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

Lecture 21: November 14

Lecturer: Alessandro Rinaldo *Scribes: Enxu Yan*

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

21.1 Non-parameteric Least-Squares

21.1.1 Recap

We assume

$$y_i = f^*(x_i) + \epsilon_i,\ i = 1...n$$

where $\epsilon_i = \sigma w_i$ for some $\sigma > 0$ and $w_1, w_2, ..., w_n$ are i.i.d. $N(0, 1)$, with fixed design $x_1, ..., x_n \in \mathcal{X} \subseteq \mathbb{R}^d$. For the current analysis we assume f^* belongs to the function class $\mathcal{F}$ considered in the least-square problem. Let

$$\hat{f} \in \underset{f \in \mathcal{F}}{argmin}\ \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2$$

be our estimated function. The goal is to relate the excess risk

$$\|\hat{f} - f^*\|_n := \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\hat{f}(x_i) - f^*(x_i))^2}$$

to the *local Gaussian Complexity* of $\mathcal{F}$ at scale $\delta > 0$:

$$\mathcal{G}_n(\mathcal{F}, \delta) := E_w \left[\sup_{f \in \mathcal{F},\ \|f\|_n \le \delta} \frac{\sigma}{n} \left| \sum_{i=1}^{n} w_i f(x_i) \right| \right].$$

21.1.2 Bound via Critical Radius

A central object of this analysis is δ that satisfies the *critical inequality*

$$\frac{\mathcal{G}_n(\mathcal{F}, \delta)}{\delta} \le \frac{\delta}{2\sigma}, \tag{21.1}$$

and the *critical radius* δ_n that satisfies the above inequality with equality, which must exist for any *star-shaped* function class $\mathcal{F}$. Recall that we say a function class $\mathcal{F}$ is *star-shaped* if

$$f \in \mathcal{F} \ \Rightarrow\ \alpha f \in \mathcal{F}$$

for any $\alpha \in [0, 1]$. Define the shifted function class $F^* = \{f - f^* | f \in \mathcal{F}\}$. We have the following theorem.

Theorem 21.1 *If $\mathcal{F}^*$ is Star-Shaped, then for any δ satisfies the* critical inequality (21.1) *and $t \geq \delta$, the nonparametric least-square estimate $\hat{f}_n$ satisfies*

$$P(\ \|\hat{f} - f^*\|_n^2 \geq 16t\delta_n\ ) \leq \exp\{-\frac{nt\delta_n}{2\sigma^2}\}, \tag{21.2}$$

which implies

$$\|\hat{f} - f^*\|^2 \lesssim \delta_n^2$$

both in expectation and with high probability.

Proof: Recall that $\frac{\mathcal{G}_n(u,\mathcal{H})}{u}$ is non-increasing if $\mathcal{H}$ is star-shaped. Now start with the basic inequality

$$\frac{1}{2}\|\hat{\Delta}_n\|_n^2 \leq \frac{\sigma}{n}\sum_{i=1}^n w_i\hat{\Delta}(x_i)$$

where $\hat{\Delta}_n := \hat{f} - f^*$. Define the bad event as

$$A(u) := \left\{\exists g \in \mathcal{H},\ \|g\|_n \geq u \ \middle|\ \Big|\frac{\sigma}{n}\sum_{i=1}^n w_i g(x_i)\Big| \geq 2u\|g\|_n\right\}$$

In Lemma 21.2, we show that for $u \geq \delta_n$,

$$P(A(u)) \leq \exp\{-\frac{nu^2}{2\sigma^2}\}.$$

Now using the lemma with $\mathcal{H} = \mathcal{F}^*$ and $u = \sqrt{t\delta_n}$ for some $t \geq \delta_n$. With probability no less than $1 - \exp\{-\frac{nt\delta_n}{2\sigma^2}\}$, we have

$$\forall g \in \mathcal{F}^* \cap \{g : \|g\|_n \geq u\},\ \ \frac{\sigma}{n}\left|\sum_{i=1}^n w_i g(x_i)\right| \leq 2\|g\|_n u.$$

Therefore, consider two cases:

Case 1: $\|\hat{\Delta}_n\|_n < \sqrt{t\delta_n}$. We obtain $\|\hat{\Delta}_n\|_n^2 \leq t\delta_n$ trivially.

Case 2: $\|\hat{\Delta}_n\|_n \geq \sqrt{t\delta_n}$.

Since $\hat{\Delta}_n \in \mathcal{F}^*$ and $\|\hat{\Delta}_n\|_n \geq \sqrt{t\delta_n}$, we have

$$\frac{1}{2}\|\hat{\Delta}_n\|^2 \ \leq\ \frac{\sigma}{n}\left|\sum_{i=1}^n w_i\hat{\Delta}(x_i)\right| \ \leq \sup_{g\in\mathcal{F}^*,\ \sqrt{t\delta_n}\leq\|g\|_n\leq\|\hat{\Delta}_n\|_n} \frac{\sigma}{n}\left|\sum_{i=1}^n w_i g(x_i)\right| \ \leq\ 2\|\hat{\Delta}_n\|\sqrt{t\delta_n}$$

and therefore $\|\hat{\Delta}_n\|^2 \leq 16t\delta_n$. ∎

Now we prove the lemma that bounds the probability of bad event.

Lemma 21.2 *Let $\mathcal{H}$ be star shaped. For $u \geq \delta_n$ (critical radius of $\mathcal{H}$), the event*

$$A(u) := \left\{\exists g \in \mathcal{H},\ \|g\|_n \geq u \ \middle|\ \Big|\frac{\sigma}{n}\sum_{i=1}^n w_i g(x_i)\Big| \geq 2u\|g\|_n\right\}$$

has

$$P(A(u)) \leq \exp\{-\frac{nu^2}{2\sigma^2}\}.$$

Proof: We show that the bad event $A(u)$ implies the maximum of a Gaussian Process deviates much from its mean. In particular, suppose there is $g \in \mathcal{H}$ s.t. $\|g\|_n \geq u$ and

$$\frac{\sigma}{n}\left|\sum_{i=1}^{n} w_i g(x_i)\right| \geq 2u\|g\|_n.$$

Then let $\tilde{g} := g\frac{u}{\|g\|_n}$, we have $\|\tilde{g}\|_n = u$ and $\tilde{g} \in \mathcal{H}$ (due to $\mathcal{H}$'s star shape). Then if $A(u)$ occurs, we will also have $\tilde{g} \in \mathcal{H}$ such that

$$\frac{\sigma}{n}\left|\sum_{i=1}^{n} w_i \tilde{g}(x_i)\right| = \frac{u}{\|g\|_n}\left|\frac{\sigma}{n}\sum_{i=1}^{n} w_i g(x_i)\right| \geq 2u^2,$$

which is equivalent to the event $Z_n(u) \geq 2u^2$ where

$$Z_n(u) := \sup_{\tilde{g}\in\mathcal{H},\ \|\tilde{g}\|_n \leq u} \frac{\sigma}{n}\left|\sum_{i=1}^{n} w_i \tilde{g}(x_i)\right|$$

is the supremun of Gaussian Process

$$\frac{\sigma}{n}\sum_{i=1}^{n} w_i \tilde{g}(x_i) \sim N(0, \frac{\sigma^2}{n}\|\tilde{g}\|_n^2).$$

Recall that if $Z = (Z_1, ..., Z_n)$ are i.i.d. $N(0, \sigma^2)$ then if $f : \mathbb{R}^n \to \mathbb{R}$ is Lipschitz with parameter L, we have

$$P(|f(Z) - E[f(Z)]| > t) \leq 2\exp\left\{-\frac{t^2}{2L^2\sigma^2}\right\}.$$

Now notice that $Z_n(u)$ is a Lipschitz function w.r.t. $(w_1, ..., w_n)$ with parameter $L = \sigma u/\sqrt{n}$, so for any $s > 0$,

$$P\left(Z_n(u) \geq E[Z_n(u)] + s\right) \leq \exp\left(-\frac{ns^2}{2u^2\sigma^2}\right).$$

Letting $s = u^2$, we obtain

$$P\left(Z_n(u) \geq E[Z_n(u)] + u^2\right) \leq \exp\left(-\frac{nu^2}{2\sigma^2}\right).$$

To bound the expectation, note $\sigma\mathcal{G}_n(u) = E[Z_n(u)]$. Since $u \geq \delta_n$ (by assumption) and $\mathcal{G}_n(u)/u$ is non-increasing w.r.t. u,

$$\sigma\frac{\mathcal{G}_n(u)}{u} \leq \sigma\frac{\mathcal{G}_n(\delta_n)}{\delta_n} = \frac{\delta_n}{2},$$

which implies

$$E[Z_n(u)] \leq u\delta_n \leq u^2.$$

In conclusion,

$$P(A(u)) \leq P(Z_n(u) \geq 2u^2) \leq P(Z_n(u) \geq E[Z_n(u)] + u^2) \leq \exp\left(-\frac{nu^2}{2\sigma^2}\right).$$

■

21.1.3 How to compute critical radius δ_n?

The critical radius is generally hard to compute. In practice, we look for upper bound of δ_n that satisfies the *critical inequality* (21.1). A very loose bound that holds for all function classes $\mathcal{F}$ is $\delta_n \leq \sigma$. In most of cases, we can obtain much tighter results by bounding the local Gaussian Complexity using, for example, Dudley integral.

Since not all function classes F^* are *star-shaped*, in general, we can find an upper bound of δ_n on the *Star Hull* of F^*:

$$star(\mathcal{F}^*) := \{\alpha f \mid f \in \mathcal{F}^*, \alpha \in [0,1]\}.$$

Define the δ-radius ball of $\mathcal{F}^*$:

$$B_n(\mathcal{F}^*, \delta) = \{h \in star(\mathcal{F}^*) \mid \|h\|_n \leq \delta\},$$

and let $N(u)$ be the u-covering number of $B_n(\mathcal{F}^*, \delta)$ in the $\|.\|_n$ norm. Then we have the following lemma.

Lemma 21.3 *Any $\delta \in (0, \sigma]$ satisfying*

$$\frac{16}{\sqrt{n}} \int_{\delta^2/4\sigma}^{\delta} \sqrt{\log N(u)} du \leq \frac{\delta^2}{4\sigma}$$

serves as an upper bound on the critical radius δ_n.

Proof: (Sketch) Let $(g_1, ..., g_N)$ be a minimal $\frac{\delta^2}{4\sigma}$-covering of $B_n(\mathcal{F}^*, \delta)$ in $\|.\|_n$. Then we have

$$\mathcal{G}_n(\delta) \leq E\left[\max_{J=1}^{N} \frac{1}{n} \left|\sum_{i=1}^{n} w_i g_J(x_i)\right|\right] + \frac{\delta^2}{4\sigma} \tag{21.3}$$

since $\sqrt{\frac{\sum_{i=1}^n w_i^2}{n}} \leq 1$ and $\sqrt{\frac{\sum_{i=1}^n (g(x_i) - g_J(x_i))^2}{n}} = \|g - g_J\|_n \leq \frac{\delta^2}{4\sigma}$. Then applying chaining argument to bound the first term in (21.3), we obtain

$$\mathcal{G}_n(\delta) \leq \frac{16}{\sqrt{n}} \int_{\delta^2/4\sigma}^{\delta} \sqrt{\log N(u)} du + \frac{\delta^2}{4\sigma},$$

which is less or equal to $\frac{\delta^2}{2\sigma}$ as desired as long as

$$\int_{\delta^2/4\sigma}^{\delta} \sqrt{\log N(u)} du \leq \frac{\delta^2}{2\sigma}.$$

■

Example (Linear Regression) Let X be an $n \times d$ design matrix with rows $\{x_i\}_{i=1}^n$. We consider the linear function class

$$\mathcal{F}_{Lin} := \left\{f(x) = \langle\theta, x\rangle \mid \theta \in \mathbb{R}^d\right\}.$$

In this example, we use the general theory to show that the least-square estimate $f_{\hat{\theta}}$ satisfies

$$\|f_{\hat{\theta}} - f_{\theta^*}\|_n^2 = \frac{\|X(\hat{\theta} - \theta^*)\|^2}{n} \lesssim \sigma^2 \frac{rank(X)}{n}.$$

Note in this special case we have $\mathcal{F}^*_{Lin} = \mathcal{F}_{Lin}$ for any choice of f^* and $\mathcal{F}_{Lin}$ is convex and hence it is also star-shaped. Now notice that $\|f_\theta\|_n$ defines a norm on $range(X)$ and $B_n(\mathcal{F}^*, \delta)$ is isomorphic to a δ-ball in $range(X)$. Therefore, by the volume ratio argument (in Ch. 5),

$$\log N(u) \leq r \log(1 + \frac{2\delta}{u})$$

where $r = rank(X) = dim(range(X))$. Then we have

$$\begin{aligned} \frac{1}{\sqrt{n}} \int_0^\delta \sqrt{\log N(u)} du &\leq \frac{1}{\sqrt{n}} \int_0^\delta \sqrt{r \log(1 + \frac{2\delta}{u})} du \\ &= \delta \sqrt{\frac{r}{n}} \int_0^1 \sqrt{\log(1 + \frac{2}{u})} du \\ &\leq c\delta \sqrt{\frac{r}{n}} \end{aligned}$$

for some constant c. And therefore, we only need an δ satisfying

$$c\delta \sqrt{\frac{r}{n}} \leq \frac{\delta^2}{4\sigma},$$

which gives an

$$\delta \asymp \sigma \sqrt{\frac{r}{n}}.$$