Title: Statistical physics of learning with two-layer neural networks

Speakers: Bruno Loureiro

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

Date: April 10, 2025 - 11:00 AM

URL: https://pirsa.org/25040093

Abstract:

Feature learning - or the capacity of neural networks to adapt to the data during training - is often quoted as one of the fundamental reasons behind their unreasonable effectiveness. Yet, making mathematical sense of this seemingly clear intuition is still a largely open question. In this talk, I will discuss a simple setting where we can precisely characterise how features are learned by a two-layer neural network during the very first few steps of training, and how these features are essential for the network to efficiently generalise under limited availability of data.

Statistical Physics of two layer-neural networks

Bruno Loureiro
@ CSD, DI-ENS & CNRS

bruno.loureiro@di.ens.fr

Theory + AI Workshop
Perimeter Institute, 10.04.2025

Motivation

What makes neural networks “good”?

Motivation

What makes neural networks “good”?

Feature learning? How to define this mathematically?

What functions are “easy” to learn with a neural net?

Motivation

Possible answer: Approximation theory.

Original contribution

Multilayer feedforward networks are universal approximators

Kurt Hornik, Maxwell Stinchcombe, Halbert White [1]

Neural Networks
Volume 2, Issue 5, 1989, Pages 359-366

Abstract

This paper rigorously establishes that standard multilayer feedforward networks with as few as one hidden layer using arbitrary squashing functions are capable of approximating any Borel measurable function from one finite dimensional space to another to any desired degree of accuracy, provided sufficiently many hidden units are available. In this sense, multilayer feedforward networks are a class of universal approximators.

c.f. [Cybenko 1989; Barron 1993; Bach 2017]

Motivation

But other ML methods are also universal approximators...

Universal Kernels

Journal of Machine Learning Research 7 (2006) 2651-2667

Charles A. Micchelli
Department of Mathematics and Statistics
State University of New York
The University at Albany
Albany, New York 12222, USA

Yuesheng Xu
Haizhang Zhang
Department of Mathematics
Syracuse University
Syracuse, NY 13244, USA

Abstract

In this paper we investigate conditions on the features of a continuous kernel so that it may approximate an arbitrary continuous target function uniformly on any compact subset of the input space. A number of concrete examples are given of kernels with this universal approximating property.

Curse of dimensionality

Let $\mathscr{D} = \{(x_i, y_i) \in \mathbb{R}^{d+1} : i = 1,\ldots,n\}$ denote data with:

$$y_i = f_\star(x_i) + \varepsilon_i$$

$$x_i \in [0,1]^d \qquad \mathbb{E}[\varepsilon_i \,|\, x_i] = 0 \qquad \mathbb{E}[\varepsilon_i^2 \,|\, x_i] = \sigma^2 < \infty$$

Curse of dimensionality

Let $\mathcal{D} = \{(x_i, y_i) \in \mathbb{R}^{d+1} : i = 1, \ldots, n\}$ denote data with:

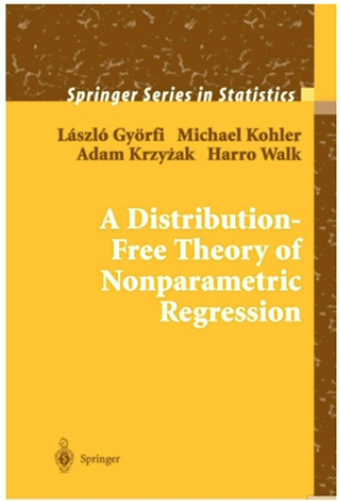
$$y_i = f_\star(x_i) + \varepsilon_i$$

$$x_i \in [0,1]^d \qquad \mathbb{E}[\varepsilon_i \,|\, x_i] = 0 \qquad \mathbb{E}[\varepsilon_i^2 \,|\, x_i] = \sigma^2 < \infty$$

Theorem (informal):

Assume $f_\star$ is α-Holder. Then:

$$\inf_{\hat{f}} \sup_{f_\star \in \mathcal{F}} \mathbb{E}[(\hat{f}(x) - f_\star(x))^2 \geq C n^{-\frac{\alpha}{2\alpha + d}}$$



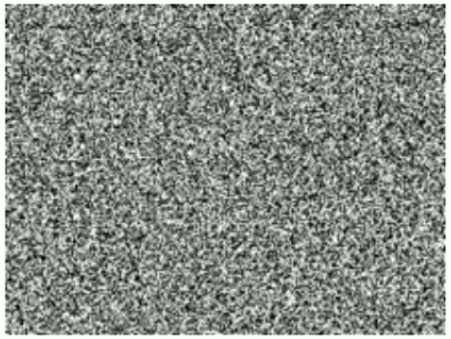
☹ Exponential dependence on d

Two cultures

To make progress, we need stronger assumptions on the data distribution.

Worst case

Typical case

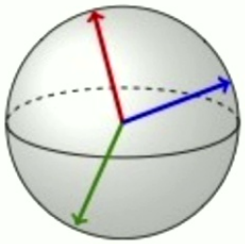


Multi-index functions

We will consider the class of Gaussian multi-index models.

$$y = g(w_1^\star x, \cdots, w_r^\star x)$$

$$x \sim \mathcal{N}(0, I_d/d) \qquad w_k \in \mathbb{S}^{d-1}(\sqrt{d})$$



Remark: Equivalently $y \sim P(y \mid W_\star x)$

Examples:

$r = 1$ *(single-index)*

$$g(z) = z$$

$$g(z) = z^2$$

$$g(z) = \text{sign}(z)$$

$r > 1$

$$g(z) = z_1 z_2 z_3 z_4$$

$$g(z) = \text{sign}(z_1 z_2 z_3)$$

$$g(z) = \sum_{k=1}^{r} a_k \sigma(z_k)$$

7

Some facts

- If $W_\star$ known, minimax rate with $d \to r$ [Stone '82; '85; '86]
- Information theoretically, estimating $W_\star \in \mathbb{R}^{k \times d}$ requires $n = O(d)$.
- Computationally, "most" GMIM can be efficiently estimated with $n = O(d)$. [Mondeli, Montanari '18; Barbier et al., '19;]

Fundamental computational limits of weak learnability in high-dimensional multi-index models

Emanuele Troiani[1], Yatin Dandi[1,2], Leonardo Defilippis[3], Lenka Zdeborová[1], Bruno Loureiro[3], and Florent Krzakala[2]

Optimal Spectral Transitions in High-Dimensional Multi-Index Models

Leonardo Defilippis[1], Yatin Dandi[2,3], Pierre Mergny[2], Florent Krzakala[2], and Bruno Loureiro[1]

Chapter I:

Learning with kernels and random features ("Initialisation")

Kernel methods

Let $\mathscr{D} = \{(x_i, y_i) \in \mathbb{R}^{d+1} : i \in [n]\}$ drawn independently from the GMIM.

Consider kernel ridge regression (KRR) on $\mathscr{D}$:

$$\min_{f \in \mathscr{H}} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - f(x_i)\right)^2 + \lambda ||f||_{\mathscr{H}}$$

where $\mathscr{H}$ is the RKHS associated with a kernel K. In particular: [Retch, Raimi 2007]

$$K_{\mathrm{RF}}(x, x') = \mathbb{E}_{w_0}\left[\sigma\left(\langle w^0, x\rangle\right)\sigma\left(\langle w^0, x'\rangle\right)\right] \approx \frac{1}{p}\sum_{k=1}^{p} \sigma\left(\langle w_k^0, x\rangle\right)\sigma\left(\langle w_k^0, x'\rangle\right)$$

Widely studied over the past ~5 years:

Mei, Montanari '19; Ghorbani, Mei, Misiakiewicz, Montanari '19, '20, '21; Gerace, **BL**, Krzakala, Mézard, Zdeborová '20; Goldt, **BL**, Reeves, Krzakala, Mézard, Zdeborová '21 Dhiffalah & Lu '20; Hu & Lu '20; Liang, Sur '20; Jacot, Simsek, Spadaro, Hongler, Gabriel '20; Liu, Liao, Suykens '21; Mei, Misiakiewicz, Montanari '22; Fan, Wang 2020; Schröder, Cui, Dmitriev, **BL** '23, 24; Defilippis, **BL**, Misiakiewicz ;24; Wang, Chen, Rosasco, Liu '25; etc...

Challenge

Predictor given by: $\hat{f}(x) = \sum_{j=1}^{p} \hat{a}_\lambda^\top \sigma(\langle w_j^0, x \rangle)$ $\hat{a}_\lambda = (\Phi^\top \Phi + \lambda I_p)^{-1} \Phi^\top y$

Morally, $\hat{a}_\lambda \approx$ projector into $\mathrm{col}(\Phi)$.

Challenge: Even if $x_i \sim \mathcal{N}(0, 1/dI_d)$, $\Phi_{ik} = \sigma(\langle w_j^0, x_i \rangle)$ is not Gaussian!

Key idea: Expand in orthogonal basis (Hermite polynomials):

$$\Phi_{ik} = \sigma(\langle w_j^0, x_i \rangle) = \sum_{k \geq 0} \mu_k h_k(\langle w_k^0, x \rangle)$$

$\sim O(d^{-k/2})$

Therefore, if $n = O(d)$:

$$\Phi_{ik} \approx \mu_0 + \mu_1 \langle w_k^0, x_i \rangle + \mu_\star z_i$$

Gaussian equivalence

Consider the following two ERM problems:

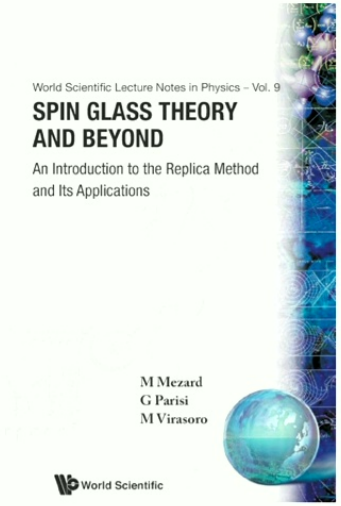
$$\hat{a}_\lambda(X,y) = \underset{a}{\operatorname{argmin}} \frac{1}{2n} \sum_{i=1}^{n} (y_i - \langle a, \sigma(W^0 x_i)\rangle)^2 + \lambda ||a||_2^2$$

$$\hat{a}_\lambda^G(X,y) = \underset{a}{\operatorname{argmin}} \frac{1}{2n} \sum_{i=1}^{n} (y_i - \langle a, \mu_0 1 + \mu_1 W^0 x_i + \mu_\star z_i\rangle)^2 + \lambda ||a||_2^2$$

Then, in the limit $d \to \infty$ with $n, p = \Theta(d)$:

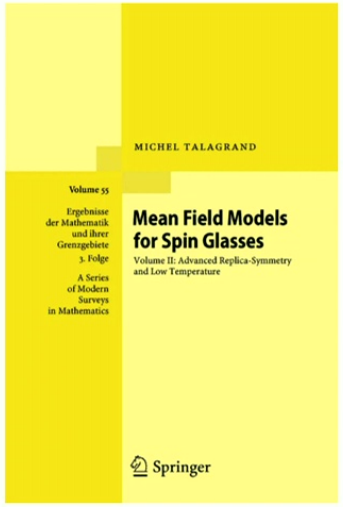
Gaussian equivalence principle (GEP)
[Goldt et al. '19, 20; Mei & Montanari '19; Hu& Lu '20]

$$|R(\hat{a}_\lambda) - R(\hat{a}_\lambda^G)| \to 0$$



$$\mu_\beta(a) = \frac{1}{Z_\beta(\Phi, y)} e^{-\beta H(a)}$$

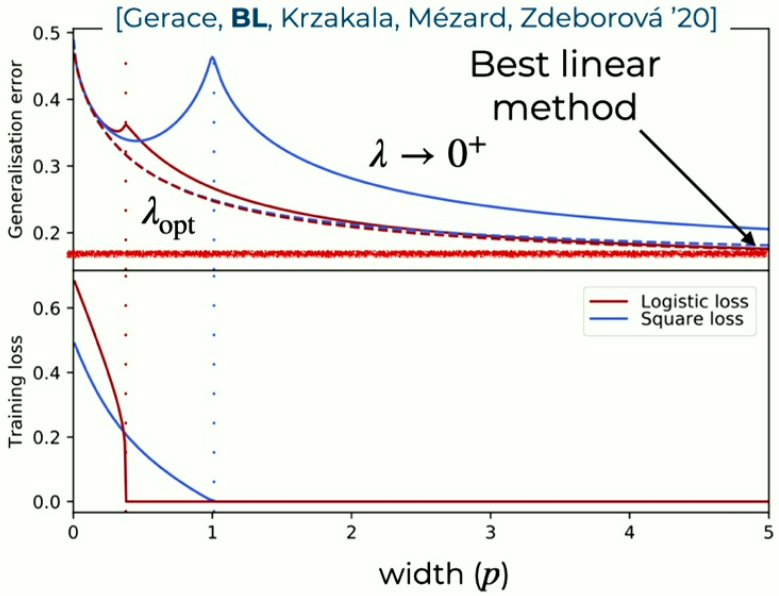
$$H(a) = \frac{1}{2n} \sum_{i=1}^{n} (y_i - \langle a, \sigma(W^0 x_i))^2 + \lambda \,||\, a \,||_2^2$$



Gaussian equivalence

Gaussian equivalence principle (GEP)
[Goldt et al. '19; Mei & Montanari '19; Hu& Lu '20]

$$|R(\hat{a}_\lambda) - R(\hat{a}^G_\lambda)| \to 0$$



Gaussian universality

↕

Equivalence to a linear model

↕

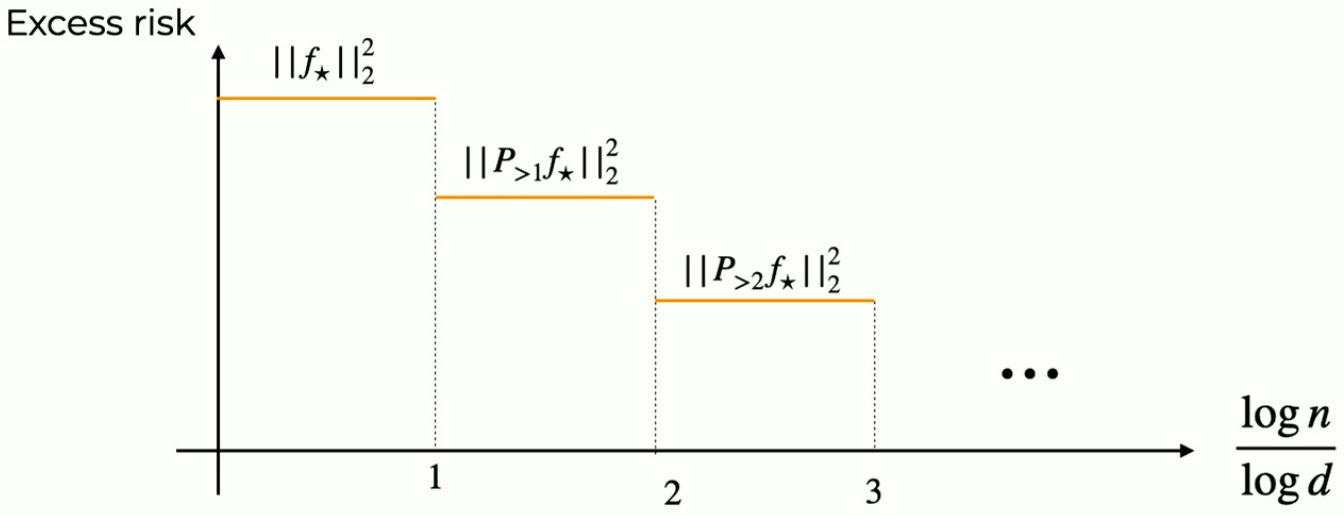
Limited expressivity

Limitations of KRR

Theorem [Mei, Misiakiewicz, Montanari '22, informal]:

For isotropic data (e.g. $x \sim \mathrm{Unif}(\mathbb{S}^{d-1})$), with $n, p = \Theta(d^\kappa)$ one can learn at best a polynomial approximation of degree κ of the target $f_\star(x)$

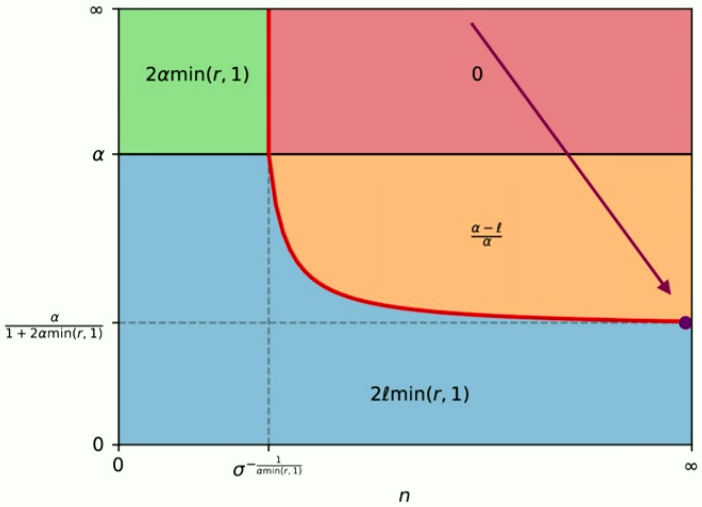
$$\mathbb{E}\,||f_\star(x) - f(x; \hat{a}_\lambda, W^0)||_2^2 = ||P_{>\kappa} f_\star||_{L_2}^2 + o_d(1)$$



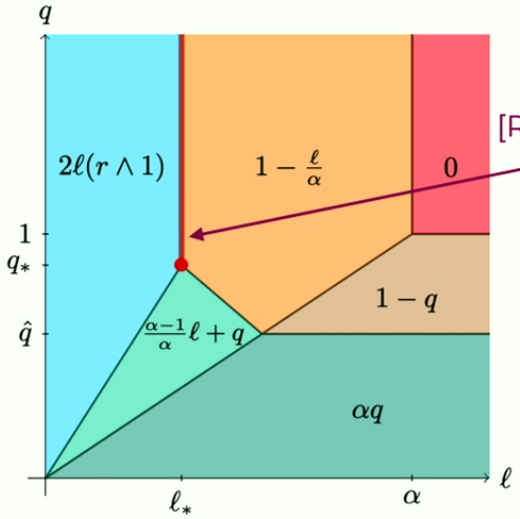
Side comment

$$p \sim n^q \qquad \eta_k \sim k^{-\alpha}$$
$$\lambda \sim n^{-\ell} \qquad f_{\star,k} = k^{-\alpha r - 1/2}$$

[Caponnetto, De Vitto '07]



[Cui et al., '20]



[Rudi, Rosasco 2017]

[Defilippis et al., '24]

$$\mathbb{E}\,||f(x) - f_\star(x)||^2 \sim n^{-\gamma}$$

Summary I

Kernels/RF are able to learn “anything”, but they need “a lot” of data.

KRR on isotropic high-d data $\approx$ Polynomial fit

In particular, no adaptativity to target structure

Chapter II:
Learning with two-layer neural networks (away from initialisation)

Giant steps

Full gradient dynamics is challenging. Instead, look at something simpler:

1. Initialise (a^0, W^0) such that $f(x; a^0, W^0) \approx 0$

2. Take t steps of SGD with fresh batches on first layer:

$$W^{t+1} = W^t - \eta \nabla_W \frac{1}{b} \sum_{i=1}^{b} (y_i - f(x_i; a^0, W^t))^2$$

3. Train the second layer on fresh batch:

$$\min_{a \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - f(x_i; a, W^t)\right)^2 + \lambda ||a||_2^2$$

What can we learn after one GD step ?

$$W^{t+1} = W^t - \eta \nabla_W \frac{1}{b} \sum_{i=1}^{b} (y_i - f(x_i; a^0, W^t))^2$$

Two flavours of results:

1. Weak learnability: How much W^t correlates with $W_\star$?

2. Generalisation: How much this improves the error?

Weak learning with SGD

Consider $r = 1$ for simplicity.

$$\mathbb{E}_x[\langle \text{grad}^t, w_\star \rangle] \propto \mathbb{E}[g(\langle w_\star, x \rangle)\sigma'(\langle w_j^t, x \rangle)\langle w_\star, x \rangle]$$

$$= \mathbb{E}[g(z_\star)\sigma'(z_j)z_\star]$$

$$= \mathbb{E}[g'(z_\star)\sigma'(z_j)]$$

$$= \sum_{k \geq 1} \sigma_k g_k \mathbb{E}[h_k(z_\star)h_{k'}(z_j)]$$

$$= \sum_{k \geq 1} \sigma_k g_k \langle w_j^t, w_\star \rangle^k$$

$$= C\langle w_j^t, w_\star \rangle^{\ell - 1} + \dots$$

Weak learning with SGD

Consider $r = 1$ for simplicity.

To achieve $O(1)$ correlation, requires:

$$n = \begin{cases} O(d) & \ell = 1 \\ O(d \log d) & \ell = 2 \\ O(d^{\ell-1}) & \ell > 2 \end{cases}$$

Where $\ell = \min\{k : \mathbb{E}[g(z)h_k(z)] \neq 0\}$

"Information exponent" [Ben Arous, Gheissari, Jagannath 2019]

Remarks:

- Can be extended to $r > 1$ direction-wise: *leap exponent*. [Abbe et al., 2022, 2023]
- One-pass SGD implements CSQ queries:

$$\mathbb{E}[y\phi(x)]$$

Indeed, can be proven this is the CSQ lower-bound

[Damian et al., 2022]

Weak learning with SGD

Examples:

		Kernels	SGD	AMP (opt. alg.)
$r = 1$				
$\ell = 1$	$g(z) = z$	$n = O(d)$	$n = O(d)$	$n = O(d)$
$\ell = 2$	$g(z) = z^2$	$n = O(d^2)$	$n = O(d \log d)$	$n = O(d)$
$\ell = 3$	$g(z) = z^3 - 3z$	$n = O(d^3)$	$n = O(d^2)$	$n = O(d)$
$r > 1$				
$\ell = 2$	$g(z) = z_1 z_2$	$n = O(d^2)$	$n = O(d \log d)$	$n = O(d)$
$\ell = 3$	$g(z) = z_1 z_2 z_3$	$n = O(d^3)$	$n = O(d^2)$	$n = O(d)$
$\ell = 3$	$g(z) = \text{sign}(z_1 z_2 z_3)$	$n = O(e^d)$	$n = O(d^2)$	???

What can we learn after one GD step ?

$$W^{t+1} = W^t - \eta \nabla_W \frac{1}{b} \sum_{i=1}^{b} (y_i - f(x_i; a^0, W^t))^2$$

Two flavours of results:

1. Weak learnability: How much W^t correlates with $W_\star$?

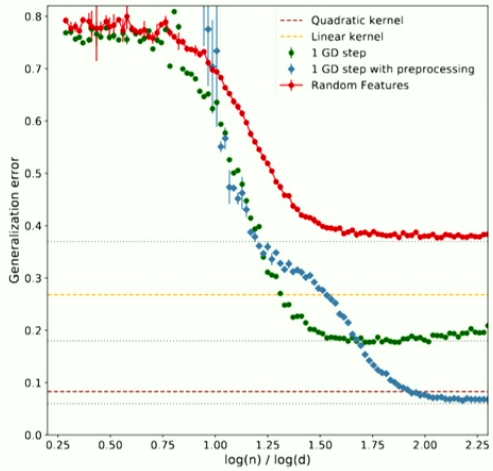
2. Generalisation: How much this improves the error?

Generalisation

We can show that at best learn non-linear functions along learned subspace:

Theorem [Dandi, Krzakala, **BL**, Pesce, Stephan '23, informal]:
Let $U \subset \mathrm{span}(w_1^\star, \dots, w_r^\star)$ be the space learned after t SGD steps.
Then, for any a such that $||a||_\infty = \Theta_d(1)$:

$$\mathbb{E}\,||f_\star(x) - f(x; a, W^t)||_2^2 \geq \mathrm{Var}(f_\star(z)\,|\,P_U z) - o_d(1)$$



What about tight results?

Summary II

With a single gradient step and

$$n, p, \eta = \Theta(d)$$

can learn at best a non-linear function of one direction

$$f_\star(x) = g(\langle \theta_\star, x \rangle)$$

🤔 Can we get sharp asymptotics for the error?

Gradient after 1 step

There is one case we can do better: $n, p, \eta = \Theta(d)$. [Cui et al., 24', '25]

Step 1: Characterisation of the gradient.

$$W^1 \approx W + ruv \qquad \kappa = \langle v, \theta_\star \rangle > 0$$

Step 2: Study a "spiked" random features model

$$f(x;\theta) = \langle a, \sigma\left(W^\top x + u\langle v, x\rangle\right)\rangle$$

Conditional Gaussian Equivalence

$$\sigma\left(\langle w^1, x\rangle\right) \approx \mu_0(\langle v, x\rangle) + \mu_1(\kappa)\langle w^0, x^\perp\rangle + \mu_\star(\kappa)\xi$$

$$\kappa = \langle v, x\rangle \qquad x = \kappa\theta_\star + x^\perp$$

27

Exact asymptotics ($a^0 = 1_p$)

$$\begin{cases} V_1 = \int \frac{d\nu(\varrho,\tau,\pi)\varrho}{\lambda + \hat{V}_1\varrho + \hat{V}_2} \\ V_2 = \int \frac{d\nu(\varrho,\tau,\pi)}{\lambda + \hat{V}_1\varrho + \hat{V}_2} \\ m = \frac{\mathbb{E}_{\kappa,y}\left[\frac{\mu_0(\kappa)(\sigma_\star(\kappa,y) - \mu_1(\kappa)\kappa\zeta)}{1+V(\kappa)}\right]}{\mathbb{E}_\kappa\left[\frac{\mu_0(\kappa)^2}{1+V(\kappa)}\right]} \\ \zeta = \hat{\zeta}\sqrt{\beta}\int d\nu(\varrho,\tau,\pi)\varrho\tau^2\frac{1}{\lambda + \hat{V}_1\varrho + \hat{V}_2} + \beta^{\frac{3}{2}}\hat{\zeta}\hat{V}_1\frac{I(\hat{V}_1,\hat{V}_2)^2}{1-\beta\hat{V}_1 I(\hat{V}_1,\hat{V}_2)} \\ \psi = \hat{\psi}\sqrt{\beta}\int \frac{d\nu(\varrho,\tau,\pi)\varrho\pi^2}{\lambda + \hat{V}_1\varrho + \hat{V}_2} \end{cases}$$

$$\begin{cases} q_1 = \int d\nu(\varrho,\tau,\pi)\varrho\frac{\left(\hat{q}_1\varrho + \hat{q}_2 + \hat{\zeta}^2\varrho\tau^2 + \hat{\psi}^2\varrho\pi^2\right)}{\left(\lambda + \hat{V}_1\varrho + \hat{V}_2\right)^2} - \beta\hat{\zeta}^2\frac{I(\hat{V}_1,\hat{V}_2)^2}{\left(1-\beta\hat{V}_1 I(\hat{V}_1,\hat{V}_2)\right)^2} \\ \qquad -\hat{\zeta}^2\frac{\int\frac{\tau^2\varrho^2 d\nu(\varrho,\tau,\pi)}{(\lambda+\hat{V}_1\varrho+\hat{V}_2)^2}\left[\left(1-\beta\hat{V}_1 I(\hat{V}_1,\hat{V}_2)\right)^2 - 1\right]}{\left(1-\beta\hat{V}_1 I(\hat{V}_1,\hat{V}_2)\right)^2} \\ q_2 = \int\frac{\left(\hat{q}_1\varrho + \hat{q}_2 + \hat{\zeta}^2\varrho\tau^2 + \hat{\psi}^2\varrho\pi^2\right)d\nu(\varrho,\tau,\pi)}{\left(\lambda + \hat{V}_1\varrho + \hat{V}_2\right)^2} \\ \qquad -\hat{\zeta}^2\int\frac{\tau^2\varrho d\nu(\varrho,\tau,\pi)}{(\lambda+\hat{V}_1\varrho+\hat{V}_2)^2}\left[1 - \frac{1}{\left(1-\beta\hat{V}_1 I(\hat{V}_1,\hat{V}_2)\right)^2}\right] \end{cases}$$

$$\begin{cases} \hat{V}_1 = \frac{\alpha}{\beta}\mathbb{E}_\kappa\frac{\rho\mu_1(\kappa)^2}{1+V(\kappa)} \\ \hat{V}_2 = \frac{\alpha}{\beta}\mathbb{E}_\kappa\frac{\rho\mu_2(\kappa)^2}{1+V(\kappa)} \\ \hat{\zeta} = \frac{\alpha}{\sqrt{\beta}}\mathbb{E}_{\kappa,y}\kappa\mu_1(\kappa)\frac{b(\kappa,y)}{1+V(\kappa)} \\ \hat{\psi} = \frac{\alpha}{\sqrt{\beta}}\mathbb{E}_{\kappa,y}\frac{y\mu_1(\kappa)b(\kappa,y) + \psi\mu_1(\kappa)^2}{1+V(\kappa)} \end{cases}$$

$$\begin{cases} \hat{q}_1 = \frac{\alpha}{\beta}\mathbb{E}_{\kappa,y}\mu_1(\kappa)^2\frac{b(\kappa,y)^2 + \rho q(\kappa) - \mu_1(\kappa)^2\psi^2}{\left(1+V(\kappa)\right)^2} \\ \hat{q}_2 = \frac{\alpha}{\beta}\mathbb{E}_{\kappa,y}\mu_2(\kappa)^2\frac{b(\kappa,y)^2 + \rho q(\kappa) - \mu_1(\kappa)^2\psi^2}{\left(1+V(\kappa)\right)^2} \end{cases}$$

$$\alpha_0 = n_B/d \qquad \beta = p/d$$
$$\alpha = n/d \qquad \tilde{\eta} = \eta/d$$
$$\kappa = \langle v, x\rangle \qquad \rho = 1 - \gamma^2$$
$$\gamma = \langle v, \theta_\star\rangle$$

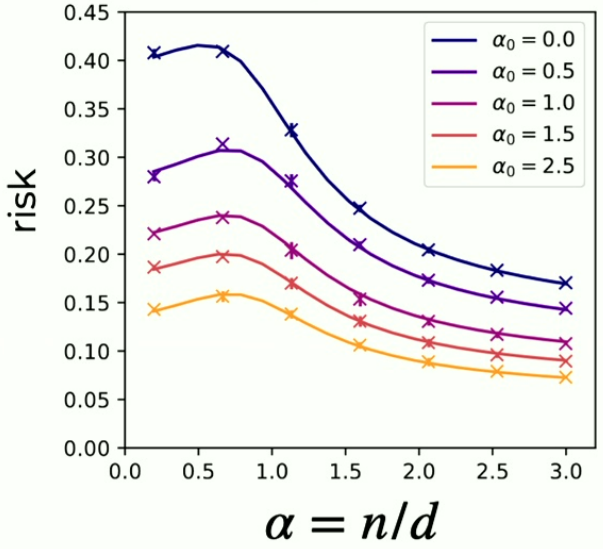
$$W = \sum_{i=1}^{\min(p,d)} \lambda_i e_i f_i^\top \qquad \Pi^\perp = I_d - vv^\top$$

$$\nu(\varrho,\tau,\pi) = \frac{1}{p}\sum_{i=1}^{\min(p,d)} \delta\left(\lambda_i - \varrho\right)\delta\left(f_i^\top v - \tau\right)\delta\left(f_i^\top\Pi^\perp\vec{\theta} - \pi\right)$$

Batch size

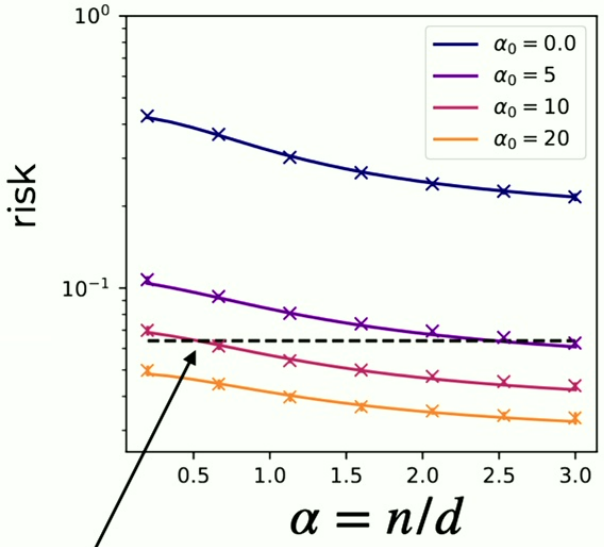
$\tilde{\eta} = 1$ $\lambda = 10^{-2}$

$\sigma = g = \tanh$



$\tilde{\eta} = 3$ $\lambda = 0.1$

$\sigma = \tanh$ $g = \text{sign}$

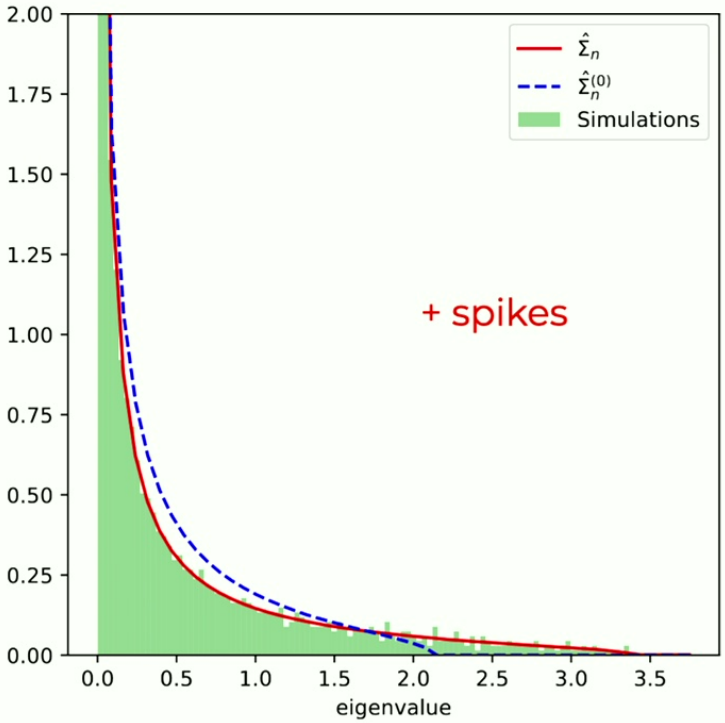


Best linear predictor

$||P_{\kappa \leq 1} f_\star||^2$

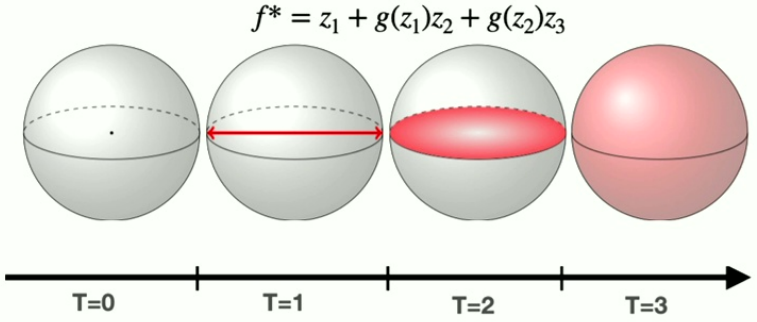
29

Spectral properties



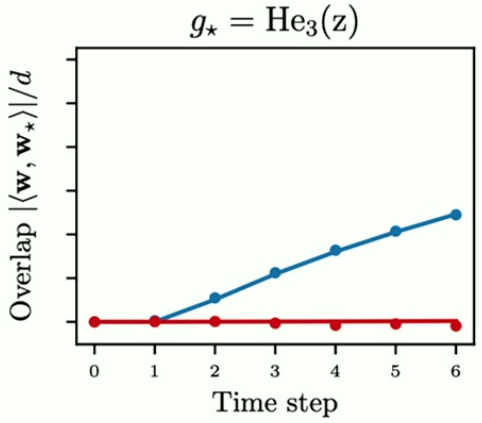
Epilogue

- Multiple steps: hierarchical learning [Dandi et al., '24]



- Reusing batches:

[Troiani et al., '24]



Conclusion

- In proportional asymptotics, kernels can learn at best a linear approximation
- With one gradient step, 2LNN learn do better than kernels along one (and only one) direction
- We can provide a sharp asymptotic description on what is learned

Thank you!


L. Zdeborová
(EPFL)


F. Krzakala
(EPFL)


L. Stephane
(EPFL)


L Arnaboldi
(EPFL)


Y.M. Lu
(Harvard U.)


H. Cui
(Harvard)


L Pesce
(EPFL)


Y. Dandi
(EPFL)


E. Troiani
(EPFL)


L. Defilippis
(DI-ENS)


S. Goldt
(SISSA)


F. Gerace
(UniBo)


M. Mézard
(Bocconi U.)