Spring 2018: STA 6448
Advanced Probability and Inference II
Lecture 22

Yun Yang

- High-dimensional linear regression

Bounds on ℓ_2-error

Conditions:

(A1) θ^* is supported on S with $|S| = s$

(A2) X satisfies the restricted eigenvalue condition over S with parameters $(\kappa, 3)$.

Theorem

Under conditions (A1) and (A2), if $\lambda_n \geq 2\|\frac{X^T w}{n}\|_\infty$, then any Lasso solution satisfies

$$\|\widehat{\theta} - \theta^*\|_2 \leq \frac{3}{\kappa}\sqrt{s}\lambda_n, \quad \textit{and}$$

$$\|\widehat{\theta} - \theta^*\|_1 \leq 4\sqrt{s}\,\|\widehat{\theta} - \theta^*\|_2.$$

Proof outline

- Denote the objective function by

$$L(\theta;\, \lambda_n) = \frac{1}{2n}\|y - X\theta\|_2^2 + \lambda_n\|\theta\|_1.$$

- Since $\widehat{\theta}$ minimizes $L(\theta;\, \lambda_n)$, we have

$$L(\widehat{\theta};\, \lambda_n) \leq L(\theta^*;\, \lambda_n).$$

- Let $\widehat{\Delta} = \widehat{\theta} - \theta^*$. Re-arranging yields the basic inequality

$$0 \leq \frac{1}{2n}\|X\widehat{\Delta}\|_2^2 \leq \frac{w^T X\widehat{\Delta}}{n} + \lambda_n\big(\|\theta^*\|_1 - \|\widehat{\theta}\|_1\big).$$

- If $\lambda_n \geq 2\|\frac{X^T w}{n}\|_\infty$, then this leads to $\widehat{\Delta} \in \mathcal{C}_3(S)$, and

$$\kappa\,\|\widehat{\Delta}\|_2^2 \leq 3\lambda_n\sqrt{s}\|\widehat{\Delta}\|_2.$$

Restricted nullspace and eigenvalues for random designs

Theorem

Consider a random matrix $X \in \mathbb{R}^{n \times d}$, in which each row $x_i \in \mathbb{R}^d$ is drawn i.i.d. from a $\mathcal{N}(0, \Sigma)$ distribution. Then there are universal positive constants $c_1 < 1 < c_2$ such that

$$\frac{\|X\theta\|_2^2}{n} \geq c_1 \|\sqrt{\Sigma}\,\theta\|_2^2 - c_2\,\rho^2(\Sigma)\,\frac{\log d}{n}\,\|\theta\|_1^2 \quad \text{for all } \theta \in \mathbb{R}^d,$$

where $\rho^2(\Sigma)$ is the maximum diagonal entry of the covariance matrix Σ.

This result implies that an RE condition (and hence a restricted nullspace condition) holds over $\mathcal{C}_3(S)$, uniformly over all subsets S of cardinality $|S| \leq c\,\frac{\lambda_{\min}(\Sigma)}{\rho^2(\Sigma)}\,\frac{n}{\log d}$.

Bounds on prediction error

In some applications, we might be interested in finding a good predictor, meaning a vector $\theta \in \mathbb{R}^d$ such that *mean-squared prediction error* below is small,

$$\frac{\|X(\theta - \theta^*)\|_2^2}{n} = \frac{1}{n}\sum_{i=1}^{n}(\langle x_i, \theta - \theta^* \rangle)^2.$$

The problem of finding a good predictor is generally easier than estimating θ^* well in ℓ_2-norm (why?).

Bounds on prediction error

Theorem

Prediction error bounds If $\lambda_n \geq 2\|\frac{X^T w}{n}\|_\infty$*, then any Lasso solution satisfies the bound*

$$\text{(Slow rates)} \qquad \frac{\|X(\theta - \theta^*)\|_2^2}{n} \leq 12\,\|\theta^*\|_1\,\lambda_n.$$

In addition, suppose θ^* *is supported on a subset* S *and the design matrix satisfies the* $(\kappa;\,3)$*-RE condition over* S*, then*

$$\text{(Fast rates)} \qquad \frac{\|X(\theta - \theta^*)\|_2^2}{n} \leq \frac{9}{\kappa}\,|S|\,\lambda_n^2.$$

Proof: Apply the basic inequality of the Lasso program.

Variable or subset selection

- In some applications, we are interested in whether or not a Lasso estimate $\widehat{\theta}$ has non-zero entries in the same positions as the true regression vector θ^*.
- More precisely, we ask the following question:

Question

Given an optimal Lasso solution $\widehat{\theta}$, when is its support set—denoted by $S(\widehat{\theta})$—exactly equal to the true support $S(\theta^*)$?

- We refer to this property as *variable selection consistency*.
- It is possible for the ℓ_2-error $\|\widehat{\theta} - \theta^*\|_2$ to be quite small even if $\widehat{\theta}$ and θ^* have different support.
- On the other hand, given the support of θ^* can be correctly recovered, we can estimate θ^* very well.
- Therefore, variable selection is harder than estimation, which is harder than prediction.

Variable selection consistency for the Lasso

Assume the design matrix X to be deterministic.

Conditions:

(A3) Lower eigenvalue:

$$\gamma_{\min}\left(\frac{X_S^T X_S}{n}\right) \geq c_{\min} > 0.$$

(A4) Mutual incoherence: There exists some $\alpha \in [0, 1)$ such that

$$\max_{j \in S^c} \|X_j^T X_S (X_S^T X_S)^{-1}\|_1 \leq \alpha.$$

Variable selection consistency for the Lasso

Let $\Pi_{S^\perp} = I_n - X_S(X_S^T X_S)^{-1} X_S^T$ denote an orthogonal projection matrix.

Theorem

Under conditions (A3) and (A4), if $\lambda_n \geq \frac{2}{1-\alpha} \|X_{S^c}^T \Pi_{S^\perp} \frac{w}{n}\|_\infty$, then

(a) *Uniqueness: There is a unique optimal solution $\widehat{\theta}$.*

(b) *No false inclusion: This solution has its support $\widehat{S}$ contained within the true support S.*

(c) *ℓ_∞-bounds:*

$$\|\widehat{\theta}_S - \theta_S^*\|_\infty \leq \underbrace{\left\| \left(\frac{X_S^T X_S}{n} \right)^{-1} X_S^T \frac{w}{n} \right\|_\infty + \left|\!\left|\!\left| \left(\frac{X_S^T X_S}{n} \right)^{-1} \right|\!\right|\!\right|_\infty \lambda_n}_{B(\lambda_n; X)} .$$

(d) *No false exclusion: The Lasso includes all indices $j \in S$ such that $|\theta_j| > B(\lambda_n; X)$, and hence is variable selection consistent if $\min_{j \in S} |\theta_j| > B(\lambda_n; X)$.*

Variable selection consistency for the Lasso

Corollary

Suppose the noise vector w has zero-mean i.i.d. σ-sub-Gaussian entries, and X satisfies (A3) and (A4), and is C-column normalized. If for some $\delta > 0$,

$$\lambda_n \geq \frac{2C\sigma}{1-\alpha}\left\{\sqrt{\frac{2\log(d-s)}{n}} + \delta\right\},$$

then for any $\varepsilon > 0$, the optimal solution $\widehat{\theta}$ is unique with its support contained within S, and satisfies the ℓ_∞-error bound

$$\|\widehat{\theta}_S - \theta_S^*\|_\infty \leq \frac{\sigma}{c_{\min}}\left\{\sqrt{\frac{2\log(d-s)}{n}} + \varepsilon\right\} + \left\|\!\left|\left(\frac{X_S^T X_S}{n}\right)^{-1}\right|\!\right\|_\infty \lambda_n,$$

all with probability at least $1 - 2\,e^{-\frac{n\delta^2}{2}} - 2\,e^{-\frac{n\varepsilon^2}{2}}$.