UNIFORM CONVERGENCE MAY BE UNABLE TO EXPLAIN GENERALIZATION IN DEEP LEARNING

VAISHNAVH NAGARAJAN[1]   ZICO KOLTER[1,2]

[1]Computer Science Department, Carnegie Mellon University   [2]Bosch Center for Artificial Intelligence, Pittsburgh

THE HIGH LEVEL MESSAGE

We study the big open question in deep learning theory: "Why do overparameterized networks generalize well even though standard intuition suggests otherwise?"

To decode this puzzle, many generalization bounds have been proposed – all based on the learning-theoretic tool of uniform convergence.

We argue that this high-level direction (of deriving uniform convergence-based bounds in deep learning) may not provide the complete answer to the generalization puzzle.

THE GENERALIZATION PUZZLE & UNIFORM CONVERGENCE (U.C)

Conventional u.c. bounds (like VC dim) fail to explain generalization in deep learning [1,2]:

$$\text{generalization gap} \leq O\left(\sqrt{\frac{\text{representational complexity of whole hypotheses class}}{\text{training set size}}}\right)$$

← depth×width ⟹ vacuous!

For tighter, more meaningful bounds, the proposed suggestion was to identify implicit bias and use it to refine u.c. bounds:

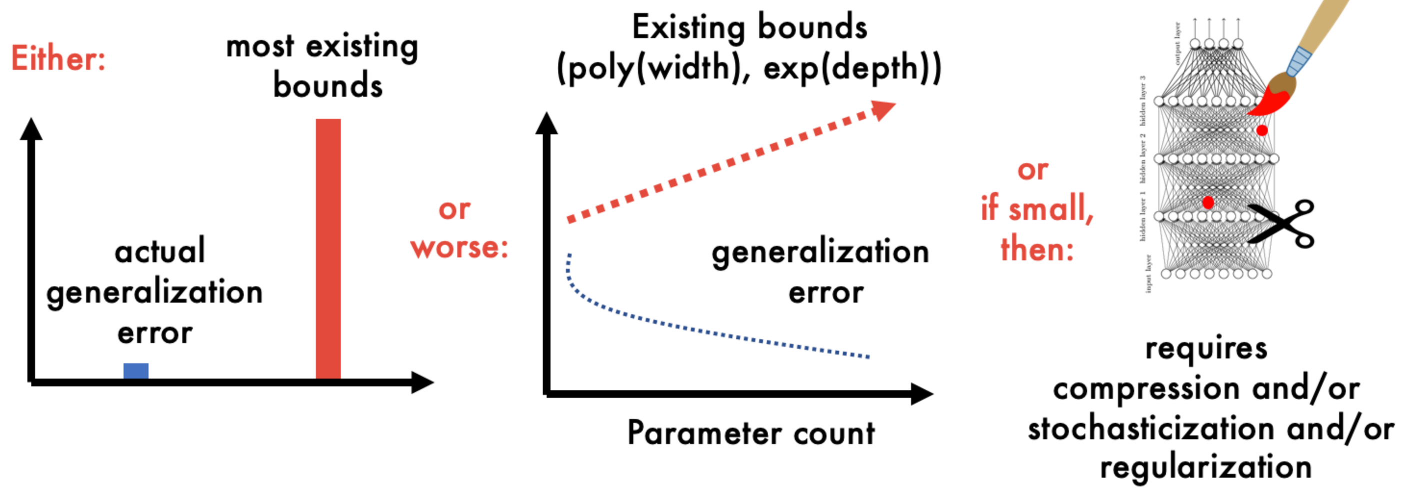
$$\text{generalization gap} \leq O\left(\sqrt{\frac{\text{representational complexity of relevant subset of hypotheses class}}{\text{training set size}}}\right)$$

Potentially depends on norms controlled by SGD e.g., dist. from init, spectral norm

PROGRESS SO FAR

Many many, novel, refined u.c. bounds have been proposed, using Rademacher complexity, covering numbers, compression, PAC-Bayes.

While each bound explains generalization in some aspect, it also fails in some other aspect.



Our 1st finding: GENERALIZATION BOUNDS ↑ WITH TRAINING SET SIZE

Notation: For input x, let logit output of network f on class k be $f(x)[k]$. On datapoint (x,y), define margin of f to be

$$\Gamma[f(x),y] \coloneqq f(x)[y] - \max_{y' \neq y} f(x)[y'].$$

Let S be training set of m examples drawn i.i.d from D. Denote network depth by d and width by h.

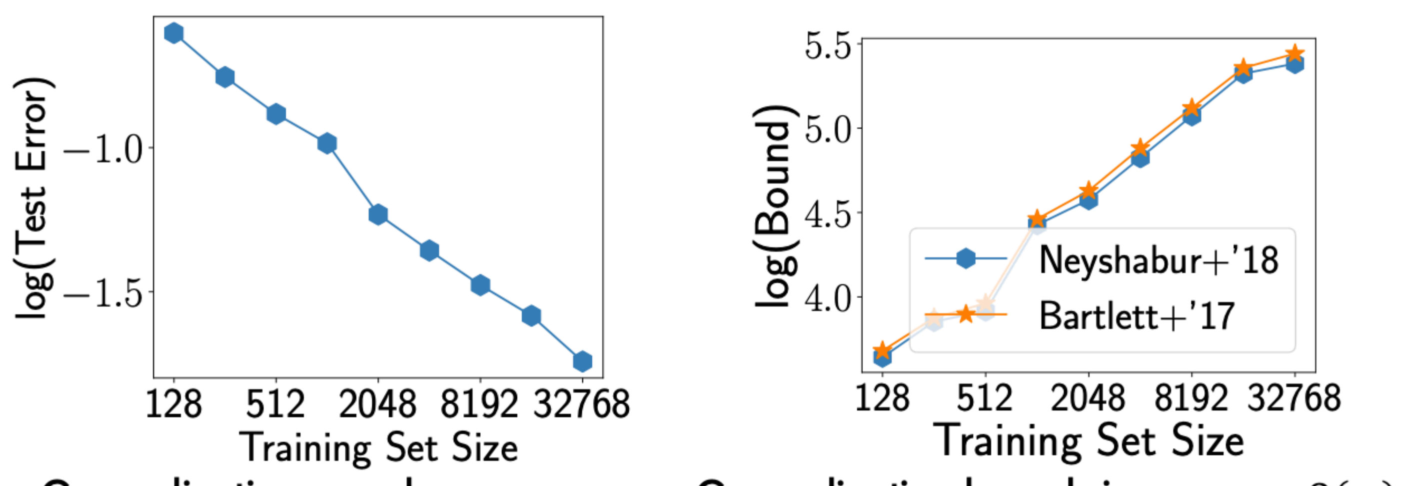
Experimental Setup: SGD with learning rate 0.1 and mini-batch size 1 until 99% of (random subset of) MNIST is classified by a margin of 10.

We evaluate bounds from [3,4], other bounds behave similarly:

$$Pr_D[\Gamma(f(x),y) \leq 0] \leq \frac{1}{m}\Sigma_{\{(x,y)\in S\}} \mathbf{1}[\Gamma(f(x),y) \leq \gamma] + O\left(\frac{\blacksquare}{\gamma\sqrt{m}}\right)$$

where:

in [3] $\blacksquare = d\sqrt{h}\prod_{k=1}^{d}||W||_2\sqrt{\Sigma_{k=1}^{d}\frac{||W_k - Z_k||_F^2}{||W||_2^2}}$ & in [4] $\blacksquare = \prod_{k=1}^{d}||W||_2\left(\Sigma_{k=1}^{d}\left(\frac{||W_k - Z_k||_F^2}{||W||_2^2}\right)^{2/3}\right)^{3/2}$



Generalization gap decreases as $O(m^{-0.43})$

Generalization bounds increase as $\Omega(m)$ (For $d = 5, h = 1024, \gamma \leftarrow 10$)

See also [5] for norms-vs-training-set-size plots in kernel learning, although for data with partially-corrupted labels and [6,7] for norm-vs-training-set-size plots.

Takeaway: Parameter-count dependence is only one part of the puzzle. We must worry about training-dataset-size dependence too!

Our 2nd finding: PROVABLE FAILURE OF UNIFORM CONVERGENCE

We show that there are situations in deep learning where any uniform convergence bound however refined, will provably fail to explain generalization.

generalization gap ≤ any refined u.c. bound

even though this is small (≈ 0)   this will be vacuous (≈ 1)

Key element in proof: TIGHTEST UNIFORM CONVERGENCE

Notations: Let $h_S \in \mathbb{H}$ be hypothesis learned on dataset $S \sim D^m$. Let $L_D(h)$ denote test 0-1 error of h & $\hat{L}_S(h)$ denote empirical 0-1 error on S.

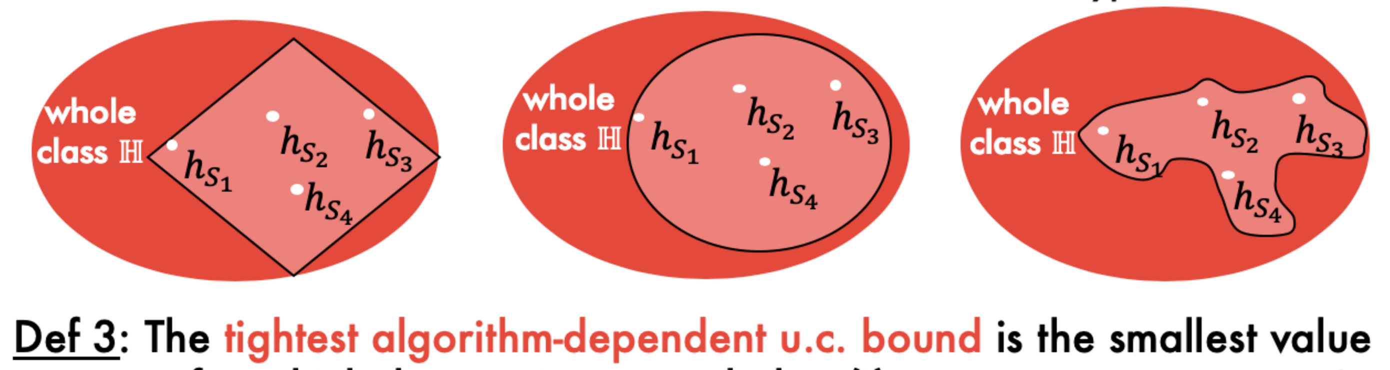
Def 1: The generalization gap is the smallest value of ϵ_{gen} s.t. with prob. $1-\delta$ over draws of $S \sim D^m$:

$$L_D(h_S) - \hat{L}_S(h_S) < \epsilon_{gen}$$

Def 2: The "conventional" u.c. bound is the smallest value of ϵ_{unif} s.t. with prob. $1-\delta$ over draws of $S \sim D^m$:

$$\sup_{h\in\mathbb{H}}|L_D(h) - \hat{L}_S(h)| < \epsilon_{unif}$$

To refine this bound, we can consider many different kinds of "relevant subsets" in $\mathbb{H}$ e.g.,

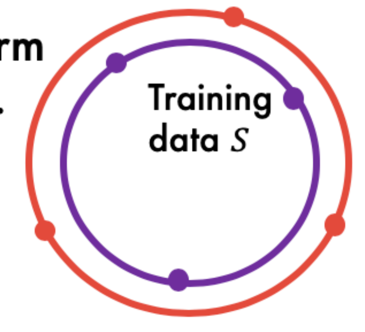


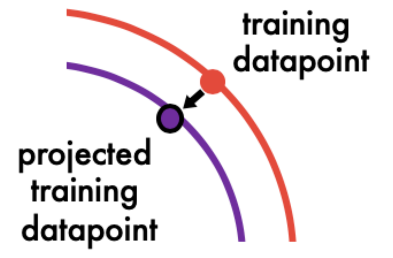
Def 3: The tightest algorithm-dependent u.c. bound is the smallest value $\epsilon_{unif-alg}$ for which there exists $\mathbb{S}_\delta$ such that (i) $Pr_{S\sim D^m}[S \notin \mathbb{S}_\delta] \leq \delta$ and (ii) $\mathbb{H}_\delta \stackrel{\text{def}}{=} \{h_S | S \in \mathbb{S}_\delta\}$ and finally (iii):

$$\sup_{S\in\mathbb{S}_\delta}\sup_{h\in\mathbb{H}_\delta}|L_D(h) - \hat{L}_S(h)| < \epsilon_{unif-alg}$$

FAILURE OF U.C. IN A HYPERSPHERE EXAMPLE

We train a 2-layer network of width $h = 100k$ using SGD to classify two concentric uniform 1000-dimensional hyperspheres of radius 1 and 1.1.



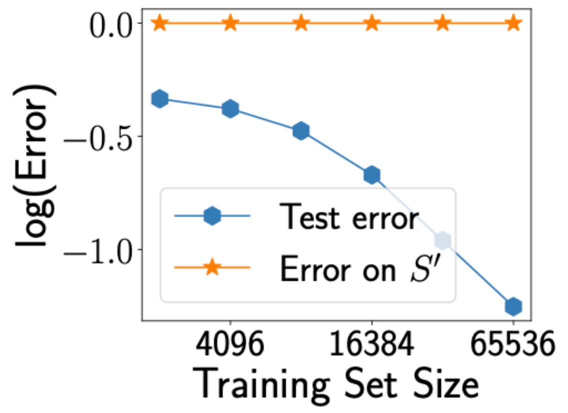


Next, we create a projected training set S'.

(by projecting each training datapoint onto its opposite hypersphere and flipping to correct label)

Observe that while generalization gap improves with m, S' is always completely misclassified (even though it's a "valid" dataset).

We show that this leads to failure of tightest u.c. (i.e., $\epsilon_{unif-alg} \approx 1$) and hence failure of all u.c. bounds.



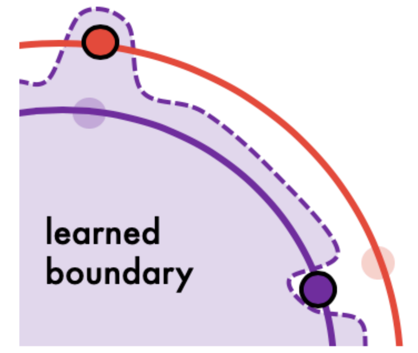
PROOF INTUITION

Key Lemma: For any given training set S, if we can design a corresponding "bad set" S' such that h_S misclassifies S' and $S' \sim D^m$ then $\epsilon_{unif-alg} \geq 1 - \epsilon_{gen}$.

(i.e., the distribution of S' without conditioning on S must be i.i.d from D)

Mathematical intuition:

- On one hand, $\forall S \in \mathbb{S}_\delta$, for the corresponding h_S both $\hat{L}_S(h_S)$ and $L_D(h_S)$ are small ⟹ small ϵ_{gen}.
- However, at the same time, $\exists\, S \in \mathbb{S}_\delta$ with a "bad" counterpart $h \in \mathbb{H}_\delta$ such that $\hat{L}_S(h)$ is large and $L_D(h)$ is small ⟹ large $\epsilon_{unif-alg}$.



Conceptual intuition: In order to classify most test data correctly, but misclassify S', the learned boundary must be representationally complex enough to "memorize" skews at S'.

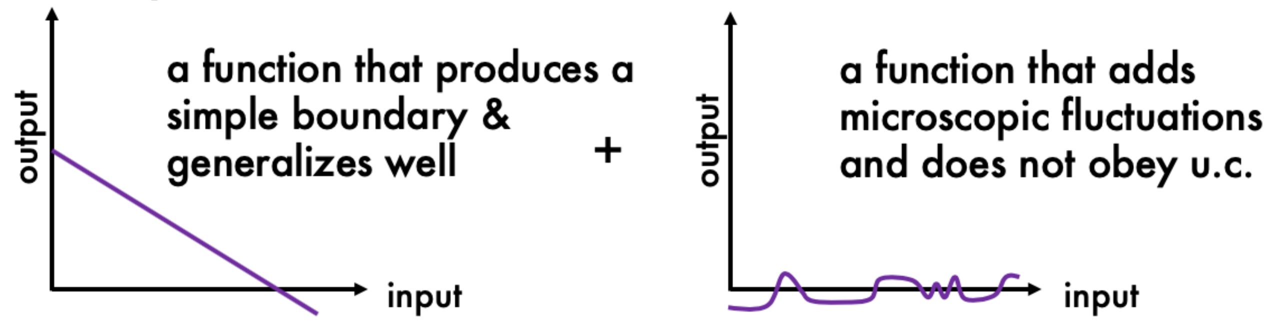
Takeaway: The decision boundary learned by SGD on overparameterized deep networks can have certain complexities which hurt u.c., without hurting generalization.

CONCLUSIONS AND FUTURE WORK

Can uniform convergence provide a complete answer to the generalization puzzle? Most likely, not.

Must go beyond uniform convergence – derive new tools using our negative examples as test cases

Conjecture: Functions learned by deep networks can be decomposed into:



REFERENCES

[1]: Zhang et al., *Understanding deep learning requires rethinking generalization*, ICLR '17
[2]: Neyshabur et al., *In Search of the Real Inductive Bias: On the Role of Implicit Regularization in Deep Learning*, ICLR '15 Workshop Track
[3]: Neyshabur et al., *A PAC-Bayesian approach to spectrally-normalized margin bounds for neural networks.* ICLR' 18
[4]: Bartlett et al., *Spectrally-normalized margin bounds for neural networks*, NeurIPS'17
[5]: Belkin et al., *To understand deep learning we need to understand kernel learning*, ICML '18
[6]: Nagarajan & Kolter, Generalization in deep networks: the role of distance from initialization, NeurIPS '17 workshop
[7]: Neyshabur et al., *Exploring Generalization in Deep Learning*, NeurIPS' 17