Guide to lectures for second two weeks.

Updated February 9, 2017

Tuesday 1/31/17 Formulation of the learning problem, continued (function estimation, agnostic case) See Chapter 3, Sections 3 and 4 of the course notes.

In Section 3.3 we see the agnostic (or model-free) model for a learning problem \((X, Y, U, P, F, \ell)\) and the corresponding definition for a learning algorithm \(A = (A_n)_{n \geq 1}\) to be PAC. If a PAC algorithm exists the problem is PAC learnable. The section concludes with pointing out that the agnostic case includes the realizable case as a special case. It also points out that the general model can be applied to a variation of a realizable case such that labels are noisy and flipped with some probability \(\eta\) (That part is summarized in Section 1 below.)

In Section 3.4 the empirical risk minimization (ERM) learning algorithm is defined for the general setup of Section 3.3. Theorem 1 is stated and proved, which shows that if the true joint distribution \(P\) is approximated uniformly well in probability (specifically, the Uniform Convergence of Empirical Means (UCEM) property holds) then the ERM procedure is PAC. So if the learning problem satisfies the UCEM condition, the problem is PAC learnable by ERM. Theorem 1 itself doesn’t have a rate of convergence associated with it, and it leaves open whether the UCEM property holds. As we’ve seen from the no free lunch theorem, the UCEM property does not hold for classification problems with infinite VC dimension.

The remainder of Section 3.4 discusses what we will be looking for in the next two chapters regarding the rate of convergence. Chapters 4 and 5 that follow are essentially due to Vapnik-Chervonenkis, and fill in the hole left at the end of Section 3, at least for the case of binary classification.

Friday 2/2/18, 1-2pm recitation session. Solutions to problems 3-5 from problem set 1 were reviewed by Amir.

Thursday 2/2/17 - Tuesday 2/7/17 See Chapter 4: Empirical Risk Minimization and Rademacher averages. Also, see Section 3 below.

In Section 4.1 we see a slick abstract formulation of the ERM algorithm and the measure of how well the empirical distribution matches the true distribution. Proposition 4.1 there essentially follows from the mismatched minimization lemma.

Section 4.2 shows that Rademacher averages can be used to measure the complexity of a learning problem. Specifically, as shown in Theorem 4.1, twice the mean Rademacher average upper bounds the mean approximation error of the empirical distribution. The proof is based on a symmetrization trick going back to Vapnik-Chervonenkis. Corollary 4.2 folds in the result we’ve seen earlier using McDiarmid’s concentration inequality to deduce an upper bound on the probability of a large approximation error based on the bound for mean approximation error. The factor 4 in Corollary 4.2 comes from two places. One factor of 2 from the lemma on minimizing an approximation of a function, and one factor of 2 from the symmetrization argument. The results in Section 4.2 motivate us to seek upper bounds on the Rademacher averages. Section 3 below summaries Sections 4.1 and 4.2 of the notes.

Section 4.3 investigates the notion of Rademacher averages in some detail, and then focuses on Lemma 4.4, the finite class lemma. It gives an upper bound on the Rademacher average of a finite set, and the proof largely follows from the union bound and Hoeffding inequality as used in the problem in problem set 1 on the maximum of a collection of subgaussian random variables. Then there is a some discussion at the end of Section 4.3 giving a pointer to the next chapter. It involves the VC dimension and Sauer’s lemma.

Tuesday 2/7/17 - Thursday 2/9/17 See Chapter 5 Vapnik-Chervonenkis classes

Section 5.1 gives the definition of VC dimension of a class of sets, or equivalently, of a class of binary valued functions. Also defined are shatter coefficients, where the \(n^{th}\) coefficient \(S_n(C)\) gives the maximum possible number of label vectors for a sample of \(n\) points.

Section 5.2 discusses some examples of computation of the VC dimension for some classes of concepts.
The most important example used in this section have to do with concept classes obtained by thresholding function from a linear or affine space, sometimes called a Dudley class of concepts.

Section 5.3 discusses a lemma called Sauer’s lemma shows that if \( V(C) = d \), then \( S_n(C) \leq \binom{n}{d} \), where \( \binom{n}{d} \) is the number of subsets of a set of \( n \) objects with cardinality \( d \) or less. For \( n \leq d \), the growth is exponential: \( \binom{n}{d} = 2^n \). For \( d \) larger, the growth becomes polynomial: \( \binom{n}{d} \leq \left( \frac{en}{d} \right)^d \) for \( n \geq d \). There are several proofs of Sauer’s lemma. In class I shall discuss a proof based on the idea of shifting. See Section 7 (based on http://www.cse.buffalo.edu/~hungngo/classes/2010/711/lectures/sauer.pdf).

The importance of Sauer’s lemma is that the shatter coefficients grow slowly enough that the finite class of Sauer’s lemma. In class I shall discuss a proof based on the idea of shifting. See Section 7 (based on http://www.cse.buffalo.edu/~hungngo/classes/2010/711/lectures/sauer.pdf).

1 Learning to classify with noisy labels

Example 1. (This is variation on Section 3.3.2 of the course notes, not covered in lecture (yet).) An agnostic classification problem can be built up from a realizable classification problem, as explained in this section. Consider a realizable classification problem with binary labels \((X, \mathcal{P}, \mathcal{C})\), under the usual 0-1 loss. Here \( \mathcal{P} \) is a family of probability measures on (the Borel subsets of) \( X \). Let \( C^* \) denote the target (i.e. true) concept and let \( C \) be another concept. Then for a given \( \mathcal{P}_X \subset \mathcal{P} \) and target concept \( C^* \), the expected loss for using concept \( C \) to classify is \( \mathcal{L}_{\mathcal{P}_X}(C^*, C) = \mathcal{L}(C^* \Delta C) \). A learning algorithm \( \mathcal{A} = (A_n)_{n \geq 1} \) is sought to produce \( \hat{C}_n = A_n(Z^n) \) that makes \( \mathcal{L}(C^* \Delta \hat{C}_n) \) small with high probability.

Let \( 0 \leq \eta < 1/2 \) denote a crossover probability. Given the above realizable classification problem, we can define a corresponding agnostic one by modeling the labels as noisy labels. Specifically, the agnostic model is denoted by \((X, Y = \{0, 1\}, \mathcal{P}, \mathcal{C})\), again under 0-1 loss. An element \( \mathcal{P}_{X,C^*} \) of \( \mathcal{P} \) corresponds to a \((\mathcal{P}_X, C^*)\) pair, such that \( \mathcal{P}_{X,C^*} \) is the joint probability distribution of \((X, Y)\) where \( X \) has probability distribution \( \mathcal{P}_X \) and \( Y = 1_{\{X \in C^*\}} \oplus W \), where \( W \) is independent of \( X \) with the Bernoulli(\( \eta \)) probability distribution, and \( \oplus \) denotes XOR (i.e. modulo 2) addition. In words, the original label \( 1_{\{X \in C^*\}} \) is toggled with probability \( \eta \) to produce the noisy label \( Y \).

Suppose that the objective of a classifier \( C \) for a given sample \( X \) is to predict the corresponding noisy label \( Y \). Then the expected loss can be written as

\[
\mathcal{L}_{\mathcal{P}_X,C^*}(C) = \mathcal{P}_X(1_{\{X \in C^*\}} \oplus W \neq 1_{\{X \in C\}})
\]

\[
= \mathcal{P}_X(W \neq 1_{\{X \in C^* \Delta C\}})
\]

\[
= (1 - \eta)\mathcal{P}_X(C^* \Delta C) + \eta(1 - \mathcal{P}_X(C^* \Delta C))
\]

\[
= \eta + (1 - 2\eta)\mathcal{P}_X(C^* \Delta C)
\]

Thus, \( \mathcal{L}_{\mathcal{P}_X,C^*}(C) \) is a (linear) increasing function of \( \mathcal{P}_X(C^* \Delta C) \). Hence, to have \( \mathcal{L}_{\mathcal{P}_X,C_n}(\hat{C}_n) \) close to its minimum possible value, we want \( \hat{C}_n \) to make \( \mathcal{L}(C^* \Delta \hat{C}_n) \) as small as possible, just as in the realizable case.

2 The mismatched minimization lemma

Suppose we’d like to find a minimizer of a function \( G \) defined on some domain \( U \), but the function \( G \) is not known. The following lemma is used often enough that it is worth naming.

Lemma 1. (Mismatched minimization lemma) Suppose that \( \hat{G} \) is an \( \epsilon \) uniform approximation of \( G \) for some \( \epsilon > 0 \), meaning that \( |G(u) - \hat{G}(u)| \leq \epsilon \) for all \( u \in U \). Suppose that \( u^* \) is a minimizer of \( \hat{G} \), meaning that \( u^* \in U \) and \( \hat{G}(u^*) \leq \hat{G}(u) \) for all \( u \in U \). Then \( G(u^*) \leq \inf_{u \in U} G(u) + 2\epsilon. \)
Proof. For any \( u \in \mathcal{U} \), \( G(u) \geq \hat{G}(u) - \epsilon \geq \hat{G}(u^*) - \epsilon \geq G(u^*) - 2\epsilon \). Therefore, \( \inf_{u \in \mathcal{U}} G(u) \geq G(u^*) - 2\epsilon \), which is equivalent to what we wanted to prove. \( \square \)

3 Expected risk minimization in abstract framework

(Summary of Sections 4.1 & 4.2 of course notes.) Consider the setup \((Z, \mathcal{P}, \mathcal{F})\) such that \( Z \) is a set, \( \mathcal{P} \) is a set of probability distributions on \( Z \) and \( \mathcal{F} \) is a family of functions on \( Z \). The learning goal is, given an unknown \( P \) selected from \( \mathcal{P} \) and given a vector of independent samples \( Z^n = (Z_1, \ldots, Z_n) \) drawn from distribution \( P^n \), produce \( \hat{f}_n = A_n(Z^n) \) to make \( L_P(f_n) \) nearly as small as possible with high probability, where \( L_P(f) = P(f) = E_P[f(Z)] = \int f(z) P(dz) \) (different notations for the same thing).

Let the minimum possible expected risk for a \( P \in \mathcal{P} \) be denoted by \( L^*_P(\mathcal{F}) = \inf_{f \in \mathcal{F}} P(f) \). The empirical risk minimization (ERM) algorithm in this setup is given by
\[
\hat{f}_n = \arg \min_{f \in \mathcal{F}} P_{Z^n}(f) = \arg \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n f(Z_i).
\]

The mismatched minimization lemma implies that \( P(\hat{f}_n) \leq L^*_P(\mathcal{F}) + 2\Delta_n(Z^n) \) where
\[
\Delta_n(Z^n) = \sup_{f \in \mathcal{F}} |P_{Z^n}(f) - P(f)| = \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(Z_i) - P(f) \right|.
\] (1)

That is, since \( P_{Z^n}(f) - P(f) \) is bounded uniformly in \( f \) by \( \Delta_n(Z^n) \), the value of \( P \) on the minimizer of \( P_{Z^n} \) is larger than the minimum of \( P(f) \) by at most twice \( \Delta_n(Z^n) \). So we’d like to have conditions under which \( \Delta_n(Z^n) \) is small with high probability.

The following definition will play an important role.

**Definition 1.** Let \( \mathcal{A} \subset \mathbb{R}^n \) with \( \mathcal{A} \) bounded. The Rademacher average of \( \mathcal{A} \), denoted by \( R_n(\mathcal{A}) \), is defined by
\[
R_n(\mathcal{A}) = E \left[ \sup_{a \in \mathcal{A}} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i a_i \right| \right],
\]
where \( \sigma_1, \ldots, \sigma_n \) are independent Rademacher (i.e., \( \pm 1 \) with equal probability) random variables.

Note that if we let \( \sigma \) denote the vector with \( i \)th coordinate \( \sigma_i \), then \( \frac{1}{\sqrt{n}} \sigma \) is a random unit length vector, and \( R_n(\mathcal{A}) = \frac{1}{\sqrt{n}} E \left[ \sup_{a \in \mathcal{A} \cup -\mathcal{A}} (a, \frac{1}{\sqrt{n}} \sigma) \right] \). Note that \( \sup_{a \in \mathcal{A} \cup -\mathcal{A}} (a, \frac{1}{\sqrt{n}} \sigma) \) can be interpreted as half the width of the smallest slab normal to \( \sigma \) that contains \( \mathcal{A} \cup -\mathcal{A} \). (FIGURE HERE WOULD BE NICE.) So \( R_n(\mathcal{A}) \) is proportional to the average half-width of \( \mathcal{A} \cup -\mathcal{A} \) in a direction chosen uniformly at random from the corners of the hypercube \([-1, 1]^n\).

To motivate the use of Rademacher averages, suppose that \( P \) and \( \mathcal{F} \) are such that if \( Z^n = (Z_1, \ldots, Z_n) \) is distributed according to \( P^n \) then
\[
\frac{1}{n} \sum_{i=1}^n f(Z_i) \approx P(f) \quad \text{for all } f \in \mathcal{F}, \text{ with high probability.} \quad (2)
\]

Condition (2) has to do with \( P, \mathcal{F} \), and \( n \) only; it doesn’t matter which random vector \( Z^n \) is used, as long as it has distribution \( P^n \). So if \( Z_{n+1}, \ldots, Z_{2n} \) are \( n \) more random variables that are independent, each with distribution \( P \), then
\[
\frac{1}{n} \sum_{i=n+1}^{2n} f(Z_i) \approx P(f) \quad \text{for all } f \in \mathcal{F}, \text{ with high probability.} \quad (3)
\]
If (2), and hence (3), are true, then the left-hand side of (2) minus the left-hand side of (3) is approximately zero, for all \( f \in \mathcal{F} \), with high probability. That can be written as:

\[
\frac{1}{n} \sum_{i=1}^{2n} \sigma_i f(Z_i) \approx 0 \quad \text{for all } f \in \mathcal{F}, \text{ with high probability,}
\]

where \( \sigma_1 = \cdots = \sigma_n = 1 \) and \( \sigma_{n+1} = \cdots = \sigma_{2n} = -1 \). Furthermore, if \( Z_1, \ldots, Z_{2n} \) are mutually independent with distribution \( P \), and if \( \pi: [2n] \to [2n] \) is a random permutation, uniformly distributed over all \( 2n! \) possibilities, and independent of \( Z^{2n} \), then

\[
\sum_{i=1}^{2n} \sigma_i f(Z_i) = \sum_{i=1}^{2n} \sigma_i f(Z_{\pi(i)}) = \sum_{i=1}^{2n} \sigma_{\pi^{-1}(i)} f(Z_i) = \sum_{i=1}^{2n} \tilde{\sigma}_i f(Z_i),
\]

where \( \tilde{\sigma}_i = \sigma_{\pi^{-1}(i)} \), so that \( \tilde{\sigma} \) is uniformly distributed over all \( \pm 1 \) vectors of length \( 2n \) with zero sum. The distribution of \( \tilde{\sigma} \) is close to the distribution of a vector of \( n \) iid Rademacher variables, if \( n \) is at least moderately large. To summarize, (2) implies that

\[
\frac{1}{2n} \sum_{i=1}^{2n} \tilde{\sigma}_i f(Z_i) \approx 0 \quad \text{for all } f \in \mathcal{F}, \text{ with high probability.}
\]

It is reasonable to think the converse is true as well. If (6) is true, it means that the lefthand sides of (2) and (3), which are independent of each other, are close to each other with high probability. It seems that forces (2) and (3) to be true. That intuition is behind to following theorem.

**Theorem 2.** For any \( P \in \mathcal{P} \), \( \mathbb{E}\left[\Delta_n(Z^n)\right] \leq 2E_{P^n}[R_n(\mathcal{F}(Z^n))] \), where \( \mathcal{F}(Z^n) = \{(f(Z_1), \ldots, f(Z_n)) : f \in \mathcal{F}\} \).

**Proof.** The idea of the proof is to use the Vapnik-Chernovenkis symmetrization argument. Fix \( P \in \mathcal{P} \) throughout the proof. By definition,

\[
\mathbb{E}\left[\Delta_n(Z^n)\right] = \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(Z_i) - f(P) \right| \right].
\]

Note that the mapping

\[
y \mapsto \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(Z_i) - y(f) \right| \right]
\]

is a convex mapping of \( y = (y(f) : f \in \mathcal{F}) \) to \( \mathbb{R} \), because the absolute value function is convex, the supremum of a set of convex functions is convex, and the expectation of a random convex function is convex. Let \( Z^n \) be an independent copy of \( Z^n \). For example, we could take \( Z_i = Z_{n+i} \) where \( Z_1, \ldots, Z_{2n} \) are distributed as above. Note that \( \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} f(Z_i)\right] = P(f) \). So by Jensen’s inequality,

\[
\mathbb{E}\left[\Delta_n(Z^n)\right] \leq \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(Z_i) - f(f(Z)) \right| \right].
\]

For each \( i \), \( (f(Z_i), f(Z_i)) \) \( \overset{d}{=} \) \( (f(Z_i), f(Z_i)) \), so that \( f(Z_i) - f(Z_i) \overset{d}{=} f(Z_i) - f(Z_i) \). That is, the distribution of \( f(Z_i) - f(Z_i) \) is symmetric. Thus, if \( \sigma_1, \ldots, \sigma_n \) are independent Rademacher random variables, \( (f(Z_i) - f(Z_i))_{1 \leq i \leq n} \overset{d}{=} (\sigma_i(f(Z_i) - f(Z_i)))_{1 \leq i \leq n} \). Thus,

\[
\mathbb{E}\left[\Delta_n(Z^n)\right] \leq \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i(f(Z_i) - f(Z_i)) \right| \right] \leq \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(Z_i) \right| \right] + \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(Z_i) \right| \right] = 2\mathbb{E}\left[R_n(\mathcal{F}(Z^n))\right].
\]
Corollary 1. For any probability distribution $P$ on $Z$ and $\delta \in (0, 1)$, the ERM algorithm satisfies

$$P(\hat{f}_n) \leq L^*_p(F) + 4E_p[\hat{R}_n(F(Z^n)) + \sqrt{2\log \left( \frac{2}{\delta} \right)}]$$

with probability at least $1 - \delta$.

Proof. As noted earlier, the mismatched minimization lemma implies $P(\hat{f}_n) \leq L^*(F) + 2\Delta_n(Z^n)$ with probability one, and by Theorem 2, $\mathbb{E}[2\Delta_n(Z^n)] \leq 4E_p[R_n(F(Z^n))]$. It thus suffices to prove the probability that $2\Delta_n(Z^n)$ exceeds its mean by more than $\epsilon = \sqrt{\frac{2\log \frac{2}{\delta}}{n}}$ is less than or equal to $\delta$. Examining the definition (1) of $\Delta(Z^n)$ shows that, as a function of $Z^n$, it has the bounded difference property for constants $c_i = \frac{1}{n}$ for all $i$. Thus, by McDiarmid’s inequality,

$$P\{2\Delta_n \geq 2\mathbb{E}[\Delta_n] + \epsilon\} = P\{\Delta_n \geq \mathbb{E}[\Delta_n] + \epsilon/2\} \leq \exp \left( -\frac{(\epsilon/2)^2}{n/2} \right) = \exp (-n\epsilon^2/2) = \delta,$$

as desired. 

4 Some properties of Rademacher averages

Here is the definition of the Rademacher average of a bounded set $A \subset \mathbb{R}^n$ used in the notes and lecture, followed by an alternative version:

$$R_n(A) \triangleq \mathbb{E} \left[ \sup_{a \in A} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i a_i \right| \right], \quad R_n^o(A) \triangleq \mathbb{E} \left[ \sup_{a \in A} \frac{1}{n} \sum_{i=1}^{n} \sigma_i a_i \right],$$

where $\sigma_1, \ldots, \sigma_n$ are independent Rademacher random variables. The alternative definition omits the absolute value.

For $A \subset \mathbb{R}^n$, the convex hull of $A$, denoted by $\text{conv } A$, and the absolute convex hull, denoted by $\text{absconv } A$, are defined by

$$\text{conv } A = \left\{ \sum_{m=1}^{N} c_m a^{(m)} : N \geq 1, c_m \geq 0 \text{ for } m \in [N], c_1 + \ldots + c_n = 1 \right\}$$

$$\text{absconv } A = \left\{ \sum_{m=1}^{N} c_m a^{(m)} : N \geq 1, |c_1| + \ldots + |c_N| = 1 \right\} = \text{conv } (A \cup (-A))$$

Some basic properties of Rademacher averages are as follows:

1. $R_n^o(A) \leq R_n(A) = R_n^o(A \cup -A)$
2. $R_n^o(A + v) = R_n^o(A)$ for any $v \in \mathbb{R}^n$,
3. $R_n(A \cup B) \leq R_n(A) + R_n(B)$
4. $R_n(A + B) = R_n(A) + R_n(B)$
5. $R_n(cA) = |c|R_n(A)$
6. $R_n(A) = R_n(\text{conv } A)$, where conv denotes convex hull.
7. $R_n(A) = R_n(\text{absconv } A)$, where absconv denotes absolute convex hull.
8. \( R_n(a) = \mathbb{E} \left[ \sup_{a, a' \in A} \frac{1}{n} \sum_{i=1}^{n} \sigma_i (a_i - a'_i) \right] \). (The right-hand side is proportional to the average width of \( A \) in the direction \( (\sigma_1, \ldots, \sigma_n) \).)

We give a proof here of the fact \( R_n(A) = R_n(\text{conv} \, A) = R_n(\text{absconv} \, A) \). Suppose \( a \in \text{conv} \, A \). Then \( a = \sum_{m=1}^{N} c_m a^{(m)} \) for some \( N \geq 1, a^{(1)}, \ldots, a^{(N)} \in A \), and probability vector \( (c_1, \ldots, c_N) \). By the convexity of the absolute value function and Jenen’s inequality,

\[
\left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i a_i \right| \leq \sum_{m=1}^{N} c_m \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i a_i^{(m)} \right| \\
\leq \sum_{m=1}^{N} c_m \sup_{b \in A} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i b_i \right| = \sup_{b \in A} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i b_i \right|
\]

Therefore,

\[
\sup_{a \in \text{conv} \, A} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i a_i \right| = \sup_{b \in A} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i b_i \right|.
\]

Taking expectations over each side of (8) proves that \( R_n(\text{conv} \, A) = R_n(A) \). Also since \( \text{absconv} \, A = \text{conv} \, A \cup (-A) \), it follows that

\[
R_n(\text{absconv} \, A) = R_n(\text{conv} \, A \cup (-A)) = R_n(A \cup (-A)) = R_n(A).
\]

**Lemma 2.** (Finite class lemma) If \( A = \{a^{(1)}, \ldots, a^{(N)}\} \subset \mathbb{R}^n \) is a finite set with \( \|a^{(j)}\| \leq L \) for \( j \in [N] \) and \( N \geq 2 \), then \( R_n(A) \leq \frac{2L^2 \log N}{n} \).

**Proof.** Note that

\[
R_n(A) = \mathbb{E} \left[ \max_{j \in [N]} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i a_i^{(j)} \right| \right] = \mathbb{E} \left[ \max \{ Y_1, -Y_1, Y_2, -Y_2, \ldots, Y_N, -Y_N \} \right],
\]

where \( Y_j = \frac{1}{n} \sum_{i=1}^{n} \sigma_i a_i^{(j)} \). A term of the form \( \sigma_i a_i^{(j)} \) has mean zero and takes values in the interval \([-a_i^{(j)}, a_i^{(j)}]\). By the Hoeffding inequality, \( \sigma_i a_i^{(j)} \) is subGaussian with proxy variance \( (a_i^{(j)})^2 \). Thus, \( Y_j \) is subGaussian with proxy variance \( \frac{1}{n} \sum_{i=1}^{n} (a_i^{(j)})^2 = \frac{\|a^{(j)}\|^2}{n} \leq \frac{L^2}{n} \). Thus, \( Y_1, -Y_1, Y_2, -Y_2, \ldots, Y_N, -Y_N \) are each subGaussian with scale parameter \( \frac{L}{\sqrt{n}} \). Thus, by a problem in problem set 1, \( R_n(A) \leq \frac{L}{\sqrt{n}} \sqrt{2 \log(2N)} \). Since \( \log(2N) = \log 2 + \log N \leq 2 \log N \) for \( N \geq 2 \), the lemma follows. \( \square \)

**Example 3.** Suppose \( A \subset \{0, 1\}^n \) Then \( \|a\| \leq \sqrt{n} \) for any \( a \in A \). Thus, Lemma 2 with \( L = \sqrt{n} \) implies \( R_n(A) \leq 2 \sqrt{\log |A| / n} \).

(a) If \( |A| = 2^n \), that is, \( A = \{0, 1\}^n \), then in the definition of \( R_n(A) \) we can take \( a_i = 1(\sigma_i = 1) \) for each \( i \), giving \( R_n(A) = 1/2 \). The upper bound gives \( R_n(A) \leq 2 \sqrt{\log 2} \approx 1.665 \). (If instead \( A = \{-1, 1\}^n \) then \( R_n(A) = 1 \) and the upper bound is the same.)

(b) If \( |A| \leq 2^n R \) for all \( n \), where \( 0 < R < 1 \), the upper bound gives \( R_n(A) \leq 2 \sqrt{2R \log 2R} \) for all \( n \), which does not converge to zero as \( n \to \infty \).

(c) If \( |A| \leq n^d \) for some positive constant \( d \), the upper bound gives \( R_n(A) \leq 2 \sqrt{d \log n / n} \to 0 \) as \( n \to \infty \).

5 The contraction principle for Rademacher averages

Next we discuss a contraction principle for Rademacher averages that will be used in later chapters. Suppose \( n \geq 1 \) and for \( 1 \leq i \leq n \), \( F_i : \mathbb{R} \to \mathbb{R} \). Let \( F \circ v = (F_1(v_1), \ldots, F_n(v_1)) \). Given a subset \( A \) of \( \mathbb{R}^n \), let \( F \circ A = \{ F \circ v : v \in A \} \).
Recall that $R_n^o(A)$ is the variation of $R_n(A)$ obtained by omitting the absolute value symbols in the definition of $R_n(A)$.

**Proposition 1.** (Contraction principles for Rademacher averages) If $A$ is a bounded subset of $\mathbb{R}^n$ and for $i \in [n]$, $F_i : \mathbb{R} \to \mathbb{R}$ is an $M$-Lipschitz continuous function, then $R_n^o(F \circ A) \leq MR_n^o(A)$. Furthermore, if $F_i(0) = 0$ for all $i$ (i.e. $F \circ 0 = 0$) then $R_n(F \circ A) \leq 2M R_n(A)$.

**Proof.** We first prove the contraction property for $R_n^o$, following a method of Kakade and Tiwari. By scaling, without loss of generality we can assume that $M = 1$. Since the functions $F_1, \ldots, F_n$ can be introduced one at a time, it suffices to consider the case that all the functions $F_i$ are equal to the identity function, except for one value of $i$. Without loss of generality, it suffices to consider the case that only $F_1$ is not the identity function. That is, it suffices to show that $R_n^o(A) = R_n^o(A_1)$, where $A_1 = \{(\phi(a_1), a_2, \ldots, a_n) : a \in A\}$ and $\phi : \mathbb{R} \to \mathbb{R}$ and $\phi$ is Lipschitz continuous with constant one. Averaging over the values of $\sigma_1$, we get

$$R_n^o(A) = \frac{1}{2n} \mathbb{E} \left[ \sup_{a \in A} \left( a + \sum_{i=2}^{n} \sigma_i a_i \right) + \sup_{a' \in A} \left( -a'_1 + \sum_{i=2}^{n} \sigma_i a'_i \right) \right]$$

$$= \frac{1}{2n} \mathbb{E} \left[ \sup_{a,a' \in A} \left( a - a'_1 + \sum_{i=2}^{n} \sigma_i a_i + \sum_{i=2}^{n} \sigma_i a'_i \right) \right]$$

$$= \frac{1}{2n} \mathbb{E} \left[ \sup_{a,a' \in A} \left( |a - a'_1| + \sum_{i=2}^{n} \sigma_i a_i + \sum_{i=2}^{n} \sigma_i a'_i \right) \right],$$

where the last equality comes from the fact that $a$ and $a'$ can be swapped. Applying the same equations with $A$ replaced by $A_1$ yields:

$$R_n^o(A_1) = \frac{1}{2n} \mathbb{E} \left[ \sup_{a,a' \in A_1} \left( |\phi(a_1) - \phi(a'_1)| + \sum_{i=2}^{n} \sigma_i a_i + \sum_{i=2}^{n} \sigma_i a'_i \right) \right].$$

Comparison of (9) and (10) and using the assumption $|\phi(a_1) - \phi(a'_1)| \leq |a_1 - a'_1|$ yields that $R_n^o(A_1) \leq R_n^o(A)$. This completes the proof of $R_n^o(F \circ A) \leq MR_n^o(A)$.

We shall show how the contraction principle for $R_n$ follows from the contraction principle for $R_n^o$. Let $0$ denote the zero vector in $\mathbb{R}^n$. We shall also use the fact that $R_n^o(A \cup B) \leq R_n^o(A) + R_n^o(B)$ if $0 \in A$ and $0 \in B$. We find

$$R_n(F \circ A) = R_n^o((F \circ A) \cup (-F \circ A) \cup \{0\})$$

$$\leq R_n^o((F \circ A) \cup \{0\}) + R_n^o((-F \circ A) \cup \{0\})$$

$$= R_n^o(F \circ (A \cup \{0\})) + R_n^o(-F \circ (A \cup \{0\}))$$

$$= 2R_n^o(F \circ (A \cup \{0\}))$$

$$\leq 2MR_n^o(A \cup \{0\}) \leq 2MR_n(A).$$

\[
\]

6 VC Dimension

Let $C$ be a family of subsets of some feature space $Z$.

**Definition 2.** A set of $n$ points $z_1, \ldots, z_n \in Z$ is said to be shattered by $C$ if for any binary vector $b = (b_1, \ldots, b_n)$, there exists $C_b \in C$ such that $b = (1_{z_1 \in C_b}, \ldots, 1_{z_n \in C_b})$. The Vapnik-Chervonenkis (VC) dimension of $C$, $V(C)$, is the largest $n$ such that there exists a set of $n$ points in $X$ that is shattered by $C$. If there exist arbitrarily large sets of points that are shattered by $C$, then $V(C)$ is infinite. If $V(C)$ is finite, $C$ is called a VC class.
The \( n \)th \textit{shatter coefficient}, \( S_n(C) \) is the maximum over all \( S = \{z_1, \ldots, z_n\} \subset \mathbb{Z} \) of the cardinality of \( \{S \cap C : C \in \mathcal{C}\} \). Note that \( V(C) = \sup \{n : S_n(C) = 2^n\} \).

See the course notes for examples of determining the VC dimensions of several classes of concepts: semi-infinite intervals, closed intervals, closed half spaces, and closed axis-parallel rectangles. A flexible and fairly general way to specify a set \( C \) of concepts is to use thresholded linear combinations of functions. The examples mentioned are all have such form.

Let \( \text{pos} \) map a vector \( v \) to the binary vector of the same dimension, such that \( \text{pos}(v)_i = 1_{\{v_i \geq 0\}} \). Similarly, given a function \( g : \mathbb{Z} \rightarrow \mathbb{R} \) let \( \text{pos}(g) : \mathbb{Z} \rightarrow \{0, 1\} \) be defined by \( \text{pos}(g)(z) = \text{pos}(g(z)) \), and if \( \mathcal{G} \) is a family of such functions let \( \text{pos}(\mathcal{G}) = \{\text{pos}(g) : g \in \mathcal{G}\} \). The \textit{(linear) span} of a set of functions \( \psi_1, \ldots, \psi_k \) on \( \mathbb{Z} \) is the set of all linear combinations of the form \( \sum_{i=1}^k c_i \psi(z) \). If \( \mathcal{G} \) is a class of functions on \( \mathbb{Z} \) and \( h : \mathbb{Z} \rightarrow \mathbb{R} \), then \( \mathcal{G} + h = \{g + h : g \in \mathcal{G}\} \).

\textbf{Definition 3.} A Dudley class for a set \( \mathbb{Z} \) is a set of \( \{0, 1\} \) valued functions (equivalent to a set of subsets of \( \mathbb{Z} \)) of the form \( \text{pos}(\mathcal{G} + h) \), where \( \mathcal{G} \) is the span of some real-valued functions \( \psi_1, \ldots, \psi_m \), and \( h : \mathbb{Z} \rightarrow \mathbb{R} \).

(Exercise: How are the examples represented as Dudley classes?)

Real-valued functions \( \psi_1, \ldots, \psi_k \) on \( \mathbb{Z} \) are \textit{linearly independent} if the only vector \( (c_i : i \in [k]) \) such that \( \sum_{i=1}^k c_i \psi_i \equiv 0 \) on \( \mathbb{Z} \) is the zero vector.

\textbf{Proposition 2.} (VC dimension of Dudley classes) Suppose \( \mathcal{G} \) is the space of functions spanned by linearly independent functions \( \psi_1, \ldots, \psi_m \) on some set \( \mathbb{Z} \). Let \( h : \mathbb{Z} \rightarrow \mathbb{R} \), with \( h \) not necessarily in \( \mathcal{G} \). (\( \text{pos}(\mathcal{G} + h) \) is a Dudley class.) Then \( V(\text{pos}(\mathcal{G} + h)) = m \).

\textit{Proof.} It will be shown here that \( V(\text{pos}(\mathcal{G})) \leq m \) (i.e. upper bound in case \( h = 0 \).) See problem set 2 for the remainder of the proof, that is, proof of the upper bound with nonzero \( h \), and proof of the matching lower bound.

Let \( z_1, \ldots, z_{m+1} \in \mathbb{Z} \). It suffices to show \( \text{pos}(\mathcal{G}) \) does not shatter this set of points. Let
\[ L(\mathcal{G}) = \{(g(z_1), \ldots, g(z_{m+1}) : g \in \mathcal{G}\} \subset \mathbb{R}^{m+1}. \]

Equivalently, \( L(\mathcal{G}) \) is the linear span (i.e. set of all linear combinations) of the \( m \) vectors \( (\psi_1(z_1), \ldots, \psi_1(z_{m+1})) \). Therefore, \( L(\mathcal{G}) \) is a linear subspace of \( \mathbb{R}^{m+1} \) with dimension \( m \). There thus exists a nonzero vector \( \nu = (\nu_1, \ldots, \nu_{m+1}) \in \mathbb{R}^{m+1} \) that is orthogonal to (every vector of) \( L(\mathcal{G}) \). That is, for any \( g \in \mathcal{G} \),
\[ \nu_1 g(z_1) + \cdots + \nu_{m+1} g(z_{m+1}) = 0. \quad (11) \]

Since \( \nu \) could be replaced by \( -\nu \) if necessary, it can be assumed without loss of generality that \( \nu_i < 0 \) for some \( i \). Let \( b \) denote the binary vector \( \nu = \text{pos}(\nu) \).

We now argue by contradiction, and suppose that \( \text{pos}(\mathcal{G}) \) does shatter \( \{z_1, \ldots, z_{m+1}\} \), so that there exists \( g \in \mathcal{G} \) such that \( \text{pos}(g(z_i)) = b_i = \text{pos}(\nu_i) \) for \( 1 \leq i \leq m + 1 \), and in particular \( \text{pos}(g(z_i)) = 0 = \text{pos}(\nu_i) \).

Thus, \( \nu_i g(z_i) \geq 0 \) for \( 1 \leq i \leq m + 1 \), and \( \nu_i g(z_i) > 0 \), contradicting \((11)\), thereby completing the proof.

\textbf{Example 4.} Let \( \mathbb{Z} = \mathbb{R}^N \). A monomial with domain \( \mathbb{Z} \) is a function of the form \( z_1^{d_1} z_2^{d_2} \cdots z_N^{d_N} \) such that \( d_1, \ldots, d_N \) are positive integers, and the degree of the monomial is \( d_1 + \cdots + d_N \). Given a positive integer \( d \), let \( \mathcal{G} \) denote the linear span of all monomials of degree less than or equal to \( d \). That is, \( \mathcal{G} \) is the set of all multivariate polynomials in \( z_1, \ldots, z_n \) with degree less than or equal to \( d \). The Dudley class \( \text{pos}(\mathcal{G}) \) consists of the linear classifiers if \( d = 1 \) and quadratic classifiers if \( d = 2 \). For example, if \( d = 4 \), \( \mathcal{G} \) includes the function \( g(z) = (||z|| - 1)^2 - \epsilon^2 \), and the corresponding concept, \( \text{pos}((||z|| - 1)^2 - \epsilon^2) \), consists of a shell of thickness \( \epsilon \) around the unit sphere in \( \mathbb{R}^N \).
7 The Sauer-Shaleh lemma

The following is a restatement of the definition of VC dimension for the case the base set consists of \([n] = \{1, \ldots, n\}\), and subsets of \([n]\) are represented as binary vectors.

Let \(U \subset \{0,1\}^n\). Given \(A \subset [n]\), if \(b \in \{0,1\}^n\) let \(\pi_A(b)\) denote the restriction of \(b\) to \(A\), and let \(\pi_A(S) = \{\pi_A(b) : b \in S\}\). A nonempty set \(A \subset [n]\) is said to be shattered by \(U\) if \(\pi_A(V)\) contains all \(2^{|A|}\) possible sequences. Let \(V(U)\) denote the VC dimension of \(U\). Let \(\binom{n}{d}\) denote the number of subsets of \([n]\) with cardinality less than or equal to \(d\).

**Lemma 3.** (Sauer-Shaleh lemma) Suppose \(U \in \{0,1\}^n\) with \(V(U) = d\). Then \(|U| \leq \binom{n}{d}\).

Also, \(\binom{n}{\leq d} \leq (n+1)^d\) and, for \(n \geq d\), \(\binom{n}{\leq d} \leq (\frac{ne}{d})^d\).

**Proof.** We give the proof based on the technique of shifting from extremal set theory. See See [http://www.cse.buffalo.edu/~hungngo/classes/2010/711/lectures/sauer.pdf](http://www.cse.buffalo.edu/~hungngo/classes/2010/711/lectures/sauer.pdf) for references. It is shown below that there exists \(V \subset \{0,1\}^n\) satisfying the following three properties:

1. \(|V| = |U|\)
2. For any \(A \subset [n]\), if \(A\) is shattered by \(V\) then \(A\) is shattered by \(U\)
3. \(V\) is downward closed, meaning if \(b, b' \in \{0,1\}^n\), \(b' \leq b\) (bitwise), and \(b \in V\) then \(b' \in V\).

This will complete the proof, because the third property implies that \(V\) shatters the support of any vector in \(V\). Therefore by property 2, \(U\) shatters the support of any vector in \(V\). Thus, any vector in \(V\) can have at most \(d\) nonzero bits. So \(|U| = |V| \leq \binom{n}{\leq d}\). It remains to describe the algorithm and show properties 1-3. The algorithm is described in pseudocode below, describing how \(U\) is transformed into \(V\). For \(i \in [n]\), \(\tau_i\) denotes the toggle function. It operates on binary vectors such that \(\tau_i(b)\) is obtained from \(b\) by flipping the \(i^{th}\) bit of \(b\) from 0 to 1 or vice versa. Here is the algorithm:

1. for \(i\) in \([n]\):
2. for \(b \in U\):
3. if \(b_i = 1\) and \(\tau_i(b) \notin U\):
4. replace \(b\) by \(\tau_i(b)\)
5. repeat steps 2-4 until no further changes occur
6. repeat steps 1-5 until no further changes occur
7. return \(V = U\).

The algorithm terminates because the total number of 1’s strictly decreases each time steps 1-5 are executed, except for the last time. Property 1 is true because \(|U|\) is never changed by the algorithm. Property 3 is true because otherwise more changes would have been possible in steps 1-5. It remains to check property 2. Consider the block of steps 2-5, executed for some \(i \in [n]\). Let \(\mathcal{U}\) denote the state just before execution of steps 2-5 and let \(\mathcal{U}'\) denote the state just after execution of those steps. Suppose \(A \subset [n]\) such that \(\mathcal{U}'\) shatters \(A\). It suffices to prove \(\mathcal{U}\) also shatters \(A\).

The only changes to \(\mathcal{U}\) made during the block of steps 2-5 is that the \(i^{th}\) bit of some vectors in \(\mathcal{U}\) might be changed from 1 to 0. So if \(i \notin A\) then \(\pi_A(\mathcal{U}) = \pi_A(\mathcal{U}')\), so that \(\mathcal{U}\) also shatters \(A\).

So suppose \(i \in A\). Let \(b\) denote an arbitrary binary vector indexed by \(A\). Since \(\mathcal{U}'\) shatters \(A\) there is a vector \(b'\) such that \(b = \pi_A(b')\). We need to show that \(b = \pi_A(b'')\) for some \(b'' \in \mathcal{U}\). If \(b_i = 1\) then, since the algorithm only turns 1’s into 0’s, \(b' \in \mathcal{U}\), so it suffices to take \(b'' = b'\). If \(b_i = 0\), then since \(\mathcal{U}'\) shatters \(A\) there must be another vector \(b''' \in \mathcal{U}'\) such that \(\pi_A(b''') = \tau_i(b)\). Since \(b''' = 1\) and \(b''' \in \mathcal{U}'\), it must be that \(b'' = b''\). Since \(b''\) was not modified by the block of steps 2-5, it must be that \(\tau_i(b'') \in \mathcal{U}\). Therefore, it suffices to take \(b'' = \tau_i(b'')\). Thus, \(\mathcal{U}\) shatters \(A\) as claimed.
Let $\text{Theorem 5.}$

The Sauer-Shaleh lemma implies that if the coefficient of $F$ is $\text{Definition 4.}$ Let $F$ be a family of functions $f : Z \to \{0, 1\}$. We say a finite set $S = \{z_1, \ldots, z_n\} \subset Z$ is shattered by $F$ if it is shattered by the class $C_S \triangleq \{C_f : f \in F\}$, and $C_f = \{z : f(z) = 1\}$. The $n^{th}$ shatter coefficient of $F$ is $S_n(F) = S_n(C_F)$ and the VC dimension of $F$ is defined as $V(F) = V(C_F)$.

The Sauer-Shaleh lemma implies that if $V(F) = d < \infty$, then for all $n \geq V(F)$, $S_n(F) \leq (n + 1)^d$ for all $n \geq d$.

**Theorem 5.** Let $F$ be a VC class of binary-valued functions $f : Z \to \{0, 1\}$ on some space $Z$. Let $Z^n$ be an iid sample of size $n$ drawn from $Z$ according to an arbitrary probability distribution $P \in \mathcal{P}(Z)$. Then

$$
\mathbb{E}[R_n(F(Z^n))] \leq 2\sqrt{\frac{V(F) \log(n+1)}{n}}
$$

**Proof.** By the finite class lemma (Lemma 2) with $L = \sqrt{n}$, $R_n(F(Z^n)) \leq 2\sqrt{\frac{\log |F(Z^n)|}{n}}$, and by the Sauer-Shaleh lemma, $|F(Z^n)| \leq (n + 1)^{V(C)}$ with probability one. \qed

**Theorem 6.** (Theorem 5 with log factor removed, based on Dudley’s chaining technique.) There exists an absolute constant $C$ such that under the conditions of the preceding theorem,

$$
\mathbb{E}[R_n(F(Z^n))] \leq C\sqrt{\frac{V(F)}{n}}
$$

Combining these theorems with Corollary 1 yields the new corollary:

**Corollary 2.** For any probability distribution $P$ on $Z$ and $\delta \in (0, 1)$, the ERM algorithm satisfies

$$
P(\hat{f}_n) \leq L^*_P(F) + 8\sqrt{\frac{V(F) \log(n+1)}{n}} + \sqrt{\frac{2\log \left(\frac{1}{\delta}\right)}{n}} \tag{12}
$$

with probability at least $1 - \delta$. There is a universal constant $C$ so that for any probability distribution $P$ on $Z$ and $\delta \in (0, 1)$, the ERM algorithm satisfies

$$
P(\hat{f}_n) \leq L^*_P(F) + C\sqrt{\frac{V(F)}{n}} + \sqrt{\frac{2\log \left(\frac{1}{\delta}\right)}{n}} \tag{13}
$$

We conclude that in the abstract framework for ERM, a problem $(Z, F, P)$ is PAC learnable if $F$ is a VC class of functions.

8 Fundamental Theorem of Concept Learning

Consider concept learning in the agnostic case. A learning problem can be specified by a triplet $(X, \mathcal{P}, \mathcal{C})$ such that
• $X$ is the feature space
• $P = \mathcal{P}_{X \times \{0,1\}}$ is a set of probability distributions on $Z \triangleq X \times \{0,1\}$
• $C$ is a set of subsets of $X$.

Recall that to simplify notation for analysis of ERM, we considered a class of functions $f$ on $Z$. For the concept learning problem, each concept $C$ corresponds to a function $\ell_C : Z \to \{0,1\}$ defined by $\ell_C(z) = 1\{y=1\{x \in C\}\}$. Let $F_C = \{\ell_C : C \in C\}$. The following lemma states that the VC dimension of the set of classifiers $C$ is the same as the VC dimension of the set of classifiers induced in the abstract setup. See problem set 3 for a proof.

**Lemma 4.** $V(C) = V(F_C)$.

Combining this lemma, Corollary 2, and the no-free lunch theorem for the converse result, yields the following theorem for concept learning.

**Theorem 7.** *(Fundamental theorem of concept learning)* Consider an agnostic concept learning problem $(X, P, C)$, and let $\delta > 0$. For any $P \in \mathcal{P}$, the ERM algorithm satisfies

$$P(C^* \triangle \hat{C}_n) \leq L^*_P(C) + 8 \sqrt{\frac{V(C) \log(n+1)}{n}} + \sqrt{\frac{2 \log \left( \frac{1}{\delta} \right)}{n}}$$  \hfill (14)

with probability at least $1 - \delta$. There is a universal constant $C$ so that for any probability distribution $P$ on $Z$ and $\delta \in (0,1)$, the ERM algorithm satisfies

$$P(C^* \triangle \hat{C}_n) \leq L^*_P(C) + C \sqrt{\frac{V(C)}{n}} + \sqrt{\frac{2 \log \left( \frac{1}{\delta} \right)}{n}}$$  \hfill (15)

Thus, if $C$ is a VC class, the concept problem is PAC learnable, and ERM is PAC. Conversely, if $V(C) = \infty$ and $\mathcal{P}$ is the set of all probability distributions on $Z = X \times \{0,1\}$, then the problem is not PAC learnable.