Can uniform convergence bounds work given best-possible data assumptions?

Ezra Winston
12/17/22

1 Introduction

We know empirically that overparameterized DNNs trained via SGD generalize well. Uniform convergence (UC) tries to explain generalization by bounding the worst case generalization gap over a set of hypotheses $\mathcal{H}$, based on some complexity measure of $\mathcal{H}$. Given $0 \leq \delta \leq 1$, the basic form of the UC bound is the smallest ϵ such that

$$\mathbb{P}_{S\sim\mathcal{D}^n}\left(\sup_{h\in\mathcal{H}} |\hat{L}_S(h) - L(h)| \leq \epsilon\right) \geq 1-\delta,$$

where the S are samples of size n from the data distribution $\mathcal{D}$, $\hat{L}_S(h)$ is the empirical loss of an hypothesis h on sample S, and $L(h)$ is the population loss.

Due to the overparameterization of DNNs used in practice, these bounds are typically vacuous when applied to the class of DNNs directly, and so a common approach is to try to restrict the class $\mathcal{H}$, e.g. by restricting the weight norm. However, [NK19] (roughly) shows this approach won't work in general, by constructing a data distributions $\mathcal{D}$ where UC bounds will fail even when $\mathcal{H} = \{$only those NNs learned by SGD on $\mathcal{D}\}$. In addition, it has been observed that CNNs can be trained to interpolate random labels on CIFAR-10 [ZBH+21], meaning that UC bounds also can't hold when $\mathcal{H} = \{$CNNs learned by SGD on CIFAR-10 with corrupted loss$\}$. [1]

A natural question is whether there are assumptions on data distributions under which UC bounds will succeed, and whether such assumptions hold for the datasets we encounter in practice. It is unclear if the “adversarial spheres" of [NK19] are indicative of a phenomena which occurs in real data. [BMDH21], for example, argues that the failure may not occur in real datasets which have angular (rather than only radial) structure.

2 Proposed approach

In this project we attempt to assess whether UC bounds could even work given the *best-possible data assumptions*: would UC bounds hold if $\mathcal{D}$ was a real data distribution, say MNIST? We don't have access to the “true" MNIST population distribution but we can approximating it by combining MNIST train and test sets to make $\hat{\mathcal{D}}$ and resampling i.i.d. training sets $S \sim \hat{\mathcal{D}}^n$.

Denote by h_S the hypothesis (DNN) learned by SGD on sample S. From now on, we consider accuracy in place of loss and denote the accuracy of h on sample S by $Acc(h, S)$, and the population accuracy of h by $Acc(h)$. We'd like to answer the following:

[1] Actually this shows that UC bounds will fail when the loss L in the bound is itself the corrupted loss. We also expect, though have not verified empirically, that models could be trained to interpolate any training set S with the *correct* labels, while simultaneously optimizing incorrect labels on examples outside of S, ensuring poor generalization error w.r.t. the original loss.

Preprint. Under review.

> *Q: For most samples S, is there a different sample S' on which SGD learns hypothesis $h_{S'}$, such that $h_{S'}$ generalizes well (i.e. $Acc(h_{S'})$ is high) but $Acc(h_{S'}, S)$ is low?*[2]

If this were true then this would imply that[3]

$$\mathbb{P}_{S\sim\mathcal{D}^n}\left(\sup_{h\in\mathcal{H}}|Acc(h,S)-Acc(h)|\leq\epsilon\right)\leq\mathbb{P}_{S\sim\mathcal{D}^n}\left(Acc(h,S)-Acc(h)\leq\epsilon\right)\ll 1.$$

Finding such an $h_{S'}$ for arbitrary S would be challenging, especially noting that we wish to only consider $h \in \mathcal{H}$ which are trained to minimize the natural loss function. (Otherwise, we could explicitly maximizing the loss on S'). Instead, we will try to assess the existence of such $h_{S'}$ indirectly.

Toy models of classifier performance The the empirical MNIST distribution $\hat{\mathcal{D}}$ is the discrete uniform distribution over the 70k train and test images. Fix a network architecture and training hyperparameters. There is an induced distribution over classifiers h, obtained by training classifiers on random samples $S \sim \hat{\mathcal{D}}^n$ (in our experiments $n = 200$), with additional randomness coming from the initialization and training order.

We will be concerned with the whether each of the 70k examples are classified correctly or incorrectly by a given classifier. For each example x_j we associate a Bernoulli rv c_j indicating if it is classified correctly. The distribution of each c_j is induced by the distribution over trained classifiers. Training a classifier and evaluating it on the 70k x_j results in a sample of all the c_j. Denote by p_j the probability of x_j being classified incorrectly, that is $p_j := \mathbb{P}(c_j = 0)$, which we refer to as the *difficulty* of x_j.

Let's consider two "extreme-case" mental models of how the c_j could be distributed:

1a) The c_j are mutually independent. That is, the probability that a classifier minclassifies an entire sample set S is

$$\mathbb{P}\left(\bigcap_{x_j\in S}c_j=0\right)=\prod_{x_j\in S}\mathbb{P}\left(c_j=0\right)=\prod_{x_j\in S}p_j.$$

2a) The c_j are ordered based on difficulty, so that $c_j \Rightarrow c_k$ for all k such that $p_j \geq p_k$. So a classifier always correctly classifies only the m least difficult examples for some $m \leq$ 70k.

We also describe two (of many possible) "less extreme" version of the above:

1b) The c_j are mutually independent except for some set of $\ll$ 70k examples which are perfectly correlated, aka equal.

2b) Each classifier falls into one of k bins, and each bin assigns a different sequence of difficulties $p_{k_1} \leq p_{k_2} \ldots$ to the examples, from which the classifier correctly classifies the m easiest.

Implications To show that UC bound will fail, it would suffice that both a) the set $\mathcal{H}$ of h generated as above all have high generalization accuracy, and b) for all (or most) $S \sim \mathcal{D}^n$, $\mathcal{H}$ contains an h with $Acc(h, S) = 0$. Another way to say condition b) is

$$\mathbb{P}_{S\sim\mathcal{D}^n}\Big(\mathbb{P}_{h\sim\mathcal{H}}\big[Acc(h,S)=0\big]>0\Big)\ll 1,$$

where $h \sim \mathcal{H}$ means h distributed according to the induced distribution.

Assume model 1a, and additionally assume that $p_j > 0$ for all j, that is, every example is misclassified by some classifier in $\mathcal{H}$. Then for all S

$$\mathbb{P}_{h\sim\mathcal{H}}\big[Acc(h,S)=0\big]=\prod_{x_j\in S}p_j>0,$$

[2] In future work, it would be good to understand the feasibility of the other direction, where $Acc(h_{S'})$ is low but $Acc(h_{S'}, S)$ is high.

[3] Note that this is not technically a failure of the *tightest possible* UC bound in terms of restriction on $\mathcal{H}$, since that requires showing that for any set of samples $\mathcal{S}_\delta$ with $\mathbb{P}_{S\sim\mathcal{D}^n}(\mathcal{S}_\delta) \geq 1-\delta$, the there is a bad hypotheses $h \in \mathcal{H}_\delta = \{$*only samples $S \in \mathcal{S}_\delta$ itself*$\}$. We defer this to future work.

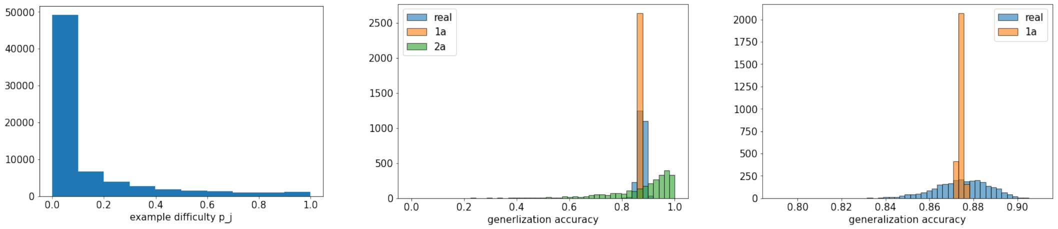
so UC bounds must fail. On the other hand, assuming model 2a, then for any S such that $x_j \in S$ and $x_k \notin S$ and $p_j \geq p_k$, there is no $h \in \mathcal{H}$ with $Acc(h, S) = 0$, so a UC bound could succeed.

3 Empirical analysis

In order to assess the extent to which the above models comport with reality, we conduct the following experiment: we sample 2641 training sets of size $n = 200$ from the 70k combined MNIST train and test examples, and on each we train a 1.2M parameter CNN until reaching perfect train accuracy.[4] Despite the tiny training set size, the models obtain nontrivial generalization accuracy of around 87% measured on the full 70k examples.

Basic characteristics It is convenient to picture this data as a 2461×70k table T with each row i corresponding to a classifier h_i and each column j corresponding to an example x_j, and $T_{ij} = 1$ if h_i classifies x_j correctly, 0 else. Then the mean of each row i is the accuracy of classifier h_i, and the mean of each column j is an estimate of $1 - p_j$, the “easiness" of example x_j. Call this table T^{real}. Using the estimated p_j we also generate synthetic versions T^{1a} and T^{2a}, corresponding to the two toy models above. Each row i of T^{1a} is generated by sampling each T^{1a}_{ij} independently with probability $1 - p_j$, and each row i of T^{2a} is generated by first sampling $q_i \sim \text{Unif}(0, 1)$ and then setting $T^{1a}_{ij} = 1$ whenever $p_j < q_i$.

By design, all three table have the same mean classifier performance, i.e. the same *mean row mean* (which is equal to the mean column mean), of ∼87.4%. Also by design, the difficulty p_j of each example (1 - column means) are the same. The distribution of example difficulty is is shown in figure (1a). By contrast, the distribution of the classifier performance, shown in figures (1b) and (1c), is very different in each case. Neither model fits the real data well: the real and model-1a histograms are approximately symmetric but model-1a has much lower variance, while model 2a is qualitatively very different.



(a) histogram of example difficulty (same for real data and models)

(b) histograms of classifier generalization accuracy

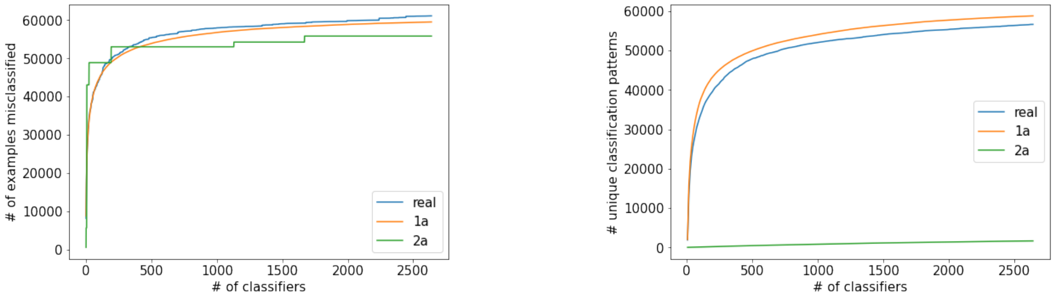
(c) same as (b), showing only real data and model 1a

Figure 1: Marginal distributions of the classifier data (blue: real data; orange: synthetic model 1a; green: synthetic model 2a)

A prerequisite for model 1a is that every example is misclassifed by *some* h. We don’t see this in our sample of classifiers: 8139 examples are always classified correctly. But in figure (2a) we plot how the total number of examples misclassified by some h grows as our set of h grows, which gives some indication that this fraction will approach one. Figure (2a) also plots the same growth for the synthetic data, which is notably less smooth for model 2a. Figure (2b) shows the growth of the number of unique “classification patterns" (i.e. unique columns of T) as we consider more classifiers. Equal columns indicates perfect correlation of two of the c_j, which would break the assumptions of model 1a. The number of unique classification patterns grows much more slowly under model 2a than for the others.

Some statistics While we don’t observer perfect independence of the c_j, we do see that the typical pairwise correlation ρ is very small for real data and under model 1a, but not under model 2a. Figures

[4] Details: Adadelta for 30 epochs; lr=1.0; batch size=64. These hyperparameters should be explored in future work. Beyond the random training set, there is additional randomness over the weight initialization and training batch order. Also, for computational expediency, the models are not trained to zero training loss, which should be explored further.

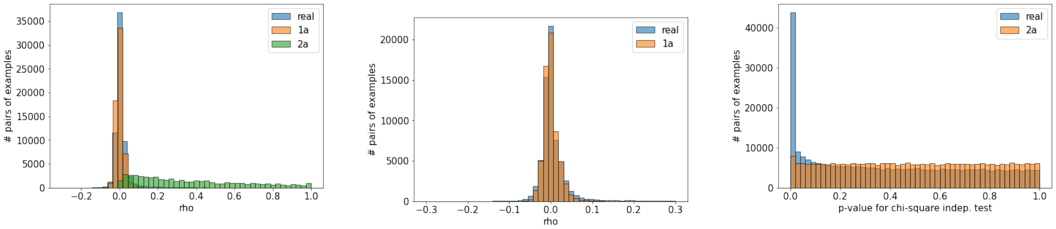


(a) total # of examples misclassified by at least one h as the number of h increases

(b) # of unique classification patterns as the number of h increases

Figure 2: As we consider more classifiers, the fraction of misclassified examples and examples with unique classification patterns may be approaching 1 for real classifiers and synthetic model 1a, with notably different behaviour for model 2a.

(3a) and (3b) show the distribution of correlation coefficients between all pairs of (c_i, c_j) for a sample of 1000 examples, for real classifiers and both models 1a and 2a. Figure (3c) shows the p-values for standard chi-square tests between pairs of examples.[5]



(a) histogram of correlation coefficients ρ between the c_j for pairs of examples

(b) same as (a) but omitting model 2a

(c) histogram of p-values for chi-square independence tests between pairs of c_j

Figure 3: Pairwise correlation and independence of the c_j

The chi-square test rejects the null hypothesis that c_i and c_j are uncorrelated when $p < 0.05$. Under model 1a there is no true dependence between c_i and c_j, and as expected, the p-values of the test for model 1a are roughly uniform; those p-values less than 0.05 represent false rejections of the null (though there are very slightly more than expected test with $p < 0.05$, which is a mystery). The p-values of the real classifiers are distributed roughly uniformly except for a subset which are near zero, indicating statistically-significant dependence. On closer inspection, we find that a large portion of the dependent pairs have the same class label, a fact which warrants further investigation.

Aside: While the chi-square test is standard, it only rejects or fails to reject the null hypothesis that the pair are independent. We would instead like to reject a null hypotheses of *dependence*. An approach for this is to use an equivalence test in the TOST framework, where the null hypothesis to be rejected is that the pair is associated with at least some strength. An approach to this using McNemar's test is described in [2]. Further investigation is required, but preliminary computations of this test indicate that the large majority of the associations in the real data, even when statistically significant, are weak. Bounding the strength of the associations could be useful because UC bound failure seems more likely if the associations are weak.

Estimating a bound under model 1b Recall that if model 1a were true, meaning the c_j are mutually independent, then every S has a classifier with $Acc(h, S) = 0$, as argued in Section 2, and so the UC bound will fail. If we instead assume, under model 1b, that the c_j are independent for

[5] This test can't be performed for model 2a since a rule of thumb is to require at least 5 examples in each quadrant of the (c_i, c_j) contingency tables; example pairs for real data and model 1a are filtered for this criteria, the effect of which requires more careful consideration in future work.

all but a small set B of the x_j, then we can show bound failure by assuming that all $x_j \in B$ are always correctly classified. Observe from figure (1c) that virtually all of $h \in \mathcal{H}$ obtain at least 0.82 generalization accuracy. We can still have trivial accuracy ≤ 0.1 on most S, meaning a generalization gap of at least 0.72 with probability $\gg 0$:

$$\mathbb{P}_{S \sim \mathcal{D}^n}\left(\sup_{h \in \mathcal{H}} |Acc(h,S) - Acc(h)| \geq 0.72\right) \geq \mathbb{P}_{S \sim \mathcal{D}^n}\left(\exists h \in \mathcal{H} | Acc(h,S) \leq 0.1\right)$$
$$\approx q(|B|)$$

where $q(|B|)$ depends on the size of B. We can compute q under the empirical distribution directly, because the distribution over sample sets will be uniform and discrete. We simply compute the number of ways of choosing a sample S with $|S| = 200$ which has ≤20 examples in B, as a fraction of the total number of ways of choosing a samples. Figure (4) shows that $q(|B|)$ remains near 1 for $|B| \leq 4000$.

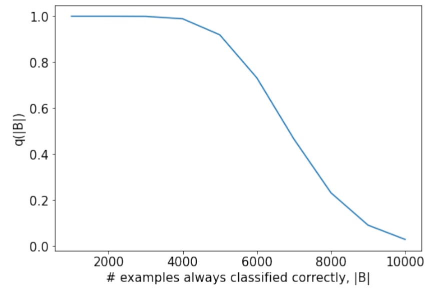


Figure 4: The probability $q(|B|)$ of a generalization failure under the UC bound estimate for model 1b, given easy example set with size $|B|$

A union-bound style approach The issue with assuming model 1a/b is that, while we can empirically verify approximate pairwise independence of the c_j, it is much harder to verify the *mutual independence* of sets of 200 examples, which is required by the model. It is unclear if there is a way to argue that the effect of mutual dependence beyond pairwise is minimal.

A potential approach to show bound failure without assuming mutual independence is the Hunter-Worsley bound [Hun76, Wor82]

$$\mathbb{P}\left(\cup_i A_i\right) \leq \sum_{i=1}^{n} p_i - \max_{\tau \in T} \sum_{(i,j) \in \tau} p_{ij}$$

which bounds the probability of the union of events A_i based on the marginal probabilities p_i and pairwise probabilities p_{ij}. The bound optimizes $\tau \in T$, the maximum spanning tree from the set T of spanning trees on the graph with nodes A_i and edge weights p_{ij}.

In our setting, the A_i are taken to be the c_j, the event that example x_j is classified correctly. Given sample S, we can estimate an upper bound on the probability that any of the c_j are classified correctly, which gives a lower bound on the event that none are. Unfortunately, we see that for most S the bound is slightly greater than 1 in our data. Further investigation is warranted, however, since this seems to be related to the existence of very easy examples: if we restrict to sets S with only difficulty of at least 0.5, we find that bound s < 1 for close to half of the S.

4 Discussion and next steps

Overall, our analysis of the misclassification data is inconclusive wrt to the question of UC bound viability. Mutually-independent misclassification, as in models 1a/b, seems difficult or impossible to verify empirically. Even pairwise independence should not be expected exactly in real data. And our existence arguments for UC bound failure depend on exact independence.

However, there is some indication that UC bound failure is likely. The distribution of misclassifications which we observe does have a notable absence of strong correlation. Roughly, we see slight but detectable pairwise correlation between classification of examples of the same class, and very weak correlation otherwise. The role of the the class label seems important to investigate further.

Our intuition is that when correlations between example pairs are more frequent and stronger, UC bounds will be more likely to succeed. But this is not guaranteed: it is conceivable that there could be strong but imperfect correlation between all pairs, yet still a nonzero probability of every sample having a misclassifier. The exact role of pairwise correlation strength should be explored. In particular, is there a relationship between the correlation strength and the Hunter-Worsley bound? This may reveal settings in which the Hunter-Worsley bound is non-vacuous.

References

[BMDH21] Gregor Bachmann, Seyed-Mohsen Moosavi-Dezfooli, and Thomas Hofmann. Uniform convergence, adversarial spheres and a simple remedy. In *International Conference on Machine Learning*, pages 490–499. PMLR, 2021.

[2] Alexis (https://stats.stackexchange.com/users/44269/alexis). Comparison of statistical tests exploring co-dependence of two binary variables. Cross Validated. URL:https://stats.stackexchange.com/q/447682 (version: 2020-02-08).

[Hun76] David Hunter. An upper bound for the probability of a union. *Journal of Applied Probability*, 13(3):597–603, 1976.

[NK19] Vaishnavh Nagarajan and J Zico Kolter. Uniform convergence may be unable to explain generalization in deep learning. *Advances in Neural Information Processing Systems*, 32, 2019.

[Wor82] KJ Worsley. An improved bonferroni inequality and applications. *Biometrika*, 69(2):297–302, 1982.

[ZBH$^+$21] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.