Title: Towards a “Theoretical Minimum” for Physicists in AI

Speakers: Yonatan Kahn

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

Date: April 09, 2025 - 2:15 PM

URL: https://pirsa.org/25040089

Abstract:

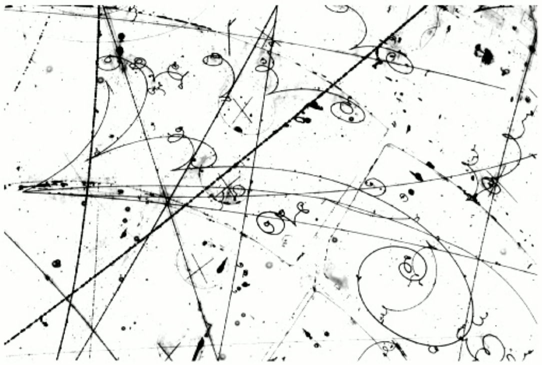
As progress in AI hurtles forward at a speed seldom seen in the history of science, theorists who wish to gain a first-principles understanding of AI can be overwhelmed by the enormous number of papers, notational choices, and assumptions in the literature. I will make a pitch for developing a “Theoretical Minimum” for theoretical physicists aiming to study AI, with the goal of getting members of our community up to speed as quickly as possible with a suite of standard results whose validity can be checked by numerical experiments requiring only modest compute. In particular, this will require close collaboration between statistical physics, condensed matter physics, and high-energy physics, three communities that all have important perspectives to bring to the table but whose notation must be harmonized in order to be accessible to new researchers. I will focus my discussion on (a) the various approaches to the infinite-width limit, which seems like the best entry point for theoretical physicists who first encounter neural networks, and (b) the need for benchmark datasets from physics complex enough to capture aspects of natural-language data but which are nonetheless “calculable” from first-principles using tools of theoretical physics.

Towards a "Theoretical Minimum" for Physicists in AI

Yoni Kahn
University of Toronto
Perimeter Institute Theory + AI workshop, 4/9/25

What this moment feels like

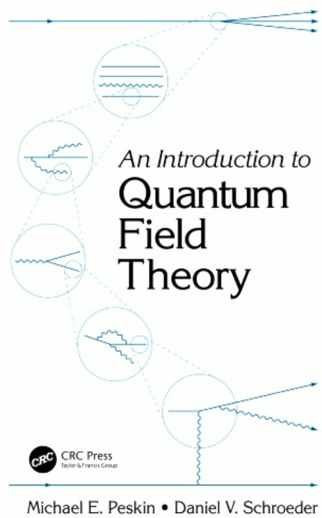
Tons of “experimental” data,
no obvious organizing principle



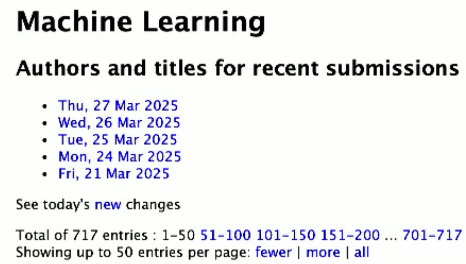
“I have heard it said that ‘the finder of a new elementary particle used to be rewarded by a Nobel Prize, but such a discovery now ought to be punished by a $10,000 fine.’ ”

- W. Lamb, 1955 Nobel lecture

No standard textbook
or curriculum…



…but O(1000) papers/week



Machine Learning

Authors and titles for recent submissions

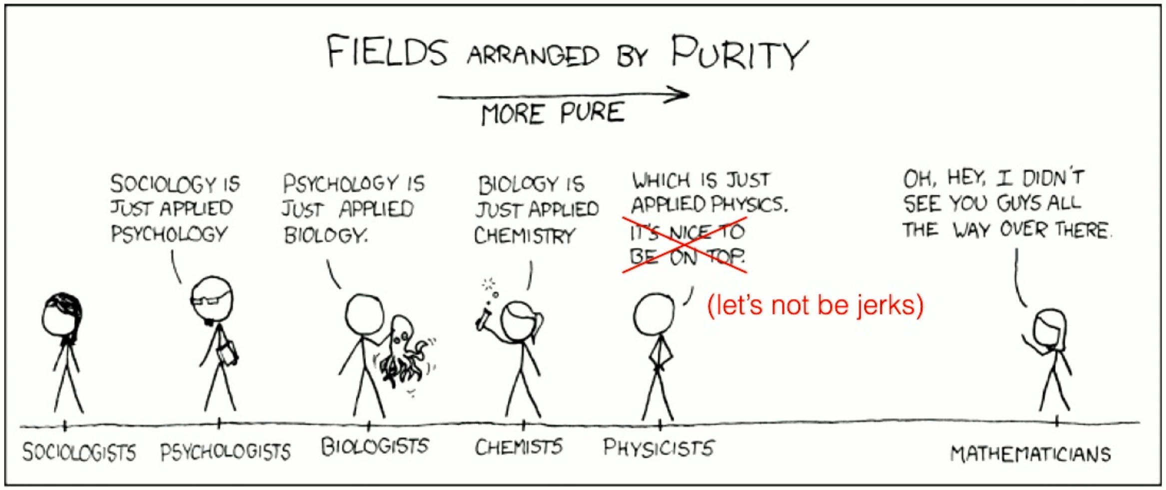
- Thu, 27 Mar 2025
- Wed, 26 Mar 2025
- Tue, 25 Mar 2025
- Mon, 24 Mar 2025
- Fri, 21 Mar 2025

See today's new changes

Total of 717 entries : 1-50 51-100 101-150 151-200 ... 701-717
Showing up to 50 entries per page: fewer | more | all

Like 100 years of modern physics
compressed into a 5-year period
(if all papers were 8 pages or less)

Different ways to study AI



trained AI as "organisms in the wild"

ANTHROP\C Interpretability

Tracing the thoughts of a large language model

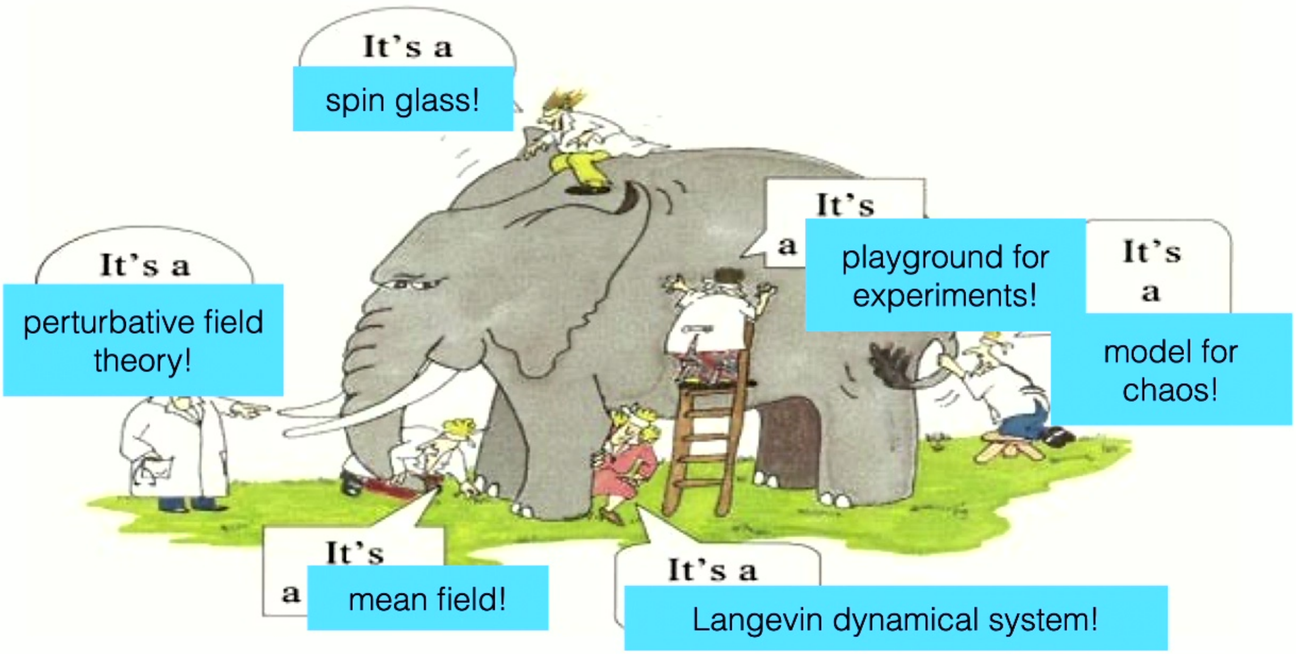
Mar 27, 2025

"We take inspiration from the field of neuroscience…"

physics approach sits in the middle: goal is a "good enough" model using both formal and intuitive reasoning

rigorous worst-case theorems, but maybe not directly applicable

Not just one kind of physics

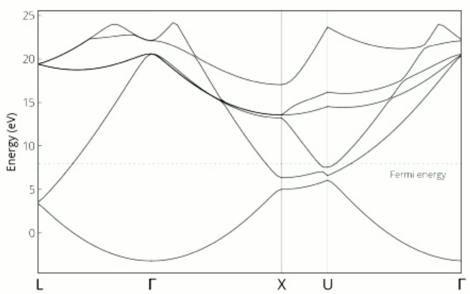


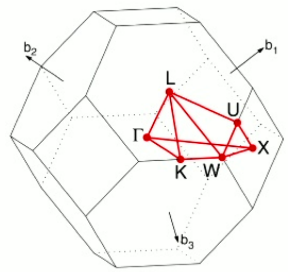
Since we know the "UV theory" of a neural network (we coded it up ourselves!), these microscopic laws should manifest in all parts of the elephant

e.g. symmetries, Noether's theorem, Goldstone's theorem (true from Newton to CMT to QFT)

Effective theories and interpretability

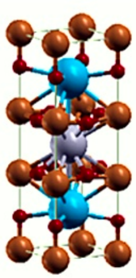
BCS superconductor…



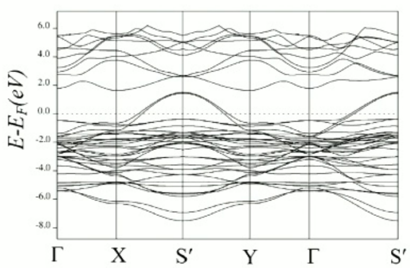


$$F_{\mathrm{GL}} = \int dV \left\{ \alpha|\psi|^2 + \frac{\beta}{2}|\psi|^4 \right\}$$

…vs. high-T$_c$ superconductor

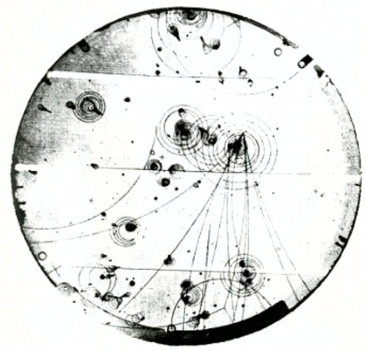


a) $YBa_2Cu_3O_6$



$$F = ???$$

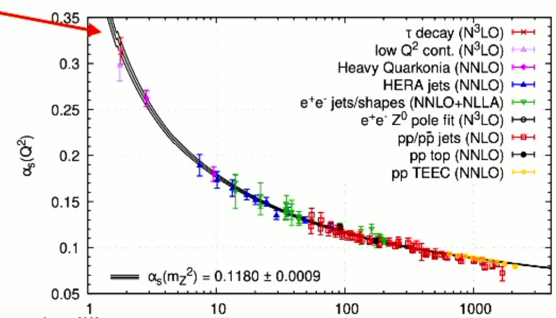
Chiral perturbation theory…



$$\mathcal{L}_{\chi\mathrm{PT}} = \frac{F_\pi^2}{4}\mathrm{Tr}\left[(D_\mu U)(D_\mu U)^\dagger\right]$$

…vs. confinement transition

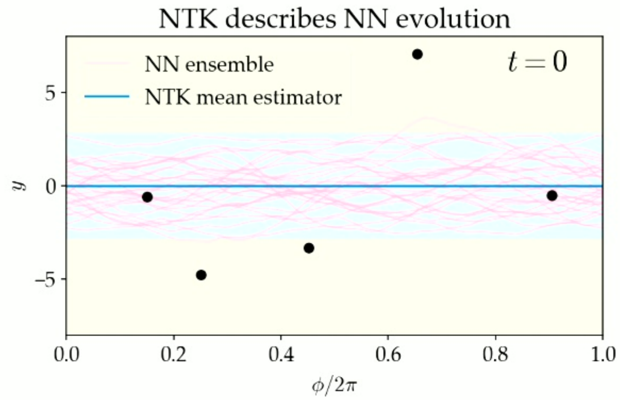
???



What we want from a Theoretical Minimum

- things the physics+AI community agrees are generally true
- at a "physics level of rigor" (**no proofs**!)
- can be checked with (fairly) simple numerical experiments
- has defined limits of validity
- relevant to the "real world" (**some** relevance to state-of-the-art models)
- can be covered in a 1-semester grad course (like the one I'm teaching next spring!)

Feynman: "What I cannot create,
I do not understand."



to a physicist,
I think examples
like this count as
"understanding"

[Animation credit: Wikipedia "Neural Tangent Kernel"]

We all know Gaussian integrals

$$\int d^n z\, e^{-\frac{1}{2}\mathbf{z}^{\mathrm{T}}\mathbf{A}\,\mathbf{z}+\mathbf{J}^{\mathrm{T}}\cdot\,\mathbf{z}} = \sqrt{\frac{(2\pi)^n}{\det \mathbf{A}}}\, e^{\frac{1}{2}\mathbf{J}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{J}}$$

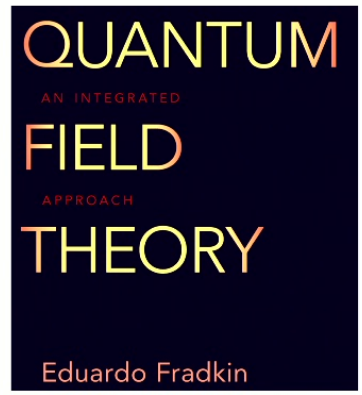
We also have a (fairly standard) notation inherited from QFT:

$$\contraction{}{\phi}{(x_1)}{\phi}\phi(x_1)\phi(x_2) = \frac{\int \mathcal{D}\phi\, e^{iS_0}\phi(x_1)\phi(x_2)}{\int \mathcal{D}\phi\, e^{iS_0}} = D_F(x_1 - x_2). \qquad (9.28)$$

Then the four-point function is simply

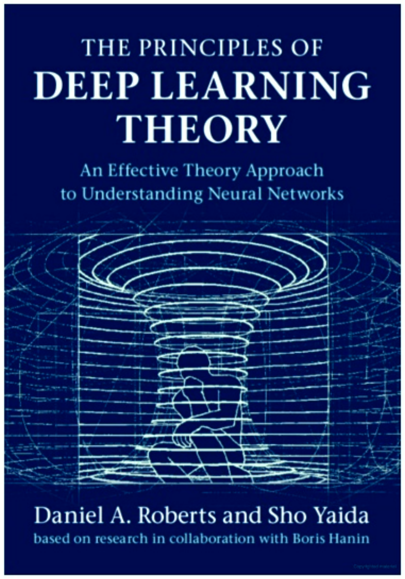
$$\begin{aligned}\langle 0|\, T\phi_1\phi_2\phi_3\phi_4\, |0\rangle &= \text{sum of all full contractions}\\ &= D_F(x_1-x_2)D_F(x_3-x_4)\\ &\quad + D_F(x_1-x_3)D_F(x_2-x_4) \qquad (9.29)\\ &\quad + D_F(x_1-x_4)D_F(x_2-x_3),\end{aligned}$$

and excellent books that treat both HEP and CM



This seems like a good starting point! (Though QFT is notoriously hard to teach…)

Some excellent places to start



arXiv:2106.10165
(471 pages)

Applications of Statistical Field Theory in Deep Learning

A gentle introduction

Suggested Citation: Zohar Ringel, Noa Rubin, Edo Mor, Moritz Helias and Inbar Seroussi (2025), "Applications of Statistical Field Theory in Deep Learning", : Vol. xx, No. xx, pp 1–18. DOI: 10.1561/XXXXXXXXX.

arXiv:2502.18553
(126 pages)

Lecture Notes on Infinite-Width Limits of Neural Networks

Cengiz Pehlevan and Blake Bordelon

June 2023

(see talk by Cengiz tomorrow for more details)

Tensor Programs IVb: Adaptive Optimization in the ∞-Width Limit

Greg Yang
xAI

Etai Littwin
Apple

arXiv:2502.18553
(87 pages)

\+

TASI Lectures on Physics for Machine Learning

Jim Halverson

Department of Physics, Northeastern University, Boston, MA 02115, USA

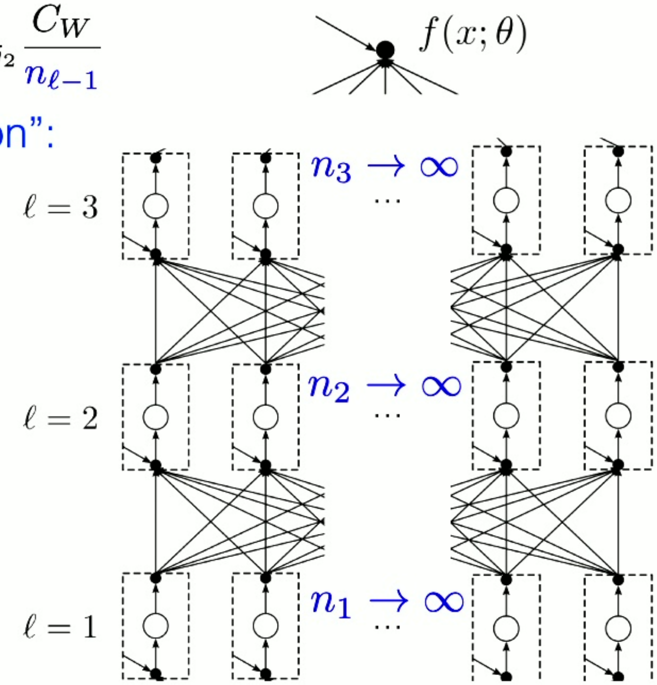
The NSF Institute for Artificial Intelligence and Fundamental Interactions

arXiv:2408.00082

Infinite-width NTK-GP correspondence

$$\mathbb{E}[W^{(\ell)}_{i_1 j_1} W^{(\ell)}_{i_2 j_2}] = \delta_{i_1 i_2}\delta_{j_1 j_2}\frac{C_W}{n_{\ell-1}}$$

"NTK parameterization":
same distribution
on output layer
as hidden layers



Output distribution is a zero-mean Gaussian process at initialization…

$$p(f|\mathcal{D}) = \exp\left(-\frac{1}{2} f(x_\alpha) K^{-1}_{\alpha\beta} f(x_\beta)\right)$$

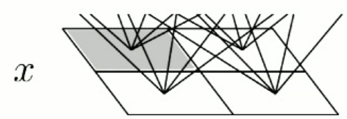
kernel

…and is still Gaussian (but with a nonzero mean and different variance) after gradient descent training

$$m^\infty_\beta \equiv \mathbb{E}[z^{(L)}_\beta] = \boldsymbol{\Theta}^{\mathrm{T}}_\beta \boldsymbol{\Theta}^{-1}_{\mathcal{D}} \mathbf{y}_{\mathcal{D}}$$

neural tangent kernel (NTK)

$$\mathcal{D} = \{x_\alpha\}$$



$$(\sigma^\infty_\beta)^2 \equiv \mathrm{Var}[z^{(L)}_\beta] = K^{(L)}_{\beta\beta} - 2\boldsymbol{\Theta}^{\mathrm{T}}_\beta \boldsymbol{\Theta}^{-1}_{\mathcal{D}} \boldsymbol{K}_\beta + \boldsymbol{\Theta}^{\mathrm{T}}_\beta \boldsymbol{\Theta}^{-1}_{\mathcal{D}} \boldsymbol{K}_{\mathcal{D}} \boldsymbol{\Theta}^{-1}_{\mathcal{D}} \boldsymbol{\Theta}_\beta$$

[R. Neal, 1996; A. Jacot, F. Gabriel, C. Hongler, NeurIPS 2018; J. Lee et al., NeurIPS 2019]

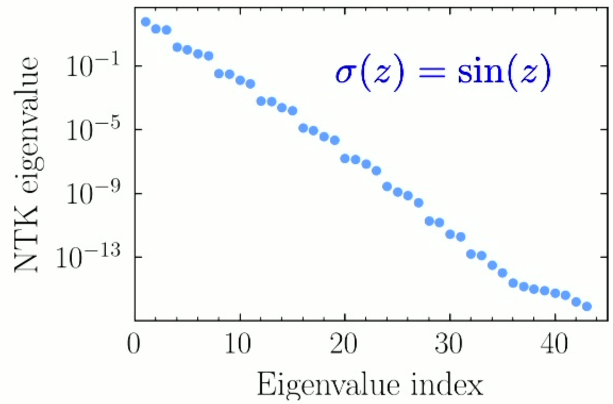
But what if the NTK isn't (numerically) invertible?

Our goal is to characterize the resulting statistics of the network activities for some set of P inputs $\mathbf{x}_\mu$. We allow these inputs, and the input dimension, to be arbitrary up to the condition that the $P \times P$ input Gram matrix

$$G_{\mu\nu} = \mathbf{x}_\mu \cdot \mathbf{x}_\nu \tag{6}$$

is invertible. In particular, we may also take the input dimension to infinity with the layer widths.

This rarely holds in practice: at minimum, can't scale P past input dimension



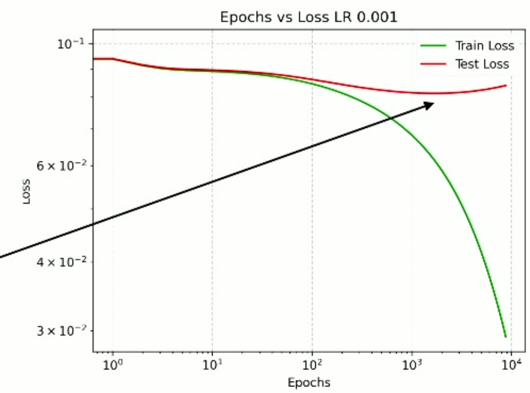
(even power law spectrum can be numerically ill-conditioned)

However, not an obstacle to training!

$$\Delta \mathbf{z}^{(L)}(x_\beta) = -\eta \sum_{x_\alpha \in \mathcal{D}} (\mathbf{z}_\alpha - \mathbf{y}_\alpha)\Theta_{\alpha\beta} + \mathcal{O}(\eta^2)$$

no inverse!

(Though, we should be careful what we mean by "fully trained".)



CIFAR (but not MNIST) with an MLP

[J. Zavatone-Veth, https://jzv.io/assets/pdf/lecture_notes_on_nngp_from_mft.pdf; **YK**, in prep]

Recovering feature learning

Finite-width NTK param.

$$\Delta \mathrm{dd}H = 0 + \mathcal{O}(1/n^2)$$ frozen during training $\left(\frac{d^3 z}{d\theta^3}\right)\left(\frac{dz}{d\theta}\right)^3$

$$\Delta \mathrm{d}H = \eta\, \mathrm{dd}H \epsilon + \mathcal{O}(1/n^2)$$ $\left(\frac{d^2 z}{d\theta^2}\right)\left(\frac{dz}{d\theta}\right)^2$

$$\Delta H = \eta\, \mathrm{d}H \epsilon + \eta^2\, \mathrm{dd}H \epsilon^2 + \mathcal{O}(1/n^2)$$ feature learning

$$\Delta z = -\eta\, H \epsilon + \eta^2\, \mathrm{d}H \epsilon^2 + \eta^3\, \mathrm{dd}H \epsilon^3 + \mathcal{O}(1/n^2)$$

predict **trained ensemble statistics** in terms of **initialization statistics**

$\epsilon = z - y$: prediction error

$$\mathrm{dd}H,\ \mathrm{d}H \sim \mathcal{O}(1/n)$$

Infinite-width μ-parameterization*

$$\mathbb{E}[W^{(L)}_{i_1 j_1} W^{(L)}_{i_2 j_2}] = \delta_{i_1 i_2}\delta_{j_1 j_2}\frac{C_W}{n^2}$$

$$\eta = n\eta_0$$

$$\Longrightarrow\ \mathrm{d}^N H = \mathcal{O}(n^0),\ N = 1, 2, 3, \ldots$$

Feature learning at infinite width!
But hierarchy doesn't truncate, so these are not the right variables to compute with

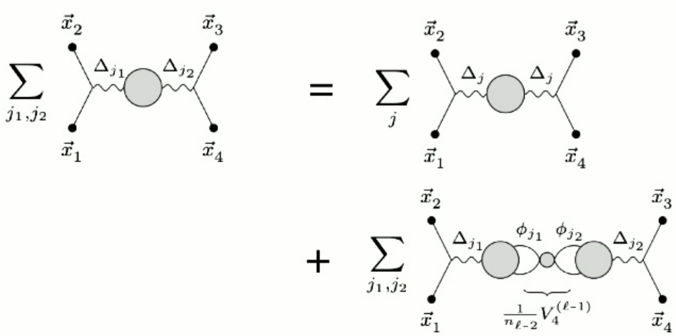
*Most people are not using this for physics applications. Should they be?

[Roberts, Yaida, Hanin, Cambridge University Press 2022; Yang and Hu, ICML 2021; Yaida, arXiv:2210.04909]

Feature Learning 1: finite-width corrections

Notation? (Schwinger vs. Feynman)

$$
\begin{aligned}
V^{(\ell+1)}_{(\alpha_1\alpha_2)(\alpha_3\alpha_4)} = &\left(C_W^{(\ell+1)}\right)^2 \left[\langle \sigma_{\alpha_1}\sigma_{\alpha_2}\sigma_{\alpha_3}\sigma_{\alpha_4}\rangle_{K^{(\ell)}} - \langle \sigma_{\alpha_1}\sigma_{\alpha_2}\rangle_{K^{(\ell)}} \langle \sigma_{\alpha_3}\sigma_{\alpha_4}\rangle_{K^{(\ell)}} \right] \qquad (4.119)\\
&+ \frac{1}{4}\left(C_W^{(\ell+1)}\right)^2 \frac{n_\ell}{n_{\ell-1}} \sum_{\beta_1,\ldots,\beta_4\in\mathcal{D}} V_{(\ell)}^{(\beta_1\beta_2)(\beta_3\beta_4)} \left\langle \sigma_{\alpha_1}\sigma_{\alpha_2}\left(z_{\beta_1}z_{\beta_2} - K^{(\ell)}_{\beta_1\beta_2}\right)\right\rangle_{K^{(\ell)}}\\
&\times \left\langle \sigma_{\alpha_3}\sigma_{\alpha_4}\left(z_{\beta_3}z_{\beta_4} - K^{(\ell)}_{\beta_3\beta_4}\right)\right\rangle_{K^{(\ell)}} + O\left(\frac{1}{n}\right),
\end{aligned}
$$



$\sigma(z) = \sin(z)$ is likely the only tractable activation:

$$\langle\sigma(z_\alpha)\sigma(z_\beta)\sigma(z_\gamma)\sigma(z_\delta)\rangle_K \qquad \langle\sigma(z_\alpha)\sigma(z_\beta)z_\gamma z_\delta z_\kappa z_\lambda\rangle_K$$

$$\langle\sigma'(z_\alpha)\sigma'(z_\beta)\sigma(z_\gamma)\sigma(z_\delta)\rangle_K \qquad \langle\sigma'(z_\alpha)\sigma'(z_\beta)\sigma'(z_\gamma)\sigma'(z_\delta)\rangle_K$$

$$\langle\sigma''(z_\alpha)\sigma'(z_\beta)\sigma'(z_\gamma)\sigma(z_\delta)\rangle_K \qquad \langle\sigma''(z_\alpha)\sigma'(z_\beta)z_\gamma\rangle_K$$

$$\langle\sigma'''(z_\alpha)\sigma'(z_\beta)\sigma'(z_\gamma)\sigma'(z_\delta)\rangle_K \qquad \langle\sigma''(z_\alpha)\sigma''(z_\beta)\sigma'(z_\gamma)\sigma'(z_\delta)\rangle_K$$

Probably impossible to track 1/n effects if numerical integrals required

[Roberts, Yaida, Hanin, Cambridge University Press 2022; Banta, Cai, Craig, Zhang, PRD 2024; Elsharkawy, **YK**, in prep]

Feature Learning 2: dynamical mean-field theory

Define a good set of
time-dependent order parameters:

$$Z = \int \prod_{\ell,\mu\alpha,ts} \frac{d\Phi^\ell_{\mu\alpha}(t,s)d\hat{\Phi}^\ell_{\mu\alpha}(t,s)}{2\pi i N^{-1}} \frac{dG_{\mu\alpha}(t,s)d\hat{G}_{\mu\alpha}(t,s)}{2\pi i N^{-1}} \frac{dA^\ell_{\mu\alpha}(t,s)dB^\ell_{\mu\alpha}(t,s)}{2\pi i N^{-1}} \exp\left(N S[\{\Phi,\hat{\Phi},G,\hat{G},A,B\}]\right)$$

$$S[\{\Phi,\hat{\Phi},G,\hat{G},A,B\}] = \sum_{\ell\mu\alpha}\int_0^\infty\int_0^\infty dtds\left[\Phi^\ell_{\mu\alpha}(t,s)\hat{\Phi}^\ell_{\mu\alpha}(t,s) + G^\ell_{\mu\alpha}(t,s)\hat{G}^\ell_{\mu\alpha}(t,s) - A^\ell_{\mu\alpha}(t,s)B^\ell_{\mu\alpha}(t,s)\right]$$
$$+ \ln \mathcal{Z}[\{\Phi,\hat{\Phi},G,\hat{G},A,B,j,v\}] \qquad (41)$$

$$\mathcal{Z} = \int \prod_{\ell\mu t} \frac{d\hat{\chi}^\ell_\mu(t)d\chi^\ell_\mu(t)}{2\pi}\frac{d\hat{\xi}^\ell_\mu(t)d\xi^\ell_\mu(t)}{2\pi}\exp\left(\sum_{\ell\mu}\int_0^\infty dt\left[\left(j^\ell_\mu(t)+i\hat{\chi}^\ell_\mu(t)\right)\chi^\ell_\mu(t) + \left(v^\ell_\mu(t)+i\hat{\xi}^\ell_\mu(t)\right)\xi^\ell_\mu(t)\right]\right)$$
$$\times \exp\left(-\frac{1}{2}\sum_{\mu\alpha}\int_0^\infty dt\int_0^\infty ds\hat{\chi}^1_\mu(t)\hat{\chi}^1_\alpha(s)K^x_{\mu\alpha} - \frac{1}{2}\sum_{\mu\alpha}\int_0^\infty dt\int_0^\infty ds\hat{\xi}^L_\mu(t)\hat{\xi}^L_\alpha(s)\right)$$
$$\times \exp\left(-\frac{1}{2}\sum_{\ell=1}^{L-1}\sum_{\mu\alpha}\int_0^\infty dt\int_0^\infty ds\left[\hat{\chi}^{\ell+1}_\mu(t)\hat{\chi}^{\ell+1}_\alpha(s)\Phi^\ell_{\mu\alpha}(t,s) + \hat{\xi}^\ell_\mu(t)\hat{\xi}^\ell_\alpha(s)G^{\ell+1}_{\mu\alpha}(t,s)\right]\right)$$
$$\times \exp\left(-\sum_{\ell=1}^{L}\sum_{\mu\alpha}\int_0^\infty dt\int_0^\infty ds\left[\phi(h^\ell_\mu(t))\phi(h^\ell_\alpha(s))\hat{\Phi}^\ell_{\mu\alpha}(t,s) + g^\ell_\mu(t)g^\ell_\alpha(s)\hat{G}^\ell_{\mu\alpha}(t,s)\right]\right)$$
$$\times \exp\left(-i\sum_{\ell=1}^{L}\sum_{\mu\alpha}\int_0^\infty dt\int_0^\infty ds\left[\phi(h^\ell_\mu(t))\hat{\xi}^\ell_\alpha(s)B^\ell_{\mu\alpha}(t,s) + \hat{\chi}^{\ell+1}_\mu(t)g^{\ell+1}_\alpha(s)A^\ell_{\mu\alpha}(t,s)\right]\right)$$

Action is dominated by a saddle point
in the infinite-width limit:

$$h^\ell_\mu(t) = \chi^\ell_\mu(t) + \gamma_0\int_0^t ds\sum_\alpha \Delta_\alpha(s)\Phi^{\ell-1}_{\mu\alpha}(t,s)z_\alpha(s)\dot{\phi}(h^\ell_\alpha(s))$$
$$= u^\ell_\mu(t) + \gamma_0\int_0^t ds\sum_\alpha\left[A^{\ell-1}_{\mu\alpha}(t,s) + \Delta_\alpha(s)\Phi^{\ell-1}_{\mu\alpha}(t,s)\right]\dot{\phi}(h^\ell_\alpha(s))z^\ell_\alpha(s)$$

pre-activations

$$z^\ell_\mu(t) = \xi^\ell_\mu(t) + \gamma_0\int_0^t ds\sum_\alpha \Delta_\alpha(s)G^{\ell+1}_{\mu\alpha}(t,s)\phi(h^\ell_\alpha(s))$$
$$= r^\ell_\mu(t) + \gamma_0\int_0^t\sum_\alpha\left[B^\ell_{\mu\alpha}(t,s) + \Delta_\alpha(s)G^{\ell+1}_{\mu\alpha}(t,s)\right]\phi(h^\ell_\alpha(s))$$

pre-gradients

$$\Phi^\ell_{\mu\alpha}(t,s) = \langle\phi(h^\ell_\mu(t))\phi(h^\ell_\alpha(s))\rangle\ ,\ G^\ell_{\mu\alpha}(t,s) = \langle g^\ell_\mu(t)g^\ell_\alpha(s)\rangle$$

feature kernels

$$A^\ell_{\mu\alpha}(t,s) = \gamma_0^{-1}\left\langle\frac{\delta\phi(h^\ell_\mu(t))}{\delta r^\ell_\alpha(s)}\right\rangle\ ,\ B^\ell_{\mu\alpha}(t,s) = \gamma_0^{-1}\left\langle\frac{\delta g^{\ell+1}_\mu(t)}{\delta u^{\ell+1}_\alpha(s)}\right\rangle$$

inter-layer
correlations

[much more on this in Blake's and Cengiz's talks]

Scaling families

Output layer is special:

Can interpolate between NTK and μP
with a 1-parameter family of initializations

$$\mathbb{E}[W^{(L)}_{i_1 j_1} W^{(L)}_{i_2 j_2}] = \delta_{i_1 i_2}\delta_{j_1 j_2}\frac{C_W}{n^{1+s}}$$

$$\varepsilon = \frac{L}{n^{1-s}}, \;\; s \in [0,1]$$

maybe independent of L for orthogonal weights?
true for s = 0 at least

In DMFT, interpolate between frozen
and dynamical NTK at strict infinite width

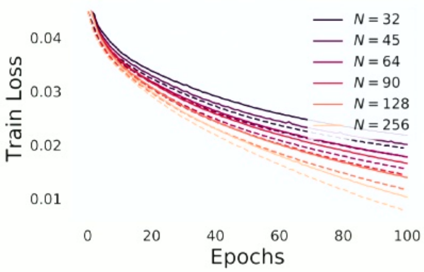
$$\mathbf{z}^{(L)} \to \frac{1}{\gamma}\mathbf{z}^{(L)}, \;\; \eta \to \gamma^2 \eta$$

$$\gamma_0 = \frac{\gamma}{\sqrt{n}} \in [0, \infty]$$

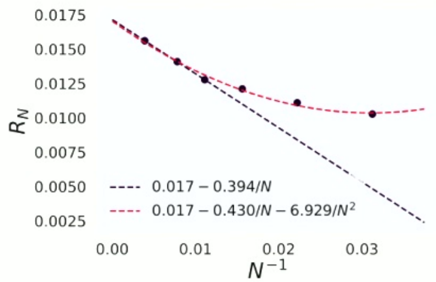
Both interpolate between perturbative and non-perturbative descriptions:
lots of rich HEP and CM analogies here!

[Yaida, arXiv:2210.04909; Day, **YK**, Roberts, arXiv:2310.07765; Bordelon and Pehlevan, NeurIPS 2022; Atanasov et al., ICLR 2025]

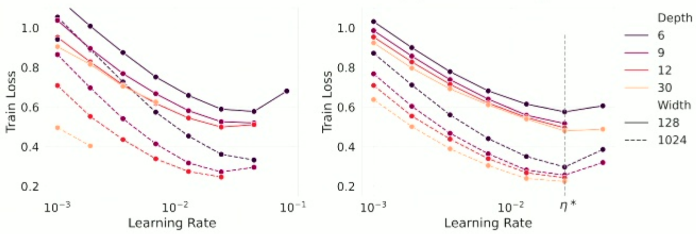
Finite vs. infinite



(a) Train Loss Across Widths



(e) Training Rates vs N



(a) ResNet, μP

(b) This work: $\frac{1}{\sqrt{L}}$-ResNet + μP

Do finite-width corrections for μP matter?

($\mathcal{O}(P^4T^4)$ seems prohibitive…)

What about finite depth?

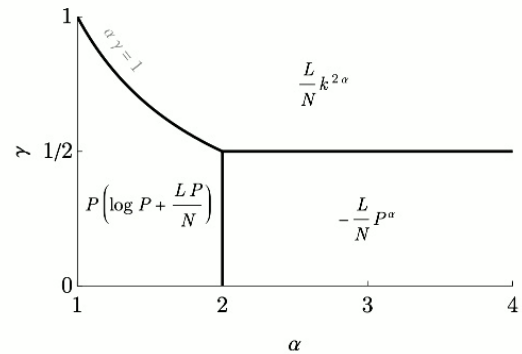


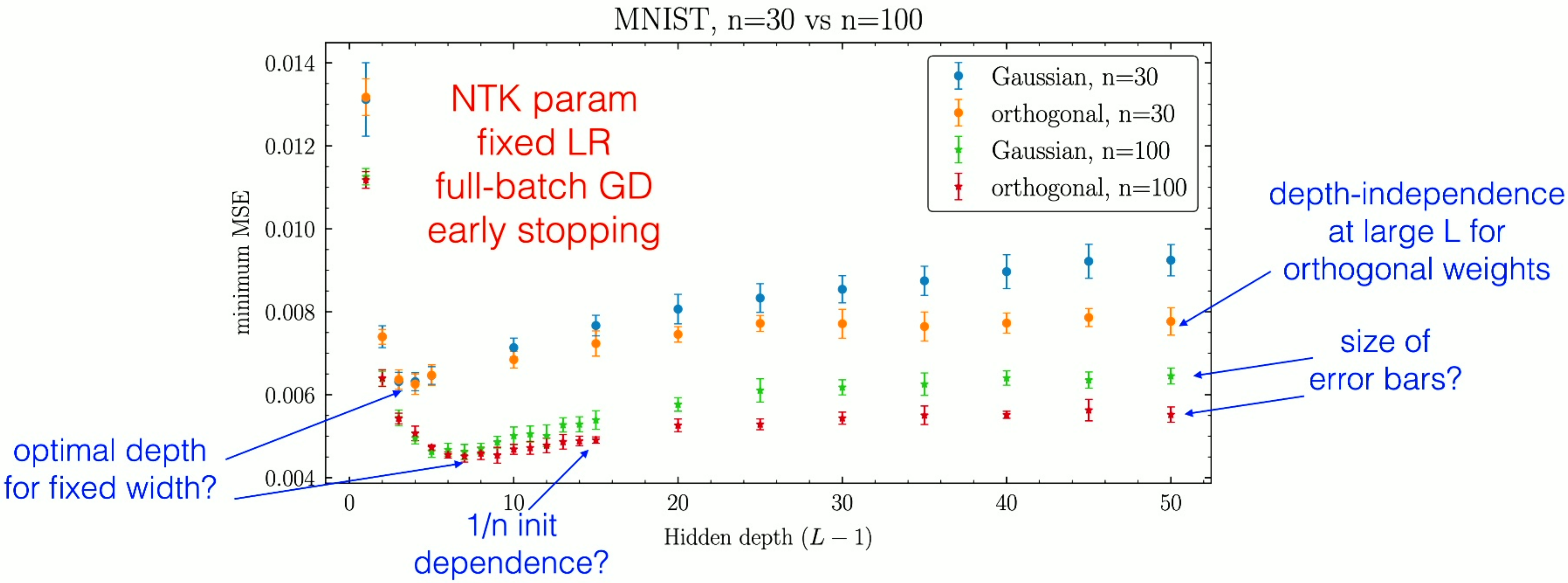
Figure 2: Phase diagram of the log evidence of a deep nonlinear network at zero temperature. The dataset covariance matrix has a power law spectrum $\lambda_j \sim j^{-\alpha}$ (2.16) and the label vector lies in the kth direction (2.17) for $k = P^{\gamma}$. The first-order in $1/N$ is perturbatively valid for $\gamma < 1/\alpha$ and $\alpha < 2$. Within the perturbative regime, depth improves the evidence; at the two boundaries of the regime, depth either increases or decreases the evidence. See Fig. 3 for the phase diagram at nonzero temperature.

Does NTK perturbation theory break down when training set is large enough?

Infinite-depth limit with orthogonal weights?

[Bordelon and Pehlevan, NeurIPS 2023; Bordelon et al, arXiv:2309.16620; Hanin and Zlokapa, arXiv:2405.16630; Ringel et al., arXiv:2502.18553]

Depth dependence

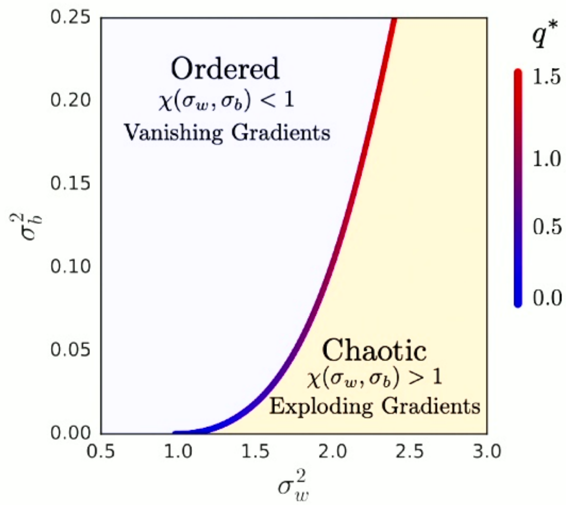


Can we explain these curves quantitatively?

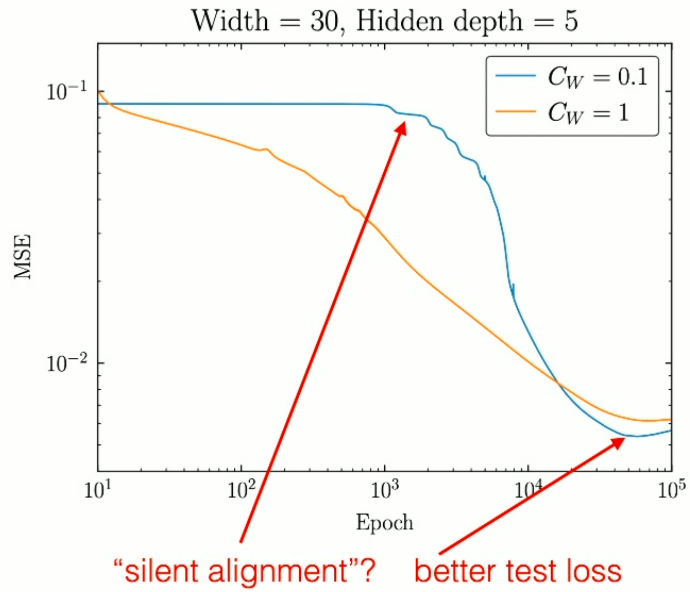
Not state-of-the-art, but a very concrete playground for testing predictions

Does initialization matter?

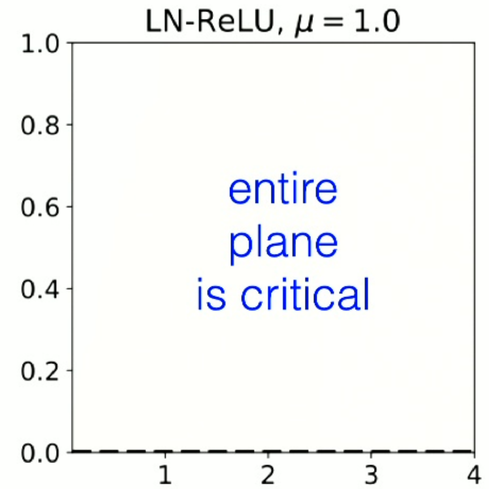
This was a great example of criticality being used in AI



(But maybe sub-critical does better?)



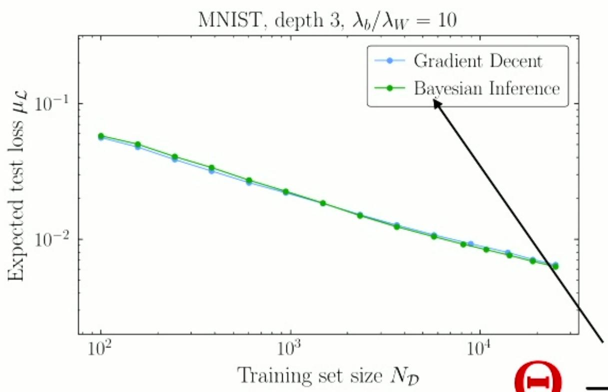
But if we apply LayerNorm to preactivations, we don't need it



[Pennington, Schoenholtz, Ganguli, NeurIPS 2017; Atanasov, Bordelon, Pehlevan, ICLR 2022; Day, **YK**, Roberts, arXiv:2310.07765; Doshi, He, Gromov, NeurIPS 2023]

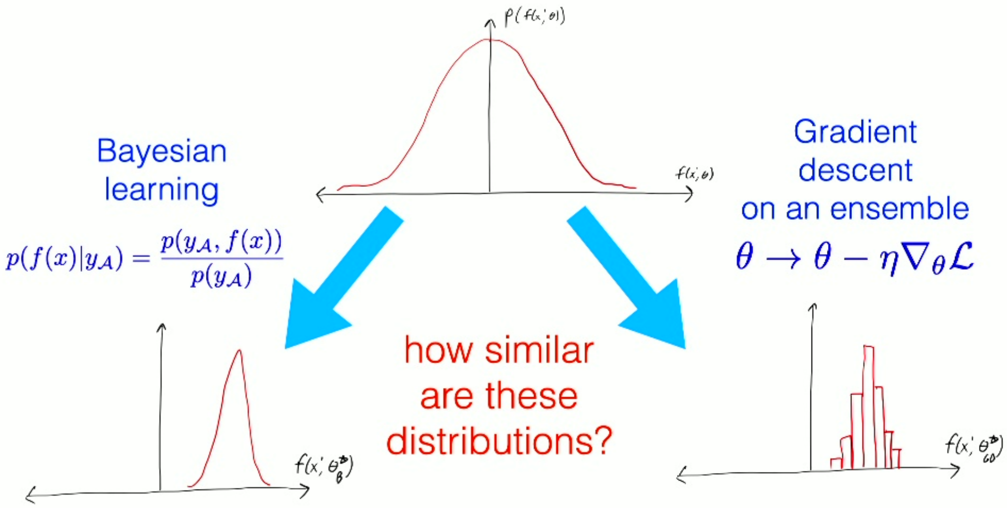
Is there a Bayesian interpretation?

Infinite-width GD
and exact Bayesian inference
are closely related for NTK init:



$\Theta \rightarrow K$

What about finite width?



Can we tell a similar story for μP/DMFT? How does all this relate to "Bayesian neural networks"?

[Elsharkawy, Hooberman, **YK**, arXiv:2503.05938; Day, Elsharkawy, **YK**, S. Roy, in prep]

What to say about transformers?

Index-free

Using tensor products, we can describe the process of applying attention as:

$$h(x) = (\mathrm{Id} \otimes W_O) \cdot (A \otimes \mathrm{Id}) \cdot (\mathrm{Id} \otimes W_V) \cdot x$$

- $(\mathrm{Id} \otimes W_O)$: Project result vectors out for each token $(h(x)_i = W_O r_i)$
- $(A \otimes \mathrm{Id})$: Mix value vectors *across* tokens to compute result vectors $(r_i = \sum_j A_{i,j} v_j)$
- $(\mathrm{Id} \otimes W_V)$: Compute value vector for each token $(v_i = W_V x_i)$

Applying the mixed product property and collapsing identities yields:

$$h(x) = (A \otimes W_O W_V) \cdot x$$

A mixes across tokens while $W_O W_V$ acts on each vector independently.

With indices

$$z_i^{(1),a} = W_{ij}^{\mathrm{emb}} x_j^a + p_i^a$$

$$z_{i,h}^{(2),a} = \Phi\left(\frac{Q_{j,h}^a K_{j,h}^b}{\sqrt{d_k}}\right) V_{i,h}^b \quad \text{(no } h \text{ summation)}$$

$$= \Phi\left(\frac{W_{lm,h}^Q z_m^{(1),a} W_{ln,h}^K z_n^{(1),b}}{\sqrt{d_k}}\right) W_{ij,h}^V z_j^{(1),b}$$

$$z_i^{(3),a} = W_{ij,h}^O z_{j,h}^{(2),a}$$

$$z_i^{(4),a} = \phi\left(W_{ij}^{(4)} z_j^{(3),a} + b_i^{(4)}\right)$$

$$z_i^{(5),a} = W_{ij}^{(5)} z_j^{(4),a} + b_i^{(5)}$$

$$f_i^a(X) = z_i^{(6),a} = W_{ij}^{\mathrm{d-emb}} z_j^{(5),a}$$

Init choices

$$\mathbb{E}\left[W_{i_1j_1}^{\mathrm{patch}} W_{i_2j_2}^{\mathrm{patch}}\right] = \left(\frac{C_{\mathrm{patch}}}{n_{\mathrm{patch}}}\right) \delta_{i_1i_2}\delta_{j_1j_2} \quad \Big\| \quad \mathbb{E}\left[W_{i_1j_1}^{\mathrm{WE}} W_{i_2j_2}^{\mathrm{WE}}\right] = (C_{\mathrm{WE}})\, \delta_{i_1i_2}\delta_{j_1j_2}\,,$$

$$\mathbb{E}\left[b_{t_1;i_1}^{\mathrm{PE}} b_{t_2;i_2}^{\mathrm{PE}}\right] = (C_{\mathrm{PE}})\, \delta_{t_1t_2}\delta_{i_1i_2}\,,$$

$$\mathbb{E}\left[Q_{c_1i_1}^{h_1} Q_{c_2i_2}^{h_2}\right] = \left(\frac{C_Q}{n}\right) \delta_{c_1c_2}\delta_{i_1i_2}\delta^{h_1h_2}\,,$$

$$\mathbb{E}\left[K_{c_1i_1}^{h_1} K_{c_2i_2}^{h_2}\right] = \left(\frac{C_K}{n}\right) \delta_{c_1c_2}\delta_{i_1i_2}\delta^{h_1h_2}\,,$$

$$\mathbb{E}\left[V_{c_1i_1}^{h_1} V_{c_2i_2}^{h_2}\right] = \left(\frac{C_V}{n}\right) \delta_{c_1c_2}\delta_{i_1i_2}\delta^{h_1h_2}\,,$$

$$\mathbb{E}\left[U_{i_1c_1}^{h_1} U_{i_2c_2}^{h_2}\right] = \left(\frac{C_U}{n}\right) \delta_{i_1i_2}\delta_{c_1c_2}\delta^{h_1h_2}\,,$$

$$\mathbb{E}\left[W_{i_1j_1} W_{i_2j_2}\right] = \left(\frac{C_W}{n}\right) \delta_{i_1i_2}\delta_{j_1j_2}\,,$$

$$\mathbb{E}\left[X_{i_1j_1} X_{i_2j_2}\right] = \left(\frac{C_X}{Mn}\right) \delta_{i_1i_2}\delta_{j_1j_2}\,,$$

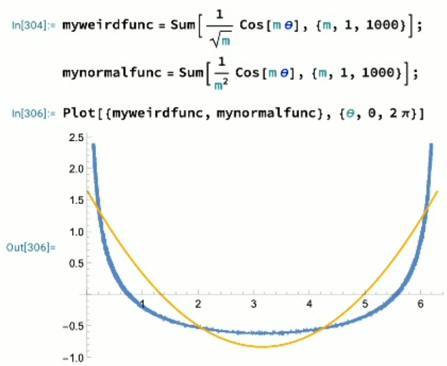
$$\mathbb{E}\left[W_{i_1j_1}^{\mathrm{head}} W_{i_2j_2}^{\mathrm{head}}\right] = \left(\frac{C_{\mathrm{head}}}{n}\right) \delta_{i_1i_2}\delta_{j_1j_2}\,,\ b_i^{\mathrm{head}} = 0 \quad \Big\| \quad \mathcal{N}_{\mathrm{rescale}} = \sqrt{\frac{1}{n}}\,,$$

Is this the simplest way to present this architecture?

[Anthropic blog post, 2021; Lavie, Gur-Ari, Ringel, ICML 2024; Dinan, Yaida, Zhang, arXiv:2304.02034]

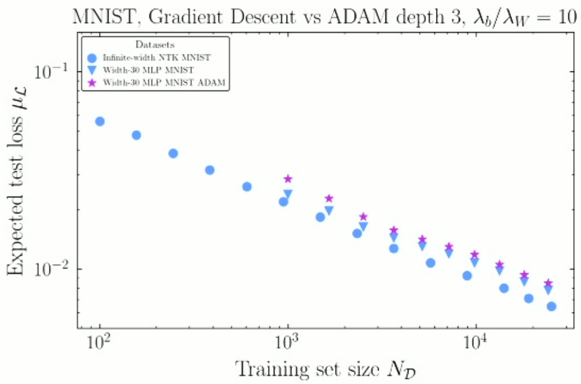
What's a good "hello world" example for feature learning?

Infinite-norm functions in the NTK eigenbasis are pretty weird

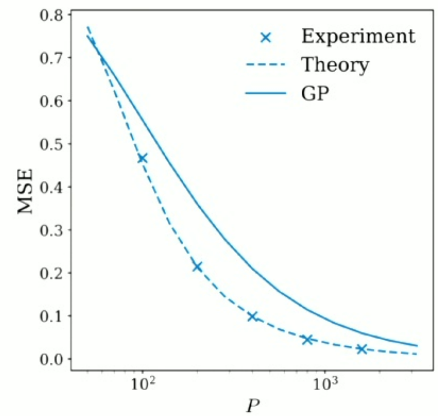


Structure at all scales, divergences, sharp edges

And even MNIST and CIFAR don't change slope



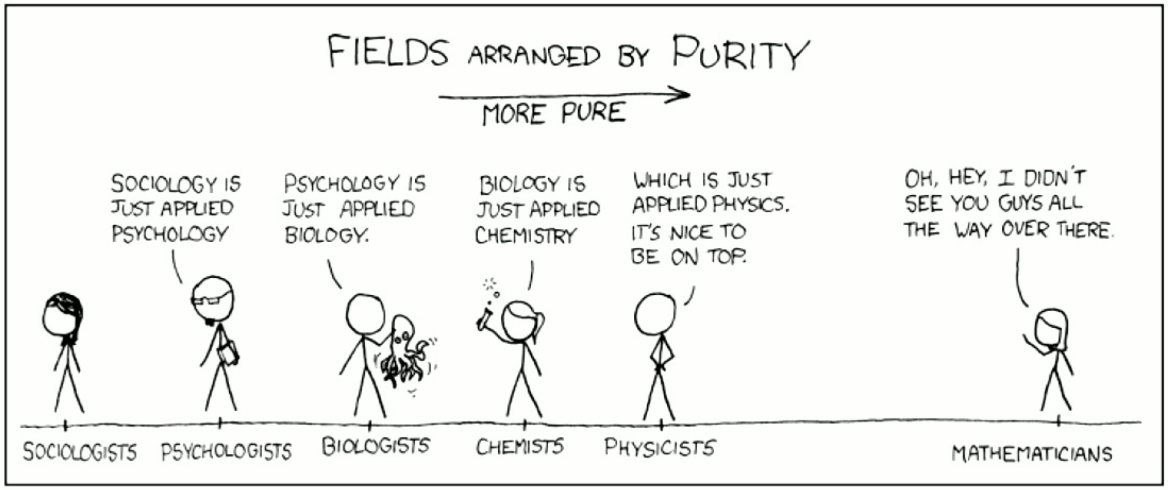
Projection tasks?



ideally something beyond single-layer linear networks

[Bordelon, Atanasov, Pehlevan, arXiv:2409.17858; Elsharkawy, Hooberman, **YK**, arXiv:2503.05938 (NeurIPS 2024 ML4PS workshop); Ringel et al., arXiv:2502.18553]

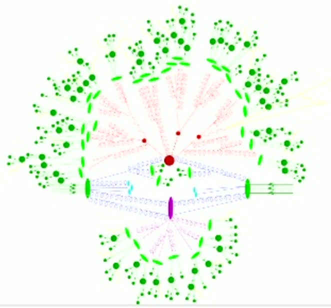
What is a “realistic” dataset?



Natural language

$p_{\text{English}}(w|\text{I went to the})$

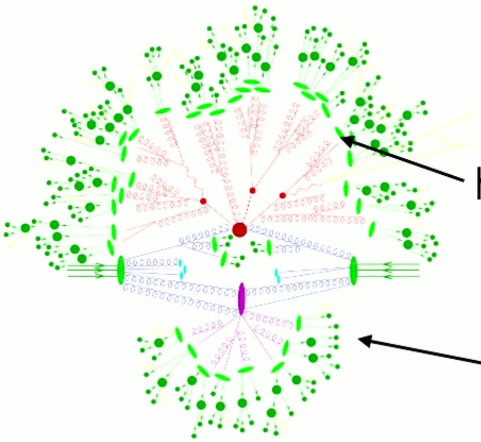
Simulated collider data



Gaussian random feature models

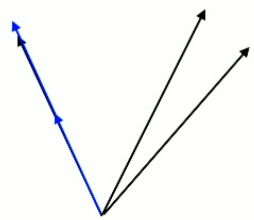
$\langle x_I x_J \rangle = \Lambda_{IJ}$

Simulated jet data



how do we classify this blob

vs. this blob?



a zero-energy ("infrared") particle might as well not be there

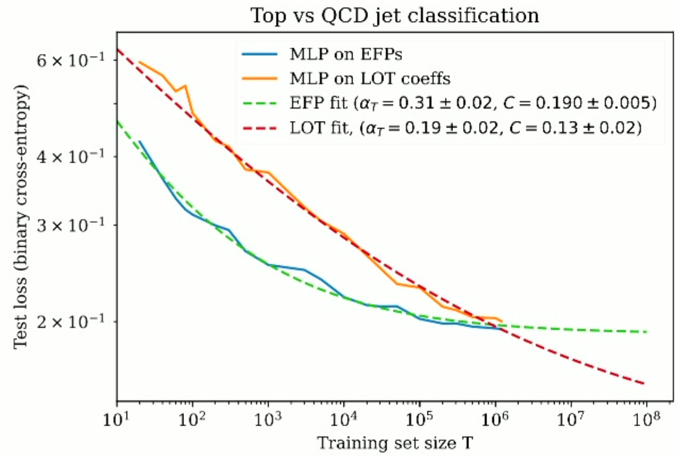
two collinear particles might as well be one

$$\mathcal{M}_{4-\text{particle}} \simeq S^8 \supset S^5$$

"How many particles" is not a well-defined question in perturbative QFT!

$$\mathcal{M}_{\text{jet}} \sim S^2 \subset S^5 \subset S^8 \cdots \subset S^{596} \subset \cdots$$

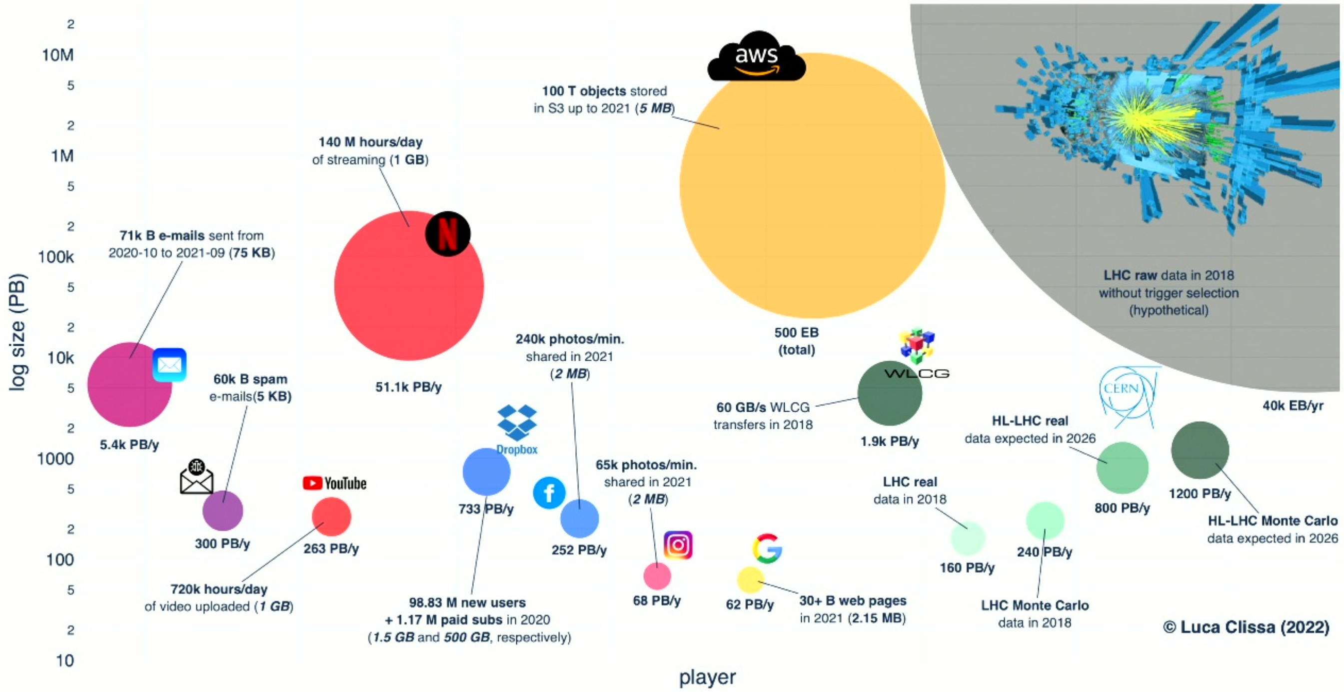
And we see nontrivial scaling laws!



same data, different slope:
depends on pre-processing,
anti-correlated with data/data covariance

Is there another good example from physics?
(e.g. projective measurements of many-qubit systems, c.f. Roger's panel talk on Monday)

HEP = big(gest) data



If scaling laws are all you need, we have plenty of this in physics!

Conclusions

We are witnessing the birth of a new field. There is so much to do!
We need a strategy to get **everyone** (students through faculty) up to speed ASAP

Standardizing things is hard, but there are examples from HEP:

Les Houches Accords

1 language

Article Talk Read Edit View history Tools

From Wikipedia, the free encyclopedia

The **Les Houches Accords** are agreements between particle physicists to standardize the interface between the matrix element programs and the event generators used to calculate different quantities. The original accord was initially formed in 2001, at a conference in Les Houches, in the French Alps, before it was subsequently expanded.

Looking forward to discussions this week!