Lecture 2: Lasso

2020年8月11日 0:32

Penalized/constrained loss minimization : The linear regression case

Lasso ("least absolute shrinkage & selection operator"), Tibshirani (1996, JRSSB)

Linear model : $y = X\beta + \varepsilon$

$n\times 1 \quad n\times p \quad p\times 1$

response — predictor/covariate/design

$\mathbb{E}(\varepsilon|X) = 0$

$\mathrm{Var}(\varepsilon|X) = \sigma^2$

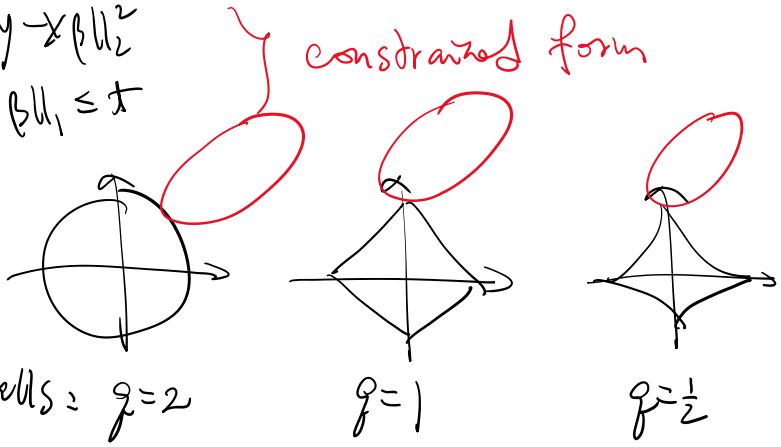
Lasso (L_1 regularization) :

$$\min_{\beta} \frac{1}{2n}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1$$ — penalized form

or

$$\min \|y - X\beta\|_2^2 \quad \text{subject to} \quad \|\beta\|_1 \le t$$ — constrained form

Why L_1 induces sparsity?



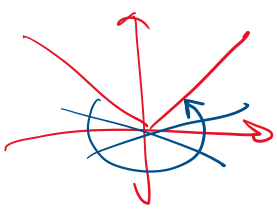
L_q balls : $q=2$ $\quad q=1$ $\quad q=\frac{1}{2}$

Prop. The Lasso estimator $\hat{\beta}^L$ satisfies

$$\left\|\frac{1}{n}X^T(y - X\hat{\beta}^L)\right\|_\infty \le \lambda.$$

Pf. The optimality condition for the Lasso problem is $0 \in \frac{1}{n}X^T(y - X\hat{\beta}^L) + \lambda\partial\|\hat{\beta}^L\|_1$, where $\partial\|\cdot\|_1 = [-1, 1]$. #

∂ : subdifferential



Related : Dantzig selector $\hat{\beta}^{DS}$ (Candes & Tao, 2007, AoS)

minimize $\|\beta\|_1$

subject to $\left\|\frac{1}{n}X^T(y - X\beta)\right\|_\infty \le \lambda$

Goal. Derive a nonasymptotic bound on the estimation/prediction error of $\hat{\beta}^L$ & $\hat{\beta}^{DS}$.

$\forall$ finite p, n

& $\hat\beta^{OLS}$. ∀ finite p, n

High-dimensional setting: $p \gg n$, but $\|\beta^*\|_0 = \#\{j : \beta^*_j \neq 0\} \triangleq s \ll n$.

p: ambient dimension; s: intrinsic dimension

Identifiability issue. When $p > n$, the Gram matrix $\Psi_n = \frac{1}{n}X^TX$ is singular, & hence β^* is not identifiable

(singular: linearly dependent)

Remedy. Assume Ψ_n is nondegenerate on certain "sparse" directions.

Restricted eigenvalue condition: $RE(s, c_0)$

$$K(s, c_0) \triangleq \min_{J \subset \{1,\dots,p\}, |J| \le s} \ \min_{\delta \neq 0, \|\delta_{J^c}\|_1 \le c_0 \|\delta_J\|_1} \frac{\|X\delta\|_2}{\sqrt{n}\|\delta_J\|_2} > 0$$

Lemma (Basic inequality) Let $S = \mathrm{supp}(\beta^*)$, $\lambda = C\sigma\sqrt{\frac{\log p}{n}}$ w/ $C > 2\sqrt{2}$. Then it holds w/ probability $\ge 1 - p^{1-c/8}$ that

$$\frac{1}{n}\|X(\hat\beta^\lambda - \beta^*)\|_2^2 + \lambda\|\hat\beta^\lambda - \beta^*\|_1 \le 4\lambda\|\hat\beta^\lambda_S - \beta^*_S\|_1 \le 4\lambda\sqrt{s}\,\|\hat\beta^\lambda_S - \beta^*_S\|_2.$$

prediction error; estimation error; can be estimated well on a sparse set

Pf. The optimality of $\hat\beta^\lambda$ implies

$$\frac{1}{2n}\|y - X\hat\beta^\lambda\|_2^2 + \lambda\|\hat\beta^\lambda\|_1 \le \frac{1}{2n}\|y - X\beta^*\|_2^2 + \lambda\|\beta^*\|_1.$$

Substituting $y = X\beta^* + \varepsilon$ gives

$$\frac{1}{2n}\|\varepsilon - X(\hat\beta^\lambda - \beta^*)\|_2^2 + \lambda\|\hat\beta^\lambda\|_1 \le \frac{1}{2n}\|\varepsilon\|_2^2 + \lambda\|\beta^*\|_1,$$

$$\|\varepsilon\|_2^2 + \|X(\hat\beta^\lambda - \beta^*)\|_2^2 - 2\varepsilon^TX(\hat\beta^\lambda - \beta^*)$$

or $$\frac{1}{2n}\|X(\hat\beta^\lambda - \beta^*)\|_2^2 \le \frac{1}{n}\varepsilon^TX(\hat\beta^\lambda - \beta^*) + \lambda\|\beta^*\|_1 - \lambda\|\hat\beta^\lambda\|_1.$$

By Gaussian concentration inequality,

$$P\left(\left\|\tfrac{1}{n}X^T\varepsilon\right\|_\infty > \tfrac{\lambda}{2}\right) \le \sum_{j=1}^p P\left(\left|\tfrac{1}{n}X_j^T\varepsilon\right| > \tfrac{\lambda}{2}\right) \le pe^{-n\lambda^2/(8\sigma^2)}.$$

$$\mathbb{P}\left(\|\tfrac{1}{n}X^T\varepsilon\|_\infty \ge \tfrac{\lambda}{2}\right) \le \sum_{j=1}^{p} \mathbb{P}\left(|\tfrac{1}{n}X_j^T\varepsilon| > \tfrac{\lambda}{2}\right) \le p e^{-n\lambda^2/(8\sigma^2)}.$$

union bound ↓ ; standardize X_j to have $\|X_j\|_2 = \sqrt{n}$

Take $\lambda = C\sigma\sqrt{\frac{\log p}{n}}$. The probability becomes

$$p e^{-\frac{C^2}{8}\log p} = p^{1-C^2/8} \to 0 \text{ since } C > 2\sqrt{2}.$$

Thus, w/ high probability the event $A = \{\|\frac{1}{n}X^T\varepsilon\|_\infty \le \frac{\lambda}{2}\}$ holds. On event A,

$$\frac{1}{n}\|X(\hat\beta^\lambda - \beta^*)\|_2^2 \le \lambda\|\hat\beta^\lambda - \beta^*\|_1 + 2\lambda(\|\beta^*\|_1 - \|\hat\beta^\lambda\|_1).$$

Adding $\lambda\|\hat\beta^\lambda - \beta^*\|_1$ to both sides yields

$$\frac{1}{n}\|X(\hat\beta^\lambda - \beta^*)\|_2^2 + \lambda\|\hat\beta^\lambda - \beta^*\|_1 \le 2\lambda(\|\hat\beta^\lambda - \beta^*\|_1 + \|\beta^*\|_1 - \|\hat\beta^\lambda\|_1)$$

$= 0$ on S^c

$$= 2\lambda(\|\hat\beta^\lambda_S - \beta^*_S\|_1 + \|\beta^*_S\|_1 - \|\hat\beta^\lambda_S\|_1) \le 4\lambda\|\hat\beta^\lambda_S - \beta^*_S\|_1. \quad \otimes$$

triangle inequality

Corollary. On event A, $\|\delta_{S^c}\|_1 \le 3\|\delta_S\|_1$, where $\delta = \hat\beta^\lambda - \beta^*$.

Pf. The basic inequality implies

$$\lambda\|\delta\|_1 \le 4\lambda\|\delta_S\|_1, \text{ or } \|\delta_{S^c}\|_1 \le 3\|\delta_S\|_1.$$

$\|\delta_S\|_1 + \|\delta_{S^c}\|_1$ — sparsity for the RE condition

Pf of Thm. By the basic inequality,

$$\frac{1}{n}\|X\delta\|_2^2 \le 4\lambda\sqrt{s}\|\delta_S\|_2.$$

On the other hand, by RE(s, 3) condition,

$$\frac{\|X\delta\|_2}{\sqrt{n}\|\delta_S\|_2} \ge \kappa > 0, \text{ or } \|\delta_S\|_2 \le \frac{\|X\delta\|_2}{\sqrt{n}\,\kappa}.$$

Combining two inequalities yields

$$\left(\frac{1}{n}\|X\delta\|_2^2\right) \le \left(4\lambda\sqrt{s}\,\frac{\|X\delta\|_2}{\sqrt{n}\,\kappa}\right), \text{ or } \frac{1}{n}\|X\delta\|_2^2 \le \frac{16\lambda^2 s}{\kappa^2}$$

$$= \frac{16C^2}{\kappa^2}\sigma^2\frac{s\log p}{n}.$$

$s\log p = o(n)$

Moreover,

$$\|\delta\|_1 \le 4\|\delta_S\|_1 \le 4\sqrt{s}\|\delta_S\|_2 \le \frac{4\sqrt{s}}{\kappa}\frac{\|X\delta\|_2}{\sqrt{n}}$$

$$\|\delta\|_1 \le 4\|\delta_S\|_1 \le 4\sqrt{s}\,\|\delta_S\|_2 \le \cdots$$

$$\le \frac{4\sqrt{s}}{\kappa}\,\frac{4C}{\kappa}\,\sigma\sqrt{\frac{s\log p}{n}} = \frac{16C}{\kappa^2}\,\sigma s\sqrt{\frac{\log p}{n}}. \qquad \#$$

$s\log p = o(n)$

$s^2\log p = o(n)$

Remark. Prediction is easier than estimation.

Question. What if our only goal is prediction? Is the Lasso estimator still prediction consistent w/o any assumption on the design matrix?

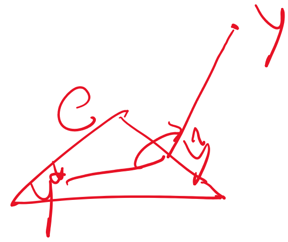
Thm. Assume $\|\beta^*\|_1 \le K$. The solution $\hat\beta$ to the Lasso problem

$$\underset{\beta}{\text{minimize}}\ \|y - X\beta\|_2^2 \quad \text{subject to } \|\beta\|_1 \le K$$

satisfies, w/ high probability, that

$$\frac{1}{n}\|X(\hat\beta - \beta^*)\|_2^2 \le CK\sigma\sqrt{\frac{\log p}{n}}.$$

Pf. By definition, $\hat y = X\hat\beta$ is the projection of y onto the compact convex set $C = \{X\beta : \|\beta\|_1 \le K\}$. Also, since $\|\beta^*\|_1 \le K$, we have $y^* = X\beta^* \in C$. Thus,



$$0 \ge (\hat y - y^*)^T(\hat y - y) = (\hat y - y^*)^T(\hat y - y^* - y + y^*)$$
$$= \|\hat y - y^*\|_2^2 - (\hat y - y^*)^T(y - y^*),$$

or

$$\frac{1}{n}\|\hat y - y^*\|_2^2 \le \frac{1}{n}(\hat y - y^*)^T(y - y^*) = \frac{1}{n}\varepsilon^T X(\hat\beta - \beta^*)$$

($\frac{1}{n}\|X(\hat\beta - \beta^*)\|_2^2$; $X(\hat\beta - \beta^*)$; ε)

$$\le \left\|\frac{1}{n}\varepsilon^T X\right\|_\infty \|\hat\beta - \beta^*\|_1 \le \frac{\lambda}{2}\cdot 2K = CK\sigma\sqrt{\frac{\log p}{n}}. \qquad \#$$

(w.h.p.; $\lambda = C\sigma\sqrt{\frac{\log p}{n}}$)

Remark. There is a rate loss in bounding $\|\hat\beta - \beta^*\|_1$.