Implicit bias produces neural scaling laws in learning curves, *from perceptrons to deep networks*

Francesco D'Amico, ***Dario Bocchi,*** ***Matteo Negri.*** *(1) Sapienza Universiy of Rome* *(2) CNR-Nanotec, Rome Unit*

arXiv: 2505.13230

Abstract

Scaling laws are empirical power-law relations between performance and resource growth. In deep learning, they emerge as striking regularities across tasks. They guide modern model design by quantifying the benefits of scaling data and model size, while also informing interpretability. Here, we go beyond convergence behavior and analyze the full training dynamics. We identify two novel dynamical scaling laws describing how performance evolves with norm-based complexity measures. Together, they recover the standard scaling of test error at convergence. These results hold across realistic networks and datasets, and are further supported analytically in a simple theoretical model, explained through the implicit bias of gradient-based training.

Perceptron setting

> Teacher - Student
> $||w^*||^2 = ||w||^2 = \lambda N$
> $P = \alpha N$ training data
> Labels $y^\mu = \mathrm{sign}(x^\mu \cdot w^*)$
> Cross-entropy loss

Deep Networks

> Simple CNN
> ResNet
> Vision Tranformer
> MNIST / CIFAR 10 / 100
> Spectral complexity norm

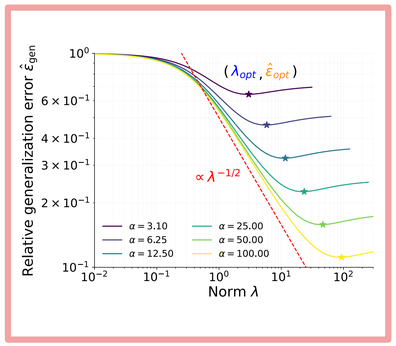
Ablation studies

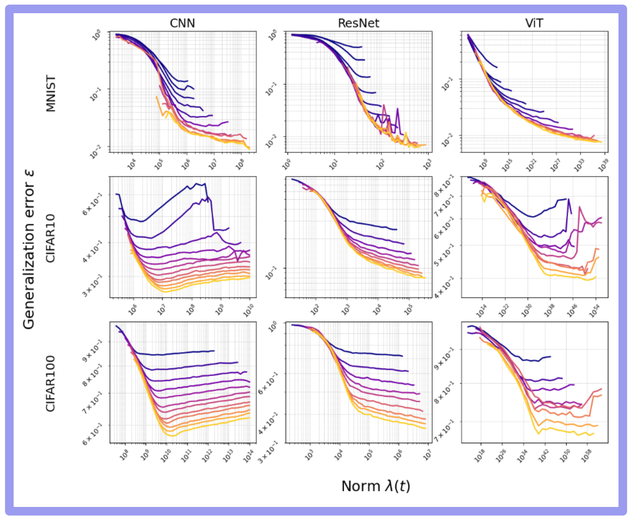
> SGD/Adam
> Weight decay
> Other norms

1 Two new scaling laws in learning curves

Early training $\hat{\epsilon}_{\mathrm{gen}} \sim k_1 \lambda^{-\gamma_1}$

Optima $\lambda_{\mathrm{opt}} \sim k_2 \alpha^{\gamma_2}$

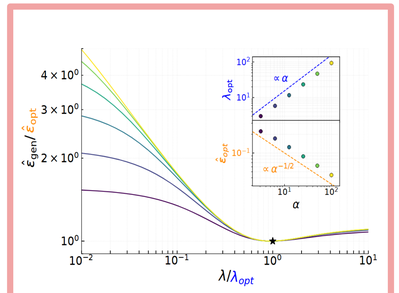


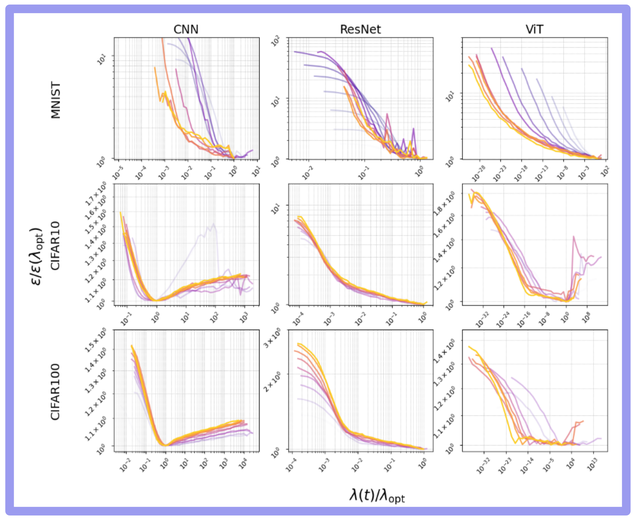


2 Entire learning curves collapse on a master curve

For $\alpha \gg 1$:

$\hat{\epsilon}_{\mathrm{gen}} / \hat{\epsilon}_{\mathrm{opt}} = \Phi(\lambda / \lambda_{\mathrm{opt}})$



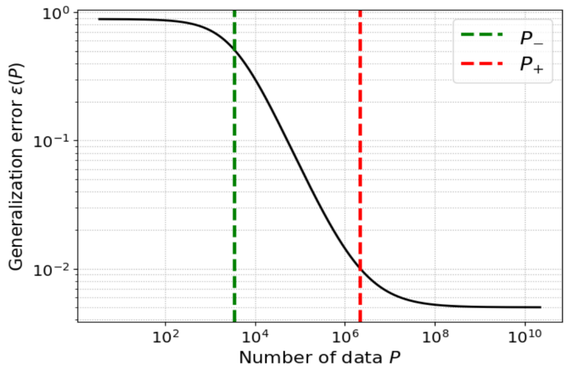


3 This phenomenology explains neural scaling law as combinations of exponents

$$\left.\begin{aligned} \epsilon_{\mathrm{gen}} &= k_1 \lambda^{-\gamma_1} + q_1 \\ \lambda_{\mathrm{opt}} &= k_2 \alpha^{\gamma_2} + q_2 \end{aligned}\right\} \epsilon_{\mathrm{gen}} = k_1 \left(k_2 P^{\gamma_2} + q_2\right)^{-\gamma_1} + q_1$$

with predicted exponent

$$\gamma_{\mathrm{pred}} = \gamma_1 \gamma_2$$



Model	Dataset	γ_{pred}	γ_{meas}	σ
CNN	MNIST	0.60	0.55	0.09
CNN	CIFAR10	0.28	0.25	0.07
CNN	CIFAR100	0.16	0.16	0.03
ResNet	MNIST	0.57	0.69	0.08
ResNet	CIFAR10	0.54	0.56	0.04
ResNet	CIFAR100	0.31	0.37	0.03
ViT	MNIST	0.47	0.54	0.03
ViT	CIFAR10	0.23	0.21	0.03
ViT	CIFAR100	0.14	0.12	0.04

Definitions

Stabilities (or margins)

$$\Delta^\mu \equiv y^\mu \left(\frac{w \cdot x^\mu}{\sqrt{\lambda N}} \right)$$

Normalized pre-activation of prediction from training data

Cross-entropy (pseudo-likelihood) loss

$$L(w) = -\left[\sum_{\mu=1}^{P} \Delta^\mu - \log 2 \cosh\left(\Delta^\mu\right) \right] = \sum_{\mu=1}^{P} V(\Delta^\mu)$$

Spectral complexity norm for a L-layers deep network with matrices A_i

- ρ_i Lipshitz constant of layer i activation function
- $\| \cdot \|_\sigma$ biggest singular value (spectral norm)
- $\| \cdot \|_{2,1}$ sum of ℓ_2 norms of columns
- M_i reference matrix (can be zero)

$$R_A = \underbrace{\left(\prod_{i=1}^{L} \rho_i \, \|A_i\|_\sigma \right)}_{\text{Maximum expansion}} \underbrace{\left(\sum_{i=1}^{L} \frac{\|A_i^\top - M_i^\top\|_{2,1}^{2/3}}{\|A_i\|_\sigma^{2/3}} \right)^{3/2}}_{\text{Effective rank}}$$

Bartlett, P. L., Foster, D. J., & Telgarsky, M. J. (2017). *Spectrally-normalized margin bounds for neural networks.*

TL;DR Implicit bias at training time: the norm acts as an order parameter of the training status