Information-theoretic reduction of deep neural networks to linear models in the overparametrized proportional regime

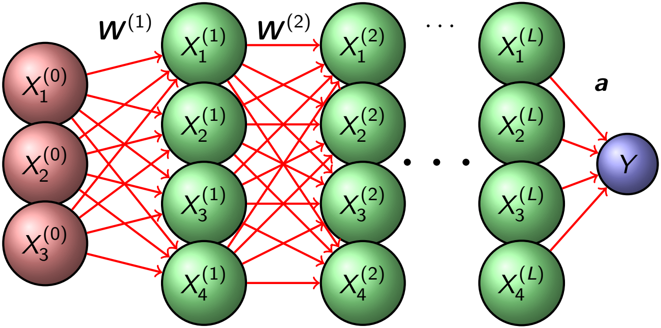
F. Camilli, **D. Tieplova**, J. Barbier, and E. Bergamin

International Center for Theoretical Physics, Trieste, Italia
($\rightarrow$ Aaurhus University, Denmark)

Log-gases in Caeli Australi, MATRIX Institute, Creswick, Australia
4-15 Aug 2025

Deep Neural Network

$\mathbf{X}^{(0)} \sim \mathcal{N}(0, d_0^{-1}\mathbb{I}_{d_0})$



$$\boldsymbol{X}^{(\ell)} = d_{\ell-1}^{-1/2}\varphi\left(\boldsymbol{W}^{(\ell)}\boldsymbol{X}^{(\ell-1)}\right) \in \mathbb{R}^{d_\ell}$$

φ – activation function

$$Y = f\left(\boldsymbol{a}^\intercal \boldsymbol{X}^{(L)}\right)$$

f –readout function.

Supervised learning: Starting from a *training set* $\mathcal{D}_n^{(L)} = \{(\boldsymbol{X}_\mu^{(0)}, Y_\mu)_{\mu=1}^n\}$, we adjust the weights $\boldsymbol{a}, \boldsymbol{W}^{(1)}, \dots, \boldsymbol{W}^{(L)}$ s.t.

$$Y_\mu \approx f\left(\boldsymbol{a}^\intercal \varphi\left(\boldsymbol{W}^{(L)} \varphi\left(\dots \varphi\left(\boldsymbol{W}^{(1)} \boldsymbol{X}_\mu^{(0)}\right)\right)\right)\right), \quad \forall \mu .$$

Main Goal

Produce the smallest possible generalization error:

$$\mathcal{E} = \left[Y_{\text{new}} - f\left(\boldsymbol{a}^\intercal \boldsymbol{X}_{\text{new}}^{(L)}\right)\right]^2$$

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

Supervised learning: Starting from a *training set* $\mathcal{D}_n^{(L)} = \{(\boldsymbol{X}_\mu^{(0)}, Y_\mu)_{\mu=1}^n\}$, we adjust the weights $\boldsymbol{a}, \boldsymbol{W}^{(1)}, \dots, \boldsymbol{W}^{(L)}$ s.t.

$$Y_\mu \approx f\left(\boldsymbol{a}^\intercal \varphi\left(\boldsymbol{W}^{(L)} \varphi\left(\dots \varphi\left(\boldsymbol{W}^{(1)} \boldsymbol{X}_\mu^{(0)}\right)\right)\right)\right), \quad \forall \mu\,.$$

Main Goal

Produce the smallest possible generalization error:

$$\mathcal{E} = \left[Y_{\rm new} - f\left(\boldsymbol{a}^\intercal \boldsymbol{X}_{\rm new}^{(L)}\right)\right]^2$$

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

Teacher-student setup

The training set is generated by a L-layer **teacher network** with matching architecture:

$$Y_\mu = f\left(\boldsymbol{a}^{*\intercal} \boldsymbol{X}_\mu^{(L)}\right) + \sqrt{\Delta} Z_\mu\,, \quad \forall \mu \leq n$$

$$\boldsymbol{X}_\mu^{(\ell)} = \varphi\left(\boldsymbol{W}^{(\ell)*} \boldsymbol{X}_\mu^{(\ell-1)}\right), \quad \ell = 1, \dots, L\,,$$

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

$$Y_\mu \sim P_{\text{out}}\left(\cdot \mid \boldsymbol{a}^{*\intercal} \boldsymbol{X}_\mu^{(L)}\right)$$

with

$$P_{\text{out}}(y \mid x) = \frac{1}{\sqrt{2\pi\Delta}} \exp\left(-\frac{1}{2\Delta}\left(f(x) - y\right)^2\right)$$

Bayes-optimal student

Definition (informal)

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

$$Y_\mu = f\left(\boldsymbol{a}^{*\intercal} \boldsymbol{X}_\mu^{(L)}\right) + \sqrt{\Delta} Z_\mu \,, \quad \forall \mu \leq n \,,$$

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

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

$$dP(\boldsymbol{\theta}^{(L)} \mid \mathcal{D}_n^{(L)}) = \frac{1}{\mathcal{Z}(\mathcal{D}_n^{(L)})} \prod_{\mu=1}^{n} P_{\mathrm{out}}\left(Y_\mu \mid \boldsymbol{a}^\intercal \boldsymbol{x}_\mu^{(L)}\right) D\boldsymbol{\theta}^{(L)}$$

where $D\boldsymbol{\theta}^{(L)} = D\boldsymbol{a} D\boldsymbol{W}^{(1)} \dots \boldsymbol{W}^{(L)}$ is the Gaussian prior on the weights.

Why Bayes-optimal?

Proposition (informal)

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

$$\mathbb{E}\mathcal{E} := \mathbb{E}\big(Y_{\mathrm{new}} - \hat{Y}(\mathcal{D}_n^{(L)}, \boldsymbol{X}_{\mathrm{new}}^{(0)})\big)^2$$

that is yielded by the BO predictor

$$\begin{aligned}\hat{Y}_{\mathrm{Bayes}}(\mathcal{D}_n^{(L)}, \boldsymbol{X}_{\mathrm{new}}^{(0)}) &= \mathbb{E}[Y_{\mathrm{new}} \mid \mathcal{D}_n^{(L)}, \boldsymbol{X}_{\mathrm{new}}^{(0)}] \\ &= \int dY\, Y\, P_{\mathrm{out}}\Big(Y \mid \boldsymbol{a}^\intercal \boldsymbol{x}_{\mathrm{new}}^{(L)}\Big) dP(\boldsymbol{\theta}^{(L)} \mid \mathcal{D}_n^{(L)}) .\end{aligned}$$

Main information theoretic quantities

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

$$\mathcal{Z}(\mathcal{D}_n^{(L)}) = \int \prod_{\mu=1}^{n} P_{\text{out}}\left(Y_\mu \mid \boldsymbol{a}^\intercal \boldsymbol{x}_\mu^{(L)}\right) D\boldsymbol{\theta}^{(L)}$$

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

$$\frac{I_n^{(L)}(\boldsymbol{\theta}^{(L)*};\mathcal{D}_n^{(L)})}{n} = \frac{H(\mathcal{D}_n^{(L)})}{n} - \frac{H(\mathcal{D}_n^{(L)} \mid \boldsymbol{\theta}^{(L)*})}{n}$$

$$= -\bar{f}_n^{(L)} + \mathbb{E}\log P_{\text{out}}\left(Y_1 \mid \boldsymbol{a}^{*\intercal}\boldsymbol{X}_1^{(L)}\right)$$

Main information theoretic quantities

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

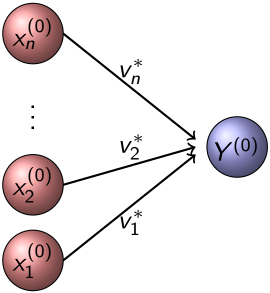
$$\mathcal{Z}(\mathcal{D}_n^{(L)}) = \int \prod_{\mu=1}^{n} P_{\mathrm{out}}\left(Y_\mu \mid \boldsymbol{a}^\intercal \boldsymbol{x}_\mu^{(L)}\right) D\boldsymbol{\theta}^{(L)}$$

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

$$\frac{I_n^{(L)}(\boldsymbol{\theta}^{(L)*};\mathcal{D}_n^{(L)})}{n} = \frac{H(\mathcal{D}_n^{(L)})}{n} - \frac{H(\mathcal{D}_n^{(L)} \mid \boldsymbol{\theta}^{(L)*})}{n}$$

$$= -\bar{f}_n^{(L)} + \mathbb{E}\log P_{\mathrm{out}}\left(Y_1 \mid \boldsymbol{a}^{*\intercal} \boldsymbol{X}_1^{(L)}\right)$$

A simpler ancestor: the GLM



The teacher Generalized Linear Model:

$$Y_\mu^{(0)} = f\left(\rho \boldsymbol{v}^{*\intercal} \boldsymbol{X}_\mu^{(0)} + \sqrt{\epsilon}\xi_\mu^*\right) + \sqrt{\Delta} Z_\mu \,,$$

or $\quad Y_\mu^{(0)} \sim P_{\rm out}\left(\cdot \mid \rho \boldsymbol{v}^{*\intercal} \boldsymbol{X}_\mu^{(0)} + \sqrt{\epsilon}\xi_\mu^*\right)$

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

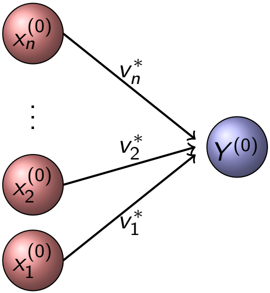
Free entropy:

$$\bar{f}_n^{(0)} = \frac{1}{n}\mathbb{E}\log \int \prod_{\mu=1}^{n} P_{\rm out}\left(Y_\mu^{(0)} \mid \rho \boldsymbol{v}^\intercal \boldsymbol{X}_\mu^{(0)} + \sqrt{\epsilon}\xi_\mu\right) D\boldsymbol{v} D\xi$$

Mutual information:

$$\frac{1}{n} I_n^{(0)}(\boldsymbol{v}^*, \xi^*; \mathcal{D}_n^{(0)}) = -\bar{f}_n^{(0)} + \mathbb{E}\log P_{\rm out}\left(Y_1^{(0)} \mid \rho \boldsymbol{v}^{*\intercal} \boldsymbol{X}_1^{(0)} + \sqrt{\epsilon}\xi_1^*\right).$$

A simpler ancestor: the GLM



The teacher Generalized Linear Model:

$$Y_\mu^{(0)} = f\left(\rho \boldsymbol{v}^{*\intercal} \boldsymbol{X}_\mu^{(0)} + \sqrt{\epsilon}\xi_\mu^*\right) + \sqrt{\Delta} Z_\mu\,,$$

$$\text{or} \quad Y_\mu^{(0)} \sim P_{\mathrm{out}}\left(\cdot \mid \rho \boldsymbol{v}^{*\intercal} \boldsymbol{X}_\mu^{(0)} + \sqrt{\epsilon}\xi_\mu^*\right)$$

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

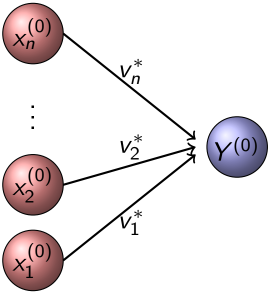
Free entropy:

$$\bar{f}_n^{(0)} = \frac{1}{n}\mathbb{E}\log \int \prod_{\mu=1}^{n} P_{\mathrm{out}}\left(Y_\mu^{(0)} \mid \rho \boldsymbol{v}^\intercal \boldsymbol{X}_\mu^{(0)} + \sqrt{\epsilon}\xi_\mu\right) D\boldsymbol{v} D\xi$$

Mutual information:

$$\frac{1}{n} I_n^{(0)}(\boldsymbol{v}^*, \xi^*; \mathcal{D}_n^{(0)}) = -\bar{f}_n^{(0)} + \mathbb{E}\log P_{\mathrm{out}}\left(Y_1^{(0)} \mid \rho \boldsymbol{v}^{*\intercal} \boldsymbol{X}_1^{(0)} + \sqrt{\epsilon}\xi_1^*\right).$$

A simpler ancestor: the GLM



The teacher Generalized Linear Model:

$$Y_\mu^{(0)} = f\left(\rho \boldsymbol{v}^{*\intercal} \boldsymbol{X}_\mu^{(0)} + \sqrt{\epsilon}\xi_\mu^*\right) + \sqrt{\Delta} Z_\mu\,,$$

$$\text{or} \quad Y_\mu^{(0)} \sim P_{\rm out}\left(\cdot \mid \rho \boldsymbol{v}^{*\intercal} \boldsymbol{X}_\mu^{(0)} + \sqrt{\epsilon}\xi_\mu^*\right)$$

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

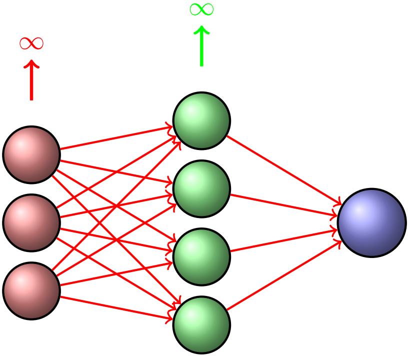
Free entropy:

$$\bar{f}_n^{(0)} = \frac{1}{n}\mathbb{E}\log\int\prod_{\mu=1}^{n} P_{\rm out}\left(Y_\mu^{(0)} \mid \rho \boldsymbol{v}^\intercal \boldsymbol{X}_\mu^{(0)} + \sqrt{\epsilon}\xi_\mu\right) D\boldsymbol{v} D\boldsymbol{\xi}$$

Mutual information:

$$\frac{1}{n} I_n^{(0)}(\boldsymbol{v}^*, \boldsymbol{\xi}^*; \mathcal{D}_n^{(0)}) = -\bar{f}_n^{(0)} + \mathbb{E}\log P_{\rm out}\left(Y_1^{(0)} \mid \rho \boldsymbol{v}^{*\intercal} \boldsymbol{X}_1^{(0)} + \sqrt{\epsilon}\xi_1^*\right).$$

Recent conjectures



Linear scalings

$$n, d_L, \ldots, d_0 \to \infty\,, \qquad \frac{n}{d_\ell} = O(1)$$

- **[Li-Sompolinsky-2021]** studied full training for **linear** networks;
- **[Ariosto-Pacelli-Pastore-Ginelli-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

Gaussian Equivalence Principles

GEP (informal)

It amounts to the following replacement:

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

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!

In the context of random kernel matrices $\mathbf{\Phi} = \varphi^\intercal(\boldsymbol{W}^*\boldsymbol{X})\varphi(\boldsymbol{W}^*\boldsymbol{X})$, also known as **Conjugate Kernel**.

- **[Louart-Liao-Couillet-2017]** found deterministic equivalent of the resolvent of $\mathbf{\Phi}$ has similar behavior with sample covariance models
- **[Pennington-Worah-2017]** spectral distribution of $\mathbf{\Phi}$
- **[Fan-Wang-2020]** spectral distribution for multilayer conjugate kernel

One layer reduction

- $\varphi, f \in C^2(\mathbb{R})$ are odd and with bounded first and second derivatives.
- $\sigma_0 = 1\,,\ \sigma_\ell = \mathbb{E}\varphi^2(Z\sqrt{\sigma_{\ell-1}})\,, \quad \rho_\ell = \mathbb{E}\varphi'(Z\sqrt{\sigma_{\ell-1}}), \quad \epsilon_\ell = \sigma_\ell - \sigma_{\ell-1}\rho_\ell^2\,,$ where Z is a standard Gaussian.

$$\left|\frac{1}{n}I_n^{(L)} - \frac{1}{n}I_n^{(L-1)}\right| = O\left(\left(1 + \sqrt{\frac{n}{d_{min}}} + \frac{n}{d_{min}}\right)\frac{1}{\sqrt{d_{min}}}\right),$$

where responses for $L-1$-layer NN are drawn as

$$Y_\mu^{(L-1)} \sim P_{out}\left(\cdot \mid \rho_L \mathbf{a}^{*\intercal}\mathbf{X}_\mu^{(L-1)} + \sqrt{\epsilon_L}\xi_\mu^*\right).$$

Mutual information equivalence

Under the same assumptions, the following holds true:

$$\left| \frac{1}{n} I_n^{(L)} - \frac{1}{n} I_n^{(0)} \right| = O\left(\left(1 + \sqrt{\frac{n}{d_{min}}} + \frac{n}{d_{min}} \right) \frac{1}{\sqrt{d_{min}}} \right),$$

where $I_n^{(0)}$ is the mutual information associated with the data set $\mathcal{D}_n^{(0)}$ with responses

$$Y_\mu^{(0)} \sim P_{out}\left(\cdot \mid \eta_0 \mathbf{a}^{*\intercal} \mathbf{X}_\mu^{(0)} + \sqrt{\gamma_0} \xi_\mu^* \right)$$

and

$$\eta_0 = \prod_{i=1}^{L} \rho_i \,, \qquad \gamma_0 = \sum_{j=1}^{L} \epsilon_j \prod_{i=j+1}^{L} \rho_i^2 \,.$$

Generalization Error

$$\widetilde{\lim} \equiv \lim_{n, d_\ell \to \infty} \text{ s.t. } \quad \left(1 + \sqrt{\frac{n}{d_{\min}}} + \frac{n}{d_{\min}}\right) \frac{1}{\sqrt{d_{\min}}} \to 0 \,.$$

Corollary

Under the same hypothesis the following holds

$$\widetilde{\lim}\, |\mathcal{E}^{(L)} - \mathcal{E}^{(0)}| = 0 \,,$$

where $\mathcal{E}^{(0)}$ is the GLM generalization error associated with $\frac{1}{n} I_n^{(0)}$.

Interpolation

The interpolation has to keep all the ingredients together:

$$S_{t\mu} := \sqrt{1-t}\left[\mathbf{a}^{*\intercal}\varphi\left(\mathbf{W}^{*(L)}\mathbf{X}_\mu^{(L-1)}\right)\right] + \sqrt{t}\left[\rho_L \mathbf{v}^{*\intercal}\mathbf{X}_\mu^{(L-1)} + \sqrt{\epsilon_L}\zeta_\mu^{*(L)}\right],$$
$$s_{t\mu} := \sqrt{1-t}\left[\mathbf{a}^{\intercal}\varphi\left(\mathbf{W}^{(L)}\mathbf{x}_\mu^{(L-1)}\right)\right] + \sqrt{t}\left[\rho_L \mathbf{v}^{\intercal}\mathbf{x}_\mu^{(L-1)} + \sqrt{\epsilon_L}\zeta_\mu^{(L)}\right],$$

Interpolating dataset:

$$\mathcal{D}_t = \{(Y_{t\mu}, \mathbf{X}_\mu^{(0)})_{\mu=1}^n\}, \quad Y_{t\mu} \sim P_{\rm out}(\cdot \mid S_{t\mu})$$

Interpolating free entropy:

$$\bar{f}_{n,t} = \frac{1}{n}\mathbb{E}_{(t)} \log \mathcal{Z}_t = \frac{1}{n}\mathbb{E}_{(t)} \log \int dP(\boldsymbol{\theta}^{(L)})\mathbb{E}_{\mathbf{v}} \prod_{\mu=1}^{n} \mathbb{E}_{\zeta_\mu^{(L)}} P_{out}\left(Y_{t\mu} \mid s_{t\mu}\right).$$

Useful concentrations

Moment control

$$\mathbb{E}\left| \|\mathbf{X}_\mu^{(\ell)}\|^2 - \sigma_\ell \right|^k \leq \frac{C_{k,\varphi}}{d_{min}^{k/2}} .$$

Quasi-orthogonality propagation

$$\mathbb{E}\left| \mathbf{X}_\mu^{(\ell)} \cdot \mathbf{X}_\nu^{(\ell)} \right|^k \leq \frac{C_k}{d_{min}^{k/2}} .$$

Free entropy concentration

There exists a non-negative constant $C(f, \varphi)$ such that

$$\mathbb{V}\left(\frac{1}{n} \log \mathcal{Z}_t(\mathcal{D}_t) \right) \leq C(f, \varphi) \left(\frac{1}{n} + \frac{1}{d_{min}} \right) .$$

Conclusions

- If $(1+\sqrt{\frac{n}{d_{\min}}}+\frac{n}{d_{\min}})\frac{1}{\sqrt{d_{\min}}} \to 0$ $(n \sim d_L \sim \ldots \sim d_0)$, training deep NN will give the same generalization error as if trained GLM.

- In order to escape this equivalence, deep neural networks must be analysed beyond the proportional regime (the number of samples n must grow much faster than d_ℓ, $n \sim \max\{d_\ell\}^2$).

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, 2018;
- **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. Pastore**, **F. Ginelli**, **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;
- **F. Camilli**, **D. Tieplova**, and **J. Barbier**, *"Fundamental limits of overparametrized shallow neural networks for supervised learning"*, arXiv:2307.05635, 2023.