Free probability as a solution to the problem of tuning neural networks

Anonymous Author(s)
Affiliation
Address
email

Abstract

Gradient descent during the learning process of a neural network can be subject to many instabilities. The spectral density of the Jacobian is a key component for analyzing stability. Following the works of Pennington et al., such Jacobians are modeled using free multiplicative convolutions from Free Probability Theory (FPT). We present a reliable and very fast method for computing the associated spectral densities, for given architecture and initialization. This method has a controlled and proven convergence. Our technique is based on an homotopy method: it is an adaptative Newton-Raphson scheme which chains basins of attraction. In order to demonstrate the relevance of our method we show that certain FPT metrics computed before training are highly correlated to final test accuracies – up to 85%. This nuances the idea that learning happens at the edge of chaos by giving evidence that a very desirable feature for neural networks is the hyperbolicity of their Jacobian at initialization.

1 Introduction

Neural network training and tuning can be wasteful in human and energy resources. For example [21, Table 1] shows that a single GPU training can have a large carbon footprint. Often, this is nothing compared to architecture search. In this context, an obvious moonshot would be estimating the performance of an architecture *before* training. This paper takes a more realistic and indirect approach. We aim for a fast and reliable computational method for estimating stability before training – and stability is a good proxy for performance.

Framework: Consider a feed-forward network of depth $L \in \mathbb{N}$, with L full-connected layers. For each depth $\ell \in \{1, 2, \ldots, L\}$, the layer has activation vector $x_\ell \in \mathbb{R}^{N_\ell}$, where N_ℓ is the current width. The vector $x_0 \in \mathbb{R}^{N_0}$ takes in the neural network's input, while $x_L \in \mathbb{R}^{N_L}$ gives the output. The vector of widths is written $\mathbf{N} := (N_0, N_1, \ldots, N_L)$ and will appear in superscript to indicate the dependence in any of the N_ℓ's. The following recurrence relation holds between layers $x_\ell = \phi_\ell \left(W_\ell^{(\mathbf{N})} x_{\ell-1} + b_\ell^{(\mathbf{N})} \right)$, where ϕ_ℓ is a choice of non-linearity applied entry-wise, $W_\ell^{(\mathbf{N})} \in M_{N_\ell, N_{\ell-1}}(\mathbb{R})$ is a weight matrix and $b_\ell^{(\mathbf{N})} \in \mathbb{R}^{N_\ell}$ is the vector of biases. We write $h_\ell := W_\ell^{(\mathbf{N})} x_{\ell-1} + b_\ell^{(\mathbf{N})}$ for the pre-activations.

The Jacobian computed during back-propagation can be written explicitly by using the chain rule:

$$J^{(\mathbf{N})} := \frac{\partial x_L}{\partial x_0} = \frac{\partial x_L}{\partial x_{L-1}} \frac{\partial x_{L-1}}{\partial x_{L-2}} \cdots \frac{\partial x_1}{\partial x_0} = D_L^{(\mathbf{N})} W_L^{(\mathbf{N})} D_{L-1}^{(\mathbf{N})} W_{L-1}^{(\mathbf{N})} \ldots D_1^{(\mathbf{N})} W_1^{(\mathbf{N})} , \quad (1)$$

where D_ℓ's are the diagonal matrices given by

$$\left[D_\ell^{(\mathbf{N})} \right]_{i,i} = \phi'_\ell \left([h_\ell]_i \right) . \quad (2)$$

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

Technically, a step of gradient descent updates weights and biases following

$$\left(W_\ell^{(\mathbf{N})}, b_\ell^{(\mathbf{N})}\right) \leftarrow \left(W_\ell^{(\mathbf{N})}, b_\ell^{(\mathbf{N})}\right) - \alpha \frac{\partial \mathcal{L}}{\partial (W_\ell^{(\mathbf{N})}, b_\ell^{(\mathbf{N})})} \ , \tag{3}$$

for each $\ell = 1, \ldots, L$. Here $\alpha > 0$ is the learning rate and $\mathcal{L}$ is the loss on a minibatch. If the minibatch has size $B \in \mathbb{N}$, and corresponds a small sample $((X_i, Y_i) \ ; \ i = 1, \ldots, B)$ of the dataset, we have $\mathcal{L} = \frac{1}{B} \sum_{i=1}^{B} d(x_L(X_i), Y_i)$. Here d is a real-valued distance or similarity function, the X_i's are the input vectors while the Y_i's are the output vectors (e.g. labels in the case of classifier, $Y_i \approx X_i$ in the case of an autoencoder etc...).

The chain rule dictates:

$$\frac{\partial \mathcal{L}}{\partial (W_\ell^{(\mathbf{N})}, b_\ell^{(\mathbf{N})})} = \frac{\partial \mathcal{L}}{\partial x_L} \frac{\partial x_L}{\partial x_{L-1}} \cdots \frac{\partial x_{\ell+1}}{\partial x_\ell} \frac{\partial x_\ell}{\partial (W_\ell^{(\mathbf{N})}, b_\ell^{(\mathbf{N})})} = \frac{\partial \mathcal{L}}{\partial x_L} J_\ell^{(\mathbf{N})} \frac{\partial h_\ell}{\partial (W_\ell^{(\mathbf{N})}, b_\ell^{(\mathbf{N})})} \ , \tag{4}$$

where

$$\frac{\partial \mathcal{L}}{\partial x_L} = \frac{1}{B} \sum_{i=1}^{B} \partial_1 d(x_L(X_i), Y_i) \in M_{1, N_L}(\mathbb{R}) \ , \tag{5}$$

$$J_\ell^{(\mathbf{N})} = D_L^{(\mathbf{N})} W_L^{(\mathbf{N})} \ldots D_{\ell+1}^{(\mathbf{N})} W_{\ell+1}^{(\mathbf{N})} D_\ell^{(\mathbf{N})} \in M_{N_L, N_\ell}(\mathbb{R}) \ . \tag{6}$$

Therefore, for the sake of simplicity, we shall focus on the Jacobian $J^{(\mathbf{N})}$ given in Eq. (1) since it has exactly the same form as the $J_\ell^{(\mathbf{N})}$ given in Eq. (6). The issue is that a large product of (even larger) matrices can easily become unstable. If many singular values are $\ll 1$, we have gradient vanishing. If many singular values are $\gg 1$, we have gradient explosion. Such a transition can be referred to as the edge of chaos.

Intuition. This instability is easily understood thanks to the naive analogy with the one-dimensional case. Indeed, the geometric progression q^n with $n \to \infty$ is the archetype of a long product and it converges extremely fast, to either 0 if $|q| < 0$ or to ∞ if $|q| > 1$.

A less naive intuition consists in recalling that mini-batch sampling in Eq. (5) is noise source in Stochastic Gradient Descent (SGD). It is fair to assume that $\frac{\partial \mathcal{L}}{\partial x_L}$ has a Gaussian behavior with covariance proportional to I_{N_L} – either because of the Central Limit Theorem if B is large enough or after time averaging, because of the mixing properties of SGD. Therefore, each gradient step $\alpha \frac{\partial \mathcal{L}}{\partial (W_\ell^{(\mathbf{N})}, b_\ell^{(\mathbf{N})})}$ in Eq. (3) should approximately be a Gaussian vector with covariance proportional to:

$$\alpha^2 \left(\frac{\partial h_\ell}{\partial (W_\ell^{(\mathbf{N})}, b_\ell^{(\mathbf{N})})} \right)^T \left(J_\ell^{(\mathbf{N})} \right)^T J_\ell^{(\mathbf{N})} \frac{\partial h_\ell}{\partial (W_\ell^{(\mathbf{N})}, b_\ell^{(\mathbf{N})})} \ .$$

Simplifying further, we see the importance of the spectrum of $\left(J_\ell^{(\mathbf{N})} \right)^T J_\ell^{(\mathbf{N})}$ for stability. Basically, eigenvectors of $\left(J_\ell^{(\mathbf{N})} \right)^T J_\ell^{(\mathbf{N})}$ are the directions along which the one-dimensional intuition applies.

Randomness. Starting from the pioneering works of Glorot and Bengio [8] on random initializations, it was suggested that the spectral properties of $J^{(\mathbf{N})}$ are an excellent indicator for stability and learning performance. In particular, an appropriate random initialization was suggested and since implemented in all modern ML frameworks [17, 1].

We make classical choices of random initializations. The biases $b_\ell^{(\mathbf{N})}$ are taken as random vectors which entries are centered i.i.d. Gaussian random variables with standard deviation σ_{b_ℓ}. For the weights, we will consider the following matrix ensembles: the $[W_\ell^{(\mathbf{N})}]_{i,j}$ are drawn from i.i.d. centered random variables with variance $\sigma_{W_\ell}^2 / N_\ell$ and finite fourth moment as in [15].

Modeling spectrum thanks to Free Probability Theory (FPT). Already, notice that the square root normalization suggested in [8] is the correct scale for the convergence of spectral measures in

FPT. This ought to be recognized and provides a new justification for that choice. Now, following the works of Pennington et al. [18], the tools of FPT can be used to more quantitatively analyze the singular values of $J^{(\mathbf{N})}$ in the large width limit. The large width limit is particularly attractive when studying large deep networks, especially because free probability appears at relatively small sizes thanks to strong concentration properties [12]. Indeed, the theory applies to random matrices of size as small as a hundred.

For the purposes of this paragraph, we restrict ourselves to square matrices and assume $N_\ell = N$ for all $\ell = 1, \dots, L$. In fact, FPT is concerned with the behavior of spectral measures as $N \to \infty$. For any diagonalizable $A_N \in M_N(\mathbb{R})$, the associated spectral measure on the real line is:

$$\mu_{A^{(N)}}(dx) := \frac{1}{N} \sum_{i=1}^{N} \delta_{a_i^{(N)}}(dx)$$

with the $a_i^{(N)}$'s being the eigenvalues of A_N. For ease of notation, the spectrum of (squared) singular values is written $\nu_{A^{(N)}} := \mu_{\left(A^{(N)}\right)^T A^{(N)}}$. A fundamental assumption for invoking tools from Free Probability Theory, is the assumption of *asymptotic freeness*. Without defining the notion, which can be found in [14], let us describe the important computation it allows, discovered in the seminal work of Voiculescu [23]. Given two sequences of square matrices $A^{(N)}$, $B^{(N)}$ in $M_N(\mathbb{R})$, with converging spectral measures:

$$\lim_{N\to\infty} \nu_{A^{(N)}} = \nu_A \ , \quad \lim_{N\to\infty} \nu_{B^{(N)}} = \nu_B \ ,$$

we have that, under the assumption of asymptotic freeness $\lim_{N\to\infty} \nu_{A^{(N)}B^{(N)}} = \nu_A \boxtimes \nu_B$, where $\boxtimes$ is a deterministic operation between measures called multiplicative free convolution. The $\boxtimes$ will be detailed in Section 2. The letter A (as well as B) does not correspond to a limiting matrix but to an abstract operator, with associated spectral measure μ_A and measure of squared singular values ν_A. For such limiting operators, we drop the superscript (N).

Under suitable assumptions which are motivated and detailed later following the works of [18, 10, 16, 15, 7], for all $\ell = 1, \dots, L$, the measures $\nu_{W_\ell^{(\mathbf{N})}}$ and $\nu_{D_\ell^{(\mathbf{N})}}$ will respectively converge to ν_{W_ℓ} and ν_{D_ℓ}. Again the W_ℓ's and D_ℓ's are abstract operators which only make sense in the infinite width regime. In the limit, asymptotic freeness will also hold. Therefore, we will see that the measure of interest is:

$$\lim_{N\to\infty} \nu_{J^{(\mathbf{N})}} = \nu_J := \nu_{D_L} \boxtimes \nu_{W_L} \boxtimes \cdots \boxtimes \nu_{D_1} \boxtimes \nu_{W_1} \ . \tag{7}$$

The goals of this paper are (1) To give a very fast and stable computation of ν_J, in the more general setup of rectangular matrices (2) Empirically demonstrate that FTP metrics computed from ν_J do correlate to the final test accuracy.

1.1 Contributions

We aim at streamlining the approach of Pennington et al. by providing the tools for a systematic use of FPT. The contributions of this paper can be categorized as:

- Theoretical: In Pennington et al., a constant width is assumed. We generalize the model to allow for varying width profiles, which is more inline with design practices. This requires us to develop a rectangular multiplicative free convolution.

  Then we propose a computational scheme for computing spectral densities, named "Newton lilypads". The method relies on adaptative inversions of S-transforms using the Newton-Raphson algorithm. If the Newton-Raphson scheme is only local, we achieve a global resolution by chaining basins of attractions, thanks to doubling strategies. As such, we have theoretical guarantees for the convergence.

  Interestingly, even in the FPT community, inverting S-transforms has been considered impossible to realize in practice. As we shall see, an S-transform is a holomorphic map with multi-valued inverse. In the words of [5, p.218], “the operations which are necessary in this method (the inversion of certain functions, the extension of certain analytic functions) are almost always impossible to realise practically.”

- Numerical: This misconception led to the use of combinatorial methods based on moments, or fixed-point algorithms via the subordination method [2, 13, 22]. In the ML community, Pennington et al. pioneered the application of FPT to the theoretical foundations of Machine Learning and did not shy away from inverting S-transforms. Their [18, Algorithm 1] is based on a generic root finding procedure, and choosing the root closest to the one found for the problem with one less layer. A major drawback of their method is that there is no guarantee of finding the correct value, unlike our chaining which always chooses the correct branch.

  Not only Newton lilypads finds the correct root, but it is also an order of magnitude faster (Fig. 1). A few standard Cython optimizations allow to gain another order of magnitude, although this can certainly be refined.

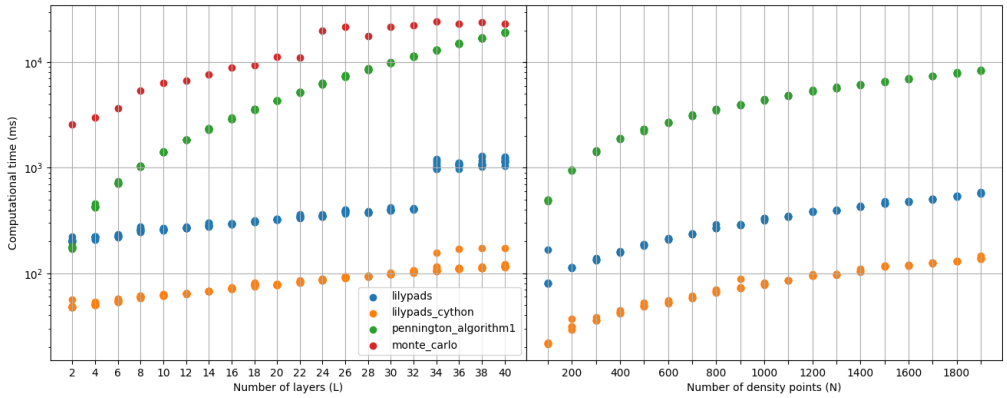
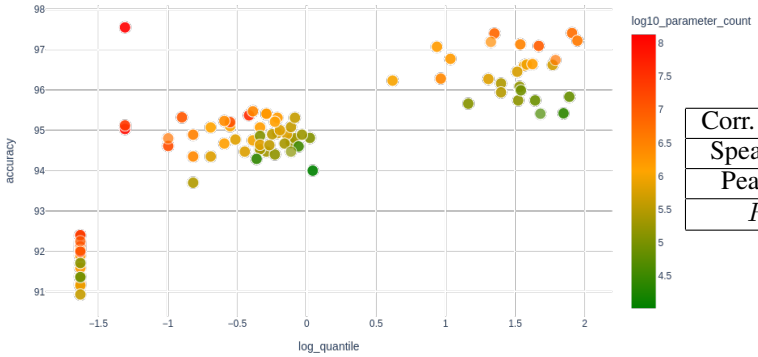


Figure 1: Computation time (in ms) for the density of ν_J w.r.t. depth L (left) and number of density points (right). Vertical axis is log-scale. The benchmarked methods are Newton lilypads in pure Python (blue), Newton lilypads with Cython optimizations (orange), Pennington et al.'s Algorithm 1 using a native root solver (green), Monte-Carlo (red). While of a different nature, Monte-Carlo is given for indication.



Corr. factor	Value	p-value
Spearman	0.83	3e-49
Pearson	0.84	9e-54
R^2	0.84	N/A

Figure 2: Scatter plot of ($\log_{10}$ of) 90th percentile of ν_J and accuracies of MLPs with randomly generated architectures. We have trained 200 MLPs on the classical MNIST dataset. The plot only shows those with $> 90\%$ accuracy. Colors indicate ($\log_{10}$ of) the number of parameters in each MLP. The p-value is based on the null hypothesis that quantile and accuracies are uncorrelated.

- Empirical: We demonstrate that FPT is not a mere abstract tool by designing a simple and unequivoqual experience. Fig. 2 shows the correlation between the accuracy of several randomly generated architectures and the quantiles of ν_J, after a fixed number of epochs.

Remarkably, we find that accuracy is very correlated to the higher quantiles. This suggests that spread-out spectral distributions ν_J are more desirable, provided of course we avoid the vanishing and explosive regimes. In the language of dynamical systems, we say that the Jacobian needs to be hyperbolic i.e. with both contracting and expanding directions. This considerably nuances the idea that learning happens at the edge of chaos. A similar point was made in the conclusion of [6], using the language of hyperbolic attractors.

Also, larger neural networks obviously perform better. But smaller networks with more spread-out ν_J perform better than larger counterparts.

For reproducibility of the numerical and empirical aspects, a complete implementation is provided in a Github repository `https://anonymous.4open.science/r/FreeNN-331E/`

1.2 Structure of the paper

We start in Section 2 by stating facts from Free Probability Theory. Most of it is available in the literature, except the setup for the product of rectangular free matrices. To the best of our knowledge, this is novel. There, we establish in the rectangular setting an analogue of Eq. (7) in Theorem 2.3.

In Section 3, we explain in detail the FPT model for random neural networks. Then thanks to the results of [16, 15] and our rectangular setting, we show that the spectral measure of the Jacobian $J^{(\mathbf{N})}$ converges to ν_J and we encode the limit in explicit generating series in Theorem 3.1. This gives how ν_J can *theoretically* be recovered.

Section 4 presents the numerical resolution which inverts the (multi-valued) generating series. By chaining different (local) basins of attractions, we obtain a global resolution method. Our algorithm is detailed in Algorithm 1 and Theorem 4.1 states the theoretical guarantees.

Finally Section 5 presents the experiment leading to Fig. 2. More details are given in Appendix F, including comments on the benchmark of Fig. 1.

2 Free Probability

2.1 Definitions and notations

Free Probability Theory provides a framework to analyze eigenvalues and singular values of large random matrices.We now introduce various complex-analytic generating series which encode the measures and the basic operations on them. First, the Cauchy-Stieltjes transform of μ, a probability measure on $\mathbb{R}_+$ is:

$$\begin{array}{rccc} G_\mu : & \mathbb{C}_+ & \to & \mathbb{C}_- \\ & z & \mapsto & \int_{\mathbb{R}_+} \frac{\mu(dv)}{z-v} \end{array},$$

where $\mathbb{C}_\pm := \{z \in \mathbb{C} \mid \pm \Im z > 0\}$. The transform G_μ encodes the measure μ and reciprocally, the measure can be recovered thanks to:

Lemma 2.1 (Cauchy-Stieltjes inversion formula – Theorem 6 in [14])**.** *We have the weak convergence of probability measures:*

$$\lim_{y \to 0} -\frac{1}{\pi} \Im G_\mu(x+iy) dx = \mu(dx) .$$

The moment generating function is

$$M_\mu(z) = zG_\mu(z) - 1 = \sum_{k=1}^{+\infty} \frac{m_k(\mu)}{z^k} , \tag{8}$$

where for all $k \in \mathbb{N}$, $m_k(\mu) := \int_{\mathbb{R}} x^k \mu(dx)$ is the k-th moment of μ. For $\mu \neq \delta_0$, M_μ is invertible in the neighborhood of ∞ and the inverse is denoted by $M_\mu^{\langle -1 \rangle}$. The S-transform of μ is defined as $S_\mu(m) = \frac{1+m}{m M_\mu^{\langle -1 \rangle}(m)}$, and is analytic in a neighborhood of $m = 0$. Furthermore, the variable z will always denote an element of $\mathbb{C}_+$, while the variables g and m will denote elements in the image of G_μ and M_μ. For a diagonalizable matrix $A^{(N)} \in M_N(\mathbb{R})$, we write $S_{A^{(N)}} := S_{\mu_{A^{(N)}}}$, $G_{A^{(N)}} := G_{\mu_{A^{(N)}}}$, $M_{A^{(N)}} := M_{\mu_{A^{(N)}}}$.

A landmark result in the field introduces free multiplicative convolution in a natural way, and shows that this operation is linearized by the S-transform:

Theorem 2.2 (Voiculescu, [23])**.** *Consider two sequences of positive matrices, each element in* $M_N(\mathbb{R})$

$$\left(A^{(N)}\,;\,N\geq 1\right)\,,\quad \left(B^{(N)}\,;\,N\geq 1\right)\,,$$

such that:

$$\lim_{N\to\infty}\mu_{A^{(N)}}=\mu_A\,,\quad \lim_{N\to\infty}\mu_{B^{(N)}}=\mu_B\,.$$

Under the assumption of asymptotic freeness for $A^{(N)}$ *and* $B^{(N)}$*, there exists a deterministic probability measure* $\mu_A\boxtimes\mu_B$ *such that* $\lim_{N\to\infty}\mu_{A^{(N)}B^{(N)}}=\mu_A\boxtimes\mu_B$*. The operation* $\boxtimes$ *is the multiplicative free convolution. Moreover*

$$S_{AB}(m)=S_A(m)S_B(m)\,. \tag{9}$$

This convergence akin to a law of large numbers is the key ingredient which allows to build the deterministic model for the back-propagation of gradients in Eq. (7).

2.2 Product of rectangular free matrices

As a generalization of Eq. (9) to rectangular matrices, we state:

Theorem 2.3. *Let* $(p_N)_{N\geq 1},(q_N)_{N\geq 1},(r_N)_{N\geq 1}$*, be three sequences of integers satisfying*

$$p_N,q_N,r_N\underset{N\to\infty}{\longrightarrow}\infty,\qquad \frac{r_N}{q_N}\underset{N\to\infty}{\longrightarrow}c>0\,.$$

Consider for all $N\geq 1$ *let* $A^{(N)}$*,* $B^{(N)}$ *be random matrices of respective sizes* $p_N\times q_N$ *and* $q_N\times r_N$ *such that the (squared) singular laws of* $A^{(N)},B^{(N)}$ *converge weakly. Assuming that* $B^{(N)}\left(B^{(N)}\right)^T$ *and* $\left(A^{(N)}\right)^T A^{(N)}$ *are asymptotically free, we have that in the limit* $N\to\infty$*:*

$$S_{(AB)^TAB}(m)=S_{A^TA}(cm)\,S_{B^TB}(m)\,.$$

Proof. See the appendix, Subsection C.1 □

Implicitly this defines a rectangular multiplicative free convolution, which could be denoted $\boxtimes_c$ in the spirit of the rectangular free additive convolution [4]. But, in the current setting, this is not a good idea. Indeed, if one defines $\mu_1\boxtimes_c\mu_2$ as the measure whose S-transform is $S_{\mu_1}(c\,\cdot)S_{\mu_2}$, then a quick computation shows that $\boxtimes_c$ is not associative i.e. for a triplet (μ_1,μ_2,μ_3) of probability measures and a pair $(c_1,c_2)\in\mathbb{R}_+^*\times\mathbb{R}_+^*$, we generically have $\mu_1\boxtimes_{c_1}(\mu_2\boxtimes_{c_2}\mu_3)\neq(\mu_1\boxtimes_{c_1}\mu_2)\boxtimes_{c_2}\mu_3$.

A better idea is to treat the dimension ratio c as part of the data via a central extension:

Definition 2.4. *On the set of pairs* (μ,c) *such that* μ *is a probability measure on* $\mathbb{R}_+$ *and* $c\in\mathbb{R}_+^*$*, define the operation* $\boxtimes$ *as:*

$$(\mu_1,c_1)\boxtimes(\mu_2,c_2):=(\nu,c_1c_2)\ ,$$

where ν *is the unique probability measure such that* $S_\nu=S_{\mu_1}(c_2\,\cdot)S_{\mu_2}$*. This extends the classical definition as the usual free convolution is recovered with* $(\mu_1,1)\boxtimes(\mu_2,1):=(\mu_1\boxtimes\mu_2,1)$*.*

Such an operation is associative and will allow a neat formulation of the measure of interest in the upcoming Theorem 3.1, entirely analogous to Eq. (7).

3 Theoretical resolution of the model

Width profile: Pennington et al. [18] consider $N_\ell=N$ for $\ell=1,2,\ldots,N$. Here, we consider that the width of layers is not constant across layers, which is mostly the case in practice. Indeed, modern architectures typically have very sophisticated topologies with layers varying in widths.

Let us assume that we are in the infinite width regime in the sense that $N_\ell\to\infty$, for all $\ell=0,1,2\ldots,L$ with: $\frac{N_{\ell-1}}{N_\ell}\underset{\mathbf{N}\to\infty}{\longrightarrow}\lambda_\ell\geq 0$. And let us denote $\Lambda_\ell:=\lim_{\mathbf{N}\to\infty}\frac{N_0}{N_\ell}=\prod_{k=1}^{\ell}\lambda_k$, with the convention $\Lambda_0=1$.

FPT limits: Let $\mathcal{N}$ be a standard Gaussian random variable on the real line $\mathbb{P}(\mathcal{N} \in dx) = \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} dx$. Here $D_\ell^{(\mathbf{N})}$ is diagonal with entries $\phi'_\ell([h_\ell]_i)$ (see Eq. (2)), and the pre-activations

$$h_\ell = W_\ell^{(\mathbf{N})} x_{\ell-1} + b_\ell^{(\mathbf{N})} = W_\ell^{(\mathbf{N})} \phi_\ell(h_{\ell-1}) + b_\ell^{(\mathbf{N})}$$

clearly depend on the previous layers. Because of this lack of independence, the standard results of FPT cannot be applied directly i.e. asymptotic freeness does not obviously hold. This is an important subtlety that is addressed in the upcoming Theorem 3.1. Based on an information propagation argument, the papers [19, 20] argue that the entries of h_ℓ behave as the i.i.d. samples of a Gaussian distribution with zero mean and variance q^ℓ. A basic law of large numbers applied to Eq. (2) gives a limit for the empirical measure $\mu_{D_\ell} = \lim_{N_\ell \to \infty} \mu_{D_\ell^{(\mathbf{N})}} = \phi'_\ell\left(\sqrt{q^\ell}\mathcal{N}\right)$. Also the recurrence for the variance is:

$$q^\ell = f_\ell\left(q^{\ell-1}\right) = \sigma_{W_\ell}^2 \mathbb{E}\left[\phi_\ell\left(\sqrt{q^{\ell-1}}\mathcal{N}\right)^2\right] + \sigma_{b_\ell}^2, \tag{10}$$

with initial condition $q^1 = \frac{\sigma_{W_1}^2}{N_1} \sum_{i=1}^{N_1} (x_0^i)^2 + \sigma_{b_1}^2$.

Recently Pastur et al. completed this heuristic thanks to a swapping trick – see [16, Lemma 3.3] and [15, Remark 3.4]. They proved that, regarding the asymptotical spectral properties of $J^{(\mathbf{N})}$, one can replace each $D_\ell^{(\mathbf{N})}$ by a diagonal matrix with independent Gaussian entries $\sqrt{q^\ell}\mathcal{N}$ independent from the rest. By combining that result and our rectangular setup from Section 2, we find:

Theorem 3.1. *In terms of the rectangular multiplicative free convolution, the measure of (squared) singular values of $J^{(\mathbf{N})}$ converges to*

$$\nu_J = (\nu_{D_L}, 1) \boxtimes (\nu_{W_L}, \lambda_L) \boxtimes \cdots \boxtimes (\nu_{D_1}, 1) \boxtimes (\nu_{W_1}, \lambda_1) \ . \tag{11}$$

Moreover, the S-transform of $J^T J$ in the infinite width regime verifies

$$S_{J^T J}(m) = \prod_{\ell=1}^{L} \left[S_{D_\ell^2}(\Lambda_\ell m)\, S_{W_\ell^T W_\ell}(\Lambda_{\ell-1} m)\right] \ . \tag{12}$$

In particular, under the assumption that the entries of W_ℓ are i.i.d. :

$$S_{J^T J}(m) = \prod_{\ell=1}^{L} \left(S_{D_\ell^2}(\Lambda_\ell m) \frac{1}{\sigma_{W_\ell}^2} \frac{1}{1+\Lambda_\ell m}\right) \ , \quad M_{J^T J}^{\langle -1 \rangle}(m) = \frac{m+1}{m} \prod_{\ell=1}^{L} \frac{\sigma_{W_\ell}^2 (1+\Lambda_\ell m)}{S_{D_\ell^2}(\Lambda_\ell m)} \ .$$

Proof. See the appendix, Subsection C.2. □

Master equation: In the end, we only need to fix width ratios and non-linearities to form $M_J^{\langle -1 \rangle}(m)$, and get the master equation which we solve numerically thanks to an adaptive Newton-Raphson scheme. The non-linearities ReLu, Hard Tanh and Hard Sine yield explicit formulas, which can be found in Table 1 of the appendix. If W_ℓ has i.i.d. entries, one finds the explicit master equation:

$$M_{J^T J}^{\langle -1 \rangle}(m) = \frac{m+1}{m} \prod_{\ell=1}^{L} \sigma_{W_\ell}^2 \left(c_\ell + \Lambda_\ell m\right) \ , \tag{13}$$

where $c_\ell = \frac{1}{2}$ when ϕ_ℓ is ReLu, $c_\ell = C_\ell = \mathbb{P}\left(0 \leq \mathcal{N} \leq \frac{1}{\sqrt{q^\ell}}\right)$ if ϕ_ℓ is Hard Tanh and $c_\ell = 1$ if ϕ_ℓ is Hard Sine.

4 Numerical resolution

Here we describe the numerical scheme aimed at computing the spectral density of $J^T J$ in Eq. (1). We use the following steps to compute the spectral density at a fixed $x \in \mathbb{R}_+$:

- Because of the Cauchy-Stieltjes inversion formula given in Lemma 2.1, pick a small $y > 0$ in order to compute: $-\frac{1}{\pi}\Im G_{J^TJ}(z = x + iy)$. The smaller the better, and in practice our method works for up to $y = 10^{-9}$. Figure 3 shows the same target distribution but convolved with various Cauchy distributions $y\mathcal{C}$ where $y \in \{1, 10^{-1}, 10^{-4}\}$. This corresponds to computing the density $\frac{-1}{\pi}\Im G_\mu(\cdot + iy)$ for different y's.
- Because of Eq. (8), we equivalently need to compute $M_{J^TJ}(z)$.
- $M^{\langle -1\rangle}_{J^TJ}(m)$ is available thanks to the master equation in Theorem 3.1. Therefore, we need to invert $m \mapsto M^{\langle -1\rangle}_{J^TJ}(m)$. This step is the crucial part: $M^{\langle -1\rangle}_{J^TJ}$ is multi-valued and one needs to choose the correct branch.

4.1 Initial setup

We first use the classical Newton-Raphson scheme to invert the equation $z = f(m)$ where $z \in \mathbb{C}_+$ is fixed and f is rational. A neat trick which leverages the fact that f is rational and that $z \in \mathbb{C}_+$ is to define:

$$\varphi_z(m) := P(m)/z - Q(z) \ . \tag{14}$$

As such, we have $z = f(m) = \frac{P(m)}{Q(m)} \iff \varphi_z(m) = 0$. There are several advantages of doing that: (1) Inversion is recast into finding the zero of a polynomial function. (2) Since we have $\lim_{z\to\infty} M(z) = 0$, if z is large in modulus, $m = 0$ is a natural starting point for the algorithm when z is large.

It is well-known that the Newton-Raphson scheme fails unless the initial guess $m_0 \in \mathbb{C}$ belongs to a basin of attraction for the method. And, provided such a guarantee, the Newton-Raphson scheme is exceptionally fast with a quadratic convergence speed. Kantorovich's seminal work in 1948 provides such a guarantee locally. For the reader's convenience, we give in Appendix D the pseudo-code for the Newton-Raphson algorithm (Algorithm 2), as well as a reference for the optimal form of the Kantorovich criterion (Theorem D.1).

Therefore, we assume that we have at our disposal a function $(z, m) \mapsto$ IS_IN_BASIN(z, m) which indicates if the Kantorovich criterion is satisfied for φ_z at any $m \in \mathbb{C}$. It is particularly easy to program with φ_z polynomial.

Algorithm 1 Newton lilypads, chaining basins of attraction

Name: NEWTON_LILYPADS
Input: Image value: $z_{objective} \in \mathbb{C}_+$, (Optional) Proxy: $(z_0, m_0) \in \mathbb{C}_+ \times \mathbb{C}$.
Output: $M(z_{objective})$
# Find a proxy $(z_0, m_0 = 0)$ using a doubling strategy, if None given
if (z_0, m_0) is None **then**
  $m \leftarrow 0$
  $z \leftarrow z_{objective}$
  while not IS_IN_BASIN(z, m) **do**
    $z \leftarrow z + i\Im(z) = \Re(z) + i2\Im(z)$ # Double imaginary part
  end while
  $m \leftarrow$ NEWTON_RAPHSON$(z, \text{Guess} = m)$
else
  $(z, m) \leftarrow (z_0, m_0)$
end if
# Starts heading towards $z_{objective}$ using dichotomy
while $|z_{objective} - z| > 0$ **do**
  $\Delta z \leftarrow z_{objective} - z$
  while not IS_IN_BASIN($z + \Delta z$, m) **do**
    $\Delta z \leftarrow 0.5 * \Delta z$
  end while
  $z \leftarrow z + \Delta z$
  $m \leftarrow$ NEWTON_RAPHSON$(z, \text{Guess} = m)$
end while
return m

4.2 Newton lilypads: Doubling strategies and chaining basins of attraction

Now we have all the (local) ingredients in order to describe a global strategy which solves in $m \in \mathbb{C}$ the equation $\varphi_z(m) = 0$.

First, one has to notice that this problem is part of a family parametrized by $z \in \mathbb{C}_+$. And the solution is $m \approx 0$ for z large. Therefore, one can find a proxy solution for $z \in \mathbb{C}_+$ high enough. This is done thanks to a doubling strategy until a basin of attraction is reached.

Second, if a proxy (z, m) is available, we can use the Newton-Raphson algorithm to find a solution $(z + \Delta z, m + \Delta m)$ starting from m. To do so, we need Δz small enough. This on the other hand is done by dichotomy.

Tying the pieces together allows to chain the different basins of attraction and leads to Algorithm 1. Notice that in the description of the algorithm, we chose to make implicit the dependence in the function f, since it is only passed along as a parameter. Technically, f is a parameter for all three functions NEWTON_RAPHSON, IS_IN_BASIN, NEWTON_LILYPADS. The discussion leading to this algorithm, combined with the Kantorovich criterion yields:

Theorem 4.1. *Given $f : m \mapsto M^{\langle -1 \rangle}(m)$ and $z \in \mathbb{C}^+$, Algorithm 1 has guaranteed convergence. Moreover it returns $m = M(z)$ i.e. the (unique) holomorphic extension of the inverse of f in the neighborhood of* 0.

5 On the experiment

To leverage the numerical scheme, we designed the following experiment whose results are reported on Fig. 2. Consider a classical Multi-Layer Perceptron (MLP) with $L = 4$ layers, feed-forward and fully-connected with ReLu non-linearities. The task at hand is classification on the classical MNIST dataset. The MLP's architecture is determined by the vector $\lambda = (\lambda_0, \lambda_1, \ldots, \lambda_L)$ while the gains at initialization are determined by the vector $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_L)$.

By randomly sampling the vector λ, we explore the space of possible architectures. In other to have balanced architectures, we chose independent λ_i's with $\mathbb{E}(\lambda_i) = 1$. Likewise, we also sample different possible gains. Hence we find ourselves with several MLPs architectures each with it's unique initialization. The spectral distributions are computed thanks to our Algorithm 1.

As shown in Fig. 2, the parameter count varies by orders of magnitude. To guarantee convergence despite the changing architectures, we intentionally chose a low learning rate and 50 epochs of training. For 200 MLPs, the experiment takes 10 hours of compute on a consumer GPU (RTX 2080Ti). This is to be contrasted with less than one minute of CPU compute for the spectral measure ν_J for all the MLPs.

To control for any stochastic variability, we also trained multiple instances for each MLP, and the shown results are averages between a few runs. Finally we calculate the correlations between the accuracy on the test set and the 90th percentile of the spectral distribution ν_J. A posteriori, the FPT regime is justified beyond initialization because the results are coherent despite various width scales.

6 Conclusion

In summary, this paper developed FPT in the rectangular setup which is the main tool for a metamodel for the stability of neural network. Then we gave a method for the numerical resolution both fast and with theoretical guarantees. Finally, confirming the initial claim that a metamodel focused on stability yields a good proxy for performance, we empirically demonstrated that the accuracy of a feed-forward neural network after training is highly correlated to higher quantiles of the *theoretically computed* spectral distribution.

Regarding the importance of hyperbolicity, we surmise that the presence of a very few large singular values allows SGD (Stochastic Gradient Descent) to escape local minima without compromising overall stability. Additionally to this conjecture, many other challenges remain in order to build a more versatile metamodel for stability: for example accomodating for skip-connections and convolutional layers fall outside the scope of the theory.

References

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[2] Octavio Arizmendi, Pierre Tarrago, Carlos Vargas, et al. Subordination methods for free deconvolution. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 56, pages 2565–2594. Institut Henri Poincaré, 2020.

[3] Serban T Belinschi and Mireille Capitaine. Spectral properties of polynomials in independent wigner and deterministic matrices. *Journal of Functional Analysis*, 273(12):3901–3963, 2017.

[4] Florent Benaych-Georges. Rectangular random matrices, related convolution. *Probability Theory and Related Fields*, 144(3-4):471–515, 2009.

[5] Florent Benaych-Georges and Mérouane Debbah. Free deconvolution: from theory to practice. *Paradigms for Biologically-Inspired Autonomic Networks and Services*, pages 201–224, 2008.

[6] Yoshua Bengio, Patrice Simard, and Paolo Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166, 1994.

[7] Benoit Collins and Tomohiro Hayase. Asymptotic freeness of layerwise jacobians caused by invariance of multilayer perceptron: The haar orthogonal case. *arXiv preprint arXiv:2103.13466*, 2021.

[8] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256, 2010.

[9] WB Gragg and RA Tapia. Optimal error bounds for the newton–kantorovich theorem. *SIAM Journal on Numerical Analysis*, 11(1):10–13, 1974.

[10] Boris Hanin and Mihai Nica. Products of many large random matrices and gradients in deep neural networks. *Communications in Mathematical Physics*, pages 1–36, 2019.

[11] LV Kantorovich. On newton's method for functional equations. In *Dokl. Akad. Nauk SSSR*, volume 59, pages 1237–1240, 1948.

[12] Cosme Louart, Zhenyu Liao, Romain Couillet, et al. A random matrix approach to neural networks. *The Annals of Applied Probability*, 28(2):1190–1248, 2018.

[13] Mylène Maïda, Tien Dat Nguyen, Thanh Mai Pham Ngoc, Vincent Rivoirard, and Viet Chi Tran. Statistical deconvolution of the free fokker-planck equation at fixed time. *arXiv preprint arXiv:2006.11899*, 2020.

[14] James A Mingo and Roland Speicher. *Free probability and random matrices*, volume 35. Springer, 2017.

[15] L. Pastur and V. Slavin. On random matrices arising in deep neural networks: General i.i.d. case, 2020.

[16] Leonid Pastur. On random matrices arising in deep neural networks. gaussian case. *arXiv preprint arXiv:2001.06188*, 2020.

[17] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

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

[19] Ben Poole, Subhaneil Lahiri, Maithra Raghu, Jascha Sohl-Dickstein, and Surya Ganguli. Exponential expressivity in deep neural networks through transient chaos. *arXiv preprint arXiv:1606.05340*, 2016.

[20] Samuel S Schoenholz, Justin Gilmer, Surya Ganguli, and Jascha Sohl-Dickstein. Deep information propagation. *arXiv preprint arXiv:1611.01232*, 2016.

[21] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in nlp. *arXiv preprint arXiv:1906.02243*, 2019.

[22] Pierre Tarrago. Spectral deconvolution of unitarily invariant matrix models. *arXiv preprint arXiv:2006.09356*, 2020.

[23] Dan Voiculescu. Multiplication of certain non-commuting random variables. pages 223–235. JSTOR, 1987.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] In the final paragraph of the conclusion, as an opening.
   (c) Did you discuss any potential negative societal impacts of your work? [No] We do not see any.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] Essentially, the free probability results (Theorems 2.3 and 3.1) require hypotheses on the nature of the randomness at initialization and the large width regime. Asymptotic freeness is not a hypothesis but a consequence of that.
   The numerical guarantees (Theorem 4.1) use the description of our method. And thus no extra hypotheses are required.
   (b) Did you include complete proofs of all theoretical results? [Yes] In the appendices / supplementary material
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] An URL to an anonymized repository is provided. There the code is released under a standard Apache license. Datasets are standard and obtained via Pytorch's dataloaders.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] Regarding Fig. 1, the algorithms we ran are deterministic except Monte-Carlo. There is some variability in the computational time, but it is very small. And it is further erased by the log-scale. We already plotted the results of 5 runs, hence the multiple markers for a fixed x.
   Regarding the experiment of Fig. 2, we made multiple runs for the same MLP. But we chose to not display error bars:
   - as it would make Fig. 2 difficult to read.
   - we are nowhere close to the CLT regime to plot error bars, as we made only 2 to 5 runs per architecture. A few runs is already a good sanity check.

- the goal is to assess a global correlation and not the result of a single training. As such noise is acceptable.
As stated in Section 5, we intentionally chose of the very low learning rate and an excessive amount of epochs to minimise the variability of the final losses and accuracies.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 5

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [No] We are using standard datasets which are automatically available via Pytorch
   (b) Did you mention the license of the assets? [No]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Appendices contain supplementary material.

A Appendix: More figures

In this section, we record non-essential yet pedagogical figures.

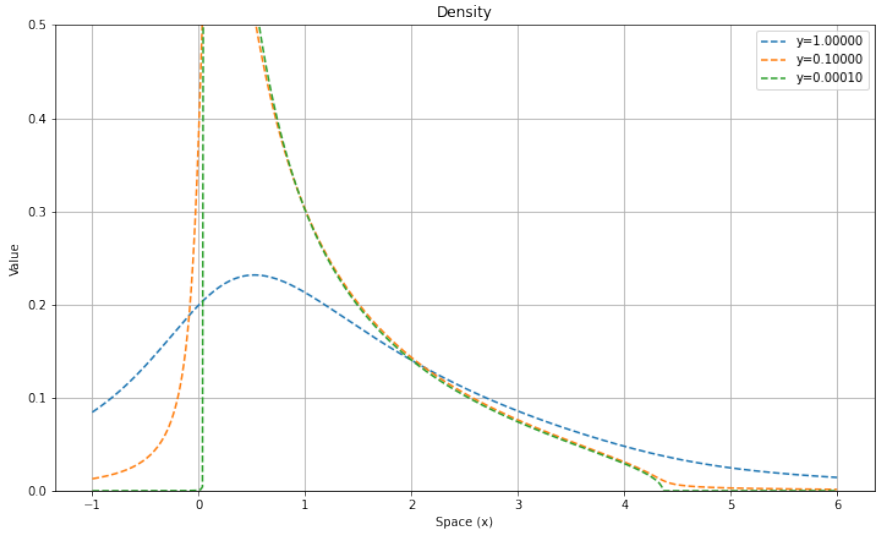


Figure 3: Plot of the probability density $\frac{-1}{\pi}\Im G_\mu\left(\cdot + iy\right)$ for $y \in \{1, 10^{-1}, 10^{-4}\}$. Here μ is the multiplicative free convolution of three Marchenko-Pastur distributions, with different parameters.

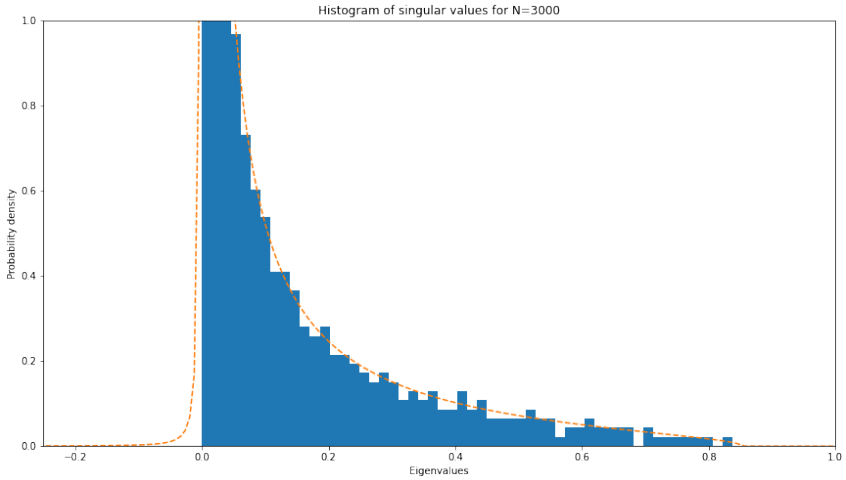


Figure 4: Plot of density of singular values for the Jacobian matrix J. The network has constant width and Relu non-linearities. Monte-Carlo sampling uses matrices of size $N_\ell = N = 3000$.

B Appendix: Table of non-linearities

Let us now collect the various required formulas, and specialize them to a selection of non-linearities: ReLu, Hard Tanh and Hard Sine. These non-linearities are very tractable hence the choice. As discussed in the introduction, the empirical distribution of $D_\ell^{(\mathbf{N})}$ converges to the law of $\phi_l'\left(\sqrt{q^\ell}\mathcal{N}\right)$.

From this observation was deduced in [18] the following formula:

$$M_{D_\ell^2}(z) = \sum_{k\geq 1} \frac{m_k(D_\ell^2)}{z^k} = \mathbb{E}\left[\frac{\phi_\ell'(\sqrt{q^\ell}\mathcal{N})^2}{z-\phi_\ell'(\sqrt{q^\ell}\mathcal{N})^2}\right] , \tag{15}$$

$$S_{D_\ell^2}(m) = \frac{1+m}{mM_{D_\ell^2}^{\langle -1\rangle}(m)} , \tag{16}$$

	Linear	ReLu	Hard Tanh	Hard sine / Triangle
$\phi_\ell(h)$	h	$[h]_+$	$[h+1]_+ - [h-1]_+ - 1$	$\frac{2}{\pi}\arcsin\circ\sin(\frac{\pi}{2}h)$
$M_{D_\ell^2}(z)$	$\frac{1}{z-1}$	$\frac{1}{2}\frac{1}{z-1}$	$C_\ell\frac{1}{z-1}$	$\frac{1}{z-1}$
$m_k(D_\ell^2)$	1	$\frac{1}{2}$	C_ℓ	1
$S_{D_\ell^2}(m)$	1	$\frac{1+m}{\frac{1}{2}+m}$	$\frac{1+m}{C_\ell+m}$	1
$g_\ell(q)$	q	$\frac{q}{2}$	$2qC_\ell - \sqrt{\frac{2q}{\pi}}e^{-\frac{1}{2q}}$	$\frac{1}{3}+\frac{4}{\pi^2}\sum_{n\geq 1}\frac{(-1)^n}{n^2}e^{-q\frac{\pi^2 n^2}{2}}$

Table 1: Table of formulas for moment generating functions and S-transforms.

In Table 1.1, the reader shall find the formulae of $M_{D_\ell^2}$, $S_{D_\ell^2}$ and the recurrence relation given by f_ℓ where

$$f_\ell\left(q^{\ell-1}\right) = \sigma_{W_\ell}^2 g_\ell(q^{\ell-1}) + \sigma_{b_\ell}^2 ,$$
$$\text{with } g_\ell(q) := \mathbb{E}\left[\phi_\ell\left(\sqrt{q}\mathcal{N}\right)^2\right] .$$

and

$$C_\ell = \mathbb{P}\left(0 \leq \mathcal{N} \leq \frac{1}{\sqrt{q^\ell}}\right) .$$

Proof of formulas in Table 1: ReLu and Hard Tanh have already been computed in [18].

Hard Sine: This is the most tricky formula to establish. If $\phi(x) = \frac{2}{\pi}\arcsin\circ\sin\left(\frac{\pi}{2}x\right)$ and $\widehat{f}(\xi) = \int dx\ f(x)e^{i\xi x}$ is the Fourier transform on f, then the application of the Plancherel formula yields:

$$\begin{aligned}
&\mathbb{E}\left[\phi\left(\sqrt{q}\mathcal{N}\right)^2\right] \\
=&\int_{-\infty}^{\infty} dx \frac{e^{-\frac{x^2}{2q}}}{\sqrt{2\pi q}}\phi^2(x) \\
=&\frac{1}{2\pi}\int_{-\infty}^{\infty} d\xi\ \widehat{\phi^2}(\xi)e^{-q\frac{\xi^2}{2}} .
\end{aligned}$$

But in term of Fourier series:

$$\phi^2(x) = \sum_{n\in\mathbb{Z}} \left(\widehat{\phi^2}\right)_n e^{i\pi n x}$$

as $\phi^2(x) = x^2$ on $[-1,1]$ and extended in order to become 2-periodic. In terms of Schwartz distributions:

$$\widehat{\phi^2}(\xi) = \sum_{n\in\mathbb{Z}} \left(\widehat{\phi^2}\right)_n (x \mapsto \widehat{e^{i\pi n x}}) = \sum_{n\in\mathbb{Z}} \left(\widehat{\phi^2}\right)_n 2\pi\delta_{-\pi n}(d\xi) .$$

Hence:

$$\mathbb{E}\left[\phi\left(\sqrt{q}\mathcal{N}\right)^2\right] = \sum_{n\in\mathbb{Z}} \left(\widehat{\phi^2}\right)_n e^{-q\frac{\pi^2 n^2}{2}} .$$

We conclude by computing the Fourier coefficients of ϕ^2.

$$\left(\widehat{\phi^2}\right)_0 = \frac{1}{2}\int_{-1}^{1} dx\, x^2 = \frac{1}{3}\,.$$

$$\begin{aligned}
\left(\widehat{\phi^2}\right)_n &= \frac{1}{2}\int_{-1}^{1} dx\, x^2\, e^{-i\pi n x} \\
&= \int_0^1 dx\, x^2\, \cos(\pi n x) \\
&= -\frac{1}{\pi n}\int_0^1 dx\, 2x\, \sin(\pi n x) \\
&= \frac{2}{(\pi n)^2}\left[x\cos(n\pi x)\right]_0^1 - \frac{2}{(\pi n)^2}\int_0^1 dx\, \cos(\pi n x) \\
&= \frac{2}{(\pi n)^2}(-1)^n\,.
\end{aligned}$$

In the end:

$$\mathbb{E}\left[\phi\left(\sqrt{q}\mathcal{N}\right)^2\right] = \frac{1}{3} + \frac{4}{\pi^2}\sum_{n\geq 1}\frac{(-1)^n}{n^2}e^{-q\frac{\pi^2 n^2}{2}}\,.$$

C Appendix: Proofs

C.1 Proof of Theorem 2.3

For $k \in \mathbb{N}$, we have :

$$\begin{aligned}
Tr\left[\left(\left(A^{(N)}B^{(N)}\right)^T A^{(N)}B^{(N)}\right)^k\right] &= Tr\left[\left(\left(B^{(N)}\right)^T\left(A^{(N)}\right)^T AB\right)^k\right] \\
&= Tr\left[(B^{(N)}\left(B^{(N)}\right)^T\left(A^{(N)}\right)^T A)^k\right]\,.
\end{aligned}$$

As $\left(A^{(N)}B^{(N)}\right)^T A^{(N)}B^{(N)} \in M_{r_N}(\mathbb{C})$ and $B^{(N)}\left(B^{(N)}\right)^T\left(A^{(N)}\right)^T A^{(N)} \in M_{q_N}(\mathbb{C})$, this shows that

$$M_{\left(A^{(N)}B^{(N)}\right)^T A^{(N)}B^{(N)}}(z) = \frac{q_N}{r_N}M_{B^{(N)}\left(B^{(N)}\right)^T\left(A^{(N)}\right)^T A^{(N)}}(z)\,,$$

and then

$$M^{\langle -1\rangle}_{\left(A^{(N)}B^{(N)}\right)^T A^{(N)}B^{(N)}}(m) = M^{\langle -1\rangle}_{B^{(N)}\left(B^{(N)}\right)^T\left(A^{(N)}\right)^T A^{(N)}}\left(\frac{r_N}{q_N}m\right).$$

Consequently,

$$\begin{aligned}
S_{\left(A^{(N)}B^{(N)}\right)^T A^{(N)}B^{(N)}}(m) =& \frac{1+m}{mM^{\langle -1\rangle}_{\left(A^{(N)}B^{(N)}\right)^T A^{(N)}B^{(N)}}(m)} \\
=& \frac{1+m}{mM^{\langle -1\rangle}_{B^{(N)}\left(B^{(N)}\right)^T\left(A^{(N)}\right)^T A^{(N)}}\left(\frac{r_N}{q_N}m\right)} \times \frac{1+\frac{r_N}{q_N}m}{\frac{r_N}{q_N}m} \times \frac{\frac{r_N}{q_N}m}{1+\frac{r_N}{q_N}m} \\
=& \frac{r_N}{q_N}\frac{1+m}{1+\frac{r_N}{q_N}m}S_{B^{(N)}\left(B^{(N)}\right)^T\left(A^{(N)}\right)^T A^{(N)}}\left(\frac{r_N}{q_N}m\right)\,. \qquad (17)
\end{aligned}$$

As $\left(A^{(N)}\right)^T A^{(N)}$ and $B^{(N)}\left(^{(N)}\right)^T$ are asymptotically free, taking the limit $N \to +\infty$ and applying Voiculescu's Theorem 2.2 , we get

$$S_{(AB)^*AB}(m) = \alpha\frac{1+m}{1+\alpha m}S_{BB^*}(\alpha m)\, S_{A^*A}(\alpha m)\,.$$

Moreover, the above equality is true replacing A with the identity I. $S_I(m) = 1$ yields:

$$S_{B^*B}(m) = \alpha \frac{1+m}{1+\alpha m} S_{BB^*}(\alpha m) \ .$$

Finally we have

$$S_{(AB)^*AB}(m) = S_{A^*A}(\alpha m)\, S_{B^*B}(m) \ .$$

This concludes the proof.

C.2 Proof of Theorem 3.1

We assume that W_ℓ have i.i.d. entries. Thanks to a swapping trick justified in [15], we can assume that the matrices D_ℓ have i.i.d. entries independent from the rest of the network, distributed as $\phi'_\ell(\sqrt{q_\ell}\mathcal{N})$. Notice that we can also replace the D_ℓ's by deterministic matrices that use the quantiles of the same distribution. This together with standard results from FPT such as [14] gives asymptotic freeness – see [3] for a more general result reflecting the current state of the art. Therefore, we can apply Theorem 2.3.

Starting from Eq. (1) , we get in the infinite width regime and by induction:

$$\begin{aligned}
&S_{\left(J^{(\mathbf{N})}\right)^T J^{(\mathbf{N})}}(m) \\
&= S_{\left(D_L^{(\mathbf{N})} W_L^{(\mathbf{N})} D_{L-1}^{(\mathbf{N})} W_{L-1}^{(\mathbf{N})} \dots D_1^{(\mathbf{N})} W_1^{(\mathbf{N})}\right)^T D_L^{(\mathbf{N})} W_L^{(\mathbf{N})} D_{L-1}^{(\mathbf{N})} W_{L-1}^{(\mathbf{N})} \dots D_1^{(\mathbf{N})} W_1^{(\mathbf{N})}}(m) \\
&= S_{\left(D_L^{(\mathbf{N})} W_L^{(\mathbf{N})} D_{L-1}^{(\mathbf{N})} W_{L-1}^{(\mathbf{N})} \dots D_2^{(\mathbf{N})} W_2^{(\mathbf{N})} D_1^{(\mathbf{N})}\right)^T D_L^{(\mathbf{N})} W_L^{(\mathbf{N})} D_{L-1}^{(\mathbf{N})} W_{L-1}^{(\mathbf{N})} \dots D_2^{(\mathbf{N})} W_2^{(\mathbf{N})} D_1^{(\mathbf{N})}}(\lambda_1 m) \\
&\qquad \times S_{\left(W_1^{(\mathbf{N})}\right)^T W_1^{(\mathbf{N})}}(m) \\
&= S_{\left(D_L^{(\mathbf{N})} W_L^{(\mathbf{N})} D_{L-1}^{(\mathbf{N})} W_{L-1}^{(\mathbf{N})} \dots D_2^{(\mathbf{N})} W_2^{(\mathbf{N})}\right)^T D_L^{(\mathbf{N})} W_L^{(\mathbf{N})} D_{L-1}^{(\mathbf{N})} W_{L-1}^{(\mathbf{N})} \dots D_2^{(\mathbf{N})} W_2^{(\mathbf{N})}}(\lambda_1 m) \\
&\qquad \times S_{\left(D_1^{(\mathbf{N})}\right)^2}(\lambda_1 m) S_{\left(W_1^{(\mathbf{N})}\right)^T W_1^{(\mathbf{N})}}(m) \\
&\vdots \\
&= \prod_{\ell=1}^{L} \left[S_{\left(D_\ell^{(\mathbf{N})}\right)^2} \left(\prod_{k=1}^{l} \lambda_k m \right) S_{\left(W_\ell^{(\mathbf{N})}\right)^T W_\ell^{(\mathbf{N})}} \left(\prod_{k=1}^{l-1} \lambda_k m \right) \right] \\
&= \prod_{\ell=1}^{L} \left[S_{\left(D_\ell^{(\mathbf{N})}\right)^2} (\Lambda_\ell m)\, S_{\left(W_\ell^{(\mathbf{N})}\right)^T W_\ell^{(\mathbf{N})}} (\Lambda_{\ell-1} m) \right] \ ,
\end{aligned}$$

with the convention $\Lambda_0 = \prod_{k=1}^{0} \lambda_k = 1$.

Under the asumption that the entries of W_ℓ are i.i.d., the Marcenko-Pastur Theorem gives

$$S_{\left(W_\ell^{(\mathbf{N})}\right)^T W_\ell^{(\mathbf{N})}}(m) = \frac{1}{\sigma_{W_\ell}^2} \frac{1}{1+\lambda_\ell m},$$

which leads to

$$S_{\left(J^{(\mathbf{N})}\right)^T J^{(\mathbf{N})}}(m) = \prod_{\ell=1}^{L} \left(S_{\left(D_\ell^{(\mathbf{N})}\right)^2} (\Lambda_\ell m) \frac{1}{\sigma_{W_\ell}^2} \frac{1}{1+\Lambda_\ell m} \right) \ .$$

We thus have

$$
\begin{aligned}
M^{\langle -1\rangle}_{\left(J^{(\mathbf{N})}\right)^T J^{(\mathbf{N})}}(m) &= \frac{m+1}{m S_{\left(J^{(\mathbf{N})}\right)^T J^{(\mathbf{N})}}(m)} \\
&= \frac{m+1}{m \prod_{\ell=1}^{L} \left(S_{\left(D_\ell^{(\mathbf{N})}\right)^2}(\Lambda_\ell m) \frac{1}{\sigma_{W_\ell}^2} \frac{1}{1+\Lambda_\ell m} \right)} \\
&= \frac{(m+1) \prod_{\ell=1}^{L} \sigma_{W_\ell}^2 (1+\Lambda_\ell m)}{m \prod_{\ell=1}^{L} S_{\left(D_\ell^{(\mathbf{N})}\right)^2}(\Lambda_\ell m)} \\
&= \frac{(m+1) \prod_{\ell=1}^{L} \sigma_{W_\ell}^2 (1+\Lambda_\ell m)}{m \prod_{\ell=1}^{L} S_{\left(D_\ell^{(\mathbf{N})}\right)^2}(\Lambda_\ell m)} \, .
\end{aligned}
$$

D Appendix: On the classical Newton-Raphson scheme

Algorithm 2 Newton-Raphson scheme for a rational function f

Name: NEWTON_RAPHSON
Input:
Numerical precision: $\varepsilon > 0$ (Default: 10^{-12}),
Image value: $z \in \mathbb{C}_+$,
Polynomials: P, Q such that $f = \frac{P}{Q}$,
(Optional) Guess: $m_0 \in \mathbb{C}$, (Default: $m_0 = 0$).
Code:
$m \leftarrow m_0$
while True **do**
  value $\leftarrow \varphi_z(m)$ # See Eq. (14)
  if $|value| < \varepsilon$ **then**
    return m
  end if
  grad $\leftarrow \varphi_z'(m)$
  $m \leftarrow m - value/grad$
end while

Here we give the optimal Kantorovich criterion from [9] adapted to this paper. Fix $z \in \mathbb{C}_+$ and recall that φ_z in Eq. (14) is the map whose zero we want to find.

Theorem D.1 (Kantorovich's criterion, [11])**.** *Consider a starting point $m_0 \in \mathbb{C}$, and define:*

$$
\delta := \left| \frac{\varphi_z(m_0)}{\varphi_z'(m_0)} \right| \, , \quad \kappa := \left| \frac{1}{\varphi_z'(m_0)} \right| \, .
$$

If the starting point satisfies $h := \delta\kappa\lambda < \frac{1}{2}$, where

$$
\lambda := \sup_{|m-m_0| \le t^*} |\varphi_z''(m)| \, , \quad t^* := \frac{2\delta}{1+\sqrt{1-h}} < 2\delta \, .
$$

Then, the Newton-Raphson scheme, starting from m_0 converges to m^* such that $\varphi_z(m^*) = 0$. Furthermore, the convergence at each step is at least quadratic.

E Appendix: Moments of J

We can reach an early understanding of the behavior of J's singular values by computing mean and variance. For ease of notation, we write:

$$
m_k^{(s)}(A) = m_k(A^T A)
$$

for any operator A, which admits a measure of singular values. We have under the assumptions that the entries of W_ℓ are i.i.d. :

$$m_1^{(s)}(J) = \prod_{\ell=1}^{L} \left(m_1^{(s)}(D_\ell) m_1^{(s)}(W_\ell) \right) \; = \; \prod_{\ell=1}^{L} (c_\ell \sigma_{W_\ell}^2) \; , \tag{18}$$

$$\begin{aligned} & m_2^{(s)}(J) - m_1^{(s)}(J) \\ &= m_1^{(s)}(J)^2 \left(\sum_{\ell=1}^{L} \Lambda_\ell \left(\frac{m_2^{(s)}(D_\ell) - m_1^{(s)}(D_\ell)^2}{m_1^{(s)}(D_\ell)^2} + \frac{m_2^{(s)}(W_\ell) - m_1^{(s)}(W_\ell)^2}{m_1^{(s)}(W_\ell)^2} \right) \right) \\ &= \left(\prod_{\ell=1}^{L} c_\ell^2 \sigma_{W_\ell}^4 \right) \left(\sum_{\ell=1}^{L} \Lambda_\ell \left(\frac{1 - c_\ell}{c_\ell} + \lambda_\ell \right) \right) \; . \end{aligned} \tag{19}$$

Under the asumption that $W_\ell^T W_\ell = \sigma_{W_\ell} I_{N_{\ell-1}}$, we find the same $m_1^{(s)}(J)$ and :

$$\begin{aligned} m_2^{(s)}(J) - (m_1^{(s)}(J))^2 &= \sum_{\ell=1}^{L} \left(\Lambda_\ell \left(\frac{m_2^{(s)}(D_\ell)}{m_1^{(s)}(D_\ell)^2} - 2 \right) \right) \prod_{\ell=1}^{L} \left(c_\ell^2 \sigma_{W_\ell}^4 \right) \\ &= \sum_{\ell=1}^{L} \left(\Lambda_\ell \left(\frac{1}{c_\ell} - 2 \right) \right) \prod_{\ell=1}^{L} \left(c_\ell^2 \sigma_{W_\ell}^4 \right) . \end{aligned} \tag{20}$$

These formulas need to be interpreted:

- Variance grows with L, showing increased instability with depth.
- Larger layers, relative to N_0, give larger Λ_ℓ's and thus the same effect.

Proof: Computations of moments. The following remark is useful in the computation of moments.

Remark E.1 (Moments)**.** *At the neighborhood of $z \sim \infty$:*

$$M_\mu(z) = \frac{m_1(\mu)}{z} + \frac{m_2(\mu)}{z^2} + \mathcal{O}\left(z^{-3}\right) \; .$$

By inversion, at the neighborhood of $m \sim 0$:

$$M_\mu^{\langle -1 \rangle}(m) = \frac{m_1(\mu)}{m} + \mathcal{O}\left(1\right) \; .$$

$$M_\mu^{\langle -1 \rangle}(m) = \frac{m_1(\mu)}{m} + \frac{m_2(\mu)}{m_1(\mu)} + \mathcal{O}\left(m\right) \; .$$

Hence:

$$\begin{aligned} S_\mu(m) = & \; \frac{1+m}{m_1(\mu) + m \frac{m_2(\mu)}{m_1(\mu)} + \mathcal{O}\left(m^2\right)} \\ = & \; \frac{1}{m_1(\mu)} (1+m) \left(1 - m \frac{m_2(\mu)}{m_1(\mu)^2} + \mathcal{O}\left(m^2\right) \right) \\ = & \; \frac{1}{m_1(\mu)} + \frac{m}{m_1(\mu)} \left(1 - \frac{m_2(\mu)}{m_1(\mu)^2} \right) + \mathcal{O}\left(m^2\right) \; . \end{aligned}$$

Thanks to this, we can prove Eq. (18) and (19). By Remark E.1 and Theorem 3.1, we have as $m \to 0$:

$$\begin{aligned}
S_{J^TJ}(m) &= \prod_{\ell=1}^{L} \left[S_{D_\ell^2}\left(\Lambda_\ell m\right) S_{W_\ell^T W_\ell}\left(\Lambda_\ell m\right) \right] \\
&= \prod_{\ell=1}^{L} \Bigg[\left(\frac{1}{m_1^{(s)}(D_\ell)} + m \frac{\Lambda_\ell}{m_1^{(s)}(D_\ell)} \left(1 - \frac{m_2^{(s)}(D_\ell)}{m_1^{(s)}(D_\ell)^2} \right) + \mathcal{O}\left(m^2\right) \right) \\
&\qquad \left(\frac{1}{m_1^{(s)}(W_\ell)} + m \frac{\Lambda_\ell}{m_1^{(s)}(W_\ell)} \left(1 - \frac{m_2^{(s)}(W_\ell)}{m_1^{(s)}(W_\ell)^2} \right) + \mathcal{O}\left(m^2\right) \right) \Bigg] \\
&= \left[\prod_{\ell=1}^{L} \frac{1}{m_1^{(s)}(D_\ell) m_1^{(s)}(W_\ell)} \right] \prod_{\ell=1}^{L} \left[1 + m\Lambda_\ell \left(2 - \frac{m_2^{(s)}(D_\ell)}{m_1^{(s)}(D_\ell)^2} - \frac{m_2^{(s)}(W_\ell)}{m_1^{(s)}(W_\ell)^2} \right) + \mathcal{O}\left(m^2\right) \right].
\end{aligned}$$

Identifying the first order term, one finds indeed Eq. (18). Continuing the previous computation:

$$S_{J^TJ}(m) = \frac{1}{m_1^{(s)}(J)} + \frac{m}{m_1^{(s)}(J)} \left(\sum_{\ell=1}^{L} \Lambda_\ell \left(2 - \frac{m_2^{(s)}(D_\ell)}{m_1^{(s)}(D_\ell)^2} - \frac{m_2^{(s)}(W_\ell)}{m_1^{(s)}(W_\ell)^2} \right) \right) + \mathcal{O}\left(m^2\right).$$

Applying Remark E.1 again for S_{J^TJ}, we get :

$$\sum_{\ell=1}^{L} \Lambda_\ell \left(2 - \frac{m_2^{(s)}(D_\ell)}{m_1^{(s)}(D_\ell)^2} - \frac{m_2^{(s)}(W_\ell)}{m_1^{(s)}(W_\ell)^2} \right) = 1 - \frac{m_2^{(s)}(J)}{m_1^{(s)}(J)^2}$$

which is equivalent to Eq. (19).

We conclude by specializing to classical weight distributions. Under the assumption that the entries of W_ℓ are i.i.d. we have $m_1^{(s)}(W_\ell) = \sigma_{W_\ell}^2$ and $m_2^{(s)}(W_\ell) = \sigma_{W_\ell}^4(1+\lambda_\ell)$ which gives

$$\begin{aligned}
m_2^{(s)}(J) &= \left(1 - \sum_{\ell=1}^{L} \left(\Lambda_\ell \left(1 - \frac{m_2^{(s)}(D_\ell)}{m_1^{(s)}(D_\ell)^2} - \lambda_\ell \right) \right) \right) \prod_{\ell=1}^{L} \left(m_1^{(s)}(D_\ell)^2 \sigma_{W_\ell}^4 \right) \\
m_2^{(s)}(J) - (m_1^{(s)}(J))^2 &= \sum_{\ell=1}^{L} \left(\Lambda_\ell \left(\frac{m_2^{(s)}(D_\ell)}{m_1^{(s)}(D_\ell)^2} + \lambda_\ell - 1 \right) \right) \prod_{\ell=1}^{L} \left(m_1^{(s)}(D_\ell)^2 \sigma_{W_\ell}^4 \right).
\end{aligned}$$

□

F Appendix: More details on the benchmarks and the experiment

Recall that we have provided an anonymized Github repository at the address:

`https://anonymous.4open.science/r/FreeNN-331E/`

F.1 On the benchmarks of Fig. 1

The figure presents the computational time required for computing the density of ν_J. Let us make the following comments:

- Pennington and al. 's algorithm has been implemented using a native root finding procedure. As such, it is much more optimized than our code.
- A closer examination of the timings shows that Newton lilypads scales sublinearly with the number of required points. This is easily understood by the fact that smaller N requires the computation of more basins of attraction per point.
- Also, the Monte-Carlo method used matrices of size $n = 3000$. Not only Monte-Carlo is imprecise because of the noise, but its performance scales very poorly with n since one needs to diagonalize ever larger matrices. Of course, Monte-Carlo remains the easiest method to implement. But the deterministic methods compute the density up to machine precision.

F.2 Description of the experiment of Fig. 2

Recall that we considered a classical Multi-Layer Perceptron (MLP) with $L = 4$ layers, feed-forward and fully-connected with ReLu non-linearities. The MLP's architecture is determined by the vector $\lambda = (\lambda_0, \lambda_1, \ldots, \lambda_L)$. The initialization follows the Xavier normal initialization [8] as implemented in Pytorch [17]. Thus the full vector of variances of this initialization $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_L)$ is determined up to a multiplicative factor called the gain.

First, we sample random architectures. We chose

- the λ_i to be i.i.d. and uniform on $\{\frac{1}{4}, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, 1.0, \frac{3}{2}, 2, 3, 4\}$. As such the mean is $\mathbb{E}\lambda_i = 1$ in order to obtain relatively balanced architectures.
- the single gain is taken as a uniform random variable on $\{\frac{1}{4}, \frac{1}{2}, 1, 2, 4\}$.

Secondly, we compute the spectral distributions predicted by FPT for each random architecture. Some architectures are very unbalanced and are discarded.

Thirdly, we train more than 200 instances of MLPs and record the learning curves and accuracies.

In the end, by considering FPT quantiles and the final test accuracy for each MLP, we obtain data which amenable to a bivariate statistical analysis.

More comments are in order:

- Other initializations than normal were tested, without a noticeable difference. In any case, FPT predicts the same outcomes.
- It was our impression that more stable architectures train better regardless of the SGD flavor. Therefore we ended up using a vanilla SGD without momentum.
- Due to the sampling procedure, the number of parameters for our MLPs can differ by orders of magnitudes. As shown in Fig. 2, the range is from 10^4 to 10^8. The largest ones do benefit from GPU acceleration.

F.3 Final remarks

Our sampling procedure does not consider large fluctuations in the λ_i 's and focuses on balanced architectures. Likewise, the gains at initialization do not deviate much from the classical [8]. It is important to recognize that without the insight of FPT, the scalings applied to such initializations are already normalized so that spectral measures do converge.

As such, we never encounter truly problematic gradient vanishing or gradient explosion, which completely sabotage the convergence of the neural network. Our refined FPT metrics are arguably "second order corrections". Nevertheless, it is surprising the 90th percentile in Fig. 2 highly correlates to the final test accuracy after training. In the end, tuning FPT metrics does not amount to second order corrections.