36-755: Advanced Statistical Theory I **Fall 2017**

Lecture 14: October 16 - Oracle Inequality for Least Squares

Lecturer: Alessandro Rinaldo *Scribes: Manjari Das*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

This lecture's notes illustrate some uses of various LaTeX macros. Take a look at this and imitate.

A few items of note

- Lasso : How to choose λ?
  Sometimes cross validation is used. For variable selection, if variables are correlated, we end up choosing more variables by using CV.
- Assumptions needed for Lasso: Strong!
  For model selection consistency, we need min β condition.
  $min_{i \in Supp(\beta^*)} \beta_i$ is "Large enough",
- Restricted Eigen value (RE) condition is perhaps the strongest.
  An earlier stronger version is the Pailrove incoherence condition.
  In $C(k) \exists c > 0$ such that $1 \le k \le d$
  $\| \frac{X^T X}{n} - I_d \|_\infty \le \frac{1}{ck}$
  $\Rightarrow$ If $c = 32$, then the RE$(\alpha = 3, k = 0.5)$ is satisfied for all $S \subset \{1, ..., d\}$ such that $|S| \le k$.
  This is satisfied if X is populated by iid sub-Gaussians.

14.1 Oracle inequalities

Model need not be correct!

Assume $Y = f(x) + \epsilon$, $\epsilon \sim SG(\sigma^2)$, f arbitrary function.
Observe $(Y_1, x_1), ..., (Y_n, x_n)$. $Y's$ are independent. $(x_1, ..., x_n)$ deterministic.
We do not assume $Y = x^T \beta + \epsilon$.

Suppose we have a dictionary of functions from $\mathbb{R}^d$ into $\mathbb{R}$.

$$\mathcal{D} = \{f_1, ..., f_M\}$$

and we are going to estimate f with a linear combination of functions in $\mathcal{D}$.

Of course if $f_j(x) = x_j$ for $j = 1, ..., M$.

Then for any vector $\theta \in \mathbb{R}^M$, $\sum_j \theta_j f_j(x) = \theta^T x$.

One possible estimator is $\hat{\theta}_{OLS}$ which minimizes

$$\frac{1}{n}\sum(Y_i - \sum_j \theta_j f_j(x_i))^2$$

and estimator of f is $f_{\hat{\theta}_{OLS}} = \sum_{j=1}^{M} \hat{\theta}_j f_j$.

To evaluate the performance of an estimator $\hat{f}$ we consider its risk
$R_f(\hat{f}) = E\left[\frac{1}{n}\sum_{i=1}^{n}\left(\hat{f}(x_i) - f(x_i)\right)^2\right]$
$= E\left[\frac{1}{n}\|\hat{f} - f\|^2\right]$, $\hat{f} = \left(\hat{f}(x_1), ..., \hat{f}(x_n)\right)^T$.

Let $K \subset \mathbb{R}^M$.

Definition 1 *The Oracle solution wrt risk R_f, $\mathcal{D}$ and K is the f_{θ^*} where $\theta^* \in K$*

$$R_f(f_{\theta^*}) \leq R_f(f_\theta) \forall \theta \in K.$$

f_{θ^*} is not necessarily a good estimator of f!!
An estimator $f_{\hat{\theta}}$, $\hat{\theta} \in K$ and depend on data satisfies Oracle inequality if

$$R_f(\hat{f}) \leq cR_f(f_{\theta^*}) + \phi(n, \mathcal{D}, f, K)$$

where $cge1$ and $\phi(n, \lceil, f, K) \longrightarrow 0$ as $n \longrightarrow \infty$.

An estimator is good when it satisfies an Oracle inequality with small c and vanishing ϕ.
[If $c = 1$, this is a sharp inequality.]

Equivalently,
$P_f\left(MSE(\hat{f}) \leq cMSE(f_{\theta^*}) + \phi(n, \mathcal{D}, f, K, \delta)\right) \geq 1 - \delta$
$\delta \in (0, 1)$. $MSE(\hat{f}) = \frac{1}{n}\|\hat{f} - f\|^2$.

Theorem 14.1 *(Oracle inequality for Least squares) Assume $(\epsilon_1, ..., \epsilon_n) \sim_{i.i.d} SG(\sigma^2)$. Then*

$P\left(MSE(f_{\hat{\theta}_{OLS}}) \leq inf_{\theta \in \mathbb{R}^M} MSE(f_\theta) + c\sigma^2 \frac{M}{n} log\left(\frac{1}{\delta}\right)\right) \geq 1 - \delta,$
$\delta \in (0, 1)$.

Proof: $Y = [Y_1, ..., Y_n]^T$, $f_\theta = [f_\theta(x_1), ..., f_\theta(x_n)]^T$ in $\mathbb{R}^n$.

Least squares is $argmin_{\theta \in \mathbb{R}^M} \frac{1}{n}\|Y - f_\theta\|^2$, $f_\theta = \sum \theta_j f_j$.
So, $\|Y - f_{\hat{\theta}_{OLS}}\|^2 \leq \|Y - f_{\theta^*}\|^2$, $Y = f + \epsilon$.
Thus

$$\frac{1}{n}\|f - \hat{f}_{th\hat{e}ta_{OLS}}\|^2 - \frac{1}{n}\|Y - f_{\theta^*}\|^2 \leq \frac{2}{n}\epsilon^T(\hat{f}_{th\hat{e}ta_{OLS}} - f_{\theta^*}).$$

LHS is $\frac{1}{n}\|f_{\hat{\theta}_{OLS}} - f_{\theta^*}\|^2 \geq 0$.
f_{θ^*} is projection of f onto $span\{f_1, ..., f_M\}$.

But $f_{\hat{\theta}} - f_{\theta^*} = \phi(\hat{\theta} - \theta^*)$, where $\phi_{n \times M}$ such that $\phi_{ij} = f_j(x_i)$.
Same proof used to derive consistency of OLS in Linear Regression model gives that

$$\frac{1}{n}\epsilon^T(f_{\hat{\theta}_{OLS}} - f_{\theta^*}) \in c\sigma^2 \frac{M}{n} log\left(\frac{1}{\delta}\right)$$

with probability $\geq 1 - \delta$. ■

Approximation error: $R_f(f_{\theta^*})$ or $MSE(f_{\theta^*})$ can only be made small with assumptions on f.

14.2 Oracle inequality for Lasso

Theorem 14.2 *Assume* $(\epsilon_1, ..., \epsilon_n) \sim_{i.i.d} SG(\sigma^2)$ *and that* $RE(3, K)$ *assumption holds for all* $S = \{1, ..., M\}$ *with* $|s| \leq K << n$.
Then if $\lambda_n \geq 2\frac{\|\Phi^T\epsilon\|_\infty}{n}$, *we have*

$$MSE(f_{\hat{\theta}}) \leq inf_{\theta \in \mathbb{R}^M, \|\theta\|_0 \leq K}\left\{\frac{1+\alpha}{1-\alpha}MSE(f_\theta) + \frac{9}{2\alpha(1-\alpha}\kappa\|\theta\|_0\lambda_n^2\right\} \forall \alpha \in (0, 1).$$

Fix α, *then*

$$MSE(f_{\hat{\theta}}) \leq cMSE(f_{\theta^*}) + k\frac{log\alpha}{n}$$

Proof: We begin with

$$\frac{1}{2n}\|Y - f_{\hat{\theta}}\|^2 + \lambda_n\|\hat{\theta}\|_1 \leq \frac{1}{2n}\|Y - f_\theta\|^2 + \lambda_n\|\theta\|_1 \forall \theta \in \mathbb{R}^M$$

Then we replace Y by $f + \epsilon$ to get

$$\frac{1}{n}\|f - f_{\hat{\theta}}\|^2 - \frac{1}{n}\|f - f_\theta\|^2 \leq 2\lambda_n\left(\|\theta\|_1 - \|\hat{\theta}\|_1\right) + 2\frac{\epsilon^T}{n}\left(f_{\hat{\theta}} - f_\theta\right) \forall \theta \in \mathbb{R}^M$$

Think of $\phi(\hat{\theta} - \theta)$ as $\lambda\hat{\Delta}$ (the proof for Lasso's fast rate).

Let $S = Supp(\theta)$ and assume $|S| \leq k$.
Then we have 2 cases

- LHS of (*) is negative
  $MSE(f_{\hat{\theta}}) \leq MSE(f_\theta)$. Nothing to show.
- If LHS of (*) is positive, then
  $MSE(f_{\hat{\theta}}) - MSE(f_\theta) \leq 2\lambda_n\|\hat{\theta} - \theta\|_1 + 2\lambda_n(\|\theta\| - \|\hat{\theta}\|_1)$

$$\left[\because \frac{\epsilon^T\phi(\hat{\theta} - \theta)}{n} \leq \frac{\|\phi^T\epsilon\|_\infty}{n}\|\hat{\theta} - \theta\|_1\right]$$

Using same proof as for the fast rates for Lasso we get
$\leq \lambda_n(3\|\hat{\Delta}_S\|_1 - \|\Delta_{S^c}\|_1)$

$\leq 3\lambda_n\sqrt{|S|}\frac{\|f_{\hat{\theta}} - f_\theta\|}{\sqrt{n}}\frac{1}{\sqrt{\kappa}}$
We use the variational inequality

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

Use the inequality with $a = \dfrac{3\lambda_n\sqrt{|S|}}{\sqrt{K}}$ and $b = \dfrac{\|f_{\hat{\theta}} - f_\theta\|}{\sqrt{n}}$ and $\alpha \in (0, 1)$.

$\leq \dfrac{1}{2\alpha}\dfrac{|S|\lambda_n^2 9}{\kappa} + \dfrac{\alpha}{2}\dfrac{\|\hat{f}_{\hat{\theta}} - f_\theta\|^2}{n}.$

Next, $\|f_{\hat{\theta}} - f_\theta\|^2 \leq 2\|f - f_{\hat{\theta}}\|^2 m + 2\|f - f_\theta\|^2$

Hence we get

$$MSE(f_{\hat{\theta}}) - MSE(f_\theta) \leq \frac{9}{2\alpha} + \alpha\left[\frac{\|\hat{f}_{\hat{\theta}} - f_\theta\|^2}{n} + \frac{\|f - f_\theta\|^2}{n}\right]$$

or, $MSE(f_{\hat{\theta}})(1-\alpha) \leq (1+\alpha)MSE(f_\theta) + \dfrac{9|S|\lambda_n^2}{2\alpha n}$

This concludes the proof. ■