

## Topic 13: Non-convex Learning + Lasso

by Sai Zhang

**Key points:** Combining the best of the two, we can use **Lasso plus Concave** method, with Lasso screening and concave component selecting variables, achieving a coordinated intrinsic two-scale learning.

**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.

We are facing a tradeoff:

- **Convex** methods: have appealing prediction power and oracle inequalities, but challenging to provide tight false sign rate control
- **Concave** methods: have good variable selection properties, but challenging to establish global properties and risk properties

Here, we take advantage of the linearity of Lasso (convex *and* concave) and try to combine it with concave regularization to get the best of both.

### 13.1 Model Setup

Again, consider a linear regression model  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ , where

- response vector ( $n \times 1$ ):  $\mathbf{y} = (y_1, \dots, y_n)'$
- design matrix ( $n \times p$ ):  $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_p)$ , with each column rescaled to have  $L_2$ -norm  $n^{1/2}$

here, we consider a scenario where

- $\boldsymbol{\beta}_0 = (\beta_{0,1}, \dots, \beta_{0,p})'$  is *sparse* (with many 0 components)
- ultra-high dimensions:  $\log p = O(n^a)$ , for some  $0 < a < 1$

and consider the penalized least squares

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ (2n)^{-1} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_0 \|\boldsymbol{\beta}\|_1 + \|p_\lambda(\boldsymbol{\beta})\|_1 \right\} \quad (13.1)$$

where

- $\lambda_0 = c \left( \frac{\log p}{n} \right)^{1/2}$  for some  $c > 0$
- $p_\lambda(\boldsymbol{\beta}) = p_\lambda(|\boldsymbol{\beta}|) = (p_\lambda(|\beta_1|), \dots, p_\lambda(|\beta_p|))'$ , with  $|\boldsymbol{\beta}| = (|\beta_1|, \dots, |\beta_p|)'$ ; the concave penalty  $p_\lambda(t)$  is defined on  $t \in [0, \infty)$ , indexed by  $\lambda \geq 0$ , increasing in both  $t$  and  $\lambda$ ,  $p_\lambda(0) = 0$

the 2 penalty components

- $L_1$ -component: minimum amount of regularization for removing noise in prediction
- concave component  $\|p_\lambda(\boldsymbol{\beta})\|_1$ : adapt model sparsity for variable selection

Under this set up, we can derive the hard-thresholding property as

**Proposition 13.1.1: Hard-Thresholding Property**

Assume the  $p_\lambda(t)$ ,  $t \geq 0$ , is **increasing and concave** with

- $p_\lambda(t) \geq p_{H,\lambda}(t) = \frac{1}{2} [\lambda^2 - (\lambda - t)_+^2]$  on  $[0, \lambda]$
- $p'_\lambda((1 - c_1)\lambda) \leq c_1\lambda$  for some  $c_1 \in [0, 1]$
- $-p''_\lambda(t)$  decreasing on  $[0, (1 - c_1)\lambda]$

then any local minimizer of 13.1 that is also a global minimizer in each coordinate has the **hard-thresholding** feature that each component is either 0 or of magnitude **larger** than  $(1 - c_1)\lambda$

Such property is shared by a wide class of concave penalties, including hard-thresholding penalty  $p_{H,\lambda}(t)$  with  $c_1 = 0$ ,  $L_0$ -penalty, and SICA (with suitable  $c_1$ ).

**How to understand this proposition?** Let  $\hat{\beta} = (\hat{\beta}_1, \dots, \hat{\beta}_p)'$ , then **each  $\hat{\beta}_j$**  is the global minimizer of the corresponding univariate penalized least-square problem along the  $j$ -th coordinate. These univariate problems share a common form with (generally) different scalars  $z$

$$\hat{\beta}(z) = \arg \min_{\beta \in \mathbb{R}} \left\{ \frac{1}{2}(z - \beta)^2 + \lambda_0 |\beta| + p_{H,\lambda}(|\beta|) \right\}$$

after we rescale all covariates to have  $L_2$ -norm  $n^{1/2}$ . The solution to these univariate problems are

$$\hat{\beta}(z) = \text{sgn}(z)(|z| - \lambda_0) \cdot \mathbf{1}_{|z| > \lambda + \lambda_0}$$

these solutions have the same feature as the hard-thresholded estimator: each component is either 0 or of magnitude larger than  $\lambda$ . This provides a better distinction between insignificant and significant covariates than soft-thresholding by  $L_1$  penalty.

With the hard-thresholding property of Prop. 13.1.1, we can prove a basic constraint for the global optimum  $\hat{\beta}$  on an event with significant probability (Fan and Lv, 2014)

$$\|\delta_2\|_1 \leq 7\|\delta_1\|_1$$

where  $\delta = \hat{\beta} - \beta_0 = (\hat{\beta}'_1, \hat{\beta}'_2)' - (\beta'_{0,1}, \beta'_{0,2})' = (\delta'_1, \delta'_2)'$ , with  $\delta_1 \in \mathbb{R}^s$ . Where does this constraint come from? For the penalized least square question 13.1

$$\min_{\beta \in \mathbb{R}^p} \{ (2n)^{-1} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda_0 \|\beta\|_1 + \|p_\lambda(\beta)\|_1 \}$$

the global minimizer  $\hat{\beta}$  leads to

$$\begin{aligned} (2n)^{-1} \|\mathbf{y} - \mathbf{X}\hat{\beta}\|_2^2 + \lambda_0 \|\hat{\beta}\|_1 + \|p_\lambda(\hat{\beta})\|_1 &= (2n)^{-1} \|\mathbf{X}\beta_0 + \epsilon - \mathbf{X}\hat{\beta}\|_2^2 + \lambda_0 \|\hat{\beta}\|_1 + \|p_\lambda(\hat{\beta})\|_1 \\ &= (2n)^{-1} \|\epsilon - \mathbf{X}(\hat{\beta} - \beta_0)\|_2^2 + \lambda_0 \|\hat{\beta}\|_1 + \|p_\lambda(\hat{\beta})\|_1 \\ &\leq (2n)^{-1} \|\mathbf{y} - \mathbf{X}\beta_0\|_2^2 + \lambda_0 \|\beta_0\|_1 + \|p_\lambda(\beta_0)\|_1 \\ &= (2n)^{-1} \|\epsilon\|_2^2 + \lambda_0 \|\beta_0\|_1 + \|p_\lambda(\beta_0)\|_1 \end{aligned}$$

then, plug in  $\delta = \hat{\beta} - \beta_0$ , we get

$$\begin{aligned} (2n)^{-1} \|\epsilon - \mathbf{X}\delta\|_2^2 + \lambda_0 \|\beta_0 + \delta\|_1 + \|p_\lambda(\beta_0 + \delta)\|_1 &\leq (2n)^{-1} \|\epsilon\|_2^2 + \lambda_0 \|\beta_0\|_1 + \|p_\lambda(\beta_0)\|_1 \\ (2n)^{-1} \|\mathbf{X}\delta\|_2^2 - n^{-1} \epsilon' \mathbf{X}\delta + \lambda_0 \|\beta_0 + \delta\|_1 + \|p_\lambda(\beta_0 + \delta)\|_1 &\leq \lambda_0 \|\beta_0\|_1 + \|p_\lambda(\beta_0)\|_1 \end{aligned}$$

since  $\beta_{0,2} = \mathbf{0}$ ,  $\delta_2 = \beta_{0,2} + \delta_2$ , we have

$$\|\beta_0 + \delta\|_1 = \|\beta_{0,1} + \beta_{0,2} + \delta_1 + \delta_2\|_1 = \|\beta_{0,1} + \delta_1 + \delta_2\|_1 \leq \|\beta_{0,1} + \delta_1\|_1 + \|\delta_2\|_1$$

hence

$$(2n)^{-1} \|\mathbf{X}\delta\|_2^2 - n^{-1} \epsilon' \mathbf{X}\delta + \lambda_0 \|\delta_2\|_1 \leq \lambda_0 \|\beta_{0,1}\|_1 - \lambda_0 \|\beta_{0,1} + \delta_1\|_1 + \|p_\lambda(\beta_0)\|_1 - \|p_\lambda(\beta_0 + \delta)\|_1$$

and by the reverse triangle inequality  $\|\beta_{0,1}\|_1 - \|\beta_{0,1} + \delta_1\|_1 \leq \|\delta_1\|_1$ , we get

$$(2n)^{-1} \|\mathbf{X}\delta\|_2^2 - n^{-1} \epsilon' \mathbf{X}\delta + \lambda_0 \|\delta_2\|_1 \leq \lambda_0 \|\delta_1\|_1 + \|p_\lambda(\beta_0)\|_1 - \|p_\lambda(\beta_0 + \delta)\|_1$$

If assume the distribution of the model error  $\epsilon$  as

$$\Pr\left(\|n^{-1} \mathbf{X}' \epsilon\|_\infty > \frac{\lambda_0}{2}\right) = O(p^{-c_0})$$

conditional on the event  $\mathcal{E} = \{\|n^{-1} \mathbf{X}' \epsilon\|_\infty \leq \lambda_0/2\}$ , we have

$$-n^{-1} \epsilon' \mathbf{X}\delta + \lambda_0 \|\delta_2\|_1 - \lambda_0 \|\delta_1\|_1 \geq -\frac{\lambda_0}{2} \|\delta\|_1 + \lambda_0 \|\delta_2\|_1 - \lambda_0 \|\delta_1\|_1 = \frac{\lambda_0}{2} \|\delta_2\|_1 - \frac{3\lambda_0}{2} \|\delta_1\|_1$$

plug this result back, get

$$\frac{1}{2n} \|\mathbf{X}\delta\|_2^2 + \frac{\lambda_0}{2} \|\delta_2\|_1 \leq \frac{3\lambda_0}{2} \|\delta_1\|_1 + \|p_\lambda(\beta_0)\|_1 - \|p_\lambda(\beta_0 + \delta)\|_1$$

Now, if we further impose 2 conditions:

- **eigenvalue condition**: for some positive constant  $\kappa_0$

$$\min_{\|\delta\|_2=1, \|\delta\|_0 \leq 2s} \frac{1}{\sqrt{n}} \|\mathbf{X}\delta\|_2 \geq \kappa_0 \quad (\mathbf{A})$$

$$\kappa = \kappa(s, 7) = \min_{\delta \neq 0, \|\delta_2\|_1 \leq 7\|\delta_1\|_1} \left\{ \frac{1}{\sqrt{n}} \frac{\|\mathbf{X}\delta\|_2}{\|\delta_1\|_2 \vee \|\tilde{\delta}_2\|_2} \right\} > 0 \quad (\mathbf{B})$$

where  $\tilde{\delta}_2$  is the subvector of  $\delta_2$  consisting of the components with the  $s$  **largest** absolute values. Here

- Condition **(A)** is a mild sparse eigenvalue condition
- Condition **(B)** combines the restricted eigenvalue assumptions in [Bickel et al. \(2009\)](#)<sup>1</sup>. The intuition is, for OLS estimation,  $\mathbf{X}'\mathbf{X}$  should be **positive definite**, that is

$$\min_{\delta \neq 0 \in \mathbb{R}^p} \left\{ \frac{1}{\sqrt{n}} \frac{\|\mathbf{X}\delta\|_2}{\|\delta\|_2} \right\} > 0$$

however, when  $p > n$ , this condition **never** holds, hence we replace  $\|\delta\|_2$  with the  $L_2$ -norm of  $\|\delta_1\|_2$ , a subvector of  $\delta$

$$\kappa = \kappa(s, 7) = \min_{\delta \neq 0, \|\delta_2\|_1 \leq 7\|\delta_1\|_1} \left\{ \frac{1}{\sqrt{n}} \frac{\|\mathbf{X}\delta\|_2}{\|\delta_1\|_2} \right\} > 0$$

and for  $L_q$  loss with  $q \in (1, 2]$ , we further bound  $\|\tilde{\delta}_2\|_2$ , which leads to condition **(B)**.

- **hard-thresholding condition**: The penalty  $p_\lambda(t)$  satisfies the conditions of Prop. 13.1.1 with

$$p'_\lambda \{(1 - c_1)\lambda\} \leq \lambda_0/4$$

$$\min_{j=1, \dots, s} |\beta_{0,j}| > \max \left\{ (1 - c_1)\lambda, 2\kappa_0^{-1} p_\lambda^{1/2}(\infty) \right\}$$

<sup>1</sup>Introduced by [Candes and Tao \(2007\)](#) for studying the oracle inequalities for the Lasso estimator and Dantzig selector.

## References

- Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, pages 1705–1732, 2009.
- Emmanuel Candes and Terence Tao. The dantzig selector: Statistical estimation when  $p$  is much larger than  $n$ . *The Annals of Statistics*, 35(6):2313–2351, 2007.
- Yingying Fan and Jinchi Lv. Asymptotic properties for combined  $l_1$  and concave regularization. *Biometrika*, 101(1):57–70, 2014.