IMPLICIT BIAS PRODUCES NEURAL SCALING LAWS IN LEARNING CURVES, FROM PERCEPTRONS TO DEEP NETWORKS

Francesco D'Amico[1,2*], **Dario Bocchi**[1,2*], **Matteo Negri**[1,2†]
[1]Physics Department, University of Rome Sapienza, Piazzale Aldo Moro 5, Rome 00185
[2]CNR - Nanotec, Rome Unit, P.le Aldo Moro 5, 00185 Rome, Italy
{francesco.damico,dario.bocchi,matteo.negri}@uniroma1.it

Francesco D'Amico

SAPIENZA
UNIVERSITÀ DI ROMA

Consiglio Nazionale
delle Ricerche

Models considered:

Analytical results in perceptron, validated in deep classifiers with image data

Focus on:

Cross-entropy implicit bias at training time

Result (1):

Norm of weights acts as an order parameter of training status

Result (2):

In $\epsilon_{\text{test}} \sim P^{-\gamma}$, we decompose $\gamma = \gamma_1 \gamma_2$,
with γ_1, γ_2 depending on spectra of weights matrices

Analytical setting

Teacher $w^* \in \mathbb{R}^N$, student $w \in \mathbb{R}^N$, spherical weights $||w^*||^2 = ||w||^2 = \lambda N$

$P = \alpha N$ random data $x^\mu \in \{\pm 1\}^N$

Norm of the weights, fixed

Labels $y^\mu = \text{sign}(x^\mu \cdot w^*)$

Cross-entropy (pseudo-likelihood) loss:

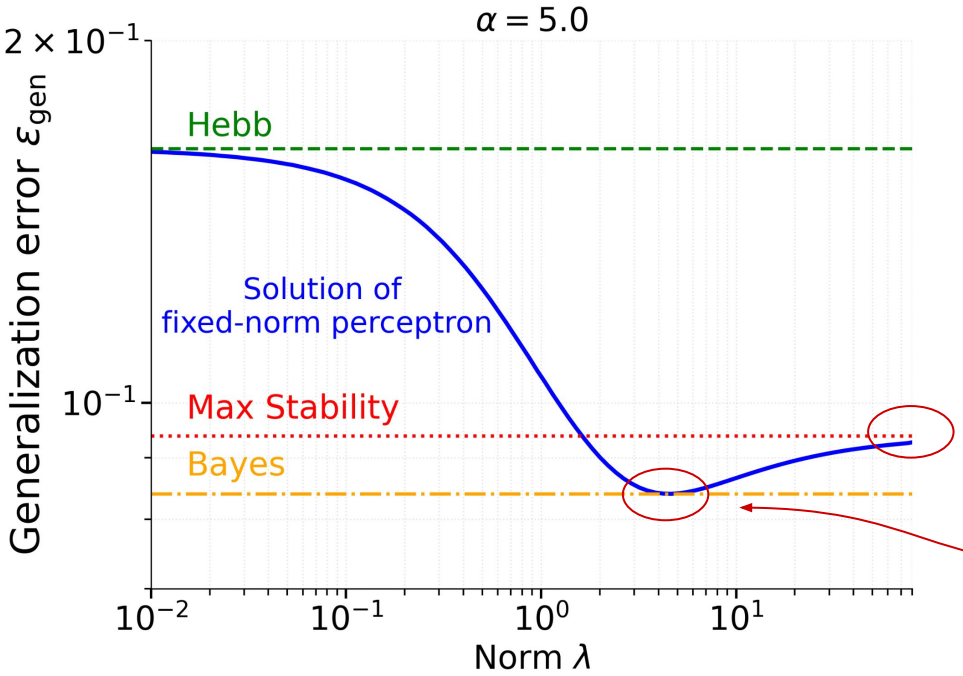
$$L(w) = -\left[\sum_{\mu=1}^{P} \Delta^\mu - \log 2\cosh\left(\Delta^\mu\right)\right] = \sum_{\mu=1}^{P} V(\Delta^\mu)$$

where 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

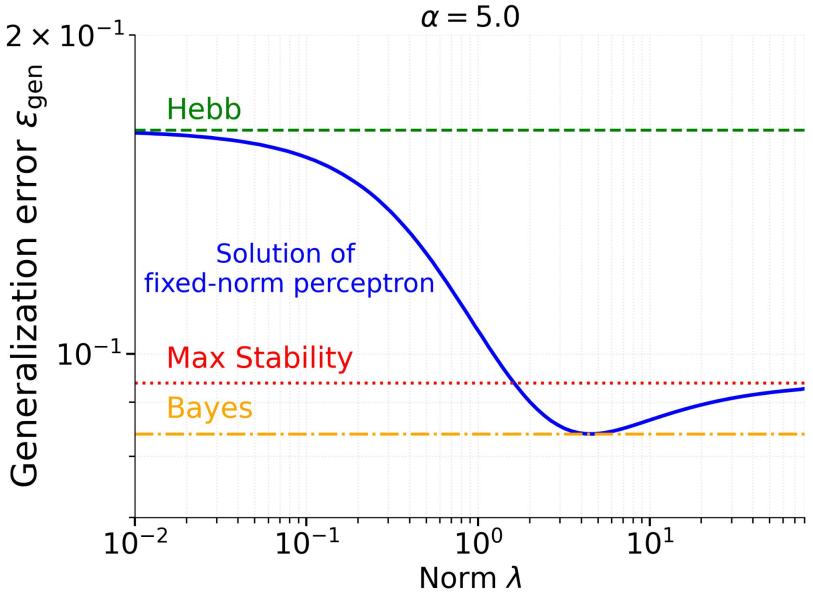
This setting is not new, previous results:



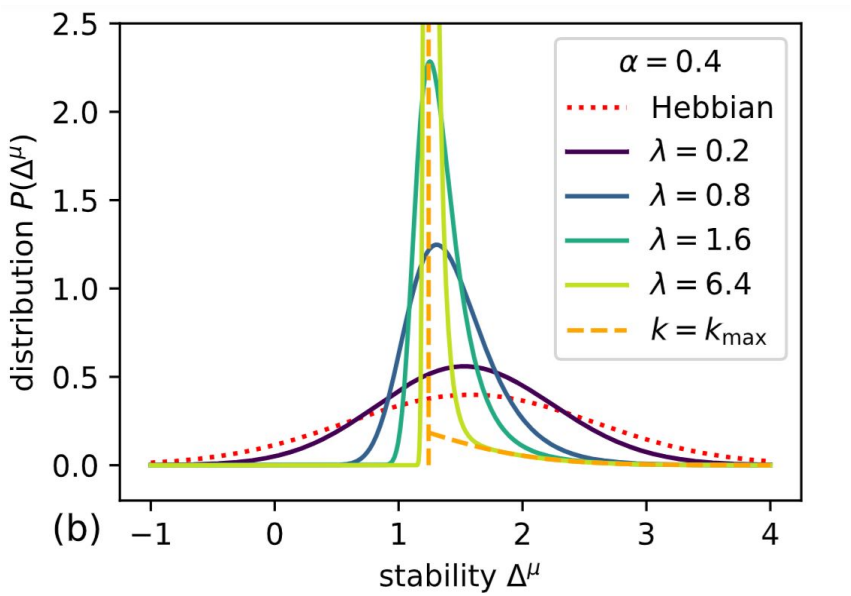
Soudry, D., Hoffer, E., Nacson, M. S., Gunasekar, S., & Srebro, N. (2018)
The implicit bias of gradient descent on separable data

Aubin, B., Krzakala, F., Lu, Y., & Zdeborová, L. (2020).
Generalization error in high-dimensional perceptrons: Approaching bayes error with convex optimization.

This setting, Teacher-Student



Random data and labels

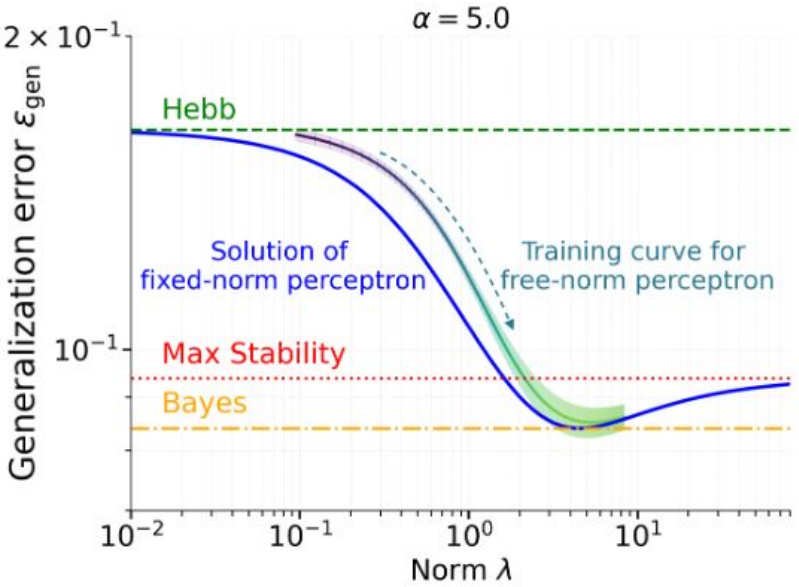


D'Amico, F., Bocchi, D., Del Bono, L. M., Rossi, S., & Negri, M. (2025).
Pseudo-likelihood produces associative memories able to generalize, even for asymmetric couplings.

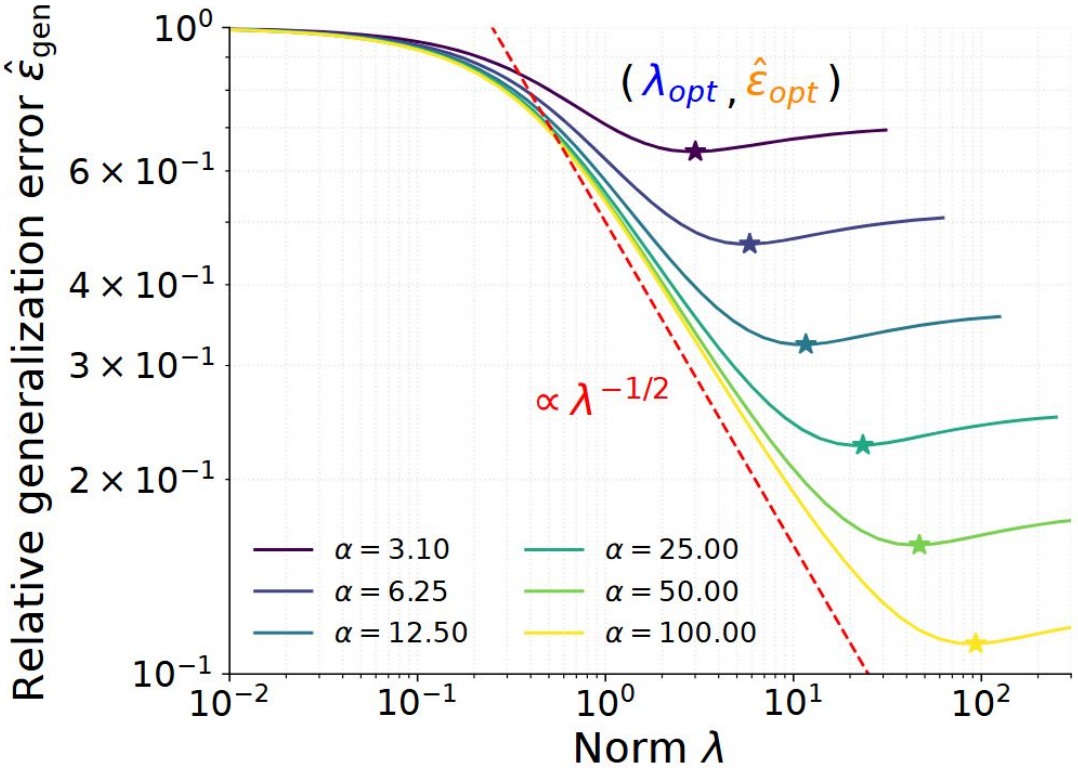
Starting observation

With unbounded norm and GD optimization, norm $\lambda(t)$ increases monotonically (Soudry et al., 2018)

Resulting $\epsilon(\lambda(t))$ curves approximated by fixed-norm $\epsilon(\lambda)$ curves



Result (1): two new scaling laws in norm λ



(1) Early training $(\lambda < \lambda_{elbow}(\alpha))$

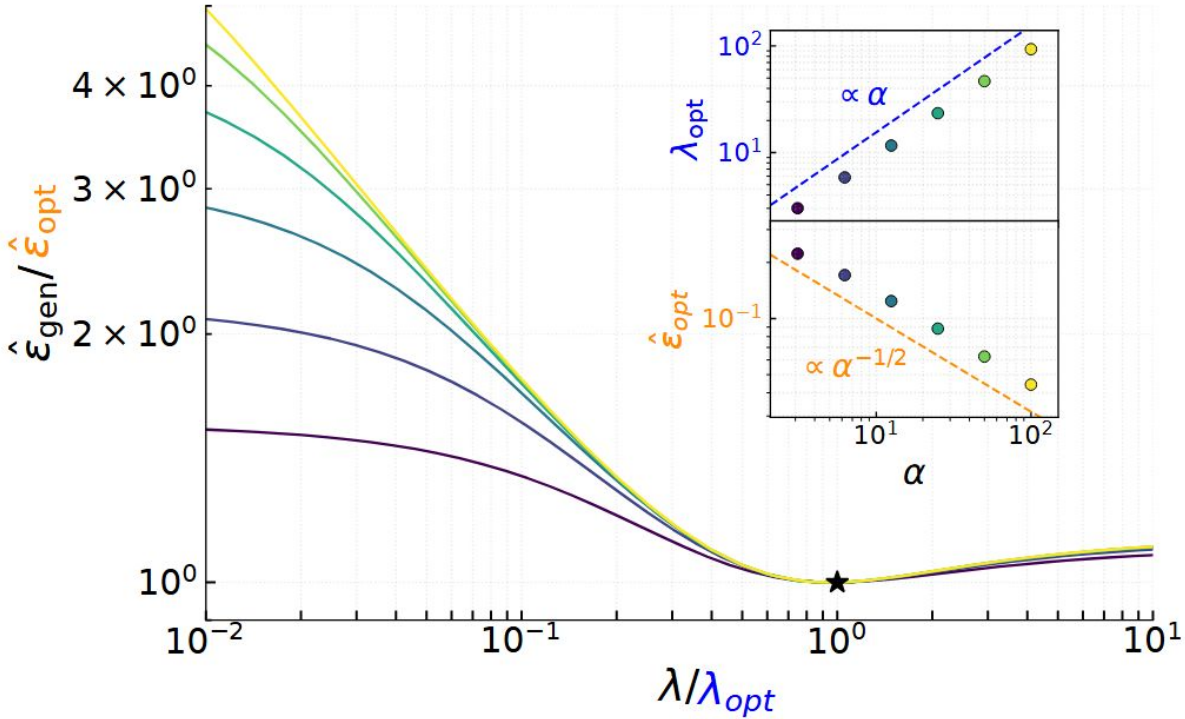
$$\hat{\epsilon}_{\text{gen}} \sim k_1 \lambda^{-\gamma_1}$$

for $\alpha \gg 1$

(2) Optima of curves

$$\lambda_{\text{opt}} \sim k_2 \alpha^{\gamma_2}$$

Result (2): collapse on a master curve Φ for $\alpha \gg 1$



Result (3):

$$\hat{\epsilon}_{\text{gen}} \sim k_1 \lambda^{-\gamma_1} \quad \text{for} \quad \lambda < \lambda_{elbow}(\alpha)$$

$$\lambda_{\text{opt}} \sim k_2 \alpha^{\gamma_2} \quad \text{for} \quad \lambda > \lambda_{elbow}(\alpha)$$

$$\hat{\epsilon}_{\text{gen}} / \hat{\epsilon}_{\text{opt}} = \Phi(\lambda / \lambda_{\text{opt}}) \quad \text{for} \quad \alpha \gg 1$$

$$\hat{\epsilon}_{\text{gen}} \sim k_1 \, k_2^{-\gamma_1} \, \alpha^{-\gamma_1 \gamma_2} \quad \text{for} \quad \alpha \gg 1$$

Meaning $\hat{\epsilon}_{\text{gen}} \sim \alpha^{-\gamma}$

with $\boxed{\gamma = \gamma_1 \gamma_2}$

Deep classifiers setup

Architectures

- CNN
- ResNet
- ViT

Image datasets

- MNIST
- CIFAR10
- CIFAR100

Norm definition for deep networks

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

Definitions:

- 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)

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

Maximum expansion

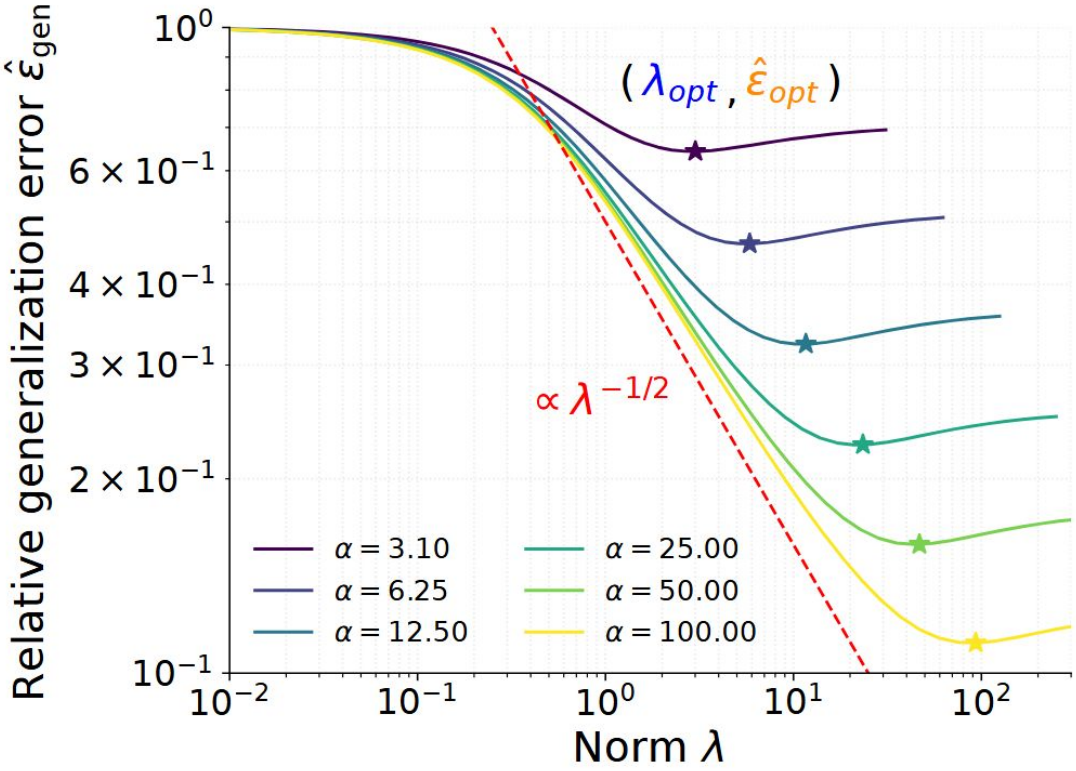
Effective rank

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

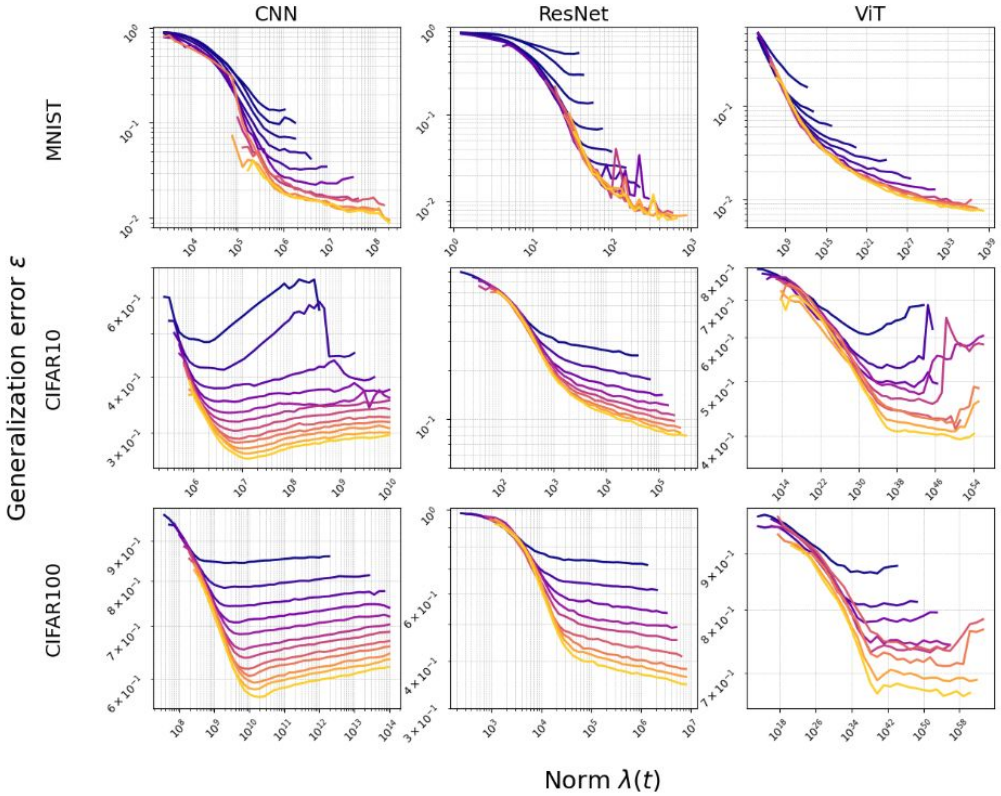
Experimental procedure

- Number of data P in the order $\mathcal{O}(10^2 - 10^4)$
- Only classifiers capable of zero training error with maximal training set
- For each architecture - dataset couple, only one set of hyperparameters (learning rate, initial variance of weights, batch size)
- No weight decay or regularization

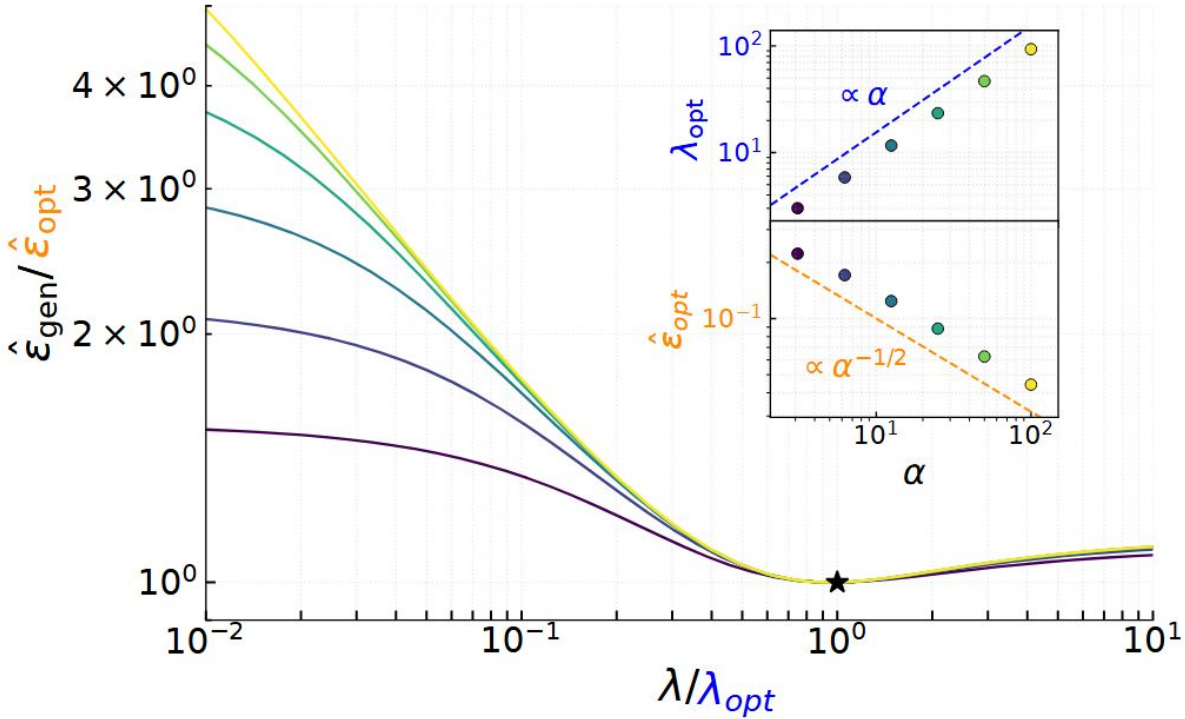
Result (1): perceptron



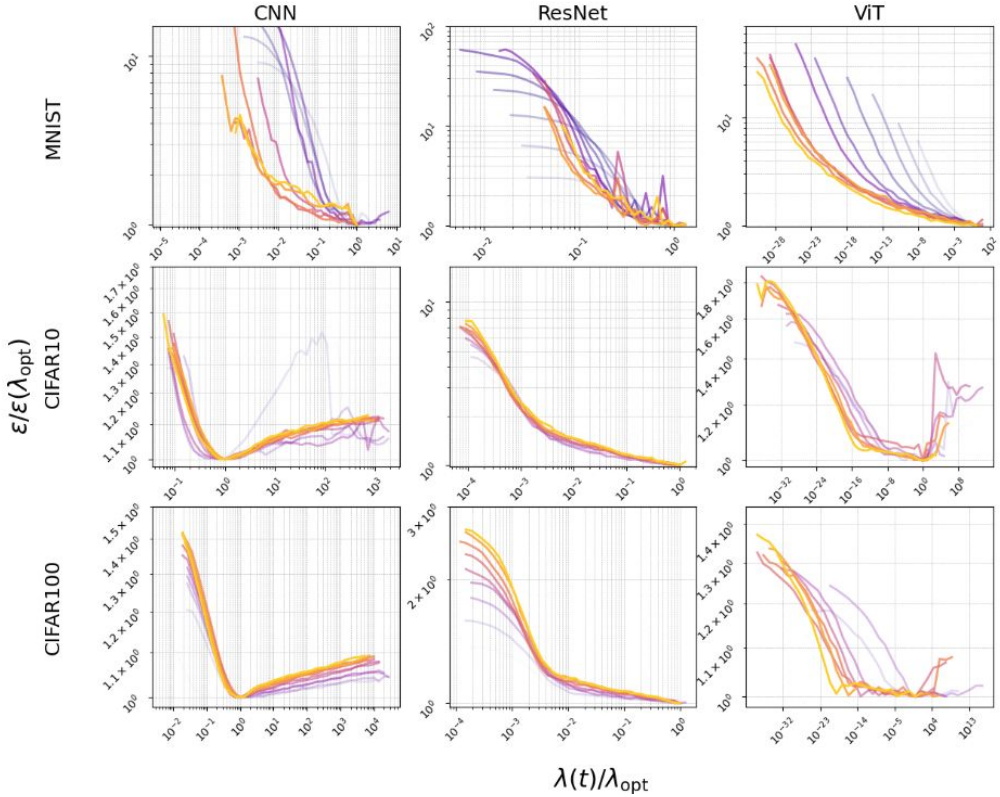
Result (1): deep classifiers



Result (2): perceptron



Result (2): deep classifiers

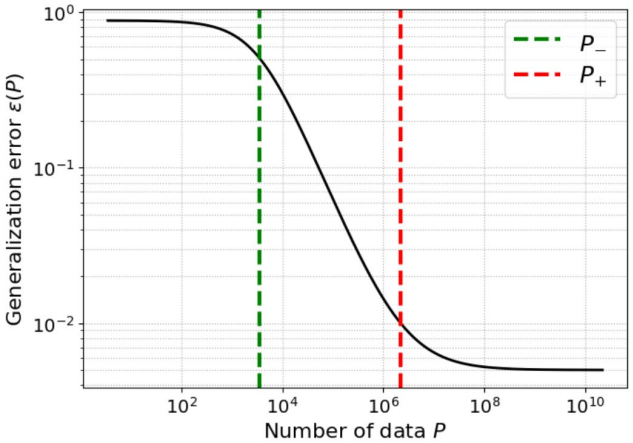


Result (3): neural scaling law

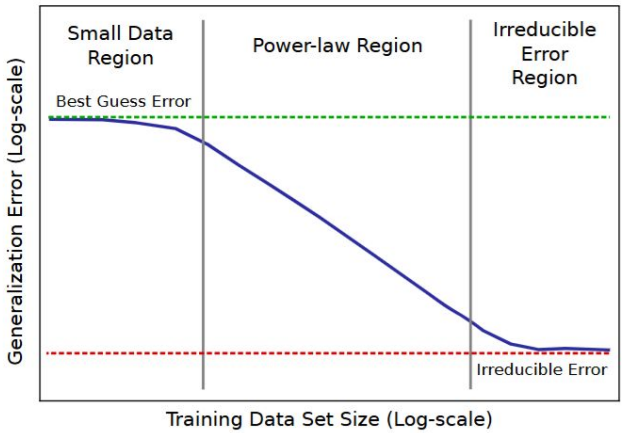
In a realistic case

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

Resulting curves



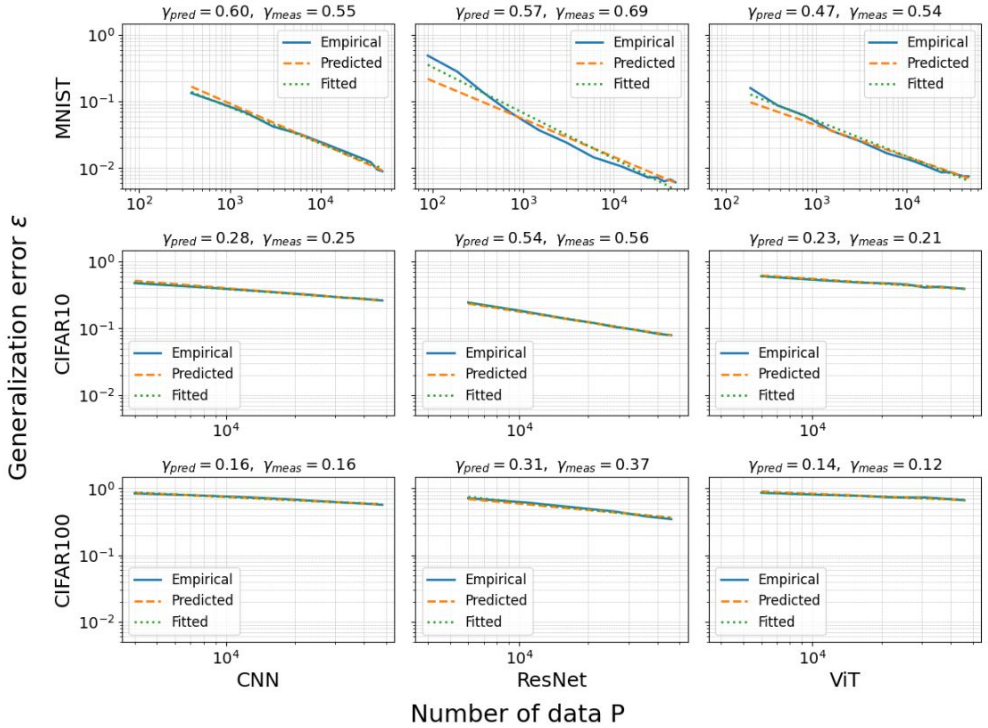
Hestness et al., (2017)



Result (3): neural scaling law

Measure γ_1, γ_2 and compute $\gamma_{\text{pred}} = \gamma_1 \gamma_2$

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



Ablation studies

1) Moderate weight decay (typical range + we request monotonicity of $\lambda(t)$)

2) SGD and Adam

3) Other norm definitions: $\ell_1, \ell_2, G_{2,1}, \text{Spectral}$

Result

- In (1) and (2) same phenomenology

- Other norms do not show the two power-laws, but they have curves collapse

Limitations and future developments

Limitation (1):
Implicit bias of cross-entropy only in image classification

Development:
Autoregressive models in language modeling

Limitation (2):
Analytical results only for perceptron. Not possible to study scaling in N.

Development:
Show same phenomenology in other settings, i.e. random features models

Limitation (3):
Analytical results at fixed norm instead of complete characterization of dynamics

Development:
DMFT in more realistic networks,
i.e. Montanari and Urbani, (2025) Dynamical Decoupling of Generalization and Overfitting in Large Two-Layer Networks

Conclusions

1. An appropriate definition of the norm acts as an order parameter of training status
2. This picture suggests an implicit bias at training time, not only for infinite epochs
3. We show two new norm-mediated scaling laws
4. Their combination produces the neural scaling law in number of data

Backup slides

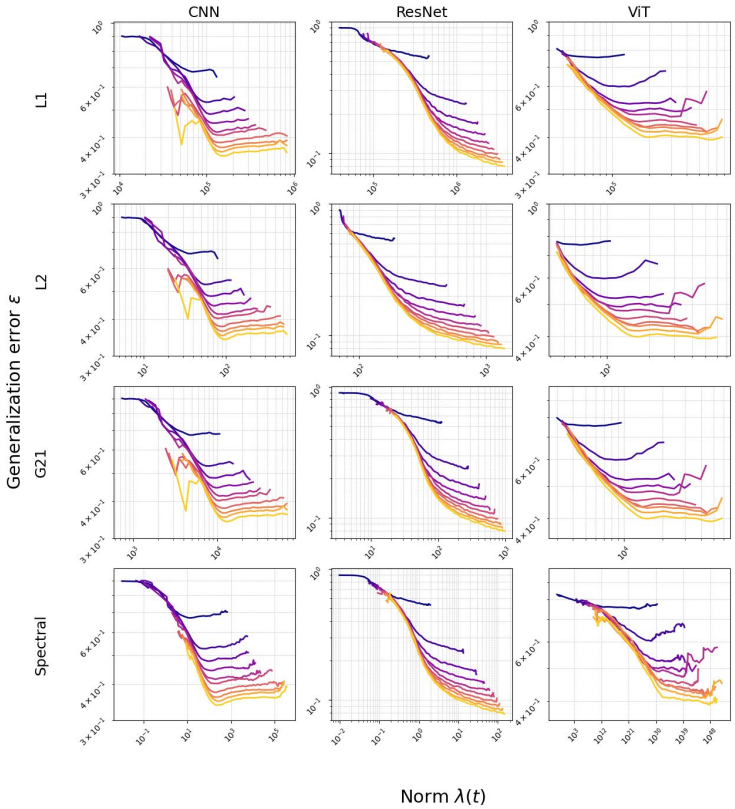
Why spectral complexity norm

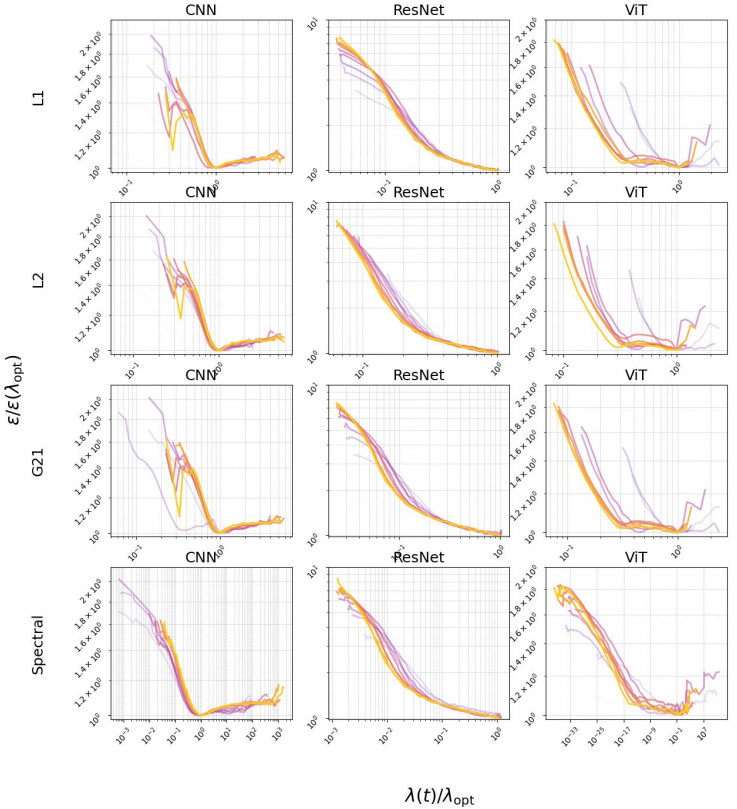
In test error margin-based decomposition

$$\Pr\left[\arg\max_j F_A(x)_j \neq y\right] \leq \hat{\mathcal{R}}_\gamma(F_A) + \hat{\mathcal{O}}\left(\frac{\|X\|_2 R_A}{\gamma n}\ln(W) + \sqrt{\frac{\ln(1/\delta)}{n}}\right)$$

Spectral norms control the relation between margins and test error

Other norms


CNN
ResNet
ViT
L1
L2
G21
Spectral
Generalization error ϵ
Norm $\lambda(t)$


CNN
ResNet
ViT
L1
L2
G21
Spectral
$\epsilon/\epsilon(\lambda_{\text{opt}})$
$\lambda(t)/\lambda_{\text{opt}}$