Machine Learning Theory, Self-Study

Selected Solutions for Mohri's Foundation to ML (2nd)
Others to be added on an ad hoc basis
Best to follow alongside lectures and notes from:
`https://home.work.caltech.edu/lectures.html`
`https://cs.nyu.edu/~mohri/mlbook/`

Guilherme Albertini

Contents

Preface

TODO blah blah

Chapter 1

PAC Learning

1.1 Priors

The appendix is quite helpful for bounds used throughout the text. As the authors state, "The book of Kearnsand Vazirani (1994) is an excellent reference dealing with most aspects of PAC-learning and several other foundational questions in machine learning. Our example of learning axis-aligned rectangles, also discussed in that reference, is originally due to Blumer et al. (1989)."

Mohri's course notes will be helpful: `https://cs.nyu.edu/~mohri/ml20/`. Note that the textbook pdf is freely available on his website, too. Familiarity with data science fundamentals, multivariable calculus, linear algebra, and algorithmic analysis is assumed for this specific text. Knowledge of convex optimization (or nonlinear programming) and real analysis would be very useful, it seems.

I include a "corrected" and (imo) clearer proof from other authors that do not assume continuity of distribution under the chapters folder: `http://compbio.fmph.uniba.sk/vyuka/ml/handouts/rectangles_correction.pdf`

For proofs throughout: For implication $p \implies q$, an antecedent (or hypothesis) p is a sufficient condition for a consequent (a conclusion) q when the truth of p alone implies the truth of q; however, p being false does not always imply that q is also false. A necessary condition is when the truth of q is guaranteed by the truth of p, or we can say that the truth of p is implied by the truth of q; in other words, p is not possible without q. Several necessary conditions may induce a condition whereas a sufficient condition is alone enough to produce the said condition. The sufficient term is the part that immediately follows "if" and the necessary term is the part that immediately follows the "then". Note that these are converses, but the converse may not always be true.

Further reading:
`https://philosophy.stackexchange.com/questions/22/what-is-the-difference-between-necessary-and-sufficient`
`https://www.kaptest.com/study/lsat/lsat-formal-logic-necessary-vs-sufficient/`
`https://pages.cs.wisc.edu/~shuchi/courses/787-F07/scribe-notes/lecture25.pdf`
`https://www.cs.princeton.edu/courses/archive/spr06/cos511/scribe_notes/0214.pdf`
`https://jeremykun.com/2014/01/02/probably-approximately-correct-a-formal-theory-of-learning/`
`https://www.cs.cornell.edu/courses/cs6781/2020sp/lectures/03-pac1.pdf`

1.2 Exercises

Exercise 2.1
Necessary and sufficient: a concept class $\mathcal{C}$ is efficiently PAC-learnable using hypothesis space $\mathcal{H}$ in the standard PAC model if and only if it is efficiently PAC-learnable using the hypothesis space $\mathcal{H} \cup \{h_0, h_1\}$ in the two-oracle PAC model.

Sufficiency: Show that if $\mathcal{H}$ is PAC-learnable in the standard, one-oracle model then so too is it in the variant.

Since $\mathcal{C}$ is efficiently PAC-learnable using $\mathcal{H}$, there exists an algorithm $\mathcal{A}$ and a polynomial p such that for any distribution $\mathcal{D}$ and any target concept $c \in \mathcal{C}$, if $\mathcal{A}$ is given a sample of size $m \geq p(1/\epsilon, 1/\delta, n, size(c))$ drawn from $\mathcal{D}$, it outputs a hypothesis h from $\mathcal{H}$ such that with probability at least $1-\delta$, the error of h on $\mathcal{D}$ is at most ϵ. Assume a distribution $\mathcal{D}$ on $\mathcal{X} \times \{-1, +1\}$. The learner's goal is to output a hypothesis with such probability over the choice of two training sets (in the stochastic scenario where the output label is a probabilistic function of the input, thus does not guarantee unique labels) and requires both $\mathbb{P}[R(h)_{x\sim\mathcal{D}_+} \leq \epsilon] \geq 1-\delta$ and $\mathbb{P}[R(h)_{x\sim\mathcal{D}_-} \leq \epsilon] \geq 1-\delta$. By the law of total proability for the error rate (or risk),

$$\begin{aligned} R(h)_{\mathcal{D}^m} &= \mathop{\mathbb{E}}_{x\sim\mathcal{D}^m}[\mathbb{1}_{h(x)\neq c(x)}] = \sum_{x\in\mathcal{D}^m} x\mathbb{P}[\mathbb{1}_{h(x)\neq c(x)}] = \mathop{\mathbb{P}}_{x\sim\mathcal{D}^m}[h(x) \neq c(x)] \\ &= \sum_{i=1}^{2} \mathop{\mathbb{P}}_{x\sim\mathcal{D}^m}[h(x) \neq c(x) \mid Oracle_i]\mathbb{P}[Oracle_i] \\ &= \mathcal{D}^m(\mathcal{X}^+)(R(h)_{\mathcal{D}_+}) + \mathcal{D}^m(\mathcal{X}^-)(R(h)_{\mathcal{D}_-}) \end{aligned}$$

Now for a weighted sampling method from the negative and positive instance distributions we can say $\mathcal{D} = \underbrace{\frac{\alpha\mathcal{D}^+ + \beta\mathcal{D}^-}{\alpha+\beta}}_{=1} = \alpha(\mathcal{D}^+ - \mathcal{D}^-) + \mathcal{D}^-$. Note that $\alpha = \beta = 0 \implies \epsilon$ is meaningless, as there are no positive and negative examples in the split dataset. Conversely, by setting $\alpha = 1$, you're only selecting positive instances so that we cannot measure false positives but can measure false negatives (errors in identifying positives). Thus, select an appropriate δ and, for simplicity, set $\alpha = \beta = \frac{1}{2}$, to get $\mathcal{D} = \frac{\mathcal{D}^+ + \mathcal{D}^-}{2}$. Note that $\epsilon/2$ is a magically chosen error bound but this is to make a nice use of union bound simplification later.

$$\begin{aligned} \mathbb{P}[R(h)_{\mathcal{D}} \leq \frac{\epsilon}{2}] &\geq 1-\delta \\ R(h)_{\mathcal{D}^m} &= \frac{1}{2}(h(x) = c(x))_+ + \frac{1}{2}(h(x) = c(x))_- \\ &= \frac{R(h)_{x\sim\mathcal{D}_-} + R(h)_{x\sim\mathcal{D}_+}}{2} \\ \implies \mathbb{P}[\frac{R(h)_{x\sim\mathcal{D}_-} + R(h)_{x\sim\mathcal{D}_+}}{2} \leq \frac{\epsilon}{2}] &= \mathbb{P}[R(h)_{x\sim\mathcal{D}_-} + R(h)_{x\sim\mathcal{D}_+} \leq \epsilon] \\ &\leq \mathbb{P}[R(h)_{x\sim\mathcal{D}_+} \leq \epsilon] + \mathbb{P}[R(h)_{x\sim\mathcal{D}_-} \leq \epsilon] \end{aligned}$$

We must have that both $\mathbb{P}[R(h)_{\mathcal{D}_-} \leq \epsilon]$, $\mathbb{P}[R(h)_{\mathcal{D}_+} \leq \epsilon] \geq 1-\delta$. We immediately notice that a biased dataset would then require us to make considerable adjustments to the overall error rate, as $\mathcal{D}^+$ will shift the weight of distribution of samples. Another thing to note is that by setting α or β to 0, we're essentially requiring that the hypothesis have zero error on positive (or negative) instances (perhaps requiring an absurdly complex hypothesis to do so). This transforms the problem into a rather stringent one-oracle PAC model focused solely on the positive (or negative) class which is not sensitive to noise. The PAC framework is designed to find a balance between minimizing errors on both classes while accounting for uncertainties in real-world data; this decoupled mode tells the learner to query the oracle for instances of one class while imposing stringent error requirements on the other class.

Necessary: Show that if $\mathcal{H}$ is PAC-learnable in the two-oracle variant, it is also PAC-learnable in the standard model.

We now can assume that $\mathcal{C}$ is efficiently PAC-learnable in the two-oracle PAC model so there exists an algorithm $\mathcal{A}$ such that for $c \in \mathcal{C}$, $\epsilon, \delta > 0$, there are m^+ and m^- in $p(1/\epsilon, 1/\delta, size(c))$, such that if we draw

at least this number of negative and positive instances with confidence of at least $1-\delta$, the hypothesis h output by the learner satisfies:

$$R(h)_{\mathcal{D}^{+,-}} \leq \epsilon$$
$$\mathbb{P}[R(h)_{\mathcal{D}^{+,-}}] \leq \mathbb{P}[\epsilon] = \epsilon$$

So, given sufficient numbers of negative and positive examples, we can generate a hypothesis h such that it has low errors on both negative and positive instances. If we draw too few examples, the conclusions about the hypothesis's performance might not hold true for the entire distribution (generalize); to bridge the gap between the variant and standard model it is then best to take $m \geq \max\{m^+, m^-\}$ and, drawing such with polynomial conditions above and using the union bound and total probability,

$$\begin{aligned}\mathbb{P}[R(h)_{\mathcal{D}}] &\leq \mathbb{P}[\mathcal{D}^+]\mathbb{P}(R(h)_{\mathcal{D}_+}) + \mathbb{P}[\mathcal{D}^-]\mathbb{P}(R(h)_{\mathcal{D}_-}) = \mathbb{P}[R(h)_{\mathcal{D}}|c(x) \neq -1]\mathbb{P}[c(x) \neq -1] + \mathbb{P}[R(h)_{\mathcal{D}}|c(x) \neq 1]\mathbb{P}[c(x) \neq 1] \\ &\leq \epsilon(\mathbb{P}[c(x) \neq -1] + \mathbb{P}[c(x) \neq 1]) \leq \epsilon\end{aligned}$$

Let X be the total number of positive examples obtained from drawing m examples, with probability of a positive example of ϵ as shown above. Note $\mathbb{E}[X] = m\epsilon$ and if $\epsilon < 1$, then $\mathbb{E}[X] < m$, indicating that you expect to obtain fewer positive examples on average than the total number of examples drawn.

$$\mathbb{P}\left[\frac{X}{m} \leq (1-\gamma)\epsilon\right] \leq e^{-\frac{m\epsilon\gamma^2}{2}} \implies \mathbb{P}\left[X > (1-\gamma)m\epsilon\right] \leq e^{-\frac{m\epsilon\gamma^2}{2}}$$

Setting (a rather lax) $\gamma = \frac{1}{2}$, required that $\gamma \in [0, 1/\epsilon - 1]$, to match conditions above and needing $m^+ = m(1-\gamma)\epsilon = \frac{m\epsilon}{2}$,

$$\mathbb{P}\left[X > m^+\right] \leq e^{-\frac{m^+}{4}} \leq \frac{\delta}{2}$$

Since we want to include the minimum number of positive instances, we set the latter expression to an appropriate bound $\delta/2$ (thinking ahead to a union bound where positive and negative instances sum to $\leq \delta$) and subsitute back to get an expression of $m \geq \min\{\frac{2m^+}{\epsilon}, \frac{8}{\epsilon}\log(2/(\delta))\}$. A similar procedure is done with the negative case to arrive at $m \geq \min\{\frac{2m^+}{\epsilon}, \frac{2m^-}{\epsilon}, \frac{8}{\epsilon}\log(2/(\delta))\}$ for a balanced dataset.

Exercise 2.2

$$\begin{aligned}\mathbb{P}_{S\sim\mathcal{D}^m}[R(h_S) > \epsilon] &\leq \mathbb{P}_{S\sim\mathcal{D}^m}[\cup_{i=1}^{2n}\{h_S \wedge r_i = \emptyset\}] \\ &\leq \sum_i \mathbb{P}_{S\sim\mathcal{D}^m}[\{h_S \wedge r_i = \emptyset\}] && \text{(union bound)} \\ &\leq 2n(1-\frac{\epsilon}{2n})^m && \mathbb{P}[r_i] \geq \frac{\epsilon}{2n} \\ &\leq (2n)e^{-\frac{m\epsilon}{2n}} \leq \delta \iff m \geq \frac{2n}{\epsilon}\log(2n/\delta)\end{aligned}$$

Exercise 2.3
Define annulus $A := \{a \leq ||x - x_0|| \leq r\}$ with $a := \sup\{r' \mid \mathbb{P}[r' \leq ||x-x_0|| \leq r] > \epsilon\}$. With this construction $\mathbb{P}[A] \geq \epsilon$. If the inner circle $\mathcal{C}'$ intersects the whole annular region for a sample, then the error (disagreement

between target and hypothesis) is at most ϵ. Using the contrapositive,

$$\begin{aligned}
\mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}[error_D(\mathcal{C}') > \epsilon] &\leq \mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}[\{\mathcal{C}' \wedge A = \emptyset\} \text{ for some } i] \\
&\leq \sum_i \mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}[\{\mathcal{C}' \wedge A = \emptyset\}] \\
&= \sum_i \mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}[\{S \wedge A = \emptyset\}] \\
&= \sum_i (1 - \mathbb{P}[A])^m \\
&\leq (1-\epsilon)^m \leq \delta
\end{aligned}$$

and the result follows.

Exercise 2.4
The thinking applied previously suggested that the true (generalization) error being at most a small ϵ was implied by the hypothesis (inner circle) intersecting the error region. The contrapostive then implies that if the error is greater than ϵ, then the hypothesis misses part of the error region. If we just look at the three points provided as the sample, we see that the error is most likely greater than ϵ even when the hypothesis intersects the error region for all instances. This invalidates the first part of the logic so the contrapositive cannot hold to generate the PAC-learning proof.

Exercise 2.5
Assume $\mathbb{P}[ABC] > \epsilon$. We can show that the triangle formed by $A'B'C'$ contains triangle $A''B''C''$. Note that the error is at most ϵ when the area intersects all three regions defined by $A''B''C''$'s segments. As $A'B'C'$ does just this as we see with previous considerations, then $R(h_S = A'B'C')_{\mathcal{D}} \leq \epsilon$. Using the contrapositive, there must be an error greater than ϵ should we miss at least one of the three regions for which $\mathbb{P}[r_i] \geq \frac{\epsilon}{3}$.

$$\begin{aligned}
\mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}[error_D(h_S) > \epsilon] &\leq \mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}[\{h_S, r_i\} = \emptyset] \\
&\leq \sum_{i=1}^{3} \mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}[\{h_S, r_i\} = \emptyset] \leq 3(1 - \frac{\epsilon}{3})^m \leq 3e^{-\frac{-m\epsilon}{3}} \leq \delta
\end{aligned}$$

And the result follows.

Exercise 2.6
The probability that each x_i in the sample will miss region r_j implies that the sample already lied outside the region and so was not flipped to negative or that the sample instance initially laid inside the region but was later flipped to negative.

$$\begin{aligned}
\mathbb{P}[\{x_i \mid x_i \notin r_j \ \cup \ (x_i \in r_j, c(x_i) * 0, c(x_i) = 1)\}] &= \mathbb{P}[\{x_i \mid x_i \notin r_j \ \cup \ (x_i \in r_j, 1 * 0)\}] \\
&= \mathbb{P}[x_i \notin r_j] + \eta\mathbb{P}[x_i \in r_j] \\
&= \mathbb{P}[x_i \notin r_j] + \eta(1 - \mathbb{P}[x_i \notin r_j]) \\
&= \mathbb{P}[x_i \notin r_j](1-\eta) + \eta \\
&\leq (1 - \frac{\epsilon}{4})(1-\eta) + \eta \\
&\leq 1 - \frac{(1-\eta)\epsilon}{4} \\
&\leq 1 - \frac{(1-\eta')\epsilon}{4}
\end{aligned}$$

Now use the usual PAC learning logic,

$$\mathbb{P}[R(R') > \epsilon] \leq \sum_i^4 1 - \frac{(1-\eta')\epsilon}{4} \leq 4(1 - \frac{(1-\eta')\epsilon}{4}) \leq 4e^{-\frac{m\epsilon(1-\eta')}{4}} \leq \delta$$

So for with probability at least $1-\delta$, we'd arrive at an approximately correct hypothesis given $m \geq \frac{4}{(1-\eta')\epsilon}\log\frac{4}{\delta}$. This decays to the original if we remove noise.

Exercise 2.7

- For any hypothesis, let $d(h)$ denote the probability that the label of a training point received by the learner disagrees with the one given by h. The probability that the label of a point be incorrect is η by definition.
- Disagreement between the target function h^* and the hypothesis occurs when the algo fails when the label is correct and when it correctly predicts an incorrect label.

$$\begin{aligned}\mathbb{P}[incorrect] &= \mathbb{P}[c(x)\ correct, c(x) \neq h(x)] + \mathbb{P}[c(x)\ incorrect, c(x) = h(x)] \\ &= R(h_S)(1-\eta) + \eta(1 - R(h_S)) = \eta + (1-2\eta)R(h_S)\end{aligned}$$

- Assuming that $R(h_S) > \epsilon$,

$$d(h) - d(h^*) = (1-2\eta)R(h_S) \geq (1-2\eta)\epsilon \geq (1-2\eta')\epsilon$$

- We use Hoeffding's inequality. Recall that $d(h^*) = \eta$ is a random variable.

$$\mathbb{P}[\hat{d}(h^*) - d(h^*) \geq \epsilon'/2] = \mathbb{P}[\frac{1}{m}(S_m - \mathbb{E}(S_m)) \geq \epsilon'/2] \leq e^{-\frac{2(m\frac{\epsilon'}{2})^2}{\sum_{i=1}^m 1^2}} = e^{-\frac{(m\epsilon')^2}{2m}} \implies m \geq \frac{2}{(\epsilon')^2}\log\frac{2}{\delta}$$

- Similar to before, but using union bounding.

$$\mathbb{P}[h \in \mathcal{H} \mid d(h) - \hat{d}(h) \geq \epsilon'/2] \leq |\mathcal{H}|e^{-\frac{2(\frac{m\epsilon'}{2})^2}{\sum_{i=1}^m 1^2}} = |\mathcal{H}|e^{-\frac{(m\epsilon')^2}{2m}}\frac{2}{(m\epsilon')^2}(\log\frac{2}{\delta} + \log|\mathcal{H}|)$$

- Using the bounding hint provided,

$$\begin{aligned}\hat{d}(h) - \hat{d}(h^*) &= \underbrace{[\hat{d}(h) - d(h)]}_{\geq -\frac{\epsilon'}{2}} + \underbrace{[d(h) - d(h^*)]}_{\geq \epsilon'} + \underbrace{[d(h^*) - \hat{d}(h^*)]}_{\geq -\frac{\epsilon'}{2}} \\ &\geq -\frac{\epsilon'}{2} + \epsilon' - \frac{\epsilon'}{2} = 0\end{aligned}$$

  The fraction of the points in S whose labels disagree with those given by any $h \in \mathcal{H}$ is at least (likely greater) than that for the target function. Thus, the learner will not select this set of poor hypotheses as it cannot minimize this difference bound. We note this learner can be used for PAC-learning despite noise in the consistent ("realizable") case where we make the strong assumption the target $c \in \mathcal{H} \subset \mathcal{C}$.

Exercise 2.8

If we let the section $[a,b]$ be the target concept and assume $\mathbb{P}[[a,b]] > \epsilon$ then we can further define two subsets, one extending from the of the concept such that $L = [a,x]$ knowing $x := \inf\{x' \mid \mathbb{P}[a,x'] > \epsilon/2\}$ and one up to the right $R = [z,b]$ knowing $z := \sup\{x' \mid \mathbb{P}[x',b] > \epsilon/2\}$. We would get at most probability of error ϵ should the hypothesis intersect both intervals, aka be in both the interiors of these regions. Should we find that the bad event of $R(h_S) > \epsilon$ with a consistent hypothesis, we know that at most all the points missed the error intervals, thus that either one or both intervals was missed:

$$\mathbb{P}[R(h_S) > \epsilon] = \mathbb{P}[\{S, r_i\} = \emptyset \text{ for some } i] \leq \sum_i \mathbb{P}[\{S, r_i\} = \emptyset] = 2(1-\epsilon/2)^m \leq 2e^{-\frac{\epsilon}{2}} \leq \delta \implies m \geq \frac{2}{\epsilon}\log\frac{2}{\delta}$$

which means closed intervals also can be considered PAC-learnable.

Exercise 2.9

For simplicity of argument, assume $a < b < c < d$. We can first have a case of the disjoint union of two intervals. If we sort the samples received in ascending order and take note of where consective positive labels are located, this tells us where the (possible) union is. The problem arises in this case: false positives are found where positive instances sampled don't cover a full interval of interest, so the learner assumes some instances are incorrectly false if they lie in the "non-covered" intervals as in the one-interval case studied previously. Also, false positives may occur when no sample instances come from the region between the intervals, which fools the learner into thinking a smaller error region exists. The other case is where both intervals overlap, which decays to the problem of one closed intervals which we saw was PAC-learnable. Analyzing the disjoint case,

$$\begin{aligned}
\mathbb{P}[R(h_S)\ errs] &= R(h_S)_{FP_{(b,c)}} + R(h_S)_{FN_{[a,b]}} + R(h_S)_{FN_{[c,d]}} \\
\mathbb{P}[R(h_S) > \epsilon] &\leq \mathbb{P}[(R(h_S)_{FP_{(b,c)}} > \epsilon/3) \cup (R(h_S)_{FN_{[a,b]}} > \epsilon/3) \cup (R(h_S)_{FN_{[c,d]}} > \epsilon/3)] \\
&\leq \mathbb{P}[R(h_S)_{FP} > \epsilon/3] + \sum_{i=1}^{2} \mathbb{P}[R(h_S)_{FN} > \epsilon/3] \\
&\leq (1-\epsilon/3)^m + 2(2(1-\epsilon/6)^m) \leq e^{-\frac{m\epsilon}{3}} + 4e^{-\frac{m\epsilon}{6}} = e^{-\frac{2m\epsilon}{6}} + 4e^{-\frac{m\epsilon}{6}} \leq 5e^{-\frac{m\epsilon}{6}} \leq \delta \implies m \geq \frac{6}{\epsilon}\log\frac{5}{\delta}
\end{aligned}$$

This shows that the union of closed intervals is also PAC learnable. To generalize to unions of $p \geq 1$ closed intervals, we need to sum both $p-1$ regions of false positives and both of the p regions of false negatives. That is,

$$\begin{aligned}
\mathbb{P}[R(h_S) > \epsilon] &\leq \sum_{i=2}^{p} \mathbb{P}[R(h_S)_{FP} > \epsilon/(2p-1)] + \sum_{i=1}^{2p} \mathbb{P}[R(h_S)_{FN} > \epsilon/(2p-1)] \\
&= (p-1)\mathbb{P}[R(h_S)_{FP} > \epsilon/(2p-1)] + 2p\mathbb{P}[R(h_S)_{FN} > \epsilon/(2p-1)] \\
&= (p-1)(1-\frac{\epsilon}{2p-1})^m + (2p)(1-\frac{\epsilon}{2(2p-1)})^m \leq (p-1)e^{-m\epsilon'} + 2pe^{-\frac{m\epsilon'}{2}} \\
&\leq (3p-1)e^{-\frac{m\epsilon}{2(2p-1)}} \\
&\leq \delta \implies m \geq \frac{2(2p-1)}{\epsilon}\log\frac{3p-1}{\delta}
\end{aligned}$$

Exercise 2.10

For simplicity, we choose the uniform distribution. When you sample uniformly at random from $\mathcal{Z}$, each example has a $1/m$ probability of being included in the random sample. By verifying the consistency of the hypothesis h with a sample of examples from $\mathcal{Z}$, you are effectively assessing the hypothesis's performance on a random subset of the data. If the hypothesis performs well on this random subset, it suggests that it's likely to perform well on the entire dataset, assuming that the sample size is sufficiently large and representative. Thus, if $R(h) \leq 1/m$, we guarantee a consistent hypothesis. To avoid a case where the algorithm makes random guesses ($\epsilon = 0.5$) without much regard for the quality of its predictions, set $\epsilon := 1/(m+1)$.

Exercise 2.11

We first have a consistent case ($\delta = 0.05$) where the target function is assumed to lie in the hypothesis space so that empirical risk is 0. Accordingly,

$$R(h_S) \leq \frac{\log(|\mathcal{H}|/\delta)}{m} = 0.055$$

This is the inconsistent case, where empirical risk is at $m'/m = 0.1$,

$$R(h) \leq \hat{R}(h_S) + \sqrt{\frac{\log(2|\mathcal{H}|/\delta)}{m}} = 0.271$$

Exercise 2.12

First note $\delta' = p(h)\delta = e^{-2m\epsilon^2} \implies \epsilon = \sqrt{\frac{\log(1/(p(h)\delta))}{2m}}$ when considering Hoeffding's inequality with m Bernoulli random variables.

$$\mathbb{P}[R(h) - \hat{R}(h) \geq \sqrt{\frac{\log \frac{1}{\delta p(h)}}{2m}}] \leq p(h)\delta$$

$$\implies \{\exists h \in \mathcal{H} \mid R(h) - \hat{R}(h) \geq \sqrt{\frac{\log \frac{1}{\delta p(h)}}{2m}}\} \leq \sum_i \mathbb{P}[R(h) - \hat{R}(h) \geq \sqrt{\frac{\log \frac{1}{\delta p(h)}}{2m}}] \leq \delta \sum_i p(h) = \delta$$

We could use prior density $1/|\mathcal{H}|$ to match the bound given in the inconsistent case for finite hypothesis sets.

Exercise 2.13

- Note $\hat{R}(h) \leq \frac{3}{4}\epsilon \implies \hat{R}(h) \leq \frac{3}{4}R(h)$ when $R(h) \geq \epsilon \implies \gamma = 0.25$.

$$\mathbb{P}[\hat{R}(h) \leq \frac{3}{4}R(h)] \leq e^{-\frac{nR(h)(1/4)^2}{2}} \leq e^{\frac{R(h)\log\frac{\delta}{2^{i+1}}}{\epsilon}} \leq \frac{\delta}{2^{i+1}}$$

- Note $\hat{R}(h) \geq \frac{3}{4}\epsilon \implies \hat{R}(h) \geq \frac{3}{2}R(h)$ when $R(h) \leq \frac{\epsilon}{2} \implies \gamma = 0.5$. Procedure similar to previous.
- Observe $\mathbb{P}[\textit{approx correct } h_i \textit{ accepted}] = \mathbb{P}[R(h_i) \leq \epsilon/2,\ h \textit{ accepted}] = \mathbb{P}[R(h_i) \leq \epsilon/2]\mathbb{P}[h \textit{ accepted}]$. Note $\frac{\delta}{2^{i+1}} \leq \frac{1}{4}$.

$$\mathbb{P}[h \textit{ accepted} \mid R(h) \leq \epsilon/2] = 1 - \mathbb{P}[R(h_i) \geq \epsilon/2] = 1 - \frac{\delta}{2^{i+1}} \geq \frac{3}{4}$$
$$\mathbb{P}[R(h_i) \leq \epsilon/2] = 1/2$$
$$\mathbb{P}[\textit{approx correct } h_i \textit{ accepted}] \geq \frac{3}{8}$$

- The complement gives us the probability that it does not halt, $\frac{5}{8}$. Observe that $\lceil x \rceil \leq x$ only for integer solutions (what we want).

$$\left(\frac{5}{8}\right)^j \leq \left(\frac{5}{8}\right)^{\frac{\log\frac{2}{\delta}}{\log\frac{8}{5}}} = e^{-\frac{\log\frac{2}{\delta}}{2\log\frac{5}{8}}} \leq e^{-\log\frac{2}{\delta}} = \frac{\delta}{2}$$

- Using the expression given,

$$\tilde{s} \geq s \iff \lfloor 2^{\frac{i-1}{\log\frac{2}{\delta}}} \rfloor \geq s$$
$$\implies 2^{\frac{i-1}{\log\frac{2}{\delta}}} \geq s \iff \frac{i-1}{\log\frac{2}{\delta}} \geq \log_2 s$$
$$\implies i \geq 1 + \log\frac{2}{\delta}(\log_2 s) \geq \lceil 1 + \log\frac{2}{\delta}(\log_2 s) \rceil$$

  Where the last step is for the approriate number of integer solutions (iterations).

- Using the statements above, we see that algo $\mathbb{B}$ will halt with probability of at least $1 - \frac{\delta}{2}$. It will then return approx correct hypothesis with error at most ϵ given the sum of the rather curious ceiling function.

Exercise 2.14

The noise is 0 in a deterministic setting as this mode assumes no randomness in the universe, thus every label is unique. The expectation would be zero for matching labels, which in this case will always be unique for all instances. See:

`https://en.wikipedia.org/wiki/Bayes_classifier`

`https://www.cs.jhu.edu/~ayuille/courses/Stat161-261-Spring14/RevisedLectureNotes2.pdf`

1.3 Summary

PAC learning makes a rather strong assumption that there exists a function (concept), which is a subset of the domain, $c : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{Y} = \{0, 1\}$. We assume that examples are independently and identically distributed (i.i.d.) according to some fixed but unknown data distribution and the learner considers a set of concepts (hypotheses, set $\mathcal{H}$) which may or may not coincide with the "true" set of concepts ($\mathcal{C}$). It receives a sample S drawn i.i.d. according to $\mathcal{D}$ as well as the labels $(c(x_1), \cdots, c(x_m))$, which are based on a specific target concept $c \in \mathcal{C}$ to learn. The task is then to use the labeled sample S to select a hypothesis $h \in \mathcal{H}$ that has a small generalization error with respect to the concept c.

While it would be great to always select a hypothesis for which we get 0 error, since we draw from a training set which have a non-zero probability of have misleading examples (thus bound such cases with δ), there may be several hypothesis consistent with such a sampling; the only way for the learner to pick the optimal solution for the target is to train on all the samples in the domain, or universe (unrealistic). It may also just overfit to memorize instances in a sampling to become a baseline classifier. We relax this assumption with the small ϵ. Note that $\epsilon = 0$ implies that the the learned hypothesis needs to perfectly match the true target function, which can lead to this overfitting and possibly require complex hypotheses. The converse $\epsilon = 1$ would allow for any hypothesis without regard to its performance, hence no learning. To further relax the assumption that (near) zero error be made by all predictors in $\mathcal{H}$ (without knowing if the target concept c lies in $\mathcal{H}$), we insist on a finite set of hypotheses, which requires some inductive bias; that is, we make an assumption that there exists a "good" set of candidate predictors which contain the target concept, $c \in \mathcal{H}$. As before, PAC learning then guarantees an approximately correct (small trues loss) hypothesis (from a possible set of minimizers) given a sufficiently sized sample. In the most (realistic) general case, there may be no hypothesis in $\mathcal{H}$ consistent with the labeled training sample. This gives us more latitude for a fixed $|\mathcal{H}|$, but to attain the same guarantee as in the consistent case, a quadratically larger labeled sample is needed in this tradeoff between error and sample size (see Mohri).

$$\underset{S\sim\mathcal{D}^m}{\mathbb{P}}[h > \epsilon \textit{ inaccurate on average}] \underset{\text{(in-sample error 0, but not for universe)}}{\iff} \underset{S\sim\mathcal{D}^m}{\mathbb{P}}[\text{algo fooled}]$$

$$\mathcal{H}_{bad} = \{\exists h \in \mathcal{C} \mid h(x) = c(x)\ \forall x \in S,\ \textit{but}\ \underset{S\sim\mathcal{D}^m}{\mathbb{P}}[h(x) \neq c(x)] \geq \epsilon\}$$

$$\mathcal{H}_{good} = \{\exists h \in \mathcal{C} \mid h(x) = c(x)\ \forall x \in S,\ \textit{and}\ \underset{S\sim\mathcal{D}^m}{\mathbb{P}}[h(x) \neq c(x)] \leq 1 - \epsilon\}$$

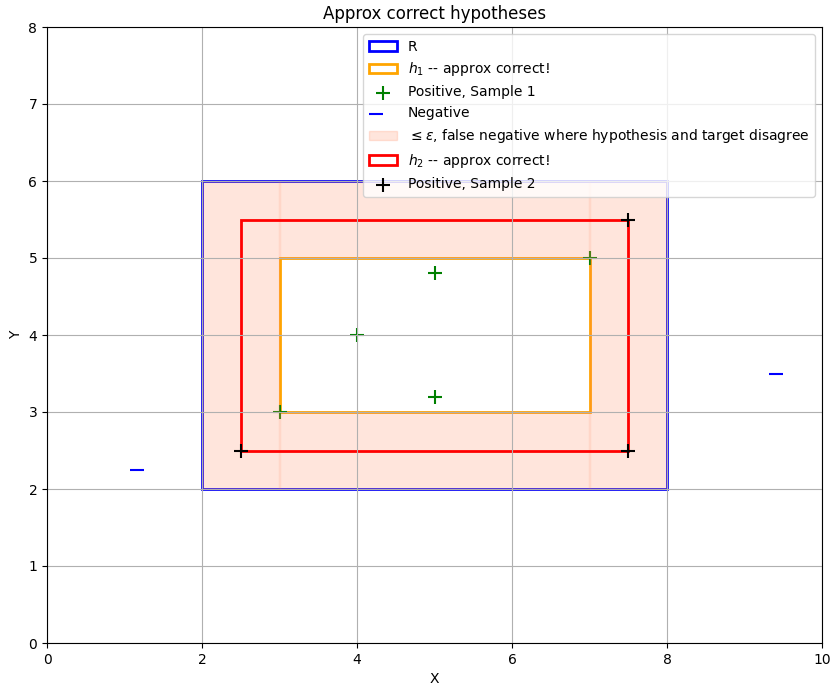
$$\underset{S\sim\mathcal{D}^m}{\mathbb{P}}[\mathcal{H}_{bad} \text{ non-empty, so algo may output one of its members}] \leq \delta$$

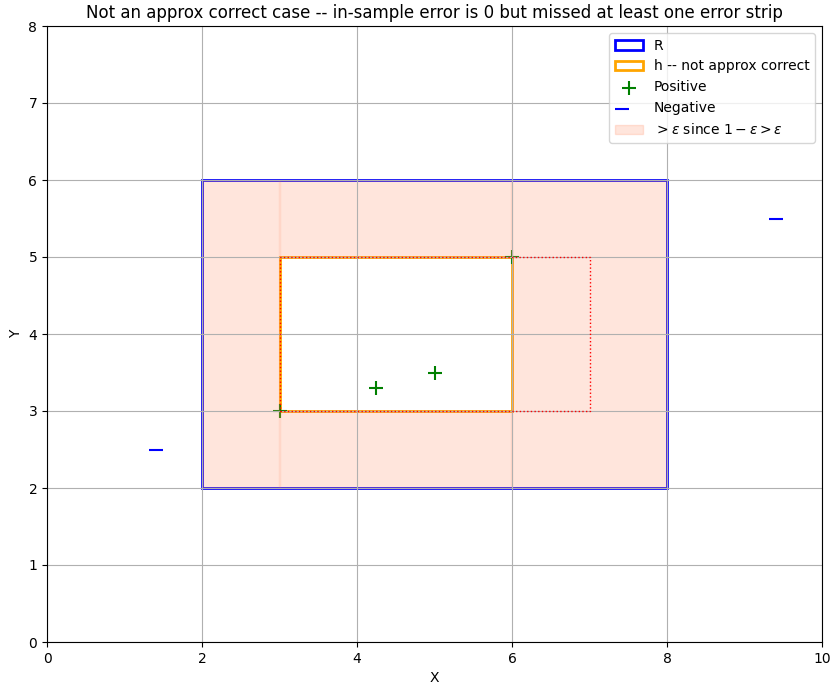
For given $S, \mathcal{H}_{good}$ is never empty as it must at least contain the target concept (later relaxed, only for "realizable" version)

Note that the proof of consistent, finite set case relied on the underlying technique of sampling from the data distribution and then determining a good generalizable hypothesis after assessing all in-sample errors. That is, there may be a set of (poor) samples (x_i's) for which the true risk under a chosen hypothesis is larger than ϵ yet have 0 empirical risk for a sample. We would want to avoid the set of all bad hypotheses for which this "bad" condition holds. The set of all samples that will lead the chosen (bad) hypothesis to produce error greater than ϵ is a subset of the set of all samples for which there exists some (bad) hypothesis that produces zero in-sample error (which we may or may not have chosen). Thus, by bounding the superset (and thus further constricting the subset of interest), we restrict such hypotheses becoming godly predictors for all instances in a sample, thus bounded by a small δ.

Wrapping up: good learners will learn with high probability and close approximation to the target concept. With high probability, we find a hypothesis that will have low error (approximately correct), requiring parameters $\epsilon, \delta > 0$. Thus, with probability of at least $1 - \delta$, the algorithm learns the concept with error at most ϵ. The ϵ is the upper bound on the accuracy $(1-\epsilon)$. We use δ to bound the probability of failure in achieving the set accuracy (a bad event) and thus want a hypothesis to be approximately correct with confidence $1-\delta$. PAC is powerful in that it requires no assumption about the underlying distribution but has the rather unrealistic requirement that the labeling rule (target concept) is in the hypothesis space. Agnostic (i.e., not realizable) PAC learning relaxes this assumption; in scenarios for which there may exist no such

concept, and instead that all distributions produce labels stochastically (and thus may not be unique),we are left with a minimal non-zero error for any hypothesis.





Chapter 2

Rademacher Complexity and VC-Dimension

RC allows us to measure the richness of a class of real-valued functions with respect to a probability distribution. The standard "intuition" is that the RC quantifies the ability of the function class $\mathcal{G}$ to fit symmetric random noise: a low value (close to 0) means that this ability is limited (and hence the capacity of this class is bounded), while a high value (close to 1 for $\{-1, 1\}$-valued classes) means that essentially any sequence of random signed bits has a perfect fit (and hence the capacity is unbounded)."Capacity" is not a formally defined term and notice that I put it in quotes. Roughly speaking this notion of capacity measures the ability of a function class to fit random noise. In statistical learning theory, a small expected Rademacher complexity indicates that, on average, the hypothesis class does not correlate strongly with random noise. This is desirable because it suggests that the class is not overly flexible and doesn't fit the noise in the data but rather captures the underlying patterns. VC-dim is a worst-case measure of this, while Rademacher is more of an average-case measure.
Further reading:
https://www.stat.cmu.edu/~larry/=sml/Concentration.pdf
https://cs.brown.edu/courses/csci1951-w/lec/lec%203%20notes.pdf
https://cstheory.stackexchange.com/questions/47879/whats-the-intuition-behind-rademacher-complexity
https://ocw.mit.edu/courses/18-465-topics-in-statistics-statistical-learning-theory-spring-2007/resources/l10/
https://people.math.binghamton.edu/qiao/math605/book/rademacher-complexity.html
https://www.cs.princeton.edu/courses/archive/spring19/cos511/scribe_notes/0220.pdf
https://www.cs.princeton.edu/courses/archive/spring19/cos511/scribe_notes/0218.pdf
https://www.cs.princeton.edu/courses/archive/spr08/cos511/scribe_notes/0220.pdf
https://www.cs.princeton.edu/courses/archive/spring19/cos511/scribe_notes/0225.pdf
https://users.cs.utah.edu/~bhaskara/courses/theoryml/scribes/lecture6.pdf
https://engineering.purdue.edu/ChanGroup/ECE595/files/Lecture26_growth.pdf
https://nowak.ece.wisc.edu/SLT09/lecture19.pdf
https://cse.iitkgp.ac.in/~saptarshi/courses/ml2018spring/vc_inequality_proof.pdf

2.1 Priors

Some "missing" proofs in the text are reproduced below and adjusted for clarity and notation where appropriate. These are fairly involved and focus on how to derive the growth function bounds directly, without using Rademacher complexity bounds first as Professor Mohri mentions. See links above for reference.
Intuition: although model class is infinite, using a finite set of training data to select a good rule effectively reduces the number of different models we need to consider (R. Nowak). For m smaller of equal to VC

dimension for that hypothesis space we have that the largest set that can be shattered by the space is 2^m. For m greater than the VC dimension for this space, then $\prod_{\mathcal{H}}(m) < 2^m$. Original proofs to follow are involved but its first key to get the effective size of the class induced by the training data by a "ghost sample" that is another sequence of data in all identical to the data generating distribution (aka permutation). Let's say $S = \{(X_1, Y_1), \ldots, (X_n, Y_n)\} \mid (X_i, Y_i) \sim P_{XY}$ in i.i.d fashion. The ghost sample follows with S' notation. Think of this ghost sample as a proxy for the generalization error. We now want to show

$$\mathbb{P}_S\big(\sup_{h\in\mathcal{H}} |\hat{R}_S(h) - R(h)| > \epsilon\big) \leq 2\mathbb{P}_{S,S'}\big(\sup_{h\in\mathcal{H}} |\hat{R}_S(h) - \hat{R}'_S(h)| > \frac{\epsilon}{2}\big)$$

The empirical risk $\hat{R}$ is the average disagreement between our proposed learning rule for the training sample and the true labeling for the joint distribution. RHS absolute value is now symmetric with two empirical risks here, and we define $\tilde{h}(S) \equiv \tilde{h}$ to be an element in $\mathcal{H}$ such that, should $|\hat{R}_S(h) - R(h)| > \epsilon$ exist, be defined as this element or, if not, an arbitrary element in this set. While it can be thought of as

$$\tilde{h} \approx \arg\max_{h\in\mathcal{H}} |\hat{R}_S(h) - R(h)|$$

though this is crude as the hypothesis set is usually infinite and there may not be an element serving as the maximizer. Now the RHS implies the following (note for reals $|x - z| > \epsilon \wedge |y - z| \leq \epsilon/2 \implies |x - y| \geq \epsilon/2$):

$$\begin{aligned}
\mathbb{P}\big(\sup_{h\in\mathcal{H}} |\hat{R}_S(h) - \hat{R}'(h)| > \frac{\epsilon}{2}\big) &\geq \mathbb{P}\big(|\hat{R}_S(h) - \hat{R}'(h)| > \frac{\epsilon}{2}\big) \\
&\geq \mathbb{P}\big(|\hat{R}_S(\tilde{h}) - R(\tilde{h})| > \epsilon \wedge |\hat{R}'_S(\tilde{h}) - R(\tilde{h})| < \frac{\epsilon}{2}\big) \\
&= \mathbb{E}_{S,S'}\big[\mathbb{1}\{|\hat{R}_S(\tilde{h}) - R(\tilde{h})| > \epsilon\}\mathbb{1}\big\{|\hat{R}'_S(\tilde{h}) - R(\tilde{h})| < \frac{\epsilon}{2}\big\}\big] \\
&= \mathbb{E}_{S,S'}\big[\mathbb{1}\{|\hat{R}_S(\tilde{h}) - R(\tilde{h})| > \epsilon\}\mathbb{E}_S\big[\mathbb{1}\big\{|\hat{R}'_S(\tilde{h}) - R(\tilde{h})| < \frac{\epsilon}{2}\big\} \mid S\big]\big] \\
&= \mathbb{E}_{S,S'}\big[\mathbb{1}\{|\hat{R}_S(\tilde{h}) - R(\tilde{h})| > \epsilon\}\mathbb{P}\big(|\hat{R}'_S(\tilde{h}) - R(\tilde{h})| < \frac{\epsilon}{2} \mid S\big)\big]
\end{aligned}$$

By conditioning on the sample, $\hat{R}'_S(\tilde{h}) - R(\tilde{h}) = \frac{1}{m}\sum_i U_i$ for $U_i = \mathbb{1}\big\{\tilde{h}(X_i) \neq Y_i\big\} - \mathbb{E}\big[\mathbb{1}\big\{\tilde{h}(X'_i) \neq Y'_i\big\}\big]$ being a zero-mean i.i.d. set RVs. We now use Chebyshev's inequality (see C.14) in the following:

$$\begin{aligned}
\mathbb{P}\big(|\hat{R}'_S(\tilde{h}) - R(\tilde{h})| < \frac{\epsilon}{2}|S\big) &= \mathbb{P}\Big(|\frac{1}{m}\sum_i^m U_i| < \frac{\epsilon}{2}|S\Big) \\
&= \mathbb{P}\Big(|\sum_i^n U_i| < \frac{m\epsilon}{2}|S\Big) \\
&\geq 1 - 4\frac{2}{m^2\epsilon^2}\operatorname{Var}\Big(|\sum U_i||S\Big) \\
&= 1 - 4\frac{2}{m^2\epsilon^2} m \operatorname{Var}(U_i|S) \\
&\geq 1 - 4\frac{2}{m^2\epsilon^2}\frac{1}{2} = 1 - \frac{1}{m\epsilon^2} \geq \frac{1}{2}
\end{aligned}$$

which assumes the denominator $m\epsilon^2 \geq 2$ to avoid a non-informative case. Thus,

$$\begin{aligned}
\mathbb{P}\big(\sup_{h\in\mathcal{H}} |\hat{R}_S(h) - \hat{R}'_S(h)| \geq \frac{\epsilon}{2}\big) & \\
&\geq \mathbb{E}_{S,S'}\big[\mathbb{1}\{|\hat{R}_S(\tilde{h}) - R(\tilde{h})| > \epsilon\}\mathbb{P}\big(|\tilde{R}'_S(\tilde{h}) - R(\tilde{h}) < \frac{\epsilon}{2}|S\big)\big] \\
&\geq \frac{1}{2}\mathbb{E}_{S,S'}\big[\mathbb{1}\{|\hat{R}_S(\tilde{h}) - R(\tilde{h})| > \epsilon\}\big] \\
&= \frac{1}{2}\mathbb{P}\big\{|\hat{R}_S(\tilde{h}) - R(\tilde{h})| > \epsilon\big\} \\
&\geq \frac{1}{2}\mathbb{P}\big\{\sup_{h\in\mathcal{H}} |\hat{R}_S(h) - R(h)| > \epsilon\big\}
\end{aligned}$$

and the result follows. Now instead of symmetrization by the ghost sample, we will symmetrize by the random signs. We observe that $\mathbb{1}\{f(X_i) \neq Y_i\}$ and $\mathbb{1}\{f(X'_i) \neq Y'_i\}$ have the same distribution and thus their difference has zero mean and is symmetric, i.e., -Z and Z having the same distribution for some RV Z.

$$\begin{aligned}
\mathbb{P}(\sup_{h\in\mathcal{H}} |\hat{R}_S(h) - \hat{R}'_S(h)| > \frac{\epsilon}{2}) &= \mathbb{P}\big(\sup_{h\in\mathcal{H}} \frac{1}{m}|\sum_{i=1}^{m} \mathbb{1}_{h(X_i)\neq Y_i} - \mathbb{1}_{h(X'_i)-Y'_i}| > \frac{\epsilon}{2}\big) \\
&= \mathbb{P}\big(\sup_{h\mathcal{H}} \frac{1}{m}|\sum_{i=1}^{m} \sigma_i(\mathbb{1}_{h(X_i)\neq Y_i} - \mathbb{1}_{h(X'_i)-Y'_i})| > \frac{\epsilon}{2}\big) \\
&\leq \mathbb{P}\Big(\sup_{h\in\mathcal{H}} \frac{1}{m}|\sum_{i} \sigma_i\mathbb{1}_{h(X_i)\neq Y_i}| > \frac{\epsilon}{4} \cup \sup_{h\in\mathcal{H}} \frac{1}{m}|\sum_{i} \sigma_i\mathbb{1}_{h(X'_i)\neq Y'_i}| > \frac{\epsilon}{4}\Big) \\
&= 2\mathbb{P}\big(\sup_{h\in\mathcal{H}} \frac{1}{m}|\sum_{i} \sigma_i\mathbb{1}_{h(X_i)\neq Y_i}| > \frac{\epsilon}{4}\big)
\end{aligned}$$

Where union bounding is used in the last steps. The Rademacher RVs are symmetric and so all of this is used to imply

$$\mathbb{P}\big(\sup_{h\in\mathcal{H}} |\hat{R}_S(h) - R(h)| > \epsilon\big) \leq 4\mathbb{P}\Big(\sup_{h\in\mathcal{H}} \frac{1}{m}|\sum_{i=1}^{m} \sigma_i\mathbb{1}\{h(X_i) \neq Y_i\}| > \frac{\epsilon}{4}\Big)$$

Since the ghost sample has now been occluded we can condition on the original training sample as it relates the the cardinality of the hypothesis set. Note that the supremum's argument on the RHS has a sequence that can take at most $\prod_{\mathcal{H}}(m)$ different values. Now, if we take the smallest subset of $\mathcal{H}$ that generates all the different prediction rules for the data (called dichotomies, $\mathcal{H}(x_1, \ldots, x_m) \subseteq \mathcal{H}$), then $|\mathcal{H}(x_1, \ldots, x_m)| \leq \prod_{\mathcal{H}}(m)$. Observe that the growth function is concerned with the number of different ways all the hypotheses in the space can classify points, while this subset measures all possible prediction rules for a particular dataset using the fewest hypotheses in that space.

$$\begin{aligned}
\mathbb{P}\big(\sup_{h\in\mathcal{H}} |\frac{1}{m}\sum_{i} \sigma_i\mathbb{1}_{h(x_i)\neq y_i}| > \frac{\epsilon}{4}\big) &= \mathbb{P}\Big(\max_{h\in\mathcal{H}(x_1,\ldots,x_m)} \frac{1}{m}|\sum_{i} \sigma_i\mathbb{1}_{h(x_i)\neq y_i}| > \frac{\epsilon}{4}\Big) \\
&= \mathbb{P}\big(\bigcup_{h\in\mathcal{H}(x_1,\ldots,x_m)} \frac{1}{m}|\sum_{i} \sigma_i\mathbb{1}_{h(x_i)\neq y_i}| > \frac{\epsilon}{4}\big) \\
&\leq \sum_{h\in\mathcal{H}(x_1,\ldots,x_m)} \mathbb{P}\Big(\frac{1}{m}|\sum_{i} \sigma_i\mathbb{1}_{h(x_i)\neq y_i}| > \frac{\epsilon}{4}\Big) \\
&\leq |\mathcal{H}(x_1, \ldots, x_m)| \sup_{h\in\mathcal{H}(x_1,\ldots,x_m)} \mathbb{P}\big(\frac{1}{m}|\sum \sigma_i\mathbb{1}_{h(x_i)\neq y_i}| > \frac{\epsilon}{4}\big) \\
&\leq \prod_{\mathcal{H}}(m) \sup_{h\in\mathcal{H}(x_1,\ldots,x_m)} \mathbb{P}\big(|\sum_{i} \sigma_i\mathbb{1}_{h(x_i)\neq y_i}| > \frac{\epsilon}{4}\big) \\
&\leq \prod_{\mathcal{H}}(m) \sup_{h\in\mathcal{H}} \mathbb{P}\big(\frac{1}{m}|\sum_{i} \sigma_i\mathbb{1}_{h(x_i)\neq y_i}| > \frac{\epsilon}{4}\big)
\end{aligned}$$

Now note we can use Hoeffding's inequlaity since the argument inside the inner sum is bounded, $[-1, 1]$, and is a sum of m independent, zero-mean RVs.

$$\mathbb{P}\Big(\frac{1}{m}|\sum_{i=1}^{m}\sigma_i\mathbb{1}_{h(x_i)\neq y_i}| > \frac{\epsilon}{4}\Big) \leq \mathbb{P}\Big(|\sum_{i=1}^{m}\sigma_i\mathbb{1}_{h(x_i)\neq y_i}| > \frac{m\epsilon}{4}\Big) \leq 2e^{-\frac{2(m\epsilon/4)^2}{\sum_{i=1}^{m}(1+1)^2}} = 2e^{-\frac{m\epsilon^2}{32}}$$

Now if we look at the RHS from before,

$$\begin{aligned}
\mathbb{P}\Big(\sup_{h\in\mathcal{H}}\frac{1}{m}|\sum_{i=1}^{m}\sigma_i\mathbb{1}_{f(X_i)\neq Y_i}| > \frac{\epsilon}{4}\Big) &= \mathbb{E}_{\sigma,S}\big[\mathbb{1}\big\{\sup_{h\in\mathcal{H}}\frac{1}{m}|\sum_{i=1}^{m}\sigma_i\mathbb{1}_{f(X_i)\neq Y_i}\big\} > \frac{\epsilon}{4}\big] \\
&= \mathbb{E}_S\Big[\mathbb{E}_\sigma\big[\mathbb{1}\Big\{\sup_{h\in\mathcal{H}}\frac{1}{m}|\sum_{i=1}^{m}\sigma_i\mathbb{1}\big\{f(X_i)\neq Y_i\big\}|\Big\}|S\big]\Big] \\
&= \mathbb{E}_S\big[\mathbb{P}\big(\sup_{h\in\mathcal{H}}\frac{1}{m}|\sum_{i=1}^{m}\sigma_i\mathbb{1}_{f(X_i)\neq Y_i}| > \frac{\epsilon}{4}|S\big)\big] \\
&\leq \prod_{\mathcal{H}}(m)\mathbb{E}_S\Big[\sup_{h\in\mathcal{H}}\mathbb{P}\big(\frac{1}{m}|\sum_{i=1}^{m}\sigma_i\mathbb{1}_{f(X_i)\neq Y_i}| > \frac{\epsilon}{4}|S\big)\Big] \\
&\leq \prod_{\mathcal{H}}(m)\mathbb{E}_S\Big[2e^{-\frac{m\epsilon^2}{32}}|S\Big] = 2\prod_{\mathcal{H}}(m)e^{-\frac{m\epsilon^2}{32}}
\end{aligned}$$

Now recall the previous relation and connect the dots,

$$\begin{aligned}
\mathbb{P}\big(\sup_{h\in\mathcal{H}}|\hat{R}_S(h) - R(h)| > \epsilon\big) &\leq 4\mathbb{P}\Big(\sup_{h\in\mathcal{H}}\frac{1}{m}|\sum_{i=1}^{m}\sigma_i\mathbb{1}\{h(X_i)\neq Y_i\}| > \frac{\epsilon}{2}\Big) \\
&\leq 8\prod_{\mathcal{H}}(m)e^{-\frac{m\epsilon^2}{32}}
\end{aligned}$$

The issue here is that Hoeffding's inequality does not use any information about the random variables except the fact that they are bounded. If the variance of the RVs is small, then we can get a sharper inequality from Bernstein's inequality (see Lafferty, et al) or use a method of permutations the original work cites (omitted here but can be found in VC paper and link to Caltech proof above). We then have finally shown how growth function bounds can be also derived directly without using Rademacher complexity bounds first. This then describes that the pessimistic or worst-case bound (upper bound) for the event where the most improper learning rule occurs (though learning is possible given enough data due to the negative exponential on the RHS), which can described by just knowing the sample points and their growth function, and not rely on the complexity of the possibly infinite hypothesis space. Note that the growth function is the number of dichotomies, which takes a hypothesis on a finite set of points and that many different hypotheses may return the same dichotomies when applied to those points. The growth function can only be either polynomial (any break point, so learning is possible) or exponential (no break point, so term may overwhelm the negative exponential and we would not be able to learn); to clarify, if I have enough examples to look at and know the hypotheses set, then I can determine the learning behavior (think: positive rays, positive intervals, or convex sets) without knowing how the hypothesis set would exactly map to the training data first. In sum, while these bounds may not be the tightest possible, given enough examples, we can generalize learning from a finite hypothesis set to its full space in probability – a great trade-off given the countably infinite set. The VC-dimension is also a purely combinatorial notion but it is often easier to compute than the growth function (or the Rademacher Complexity), as Mohri describes.

The VC dimension is the largest number of data points which the growth function is 2^N (i.e., all classifications of the positive and negative class are possible). You just need at least one set of such points to be fully realized. Another thing to notice is that, $N \leq d_{VC} \implies \mathcal{H}$ can shatter N points. If $N > d_{VC} \implies N$ is a break point for $\mathcal{H}$. In other words if $d_{VC} = \infty$, we always get exponential growth

function; however, if $d_{VC} < \infty$, the growth function increases exponentially up to d and polynomially for $N > d$, as given by Sauer's lemma. We are guaranteed to be able to learn if we have the polynomial case when we look at the related Hoeffding's inequality. Thus, VC says that if there exists a set of N points that can be shattered by the classifier (note: notice how it doesn't say any set of shattered N points) and there is no set of $N + 1$ points that can be shattered by the classifier, then the VC dimension of the classifier is N. If a classifier's VC dimension is 3, it does not have to shatter all possible arrangements of 3 points. If of all arrangements of 3 points you can find at least one such arrangement that can be shattered by the classifier, yet cannot find 4 points that can be shattered, then VC dimension is 3.

2.2 Exercises

Exercise 3.1
TBD