**36-755: Advanced Statistical Theory 1** **Fall 2016**

# Lecture 24: November 21

*Lecturer: Alessandro Rinaldo* *Scribes: Adarsh Prasad*

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



**(Recap)Main Result for Oracle Inequality for Nonparametric Least Squares**

**Theorem 24.1** *Assume $\partial\mathcal{F} = \{\mathcal{F} - \mathcal{F}\}$ to be star-shaped and let $\delta_n$ be any solution of the critical inequality:*

$$\frac{\mathcal{G}(\delta, \partial\mathcal{F})}{\delta} \leq \frac{\delta}{2\sigma}$$

*Then $\exists c_0, c_1, c_2 > 0$, such that for any $t \geq \delta_n$ and for all $f \in \mathcal{F}$ we have:*

$$\left\| \widehat{f}_n - f^* \right\|_n^2 \leq \inf_{\gamma \in (0,1)} \left\{ \frac{1+\gamma}{1-\gamma} \left\| f - f^* \right\|_n^2 + \frac{c_0 t \delta_n}{\gamma(1-\gamma)} \right\} \tag{24.1}$$

$$w.p. \geq 1 - c_1 exp\left( \frac{-c_2 n t \delta_n}{\sigma^2} \right)$$

**Proof:** *Proof can be found in Lecture 23.* ■

**Remarks** Note that Theorem 24.1 gives a family of bounds and setting $t = \delta_n$ in Theorem 24.1, yields an upper bound of the form:

$$\left\| \widehat{f} - f^* \right\|_n^2 \precsim \inf_{f \in \mathcal{F}} \left\| f - f^* \right\|_n^2 + \delta_n^2 \tag{24.2}$$

## 24.1 Uses of Oracle Inequality for Nonparametric least squares

### 24.1.1 Orthogonal series expansion

Let $P$ be a distribution on $\mathscr{X}$ and let $\{\phi_m\}_{m=1}^{\infty}$ be an orthonormal basis for $L^2(P)$ i.e. $\int \phi_m^2(x)dP = 1$ and $\int \phi_m(x)\phi_{m'}(x)dP = 0$. For all integers $T = 1, 2 \ldots$, consider the the function class:

$$\mathcal{F}(1,T) = \left\{ f \in L^2(P) \middle| \, f = \sum_{m=1}^{T} \theta_m \phi_m, \quad \sum_{m=1}^{T} \theta_m^2 \leq 1 \right\}$$

Then $f_{\widehat{\theta}}$ be the constrained least-squares estimate over this class which can be computed by solving the following version of ridge regression:

$$\widehat{\theta} \in \underset{\theta in \mathbb{R}^T}{\operatorname{argmin}} \frac{1}{2} \left\| Y - X\theta \right\|_2^2 + \lambda_n \left\| \theta \right\|_2^2 \tag{24.3}$$

$$where \;\; [X_{n \times T}]_{ij} = \phi_j(x_i)$$

Let $f^*$ be the true function which lies in the unit ball in $L^2(P)$. Since $\{\phi_m\}_{m=1}^{\infty}$ is an orthonormal basis for $L^2(P)$, we have $f^* = \sum_m \phi_m \theta_m^*$ such that from Parseval's theorem $\|f^*\|_2^2 = \sum (\theta_m^*)^2 \leq 1$.
Then,

$$\inf_{f \in \mathcal{F}(1,T)} \|f - f^*\|_{L^2(P)}^2 = \sum_{m=T+1}^{\infty} (\theta_m^*)^2 \text{ for each } T = 1, 2, \ldots$$

and the infimum is achieved by the truncated function $\tilde{f} = \sum_{m=1}^{T} \theta_m^* \phi_m$.

- For this problem(Equation 24.3), the critical radius $\delta_n \precsim \frac{\sigma^2 T}{n}$ (HW!!).
- Set $f = \tilde{f}$ in the oracle inequality in Equation 24.2, we get:

$$\left\|f_{\widehat{\theta}} - f^*\right\|_n^2 \precsim \underbrace{\sum_{m=T+1}^{\infty} (\theta_m^*)^2}_{\text{Approximation Error}} + \underbrace{\frac{\sigma^2 T}{n}}_{\text{Estimation Error}} \tag{24.4}$$

  As $n \to \infty$, we can let $T = T(n) \to \infty$, and we choose the optimal $T$ by balancing the approximation and estimation terms.

- In many cases the coefficients $\theta_m^*$ exhibit polynomial decay such that:

$$\sum_{m=T+1}^{\infty} \theta_m^* \precsim \frac{c}{T^{2\alpha}} \quad \alpha \geq 1, \alpha \in \mathbb{N}$$

  This is the case if $f^*$ is $\alpha-$times differentiable and it's $\alpha-$order derivative is square integrable. In this case, we can obtain the optimal $T$, balancing both terms in Equation 24.4,

$$\frac{c}{T^{2\alpha}} = \frac{\sigma^2 T}{n} \implies \left(\frac{cn}{\sigma^2}\right)^{\frac{1}{2\alpha+1}}$$

  Using this, we get the final rate as $\left(\frac{\sigma^2 T}{n}\right)^{\frac{2\alpha}{2\alpha+1}}$.

### 24.1.2 Best Sparse Approximation.

Consider the standard linear model $y_i = f_{\theta^*}(x_i) + \sigma w_i$, where $f_{\theta^*}(x) := \langle \theta^*, x \rangle$ is an unknown linear regression function, and $w_i \sim \mathcal{N}(0, 1)$ is an i.i.d. noise sequence. For a fixed sparsity index $s \in \{1, 2, \ldots, d\}$, consider the class of all linear regression functions based on $s$-sparse vectors, the class:

$$\mathcal{F}_{\text{spar}}(s) := \left\{ f_\theta | \theta \in \mathbb{R}^d, \|\theta\|_0 \leq s \right\}$$

Consider the estimator $\widehat{\theta}$ corresponding to performing least-squares over the set of all regression vectors with at most s non-zero coefficients:

$$f_{\widehat{\theta}} \in \operatorname*{argmin}_{f_\theta \in \mathcal{F}_{\text{spar}}(s)} \frac{1}{2n} \|Y - f_\theta\|_2^2 \tag{24.5}$$

Using Equation 24.2 for this problem, we get:

$$\left\|f_{\widehat{\theta}} - f^*\right\|_n^2 \precsim \inf_{f \in \mathcal{F}_{\text{spar}}(s)} \|f - f^*\|_n^2 + \underbrace{\sigma^2 \frac{s \log(ed/s)}{n}}_{\delta_n^2} \tag{24.6}$$

where we devote the rest of section to proving that $\delta_n^2 = \sigma^2 \frac{s \log(ed/s)}{n}$.

- Firstly note that $\partial\mathcal{F}_{\text{spar}}(s) = \mathcal{F}_{\text{spar}}(s) - \mathcal{F}_{\text{spar}}(s) \subset \mathcal{F}_{\text{spar}}(2s)$. Therefore, we have $\mathscr{G}_n(\delta, \partial\mathcal{F}_{\text{spar}}(s)) \leq \mathscr{G}_n(\delta, \partial\mathcal{F}_{\text{spar}}(2s))$.
- Now, let $S \subseteq \{1, 2, \dots, d\}$ with $|S| = 2s \leq d$. Let $X_{n,d}$ with $i^{th}$ row given by $x_i^T$. And let $X_S \in \mathbb{R}^{n\times 2s}$ be the sub-matrix with columns indexed by $S$. We can then write:
- Then,

$$\mathscr{G}_n(\delta, \partial\mathcal{F}_{\text{spar}}(2s)) = E_w\left[\max_{|S|=2s} Z_n(S)\right] \quad \textit{where} \quad Z_n(S) = \sup_{\substack{\theta_S\in\mathbb{R}^{2s} \\ \|X_S\theta_S\|_2 \leq \delta\sqrt{n}}} \left|\frac{w^T X_S \theta_S}{n}\right|$$

  - Now, observe that for a fixed $S$, $Z_n(S)$ is a Lipschitz function of $w_1, \dots, w_n$ with Lipschitz constant $\frac{\delta}{\sqrt{n}}$. So, by concentration of Lipschitz function for gaussian r.v.'s, we have that:

$$P(Z_n(S)) \geq E\left[Z_n(S)\right] + t\delta \leq e^{\frac{-nt^2}{2}} \ \ \forall t > 0. \tag{24.7}$$

  - So, we just need to bound $E\left[Z_n(S)\right]$. To do this, let $X_S = UDV^T$ be the SVD decomposition of $X_S$, where $U \in \mathbb{R}^{n\times 2s}$ and $V \in \mathbb{R}^{d\times 2s}$ are the left and right singular matrices, and $D \in \mathbb{R}^{2s\times 2s}$ is a diagonal matrix of singular values.
  - Noting that $\frac{\|X_S\theta_S\|_2}{\sqrt{n}} = \frac{\|DV^T\theta_S\|_2}{\sqrt{n}}$, let $\beta = \frac{DV^T\theta_S}{\sqrt{n}}$, we get:

$$E\left[Z_n(S)\right] \leq E\left[\sup_{\substack{\beta\in\mathbb{R}^{2s} \\ \|\beta\|_2\leq\delta}} \left|\frac{1}{\sqrt{n}}\langle U^T w, \beta\rangle\right|\right] \leq \frac{\delta}{\sqrt{n}} E\left[\left\|U^T w\right\|_2\right]$$

  - Now, by observing that $U^T w \sim \mathcal{N}(0, I_{2s})$, and using Jensen's inequality, we get that $\left\|U^T w\right\|_2 \leq \sqrt{2s}$. Therefore $E\left[Z_n(S)\right] \leq \frac{\delta\sqrt{2s}}{\sqrt{n}}$.
  - Now combining, the upper bound on expectation with the tail bounds of Equation 24.7, along with a union bound over all subsets of size $2s$, we get:

$$P\left[\max_{|S|=2s} Z_n(S) \geq \delta\left(\sqrt{\frac{2s}{n}} + t\right)\right] \leq \binom{d}{2s} e^{-\frac{nt^2}{2}}, \quad \text{valid for all t} \geq 0 \tag{24.8}$$

- By integrating this tail bound, we get:

$$\frac{E_w\left[\max_{|S|=2s} Z_n(S)\right]}{\delta} = \frac{\mathscr{G}_n(\delta, \partial\mathcal{F}_{\text{spar}}(2s))}{\delta} \precsim \sqrt{\frac{s}{n}} + \sqrt{\frac{\log\left(\binom{d}{2s}\right)}{n}} \precsim \sqrt{\frac{s\log(ed/s)}{n}}. \tag{24.9}$$

- Hence, from Equation 24.9, we get that the critical inequality is satisfied for $\delta_n^2 \simeq \sigma^2\frac{s\log(ed/s)}{n}$

## 24.2 Introduction to U-statistics

U-statistics were invented by Hoeffding in 1948, although not in a high dimensional setting. Let $\mathscr{P}$ be a family of distributions on $(\mathscr{X}, \mathscr{B})$, and let $\theta : \mathscr{P} \mapsto \mathbb{R}$. If $P \in \mathscr{P}$, then $\theta(P)$ is it's parameter.
A parameter $\theta$ is estimable based on $m$ iid realizations $X_1, X_2, \dots, X_m \sim P$ if there exists a kernel function $h : \mathcal{X}^M \mapsto \mathbb{R}$ such that $\theta(P) = E\left[h(X_1, X_2, \dots, X_m)\right]$. The smallest such $m$ is called the degree of the parameter $\theta$.

**Symmetry.** WLOG we may take $h$ to be symmetric in its arguments. If $h$ is not symmetric, we may symmetrize by considering the following function $\tilde{h}$ obtained by averaging over all permutations of the input:

$$\tilde{h}(X_1, \dots, X_m) = \frac{1}{m!} \sum_{\sigma \in S_m} h(X_{\sigma_1}, X_{\sigma_2}, \dots, X_{\sigma_m}) \tag{24.10}$$

and $E[\tilde{h}(X_1, \dots, X_m)] = E[h(X_1, X_2, \dots, X_m)]$.

**Estimation.** Suppose we obtain $n > m$ samples $X_1, X_2, \dots, X_n \overset{iid}{\sim} P$. For a symmetric kernel function $h$ which estimates $\theta(P)$ unbiasedly, the corresponding U-statistic estimator is given by:

$$U_n = U_n(h) = \binom{n}{m}^{-1} \sum_{i_1 < \dots < i_m} h(X_{i_1}, \dots, X_{i_m}) \tag{24.11}$$

where $m$ is the order/degree of the parameter $\theta(P)$. Clearly $E[U_n] = \theta(P)$, hence $U_n$ is an unbiased estimator of the parameter $\theta(P)$.

**Motivation and Intuition.** Let $S_n = S(X_1, \dots, X_n)$ be an unbiased estimator of $\theta(P)$. Can we somehow reduce its variance?
Let $U_n$ be the corresponding U-statistic i.e.:

$$U_n = \frac{1}{n!} \sum_{\sigma \in S_n} S(X_{\sigma_1}, \dots, X_{\sigma_n}) \tag{24.12}$$

Then, $U_n = E[S_n | X_{(1)}, \dots, X_{(n)}]$, assuming, $\theta = 0$, then the variance of $U_n$ is given by:

$$\begin{aligned} E[U_n^2] &= E\left[E[S_n | X_{(1)}, \dots, X_{(n)}]^2\right] \\ &\leq E\left[E[S_n^2 | X_{(1)}, \dots, X_{(n)}]^2\right] \ (Using Jensen's) \\ &\leq E[S_n^2] \end{aligned}$$

Hence, we obtained an unbiased estimator whose variance is smaller than the initial variance.

# References

[1] M. Wainwright, High-dimensional statistics: A non-asymptotic viewpoint