Title: Solvable models of scaling and emergence in deep learning

Speakers: Cengiz Pehlevan

Collection/Series: Theory + AI Workshop: Theoretical Physics for AI

Date: April 10, 2025 - 9:00 AM

URL: https://pirsa.org/25040091

Solvable Models of Scaling and Emergence in Deep Learning

Cengiz ("Jen-ghiz") Pehlevan

Harvard John A. Paulson
School of Engineering
and Applied Sciences

Kempner
INSTITUTE

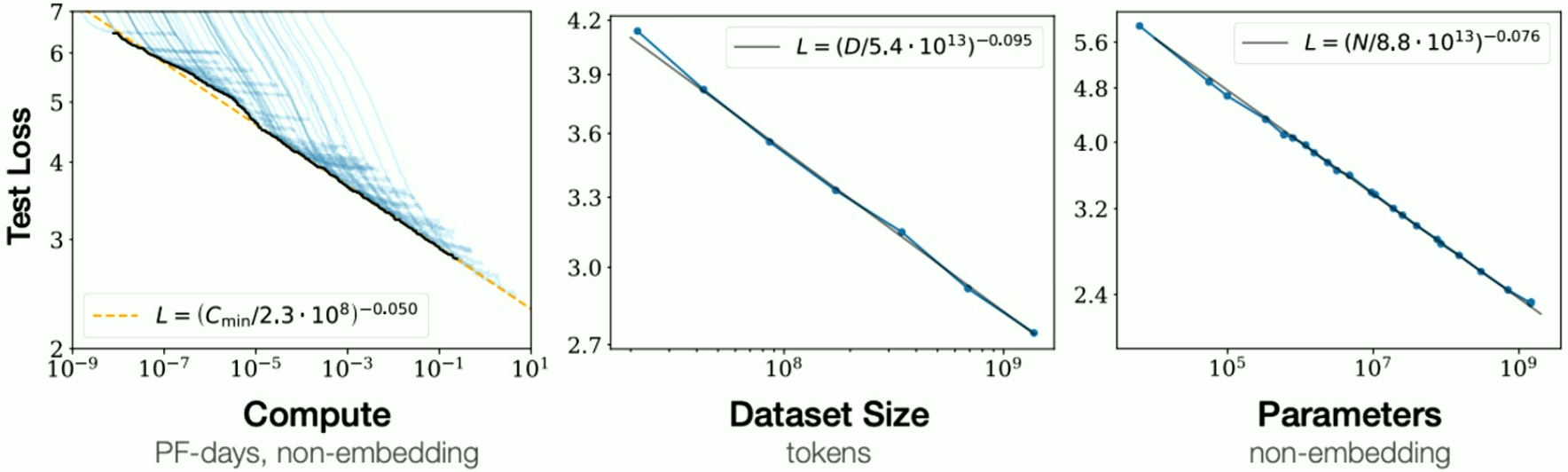
For the Study of Natural
& Artificial Intelligence
at Harvard University

Dominant AI Paradigm: Scaling

Kaplan et al 2020, “Scaling Laws for Neural Language Models”

Dominant AI Paradigm: Scaling

Kaplan et al 2020, “Scaling Laws for Neural Language Models”



GPT-4 Technical Report

OpenAI*

3 Predictable Scaling

A large focus of the GPT-4 project was building a deep learning stack that scales predictably. The primary reason is that for very large training runs like GPT-4, it is not feasible to do extensive model-specific tuning. To address this, we developed infrastructure and optimization methods that have very predictable behavior across multiple scales. These improvements allowed us to reliably predict some aspects of the performance of GPT-4 from smaller models trained using $1,000\times$ – $10,000\times$ less compute.

Key questions

- What are the limits of scaling? How do we describe and understand infinitely-sized networks? Is there a “thermodynamic” description that allows reduced descriptions?

- Predictable scaling and hyperparameter transfer from small to large scale are fundamental to efficiency. How do we achieve it?

- What defines these scaling laws? Data, Architecture, Optimization. Why power laws?

- At what scale new capabilities emerge?

Key questions

- What are the limits of scaling? How do we describe and understand infinitely-sized networks? Is there a “thermodynamic” description that allows reduced descriptions?

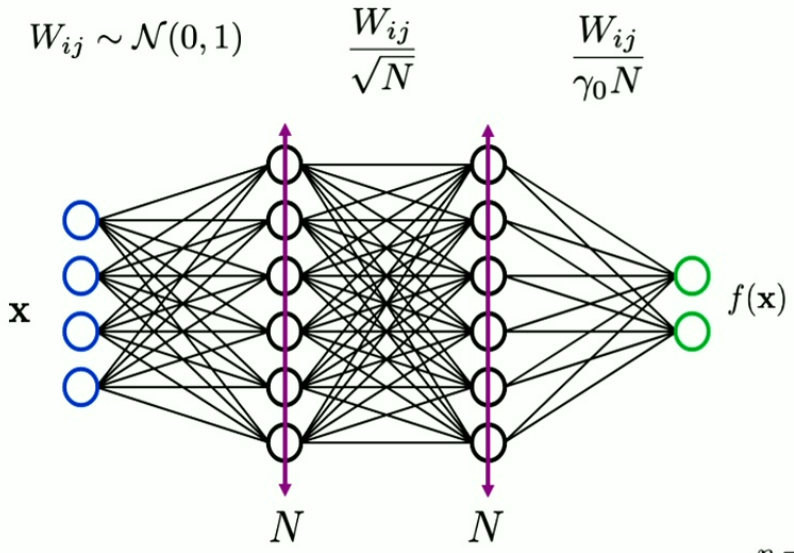
- Predictable scaling and hyperparameter transfer from small to large scale are fundamental to efficiency. How do we achieve it?

Blake’s talk

- What defines these scaling laws? Data, Architecture, Optimization. Why power laws?

- At what scale new capabilities emerge?

Different parameter scalings lead to different limiting behavior

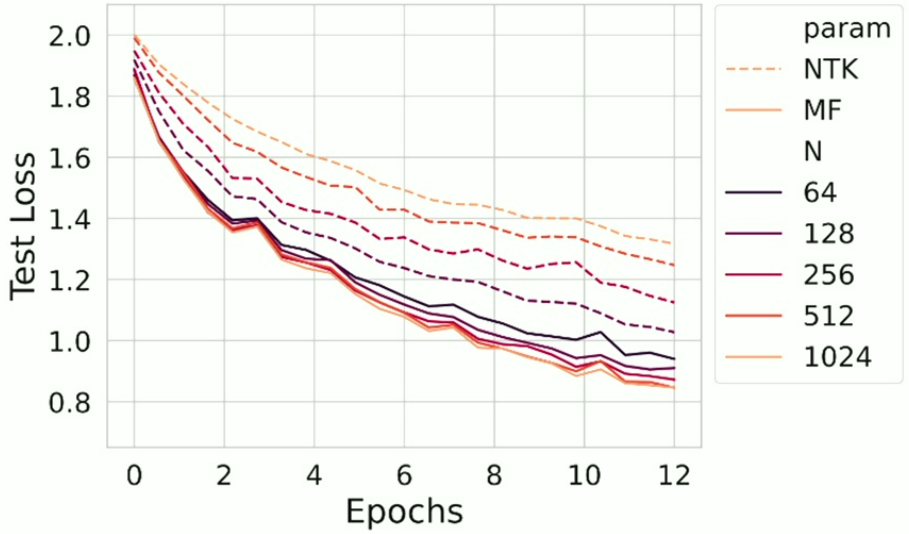


$\eta = \eta_0 \gamma_0^2 N$

Neural Tangent (Jacot et al, 2018; Lee et al, 2019): $\gamma_0 \sim \mathcal{O}(1/\sqrt{N})$

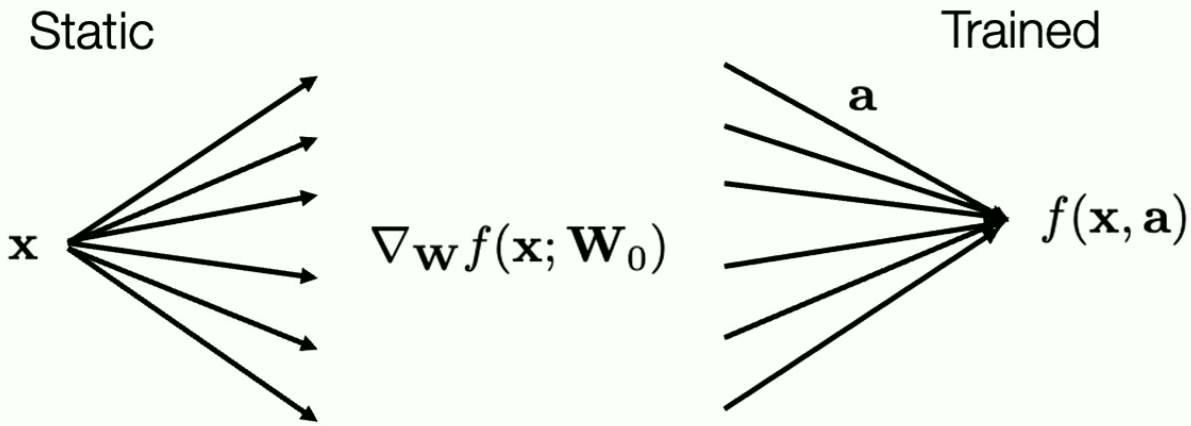
Mean field scaling (Mei et al 2018, Yang et al 2021): $\gamma_0 \sim \mathcal{O}(1)$

Depth 12 ResNet on CIFAR-10 SGD



Simulations by Bordelon

The "Lazy" Infinite-Width Limit and Neural Tangent Kernel



Kernel machine with the Neural Tangent Kernel

$$K(\mathbf{x}, \mathbf{x}') \equiv \nabla_{\mathbf{w}} f(\mathbf{x}; \mathbf{W}_0) \cdot \nabla_{\mathbf{w}} f(\mathbf{x}'; \mathbf{W}_0)$$

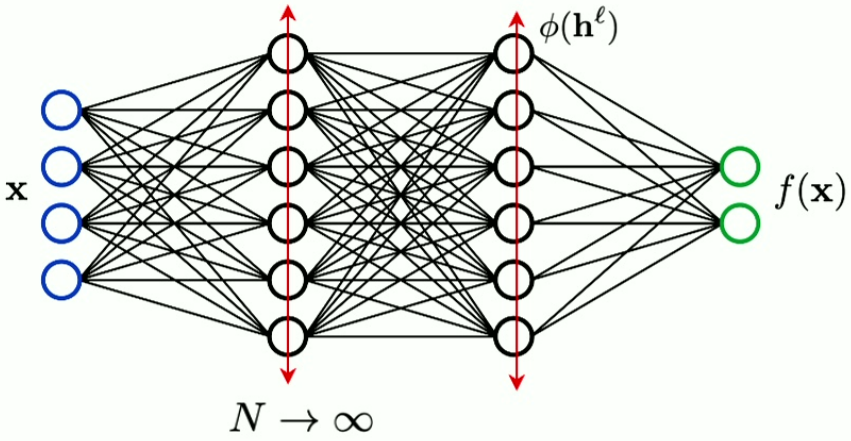
Jacot et al., 2018; Arora et al., 2019; Lee et al., 2019

The "Rich" Infinite-Width Limit and DMFT

Dynamical Mean Field Theory (DMFT) analysis of infinite neural networks yields self-consistent summary statistics for learning, mirroring how macroscopic variables emerge in thermodynamics at the thermodynamic limit.

We developed DMFTs for MLPs, CNNs, ResNets, Transformers, involving various infinite limits (width, depth, key/query dimension, input dimension and dataset size) with real data or teacher-student setups (See Blake's talk).

Example: DMFT for MLPs in the infinite-width limit



Summary statistics are given by feature and gradient kernels

$$\Phi^\ell_{\mu\nu}(t,t') = \frac{1}{N}\phi(\boldsymbol{h}^\ell_\mu(t)) \cdot \phi(\boldsymbol{h}^\ell_\nu(t'))\,, \quad G^\ell_{\mu\nu}(t,t') = \frac{1}{N}\boldsymbol{g}^\ell_\mu(t) \cdot \boldsymbol{g}^\ell_\nu(t')$$

They fully describe learning trajectory of the network's output

$$\frac{d}{dt} f(\mathbf{x}) = F\left(\{\Phi^\ell(t,s), G^\ell(t,s)\}_{\ell=0}^{L}\right)$$

Bordelon, Pehlevan 2018

Scaling exponents

Kaplan et al., 2020

$$\mathcal{L}(N,T) = \left[\left(\frac{N_c}{N}\right)^{\alpha_N/\alpha_T} + \frac{T_c}{T}\right]^{\alpha_T} \qquad \alpha_N = 0.076, \qquad \alpha_T = 0.095$$

Hoffman et al., 2022

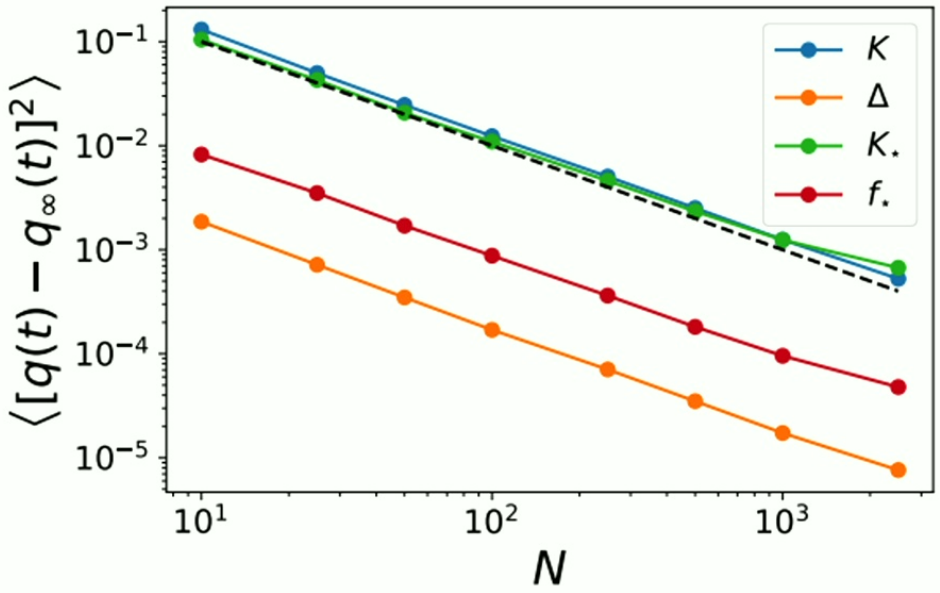
$$\mathcal{L}(N,T) = E + N^{-\alpha_N} + T^{-\alpha_T} \qquad \alpha_N = 0.39, \qquad \alpha_T = 0.28$$

Compute proportional to NT

Dynamics of Finite Width Kernel and Prediction Fluctuations in Mean Field Neural Networks

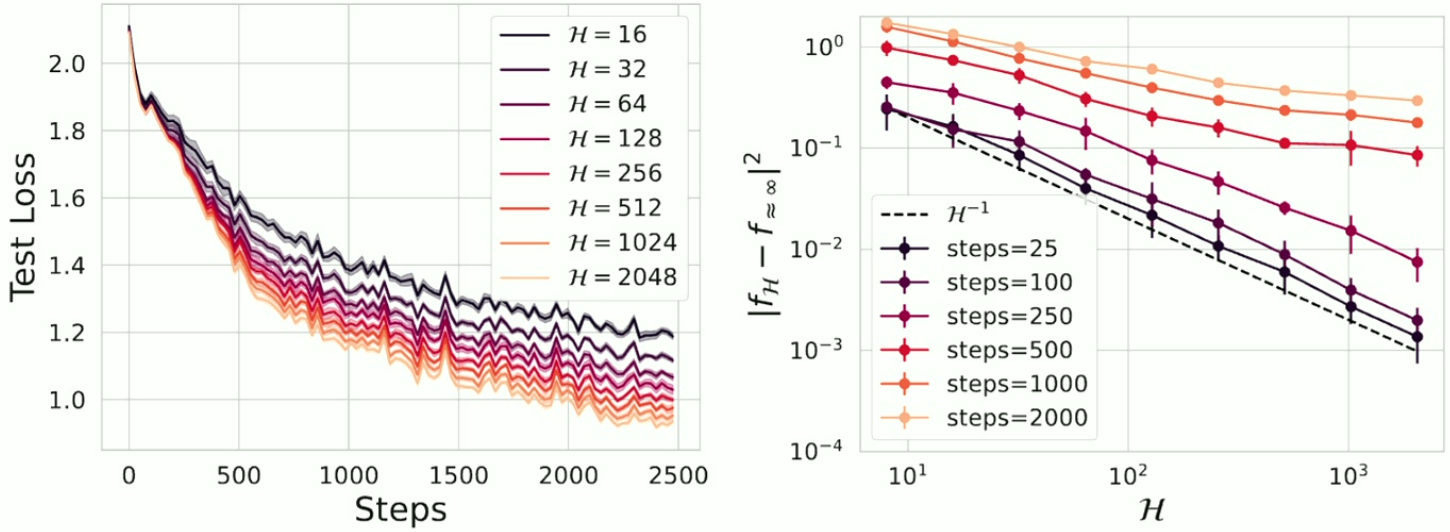
Blake Bordelon & Cengiz Pehlevan

NeurIPS 2023



Deviations from consistent behavior across widths

SGD dynamics for a vision transformer trained on CIFAR-5M. H – number of attention heads



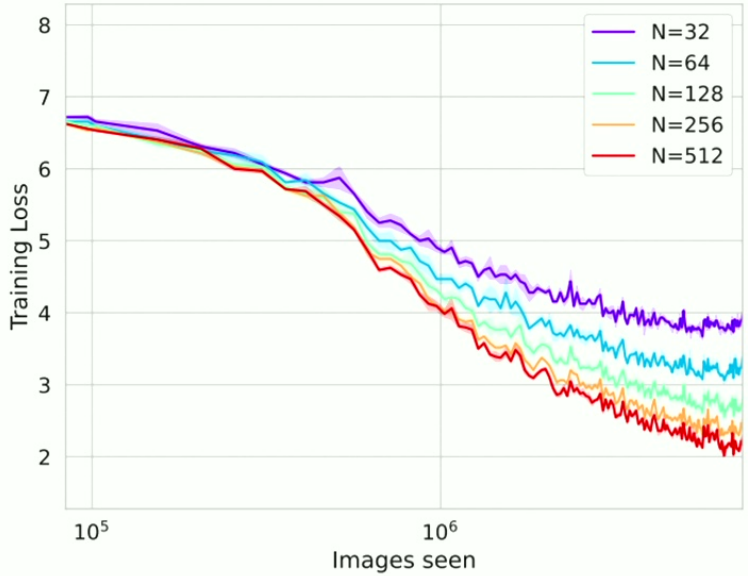
(a) Training Dynamics Varying $\mathcal{H}$

(b) Convergence to $\mathcal{H} \to \infty$ limit

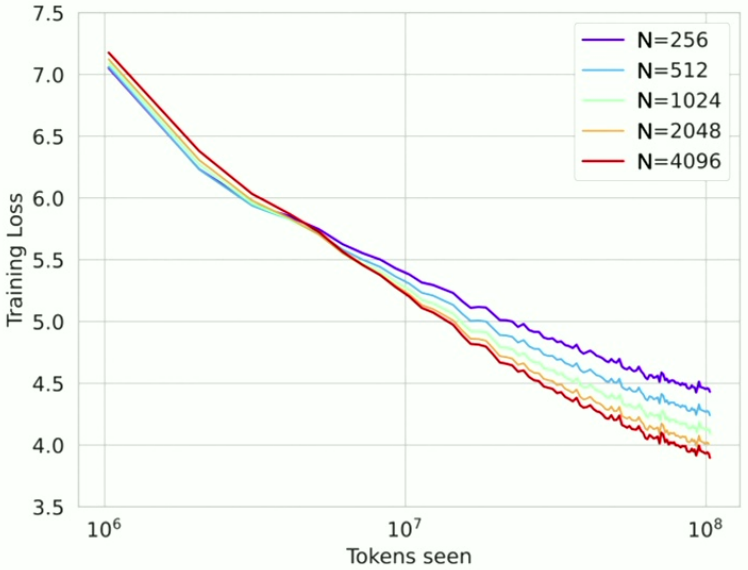
Bordelon, Chaudhry, Pehlevan

Infinite Limits of Multi-head Transformer Dynamics (NeurIPS 2024)

Deviations from consistent behavior across widths for more complex datasets



(b) Imagenet

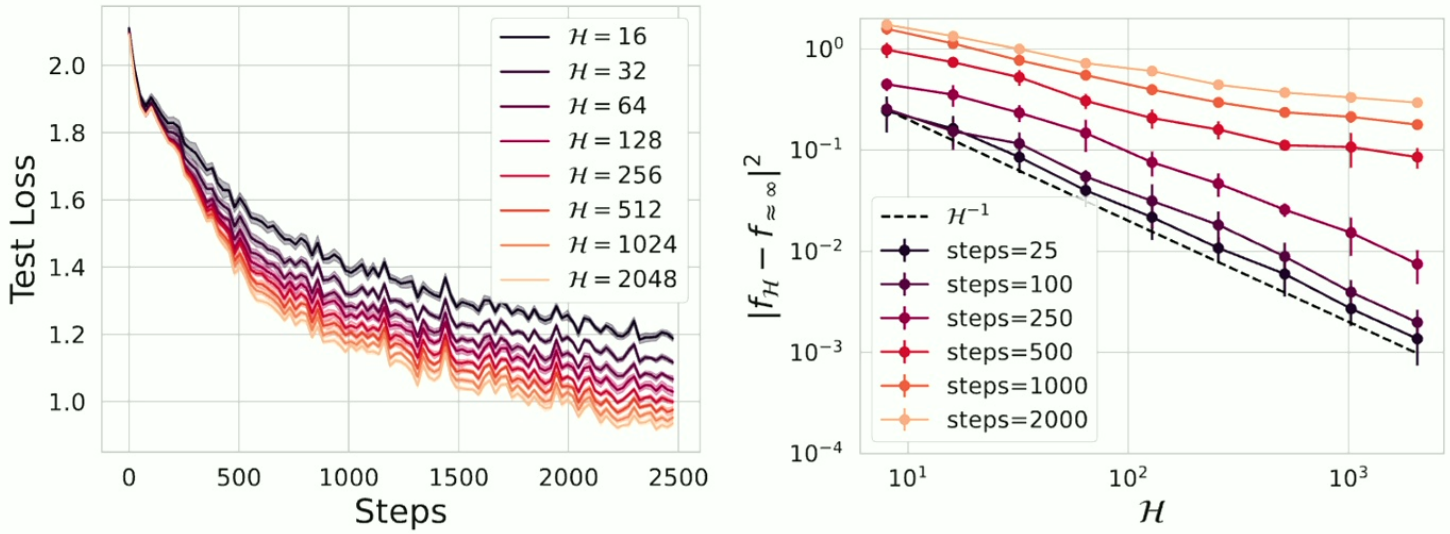


(c) Wikitext-103

More experiments in Nikhil Vyas*, Alex Atanasov*, Blake Bordelon*, Depen Morwani, Sab Sainathan, Cengiz Pehlevan, NeurIPS, 2023

Deviations from consistent behavior across widths

SGD dynamics for a vision transformer trained on CIFAR-5M. H – number of attention heads



(a) Training Dynamics Varying $\mathcal{H}$

(b) Convergence to $\mathcal{H} \to \infty$ limit

Bordelon, Chaudhry, Pehlevan

Infinite Limits of Multi-head Transformer Dynamics (NeurIPS 2024)

A Dynamical Model of Neural Scaling Laws

Blake Bordelon [1 2] **Alexander Atanasov** [3 2] **Cengiz Pehlevan** [1 2]

ICML 2024

- Related “static” works by: Caponnetto & De Vito 2007; Bordelon et al., 2020; Spigler et al., 2020; Bahri et al. 2021; Mel & Ganguli 2021; Favero et al. 2021; Maloney et al. 2022; Atanasov et al. 2022; Cui et al. 2022; Cagnetta et al. 2023; Simon et al. 2023; Dohmatob et al. 2024; Defilippis et al. 2024

- More recent “dynamic” works: Paquette et al. 2024; Lin et al. 2024

Main ideas of the model

- We are looking for a simple model where we can vary parameters (N), dataset size (P) and training time (T)

- For analytical tractability, we will consider the lazy limit of neural network training. (I will later generalize to feature learning).

- In the lazy limit, neural networks are kernel machines (NTK in the infinite width, Jacot et al. 2018)

- Main modeling idea: Kernels of finite-width networks are “noisy” versions of the infinite-width models.

Bahri et al. 2021, Atanasov et al. 2022, Maloney et al. 2022

Teacher-Student Setup

Eigenvalues and eigenfunctions of the infinite-width NTK are a complete basis:

$$K_\infty(\mathbf{x}, \mathbf{x}') = \sum_k \psi_k^\infty(\mathbf{x}) \psi_k^\infty(\mathbf{x}') \qquad \int d\mathbf{x}\, p(\mathbf{x}) \psi_k^\infty(\mathbf{x}) \psi_l^\infty(\mathbf{x}) = \lambda_k \delta_{kl}$$

Finite-width model: $f(\mathbf{x}) = \sum_k w_k \psi_k^N(\mathbf{x})$

(student, width N)

Expand: $\psi_k^N(\mathbf{x}) = \frac{1}{\sqrt{N}} \sum_l A_{kl} \psi_l^\infty(\mathbf{x})$

$$\lim_{N\to\infty} \frac{1}{N} \mathbf{A}^\top \mathbf{A} = \mathbf{I} \quad A_{ij} \sim \mathcal{N}(0,1)$$

Teacher: $y(\mathbf{x}) = \sum_k \bar{w}_k \psi_k^\infty(\mathbf{x})$

Data: $\mathcal{D} = (\mathbf{x}_\mu, y_\mu)_{\mu=1}^P \qquad \mathbf{x}_\mu \sim p(\mathbf{x}) \qquad y_\mu = y(\mathbf{x}_\mu) + \epsilon_\mu \qquad \Psi_{\mu k} \equiv \psi_k^\infty(\mathbf{x}_\mu)$

Teacher-Student Setup

Teacher: $y(x) = \sum_k \bar{w}_k \psi_k^\infty$

Student: $f(x) = \sum_l \frac{1}{\sqrt{N}} \sum_k w_k A_{kl} \psi_l^\infty(x)$

Consider gradient-flow on MSE loss

$$v_k^0 \equiv \bar{w}_k - \frac{1}{\sqrt{N}} \sum_{l=1}^{N} A_{lk} w_l$$

$$\frac{d}{dt}\mathbf{v}^0(t) = -\left(\frac{1}{N}\mathbf{A}^\top \mathbf{A}\right)\left(\frac{1}{P}\mathbf{\Psi}^\top \mathbf{\Psi}\right)\mathbf{v}^0(t)$$

- Using path integral (MSR) methods and Gaussian equivalence, we study the P, N >> 1 regime.
- Averages over two types of disorder: initialization and data.
- We derive a mean field theory for full asymptotic learning dynamics. See paper
- We can also analyze stochastic gradient descent and momentum.

Power law in – power law out

$$K_\infty(\mathbf{x},\mathbf{x}') = \sum_k \psi_k^\infty(\mathbf{x})\psi_k^\infty(\mathbf{x}') \qquad \int d\mathbf{x}\, p(\mathbf{x})\psi_k^\infty(\mathbf{x})\psi_l^\infty(\mathbf{x}) = \lambda_k \delta_{kl} \qquad y(x) = \sum_k \bar{w}_k \psi_k^\infty$$

Source-and-capacity constraints: $(\bar{w}_k)^2 \lambda_k \sim k^{-a}, \quad \lambda_k \sim k^{-b}$

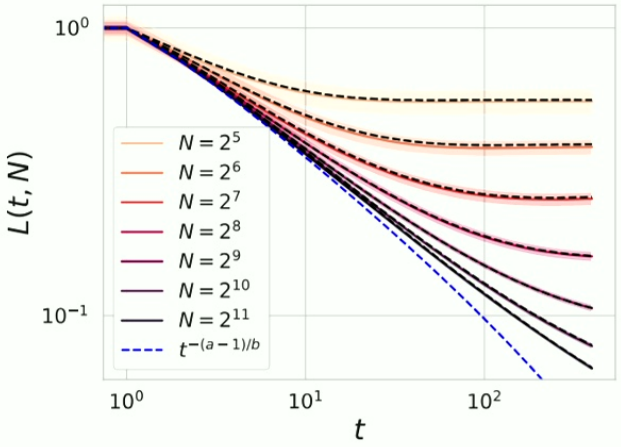
$$\mathcal{L}(t,P,N) \approx \begin{cases} t^{-(a-1)/b}\ ,\ P,N\to\infty\ ,\ \text{(Time-Bottleneck)} \\ P^{-(a-1)}\ ,\ t,N\to\infty\ ,\ \text{(Data-Bottleneck)} \\ N^{-(a-1)}\ ,\ t,P\to\infty\ ,\ \text{(Model-Bottleneck)} \end{cases}$$

Caponnetto & De Vito 2007;
Bordelon et al., 2020;
Spigler et al., 2020;
Bahri et al. 2021;
Favero et al. 2021;
Maloney et al. 2022;
Cui et al. 2022;
Cagnetta et al. 2023;
Simon et al. 2023;
Dohmatob et al. 2024;
Defilippis et al.; 2024
…

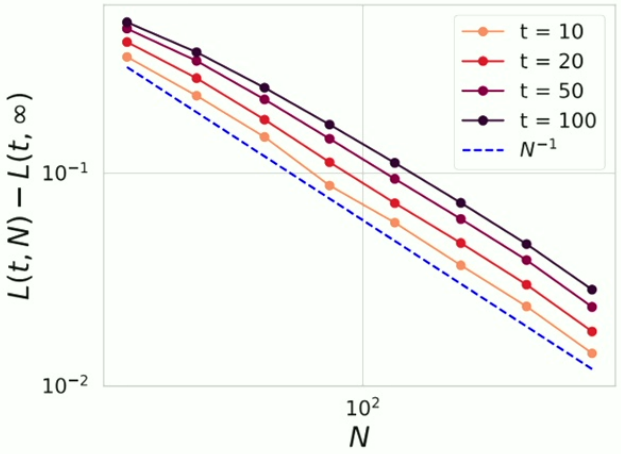
$$(\bar{w}_k)^2 \lambda_k \sim k^{-a}, \quad \lambda_k \sim k^{-b}$$

$$\mathcal{L}(t, P, N) \approx \begin{cases} t^{-(a-1)/b}\ , \ P, N \to \infty\ , \ \text{(Time-Bottleneck)} \\ P^{-(a-1)}\ , \ t, N \to \infty\ , \ \text{(Data-Bottleneck)} \\ N^{-(a-1)}\ , \ t, P \to \infty\ , \ \text{(Model-Bottleneck)} \end{cases}$$

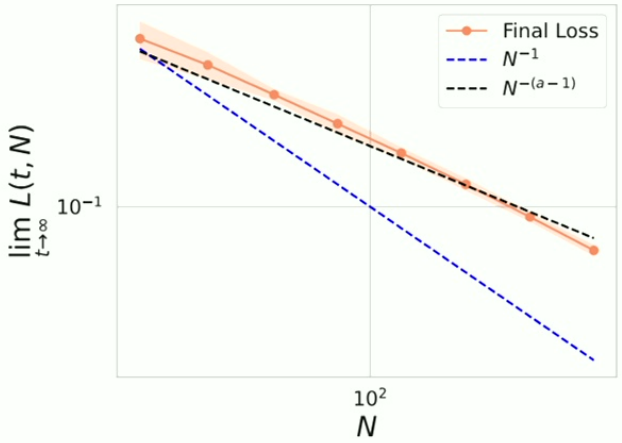
$$a = 1.5, \quad b = 1.25$$



(a) $P = 1000$ Test Loss Dynamics



(b) Early Time Model Convergence



(c) Late Time Model Bottleneck

Bottlenecks as rank-constraints

$$\mathcal{L}(t,P,N) \approx \begin{cases} t^{-(a-1)/b}\ ,\ P,N \to \infty\ , \text{ (Time-Bottleneck)} \\ P^{-(a-1)}\ ,\ t,N \to \infty\ , \text{ (Data-Bottleneck)} \\ N^{-(a-1)}\ ,\ t,P \to \infty\ , \text{ (Model-Bottleneck)} \end{cases}$$

$$\mathcal{L} \approx \sum_{k>k_\star} (\bar{w}_k)^2 \lambda_k \approx k_\star^{-(a-1)}$$

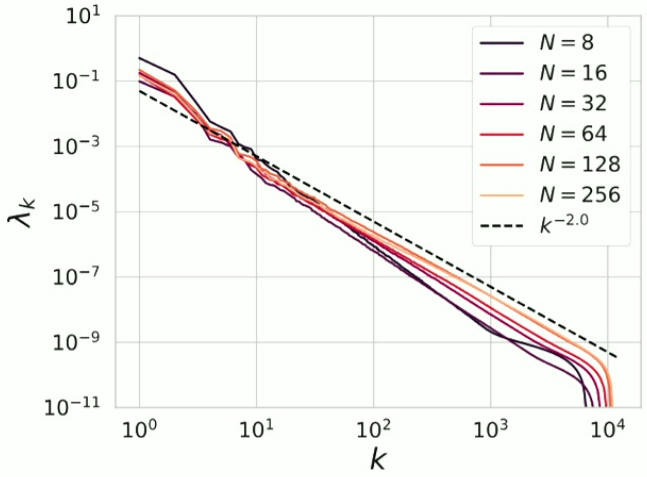
Data Bottleneck: $k_\star = P$

Spigler et al., 2020

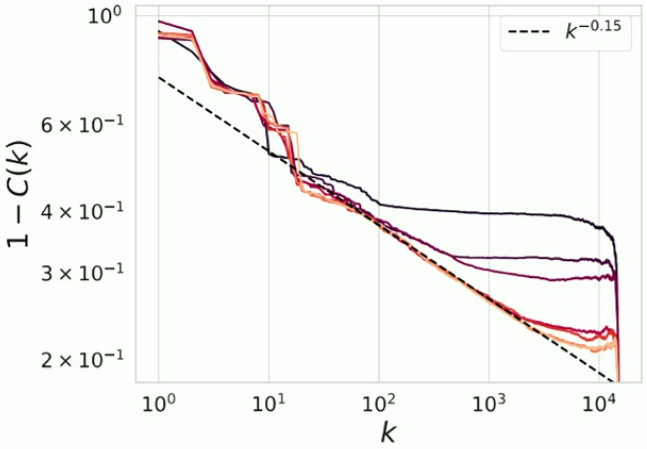
Model Bottleneck: $k_\star = N$

Time Bottleneck: $k_\star = t^{1/b}$ $\qquad (\lambda_k = k^{-b},\ \tau_k \sim k^b)$

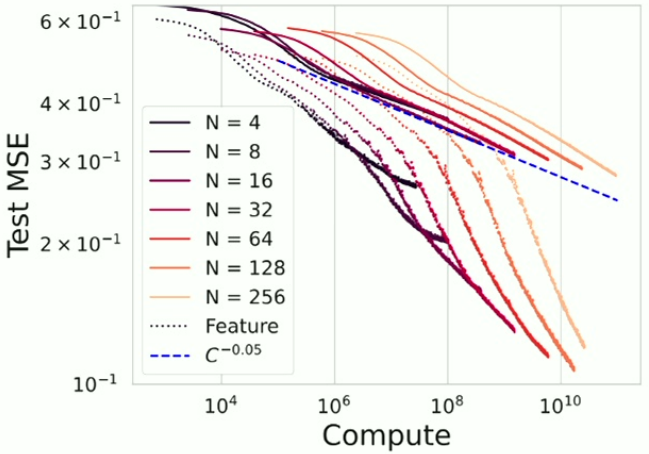
ResNets on CIFAR-5M



(a) NTK Spectra for Varying Widths



(b) Task-Power Decay



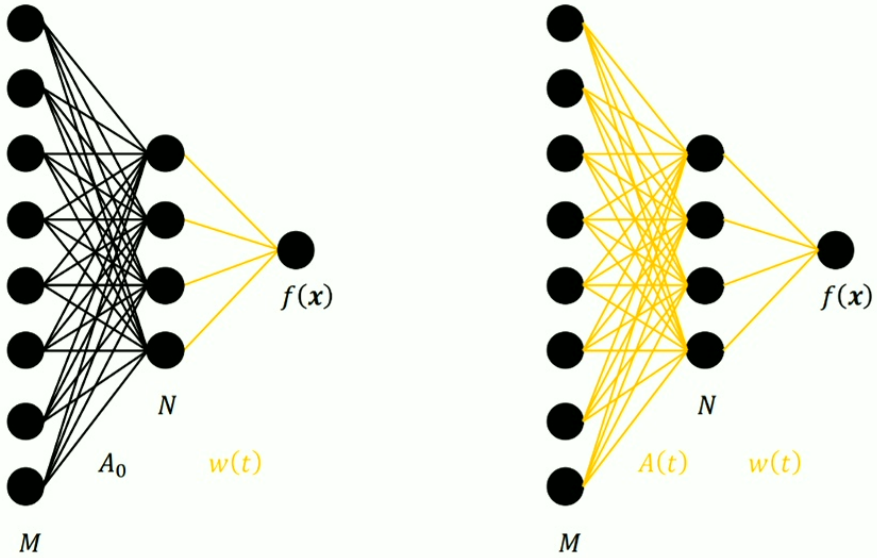
(c) Lazy vs Feature Compute Scaling

$$C(k) \equiv \frac{\sum_{i \leq k} \lambda_i (\bar{w}_i)^2}{\sum_i \lambda_i (\bar{w}_i)^2}$$

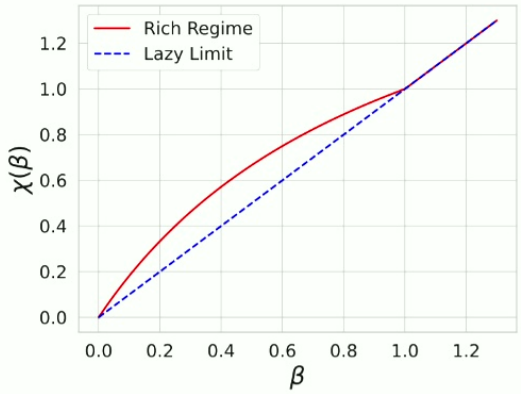
How Feature Learning Can Improve Neural Scaling Laws

Blake Bordelon*, Alexander Atanasov* , Cengiz Pehlevan

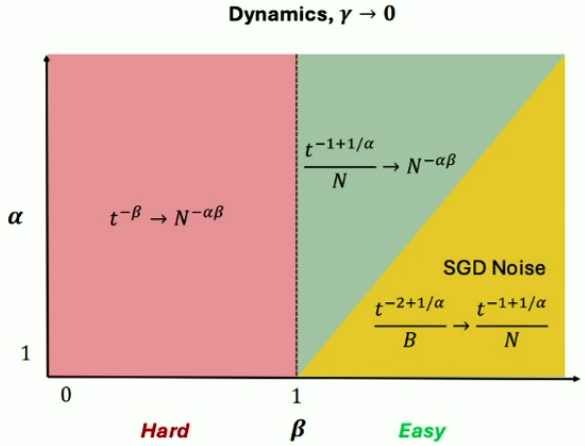
ICLR 2025



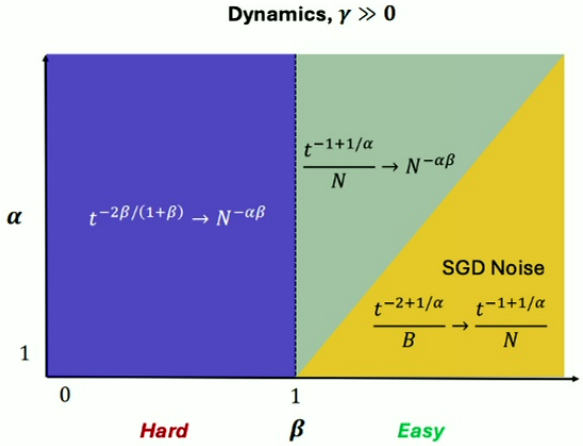
Improved scaling for hard, but not easy tasks from feature learning



(a) $\lim_{N\to\infty} \mathcal{L}(t, N) \sim t^{-\chi(\beta)}$



(b) Lazy Limit



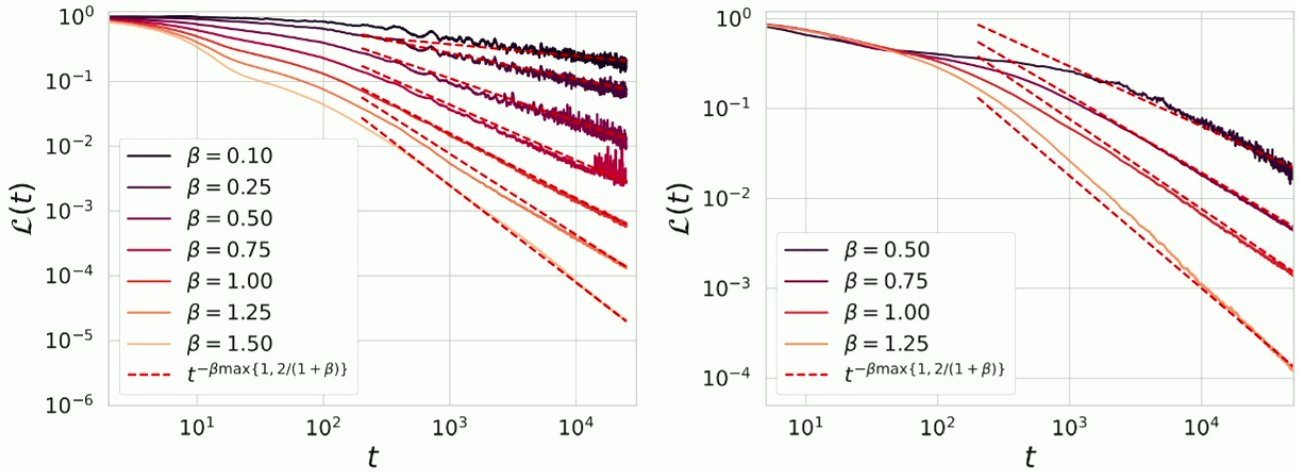
(c) Rich Regime

Hard task: out of RKHS

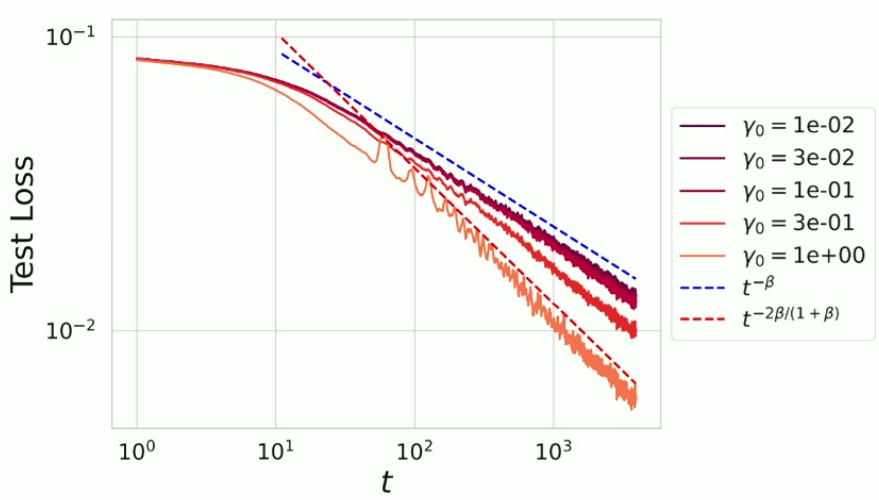
$$|y|_{\mathcal{H}}^2 = \sum_k (w_k^\star)^2 = \sum_k k^{-\alpha(\beta-1)-1} \approx \begin{cases} \frac{1}{\alpha(\beta-1)} & \beta > 1 \\ \infty & \beta < 1 \end{cases}$$

Theory predicts correct power exponents for deep ReLU networks

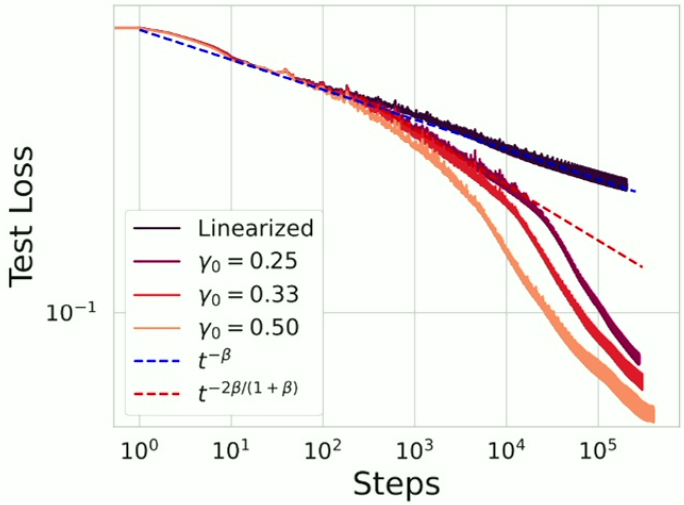
$$y(\theta) = \sum_{k=1}^{\infty} k^{-q} \cos(k\theta) \ , \ K(\theta, \theta') = \sum_{k=1}^{\infty} \lambda_k \cos(k(\theta - \theta'))$$



(a) ReLU ($q_\phi = 1.0$) with varying $\beta = \frac{2q-1}{2q_\phi}$ (b) Varying q_ϕ with fixed target ($q = 1.4$)



(a) CNNs on MNIST-1M, $\beta = 0.30$



(b) CNNs on CIFAR-5M $\beta = 0.075$

Interim Conclusion

- Scaling laws derive from structures is data
- Scaling in lazy learning regime is well-understood
- Scaling in rich regimes is still open

Acknowledgments

Blake Bordelon
Hamza Chaudhry
Ganesh Kumar
Clarissa Lauditi
Adam Lee
Mary Letey
Alex Meterez
Mo Osman
Billy Qian
Ben Ruben
Sabarish Sainathan
William Tong
Ningjing Xia

Alumni:
Abdul Canatar
Alex Atanasov
Jacob Zavatone-Veth
Anindita Maiti

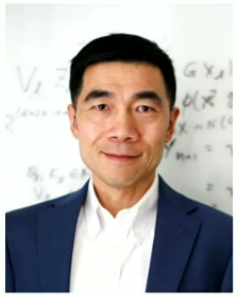
Collaborators:
Yue Lu

Blake Bordelon

Jacob Zavatone-Veth

Alex Atanasov

Hamza Chaudhry



Yue Lu

Mary Letey

Anindita Maiti

Harvard John A. Paulson School of Engineering and Applied Sciences

Kempner Institute FOR THE STUDY OF NATURAL & ARTIFICIAL INTELLIGENCE

ALFRED P. SLOAN FOUNDATION

HDSI

intel

Google

NIH

NSF

DARPA