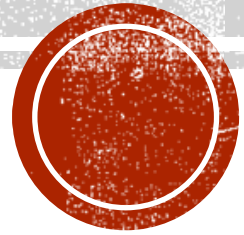
NONPARAMETRIC DENSITY ESTIMATION AND CONVERGENCE OF GANS UNDER BESOV IPM LOSSES

Team: Pradyum Soni, 170020001



OUTLINE

1. Generalization in Deep Learning
2. Uniform Convergence in Generalization
3. AIM of the paper
4. Wasserstein and Kolmogorov-Smirnov distances
5. Experiments:
    - Part 1 – Existing bounds increase with dataset size
    - Part 2 – The provable failure of Uniform Convergence

GENERALIZATION IN DEEP LEARNING

- Generalization is the gap between error on the training and test set drawn from the same distribution.
- To theoretically study the generalization behavior of a model, one must derive upper bounds on the gap between the error of the model on test data (unseen data) and the training data.

$$\text{Test error} \leq \text{Train error} + \mathcal{O}\left(\frac{\text{some complexity term}}{\sqrt{\text{training set size}}}\right)$$

UNIFORM CONVERGENCE IN GENERALIZATION

Deep artificial neural networks often have far more trainable model parameters than the number of samples they are trained on

Some of these models exhibit remarkably small generalization error, i.e., difference between “training error” and “test error”.

What is it then that distinguishes neural networks that generalize well from those that don’t?

Statistical learning theory has proposed a number of different complexity measures that are capable of controlling generalization error

These include VC dimension (Vapnik, 1998), Rademacher complexity (Bartlett & Mendelson, 2003), and uniform stability (Mukherjee et al., 2002; Bousquet & Elisseeff, 2002; Poggio et al., 2004).

CONTINUED: CLASSICAL APPROACH

- Standard generalization bounds (like the VC dimension, Vapnik) compute the "some complexity term" by considering a *uniform convergence* bound on the class of functions representable by a deep network.
- When we apply 'uniform convergence' on a class of functions, we basically compute how representationally rich the function class is. This approach has been proven to fail miserably!
- This is because the same deep network that generalizes well on a nice dataset, can also be shown to be representationally rich enough to fit noisy labels on that dataset.
- The below shown term, to which uniform convergence resolves to, is vacuous in overparameterized settings. Hence, it fails.

$$\text{Test error} \leq \text{Train error} + \mathcal{O}\left(\frac{\text{depth} \times \text{width}}{\sqrt{\text{training set size}}}\right).$$

CONTINUED: MODERN APPROACH

- Zhang et al, 2017: "search for the inductive bias" of SGD in deep learning

$$\text{Test error} \leq \text{Train error} + \mathcal{O}\left(\frac{\text{weight norms controlled by SGD for the given data distribution}}{\sqrt{\text{training set size}}}\right)$$

- Existing attempts at deriving stated generalization bounds span many different learning-theoretic techniques: PAC-Bayes, Rademacher complexity, covering numbers, compression.
- A lot of ongoing efforts in this space has been focused on developing more novel or refined variations of such uniform-convergence-based generalization bounds to better "explain generalization" in deep learning. However, the goal of providing a complete explanation of generalization through these bounds appears to be elusive. **The existing generalization bounds records progress in a particular aspect, but also falls short of explaining generalization in some other aspect.**

AIM OF THE PAPER

- The paper is aimed at explaining the surprisingly good generalization behavior of overparameterized deep networks, recent works have developed a variety of generalization bounds for deep learning, all based on the fundamental learning-theoretic technique of uniform convergence. While it is well-known that many of these existing bounds are numerically large, through numerous experiments, we bring to light a more concerning aspect of these bounds: **in practice, these bounds can increase with the training dataset size.**
- Guided by the authors' observations, they present examples of overparameterized linear classifiers and neural networks trained by gradient descent (GD) where **uniform convergence provably cannot 'explain generalization'** even if we take into account the implicit bias of GD

WASSERSTEIN AND KOLMOGOROV SMIRNOV DISTANCES

WASSERSTEIN DISTANCE:

- It is used to compare the probability distributions of two variables *X* and *Y*, where one variable is derived from the other by small, non-uniform perturbations (random or deterministic).
- In their paper 'Wasserstein GAN', Arjovsky et al. use the Wasserstein-1 metric as a way to improve the original framework of Generative Adversarial Networks (GAN), to alleviate the vanishing gradient and the mode collapse issues.

KOLMOGOROV SMIRNOV DISTANCE:

- In statistics, the **Kolmogorov–Smirnov test** is a nonparametric test of the equality of continuous one-dimensional probability distributions that can be used to compare a sample with a reference probability distribution.

$$\begin{aligned}\operatorname{Kolm}(\mu,\nu) &:= \sup_{x\in\mathbb{R}} |\mu((-\infty,x]) - \nu((-\infty,x])| \\ &\leq \operatorname{TV}(\mu,\nu).\end{aligned}$$

EXPERIMENT: EXISTING BOUNDS INCREASE WITH DATASET SIZE

- Focus on 3 generalization bounds:
  1. The PAC-Bayesian bound: here
  2. The covering number bound: here
  3. Rademacher complexity bound: here
- The above generalization bounds can be written as follows, ignoring some minor variations. For any training input (x,y) (where x is the feature vector and y∈{1,2,...K} is the class label), let f(x)[k] ∈ R be the output of the network on class k. If D denotes the underlying data distribution, and S the training set of m datapoints, then, for any γ≥0, we can upper bound the test error as follows:

$$\underbrace{Pr_{(x,y)\sim\mathcal{D}}[f(x)[y] < \max_{y\neq k} f(x)[k]]}_{\text{test 0-1 error}} \leq \underbrace{\frac{1}{m}\sum_{(x,y)\in S} \mathbf{1}[f(x)[y] < \max_{y\neq k} f(x)[k]+\gamma]}_{\text{proportion of training points misclassified by a "margin" of } \gamma} + \mathcal{O}\left(\frac{\text{some weight norms}}{\gamma\sqrt{m}}\right)$$

CONTINUED: THEORY

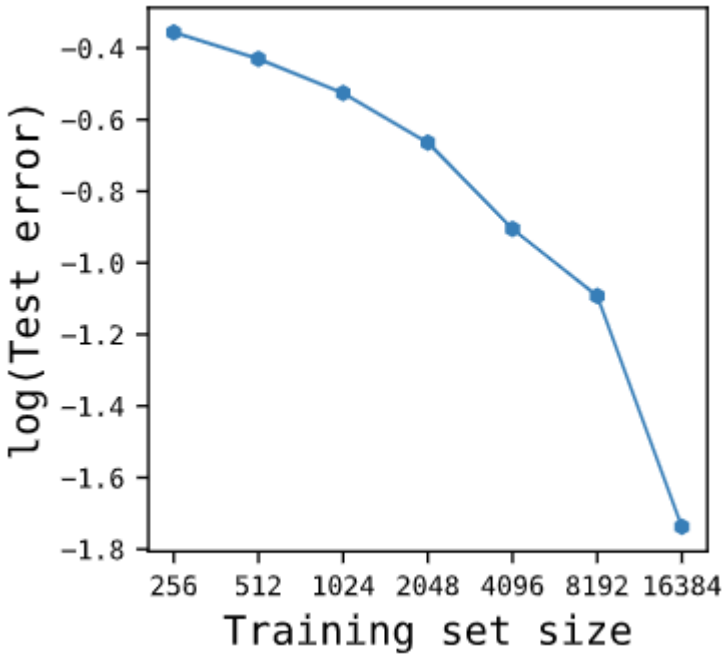
- Training Data: Assume we are given two classes distributed uniformly on two concentric hyperspheres in 100 dimensions. The inner sphere has radius 1 and outer sphere has radius 1.2

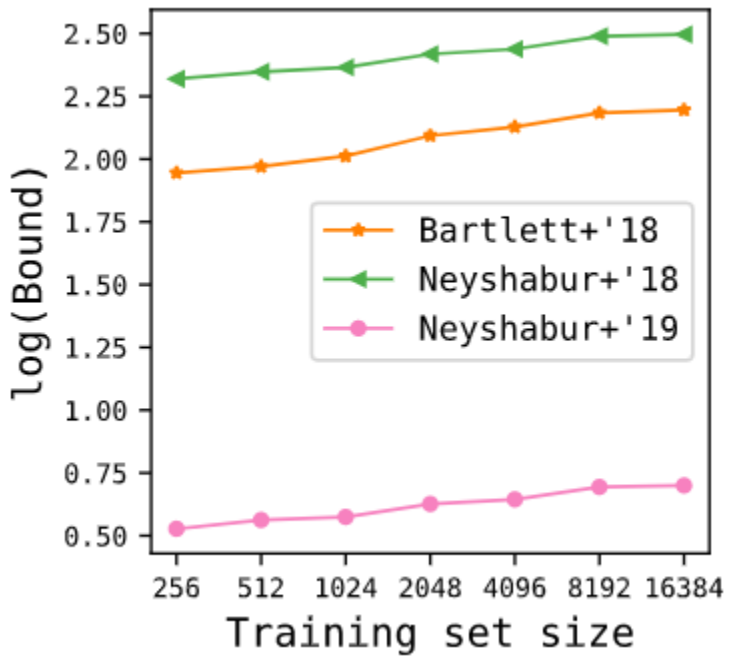
CONTINUED: THEORY

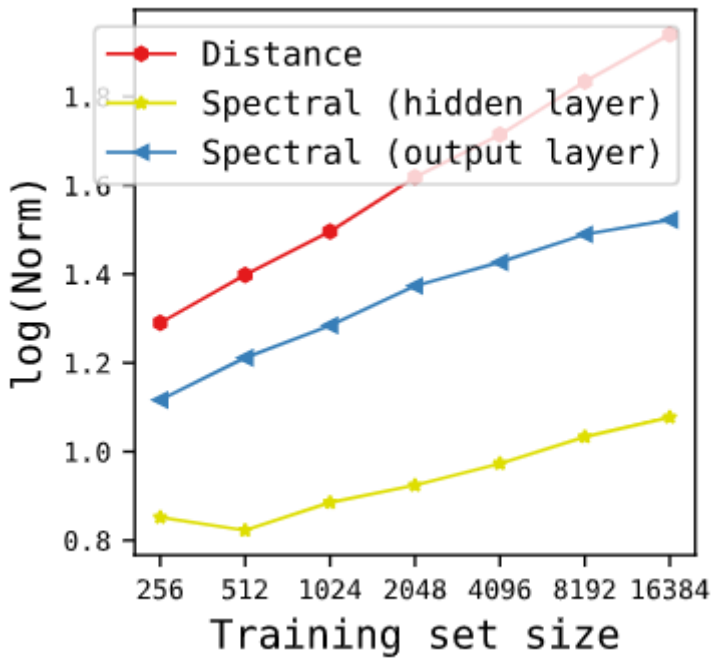
- We train a two-layer ReLU network of 1000 hidden units using SGD with batch size 1 (our observations hold more prominently for smaller batch sizes) and learning rate 0.2 to minimize the cross-entropy loss. Since the cross entropy loss is minimized only when the parameters approach infinity, we need to fix a stopping criterion. Keeping the margin-based bounds in mind, we stop training when a fixed proportion of the training set (99%, to be specific).
- Next, we evaluate the margin-based generalization bounds. In order to do so, we must set the knob γ to some value. We first set it to be the hyperparameter $\gamma_\star$ i.e., 1. Since we stop training when about 0.99 fraction of the data is classified by a margin of $\gamma_\star$, the bound would take the following form:

$$\underbrace{Pr_{(x,y)\sim\mathcal{D}}[f(x)[y] < \max_{y\neq k} f(x)[k]]}_{\text{test 0-1 error}} \leq 0.01 + \frac{\text{some weight norms}}{1\sqrt{m}}$$

CONTINUED: PLOTS







CONTINUED: INSIGHTS

- Simple weight norms like distance from initialization and/or spectral norms have undesirable dependence on the dataset size.
- *All* existing neural network bounds depend on these kinds of weight norms, or some small variations of these. Hence, this means that our concern about the undesirable dataset size dependence is not only limited to the three bounds we experimented with, but all existing bounds in deep learning.
- **Conclusion:** Existing norm-based generalization guarantees (falsely) indicate that, as we add more datapoints to the training set, the network simply memorizes these datapoints. Therefore, to truly understand generalization in deep learning, it is not only important to derive parameter-count-independent generalization guarantees, but also important to derive bounds whose dataset-size-dependence reflects that of the actual generalization error, at least to a reasonable extent.

- $D \rightarrow$ A distribution known to us.
- $S \rightarrow$ From D, training set S of m examples is drawn i.i.d.
  $\therefore S \sim D^m$.
- An algorithm when given S, outputs a hypothesis h_S picked from a class H. h_S has zero error on S. For any h, we let $L_D(h)$ denote test error of h (expected error over D) and $\hat{L}_S(h)$ the average empirical error over the dataset S.
- The generalization error of the above is stated below,

$$\epsilon_{gen} \geq \Pr_{S \sim D^m} \left[\underbrace{L_D(h_S)}_{\text{test error}} - \underbrace{\hat{L}_S(h_S)}_{\text{training error}} \right]$$

$\epsilon_{gen} \leq 1 - \delta \rightarrow$ unlucky dataset draws

EXPERIMENT

Provable failure of uniform convergence

→ Uniform Convergence Bound. (loose upper bound)

$$\Pr_{S\sim D^m}\left[\sup_{h\in H} |L_D(h) - \hat{L}_S(h)| \le \epsilon_{unif}\right] \ge 1-\delta.$$

Note ⇒ For most S, ϵ_{unif} bounds the difference between test and empirical error simultaneously $\forall$ HYPOTHESIS in H.

→ Now, how would one arrive at the TIGHTEST uniform convergence bound?

- A hypothesis class that contains those and only those hypotheses picked by algo. across different training sets. Let's call this H^*.

$$H^* = \{h_S \mid S \sim D^m\}$$

- Improved $\epsilon_{unif-alg.}$:-

$$\Pr_{S\sim D^m}\left[\sup_{h\in H^*} |L_D(h) - \hat{L}_S(h)| \le \epsilon_{unif-alg.}\right] \ge 1-\delta$$

(*) EVEN THIS CLEVER APPROACH FALLS APART.

EXPERIMENT

Provable failure of uniform convergence

CONTINUED

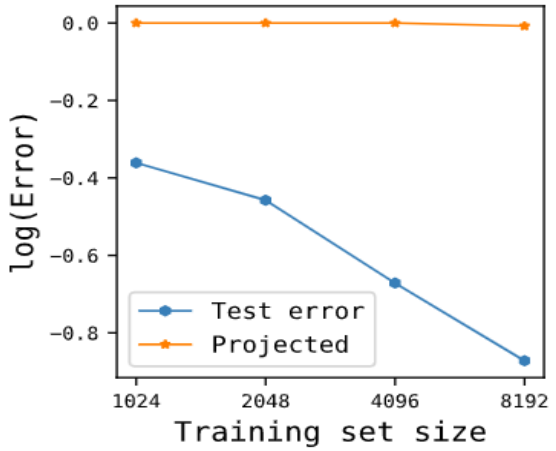
Based on a Lemma and a projected dataset S' the paper defines a Stochastic Gradient Descent based neural network and plots the test error of the network and the error on the projected dataset S′ for varying dataset sizes.

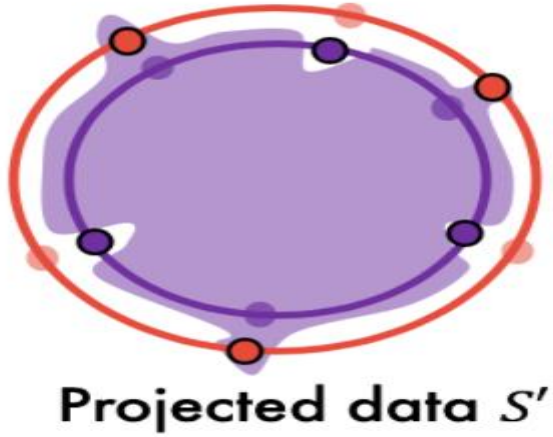
Lemma 1 Consider an algorithm that has zero training error and a test error of ϵ (on almost all draws of the training set). Assume the algorithm is such that, for nearly all draws of S, one can construct another dataset S' of size m (which can be depend on S) such that:

1. h_S misclassifies S' completely.
2. the distribution of S' satisfies $S' \sim \mathcal{D}^m$ (note that S' is a random variable that is a function of S, and here we want the distribution of S', with S marginalized out, to be equal to $\mathcal{D}^m$).

If the algorithm satisfies these requirements, then:

$$\epsilon_{\text{unif-alg}} \geq 1 - \epsilon$$





INSIGHTS

- We observe that even though the test error decreases with dataset size, the error on S' remains constant at (or almost near) 1.00, implying that ϵunif-alg is vacuous. Note that we did not explicitly modify the training in any way to make the network misclassify S'. It just so happens that SGD trains the network in a particular way that leads to misclassification of S'.

- Overparameterized models trained by gradient descent can have certain complexities in their decision boundary; on one hand, these complexities need not hurt generalization error. However, they can severly hurt uniform convergence based bounds. Thus, uniform convergence may not be able to fully explain generalization in deep learning.

REFERENCES

- Original paper and relevant provided: One, Two
- IN SEARCH OF THE REAL INDUCTIVE BIAS : ON THE ROLE OF IMPLICIT REGULARIZATION IN DEEP LEARNING: Behnam Neyshabur, Ryota Tomioka & Nathan Srebro: here
- UNDERSTANDING DEEP LEARNING REQUIRES RETHINKING GENERALIZATION (Zhang et al 2017): here
- UNIFORM CONVERGENCE OF VAPNIK–CHERVONENKIS CLASSES UNDER ERGODIC SAMPLING : here
- A Statistical Distance Derived From The Kolmogorov-Smirnov Test: specification, reference measures (benchmarks) and example uses: here