

## 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  $c_1$ ).

With Prop. 13.1.1,