Uniform convergence may be unable to explain generalization in deep learning

David Admète & Benjamin Cohen

Generalization properties in deep neural networks

Generalization of overparametrized networks without regularization ?

- Implicit Bias of GD
- Noise
- Initialization...

Find an upper bound for :

$$\left| \mathscr{L}_{\mathscr{D}}(h) - \hat{\mathscr{L}}_S(h) \right|$$

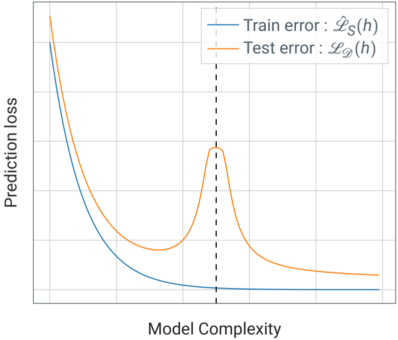


Figure: Double descent phenomenon for a deep neural network without regularization.

From generalization to uniform convergence

Generalization error :

$$\mathscr{L}_{\mathscr{D}}(h_S) - \hat{\mathscr{L}}_S(h_S) \underset{S}{\overset{1-\delta}{\leqslant}} \epsilon_{\text{gen}}(m, \delta).$$

Uniform convergence bound :

$$\sup_h \left| \mathscr{L}_{\mathscr{D}}(h) - \hat{\mathscr{L}}_S(h) \right| \underset{S}{\overset{1-\delta}{\leqslant}} \epsilon_{\text{unif}}(m, \delta).$$

What is a good u-c bound ?

(i) Small and non-vacuous;
(ii) Decrease with increasing width/depth;
(iii) Apply without explicit regularization;
(iv) Increase with memorization;
(v) Decrease with the increasing dataset size (same rate of generalization error).

Table of contents

Notations

Main notations

$(X,y) \sim \mathscr{D}$	sample
$y \in \{-1,+1\}$	
m	dataset size
$S = \left\{(X^{(i)},y^{(i)})\right\}$	dataset
$h \in \mathscr{H}$	hypothesis
h_S	hyp. trained on S

Expected loss :

$$\mathscr{L}_{\mathscr{D}}(h) := \mathbb{E}_{(x,y)\sim\mathscr{D}}[\mathscr{L}(h(x),y)]$$

Empirical loss :

$$\hat{\mathscr{L}}_S(h) := \frac{1}{m}\sum_{i=1}^{m}\mathscr{L}(h(x^{(i)}),y^{(i)})$$

The γ-loss $\mathscr{L} : \mathbb{R} \times \{-1,1\} \to [0,1]$:

$$\mathscr{L}(y',y) = \begin{cases} 1 & \text{if } yy' \leqslant 0 \\ 1 - \dfrac{yy'}{\gamma} & \text{if } yy' \leqslant (0,\gamma) \\ 0 & \text{if } yy' \geqslant 0. \end{cases}$$

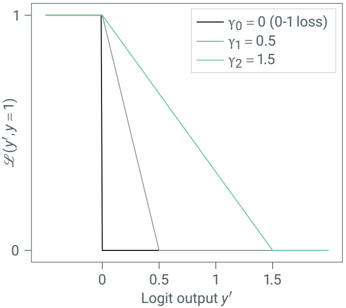


Figure: γ-loss for $\gamma \in \{\gamma_0, \gamma_1, \gamma_2\}$ and $y = 1$.

Generalization error

Definition : Generalization error

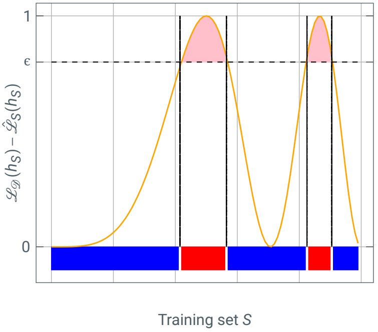
$$\Pr_{S\sim\mathcal{D}^m}\left[\mathcal{L}_{\mathcal{D}}(h_S)-\hat{\mathcal{L}}_S(h_S)\leqslant\epsilon_{\text{gen}}(m,\delta)\right]\geqslant 1-\delta.$$

Reminder

$(X,y)\sim\mathcal{D}$
$S=\left\{(X^{(i)},y^{(i)})\right\}\sim\mathcal{D}^m$

$\mathcal{L}_{\mathcal{D}}(h):=$
$\mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathcal{L}(h(x),y)]$

$\hat{\mathcal{L}}_S(h):=$
$\frac{1}{m}\sum_{i=1}^{m}\mathcal{L}(h(x^{(i)}),y^{(i)})$



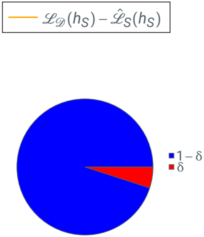


Figure: Illustration of the generalization error.

Disclaimer : The functions shown are of course not continuous in reality.

Uniform convergence bound

Definition : Uniform convergence bound

$$\Pr_{S\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}}\left|\mathcal{L}_{\mathcal{D}}(h)-\hat{\mathcal{L}}_S(h)\right|\leqslant \epsilon_{\text{unif}}(m,\delta)\right]\geqslant 1-\delta.$$

Reminder

$(X,y)\sim\mathcal{D}$
$S=\left\{(X^{(i)},y^{(i)})\right\}\sim\mathcal{D}^m$

$\mathcal{L}_{\mathcal{D}}(h):=$
$\mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathcal{L}(h(x),y)]$

$\hat{\mathcal{L}}_S(h):=$
$\frac{1}{m}\sum_{i=1}^m\mathcal{L}(h(x^{(i)}),y^{(i)})$

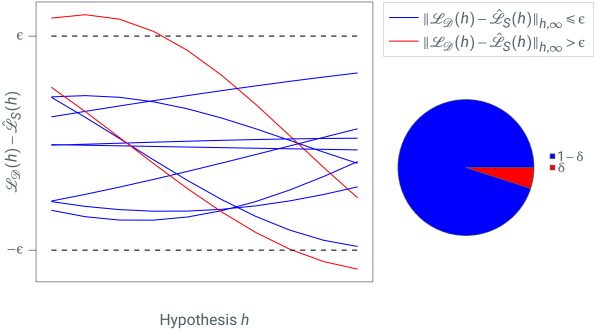


Figure: Illustration of the uniform convergence bound.

Disclaimer : The functions shown are of course not continuous in reality.

A ReLU network

Parameters

Fully connected network with :

Inputs	MNIST
Depth	$d = 5$
Width	$h = 1024$
Optimizer	SGD with rate 0.1
Activations	ReLU
Loss	Crossentropy
Batch size	1

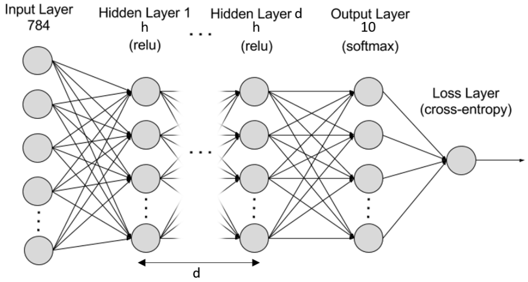
Stop criterion :

- 99% of training data classified correctly
- by a margin of $\gamma = 10$ with

$$\gamma = \max \Gamma(f(x), y)$$

and

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



First issue : need to prune the hypothesis class

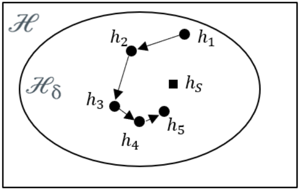
Remark

Uniform convergence bound is inadequate when $\mathcal{H}$ is large.

e.g. : Our fully connected network $|\mathcal{H}| = d \times h$.

<u>Solution :</u> Pruning $\mathcal{H}$.

$$\sup_{h \in \mathcal{H}} \xrightarrow[\text{implicit bias of SGD}]{\text{Taking into account}} \sup_{S \in S_\delta}$$



Definition : Tightest algorithm-dependent uniform convergence bound

Smallest $\epsilon_{\text{u-a}}$ such that :

$$\exists S_\delta, \Pr_{S \sim \mathcal{D}^m}[S \in S_\delta] \geq 1 - \delta \quad \text{and} \quad \epsilon_{\text{u-a}}(m, \delta) \geq \sup_{(S,S') \in S_\delta^2} \left| \mathcal{L}_{\mathcal{D}}(h_{S'}) - \hat{\mathcal{L}}_S(h_{S'}) \right|.$$

e.g. : Consider $\|w\|$ instead of $|\mathcal{H}| = d \times h$.

Bounds growing with m

Weights and bounds

Stop criterion implies :

$$\epsilon_{\text{u-a}} = \mathcal{O}\left(\frac{Bd\sqrt{h}}{\gamma\sqrt{m}} \prod_{k=1}^{d} \|W_k\|_2 \times \text{dist}_i\right)$$

with

$$\left[\begin{array}{l} \text{dist}_F = \sqrt{\sum_{k=1}^{d} \frac{\|W_k - Z_k\|_F^2}{\|W_k\|_2^2}} \\ \text{dist}_{2,1} = \frac{1}{d\sqrt{h}} \left(\sum_{k=1}^{d} \left(\frac{\|W_k - Z_k\|_{2,1}}{\|W_k\|_2}\right)^{2/3}\right)^{3/2} \end{array}\right.$$

Reminder : for $A \in \mathcal{M}_{mn}$,

$$\|A\|_{2,1} = \sum_{i=1}^{m} \sqrt{\sum_{j=1}^{n} a_{ij}^2}.$$

Practical bound

In practice, $\epsilon_{\text{u-a}} = \Omega\left(m^{0.68}\right)$.

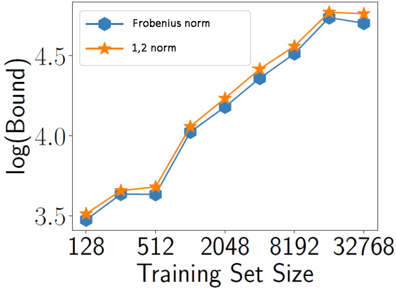


Figure: Tightest bounds from Frobenius norm and $L^{1,2}$ norm, versus the training set size m.

Three examples of overparametrized models

Linear classifier	Wide ReLU network	Infinite width exponential activations network
Input dim $K+D$	Input dim 1000	Input dim $2D$
x_1 deterministic on y x_2 corresponds to noise	Two hyperspheres of radius 1 and 1.1	x_1 deterministic on y x_2 corresponds to noise
Two classes	Two classes	Two classes
m depends on D	$m \in [4k, 65k]$	m depends on D
$h = 1$	$h = 100k$	$h = +\infty$
$d = 1$	$d = 2$	$d = 1$ (only output trainable)
No activation	ReLU activations	Exponential activations
GD step	Acc. 99% with $\gamma = 10$	GD step

Conclusion

Overview

1. Dependence on m of the studied bounds,
2. Three setups where tightest UC bounds become nearly vacuous.

About the reviews

- → Four reviewers highlighted a thorough work.
- → Theoretical results supported by numerical results.
- → Several errors that reviewers did not notice.

Limits and follow-up work

- → No set of tools introduced to replace UC.
- → Focus on setups without explicit regularization.
- → Could work on SRM to limit the parameter norms.

Conclusion

Links and differences with the article

- [Nag21] Identify limits of UC to describe generalization. Empirical technique to get generalization using unlabeled data without UC based complexity.
- [YBM21] UC for nonlinear random feature model. Difference between the test errors and UC bounds for various interpolators.
- [TMY21] Stability derived bounds. Decompose the excess risk to use stability bounds only on the noise part (linear and non-linear models).
- [VPL20] VC bounds → PAC bounds. Marginal Likelihood PAC bound.
- [NDR20] Random class' tight uniform bound. Bound the risk of a predictor using substitutes constructed by conditioning and denoising random predictors.

References I

[BEHW89] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K Warmuth, *Learnability and the vapnik-chervonenkis dimension*, Journal of the ACM (JACM) **36** (1989), no. 4, 929–965.

[BFT17] Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky, *Spectrally-normalized margin bounds for neural networks*, Advances in neural information processing systems **30** (2017).

[Nag21] Vaishnavh Nagarajan, *Explaining generalization in deep learning: progress and fundamental limits*, arXiv preprint arXiv:2110.08922 (2021).

[NBMS17] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro, *Exploring generalization in deep learning*, Advances in neural information processing systems **30** (2017).

[NBS18] Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro, *A pac-bayesian approach to spectrally-normalized margin bounds for neural networks*, 2018.

References II

[NDR20] Jeffrey Negrea, Gintare Karolina Dziugaite, and Daniel Roy, *In defense of uniform convergence: Generalization via derandomization with an application to interpolating predictors*, International Conference on Machine Learning, PMLR, 2020, pp. 7263–7272.

[NK19] Vaishnavh Nagarajan and J Zico Kolter, *Generalization in deep networks: The role of distance from initialization*, arXiv preprint arXiv:1901.01672 (2019).

[SSSSS10] Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan, *Learnability, stability and uniform convergence*, The Journal of Machine Learning Research **11** (2010), 2635–2670.

[TMY21] Jiaye Teng, Jianhao Ma, and Yang Yuan, *Towards understanding generalization via decomposing excess risk dynamics*, 2021.

[VPL20] Guillermo Valle-Pérez and Ard A Louis, *Generalization bounds for deep learning*, arXiv preprint arXiv:2012.04115 (2020).

[YBM21] Zitong Yang, Yu Bai, and Song Mei, *Exact gap between generalization error and uniform convergence in random feature models*, 2021.

Exercise : Uniform convergence and generalization bounds

Setup overview

- Dataset :

$$S = \left(x^{(i)}, y^{(i)}\right)_{i \in [\![1,m]\!]} \in \mathbb{R}^{K+D} \times \{-1, 1\}.$$

with

$$\left| \begin{array}{l} x = (x_1, x_2) \\ x_1 = 2y \cdot u \\ x_2 \sim \mathcal{N}(0, 16/D \cdot I_D). \end{array} \right.$$

- Hypothesis :

$$h(x) = w_1 \cdot x_1 + w_2 \cdot x_2$$

- Learning algorithm :

$$\max_w \left(h_S(x) \cdot y\right)$$

by GD.

Aim

For $\epsilon > 0$, $\delta > 0$, and D large enough,

1. Prove that ϵ_{gen} is upper bounded :

$$\Pr_{S \sim \mathscr{D}^m} \left[\mathscr{L}_{\mathscr{D}}(h_S) - \hat{\mathscr{L}}_S(h_S) \leq \epsilon\right] \geq 1 - \delta$$

2. Prove that $\epsilon_{u\text{-}a}$ is lower bounded :

$$\sup_{(S,S') \in S_\delta^2} \left|\mathscr{L}_{\mathscr{D}}(h_{S'}) - \hat{\mathscr{L}}_S(h_{S'})\right| \geq 1 - \epsilon$$

- Two concentration inequalities admitted,
- Possibility to skip questions by admitting the results,
- Hints provided for the most complex questions,
- From 1h30 to 2h.