Jacob Snyder

CS156a Set 4

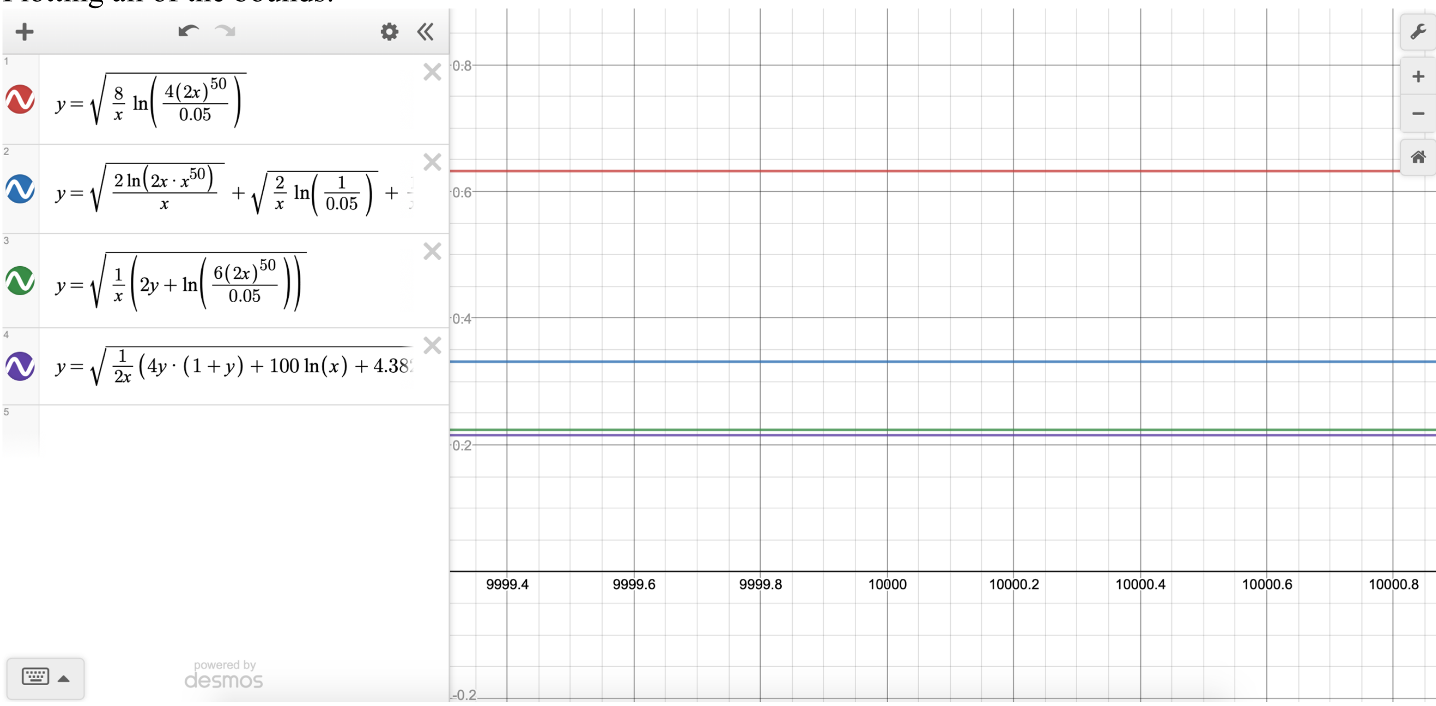
1. D

$$0.05 = \sqrt{\frac{8}{N} \ln \frac{4 * (2N)^{10}}{0.05}}$$

Solving for N, we get $N = 452957$ which is closest to D.

2. D

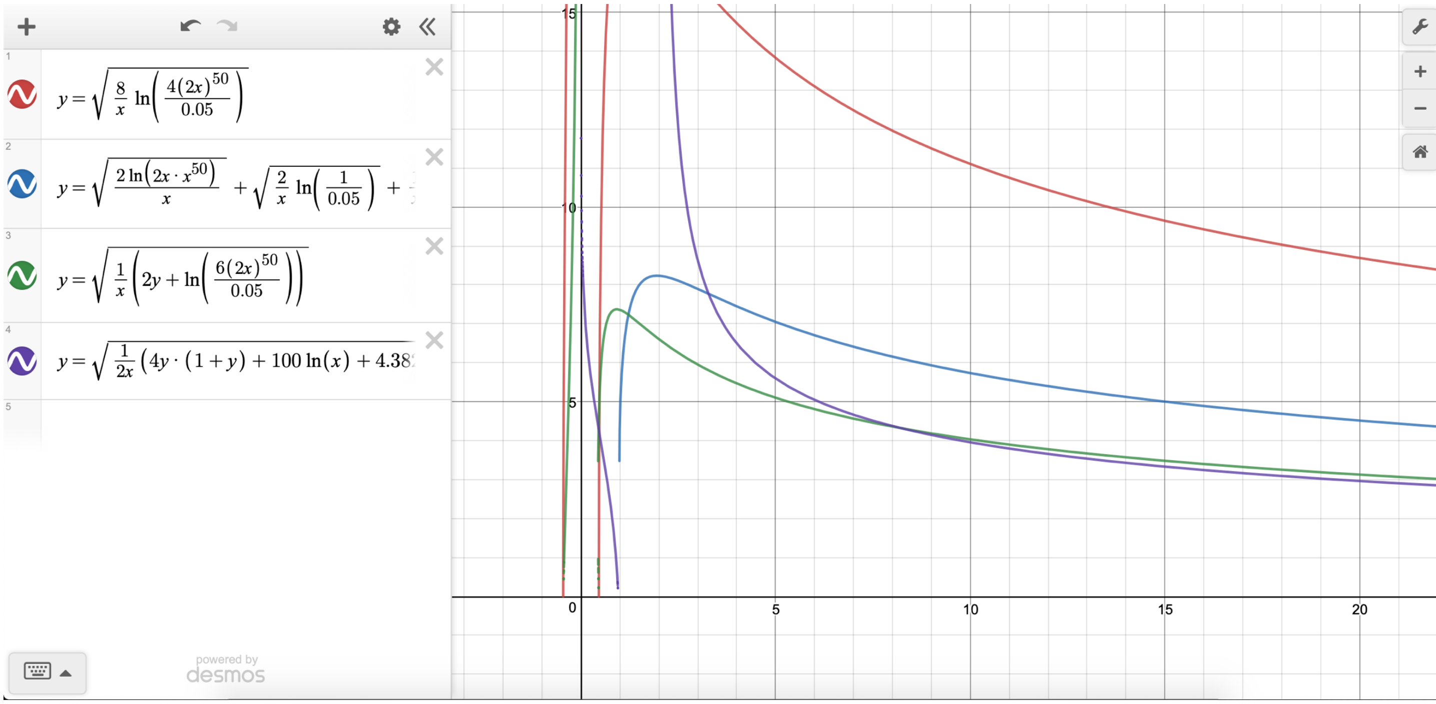
Plotting all of the bounds:



We can see that at $N = 10000$, the smallest bound is the Devroye bound. So the answer is D.

3. C

Plotting all of the bounds:



At $N = 5$ we can see that the Parrondo and Van den Broek bounds are the lowest, so the answer is C.

4. E
We start with the error function for 2 points:
$$E = \frac{1}{2}\sum_{i=1}^{2}(\sin(\pi x_i) - a x_i)^2$$
$$\frac{dE}{da} = -\sum_{i=1}^{2} x_i(\sin(\pi x_i) - a x_i)$$
By setting $\frac{dE}{da} = 0$ we can find the minimum error by varying a:
$$a = \frac{x_1\sin(\pi x_1) + x_2\sin(\pi x_2)}{x_1^2 + x_2^2}$$
Now we plug in many randomly generated x_1 and x_2 to find an average value for a (done in attached Jupyter notebook).
We get $a \approx 1.425$, which matches none of A-D. Therefore the answer is E.

5. B
We can get the bias by taking the integral from -1 to 1 of the bias function:
$$\int_{-1}^{1}(1.425x - \sin(\pi x))^2 dx$$
Using Wolfram Alpha, this evaluates to 0.5394. Taking the average, we have
$$\frac{0.5394}{2} = 0.2697$$
This is closest to 0.3, so the answer is B.

6. A
We will calculate variance by generating many hypotheses. The code is attached in the Jupyter Notebook.
We get 0.237 for variance. This is closest to 0.2 so the answer is A.

7. B

We know the bias and variance for option B: expected out-of-sample error = 0.237 + 0.2697 = 0.5067. We compute the bias for option A (using the same method as in problem 5) as 0.50 and the variance as 0.26 (work in Jupyter Notebook) 0.5+0.26 = 0.76 > 0.5067. Therefore A cannot have the lowest expected out-of-sample error out of the options given. Following the same logic for parts C and E, adding a b term will drastically increase the bias. This means that our only possible viable options are B or D. For D, computing the bias we have 0.50 and variance of 0.081 (work in Jupyter Notebook). 0.5 + 0.081 = 0.581 > 0.5067. Therefore the lowest expected out of sample error is $h(x) = ax$, so the answer is B.

8. C

   Rewrite the growth function as follows:

$$m_{\mathcal{H}}(N) = 2m_{\mathcal{H}}(N-1) - \binom{N-1}{q}$$

   For any $N \leq q$, $\binom{N-1}{q} = 0$ so $m_{\mathcal{H}}(N) = 2m_{\mathcal{H}}(N-1) = 2^N$

   But when $N = q + 1$ we have $\binom{N+1}{q} = \binom{q}{q} = 1$ therefore $m_{\mathcal{H}}(N) = 2m_{\mathcal{H}}(N-1) - 1 = 2^N - 1$.

   Therefore the largest value of N for which $m_{\mathcal{H}}(N) = 2^N$ is $N = q$. Therefore the answer is C.

9. B

   The worst case is where none of the hypothesis sets have anything in common and the intersection of the sets yields an empty set. This would give a VC dimension of 0, so the lower bound must be zero. The upper bound must be the minimum VC dimension of all the sets: suppose we found an group of hypothesis sets that when intersected yielded a VC dimension greater than the minimum of VC dimensions of the hypothesis sets, call it h. Then, by definition of intersection, all hypothesis sets used in the intersection function must also have a VC dimension of at least h (because it must contain at least the same contents as the intersection of all of the sets). However, this is a contradicition, because if every set has a VC dimension of at least h, then the minimum VC dimension cannot be below h. Therefore the minimum VC dimension must be h and therefore the upper bound is the minimum VC dimension of the sets. Therefore the answer is B.

10. E

   When taking the union of our hypothesis sets, we are guaranteed to have d_{vc} of at least the largest d_{vc} from the component hypothesis sets, because they'll all be included in the resulting union. Therefore the answer must be D or E. We can argue using a basic example that it is possible for a union of multiple hypothesis sets to have a larger d_{vc} than the sum of the d_{vc} of the component hypothesis sets:

   Let $\mathcal{H}_1 = \{0\ 0\ 0, 0\ 0\ 1, 0\ 1\ 0, 1\ 0\ 0\}$ where the space is 3 coordinates in space that can be set to either 1 or 0. $\mathcal{H}_1$ covers all possible outcomes of randomly placing a point with a random value. Therefore d_{vc} of this set is 1.

   Let $\mathcal{H}_2 = \{1\ 1\ 1, 1\ 1\ 0, 1\ 0\ 1, 0\ 1\ 1\}$ where the space is 3 coordinates in space that can be set to either 1 or 0. $\mathcal{H}_2$ covers all possible outcomes of randomly placing a point with a random value. Therefore d_{vc} of this set is 1.

   $\mathcal{H}_1 \cup \mathcal{H}_2$ gives us the set $\{0\ 0\ 0, 0\ 0\ 1, 0\ 1\ 0, 0\ 1\ 1, 1\ 0\ 0, 1\ 0\ 1, 1\ 1\ 0, 1\ 1\ 1\}$. Note that this set covers all possible outcomes of randomly placing 3 points with random values. Therefore d_{vc} of this set is 3.

   Therefore d_{vc} of the union of the sets is greater than the sum of the d_{vc} of the sets. Therefore D cannot be the answer, therefore E is the correct answer.