Stats 300b: Theory of Statistics **Winter 2019**

Lecture 12 – February 14

Lecturer: John Duchi *Scribe: Benjamin Seiler*

Warning: *these notes may contain factual errors*

Reading: HDP Ch.8, VdV 18-19

Outline:

- Uniform Entropy Bounds
  - VC Classes
  - VC Function Classes
- Chaining
  - Entropy Integrals
  - Sub-Gaussian Processes

Recap: VC Classes of Sets

Definition 0.1. *Given $\mathcal{C}$ a collection of sets, the shattering coefficient of $\mathcal{C}$ on $x_1, x_2, ...x_n$ is $\Delta_n(\mathcal{C}, x_{1:n}) := card\{A \cap x_1, ...x_n : A \in \mathcal{C}\}$ = the number of labelings $\mathcal{C}$ can realize on $x_{1:n}$.*

Definition 0.2. *The VC-dimension (Vapnik-Chervonenkis) of $\mathcal{C}$ is*
$VC(\mathcal{C}) := \sup\{n \in \mathbb{N} : \max_{X_{1:n} \in \mathcal{X}^n} \Delta_n(\mathcal{C}, x_{1:n}) = 2^n\}$ = the size of the largest set of points that $\mathcal{C}$ can shatter.

Fact 1. *Half Spaces $\mathcal{C}$ in $\mathbb{R}^d$ have $VC(\mathcal{C}) = d + 1$*

Lemma 2. *Sauer-Shelah lemma For any class $\mathcal{C}$ of sets,*

$$\max_{x_{1:n} \in X^n} \Delta_n(\mathcal{C}, x_{1:n}) \leq \sum_{j=0}^{VC(\mathcal{C})} \binom{n}{j} = O(n^{VC(\mathcal{C})})$$

Consequence: *If $\max_{x_{1:n} \in X^n} \Delta_n(\mathcal{C}, x_{1:n}) < 2^n$, then $VC(\mathcal{C}) < n$ and*

$$\Delta_n(\mathcal{C}, x_{1:n}) \leq O(1) \cdot n^{VC(\mathcal{C})}$$

. Additional lectures notes on the course website provide a further reference on this topic.

New Material: ***Uniform Entropy Bounds (continued)***

Definition 0.3. *We will define the $L_r(P)$ norm on sets using indicator functions as follows: $||\mathbb{1}_A - \mathbb{1}_B||_{L_r(P)} = (\int |\mathbb{1}_A - \mathbb{1}_B|^r dP)^{\frac{1}{r}}$*

Theorem 3. (Uniform covering numbers in $L_r(P)$) *For a collection of sets $\mathcal{C}$, there $\exists$ constant $K < \infty$ s.t.,*

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

i.e.

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

Note: this is true for all P simultaneously and here $e \approx 2.718281828459045$

Sketch of Proof *The rough idea is that if we take our space and cover it with ϵ-balls of probability we can only shatter so many $0, 1$ functions on these balls. We have $\frac{1}{\epsilon}$ such balls so $\mathcal{C}$ can only realize $\frac{1}{\epsilon}^{VC(\mathcal{C})}$ different $0, 1$-valued functions on them and the covering number must be less than the total number of such functions.*

Note: Much like Paul "The Truth" Pierce who could not win an NBA championship on his own, a full treatment of this proof requires outside help and can be found on the course website in the additional notes section.

□

Example 1: **Lower Left Rectangles:** *Let $\mathcal{F} = \{f(x) = \mathbb{1}_{\{X \leq t\}}, t \in \mathbb{R}^d\}$. Then $VC(\mathcal{F})$=$O(d)$ and $\exists C < \infty$ s.t.*

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

As a consequence, we have the classical Glivenko Cantelli theorem in $\mathbb{R}^d$:

$$\begin{aligned}\mathbb{E}[\sup_{t\in\mathbb{R}^d} |\mathbb{P}_n(X \leq t) - \mathbb{P}(X \leq t)|] &= \mathbb{E}[\sup_{f\in\mathcal{F}} |\mathbb{P}_n f - \mathbb{P}f|] \\ &\leq C(\frac{1}{\sqrt{n}}\sqrt{\log N(\mathcal{F}, L_1(P), \epsilon)} + \epsilon) \\ &\leq C(\sqrt{\frac{d}{n}\log\frac{1}{\epsilon}} + \epsilon)\end{aligned}$$

Let $\epsilon = \sqrt{d/n}$

$$\leq C\sqrt{\frac{d}{n}\log\frac{n}{d}}$$

Note: this is not the tightest bound we will get (with chaining we will be able to lose the logn term) ♣

VC Classes of Functions:

Definition 0.4. *The subgraph of a function $f : \mathbb{X} \to \mathbb{R}$ is the set $subf \subset \mathbb{X} \times \mathbb{R}$ s.t. $subf = \{(x, t) : f(x) > t\}$.*

Definition 0.5. *$\mathcal{F} \subset \{\mathbb{X} \to \mathbb{R}\}$ is a VC-class if $\mathcal{C} = \{subf : f \in \mathcal{F}\}$ forms a VC collection in $\mathbb{X} \times \mathbb{R}$ i.e. $VC(\mathcal{C}) < \infty$. and $VC(\mathcal{F}) := VC(\{subf : f \in \mathcal{F}\})$.*

Theorem 4. *If $VC(\mathcal{F}) < \infty$ and $\mathcal{F}$ has an envelope function F i.e. $F(x) \geq |f(x)|\ \forall f \in \mathcal{F}$ where $F \in L_r(P)$ i.e. $\mathbb{E}[F^r(x)] \leq \infty$, then $\exists K \leq \infty$ s.t.*

$$\sup_P N(\mathcal{F}, L_r(P), ||F||_{L_r(P)}\epsilon) \leq KVC(\mathcal{F})(16e)^{VC(\mathcal{F})}(\frac{1}{\epsilon})^{rVC(\mathcal{F})}$$

if $0 < \epsilon < 1$

Sketch of Proof *Form subgraphs to approximate f and then apply the previous theorem.* □

Example 2: **Classification Problem**
Suppose we have data of the form $(x, y) \in \mathbb{R}^d \times \{-1, 1\}$ and we train a binary linear classifier on the data to predict y. We compute some $\theta \in \mathbb{R}^d$ s.t. our predictions are of the form $sgn(\theta^T x) = \mathbb{1}_{\{\theta^T x > 0\}} - \mathbb{1}_{\{\theta^T x \leq 0\}}$. Our goal then is to find θ s.t. $P(sgn(\theta^T x) \neq y)$ is small.
Consider $\mathcal{F} = \{f(x) = \theta^T x, \theta \in \mathbb{R}^d\}$. $\mathcal{F}$ is a VC-class of functions with $VC(\mathcal{F}) = d + 1$. Given a sample $(x_i, y_i) \overset{iid}{\sim} P, (i = 1, 2, \ldots, n)$, we can see based on the previous example that there $\exists C \leq \infty$ s.t.

$$\mathbb{E}[\sup_{f \in \mathcal{F}} |P_n(sgn(\theta^T x) \neq y) - P(sgn(\theta^T x) \neq y)|] \leq C\sqrt{\frac{d}{n}\log\frac{n}{d}}$$

♣

Calculus of VC-Properties:

- *If $\mathcal{F}$ is a linear space of functions with $dim(\mathcal{F}) \leq \infty$ then $VC(\mathcal{F}) = O(1)dim(\mathcal{F})$*
- **Preservation** *If $\mathcal{C}$ and $\mathbb{D}$ are VC classes of functions then:*
  $\mathcal{C} \sqcup \mathbb{D} := \{C \cup D : C \in \mathcal{C}, D \in \mathbb{D}\}$ *is VC*
  $\mathcal{C} \sqcap \mathbb{D} := \{C \cap D : C \in \mathcal{C}, D \in \mathbb{D}\}$ *is VC*
- **Composition** *If $\mathcal{F}$ is a VC collection of functions and $\phi : \mathbb{R} \to \mathbb{R}$ is monotone, then $\phi \circ \mathcal{F}$ is VC.*

Chaining

Goal: *Achieve tighter/sharper bounds on $\mathbb{E}[\sup_{f \in \mathcal{F}} |\mathbb{P}_n f - \mathbb{P}f|]$ or more specifically for sub-gaussian processes $\{X_t\}_{t \in \mathcal{T}}$, we want sharp/tight bounds on $\mathbb{E}[\sup_{t \in \mathcal{T}} |X_t|]$.*
Recall: *$\{X_t\}_{t \in \mathcal{T}}$ is a sub-Gaussian process for a metric don the space $\mathcal{T}$ if $\mathbb{E}[\exp(\lambda(x_s - x_t))] \leq \exp(\frac{\lambda^2 d(s,t)^2}{2})\ \forall s, t \in \mathcal{T}, \lambda \in \mathbb{R}$ and that for our purposes, we can assume that $\{X_t\}$ is a seperable process i.e. $\exists \mathcal{T}'$ s.t. $\mathcal{T}'$ is countable and $\sup_{t \in \mathcal{T}} X_t = \sup_{t \in \mathcal{T}'} X_t$.*
Naive Approach: *We can approximate $\mathbb{E}[\sup_{t \in \mathcal{T}} |X_t|]$ with $\mathbb{E}[\max_{t \in \mathcal{T}_\epsilon} |X_t|]$ where T_ϵ is a discretization of $\mathcal{T}$. Issues with this approach are that we do not know how finely to discretize and cannot guarentee a correct discritization level. Instead we consider chaining:*
New Approach *Let $\{X_t\}_{t \in \mathcal{T}}$ be sub-Gaussian for a metric d, separable, and mean-zero, i.e. $\mathbb{E}[X_t] = 0$. The idea is to control $sup_{t \in \mathcal{T}} X_t$ by finer and finer approximations to the supremum. We can do this because the process is separable. Let $\mathcal{T}_0, \mathcal{T}_1, \mathcal{T}_2, ...\mathcal{T}$ be a sequence of covers of $\mathcal{T}$, where $\mathcal{T}$ = minimal 2^{-k} $diam(\mathcal{T})$ cover of $\mathcal{T}$ in the metric (or semimetric) d, where*

$diam(\mathcal{T}) := \sup_{s,t\in\mathcal{T}} d(s,t)$ (assumed finite), $\mathcal{T}_0 = \{t_0\}$, and $d(t_0,t) \leq diam(\mathcal{T})\ \forall t \in \mathcal{T}$.

For any $t \in \mathcal{T}$, consider sequences $t_0, t_1, ..., t_k, ... \to t$ where $t_k \in \mathcal{T}_k\ \forall k \in \mathbb{N}$. Let $\pi_i(t) = \arg\min_{t_i\in\mathcal{T}_i} \rho(t_i,t)$ be the closest point to t in $\mathcal{T}_i$. Fix any $k \in \mathbb{N}$. Then $x_i = x_{\pi_{k-1}(t)} + x_t - x_{\pi_{k-1}t)}$.

Let $\pi^i(t) := \pi_i(\pi_{i+1}(...(\pi_{k-1}(t))...)$ (a concatenation of projections). Observe that

$$x_t = \sum_{i=1}^{k} x^i_{\pi_k}(t) - x^{i-1}_{\pi_k}(t) + x^0_\pi(t) = \sum_{i=1}^{k} x^i_{\pi_k}(t) - x^{i-1}_{\pi_k}(t) + x_{t_0}$$

as $\pi^k_k(t) = t$. This is the "chain."

Remark *For any $k \in \mathbb{N}$, $\max_{t\in\mathcal{T}}(x_t) \leq \max_{t\in\mathcal{T}}(x^i_{\pi_k}(t) - x^{i-1}_{\pi_k}(t)) + x^0_\pi(t)$. How many points are there in this maximum? $\pi^i_k(t)$ takes values in $\mathcal{T}_i$ and $\pi^{i-1}_k(t) = \pi_{i-1}(\pi_k{}^i(t))$ is a deterministic function of $\pi^i_k(t)$. So this is really, at "worst", a maximum over points in a set $\mathcal{T}_i$.*

We know that if $D = diam(\mathcal{T})$, $d(\pi^i_i(t), \pi^{i-1}_k(t)) \leq 2^{1-i}D$ as $\pi^{i-1}_k(t) = \pi_{i-1}(\pi^i_k(t))$, $\mathcal{T}_{i-1}$ is a 2^{1-i} diameter cover of $\mathcal{T}$. Then,

$$\max_{t\in\mathcal{T}} x_t \leq \sum_{i=1}^{k} \max_{t\in\mathcal{T}}(x_t - x_{\pi_{i-1}}(t)) + x_0$$

where $t \in \mathcal{T} \max(x_t - x_{\pi_{i-1}}(t))$ is a finite maximum of $2^{1-i}D$-sub-Gaussian random variables. Recall that if $\{Y_i\}_{i=1}^N$ are σ^2-sub-Gaussian, then

$$\mathbb{E}[\max_i(Y_i)] \leq \sqrt{(2\sigma^2 \log(N))}$$

$$\mathbb{E}[\max_{t\in T_i}(x_t - x_{\pi_{i-1}}(t))] \leq \sqrt{4^{1-i}2D^2 \log|\mathcal{T}_i|}$$

where $Card(\mathcal{T}_i) = \mathcal{N}(\mathcal{T}, d, 2^{-i}D)$. Then,

$$\mathbb{E}[\max_{t\in\mathcal{T}_k}(x_t)] \leq \sum_{i=1}^{k} \sqrt{8 \cdot 4^{-1} D^2 \log \mathcal{N}(2^{-i}D)}$$

$$= 2\sqrt{(2)}D \sum_{i=1}^{k} 2^{-i} \sqrt{\log \mathcal{N}(D, 2^{-i})}$$

Note tht we can think of this as a Riemann integral, so

$$\mathbb{E}[\max_{t\in\mathcal{T}_k}(x_t)] \leq 2\sqrt{(2)}D \sum_{i=1}^{k} 2^{-i} \sqrt{\log \mathcal{N}(D, 2^{-i})}$$

$$\leq 4\sqrt{2}D \sum_{i=1}^{\infty} \int_{2^{-i+1}}^{2^{-i}} \sqrt{\log \mathcal{N}(D_\epsilon)} d\epsilon$$

$$= 4\sqrt{2}D \int_0^1 \sqrt{\log \mathcal{N}(D_\epsilon)} d\epsilon$$

$$= 4\sqrt{2}\int_0^{diam(T)} \sqrt{\log \mathcal{N}(\mathcal{T}, d, \epsilon)} d\epsilon$$

where the last equality comes from substituting ϵ for D_ϵ and letting $D = diam(\mathcal{T})$. Finally, note that $\max_{t \in \mathcal{T}_k \cup \mathcal{T}_0}(x_t - x_{t_0})$ is non-negative, so Fatou's lemma implies that

$$\mathbb{E}[\sup_{t \in \mathcal{T}_k}(x_t)] \leq 4\sqrt{2}\int_0^{diam(\mathcal{T})} \sqrt{\log \mathcal{N}(\mathcal{T}, d, \epsilon)} d\epsilon$$

and we can use MCT to go from $\mathcal{T}_k$ to $\mathcal{T}$.

Definition 0.6. *For a metric space $(\mathcal{T}, \rho)$ with finite ρ-diameter $J(\mathcal{T}, d) := \int_0^{diam(T)} \sqrt{\log \mathcal{N}(\mathcal{T}, d, \epsilon)} d\epsilon$ is Dudley's entropy integral.*

Theorem 5. *Let $\{X_t\}_{t \in \mathcal{T}}$ be a separable d-sub-Gaussian process. Then $\mathbb{E}[\sup_{t \in T}(X_t)] \leq C \cdot J(\mathcal{T}, d)$, where $C < \infty$ is a numerical constant.*