Chaining Method to Improve Rademacher Bound

May 11, 2017

1 Introduction

"What must one know a priori about an unknown functional dependency in order to estimate on the basis of observations?" Vapnik stated as the question which sums up the problem which is learning theory. This question is answered by asserting bounds for the number of samples necessary given a desired confidence that a given error is achieved. In the notes by Maxim Raginsky [1], this was often achieved using methods of Rademacher complexity. The paper by F. Cucker and S. Smale [2], "On the Mathematical Foundations of Learning", however, uses the concept of "covering numbers" to establish these bounds. In the Maxim's notes [1] there is mention of a better bound for

$$\mathbb{E}\left[R_n(\mathcal{F}(Z^n))\right] \leq C\sqrt{\frac{V(\mathcal{F})}{n}}, \tag{1}$$

where $\mathcal{F}$ is a space of binary classifiers. But C was not given and neither was the proof. The proof for this bound is attributed to Dudley [3] , and uses what is called the "chaining method" which utilizes concepts of both covering numbers and Rademacher averages. The project goal is to prove this bound and calculate such a value C.

2 Preliminary

A space is sequentially compact if for every sequence there exists a convergent subsequence. A space is compact if for every open cover there exists a finite subcover. In a metric space these two concepts are identical, but not in general. Assume that X is a compact space or a manifold in Euclidean space and that $Y = \mathbb{R}^n$. In applications elements of X are the input data which are to be interpreted while elements of Y are vectors of probabilities associated with the different possible interpretations. Let ρ be a Borel probability measure on the probability space $Z = X \times Y$ whose regularity properties will be assumed as needed. The goal in learning theory is to approximate the function best suited for interpreting our input data.

The only constraint we require on each function $f : X \to Y$ is that
$\int_Z (f(x)-y)d\rho = 0$. Notice that this forms a subspace, call it S, of functions. Since the expectation of $f(x)-y$ is zero our main concern is the variance. Define the error (in the least squares sense) of a function f as

$$\mathcal{E}(f) \triangleq \mathcal{E}_\rho(f) := \int_Z (f(x)-y)^2 d\rho \text{ for } f : X \to Y.$$

Ideally we would like to have the minimizer

$$\hat{f} = \underset{f \in S}{\operatorname{argmin}} \int_Z (f(x)-y)^2 d\rho \text{ for } f : X \to Y.$$

To that end consider the conditional probability measure $\rho(y|x)$ on Y for given $x \in X$, and the marginal probability measure ρ_X on X. That is $\rho_X(S) = \rho(\pi^{-1}(S))$ where π is simply the projection map $\pi : X \times Y \to$

X. Therefore for any integrable function $\varphi : X \times Y \to \mathbb{R}$,

$$\int_{X\times Y} \varphi(x,y)d\rho = \int_X \left(\int_Y \varphi(x,y)d\rho(y|x) \right) d\rho_X,$$

by Fubini's Theorem. Now let us define the *regression function* $f_\rho : X \to \mathbb{R}$ by

$$f_\rho(x) \triangleq \int_Y y d\rho(y|x).$$

Any regularity hypothesis on ρ will induce regularity properties on f_ρ. Henceforth, assume f_ρ is a bounded function. It is clear that $f_\rho \in S$, by definition and Fubini's Theorem. An important constant in this paper is

$$\sigma_\rho^2 \triangleq \int_X \left(\int_Y (f_\rho - y)^2 d\rho(y|x) \right) d\rho_X \text{ and so } \mathcal{E}(f_\rho) = \sigma_\rho^2,$$

the importance of σ_ρ^2 can be seen in Proposition 3.1.

3 Empirical Error

Proposition 3.1. *For every $f : X \to Y$,*

$$\mathcal{E}(f) = \int_X (f(x) - f_\rho(x))^2 + \sigma_\rho^2.$$

Proof of Proposition 3.1.

$$\begin{aligned}
\mathcal{E}(f) &= \int_Z (f(x) - f_\rho(x) + f_\rho(x) - y)^2 \\
&= \int_Z (f(x) - f_\rho(x))^2 + 2\int_Z (f(x) - f_\rho(x))(f_\rho(x) - y) + \int_Z (f_\rho(x) - y)^2 \\
&= \int_X (f(x) - f_\rho(x))^2 + 2\int_Y (f_\rho(x) - y)\left(\int_X (f(x) - f_\rho(x)) \right) + \sigma_\rho^2 \\
&= \int_X (f(x) - f_\rho(x))^2 + 2\int_X (f(x) - f_\rho(x)) \cdot 0 + \sigma_\rho^2 \\
&= \int_X (f(x) - f_\rho(x))^2 + \sigma_\rho^2.
\end{aligned}$$

By Fubini's Theorem and the definitions of f_ρ and σ_ρ. □

Thus, by Proposition 3.1, it is clear that the goal is to find a good approximation of f_ρ by taking random samples on X instead of Z. By Proposition 3.1 the minimum of $\mathbb{E}(f)$ is our constant σ_ρ^2, because $\int_X (f(x) - f_\rho(x))^2 \geq 0$ with equality at f_ρ. σ_ρ^2 can be thought of as a conditioning number of ρ [2], and $\sigma_\rho^2 = 0$ is perfect conditioning. Consider a sample $z \in Z^m$, and denote the *empirical error of* f w.r.t. z as

$$\mathcal{E}_z(f) \triangleq \frac{1}{m}\sum_{i=1}^{m} (f(x_i) - y_i)^2, \text{ with } z = ((x_1, y_1), ..., (x_m, y_m)).$$

The empirical error of f is, as a matter of fact, the most important error in this paper, because it will tell us how well our function is fitted to an actual sample set. Denote $f_z = f_{z,\mathcal{H}}$ as the minimizer of $\mathcal{E}_z(f)$. Cucker and Smale's first main result is to show that $\mathcal{E}_z(f)$ can be approximated by $\mathcal{E}(f)$ for a large enough number of samples. Define the *defect function* of f as $L_z(f) \triangleq \mathcal{E}(f) - \mathcal{E}_z(f)$. In class we referred to the uniform deviation, $\Delta_n(Z^n) := \sup_{f\in\mathcal{H}} |L_z(f)|$. Before continuing we first need a couple of inequalities.

Proposition 3.2. *Let ξ be a random variable on a probability space Z with $\mathbb{E}(\xi) = \mu$ and variance $\mathrm{Var}(\xi) = \sigma^2$.* [Chebyshev] *For all $\epsilon > 0$,*

$$\operatorname*{Prob}_{z\in Z^m}\left\{\left|\frac{1}{m}\sum_{i=1}^{m}\xi(z_i) - \mu\right| \geq \epsilon\right\} \leq \frac{\sigma^2}{m\epsilon^2}.$$

[Bernstein] *If $|\xi(z) - \mu| \leq M$ for almost all $z \in Z$, then for all $\epsilon > 0$,*

$$\operatorname*{Prob}_{z\in Z^m}\left\{\left|\frac{1}{m}\sum_{i=1}^{m}\xi(z_i) - \mu\right| \geq \epsilon\right\} \leq 2e^{-\frac{m\epsilon^2}{2(\sigma^2+\frac{1}{3}M^2\epsilon)}}.$$

For Chebyshev's inequality just use Jensen's inequality and the convexity of the square function. For Bernstein's inequality see [4]. For any function $f : X \to Y$ denote by f_Y the function

$$\begin{aligned} f_Y : X \times Y &\to Y \\ (x, y) &\mapsto f(x) - y. \end{aligned}$$

Theorem 3.3. *Let $M > 0$ and $f : X \to Y$ be such that $|f(x) - y| \leq M$ almost everywhere. Then, for any $\epsilon > 0$,*

$$\operatorname*{Prob}_{z\in Z^m}\{|L_z(f)| \leq \epsilon\} \geq 1 - 2e^{-\frac{m\epsilon^2}{2(\sigma^2+\frac{1}{3}M^2\epsilon)}} \tag{2}$$

where σ^2 is the variance of f_Y^2.

Proof of Theorem 3.3. Use Bernstein's inequality for $\xi = f_Y^2$. □

Note that $m \geq \frac{2(\sigma^2+\frac{1}{3}M^2\epsilon)}{\epsilon^2}$ is sufficient for the right hand side of equation 2 to be less than 1.

4 Hypothesis Space and Target Functions

Consider the Banach space $(C(X), \|\cdot\|_\infty)$, of continuous functions on X so that

$$\|f\|_\infty = \sup_{x\in X}|f(x)|.$$

Let $\mathcal{H}$ be a compact (and often convex) subset of $\mathcal{C}(X)$, which we will call our hypothesis space. Here we will approximate f_ρ as best as possible, but must first define two other functions. Define our target function as

$$f_{\mathcal{H}} \triangleq \operatorname*{argmin}_{f\in\mathcal{H}} \mathcal{E}(f)$$

By Proposition 3.1, $f_{\mathcal{H}} = \operatorname{argmin}_{f\in\mathcal{H}} \int_X (f - f_\rho)^2$, so $f_{\mathcal{H}}$ is the closest function in $\mathcal{H}$ to f_ρ, in the least squares sense. Notice that $\mathcal{C}(X)$ is a Banach space not an IPS, and that $\mathcal{H}$ is compact and thus complete (compact sets are complete) but is not necessarily a subspace. Thus, Theorem 7.5 from the course notes, see [2], does not apply. Yet, we still have existence. Since $\mathcal{H}$ is compact and $\mathcal{E} : \mathcal{C}(X) \to \mathbb{R}$ is continuous, its image is compact thus $f_{\mathcal{H}}$ exists. Denote the *error in $\mathcal{H}$* of a function $f \in \mathcal{H}$ as the normalized error

$$\mathcal{E}_{\mathcal{H}}(f) \triangleq \mathcal{E}(f) - \mathcal{E}(f_{\mathcal{H}}) = \int_X (f - f_\rho)^2 - \int_X (f_{\mathcal{H}} - f_\rho)^2.$$

Theorem 3.3 allows us to approximate $\mathcal{E}_z(f)$ with $\mathcal{E}(f)$, and with Proposition 3.1,

$$\mathcal{E}_{\mathcal{H}}(f_z) + \mathcal{E}(f_{\mathcal{H}}) = \int_X (f_z - f_\rho)^2 + \sigma_\rho^2,$$

where the first error term $\mathcal{E}_{\mathcal{H}}(f_z)$ is called the *sample error* and takes functions only from $\mathcal{H}$ while the second error term $\mathcal{E}(f_{\mathcal{H}})$ is called the *approximation error* and is independent of sampling altogether. Notice that for fixed sample size m, if we were to enlarge $\mathcal{H}$ then our approximation error will decrease while our sampling error will increase, this is called "bias-variance" trade off. Let us introduce the notion of covering numbers, where $\mathcal{N}(r, U)$ is the minimal number of balls of radius r which cover a set U. This number is always finite for compact sets like $\mathcal{H}$.

5 Using Covering Numbers to Bound Uniform Deviation

Proposition 5.1. *If $|f_j(x) - y| \leq M$ on a set $U \subset Z$ of full measure for $j = 1, 2$, then for $\mathbf{z} \in U^m$*

$$|L_{\mathbf{z}}(f_1) - L_{\mathbf{z}}(f_2)| \leq 4M\|f_1 - f_2\|_\infty$$

Proof. First notice that

$$\begin{aligned}|(f_1(x) - y)^2 - (f_2(x) - y)^2| &= |(f_1(x) - f_2(x))(f_1(x) + f_2(x) - 2y)| \\ &\leq \|f_1 - f_2\|_\infty |(f_1(x) + f_2(x) - 2y)| \\ &\leq 2M\|f_1 - f_2\|_\infty\end{aligned}$$

$$\begin{aligned}|L_{\mathbf{z}}(f_1) - L_{\mathbf{z}}(f_2)| &\leq |\,\mathcal{E}(f_1) - \mathcal{E}_{\mathbf{z}}(f_1) - \mathcal{E}(f_2) + \mathcal{E}_{\mathbf{z}}(f_2)| \\ &\leq \left|\int_Z (f_1(x) - y)^2 - (f_2(x) - y)^2 d\rho\right| + \left|\frac{1}{m}\sum_{i=1}^{m}(f_1(x_i) - y_i)^2 - (f_2(x_i) - y_i)^2\right| \\ &\leq 4M\|f_1 - f_2\|_\infty\end{aligned}$$

□

Lemma 5.2. *Let $\mathcal{H} = S_1 \cup ... \cup S_l$ and $\epsilon > 0$. Then*

$$\operatorname*{Prob}_{\mathbf{z}\in Z^m}\left\{\sup_{f\in\mathcal{H}}|L_{\mathbf{z}}(f)| \geq \epsilon\right\} \leq \sum_{j=1}^{l}\operatorname*{Prob}_{\mathbf{z}\in Z^m}\left\{\sup_{f\in S_j}|L_{\mathbf{z}}(f)| \geq \epsilon\right\}$$

Proof.

$$\begin{aligned}\operatorname*{Prob}_{\mathbf{z}\in Z^m}\left\{\sup_{f\in\mathcal{H}}|L_{\mathbf{z}}(f)| \geq \epsilon\right\} &= \operatorname*{Prob}_{\mathbf{z}\in Z^m}\left\{\bigcup_{j=1}^{l}\left\{\sup_{f\in S_j}|L_{\mathbf{z}}(f)| \geq \epsilon\right\}\right\} \\ &\leq \sum_{j=1}^{l}\operatorname*{Prob}_{\mathbf{z}\in Z^m}\left\{\sup_{f\in S_j}|L_{\mathbf{z}}(f)| \geq \epsilon\right\}\end{aligned}$$

□

Theorem 5.3. *Let $\mathcal{H}$ be a compact subset of $\mathcal{C}(X)$. Assume that, for all $f \in \mathcal{H}$, $|f(x) - y| \leq M$ almost everywhere. Then, for all $\epsilon > 0$,*

$$\operatorname*{Prob}_{\mathrm{z}\in\mathrm{Z}^{\mathrm{m}}}\{\Delta_{\mathrm{n}}(\mathrm{Z}^{\mathrm{n}}) \leq \epsilon\} \geq 1 - \mathcal{N}\left(\frac{\epsilon}{8M}, \mathcal{H}\right) 2e^{-\frac{m\epsilon^2}{4(2\sigma^2 + \frac{1}{3}M^2\epsilon)}}.$$

Here $\sigma^2 = \sup_{f\in\mathcal{H}}\sigma^2(f_Y^2)$.

Proof. Let $l := \mathcal{N}(\frac{\epsilon}{8M}, \mathcal{H})$, then there exists $f_1, ..., f_l \in \mathcal{H}$ so that the disks, D_i, centered at f_i with radius $\frac{\epsilon}{8M}$ cover $\mathcal{H}$. Let U be any set of probability 1 and $|f(x) - y| \leq M$. So by proposition 5.1

$$|L_{\mathbf{z}}(f) - L_{\mathbf{z}}(f_j)| \leq 4M\|f - f_j\|_\infty \leq 4M\frac{\epsilon}{8M} = \epsilon/2.$$

Thus we have that for any $f \in D_j$ and $\mathbf{z} \in U^m$,

$$\sup_{f \in D_j} |L_{\mathbf{z}}(f)| \geq \epsilon \Rightarrow |L_{\mathbf{z}}(f_j)| \geq \epsilon/2.$$

So

$$\begin{aligned}\operatorname*{Prob}_{\mathbf{z}\in Z^m}\left\{\sup_{f\in D_j}|L_{\mathbf{z}}(f)| \geq \epsilon\right\} &\leq \operatorname*{Prob}_{\mathbf{z}\in Z^m}\{|L_{\mathbf{z}}(f_j)| \geq \epsilon/2\} \\ &\leq 2\exp\frac{-m\epsilon^2}{4(2\sigma^2(f_{jY}^2) + \frac{1}{3}M^2\epsilon)}\end{aligned}$$

Now applying lemma 5.2 we have

$$\operatorname*{Prob}_{\mathrm{z}\in\mathrm{Z}^{\mathrm{m}}}\left\{\sup_{f\in\mathcal{H}}|L_z(f)| \leq \epsilon\right\} \geq 1 - \mathcal{N}\left(\frac{\epsilon}{8M}, \mathcal{H}\right)2e^{-\frac{m\epsilon^2}{4(2\sigma^2+\frac{1}{3}M^2\epsilon)}}.$$

□

Lemma 5.4. *Let $\mathcal{H}$ be a compact subset of C(X). Let $\epsilon > 0$ and $0 < \delta < 1$ so that*

$$\operatorname*{Prob}_{\mathrm{z}\in\mathrm{Z}^{\mathrm{m}}}\left\{\sup_{f\in\mathcal{H}}|L_z(f)| \leq \epsilon\right\} \geq 1 - \delta.$$

Then

$$\operatorname*{Prob}_{\mathrm{z}\in\mathrm{Z}^{\mathrm{m}}}\{\mathcal{E}_{\mathcal{H}} \leq 2\epsilon\} \geq 1 - \delta.$$

Proof of Lemma 5.4. Then an event exists with probability $1 - \delta$, so that both

$$\mathcal{E}(f_z) \leq \mathcal{E}_{\mathrm{z}}(f_z) + \epsilon \text{ and } \mathcal{E}_{\mathrm{z}}(f_{\mathcal{H}}) \leq \mathcal{E}(f_{\mathcal{H}}) + \epsilon.$$

Now since f_z minimizes $\mathcal{E}_z$ on $\mathcal{H}$,

$$\mathcal{E}_{\mathrm{z}}(f_z) \leq \mathcal{E}_{\mathrm{z}}(f_{\mathcal{H}}).$$

Thus on the aforementioned event,

$$\mathcal{E}(f_z) \leq \mathcal{E}_{\mathrm{z}}(f_z) + \epsilon \leq \mathcal{E}_{\mathrm{z}}(f_{\mathcal{H}}) + \epsilon \leq \mathcal{E}(f_{\mathcal{H}}) + 2\epsilon$$

or

$$\mathcal{E}_{\mathcal{H}}(f_z) \leq \mathcal{E}(f_z) - \mathcal{E}(f_{\mathcal{H}}) \leq 2\epsilon.$$

□

Theorem 5.5. *Let $\mathcal{H}$ be a compact subset of C(X). Assume that, for all $f \in \mathcal{H}$, $|f(x) - y| \leq M$ almost everywhere. Let $\sigma^2 = \sup_{f\in\mathcal{H}}\sigma^2(f_Y^2)$. Then, for all $\epsilon > 0$,*

$$\operatorname*{Prob}_{\mathrm{z}\in\mathrm{Z}^{\mathrm{m}}}\{\mathcal{E}_{\mathcal{H}}(f_z) \leq \epsilon\} \geq 1 - \mathcal{N}\left(\frac{\epsilon}{16M}, \mathcal{H}\right)2e^{-\frac{m\epsilon^2}{8(4\sigma^2+\frac{1}{3}M^2\epsilon)}}.$$

Proof of Theorem 5.5. Replace ϵ with $\epsilon/2$, then use Lemma 5.4 and Theorem 5.3. □

A set S is convex if for any $u, v \in S$ and $\alpha \in [0,1]$, $\alpha u + (1-\alpha)v \in S$.

Lemma 5.6. *Let $\mathcal{H}$ be a convex subset of $\mathcal{C}(X)$ such that $f_\mathcal{H}$ exists. Then $f_\mathcal{H}$ is unique as an element in $\mathcal{L}_p^2(X)$ and, for all $f \in \mathcal{H}$,*

$$\int_X (f_\mathcal{H} - f)^2 \leq \mathcal{E}_\mathcal{H}(f).$$

Proof. Consider the line segment $\overline{f_\mathcal{H} f}$, for some fixed $f \in \mathcal{H}$. Since $\mathcal{H}$ is convex $\overline{f_\mathcal{H} f} \subset \mathcal{H}$. So $\forall g \in \overline{f_\mathcal{H} f}$, $g \in \mathcal{H}$ and thus $\|f_\mathcal{H} - f_\rho\|_\rho \leq \|g - f_\rho\|_\rho$. This gives that $\widehat{f_\rho f_\mathcal{H} f}$ is obtuse, which implies

$$\|f_\mathcal{H} - f\|_\rho^2 + \|f_\mathcal{H} - f_\rho\|_\rho^2 \leq \|f - f_\rho\|_\rho^2.$$

Thus $\int_X (f_\mathcal{H} - f)^2 \leq \mathbb{E}(f) - \mathbb{E}(f_\mathcal{H})$. ✓
Suppose we have two minimizers f' and f''. Then

$$\|f' - f''\|_\rho^2 + \|f' - f_\rho\|_\rho^2 \leq \|f'' - f_\rho\|_\rho^2 \text{ and}$$
$$\|f'' - f'\|_\rho^2 + \|f'' - f_\rho\|_\rho^2 \leq \|f' - f_\rho\|_\rho^2,$$

which implies that $f' = f''$. ✓ □

Proposition 5.7. *For all $\epsilon > 0$ and $0 < \alpha \leq 1$,*

$$\operatorname*{Prob}_{\mathbf{z} \in Z^m} \left\{ \sup_{f \in \mathcal{H}} \frac{\mathcal{E}_\mathcal{H}(f) - \mathcal{E}_{\mathcal{H},z}(f)}{\mathcal{E}_\mathcal{H}(f) + \epsilon} \geq 3\alpha \right\} \leq \mathcal{N}\left(\frac{\alpha\epsilon}{4M}, \mathcal{H}\right) e^{-\frac{\alpha^2 \epsilon m}{8M^2}}.$$

Theorem 5.8. *Let $\mathcal{H}$ be a compact and convex subset of $\mathcal{C}(X)$. Assume that, for all $f \in \mathcal{H}$, $|f(x) - y| \leq M$ almost everywhere. Then, for all $\epsilon > 0$,*

$$\operatorname*{Prob}_{\mathbf{z} \in Z^m} \{\mathcal{E}_\mathcal{H}(f_z) \leq \epsilon\} \geq 1 - \mathcal{N}\left(\frac{\epsilon}{24M}, \mathcal{H}\right) e^{-\frac{\epsilon m}{288M^2}}.$$

Theorems 5.5 and 5.8 show that our *sample error* $\mathcal{E}_\mathcal{H}(f_z)$ converges to 0 exponentially in probability with respect to the number of samples. Thus, our choice on $\mathcal{H}$ is essential in learning theory, because if $\mathcal{H}$ is relatively large then the covering number will be large so that $\mathcal{E}_\mathcal{H}(f_z)$ is bounded by ϵ on an event with unacceptably small probability, this is called overfitting. But it will cost us many more samples to rectify this error. On the other hand if we choose a nice, simple hypothesis space, $\mathcal{H}$, then our sample error may be low (even with relatively few samples) but our approximation error may still be too large. This is a problem of regression also known as "bias-variance" problem.

6 Chaining Method

In this section we will use the "chaining method" to improve the bound of $R_n(\mathcal{F}(Z^n))$. We will use a more general notion of covering number.

Definition 6.1. *We say that $V \subset \mathcal{R}^T$ is an α-cover on x^T with respect to $\|\cdot\|_p$ with $1 \leq p \leq \infty$, if $\forall f \in \mathcal{F}$, $\exists \boldsymbol{v} \in V$ so that $\left(\frac{1}{T}\sum_{i=1}^T (f(x_i) - v_i)^p\right)^{1/p} \leq \alpha$ or $\frac{1}{\sqrt[p]{T}}\|f(x^T) - v\|_p \leq \alpha$.*

The α-cover number on x^T with respect to $\|\cdot\|_p$ is

$$\mathcal{N}_p(\alpha, \mathcal{F}, x^T) := \min\{|V| : V \text{ } \alpha\text{-covers } x^T \text{ with respect to } \|\cdot\|_p\}.$$

The following theorem was proved, using the Sauer-Shelah Lemma, by Mendelson [5] (page 14, his Theorem 2.14).

Theorem 6.2. *Given a class $\mathcal{F}$ of $\{0,1\}$- valued functions, x^T, $\alpha > 0$, and VC-dimension $V(\mathcal{F}) = d$ then*

$$\mathcal{N}_2(\alpha, \mathcal{F}, x^T) \leq \left((4e^2)\log\left(\frac{2e^2}{\alpha}\right)\right)^d \left(\frac{1}{\alpha}\right)^{2d}.$$

The following is (essentially) Dudley's Chaining Theorem [3].

Theorem 6.3. *Suppose that $\mathcal{F}(Z^n) \subset \mathbb{R}^n$. Then*

$$R_n(\mathcal{F}(Z^n)) \leq 12\sqrt{2}\int_0^\infty \sqrt{\frac{\log \mathcal{N}_2(\alpha, \mathcal{F}, x^n)}{n}}d\alpha$$

Proof. Let

$$B := \sup_{f\in\mathcal{F}(Z^n)} \sqrt{\frac{1}{n}\sum_{i=1}^n f_i^2},$$

and $\alpha_j := 2^{-j}B$ for $j \in \mathbb{N}$, so $\alpha_i \to 0$ as $i \to \infty$. Let T_j be a minimal α_j-cover of x^n in the 2-norm. Let $T_0 = \{0\}$ which is a α_0-cover of x^n for $\alpha_0 = B$. Denote the element $c_j(f) \in T_j$ to be the closest element in T_j to f. This means that $\|f - c_j(f)\|_2 \leq \sqrt{n}\alpha_j$, for each j.

The first step is to use telescoping. Notice that $f = f - c_N(f) + \sum_{j=1}^N (c_j(f) - c_{j-1}(f))$, since $c_0(f) = 0$. Now

$$\begin{aligned}
R_n(\mathcal{F}(Z^n)) &= \mathbb{E}\left[\sup_{f\in\mathcal{F}(Z^n)}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i f_i\right|\right], \text{Rademacher definition,}\\
&= \mathbb{E}\left[\sup_{f\in\mathcal{F}(Z^n)}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i\left(f_i - (c_N(f))_i + \sum_{j=1}^N((c_j(f))_i - (c_{j-1}(f))_i)\right)\right|\right], \text{telescoping,}\\
&\leq \mathbb{E}\left[\sup_{f\in\mathcal{F}(Z^n)}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i\left(f_i - (c_N(f))_i\right)\right|\right] + \mathbb{E}\left[\sup_{f\in\mathcal{F}(Z^n)}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i\left(\sum_{j=1}^N((c_j(f))_i - (c_{j-1}(f))_i)\right)\right|\right]\\
&\leq \mathbb{E}\left[\sup_{f\in\mathcal{F}(Z^n)}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i\left(f_i - (c_N(f))_i\right)\right|\right] + \sum_{j=1}^N \mathbb{E}\left[\sup_{f\in\mathcal{F}(Z^n)}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i((c_j(f))_i - (c_{j-1}(f))_i)\right|\right]\\
&\leq \alpha_N + \sum_{j=1}^N \mathbb{E}\left[\sup_{f\in\mathcal{F}(Z^n)}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i((c_j(f))_i - (c_{j-1}(f))_i)\right|\right], \text{ by Cauchy-Schwarz.}
\end{aligned}$$

By the Finite Class Lemma, the triangle inequality, and $|T_j| \geq |T_{j-1}|$, we have

$$\begin{aligned}
\mathbb{E}\left[\sup_{f\in\mathcal{F}(Z^n)}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i((c_j(f))_i - (c_{j-1}(f))_i)\right|\right] &\leq 2\frac{\sup_{f\in\mathcal{F}(Z^n)}\|c_j(f) - c_{j-1}(f)\|_2\sqrt{\log|T_j||T_{j-1}|}}{n}, F.C.L.\\
&\leq 2\frac{\sup_{f\in\mathcal{F}(Z^n)}(\|c_j(f) - f\|_2 + \|f - c_{j-1}(f)\|_2)\sqrt{\log|T_j||T_{j-1}|}}{n}\\
&\leq 2\frac{\sqrt{n}(\alpha_j + \alpha_{j-1})\sqrt{\log|T_j||T_{j-1}|}}{n}, \text{definition } \alpha_j\text{-cover,}\\
&\leq 6\frac{\alpha_j\sqrt{2\log|T_j|}}{\sqrt{n}}, \text{definition } \alpha_j \text{ and } |T_j| \geq |T_{j-1}|,\\
&= 6\sqrt{2}\frac{\alpha_j\sqrt{\log|T_j|}}{\sqrt{n}}.
\end{aligned}$$

So now we have

$$\begin{aligned}
R_n(\mathcal{F}(Z^n)) &\leq \alpha_N + \sum_{j=1}^{N} \mathbb{E}\left[\sup_{f\in\mathcal{F}(Z^n)}\left(\frac{1}{n}\sum_{i=1}^{n}\sigma_i((c_j(f))_i - (c_{j-1}(f))_i)\right)\right] \\
&\leq \alpha_N + 6\sqrt{2}\sum_{j=1}^{N}\frac{\alpha_j\sqrt{\log|T_j|}}{\sqrt{n}}, \text{shown,} \\
&= \alpha_N + 6\sqrt{2}\sum_{j=1}^{N}\alpha_j\frac{\sqrt{\log\mathcal{N}_2(\alpha_j,\mathcal{F},x^n)}}{\sqrt{n}}, \text{definition } T_j, \\
&= \alpha_N + 12\sqrt{2}\sum_{j=1}^{N}(\alpha_j-\alpha_{j+1})\frac{\sqrt{\log\mathcal{N}_2(\alpha_j,\mathcal{F},x^n)}}{\sqrt{n}}, \text{since } \alpha_j = 2(\alpha_j-\alpha_{j+1}), \\
&\leq \alpha_N + 12\sqrt{2}\int_{\alpha_{N+1}}^{\alpha_1}\frac{\sqrt{\log\mathcal{N}_2(\alpha,\mathcal{F},x^n)}}{\sqrt{n}}d\alpha, \text{since } \mathcal{N}_2(\alpha,\mathcal{F},x^n) \text{ is decreasing in } \alpha, \\
&\leq \alpha_N + 12\sqrt{2}\int_{\alpha_{N+1}}^{\infty}\sqrt{\frac{\log\mathcal{N}_2(\alpha,\mathcal{F},x^n)}{n}}d\alpha, \text{positivity of integrand,}
\end{aligned}$$

for any N. Now letting N approach ∞, giving us

$$R_n(\mathcal{F}(Z^n)) \leq 12\sqrt{2}\int_0^{\infty}\sqrt{\frac{\log\mathcal{N}_2(\alpha,\mathcal{F},x^n)}{n}}d\alpha.$$

□

Corollary 6.4. *If $\mathcal{F}$ is a VC class of binary functions, then*

$$R_n(\mathcal{F}(Z^n)) \leq 160\sqrt{\frac{V(\mathcal{F})}{n}}$$

Proof. For binary functions we have that

$$\sup_{f\in\mathcal{F}(Z^n)}\sqrt{\frac{1}{n}\sum_{i=1}^{n}f_i^2} = 1.$$

Recall theorem (6.2) where $\mathcal{N}_2(\alpha,\mathcal{F},x^n) \leq \left((4e^2)\log\left(\frac{2e^2}{\alpha}\right)\right)^{V(\mathcal{F})}\left(\frac{1}{\alpha}\right)^{2V(\mathcal{F})}$ and notice $\mathcal{N}_2(\alpha,\mathcal{F},x^n) = 1$ for

$\alpha > 1$. So by theorem 6.3, we have that

$$
\begin{aligned}
R_n(\mathcal{F}(Z^n)) &\leq 12\sqrt{2}\int_0^\infty \sqrt{\frac{\log \mathcal{N}_2(\alpha, \mathcal{F}, x^T)}{n}}d\alpha, \text{theorem 6.2,} \\
&\leq 12\sqrt{2}\int_0^1 \sqrt{\frac{\log[\left((4e^2)\log\left(\frac{2e^2}{\alpha}\right)\right)^{V(\mathcal{F})}\left(\frac{1}{\alpha}\right)^{2V(\mathcal{F})}]}{n}}d\alpha, \text{theorem 6.3,} \\
&= 12\sqrt{2}\sqrt{\frac{V(\mathcal{F})}{n}}\int_0^1 \sqrt{\log[\left((4e^2)\log\left(\frac{2e^2}{\alpha}\right)\right)\left(\frac{1}{\alpha}\right)^2]}d\alpha, \text{log power property,} \\
&\overset{*}{\leq} 12\sqrt{2}\sqrt{\frac{V(\mathcal{F})}{n}}\int_0^1 \sqrt{\log[(4e^2\log\left(2e^2\right))\left(\frac{1}{\alpha}\right)^3]}d\alpha, \text{since } \log\left[\tfrac{2e^2}{\alpha}\right] \leq \log\left[2e^2\right]\left(\tfrac{1}{\alpha}\right), \\
&= 12\sqrt{6}\sqrt{\frac{V(\mathcal{F})}{n}}\int_0^1 \sqrt{\log[(4e^2\log(2e^2))^{1/3}\left(\frac{1}{\alpha}\right)]}d\alpha, \text{log power property,} \\
&= 12\sqrt{6}\left(4e^2\log(2e^2)\right)^{1/3}\sqrt{\frac{V(\mathcal{F})}{n}}\int_0^{1/\left(4e^2\log(2e^2)\right)^{1/3}} \sqrt{\log\left(\frac{1}{\beta}\right)}d\beta, \text{with } \beta = \alpha/\left(4e^2\log(2e^2)\right)^{1/3}, \\
&< 46\sqrt{\frac{V(\mathcal{F})}{n}}.
\end{aligned}
$$

□

Much of my work came from work by Rakhlin [6], and notes by a Dr. Kakade. If you integrate directly at the $*$ inequality above, you still do no better then 43. So 46 is a good bound.

References

[1] M. Raginsky, *Statistical Learning Theory*. UIUC, 2017.

[2] F. Cucker and S. Smale, "On the mathematical foundations of learning," *American Mathematical Society*, vol. 39, no. 1, pp. 1–49, 2002.

[3] R. M. Dudley, "Central limit theorems for empirical measures," *The Annals of Probability*, pp. 899–929, 1978.

[4] D. Pollard, *Convergence of stochastic processes*. Springer Science & Business Media, 2012.

[5] S. Mendelson, "A few notes on statistical learning theory," in *Advanced lectures on machine learning*, pp. 1–40, Springer, 2003.

[6] A. Rakhlin and K. Sridharan, "Statistical learning theory and sequential prediction," *Lecture Notes in University of Pennsyvania*, 2012.