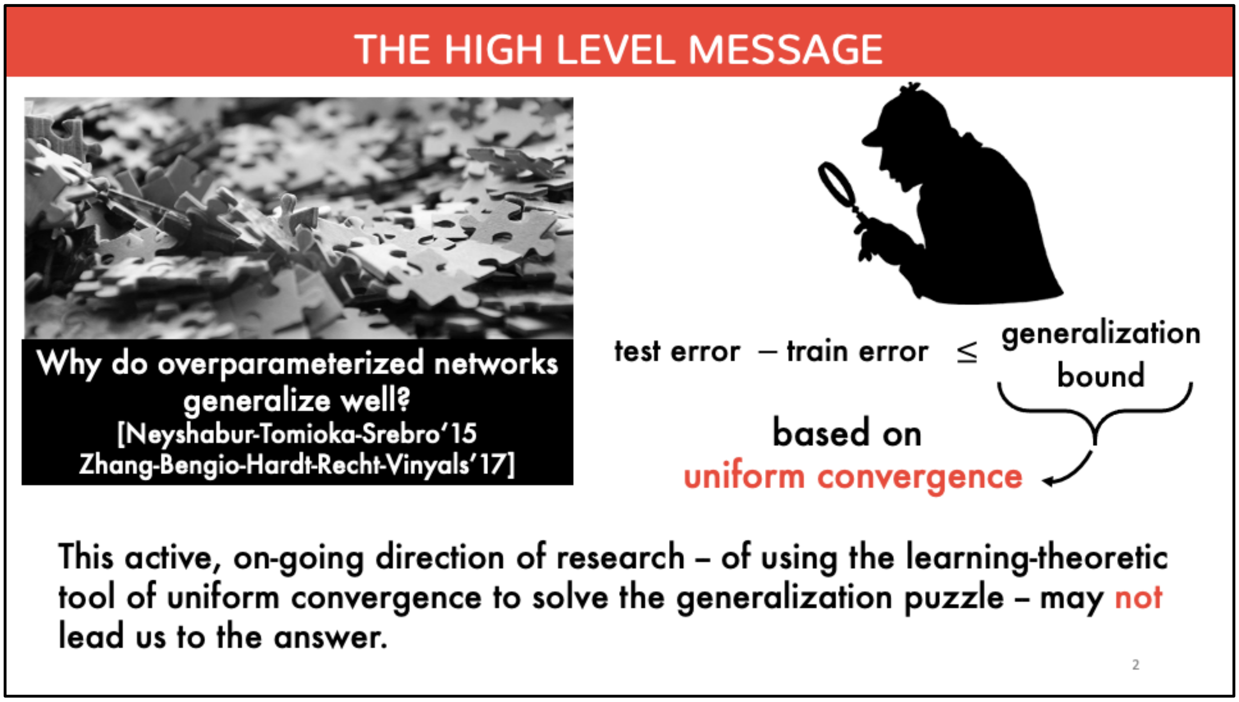
UNIFORM CONVERGENCE MAY BE UNABLE TO EXPLAIN GENERALIZATION IN DEEP LEARNING

Vaishnavh Nagarajan* J. Zico Kolter*†

* Computer Science Department,
Carnegie Mellon University

† Bosch Center for AI
Pittsburgh

1



One of the biggest open challenges in deep learning theory is the generalization puzzle. Classical learning theory suggests that models that have many many more parameters than training datapoints, should not really generalize well. But, deep network models generalize very well inspite of heavy overparameterization. What explains this counter-intuitive behavior?

Theoretical works have tried to understand this by deriving upper bounds on the generalization gap of deep networks. Notably, most of these bounds are based on the same learning-theoretic idea of uniform convergence. Now, despite a lot of work in this space, a tight generalization bound has so far proven to be elusive.

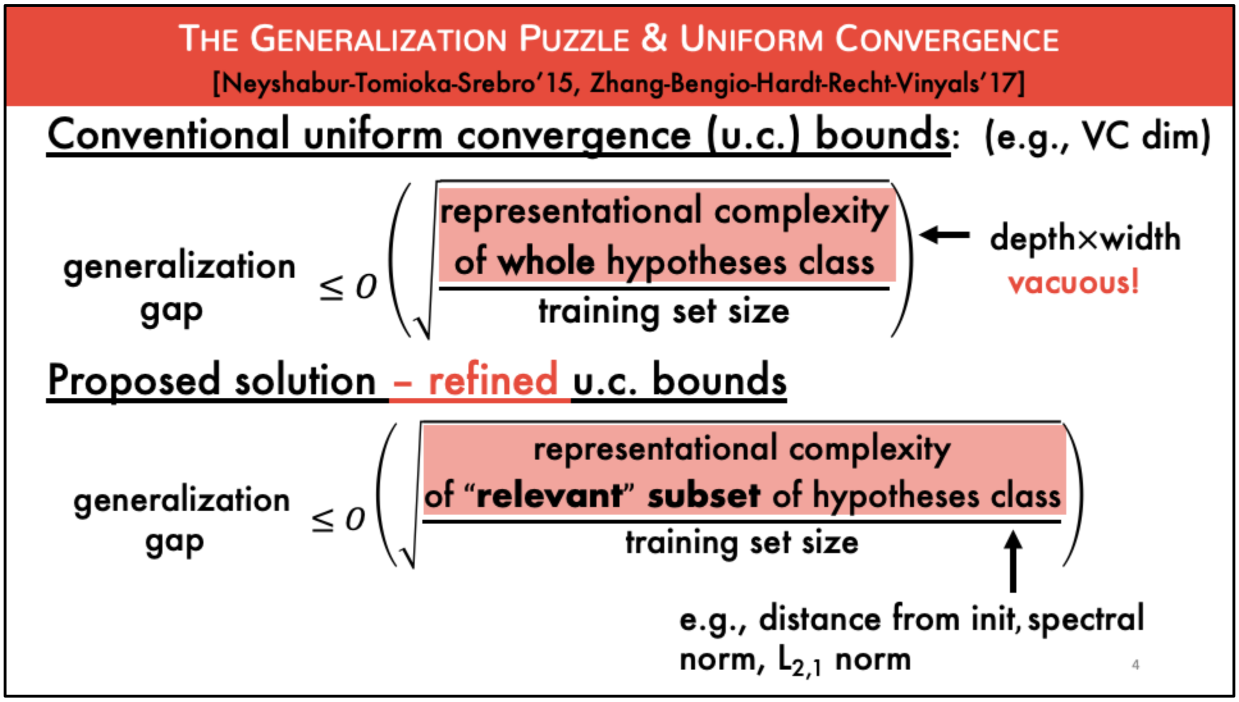
In this work, we take a step back, and argue that this high level direction of pursuing uniform convergence-based bounds may not actually lead us to the complete solution of this puzzle.

OUTLINE

- PAST WORK
- OUR FIRST FINDING: **Bounds grow with training set size**
- OUR SECOND FINDING: **Provable failure of uniform convergence**

3

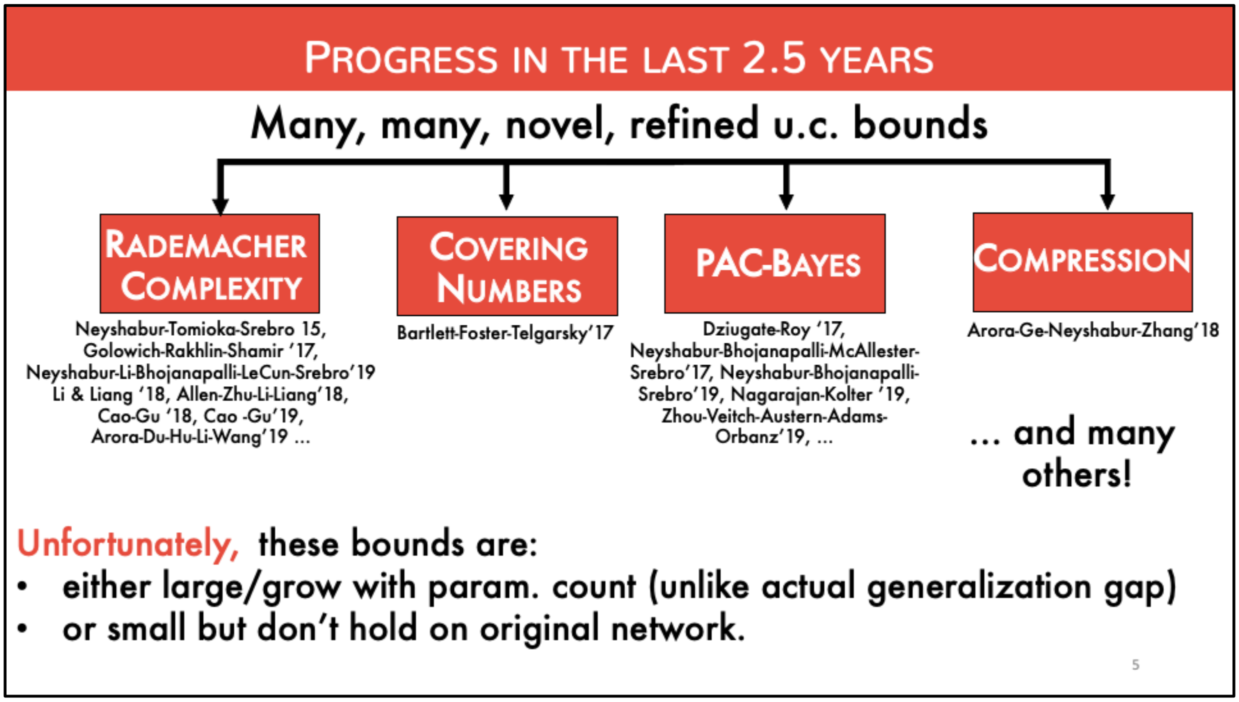
For this talk, we'll first go over some background work and then look at the two main findings which are limitations of u.c. bounds.



To recap, here's what we already know about uniform convergence before this work.

Standard uniform convergence bounds measure the representational complexity of the whole class of functions representable by a deep network. However, a deep network is an extremely expressive model, as a result of which these bounds are vacuous. Mathematically, the numerator here grows with the parameter count, and is hence larger than the denominator in the overparameterized setting, leading to a vacuous bound.

To fix this, the solution proposed was to take into account the **implicit bias** of SGD. That is, we must derive these bounds by "ignoring extraneous hypothesis" and focusing only on those that are relevant to the algorithm and the data distribution The hope was this would yield tighter bounds typically by depending on the weight norms of the network that are controlled by SGD, like its spectral norms, distance from initialization etc.,



This proposal triggered an exciting and active area of research resulting in a wide variety of
refined uniform convergence bounds over the last couple of years: PAC-Bayesian to Rademacher to covering number to compression based bounds.
All these works shed a lot of varying insights into why deep networks generalize well. I.e. each of these bounds explained certain aspects of generalization.

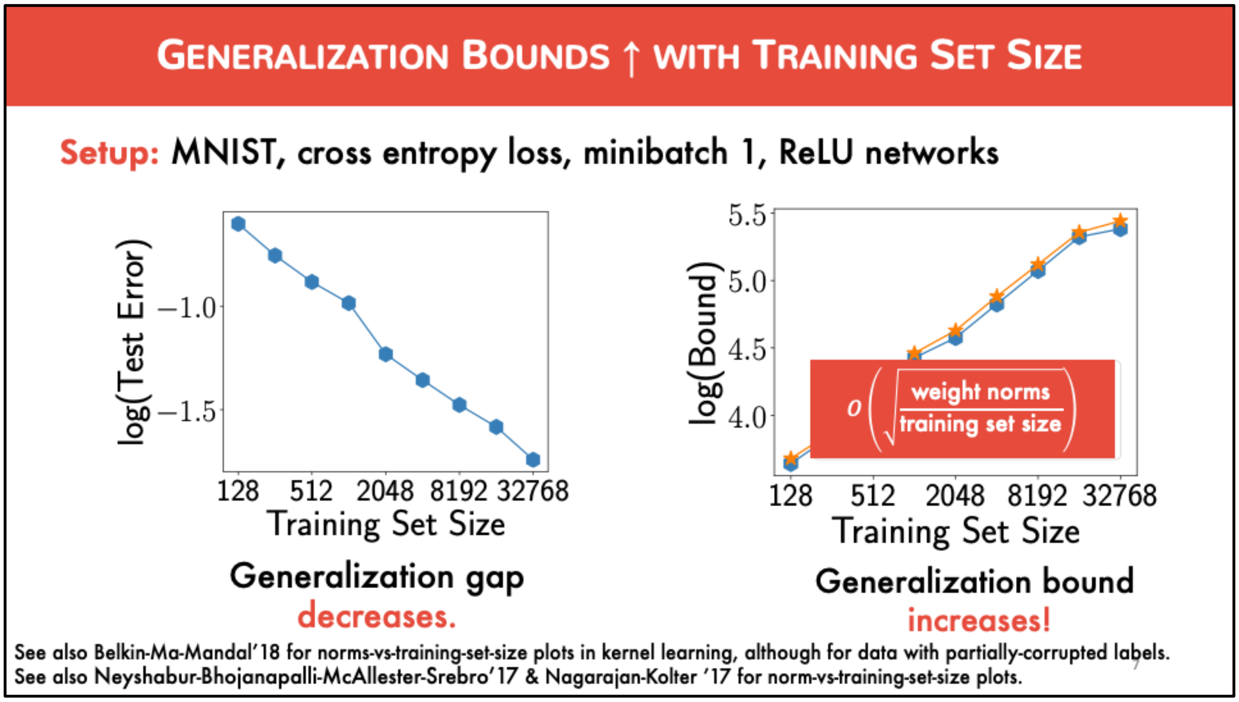
Unfortunately, while these papers explained generalization in one way or the other, they also failed to explain eneralizationi in some other way. These are either too large or grow with parameter count unlike the actual generalization gap. The ones that are small require the network learned by SGD to be modified say, by compression or explicit regularization.

OUTLINE

- BACKGROUND & PAST WORK
- OUR FIRST FINDING: **Bounds grow with training set size**
- OUR SECOND FINDING: **Provable failure of uniform convergence**

6

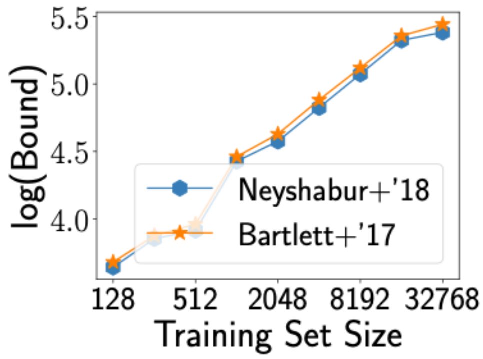
In this backdrop, we take a step back and study limitations of uniform convergence based bounds. Our first finding is an empirical limitation of these bounds.



, we observe that there are certain hyperparameter configurations for which, on one hand, the test error and the generalization gap decrease with training set size, as expected. However, unfortunately, existing generalization bounds increase with the training set size. And this holds despite the fact that the denominator in these bounds grow with training set size. And this is so, because the numerator here has certain weight norms which grow drastically with the training set size. We present many other related observations about this in the paper, but the main take away is that:

Generalization Bounds ↑ with Training Set Size

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



8

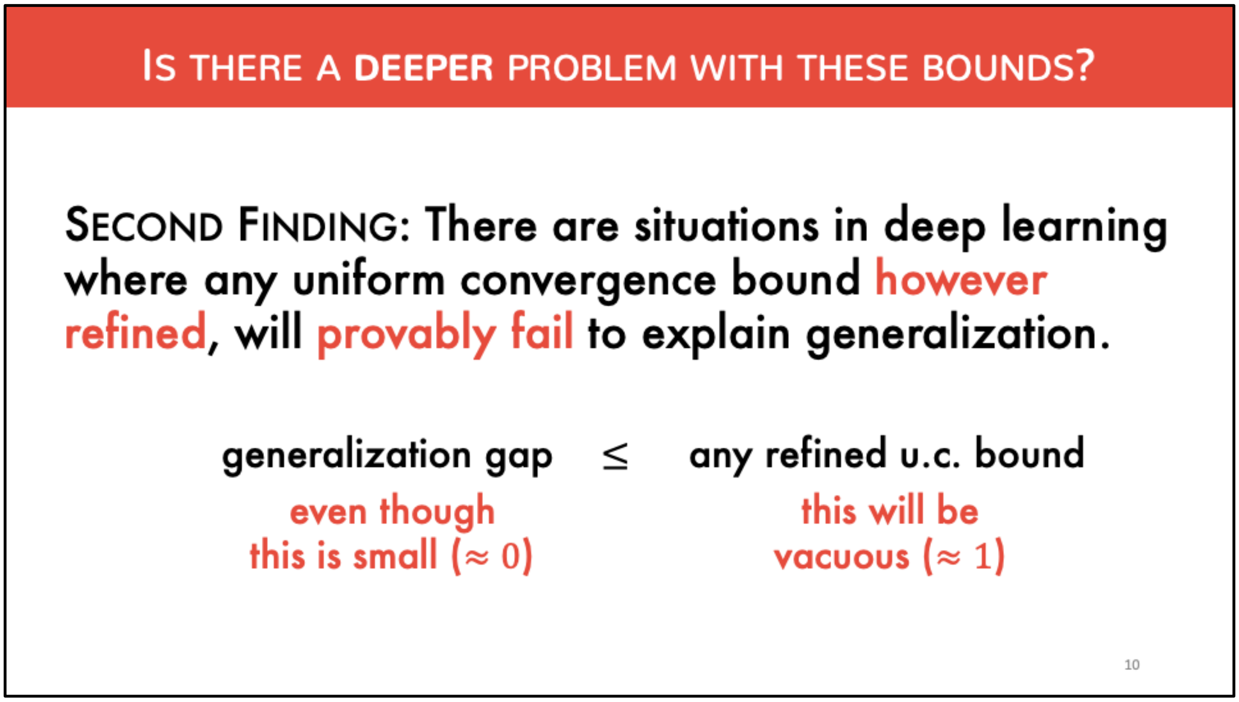
The main takeaway from our first finding here is that, on one hand we have all been focusing on the parameter count dependence of these generalization bounds. At the same time, our finding highlights that we should also worry about deriving generalization bounds that have at least a reasonable kind of dependence on the training set size… as that is another aspect of generalization that ours bounds should be able to explain.

OUTLINE

- PAST WORK
- OUR FIRST FINDING: **Bounds grow with training set size**
- OUR SECOND FINDING: **Provable failure of uniform convergence**

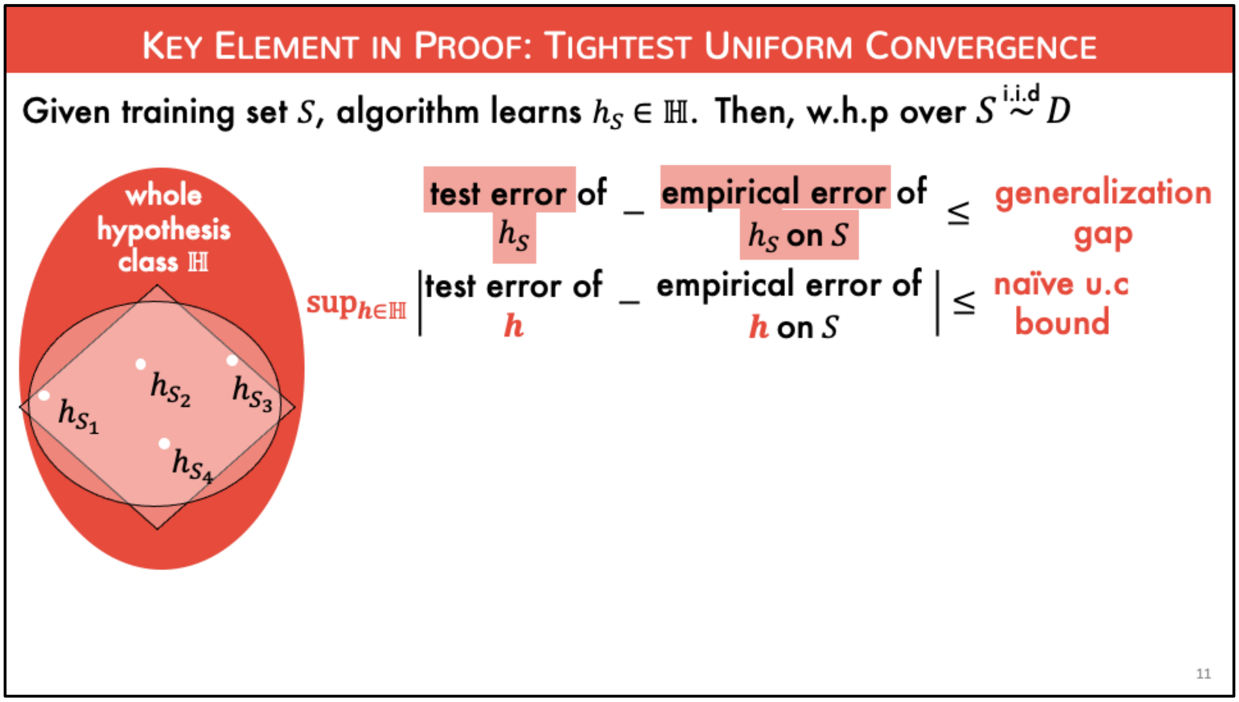
9

So, clearly, uniform convergence bounds, at least the ones that exists, seem to suffer from problems in practice. But one might still wonder, if it’s possible somehow cleverly refine these bounds in a way that we can overcome all these problems. To this, we present our second finding which is a provable failure of u.c.



Specifically, we show that any uniform convergence bounds, including all kinds of refined uniform convergence bounds, will provably fail to explain generalization in certain situations in deep learning.

By this, we mean that even though the generalization gap in these settings is really small, any refined uniform convergence bound would be vacuous in these settings.

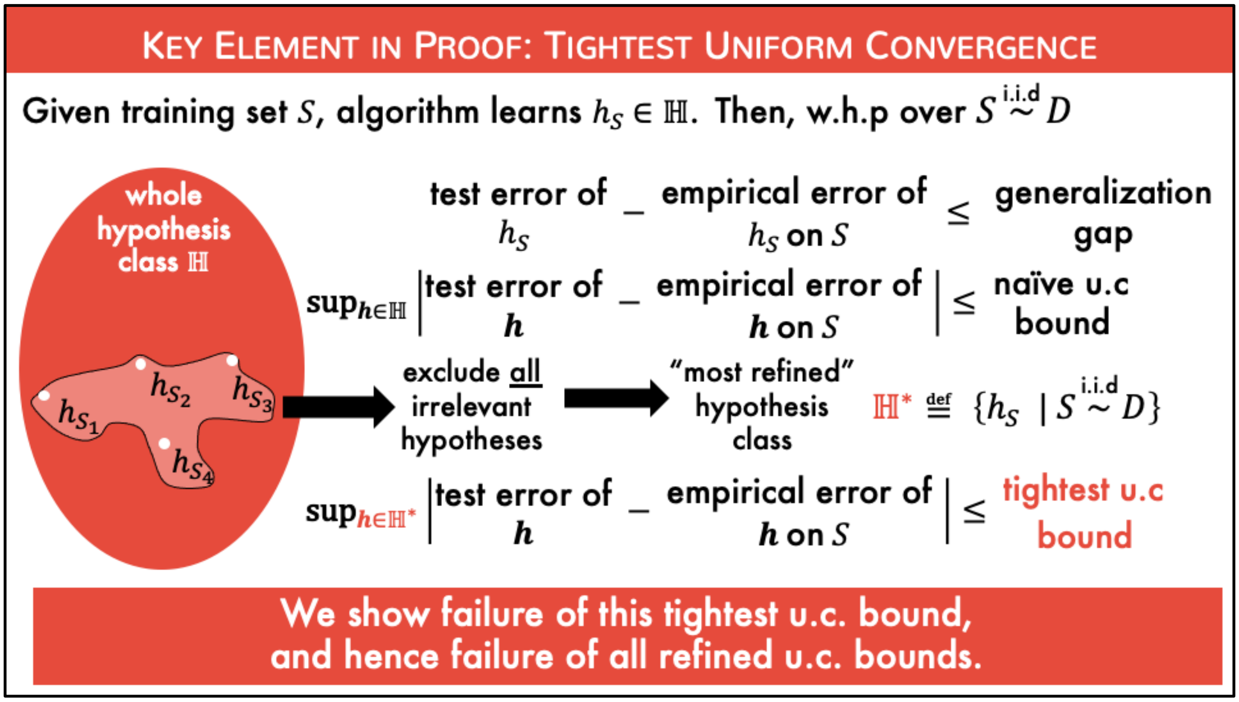


How do we show this? A crucial element in our proof is what we define as the tightest uniform convergence bound, which we eventually show is vacuous. To see what we mean by this term, let us go over some quick technical definitions.

Given a training set S, let us denote the hypothesis learned by the algorithm by h_S. Then, w.h.p over the training set drawn iid from an underlying distribution D, the generalization gap is the difference between the test error and the empirical error on the dataset S, for the hypothesis h_S learned on S. In other words, the difference between the test and the training error.

Now, a naïve u.c. bound is an upper bound this quantity. And this essentially corresponds to looking at the difference between the empirical and test error uniformly overall hypothesis in the hypothesis class H, instead of just the hypothesis learned on the dataset S.

However, this is pretty loose. And to refine this, we want to exclude irrelevant hyptoehsis. There would be many ways to do this, For example, by focusing on an l2 norm ball containing the relevant hypothesis or an l1 norm ball. And this is precisely the kind of pursuit that has been happening in deep learning: we want to identify a nice kind of refinement which would lead to a meaningful bound.

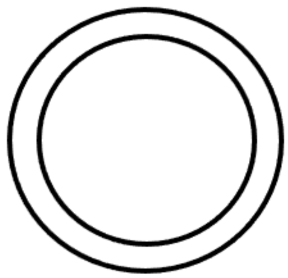


But, let us take this to the extreme and think about the tightest possible refinement. To get the tightest possible refinement, we need to ignore all irrelevant hypothesis on focus on a subset of hypotheses that includes only those hypothesis learned by the **given algorithm** for the **given data distribution**. Let's call this hypothesis call H^*. Then, the tightest uniform convergence bounds would be one where the sup is restricted to H^*. and all other u.c. bounds would be llooser than this bound.

Having defined this quanity, we essentially show that even this tightest u.c. bound would become vacuous in certain settings, and so does all other u.c. bounds

SECOND FINDING: PROVABLE FAILURE OF U.C.

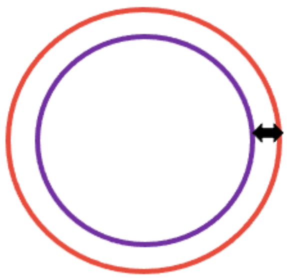
Setup: 1000-dimensional hypersphere classification task; 1-hidden-layer ReLU, 100k units, SGD, trained to zero error.



Here’s the setting where we show failure of u.c. Consider a binary classification task where we are given two uniform hypersphere distributions in 1000 dimensions. The two hyperspheres are close to each other and we must learn a decision boundary that separates them.

SECOND FINDING: PROVABLE FAILURE OF U.C.

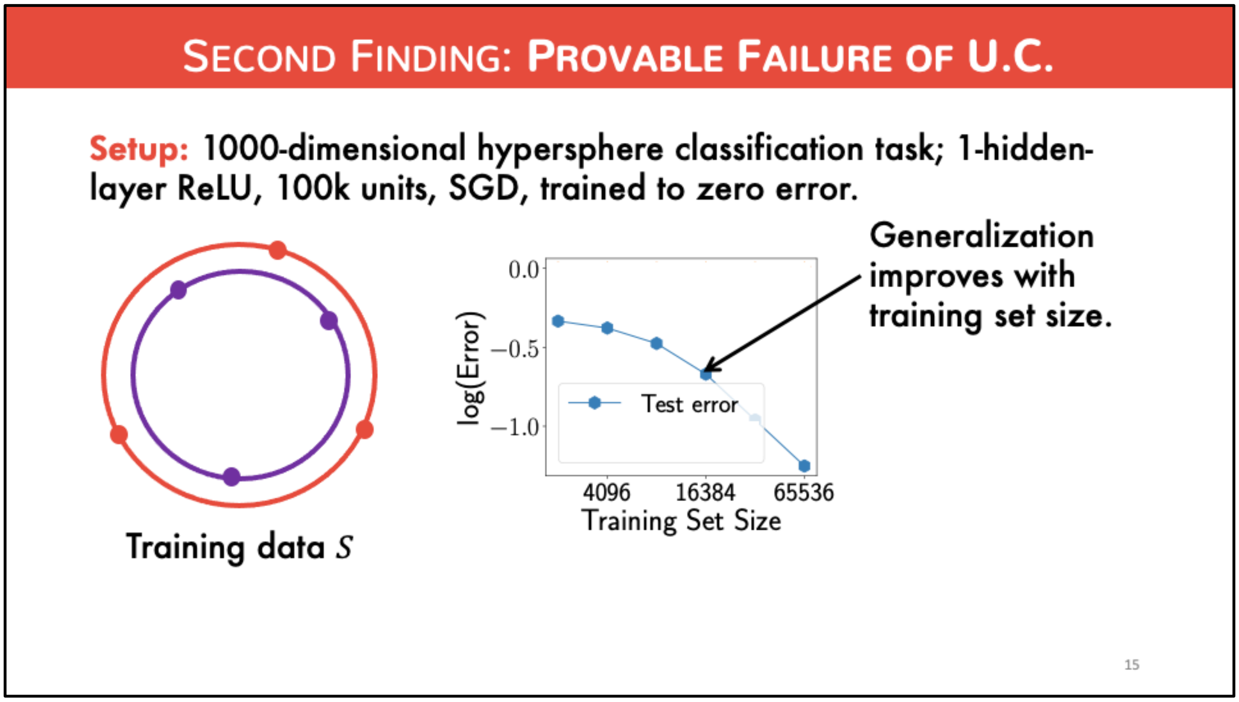
Setup: 1000-dimensional hypersphere classification task; 1-hidden-layer ReLU, 100k units, SGD, trained to zero error.



close to each other,
but perfectly separable
(no label noise)

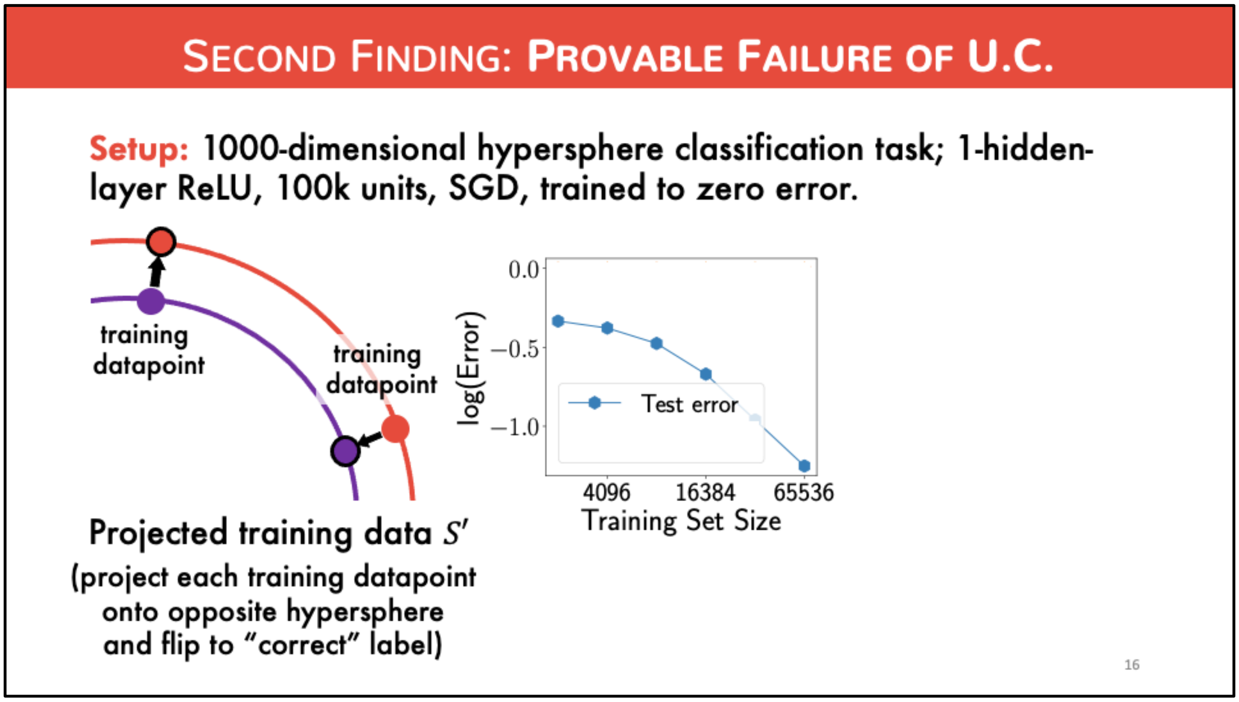
14

These hyperspheres are close to each other, but they are perfectly separable since there is no label noise.



First, we observe that as we have more and more training data, as seen in this plot, the network improves it generalization.

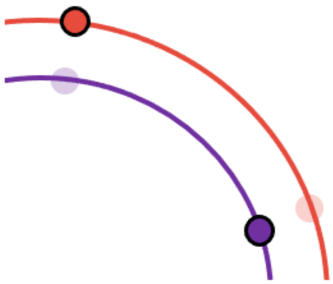
Now to show failure of uniform convergence, there are two key steps.



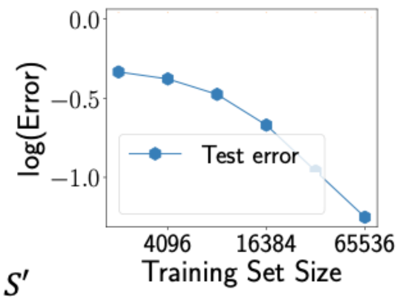
Our first step is to create a dataset S' which is obtained by projecting each training point onto its opposite hypersphere and flipping its label to match the opposite hypersphere.

SECOND FINDING: PROVABLE FAILURE OF U.C.

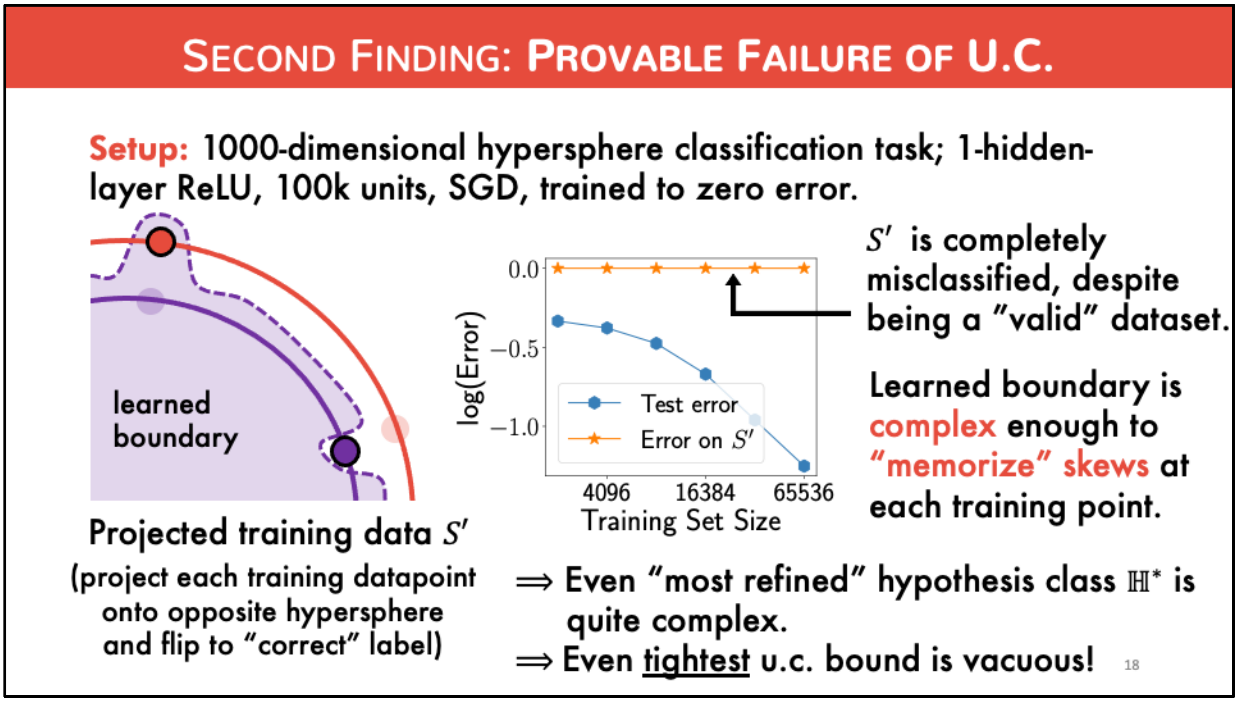
Setup: 1000-dimensional hypersphere classification task; 1-hidden-layer ReLU, 100k units, SGD, trained to zero error.



Projected training data S'
(project each training datapoint onto opposite hypersphere and flip to "correct" label)



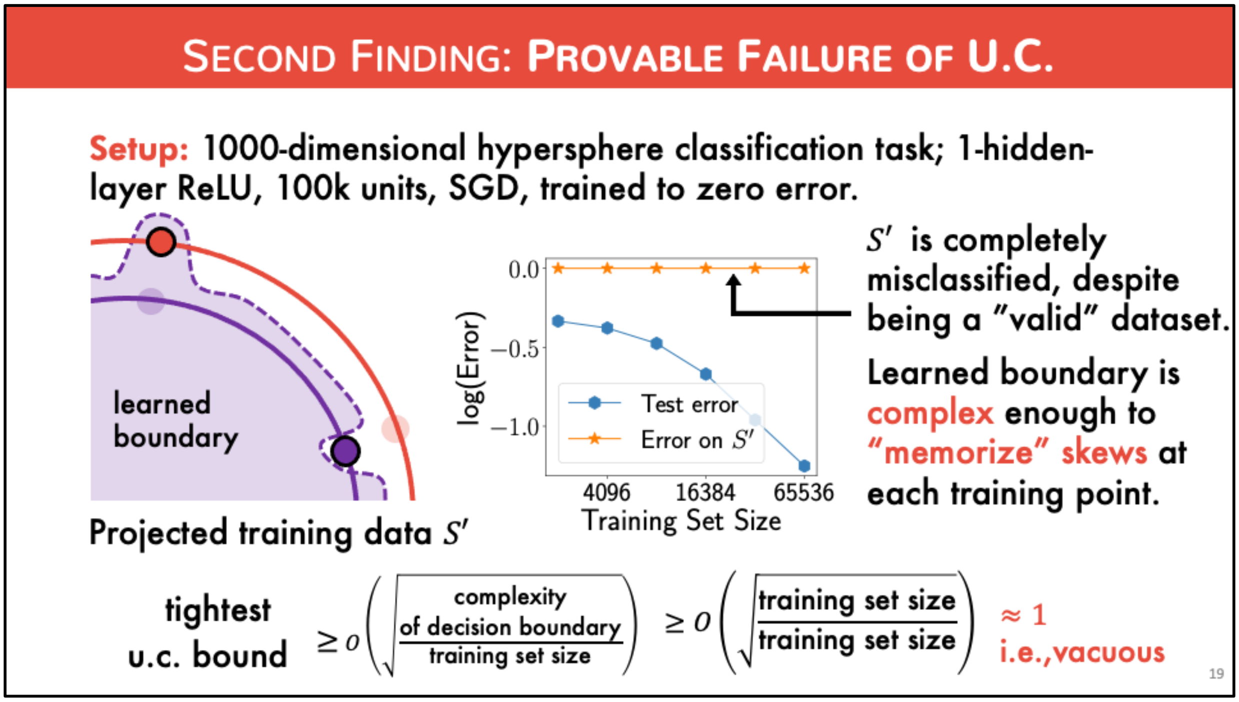
17



As a crucial second step, we demonstrate that even though the test error and the training error is very small, the dataset S' is completely misclassified as seen in this plot.
Intuitively, this indicates that there are skews in the decision boundary located specifically around the training datapoints which results in misclassification of the projected datapoints S'.

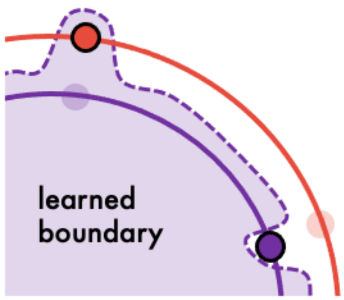
Or in other words, the decision boundary is itself inherently quite complex -- complex enough to memorize skews in
the locations of the training datapoints.

We then mathematically show that this sort of inherent complexity renders even the most refined hypothesis class H^*, which we defined a couple of slides ago, to be very complex. And so all uniform convergence bounds, included the tightest one, is limited/lower-bounded by this inherent complexity. Thus, all these bounds become vacuous in this particular setting.



To be a bit more mathematical, we would have that the numerator in the tightest u.c. bound is as large as the training set, and hence the bound becomes vacuous. This is the outline of our proof.

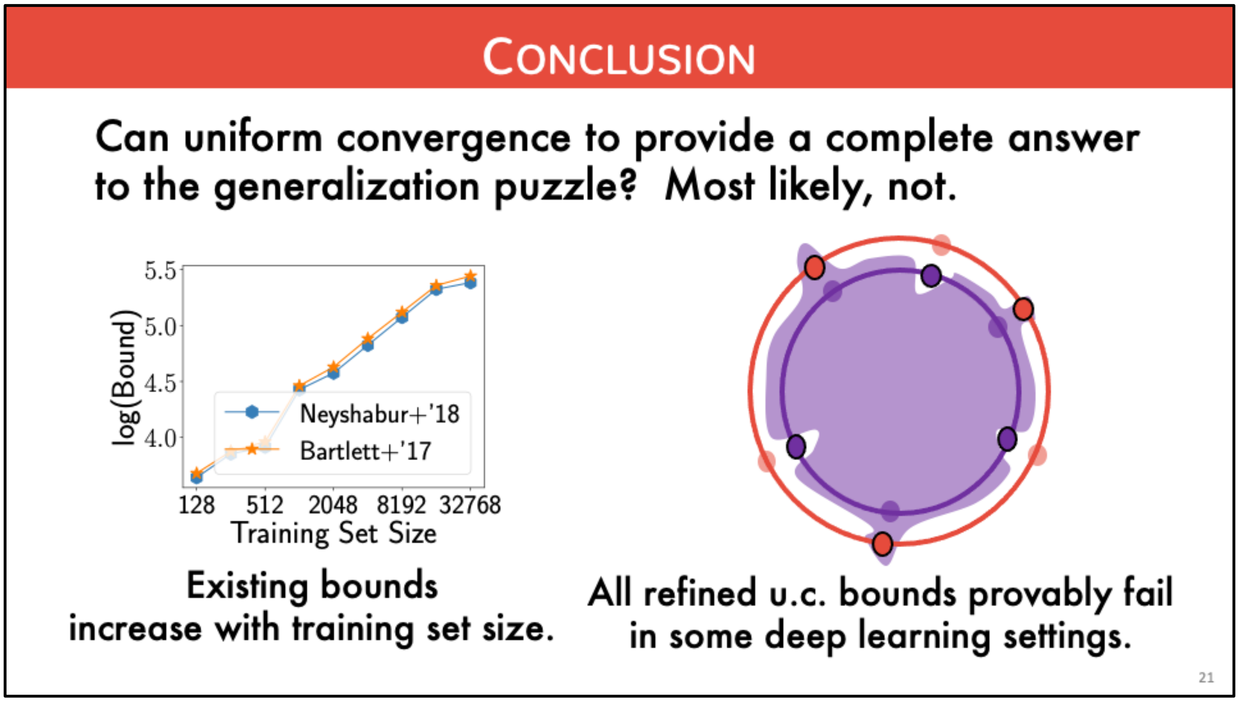
Second Finding: Provable Failure of U.C.



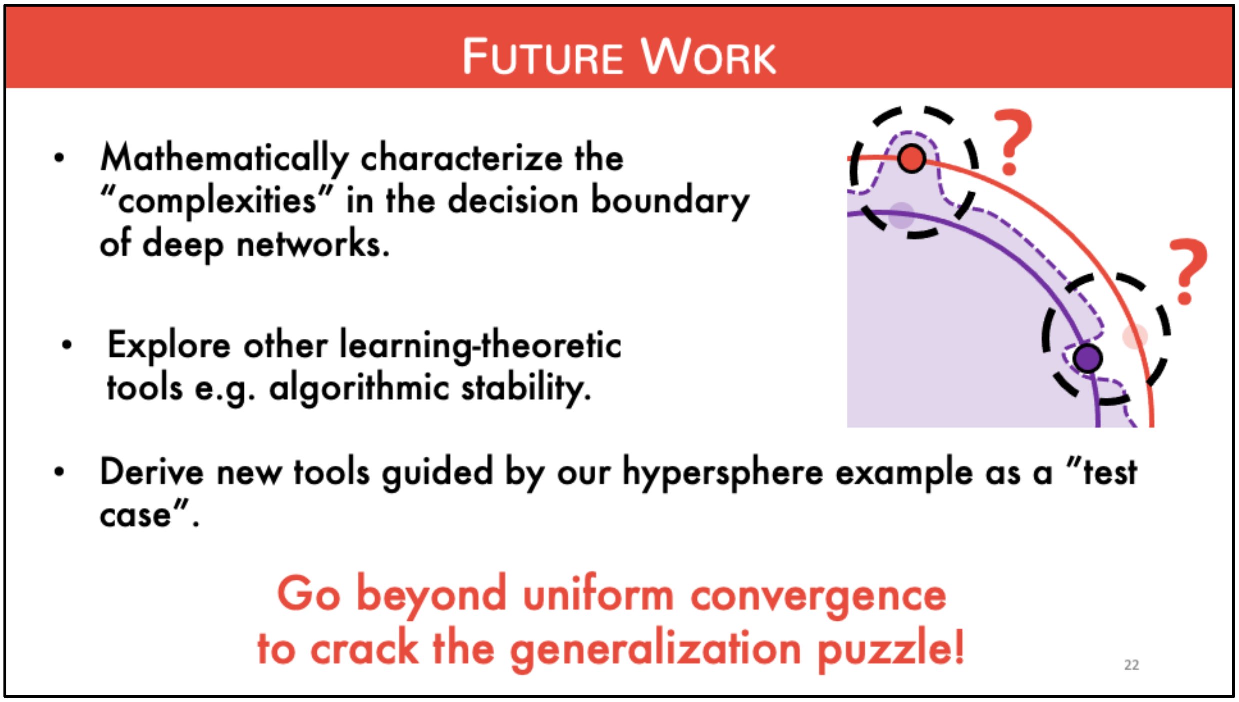
The decision boundary learned by SGD on overparameterized deep networks can have certain complexities which hurt uniform convergence, without hurting generalization.

20

The main takeaway from this second finding is the following: the decision boundary learned by SGD on overparamtereize deep networks can be inherently complex in certain ways, and due to these complexities, uniform convergence can provably fail. At the same, these complexities do not hurt generalization error.



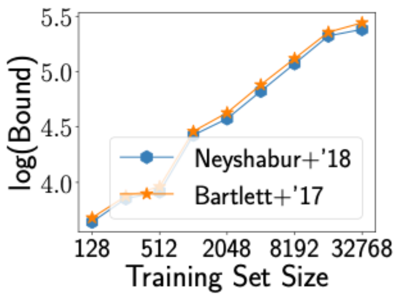
In conclusion, in this work, we cast doubt on the power of uniform convergence bounds to fully explain generalization in deep learning. First, we highlight that explaining the training-set-size dependence of the generalization error is apparently just as non-trivial as explaining its parameter-count dependence. Furthermore, we also showed that there are scenarios where all uniform convergence bounds, however cleverly applied, become vacuous.

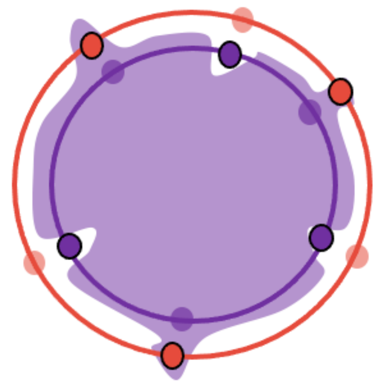


In order to get a better grasp of generalization in deep learning, we believe that it is essential to better understand the complexities that we observed in the decision boundaries learned by deep networks. Furthermore, it may be useful to more carefully explore other learning-theoretic tools like algorithmic stability. Perhaps the most exciting direction would be to derive new learning-theoretic tools. And to do this, our negative examples may be useful test cases.
Overall, we belive that going beyond uniform convergence may be essential to fully explaining generalization in deep learning.

THANK YOU!

Poster #229
10:45 AM – 12:45 PM
@East Exhibition Hall B + C #229





23