Theoretical principles of deep learning
Final Report

Review of a scientific paper

Uniform Convergence May be Unable to Explain Generalization in Deep Learning by Nagarajan and Kolter (2019)

Marie Generali & Lucas Biechy
Professor: Hédi Hadiji

Institut Mathématique d'Orsay (IMO)
Paris-Saclay University

June 4, 2024

Abstract

Understanding generalization is one of the fundamental unsolved problems in deep learning. Why does a model performing well on training data is able to perform almost as well on data it has never seen?
In order to explain this generalization, a lot of work is looking in the direction of uniform convergence. This article nuances the enthusiasm toward this research by showing both theoretically and experimentally that in the case of linear models and by extension of current networks, uniform convergence can fail at explaining generalization for overparametrized networks.
Our report is a summary of this article. After briefly outlining all the contributions of this paper, we will develop the mathematical theorem linking uniform convergence and generalization bounds for linear models. Additionally, we will conduct our own implementation of this theorem.

Keywords: Deep Learning · Uniform Convergence · Generalization

Contents

Glossary

S	dataset of m points.
hypothesis output on S	h_S .
smoothed binary loss	$$\mathcal{L}^{(\gamma)}(y, y') = \begin{cases} 1 & \text{if } yy' \leq 0 \\ 1 - \frac{yy'}{\gamma} & \text{if } 0 < yy' < \gamma \\ 0 & \text{if } yy' \geq \gamma \end{cases}$$.
expectation over S	$\mathcal{L}_D(h_S) := \mathbb{E}_{(\mathbf{x},y)\sim D}[\mathcal{L}(h(x), y)]$.
empircal expectation over S	$\mathcal{L}_S\hat{(}h_S) := \frac{1}{m}\sum_{(\mathbf{x},y)\in S} \mathcal{L}(h(x), y)$.

1 Introduction

Due to the bias-variance decomposition of the generalization error, it is surprising that overparameterized neural networks manage to generalize effectively. This observation has sparked renewed interest in current research, leading to efforts to reform theoretical learning approaches by identifying and incorporating the implicit bias/regularization of stochastic gradient descent (SGD). Subsequently, generalization bounds for deep networks have been developed, primarily relying on uniform convergence. This allows for bounds on the generalization error that are both (a) small and non-vacuous (i.e., < 1), (b) inversely proportional to the width/depth of the network, (c) applicable to the network learned by SGD (without modification or explicit regularization), and (d) increasing with the number of randomly flipped labels used during training.

While many bounds satisfy some of these criteria, none are known to satisfy all simultaneously. This paper highlights another fundamental shortcoming of existing bounds by demonstrating that they violate a natural but widely overlooked criterion for explaining generalization: (e) they do not decrease with the dataset size at the same rate as the generalization error. This paper empirically shows that these bounds increase with dataset size. Consequently, it questions the efficacy of uniform convergence in explaining generalization and even proves its intrinsic limitations in the overparameterized regime.

2 Contribution of the paper

2.1 Theory

First Contribution They showed that in practice, the weighted norm of deep ReLU networks (such as the distance at initialization) grows polynomially with the number of training data m. Consequently, the generalization bounds that depend on these norms do not adequately capture this dependence on m relative to the test error, violating criterion (e). For small lot sizes, these bounds even increase with the number of examples. Thus, there is a significant conceptual gap in our understanding of networks, particularly due to this void, which is independent of the number of parameters.

Second Contribution The authors examined three examples of overparameterized models trained with stochastic gradient descent: a linear classifier, a very wide ReLU neural network, and finally an infinitely wide network with exponential activations (with frozen hidden layer weights). These settings also mimic the observation that norms such as distance from initialization increase with dataset size m. More importantly, under these conditions, any bilateral uniform convergence limit would yield a (nearly) vacuous generalization bound.

We will focus on the central configuration of this paper, which is the linear classifier. The authors develop a theorem which we will elaborate on throughout our report below. The other two sections on neural networks are merely applications to explore the impact of the theorem in modern use cases. The authors manage so to identify a possible inconsistency between the explanation of generalization and the bound of uniform convergence in current networks.

To demonstrate the theorem, which is the main contribution of the paper, the author introduces for the dataset S the smoothed binary loss, its expectation over S, its empircal expectation over S and h_S the hypothesis output on S (cf glossary).

The generalization error is defined as,

$$\epsilon_{gen}(m,\delta) := \underset{\epsilon\in\mathbb{R}}{\operatorname{argmin}} \left\{ \mathbb{P}_{S\sim D^m} \left(\mathcal{L}_D(h_S) - \hat{\mathcal{L}}_S(h_S) \leq \epsilon \right) \geq 1-\delta \right\}$$

Then the article defines the "tightest algorithm-dependent uniform convergence bound". Such a definition aims to get rid of the hypotheses of that would never be chosen according to the distribution D, in order to show that even in the "most appropriate" case, uniform convergence doesn't explain the generalization (which implies that it wouldn't be able to explain it in less appropriate settings). The point is to find an $\epsilon_{unif-alg}(m,\delta)$ that is very tight and for which a set of sample set, S_δ, exists for which $\epsilon_{unif-alg}(m,\delta)$ is an acceptable bound. According to such a set of sample set, we define the set of hypotheses defined by the algorithm, $\mathcal{A}$ on $\mathcal{S}_\delta$ as $\mathcal{H}_\delta := \cup_{S\in\mathcal{S}_\delta} h_S$.

$$\epsilon_{unif-alg}(m,\delta) := \underset{\epsilon\in\mathbb{R}}{\operatorname{argmin}} \left\{ \mathbb{P}_{S\sim D^m} \left(S \in \mathcal{S}_\delta \right) \geq 1-\delta \right\}$$

$$\sup_{S\in\mathcal{S}_\delta} \sup_{h\in H_\delta} \left| \mathcal{L}_D(h) - \hat{\mathcal{L}}_S(h) \right| \leq \epsilon_{unif-alg}(m,\delta)$$

The proof of the theorem is done in the context of a linear model; indeed, such models are said to give tighter bounds on uniform convergence than more complicated ones-meaning that the proof is developed in the "most appropriate" case for uniform convergence to explain the generalization. In addition, Jacot et al. (2020) showed that asymptotically, under some conditions, all neural wide models tend to a linear one.
We consider the linear model where the input is $\mathbf{x} = (x_1, x_2)$ with $x_1 \in \mathbb{R}^K$ and $x_2 \in \mathbb{R}^D$, where K is a small constant and $D \ll m$ and the weights are w_1 and w_2.

With this setting introduced, the following theorem is demonstrated.

Theorem 2.1. *$\forall\epsilon > 0, \forall\delta \in]0, \frac{1}{4}[$ et $D(\epsilon,\delta)$ big enough, we have $\forall\gamma \geq 0$ for the $\mathcal{L}^{(\gamma)}$ loss, $\epsilon_{unif-alg}(m,\delta) \geq 1 - \epsilon_{gen}(m,\delta)$. And for $\gamma \in]0,1[$, $\epsilon_{gen}(m,\delta) \leq \epsilon$ so $\epsilon_{unif-alg}(m,\delta) \geq 1-\epsilon$*

2.2 Implementation

No specific code was shared with the paper. Therefore, we decided to implement it our way, and you can find our implementation on our GitHub repository. While we did not manage to implement the linear classifier directly as mentioned in the theorem, we believed it was crucial to create a model that closely resembles it. To achieve this, we started with the ReLU network discussed in the attributed section, which already demonstrates a corollary application of the theorem according to the paper, and attempted to modify it to be as close as possible to the linear classifier described in the theorem.

Setup First, we do not have specific dimensions in the data that are noisy and second, the data dimensionality here as such is a constant less than m. We vary the number of training examples from $4k$ to $60k$ using the MNIST dataset. The aim is to classify images in a binary manner by predicting whether the digit shown in the image is even or odd. We train a one linear layer network to minimize cross entropy loss with a batch size of 1 as in the theorem, using SGD with a learning rate 0.1 and with 15 epochs to overparameterize the models enough. We choose the Binary Cross Entropy as loss function and the 0-1 error, i.e. $\mathcal{L}^{(0)}$, as generalization error.

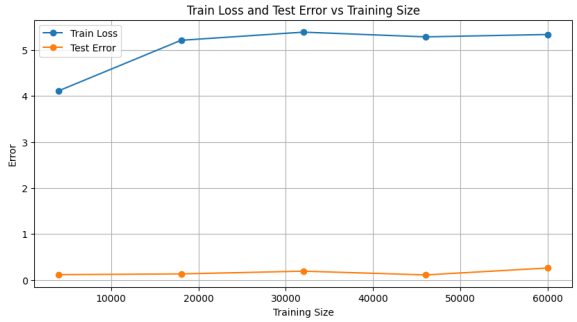

Figure 1: Implementation results

Result The result is not as clear as hoped. We do not observe any trend towards having a generalization and convergence inversely proportional and dependent on the dataset size m, as expected in the paper. Perhaps the model is not over-parameterised enough. Moreover, we do see a clear separation between generalization and loss bound. Therefore, no information from this implementation allows us to refute (or validate) the theorem.

3 Proof

The proof relies on two lemmas:

- **First lemma for upper bounding generalization error**: In Lemma E.1, it is shown that the learned parameters lead to zero training loss. The classifier aligns well with the training data and has low training error. Lemma E.2 establishes a lower bound on uniform convergence by arguing that, with high probability, the classifier misclassifies the training data when the noise vectors are negated.
- **Second lemma for lower bounding uniform convergence**: The proof of Lemma E.2 involves demonstrating the existence of a "bad" dataset $\tilde{S}$ such that the classifier has low test error but high empirical error on $\tilde{S}$. The union bound is used to argue that there is a non-zero probability of picking such a "bad" $\tilde{S}$, leading to a positive probability of failure for uniform convergence.

The theorem's statement involves conditions on the size of the dataset D for the failure of uniform convergence. Specific constants c_1, c_2, c_3, c_4 are defined. The theorem asserts that for certain conditions on D is large enough, then the learning algorithm fails to achieve uniform convergence for the $\mathcal{L}^{(\gamma)}$ loss.

The demonstration is done choosing a specific D verifying the following equations

$$
\begin{aligned}
D &\geq \frac{1}{c_1} \ln \frac{6m}{\delta}, \\
D &\geq m \left(\frac{4c_4 c_3}{c_2^2} \right)^2 \ln \frac{6m}{\delta}, \\
D &\geq m \left(\frac{4c_4 c_3}{c_2^2} \right)^2 \cdot 2 \ln \frac{2}{\epsilon},
\end{aligned}
$$

Lemma 3.1. *When $\gamma \in [0,1]$ for $\mathcal{L}^{(\gamma)}, \epsilon_{gen}(m,\delta) \leq \epsilon$*

Proof. The two corollaries demonstrated in appendix yield the two following equations with probability $1 - \frac{\delta}{3m}$ over the respective draws of $\mathbf{x}_2^{(i)}$ and $\sum_{j\neq i} y^{(j)}\mathbf{x}_2^{(j)}$. According to the first equation verified by D.

$$c_2 \leq \frac{1}{2\sqrt{2}} c_2 \left\|\mathbf{x}_2^{(i)}\right\| \leq c_3.$$

And,

$$\left|\mathbf{x}_2^{(i)} \cdot \sum_{j\neq i} y^{(j)}\mathbf{x}_2^{(j)}\right| \leq c_4 \left\|\mathbf{x}_2^{(i)}\right\| \frac{2\sqrt{2}\cdot\sqrt{m}}{c_2\sqrt{D}} \sqrt{\ln\frac{6m}{\delta}}$$

Then, with probability $1 - \frac{2}{3}\delta$ over the draws of the training dataset we have for all i ,

$$\begin{aligned} y^{(i)} h\left(\mathbf{x}^{(i)}\right) &= y^{(i)}\mathbf{w}_1 \cdot \mathbf{x}_1^{(i)} + y^{(i)} \cdot y^{(i)} \left\|\mathbf{x}_2^{(i)}\right\|^2 + y^{(i)} \cdot \mathbf{x}_2^{(i)} \cdot \sum_{j\neq i} y^{(j)}\mathbf{x}_2^{(j)} \\ &= 4 + \underbrace{\left\|\mathbf{x}_2^{(i)}\right\|^2}_{\text{because of the first equation}} + \underbrace{y^{(i)}\mathbf{x}_2^{(i)} \cdot \sum_{j\neq i} y^{(j)}\mathbf{x}_2^{(j)}}_{\text{because of the second equation}} \\ &\geq 4 + 4\cdot 2 - c_4 \frac{2\sqrt{2}c_3}{c_2} \cdot \underbrace{\frac{2\sqrt{2}\cdot\sqrt{m}}{c_2\sqrt{D}} \sqrt{\ln\frac{6m}{\delta}}}_{\text{from the second equation verified by D}} \\ &\geq 4 + 8 - 2 = 10 > 1. \end{aligned}$$

We deduce from this result that the loss on the training dataset is zero since it classifies well all the samples. We can then deduce from the first corollary with the same probability,

$$c_2\sqrt{m} \leq \frac{1}{2\sqrt{2}} c_2 \left\|\sum y^{(i)}\mathbf{x}_2^{(i)}\right\| \leq c_3\sqrt{m}.$$

Then, we define a draw of test data point (**z**, y) over which, with $\epsilon' > 0$ we have with probability $1 - \epsilon'$ according to the second corollary,

$$\left|\mathbf{z}_2 \cdot \sum y^{(i)}\mathbf{x}_2^{(i)}\right| \leq c_4 \left\|\sum y^{(i)}\mathbf{x}_2^{(i)}\right\| \cdot \frac{2\sqrt{2}}{c_2\sqrt{D}} \cdot \ln\frac{1}{\epsilon'}.$$

Using this, we have that with probability $1 - 2\exp\left(-\frac{1}{2}\left(\frac{c_2^2}{4c_4c_3}\sqrt{\frac{D}{m}}\right)^2\right)$ over the draws of a test data point, $(\mathbf{z}, y)$, we have that for $\gamma \in [0,1]$, the $\mathcal{L}^{(\gamma)}$ loss of the classifier on the distribution $\mathcal{D}$ is

$2\exp\left(-\frac{1}{2}\left(\frac{c_2^2}{4c_4c_3}\sqrt{\frac{D}{m}}\right)^2\right)$ which is at most ϵ, indeed,

$$\begin{aligned} yh(\mathbf{x}) &= y\mathbf{w}_1 \cdot \mathbf{z}_1 + \underbrace{y \cdot \mathbf{z}_2 \cdot \sum_j y^{(j)}\mathbf{x}_2^{(j)}}_{\text{last equation showed}} \\ &\geq 4 - c_4 \underbrace{\left\|\sum^{(i)} \mathbf{x}_2^{(i)}\right\|}_{\text{penultimate result showed}} \cdot \frac{2\sqrt{2}}{c_2\sqrt{D}} \frac{c_2^2}{4c_4c_3}\sqrt{\frac{D}{m}} \\ &\geq 4 - 2 \geq 2. \end{aligned}$$

In other words, the absolute difference between the distribution loss and the train loss is at most ϵ and this holds for at least $1-\delta$ draws of the samples S. Then, by the definition of ϵ_{gen} we have the result.

□

We have showed that we have a bound on the generalization bound. Now, in order to prove our results, we need to show that this bound doesn't induce a bound on the uniform error. Therefor, we demonstrate the second lemma,

Lemma 3.2. *For any $\epsilon > 0$ and for any $\delta \leq 1/4$, and for the same lower bounds on D, and for any $\gamma \geq 0$, we have that*

$$\epsilon_{unif\text{-}alg}\left(m, \delta\right) \geq 1 - \epsilon_{gen}\left(m, \delta\right)$$

for the $\mathcal{L}^{(\gamma)}$ loss.

The idea of the proof of this theorem is to show that for any set from S_δ used for the definition of the tightest bound -the set minimizing the generalization error- we can find a set, S_* following the same distribution -e.g the distribution of D- and belonging to S_δ. The goal is to show that even if for the hypothesis output by the algorithm on the dataset S_* we can control the test error, the empirical test error on a noised dataset is very high.

Proof. After defining the noised-negated sample associated to S,
$S' = \{((\mathbf{x}_1, -\mathbf{x}_2), u) \mid ((\mathbf{x}_1, \mathbf{x}_2), y) \in S\}$. We can show[1] that for any $\mathbf{x}_{\text{neg}}^{(i)} = \left(\mathbf{x}_1^{(i)}, -\mathbf{x}_2^{(i)}\right)$, we have $y^{(i)}h\left(\mathbf{x}_{\text{neg}}^{(i)}\right) < 0$. Which implies that the loss over S' is 1 -the set is completely missclassified.

Now the goal is to show that for any set $S \in S_\delta$ we can find $S_* \in S_\delta$ such that its noised version $S'_* \in S_\delta$, and even though the hypotheses over this set has a controlled test error, it completely misclassifies the noised-negated set. Indeed, considering the prior of the existence of such a set we have,

$$\begin{aligned} &\Pr_{S\sim\mathcal{D}^m}\left[S \in \mathcal{S}_\delta, S' \in \mathcal{S}_\delta, \mathcal{L}_\mathcal{D}\left(h_S\right) \leq \epsilon_{\text{gen}}(m,\delta), \hat{\mathcal{L}}_{S'}\left(h_S\right) = 1\right] \\ &\geq 1 - \Pr_{S\sim\mathcal{D}^m}\left[S \notin \mathcal{S}_\delta\right] - \Pr_{S\sim\mathcal{D}^m}\left[S' \notin \mathcal{S}_\delta\right] \\ &- \Pr_{S\sim\mathcal{D}^m}\left[\mathcal{L}_\mathcal{D}\left(h_S\right) > \epsilon_{\text{gen}}(m,\delta)\right] - \Pr_{S\sim\mathcal{D}^m}\left[\hat{\mathcal{L}}_{S'}\left(h_S\right) \neq 1\right]. \end{aligned}$$

[1]The demostration, very similar to the one for the first lemma, is given in A.1

We have,

- $\Pr_{S\sim\mathcal{D}^m}\left[S \notin \mathcal{S}_\delta\right] \leq \delta$ by definition of S'
- $\Pr_{S\sim\mathcal{D}^m}\left[\mathcal{L}_\mathcal{D}\left(h_S\right) > \epsilon_{\text{gen}}\left(m,\delta\right)\right] \leq \delta$ by definition of the generalization error
- $\Pr_{S\sim\mathcal{D}^m}\left[\hat{\mathcal{L}}_{S'}\left(h_S\right) \neq 1\right] \leq 2\delta/3$ according to 3
- $\Pr_{S\sim\mathcal{D}^m}\left[S' \notin \mathcal{S}_\delta\right] < \delta$ since the noised negated set follows the same distribution as S, by isotropic property of the Gaussian noise.

Hence, we have,

$$\Pr_{S\sim\mathcal{D}^m}\left[S \in \mathcal{S}_\delta, S' \in \mathcal{S}_\delta, \mathcal{L}_\mathcal{D}\left(h_S\right) \leq \epsilon_{\text{gen}}(m,\delta), \hat{\mathcal{L}}_{S'}\left(h_S\right) = 1\right] \geq 1 - 4\delta > 0$$

This yields the existence of S_* for any choice of S. And finally, we have,

$$\begin{aligned}
\epsilon_{\text{unif-alg}}\left(m,\delta\right) &= \sup_{S\in\mathcal{S}_\delta}\sup_{h\in\mathcal{H}_\delta}\left|\mathcal{L}_\mathcal{D}(h) - \hat{\mathcal{L}}_S(h)\right| \\
&\geq \left|\mathcal{L}_\mathcal{D}\left(h_{S_\star}\right) - \hat{\mathcal{L}}_{S'_\star}(h)\right| \\
&= |\epsilon - 1| \\
&= 1 - \epsilon.
\end{aligned}$$

□

Subsequentially of those two lemma, we can show the final theorem,

Proof. The proof is immediate by combining the two previous lemmas. □

4 Discussion

The aim of the article to demonstrate that uniform convergence can fail at explaining generalization is fulfilled: the author managed to construct a case where, even with a uniform convergence bound void, the model did generalize well. This result is interesting considering that uniform convergence is often used to assess the generalization power of a model. However, the hypotheses in which the proof is developed are rather tight: first, the choice of D is quite precise, then, we take a special dataset that ensures the tightest generalization bound before exhibiting a dataset that could be the worst in terms of uniform convergence. Therefore, the impact of this theorem on currently commonly used neural networks appears to be doubtful. That is why, even if the experiments seem to be concluding, the author himself seems to be saying that it is not to be sure that the theoretical result he showed is really explaining the observed phenomenon.

From our perspective, this article left us somewhat unsatisfied for two reasons. Firstly, the absence of provided computer code hinders the verification of everything described and understanding the utilized configurations in depth. Secondly, and most importantly, the article is initially unclear and quickly becomes unnecessarily complex, raising doubts about the validity of the results and the relevance of the findings. For future work, it would be beneficial to rephrase this article in a clearer manner, potentially uncovering theorems applicable to a broader range of cases.

References

Jacot, A., Gabriel, F., and Hongler, C. (2020). Neural tangent kernel: Convergence and generalization in neural networks.

Nagarajan, V. and Kolter, J. Z. (2019). Uniform convergence may be unable to explain generalization in deep learning. *Advances in Neural Information Processing Systems*, 32.

A Sub-lemmas for the demonstration of the second lemma

Lemma A.1. *According to the definitions introduced for the demonstration of the lemma, we have, with high probability* $1 - 2\delta/3$ *over the draws of* S*,*

$$y^{(i)} h\left(\mathbf{x}_{neg}^{(i)}\right) < 0$$

Proof.

$$\begin{aligned}
y^{(i)} h\left(\mathbf{x}_{\text{neg}}^{(i)}\right) &= y^{(i)} \mathbf{w}_1 \cdot \mathbf{x}_1^{(i)} - y^{(i)} \cdot y^{(i)} \left\|\mathbf{x}_2^{(i)}\right\|^2 - y^{(i)} \cdot \mathbf{x}_2^{(i)} \cdot \sum_{j \neq i} y^{(j)} \mathbf{x}_2^{(j)} \\
&= 4 - \underbrace{\left\|\mathbf{x}_2^{(i)}\right\|^2}_{\text{because } c_2 \leq \frac{1}{2\sqrt{2}} c_2 \left\|\mathbf{x}_2^{(i)}\right\| \leq c_3} - y^{(i)} \mathbf{x}_2^{(i)} \cdot \sum_{j \neq i} y^{(j)} \mathbf{x}_2^{(j)} \\
&\leq 4 - 4 \cdot 2 + c_4 \frac{2\sqrt{2} c_3}{c_2} \cdot \underbrace{\frac{2\sqrt{2} \cdot \sqrt{m}}{c_2 \sqrt{D}} \ln \frac{3m}{\delta}}_{\text{because } D \geq m \left(\frac{4 c_4 c_3}{c_2^2}\right)^2 \ln \frac{6m}{\delta}} \\
&\leq 4 - 8 + 2 = -2 < 0.
\end{aligned}$$

□