Note on Refined Dudley Integral Covering Number Bound

We provide a refined version of Dudley's covering number bound that give tighter bounds for rademacher complexity of function classes with certain covering number bounds. Much of the proof in these notes follow from Peter Bartlett's notes on Covering numbers, chaining and Dudley's integral.

1 Preliminaries

Definition 1. *Let (M, ρ) be a metric space. A subset $\hat{T} \subseteq M$ is called an ϵ cover of $T \subseteq M$ if for every $m \in T$, there exists an $m' \in \hat{T}$ such that $\rho(m, m') \leq \epsilon$. $\hat{T}$ is called a proper cover if $\hat{T} \subset T$. The ϵ covering number of T is the cardinality of the smallest ϵ cover of T, that is*

$$\mathcal{N}(\epsilon, T, \rho) = \min\{|\hat{T}| : \hat{T} \text{ is an } \epsilon \text{ cover of } T\}$$

Let $(\mathcal{F}_{\mathbf{x}_1,...,\mathbf{x}_n}, L_2(P_n))$ stand for the data dependent L_2 metric space given by metric

$$\rho(f, f') = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f(\mathbf{x}_i) - f'(\mathbf{x}_i))^2}$$

where $\mathbf{x}_1, ..., \mathbf{x}_n$ are samples from space $\mathcal{X}$ and $\mathcal{F}_{\mathbf{x}_1,...,\mathbf{x}_n}$ stands for the restriction of function class $\mathcal{F}$ to that sample.

Also let $\hat{R}_n(\mathcal{F})$ be the empirical rademacher complexity of function class $\mathcal{F}$, defined as

$$\hat{R}_n(\mathcal{F}) = \frac{1}{n}\mathbf{E}_\sigma\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}\sigma_i f(\mathbf{x}_i)\right]$$

where $\sigma_1, ..., \sigma_n \in \{-1, +1\}$ are rademacher random variables that have equal probability of being 1 or -1. Also let $R_n(\mathcal{F}) = \mathbb{E}\left[\hat{R}_n(\mathcal{F})\right]$ be the Rademacher complexity of function class $\mathcal{F}$ where the expectation is w.r.t. sample $\mathbf{x}_1, ..., \mathbf{x}_n$. Finally let $B = \sup_{f\in\mathcal{F}}\{\sup_{\mathbf{x}\in\mathcal{X}}|f(x)|\}$. Clearly for any sample $\mathbf{x}_1, ..., \mathbf{x}_n$ we have that $\sup_{f\in\mathcal{F}}\|f\|_{L_2(P_n)} \leq B$

2 Result

Theorem 2. *For any function class $\mathcal{F}$ containing functions $f : \mathcal{X} \mapsto \mathbb{R}$, we have that*

$$\hat{R}_n(\mathcal{F}) \leq \inf_{\epsilon \geq 0} \left\{ 4\epsilon + 12 \int_{\epsilon}^{\sup_{f\in F} \sqrt{\hat{\mathbb{E}}[f^2]}} \sqrt{\frac{\log \mathcal{N}(\tau, \mathcal{F}, L_2(P_n))}{n}} d\tau \right\}$$

Proof. Let $\alpha_0 = \sup_{f\in F} \sqrt{\hat{\mathbb{E}}\left[f^2\right]}$ and for any $j \in \mathbb{Z}_+$ let $\alpha_j = 2^{-j} \sup_{f\in F} \sqrt{\hat{\mathbb{E}}\left[f^2\right]}$. The basic trick here is the idea of chaining. For each j let T_i be a (proper) α_i-cover of $\mathcal{F}$ w.r.t. $L_2(P_n)$. For each $f \in \mathcal{F}$ and j, pick an $\hat{f}_i \in T_i$ such that $\hat{f}_i$ is an α_i approximation of f. Now for any N, we express f by chaining as

$$f = f - \hat{f}_N + \sum_{i=1}^{N} \left(\hat{f}_i - \hat{f}_{i-1} \right)$$

where $\hat{f}_0 = 0$. Hence for any N we have that

$$\begin{aligned}
\hat{R}_n(\mathcal{F}) &= \frac{1}{n}\mathbf{E}_\sigma \left[\sup_{f\in\mathcal{F}} \sum_{i=1}^{n} \sigma_i \left(f(\mathbf{x}_i) - \hat{f}_N(\mathbf{x}_i) + \sum_{j=1}^{N} \left(\hat{f}_j(\mathbf{x}_i) - \hat{f}_{j-1}(\mathbf{x}_i) \right) \right) \right] \\
&\leq \frac{1}{n}\mathbf{E}_\sigma \left[\sup_{f\in\mathcal{F}} \sum_{i=1}^{n} \sigma_i \left(f(\mathbf{x}_i) - \hat{f}_N(\mathbf{x}_i) \right) \right] + \sum_{j=1}^{N} \frac{1}{n}\mathbf{E}_\sigma \left[\sup_{f\in\mathcal{F}} \sum_{i=1}^{n} \sigma_i \left(\hat{f}_j(\mathbf{x}_i) - \hat{f}_{j-1}(\mathbf{x}_i) \right) \right] \\
&\leq \|\sigma\|_{L_2(P_n)} \sup_{f\in\mathcal{F}} \|f - \hat{f}_N\|_{L_2(P_n)} + \sum_{j=1}^{N} \frac{1}{n}\mathbf{E}_\sigma \left[\sup_{f\in\mathcal{F}} \sum_{i=1}^{n} \sigma_i \left(\hat{f}_j(\mathbf{x}_i) - \hat{f}_{j-1}(\mathbf{x}_i) \right) \right] \\
&\leq \alpha_N + \sum_{j=1}^{N} \frac{1}{n}\mathbf{E}_\sigma \left[\sup_{f\in\mathcal{F}} \sum_{i=1}^{n} \sigma_i \left(\hat{f}_j(\mathbf{x}_i) - \hat{f}_{j-1}(\mathbf{x}_i) \right) \right] \qquad (1)
\end{aligned}$$

where the step before last is due to cauchy-shwartz inequality and $\sigma = [\sigma_1, ..., \sigma_n]^\top$. Now note that

$$\begin{aligned}
\|\hat{f}_j - \hat{f}_{j-1}\|^2_{L_2(P_n)} &= \|\hat{f}_j - f + f - \hat{f}_{j-1}\|^2_{L_2(P_n)} \\
&\leq \left(\|\hat{f}_j - f_{L_2(P_n)} + \|f - \hat{f}_{j-1}\|_{L_2(P_n)} \right)^2 \\
&\leq (\alpha_j + \alpha_{j-1})^2 = (\alpha_j + 2\alpha_j)^2 = 3\alpha_j^2
\end{aligned}$$

Now Massart's finite class lemma states that if for any function class $\mathcal{G}$, $\sup_{g\in\mathcal{G}} \|g\|_{L_2(P_n)} \leq R$, then $\hat{R}_n(\mathcal{G}) \leq \sqrt{\frac{2R^2 \log(|\mathcal{G}|)}{n}}$. Applying this to function classes $\{f - f' : f \in T_j m f' \in T_{j-1}\}$ (for each j) we get from Equation 1 that

for any N,

$$\begin{aligned}
\hat{R}_n(\mathcal{F}) &\leq \alpha_N + \sum_{j=1}^{N} 3\alpha_j \sqrt{\frac{2\log(|T_j|\ |T_{j-1}|)}{n}} \\
&\leq \alpha_N + 6\sum_{j=1}^{N} \alpha_j \sqrt{\frac{\log\ |T_j|}{n}} \\
&\leq \alpha_N + 12\sum_{j=1}^{N} (\alpha_j - \alpha_{j+1}) \sqrt{\frac{\log\ |T_j|}{n}} \\
&\leq \alpha_N + 12\sum_{j=1}^{N} (\alpha_j - \alpha_{j+1}) \sqrt{\frac{\log\ \mathcal{N}(\alpha_j, \mathcal{F}, L_2(P_n))}{n}} \\
&\leq \alpha_N + 12\int_{\alpha_{N+1}}^{\alpha_0} \sqrt{\frac{\log\ \mathcal{N}(\tau, \mathcal{F}, L_2(P_n))}{n}} d\tau
\end{aligned}$$

where the third step is because $2(\alpha_j - \alpha_{j+1}) = \alpha_j$. Now for any $\epsilon > 0$, pick $N = \sup\{j : \alpha_j > 2\epsilon\}$. In this case we see that by our choice of N, $\alpha_{N+1} \leq 2\epsilon$ and so $\alpha_N = 2\alpha_{N+1} \leq 4\epsilon$. Also note that since $\alpha_N > 2\epsilon$, $\alpha_{N+1} = \frac{\alpha_N}{2} > \epsilon$. Hence we conclude that

$$\hat{R}_n(\mathcal{F}) \leq 4\epsilon + 12\int_{\epsilon}^{\sup_{f\in F}\sqrt{\hat{\mathbb{E}}[f^2]}} \sqrt{\frac{\log\ \mathcal{N}(\tau, \mathcal{F}, L_2(P_n))}{n}} d\tau$$

Since the choice of ϵ was arbitrary we take an infimum over ϵ. □

Consider the function class bounded by B. Say $\sqrt{\log\ \mathcal{N}(\tau, \mathcal{F}, L_2(P_n))}$ is upper bounded by some analytic function $g_n(\tau)$ and let G_n be the analytic function whose derivative at τ is $g_n(\tau)$ then by FTC we see that

$$\hat{R}_n(\mathcal{F}) \leq \frac{12G_n(B)}{\sqrt{n}} + \inf_{\epsilon}\{4\epsilon - \frac{12G_n(\epsilon)}{\sqrt{n}}\}$$

Note that in natural examples that occur in machine learning, $\log\ \mathcal{N}(\tau, \mathcal{F}, L_2(P_n))$ depends only logarithmically on n and so ignoring the logarithmic factor, assume $g_n(\tau) = g(\tau)$ which is some function that monotonically increases with $\frac{1}{\tau}$. Hence we write the rademacher bound as

$$\hat{R}_n(\mathcal{F}) \leq \mathcal{O}\left(\frac{1}{\sqrt{n}}\right) + \inf_{\epsilon}\{4\epsilon - \frac{12G(\epsilon)}{\sqrt{n}}\}$$

2.1 Examples

Say $\log\ \mathcal{N}(\tau, \mathcal{F}, L_2(P_n)) = \mathcal{O}\left(\frac{1}{\epsilon^p}\right)$, that is $g(\epsilon) = \frac{1}{\epsilon^{p/2}}$. In this case we see that for any $p > 2$, $G(\epsilon) = \frac{2}{(2-p)\epsilon^{p/2-1}}$ and so we see that

$$\inf_{\epsilon}\{4\epsilon - \frac{12\ G(\epsilon)}{\sqrt{n}}\} = \inf_{\epsilon}\{4\epsilon + \frac{24}{\sqrt{n}(2-p)\epsilon^{p/2-1}}\} = \mathcal{O}\left(\frac{1}{n^{1/p}}\right)$$

Hence we can conclude that if $p > 2$,

$$\hat{R}_n(\mathcal{F}) = \mathcal{O}\left(\frac{1}{n^{1/p}}\right)$$

Notice that in this range of p, the original Dudely's integral blows up at 0 (ie $G(0) = \infty$) and hence is unusable. On the other hand if we use the bound

$$\hat{R}_n(\mathcal{F}) = \inf_{\epsilon}\left\{\epsilon + \sqrt{\frac{2\log\ \mathcal{N}(\tau, \mathcal{F}, L_2(P_n))}{n}}\right\} = \inf_{\epsilon}\left\{\epsilon + \frac{1}{\sqrt{n}\epsilon^{p/2}}\right\} = \mathcal{O}\left(\frac{1}{n^{1/(p+2)}}\right)$$

This shows us that the bound we have is qualitatively better when $p > 2$.

Now consider the case when $p < 2$, in this case note that

$$\begin{aligned}
\hat{R}_n(\mathcal{F}) &= \frac{G(B)}{\sqrt{n}} + \inf_{\epsilon}\left\{4\epsilon - \frac{12\ G(\epsilon)}{\sqrt{n}}\right\} \\
&= \frac{24\ B^{1-p/2}}{\sqrt{n}\ (2-p)} + \inf_{\epsilon}\left\{4\epsilon - \frac{24\epsilon^{1-\frac{p}{2}}}{\sqrt{n}(2-p)}\right\} \\
&= \mathcal{O}\left(\frac{1}{\sqrt{n}}\right)
\end{aligned}$$

Here again note that if we use the bound $\hat{R}_n(\mathcal{F}) = \inf_{\epsilon}\{\epsilon + \sqrt{\frac{2\log\ \mathcal{N}(\tau,\mathcal{F},L_2(P_n))}{n}}\}$ we get a qualitatively worse bound.

For the case when $p = 2$, while using bound $\hat{R}_n(\mathcal{F}) = \inf_{\epsilon}\{\epsilon + \sqrt{\frac{2\log\ \mathcal{N}(\tau,\mathcal{F},L_2(P_n))}{n}}\}$ gives a rate of $\mathcal{O}(\frac{1}{n^{1/4}})$ while Our refined dudley integral results in $\mathcal{O}(\sqrt{\frac{\log n}{n}})$. Thus we see that the bound we have is always qualitatively better than using the bound, $\hat{R}_n(\mathcal{F}) = \inf_{\epsilon}\{\epsilon + \sqrt{\frac{2\log\ \mathcal{N}(\tau,\mathcal{F},L_2(P_n))}{n}}\}$. Clearly the bound is also better than Dudley's integral bound.

To summarize, let us use $\mathcal{O}^*$ to hide any factor that is sub-polynomial. Then we basically have that if $\log\ \mathcal{N}(\tau, \mathcal{F}, L_2(P_n)) = \mathcal{O}^*\left(\frac{1}{\epsilon^p}\right)$ then

$$\hat{R}_n = \begin{cases} \mathcal{O}\left(\frac{1}{\sqrt{n}}\right) & p < 2 \\ \mathcal{O}^*\left(\frac{1}{\sqrt{n}}\right) & p = 2 \\ \mathcal{O}^*\left(\frac{1}{n^{1/p}}\right) & p > 2 \end{cases}$$