline{\mathcal{A}}_m = \{j \mid \beta_j^{[m]} = 0\}$.

- $\boldsymbol{\beta}^{[m+0.5]} \leftarrow$ Active Coordinate Minimization over $\mathcal{A}_m$.
- Randomly select $k_m \in \overline{\mathcal{A}}_m$ such that $|\nabla_k \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}^{[m+0.5]})| \geqslant \delta\lambda$.
- $\beta_{k_m}^{[m+1]} \leftarrow \mathcal{T}_{k_m,\lambda}(\boldsymbol{\beta}^{[m+0.5]})$ and $\boldsymbol{\beta}_{\setminus k_m}^{[m+1]} \leftarrow \boldsymbol{\beta}_{\setminus k_m}^{[m+0.5]}$.

Remark: Conservative coordinate selection.

Active Set Identification (Middle Loop)

For notational simplicity, the outer loop index K is omitted.

Truncated Cyclic Selection:

At the m-th iteration, we have $\boldsymbol{\beta}^{[m]}$ and define
$\mathcal{A}_m = \{j \mid \beta_j^{[m]} \neq 0\}$ and $\overline{\mathcal{A}}_m = \{j \mid \beta_j^{[m]} = 0\}$.

- $\boldsymbol{\beta}^{[m+0.5]} \leftarrow$ Active Coordinate Minimization over $\mathcal{A}_m$.
- For all $k \in \overline{\mathcal{A}}_m$, take
$$\beta_k^{[m+0.5]} \leftarrow \begin{cases} \mathcal{T}_{k,\lambda}(\beta^{[m+0.5]}) & \text{if } |\nabla_k \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}^{[m+0.5]})| \geqslant \delta\lambda, \\ \beta_k^{[m+0.5]} & \text{if } |\nabla_k \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}^{[m+0.5]})| < \delta\lambda. \end{cases}$$
- $\beta^{[m+1]} \leftarrow \boldsymbol{\beta}^{[m+0.5]}$.

Remark: Prevent from the overselection of irrelevant variables.

Active Set Identification (Middle Loop)

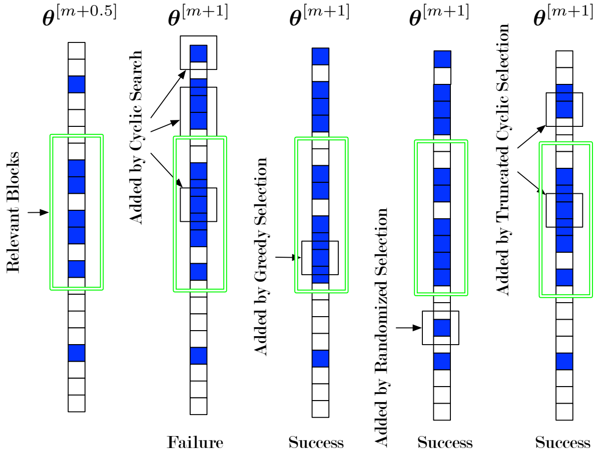

Figure: 4. The cyclic search in Friedman et al. 2007, Mazumder et al. 2011 may overselect irrelevant variables.

Active Coordinate Minimization (Inner Loop)

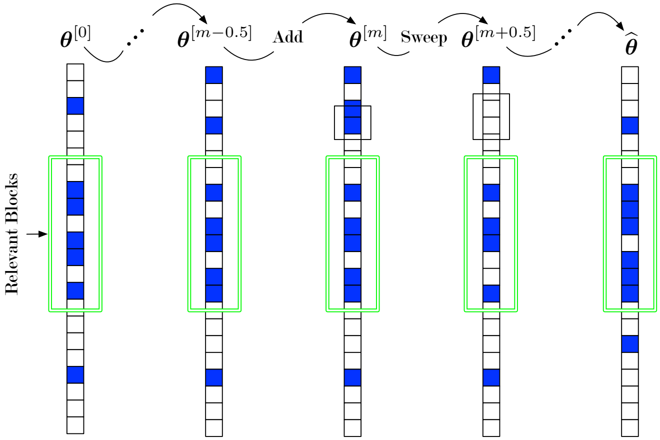

Figure: 5. The active coordinate minimization not only decreases the objective value, but also sweeps some variables from the active set.

Computational and Statistical Theories

Preliminaries

Definition [Sparse Eigenvalues]: Given an integer $s \geqslant 1$,

$$\rho_+(s) = \sup_{\|\mathbf{v}\|_0 \leqslant s} \frac{\mathbf{v}^T \nabla^2 \mathcal{L}(\boldsymbol{\beta}) \mathbf{v}}{\|\mathbf{v}\|_2^2}, \quad \rho_-(s) = \inf_{\|\mathbf{v}\|_0 \leqslant s} \frac{\mathbf{v}^T \nabla^2 \mathcal{L}(\boldsymbol{\beta}) \mathbf{v}}{\|\mathbf{v}\|_2^2},$$

and $\widetilde{\rho}_-(s) = \rho_-(s) - \alpha$.

Lemma [Restricted Curvature]: Given $\rho_+(s) > \rho_-(s) > \alpha$, for any $\boldsymbol{\beta},\ \boldsymbol{\beta}' \in \mathbb{R}^d$ such that $|\{j \mid \beta_j \neq 0 \text{ or } \beta_j' \neq 0\}| \leqslant s$,

$$\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}') - \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) - (\boldsymbol{\beta}' - \beta)^T \nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) \leqslant \frac{\rho_+(s)}{2} \|\boldsymbol{\beta}' - \beta\|_2^2,$$

$$\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}') - \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) - (\boldsymbol{\beta}' - \beta)^T \nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) \geqslant \frac{\widetilde{\rho}_-(s)}{2} \|\boldsymbol{\beta}' - \beta\|_2^2,$$

where $\mathcal{H}_\lambda(\boldsymbol{\beta}') - \mathcal{H}_\lambda(\boldsymbol{\beta}') - (\boldsymbol{\beta}' - \beta)^T \nabla \mathcal{H}_\lambda(\boldsymbol{\beta}) \geqslant -\frac{\alpha}{2} \|\boldsymbol{\beta}' - \beta\|_2^2.$

Preliminaries

Assumption A: $\lambda_N \geqslant 4\left\|\nabla\mathcal{L}(\beta^*)\right\|_\infty$ and $\eta \in [23/24, 1)$.

The regularization parameters are LARGE enough to eliminate irrelevant variables (Negahban et al. 2012).

Assumption B: Given $\|\beta^*\|_0 \leqslant s^*$, there exists an $\widetilde{s}$ such that

$$
\begin{aligned}
&(1)\ \ \widetilde{s} \geqslant (484\kappa^2 + 100\kappa)s^*,\\
&(2)\ \ \rho_+(s^* + 2\widetilde{s} + 2) < +\infty,\\
&(3)\ \ \widetilde{\rho}_-(s^* + 2\widetilde{s} + 2) > 0,
\end{aligned}
$$

where $\kappa = \rho_+(s^* + 2\widetilde{s} + 2)/\widetilde{\rho}_-(s^* + 2\widetilde{s} + 2)$.

The algorithm can tolerate AT MOST $\widetilde{s} + 1$ nonzero irrelevant variables throughout all iterations (Bickel, 2009; Zhang, 2009).

Globally Convergence (Greedy-PICASSO)

Suppose that Assumptions A and B hold. We have the following results:

- (**Solution Sparsity**) Through all iterations of PICASSO, any solution β satisfies $\|\beta_{\overline{\mathcal{S}}}\|_0 \leqslant \widetilde{s} + 1$.
- (**Sparse Optimum**) At the K-th iteration of the outer loop, PICASSO converges to a unique sparse local optimum $\bar{\beta}^{\lambda_K}$ satisfying $\left\|\bar{\beta}^{\lambda_K}_{\overline{\mathcal{S}}}\right\|_0 \leqslant \widetilde{s}$ and $\min_{\xi \in \partial \left\|\bar{\beta}^{\lambda_K}\right\|_1} \left\|\nabla \widetilde{\mathcal{L}}_{\lambda_K}(\bar{\beta}^{\lambda_K}) + \lambda \xi\right\|_\infty = 0$.
- (**Logarithm Iteration Complexity**) To attain $\mathcal{F}_{\lambda_N}(\widehat{\beta}^{\{N\}}) - \mathcal{F}_{\lambda_N}(\bar{\beta}^{\lambda_N}) \leqslant \epsilon$, the number of active set identification iterations is at most $\mathcal{O}\left(N \cdot \log(1/\epsilon)\right)$.

Two-step Method

- **Step 1 – Convex Relaxation**: Obtain β^{relax} satisfying

$$\min_{\xi \in \partial \|\beta^{\text{relax}}\|_1} \left\| \nabla \mathcal{L}(\beta^{\text{relax}}) + \lambda_0 \xi \right\|_\infty \leqslant \frac{\lambda_0}{8}.$$

- **Step 2 – PICASSO**: Solve the optimization problem with PICASSSO, and use β^{relax} as initialization for λ_0.

Remark: The low precision makes Step 1 very efficient.

Remark: The restricted strong convexity holds for $\|\beta - \beta^*\|_2 \leqslant R$ (e.g. logistic loss, huber loss), where R is a constant and does not scale with (n, d, s^*). All previous theoretical results hold.

Nearly Unbiased Estimation

Suppose that Assumptions A and B. We have

$$\left\|\widehat{\beta}^{\{N\}} - \beta^*\right\|_2 = \mathcal{O}\Bigg(\underbrace{\frac{\left\|\nabla_{\mathcal{S}_1}\mathcal{L}(\beta^*)\right\|_2}{\widetilde{\rho}_-(s^* + 2\widetilde{s})}}_{\text{Strong Signals}} + \underbrace{\frac{\lambda_N\sqrt{|\mathcal{S}_2|}}{\widetilde{\rho}_-(s^* + \widetilde{s})}}_{\text{Weak Signals}}\Bigg),$$

where $\mathcal{S}_1 = \{j \mid |\beta_j^*| \geqslant \gamma\lambda_N\}$ and $\mathcal{S}_2 = \{j \mid 0 < |\beta_j^*| < \gamma\lambda_N\}$.

Clarification: To establish the theoretical analysis for each individual problem, we need to assume that the model is CORRECTLY specified. This is a very common assumption in high dimensional statistical theories.

Model Specification

Sparse Linear Regression: We consider a linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} \sim N(0, \sigma^2 \mathbf{I}_n)$ is the observational noise vector.

Sparse Logistic Regression: We consider a logistic model

$$y_i \sim \text{Bernoulli}\left(\frac{\exp(\mathbf{X}_{i*}^T \boldsymbol{\beta}^*)}{1 + \exp(\mathbf{X}_{i*}^T \boldsymbol{\beta}^*)}\right)$$

for $i = 1, ..., n$.

Application to Spare Linear Regression

Verify Assumption A: Given $\lambda_N = 8\sigma\sqrt{\log d/n}$, with HIGH PROBABILITY, we have

$$\lambda_N \geqslant 4\|\nabla\mathcal{L}(\beta^*)\|_\infty .$$

Verify Assumption B: Suppose that each column of $\mathbf{X}$ is independently sampled from a sub-Gaussian distribution with mean $\mathbf{0}$ and covariance $\boldsymbol{\Sigma}$, where $\Lambda_{\min}(\boldsymbol{\Sigma}) \geqslant \psi_{\min}$ and $\Lambda_{\max}(\boldsymbol{\Sigma}) \geqslant \psi_{\max}$. Given $\alpha = \psi_{\min}/4$, there exists an $\widetilde{s}$ such that for large enough n, HIGH PROBABILITY, we have

$$\begin{aligned}
&(1)\quad \widetilde{s} \geqslant [484\kappa^2 + 100\kappa] \cdot s^*,\\
&(2)\quad \widetilde{\rho}_-(s^* + 2\widetilde{s} + 2) \geqslant \psi_{\min}/4,\\
&(3)\quad \rho_+(s^* + 2\widetilde{s} + 2) \leqslant 3\psi_{\max}/2.
\end{aligned}$$

Application to Spare Linear Regression

Parameter Estimation:

Given $\alpha = \psi_{\min}/4$ and $\lambda_N = 8\sigma\sqrt{\log d/n}$, we have

$$\left\|\widehat{\boldsymbol{\beta}}^{\{N\}} - \boldsymbol{\beta}^*\right\|_2 = \mathcal{O}_P\left(\underbrace{\sigma\sqrt{\frac{s_1^*}{n}}}_{\text{Strong Signals}} + \underbrace{\sigma\sqrt{\frac{s_2^*\log d}{n}}}_{\text{Weak Signals}}\right),$$

where $s_1^* = |\{j \mid |\beta_j^* \geqslant \gamma\lambda_N\}|$ and $s_2^* = |\{j \mid 0 < |\beta_j^*| < \gamma\lambda_N\}|$.

MCP v.s. ℓ_1:

$$\left\|\widehat{\boldsymbol{\beta}}^{\ell_1} - \boldsymbol{\beta}^*\right\|_2 = \mathcal{O}_P\left(\sigma\sqrt{\frac{s^*\log d}{n}}\right).$$

Application to Spare Linear Regression

Minimum Signal Strength: $\min_{j \in \mathcal{S}} |\beta_j^*| \geqslant \frac{C'\sigma}{\psi_{\min}} \sqrt{\frac{\log d}{n}}$.

Support Recovery:

Given $\alpha = \psi_{\min}/4$ and $\lambda_N = 8\sigma\sqrt{\log d/n}$, we have

$$\bar{\beta}^{\lambda_N} = \arg\min_{\beta} \frac{1}{2n} \|\boldsymbol{y} - \mathbf{X}\beta\|_2^2 \ \text{ subject to } \beta_{\overline{\mathcal{S}}} = \mathbf{0}$$

with high probability.

MCP v.s. ℓ_1: Restricted Strong Convexity v.s. Irrepresentablity.

Application to Spare Logistic Regression

Verify Assumption A: Given $\lambda_N = 8\sqrt{\log d/n}$, with high probability we have

$$\lambda_N \geqslant 4\left\|\nabla\mathcal{L}(\boldsymbol{\beta}^*)\right\|_\infty.$$

Verify Assumption B: Suppose that each column of $\mathbf{X}$ is independently sampled from a sub-Gaussian distribution with mean $\mathbf{0}$ and covariance $\boldsymbol{\Sigma}$, where $\Lambda_{\min}(\boldsymbol{\Sigma}) \geqslant \psi_{\min}$ and $\Lambda_{\max}(\boldsymbol{\Sigma}) \geqslant \psi_{\max}$. Given $\alpha = \psi_{\min}/4$, there exists an $\widetilde{s}$ such that for large enough n and any $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\| \leqslant R$, with high probability, we have

$$\begin{aligned}
&(1)\ \ \widetilde{s} \geqslant [484\kappa^2 + 100\kappa] \cdot s^*,\\
&(2)\ \ \widetilde{\rho}_-(s^* + 2\widetilde{s} + 2) \geqslant \psi_{\min}/4,\\
&(3)\ \ \rho_+(s^* + 2\widetilde{s} + 2) \leqslant 3\psi_{\max}/2.
\end{aligned}$$

Application to Spare Logistic Regression

Parameter Estimation:

Given $\alpha = \psi_{\min}/4$ and $\lambda_N = 8\sqrt{\log d/n}$, we have

$$\left\|\widehat{\beta}^{\{N\}} - \beta^*\right\|_2 = \mathcal{O}_P\left(\underbrace{\sqrt{\frac{s_1^*}{n}}}_{\text{Strong Signals}} + \underbrace{\sqrt{\frac{s_2^* \log d}{n}}}_{\text{Weak Signals}}\right),$$

where $s_1^* = |\{j \mid |\beta_j^* \geqslant \gamma\lambda_N\}|$ and $s_2^* = |\{j \mid 0 < |\beta_j^*| < \gamma\lambda_N\}|$.

MCP v.s. ℓ_1:

$$\left\|\widehat{\beta}^{\ell_1} - \beta^*\right\|_2 = \mathcal{O}_P\left(\sqrt{\frac{s^* \log d}{n}}\right).$$

Numerical Simulations

Numerical Simulations

- PICASSO with Greedy selection, denoted by "G-PICASSO".
- PICASSO with Randomized selection, denoted by "R-PICASSO".
- PICASSO with Truncated Cyclic selection, denoted by "TC-PICASSO".
- SPARSENET proposed in Mazumder et al. 2011.
- PISTA proposed in Wang et al. 2014.

Numerical Simulations

Table: 1. Quantitive comparison on sparse linear regression $(N = 100, n = 60, d = 1000, \sigma = 1, \lambda_N = 0.25\sqrt{\log d/n}), \gamma = 1.05$.

Method	$\|\widehat{\beta} - \beta^*\|_2$	$\|\widehat{\beta}_{\mathcal{S}}\|_0$	$\|\widehat{\beta}_{\mathcal{S}^c}\|_0$	Correct Selection	Timing
G-PICASSO	**0.8003**(0.8908)	**2.812**(0.4997)	**0.844**(2.066)	**666/1000**	0.0169(0.0027)
R-PICASSO	0.8102(0.9663)	2.791(0.5355)	0.902(2.353)	653/1000	0.0186(0.0034)
TC-PICASSO	0.8057(0.8374)	2.800(0.4839)	0.888(2.038)	645/1000	0.0167(0.0024)
SPARSENET	1.1260(1.2708)	2.669(0.6942)	1.678(3.191)	514/1000	0.0171(0.0025)
PISTA	0.8135(0.8998)	2.797(0.5115)	0.881(2.112)	664/1000	2.1771(0.3805)

Conclusions

Conclusions

- Multistage convex relaxation (Zhang, 2010; Zhang 2012): No theoretical guarantee on the iteration complexity; Needs to be combined with an efficient solver for each subproblem.
- One-step convex relaxation method (Zou and Li, 2008; Wang and Li, 2013; Fan et al. 2014): Attains suboptimal statistical rates of convergence; Requires a stronger minimum signal strength assumption; Needs to be combined with an efficient solver for each subproblem.
- Path-following proximal gradient algorithm (Wang et al. 2014): Worse empirical computational performance; Requires $\|\beta^*\|_2 \leqslant R/2$ for sparse generalized linear model estimation.

Conclusions

- Proximal gradient algorithm (Loh and Wainwright, 2013): Solves

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^d} \mathcal{L}(\boldsymbol{\beta}) + \mathcal{R}_\lambda(\boldsymbol{\beta}) \quad \text{subject to } \|\boldsymbol{\beta}\|_1 \leqslant R/2. \tag{1}$$

  Sophisticated parameter tuning; Inexact convergence; Slower parameter estimation rates of convergence; Requires $\|\boldsymbol{\beta}^*\|_1 \leqslant R/2$ for all nonconves sparse learning problems.

- Pathwise Calibrated Sparse Shooting Algorithm: Concrete theoretical guarantees; Empirically very efficient; Weaker Assumptions.

Thank You! Questions?