ICLR

UNIVERSITÄT LEIPZIG

Max Planck Institute for Mathematics in the Sciences

Analyzing Neural Scaling Laws in Two-Layer Networks with Power-Law Data Spectra

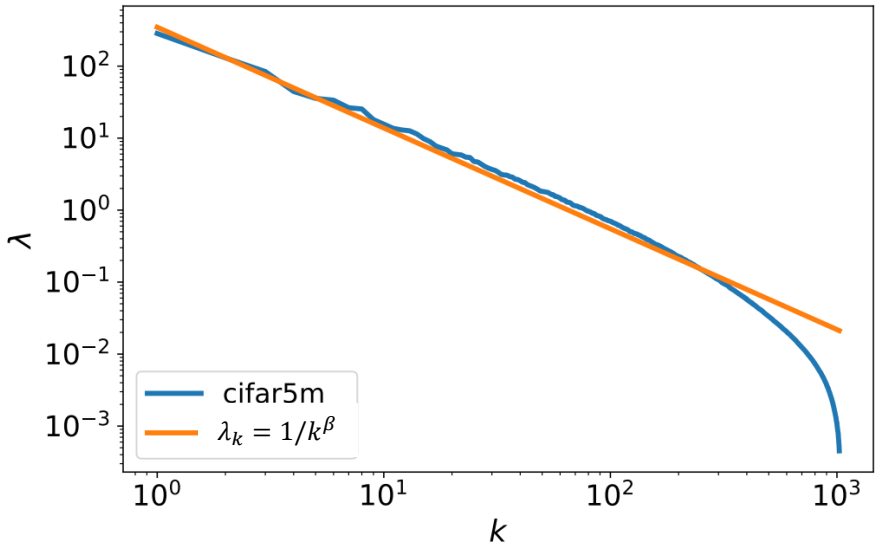
Roman Worschech[1,2] Bernd Rosenow[1]

[1] Institut für Theoretische Physik, Universität Leipzig

[2] Max Planck Institute for Mathematics in the Sciences

Introduction

- Neural scaling laws: Empirically found power-law scaling of test error (Kaplan et al, 2020)
- Various datasets exhibit power-law spectra for covariance matrix
- Previous theoretical studies focused power-law distributed data and linear models:
  - (Maloney et al., 2022), (Bahri et al., 2024) and (Bordelon et al., 2024), (Bordelon et al., 2022) and (Lin et al., 2024)



- Spectrum λ_k of cifar5m dataset

Soft Committee Machine

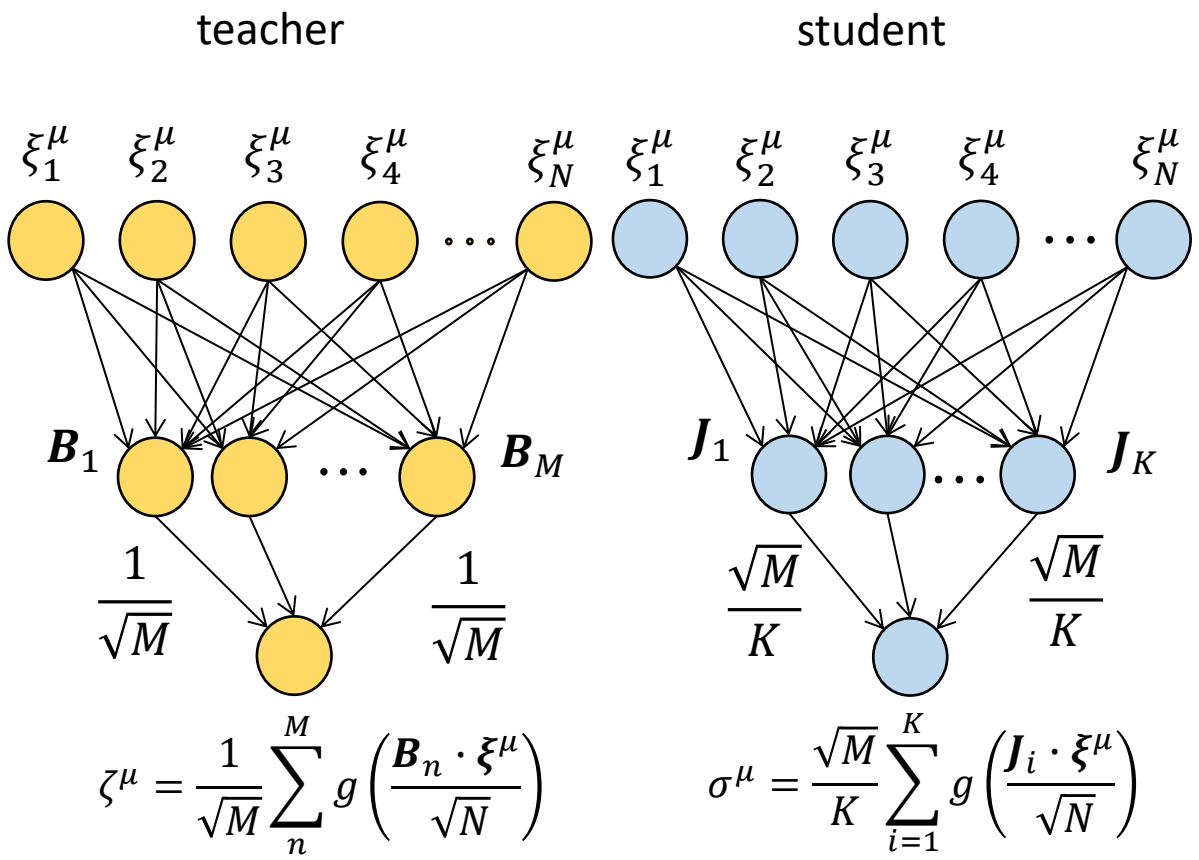
- Model of real data: $\boldsymbol{\xi}^{\mu} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$ with $\mu \in (1, \dots, p)$
- $\boldsymbol{\Sigma}$ with power-law spectrum:

$$\lambda_l = \frac{\lambda_+}{l^{1+\beta}}, \qquad l \in \{1, \dots, L\}$$

- L distinct eigenvalues
- Loss function: $\epsilon(\boldsymbol{\xi}^{\mu}) = \frac{1}{2}[\sigma(\boldsymbol{\xi}^{\mu}) - \zeta(\boldsymbol{\xi}^{\mu})]^2$
- One-pass stochastic gradient descent:

$$\boldsymbol{J}_i^{\mu+1} = \boldsymbol{J}_i^{\mu} - \eta \nabla_{\boldsymbol{J}_i} \epsilon(\boldsymbol{J}^{\mu}, \boldsymbol{\xi}^{\mu})$$

- $B_{i,a} \sim \mathcal{N}(0,1)$, $J_{i,a}^0 \sim \mathcal{N}(0, \sigma_J^2)$, for $a \in (1, \dots, N)$



Dynamical Equations

- Order parameters:

$$R_{in}^{(l)} = \frac{\boldsymbol{J}_i(\boldsymbol{\Sigma})^l \boldsymbol{B}_n}{N},\ Q_{ij}^{(l)} = \frac{\boldsymbol{J}_i(\boldsymbol{\Sigma})^l \boldsymbol{J}_j}{N},\ T_{nm}^{(l)} = \frac{\boldsymbol{B}_n(\boldsymbol{\Sigma})^l \boldsymbol{B}_m}{N}$$

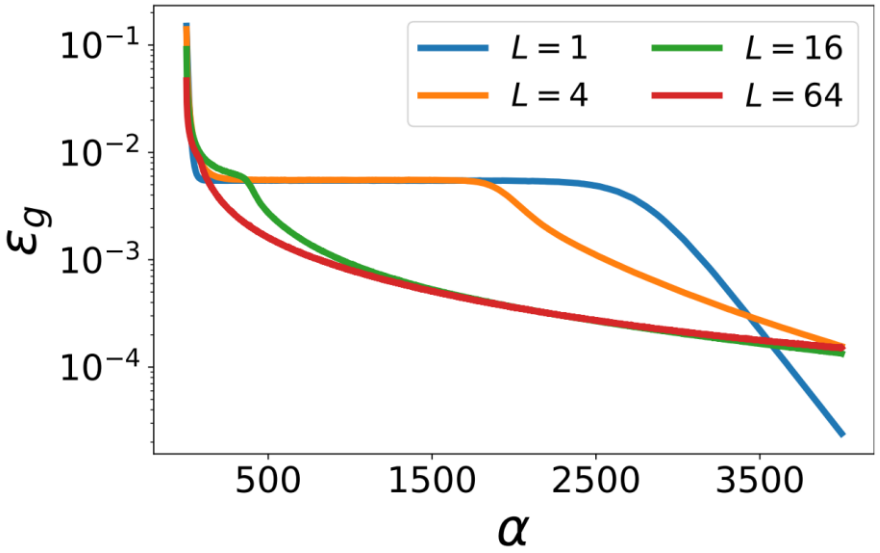
- Generalisation error: $\epsilon_g\left(\boldsymbol{Q}^{(1)}, \boldsymbol{R}^{(1)}, \boldsymbol{T}^{(1)}\right) = \langle \epsilon(\boldsymbol{J}^\mu, \boldsymbol{\xi}^\mu) \rangle_{\{\boldsymbol{\xi}\}}$

- Consider $N, p \to \infty$, $\alpha = \frac{p}{N}$ finite (Yoshida & Okada, 2019)

$$\frac{d\boldsymbol{R}^{(l)}}{d\alpha} = \frac{\eta}{K} F_1\left(\boldsymbol{R}^{(l+1)}, \boldsymbol{Q}^{(l+1)}\right)$$

$$\frac{d\boldsymbol{Q}^{(l)}}{d\alpha} = \frac{\eta}{K} F_2\left(\boldsymbol{R}^{(l+1)}, \boldsymbol{Q}^{(l+1)}\right) + O(\eta^2)$$

Simulation for $K = M = 2,\ \beta = 1,\ N = 1024,\ \eta = 0.1,\ \sigma_J = 0.01$



- Isotropic case $L = 1$ analyzed by (Saad & Solla, 1995)

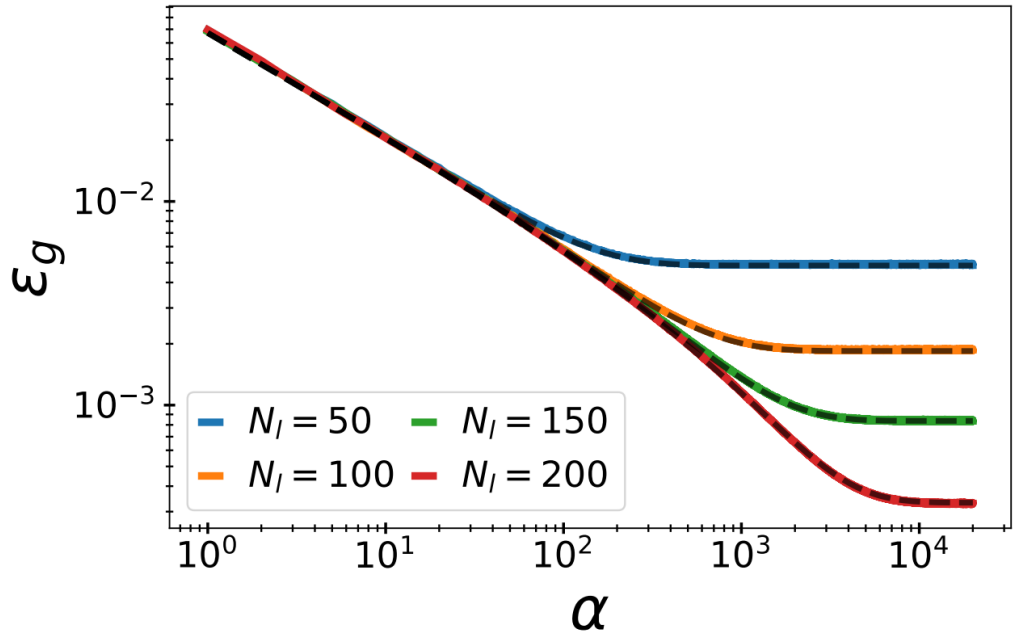
- Minimal polynomial with coefficients c_i to close ODEs:

$\sum_{k=0}^{L} c_k R_{in}^{(k)} = \sum_{k=0}^{L} c_k Q_{ij}^{(k)} = 0$ ⇒ Leading to $L \times (K^2 + MK)$ coupled ODEs

Linear Model

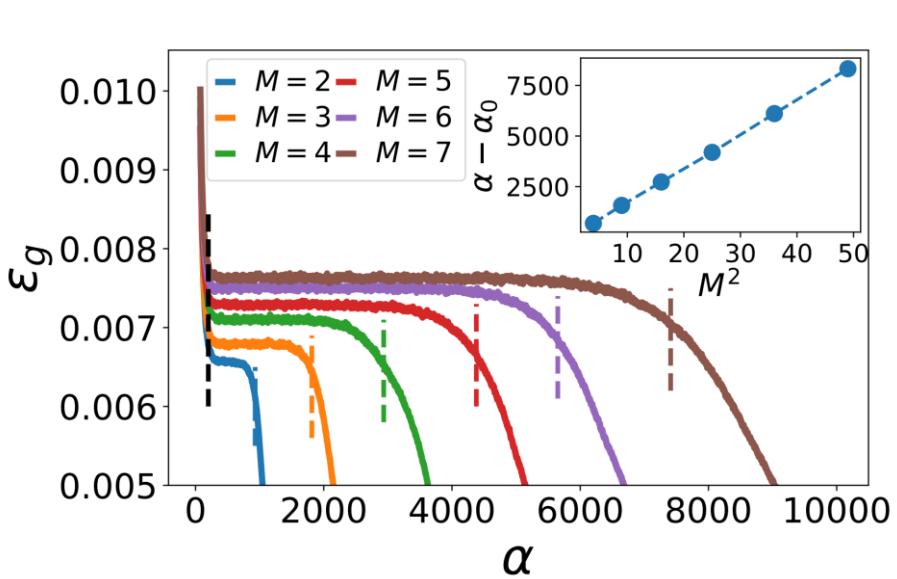
- $g(x) = x$
- Student and teacher perceptron: $K = M = 1$
- Model: only N_l components of $\boldsymbol{J}$ are trainable
- Scaling laws:
  - $\langle \epsilon_g \rangle \sim \alpha^{\frac{-\beta}{1+\beta}}$
  - $\langle \epsilon_g \rangle \sim N_l^{-\beta}$ (Asymptotic plateau)
- Similar to (Lin et al., 2024)

Simulation for $\beta = 1,\ L = N = 256,\ \eta = 0.05,\ \sigma_J = 0.01$

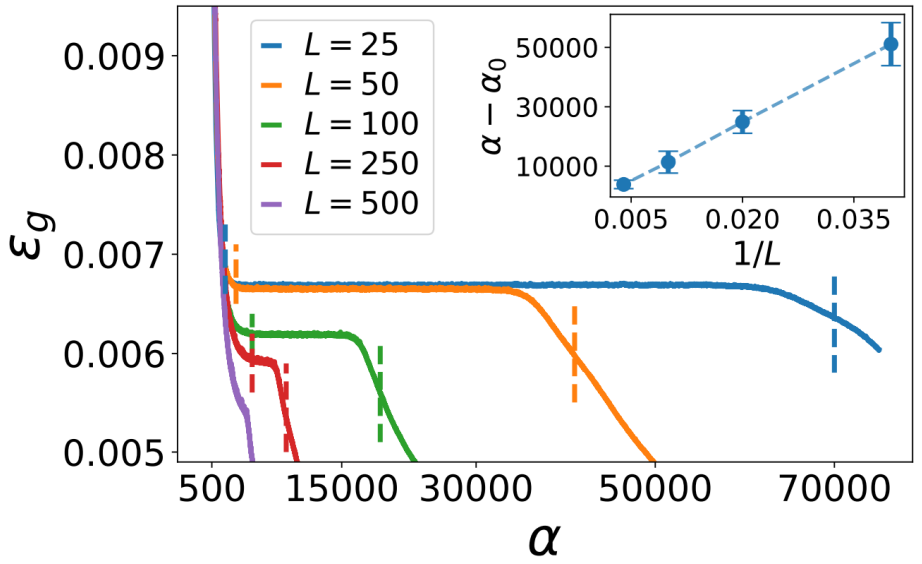


$$\langle \epsilon_g \left(N_l, \alpha\right) \rangle_{J_a^0, B_a} \underset{\eta \to 0}{\approx} \frac{1+\sigma^2}{2} \frac{\lambda_+}{\beta L} \left(\frac{1}{N_l^\beta} - \frac{1}{L^\beta} \right) + \lambda_+ \frac{1+\sigma_J^2}{2L} \frac{(2\eta\lambda_+\alpha)^{-\frac{\beta}{1+\beta}}}{1+\beta} \left[\Gamma\left(\frac{\beta}{1+\beta}, \frac{2\eta\lambda_+\alpha}{L^{\beta+1}} \right) - \Gamma\left(\frac{\beta}{1+\beta}, 2\eta\lambda_+\alpha \right) \right]$$

Plateau phase



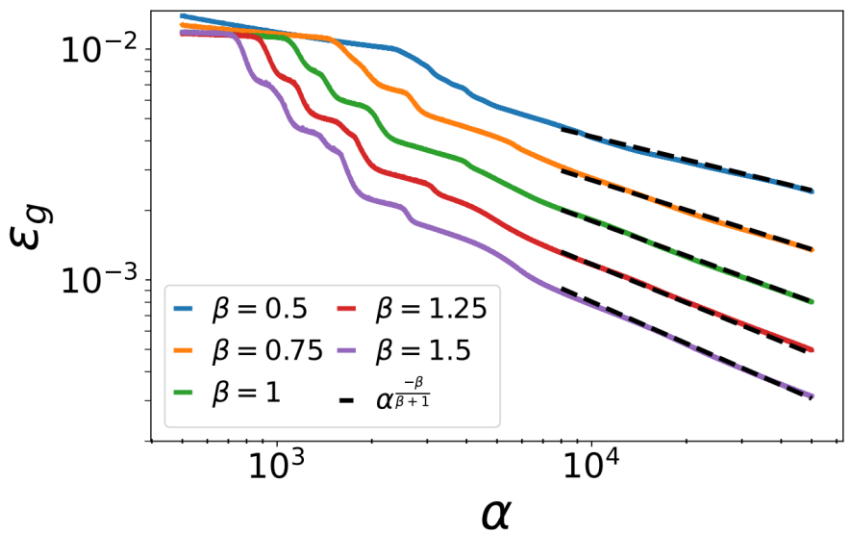
Simulation for $K = M,\ L = 4,\ N = 1024,\ \eta = 0.1,$ $\sigma_J = 0.01,\ \beta = 1$



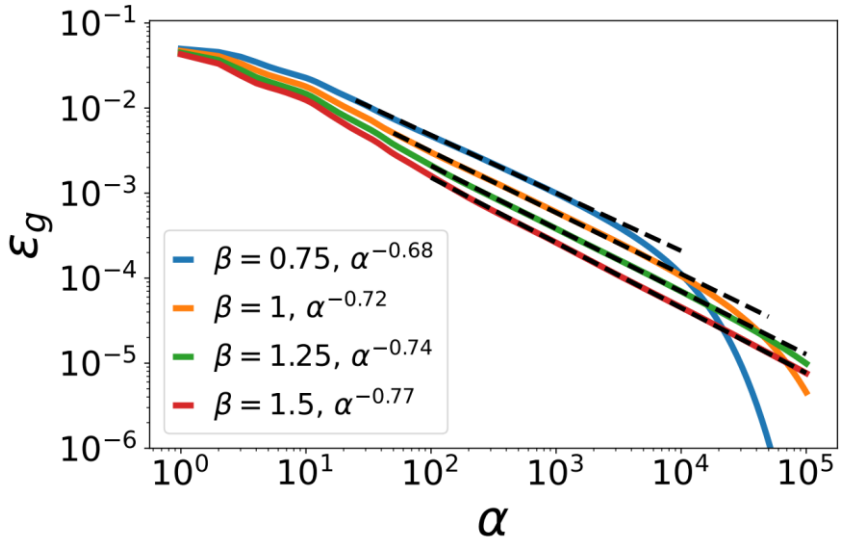
Simulation for $N = 500,\ \eta = 0.01,\ \sigma_J = 10^{-6},$ $K = M = 6,\ \beta = 0.25$

- $g(x) = \mathrm{erf}\left(\frac{x}{\sqrt{2}}\right)$
- Escape time from the plateau: $\tau_{esc} \sim \frac{\eta M^2}{L}$
- Same behavior for $K > M$

Asymptotic phase



Simulation for $L = N = 512,\ \eta = 0.01,$
$\sigma_J = 10^{-6},\ K = M = 40$

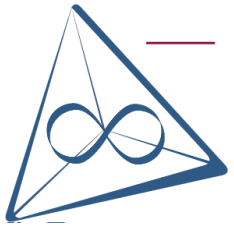


Solution of ODEs for $L = N = 1024,\ \eta = 0.001,\ \sigma_J = 0.1$

- $g(x) = \mathrm{erf}\left(\frac{x}{\sqrt{2}}\right)$
- Scaling law: $\langle \epsilon_g \rangle \sim \alpha^{\frac{-\beta}{1+\beta}}$

- New model: $\zeta^\mu = g\left(\frac{\boldsymbol{B}\cdot\boldsymbol{\xi}^\mu}{\sqrt{N}}\right),\ \sigma^\mu = cg\left(\frac{\boldsymbol{J}\cdot\boldsymbol{\xi}^\mu}{\sqrt{N}}\right)$
- Trainable parameters $\boldsymbol{J}$ and c
- $g(x) = \max(0, x)$

UNIVERSITÄT
LEIPZIG



Max Planck Institute for
Mathematics
in the Sciences

Thank you for your attention!

References

- Alexander Maloney, Daniel A Roberts, and James Sully. A solvable model of neural scaling laws. arXiv preprint arXiv:2210.16859, 2022.
- Blake Bordelon and Cengiz Pehlevan. Learning curves for SGD on structured features. In International Conference on Learning Representations, 2022.
- Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining neural scaling laws. Proceedings of the National Academy of Sciences, 121(27):e2311878121, 2024.
- Licong Lin, Jingfeng Wu, Sham M Kakade, Peter L Bartlett, and Jason D Lee. Scaling laws in linear regression: Compute, parameters, and data. arXiv preprint arXiv:2406.08466, 2024.
- Blake Bordelon, Alexander Atanasov, and Cengiz Pehlevan. A dynamical model of neural scaling laws. In Forty-first International Conference on Machine Learning, 2024

References

- Yuki Yoshida and Masato Okada. Data-dependence of plateau phenomenon in learning with neural network — statistical mechanical analysis. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'AlchéBuc, E. Fox, and R. Garnett (eds.), Advances in Neural Information Processing Systems, volume 32. Curran Associates, Inc., 2019.
- Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. arXiv preprint arXiv:2001.08361, 2020.
- David Saad and Sara A. Solla. On-line learning in soft committee machines. Phys. Rev. E, 52: 4225–4243, Oct 1995.