Theory of Machine Learning: Homework 2

1. The paper's title implies a pretty sweeping claim: "Uniform convergence may be unable to explain generalization in deep learning." From what I could gather their theoretical results cover examples of linear models and their empirical results use a 2-layer ReLU network in their setup. The theoretical results seem to prove that UC can fail to explain generalization on a linear model. These results are used to imply failure of UC in their tests with the ReLU network. The applicability of the linear example to the ReLU network was not entirely clear to me, though I am not fluent in these concepts. Roughly, it seemed like the linear argument was based in part on the existence of some bad datasets, and the ReLU network test uses some purposefully constructed bad datasets to show failure of UC, presumably by the same logic of the linear example. A remark is also made about how recent work has shown that deep networks converge to linear models as the network width goes to infinity. I guess the connection between these two results, theoretical and empirical, did not seem entirely rigorous and seemed to rest partly on intuition but the reviewers did not seem to complain. I think these results provably show failure of UC in two particular instances with very specific setups, which probably calls UC into question for similar examples. But I am not sure if it is reasonable to say that UC is unable to explain generalization in deep learning in general, given the wide variety of deep learning architectures. I am a little doubtful that a 2-layer ReLU network is representative of most deep learning models.

   Figure 1 seemed very compelling, especially the comparison between decreasing test error and increasing generalization bounds. Those generalization bounds, from what I could understand, seemed to depend on the norms of the weight matrices. It seems like these figures cast doubt over uniform convergence's ability to explain generalization in deep learning, but the authors do admit that UC may work if gradient descent is run with explicit regularization. Within this context (and some leftover knowledge from previous classes) I understand regularization to entail penalizing large weights. Despite this disclaimer, I found this detail a little strange: "a regularized setting however, is not the main focus of the generalization puzzle." Regularization seems to solve some of the problems presented but is dismissed without much acknowledgement besides this (to be fair, I did not look at the references linked with this statement). The paper mentions implicit regularization due to GD not exploring the full parameter space, but admits that this phenomenon is not fully understood. So I think I agree with some of the comments made by Reviewer 3 in the linked NeurIPS reviews. The reviewer remarks that, with regularization, there are different deep learning algorithms with just as good generalization for which UC may still hold. So the paper demonstrates their claim on what seems to be a small class of examples. Obviously, it is still concerning that existing bounds grow with m. I thought this was a pretty good point Reviewer 3 made. Some of the other reviews do not appear to be as productive. Reviewer 2's response of two sentences says that the paper is written well but that the authors' claim is incorrect, without providing much else in terms of elaboration.

2. For the following examples we will use the Lebesgue measure λ. We may construct any sequence of measurable sets $A_i \in \mathbb{R}$. First, consider

$$A_i = (i, i + 0.5)$$

   All A_i are Borel sets which are measurable w.r.t. the Lebesgue measure. Clearly, $\lambda(A_i) = 0.5 \, \forall i$. We expect that $\lambda\left(\bigcup_i^\infty A_i\right) = \sum_i^\infty (0.5) = \infty$. But it is also the case that $\lim_{i\to\infty} \lambda(A_i) = 0.5$. So we see that $\lambda\left(\bigcup_i^\infty A_i\right) \neq \lim_{i\to\infty} \lambda(A_i)$.

   For intersections, we can use a fairly similar example

$$A_i = \begin{cases} (0,1) & i \text{ even} \\ (1,2) & i \text{ odd} \end{cases}$$

   Again, all A_i are Borel sets and therefore measureable w.r.t. the Lebesgue measure. Since $(0,1) \cap (1,2) = \emptyset$, we expect that $\lambda\left(\bigcap_i^\infty A_i\right) = \lambda(\emptyset) = 0$. But for any i, we see that $\lambda(A_i) = 1$. So $\lim_{i\to\infty} \lambda(A_i) = 1$. So $\lambda\left(\bigcap_i^\infty A_i\right) \neq \lim_{i\to\infty} \lambda(A_i)$.

   Without an assumption that the A_i are nested or finite there are no limit properties like these.