36-710: Advanced Statistical Theory **Fall 2018**

Lecture 1: October 24

Lecturer: Alessandro Rinaldo *Scribes: Shamindra Shrotriya*

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.

1.1 Oracle Inequalities

Here we do not assume a linear model, just:

$$Y = f^*(x) + \epsilon$$

Where $f^* : \mathbb{R}^d \to \mathbb{R}$ and $\epsilon \sim (0, \sigma^2)$

We observe n pairs $\{(Y_i, x_i)\}_{i=1}^n$ where $(x_1, \ldots, x_n)$ are **fixed** in $\mathbb{R}^d$. Suppose that we have a dictionary:

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

of M functions $f_j : \mathbb{R}^d \to \mathbb{R}$.

And suppose further that we want to estimate f^* using a linear combination of functions in $\mathcal{D}$.

$$\sum_{j=1}^{M} \theta_j f_j(.) \text{ for } (\theta_1, \ldots, \theta_M) \in \mathbb{R}^M$$

.

Remark. In this approach we note the following:

1. We can recover the linear case by setting $M = d$ and $f_j(x) = x_j$ where x_j is the jth coordinate of $x \in \mathbb{R}^d$. Then we have that
$$x \mapsto \sum_{j=1}^{M} f_j(x) = \theta^T x$$
2. We may want to restrict the coefficient $(\theta_1, \ldots, \theta_M) \in K \subseteq \mathbb{R}^M$

For any $f : \mathbb{R}^d \to \mathbb{R}$ let

$$\begin{aligned}\text{MSE}(f) &= \frac{1}{n} \sum_{j=1}^{M} (f(x_i) - f^*(x_i))^2 \\ &= \mathbb{E}_n \|f - f^*\|_2^2\end{aligned}$$

Where E_n is the expectation with respect to the empirical measure corresponding to $(x_1, \ldots, x_n)$. If $\hat{f}$ is an estimator then the $\text{MSE}(\hat{f})$ is random.

Definition. The Oracle approximation to f^* with respect to K is the function:

$$f_{\theta^{\mathrm{OR}}} = \sum_{j=1}^{M} \theta_J^{\mathrm{OR}} f_j \tag{1.1}$$

$$\text{s.t. } \mathrm{MSE}(f_{\theta^{\mathrm{OR}}}) = \inf_{\theta \in K} \mathrm{MSE}(f_\theta) \tag{1.2}$$

Note that $f_\theta = \sum_{j=1}^{M} \theta_j f_j$ and $\mathrm{MSE}(f_\theta) = \frac{1}{n} \sum_{j=1}^{M} (f_\theta(x_i) - f^*(x_i))^2$.

We further note that $f_{\theta^{\mathrm{OR}}}$ need not be unique and that $f_{\theta^{\mathrm{OR}}}$ may be a terrible approximation of f^*.

We would like to do as well as as the Oracle (who has access to f^* to compute $\min_{\theta \in K} MSE(f_\theta)$. An estimator $\hat{f}$ of f^* satisfies an Oracle inequality with respect to $\mathcal{D}, K$ and the choice of the loss function if:

$$\mathbb{E}\Big(\mathrm{MSE}(\hat{f})\Big) \leq C \inf_{\theta \in K} \mathrm{MSE}(f_\theta) + \underbrace{\phi(n, \mathcal{D}, K, f^*)}_{\text{random fluctuations}} \tag{1.3}$$

Where $C > 0$ and $\phi_n > 0$ and hopefully $\phi_n \to 0$ as $n \to \infty$. Typically $C \geq 1$ and if $C = 1$ this Oracle inequality is sharp.

Alternatively we could get a high probability bound:

$$\mathbb{P}\Bigg(\mathrm{MSE}(\hat{f}) \geq C \inf_{\theta \in K} \mathrm{MSE}(f_{\theta^{\mathrm{OR}}}) + \phi(n, \mathcal{D}, K, f^*, \delta)\Bigg) \leq \delta \text{ small} \tag{1.4}$$

1.2 Oracle Inequality for Least Squares

Theorem (Oracle Inequality for Least Squares)**.** *Let $K = \mathbb{R}^n$ and assume $(\epsilon_1, \ldots, \epsilon_n) \in SG(\sigma^2)$. Then with probability $\geq 1 - \delta$, $\delta \in (0,1)$ small we have:*

$$MSE\left(\hat{f}^{OLS}\right) \leq \inf_{\theta \in \mathbb{R}^M} MSE(f_\theta) + C\left(\sigma^2 \frac{M}{n} + \log\left(\frac{1}{\delta}\right)\right) \tag{1.5}$$

Where $f_J(x_i) := X_{ij} \quad \forall i \in \{1, \ldots, n\}, j \in \{1, \ldots, M\}$. We also have

$$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} \in \mathbb{R}^n$$

$$\text{and} f_j = \begin{bmatrix} f_j(x_1) \\ \vdots \\ f_j(x_n) \end{bmatrix} \in \mathbb{R}^n$$

We have

$$\hat{\theta}^{\mathrm{OLS}} = \arg\min_{\theta \in \mathbb{R}^M} \|Y - X\theta\|_2^2$$

Proof. We start with the basic inequality:

$$\frac{1}{n}\|Y - X\hat{\theta}^{\text{OLS}}\|_2^2 \leq \frac{1}{n}\|Y - X\hat{\theta}^{\text{OR}}\|_2^2$$

Note that $X\hat{\theta}^{\text{OR}}$ is the orthogonal projection of $Y^* = f^*$ onto $\text{span}\{f_1, \dots, f_M\}$. Next we write $Y = f^* + \epsilon$ where $\epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$. We then plug this back into the basic inequality to obtain

$$\frac{1}{n}\left[\|Y^* - X\hat{\theta}^{\text{OLS}}\|_2^2 - \frac{1}{n}\|Y - X\hat{\theta}^{\text{OR}}\|_2^2\right] \leq 2\epsilon^T(X\hat{\theta}^{\text{OLS}} - X\hat{\theta}^{\text{OR}})$$

Since $f^* - f^{\text{OR}}$ is orthogonal to $\text{span}\{f_1, \dots, f_M\}$. It is orthogonal to $\hat{f}^{\text{OLS}}$ and $\hat{f}^{\text{OR}}$. We then use the Pythagorean theorem to conclude that:

$$\begin{aligned}
\|f^* - \hat{f}^{\text{OLS}}\|_2^2 - \|f^* - f^{\text{OR}}\|_2^2 &= \|\hat{f}^{\text{OLS}} - \hat{f}^{\text{OR}}\|_2^2 \\
\implies \frac{1}{n}\|\hat{f}^{\text{OLS}} - \hat{f}^{\text{OR}}\|_2^2 &\leq \frac{2}{n}\epsilon^T(\hat{f}^{\text{OLS}} - \hat{f}^{\text{OR}}) \\
\implies \frac{1}{n}\|X\hat{\theta}^{\text{OLS}} - X\hat{\theta}^{\text{OR}}\|_2^2 &\leq C\left[\sigma^2\frac{M}{n} + \log\left(\frac{1}{\delta}\right)\right]
\end{aligned}$$

The final line follows since:

- $\hat{f}^{\text{OLS}} = X\hat{\theta}^{\text{OLS}}$
- $\hat{f}^{\text{OR}} = X\hat{\theta}^{\text{OR}}$
- $\frac{1}{n}\|X\hat{\theta}^{\text{OLS}} - X\hat{\theta}^{\text{OR}}\|_2^2 \leq C\left[\sigma^2\frac{M}{n} + \log\left(\frac{1}{\delta}\right)\right]$ by the last least squares proof

□

Remark. $\frac{1}{n}\|\hat{f}^{\text{OR}} - f^*\|_2^2$ is the approximation error. If we do not have information about f^* this approximation error is unavoidable and may be very large. It is non-stochastic given $\mathcal{D}$ and K.

1.3 Sparse Oracle Inequality for the LASSO

Theorem (Sparse Oracle Inequality for the LASSO)**.** *Assume that for all subsets $S \subseteq \{1, \dots, m\}$ with $|S| \leq s$ and that the RE$(3, k)$ holds for $X = (f_j(x_i)))\, \forall i \in \{1, \dots, n\}, J \in \{1, \dots, M\}$. Then for $\lambda_n \geq \frac{2\|\epsilon^T X\|_\infty}{n}$ and $\forall \alpha \in (0, 1)$. We have that:*

$$MSE(f_{\hat{\theta}^{LASSO}}) \leq \inf_{\substack{\theta \in \mathbb{R}^M \\ \|\theta\|_0 \leq s}} \left\{\frac{1+\alpha}{1-\alpha}MSE(f_\theta) + 9\left(\frac{1}{2\alpha(1-\alpha)}\frac{S}{k}\lambda_n^2\right)\right\}$$