On the Lipschitz Constant of Deep Networks and Double Descent

Matteo Gamba, Hossein Azizpour, Mårten Björkman

BMVC 2023

Motivation

State of the art neural networks operate in the **overparameterized regime**, and are large enough to **interpolate the training set**, at the same time showing remarkable **generalization** performance.

Despite the increased expressivity afforded by overparameterization, the **effective complexity** of neural networks is **constrained in practice**.

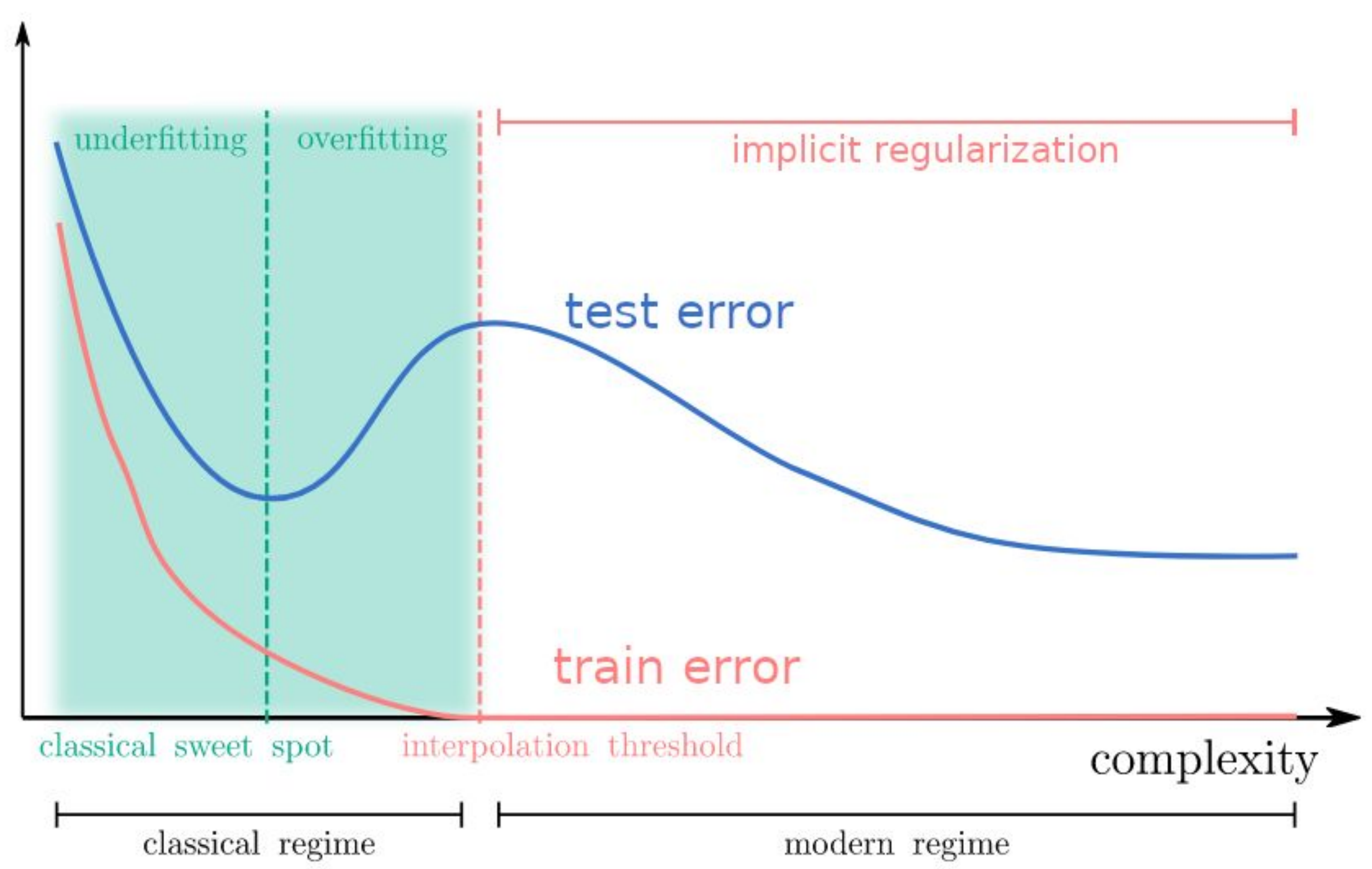
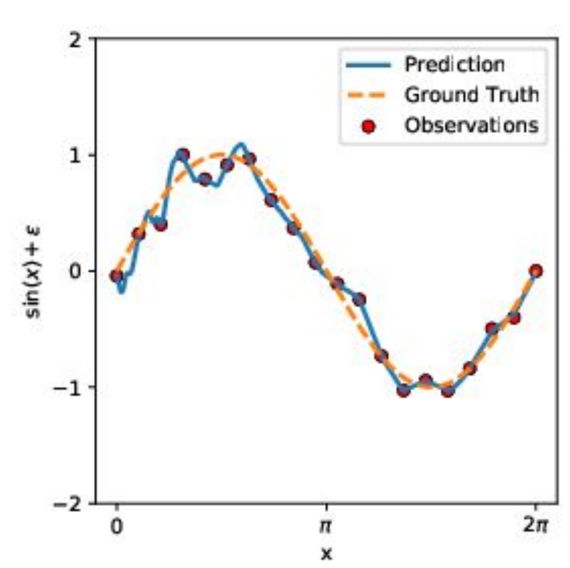


Figure: Berner et al. (2021). Double descent curve of the test error, for networks of increasing model size, interpreted as model complexity.

Research question

We study **effective complexity** of deep networks through the lens of **smooth interpolation of the training data**, to quantify regularity of neural networks trained in practice.

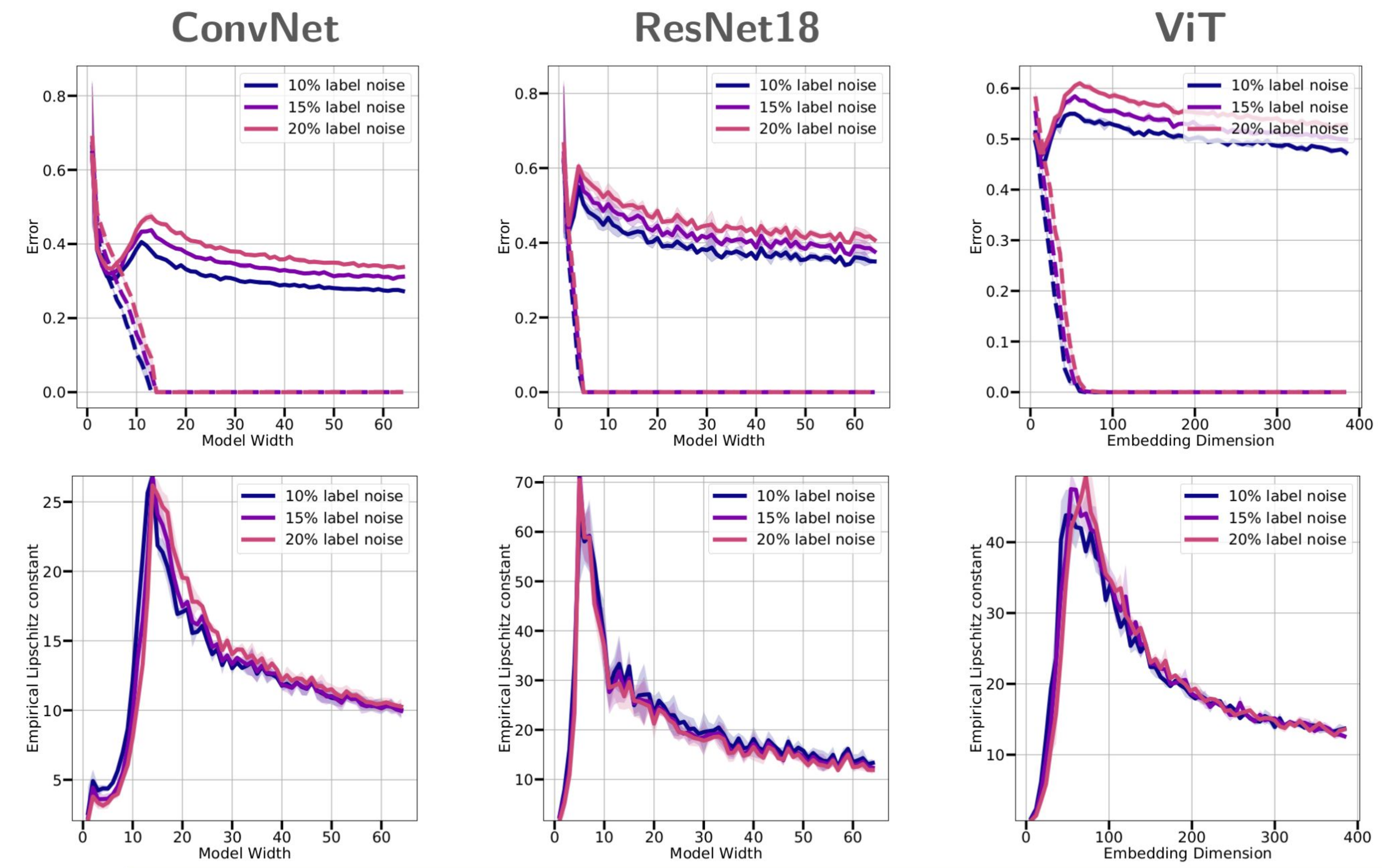


Definitions

For ReLU networks **f** of parameter $\boldsymbol{\theta}$, we quantify smoothness of interpolation via the **input Jacobian norm** of the neural network model function **f**, capturing **local complexity around each training point.**

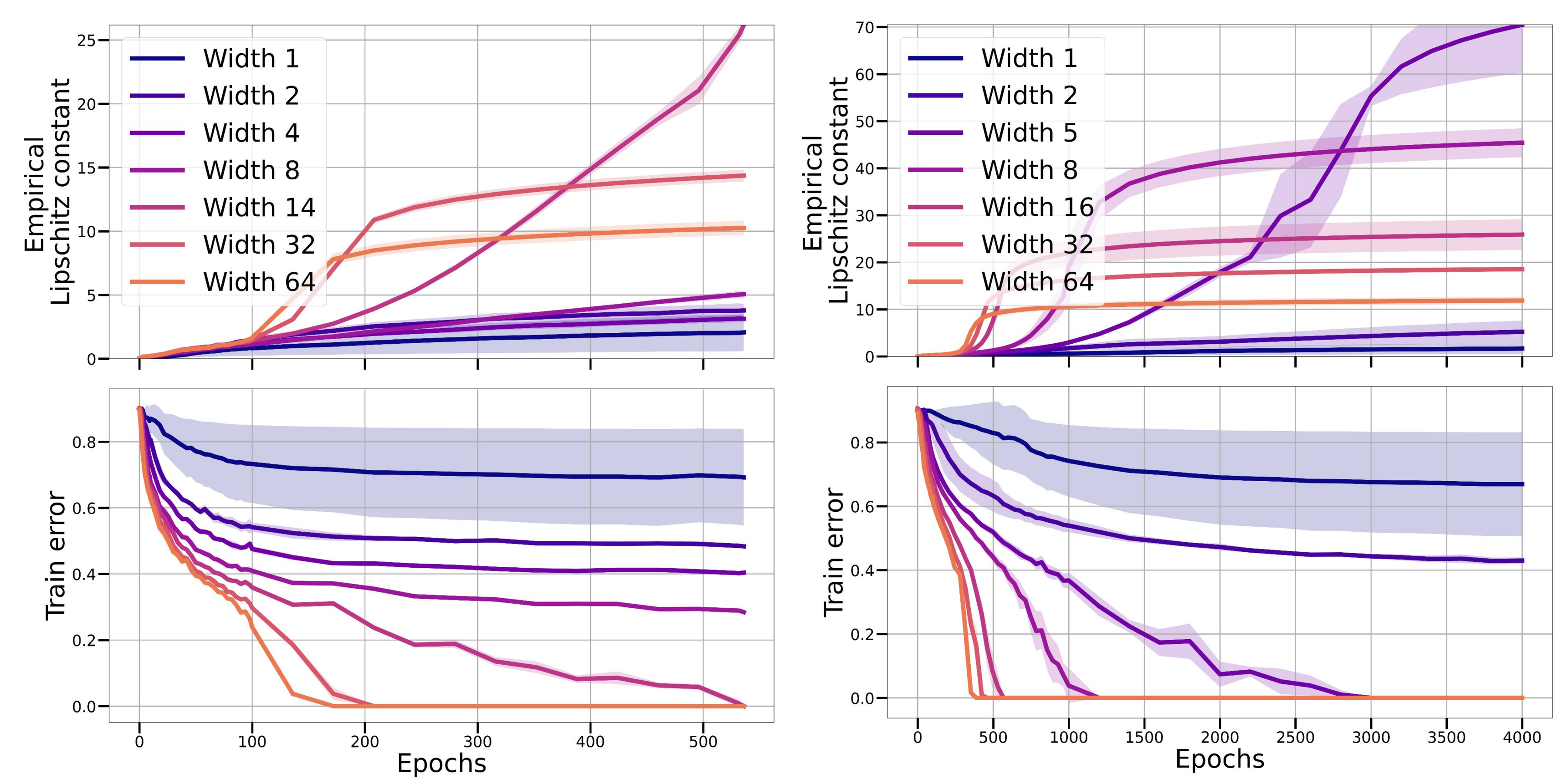
$$(\mathbb{E}_{\mathcal{D}}\|\nabla_{\mathbf{x}}\mathbf{f}_{\boldsymbol{\theta}}\|_2^2)^{\frac{1}{2}} := \left(\frac{1}{N}\sum_{n=1}^{N}\sup_{\mathbf{x}:\|\mathbf{x}\|\neq 0}\frac{\|\boldsymbol{\theta}_{\varepsilon_n}\mathbf{x}\|_2^2}{\|\mathbf{x}\|_2^2}\right)^{\frac{1}{2}}$$

Smooth interpolation and double descent

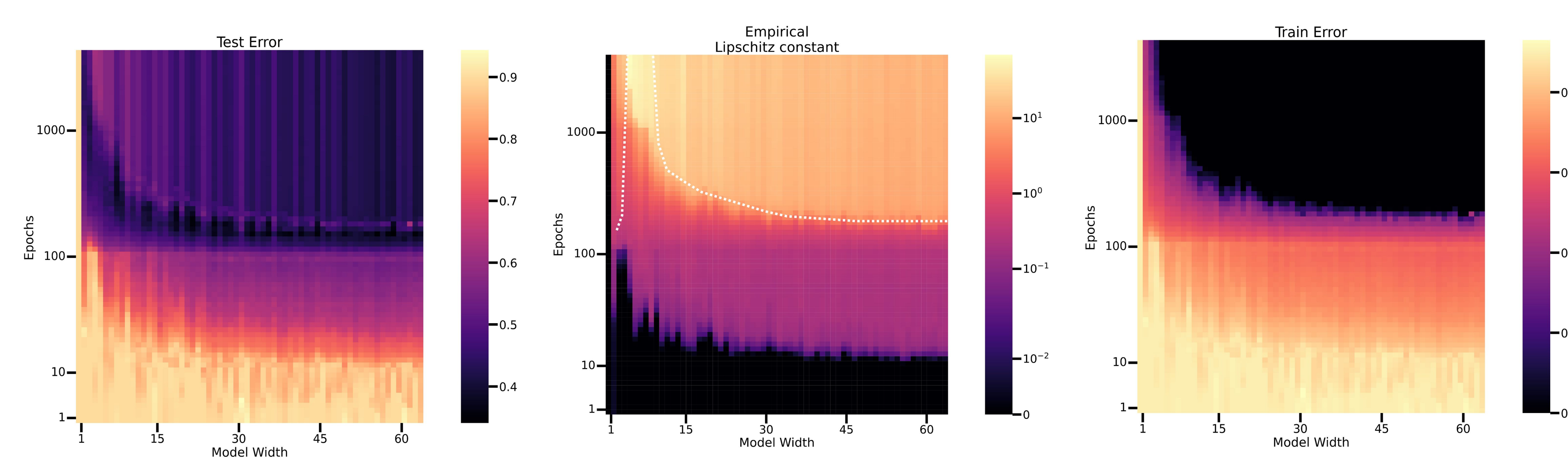


(Top) **Double descent curves for the test error** (solid) and interpolation of training data (dashed). (Bottom) **Input smoothness mirrors double descent** as model size increases.

Overparameterization accelerates interpolation



(Top) **Input smoothness over epochs** for representative models. (Bottom) **Train error** for the same models. In the overparameterized regime, **large models achieve interpolation faster**, thereby retaining low complexity.



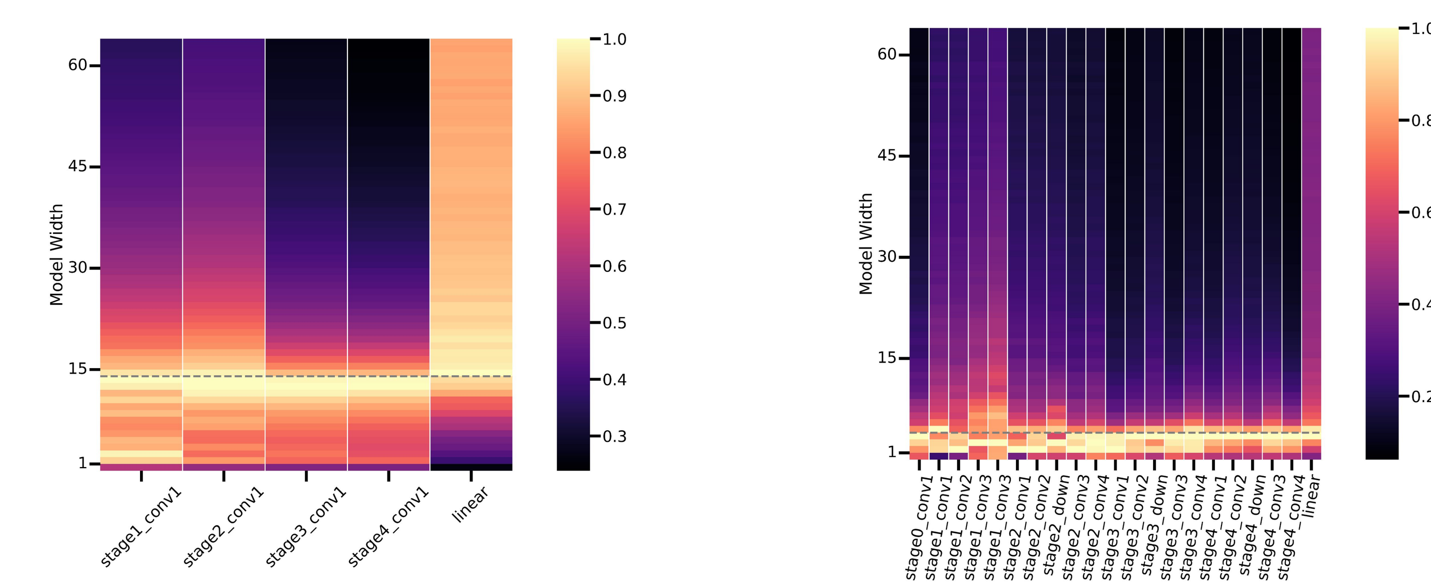
Implicit regularization

Theorem 2. *Let $\boldsymbol{\theta}^*$ be a critical point for the loss $\mathcal{L}(\boldsymbol{\theta},\mathbf{x},y)$ on $\mathcal{D}$. Let $\mathbf{f}_{\boldsymbol{\theta}}$ denote a neural network with at least one hidden layer, with $\|\boldsymbol{\theta}^1\| > 0$. Then,*

$$\frac{x_{\min}^2}{\|\boldsymbol{\theta}^1\|_2^2}\mathbb{E}_{\mathcal{D}}\|\nabla_{\mathbf{x}}\mathcal{L}\|_2^2 \leq 2\mathcal{L}_{\max}(\boldsymbol{\theta})\Delta(\mathcal{L}(\boldsymbol{\theta})) + o(\mathcal{L}(\boldsymbol{\theta}))$$

with $\Delta(\mathcal{L}(\boldsymbol{\theta})) := \mathrm{tr}(H)$ denoting the Laplace operator, $H := \mathbb{E}_{\mathcal{D}}[\frac{\partial^2\mathcal{L}}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}]$ denoting the expected parameter-space Hessian of $\mathcal{L}$, and $\mathcal{L}_{\max}(\boldsymbol{\theta}) := \max_{(\mathbf{x}_n,y_n)\in\mathcal{D}}\mathcal{L}(\boldsymbol{\theta},\mathbf{x}_n,y_n)$.

Globally constrained complexity



The distance from initialization of each layer's parameters mirrors double descent as model size increases, showing globally bounded complexity beyond the training data for large models.

Conclusions

1. Overparameterized networks retain **low complexity by smoothly interpolating** the training data.
2. Parameter-space gradients **implicitly regularize** interpolation smoothness via the input Jacobian for generalizing networks.
3. **Overparameterization accelerates interpolation**, resulting in reduced distance from initialization of each layer.
4. Taken together, the results show that **overparameterization controls complexity globally.**

References

S. Bubeck and M.Sellke (2021). "A universal law of robustness via isoperimetry" In: Advances in Neural Information Processing Systems, 34.

M. Gamba, E. Englesson, M. Björkman, H. Azizpour (2023). "Deep Double Descent via Smooth Interpolation". In: Transactions on Machine Learning Research.

C. Ma and L. Ying. (2021). "On linear stability of sgd and input-smoothness of neural networks." In: Advances in Neural Information Processing Systems 34.

S. P. Singh, A. Lucchi, T. Hofmann, B. Schölkopf (2022). "Phenomenology of double descent in finite-width neural networks". In: International Conference on Learning Representations.