High-dimensional Statistics — 7th Lecture

A. DALALYAN

MASTER MVA

0. RECAP

We observe $y = (y_1, \dots, y_n)^T \in \mathbb{R}^n$ and a matrix $\Phi \in \mathbb{R}^{n\times p}$ s.t.

(1) $$y = \Phi \cdot \beta^* + \xi \quad ; \quad \xi \sim \mathcal{N}(0, \sigma^2 I_n)$$

The unknown parameter β^* is high dimensional but sparse. This means that p is large but $\|\beta^*\|_0 := s$ is small. We assume that the columns of Φ are normalized in such a way that $\frac{1}{n}\|\Phi^j\|_2^2 = 1$ où Φ^j is the j^{th} column of Φ.

We estimate β^* by the Lasso:

(2) $$\hat\beta_\lambda^{Lasso} \in \arg\min_{\beta\in\mathbb{R}^p} \left\{ \frac{1}{2n}\|y - \Phi\beta\|_2^2 + \lambda\cdot\|\beta\|_1 \right\}$$

The quality of this estimator will be measured by the risk:

$$\ell_n(\hat\beta, \beta^*) = \frac{1}{n}\|\Phi(\hat\beta - \beta^*)\|_2^2.$$

We proved last week that if $\lambda = \sigma\sqrt{\frac{2}{n}\ln(p/\delta)}$, then

(3) $$\ell_n(\hat\beta, \beta^*) \leq \inf_{\beta\in\mathbb{R}^p} \left\{ \ell_n(\beta, \beta^*) + 4\sigma\|\beta\|_1 \sqrt{\frac{2\ln(p/\delta)}{n}} \right\}$$

with probability $\geq 1 - \delta$. Since the right-hand side is smaller than the same expression with β^* instead of β, we get

$$\ell_n(\hat\beta, \beta^*) \leq 4\sigma\|\beta^*\|_\infty \times \underbrace{\frac{s}{\sqrt{n}} \times \sqrt{2\ln(p/\delta)}}_{\text{"slow" rate.}},$$

<u>Question</u>: Is it possible to improve the rate of convergence of the Lasso?

1. FAST RATES FOR LASSO

Unfortunately, it is impossible to improve the rate of convergence of the Lasso without additional assumptions on Φ. Indeed, it is shown in DALALYAN, HEBIRI & LEDERER (2014) that if $p = \sqrt{8n}$ and

$$\Phi = \sqrt{\frac{n}{2}} \times \begin{bmatrix} 1 & 1 & 1 & 1 & \cdots & 1 & 1 \\ 1 & -1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & -1 & & & \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & -1 \end{bmatrix}$$

with $\beta^* = (1, 1, 0, \dots, 0)^T$ and $\sigma = 1$, then for any $\lambda > 0$

$$\mathbb{P}\left(\ell_n(\hat\beta_\lambda^{\text{Lasso}}, \beta^*) \geq \frac{1}{\sqrt{8n}} \right) \geq \frac{1}{2}.$$

This shows that the risk of the Lasso cannot be smaller than $c \cdot n^{-1/2}$ irrespectively of the choice of the tuning parameter λ.

However, an attractive property of the Lasso is that it does converge with a faster rate if the matrix Φ satisfies some assumptions. Here, we will only focus on matrices Φ that satisfy the Restricted Eigenvalue (RE) condition [Bickel et al. 2009]

DEF. We say that Φ satisfies the condition $RE(c_0, s)$, where $c_0 > 0$ and $s \in \mathbb{N}$, if $\exists\, \varkappa > 0$ such that for all $J \subset \{1, \dots, p\}$ satisfying $\mathrm{Card}(J) \leq s$ we have

$$\|u_{J^c}\|_1 \leq c_0 \|u_J\|_1 \quad \Longrightarrow \quad \frac{\|\Phi u\|_2^2}{n \|u_J\|_2^2} \geq \varkappa$$

Here, $J^c = \{1, \dots, p\} \setminus J$ and u_J is the restriction of the vector u on the index-set J.

Remarks

① If Φ is orthogonal, i.e., $\frac{1}{n}\Phi^T\Phi = I_p$, then Φ satisfies $RE(+\infty, p)$ with $\varkappa = 1$.

② The classe $\mathcal{F}(c_0, s)$ of matrices Φ satisfying $RE(c_0, s)$ is decreasing w.r.t. both c_0 and s, that is

$$c_0 \leq c_0' \ \& \ s \leq s' \implies \mathcal{F}(c_0, s) \supset \mathcal{F}(c_0', s').$$

THEOREM (Fast rates)

If the matrix Φ satisfies the condition $RE(3, s)$, and the tuning parameter is chose so that $\lambda \geq 2\sigma\sqrt{\frac{2}{n}\log(p/\delta)}$, then with probability $\geq 1-\delta$ it holds that

$$\ell_n(\hat{\beta}^L, \beta^*) \leq \inf_{\beta \in \mathbb{R}^p} \min_{|J| \leq s} \left\{ \ell_n(\beta, \beta^*) + 4\lambda \|\beta_{J^c}\|_1 + \frac{9\lambda^2 s}{4\varkappa} \right\}.$$

Proof. Since $\hat{\beta}^L$ is a solution of (2), we have

$$0 \in \partial \left\{ \frac{1}{2n} \|y - \Phi\hat{\beta}^L\|_2^2 + \lambda \|\hat{\beta}^L\|_1 \right\}$$

where ∂ designes the subdifferential. Since both terms above are convex functions of β, the subdifferential of the sum is the sum of subdifferentials. This leads to

$$(5) \quad 0 \in -\frac{1}{n}\Phi^T(y - \Phi\hat{\beta}^L) + \lambda \cdot \mathrm{sign}(\hat{\beta}^L)$$

where $\mathrm{sign}(\beta)$ is the subdifferential of $\beta \mapsto \|\beta\|_1$ defined as

$$\mathrm{sign}(\beta) = \left\{ u \in \mathbb{R}^p : \begin{array}{l} u_j = 1 \text{ if } \beta_j > 0, \ u_j = -1 \text{ if } \beta_j < 0 \text{ and} \\ u_j \in [-1, 1] \text{ if } \beta_j = 0 \end{array} \right\}$$

As a consequence of (5), there is $u \in \mathrm{sign}(\hat{\beta})$ such that (in what follows we write $\hat{\beta}$ instead of $\hat{\beta}^{Lasso} = \hat{\beta}^L$)

$$(6) \qquad \frac{1}{n} \Phi^T (y - \Phi\hat{\beta}) = \lambda \cdot u$$

Taking the scalar product with any $\beta \in \mathbb{R}^p$, we get

$$(7) \quad \frac{1}{n} \beta^T \Phi^T (y - \Phi\hat{\beta}) = \lambda \cdot \beta^T u \leq \lambda \cdot \|\beta\|_1 \quad (\text{since } |u_j| \leq 1)$$

Similarily, taking the scalar product with $\hat{\beta}$, we get

$$(8) \quad \frac{1}{n} \hat{\beta}^T \Phi^T (y - \Phi\hat{\beta}) = \lambda \cdot \hat{\beta}^T u = \lambda \cdot \|\hat{\beta}\|_1 \quad (\text{since } u \in \mathrm{sign}(\hat{\beta}))$$

Combining (7) and (8), we arrive at

$$(9) \quad \frac{1}{n} (\beta - \hat{\beta})^T \Phi^T (y - \Phi\hat{\beta}) \leq \lambda \left(\|\beta\|_1 - \|\hat{\beta}\|_1 \right) \quad \forall \beta \in \mathbb{R}^p$$

We replace now y by its expression $\Phi\beta^* + \xi$ and rearrange the terms:

$$\frac{1}{n} (\beta - \hat{\beta})^T \Phi^T \Phi (\beta^* - \hat{\beta}) \leq \frac{1}{n} (\hat{\beta} - \beta)^T \Phi^T \xi + \lambda \left(\|\beta\|_1 - \|\hat{\beta}\|_1 \right)$$

Using the duality inequality, we find that on the event

$$\mathcal{B} = \left\{ \frac{1}{n} \|\Phi^T \xi\|_\infty \leq \sigma \sqrt{\frac{2}{n} \ln(p/\delta)} \right\} = \left\{ \frac{1}{n} \|\Phi^T \xi\|_\infty \leq \lambda/2 \right\}$$

(We have proved during the previous lecture that $\mathbb{P}(\mathcal{B}) \geq \delta$)

we have

$$\frac{1}{n} (\hat{\beta} - \beta)^T \Phi^T \xi \leq \|\hat{\beta} - \beta\|_1 \times \frac{1}{n} \|\Phi^T \xi\|_\infty \leq \frac{\lambda}{2} \|\hat{\beta} - \beta\|_1 .$$

This leads to

$$(10) \quad \frac{1}{n} (\beta - \hat{\beta})^T \Phi^T \Phi (\beta^* - \hat{\beta}) \leq \frac{\lambda}{2} \left\{ \|\hat{\beta} - \beta\|_1 + 2\|\beta\|_1 - 2\|\hat{\beta}\|_1 \right\}$$

One can notice that the left-hand side in (10) is:

$$\frac{1}{n}(\beta-\hat\beta)^T\Phi^T\Phi(\beta^*-\hat\beta) = \frac{1}{2n}\|\Phi(\beta^*-\hat\beta)\|_2^2 + \frac{1}{2n}\|\Phi(\beta-\hat\beta)\|_2^2 - \frac{1}{2n}\|\Phi(\beta^*-\beta)\|_2^2$$

$$= \frac{1}{2}\left\{ l_n(\hat\beta,\beta^*) - l_n(\beta,\beta^*) + \frac{1}{n}\|\Phi(\beta-\hat\beta)\|_2^2 \right\}$$

Therefore, we arrive at

$$(11)\quad l_n(\hat\beta,\beta^*) \leq l_n(\beta,\beta^*) + \lambda\left\{\|\hat\beta-\beta\|_1 + 2\|\beta\|_1 - 2\|\hat\beta\|_1\right\} - \frac{1}{n}\|\Phi(\beta-\hat\beta)\|_2^2$$

Let us set $u = \hat\beta-\beta$ and let $J \subset \{1,\dots,p\}$ be any set of $\mathrm{card}(J) \leq s$. One easily check that

$$\|\hat\beta-\beta\|_1 + 2\|\beta\|_1 - 2\|\hat\beta\|_1 = \|u_J\|_1 + \|u_{J^c}\|_1 + 2\|\beta_J\|_1 - 2\|\hat\beta_J\|_1 + 2\|\beta_{J^c}\|_1 - 2\|\hat\beta_{J^c}\|_1$$

$$\leq \|u_J\|_1 + \|u_{J^c}\|_1 + 2\|u_J\|_1 - 2\|u_{J^c}\|_1 + 4\|\beta_{J^c}\|_1$$

$$(12)\quad = 3\|u_J\|_1 - \|u_{J^c}\|_1 + 4\|\beta_{J^c}\|_1 .$$

We consider two cases.

If $3\|u_J\|_1 \leq \|u_{J^c}\|_1$, then (11) implies that

$$l_n(\hat\beta,\beta^*) \leq l_n(\beta,\beta^*) + 4\lambda\|\beta_{J^c}\|_1 .$$

If $\|u_{J^c}\|_1 < 3\|u_J\|_1$, then $RE(3,s)$ yields

$$\|u_J\|_2^2 \leq \frac{1}{n\varkappa}\|\Phi u\|_2^2$$

Hence, in view of (11), and (12), we get

$$l_n(\hat\beta,\beta^*) \leq l_n(\beta,\beta^*) + 4\lambda\|\beta_{J^c}\|_1 + 3\lambda\|u_J\|_1 - \frac{1}{n}\|\Phi u\|_2^2$$

$$\leq l_n(\beta,\beta^*) + 4\lambda\|\beta_{J^c}\|_1 + 3\lambda\sqrt{s}\cdot\|u_J\|_2 - \frac{1}{n}\|\Phi u\|_2^2$$

$$\leq l_n(\beta,\beta^*) + 4\lambda\|\beta_{J^c}\|_1 + \frac{3\lambda\sqrt{s}}{\sqrt{n\varkappa}}\cdot\|\Phi u\|_2 - \frac{1}{n}\|\Phi u\|_2^2 .$$

If we set $x = \frac{1}{\sqrt{n}} \|\Phi u\|_2$, then the last two terms are

$$3\lambda \sqrt{\frac{s}{\varkappa}} \cdot x - x^2 = \left(\frac{3\lambda}{2}\sqrt{\frac{s}{\varkappa}}\right)^2 - \left(\frac{3\lambda}{2}\sqrt{\frac{s}{\varkappa}} - x\right)^2$$

$$\leq \frac{9\lambda^2 s}{4\varkappa} .$$

Thus, in this second case,

$$\ell_n(\hat{\beta}, \beta^*) \leq \ell_n(\beta, \beta^*) + 4\lambda \|\beta_{J^c}\|_1 + \frac{9\lambda^2 s}{4\varkappa} . \quad \blacksquare$$