Mildly Overparameterized ReLU Networks on Orthogonal Data: Incremental Learning and Implicit Bias

James Town Etienne Boursier Ben Lewis Matthias Englert Ranko Lazić

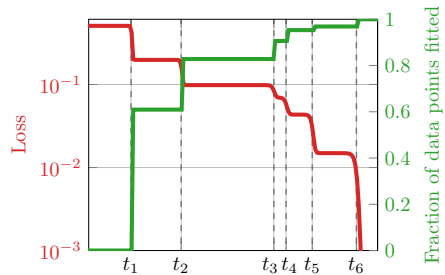
University of Warwick INRIA, Université Paris-Saclay

NeurIPS 2026

Interesting phenomena in the training of neural networks

- Modern networks have **far more parameters than data**, yet gradient methods find solutions that generalise
- **Implicit bias:** training prefers "simple" fits, e.g. of small norm
- **Incremental learning:** plateaus in the loss, then sudden drops, as features are learned one at a time

This work: incremental learning can be the **mechanism** behind a favourable implicit bias



Loss (red) and fraction of data fitted (green) during training

Setting

- Two-layer ReLU network $h_\theta(x) = \sum_{j=1}^m a_j \, \sigma(w_j^\top x)$, trained by **gradient flow**
- **Orthogonal data** $(x_i)_{i=1}^n$, **small initialization** $\theta^\alpha(0)$ of scale $\alpha \to 0$

Boursier, Pillaud-Vivien & Flammarion (NeurIPS 2022)

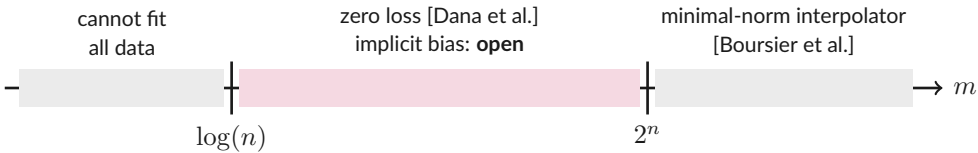
If, at initialization, some neuron is active on **all** positive data, and some neuron on all negative data, then

$$\lim_{t\to\infty} \theta^\alpha(t) = \theta^\alpha_\infty \text{ with } \mathcal{L}(\theta^\alpha_\infty) = 0, \qquad \lim_{\alpha\to 0} \theta^\alpha_\infty = \theta^* \in \operatorname*{arg\,min}_{\mathcal{L}(\theta)=0} \|\theta\|_2^2.$$

- The minimal-norm interpolator has $\frac{1}{2}\|\theta^*\|_2^2 \sim \sqrt{n}$
- But activation patterns are fixed during training, so this needs $m \gtrsim 2^n$ neurons

The gap we fill

- **Dana, Pillaud-Vivien & Bach (NeurIPS 2025):** with $m \gtrsim \log(n)$ neurons, gradient flow still converges to **zero loss**, with high probability
- … but **which** interpolator it finds was left open

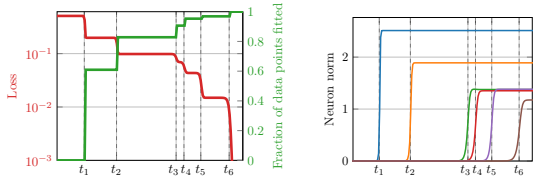


Research problem

In the more realistic mildly overparameterized regime $\log(n) \lesssim m \ll 2^n$, what is the implicit bias of the learned interpolator?

Result 1: saddle-to-saddle dynamics

- As $\alpha \to 0$, gradient flow moves **from saddle to saddle** of the loss
- At each jump, **one new neuron** grows on and fits all the remaining data it is active on
- The whole trajectory, jump times included, is given by a **simple algorithm, without training**
- First rigorous account of incremental learning for a network with a **non-linear activation**



Dashed lines: jump times predicted by the algorithm.

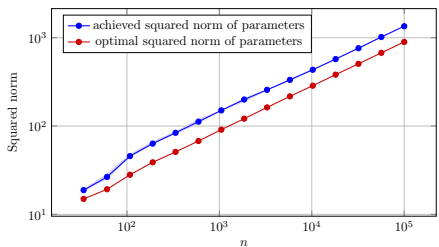
Right: one neuron grows at each drop.

Result 2: near-optimal implicit bias

With high probability over the initialization, the learned network θ^* satisfies

$$\tfrac{1}{2}\|\theta^*\|_2^2 \;\le\; 4\big(\sqrt{n_+}+\sqrt{n_-}\big)\max_i |y_i|$$

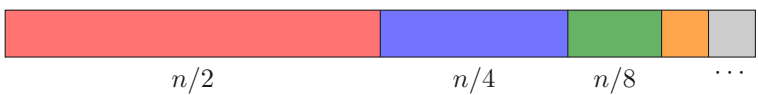
- Same $\sqrt{n}$ scaling as the **minimal-norm interpolator**
- With only $\approx \log(n)$ neurons, instead of $\approx 2^n$



Learned and optimal squared norms grow like $\sqrt{n}$

Why: each neuron takes half of what is left

- Each neuron is active on each data point with probability $\frac{1}{2}$ at initialization
- So each new neuron fits **about half of the remaining data**:



- A neuron fitting k points costs about $\sqrt{k}$, so the norm is a **geometric series**:

$$\tfrac{1}{2}\|\theta^*\|^2 \;\approx\; \sqrt{n/2} + \sqrt{n/4} + \sqrt{n/8} + \cdots \;\sim\; \sqrt{n}$$

Takeaways

- With small initialization, training **activates on one neuron at a time**, and the whole trajectory is predictable without training
- This incremental learning gives a learned network with **near-optimal norm** using only $\approx \log(n)$ neurons
- Incremental learning can provably occur throughout training of neural networks with non-linear activation and lead to a favourable implicit bias

Please do contact us for any discussions about this work!