# BEYOND IID WEIGHTS: SPARSE AND LOW-RANK DEEP NEURAL NETWORKS ARE ALSO GAUSSIAN PROCESSES

**Thiziri Nait Saada, Alireza Naderi & Jared Tanner**
Mathematical Institute
University of Oxford
{naitsaadat, naderi, tanner}@maths.ox.ac.uk

## ABSTRACT

The infinitely wide neural network has been proven a useful and manageable mathematical model that enables the understanding of many phenomena appearing in deep learning. One example is the convergence of random deep networks to Gaussian processes that allows a rigorous analysis of the way the choice of activation function and network weights impacts the training dynamics. In this paper, we extend the seminal proof of Matthews et al. (2018) to a larger class of initial weight distributions (which we call PSEUDO-IID), including the established cases of IID and orthogonal weights, as well as the emerging low-rank and structured sparse settings celebrated for their computational speed-up benefits. We show that fully-connected and convolutional networks initialized with PSEUDO-IID distributions are all effectively equivalent up to their variance. Using our results, one can identify the Edge-of-Chaos for a broader class of neural networks and tune them at criticality in order to enhance their training.

## 1 INTRODUCTION

Deep neural networks are often studied at random initialization, in the limit of infinite width, where they have been shown to generate intermediate entries which approach Gaussian processes. Seemingly this was first studied for one-layer networks in Neal (2012) when the weight matrices have identically and independently distributed (IID) entries and became a popular model for deep networks following the seminal results for fully-connected networks in Lee et al. (2017) where orthogonal matrices were also studied. Specifically, Lee et al. (2017) gave a framework to compute the Gaussian process behaviour as a function of the nonlinear activation as well as the variance of the network weights and biases. This model was then used to explain the exploding and vanishing gradient phenomenon Schoenholz et al. (2017) and Pennington et al. (2018) amongst other network properties. The Gaussian process limit has since been extended to a broad class of network architectures and scaling limits, see Section 1.1.

In this paper, we extend the simultaneous scaling proof of the Gaussian process in Matthews et al. (2018) to a larger class of initial weight distributions (which we call PSEUDO-IID). The PSEUDO-IID distribution includes structured low-dimensional matrices such as low-rank and structured sparse settings celebrated for their computational efficiency and regularizing properties, as well as the already established cases of IID and orthogonal weights. The PSEUDO-IID distribution is defined in Definition 1 from the exchangeable distribution Definition 2 combined with a specified variance and a bounded high order moment.

**Definition 1** (PSEUDO-IID)**.** *Let $m, n$ be two integers. We will say that the random matrix $W = (W_{ij}) \in \mathbb{R}^{m\times n}$ is in the* PSEUDO-IID *distribution with parameter $\sigma^2$ if*

*(i) the matrix is row-exchangeable and column-exchangeable,*

*(ii) its entries are centered, uncorrelated, with variance $\mathbb{E}(W_{ij}^2) = \frac{\sigma^2}{n}$,*

*(iii) $\mathbb{E}\big|\sum_{j=1}^n a_j W_{ij}\big|^8 = K\|\mathbf{a}\|_2^8 n^{-4}$ for some constant $K$,*

*(iv) and $\lim_{n\to\infty} \frac{n^2}{\sigma^4}\mathbb{E}(W_{i_a,j}W_{i_b,j}W_{i_c,j'}W_{i_d,j'}) = \delta_{i_a,i_b}\delta_{i_c,i_d}$, for all $j \neq j'$.*

When $W^{(1)}$ has IID Gaussian entries and the other weight matrices $W^{(\ell)}$; $2 \leq \ell \leq L+1$, of a neural network (see 1) are drawn from a PSEUDO-IID distribution, we will say that the network is under the PSEUDO-IID regime.

Definition 2 (Exchangeability.) Let $X_1, \ldots, X_n$ be scalar or vector-valued random variables. We say $(X_i)_{i=1}^n$ are exchangeable if their joint distribution is invariant under permutations, i.e. $(X_1, \ldots, X_n) \overset{d}{=} (X_{\pi(1)}, \ldots, X_{\pi(n)})$ for all permutations $\pi : [n] \to [n]$. A random matrix is called row- (column-) exchangeable if its rows (columns) are exchangeable random vectors, respectively.

A row-exchangeable and column-exchangeable matrix $W \in \mathbb{R}^{m \times n}$ is not in general entrywise exchangeable, which means its distribution is not typically invariant under arbitrary permutations of its entries; particularly, out of $(mn)!$ possible permutations of its entries, $W$ only needs to be invariant under $m!n!$ of them — an exponentially smaller number. Matrices drawn uniformly from the Grassmanian of orthogonal matrices are a special case of random matrices satisfying row- and column-exchangeability and also entrywise exchangeability but are not independently distributed.

The conditions of PSEUDO-IID are sufficient to prove that fully-connected and convolutional networks are Gaussian processes, Theorems 1 and 2 respectively. Moreover, we are able to verify the PSEUDO-IID conditions for common initializations as well as parsimonious (e.g. sparse and/or low-rank) ones (see Section 3.1). We will illustrate these conditions further in Section 3.1 via some examples. However, the sharpest variant of Definition 1 condition (iii) remains an open question; its importance is expanded upon in Appendix D.

### 1.1 RELATED WORK

To the best of our knowledge, the Gaussian Process behaviour in the infinite width regime was first established by Neal (2012) in the case of one-layer fully-connected networks when the weights are sampled from standard distributions. The result has then been extended in Matthews et al. (2018) to deep fully-connected networks, where the depth is fixed, the weights distributed as centered Gaussians and the hidden layers widths growing jointly to infinity. Jointly scaling the network width substantially distinguishes their method of proof from the approach taken in Lee et al. (2017), where the authors considered a sequential limit analysis through layers. That is, analyzing the limiting distribution at one layer when the previous ones have already converged to their limiting distributions as in Lee et al. (2017), is significantly different from analyzing the limiting distribution at a current layer when the previous layers are jointly converging to their limits at the same time, as is done in Matthews et al. (2018).

Since this Gaussian Process behaviour has been established, two main themes of research have further been developed. The first one consists in the extension of such results for more general and complex architectures such as convolutional networks with many channels [Novak et al. (2020), Garriga-Alonso et al. (2019)] or any modern architectures composed of fully-connected, convolutional or residual connections, as summarized in Yang (2021), using the Tensor Program terminology. The second research theme concerns the generalization of this Gaussian Process behaviour to other possible weight distributions, such as orthogonal weights in Huang et al. (2021) or, alternatively, any IID weights with finite moments as derived in Hanin (2021). Note that the orthogonal case does not fit into the latter as entries are exchangeable but not independent (the first column of an orthogonal matrix cannot be independent from the second one in order to satisfy the orthogonality constraints). The same kind of results for general architectures in this setting have been derived by Yang in his Tensor Program framework Golikov & Yang (2022) and Yang (2021). Our contribution fits into this line of research, where we relax the independence requirement of the weight matrix entries and consider instead the PSEUDO-IID distribution of uncorrelated and exchangeable random variables. This broader distribution, Definition 1, enables us to present a unified proof that generalizes the approaches taken so far and encompasses all of them, for two types of architectures, namely, fully-connected and convolutional networks.

We conclude this section by mentioning that this Gaussian Process behaviour is a special case of a more general result about the convergence of wide neural networks towards a symmetrically distributed stochastic process Peluchetti et al. (2020) where independent, but not identically distributed, entries were considered in contrast to the setting of the present paper with identically distributed but not independent entries.

### 1.2 Organization of the paper

In Section 2, we focus on fully-connected neural networks, formally stating the associated Gaussian Process in Theorem 1, and outlining its proof in Section 2.2 with further technical details of its proof relegated to Appendix A. Our PSEUDO-IID regime unifies the previously studied settings of IID and orthogonal weights, while also allowing for novel settings such as low-rank Gaussian weights. We extend our results to the convolutional neural networks (CNNs) in Section 2.3. In Section 3, we provide examples of PSEUDO-IID distributions in practice, supporting our theoretical results with numerical simulations. Moreover, we explore the problem of stable initialization of deep networks, on the so-called Edge-of-Chaos (EOC), in our more expansive PSEUDO-IID regime. Lastly, in Section 4, we review our main contributions and put forward some further research directions.

## 2 Gaussian Process behaviour in the Pseudo-IID regime

We consider an untrained fully-connected neural network with width $N_\ell$ at layer $\ell \in \{1, \dots, L+1\}$. Its weights $W^{(\ell)} \in \mathbb{R}^{N_\ell \times N_{\ell-1}}$ and biases $b^{(\ell)} \in \mathbb{R}^{N_\ell}$ at layer $\ell$ are sampled from a centered probability distribution, respectively $\mathcal{P}_W^{(\ell)}$ and $\mathcal{P}_b^{(\ell)}$. Most commonly, the weights and biases are sampled IID Gaussian. Starting with such a network, with nonlinear activation $\phi : \mathbb{R} \to \mathbb{R}$, the propagation of any input data vector $z^{(0)}(x) := x \in \mathcal{X} \subset \mathbb{R}^{N_0}$ through the network is given by the following equations,

$$h_i^{(\ell)}(x) = \sum_{j=1}^{N_{\ell-1}} W_{ij}^{(\ell)} z_j^{(\ell-1)}(x) + b_i^{(\ell)}, \qquad z_j^{(\ell)}(x) = \phi(h_j^{(\ell)}(x)); \tag{1}$$

where $h^{(\ell)}(x) \in \mathbb{R}^{N_\ell}$ is referred to as the pre-activation vector at layer $\ell$, or the feature maps.

Throughout this paper, we will consider a specific set of activation functions that satisfy the so-called linear envelope property, Definition 3, which is satisfied by most activation functions used in practice (ReLu, Softmax, Tanh, HTanh, etc.).

Definition 3. (Linear envelope property) A function $\phi : \mathbb{R} \to \mathbb{R}$ is said to satisfy the linear envelope property if there exist $c, M \geq 0$ such that, for any $x \in \mathbb{R}$,

$$|\phi(x)| \leq c + M|x|. \tag{2}$$

### 2.1 The Pseudo-IID regime for fully-connected networks

Our proofs of PSEUDO-IID networks converging to Gaussian processes are done in the more sophisticated simultaneous width growth limit as pioneered by Matthews et al. (2018). For a review of the literature on deep networks with sequential vs. simultaneous scaling see Section 1.1. One way of characterizing such a simultaneous convergence over all layers is to consider that all widths $N_\ell$ are increasing functions of one parameter, let us say $n$, such that, as $n$ grows, all layers' widths grow: $\forall \ell \in \{1, \dots, L+1\}, N_\ell := N_\ell[n]$. We emphasize this dependence on $n$ by adding a suffix $[n]$ to the random variables $h$'s when $n$ is finite and denote by $h[\infty]$ its limiting distribution, corresponding to $n \to \infty$. The width of the first layer $N_0$ is fixed by the input dimension and the final output layer dimension $N_{L+1}$ do not scale with $n$ either. Moreover, the input data are assumed to come from a countably infinite input space $\mathcal{X}$. Equation 1 can thus be rewritten, for any $x \in \mathcal{X}$, as

$$h_i^{(\ell)}(x)[n] = \sum_{j=1}^{N_{\ell-1}[n]} W_{ij}^{(\ell)} z_j^{(\ell-1)}(x)[n] + b_i^{(\ell)}, \qquad z_j^{(\ell)}(x)[n] = \phi(h_j^{(\ell)}(x)[n]); \tag{3}$$

and the associated Gaussian process limit is given in Theorem 1.

Theorem 1 (GP limit for fully-connected PSEUDO-IID networks) Suppose a fully-connected neural network as in equation 3 is under the PSEUDO-IID regime with parameter $\sigma_W^2$ and the activation satisfies the linear envelope property Def. 3. Let $\mathcal{X}$ be a countably-infinite set of inputs. Then, for every layer $1 \leq \ell \leq L+1$, the sequence of random fields $(i, x) \in [N_\ell] \times \mathcal{X} \mapsto h_i^{(\ell)}(x)[n] \in \mathbb{R}^{N_\ell}$ converges in distribution to a centered Gaussian process $(i, x) \in [N_\ell] \times \mathcal{X} \mapsto h_i^{(\ell)}(x)[\infty] \in \mathbb{R}^{N_\ell}$, whose covariance function is given by

$$\mathbb{E}\left[h_i^{(\ell)}(x)[\infty]\, h_j^{(\ell)}(x')[\infty]\right] = \delta_{i,j} K^{(\ell)}(x, x'); \tag{4}$$

where

$$K^{(\ell)}(x, x') = \begin{cases} \sigma_b^2 + \sigma_W^2 \, \mathbb{E}_{(u,v)\sim N(0, K^{(\ell-1)}(x,x'))}[\phi(u)\phi(v)], & \ell \geq 1 \\ \sigma_b^2 + \frac{\sigma_W^2}{n_0} \langle x, x' \rangle, & \ell = 0 \end{cases}. \tag{5}$$

### 2.2 Sketch of the proof of Theorem 1: GP limit of fully-connected Pseudo-iid networks

This section includes the outline of our proof, which closely follows the steps taken in the original proof in Matthews et al. (2018) in the Gaussian setting for fully-connected networks. We refer the reader to Appendix A for its complete proof. Some of the results we use are shown in Matthews et al. (2018), so we intentionally choose our notation similarly to aid readers familiar with the prior paper. These steps are as follows:

1. The first step consists in reducing the problem of showing the convergence of the stochastic process $(h_i^{(\ell)}(x)[n])_{i \geq 1, [n] \subset \mathbb{N}, x \in \mathcal{X}}$ to a Gaussian Process, defined on a countably-infinite input space $\mathbb{N} \times \mathcal{X}$ (e.g. $\mathcal{X} = \{x_j\}_{j \in \mathbb{N}}$), to the convergence of a finite-dimensional vector $(h_i^{(\ell)}(x)[\,])_{(i;x) \in L}$, where $|L| < \infty$, to a multidimensional Gaussian. This is possible as the convergence towards the Gaussian Process is ensured with respect to the topology generated by a specific metric introduced in Matthews et al. (2018). We will therefore restrict our attention to showing the convergence in distribution of a finite-dimensional vector $(h_i^{(\ell)}(x)[n])_{(i;x) \in L}$, where $i$ refers to the neuron index and $x$ to the input data.
2. Given this finite-dimensional random vector, the problem is once again reduced to proving the convergence in distribution of any of its linear projections, which are scalars, to the corresponding linearly projected limiting Gaussians (see Cramér & Wold (1936)). We will thus consider the totality of these one-dimensional linear projections of the unbiased feature maps $h_i^{(\ell)}(x) - b_i^{(\ell)}$ onto $\alpha_{(i;x)}$,

$$T^{(\ell)}(\alpha; L)[n] := \sum_{(i;x) \in L} \alpha_{(i;x)} \left( h_i^{(\ell)}(x)[n] - b_i^{(\ell)} \right); \tag{6}$$

   where the suffix $[n]$ emphasizes on the joint limit to be taken simultaneously, as detailed in Section 2.1. Given the recursion formulae (equation 3) satisfied by the feature maps, the latter can be rewritten as

$$T^{(\ell)}(\alpha; L)[n] = \frac{1}{\sqrt{N_{\ell-1}[n]}} \sum_{j=1}^{N_{\ell-1}[n]} \gamma_j^{(\ell)}(\alpha; L)[n]; \tag{7}$$

   where, considering the renormalization $\tilde{w}_{ij}^{(\ell)} := w_{ij}^{(\ell-1)} \sqrt{N_{\ell-1}} W_{ij}^{(\ell)}$, the summands are defined as,

$$\gamma_j^{(\ell)}(\alpha; L)[n] := \sum_{(i;x) \in L} \alpha_{(i;x)} \tilde{w}_{ij}^{(\ell)} z_j^{(\ell-1)}(x)[n]. \tag{8}$$

3. The third step is to apply a version of the Central Limit Theorem (CLT) in the case where the random variables are exchangeable rather than independent. This CLT is often attributed to Blum & Rosenblatt (1956). The reason we need such an extension theorem is that, as opposed to taking the width limit sequentially, the distribution at the previous layer has not reached its limit, so when taking the joint limit we cannot claim the independence of the previous activities anymore and the standard CLT is no longer valid.
4. The last step is to proceed by induction through layers, and verifying at each layer that the moment assumptions of the exchangeable CLT hold in our Pseudo-iid regime.[1]

Note that these steps can also be found in Garriga-Alonso et al. (2019), where the proof technique of Matthews et al. (2018) is adapted to CNNs, with weights drawn from a Gaussian. In the following section, we define the Pseudo-iid regime for the CNNs and present an analogue of Theorem 1.

[1] The base case is carried out by the Gaussian condition on the weights of the first layer, guaranteed by our Pseudo-iid regime (see Definition 1).

### 2.3 THE PSEUDO-IID REGIME FOR CNNS

We consider a CNN with $C_\ell$ number of channels at layer $\ell \in \{1, \dots, L+1\}$ and two-dimensional convolutional filters $U^{(\ell)}_{i;j} \in \mathbb{R}^{k \times k}$ mapping the input channel $j \in \{1, \dots, C_{\ell-1}\}$ to the output channel $i \in \{1, \dots, C_\ell\}$. The input signal $X$ (also two-dimensional) has $C_0$ channels $x_1, \dots, x_{C_0}$ and its propagation through the network is given by

$$h^{(\ell)}_i(X)[n] = \begin{cases} b^{(1)}_i 1 + \sum_{j=1}^{C_0} U^{(1)}_{i;j} \star x_j\,, & \ell = 1 \\ b^{(\ell)}_i 1 + \sum_{j=1}^{C_{\ell-1}} U^{(\ell)}_{i;j} \star z^{(\ell-1)}_j(X)[n], & \ell \geq 2 \end{cases}; \qquad z^{(\ell)}_i(X)[n] = \phi(h^{(\ell)}_i(X))[n]. \tag{9}$$

In equation 9, $1$ should be understood as having the same size as the convolution output, non-linearity $\phi(\cdot)$ is applied entrywise, and we emphasize the simultaneous scaling with $n$ and the addition of $[n]$ to the feature maps $h^{(\ell)}_i(X)[n]$. We denote spatial (multi-) indices by boldface Greek letters, $\alpha, \beta$, etc., that are ordered pairs of integers taking values in the range of the size of the array. For example, if $X$ is an RGB ($C_0 = 3$) image of $H \times D$ pixels, $i = 2$, and $\alpha = (\alpha_1, \alpha_2)$, then $X_{i;\alpha}$ returns the Green intensity of the $(\alpha_1, \alpha_2)$ pixel. Moreover, we define $J \alpha K$ to be the patch centered at the pixel $\alpha$ covered by the filter, e.g. if $\alpha = (\alpha_1, \alpha_2)$ and the filter covers $k \times k = (2k_0 + 1) \times (2k_0 + 1)$ pixels, then $J \alpha K = \{(\alpha'_1, \alpha'_2) : \alpha_1 - k_0 \leq \alpha'_1 \leq \alpha_1 + k_0, \; \alpha_2 - k_0 \leq \alpha'_2 \leq \alpha_2 + k_0\}$, with the usual convention of zero-padding for the out-of-range indices. Sufficient conditions for PSEUDO-IID CNNs to converge to a Gaussian process in the simultaneous scaling limit are given in Definition 4.

**Definition 4** (PSEUDO-IID for CNNs). Consider a CNN with random filters and biases $\{U^{(\ell)}_{i;j}\}$ and $\{b^{(\ell)}_i\}$ as in equation 9. It is said to be in the PSEUDO-IID regime with parameter $\sigma_W^2$ if $U^{(1)}$ has IID $N(0; \frac{\sigma_W^2}{C_0})$ entries and, for $2 \leq \ell \leq L+1$,

(i) the convolutional kernel $U^{(\ell)} \in \mathbb{R}^{C_\ell \times C_{\ell-1} \times k \times k}$ is row-exchangeable and column-exchangeable, that is its distribution is invariant under permutations of first and second indices,

(ii) filters' entries are centered, uncorrelated, with variance $\mathbb{E}[(U^{(\ell)}_{i;j;\alpha})^2] = \sigma_W^2 / C_{\ell-1}$,

(iii) $\mathbb{E}\left[\left(\sum_{j=1}^{C_{\ell-1}} a_j U^{(\ell)}_{i;j;\beta}\right)^8\right] = K \|a\|_2^8 (C_{\ell-1})^{-4}$ for some constant $K$,

(iv) and $\lim_{n \to \infty} \frac{C_{\ell-1}[n]^2}{n^4} \mathbb{E}\left[U^{(\ell)}_{i_a;j;\alpha_a} U^{(\ell)}_{i_b;j;\alpha_b} U^{(\ell)}_{i_c;j';\alpha_c} U^{(\ell)}_{i_d;j';\alpha_d}\right] = \delta_{i_a;i_b} \delta_{i_c;i_d} \delta_{\alpha_a;\alpha_b} \delta_{\alpha_c;\alpha_d}$, for all $j \neq j'$.

**Theorem 2** (GP limit for CNN PSEUDO-IID networks). Suppose a CNN as in equation 9 is under the PSEUDO-IID regime with parameter $\sigma_W^2$ and the activation satisfies the linear envelope property Def. 3. Let $X$ be a countably-infinite set of inputs and $\alpha \in I$ denote a spatial (multi-) index. Then, for every layer $1 \leq \ell \leq L+1$, the sequence of random fields $(i, X; \alpha) \in [C_\ell] \times X \times I \mapsto h^{(\ell)}_{i;\alpha}(X)[n]$ converges in distribution to a centered Gaussian process $(i, X; \alpha) \in [C_\ell] \times X \times I \mapsto h^{(\ell)}_{i;\alpha}(X)[\,]$, whose covariance function is given by

$$\mathbb{E}\left[h^{(\ell)}_{i;\alpha}(X)[\,]\, h^{(\ell)}_{j;\beta}(X')[\,]\right] = \delta_{i;j} \left(\sigma_b^2 + \sigma_W^2 \sum_{\delta \in J \alpha K \cap J \beta K} K^{(\ell)}(X; X')\right), \tag{10}$$

where

$$K^{(\ell)}(X; X') = \begin{cases} \mathbb{E}_{(u,v) \sim N(0; K^{(\ell-1)}(X;X'))}[\phi(u)\phi(v)], & \ell \geq 1 \\ \frac{1}{C_0} \sum_{i=1}^{C_0} X_{i;\alpha} X'_{i;\beta}, & \ell = 0 \end{cases}. \tag{11}$$

Our proof of Theorem 2 for PSEUDO-IID CNNs is derived similarly to that of fully-connected networks, see Appendix B; Garriga-Alonso et al. (2019) developed a more restrictive proof to the setting of IID Gaussian CNNs. Note the extra index on the patch and how the fully-connected case is recovered when the filter kernel size is reduced to $k = 1$.

## 3 PSEUDO-IID IN PRACTICE

Untrained networks are typically initialized with IID weights, for example, the Gaussian and the uniform distributions used as default in PYTORCH. An important non-IID case that also leads to the Gaussian process limit is the random orthogonal initialization Huang et al. (2021). Our proposed PSEUDO-IID regime encompasses both IID and orthogonal weights as special cases and also allows for a broader class of weight distributions such as random low-rank orthogonal matrices Saada & Tanner (2023) and structured sparse matrices Dao et al. (2022a) and Dao et al. (2022b).

### 3.1 EXAMPLES OF PSEUDO-IID DISTRIBUTIONS

Here we give examples of PSEUDO-IID distributions that do not fit in the setting of the prior proofs of Gaussian process reviewed in Section 1.1.

IID weights. If $A = (A_{ij}) \in \mathbb{R}^{m \times n}$ has IID entries $A_{ij} \overset{iid}{\sim} D$, then it is automatically row- and column-exchangeable, and the entries are uncorrelated. Therefore, as long as the distribution of the weights $D$ satisfies the moment conditions of Definition 1, then the network is in the PSEUDO-IID regime. A sufficient condition is that $A_{ij}$ be sub-gaussian with parameter $O(n^{-1/2})$. Then, given the independence of entries, the random variable $\sum_{j=1}^{p} a_j A_{ij}$ would be sub-gaussian with parameter $\|a\|_2 n^{-1/2}$, and its $p^{th}$ moment is known to be $O(\|a\|_2^p n^{-p/2})$; see Vershynin (2018). Appropriately scaled Gaussian and uniform IID weights, for example, meet the sub-gaussianity criterion and therefore fall in the PSEUDO-IID class. Condition (iv) is trivial in this case.

Orthogonal weights. Let $A = (A_{ij}) \in \mathbb{R}^{n \times n}$ be drawn from the uniform (Haar) measure on the group of orthogonal matrices $O(n)$. While the entries are not independent, they are uncorrelated, and the rows and columns are exchangeable. To bound the moments, we may employ the concentration of the Lipschitz functions on the sphere, since one individual row of $A$, let us say the $i$-th one $(A_{i,1}, \ldots, A_{i,n})$, is drawn uniformly from $S^{n-1}$. Let $f(A_{i,1}, \ldots, A_{i,n}) := \sum_{j=1}^{n} a_j A_{ij}$, then $f$ is Lipschitz with constant $\|a\|_2$. By Theorem 5.1.4 of Vershynin (2018), the random variable $f(A_{i,1}, \ldots, A_{i,n}) = \sum_{j=1}^{n} a_j A_{ij}$ is sub-gaussian with parameter $k_{Lip} n^{-1/2} = \|a\|_2 n^{-1/2}$, which also implies $\mathbb{E}\left[\sum_{j=1}^{n} a_j A_{ij}\right]^p = O(\|a\|_2^p n^{-p/2})$. Moreover, the exact expressions of the $p^{th}$ moments are known for orthogonal matrices (see Collins et al. (2021) for an introduction to the calculus of Weingarten functions) and satisfy condition (iv).

Low-rank weights. Low-rank structures are widely recognized for speeding up matrix multiplications and can be used to reduce memory requirements of feature maps Price & Tanner (2023). Whilst such structures inevitably impose dependencies between the weight matrix entries $A \in \mathbb{R}^{m \times n}$, thus breaking the IID assumption, Saada & Tanner (2023) introduced a low-rank framework that falls within our PSEUDO-IID regime. Let $C := [C_1, \ldots, C_r] \in \mathbb{R}^{m \times r}$ be a uniformly drawn orthonormal basis for a random $r$-dimensional subspace. Let $P = (P_{ij}) \in \mathbb{R}^{r \times n}$ has IID entries $P_{ij} \overset{iid}{\sim} D$. If we set $A := CP$, it is easy to see that each column of $A$ is a linear combination of the columns given by $C$, with coefficients given by the matrix $P$. The row and column exchangeability of $A$ follows immediately from that of $C$ and $P$ and the moment conditions are controlled by the choice of distribution $D$. Direct computation of the four-cross product that appears in condition (iv) gives us

$$\mathbb{E}(A_{i_a;1} A_{i_b;1} A_{i_c;2} A_{i_d;2}) = s^2 \sum_{1 \le k,k' \le r} \mathbb{E}\left[C_{i_a;k} C_{i_b;k} C_{i_c;k'} C_{i_d;k'}\right],$$

where $s := \mathbb{E}(P_{1,1}^2)$. Using the expression in (Huang et al., 2021, Lemma 3) we can calculate the above expectation and deduce condition (iv) when $r$ is linearly proportional to $m$.

Permuted block-sparse weights. Block-wise pruned networks have recently been under extensive study for their efficient hardware implementation (Dao et al. (2022b), Dao et al. (2022a)). Once the sparsity pattern is fixed, we may apply random row and column permutations on the weight matrices without compromising the accuracy or the computational benefit. Let $A = (A_{ij}) \in \mathbb{R}^{m \times n}$ has IID entries $A_{ij} \overset{iid}{\sim} D$ and $B \in \mathbb{R}^{m \times n}$ be the binary block-sparse mask. Let $\tilde{A} = P_m (A \odot B) P_n$, where $P_m$ and $P_n$ are random permutation matrices of size $m$ and $n$ respectively, and $\odot$ represents entrywise multiplication. Then, by construction, $\tilde{A}$ is row- and column-exchangeable and,

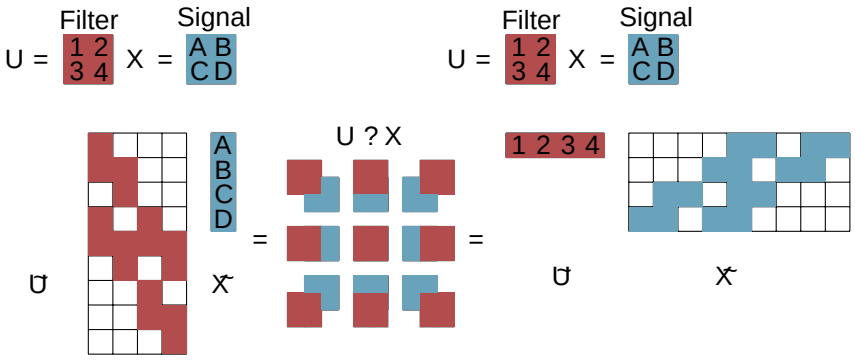


Figure 1: There exist multiple ways to compute convolutions between a tensor $U$ lter and a 2-dimensional signal $X$ (shown in the middle) based on matrix multiplications. We illustrate the approach taken in Garriga-Alonso et al. (2019) on the left, where the reshaping procedure is applied to the lter before computing a matrix multiplication, whilst the method we followed, shown in the right hand side of the gure consists in reshaping the signal in order to de ne special structures on the CNN lters such as orthogonality, sparsity and low-rank.

for suitable choices of underlying distribution $D$, it satis es the moment conditions of De nition 1. Some examples of realisations from such distribution are provided in Appendix E.

Orthogonal CNN lters. Unlike the fully-connected case, it is not obvious how to de ne the orthogonality of a convolutional layer and, once de ned, how to randomly generate such layers for initialization. Xiao et al. (2018) de nes an orthogonal convolutional kernel $U \in \mathbb{R}^{c_{out} \times c_{in} \times k \times k}$ made of $c_{out}$ lters of size $k$ by $k$ via the energy preserving property $\|U ? X\|_2 = \|X\|_2$, for any signal $X$ with $c_{in}$ input channels. Wang et al. (2020) requires the matricised version of the kernel to be orthogonal, while Qi et al. (2020) gives a more stringent de nition imposing isometry, i.e.

$$\sum_{i=1}^{c_{out}} U_{i,j} ? U_{i,j'} = \begin{cases} \cdot, & j = j' \\ 0, & \text{otherwise} \end{cases}.$$

Another de nition in Huang et al. (2021) calls for orthogonality of "spatial" slices $U_{\cdot,\cdot,\alpha,\beta} \in \mathbb{R}^{c_{out} \times c_{in}}$, for all positions $\alpha, \beta$.

We take a different approach than Wang et al. (2020) for matricising the tensor convolution operator, setting the stride to 1 and padding to 0: reshape the kernel into a matrix $\tilde{U} \in \mathbb{R}^{c_{out} \times k^2 c_{in}}$ and unfold the signal $X$ into $\tilde{X} \in \mathbb{R}^{k^2 c_{in} \times d}$, where $d$ is the number of patches depending on the sizes of the signal and the lter. This allows $\tilde{U}$ to be an arbitrary unstructured matrix rather than the doubly block-Toeplitz matrix in Wang et al. (2020), as shown in Figure 1. Matricising the tensor convolution operator, imposes the structure on the signal $\tilde{X}$ rather than the lter $U$. Orthogonal (i.e. energy-preserving) kernels can then be drawn uniformly random with orthogonal columns, such that

$$\tilde{U}^\top \tilde{U} = \frac{1}{k^2} I \tag{12}$$

and then reshaped into the original tensor kernel. Note that this construction is only possible when $\tilde{U}$ is a tall matrix with trivial null space, that is when $c_{out} \geq k^2 c_{in}$, otherwise the transpose might be considered instead. We emphasize that equation 12 is a suf cient (and not necessary) condition for $U$ to be energy-preserving, since $\tilde{X}$, by construction, has a very speci c structure. The set $\mathbb{R}^{k^2 c_{in} \times d}$, and, therefore, $\tilde{U}$ only needs to preserve norm on $\tilde{X}$ (and not everywhere). Therefore, we do not claim the generated orthogonal convolutional kernels is "uniformly distributed" over the set of all such kernels.

Now let us verify the conditions of De nition 4. Each ltU$_{i;j}$ is attened as$U_{i;j} \in \mathbb{R}^{1 \times k^2}$ and forms part of a row of$U$ as shown below:

$$U = \begin{bmatrix} U_{1;1} & U_{1;2} & \cdots & U_{1;c_{in}} \\ U_{2;1} & U_{2;2} & \cdots & U_{2;c_{in}} \\ \vdots & \vdots & \ddots & \vdots \\ U_{c_{out};1} & U_{c_{out};2} & \cdots & U_{c_{out};c_{in}} \end{bmatrix}. \tag{13}$$

Applying permutations on the indicesi andj translates to permuting rows and "column blocks" of the orthogonal matrix$U$, which does not affect the joint distribution of its entries. Hence, the kernel's distribution is unaffected, that is$U$ is row- and column-exchangeable. The moment conditions are both straightforward to check as$U_{i;j;\ell} = U_{i;(j-1)k^2+\ell}$; $\ell \in \{1, \ldots, k^2\}$, where is the counting number of the pixel . To check condition (iv), note that$\mathbb{E}\left[U^{(\grave{})}_{i_a;1;\alpha} U^{(\grave{})}_{i_b;1;\beta} U^{(\grave{})}_{i_c;2;\gamma} U^{(\grave{})}_{i_d;2;\delta}\right] = \mathbb{E}\left[U_{i_a;\alpha} U_{i_b;\beta} U_{i_c;k^2+\gamma} U_{i_d;k^2+\delta}\right]$, that is a four-cross product of the entries of an orthogonal matrix, whose expectation is explicitly known to be$\frac{C^{\cdot}+1}{(C^{\cdot}-1)C^{\cdot}(C^{\cdot}+2)} \delta_{i_a;i_b}\delta_{i_c;i_d}\delta_{\alpha;\beta}\delta_{\gamma;\delta}$ (Huang et al., 2021, Lemma 3).

### 3.2 Simulation of the Gaussian processes in Theorems 1 for fully-connected networks with Pseudo-iid

Theorem 1 establishes that in the in nite width simultaneous scaling the fully-connected Pseudo-iid networks converge to Gaussian processes. Here we conduct numerical simulations which validate this for modest dimensions of width$N^{\cdot} = n$ for $n = 30$, and300. Fig. 2 shows histograms of entries equation 3 weight matrices with uniform entries, Gaussian and with dropout, Pseudo-iid low-rank matrices, and Pseudo-iid structured sparse matrices.[2] Even at$n = 30$ there is excellent agreement of the histogram and the variance$h^{(\grave{})}_i(x)[\ ]$ in the in nite width limit. These histograms in Fig. 2 are quanti ed with Q-Q plots in Appendix F.

Fig. 3 explores the rate with which two independent inputs$x_a$ and$x_b$, generate uncorrelated Gaussian processes for the same neuron. This is done by plotting the joint distribution of$h^{(\grave{})}_i(x_a)[n]$ and $h^{(\grave{})}_i(x_b)[n]$ for the same value of and ` with the same network. Convergence to the limiting correlation between$h^{(\grave{})}_i(x_a)[\ ]$ and$h^{(\grave{})}_i(x_b)[\ ]$ given by Theorem 1 is also shown with the overlayed level curves. These experiments are conducted for weight matrices with uniform entries with dropout as well as Pseudo-iid orthogonal and Pseudo-iid Gaussian low-rank and structured sparse matrices. Interestingly the Pseudo-iid orthogonal converge to the large width distribution most quickly with good agreement at even$n = 3$. The other distributions considered show good agreement at$n = 30$ which improves at$n = 300$. The horizontal and vertical axis in each subplot of Fig. 7 are $h^{(5)}_1(x_a)$ and$h^{(5)}_1(x_b)$ respectively with$x_a$ and$x_b$ drawn independently.

### 3.3 Propagation of Gaussian processes through deep networks

Poole et al. (2016) and Xiao et al. (2018) developed formulae for the dynamics of the co-variance matrix of a Gaussian process through layers of fully-connected and convolutional networks respectively. These formulae determine xed points of the covariance matrices along with sensitivity of the networks to small perturbations of inputs. Speci cally, they derive the Edge-of-Chaos (EoC) condition for which the network is stable to perturbations. This same EoC condition was subsequently shown in Schoenholz et al. (2017) and Pennington et al. (2018) to avoid the exploding and vanishing gradient phenomenon.

All of the aforementioned results require the network to generate a Gaussian process in the large width limit. Theorems 1 and 2 extend this theory to networks in the Pseudo-iid regime, allowing for the rst initialization conditions for the rst time a rigorous EoC analysis of random networks with

[2] Experiments conducted for Fig. 2-7 used a fully-connected network with activation$\phi(x) = \tanh(x)$, weight variance$\sigma_w = 2$ and without bias. Dropout used probability 0.2 of setting an entry to zero, low-rank used rank$d = 2$, and block-sparsity used randomly permuted block-diagonal matrices with block size 5. The code to reproduce all these gures can be found at https://shorturl.at/gNOQ0 .

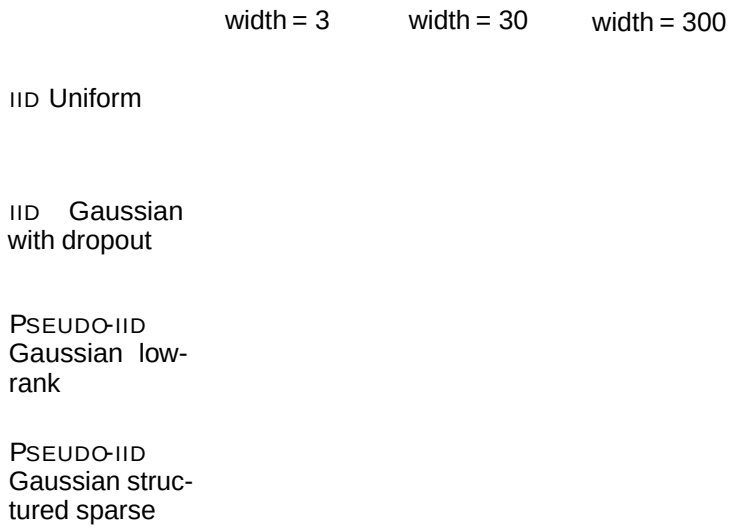


Figure 2: For different instances of the PSEUDO-IID regime, as the width of a fully-connected network grows, the pre-activation given in the first neuron at the fifth layer tend to a Gausssian whose moments are given by Theorem 1. The experiments were conducted 10000 times on a network whose depth is set to be 7 and the input data is sampled from $S^8$.

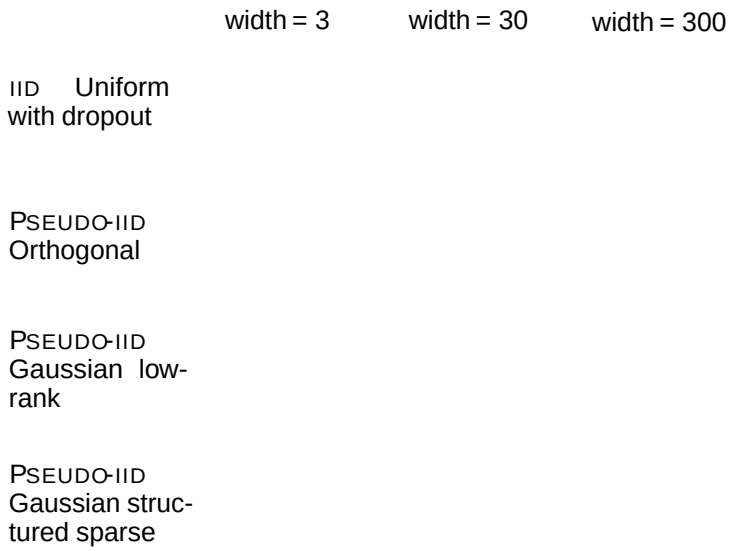


Figure 3: The empirical joint distribution of two pre-activation values at the same neuron resulting from two distinct inputs. Different PSEUDO-IID regimes for fully-connected networks are simulated and their large width limiting distribution from Theorem 1 included as level curves. The input data are taken from $S^9$ and 10000 experiments were conducted on 7-layer deep network.

parsimonious initialization as developed in Saada & Tanner (2023). More precisely, Theorems 1 and 2 can be expected to form the foundation of developing initialization theory for networks designed for greater ef ciency (and accuracy) by using structured sparse and low-rank weight matrices.

## 4 CONCLUSION

Here we have presented the  rst Gaussian process theory for deep networks with weight matrices having low-dimensional dependent entries. Theorems 1 and 2 for fully-connected and convolutional networks allow calculation of conditions necessary for initialization of these networks which train ef ciently Saada & Tanner (2023). We anticipate these theorems will be extended to additional network architectures and may serve as a roadmap for yet other network regularizers. Moreover, the theory presented here can be further re ned to determine  nite dimensional corrections, following the approach of Roberts et al. (2022), such as the rate of convergence of these quantities and the variance of these quantities for  nite dimensions.

## ACKNOWLEDGMENTS

Thiziri Nait Saada is supported by the UK Engineering and Physical Sciences Research Council (EPSRC) through the grant EP/W523781/1. Jared Tanner is supported by the Hong Kong Innovation and Technology Commission (InnoHK Project CIMDA).

## REFERENCES


Patrick Billingsley. Convergence of probability measures; 2nd ed. Wiley series in probability and statistics. Wiley, Hoboken, NJ, 1999. URL https://cds.cern.ch/record/1254129.

Julius R. Blum and Murray Rosenblatt. A class of stationary processes and a central limit theorem. Proceedings of the National Academy of Sciences of the United States of America, 42:7:412–3, 1956.

Benoit Collins, Sho Matsumoto, and Jonathan Novak. The weingarten calculus, 2021.

H. Cramér and H. Wold. Some theorems on distribution functions. Journal of the London Mathematical Society, s1-11(4):290–294, 1936. doi: https://doi.org/10.1112/jlms/s1-11.4.290. URL https://londmathsoc.onlinelibrary.wiley.com/doi/abs/10.1112/jlms/s1-11.4.290.

Tri Dao, Beidi Chen, Kaizhao Liang, Jiaming Yang, Zhao Song, Atri Rudra, and Christopher Ré. Pixelated butterfly: Simple and efficient sparse training for neural network models, 2022a.

Tri Dao, Beidi Chen, Nimit Sohoni, Arjun Desai, Michael Poli, Jessica Grogan, Alexander Liu, Aniruddh Rao, Atri Rudra, and Christopher Ré. Monarch: Expressive structured matrices for efficient and accurate training, 2022b.

Adrià Garriga-Alonso, Carl Edward Rasmussen, and Laurence Aitchison. Deep convolutional networks as shallow gaussian processes, 2019.

Eugene Golikov and Greg Yang. Non-gaussian tensor programs. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), Advances in Neural Information Processing Systems, 2022. URL https://openreview.net/forum?id=AcHUIG2wA8-.

Boris Hanin. Random neural networks in the infinite width limit as gaussian processes, 2021.

Wei Huang, Weitao Du, and Richard Yi Da Xu. On the neural tangent kernel of deep networks with orthogonal initialization, 2021.

Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S. Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. Deep neural networks as gaussian processes, 2017. URL https://arxiv.org/abs/1711.00165.

Alexander G. de G. Matthews, Mark Rowland, Jiri Hron, Richard E. Turner, and Zoubin Ghahramani. Gaussian process behaviour in wide deep neural networks, 2018. URL https://arxiv.org/abs/1804.11271.

Radford M Neal. Bayesian learning for neural networks, volume 118. Springer Science & Business Media, 2012.

Roman Novak, Lechao Xiao, Jaehoon Lee, Yasaman Bahri, Greg Yang, Jiri Hron, Daniel A. Abola a, Jeffrey Pennington, and Jascha Sohl-Dickstein. Bayesian deep convolutional networks with many channels are gaussian processes, 2020.

Stefano Peluchetti, Stefano Favaro, and Sandra Fortini. Stable behaviour of infinitely wide deep neural networks. In Silvia Chiappa and Roberto Calandra (eds.), Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, volume 108 of Proceedings of Machine Learning Research, pp. 1137–1146. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/peluchetti20b.html.

Jeffrey Pennington, Samuel Schoenholz, and Surya Ganguli. The emergence of spectral universality in deep networks. In International Conference on Artificial Intelligence and Statistics, pp. 1924–1932. PMLR, 2018.

Ben Poole, Subhaneil Lahiri, Maithra Raghu, Jascha Sohl-Dickstein, and Surya Ganguli. Exponential expressivity in deep neural networks through transient chaos, 2016. URL https://arxiv.org/abs/1606.05340.

Ilan Price and Jared Tanner. Improved projection learning for lower dimensional feature maps. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5. IEEE, 2023.

Haozhi Qi, Chong You, Xiaolong Wang, Yi Ma, and Jitendra Malik. Deep isometric learning for visual recognition. In International conference on machine learning, pp. 7824–7835. PMLR, 2020.

Daniel A. Roberts, Sho Yaida, and Boris Hanin. The Principles of Deep Learning Theory. Cambridge University Press, 2022. https://deeplearningtheory.com .

Thiziri Nait Saada and Jared Tanner. On the initialisation of wide low-rank feedforward neural networks, 2023. URL https://arxiv.org/abs/2301.13710 .

Samuel S. Schoenholz, Justin Gilmer, Surya Ganguli, and Jascha Sohl-Dickstein. Deep information propagation. In International Conference on Learning Representations, 2017. URL https://openreview.net/forum?id=H1W1UN9gg .

Roman Vershynin. High-dimensional probability: An introduction with applications in data science, volume 47. Cambridge university press, 2018.

Jiayun Wang, Yubei Chen, Rudrasis Chakraborty, and Stella X Yu. Orthogonal convolutional neural networks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 11505–11515, 2020.

Lechao Xiao, Yasaman Bahri, Jascha Sohl-Dickstein, Samuel Schoenholz, and Jeffrey Pennington. Dynamical isometry and a mean field theory of CNNs: How to train 10,000-layer vanilla convolutional neural networks. In Jennifer Dy and Andreas Krause (eds.), Proceedings of the 35th International Conference on Machine Learning, volume 80 of Proceedings of Machine Learning Research, pp. 5393–5402. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/xiao18a.html .

Greg Yang. Tensor programs i: Wide feedforward or recurrent neural networks of any architecture are gaussian processes, 2021.

# A Proof of Theorem 1: Gaussian Process behaviour in Fully-Connected Networks in the Pseudo-iid regime

## A.1 Step 1: Reduction of the problem from countably-infinite to finite dimensional

Firstly, we must clarify in what sense a sequence of stochastic processes $h_i[\cdot]$, $i \in \mathbb{N}$ converges in distribution to its limit $h_i[\cdot]$. For a sequence of real-valued random variables, we can define convergence in distribution $X_n \xrightarrow{d} X$ by the following condition: $Ef(X_n) \to Ef(X)$ for all continuous functions $f : \mathbb{R} \to \mathbb{R}$. Similarly, we can define weak convergence for random objects taking values in $\mathbb{R}^{\mathbb{N}}$ (countably-indexed stochastic processes), provided that we equip $\mathbb{R}^{\mathbb{N}}$ with a "good" topology[3] that maintains a sufficiently rich class of continuous functions. The metric on the space of real sequences $\mathbb{R}^{\mathbb{N}}$ defined by

$$(h, h') := \sum_{i=1}^{\infty} 2^{-i} \min(1, |h_i - h'_i|) \tag{14}$$

is one example. Thus, we can speak of the weak convergence $h_i[\cdot] \xrightarrow{d} h_i[\cdot]$ in the sense that $Ef(h_i[n]) \to Ef(h_i[\cdot])$ for all $f : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}$ continuous with respect to the metric

Fortunately, to prove the weak convergence of in finite-dimensional distributions, it is sufficient to show the convergence of their finite-dimensional marginals Billingsley (1999).

## A.2 Step 2: Reduction of the problem from multidimensional to one-dimensional

Let $L = \{(i_1, x_1), \ldots, (i_P, x_P)\}$ be a finite subset of the index set $\mathbb{N} \times \mathbb{N}$. We need to show that the vector $h^{(\ell)}_{i_p}(x_p)[n] \in \mathbb{R}^P$ converges in distribution to $h^{(\ell)}_{i_p}(x_p)[\cdot] \in \mathbb{R}^P$. By the Cramér-Wold theorem (Cramér & Wold (1936)), we may equivalently show the weak convergence of an arbitrary linear projection

$$T^{(\ell)}(\cdot; L)[n] = \sum_{(i;x) \in L} h^{(\ell)}_i(x)[n] \, b^{(\ell)}_i \tag{15}$$

$$= \sum_{(i;x) \in L} \sum_{j=1}^{N_{\ell-1}[n]} W^{(\ell)}_{ij} z^{(\ell-1)}_j(x)[n] \tag{16}$$

$$= \sqrt{\frac{1}{N_{\ell-1}[n]}} \sum_{j=1}^{N_{\ell-1}[n]} \nu^{(\ell)}_j(\cdot; L)[n]; \tag{17}$$

where

$$\nu^{(\ell)}_j(\cdot; L)[n] := \sum_{(i;x) \in L} w^{(\ell)}_{ij} z^{(\ell-1)}_j(x)[n]; \tag{18}$$

and $w_{ij}$ is centered and normalized, i.e. $E w_{ij} = 0$; $E w^2_{ij} = 1$. We will be using a suitable version of Central Limit Theorem (CLT) to prove the weak convergence of the series in equation 17 to a Gaussian random variable.

## A.3 Step 3: Use of an exchangeable Central Limit Theorem

The classical Central Limit Theorem (CLT) establishes that for a sequence of iid random variables, the properly scaled sample mean converges to a Gaussian random variable in distribution. Here we recall an extension of the Central Limit Theorem introduced in Blum & Rosenblatt (1956), where the independence assumption on the summands is relaxed and replaced by an exchangeability condition. The following statement is an adapted version derived in Matthews et al. (2018), which is more suited to our case.

[3] In fact, it needs to be a Polish space, i.e. a complete separable metric space.

Theorem 3 (Matthews et al. (2018), Lemma 10) For each positive integer $n$, let $(X_{n;j} ; j \in \mathbb{N})$ be an infinitely exchangeable process with mean zero, finite variance, and finite absolute third moment. Suppose also that the variance has a limit, $\lim_{n\to\infty} \sigma_n^2 = \sigma^2$. Define

$$S_n := \frac{1}{\sqrt{N[n]}} \sum_{j=1}^{N[n]} X_{n;j} ; \tag{19}$$

where $N : \mathbb{N} \to \mathbb{N}$ is a strictly increasing function. If

(a) $\mathbb{E}[X_{n;1} X_{n;2}] = 0$,

(b) $\lim_{n\to\infty} \mathbb{E}[X_{n;1}^2 X_{n;2}^2] = \sigma^4$,

(c) $\mathbb{E}[|X_{n;1}|^3] = o(\sqrt{N[n]})$,

then $S_n$ converges in distribution to $\mathcal{N}(0; \sigma^2)$.

Comparing equation 17 with equation 19, we need to check if the summands $X_{n;j} = z_j^{(\ell)}(\cdot\,; L)[n]$ satisfy the conditions of Theorem 3. We will carefully verify each condition in the following sections.

### A.3.1 Exchangeability of the summands

To apply Theorem 3, we must first prove that the random variables $z_j^{(\ell)}(\cdot\,; L)[n]$ are exchangeable. Let us expand the expression for $z_j^{(\ell)}(\cdot\,; L)[n]$ further and write it in terms of the pre-activations of layer $\ell - 2$, i.e.

$$\begin{aligned} z_j^{(\ell)}(\cdot\,; L)[n] &= \sum_{(i;x)\in\mathcal{L}} w_{ij}^{(\ell)} \left(h_j^{(\ell-1)}(x)[n]\right) \\ &= \sum_{(i;x)\in\mathcal{L}} w_{ij}^{(\ell)} \sum_{k=1}^{N^{z^2}[n]} W_{jk}^{(\ell-1)} z_k^{(\ell-2)}(x)[n] + b_j^{(\ell-1)} : \end{aligned}$$

If we apply a random permutation on the index set, let us say $\pi : \{1; \ldots; N_{\ell-1}\} \to \{1; \ldots; N_{\ell-1}\}$, then the row-exchangeability and column-exchangeability of the PSEUDO-IID weights ensure that $W_{i\,(j)}^{(\ell)}$ has same distribution as $W_{ij}^{(\ell)}$ and that $W_{(j)k}^{(\ell-1)}$ has same distribution as $W_{jk}^{(\ell-1)}$, for any $i \in \{1; \ldots; N_\ell\}$, $k \in \{1; \ldots; N_{\ell-2}\}$. Note that this extends easily to the normalized versions of the weights $w_{ij}^{(\ell)}$. Additionally, as the biases are set to be Gaussians, they are a fortiori exchangeable and their distributions remain unchanged when considering random permutations of indices. Therefore, $z_j^{(\ell)}(\cdot\,; L)[n]$ is equal to $z_{(j)}^{(\ell)}(\cdot\,; L)[n]$ in distribution, hence exchangeability.

### A.3.2 Moment conditions

We mean by moment conditions the existence of a limiting variance, as well as the conditions (a)-(c) in Theorem 3. We will prove the moment conditions by induction on the layer number $\ell$.

Existence of the limiting variance of the summands. To show the existence of the limiting variance of the summands, let us first write down such a variance at a finite width. Since $z_j^{(\ell)}$'s are exchangeable (see A.3.1), their distribution is identical and we may simply calculate the variance of

$\overset{(\ell)}{1}$ as follows.

$$
\begin{aligned}
(\overset{(\ell)}{}{}^2(;L)[n] &:= E\left[\overset{(\ell)}{1}(;L)[n]^2\right] = E\Big[\Big(\sum_{(i;x)} w^{(i;x)2L}\overset{(\ell)}{i;1} z_1^{(\ell-1)}(x)[n]\Big)^2\Big] \\
&= \sum_{\substack{(i_a;x_a)2L\\(i_b;x_b)2L}} w^2_{(i_a;x_a)\,(i_b;x_b)} E\left[\overset{(\ell)}{i_a;1}\overset{(\ell)}{i_b;1}\right] E\left[z_1^{(\ell-1)}(x_a)[n]\; z_1^{(\ell-1)}(x_b)[n]\right] \\
&= \sum_{\substack{(i_a;x_a)2L\\(i_b;x_b)2L}} w^2_{(i_a;x_a)\,(i_b;x_b)}\, \delta_{i_a;i_b} E\left[z_1^{(\ell-1)}(x_a)[n]\; z_1^{(\ell-1)}(x_b)[n]\right]; \qquad (20)
\end{aligned}
$$

where we first considered the independence between the normalized weights at layer $\ell$ and the activations at layer $\ell-1$; then used the fact that the normalized weights are uncorrelated in the PSEUDO-IID regime.

The convergence of the second moment of the summands is thus dictated by the convergence of the covariance of the activations of the last layer. By the induction hypothesis, the feature maps $h_j^{(\ell-1)}(x)[n]$ converge in distribution, so the continuous mapping theorem guarantees the existence of a limiting distribution for the activations $z_1^{(\ell-1)}(x_a)[n]$ and $z_1^{(\ell-1)}(x_b)[n]$ as $n$ tends to infinity. Note that this result holds even if the activation function has a set of discontinuity points of Lebesgue measure zero, e.g. the step function. Thus, the product inside the expectation in equation 20 converges in distribution to a limiting random variable $z_1^{(\ell-1)}(x_a)[\,]z_1^{(\ell-1)}(x_b)[\,]$.

From Billingsley (1999), one knows that if a sequence weakly converges to a limiting distribution and is uniformly integrable, then we can swap the order of taking the limit and the expectation. Thus, by Proposition 1, we have

$$
\begin{aligned}
(\overset{(\ell)}{}{}^2(;L)[\,] &:= \lim_{n\to\infty} (\overset{(\ell)}{}{}^2(;L)[n] \\
&= \sum_{\substack{(i_a;x_a)2L\\(i_b;x_b)2L}} w^2_{(i_a;x_a)\,(i_b;x_b)}\, \delta_{i_a;i_b} E\left[z_1^{(\ell-1)}(x_a)[\,]\; z_1^{(\ell-1)}(x_b)[\,]\right]: \qquad (21)
\end{aligned}
$$

Condition (a). At a given layer $\ell$, we need to show that $X_{n;1} := \overset{(\ell)}{1}(;L)[n]$ and $X_{n;2} := \overset{(\ell)}{2}(;L)[n]$ are uncorrelated. We have

$$
\begin{aligned}
E\left[X_{n;1}X_{n;2}\right] &= E\Big[\Big(\sum_{(i;x)} w^{(i;x)2L}\overset{(\ell)}{i;1} z_1^{(\ell-1)}(x)[n]\Big)\Big(\sum_{(i;x)} w^{(i;x)2L}\overset{(\ell)}{i;2} z_2^{(\ell-1)}(x)[n]\Big)\Big] \\
&= \sum_{\substack{(i_a;x_a)2L\\(i_b;x_b)2L}} w^2_{(i_a;x_a)\,(i_b;x_b)} E\left[\overset{(\ell)}{i_a;1}\overset{(\ell)}{i_b;2}\right] E\left[z_1^{(\ell-1)}(x_a)[n]\; z_2^{(\ell-1)}(x_b)[n]\right] \\
&= 0;
\end{aligned}
$$

since $\overset{(\ell)}{i_a;1}$ and $\overset{(\ell)}{i_b;2}$ are uncorrelated by the PSEUDO-IID assumption.

Condition (b).

$$\begin{aligned}
\mathbb{E}\left[X^2_{n;1}X^2_{n;2}\right] &= \mathbb{E}\left[\Big(\sum_{(i;x)} w_{i;1}^{(\grave{})} z_1^{(\grave{}-1)}(x)[n]\Big)^2 \Big(\sum_{(i;x)} w_{i;2}^{(\grave{})} z_2^{(\grave{}-1)}(x)[n]\Big)^2\right] \\
&= \sum_{(i_t;x_t)\in L \atop t\in\{a;b;c;d\}} w_{i_t;x_t}^4 \, \mathbb{E}\left[\prod_{t\in\{a;b\}} z_1^{(\grave{}-1)}(x_t)[n] \prod_{t\in\{c;d\}} z_2^{(\grave{}-1)}(x_t)[n]\right] \\
&= \sum_{(i_t;x_t)\in L \atop t\in\{a;b;c;d\}} w_{i_t;x_t}^4 \, F_n(i_a;i_b;i_c;i_d) \, \mathbb{E}\left[\prod_{t\in\{a;b\}} z_1^{(\grave{}-1)}(x_t)[n] \prod_{t\in\{c;d\}} z_2^{(\grave{}-1)}(x_t)[n]\right];
\end{aligned} \tag{22}$$

where

$$F_n(i_a;i_b;i_c;i_d) := \mathbb{E}_{i_{a;1}\, i_{b;1}\, i_{c;2}\, i_{d;2}}. \tag{23}$$

We justify the convergence in distribution of the random variable inside the expectation in the exact same way as in the previous section, referring to the continuous mapping theorem and the induction hypothesis. By Proposition 1, as $n \to \infty$, the above expectation converges to the expectation of the limiting Gaussian process, i.e.

$$\begin{aligned}
&\lim_{n\to\infty} \mathbb{E}\left[\prod_{t\in\{a;b\}} z_1^{(\grave{}-1)}(x_t)[n] \prod_{t\in\{c;d\}} z_2^{(\grave{}-1)}(x_t)[n]\right] \\
&\quad = \mathbb{E}\left[\prod_{t\in\{a;b\}} z_1^{(\grave{}-1)}(x_t)[\infty] \prod_{t\in\{c;d\}} z_2^{(\grave{}-1)}(x_t)[\infty]\right].
\end{aligned}$$

Moreover, condition (iv) of De nition 1 implies that

$$\lim_{n\to\infty} F_n(i_a;i_b;i_c;i_d) = \delta_{i_a;i_b}\,\delta_{i_c;i_d}.$$

Substituting the two limits back in the equation 22 and using the independence of the activations at layer $\grave{}-1$ given by the induction hypothesis, we get

$$\begin{aligned}
\lim_{n\to\infty} \mathbb{E}\left[X^2_{n;1}X^2_{n;2}\right] &= \sum_{(i_t;x_t)\in L \atop t\in\{a;b;c;d\}} w_{i_t;x_t}^4 \, \delta_{i_a;i_b}\,\delta_{i_c;i_d} \, \mathbb{E}\left[\prod_{t\in\{a;b\}} z_1^{(\grave{}-1)}(x_t)[\infty] \prod_{t\in\{c;d\}} z_2^{(\grave{}-1)}(x_t)[\infty]\right] \\
&= \sum_{(i_t;x_t)\in L \atop t\in\{a;b\}} w_{i_t;x_t}^4 \, \delta_{i_a;i_b} \, \mathbb{E}\left[\prod_{t\in\{a;b\}} z_1^{(\grave{}-1)}(x_t)[\infty]\right] \sum_{(i_t;x_t)\in L \atop t\in\{c;d\}} w_{i_t;x_t}^4 \, \delta_{i_c;i_d} \, \mathbb{E}\left[\prod_{t\in\{c;d\}} z_1^{(\grave{}-1)}(x_t)[\infty]\right] \\
&= \mu_4(\grave{}) \, \mu_4(\grave{};L)[\infty].
\end{aligned}$$

Condition (c). To show that the third absolute moment of the $\mu_1^{(\grave{})}(\grave{};L)[n]$ grows slower than $\sqrt{N_{\grave{}-1}[n]}$ as $n \to \infty$, it is sufficient to bound it by a constant. Applying Hölder's inequality on $X = \mu_1^{(\grave{})}(\grave{};L)[n]^3$; $Y = 1$; $p = 4/3$; $q = 4$, we obtain

$$\mathbb{E}\left[\mu_1^{(\grave{})}(\grave{};L)[n]^3\right] \leq \mathbb{E}\left[\mu_1^{(\grave{})}(\grave{};L)[n]^4\right]^{\frac{3}{4}} \leq 1.$$

Therefore, condition (c) boils down to showing that $\lim_{n\to\infty} E\left[\tilde{\Sigma}_1^{(\ell)}(\cdot\,; L)[n]\right]^4$ is finite, for which we will once again make use of the uniform integrability of the feature maps derived in C. Define

$$G_n(i_a, i_b, i_c, i_d) := E\left[W_{i_a,1} W_{i_b,1} W_{i_c,1} W_{i_d,1}\right], \tag{24}$$

and observe that

$$\begin{aligned} E\left[\tilde{\Sigma}_1^{(\ell)}(\cdot\,; L)[n]\right]^4 &= E\left[\sum_{(i,x)} w_{(i,x)}\, \tilde{z}_{i,1}^{(\ell-1)}(x)[n]\right]^4 \\ &= \sum_{\substack{(i_t, x_t) \in L \\ t \in \{a,b,c,d\}}} \prod_{t \in \{a,b,c,d\}} w_{(i_t,x_t)}\, G_n(i_a, i_b, i_c, i_d) E\left[\prod_{t \in \{a,b,c,d\}} \tilde{z}_1^{(\ell-1)}(x_t)[n]\right]. \end{aligned}$$

Using Cauchy-Schwarz inequality, we can bound $G_n$ by the fourth moment of the normalized weights, i.e.

$$\begin{aligned} G_n &\le \sqrt{\mathrm{Var}(W_{i_a,1} W_{i_b,1}) \mathrm{Var}(W_{i_c,1} W_{i_d,1})} \le \mathrm{Var}(W_{i_a,1} W_{i_b,1}) \\ &\le E\left[W_{i_a,1}^2 W_{i_b,1}^2\right], \quad \mathrm{Var}(W_{i_a,1}^2) \le E\left[W_{i_a,1}^4\right] - 1. \end{aligned}$$

Then, we use condition (iii) of the Definition 1 with $p = 4$ and $a = (1, 0, \ldots, 0)^\top$ to bound the fourth moment:

$$\begin{aligned} E\left[W_{i_a,1}^4\right] &= \frac{n^2}{4} E\left[W_{i_a,1}^4\right] \\ &= \frac{n^2}{4} K_4 \|a\|_4^4 n^{-2} = \frac{K_4}{4} = o_n(1). \end{aligned}$$

Furthermore, the induction hypothesis gives the convergence in distribution of the feature maps from the last layer, and combined with the continuous mapping theorem we get the convergence in distribution of the above product inside expectation. Using Lemma 1, the uniform integrability of the activations follows, and Billingsley's theorem (Lemma 2) enables us to swap the limit and the expectation. Thus,

$$\lim_{n\to\infty} E\left[\prod_{t \in \{a,b,c,d\}} \tilde{z}_1^{(\ell-1)}(x_t)[n]\right] = E\left[\prod_{t \in \{a,b,c,d\}} \tilde{z}_1^{(\ell-1)}(x_t)[\,]\right].$$

To bound the product of four different random variables, it is sufficient to bound the fourth order moment of each (see Lemma 3). We can do so using the linear envelope property (Definition 3) satisfied by the activation function to get, for any $(i,x) \in L$,

$$E\left[\tilde{z}_1^{(\ell-1)}(x)[\,]\right]^4 \le 2^{4-1} E\left[c^4 + M^4 h_1^{(\ell-1)}(x)[\,]^4\right].$$

The induction hypothesis indicates that $h_1^{(\ell-1)}(x)[\,]$ follows a Gaussian distribution, whose fourth moment is bounded. Using the fact that we chose the set $L$ to be finite, we can take the supremum over all $x$. Therefore, we have

$$E\left[\tilde{z}_1^{(\ell-1)}(x_t)[\,]\right]^4 \le 2^{4-1} \sup_{(i,x) \in L} E\left[c^4 + M^4 h_1^{(\ell-1)}(x)[\,]^4\right] = o_n(1).$$

Combining the above bounds, we then have

$$\lim_{n\to\infty} E\left[\tilde{\Sigma}_1^{(\ell)}(\cdot\,; L)[n]\right]^4 < 1\,.$$

### A.3.3 Conclusion from the exchangeable CLT

In the above sections, we showed by induction that at any depth, if the feature maps from previous layers converge in distribution to GP Gaussian processes, then the assumptions of Theorem 3 hold and the one-dimensional projection of the feature maps at the current layer also converges in distribution to a Gaussian random variable with a specified variance. More precisely, we showed that for any finite set $L$ and projection vector $\lambda$, any linear one-dimensional projection of the feature maps at the current layer, $T^{(\ell)}(\lambda; L)[n]$, converges in distribution to a Gaussian $\mathcal{N}(0, \Sigma^{(\ell)}(\lambda; L)[n])$ as $n$ grows. This gives the convergence of the feature maps at layer $\ell$ to Gaussian processes.

Considering the unbiased quantity

$$T^{(\ell)}(\lambda; L)[n] := \sum_{(i;x)\in L} \lambda_{(i;x)} \left( h_i^{(\ell)}(x)[n] - b_i^{(\ell)} \right),$$

we can compute its variance:

$$\mathbb{E}\left[ T^{(\ell)}(\lambda; L)[n] \right]^2 = \sum_{\substack{(i_a;x_a)\in L \\ (i_b;x_b)\in L}} \lambda_{(i_a;x_a)} \lambda_{(i_b;x_b)} \left( \mathbb{E}\left[ h_{i_a}^{(\ell)}(x_a)[n] \, h_{i_b}^{(\ell)}(x_b)[n] \right] - \sigma_b^2 \right)_{i_a, i_b}.$$

As we saw in the previous sections,

$$\Sigma^{(\ell)}(\lambda; L)[n] = \sigma_W^2 \sum_{\substack{(i_a;x_a)\in L \\ (i_b;x_b)\in L}} \lambda_{(i_a;x_a)} \lambda_{(i_b;x_b)} \left( \mathbb{E}\left[ z_1^{(\ell-1)}(x_a)[n] \, z_1^{(\ell-1)}(x_b)[n] \right] \right)_{i_a, i_b}.$$

Thus, by identification, and using the inductive hypothesis, one recovers the recursion formula for the variance as described in Theorem 1: For any $i_a, i_b \in \mathbb{N}$, $x_a, x_b \in \mathbb{X}$,

$$\begin{aligned} \mathbb{E}\left[ h_{i_a}^{(\ell)}(x_a)[n] \, h_{i_b}^{(\ell)}(x_b)[n] \right] &= \delta_{i_a, i_b} \left( \sigma_W^2 \, \mathbb{E}\left[ z_1^{(\ell-1)}(x_a)[n] \, z_1^{(\ell-1)}(x_b)[n] \right] + \sigma_b^2 \right) \\ &= \delta_{i_a, i_b} \left( \sigma_W^2 \, \mathbb{E}_{(u,v)\sim \mathcal{N}\left(0, K^{(\ell-1)}(x,x')\right)} \left[ \phi(u) \, \phi(v) \right] + \sigma_b^2 \right). \end{aligned}$$

### A.4 Identical distribution and independence over neurons

As we saw, for any $n$, the feature maps $h_j^{(\ell)}(\lambda; L)[n]$ are exchangeable, and, in particular, identically distributed. This still holds after taking the limit, that is $h_j^{(\ell)}[n]$ and $h_k^{(\ell)}[n]$ have the same distribution for any $j, k \in \mathbb{N}$.

It still remains to show the independence between $h_i^{(\ell)}[n]$ and $h_j^{(\ell)}[n]$ for $i \neq j$. As we now know the limiting distribution is Gaussian, it suffices to analyze their covariance to conclude about their independence. As derived in the previous section, for any $x, x' \in \mathbb{X}$, $i \neq j$, $\mathbb{E}\left[ h_i^{(\ell)}(x)[n] \, h_j^{(\ell)}(x')[n] \right] = 0$, hence the independence.

## B Proof of Theorem 2: Gaussian Process behaviour in Convolutional Neural Networks in the Pseudo-IID regime

We apply the same machinery to show the Gaussian Process behaviour in CNNs under the Pseudo-IID regime, closely following the steps detailed in the fully-connected case. To reduce the problem to a simpler one, one can proceed as previously by considering a finite subset of the feature maps at layer $\ell$, $L = \{(i_1; X_1, \alpha_1), \ldots, (i_P; X_P, \alpha_P)\} \subset [C_\ell] \times \mathbb{X} \times I$, where $I$ consists of all the spatial multi-indices. We will follow the same strategy outlined in Sections A.1-A.3. Given a finite set $L$ and the projection vector $\lambda \in \mathbb{R}^{|L|}$, we may form the unbiased one-dimensional projection as

$$T^{(\ell)}(\lambda; L)[n] := \sum_{(i; X, \alpha)\in L} \lambda_{(i; X, \alpha)} \left( h_{i;\alpha}^{(\ell)}(X)[n] - b_i^{(\ell)} \right), \tag{25}$$

which can be rewritten, using equation 9, as the sum

$$T^{(\ell)}(\,;\,L)[n] = \frac{1}{\sqrt{C^{(\ell-1)}[n]}} \sum_{j=1}^{C^{(\ell-1)}[n]} {}^{(\ell)}_{j}(\,;\,L)[n]; \tag{26}$$

where the summands are

$${}^{(\ell)}_{j}(\,;\,L)[n] := \sum_{(i;X;\,)\in L} \sum_{\in J\ K} E^{(\ell)}_{i;j;\,} z^{(\ell-1)}_{j;\,}(X)[n]: \tag{27}$$

As before, we introduced the renormalized version $E^{(\ell)}$ of the lter $U^{(\ell)}$ such that $E^{(\ell)}_{i;j;} := \frac{1}{w}\sqrt{C^{(\ell-1)}}\,U^{(\ell)}_{i;j;}$.

We will proceed once again by induction forward through the network's layer verifying the assumptions of the exchangeable CLT (Theorem 3) and using it on $T^{(\ell)}(\,;\,L)[n]$ to conclude its convergence in distribution to a Gaussian random variable.

The exchangeability of the summands. Similar to the fully-connected case, we employ the row- and column-exchangeability of the PSEUDO-IID convolutional kernel to show the exchangeability of ${}^{(\ell)}_{j}$'s. Let us expand equation 27 and write

$$\begin{aligned} {}^{(\ell)}_{j}(\,;\,L)[n] &= \sum_{(i;X;\,)\in L} \sum_{\in J\ K} E^{(\ell)}_{i;j;\,} \left(h^{(\ell-1)}_{j;\,}(X)[n]\right) \\ &= \sum_{(i;X;\,)\in L} \sum_{\in J\ K} E^{(\ell)}_{i;j;\,} \sum_{k=1}^{C^{(\ell-2)}[n]} \sum_{\in J\ K} U^{(\ell-1)}_{j;k;\,} z^{(\ell-2)}_{k;\,}(X)[n] \ : \end{aligned}$$

The joint distributions of $U^{(\ell)}_{i;j;\ j=1}^{C^{(\ell-1)}}$ and $U^{(\ell-1)}_{j;k;\ j=1}^{C^{(\ell-1)}}$ are invariant under any permutation $\in S_\infty$ (j), therefore ${}^{(\ell)}_{j}$'s are exchangeable.

Moment conditions. Condition (a) is straightforward as the lters' entries are uncorrelated by the PSEUDO-IID assumption $E\left[U^{(\ell)}_{i;j;}\,U^{(\ell)}_{i^0;j^0;\ 0}\right] = \frac{2}{C^{(\ell-1)}}\,{}_{i;i^0}\,{}_{j;j^0}\,{}_{;\ 0}$.

The moment conditions are shown to be satis ed by induction through the network and the proofs boil down to showing the uniform integrability of the activation vectors to be able to swap limit and expectation. This uniform integrability in the CNN case under PSEUDO-IID weights is rigorously demonstrated in proposition 2. One can compute the variance of one representative of the summands, let us say the rst one, as follows:

$$
\begin{aligned}
(\tilde{\Sigma}^{(\ell)})^2(x,x')[n] &:= \mathbb{E}\left[\tilde{Z}^{(\ell)}_1(x,x')[n]^2\right] = \mathbb{E}\Bigg[\sum_{(i,X,\alpha)\in L^\ell} \sum_{j=1}^{C_\ell} E^{(\ell)}_{i,1,\alpha}\, z^{(\ell-1)}_{1,\alpha}(X)[n]\Bigg]^2 \\
&= \sum_{(i_a,X_a,\alpha_a)\in L^\ell} \sum_{(i_b,X_b,\alpha_b)\in L^\ell} \sum_{a=1}^{C_\ell}\sum_{b=1}^{C_\ell} \mathbb{E}\left[E^{(\ell)}_{i_a,1,\alpha_a} E^{(\ell)}_{i_b,1,\alpha_b}\right] \mathbb{E}\left[z^{(\ell-1)}_{1,\alpha_a}(X_a)[n]\; z^{(\ell-1)}_{1,\alpha_b}(X_b)[n]\right] \\
&= \sum_{(i_a,X_a,\alpha_a)\in L^\ell} \sum_{(i_b,X_b,\alpha_b)\in L^\ell} \sum_{a=1}^{C_\ell}\sum_{b=1}^{C_\ell} \mathbb{E}\left[E^{(\ell)}_{i_a,1,\alpha_a} E^{(\ell)}_{i_b,1,\alpha_b}\right] \mathbb{E}\left[z^{(\ell-1)}_{1,\alpha_a}(X_a)[n]\; z^{(\ell-1)}_{1,\alpha_b}(X_b)[n]\right] \\
&= \sum_{(i_a,X_a,\alpha_a)\in L^\ell} \sum_{(i_b,X_b,\alpha_b)\in L^\ell} \sum_{a=1}^{C_\ell}\sum_{b=1}^{C_\ell} \delta_{i_a,i_b}\delta_{\alpha_a,\alpha_b}\, \mathbb{E}\left[z^{(\ell-1)}_{1,\alpha_a}(X_a)[n]\; z^{(\ell-1)}_{1,\alpha_b}(X_b)[n]\right].
\end{aligned}
$$

Given the uniform integrability of the product the activations in a CNN, we may swap the order of taking the limit and the expectation to have

$$
\begin{aligned}
(\tilde{\Sigma}^{(\ell)})^2(x,x')[\infty] &:= \lim_{n\to\infty} (\tilde{\Sigma}^{(\ell)})^2(x,x')[n] \\
&= \sum_{(i_a,X_a,\alpha_a)\in L^\ell} \sum_{(i_b,X_b,\alpha_b)\in L^\ell} \sum_{a=1}^{C_\ell}\sum_{b=1}^{C_\ell} \delta_{i_a,i_b}\delta_{\alpha_a,\alpha_b}\, \mathbb{E}\left[z^{(\ell-1)}_{1,\alpha_a}(X_a)[\infty]\; z^{(\ell-1)}_{1,\alpha_b}(X_b)[\infty]\right].
\end{aligned}
$$

As in the fully-connected case, we conclude that the feature maps at the next layer converge to Gaussians processes, whose covariance function is given by

$$
\mathbb{E}\left[h^{(\ell)}_{i_a,\alpha_a}(X_a)[\infty]\; h^{(\ell)}_{i_b,\alpha_b}(X_b)[\infty]\right] = \delta_{i_a,i_b}\sigma_b^2 + \sigma_W^2 \sum_{(a,b)\in\{1,\dots,C_\ell\}^2} K^{(\ell)}(X_a,X_b)\,,
$$

where

$$
K^{(\ell)}(X_a,X_b) = \begin{cases} \mathbb{E}_{(u,v)\sim N(0,K^{(\ell-1)}(X_a,X_b))}[\phi(u)\phi(v)], & \ell \geq 1 \\ \frac{1}{C_0}\sum_{j=1}^{C_0} (X_a)_{j,\alpha}(X_b)_{j,\alpha}, & \ell = 0 \end{cases}.
$$

## C LEMMAS USED IN THE PROOF OF THEOREMS 1 AND 2

We will present in this section the lemmas used in the derivation of our results. The proofs are omitted in cases where they can easily be found in the literature.

**Lemma 1** (Hölder's inequality) For a probability space $(\Omega, \mathcal{F}, P)$, let $X, Y$ be two random variables on $\Omega$ and $p, q > 1$ such that $p^{-1} + q^{-1} = 1$. Then,

$$
\mathbb{E}|XY| \leq \left(\mathbb{E}|X|^p\right)^{\frac{1}{p}} \left(\mathbb{E}|Y|^q\right)^{\frac{1}{q}}.
$$

Lemma 2 (Bellingsley's theorem) Let $X[n]$ be a sequence of random variables, for $n \in \mathbb{N}$ and $X$ another random variable. If $X[n]$ is uniformly integrable and the sequence $X[n]$ converges in distribution to $X[\,] = X$, then $X$ is integrable and one can swap the limit and the expectation, i.e.

$$\lim_{n \to \infty} \mathbb{E}X[n] = \mathbb{E}X[\,].$$

We adapt Lemma 18 from Matthews et al. (2018) to our setting and the proof can trivially be obtained from the proof derived in the cited article.

Lemma 3 (Sufficient condition to uniformly bound the expectation of a four-cross product) Let $X_1$, $X_2$, $X_3$, and $X_4$ be random variables on $\mathbb{R}$ with the usual Borel $\sigma$-algebra. Assume that $\mathbb{E}|X_i|^4 < \infty$ for all $i \in \{1, 2, 3, 4\}$. Then, for any choice of $p_i \in \{0, 1, 2\}$ (where $i \in \{1, 2, 3, 4\}$), the expectations $\mathbb{E}\prod_{i=1}^{4} |X_i|^{p_i}$ are uniformly bounded by a polynomial in the eighth moments $\mathbb{E}|X_i|^8 < \infty$.

This Lemma can be easily derived from standard Pearson correlation bounds.

We now have the tools to show that all four-cross products of the activations are uniformly integrable in the PSEUDO-IID setting, which enables us to swap the limit and the expectation in the previous sections of the appendix.

Proposition 1 (Uniform integrability in the PSEUDO-IID regime – fully-connected networks) Consider a fully-connected neural network in the PSEUDO-IID regime. Consider the random activations $z_i^{(\ell)}(x_a)[n], z_j^{(\ell)}(x_b)[n], z_k^{(\ell)}(x_c)[n], z_l^{(\ell)}(x_d)[n]$ with any $i, j, k, l \in \mathbb{N}, x_a, x_b, x_c, x_d \in \mathcal{X}$, neither necessarily distinct, as in 3. Then, the family of random variables

$$z_i^{(\ell)}(x_a)[n] z_j^{(\ell)}(x_b)[n] z_k^{(\ell)}(x_c)[n] z_l^{(\ell)}(x_d)[n],$$

indexed by $n$ is uniformly integrable for any $\ell = 1, \ldots, L+1$.

Proof. The proof is adapted from Matthews et al. (2018) to the PSEUDO-IID regime in fully-connected neural networks described in equation 3.

If a collection of random variables is uniformly $L^p$-bounded for $p > 1$, then it is uniformly integrable. So, we will show that our family of random variables is uniformly $L^{1+\varepsilon}$ bounded for some $\varepsilon > 0$, i.e. there exists $K < \infty$ independent of $n$ such that,

$$\mathbb{E}\left| z_i^{(\ell)}(x_a)[n] z_j^{(\ell)}(x_b)[n] z_k^{(\ell)}(x_c)[n] z_l^{(\ell)}(x_d)[n] \right|^{1+\varepsilon} \leq K,$$

which is equivalent to

$$\mathbb{E}\left[ \left| z_i^{(\ell)}(x_a)[n] \right|^{1+\varepsilon} \left| z_j^{(\ell)}(x_b)[n] \right|^{1+\varepsilon} \left| z_k^{(\ell)}(x_c)[n] \right|^{1+\varepsilon} \left| z_l^{(\ell)}(x_d)[n] \right|^{1+\varepsilon} \right] \leq K.$$

To do so, Lemma 3 gives us a sufficient condition: bounding the moment of order $4(1+\varepsilon)$ of each term in the product by a constant independent of $n$. For any $x_t \in \mathcal{X}$ and $i \in \mathbb{N}$, this moment can be rewritten in terms of the feature maps using the linear envelope property 3 and the convexity of the map $x \mapsto x^{4(1+\varepsilon)}$ as

$$\mathbb{E}\left[ \left| z_i^{(\ell)}(x_t)[n] \right|^{4(1+\varepsilon)} \right] \leq 2^{4(1+\varepsilon)-1} \mathbb{E}\left[ c^{4(1+\varepsilon)} + M^{4(1+\varepsilon)} \left| h_i^{(\ell)}(x_t)[n] \right|^{4(1+\varepsilon)} \right].$$

Thus it is sufficient to show that the absolute feature maps $|h_i^{(\ell)}(x_t)[n]|$ have a finite moment of order $4 + \varepsilon$, independent of $x_t$, $i$, and $n$, for some $\varepsilon$. For the sake of simplicity, we will show this by induction on $\ell$ for $\varepsilon = 1$.

Base case. For $\ell = 1$, the feature maps $h_i^{(1)} = \sum_{j=1}^{N_0} W_{i,j}^{(1)} x_j + b_i^{(1)}$ are identically distributed for all $i \in \{1, \ldots, N_1[n]\}$ from the row-exchangeability of the weights. Moreover, from the moment condition of Definition 1, there exists $p = 8$ such that

$$\mathbb{E}\left| h_i^{(1)}(x)[n] \right|^p \leq \mathbb{E}\left| \sum_{j=1}^{N_0} x_j W_{i,j}^{(1)} \right|^p + \mathbb{E}\left| b_i^{(1)} \right|^p$$

$$= K \|x\|_2^p N_0^{p/2} + \mathbb{E}\left| b_i^{(1)} \right|^p.$$

The RHS of the above equality is independent of $N_1$ (and $n$). Therefore, for all $n \in \mathbb{N}$ and $i \in \{1, \ldots, N_1[n]\}$, we have found a constant bound for the feature map's moment of order $p = 8$.

Inductive step. Let us assume that for any $\{x_t\}_{t=1}^4$ and $i \in \mathbb{N}$, the eigth-order moment of $h_i^{(\ell-1)}(x_t)[n]$ is bounded by a constant independent from $n$.

We will show that this implies

$$\mathbb{E}\left[h_i^{(\ell)}(x_t)[n]\right]^8 < 1 :$$

Considering the vector of activations $h^{(\ell-1)}(x_t)[n]$, we have, from the moment condition (iii) of the PSEUDO-IID regime the following conditional expectation,

$$\mathbb{E}\left[\sum_{j=1}^{N_{\ell-1}[n]} W_{ij}^{(\ell)}\left[h_j^{(\ell-1)}(x_t)[n]\right]^8 \,\middle|\, h^{(\ell-1)}(x_t)[n]\right] = K\,\|h^{(\ell-1)}(x_t)[n]\|_2^8 N_{\ell-1}[n]^{\frac{8}{2}}.$$

Thus using the recursion formulae for the data propagation in such architecture given by 3, the convexity of $x \mapsto x^8$ on $\mathbb{R}^+$ and taking conditional expectations,

$$\begin{aligned}
\mathbb{E}\left[h_i^{(\ell)}(x_t)[n]\right]^8 &\leq 2^{8-1}\mathbb{E}\left[b_i^{(\ell)8} + \left[\sum_{j=1}^{N_{\ell-1}[n]} W_{ij}^{(\ell)}\left[h_j^{(\ell-1)}(x_t)[n]\right]^8\right]^8\right] && (28)\\
&= 2^{8-1}\mathbb{E}\left[b_i^{(\ell)8}\right] + \mathbb{E}\left[\mathbb{E}\left[\sum_{j=1}^{N_{\ell-1}[n]} W_{ij}^{(\ell)}\left[h_j^{(\ell-1)}(x_t)[n]\right]^8 \,\middle|\, h^{(\ell-1)}(x_t)[n]\right]^8\right]\\
&= 2^{8-1}\mathbb{E}\left[b_i^{(\ell)8}\right] + KN_{\ell-1}[n]^4\mathbb{E}\left[\|h^{(\ell-1)}(x_t)[n]\|_2^8\right]. && (29)
\end{aligned}$$

As the biases are IID Gaussians, the first expectation involving the bias is trivially finite and does not depend on $i$ as they are identically distributed. For the second expectation, we compute,

$$\begin{aligned}
\mathbb{E}\|h^{(\ell-1)}(x_t)[n]\|_2^8 &= \mathbb{E}\left[\sum_{j=1}^{N_{\ell-1}[n]} h_j^{(\ell-1)}(x_t)[n]^2\right]^{\frac{8}{2}} && (30)\\
&\mathbb{E}\left[\sum_{j=1}^{N_{\ell-1}[n]} c + M\,h_j^{(\ell-1)}(x_t)[n]^2\right]^{\frac{8}{2}}\\
&= \mathbb{E}\left[\sum_{j=1}^{N_{\ell-1}[n]} c^2 + M^2\,h_j^{(\ell-1)}(x_t)^2[n] + 2cM\,h_j^{(\ell-1)}(x_t)[n]\right]^4.
\end{aligned}$$

This last expression can be written as a linear combination of $N_{\ell-1}[n]^4$ quantities of the form

$$\mathbb{E}\left[h_{j_1}^{(\ell-1)}(x_t)^{p_1}[n]\; h_{j_2}^{(\ell-1)}(x_t)^{p_2}[n]\; h_{j_3}^{(\ell-1)}(x_t)^{p_3}[n]\; h_{j_4}^{(\ell-1)}(x_t)^{p_4}[n]\right];$$

and we bound each of them making use of Lemma 3. The factor $N_{\ell-1}[n]^4$ in 29 thus cancels out with the number of terms in the sum 30. As the pre-activations are exchangeable, they are identically distributed so the dependence on the neuron index $i$ can be ignored, and taking the supremum over the finite set of input data $\{x_t\}$ does not affect the uniformity of the bound; which concludes the proof.

□

Proposition 2 (Uniform integrability in the PSEUDO-IID regime – CNN.) Consider a convolutional neural network in the PSEUDO-IID regime. Consider a collection of random variables

$z_i^{(\ell)}(x_a)[n], z_j^{(\ell)}(x_b)[n], z_k^{(\ell)}(x_c)[n], z_l^{(\ell)}(x_d)[n]$ with any $i, j, k, l \in \mathbb{N}, X_a, X_b, X_c, X_d \in \mathcal{X}$, neither necessarily distinct, obtained by the recursion 9. Then, the family of random variables

$$z_i^{(\ell)}(X_a)[n] z_j^{(\ell)}(X_b)[n] z_k^{(\ell)}(X_c)[n] z_l^{(\ell)}(X_d)[n],$$

indexed by $n$ is uniformly integrable for any $\ell \in \{1, \ldots, L+1\}$.

Proof. As previously, this proposition holds some novelty of this paper, extending already known proofs in the standard Gaussian setting to the PSEUDO-IID regime in CNNs. Note that this directly implies the universality of the Gaussian Process behaviour for CNNs in this regime, which has been established so far only by Yang (2021) to the best of our knowledge. Our result goes beyond. We recall that the data propagation is described by Equation 9.

Once again, we observe it is sufficient to show such that the moment of order 8 of the feature maps is uniformly bounded. We do this again by induction.

Base case. Since $h_{i,}^{(1)}(X)[n] = b_i^{(1)} + \sum_{j=1}^{C_0} \sum_{u \in \mathcal{J}\, \mathcal{K}} U_{i,j,}^{(1)}\, x_{j,}$, and weights and biases are independent, for some $p \geq 8$ we can write

$$\begin{aligned} \mathbb{E}\, h_{i,}^{(1)}(X)[n]^{\,p} &\leq \mathbb{E} \Big| \sum_{j=1}^{C_0} \sum_{u \in \mathcal{J}\, \mathcal{K}} x_{j,}\, U_{i,j,}^{(1)} \Big|^{p} + \mathbb{E}\, b_i^{(1)\,p} \\ &= K_p \|X_{\mathcal{J}\, \mathcal{K}}\|_2^p C_0^{\,p/2} + \mathbb{E}\, b_i^{(1)\,p}, \end{aligned}$$

where $X_{\mathcal{J}\, \mathcal{K}}$ is the part of signal around the pixel and we have used the condition (iii) of Definition 4 in the last line. Observe that the RHS is independent of $n$.

Inductive step. Let us assume that for any $\{X_t\}_{t=1}^4$, $\ell \in \{1, \ldots, L\}$ and $i \in \mathbb{N}$, there exists $c_0 \in (0, 1)$ such that the eighth moment of the pre-activations from the previous layer $h_{i,}^{(\ell-1)}(X_t)[n]$ is bounded by a constant independent from $i \in \mathbb{N}$, $\ell \in \{1, \ldots, L\}$, and $n$, for all $X_t \in \mathcal{X}$.

Mirroring our proof in the fully-connected case, we will show that this propagates to the next layer,

$$\mathbb{E}\, h_{i,}^{(\ell)}(X_t)[n]^{\,8} < 1.$$

From the third condition of the PSEUDO-IID regime, we can compute the expectation conditioned on the vector of activations $h^{(\ell-1)}(X_t)[n]$,

$$\begin{aligned} \mathbb{E}\Big[ \Big| \sum_{j=1}^{C_{\ell-1}[n]} \sum_{u \in \mathcal{J}\, \mathcal{K}} U_{i,j,}^{(\ell)}\, h_{j,}^{(\ell-1)}(X_t)[n] \Big|^{8} \Big] &= \mathbb{E}\, \mathbb{E}\Big[ \Big| \sum_{j=1}^{C_{\ell-1}[n]} \sum_{u \in \mathcal{J}\, \mathcal{K}} U_{i,j,}^{(\ell)}\, h_{j,}^{(\ell-1)}(X_t)[n] \Big|^{8} \Big| \\ &\qquad \Big| h^{(\ell-1)}(X_t)[n] \Big] \\ &= K C_{\ell-1}[n]^{\,4} \mathbb{E} \big\| h^{(\ell-1)}(X_t)[n] \big\|_2^8, \end{aligned}$$

where the norm of the activations can be computed using the linear envelope property,

$$\begin{aligned} \mathbb{E} \big\| h^{(\ell-1)}(X_t)[n] \big\|_2^8 &\leq \mathbb{E} \Big[ \sum_{j=1}^{C_{\ell-1}[n]} \sum_{u \in \mathcal{J}\, \mathcal{K}} h_{j,}^{(\ell-1)}(X_t)[n] \Big]^{4} \\ &\leq \mathbb{E} \Big[ \sum_{j=1}^{C_{\ell-1}[n]} \sum_{u \in \mathcal{J}\, \mathcal{K}} c^2 + M^2 h_{j,}^{(\ell-1)}(X_t)^2[n] + 2cM h_{j,}^{(\ell-1)}(X_t) \Big]^{4}. \end{aligned}$$

This last quantity turns out to be the weighted sum (over $k$ of $C_{\ell-1}[n])^4$ terms (recall $k$ being the filter size, which is finite) of the form

$$E\left[h^{(\ell-1)}_{j_1,i}(X_t)^{p_1}[n]h^{(\ell-1)}_{j_2,i}(X_t)^{p_2}[n]h^{(\ell-1)}_{j_3,i}(X_t)^{p_3}[n]h^{(\ell-1)}_{j_4,i}(X_t)^{p_4}[n]\right],$$

that can be bounded using Lemma 3 combined with our inductive hypothesis. Observe how the factors $C_{\ell-1}[n]^4$ cancel out and lead to a bound independent from $n$.

Using the recursion formulae for the data propagation in the CNN architecture recalled in 9 and the convexity of the map $x \mapsto x^8$ on $\mathbb{R}^+$, we have

$$E\left[h^{(\ell)}_{i,i}(X_t)[n]\right]^8 \leq 2^{8-1}E\left[b^{(\ell)}_i\right]^8 + C_{\ell-1}[n]\sum_{j=1}^{n_{\ell-1}}\sum_{2\leq|J|\leq K} U^{(\ell)}_{i;J}\left[h^{(\ell-1)}_{j,i}(X_t)\right]^8[n]\Big].$$

The first expectation is finite and uniformly bounded as the biases are Gaussians in the PSEUDO-IID regime and we have just shown above the boundedness of the second expectation.

Therefore,

$$E\left[h^{(\ell-1)}_{j,i}(X_t)\right]^8[n] < \infty, \tag{31}$$

and taking the supremum over the finite input data set does not change the uniformly bounded property, which was needed to be shown.

□

## D EXAMPLES CONCERNING THE BOUNDED MOMENT CONDITION (III) OF THE PSEUDO-IID DISTRIBUTION

The bounded moment condition (iii) of the PSEUDO-IID distribution in Definition 1 is a key condition in the distinct proof of IID matrices taken in Hanin (2021) (Lemma 2.9). We show some examples of distributions in Figure 4 which verify or violate the conditions we identified as sufficient to rigorously prove the convergence of random neural networks to gaussian processes in the large width limit.

## E STRUCTURED SPARSE WEIGHT MATRICES IN THE FULLY-CONNECTED SETTING

Fig. 5 shows examples of permuted block-sparse weight matrices used to initialize a fully-connected network in order to produce the plots given in Fig. 2-7.

## F FURTHER NUMERICAL SIMULATIONS VALIDATING THEOREM 1

Fig. 6 shows $Q-Q$ plots for the histograms in Fig. 2 as compared to their infinite with Gaussian limit. These $Q-Q$ plots show how the PSEUDO-IID networks approach the Gaussian process at somewhat different rates with IID uniform approaching the fastest.

Fig. 7 explores the growing independence of entries $h^{(\ell)}_i(x)[n]$ for different $i$ by showing their joint distributions for two distinct choices of $i$, moreover, convergence to the limiting isotropic Gaussian distribution $h^{(\ell)}_i(x)[\,]$ is overlayed in the same plots. Uniform IID converges the quickest, while PSEUDO-IID Gaussian low-rank and structured sparse converge towards an isotropic distribution somewhat slower, albeit already showing good agreement at $n = 30$. The horizontal and vertical axis in each subplot of Fig. 7 are $h^{(5)}_i(x)$ for $i = 1$ and 2 respectively.

Figure 4: Different cases where conditions (ii) and (iii) of the pseudo IID regime are either satisfied or violated. Condition (iii) is considered with $n = 8$ and $a = (1; \ldots; 1)$ such that it becomes $E\left[\left(\sum_{j=1}^{n} X_j\right)^8\right] = K$, where $X = (X_1; \ldots; X_N)$ is regarded as one row of the weight matrix. The chosen distribution for the vector $X$ impacts whether the network is in the PSEUDO-IID regime. In the identical coordinates case, $X = (X_1; \ldots; X_1)$ is the concatenation of the same realisation $X_1$ sampled from a standard normal. Not only the traditional assumption is broken as the coordinates are obviously dependent but also condition (iii) is violated, thus resulting in an unbounded expectation when growing the dimension $N$. The autoregressive process $(X_1; \ldots; X_N)$ shown is obtained by $X_i = \varepsilon_i + X_{i-1}$, where $\varepsilon_i$ are IID multivariate gaussian noises of dimension $N$. The correlation between the coordinates does not decrease fast enough with the dimension to get a bound on the computed expectation and condition (iii) is once again violated. On the contrary, from the plot produced by sampling IID Cauchy distributions, it is not obvious whether condition (iii) holds. Nonetheless, condition (ii) which ensures the finiteness of the variance is not, thus a random network initialized with IID Cauchy weights falls outside the scope of our identified broad class of distributions to ensure a convergence towards a gaussian process. The last cases of samples taken either from scaled multivariate normals in red or uniformly sampled from the unit sphere in orange (with appropriate scaling such that condition (ii) holds) show that there exists a bound independent from the dimension on the expectation of interest. The empirical expectations are taken considering averages over 100000 samples.

width = 3 width = 30 width = 300

Figure 5: Example of a permuted block-sparse weight matrix at initialization of a fully-connected network with increasing width $N$. The matrix is initialized with identically and independently sampled diagonal blocks from a scaled Gaussian. Its rows and columns are then randomly permuted in order to satisfy the PSEUDO-IID conditions. The block size is set to be $2\sqrt{N}$. Entries in yellow are zero and entries in black are nonzero..

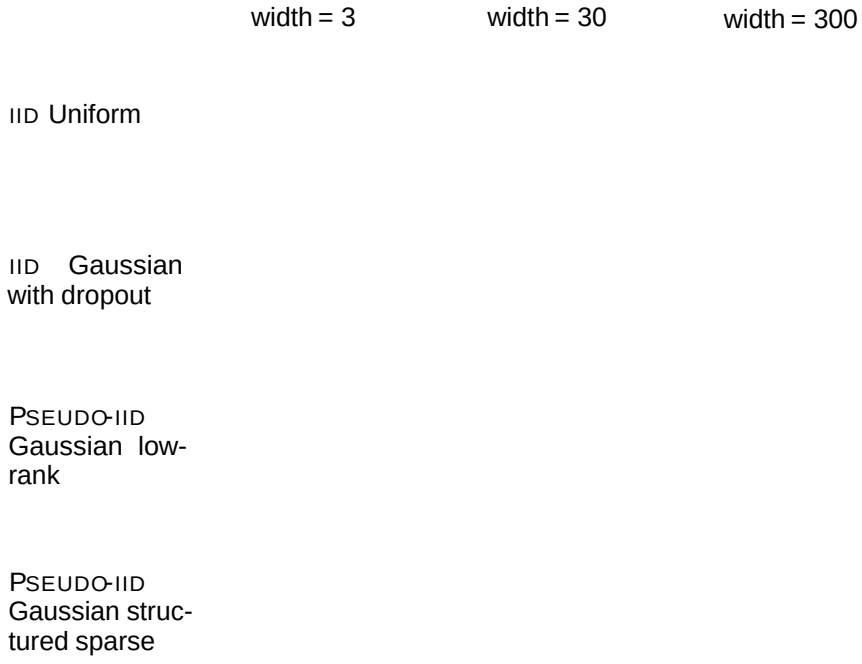


Figure 6: Q-Q plots of the pre-activations values in Fig. 2 as an alternative way of showing the convergence of the pre-activation of a fully-connected network to a Gaussian as fully characterized in Theorem 1. The settings of the experiment are the same as those in Fig. 2.