Sparse Regression with Non-Convex Regularization

Tong Zhang

Rutgers University

Background

Convex methods have become tremendously popular

- interesting **formulations**
- **computation**: can be solved efficiently
- formulations can be separated from computation
  - different computational procedures lead to the same solutions
- some **strong theoretical resutls** can be proved
  - working with the **KKD** condition at the solution

Background

Convex methods have become tremendously popular

- interesting **formulations**
- **computation**: can be solved efficiently
- formulations can be separated from computation
  - different computational procedures lead to the same solutions
- some **strong theoretical resutls** can be proved
  - working with the **KKD** condition at the solution

Nonconvex methods are much more difficult to analyze

- **formulation and computation needs to be considered together**
  - different computational procedures lead to different solutions
  - rigorously speaking, one cannot study one particular solution and its KKD condition
- may suffer from stability problems (multiple local solutions)

Background

Convex methods have become tremendously popular

- interesting **formulations**
- **computation**: can be solved efficiently
- formulations can be separated from computation
  - different computational procedures lead to the same solutions
- some **strong theoretical resutls** can be proved
  - working with the **KKD** condition at the solution

Nonconvex methods are much more difficult to analyze

- **formulation and computation needs to be considered together**
  - different computational procedures lead to different solutions
  - rigorously speaking, one cannot study one particular solution and its KKD condition
- may suffer from stability problems (multiple local solutions)

However, nonconvex formulations are natural for sparse learning.

This Talk: analyzing nonconvex methods

Need to study formulation and computation together

This Talk: analyzing nonconvex methods

Need to study formulation and computation together

- Natural formualtion requires **nonconvex penalty**
- Under certain assumptions (RIP), convex methods are not optimal
  - can be fixed by nonconvex procedures

This Talk: analyzing nonconvex methods

Need to study formulation and computation together

- Natural formualtion requires **nonconvex penalty**
- Under certain assumptions (RIP), convex methods are not optimal
  - can be fixed by nonconvex procedures
- What's known before: **exists a good local minimum solution** (better than Lasso)
  - but it is not clear one can find such a local solution efficiently

This Talk: analyzing nonconvex methods

Need to study formulation and computation together

- Natural formualtion requires **nonconvex penalty**
- Under certain assumptions (RIP), convex methods are not optimal
  - can be fixed by nonconvex procedures
- What's known before: **exists a good local minimum solution** (better than Lasso)
  - but it is not clear one can find such a local solution efficiently
- A specific computational procedure for nonconvex methods
  - we prove **the procedure lead to good local solution better than Lasso** (under reasonable conditions)
- A more general theory

Sparse Regression

$$Y = X\bar{\beta} + \epsilon$$

L_1 regularization: convex relaxation (computationally efficient)

$$\hat{\beta}_{L_1} = \arg\min_{\beta} \left[\|Y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right]$$

Sparse Regression

$$Y = X\bar{\beta} + \epsilon$$

L_1 regularization: convex relaxation (computationally efficient)

$$\hat{\beta}_{L_1} = \arg\min_{\beta} \left[\| Y - X\beta \|_2^2 + \lambda \|\beta\|_1 \right]$$

Theoretical question: recovery performance

- Variable selection (**can we find nonzero variables**):

$$\mathrm{supp}(\hat{\beta}) \approx \mathrm{supp}(\bar{\beta})?$$

- Parameter estimation (**how well we can estimate $\bar{\beta}$**):

$$\|\hat{\beta} - \bar{\beta}\|_2^2 \leq ?$$

RIP Condition

Definition (RIP — Sparse Eigenvalue Condition)

X satisfies the sparse eigenvalue condition at sparsity level s if

$$\inf\{n^{-1}\|X\beta\|_2^2 : \|\beta\|_2 = 1, \|\beta\|_0 \leq s\} > c_-,$$

$$\sup\{n^{-1}\|X\beta\|_2^2 : \|\beta\|_2 = 1, \|\beta\|_0 \leq s\} < c_+.$$

for constants $c_- > 0$ and $c_+ < \infty$

- requires the condition to hold at $s = O(\|\bar{\beta}\|_0)$
- Slightly more general than original RIP of Candes-Tao for compressed sensing.
- High dimensional **generalization of classical regularity condition** of design matrix being rank-p

Results under Restricted Isometry Property

- Variable selection guarantees:
  - Lasso is not variable selection consistent under noise
- Parameter estimation (oracle property):
  - Under variable selection consistency, we expect:

$$\|\bar{\beta} - \hat{\beta}\|^2 = O(\sigma^2 \|\bar{\beta}\|_0 / n)$$

  - Lasso: bias shows up as $\ln p$ factor

$$\|\bar{\beta} - \hat{\beta}\|^2 = O(\sigma^2 \|\bar{\beta}\|_0 \ln \mathbf{p} / n)$$

  high dimensional version of Lasso bias first discussed by Fan and Li.

Can we do better under RIP?

- Want to: **achieve optimal results under RIP**
  - variable selection consistency and parameter estimation without bias
- L_1 **not good enough approximation for** L_0 **regularization**

Can we do better under RIP?

- Want to: **achieve optimal results under RIP**
  - variable selection consistency and parameter estimation without bias
- **L_1 not good enough approximation for L_0 regularization**
- Improve convex relaxation:
  - **require nonconvex optimization**
  - difficult to analyze
  - computational efficiency statement for nonconvex optimization

Can we do better under RIP?

- Want to: **achieve optimal results under RIP**
  - variable selection consistency and parameter estimation without bias
- L_1 **not good enough approximation for** L_0 **regularization**
- Improve convex relaxation:
  - **require nonconvex optimization**
  - difficult to analyze
  - computational efficiency statement for nonconvex optimization

  This lecture:
  - **a special computational procedure**: multi-stage convex relaxation
  - **a general theory** of nonconvex regularization

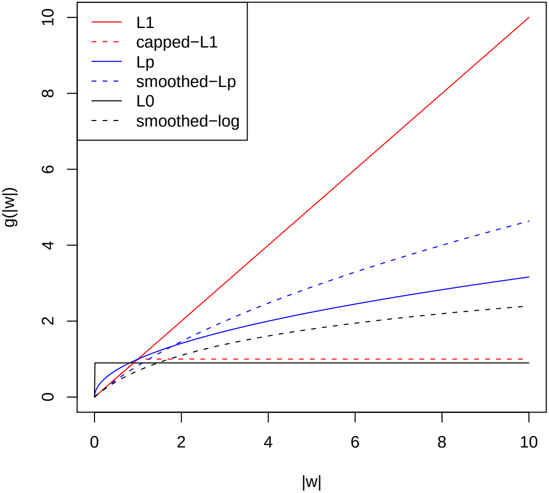
Non-convex formulation

- Approximate L_0 by **smooth concave sparse regularization** g
- Find local minimum by solving nonconvex problem

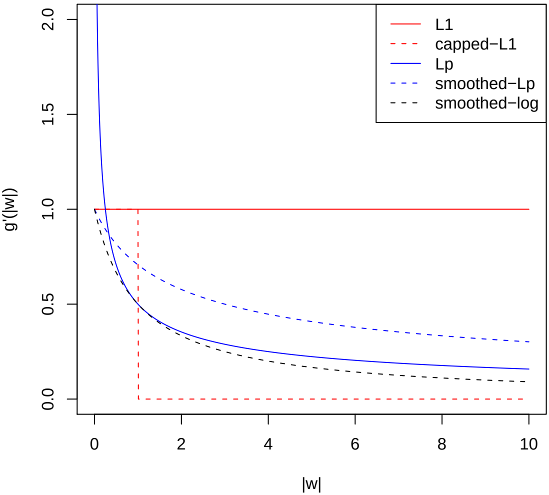
$$\hat{\beta}_g = \arg\min_{\beta} \left[\|Y - X\beta\|_2^2 + \lambda g(\beta) \right]$$

  want $g(\beta)$ to be closer to L_0 regularization than L_1 regularization
- Examples
  - L_p regularization: $g(\beta) = \sum_j |\beta_j|^p$ ($p < 1$)
  - smoothed L_p regularization: $g(\beta) = \sum_j [(\alpha + |\beta_j|)^p - \alpha]/(p\alpha^{p-1})$ ($p < 1$)
  - capped L_1 regularization: $g(\beta) = \sum_j \min(\alpha, |\beta_j|)$.

Sparse Regularizers (component-wise)



Derivative of Sparse Regularizers



Convex relaxation

- Want to optimize:

$$\hat{\beta} = \arg\min_{\beta \in R^p} [n^{-1}\|X\beta - Y\|_2^2 + \lambda g(\beta)], \tag{1}$$

- $g(\beta)$ concave with respect to element-wise vector function $\mathbf{h}(\beta)$ (e.g. $\mathbf{h}(\beta) = |\beta|$): exists g^* so that

$$g(\beta) = \inf_{\mathbf{v} \in R^p} \left[\mathbf{v}^T \mathbf{h}(\beta) + g^*(\mathbf{v}) \right].$$

- Rewrite (1) as

$$[\hat{\beta}, \hat{\mathbf{v}}] = \arg\min_{\beta, \mathbf{v} \in R^d} \left[n^{-1}\|X\beta - Y\|_2^2 + \lambda[\mathbf{v}^T \mathbf{h}(\beta) + g^*(\mathbf{v})] \right],$$

with auxiliary convex relaxation parameter **v**.

Multi-stage Convex Relaxation

- Numerical algorithm for solving

$$[\hat{\beta}, \hat{\mathbf{v}}] = \arg \min_{\beta, \mathbf{v} \in R^d} \left[n^{-1} \|X\beta - Y\|_2^2 + \lambda [\mathbf{v}^T \mathbf{h}(\beta) + g^*(\mathbf{v})] \right].$$

- Alternating Optimization: iterate from stage $\ell = 1, 2, \ldots$
  - fix $\mathbf{v}$ and optimize β:

$$\hat{\beta}^{(\ell)} = \arg \min_{\beta \in R^d} \left[n^{-1} \|X\beta - Y\|_2^2 + \lambda \hat{\mathbf{v}}_{old}^T \mathbf{h}(\beta) \right],$$

    solving weighted Lasso in β
  - fix β and optimize $\mathbf{v}$:

$$\hat{\mathbf{v}}_{new} = \arg \min_{\mathbf{v} \in R^d} [\mathbf{v}^T \mathbf{h}(\beta^{(\ell)}) + g^*(\mathbf{v})], \qquad (2)$$

    with closed form solution, leading to better and better convex relaxation.

Algorithm for $\mathbf{h}(\beta) = |\beta|$

Algorithm

- Initialization: $v_j^{(0)} = \lambda \ (j = 1, \ldots, p)$
- Iterate $\ell = 1, 2, \ldots$

$$\hat{\beta}^{(\ell)} = \arg\min_{\beta \in R^p} \left[\frac{1}{n} \|X\beta - Y\|_2^2 + \sum_{j=1}^{p} v_j^{(\ell-1)} |\beta_j| \right]$$

$$v_j^{(\ell)} = \lambda g'(|\hat{\beta}_j^{(\ell)}|) \qquad (j = 1, \ldots, p).$$

Remarks:

- Computationally **efficient**
  (solving convex/closed form solution problems each iteration)
- Converge to a **local minimum** of non-convex formulation
- Equivalent to local linear approximation of (Zou and Li)

Algorithm

- Initialization: $v_j^{(0)} = \lambda$ ($j = 1, \ldots, p$)
- Iterate $\ell = 1, 2, \ldots$

$$\hat{\beta}^{(\ell)} = \arg\min_{\beta \in R^p} \left[\frac{1}{n} \|X\beta - Y\|_2^2 + \sum_{j=1}^{p} v_j^{(\ell-1)} |\beta_j| \right]$$

$$v_j^{(\ell)} = \lambda g'(|\hat{\beta}_j^{(\ell)}|) \qquad (j = 1, \ldots, p).$$

Remarks:

- Computationally **efficient**
  (solving convex/closed form solution problems each iteration)
- Converge to a **local minimum** of non-convex formulation
- Equivalent to local linear approximation of (Zou and Li)

Key question: is the local minimum good in high dimension?

High Dimensional Theory under RIP

Theorem (T.Z. 10 & 12)

Under RIP, multi-stage convex relaxation with appropriate nonconvex regularizer $g(\beta)$ gives a solution

$$\mathrm{supp}(\hat{\beta}) = \mathrm{supp}(\bar{\beta}), \qquad \|\hat{\beta} - \bar{\beta}\|_2^2 \leq O(\sigma^2 \|\bar{\beta}\|_0 / n)$$

after $\log(\|\bar{\beta}\|_0)$ *stages; if for somce constant c:*

$$\min_{j \in \mathrm{supp}(\bar{\beta})} |\bar{\beta}_j| \geq c\sigma \sqrt{\ln p/n}.$$

- **local minimum found by the algorithm is good under RIP**
- Two-stage version is adaptive Lasso (Zou), which suffers from bias (C.H. Zhang), and sub-optimal for variable selection under RIP:

$$\min_{j \in \mathrm{supp}(\bar{\beta})} |\bar{\beta}_j| \geq c\sigma \sqrt{\|\bar{\beta}\|_0 \ln p/n}.$$

An Illustrative Example

- 500 variables and 100 data points
- True coefficients (5 nonzeros)

	coefficient	2-norm error
truth	$[8.2, 1.7, 5.4, 6.9, 5.7, 0.0, 0.0, 0.0, 0.0, \cdots]$	0
Stage 1	$[6.0, 0.0, 4.7, 4.8, 3.9,$ **0.6**, **0.7**, **1.2**, $0.0, \ldots]$	4.4
Stage 2	$[7.7, 0.4, 5.7, 6.3, 5.7, 0.0, 0.0,$ **0.2**, $0.0, \ldots]$	1.6
Stage 3	$[7.8, 1.2, 5.7, 6.6, 5.7, 0.0, 0.0, 0.0, 0.0, \ldots]$	0.98

- The result is with capped-L_1 regularization: stablizes after stage 3.
- **Errors** are highlighted.

Summary of Multi-stage Convex Relaxation

- A **specialized procedure** for solving concave regularizaton.
  - lead to a **local minimum**
  - **optimal performance** of the local minimum under RIP
    - optimal for parameter estimation and variable selection

Summary of Multi-stage Convex Relaxation

- A **specialized procedure** for solving concave regularizaton.
  - lead to a **local minimum**
  - **optimal performance** of the local minimum under RIP
    - optimal for parameter estimation and variable selection
- Similar results hold for other procedures, and in particular forward-backward procedure by T.Z. and MC+ by C.H. Zhang.

Summary of Multi-stage Convex Relaxation

- A **specialized procedure** for solving concave regularizaton.
  - lead to a **local minimum**
  - **optimal performance** of the local minimum under RIP
    - optimal for parameter estimation and variable selection
- Similar results hold for other procedures, and in particular forward-backward procedure by T.Z. and MC+ by C.H. Zhang.
- Can we prove something more general?
  - What's the **relationship among local minima** from different procedures?
  - What's the property of **global optimal** solution?
  - Can we **find global optimal solution efficiently** of nonconvex sparse regularization under RIP type condition?

General Theory: outline

- Oracle Least Squares Solution
  - least squares when support is known
  - the target solution we hope to achieve using nonconvex regularization
- Theory of L_0 regularization
  - property of global solution
  - local solution and algorithm (forward-backward greedy procedure)
- Smooth nonconvex penalty
  - sparse local solution
  - global solution
  - approximate global solution and a numerical procedure

Oracle Least Squares

Least squares solution under the oracle of **knowing the true support** $\bar{F} = \mathrm{supp}(\bar{\beta})$

$$\hat{\beta}_{oracle} = \arg\min_{\beta} \|Y - X\beta\|_2^2, \qquad \text{subject to } \mathrm{supp}(\beta) \subset \bar{F}.$$

- Not a practical solution, but introduced **for theoretical analysis**
  - it is variable selection sign consistent when $|\beta_j| \geq c\sigma\sqrt{\ln p/n}$ for $j \in \bar{F}$.
  - it has oracle proprty for parameter estimation
- Goal of nonconvex penalty: **close to** $\hat{\beta}_{oracle}$ as much as possible.

Theory of L_0 Regularizaton: global solution

If we can compute the global solution of L_0 regularization problem

$$\hat{\beta}_{L_0} = \arg\min_{\beta} \left[n^{-1} \| Y - X\beta \|_2^2 + \lambda \|\beta\|_0 \right]$$

what's the theoretical guarantees?

Theory of L_0 Regularizaton: global solution

If we can compute the global solution of L_0 regularization problem

$$\hat{\beta}_{L_0} = \arg\min_{\beta} \left[n^{-1} \|Y - X\beta\|_2^2 + \lambda \|\beta\|_0 \right]$$

what's the theoretical guarantees?

- is the global solution sparse?

Theory of L_0 Regularizaton: global solution

If we can compute the global solution of L_0 regularization problem

$$\hat{\beta}_{L_0} = \arg\min_{\beta} \left[n^{-1} \| Y - X\beta \|_2^2 + \lambda \|\beta\|_0 \right]$$

what's the theoretical guarantees?

- is the global solution sparse?
- how good is global solution?

Theory of L_0 Regularizaton: global solution

If we can compute the global solution of L_0 regularization problem

$$\hat{\beta}_{L_0} = \arg\min_{\beta} \left[n^{-1} \|Y - X\beta\|_2^2 + \lambda \|\beta\|_0 \right]$$

what's the theoretical guarantees?

- is the global solution sparse?
- how good is global solution?
- does it recover support

Theory of L_0 Regularizaton: global solution

If we can compute the global solution of L_0 regularization problem

$$\hat{\beta}_{L_0} = \arg\min_{\beta} \left[n^{-1} \|Y - X\beta\|_2^2 + \lambda \|\beta\|_0 \right]$$

what's the theoretical guarantees?

- is the global solution sparse?
- how good is global solution?
- does it recover support
- what is the relationship with oracle least squares solution?

Theoretical Results

Theorem (C.H. Zhang and T.Z. 12)

Assume $\lambda \geq c\sigma\sqrt{\ln p/n}$ *for some constant* $c > 0$. *The* **global solution of L_0 regularization is sparse***:*

$$\|\hat{\beta}_{L_0}\|_0 \leq \frac{1+\eta^2}{1-\eta^2}\|\bar{\beta}\|_0, \qquad \|X\hat{\beta}_{L_0} - X\bar{\beta}\|_2^2 \leq \frac{(1+\eta)\lambda^2\|\bar{\beta}\|_0}{1-\eta}.$$

Let $s = 2\|\bar{\beta}\|_0/(1-\eta^2)$ *and* $\hat{\beta}_{oracle}$ *be oracle least squares solution. Let* $\delta^o = \#\{j \in \bar{F} : |\bar{\beta}_j| = O(\lambda)\}$*, then*

$$|\bar{F} - \mathrm{supp}(\hat{\beta})| + |\mathrm{supp}(\hat{\beta}) - \bar{F}| = O(\delta^o), \qquad \|X(\hat{\beta}_{L_0} - \hat{\beta}_{oracle})\|_2^2 \leq 2\lambda^2\delta^o.$$

- if $|\bar{\beta}_j| \geq c\sigma\sqrt{\ln p/n}$ for some c, then $\delta^o = 0$, which means

$$\mathrm{supp}(\hat{\beta}) = \bar{F} \qquad \hat{\beta}_{L_0} = \hat{\beta}_{oracle}.$$

- how to find local/global solution for L_0 regularization?

Theory of Smooth Nonconvex Penalty

- L_0 penalty is discountinuous.
  - Can be tricky to optimize using traditional numerical methods
  - Although some special procedures (FoBa) can be employed
- What if we have a smooth regularizer
  - piece-wise differentiable

Theory of Smooth Nonconvex Penalty

- L_0 penalty is discountinuous.
  - Can be tricky to optimize using traditional numerical methods
  - Although some special procedures (FoBa) can be employed
- What if we have a smooth regularizer
  - piece-wise differentiable
- Key property of smooth regularizer:
  - **well-defined local optimal solution**
  - stuiable for traditional numerical methods

Sparsity of Global Solution

Consider concave regularization

$$\hat{\beta} = \arg\min_{\beta} \left[n^{-1}\|Y - X\beta\|_2^2 + \lambda \sum_{j=1}^{p} g(\beta_j) \right]$$

Theorem (C.H. Zhang and T.Z. 12)

Under appropriate conditions, with $\lambda \geq c\sigma\sqrt{\ln p/n}$ for some constant $c > 0$. The **global solution is sparse***:*

$$|\mathrm{supp}(\hat{\beta})| = O(|\bar{F}|)$$

The sparsity of global solution allows us to show its relationship to a sparse local soluiton.

Sparse Local Solution

Consider concave regularization

$$\hat{\beta} = \arg\min_{\beta} \left[n^{-1} \| Y - X\beta \|_2^2 + \lambda \sum_{j=1}^{p} g(\beta_j) \right]$$

$\tilde{\beta} \in R^p$ is a local solution if for all $j \in R^p$

$$X_j^\top (X\tilde{\beta} - Y)/n + \lambda g'(\tilde{\beta}_j) = 0.$$

Theorem (C.H. Zhang and T.Z. 12)

Suppose $g'(t) = 0$ for some $t > \min_{j \in \bar{F}} |\bar{\beta}_j| \geq c\lambda$ for some constant c. Then, there exists a **unique sparse local solution** *$\tilde{\beta}$ at sparsity level $\|\tilde{\beta}\|_0 = O(|\bar{F}|)$ such that $\mathrm{sgn}(\tilde{\beta}) = \mathrm{sgn}(\bar{\beta})$ and $\tilde{\beta} = \hat{\beta}_{oracle}$. Moreover, $\tilde{\beta}$ is the global solution.*

Approximate Global Optimal

Similar results hold for **approximate local solution**

$$\|X^\top(X\tilde{\beta} - Y)/n + \lambda \nabla g(\tilde{\beta})\|_2 \le \nu.$$

Also define **approximate global solution**

$$\left[\frac{1}{2n}\|X\tilde{\beta} - \mathbf{y}\|_2^2 + \lambda g(\tilde{\beta})\right] \le \left[\frac{1}{2n}\|X\bar{\beta} - \mathbf{y}\|_2^2 + \lambda g(\bar{\beta})\right] + \nu.$$

Theorem (C.H. Zhang and T.Z 12)

The **Lasso solution** *(L_1 regularization)* **is an approximate global solution** *with $\nu = O(\lambda^2|\bar{F}|)$, and any approximate global solution with $\nu = O(\lambda^2|\bar{F}|)$ which is also an approximate local minimum* **is sparse***:*

$$|\mathrm{supp}(\hat{\beta}) \setminus \bar{F}| = O(|\bar{F}|).$$

Smooth Regularizer: Putting Things Together

If $g'(t) = 0$ for $t \geq c\sigma\sqrt{\ln p/n}$, then under appropriate conditions:

- global solution is sparse
- **approximate global solution is sparse** if it is also a local minimum
- approximate global solution **can be achieved by Lasso**
- **sparse local solution is unique**
- sparse local solution has **appropriate oracle property**
  - optimal (up to a constant depending on RIP condition) both for estimation and variable selection

Smooth Regularizer: Putting Things Together

If $g'(t) = 0$ for $t \geq c\sigma\sqrt{\ln p/n}$, then under appropriate conditions:

- global solution is sparse
- **approximate global solution is sparse** if it is also a local minimum
- approximate global solution **can be achieved by Lasso**
- **sparse local solution is unique**
- sparse local solution has **appropriate oracle property**
  - optimal (up to a constant depending on RIP condition) both for estimation and variable selection
- Computational idea:
  - start with Lasso
  - do gradient descent to decrease objective function.
  - eventually converges to sparse local minimum which is global optimal

Simple Computational Procedure

- Start with Lasso solution $\hat{\beta}_{L_1}$
- Using gradient descent to decrease objective value with appropriate non-convex penalty until convergence.

Corollary

Under appropriate conditions, the solution from above procedure **converges to the unique global solution that is sparse***, and thus has appropriate* **oracle properties***.*

References

- J Fan and R Li, Variable selection via nonconcave penalized likelihood and its oracle properties, JASA, 2001.
- C-H Zhang, Nearly unbiased variable selection under minimax concave penalty, The Annals of Statistics, 2010.
- T. Zhang, Analysis of multi-stage convex relaxation for sparse regularization, JMLR, 2010.
- C-H Zhang and T Zhang, A general theory of concave regularization for high-dimensional sparse estimation problems, Statistical Science, 2012.