Stats 300b: Theory of Statistics **Winter 2018**

Lecture 13 – February 20

Lecturer: John Duchi *Scribe: Andrea Zanette*

Warning: *these notes may contain factual errors*

Reading: VDV 18-19

1 VC Classes and Dimension

Definition 1.1. (Vapnik-Chervonenkis classes) *$\mathcal{C}$(= collection of sets) shatters $x_1, ..., x_n$ if for all labelings $y \in \{\pm 1\}^n$ of $\{x_i\}$ $\exists C \in \mathcal{C}$, s.t.*

$$\begin{cases} y_i = 1 & x_i \in C \\ y_i = 0 & x_i \notin C \end{cases}$$

Definition 1.2. (Vapnik-Chervonenkis dimension) *VC($\mathcal{C}$)=size of largest set $x_1, ..., x_n$ shattered by $\mathcal{C}$.*

Theorem 1. (Uniform covering numbers in $L_r(P)$) *For sets A,B, let dist(A,B)=$||\mathbb{1}_A - \mathbb{1}_B||_{L_r(P)} = (\int |\mathbb{1}_A - \mathbb{1}_B|^r dP)^{\frac{1}{r}}$. Then $\exists$ constant $K < \infty$,*

$$\sup_P N(\mathcal{C}, Lr(P), \epsilon) \leq KVC(\mathcal{C})(4e)^{VC(\mathcal{C})}(\frac{1}{\epsilon})^{rVC(\mathcal{C})}$$

i.e.

$$\log N(\mathcal{C}, Lr(P), \epsilon) \lesssim rVC(\mathcal{C}) \log(\frac{1}{\epsilon})$$

Example 1: Let $\mathcal{F} = \{f(x) = \mathbb{1}_{X \leq t}, t \in \mathbb{R}^d\}$. Then $VC(\mathcal{F})$=$O(d)$.

$$\sup_P \log N(\mathcal{F}, L_2(P), \epsilon) \leq Kd \log(\frac{1}{\epsilon})$$

As a consequence, we have the classical Glivenko Cantelli theorem:

$$\begin{aligned} \mathbb{E}[\sup_{t \in \mathbb{R}^d} |\mathbb{P}_n(X \leq t) - \mathbb{P}(X \leq t)|] &= \mathbb{E}[\sup_{f \in \mathcal{F}} |\mathbb{P}_n f - \mathbb{P} f|] \\ &\leq \frac{2}{\sqrt{n}} \mathbb{E}[|\sup_{f \in \mathcal{F}|} \frac{1}{\sqrt{n}} \sum_{i=1}^n \epsilon_i f(X_i)] (\epsilon_i \overset{i.i.d}{\sim} \{\pm 1\}) \\ &\overset{Dudley}{\leq} \frac{const}{\sqrt{n}} \int_0^1 \sqrt{\log N(\mathcal{F}, L_2(P_n), \epsilon)} d\epsilon \\ &= \frac{const\sqrt{d}}{\sqrt{n}} \int_0^1 \sqrt{\log(1/\epsilon)} d\epsilon \\ &\leq \frac{\tilde{const}\sqrt{d}}{\sqrt{n}} \end{aligned}$$

♣

Example 2: Hyperplanes in $\mathbb{R}^d$ are sets of the form $\{x \in \mathbb{R}^d \mid w^\top x \leq b, w \in \mathbb{R}^d, b \in \mathbb{R}\}$. In $\mathbb{R}^2$ we can shatter 3 points. ♣

Definition 1.3. *The subgraph of a function:* $\mathcal{X} \to \mathbb{R}$*:*

$$subf := \{(x,t) : t < f(x)\} = (epif)^c$$

Note: sub$f \subseteq \mathcal{X} \subseteq \mathbb{R}$.

Definition 1.4. $\mathcal{F}$ *is a VC-class (VC-subgraph-class) if* $subf : f \in \mathcal{F}$ *is VC.*

2 Rates of Convergence of Estimators and Modulus of Continuity

We want to understand and use continuity properties of empirical processes and estimators. Consider the following M-estimator problem of minimizing the risk $R(\theta)$. Let $\theta_0 = \text{argmin}_\theta R(\theta)$. Assume that $R(\theta) \geq R(\theta_0) + \lambda d(\theta, \theta_0)^\alpha$ for distance $d : \theta \times \theta \to R, \alpha > 0$. Suppose we have R_n such that $R_n \to R$ in some sense. How do we get a handle on the convergence of θ_n?
Example 3: Let $X \sim P_{\theta_0}, l(\theta, x) = -\log P_\theta(x), R(\theta) = E_{P_{\theta_0}} l(\theta, x)$. Then

$$R(\theta) - R(\theta_0) = \frac{1}{2}(\theta - \theta_0)^\top \nabla^2 R(\theta_0)(\theta - \theta_0) + O(\|\theta - \theta_0\|^2)$$

Here we could take $\alpha = 2, d(\theta, \theta_0) = \|\theta - \theta_0\|$. Assume θ_n is consistent under the distance d. We define the localized empirical process

$$\Delta_n(\theta) = (R_n(\theta) - R(\theta)) - (R_n(\theta_0) - R(\theta_0))$$

Definition 2.1. *The modulus of continuity of* R_n *at point* θ_0 *is* $w_= \sup_{d(\theta,\theta_0)\leq\delta} |\Delta_n|$

♣

With this is mind we present our main result of the lecture:

Theorem 2. *Suppose* $M(\theta_0) \geq M(\theta) + d(\theta, \theta_0)^2$ *near* θ_0.
Let ϕ *be such that* $\phi(c\delta) \leq c^\alpha \phi(\delta)$ *for some* $\alpha \in (0, 2)$.
Assume

$$\mathbb{E}\Big[\sup_{d(\theta,\theta_0)\leq\delta} \big|M_n(\theta) - M(\theta) - (M_n(\theta_0) - M(\theta_0))\big|\Big] \leq \frac{\phi(\delta)}{\sqrt{n}}$$

Let $r_n \to +\infty$ *such that* $r_n^2 \phi(\frac{1}{r_n}) \leq \sqrt{n}$.
If $\hat{\theta}_n \to \theta_0$ *in probability, then* $r_n d(\hat{\theta}_n, \theta_0) = O_P(1)$

Idea: Estimation errors can scale at most as $\frac{\phi(\delta)}{\sqrt{n}} \sim \frac{d(\theta,\theta_0)^\alpha}{\sqrt{n}}$ with $\alpha < 2$. But the growth of the objective function is quadratic: $d(\theta, \theta_0)^2$.
We solve $\delta^2 = \frac{\delta^\alpha}{\sqrt{n}}$: it implies $\delta = \big(\frac{1}{n}\big)^{\frac{1}{2(2-\alpha)}}$

Proof Let $\eta \leq \delta$. Then $\mathbb{P}\big(d(\hat{\theta}_n, \theta_0) \geq \eta\big) = o(1)$ (consistency of the estimator).
Define the shell: $S_{nj} = \{\theta : 2^{j-1} \leq r_n d(\theta, \theta_0) \leq 2^j\}$.

Consider the event

$$d(\hat{\theta}_n, \theta_0) \geq \frac{2^t}{r_n}$$

for some $t \geq 0$.

Then $\exists j \geq t$ such that $\hat{\theta}_n \in S_{jn}$.

Therefore

$$\begin{aligned}
\mathbb{P}(r_n d(\hat{\theta}_n, \theta_0) \geq 2^t) &\leq \mathbb{P}(r_n d(\hat{\theta}_n, \theta_0) \geq 2^t, d(\hat{\theta}_n, \theta_0) \leq \eta) + \mathbb{P}(d(\hat{\theta}_n, \theta_0) > \eta) \\
&= \Big(\sum_{j \leq t,\, 2^{j-1} \leq r_n \eta} \mathbb{P}(\hat{\theta}_n \in S_{jn}) \Big) + \mathbb{P}(d(\hat{\theta}_n, \theta_0) > \eta) \\
&\leq \Big(\sum_{j \leq t,\, 2^{j-1} \leq r_n \eta} \mathbb{P}(\exists \theta \in S_{jn}, M_n(\theta) \geq M_n(\theta_0)) \Big) + o(1)
\end{aligned}$$

But if $M_n(\theta) \geq M_n(\theta_0)$ for some $\theta \in S_{jn}$ then

$$\begin{aligned}
M_n(\theta) - M(\theta) &\geq M_n(\theta_0) - M(\theta_0) + M(\theta_0) - M(\theta) \\
&\geq M_n(\theta_0) - M(\theta_0) + d(\theta, \theta_0)^2
\end{aligned}$$

So

$$\sup_{\theta \in S_{jn}} \big| (M_n(\theta) - M(\theta)) - (M_n(\theta_0) - M(\theta_0)) \big| \geq \frac{2^{2j}}{r_n^2}$$

And

$$\begin{aligned}
&\mathbb{P}(\exists \theta \in S_{jn}, M_n(\theta) \geq M_n(\theta_0)) \\
&\qquad \leq \mathbb{P}\big(\sup_{\theta \in S_{jn}} \big| (M_n(\theta) - M(\theta)) - (M_n(\theta_0) - M(\theta_0)) \big| \geq \frac{4^j}{r_n^2} \big) \\
&\qquad \leq \frac{r_n^2}{4^j} \, \mathbb{E}\big[\sup_{\theta \in S_{jn}} \big| (M_n(\theta) - M(\theta)) - (M_n(\theta_0) - M(\theta_0)) \big| \big] \\
&\qquad \leq \frac{r_n^2}{4^j \sqrt{n}} \, \phi\big(\frac{2^j}{r_n}\big) \;\leq\; \frac{r_n^2 \, 2^{\alpha j}}{4^j \sqrt{n}} \phi\big(\frac{1}{r_n}\big) \;\leq\; 2^{j(\alpha - 2)}
\end{aligned}$$

So we bound the inequality above with this result and get:

$$\mathbb{P}(d(\hat{\theta}_n, \theta_0) \geq \frac{2^t}{r_n}) \leq \sum_{j \geq t} 2^{-j(2-\alpha)} + o(1)$$

So by taking t large enough we get that $r_n d(\hat{\theta}_n, \theta_0) = O_P(1)$

□