Topics in High-Dimensional Statistics

Lecture 2: Performance analysis of the LASSO estimator

Contents

This lecture explores the performance of the LASSO estimator. Recall from Lecture 1 that, in matrix notation, the LASSO estimator of μ^* is defined by

$$\hat{\mu}^{\text{lasso}} = \mathbf{X}\hat{\beta}^{\text{lasso}} \quad \text{where} \quad \hat{\beta}^{\text{lasso}} \in \arg\min_{\beta \in \mathbb{R}^d} \{ \frac{1}{n}\|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + 2\lambda\|\beta\|_1 \}.$$

We argued in Lecture 1 that the LASSO estimator was a computationally friendly alternative to the BIC estimator with comparable statistical performance. In this lecture, we'll focus on describing this statistical performance, and show how it depends the structure of the design matrix $\mathbf{X}$.

1 First order optimality condition

The analysis relies on a few general results from convex analysis. We review them briefly in this section.

Recall that a subset $C \subset \mathbb{R}^d$ is convex if, for all $x, y \in C$ and for all $\lambda \in [0,1]$, $(1-\lambda)x + \lambda y \in C$. Next is an important characterization of convex sets.

Theorem 1.1 (Supporting hyperplane)**.** *Let $C \subset \mathbb{R}^d$ be a convex set and $x_0 \in \partial C$ be a point on its boundary. Then, there exists $u \in \mathbb{R}^d$, $u \neq 0$, such that for all $x \in C$, $u^\top x_0 \leq u^\top x$.*

Given a convex set $C \subset \mathbb{R}^d$, recall that a function $f : C \to \mathbb{R}$ is convex if, for all $x, y \in C$ and for all $\lambda \in [0,1]$,

$$f((1-\lambda)x + \lambda y) \leq (1-\lambda)f(x) + \lambda f(y).$$

One checks that the function $f : C \to \mathbb{R}$ is convex if and only if the epigraph of f, i.e.

$$\text{epi}(f) = \{(x,t) \in C \times \mathbb{R} : f(x) \leq t\},$$

is a convex subset of $\mathbb{R}^d \times \mathbb{R}$.

Definition 1.2 (Subgradients)**.** *Given a set $C \subset \mathbb{R}^d$ and a function $f : C \to \mathbb{R}$, a vector $g \in \mathbb{R}^d$ is called a subgradient of f at $x \in C$ if,*

$$\forall y \in C, \quad f(y) - f(x) \geq g^\top (y - x).$$

The set of all subgradients of f at x is denoted $\partial f(x)$ and called the subdifferential of f at x.

Theorem 1.3. *Let $C \subset \mathbb{R}^d$ be a convex set and $f : C \to \mathbb{R}$ be a function.*
(1) *The function f is convex if, for all $x \in C$, $\partial f(x) \neq \emptyset$.*
(2) *If f is convex then, for all $x \in \mathrm{int}(C)$, $\partial f(x) \neq \emptyset$.*
(3) *If f is convex and differentiable, then for all $x \in \mathrm{int}(C)$, $\partial f(x) = \{\nabla f(x)\}$.*
(4) *If f is convex, then for all $x, y \in \mathrm{int}(C)$, all $g_x \in \partial f(x)$ and all $g_y \in \partial f(y)$,*

$$(g_x - g_y)^\top (x - y) \geq 0.$$

Proof. (1) Let $x, y \in C$ and $\lambda \in [0, 1]$. Since there exists $g \in \partial f((1-\lambda)x+\lambda y)$, it follows by definition of a subgradient that

$$f(x) - f((1-\lambda)x + \lambda y) \geq \lambda g^\top (y - x),$$

and

$$f(y) - f((1-\lambda)x + \lambda y) \geq (1-\lambda) g^\top (x - y).$$

Multiplying the first inequality by $(1-\lambda)$, the second by λ and summing the obtained inequalities, we obtain that $f((1-\lambda)x+\lambda y) \leq (1-\lambda)f(x)+\lambda f(y)$. Since this holds for all $x, y \in C$ and all $\lambda \in [0, 1]$, we deduce that f is convex.
(2) Let $x \in C$. The point $(x, f(x))$ belongs to $\partial \mathrm{epi}(f)$. Since $\mathrm{epi}(f)$ is a convex set, we deduce from Theorem 1.1 that there exists $(a, b) \in \mathbb{R}^d \times \mathbb{R}$, $(a, b) \neq (0, 0)$, such that

$$\forall (y, t) \in \mathrm{epi}(f), \quad a^\top x + b f(x) \geq a^\top y + bt. \tag{1.1}$$

Observe that $(y, t) \in \mathrm{epi}(f)$ implies that $(y, t') \in \mathrm{epi}(f)$ for all $t' \geq t$. Hence, for any $y \in C$ the above inequality should hold true for any $t \geq f(y)$ and in particular when $t \to +\infty$ which imposes that $b \leq 0$. Now suppose that $x \in \mathrm{int}(C)$. Then, for $\varepsilon > 0$ small enough, the point $z = x + \varepsilon a$ belongs to C so that, for all $t \geq f(z)$,

$$a^\top x + b f(x) \geq a^\top z + bt \Leftrightarrow b f(x) \geq \varepsilon \|a\|^2 + bt.$$

If $b = 0$ we deduce that $a = 0$ which is a contradiction. Hence $b < 0$. Now for any $y \in C$, writing (1.1) for $t = f(y)$ implies that

$$f(y) - f(x) \geq \frac{a^\top (y - y)}{|b|},$$

which shows that $a/|b| \in \partial f(x)$.
(3) Suppose that f is convex, differentiable and take $x \in \mathrm{int}(C)$. For any

$h \in \mathbb{R}^d$ and $t \in \mathbb{R}$ small enough so that both $x \pm th \in C$, a Taylor expansion of f around x reveals that

$$f(x \pm th) = f(x) \pm t\nabla f(x)^\top h + \mathrm{o}(t).$$

Now for any $g \in \partial f(x)$, we have by definition of a subgradient that

$$f(x \pm th) \geq f(x) \pm tg^\top h.$$

In particular, we deduce that

$$\pm t\nabla f(x)^\top h + \mathrm{o}(t) \geq \pm tg^\top h.$$

This imposes finally that, for all $h \in \mathbb{R}^d$, $\nabla f(x)^\top h = g^\top h$ which implies that $g = \nabla f(x)$.
(4) For all $x, y \in \mathrm{int}(C)$, all $g_x \in \partial f(x)$ and all $g_y \in \partial f(y)$, summing the inequalities $f(x) - f(y) \geq g_y^\top (x - y)$ and $f(y) - f(x) \geq g_x^\top (y - x)$ easily provides the last property. □

Theorem 1.4 (First order optimality condition)**.** *Let $C \subset \mathbb{R}^d$ be a convex set and $f : C \to \mathbb{R}$ be a convex function. Then*

$$x^* \in \arg\min_{x \in C} f(x) \quad \Leftrightarrow \quad 0 \in \partial f(x^*).$$

Proof. Both conditions are equivalent to the fact that $f(x) \geq f(x^*) + 0^\top (x - x^*)$, for all $x \in C$. □

Example 1.5. *For any $t \in \mathbb{R}$, denote*

$$\mathrm{sign}(t) := \begin{cases} \{+1\} & \text{if } t > 0, \\ [-1, 1] & \text{if } t = 0, \\ \{-1\} & \text{if } t < 0. \end{cases}$$

For $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$, define

$$\mathrm{sign}(x) := \prod_{j=1}^{d} \mathrm{sign}(x_j).$$

Consider $f(x) = \|x\|_1$. Then, for all $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$.

$$\partial f(x) := \mathrm{sign}(x).$$

2 A general result

We introduce some notation. Recall that $[d] = \{1, \dots, d\}$. For any vector $\beta \in \mathbb{R}^d$ and any $J \subset [d]$, we denote $J^c = [d] \setminus J$ and $\beta_J \in \mathbb{R}^d$ the vector defined, for $j \in [d]$, by

$$(\beta_J)_j := \begin{cases} \beta_j & \text{if } j \in J, \\ 0 & \text{if } j \notin J. \end{cases}$$

Then, for any $c > 0$ and any $J \subset [d]$, we define the restricted eigenvalue constant $\mathrm{RE}(c, J)$ with parameters c and J by

$$\mathrm{RE}(c, J) := \inf \left\{ \frac{\|\mathbf{X}\beta\|_2^2}{n\|\beta_J\|_2^2} : \beta \in \mathbb{R}^d,\ \|\beta_{J^c}\|_1 < c\|\beta_J\|_1 \right\}.$$

We are now in position to state a first important result.

Theorem 2.1. *Fix $\lambda > 0$. Then, on the event*

$$\left\{ \frac{1}{n}\|\mathbf{X}^\top \boldsymbol{\xi}\|_\infty \le \frac{\lambda}{2} \right\}, \tag{2.1}$$

the Lasso estimator $\hat{\beta} = \hat{\beta}^{\mathrm{lasso}}$ satisfies, for all $n \ge 1$,

$$\begin{aligned} &\frac{1}{n}\|\mathbf{X}(\hat{\beta} - \beta^*)\|_2^2 \\ &\le \inf_{\beta \in \mathbb{R}^d,\, J \subset [d]} \left\{ \frac{1}{n}\|\mathbf{X}(\beta - \beta^*)\|_2^2 + 4\lambda\|\beta_{J^c}\|_1 + \frac{9}{4}\frac{\lambda^2|J|}{\mathrm{RE}(3, J)} \right\}. \end{aligned}$$

Proof. First, denote $\mathcal{L} : \mathbb{R}^d \to \mathbb{R}$ the LASSO objective function, i.e.,

$$\mathcal{L}(\beta) = \frac{1}{n}\|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + 2\lambda\|\beta\|_1.$$

The function $\mathcal{L}$ is convex. It then follows from Theorem 1.4 that

$$0 \in \partial\mathcal{L}(\hat{\beta}). \tag{2.2}$$

Using Example 1.5, we obtain[1] for all $\beta \in \mathbb{R}^d$,

$$\partial\mathcal{L}(\beta) = \frac{2}{n}\mathbf{X}^\top(\mathbf{X}\beta - \mathbf{Y}) + 2\lambda\,\mathrm{sign}(\beta). \tag{2.3}$$

[1] For $a \in \mathbb{R}^d$ and $B \subset \mathbb{R}^d$, notation $a + B$ denotes $\{a + b : b \in B\}$.

Combining equations (2.2) and (2.3) yield

$$\frac{1}{n}\mathbf{X}^\top(\mathbf{Y}-\mathbf{X}\hat{\beta}) \in \lambda \operatorname{sign}(\hat{\beta}). \tag{2.4}$$

Considering the inner product with $\hat{\beta}$ in equation (2.4), we obtain that

$$\frac{1}{n}\hat{\beta}^\top\mathbf{X}^\top(\mathbf{Y}-\mathbf{X}\hat{\beta}) = \lambda\|\hat{\beta}\|_1. \tag{2.5}$$

Also, for all $\beta \in \mathbb{R}^d$, it follows from (2.4) that

$$\frac{1}{n}\beta^\top\mathbf{X}^\top(\mathbf{Y}-\mathbf{X}\hat{\beta}) \in [-\lambda\|\beta\|_1, \lambda\|\beta\|_1]. \tag{2.6}$$

As a result, for any $\beta \in \mathbb{R}^d$, substracting (2.5) to (2.6) we get that,

$$\frac{1}{n}(\beta-\hat{\beta})^\top\mathbf{X}^\top(\mathbf{Y}-\mathbf{X}\hat{\beta}) \leq \lambda(\|\beta\|_1 - \|\hat{\beta}\|_1). \tag{2.7}$$

Now by expanding $\mathbf{Y} = \mathbf{X}\beta^* + \boldsymbol{\xi}$ on the left hand side of (2.7), we obtain that for any $\beta \in \mathbb{R}^d$,

$$\begin{aligned}
&\frac{1}{n}(\beta-\hat{\beta})^\top\mathbf{X}^\top\mathbf{X}(\beta^*-\hat{\beta}) \\
&\leq \frac{1}{n}(\hat{\beta}-\beta)^\top\mathbf{X}^\top\boldsymbol{\xi} + \lambda(\|\beta\|_1 - \|\hat{\beta}\|_1) \\
&\leq \frac{1}{n}\|\mathbf{X}^\top\boldsymbol{\xi}\|_\infty\|\hat{\beta}-\beta\|_1 + \lambda(\|\beta\|_1 - \|\hat{\beta}\|_1).
\end{aligned} \tag{2.8}$$

Lets now work on the event $\{\|\mathbf{X}^\top\boldsymbol{\xi}\|_\infty \leq n\lambda/2\}$. On this event, we get that

$$\begin{aligned}
&\frac{1}{n}(\beta-\hat{\beta})^\top\mathbf{X}^\top\mathbf{X}(\beta^*-\hat{\beta}) \\
&\leq \frac{\lambda}{2}(\|\hat{\beta}-\beta\|_1 + 2\|\beta\|_1 - 2\|\hat{\beta}\|_1).
\end{aligned} \tag{2.9}$$

Now, using the notation introduced before the Theorem, observe that for all $J \subset [d]$:

$$\begin{aligned}
&\|\hat{\beta}-\beta\|_1 + 2\|\beta\|_1 - 2\|\hat{\beta}\|_1 \\
&= \|(\hat{\beta}-\beta)_J\|_1 + \|(\hat{\beta}-\beta)_{J^c}\|_1 + 2\|\beta_J\|_1 \\
&+ 2\|\beta_{J^c}\|_1 - 2\|\hat{\beta}_J\|_1 - 2\|\hat{\beta}_{J^c}\|_1.
\end{aligned} \tag{2.10}$$

From the inequalities $\|\beta_J\|_1 - \|\hat{\beta}_J\|_1 \leq \|(\hat{\beta} - \beta)_J\|_1$ and $\|\hat{\beta}_{J^c}\|_1 \geq \|(\hat{\beta} - \beta)_{J^c}\|_1 - \|\beta_{J^c}\|_1$, we deduce from equation (2.10) that

$$\begin{aligned} &\|\hat{\beta} - \beta\|_1 + 2\|\beta\|_1 - 2\|\hat{\beta}\|_1 \\ &\leq 3\|(\hat{\beta} - \beta)_J\|_1 - \|(\hat{\beta} - \beta)_{J^c}\|_1 + 4\|\beta_{J^c}\|_1. \end{aligned} \tag{2.11}$$

Combining inequalities (2.9) and (2.11) yields that, on the event $\{\|\mathbf{X}^\top \boldsymbol{\xi}\|_\infty \leq n\lambda/2\}$, and for any $J \subset [d]$ and any $\beta \in \mathbb{R}^d$,

$$\begin{aligned} &\frac{1}{n}(\beta - \hat{\beta})^\top \mathbf{X}^\top \mathbf{X}(\beta^* - \hat{\beta}) \\ &\leq \frac{\lambda}{2}\left\{3\|(\hat{\beta} - \beta)_J\|_1 - \|(\hat{\beta} - \beta)_{J^c}\|_1\right\} + 2\lambda\|\beta_{J^c}\|_1. \end{aligned} \tag{2.12}$$

Then, by definition of the restricted eigenvalue constant $\mathrm{RE}(3, J)$, we deduce that, if $3\|(\hat{\beta} - \beta)_J\|_1 > \|(\hat{\beta} - \beta)_{J^c}\|_1$,

$$\begin{aligned} 3\|(\hat{\beta} - \beta)_J\|_1 - \|(\hat{\beta} - \beta)_{J^c}\|_1 &\leq 3\|(\hat{\beta} - \beta)_J\|_1 \\ &\leq 3\sqrt{|J|} \cdot \|(\hat{\beta} - \beta)_J\|_2 \\ &\leq 3\sqrt{\frac{|J|}{n\mathrm{RE}(3, J)}} \cdot \|\mathbf{X}(\hat{\beta} - \beta)\|_2. \end{aligned}$$

Notice next that if $3\|(\hat{\beta} - \beta)_J\|_1 \leq \|(\hat{\beta} - \beta)_{J^c}\|_1$ then the last inequality trivially holds. As a result, it follows from (2.12) that, for all $J \subset [d]$ and all $\beta \in \mathbb{R}^d$,

$$\begin{aligned} &\frac{1}{n}(\beta - \hat{\beta})^\top \mathbf{X}^\top \mathbf{X}(\beta^* - \hat{\beta}) \\ &\leq \frac{3\lambda}{2}\sqrt{\frac{|J|}{n\mathrm{RE}(3, J)}}\|\mathbf{X}(\hat{\beta} - \beta)\|_2 + 2\lambda\|\beta_{J^c}\|_1. \end{aligned} \tag{2.13}$$

To complete the proof, a few more tricks are in order. First, using in (2.13) the identity $2u^\top v = \|u\|_2^2 + \|v\|_2^2 - \|u - v\|_2^2$ with $u = \mathbf{X}(\beta - \hat{\beta})$ and $v = \mathbf{X}(\beta^* - \hat{\beta})$, yields

$$\begin{aligned} &\frac{1}{2n}\|\mathbf{X}(\hat{\beta} - \beta^*)\|_2^2 \\ &\leq \frac{1}{2n}\|\mathbf{X}(\beta - \beta^*)\|_2^2 + 2\lambda\|\beta_{J^c}\|_1 \\ &+ \frac{3\lambda}{2}\sqrt{\frac{|J|}{n\mathrm{RE}(3, J)}}\|\mathbf{X}(\hat{\beta} - \beta)\|_2 - \frac{1}{2n}\|\mathbf{X}(\hat{\beta} - \beta)\|_2^2. \end{aligned} \tag{2.14}$$

Finally, using the inequality $ax - bx^2 \leq a^2/4b$, we may upper bound the sum of the last two terms in (2.14) by

$$\frac{9}{8}\frac{|J|\lambda^2}{\mathrm{RE}(3, J)},$$

and the proof is complete. □

3 Slow rates for the LASSO

Theorem 2.1 is very general and can be used to deduce more explicit results. The first of these results is presented below. We denote

$$\varkappa := \max_j \|\mathbf{X}_j\|_2$$

where $\mathbf{X}_j \in \mathbb{R}^n$ denotes the j-th column of the design matrix $\mathbf{X}$.

Lemma 3.1. *Suppose that the noise vector $\boldsymbol{\xi} \in \mathbb{R}^n$ is sub-gaussian with variance proxy upper bounded by $\sigma^2 > 0$. Fix $\delta \in (0, 1)$. Then, for any*

$$\lambda \geq \frac{2\varkappa\sigma}{n}\sqrt{2\log\left(\frac{2d}{\delta}\right)},$$

we have

$$\mathbb{P}\left(\frac{1}{n}\|\mathbf{X}^\top\boldsymbol{\xi}\|_\infty \leq \frac{\lambda}{2}\right) \geq 1 - \delta.$$

Proof. Left as an exercise. □

Theorem 3.2 (Slow rates)**.** *Suppose that the noise vector $\boldsymbol{\xi} \in \mathbb{R}^n$ is sub-gaussian with variance proxy upper bounded by $\sigma^2 > 0$. Suppose that $\varkappa \leq \sqrt{n}$. Fix $\delta \in (0, 1)$. Then, for*

$$\lambda = 2\sigma\sqrt{\frac{2}{n}\log\left(\frac{2d}{\delta}\right)},$$

the LASSO estimator with parameter λ satisfies

$$\frac{1}{n}\|\mathbf{X}(\hat{\beta} - \beta^*)\|_2^2 \leq 8\sigma\|\beta^*\|_1\sqrt{\frac{2}{n}\log\left(\frac{2d}{\delta}\right)},$$

with probability at least $1 - \delta$.

Proof. Since $\inf \emptyset = +\infty$, it follows that $\textsc{re}(3, \emptyset) = +\infty$ for all $c > 0$. Hence, for any $\lambda \geq 0$, selecting $\beta = \beta^*$ and $J = \emptyset$ (and therefore $J^c = [d]$), it follows that

$$\begin{aligned}
&\inf_{\beta \in \mathbb{R}^d,\, J \subset [d]} \left\{ \frac{1}{n} \|\mathbf{X}(\beta - \beta^*)\|_2^2 + 4\lambda \|\beta_{J^c}\|_1 + \frac{9}{4} \frac{\lambda^2 |J|}{\textsc{re}(3, J)} \right\} \\
&\leq \frac{1}{n} \|\mathbf{X}(\beta^* - \beta^*)\|_2^2 + 4\lambda \|\beta^*\|_1 + \frac{9}{4} \frac{\lambda^2 |\emptyset|}{\textsc{re}(3, \emptyset)} \\
&= 4\lambda \|\beta^*\|_1.
\end{aligned}$$

The result then follows by combining Theorem 2.1 with Lemma 3.1. □

4 Fast rates for the LASSO

Denote $J^* \subset [d]$ the support of β^*, i.e., the set of non-zero coordinates of β^*. Then, we have the following.

Theorem 4.1 (Fast rates)**.** *Suppose that the noise vector $\boldsymbol{\xi} \in \mathbb{R}^n$ is sub-gaussian with variance proxy upper bounded by $\sigma^2 > 0$. Suppose that $\varkappa \leq \sqrt{n}$. Fix $\delta \in (0, 1)$. Then, for*

$$\lambda = 2\sigma \sqrt{\frac{2}{n} \log\left(\frac{2d}{\delta}\right)},$$

the LASSO estimator with parameter λ satisfies

$$\frac{1}{n} \|\mathbf{X}(\hat{\beta} - \beta^*)\|_2^2 \leq \frac{18\sigma^2 \|\beta^*\|_0}{n \textsc{re}(3, J^*)} \log\left(\frac{2d}{\delta}\right),$$

with probability at least $1 - \delta$.

Proof. By definition of the support J^* of β^*, we have $|J^*| = \|\beta^*\|_0$ and $\|\beta^*_{(J^*)^c}\|_1 = 0$. Hence, for any $\lambda \geq 0$, selecting $\beta = \beta^*$ and $J = J^*$, we get

$$\begin{aligned}
&\inf_{\beta \in \mathbb{R}^d,\, J \subset [d]} \left\{ \frac{1}{n} \|\mathbf{X}(\beta - \beta^*)\|_2^2 + 4\lambda \|\beta_{J^c}\|_1 + \frac{9}{4} \frac{\lambda^2 |J|}{\textsc{re}(3, J)} \right\} \\
&\leq \frac{1}{n} \|\mathbf{X}(\beta^* - \beta^*)\|_2^2 + 4\lambda \|\beta^*_{(J^*)^c}\|_1 + \frac{9}{4} \frac{\lambda^2 |J^*|}{\textsc{re}(3, J^*)} \\
&= \frac{9}{4} \frac{\lambda^2 \|\beta^*\|_0}{\textsc{re}(3, J^*)}.
\end{aligned}$$

The result then follows by combining Theorem 2.1 with Lemma 3.1. □

5 Recommended literature

Theorem 2.1 follows mainly [4] and [1]. Similar results can be found in [2] or [3].

References

[1] A. S. Dalalyan, M. Hebiri, and J. Lederer. On the prediction performance of the lasso. *Bernoulli*, 23(1):552–581, 2017.

[2] C. Giraud. *Introduction to High-dimensional Statistics*. CRC Press, 2015.

[3] P. Rigollet and J.-C. Hütter. High-dimensional statistics. Lecture notes, 2017.

[4] T. Sun and C.-H. Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 2012.