CS281B/Stat241B (Spring 2003) Statistical Learning Theory **Lecture: 19**

Chaining, Dudley's entropy integral, covering numbers of VC-classes

Lecturer: Peter L. Bartlett *Scribe: Zile Wei*

Theorem 1. (Dudley)

$$\hat{R}_n(\mathcal{F}) = \mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(x_i) \leq 12 \int_0^\infty \sqrt{\frac{\log N(\epsilon, \mathcal{F}, L_2(P_n))}{n}} d\epsilon$$

Proof. Set $B = \sup_{f \in \mathcal{F}} \|f\|_{L_2(P_n)}$. Choose $\alpha_0 = B$, $\alpha_i = B \times 2^{-i}, i = 0, 1, 2, \cdots$. Let T_i be a proper α_i-cover of $\mathcal{F}$ with respect to $L_2(P_n)$. For each $f \in \mathcal{F}$, choose $\hat{f}_i \in T_i$ such that $\|f - \hat{f}_i\|_{L_2(Pn)} \leq \alpha_i$

Chaining sequence
For any N, notice that $f = f - \hat{f}_N + \sum_{j=1}^{N} (\hat{f}_j - \hat{f}_{j-1})$. We can choose $T_0 = \{\hat{f}_0\} = \{0\}$ since $\|f - 0\| \leq B = \alpha_0$.

$$\begin{aligned}
\hat{R}_n(F) \leq & \mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \big(f(X_i) - \hat{f}_N(X_i)\big) + \sum_{j=1}^{N} \mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \big(\hat{f}_j(X_i) - \hat{f}_{j-1}(x_i)\big) \\
= & \mathbb{E} \sup_{f \in \mathcal{F}} \left\langle \epsilon, f - \hat{f}_N \right\rangle_{L_2(P_n)} + \sum_{j=1}^{N} \mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \big(\hat{f}_j(X_i) - \hat{f}_{j-1}(x_i)\big).
\end{aligned}$$

For the first term, we have

$$\begin{aligned}
\mathbb{E} \sup_{f \in \mathcal{F}} \left\langle \epsilon, f - \hat{f}_N \right\rangle_{L_2(P_n)} &\leq \mathbb{E} \sup_{f \in \mathcal{F}} \|\epsilon\| \|f - \hat{f}_N\|_{L_2(P_n)} \\
&\leq \alpha_N.
\end{aligned}$$

We can use Massart's finite lemma to bound the second term. Recall:

Lemma 2. (Finite Lemma) If $|\mathcal{F}|$ is finite and $\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} f^2(X_i) \leq R^2$, then

$$\mathbb{E}\Big(\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(x_i) \Big) \leq \sqrt{\frac{2R^2 \log |\mathcal{F}|}{n}}$$

To use the lemma, we need to bound the norm in $L_2(P_n)$ of $\hat{f}_j - \hat{f}_{j-1}$. We have

$$\begin{aligned}
\big(\hat{f}_j(X_i) - \hat{f}_{j-1}(X_i)\big)^2 &= \left\| \hat{f}_j - \hat{f}_{j-1} \right\|^2 \\
&\leq \Big(\left\| \hat{f}_j - f \right\| + \left\| f - \hat{f}_{j-1} \right\| \Big)^2 \\
&\leq (\alpha_j + \alpha_{j-1})^2 \\
&= 9\alpha_j^2.
\end{aligned}$$

Using Massart's lemma, we have

$$\begin{aligned}
&\mathbb{E}\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\big(\hat{f}_j(X_i)-\hat{f}_{j-1}(X_i)\big)\\
\le&\sqrt{\frac{2\log|T_j||T_{j-1}|}{n}}3\alpha_j\\
\le&6\alpha_j\sqrt{\frac{\log|T_i|}{n}}
\end{aligned}$$

Hence,

$$\begin{aligned}
\hat{R}_n(\mathcal{F})\le&\alpha_n+6\sum_{j=1}^{N}\alpha_j\sqrt{\frac{\log|T_i|}{n}}\\
=&\alpha_N+12\sum_{j=1}^{N}(\alpha_j-\alpha_{j-1})\sqrt{\frac{\log N\big(\alpha_j,\mathcal{F},L_2(P_n)\big)}{n}}\\
\le&\alpha_N+12\int_{\alpha_{N+1}}^{\infty}\sqrt{\frac{\log N\big(\alpha,\mathcal{F},L_2(P_n)\big)}{n}}d\alpha.
\end{aligned}$$

Taking the limit as $N\to\infty$ completes the proof. $\square$

Example. The covering number in non-decreasing function class from $\mathbb{R}$ to [-1,1] is bounded by $n^{1/\alpha}$. By Theorem 1, we can have

$$\begin{aligned}
\hat{R}_n(F)\le&12\int_0^{\infty}\sqrt{\frac{\log N\big(\alpha,\mathcal{F},L_2(P_n)\big)}{n}}d\alpha\\
\le&12\sqrt{\frac{\log n}{n}}\int_0^1\frac{d\alpha}{\sqrt{\alpha}} \qquad (1)\\
=&24\sqrt{\frac{\log n}{n}}
\end{aligned}$$

This is the correct rate. Using a cruder approach (without chaining), we earlier obtained the worse rate of $\left(\frac{\log n}{n}\right)^{1/3}$.

Example. $F\subseteq[-1,1]^X$ has $\log N\big(\alpha,F,L_2(P_n)\big)\le d\log\frac{1}{\alpha}$.

$$\begin{aligned}
\hat{R}_n(F)\le&12\sqrt{\frac{d}{n}}\int_0^1\sqrt{\log\frac{1}{\epsilon}}d\epsilon\\
=&12\sqrt{\frac{d}{n}}\int_{-\infty}^{0}e^{-y^2}dy \qquad (2)\\
=&12\sqrt{\frac{\pi}{2}}\cdot\sqrt{\frac{d}{n}}
\end{aligned}$$

In previous lecture, we got $c\sqrt{\frac{d\log n}{n}}$ without chaining.

Example. The converse result of Rademacher average (Sudakov)

The lower bound of Rademacher average is

$$\hat{R}_n(F) \geq \frac{c}{\log n} \sup_{\alpha > 0} \alpha \sqrt{\frac{\log N(\alpha, F, L_2(P_n))}{n}}$$

Gaussian averages:
Instead of using uniformly binary random variable ϵ_i in Rademacher average, we can replace it with a Gaussian distribution random variable. E.g.

$$G_n(F) = \mathbb{E} \sup_{f \in F} \frac{1}{n} \sum_{i=1}^{n} Z_i f(X_i), \quad Z_i \sim N(0,1).$$

In this case, Sudakov's lower bound is neater:

$$\hat{G}_n(F) \geq c \sup_{\alpha > 0} \alpha \sqrt{\frac{\log N(\alpha, F, L_2(P_n))}{n}}.$$

Example. VC-class

Since $\frac{1}{n} \sum_{i=1}^{n} (u_i - v_i)^2 \leq \max_i |u_i - v_i|$, it implies that the max-norm is always no less than the L_2 norm. Hence,

$$N(\epsilon, F, L_2(P_n)) \leq N(\epsilon, F, L_\infty(P_n)) \leq |F_{|D_n}| \tag{3}$$

In the last inequality, we use the fact that covering number is always smaller than set size. For VC-class, it's less than $\Pi_F(n)$. In the previous lecture, we had the result: $\hat{R}_n(F) \leq c\sqrt{\frac{d \log(en/d)}{n}}$, where $d_{VC}(F) = d$. This result can be improved by the following theorem. The theorem is due to Dudley; Haussler improved the constants.

Theorem 3. (Haussler)

For $F \subseteq \{\pm 1\}^X$, $d_{VC}(F) = d$, any P on X, then, $N(\epsilon, F, L_2(P_n)) \leq \left(\frac{c}{\epsilon}\right)^{2d}$. Hence $\hat{R}_n(F) \leq c' \sqrt{\frac{d}{n}}$.

To prove the theorem, we need some concepts about packing number and one-inclusion graphs of VC class, which are of independent interest.

Definition. Packing Number: For a metric space (S, ρ) and $T \subseteq S$, we say $\hat{T} \subseteq T$ is an ϵ-packing, if $\forall t_1, t_2 \in \hat{T}$, $\rho(t_1, t_2) \geq \epsilon$.

The ϵ-packing number of T is

$$M(\epsilon, T, \rho) = \max\{|\hat{T}| : \hat{T} \subseteq T \text{ is an } \epsilon\text{-packing }\}$$

Lemma 4.

$$M(2\epsilon, T, \rho) \leq N(\epsilon, T, \rho) \leq M(\epsilon, T, \rho)$$

Proof.
Step 1: $M(2\epsilon) \leq N(\epsilon)$

Firstly, we fix an ϵ-cover $\mathcal{C} \subseteq T$ and a 2ϵ-packing $\mathcal{P} \subseteq T$. $\forall p \in \mathcal{P}, \exists c \in \mathcal{C}, \rho(c, p) < \epsilon$. Suppose $|\mathcal{P}| > |\mathcal{C}|$, then, there exists $c \in \mathcal{C}$, which has two distinct $p_1, p_2 \in \mathcal{P}$ with $\rho(c, p_1) < \epsilon, \rho(c, p_2) < \epsilon$. By triangle inequality in metric space, $\rho(p_1, p_2) < 2\epsilon$. So $\mathcal{P}$ is not a maximal packing. Thus, a maximal packing $\mathcal{P}$ has $|\mathcal{P}| \leq |\mathcal{C}|$.

Step 2: $N(\epsilon) \leq M(\epsilon)$

Suppose $\mathcal{P}$ is an ϵ-packing, but not an ϵ-cover. This implies $\exists p' \in T, \forall p \in \mathcal{P}, \rho(p, p') \geq \epsilon$. So: $P \cup \{p\}$ is another ϵ-packing and larger than $\mathcal{P}$. So $\mathcal{P}$ is not a maximal packing. Thus, any maximal packing is also a cover. $\square$

Theorem 5. If $V \subseteq \{0,1\}^n$ has $d_{VC}(V) = d$, $d_{VC} = \max\{|I| : I \subseteq \{1, \cdots, n\}, |V_{|I}| = 2^{|I|}\}$ and define the metric d_1 on $\{0,1\}^n$ as $d_1(a,b) = \frac{1}{n}\sum_{i=1}^{n} |a_i - b_i|$, then

$$M(\epsilon, V, d_1) \leq 2\Big(\frac{4\epsilon}{n}\Big)^d$$

The idea of proof is related to randomly choosing m coordinates and projecting into those coordinate.

$$I_m = \{i_1, \cdots, i_m\} \subseteq \{1, \cdots, n\}$$

which is uniformly chosen from subset of m size. Here are two ingredients in the proof.

$\diamond$ Choose $Y, Y' \in V$ uniformly and independently. Project them to m coordinates.

$\diamond$ Given Y and Y' agree on m-1 coordinates, use posterior distribution to guess that m^{th} coordinate. We will show that

(1) Probability that samples from posterior disagree on m^{th} coordinates $\leq \frac{2d}{m}$
(2) Probability that samples from posterior disagree on m^{th} coordinates $\geq$ increasing function of $M(\epsilon, V, d)$.

Here are some notations for the following lemma.

$\Pi_m(a) = (a_{i_1}, \cdots, a_{i_m})$
$W = \Pi_m(V) \subseteq \{0,1\}^m$
Y and Y' are independent and uniform on V
$X = \Pi_m(Y), X' = \Pi_m(Y')$

Lemma 6. For all Π_m,

$$\frac{1}{m}\sum_{i=1}^{m} P(X_i \neq X_i' | \Pi_m, X_j = X_j' for j \neq i) \leq \frac{2d}{m}$$

Proof. Construct 1-inclusion graph (V,E) of $V \subseteq \{0,1\}^n$ with node set V and edges E, where $\{u, v\} \in E$ if $u, v \in V$ and satisfy $d_1(u,v) = 1/n$, which means u and v have only one difference in all directions.

Let (W,F) be a subgraph with $W = \Pi_m(V)$ (a projection over V). $F_i = \{(u,v) \in F : u_j = v_j, \text{ for } j \neq i\}$

contains edges only different in i^{th} direction. Then for any Π_m and i,

$$\begin{aligned}
&P(X_i \neq X_i' | X_j = X_j', j \neq i, \Pi_m) \\
=&\sum_{w \in W} P(X = w) P(X_i \neq X_i' | X_j = X_j' = w_j, \text{ for } j \neq i, \text{ and } X = w) \\
=&\sum_{\{v,w\} \in F_i} P(X = v) P(X' = w | X' \in \{v, w\}) + P(X = w) P(X' = v | X' \in \{v, w\}) \\
=&2 \sum_{\{v,w\} \in F_i} \frac{P(v)P(w)}{P(v) + P(w)} \\
\leq&2 \sum_{\{v,w\} \in F_i} \min\{P(v), P(w)\} \qquad (4) \\
\leq&2 \sum_{(v,w) \in F_i} P(v) \\
=&2 \sum_{v \in W} P(v) \text{ outdegree}(v) \\
\leq&2 \max_{v \in W} \text{ outdegree}(v) \\
\leq&2d
\end{aligned}$$

$\square$

(continued in next lecture...)