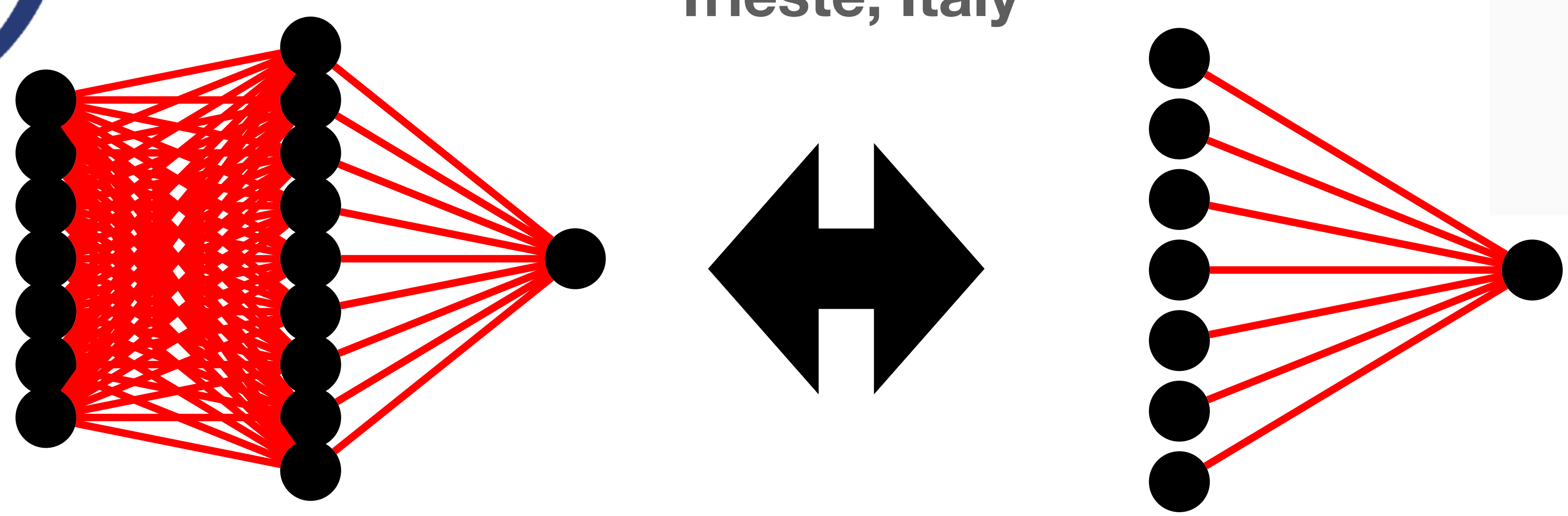
Fundamental limits of shallow neural networks in supervised learning

Jean Barbier

International Center for Theoretical Physics
Trieste, Italy

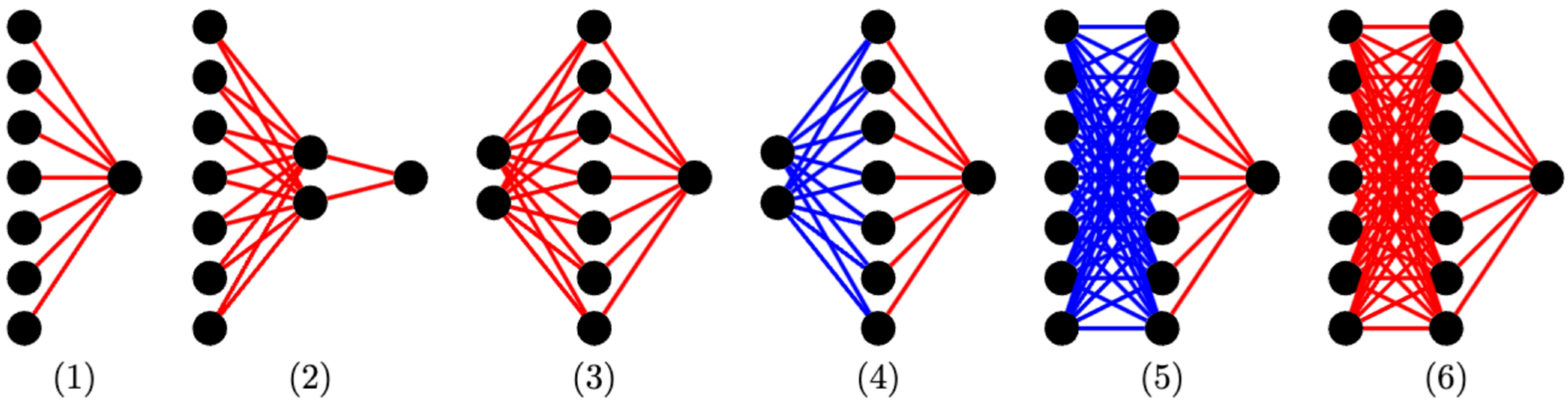


Joint with
Francesco Camilli
and Daria Tieplova
(ICTP)

Neural Networks Zoology

Committee machine

Random features models



GLM/perceptron

Mean field regime

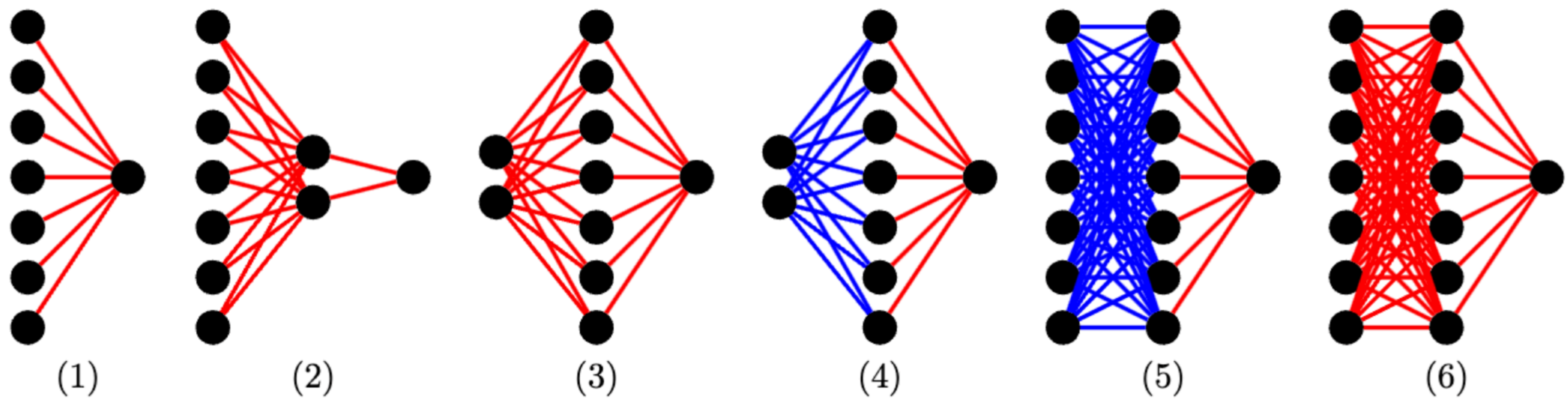
Fully learnable 2LNN
Proportional regime

Neural Networks Zoology

But not « too much data»

Committee machine

Random features models



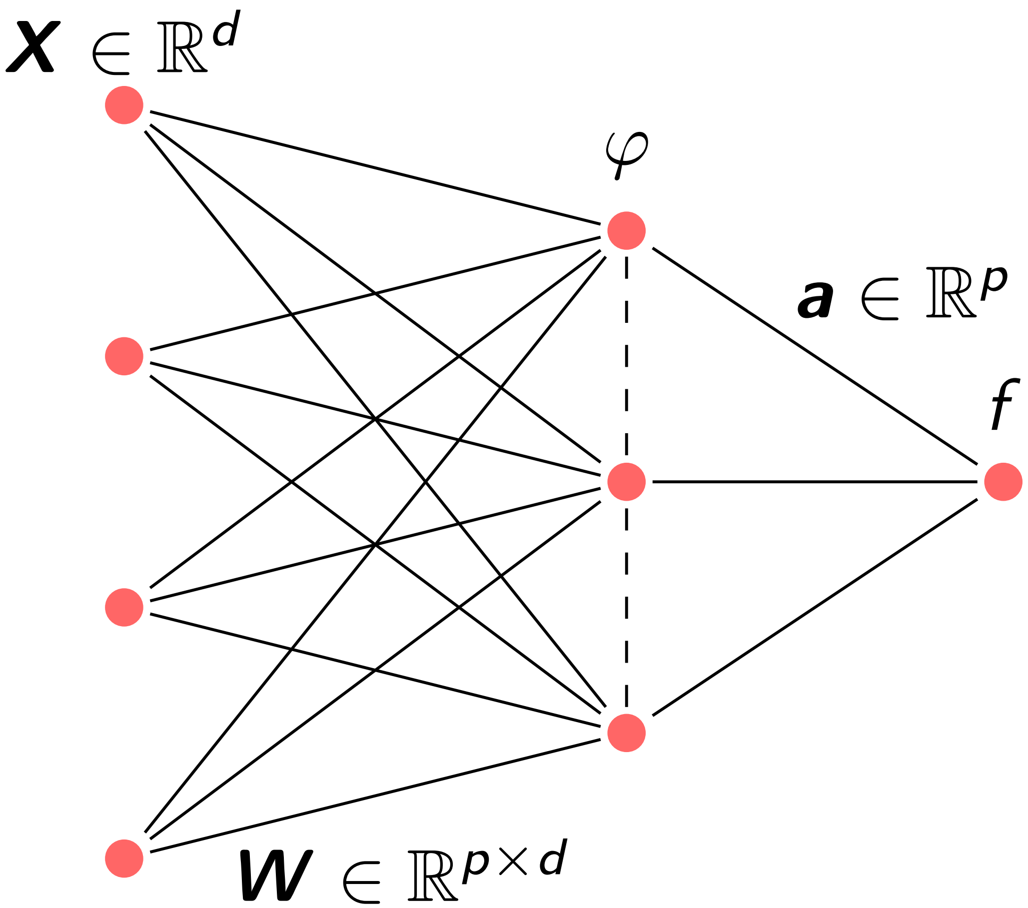
GLM/perceptron

Mean field regime

Fully learnable 2LNN
Proportional regime

Set-up

A two-layer NN



$$Y = f_{\boldsymbol{A}}\Big(\frac{\boldsymbol{a}^\intercal}{\sqrt{p}}\varphi\Big(\frac{\boldsymbol{W}\boldsymbol{X}}{\sqrt{d}}\Big)\Big)\,.$$

Supervised learning: Starting from a *training set* $\mathcal{D}_n = \{(\boldsymbol{X}_\mu, Y_\mu)_{\mu=1}^n\}$, we adjust the weights $\boldsymbol{a}$, $\boldsymbol{W}$ s.t.

$$Y_\mu \approx f_{\boldsymbol{A}_\mu}\Big(\frac{\boldsymbol{a}^\intercal}{\sqrt{p}}\varphi\Big(\frac{\boldsymbol{W}\boldsymbol{X}_\mu}{\sqrt{d}}\Big)\Big)\,, \quad \forall \mu\,.$$

Main Goal

Produce the smallest possible generalization error:

$$\mathcal{E} = \Big(Y_{\rm new} - f_{\boldsymbol{A}_{\rm new}}\Big(\frac{\boldsymbol{a}^\intercal}{\sqrt{p}}\varphi\Big(\frac{\boldsymbol{W}\boldsymbol{X}_{\rm new}}{\sqrt{d}}\Big)\Big)\Big)^2$$

for a new couple $(\boldsymbol{X}_{\rm new}, Y_{\rm new})$.

What affects $\mathcal{E}$?

Among the many factors that can affect it, the most relevant ones for us are:

- The size of the training set n: the more data points the more accurate we expect to be;
- The size of the network itself, parameterized by p;
- The dimensionality of the input d;
- The method used to train the network (e.g. ERM, SGD, Bayes)
- **The nature of the dataset**, i.e. what is the true underlying function $Y = f(\boldsymbol{X})$ the NN aims at approximating
- many more...

Central questions

What is the least possible $\mathcal{E}$? When is it achieved?

Teacher-student setup

We assume that the training set is generated itself by a 2-layer **teacher network** with matching architecture:

$$Y_\mu = f_{\boldsymbol{A}_\mu}\Big(\frac{\boldsymbol{a}^{*\intercal}}{\sqrt{p}}\varphi\Big(\frac{\boldsymbol{W}^*\boldsymbol{X}_\mu}{\sqrt{d}}\Big)\Big) + \underbrace{\sqrt{\Delta}Z_\mu}_{\text{label noise}}, \quad \forall \mu \leq n,$$

for some $\Delta > 0$, $Z_\mu \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$. **Prior on the weights**: $a_i^*, W_{ij}^* \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$. Same as

$$Y_\mu \sim P_{\text{out}}\Big(\cdot \mid \frac{\boldsymbol{a}^{*\intercal}}{\sqrt{p}}\varphi\Big(\frac{\boldsymbol{W}^*\boldsymbol{X}_\mu}{\sqrt{d}}\Big)\Big)$$

with

$$P_{\text{out}}(y \mid x) = \int \frac{dP_A(\boldsymbol{A})}{\sqrt{2\pi\Delta}} \exp\Big(-\frac{1}{2\Delta}\big(f_{\boldsymbol{A}}(x) - y\big)^2\Big)$$

Main theoretical restriction

We consider $\boldsymbol{X}_\mu \overset{\text{iid}}{\sim} \mathcal{N}(0, I_d)$, *structureless* input data.

Bayes-optimal student

Definition (informal)

A student network is Bayes-optimal if it is completely aware of the generative model

$$Y_\mu = f_{\boldsymbol{A}_\mu}\Big(\frac{\boldsymbol{a}^{*\mathsf{T}}}{\sqrt{p}}\varphi\Big(\frac{\boldsymbol{W}^*\boldsymbol{X}_\mu}{\sqrt{d}}\Big)\Big) + \sqrt{\Delta}Z_\mu\,, \quad \forall\mu \leq n\,,$$

and it matches the teacher's architecture. In other words: **a part from the true weights, it knows everything there is know**.

A Bayes-optimal student has access to the **Bayes-posterior**:

$$dP(\boldsymbol{\theta} \mid \mathcal{D}_n) = \frac{1}{\mathcal{Z}(\mathcal{D}_n)}\prod_{\mu=1}^{n} P_{\text{out}}\Big(Y_\mu \mid \frac{\boldsymbol{a}^{\mathsf{T}}}{\sqrt{p}}\varphi\Big(\frac{\boldsymbol{W}\boldsymbol{X}_\mu}{\sqrt{d}}\Big)\Big)D\boldsymbol{\theta}$$

where $D\boldsymbol{\theta} = D\boldsymbol{a}D\boldsymbol{W}$ is the Gaussian prior on the weights.

Proposition (informal)

A Bayes-optimal student NN achieves the lowest expected generalization error

$$\mathbb{E}\mathcal{E} := \mathbb{E}\big(Y_{\text{new}} - \hat{Y}(\mathcal{D}_n, \boldsymbol{X}_{\text{new}})\big)^2$$

that is yielded by the BO predictor

$$\begin{aligned}\hat{Y}_{\text{Bayes}}(\mathcal{D}_n, \boldsymbol{X}_{\text{new}}) &= \mathbb{E}[Y_{\text{new}} \mid \mathcal{D}_n, \boldsymbol{X}_{\text{new}}] \\ &= \int dY\, Y\, P_{\text{out}}\Big(Y \mid \frac{\boldsymbol{a}^\intercal}{\sqrt{p}}\varphi\Big(\frac{\boldsymbol{W}\boldsymbol{X}_{\text{new}}}{\sqrt{d}}\Big)\Big) dP(\boldsymbol{\theta} \mid \mathcal{D}_n)\,.\end{aligned}$$

Main information theoretic quantities

Recall

$$dP(\boldsymbol{a}, \boldsymbol{W} \mid \mathcal{D}_n) = \frac{1}{\mathcal{Z}(\mathcal{D}_n)} \prod_{\mu=1}^{n} P_{\mathrm{out}}\Big(Y_\mu \mid \frac{\boldsymbol{a}^\intercal}{\sqrt{p}}\varphi\Big(\frac{\boldsymbol{W}\boldsymbol{X}_\mu}{\sqrt{d}}\Big)\Big) D\boldsymbol{a} D\boldsymbol{W} \to \langle \cdot \rangle \,.$$

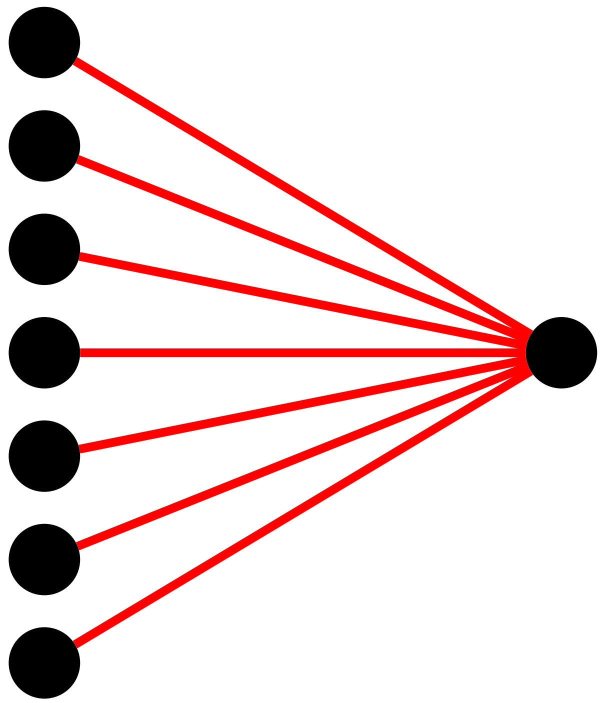
- **Partition function** or **evidence**:

$$\mathcal{Z}(\mathcal{D}_n) = \int \prod_{\mu=1}^{n} P_{\mathrm{out}}\Big(Y_\mu \mid \frac{\boldsymbol{a}^\intercal}{\sqrt{p}}\varphi\Big(\frac{\boldsymbol{W}\boldsymbol{X}_\mu}{\sqrt{d}}\Big)\Big) D\boldsymbol{a} D\boldsymbol{W}$$

- **free entropy**: $\bar{f}_n = \frac{1}{n}\mathbb{E}\log \mathcal{Z}(\mathcal{D}_n)$
- **Mutual Information** per data point:

$$\begin{aligned}\frac{I_n(\boldsymbol{a}^*, \boldsymbol{W}^*; \mathcal{D}_n)}{n} &= \frac{H(\mathcal{D}_n)}{n} - \frac{H(\mathcal{D}_n \mid \boldsymbol{a}^*, \boldsymbol{W}^*)}{n} \\ &= -\bar{f}_n + \mathbb{E}\log P_{\mathrm{out}}\Big(Y_1 \mid \frac{\boldsymbol{a}^{*\intercal}}{\sqrt{p}}\varphi\Big(\frac{\boldsymbol{W}^*\boldsymbol{X}_1}{\sqrt{d}}\Big)\Big)\end{aligned}$$

A simpler ancestor: the GLM



The teacher Generalized Linear Model is given by:

$$Y_\mu^\circ = f_{\boldsymbol{A}_\mu}\Big(\rho \frac{\boldsymbol{v}^{*\intercal}\boldsymbol{X}_\mu}{\sqrt{d}} + \sqrt{\epsilon}\xi_\mu^*\Big) + \sqrt{\Delta} Z_\mu\,,$$

$$\text{or} \quad Y_\mu^\circ \sim P_{\rm out}\Big(\cdot \mid \rho \frac{\boldsymbol{v}^{*\intercal}\boldsymbol{X}_\mu}{\sqrt{d}} + \sqrt{\epsilon}\xi_\mu^*\Big)$$

with $v_i^*, \xi_\mu^* \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, $\rho, \epsilon \geq 0$.

Free entropy:

$$\bar{f}_n^\circ = \frac{1}{n}\mathbb{E}\log \int \prod_{\mu=1}^{n} P_{\rm out}\Big(Y_\mu \mid \rho \frac{\boldsymbol{v}^\intercal \boldsymbol{X}_\mu}{\sqrt{d}} + \sqrt{\epsilon}\xi_\mu\Big) D\boldsymbol{v} D\boldsymbol{\xi}$$

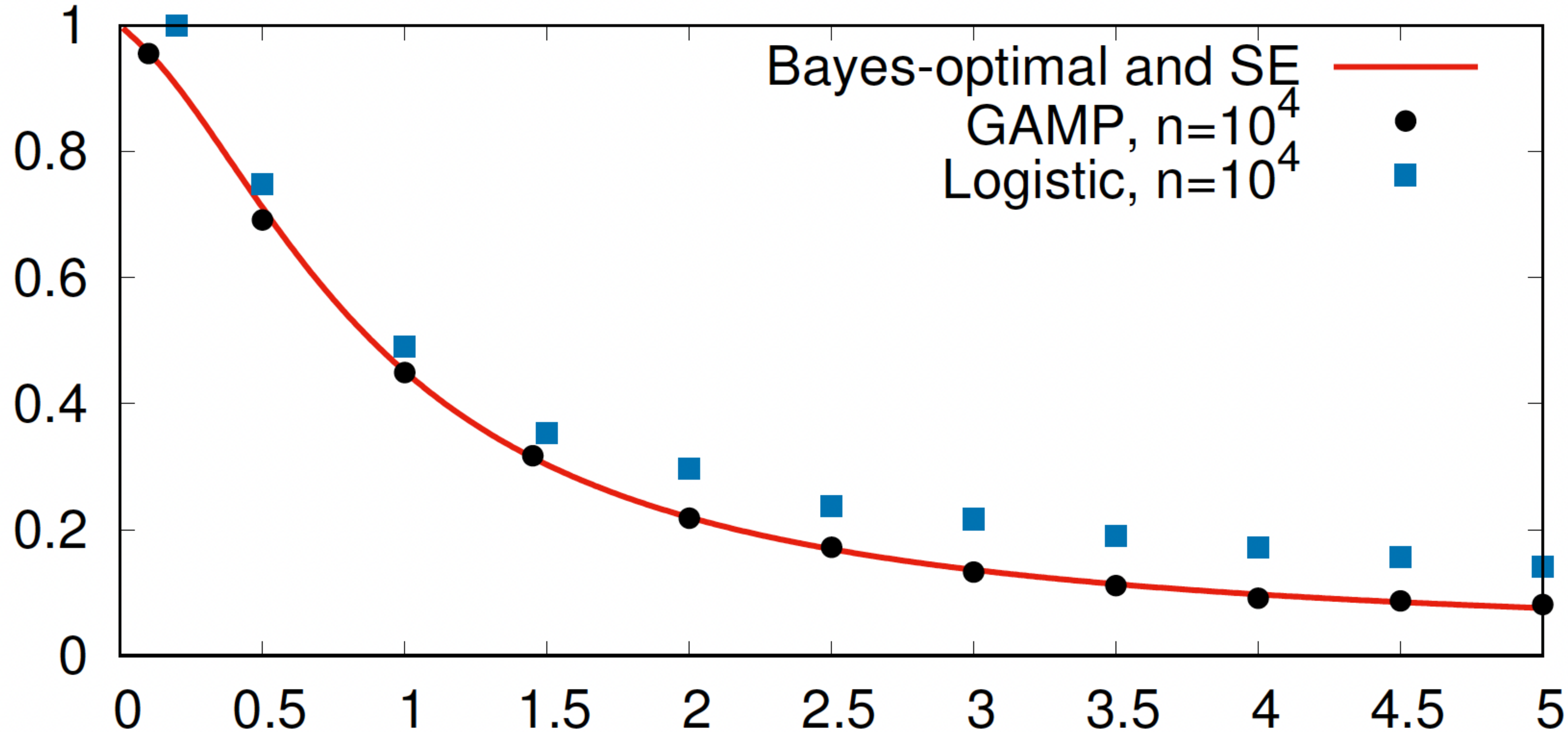
Mutual information:

$$\frac{1}{n} I_n^\circ(\boldsymbol{v}^*, \boldsymbol{\xi}^*; \mathcal{D}_n^\circ) = -\bar{f}_n^\circ + \mathbb{E}\log P_{\rm out}\Big(Y_1 \mid \rho \frac{\boldsymbol{v}^{*\intercal}\boldsymbol{X}_1}{\sqrt{d}} + \sqrt{\epsilon}\xi_1^*\Big)\,.$$

Barbier, Miolane, Macris, Krzakala, Zdeborova PNAS 19'

Relevant scalings in the GLM

In the GLM the relevant scaling is $\alpha = \frac{n}{d} = O(1)$, and the free entropy (and MI) are given by an RS formula.
The $\mathbb{E}\mathcal{E}$ is related through a derivative to the MI.



α

Barbier, Miolane, Macris, Krzakala, Zdeborova PNAS 19'

Results

Free entropy equivalence

Recall

$$Y_\mu \sim P_{\mathrm{out}}\Big(\cdot \mid \frac{\boldsymbol{a}^{*\intercal}}{\sqrt{p}}\varphi\Big(\frac{\boldsymbol{W}^{*}\boldsymbol{X}_\mu}{\sqrt{d}}\Big)\Big), \quad Y^\circ_\mu \sim P_{\mathrm{out}}\Big(\cdot \mid \rho\frac{\boldsymbol{v}^{*\intercal}\boldsymbol{X}_\mu}{\sqrt{d}} + \sqrt{\epsilon}\xi^*_\mu\Big)$$

Theorem (Barbier, Camilli, Tieplova 23')

Let $\rho := \mathbb{E}_{\mathcal{N}(0,1)}\varphi'$ and $\epsilon^2 := \mathbb{E}_{\mathcal{N}(0,1)}\varphi^2 - \rho^2$. Assume

A1) $\varphi \in C^3(\mathbb{R},\mathbb{R})$ is odd, $|\varphi'|, |\varphi''|, |\varphi'''| \leq \bar{K}$

A2) $|f|, |f'|, |f''| \leq \bar{K}$ P_A-almost surely.

Then:

$$|\bar{f}_n - \bar{f}^\circ_n| = O\Bigg(\sqrt{\Big(1+\frac{n}{d}\Big)\Big(\frac{n}{p}+\frac{n}{d^{3/2}}+\frac{1}{\sqrt{d}}\Big)}\Bigg).$$

Mutual information

Corollary

Under the same hypothesis, the following holds:

$$\left|\frac{1}{n} I_n(\boldsymbol{a}^*, \boldsymbol{W}^*; \mathcal{D}_n) - \frac{1}{n} I_n^\circ(\boldsymbol{v}^*, \boldsymbol{\xi}^*; \mathcal{D}_n^\circ)\right| = O\left(\sqrt{\left(1 + \frac{n}{d}\right)\left(\frac{n}{p} + \frac{n}{d^{3/2}} + \frac{1}{\sqrt{d}}\right)}\right).$$

We can thus identify a scaling in which we expect the two-layer NN to collapse into the equivalent GLM:

$$\widetilde{\lim} \equiv \lim_{n,p,d\to\infty} \text{ s.t. } \quad \left(1 + \frac{n}{d}\right)\left(\frac{n}{p} + \frac{n}{d^{3/2}} + \frac{1}{\sqrt{d}}\right) \to 0\,.$$

Mutual information

Corollary

Under the same hypothesis, the following holds:

$$\left|\frac{1}{n} I_n(\boldsymbol{a}^*, \boldsymbol{W}^*; \mathcal{D}_n) - \frac{1}{n} I_n^\circ(\boldsymbol{v}^*, \boldsymbol{\xi}^*; \mathcal{D}_n^\circ)\right| = O\left(\sqrt{\left(1+\frac{n}{d}\right)\left(\frac{n}{p}+\frac{n}{d^{3/2}}+\frac{1}{\sqrt{d}}\right)}\right).$$

We can thus identify a scaling in which we expect the two-layer NN to collapse into the equivalent GLM:

$$\widetilde{\lim} \equiv \lim_{n,p,d\to\infty} \text{ s.t. } \quad \left(1+\frac{n}{d}\right)\left(\frac{n}{p}+\frac{n}{d^{3/2}}+\frac{1}{\sqrt{d}}\right) \to 0\,.$$

p ~ d allowed! as long as p>>n

Generalization Error

$$\widetilde{\lim} \equiv \lim_{n,p,d\to\infty} \quad \text{s.t.} \quad \left(1+\frac{n}{d}\right)\left(\frac{n}{p}+\frac{n}{d^{3/2}}+\frac{1}{\sqrt{d}}\right) \to 0\,.$$

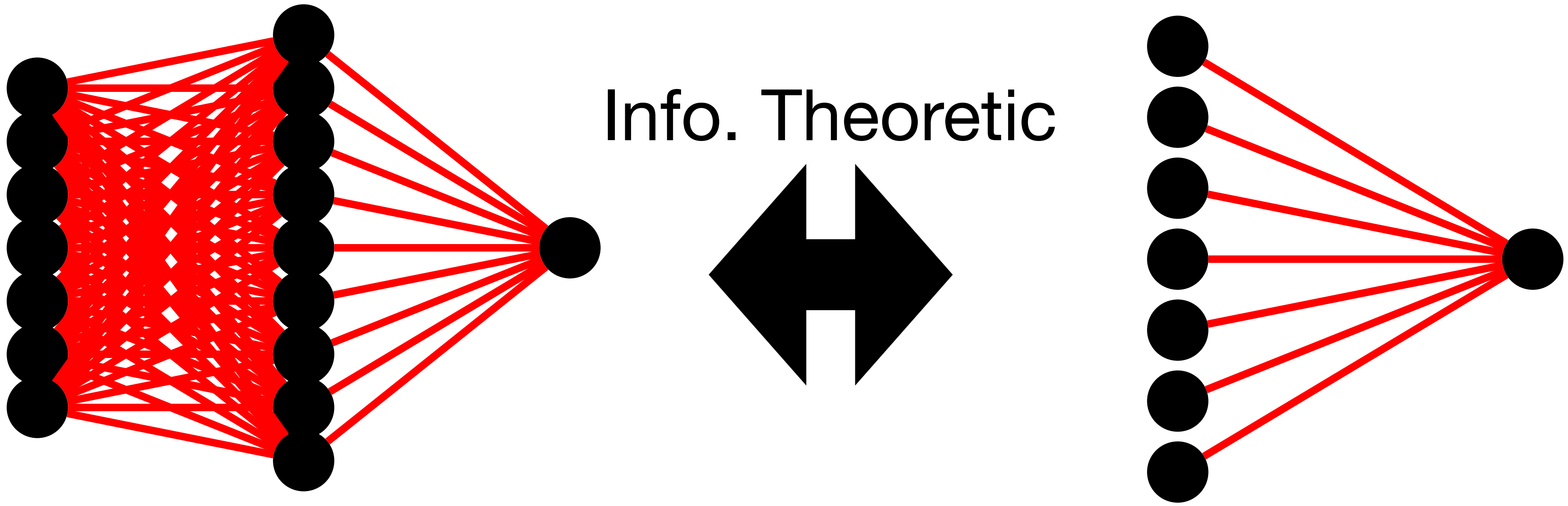
Corollary

Under the same hypothesis, the following holds:

$$\widetilde{\lim}|\mathbb{E}\mathcal{E} - \mathbb{E}\mathcal{E}^\circ| = 0\,,$$

i.e. the 2-layer NN and the GLM share the same generalization error in the $\widetilde{\lim}$.

p ~ d allowed! as long as p>>n



$$Y_\mu \sim P_{\mathrm{out}}\Big(\cdot \mid \frac{\boldsymbol{a}^{*\intercal}}{\sqrt{p}}\varphi\Big(\frac{\boldsymbol{W}^*\boldsymbol{X}_\mu}{\sqrt{d}}\Big)\Big) \qquad Y_\mu^\circ \sim P_{\mathrm{out}}\Big(\cdot \mid \rho\frac{\boldsymbol{v}^{*\intercal}\boldsymbol{X}_\mu}{\sqrt{d}} + \sqrt{\epsilon}\xi_\mu^*\Big)$$

$$\mathcal{D}_n = \{(\boldsymbol{X}_\mu, Y_\mu)_{\mu=1}^n\} \qquad \mathcal{D}_n^\circ = \{(\boldsymbol{X}_\mu, Y_\mu^\circ)_{\mu=1}^n\}$$

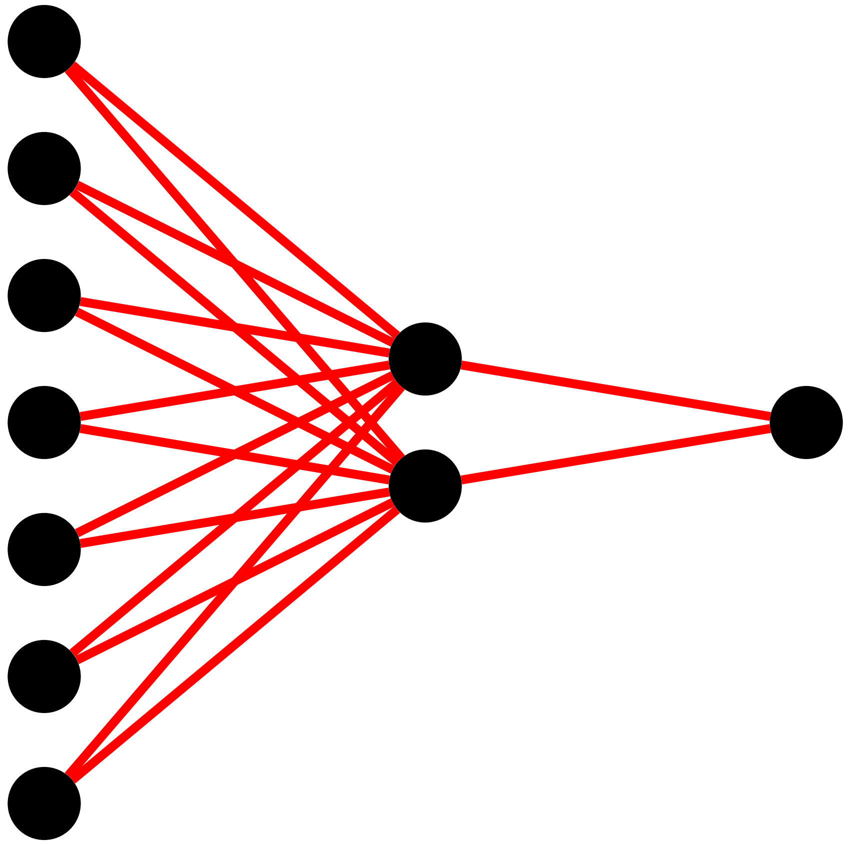
Remark

Our results **do not** imply that training a two-layer NN on $\mathcal{D}_n^\circ$ yields BO generalization error. (or vice versa)

Various scalings

Committee Machines

$$\widetilde{\lim} \equiv \lim_{n,p,d\to\infty} \text{ s.t. } \quad \left(1+\frac{n}{d}\right)\left(\frac{n}{p}+\frac{n}{d^{3/2}}+\frac{1}{\sqrt{d}}\right) \to 0\,.$$



$$p = O(1)\,, \quad \frac{n}{d} = O(1)$$

Warning!

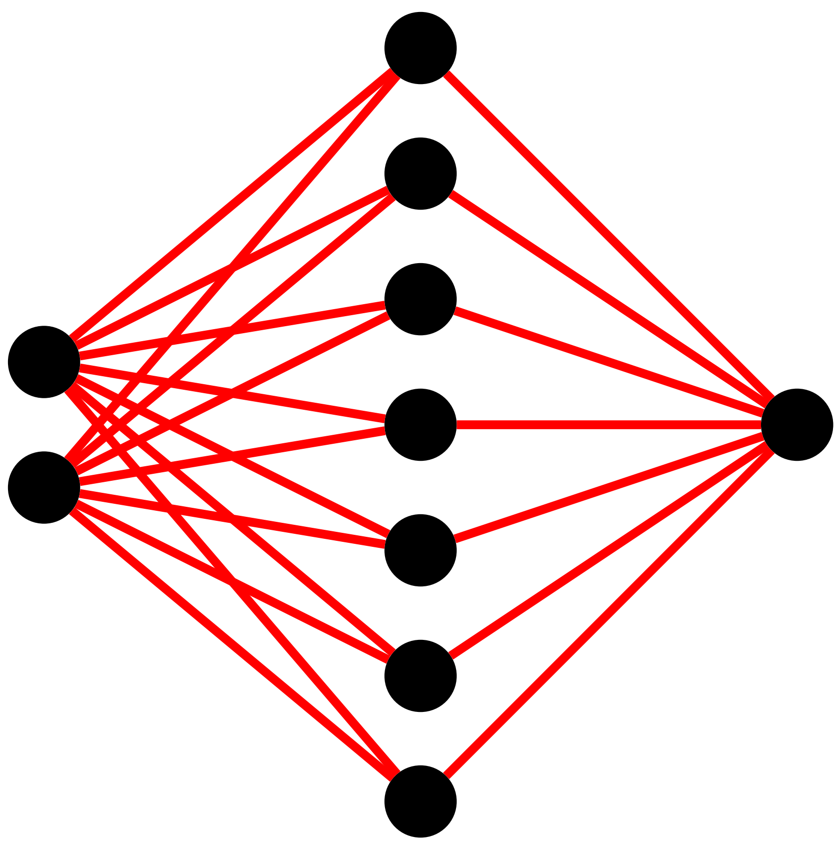
Only $n, d \to \infty$. Different from the our scaling in $\widetilde{\lim}$

If the data are really high-dimensional, the middle layer compresses them too much.

Aubin, Maillard, Barbier, Macris, Krzakala, Zdeborova NeurIPS 18'

Mean field regime

$$\widetilde{\lim} \equiv \lim_{n,p,d\to\infty} \text{ s.t. } \quad \left(1+\frac{n}{d}\right)\left(\frac{n}{p}+\frac{n}{d^{3/2}}+\frac{1}{\sqrt{d}}\right)\to 0\,.$$



In the hypothesis $p \gg d$, one can track the empirical distribution of the weights of the network by means of a distributional equation **[Mei-Montanari-Nguyen-2008]**, regardless of n.

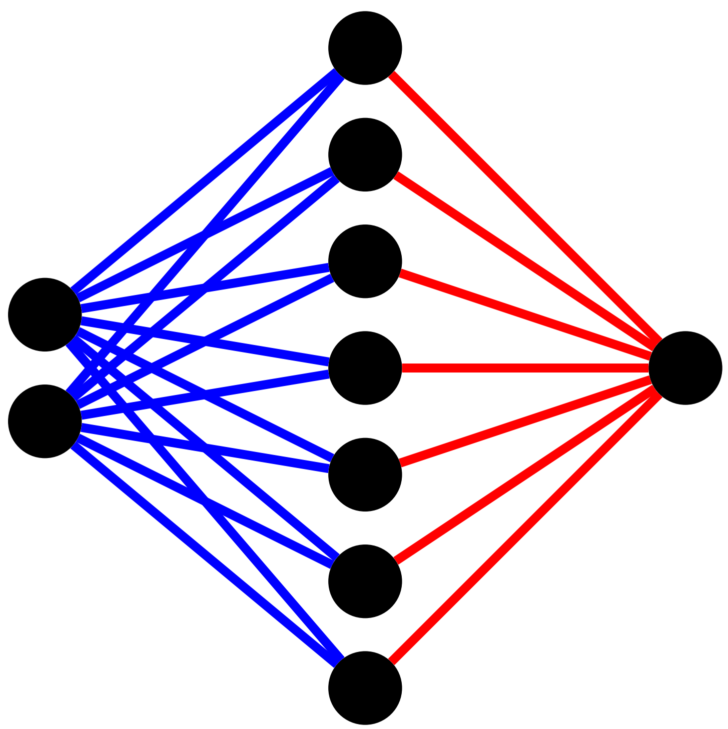
Chizat, Bach; ...

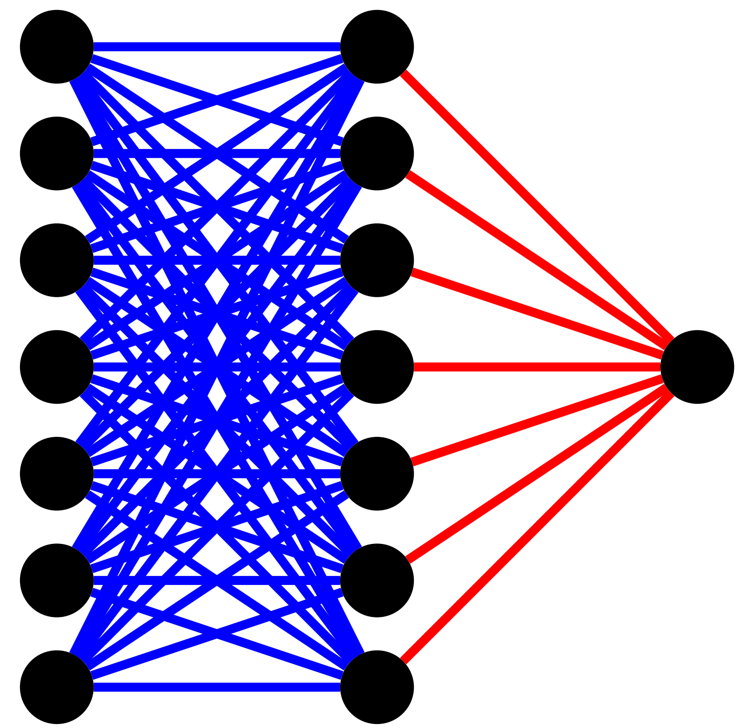
Remark

In our proof and in **[Mei-Montanari-Nguyen-2008]** p seems to play a crucial role!

Frozen hidden weights

Blue weights are **quenched**, they vary negligibly w.r.t. red weights, that are learnable or **annealed**.





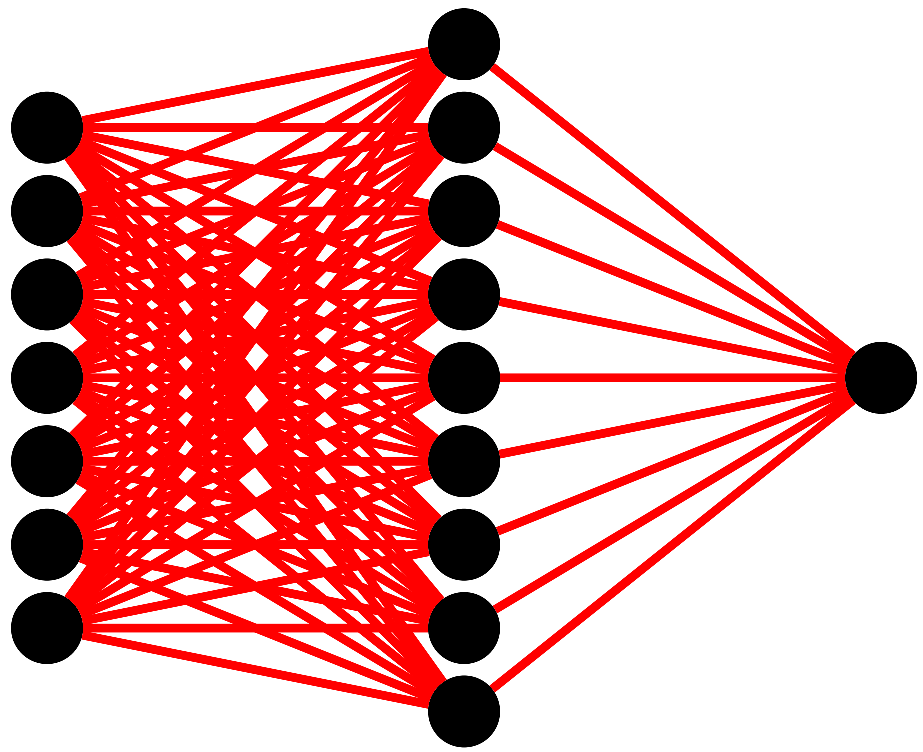
- SGD in MF at initial stage
- NTK   Jacot, Hongler

- Random features models   Rahmini et al
- NNs as Gaussian processes   Neal

Parameter counting

Only p parameters to be learnt, w.r.t. $dp + p$ in our case.

Recent conjectures



Even for deeper networks.

Linear scalings

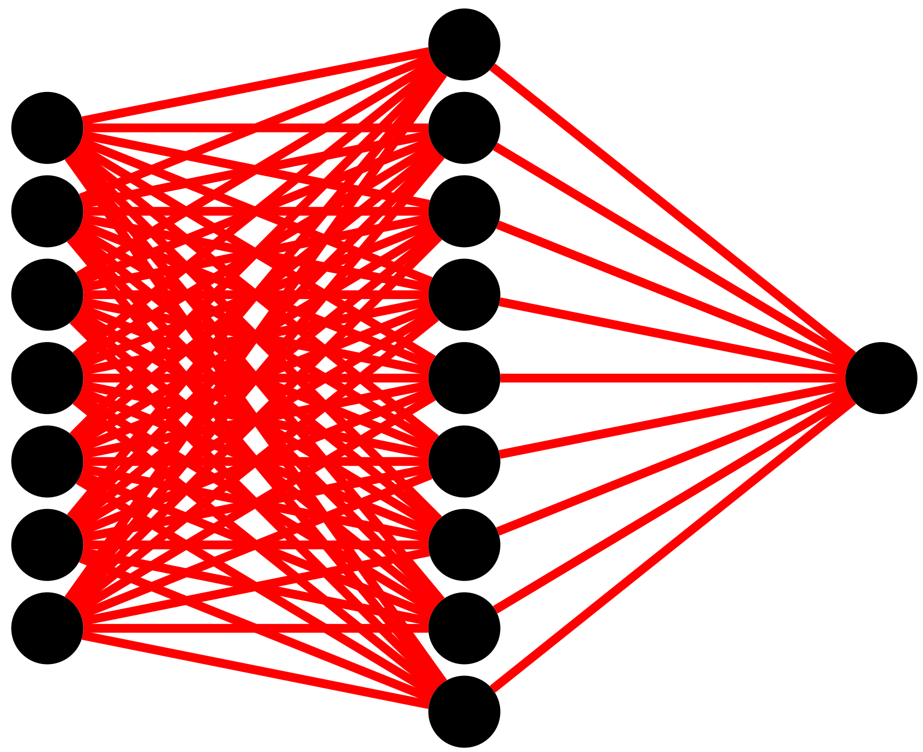
$$n, d, p \to \infty\,, \qquad \frac{n}{p} = O(1) = \frac{d}{p}$$

$\widetilde{\lim}$ does **not** catch these scalings.

We need p>>n

- **[Li-Sampolinsky-2021]** studied full training for **linear** networks;
- **[Ariosto-Pacelli-Gherardi-Rotondo-2022]** conjectures a formula for the ERM - generalization error;
- **[Cui-Krzakala-Zdeborová-2023]** builds on a **Gaussian Equivalence Principle** to compute the Bayes-optimal limits as we do

Recent conjectures



Even for deeper networks.

Linear scalings

$$n, d, p \to \infty \,, \qquad \frac{n}{p} = O(1) = \frac{d}{p}$$

$\widetilde{\lim}$ does **not** catch these scalings.

We need p>>n

- **[Li-Sampolinsky-2021]** studied full training for **linear** networks;
- **[Ariosto-Pacelli-Gherardi-Rotondo-2022]** conjectures a formula for the ERM - generalization error;
- **[Cui-Krzakala-Zdeborová-2023]** builds on a **Gaussian Equivalence Principle** to compute the Bayes-optimal limits as we do

Q: is our proof sub-optimal or is the replica prediction of Cui et al. in the linear scaling not exactly correct?

Proof idea

Universality: Gaussian Equivalence Principles

The most annoying part in the analysis is for sure the non-linearity in the middle layer. To get rid of it, some authors have noticed that

GEP (informal)

It amounts to the following replacement:

$$\varphi\left(\frac{\boldsymbol{W}^*\boldsymbol{X}_\mu}{\sqrt{d}}\right) \approx \rho\frac{\boldsymbol{W}^*\boldsymbol{X}_\mu}{\sqrt{d}} + \sqrt{\epsilon}\xi^*_\mu$$

with ξ^*_μ an independent standard Gaussian noise and

$$\rho = \mathbb{E}_{\mathcal{N}(0,1)}\varphi' \,, \quad \epsilon = \mathbb{E}_{\mathcal{N}(0,1)}\varphi^2 - (\mathbb{E}_{\mathcal{N}(0,1)}\varphi')^2$$

In our setting, it is not clear to what extent this is applicable!

Interpolation

The interpolation has to keep all the ingredients together:

$$S_{t\mu} := \sqrt{1-t}\frac{\boldsymbol{a}^{*\intercal}}{\sqrt{p}}\varphi\Big(\frac{\boldsymbol{W}^*\boldsymbol{X}_\mu}{\sqrt{d}}\Big) + \sqrt{t}\rho\frac{\boldsymbol{v}^{*\intercal}\boldsymbol{X}_\mu}{\sqrt{d}} + \sqrt{t\epsilon}\xi^*_\mu\,,$$

$$s_{t\mu} := \sqrt{1-t}\frac{\boldsymbol{a}^{\intercal}}{\sqrt{p}}\varphi\Big(\frac{\boldsymbol{W}\boldsymbol{X}_\mu}{\sqrt{d}}\Big) + \sqrt{t}\rho\frac{\boldsymbol{v}^{\intercal}\boldsymbol{X}_\mu}{\sqrt{d}} + \sqrt{t\epsilon}\xi_\mu\,.$$

Interpolating dataset:

$$\mathcal{D}_{n,t} = \{(\boldsymbol{X}_\mu, Y_{t\mu})_{\mu=1}^n\}\,, \quad Y_{t\mu} \sim P_{\mathrm{out}}\big(\cdot \mid S_{t\mu}\big)$$

Interpolating free entropy:

$$\bar{f}_n(t) = \frac{1}{n}\mathbb{E}_{\mathcal{D}_{n,t}} \log \int D\boldsymbol{v} D\boldsymbol{\xi} D\boldsymbol{a} D\boldsymbol{W} \prod_{\mu=1}^n P_{\mathrm{out}}\big(Y_{t\mu} \mid s_{t\mu}\big)$$

Interpolation

Interpolating free entropy:

$$\bar{f}_n(t) = \frac{1}{n}\mathbb{E}_{\mathcal{D}_{n,t}} \log \underbrace{\int D\boldsymbol{v} D\boldsymbol{\xi} D\boldsymbol{a} D\boldsymbol{W} \prod_{\mu=1}^{n} P_{\mathrm{out}}\big(Y_{t\mu} \mid s_{t\mu}(\boldsymbol{a}, \boldsymbol{W}, \boldsymbol{v}, \xi_\mu)\big)}_{\mathcal{Z}_t}$$

Main Goal

Prove that

$$\frac{d}{dt}\bar{f}_n(t) = O\Big(\sqrt{\Big(1 + \frac{n}{d}\Big)\Big(\frac{n}{p} + \frac{n}{d^{3/2}} + \frac{1}{\sqrt{d}}\Big)}\Big),$$

uniformly in $t \in [0, 1]$.

Cross terms

$$\frac{d}{dt}\bar{f}_n(t) = -A_1 + A_2 + A_3 + B$$

Let $u_y(x) = \log P_{\mathrm{out}}(y \mid x)$, then

$$A_1 := \frac{1}{2n}\mathbb{E}_{(t)} \log \mathcal{Z}_t \sum_{\mu=1}^{n} u'_{Y_{t\mu}}(S_{t\mu}) \frac{\boldsymbol{a}^{*\top}}{\sqrt{(1-t)p}} \varphi\Big(\frac{\boldsymbol{W}^*\boldsymbol{X}_\mu}{\sqrt{d}}\Big),$$

$$A_2 := \frac{1}{2n}\mathbb{E}_{(t)} \log \mathcal{Z}_t \sum_{\mu=1}^{n} u'_{Y_{t\mu}}(S_{t\mu}) \rho \frac{\boldsymbol{v}^{*\top}\boldsymbol{X}_\mu}{\sqrt{td}},$$

$$A_3 := \frac{1}{2n}\mathbb{E}_{(t)} \log \mathcal{Z}_t \sum_{\mu=1}^{n} u'_{Y_{t\mu}}(S_{t\mu}) \sqrt{\frac{\epsilon}{t}} \xi^*_\mu,$$

$$B := \frac{1}{n}\mathbb{E}_{(t)} \Big\langle \sum_{\mu=1}^{n} u'_{Y_{t\mu}}(s_{t\mu}) \frac{ds_{t\mu}}{dt} \Big\rangle.$$

Let $u_y(x) = \log P_{\mathrm{out}}(y \mid x)$, then

$$\frac{1}{2n}\mathbb{E}_{(t)} \log \mathcal{Z}_t \sum_{\mu=1}^{n} u'_{Y_{t\mu}}(S_{t\mu})\Big[\frac{\boldsymbol{a}^{*\intercal}}{\sqrt{(1-t)p}}\varphi\Big(\frac{\boldsymbol{W}^{*}\boldsymbol{X}_{\mu}}{\sqrt{d}}\Big) - \rho\frac{\boldsymbol{a}^{*\intercal}\boldsymbol{W}^{*}\boldsymbol{X}_{\mu}}{\sqrt{pd}}\Big] +$$

$$+ \frac{1}{2n}\mathbb{E}_{(t)} \log \mathcal{Z}_t \sum_{\mu=1}^{n} u'_{Y_{t\mu}}(S_{t\mu})\rho\frac{\boldsymbol{a}^{*\intercal}\boldsymbol{W}^{*}\boldsymbol{X}_{\mu}}{\sqrt{pd}}$$

Let $u_y(x) = \log P_{\rm out}(y \mid x)$, then

$$\frac{1}{2n}\mathbb{E}_{(t)} \log \mathcal{Z}_t \sum_{\mu=1}^{n} u'_{Y_{t\mu}}(S_{t\mu}) \Big[\frac{\boldsymbol{a}^{*\intercal}}{\sqrt{(1-t)p}} \varphi\Big(\frac{\boldsymbol{W}^*\boldsymbol{X}_\mu}{\sqrt{d}}\Big) - \rho \frac{\boldsymbol{a}^{*\intercal}\boldsymbol{W}^*\boldsymbol{X}_\mu}{\sqrt{pd}} \Big] +$$

$$+ \frac{1}{2n}\mathbb{E}_{(t)} \log \mathcal{Z}_t \sum_{\mu=1}^{n} u'_{Y_{t\mu}}(S_{t\mu}) \rho \frac{\boldsymbol{a}^{*\intercal}\boldsymbol{W}^*\boldsymbol{X}_\mu}{\sqrt{pd}}$$

Gaussian integration by part (here w.r.t. a*)

$$= \frac{1}{2n}\mathbb{E}_{(t)} \log \mathcal{Z}_t \sum_{\mu,\nu=1}^{n} U_{\mu\nu} \Big[\frac{\varphi(\boldsymbol{\alpha}_\mu)^\intercal \varphi(\boldsymbol{\alpha}_\nu) - \rho \boldsymbol{\alpha}_\mu^\intercal \varphi(\boldsymbol{\alpha}_\nu)}{p} \Big]$$

$$\boldsymbol{\alpha}_\mu := \frac{\mathbf{W}^*\mathbf{X}_\mu}{\sqrt{d}}$$

Let $u_y(x) = \log P_{\text{out}}(y \mid x)$, then

$$\frac{1}{2n}\mathbb{E}_{(t)} \log \mathcal{Z}_t \sum_{\mu=1}^{n} u'_{Y_{t\mu}}(S_{t\mu})\Big[\frac{\boldsymbol{a}^{*\intercal}}{\sqrt{(1-t)p}}\varphi\Big(\frac{\boldsymbol{W}^*\boldsymbol{X}_\mu}{\sqrt{d}}\Big) - \rho\frac{\boldsymbol{a}^{*\intercal}\boldsymbol{W}^*\boldsymbol{X}_\mu}{\sqrt{pd}}\Big] +$$

$$+\frac{1}{2n}\mathbb{E}_{(t)} \log \mathcal{Z}_t \sum_{\mu=1}^{n} u'_{Y_{t\mu}}(S_{t\mu})\rho\frac{\boldsymbol{a}^{*\intercal}\boldsymbol{W}^*\boldsymbol{X}_\mu}{\sqrt{pd}}$$

Gaussian integration by part (here w.r.t. a*)

$$= \frac{1}{2n}\mathbb{E}_{(t)} \log \mathcal{Z}_t \sum_{\mu,\nu=1}^{n} U_{\mu\nu}\Big[\frac{\varphi(\boldsymbol{\alpha}_\mu)^\intercal\varphi(\boldsymbol{\alpha}_\nu) - \rho\boldsymbol{\alpha}_\mu^\intercal\varphi(\boldsymbol{\alpha}_\nu)}{p}\Big]$$

$$\boldsymbol{\alpha}_\mu := \frac{\mathbf{W}^*\mathbf{X}_\mu}{\sqrt{d}}$$

Getting rid of non-linearities by approximation

$$\mathbb{E}_{\mathbf{W}^*}\varphi'(\alpha_{\mu i}) = \rho + O\Big(\frac{\|\mathbf{X}_\mu\|^2}{d} - 1\Big),$$

$$\mathbb{E}_{\mathbf{W}^*}\varphi^2(\alpha_{\mu i}) = \mathbb{E}_{\mathcal{N}(0,1)}\varphi^2 + O\Big(\frac{\|\mathbf{X}_\mu\|^2}{d} - 1\Big),$$

$$\mathbb{E}_{\mathbf{W}^*}\varphi(\alpha_{\mu i})\varphi(\alpha_{\nu i}) = \mathbb{E}_{\mathcal{N}(0,1)}\varphi'\frac{\mathbf{X}_\mu^\intercal\mathbf{X}_\nu}{d} + O\Big(\frac{\mathbf{X}_\mu^\intercal\mathbf{X}_\nu}{d}\Big(\frac{\|\mathbf{X}_\mu\|^2}{d} - 1\Big)\Big) + O\Big(\Big(\frac{\mathbf{X}_\mu^\intercal\mathbf{X}_\nu}{\|\mathbf{X}_\nu\|^2}\Big)^2\Big) + O\Big(\frac{(\mathbf{X}_\mu^\intercal\mathbf{X}_\nu)^2}{\|\mathbf{X}_\nu\|^2 d}\Big),$$

Let $u_y(x) = \log P_{\text{out}}(y \mid x)$, then

$$\frac{1}{2n}\mathbb{E}_{(t)} \log \mathcal{Z}_t \sum_{\mu=1}^{n} u'_{Y_{t\mu}}(S_{t\mu})\Big[\frac{\boldsymbol{a}^{*\intercal}}{\sqrt{(1-t)p}}\varphi\Big(\frac{\boldsymbol{W}^*\boldsymbol{X}_\mu}{\sqrt{d}}\Big) - \rho\frac{\boldsymbol{a}^{*\intercal}\boldsymbol{W}^*\boldsymbol{X}_\mu}{\sqrt{pd}}\Big] +$$

$$+\frac{1}{2n}\mathbb{E}_{(t)} \log \mathcal{Z}_t \sum_{\mu=1}^{n} u'_{Y_{t\mu}}(S_{t\mu})\rho\frac{\boldsymbol{a}^{*\intercal}\boldsymbol{W}^*\boldsymbol{X}_\mu}{\sqrt{pd}}$$

Gaussian integration by part (here w.r.t. a*)

$$= \frac{1}{2n}\mathbb{E}_{(t)} \log \mathcal{Z}_t \sum_{\mu,\nu=1}^{n} U_{\mu\nu}\Big[\frac{\varphi(\boldsymbol{\alpha}_\mu)^\intercal\varphi(\boldsymbol{\alpha}_\nu) - \rho\boldsymbol{\alpha}_\mu^\intercal\varphi(\boldsymbol{\alpha}_\nu)}{p}\Big]$$

$$\boldsymbol{\alpha}_\mu := \frac{\mathbf{W}^*\mathbf{X}_\mu}{\sqrt{d}}$$

Getting rid of non-linearities by approximation

$$\mathbb{E}_{\mathbf{W}^*}\varphi'(\alpha_{\mu i}) = \rho + O\Big(\frac{\|\mathbf{X}_\mu\|^2}{d} - 1\Big),$$

$$\mathbb{E}_{\mathbf{W}^*}\varphi^2(\alpha_{\mu i}) = \mathbb{E}_{\mathcal{N}(0,1)}\varphi^2 + O\Big(\frac{\|\mathbf{X}_\mu\|^2}{d} - 1\Big),$$

$$\mathbb{E}_{\mathbf{W}^*}\varphi(\alpha_{\mu i})\varphi(\alpha_{\nu i}) = \mathbb{E}_{\mathcal{N}(0,1)}\varphi'\frac{\mathbf{X}_\mu^\intercal\mathbf{X}_\nu}{d} + O\Big(\frac{\mathbf{X}_\mu^\intercal\mathbf{X}_\nu}{d}\Big(\frac{\|\mathbf{X}_\mu\|^2}{d} - 1\Big)\Big) + O\Big(\Big(\frac{\mathbf{X}_\mu^\intercal\mathbf{X}_\nu}{\|\mathbf{X}_\nu\|^2}\Big)^2\Big) + O\Big(\frac{(\mathbf{X}_\mu^\intercal\mathbf{X}_\nu)^2}{\|\mathbf{X}_\nu\|^2 d}\Big),$$

It is possible that the replica prediction is a good approximation in the proportional regimes, but only exact when inputs are orthogonal (while gaussians have a weak overlap)

$$\frac{1}{2}\mathbb{E}_{(t)}\overbrace{\frac{(\log \mathcal{Z}_t - \mathbb{E}_{(t)}\log \mathcal{Z}_t)}{n}}^{O\left(\sqrt{\frac{1}{n}+\frac{1}{d}}\right)}\sum_{\mu,\nu=1}^{n} U_{\mu\nu}\Big[\frac{1}{p}\varphi\Big(\frac{\boldsymbol{W}^*\boldsymbol{X}_\mu}{\sqrt{d}}\Big)^\intercal\varphi\Big(\frac{\boldsymbol{W}^*\boldsymbol{X}_\nu}{\sqrt{d}}\Big) - \frac{\rho}{p}\varphi\Big(\frac{\boldsymbol{W}^*\boldsymbol{X}_\mu}{\sqrt{d}}\Big)^\intercal\frac{\boldsymbol{W}^*\boldsymbol{X}_\nu}{\sqrt{d}}\Big]$$

Finally...

$$\sum_{\mu,\nu} U_{\mu\nu}[RED - BLUE]_{\mu\nu} \sim O\left(n\sqrt{\frac{1}{p}+\frac{1}{d^{3/2}}}\right)$$

Conclusion

Conclusion and perspectives (Work in progress)

- Well... what about $n/p = O(1)$? Are we actually able to reach physicists' conjectured regimes?
- Say we have partial information on $\boldsymbol{W}^*$ through a Gaussian channel: $\tilde{\boldsymbol{W}} = \sqrt{\sigma}\boldsymbol{W}^* + \boldsymbol{Z}$.
  - When $\sigma \to \infty \Rightarrow$ RF model in BO setting
  - When $\sigma = 0 \Rightarrow$ our setting

  We could interpolate between the two!
- Why not more than two layers?
  - We only miss concentration of the free entropy, then an inductive argument and concentration by parts should yield the result in a similar scaling.
- What happens if we add strcture to the data? $\boldsymbol{X} \sim \mathcal{N}(0, \Sigma)$ or mixtures.

(Some) References

- **Song Mei**, **Andrea Montanari** and **Phan-Minh Nguyen**, *"A mean field view of the landscape of two-layer neural networks"*, Proceedings of the National Academy of Sciences, v. 115, 2008;
- **Qianyi Li** and **Haim Sompolinsky**, *"Statistical Mechanics of Deep Linear Neural Networks: The Backpropagating Kernel Renormalization"*, Phys. Rev. X, v. 11, 2021;
- **S. Ariosto**, **R. Pacelli**, **M. Gherardi** and **P. Rotondo**, *"Statistical mechanics of deep learning beyond the infinite-width limit"*, arXiv:2209.04882, 2022;
- **Hugo Cui**, **Florent Krzakala** and **Lenka Zdeborová**, *"Optimal Learning of Deep Random Networks of Extensive-width"*, arXiv:2302.00375, 2023;
- **Hong Hu** and **Yue M Lu**, *"Universality laws for high-dimensional learning with random features"*, IEEE Transactions on Information Theory, 2022;
- **Sebastian Goldt**, **Bruno Loureiro**, **Galen Reeves**, **Florent Krzakala**, **Marc Mézard** and **Lenka Zdeborová**, *"The Gaussian equivalence of generative models for learning with shallow neural networks"*, Mathematical and Scientific Machine Learning PMLR, 2022.