HARVARD UNIVERSITY

HARVARD UNIVERSITY
CENTER OF MATHEMATICAL SCIENCES AND APPLICATIONS

Training Dynamics in Large Networks:

From Super-Wide to the Scaling Law Regime

Blake Bordelon

ProbAI Scaling Workshop

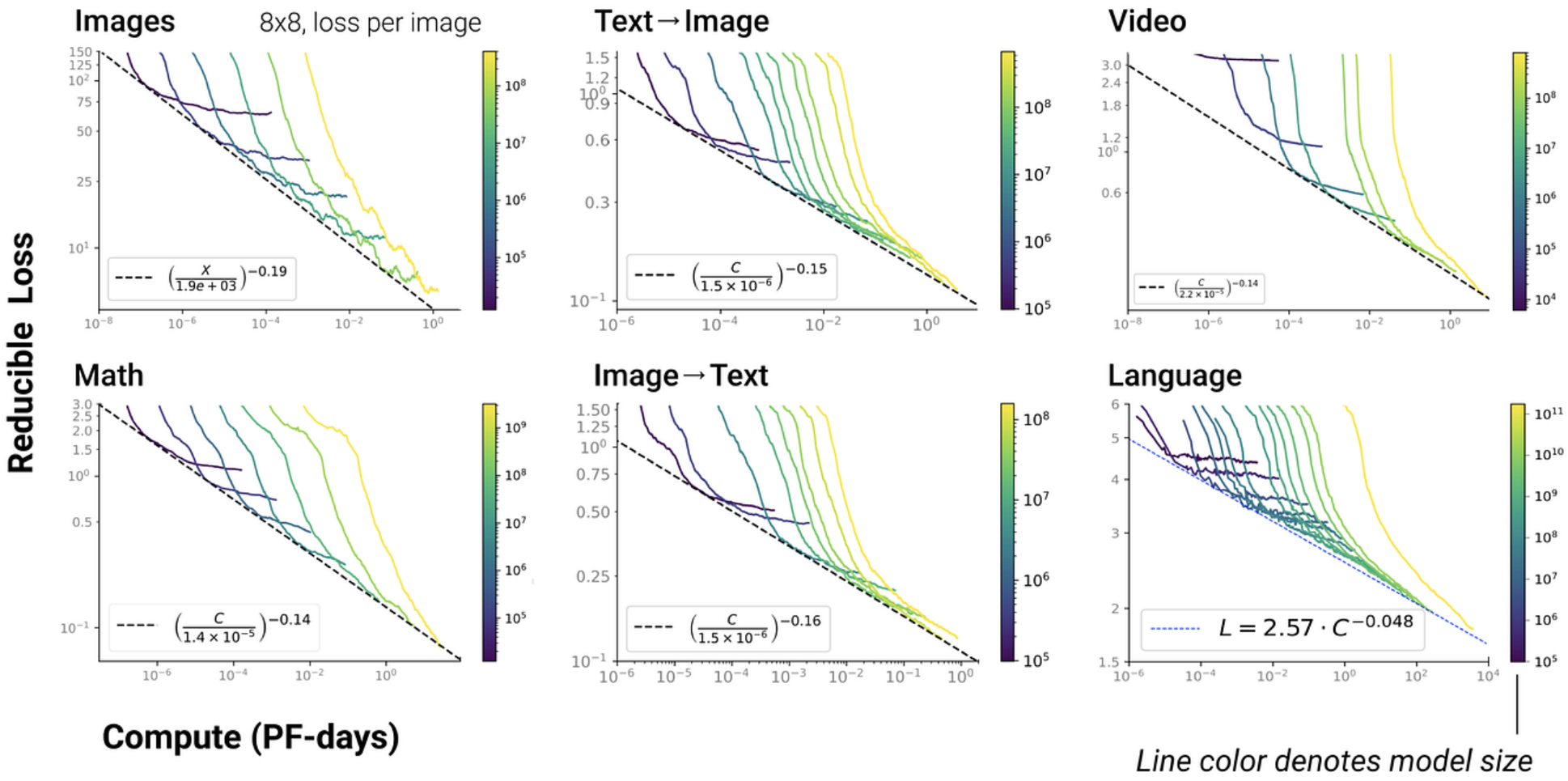
June 2026

TEXAS
The University of Texas at Austin

INSTITUTE FOR COMPUTATIONAL ENGINEERING & SCIENCES
ODEN

Scale as a Harsh Reality in Modern Models

Neural Scaling Laws -> investment in larger models trained on more data!



Scale of frontier models is utterly baffling

Example: Claude Mythos announced recently, rumored 10T parameter MOE model

Cost of Artemis 2 launch?

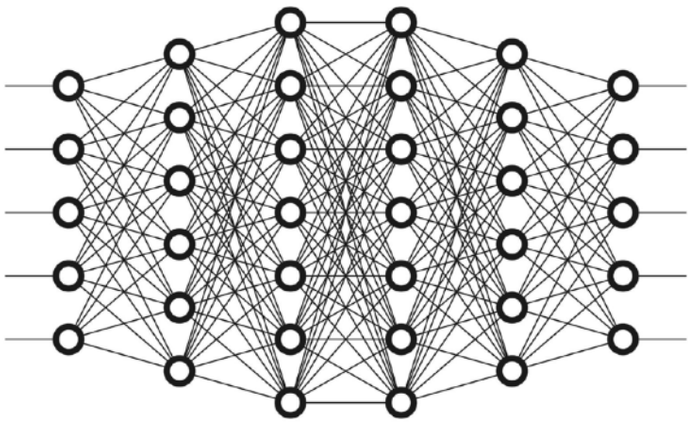
~ $4B USD per launch

Rumored cost for training Mythos ~$10B USD

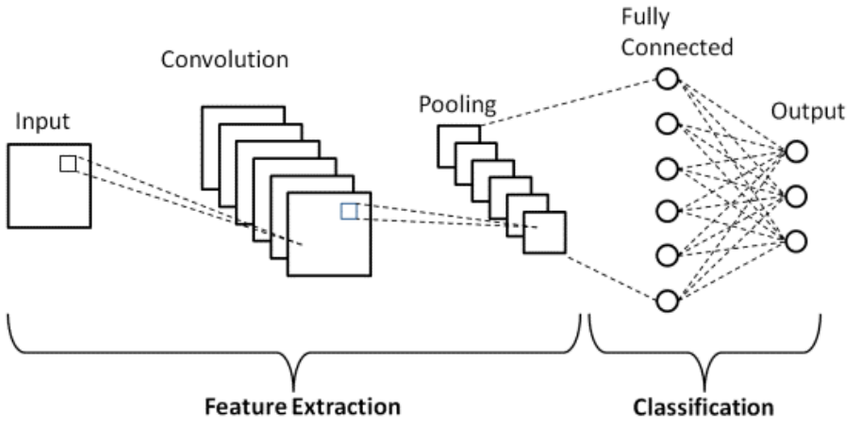
Modern training runs are ***Moonshots!***

Architectural Challenges for Modern Deep Learning Theory

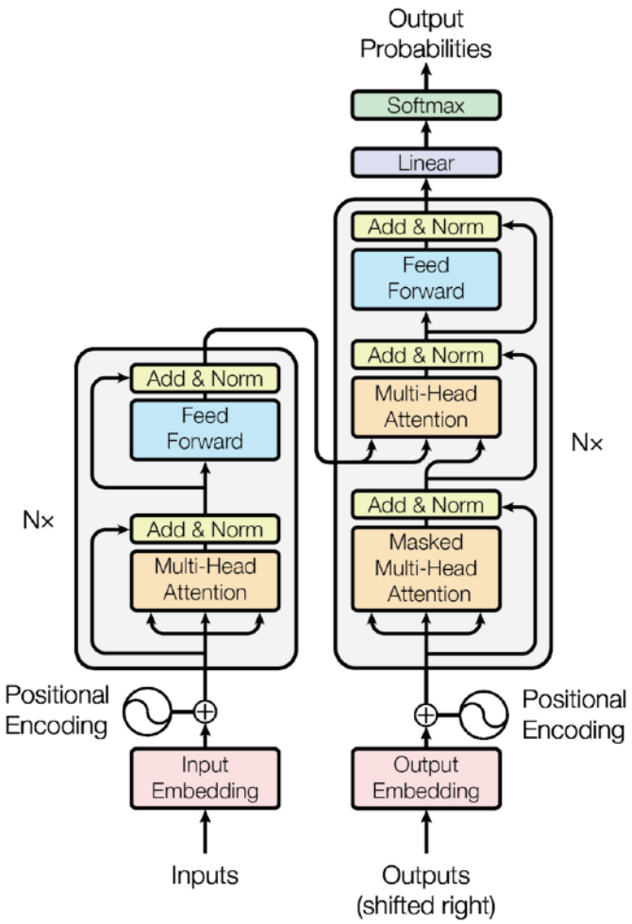
Complex deep architectures *with billions or trillions of degrees of freedom!*



Fully Connected (MLP)



Convolutional network (weights shared across spatial positions)



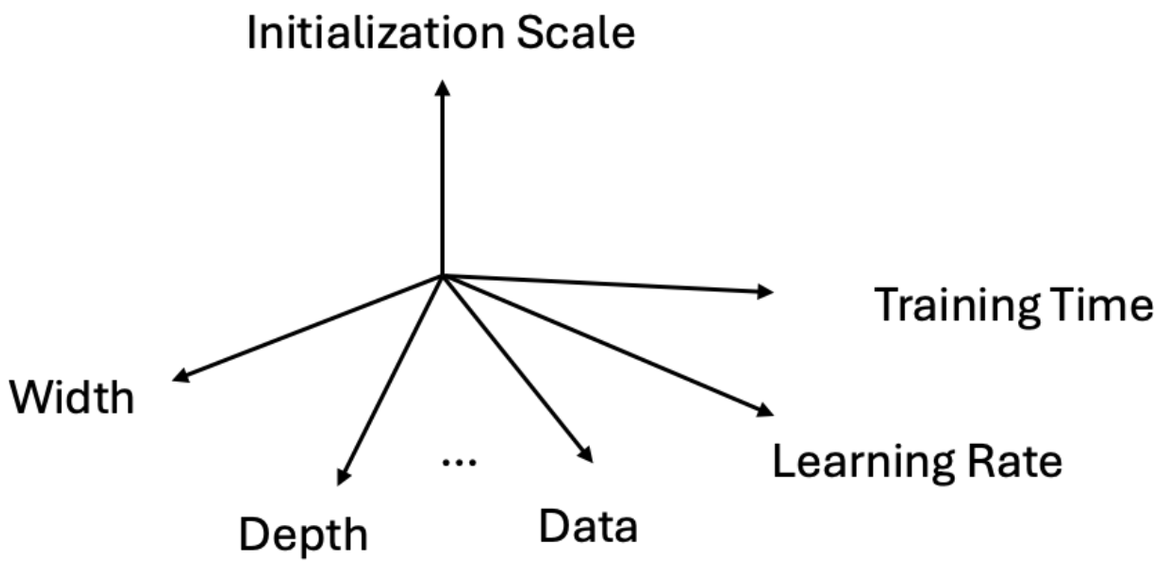
Transformer (learnable attention maps across spatial positions)

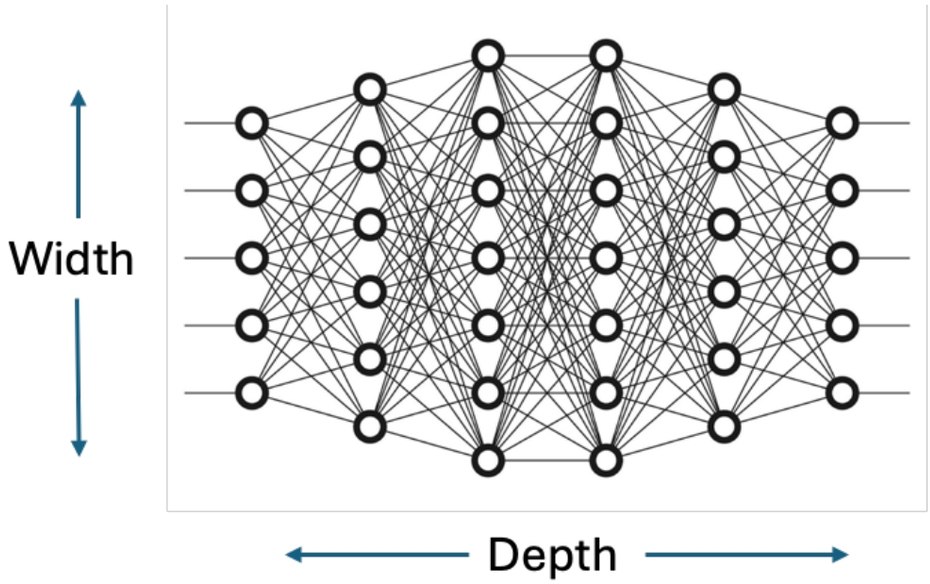
How do these models behave as they become very large?
What do we mean by large (which scaling dimensions)?

When / how do intelligent behaviors emerge at the macroscopic scale from microscopic parameter adjustments (learning)?

Today's Talk

1. Some Example Scaling Limits of Neural Networks (width and depth)





How to scale up to get well defined infinite parameter limits? What do limits look like?

Dynamical mean field theory (DMFT) for deep learning networks

Neurons (~ particles) interacting at finite width N

$\lim_{N \to \infty}$ ⟶ *independent neurons (particles) coupled to population averages*

2. Application to more modern and realistic architectures (transformers w/ MOE blocks)

Not apriori obvious what/how to scale to get well defined limits / predictable behavior

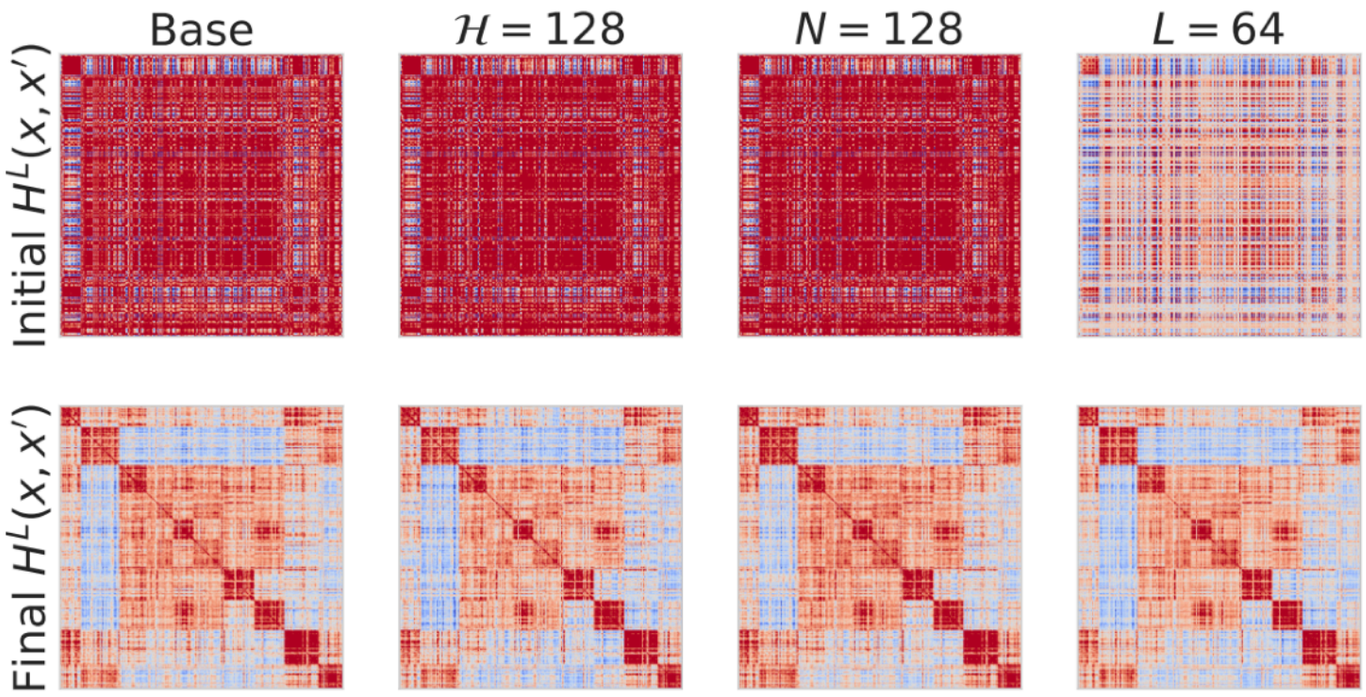
3. Practical extensions: Hyperparameter transfer to reduce training costs during scaling

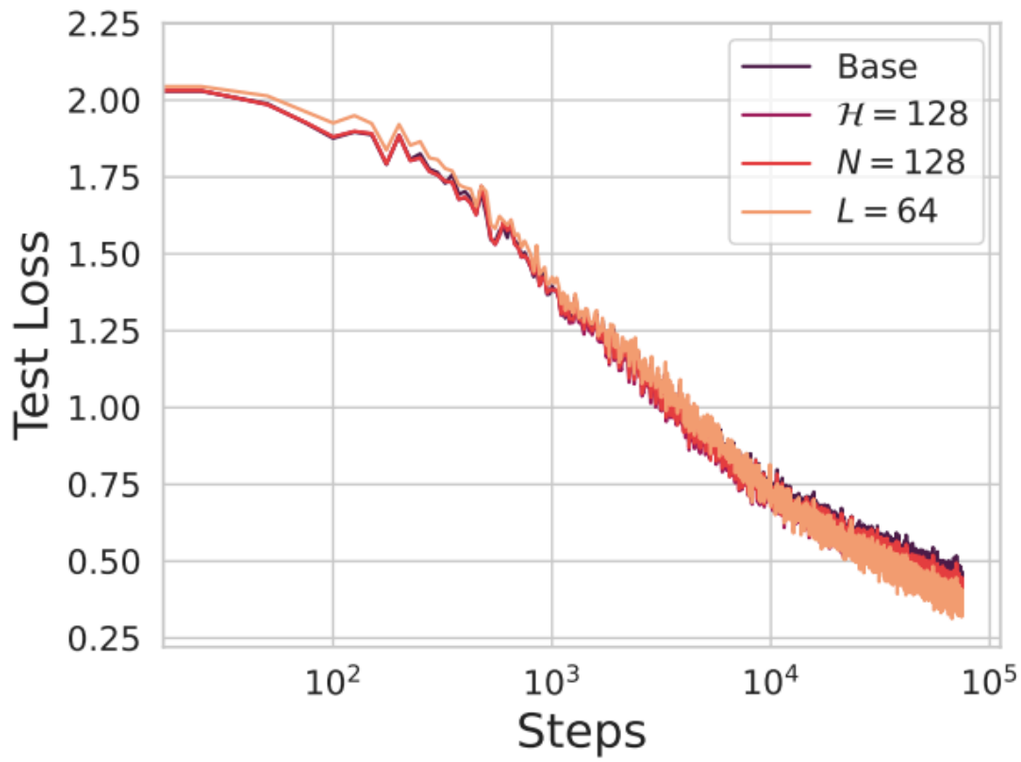
What are models scaling towards?

Claim: *Properly scaled networks converge to universal training dynamics*

Universal = family of sensible joint scalings of randomly initialized LLMs converge to same output dynamics

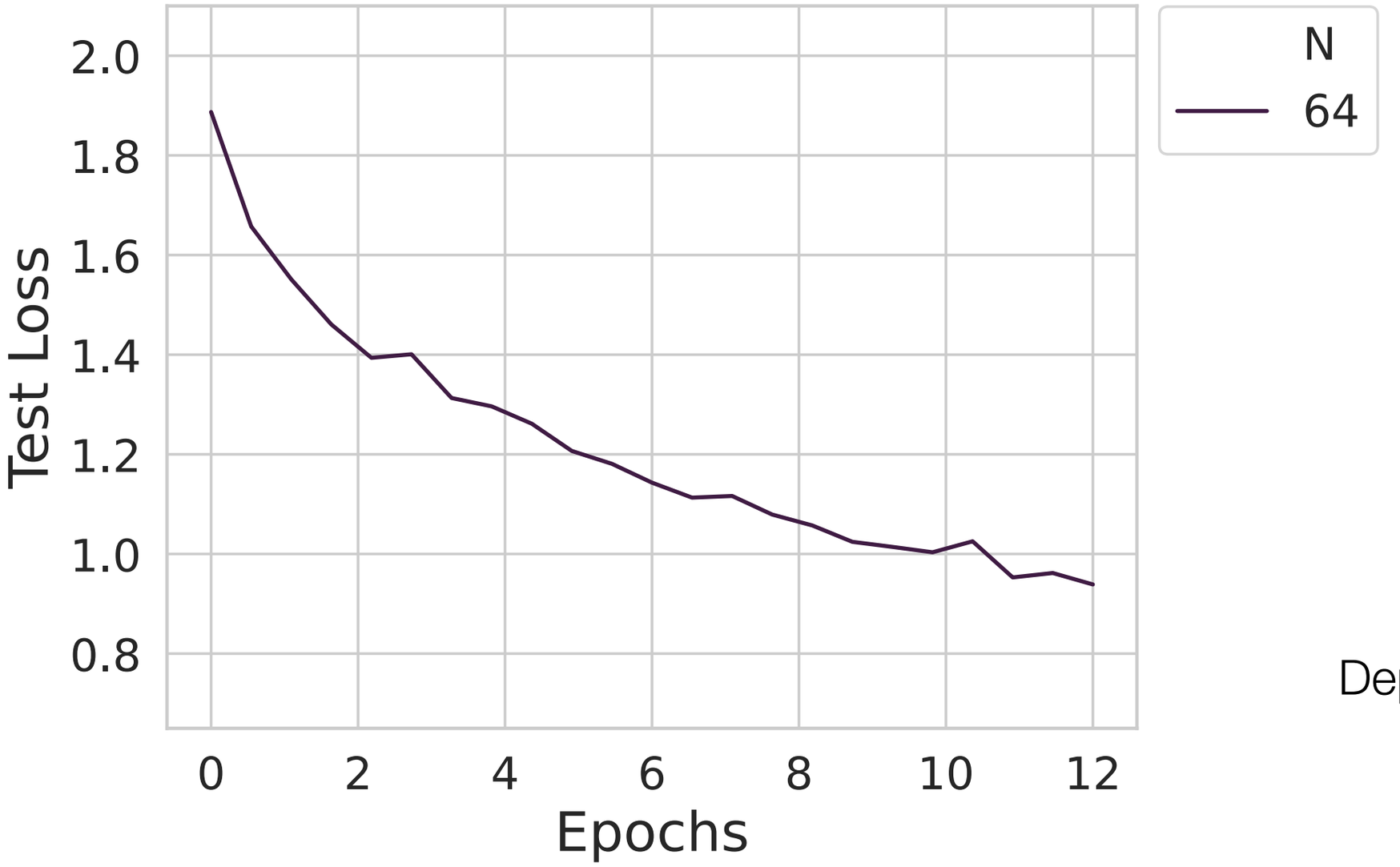
Limit: *The dynamics in this limit are described by an* ***evolving representational level geometry*** *(dynamical similarity kernels)*





Scaling Laws: *Finite size deviations from limiting dynamics generate neural scaling laws*

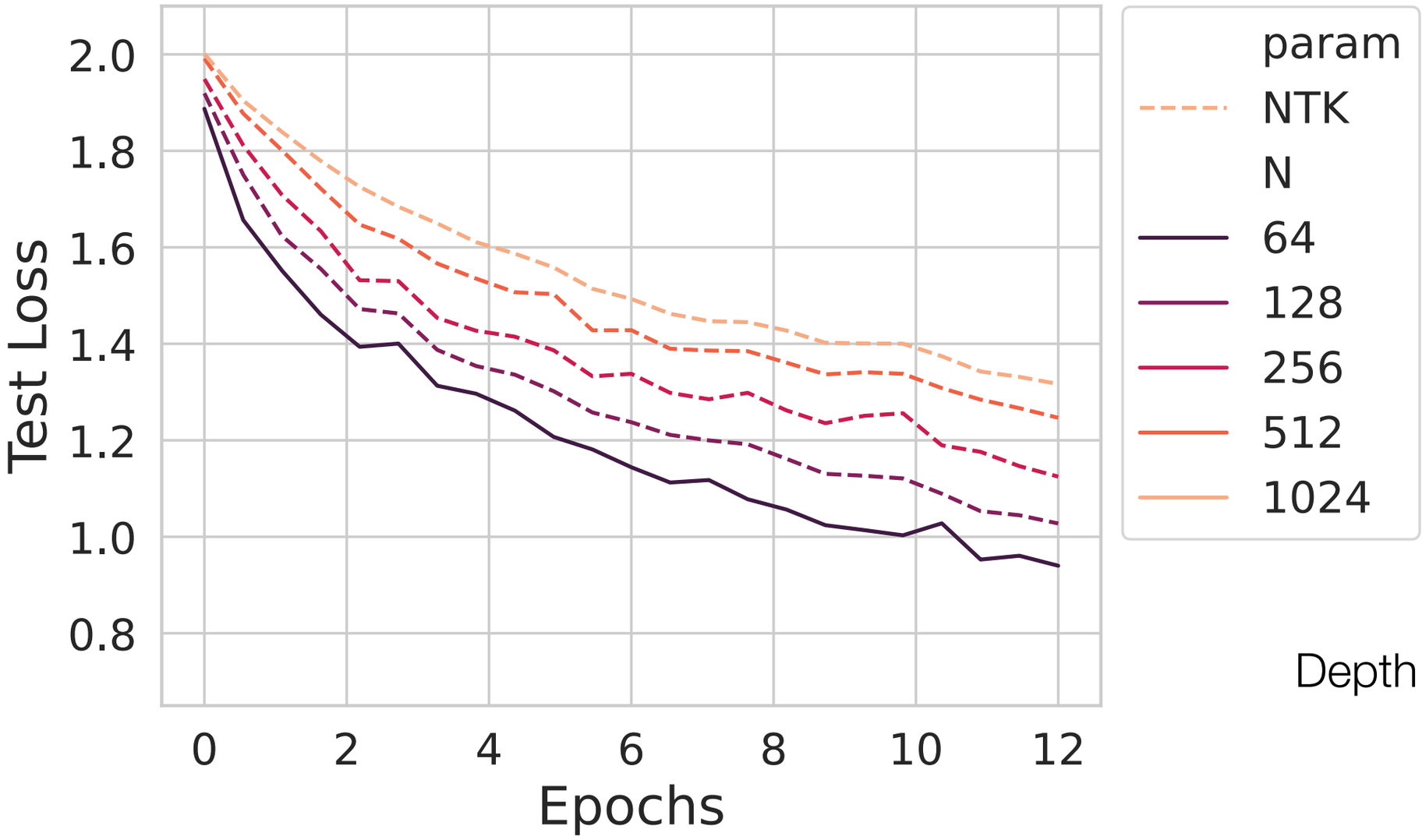
How you Scale Up Matters!



$$f = \frac{1}{\gamma\sqrt{N}} \sum_{j=1}^{N} w_i^L \phi(h_i^L)$$

Depth 12 ResNet on CIFAR-10
SGD training

How you Scale Up Matters!

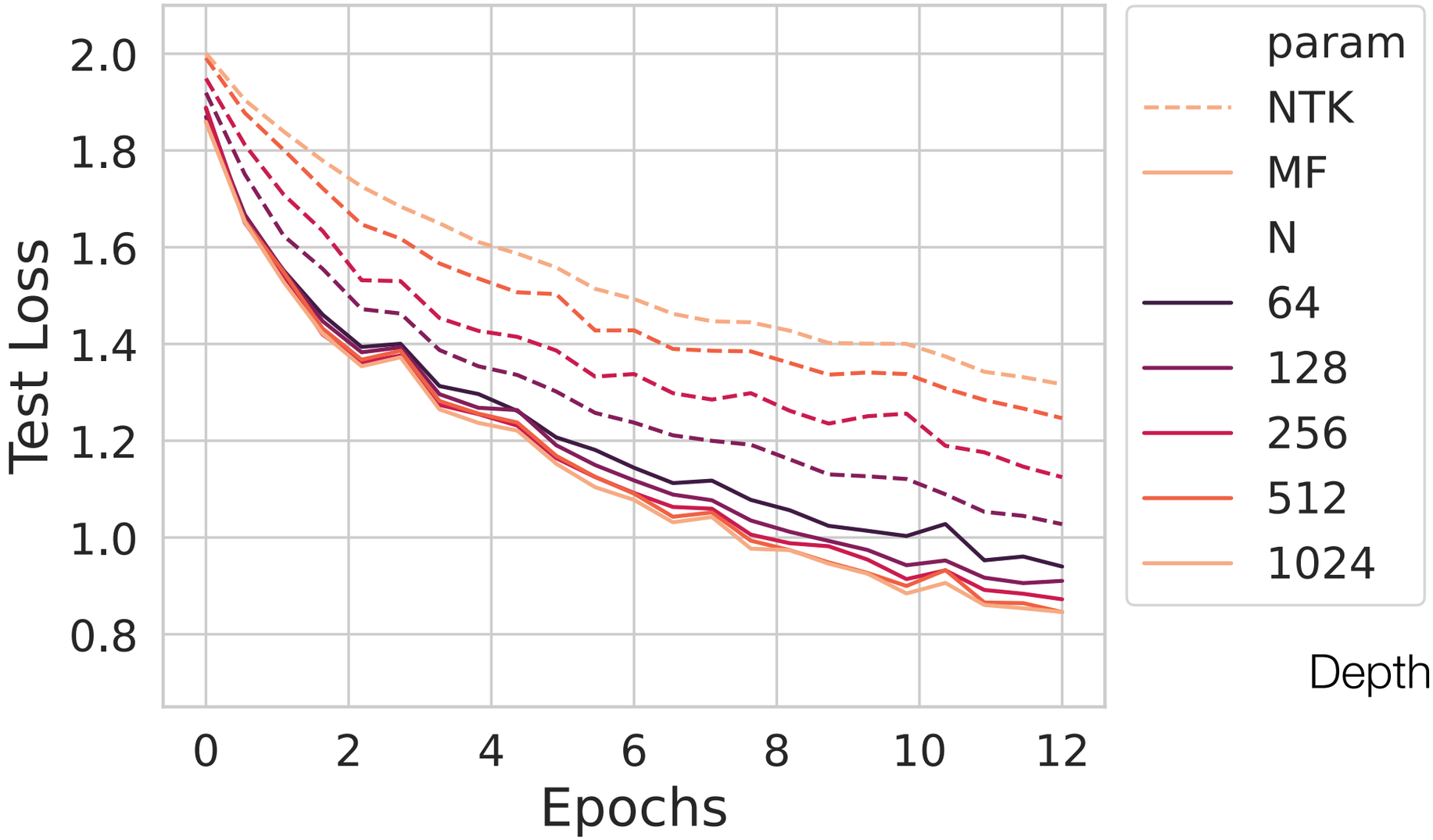


$$f = \frac{1}{\gamma\sqrt{N}} \sum_{j=1}^{N} w_i^L \phi(h_i^L)$$

Depth 12 ResNet on CIFAR-10
SGD training

Increasing width in NTK param $\gamma = \mathcal{O}(1)$ $\Longrightarrow$ Worse performance

How you Scale Up Matters!



$$f = \frac{1}{\gamma\sqrt{N}} \sum_{j=1}^{N} w_i^L \phi(h_i^L)$$

Depth 12 ResNet on CIFAR-10
SGD training

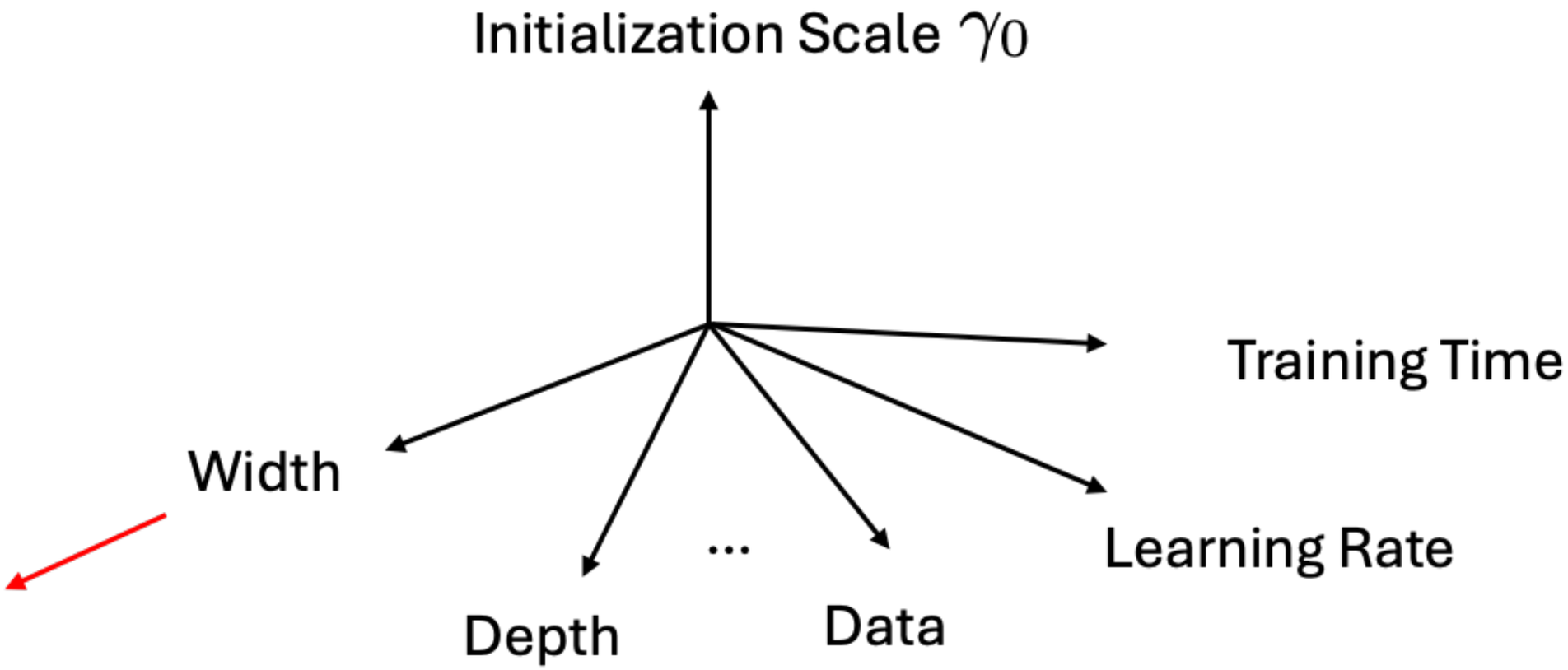
Common scaling practice: $\gamma = \mathcal{O}(1)$ $\implies$ Slower training as N increases

Increasing width based on mean-field theory $\gamma = \mathcal{O}\left(\sqrt{N}\right)$ $\implies$ Similar performance

Mean field also displays faster convergence to limiting behavior

Proposal: study this scaling rule for infinite width networks! $\gamma = \gamma_0\sqrt{N}$

Large Width Scaling Limits

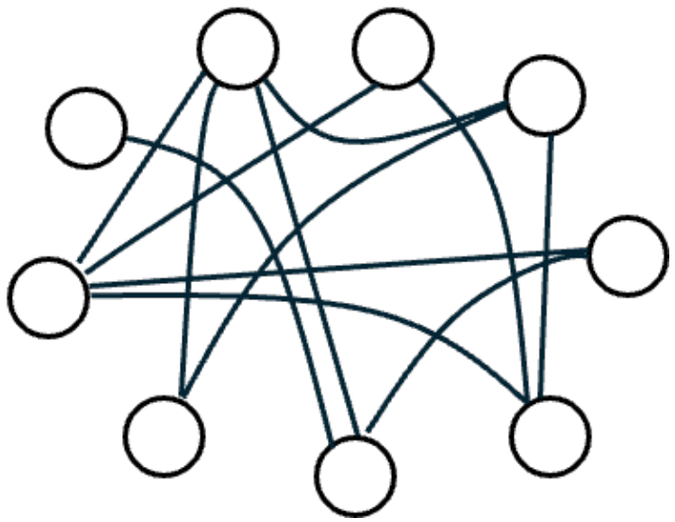


To take this limit, we need some inspiration from physics… brief detour

Primer on Dynamical Mean Field Theory

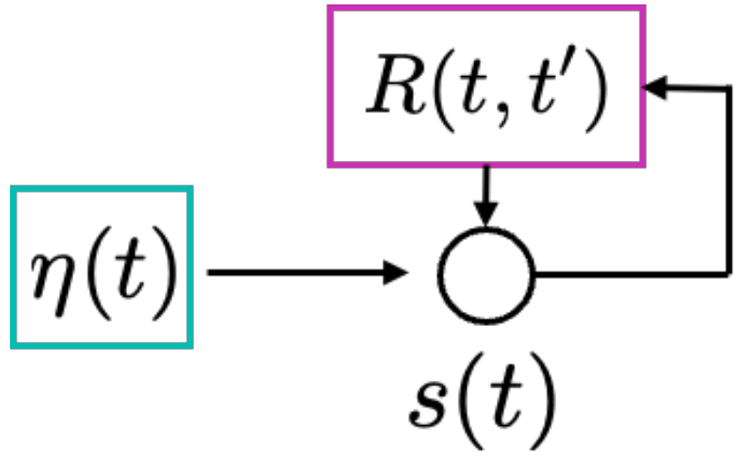
Random Coupled Dynamical System -> Uncoupled System in the Limit

$$\partial_t s_i(t) = \frac{1}{\sqrt{N}} \sum_{j=1}^{N} J_{ij} s_j(t) - \lambda(t) s_i(t) + j_i(t) \qquad J_{ij} = J_{ji} \sim \mathcal{N}(0,1)$$



$$\sim$$

$$N \to \infty$$



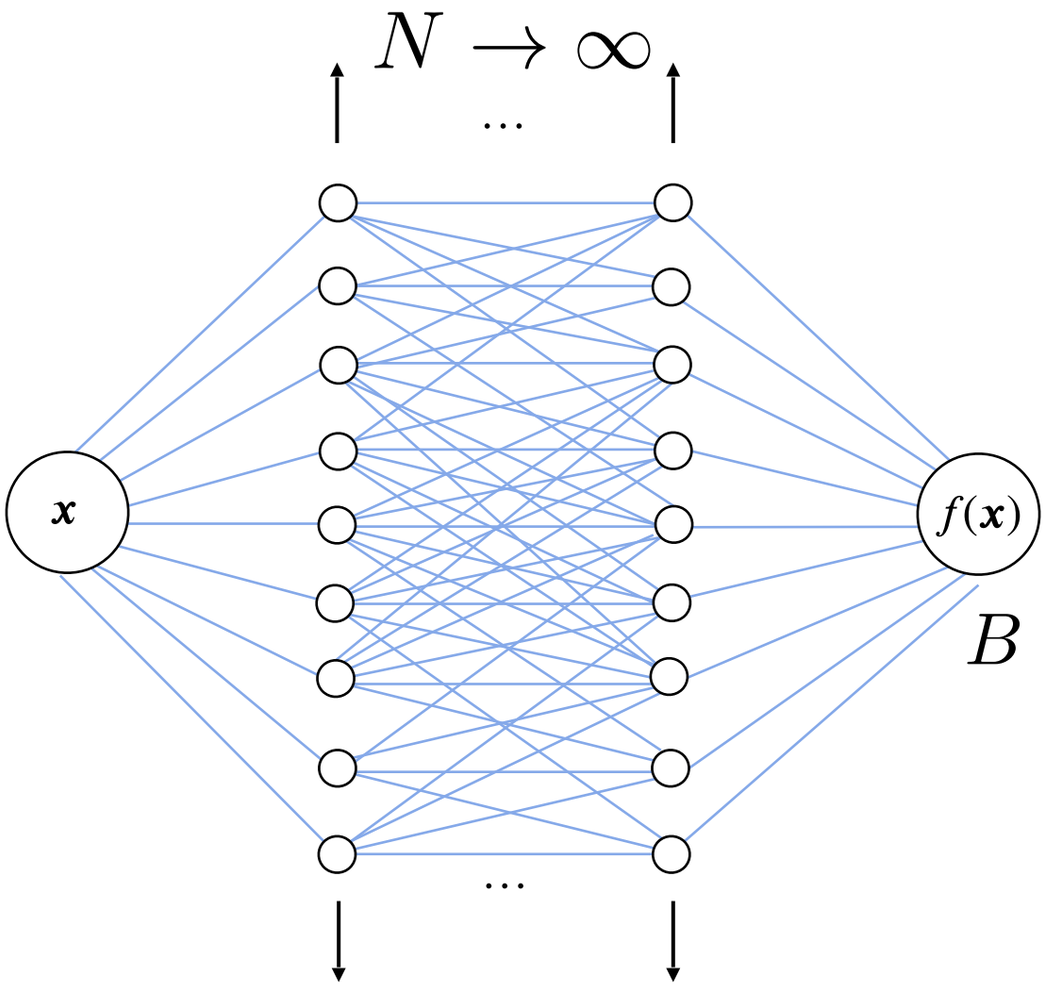
All sites decouple: effectively a one dimensional stochastic process (dynamical mean field)

$$\partial_t s(t) = -\lambda(t) s(t) + \underbrace{\eta(t)}_{\text{colored noise}} + \underbrace{\int dt' R(t,t') s(t')}_{\text{memory term}}$$

Correlation and Response Form Closed System from Single Site Picture

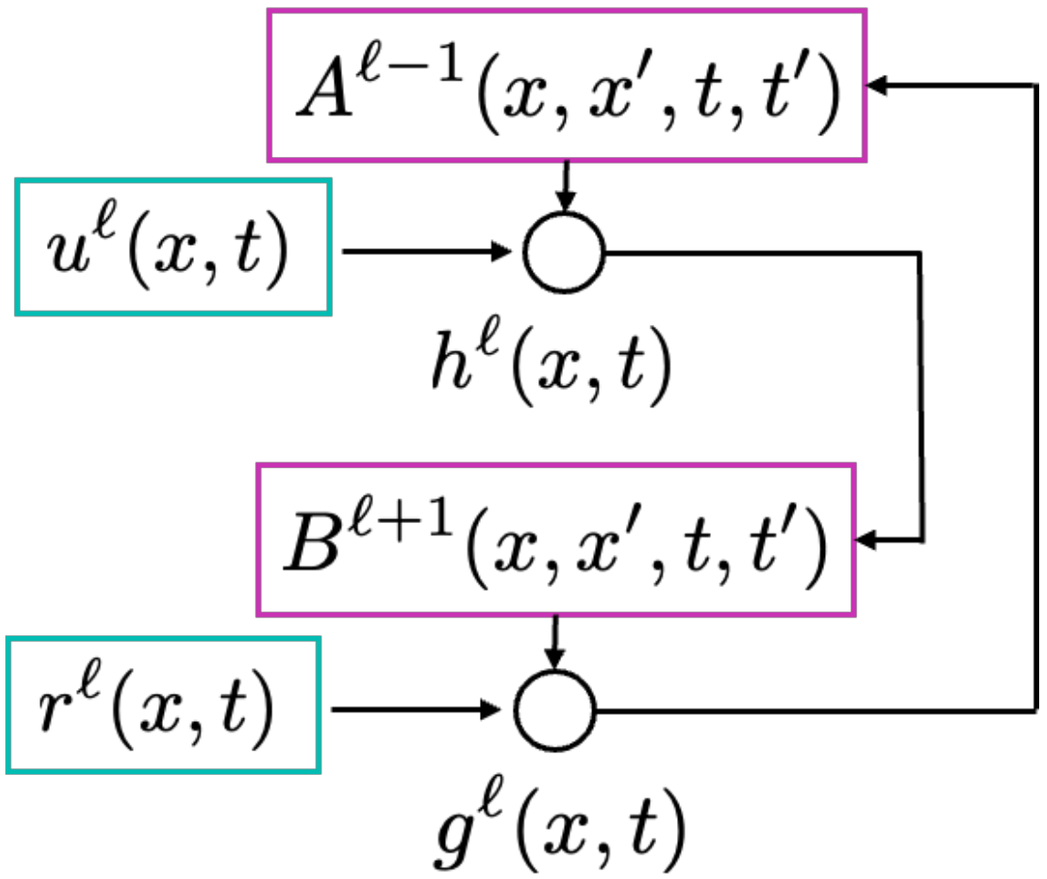
$$C(t,t') = \langle s(t) s(t') \rangle \qquad R(t,t') = \left\langle \frac{\delta s(t)}{\delta \eta(t')} \right\rangle \qquad \eta(t) \sim \mathcal{GP}(0, C(t,t'))$$

Mean Field Theory for Deep Network Training



$$\sim$$

$$N \to \infty$$
$$B \times T = \mathcal{O}(1)$$



Gradient fields: $\mathbf{g}^\ell(x,t) \equiv N\gamma_0 \frac{\partial f(x,t)}{\partial \mathbf{h}^\ell(x,t)}$

Correlation and Response: As $N \to \infty$ learning dynamics completely summarized by

Dynamical Feature kernels

$$\Phi^\ell(x, x', t, t') = \left\langle \phi(h^\ell(x,t))\phi(h^\ell(x',t')) \right\rangle$$

Gradient kernels

$$G^\ell(x, x', t, t') = \left\langle g^\ell(x,t) g^\ell(x',t') \right\rangle$$

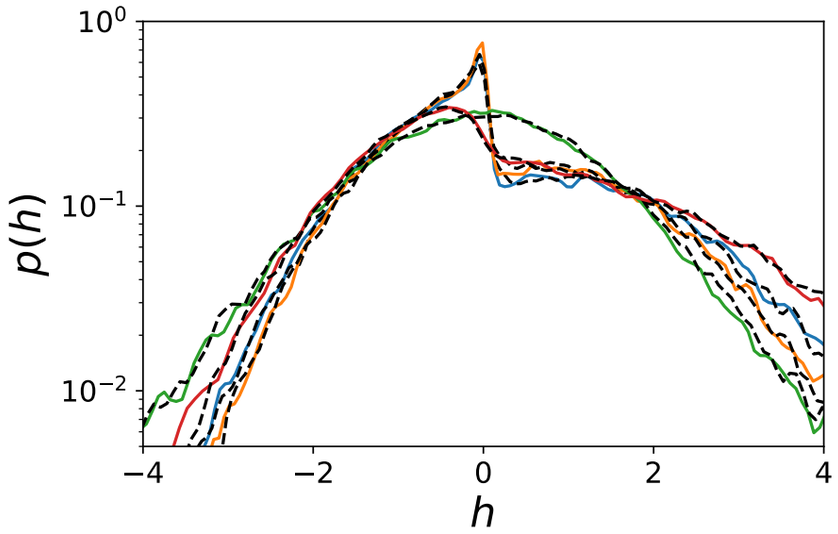
B & Pehlevan ’22, ’23, ’24

$$A^\ell(x, x', t, t') = \left\langle \frac{\delta \phi(h^\ell(x,t))}{\delta r^\ell(x',t')} \right\rangle \quad B^\ell(x, x', t, t') = \left\langle \frac{\delta g^\ell(x,t)}{\delta u^\ell(x',t')} \right\rangle$$

Saddle Point Equations (the $N \to \infty$ limit)

Single-Site Dynamics: Each neuron is independent & follows a single-site stochastic process

$$p(\mathbf{h}^\ell) \sim \prod_{i=1}^{N} p(h_i^\ell)$$



$$h^\ell(x,t) = \underbrace{u^\ell(x,t)}_{\text{Gaussian Process}} + \underbrace{\gamma_0 \mathbb{E}_{x'} \int_0^t ds \left[A^{\ell-1}(x,x',t,s) + p(x')\Delta(x',s')\Phi^{\ell-1}(x,x',t,s)\right] g^\ell(x',s)}_{\text{Feature Learning Correction}}$$

Correlation Functions: Averages over neurons replaced with averages over this process

Correlation functions (kernels):

$$\Phi^\ell(x,x',t,s) = \left\langle \phi(h^\ell(x,t))\phi(h^\ell(x',s)) \right\rangle$$

$$G^\ell(x,x',t,t') = \left\langle g^\ell(x,t) g^\ell(x',t') \right\rangle$$

Output Dynamics: The outputs of the network evolve in terms of these correlation functions

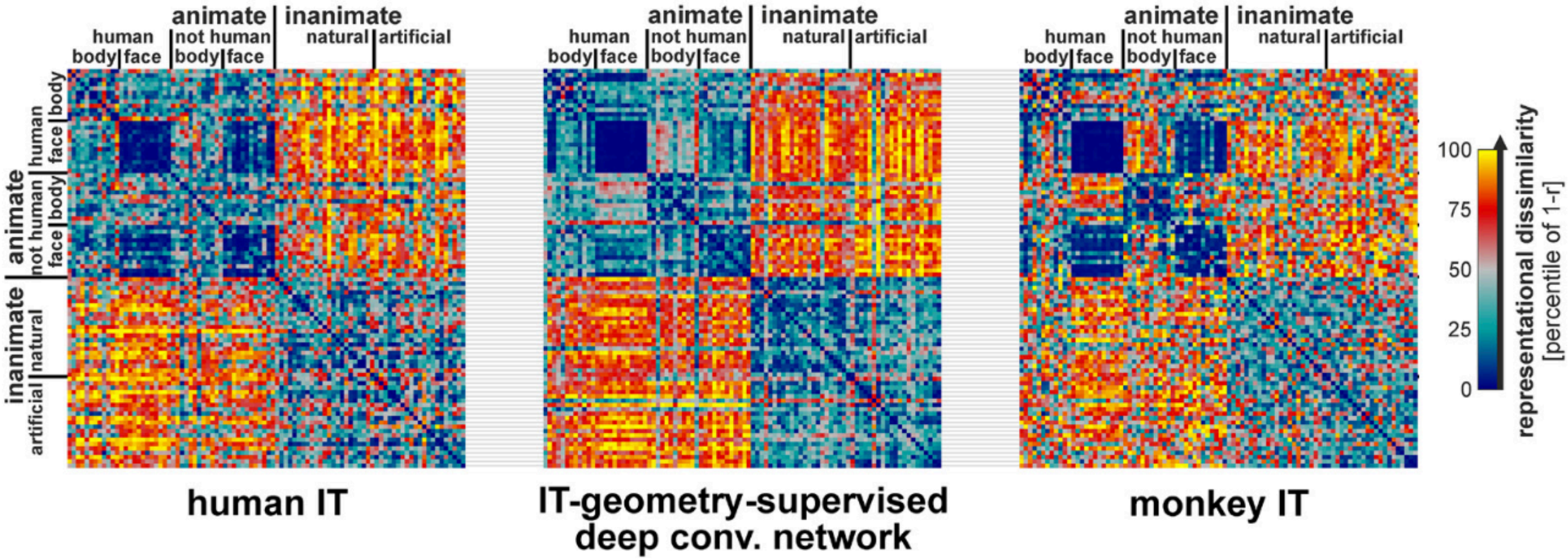
$$\frac{d}{dt} f(x,t) = -\mathbb{E}_{x'} \sum_{\ell=1}^{L} G^{\ell+1}(x,x',t)\Phi^\ell(x,x',t) \frac{\partial \mathcal{L}}{\partial f(x',t)}$$

B, Pehlevan ‘22

Feature Kernels Are Natural "Order Parameters" for Brain and Machines

Representational Similarity Matrices (Kernels) in Cortex and CNNs

$$\Phi^{\ell}(x, x') = \frac{1}{N}\phi(\mathbf{h}^{\ell}(x)) \cdot \phi(\mathbf{h}(x'))$$



(Adapted from Khaligh & Kriegeskorte 2014)

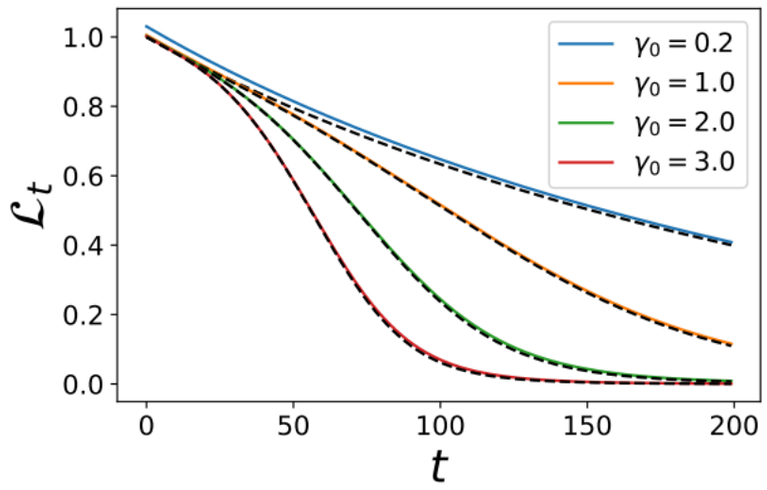
DMFT Implications: *may need to understand time x time correlations as well!*

Summary statistics of learning link changing neural representations to behavior

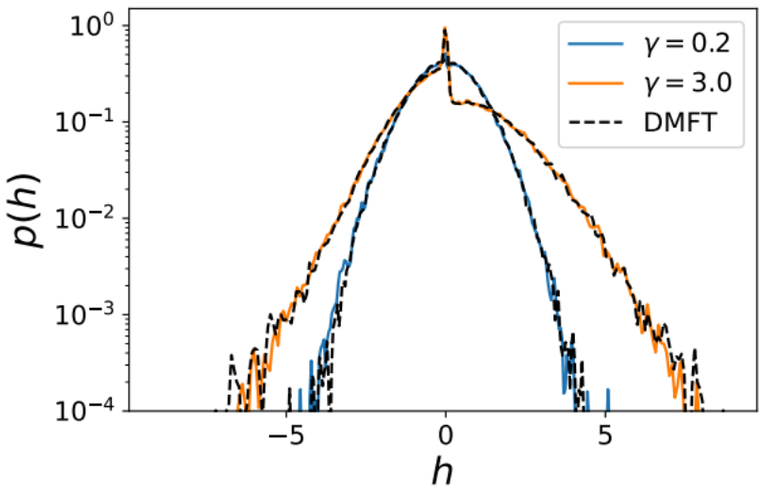
Jacob A. Zavatone-Veth, Blake Bordelon, Cengiz Pehlevan

Lazy vs Rich Operating Regimes

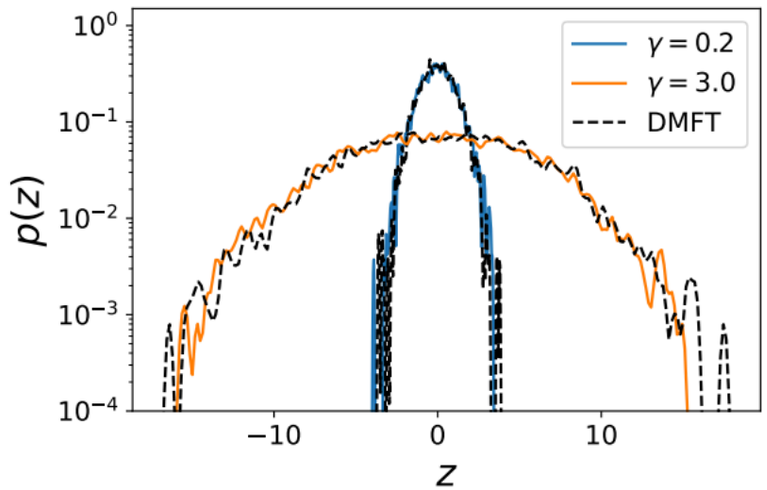
Richness: Infinite width equations depend crucially on an output multiplier γ_0



Loss Dynamics



Final h Distribution



Final z Distribution

Kernel / NTK ⟷ Intermediate (random + spikes) ⟷ Neural Collapse

New Weight Space Picture: BBP for μP

Weight Space: Weights are random bulk + *statistically coupled spikes (not the usual BBP)*

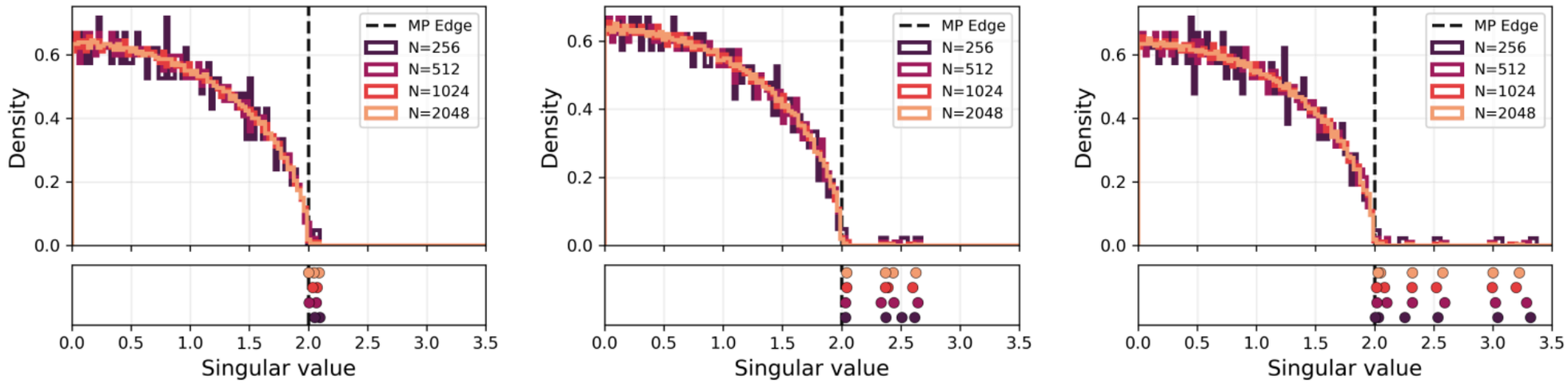
$$\boldsymbol{W}(S) = \boldsymbol{W}(0) + \frac{1}{\sqrt{N_0}} \sum_{t \in [S]} \boldsymbol{g}(t)\boldsymbol{\phi}(t)^\top$$ Spikes S (online training $S = B \times T$)

Spike dynamics: Spikes depend on history of projections onto bulk, iid entries asymptotically

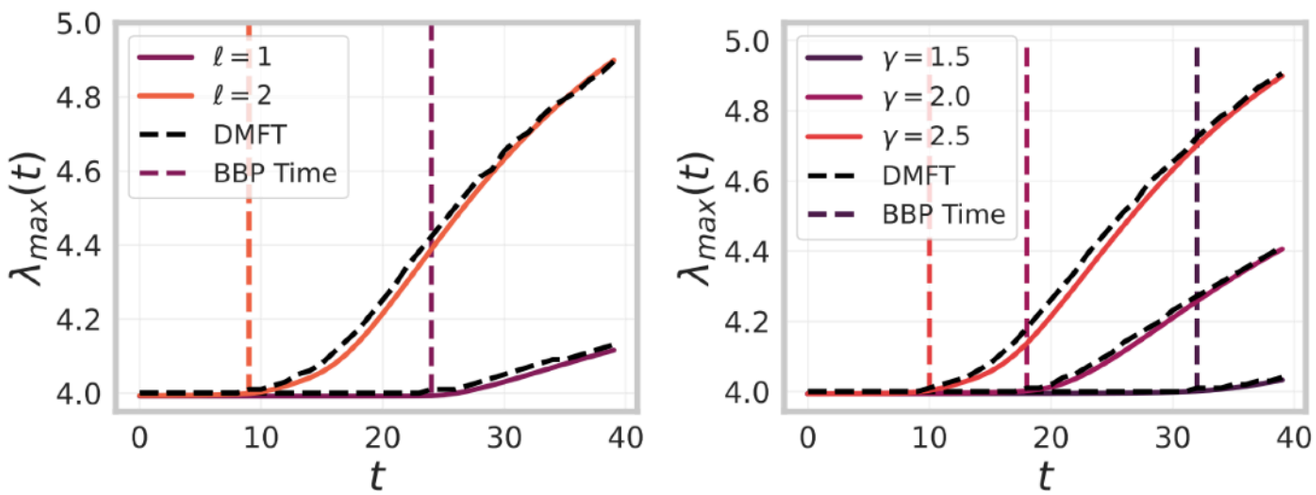
$$\boldsymbol{\chi}_\mu^{\ell+1}(t) = \frac{1}{\sqrt{N}} \boldsymbol{W}^\ell(0)\boldsymbol{\phi}_\mu^\ell(t) \; , \; \boldsymbol{\xi}_\mu^\ell(t) = \frac{1}{\sqrt{N}} \boldsymbol{W}^\ell(0)^\top \boldsymbol{g}_\mu^{\ell+1}(t)$$

$$\boldsymbol{\phi}_\mu^\ell(t) \sim_{N\to\infty} \phi_{\mu,t}^\ell\left(\{\boldsymbol{\chi}_\nu^\ell(s), \boldsymbol{\xi}_\nu^\ell(s)\}_{\nu; s<t}; \gamma\right) \; , \; \boldsymbol{g}_\mu^\ell(t) \sim_{N\to\infty} g_{\mu,t}^\ell\left(\{\boldsymbol{\chi}_\nu^\ell(s), \boldsymbol{\xi}_\nu^\ell(s)\}_{\nu; s<t}; \gamma\right)$$

CIFAR-10:



(a) $T = 100$ (b) $T = 250$ (c) $T = 1000$

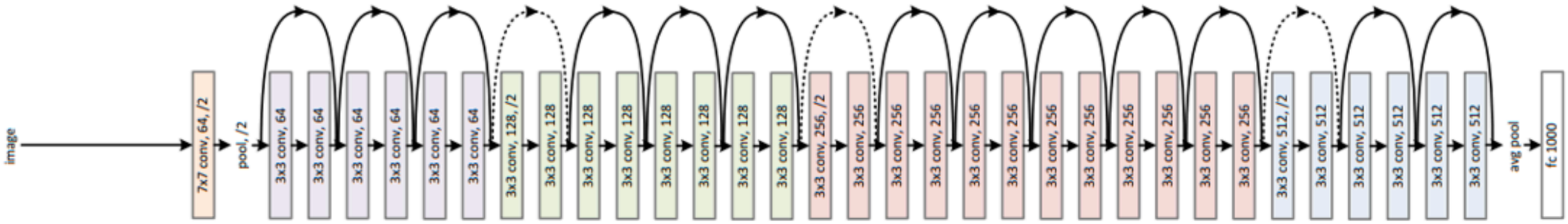


Spectral Dynamics in Deep Networks: Feature Learning, Outlier Escape, and Learning Rate Transfer

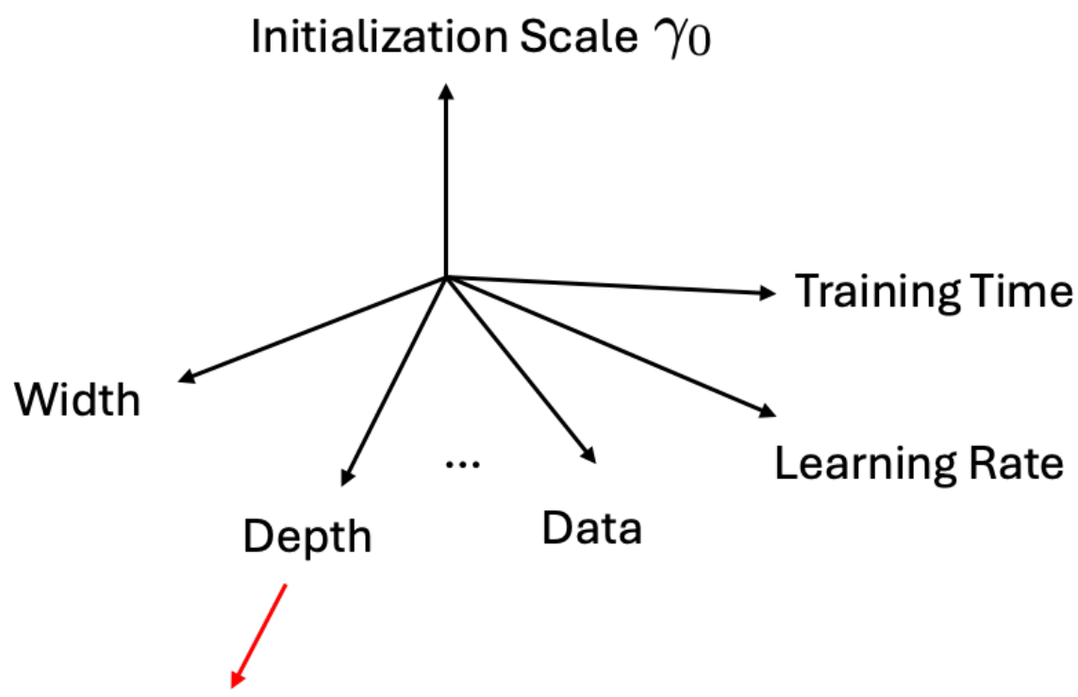
Clarissa Lauditi[1], Cengiz Pehlevan[1,2]*, Blake Bordelon[3,4]*

What about Large Depth?

Practitioners routinely train models with $L \sim 100$ layers (GPT-4 ≈ 120 block layers)



Can we characterize the training dynamics as $L \to \infty$?



Existing common practice does not yield a limit… but for **scaled residual networks**, yes!

$$\mathbf{h}^{\ell+1} = \mathbf{h}^{\ell} + \frac{1}{L^{\alpha}\sqrt{N}}\mathbf{W}^{\ell}\phi(\mathbf{h}^{\ell}) \qquad \alpha \in \left[\frac{1}{2}, 1\right]$$

B*, Noci*, Li, Hanin, Pehlevan, ‘24 Result: a dynamical system across training time and layers!

The Large Depth (SDE) and Width Limit

Solution: Res-Nets with scaled branches $$\mathbf{h}^{\ell+1} = \mathbf{h}^{\ell} + \frac{\beta}{\sqrt{NL}} \mathbf{W}^{\ell} \phi(\mathbf{h}^{\ell})$$

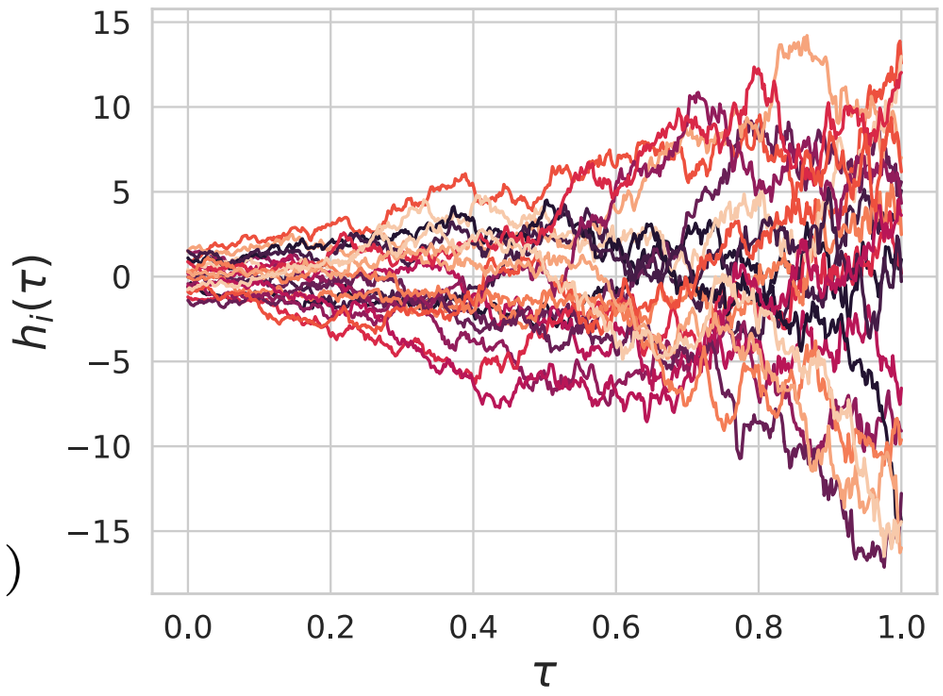
Result: Non-random limit for all DMFT observables $$q_{\infty,\infty} = \lim_{N,L\to\infty} q_{N,L}$$

Intuition pump: characterize initialization

Neurons follow geometric brownian motion

$$H^{\ell} = \left\langle (h^{\ell})^2 \right\rangle$$

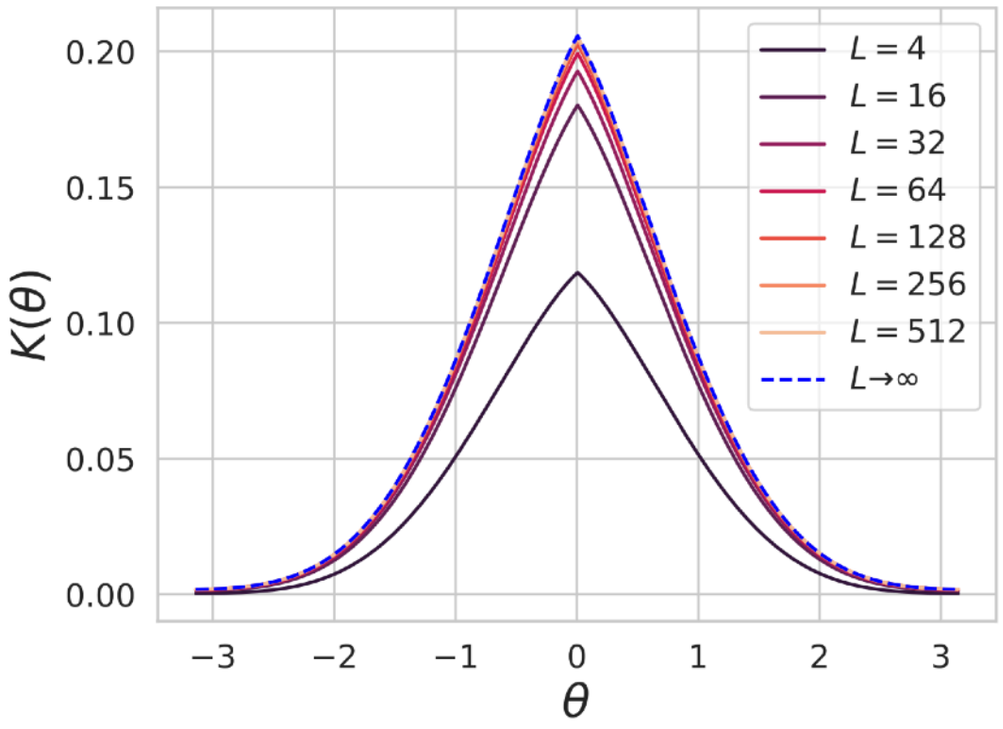
$$H^{\ell+1} = H^{\ell} + \frac{\beta^2}{L} \left\langle \phi(h)^2 \right\rangle_{h\sim\mathcal{N}(0,H^{\ell})}$$



Kernels converge

$$K(\theta) = \frac{1}{L} \sum_{\ell=1}^{L} G^{\ell+1}(x,x') \Phi^{\ell}(x,x')$$

$$\cos(\theta) = \frac{x \cdot x'}{|x||x'|}$$



Feature Learning at Infinite Width and Depth

What happens in the network during training? Start with $\alpha = 1/2$

Averages over a stochastic process in "layer time" $\tau = \dfrac{\ell}{L} \in [0,1]$ and gradient flow time t

$$h(\tau, x, t) = h(0, x, t) + \int_0^\tau du(\tau', x, t)$$

Brownian motion (from initial weights)

$$+\gamma_0 \int_0^\tau d\tau' \int_0^t ds \int dx' \; [A(\tau', x, x', t, s) + \Phi(\tau', x, x', t, s) p(x') \Delta(x', s)] g(\tau', x', s)$$

Feature Learning Corrections

Brownian motion has covariance $\langle du(\tau, x, t) du(\tau', x', s) \rangle = \delta(\tau - \tau') \Phi(\tau, x, x', t, s) d\tau$

Kernels are still single-site averages $\Phi(\tau, x, x', t, s) = \langle \phi(h(\tau, x, t)) \phi(h(\tau, x', s)) \rangle$

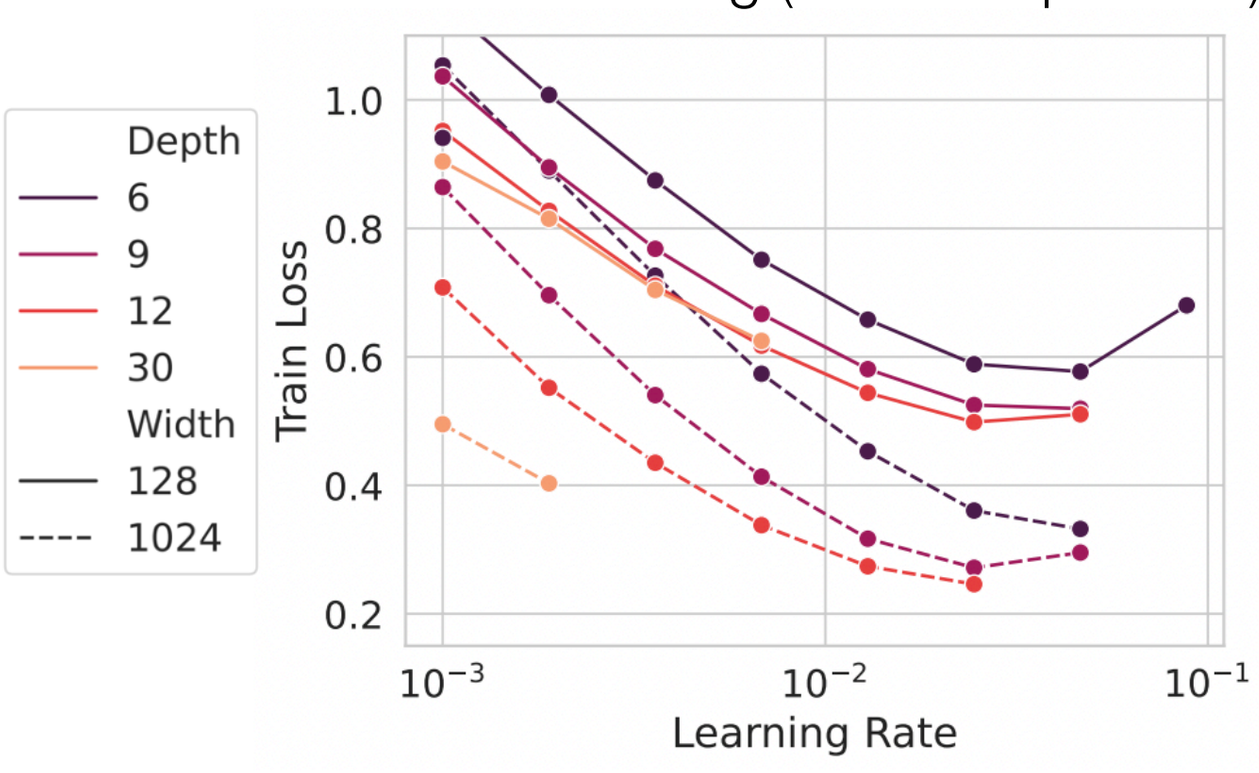
Finite depth networks are Euler–Maruyama approximations of these τ dynamics with step $1/L$

Practical Application: Hyperparameter Transfer

Sweep HPs in small models and then scale up with improved performance (Yang et al 2022)

Possible for both **width and depth** (**B***, Noci*, Li, Hanin, Pehlevan, '24)

No branch scaling (common practice)

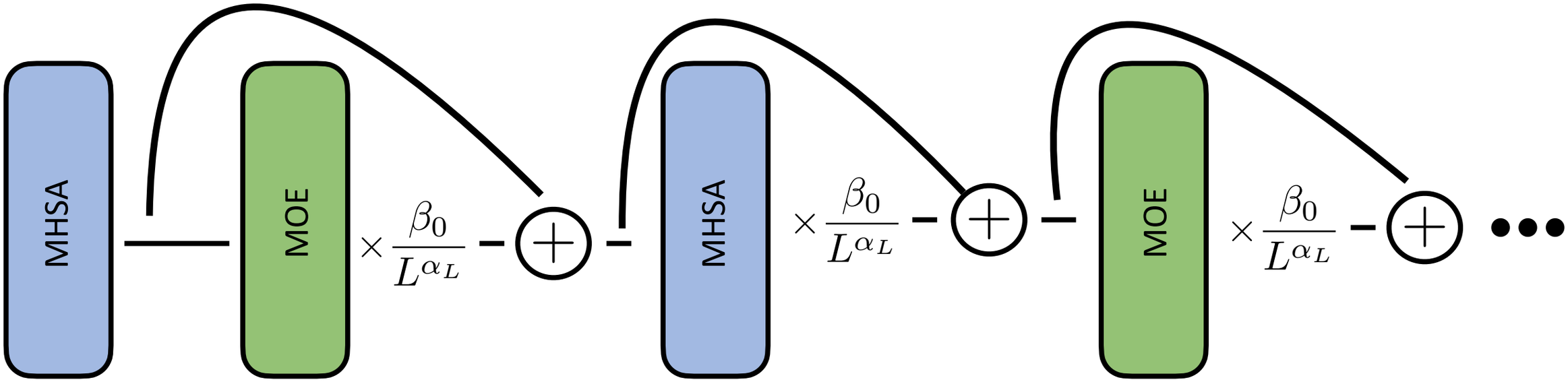


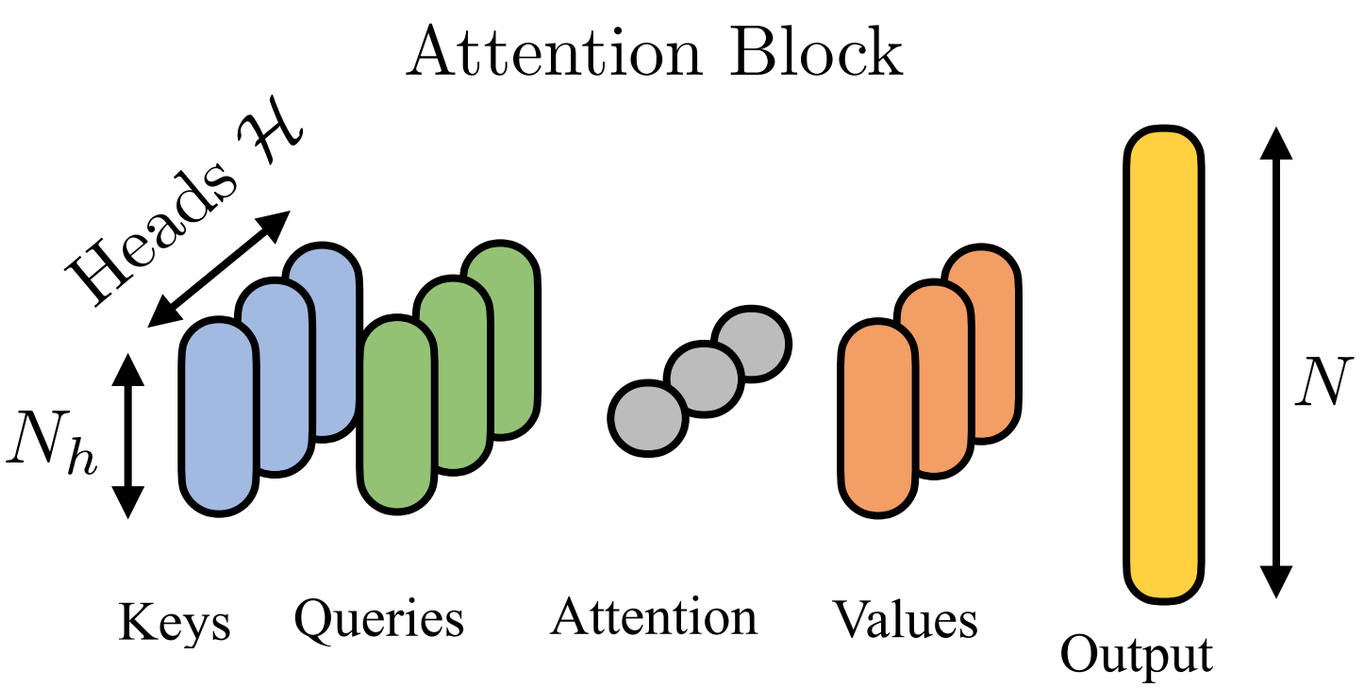
Optimal hyperparams (HP) are not the same for different depths

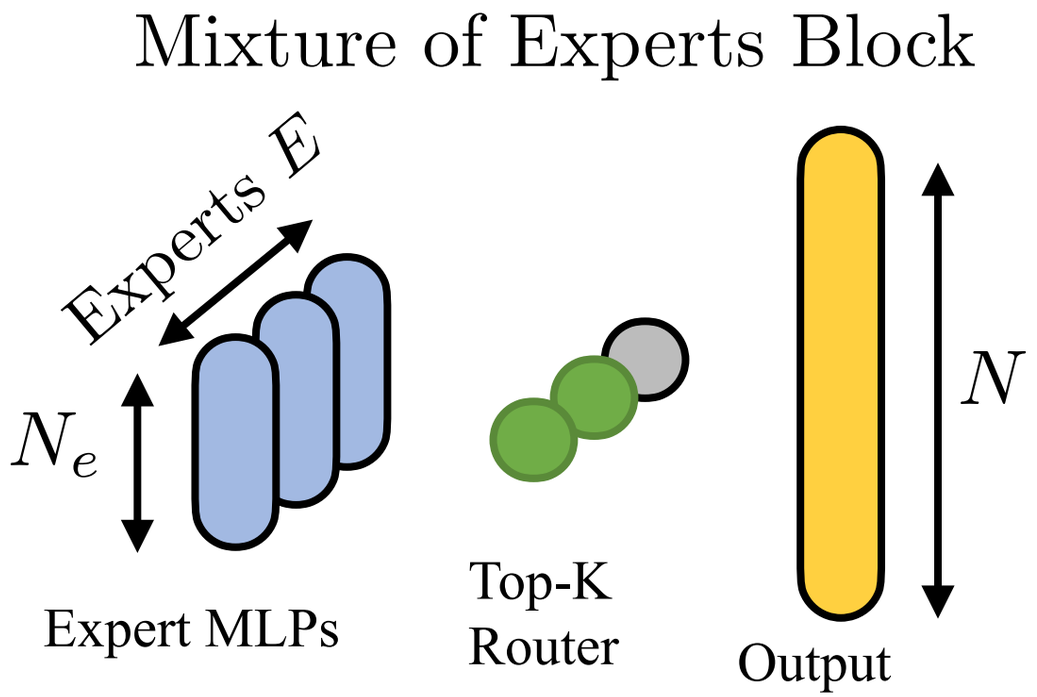
Hyperparameters *transfer* across widths and depths

Can save $ since you only have to do search for good learning rates etc in *small* models

Attention + MOE Layers





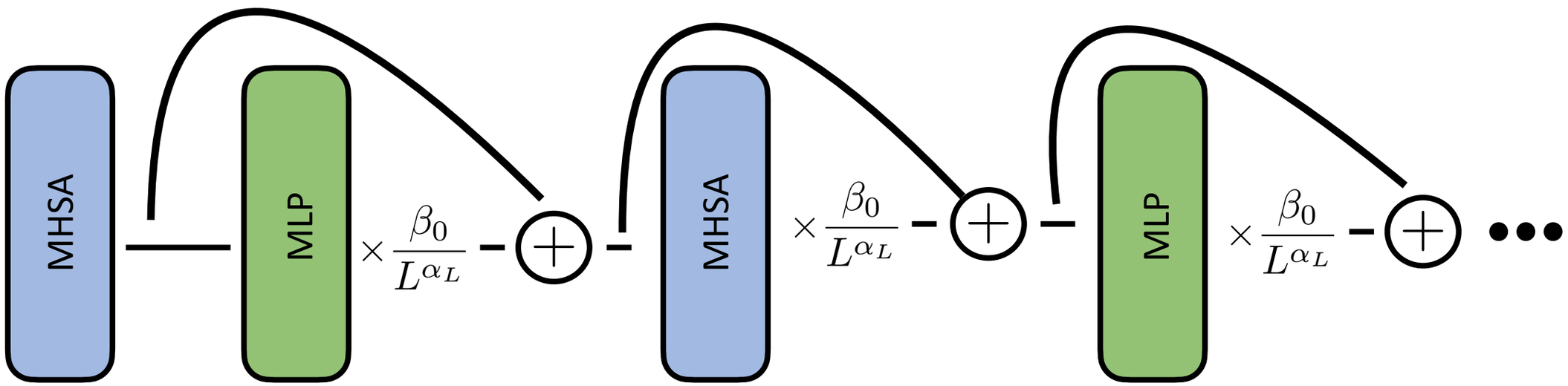


B, Chaudhry, Pehlevan ’24

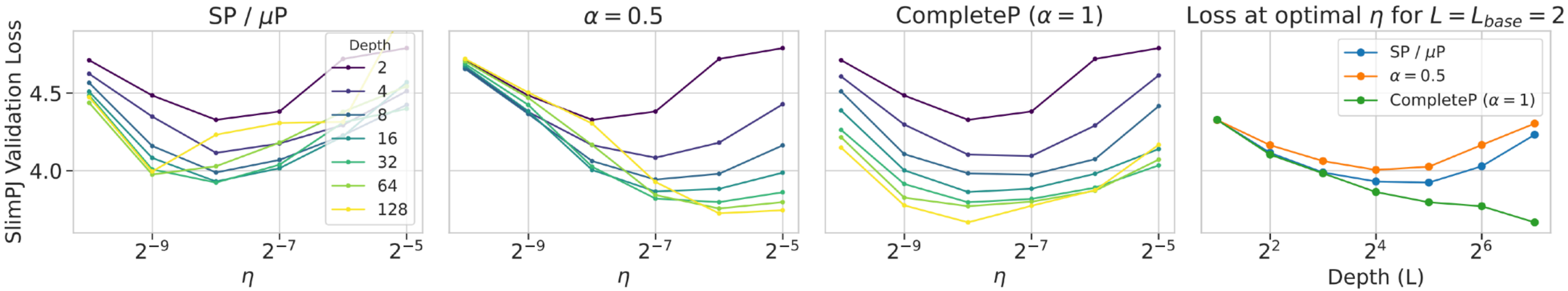
Jiang, **B**, Pehlevan, Hanin ’26

For Transformers: Neural ODE > Neural SDE

How to Choose Depth Exponent?



Well defined feature learning limit for any $\alpha_L \in \left[\frac{1}{2}, 1\right]$ $\quad \delta h^{\ell} \sim \mathcal{O}_L(1)$



(Dey, Zhang, Noci, Li, **B**, Bergsma, Pehlevan, Hanin, Hestness ’25)

Why $\alpha = 1$? *Complete Feature Learning*

More generic residual blocks that depend on θ^ℓ in each layer

$$h^{\ell+1} = h^\ell + \frac{1}{L^\alpha}\,\mathcal{F}(h^\ell, \theta^\ell)$$

Our theory from (**B**, Chaudhry, Pehlevan,'24) indicates that $\theta^\ell(t) - \theta^\ell(0) = \Theta(L^{\alpha-1})$

To retain nonlinearity as $L \to \infty$, we need to take $\alpha = 1$

Neural SDE $\Longrightarrow$ Local Linearization (but NN is globally *not a kernel method*)

$\alpha = \frac{1}{2}$ $\qquad h^{\ell+1} \sim h^\ell + \underbrace{\frac{1}{\sqrt{L}}\mathcal{F}(h^\ell, \theta_0^\ell)}_{\text{Brownian Motion}} + \underbrace{\frac{1}{\sqrt{L}}\nabla_\theta \mathcal{F}(h^\ell, \theta_0^\ell)\cdot(\theta^\ell - \theta_0^\ell)}_{\text{Learning Update}}$

Neural ODE $\Longrightarrow$ Nonlinear, but no Brownian motion

$\alpha = 1$ $\qquad h^{\ell+1} \sim h^\ell + \frac{1}{L}\sum_{k=1}^{\infty} \frac{d^k}{d\theta^k}\mathcal{F}(h^\ell, \theta_0^\ell)(\theta^\ell - \theta_0^\ell)^k$

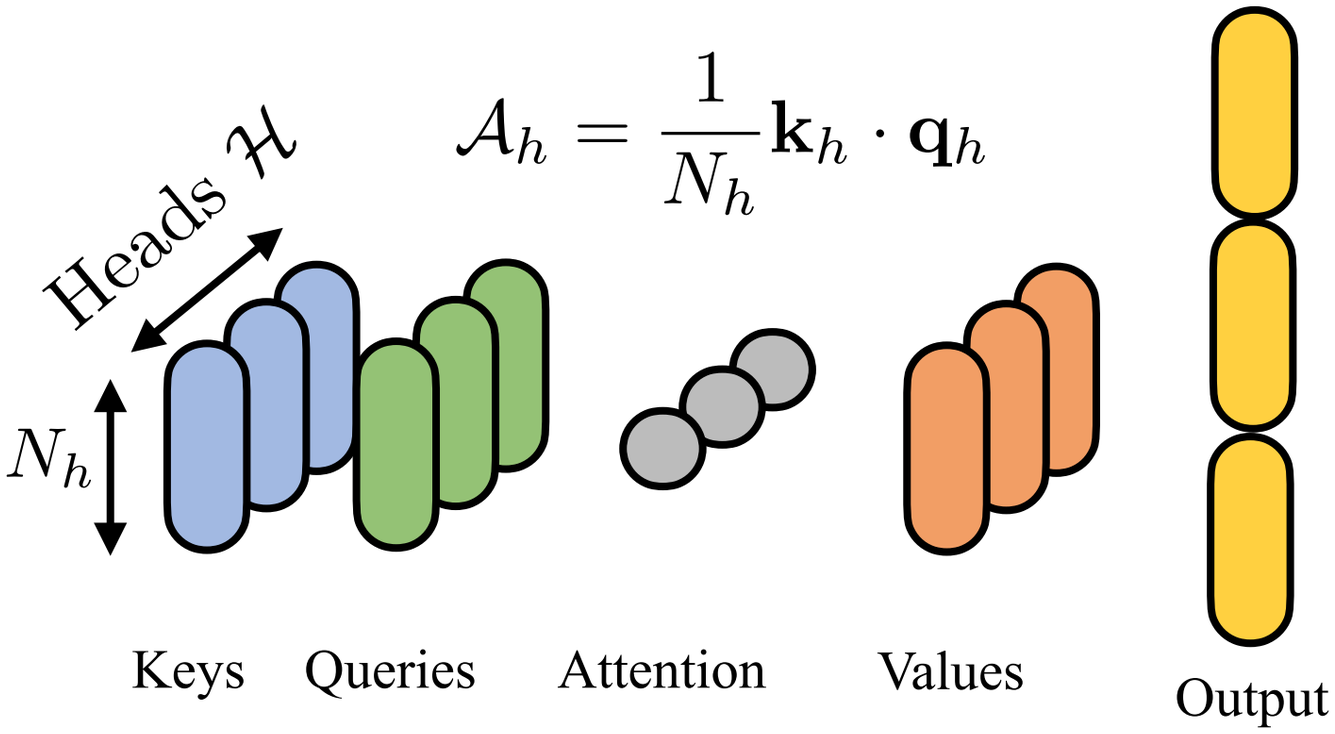
Both "learn features" in sense of $\delta h \sim \Theta(1)$

Complete Feature Learning: don't lose any nonlinearity in any sub-block of the network!

(Dey, Zhang, Noci, Li, **B**, Bergsma, Pehlevan, Hanin, Hestness '25)

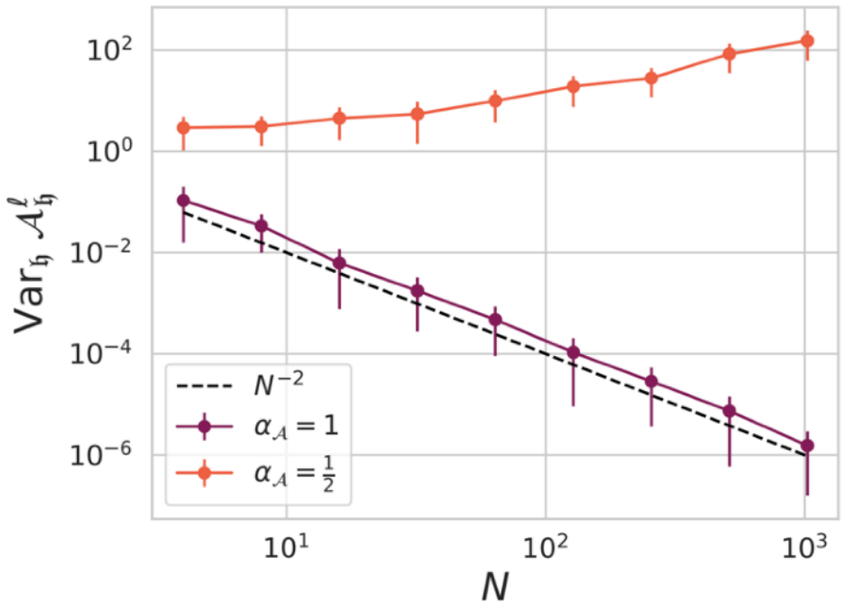
Two Distinct Width Limits of Multi-head Transformers

Dynamics of randomly initialized transformer architectures (**B**, Chaudhry, Pehlevan '24)



Limit 1: Infinite head size $N_h \to \infty$

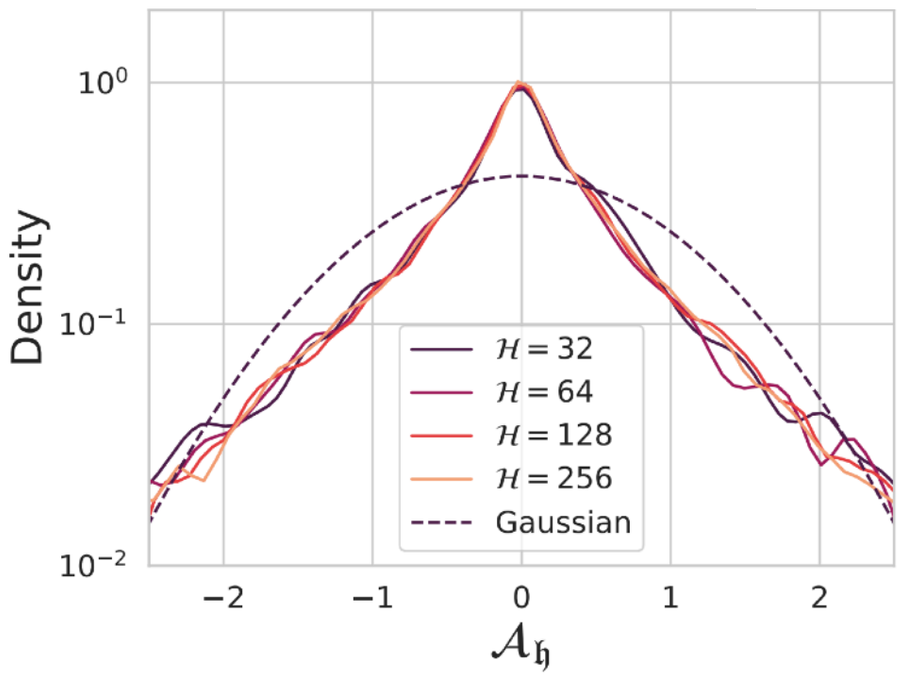
Attention matrices concentrate
(But all collapse to same dynamics!)

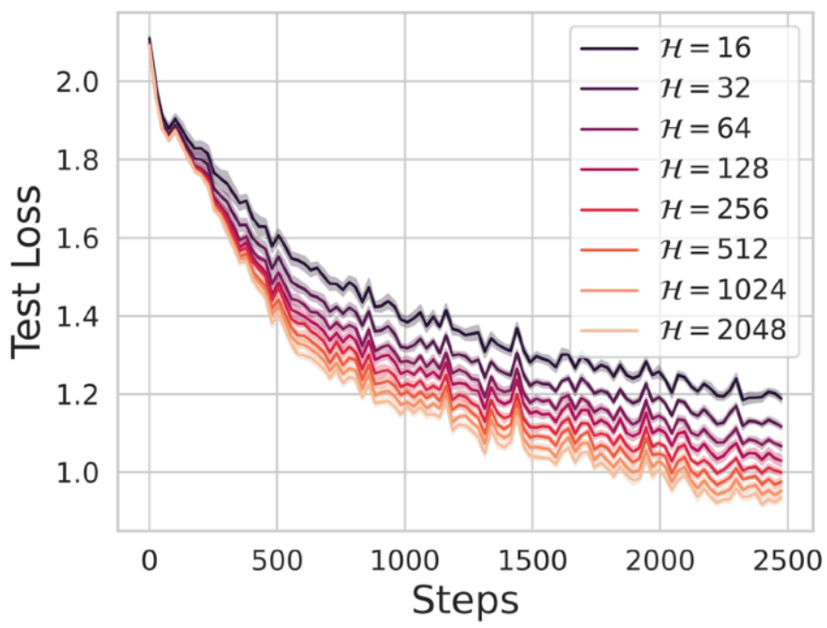


Limit 2: Infinite heads $\mathcal{H} \to \infty$

*Mean field **distribution** over attention heads*

Monotonic improvements with heads

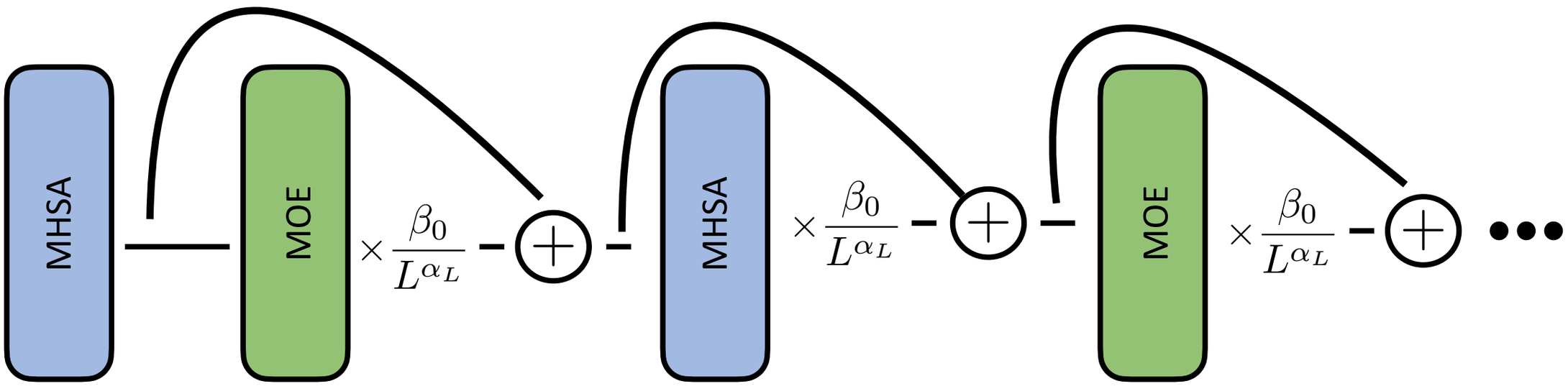




Limit 3: Joint scaling with diversity (open)

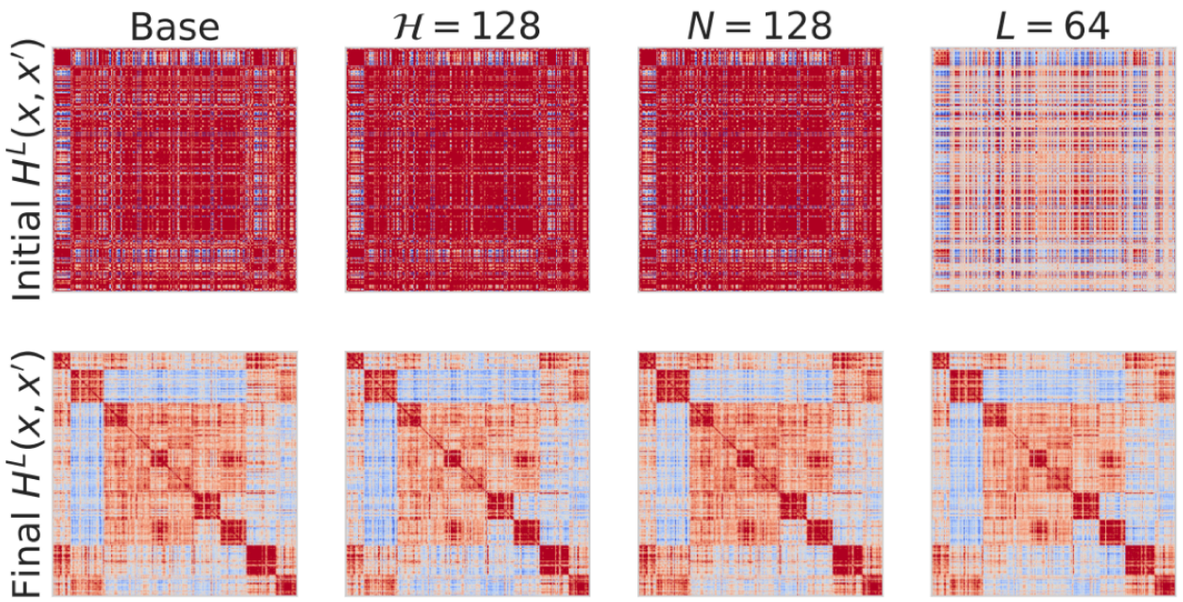
Enormous # of Correlations to Track for Transformers

Depth L Transformer

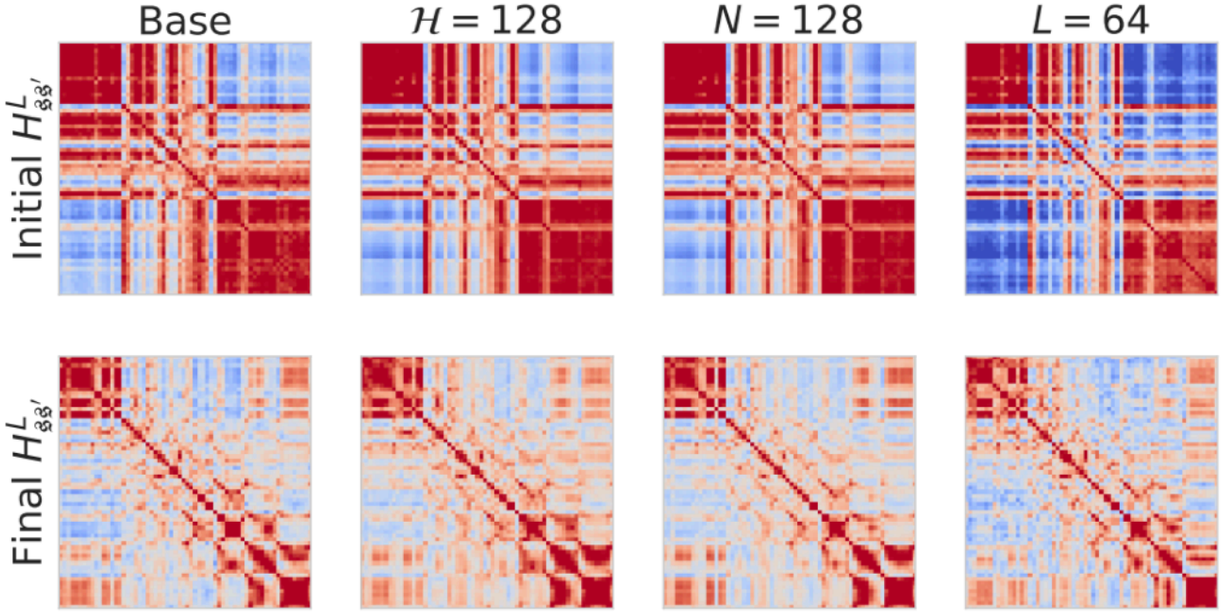


$$C_h = \frac{1}{N}\mathbf{h}_s^\ell(x,t)\cdot\mathbf{h}_{s'}^\ell(x',t') \to \left\langle h_s^\ell(x,t)h_{s'}^\ell(x',t')\right\rangle$$

VIT Data x Data Similarity Structure



VIT Space x Space Similarity Structure



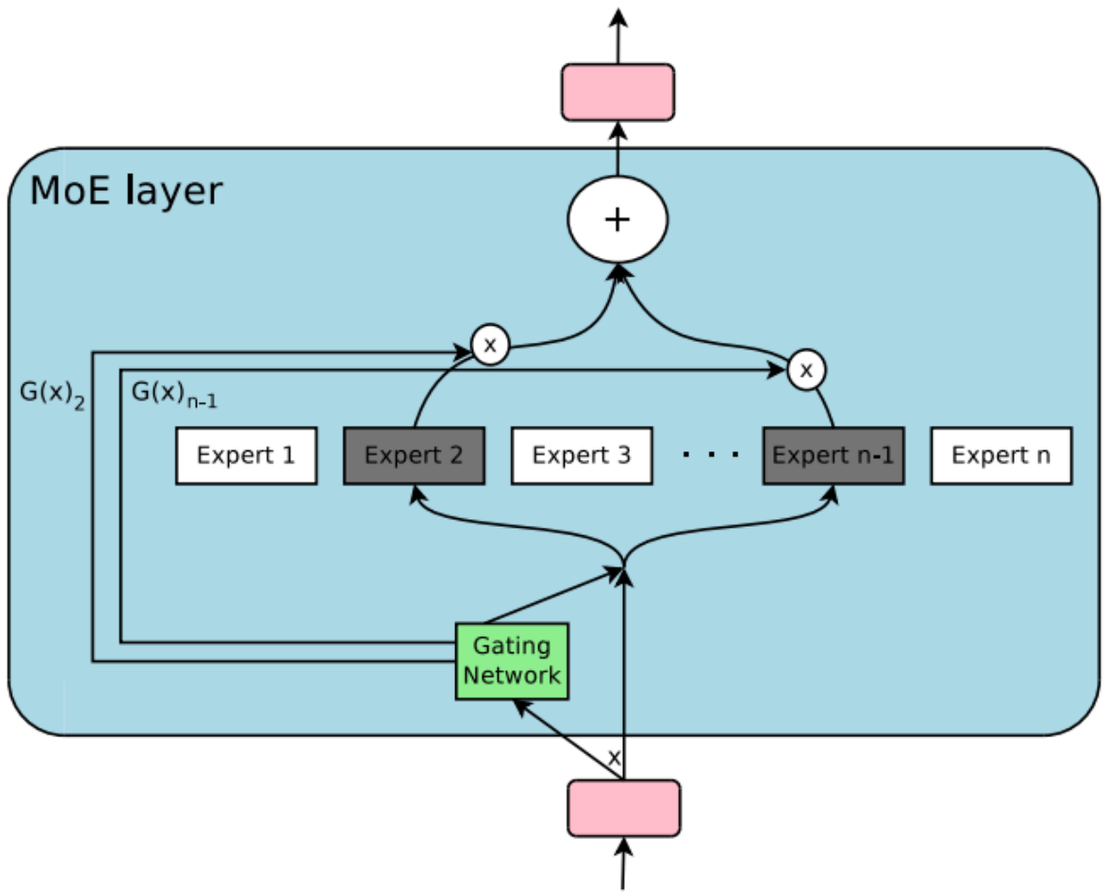
In LLM pretraining this is steps x steps x batch_size x batch_size x seq_len x seq_len

Not practical to run these equations at scale (focus on extracting high level insights)

Why MOE?

Dense transformer computation often dominated by MLP blocks (90% of compute in Palm3)

***MOE Main Idea**: break up the MLP block into experts with a gating/routing mechanism. Each token activates a subset of the experts (Shazeer et al 2017)*

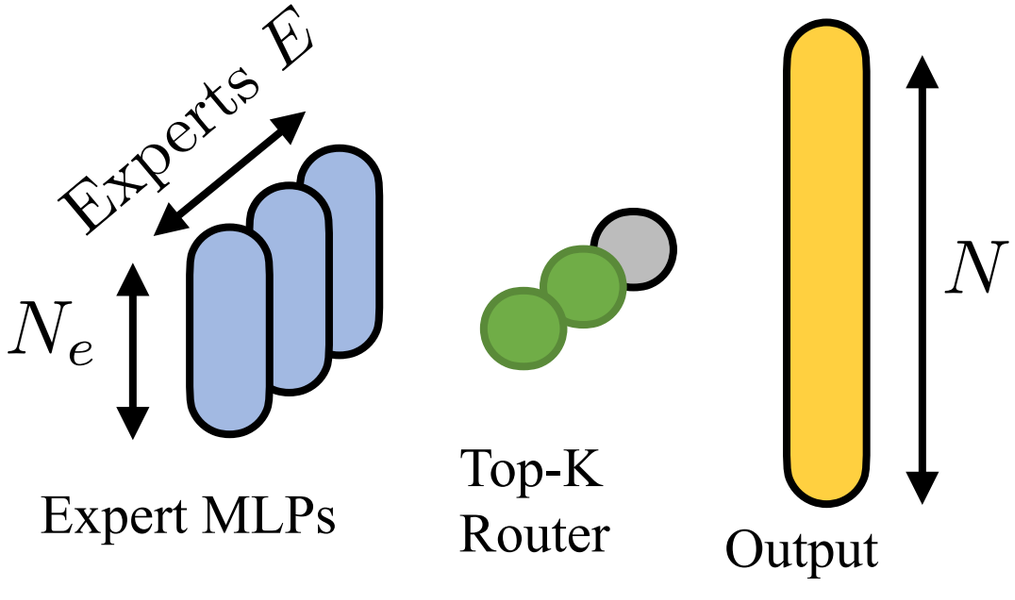


The SOTA LLM models are now all MOE (GPT-4+, Grok-1, Gemini 1.5, Claude Opus, etc)

***Question:** How to stably scale up/down MOE model training, converge to valid limit, and enable HP transfer? What are the relevant scaling variables? How sparse can we make it?*

MOE Blocks

Mixture of Experts Block



Layer Forward Pass

$$\frac{\sqrt{N}}{EN_e}\sum_{k=1}^{E}\underbrace{\Theta(p_k+b_k-q_\star)\sigma(p_k)}_{\text{hard gate}}\mathbf{W}_k^{(2)}\phi\left(\frac{1}{\sqrt{N}}\mathbf{W}_k^{(1)}\mathbf{h}\right)$$

Router Scores $\quad p_k=\frac{1}{N}\mathbf{r}_k\cdot\mathbf{h}$

Router Bias b_k

Top-K Activated Experts:

threshold $q_\star$ chosen so that the top K of the E experts are active

(Shazeer et al 2017, Fedus et al 2022)

Load Balance Goal:$\forall k,\ell$ # tokens @ router k = # tokens @ router ℓ

Common Strategy: add aux. loss to promotes balancing (hard to make this scale, nonlocal, etc)

Update Rule for Bias Encourages Load Balance

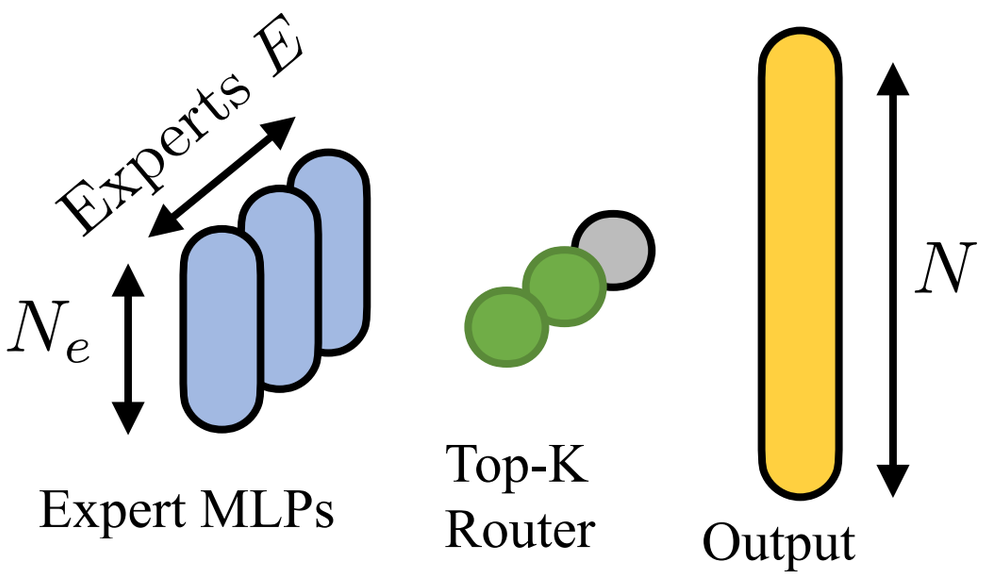
$b_k(t+1)-b_k(t)\propto(K/E-\text{Fraction of Batch Assigned to Expert }k)$

(Liu et al 2024, Dai et al 2024)

MOE Blocks: Three Level Mean Field Limit

We characterize the dynamics of randomly initialized MOE (Jiang, **B**, Pehlevan, Hanin '26)

Mixture of Experts Block



Assume Fixed Sparsity

$$\kappa \equiv K/E = \text{constant}$$

Universal Joint Scaling Condition

$$N_e(N), E(N), L(N) \text{ diverging with } N$$

$$\lim_{N\to\infty} \frac{N}{N_e(N)E(N)L(N)} = 0$$

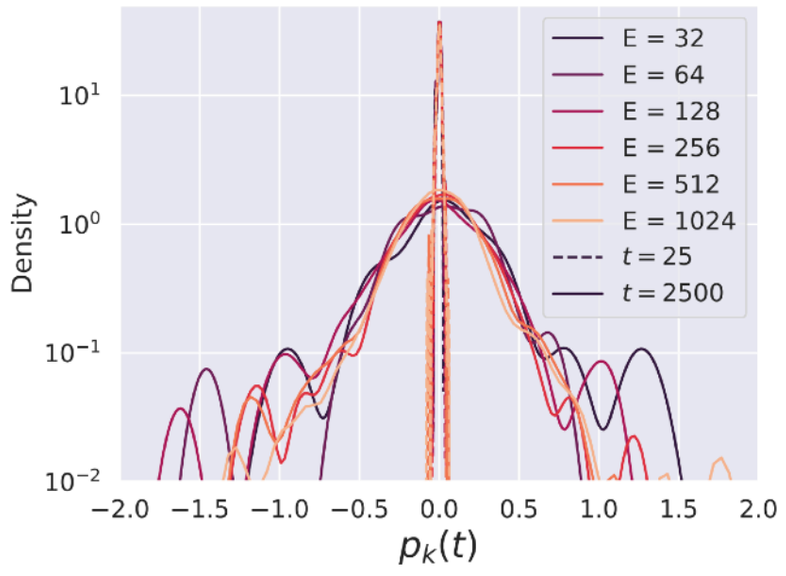
SGD Dynamics in this Limit are Governed by a Three Level DMFT

Level 1: The neurons and kernels on the residual stream

Level 2: The distribution over Expert States $\{p, b, ...\}$

top-K is a quantile threshold

Level 3: The distribution over Within-Expert Neurons



In this limit, the NN outputs only depend on various averages over these levels

MSR Saddle Point Equations

$$\mathcal{S} = \widehat{f}\left(f - \Phi^L \Delta - R^L_{h\xi}\right) - \widehat{R}^L_{h\xi} R^L_{h\xi} + \sum_\ell \left[C^\ell_h \widehat{C}^\ell_h + C^\ell_g \widehat{C}^\ell_g\right]$$

$$+ \nu \sum_\ell \left[\widehat{M}^\ell_{\sigma\sigma C_\phi} M^\ell_{\sigma\sigma C_\phi} + \widehat{M}^\ell_{\sigma\sigma C_g} M^\ell_{\sigma\sigma C_g} + \widehat{M}^\ell_{\dot\sigma\dot\sigma\mathcal{A}\mathcal{A}} M^\ell_{\dot\sigma\dot\sigma\mathcal{A}\mathcal{A}}\right] - \frac{1}{\alpha_\star L} \sum_\ell \left[N_e \widehat{\bar{R}}^\ell_{\phi\xi} \bar{R}^\ell_{\phi\xi} + N_e \widehat{\bar{R}}^\ell_{g\chi} \bar{R}^\ell_{g\chi}\right]$$

$$+ \ln \mathcal{Z}_{\text{res}} + \nu \sum_{\ell=1}^{L} \ln \mathcal{Z}^\ell_{\text{exp}}$$

$$\nu = \frac{E}{N}\,,\ \alpha_\star = \frac{N}{N_e E L}$$

$$\mathcal{Z}_{\text{res}} = \int \prod_\ell \mathcal{D}\bar\chi^\ell \mathcal{D}\widehat{\bar\chi}^\ell \mathcal{D}\bar\xi^\ell \mathcal{D}\widehat{\bar\chi}^\ell \mathcal{D}h^\ell \mathcal{D}\widehat{h}^\ell \mathcal{D}g^\ell \mathcal{D}\widehat{g}^\ell \exp\left(-\frac{\alpha_\star L}{2} \sum_\ell \left[\widehat{\bar\chi}^\ell \widehat{\bar\chi}^\ell M^\ell_{\sigma\sigma C_\phi} + \widehat{\bar\xi}^\ell \widehat{\bar\xi}^\ell M^\ell_{\sigma\sigma C_g}\right]\right)$$

$$\exp\left(-i\widehat{R}^L_{h\xi} h^L \xi^L + i\sum_\ell \widehat{\bar\chi}^\ell \left[\bar\chi^\ell - \bar{R}^\ell_{\phi\xi} g^\ell\right] + i\sum_\ell \widehat{\bar\xi}^\ell \left[\bar\xi^\ell - \bar{R}^\ell_{g\chi} h^\ell\right]\right)$$

$$\exp\left(i\sum_\ell \widehat{h}^{\ell+1}\left[h^{\ell+1} - h^\ell - L^{-1}\bar\chi^\ell - \gamma_0 L^{-1}\Delta M^\ell_{\sigma\sigma C_\phi} g^{\ell+1}\right]\right)$$

$$\exp\left(i\sum_\ell \widehat{g}^\ell\left[g^\ell - g^{\ell+1} - L^{-1}\bar\xi^\ell - \gamma_0 L^{-1}\Delta\left(M^\ell_{\sigma\sigma C_g} + M^\ell_{\dot\sigma\dot\sigma\mathcal{A}\mathcal{A}}\right) h^\ell\right]\right)$$

$$\alpha_\star \equiv \frac{N}{N_e E L}$$

$$\mathcal{Z}^\ell_{\text{exp}} = \int \mathcal{D}p^\ell \mathcal{D}\widehat{p}^\ell \mathcal{D}\mathcal{A}^\ell \mathcal{D}\widehat{\mathcal{A}}^\ell \mathcal{D}C_{\phi_k} \mathcal{D}\widehat{C}_{\phi_k} \mathcal{D}C^\ell_{g_k} \mathcal{D}\widehat{C}_{g_k} \mathcal{D}C^\ell_{h_k\xi_k} \mathcal{D}\widehat{C}_{h_k\xi_k}$$

$$\exp\left(-\widehat{M}^\ell_{\sigma\sigma C_\phi}\sigma^\ell\sigma^\ell C^\ell_{\phi_k} - \widehat{M}^\ell_{\sigma\sigma C_g}\sigma^\ell\sigma^\ell C^\ell_{g_k} - \widehat{M}^\ell_{\dot\sigma\dot\sigma\mathcal{A}\mathcal{A}}\dot\sigma^\ell\dot\sigma^\ell\mathcal{A}^\ell\mathcal{A}^\ell\right)$$

$$\exp\left(i\widehat{p}^\ell_k\left[p^\ell_k - \gamma_0\Delta\dot\sigma^\ell_k\mathcal{A}^\ell_k C^\ell_h\right] + i\widehat{\mathcal{A}}^\ell_k\left[\mathcal{A}^\ell_k - C^\ell_{h_k\xi_k} - \gamma_0\Delta\sigma^\ell_k C^{\ell+1}_g C^\ell_{\phi_k}\right]\right)$$

$$\exp\left(C_{\phi_k}\widehat{C}_{\phi_k} + C_{g_k}\widehat{C}_{g_k} + C_{h_k\xi_k}\widehat{C}_{h_k\xi_k} + N_e \ln \mathcal{Z}^\ell_{\text{within-exp}}\right)$$

$$\frac{\partial\mathcal{S}}{\partial\widehat{C}^\ell_h} = C^\ell_h - \left\langle h^\ell h^\ell\right\rangle = 0$$

$$\frac{\partial\mathcal{S}}{\partial\widehat{C}^\ell_g} = C^\ell_g - \left\langle g^\ell g^\ell\right\rangle = 0$$

$$\frac{\partial\mathcal{S}}{\partial\widehat{M}^\ell_{\sigma\sigma C_{\phi_k}}} = \nu M^\ell_{\sigma\sigma C_{\phi_k}} - \nu\left[\sigma^\ell\sigma^\ell\{\phi(h^\ell_k)\phi(h^\ell_k)\}\right] = 0$$

$$\frac{\partial\mathcal{S}}{\partial\widehat{M}^\ell_{\sigma\sigma C_{g_k}}} = \nu M^\ell_{\sigma\sigma C_{g_k}} - \nu\left[\sigma^\ell\sigma^\ell\{g^\ell_k g^\ell_k\}\right] = 0$$

$$\frac{\partial\mathcal{S}}{\partial\widehat{M}^\ell_{\dot\sigma\dot\sigma\mathcal{A}\mathcal{A}}} = \nu M^\ell_{\dot\sigma\dot\sigma\mathcal{A}\mathcal{A}} - \nu\left[\dot\sigma^\ell\dot\sigma^\ell\mathcal{A}^\ell\mathcal{A}^\ell\right] = 0$$

$$\frac{\partial\mathcal{S}}{\partial\widehat{\bar{R}}^\ell_{\phi\xi}} = -\nu\bar{R}^\ell_{\phi\xi} - i\nu\left[\sigma^\ell\left\{\phi(h^\ell_k)\widehat{\xi}^\ell_k\right\}\right] = 0$$

$$\frac{\partial\mathcal{S}}{\partial\widehat{\bar{R}}^\ell_{g\chi}} = -\nu\bar{R}^\ell_{g\chi} - i\nu\left[\sigma^\ell\left\{g^\ell_k\widehat{\chi}^\ell_k\right\}\right] = 0$$

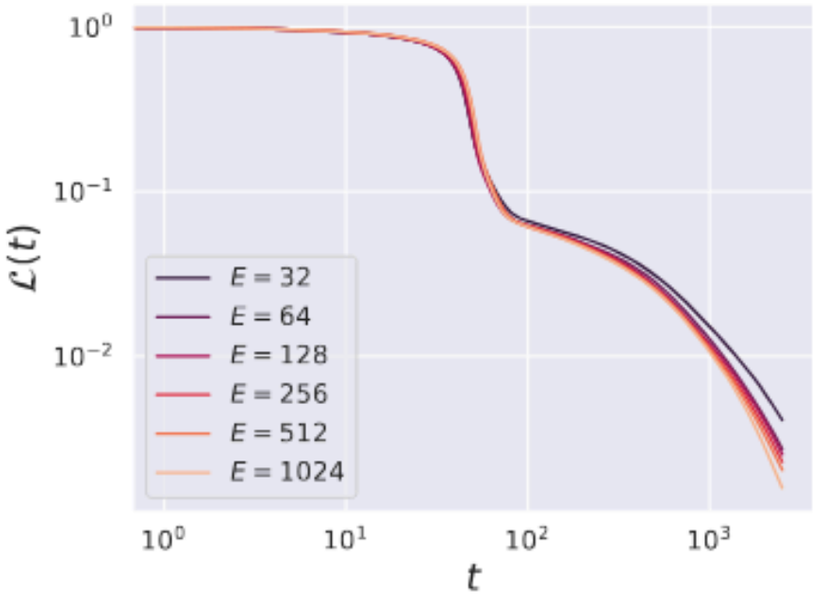
$$\frac{\partial\mathcal{S}}{\partial\bar{R}^\ell_{\phi\xi}} = -\frac{1}{\alpha_\star L}\widehat{\bar{R}}^\ell_{\phi\xi} - i\left\langle\widehat{\bar\chi}^\ell g^\ell\right\rangle = 0$$

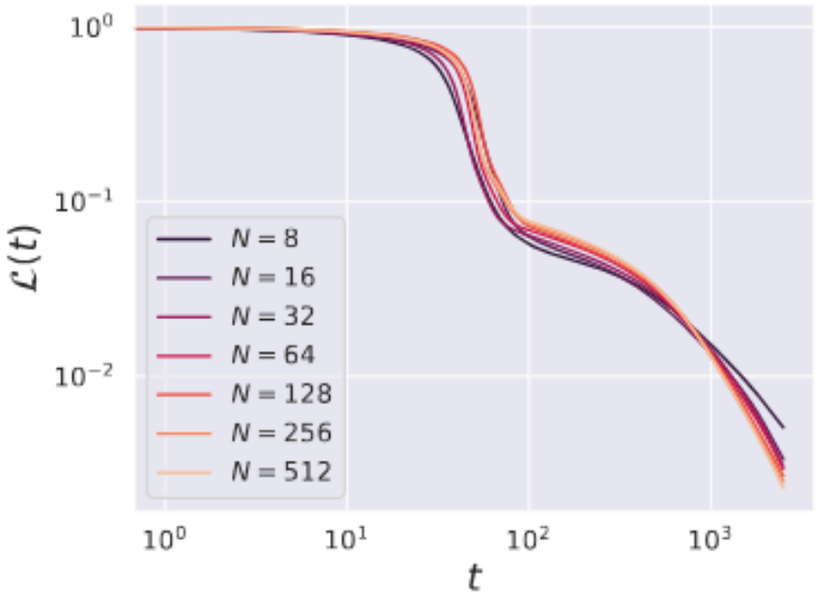
$$\frac{\partial\mathcal{S}}{\partial\bar{R}^\ell_{g\chi}} = -\frac{1}{\alpha_\star L}\widehat{\bar{R}}^\ell_{g\chi} - i\left\langle\widehat{\bar\xi}^\ell h^\ell\right\rangle = 0$$

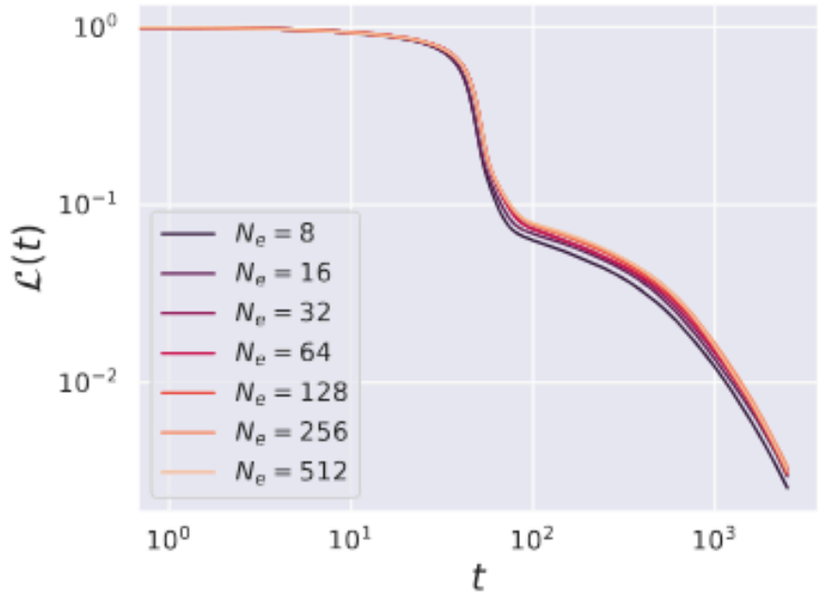
Top-k condition gives deterministic Implicit function $q = \sigma(p) + b$

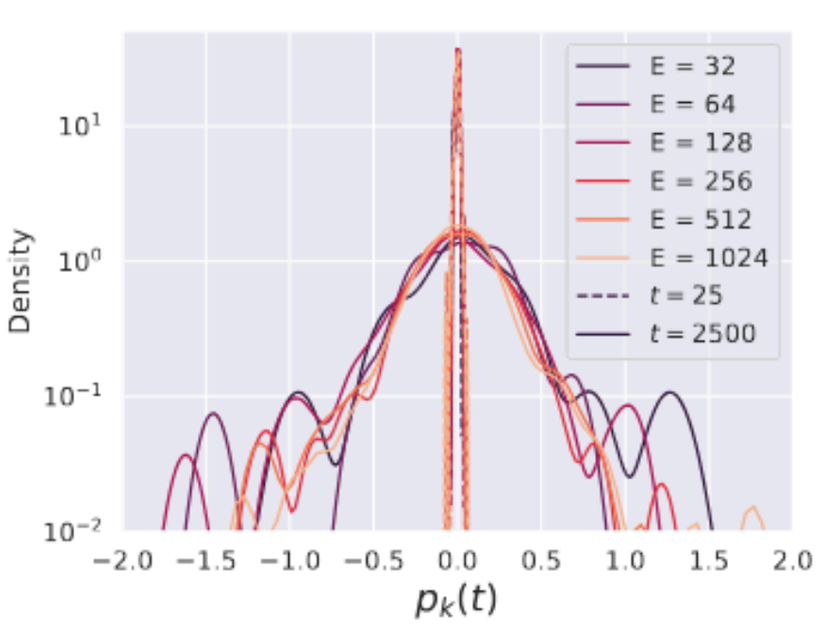
$$\left[1_{q\geq q_\star(\kappa)}\right] = \kappa$$

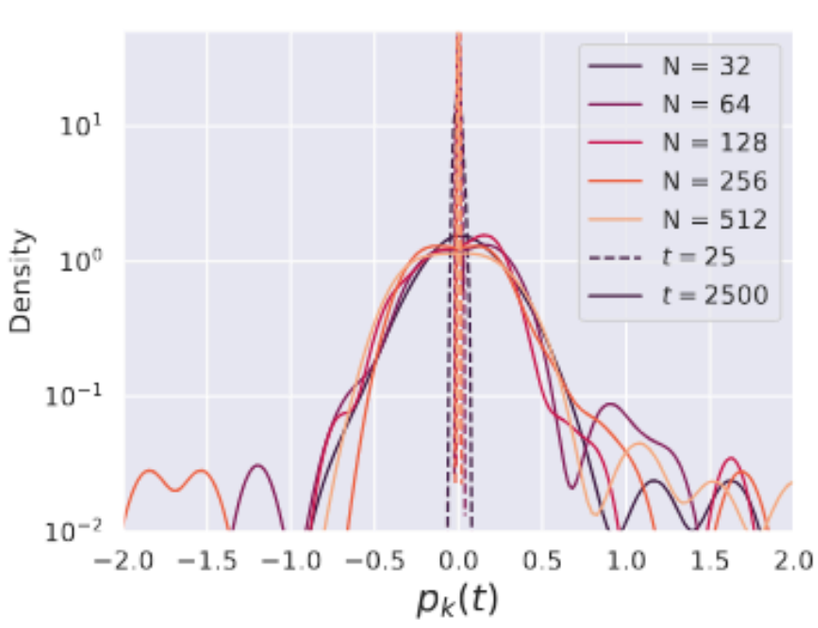
Convergent Dynamics in a Toy Soft Routing Model

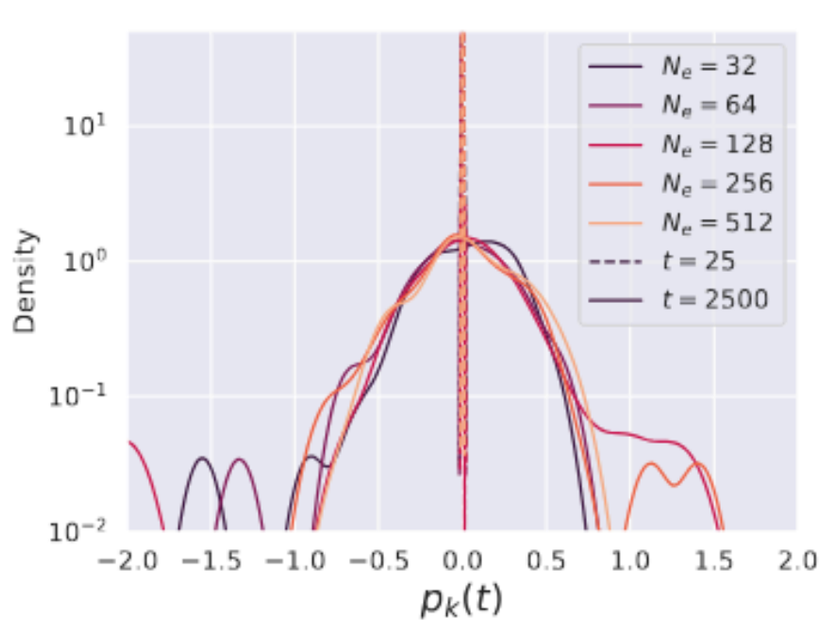








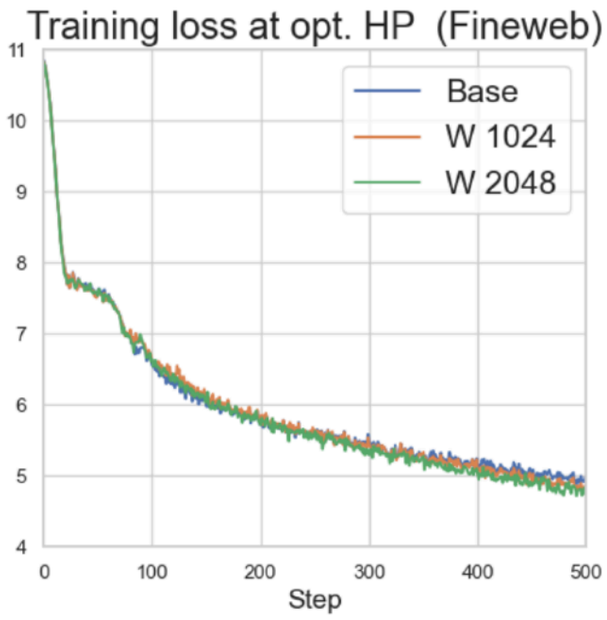


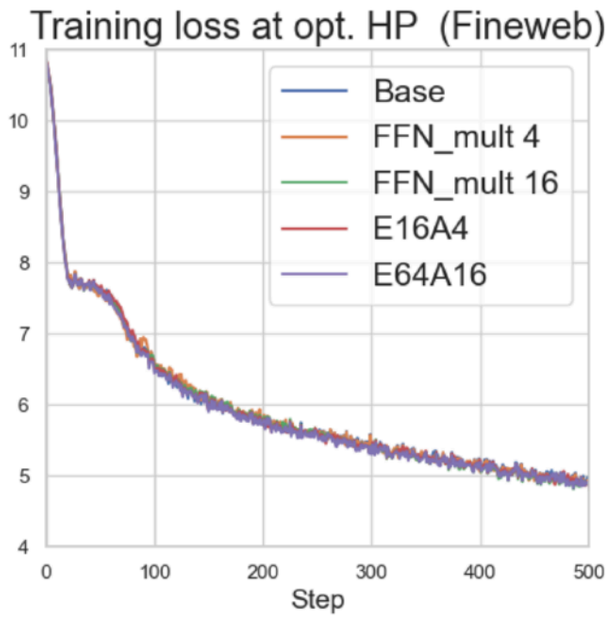


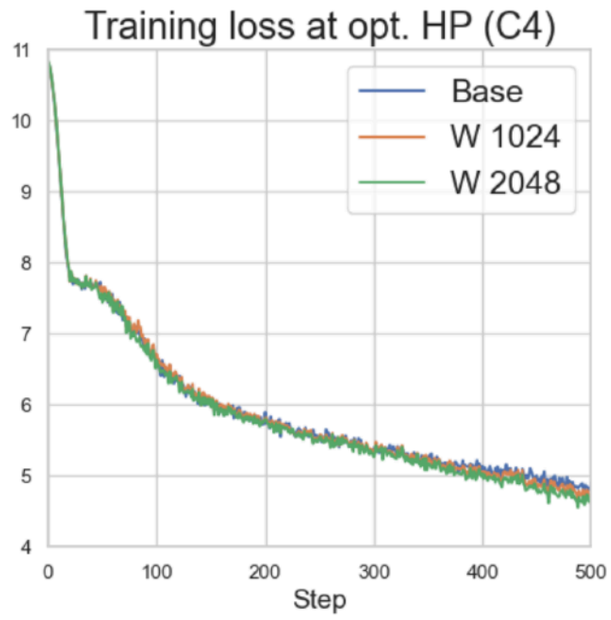
Simulations of a simple supervised learning task (multi-index polynomial model)

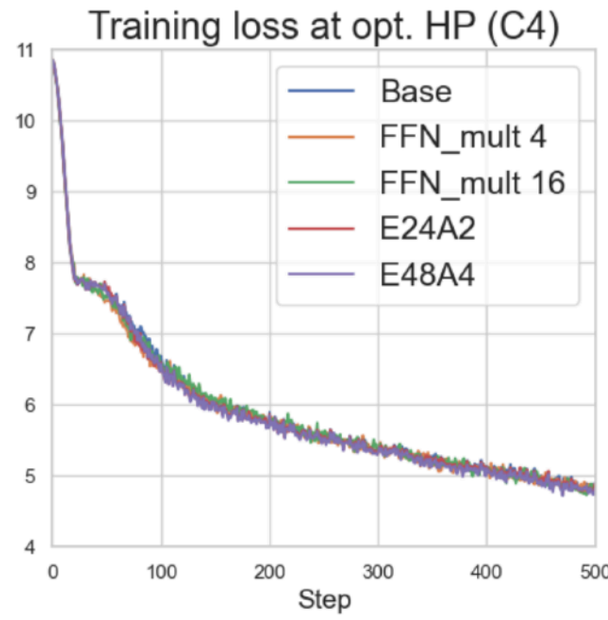
Sparse MOE: Convergent Universal Dynamics + HP Transfer

Consistent Early Pre-training Dynamics across all Scaling Directions

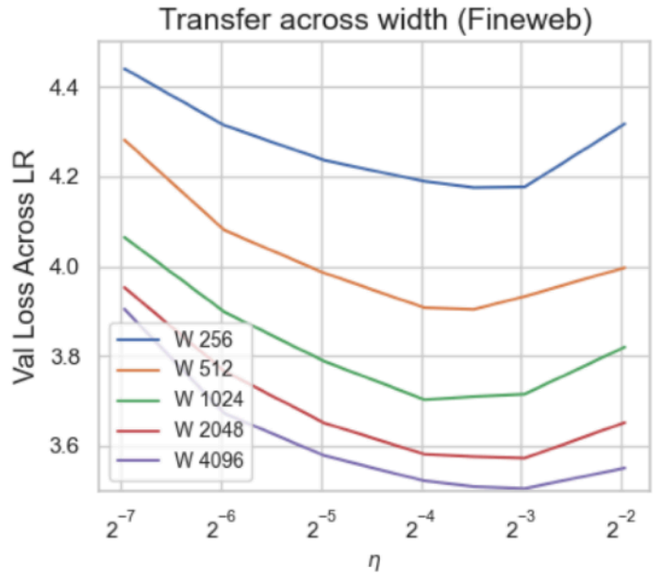


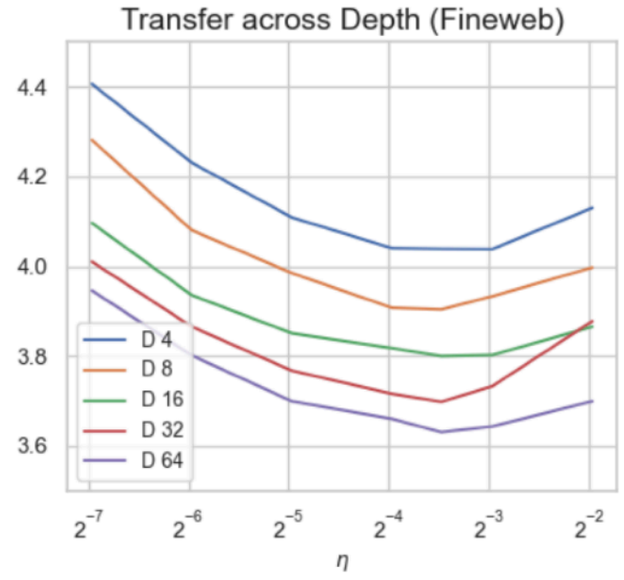


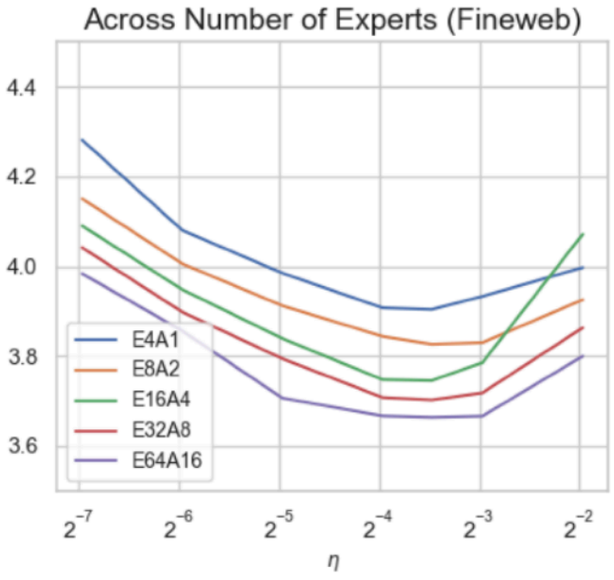


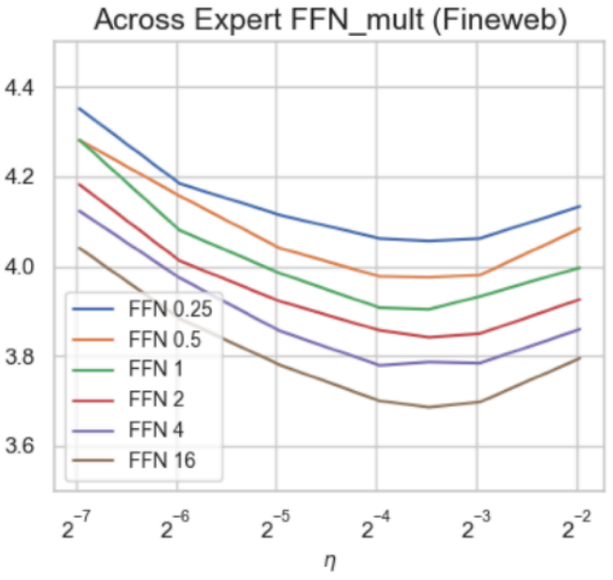


Learning Rate Transfer Across all Scaling Variables







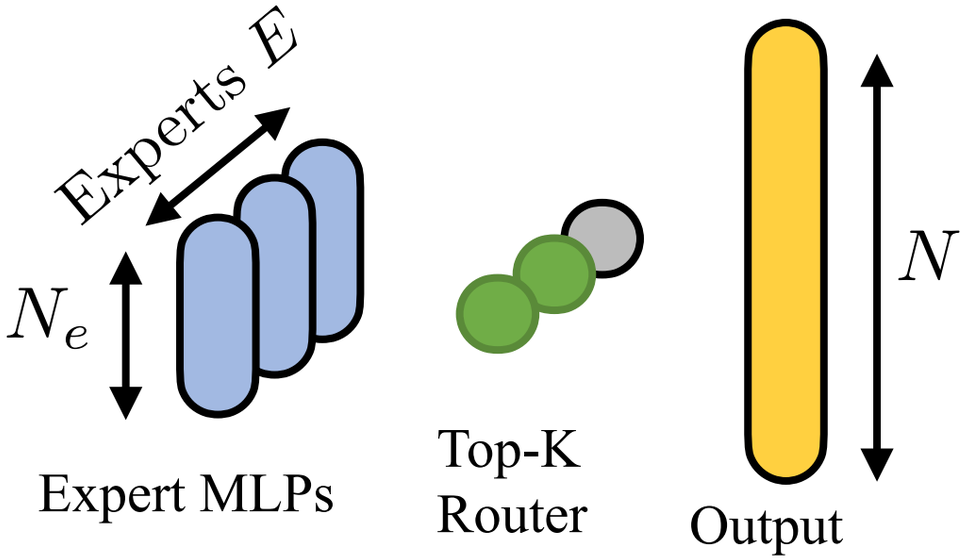


Transfer across all scaling variables possible since the limiting neural ODE surprisingly independent of FFN ratio, # experts, width or depth, only on $\kappa = K/E$

Jiang, **B**, Pehlevan, Hanin ’26

In Progress: Complete FL SDE Under Joint Scaling

1/L Parameterization but Different Joint Scaling of All Scaling Variables



$$\lim_{N\to\infty} \frac{N}{N_e(N)E(N)L(N)} = \alpha_\star > 0$$

$$dh(\tau) = \underbrace{\sqrt{\alpha_\star}\, du(\tau)}_{\text{Brownian Motion}} + \gamma_0 \underbrace{d\tau \mathcal{K}_h(\tau, \alpha_\star) g(\tau)}_{\text{Feature Learning}}$$

Can see this in a fully dense (non-MOE) model as well

$$\mathbf{h}^{\ell+1} = \mathbf{h}^\ell + \frac{\sqrt{N}}{N_e L}\mathbf{W}^{\ell,2}\phi\left(\frac{1}{\sqrt{N}}\mathbf{W}^{\ell,1}\mathbf{h}^\ell\right) \qquad \begin{aligned}\mathbf{W}^{\ell,1} &\in \mathbb{R}^{N_e\times N}\\ \mathbf{W}^{\ell,2} &\in \mathbb{R}^{N\times N_e}\end{aligned}$$

$\alpha_\star = \lim_{N\to\infty} \frac{N}{N_e(N)L(N)}$ If nonzero, then you get **SDE behavior** in the joint limit

Convergence to the Neural ODE requires controlling $\frac{N}{N_e L}$ see Chaintron ’26 & Jiang et al ‘26

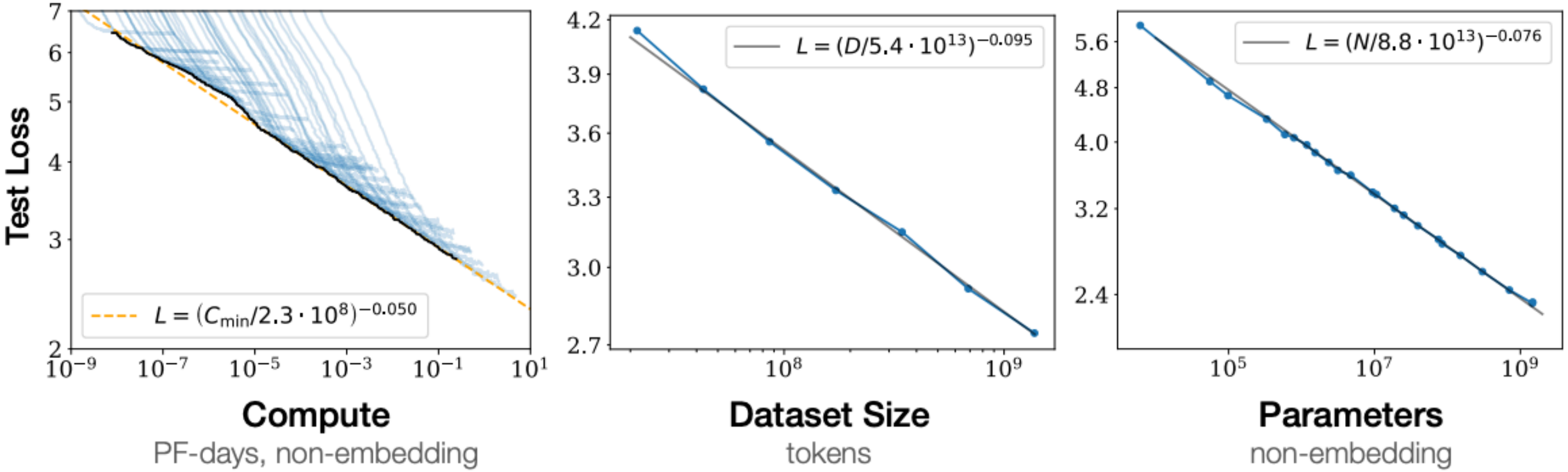
Approximation error to an **SDE** is *smaller* if $\frac{N}{N_e L}$ is fixed

Open: Is transfer across a large range of model sizes best by fixing $\alpha_\star$?

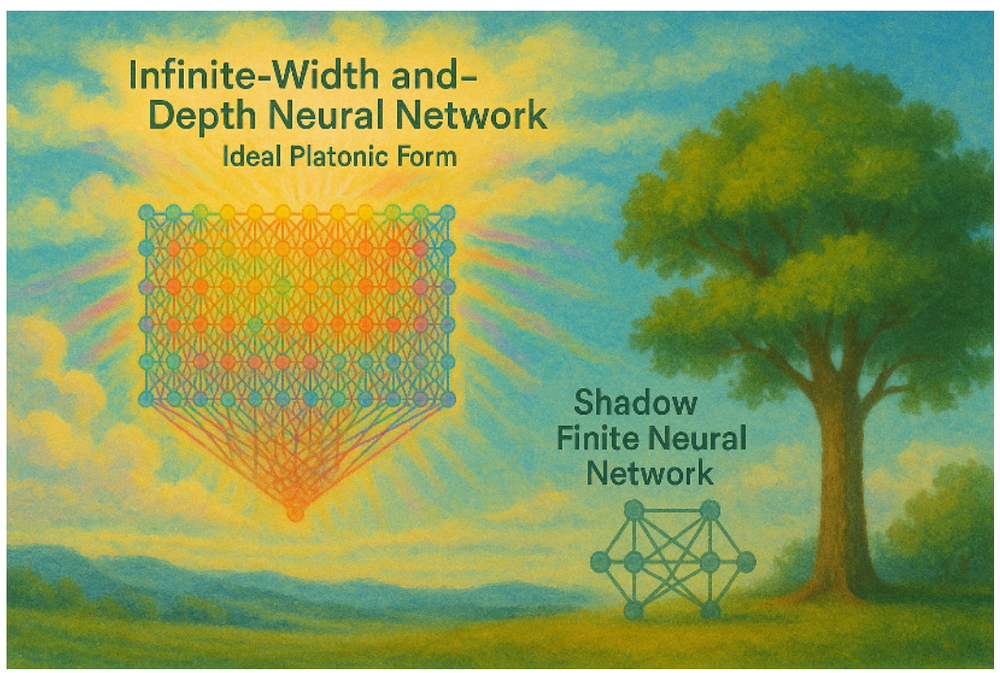
Common use of completeP: $\frac{N}{N_e L} \sim \frac{1}{N}$ In prep with Datong Zhou and Zhen Yang

We Talked About Infinite Limits… What About Scaling Laws?

How do finite size effects influence the model's training dynamics and scaling law?



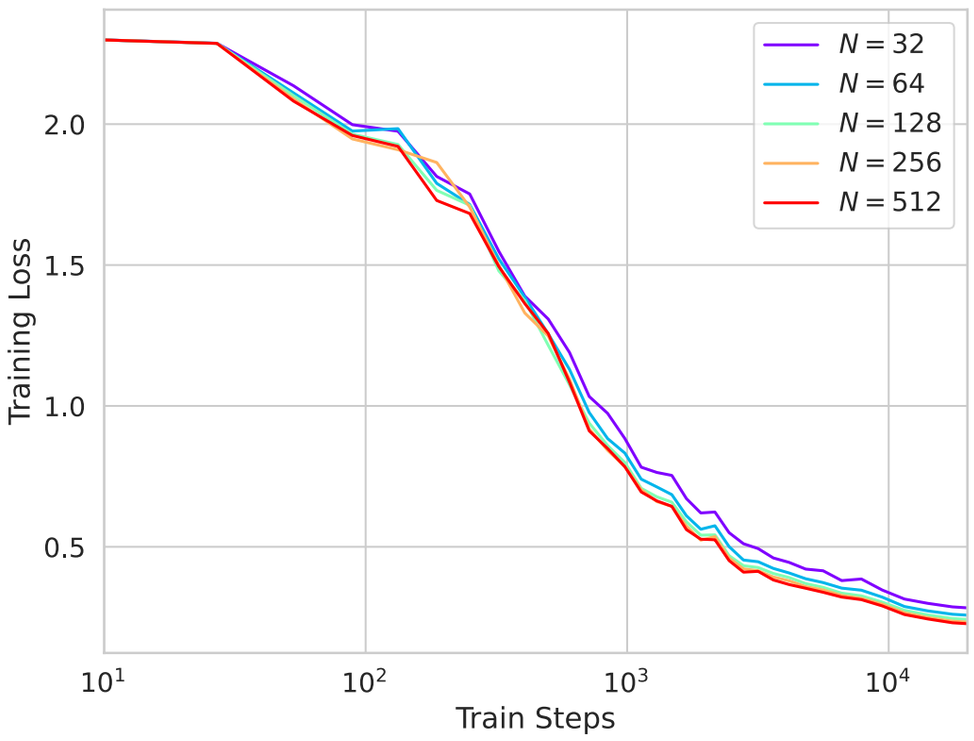
Infinite limits are like platonic forms of NNs … what we can fit in memory is a mere shadow



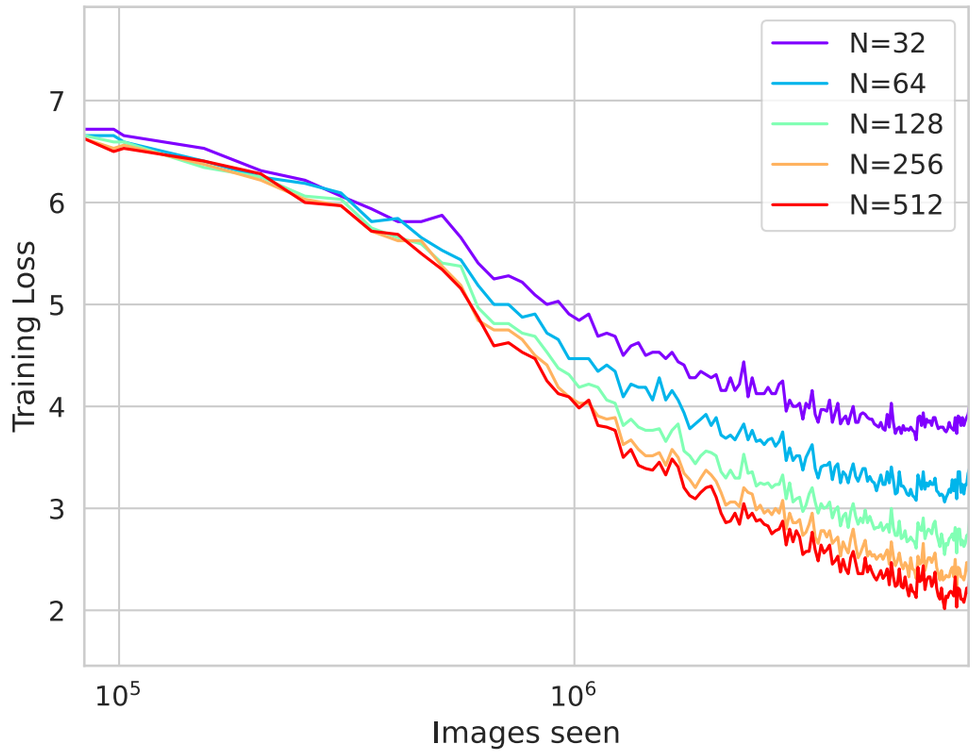
To make practical decisions,
we need to understand loss
to performance from "finiteness"

Empirically Stress Testing the Mean Field Limit

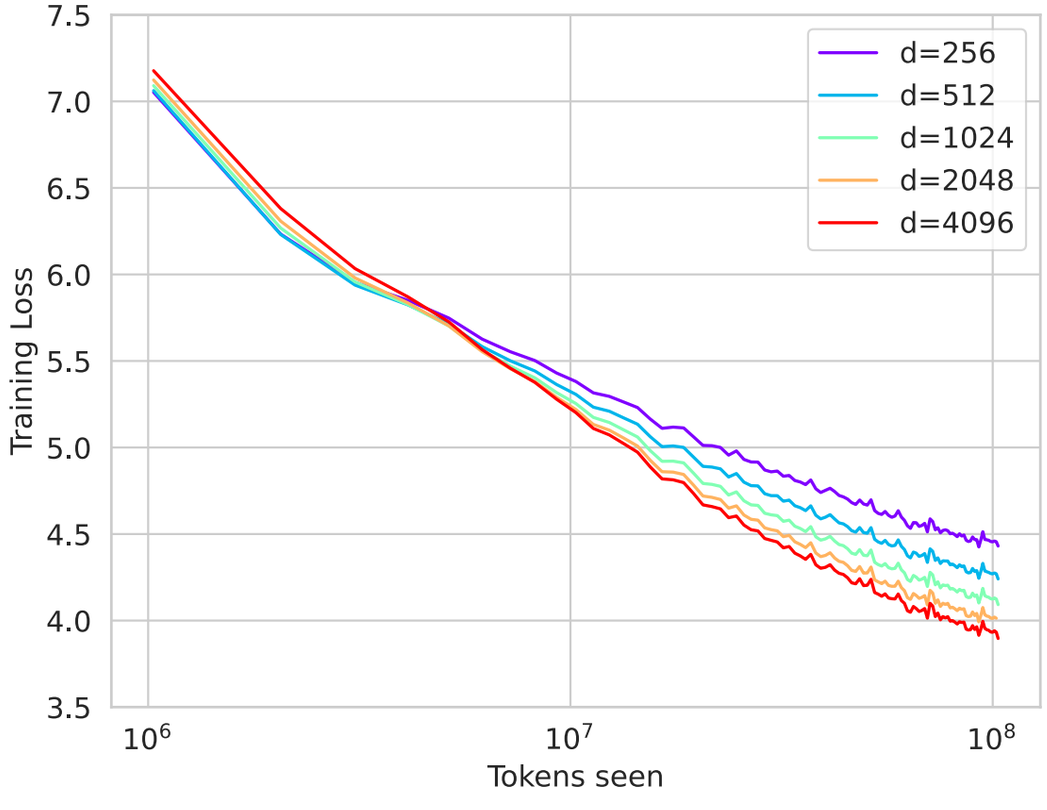
ResNets on CIFAR-5M



ImageNet



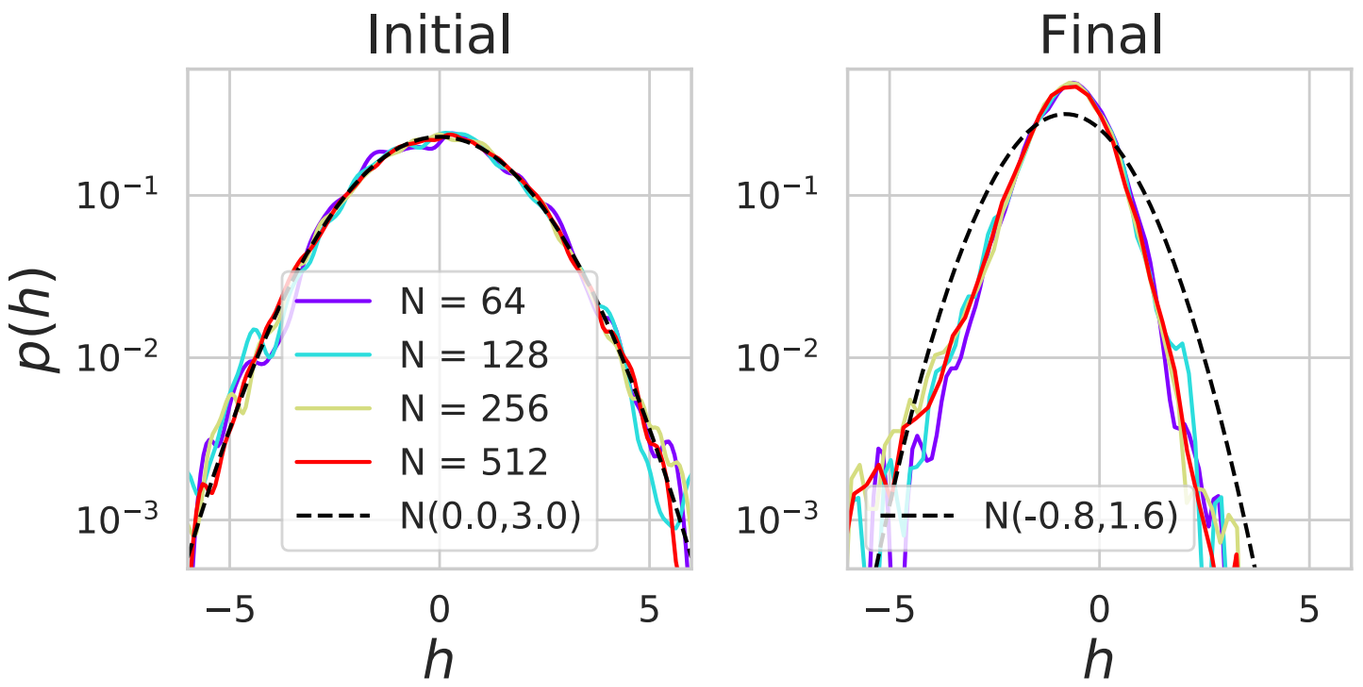
Transformer on Wikitext-103



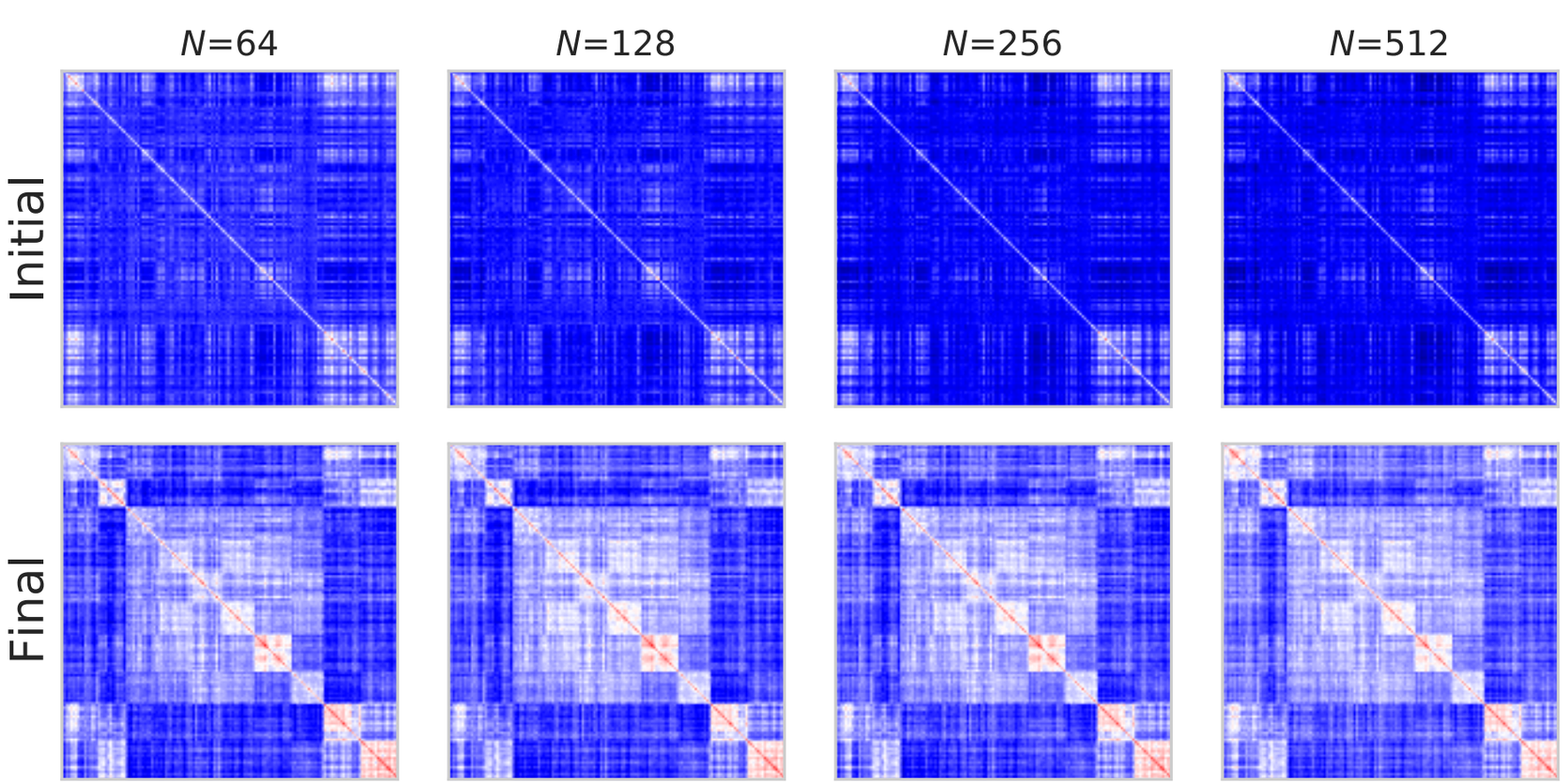
Vyas*, Atanasov*, **B***, Morwani, Sainathan, Pehlevan ‘23

When the MF limit works … it really works!

Non-Gaussian preactivation distributions conserved across widths.



Feature kernel dynamics also conserved across widths

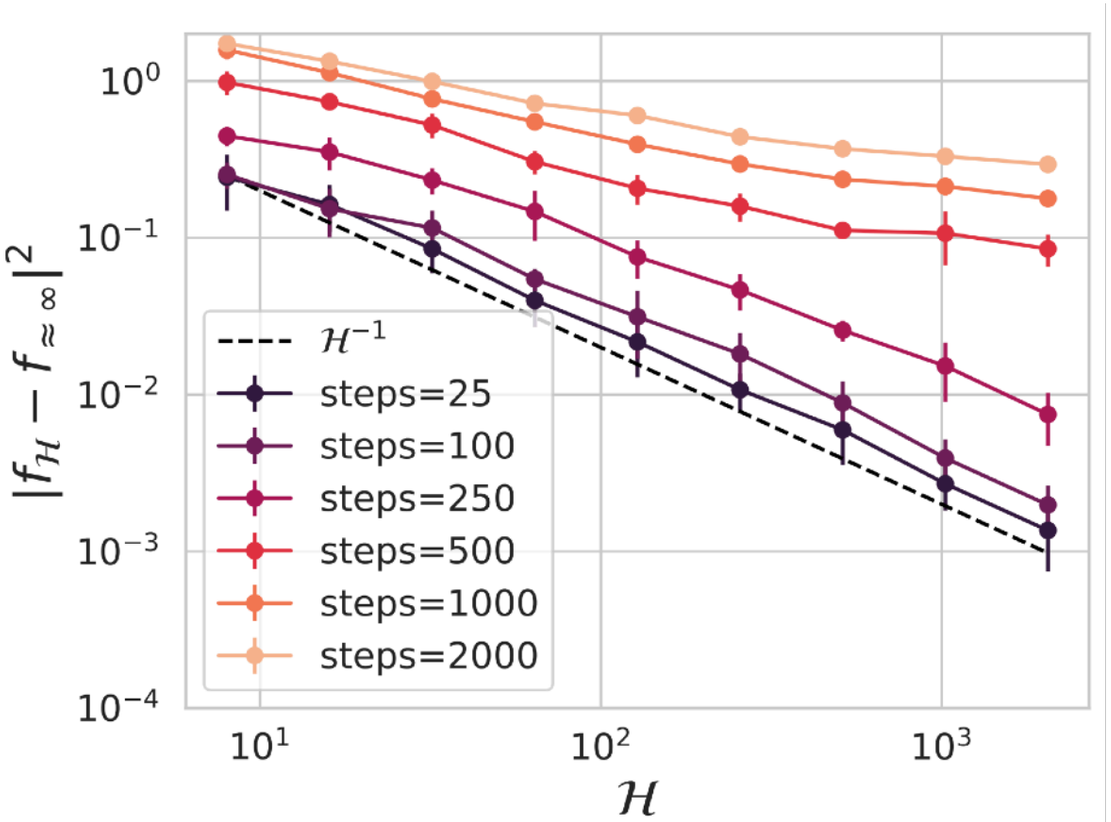


Vyas*, Atanasov*,**B***, Morwani, Sainathan, Pehlevan ‘23

What about Tasks Where It Doesn't Work? Scaling Laws…

Early dynamics converge at
Rate $\mathcal{O}(\mathcal{H}^{-1})$

Later dynamics converge at
Rate $\mathcal{O}(\mathcal{H}^{-\alpha})$



After long training, power law scaling
In network width N

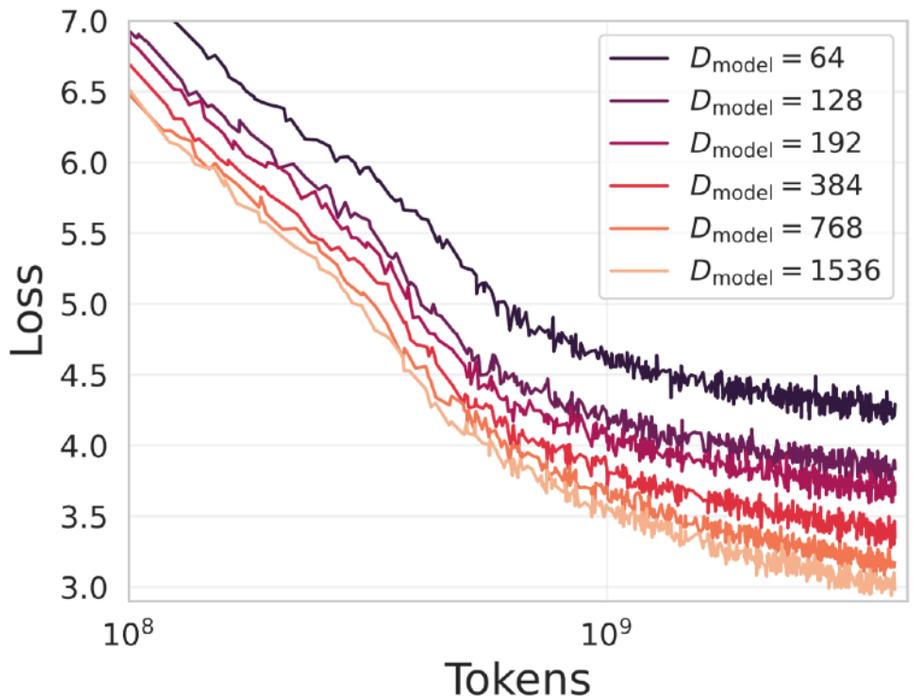
Scaling exponents non-universal…
task + architecture dependent

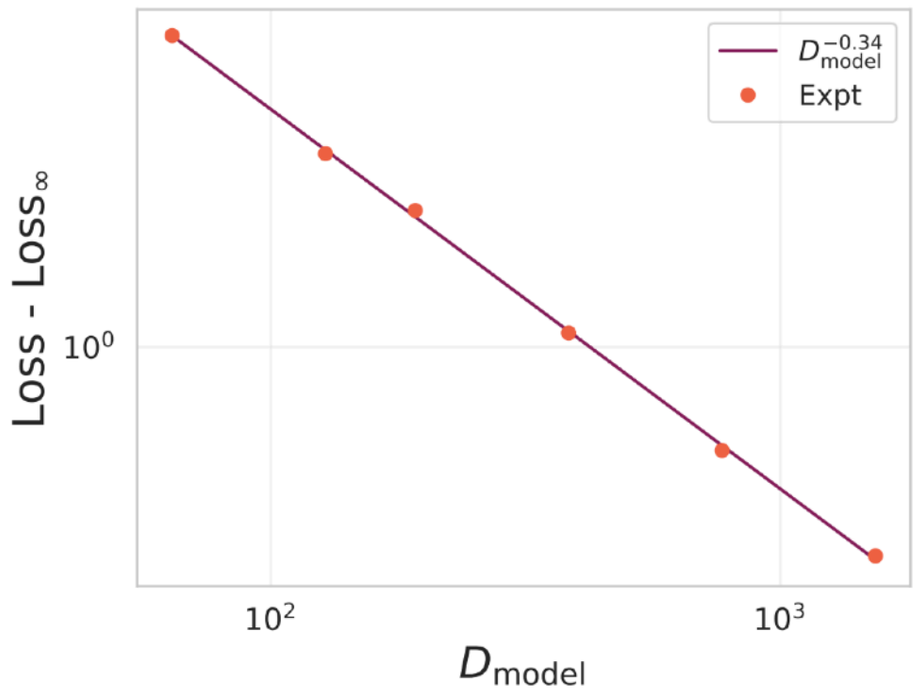
We need to understand this
in a theoretical model!

What joint data/width/depth limit describes modern LLM regime?

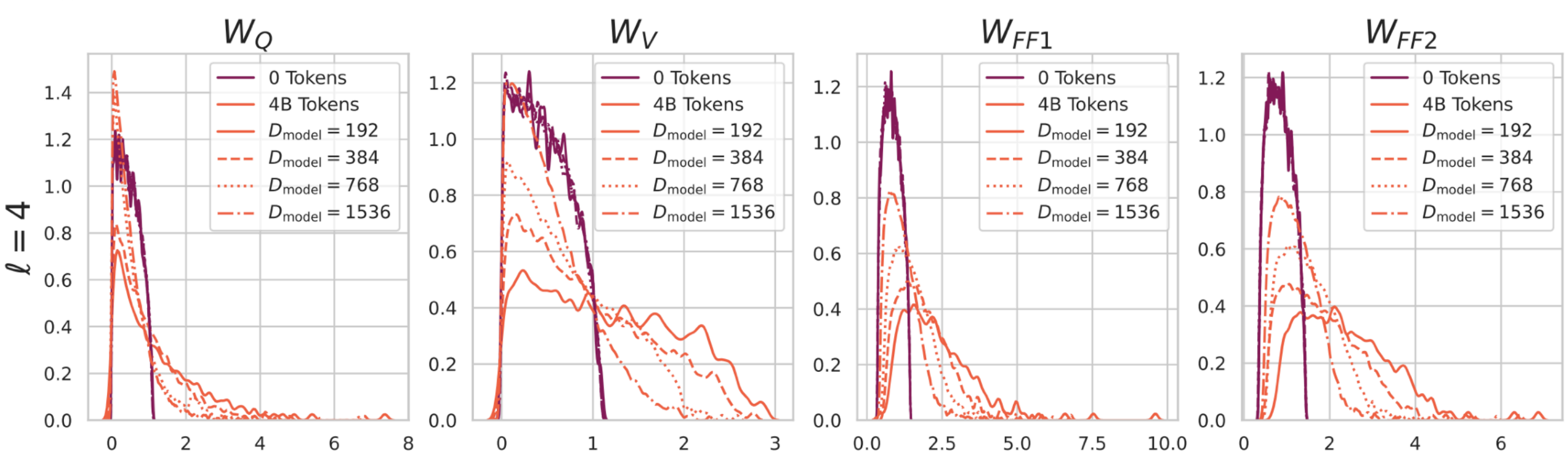
Scaling regime:

LMs trained on C4





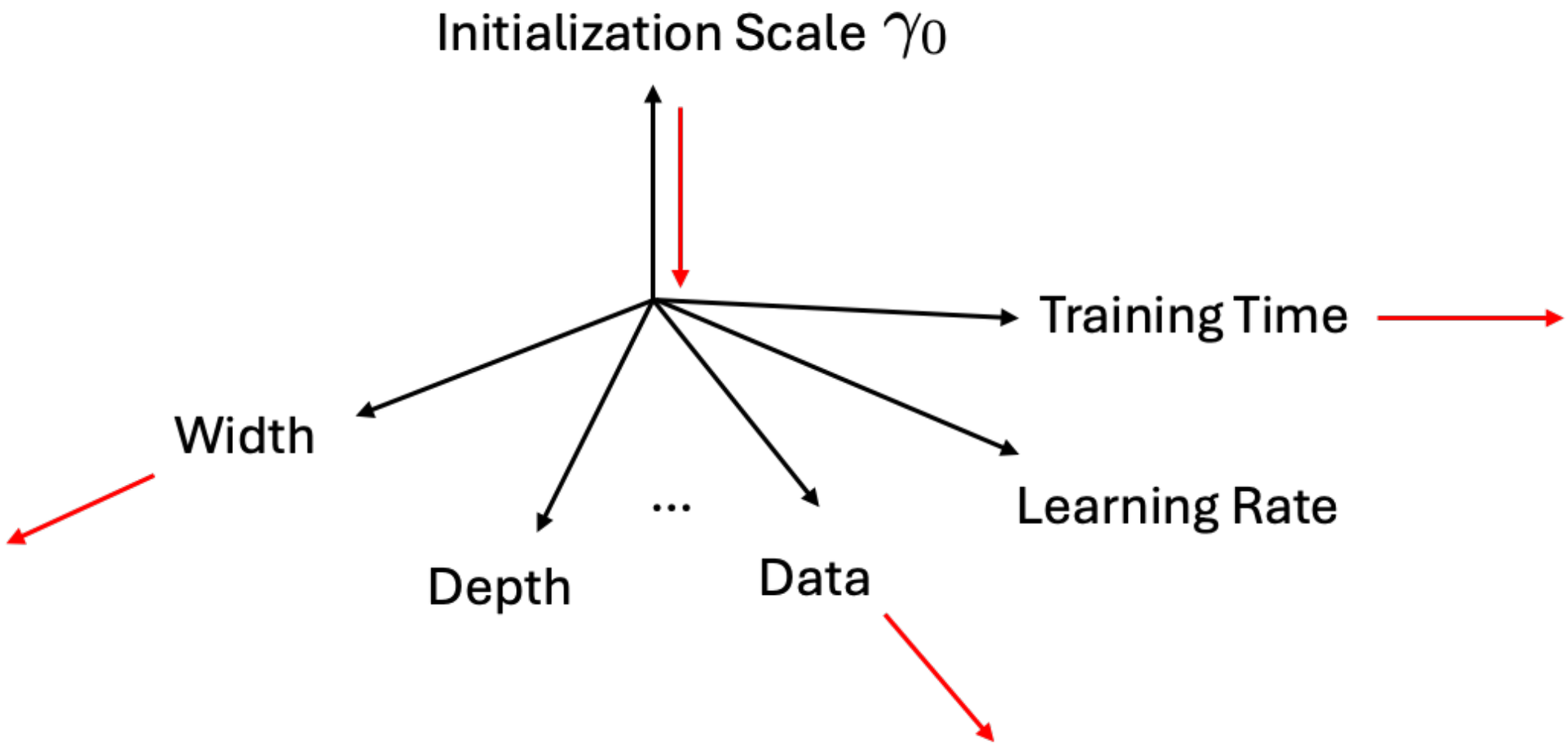
Weight spectra not bulk+spike: at width ~1000, Chinchilla optimal is around ~1B tokens



Extensive rank theories? Can we build theories that track large tokens compared to width?

A Theory of Neural Scaling Laws for the Lazy Limit

Typical case behavior of this limit over *random draws of data and initial parameters*



This simple theory captures how the behavior of the model depends on computational/ statistical resources (width, training time, and total data)

Many works on statics of random feature models: Mei & Montanari ’21, Bahri et al ’22, Maloney et al ’22, Xiao et al ’22, Atanasov*, B*, Pehlevan ‘22 Simon et al ’23,…

We will also look at *dynamics,* which are a bit more involved

Dynamical Model of Neural Scaling Laws

Lazy Limit: To gain insight in a tractable model we first consider lazy training $\gamma_0 \to 0$

Diagonalize Kernel: Find eigenvalues and eigenfunctions of limiting kernel $K_\infty(x, x')$

$$K_\infty(x, x') = \sum_k \psi_k^\infty(x)\psi_k^\infty(x') \qquad \int dx\ p(x)\psi_k^\infty(x)\psi_\ell^\infty(x) = \lambda_k \delta_{k\ell}$$

Expand Finite Width Features: finite kernel eigenfunctions can be expanded in basis

$$\psi_k^N(x) = \frac{1}{\sqrt{N}} \sum_\ell A_{k\ell}\psi_\ell^\infty(x) \qquad k \in \{1, ..., N\}$$

$$K_N(x, x') = \sum_k \psi_k^N(x)\psi_k^N(x')$$

Key properties: $\mathbf{A}$ random over inits $\lim_{N\to\infty} \frac{1}{N}\mathbf{A}^\top\mathbf{A} = \mathbf{I}$ $K_N \to K$

Expand NN function and Target Function:

$$f(x) = \sum_\ell w_\ell \psi_\ell^N(x) = \frac{1}{\sqrt{N}} \sum_{k\ell} A_{\ell k}\ w_\ell\ \psi_k^\infty(x) \qquad y(x) = \sum_k w_k^\star \psi_k^\infty(x)$$

Dynamical Model of Neural Scaling Laws

Residual errors: $$f(x) - y(x) = \sum_k v_k^0 \psi_k(x) \qquad v_k^0 = w_k^\star - \frac{1}{\sqrt{N}} \sum_{\ell=1}^{N} A_{\ell k} w_\ell$$

Sample Random Data: $$\mathcal{D} = \{x_\mu\}_{\mu=1}^{P} \qquad \Psi_{\mu k} = \psi_k^\infty(x_\mu)$$

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

Test Loss: $\mathcal{L}(t) = \sum_k \lambda_k \left\langle v_k^0(t)^2 \right\rangle$ Equivalent to structured random feature model

Statics: Bahri et al ’22, Maloney et al ’22, Simon et al ’23, Atanasov*, **B***, Sainathan, Pehlevan ’22, …

Key insight: *random low rank projections (low rank kernels) limit generalization*

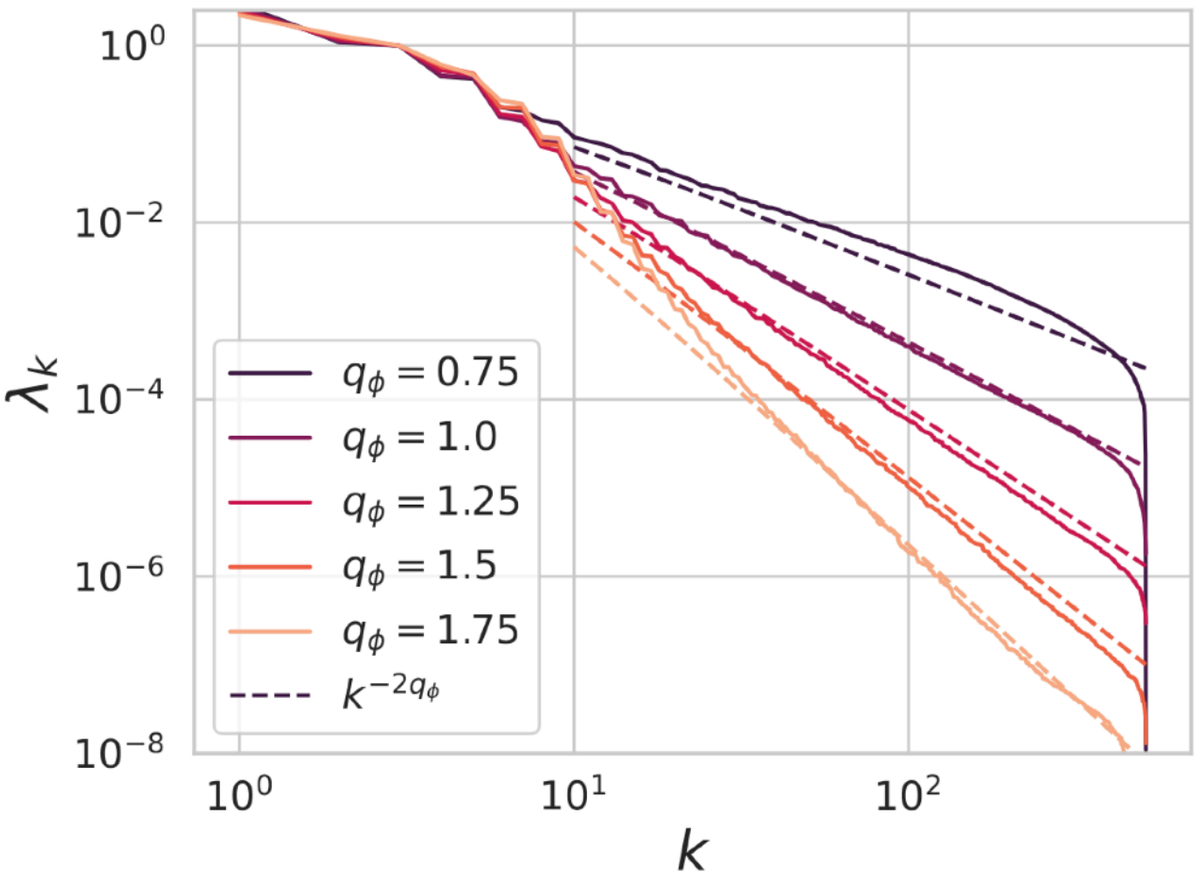
Finite time, data or finite model size can limit performance

Can we get an accurate description for random matrices?

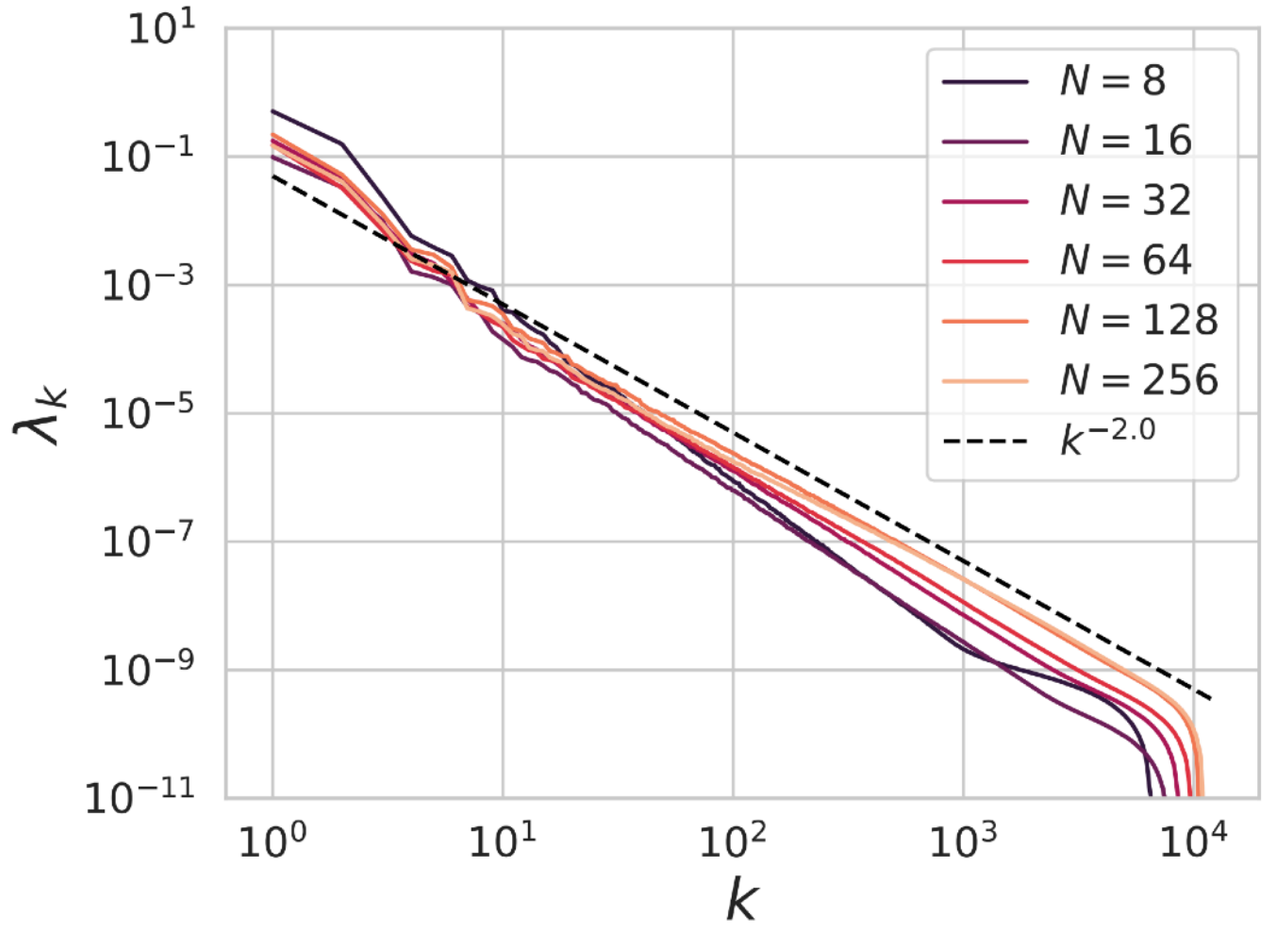
B, Atanasov, Pehlevan ICML ‘24

Dynamical Model of Neural Scaling Laws

Observation: Limiting NTK often has *Power Law Spectrum*



MLP with data uniformly distributed on the circle



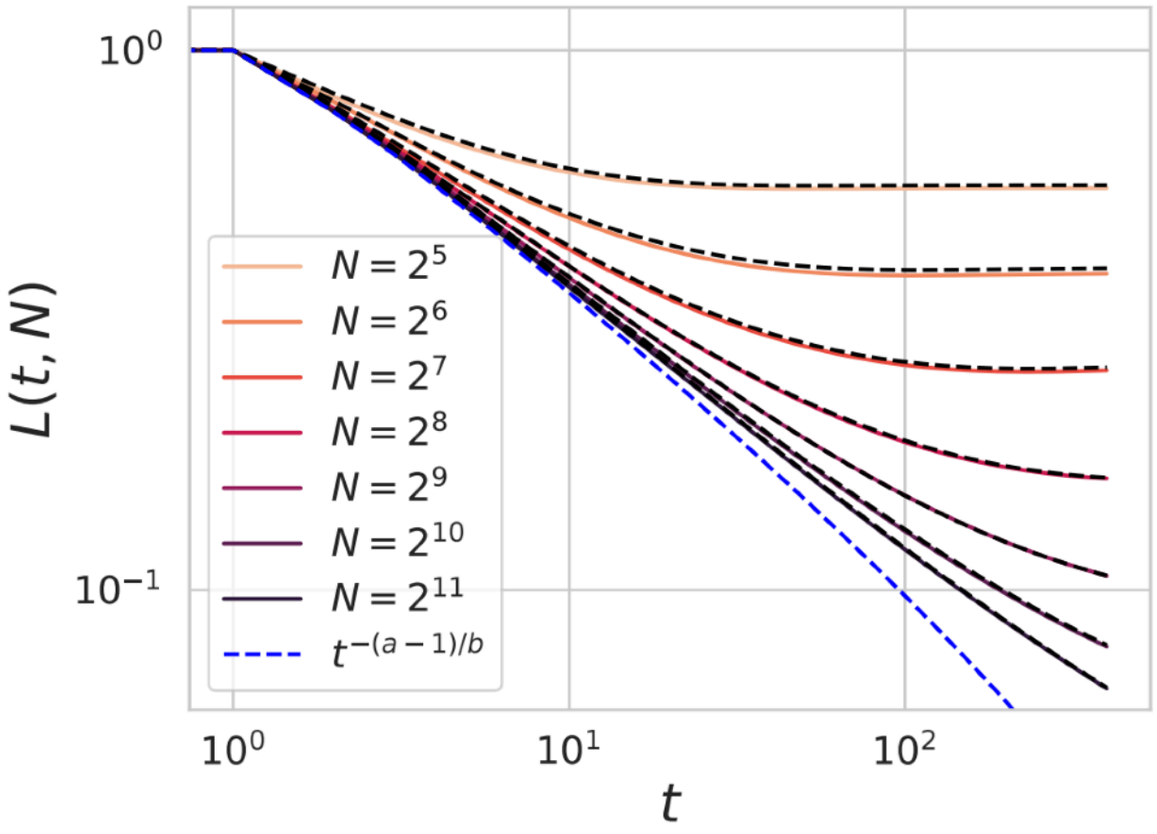
CNN with CIFAR-5M images

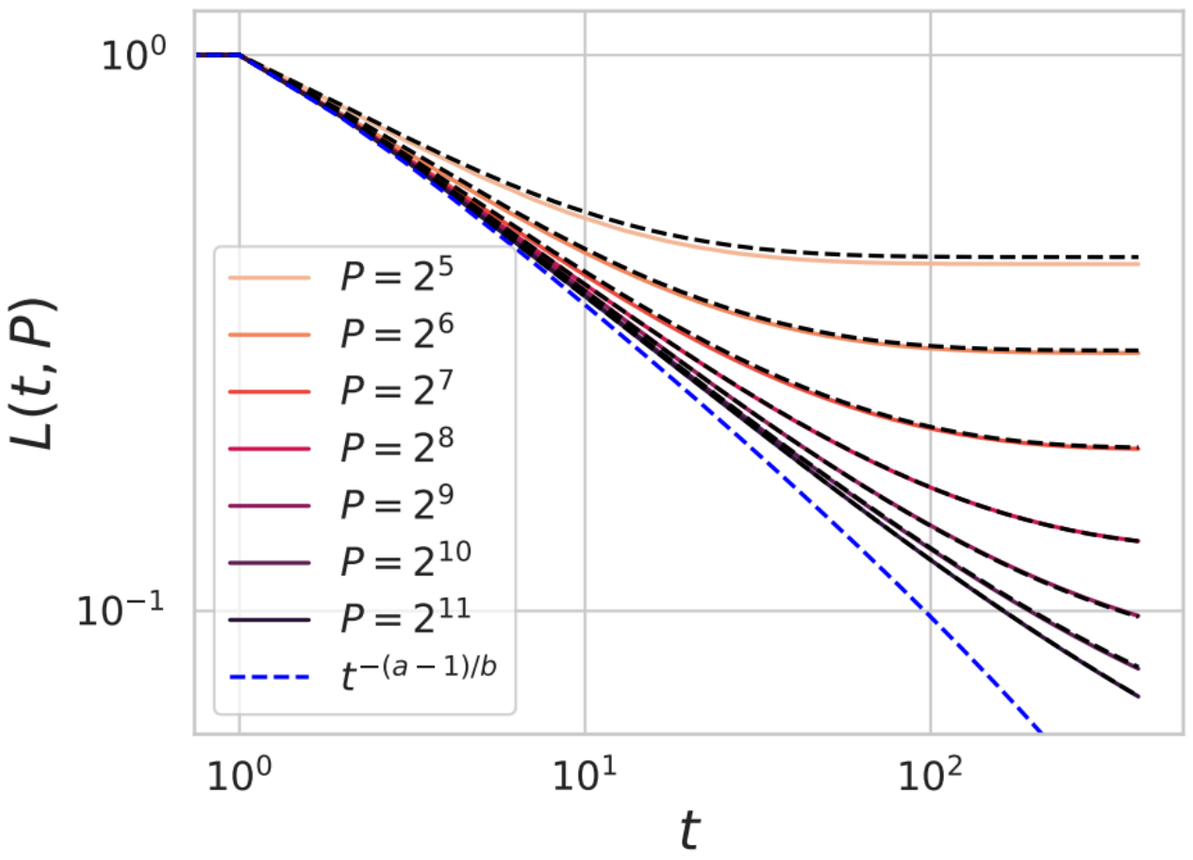
$$y_k^2 \sim k^{-a} \ , \ \lambda_k \sim k^{-b}$$

B, Atanasov, Pehlevan ‘24

Lazy Limit as a Toy Model of Neural Scaling Laws

Power Law Scalings Due to Rank-Limited Correlation Functions!





Bottleneck Scaling Laws

$$\mathcal{L}(t, N, P) \sim \begin{cases} t^{-(a-1)/b} & t^{1/b} \ll N, P \\ N^{-(a-1)} & N \ll t^{1/b}, P \\ P^{-(a-1)} & P \ll t^{1/b}, N \end{cases}$$

Intuition: GD captures top

$$k^\star \sim \min\{t^{1/b}, N, P\}$$

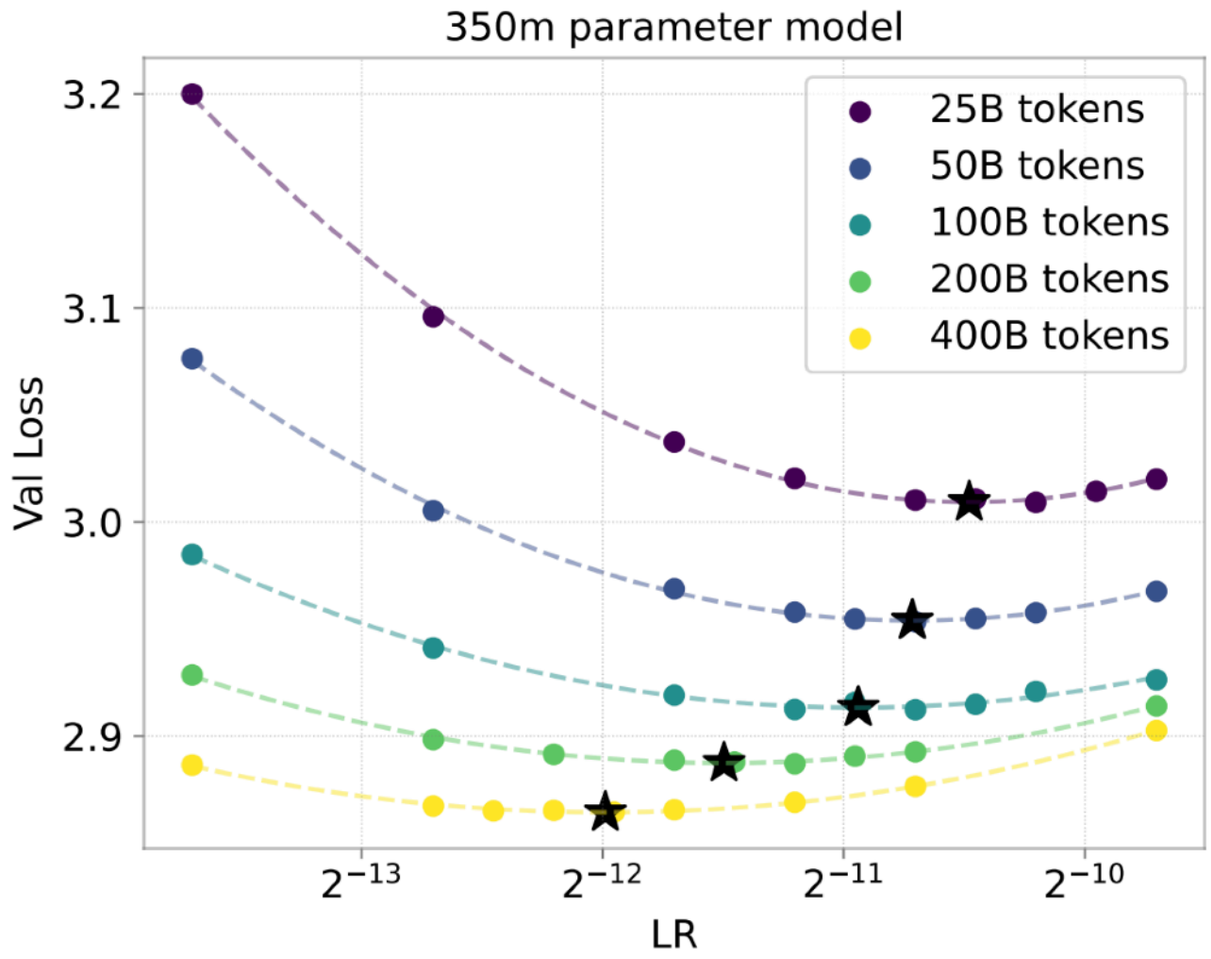
spectral components of the target fn

$$y_k^2 \sim k^{-a} \ , \ \lambda_k \sim k^{-b}$$

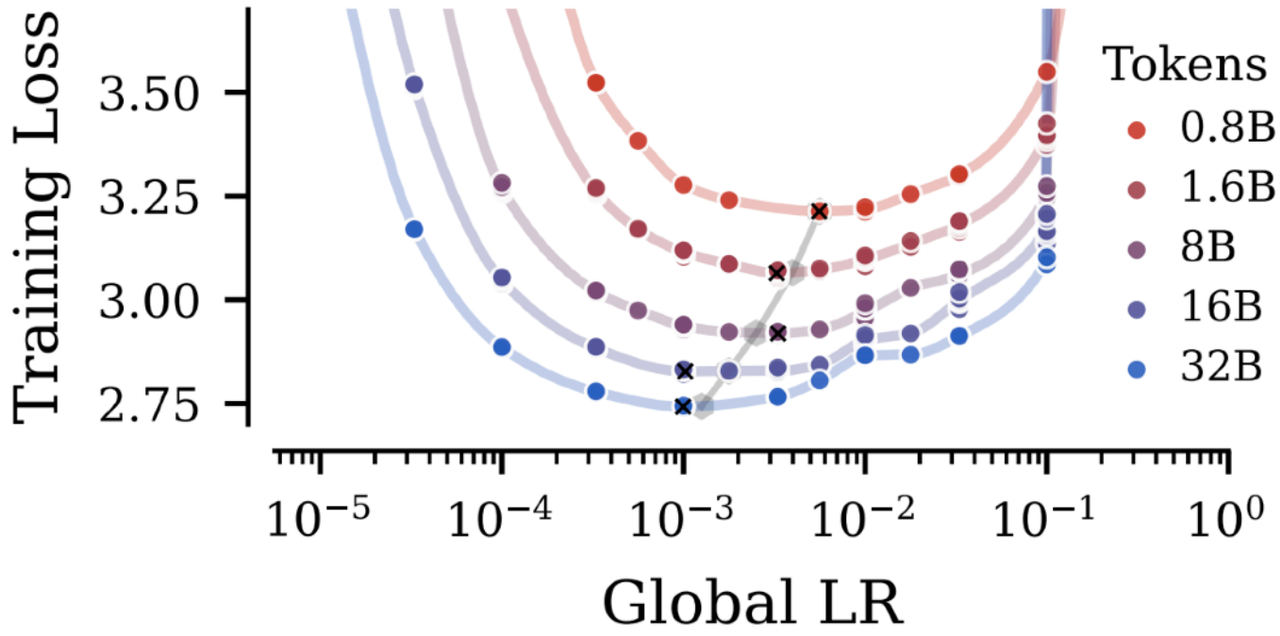
B, Atanasov, Pehlevan ‘24

A Key Open Problem: Transfer Across Training Horizon

While we can transfer across model sizes, how to transfer across training time?



Bjork et al ‘25



Mlodozeniec et al ‘25

Can we develop a toy model of this effect? How to optimally set LR (+schedule?)

Optimal Control of the Dynamical Model of Neural Scaling Laws

In the DSL Model, formulate an optimal control problem for the learning rate

$$\eta_T^\star(t) = \mathrm{argmin}_{\eta(t)}\ \mathcal{L}_T[\eta(t)]$$ Mori & **B** 2026

At large width, SGD can be approximated as

$$\mathcal{L}_T[\eta(t)] = \chi(0)^{-\frac{a-1}{b}} + \frac{\sigma^2}{B}\int_0^T dt\ \dot{\chi}(t)^2\ \chi(t)^{-2+1/b}$$

$$\chi(t) \equiv \int_t^T dt\ \eta(t)$$

Phase I (Bias Dominated): $a < b \quad \eta_T^*(t) \sim \begin{cases} \eta_{\max} & t < t_s\ , \\ \eta_{\max}\left(\frac{1-t/T}{1-t_s/T}\right)^{2b-1} & t > t_s\ , \end{cases}$

Base learning rate constant, annealing after t_s steps $\quad 1 - t_s/T \sim T^{-\frac{b-a}{2b-1}}$

Phase II (SGD Dominated): $a > b \qquad \eta_T^\star(t) = T^{-(a-b)/a}\,(1-t/T)^{2b-1}$

Base learning rate decreases with training horizon, annealing is polynomial with exponent 2b-1

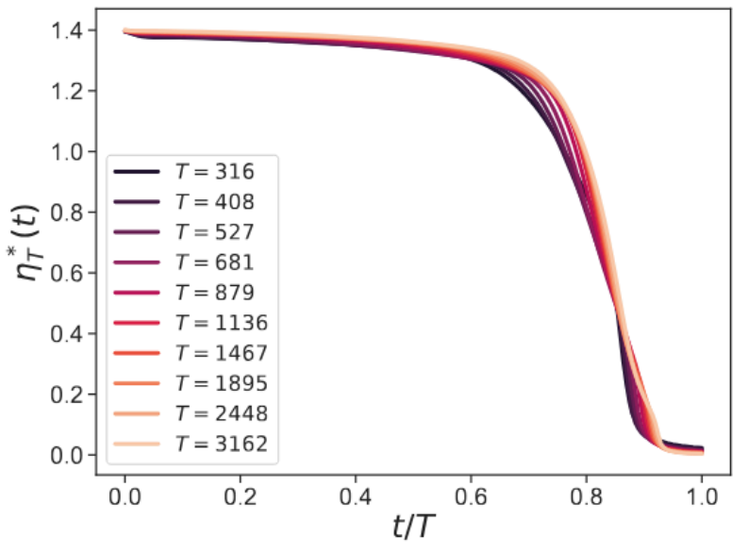
Optimal Control of the Dynamical Model of Neural Scaling Laws

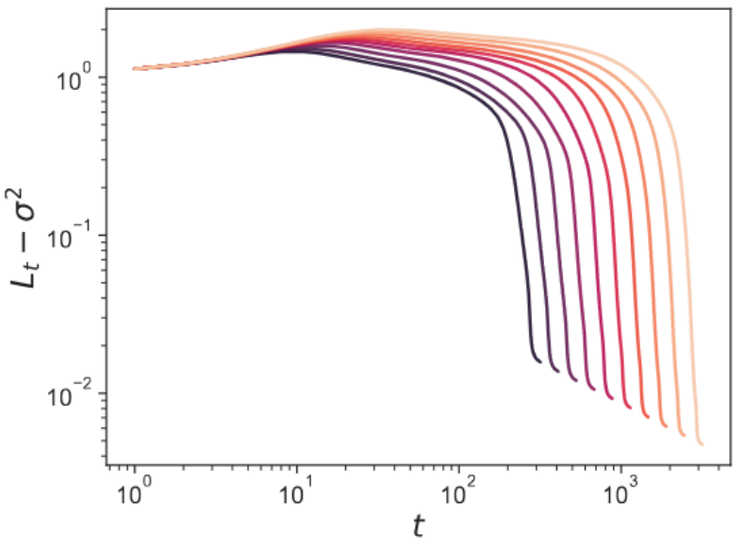
In the DSL Model, formulate an optimal control problem for the learning rate

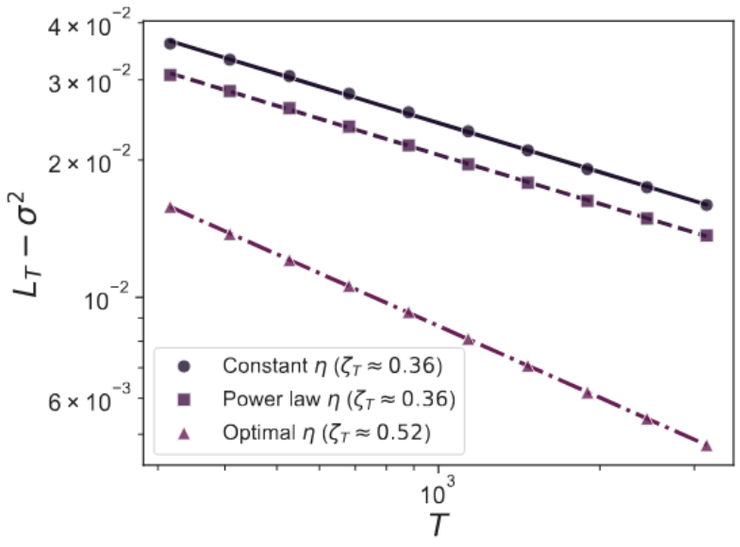
$$\eta_T^\star(t) = \text{argmin}_{\eta(t)}\ \mathcal{L}_T[\eta(t)]$$

Mori & **B** 2026

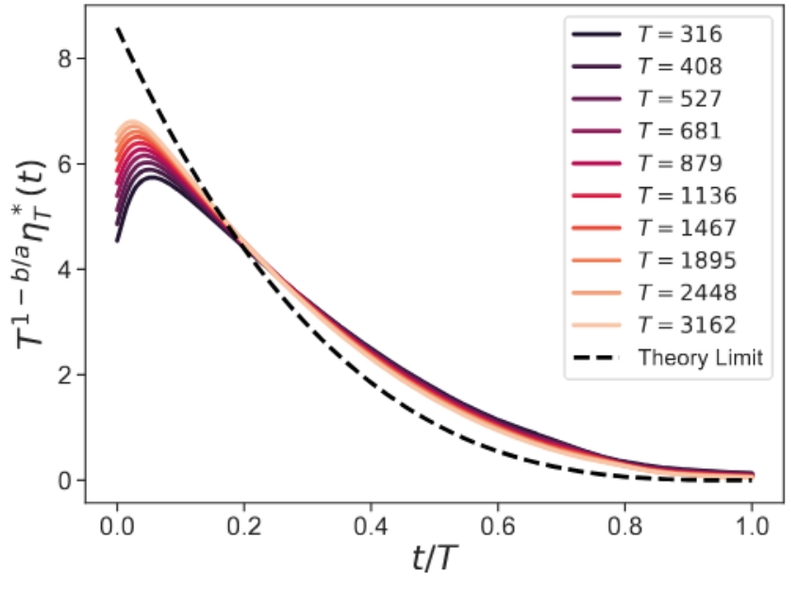
Phase I (Bias Dominated):

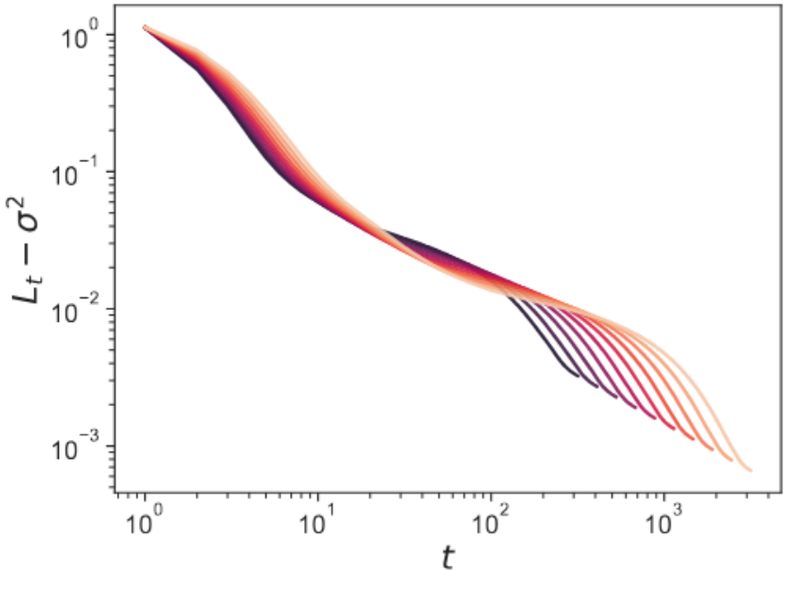


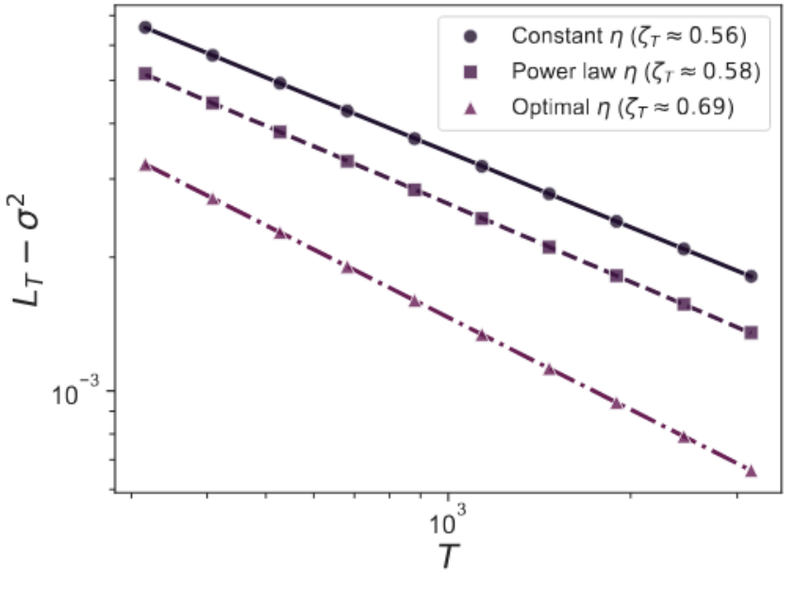




Phase II (SGD Dominated): $a > b$





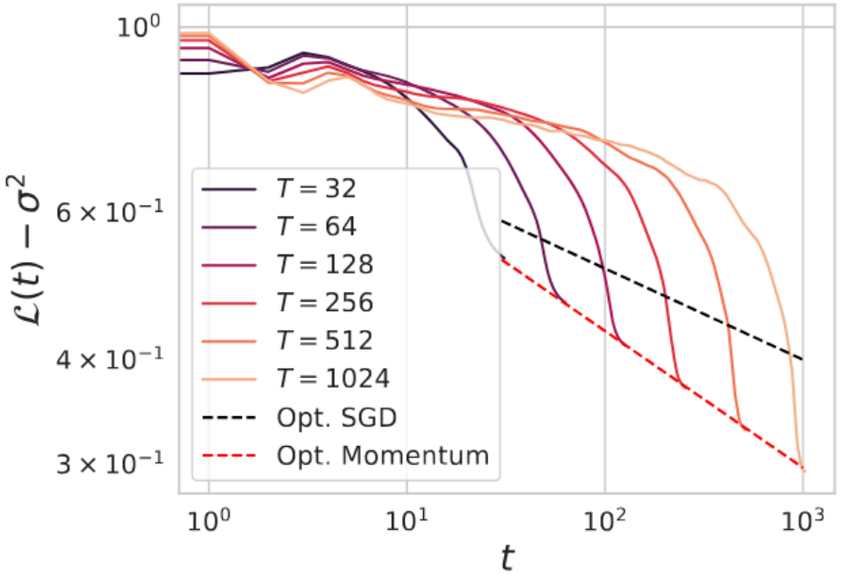


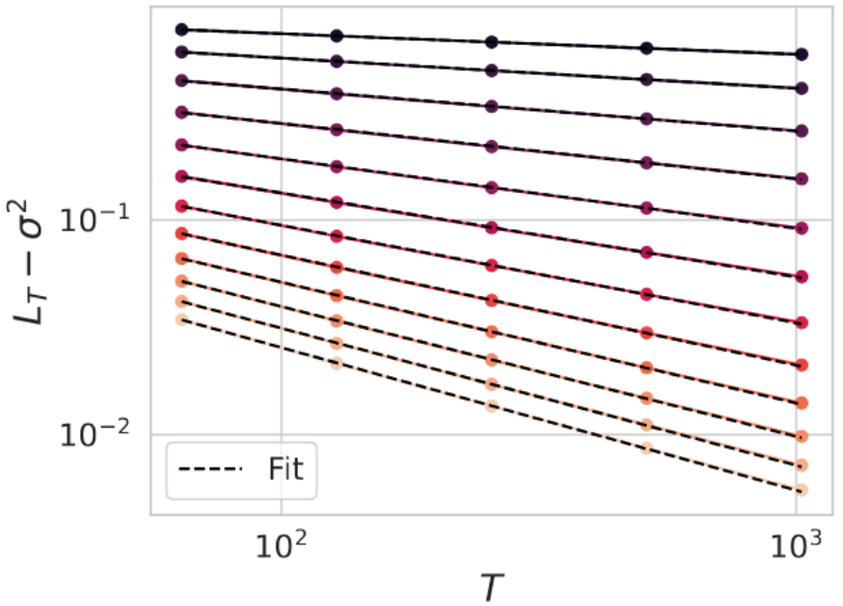
Optimal Momentum Boosts Bias Dominated Phase

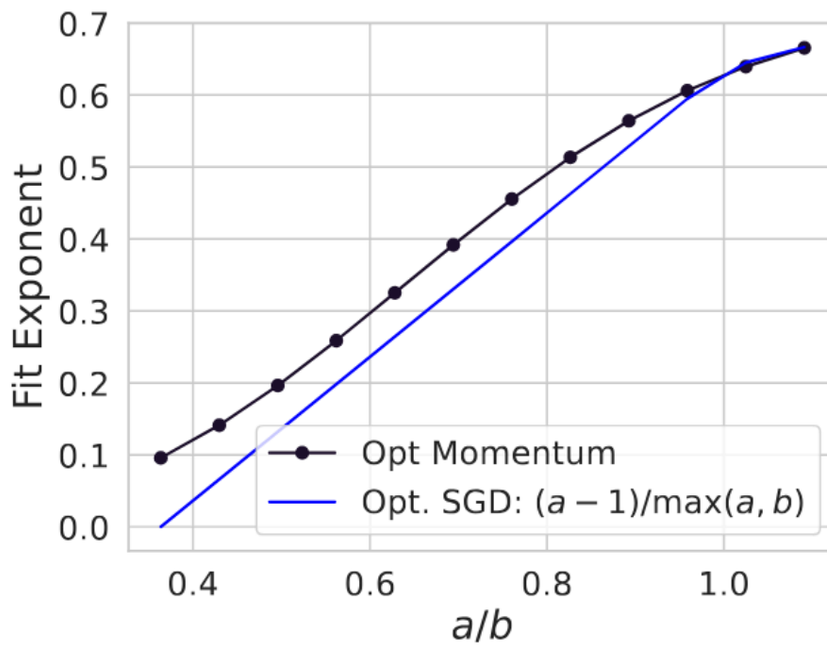
Joint optimal Control over Momentum and Learning Rate Schedules!

$$\eta_T^\star(T), \beta_T^\star(T) = \text{argmin}_{\{\eta(t),\beta(t)\}} \mathcal{L}_T[\eta(t), \beta(t)]$$

$$\mathbf{w}(t+1) = \mathbf{w}(t) - \eta(t)\mathbf{v}(t) \ , \ \mathbf{v}(t) = (1-\beta(t))\mathbf{v}(t-1) + \beta(t)\mathbf{g}(t)$$



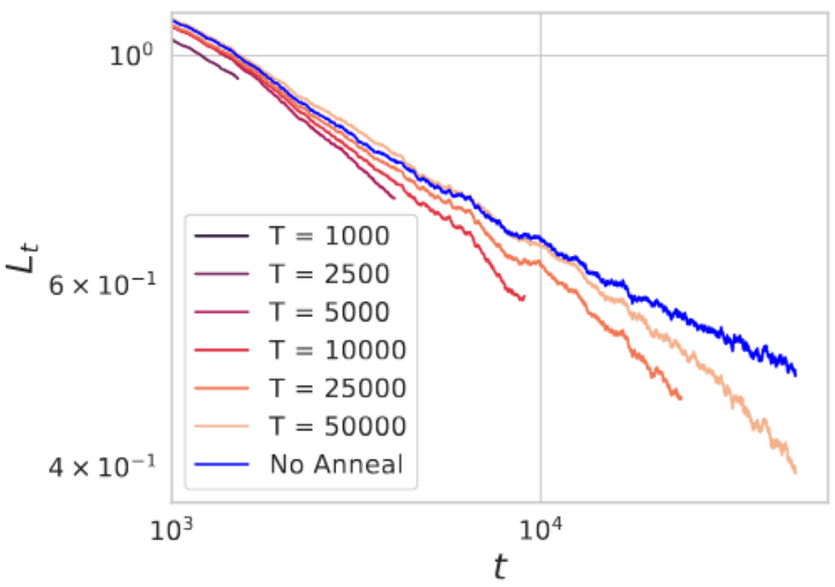


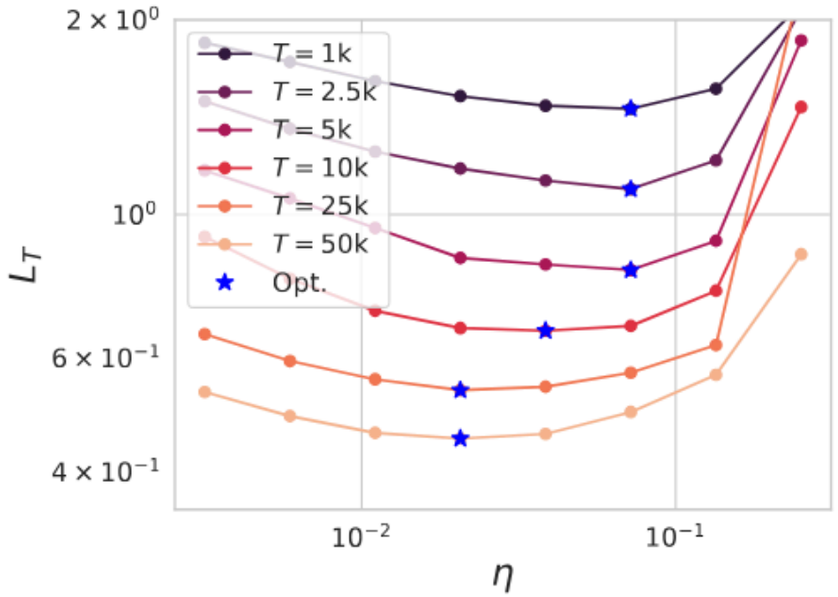


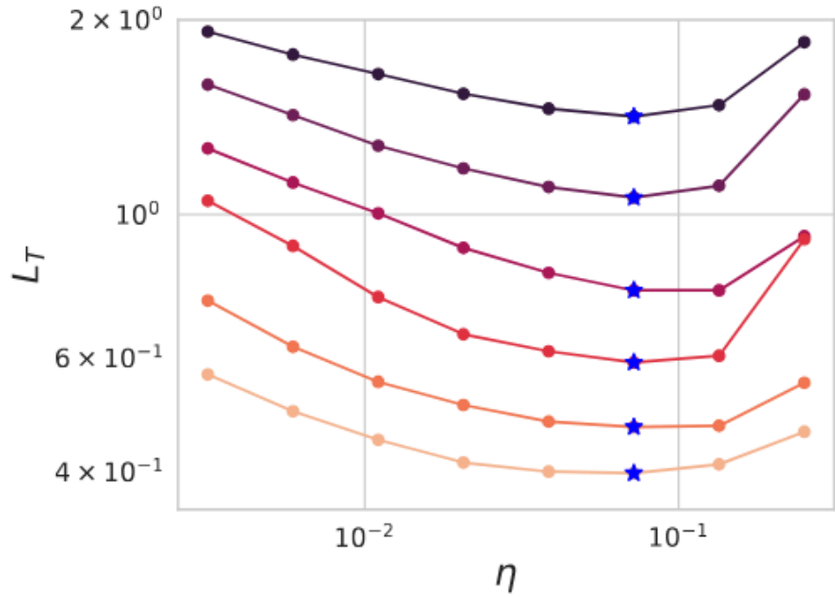
Optimal momentum beats optimal SGD in bias dominant regime, but does not saturate the optimal statistical rate.

Bias and SGD Dominated Phases in the Wild

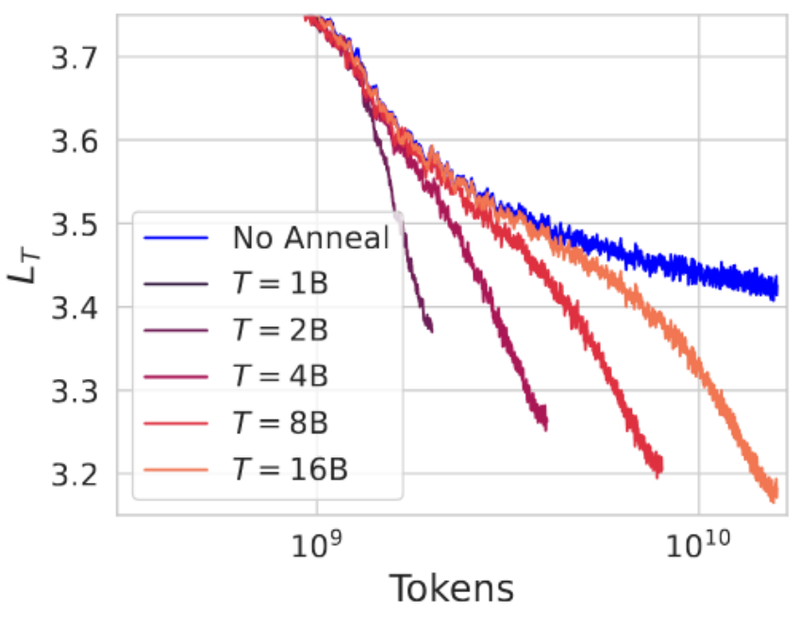
ResNets on CIFAR Image Classification ~ Bias Dominated Phase

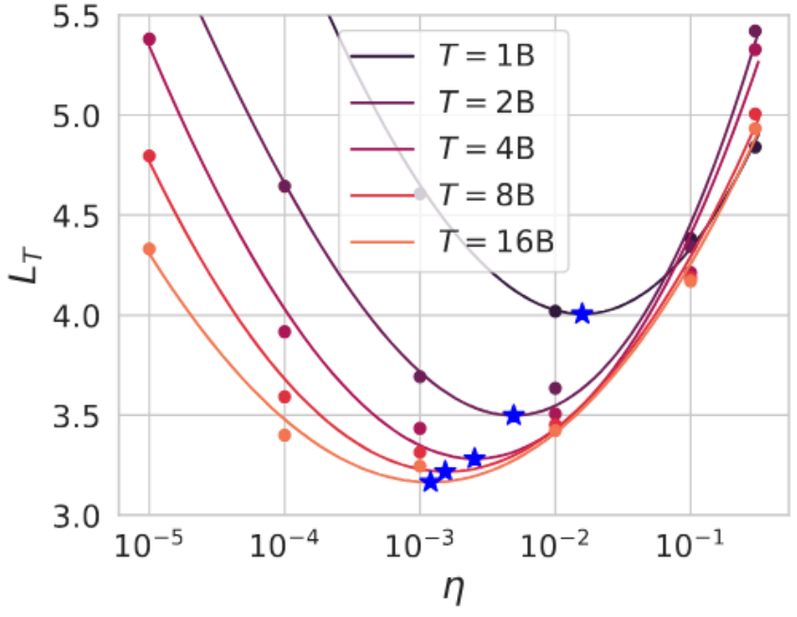


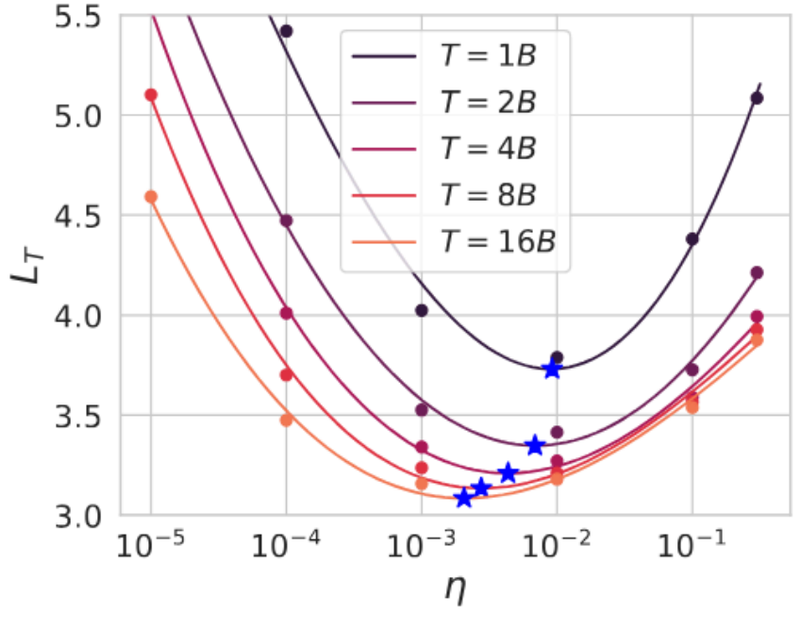




Transformers on Next Token Prediction ~ SGD Dominated

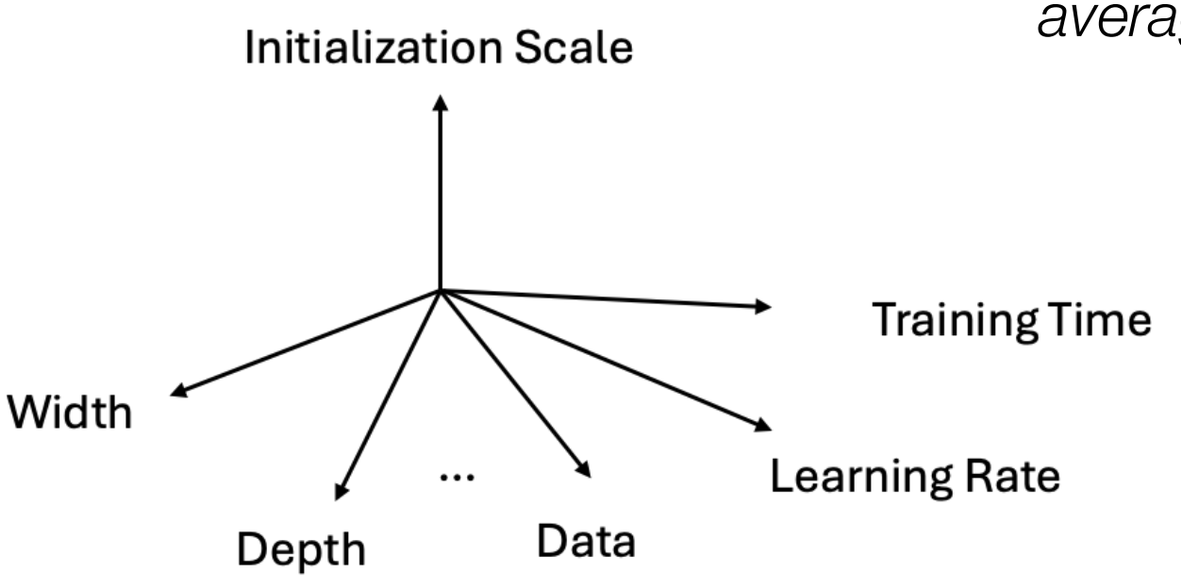




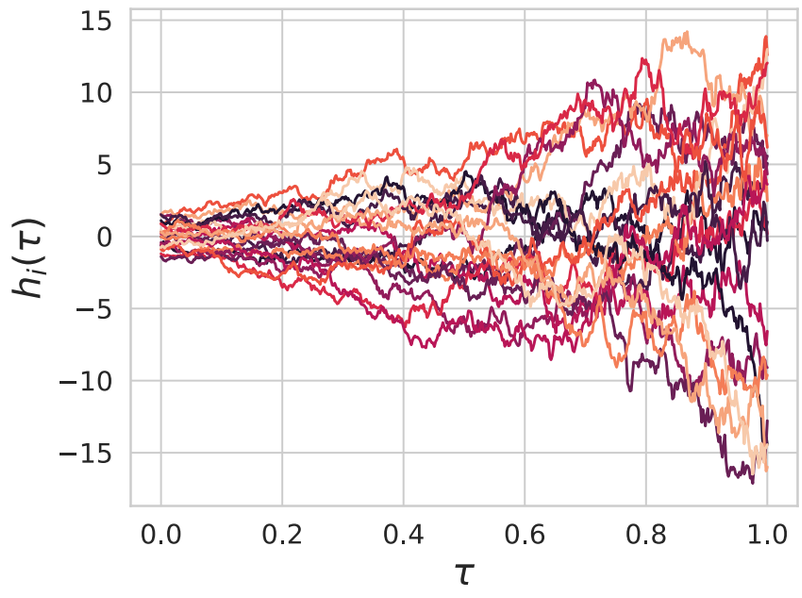


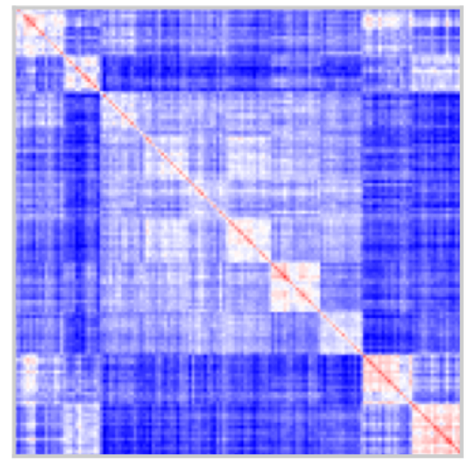
Takeaways

Scaling Limits of Neural Networks



Infinite width+depth limits of NNs are described by averages over a stochastic process (kernels, logits,etc)





Wide NNs exhibit deterministic macroscopic behavior independent of microscopic details!

This Line of Theoretical Inquiry Has Practical Consequences

Enables hyperparameter transfer (consistent optimal learning rates) and guides design choices

Much more to do on this front! *One example: How to stably scale up context length?*

Some important missing ingredients

*Limit assumes **massive** network sizes (concentration holds for widths >> batch * seq_len * steps = tokens)*

-> theory that describes opt. (or over-)trained networks params ~ data (worse perf but capture scaling law)

*HP transfer and optimal schedules **across token horizons*** (for solvable toy model, see **B,** Mori ’26)

Acknowledgments

Cengiz Pehlevan

Boris Hanin

Tianze Jiang

Clarissa Lauditi

Hamza Chaudhry

Alex Atanasov

Lorenzo Noci

Mufan (Bill) Li

Nolan Dey

Claire Zhang

Shane Bergsma

Joel Hestness

Francesco Mori

Google

cerebras

NSF

Harvard John A. Paulson School of Engineering and Applied Sciences

NIH

Kempner INSTITUTE

For the Study of Natural & Artificial Intelligence at Harvard University

Center for Brain Science

HDSI

Harvard University Center of Mathematical Sciences and Applications

TEXAS

The University of Texas at Austin