

## Topic 14: Regularization Methods in Thresholded Parameter Space

by Sai Zhang

**Key points:** The connections and differences of all regularization methods and some interesting phase transition phenomena.

**Disclaimer:** The note is built on Prof. [Jinchi Lv](#)'s lectures of the course at USC, DSO 607, High-Dimensional Statistics and Big Data Problems.

### 14.1 Model Setup

Now, consider a generalized linear model (GLM) linking a  $p$ -dimensional predictor  $\mathbf{x}$  to a scalar response  $Y$ . With canonical link, the conditional distribution of  $Y$  given  $\mathbf{x}$  has density

$$f(y; \theta, \phi) = \exp [y\theta - b(\theta) + c(y, \phi)]$$

where  $\theta = \mathbf{x}'\boldsymbol{\beta}$  with  $\boldsymbol{\beta}$  a  $p$ -dimensional regression coefficient vector,  $b(\cdot)$  and  $c(\cdot, \cdot)$  are known functions and  $\phi$  is dispersion parameter. Again,  $\boldsymbol{\beta} = (\beta_{0,1}, \dots, \beta_{0,p})'$  is sparse with many zero components, and  $\log p = O(n^a)$  for some  $0 < a < 1$ .

The penalized negative log-likelihood is

$$Q_n(\boldsymbol{\beta}) = -n^{-1} [\mathbf{y}'\mathbf{X}\boldsymbol{\beta} - \mathbf{1}'\mathbf{b}(\mathbf{X}\boldsymbol{\beta})] + \|p_\lambda(\boldsymbol{\beta})\|_1$$

where

- $\mathbf{y} = (y_1, \dots, y_n)'$ ,  $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)'$ , each column of  $\mathbf{X}$  is rescaled to have  $L_2$ -norm  $\sqrt{n}$
- $\mathbf{b}(\boldsymbol{\theta}) = (b(\theta_1), \dots, b(\theta_n))'$  with  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_n)'$
- $\|p_\lambda(\boldsymbol{\beta})\|_1 = \sum_{j=1}^p p_\lambda(|\beta_j|)$

Next, define **robust spark**  $\kappa_c$

#### Definition 14.1.1: Robust spark $\kappa_c$

The robust spark  $\kappa_c$  of the  $n \times p$  design matrix  $\mathbf{X}$  is defined as the smallest possible positive integer s.t. there exists an  $n \times \kappa_c$  submatrix of  $\frac{1}{\sqrt{n}}\mathbf{X}$  having a singular value less than a given positive constant  $c$  ([Zheng et al., 2014](#)), and

$$\kappa_c \leq n + 1$$

Bounding sparse model size can control collinearity and ensure model identifiability and stability, and as  $c \rightarrow 0+$ ,  $\kappa_c$  approaches the spark. Robust spark can be some large number diverging with  $n$ :

#### Proposition 14.1.2: Order of $\kappa_c$

Assume  $\log p = o(n)$  and that the rows of the  $n \times p$  random design matrix  $\mathbf{X}$  are i.i.d. as  $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ , where  $\boldsymbol{\Sigma}$  has smallest eigenvalue bounded from below by some positive constant. Then there exist

positive constants  $c$  and  $\tilde{c}$  s.t. with asymptotic probability one,  $\kappa_c \geq \frac{\tilde{c}n}{\log p}$

Next, we define a thresholded parameter space

**Definition 14.1.3: Thresholded parameter space**

$$\mathcal{B}_{\tau,c} = \left\{ \beta \in \mathbb{R}^p : \|\beta\|_0 < \frac{\kappa_c}{2}, \text{ and for each } j, \beta_j = 0 \text{ or } |\beta_j| \geq \tau \right\}$$

where  $\beta = (\beta_1, \dots, \beta_p)'$ .  $\tau$  is some positive threshold on parameter magnitude:

Here,  $\tau$  is very important:

- $\tau$  is key to distinguishing between important covariates and noise covariates for the purpose of variable selection
- $\tau$  typically needs to satisfy  $\tau \sqrt{n/\log p} \xrightarrow{n \rightarrow \infty} \infty$

It turns out that the solution to the regularization problem has the (very natural) hard-thresholding property:

**Proposition 14.1.4: Hard-thresholding property**

or the  $L_0$ -penalty  $p_\lambda(t) = \lambda \mathbf{1}_{t \neq 0}$ , the global minimizer  $\hat{\beta} = (\hat{\beta}_1, \dots, \hat{\beta}_p)'$  of the regularization problem over  $\mathbb{R}^p$  satisfies that each component  $\hat{\beta}_j$  is either 0 or has magnitude larger than some positive threshold

This hard-thresholding property is shared by many other penalties such as SICA penalties. This property guarantees sparsity of the model: weak signals are generally difficult to stand out comparing to noise variables due to impact of high dimensionality

## 14.2 Asymptotic Equivalence of Regularization Methods

For a universal  $\lambda = c_0 \sqrt{\log p/n}$  with  $c_0 > 0$  and  $p$  implicitly as  $n \vee p$ , consider 2 key events:

$$\mathcal{E} = \left\{ \|n^{-1} \mathbf{X}' \epsilon\|_\infty \leq \lambda/2 \right\} \quad \mathcal{E}_0 = \left\{ \|n^{-1} \mathbf{X}'_{\alpha_0} \epsilon\|_\infty \leq c_0 \sqrt{\log n/n} \right\}$$

where  $\epsilon = \mathbf{y} - \mathbb{E}\mathbf{y}$ ,  $\mathbf{X}_\alpha$  is a submatrix of  $\mathbf{X}$  consisting of columns in  $\alpha$ . Here, let  $\alpha_0 = \text{supp}(\beta_0)$  (non-zero variables in the true model).

For this setting, consider the following technical conditions:

- C1 **Error tail distribution**:  $\Pr(\mathcal{E}^c) = O(p^{-c_1})$  and  $\Pr(\mathcal{E}_0^c) = O(n^{-c_1})$  for some positive constant  $c_1$  that can be sufficiently large for large enough  $c_0$
- C2 **Bounded variance**:  $b(\theta)$  satisfies that  $c_2 \leq b''(\theta) \leq c_2^{-1}$  in its domain, where  $c_2$  is some positive constant
- C3 **Concave penalty function**:  $p_\lambda(t)$  is increasing and concave in  $t \in [0, \infty)$  with  $p_\lambda(0) = 0$ , and is differentiable with  $p'_\lambda(0+) = c_3 \lambda$  for some positive constant  $c_3$ <sup>1</sup>
- C4 **Ultra-high dimensionality**:  $\log p = O(n^a)$  for some constant  $a \in (0, 1)$

<sup>1</sup>A wide class of penalties, including  $L_1$ -penalty in Lasso, SCAD, MCP and SICA, satisfy this condition.

C5 **True parameter vector**:  $s = o(n^{1-a})$  and  $\exists c > 0$  s.t. the **robust spark**  $\kappa_c > 2s$ . Moreover,  $\min_{1 \leq j \leq s} |\beta_{0,j}| \gg \sqrt{\log p/n}$

Given these 5 conditions, we have that the global minimizer  $\hat{\beta} = \arg \min_{\beta \in \mathcal{B}_\tau} Q_n(\beta)$  exists and satisfies oracle inequalities:

#### Theorem 14.2.1: Oracle Inequalities

Assume that Condition 1-5 hold and  $\tau$  is chosen s.t.  $\tau < \min_{1 \leq j \leq s} |\beta_{0,j}|$  and  $\lambda = c_0 \sqrt{\log p/n} = o(\tau)$ , then the global minimizer exists, and any such global minimizer satisfies that with probability at least  $1 - O(p^{-c_1})$ , it holds simultaneously that

- **False sign**:

$$FS(\hat{\beta}) \leq \frac{Cs\lambda^2\tau^{-2}}{1 - C\lambda^2\tau^{-2}}$$

- **Estimation losses**:

$$\begin{aligned} \|\hat{\beta} - \beta_0\|_q &\leq C\lambda s^{1/q}(1 - C\lambda^2\tau^{-2})^{-1/q} \\ \|\hat{\beta} - \beta_0\|_\infty &\leq C\lambda s^{1/2}(1 - C\lambda^2\tau^{-2})^{-1/2} \end{aligned} \quad \forall q \in [1, 2]$$

- **Prediction loss**:

$$\frac{1}{\sqrt{n}} \|\mathbf{X}(\hat{\beta} - \beta_0)\|_2 \leq C\lambda s^{1/2}(1 - C\lambda^2\tau^{-2})^{-1/2}$$

where  $C$  is some positive constant.

#### How to understand Thm.14.2.1

- These results hold uniformly over the set of all possible global minimizers
- $c_1$  in probability bound can be chosen arbitrarily large, affecting **only**  $C$
- $FS(\hat{\beta}) = o(s)$  since  $\lambda = o(\tau)$ , while  $\|\hat{\beta}\|_0 = O(\phi_{\max}s)$  where  $\phi_{\max}$  is the largest eigenvalue of  $\frac{1}{n}\mathbf{X}'\mathbf{X}$
- $\forall q \in [1, 2]$ , the convergence rates of estimation losses

$$\begin{aligned} \|\hat{\beta} - \beta_0\|_q &= O\left\{s^{1/q}\sqrt{\frac{\log p}{n}}\right\} \\ \frac{1}{\sqrt{n}}\|\mathbf{X}(\hat{\beta} - \beta_0)\|_2 &= O\left(\sqrt{\frac{s \log p}{n}}\right) \end{aligned}$$

are consistent with Lasso.

We also have a sign consistency result:

#### Theorem 14.2.2: Sign Consistency and Oracle Inequalities

Assume the same conditions of Thm.14.2.1, further assume  $\min_{1 \leq j \leq s} |\beta_{0,j}| \geq 2\tau$  and  $\lambda = c_0 \sqrt{\log p/n} = o(s^{-1/2}\tau)$ , and  $\gamma_n = o\left(\tau \sqrt{\frac{n}{s \log n}}\right)$ , then any global minimizer  $\hat{\beta}$  defined satisfies that with probability at least  $1 - O(n^{-c_1})$ , it holds simultaneously that

- **Sign consistency**:  $\text{sgn}(\hat{\beta}) = \text{sgn}(\beta_0)$
- **Estimation and prediction losses**: If the penalty function further satisfies  $p'_\lambda(\tau) = O\left(\frac{\log n}{n}\right)$ , then  $\forall q \in [1, 2]$ ,

$$\|\hat{\beta} - \beta_0\|_q \leq C s^{1/q} \sqrt{\frac{\log n}{n}} \quad \|\hat{\beta} - \beta_0\|_\infty \leq C \gamma_n^* \sqrt{\frac{\log n}{n}} \quad n^{-1} D(\hat{\beta}) \leq C \frac{s \log n}{n}$$

where  $\gamma_n^*$  is a constant showing the behavior of  $\left\| \left[ \frac{1}{n} \mathbf{X}'_{\alpha_0} \mathbf{H}(\beta_1, \dots, \beta_n) \mathbf{X}_{\alpha_0} \right]^{-1} \right\|_\infty$  in a small neighborhood of  $\beta_0$ ,  $D(\hat{\beta})$  is the Kullback-Leibler divergence, and  $C$  is some positive constant

**How to understand Thm.14.2.2** Consider a linear model, where

$$\gamma_n^* = \left\| \left( \frac{1}{n} \mathbf{X}'_{\alpha_0} \mathbf{X}_{\alpha_0} \right)^{-1} \right\|_\infty \leq \sqrt{s} \left\| \left( \frac{1}{n} \mathbf{X}'_{\alpha_0} \mathbf{X}_{\alpha_0} \right)^{-1} \right\|_2 \leq \frac{\sqrt{s}}{c} \quad \gamma_n = \sup_{\alpha \in \{s+1, \dots, p\}, |\alpha| \leq s} \left\| \frac{1}{n} \mathbf{X}'_{\alpha_0} \mathbf{X}_\alpha \right\|_\infty$$

when all true covariates are orthogonal to each other,  $\gamma_n^* = 1$  and

$$\|\hat{\beta} - \beta_0\|_\infty \leq C \sqrt{\frac{\log n}{n}}$$

within a logarithmic factor  $\log n$  or oracle rate. Meanwhile, the penalty function condition  $p'_\lambda(\tau) = O\left(\frac{\log n}{n}\right)$  can be easily satisfied by concave penalties such as SCAD and SICA, having convergence rates improved with  $\log n$  in place of  $\log p$ .

**Phase transition phenomenon** Combining Thm.14.2.1 and 14.2.2, it's shown that

- for  $p = O(n^a)$ , Lasso and concave regularization methods are **asymptotically equivalent**, having the same convergence rates in the oracle inequalities, with a logarithmic factor of  $\log n$
- for  $\log p = O(n^a)$ , concave regularization methods are **asymptotically equivalent** and still enjoy the same convergence rates in the oracle inequalities, with a logarithmic factor of  $\log n^2$ .

A phase diagram on how the performance of regularization methods, in the thresholded parameter space, evolves with dimensionality and penalty function.

Further, we have the following **oracle risk inequalities** of the global minimizer

#### Theorem 14.2.3: Oracle Risk Inequalities

Assume that conditions of Thm.14.2.2 hold and the fourth moments of errors  $\mathbb{E}\epsilon_i^4$  are **uniformly bounded**. Then any global minimizer  $\hat{\beta}$  defined satisfies that

<sup>2</sup>For Lasso, the condition  $p'_\lambda(\tau) = O\left(\frac{\log n}{n}\right)$  and the choice of  $\lambda = c_0 \sqrt{\frac{\log p}{n}}$  are **incompatible** with each other in this ultra-high dimensional case, and the convergence rates for Lasso (of  $\log p$ ) are slower than those for concave regularization methods.

- **Sign risk**

$$\mathbb{E} \left[ \text{FS}(\hat{\beta}) \right] = \frac{1}{p_\lambda(\tau)} \left[ \left( \|p_\lambda(\beta_0)\|_1 + s\lambda^2 \right) O(n^{-c_1}) + O(p^{-c_1/2})\kappa_c \right]$$

- **Estimation and prediction risks**: If the pnnalty function further satisfies  $p'_\lambda(\tau) = O\left(\sqrt{\frac{\log n}{n}}\right)$ , then  $\forall q \in [1, 2]$

$$\mathbb{E} \|\hat{\beta} - \beta_0\|_q^q \leq Cs \left( \frac{\log n}{n} \right)^{q/2} \quad \mathbb{E} \|\hat{\beta} - \beta_0\|_\infty \leq C\gamma_n^* \sqrt{\frac{\log n}{n}} \quad \mathbb{E} \left[ \frac{1}{n} D(\hat{\beta}) \right] \leq Cs \frac{\log n}{n}$$

where C is some positive constant.

### How to understand Thm.14.2.3

- $\mathbb{E} \left[ \text{FS}(\hat{\beta}) \right]$  converges to 0 at a polynomial rate of  $n$
- Consistent with the risk bounds  $O\left(\frac{s \log n}{n}\right)$  of the regularized estimators under the  $L_2$ -loss in wavelets setting with orthogonal design
- No additional cost in risk bounds for generalizing to the ultra-high dimensional nonlinear model setting of GLM

## 14.3 Computability and Implementation

These properties are quite nice, but what about the computability? Specifically, what if computable solutions produced by an algorithm are not actually the **global** minimizer?

### Theorem 14.3.1: Asymptotic Properties of Computable solutions

Let  $\hat{\beta} \in \mathcal{B}_\tau$  be a computable solution to the minimization problem produced by any algorithm that is the global minimizer when **constrained on the subspace given by**  $\text{supp}(\hat{\beta})$  and  $\eta_n = \left\| \frac{1}{n} \mathbf{X}' [\mathbf{y} - \mu(\mathbf{X}\hat{\beta})] \right\|_\infty$ . Assume in addition that  $\exists c_4 > 0$  s.t.

$$\left\| \frac{1}{n} \mathbf{X}'_\alpha [\mu(\mathbf{X}\beta) - \mu(\mathbf{X}\beta_0)] \right\|_2 \geq c_4 \|\beta - \beta_0\|_2, \quad \forall \beta \in \mathcal{B}_\tau, \alpha = \text{supp}(\beta) \cup \text{supp}(\beta_0)$$

if the model is nonlinear. If  $\eta_n + \lambda = o(\tau)$  and  $\min_{1 \leq j \leq s} |\beta_{0,j}| > c_5 \sqrt{s}(\eta_n + \lambda)$  with a sufficiently large positive constant  $c_5$ , then  $\hat{\beta}$  enjoys the same asymptotic properties as for any global minimizer in Thm.14.2.1, 14.2.2, 14.2.3 under the same conditions therein.

With Thm.14.3.1, we have that a **computable solution** produced by any algorithm can share the same nice asymptotic properties as for any global minimizer, when the maximum correlation between the covariates and the residual vector  $\mathbf{y} - \mu(\mathbf{X}\hat{\beta})$  is a smaller order of the threshold  $\tau$ , where  $\mu(\theta) = (b'(\theta_1), \dots, b'(\theta_n))'$ .

**Implementation** Since computable solutions also have the nice asymptotic properties, we can then implement this algorithm. There are several ways to do so

- **Lasso-type methods**: LARS algorithm (Efron et al., 2004)
- **nonconcave penalized likelihood methods**: LQA algorithm (Fan and Li, 2001) and LLA algorithm (Zou and Li, 2008)
- **coordinate optimization** (Wu and Lange, 2008)

Another way is **ICA algorithm**: implementing **nonconcave penalized likelihood** methods with **second-order quadratic approximation** of the likelihood function and the coordination optimization (Fan and Lv, 2011).

- For each coordinate within each iteration, solve the univariate penalized least-squares problem with the corresponding quadratic approximation of the likelihood function, and update this coordinate only when the global minimizer has magnitude above the given threshold  $\tau$
- Thresholded parameter space naturally puts a constraint on each component, while also inducing additional sparsity of regularized estimate, making the algorithm converge faster

- **Stability**

- ★ Assume  $p_\lambda(t)$  has *maximum concavity*

$$\rho(p_\lambda) = \sup_{0 < t_1 < t_2 < \infty} \left\{ -\frac{p'_\lambda(t_2) - p'_\lambda(t_1)}{t_2 - t_1} < c \cdot c_2 \right\}$$

with constants  $c, c_2$

- ★ This ensures that  $Q_n(\beta)$  is *strictly convex* on a union of coordinate subspaces  $\{\beta \in \mathbb{R}^p : \|\beta\|_0 < \kappa_c\}$ , which is key to stability of sparse solution found by any algorithm
- ★ This condition holds for many concave penalties
  - $L_1$ -penalty  $p_\lambda(t) = \lambda t$  has maximum concavity

$$\rho(p_\lambda) = 0$$

- SCAD  $p_\lambda(t)$  has

$$\rho(p_\lambda) = (a - 1)^{-1}$$

- SICA  $p_\lambda(t; a) = \frac{\lambda(a+1)t}{a+t}$  has

$$\rho(p_\lambda) = 2\lambda(a^{-1} + a^{-2})$$

## References

- Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407 – 499, 2004. doi: 10.1214/009053604000000067. URL <https://doi.org/10.1214/009053604000000067>.
- Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.
- Jianqing Fan and Jinchi Lv. Nonconcave penalized likelihood with np-dimensionality. *IEEE Transactions on Information Theory*, 57(8):5467–5484, 2011.
- Tong Tong Wu and Kenneth Lange. Coordinate descent algorithms for lasso penalized regression. *The Annals of Applied Statistics*, pages 224–244, 2008.
- Zemin Zheng, Yingying Fan, and Jinchi Lv. High dimensional thresholded regression and shrinkage effect. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, pages 627–649, 2014.
- Hui Zou and Runze Li. One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics*, pages 1509–1533, 2008.