Notes on Covering Number Bounds

Jon A. Wellner

December 15, 2000

Abstract

We give Dudley's (1978) proof of bounds for $L_r(Q)$ covering numbers for VC-classes, discuss the improvement due to Haussler (1995), and then state bounds on convex hulls of VC-subgraph classes due to Dudley (1987), Ball and Pajor (1990), and Van der Vaart and Wellner (1996).

1. Covering Number Bounds for VC-classes

Recall that for a collection of subsets $\mathcal{C}$ of a set $\mathcal{X}$, and points $x_1, \ldots, x_n \in \mathcal{X}$,

$$\Delta_n^{\mathcal{C}}(x_1, \ldots, x_n) \equiv \#\{C \cap \{x_1, \ldots, x_n\} : \ C \in \mathcal{C}\};$$

so that $\Delta_n^{\mathcal{C}}(x_1, \ldots, x_n)$ is the number of subsets of $\{x_1, \ldots, x_n\}$ picked out by the collection $\mathcal{C}$. Also we define

$$m^{\mathcal{C}}(n) \equiv \max_{x_1,\ldots,x_n} \Delta_n^{\mathcal{C}}(x_1, \ldots, x_n)\,.$$

Let

$$V(\mathcal{C}) \equiv \inf\{n : \ m^{\mathcal{C}}(n) < 2^n\},$$

where the infimum over the empty set is taken to be infinity. Thus $V(\mathcal{C}) = \infty$ if and only if $\mathcal{C}$ shatters sets of arbitrarily large size. A collection $\mathcal{C}$ is called a *VC - class* if $V(\mathcal{C}) < \infty$.

Lemma. (VC - Sauer - Shelah). For a VC - class of sets with VC index $V(\mathcal{C})$, set $S \equiv S(\mathcal{C}) \equiv V(\mathcal{C}) - 1$. Then for $n \geq S$,

$$m^{\mathcal{C}}(n) \leq \sum_{j=0}^{S} \binom{n}{j} \leq \left(\frac{ne}{S}\right)^S . \tag{1.1}$$

Proof. For the first inequality, see Van der Vaart and Wellner (1996), pages 135-136. To see the second inequality, note that with $Y \sim$Binomial$(n, 1/2)$,

$$\begin{aligned}
\sum_{j=0}^{S} \binom{n}{j} &= 2^n \sum_{j=0}^{S} \binom{n}{j} (1/2)^n = 2^n P(Y \leq S) \\
&\leq 2^n E r^{Y-S} \qquad \text{for any} \quad r \leq 1 \\
&= 2^n r^{-S} (\frac{1}{2} + \frac{r}{2})^n \\
&= r^{-S}(1+r)^n \\
&= \left(\frac{n}{S}\right)^S (1 + \frac{S}{n})^n \qquad \text{by choosing} \quad r = S/n \\
&\leq \left(\frac{n}{S}\right)^S e^S ,
\end{aligned}$$

and hence (1.1) holds.

□

Theorem 1. There is a universal constant K such that for any probability measure Q, any VC-class of sets $\mathcal{C}$, and $r \geq 1$, and $0 < \epsilon \leq 1$,

$$N(\epsilon, \mathcal{C}, L_r(Q)) \leq \left(\frac{K \log(K/\epsilon^r)}{\epsilon^r}\right)^{V(\mathcal{C})-1} \leq \left(\frac{K'}{\epsilon}\right)^{r(V(\mathcal{C})-1)+\delta}, \qquad \delta > 0\,; \tag{1.2}$$

here $K = 3e^2/(e-1) \approx 12.9008...$ works.
Moreover,

$$N(\epsilon, \mathcal{C}, L_r(Q)) \leq \widetilde{K} V(\mathcal{C})(4e)^{V(\mathcal{C})} \left(\frac{1}{\epsilon}\right)^{r(V(\mathcal{C})-1)} . \tag{1.3}$$

where $\tilde{K}$ is universal.

The inequality (1.2) is due to Dudley (1978); the inequality (1.3) is due to Haussler (1995). Here we will (re-)prove (1.2), but not (1.3). For the proof of (1.3), see Haussler (1995) or van der Vaart and Wellner (1996), pages 136-140.

Proof. Fix $0 < \epsilon \leq 1$. Let $m = D(\epsilon, \mathcal{C}, L_1(Q))$, the $L_1(Q)$ packing number for the collection $\mathcal{C}$. Thus there exist sets $C_1, \ldots, C_m \in \mathcal{C}$ which satisfy

$$Q(C_i \Delta C_j) = E_Q |1_{C_i} - 1_{C_j}| > \epsilon \qquad \text{for} \quad i \neq j .$$

Let $X_1, \ldots, X_n$ be i.i.d Q. Now C_i and C_j pick out the same subset of $\{X_1, \ldots, X_n\}$ if and only if no $X_k \in C_i \Delta C_j$. If every $C_i \Delta C_j$ contains some X_k, then all C_i's pick out different subsets, and $\mathcal{C}$ picks out at least m subsets from $\{X_1, \ldots, X_n\}$. Thus we compute

$$\begin{aligned}
& Q([X_k \in C_i \Delta C_j \ \text{ for some } k, \ \text{for all } \ i \neq j]^c) \\
&= Q([X_k \notin C_i \Delta C_j \ \text{ for all } k \leq n, \ \text{for some } i \neq j]) \\
&\leq \sum_{i<j} Q([X_k \notin C_i \Delta C_j \ \text{ for all } \ k \leq n]) \\
&\leq \binom{m}{2} \max[1 - Q(C_i \Delta C_j)]^n \\
&\leq \binom{m}{2} (1-\epsilon)^n < 1 \qquad \text{for } \ n \ \text{ large enough} .
\end{aligned} \tag{1.4}$$

In particular this holds if

$$n > \frac{-\log \binom{m}{2}}{\log(1-\epsilon)} = \frac{\log(m(m-1)/2)}{-\log(1-\epsilon)} .$$

Since $-\log(1-\epsilon) < \epsilon$, (1.4) holds if

$$n = \lfloor 3 \log m/\epsilon \rfloor .$$

for this n,

$$Q([X_k \in C_i \Delta C_j \ \text{ for some } k \leq n, \ \text{for all } \ i \neq j]) > 0 .$$

Hence there exist points $X_1(\omega), \ldots, X_n(\omega)$ such that

$$\begin{aligned}
m &\leq \Delta_n^{\mathcal{C}}(X_1(\omega), \ldots, X_n(\omega)) \\
&\leq \max_{x_1, \ldots, x_n} \Delta_n^{\mathcal{C}}(x_1, \ldots, x_n) \\
&\leq \left(\frac{en}{S}\right)^S
\end{aligned} \tag{1.5}$$

where $S \equiv S(\mathcal{C}) \equiv V(\mathcal{C}) - 1$ by the VC - Sauer - Shelah lemma. With $n = \lfloor 3 \log m/\epsilon \rfloor$, (1.5) implies that

$$m \leq \left(\frac{3e \log m}{S\epsilon} \right)^S .$$

Equivalently,

$$\frac{m^{1/S}}{\log m} \leq \frac{3e}{S\epsilon} ,$$

or, with $g(x) \equiv x/\log x$,

$$g(m^{1/S}) \leq \frac{3e}{\epsilon} . \tag{1.6}$$

This implies that

$$m^{1/S} \leq \frac{e}{e-1} \frac{3e}{\epsilon} \log \left(\frac{3e}{\epsilon} \right) , \tag{1.7}$$

or

$$D(\epsilon, \mathcal{C}, L_1(Q)) = m \leq \left\{ \frac{e}{e-1} \frac{3e}{\epsilon} \log \left(\frac{3e}{\epsilon} \right) \right\}^S . \tag{1.8}$$

Since $N(\epsilon, \mathcal{C}, L_1(Q)) \leq D(\epsilon, \mathcal{C}, L_1(Q))$, (1.2) holds for $r = 1$ with $K = 3e^2/(e-1)$.

Here is the argument for (1.6) implies (1.7): note that the inequality

$$g(x) = \frac{x}{\log x} \leq y$$

implies

$$x \leq \frac{e}{e-1} y \log y .$$

To see this, note that $g(x) = x/\log x$ is minimized by $x = e$ and is $\uparrow$. Furthermore $y \geq g(x)$ for $x \geq e$ implies that

$$\log y \geq \log x - \log \log x = \log x \left(1 - \frac{\log \log x}{\log x} \right) > \log x (1 - \frac{1}{e}) ,$$

so

$$x \leq y \log x < y \log y (1 - 1/e)^{-1} .$$

For $L_r(Q)$ with $r > 1$, note that

$$\|1_C - 1_D\|_{L_1(Q)} = Q(C \Delta D) = \|1_C - 1_D\|^r_{L_r(Q)} ,$$

so that

$$N(\epsilon, \mathcal{C}, L_r(Q)) = N(\epsilon^r, \mathcal{C}, L_1(Q)) \leq \left(K \epsilon^{-r} \log \left(\frac{K}{\epsilon^r} \right) \right)^S .$$

This completes the proof.

□

Definition. The *subgraph* of $f : \mathcal{X} \times R$ is the subset of $\mathcal{X} \times R$ given by $\{(x,t) \in \mathcal{X} \times R : \ t < |f(x)|\}$. A collection of functions $\mathcal{F}$ from $\mathcal{X}$ to R is called a *VC - subgraph class* if the collection of subgraphs in $\mathcal{X} \times R$ is a VC -class of sets. For a VC - subgraph class, let $V(\mathcal{F}) \equiv V(\text{subgraph}(\mathcal{F}))$.

Theorem 2. For a VC-subgraph class with envelope function F and $r \geq 1$, and for any probability measure Q with $\|F\|_{L_r(Q)} > 0$,

$$N(2\epsilon\|F\|_{Q,r}, \mathcal{F}, L_r(Q)) \leq KV(\mathcal{F})(16e)^{V(\mathcal{F})} \left(\frac{1}{\epsilon}\right)^{r(V(\mathcal{F})-1)}$$

for a universal constant K and $0 < \epsilon \leq 1$.

Proof. Let $\mathcal{C}$ be the set of all subgraphs C_f of functions $f \in \mathcal{F}$. By Fubini's theorem,

$$Q|f-g| = (Q \times \lambda)(C_f \Delta C_g)$$

where λ is Lebesgue measure on R. Renormalize $Q \times \lambda$ to be a probability measure on $\{(x,t) : |t| \leq F(x)\}$ by defining $P = (Q \times \lambda)/2Q(F)$. Then by the result for sets,

$$N(\epsilon 2Q(F), \mathcal{F}, L_1(Q)) = N(\epsilon, \mathcal{C}, L_1(P)) \leq KV(\mathcal{F}) \left(\frac{4e}{\epsilon}\right)^{V(\mathcal{F})-1}.$$

For $r > 1$, note that

$$Q|f-g|^r \leq Q|f-g|(2F)^{r-1} = 2^{r-1}R|f-g|Q(F^{r-1})$$

for the probability measure R with density $F^{r-1}/Q(F^{r-1})$ with respect to Q. Thus the $L_r(Q)$ distance is bounded by the distance $2(Q(F^{r-1})^{1/r}\|f-g\|_{R,1}^{1/r}$. Elementary manipulations yield

$$\begin{aligned} N(\epsilon 2\|F\|_{Q,r}, \mathcal{F}, L_r(Q)) &\leq N(\epsilon^r RF, \mathcal{F}, L_1(R)) \\ &\leq KV(\mathcal{F}) \left(\frac{8e}{\epsilon^r}\right)^{V\mathcal{F})-1} \end{aligned}$$

by the inequality (1.3).

□

2. Convex Hulls and VC-hull classes

Definition. The *convex hull*, $\text{conv}(\mathcal{F})$ of a class of functions $\mathcal{F}$ is defined as the set of functions $\sum_{i=1}^m \alpha_i f_i$ with $\sum_{i=1}^m \alpha_i \leq 1$, $\alpha_i \geq 0$ and each $f_i \in \mathcal{F}$. The *symmetric convex hull*, denoted by $\text{sconv}(\mathcal{F})$, of a class of functions $\mathcal{F}$ is defined as the set of functions $\sum_{i=1}^m \alpha_i f_i$ with $\sum_{i=1}^m |\alpha_i| \leq 1$ and each $f_i \in \mathcal{F}$. A set of measurable functions $\mathcal{F}$ is a *VC - hull class* if it is in the pointwise sequential closure of the symmetric convex hull of a VC class of of functions, $\mathcal{F} \subset \overline{\text{sconv}}(\mathcal{G})$, $\mathcal{G}$ a VC-class.

Theorem 3. (Dudley, Ball and Pajor). Let Q be a probability mesaure on $(\mathcal{X}, \mathcal{A})$, and let $\mathcal{F}$ be a class of measurable functions with measurable square- integrable envelope F such that $QF^2 < \infty$. and

$$N(\epsilon \|F\|_{Q,2}, \mathcal{F}, L_2(Q)) \leq C \left(\frac{1}{\epsilon}\right)^V , \qquad 0 < \epsilon \leq 1 .$$

Then there is a K depending on C and V only such that

$$\log N(\epsilon \|F\|_{Q,2}, \overline{\text{conv}}(\mathcal{F}), L_2(Q)) \leq K \left(\frac{1}{\epsilon}\right)^{2V/(V+2)} .$$

Note that $2V/(V+2) < 2$ for $V < \infty$. Dudley (1987) proved that for any $\delta > 0$

$$\log N(\epsilon \|F\|_{Q,2}, \overline{\text{conv}}(\mathcal{F}), L_2(Q)) \leq K \left(\frac{1}{\epsilon}\right)^{2V/(V+2)+\delta} .$$

Proof. See Ball and Pajor (1990) or van der Vaart and Wellner (1996), 142 - 145. See also Carl (1997).

□

Example 1. (Monotone functions on R). For $\mathcal{F} = \{1_{[t,\infty)}(x) : \ t \in R\}$, $\mathcal{F}$ is VC, so by Theorem 2, with $F \equiv 1$, $V(\mathcal{F}) = 2$,

$$N(\epsilon, \mathcal{F}, L_2(Q)) \leq K\epsilon^{-2}, \qquad 0 < \epsilon \leq 1.$$

Now

$$\mathcal{G} \equiv \{g : R \to [0,1] | g \nearrow\} \subset \overline{\text{conv}}(\mathcal{F}) .$$

Hence by Theorem 3

$$\log N(\epsilon, \mathcal{G}, L_2(Q)) \leq \frac{K}{\epsilon}, \qquad 0 < \epsilon \leq 1 .$$

Example 2. (Distribution functions on R^d.) For $\mathcal{F} = \{1_{[t,\infty)}(x) : \ t \in R^d\}$, $\mathcal{F}$ is VC with $V(\mathcal{F}) = d + 1$. By Theorem 2 with $F \equiv 1$,

$$N(\epsilon, \mathcal{F}, L_2(Q)) \leq K\epsilon^{-2d}, \qquad 0 < \epsilon \leq 1.$$

Now

$$\mathcal{G} \equiv \{g : R^d \to [0,1] | g \text{ is a d.f. on } \ R^d\} \subset \overline{\text{conv}}(\mathcal{F}) .$$

Hence by Theorem 3

$$\log N(\epsilon, \mathcal{G}, L_2(Q)) \leq K\epsilon^{-2d/(d+1)}, \qquad 0 < \epsilon \leq 1 .$$

In particular, for $d = 2$,

$$\log N(\epsilon, \mathcal{G}, L_2(Q)) \leq K\epsilon^{-4/3}, \qquad 0 < \epsilon \leq 1 .$$

REFERENCES

Ball, K. and Pajor, A. (1990). The entropy of convex bodies with "few" extreme points. *Geometry of Banach spaces, Proceedings of the conference held in Strobl, Austria, 1989*, (eds., P.F.X. Müller and W. Schachermayer). London Mathematical Society Lecture Note Series **158**, 25 - 32.

Carl, B. (1997). Metric entropy of convex hulls in Hilbert space. *Bull. London Math. Soc.* **29**, 452-458.

Dudley, R. M. (1978). Central limit theorems for empirical measures. *Ann. Probability* **6**, 899 - 929.

Dudley, R. M. (1987). Universal Donsker classes and metric entropy. *Ann. Probability* **15**, 1306 - 1326.

Haussler, D. (1995). Sphere packing numbers for subsets of the Boolean $n-$cube with bounded Vapnik-Chervonenkis dimension. *J. Comb. Theory* **A 69**, 217 - 232.

Pisier, G. (1981). Remarques sur un résultat non publié de B. Maurey. *Séminaire d'analyse Fonctionelle, 1980-1981*, Exposé No. 5. École Polytechnique, Palaiseau.

Van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes.* Springer-Verlag, New York.

Department of Statistics
University of Washington
P.O. Box 354322
Seattle, Washington 98195-4322
U.S.A.
e-mail: jaw@stat.washington.edu