Generalization in Deep Networks: The Role of Distance from Initialization

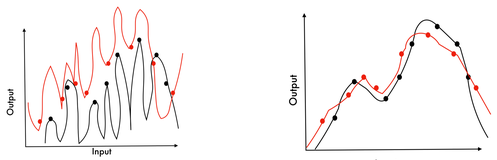
Vaishnavh Nagarajan, J. Zico Kolter

Carnegie Mellon University Computer Science Department

To explain generalization in deep learning, we highlight the need to study the effective model capacity for a *given initialization*, and argue that *distance from initialization* plays a key role in generalization.

Generalization in Deep Learning

How do large deep networks learn simple patterns from real-world data sets even though they can memorize randomly labeled data sets of same size? [1]



What does **not** happen — What does happen

Intuitively: functions learned on two different random draws of training sets (shown in red and black) from the same distribution are "similar" to each other.

Key question: How does stochastic gradient descent (SGD) lead to network-size-independent generalization behavior?

Known Approaches

Active area of research over the last year and different directions of exploration:

I. Solutions lie in regions of flat minima
II. Algorithmic stability: how does algorithm react to change in training datapoint?
III. Implicit regularization offered by SGD: What is it?

[2] investigate different norms of the final network – no norm conclusively explains generalization.

Problem Setting

We focus on networks of d hidden layers (d>1), with H hidden ReLu units in each hidden layer. The weights are initialized with Xavier initialization, and biases initialized to zero.

$\sim \mathcal{N}(0, 1/\sqrt{H})$

Effective Model Capacity for a Given Initialization

It is sufficient to bound the learning-theoretic complexity of:

The set of hypotheses explored on different random draws of *m* training points w.h.p for a given

$$\mathcal{H}_{m,\delta}[\mathcal{D}, (\mathcal{Z}, \mathcal{C}), \mathcal{A}]$$

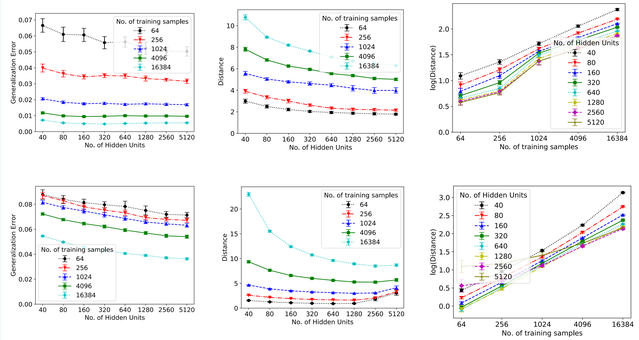
distribution — random initialization — algorithm

Our conjecture: For a fixed initialization, the effective capacity of SGD is contained in an L2 ball around the initialization, with radius independent of H

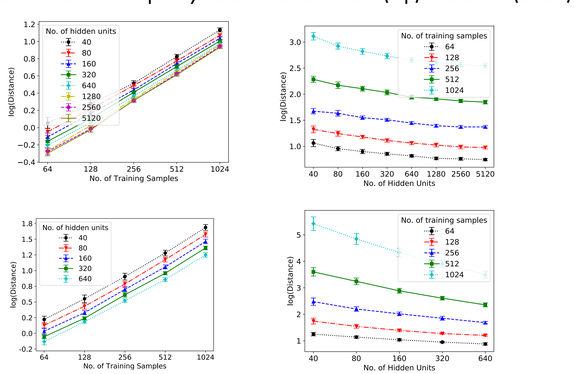
The distance from initialization norm was in fact used in [3] to arrive at non-vacuous generalization bounds.

Empirical Evidence

Minimizing squared error loss (until a fixed threshold) on MNIST (top) and CIFAR (below) demonstrates that distance from initialization is indeed regularized.

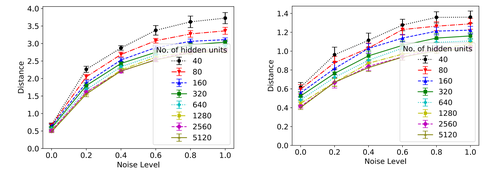


The distance from initialization grows more rapidly with m when minimizing squared error loss on completely random labels. MNIST (top) and CIFAR (below):



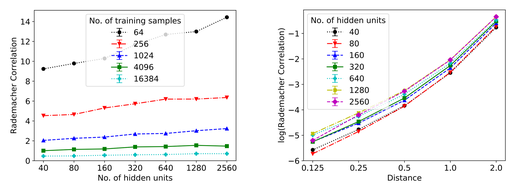
Empirical Evidence (contd.)

Progressively increasing noise in the labels, results in greater distance traveled from initialization to reach the same loss threshold.



We measure *Empirical* Empirical Rademacher Complexity: Sample random label vectors, and use SGD to maximize the correlation with these labels.

The increase with no. of hidden units is mild, and at most logarithmic, less pronounced for larger datasets.



Theoretical Evidence

(Informal) Lemma: An L2 ball of H-independent radius around a Xavier initialization has some nice properties (with high probability):

1. The output of the network can be bounded independent of H
2. Gradients with respect to weights can be bounded independent of H.

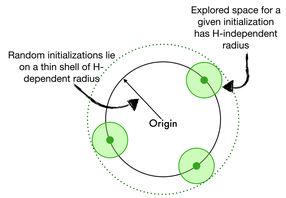
Theorem 1: The Empirical Rademacher Complexity of the L2 ball of radius r for a **linear network** can be bounded independent of number of hidden units H:

$$\mathbb{E}_{\xi}\left[\sup_{(\mathcal{W},\mathcal{B}):\|(\mathcal{Z},\mathcal{C})-(\mathcal{W},\mathcal{B})\|_F\le r}\sum_{i=1}^{m}\xi_i f_{(\mathcal{W},\mathcal{B})}(\mathbf{x}_i)\right] = \tilde{\mathcal{O}}\left(\frac{dc^d(r+1)^{d-1}(r+\sqrt{n})\max_i\|\mathbf{x}_i\|}{\sqrt{m}}\right),$$

Proposition: A system with initial (non-negative) loss L, that takes infinitesimal gradient descent steps at each time instant until the norm of its gradient diminishes below a threshold c, moves a distance of at most L/c i.e., independent of no. of parameters.

On Initialization-Independent Norms

Why can't L2 norm of the weights $\prod_{i=1}^{d}\|\mathbf{W}_k\|_F^2$ [2] explain generalization?



Proposition: Even though the *untrained* network has H-independent generalization error of $\tilde{\mathcal{O}}(1/\sqrt{m})$ its L2 norm grows as $\tilde{\Omega}(H^{d-2})$, so the existing generalization error bounds from [4] are H-dependent.

Some papers, including [2] study the spectral norm $\prod_{k=1}^{d}\|\mathbf{W}_k\|_2$ but if distance-from-initialization is H-independent the, it implies spectral norm is H-independent too.

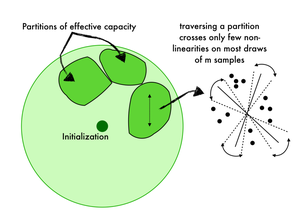
Proposition: $\prod_{k=1}^{d}\|\mathbf{W}_k\|_2 \le \tilde{\mathcal{O}}\left(c^d(1+\|(\mathcal{W},\mathcal{B})-(\mathcal{Z},\mathcal{C})\|_F)^{d-1}(\sqrt{n}+\|(\mathcal{W},\mathcal{B})-(\mathcal{Z},\mathcal{C})\|_F)\right)$

Thus regularization of distance from initialization is more powerful than the regularization of the spectral norm.

Conclusion And Open Questions

The effective capacity of the model for a given initialization is limited by distance from initialization.

A. Is this observation sufficient? Can we extend H-independent generalization error in Theorem 1 to non-linear ReLu networks?



B.Or, a more precise characterization? E.g: effective capacity can be divided into poly(m) continuous subsets such that w.h.p over a draw of m samples, traversing within each parameter subset, no/only few non-linearities is crossed?

References

1. Zhang et al., Understanding deep learning requires rethinking generalization (ICLR 2017)
2. Neyshabur et al., Exploring Generalization in Deep Learning (NIPS 2017)
3. Dziugaite and Roy, Computing Non-vacuous Generalization bounds for Deep Stochastic Neural Networks with Many more parameters than training data (UAI 2017)
4. Neyshabur et al., Norm-based capacity control in neural networks (COLT 2015)