36-709: Advanced Statistical Theory **Spring 2019**

Lecture 20: April 02

Lecturer: Alessandro Rinaldo *Scribes: Laixi Shi*

In section 20.1, we review oracle inequality and give the proof of the Oracle inequality for Ordinary Least Squares.

In section 20.2, we introduce the Oracle inequality for Lasso.

20.1 Oracle inequality for Ordinary Least Squares

Here are some notes for Oracle inequality

- Only consider estimator of the form $f_\theta = \sum_{j=1}^m \theta_j f_j(\cdot)$, where $(\theta_1, \cdots, \theta_m) \in \mathcal{K} \subseteq \mathbb{R}^m$
- Quantify the performance of an estimator $\hat{f} = f_{\hat{\theta}} = \sum_{j=1}^m \hat{\theta}_j f_j(\cdot)$. In Oracle inequality, we compare $\mathrm{MSE}(\hat{f})$ with $\mathrm{MSE}(f^{\mathrm{oracle}})$, where $\mathrm{MSE}(f^{\mathrm{oracle}}) = \inf\limits_{\theta \in \mathcal{K}} \mathrm{MSE}(f_\theta)$

For the Oracle inequality for Ordinary Least Squares (OLS) $\hat{f}_{\mathrm{OLS}}$,

$$\hat{f}_{\mathrm{OLS}} = \underset{\theta \in \mathbb{R}^m}{\mathrm{argmin}} f_\theta = \frac{1}{n}\sum_{i=1}^n (y_i - f_\theta(x_i))^2 = \frac{1}{n}(y - X\theta)^2$$

where $X_{ij} = f_j(x_i), f : \mathbb{R}^d \to \mathbb{R}, f = (f(x_1), \cdots, f(x_n))^\top$

Lemma 20.1 *Assume for each $i, \epsilon_i \in SG(\sigma^2)$, then for $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$MSE(\hat{f}_{OLS}) \leq \underbrace{\inf_{\theta \in \mathbb{R}^m} MSE(f_\theta)}_{MSE(f_{Oracle})} + C\sigma^2 \frac{m + \log(\frac{1}{\delta})}{n}$$

where $C > 0$ is universal constant.

Proof: Firstly, use basic inequality:

$$\frac{1}{n}(y - \hat{f}_{OLS})^2 \leq \frac{1}{n}(y - \hat{f}_{oracle})^2$$

Plug in $y = f^* + \epsilon$, we have

$$\frac{1}{n}(f^* - \hat{f}_{OLS})^2 - \frac{1}{n}(f^* - f_{oracle})^2 \leq \frac{2\epsilon^\top}{n}(\hat{f}_{OLS} - f_{oracle})$$
$$\to \frac{1}{n}(f_{oracle} - \hat{f}_{OLS})^2 \leq \frac{2\epsilon^\top}{n}(\hat{f}_{OLS} - f_{oracle})$$

Since by Trigonometric theorem: $a^2 + b^2 = c^2$, we have

$$(f^* - \hat{f}_{OLS})^2 - (f^* - f_{oracle})^2 = (\hat{f}_{OLS} - f_{oracle})^2$$

So, we obtain

$$\begin{aligned}\frac{1}{n}\left\|f_{oracle} - \hat{f}_{OLS})\right\|_2 &\leq \frac{2\epsilon^\top}{n}\frac{(\hat{f}_{OLS} - f_{oracle})}{\left\|(\hat{f}_{OLS} - f_{oracle})\right\|_2} \\ &\leq C\sigma^2\frac{m + \log(1/\delta)}{n}\end{aligned}$$

Using the prrof of MSE bound for OLS, we get the result. ■

20.2 Oracle Inequality for Lasso

In this section, we give the oracle inequality for Lasso.

Lemma 20.2 *Assume*

1) for each $i, \epsilon_i \in SG(\sigma^2)$ independently,

2) for any $S \subseteq \{1, \cdots, d\}, S \neq \emptyset, s.t \quad |S| \leq k$, The design matrix satisfy the $RE(\alpha, \kappa, S)$, which means $\frac{\|X\Delta\|^2}{n} \geq \kappa\|\Delta\|^2$, for $\Delta \in C_\alpha(S) = \{\Delta : \|\Delta_{S^c}\|_1 \leq \alpha\|\Delta_S\|_1\}$

Then, if $\lambda_n \geq \frac{2\left\|X^\top\epsilon\right\|_\infty}{n}$

$$MSE(\hat{f}_{Lasso}) \leq \inf_{\theta\in\mathbb{R}^m, s.t\|\theta\|_0\leq k}\{\frac{1+\alpha}{1-\alpha}MSE(f_\theta) + \frac{9}{2\alpha(1-\alpha)\kappa}\|\theta\|_0\lambda_n^2\}$$

where λ is any number in $(0, 1)$

So by this lemma, we can know Lasso is not a good tool to do subset selection, but is good at prediction. Lasso is as good as best subset selection under sparse linear model.

Proof: As usual, let's start with basic inequality: that holds for any $\theta \in \mathbb{R}^m$:

$$\frac{1}{2n}\left\|y - \hat{f}_\theta\right\|^2 + \lambda_n\left\|\hat{\theta}\right\|_1 \leq \frac{1}{2n}\|y - f_\theta\|^2 + \lambda_n\|\theta\|_1$$

Plug in $y = f^* + \epsilon$, we obtain

$$\frac{1}{n}\{\left\|f^* - \hat{f}_\theta\right\|^2 - \|f^* - f_\theta\|^2\} \leq \underbrace{\lambda_n\left[\|\theta\|_1 - \left\|\hat{\theta}\right\|_1\right] + \frac{\epsilon^\top}{n}(f_{\hat{\theta}} - f_\theta)}_{\mathcal{A}}$$

Recall $f_{\hat{\theta}} - f_\theta = X(\hat{\theta} - \theta) = X\Delta$.

If RHS of $\mathcal{A} \leq 0$,, the proof will be done. If not, then RHS of $\mathcal{A} > 0$, we will follow the proof of the fast rate for Lasso.

To conclude that $3\left\|\Delta_S\right\|_1 - \left\|\Delta_{S^C}\right\| > 0, S =$ support of θ, so $\Delta = \hat{\theta} - \theta \in C_3(S)$. If $0 < |S| \leq k$, then use the assumption about RE condition to upper bound of $\mathcal{A}$ by $3\lambda\sqrt{|S|} \cdot \frac{\|X\Delta\|}{\sqrt{n}\sqrt{\kappa}}$.

Next: use variation inequality:

$$ab \leq \frac{a^2}{2\beta} + \frac{\beta b^2}{2}, \forall \beta > 0, a, b \in \mathbb{R}$$

Plug in $a = \frac{3\sqrt{|S|}\lambda}{\sqrt{\kappa}}, b = \frac{\|X\Delta\|}{\sqrt{n}} = \frac{\left\|f_{\hat{\theta}} - f_\theta\right\|}{\sqrt{n}}$, and we have

$$\left\|f_{\hat{\theta}} - f_\theta\right\|^2 \leq 2\left[\left\|f^* - f_\theta\right\|^2 + \left\|f^* - f_\theta\right\|^2\right]$$

$$\mathcal{A} \leq \frac{9}{2\alpha}\frac{|S|\,\lambda^2}{\kappa} + \frac{\lambda}{n}\left[\left\|f^* - f_\theta\right\|^2 + \left\|f^* - f_\theta\right\|^2\right]$$

Rearranging and taking inf over all $\theta, s.t \left\|\theta\right\|_0 \leq k$, we get the result. ∎