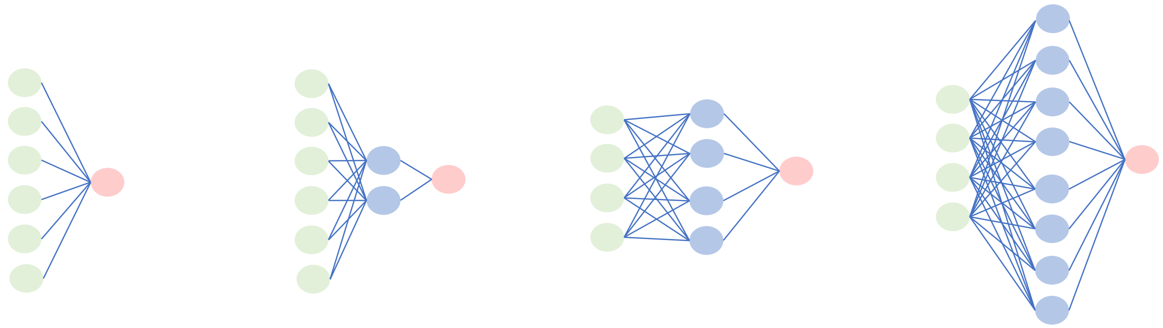
# Bayes-optimal learning of Deep Random Networks of Extensive-width

Hugo Cui, Florent Krzakala & Lenka Zdeborová

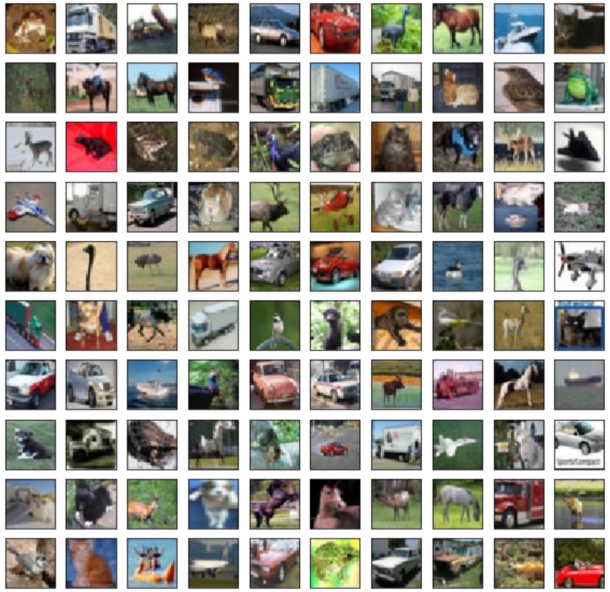
*EPFL, Switzerland*



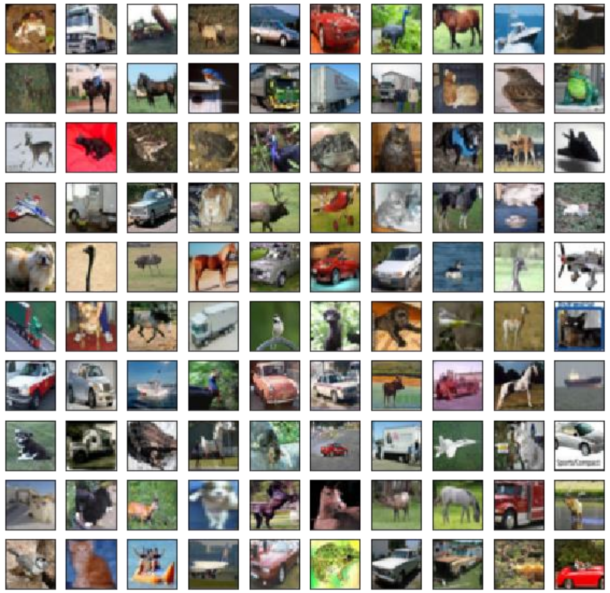
EPFL

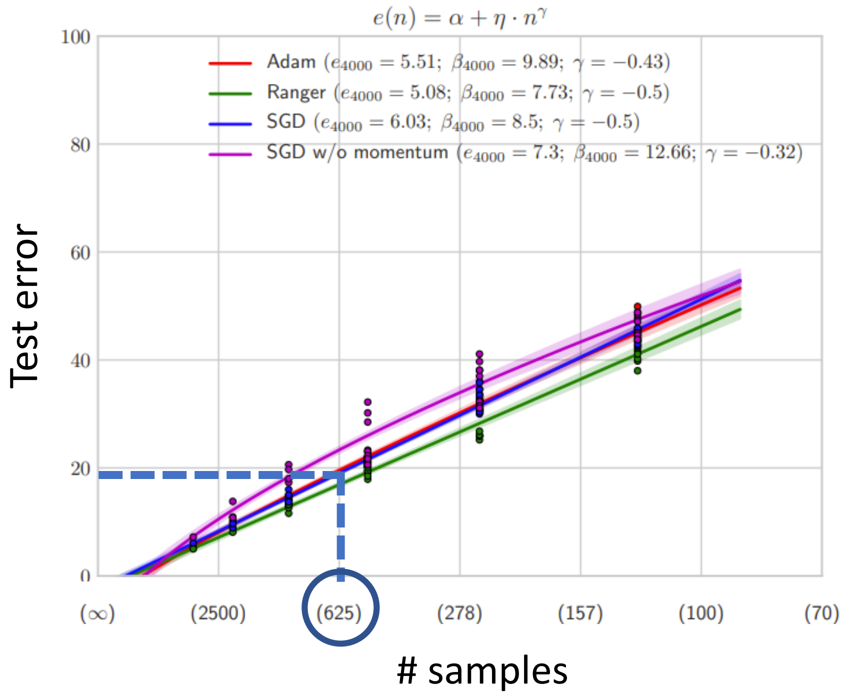
SPOC
Statistical Physics Of Computation

ICML
International Conference
On Machine Learning



***Question*: *What is the best accuracy*** one can achieve from 600 training samples?



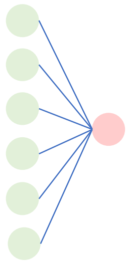



***Question*: *What is the best accuracy*** one can achieve from 600 training samples?

***(Empirical) Answer*:** Probably ≈ 82%, using good networks.

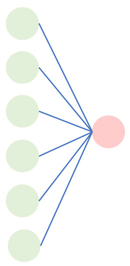
Hoeim et al., *Learning curves for Analysis of Deep Networks*, ICML 2020

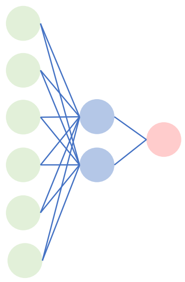
## Theoretical testbeds: ***random neural networks***



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

Theoretical testbeds: ***random neural networks***



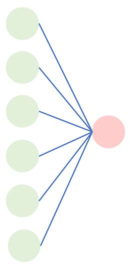


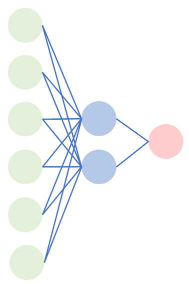
$width \ll dimension$

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

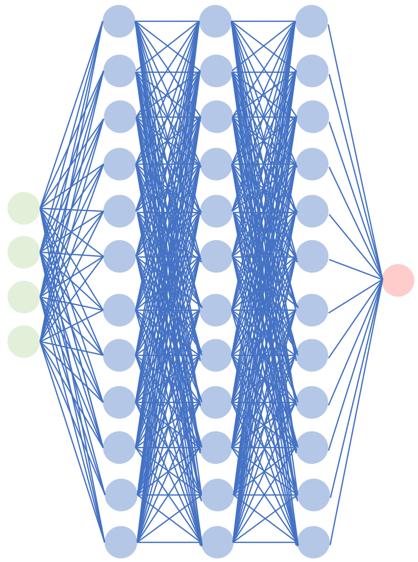
Aubin et al, *The committee machine: Computational to statistical gaps,* NeurIPS 2019

## Theoretical testbeds: ***random neural networks***





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



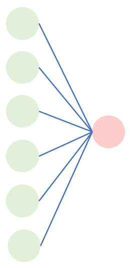
$\textit{width} \gg \textit{dimension}$

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

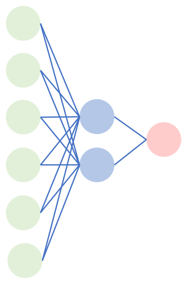
Aubin et al, *The committee machine: Computational to statistical gaps,* NeurIPS 2019

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

Theoretical testbeds: ***random neural networks***

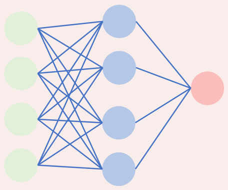


Barbier et al, *Optimal errors and phase transitions in high-dimensional generalized linear models,* PNAS 2017



*width* $\ll$ *dimension*

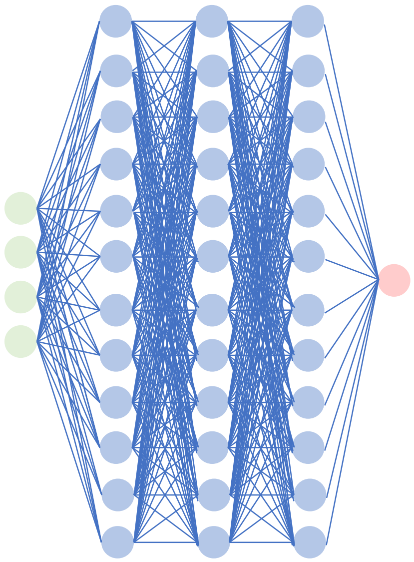
Aubin et al, *The committee machine: Computational to statistical gaps,* NeurIPS 2019



*width* $\sim$ *dimension*

Li and Sompolinsky, *Statistical Mechanics of Deep Linear Networks,* PRX 2020

Ariosto et al., *Statistical Mechanics of Deep Learning Beyond the Infinite Width limit,* 2023



*width* $\gg$ *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

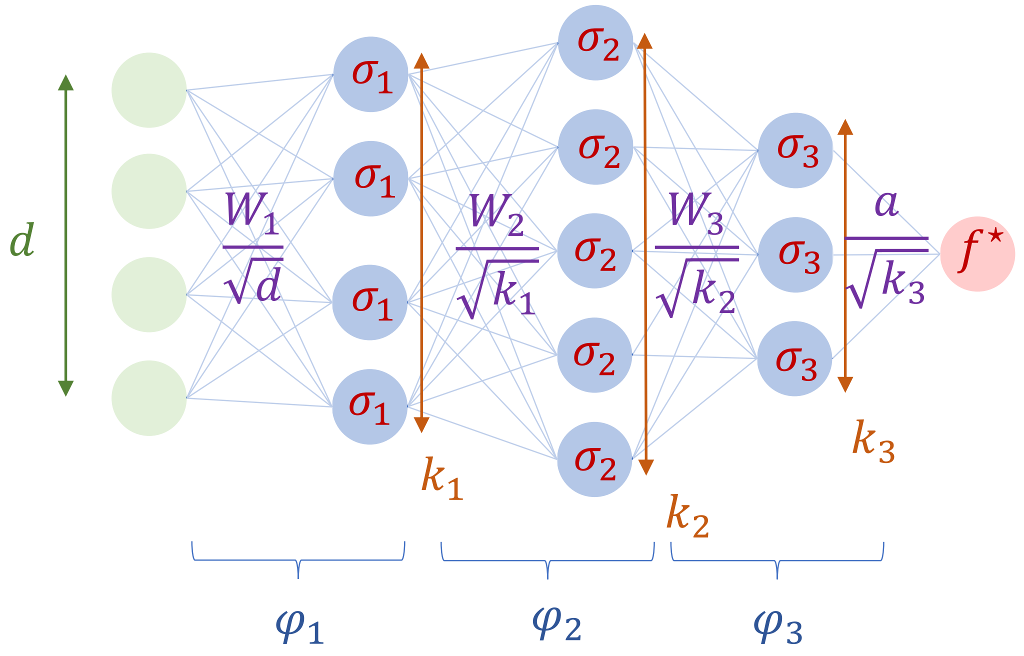
*(Data)* Gaussian data: $x \sim \mathcal{N}(0, \Sigma)$

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

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

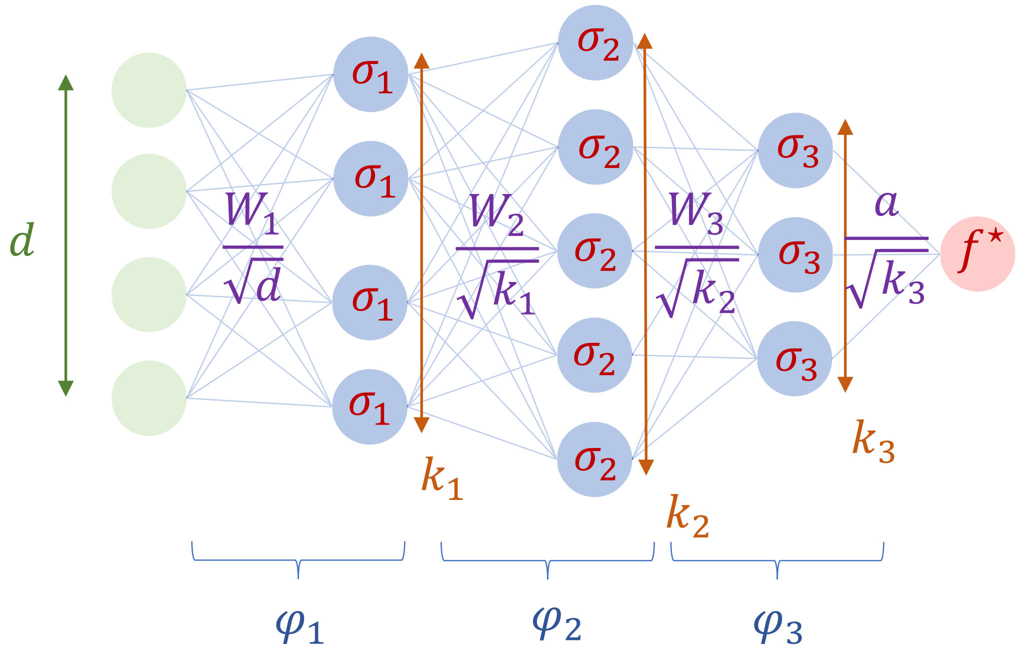


*(Data)* Gaussian data: $x \sim \mathcal{N}(0, \Sigma)$

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

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



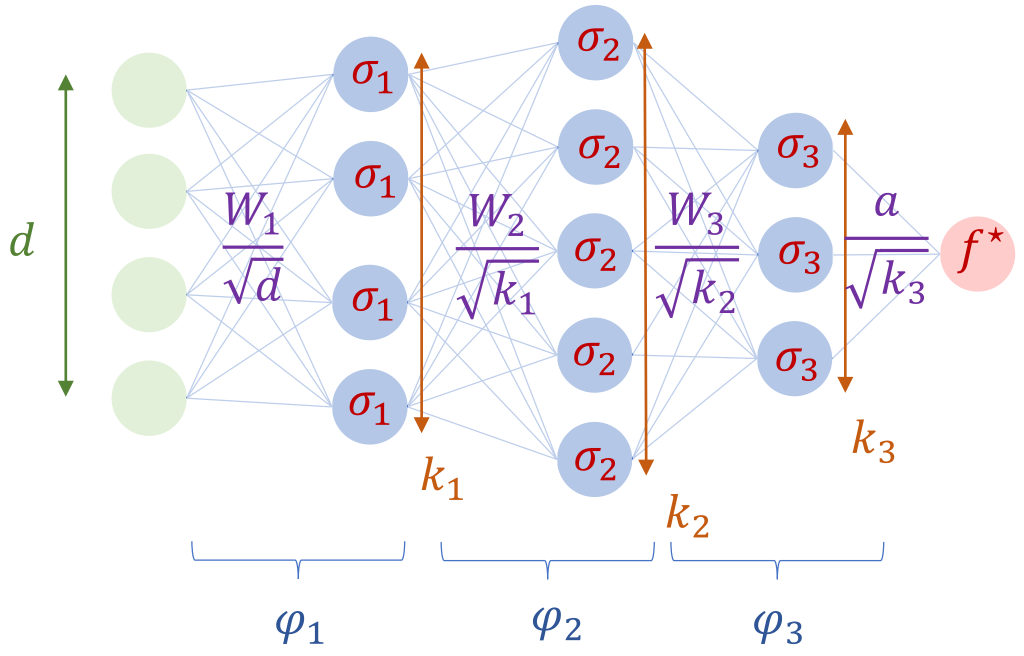

*(Train set)* Supervised learning with $n$ i.i.d samples $\mathcal{D} = \{x^\mu, y^\star(x^\mu)\}_{\mu=1}^n$

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

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

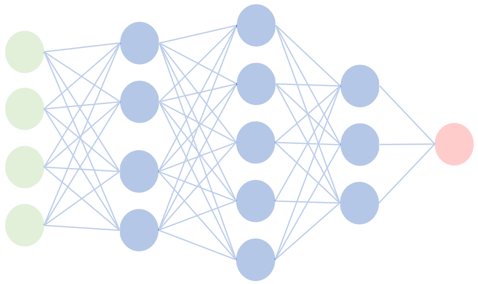



*(Train set)* Supervised learning with $n$ i.i.d samples $\mathcal{D} = \{x^\mu, y^\star(x^\mu)\}_{\mu=1}^n$

***Proportional extensive-width limit***

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

*same Bayes optimal errors*



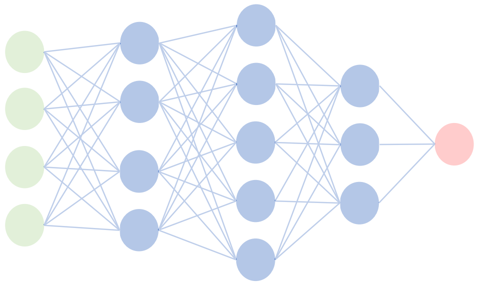
$\approx$



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

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

*same Bayes optimal errors*



$$\approx$$



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

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

$\boldsymbol{\rho}$, $\boldsymbol{\epsilon_r}$ depend on the architecture and activations of the original network.

*Regression*

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

*Classification*

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

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

*Regression*

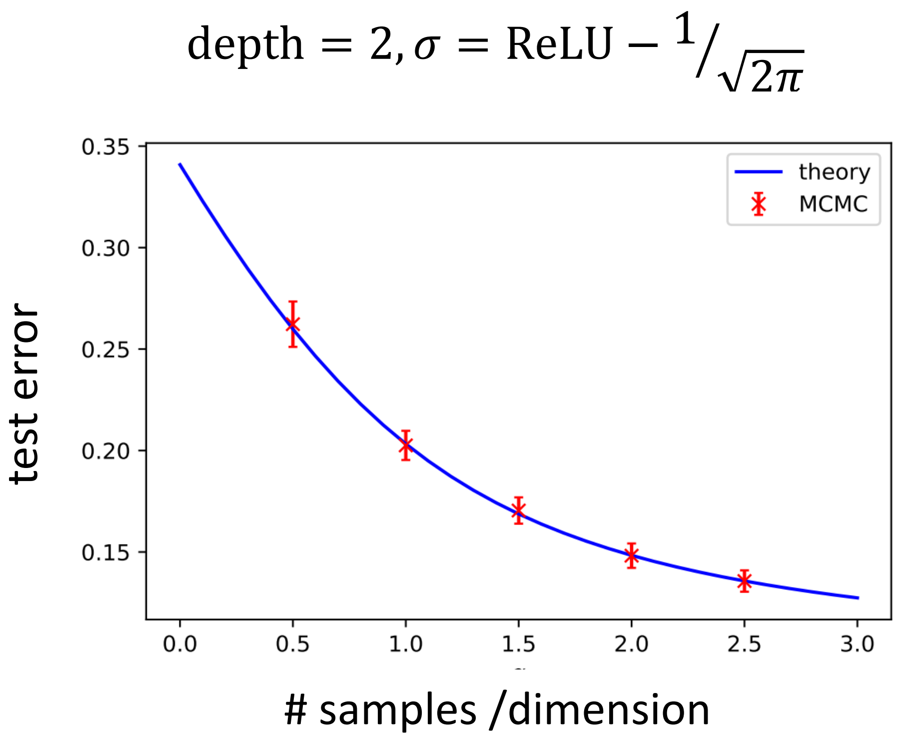
$$\epsilon_{g,\mathrm{reg}}^{\mathrm{BO}}=\prod_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2\Big(\Delta_a\big(\int z\,\mathrm{d}\mu(z)\big)\prod_{\ell=1}^{L}\Delta_\ell-q\Big)+\epsilon_r$$

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

*Classification*

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

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

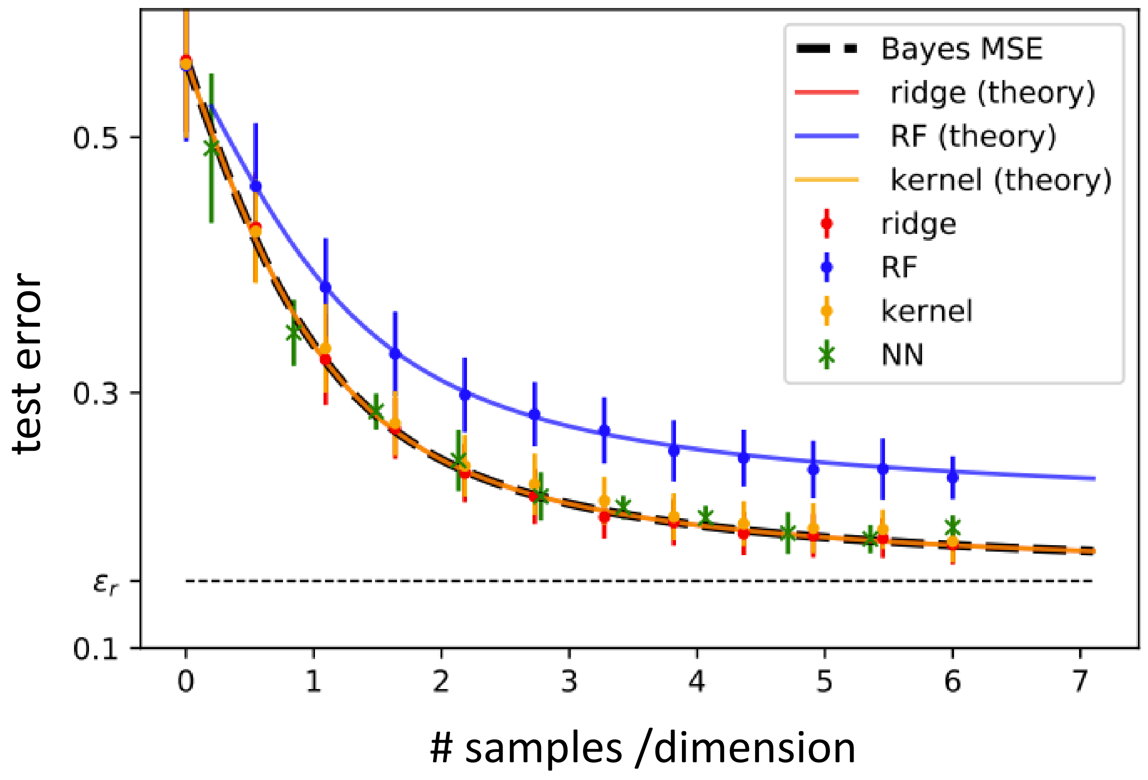


✓ **Q1**. Can one provide a sharp asymptotic characterization of the Bayes-optimal error?

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

## *Regression*

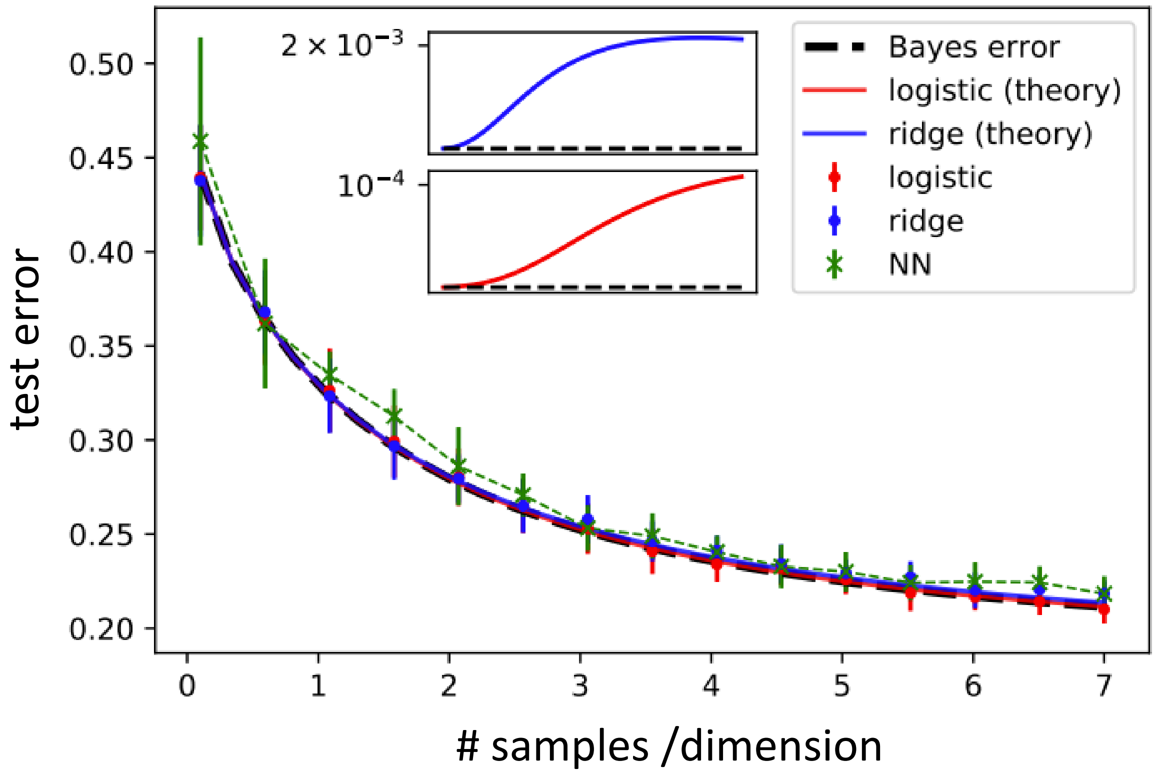
$$\mathrm{depth} = 3, \sigma = tanh$$



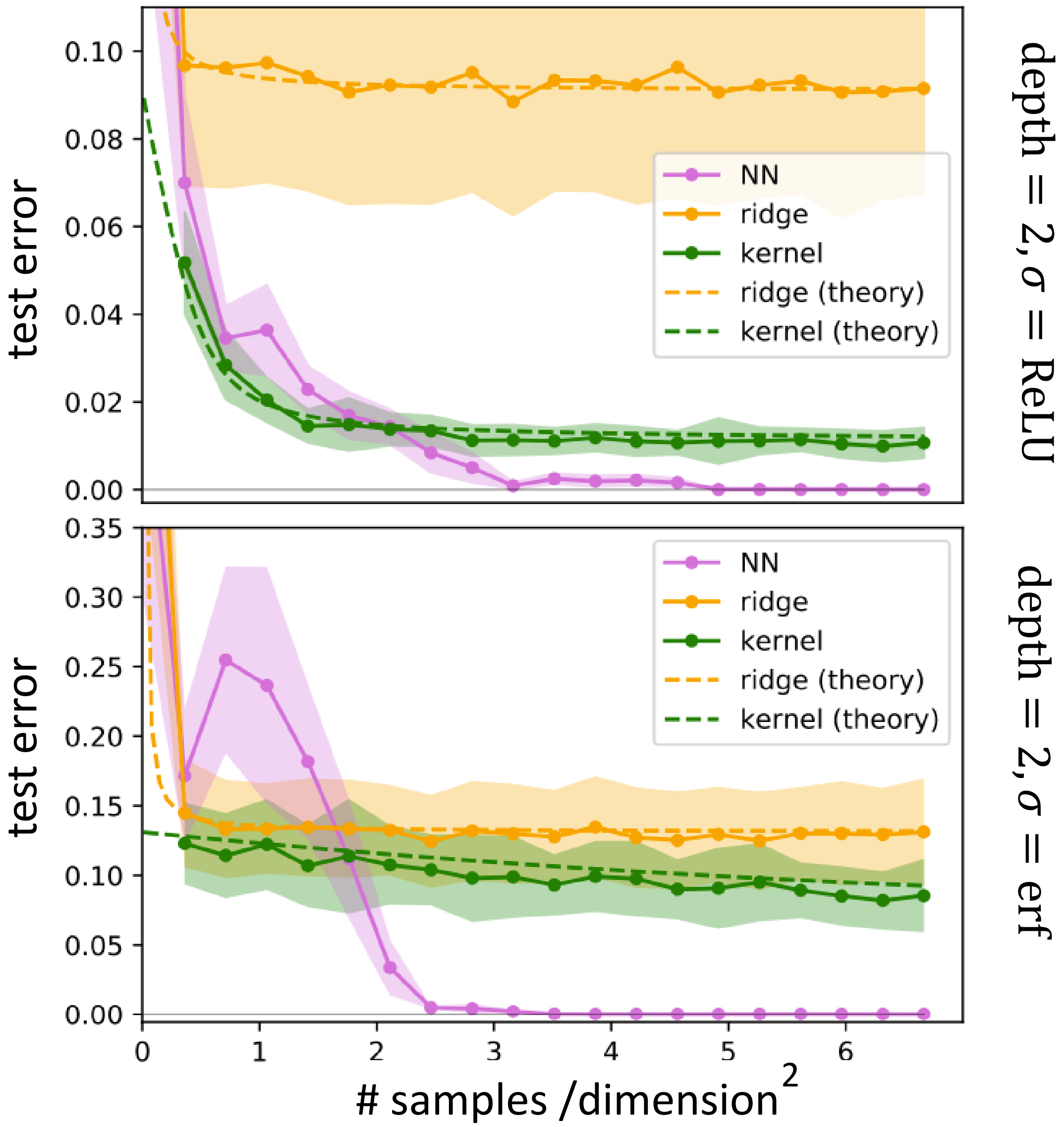

Optimally regularized ridge regression and kernel regression ***are Bayes optimal***.

## *Classification*

$\text{depth} = 3, \sigma = tanh$




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




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*:

- We conjecture closed-form formulas for the Bayes-optimal test errors when learning data generated by a deep non-linear random network.
- This optimal error is achieved by very simple ERM methods.

*Challenge /Future work:*

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

## *Thank you for your attention !*

See you at posters:

# 221 on Thu. 10.30 (*this work*)

# 814 on Wed 14.00 (*learning with deep random nets*)