On the Lipschitz Constant of Deep Networks and Double Descent

M. Gamba, H. Azizpour, M. Björkman

KTH Royal Institute of Technology
Stockholm, Sweden

Motivation

- Deep networks operate in the interpolating regime

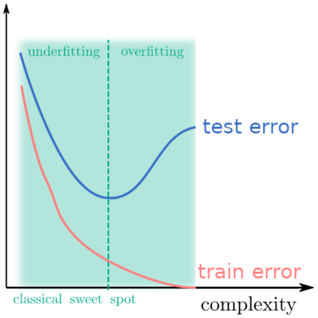


Figure: Berner et al. (2022)

Motivation

- Deep networks operate in the interpolating regime
- Open question: why do they generalize so well?

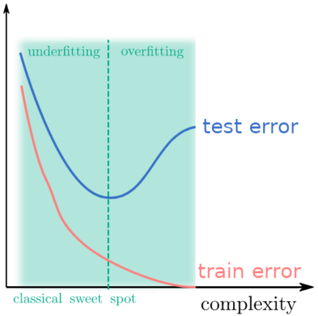


Figure: Berner et al. (2022)

Motivation

- Deep networks operate in the interpolating regime
- Open question: why do they generalize so well?

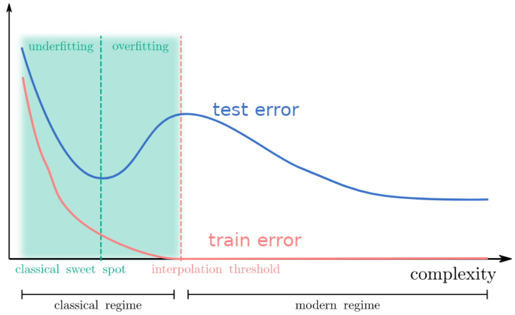


Figure: Berner et al. (2022)

Motivation

- Deep networks operate in the interpolating regime
- Open question: why do they generalize so well?

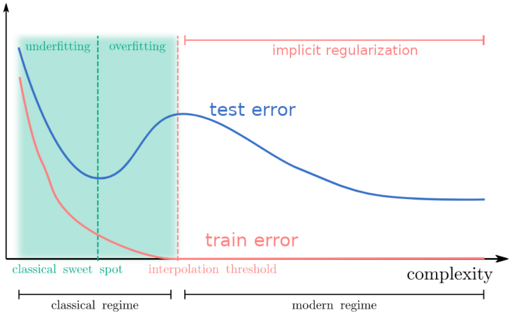


Figure: Berner et al. (2022)

Related works: Singh et al. (2022); Bubeck & Sellke (2021); Ma & Ying (2021); Novak et al. (2018)

Regularity of interpolation

Research question

- Study local geometry of the input/output mapping

$$\mathbf{f}_{\theta} : \mathcal{X} \rightarrow \mathcal{Y}$$

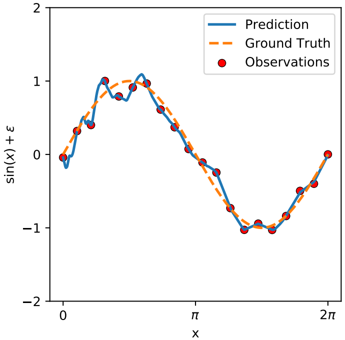


Figure: Gamba et al. (2023b)

Regularity of interpolation

Research question

- Study local geometry of the input/output mapping

$$\mathbf{f}_{\boldsymbol{\theta}} : \mathcal{X} \rightarrow \mathcal{Y}$$

- Study interpolation smoothness through the Jacobian norm

$$\mathbb{E}_{\mathcal{D}} \|\nabla_{\mathbf{x}} \mathbf{f}_{\boldsymbol{\theta}}\|$$

on the training set $\mathcal{D}$

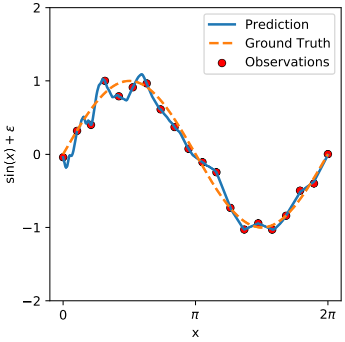


Figure: Gamba et al. (2023b)

Regularity of interpolation

Research question

- Study local geometry of the input/output mapping

$$\mathbf{f}_{\boldsymbol{\theta}} : \mathcal{X} \to \mathcal{Y}$$

- Study interpolation smoothness through the Jacobian norm

$$\mathbb{E}_{\mathcal{D}} \|\nabla_{\mathbf{x}} \mathbf{f}_{\boldsymbol{\theta}}\|$$

  on the training set $\mathcal{D}$

- Hereafter: *empirical Lipschitz constant*

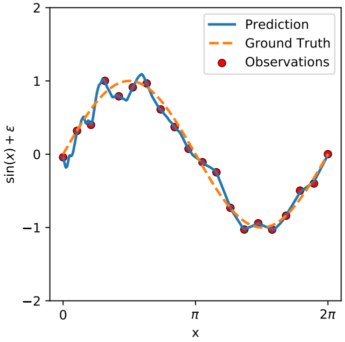
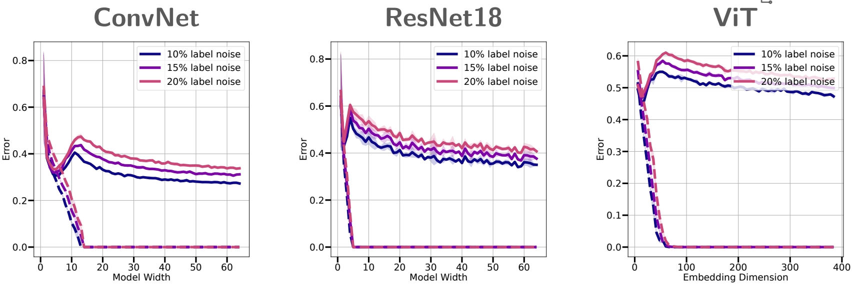
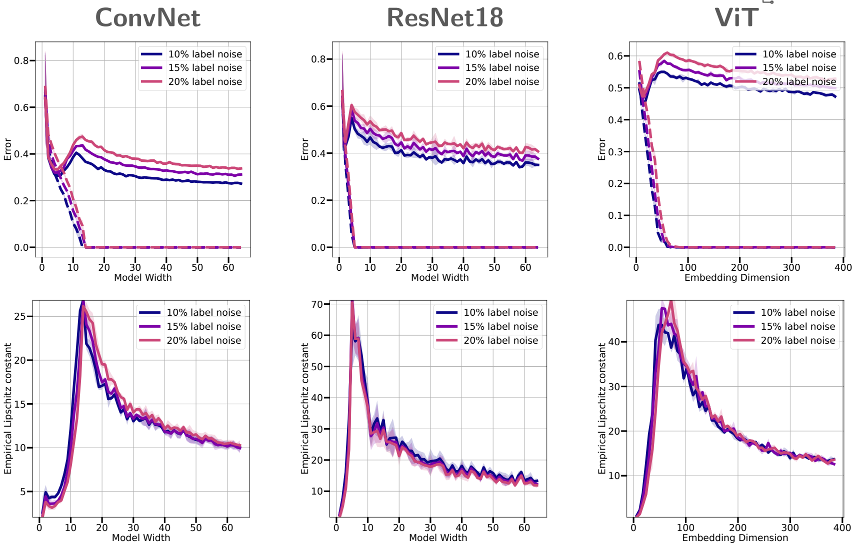


Figure: Gamba et al. (2023b)

Input smoothness mirrors double descent



Input smoothness mirrors double descent



Implicit regularization

Implicit regularization mechanism, at each layer ℓ:

$$\mathbb{E}_{\mathcal{D}}\left\|\frac{\partial \mathbf{f}_{\boldsymbol{\theta}}}{\partial \mathbf{x}_{\ell-1}}\right\|_2 \leq \frac{\|\boldsymbol{\theta}_\ell\|}{h}\mathbb{E}_{\mathcal{D}}\left\|\frac{\partial \mathbf{f}_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}_\ell}\right\|$$

- At each layer, parameter gradient bounds growth of $\|\nabla_{\mathbf{x}}\mathbf{f}_{\boldsymbol{\theta}}\|$

Implicit regularization

Implicit regularization mechanism, at each layer ℓ:

$$\mathbb{E}_{\mathcal{D}}\left\|\frac{\partial \mathbf{f}_{\boldsymbol{\theta}}}{\partial \mathbf{x}_{\ell-1}}\right\|_2 \leq \frac{\|\boldsymbol{\theta}_\ell\|}{h}\mathbb{E}_{\mathcal{D}}\left\|\frac{\partial \mathbf{f}_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}_\ell}\right\|$$

implicit regularization

- At each layer, parameter gradient bounds growth of $\|\nabla_{\mathbf{x}} \mathbf{f}_{\boldsymbol{\theta}}\|$
- *Implicit control on input smoothness* for generalizing networks

Smooth interpolation

Intuition

1. Input Jacobian of ReLU networks

$$\|\nabla_{\mathbf{x}} \mathbf{f}_{\boldsymbol{\theta}}\| = \left\| \prod_{\ell=1}^{L} \boldsymbol{\theta}_{\ell} A_{\ell}(\mathbf{x}) \right\|$$

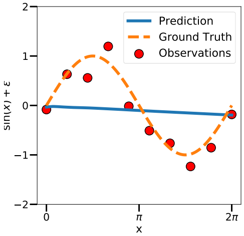
2. Modern weight initialization (He et al., 2015; Glorot & Bengio, 2010):

$$\begin{cases} \theta_i & \sim \mathcal{N}(0, \frac{1}{\alpha}) \quad \alpha \gg 1 \\ b_i & = 0 \end{cases}$$

Smooth interpolation

Intuition

2. Modern weight initialization (He et al., 2015; Glorot & Bengio, 2010):



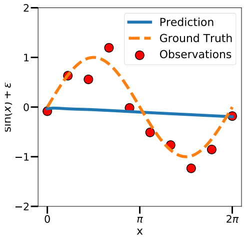
trivially smooth network (with bad generalization)

$$\begin{cases} \theta_i & \sim \mathcal{N}(0, \frac{1}{\alpha}) \quad \alpha \gg 1 \\ b_i & = 0 \end{cases}$$

Smooth interpolation

Intuition

3. At each training step:

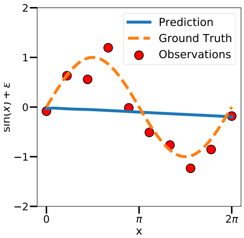


$$\Delta\theta_{ij} \propto \|\nabla_{\theta_{ij}} \mathbf{f}_{\boldsymbol{\theta}}\|$$

Smooth interpolation

Intuition

3. At each training step:



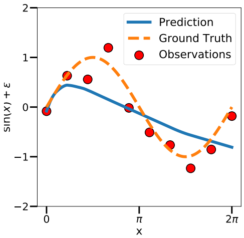
param gradients control smoothness change

$$\mathbb{E}_{\mathcal{D}}\left\|\frac{\partial \mathbf{f}_{\boldsymbol{\theta}}}{\partial \mathbf{x}_{\ell-1}}\right\|_2 \leq \frac{\|\boldsymbol{\theta}_\ell\|}{h}\mathbb{E}_{\mathcal{D}}\left\|\frac{\partial \mathbf{f}_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}_\ell}\right\|$$

Smooth interpolation

Intuition

3. At each training step:



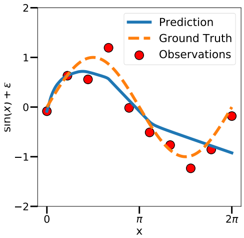
param gradients control smoothness change

$$\mathbb{E}_{\mathcal{D}} \left\| \frac{\partial \mathbf{f}_{\boldsymbol{\theta}}}{\partial \mathbf{x}_{\ell-1}} \right\|_2 \leq \frac{\|\boldsymbol{\theta}_\ell\|}{h} \mathbb{E}_{\mathcal{D}} \left\| \frac{\partial \mathbf{f}_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}_\ell} \right\|$$

Smooth interpolation

Intuition

3. At each training step:



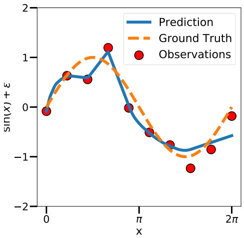
param gradients control smoothness change

$$\mathbb{E}_{\mathcal{D}} \left\| \frac{\partial \mathbf{f}_{\boldsymbol{\theta}}}{\partial \mathbf{x}_{\ell-1}} \right\|_2 \leq \frac{\|\boldsymbol{\theta}_\ell\|}{h} \mathbb{E}_{\mathcal{D}} \left\| \frac{\partial \mathbf{f}_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}_\ell} \right\|$$

Smooth interpolation

Intuition

3. At each training step:



param gradients control smoothness change

$$\mathbb{E}_{\mathcal{D}} \left\| \frac{\partial \mathbf{f}_{\boldsymbol{\theta}}}{\partial \mathbf{x}_{\ell-1}} \right\|_2 \leq \frac{\|\boldsymbol{\theta}_\ell\|}{h} \mathbb{E}_{\mathcal{D}} \left\| \frac{\partial \mathbf{f}_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}_\ell} \right\|$$

Smooth interpolation

Intuition

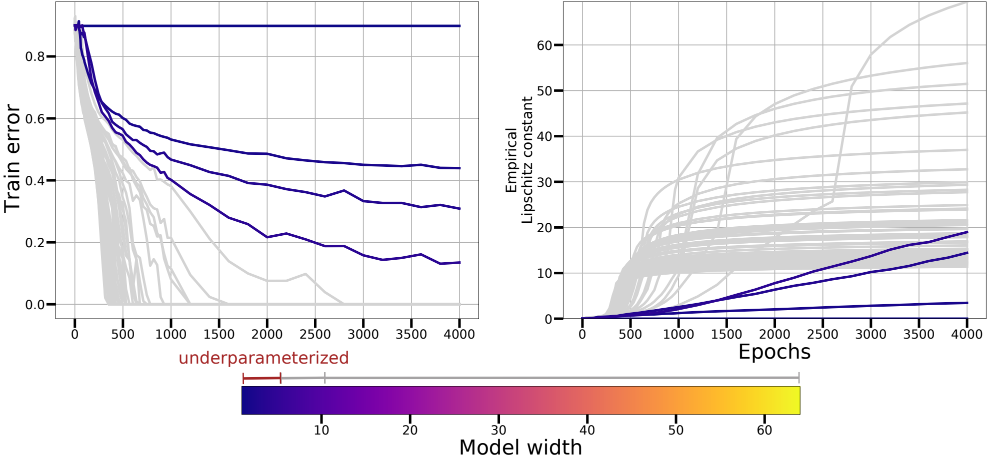
4. Overparameterization:
   - faster interpolation $\rightarrow$ reduced effective complexity

$$\mathbb{E}_{\mathcal{D}} \left\| \frac{\partial \mathbf{f}_{\boldsymbol{\theta}}}{\partial \mathbf{x}_{\ell-1}} \right\|_2 \leq \frac{\|\boldsymbol{\theta}_{\ell}\|}{h} \mathbb{E}_{\mathcal{D}} \left\| \frac{\partial \mathbf{f}_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}_{\ell}} \right\|$$

implicit regularization

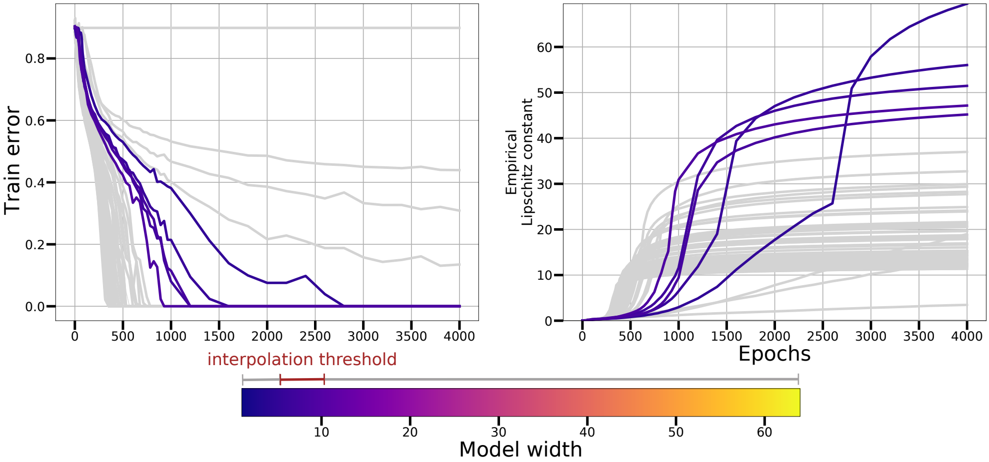
Implications

Implicit acceleration



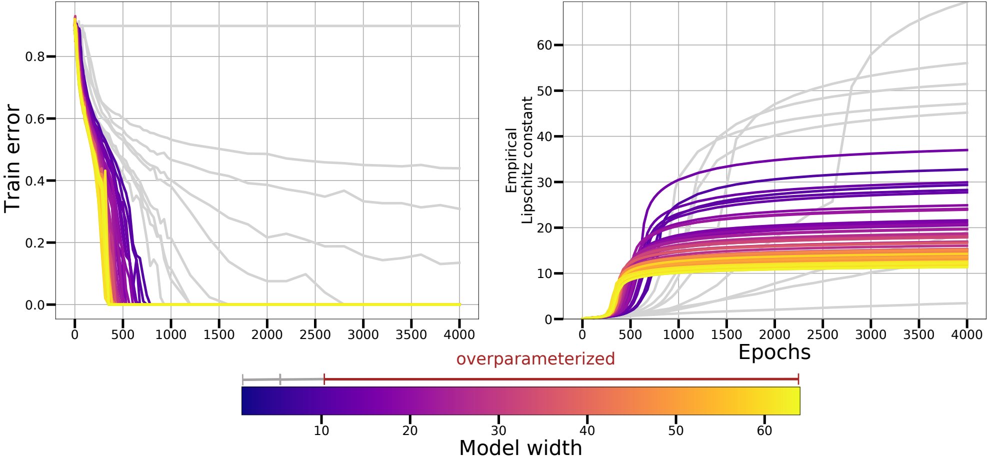
Implications

Implicit acceleration



Implications

Implicit acceleration



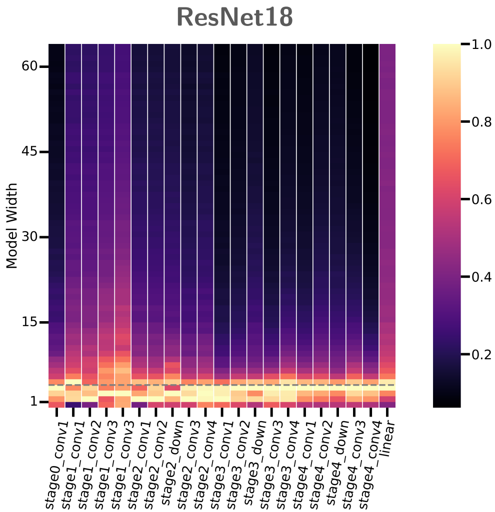
Implications

Reduced complexity

- Distance from initialization:

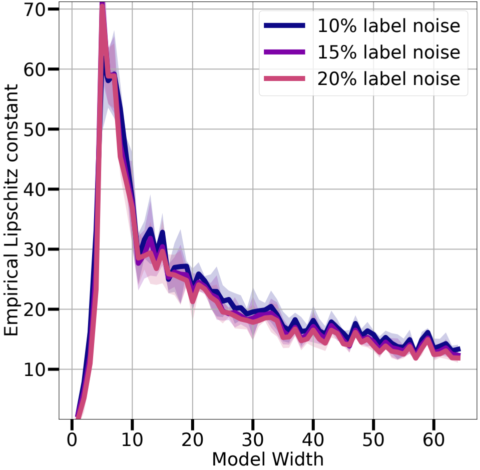
$$d_{\ell,T} = \frac{\|\boldsymbol{\theta}_0^\ell - \boldsymbol{\theta}_T^\ell\|}{\|\boldsymbol{\theta}_0^\ell\|}$$

- Bounded global complexity



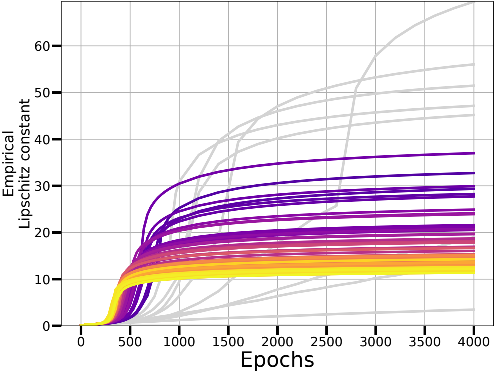
Summary

1. Overparameterization promotes smooth interpolation



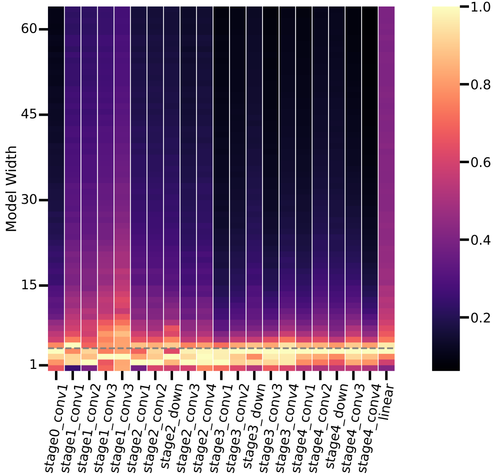
Summary

1. Overparameterization promotes smooth interpolation
2. Overparameterization accelerates interpolation



Summary

1. Overparameterization promotes smooth interpolation
2. Overparameterization accelerates interpolation
3. Overparameterization restricts model complexity



Matteo Gamba

on the job market

Hossein Azizpour

Mårten Björkman

Paper

References I

Julius Berner, Philipp Grohs, Gitta Kutyniok, and Philipp Petersen. *The Modern Mathematics of Deep Learning*, pp. 1–111. Cambridge University Press, 2022.

Sébastien Bubeck and Mark Sellke. A universal law of robustness via isoperimetry. *Advances in Neural Information Processing Systems*, 34, 2021.

Matteo Gamba, Hossein Azizpour, and Mårten Björkman. On the lipschitz constant of deep networks and double descent. In *Procedings of the British Machine Vision Conference 2023*. British Machine Vision Association, 2023a.

Matteo Gamba, Erik Englesson, Mårten Björkman, and Hossein Azizpour. Deep double descent via smooth interpolation. In *Transactions on Machine Learning Research*, 2023b.

Behrooz Ghorbani, Shankar Krishnan, and Ying Xiao. An investigation into neural net optimization via hessian eigenvalue density. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2232–2241. PMLR, 09–15 Jun 2019.

References II

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 249–256. JMLR Workshop and Conference Proceedings, 2010.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1026–1034, 2015.

Matt Jordan and Alexandros G Dimakis. Exactly computing the local lipschitz constant of relu networks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 7344–7353. Curran Associates, Inc., 2020.

Chao Ma and Lexing Ying. On linear stability of sgd and input-smoothness of neural networks. *Advances in Neural Information Processing Systems*, 34: 16805–16817, 2021.

References III

Roman Novak, Yasaman Bahri, Daniel A Abolafia, Jeffrey Pennington, and Jascha Sohl-Dickstein. Sensitivity and generalization in neural networks: an empirical study. In *International Conference on Learning Representations*, 2018.

Sidak Pal Singh, Aurelien Lucchi, Thomas Hofmann, and Bernhard Schölkopf. Phenomenology of double descent in finite-width neural networks. In *International Conference on Learning Representations*, 2022.

Valentin Thomas, Fabian Pedregosa, Bart van Merriënboer, Pierre-Antoine Manzagol, Yoshua Bengio, and Nicolas Le Roux. On the interplay between noise and curvature and its effect on optimization and generalization. In Silvia Chiappa and Roberto Calandra (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 3503–3513. PMLR, 26–28 Aug 2020.

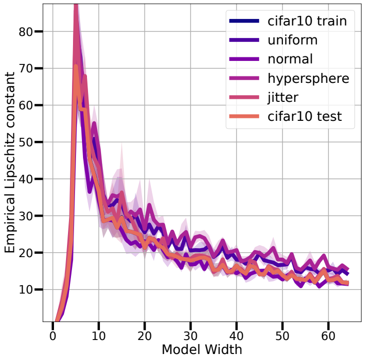
References IV

Aladin Virmaux and Kevin Scaman. Lipschitz regularity of deep neural networks: analysis and efficient estimation. *Advances in Neural Information Processing Systems*, 31, 2018.

Implications

Beyond data manifold

- Probe networks with unseen random data
- Globally bounded smoothness, away from data manifold



Lipschitz continuity

Lipschitz constant:

- $\gamma := \sup_{\mathbf{x} \in \mathcal{X}} \|\nabla_{\mathbf{x}} \mathbf{f}_{\boldsymbol{\theta}}\|$

Lipschitz continuity

Lipschitz constant:

- $\gamma := \sup_{\mathbf{x} \in \mathcal{X}} \|\nabla_{\mathbf{x}} \mathbf{f}_{\boldsymbol{\theta}}\|$
- NP-hard to estimate for neural networks (Jordan & Dimakis, 2020; Virmaux & Scaman, 2018)

Lipschitz continuity

Lipschitz constant:

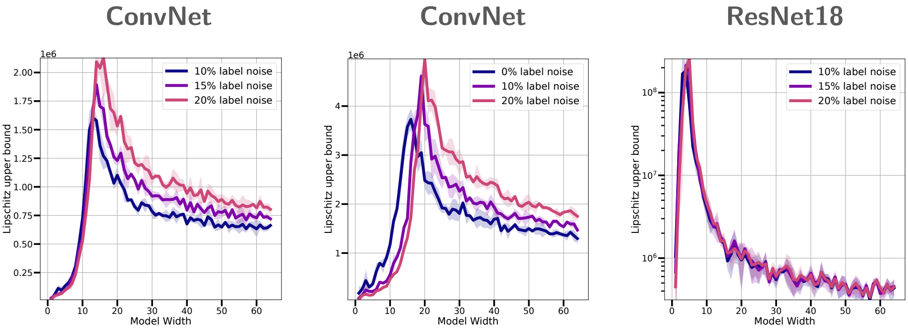
- $\gamma := \sup_{\mathbf{x} \in \mathcal{X}} \|\nabla_{\mathbf{x}} \mathbf{f}_{\boldsymbol{\theta}}\|$
- NP-hard to estimate for neural networks (Jordan & Dimakis, 2020; Virmaux & Scaman, 2018)
- For ReLU networks $\gamma \leq \prod_{\ell=1}^{L} \|\boldsymbol{\theta}_\ell\|_2$

Lipschitz continuity

Lipschitz constant:

- $\gamma := \sup_{\mathbf{x} \in \mathcal{X}} \|\nabla_{\mathbf{x}} \mathbf{f}_{\boldsymbol{\theta}}\|$
- NP-hard to estimate for neural networks (Jordan & Dimakis, 2020; Virmaux & Scaman, 2018)
- For ReLU networks $\gamma \leq \prod_{\ell=1}^{L} \|\boldsymbol{\theta}_\ell\|_2$
- Hence: $\mathbb{E}_{\mathcal{D}} \|\nabla_{\mathbf{x}} \mathbf{f}_{\boldsymbol{\theta}}\|_2 \leq \gamma \leq \prod_{\ell=1}^{L} \|\boldsymbol{\theta}_\ell\|_2$

Lipschitz continuity



Loss landscape and overparameterization

We extend our results to the loss landscape:

$$\frac{h}{\|\boldsymbol{\theta}_1\|}\mathbb{E}_{\mathcal{D}}\|\nabla_{\mathbf{x}}\mathcal{L}(\boldsymbol{\theta}), \mathbf{x}, \boldsymbol{y})\|_2 \leq \mathbb{E}_{\mathcal{D}}\|\nabla_{\boldsymbol{\theta}_1}\mathcal{L}(\boldsymbol{\theta}, \mathbf{x}, y)\|$$
$$\leq \mathcal{L}_{\max}(\boldsymbol{\theta})\Delta\mathcal{L}(\boldsymbol{\theta})$$

Loss landscape and overparameterization

Implications

1. Cross-entropy loss is degenerate at interpolation:

   $\nabla^2_{\mathbf{f}_\theta} \mathcal{L}(\theta, \mathbf{x}, y) = \mathrm{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^T$

Loss landscape and overparameterization

Implications

1. Cross-entropy loss is degenerate at interpolation:
   $\nabla^2_{\mathbf{f}_\theta}\mathcal{L}(\theta, \mathbf{x}, y) = \text{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^T$
2. SGD is aligned with directions of maximum curvature in loss landscape (Thomas et al., 2020; Ghorbani et al., 2019)

Loss landscape and overparameterization

Implications

1. Cross-entropy loss is degenerate at interpolation:
   $\nabla^2_{\mathbf{f}_{\boldsymbol{\theta}}} \mathcal{L}(\boldsymbol{\theta}, \mathbf{x}, y) = \mathrm{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^T$
2. SGD is aligned with directions of maximum curvature in loss landscape (Thomas et al., 2020; Ghorbani et al., 2019)
3. If top directions converge, SGD follows the largest non-zero eigenvalue

Loss landscape and overparameterization

Implications

1. Cross-entropy loss is degenerate at interpolation:
   $\nabla^2_{\mathbf{f}_\theta} \mathcal{L}(\boldsymbol{\theta}, \mathbf{x}, y) = \text{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^T$
2. SGD is aligned with directions of maximum curvature in loss landscape (Thomas et al., 2020; Ghorbani et al., 2019)
3. If top directions converge, SGD follows the largest non-zero eigenvalue
4. Asymptotically, for $\boldsymbol{\theta}_t \to \boldsymbol{\theta}_*$, the smallest non-zero eigenvalue controls regularization