ONE DIMENSIONAL CONCENTRATION INEQUALITIES*

11.1. Law of Large Numbers

In this lecture, we will look at concentration inequalities or law of large numbers for a fixed function. Let $(\Omega, \mathcal{L}, \mu)$ be a probability space. Let $x_1, ..., x_n$ be real random variables on Ω. A sequence of random variables y_n converges almost surely to a random variable Y iff $\mathbb{P}(y_n \to Y) = 1$. A sequence of random variables y_n converges in probability to a random variable Y iff for every $\epsilon > 0$, $\lim_{n\to\infty} \mathbb{P}(|y_n - Y| > \epsilon) = 0$. Let $\hat{\mu}_n := n^{-1} \sum_{i=1}^n x_n$. The sequence $x_1, ..., x_n$ satisfies the strong law of large numbers if for some constant c, $\hat{\mu}_n$ converges to c almost surely. The sequence $x_1, ..., x_n$ satisfies the weak law of large numbers iff for some constant c, $\hat{\mu}_n$ converges to c in probability. In general the constant c will be the expectation of the random variable $\mathbb{E}x$.

A given function $f(x)$ of random variables x concentrates if the deviation between its empirical average, $n^{-1} \sum_{i=1}^n f(x_i)$ and expectation, $\mathbb{E}f(x)$, goes to zero as n goes to infinity. That is $f(x)$ satisfies the law of large numbers.

11.2. Polynomial inequalities

Theorem (Jensen)**.** *If ϕ is a convex function then $\phi(\mathbb{E}x) \leq \mathbb{E}\phi(x)$.*

Theorem (Bienaymé-Chebyshev)**.** *For any random variable x, $\epsilon > 0$*

$$\mathbb{P}(|x| \geq \epsilon) \leq \frac{\mathbb{E}x^2}{\epsilon^2}.$$

Proof.

$$\mathbb{E}x^2 \geq E(x^2 I_{\{|x|\geq\epsilon\}}) \geq \epsilon^2 \mathbb{P}(|x| > \epsilon). \ \square$$

Theorem (Markov)**.** *For any random variable x, $\epsilon > 0$*

$$\mathbb{P}(|x| \geq \epsilon) \leq \frac{\mathbb{E}e^{\lambda x}}{e^{\lambda\epsilon}}$$

and

$$\mathbb{P}(|x| \geq \epsilon) \leq \inf_{\lambda<0} e^{-\lambda\epsilon} \mathbb{E}e^{\lambda x}.$$

Proof.

$$\mathbb{P}(x > \epsilon) = \mathbb{P}(e^{\lambda x} > e^{\lambda \epsilon}) \leq \frac{\mathbb{E}e^{\lambda x}}{e^{\lambda \epsilon}}. \ \square$$

11.3. Exponential inequalities

For the sums or averages of independent random variables the above bounds can be improved from polynomial in $1/\epsilon$ to exponential in ϵ.

Theorem (Bennet)**.** *Let $x_1, ..., x_n$ be independent random variables with $\mathbb{E}x = 0$, $\mathbb{E}x^2 = \sigma^2$, and $|x_i| \leq M$. For $\epsilon > 0$*

$$\mathbb{P}\left(|\sum_{i=1}^{n} x_i| > \epsilon \right) \leq 2e^{\frac{-n\sigma^2}{M^2}\phi\left(\frac{\epsilon M}{n\sigma^2}\right)},$$

where

$$\phi(z) = (1+z)\log(1+z) - z.$$

Proof. We will prove a bound on one-side of the above theorem

$$\mathbb{P}\left(\sum_{i=1}^{n} x_i > \epsilon \right).$$

$$\begin{aligned}\mathbb{P}\left(\sum_{i=1}^{n} x_i > \epsilon \right) &\leq e^{-\lambda\epsilon}\mathbb{E}e^{\lambda \sum x_i} = e^{-\lambda\epsilon}\Pi_{i=1}^{n}\mathbb{E}e^{\lambda x_i} \\ &= e^{-\lambda\epsilon}(\mathbb{E}e^{\lambda x})^n.\end{aligned}$$

$$\begin{aligned}\mathbb{E}e^{\lambda x} &= \mathbb{E}\sum_{k=0}^{\infty}\frac{(\lambda x)^k}{k!} = \sum_{k=0}^{\infty}\lambda^k\frac{\mathbb{E}x^k}{k!} \\ &= 1 + \sum_{k=2}^{\infty}\frac{\lambda^k}{k!}\mathbb{E}x^2x^{k-2} \leq 1 + \sum_{k=2}^{\infty}\frac{\lambda^k}{k!}M^{k-2}\sigma^2 \\ &= 1 + \frac{\sigma^2}{M^2}\sum_{k=2}^{\infty}\frac{\lambda^k M^k}{k!} = 1 + \frac{\sigma^2}{M^2}(e^{\lambda M} - 1 - \lambda M) \\ &\leq e^{\frac{\sigma^2}{M^2}(e^{\lambda M} - \lambda M - 1)}.\end{aligned}$$

The last line holds since $1 + x \leq e^x$.

Therefore,

$$\mathbb{P}\left(\sum_{i=1}^{n} x_i > \epsilon \right) \leq e^{-\lambda\epsilon}e^{\frac{\sigma^2}{M^2}(e^{\lambda M} - \lambda M - 1)}. \tag{11.5}$$

We now optimize with respect to λ by taking the derivative with respect to λ

$$\begin{aligned}0 &= -\epsilon + \frac{n\sigma^2}{M^2}(Me^{\lambda M} - M), \\ e^{\lambda M} &= \frac{\epsilon M}{n\sigma^2} + 1, \\ \lambda &= \frac{1}{M}\log\left(1 + \frac{\epsilon M}{n\sigma^2}\right).\end{aligned}$$

The theorem is proven by substituting λ into equation (11.5). $\square$

The problem with Bennet's inequality is that it is hard to get a simple expression for ϵ as a function of the probability of the sum exceeding ϵ.

Theorem (Bernstein)**.** *Let $x_1, ..., x_n$ be independent random variables with $\mathbb{E}x = 0$, $\mathbb{E}x^2 = \sigma^2$, and $|x_i| \leq M$. For $\epsilon > 0$*

$$\mathbb{P}\left(|\sum_{i=1}^{n} x_i| > \epsilon\right) \leq 2e^{-\frac{\epsilon^2}{2n\sigma^2 + \frac{2}{3}\epsilon M}}.$$

Proof.
Take the proof of Bennet's inequality and notice

$$\phi(z) \geq \frac{z^2}{2 + \frac{2}{3}z}. \ \square$$

Remark. With Bernstein's inequality a simple expression for ϵ as a function of the probability of the sum exceeding ϵ can be computed

$$\sum_{i=1}^{n} x_i \leq \frac{2}{3}uM + \sqrt{2n\sigma^2 u}.$$

Outline.

$$\mathbb{P}\left(\sum_{i=1}^{n} x_i > \epsilon\right) \leq 2e^{-\frac{\epsilon^2}{2n\sigma^2 + \frac{2}{3}\epsilon M}} = e^{-u},$$

where

$$u = \frac{\epsilon^2}{2n\sigma^2 + \frac{2}{3}\epsilon M}.$$

we now solve for ϵ

$$\epsilon^2 - \frac{2}{3}\epsilon M - 2n\sigma^2\epsilon = 0$$

and

$$\epsilon = \frac{1}{3}uM + \sqrt{\frac{u^2M^2}{9} + 2n\sigma^2 u}.$$

Since $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$

$$\epsilon = \frac{2}{3}uM + \sqrt{2n\sigma^2 u}.$$

So with large probability

$$\sum_{i=1}^{n} x_i \leq \frac{2}{3}uM + \sqrt{2n\sigma^2 u}. \ \triangle$$

If we want to bound

$$|n^{-1}\sum_{i=1}^{n} f(x_i) - \mathbb{E}f(x)|$$

we consider

$$|f(x_i) - \mathbb{E}f(x)| \leq 2M.$$

Therefore

$$\sum_{i=1}^{n}(f(x_i) - \mathbb{E}f(x)) \leq \frac{4}{3}uM + \sqrt{2n\sigma^2 u}$$

and

$$n^{-1}\sum_{i=1}^{n} f(x_i) - \mathbb{E}f(x) \leq \frac{4}{3}\frac{uM}{n} + \sqrt{\frac{2\sigma^2 u}{n}}.$$

Similarly,

$$\mathbb{E}f(x) - n^{-1}\sum_{i=1}^{n} f(x_i) \geq \frac{4}{3}\frac{uM}{n} + \sqrt{\frac{2\sigma^2 u}{n}}.$$

In the above bound

$$\sqrt{\frac{2\sigma^2 u}{n}} \geq \frac{4uM}{n}$$

which implies $u \leq \frac{n\sigma^2}{8M^2}$ and therefore

$$|n^{-1}\sum_{i=1}^{n} f(x_i) - \mathbb{E}f(x)| \lesssim \sqrt{\frac{2\sigma^2 u}{n}} \text{ for } u \lesssim n\sigma^2,$$

which corresponds to the tail probability for a Gaussian random variable and is predicted by the Central Limit Theorem (CLT) Condition that $\lim_{n\to\infty} n\sigma^2 \to \infty$. If $\lim_{n\to\infty} n\sigma^2 = C$, where C is a fixed constant, then

$$|n^{-1}\sum_{i=1}^{n} f(x_i) - \mathbb{E}f(x)| \lesssim \frac{C}{n}$$

which corresponds to the tail probability for a Poisson random variable.

We now look at an even simpler exponential inequality where we do not need information on the variance.

Theorem (Hoeffding)**.** *Let $x_1, ..., x_n$ be independent random variables with $\mathbb{E}x = 0$ and $|x_i| \leq M_i$. For $\epsilon > 0$*

$$\mathbb{P}\left(|\sum_{i=1}^{n} x_i| > \epsilon\right) \leq 2e^{-\frac{2\epsilon^2}{\sum_{i=1}^{n} M_i^2}}.$$

Proof.

$$\mathbb{P}\left(\sum_{i=1}^{n} x_i > \epsilon\right) \leq e^{-\lambda\epsilon}\mathbb{E}e^{\lambda\sum_{i=1}^{n} x_i} = e^{-\lambda\epsilon}\Pi_{i=1}^{n}\mathbb{E}e^{\lambda x_i}.$$

It can be shown (Homework problem)

$$\mathbb{E}(e^{\lambda x_i}) \leq e^{\frac{\lambda^2 M_i^2}{8}}.$$

The bound is proven by optimizing the following with respect to λ

$$e^{-\lambda\epsilon}\Pi_{i=1}^{n} e^{\frac{\lambda^2 M_i^2}{8}}. \ \square$$

Applying Hoeffding's inequality to

$$n^{-1}\sum_{i=1}^{n} f(x_i) - \mathbb{E}f(x)$$

we can state that with probability $1 - e^{-u}$

$$n^{-1}\sum_{i=1}^{n} f(x_i) - \mathbb{E}f(x) \leq \sqrt{\frac{2Mu}{n}},$$

which is a sub-Gaussian as in the CLT but without the variance information we can never achieve the $\frac{1}{n}$ rate we achieved when the random variable has a Poisson tail distribution.

We will use the following version of Hoeffding's inequality in later lectures on Kolmogorov chaining and the Dudley's entropy integral.

Theorem (Hoeffding)**.** *Let $x_1, ..., x_n$ be independent random variables with $\mathbb{P}(x_i = M_i) = 1/2$ and $\mathbb{P}(x_i = -M_i) = 1/2$. For $\epsilon > 0$*

$$\mathbb{P}\left(|\sum_{i=1}^{n} x_i| > \epsilon\right) \leq 2e^{-\frac{2\epsilon^2}{\sum_{i=1}^{n} M_i^2}}.$$

Proof.

$$\mathbb{P}\left(\sum_{i=1}^{n} x_i > \epsilon\right) \leq e^{-\lambda\epsilon}\mathbb{E}e^{\lambda\sum_{i=1}^{n} x_i} = e^{-\lambda\epsilon}\Pi_{i=1}^{n}\mathbb{E}e^{\lambda x_i}.$$

$$\begin{aligned} \mathbb{E}(e^{\lambda x_i}) &= \frac{1}{2}e^{\lambda M_i} + \frac{1}{2}e^{-\lambda M_i}, \\ \frac{1}{2}e^{\lambda M_i} + \frac{1}{2}e^{-\lambda M_i} &= \sum_{k=0}^{\infty}\frac{(M_i\lambda)^{2k}}{(2k)!} \leq e^{\frac{\lambda^2 M_i^2}{2}}. \end{aligned}$$

Optimize the following with respect to λ

$$e^{-\lambda\epsilon}\Pi_{i=1}^{n} e^{\frac{\lambda^2 M_i^2}{2}}. \ \square$$

11.4. Martingale inequalities

In the previous section we stated some concentration inequalities for sums of independent random variables. We now look at more complicated functions of independent random variables and introduce a particular Martingale inequality to prove concentration.

Let $(\Omega, \mathcal{L}, \mu)$ be a probability space. Let $x_1, ..., x_n$ be real random variables on Ω. Let the function $Z(x_1, ..., x_n) : \Omega^n \to \mathbb{R}$ be a map from the random variables to a real number.

The function Z concentrates if the deviation between $Z(x_1, ..., x_n)$ and $\mathbb{E}_{x_1,..,x_n} Z(x_1, .., x_n)$ goes to zero as n goes to infinity.

Theorem (McDiarmid)**.** *Let $x_1, ..., x_n$ be independent random variables let $Z(x_1, ..., x_n) : \Omega^n \to \mathbb{R}$ such that*

$$\forall x_1, ..., x_n, x_1', ..., x_n' \quad |Z(x_1, .., x_n) - Z(x_1, ..., x_{i-1}, x_i', x_{i+1}, x_n)| \leq c_i,$$

then

$$\mathbb{P}\left(Z - \mathbb{E}Z > \epsilon\right) \leq e^{-\frac{\epsilon^2}{2\sum_{i=1}^{n} c_i^2}}.$$

Proof.

$$\mathbb{P}(Z - \mathbb{E}Z > \epsilon) = \mathbb{P}(e^{\lambda(Z-\mathbb{E}Z)} > e^{\lambda\epsilon}) \leq e^{-\lambda\epsilon}\mathbb{E}e^{\lambda(Z-\mathbb{E}Z)}.$$

We will use the following very useful decomposition

$$\begin{aligned} Z(x_1,...,x_n) - \mathbb{E}_{x'_1,..,x'_n} Z(x'_1,..,x'_n) &= [Z(x_1,...,x_n) - E_{x'_1} Z(x'_1,x_2,...,x_n)] \\ &+ [E_{x'_1} Z(x'_1,x_2,...,x_n) - E_{x'_1,x'_2} Z(x'_1,x'_2,x_3,...,x_n)] \\ &+ ... \\ &+ [E_{x'_1,...,x'_{n-1}} Z(x'_1,x'_2,...x'_{n-1},x_n) - E_{x'_1,...,x'_n} Z(x'_1,...,x'_n)]. \end{aligned}$$

We denote the random variable

$$z_i(x_i,...,x_n) := \mathbb{E}_{x'_1,...,x'_{i-1}} Z(x'_1,...,x'_{i-1},x_i,...,x_n) - \mathbb{E}_{x'_1,...,x'_i} Z(x'_1,...,x'_i,x_{i+1},...,x_n),$$

and

$$Z(x_1,...,x_n) - \mathbb{E}_{x'_1,..,x'_n} Z(x'_1,..,x'_n) = z_1 + ... + z_n.$$

The following inequality is true (see the following Lemma for a proof)

$$\mathbb{E}_{x_i} e^{\lambda z_i} \leq e^{\lambda^2 c_i^2/2} \ \ \forall \lambda \in \mathbb{R}.$$

$$\begin{aligned} \mathbb{E}e^{\lambda(Z-\mathbb{E}Z)} &= \mathbb{E}e^{\lambda(z_1+...,+z_n)} \\ \mathbb{E}\mathbb{E}_{x_1} e^{\lambda(z_1+...,+z_n)} &= \mathbb{E}e^{\lambda(z_2+...,+z_n)} \mathbb{E}_{x_1} e^{\lambda z_1} \\ &\leq \mathbb{E}e^{\lambda(z_2+...,+z_n)} e^{\lambda c_1^2/2}, \end{aligned}$$

by induction

$$\mathbb{E}e^{\lambda(Z-\mathbb{E}Z)} \leq e^{\lambda^2 \sum_{i=1}^n c_i^2/2}.$$

To derive the bound we optimize with respect to λ

$$e^{-\lambda\epsilon + \lambda^2 \sum_{i=1}^n c_i^2/2}. \ \square$$

Lemma. *For all* $\lambda \in \mathbb{R}$

$$\mathbb{E}_{x_i} e^{\lambda z_i} \leq e^{\lambda^2 c_i^2/2}.$$

Proof.
For any $t \in [-1,1]$ the function $e^{\lambda t}$ is convex with respect to λ.

$$\begin{aligned} e^{\lambda t} &= e^{\lambda(\frac{1+t}{2}) - \lambda(\frac{1-t}{2})} \\ &\leq \frac{1+t}{2} e^{\lambda} + \frac{1-t}{2} e^{-\lambda} \\ &= \frac{e^{\lambda} + e^{-\lambda}}{2} + t \frac{e^{\lambda} - e^{-\lambda}}{2} \\ &\leq e^{\lambda^2/2} + t \, \mathrm{sh}(\lambda). \end{aligned}$$

Set $t = \frac{z_i}{c_i}$ and notice that $\frac{z_i}{c_i} \in [-1,1]$ so,

$$e^{\lambda z_i} = e^{\lambda c_i \frac{z_i}{c_i}} \leq e^{\lambda^2 c_i^2/2} + \frac{z_i}{c_i} \mathrm{sh}(\lambda c_i),$$

and

$$\mathbb{E}_{x_i} e^{\lambda z_i} \leq e^{\lambda^2 c_i^2/2}. \ \square$$

Example. *We can use McDiarmid's inequality to prove the concentration of the empirical minima. Given a dataset* $\{v_1 = (x_1,y_1),...,v_n = (x_n,y_n)\}$ *the empirical minima is*

$$Z(v_1,...,v_n) = \min_{f \in \mathcal{H}_K} n^{-1} \sum_{i=1}^n V(f(x_i),y_i).$$

If the loss function is bounded one can show that for all $(v_1, ..., v_n, v_i')$

$$|Z(v_1, ..., v_n) - Z(v_1, ..., v_{i-1}, v_i', ...v_n)| \leq \frac{k}{n}.$$

Therefore with probability $1 - e^{-u}$

$$|Z - \mathbb{E}Z| \leq \sqrt{\frac{2ku}{n}}.$$

So the empirical minima concentrates.

LECTURE 12
Vapnik-Červonenkis theory

12.1. Uniform law of large numbers

In the previous lecture we considered law of large numbers for a single or fixed function. We termed this as one dimensional concentration inequalities. We now look at uniform law of large numbers, that is a law of large numbers that holds uniformly over a class of functions.

The point of these uniform limit theorems is that if the law of large numbers holds for all functions in a hypothesis space then it holds for the empirical minimizer.

The reason this chapter is called Vapnik-Červonenkis theory is that they provided some of the basic tools to study these classes.

12.2. Generalization bound for one function

Before looking at uniform results we prove generalization results when the hypothesis space $\mathcal{H}$ consists of one function, f_1.

In this case the empirical risk minimizer is f_1

$$f_1 = f_S := \arg\min_{f\in\mathcal{H}} \left[n^{-1} \sum_{i=1}^{n} V(f, z_i) \right].$$

Theorem. *Given $0 \le V(f_1, z) \le M$ for all z and $S = \{z_i\}_{i=1}^n$ drawn i.i.d. then with probability at least $1 - e^{-t}$ $(t > 0)$*

$$\mathbb{E}_z V(f_1, z) \le n^{-1} \sum_{i=1}^{n} V(f_1, z_i) + \sqrt{\frac{M^2 t}{n}}.$$

Proof.

By Hoeffding's inequality

$$\mathbb{P}\left(\mathbb{E}_z V(f_1, z) - n^{-1} \sum_{i=1}^{n} V(f_1, z_i) > \varepsilon \right) \le e^{-n\varepsilon^2/M^2}$$

so

$$\mathbb{P}\left(\mathbb{E}_z V(f_1, z) - n^{-1} \sum_{i=1}^{n} V(f_1, z_i) \le \varepsilon \right) > 1 - e^{-n\varepsilon^2/M^2}.$$

Set $t = n\varepsilon^2/M^2$ and the result follows. □

12.3. Generalization bound for a finite number of functions

We now look at the case of ERM on a hypothesis space $\mathcal{H}$ with a finite number of functions, $k = |\mathcal{H}|$. In this case, the empirical minimizer will be one of the k functions.

Theorem. *Given $0 \leq V(f_j, z) \leq M$ for all $f_j \in \mathcal{H}, z$ and $S = \{z_i\}_{i=1}^n$ drawn i.i.d. then with probability at least $1 - e^{-t}$ $(t > 0)$ for the empirical minimizer, f_S,*

$$\mathbb{E}_z V(f_S, z) < n^{-1} \sum_{i=1}^n V(f_S, z_i) + \sqrt{\frac{M^2(\log K + t)}{n}}.$$

Proof.

The follow implication of events holds

$$\left\{ \max_{f_j \in \mathcal{H}} \; \mathbb{E}_z V(f_j, z) - n^{-1} \sum_{i=1}^n V(f_j, z_i) < \varepsilon \right\} \Rightarrow \left\{ \mathbb{E}_z V(f_S, z) - n^{-1} \sum_{i=1}^n V(f_S, z_i) < \varepsilon \right\}.$$

$$\begin{aligned}
& \mathbb{P}\left(\max_{f_j \in \mathcal{H}} \; \mathbb{E}_z V(f_j, z) - n^{-1} \sum_{i=1}^n V(f_j, z_i) \geq \varepsilon \right) \\
= \; & \mathbb{P}\left(\bigcup_{f \in \mathcal{H}} \left\{ \mathbb{E}_z V(f, z) - n^{-1} \sum_{i=1}^n V(f, z_i) \geq \varepsilon \right\} \right) \\
\leq \; & \sum_{f_j \in \mathcal{H}} \mathbb{P}\left(\mathbb{E}_z V(f_j, z) - n^{-1} \sum_{i=1}^n V(f_j, z_i) \geq \varepsilon \right) \\
\leq \; & k e^{-n\varepsilon^2/M^2},
\end{aligned}$$

the last step comes from our single function result. Set $e^{-t} = k e^{-n\varepsilon^2/M^2}$ and the result follows. □

12.4. Generalization bound for compact hypothesis spaces

We now prove a sufficient condition for the generalization of hypothesis spaces with an infinite number of functions and then give some examples of such spaces.

We first assume that our hypothesis space is a subset of the space of continuous functions, $\mathcal{H} \subset \mathcal{C}(\mathcal{X})$.

Definition. *A metric space is compact if and only if it is totally bounded and complete.*

Definition. *Let R be a metric space and ϵ any positive number. Then a set $A \subset R$ is said to be an ϵ-net for a set $M \subset R$ if for every $x \in M$, there is at least one point $a \in A$ such that $\rho(x, a) < \epsilon$. Here $\rho(\cdot, \cdot)$ is a norm.*

Definition. *Given a metric space R and a subset $M \subset R$ suppose M has a finite ϵ-net for every $\epsilon > 0$. Then M is said to be totally bounded.*

Proposition. *A compact space has a finite ϵ-net for all $\epsilon > 0$.*

For the remainder of this section we will use the supnorm,

$$\rho(a, b) = \sup_{x \in \mathcal{X}} |a(x) - b(x)|.$$

Definition. *Given a hypothesis space $\mathcal{H}$ and the supnorm, the covering number $\mathcal{N}(\mathcal{H},\epsilon)$ is the minimal number $\ell \in \mathbb{N}$ such that for every $f \in \mathcal{H}$ there exists functions $\{g_i\}_{i=1}^{\ell}$ such that*

$$\sup_{x\in\mathcal{X}} |f(x) - g_i(x)| \leq \epsilon \text{ for some } i.$$

We now state a generalization bound for this case. In the bound we assume $V(f,z) = (f(x)-y)^2$ but the result can be easily extended for any Lipschitz loss

$$|V(f_1,z) - V(f_2,z)| \leq C||f_1(x) - f_2(x)||_\infty \ \forall z.$$

Theorem. *Let $\mathcal{H}$ be a compact subset of $\mathcal{C}(\mathcal{X})$. Given $0 \leq |f(x)-y| \leq M$ for all $f \in \mathcal{H}, z$ and $S = \{z_i\}_{i=1}^n$ drawn i.i.d. then with probability at least $1-e^{-t}$ ($t>0$) for the empirical minimizer, f_S,*

$$\mathbb{E}_{x,y}(f_S(x)-y)^2 < n^{-1}\sum_{i=1}^n (f_S(x_i)-y_i)^2 + \sqrt{\frac{M^2(\log\mathcal{N}(\mathcal{H},\varepsilon/8M)+t)}{n}}.$$

We first prove two useful lemmas. Define

$$D(f,S) := \mathbb{E}_{x,y}(f(x)-y)^2 - n^{-1}\sum_{i=1}^n (f(x_i)-y_i)^2.$$

Lemma. *If $|f_j(x)-y| \leq M$ for $j=1,2$ then*

$$|D(f_1,S) - D(f_2,S)| \leq 4M||f_1-f_2||_\infty.$$

Proof. Note that

$$(f_1(x)-y)^2 - (f_2(x)-y)^2 = (f_1(x)-f_2(x))(f_1(x)+f_2(x)-2y)$$

so

$$\begin{aligned} |\mathbb{E}_{x,y}(f_1(x)-y)^2 - \mathbb{E}_{x,y}(f_2(x)-y)^2| &= \left|\int (f_1(x)-f_2(x))(f_1(x)+f_2(x)-2y)d\mu(x,y)\right| \\ &\leq ||f_1-f_2||_\infty \int |f_1(x)-y+f_2(x)-y|du(x,y) \\ &\leq 2M||f_1-f_2||_\infty, \end{aligned}$$

and

$$\begin{aligned} |n^{-1}\sum_{i=1}^n [(f_1(x_i)-y_i)^2 - (f_2(x_i)-y_i)^2]| &= n^{-1}\left|\sum_{i=1}^n (f_1(x_i)-f_2(x_i))(f_1(x_i)+f_2(x_i)-2y_i)\right| \\ &\leq ||f_1-f_2||_\infty \frac{1}{n}\sum_{i=1}^n |f_1(x_i)-y_i+f_2(x_i)-y_i| \\ &\leq 2M||f_1-f_2||_\infty. \end{aligned}$$

The result follows from the above inequalities. □

Lemma. *Let $\mathcal{H} = B_1 \bigcup ... \bigcup B_\ell$ and $\varepsilon > 0$. Then*

$$\mathbb{P}\left(\sup_{f\in\mathcal{H}} D(f,S)\right) \leq \sum_{j=1}^{\ell} \mathbb{P}\left(\sup_{f\in B_j} D(f,S)\right).$$

Proof.

The result follows from the following equivalence and the union bound

$$\sup_{f\in\mathcal{H}} D(f,S) \geq \varepsilon \Longleftrightarrow \exists j \leq \ell \text{ s.t. } \sup_{f\in B_j} D(f,S) \geq \varepsilon. \ \square$$

We now prove Theorem 12.4.

Let $\ell = \mathcal{N}\left(\mathcal{H}, \frac{\varepsilon}{4M}\right)$ and the functions $\{g_j\}_{j=1}^{\ell}$ have the property that the disks B_j centered at f_j with radius $\frac{\varepsilon}{4M}$ cover $\mathcal{H}$. By the first lemma for all $f \in B_j$

$$|D(f,S) - D(f_j,S)| \leq 4M||f - f_j||_\infty \leq 4M\frac{\varepsilon}{4M} = \varepsilon,$$

this implies that for all $f \in B_j$

$$\sup_{f\in B_j} |D(f,S)| \geq 2\varepsilon \Rightarrow |D(f_j,S)| \geq \varepsilon.$$

So

$$\mathbb{P}\left(\sup_{f\in B_j} |D(f,S)| \geq 2\varepsilon\right) \leq \mathbb{P}\left(|D(f_j,S)| \geq \varepsilon\right) \leq 2e^{-\varepsilon^2 n/M^2}.$$

This combined with the second lemma implies

$$\mathbb{P}\left(\sup_{f\in\mathcal{H}} D(f,S)\right) \leq \mathcal{N}\left(\mathcal{H}, \frac{\varepsilon}{8M}\right) e^{-\varepsilon^2 n/M^2}.$$

Since the following implication of events holds

$$\left\{\sup_{f\in\mathcal{H}} \ \mathbb{E}_z V(f_j,z) - n^{-1}\sum_{i=1}^{n} V(f_j,z_i) < \varepsilon\right\} \Rightarrow \left\{\mathbb{E}_z V(f_S,z) - n^{-1}\sum_{i=1}^{n} V(f_S,z_i) < \varepsilon\right\}$$

the result is obtained by setting $e^{-t} = \mathcal{N}\left(\mathcal{H}, \frac{\varepsilon}{8M}\right) e^{-n\varepsilon^2/M^2}$. $\square$

A result of the above theorem is the following sufficient condition for uniform convergence and consistency of ERM.

Corollary. *For a Lipschitz loss function ERM is consistent if for all $\varepsilon > 0$*

$$\lim_{n\to\infty} \frac{\log \mathcal{N}\left(\mathcal{H},\varepsilon\right)}{n} = 0.$$

Proof.

This follows directly from the statement

$$\mathbb{P}\left(\sup_{f\in\mathcal{H}} D(f,S)\right) \leq \mathcal{N}\left(\mathcal{H}, \frac{\varepsilon}{8M}\right) e^{-\varepsilon^2 n/M^2}. \ \square$$

We now compute covering numbers for a few types of hypothesis spaces.

We also need the definition of packing numbers.

Definition. *Given a hypothesis space $\mathcal{H}$ and the supnorm, ℓ functions $\{g_i\}_{i=1}^{\ell}$ are ϵ-separated if*

$$\sup_{x\in\mathcal{X}} |g_j(x) - g_i(x)| > \epsilon \ \ \forall i \neq j.$$

The packing number $\mathcal{P}(\mathcal{H},\epsilon)$ is the maximum cardinality of ϵ-separated sets.

The following relationship between packing and covering numbers is very useful.

Lemma. *Given a metric space (A, ρ). Then for all $\epsilon > 0$ and for every $W \subset A$, the covering numbers and packing numbers satisfy*

$$\mathcal{P}(W, 2\epsilon, \rho) \leq \mathcal{N}(W, \epsilon, \rho) \leq \mathcal{P}(W, \epsilon, \rho).$$

Proof.

For the second inequality suppose P is an ϵ-packing of maximal cardinality, $\mathcal{P}(W, \epsilon, d)$. Then for any $w \in W$ there must be a $u \in P$ with $\rho(u, w) < \epsilon$, otherwise w is not an element of P and $P \cup w$ is an ϵ-packing. This contradicts the assumption that P is a maximal ϵ-packing. So any maximal ϵ packing is an ϵ-cover.

For the first inequality suppose C is an ϵ-cover for W and and that P is a 2ϵ-packing of W with maximum cardinality $\mathcal{P}(W, \epsilon, d)$. We will show that $|P| \leq |C|$. Assume that $|C| > |P|$. Then for two points $w_1, w_2 \in P$ and one point $u \in C$ the following will hold

$$\rho(w_1, u) \leq \epsilon \text{ and } \rho(w_2, u) \leq \epsilon \Longrightarrow \rho(w_1, w_2) \leq 2\epsilon.$$

This contradicts the fact that the points in P are 2ϵ-separated. □

In general we will compute packing numbers for hypothesis spaces and use the above lemma to obtain the covering number.

The following proposition will be useful.

Proposition. *Given $x \in \mathbb{R}^d$, the restriction the space to the unit ball $B = \{x : ||x|| \leq M\}$, and the standard Euclidean metric $\rho(x, y) = ||x - y||$, then for $\epsilon \leq M$*

$$\mathcal{P}(B, \epsilon, \rho) \leq \left(\frac{3M}{\epsilon}\right)^d.$$

Proof.

The ℓ points $w_1, .., w_\ell$ form an optimal ϵ-packing so

$$\begin{aligned}
\text{Vol}\left(M + \frac{\epsilon}{2}\right) &= C_d \left(M + \frac{\epsilon}{2}\right)^d \\
\text{Vol}\left(\frac{\epsilon}{2}\right) &= C_d \left(\frac{\epsilon}{2}\right)^d \\
\ell[C_d \left(\frac{\epsilon}{2}\right)^d] &= C_d \left(M + \frac{\epsilon}{2}\right)^d \\
\ell &\leq \left(\frac{2M + \epsilon}{\epsilon}\right)^d \\
&\leq \left(\frac{3M}{\epsilon}\right)^d \text{ for all } \epsilon \leq M. \ \square
\end{aligned}$$

Example. *Covering numbers for a finite dimensional RKHS.*

For a finite dimensional bounded RKHS

$$\mathcal{H}_K = \left\{ f : f(x) = \sum_{p=1}^{m} c_p \phi_p(x) \right\},$$

with $\|f\|_K^2 \leq M$.

By the reproducing property and Cauchy-Schwartz inequality, the supnorm can be bound by the RKHS norm:

$$\begin{aligned} ||f(\mathbf{x})||_\infty &= ||\langle K(\mathbf{x},\cdot), f(\cdot)\rangle_K||_\infty \\ &\leq ||K(\mathbf{x},\cdot)||_K||f||_K \\ &= \sqrt{\langle K(x,\cdot), K(x,\cdot)\rangle}||f||_K \\ &= \sqrt{K(\mathbf{x},\mathbf{x})}||f||_K \\ &\leq \kappa||f||_K \end{aligned}$$

This means that if we can cover with the RKHS norm we can cover with the supnorm.

Each function in our cover, $\{g_i\}_{i=1}^{\ell}$ *can be written as*

$$g_i(x) = \sum_{p=1}^{m} d_{ip}\phi_p(x)$$

So if we find ℓ *vectors* d_i *for which for all* $c : \sum_{p=1}^{m} \frac{c_p^2}{\lambda_p} \leq M$ *there exists a* d_i *such that*

$$\sum_{p=1}^{m} \frac{(c_p - d_{ip})^2}{\lambda_p} < \epsilon^2,$$

we have a cover at scale ϵ*. The above is simply a weighted Euclidean norm and can be reduced to the problem of covering a ball of radius* M *in* $\mathbb{R}^m$ *using the Euclidean metric. Using proposition 12.4 we can bound the packing number with the RKHS norm and the supnorm*

$$\begin{aligned} \mathcal{P}(\mathcal{H}, \epsilon, ||\cdot||_{\mathcal{H}_k}) &\leq \left(\frac{3M}{\epsilon}\right)^m, \\ \mathcal{P}(\mathcal{H}, \epsilon, ||\cdot||_\infty) &\leq \left(\frac{3M}{\kappa\epsilon}\right)^m. \end{aligned}$$

Using lemma 12.4 we can get a bound on the covering number

$$\mathcal{N}(\mathcal{H}, \epsilon, ||\cdot||_\infty) \leq \left(\frac{3M}{\kappa\epsilon}\right)^m.$$

We have shown that for $\mathcal{H} \subset \mathcal{C}(\mathcal{X})$ that is compact with respect to the supnorm we have uniform convergence. This requirement is to strict to determine necessary conditions. A large class of functions that these conditions do not apply to are indicator functions $f(x) \in \{0, 1\}$.

12.5. Generalization bound for hypothesis spaces of indicator functions

In this section we derive necessary and sufficient conditions for uniform convergence of indicator functions and as a result generalization bounds for indicator functions, $f(x) \in \{0, 1\}$.

As in the case of compact functions we will take a class of indicator functions $\mathcal{H}$ and reduce this to some finite set of functions. In the case of indicator functions this is done via the notion of a growth function which we now define.

Definition. *Given a set of n points $\{x_i\}_{i=1}^n$ and a class of indicator functions $\mathcal{H}$ we say that a function $f \in \mathcal{H}$ picks out a certain subset of $\{x_i\}_{i=1}^n$ if this set can be formed by the operation $f \cap \{x_i\}_{i=1}^n$. The cardinality of the number of subsets that can be picked out is called the growth function:*

$$\triangle_n(\mathcal{H}, \{x_i\}_{i=1}^n) = \#\left\{f \cap \{x_i\}_{i=1}^n : f \in \mathcal{H}\right\}.$$

We will now state a lemma which will look very much like the generalization results for the compact or finite dimensional case.

Lemma. *Let $\mathcal{H}$ be a class of indicator functions and $S = \{z_i\}_{i=1}^n$ drawn i.i.d. then with probability at least $1 - e^{-t/8}$ $(t > 0)$ for the empirical minimizer, f_S,*

$$\mathbb{E}_{x,y} I_{\{f_S(x) \neq y\}} < n^{-1} \sum_{i=1}^n I_{\{f_S(x_i) \neq y_i\}} + \sqrt{\frac{(8 \log 8 \triangle_n(\mathcal{H}, S) + t)}{n}},$$

where $\triangle_n(\mathcal{H}, S)$ is the growth function given S observations.

Note: the above result depends upon a particular draw of $2n$ samples through the growth function. We will remove this dependence soon.

We first prove two useful lemmas. Define

$$D(f, S) := \mathbb{E}_{x,y} I_{\{f(x) \neq y\}} - n^{-1} \sum_{i=1}^n I_{\{f(x_i) \neq y_i\}}.$$

The first lemma is based upon the idea of symmetrization and replaces the deviation between the empirical and expected error to the difference between two empirical errors.

Lemma. *Given two independent copies of the data $S = \{z_i\}_{i=1}^n$ and $S = \{z_i'\}_{i=1}^n$ then for any fixed $f \in \mathcal{H}$ if $n \geq 2/\epsilon^2$*

$$\mathbb{P}\left(|D(f, S)| > \epsilon\right) \leq 2\, \mathbb{P}\left(|D(f, S) - D(f, S')| > \epsilon/2\right),$$

where

$$|D(f, S) - D(f, S')| = n^{-1} \sum_{i=1}^n I_{\{f(x_i) \neq y_i\}} - n^{-1} \sum_{i=1}^n I_{\{f(x_i') \neq y_i'\}}.$$

Proof. We first assume that

$$\mathbb{P}(|D(f, S')| \leq \epsilon/2 \,|\, S) \geq 1/2,$$

where we have conditioned on S. Since S and S' are independent we can integrate out

$$1/2\, \mathbb{P}(|D(f, S')| > \epsilon) \leq \mathbb{P}(|D(f, S')| \leq \epsilon/2, |D(f, S')| > \epsilon).$$

By the triangle inequality $|D(f, S)| > \epsilon$ and $|D(f, S')| \leq \epsilon/2$ implies

$$|D(f, S) - D(f, S')| \geq \epsilon/2,$$

so

$$\mathbb{P}(|D(f, S')| \leq \epsilon/2, |D(f, S')| > \epsilon) \leq \mathbb{P}(|D(f, S) - D(f, S')| \geq \epsilon/2).$$

To complete the proof we need to show our initial assumption holds. Since $\mathcal{H}$ is a class of indicator functions the elements in the sum are binomial random variables and the variance of n of them will be at most $1/4n$. So by the Bienaymé-Chebyshev inequality

$$\mathbb{P}(|D(f, S')| > \epsilon/2) \geq 1/4n\epsilon^2,$$

which implies the initial assumption when $n \geq 2/\epsilon^2$. $\square$

By symmetrizing we now have a term that depends only on samples. The problem is that it depends on the samples we have but also an independent compy. This nuisance is removed by a second step of symmetrization.

Lemma. *Let σ_i be a Rademacher random variable ($\mathbb{P}(\sigma_i = \pm 1) = 1/2$) then*

$$\mathbb{P}\left(|D(f,S) - D(f,S')| > \epsilon/2\right) \leq 2\mathbb{P}\left(\left|n^{-1}\sum_{i=1}^{n}\sigma_i I_{\{f(x_i)\neq y_i\}}\right| > \epsilon/4\right).$$

Proof.

$$\begin{aligned}
\mathbb{P}\left(|D(f,S) - D(f,S')| > \epsilon/2\right) &= \mathbb{P}\left(\left|n^{-1}\sum_{i=1}^{n}\sigma_i I_{\{f(x_i)\neq y_i\}} - n^{-1}\sum_{i=1}^{n}\sigma_i I_{\{f(x_i')\neq y_i'\}}\right| > \epsilon/2\right) \\
&\leq \mathbb{P}\left(\left|n^{-1}\sum_{i=1}^{n}\sigma_i I_{\{f(x_i)\neq y_i\}}\right| > \epsilon/4\right) + \\
&\quad \mathbb{P}\left(\left|n^{-1}\sum_{i=1}^{n}\sigma_i I_{\{f(x_i')\neq y_i'\}}\right| > \epsilon/4\right) \\
&\leq 2\mathbb{P}\left(\left|n^{-1}\sum_{i=1}^{n}\sigma_i I_{\{f(x_i)\neq y_i\}}\right| > \epsilon/4\right). \ \square
\end{aligned}$$

The second symmetrization step allows is to bound the deviation between the empirical and expected errors based upon a quantity computed just on the empirical data.

We now prove Lemma 12.5.

By the symmetrization lemmas for $n \geq 8/\epsilon^2$

$$\mathbb{P}(|D(f,S)| > \epsilon) \leq 4\,\mathbb{P}\left(\left|n^{-1}\sum_{i=1}^{n}\sigma_i I_{\{f(x_i)\neq y_i\}}\right| > \epsilon/4\right).$$

By the Rademacher version of Hoeffdings inequality

$$\mathbb{P}\left(\left|n^{-1}\sum_{i=1}^{n}\sigma_i I_{\{f(x_i)\neq y_i\}}\right| > \epsilon\right) \leq 2e^{-2\epsilon^2}.$$

Combining the above for a single function

$$\mathbb{P}(|D(f,S)| > \epsilon) \leq 8e^{-\epsilon^2/8}.$$

Given data S the growth function characterizes the cardinality of subsets that can be "picked out" which is a bound on the number of possible labellings or realizable functions, $\ell = \triangle_n(\mathcal{H}, S)$. We index the possible labelings by f_j where $j = 1, ..., \ell$.

We now proceed as in the case of a finite number of functions

$$\begin{aligned}\mathbb{P}\left(\sup_{f\in\mathcal{H}}|D(f,S)|\geq\epsilon\right) &= \mathbb{P}\left(\bigcup_{f\in\mathcal{H}}|D(f,S)|\geq\epsilon\right)\\ &\leq \sum_{i=1}^{\ell}\mathbb{P}\left(|D(f_j,S)|\geq\varepsilon\right)\\ &\leq 8\triangle_n(\mathcal{H},S)e^{-n\epsilon^2/8}.\end{aligned}$$

Setting $e^{-t/8}=8\triangle_n(\mathcal{H},S)e^{-n\epsilon^2/8}$ completes the proof. □

This bound is not uniform since the growth function depends on the data S. We can make the bound uniform by defining a uniform notion of the growth function.

Definition. *The uniform growth function is*

$$\triangle_n(\mathcal{H})=\max_{x_1,\ldots,x_n\in\mathcal{X}}\triangle_n(\mathcal{H},\{x_i\}_{i=1}^n).$$

Corollary. *Let $\mathcal{H}$ be a class of indicator functions and $S=\{z_i\}_{i=1}^n$ drawn i.i.d. then with probability at least $1-e^{-t/8}$ $(t>0)$ for the empirical minimizer, f_S,*

$$\mathbb{E}_{x,y}I_{\{f_S(x)\neq y\}}<n^{-1}\sum_{i=1}^{n}I_{\{f_S(x_i)\neq y_i\}}+\sqrt{\frac{(\log 8\triangle_n(\mathcal{H})+t)}{n}},$$

where $\triangle_n(\mathcal{H})$ is the uniform growth function.

Corollary. *For a class of indicator functions ERM is consistent if and only if for all $\varepsilon>0$*

$$\lim_{n\to\infty}\frac{8\log\triangle_n(\mathcal{H})}{n}=0.$$

We now characterize conditions under which the uniform growth function grows polynomially. To do this we need a few definitions.

Definition. *A hypothesis space, $\mathcal{H}$, shatters a set $\{x_1,\ldots,x_n\}$ if each of its 2^n subsets can be "picked out" by $\mathcal{H}$. The Vapnik-Červonenkis (VC) dimension, $v(\mathcal{H})$, of a hypothesis space is the largest n for which all sets of size n are shattered by $\mathcal{H}$*

$$v(\mathcal{H})=\sup\left\{n:\triangle_n(\mathcal{H})=2^n\right\},$$

if the exists no such n then the VC dimension is infinite.

Definition. *A hypothesis space of indicator functions $\mathcal{H}$ is a VC class if and only if it has finite VC dimension.*

Examples.

The VC dimension controls the growth function via the following lemma.

Lemma. *For a hypothesis space $\mathcal{H}$ with VC dimension d and $n>d$*

$$\triangle_n(\mathcal{H})\leq\sum_{i=1}^{d}\binom{n}{i}.$$

Proof.

The proof will be by induction on $n+d$. We define $\binom{n}{i} := 0$ if $i < 0$ or $i > n$. In addition one can check

$$\binom{n}{i} = \binom{n-1}{i-1} + \binom{n-1}{i}.$$

When $d = 0$ $|\mathcal{H}| = 1$ since no points can be shattered so for all n

$$\triangle_n(\mathcal{H}) = 1 = \binom{n}{0} = \Phi_0(n).$$

When $n = 0$ there is only one way to label 0 examples so

$$\triangle_0(\mathcal{H}) = 1 = \sum_{i=1}^{d} \binom{0}{i} = \Phi_d(0).$$

Assume the lemma to hold for n', d' such that $n' + d' < n + d$.

Given $S = \{x_1, ..., x_n\}$ and $S_n = \{x_1, ..., x_{n-1}\}$. We now define three hypothesis spaces $\mathcal{H}_0$, $\mathcal{H}_1$, and $\mathcal{H}_2$:

$$\begin{aligned} \mathcal{H}_0 &:= \{f_i : i = 1, ..., \triangle_n(\mathcal{H}, S)\} \\ \mathcal{H}_1 &:= \{f_i : i = 1, ..., \triangle_{n-1}(\mathcal{H}, S_n)\} \\ \mathcal{H}_2 &:= \mathcal{H}_0 - \mathcal{H}_1, \end{aligned}$$

where each f_i in set $\mathcal{H}_0$ is a possible labeling of S by $\mathcal{H}$, each f_i in set $\mathcal{H}_1$ is a possible labeling of S_n by $\mathcal{H}$.

For the set $\mathcal{H}_1$ over S_n: $n_1 = n-1$ since there is one fewer sample and $v(\mathcal{H}_1) \leq d$ since reducing the number of hypotheses cannot increase the VC dimension.

For the set $\mathcal{H}_2$ over S_n: $n_1 = n-1$ since there is one fewer sample and $v(\mathcal{H}_2) \leq d-1$. If $S' \subseteq S_n$ is shattered by $\mathcal{H}_2$ then all labellings of S' must occur both in $\mathcal{H}_1$ and $\mathcal{H}_2$ but with different labels on x_n. So $S' \cup \{x_n\}$ which has cardinality $|S'| + 1$ is shattered by $\mathcal{H}$ and so $|S'|$ cannot be more than $d-1$.

By induction $\triangle_{n-1}(\mathcal{H}_1, S_n) \leq \Phi_d(m-1)$ and $\triangle_{n-1}(\mathcal{H}_2, S_n) \leq \Phi_{d-1}(m-1)$.

By construction

$$\begin{aligned} \triangle_n(\mathcal{H}, S) &= |\mathcal{H}_1| + |\mathcal{H}_2| = \triangle_{n-1}(\mathcal{H}_1, S_n) + \triangle_{n-1}(\mathcal{H}_2, S_n) \\ &\leq \Phi_d(n-1) + \Phi_{d-1}(n-1) \\ &= \sum_{i=0}^{d} \binom{n-1}{i} + \sum_{i=0}^{d-1} \binom{n-1}{i} \\ &= \sum_{i=0}^{d} \binom{n-1}{i} + \sum_{i=0}^{d} \binom{n-1}{i-1} \\ &= \sum_{i=0}^{d} \left[\binom{n-1}{i} + \binom{n-1}{i-1} \right] \\ &= \sum_{i=0}^{d} \binom{n}{i}. \ \square \end{aligned}$$

Lemma. *For $n \geq d \geq 1$*

$$\sum_{i=1}^{d} \binom{n}{i} < \left(\frac{en}{d}\right)^d.$$

Proof.

For $0 \leq i \leq d$ and $n \geq d$

$$(m/d)^d (d/m)^i \geq 1,$$

so

$$\sum_{i=1}^{d} \binom{n}{i} \leq (n/d)^d \sum_{i=1}^{d} \binom{n}{i} (d/n)^i \leq (n/d)^d (1 + d/n)^n < (ne/d)^d. \quad \square$$

This now lets state the generalization bound in terms of VC dimension.

Theorem. *Let $\mathcal{H}$ be a class of indicator functions with VC dimension d and $S = \{z_i\}_{i=1}^n$ drawn i.i.d. then with probability at least $1 - e^{-t/8}$ $(t > 0)$ for the empirical minimizer, f_S,*

$$\mathbb{E}_{x,y} I_{\{f_S(x) \neq y\}} < n^{-1} \sum_{i=1}^{n} I_{\{f_S(x_i) \neq y_i\}} + 2\sqrt{\frac{(8d \log(8en/d) + t)}{n}}.$$

Proof. From the proof of lemma we have

$$\mathbb{P}\left(\sup_{f \in \mathcal{H}} |D(f, S)| \geq \epsilon \right) \leq 8 \triangle_n(\mathcal{H}, S) e^{-n\epsilon^2/8},$$

therefore since $\triangle_n(\mathcal{H}, S) \leq \left(\frac{en}{d}\right)^d$, we have

$$\mathbb{P}\left(\sup_{f \in \mathcal{H}} |D(f, S)| \geq \epsilon \right) \leq 8 \left(\frac{en}{d}\right)^d e^{-n\epsilon^2/8},$$

and setting $e^{-t/8} = 8 \left(\frac{en}{d}\right)^d e^{-n\epsilon^2/8}$ gives us

$$\mathbb{E}_{x,y} I_{\{f_S(x) \neq y\}} < n^{-1} \sum_{i=1}^{n} I_{\{f_S(x_i) \neq y_i\}} + \sqrt{\frac{(8d \log(8en/d) + t + 8 \log 8)}{n}},$$

for $n > 2$ and $d > 1$ $8 \log 8 < 8d \log(en/d)$ so

$$\sqrt{\frac{(8d \log(8en/d) + t + 8 \log 8)}{n}} < 2\sqrt{\frac{(8d \log(8en/d) + t)}{n}},$$

which proves the theorem. $\square$

Theorem. *For a class of indicator functions ERM the following are equivalent*

1. *ERM is consistent*
2. *for all $\varepsilon > 0$*
$$\lim_{n \to \infty} \frac{8 \log \triangle_n(\mathcal{H})}{n} = 0.$$
3. *the VC dimension $v(\mathcal{H})$ is finite.*

12.6. Kolmogorov chaining

In this section we introduce Kolmogorov chaining which is a much more efficient way of constructing a cover. In the process we derive Dudley's entropy integral.

We first define an empirical norm.

Definition. *Given $S = \{x_1, ..., x_n\}$ the empirical ℓ_2 norm is*

$$\rho_S(f,g) = \left(n^{-1} \sum_{i=1}^n (f(x_i) - g(x_i))^2 \right)^{1/2}.$$

We can define a cover given the empirical norm

Definition. *Given a hypothesis space $\mathcal{H}$ and the above norm, the covering number $\mathcal{N}(\mathcal{H}, \epsilon, \rho_S)$ is the minimal number $\ell \in \mathbb{N}$ such that for every $f \in \mathcal{H}$ there exists functions $\{g_i\}_{i=1}^{\ell}$ such that*

$$\rho_S(f, g_i) \leq \epsilon \text{ for some } i.$$

The proof of the following theorem is identical to the proof of lemma 12.5.

Theorem. *Given the square loss and $\mathcal{H}$ be a functions such that $-1 \leq f(x) \leq 1$, $y \in [-1, 1]$ and $S = \{z_i\}_{i=1}^n$ drawn i.i.d. then with probability at least $1 - e^{-t/8}$ $(t > 0)$ for the empirical minimizer, f_S,*

$$\mathbb{E}_{x,y}(f_S(x) - y)^2 < n^{-1} \sum_{i=1}^n (f_S(x_i) - y_i)^2 + \sqrt{\frac{(8 \log \mathcal{N}(\mathcal{H}, \varepsilon/8M, \rho_S) + t)}{n}},$$

where $\mathcal{N}(\mathcal{H}, \varepsilon/8M, \rho_S)$ is the empirical cover.

The key idea in the proof of both lemma 12.5 and the above theorem is that

$$\mathbb{P}(|D(f,S)| > \epsilon) \leq 4\, \mathbb{P}\left(\left| n^{-1} \sum_{i=1}^n \sigma_i f(x_i) \right| > \epsilon/4 \right),$$

where

$$D(f,S) := \mathbb{E}_{x,y}(f(x) - y)^2 - n^{-1} \sum_{i=1}^n (f(x_i) - y_i)^2,$$

and σ_i is a Rademacher random variable.

We now prove the chaining theorem.

Theorem. *Given a hypothesis space $\mathcal{H}$ where for all $f \in \mathcal{H}$ $-1 \leq f(x) \leq 1$ if we define*

$$R(f) = n^{-1} \sum_{i=1}^n \sigma_i f(x_i),$$

then

$$\mathbb{P}\left(\forall f \in \mathcal{H},\ \ R(f) \leq \frac{2^{9/2}}{\sqrt{n}} \int_0^{d(0,f)} \log^{1/2} \mathcal{P}(\mathcal{H}, \varepsilon, \rho_S) d\varepsilon + 2^{7/2} d(0,f) \sqrt{\frac{u}{n}} \right) \geq 1 - e^{-u},$$

where $\mathcal{P}(\mathcal{H}, \varepsilon, \rho_S)$ is the empirical packing number and

$$\int_0^{d(0,f)} \log^{1/2} \mathcal{P}(\mathcal{H}, \varepsilon, \rho_S) d\varepsilon$$

is Dudley's entropy interal.

Proof.

Without loss of generality we will assume that the zero function $\{0\}$ is in $\mathcal{H}$. We will construct a nested sets of functions

$$\{0\} = \mathcal{H}_0 \subseteq \mathcal{H}_1 \subseteq \mathcal{H}_2 ... \subseteq \mathcal{H}_j \subseteq ... \mathcal{H}.$$

These subsets will have the following properties

(1) $\forall f, g \in \mathcal{H}_j \;\; \rho_S(f,g) > 2^{-j}$
(2) $\forall f \in \mathcal{H} \;\; \exists \;\; f \in \mathcal{H}_j$ such that $\rho_S(f,g) \leq 2^{-j}$.

Given a set $\mathcal{H}_j$ we can construct $\mathcal{H}_{j+1}$ via the following procedure:

(1) $\mathcal{H}_{j+1} := \mathcal{H}_j$
(2) Find all $f \in \mathcal{H}$ such that for all $g \in \mathcal{H}_{j+1} \;\; \rho_S(f,g) > 2^{-(j+1)}$
(3) Add the above f to $\mathcal{H}_{j+1}$.

We now define a projection operation $\pi_j : \mathcal{H} \to \mathcal{H}_j$ where given $f \in \mathcal{H}$

$$\pi_j(f) := g \text{ where } g \in \mathcal{H}_j \text{ such that } \rho_S(g,f) \leq 2^{-j}.$$

For all $f \in \mathcal{H}$ the following chaining holds

$$\begin{aligned} f &= \pi_0(f) + (\pi_1(f) - \pi_0(f)) + (\pi_2(f) - \pi_1(f)) + ... \\ &= \sum_{j=1}^{\infty} (\pi_j(f) - \pi_{j-1}(f)), \end{aligned}$$

and

$$\begin{aligned} \rho_S(\pi_{j-1}(f), \pi_j(f)) &\leq \rho(\pi_{j-1}(f), f) + \rho_S(\pi_j(f), f) \\ &\leq 2^{-(j-1)} + 2^{-j} = 3 \cdot 2^{-j} \leq 2^{-j+2}. \end{aligned}$$

$R(f)$ is a linear function, so

$$R(f) = \sum_{j=1}^{\infty} (\pi_j(f) - \pi_{j-1}(f)).$$

The set of links in the chain between two levels are defined as follows

$$L_{j-1,j} := \{f - g : f \in \mathcal{H}_j,\, g \in \mathcal{H}_{j-1} \text{ and } \rho_S(f,g) \leq 2^{-j+2}\}.$$

For a fixed link $\ell \in L_{j-1,j}$

$$R(l) = n^{-1} \sum_{i=1}^{n} \sigma_i \ell(x_i),$$

and $|\ell(x_i)| \leq 2^{-j+2}$ so by Hoeffding's inequality

$$\begin{aligned} \mathbb{P}(R(\ell) \geq t) &\leq e^{-nt^2/(\frac{2}{n}\sum_{i=1}^{n} \ell^2(x_i))} \\ &\leq e^{-nt^2/(2 \cdot 2^{-2j+4})}. \end{aligned}$$

The cardinality of the set of links is

$$|L_{j-1,j}| \leq |\mathcal{H}_j| \cdot |\mathcal{H}_{j-1}| \leq (|\mathcal{H}_j|)^2.$$

So

$$\mathbb{P}\left(\forall \ell \in L_{j-1,j},\; R(\ell) \leq t\right) \geq 1 - \left(|\mathcal{H}_j|\right)^2 e^{-nt^2/2^{-2j+5}},$$

setting

$$t = \sqrt{\frac{2^{-2j+5}}{n}(4 \log |\mathcal{H}_j| + u)} \leq \sqrt{\frac{2^{-2j+5}}{n} 4 \log |\mathcal{H}_j|} + \sqrt{\frac{2^{-2j+5} u}{n}},$$

gives us

$$\mathbb{P}\left(\forall \ell \in L_{j-1,j},\; R(\ell) \leq \frac{2^{7/2} 2^{-j} \log^{1/2} |\mathcal{H}_j|}{\sqrt{n}} + 2^{5/2} 2^{-j} \sqrt{\frac{u}{n}}\right) \geq 1 - \frac{1}{|\mathcal{H}_j|} e^{-u}.$$

If $\mathcal{H}_{j-1} = \mathcal{H}_j$ then

$$\pi_{j-1}(f) = \pi_j(f) \text{ and } L_{j-1,j} = \{0\}.$$

So over all levels and links with probability at least $1 - \sum_{j=1}^{\infty} \frac{1}{|\mathcal{H}_j|} e^{-u}$

$$\forall j \geq 1,\ \forall \ell \in L_{j-1,j},\ R(\ell) \leq \frac{2^{7/2} 2^{-j} \log^{1/2} |\mathcal{H}_j|}{\sqrt{n}} + 2^{5/2} 2^{-j} \sqrt{\frac{u}{n}},$$

and

$$1 - \sum_{j=1}^{\infty} \frac{1}{|\mathcal{H}_j|} e^{-u} \geq 1 - \sum_{j=1}^{\infty} \frac{1}{j^2} e^{-u} = 1 - \left(\frac{\pi^2}{6} - 1 \right) e^{-u} \geq 1 - e^{-u}.$$

For some level k

$$2^{-(k+1)} \leq d(0, f) \leq 2^{-k}$$

and

$$0 = \pi_0(f) = \pi_1(f) = \cdots = \pi_k(f).$$

So

$$\begin{aligned} R(f) &= \sum_{j=k+1}^{\infty} R(\pi_j(f) - \pi_{j-1}(f)) \\ &\leq \sum_{j=k+1}^{\infty} \left(\frac{2^{7/2} 2^{-j}}{\sqrt{n}} \log^{1/2} |\mathcal{H}_j| + 2^{5/2} 2^{-j} \sqrt{\frac{u}{n}} \right) \\ &\leq \sum_{j=k+1}^{\infty} \left(\frac{2^{7/2} 2^{-j}}{\sqrt{n}} \log^{1/2} \mathcal{P}(\mathcal{H}, 2^{-j}, \rho_S) \right) + 2^{5/2} 2^{-k} \sqrt{\frac{u}{n}}. \end{aligned}$$

Since $2^{-k} < 2d(f, 0)$ we get the second term in the theorem

$$2^{7/2} d(0, f) \sqrt{\frac{u}{n}}.$$

For the first term

$$\begin{aligned} \sum_{j=k+1}^{\infty} \frac{2^{7/2} 2^{-j}}{\sqrt{n}} \log^{1/2} \mathcal{P}(\mathcal{H}, 2^{-j}, \rho_S) &\leq \frac{2^{9/2}}{\sqrt{n}} \sum_{j=k+1}^{\infty} 2^{-(j+1)} \log^{1/2} \mathcal{P}(\mathcal{H}, 2^{-j}, \rho_S) \\ &\leq \frac{2^{9/2}}{\sqrt{n}} \int_0^{2^{-(k+1)}} \log^{1/2} \mathcal{P}(\mathcal{H}, \varepsilon, \rho_S) d\varepsilon \\ &\leq \frac{2^{9/2}}{\sqrt{n}} \int_0^{d(0,f)} \log^{1/2} \mathcal{P}(\mathcal{H}, \varepsilon, \rho_S) d\varepsilon, \end{aligned}$$

the above quantity is Dudley's entropy integral. □

12.7. Covering numbers and VC dimension

In this section we will show how to bound covering numbers via VC dimension. Covering numbers as we have introduced them have been in general for real-valued functions and not indicator functions.

The notion of VC-dimension and VC classes can be extended to real-valued functions in a variety of mappings. The most standard extension is the notion of VC subgraph classes.

Definition. *A subgraph of function $f(x)$ where $f : \mathcal{X} \to \mathbb{R}$ is the set*

$$\mathcal{F}_f = \{(x,t) \in \mathcal{X} \times \mathbb{R} : 0 \le t \le f(x) \text{ or } f(x) \le t \le 0\}.$$

Definition. *The subgraph of a class of funtions $\mathcal{H}$ are the sets*

$$\mathcal{F} = \{\mathcal{F}_f : f \in \mathcal{H}\}.$$

Definition. *If $\mathcal{F}$ is a VC class of sets then $\mathcal{H}$ is a VC subgraph class of functions and $v(\mathcal{H}) = v(\mathcal{F})$.*

We now show that we can upper-bound the covering number with the empirical ℓ_1 norm with a function of then VC dimension for a hypothesis spaces with finite VC dimension.

Theorem. *Given a VC subgraph class $\mathcal{H}$ where $-1 \le f(x) \le 1$ $\forall f \in \mathcal{H}$ and $x \in \mathcal{X}$ with $v(\mathcal{H}) = d$ and $\rho_S(f,g) = n^{-1}\sum_{i=1}^n |f(x_i) - g(x_i)|$ then*

$$\mathcal{P}(\mathcal{H}, \varepsilon, \rho_s) \le \left(\frac{8e}{\varepsilon}\log\frac{7}{\varepsilon}\right)^d.$$

The bound in the above theorem can be improved to

$$\mathcal{P}(\mathcal{H}, \varepsilon, \rho_s) \le \left(\frac{K}{\varepsilon}\right)^d,$$

however, the proof is more complicated so we prove the weaker statement.

Proof.

Set $m = \mathcal{P}(\mathcal{H}, \varepsilon, \rho_s)$ so $\{f_1, ..., f_m\}$ are ε-separated and each function f_k has its respective subgraph $\mathcal{F}_{f_k}$.

Sample uniformly from $\{x_1, ..., x_n\}$ k elements $\{z_1, ...z_k\}$ and uniformly on $[-1,1]$ k elements $\{t_1, ..., t_k\}$.

We now bound the probability that the subgraphs of two ε-separated functions pick out different subsets of $\{(z_1,t_1), ..., (z_k,t_k)\}$

$$\begin{aligned} & \mathbb{P}\left(\mathcal{F}_{f_k} \text{ and } \mathcal{F}_{f_l} \text{ pick out different subsets of } \{(z_1,t_1), ..., (z_k,t_k)\}\right) \\ = \; & \mathbb{P}\left(\text{at least one } (z_i,t_i) \text{ is picked out by either } \mathcal{F}_{f_k} \text{ or } \mathcal{F}_{f_l} \text{ but not the other}\right) \\ = \; & 1 - \mathbb{P}\left(\text{all } (z_i,t_i) \text{ are picked out by both or none}\right). \end{aligned}$$

The probability that (z_i,t_i) is either picked out by either both $\mathcal{F}_{f_k}, \mathcal{F}_{f_l}$ or by neither

12.8. Symmetrization and Rademacher complexities

In the previous lectures we have considered various complexity measures, such as covering numbers. But what is the right notion of complexity for the learning problem we posed? Consider the covering numbers for a moment. Take a small function class and take its convex hull. The resulting class can be extremely large. Nevertheless, the supremum of the difference of expected and empirical errors will be attained at the vertices, i.e. at the base class. In some sense, the "inside" of the class does not matter. The covering numbers take into account the whole class, and therefore become very large for the convex hull, even though the essential complexity is that of the base class. This suggests that the covering numbers are not the ideal complexity measure. In this lecture we introduce another notion (Rademacher averages), which can be claimed to be the "correct" one. In particular,

the Rademacher averages of a convex hull will be equal to those of the base class. This notion of complexity will be shown to have other nice properties.

Instead of jumping right to the definition of Rademacher Averages, we will take a longer route and show how these averages arise. Results on this topic can be found in the Theory of Empirical Processes, and so we will give some definitions from it.

Let $\mathcal{F}$ be a class of functions. Then $(Z_i)_{i\in\mathcal{I}}$ is a random process indexed by $\mathcal{F}$ if $Z_i(f)$ is a random variable for any i.

As before, μ is a probability measure on Ω, and data $x_1, ..., x_n \sim \mu$. Then μ_n is the empirical measure supported on $x_1, ..., x_n$:

$$\mu_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}.$$

Define $Z_i(\cdot) = (\delta_{x_i} - \mu)(\cdot)$, i.e.

$$Z_i(f) = f(x_i) - \mathbb{E}_\mu(f).$$

Then $Z_1, ..., Z_n$ is an i.i.d. process with 0 mean.

In the previous lectures we looked at the quantity

$$\sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^{n} f(x_i) - \mathbb{E}f \right|, \tag{12.1}$$

which can be written as $n \sup_{f\in\mathcal{F}} |\sum_{i=1}^{n} Z_i(f)|$.

Recall that the difficulty with (12.1) is that we do not know μ and therefore cannot calculate $\mathbb{E}f$. The classical approach of covering $\mathcal{F}$ and using the union bound is too loose.

Proposition. *Symmetrization: If $\frac{1}{n}\sum_{i=1}^{n} f(x_i)$ is close to $\mathbb{E}f$ for data $x_1, ..., x_n$, then $\frac{1}{n}\sum_{i=1}^{n} f(x_i)$ is close to $\frac{1}{n}\sum_{i=1}^{n} f(x_i')$, the empirical average on $x_1', ..., x_n'$ (an independent copy of $x_1, ..., x_n$). Therefore, if the two empirical averarages are far from each other, then empirical error is far from expected error.*

Now fix one function f. Let $\epsilon_1, ..., \epsilon_n$ be i.i.d. Rademacher random variables (taking on values 0 or 1 with probability 1/2). Then

$$\begin{aligned}
\mathbb{P}\left[\left|\sum_{i=1}^{n}(f(x_i) - f(x_i'))\right| \geq t\right] &= \mathbb{P}\left[\left|\sum_{i=1}^{n}\epsilon_i(f(x_i) - f(x_i'))\right| \geq t\right] \\
&\leq \mathbb{P}\left[\left|\sum_{i=1}^{n}\epsilon_i f(x_i)\right| \geq t/2\right] + \mathbb{P}\left[\left|\sum_{i=1}^{n}\epsilon_i f(x_i')\right| \geq t/2\right] \\
&= 2\mathbb{P}\left[\left|\sum_{i=1}^{n}\epsilon_i f(x_i)\right| \geq t/2\right]
\end{aligned}$$

Together with symmetrization, this suggests that controlling $\mathbb{P}\left(|\sum_{i=1}^{n}\epsilon_i f(x_i)| \geq t/2\right)$ is enough to control $\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} f(x_i) - \mathbb{E}f\right| \geq t\right)$. Of course, this is a very simple example. Can we do the same with quantities that are uniform over the class?

Definition. *Suprema of an Empirical process:*

$$Z(x_1, ..., x_n) = \sup_{f\in\mathcal{F}} \left[\mathbb{E}f - \frac{1}{n}\sum_{i=1}^{n} f(x_i)\right].$$

Definition. *Suprema of a Rademacher Process:*

$$R(x_1, ..., x_n, \epsilon_1, ..., \epsilon_n) = \sup_{f \in \mathcal{F}} \left[\frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(x_i) \right].$$

Proposition. *The expectation of the Rademacher process bounds the expectation of the empirical process:*

$$\mathbb{E}Z \leq 2\mathbb{E}R^1.$$

Proof.

$$\begin{aligned}
\mathbb{E}Z &= \mathbb{E}_x \sup_{f \in \mathcal{F}} \left[\mathbb{E}f - \frac{1}{n} \sum_{i=1}^{n} f(x_i) \right] \\
&= \mathbb{E}_x \sup_{f \in \mathcal{F}} \left[\mathbb{E}_{x'} \left(\frac{1}{n} \sum_{i=1}^{n} f(x_i') \right) - \frac{1}{n} \sum_{i=1}^{n} f(x_i) \right] \\
&\leq \mathbb{E}_{x,x'} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (f(x_i') - f(x_i)) \\
&= \mathbb{E}_{x,x',\epsilon} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i (f(x_i') - f(x_i)) \\
&\leq \mathbb{E}_{x,x',\epsilon} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(x_i') + \mathbb{E}_{x,x',\epsilon} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (-\epsilon_i) f(x_i) \\
&= 2\mathbb{E}R \quad \square.
\end{aligned}$$

As we discussed previously, we would like to bound the empirical process Z since this will imply "generalization" for any function in $\mathcal{F}$. We will bound Z by the Rademacher average $\mathbb{E}R$ which we will see has some nice properties.

Theorem. *If the functions in $\mathcal{F}$ are uniformly bounded between $[a, b]$ then with probability $1 - e^{-u}$*

$$Z \leq 2\mathbb{E}R + \sqrt{\frac{2u(b-a)}{n}}.$$

Proof. The inequality involves two steps

(1) the concentration of Z around its mean $\mathbb{E}Z$
(2) applying the bound $\mathbb{E}Z \leq 2\mathbb{E}R$

We will use McDiarmid's inequality for the first step. We define the following two variables $Z := Z(x_1, ..., x_i, ..., x_n)$ and $Z^i := Z(x_1, ..., x_i', ..., x_n)$. Since $a \leq f(x) \leq b$ for all x and $f \in \mathcal{F}$:

$$\begin{aligned}
\left| Z^i - Z \right| &= \left| \sup_{f \in \mathcal{F}} |\mathbb{E}f - n^{-1} \sum_{j=1}^{n} f(x_j) + \left(n^{-1} f(x_i) - n^{-1} f(x_i') \right)| - \sup_{f \in \mathcal{F}} |\mathbb{E}f - n^{-1} \sum_{j=1}^{n} f(x_j)| \right| \\
&\leq \sup_{f \in \mathcal{F}} \frac{1}{n} |f(x_i) - f(x_i')| \leq \frac{b-a}{n} = c_i.
\end{aligned}$$

[1]The quantity $\mathbb{E}R$ is called a *Rademacher average*.

This bounds the Martingale difference for the empirical process. Given the difference bound McDiarmid's inequality states

$$\mathbb{P}\left(Z - \mathbb{E}Z > t\right) \leq \exp\left(\frac{-t^2}{2\sum_{i=1}^n \frac{(b-a)^2}{n^2}}\right) = \exp\left(\frac{-nt^2}{2(b-a)^2}\right).$$

Therefore, with probability at least $1 - e^{-u}$,

$$Z - \mathbb{E}Z < \sqrt{\frac{2u(b-a)}{n}}.$$

So as the number of samples, n, grows, Z becomes more and more concentrated around $\mathbb{E}Z$.

Applying symmetrization proves the theorem. With probability at least $1-e^{-u}$.

$$Z \leq \mathbb{E}Z + \sqrt{\frac{2u(b-a)}{n}} \leq 2\mathbb{E}R + \sqrt{\frac{2u(b-a)}{n}}. \quad \square$$

McDiamid's inequality does not incorporate a notion of variance so it is possible to obtain a sharper inequality using see Talagrand's inequality for the suprema of empirical processes.

We are now left with bounding the Rademacher average. Implicit in the previous lecture on on Kolmogorov chaining was such a bound. Before we restate that result and give some examples we state some nice and useful properties of Rademacher averages.

Properties. *Let $\mathcal{F}$, $\mathcal{G}$ be classes of real-valued functions. Then for any n,*

1. *If $\mathcal{F} \subseteq \mathcal{G}$, then $\mathbb{E}R(\mathcal{F}) \leq \mathbb{E}R(G)$*
2. *$\mathbb{E}R(\mathcal{F}) = \mathbb{E}R(conv\mathcal{F})$*
3. *$\forall c \in \mathbb{R}$, $\mathbb{E}R(c\mathcal{F}) = |c|\mathbb{E}R(\mathcal{F})$*
4. *If $\phi : \mathbb{R} \to \mathbb{R}$ is L-Lipschitz and $\phi(0) = 0$, then $\mathbb{E}R(\phi(\mathcal{F})) \leq 2\,L\,\mathbb{E}R(\mathcal{F})$*
5. *For RKHS balls, $c\left(\sum_{i=1}^\infty \lambda_i\right)^{1/2} \leq \mathbb{E}R(\mathcal{F}_k) \leq C(\sum_{i=1}^\infty \lambda_i)^{1/2}$, where λ_i's are eigenvalues of the corresponding linear operator in the RKHS.*

Theorem. *The Rademacher average is bounded by Dudley's entropy integral*

$$\mathbb{E}_\epsilon R \leq c\frac{1}{\sqrt{n}}\int_0^D \sqrt{\log \mathcal{N}(\epsilon, \mathcal{F}, L_2(\mu_n))}d\epsilon,$$

where $\mathcal{N}$ denotes the covering number.

Example. *Let $\mathcal{F}$ be a class with finite VC-dimension V. Then*

$$\mathcal{N}(\epsilon, \mathcal{F}, L_2(\mu_n)) \leq \left(\frac{2}{\epsilon}\right)^{kV},$$

for some constant k. The entropy integral above is bounded as

$$\begin{aligned}\int_0^1 \sqrt{\log \mathcal{N}(\epsilon, \mathcal{F}, L_2(\mu_n))}d\epsilon &\leq \int_0^1 \sqrt{kV \log 2/\epsilon}\; d\epsilon \\ &\leq k'\sqrt{V}\int_0^1 \sqrt{\log 2/\epsilon}\; d\epsilon \leq k\sqrt{V}.\end{aligned}$$

Therefore, $\mathbb{E}_\epsilon R \leq k\sqrt{\frac{V}{n}}$ for some constant k.