Probability Bounds

John Duchi

This document starts from simple probalistic inequalities (Markov's Inequality) and builds up through several stronger concentration results, developing a few ideas about Rademacher complexity, until we give proofs of the main Vapnik-Chervonenkis complexity for learning theory. Many of these proofs are based on Peter Bartlett's lectures for CS281b at Berkeley or Rob Schapire's lectures at Princeton. The aim is to have one self-contained document some of the standard uniform convergence results for learning theory.

1 Preliminaries

We begin this document with a few (nearly trivial) preliminaries which will allow us to make very strong claims on distributions of sums of random variables.

Theorem 1 (Markov's Inequality)**.** *For a nonnegative random variable X and $t > 0$,*

$$\mathbb{P}[X \geq t] \leq \frac{\mathrm{E}[X]}{t}.$$

Proof For $t > 0$,

$$\mathrm{E}[X] = \int_X x\mathbb{P}(dx) \geq \int_t^\infty x\mathbb{P}(dx) \geq \int_t^\infty t\mathbb{P}(dx) = t\mathbb{P}[X \geq t].$$

□

One very powerful consequence of Markov's Inequality is the Chernoff method, which uses the fact that for any $s \geq 0$,

$$\mathbb{P}(X \geq t) = \mathbb{P}(e^{sX} \geq e^{st}) \leq \frac{\mathrm{E}[e^{sX}]}{e^{st}}. \tag{1}$$

The inequality above is a simple consequence of $e^z > 0$ for all $z \in \mathbb{R}$.

2 Hoeffding's Bounds

Lemma 2.1 (Hoeffding's Lemma)**.** *Given a random variable X, $a \leq X \leq b$, and $\mathrm{E}[X] = 0$, then for any $s > 0$, we have*

$$\mathrm{E}\left[e^{sX}\right] \leq e^{\frac{s^2(b-a)^2}{8}}$$

Proof Given any x such that $a \leq x \leq b$, we can define $\lambda \in [0, 1]$ as

$$\lambda = \frac{b-x}{b-a}.$$

Thus, we see that $(b-a)\lambda = b - x$, so that $x = b - \lambda(b-a) = \lambda a + (1-\lambda)b$. As such, we know that $sx = s\lambda a + s(1-\lambda)b$. So the convexity of $\exp(\cdot)$ implies

$$e^{sx} = e^{\lambda sa + (1-\lambda)sb} \leq \lambda e^{sa} + (1-\lambda)e^{sb} = \frac{b-x}{b-a}e^{sa} + \frac{x-a}{b-a}e^{sb}$$

Using the above and the fact that $\mathrm{E}[X] = 0$,

$$\mathrm{E}\left[e^{sX}\right] \leq \mathrm{E}\left[\frac{b-X}{b-a}e^{sa} + \frac{X-a}{b-a}e^{sb}\right] = \frac{b}{b-a}e^{sa} - \frac{a}{b-a}e^{sb}. \tag{2}$$

Now, we let $p = -\frac{a}{b-a}$ (noting that $a \leq 0$ as $\mathrm{E}[X] = 0$ and hence $p \in [0,1]$), and we have $1 - p = \frac{b}{b-a}$, giving

$$\frac{b}{b-a}e^{sa} - \frac{a}{b-a}e^{sb} = (1-p)e^{sa} + pe^{sb} = (1 - p + pe^{sb-sa})e^{sa}.$$

Solving for a in $p = -\frac{a}{b-a}$, we find that $a = -p(b-a)$, so

$$(1 - p + pe^{sb-sa})e^{sa} = (1 - p + pe^{s(b-a)})e^{-ps(b-a)}.$$

Defining $u = s(b-a)$ and

$$\phi(u) \triangleq -ps(b-a) + \log(1 - p + pe^{s(b-a)}) = -pu + \log(1 - p + pe^{u})$$

and using equation (2), we have that

$$\mathrm{E}\left[e^{sX}\right] \leq e^{\phi(u)}.$$

If we can upper bound $\phi(u)$, then we are done. Of course, by Taylor's theorem, there is some $z \in [0, u]$ such that

$$\phi(u) = \phi(0) + u\phi'(0) + \frac{1}{2}u^2\phi''(z) \leq \phi(0) + u\phi'(0) + \sup_z \frac{1}{2}u^2\phi''(z) \tag{3}$$

Taking derivatives,

$$\phi'(u) = -p + \frac{pe^u}{1-p+pe^u}, \quad \phi''(u) = \frac{pe^u}{1-p+pe^u} - \frac{p^2e^{2u}}{(1-p+pe^u)^2} = \frac{p(1-p)e^u}{(1-p+pe^u)^2}$$

Since $\phi'(0) = -p + p = 0$, and $\phi(0) = 0$, we maximize $\phi''(u)$. Substituting z for e^u, we see that $\phi''(u)$ is concave for $z > 0$ as it is linear over quadratic. Thus

$$\begin{aligned}
\frac{d}{dz}\frac{p(1-p)z}{(1-p+pz)^2} &= \frac{p(1-p)}{(1-p+pz)^2} - \frac{2p^2(1-p)z}{(1-p+pz)^3} \\
&= \frac{p(1-p)(1-p+p+z) - 2p^2(1-p)z}{(1-p+pz)^3} \\
&= \frac{p^2(1-p)z - p^2(1-p)z - p^2(1-p)z + p(1-p)^2}{(1-p+pz)^3} \\
&= \frac{p(1-p)(1-p-zp)}{(1-p+pz)^3}
\end{aligned}$$

so that the critical point is at $z = e^u = \frac{1-p}{p}$. Substituting,

$$\phi''(u) \leq \frac{p(1-p) \cdot \frac{1-p}{p}}{(1-p+p \cdot \frac{1-p}{p})^2} = \frac{(1-p)^2}{4(1-p)^2} = \frac{1}{4}.$$

Using equation (3), it is evident that $\phi(u) \leq \frac{1}{2}u^2 \cdot \frac{1}{4} = \frac{1}{8}s^2(b-a)^2$. This completes the proof of the lemma, as we have

$$\mathrm{E}\left[e^{sX}\right] \leq e^{\phi(u)} \leq e^{\frac{s^2(b-a)^2}{8}}.$$

□

Now we prove a slightly more general result than the standard Hoeffding bound using Chernoff's method.

Theorem 2 (Hoeffding's Inequality)**.** *Suppose that $X_1, \ldots, X_m$ are independent random variables with $a_i \leq X_i \leq b_i$. Then*

$$\mathbb{P}\left(\frac{1}{m}\sum_{i=1}^{m} X_i - \frac{1}{m}\sum_{i=1}^{m} \mathrm{E}[X_i] \geq \varepsilon\right) \leq \exp\left(\frac{-2\varepsilon^2 m^2}{\sum_{i=1}^{m}(b_i - a_i)^2}\right)$$

Proof The proof is fairly straightforward using lemma 2.1. First, define $Z_i = X_i - \mathrm{E}[X_i]$, so that $\mathrm{E}[Z_i] = 0$ (and we can assume without loss of generality that the bound on Z_i is still $[a_i, b_i]$). Then

$$\begin{aligned}
\mathbb{P}\left(\sum_{i=1}^{m} Z_i \geq t\right) &= \mathbb{P}\left(\exp\left(s\sum_{i=1}^{m} Z_i\right) \geq \exp(st)\right) \leq \frac{\mathrm{E}\left[\prod_{i=1}^{m}\exp(sZ_i)\right]}{e^{st}} \\
&= \frac{\prod_{i=1}^{m}\mathrm{E}[\exp(sZ_i)]}{e^{st}} \leq e^{-st}\prod_{i=1}^{m} e^{s^2(b_i-a_i)^2/8} \\
&= \exp\left(\frac{s^2}{8}\sum_{i=1}^{m}(b_i - a_i)^2 - st\right).
\end{aligned}$$

The first line is an application of the Chernoff method and the second the application of lemma 2.1 and the fact that the Z_i are independent.

If we substitute $s = 4t/\left(\sum_{i=1}^{m}(b_i - a_i)^2\right)$, which is evidently > 0, we find that

$$\mathbb{P}\left(\sum_{i=1}^{m} Z_i \geq t\right) \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^{m}(b_i - a_i)^2}\right).$$

Finally, letting $t = \varepsilon m$ gives our result. □

We note that the above proof can be extendend using the union bound and reproving a bound on $\leq -\varepsilon$ (setting $Z_i' = 1 - Z_i$) to give

$$\mathbb{P}\left(\left|\sum_{i=1}^{m} X_i - \sum_{i=1}^{m}\mathrm{E}[X_i]\right| \geq m\varepsilon\right) \leq 2\exp\left(\frac{-2\varepsilon^2 m^2}{\sum_{i=1}^{m}(b_i - a_i)^2}\right).$$

3 McDiarmid's Inequality

This more general result, of which Hoeffding's Inequality can be seen as a special case, is very useful in learning theory and other domains. The statement of the theorem is this:

Theorem 3 (McDiarmid's Inequality)**.** *Let $X = X_1, \ldots, X_m$ be m independent random variables taking values from some set A, and assume that $f : A^m \to \mathbb{R}$ satisfies the following boundedness condition (bounded differences):*

$$\sup_{x_1,\ldots,x_m,\hat{x}_i} |f(x_1, x_2, \ldots, x_i, \ldots, x_m) - f(x_1, x_2, \ldots, \hat{x}_i, \ldots, x_m)| \leq c_i$$

for all $i \in \{1, \ldots, m\}$. Then for any $\varepsilon > 0$, we have

$$\mathbb{P}\left[f(X_1, \ldots, X_m) - \mathrm{E}[f(X_1, \ldots, X_m)] \geq \varepsilon\right] \leq \exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^{m} c_i^2}\right).$$

Proof The proof of this result begins by introducing some notation. First, let $X = \{X_1, \ldots, X_m\}$ and $X_{i:j} = \{X_i, \ldots, X_j\}$. Further, let $Z_0 = \mathrm{E}[f(X)]$, $Z_i = \mathrm{E}[f(X) \mid X_1, \ldots, X_i]$, and (naturally) $Z_m = f(X)$. We now prove the following claim:

Claim 3.1.

$$\mathrm{E}\left[\exp(s(Z_k - Z_{k-1})) \mid X_1, \ldots, X_{k-1}\right] \leq \exp\left(\frac{s^2 c_k^2}{8}\right)$$

Proof of Claim 3.1 First, let

$$\begin{aligned} U_k &= \sup_u \{\mathrm{E}[f(X) \mid X_1, \ldots, X_{k-1}, u] - \mathrm{E}[f(X) \mid X_1, \ldots, X_{k-1}]\} \\ L_k &= \inf_l \{\mathrm{E}[f(X) \mid X_1, \ldots, X_{k-1}, l] - \mathrm{E}[f(X) \mid X_1, \ldots, X_{k-1}]\} \end{aligned}$$

and note that

$$\begin{aligned} U_k - L_k &\leq \sup_{l,u} \{\mathrm{E}[f(X) \mid X_1, \ldots, X_{k-1}, u] - \mathrm{E}[f(X) \mid X_1, \ldots, X_{k-1}, l]\} \\ &\leq \sup_{l,u} \left\{ \int_{y_{k+1:m}} [f(X_{1:k-1}, u, y_{k+1:m}) - f(X_{1:k-1}, l, y_{k+1:m})] \prod_{j=k+1}^{m} p(X_j = y_j) \right\} \\ &\leq \int_{y_{k+1:m}} c_k \prod_{j=k+1}^{m} p(X_j = y_j) = c_k \end{aligned}$$

The second line follows because $X_1, \ldots, X_m$ are independent, and the last line follows by Jensen's inequality and the boundedness condition on f. Thus, $L_k \leq Z_k - Z_{k-1} \leq U_k$, so $Z_k - Z_{k-1} \leq c_k$. By lemma 2.1, as

$$\begin{aligned} \mathrm{E}[Z_k - Z_{k-1} \mid X_{1:k-1}] &= \mathrm{E}_{X_{k:m}}[\mathrm{E}[f(X) \mid X_{1:k}] - \mathrm{E}[f(X) \mid X_{1:k-1}]] \\ &= \mathrm{E}_{X_{k:m}}[\mathrm{E}[f(X) \mid X_{1:k-1}] - \mathrm{E}[f(X) \mid X_{1:k-1}]] = 0, \end{aligned}$$

our claim follows. □

Now we simply proceed through a series of inequalities.

$$\begin{aligned} \mathbb{P}[f(X) - \mathrm{E}[f(X)] \geq \varepsilon] &\leq e^{-s\varepsilon} \mathrm{E}\left[\exp(s(f(X) - \mathrm{E}[f(X)]))\right] \\ &= e^{-s\varepsilon} \mathrm{E}\left[\exp(s(Z_m - Z_{m-1} + Z_{m-1} - Z_0))\right] = e^{-s\varepsilon} \mathrm{E}\left[\exp\left(s \sum_{i=1}^{m} (Z_i - Z_{i=1})\right)\right] \\ &= e^{-s\varepsilon} \mathrm{E}\left[\mathrm{E}\left[\exp\left(s \sum_{i=1}^{m} (Z_i - Z_{i=1})\right) \mid X_{1:m-1}\right]\right] \\ &= e^{-s\varepsilon} \mathrm{E}\left[\exp\left(s \sum_{i=1}^{m-1} (Z_i - Z_{i-1})\right) \mathrm{E}\left[e^{s(Z_m - Z_{m-1})} \mid X_{1:m-1}\right]\right] \\ &\leq e^{-s\varepsilon} \mathrm{E}\left[\exp\left(s \sum_{i=1}^{m-1} (Z_i - Z_{i-1})\right) e^{\frac{s^2 c_m^2}{8}}\right] \\ &\leq e^{-s\varepsilon} \prod_{i=1}^{m} \exp\left(\frac{s^2 c_i^2}{8}\right) = \exp\left(-s\varepsilon + s^2 \sum_{i=1}^{m} \frac{c_i^2}{8}\right) \end{aligned}$$

The third line follows because of the properties of expectation (that is, that $\mathrm{E}[g(X,Y)] = \mathrm{E}_X[\mathrm{E}_Y[g(X,Y) \mid X]]$), the fourth because of our independence assumptions, and the fifth and sixth by repeated applications of claim 3.1.

Minimizing the last equation with respect to s, we take the derivative and find that

$$\frac{d}{ds} \exp\left(-s\varepsilon + s^2 \sum_{i=1}^{m} \frac{c_i^2}{8}\right) = \exp\left(-s\varepsilon + s^2 \sum_{i=1}^{m} \frac{c_i^2}{8}\right) \left(-\varepsilon + 2s \sum_{i=1}^{m} \frac{c_i^2}{8}\right)$$

which has a critical point at $s = 4\varepsilon / \sum_{i=1}^m c_i^2$. Substituting, we see that

$$\exp\left(-s\varepsilon + s^2 \sum_{i=1}^m \frac{c_i^2}{8}\right) = \exp\left(-\frac{4\varepsilon^2}{\sum_{i=1}^m c_i^2} + \frac{16\varepsilon^2 \sum_{i=1}^m c_i^2}{8\left(\sum_{i=1}^m c_i^2\right)^2}\right) = \exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^m c_i^2}\right)$$

This completes our proof. □

As a quick note, it is worth mentioning that Hoeffding's inequality follows by applying McDiarmid's inequality to the function $f(x_1, \ldots, x_m) = \sum_{i=1}^m x_i$.

4 Glivenko-Cantelli Theorem

In this section, we give a proof of the Glivenko-Cantelli theorem, which gives uniform convergence of the empirical distribution function for a random variable X, to the true distribution. There are many ways of proving this; for another example, see [1, Theorem 20.6]. Our proof makes use of Rademacher random variables and gives a convergence rate as well. First, though, we need a Borel-Cantelli lemma.

Theorem 4 (Borel-Cantelli Lemma I). *Let A_n be a sequence of subsets of some probability space Ω. If $\sum_{n=1}^\infty \mathbb{P}(A_n) < \infty$, then $\mathbb{P}(A_n \text{ i.o.}) = 0$.*

Proof Let $N = \sum_{n=1}^\infty \mathbf{1}\{A_n\}$, the number of events that occur. By Fubini's theorem, we have $\mathbb{E}N = \sum_{n=1}^\infty \mathbb{P}(A_n) < \infty$, so $N < \infty$ almost surely. □

Now for the real proof. Let $F_n(x)$ be the empirical distribution function of a sequence of i.i.d. random variables $X_1, \ldots, X_n$, that is,

$$F_n(x) \triangleq \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{X_i \le x\}$$

and let F be the true distribution. We have the following theorem.

Theorem 5 (Glivenko-Cantelli). *As $n \to \infty$, $\sup_x |F_n(x) - F(x)| \overset{\text{a.s.}}{\to} 0$.*

Proof Our proof has three main parts. First is McDiarmid's concentration inequality, then we symmetrize using Rademacher random variables, and lastly we show the class of functions we use is "small" by ordering the data we see.

To begin, define the function class

$$\mathcal{G} \triangleq \{g : x \mapsto \mathbf{1}\{x \le \theta\}, \theta \in \mathbb{R}\}$$

and note that there is a one-to-one mapping between $\mathcal{G}$ and $\mathbb{R}$. Now, define $\mathbb{E}_n g = \frac{1}{n}\sum_{i=1}^n g(X_i)$. The theorem is equivalent to $\sup_{g\in\mathcal{G}} |\mathbb{E}_n g - \mathbb{E}g| \overset{\text{a.s.}}{\to} 0$ for any probability measure $\mathbb{P}$, and the rates of convergence will be identical.

We begin with the concentration result. Let $f(X_1, \ldots, X_n) = \sup_{g\in\mathcal{G}} |\mathbb{E}_n g - \mathbb{E}g|$, and note that changing any 1 of the n data points arbitrarily makes the empirical distribution g change by at most $1/n$. Thus, McDiarmid's inequality (theorem 3) implies that

$$\sup_{g\in\mathcal{G}} |\mathbb{E}_n g - \mathbb{E}g| \le \mathbb{E}\sup_{g\in\mathcal{G}} |\mathbb{E}_n g - \mathbb{E}g| + \varepsilon \tag{4}$$

with probability at least $1 - e^{-2\varepsilon^2 n}$.

We now use Rademacher random variables and symmetrization to get a handle on the term

$$\mathbb{E}\sup_{g\in\mathcal{G}} |\mathbb{E}_n g - \mathbb{E}g| \tag{5}$$

It is hard to directly show that this converges to zero, but we can use symmetrization to upper bound Eq. (5). To this end, let $Y_1, \ldots, Y_n$ be n independent copies of $X_1, \ldots, X_n$. We have

$$\begin{aligned}
\mathbb{E}\sup_{g\in\mathcal{G}}|\mathbb{E}_n g - \mathbb{E}g| &= \mathbb{E}\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^n g(X_i) - \frac{1}{n}\sum_{i=1}^n \mathbb{E}g(Y_i)\right| \\
&= \mathbb{E}\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^n (g(X_i) - \mathbb{E}[g(Y_i) \mid X_1, \ldots, X_n])\right| \\
&= \mathbb{E}\sup_{g\in\mathcal{G}}\left|\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n (g(X_i) - g(Y_i)) \mid X_1, \ldots, X_n\right]\right| \\
&\le \mathbb{E}\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^n (g(X_i) - g(Y_i))\right| \mid X_1, \ldots, X_n\right] = \mathbb{E}\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^n g(X_i) - g(Y_i)\right|.
\end{aligned}$$

The second and third lines are almost sure and follow by properties of conditional expectation, and the last inequality follows via convexity of $|\cdot|$ and sup.

We now proceed to remove dependence on $g(Y_i)$ by the following steps. First, note that $g(X_i) - g(Y_i)$ is symmetric around 0, so if $\sigma_i \in \{-1, 1\}$, $\sigma_i(g(X_i) - g(Y_i))$ has identical distribution. Thus, we can continue our inequalities:

$$\begin{aligned}
\mathbb{E}\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^n g(X_i) - g(Y_i)\right| &= \mathbb{E}\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i(g(X_i) - g(Y_i))\right| \le \mathbb{E}\sup_{g\in\mathcal{G}}\left[\left|\frac{1}{n}\sum_{i=1}^n \sigma_i g(X_i)\right| + \left|\frac{1}{n}\sum_{i=1}^n \sigma_i g(Y_i)\right|\right] \\
&\le \mathbb{E}\left[\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i g(X_i)\right| + \sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i g(Y_i)\right|\right] \\
&= 2\mathbb{E}\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i g(X_i)\right|. \qquad (6)
\end{aligned}$$

The last expectation has a maximum inner product between the vectors $[\sigma_1 \cdots \sigma_n]^\top$ and $[g(X_1) \cdots g(X_n)]^\top$. This an indication of how well the class of vectors $\{[g(X_1) \cdots g(X_n)]^\top : g \in \mathcal{G}\}$ can align with random directions $\sigma_1, \ldots, \sigma_n$, which are uniformly distributed on the corners of the n-cube.

Now what remains is to bound $\mathbb{E}\sup_{g\in\mathcal{G}}|\sum_i \sigma_i g(X_i)|$, which we do by noting that $\mathcal{G}$ is in a sense simple. Fix the data set as $(x_1, \ldots, x_n)$ and consider the order statistics $x_{(1)}, \ldots, x_{(n)}$. Note that $x_{(i)} \le x_{(i+1)}$ implies that $g(x_{(i)}) = \mathbf{1}\left\{x_{(i)} \le \theta\right\} \ge \mathbf{1}\left\{x_{(i+1)} \le \theta\right\} = g(x_{(i+1)})$, so

$$[g(x_{(1)}) \cdots g(x_{(n)})]^\top \in \left\{[0 \cdots 0]^\top, [1\ 0 \cdots 0]^\top, \ldots, [1 \cdots 1]^\top\right\}.$$

The bijection between $(x_1, \ldots, x_n)$ and $(x_{(1)}, \ldots, x_{(n)})$ implies that the cardinality of the set $\{[g(x_1), \ldots, g(x_n)]^\top : g \in \mathcal{G}\}$ is at most $n+1$. We can thus use bounds relating the size of the class of functions to bound the earlier expectations.

Lemma 4.1 (Massart's finite class lemma)**.** *Let $A \subset \mathbb{R}^n$ have $|A| < \infty$, $R = \max_{a\in A}\|a\|$, and σ_i be independent Rademacher variables. Then*

$$\mathbb{E}\max_{a\in A}\left[\frac{1}{n}\sum_{i=1}^n \sigma_i a_i\right] \le \frac{R\sqrt{2\log|A|}}{n}$$

Proof of Lemma Let $s > 0$ and define $Z_a \triangleq \sum_{i=1}^n \sigma_i a_i$. Because exp is convex and positive,

$$\exp\left(s\mathbb{E}\max_{a\in A} Z_a\right) \le \mathbb{E}\exp\left(s\max_{a\in A} Z_a\right) = \mathbb{E}\max_{a\in A}\exp(sZ_a) \le \mathbb{E}\sum_{a\in A}\exp(sZ_a).$$

Now we apply Hoeffding's lemma (lemma 2.1) by noting that $\sigma_i a_i \in [-a_i, a_i]$, and have

$$\sum_{a\in A} \mathbb{E}\exp(sZ_a) \le \sum_{a\in A} \exp\left(s^2 \sum_{i=1}^n a_i^2/2\right) \le \sum_{a\in A} \exp(s^2R^2/2) = |A|\exp(s^2R^2/2).$$

Combining the above bound with the first string of inequalities, we have $\exp(s\mathbb{E}\max_{a\in A} Z_a) \le |A|\exp(s^2R^2/2)$, or

$$\mathbb{E}\max_{a\in A} Z_a \le \inf_{s>0}\left(\frac{\log|A|}{s} + \frac{sR^2}{2}\right)$$

Setting $s = \sqrt{2\log|A|/R^2}$, we have

$$\mathbb{E}\max_{a\in A} Z_a \le \left(\frac{R\log|A|}{\sqrt{2\log|A|}} + \frac{R\sqrt{2\log|A|}}{2}\right) = R\sqrt{2\log|A|}$$

and dividing by n gives the final bound. □

Now, letting $A = \{[g(X_1), \dots, g(X_n)]^\top : g \in \mathcal{G}\}$, we note that $|A| \le n+1$ and $R = \max_{a\in A}\|a\| \le \sqrt{n}$. Thus,

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i g(X_i)\right|\right] = \mathbb{E}\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i g(X_i)\right| \mid X_1,\dots,X_n\right] \le \sqrt{\frac{2\log(n+1)}{n}}.$$

By the above equation, Eq. (6), and the McDiarmid inequality application in Eq. (4),

$$\begin{aligned}\mathbb{P}\left(\sup_{g\in\mathcal{G}}|\mathbb{E}_n g - \mathbb{E}g| > \varepsilon + 2\sqrt{\frac{2\log(n+1)}{n}}\right) &\le \mathbb{P}\left(\sup_{g\in\mathcal{G}}|\mathbb{E}_n g - \mathbb{E}g| > \varepsilon + 2\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i g(X_i)\right|\right]\right)\\ &\le \mathbb{P}\left(\sup_{g\in\mathcal{G}}|\mathbb{E}_n g - \mathbb{E}g| > \varepsilon + \mathbb{E}\sup_{g\in\mathcal{G}}|\mathbb{E}_n g - \mathbb{E}g|\right)\\ &\le 2\exp(-2\varepsilon^2 n).\end{aligned}$$

This implies that $\sup_{g\in\mathcal{G}}|\mathbb{E}_n g - \mathbb{E}g| \xrightarrow{p} 0$. To get almost sure convergence, choose n so that $2\sqrt{2\log(n+1)/n} < \varepsilon$ and let $A_n = \{\sup_{g\in\mathcal{G}}|\mathbb{E}_n g - \mathbb{E}g| > 2\varepsilon\}$. Then $\sum\mathbb{P}(A_n) < \infty$, so A_n happens only finitely many times. □

5 Rademacher Averages

Now we explore uses of Rademacher random variabels to measure the complexity of a class of functions. We also use them to masure (to some extent) generalization ability of a function from data to the true distribution.

Definition 5.1 (Rademacher complexity)**.** *Let $\mathcal{F}$ be a function class with domain X, i.e. $\mathcal{F} \subseteq \{f : \mathcal{X} \to \mathbb{R}\}$, and let $S = \{X_1, \dots, X_n\}$ be a set of samples generated by a distribution $\mathbb{P}$ on X. The* **empirical Rademacher complexity** *of $\mathcal{F}$ is*

$$\hat{R}_n(\mathcal{F}) = \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i f(X_i)\right| \mid X_1,\dots,X_n\right]$$

where σ_i i.i.d. uniform random variables (Rademacher variables) on ±1. The **Rademacher complexity** *of $\mathcal{F}$ is*

$$R_n(\mathcal{F}) = \mathbb{E}\hat{R}_n(\mathcal{F}) = \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i f(X_i)\right|\right].$$

Lemma 5.1.

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\mathbb{E}f - \frac{1}{n}\sum_{i=1}^{n} f(X_i)\right|\right] \leq 2R_n(\mathcal{F})$$

Proof As we did for the Glivenko-Cantelli theorem, introduce i.i.d. random variables Y_i, $i \in \{1, \dots, n\}$ independent of the X_is. Letting $\mathbb{E}_Y$ denote expectation with respect to the Y_i and σ_i be Rademacher variables,

$$\begin{aligned}
\mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}f - f(X_i)\right| &= \mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}f(Y_i) - f(X_i)\right| \leq \mathbb{E}_X\mathbb{E}_Y\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} g(X_i) - g(Y_i)\right| \\
&= \mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i(g(X_i) - g(Y_i))\right| \\
&\leq \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i g(X_i)\right| + \sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i g(Y_i)\right|\right] = 2R_n(\mathcal{F}).
\end{aligned}$$

The first inequality follows from the convexity of $|\cdot|$ and sup, the second by the triangle inequality. □

We can also use Rademacher complexity to bound the expected value of certain functions, which is often used in conjunction with loss functions or expected risks. For example, we have the following theorem dealing with bounded functions. Recall that $\mathbb{E}_n f(X) = \frac{1}{n}\sum_{i=1}^{n} f(X_i)$, where the X_i are given as a sample.

Theorem 6. *Let $\delta \in (0,1)$ and $\mathcal{F}$ be a class of functions mapping X to $[0,1]$. Then with probability at least $1-\delta$, all $f \in \mathcal{F}$ satisfy.*

$$\mathbb{E}f(X) \leq \mathbb{E}_n f(X) + 2R_n(\mathcal{F}) + \sqrt{\frac{\log 1/\delta}{2n}}$$

Also with probability at least $1-\delta$, all $f \in \mathcal{F}$ satisfy

$$\mathbb{E}f(X) \leq \mathbb{E}_n f(X) + 2\hat{R}_n(\mathcal{F}) + 5\sqrt{\frac{\log 2/\delta}{2n}}.$$

Proof Fix $f \in \mathcal{F}$. Then we clearly have (by choosing f in the sup)

$$\mathbb{E}f(X) \leq \mathbb{E}_n f(X) + \sup_{g\in\mathcal{F}}\left(\mathbb{E}g(X) - \mathbb{E}_n g(X)\right).$$

As $f(X_i) \in [0,1]$, modifying one of the X_i can change $\mathbb{E}_n g(X)$ by at most $1/n$. McDiarmid's inequality (Theorem 3) thus implies that

$$\mathbb{P}\left(\sup_{g\in\mathcal{F}}(\mathbb{E}g(X) - \mathbb{E}_n g(X)) - \mathbb{E}[\sup_{g\in\mathcal{F}}(\mathbb{E}g(X) - \mathbb{E}_n g(X))] \geq \varepsilon\right) \leq \exp(-2\varepsilon^2 n).$$

Setting the right hand side bound equal to δ and solving for ε, we have

$$-2\varepsilon^2 n = \log\delta \quad \text{or} \quad \frac{1}{2}\frac{\log 1/\delta}{n} = \varepsilon^2 \quad \text{so} \quad \varepsilon = \sqrt{\frac{\log 1/\delta}{2n}}.$$

That is, with probability at least $1-\delta$, we have

$$\sup_{g\in\mathcal{F}}\left(\mathbb{E}g(X) - \mathbb{E}_n g(X)\right) \leq \mathbb{E}\left[\sup_{g\in\mathcal{F}}\left(\mathbb{E}g(X) - \mathbb{E}_n g(X)\right)\right] + \sqrt{\frac{\log 1/\delta}{2n}}.$$

Applying lemma 5.1, we immediately see that the right hand expectation is bounded by $R_n(\mathcal{F})$. This completes the proof of the first inequality in the theorem.

Now we need to bound $R_n(\mathcal{F})$ with high probability using $\hat{R}_n(\mathcal{F})$. First note that the above reasoning could have been done using probability $\delta/2$, giving a bound with $2R_n(\mathcal{F})$ and $\sqrt{\log(2/\delta)/(2n)}$ instead. Now we note that changing one example X_i changes $|R_n(\mathcal{F}) - \hat{R}_n(\mathcal{F})|$ by at most $2/n$ (because one sign inside of $\hat{R}_n(\mathcal{F})$ can change). Letting $c_i = 2/n$ in McDiarmid's inequality, we have

$$\mathbb{P}(\hat{R}_n(\mathcal{F}) - R_n(\mathcal{F}) \geq \varepsilon) \leq \exp\left(-\frac{1}{2}\varepsilon^2 n\right).$$

Again setting this equal to $\delta/2$ and solving, we have $\frac{1}{2}\varepsilon^2 = \log(2/\delta)/n$ so that $\varepsilon = 2\sqrt{\log(2/\delta)/(2n)}$. We thus have with probability $\geq 1 - \delta/2$,

$$R_n(\mathcal{F}) \leq \hat{R}_n(\mathcal{F}) + 2\sqrt{\frac{\log 2/\delta}{2n}}.$$

Using the union bound with two events of probability at least $1-\delta/2$ gives the desired second inequality. □

Theorem 7 (Ledoux-Talagrand contraction)**.** *Let $f : \mathbb{R}_+ \to \mathbb{R}_+$ be convex and increasing. Let $\phi_i : \mathbb{R} \to \mathbb{R}$ satisfy $\phi_i(0) = 0$ and be Lipschitz with constant L, i.e., $|\phi_i(a) - \phi_i(b)| \leq L|a-b|$. Let σ_i be independent Rademacher random variables. For any $T \subseteq \mathbb{R}^n$,*

$$\mathbb{E}f\left(\frac{1}{2}\sup_{t\in T}\left|\sum_{i=1}^n \sigma_i\phi_i(t_i)\right|\right) \leq \mathbb{E}f\left(L\cdot\sup_{t\in T}\left|\sum_{i=1}^n \sigma_i t_i\right|\right).$$

Proof First, note that if T is unbounded, there will be some setting of σ_i so that $\sup_{t\in T}|\sum_{i=1}^n \sigma_i t_i| = \infty$. This event is not a probability zero event, and f is increasing and convex and so will also be infinite, so the right expectation will be infinite. We can thus focus on bounded T.

We begin by showing a similar statement to the proof, that is, that if $g : \mathbb{R} \to \mathbb{R}$ is convex and increasing, then

$$\mathbb{E}g\left(\sup_{t\in T}\sum_{i=1}^n \sigma_i\phi_i(t_i)\right) \leq \mathbb{E}g\left(L\sup_{t\in T}\sum_{i=1}^n \sigma_i t_i\right) \tag{7}$$

By conditioning, we note that if we prove for $T \subseteq \mathbb{R}^2$

$$\mathbb{E}g\left(\sup_{t\in T}(t_1 + \sigma_2\phi_2(t_2))\right) \leq \mathbb{E}g\left(\sup_{t\in T}(t_1 + L\sigma_2 t_2)\right) \tag{8}$$

we are done. This follows because we will almost surely have

$$\mathbb{E}\left[g\left(\sup_{t\in T}(\sigma_1\phi_1(t_1) + \sigma_2\phi_2(t_2))\right) \mid \sigma_1\right] \leq \mathbb{E}\left[g\left(\sup_{t\in T}(\sigma_1\phi_1(t_1) + L\sigma_2 t_2)\right) \mid \sigma_1\right]$$

as $\sigma_1\phi_1(t_1)$ simply transforms T (and is still bounded). By conditioning, this implies that

$$\mathbb{E}g\left(\sup_{t\in T}(\sigma_1\phi_1(t_1) + \sigma_2\phi_2(t_2))\right) \leq \mathbb{E}g\left(\sup_{t\in T}(\sigma_1\phi_1(t_1) + L\sigma_2 t_2)\right)$$

and we can iteratively apply this.

Thus we focus on proving Eq. (8). Define $I(t,s) \triangleq \frac{1}{2}g(t_1 + \phi(t_2)) + \frac{1}{2}g(s_1 - \phi(s_2)$; if we show that the right side of Eq. (8) is larger than $I(t,s)$ for all $t, s \in T$, clearly we are done (as it is the expectation with respect to the Rademacher random variable σ_2). Noting that we are taking a supremum over t and s in I, we can assume w.l.o.g. that

$$t_1 + \phi(t_2) \geq s_1 + \phi(s_2) \quad \text{and} \quad s_1 - \phi(s_2) \geq t_1 - \phi(t_2). \tag{9}$$

We define four quantities and then proceed through four cases to prove Eq. (8):

$$a = s_1 - \phi(s_2), \quad b = s_1 - Ls_2, \quad a' = t_1 + Lt_2, \quad b' = t_1 + \phi(t_2).$$

We would like to show that $2I(t,s) = g(a) + g(b') \leq g(a') + g(b)$.

Case I. Let $t_2 \geq 0$ and $s_2 \geq 0$. We know that, as $\phi(0) = 0$, $|\phi(s_2)| \leq Ls_2$. This implies that $a \geq b$ and Eq. (9) implies that $b' = t_1 + \phi(t_2) \geq s_1 + \phi(s_2) \geq s_1 - Ls_2 = b$. Now assume that $t_2 \geq s_2$. In this case,

$$b' + a - b = t_1 + \phi(t_2) + s_1 - \phi(s_2) - s_2 + Ls_2 \leq t_1 + Lt_2 + Ls_2 - \phi(s_2) \leq t_1 + Lt_2 = a'$$

since $|\phi(t_2) - \phi(s_2)| \leq L|t_2 - s_2| = L(t_2 - s_2)$. Thus $a - b \leq a' - b'$. Note that $g(y + x) - g(y)$ is increasing in y if $x \geq 0$.[1] Letting $x = a - b \geq 0$ and noting that $b' \geq b$,

$$g(a) - g(b) = g(b + x) - g(b) \leq g(b' + x) - g(b') = g(b' + a - b) - g(b') \leq g(a') - g(b')$$

so that $g(a) + g(b') \leq g(a') + g(b)$. If $s_2 \geq t_2$, then we use $-\phi$ instead of ϕ and switch the roles of s and t, giving a similar proof.

Case II. Let $t_2 \leq 0$ and $s_2 \leq 0$. This is similar to the above case, switching signs as necessary, so we omit it.

Case III. Let $t_2 \geq 0$ and $s_2 \leq 0$. We have $\phi(t_2) \leq Lt_2$ and $-\phi(s_2) \leq -Ls_2$ by the Lipschitz condition on ϕ. This implies that

$$2I(t,s) = g(t_1 + \phi(t_2)) + g(s_1 - \phi(s_2)) \leq g(t_1 + Lt_2) + g(s_1 - Ls_2).$$

Case IV. Let $t_2 \leq 0$ and $s_2 \geq 0$. Similar to above, we have $-\phi(s_2) \leq Ls_2$ and $\phi(t_2) \leq -Lt_2$, so $2I(t,s) =\leq g(t_1 - Lt_2) + g(s_1 + Ls_2)$, which is symmetric to the above. We have thus proved Eq. (7).

We now conclude the proof. Denoting $[x]_+ = x$ if $x \geq 0$ and $[x]_- = -x$ if $x \leq 0$, we note that since f is increasing and convex that

$$f\left(\frac{1}{2}\sup_{x\in X}|x|\right) = f\left(\frac{1}{2}\sup_{x\in X}([x]_+ + [x]_-)\right) \leq f\left(\frac{1}{2}\sup_{x\in X}[x]_+ + \frac{1}{2}\sup_{x\in X}[x]_-\right) \leq \frac{1}{2}f\left(\sup_{x\in X}[x]_+\right) + \frac{1}{2}f\left(\sup_{x\in X}[x]_-\right).$$

The above equation implies

$$\begin{aligned} \mathbb{E}f\left(\frac{1}{2}\sup_{t\in T}\left|\sum_{i=1}^n \sigma_i\phi_i(t_i)\right|\right) &\leq \frac{1}{2}\mathbb{E}f\left(\sup_{t\in T}\left[\sum_{i=1}^n \sigma_i\phi_i(t_i)\right]_+\right) + \frac{1}{2}\mathbb{E}f\left(\sup_{t\in T}\left[\sum_{i=1}^n \sigma_i\phi_i(t_i)\right]_-\right) \\ &\leq \mathbb{E}f\left(\sup_{t\in T}\left[\sum_{i=1}^n \sigma_i\phi_i(t_i)\right]_+\right). \end{aligned}$$

The last step uses the symmetry of σ_i and the fact that $[-x]_- = [x]_+$.

Finally, note that $f([\cdot]_+)$ is convex, increasing on $\mathbb{R}$, and $f(\sup_x [x]_+) = f([\sup_x x]_+)$. Applying Eq. (7), we have

$$\mathbb{E}f\left(\left[\sup_{t\in T}\sum_{i=1}^n \sigma_i\phi_i(t_i)\right]_+\right) \leq \mathbb{E}f\left(\left[L\sup_{t\in T}\sum_{i=1}^n \sigma_i t_i\right]_+\right) \leq \mathbb{E}f\left(L\sup_{t\in T}\left|\sum_{i=1}^n \sigma_t t_i\right|\right).$$

□

This is a simple extension of Theorem 4.13 of [2], but I include the entire theorem here because its proof is somewhat interesting, and it is often cited. For instance, it gives the following corollary:

[1] To see this, note that the slope of g (the right or left derivative or the subgradient set) is always increasing, so for $x, d > 0$, we have $g(y + d + x) - g(y + x) \geq g(y + d) - g(y)$.

Corollary 5.1. *Let ϕ be an L-Lipschitz map from $\mathbb{R}$ to $\mathbb{R}$ with $\phi(0) = 0$ and $\mathcal{F}$ be a function class with domain X. Let $\phi \circ \mathcal{F} = \{\phi \circ f : f \in \mathcal{F}\}$ denote their composition. Then*

$$R_n(\phi \circ \mathcal{F}) \leq 2LR_n(\mathcal{F}).$$

The corollary follows by taking the convex increasing function in Theorem 7 to be the identity and letting the space $T \subseteq \mathbb{R}^n = \{f(x) : f \in \mathcal{F}, x \in X\}$.

6 Growth Functions, VC Theory, and Rademacher Complexity

In this section, we will be using a sample set $S = (x_1, \dots, x_n)$, a hypothesis class $\mathcal{H}$ of functions mapping the sample space X to $\{-1, 1\}$. We focus on what is known as the growth function $\Pi_{\mathcal{H}}$. We define $\Pi_{\mathcal{H}}(S)$ to be the set of dichotomies of $\mathcal{H}$ on the set S, that is,

$$\Pi_{\mathcal{H}}(S) \triangleq \{\langle h(x_1), \dots, h(x_n)\rangle : h \in \mathcal{H}\}.$$

With this, we make the following definition

Definition 6.1. *The* **growth function** $\Pi_{\mathcal{H}}(n)$ *of a hypothesis class $\mathcal{H}$ is the number of dichotomies of the hypothesis class $\mathcal{H}$ on a sample of size S, that is,*

$$\Pi_{\mathcal{H}}(n) \triangleq \max_{S:|S|=n} |\Pi_{\mathcal{H}}(S)|$$

Clearly, we have $\Pi_{\mathcal{H}}(n) \leq |\mathcal{H}|$ and $\Pi_{\mathcal{H}}(n) \leq 2^n$. With the growth function in mind, we can bound the Rademacher complexity of certain function classes.

Lemma 6.1. *Let $\mathcal{H}$ be a class of functions mapping from X to $\{-1, 1\}$. If $\mathcal{H}$ satisfies $h \in \mathcal{H} \Rightarrow -h \in \mathcal{H}$,*

$$R_n(\mathcal{H}) \leq \sqrt{\frac{2\log \Pi_{\mathcal{H}}(n)}{n}}.$$

If $\mathcal{H}$ does not satisfy $h \in \mathcal{H}$ implies $-h \in \mathcal{H}$,

$$R_n(\mathcal{H}) \leq \sqrt{\frac{2\log 2\Pi_{\mathcal{H}}(n)}{n}}.$$

Proof Note that if we let $A = \{[h(X_1) \cdots h(X_n)]^\top : h \in \mathcal{H}\}$ and $-A = \{-a : a \in A\}$, then

$$\sup_{h\in\mathcal{H}} \left|\sum_{i=1}^n \sigma_i h(X_i)\right| = \max_{a\in A} \left|\sum_{i=1}^n \sigma_i a_i\right| = \max_{a\in A\cup -A} \sum_{i=1}^n \sigma_i a_i$$

so that $\|a\| = \sqrt{n}$ for $a \in A$ and Massart's finite class lemma (lemma 4.1) imply

$$\mathbb{E}\left[\sup_{h\in\mathcal{H}} \left|\frac{1}{n}\sum_{i=1}^n \sigma_i h(X_i)\right| \mid X_1, \dots, X_n\right] \leq \frac{\sqrt{n}\sqrt{2\log\left(2|\{[h(X_1)\cdots h(X_n)]^\top : h \in \mathcal{H}\}|\right)}}{n} \leq \sqrt{\frac{2\log 2\Pi_{\mathcal{H}}(n)}{n}}.$$

Thus $\mathbb{E}\hat{R}_n(\mathcal{H}) = R_n(\mathcal{H})$ implies the theorem. □

Definition 6.2. *A hypothesis class $\mathcal{H}$* **shatters** *a finite set $S \subseteq X$ if $|\Pi_{\mathcal{H}}(S)| = 2^{|S|}$.*

Intuitively, $\mathcal{H}$ shatters a set S if for every labeling of S, there is an $h \in \mathcal{H}$ that realizes that labeling. This notion of shattering leads us to a new notion of the complexity of a hypothesis class.

Definition 6.3. *The* **Vapnik-Chervonenkis dimension***, or* **VC dimension** *of a hypothesis class $\mathcal{H}$ on a set X is the cardinality of the largest set shattered by $\mathcal{H}$, that is, the largest n such that there exists a set $S \subseteq X$, $|S| = n$ that $\mathcal{H}$ shatters.*

As a shorthand, we will use $d_{\mathrm{VC}}(\mathcal{H})$ to denote the VC-dimension of a class of functions.

Theorem 8 (Sauer's lemma)**.** *Let $\mathcal{H}$ be a class of functions mapping X to $\{-1, 1\}$ and let $d_{\mathrm{VC}}(\mathcal{H}) = d$. Then*

$$\Pi_{\mathcal{H}}(n) \leq \sum_{i=0}^{d} \left(\begin{array}{c} n \\ i \end{array} \right)$$

and for $n \geq d$,

$$\Pi_{\mathcal{H}}(n) \leq \left(\frac{en}{d} \right)^d .$$

Proof The proof of Sauer's lemma is a completely combinatorial argument. We prove the lemma by induction on the sum $n + d$, beginning from $n = 0$ or $d = 0$ as our base cases. For notational convenience, we first define $\Phi_d(n) = \sum_{i=0}^{d} {}_nC_i$.

Suppose that $n = 0$. Then $\Pi_{\mathcal{H}}(n) = \Pi_{\mathcal{H}}(0) = 1$, the degenerate labeling of the empty set, and $\Phi_d(0) = {}_0C_0 = 1$. When $d = 0$, no datasets can be shattered at all, so $\Pi_{\mathcal{H}}(S)$ is simply a labeling from one function and $\Pi_{\mathcal{H}}(n) = 1$.

Now we assume that for any n', d' with $n' + d' < n + d$, the first inequality holds. We want to construct hypothesis spaces $\mathcal{H}_i$ that are smaller than $\mathcal{H}$ so that we can use our inductive hypothesis. To this end, we represent the labelings of $\mathcal{H}$ as a table and perform operations on said table. So let $S = \{x_1, \ldots, x_n\}$ be the dataset, and let $S_1 = \{x_1, \ldots, x_{n-1}\}$ be S shrunk by removing x_n. Now let $\mathcal{H}_1$ be the set of hypotheses restricted to S_1, as in Fig. 1. We see that $d_{\mathrm{VC}}(\mathcal{H}_1) \leq d_{\mathrm{VC}}(\mathcal{H})$, because any set that $\mathcal{H}_1$ shatters $\mathcal{H}$ must be able to shatter. Thus, by induction, we have $|\Pi_{\mathcal{H}_1}(S_1)| \leq \Phi_d(n-1)$.

Now let $\mathcal{H}_2$ be the collection of hypotheses that were "collapsed" going from $\mathcal{H}$ to $\mathcal{H}_1$. In the example of Fig. 1, $\mathcal{H}_2$ is h_1 and h_4, as they were collapsed. In particular, the collapsed hypotheses have that in mcH, there was an $h \in \mathcal{H}$ with $h(x_n) = 1$ and another $h \in \mathcal{H}$ with $h(x_n) = -1$, wherease un-collapsed hypotheses do not have this. The hypotheses are also restricted to $S_2 = S_1$, and $|\Pi_{\mathcal{H}_2}(S_2)| = |\mathcal{H}_2|$. Since the original $\mathcal{H}$ had hypotheses to label x_n as ±1, any dataset T that $\mathcal{H}_2$ shatters will also have $T \cup \{x_n\}$ shattered by $\mathcal{H}$, but $\mathcal{H}_2$ cannot shatter $T \cup \{x_n\}$ as the dichotomies on x_n were collapsed. In other words, the VC-dimension of $\mathcal{H}$ is strictly greater than that of $\mathcal{H}_2$, so that $d_{\mathrm{VC}}(\mathcal{H}_2) \leq d - 1$. By the inductive hypothesis, $|\Pi_{\mathcal{H}_2}(S_2)| \leq \Phi_{d-1}(n-1)$.

	$\mathcal{H}$							$\mathcal{H}_1$						$\mathcal{H}_2$			
	$x_1, \ldots, x_n$							$x_1, \ldots, x_{n-1}$						$x_1, \ldots, x_{n-1}$			
h_1	-1	1	1	-1	-1	$\rightarrow$	h_1	-1	1	1	-1	$\rightarrow$	h_1	-1	1	1	-1
h_2	-1	1	1	-1	1	$\nearrow$											
h_3	-1	1	1	1	-1	$\rightarrow$	h_3	-1	1	1	1						
h_4	1	-1	-1	1	-1	$\rightarrow$	h_4	1	-1	-1	1	$\rightarrow$	h_4	1	-1	-1	1
h_5	1	-1	-1	1	1	$\nearrow$											
h_6	1	1	-1	-1	1	$\rightarrow$	h_6	1	1	-1	-1						

Figure 1: Hypothesis tables for the proof of Sauer's Lemma

Combining the previous two paragraphs and noting that by construction the number of labelings $|\Pi_{\mathcal{H}}(S)|$

of $\mathcal{H}$ on S is simply the size of $\mathcal{H}_1$ and $\mathcal{H}_2$,

$$\begin{aligned}
|\Pi_{\mathcal{H}}(S)| &= |\mathcal{H}_1| + |\mathcal{H}_2| \le \Phi_d(n-1) + \Phi_{d-1}(n-1) \\
&= \sum_{i=0}^{d} \left(\begin{array}{c} n-1 \\ i \end{array} \right) + \sum_{i=0}^{d-1} \left(\begin{array}{c} n-1 \\ i \end{array} \right) = \sum_{i=0}^{d} \left(\begin{array}{c} n-1 \\ i \end{array} \right) + \sum_{i=0}^{d} \left(\begin{array}{c} n-1 \\ i-1 \end{array} \right) \\
&= \sum_{i=0}^{d} \left(\begin{array}{c} n \\ i \end{array} \right) = \Phi_d(n).
\end{aligned}$$

The equality in the second line follows because ${}_{n-1}C_{-1} = 0$, and the third line follows via the combinatorial identity ${}_{n-1}C_i + {}_{n-1}C_{i-1} = {}_nC_i$. As S was arbitrary, this completes the proof of the first part of the theorem.

Now suppose that $n \ge d \ge 1$. Then $\Phi_d(n)(d/n)^d = \sum_{i=0}^{d} {}_nC_i(d/m)^d$ and ${}_nC_i(d/n)^d \le {}_nC_i(d/n)^i$ since $d \le n$. Thus,

$$\begin{aligned}
\Phi_d(n) \left(\frac{d}{n}\right)^d &\le \sum_{i=0}^{d} \left(\begin{array}{c} n \\ i \end{array} \right) \left(\frac{d}{n}\right)^i \le \sum_{i=0}^{n} \left(\begin{array}{c} n \\ i \end{array} \right) \left(\frac{d}{n}\right)^i \\
&= \left(1 + \frac{d}{n}\right)^n \le \left(e^{d/n}\right)^n = e^d
\end{aligned}$$

The second line follows via an application of the binomial theorem, and its inequality is a consequence of $1 + x \le e^x$ for all x. Multiplying both sides by $(n/d)^d$ gives the desired result. □

By combining Sauer's lemma and a simple application of the Ledoux-Talagrand contraction (via Corollary 5.1), we can derive bounds on the expected loss of a classifier. Let $\mathcal{H}$ be a class of $\{-1, 1\}$ valued functions, and let examples be drawn from a distribution $\mathbb{P}(X, Y)$ where $Y \in \{-1, 1\}$ are labels for X. Then a classifier $h \in \mathcal{H}$ makes a mistake if and only if $Yh(X) = -1$. As such, the function $[1 - Yh(X)]_+ \ge \mathbf{1}\{Y \ne h(X)\}$, and $[\cdot]_+$ has Lipschitz constant 1. Thus, we have

$$\mathbb{P}(h(X) \ne Y) = \mathbb{E}\mathbf{1}\{Y \ne h(X)\} \le \mathbb{E}\left[1 - Yh(X)\right]_+ . \tag{10}$$

Further, for a Rademacher random variable σ_i, we have $\sigma_i Yh(X)$ symmetric around 0, so that $Yh(X)$ has the same distribution. Thus, $R_n(Y \cdot \mathcal{H}) = R_n(\mathcal{H})$. Further, $\phi(x) = [1 - x]_+ - 1$ is 1-Lipschitz and satisfies $\phi(0) = 0$, so Corollary 5.1 implies

$$R_n(\phi \circ (Y \cdot \mathcal{H})) \le 2R_n(Y \cdot \mathcal{H}) = 2R_n(\mathcal{H}).$$

Combining this with Eq. (10), $\mathbb{P}(h(X) \ne Y) - 1 = \mathbb{E}\mathbf{1}\{Y \ne h(X)\} - 1 \le \mathbb{E}\phi(Yh(X))$, which by Theorem 6 gives that with probability at least $1 - \delta$,

$$\mathbb{P}(h(X) \ne Y) - 1 \le \mathbb{E}_n\left[[1 - Yh(X)]_+ - 1\right] + 2R_n(\phi \circ (Y \cdot \mathcal{H})) + \sqrt{\frac{\log 1/\delta}{2n}}.$$

Clearly, we can add 1 to both sides of the above equation, and the empirical probability of a mistake $\hat{\mathbb{P}}(h(X) \ne Y) = \mathbb{E}_n\left[1 - Yh(X)\right]_+$. Combining Sauer's lemma and the above two equations, we have proved the following theorem.

Theorem 9. *Let $\mathcal{H}$ be a class of hypotheses on a space X with labels Y drawn according to a joint distribution $\mathbb{P}(X, Y)$. Then for any $h \in \mathcal{H}$ and given any sample $S = \{\langle x_1, y_1\rangle, \ldots, \langle x_n, y_n\rangle\}$ drawn i.i.d. according to $\mathbb{P}$, with probability at least $1 - \delta$ over the sample S drawn,*

$$\begin{aligned}
\mathbb{P}(h(X) \ne Y) &\le \hat{\mathbb{P}}(h(x_i) \ne y_i) + 4R_n(\mathcal{H}) + \sqrt{\frac{\log 1/\delta}{2n}} \\
&\le \hat{\mathbb{P}}(h(x_i) \ne y_i) + 4\sqrt{\frac{2d\log(en) - 2d\log d}{n}} + \sqrt{\frac{\log 1/\delta}{2n}}
\end{aligned}$$

In short, we have everyone's favorite result that the estimated probability is close to the true probability:

$$\mathbb{P}(h(X) \neq Y) = \hat{\mathbb{P}}(h(X) \neq Y) + O\left(\sqrt{\frac{d \log n + \log 1/\delta}{n}}\right).$$

References

[1] P. Billingsley, *Probability and Measure*, Third Edition, Wiley 1995.

[2] M. Ledoux and M. Talagrand, *Probability in Banach Spaces*, Springer Verlag 1991.