Title: Scaling Limits for Learning: Dynamics and Statics

Speakers: Blake Bordelon

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

Date: April 09, 2025 - 9:45 AM

URL: https://pirsa.org/25040085

Abstract:

In this talk, I will discuss how physics can help improve our understanding of deep learning systems and guide improvements to their scaling strategies. I will first discuss mathematical results based on mean-field techniques from statistical physics to analyze the feature learning dynamics of neural networks as well as posteriors of large Bayesian neural networks. This theory will provide insights to develop initialization and optimization schemes for neural networks that admit well defined infinite width and depth limits and behave consistently across model scales, providing practical advantages. These limits also enable a theoretical characterization of the types of learned solutions reached by deep networks, and provide a starting point to characterize generalization and neural scaling laws (see Cengiz Pehlevan's talk).

Scaling Limits for Learning: Dynamics and Statics

Blake Bordelon
Perimeter Institute
April 2025

AI is Changing The World

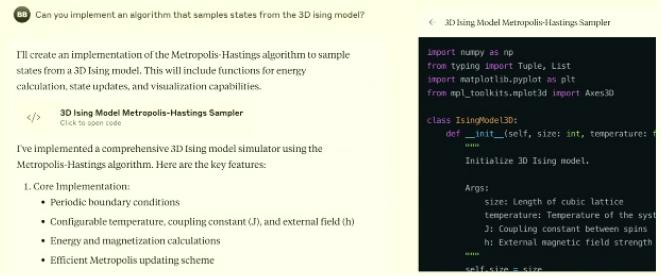
Language Models for Text and Code Generation

ChatGPT

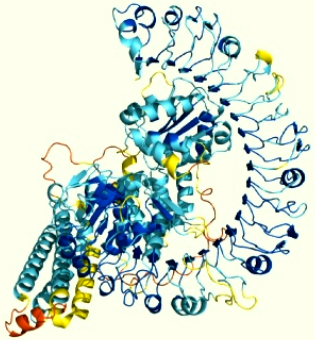
Claude

Gemini

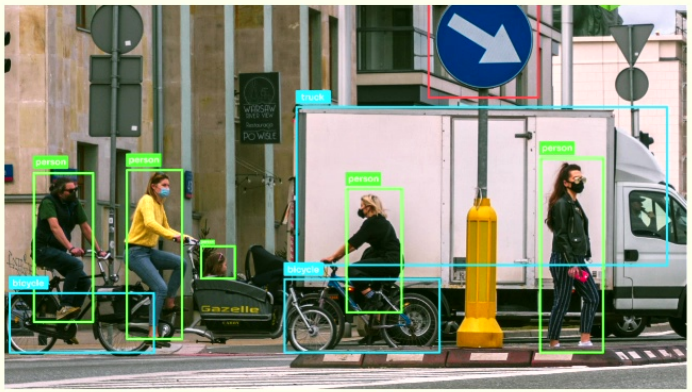
LLaMA by Meta



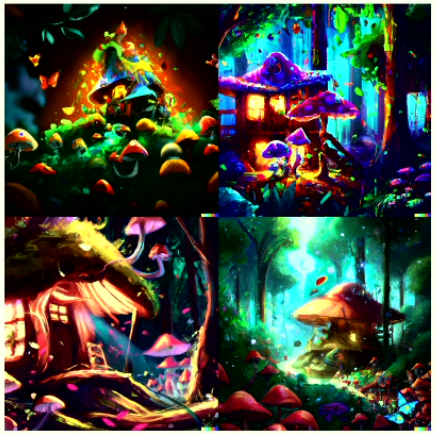
Biology (Protein Folding)



Vision Models (Object Recognition)



(Generative Image/Video Models)

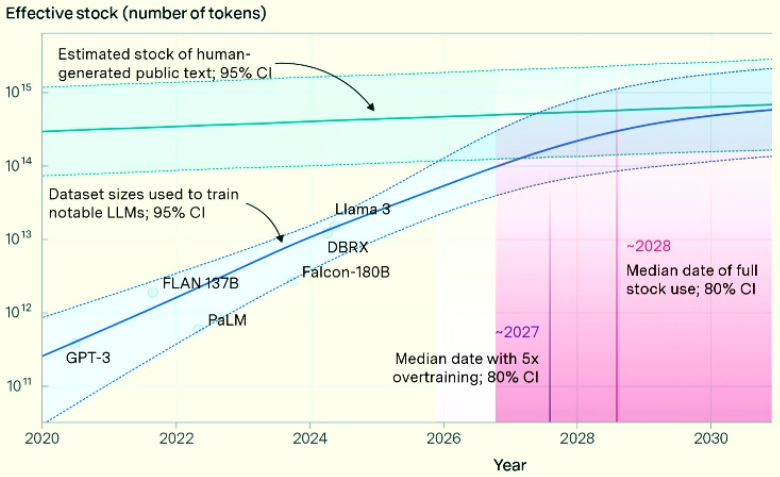


Challenges of Modern Machine Learning

Expensive: *Current approaches very data and compute hungry*

Data: *Biggest models trained on ~10 trillion out of ~100 trillion words on internet*

Human being reading 100 pages per day for 100 years would get < 3 billion words



Villalobos et al 2024

Energy: Training a single model ~500 MWh, which annual consumption of 50 US households. Even more for inference costs

Georgia Institute of Technology

AI's Energy Demands Spark Nuclear Revival

The demand for electricity to power AI data centers is skyrocketing, placing immense pressure on traditional energy sources.

2 weeks ago

Financial Times

AI set to fuel surge in new US gas power plants

The US is on the cusp of a natural gas power plant construction boom, as Big Tech turns to fossil fuels to meet the huge electricity needs...

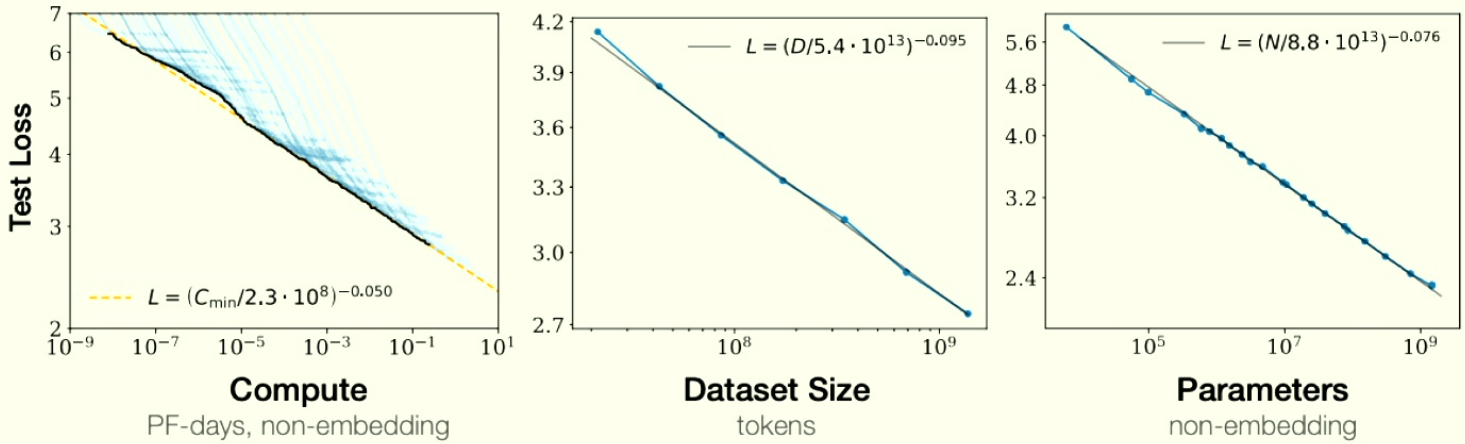
1 week ago

Money: SOTA LLMs cost ~$100M USD. Decent part of the cost is due to performing *large number of training runs*

Can we reduce some of these costs by making training more stable / predictable?

Era of Scaling in Deep Learning

How does performance depend on model size and training time? **Neural Scaling Laws**



Following these trends, 10x of compute leads to 10% reduction in loss

What are the limits of this scaling paradigm? ***Infinitely large models***

Today's talk: Statistical mechanics theory of *large* neural networks

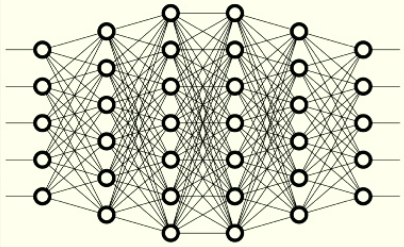
What sets these power laws? What do they depend on?

Cengiz' talk (tomorrow): Compute optimal scaling laws (convergence rates to the limits)

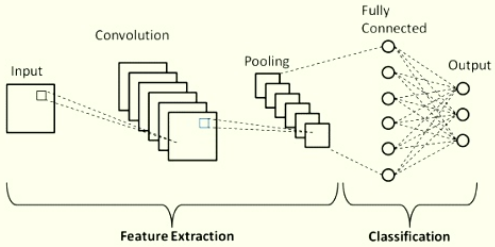
Challenges of Modern Machine Learning

Theoretical Challenges: *predictability / interpretability / principled design choices*

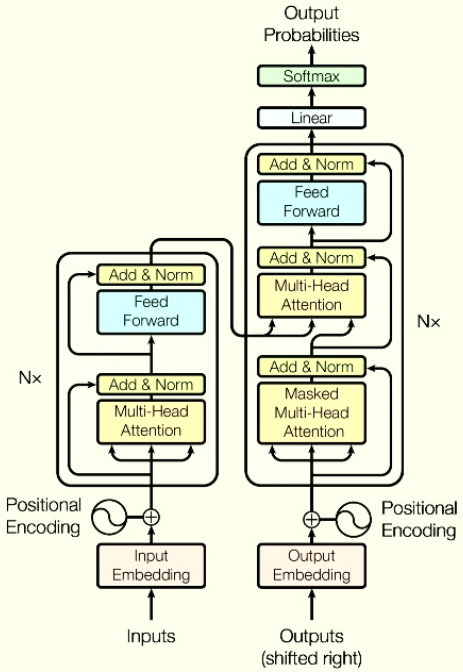
Complex architectures with billions of parameters!



Fully Connected (MLP)



Convolutional network (weights shared across spatial positions)



Transformer (learnable attention maps across spatial positions)

How to initialize and optimize models to be predictable, and monotonically improving under scaling?

Why Physics for Modern ML Problems?

Philosophy: Accurate Approximations + Insights of Simple Models >> Rigor

Leo Breiman
Statistics Department, University of California, Berkeley, CA 94305;
e-mail: leo@stat.berkeley.edu

1995!

Reflections After Refereeing Papers for Neurips

- Why don't heavily parameterized neural networks overfit the data?
- What is the effective number of parameters?
- Why doesn't backpropagation head for a poor local minima?
- When should one stop the backpropagation and use the current parameters?

These are problems about *dynamics*, *optimization*, and *statistics* better suited to the techniques and ideas of physics.

Understanding deep learning is also a job for physicists

Automated learning from data by means of deep neural networks is finding use in an ever-increasing number of applications, yet key theoretical questions about how it works remain unanswered. A physics-based approach may help to bridge this gap.

Lenka Zdeborová

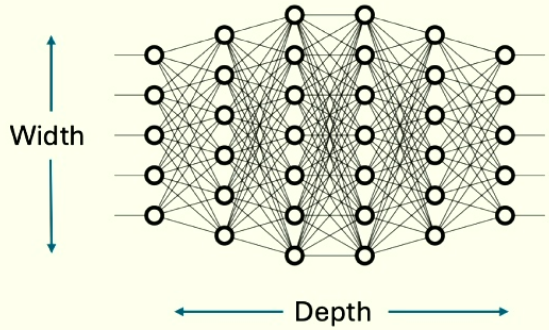
Focus on science: don't (just) chase empirical benchmarks, tight feedback between experiment and theory

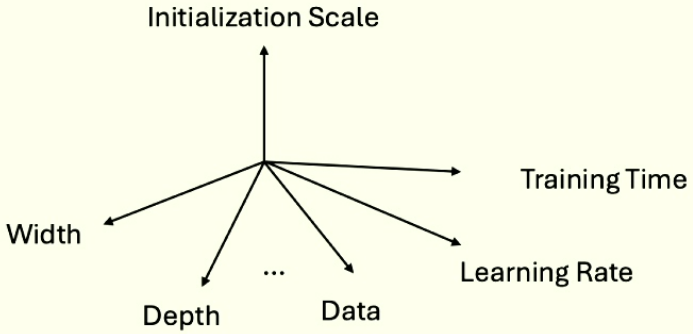
Don't fear the infinite: microscopic -> macroscopic descriptions

Physics Approach: Find theoretical descriptions of the solutions that randomly initialized networks trained with optimization algorithms *actually converge to* in typical/average case

Modern Machine Learning Problems

Scaling Limits of Neural Networks





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

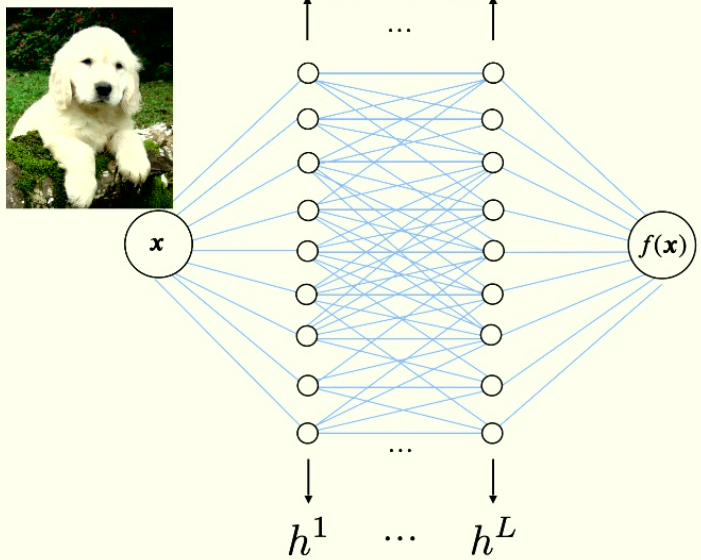
Neurons (particles) interacting at finite width N

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

Practical extensions: Hyperparameter transfer to reduce training costs during scaling

Improved theory and practice for transformer scaling

Training Wide Neural Networks



~~"Cat"~~

"Dog"

$$h_i^1 = \frac{1}{\sqrt{D}} \sum_{j=1}^{D} W_{ij}^0 \, x_j$$

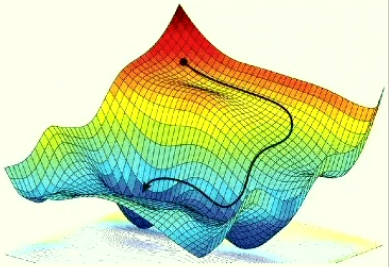
$$h_i^{\ell+1} = \frac{1}{\sqrt{N}} \sum_{j=1}^{N} W_{ij}^{\ell} \phi(h_j^{\ell})$$

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

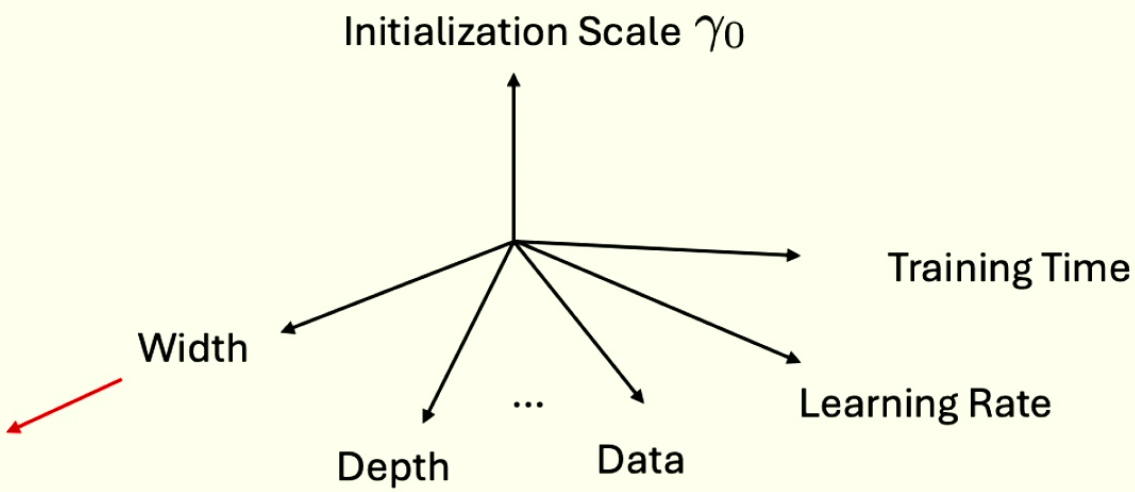
Random Initial Weights: $W_{ij}^{\ell}(0) \sim \mathcal{N}(0,1)$ at initialization.

Non-convex High Dimensional Optimization: Weights are updated so the network *fits data!*

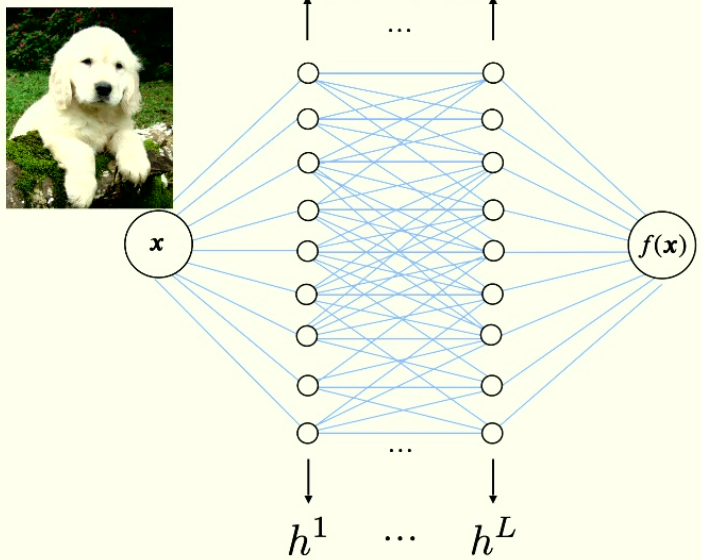
$$\frac{d}{dt} W(t) = -\frac{d}{dW} \underbrace{L(\{W^{\ell}\})}_{\text{Loss Function}} \qquad L = \sum_{\mu=1}^{P} \left(\underbrace{f(x_\mu)}_{\text{output}} - \underbrace{y(x_\mu)}_{\text{target}} \right)^2$$



Large Width Limits



Training Wide Neural Networks



~~"Cat"~~

"Dog"

$$h_i^1 = \frac{1}{\sqrt{D}} \sum_{j=1}^{D} W_{ij}^0 \, x_j$$

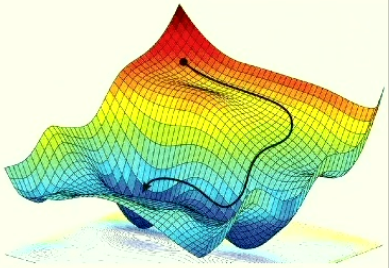
$$h_i^{\ell+1} = \frac{1}{\sqrt{N}} \sum_{j=1}^{N} W_{ij}^{\ell} \phi(h_j^{\ell})$$

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

Random Initial Weights: $W_{ij}^{\ell}(0) \sim \mathcal{N}(0,1)$ at initialization.

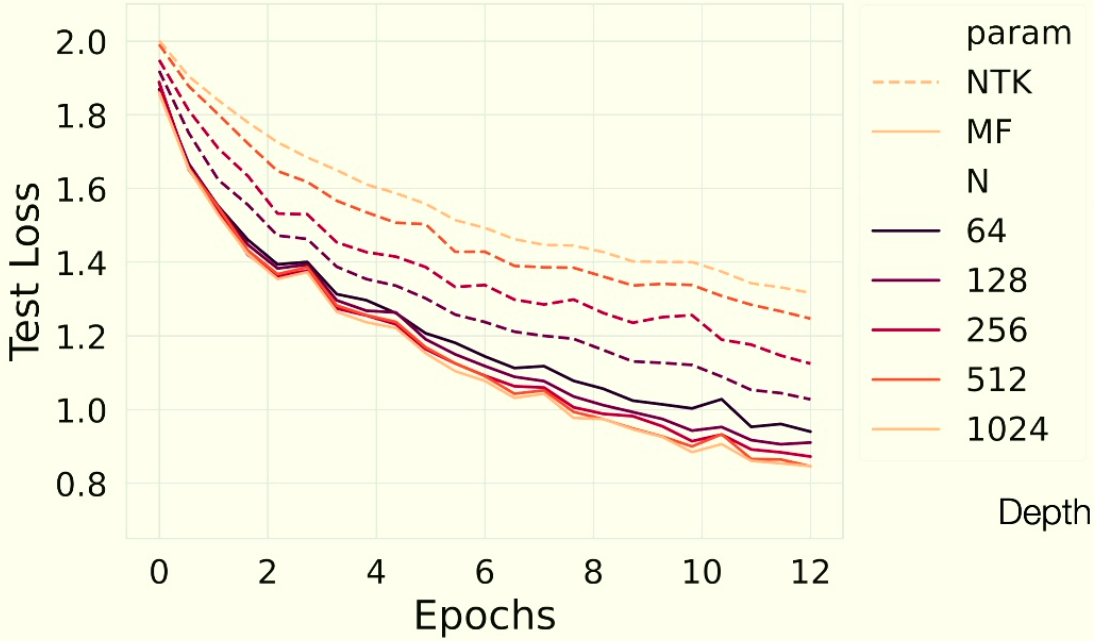
Non-convex High Dimensional Optimization: Weights are updated so the network *fits data!*

$$\frac{d}{dt} W(t) = -\frac{d}{dW} \underbrace{L(\{W^{\ell}\})}_{\text{Loss Function}} \qquad L = \sum_{\mu=1}^{P} \left(\underbrace{f(x_\mu)}_{\text{output}} - \underbrace{y(x_\mu)}_{\text{target}} \right)^2$$



How to characterize/predict/summarize what the model learns?

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

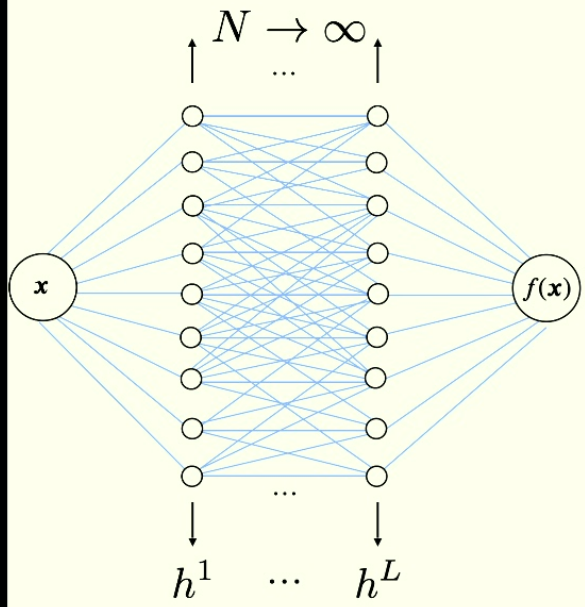
Common scaling practice: $\gamma = \mathcal{O}(1)$ $\Longrightarrow$ Slower training as N increases

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

Mean field also displays faster convergence to limiting behavior

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

How to Scale Up Width? Dimensional Analysis



$$h_i^1 = \frac{1}{N^{a_0}} \sum_{j=1}^{D} W_{ij}^0 x_j \qquad W_{ij}^0(0) \sim \mathcal{N}(0, N^{-2b_0})$$

$$h_i^{\ell+1} = \frac{1}{N^{a_\ell}} \sum_{j=1}^{N} W_{ij}^\ell \phi(h_j^\ell) \quad W_{ij}^\ell(0) \sim \mathcal{N}(0, N^{-2b_\ell})$$

$$f = \frac{1}{\gamma N^{a_L}} \sum_{i=1}^{N} w_i^L \phi(h_i^L) \quad w_i^L(0) \sim \mathcal{N}(0, N^{-2b_L})$$

$$\frac{d}{dt} W_{ij}^\ell(t) = -\eta_0 \gamma^2 \frac{\partial}{\partial W_{ij}^\ell} \mathcal{L} \qquad \gamma = \gamma_0 N^c$$

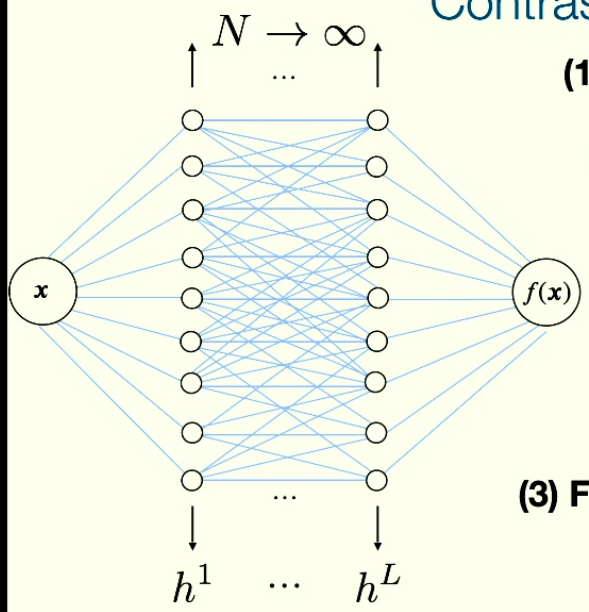
(1) Stable Forward & Backward Passes

$$\left\langle (h_i^\ell)^2 \right\rangle = \Theta(1) \qquad a_0 + b_0 = 0 \qquad a_\ell + b_\ell = \frac{1}{2} \qquad \ell \in \{1, ..., L\}$$

(2) Function Learning: $\frac{d}{dt} f = \Theta(1) \quad a_L = \frac{1}{2} \quad a_\ell + b_L = \frac{1}{2} \quad a_0 + b_L = 0$

(3) Feature Learning: $\frac{d}{dt} h_i^\ell = \Theta(1) \implies c = \frac{1}{2}$ See Yang & Hu '21, **B** & Pehlevan '22

Contrasting Parameterizations



(1) Stable Forward & Backward

$$a_0 + b_0 = 0 \qquad a_\ell + b_\ell = \frac{1}{2} \qquad \ell \in \{1, ..., L\}$$

(2) Function Learning: $\frac{d}{dt} f = \Theta(1)$

$$a_L = \frac{1}{2} \quad a_\ell + b_L = \frac{1}{2} \quad a_0 + b_L = 0$$

(3) Feature Learning: $\frac{d}{dt} h_i^\ell = \Theta(1) \implies c = \frac{1}{2}$

Standard Parameterization (SP) *(PyTorch Default with no LR Tuning)*

$a_0 = 0, b_0 = 0, a_\ell = 0, b_\ell = \frac{1}{2}, c = 0$ Satisfies (1) but not (2) or (3) , unstable

NTK Parameterization *(Jacot et al '19)*

$a_0, b_0 = 0$, $a_\ell = \frac{1}{2}, b_\ell = 0, c = 0$

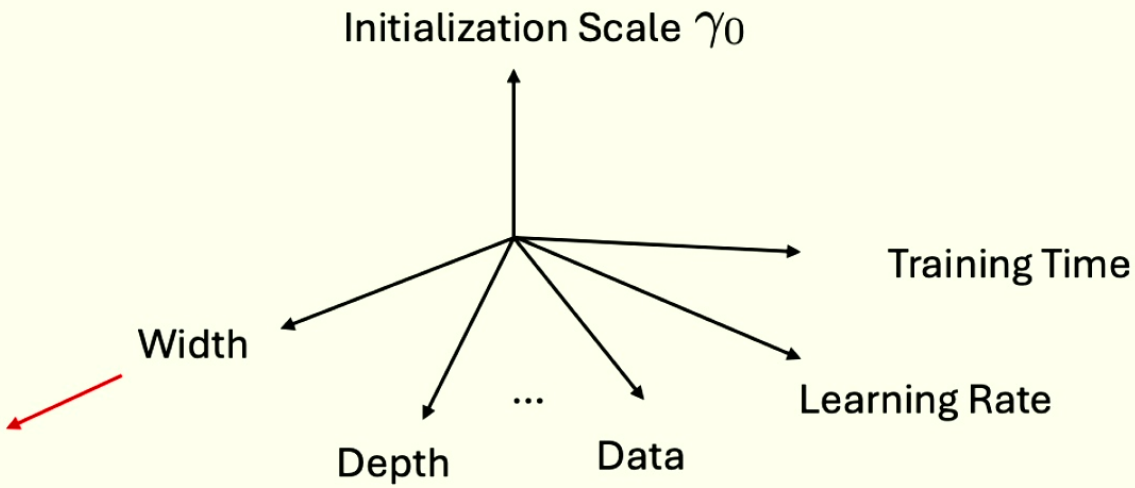
Satisfies (1) and (2) but not (3)
(stable but no feature learning in the limit)

MF / muP *(Geiger et al '19, Yang & Hu '21)*

$a_0, b_0 = 0, a_\ell = \frac{1}{2}, b_\ell = 0, c = \frac{1}{2}$

See Yang & Hu '21, **B** & Pehlevan '22

Large Width Limits



How to mathematically characterize the dynamics of training in the infinite width limit that satisfies all 3 constraints?

We need some physics!

Primer on Dynamical Mean Field Theory

Random Coupled Systems in High Dimensions

Spin Glass Example: $\mathcal{H}(\{s_i\}) = -\frac{1}{2\sqrt{N}} \sum_{ij} J_{ij} s_i s_j \qquad J_{ij} = J_{ji} \sim \mathcal{N}(0,1)$

Langevin Dynamics On sphere $\partial_t s_i(t) = \frac{1}{\sqrt{N}} \sum_{j=1}^{N} J_{ij} s_j(t) - \lambda(t) s_i(t) + j_i(t)$

Correlation Function

$$C(t,t') = \frac{1}{N} \sum_{i=1}^{N} \langle s_i(t) s_i(t') \rangle$$

Response Function

$$R(t,t') = \frac{1}{N} \sum_{i=1}^{N} \left\langle \frac{\delta s_i(t)}{\delta j_i(t')} \right\rangle$$

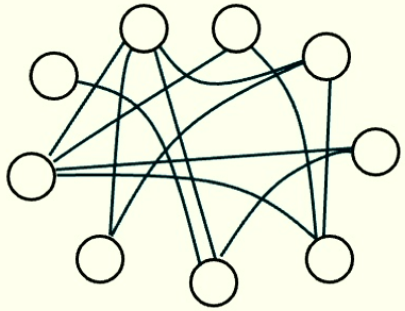
Sompolinsky & Zippelius '82, Kurchan & Cugliandolo '93, Bouchaud et al '97

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

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

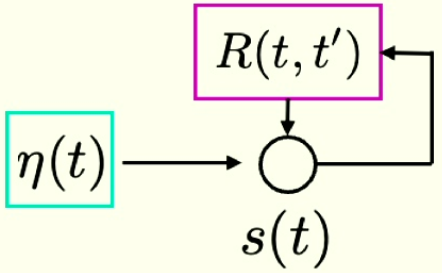
Primer on Dynamical Mean Field Theory

Random Coupled -> Uncoupled System in the Limit



$$\sim$$

$$N \to \infty$$



Correlation and Response Form Closed System from Single Site Picture

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

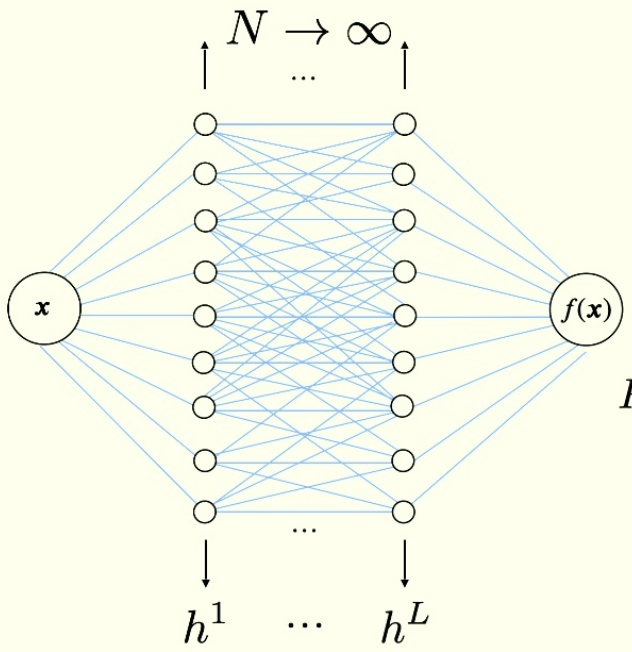
Many theoretical methods give this result

1. Saddle point of a Martin Siggia Rose Path integral $Z = \int dC dR \exp\left(-N\mathcal{S}(C,R)\right)$

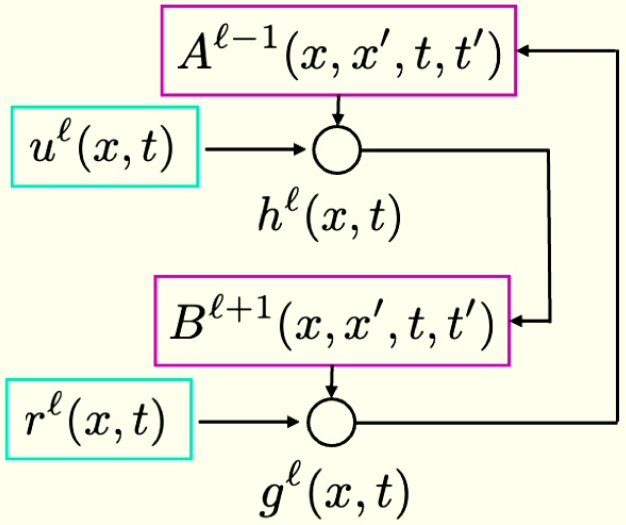
2. Cavity (add new site) argument, compute self-feedback through other sites

3. When dynamical system is *linear*, can use random matrix theory / deterministic equivalence

Mean Field Theory for Deep Network Training



$$\sim \quad N \to \infty \quad P = \Theta_N(1)$$



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

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

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

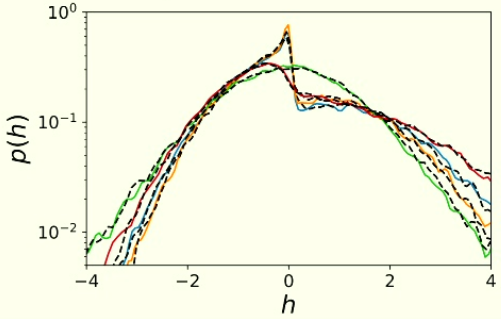
B & Pehlevan '22, '23, '24

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

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):

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

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

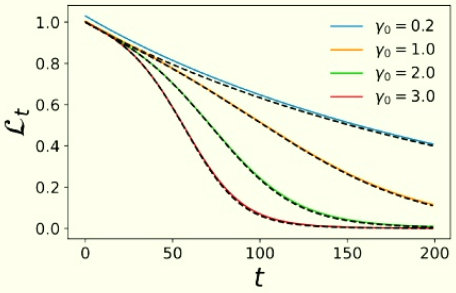
Output Dynamics: The outputs of the network evolve in terms of these correlation functions

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

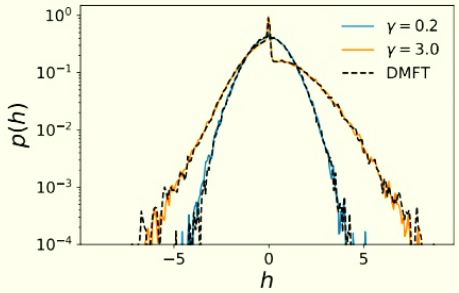
B, Pehlevan ‘22

Lazy vs Rich Operating Regimes

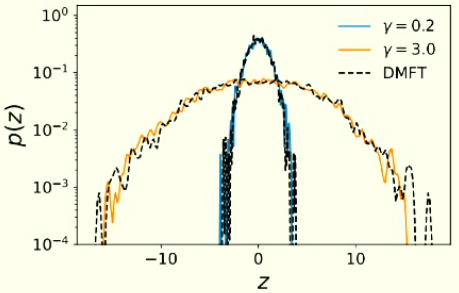
Richness: Infinite width equations depend crucially on an output multiplier γ_0



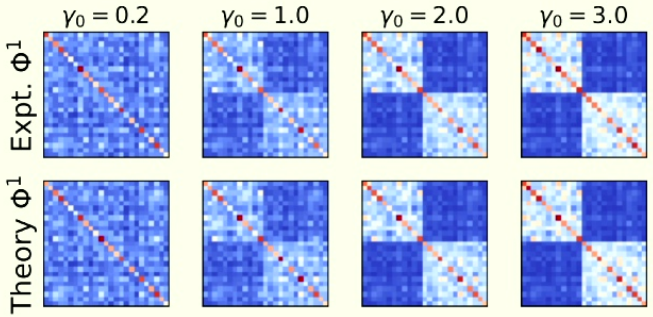
Loss Dynamics



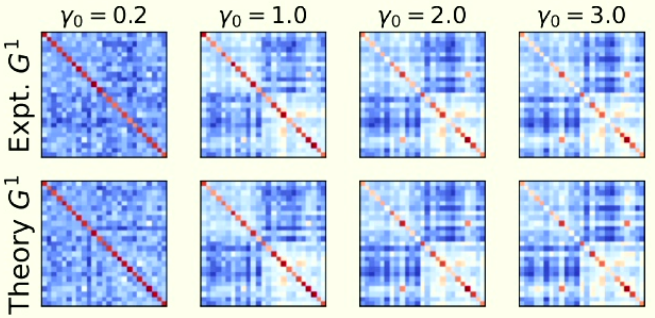
Final h Distribution



Final z Distribution



Final Φ^1 Kernels



Final G^1 Kernels

Lazy Learning = Constant Neural Tangent Kernel

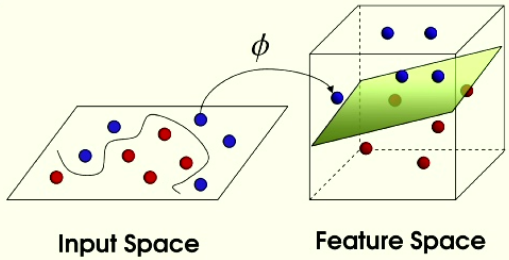
Lazy Limit: the $\gamma_0 \to 0$ limit gives a dramatic simplification to the DMFT equations

Internal variables constant $$h^\ell(x,t) = u^\ell(x) \sim \mathcal{N}(0, \Phi^{\ell-1}(x,x'))$$

Linear dynamics for outputs $$\frac{d}{dt} f(x,t) = -\sum_{\mu=1}^{P} K(x,x_\mu)\left(f(x_\mu,t) - y_\mu\right)$$

Neural Tangent Kernel
(Jacot et al 2019) $$K(x,x') = \sum_\ell G^{\ell+1}(x,x')\Phi^\ell(x,x')$$

Linear method in infinite dimensions



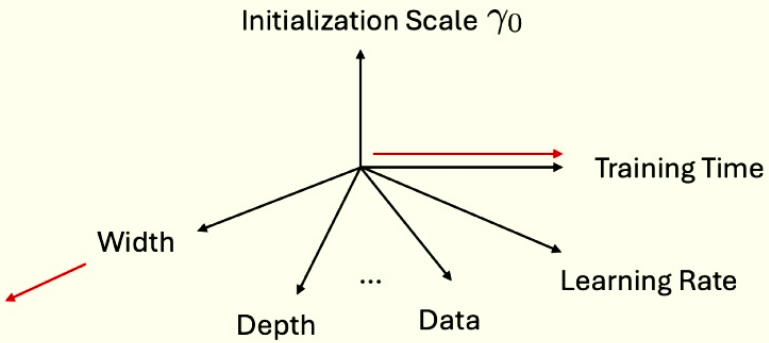
Many theoretical works on this limit

Jacot et al 2019, Hanin 2019, Arora et al 2020, Lee et al 2020, …

Weak feature expansion: perturbation theory in γ_0

Dyer & Gur Ari 2021, Roberts, Yaida, Hanin 2021, **B** & Pehlevan ’22, …

Statics: Is there an equilibrium distribution?



DMFT studies $\lim_{N\to\infty}$ *with* t *fixed*

Can we say something about $t \to \infty$?

Langevin Dynamics: Add noise and weight decay to the dynamics

$$\frac{d}{dt}W_{ij}^{\ell}(t) = -\eta\gamma^2 \frac{\partial}{\partial W_{ij}^{\ell}}\mathcal{L}(\{\mathbf{W}^{\ell}\}) - \beta^{-1}W_{ij}^{\ell}(t) + \sqrt{2\beta^{-1}}\epsilon_{ij}^{\ell}(t)$$

Equilibrium Distribution: Take the $t \to \infty$ limit first, converges to a Gibbs measure

$$p(\{\mathbf{W}^{\ell}\}) \propto \exp\left(-\beta\gamma^2\mathcal{L}(\{\mathbf{W}^{\ell}\}) - \frac{1}{2}\sum_{\ell}|\mathbf{W}^{\ell}|^2\right)$$

Prior work on Statics

Equilibrium Distribution: Take the $t \to \infty$ limit first, converges to a Gibbs measure

$$p(\{\mathbf{W}^\ell\}) \propto \exp\left(-\beta\gamma^2 \mathcal{L}(\{\mathbf{W}^\ell\}) - \frac{1}{2}\sum_\ell |\mathbf{W}^\ell|^2\right)$$

Lazy Limit of Bayesian Networks (NNGP): $N \to \infty$ Fixed γ, P

$$\Phi^\ell_{\mu\nu} = \langle \phi(h_\mu)\phi(h_\nu)\rangle_{\mathbf{h}\sim\mathcal{N}(0,\mathbf{\Phi}^{\ell-1})}$$ Kernels behave same as prior

Neal '95, Lee, Bahri et al 2018, Novak, Xiao et al 2019, …

Weak Feature Expansion: leading order corrections in γ^2/N

$$\Phi = \Phi_0 + \frac{\gamma^2}{N}\Phi_1 + \frac{\gamma^4}{N^2}\Phi_2 + \ldots$$

Zavatone-Veth et al 2021, Yaida 2021, Roberts Yaida Hanin 2021, …

Proportional Limit: Scaling limit where data P and width N diverge with $P = \alpha N$

Scale renormalization effect $\Phi \sim c(\alpha)\Phi_0$

New Results on muP/MF Statics

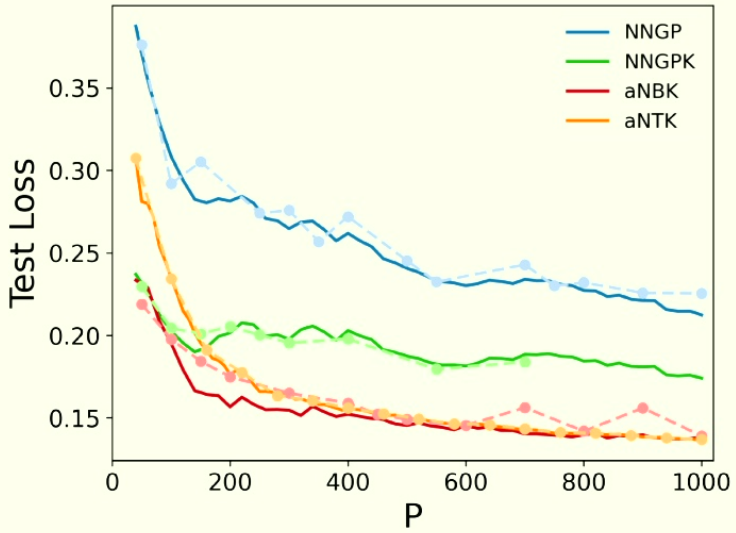
Equilibrium Distribution: Take the $t \to \infty$ limit first, converges to a Gibbs measure

$$p(\{\mathbf{W}^\ell\}) \propto \exp\left(-\beta\gamma^2 \mathcal{L}(\{\mathbf{W}^\ell\}) - \frac{1}{2}\sum_\ell |\mathbf{W}^\ell|^2\right)$$

muP Scaling Limit for BNN: $N \to \infty \quad \gamma = \gamma_0\sqrt{N}$

Kernels solve a set of saddle point equations
Similar to the DMFT equations without response

$$\min_{\{\Phi^\ell\}} \max_{\{\hat{\Phi}^\ell\}} \mathcal{S}[\{\Phi^\ell, \hat{\Phi}^\ell\}]$$



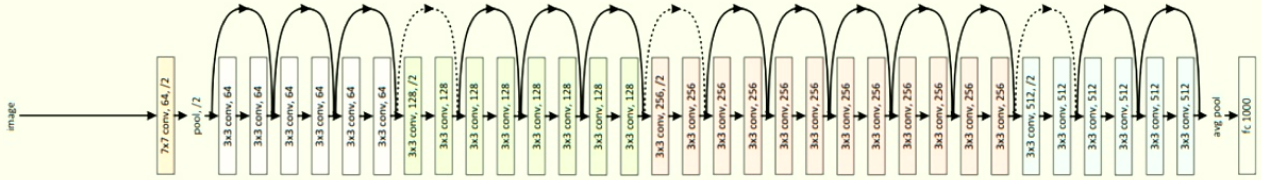
Mean field equations exact for this limit

This limit performs much better than NNGP on CIFAR (image data)

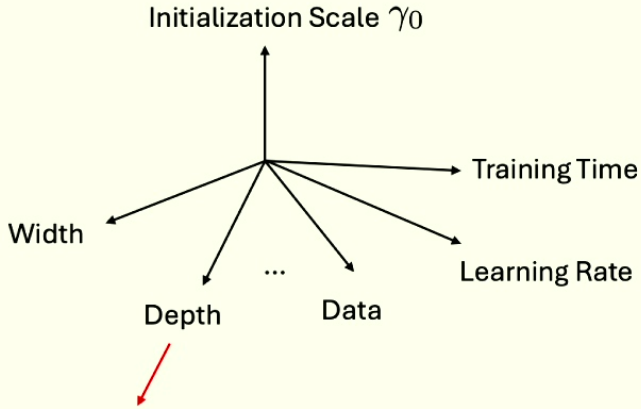
Lauditi, **B,** Pehlevan 2025

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!

$$\mathbf{h}^{\ell+1} = \mathbf{h}^{\ell} + \frac{\beta}{\sqrt{NL}} \mathbf{W}^{\ell} \phi(\mathbf{h}^{\ell})$$

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

The Large Depth and Width Limit

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

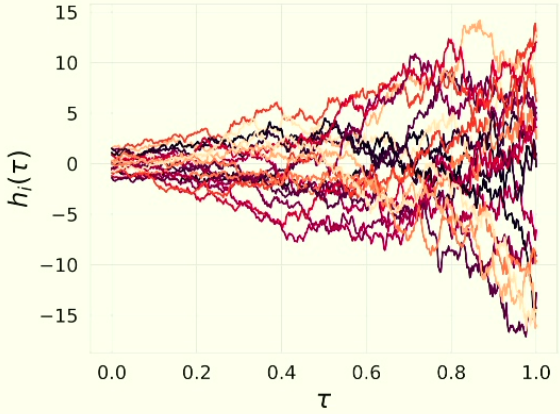
Result: Non-random limit for all DMFT observables $q_{\infty,\infty} = \lim_{N,L\to\infty} q_{N,L}$

Intuition pump: characterize initialization

Neurons follow geometric brownian motion

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

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

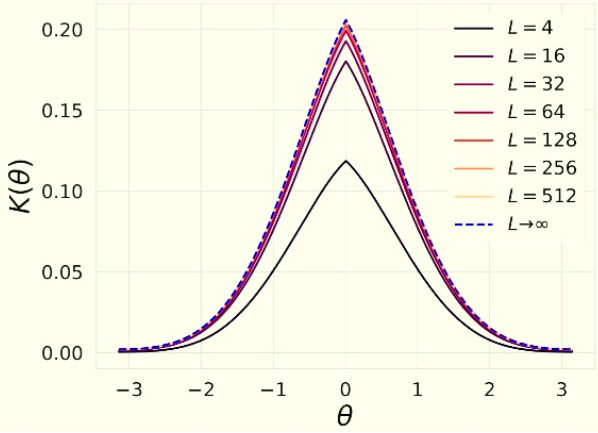


Introduce "layer time" $\tau = \frac{\ell}{L} \in [0,1]$

$$\lim_{L\to\infty} H^{L\tau} \equiv H(\tau)$$

Finite Difference to Differential Equation

$$\partial_{\tau} H(\tau) = \beta^2 \left\langle \phi(h)^2 \right\rangle_{h\sim\mathcal{N}(0,H(\tau))}$$

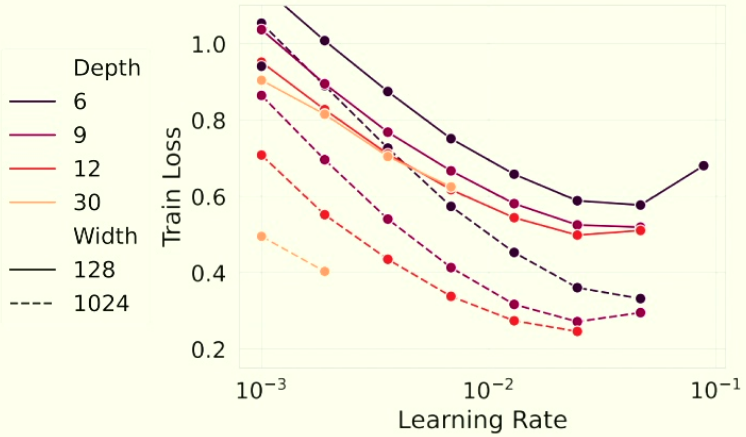


Practical Application: Hyperparameter Transfer

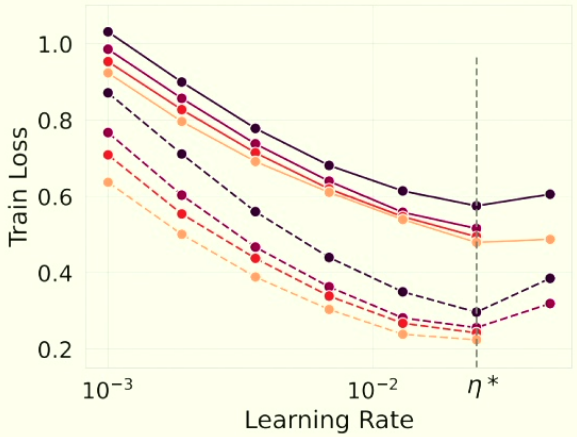
Sweep HPs in small models and then scale up with improved performance (Yang et al 2022)

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

No branch scaling (common practice)



MF+ scaled branches



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

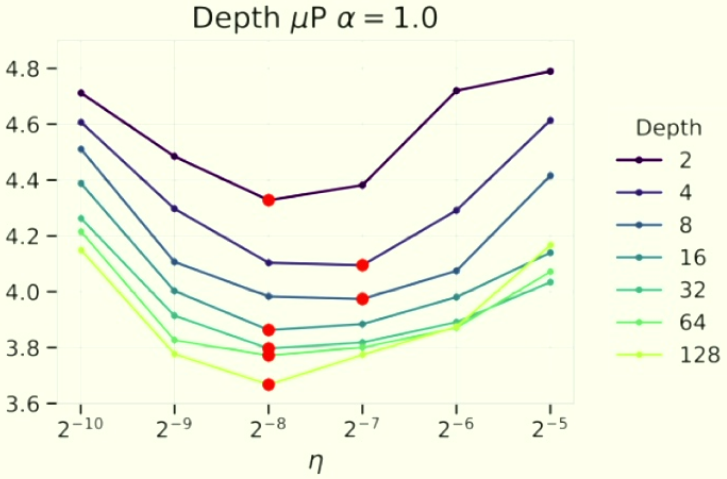
Hyperparameters *transfer* across widths and depths

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

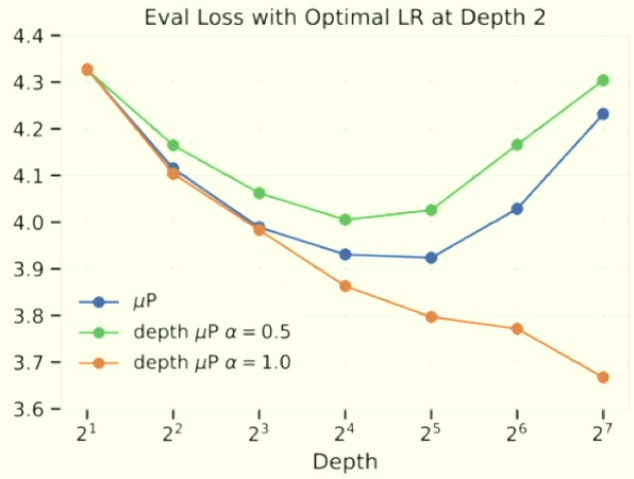
Industry starting to pursue this direction (Open AI, Google Deepmind, Cerebras, etc)

Depth Limits of Transformers

Stable HPs



Improved deep models

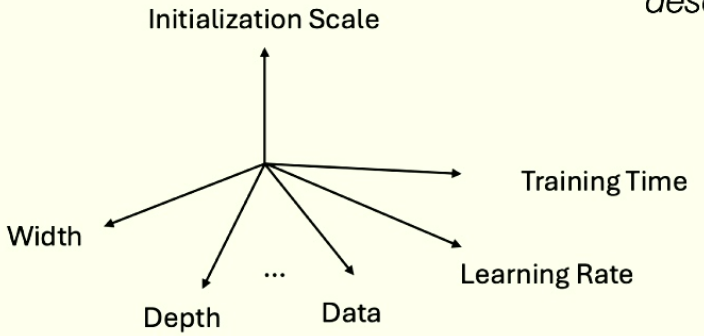


Our scaling ideas are useful for LLMs at large scale (in prep)!

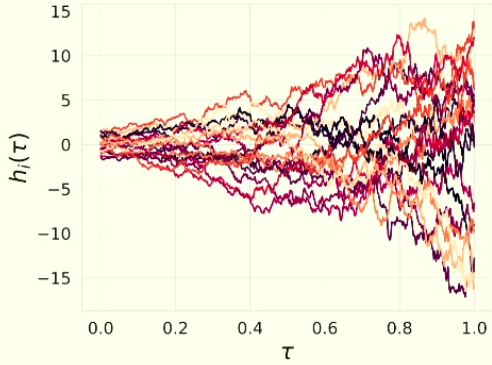
cerebras

Takeaways

Scaling Limits of Neural Networks



Infinite width + depth limits of neural networks are described by stochastic processes for each neuron



Averages over neurons become deterministic and determine the macroscopic behavior of the network

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!

How Are Finite Models Different Than these Scaling Limits?

Come back tomorrow for Cengiz' talk , a simple solvable model of neural scaling laws

Acknowledgments

Cengiz Pehlevan
Abdul Canatar
Hamza Chaudhry
Lorenzo Noci
Mufan (Bill) Li
Boris Hanin
Nolan Dey
Claire Zhang
Shane Bergsma
Joel Hestness

Google

cerebras

NSF

Harvard John A. Paulson School of Engineering and Applied Sciences

NIH

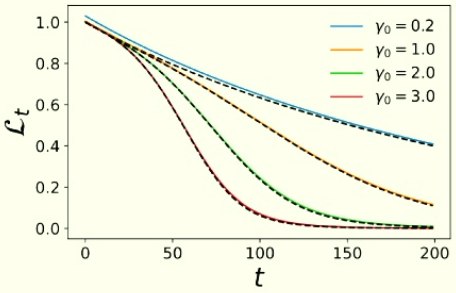
Kempner INSTITUTE
For the Study of Natural & Artificial Intelligence at Harvard University

Center for Brain Science

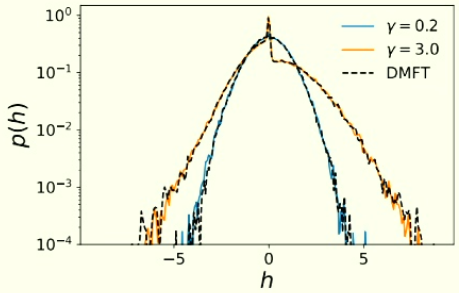
HDSI

Lazy vs Rich Operating Regimes

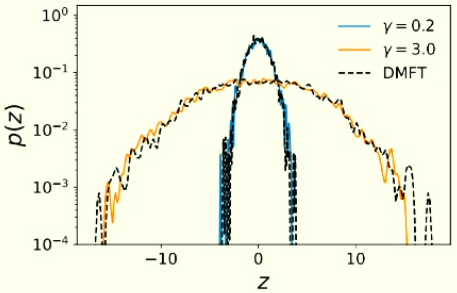
Richness: Infinite width equations depend crucially on an output multiplier γ_0



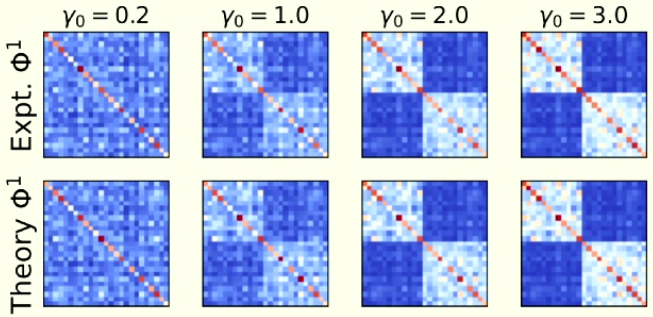
Loss Dynamics



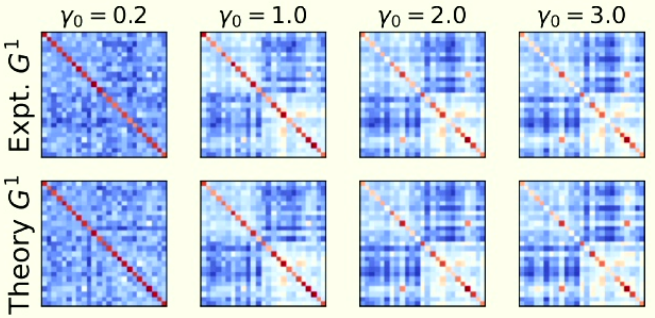
Final h Distribution



Final z Distribution



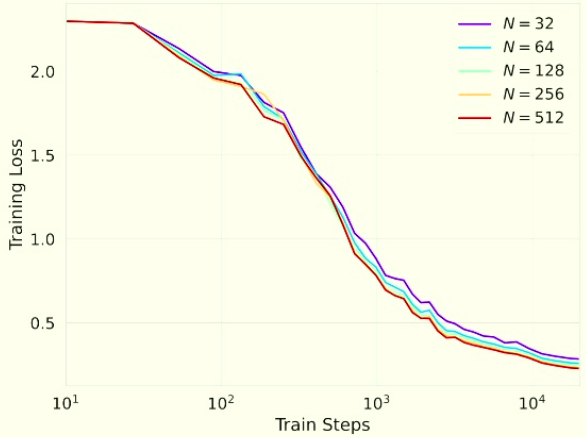
Final Φ^1 Kernels



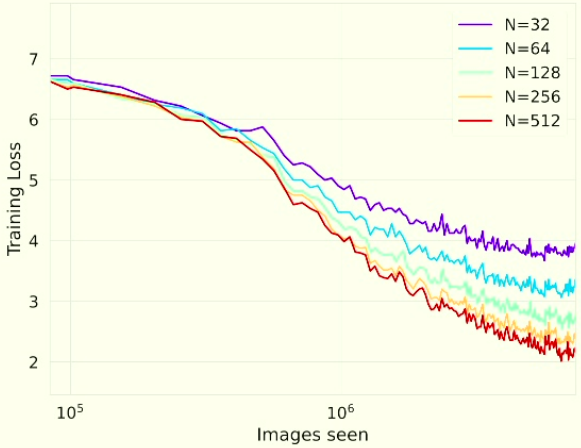
Final G^1 Kernels

Realistic Architectures and Datasets in Online Training

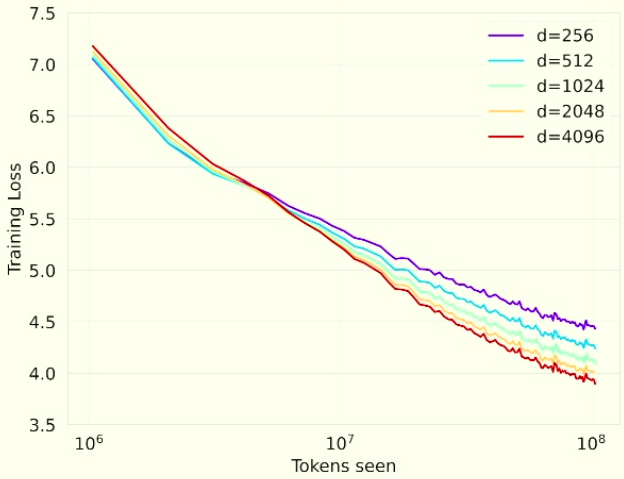
ResNets on CIFAR-5M



ImageNet



Transformer on Wikitext-103



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