Homework 3

Sara Beery
CS 156A - Learning Systems

October 29, 2018

1 Generalization Error

For an H with $d_{vc} = 10$, if you want 95% confidence that your generalization error is at most 0.05, what is the closest numerical approximation of the sample size that the VC generalization bound predicts?
First, recall

$$P[|Ein(g) - Eout(g)| > \epsilon] \leq 4m_H(N)e^{-\frac{1}{8}\epsilon^2 N}, \ \forall \epsilon > 0.$$

We are told to use the approximate bound $m_H(N) \leq N^{d+vc}$ for $N > d_{vc}$. We set $\epsilon = 0.05$ (corresponding to a 95% confidence) and want to find N such that the probability bound $4m_H(N)e^{-\frac{1}{8}\epsilon^2 N}$ is at most 0.05. In order to do this, we solve

$$N^{d_{vc}} = e^{\frac{1}{8}(0.5)^2 N}\left(\frac{0.5}{4 * 2^{d_{vc}}}\right),$$

numerically using Mathematica and find the most reasonable root is $N = 452,957$ which is closest to

1. **[d]** 460,000

There are a number of bounds on the generalization error ϵ, all holding with probability at least $1 - \delta$. Fix $d_{vc} = 50$ and $\delta = 0.05$ and plot these bounds as a function of N. Which bound is the smallest for very N when N is either small or large? Note that [c] and [d] are implicit bounds in ϵ, and note that for $N > d_v c$ we should use $m_H(N) \leq N^{d+vc}$, and for $N \leq d_v c$ we should use $m_H(N) = 2^N$

2. **[d]** Devroye, as seen in Fig. 1.

3. **[c]** Parrondo and Van den Broek, as seen in Fig. 2.

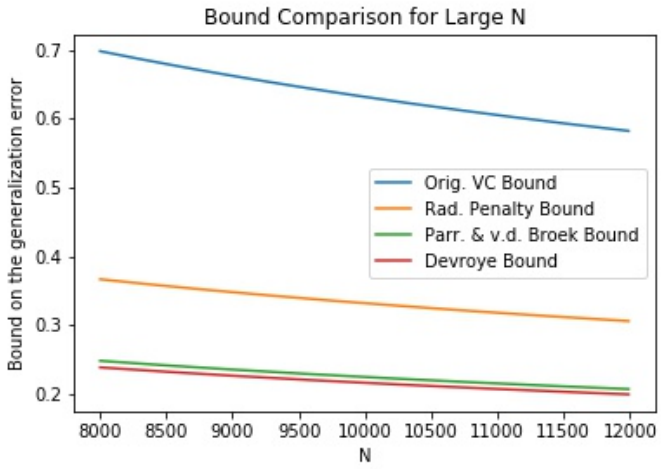

Figure 1: Looking at the bounds for large N (specifically $N = 10,000$), we see the smallest bound is the Devroye bound.

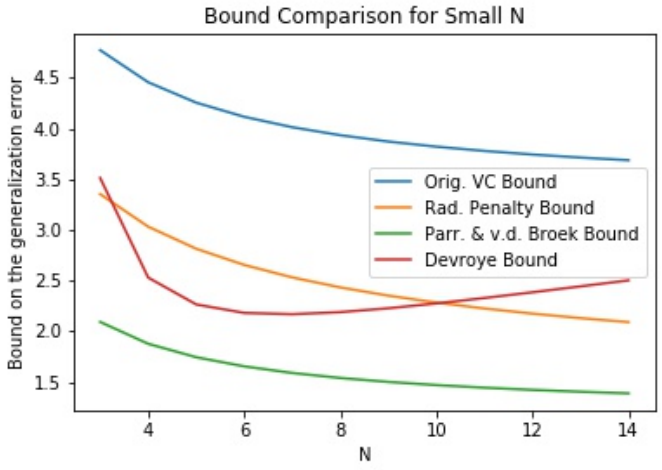

Figure 2: Looking at the bounds for small N (specifically $N = 5$), we see the smallest bound is the Parrondo and Van den Broek bound.

2 Bias and Variance

Consider the case where the target function $f : [1, 1] \leftarrow \mathbb{R}$ is given by $f(x) = sin(\pi x)$ and the input probability distribution is uniform on $[1, 1]$. Assume that the training set has only two examples (picked independently), and that the learning algorithm produces the hypothesis that minimizes the mean squared error on the examples.

Assume the learning model consists of all hypotheses of the form $h(x) = ax$. What is the expected value, $\overline{g}(x)$, of the hypothesis produced by the learning algorithm (expected value with respect to the data set)?

First, let us approximate

$$\overline{g}(x) = \frac{1}{K} \sum_{k=1}^{K} a_{D_k} x$$

Note that for each dataset D we get two input points, x_1 and x_2, which are drawn uniformly from $[1, 1]$, and we want to find

$$a = argmin_{a \in \mathbb{R}}(ax_1 - sin(\pi x_1)^2 + (ax_2 - sin(\pi x_2)^2$$

Which can be found by solving for the roots of the derivative of the above with respect to a, namely solving for a in

$$0 = 2x_1(ax_1 - sin(\pi x_1) + 2x_2(ax_2 - sin(\pi x_2),$$

which becomes

$$a = \frac{x_1 sin(\pi x_1) + x_2 sin(\pi x_2)}{x_1^2 + x_2^2}.$$

Now, we can solve for the expected value over D explicitly with

$$\mathbb{E}_D[a] = \int_{-1}^{1} \int_{-1}^{1} \frac{x_1 sin(\pi x_1) + x_2 sin(\pi x_2)}{x_1^2 + x_2^2} x_1 x_2 dx_1 dx_2$$

or we can approximate the expected value over D empirically by taking K to be large, and averaging over a_k for $k \in K$.

We chose to do the latter, and found that $\mathbb{E}_D[a] = \overline{a} = 0.42$. Thus:

4. **[e]** None of the above.

We also chose to empirically calculate bias and variance, with

$$bias = average_i(\overline{a}x_i - sin(\pi x_i)^2$$

and

$$variance = average_i(average_{k \in K}(a_k x_i - \overline{a}x_i)^2)$$

and find $bias = 0.273$ and $variance = 0.238$, which is closest to

5. [**b**] 0.3 and **6.** [**a**] 0.2.

Next, to further understand bias/variance tradeoff, we consider alternative hypothesis sets.

Which of the following learning hypothesis target functions has the least expected value of out-of-sample error?

$$(a)\ h(x) = b$$
$$(b)\ h(x) = ax$$
$$(c)\ h(x) = ax + b$$
$$(d)\ h(x) = ax^2$$
$$(e)\ h(x) = ax^2 + b$$

We consider each empirically, and find that $h(x) = ax$ has the smallest expected out-of-sample error (see Fig. 3).

Thus **7.** [**b**] $h(x) = ax$

3 VC Dimension

Let $q \geq 1$ be an integer, and assume that $m_H(1) = 2$. What is the VC dimension of a hypothesis set whose growth function for all $N \geq 1$ satisfies: $m_H(N+1) = 2m_H(N) - \binom{N}{q}$? Recall that $\binom{M}{m} = 0\ if M < m$.

There are 3 cases to investigate: $q > N$, $q = N$, $q < N$. Assume (without loss of generality, since $m_H(N)$ is defined recursively) that we have $d_vc > N - 1$, which implies $m_H(N-1) = 2^{N-1}$.

If $q > N$ we have

$$m_H(N) = 2 * 2^{N-1} - \binom{N}{q} = 2^N - 0 = 2^N \tag{1}$$

which means we can shatter N points, and $d_vc \geq N$

If $q = N$ have

$$m_H(N) = 2 * 2^{N-1} - \binom{N}{q} = 2^N - \binom{N}{N} = 2^N - 1 \leq 2^N \tag{2}$$

which means that we cannot shatter N points, and $d_vc < N$. The combination of the these two means that $d_vc = q - 1$, and we do not need to consider the third case. The largest number of points that can be shattered by our hypothesis set will always be $q - 1$. Thus:

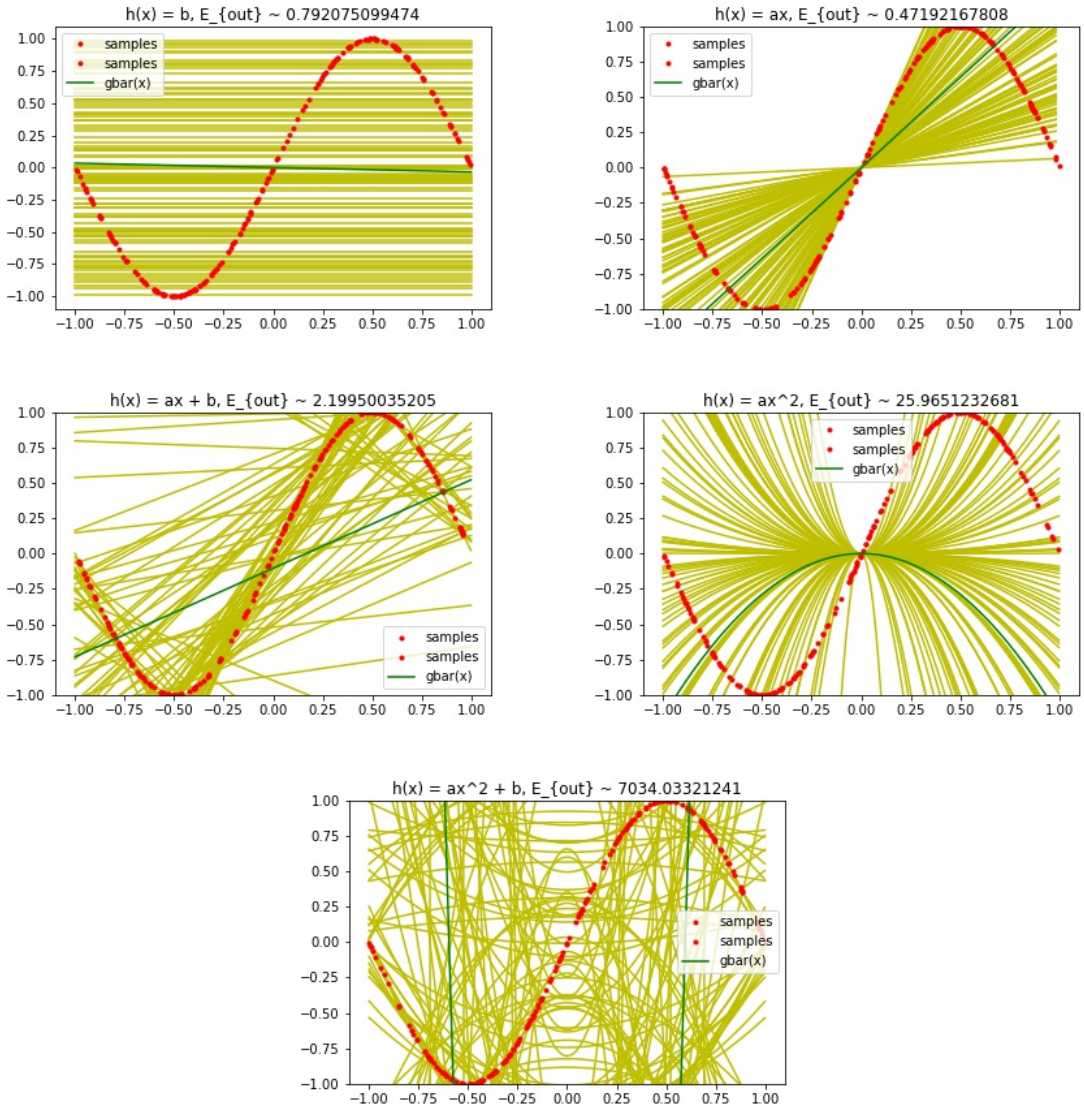

Figure 3: Bias/variance tradeoff for various hypothesis target functions, using an empirical average over 100 datasets

8. **[b]** $q - 1$

For hypothesis sets $H_1, H_2, ..., H_K$ with finite, positive VC dimensions $d_{vc}(H_k)$ (same input space X), some of the following bounds are correct and some are not. Which, among the correct ones, is the tightest bound (the smallest range of values) on the VC dimension of the intersection of the sets: $d_{vc}(\cap_{k=1}^{K} H_k)$? (The VC dimension of an empty set or a singleton set is taken as zero)

If k such that $d_{vc}(H_k) = 0$ then $d_{vc}(\cap_{k=1}^{K} H_k) = 0$, and $d_{vc} \geq 0$ by construction, so the tightest lower bound will be 0.

Note that

$$\cap_{k=1}^{K} H_k \subseteq H_k \ \forall k.$$

Therefore

$$d_{vc}(\cap_{k=1}^{K} H_k) \leq d_{vc}(H_k) \ \forall k$$

and thus

$$d_{vc}(\cap_{k=1}^{K} H_k) \leq min_k\{d_{vc}(H_k)\}$$

and we see:

9. **[b]** $0 \leq d_{vc}(\cap_{k=1}^{K} H_k) \leq min_k\{H_k\}$

Now consider $d_{vc}(\cup_{k=1}^{K} H_k)$, and note that similarly we have

$$H_k \subseteq \cup_{k=1}^{K} H_k \ \forall k.$$

Therefore

$$d_{vc}(H_k) \leq d_{vc}(\cup_{k=1}^{K} H_k) \ \forall k.$$

and thus

$$max_k\{d_{vc}(H_k)\} \leq d_{vc}(\cup_{k=1}^{K} H_k)$$

is a lower bound. An upper bound is more challenging to derive, but we can show that $\sum_{k=1}^{K} d_{vc}(H_k)$ is not an upper bound. Consider for the sake of counterexample the hypotheses sets $H_1 = \{h(x) = +1\}$ and $H_2 = \{h(x) = -1\}$. Note that $d_{vc}(H_1) = 0$ and $d_{vc}(H_2) = 0$, but that $\cup_{k=1}^{2} H_k = \{h(x) = +1, h(x) = -1\}$ and $d_{vc}(\cup_{k=1}^{K} H_k) = 1 > \sum_{k=1}^{2} d_{vc}(H_k) = 0$.

Now, looking at the above counterexample, we see that the additional VC dimension comes about through a *contradiction*, which gives you the ability to choose from a set of hypotheses based on the data points shown and lets you shatter what would have been impossible with any one of the hypotheses by allowing you the choice between the hypotheses (essentially an additional degree of freedom). When we take a union over a set of hypothesis sets, if we think of it sequentially, we can interpret this as:

$$\begin{aligned} &step\ 1: && d_{vc}^{(1)} = d_{vc}(H_1), \\ &step\ 2: && d_{vc}^{(2)} \leq d_{vc}^{(1)} + d_{vc}(H_2) + 1, \end{aligned}$$

where the 1 comes from our ability to choose EITHER to take a hypothesis from H_1 or from H_2, and will occur if we have contradictory hypotheses in the two sets that open up additional dichotomies that neither set alone could create (like in the above counterexample).

$$step\ 3: \qquad d_{vc}^{(3)} \leq d_{vc}^{(2)} + d_{vc}(H_3) + 1,$$

where the 1 comes from our ability to choose EITHER to take a hypothesis from $\cup_{k=1}^{2} H_k$ or from H_3,

$$step\ 4: \qquad d_{vc}^{(2)} \leq d_{vc}^{(3)} + d_{vc}(H_4) + 1,$$

where the 1 comes from our ability to choose EITHER to take a hypothesis from $\cup_{k=1}^{3} H_k$ or from H_4, and in the general case at any step we have:

$$step\ n: \qquad d_{vc}^{(n)} \leq d_{vc}^{(n-1)} + d_{vc}(H_n) + 1.$$

In the worst case, if we have equality to the bound for each step, this will recursively give us:

$$step\ n: \qquad d_{vc}^{(n)} = K - 1 + \sum_{k=1}^{2} d_{vc}(H_k),$$

and thus:

10. **[e]** $max_k\{d_{vc}(H_k)\} \leq d_{vc}(\cup_{k=1}^{K} H_k) \leq K - 1 + \sum_{k=1}^{2} d_{vc}(H_k)$