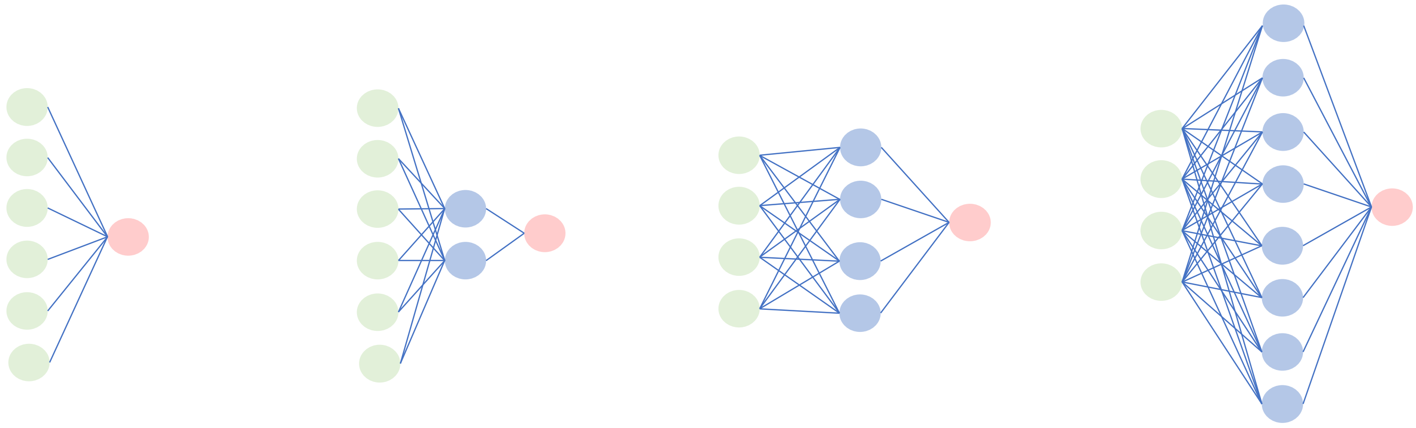
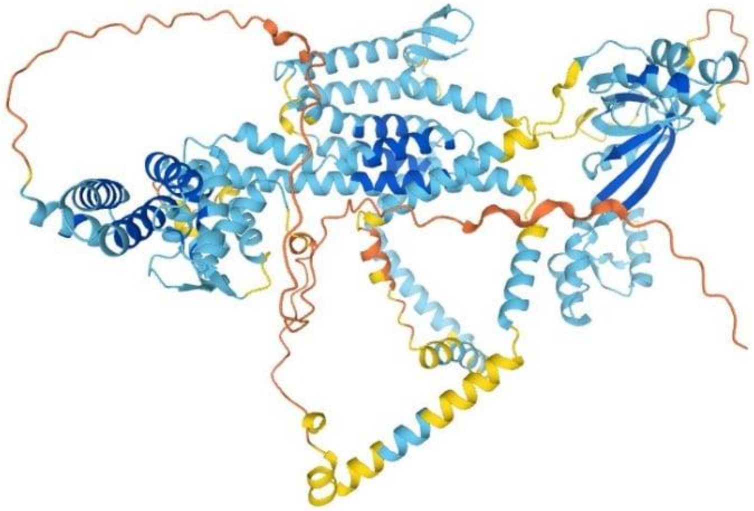
Learning (with) deep random networks

Hugo Cui

SPOC lab, EPFL, Switzerland

Wed. April 5 2023





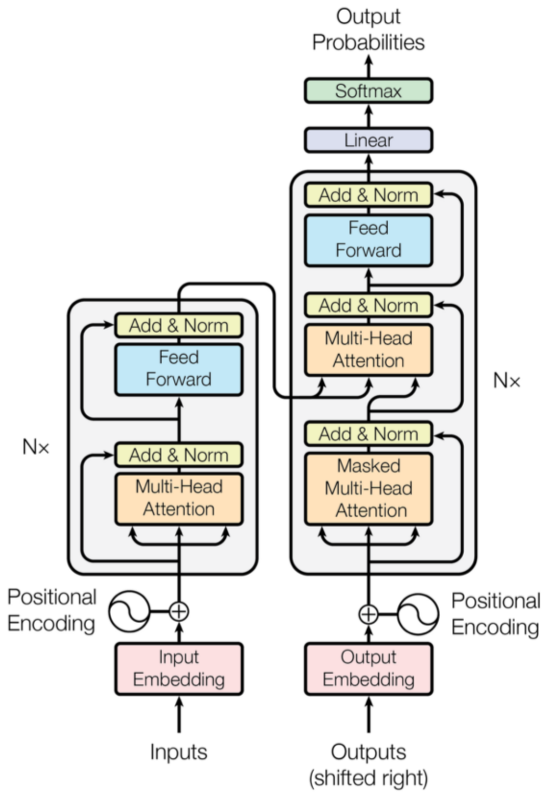
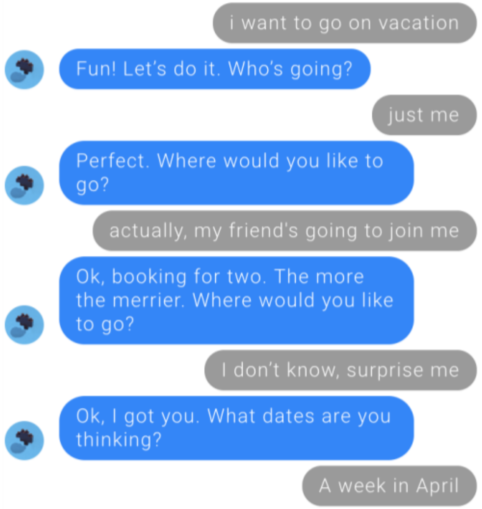


Figure 1: The Transformer - model architecture.



Bruno Loureiro
ENS

Cédric Gerbelot
NYU

Sebastian Goldt
SISSA

Dominik Schröder
ETH

Florent Krzakala
EPFL

Lenka Zdeborová
EPFL

Marc Mézard
Bocconi

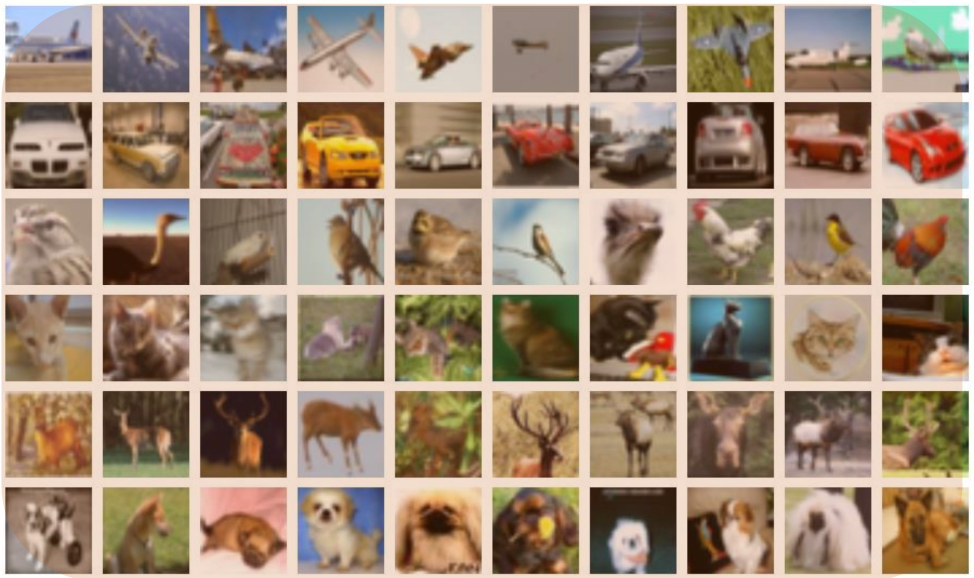
Daniil Dmitriev
ETH

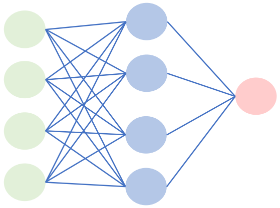
training

labels

+1
+1
-1
-1
-1
-1

Training data





training

labels

+1

+1

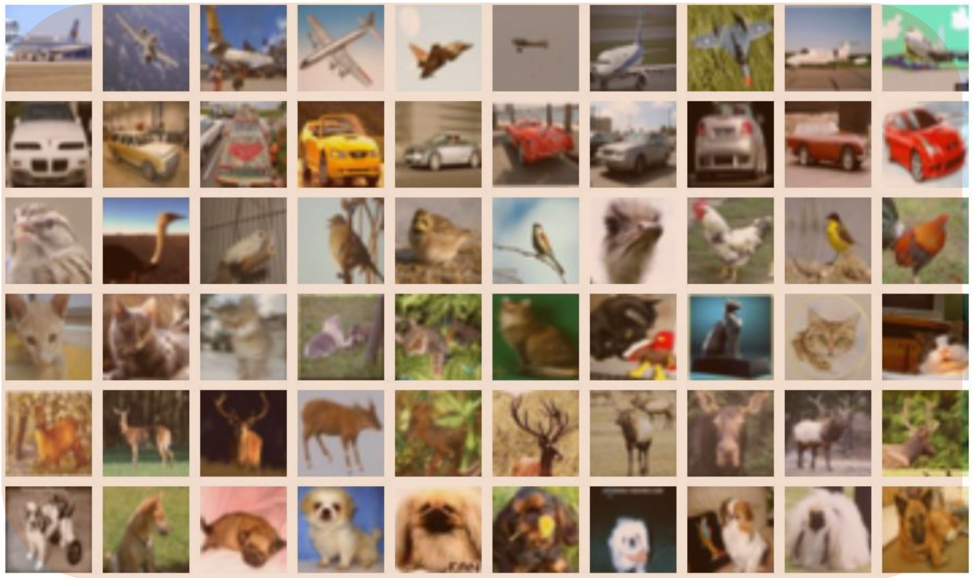
-1

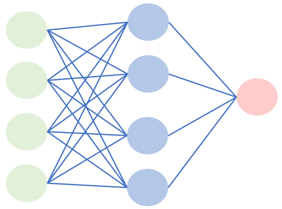
-1

-1

-1

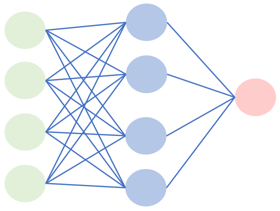
Training data





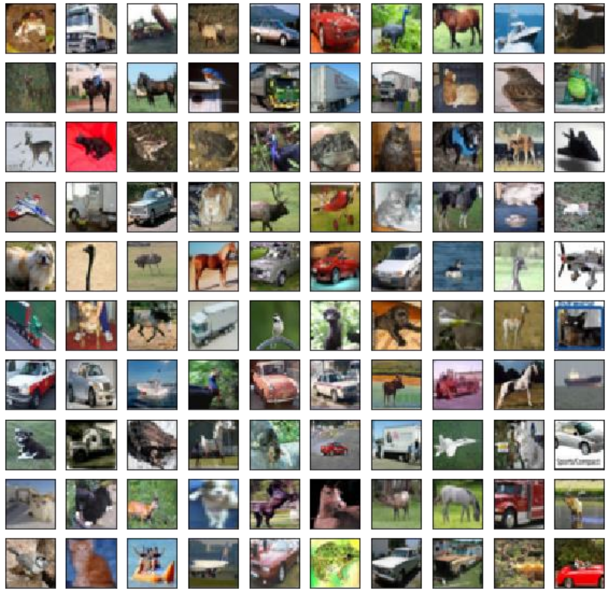
testing

Test data

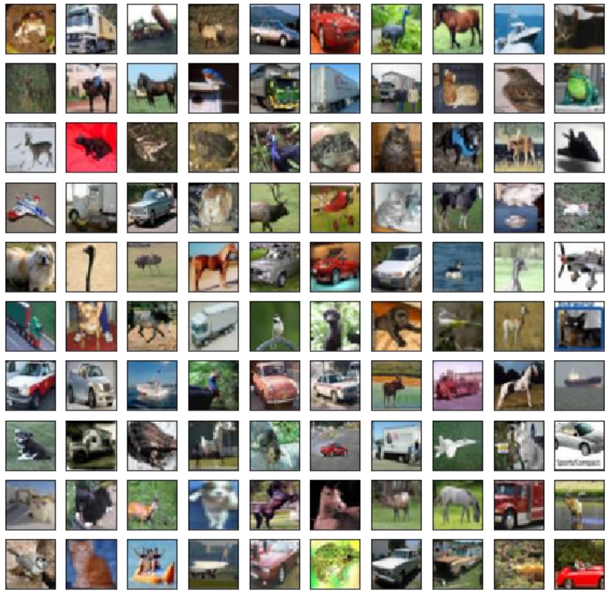


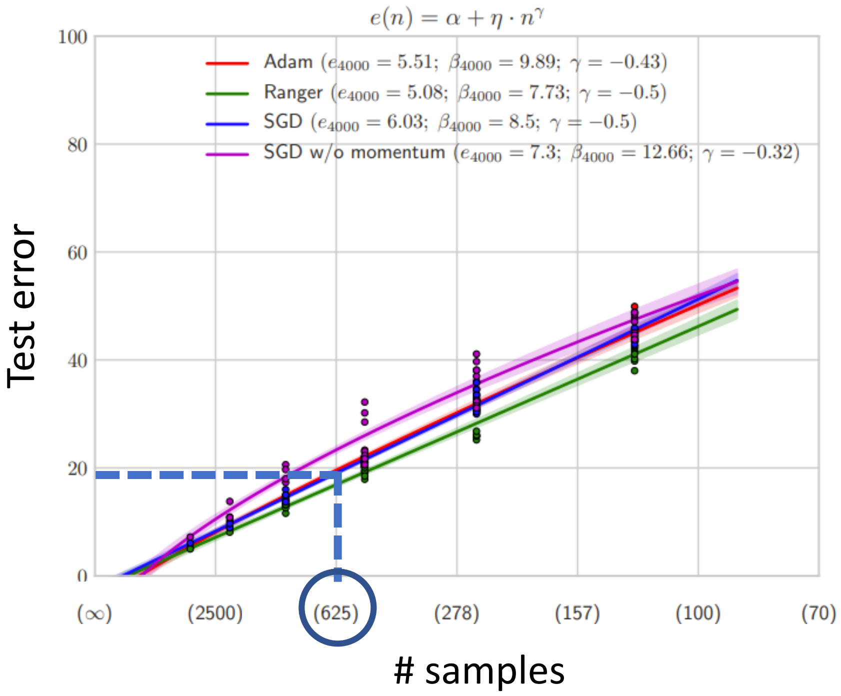
-1

Predicted label



Question: ***How large does your train set needs to be*** to learn CIFAR 10 to test error <20%?





Question*: *How large does your train set needs to be to learn CIFAR 10 to test error <20%?

***(Empirical) Answer*:** Probably ≈ 600, using good networks.

Hoeim et al., *Learning curves for Analysis of Deep Networks*, ICML 2020

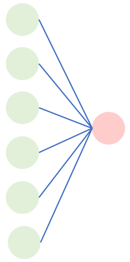
(Q1) Given a target function $y^\star: \mathbb{R}^d \to \mathbb{R}$, and a data distribution ν over $\mathbb{R}^d$, ***how many i.i.d training samples*** $x^\mu \sim \nu$ does one need to sample so that from a train set $\mathcal{D} = \{x^\mu, y^\star(x^\mu)\}_{\mu=1}^n$ the target can be learnt ***up to test error*** $\epsilon_g < \epsilon$?

(Q1) Given a target function $y^{\star}: \mathbb{R}^d \to \mathbb{R}$, and a data distribution ν over $\mathbb{R}^d$, ***how many i.i.d training samples*** $x^{\mu} \sim \nu$ does one need to sample so that from a train set $\mathcal{D} = \{x^{\mu}, y^{\star}(x^{\mu})\}_{\mu=1}^{n}$ the target can be learnt ***up to test error*** $\boldsymbol{\epsilon_g < \epsilon}$?

or, equivalently:

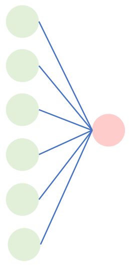
(Q1') For a train set $\mathcal{D} = \{x^{\mu}, y^{\star}(x^{\mu})\}_{\mu=1}^{n}$ of given size n, what is ***the lowest achievable test error*** $\boldsymbol{\epsilon_g}$ one can hope to achieve with typical algorithms, e.g. ERM?

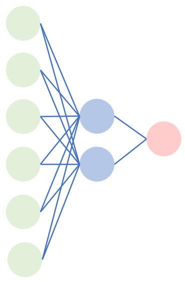
Theoretical testbeds: random neural networks



Barbier et al, *Optimal errors and phase transitions in high-dimensional generalized linear models,* PNAS 2017

Theoretical testbeds: random neural networks



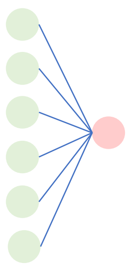


$\color{#C55A11}{width} \ll \color{#548235}{dimension}$

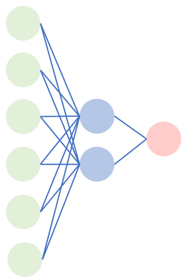
Barbier et al, *Optimal errors and phase transitions in high-dimensional generalized linear models,* PNAS 2017

Aubin et al, *The committee machine: Computational to statistical gaps,* NeurIPS 2019

Theoretical testbeds: random neural networks

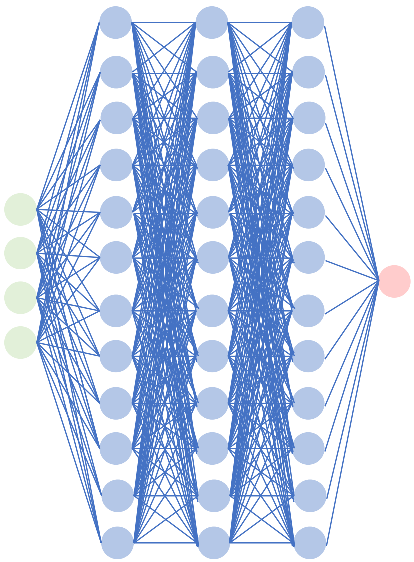


Barbier et al, *Optimal errors and phase transitions in high-dimensional generalized linear models,* PNAS 2017



$\textit{width} \ll \textit{dimension}$

Aubin et al, *The committee machine: Computational to statistical gaps,* NeurIPS 2019



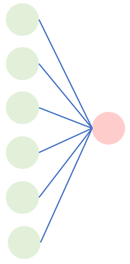
$\textit{width} \gg \textit{dimension}$

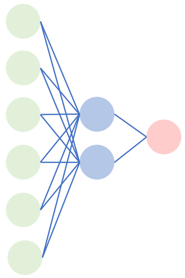
Neal, *Priors for infinite nets*, Uni. Toronto 1996

Williams, *Computing with infinite networks, NeurIPS 1996*

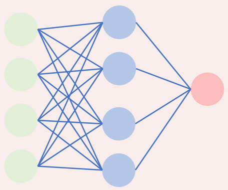
Lee et. al., *Deep Neural Networks as GPs*, ICLR 2018

Theoretical testbeds: random neural networks

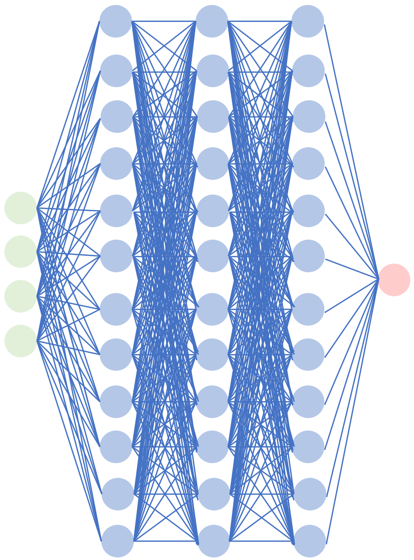




$width \ll dimension$



$width \sim dimension$



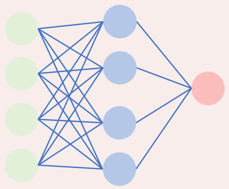
$width \gg dimension$

Barbier et al, *Optimal errors and phase transitions in high-dimensional generalized linear models,* PNAS 2017

Aubin et al, *The committee machine: Computational to statistical gaps,* NeurIPS 2019

Neal, *Priors for infinite nets*, Uni. Toronto 1996
Williams, *Computing with infinite networks, NeurIPS 1996*
Lee et. al., *Deep Neural Networks as GPs*, ICLR 2018

Some related works:



$width \sim dimension$

High-dimensional formulae for sign/ReLU Bayes regression

Li and Sompolinsky, *Statistical mechanics of deep linear neural networks: The backpropagating kernel renormalization,* PRX, 2021.

Ariosto et al., *Statistical mechanics of deep learning beyond the infinite-width limit*. ArXiv, abs/2209.04882, 2022.

(Non)-asymptotics for linear networks

Zavatone-Veth, Tong and Pehlevan, *Contrasting random and learned features in deep bayesian linear regression*, PRE 2022

Hanin and Zlokapa, *Bayesian interpolation with deep linear networks*. ArXiv, abs/2212.14457, 2022

(Data) Gaussian data: $x \sim \mathcal{N}(0, \Sigma)$

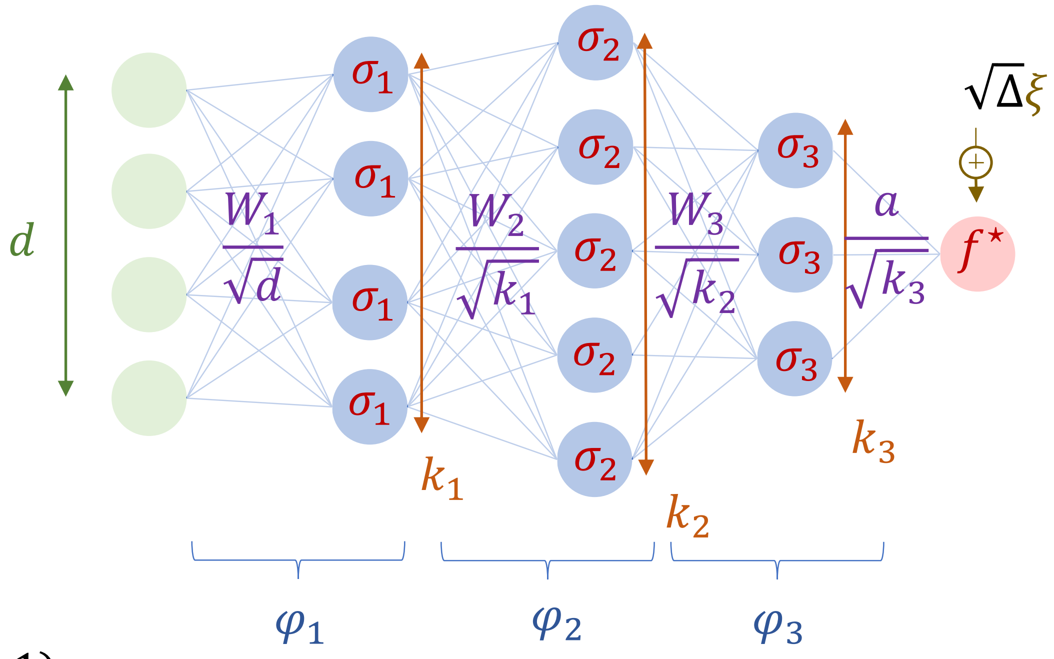
(Data) Gaussian data: $x \sim \mathcal{N}(0,\Sigma)$

(Target) $y^\star(x) = f^\star\left(\frac{a^\top}{\sqrt{k_L}}\varphi_L \circ \cdots \circ \varphi_1(x) + \sqrt{\Delta}\xi\right)$

with layers $\varphi_\ell(h) = \sigma_\ell\left(\frac{W_\ell}{\sqrt{k_{\ell-1}}}h\right)$

Odd activations σ_ℓ

$(W_\ell)_{ij} \sim \mathcal{N}(0,\Delta_\ell)$, $a_i \sim \mathcal{N}(0,\Delta_a)$, $\xi \sim \mathcal{N}(0,1)$



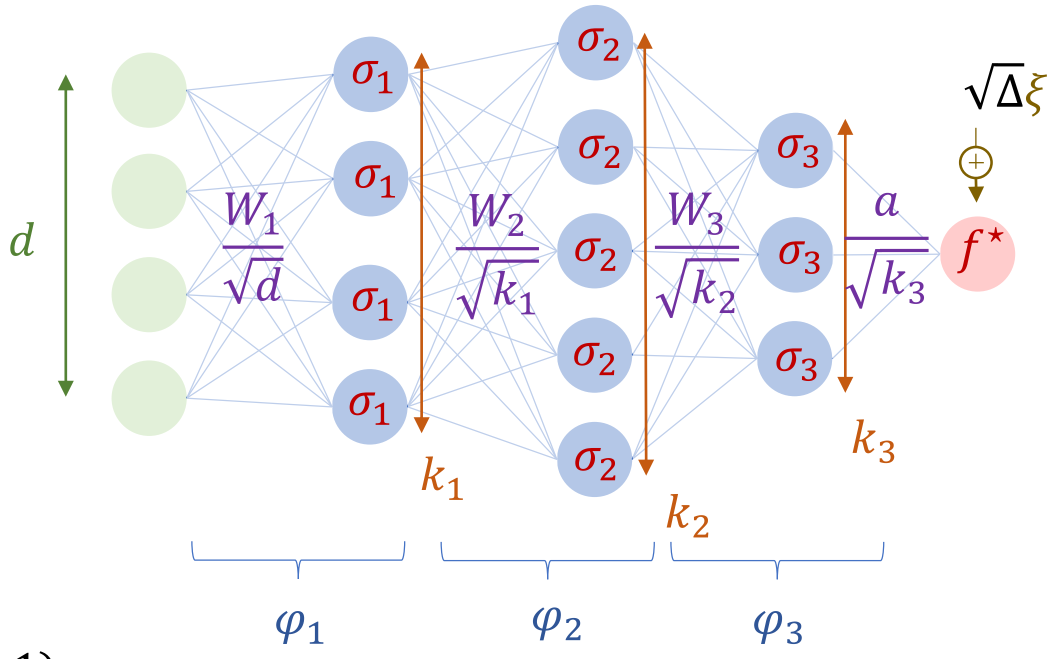
(Data) Gaussian data: $x \sim \mathcal{N}(0, \Sigma)$

(Target) $$y^\star(x) = f^\star\left(\frac{a^\top}{\sqrt{k_L}} \varphi_L \circ \cdots \circ \varphi_1(x) + \sqrt{\Delta}\xi\right)$$

with layers $$\varphi_\ell(h) = \sigma_\ell\left(\frac{W_\ell}{\sqrt{k_{\ell-1}}} h\right)$$

Odd activations σ_ℓ

$$(W_\ell)_{ij} \sim \mathcal{N}(0, \Delta_\ell), \; a_i \sim \mathcal{N}(0, \Delta_a), \; \xi \sim \mathcal{N}(0,1)$$



(Train set) Supervised learning with n i.i.d samples $\mathcal{D} = \{x^\mu, y^\star(x^\mu)\}_{\mu=1}^n$

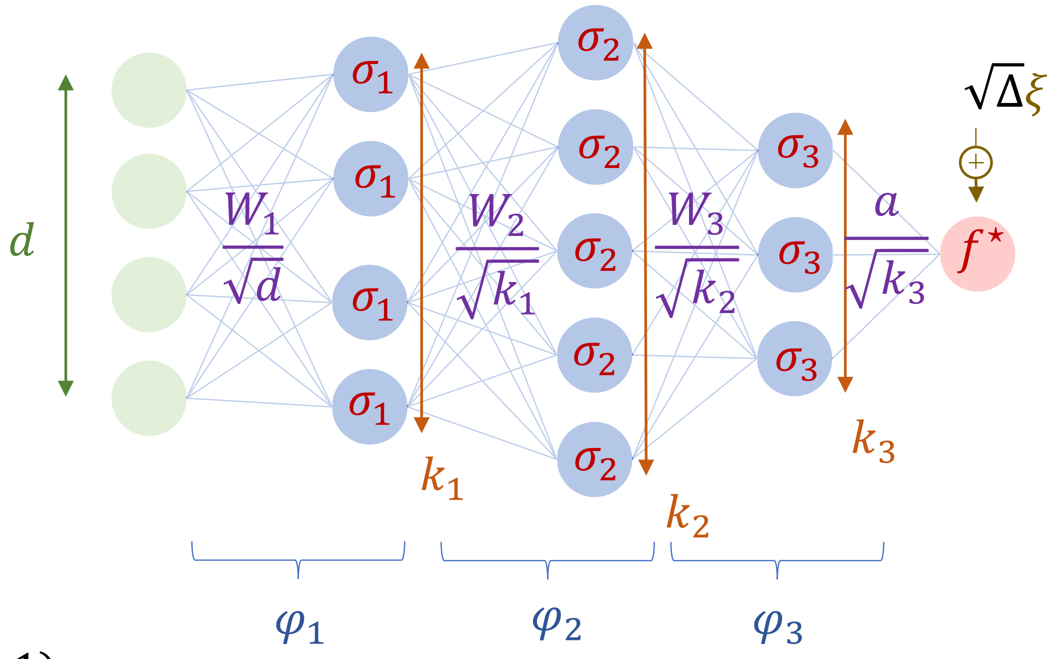
(Data) Gaussian data: $x \sim \mathcal{N}(0,\Sigma)$

(Target) $$y^\star(x) = f^\star\left(\frac{a^\top}{\sqrt{k_L}}\varphi_L \circ \cdots \circ \varphi_1(x) + \sqrt{\Delta}\xi\right)$$

with layers $$\varphi_\ell(h) = \sigma_\ell\left(\frac{W_\ell}{\sqrt{k_{\ell-1}}}h\right)$$

Odd activations σ_ℓ

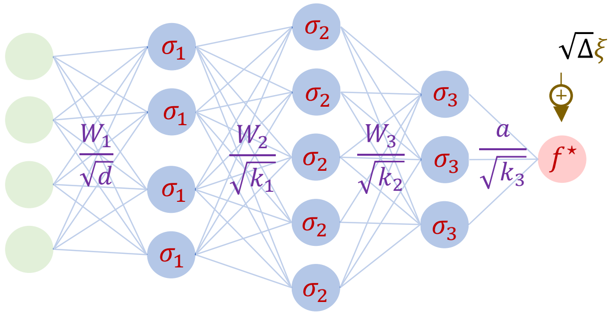
$$(W_\ell)_{ij} \sim \mathcal{N}(0,\Delta_\ell),\ \ a_i \sim \mathcal{N}(0,\Delta_a),\ \ \xi \sim \mathcal{N}(0,1)$$



(Train set) Supervised learning with n i.i.d samples $\mathcal{D} = \{x^\mu, y^\star(x^\mu)\}_{\mu=1}^n$

Proportional extensive-width limit

$n, d, k_1, \ldots, k_L \longrightarrow \infty$ with $\alpha = \frac{n}{d}, \gamma_\ell = \frac{k_\ell}{d} = \mathcal{O}(1)$

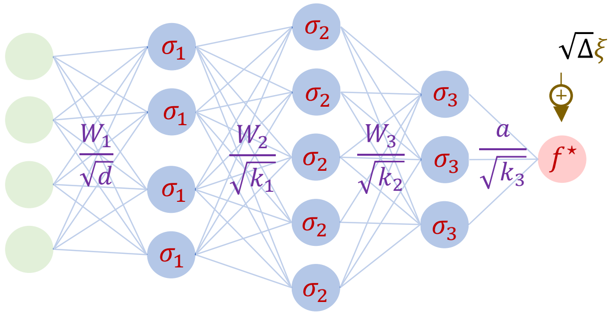


Suppose the architecture, priors, activations are known.
The best test error is then given by *Bayesian inference*:

Bayes posterior

$$\mathbb{P}\left(a, \{W_\ell\}_{\ell=1}^L \middle| \mathcal{D}\right) \propto e^{-\frac{||a||^2}{2\Delta_a} - \sum_{\ell=1}^L \frac{||W_\ell||_F^2}{2\Delta_\ell}} \times \prod_{\ell=1}^{L} \int \frac{d\xi e^{-\frac{\xi^2}{2}}}{\sqrt{2\pi}} \delta\left(y^\star(x^\mu) - f^\star\left(\frac{a^\top}{\sqrt{k_L}} \varphi_L \circ \cdots \circ \varphi_1(x) + \sqrt{\Delta}\xi \right)\right)$$

Suppose the architecture, priors, activations are known.
The best test error is then given by *Bayesian inference*:



Bayes posterior

$$\mathbb{P}\left(a, \{W_\ell\}_{\ell=1}^{L} \middle| \mathcal{D}\right) \propto e^{-\frac{||a||^2}{2\Delta_a} - \sum_{\ell=1}^{L} \frac{||W_\ell||_F^2}{2\Delta_\ell}} \times \prod_{\ell=1}^{L} \int \frac{d\xi e^{-\frac{\xi^2}{2}}}{\sqrt{2\pi}} \delta\left(y^\star(x^\mu) - f^\star\left(\frac{a^\top}{\sqrt{k_L}} \varphi_L \circ \cdots \circ \varphi_1(x) + \sqrt{\Delta}\xi\right)\right)$$

Regression ($f^\star = id$)

$$\epsilon_{g,\text{reg}}^{\text{BO}} = \mathbb{E}_{\mathcal{D},\{W_\ell^\star\}_{\ell=1}^{L},\mathbf{a}_\star} \mathbb{E}_{\mathbf{x},y}\left[\left(y - \langle \hat{y}(\mathbf{x})\rangle_{\mathbf{a},\{W_\ell\}_{\ell=1}^{L}\sim\mathbb{P}}\right)^2\right]$$

Classification ($f^\star = \text{sign}$)

$$\epsilon_{g,\text{class}}^{\text{BO}} = \mathbb{E}_{\mathcal{D},\{W_\ell^\star\}_{\ell=1}^{L},\mathbf{a}_\star} \mathbb{P}_{\mathbf{x},y}\left[y \neq \text{sign}\left(\langle \text{sign}(\hat{y}(\mathbf{x}))\rangle_{\mathbf{a},\{W_\ell\}_{\ell=1}^{L}\sim\mathbb{P}}\right)\right].$$

Q1. Can one provide a sharp asymptotic characterization of the Bayes-optimal error?

Q2. How do the test errors achieved by ERM algorithms in practice compare?

Outline

Preliminaries: Second-order statistics of random(-ish) neural nets

A1 Bayes-optimal test errors

A2 ERM test errors

HC, Krzakala and Zdeborová, *Optimal learning of random networks of extensive width,* arXiv:2302.00375 (2023).

Loureiro, Gerbelot, **HC**, Goldt, Krzakala, Mézard and Zdeborová, *Learning curves of generic feature maps for realistic datasets with a teacher-student model,* NeurIPS (2021).

Schröder, **HC,** Dmitriev and Loureiro, *Deterministic equivalent and error universality of deep random features learning,* arXiv:2302.00401 (2023).

Preliminaries: Second-order statistics of random(-ish) neural nets

Why second order statistics?

1. Appear naturally in the replica computation.

2. **Gaussian universality :** in a number of simple ERM settings, the test error only depends on the second order statistics of the data (*more later*)

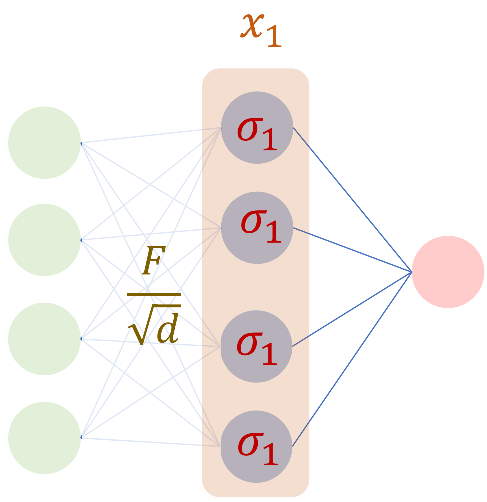
Song Mei and Andrea Montanari. *Generalization Error of Random Features Regression: Precise Asymptotics and the Double Descent Curve. Commun*. Pure Appl. Math.,, 2022

Hong Hu and Yue M. Lu. *Universality Laws for High-Dimensional Learning with Random Features*. IEE Trans. Inf. Theory

Federica Gerace, Bruno Loureiro, Florent Krzakala, Marc Mezard, and Lenka Zdeborova. *Generalisation error in learning with random features and the hidden manifold model*. ICML 2020

The shallow L=1 hidden layer case

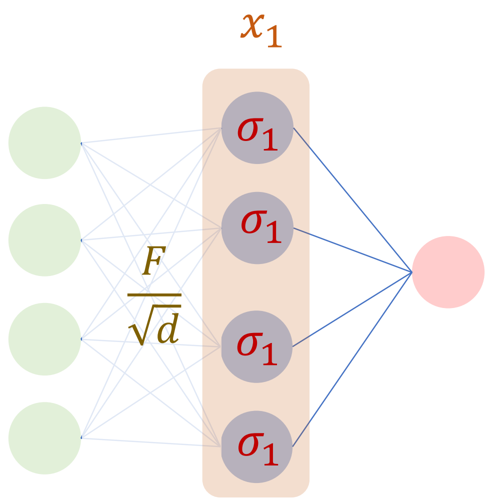
$x \sim \mathcal{N}(0, \Sigma)$



For fixed F, what is the covariance $\Omega = \langle x_1 x_1^\top \rangle_x$ *of the last layer post-activation wrt the Gaussian input randomness?*

The shallow L=1 hidden layer case

$x \sim \mathcal{N}(0, \Sigma)$



For fixed F, what is the covariance $\Omega = \langle x_1 x_1^\top \rangle_x$ *of the last layer post-activation wrt the Gaussian input randomness?*

(Gaussian Equivalence Property)

Defining

$$\kappa_1 = \mathbb{E}_{z\sim\mathcal{N}(0,1)}[\sigma_1(z)z]$$

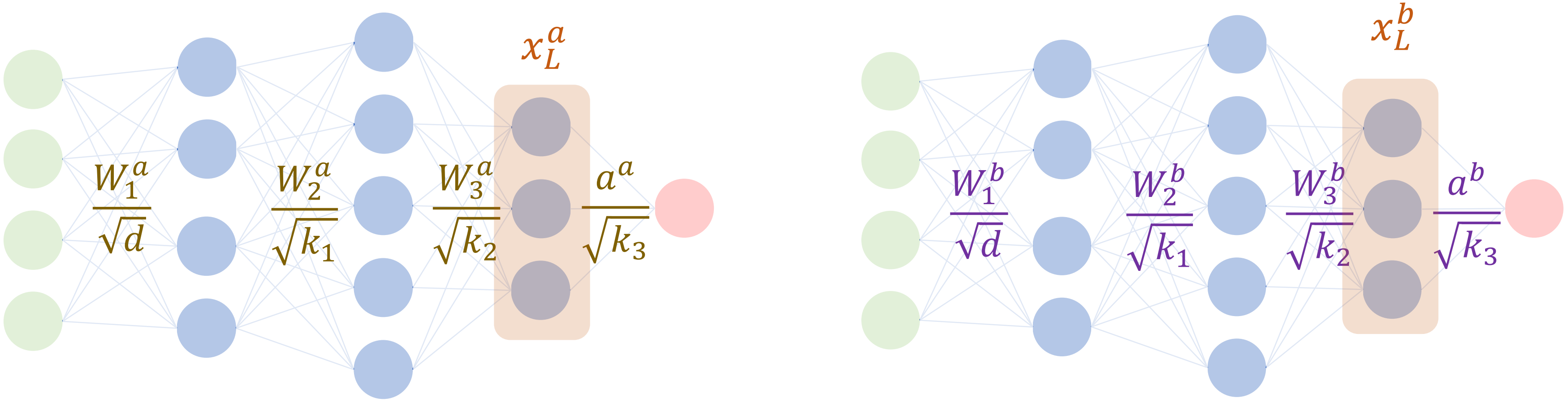
$$\kappa_* = \sqrt{\mathbb{E}_{z\sim\mathcal{N}(0,1)}[\sigma_1(z)^2] - \kappa_1^2}$$

then simply

$$\Omega = \kappa_1^2 \frac{F\Sigma F^\top}{d} + \kappa_*^2 \mathbb{I}_k$$

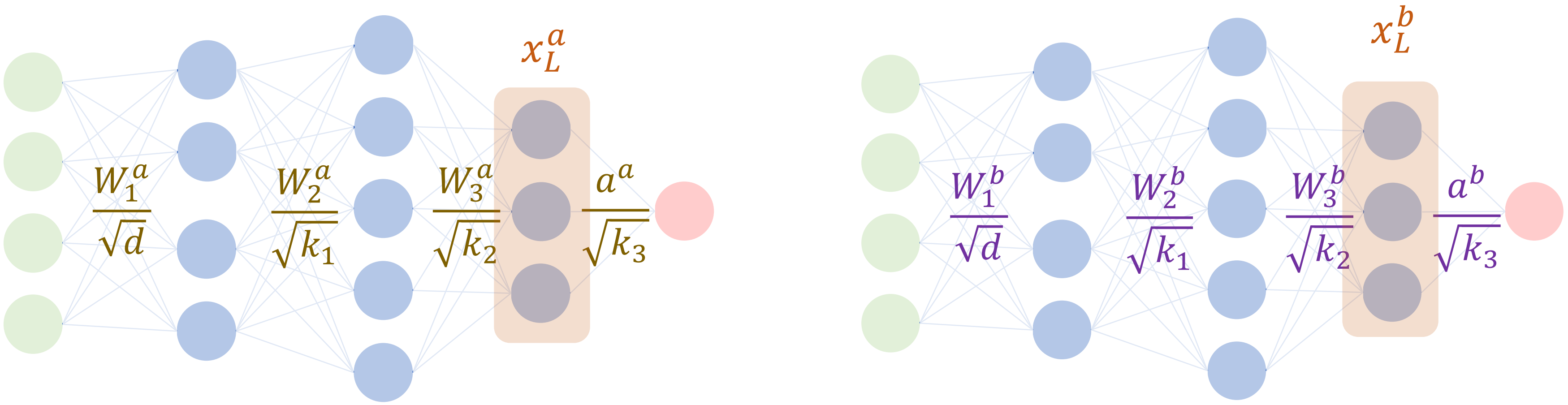
Goldt, Mézard, Krzakala, and Zdeborová, *Modelling the influence of data structure on learning in neural networks*. PRX 2020

Draw two networks $W_1^a, \dots, W_L^a, a^a$ and $W_1^b, \dots, W_L^b, a^b$ i.i.d from the Bayes posterior.



What is the covariance $\Omega_L^{ab} = \left\langle x_L^a x_L^{b\top} \right\rangle_x$?

Draw two networks $W_1^a, \dots, W_L^a, a^a$ and $W_1^b, \dots, W_L^b, a^b$ i.i.d from the Bayes posterior.



What is the covariance $\Omega_L^{ab} = \langle x_L^a x_L^{b\top} \rangle_x$?

(Deep Bayes GEP)

Defining

$$r_{\ell+1} = \Delta_{\ell+1} \mathbb{E}_{z\sim\mathcal{N}(0,r_\ell)} \left[\sigma_\ell(z)^2\right],$$

$$\kappa_1^{(\ell)} = \frac{1}{r_\ell} \mathbb{E}_{z\sim\mathcal{N}(0,r_\ell)} \left[z\sigma_\ell(z)\right],$$

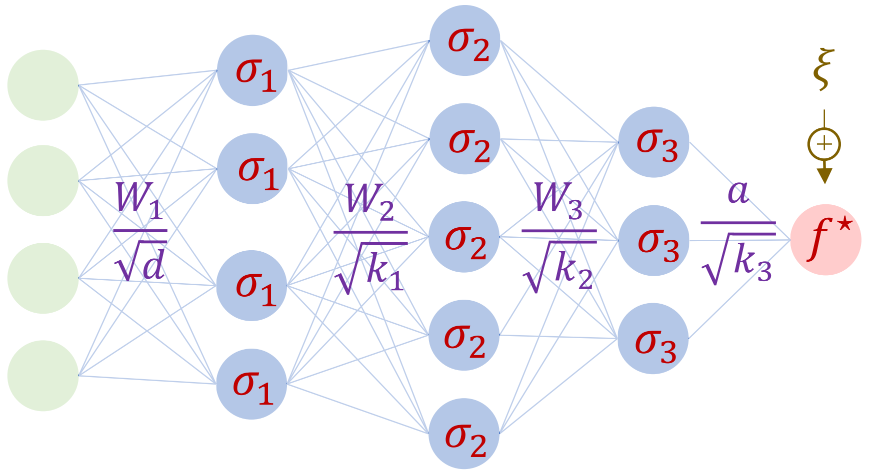
$$\kappa_*^{(\ell)} = \sqrt{\mathbb{E}_{z\sim\mathcal{N}(0,r_\ell)} \left[\sigma_\ell(z)^2\right] - r_\ell \left(\kappa_1^{(\ell)}\right)^2},$$

Ω_L^{ab} is given by the Lth term of the recursion

$$\Omega_\ell^{ab} = \left(\kappa_1^{(\ell)}\right)^2 \frac{W_\ell^a \Omega_{\ell-1}^{ab} W_\ell^{b\top}}{k_{\ell-1}} + \delta_{ab} \left(\kappa_*^{(\ell)}\right)^2 \mathbb{I}_{k_\ell}$$

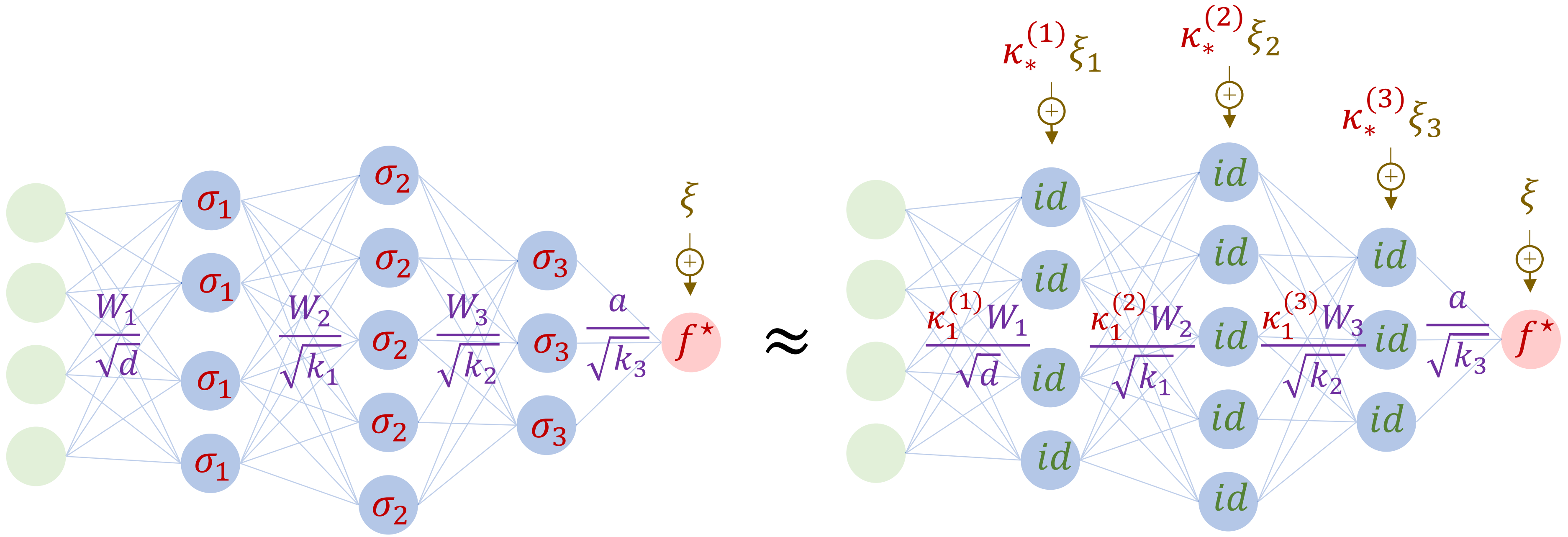
HC, Krzakala and Zdeborová, *Optimal learning of random networks of extensive width*, arXiv:2302.00375 (2023).

In terms of second-order activation statistics,



Non-linear deep network

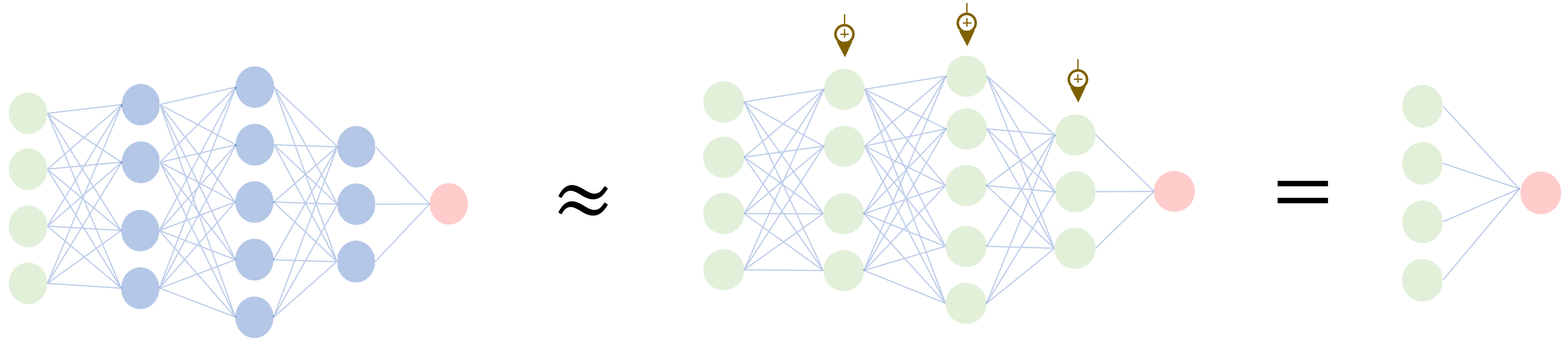
In terms of second-order activation statistics,



Non-linear deep network

Noisy, linear deep network

Q1. Can one provide a sharp asymptotic characterization of the Bayes-optimal error?



$$y^\star(x) = f^\star\left(\frac{a^\top}{\sqrt{k_L}}\varphi_L \circ \cdots \circ \varphi_1(x) + \sqrt{\Delta}\mathcal{N}(0,1)\right)$$

With layers $\varphi_\ell(h) = \sigma_\ell\left(\frac{W_\ell}{\sqrt{k_{\ell-1}}}\, h\right)$

$(W_\ell)_{ij} \sim \mathcal{N}(0, \Delta_\ell)$, $a_i \sim \mathcal{N}(0, \Delta_a)$

$$y^{\mathrm{eq}}(x) = f^\star\left(\rho\frac{\theta^\top x}{\sqrt{d}} + \epsilon_r \mathcal{N}(0,1)\right)$$

With

$$\epsilon_r \equiv \sum_{\ell_0=1}^{L-1} \left(\kappa_*^{(\ell_0)}\right)^2 \Delta_a \prod_{\ell=\ell_0+1}^{L} \left(\kappa_1^{(\ell)}\right)^2 \Delta_\ell + \left(\kappa_*^{(L)}\right)^2 \Delta_a + \Delta$$

$$\rho \equiv \Delta_a \prod_{\ell=1}^{L} \left(\kappa_1^{(\ell)}\right)^2 \Delta_\ell$$

$\theta_i \sim \mathcal{N}(0,1)$

Can be shown to be characterized by the same replica free entropy, and the ***same Bayes optimal errors***

Regression

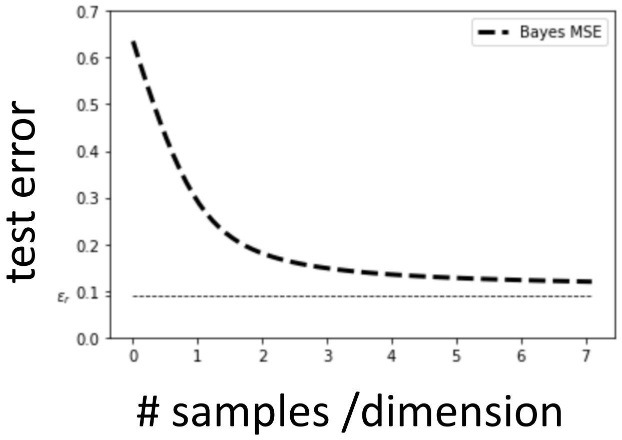
$$\epsilon_{g,\mathrm{reg}}^{\mathrm{BO}}=\prod_{\ell=1}^{L}(\kappa_1^{(\ell)})^2\Big(\Delta_a\Big(\int z\mathrm{d}\mu(z)\Big)\prod_{\ell=1}^{L}\Delta_\ell-q\Big)+\epsilon_r \qquad q=\frac{1}{2}\int\frac{\alpha\prod_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2 z^2\Delta_a^2\prod_{\ell=1}^{L}\Delta_\ell^2}{\epsilon_{g,\mathrm{reg}}^{\mathrm{BO}}+\alpha\prod_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2 z\Delta_a\prod_{\ell=1}^{L}\Delta_\ell}\mathrm{d}\mu(z).$$

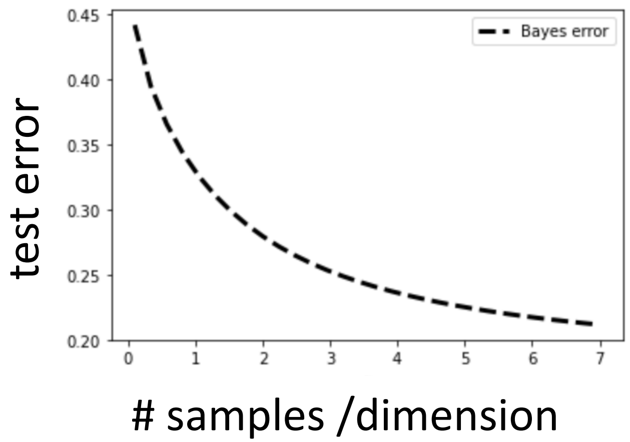
Classification

$$\epsilon_{g,\mathrm{class}}^{\mathrm{BO}}=\frac{1}{\pi}\arccos\left[\frac{\sqrt{\prod_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2 q}}{\sqrt{\Delta_a\int z\mathrm{d}\mu(z)\prod_{\ell=1}^{L}(\kappa_1^{(\ell)})^2\Delta_\ell+\epsilon_r}}\right]$$

$$\begin{cases} q=\int\frac{\hat{q}\Delta_a^2\prod_{\ell=1}^{L}\Delta_\ell^2 z^2}{\hat{q}z\Delta_a\prod_{\ell=1}^{L}\Delta_\ell+1}\mathrm{d}\mu(z) \\ \hat{q}=\frac{2\alpha\prod_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2}{\Delta_a\int z\mathrm{d}\mu(z)\prod_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2\Delta_\ell+\epsilon_r-\prod_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2 q} \\ \int\frac{d\xi}{(2\pi)^{\frac{3}{2}}}\frac{2e^{-\frac{1}{2}\frac{\Delta_a\int z\mathrm{d}\mu(z)\prod_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2\Delta_\ell+\epsilon_r+\prod_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2 q}{\Delta_a\int z\mathrm{d}\mu(z)\prod_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2\Delta_\ell+\epsilon_r-\prod_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2 q}\xi^2}}{1-\mathrm{erf}\left(\frac{\prod_{\ell=1}^{L}\kappa_1^{(\ell)}\sqrt{q}\xi}{\sqrt{2\left(\Delta_a\int z\mathrm{d}\mu(z)\prod_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2\Delta_\ell+\epsilon_r-\prod_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2 q\right)}}\right)} \end{cases}$$

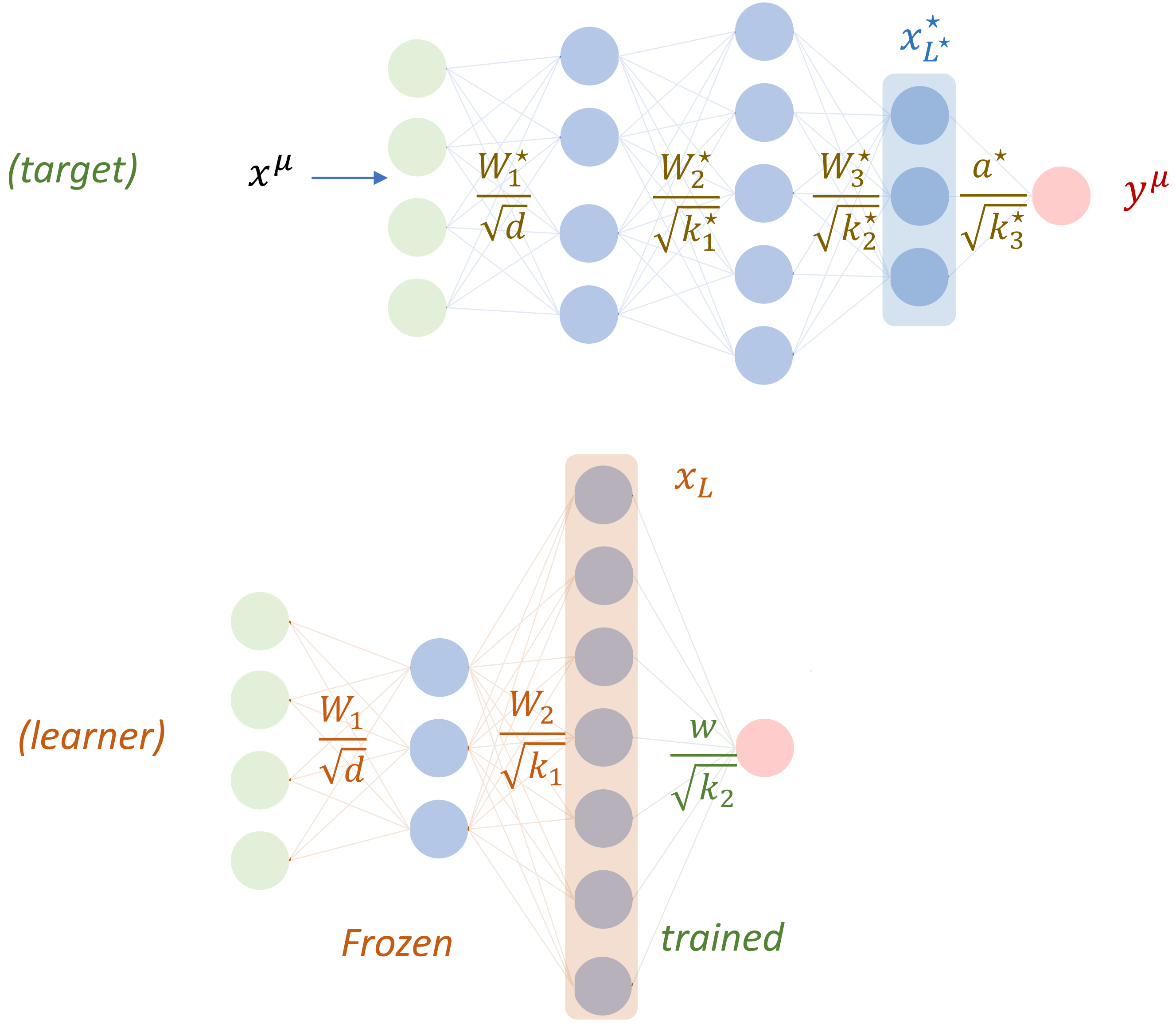
depth $=3, \sigma=tanh$



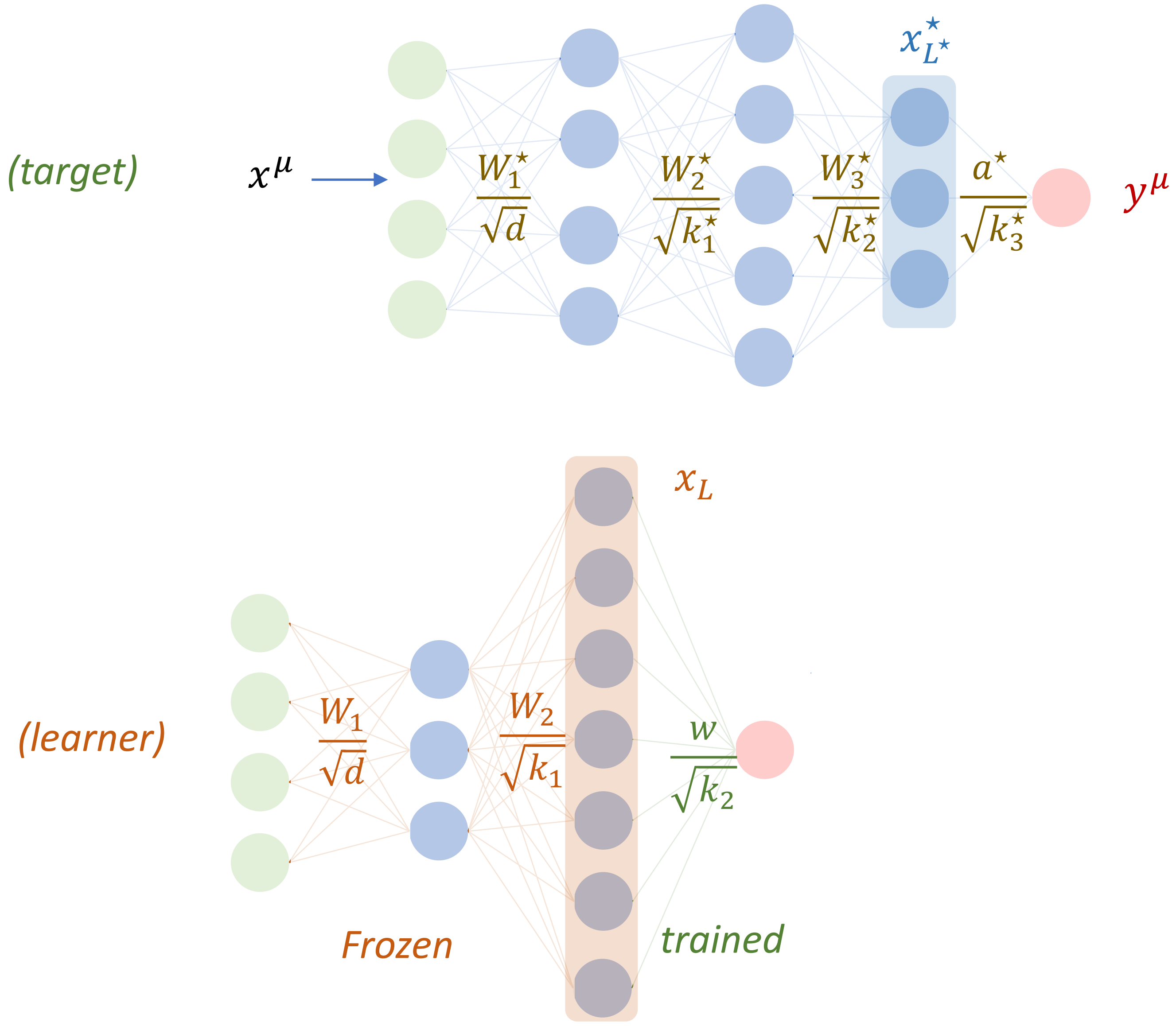


✓ **Q1**. Can one provide a sharp asymptotic characterization of the Bayes-optimal error?

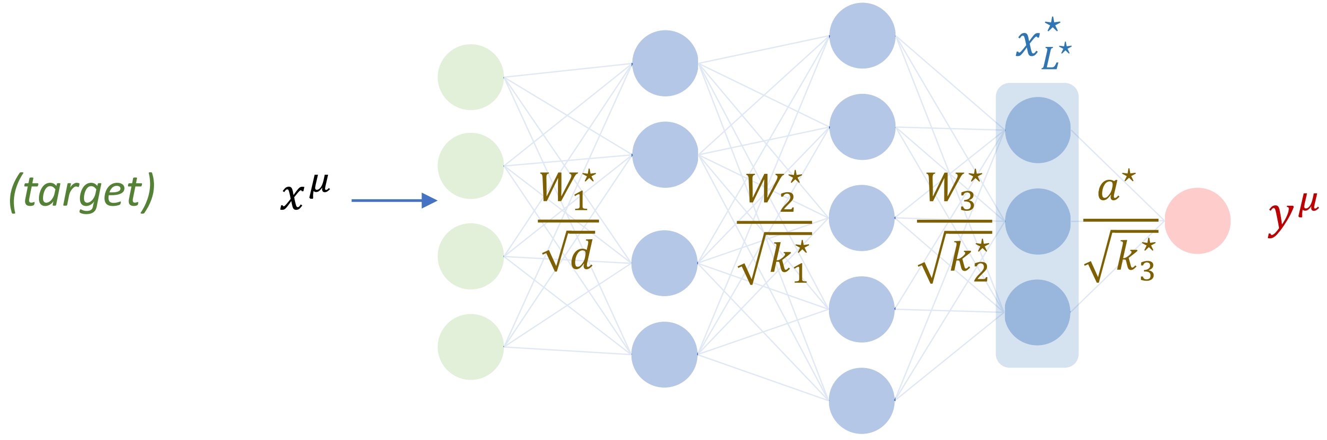
Q2. How do the test errors achieved by ERM algorithms in practice compare?

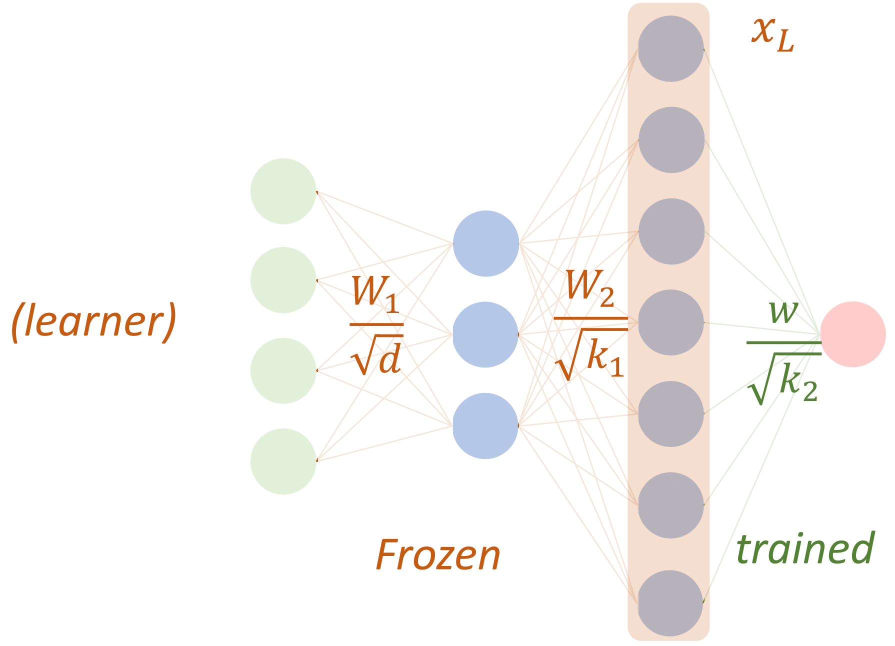


(target)
x^μ
$\frac{W_1^\star}{\sqrt{d}}$
$\frac{W_2^\star}{\sqrt{k_1^\star}}$
$\frac{W_3^\star}{\sqrt{k_2^\star}}$
$x_{L^\star}^\star$
$\frac{a^\star}{\sqrt{k_3^\star}}$
y^μ
(learner)
$\frac{W_1}{\sqrt{d}}$
$\frac{W_2}{\sqrt{k_1}}$
x_L
$\frac{w}{\sqrt{k_2}}$
Frozen
trained

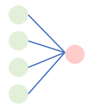


$$\hat{w} = \underset{w}{\operatorname{argmin}} \left(\sum_{\mu=1}^{n} g\left(y^\mu, \frac{w^\top x_L^\mu}{\sqrt{k_L}} \right) + r(w) \right)$$

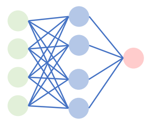




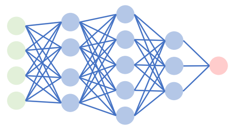
$$\hat{w} = \underset{w}{\operatorname{argmin}} \left(\sum_{\mu=1}^{n} g\left(y^\mu, \frac{w^\top x_L^\mu}{\sqrt{k_L}}\right) + r(w) \right)$$



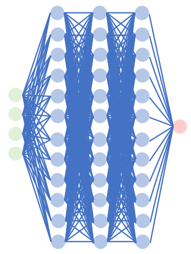
- Ridge, LASSO, elastic net…
- Logistic / hinge/ ridge classification



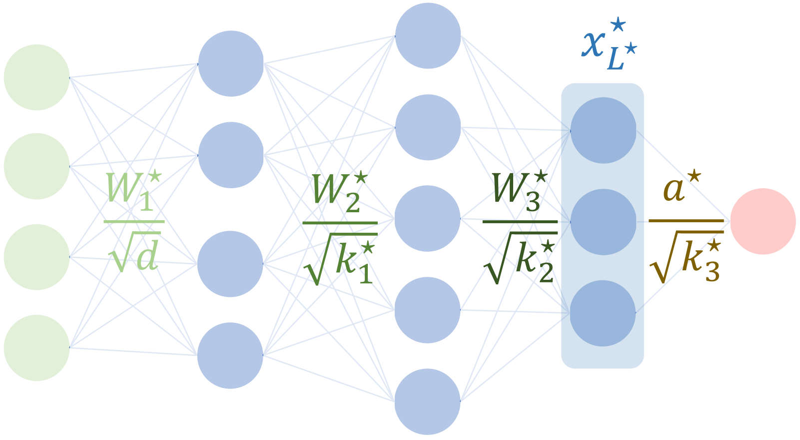
- Random Features



- Deep Random Features

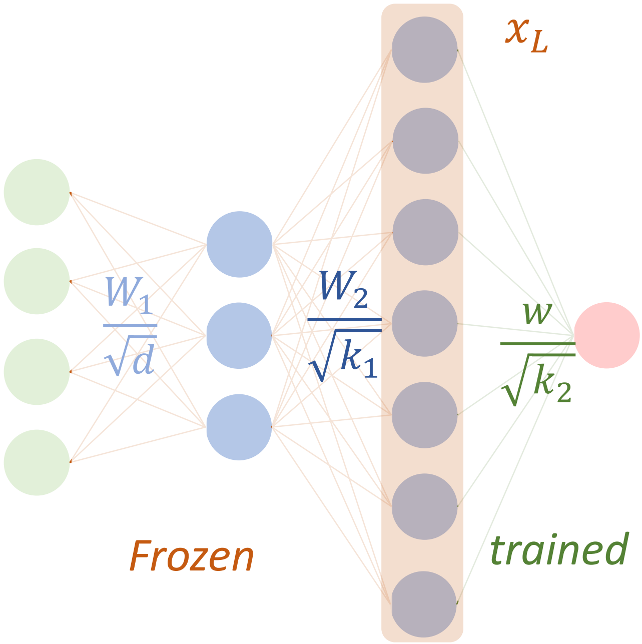


- Kernel regression/classification

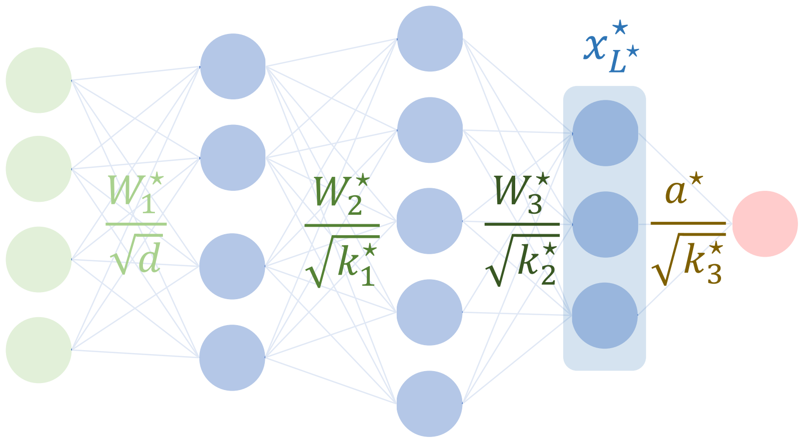


Introduce the ***Gaussian clones*** u, v of $x_L, x^\star_{L^\star}$

$$u, v \sim \mathcal{N}\left(0, \begin{bmatrix} \langle x_L x_L^\top \rangle & \langle x_L x^{\star\,\top}_{L^\star} \rangle \\ \langle x^\star_{L^\star} x_L^\top \rangle & \langle x^\star_{L^\star} x^{\star\,\top}_{L^\star} \rangle \end{bmatrix}\right)$$



Schröder, **HC,** Dmitriev and Loureiro, *Deterministic equivalent and error universality of deep random features learning,* arXiv:2302.00401 (2023).

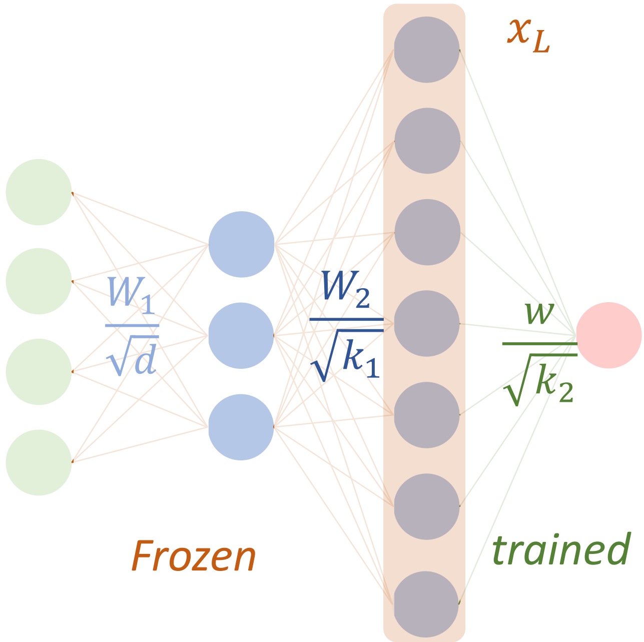


Introduce the ***Gaussian clones*** u, v of $x_L, x^\star_{L^\star}$

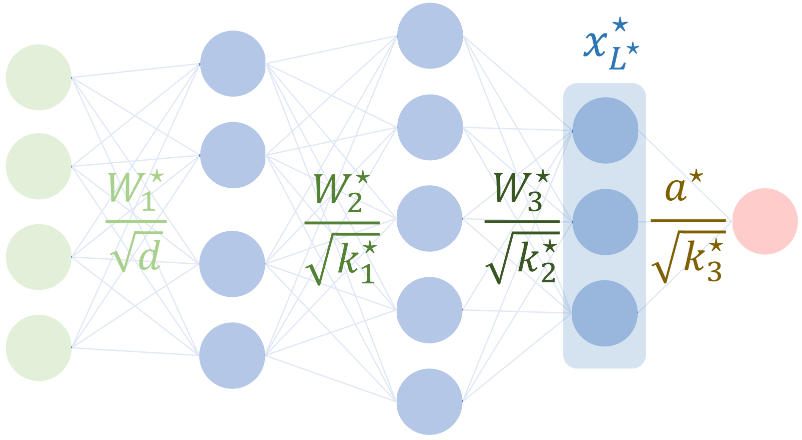
$$u, v \sim \mathcal{N}\left(0, \begin{bmatrix} \langle x_L x_L^\top \rangle & \langle x_L {x^\star_{L^\star}}^\top \rangle \\ \langle x^\star_{L^\star} x_L^\top \rangle & \langle x^\star_{L^\star} {x^\star_{L^\star}}^\top \rangle \end{bmatrix}\right)$$

(ERM) $$\mathcal{D} = \left\{x^\mu, y^\mu = f^\star\left(\frac{a_\star^\top {x^\star_{L^\star}}^\mu}{\sqrt{k^\star_{L^\star}}}\right)\right\} \quad \widehat{w} = \underset{w}{\mathrm{argmin}}\left(\sum_{\mu=1}^n g\left(y^\mu, \frac{w^\top x_L^\mu}{\sqrt{k_L}}\right) + r(w)\right)$$

(ERMg) $$\mathcal{D}^G = \left\{u^\mu, y^\mu = f^\star\left(\frac{a_\star^\top v^\mu}{\sqrt{k^\star_{L^\star}}}\right)\right\} \quad \widehat{w} = \underset{w}{\mathrm{argmin}}\left(\sum_{\mu=1}^n g\left(y^\mu, \frac{w^\top u^\mu}{\sqrt{k_L}}\right) + r(w)\right)$$



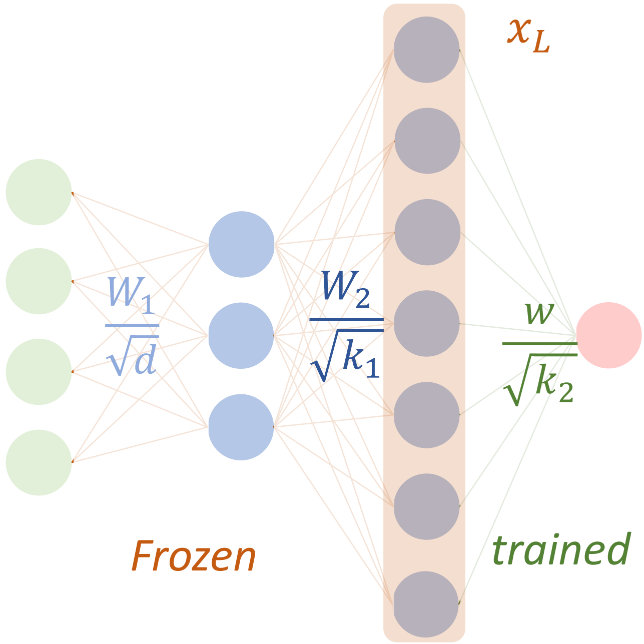
Schröder, **HC,** Dmitriev and Loureiro, *Deterministic equivalent and error universality of deep random features learning,* arXiv:2302.00401 (2023).



Introduce the ***Gaussian clones*** u, v of $x_L, x_{L^\star}^\star$

$$u, v \sim \mathcal{N}\left(0, \begin{bmatrix} \langle x_L x_L^\top \rangle & \langle x_L {x_{L^\star}^\star}^\top \rangle \\ \langle x_{L^\star}^\star x_L^\top \rangle & \langle x_{L^\star}^\star {x_{L^\star}^\star}^\top \rangle \end{bmatrix}\right)$$

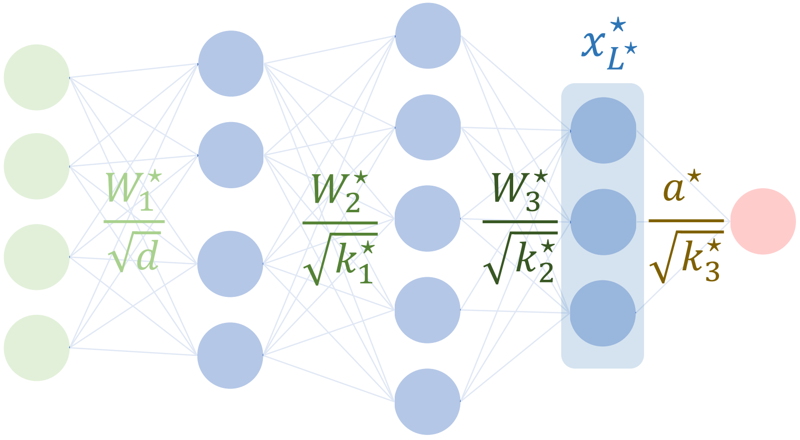
(ERM) $$\mathcal{D} = \left\{x^\mu, y^\mu = f^\star\left(\frac{a_\star^\top {x_{L^\star}^\star}^\mu}{\sqrt{k_{L^\star}^\star}}\right)\right\} \quad \hat{w} = \underset{w}{\mathrm{argmin}}\left(\sum_{\mu=1}^{n} g\left(y^\mu, \frac{w^\top x_L^\mu}{\sqrt{k_L}}\right) + r(w)\right)$$



(ERMg) $$\mathcal{D}^G = \left\{u^\mu, y^\mu = f^\star\left(\frac{a_\star^\top v^\mu}{\sqrt{k_{L^\star}^\star}}\right)\right\} \quad \hat{w} = \underset{w}{\mathrm{argmin}}\left(\sum_{\mu=1}^{n} g\left(y^\mu, \frac{w^\top u^\mu}{\sqrt{k_L}}\right) + r(w)\right)$$

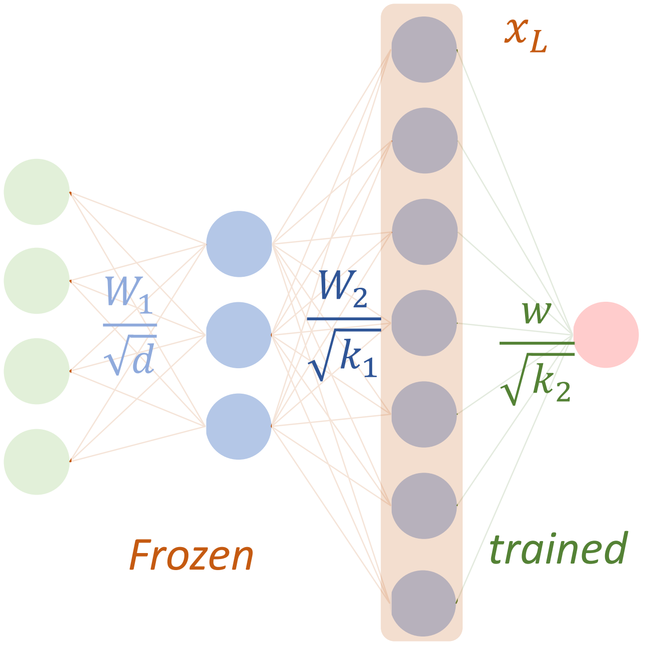
Conjecture: (part 1) **(Gaussian universality)** The learning problems (ERM) and (ERMg) lead to the same test error and training loss.

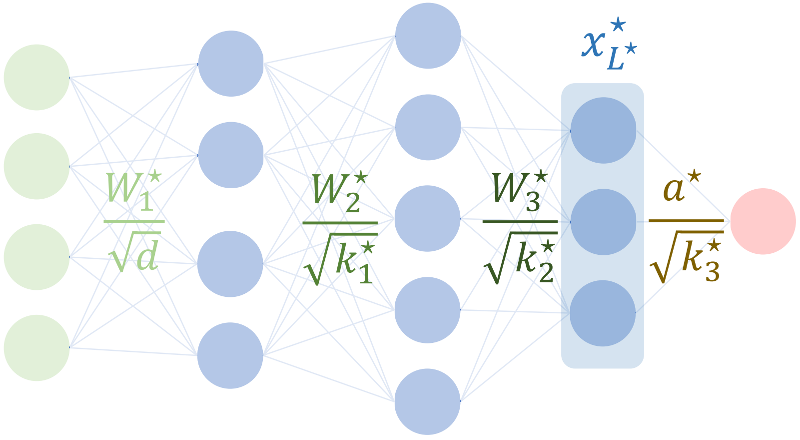
Schröder, **HC,** Dmitriev and Loureiro, *Deterministic equivalent and error universality of deep random features learning*, arXiv:2302.00401 (2023).



$$u, v \sim \mathcal{N}\left(0, \begin{bmatrix} \langle x_L x_L^\top \rangle & \langle x_L {x^\star_{L^\star}}^\top \rangle \\ \langle x^\star_{L^\star} x_L^\top \rangle & \langle x^\star_{L^\star} {x^\star_{L^\star}}^\top \rangle \end{bmatrix}\right)$$

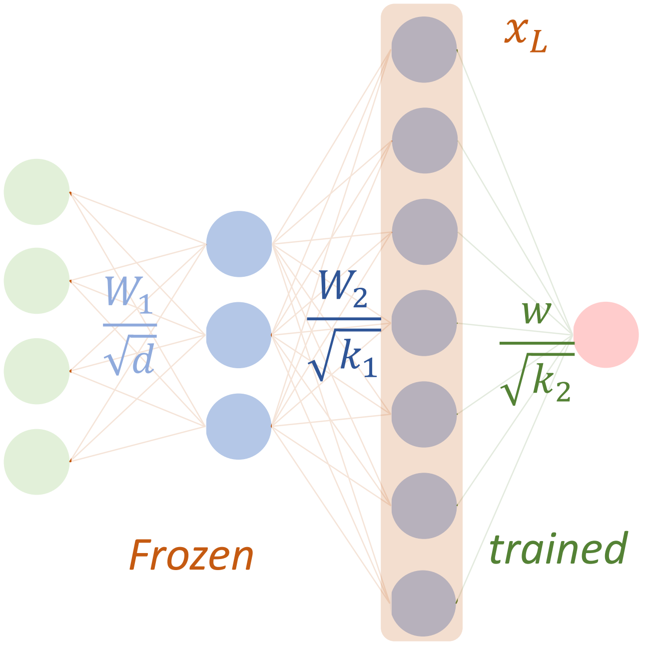
Conjecture: (part 2) Furthermore, the covariances $\langle x_L x_L^\top \rangle$, $\langle x^\star_{L^\star} {x^\star_{L^\star}}^\top \rangle$ and $\langle x^\star_{L^\star} x_L^\top \rangle$ can be computed simply with the noisy equivalent model.





$$u, v \sim \mathcal{N}\left(0, \begin{bmatrix} \langle x_L x_L^\top \rangle & \langle x_L x_{L^\star}^{\star\top} \rangle \\ \langle x_{L^\star}^\star x_L^\top \rangle & \langle x_{L^\star}^\star x_{L^\star}^{\star\top} \rangle \end{bmatrix}\right)$$

Conjecture: (part 2) Furthermore, the covariances $\langle x_L x_L^\top \rangle$, $\langle x_{L^\star}^\star x_{L^\star}^{\star\top} \rangle$ and $\langle x_{L^\star}^\star x_L^\top \rangle$ can be computed simply with the noisy equivalent model.



Here for instance

$$\langle x_L x_L^\top \rangle = {\kappa_1^{(1)}}^2 {\kappa_1^{(2)}}^2 \frac{W_2 W_1 \Sigma W_1^\top W_2^\top}{d k_1} + {\kappa_*^{(1)}}^2 {\kappa_1^{(2)}}^2 \frac{W_2 W_2^\top}{k_1} + {\kappa_*^{(2)}}^2 \mathbb{I}_{k_1}$$

$$\langle x_{L^\star}^\star x_{L^\star}^{\star\top} \rangle = {\kappa_1^{\star(1)}}^2 {\kappa_1^{\star(2)}}^2 {\kappa_1^{\star(3)}}^2 \frac{W_3^\star W_2^\star W_1^\star \Sigma \; W_1^{\star\top} W_2^{\star\top} W_3^{\star\top}}{d k_1^\star k_2^\star} + {\kappa_*^{\star(1)}}^2 {\kappa_1^{\star(2)}}^2 {\kappa_1^{\star(3)}}^2 \frac{W_3^\star W_2^\star W_2^{\star\top} W_3^{\star\top}}{k_1^\star k_2^\star} + {\kappa_*^{\star(2)}}^2 {\kappa_1^{\star(3)}}^2 \frac{W_3^\star W_3^{\star\top}}{k_2^\star} + {\kappa_*^{\star\,(2)}}^2 \mathbb{I}_{k_2^\star}$$

$$\langle x_{L^\star}^\star x_L^\top \rangle = {\kappa_1}^{(1)} {\kappa_1}^{(2)} \; {\kappa_1}^{\star(1)} \; {\kappa_1}^{\star(2)} \; {\kappa_1}^{\star(3)} \frac{W_3^\star W_2^\star W_1^\star \Sigma \; W_1^\top W_2^\top}{d\sqrt{k_1 k_1^\star k_2^\star}}$$

So one just needs to solve the proxy ERM

$$u, v \sim \mathcal{N}\left(0, \begin{bmatrix} \langle x_L x_L^\top \rangle & \langle x_L x_{L^\star}^{\star\top} \rangle \\ \langle x_{L^\star}^\star x_L^\top \rangle & \langle x_{L^\star}^\star x_{L^\star}^{\star\top} \rangle \end{bmatrix}\right)$$

(ERMg)

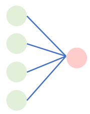
$$\mathcal{D}^G = \left\{ u^\mu, y^\mu = f^\star\left(\frac{a_\star^\top v^\mu}{\sqrt{k_{L^\star}^\star}}\right) \right\}$$

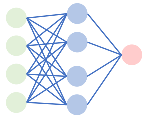
$$\hat{w} = \underset{w}{\operatorname{argmin}} \left(\sum_{\mu=1}^{n} g\left(y^\mu, \frac{w^\top u^\mu}{\sqrt{k_L}}\right) + r(w) \right)$$

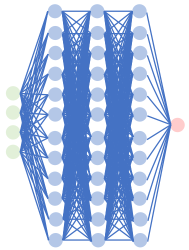
Theorem (informal) : The test error of the problem (ERMg) can be characterized in terms of three order parameters q, m, V given as the solution of a system of self-consistent equations.

$$\begin{cases} V = \mathbb{E}_{(\omega,\bar{\theta})\sim\mu}\left[\frac{\omega}{\lambda+\hat{V}\omega}\right] \\ m = \frac{\hat{m}}{\sqrt{\gamma}}\mathbb{E}_{(\omega,\bar{\theta})\sim\mu}\left[\frac{\bar{\theta}^2}{\lambda+\hat{V}\omega}\right] \\ q = \mathbb{E}_{(\omega,\bar{\theta})\sim\mu}\left[\frac{\hat{m}^2\bar{\theta}^2\omega+\hat{q}\omega^2}{(\lambda+\hat{V}\omega)^2}\right] \end{cases}, \qquad \begin{cases} \hat{V} = \frac{\alpha}{V}(1-\mathbb{E}_{s,h\sim\mathcal{N}(0,1)}[f_g'(V,m,q)]) \\ \hat{m} = \frac{1}{\sqrt{\rho\gamma}}\frac{\alpha}{V}\mathbb{E}_{s,h\sim\mathcal{N}(0,1)}\left[s f_g(V,m,q) - \frac{m}{\sqrt{\rho}} f_g'(V,m,q)\right] \\ \hat{q} = \frac{\alpha}{V^2}\mathbb{E}_{s,h\sim\mathcal{N}(0,1)}\left[\left(\frac{m}{\sqrt{\rho}}s + \sqrt{q-\frac{m^2}{\rho}}h - f_g(V,m,q)\right)^2\right] \end{cases}$$

Loureiro, Gerbelot, **HC**, Goldt, Krzakala, Mézard and Zdeborová, *Learning curves of generic feature maps for realistic datasets with a teacher-student model,* NeurIPS 2021







$$\epsilon_g = \rho \int z \mathrm{d}\mu(z) + q - 2 \prod_{\ell=1}^{L} \kappa_1^{(\ell)} m + \epsilon_r$$

$$\begin{cases} \hat{V} = \frac{\alpha}{1+V} \\ \hat{q} = \alpha \frac{\epsilon_g}{(1+V)^2} \\ \hat{m} = \frac{\prod_{\ell=1}^{L} \kappa_1^{(\ell)} \alpha}{1+V} \end{cases} \quad \begin{cases} V = \int \frac{z}{\lambda + \hat{V} z} \mathrm{d}\mu(z) \\ q = \int \frac{\Delta_a \prod_{\ell=1}^{L} \Delta_\ell \hat{m}^2 z^3 + \hat{q} z^2}{(\lambda + \hat{V} z)^2} \mathrm{d}\mu(z) \\ m = \Delta_a \prod_{\ell=1}^{L} \Delta_\ell \hat{m} \int \frac{z^2}{\lambda + \hat{V} z} \mathrm{d}\mu(z) \end{cases}$$

$$\epsilon_g = \rho \int z \mathrm{d}\mu(z) + q - 2 \prod_{\ell=1}^{L} \kappa_1^{(\ell)} m + \epsilon_r$$

$$\begin{cases} \hat{V} = \frac{\frac{\alpha}{\gamma}}{1+V} \\ \hat{q} = \frac{\alpha}{\gamma} \frac{\epsilon_g}{(1+V)^2} \\ \hat{m} = \sqrt{\Delta_a \prod_{\ell=1}^{L} \Delta_\ell} \sqrt{\gamma} \frac{\prod_{\ell=1}^{L} \kappa_1^{(\ell)} \frac{\alpha}{\gamma}}{1+V} \end{cases}$$

$$\begin{cases} V = \frac{1}{\hat{V}} - \frac{\lambda}{\hat{V}^2 \kappa_1^2} g\left(-\frac{\lambda + \hat{V}\kappa_*^2}{\hat{V}\kappa_1^2}\right) \\ q = \frac{\hat{m}^2 + \hat{q}}{\hat{V}^2} - \frac{1}{\kappa_1^2 \hat{V}^2} \left(\frac{2\lambda(\hat{m}^2 + \hat{q})}{\hat{V}} + \hat{m}^2 \kappa_*^2\right) g\left(-\frac{\lambda + \hat{V}\kappa_*^2}{\hat{V}\kappa_1^2}\right) \\ \qquad + \frac{\lambda}{\kappa_1^4 \hat{V}^3} \left(\frac{\lambda(\hat{m}^2 + \hat{q})}{\hat{V}} + \hat{m}^2 \kappa_*^2\right) g'\left(-\frac{\lambda + \hat{V}\kappa_*^2}{\hat{V}\kappa_1^2}\right) \\ m = \sqrt{\gamma} \frac{\hat{m}}{\hat{V}} \left[1 - \frac{1}{\kappa_1^2} \left(\frac{\lambda}{\hat{V}} + \kappa_*^2\right) g\left(-\frac{\lambda + \hat{V}\kappa_*^2}{\hat{V}\kappa_1^2}\right)\right]. \end{cases}$$

$$\epsilon_g = \rho \int z \mathrm{d}\mu(z) + q - 2 \prod_{\ell=1}^{L} \kappa_1^{(\ell)} m + \epsilon_r$$

$$\begin{cases} \hat{V} = \frac{\alpha}{1+V} \\ \hat{q} = \alpha \frac{\epsilon_g}{(1+V)^2} \\ \hat{m} = \alpha \frac{\prod_{\ell=1}^{L} \kappa_1^{(\ell)}}{1+V} \end{cases} \quad \begin{cases} V = \frac{\kappa_*^2}{\lambda} + \frac{\kappa_1^2}{\lambda + \hat{V}\kappa_1^2} \\ q = \frac{\Delta_a \prod_{\ell=1}^{L} \Delta_\ell \hat{m}^2 \kappa_1^4 + \hat{q}\kappa_1^4}{(\lambda + \hat{V}\kappa_1^2)^2} \\ m = \Delta_a \prod_{\ell=1}^{L^\star} \Delta_\ell \hat{m} \frac{\kappa_1^2}{\lambda + \hat{V}\kappa_1^2} \end{cases}$$

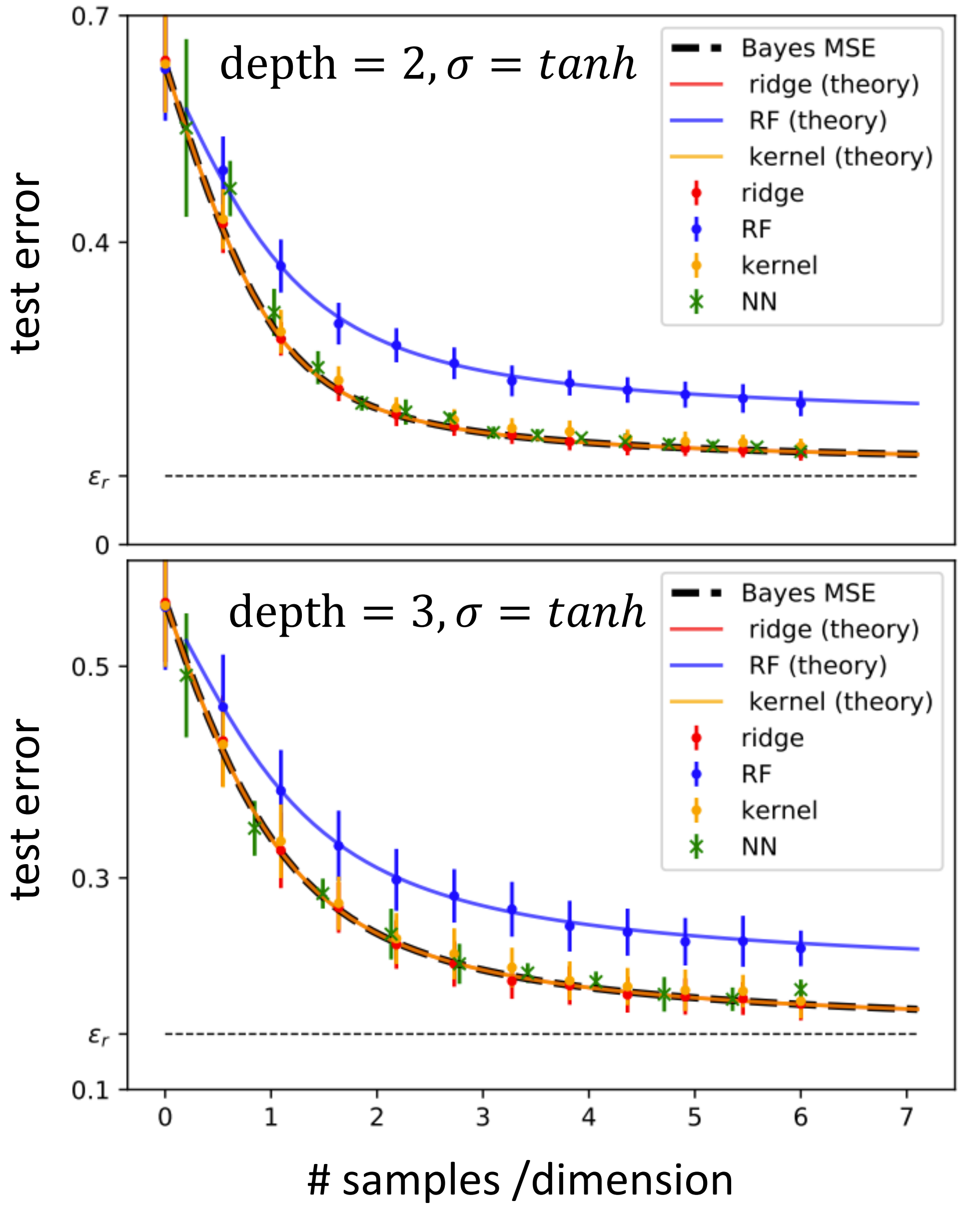
Summary:

✓ **Q1** We have sharp asymptotics for the Bayes optimal error of a deep, random network
= *lowest information theoretically achievable error*

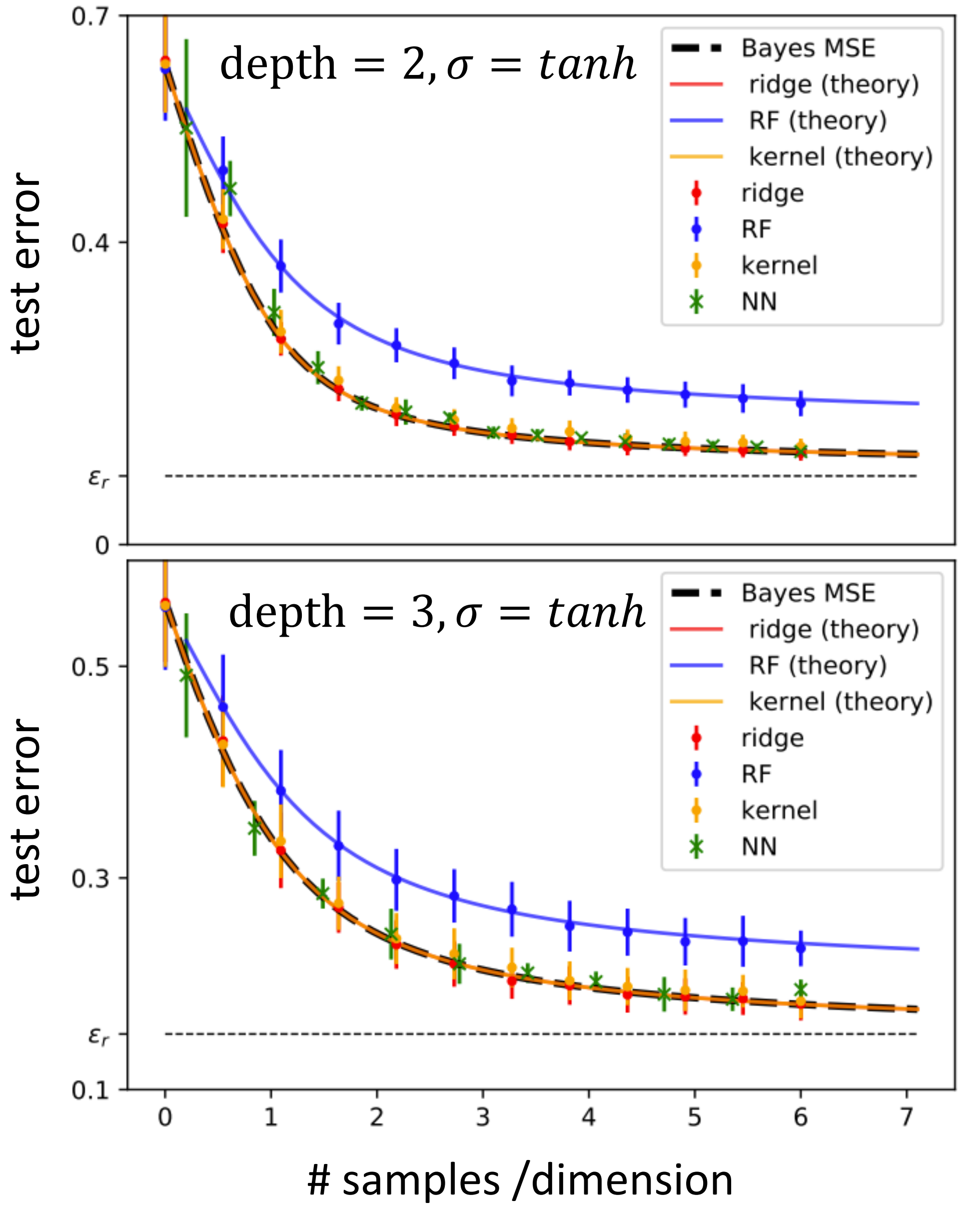
✓ **Q2a** We have sharp asymptotics for test error of a large class of ERM algorithms on the same target.

Q2b How do they compare?

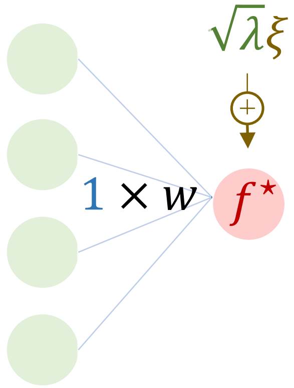
Regression

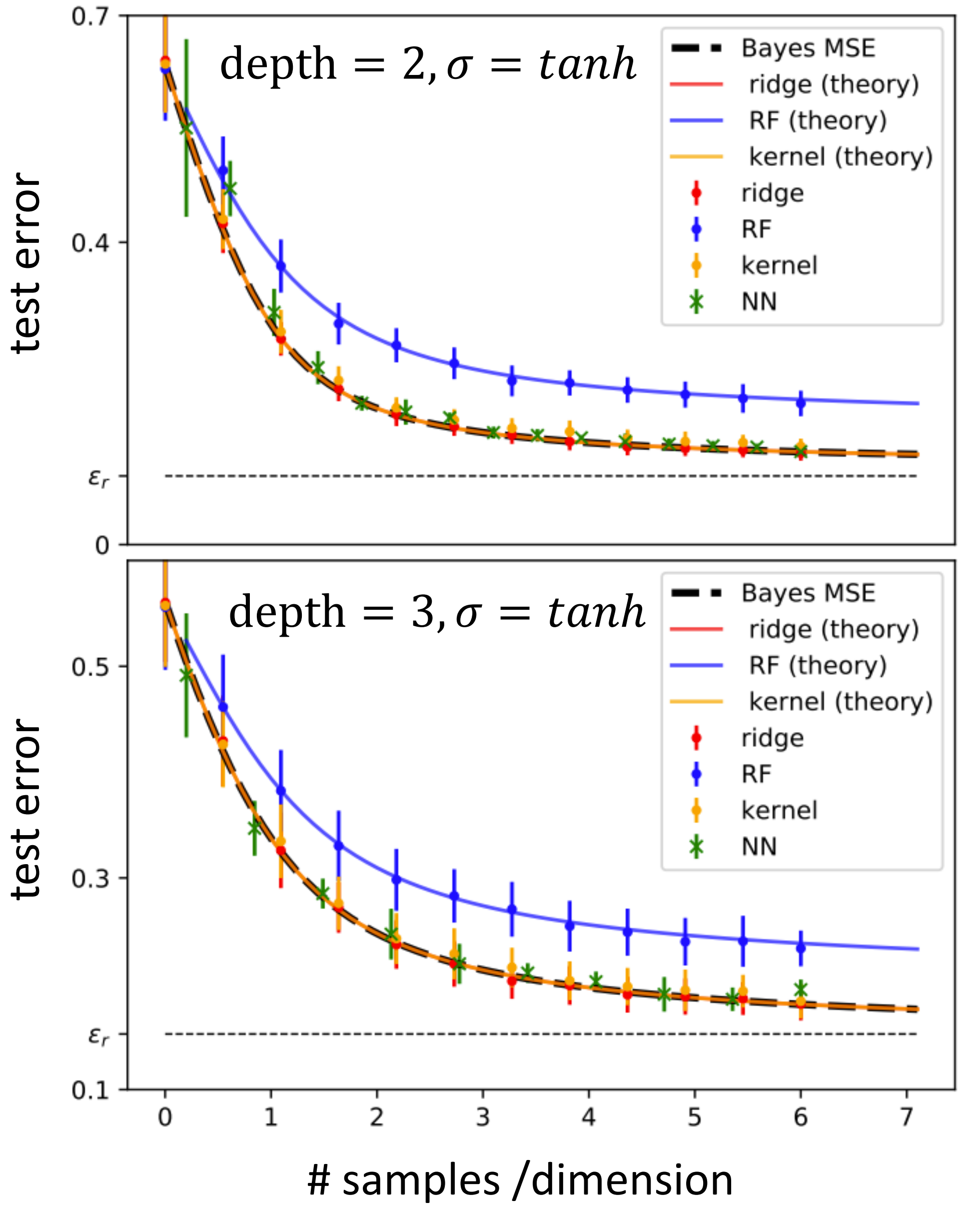


Optimally regularized ridge regression and kernel regression ***are Bayes optimal***.

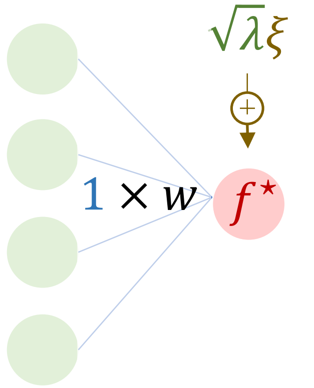


Ridge regression with ℓ_2 regularization λ is equivalent to Bayesian inference assuming the dataset comes from

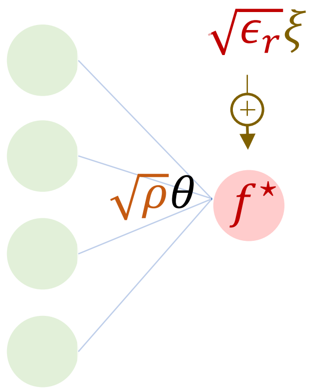




Ridge regression with ℓ_2 regularization λ is equivalent to Bayesian inference assuming the dataset comes from

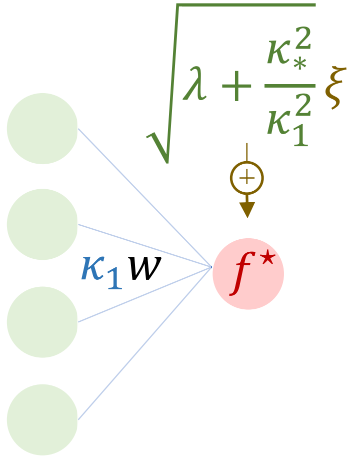


Since the target is equivalent to



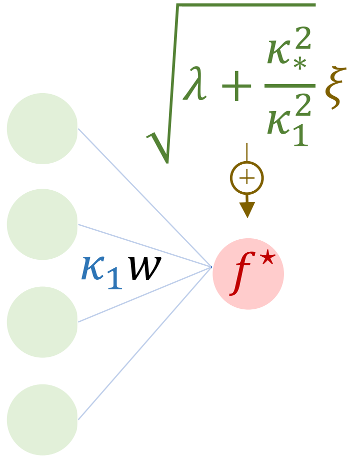
Bayes optimality is reached when taking $\frac{\lambda_{opt}}{1} = \frac{\epsilon_r}{\rho}$

Kernel regression with ℓ_2 regularization λ is equivalent to Bayesian inference assuming the dataset comes from

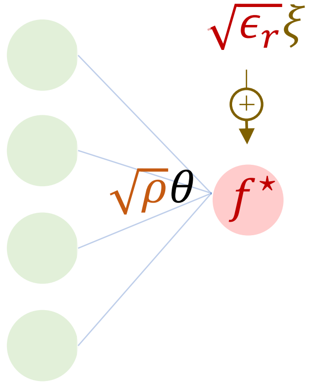


Mei, Misiakiewicz and Montanari, *Generalization error of random feature and kernel methods: hypercontractivity and kernel matrix concentration.* Applied and Computational Harmonic Analysis, 2021.

Kernel regression with ℓ_2 regularization λ is equivalent to Bayesian inference assuming the dataset comes from



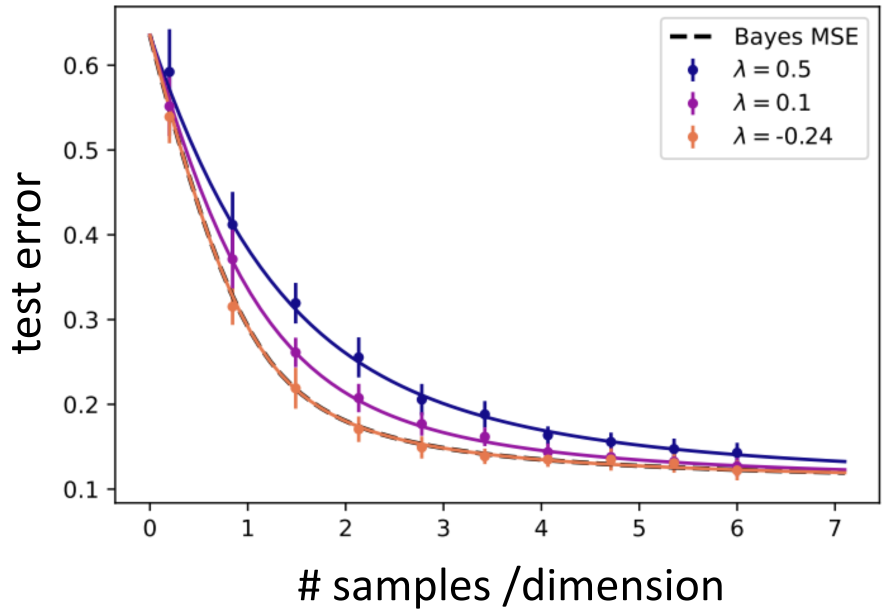
Since the target is equivalent to



Bayes optimality is reached when taking $\lambda_{opt} = \kappa_1^2\left(\frac{\epsilon_r}{\rho} - \frac{\kappa_*^2}{\kappa_1^2}\right)$

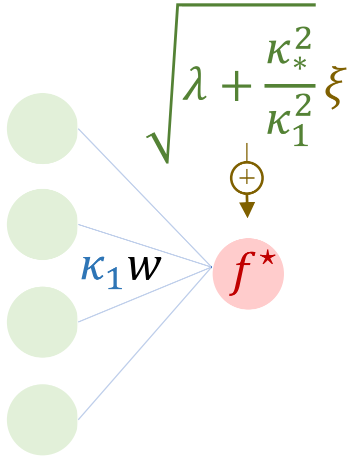
Mei, Misiakiewicz and Montanari, *Generalization error of random feature and kernel methods: hypercontractivity and kernel matrix concentration.* Applied and Computational Harmonic Analysis, 2021.

$\text{depth} = 2, \sigma = tanh$

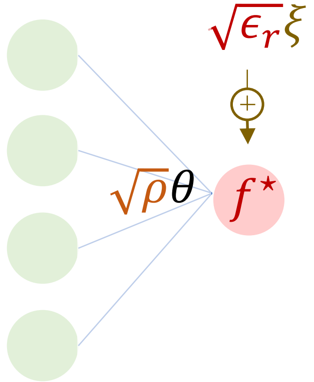


Mei, Misiakiewicz and Montanari, *Generalization error of random feature and kernel methods: hypercontractivity and kernel matrix concentration.* Applied and Computational Harmonic Analysis, 2021.

Kernel regression with ℓ_2 regularization λ is equivalent to Bayesian inference assuming the dataset comes from



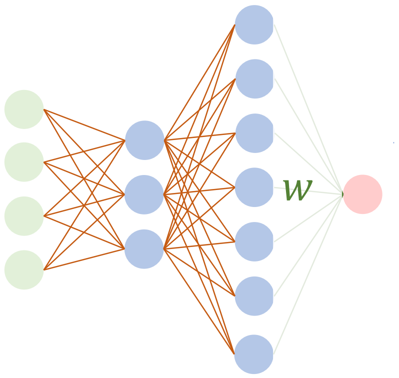
Since the target is equivalent to



Bayes optimality is reached when taking $$\lambda_{opt} = \kappa_1^2\left(\frac{\epsilon_r}{\rho} - \frac{\kappa_*^2}{\kappa_1^2}\right)$$

Remark : The optimal regularization can be negative

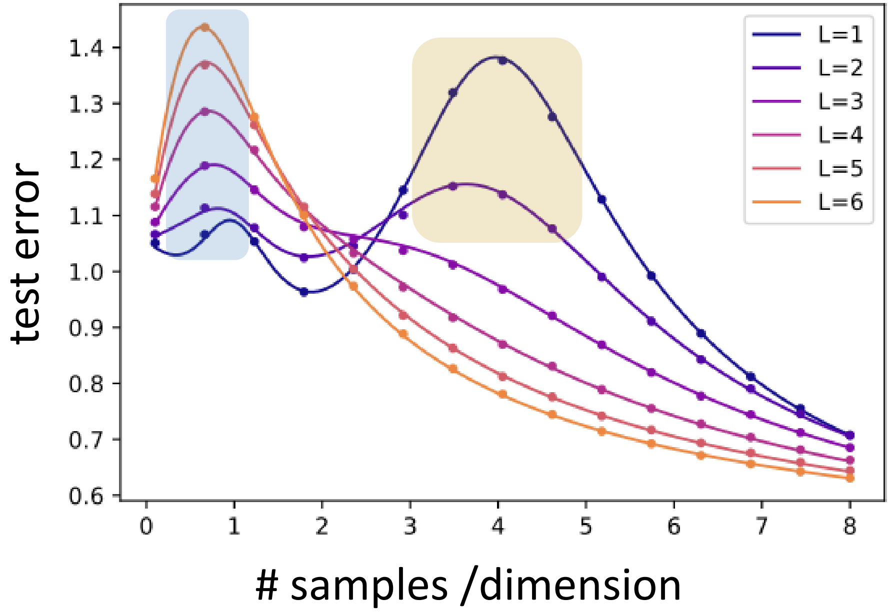
(learner)

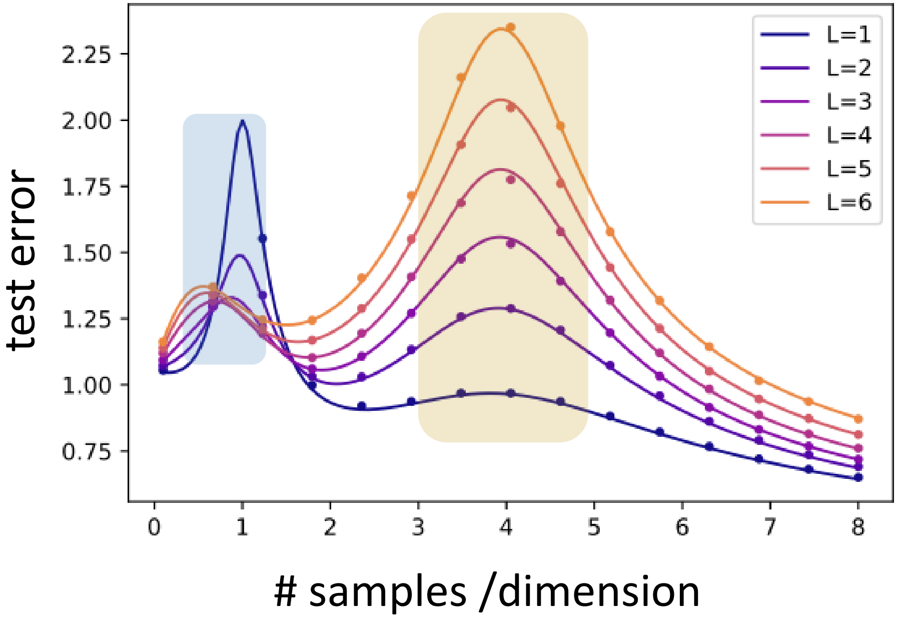


(GEP) $\approx$ $w^{\mathsf{T}} \underbrace{Ax}_{\text{signal}} + \underbrace{\xi}_{\text{noise}}$

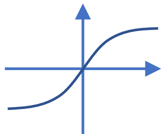
When the signal is used to interpolate, the noise behaves as an depth-induced ***implicit regularization***.

A second peak appears when the noise is used to interpolate the train set.

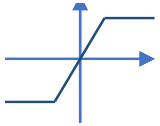




$\sigma =$



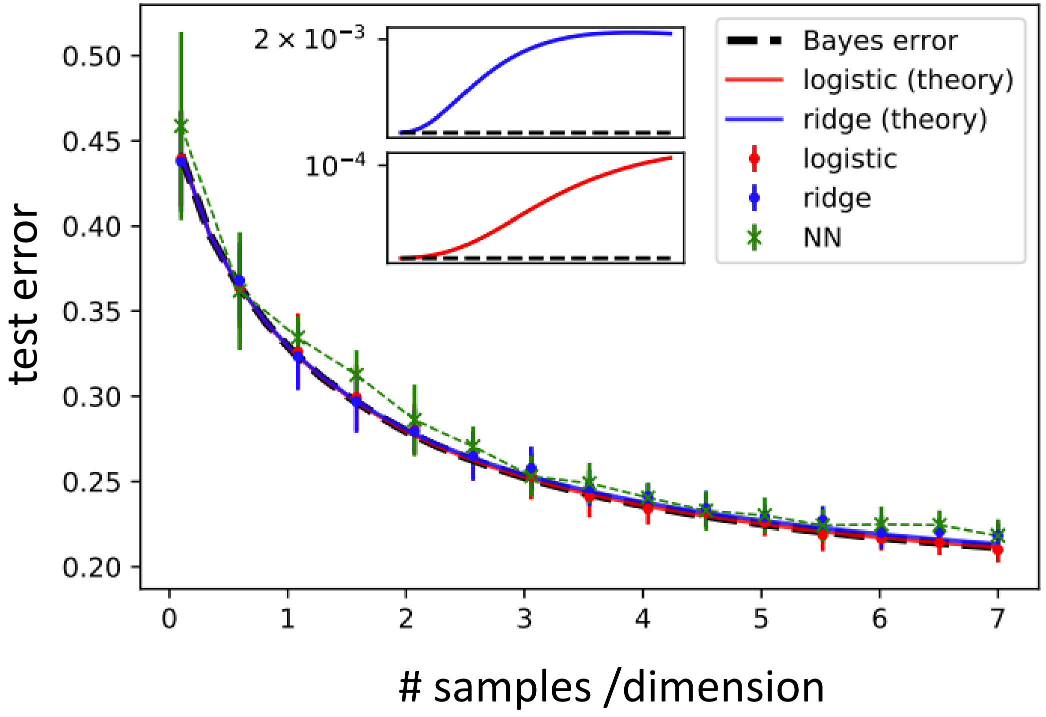
$\sigma =$



D'Ascoli, Sagun and Biroli. *Triple descent and the two kinds of overfitting* J. Stat. Mech. 2021

Classification

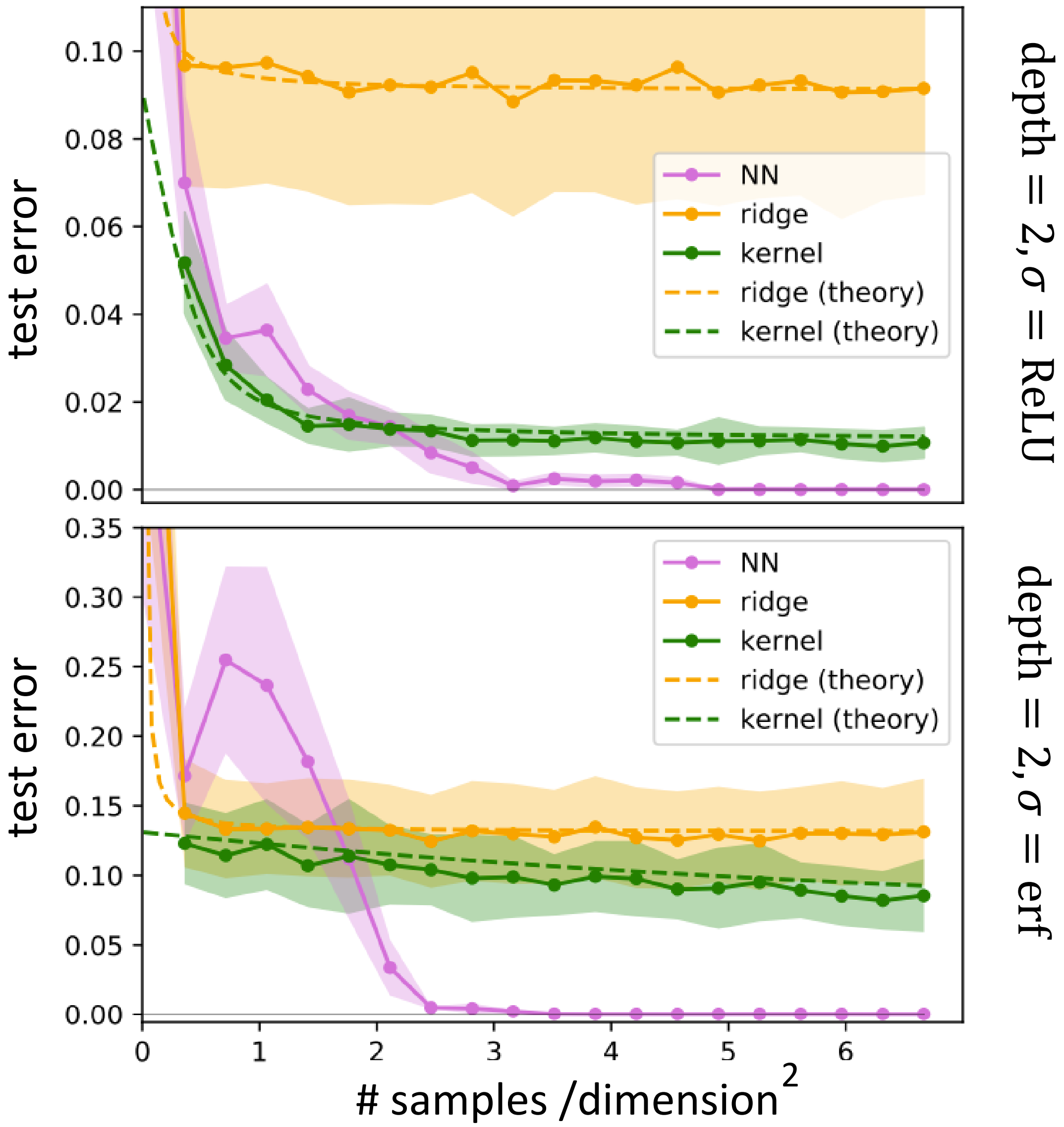
depth $= 3, \sigma = tanh$



Optimally regularized logistic and ridge classification ***are close to Bayes optimal***.

Q2 Can ERM methods achieve the Bayes error?

A2 Yes, because in the $n \sim d$ regime ***only second-order statistics*** seem to be learnt, and in terms of those the target is equivalent to a single-layer network.



When $n \sim d^2$, ***higher-order statistics are learnt***, the Gaussian equivalences break down.

Hong Hu and Yue M. Lu. *Sharp asymptotics of kernel ridge regression beyond the linear regime*. arXiv:2205.06798, 2022

Bordelon, Canatar, Pehlevan. *Spectrum dependent learning curves in kernel regression and wide neural networks* ICML 2020

Takeaways:

- In terms of ***second order statistics*** wrt a Gaussian input, a deep non-linear network is equivalent to a noisy linear network.
- Hence, In the $n \sim d$ regime, they are ***characterized by the same Bayes / ERM errors.***
- Thus, single-layer ERM learners are Bayes optimal.

Challenge /Future work:

There is a need for a theory of finite-width architectures in ***super linear regimes.***

Takeaways:

- In terms of ***second order statistics*** wrt a Gaussian input, a deep non-linear network is equivalent to a noisy linear network.
- Hence, In the $n \sim d$ regime, they are ***characterized by the same Bayes / ERM errors.***
- Thus, single-layer ERM learners are Bayes optimal.

Challenge /Future work:

There is a need for a theory of finite-width architectures in ***super linear regimes.***

Thank you for your attention !