36-709: Advanced Statistical Theory I **Spring 2020**

Lecture 21: April 9

Lecturer: Siva Balakrishnan *Scribe: Xuejian Wang*

21.1 Recap

1. Yatracos' method and robust estimation:
Even when model is misspecified, if $n\epsilon^2 \gg \log N(\epsilon)$,

$$\mathbb{P}\left(\text{TV}(\widehat{p}, p_0) \geq 3 \inf_{p \in \mathcal{P}} \text{TV}(p, p_0) + \epsilon\right) \leq C \exp\{-cn\epsilon^2\}$$

2. Non-parametric least squares:
We have $\{(x_1, y_1), ..., (x_n, y_n)\} \sim P_{XY}$ and

$$y_i = f^\star(x_i) + \sigma w_i, \quad w_i \sim \mathcal{N}(0, 1).$$

We want

$$\widehat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_i (y_i - f(x_i))^2.$$

3. Main claim:
Local Gaussian width here is defined as:

$$\mathcal{G}_n(\sigma) = \mathbb{E} \sup_{\substack{g \in \widetilde{\mathcal{F}} \\ \|g\|_n \leq \delta}} \left| \frac{1}{n} \sum_i w_i g(x_i) \right|.$$

If δ_n satisfies critical inequality:

$$\frac{\mathcal{G}}{\delta} \leq \frac{\delta}{2\sigma},$$

then $\forall t \geq \delta_n$, we have the following bounds:

$$\mathbb{P}(\|\widehat{f} - f\|_n^2 \leq 16t\delta) \leq \exp\{\frac{-nt\delta}{2\sigma^2}\}.$$

21.2 Bounds via metric entropy

For any function class $\mathcal{F}$, we define $\mathbb{B}_n(\delta; \mathcal{F}) = \{h \in \text{star}(\mathcal{F}) \mid \|h\|_n \leq \delta\}$, and we let $N_n(t; \mathbb{B}_n(\delta; \mathcal{F}))$ denote the t-covering number of $\mathbb{B}_n(\delta; \mathcal{F})$ in the norm $\|\cdot\|_n$.

Theorem 21.1 *Under the condition of Theorem 20.1, any $\delta \in (0, \sigma]$ such that*

$$\frac{16}{\sqrt{n}} \int_{\frac{\delta^2}{4\sigma}}^{\delta} \sqrt{\log N_n(t; \mathbb{B}_n(\delta; \mathcal{F})} dt \leq \frac{\delta^2}{4\sigma}$$

satisfies the critical inequality, and hence can be used in the conclusion of Theorem 20.1.

21.2.1 Examples:

1. Non-parametric Least Squares over class of L-Lipschitz functions:

$$\mathcal{F}_L = \{f : [0,1] \to \mathbb{R} \;\big|\; f(0) = 0, |f(x) - f(y)| \leq L|x - y|\}$$

The global metric entropy of this class that we computed before, scales as $N(\mathcal{F}, u) \asymp \frac{L}{u}$. We need to find δ such that:

$$\frac{C}{\sqrt{n}} \int_0^{\delta} \sqrt{\frac{L}{u}} du \leq \frac{\delta^2}{4\sigma} \tag{21.1}$$

Such a δ will satisfy the critical inequality, solve it out and we have:

$$\frac{C}{\sqrt{n}} \sqrt{L} \times \sqrt{\delta} \leq \frac{\delta^2}{4\sigma} \tag{21.2}$$

$$\Rightarrow \quad \delta^{3/2} \gtrsim \frac{\sigma\sqrt{L}}{\sqrt{n}} \tag{21.3}$$

$$\delta \gtrsim \left(\sigma\sqrt{\frac{L}{n}}\right)^{2/3}. \tag{21.4}$$

So

$$\mathbb{E}\|\widehat{f} - f^\star\|_n^2 \precsim \delta^2 \precsim \left(\frac{\sigma^2 L}{n}\right)^{2/3}. \tag{21.5}$$

Note that here we bounded by the global metric entropy instead of local metric entropy, and it turns out for most non-parametric classes, they are the same up to a constant. So it won't matter much when we use global metric entropy here, but it matters in the following case.

2. Parametric problems: linear regression settings

$$\mathcal{F} = \{f_\theta =< \theta, x >, \theta \in \mathbb{R}^d\}$$
$$\|X\widehat{\theta} - X\theta^\star\|_n^2 \precsim \frac{\sigma^2 \text{rank}(X)}{n}$$

Based on previous heuristics,

$$n\epsilon^2 \asymp \log N(\epsilon)$$
$$n\epsilon^2 \asymp d \log(\frac{1}{\epsilon})$$

Find δ such that:

$$\frac{C}{\sqrt{n}} \int_{\delta^2}^{\delta} \sqrt{\log\left[\left(1 + \frac{\delta}{u}\right)^d\right]} du \leq \frac{\delta^2}{4\sigma} \tag{21.6}$$

Let $v = \frac{u}{\delta}$ and rearrange the terms, we have:

$$\delta\sqrt{\frac{d}{n}} \int_0^1 \sqrt{\log\left(1 + \frac{1}{v}\right)} dv \precsim \frac{\delta^2}{4\sigma}. \tag{21.7}$$

The integral part turns out to be some constant, and

$$\delta \geq \sigma\sqrt{\frac{d}{n}} \tag{21.8}$$
$$\mathbb{E}\|X\widehat{\theta} - X\theta^\star\|_n^2 \precsim \frac{\sigma^2 d}{n} \tag{21.9}$$

21.3 A more rigorous proof of Theorem 20.1

Proof: We start with the basic inequality first:

$$\frac{\|\widehat{\Delta}\|_n^2}{2} \leq \frac{\sigma}{n}\sum_{i=1}^n w_i\widehat{\Delta}(x_i) \tag{21.10}$$

And we define the following event:

$$\mathcal{A}(u) = \{\exists g \in \widetilde{\mathcal{F}}, \|g\|_n \geq u, \left|\frac{\sigma}{n}\sum_{i=1}^n w_i g(x_i)\right| \geq 2\|g\|_n u\}, \;\; \widetilde{\mathcal{F}} = \mathcal{F} - f^\star. \tag{21.11}$$

The main idea is that $\mathcal{A}(u)$ is very unlikely, i.e., for all the functions outside a radius from $f^\star$, the Gaussian width will be somehow upperbounded. Roughly, you can rescale any function outside the radius back to the ring according to star-shaped property.

We want to show if $u \geq \delta_n$:

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

Pick $u = \sqrt{t\delta_n}, t \geq \delta_n$, then on $\mathcal{A}(u)^c$:

$$\{\text{No function}, \;\; \|g\|_n \geq \sqrt{t\delta_n}, \;\; |\frac{\sigma}{n}\sum_i w_i g(x_i) \geq 2\|g\|_u\},$$

there are two cases:
(1) $\|\Delta\|_n \leq \sqrt{t\delta_n}$: in this case we are fine.
(2) $\|\Delta\|_n \geq \sqrt{t\delta_n}$: then

$$\left|\frac{\sigma}{n}\sum_i w_i\Delta(x_i)\right| \leq 2\|\Delta\| \times \sqrt{t\delta_n}$$

Use basic inequality:

$$\frac{\|\Delta\|_n^2}{2} \leq \left|\frac{\sigma}{n}\sum_i w_i\Delta(x_i)\right| \leq 2\|\Delta\| \times \sqrt{t\delta_n} \tag{21.12}$$

$$\|\Delta\|_n^2 \leq 16t\delta_n \tag{21.13}$$

So far, we haven't reached anything substantial, since we assumed $\mathcal{A}(u)^c$. In the next class we will prove that $\mathcal{A}(u)$ is a low-probability event. ■