Field theory approach to DNNs, a potential common language

Zohar Ringel | 19 Mar 2025 | Oxford |

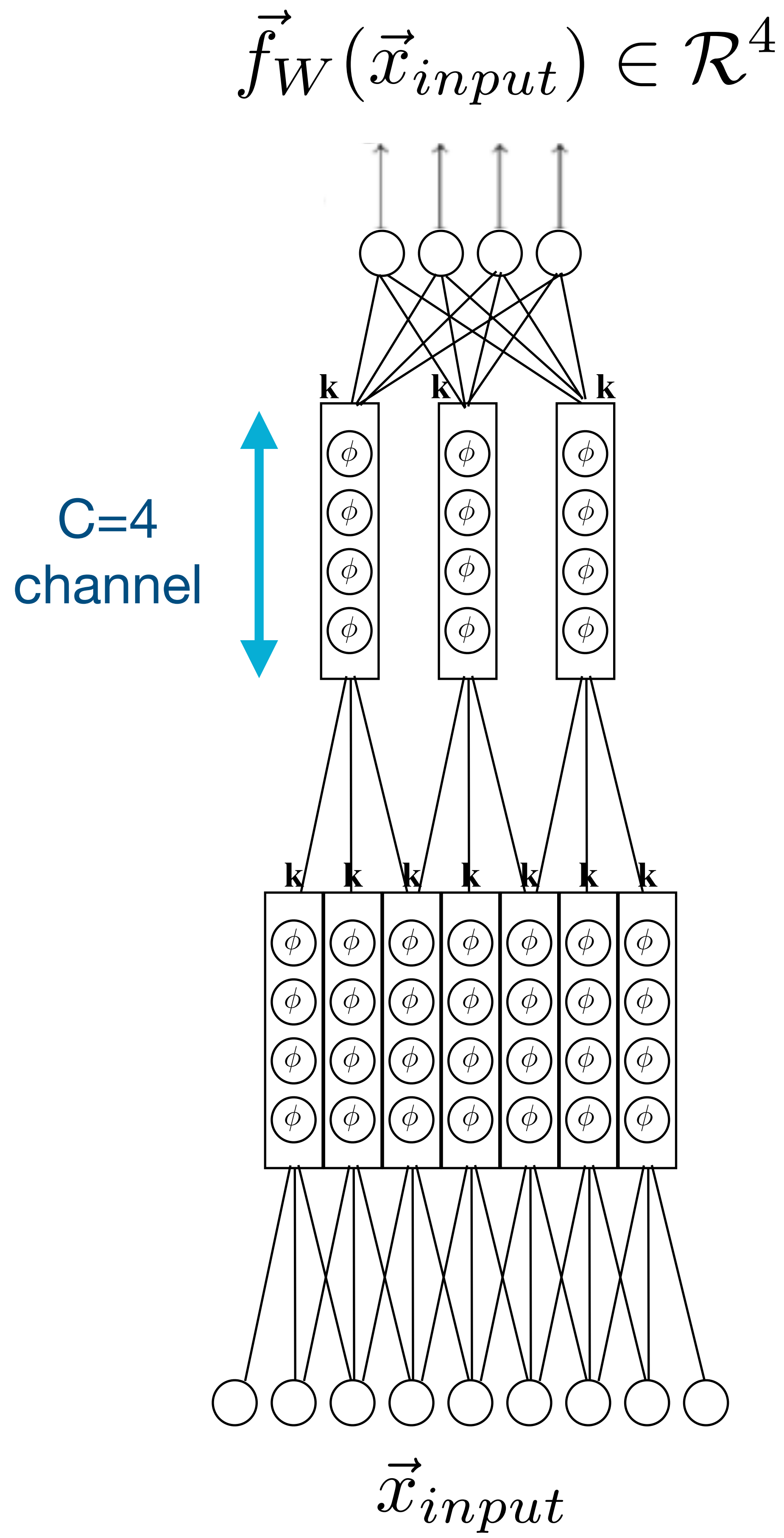
Applications of Statistical Field Theory in Deep Learning

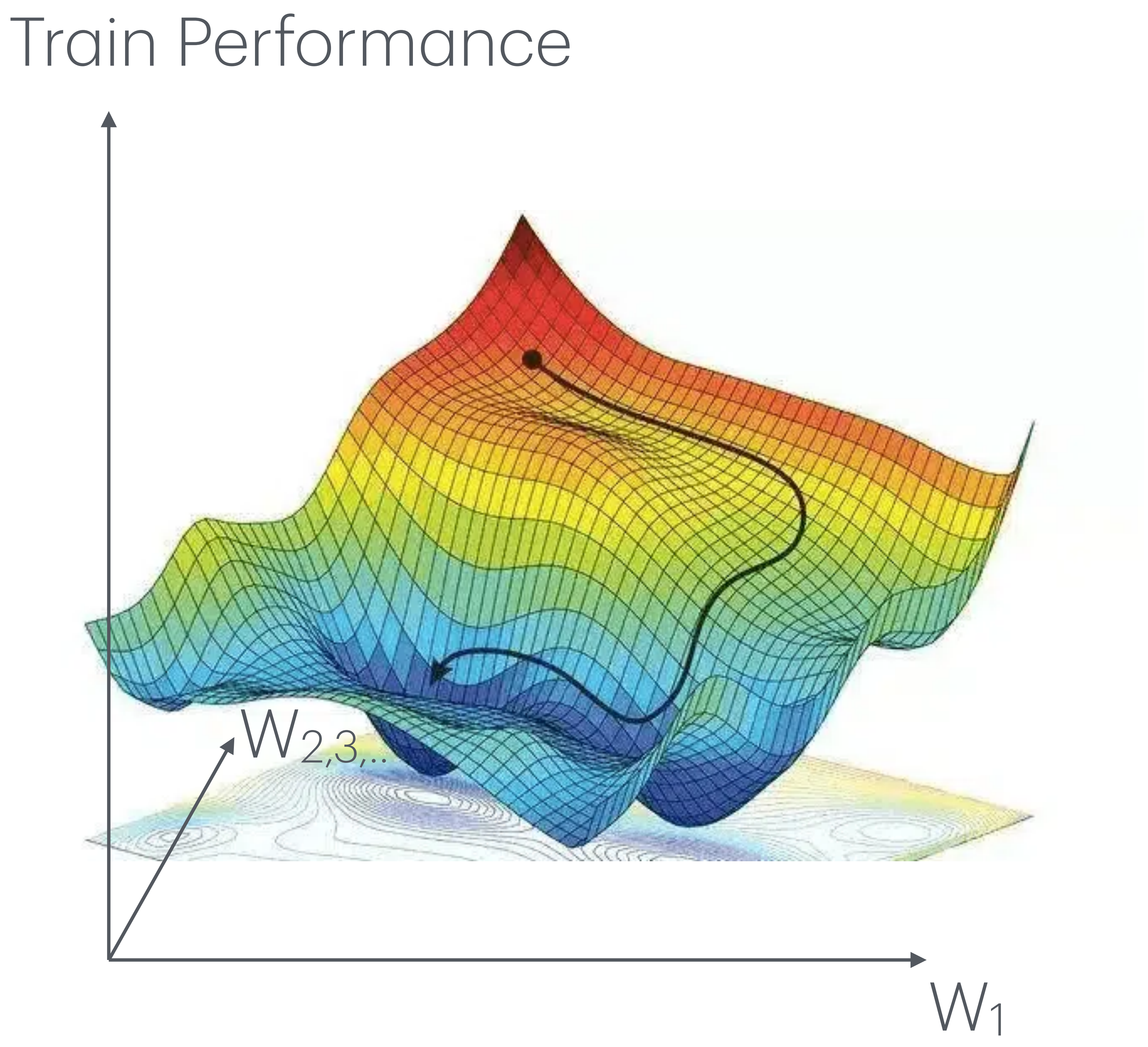
Zohar Ringel, Noa Rubin, Edo Mor, Moritz Helias, Inbar Seroussi

https://arxiv.org/abs/2502.18553

Deep Learning

- **Redundant, brain-inspired ansatz for functions** + **Many Examples** + **local optimization**





= Good generalization/performance

Theoretical Questions

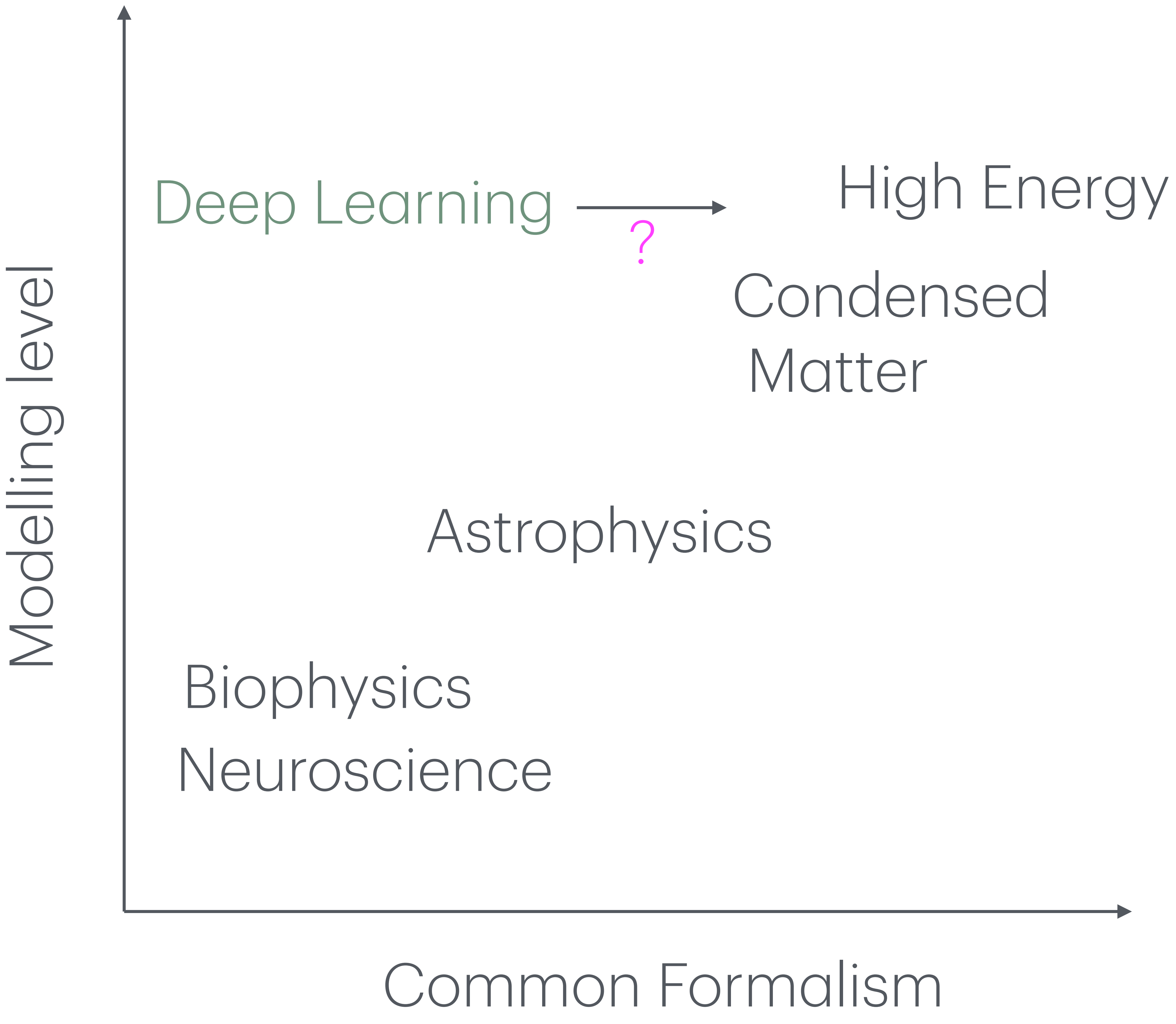
- **Optimization:** Why is optimization relatively easy despite the highly non-convex landscape?
- **Generalization:** Why does it find a good solution while having many more model parameters compared to training data?
- **Alignment/Interpretation:** What features of the data are used for prediction? Is it aligned with the proper way of thinking on problems?
- **Complexity classes and scaling:** How smart would ChatGPT with x100 compute be? Are Humans in the same sample-complexity class as ChatGPT?

Science on the formality scale

In more mature fields

Partial List of Theoretical Techniques
In Deep learning

Random Matrix Theory
Gaussian Process
(Deep) Linear models
Saad & Sola
DMFT
Stat. Mech.
Spin-Glass for Capacity
Stochastic Processes in time
Stat. Learning Theory
One-Step GD
....



Our emphasis

1. **Generic formalism** [not conditioned on one-trainable layer, width, etc..]

2. **Physics style** [approximations, typical case thinking, no special focus on exact asymptotic limits]

3. **Fundamental Science** [Prioritize understanding, long-term]

4. **Bottom-up/Microscopic** [as opposed to biological/data-science approaches]

5. **Expanding the toolkit** [Replicas, Diagrams, RG, Field-Theory, Representation-Theory,..]

Collaborators

Inbar Seroussi

Moritz Helias

Gad Naveh

Itay Lavie

Noa Rubin

Learning curves for deep neural network a gaussian field theory approach **Cohen, Malka, ZR** (2019)

Separation of scales + thermodynamic description of feature learning in some CNNs **Seroussi, Naveh, ZR** (2021)

Grokking as a First Order Phase transition in two layer Neural Networks **Rubin, Seroussi, ZR** (2023)

Wilsonian Renormalization of Neural Network Gaussian Processes **Howard, Maiti, Jefferson, ZR** (2024)

Towards Understanding Inductive Bias In Transformers.... **Lavie, Gur-Ari, Ringel** (2024)

A unified approach to feature learning in Bayesian Neural Networks **Rubin, Seroussi, ZR, Helias** (2024)

Strategy I

Let’s first understand equilibrium/Bayesian

Grounds

1. True out-of-equilibrium physics is extremely challenging…

2. While life is an inherently out-of-equilibrium phenomena, there is little reason to think deep learning is an inherently out-of-equilibrium.

3. DNNs are often over-parametrized, no reason to expect Glassy-behavior or exponentially large equilibrium times. SGD has been argued to be roughly Bayesian*.

4. A lot of the formalism and approximations generalize to dynamics via MSRDJ.

5. Equilibrium relates to Bayesian learning, which is a holy grail in inference.

* e.g. C. Mingard et. al. 2021

Equilibrium and Bayesian

Equilibrium

$z(x) = DNN_\theta(x)$ $(x_1, y_1)\ldots(x_n, y_n)$ $\dot{\theta} = -\gamma\theta - \partial_\theta L + \xi$

$$L[z] = \sum_{\mu=1}^{P} \frac{(z(x_\mu) - y_\mu)^2}{2}$$

Langevin training with weight decay ($\boldsymbol{\gamma}$)

and noise with variance T

$$Z_{t\to\infty} = \int d\theta e^{-(L[z]+\gamma|\theta|^2/2)/T} = \int d\theta e^{-\frac{\gamma|\theta|^2}{2T}} e^{-L[z]/T}$$

Bayesian Interpretation

Gaussian Weight Prior

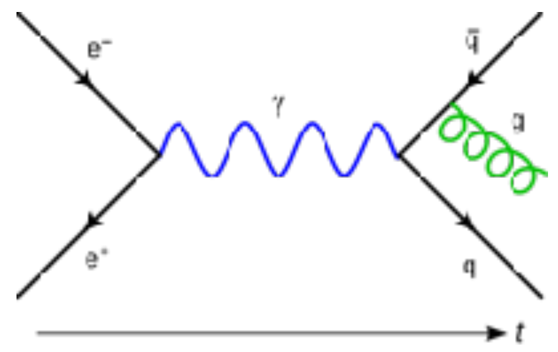
Likelihood under noisy observations
With variance T

Strategy II

Fields rather than weights

The field theory viewpoint

On random/at-init networks



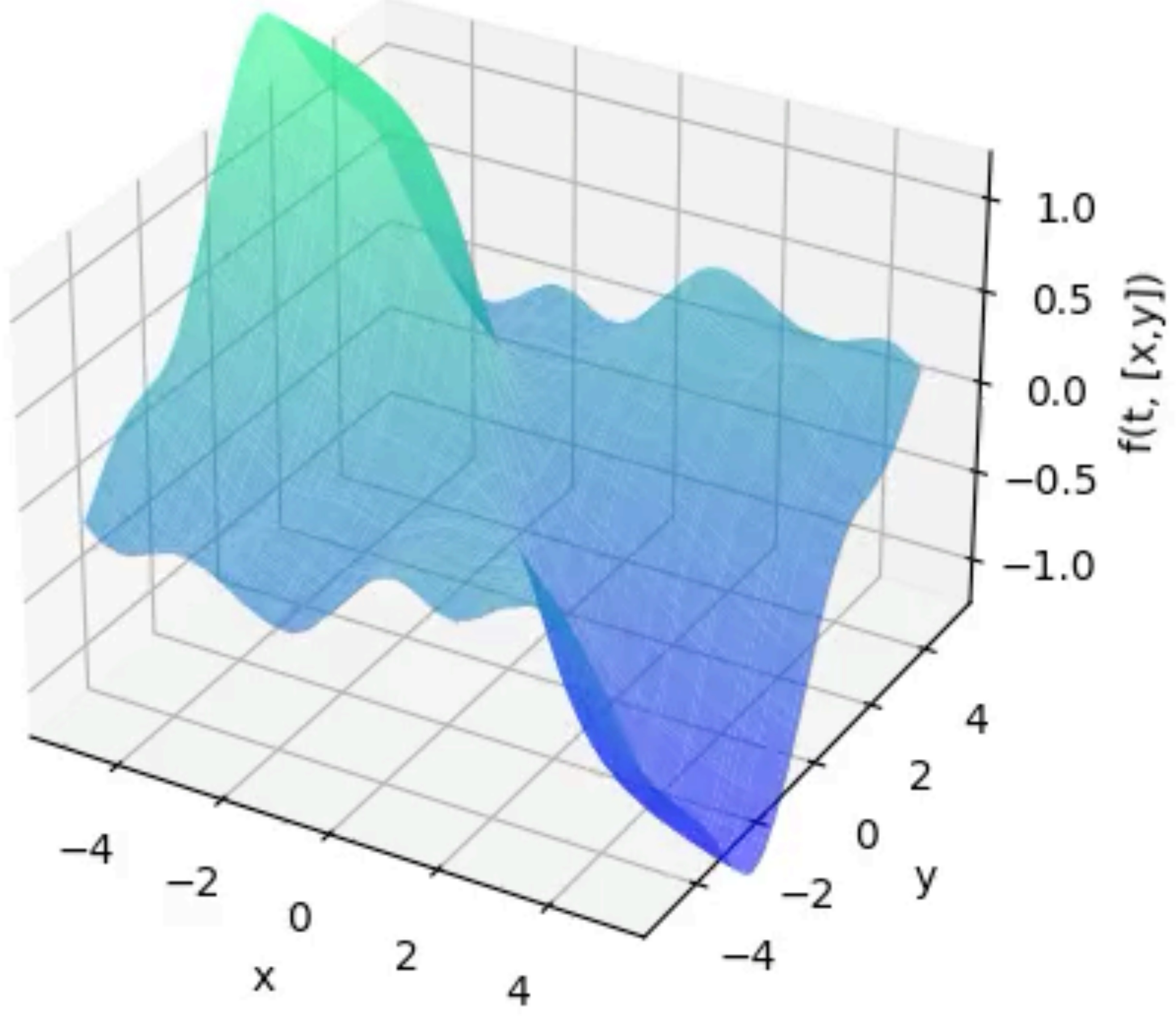
$$z_{w,a}(x) = \sum_{c=1}^{N} a_c Erf(w_c^T x) \quad x \in \mathrm{R}^2$$

$+$

Random Weights (a_c, w_c)

$\longrightarrow$

Random Function



Kernel

$$K(x, x') = \langle z_{w,a}(x) z_{w,a}(x') \rangle_{uniform(w,a)}$$

$\longrightarrow$

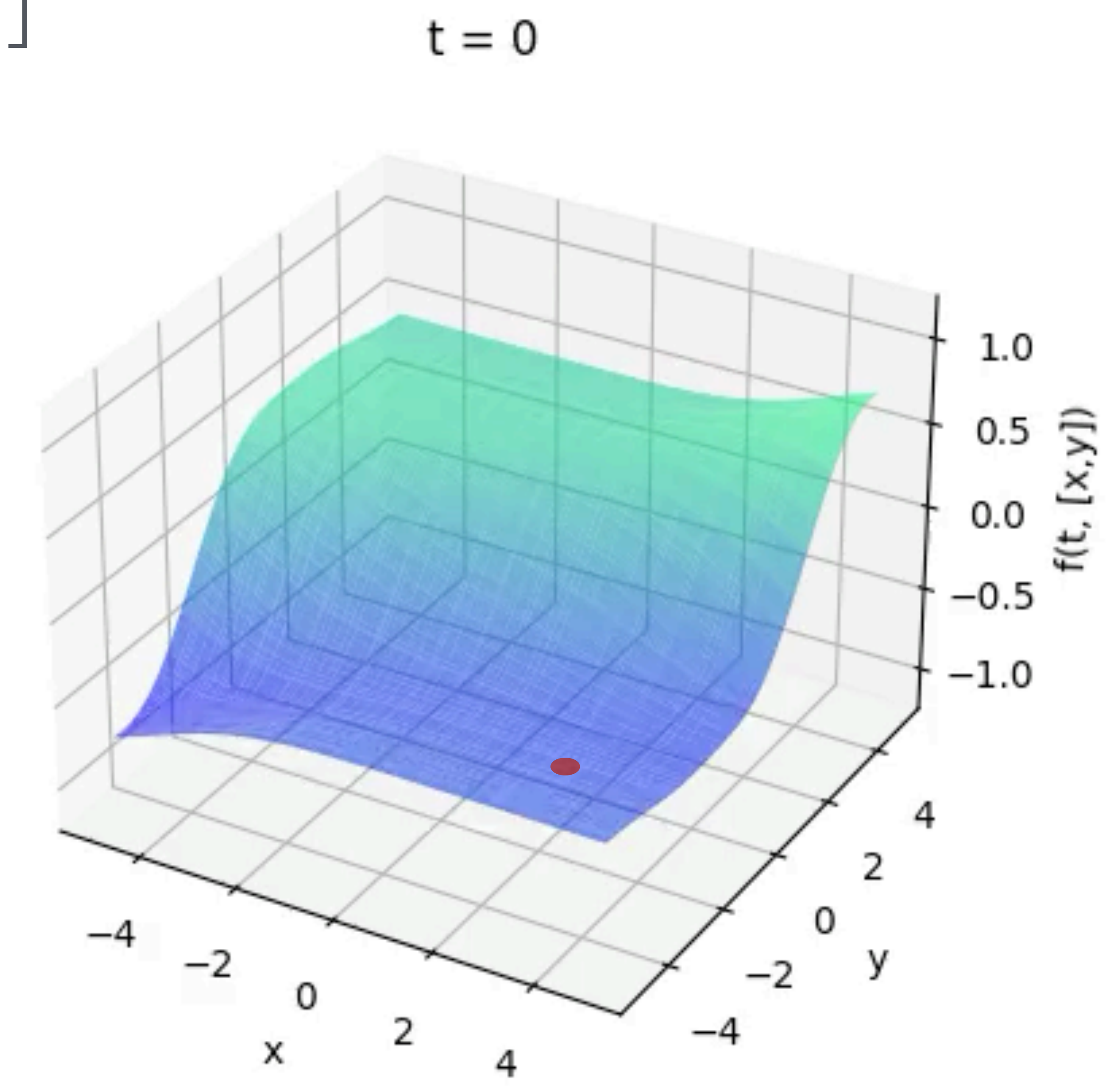
Green's function

$$G(x, x')$$

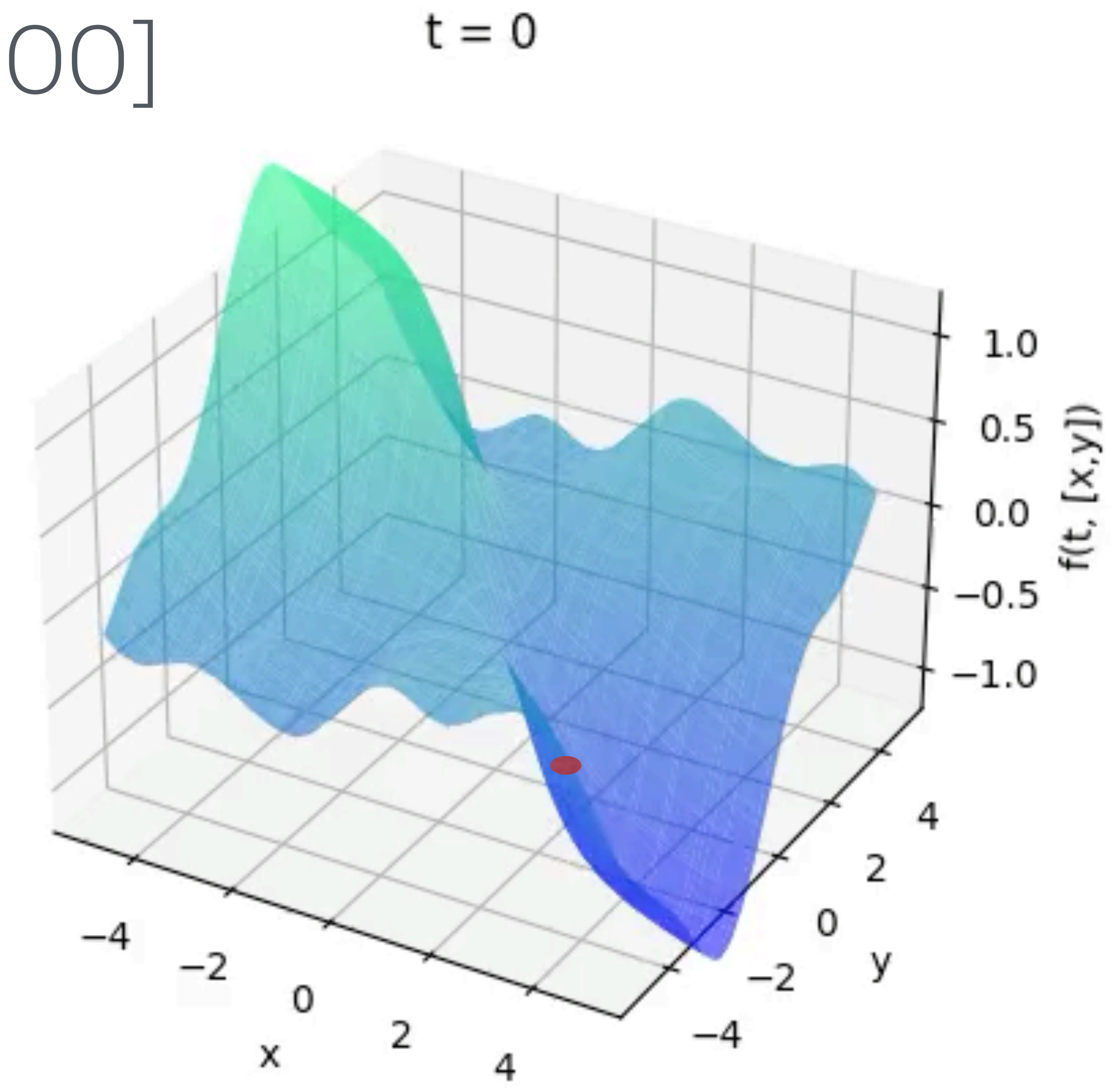
Cohen, Malka, ZR (2019) ; Halverson, Maiti, Stoner (2020); Helias Dahmen (2020)

Random DNNs induce a probability on function Space

Narrow [N=5]



Wide [N=500]



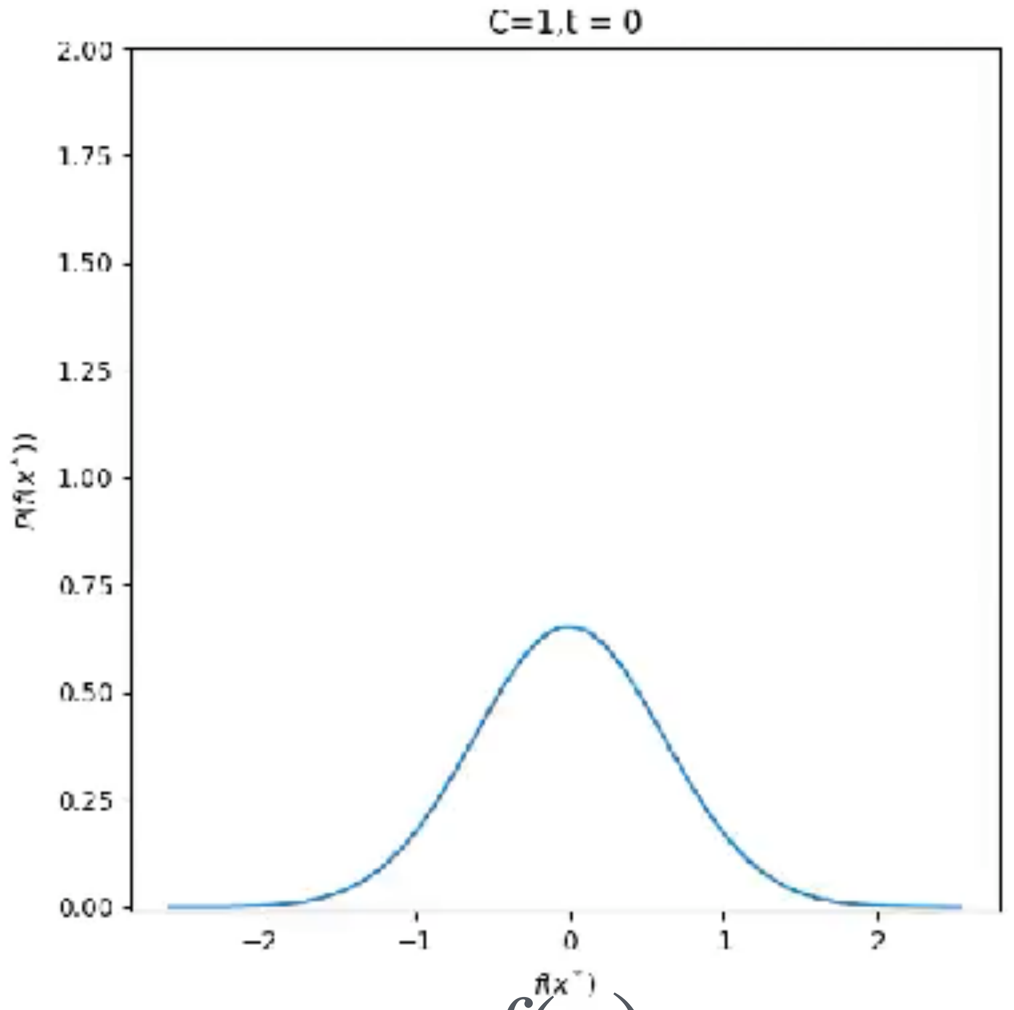
$$P[f] = \int_{-d^{-1/2}}^{d^{-1/2}} Dw_{11} \ldots \int_{-N^{-1/2}}^{N^{-1/2}} Da_1 \ldots \delta\left[f(\,.\,) - z_{w,a}(\,.\,)\right]$$

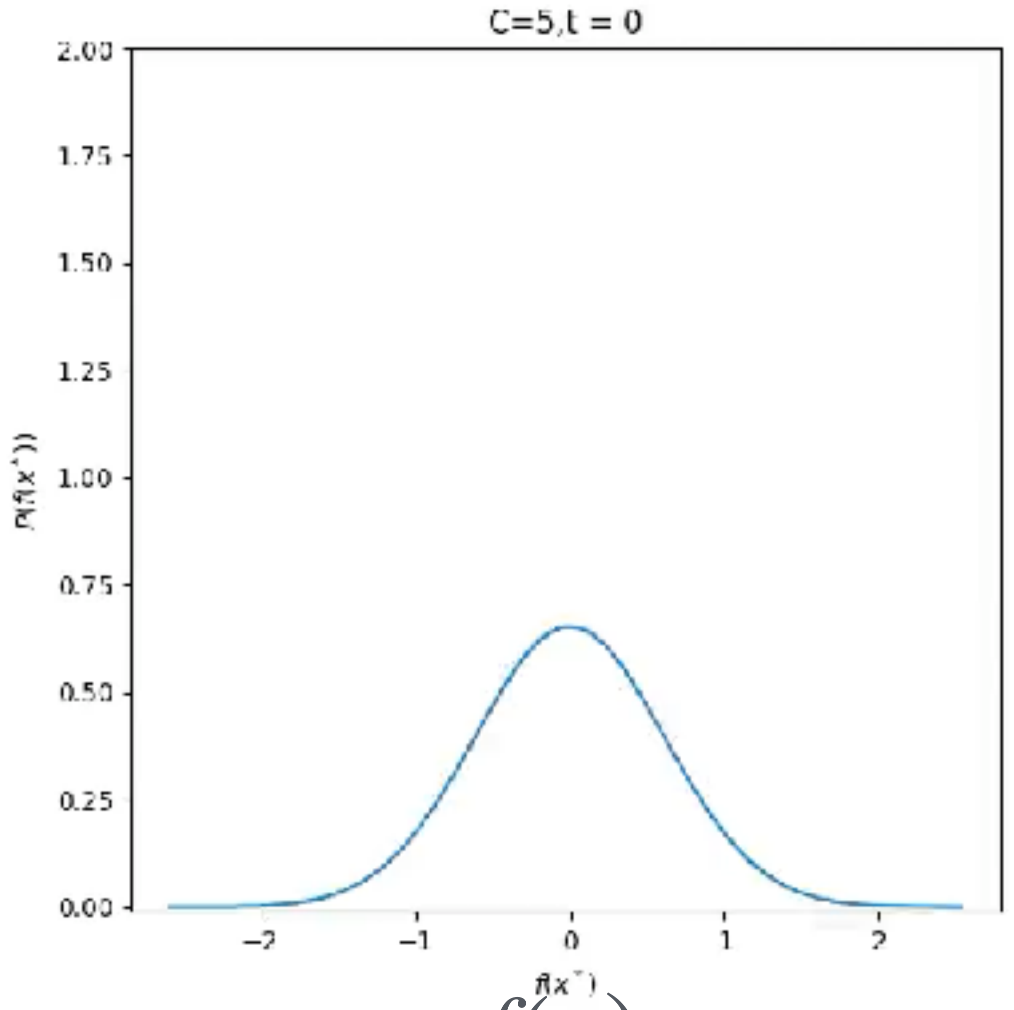
$$z_{w,a}(x) = \sum_{c=1}^{N} a_c Erf(w_c^T x) \quad x \in \mathrm{R}^2$$

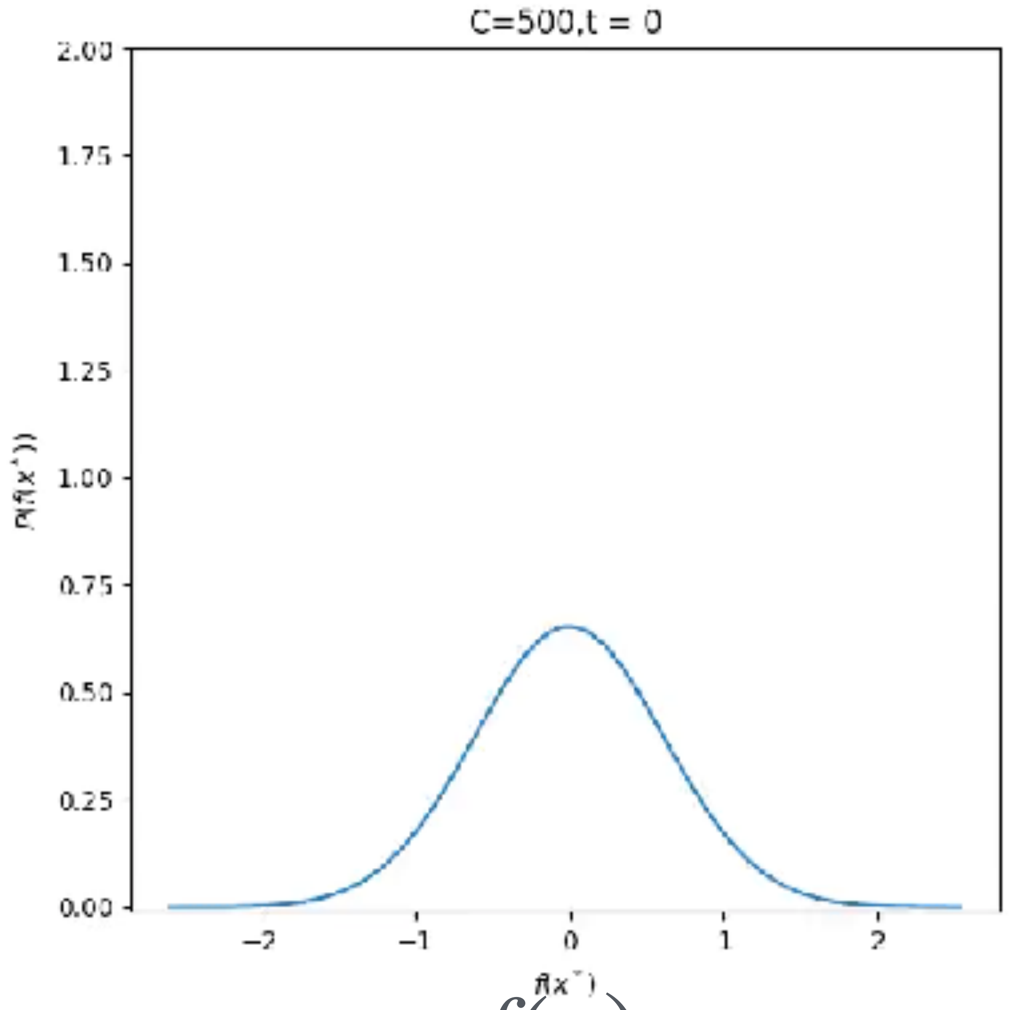
Random Infinite width DNNs - Free Field Theory

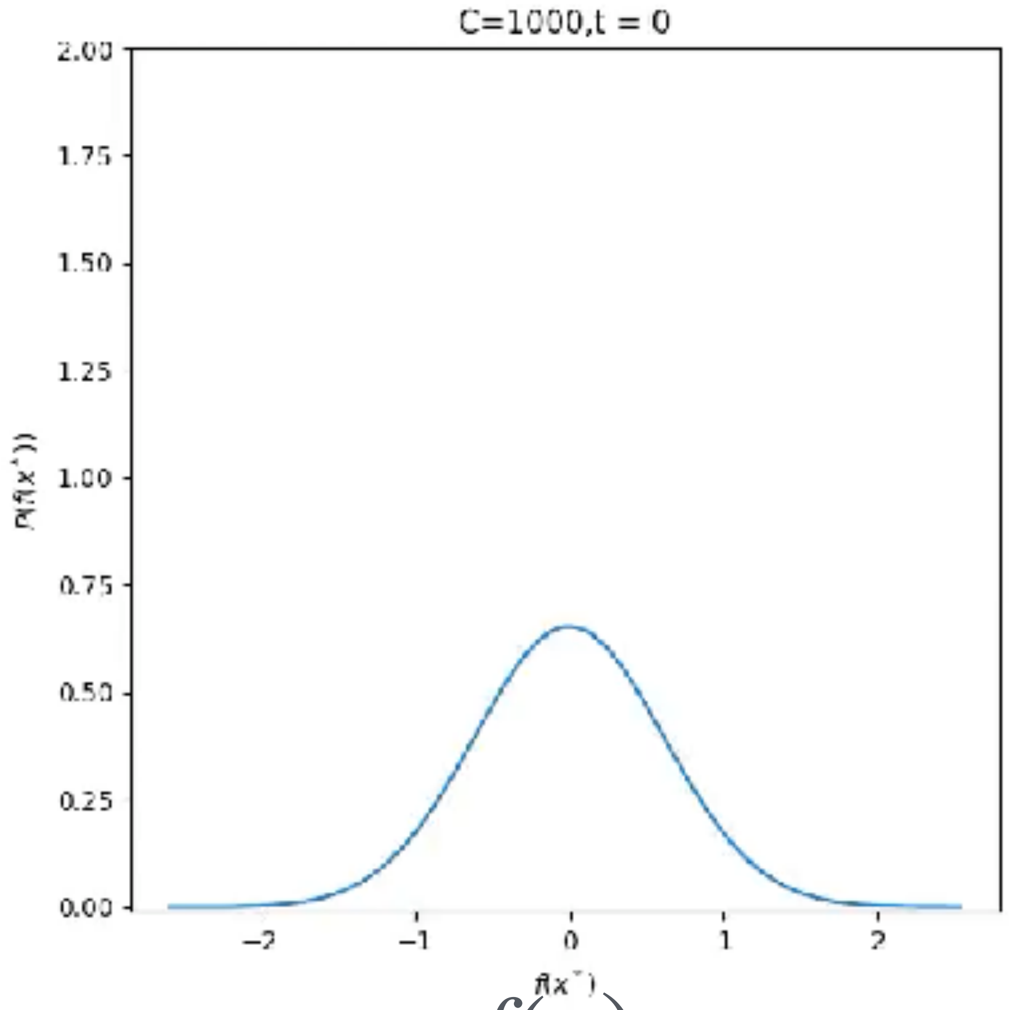
Width N=1 | N=5 | N=500 | N=1000

$P(f(\bullet))$

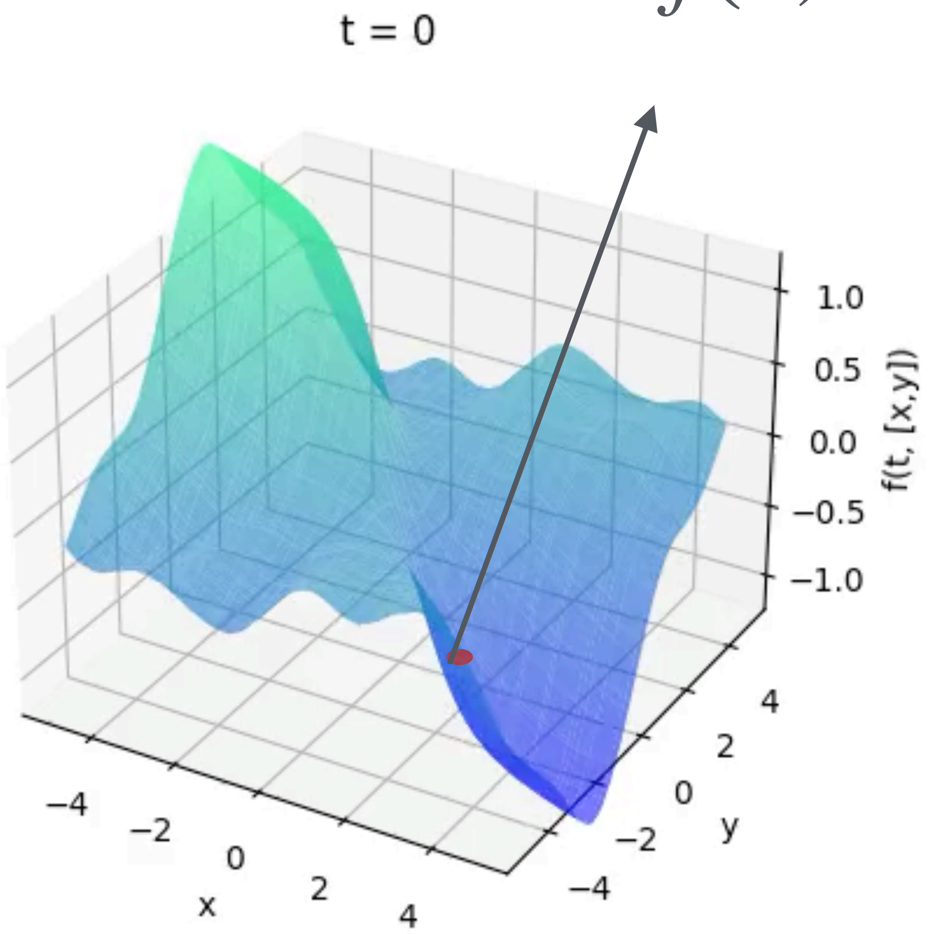








$f(\bullet)$ $f(\bullet)$ $f(\bullet)$ $f(\bullet)$



$$P[f] = \int_{-d^{-1/2}}^{d^{-1/2}} Dw_{11} \dots \int_{-N^{-1/2}}^{N^{-1/2}} Da_1 \dots \delta\left[f(\,.\,) - z_{w,a}(\,.\,)\right]$$

$$z_{w,a}(x) = \sum_{c=1}^{N} a_c Erf(w_c^T x) \quad x \in R^2$$

Random Infinite width DNNs - Analytical Results

- Infinite randomized DNNs generate a free field theory (Gaussian Process [R. Neal 1996])

- $z_{w,a}(x) = \sum_{c=1}^{N} a_c \phi(w_c^T x) \quad x \in \mathrm{R}^d$

Entropic term!

$$P[f] = \int_{-d^{-1/2}}^{d^{-1/2}} Dw_{11} \dots Dw_{Nd} \int_{-N^{-1/2}}^{N^{-1/2}} Da_1 \dots Da_N \delta\left[f(\,.\,) - z_{w,a}(\,.\,)\right] \propto_{N\to\infty} e^{-\frac{1}{2}\int dx dx' f(x) K^{-1}(x,y) f(x')}$$

$$K(x,x') = \langle z_{w,a}(x) z_{w,a}(x') \rangle_{uniform(w,a)}$$

- **Entropy generates a bias/entropic-force towards network output function which have large K(x,x') eigenvalues. For this simple network, this means a bias towards low order polynomials**

Field theory of an infinite network

$$Z_{t\to\infty} = \int d\theta e^{-(L[z]+\gamma|\theta|^2/2)/T} = \int d\theta e^{-\frac{\gamma|\theta|^2}{2T}} e^{-L[z]/T}$$

$$\underbrace{Z_{t\to\infty}}_{\text{Boltzmann}} \equiv \int d\theta e^{-(L[z]+\gamma|\theta|^2/2)/T} = \int Df \underbrace{\left(\int d\theta e^{-\frac{\gamma|\theta|^2}{2T}} \delta[f-z]\right)}_{\text{Output dist. Of random DNN}} e^{-L[f]/T} \equiv \int Df e^{-S}$$

$$S = -\log(P_{prior}) + \sum_{\mu=1}^{P} [f(x_\mu) - y(x_\mu)]^2/2T = \frac{1}{2}\int\int f K^{-1} f + \sum_{\mu=1}^{P} [f(x_\mu) - y(x_\mu)]^2/2T$$

It's A Gaussian Process!

Physical Analogy: Pinned Elastic Membrane with non-local elastic modulus

$$Z_{t\to\infty} \propto_{N\to\infty} \int Df e^{-\frac{1}{2}\int\int f K^{-1} f - \frac{1}{2T}\sum_{\mu=1}^{P}[f(x_\mu)-y_\mu]^2}$$

Elasticity <-> Entropy of weights given f

Pinning Potential <-> Training data

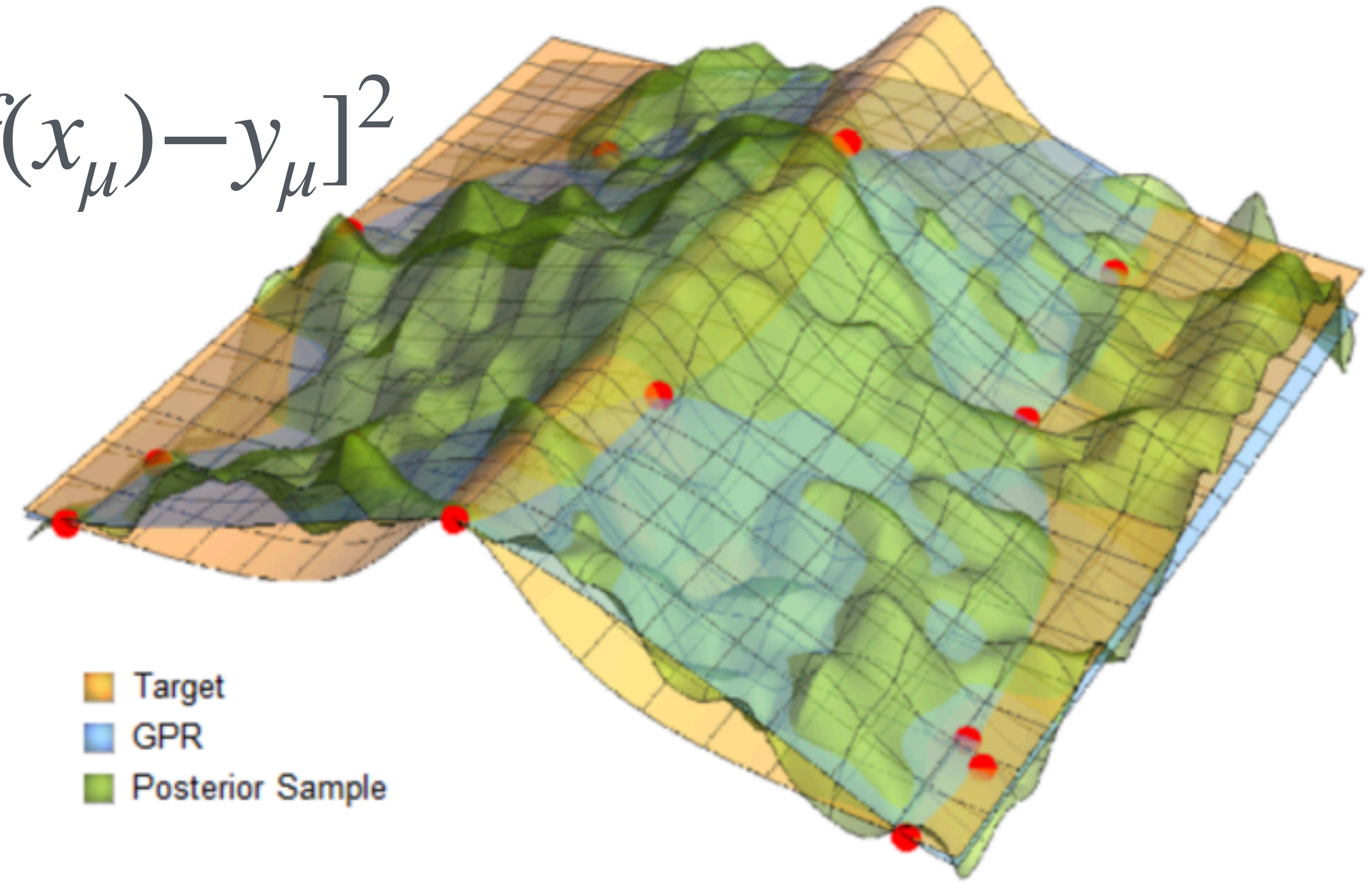


FIG. 1. **A physical picture of supervised deep learning.** The output of the DNN, as a function of input data, can be seen as an elastic membrane (surface) which relaxes to its equilibrium distribution during training. In this steady state

* From Cohen, Malka, ZR (2019)

** Silverman (1984); P. Sollich (2001); Bartlett et .al.(2019); Cohen, Malka, ZR (2019); Canatar, Bordelon, Cengiz (2019)

Strategy III

Truly analytic predictions require dataset averaging

GP limit and Dataset averaging

$$S_{N\to\infty,s.scaling} = \frac{1}{2}\int d\mu_x d\mu_y f(x)K^{-1}(x,y)f(y) + \sum_{\mu=1}^{P}[f(x_\mu) - y(x_\mu)]^2/2T = \frac{|f|^2_{RKHS}}{2} + \sum_{\mu=1}^{P}\frac{(f(x_\mu) - y(x_\mu))^2}{2T}$$

GP limit and Dataset averaging

$$S_{N\to\infty,s.scaling} = \frac{1}{2}\int d\mu_x d\mu_y f(x)K^{-1}(x,y)f(y) + \sum_{\mu=1}^{P}[f(x_\mu) - y(x_\mu)]^2/2T = \frac{|f|_{RKHS}^2}{2} + \sum_{\mu=1}^{P}\frac{f(x_\mu) - y(x_\mu)^2}{2T}$$

- Taking an extremum yields standard GPR predictor
- Disorder average (omitting replicas for clarity)

$$\left\langle Z_{t\to\infty}\right\rangle_{data} = \left\langle \int e^{-\frac{|f|_{RKHS}^2}{2}+L/T}\right\rangle_{data} = \int e^{-\frac{|f|_{RKHS}^2}{2}}\left\langle e^{-L/T}\right\rangle_{data} = \int e^{-\frac{|f|_{RKHS}^2}{2}}\exp\left(P\int d\mu_x e^{-L_x/T}\right)$$

$$\boxed{S_{N\to\infty,s.scaling,dataAv.} = \frac{1}{2}\int d\mu_x d\mu_y f(x)K^{-1}(x,y)f(y) - P\int d\mu_x e^{-[f(x)-y(x)]^2/2T}}$$

Let's understand (the absence of) overfitting

Three different treatments of the GP average action

Perturbative expansion [Cohen, Malka, Ringel (2019)]

$$S_{\dots} = \int d\mu_x d\mu_y f(x) K^{-1}(x,y) f(y) + \frac{n}{T}\int d\mu_x [f(x) - y(x)]^2 - n + O(1/T^2)$$

a.k.a. Equivalent Kernel [Silverman (1982)]

Gaussian Discrepancy Approximation [Canatar, Bordelon, Cengiz (2020)]

$$S_{\dots} \approx \int d\mu_x d\mu_y f(x) K^{-1}(x,y) f(y) - n \log\left(T + \int d\mu_x (f-y)^2\right)$$

Can be view as an RMT result
Simons et.al. (2023)

$$\approx \int d\mu_x d\mu_y f(x) K^{-1}(x,y) f(y) - \frac{n}{T + \langle \int d\mu_x (f-y)^2 \rangle_{MF}} \frac{\int d\mu_x (f-y)^2}{2}$$

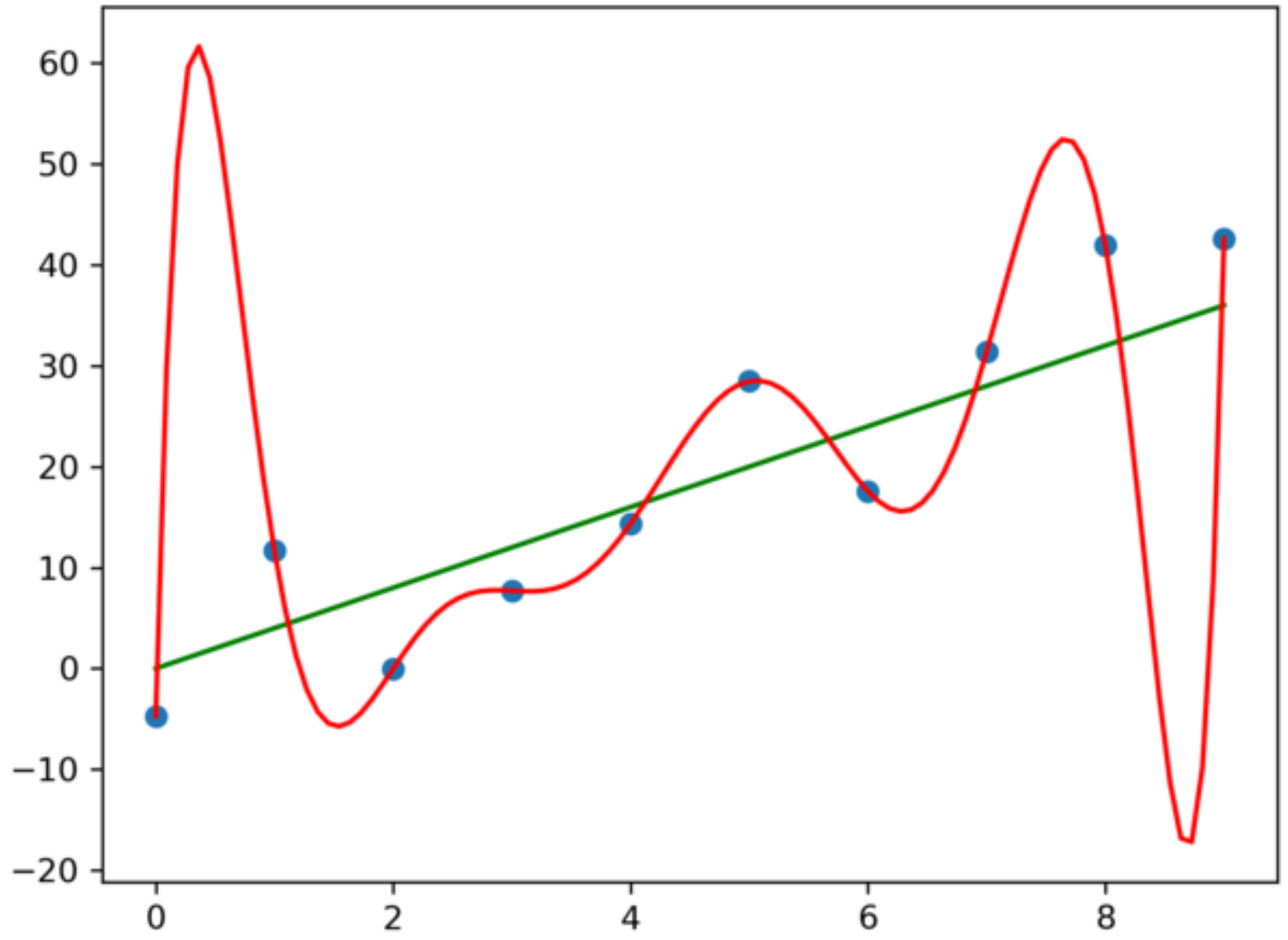
Renormalization-Group Flow [Howard, Maiti, Jefferson, Ringel (2024)]

$$S_{\dots,\Lambda} = \int d\mu_x d\mu_y f(x) K_\Lambda^{-1}(x,y) f(y) - n\int d\mu_x e^{-[f(x)-y(x)]^2/T(\Lambda)} + O(\lambda_\Lambda / T(\Lambda))^2$$

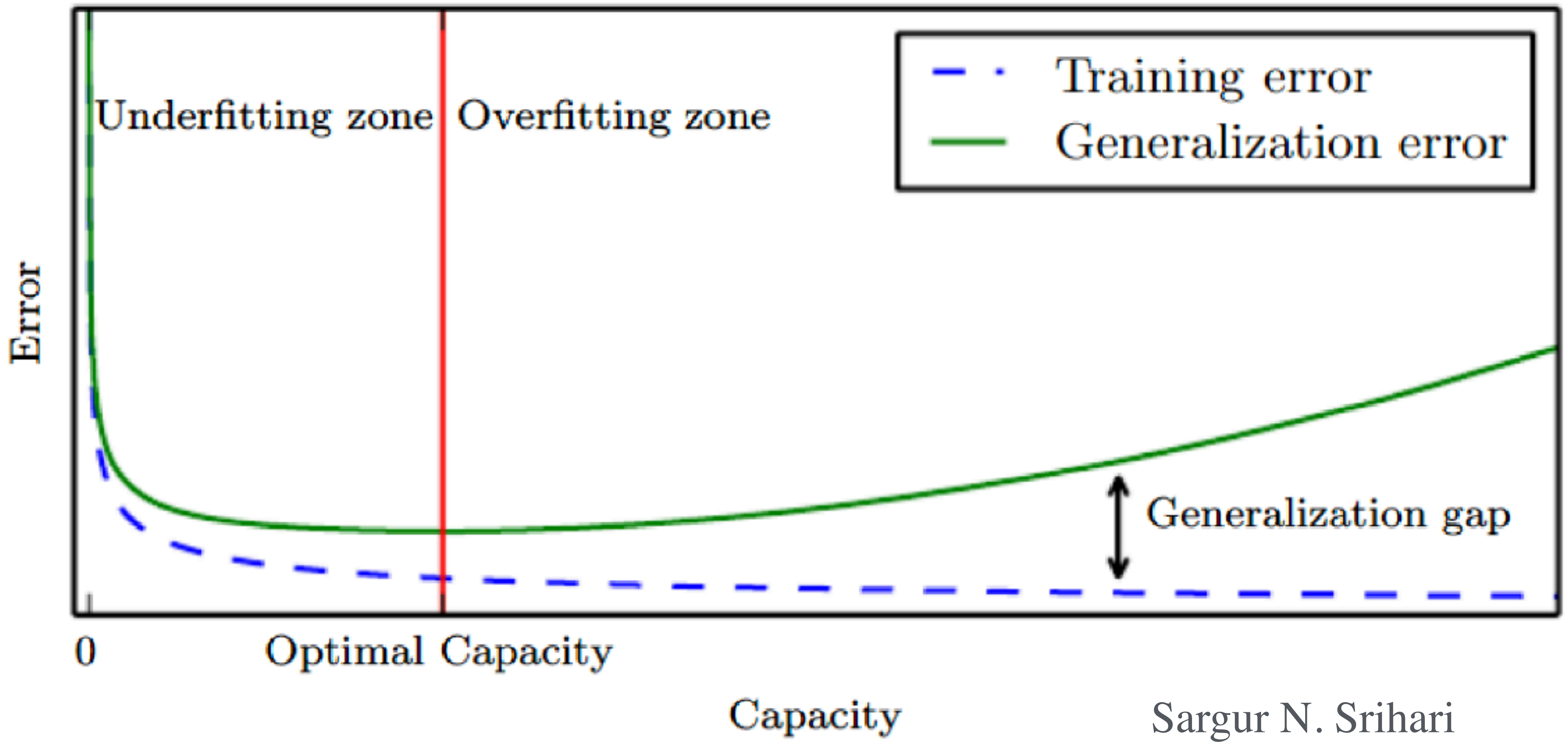
MSR + Disorder average [Helias, Dahmen 2020 Springer lecture notes in physics]

$$S_{N\to\infty, s.scaling, dataAv.} = \int d\mu_x d\mu_y f(x) K^{-1}(x,y) f(y) - n\int d\mu_x e^{-[f(x)-y(x)]^2/T}$$

“Classical” Thinking - Too much model capacity is bad.



Wiki



Sargur N. Srihari
Lecture Notes

Some saw through this

In the early 90's

For instance, there are many important questions regarding neural networks which are largely unanswered. There seem to be conflicting stories regarding the following issues:

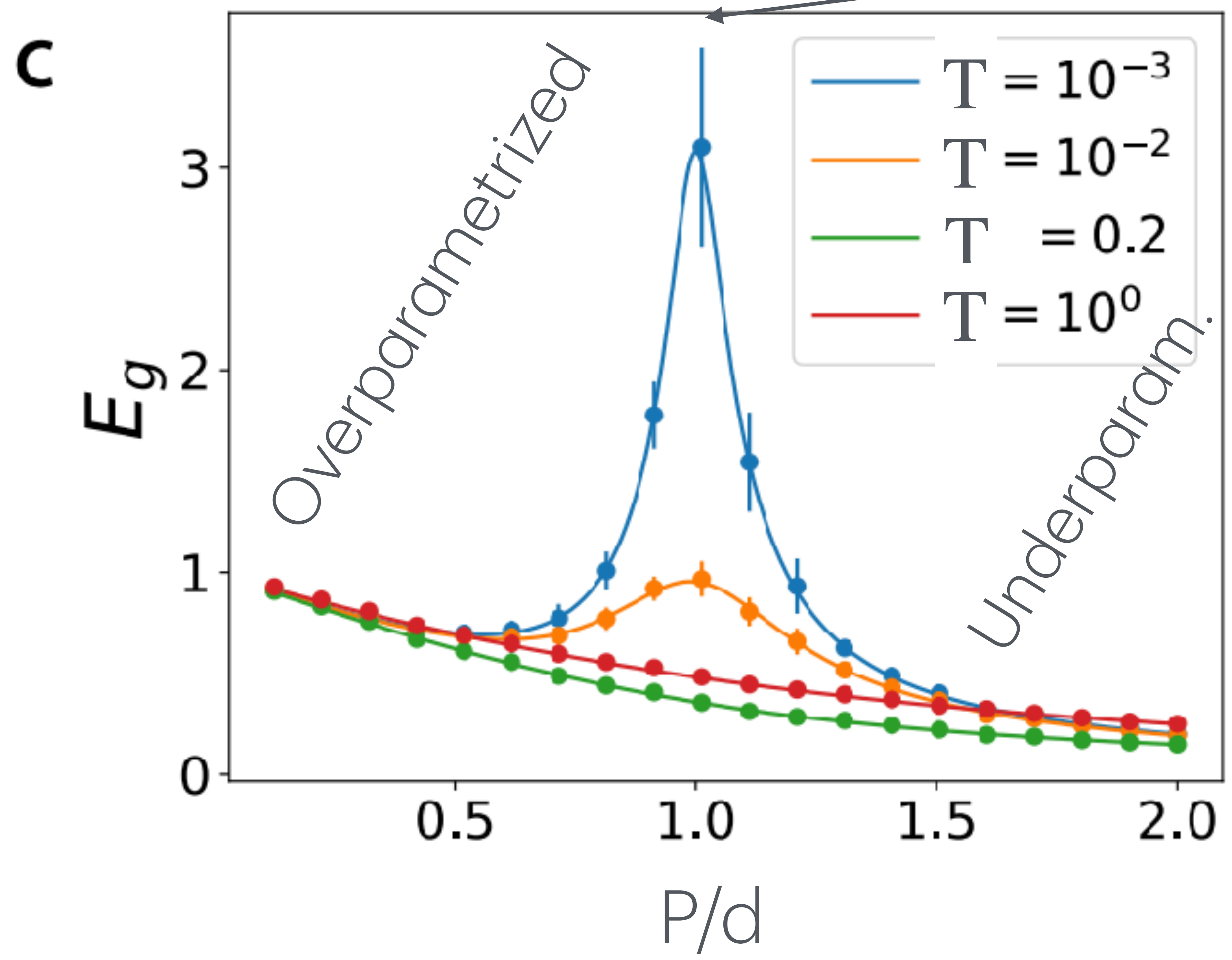
- Why don't heavily parameterized neural networks overfit the data?
- What is the effective number of parameters?
- Why doesn't backpropagation head for a poor local minima?
- When should one stop the backpropagation and use the current parameters?

Reflection after refereeing paper for NIPS,
Leo Breiman, 1995,

A classical example of overfitting

Finite rank degenerate kernels and noisy target

Interpolation Threshold



From Canatar et. al. 2020

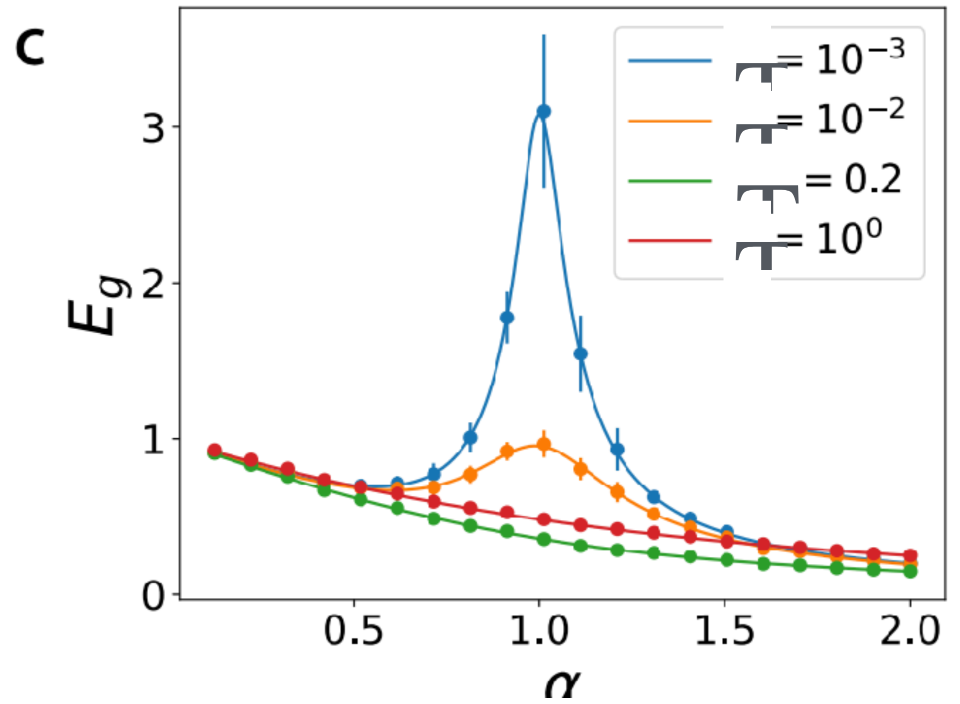
$$K(x, x') = \lambda \sum_{k=1}^{d} \phi_k(x)\phi_k(x') \quad y(x) = y_{clean}(x) + \xi(x)$$

Real Kernels are different

$$K(x, x') = \lambda \sum_{k=1}^{d} \phi_k(x)\phi_k(x') \Rightarrow \sum_{k=1}^{\infty} \lambda_k \phi_k(x)\phi_k(x') \qquad \lambda_k \propto k^{-1-\alpha}$$

Also input dimension and input entropy is typically very large.

Main insight — T kills the peak around the interpolation threshold. Many very small kernel modes can induce an effective temperature/T.



First step: Large T behavior - Eigenlearning

$$S_{N\to\infty,s.scaling,dataAv.} = \frac{1}{2}\int d\mu_x d\mu_y f(x)K^{-1}(x,y)f(y) - P\int d\mu_x e^{-[f(x)-y(x)]^2/2T}$$

$$\approx_{T\gg Discrepancy} \frac{1}{2}\int d\mu_x d\mu_y f(x)K^{-1}(x,y)f(y) + \frac{P}{2T}\int d\mu_x [f(x)-y(x)]^2$$

$$=_{Diagonalization} \frac{1}{2}\sum_k \lambda_k^{-1} f_k^2 + \frac{P}{T}(f_k - y_k)^2$$

Learning decouples in kernel eigenfunction space. We learn well eigenfunctions for which

$$\langle f_k \rangle = \frac{\lambda_k}{\lambda_k + T/P} y_k \quad Var[f_k] = \frac{\lambda_k T/P}{\lambda_k + T/P}$$

J. B. Simon et. al. 2023

Low T behavior - Non-perturbative approaches

$$S_{N\to\infty,s.scaling,dataAv.} = \frac{1}{2}\int d\mu_x d\mu_y f(x)K^{-1}(x,y)f(y) - P\int d\mu_x e^{-[f(x)-y(x)]^2/2T}$$

$$\approx_{T\gg Discrepancy} \frac{1}{2}\int d\mu_x d\mu_y f(x)K^{-1}(x,y)f(y) + \frac{P}{2T}\int d\mu_x [f(x)-y(x)]^2$$

Gaussian Discrepancy Approximation [Canatar, Bordelon, Cengiz (2020)]

$$S_{...} \approx \int d\mu_x d\mu_y f(x)K^{-1}(x,y)f(y) + P\log\left(T + \int d\mu_x (f-y)^2/2\right)$$

T->0 Strong interactions

$$\approx \int d\mu_x d\mu_y f(x)K^{-1}(x,y)f(y) + \frac{P}{T + \frac{1}{2}\langle \int d\mu_x (f-y)^2\rangle_{MF}} \frac{\int d\mu_x (f-y)^2}{2}$$

Renormalization-Group Flow [Howard, Maiti, Jefferson, Ringel (2024)]

Baby version also appeared in [Cohen, Malka, Ringel (2019)]

$$S_{...,\Lambda} = \int d\mu_x d\mu_y f(x)K_\Lambda^{-1}(x,y)f(y) - P\int d\mu_x e^{-[f(x)-y(x)]^2/T(\Lambda)} + O(\lambda_\Lambda/T(\Lambda))^2$$

Experiments:

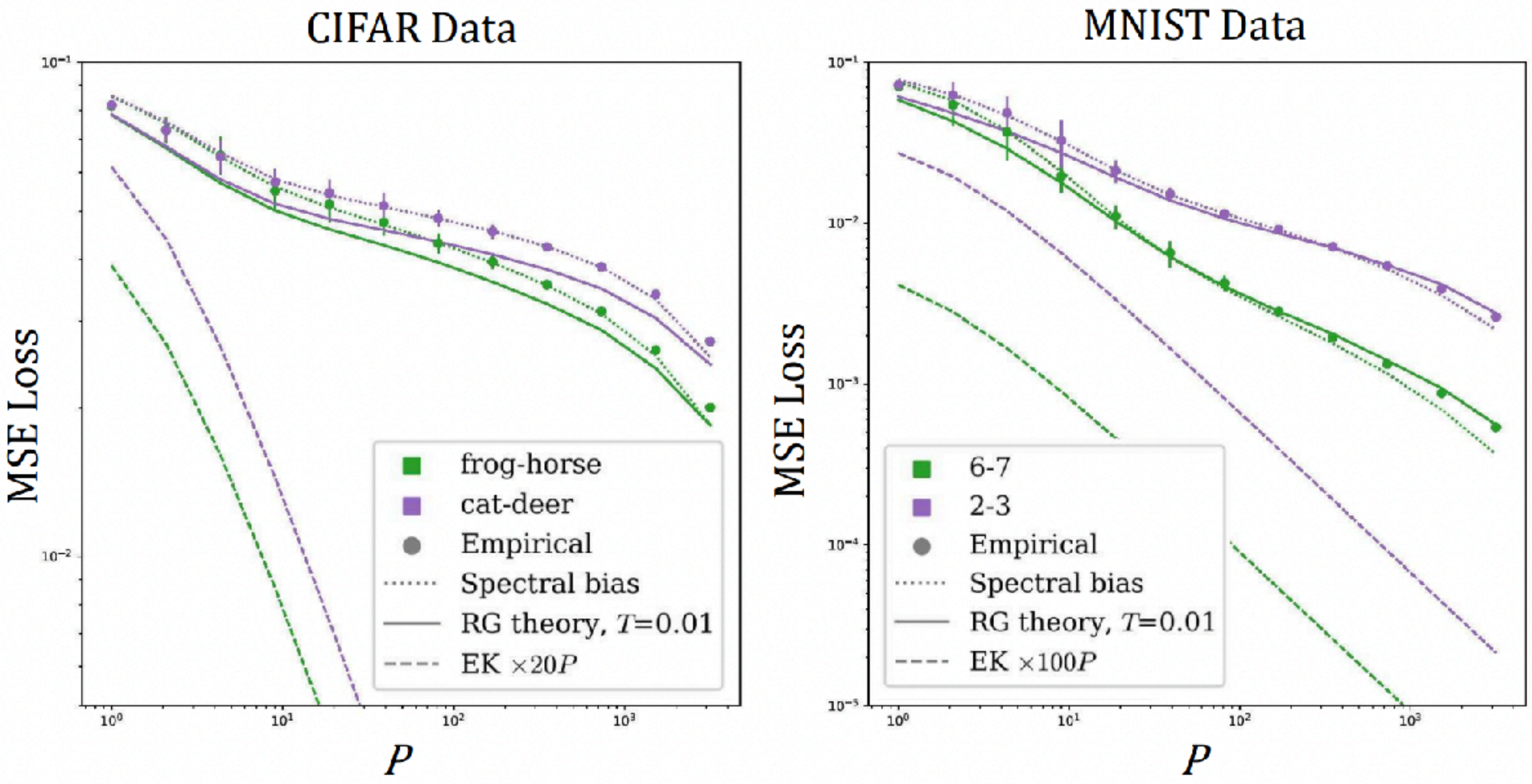


Figure 2.1: Gaussian Processes Regression on four 10k binary CIFAR and MNIST datasets, at $\kappa^2 = 1e - 8$. Experimental results (dots) match well both the effective ridge theory and the RG theory. In the latter, we took 0.01-learnability as marking the RG cut-off. We comment that results are similarly accurate for $T = 0.001$ and $T = 0.1$. The Equivalent Kernel estimator is expected to become accurate when the loss reaches the scale of κ^2, explaining its poor performance in the shown range of P.

Applications of Statistical Field Theory in Deep Learning

Zohar Ringel, Noa Rubin, Edo Mor, Moritz Helias, Inbar Seroussi

https://arxiv.org/pdf/2502.18553

GP limits: Open Questions

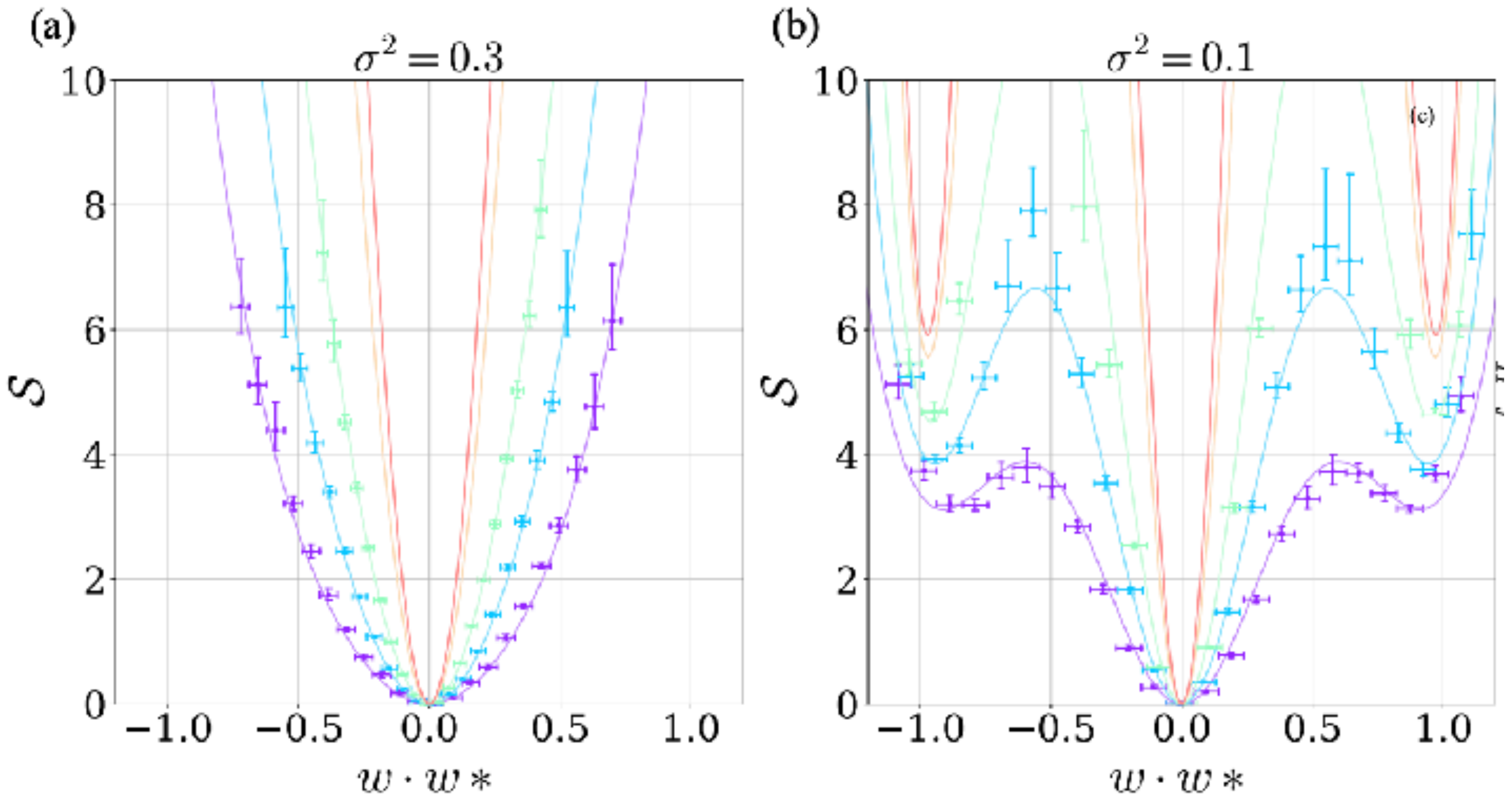
- What lays beyond the Gaussian Discrepancy Approximation? [e.g. Howard et. al. find position dependent temperature/ridge]
- GP Limits of Diffusion Models
- GP Limits of Physically Informed Neural Networks [some first steps carried in Miron, Seroussi, Ringel 2023]

Feature/Representation Learning

AlexNet

Bridging Feature Learning and Mechanistic Interpretation

Theory



Order Parameter:

Linear super-position of neurons in input/middle layer

Practice

Claude

At the end of May, Anthropic released Golden Gate Claude to the public for 24 hours. This manipulated version of Anthropic's *Claude 3 Sonnet* model had an obvious obsession with the Golden Gate Bridge.

> *If you ask this "Golden Gate Claude" how to spend $10, it will recommend using it to drive across the Golden Gate Bridge and pay the toll. If you ask it to write a love story, it'll tell you a tale of a car who can't wait to cross its beloved bridge on a foggy day. If you ask it what it imagines it looks like, it will likely tell you that it imagines it looks like the Golden Gate Bridge.*

Order Parameter:

super-position of neurons in middle layer

- Rubin, Seroussi, ZR, ICLR, (2023).

Understanding Anthropic's Golden Gate Claude, Davis, Medium (2024)

Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet, **Anthropic**, Transformer-Circuit (2024)

Feature Learning Regime - Data averaged

Finite N or MF.Scaling

- Focus on a two-layer network for simplicity $$z(x) = \sum_{c=1}^{N} a_c \phi(w_c^T x)$$
- Introduce two auxiliary fields for every layer except the input layer by introducing functional delta functions
- Integrate out all weights expect input layer weights

$$S = \sum_c \frac{d\,|w_c|^2}{2\sigma_w^2} + \frac{\sigma_a^2}{N}\sum_{c=1}^{N}\int d\mu_x d\mu_y t(x)\phi(w_c^T x)\phi(w_c^T y)t(y) + i\int d\mu_x tf - P\int d\mu_x e^{-[f(x)-y(x)]^2/T}$$

A list of actions

- GP limit, any networks, any depth

$$S = \frac{1}{2}\int d\mu_x d\mu_y f(x) K^{-1}(x,y) f(y) - P\int d\mu_x e^{-[f(x)-y(x)]^2/T}$$

- Two layer DNN

$$S = \sum_c \frac{d\,|w_c|^2}{2\sigma_w^2} + \frac{\sigma_a^2}{N}\sum_{c=1}^{N}\left(\int d\mu_x t(x)\phi(w_c^T x)\right)^2 + i\int d\mu_x t f - P\int d\mu_x e^{-[f(x)-y(x)]^2/T}$$

- Three layer DNN

$$S = \frac{d\,|w^{(0)}|^2}{2} - i\int d\mu_x \left[\tilde{f}(x) f(x) + \sum_i \tilde{h}_i^{(1)}(x) h_i^{(1)}(x)\right] + \frac{1}{2N^{(1)}}\sum_i \left(\int d\mu_x \tilde{f}(x)\sigma(h_i^{(1)}(x))\right)^2 + \frac{1}{2N^{(0)}}\sum_{ij}\left(\int d\mu_x \tilde{h}_i^{(1)}(x)\sigma(w_j^{(0)}\cdot x)\right)^2 - P\int d\mu_x e^{-[f(x)-y(x)]^2/T}$$

Recovering the NNGP limit

$$S = \sum_c \frac{d\,|w_c|^2}{2\sigma_w^2} + \frac{\sigma_a^2}{N}\sum_{c=1}^{N}\int d\mu_x d\mu_y t(x)\phi(w_c^T x)\phi(w_c^T y)t(y) + i\int d\mu_x tf - P\int d\mu_x e^{-[f(x)-y(x)]^2/2T}$$

- As $N \to \infty$ each individual $\mathbf{w_c}$ feels t(x) less and less, hence stays in its Gaussian prior
- However t(x)t(y) see the aggregate effect of all $\phi(w_c^T x)\phi(w_c^T y)$ leading to

$$S = \sum_c \frac{d\,|w_c|^2}{2\sigma_w^2} + \frac{1}{2}\int d\mu_x d\mu_y t(x)\langle \sigma_a^2 \phi(w^T x)\phi(w^T y)\rangle_{w\sim\mathcal{N}} t(y) + i\int d\mu_x tf - \dots$$

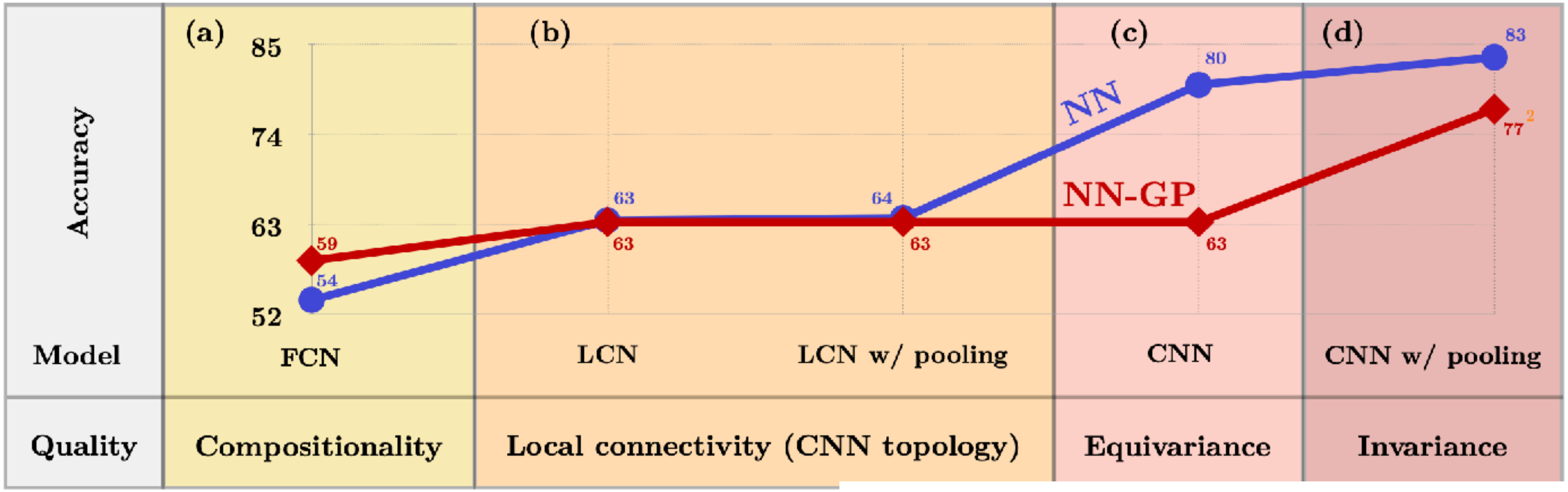
K(x,y)

- Integrating out t using square completion we obtain our previous NNGP action for the output

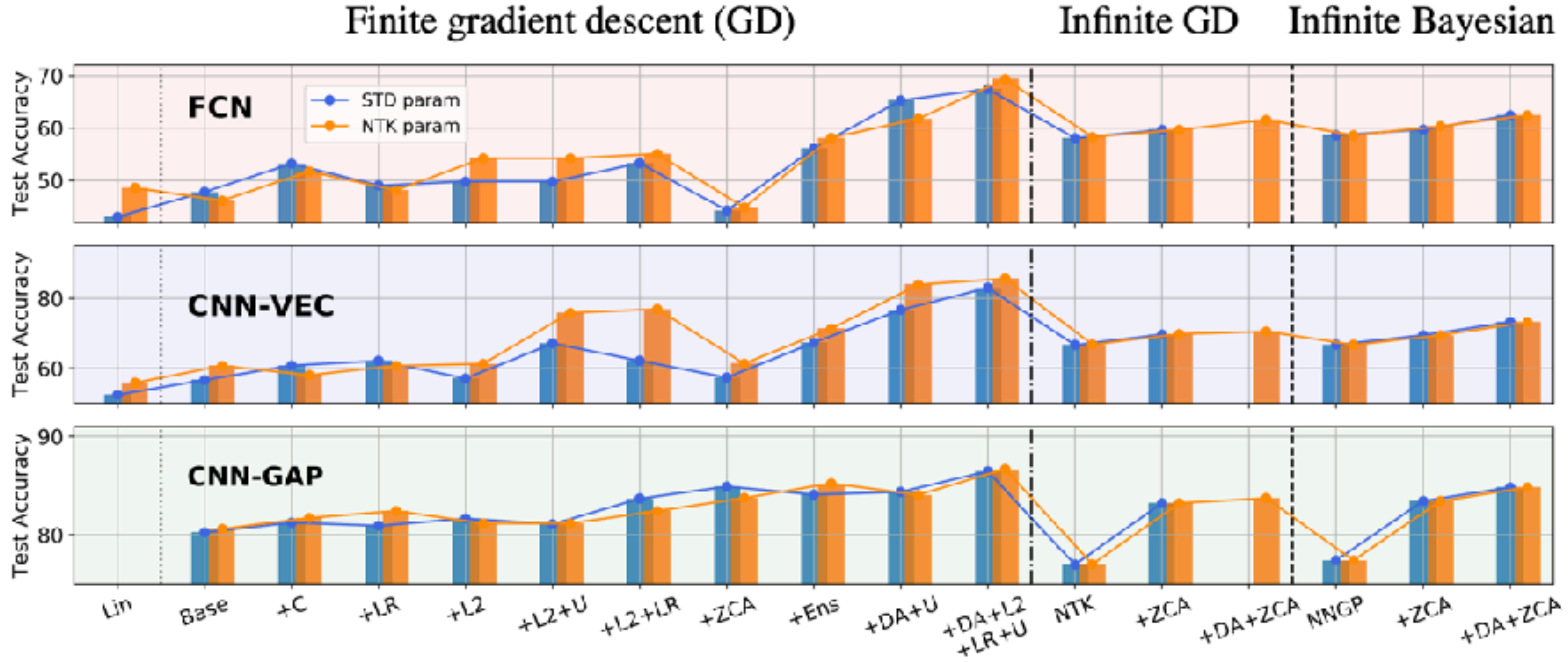
$$S_{N\to\infty, s.scaling} = \frac{1}{2}\int d\mu_x d\mu_y f(x)K^{-1}(x,y)f(y) - P\int d\mu_x e^{-[f(x)-y(x)]^2/2T}$$

Are there qualitative performance differences between the GP limit and feature learning regime?

The case in favor of Gaussian Processes



Novak et. al. 2019 (see also Lee. 2020)



The case against GPs

- The common lore is that feature learning is an essential ingredient in deep learning.
- We know that converges to the GP limit only happens when width=N>>P, which is very unrealistic
- We know of toy settings where GP qualitatively underperforms real DNNs
- GPs are also very-inefficient at large datasets, as they require inverting a P by P matrix.

Example: staircase learning

$$z(x) = \sum_{c=1}^{N} a_c Erf(w_c^T x) \qquad y(x) = (w_* \cdot x) + \epsilon (w_* \cdot x)^3 \qquad x \in R^d$$

Sample Complexity Question: What is the scaling of P with d, required to learn 90% of the cubic component?

Intuition:

1. The Gaussian Process, being linear in the target, ``learns" the linear part and cubic parts separately. For symmetric kernels and datasets, learning the cubic part requires $\mathrm{P{=}d^3}$

2. The actual DNN, being non-linear in the target, can learn to focus on the w_* direction based on the linear part, so much that the effective data dimension becomes $\mathrm{O(1)}$, in which case learning a cubic function is not very hard.

E Abbe et. al. 2021

GP solution and why $P = O(d^3)$

A dash of representation theory

$$z(x) = \sum_{c=1}^{N} a_c Erf(w_c^T x) \qquad y(x) = (w_* \cdot x) + \epsilon (w_* \cdot x)^3 \qquad x \in S^d$$

$$K(x, x') \equiv \langle z(x) z(x') \rangle_{a,w \sim \mathcal{N}} = F(|x|, |x'|, x \cdot x') \Rightarrow K(Ox, Ox') = K(x, x') \quad O \in \mathrm{O}(d)$$

$$\Swarrow$$

$$K(x, x') = \sum_{lm} \lambda_l Y_{lm}(x) Y_{lm}(x') \qquad m \in [1..O(d^l)]$$

$$y(x) = \sum_{m \in [1..d]} a_m Y_{1m}(x) + \sum_{m \in [1..O(d^3)]} c_m Y_{3m}(x)$$

Since the cubic part is determined by O(d^l) coefficients, which are equally probable in the prior + linear part doesn't help — we find **P=O(d^3)** data-points are required to fix these coefficients.

Beyond GP: Kernel adaptation approximation

$$S = \sum_c \frac{d|w_c|^2}{2\sigma_w^2} + \frac{\sigma_a^2}{N}\sum_{c=1}^{N}\int d\mu_x d\mu_y t(x)\phi(w_c^T x)\phi(w_c^T y)t(y) + i\int d\mu_x tf - P\int d\mu_x e^{-[f(x)-y(x)]^2/2T}$$

Kernel Adaptation [Seroussi (2021) ; also Aitchison (2019), Cengiz (2022), Helias (2024), Mallet (2024)]

$$\int\int t(x)\phi(w_c^T x)\phi(w_c^T y)t(y) \approx \int\int \langle t(x)\rangle_{MF}\phi(w_c^T x)\phi(w_c^T y)\langle t(y)\rangle_{MF} + \int\int t(x)\langle\phi(w_c^T x)\phi(w_c^T y)\rangle_{MF} t(y)$$

Backprop effects on input weights $\quad$ $K_{MF}(x,y)$

$$\langle t(x)\rangle = \frac{P}{T}\left[\langle f(x)\rangle - y(x)\right]$$

- *For kernel scale as order parameter*: Li & Sompolinsky PRX (2021) Hanin et. al. (2023) [linear FCNs]; Arioso et. al. Nat. ML. (2023);

Kernel Adaptation in some more detail

Data average case

$$S = \sum_c \frac{d|w_c|^2}{2\sigma_w^2} + \frac{\sigma_a^2}{N}\sum_{c=1}^{N}\int d\mu_x d\mu_y t(x)\phi(w_c^T x)\phi(w_c^T y)t(y) + i\int d\mu_x tf - n\int d\mu_x e^{-[f(x)-y(x)]^2/T}$$

Adaptive Kernel decoupling + Leading order expansion in (f-y)^2/T [a.k.a Equivalent Kernel]

$$S = \sum_{c=1}^{N} S_{MF}[w_c] + S_{MF}[f,t] =_{\int Dt} \sum_{c=1}^{N} S_{MF}[w_c] + S_{MF}[f]$$

$$S_{MF}(w) = \frac{d|w|^2}{2\sigma_w^2} + \frac{\sigma_a^2}{2N}\int\int \langle t(x)\rangle_{MF}\langle t(y)\rangle_{MF}\phi(w^T x)\phi(w^T y)$$

$$S_{MF}(f) = \frac{1}{2}\int\int fK_{MF}^{-1}f + L[f] \qquad \langle t\rangle_{MF} = [K_{MF} + \sigma^2/P]^{-1}y \qquad K_{MF}(x,y) = \langle\phi(w^T x)\phi(w^T y)\rangle_{MF}$$

Kernel Adaptation in some more detail

Data average case

$$S = \sum_c \frac{d|w_c|^2}{2\sigma_w^2} + \frac{\sigma_a^2}{N}\sum_{c=1}^{N}\int d\mu_x d\mu_y t(x)\phi(w_c^T x)\phi(w_c^T y)t(y) + i\int d\mu_x tf - n\int d\mu_x e^{-[f(x)-y(x)]^2/T}$$

Adaptive Kernel decoupling + Leading order expansion in (f-y)^2/T [a.k.a Equivalent Kernel]

$$S = \sum_{c=1}^{N} S_{MF}[w_c] + S_{MF}[f,t] =_{\int Dt} \sum_{c=1}^{N} S_{MF}[w_c] + S_{MF}[f]$$

$$S_{MF}(w) = \frac{d|w|^2}{2\sigma_w^2} + \frac{\sigma_a^2}{2N}\int\int \langle t(x)\rangle_{MF}\langle t(y)\rangle_{MF}\phi(w^T x)\phi(w^T y) \overset{\text{GFL}}{\approx}_{d,n\to\infty} \frac{1}{2}w^T\Sigma_{MF}^{-1}w$$

$$S_{MF}(f) = \frac{1}{2}\int\int fK_{MF}^{-1}f + L[f] \qquad \langle t\rangle_{MF} = [K_{MF} + \sigma^2/P]^{-1}y \qquad K_{MF}(x,y) = \langle\phi(w^T x)\phi(w^T y)\rangle_{w\sim\mathcal{N}(\Sigma_{MF})}$$

GFL - Gaussian Feature Learning

Article https://doi.org/10.1038/s41467-023-36361-y

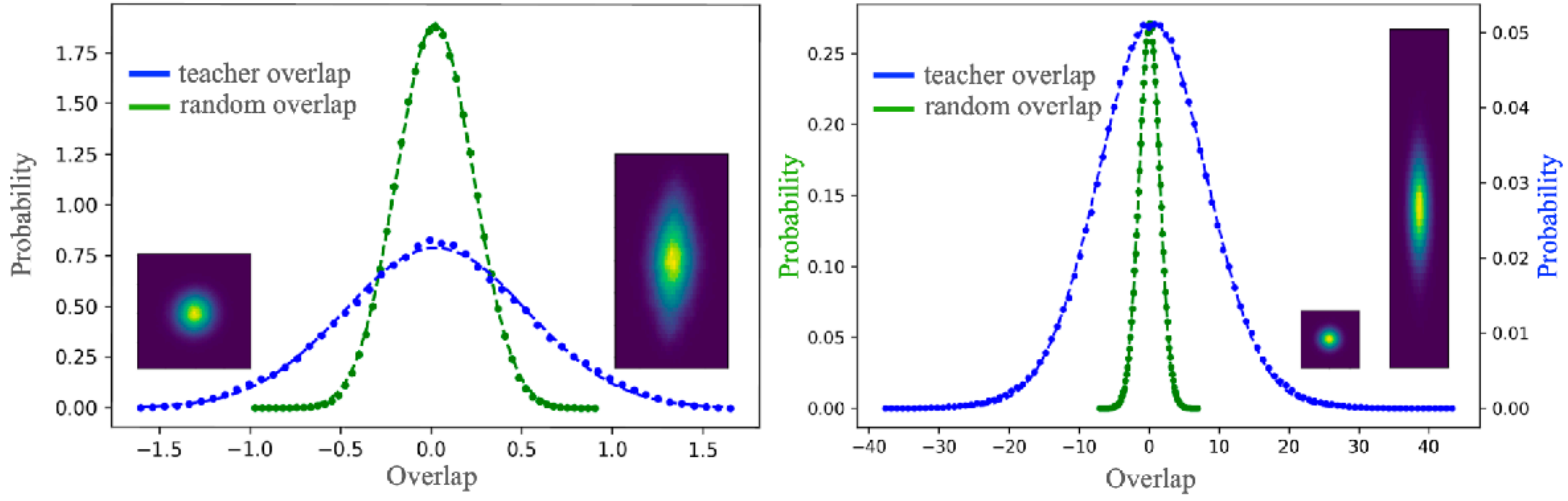
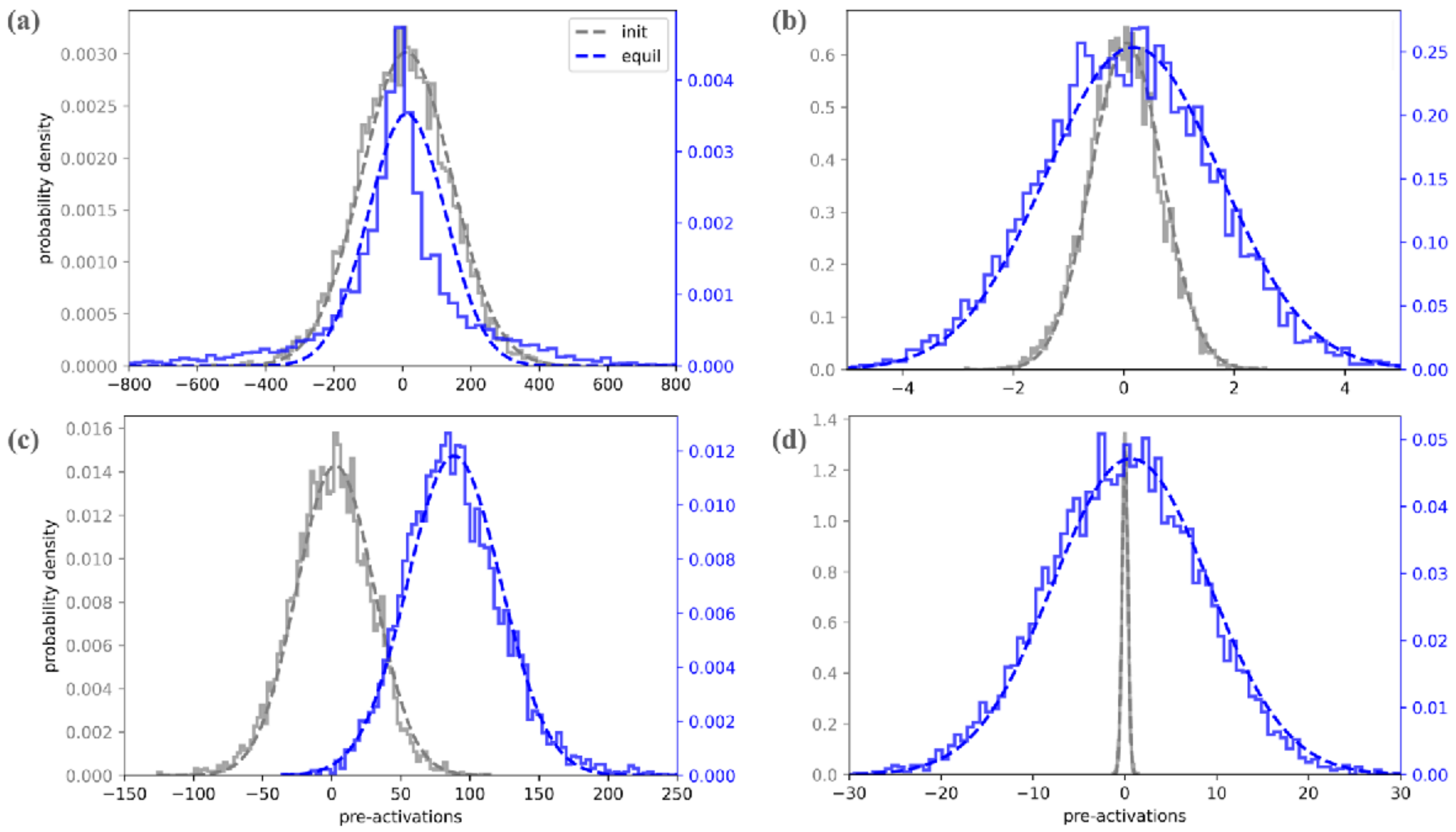


Fig. 4 | Pre-activation statistics in the 3-layer student-teacher setting. Left. Histogram of student input weight vector, dotted with the normalized teacher weight (blue) and normalized random weight (green). Right. Histogram of student hidden layer pre-activations, dotted with the normalized teacher pre-activations (blue) and normalized pre-activations of a random teacher (green). Dots are empirical values and dashed lines are Gaussian fits. Insets: 2d histograms along the same vectors before (left) and after (right) training. Within our framework, these variances are determined by $\mathbf{v}^{\top}K^{(l)}\mathbf{v}$ with $\mathbf{v}$ being either random unit vector or $h^{(l)}$ of the single channel teacher. Remarkably, despite strong changes to the kernels and various non-linearities in the action, the pre-activation is almost perfectly Gaussian.

3 layer non-linear CNN, student teacher setting

Seroussi, Naveh, ZR (2021)

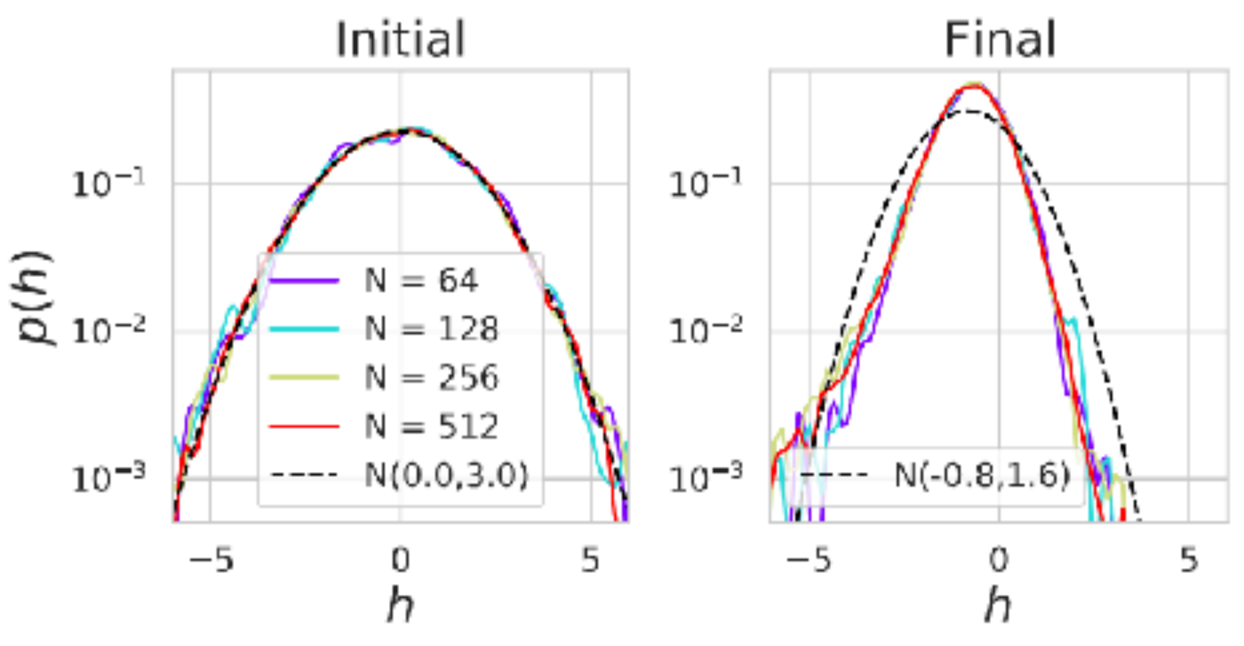
GFL - Real Networks and real datasets



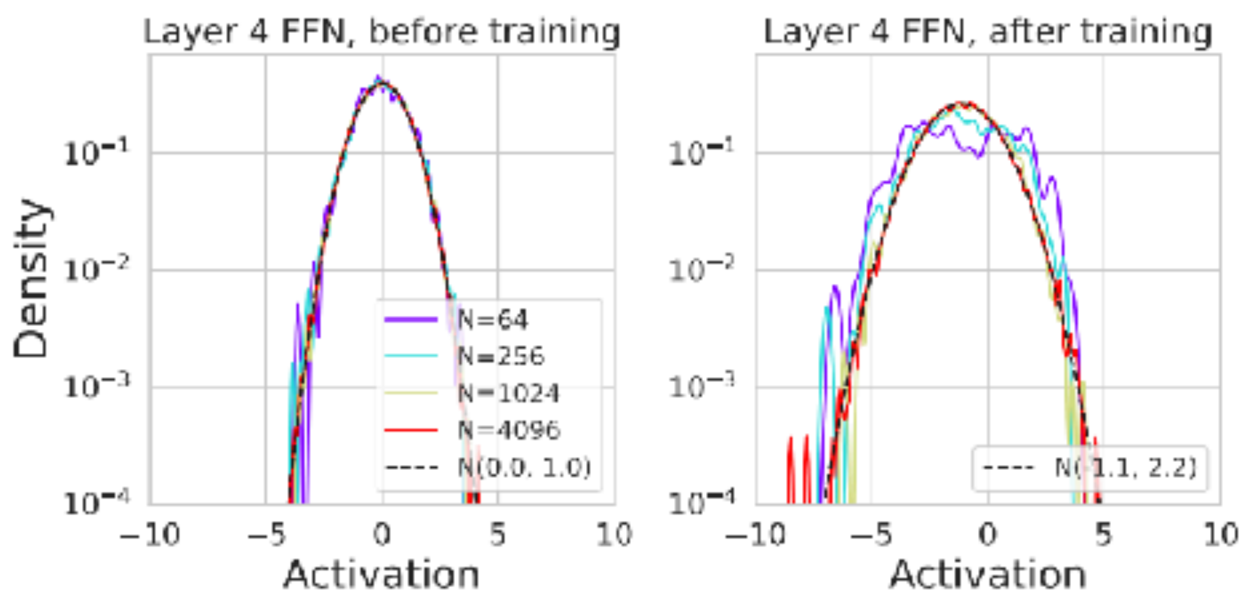
5 Layer non-linear CNN with pooling on CIFAR-10

Seroussi, Naveh, ZR (2021)

GFL - Real Networks and Real datasets



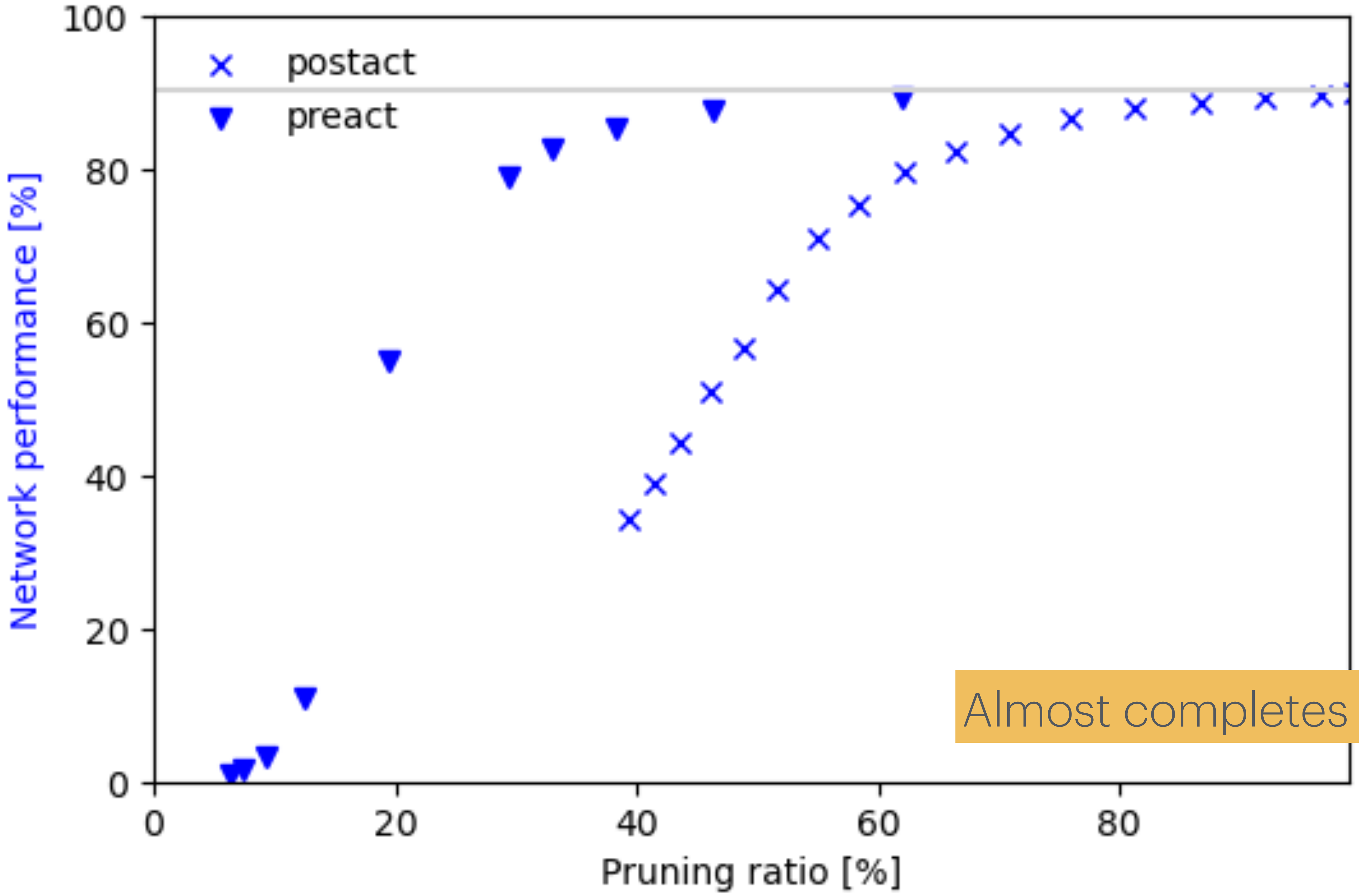
(a) Preactivations CIFAR-5M



(d) FFN Preactivations Wikitext

Bordellon et. Al. (2023)

GFL - Evidence from pruning



Almost completes with Zhang et .al. 2017

VGG-19 [from Torch-vision] pruned by projecting out low latent layer kernel eigenvalues

K. Fisher, M. Helias, Z. Ringel [to be published]

Kernel Adaptation: Exploiting symmetries

$$S = \sum_{c=1}^{N} S_{MF}[w_c] + S_{MF}[f, t] =_{\int Dt} \sum_{c=1}^{N} S_{MF}[w_c] + S_{MF}[f]$$

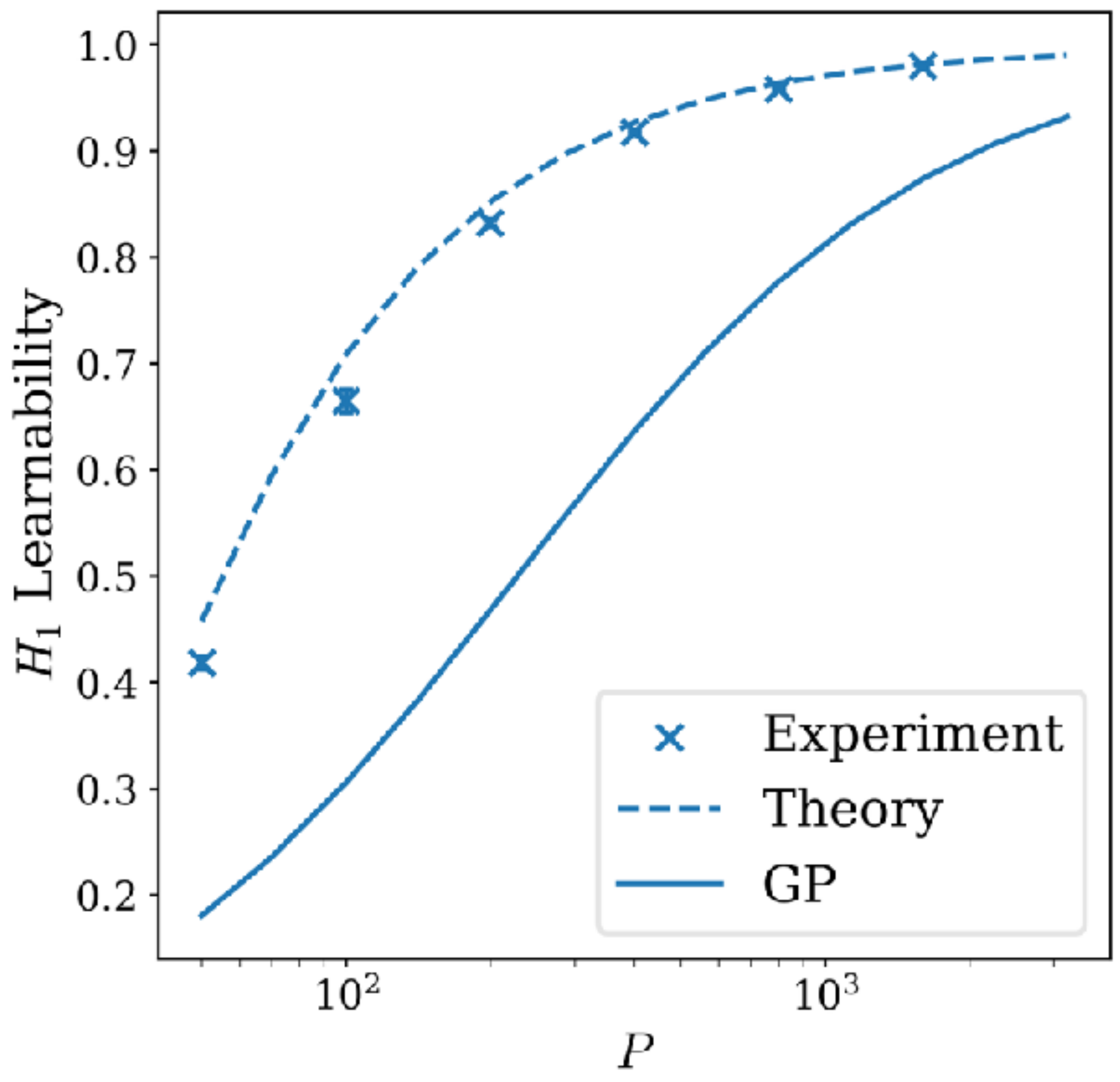
$$S_{MF}(w) = \frac{d|w|^2}{2\sigma_w^2} + \frac{\sigma_a^2}{2N} \int\int \langle t(x)\rangle_{MF} \langle t(y)\rangle_{MF} \phi(w^T x)\phi(w^T y) \overset{\text{GFL}}{\approx}_{d,n\to\infty} \frac{1}{2} w^T \Sigma_{MF}^{-1} w$$

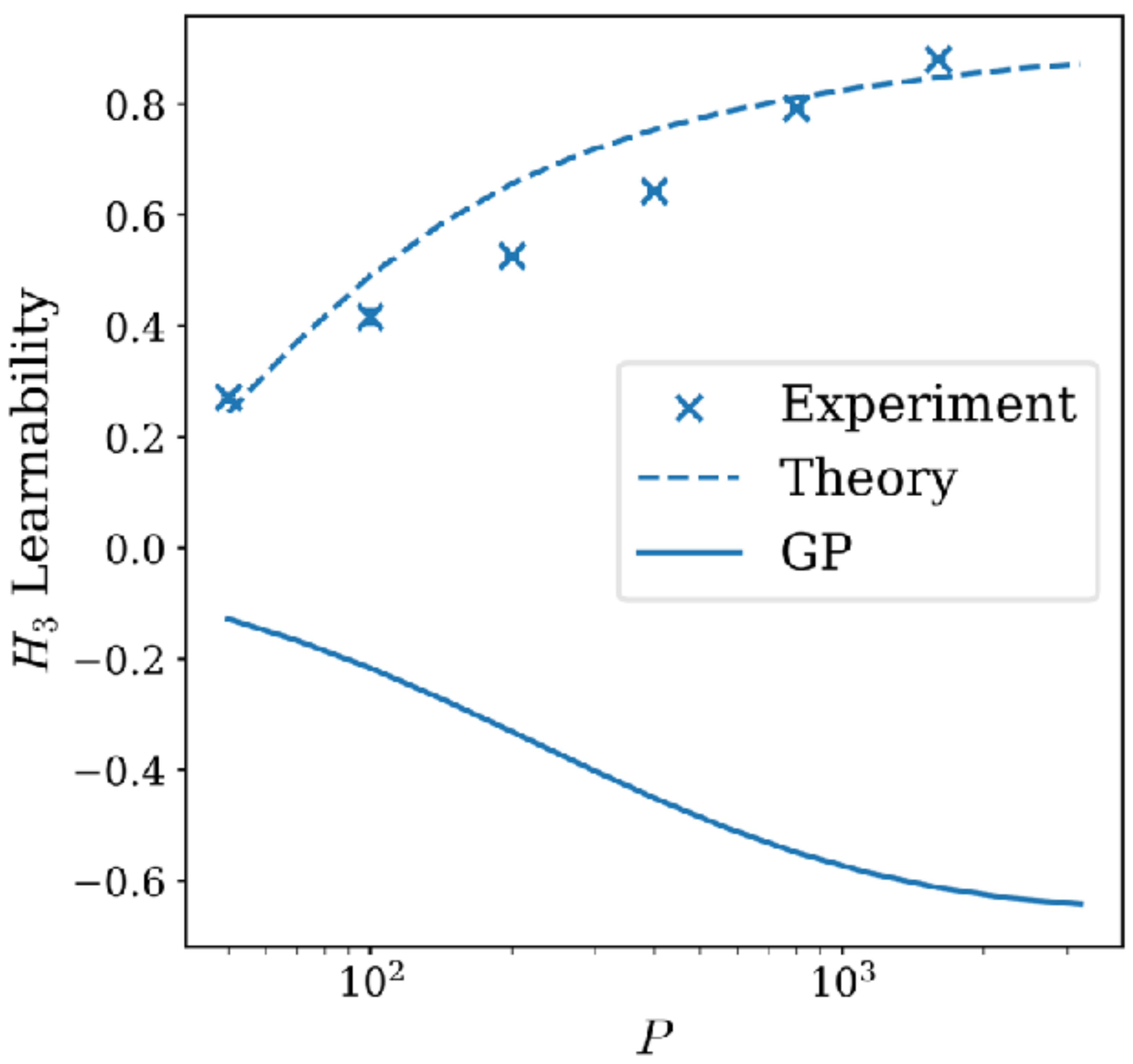
$$S_{MF}(f) = \frac{1}{2} \int\int f K_{MF}^{-1} f + L[f] \qquad \langle t \rangle_{MF} = [K_{MF} + \sigma^2/n]^{-1} y \qquad K_{MF}(x, y) = \langle \phi(w^T x)\phi(w^T y)\rangle_{w\sim\mathcal{N}(\Sigma_{MF})}$$

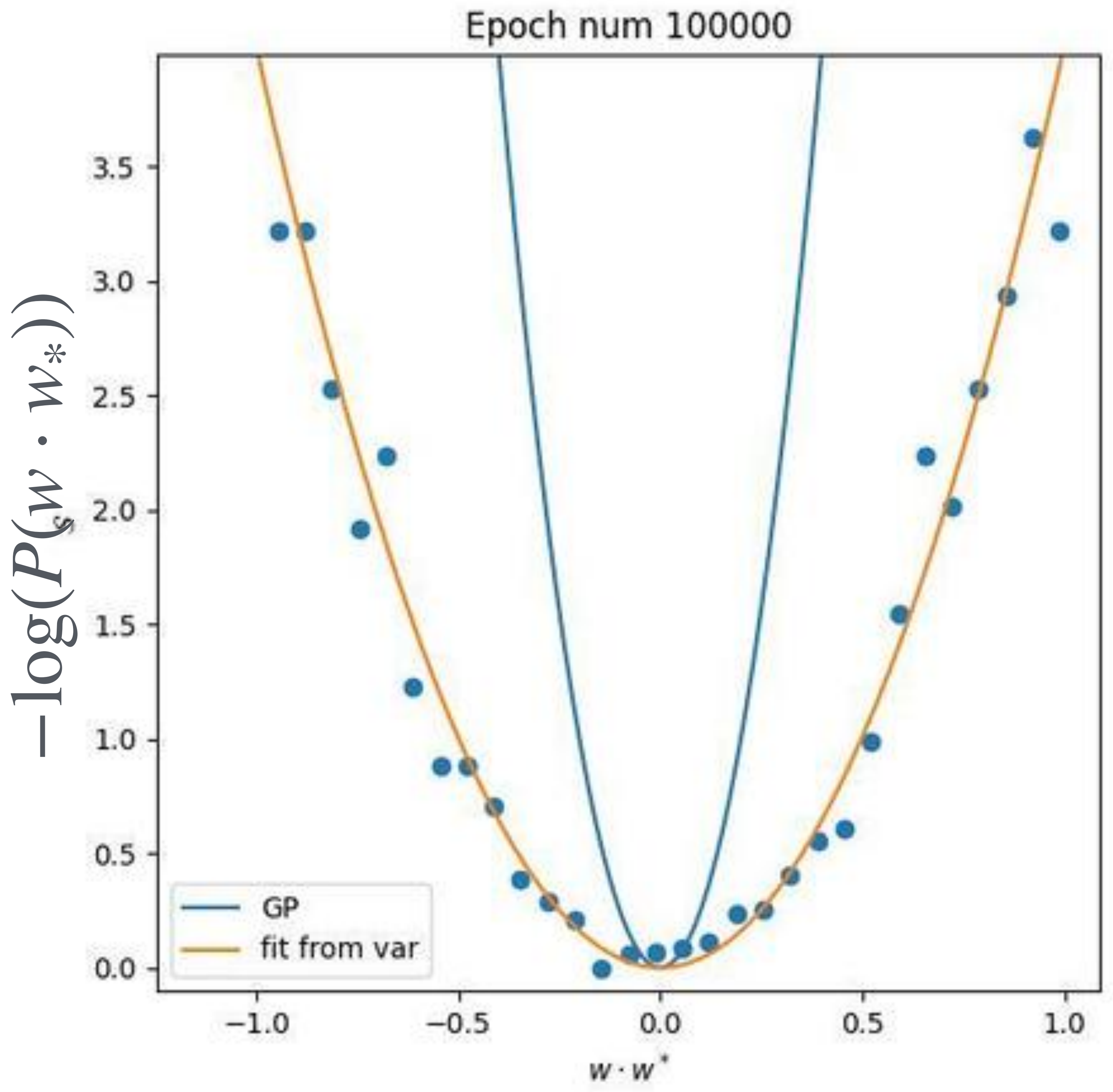
For $y(x) = y(w_*^T x)$ $\quad \mu_x = \mu_{gx} \quad g \in O(d)$ then solution must obey $\Sigma = aP_\perp + bw_* w_*^T$

We Thus obtain a non-linear equation in only two variables [a,b].

Experimental results

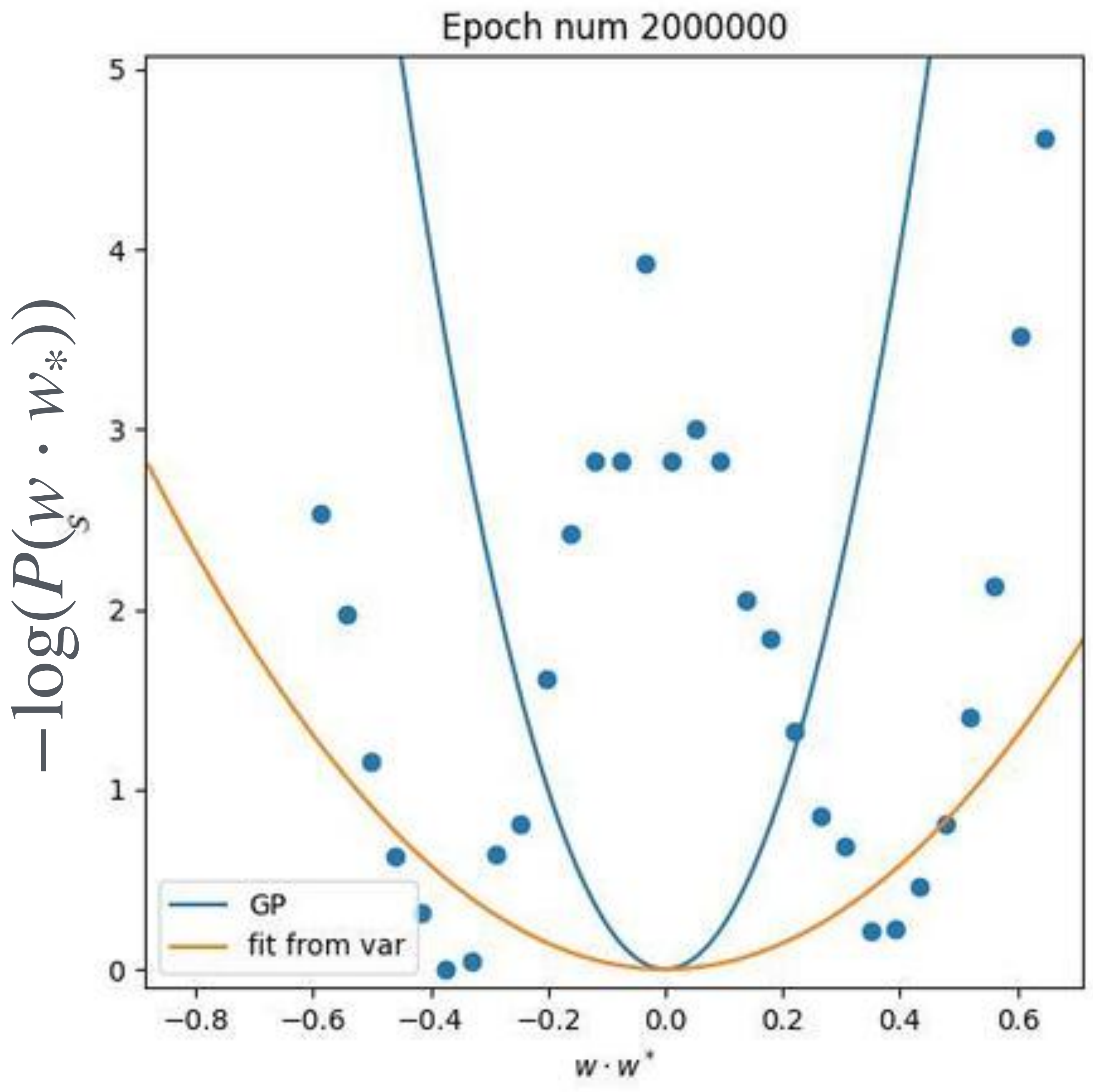






d=50 | Mean-field Scaling | N=Width=1000 | $y(x) = H_1(w_* \cdot x) - 0.05H_3(w_* \cdot x)$

Experimental results - Proximity to a phase transition

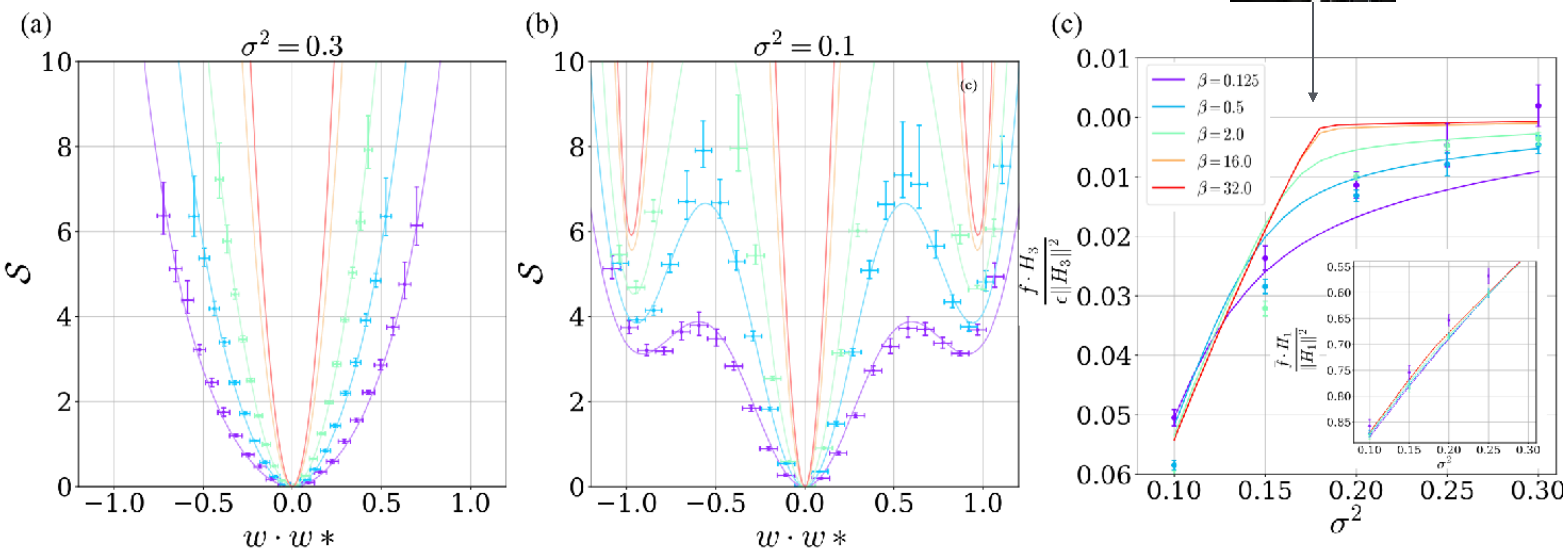


d=50 | Mean-field Scaling | N=Width=1000 | $y(x) = H_1(w_* \cdot x) + 0.05 H_3(w_* \cdot x)$

Phase transition can be 1st order - Related to Grokking

GFL/VGA phase

GMFL/Coexistence phase



Rubin, Seroussi, Ringel (ICML 2023)

Applications for CNNs (actually simpler)

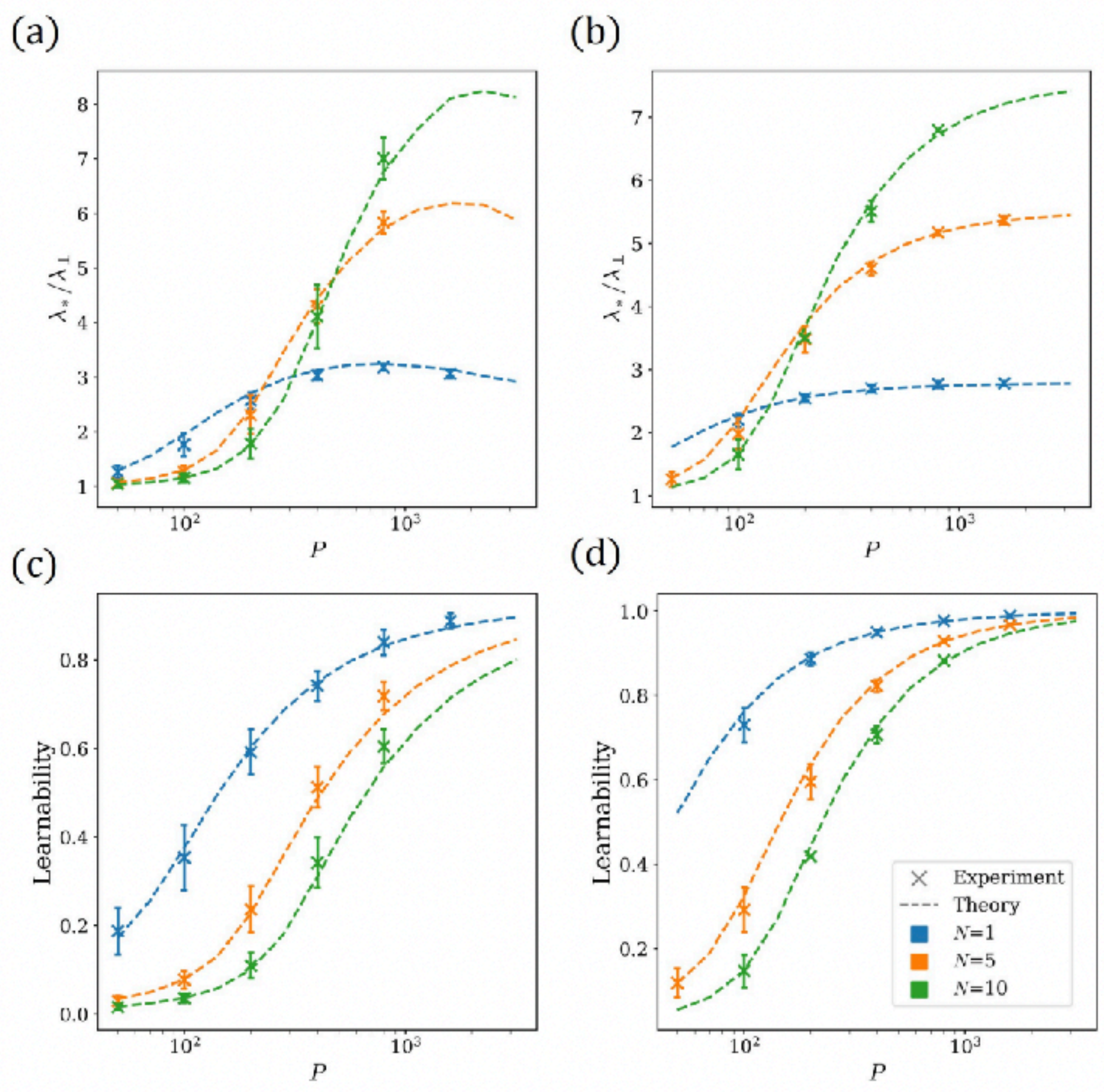


Figure 3.3: In this figure we compare a linear network trained on a single index linear teacher, with an Erf network trained on a cubic single index teacher ($y(x) = w_* \cdot x + 0.1 H_3(w_* \cdot x)$, where H_3 is the third Hermite polynomial). The ratio between the teacher direction eigenvalue of the kernel to the eigenvalues corresponding to orthogonal directions for the Erf and linear networks is shown in panels (a) and (b) respectively. In panels (c), (d) the learnability ($f \cdot y / y \cdot y$) is shown for the Erf and linear network respectively. Network parameters: $\chi = 100$, $N = 1, 5, 10$, $S = 50$, $C = 1000$.

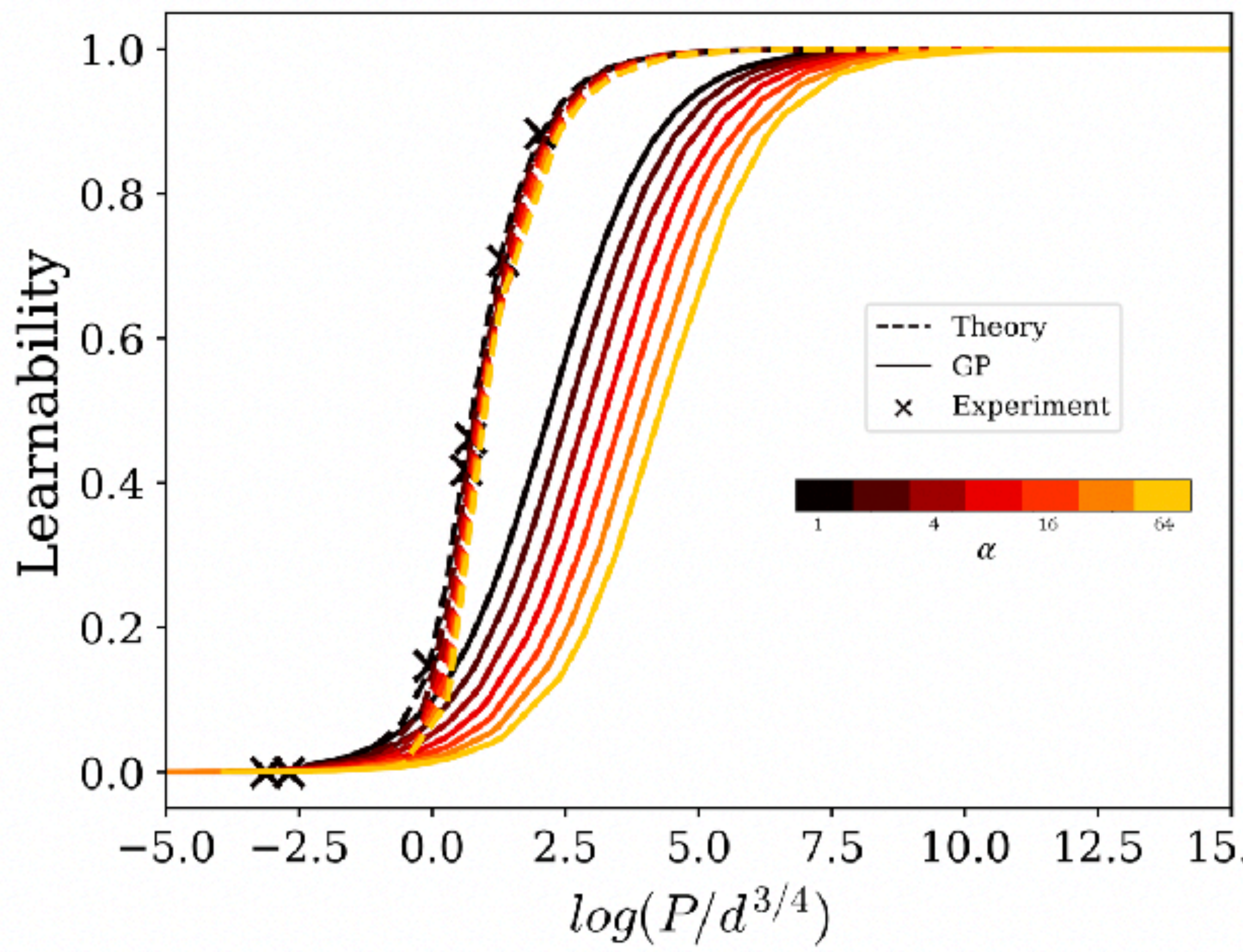


Figure 3.2: Learnability of linear CNNs as a function of P. We take $S, N, C \propto \alpha$, and consider different α scales of these parameters. Here the network is observed to learn the target at $P \propto d^{3/4}$, regardless of the parameter scale, as opposed to the GP predictions which predict learning at $P \propto d$. Parameters: $\chi = 100$, $N = 10\alpha, S = 50\alpha, C = 1000\alpha$.

Applications of Statistical Field Theory in Deep Learning

Zohar Ringel, Noa Rubin, Edo Mor, Moritz Helias, Inbar Seroussi

https://arxiv.org/abs/2502.18553

Feature Learning - Open questions

- Equilibrium Phase diagram of feature learning [GFL, GMFL-I,GMFL-II, Specialization,Feature Compression]
- Feature learning in the Neural Scaling Laws regime.
- Developing techniques for approaching the interpolation threshold [e.g. Relating SAE with Kernel Adaptation]
- Limitations, implicit biases, and overfitting of feature learning

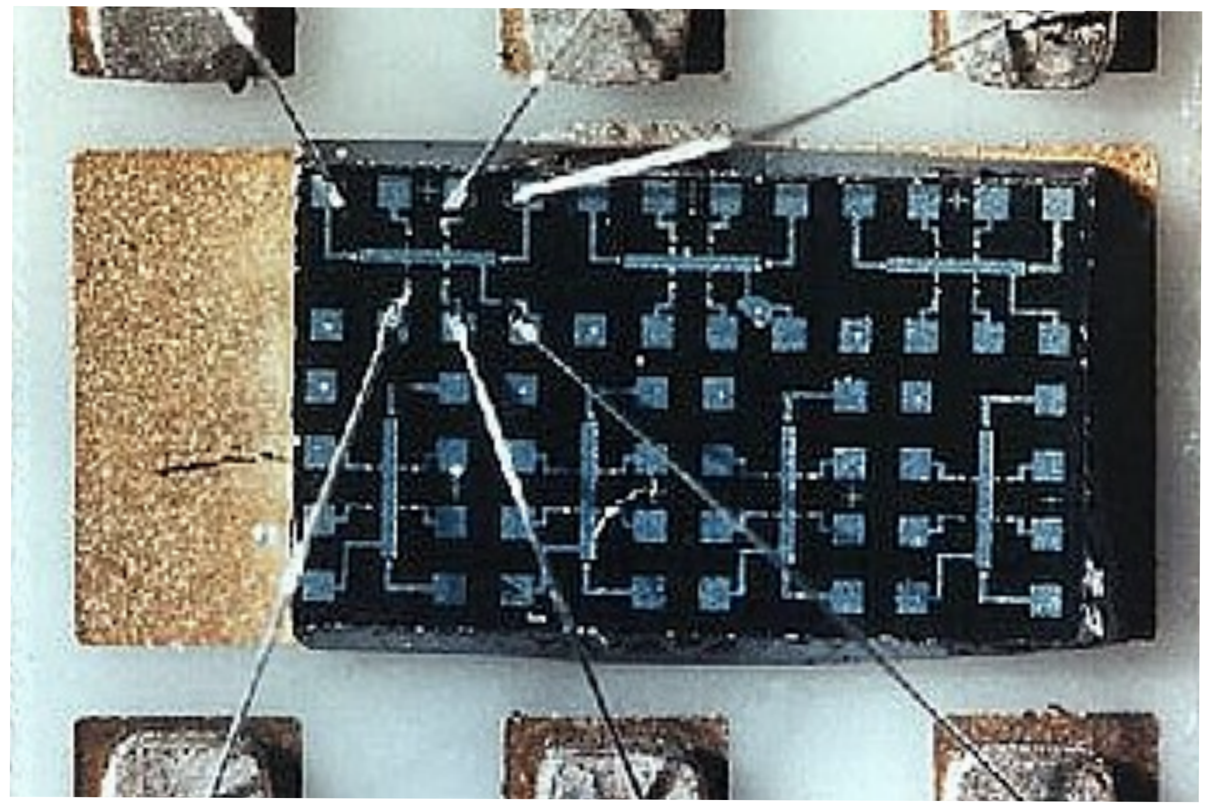
Now to the bad news: Explainability paradox

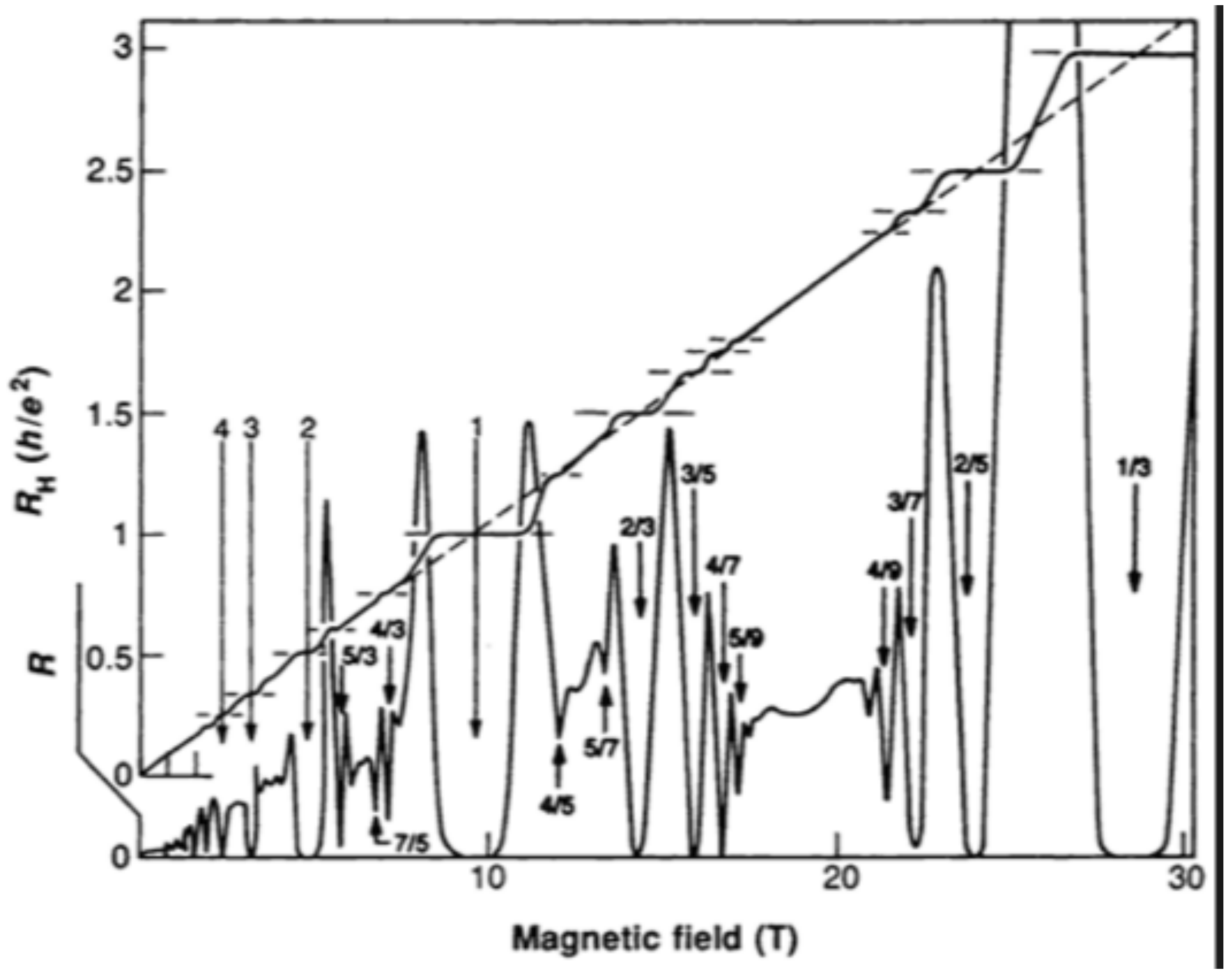
Or how far can we take such analytic analysis?

- Assume we found a set of analytically solvable equations describing the outputs of a DNN trained on a specific data-set
- Analytically solvable means that the complexity of inferring predictions from these equations is O(1).
- Instead of training the DNN, we may just solve the equations.
- We thus found a simple O(1) training algorithm for this DNN. We also obtained a good classifier analytically. Both are highly unlikely....

In contrast in physics,

Quantization of the hall conductance





$$\sigma_h = n\frac{e^2}{h} \quad n \in \mathrm{N}$$

1998

Theory's main hope here: Universality/Irrelevance

- Toy models may capture a greater truth
- Some aspects of DNNs (e.g. initialization) might decouple (Modularity)
- Dimensionality reduction - The effective number of hyper-parameters and parameters may be much smaller than it seems.

- The 1st and 3rd items are formalized in physics via the **Renormalization Group Approach**

The Renormalization Group, Scale-Freeness, and Neural Scaling Laws

Collaborators

Moritz Helias

Ro Jefferson

Jessica Howard

Anindita Maiti

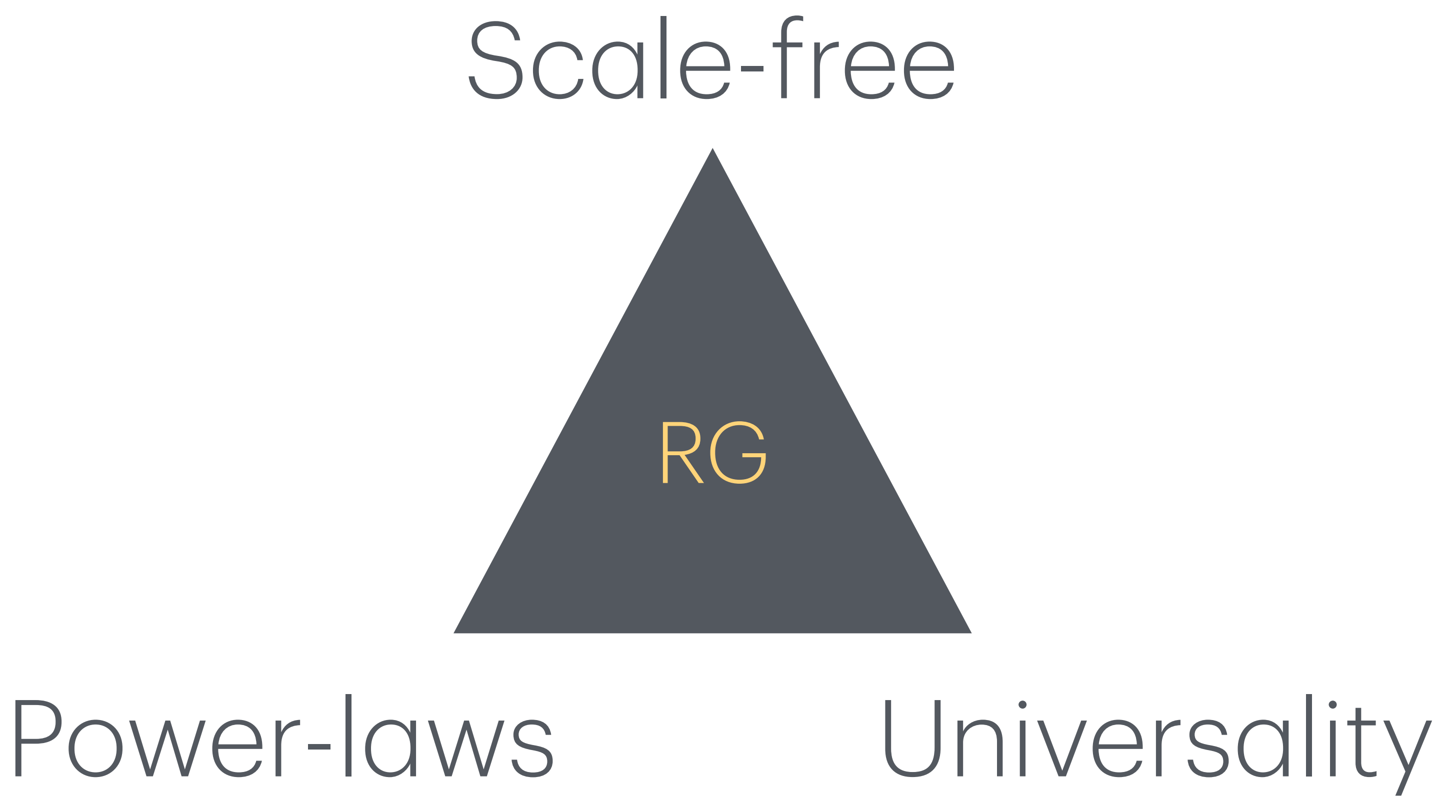
Gorka P. Coppola

Learning curves for deep neural network a gaussian field theory approach **Cohen, Malka, ZR** (2019)

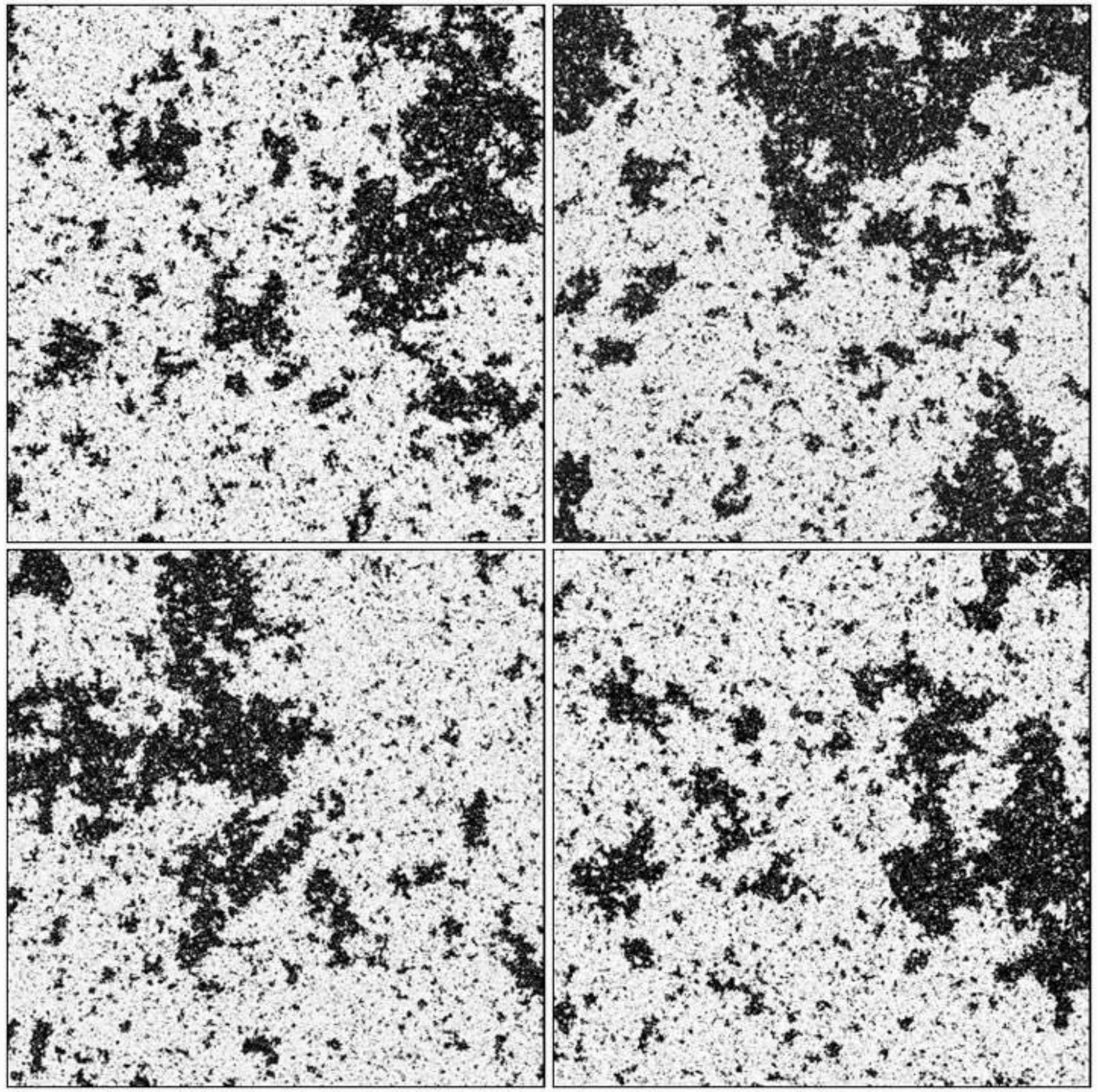
Wilsonian Renormalization of Neural Network Gaussian Processes **Howard, Maiti, Jefferson, ZR** (2024)

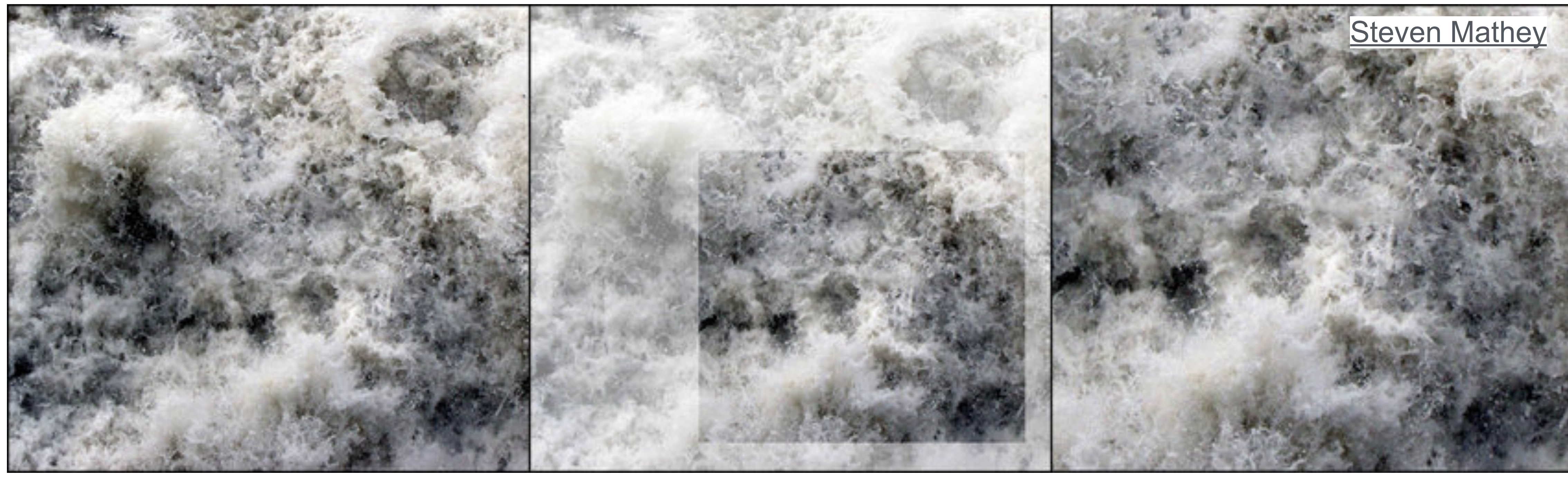
,Universality and finite-data effects in deep neural networks trained on scale-free data. **Coppola, Helias, Ringel** (TBP)

The Natural World

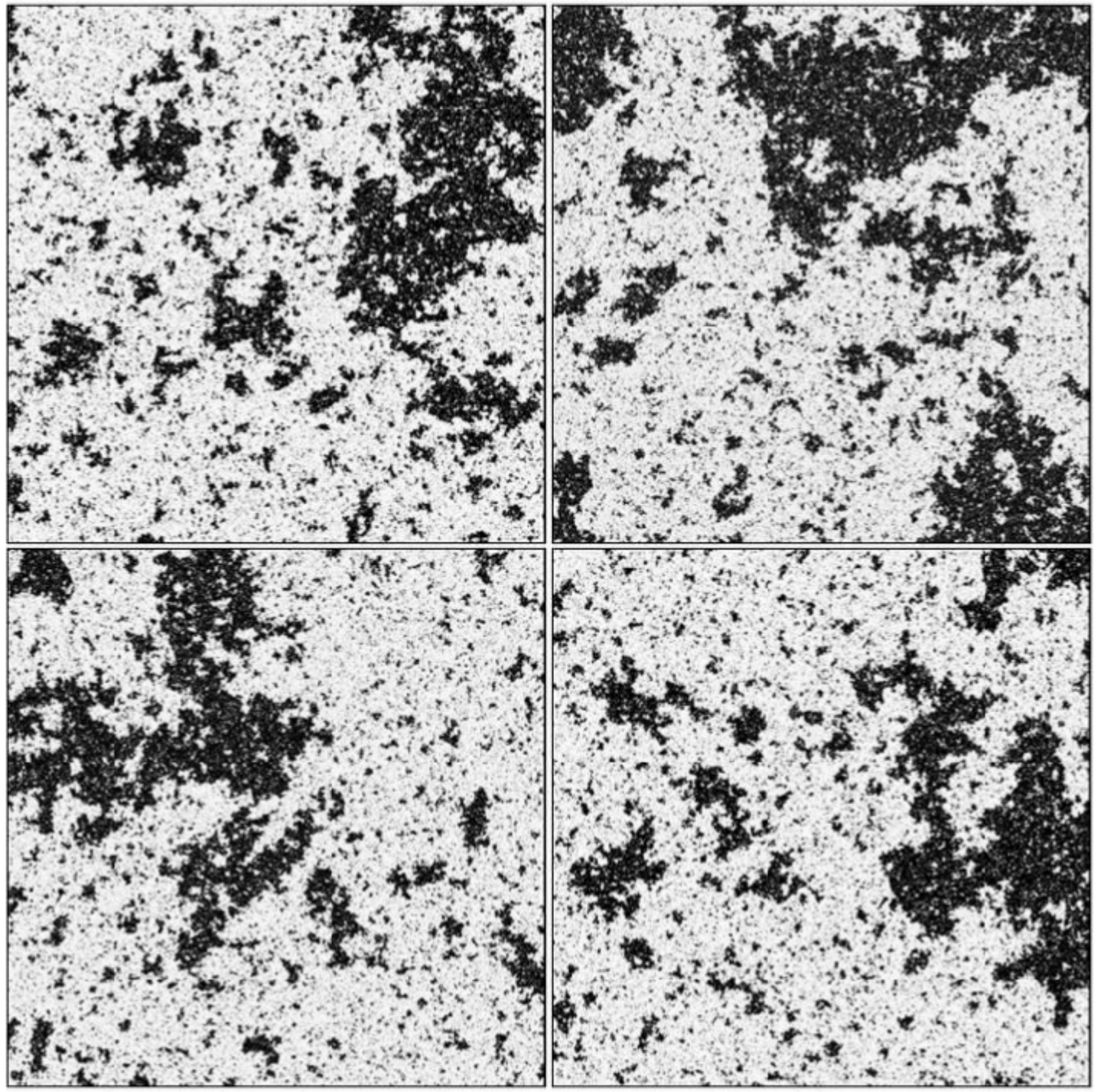


Scale-free Phenomena





Power laws



Critical 2d Ising Model

$$\langle f(0) f(x) \rangle \propto \frac{1}{x^{\eta}}$$

$$\langle f_k f_{-k} \rangle \propto \frac{1}{k^{2-\eta}}$$

Universality

Ising Model

Ising Model + All local symmetry respecting perturbations

Liquid-Gas at the tri-critical point

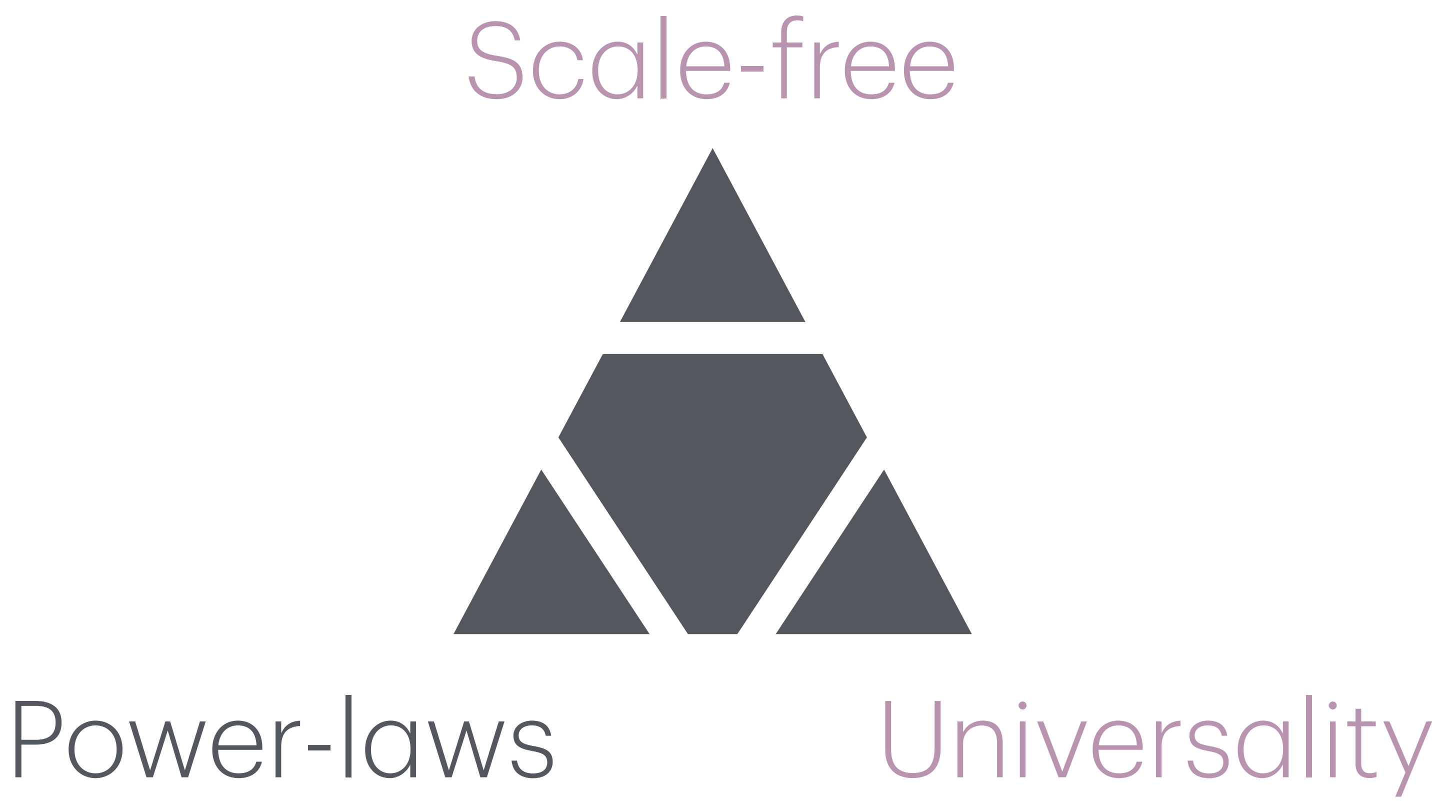
Quantum systems at d-1 with a Z2 symmetry

...

$$\langle f(0)f(x)\rangle = C\frac{1}{x^\eta}$$

$$\eta = 1/4 \quad C = O(1)$$

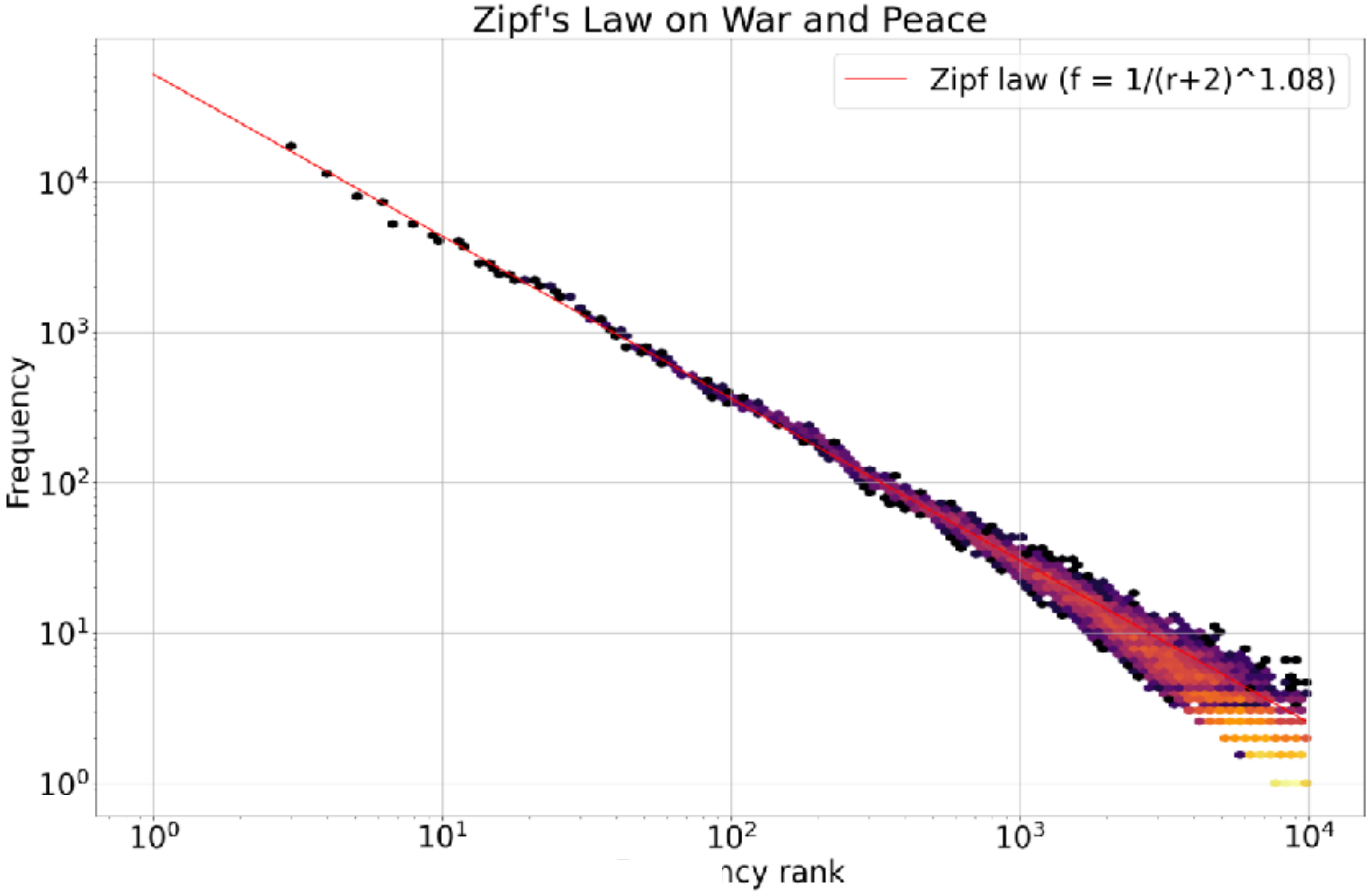
Scale-free phenomena in ML

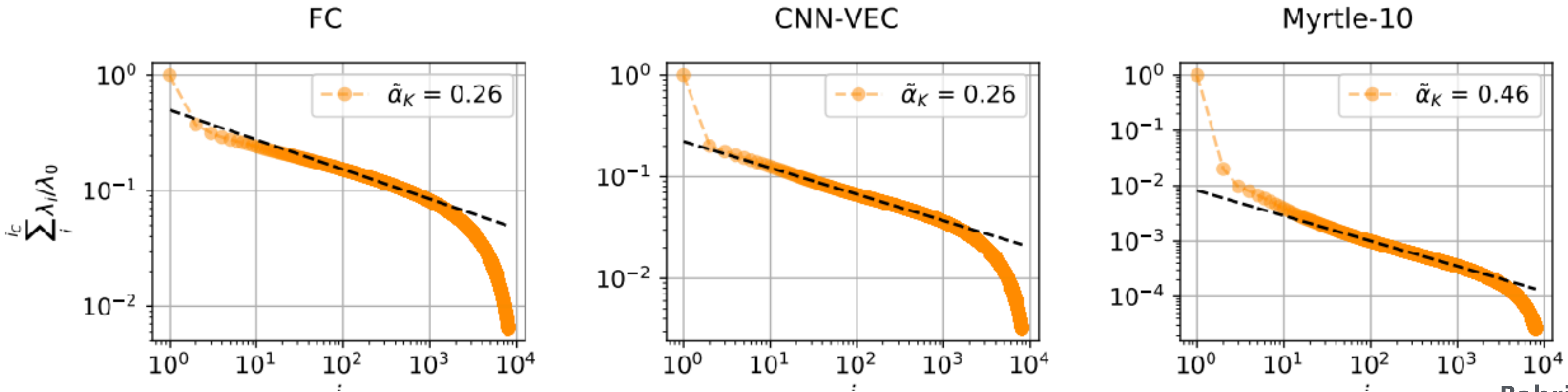


Power laws in PCA and kernel-PCA

$$\text{frequency} \propto \frac{1}{(\text{rank} + b)^a}$$

where a, b are fitted parameters, with $a \approx 1$, and $b \approx 2.7$.[1]





Bahri et. al., Explaining Neural Scaling Laws, 2021

Neural Scaling Laws (Power laws in learning curves)

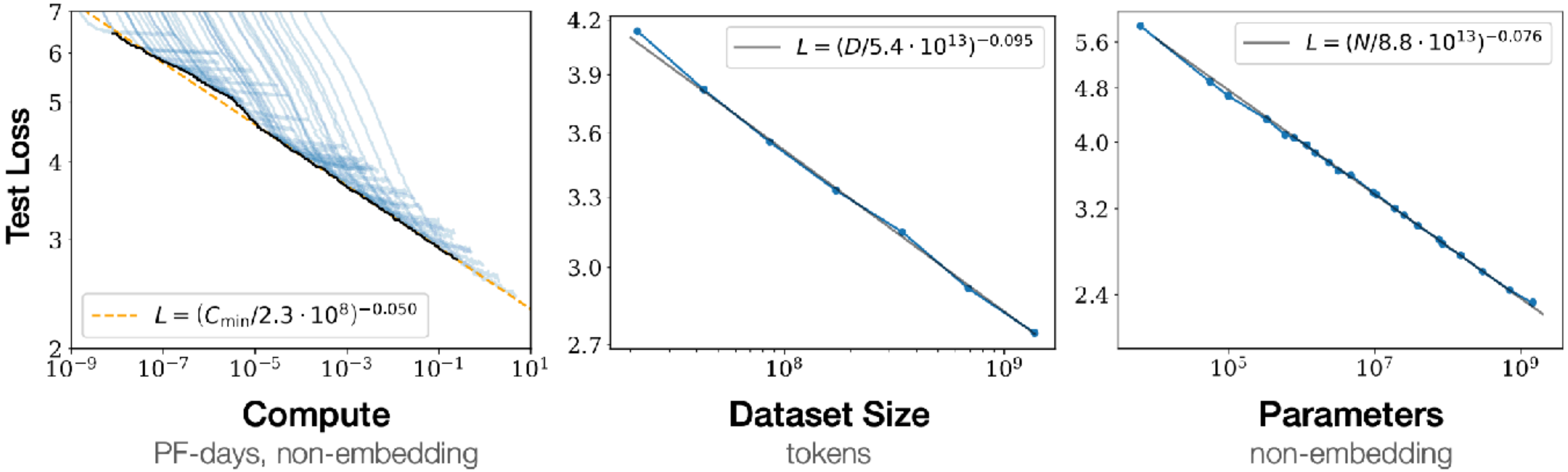
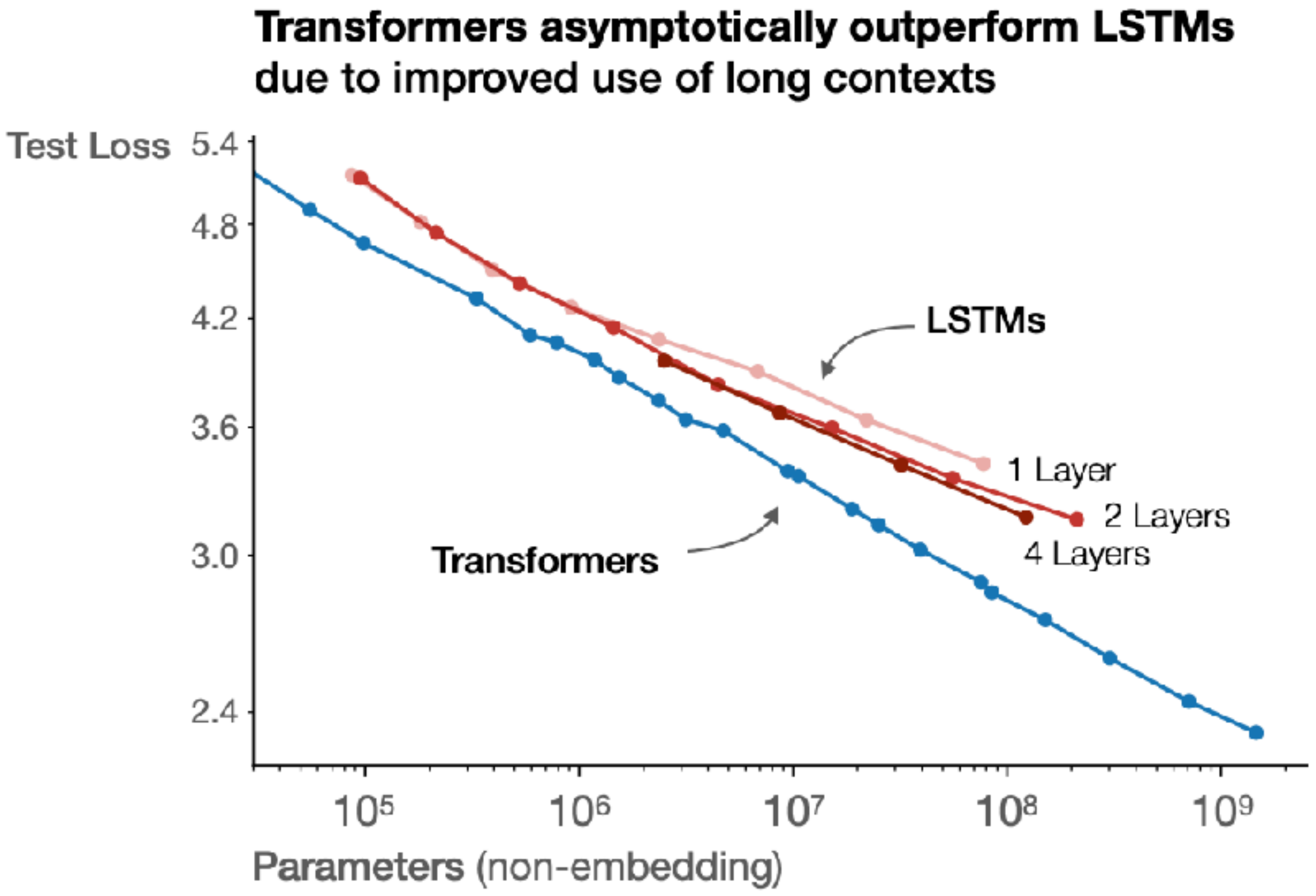


Figure 1 Language modeling performance improves smoothly as we increase the model size, datasetset size, and amount of compute[2] used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.

Kaplan et. al. Scaling Laws for Neural Language Models (2020)

Universality - neural networks

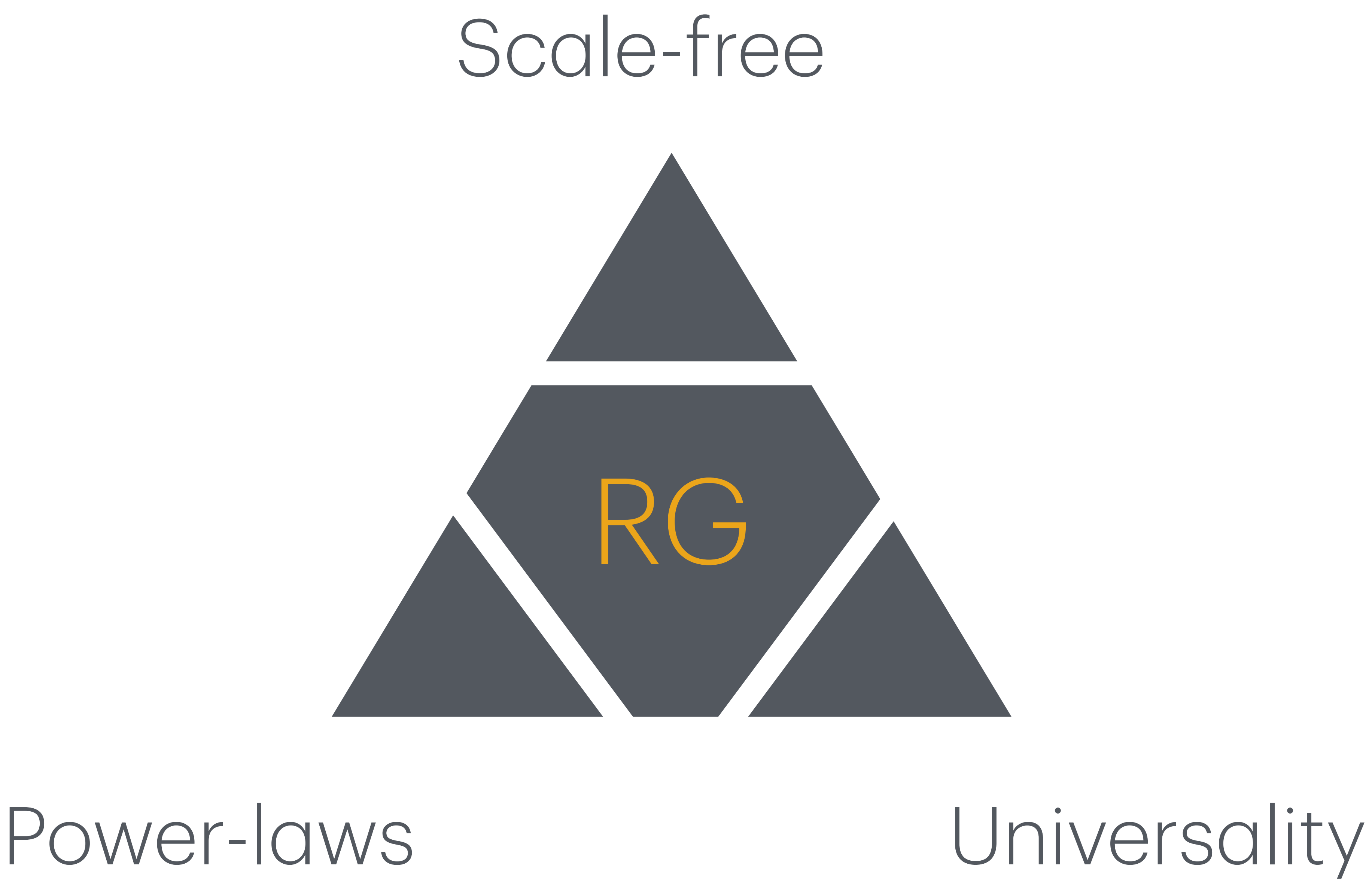
- The fact that many changes to training and hyper-parameters matter only in a few percent and changes to architecture drive most innovation.
- Hyper-parameter transfer protocols
- Similar scaling curves for LSTMs and Transformers



Kaplan et. al. Scaling Laws for Neural Language Models (2020)

The Big Questions

- **Is there an input ``scale" associated with power-laws in data and learning?**
- **Are power-law indicative of universality?**
- **If there is universality, what is the minimal model?**
- **How can we adapt RG from physics to ML?**



The Renormalization Group Approach 101

Wilsonian RG - As a Greedy Algorithm

- **General Setting**: We wish to compute some average under some complicated probability
- $P(f_1 ..f_k ..f_\Lambda) \propto \exp\left(-S(f_1 ..f_k ..f_\Lambda)\right)$ where f_k are Fourier Modes of some field
- **This is hard,** so we focus observables which are function of $k < k_{Observation}$ ("low energy"/ "IR" sector) and that modes with $k > k_{Observation}$ are weakly coupled to the rest ("high energy"/"UV" sector).
- **We gradually remove large k modes (Decimation)**
  $$P_{\Lambda-1}(f_1 ..f_k ..f_\Lambda) \propto \int df_\Lambda \exp\left(-S(f_1 ..f_k ..f_\Lambda)\right) \equiv \exp\left(-S_{\Lambda-1}(f_1 ..f_k ..f_{\Lambda-1})\right)$$
- **We then rescale the k/wave-number/momentum "index" (Rescale space)**

Wilsonian RG - The Flow

- **We found a mapping between the S and S′ actions**

$$P_{\Lambda-1}(f_1..f_k..f_\Lambda) \propto \int df_\Lambda \exp\left(-S(f_1..f_k..f_\Lambda)\right) := \exp\left(-S_{\Lambda-1}(f_1..f_k..f_{\Lambda-1})\right) := \exp\left(-S'(\tilde{f}_1..\tilde{f}_k..\tilde{f}_\Lambda)\right)$$

- **Since we integrate out only a single mode which is weakly coupled to the rest S should be very similar to S′.**

- **This can be phrased as the following functional diffusion-like equation (Polchinski's equation)**

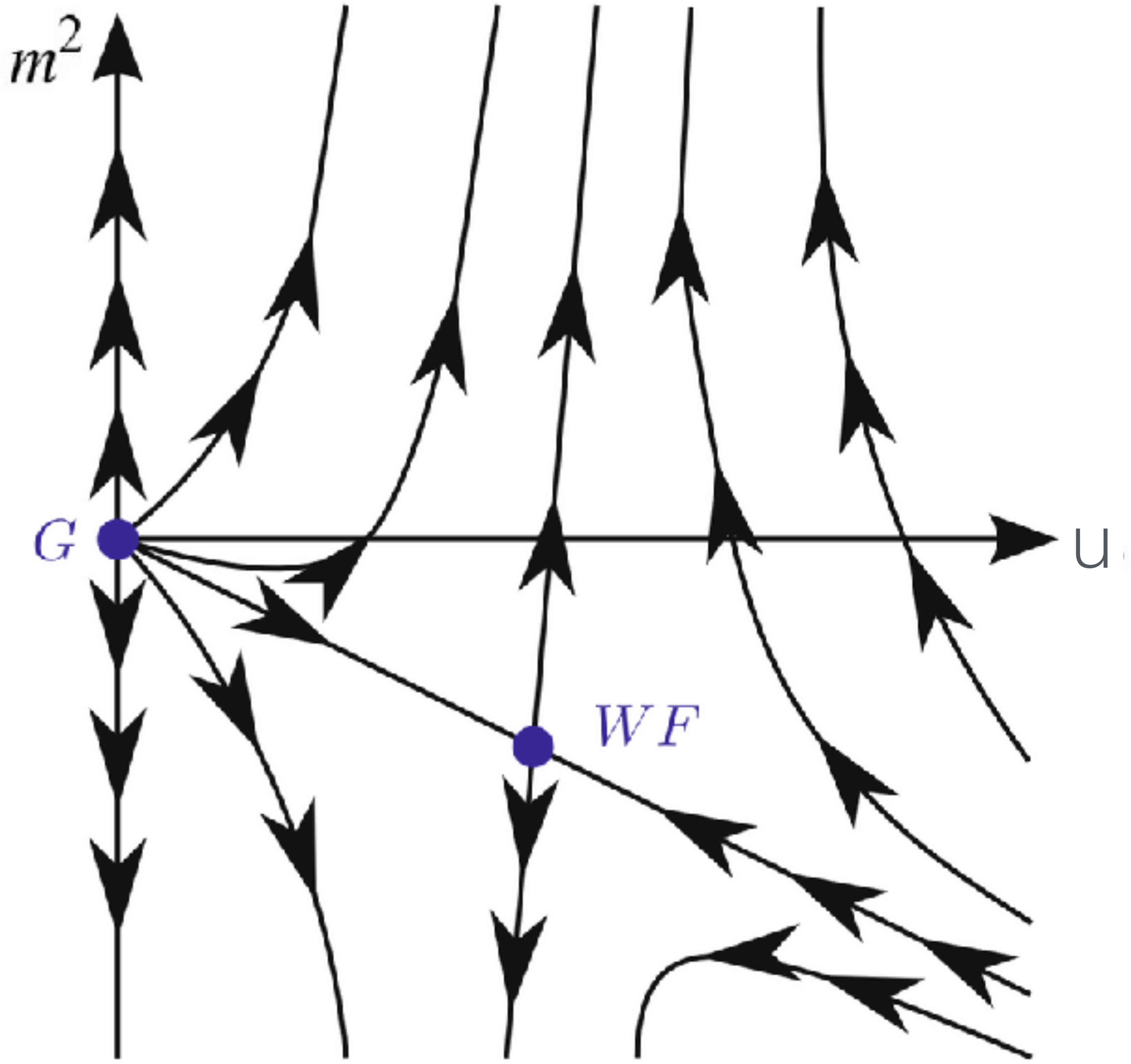
- $$\frac{dS(f_1..f_k..f_\Lambda)}{d\Lambda} = L[S]$$

- **In many physically relevant cases, this functional equation simplifies to a finite set of non-linear ODEs.**

Wilsonian RG - The Ising Model and Universality

$$S[f(x)] = \int dx(\nabla f(x))^2 + m^2 f^2(x) + u\int dx f(x)^4 = \int dk[k^2 + m^2] f_k^2 + u\int dx f(x)^4$$

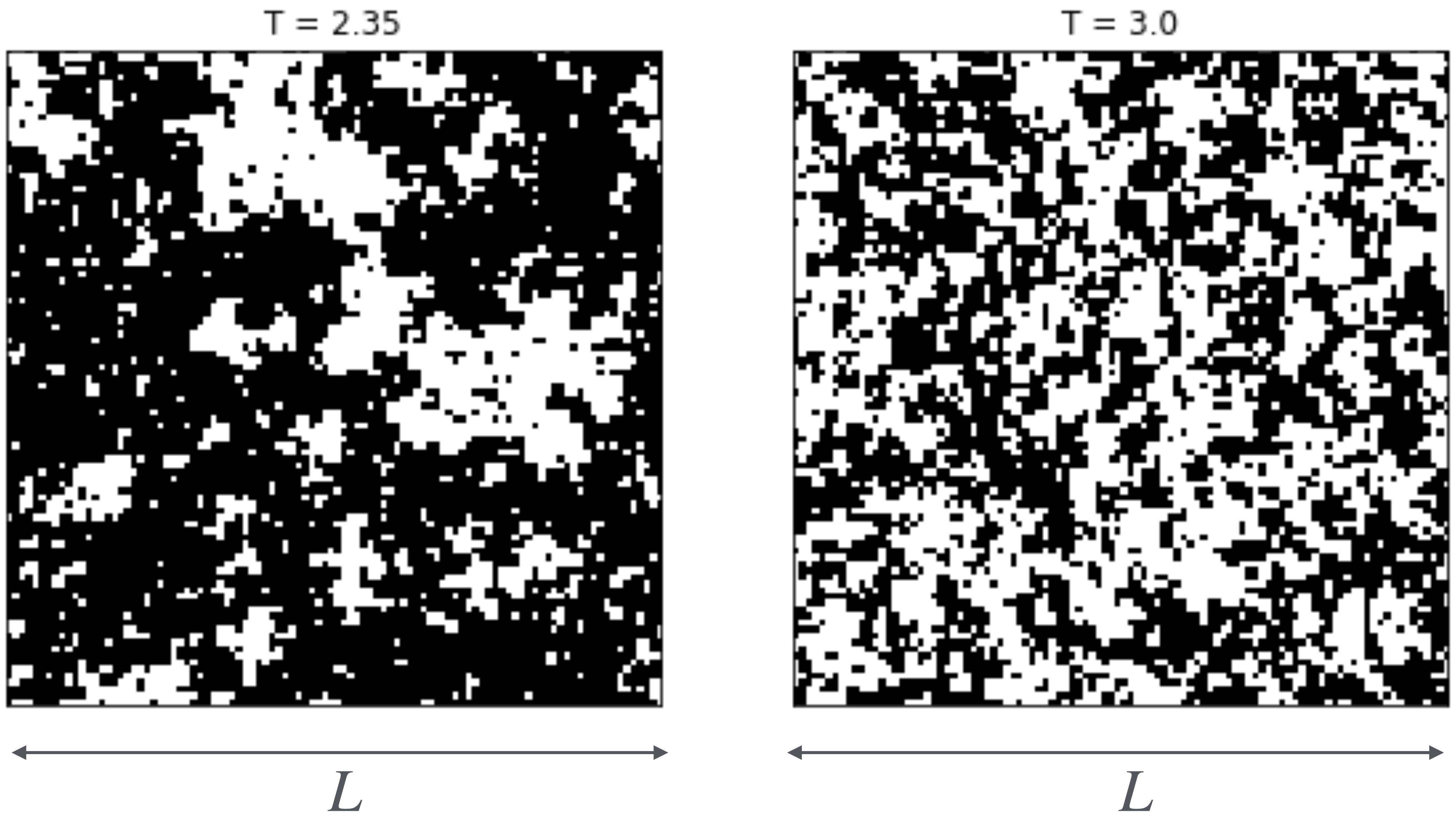
$$+v\int dx f^6(x)$$

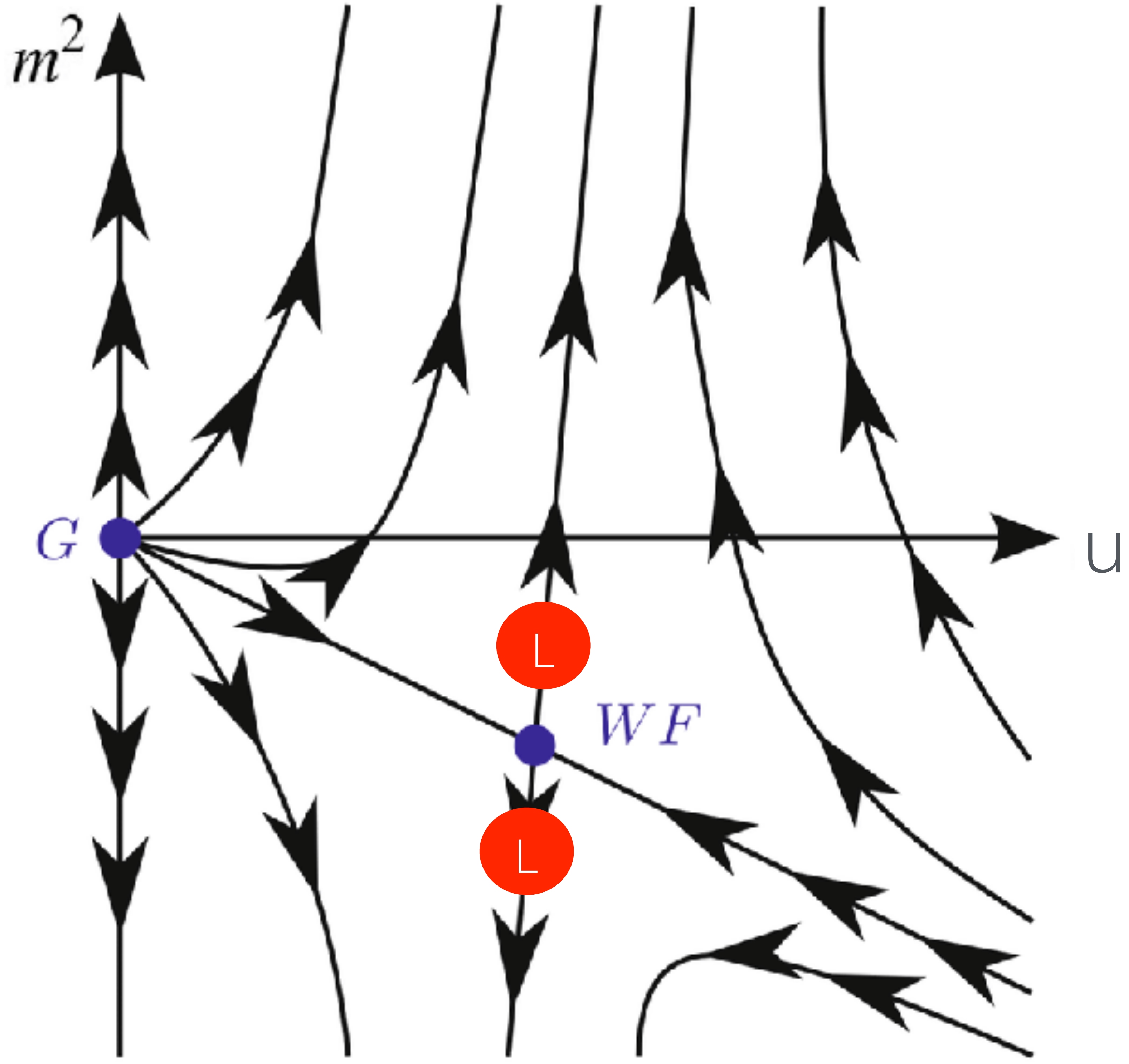


Wilsonian RG - finite size effects

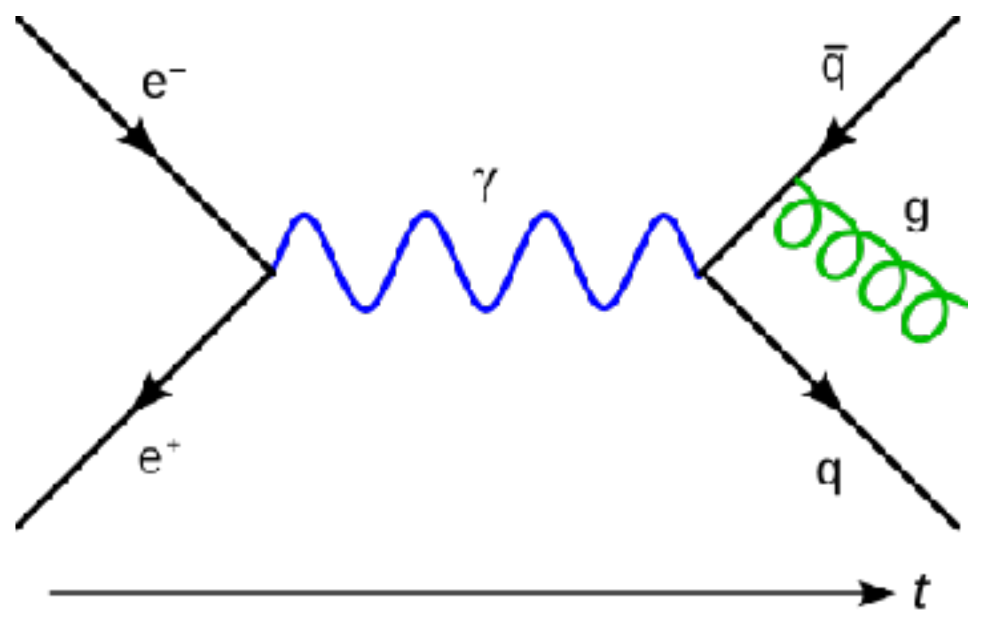
Detuning From Criticality = Finite Size System (L)

$$|T_c - T|^{-\nu} \propto L$$





Applying RG to Learning:
Step (1) deep Learning As a Field Theory



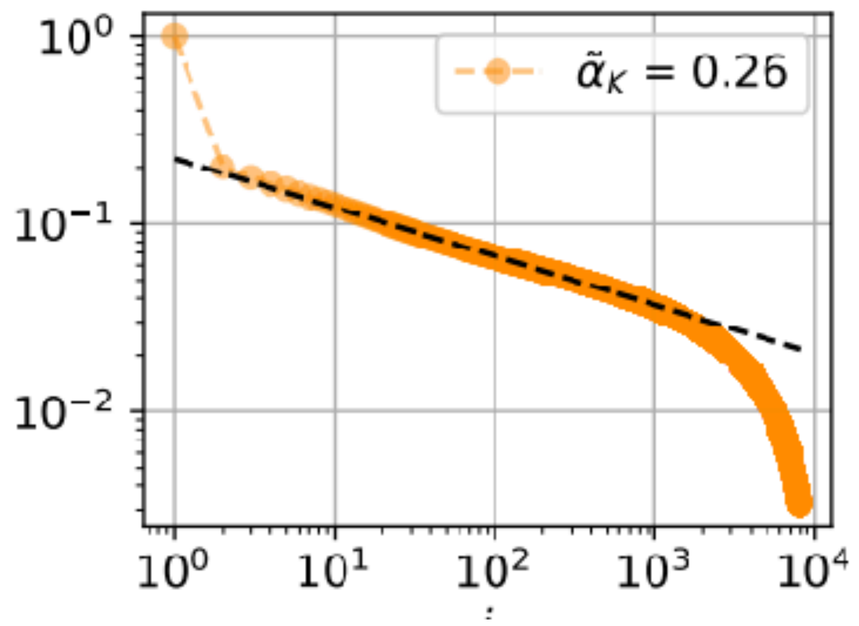
Let's start modestly, from the GP limit

$$S = \frac{1}{2}\int d\mu_x d\mu_y f(x) K^{-1}(x,y) f(y) - P\int d\mu_x e^{-[f(x)-y(x)]^2/T}$$

$$S_{GP}[f(x)] = \sum_k \lambda_k^{-1} f_k^2 - P\int d\mu_x e^{-\frac{(f(x)-y(x))^2}{2\kappa^2}}$$

- For real world data we often have

$$S_{GP}[f(x)] \approx \sum_k k^{\alpha} f_k^2 - P\int d\mu_x e^{-\frac{(f(x)-y(x))^2}{2\kappa^2}}$$



Bahri et. al., Explaining Neural Scaling Laws, 2021

GPR Field Theory Compared to Ising Field Theory

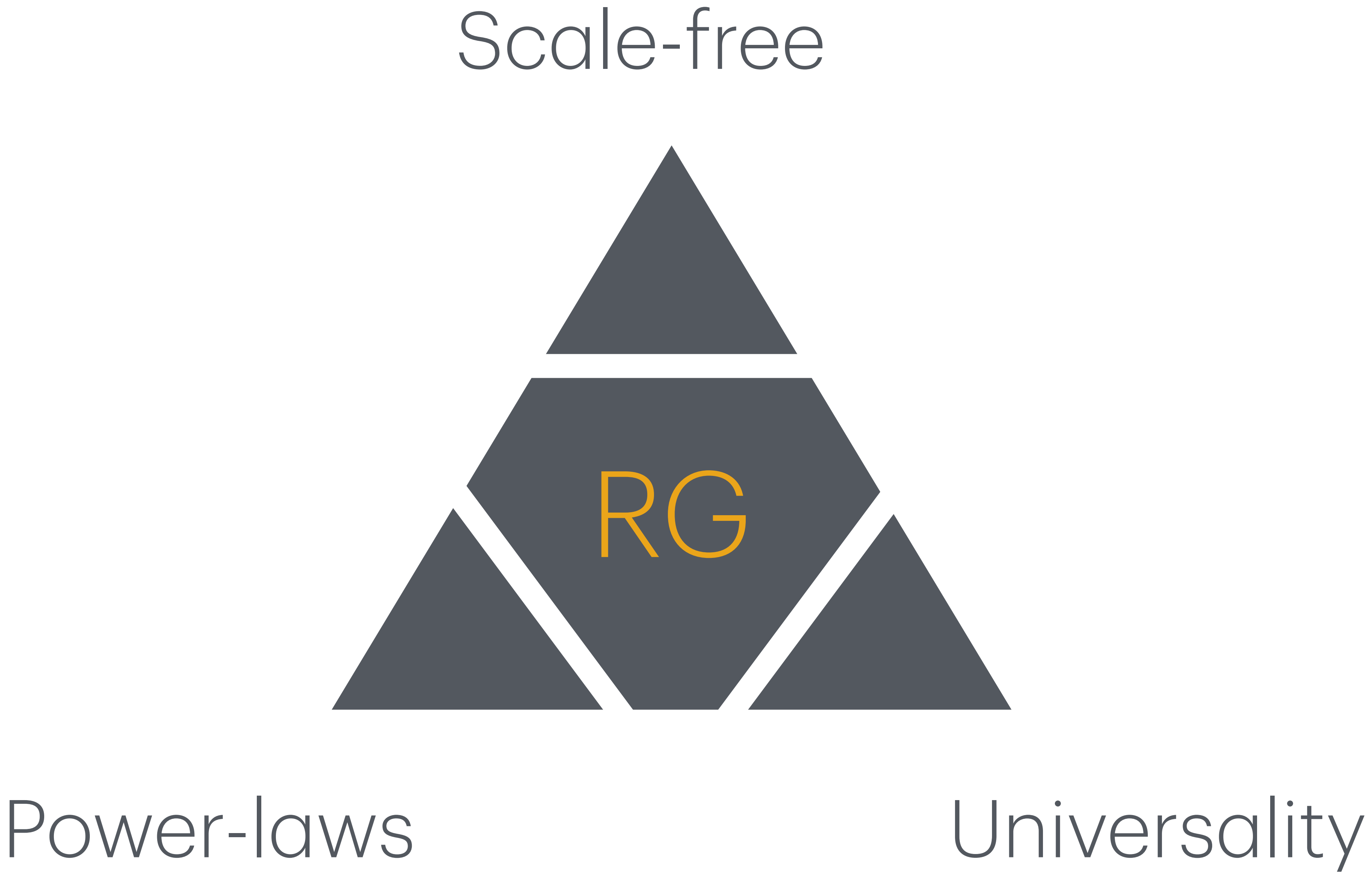
$$S_{GP}[f(x)] = \sum_k k^{\alpha} f_k^2 + \frac{P}{T} \int d\mu_x \left[\frac{(f(x) - y(x))^2}{2T} \right]$$

$$S_{Ising}[f(x)] = \int dk[k^2 + m^2] f_k^2 + u \int dx f(x)^4$$

- Different power of k —> Let's take $\alpha = 2$
- Different local interaction —> No biggy
- Discrete summation over k instead of an integral —> Like what happens in a finite system.
- No translation invariance, k is not momentum, first term is non-local —> A bit scary...

The ~~Big~~ Concrete Questions

- **Can we track the RG flow of** $$S_{GP}[f(x)] = \sum_k [k^\alpha + \frac{P}{T}] f_k^2 - P\left[\int d\mu_x e^{-\frac{(f(x)-y(x))^2}{2T}} + \frac{(f(x)-y(x))^2}{2T}\right]$$
- **How many coupling constants are required to describe the flow?**
- **How does the RG flow relate to the learning curve?**
- **Is there some form of universality?**
- **Are there RG fixed points?**



Decimation only RG (EFT style RG)

Howard, Maiti, Jefferson, ZR (2024)

Wilsonian RG - A Greedy Algorithm

- **General Setting**: We wish to compute some average under some complicated probability
- $P(f_1 \,..f_k \,..f_\Lambda) \propto \exp\left(-S(f_1 \,..f_k \,..f_\Lambda)\right)$ where f_k are Fourier Modes of some field
- **This is hard.** So we focus observables which are function of $k < k_{Observation}$ ("low energy"/ "IR" sector) and that modes with $k > k_{Observation}$ are weakly coupled to the rest ("high energy"/"UV" sector).
- **We gradually remove large k modes (Decimation)**
$$P_{\Lambda-1}(f_1 \,..f_k \,..f_\Lambda) \propto \int df_\Lambda \exp\left(-S(f_1 \,..f_k \,..f_\Lambda)\right) := \exp\left(-S_{\Lambda-1}(f_1 \,..f_k \,..f_{\Lambda-1})\right)$$
- ~~**We then rescale the k momentum "index" (Rescale space)**~~

Decimation-only RG - Analytical Results

- Generically, at large input dimension

$$S_{GP}[f(x)] = \sum_{k=1}^{\Lambda} \lambda_k^{-1} f_k^2 - P \int d\mu_x e^{-\frac{(f(x)-y(x))^2}{2T}}$$

$$\Rightarrow_{RG} \sum_{k=1}^{\Lambda'} \lambda_k^{-1} f_k^2 - P \int d\mu_x e^{-\frac{(f(x)-y(x))^2}{2T(\Lambda')}}$$

$$\approx_{\Lambda' \ll \Lambda} \sum_{k=1}^{\Lambda'} \lambda_k^{-1} f_k^2 + P \int d\mu_x \frac{(f(x)-y(x))^2}{2T(\Lambda')}$$

$$T(\Lambda') = T + \sum_{\Lambda'}^{\Lambda} \lambda_k$$

Intuitively: unlearnable modes look like observation-noise/regulator

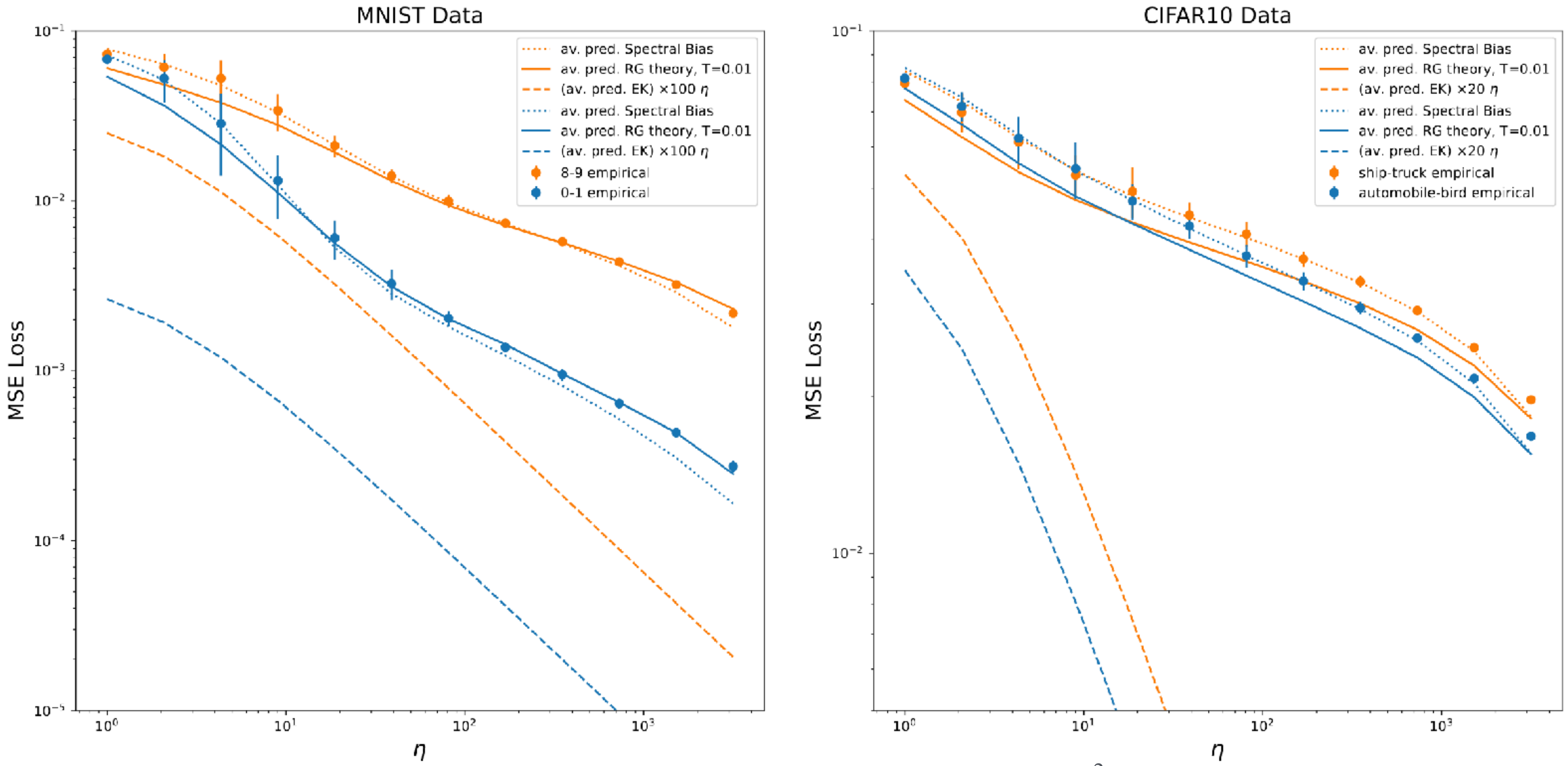
[reminiscent of Bartlett et. al. Benign Overfitting... (2020)]

Decimation only RG - gained insights

- A single parameter ($T(\Lambda')$) tracks the flow!
- Integrating out leads to a more Gaussian theory (weak coupling regime)
- Modes with $\lambda_k^{-1} = k^\alpha \ll P/T(\Lambda')$ are effectively frozen to $y(x)$ [perfectly learnable]
- We may therefore view $P^{1/\alpha}$ as setting the minimal allow k which in physics is the inverse system size ($k_{min} \propto L^{-1}$). RG Machinery for finite-size correction can then be important to finite P corrections.

$$\sum_{k=1}^{\Lambda'} \lambda_k^{-1} f_k^2 - P \int d\mu_x e^{-\frac{(f(x)-y(x))^2}{2T(\Lambda')}} \approx_{\Lambda' \ll \Lambda} \sum_{k=1}^{\Lambda'} \lambda_k^{-1} f_k^2 + P \int d\mu_x \frac{(f(x)-y(x))^2}{2T(\Lambda')}$$

Decimation-only RG - Numerical Results



$$\sum_{k=1}^{\Lambda'} \lambda_k^{-1} f_k^2 - P\int d\mu_x e^{-\frac{(f(x)-y(x))^2}{2\kappa_{\Lambda'}^2}} \approx_{\Lambda'\ll\Lambda} \sum_{k=1}^{\Lambda'} \lambda_k^{-1} f_k^2 + P\int d\mu_x \frac{(f(x)-y(x))^2}{2\kappa_{\Lambda'}^2}$$

The ~~Big~~ Concrete Questions

Revisited

- **Can we track the RG flow of** $S_{GP}[f(x)] = \sum_k k^\alpha f_k^2 + P \int d\mu_x e^{-\frac{(f(x)-y(x))^2}{2\kappa^2}}$ Yes at large input dim.
- **How many coupling constants are required to describe the flow?** One
- **Is the RG flow indicative of the learning curve?** P sets the system size. Learning curve scales like the loss of a theory with P=0 within this “system size:
- **Is there some form of universality?**
- **Are there RG fixed points?**

Full Wilsonian RG (Adding the re-scaling step)

$$f_1 \cdot \cdot f_{\Lambda-1} \to f_1 \cdot \cdot f_\Lambda \qquad\qquad k \to k\Lambda/\Lambda'$$

$$S_{GP}[f(x)] = \sum_{k=1}^{\Lambda} \lambda_k^{-1} f_k^2 - P \int d\mu_x e^{-\frac{(f(x)-y(x))^2}{2\kappa^2}} \to \int dk \lambda_k^{-1} f_k^2 - P \int d\mu_x e^{-\frac{(f(x)-y(x))^2}{2\kappa^2}}$$

Coppola, Helias, Ringel (To be published)

The technical challenges

- The theory has a background field (y(x))
- The average of f(x) scales differently than the std. under-rescaling
- Due to lack of locality, seemingly standard interaction terms contain delta functions in addition to the trivial one from field-theory $\delta(k_1 + k_2 + k_3 + k_4)$
- The scaling of Feynman diagrams becomes trickier and doesn't trivially depend on the naive operator scaling dimension.

A quick overview of preliminary results

GPR with MSE+MQE+large ridge: $S_{GP}[f(x)] = \sum_{k=1}^{\Lambda} \lambda_k^{-1} f_k^2 - P \int d\mu_x e^{-\frac{(f(x)-y(x))^2 + u(f(x)-y(x))^2}{2\kappa^2}}$

Large-P/UV universality/asymptotic-freedom:

Most reasonable perturbation to the model vanish at large P

Learning Curves from scaling dimensions:

Simple relations exist between 1/P expansion of the learning curves with the scaling dimensions obtained from RG

The ~~Big~~ Concrete + next Questions

- **Can we track the RG flow of** $S_{GP}[f(x)] = \sum_k k^\alpha f_k^2 + P \int d\mu_x e^{-\frac{(f(x)-y(x))^2}{2\kappa^2}}$ Yes at large input dim.
- **How many coupling constants are required to describe the flow?** One or few at large P
- **Is the RG flow indicative of the learning curve?** P sets the system size + Loss="P$^{\text{scaling-dimensions}}$"
- **Is there some form of universality?** Yes but only at large P (most perturbations are RG-relevant)
- **Are there useful RG fixed points?** Still checking this.
- **Is there a minimal model which faithfully captures large P behavior of a realistic network?**
- **Can importing RG machinery help us derive scaling laws in the feature learning regime?**

Thank you for your attention!

Join the phys4ml mailing list -

get your work emailed to 120+ physicists!

https://lists.fz-juelich.de/mailman/listinfo/phys4ml

Read and gives us comments on our review

Applications of Statistical Field Theory in Deep Learning

Zohar Ringel, Noa Rubin, Edo Mor, Moritz Helias, Inbar Seroussi

https://arxiv.org/abs/2502.18553

Extra slides

Adaptive Kernel Formulation

$$S_{MF}(w) = \frac{d|w|^2}{2\sigma_w^2} + \frac{\sigma_a^2}{2N}\int\int \langle t(x)\rangle_{MF}\langle t(x')\rangle_{MF}\phi(w^T x)\phi(w^T x') \approx_{d,n\to\infty} \frac{1}{2}w^T\Sigma_{MF}^{-1}w$$

$$S_{MF}(f) = \frac{1}{2}\int\int fK_{MF}^{-1}f + L[f] \qquad \langle t\rangle_{MF} = [K_{MF} + \sigma^2/n]^{-1}y$$

We find that the following ansatz for t(x) works at high dimension

$$\langle t(x)\rangle_{MF} = bH_1(w_*^T x) + cH_3(w_*^T x) \qquad \boldsymbol{y(x) = H_1(w^* \cdot x) + \epsilon H_3(w^* \cdot x\ )}$$

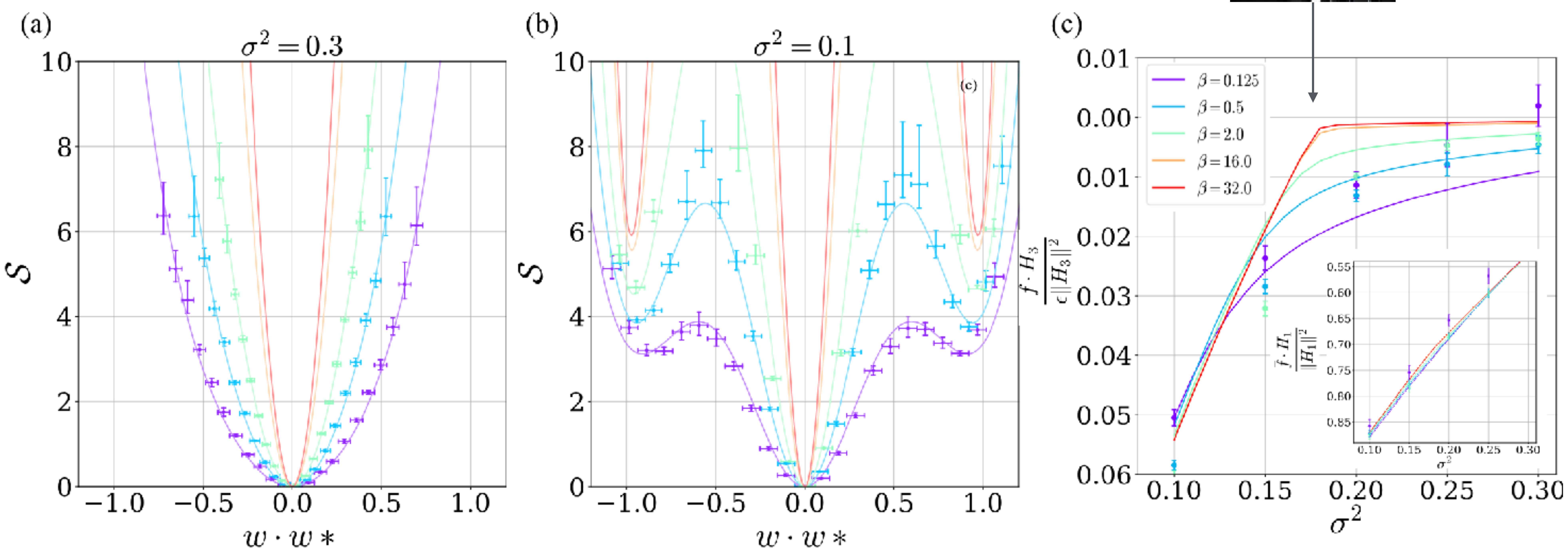
Calculating the integrals in $S_{MF}(w)$ yields (also at high dimension) the decoupled actions

$$S_{MF}(w^T w_{\perp,*}) = \frac{d|w^T w_{\perp,*}|^2}{2\sigma_w^2} \qquad \mathcal{S}[w\cdot w^*] = d\left(\frac{(w\cdot w^*)^2}{2\sigma_w^2} - \frac{2n^2\sigma_a^2}{\pi\sigma^4 dN}\frac{(w\cdot w^*)^2}{1+2\left(\sigma_w^2+(w\cdot w^*)^2\right)}\left(b - \frac{2c(w\cdot w^*)^2}{1+2\left(\sigma_w^2+(w\cdot w^*)^2\right)}\right)^2\right)$$

Our "ϕ^6" Landau theory ; $\phi = w_*^T w$

Theory - Experiment comparison

GFL/VGA phase GMFL/Coexistence phase



$d = 150\sqrt{\beta}$ $n = 3000\beta$ $N = 700\beta$ $\sigma^2 \propto \sqrt{\beta}$ FCN at MF-scaling

* Note that $n_{effective} = n/\sigma^2 \propto d$ - We find a change in complexity class!

Summary

Beyond Spectral Bias
Using Wilsonian RG

EK limit and beyond

Spectral Bias [Canatar et. Al.]

Multispectral kernel

Large N, large T

Large N, large d

$$S = \sum_c \frac{d|w_c|^2}{2\sigma_w^2} + \frac{\sigma_a^2}{N}\sum_{c=1}^{N}\int d\mu_x d\mu_y t(x)\phi(w_c^T x)\phi(w_c^T y)t(y) + i\int d\mu_x tf - n\int d\mu_x e^{-[f(x)-y(x)]^2/T}$$

$N \propto n$

$N \not\propto n$
standard-scaling

Adaptive Kernel

Kernel Scaling

Equilibrium Grokking
And phase transition
in representation learning

Sample Complexity
Effects and Soft Grokking

Erf+Beyond VGA

ReLU within VGA

Further evidence for complexity change

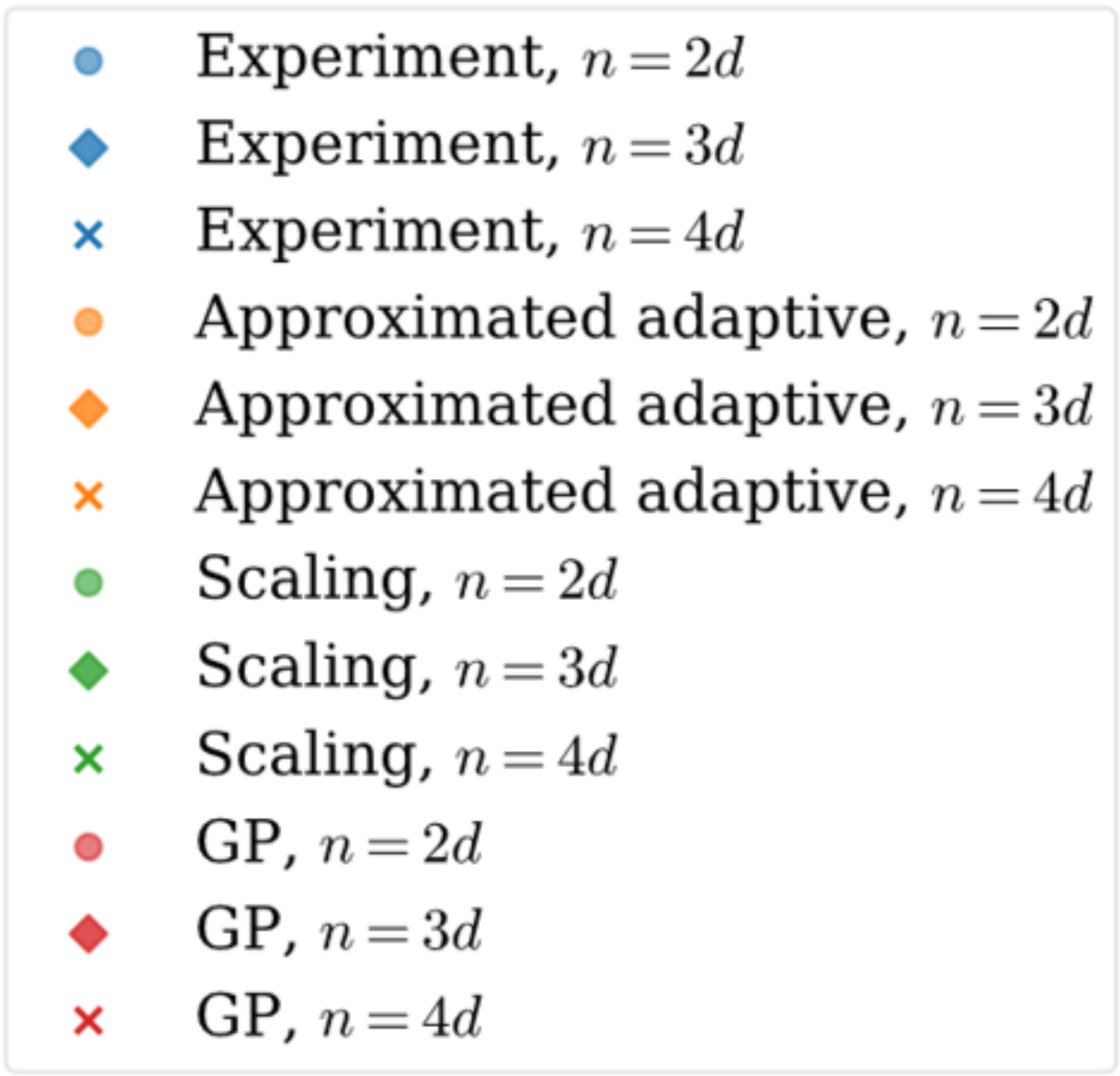
This time within VGA $[S[w] \approx w^T \Sigma_{MF}^{-1} w]$ and with two layer ReLU networks

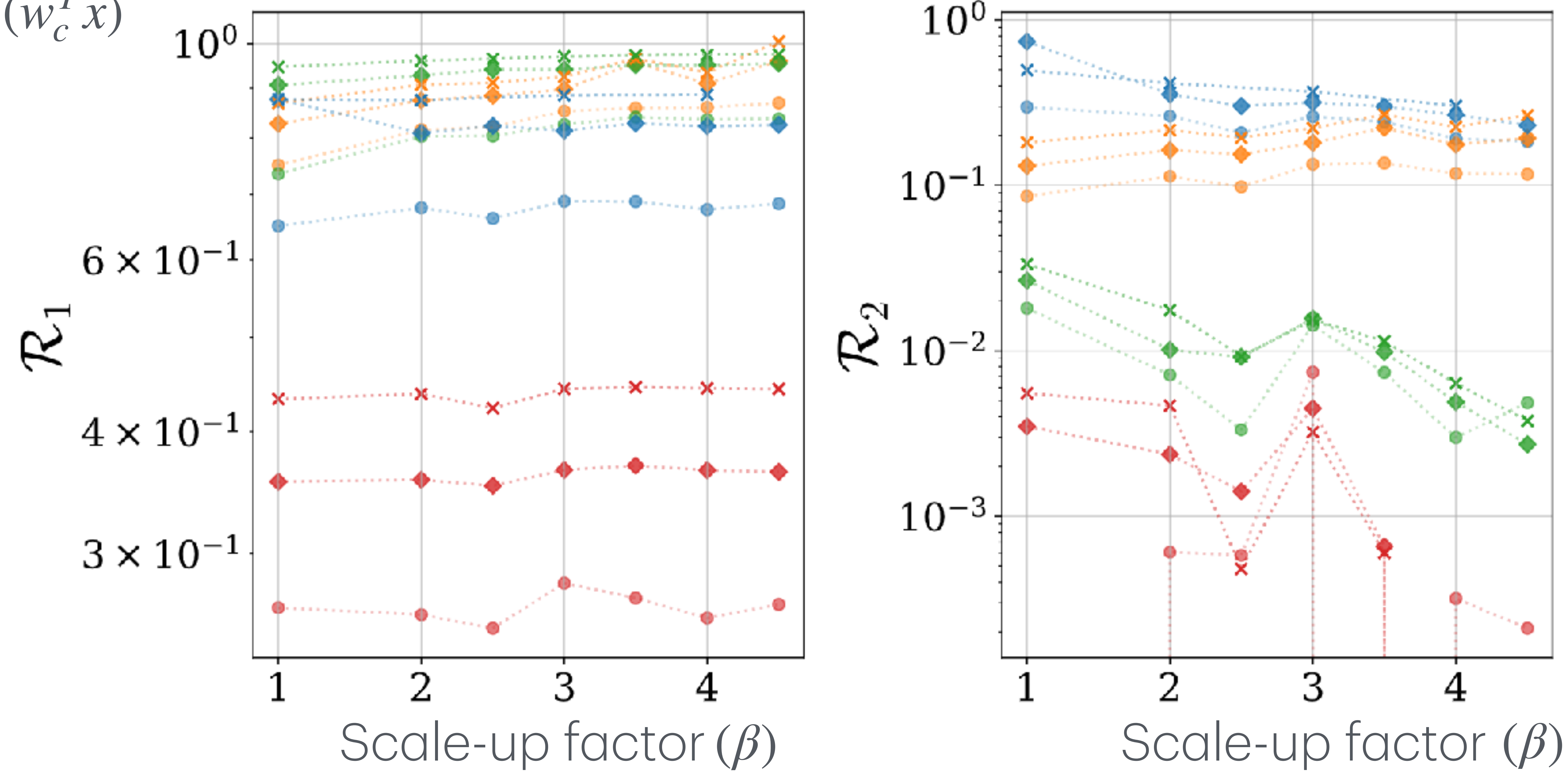
Target $\quad y(x) = v^{*\mathrm{T}} x + \epsilon \left(|v^{*\mathrm{T}} x| - \sqrt{2/\pi} \right)$

Learnabilities $\quad \mathcal{R}_p = \dfrac{\bar{f}(X) \cdot H_p(Xv^*)}{y(X) \cdot H_p(Xv^*)}$

Network $\quad f(x) = \sum_c a_c ReLU(w_c^T x)$

+ Mean-field scaling





$N = 750\beta \quad d = 96\beta$

A unified approach to feature learning in Bayesian Neural Networks **Rubin, Seroussi, ZR, Helias** (In preparation)

Kernel Adaptation predicts change in complexity class

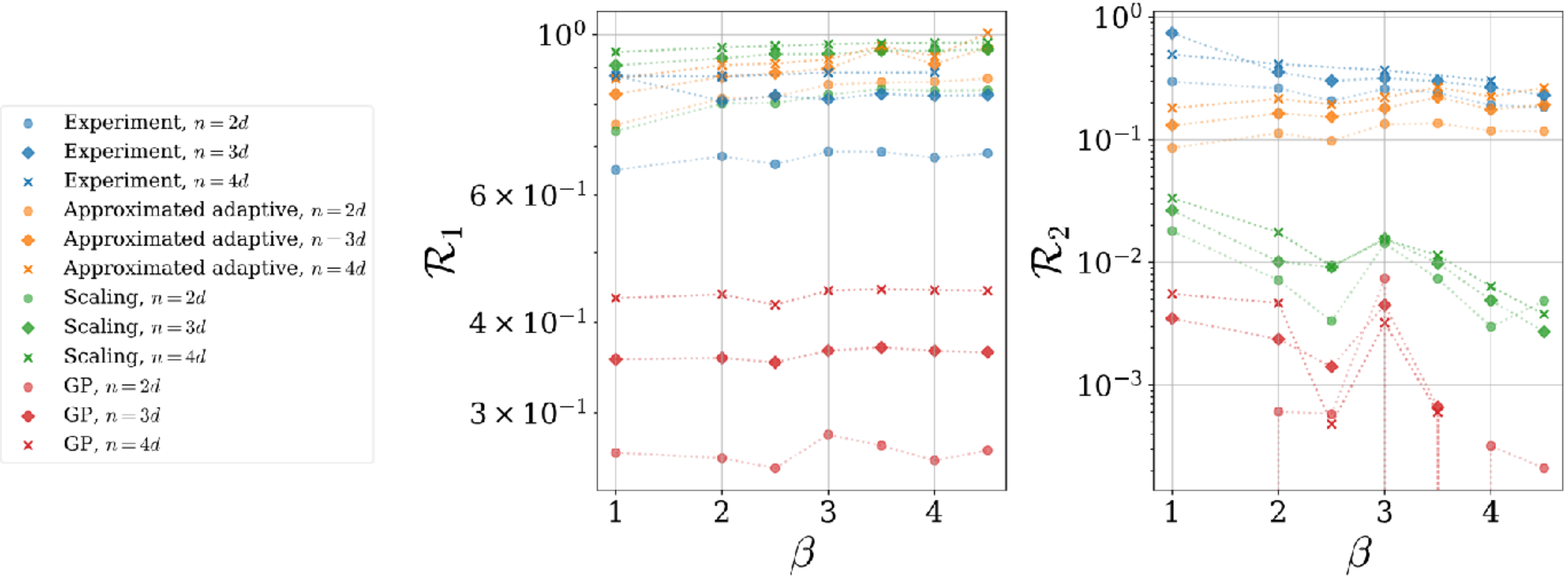


Figure 2: Learnability ratios $\mathcal{R}_{1,2}$ predicted by the approximate adaptive approach (orange (54)), scaling approach (green (6)), GP (red) compared to experimental values (blue), as a function of the scaling factor β, for different values of n/d.

$$y(x) = v^{*\mathrm{T}}x + \epsilon\left(|v^{*\mathrm{T}}x| - \sqrt{2/\pi}\right)$$

Noa Rubin, Zohar Ringel, Inbar Seroussi, Moritz Helias (2024) HiLD workshop