Self-Regularity of Non-Negative Output Weights for Overparameterized Two-Layer Neural Networks

Eren C. Kızıldağ (MIT)

Joint work with David Gamarnik (MIT) and Ilias Zadik (NYU)

https://arxiv.org/abs/2103.01887

2021 IEEE International Symposium on Information Theory

Overview

1. Motivation, Prior Work, and Setup
2. Contributions
   - Self-Regularity Results
   - Generalization Guarantees
3. Conclusion and Future Work

Motivation

NN models achieved great practical success:

Image recognition [HZRS16], image classification [KSH12], speech recognition [MDH11], natural language processing [CW08], game playing [SSS+17],...

Overparameterization & Generalization:

- $\#$ Parameters $\gg$ $\#$ Training Data.
- *Conventional wisdom:* Overfit, **poor** generalization.
- **Exact opposite** for NN models: [ZBH+16, BHMM19, ADH+19]...

Why?

Overparameterization & Generalization: Prior Work

Standard **VC Theory** does not help: [HLM17, BHLM19].

Algorithm-independent front:

Norms of weights [NTS15, BFT17, LPRS17, GRS17, DR17, WZ$^+$17], PAC-Bayes theory [NBS17, NBMS17], compression-based bounds [AGNZ18],...

Drawback: Mainly **a posteriori.** Need training to complete.

A priori guarantees: Algorithm-dependent front (soon).

Overparameterization & Generalization: Prior Work

Self-Regularization.

- Many parameter choices (near) **perfectly** interpolating data.
- Algorithms "prefer" *regularized* solutions: *e.g.*, small norm.

Algorithm-specific front: analyze end results.

Gradient Descent [BG17, FCG19], Stochastic Gradient Descent [HRS16, LL18, AZLS19, CG19], Langevin dynamics [MWZZ18],....

Our Work.

Algorithm-independent route.
Well-controlled norm, under a certain non-negativity assumption.
Good generalization, through *fat-shattering dimension*.

Setup and Main Assumption

- Two-layer $\mathtt{NN}$ $(a, W) \in \mathbb{R}^{\overline{m}} \times \mathbb{R}^{\overline{m} \times d}$. j^{th} row of W, $w_j \in \mathbb{R}^d$.
- Width $\overline{m}$ & activation $\sigma(\cdot)$.
- For $X \in \mathbb{R}^d$ computes

$$\sum_{1 \le j \le \overline{m}} a_j \sigma\left(w_j^T X\right), \qquad \sigma \in \{\mathtt{ReLU}, \mathtt{SGM}, \mathtt{Step}\}.$$

- Output weights, $a = (a_j : 1 \le j \le \overline{m})$. **Outer norm**: $\|a\|_1$.

Assumption (Non-Negativity)

$a_j \ge 0$ *for* $j \in [\overline{m}] \triangleq \{1, 2, \ldots, \overline{m}\}$.

Non-Negativity Assumption

Non-negativity of a_j:

- Employed often in literature

  [GLM17, DKKZ20, LMZ20, DL18, SS18, ZYWG19, GKM18],...
- Inherent to **real data** (e.g. audio, muscular activity) [SV17].
- Related to *non-negative matrix factorization (NMF)*.

Non-Negative Matrix Factorization

Given: Non-negative $M \in \mathbb{R}^{n \times m}$ and an $r \in \mathbb{N}$.
Goal: Find non-negative $A \in \mathbb{R}^{n \times r}$, $W \in \mathbb{R}^{r \times m}$ s.t. $\|M - AW\|$ small.

Many applications of NMF:

Info retrieval, document clustering, segmentation, demography, chemometrics,... [AGKM16].

Setup and Distributional Assumptions

Given training data $(X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$, $1 \le i \le N$, find a $\texttt{NN}$ with small **training error**:

$$\widehat{\mathcal{L}}(a, W) \triangleq \frac{1}{N} \sum_{1 \le i \le N} \left(Y_i - \sum_{1 \le j \le m} a_j \sigma(w_j^T X_i) \right)^2 .$$

Run any training algorithm (e.g. GD, SGD, MD).

Assumption (Distributional)

Input/label $(X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$, $1 \le i \le N$, *i.i.d.*

- **Input:** $\exists C > 0$, $\mathbb{P}(\|X\|_2^2 \le Cd) \ge 1 - \exp(-\Theta(d))$.
- **Label:** $\mathbb{E}[|Y|] \triangleq M < \infty$.

X need **not** have **independent** coordinates. Real data have **bounded** labels [DLL+18].

Overview

1. Motivation, Prior Work, and Setup
2. Contributions
   - Self-Regularity Results
   - Generalization Guarantees
3. Conclusion and Future Work

Self-Regularity: ReLU Networks

- Activation: $\texttt{ReLU}(x) = \max\{x, 0\} = (x + |x|)/2$.
- Positive homogenous: $\forall c \ge 0$, $\texttt{ReLU}(cx) = c\texttt{ReLU}(x)$. WLOG, $\|w_j\|_2 = 1$.
- Data (X_i, Y_i), $1 \le i \le N$, i.i.d. with

$$\inf_{w:\|w\|_2=1} \mathbb{E}[\texttt{ReLU}(w^T X)] \ge \boldsymbol{\mu}^* \quad \text{and} \quad \mathbb{E}[|Y|] = M < \infty.$$

- Fix $\delta > 0$ and $\overline{m} \in \mathbb{N}$. Set,

$$\mathcal{G}(\overline{m}, \delta) \triangleq \left\{(a, W) \in \mathbb{R}^{\overline{m}}_{\ge 0} \times \mathbb{R}^{\overline{m} \times d} : \|w_j\|_2 = 1, 1 \le j \le \overline{m};\ \widehat{\mathcal{L}}(a, W) \le \delta^2\right\}.$$

  $\mathcal{G}(\overline{m}, \delta)$: two-layer $\texttt{ReLU NN}$. Width $\overline{m}$ & training error δ^2.
- Set $\mathcal{G}(\delta) \triangleq \bigcup_{\overline{m} \in \mathbb{N}} \mathcal{G}(\overline{m}, \delta)$.

Self-Regularity: ReLU Networks

Theorem (Gamarnik, K., and Zadik, 2021)

We have

$$\sup_{(a,W)\in\mathcal{G}(\delta)} \|a\|_1 \le 4(\delta+2M)(\boldsymbol{\mu}^*)^{-1}.$$

with probability at least $1-\left(12\sqrt{Cd}/\boldsymbol{\mu}^*\right)^d \exp\left(-\Theta(N)\right) - N\exp\left(-\Theta(d)\right) - o_N(1)$.

Suffices to have **near-linear** N: $N = \Theta(d\log d)$.

- For any `ReLU NN` with small $\widehat{\mathcal{L}}(a,W)$ (and $a_j \ge 0$), $\|a\|_1 = O(1)$.
- Oblivious to **training** algorithm.
- Oblivious to **width** $\overline{m}$. Assume *teacher/student* setting:
  - Data (X_i, Y_i) generated by a teacher `NN`.
  - Any student `NN` (potentially overparameterized) has $\|a\|_1 = O(1)$, provided $\widehat{\mathcal{L}}(\cdot)$ is small.

Self-Regularity: Technical Remarks

Probability term $o_N(1)$. Can be made explicit.

- $O(1/N)$: If $\mathbb{E}[Y^2] < \infty$
- $\exp(-\Theta(N))$: If Y_i, $1 \le i \le N$ satisfies large deviations estimates.
- Dropped altogether, if $|Y| \le M$ almost surely.

$\boldsymbol{\mu}^*$ term:

$$\inf_{w:\|w\|_2=1} \mathbb{E}[\mathtt{ReLU}(w^T X)] \ge \boldsymbol{\mu}^*.$$

Suppose $X \stackrel{d}{=} \mathcal{N}(0, I_d)$. Suffices to take $\boldsymbol{\mu}^* = 1/\sqrt{2\pi}$.

Self-Regularity: Sigmoid and Step Networks

- Activations: $\mathtt{SGM}(x) = 1/(1+\exp(-x))$ and $\mathtt{Step}(x) = \mathbb{1}\{x \ge 0\}$.
- Let $\delta, R > 0$ and $\overline{m} \in \mathbb{N}$.
- For $\sigma = \mathtt{SGM}(x)$, define

$$\mathcal{S}\left(\overline{m}, \delta, R\right) = \left\{(a, W) \in \mathbb{R}^{\overline{m}}_{\ge 0} \times \mathbb{R}^{\overline{m} \times d} : \max_{1 \le j \le \overline{m}} \|w_j\|_2 \le R,\ \widehat{\mathcal{L}}\left(a, W\right) \le \delta^2\right\}.$$

- For $\sigma = \mathtt{Step}(x)$, define

$$\mathcal{H}\left(\overline{m}, \delta\right) = \left\{(a, W) \in \mathbb{R}^{\overline{m}}_{\ge 0} \times \mathbb{R}^{\overline{m} \times d} : \|w_j\|_2 = 1, 1 \le j \le \overline{m};\ \widehat{\mathcal{L}}\left(a, W\right) \le \delta^2\right\}.$$

- Set

$$\mathcal{S}(\delta, R) = \bigcup_{\overline{m} \in \mathbb{N}} \mathcal{S}\left(\overline{m}, \delta, R\right) \quad \text{and} \quad \mathcal{H}(\delta) = \bigcup_{\overline{m} \in \mathbb{N}} \mathcal{H}(\overline{m}, \delta).$$

Self-Regularity: Sigmoid and Step Networks

Theorem (Gamarnik, K., and Zadik, 2021)

With high probability, we have

$$\sup_{(a,W)\in\mathcal{S}(\delta,R)} \|a\|_1 \le 3(1+e)(\delta+2M) \quad \textit{and} \quad \sup_{(a,W)\in\mathcal{H}(\delta)} \|a\|_1 \le 2(\delta+2M)\eta^{-1}.$$

Same remarks apply. Additionally,

- `SGM` is not homogeneous: Control parameter R, $\max_j \|w_j\|_2 \le R$.
- $\|a\|_1 = O(1)$, even when $R = \exp(\mathrm{Poly}(d))$ (if $N = \mathrm{poly}(d)$).
- For $X \stackrel{d}{=} \mathcal{N}(0, I_d)$, $\eta = 0.3$ suffices.

Other activations: Softplus ($\ln(1+e^x)$), Gaussian ($\exp(-x^2)$), ...

Generalization Guarantees: Fat-Shattering Dimension

So far: Small $\widehat{\mathcal{L}} \implies$ Controlled $\|a\|_1$ (if $a_i \geq 0$ & $N = \mathrm{Poly}(d)$).

Prior Work [BLW96, Bar98]: Controlled $\|a\|_1 \implies$ Good generalization.

- Through *fat-shattering dimension (FSD)* [KS94]
- A (scale-sensitive) measure of complexity (of model class).

Generalization Guarantees: Fat-Shattering Dimension

Theorem (Bartlett, 1998 [Bar98])

Let $\mathcal{M} > 0$; $\sigma : \mathbb{R} \to [-\mathcal{M}/2, \mathcal{M}/2]$ be non-decreasing. Define sets:

$$F \triangleq \left\{X \mapsto \sigma(w^T X + w_0) : w \in \mathbb{R}^d, w_0 \in \mathbb{R}\right\},$$

$$H(A) \triangleq \{\sum_{j=1}^{\overline{m}} a_j f_j : \overline{m} \in \mathbb{N},\ f_j \in F,\ \|a\|_1 \le A\}$$

where $A \ge 1$. Then for $\gamma \le \mathcal{M}A$,

$$\mathrm{FSD}_{H(A)}(\gamma) \le \widetilde{O}(\mathcal{M}^2 A^2 d/\gamma^2).$$

$H(A)$: two-layer `NN` with **outer norm** at most A.

$\therefore$ Two-layer `NN` with bounded $\|a\|_1$ has "low complexity".

Learning Setting

Data: $\mathcal{D}$ on $\mathbb{R}^d \times \mathbb{R}$. $(X_i, Y_i) \sim \mathcal{D}$, $1 \le i \le N$ i.i.d.

Bounded Y_i**:** $|Y_i| \le M$ almost surely.

Focus: Any $(a, W) \in \mathbb{R}^{\overline{m}}_{\ge 0} \times \mathbb{R}^{\overline{m} \times d}$ with small $\widehat{\mathcal{L}}(\cdot, \cdot)$:

$$\widehat{\mathcal{L}}(a, W) \triangleq \frac{1}{N} \sum_{1 \le i \le N} \left(Y_i - \sum_{1 \le j \le \overline{m}} a_j \sigma(w_j^T X_i) \right)^2 \le \delta^2.$$

Use "learned" (a, W) to **predict** unseen data. Quantified by **Generalization Error:**

$$\mathcal{L}(a, W) \triangleq \mathbb{E}_{(X,Y) \sim \mathcal{D}} \left[\left(Y - \sum_{1 \le j \le \overline{m}} a_j \sigma(w_j^T X) \right)^2 \right].$$

Generalization Guarantee: Main Result

$\mathscr{S}(\delta)$: placeholder for $\mathcal{S}(\delta, R)$ (`SGM` case), $\mathcal{G}(\delta)$ (`ReLU` case), and $\mathcal{H}(\delta)$ (`Step` case).

α: controls *generalization gap* $|\widehat{\mathcal{L}}(a, W) - \mathcal{L}(a, W)|$.

Theorem (Gamarnik, K., and Zadik, 2021)

Let $N = \mathrm{Poly}(d, \alpha^{-1})$. *Then, with high probability over* (X_i, Y_i), $1 \le i \le N$,

$$\sup_{(a,W)\in\mathscr{S}(\delta)} \mathcal{L}(a, W) \le \alpha + \delta^2.$$

Shown by combining our outer norm bounds + [Hau92, BLW96, ABDCBH97, Bar98].

- Complication for `ReLU`: unbounded output. Consider *saturated* version.
- $\texttt{S-ReLU}(x) = \texttt{ReLU}(x)$ for $x \le 1$; and $\texttt{S-ReLU}(x) = 1$ for $x > 1$.

Overview

1. Motivation, Prior Work, and Setup
2. Contributions
   - Self-Regularity Results
   - Generalization Guarantees
3. Conclusion and Future Work

Main Contributions

Two-layer NN with ReLU, SGM, and Step activations.

Assume $a_j \geq 0$.

Self-Regularity:

- $\|a\|_1 = O(1)$ w.h.p. for any (a, W) achieving small $\widehat{\mathcal{L}}(\cdot)$ (on $N = \mathrm{poly}(d)$ data).
- Independent of **width** and **training algorithm**.
- **Mild** data assumption. **Elementary** proof: $\epsilon-$net argument.

Generalization:

- Small $\widehat{\mathcal{L}}(\cdot, \cdot) \implies \|a\|_1 = O(1) \implies$ Good Generalization.

Future Work

- **Different activations.**
- **Non-negativity necessary?:** Yes, strictly speaking.
  - Teacher network, m^* neurons.
  - Student network $\overline{m} \geq m^*$ neurons.
  - Introduce "sign cancellations".
  - Zero training error, but unbounded outer norm.
- **Deeper networks?**

Thank you!

References I

- Noga Alon, Shai Ben-David, Nicolo Cesa-Bianchi, and David Haussler, *Scale-sensitive dimensions, uniform convergence, and learnability*, Journal of the ACM (JACM) **44** (1997), no. 4, 615–631.
- Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Russ R Salakhutdinov, and Ruosong Wang, *On exact computation with an infinitely wide neural net*, Advances in Neural Information Processing Systems, 2019, pp. 8139–8148.
- Sanjeev Arora, Rong Ge, Ravi Kannan, and Ankur Moitra, *Computing a nonnegative matrix factorization—provably*, SIAM Journal on Computing **45** (2016), no. 4, 1582–1611.
- Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang, *Stronger generalization bounds for deep nets via a compression approach*, arXiv preprint arXiv:1802.05296 (2018).

References II

- Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song, *A convergence theory for deep learning via over-parameterization*, International Conference on Machine Learning, PMLR, 2019, pp. 242–252.
- Peter L Bartlett, *The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network*, IEEE transactions on Information Theory **44** (1998), no. 2, 525–536.
- Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky, *Spectrally-normalized margin bounds for neural networks*, Advances in Neural Information Processing Systems, 2017, pp. 6240–6249.
- Alon Brutzkus and Amir Globerson, *Globally optimal gradient descent for a convnet with gaussian inputs*, Proceedings of the 34th International Conference on Machine Learning-Volume 70, JMLR. org, 2017, pp. 605–614.

References III

Peter L Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian, *Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks.*, Journal of Machine Learning Research **20** (2019), no. 63, 1–17.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal, *Reconciling modern machine-learning practice and the classical bias–variance trade-off*, Proceedings of the National Academy of Sciences **116** (2019), no. 32, 15849–15854.

Peter L Bartlett, Philip M Long, and Robert C Williamson, *Fat-shattering and the learnability of real-valued functions*, journal of computer and system sciences **52** (1996), no. 3, 434–452.

Yuan Cao and Quanquan Gu, *Generalization bounds of stochastic gradient descent for wide and deep neural networks*, Advances in Neural Information Processing Systems, 2019, pp. 10836–10846.

References IV

- Ronan Collobert and Jason Weston, *A unified architecture for natural language processing: Deep neural networks with multitask learning*, Proceedings of the 25th international conference on Machine learning, ACM, 2008, pp. 160–167.
- Ilias Diakonikolas, Daniel M Kane, Vasilis Kontonis, and Nikos Zarifis, *Algorithms and sq lower bounds for pac learning one-hidden-layer relu networks*, Conference on Learning Theory, PMLR, 2020, pp. 1514–1539.
- Simon S Du and Jason D Lee, *On the power of over-parametrization in neural networks with quadratic activation*, arXiv preprint arXiv:1803.01206 (2018).
- Simon S Du, Jason D Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai, *Gradient descent finds global minima of deep neural networks*, arXiv preprint arXiv:1811.03804 (2018).

References V

- Gintare Karolina Dziugaite and Daniel M Roy, *Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data*, arXiv preprint arXiv:1703.11008 (2017).
- Spencer Frei, Yuan Cao, and Quanquan Gu, *Algorithm-dependent generalization bounds for overparameterized deep residual networks*, Advances in Neural Information Processing Systems, 2019, pp. 14797–14807.
- Surbhi Goel, Adam Klivans, and Raghu Meka, *Learning one convolutional layer with overlapping patches*, International Conference on Machine Learning, PMLR, 2018, pp. 1783–1791.
- Rong Ge, Jason D Lee, and Tengyu Ma, *Learning one-hidden-layer neural networks with landscape design*, arXiv preprint arXiv:1711.00501 (2017).

References VI

- Noah Golowich, Alexander Rakhlin, and Ohad Shamir, *Size-independent sample complexity of neural networks*, arXiv preprint arXiv:1712.06541 (2017).
- David Haussler, *Decision theoretic generalizations of the pac model for neural net and other learning applications*, Information and computation **100** (1992), no. 1, 78–150.
- Nick Harvey, Christopher Liaw, and Abbas Mehrabian, *Nearly-tight vc-dimension bounds for piecewise linear neural networks*, Conference on Learning Theory, 2017, pp. 1064–1068.
- Moritz Hardt, Ben Recht, and Yoram Singer, *Train faster, generalize better: Stability of stochastic gradient descent*, International Conference on Machine Learning, PMLR, 2016, pp. 1225–1234.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, *Deep residual learning for image recognition*, Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

References VII

- Michael J Kearns and Robert E Schapire, *Efficient distribution-free learning of probabilistic concepts*, Journal of Computer and System Sciences **48** (1994), no. 3, 464–497.
- Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, *Imagenet classification with deep convolutional neural networks*, Advances in neural information processing systems, 2012, pp. 1097–1105.
- Yuanzhi Li and Yingyu Liang, *Learning overparameterized neural networks via stochastic gradient descent on structured data*, Advances in Neural Information Processing Systems, 2018, pp. 8157–8166.
- Yuanzhi Li, Tengyu Ma, and Hongyang R Zhang, *Learning over-parametrized two-layer neural networks beyond ntk*, Conference on Learning Theory, PMLR, 2020, pp. 2613–2682.
- Tengyuan Liang, Tomaso Poggio, Alexander Rakhlin, and James Stokes, *Fisher-rao metric, geometry, and complexity of neural networks*, arXiv preprint arXiv:1711.01530 (2017).

References VIII

Abdel-rahman Mohamed, George E Dahl, and Geoffrey Hinton, *Acoustic modeling using deep belief networks*, IEEE transactions on audio, speech, and language processing **20** (2011), no. 1, 14–22.

Wenlong Mou, Liwei Wang, Xiyu Zhai, and Kai Zheng, *Generalization bounds of sgld for non-convex learning: Two theoretical viewpoints*, Conference on Learning Theory, PMLR, 2018, pp. 605–638.

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro, *Exploring generalization in deep learning*, Advances in Neural Information Processing Systems, 2017, pp. 5947–5956.

Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro, *A pac-bayesian approach to spectrally-normalized margin bounds for neural networks*, arXiv preprint arXiv:1707.09564 (2017).

References IX

- Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro, *Norm-based capacity control in neural networks*, Conference on Learning Theory, 2015, pp. 1376–1401.
- Itay Safran and Ohad Shamir, *Spurious local minima are common in two-layer relu neural networks*, International Conference on Machine Learning, PMLR, 2018, pp. 4433–4441.
- David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al., *Mastering the game of go without human knowledge*, Nature **550** (2017), no. 7676, 354.
- Paris Smaragdis and Shrikant Venkataramani, *A neural network alternative to non-negative audio models*, 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2017, pp. 86–90.
- Lei Wu, Zhanxing Zhu, et al., *Towards understanding generalization of deep learning: Perspective of loss landscapes*, arXiv preprint arXiv:1706.10239 (2017).

References X

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals, *Understanding deep learning requires rethinking generalization*, arXiv preprint arXiv:1611.03530 (2016).

Xiao Zhang, Yaodong Yu, Lingxiao Wang, and Quanquan Gu, *Learning one-hidden-layer relu networks via gradient descent*, The 22nd International Conference on Artificial Intelligence and Statistics, PMLR, 2019, pp. 1524–1534.