Fonds National de la Recherche Luxembourg

UNIVERSITÉ DU LUXEMBOURG

CLTs in Deep Neural Networks

Quantitative Bounds through Coupling, Stein's Method and Entropy

Giovanni Peccati (Luxembourg University)

Joint works with: S. Favaro (Turin), B. Hanin (Princeton), D. Marinucci (Rome) and I. Nourdin (Luxembourg)

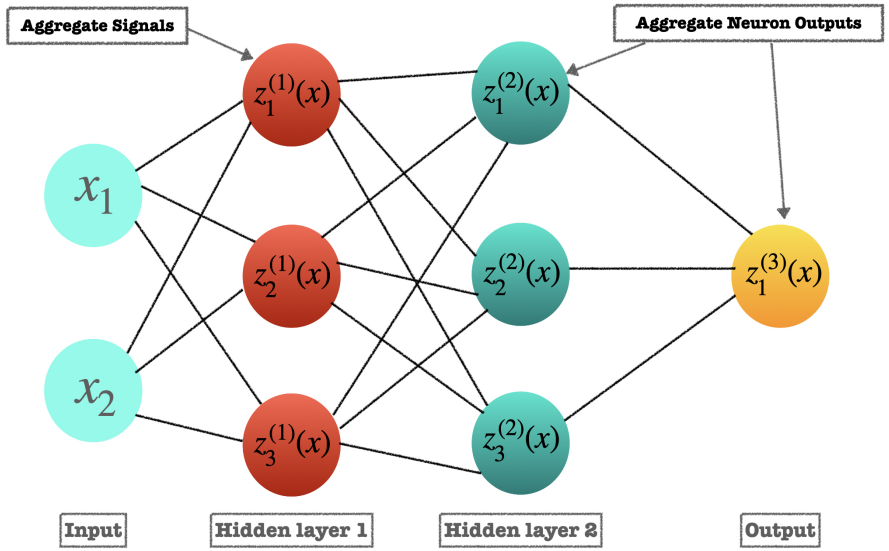
&

L Celli (Luxembourg)

Al-Khwarizmi Webinar

April 28th, 2026

Overview



A fully connected, feed-forward **deep neural network** ("DNN", see *Rosenblatt* (1958)) is a parametrized (*biologically inspired*) mapping

$$z : \mathbb{R}^N \to \mathbb{R}^M.$$

Crucial features: the **depth** L (number of hidden layers, $L = 2$ *in the picture*) and **width** n (maximal hidden layer size, $n = 3$ *in the picture*).

Our aim: analysis of **random DNNs** in the **infinite width regime**.

Formal Definition

We fix a **depth** $L \geq 1$ and consider a **fully connected neural network**

$$z^{(L+1)} : \mathbb{R}^{n_0} \to \mathbb{R}^{n_{L+1}} : x \mapsto (z_1^{(L+1)}(x), ..., z_{n_{L+1}}^{(L+1)}(x)) = z^{(L+1)}(x)$$

recursively defined as:

$$z^{(\ell)}(x) = z^{(\ell)}(x; \vartheta_0) = \begin{cases} W^{(1)}x + b^{(1)}, & \ell = 1, \\ W^{(2)}\sigma(z^{(1)}(x)) + b^{(2)}, & \ell = 2, \\ W^{(\ell)}\sigma(z^{(\ell-1)}(x)) + b^{(\ell)}, & \ell = 3, \dots, L+1. \end{cases}$$

Here:

- $\star$ $\sigma : \mathbb{R} \to \mathbb{R}$ is an (entrywise) **activation function**, e.g.

$$\sigma(x) = \max(x, 0) = \textbf{ReLU} \quad ; \quad \sigma(x) = \tanh(x).$$

- $\star$ The **weights** $\{W^{(\ell)}\}_{\ell=1,...,L+1}$ and **biases** $\{b^{(\ell)}\}_{\ell=1,...,L+1}$, are such that, for $n_1, ..., n_L \geq 1$,

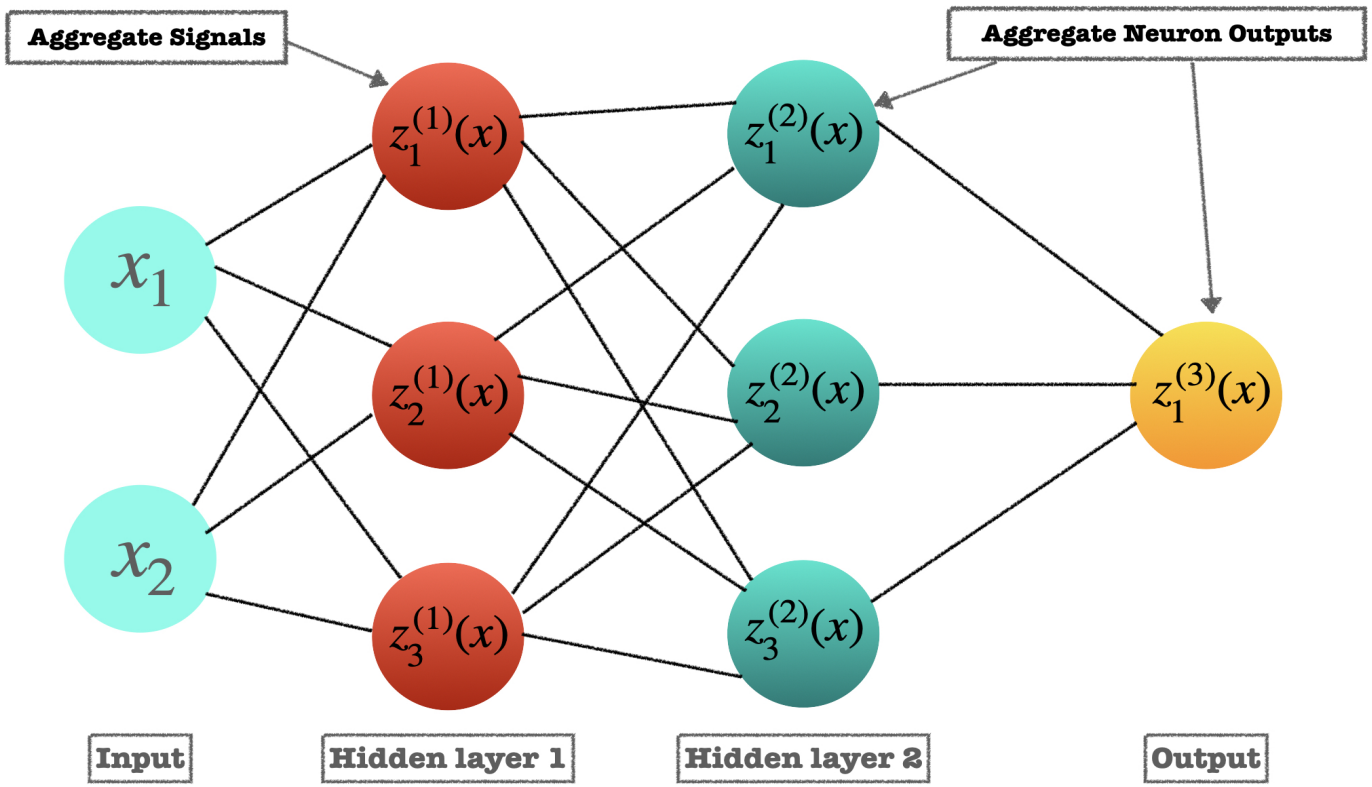
$$W^{(\ell)} \in \mathbb{R}^{n_\ell \times n_{\ell-1}}, \quad b^{(\ell)} \in \mathbb{R}^{n_\ell}.$$

[**Trainable parameters** at initialization, noted $\{W^{(\ell)}, b^{(\ell)}\} := \vartheta_0$]

Compact Description

In short, noting $A^{(\ell)}(y) := b^{(\ell)} + W^{(\ell)}y$, $\ell = 1, ..., L+1$,

$$z^{(L+1)}(x) = A^{(L+1)}\sigma(A^{(L)}\sigma(\cdots A^{(3)}\sigma(A^{(2)}\sigma(A^{(1)}x))).$$



In Practice

Typical use: approximate an unknown function $f : \mathbb{R}^{n_0} \to \mathbb{R}^{n_{L+1}}$, starting from a training dataset

$$\{(x_j, f(x_j)) : j = 1, \ldots, k\}.$$

Goal: find weights and biases so that

$$z^{(L+1)}(x) \approx f(x),$$

for x in the dataset *and* for inputs not in the training data.

This optimization is typically done in two steps:

(1) **Randomly initialize** the network weights and biases, as above;

(2) Optimize the weights and biases by some variant of **gradient descent** on an empirical loss such as the squared error:

$$\vartheta \mapsto \sum_{j=1}^{k} \|z^{(L+1)}(x_j; \vartheta) - f(x_j)\|_2^2.$$

Our Setting (Randomness & Smoothness)

Assumptions

1. **Weight** *and* **biases** *(from distinct layers and within the same layer) are independent Gaussian random variables with distribution*

$$W_{ij}^{(\ell)} \sim \mathcal{N}\left(0, \frac{C_W}{n_{\ell-1}}\right), \quad b^{(\ell)} \sim \mathcal{N}(0, C_b),$$

   for fixed $C_W > 0$, $C_b \geq 0$.

2. *For some* $r \geq 0$, *the* **activation** $\sigma : \mathbb{R} \to \mathbb{R}$ *is of class* $C^r(\mathbb{R})$, *and the* $(r+1)$*th derivative* $D^{r+1}\sigma$ *exists almost everywhere and is polynomially bounded.*

Our aim is to study the fluctuations of $z^{(L+1)}$, whenever

$$n_1, ..., n_L \asymp n,$$

for some large integer n.

Neural Tangent Kernels

⋆ (*Vanishing second-order structure*) The **NTK theory** (*Jacot & al.* (2018), *Du & al.* (2019), ...) implies that, for $n \gg 1$, the optimization trajectory of $\vartheta \mapsto z^{(L+1)}(x;\vartheta)$ can be tuned to coincide (with overwhelming probability) with that of its **linearization**

$$\vartheta \mapsto z^{\mathbf{lin}}(x;\vartheta) := z^{(L+1)}(x;\vartheta_0) + \langle \nabla_\vartheta z^{(L+1)}(x;\vartheta_0), \vartheta - \vartheta_0 \rangle,$$

where ϑ_0 = **parameters at intializiation**.

⋆ (*Deterministic first-order structure*) In particular, in this regime both the training dynamics and the out-of-sample performance of the network are determined by the **deterministic NTK**

$$\Theta^{(L+1)}(x,y) := \mathbb{P} - \lim_{n\to\infty} \langle \nabla_\vartheta z^{(L+1)}(x;\vartheta_0), \nabla_\vartheta z^{(L+1)}(y;\vartheta_0) \rangle.$$

Functional CLT

Theorem (Neal (1996), Matthews et al. (2018), Hanin (2023), ...)

As $n_1, \dots n_L \to \infty$, *the random field* $z^{(L+1)} \in \mathbb{R}^{n_{L+1}}$ *converges weakly to a* **Gaussian process** *with* n_{L+1} *iid components* $\{G_i^{(L+1)} : i = 1, \dots, n_{L+1}\}$ *with a* **limit covariance** $K^{(L+1)}$ *recursively defined by:*

$$K^{(\ell+1)}(x,y) = \begin{cases} C_b + C_W \mathbb{E}[\sigma(G^{(\ell)}(x)) \times \sigma(G^{(\ell)}(y))], & \ell = 1, ..., L+1, \\ C_b + \frac{C_W}{n_0} x \cdot y, & \ell = 0. \end{cases},$$

where $G^{(\ell)}$ *has covariance* $K^{(\ell)}$.

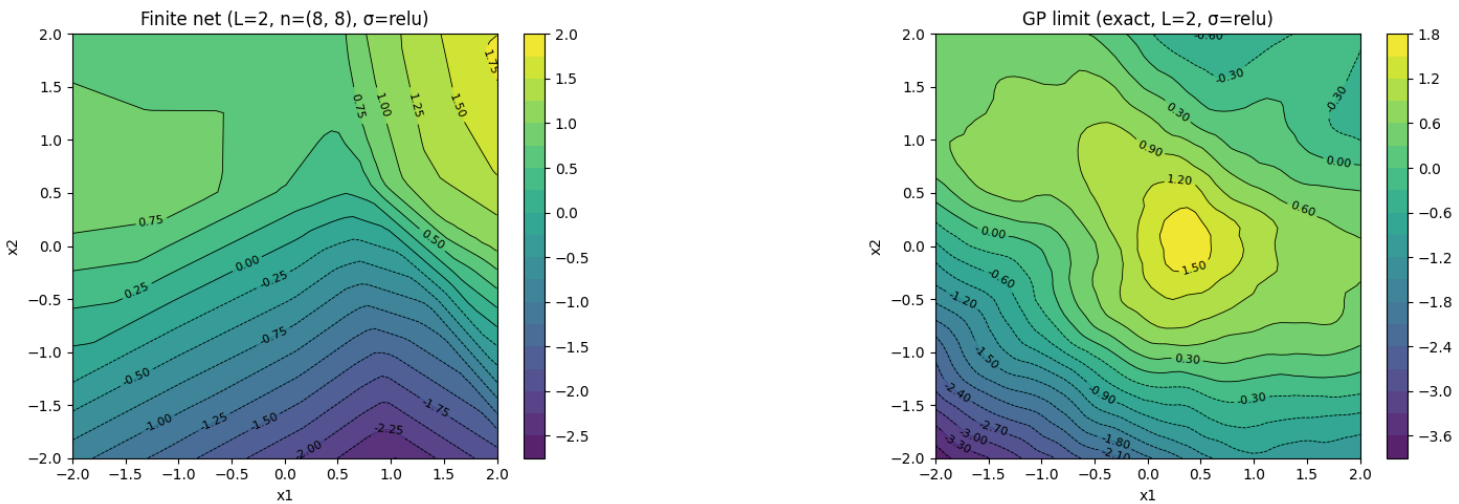


Figure 1: A **ReLU network** (for $L = 2$, & $n = 8, 64, 256, 1024, 4096$) and its **uncoupled Gaussian limit**.

Functional CLT

Theorem (Neal (1996), Matthews et al. (2018), Hanin (2023), ...)

As $n_1, \dots n_L \to \infty$, *the random field* $z^{(L+1)} \in \mathbb{R}^{n_{L+1}}$ *converges weakly to a* **Gaussian process** *with* n_{L+1} *iid components* $\{G_i^{(L+1)} : i = 1, \dots, n_{L+1}\}$ *with a* **limit covariance** $K^{(L+1)}$ *recursively defined by:*

$$K^{(\ell+1)}(x,y) = \begin{cases} C_b + C_W \mathbb{E}[\sigma(G^{(\ell)}(x)) \times \sigma(G^{(\ell)}(y))], & \ell = 1, ..., L+1, \\ C_b + \frac{C_W}{n_0} x \cdot y, & \ell = 0. \end{cases},$$

where $G^{(\ell)}$ *has covariance* $K^{(\ell)}$.

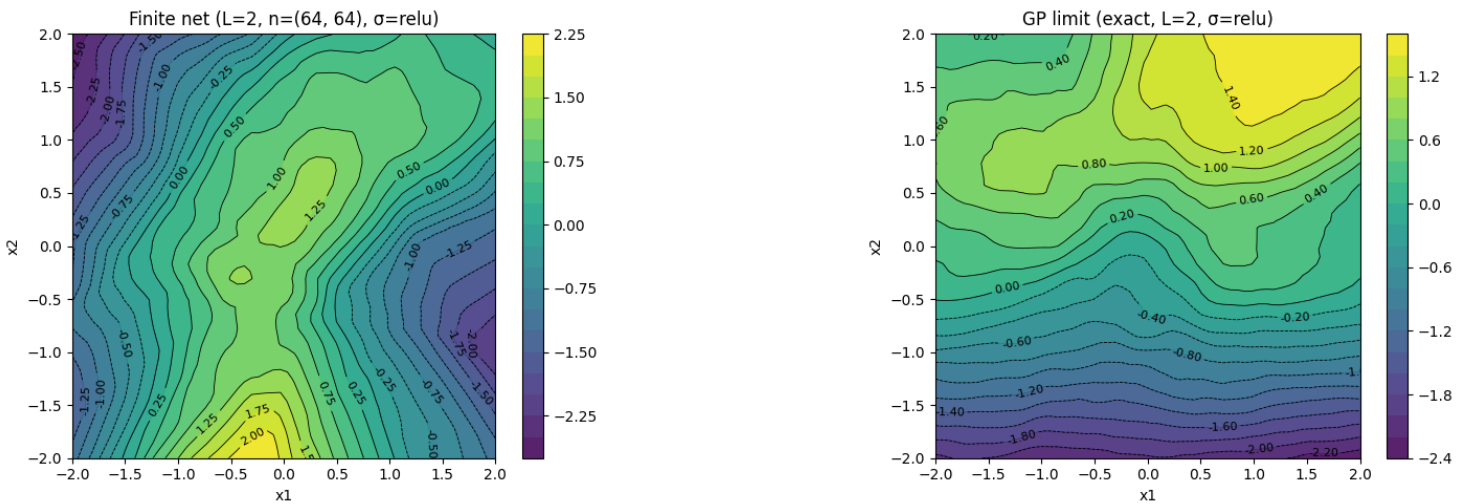


Figure 1: A **ReLU network** (for $L = 2$, & $n = 8, 64, 256, 1024, 4096$) and its **uncoupled Gaussian limit**.

Functional CLT

Theorem (Neal (1996), Matthews et al. (2018), Hanin (2023), ...)

As $n_1, \dots n_L \to \infty$, *the random field* $z^{(L+1)} \in \mathbb{R}^{n_{L+1}}$ *converges weakly to a* **Gaussian process** *with* n_{L+1} *iid components* $\{G_i^{(L+1)} : i = 1, \dots, n_{L+1}\}$ *with a* **limit covariance** $K^{(L+1)}$ *recursively defined by:*

$$K^{(\ell+1)}(x,y) = \begin{cases} C_b + C_W \mathbb{E}[\sigma(G^{(\ell)}(x)) \times \sigma(G^{(\ell)}(y))], & \ell = 1, ..., L+1, \\ C_b + \frac{C_W}{n_0} x \cdot y, & \ell = 0. \end{cases},$$

where $G^{(\ell)}$ *has covariance* $K^{(\ell)}$.

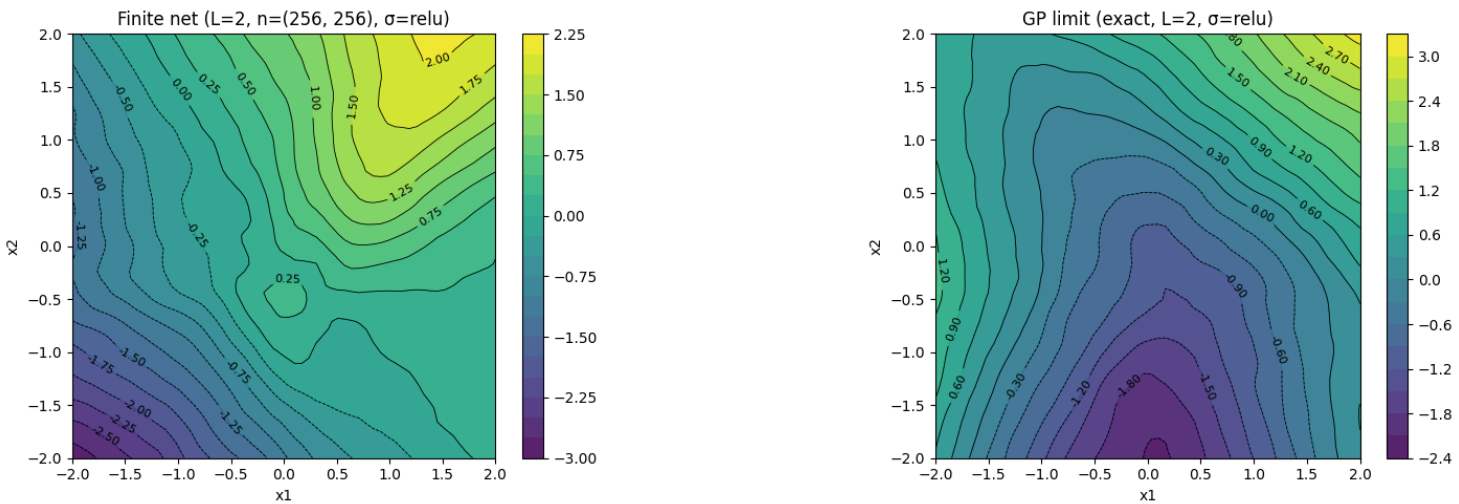


Figure 1: A **ReLU network** (for $L = 2$, & $n = 8, 64, 256, 1024, 4096$) and its **uncoupled Gaussian limit**.

Functional CLT

Theorem (Neal (1996), Matthews et al. (2018), Hanin (2023), ...)

As $n_1, \dots n_L \to \infty$, *the random field* $z^{(L+1)} \in \mathbb{R}^{n_{L+1}}$ *converges weakly to a* **Gaussian process** *with* n_{L+1} *iid components* $\{G_i^{(L+1)} : i = 1, \dots, n_{L+1}\}$ *with a* **limit covariance** $K^{(L+1)}$ *recursively defined by:*

$$K^{(\ell+1)}(x,y) = \begin{cases} C_b + C_W \mathbb{E}[\sigma(G^{(\ell)}(x)) \times \sigma(G^{(\ell)}(y))], & \ell = 1, ..., L+1, \\ C_b + \frac{C_W}{n_0} x \cdot y, & \ell = 0. \end{cases},$$

where $G^{(\ell)}$ *has covariance* $K^{(\ell)}$.

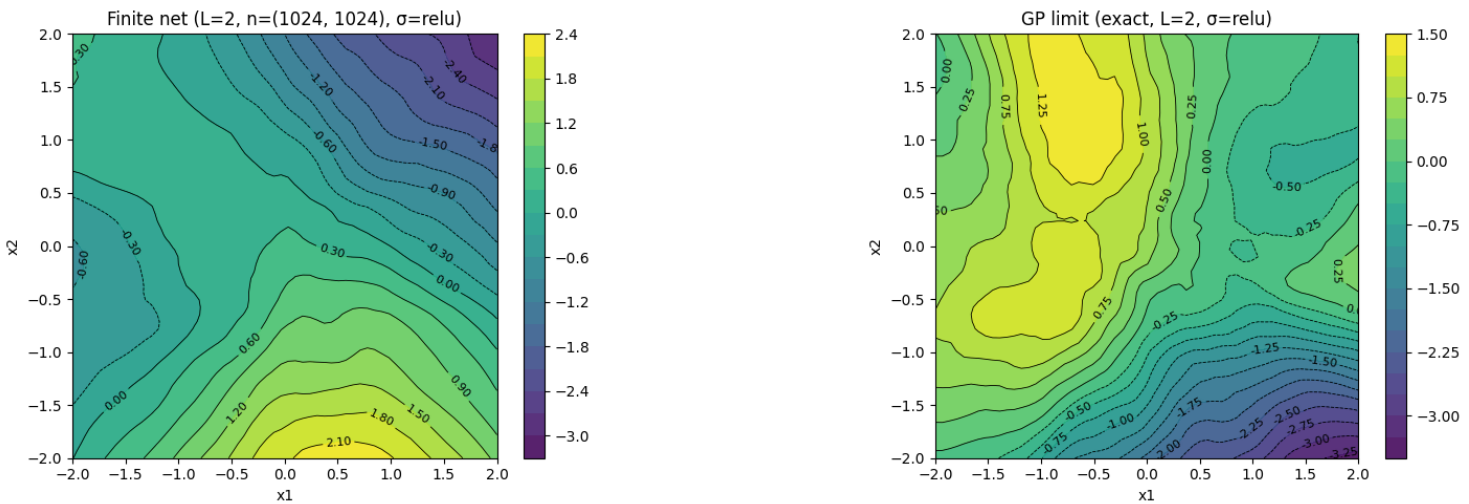


Figure 1: A **ReLU network** (for $L = 2$, & $n = 8, 64, 256, 1024, 4096$) and its **uncoupled Gaussian limit**.

Functional CLT

Theorem (Neal (1996), Matthews et al. (2018), Hanin (2023), ...)

As $n_1, \dots n_L \to \infty$, *the random field* $z^{(L+1)} \in \mathbb{R}^{n_{L+1}}$ *converges weakly to a* **Gaussian process** *with* n_{L+1} *iid components* $\{G_i^{(L+1)} : i = 1, \dots, n_{L+1}\}$ *with a* **limit covariance** $K^{(L+1)}$ *recursively defined by:*

$$K^{(\ell+1)}(x,y) = \begin{cases} C_b + C_W \mathbb{E}[\sigma(G^{(\ell)}(x)) \times \sigma(G^{(\ell)}(y))], & \ell = 1, ..., L+1, \\ C_b + \frac{C_W}{n_0} x \cdot y, & \ell = 0. \end{cases},$$

where $G^{(\ell)}$ *has covariance* $K^{(\ell)}$.

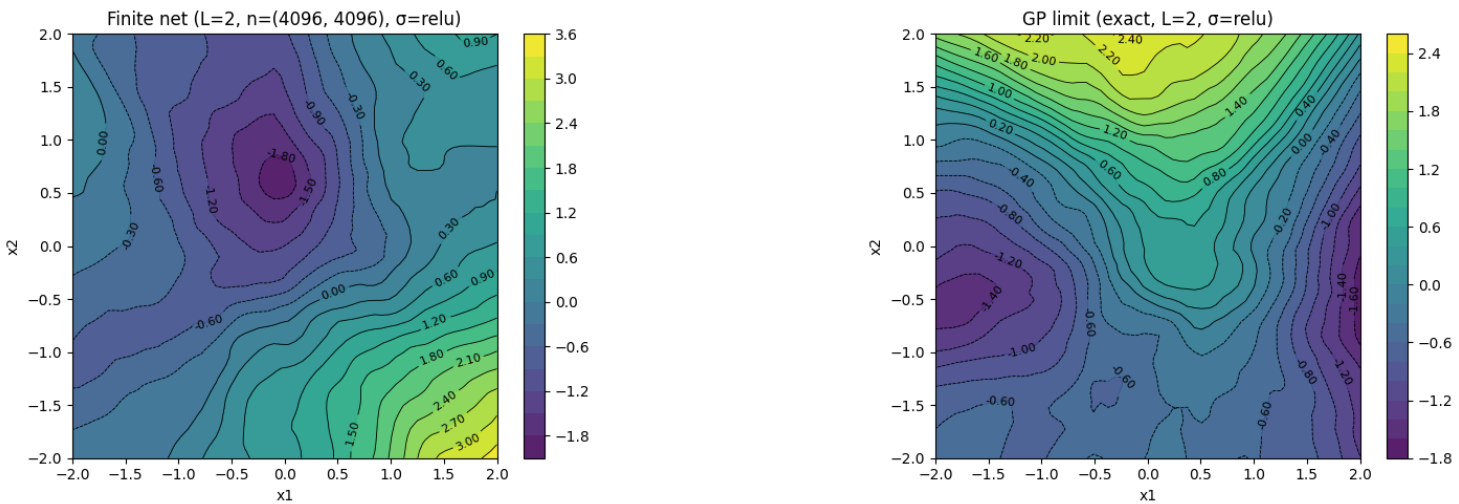


Figure 1: A **ReLU network** (for $L = 2$, & $n = 8, 64, 256, 1024, 4096$) and its **uncoupled Gaussian limit**.

SALIENT PROPERTIES OF DNNS

Remark: From now on, we will assume $n_{L+1} = 1$ (one-dimensional output).

Elementary features:

- ⋆ Conditionally to $\{z_j^{(L)}\}$ the field $z^{(L+1)}$ is **centered and Gaussian**, with covariance

$$\Sigma^{(L)}(x,y) := C_b + \frac{C_W}{n_L}\sum_{j=1}^{n_L}\sigma(z_j^{(L)}(x))\sigma(z_j^{(L)}(y)).$$

- ⋆ As a consequence, given a vector $Z := (z^{(L+1)}(x_1), ..., z^{(L+1)}(x_m))$, the corresponding random covariance $\mathbb{S} := \{\Sigma^{(L)}(x_i, x_j) : i, j = 1, ..., m\}$ acts as a **Stein's kernel** for Z: for every smooth mapping φ,

$$\mathbb{E}[Z \cdot \nabla\varphi(Z)] = \mathbb{E}[\langle \mathbb{S}, \mathrm{Hess}(\varphi)(Z)\rangle_{H.S.}].$$

- ⋆ The previous **Central Limit Theorem** emerges from the fact that

$$\Sigma^{(L)} \longrightarrow K^{(L+1)} \text{ (deterministic)}, \quad \text{as } n \to \infty.$$

- ⋆ For every open ball $\mathbb{U} \subset \mathbb{R}^{n_0}$, almost surely, $z^{(L+1)} \in \mathbb{W}^{2;r}(\mathbb{U})$ = **Sobolev space** of r times weakly differentiable mappings.

Quantities of Interest

⋆ Given random vectors $\mathbf{X} = (X_1, ..., X_m)$ and $\mathbf{Y} = (Y_1, ..., Y_m)$ the **total variation distance** between the laws of $\mathbf{X}$ and $\mathbf{Y}$ is

$$\mathbf{TV}(\mathbf{X}, \mathbf{Y}) := \sup_{C \text{ Borel}} \left| \mathbb{P}(X \in C) - \mathbb{P}(\mathbf{Y} \in C) \right|.$$

⋆ For $q \geq r$, the **trivial coupling** (*not* optimal!) of $z^{(L+1)}$ and $G^{(L+1)}$ in $\mathbb{W}^{2;q}(\mathbb{U})$ is given by

$$\left(\widetilde{z}^{(L+1)}, \widetilde{G}^{(L+1)} \right) := \left(\sqrt{\Sigma^{(L)}} \star \zeta, \sqrt{K^{(L+1)}} \star \zeta \right),$$

where ζ is a common **white noise** on $\mathbb{W}^{2;q}(\mathbb{U})$, **independent** of $\Sigma^{(L)}$.

⋆ One has

$$\mathbb{E}\left[\| \widetilde{z}^{(L+1)} - \widetilde{G}^{(L+1)} \|^2_{\mathbb{W}^{2;q}(\mathbb{U})} \right] = \mathbb{E}\left[\| \sqrt{\Sigma^{(L)}} - \sqrt{K^{(L+1)}} \|^2_{HS} \right].$$

Finite Dimension

Theorem A (Celli & Peccati, 2025)

Under a certain non-degeneracy assumption on the limiting covariance, for fixed inputs $\mathcal{X} := \{x_1, ..., x_m\} \subset \mathbb{R}^{n_0}$, *and multiindices* $\mathbb{J} := \{J_1, ..., J_m\} \subseteq [r]$, *writing*

$$D^{\mathbb{J}} f(\mathcal{X}) := (D^{J_1} f(x_1), ..., D^{J_m} f(x_m)),$$

we have that,

$$\mathbf{TV}\left(D^{\mathbb{J}} z^{(L+1)}(\mathcal{X}) ; D^{\mathbb{J}} G^{(L+1)}(\mathcal{X})\right) \leq \frac{C}{n}.$$

The bound is **optimal** *(matching lower bound).*

Remark: (i) The result extends to the case of **degenerate covariance structures**.
(ii) *One-dimensional case*: *Favaro et al.* (2025), via **Stein's method**.
(iii) *Convex distance*: rates $\asymp n^{-1/2}$ in *Favaro et al.* (2025), via **Stein's method**.
(iv) *Transport distances*: rates $\asymp n^{-1}$ in *Basteri & Trevisan* (2024)
(v) *Non-Gaussian weights & Edgeworth expansions* : *Celli* (2025, 2026)

Functional Bounds

Theorem B (Favaro, Hanin, Marinucci, Nourdin & Peccati, 2025)
Fix an open ball $\mathbb{U} \subset \mathbb{R}^{n_0}$, and $q \leq r$.

1. *Assume that the eigenvalues of the covariance of $G^{(L+1)}$ (as a Gaussian element in $\mathbb{W}^{2;q}$) are such that*
$$\sum_k (\lambda_k)^{1/2} < \infty.$$
Then, under a certain non-degeneracy assumption ,
$$\mathbb{E}\left[\|\widetilde{z}^{(L+1)} - \widetilde{G}^{(L+1)}\|^2_{\mathbb{W}^{2;q}(\mathbb{U})}\right]^{1/2} \leq \frac{C}{n^{1/8}}$$

2. *If σ is* **smooth**, *then, for every $q \geq 1$ there exists a coupling $(\bar{z}^{(L+1)}, \bar{G}^{(L+1)})$ such that*
$$\mathbb{E}\left[\|\bar{z}^{(L+1)} - \bar{G}^{(L+1)}\|^2_{\mathbb{C}^q(\mathbb{U})}\right]^{1/2} \leq \frac{C}{n^{1/8}}.$$

Remark: Rates are plausibly **not optimal**.

REMARKS

- ⋆ The condition $\sum_{k=1}^{\infty}(\lambda_k)^{1/2} < \infty$ is verified for all $q \geq 0$ for σ **smooth** (see *Dierickx, Nourdin, Peccati & Rossi* (2023)).

- ⋆ For **ReLU activations** restricted to the **sphere**, one has that $\sum_{k=1}^{\infty}(\lambda_k)^{p} < \infty$, when $q = 0$ and

$$p > 1/(2 + n_0)$$

(*Bietti & Bach* (2021)). This yields a bound of the order $n^{-\frac{1}{8}}$ in the space $\mathbb{W}^2(\mathbb{S}^{n_0-1}) = L^2(\mathbb{S}^{n_0-1})$.

- ⋆ See also *Di Lillo, Marinucci, Salvi & Vigogna* (2025); *Di Lillo, Maini & Marinucci* (2026).

Spherical ReLU (Simulations)

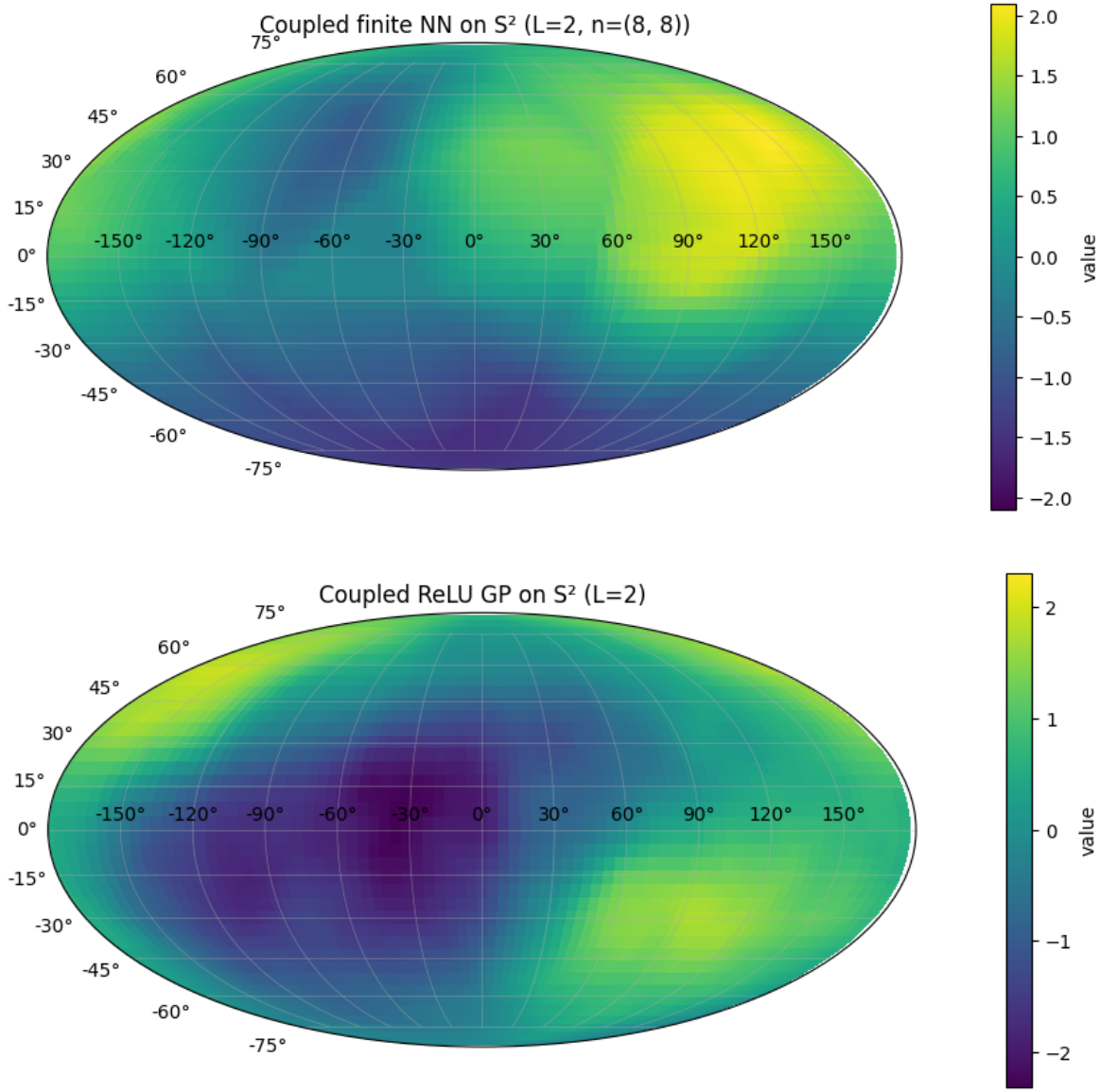


Figure 2: **Spherical ReLU** (for $L = 2$ and $n = 8 \longrightarrow 12000$) and its **coupled Gaussian limit**.

Spherical ReLU (Simulations)

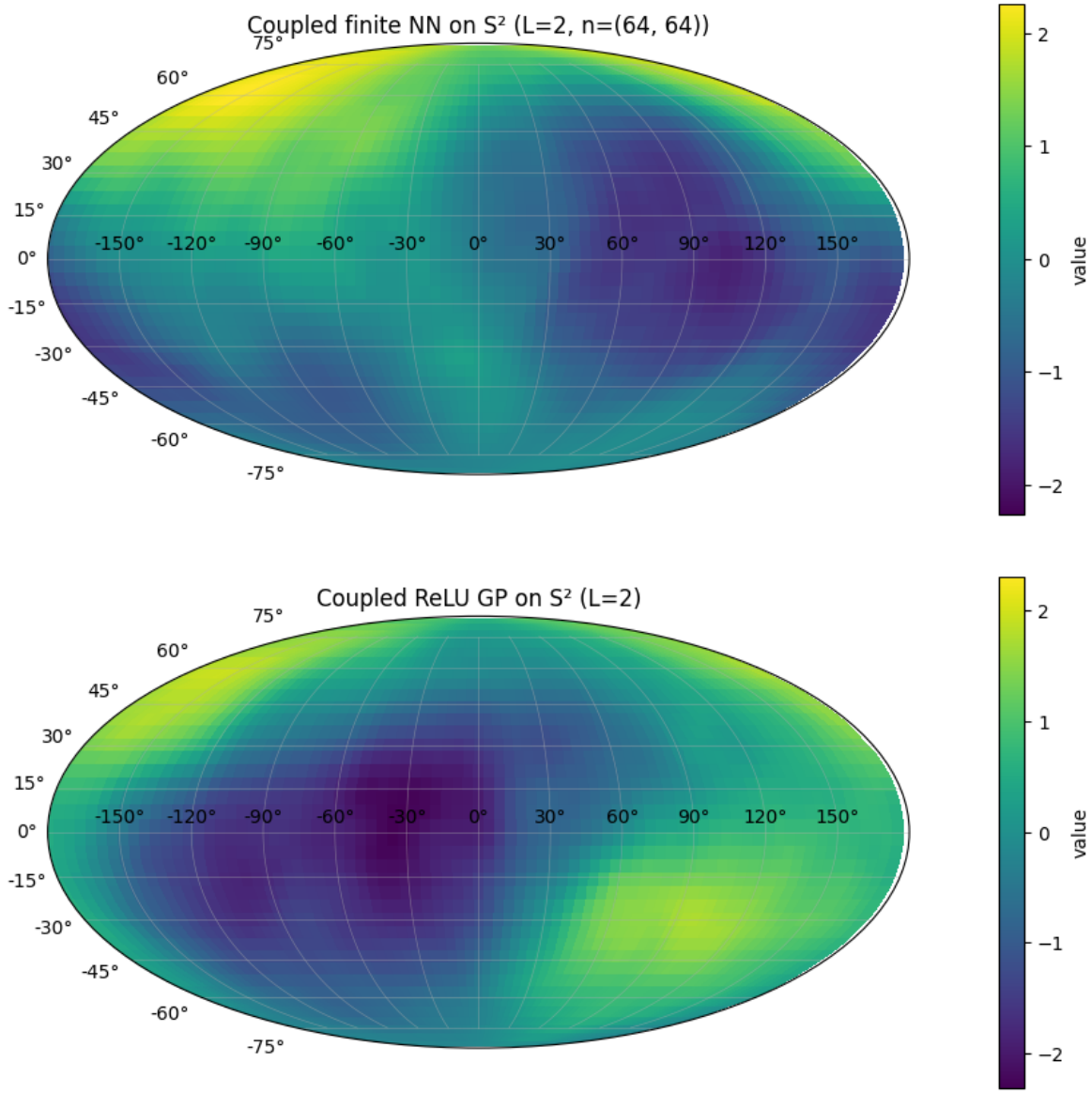


Figure 2: **Spherical ReLU** (for $L = 2$ and $n = 8 \longrightarrow 12000$) and its **coupled Gaussian limit**.

SPHERICAL RELU (SIMULATIONS)

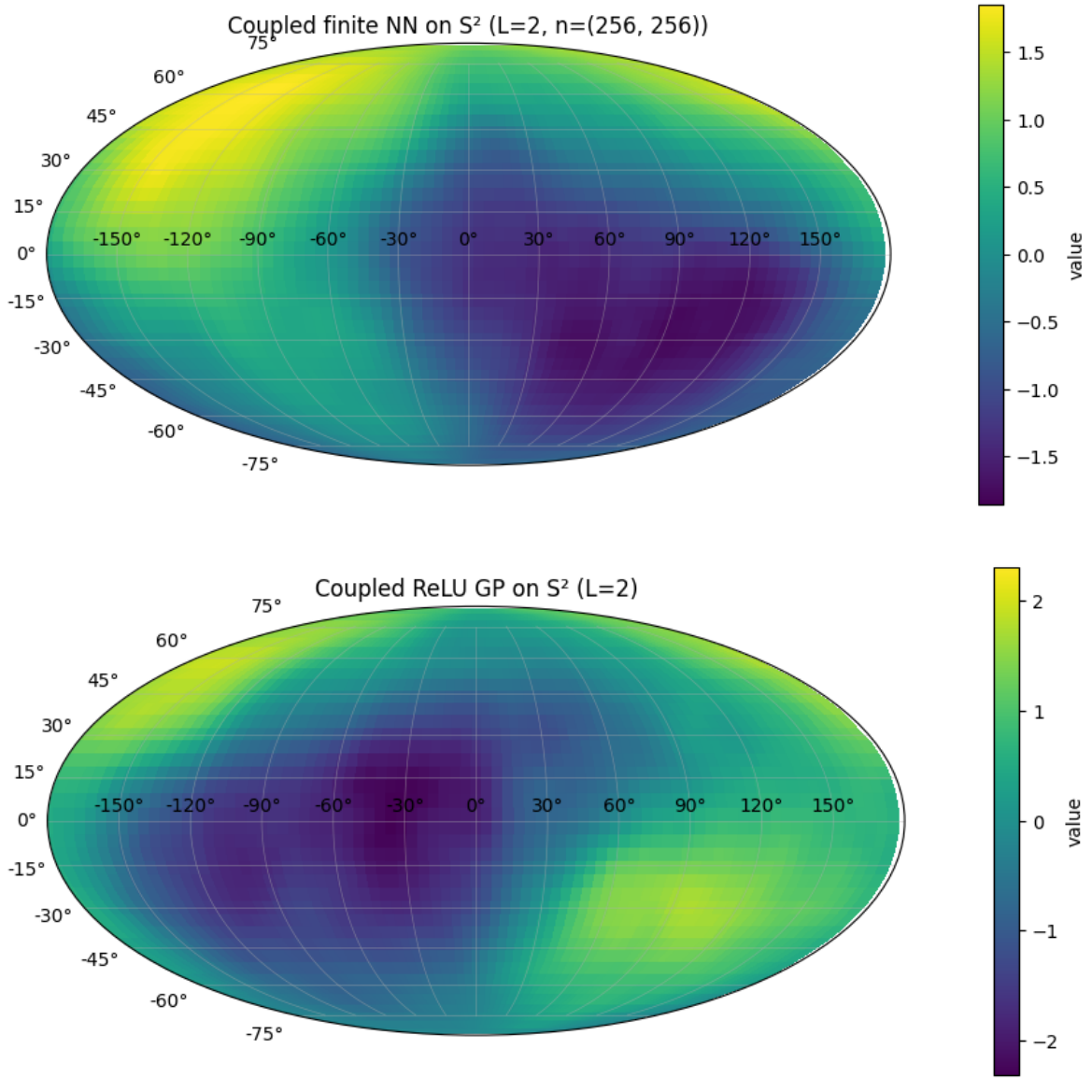


Figure 2: **Spherical ReLU** (for $L = 2$ and $n = 8 \longrightarrow 12000$) and its **coupled Gaussian limit**.

Spherical ReLU (Simulations)

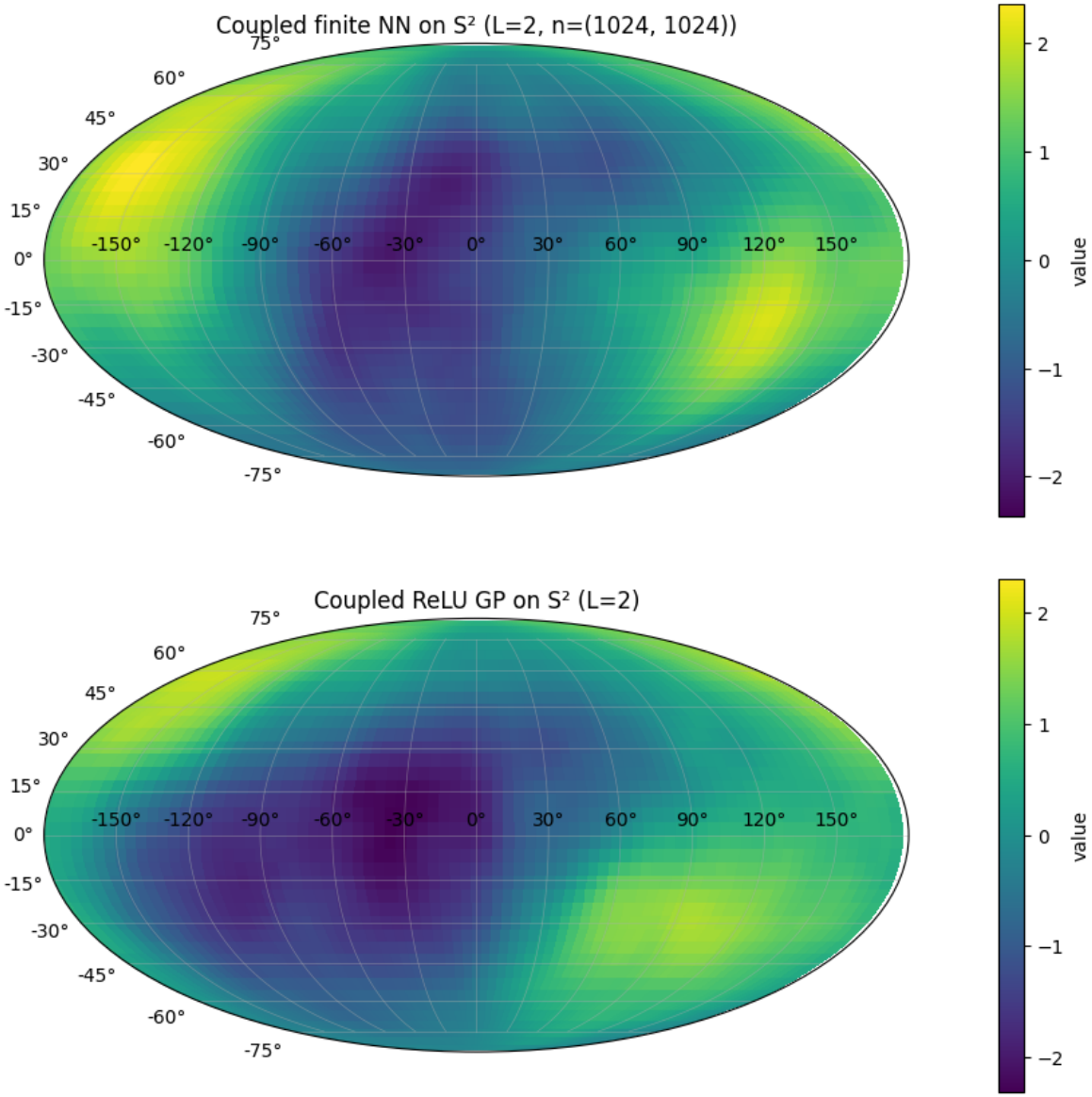


Figure 2: **Spherical ReLU** (for $L = 2$ and $n = 8 \longrightarrow 12000$) and its **coupled Gaussian limit**.

Spherical ReLU (Simulations)

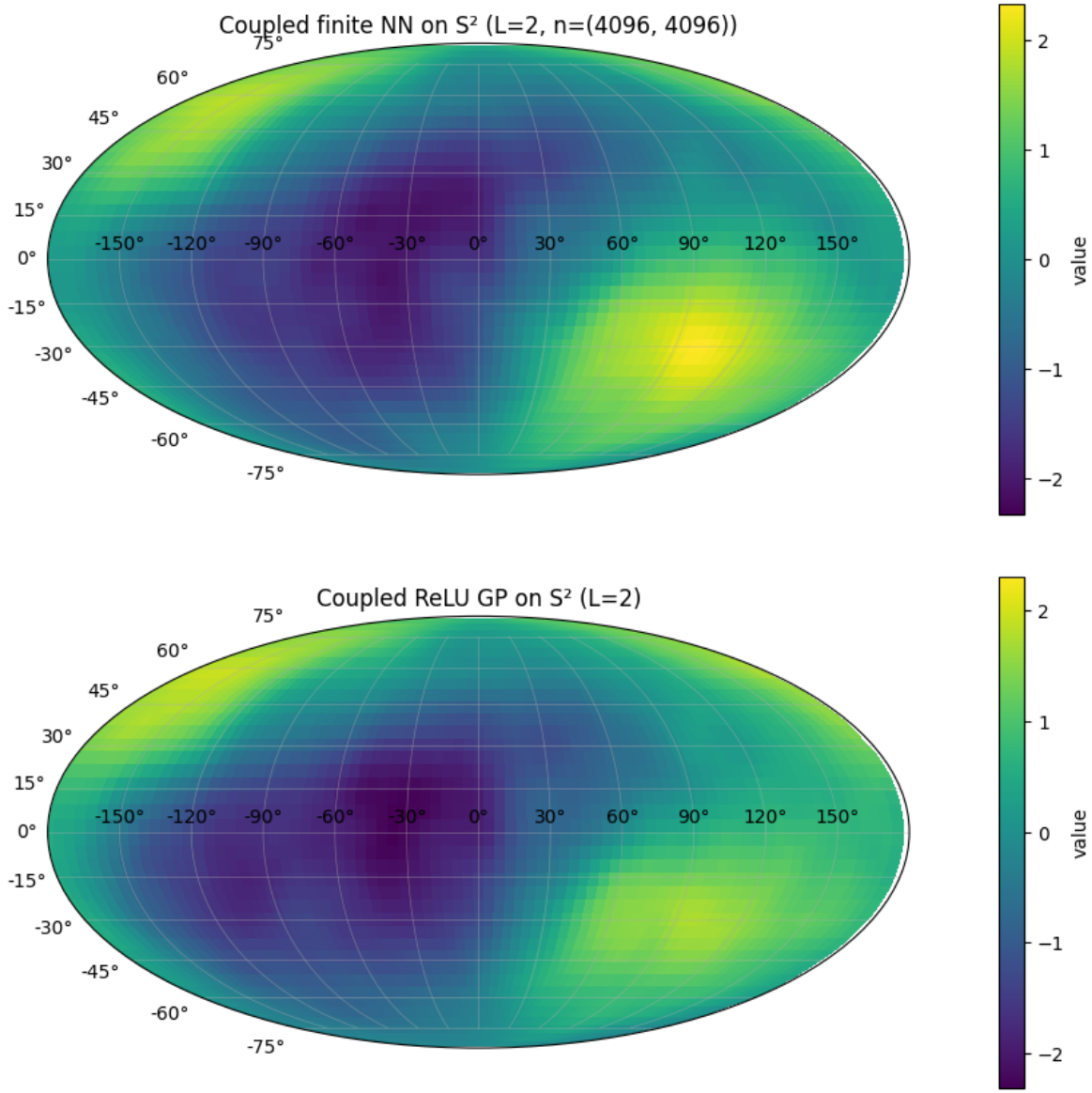


Figure 2: **Spherical ReLU** (for $L = 2$ and $n = 8 \longrightarrow 12000$) and its **coupled Gaussian limit**.

Spherical ReLU (Simulations)

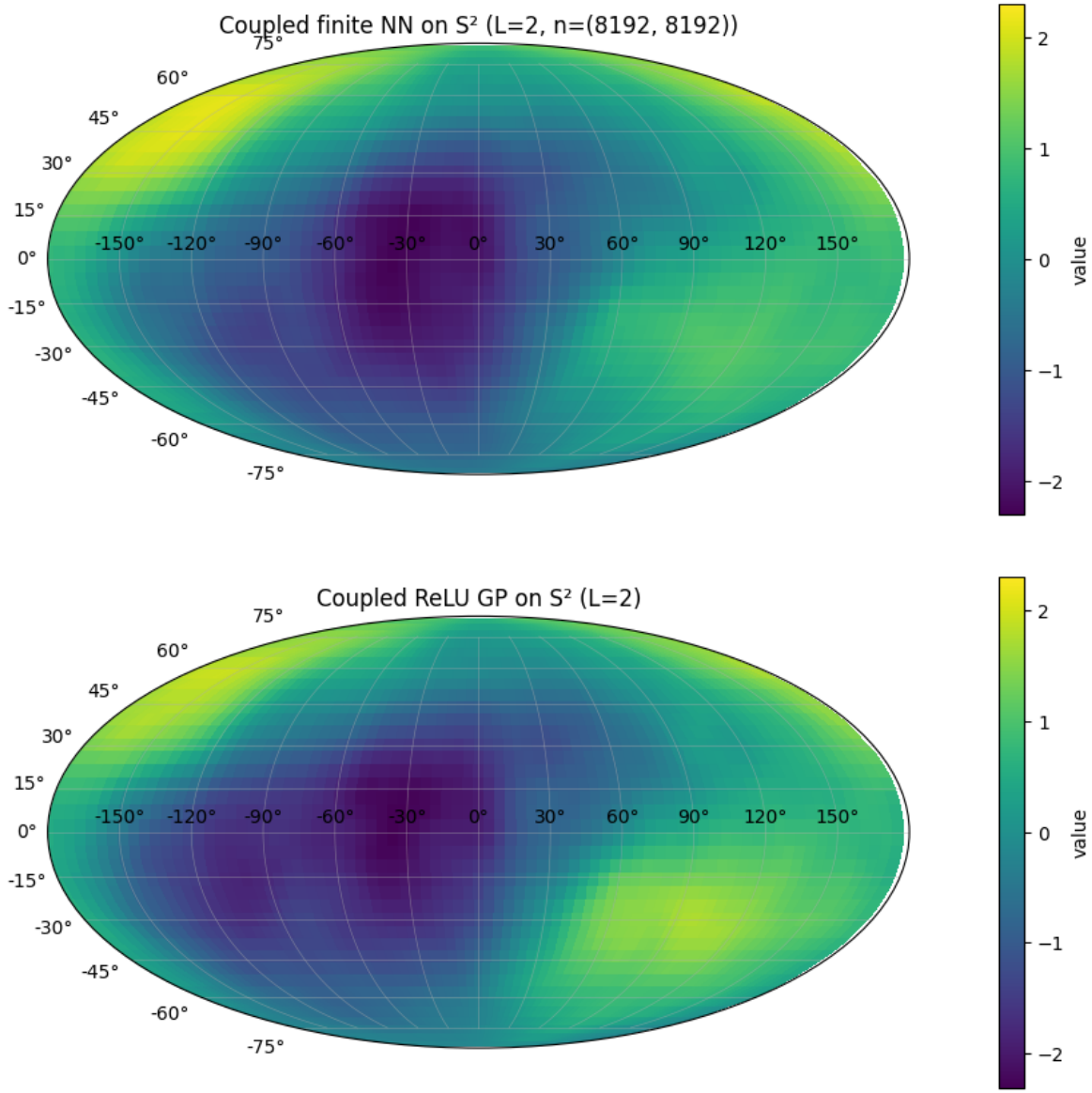


Figure 2: **Spherical ReLU** (for $L = 2$ and $n = 8 \longrightarrow 12000$) and its **coupled Gaussian limit**.

SPHERICAL RELU (SIMULATIONS)

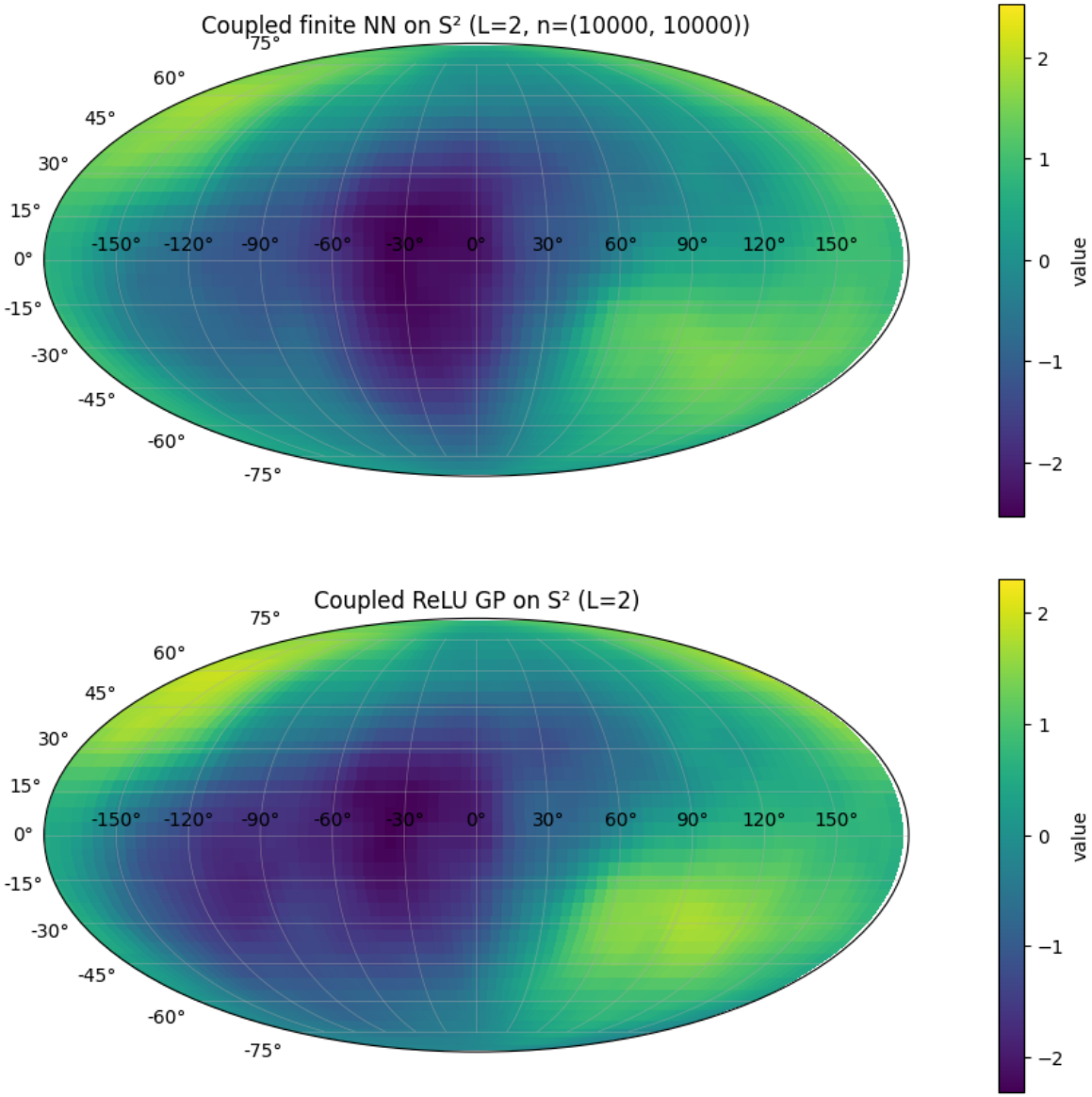


Figure 2: **Spherical ReLU** (for $L = 2$ and $n = 8 \longrightarrow 12000$) and its **coupled Gaussian limit**.

Spherical ReLU (Simulations)

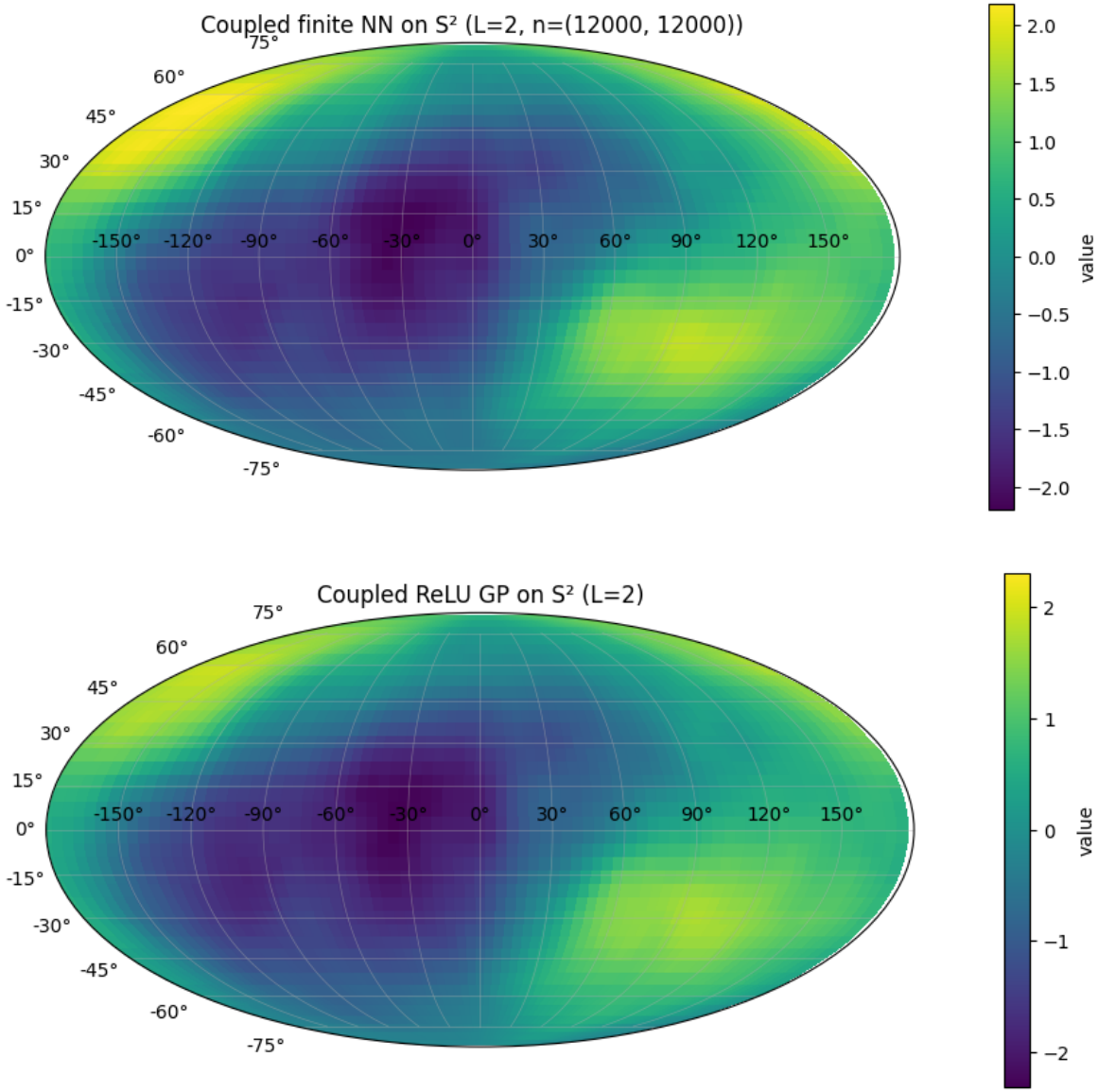


Figure 2: **Spherical ReLU** (for $L = 2$ and $n = 8 \longrightarrow 12000$) and its **coupled Gaussian limit**.

Spherical ReLU (Log-Log Plots)

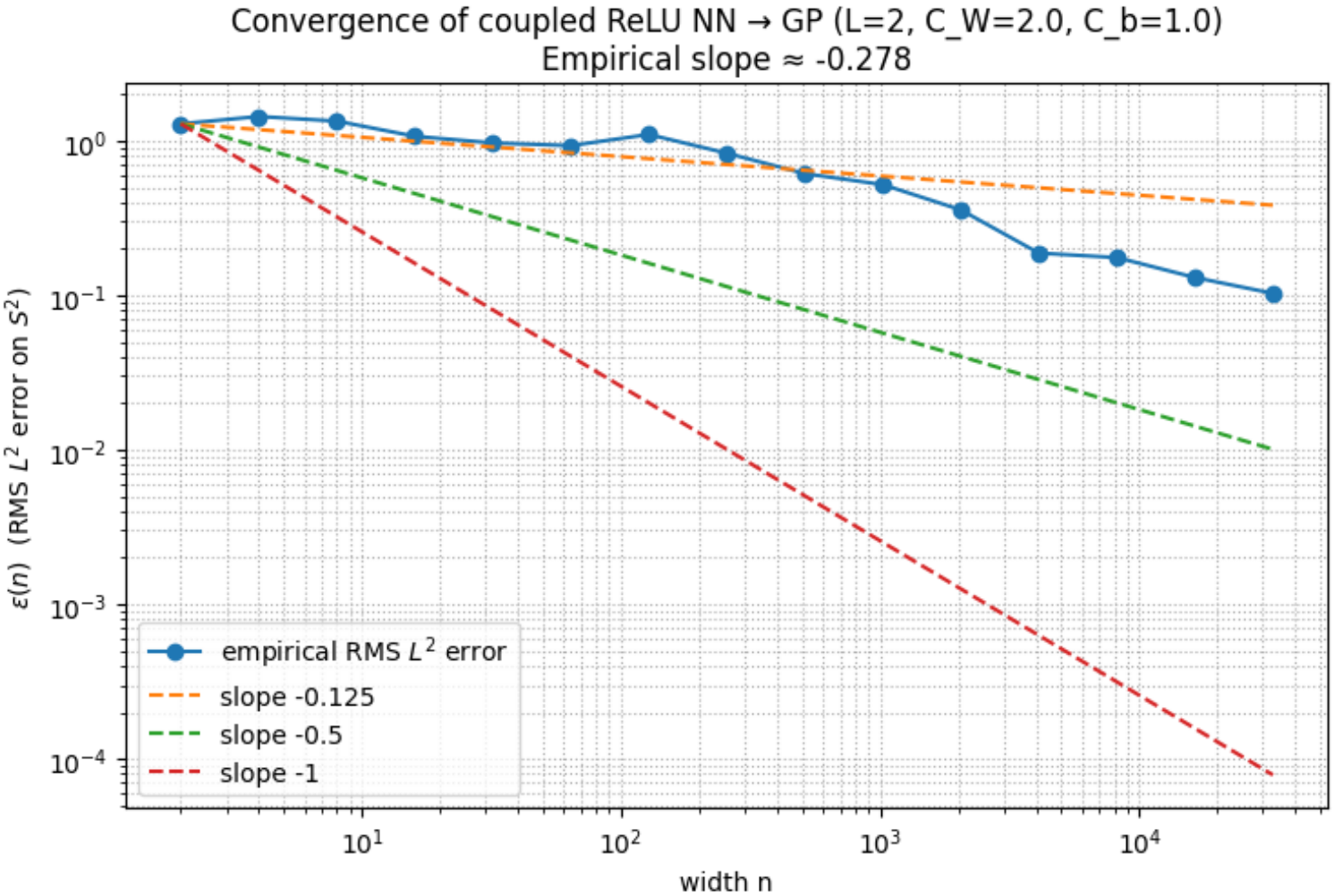


Figure 3: Estimated **ReLU coupling errors** on the sphere (NN vs. GP limit).

Spherical ReLU (Log-Log Plots)

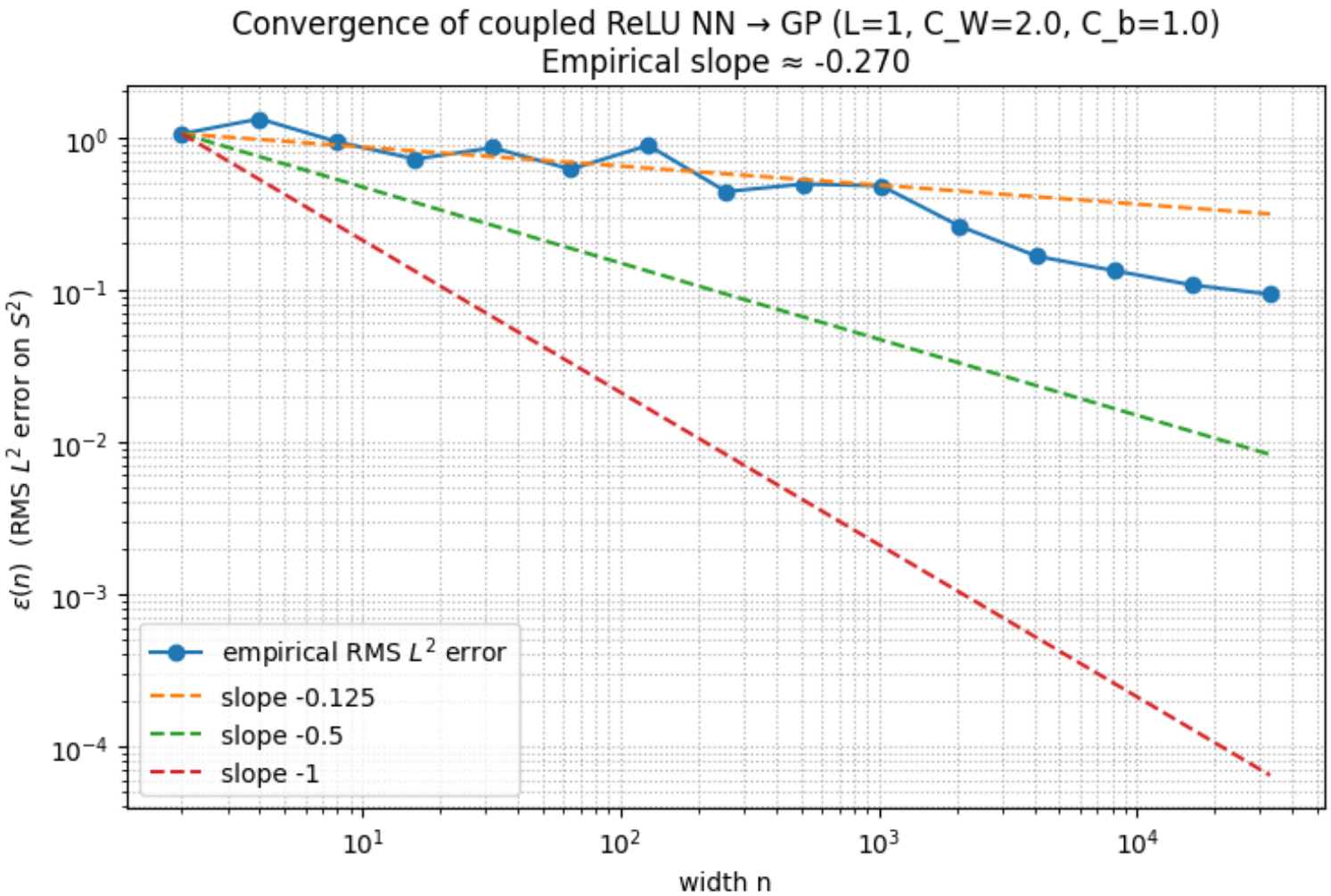
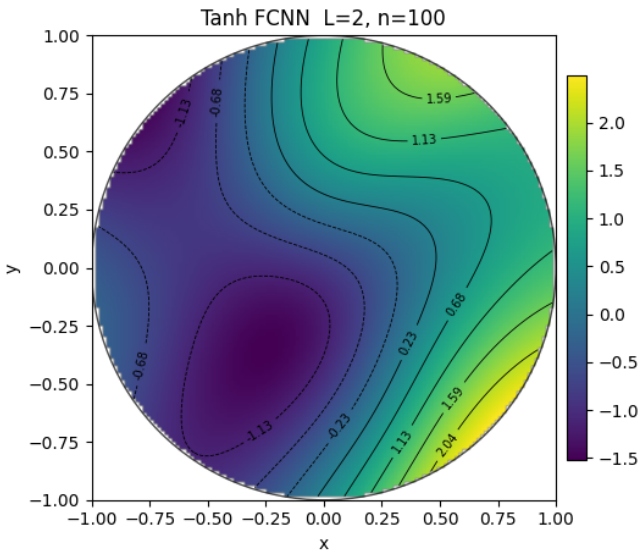
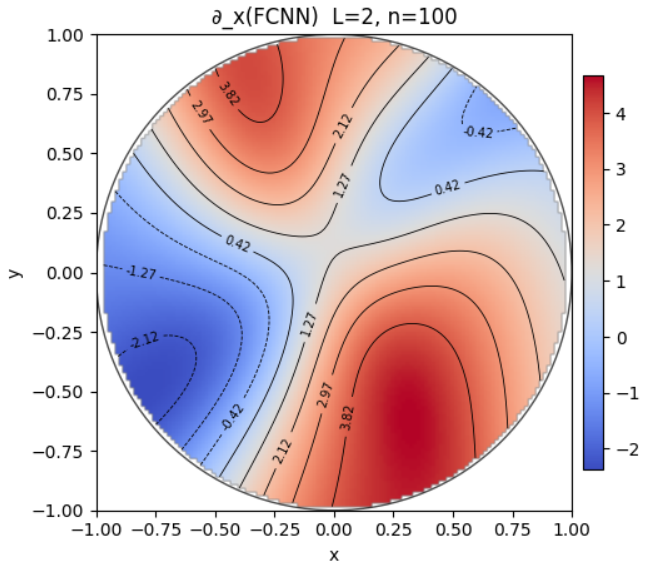
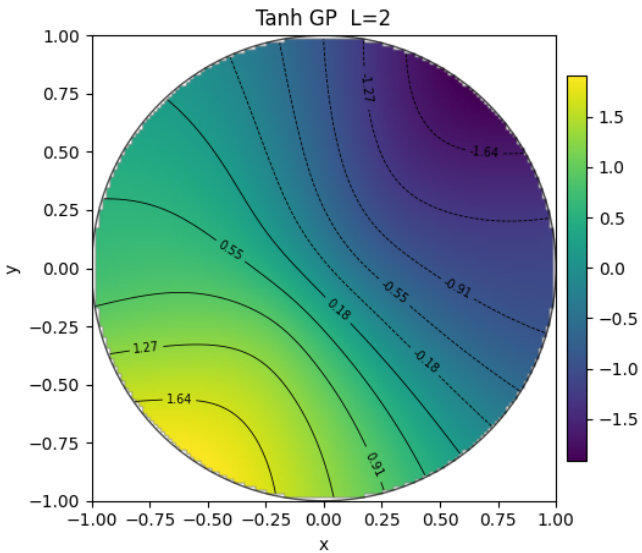


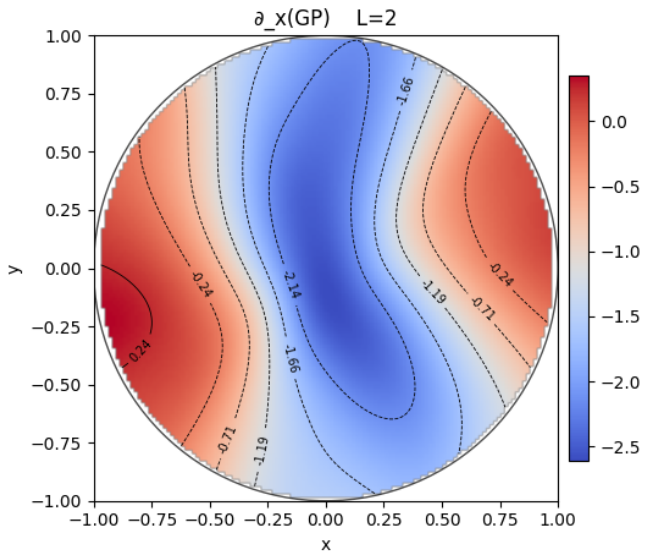
Figure 3: Estimated **ReLU coupling errors** on the sphere (NN vs. GP limit).

Tanh Networks on a Disk

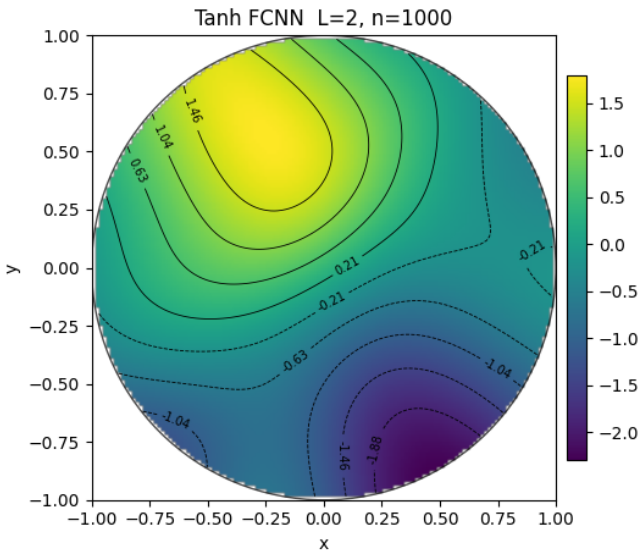


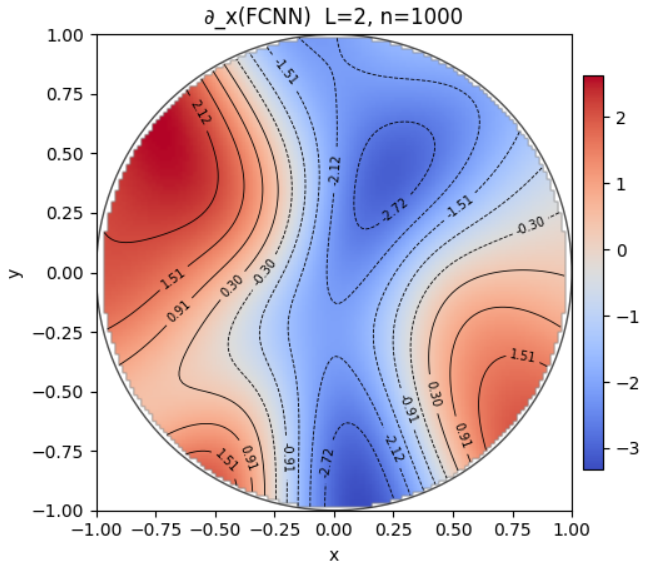


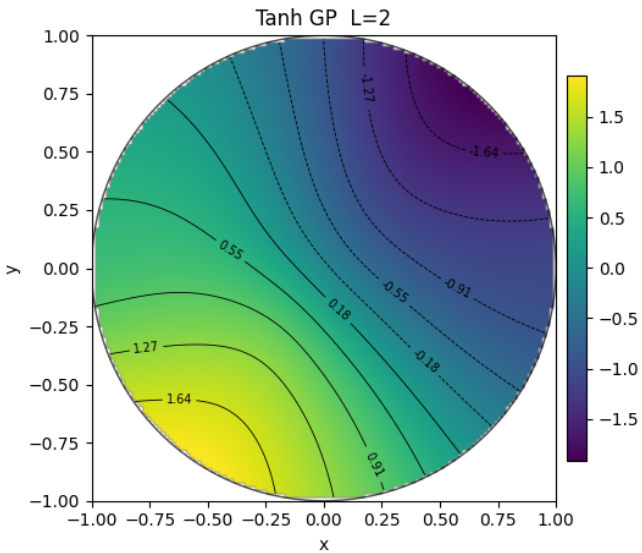


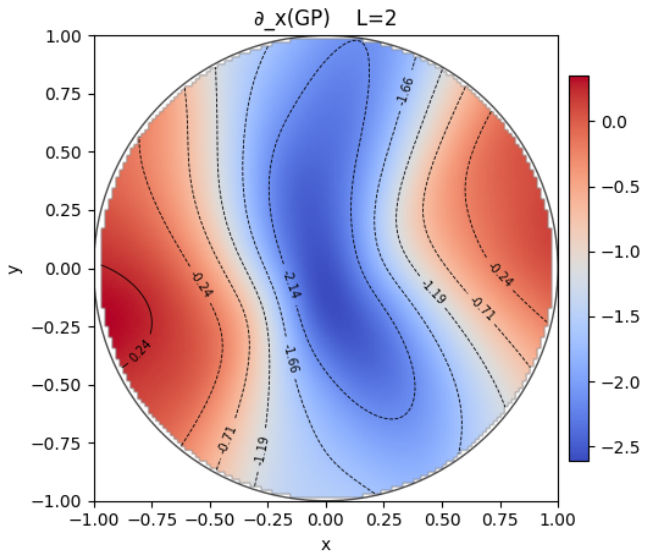


Tanh Networks on a Disk

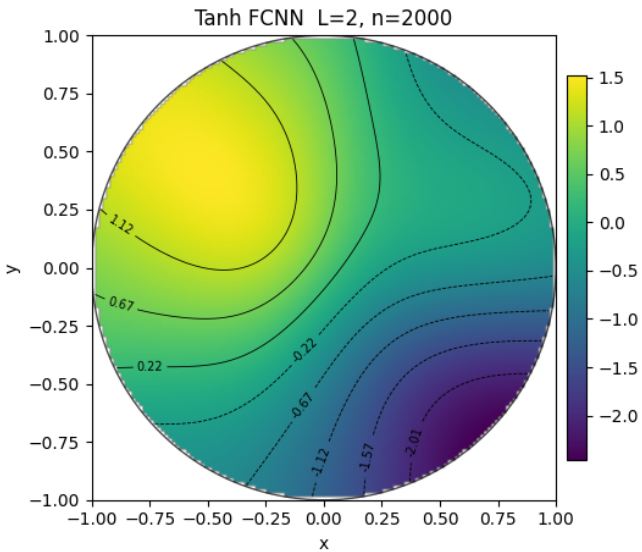


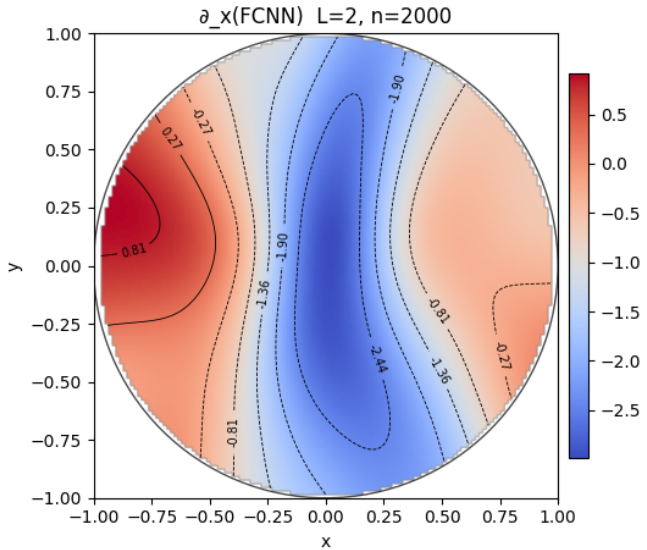


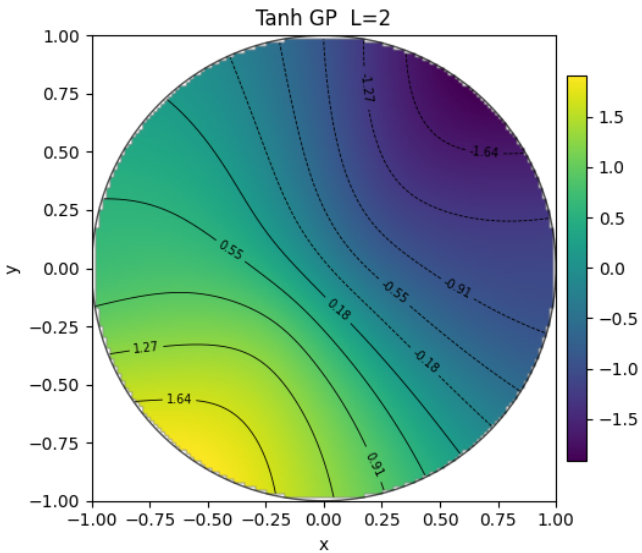


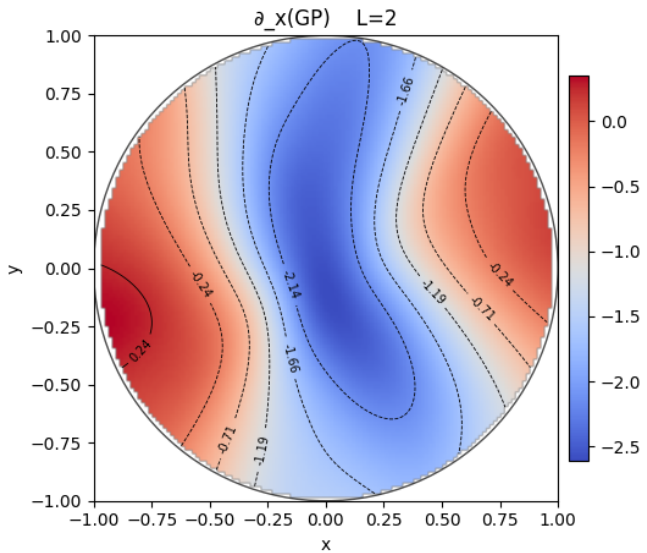


Tanh Networks on a Disk

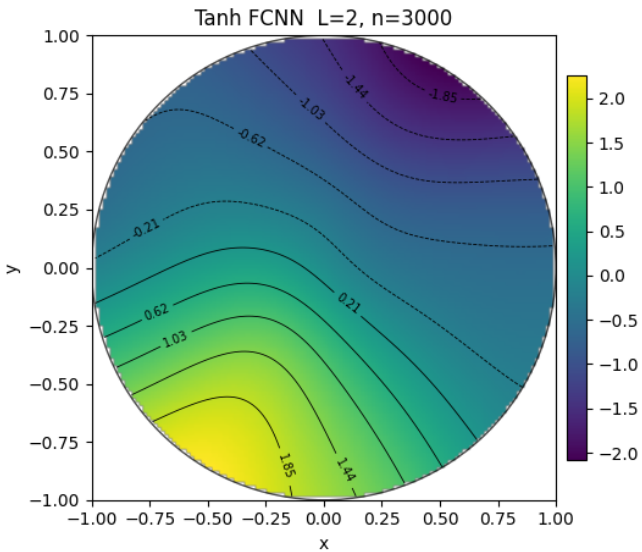


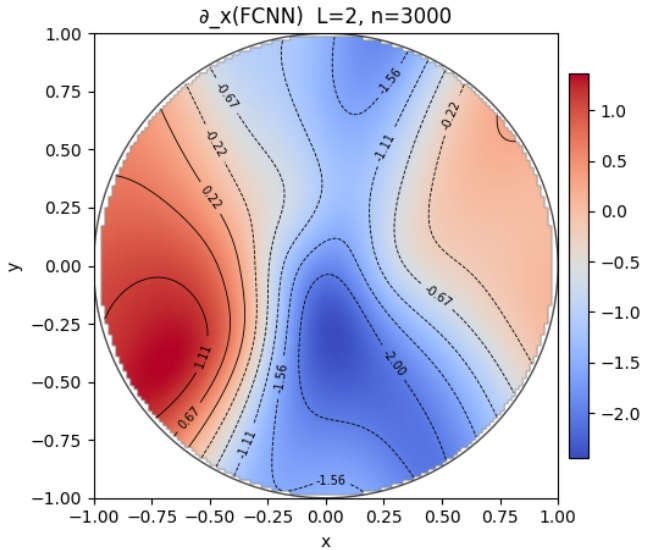


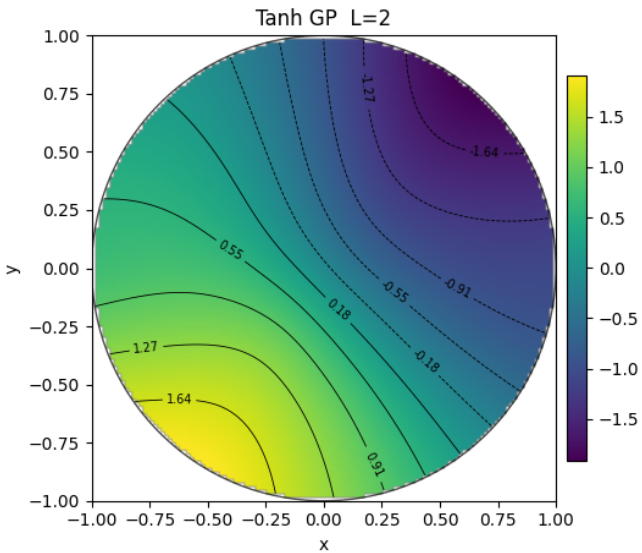


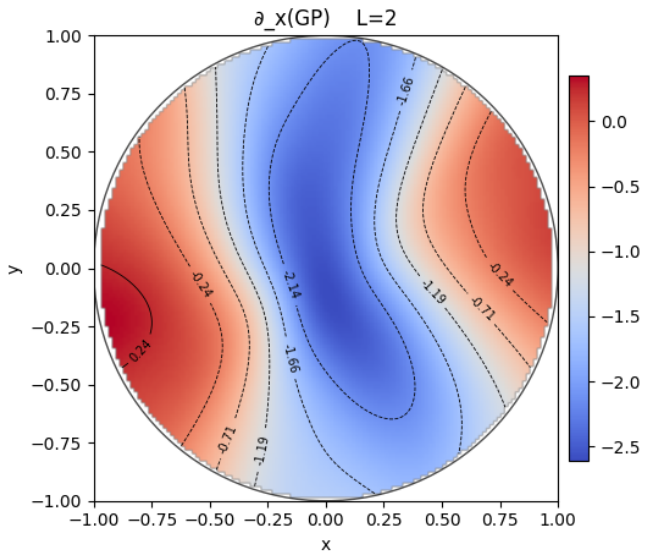


Tanh Networks on a Disk

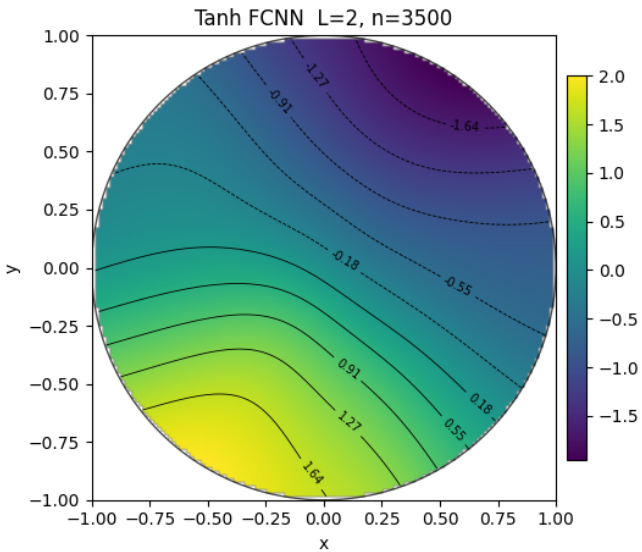


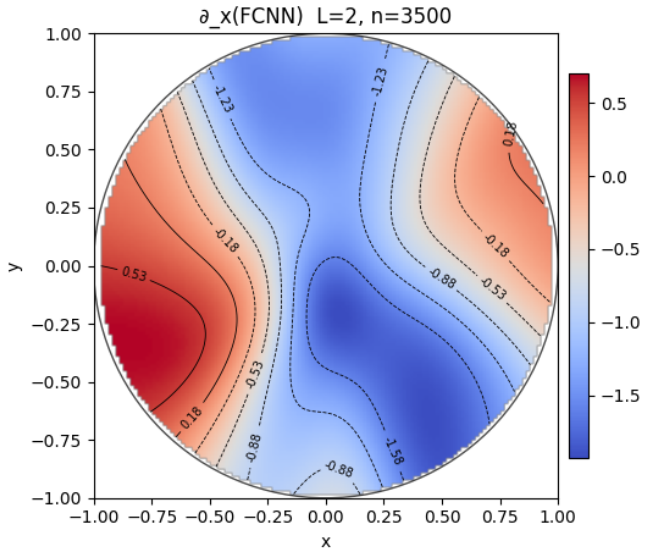


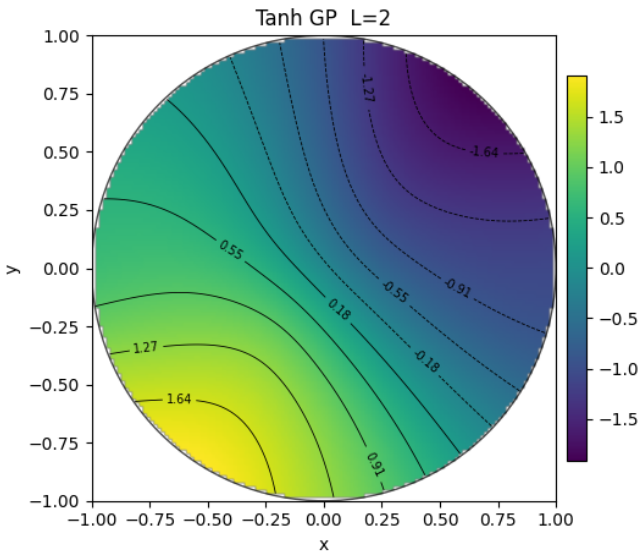


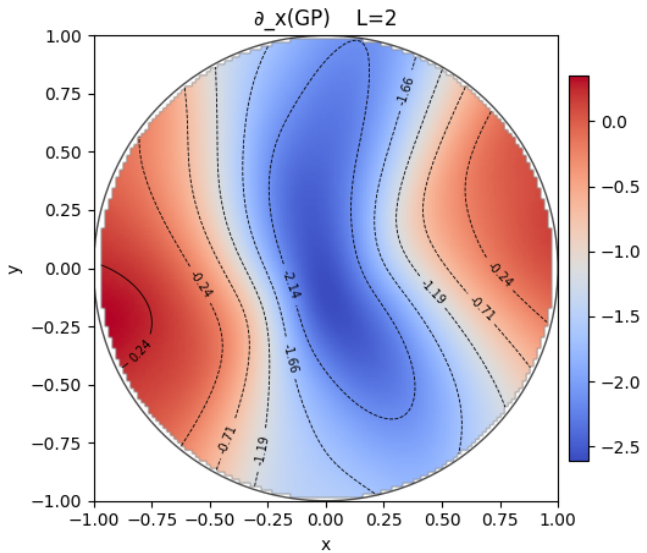


Tanh Networks on a Disk

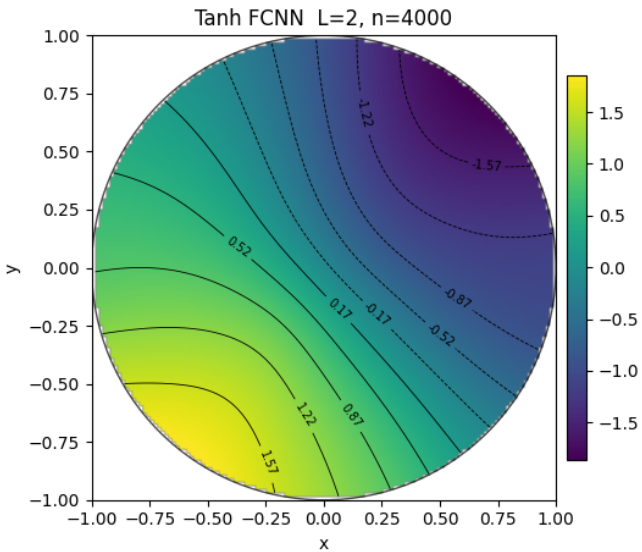


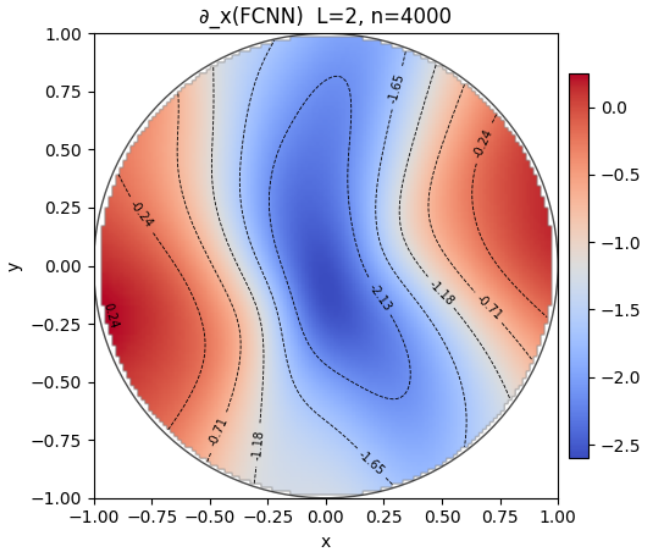


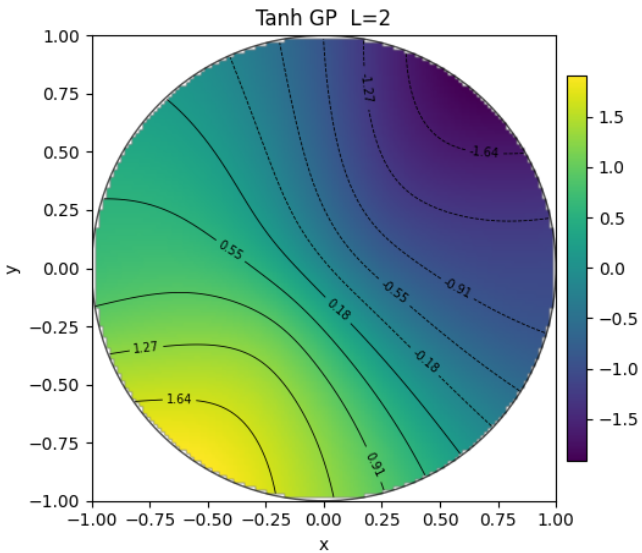


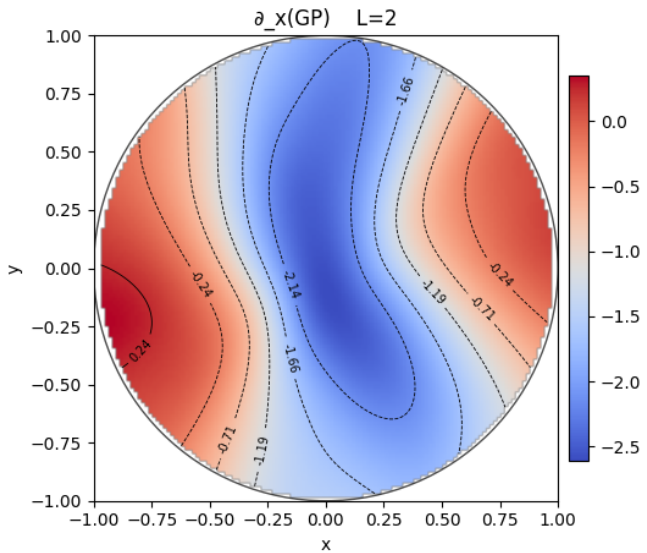


Tanh Networks on a Disk









LITERATURE

Finite-Dimensional

- ⋆ *Basteri & Trevisan* (2022); *Trevisan* (2023). **Arbitrary $L \geq 1$. Wass$_2$** bounds $\asymp n^{-1/2} \to n^{-1}$ (*coupling of Gaussian vectors*).
- ⋆ *Bordino, Favaro & Fortini* (2023). **$L = 1, 2$, scalar input/output.** Total variation bounds $\asymp n^{-1/2}$ ($L=1$), $\asymp n^{-1/4}$ ($L=2$) (2^{nd}*-order Poincaré*).
- ⋆ *Apollonio, De Canditiis, Franzina, Stolf & Torrisi* (2023). **Arbitrary L, one input.** Convex distance $\asymp n^{-1/2}$ (*smoothing & Stein*).
- ⋆ *Mosig, Agazzi & Trevisan* (2025). $L = 1$. Quantitative CLT **after training**.
- ⋆ *Celli* (2025-26). $L \geq 1$. **Non-Gaussian init.** & **Edgeworth expansions** .

Functional

- ⋆ *Eldan, Mikulincer & Schramm* (2021). $L=1$, **non-Gaussian init.** On sphere: **Wass$_2$** $\asymp 1/\log n$ (ReLU), **Wass$_\infty$** $\asymp n^{-1/2}$ (poly) (*transport CLTs*).
- ⋆ *Klukowski* (2021). $L = 1$, **non-Gaussian init.** **Wass$_2$** $\asymp n^{-1/2}$ (poly), $\asymp n^{-3/(2n_0)}$ (ReLU) (*Stein kernels*).
- ⋆ *Cammarota, Marinucci, Salvi & Vigogna* (2023). $L = 1$, **Gaussian init.** Smooth distance $\asymp n^{-1/2}$ (poly), $\asymp (\log n)^{-3/4}$ (ReLU) (*Stein*).
- ⋆ *Balasubramanian, Goldstein, Ross & Salim* (2024). **Arbitrary $L \geq 1$, Lipschitz σ. Wass$_\infty$** bounds with $\asymp \frac{n_{\ell+1}^4}{n_\ell}$ (*functional Stein*).

Fourth Cumulants (Variance of Covariance)

Proposition (Hanin, 2024)

Fix a compact $A \subset \mathbb{R}^{n_0}$, as well as multiindices I, J such that $|I|, |J| \leq r$. Then, under a certain non-degeneracy assumption on the limit covariance structure, there exists a constant C (depending on $A, L, r, \sigma, C_W, C_b$) such that

$$\mathbf{Var}(D^I D^J \Sigma^{(L)}(x,y)),$$

$$\left| \mathbb{E}[D^I D^J \Sigma^{(L)}(x,y)] - D^I D^J K^{(L+1)}(x,y) \right| \leq \frac{C}{n}.$$

Remark. The result extends to **arbitrary joint cumulants**. Part of a general "perturbative analysis" of DNNs, based on tight layer-to-layer recursive relations.

Reinforced Stein's Bounds

Proposition (Favaro, Hanin, Marinucci, Nourdin & Peccati, 2025)
Let F be a **conditionally Gaussian** *random variable, with random conditional variance A such that $\mathbb{E}[A^2] < \infty$, and $\mathbb{E}[A] = \sigma^2 > 0$. Then, for $Z \sim \mathcal{N}(0, \sigma^2)$, one has*

$$\mathbf{TV}(F, Z) \leq \frac{8}{\sigma^4}\mathbf{Var}(A) = \frac{8}{\sigma^4}\mathbf{Cum}_4(F).$$

Remark. The "usual" Stein's bounds yield

$$\mathbf{TV}(F, Z) \lesssim \mathbf{Var}(A)^{1/2}.$$

Bounds Based on Entropy

Theorem (Celli & Peccati, 2025)

Assume that $G \sim \mathcal{N}(0, K)$ *in* $\mathbb{R}^d$, *with* $K \in \mathbb{R}^{d\times d}$ **invertible**; F *a* **conditionally Gaussian random vector** *with conditional covariance* $A \in \mathbb{R}^{d\times d}$.

- *If* $\mathbb{E}[\|A\|_{HS}^8] < \infty$, $\mathbb{P}(\det(A) > 0) = 1$, *and* $\mathbb{E}[\|A^{-1}\|_{HS}^2] < \infty$, *then*

$$D(F\|G) \leq C_1 \, \|\mathbb{E}[A] - K\|_{HS}^2 + C_2 \, \mathbb{E}[\|A - K\|_{HS}^8]^{1/2},$$

  where $D(\bullet\|\bullet)$ = **relative entropy**.

- *If* $\mathbb{E}[\|A\|_{HS}^8] < \infty$, *then*

$$\mathbf{TV}(F, G) \leq C_3 \, \|\mathbb{E}[A] - K\|_{HS} + C_4 \, \mathbb{E}[\|A - K\|_{HS}^8]^{1/4}.$$

The constants C_1, C_2, C_3, C_4 *depend on* d *and* K.

Remark. (i) Analogous bounds in the 2-**Wasserstein** distance.
(ii) Proofs rely on **Pinsker's** and **Talagrand's** inequalities.

Wasserstein Distances on Hilbert Spaces

- ⋆ For $q \leq r$, $z^{(L+1)}$ and $G^{(L+1)}$ are $\mathbb{W}^{2;q}(\mathbb{U})$-**valued random elements**. Write $\Sigma^{(L)}$ and $K^{(L+1)}$ for their covariance kernels ($\Sigma^{(L)}$ is random!).
- ⋆ We know that, for the **trivial coupling**,

$$\mathbb{E}\left[\|\widetilde{z}^{(L+1)} - \widetilde{G}^{(L+1)}\|^2_{\mathbb{W}^{2;q}(\mathbb{U})}\right] = \mathbb{E}\left[\|\sqrt{\Sigma^{(L)}} - \sqrt{K^{(L+1)}}\|^2_{HS}\right].$$

- ⋆ **Problem**: $\|\sqrt{\Sigma^{(L)}} - \sqrt{K^{(L+1)}}\|^2_{HS}$ is *not* directly amenable to analysis. For instance, the classical estimate (*Powers-Størmer*):
  if $K^{(L+1)} \geq \epsilon\mathbb{I}$, *then*

$$\|\sqrt{\Sigma^{(L)}} - \sqrt{K^{(L+1)}}\|_{HS} \leq \epsilon^{-1/2}\|\Sigma^{(L)} - K^{(L+1)}\|_{HS},$$

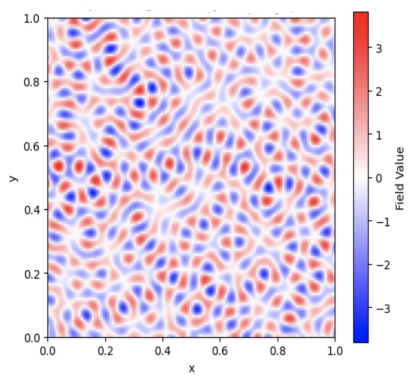
  is moot in our setting.

Powers-Størmer-Type Inequalities

Proposition (Dierickx, Nourdin, Peccati & Rossi, 2023)

Then,

$$\|\sqrt{\Sigma^{(L)}} - \sqrt{K^{(L+1)}}\|_{HS} \leq \left|\mathrm{Tr}\left(\Sigma^{(L)}\right) - \mathrm{Tr}\left(K^{(L+1)}\right)\right|^{1/2}$$
$$+\sqrt{2}\|\Sigma^{(L)} - K^{(L+1)}\|_{HS}^{\frac{1}{4}} \min\left\{\mathrm{Tr}\left(\sqrt{K^{(L+1)}}\right), \mathrm{Tr}\left(\sqrt{\Sigma^{(L)}}\right)\right\}^{1/2}.$$



Remark. The previous inequality is used in our recent study of **Gaussian random waves**. Context: **Universality of Berry's Random Waves** (see *Dierickx, Nourdin, Peccati & Rossi* (2023))

Bayesian Bounds

Write $\Theta := \{W^{(\ell)}, b^{(\ell)}\}$ for the **trainable parameter**,

$$D = \{(y_i, x_i) : i = 1, ..., p\} \subset \mathbb{R}^{n_{L+1}+n_0}$$

for a **training dataset**. Fix a bounded a strictly positive mapping

$$g : \mathbb{R}^{n_{L+1}+n_0} \to \mathbb{R}_+ : (y, x) \mapsto g(y, x),$$

and let $\{x_k^* : k = 1, ..., m\}$ be a set of **unseen inputs**.

From a Bayesian standpoint:

- $\star$
$$\textbf{Law}_\Theta(d\vartheta) = \textbf{prior distribution}$$
- $\star$
$$\prod_{i=1}^{p} g(z^{(L+1)}(x_i; \vartheta), y_i)\, \textbf{Law}_\Theta(d\vartheta) \propto p(d\vartheta \mid D) = \textbf{posterior}.$$
- $\star$
$$(B_1, ..., B_m) \mapsto \prod_{i=1}^{m} \mathbf{1}_{(z^{(L+1)}(x_i^*; \vartheta) \in B_i)}\, p(d\vartheta \mid D) = \textbf{(posterior) predictive.}$$

Bayesian Bounds

Theorem (Celli & Peccati, 2025)

Suppose $g(y,x)$ *is bounded and write*

$$\mathcal{X} := (x_1, ..., x_p) \quad \textit{and} \quad \mathcal{X}^* = (x_1^*, ..., x_m^*).$$

Then, there exist (explicit!) vectors

$$\mathbf{G} = (G_1, ..., G_p) \quad \textit{and} \quad \mathbf{G}^* = (G_1^*, ..., G_m^*)$$

such that

$$\mathbf{TV}\Big(\{\textbf{posterior of } z^{(L+1)}(\mathcal{X})\}, \mathbf{G}\Big) \lesssim \frac{1}{n}$$

and

$$\mathbf{TV}\Big(\{\textbf{predictive for } z^{(L+1)}(\mathcal{X}^*)\}, \mathbf{G}^*\Big) \lesssim \frac{1}{n}.$$

Moreover, $\mathbf{G}$, $\mathbf{G}^*$ *are Gaussian when* $g(y,x) = \exp\{-(y-x)^2\}$.

Some References and Final Words

1. L. Celli and G. Peccati, Entropic bounds for conditionally Gaussian vectors and applications to neural networks, ArXiv, 2025.

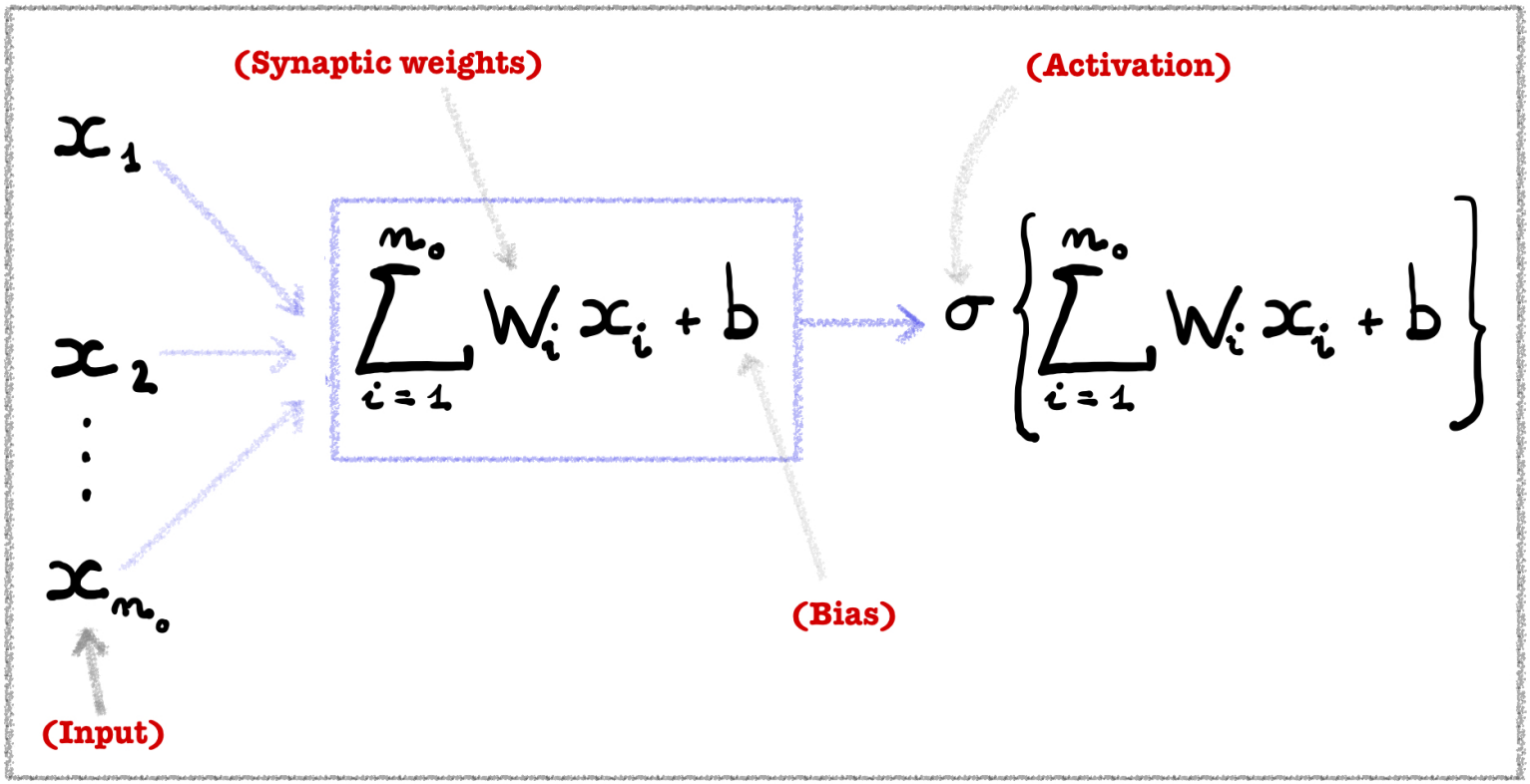
2. S. Favaro, B. Hanin, D. Marinucci, I. Nourdin, and G. Peccati, Quantitative CLTs in deep neural networks, *Probability Theory and Related Fields*, 2025.

3. B. Hanin, Random fully connected neural networks as perturbatively solvable hierarchies, *Journal of Machine Learning Research*, 2024.

4. D. Trevisan, Wide deep neural networks with Gaussian weights are very close to Gaussian processes, *arXiv:2312.11737*, 2023.

— THANK YOU FOR YOUR ATTENTION! —

Appendix: Neurons (No Depth!)

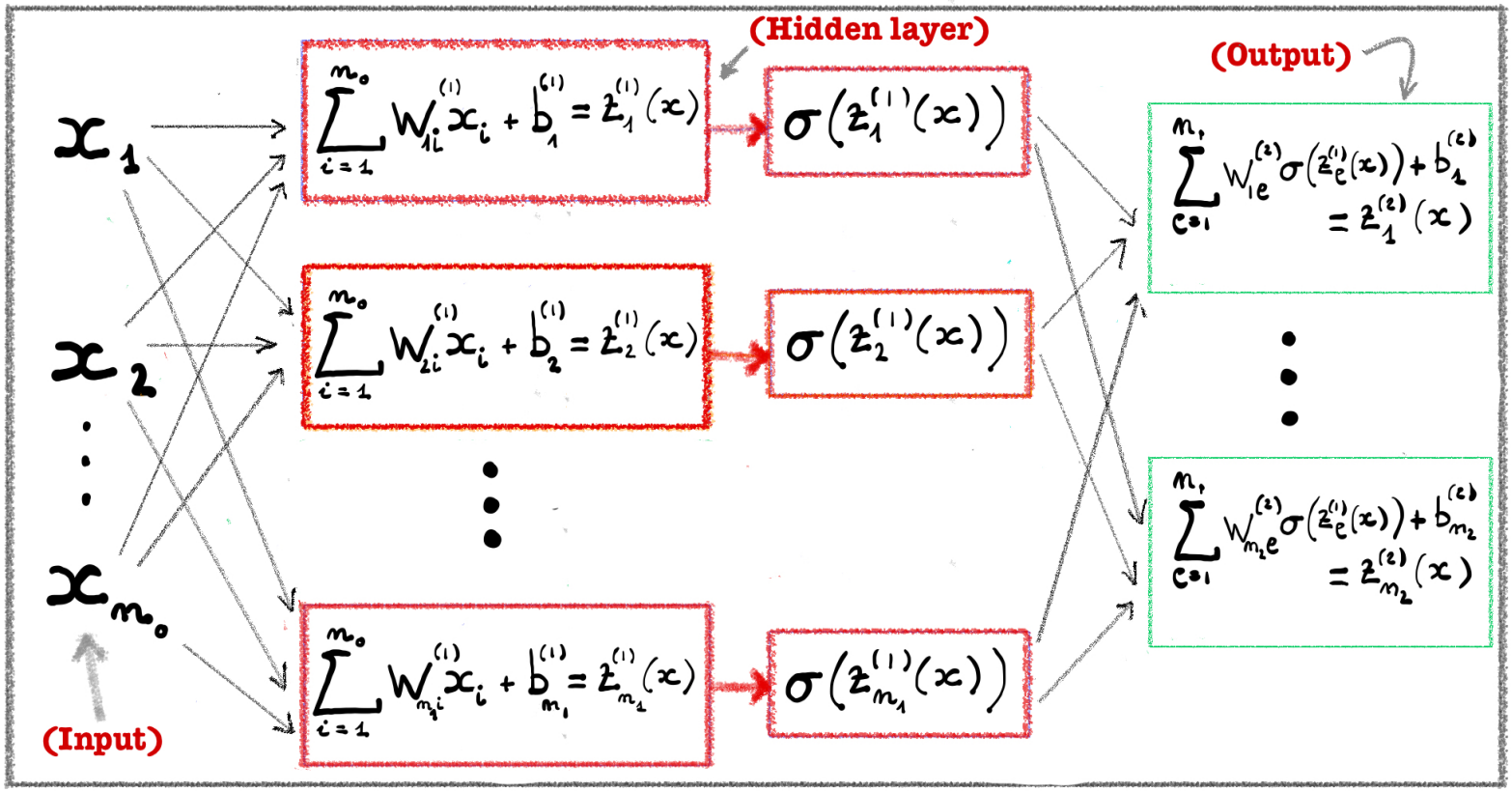


A **"neuron"** is a mapping from $\mathbb{R}^{n_0}$ (**input space**) into $\mathbb{R}$, composing an **affine transformation**

$$A : \mathbb{R}^{n_0} \to \mathbb{R} : x \mapsto Ax := Wx + b$$

with a **non-linear "activation"** $\sigma : \mathbb{R} \to \mathbb{R}$.

Appendix: Shallow Networks (Depth = 1)



A **"shallow neural network"** is the composition of a **hidden layer** of n_1 neurons, and a further **affine transformation**. We can write its output as:

$$z^{(2)}(x) = A^{(2)}(\sigma(z^{(1)}(x))) = A^{(2)}(\sigma(A^{(1)}x)),$$

where $A^{(1)} : \mathbb{R}^{n_0} \to \mathbb{R}^{n_1} : x \mapsto W^{(1)}x + b^{(1)}$, and

$$A^{(2)} : \mathbb{R}^{n_1} \to \mathbb{R}^{n_2} : z \mapsto W^{(2)}z + b^{(2)}.$$

Appendix: Entropy

⋆ **Definition (Relative entropy):** For X, Y as above let ν_X and ν_Y denote the laws of X and Y, respectively. Writing $\frac{d\nu_Y}{d\nu_X}$ for the density of the law of Y with respect to the law of X, the **relative entropy** of Y with respect to X is defined as

$$D(Y||X) := \int_{\mathbb{R}^d} \log\left(\frac{d\nu_Y}{d\nu_X}(z)\right)\nu_Y(dz) = E\left[\frac{d\nu_Y}{d\nu_X}(X)\log\left(\frac{d\nu_Y}{d\nu_X}(X)\right)\right],$$

with the convention that $0\log 0 = 0$.

⋆ **Theorem (Pinsker-Csizsar-Kullback inequality):** If X and Y are two random vectors in $\mathbb{R}^d$ and the law of X has a density with respect to the law of Y, then

$$\mathbf{TV}(X, Y) \leq \sqrt{\frac{1}{2}D(X||Y)}.$$

⋆ **Theorem (Talagrand's inequality):** Let $Y \sim \mathcal{N}_d(0, I_d)$ be a random vector in $\mathbb{R}^d$, where I_d is the $d \times d$ identity matrix, and let X be a random vector in $\mathbb{R}^d$ with $E[\|X\|^2] < \infty$. Then

$$W_2(X, Y) \leq \sqrt{2D(X||Y)}.$$