

Scaling ResNets in the Large-depth Regime

Pierre Marion

PIERRE.MARION@SORBONNE-UNIVERSITE.FR

*Sorbonne Université, CNRS, Laboratoire de Probabilités, Statistique et Modélisation
F-75005 Paris, France*

Adeline Fermanian

ADELINE.FERMANIAN@MINES-PARISTECH.FR

*MINES ParisTech, PSL Research University, CBIO, F-75006 Paris, France
Institut Curie, PSL Research University, F-75005 Paris, France
INSERM, U900, F-75005 Paris, France*

G erard Biau

GERARD.BIAU@SORBONNE-UNIVERSITE.FR

*Sorbonne Universit , CNRS, Laboratoire de Probabilit s, Statistique et Mod lisation
F-75005 Paris, France*

Jean-Philippe Vert

JPVERT@GOOGLE.COM

Google Research, Brain team, Paris, France

Abstract

Deep ResNets are recognized for achieving state-of-the-art results in complex machine learning tasks. However, the remarkable performance of these architectures relies on a training procedure that needs to be carefully crafted to avoid vanishing or exploding gradients, particularly as the depth L increases. No consensus has been reached on how to mitigate this issue, although a widely discussed strategy consists in scaling the output of each layer by a factor α_L . We show in a probabilistic setting that with standard i.i.d. initializations, the only non-trivial dynamics is for $\alpha_L = 1/\sqrt{L}$ —other choices lead either to explosion or to identity mapping. This scaling factor corresponds in the continuous-time limit to a neural stochastic differential equation, contrarily to a widespread interpretation that deep ResNets are discretizations of neural ordinary differential equations. By contrast, in the latter regime, stability is obtained with specific correlated initializations and $\alpha_L = 1/L$. Our analysis suggests a strong interplay between scaling and regularity of the weights as a function of the layer index. Finally, in a series of experiments, we exhibit a continuous range of regimes driven by these two parameters, which jointly impact performance before and after training.

Keywords: ResNets, deep learning theory, neural ODE, neural network initialization, continuous-time models

1. Introduction

1.1 Deep residual neural networks

Residual neural networks (ResNets), introduced by [He et al. \(2016a\)](#) in the field of computer vision, were the first deep neural network models successfully trained with several thousand layers. Since then, extensive empirical evidence has shown that increasing the depth leads to significant improvements in performance, while raising new challenges in terms of training (e.g., [Wang et al., 2022](#)). From a high-level perspective, the key feature of ResNets is the presence of skip connections between successive layers. In mathematical terms, this means that the $(k + 1)$ -th hidden state $h_{k+1} \in \mathbb{R}^d$ follows sequentially from the previous hidden

state via the recurrence relation

$$h_{k+1} = h_k + f(h_k, \theta_{k+1}), \quad 0 \leq k \leq L - 1, \quad (1)$$

where $f(\cdot, \theta_{k+1})$ is the layer function parameterized by $\theta_{k+1} \in \mathbb{R}^p$ and L is the number of layers. The skip connection corresponds to the addition of h_k on the right-hand side of (1), which is absent in classical feedforward networks. This refinement prevents instability issues during training when L is large, provided training is performed in a careful way (He et al., 2015). The idea of adding skip connections has become common practice in the field of deep learning, and is today incorporated in many other models such as Transformers in natural language processing (Vaswani et al., 2017). For simplicity, in the rest of the paper, we continue to use the terminology ResNets to denote any architecture of the form (1), keeping in mind that this framework goes beyond the original model of He et al. (2016a).

The most common architectures have 50-150 layers, but ResNets can be trained with depths up to the order of thousand layers (He et al., 2016b). Yet, the training procedure needs to be carefully crafted to avoid vanishing or exploding gradients, particularly as the depth increases. As pointed out by, e.g., Shao et al. (2020), these instabilities are related to a shift in the magnitude of the variance of a signal as it passes through the network. In the original approach of He et al. (2016a), the issue was mitigated by adding a normalization step, called batch normalization (Ioffe and Szegedy, 2015), which rescales the output of each layer via centering and unit variance normalization. However, this normalization stage introduces practical and theoretical difficulties, among which computational overhead and strong dependence on the batch size (see Brock et al., 2021, and the references therein). A widespread alternative to stabilize training in deep models, explored for example by Yang and Schoenholz (2017), Arpit et al. (2019), Zhang et al. (2019b), and De and Smith (2020), is to incorporate a scaling factor α_L in front of the residual term in (1), yielding the model

$$h_{k+1} = h_k + \alpha_L f(h_k, \theta_{k+1}), \quad 0 \leq k \leq L - 1. \quad (2)$$

There is strong evidence that this scaling factor α_L should depend on L , without however any consensus to date on the exact form of this dependence, nor on the mathematical grounding of the approach. Thus, despite progresses on the empirical side, the mathematical forces in action behind the stability of deep ResNets are still poorly understood, although they are key to unlock training at arbitrary depth.

Our goal in the present paper is to take a step forward towards a better theoretical understanding of deep ResNets by providing a thorough probabilistic analysis of the sequence $(h_k)_{0 \leq k \leq L}$ at initialization when L is large, and by leveraging a continuous-time interpretation of model (2) via the so-called neural ordinary differential equation (neural ODE, Chen et al., 2018) paradigm. In a nutshell, our results highlight the intimate connection that exists at initialization between stability of the learning process, the regularity of the weights, and the scaling factor α_L . We offer in particular a proper mathematical grounding on why and how to choose the parameter α_L as a function of the depth L and the distribution of the weights.

1.2 Our contributions

Scaling at initialization. The optimal parameters of ResNets are learnt by minimizing some empirical risk function via a gradient descent algorithm. As highlighted for example

by Yang and Schoenholz (2017), Hanin and Rolnick (2018), and Arpit et al. (2019), a good parameter initialization of this learning phase plays a major role in the quality of the learnt model, in particular to avoid vanishing gradients and deadlock at initialization, or exploding gradients and quick divergence of the model parameters at the beginning of training. Moreover, a good initialization allows the use of larger learning rates, which have been shown to correlate with better generalization (Jastrzebski et al., 2017). It is thus of great interest to study and understand the role played by scaling of deep ResNets at initialization. This is the context in which we place ourselves in the sequel.

At initialization stage, the weights $(\theta_k)_{1 \leq k \leq L}$ are usually chosen as (realizations of) independent and identically distributed (i.i.d.) random variables, which typically follow a uniform or Gaussian distribution on \mathbb{R}^p . Accordingly, the sequence $(h_k)_{0 \leq k \leq L}$ that results from the recursion (2) for a given input to the network takes the form of a sequence of random variables that are not i.i.d. but are actually a martingale. Thus, denoting informally by \mathcal{L} the differentiable loss associated with the learning task (classification or regression), the distributions of $(h_k)_{0 \leq k \leq L}$ and $(\frac{\partial \mathcal{L}}{\partial h_k})_{0 \leq k \leq L}$ as L becomes large carry useful information on the stability of training. For instance, exploding gradients in the backpropagation phase of learning correspond to the fact that, with high probability, $\|\frac{\partial \mathcal{L}}{\partial h_0}\| \gg \|\frac{\partial \mathcal{L}}{\partial h_L}\|$, where $\|\cdot\|$ denotes the Euclidean norm. Our first contribution, in Section 2, is to provide thorough mathematical statements on the behavior of these distributions (both for finite and infinite L), depending on the value of α_L . Among other results, we show that only the choice $\alpha_L \approx 1/\sqrt{L}$ yields a non-trivial behavior at initialization, thereby confirming empirical findings in the literature (Arpit et al., 2019; De and Smith, 2020). For $\alpha_L \gg 1/\sqrt{L}$, the norms explode exponentially fast with L , which is inappropriate for training. For $\alpha_L \ll 1/\sqrt{L}$, the network is almost equivalent to identity, that is, $h_L \approx h_0$. The analysis of the different cases as a function of α_L is mathematically involved and makes extensive use of concentration tools from random matrix theory.

The continuous approach. As noticed by several authors (Chen et al., 2018; Thorpe and van Gennip, 2018; E et al., 2019), model (2) with a scaling factor $\alpha_L = 1/L$ (and not $1/\sqrt{L}$) is formally similar to the discretization of a differential equation. Thus, when L tends to infinity, the weights and hidden states change continuously with the layer according to the equation

$$\frac{dH_t}{dt} = f(H_t, \Theta_t), \quad t \in [0, 1]. \quad (3)$$

Here, time t is the continuous analogue of the layer index k , $H : [0, 1] \rightarrow \mathbb{R}^d$ is a continuous-time hidden state, and $\Theta : [0, 1] \rightarrow \mathbb{R}^p$ a continuous-time parameter. This important connection between ResNets and differential equations has been identified in the past years under the umbrella name of neural ODE. Since the original article of Chen et al. (2018), this point of view has led to the development of a variety of new continuous-time models, together with innovative architectures and efficient training algorithms (Chang et al., 2019; Grathwohl et al., 2019; Kidger et al., 2021). The neural ODE paradigm also enabled to leverage the rich theory of differential equations to better understand the mechanisms at work behind deep ResNets (E et al., 2019; Fermanian et al., 2021). However, there is a debated question in the neural ODE community about the choice $\alpha_L = 1/L$, which guarantees convergence of the discrete model (2) to its continuous-time counterpart (3). As a matter of fact, it seems

that this choice is guided by more mathematical than practical considerations, and several authors have suggested that it is inconsistent with what is done in practice (Cohen et al., 2021; Bayer et al., 2022). Moreover, letting $\alpha_L = 1/L$ is somewhat contradictory with the results discussed above, which highlighted that the only non-trivial limit at initialization is $\alpha_L = 1/\sqrt{L}$. Thus, as a second contribution, we clarify the problem in Section 3 by leveraging our previous results on stability. We show that the value $\alpha_L = 1/\sqrt{L}$ corresponds in the continuous world to a neural stochastic differential equation (SDE) of the form (3), where now $\Theta : [0, 1] \rightarrow \mathbb{R}^p$ takes the form of a continuous-time stochastic process, typically a Brownian motion. By contrast, we also prove that the neural ODE regime with $\alpha_L = 1/L$ corresponds to the limit of a ResNet, not with i.i.d. weights as considered before, but with more complex and correlated weight distributions. For these weight distributions, the scaling $\alpha_L = 1/L$ is also a critical value between explosion and identity.

Going further, our third contribution is to exhibit in Section 4 a continuous range of regimes that are controlled by the choice of α_L (beyond the cases $1/\sqrt{L}$ and $1/L$) and the distribution of $(\theta_k)_{1 \leq k \leq L}$ at initialization, derived from a continuous-time process Θ with a regularity different from a Brownian motion. More precisely, we show experimentally that there is a strong interplay (with the same three cases—explosion, identity mapping, non-trivial behavior) between the choice of α_L and the regularity of $(\theta_k)_{1 \leq k \leq L}$ as a function of the layer index k . In addition, empirical evidence suggests that this interplay impacts both the behavior and performance of the networks during training, beyond initialization.

1.3 Related work

The choice of scaling for ResNets has been discussed in many papers, without however reaching a clear consensus on the form this scaling factor should take. For instance, Hanin and Rolnick (2018) state that stability requires $\alpha_L \leq 1/L$, while Zhang et al. (2019b) show that $\alpha_L \leq 1/\sqrt{L}$ is enough to ensure stability. On the other hand, Cohen et al. (2021) claim that the scaling factor observed in practice in trained ResNets is of the form $1/L^\beta$ with $\beta \approx 0.7$. Other authors have proposed more complex choices for α_L (e.g., Zhang et al., 2019a; Shao et al., 2020). Taking another point of view, De and Smith (2020) observe that batch normalization is empirically equivalent to taking a $1/\sqrt{L}$ normalization factor. Bachlechner et al. (2021) suggest to learn a scaling parameter α_k that is allowed to vary from one layer to another, whereas, in (4), α_L is kept constant across layers. These authors observe a great acceleration for training compared to traditional ResNets with no scaling. They also suggest a similar architecture for Transformers and then notice that $\alpha_k \approx 1/L$ at the end of training.

Closest to our analysis at initialization are the papers of Arpit et al. (2019) and Zhang et al. (2019b). Arpit et al. (2019) develop a theoretical analysis based on mean field approximation that suggests that a scaling factor $\alpha_L = 1/\sqrt{L}$ prevents vanishing/exploding gradients at initialization, and provide experimental evidence that this approach is competitive with batch normalization. However, the authors do not provide rigorous mathematical statements for the three different cases $\alpha_L \ll 1/\sqrt{L}$, $\alpha_L \approx 1/\sqrt{L}$, and $\alpha_L \gg 1/\sqrt{L}$, nor do they highlight the connection with the continuous-time interpretation. Interestingly, the idea of exploiting the martingale structure to analyze the magnitude of the hidden states is present in Zhang et al. (2019b), who study the convergence of gradient descent for over-parameterized ResNets

with different values of α_L . Nevertheless, they consider a specific model with Gaussian weights, and only provide asymptotic results when both width and depth tend to infinity.

The connection between the choice of scaling and the continuous-time point of view has previously been noticed by Zhang et al. (2019c), then studied in detail by Cohen et al. (2021). The latter show that, under assumptions on the form of the weights, it is possible to derive limiting (stochastic or ordinary) differential equations for the hidden states. However, they do not discuss the transition between these two regimes, nor do they link differential equations regimes with the stability of the network.

2. Scaling at initialization

Our goal in this section is to study the effect of the scaling factor α_L on the stability of ResNets at initialization, assuming that the weights are i.i.d. random variables. We start by making more precise the model and the learning problem introduced in (1).

2.1 Model and assumptions

Model. The data is a sample of n pairs $(x_i, y_i)_{1 \leq i \leq n}$, where x_i is the input vector in $\mathbb{R}^{n_{\text{in}}}$ and $y_i \in \mathbb{R}^{n_{\text{out}}}$ is the output vector to be predicted. This setting includes regression and classification (after one-hot encoding of the labels). Specifying the informal recurrence (1), for any input $x \in \mathbb{R}^{n_{\text{in}}}$, we consider the output $F_\pi(x) \in \mathbb{R}^{n_{\text{out}}}$ of the ResNet model defined by

$$\begin{aligned} h_0 &= Ax, \\ h_{k+1} &= h_k + \alpha_L V_{k+1} g(h_k, \theta_{k+1}), \quad 0 \leq k \leq L-1, \\ F_\pi(x) &= Bh_L, \end{aligned} \tag{4}$$

where $\alpha_L > 0$ is the scaling factor of the ResNet and $\pi = (A, B, (\theta_k)_{1 \leq k \leq L}, (V_k)_{1 \leq k \leq L})$ are its parameters, with $A \in \mathbb{R}^{d \times n_{\text{in}}}$, $B \in \mathbb{R}^{n_{\text{out}} \times d}$, $\theta_k \in \mathbb{R}^p$ and $V_k \in \mathbb{R}^{d \times d}$ for $k = 1, \dots, L$. The almost-everywhere differentiable function $g : \mathbb{R}^d \times \mathbb{R}^p \rightarrow \mathbb{R}^d$ encodes the choice of architecture. We note that the model includes initial and final linear layers in order to map the input space $\mathbb{R}^{n_{\text{in}}}$ into the space of hidden states \mathbb{R}^d , and symmetrically to map the last hidden state h_L into the output space $\mathbb{R}^{n_{\text{out}}}$. These two transformations are of little interest to us, since we mostly focus on the behavior of the sequence of hidden states $(h_k)_{0 \leq k \leq L}$. Let us finally notice that the results of this section can be adapted to hidden layers that do not have the same width, at the cost of increased technicality.

An important feature of model (4) is that the layer function takes the form of a matrix-vector multiplication, which will prove crucial to make use of concentration results on random matrices. We stress that this setting is standard in practice and that it encompasses many different types of ResNets. It includes for example simple ResNets where $g(h, \theta) = \sigma(h)$ with σ the activation function, and the original ResNets from He et al. (2016a), which have

$$g(h, \theta) = \text{ReLU}(Wh + b),$$

where the parameter is a pair $\theta = (W, b)$ with $W \in \mathbb{R}^{d \times d}$ a weight matrix and $b \in \mathbb{R}^d$ a bias, and $\text{ReLU}: x \mapsto \max(x, 0)$ is applied element-wise. This setting also includes attention layers, where g corresponds to the scaled dot-product between keys and queries, as well as

convolutional layers. Although the assumptions we make later have to be slightly modified to cover this context, the rationale should extend. We leave this extension for future work.

Throughout the article, we let $\ell : \mathbb{R}^{n_{\text{out}}} \times \mathbb{R}^{n_{\text{out}}} \rightarrow \mathbb{R}_+$ be a loss function, differentiable w.r.t. its first parameter, for example the squared loss or the cross-entropy loss. The objective of learning is to find the optimal parameter π that minimizes the empirical risk $\mathcal{L}(\pi) = \sum_{i=1}^n \ell(F_\pi(x_i), y_i)$.

Probabilistic setting at initialization. The minimization of the empirical risk is usually performed by stochastic gradient descent or one of its variants (Goodfellow et al., 2016, Chapter 8). The gradient descent is initialized by choosing the weights as (realizations of) i.i.d. random variables. The parameters $\theta_1, V_1, \dots, \theta_L, V_L$ in model (4) are therefore assumed to be an i.i.d. collection of random variables, where we recall that $\theta_k \in \mathbb{R}^p$ and $V_k \in \mathbb{R}^{d \times d}$ parameterize the k -th layer of the network. In this stochastic context, the successive hidden states h_0, \dots, h_L given a fixed input x are also random variables, but their distribution is not i.i.d.—in fact, under our assumptions, this sequence is a martingale. To avoid unnecessary technicalities, we assume that the sequence $(h_k)_{0 \leq k \leq L}$ is non-atomic. This is for example the case if the distribution of the parameters is absolutely continuous w.r.t. the Lebesgue measure. In particular, this ensures that the sequence $(h_k)_{0 \leq k \leq L}$ almost surely does not hit the non-differentiability points of g .

It is stressed that the distribution of the parameters are assumed to be independent of the depth, so that all the dependence on L is captured in the scaling factor α_L . This model enables us to consider multiple architectures at once, via the function g . By contrast, some authors formulate the problem of scaling as a choice of the variance at initialization (e.g., Yang and Schoenholz, 2017; Wang et al., 2022), which makes the analysis architecture-dependent. However, for a given architecture, these two approaches are essentially equivalent since $\text{Var}(\alpha_L V_k) = \alpha_L^2 \text{Var}(V_k)$.

The quantity $\|h_L - h_0\|/\|h_0\|$ carries key information on the behavior of the network at initialization. On the one hand, if $\|h_L - h_0\| \ll \|h_0\|$, the network is essentially equal to the identity function. On the other hand, if $\|h_L - h_0\| \gg \|h_0\|$, the output of the network explodes. An intermediate situation is when $\|h_L - h_0\| \approx \|h_0\|$. In addition, another source of information is provided by the gradients of the hidden states with respect to the empirical risk \mathcal{L} . If $\|\frac{\partial \mathcal{L}}{\partial h_0} - \frac{\partial \mathcal{L}}{\partial h_L}\| \ll \|\frac{\partial \mathcal{L}}{\partial h_L}\|$, the gradients do not change as they flow through the network, which means that the exact same information is backpropagated throughout the network. Conversely, if $\|\frac{\partial \mathcal{L}}{\partial h_0} - \frac{\partial \mathcal{L}}{\partial h_L}\| \gg \|\frac{\partial \mathcal{L}}{\partial h_L}\|$, the gradients explode during backpropagation. By exploiting the martingale structure of $(\|h_k\|)_{0 \leq k \leq L}$, as well as state-of-the-art concentration inequalities for random matrices with sub-Gaussian entries, we provide in this section probabilistic bounds on the magnitude of these various quantities.

Assumptions. Some assumptions are needed on the choices of architecture and initialization. Recall that a real-valued random variable X is said to be s^2 sub-Gaussian (van Handel, 2016, Chapter 3) if for all $\lambda \in \mathbb{R}$, $\mathbb{E}(\exp(\lambda X)) \leq \exp(\lambda^2 s^2/2)$. The sub-Gaussian property is a constraint on the tail of the probability distribution. As an example, Gaussian random variables on the real line are sub-Gaussian and so are bounded random variables.

The following assumptions will be needed throughout the section: for any $1 \leq k \leq L$,

- (A₁) For some $s \geq 1$, the entries of $\sqrt{d}V_k$ are symmetric i.i.d. s^2 sub-Gaussian random variables, independent of d and L , with unit variance.

(A₂) For some $C > 0$, independent of d and L , and for any $h \in \mathbb{R}^d$,

$$\frac{\|h\|^2}{2} \leq \mathbb{E}(\|g(h, \theta_k)\|^2) \leq \|h\|^2 \quad \text{and} \quad \mathbb{E}(\|g(h, \theta_k)\|^8) \leq C\|h\|^8.$$

Assumption (A₁) is mild and satisfied by all initializations used in practice. For example, the classical Glorot initialization (Glorot and Bengio, 2010)—which is the default implementation in the Keras package (Chollet et al., 2015)—takes the entries of V_{k+1} as uniform $\mathcal{U}(-\sqrt{3/d}, \sqrt{3/d})$ variables. This means that $\sqrt{d}V_{k+1}$ is initialized with $\mathcal{U}(-\sqrt{3}, \sqrt{3})$ random variables, which satisfy (A₁). Other examples include the Gaussian $\mathcal{N}(0, 1/d)$ initialization of He et al. (2015) and, for example, initialization with Rademacher variables.

The first part of Assumption (A₂) ensures that $g(\cdot, \theta_{k+1})$ is not too far away from being an isometry in expectation. The second part is more technical and, roughly, allows to upper bound the deviations of the norm of $g(h_k, \theta_{k+1})$. Our next Proposition 1 shows that most classical ResNet architectures verify Assumption (A₂). For the sake of readability, these models, together with their parameters, are summarized in Table 1 below.

	Name	Recurrence relation	Parameters
res-1	Simple ResNet	$h_{k+1} = h_k + \alpha_L V_{k+1} \sigma(h_k)$	$\theta_{k+1} = \emptyset$
res-2	Parametric ResNet	$h_{k+1} = h_k + \alpha_L V_{k+1} \sigma(W_{k+1} h_k)$	$\theta_{k+1} = W_{k+1}$
res-3	Classical ResNet	$h_{k+1} = h_k + \alpha_L V_{k+1} \text{ReLU}(W_{k+1} h_k)$	$\theta_{k+1} = W_{k+1}$

Table 1: Examples of ResNet architectures considered in the paper. In the first two cases, the activation function σ is such that, for all $x \in \mathbb{R}$, $a|x| \leq |\sigma(x)| \leq b|x|$, $1/\sqrt{2} \leq a < b \leq 1$. In the last two cases, $W_{k+1} \in \mathbb{R}^{d \times d}$.

Proposition 1 *Let res-1, res-2, and res-3 be the models defined in Table 1. Then*

- (i) *Assumption (A₂) is satisfied for res-1.*
- (ii) *Assumption (A₂) is satisfied for res-2 and res-3, as soon as the entries of $\sqrt{d}W_{k+1}$, $0 \leq k \leq L - 1$, are symmetric i.i.d. sub-Gaussian random variables, independent of d and L , with unit variance.*

In the models **res-1** and **res-2**, σ can be, for instance, taken as the parametric ReLU function, i.e., $\sigma(x) = x_+ + sx_-$, where x_+ (resp. x_-) denotes the positive (resp. negative) part and the slope $s \in [1/\sqrt{2}, 1]$ is a parameter of the model. Observe also that **res-2** differs from **res-3** since the classical ReLU function is defined by $\text{ReLU}(x) = x_+$ and thus does not satisfy the condition $|\sigma(x)| \geq a|x|$. Note that there is no bias term in these three models, as this term is commonly initialized to zero, and we are interested in the behavior at initialization.

2.2 Probabilistic bounds on the norm of the hidden states

The next two propositions describe how the quantity $\|h_L - h_0\|/\|h_0\|$ changes as a function of $L\alpha_L^2$. Proposition 2 provides a high-probability bound of interest when $L\alpha_L^2 \ll 1$. In this

case, we see that, with high probability, the network acts as the identity function, directly mapping h_0 to h_L . On the other hand, Proposition 3 provides information in the two cases $L\alpha_L^2 \gg 1$ and $L\alpha_L^2 \approx 1$. When $L\alpha_L^2 \gg 1$, the lower bound (i) indicates an explosion with high probability of the norm of the last hidden state. On the other hand, when $L\alpha_L^2 \approx 1$, the bounds (i) and (ii) show that h_L randomly varies around h_0 with fluctuation sizes bounded from below and above.

Proposition 2 Consider a ResNet (4) such that Assumptions (A₁) and (A₂) are satisfied. If $L\alpha_L^2 \leq 1$, then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$\frac{\|h_L - h_0\|^2}{\|h_0\|^2} \leq \frac{2L\alpha_L^2}{\delta}.$$

Proposition 3 Consider a ResNet (4) such that Assumptions (A₁) and (A₂) are satisfied.

(i) Assume that $d \geq 64$ and $\alpha_L^2 \leq \frac{2}{(\sqrt{C}s^4 + 4\sqrt{C+16s^4})d}$. Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$\frac{\|h_L - h_0\|^2}{\|h_0\|^2} > \exp\left(\frac{3L\alpha_L^2}{8} - \sqrt{\frac{11L\alpha_L^2}{d\delta}}\right) - 1,$$

provided that

$$2L \exp\left(-\frac{d}{64\alpha_L^2 s^2}\right) \leq \frac{\delta}{11}. \quad (5)$$

(ii) Assume that $\alpha_L^2 \leq \frac{1}{\sqrt{C}(d+128s^4)}$. Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$\frac{\|h_L - h_0\|^2}{\|h_0\|^2} < \exp\left(L\alpha_L^2 + \sqrt{\frac{5L\alpha_L^2}{d\delta}}\right) + 1.$$

Note that the assumptions of Proposition 3 on d and α_L are mild, since in the learning tasks where deep ResNets are involved, one typically has $\alpha_L = 1/L^\beta$ with $\beta > 0$, $d \geq 10^2$ and $L \geq 10^2$. Note also that condition (5) is not severe since, when d and L are large, it encompasses all reasonable values of δ . Propositions 2 and 3 are interesting in the sense that they provide finite-depth high-probability bounds on the behavior of the hidden states, depending on the magnitude of $L\alpha_L^2$. The results become clearer by letting $\alpha_L = 1/L^\beta$, with $\beta > 0$, as shown in the following corollary.

Corollary 4 Consider a ResNet (4) such that Assumptions (A₁) and (A₂) are satisfied, and let $\alpha_L = 1/L^\beta$, with $\beta > 0$.

(i) If $\beta > 1/2$, then

$$\frac{\|h_L - h_0\|}{\|h_0\|} \xrightarrow[L \rightarrow \infty]{\mathbb{P}} 0.$$

(ii) If $\beta < 1/2$ and $d \geq 9$, then

$$\frac{\|h_L - h_0\|}{\|h_0\|} \xrightarrow[L \rightarrow \infty]{\mathbb{P}} \infty.$$

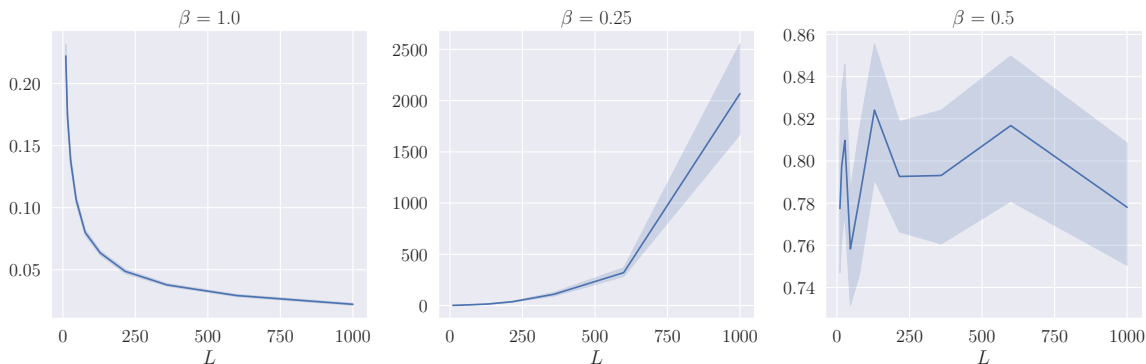


Figure 1: Evolution of $\|h_L - h_0\|/\|h_0\|$ as a function of L for different values of β and an i.i.d. $\mathcal{U}(-\sqrt{3/d}, \sqrt{3/d})$ initialization of model **res-3**, with $d = 40$. The input is a random Gaussian observation x in dimension $n_{\text{in}} = 64$. The experiment is repeated with 50 independent randomizations.

(iii) If $\beta = 1/2$, $d \geq 64$, $L \geq (\frac{1}{2}\sqrt{C}s^4 + 2\sqrt{C} + 8s^4)d + 96\sqrt{C}s^4$, then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$\exp\left(\frac{3}{8} - \sqrt{\frac{22}{d\delta}}\right) - 1 < \frac{\|h_L - h_0\|^2}{\|h_0\|^2} < \exp\left(1 + \sqrt{\frac{10}{d\delta}}\right) + 1,$$

provided that

$$2L \exp\left(-\frac{Ld}{64s^2}\right) \leq \frac{\delta}{11}.$$

Corollary 4 highlights three different asymptotic behaviors for $\|h_L\|$, depending on the values of β . For $\beta > 1/2$, statement (i) tells that h_L converges towards h_0 in probability, as L tends to infinity, which means that the neural network is essentially equivalent to an identity mapping. On the other hand, for $\beta < 1/2$, the norm of h_L explodes with high probability. Finally, for the critical value $\beta = 1/2$, we see that h_L fluctuates around h_0 , with a fluctuation size independent of L . Observe that the lower bound in (iii) is not trivial as soon as $\exp(3/8 - \sqrt{11/d\delta}) > 1$, i.e., $d > 99/64\delta$. The message of Corollary 4 is that the only scaling leading to a non-degenerate distribution at initialization is for $\beta = 1/2$.

The three statements of Corollary 4 are illustrated in Figure 1. In this experiment, we consider model **res-3**, a random Gaussian observation x in dimension $n_{\text{in}} = 64$, and parameters initialized with a uniform distribution $\mathcal{U}(-\sqrt{3/d}, \sqrt{3/d})$. We refer to Appendix E for a detailed setup of all the experiments of the paper. Figure 2a shows the empirical distribution of $\|h_L\|/\|h_0\|$ when $\beta = 1/2$ for a large number of realizations. This figure illustrates in particular that our bounds are reasonably sharp, since the bounds indicate that the first quartile of the distribution is larger than 0.87 (whereas the first quartile of the empirical histogram is equal to 1.21) and the third quartile is less than 2.06 (whereas the third quartile of the empirical histogram is equal to 1.34). Determining the exact distribution of $\|h_L\|/\|h_0\|$ is an interesting avenue for future research that is beyond the scope of the present article. There is however a strong indication that the ratio follows a log-normal distribution, as confirmed by a normality test on (the log of) the empirical distribution.

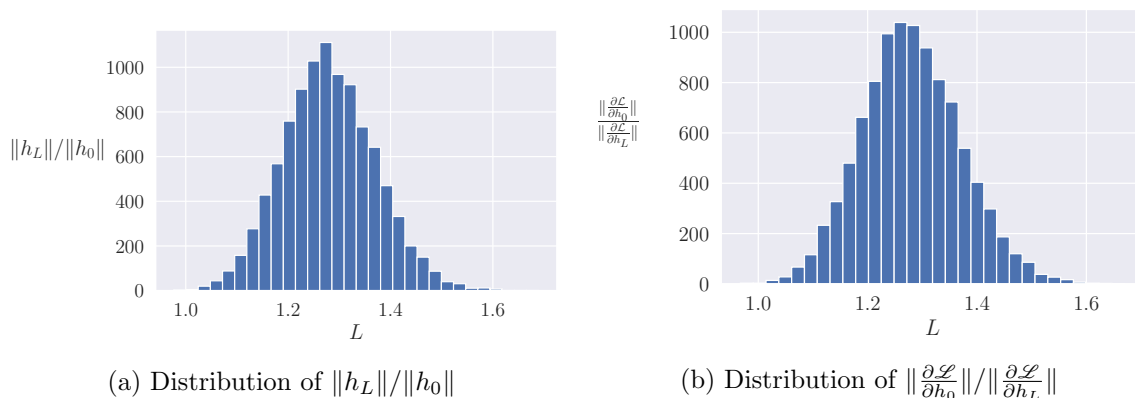


Figure 2: Empirical distributions of the norms for $\beta = 1/2$, $L = 10^3$, $d = 100$. The experiment is repeated with 10^4 independent randomizations.

In a nutshell, the proofs of Propositions 2 and 3 rest upon controlling of the norm of the hidden states, which obeys the recurrence

$$\|h_{k+1}\|^2 = \|h_k\|^2 + \alpha_L^2 \|V_{k+1}g(h_k, \theta_{k+1})\|^2 + 2\alpha_L \langle h_k, V_{k+1}g(h_k, \theta_{k+1}) \rangle, \quad (6)$$

where $\langle \cdot, \cdot \rangle$ denotes the standard scalar product in \mathbb{R}^d . Taking the expectations on both side, one deduces with Assumptions (A_1) and (A_2) that

$$\begin{aligned} \mathbb{E}(\|V_{k+1}g(h_k, \theta_{k+1})\|^2) &= \mathbb{E}\left(\mathbb{E}(\|V_{k+1}g(h_k, \theta_{k+1})\|^2) \mid h_k, \theta_{k+1}\right) \\ &= \mathbb{E}(\|g(h_k, \theta_{k+1})\|^2) \approx \|h_k\|^2 \end{aligned} \quad (7)$$

and

$$\mathbb{E}(\langle h_k, V_{k+1}g(h_k, \theta_{k+1}) \rangle) = \mathbb{E}\left(\mathbb{E}(\langle h_k, V_{k+1}g(h_k, \theta_{k+1}) \rangle \mid h_k, \theta_{k+1})\right) = 0. \quad (8)$$

The equalities (7) and (8) allow deriving without further work bounds in expectation on $\|h_L\|$, as already observed by Arpit et al. (2019). However, the results we are after are stronger since they involve high-probability bounds. A finer control of the deviations of $\|V_{k+1}g(h_k, \theta_{k+1})\|^2$ and $\langle h_k, V_{k+1}g(h_k, \theta_{k+1}) \rangle$ is then needed. This involves concentration inequalities on random matrices with sub-Gaussian entries.

2.3 Probabilistic bounds on the gradients

Propositions 2 and 3 provide insights on the output of the network when L is large. However, they do not carry information on the backwards dynamics of propagation of the gradients of the loss $p_k = \frac{\partial \mathcal{L}}{\partial h_k} \in \mathbb{R}^d$. Assessing the dynamics of the $(p_k)_{0 \leq k \leq L}$ as a function of L is important since the behavior of this sequence impacts trainability of the network at initialization. Thus, in this subsection, we are interested in quantifying the magnitude of $\|p_0 - p_L\|/\|p_L\|$, when L is large. Notice that, contrarily to the previous subsection where we were mostly interested in the last hidden state h_L , the quantity of interest is now p_0 (not p_L), the gradient at index 0. The reason is that the sequence $(p_k)_{0 \leq k \leq L}$ is defined backwardly, as we will see below. We also stress that $(p_k)_{0 \leq k \leq L}$ is the sequence of derivatives of the loss

w.r.t. the hidden states h_k , and not w.r.t. the parameters. The reason for considering this sequence is that the p_k are involved in the backpropagation algorithm and are therefore essential for assessing the stability of the gradient descent (see, e.g., Arpit et al., 2019).

Analyzing the behavior of the sequence $(p_k)_{0 \leq k \leq L}$ is challenging since, according to the backpropagation (or reverse-mode differentiation) formula, one has

$$p_k = p_{k+1} + \alpha_L \frac{\partial g(h_k, \theta_{k+1})^\top}{\partial h} V_{k+1}^\top p_{k+1}.$$

Taking the norm,

$$\|p_k\|^2 = \|p_{k+1}\|^2 + \alpha_L^2 \left\| \frac{\partial g(h_k, \theta_{k+1})^\top}{\partial h} V_{k+1}^\top p_{k+1} \right\|^2 + 2\alpha_L \left\langle p_{k+1}, \frac{\partial g(h_k, \theta_{k+1})^\top}{\partial h} V_{k+1}^\top p_{k+1} \right\rangle.$$

Although the equation looks qualitatively similar to (6), it has the unpleasant feature that $\frac{\partial g(h_k, \theta_{k+1})}{\partial h}$ depends on h_k , hence on $\theta_1, V_1, \dots, \theta_k, V_k$, while p_{k+1} depends on $\theta_{k+2}, V_{k+2}, \dots, \theta_{L-1}, V_{L-1}$. This forbids applying directly the same proof techniques as for the hidden states. Therefore, to extract useful information from this recurrence equation, one needs to characterize the dependence of the distribution of $\frac{\partial g(h_k, \theta_{k+1})}{\partial h}$ with respect to h_k . To do so, it is sometimes assumed that these two quantities are independent (see, e.g., Yang and Schoenholz, 2017). However, assuming independence remains a strong requirement, which is not verified for many network architectures (for example model `res-1`). We tackle the problem from a different point of view and propose an alternative approach based on forward-mode differentiation, valid under a much weaker assumption. The cost we pay is that we obtain results in expectation and not in high probability.

Let us sketch our approach before stating the results more formally. Let $z \in \mathbb{R}^d$, and, for any $0 \leq i, j \leq L$, let $\frac{\partial h_j}{\partial h_i} \in \mathbb{R}^{d \times d}$ be the Jacobian matrix of h_j with respect to h_i . Recall that the (m, n) -th entry of this matrix equals the derivative of the m -th coordinate of h_j w.r.t. the n -th coordinate of h_i . Then, letting $q_k(z) = \frac{\partial h_k}{\partial h_0} z$, we have, by the chain rule,

$$q_{k+1}(z) = \frac{\partial h_{k+1}}{\partial h_k} q_k(z) = q_k(z) + \alpha_L V_{k+1} \frac{\partial g(h_k, \theta_{k+1})}{\partial h} q_k(z). \quad (9)$$

Identity (9), which is similar to (4), expresses $q_{k+1}(z)$ as a function of $q_k(z)$, and therefore respects the flow of information. Next, assuming that z is random with a Gaussian distribution, it is possible to express one of our quantities of interest, $\|p_0\|/\|p_L\|$, as a function of the last vector $q_L(z)$. Indeed,

$$\frac{\|p_0\|^2}{\|p_L\|^2} = \frac{1}{\|p_L\|^2} \mathbb{E}_{z \sim \mathcal{N}(0, I_d)} \left(|p_0^\top z|^2 \right) = \mathbb{E}_{z \sim \mathcal{N}(0, I_d)} \left(\left| \left(\frac{p_L}{\|p_L\|} \right)^\top q_L(z) \right|^2 \right), \quad (10)$$

where I_d is the identity matrix in \mathbb{R}^d and the second equality is a consequence of

$$p_0^\top z = \left(\frac{\partial \mathcal{L}}{\partial h_0} \right)^\top z = \left(\frac{\partial \mathcal{L}}{\partial h_L} \right)^\top \frac{\partial h_L}{\partial h_0} z = p_L^\top q_L(z).$$

In summary, the recurrence (9) allows us to derive bounds on the norm of $q_L(z)$, which can then transfer to $\|p_0\|/\|p_L\|$ via (10). For this, it is necessary to make the following assumption on the ratio $p_L/\|p_L\|$:

(A₃) Let $b = p_L/\|p_L\|$. Then $\mathbb{E}(b|h_L) = 0$ and $\mathbb{E}(b^\top b|h_L) = I_d/d$.

It is a mild assumption, which is verified for instance if $n_{\text{out}} = 1$ with squared error (for regression) or cross-entropy (for binary classification). In these cases, $p_L/\|p_L\| = B^\top/\|B\|_F$, where $\|\cdot\|_F$ is the Frobenius norm and B is the weight matrix of the last layer. We finally need the following assumption, which is the equivalent of Assumption (A₂) for the gradients.

(A₄) One has, almost surely,

$$\frac{\|q_k\|^2}{2} \leq \mathbb{E}\left(\left\|\frac{\partial g(h_k, \theta_{k+1})}{\partial h} q_k\right\|^2 \middle| h_k, q_k\right) \leq \|q_k\|^2.$$

Assumption (A₄) is satisfied by all the standard architectures listed in Table 1, as shown by the next proposition.

Proposition 5 *Let res-1, res-2, and res-3 be the models defined in Table 1. Assume that (A₁) is satisfied and σ is almost everywhere differentiable, with $a \leq \sigma' \leq b$. Then*

- (i) *Assumption (A₄) is satisfied for res-1.*
- (ii) *Assumption (A₄) is satisfied for res-2 and res-3, when the entries of $\sqrt{d}W_{k+1}, 0 \leq k \leq L-1$, are symmetric i.i.d. random variables, independent of d and L , with unit variance.*

The next two propositions are the counterparts of Proposition 2 and Proposition 3 for the gradient dynamics.

Proposition 6 *Consider a ResNet (4) such that Assumptions (A₁)-(A₄) are satisfied. If $L\alpha_L^2 \leq 1$, then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,*

$$\frac{\|p_0 - p_L\|^2}{\|p_L\|^2} \leq \frac{2L\alpha_L^2}{\delta}.$$

Proposition 7 *Consider a ResNet (4) such that Assumptions (A₁)-(A₄) are satisfied. Then*

$$\left(1 + \frac{1}{2}\alpha_L^2\right)^L - 1 \leq \mathbb{E}\left(\frac{\|p_0 - p_L\|^2}{\|p_L\|^2}\right) \leq (1 + \alpha_L^2)^L - 1.$$

A simple corollary of the propositions above is as follows.

Corollary 8 *Consider a ResNet (4) such that Assumptions (A₁)-(A₄) are satisfied, and take $\alpha_L = 1/L^\beta$, with $\beta > 0$. Then*

- (i) *If $\beta > 1/2$,*

$$\frac{\|p_0 - p_L\|}{\|p_L\|} \xrightarrow[L \rightarrow \infty]{\mathbb{P}} 0.$$

- (ii) *If $\beta < 1/2$,*

$$\mathbb{E}\left(\frac{\|p_0 - p_L\|^2}{\|p_L\|^2}\right) \xrightarrow[L \rightarrow \infty]{} \infty.$$

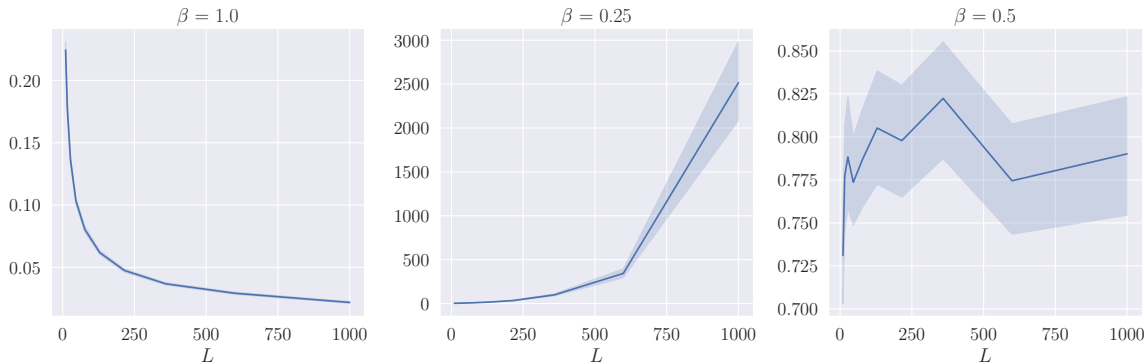


Figure 3: Evolution of $\|p_0 - p_L\|/\|p_L\|$ as a function of L for different values of β and an i.i.d. $\mathcal{U}(-\sqrt{3/d}, \sqrt{3/d})$ initialization of model `res-3`, with $d = 40$. The input is a random Gaussian observation x in dimension $n_{\text{in}} = 64$. The experiment is repeated with 50 independent randomizations.

(ii) If $\beta = 1/2$,

$$\exp\left(\frac{1}{2}\right) - 1 \leq \mathbb{E}\left(\frac{\|p_0 - p_L\|^2}{\|p_L\|^2}\right) \leq \exp(4) - 1.$$

Corollary 8 is illustrated in Figure 3. The experimental protocol is the same as in Figure 1, but we now track p_0 and p_L , the gradients of the loss \mathcal{L} with respect to the first and the last hidden states. In accordance with our results, when $\beta > 1/2$, the gradient remains the same from one layer to another (left plot). On the other hand, the middle plot clearly shows that when $\beta < 1/2$ the gradient explodes. Once again, the case $\beta = 1/2$ (right plot) is the only one for which the distribution of gradients at initialization is non-trivial. Figure 2b illustrates that the empirical distribution of gradients in this case also seems to be log-normal.

In summary, this and the previous subsection both point towards the same conclusion: there are three different cases, depending on the value of β —explosion when $\beta < 1/2$, non-degenerate limit when $\beta = 1/2$, and identity when $\beta > 1/2$. In the explosion case, it is well known that the network cannot be trained (Yang and Schoenholz, 2017). The theory thus points out that the value $1/2$ plays a pivotal role. Remarkably, this value has a specific interpretation in the continuous-time point of view of ResNets, in terms of SDE. This is the topic that we address in the next section.

3. Scaling in the continuous-time setting

Starting with the discrete ResNet (4), it is tempting to let L go to infinity and consider the network as the discretization of a differential equation where the layer index $k \in \{0, \dots, L\}$ is replaced by the time index $t \in [0, 1]$. This interpretation of deep neural networks has been popularized by Chen et al. (2018) and is often referred to as the neural ODE paradigm. Notice that this setting is different from the so-called mean-field analysis, where the width of the network is assumed to be infinite beforehand. In our setting, the width d is assumed to be finite and fixed.

3.1 Convergence towards an SDE in the large-depth regime

One of the main messages of Section 2 is that the standard initialization with i.i.d. parameters leads to a non-degenerate model for large values of L only if $L\alpha_L^2 \approx 1$ (Propositions 2 and 3), or, equivalently, if $\beta = 1/2$ when $\alpha_L = 1/L^\beta$ (Corollary 4). Remarkably, in the continuous-time limit, this regime corresponds to the discretization of an SDE. Indeed, consider for simplicity the (discrete) ResNet model `res-1`

$$h_{k+1} = h_k + \frac{1}{\sqrt{L}} V_{k+1} \sigma(h_k), \quad 0 \leq k \leq L-1, \quad (11)$$

where the entries of V_{k+1} are assumed to be i.i.d. $\mathcal{N}(0, 2/d)$. Recall the following definition:

Definition 9 *A one-dimensional Brownian motion $(B_t)_{t \in [0,1]}$ is a continuous-time stochastic process with $B_0 = 0$, almost surely continuous, with independent increments, and such that for any $0 \leq s < t \leq 1$, $B_t - B_s \sim \mathcal{N}(0, t - s)$.*

Now, let $\mathbf{B} : [0, 1] \rightarrow \mathbb{R}^{d \times d}$ be a $(d \times d)$ -dimensional Brownian motion, in the sense that the $(B_{ij})_{1 \leq i, j \leq d}$ are independent one-dimensional Brownian motions. Thus, for any $0 \leq k \leq L-1$ and any $1 \leq i, j \leq d$, we have

$$\mathbf{B}_{(k+1)/L, i, j} - \mathbf{B}_{k/L, i, j} \sim \mathcal{N}\left(0, \frac{1}{L}\right),$$

and the increments for different values of (i, j, k) are independent. As a consequence, the recurrence (11) is equivalent in distribution to the recurrence

$$h_{k+1}^\top = h_k^\top + \sqrt{\frac{2}{d}} \sigma(h_k^\top) (\mathbf{B}_{(k+1)/L} - \mathbf{B}_{k/L}), \quad 0 \leq k \leq L-1.$$

(Note that this is true because V_{k+1} has the same distribution as V_{k+1}^\top .) We recognize the Euler-Maruyama discretization (Kloeden and Platen, 1992) on the $\{k/L, 0 \leq k \leq L\}$ mesh of the SDE

$$dH_t^\top = \sqrt{\frac{2}{d}} \sigma(H_t^\top) d\mathbf{B}_t, \quad t \in [0, 1], \quad (12)$$

where the output of the network is now a function of the final value of H , that is, H_1 . The link between the discrete ResNet (11) and the SDE (12) is formalized in the next proposition.

Proposition 10 *Consider the `res-1` model, where the entries of V_{k+1} are i.i.d. Gaussian $\mathcal{N}(0, 2/d)$ random variables. Assume that the activation function σ is Lipschitz continuous. Then the SDE (12) has a unique solution H and, for any $0 \leq k \leq L$,*

$$\mathbb{E}(\|H_{k/L} - h_k\|) \leq \frac{c}{\sqrt{L}},$$

for some $c > 0$.

Notice that the requirement that σ is Lipschitz continuous is satisfied by most classical activation functions, including ReLU. This proposition is interesting for several reasons. First, the scaling $\beta = 1/2$, which is exactly the one that yields a non-trivial dynamics at initialization, corresponds in the continuous world to a remarkably ‘simple’ model of diffusion. This shows that very deep neural networks properly initialized with i.i.d. weights are equivalent to solutions of SDE. This analogy opens interesting perspectives for training deep networks using automatic differentiation for solutions of neural SDE (Li et al., 2020).

Second, we stress that the emergence of an SDE instead of an ODE carries an important message. Several authors (including, e.g., Thorpe and van Gennip, 2018) have shown that, under appropriate assumptions, a deep ResNet converges in the large depth limit to an ODE and not an SDE. The reason why we obtain an SDE here is intrinsically connected with the choice of i.i.d. initialization for the weights, which makes a Brownian motion appear at the limit, as highlighted above. In other words, the i.i.d. initialization, the choice $\beta = 1/2$ (the relevant critical value exhibited in Section 2), and the emergence of an SDE are intimately linked together. On the other hand, the case $\beta = 1$ matches with an ODE if the initialization is not i.i.d., as we will see in Subsection 3.2.

Finally, we point out that Proposition 10 states the convergence of a ResNet towards an SDE for the basic architecture **res-1** and for Gaussian initialization. The extension to more general settings is an interesting direction of research, although clearly beyond the scope of the present paper (see, e.g., Peluchetti and Favaro, 2020, and Cohen et al., 2021, for results in this direction).

3.2 Scaling in the neural ODE setting

Convergence towards an ODE. The basic message of our Proposition 10 is that an i.i.d. initialization, together with $\beta = 1/2$, leads to an SDE rather than an ODE. A natural question is then whether a different choice of weight distributions (at initialization) and scaling can lead to a classical neural ODE.

To answer this question and leave the world of i.i.d. initialization, we assume that the weights $(V_k)_{1 \leq k \leq L}$ and $(\theta_k)_{1 \leq k \leq L}$ are discretizations of smooth functions $\mathcal{V} : [0, 1] \rightarrow \mathbb{R}^{d \times d}$ and $\Theta : [0, 1] \rightarrow \mathbb{R}^p$. We then consider the general iteration (4) with $\alpha_L = 1/L$, that is,

$$h_0 = Ax, \quad h_{k+1} = h_k + \frac{1}{L} V_{k+1} g(h_k, \theta_{k+1}), \quad 0 \leq k \leq L - 1, \quad (13)$$

where $V_k = \mathcal{V}_{k/L}$ and $\theta_k = \Theta_{k/L}$. Of course, it is still possible to consider $(V_k)_{1 \leq k \leq L}$ (resp. $(\theta_k)_{1 \leq k \leq L}$) as random variables, by letting $(\mathcal{V}_t)_{t \in [0,1]}$ (resp. $(\Theta_t)_{t \in [0,1]}$) be a continuous-time stochastic process. In this model, we shall need the following assumption:

(A₅) For any $0 \leq k \leq L - 1$, one has $V_k = \mathcal{V}_{k/L}$ and $\theta_k = \Theta_{k/L}$, where the stochastic processes \mathcal{V} and Θ are almost surely Lipschitz continuous and bounded.

More precisely, almost surely, there exist $K_{\mathcal{V}}, K_{\Theta}, C_{\mathcal{V}}, C_{\Theta} > 0$, such that, for any $s, t \in [0, 1]$,

$$\|\mathcal{V}_t - \mathcal{V}_s\| \leq K_{\mathcal{V}} |t - s|, \quad \|\Theta_t - \Theta_s\| \leq K_{\Theta} |t - s|, \quad \|\mathcal{V}_t\| \leq C_{\mathcal{V}}, \quad \|\Theta_t\| \leq C_{\Theta}.$$

A typical model that satisfies Assumption (A₅) is obtained by letting the entries of \mathcal{V} and Θ be independent Gaussian processes with expectation zero and squared exponential covariance $K(x, x') = \exp(-\frac{(x-x')^2}{2\ell^2})$, where $\ell > 0$.

We shall also need the following requirement on g , which is satisfied by all our models as soon as σ is Lipschitz continuous:

(A₆) The function g is Lipschitz continuous on compact sets, in the sense that for any compact $\mathcal{D} \subseteq \mathbb{R}^p$, there exists $K_{\mathcal{D}} > 0$ such that, for all $h, h' \in \mathbb{R}^d$, $\theta \in \mathcal{D}$,

$$\|g(h, \theta) - g(h', \theta)\| \leq K_{\mathcal{D}} \|h - h'\|,$$

and for any compact $\mathcal{D} \subseteq \mathbb{R}^d$, there exists $K_{\mathcal{D}, \mathcal{D}} > 0$ such that, for all $h \in \mathcal{D}$, $\theta, \theta' \in \mathcal{D}$,

$$\|g(h, \theta) - g(h, \theta')\| \leq K_{\mathcal{D}, \mathcal{D}} \|\theta - \theta'\|.$$

Under Assumptions (A₅) and (A₆), the recurrence (13) almost surely converges towards the neural ODE given by

$$H_0 = Ax, \quad dH_t = \mathcal{V}_t g(H_t, \Theta_t) dt, \quad t \in [0, 1], \quad (14)$$

as shown by the proposition below.

Proposition 11 *Consider model (13) such that Assumptions (A₅) and (A₆) are satisfied. Then the ODE (14) has a unique solution H , and, almost surely, there exists some $c > 0$ such that, for any $0 \leq k \leq L$,*

$$\|H_{k/L} - h_k\| \leq \frac{c}{L}.$$

It should be stressed that the transition from the discrete recurrence (13) to the continuous-time differential equation (14) relies on the assumptions that the weight sequences $(\theta_k)_{1 \leq k \leq L}$ and $(V_k)_{1 \leq k \leq L}$ are the discretizations of smooth limiting processes Θ and \mathcal{V} on the one hand, and that the scaling α_L is chosen as $1/L$ on the other hand. From a practical perspective, Proposition 11 shows that it is possible to initialize ResNets in the ODE regime, by choosing a smooth stochastic process, discretizing it at each layer, and taking a $1/L$ scaling. This is in sharp contrast with the results of Sections 2 and 3.1, which show that the usual i.i.d. procedure leads to a neural SDE.

Stability and scaling. Assuming that the weights of the network are discretizations of a smooth function (Assumption (A₅)), it is possible to obtain stability results, depending on the value of β , similarly to what has been done in Section 2. We show below that $\beta = 1$ is a critical value, by examining the hidden states, in the same way as $\beta = 1/2$ is a critical value in the i.i.d. setting. Similar results can be shown for the gradients. We begin by a proposition handling the cases $\beta > 1$ and $\beta = 1$.

Proposition 12 *Consider a ResNet (4) such that Assumptions (A₅) and (A₆) are satisfied. Let $\alpha_L = 1/L^\beta$, with $\beta > 0$.*

(i) *If $\beta > 1$, then, almost surely,*

$$\frac{\|h_L - h_0\|}{\|h_0\|} \xrightarrow{L \rightarrow \infty} 0.$$

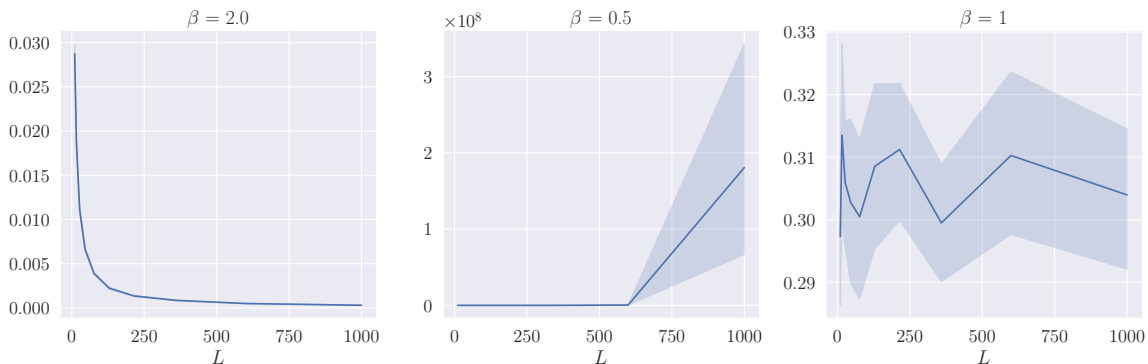


Figure 4: Evolution of $\|h_L - h_0\|/\|h_0\|$ as a function of L for different values of β and a smooth initialization of model **res-3**, with $d = 40$. The input is a random Gaussian observation x in dimension $n_{\text{in}} = 64$. The experiment is repeated with 50 independent randomizations.

(ii) If $\beta = 1$, then, almost surely, there exists some $c > 0$ such that

$$\frac{\|h_L - h_0\|}{\|h_0\|} \leq c.$$

The explosion case ($\beta < 1$) is more delicate to deal with. We prove it for a linear model, and leave for future work the extension to more general cases.

Proposition 13 *Consider the **res-1** model, taking σ as the identity function. Assume that Assumption (A₅) is satisfied and that \mathcal{V}_0^T has a positive eigenvalue. Let $\alpha_L = 1/L^\beta$, with $\beta \in (0, 1)$. Then, almost surely,*

$$\max_k \frac{\|h_k - h_0\|}{\|h_0\|} \xrightarrow{L \rightarrow \infty} \infty.$$

The assumption of the existence of a positive eigenvalue for \mathcal{V}_0^T is mild. For instance, if the entries of \mathcal{V}_0 are i.i.d. random variables with finite moments of all order, [Götze and Jalowy \(2021\)](#) show that such an eigenvalue exists with probability at least $1 - 1/d$ for d large enough.

In this setting, we observe experimentally a behavior of the output and of the gradients when L grows large similar to the one explored in Section 2. This is illustrated in Figures 4 and 5, which mirror Figures 1 and 3 in Section 2. The figures clearly show that there exist three cases for the output and for the gradients: an identity case (left plots), an explosion case (middle), and a non-trivial case separating explosion and identity (right). However, the remarkable point is that the separation occurs for $\beta = 1$, and not $\beta = 1/2$, as predicted by Propositions 12 and 13.

4. Experiments

We experimentally investigate in this section two questions. The first one is to know whether there exists a range of scaling factors $\beta > 0$ and weight initializations, beyond the i.i.d. and

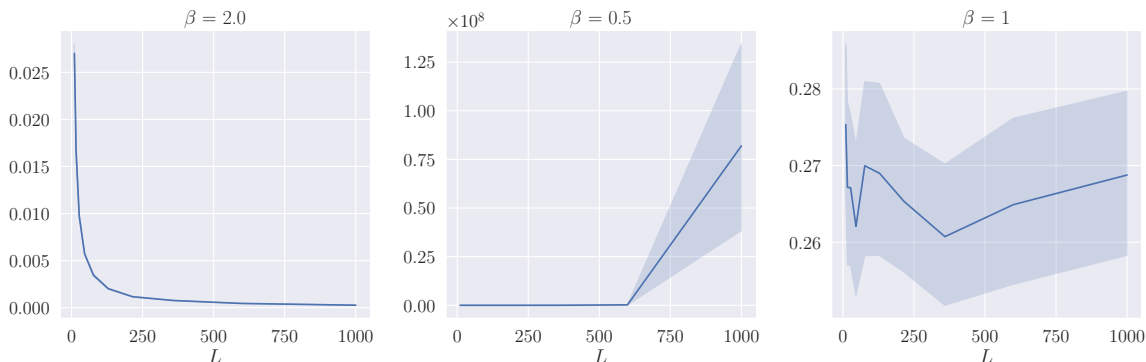


Figure 5: Evolution of $\|p_0 - p_L\|/\|p_L\|$ as a function of L for different values of β and a smooth initialization of model **res-3**, with $d = 40$. The input is a random Gaussian observation x in dimension $n_{\text{in}} = 64$. The experiment is repeated with 50 independent randomizations.

the smooth regimes. The second question is whether our analysis, which pertains to the initialization phase, provides insights into the training phase, beyond initialization.

4.1 Intermediate regimes

In order to describe the transition between the i.i.d. and smooth cases, a possible route is to consider that the weights are increments of a γ -Hölder stochastic process. This model is interesting insofar as the Brownian motion (SDE regime) is $(1/2 - \varepsilon)$ -Hölder ($\varepsilon > 0$) and a Lipschitz continuous stochastic process (ODE regime) is 1-Hölder.

In line with the above, in a series of experiments, we initialize the weights as increments of a fractional Brownian motion $(B_t^H)_{t \in [0,1]}$. Recall that B^H is a continuous-time Gaussian process, starting at zero, with zero expectation for all $t \in [0, 1]$, and covariance function

$$\mathbb{E}(B_s^H B_t^H) = \frac{1}{2}(|s|^{2H} + |t|^{2H} - |t - s|^{2H}), \quad 0 \leq s, t \leq 1,$$

where $H \in (0, 1)$ is called the Hurst index. This index describes the raggedness of the process, with a higher value leading to a smoother process. When $H = 1/2$, the process is a standard Brownian motion (Definition 9), whose increments are independent by construction. When $H > 1/2$, the increments of the process are positively correlated, while if $H < 1/2$ the increments are negatively correlated. Importantly, a fractional Brownian motion with Hurst index H is $(H - \varepsilon)$ -Hölder continuous for any $\varepsilon > 0$. In the limit when $H \rightarrow 1$, the trajectories converge to linear functions (whose increments satisfy (A_5)). As an illustration, Figure 6 depicts three realizations of a fractional Brownian motion with $H = 0.2$ (left), $H = 0.5$ (middle), and $H = 0.8$ (right).

In order to assess the effect of the scaling factor β and the Hurst index H , we initialize a neural network **res-3** with $d = 40$, $L = 1000$, various values of $\beta \in [0.2, 1.3]$, and with weights taken as increments of fractional Brownian motions with various Hurst indices $H \in (0, 1)$. Figure 7 depicts the empirical magnitude of the output and the gradients at initialization as a function of the Hurst index H and the scaling factor β . First note that we

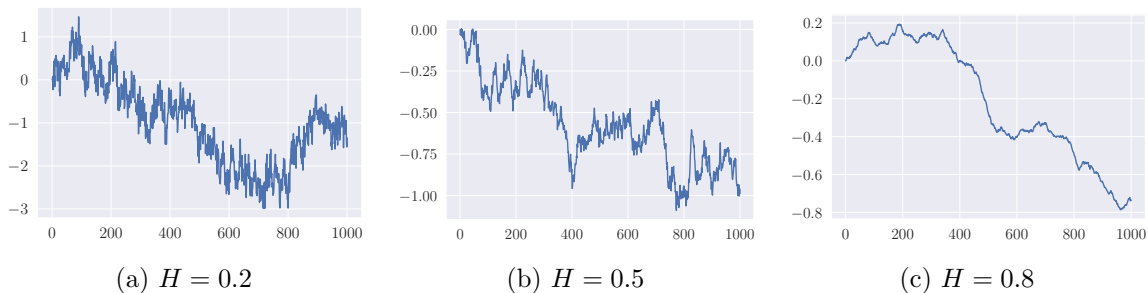


Figure 6: Examples of realizations of a fractional Brownian motion B^H for different Hurst indexes H . Note that the smaller the value of H , the more irregular the trajectory is.

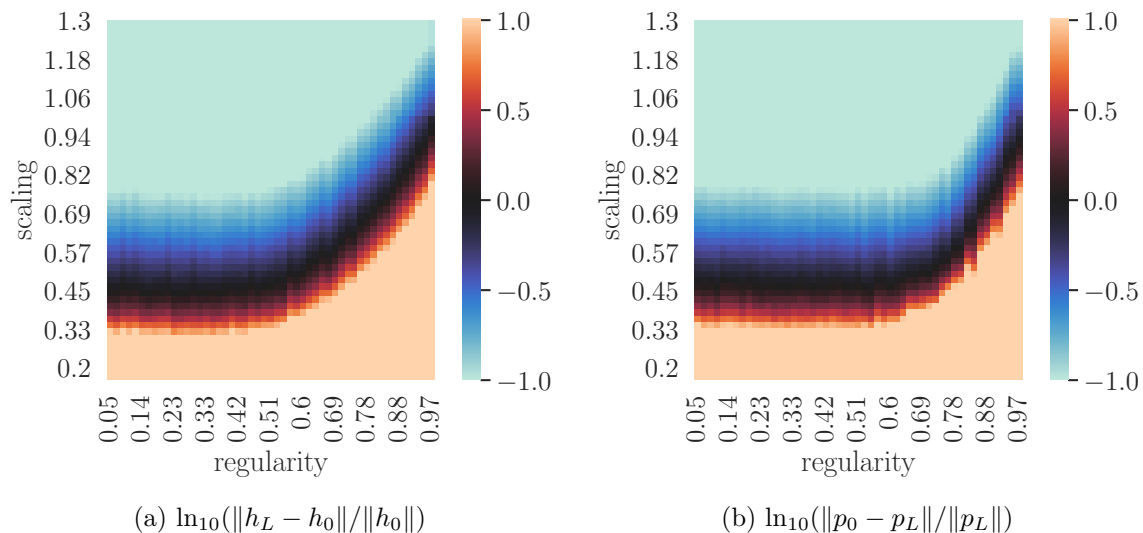


Figure 7: Magnitude of the outputs and of the gradients as a function of the regularity of the weights (Hurst index H) and of the scaling factor β . The orange zone corresponds to the explosion regime, i.e., $\|h_L - h_0\| \gg \|h_0\|$ and $\|p_0 - p_L\| \gg \|p_L\|$. The blue zone corresponds to the identity regime, i.e., $\|h_L - h_0\| \ll \|h_0\|$ and $\|p_0 - p_L\| \ll \|p_L\|$. Finally, the black zone is an intermediate regime, where $\|h_L - h_0\| \approx \|h_0\|$ and $\|p_0 - p_L\| \approx \|p_L\|$.

recover the two regimes (i.i.d. and smooth) discussed so far. For $H = 1/2$, the i.i.d. regime kicks in, with explosion ($\beta < 1/2$, orange zone), non-trivial behavior ($\beta = 1/2$, black zone), and identity ($\beta > 1/2$, blue zone). Likewise, we see at $H = 1$ a similar pattern in the smooth regime, with, as predicted by Proposition 12, a critical value $\beta = 1$. Beyond these two specific cases, we observe for an index H varying in $(1/2, 1)$ a whole range of intermediate situations, where the transition between identity and explosion seems to happen for a critical $\beta = H$. Interestingly, for $H < 1/2$, the transition seems to saturate at the value $\beta = 1/2$.

The take-home message is that the choice of the scaling of a ResNet seems to be closely linked to the regularity of the weights as a function of the layer. More precisely, for all regimes, the critical scaling factor between explosion and identity seems to have a natural interpretation as the (Hölder) regularity of the underlying continuous-time stochastic process.

We believe that the mathematical understanding of this connection, beyond the fractional Brownian motion case, is a promising research direction for the future. Finally, these experiments suggest that it is sensible to initialize a ResNet for any value of the scaling $\beta \in (1/2, 1)$, while avoiding the identity and explosion situations, by simulating a fractional Brownian motion of Hurst index $H = \beta$ and initializing the weights as the increments of this process.

4.2 Beyond initialization

At initialization, before the gradient descent, the distribution of the weights $(\theta_k)_{1 \leq k \leq L}$ and $(V_k)_{1 \leq k \leq L}$ is chosen by the practitioner. By contrast, during and after training, control is lost on these distributions, making the picture more complex. In particular, the existence and characterization of a continuous-time stochastic process whose discretization matches the trained ResNet is an interesting but difficult problem. Attacking this question requires a fine understanding of the interaction between training dynamics and the regularity of the sequence of the weights during the gradient descent. However, there is experimental evidence that the trained weights exhibit strong structure as a function of the layer index k (Cohen et al., 2021; Bayer et al., 2022), and that their regularity strongly depends on the choice of initialization. Figure 8 depicts this mechanism by plotting a given coordinate of θ_k as a function of the layer index k ranging from 1 to the depth $L = 1000$, after training.

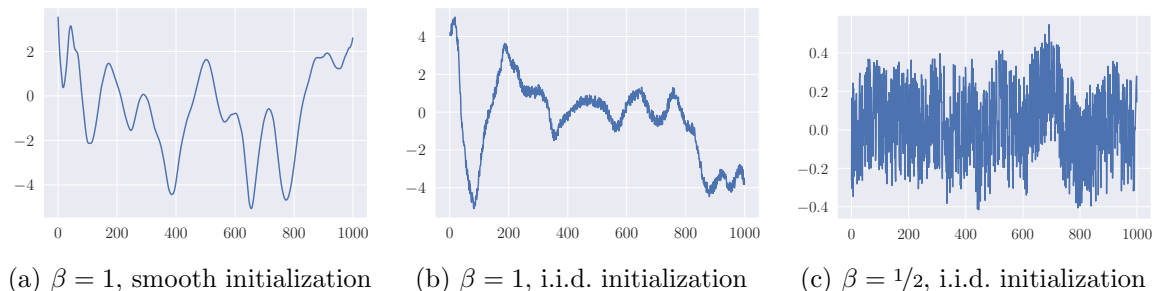


Figure 8: Plot of a given coordinate of θ_k , after training, as a function of the layer index k ranging from 1 to the depth $L = 1000$ for three different choices of β and initializations.

To investigate the link between regularity of the weights at initialization, scaling, and performance after training, we train ResNets on the datasets MNIST (Deng, 2012) and CIFAR-10 (Krizhevsky, 2009). As in Subsection 4.1, we initialize the ResNets with various scaling factors and weights that are increments of fractional Brownian motions with different regularities. Then, for each combination of weight initialization and scaling factor, the ResNet is trained using the Adam optimizer (Kingma and Ba, 2015) for 10 epochs. The results in terms of accuracy are presented in Figure 9 (light orange = good performance, blue = bad performance). We observe a pattern similar to the one of Figure 7, however shifted downwards. This means that, for a given regularity, the network is unable to learn if it is initialized with a scaling too far below the critical value, which of course is connected with the gradient explosion issue discussed previously. On the other hand, and perhaps more surprisingly, the performance seems to be more or less stable in the identity region, with perhaps a small degradation in the case of CIFAR-10. This somewhat contrasts with

the results from [Yang and Schoenholz \(2017\)](#), who exhibit a decrease in performance for i.i.d. initialization and a large scaling factor β . Note however that, in our experiments, we adapted the learning rate of the gradient descent on a grid by cross-validation. This was done to prevent a slowdown in training when the scaling factor β is too large, since, in the gradient descent, the gradients updates are also scaled by the factor $1/L^\beta$ —which gets smaller as β increases. The interplay between the learning rate and the scaling factor is one of the keys to better assess how the performance of the trained network is connected with the scaling.

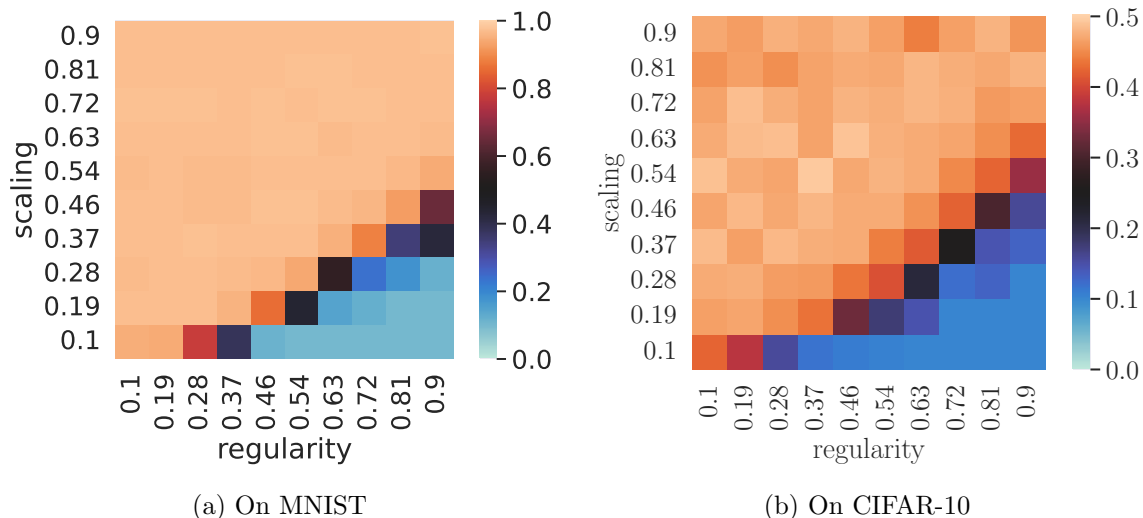


Figure 9: Accuracy after training as a function of the regularity of the weights at initialization and scaling. For each point of the heatmap, the model was trained on a grid of learning rates, and the best performance is shown.

Acknowledgments

The authors thank S. Schoenholz for fruitful discussion. P. Marion has been supported by a grant from Région Île-de-France and by a Google PhD Fellowship award.

A. Proofs

Throughout the proofs, the i -th coordinate of a vector v is denoted by v_i . Similarly, the i -th row of a matrix M is denoted by M_i , and its (i, j) -th entry by M_{ij} .

A.1 Proof of Proposition 1

Statement (i) is clear (with $C = 1$) since, for any $h \in \mathbb{R}^d$,

$$\|\sigma(h)\|^2 \in [a^2\|h\|^2, b^2\|h\|^2] \subseteq [\frac{1}{2}\|h\|^2, \|h\|^2].$$

With respect to statement (ii), it is enough to show that for any $h \in \mathbb{R}^d$ and any random matrix W satisfying the assumptions of the proposition, one has

$$\frac{\|h\|^2}{2} \leq \mathbb{E}(\|\sigma(Wh)\|^2) \leq \|h\|^2 \quad \text{and} \quad \mathbb{E}(\|\sigma(Wh)\|^8) \leq C\|h\|^8,$$

as well as

$$\mathbb{E}(\|\text{ReLU}(Wh)\|^2) = \frac{\|h\|^2}{2} \quad \text{and} \quad \mathbb{E}(\|\text{ReLU}(Wh)\|^8) \leq C\|h\|^8.$$

The two claims with the squared norms are consequences of Lemmas 16 and 17 in Appendix B, together with the fact that the variance of the entries of W equals $1/d$. In order to prove the other two statements, first note that $\mathbb{E}(\|\sigma(Wh)\|^8) \leq \mathbb{E}(\|Wh\|^8)$ and $\mathbb{E}(\|\text{ReLU}(Wh)\|^8) \leq \mathbb{E}(\|Wh\|^8)$. The results are then consequences of Lemma 21 in Appendix C, which states that

$$\mathbb{E}\left(\frac{\|Wh\|^8}{\|h\|^8}\right) \leq 1 + \frac{384s^4}{d} + \frac{3072s^6}{d^2} \leq 1 + 384s^4 + 3072s^6.$$

A.2 Proof of Proposition 2

According to Lemma 14 below, one has

$$\mathbb{E}\left(\frac{\|h_L - h_0\|^2}{\|h_0\|^2}\right) \leq ((1 + \alpha_L^2)^L - 1).$$

But, for $L\alpha_L^2 \leq 1$, we have $(1 + \alpha_L^2)^L - 1 \leq \exp(L\alpha_L^2) - 1 \leq 2L\alpha_L^2$. Therefore,

$$\mathbb{E}\left(\frac{\|h_L - h_0\|^2}{\|h_0\|^2}\right) \leq 2L\alpha_L^2,$$

and the result follows from Markov's inequality.

Lemma 14 *Consider a ResNet (4) such that Assumptions (A₁) and (A₂) are satisfied. Then*

$$\left(\left(1 + \frac{\alpha_L^2}{2}\right)^L - 1\right) \leq \mathbb{E}\left(\frac{\|h_L - h_0\|^2}{\|h_0\|^2}\right) \leq \left((1 + \alpha_L^2)^L - 1\right).$$

Proof (Lemma 14) Taking the squared norm of the forward update rule (4) and dividing by $\|h_0\|^2$ yields

$$\frac{\|h_{k+1}\|^2}{\|h_0\|^2} = \frac{1}{\|h_0\|^2} \left(\|h_k\|^2 + \alpha_L^2 \|V_{k+1}g(h_k, \theta_{k+1})\|^2 + 2\alpha_L \langle h_k, V_{k+1}g(h_k, \theta_{k+1}) \rangle \right). \quad (15)$$

We deduce by Assumptions (A₁) and (A₂) that

$$\left(1 + \frac{\alpha_L^2}{2}\right) \mathbb{E}\left(\frac{\|h_k\|^2}{\|h_0\|^2}\right) \leq \mathbb{E}\left(\frac{\|h_{k+1}\|^2}{\|h_0\|^2}\right) \leq (1 + \alpha_L^2) \mathbb{E}\left(\frac{\|h_k\|^2}{\|h_0\|^2}\right).$$

Therefore, by recurrence, we are led to

$$\left(1 + \frac{\alpha_L^2}{2}\right)^k \leq \mathbb{E}\left(\frac{\|h_k\|^2}{\|h_0\|^2}\right) \leq (1 + \alpha_L^2)^k. \quad (16)$$

Now, observe that $h_L = h_0 + \alpha_L \sum_{k=0}^{L-1} V_{k+1}g(h_k, \theta_{k+1})$. Thus, we have

$$\mathbb{E}\left(\frac{\|h_L - h_0\|^2}{\|h_0\|^2}\right) = \alpha_L^2 \sum_{k,k'=0}^{L-1} \mathbb{E}\left(\frac{g(h_k, \theta_{k+1})^\top V_{k+1}^\top V_{k'+1} g(h_{k'}, \theta_{k'+1})}{\|h_0\|^2}\right).$$

By conditioning on all random variables except $V_{k'+1}$ for $k < k'$ (and V_{k+1} for $k > k'$), it is easy to see that the only non-zero terms are when $k = k'$. This yields

$$\begin{aligned} \mathbb{E}\left(\frac{\|h_L - h_0\|^2}{\|h_0\|^2}\right) &= \alpha_L^2 \sum_{k=0}^{L-1} \mathbb{E}\left(\frac{\|V_{k+1}g(h_k, \theta_{k+1})\|^2}{\|h_0\|^2}\right) \\ &\leq \alpha_L^2 \sum_{k=0}^{L-1} \mathbb{E}\left(\frac{\|h_k\|^2}{\|h_0\|^2}\right) \\ &\quad \text{(by Assumptions } A_1 \text{ and } A_2) \\ &\leq \alpha_L^2 \sum_{k=0}^{L-1} (1 + \alpha_L^2)^k \\ &\quad \text{(by (16))} \\ &= ((1 + \alpha_L^2)^L - 1). \end{aligned}$$

Similarly,

$$\begin{aligned} \mathbb{E}\left(\frac{\|h_L - h_0\|^2}{\|h_0\|^2}\right) &\geq \frac{\alpha_L^2}{2} \sum_{k=0}^{L-1} \mathbb{E}\left(\frac{\|h_k\|^2}{\|h_0\|^2}\right) \\ &= \left(\left(1 + \frac{\alpha_L^2}{2}\right)^L - 1\right). \end{aligned}$$

■

A.3 Proof of Proposition 3

Dividing (15) by $\|h_k\|^2$ and taking the logarithm leads to

$$\ln(\|h_{k+1}\|^2) = \ln(\|h_k\|^2) + \ln\left(1 + \alpha_L^2 \frac{\|V_{k+1}g(h_k, \theta_{k+1})\|^2}{\|h_k\|^2} + 2\alpha_L \left\langle \frac{h_k}{\|h_k\|}, \frac{V_{k+1}g(h_k, \theta_{k+1})}{\|h_k\|} \right\rangle\right).$$

Let

$$Y_{k,1} = \alpha_L^2 \frac{\|V_{k+1}g(h_k, \theta_{k+1})\|^2}{\|h_k\|^2}, \quad Y_{k,2} = 2\alpha_L \left\langle \frac{h_k}{\|h_k\|}, \frac{V_{k+1}g(h_k, \theta_{k+1})}{\|h_k\|} \right\rangle,$$

and $Y_k = Y_{k,1} + Y_{k,2}$. The proof of Proposition 3 strongly relies on the following lemma, which provides technical information on the moments of $Y_{k,1}$ and $Y_{k,2}$. For the sake of clarity, its proof is postponed to Appendix B.

Lemma 15 *Assume that Assumptions (A₁) and (A₂) are satisfied. Then*

$$\begin{aligned}
 (E_1) \quad \mathbb{E}(Y_{k,2}|h_k) &= 0. & (E_5) \quad \mathbb{E}(Y_{k,2}^4|h_k) &\leq 2048 \frac{s^4 \alpha_L^4}{d^2}. \\
 (E_2) \quad \frac{\alpha_L^2}{2} &\leq \mathbb{E}(Y_{k,1}|h_k) \leq \alpha_L^2. & (E_6) \quad \mathbb{E}(Y_{k,1}^4|h_k) &\leq C \left(3072 \frac{s^6}{d^2} + 384 \frac{s^4}{d} + 1 \right) \alpha_L^8. \\
 (E_3) \quad \mathbb{E}(Y_{k,1} Y_{k,2}|h_k) &= 0. & (E_7) \quad \mathbb{E}(Y_{k,1}^2|h_k) &\leq \sqrt{C} \left(128 \frac{s^4}{d} + 1 \right) \alpha_L^4. \\
 (E_4) \quad \mathbb{E}(Y_{k,2}^2|h_k) &\leq 4 \frac{\alpha_L^2}{d}.
 \end{aligned}$$

For $c > 0$, we have

$$\begin{aligned}
 \mathbb{P}\left(\frac{\|h_L\|^2}{\|h_0\|^2} \geq c\right) &= \mathbb{P}\left(\ln(\|h_L\|^2) - \ln(\|h_0\|^2) \geq \ln(c)\right) \\
 &= \mathbb{P}\left(\sum_{k=0}^{L-1} \ln(1 + Y_k) \geq \ln(c)\right) \\
 &\leq \mathbb{P}\left(\sum_{k=0}^{L-1} Y_k \geq \ln(c)\right) \\
 &\quad (\text{using } \ln(1+x) \leq x \text{ for } x > -1).
 \end{aligned}$$

Let $S = \sum_{k=0}^{L-1} Y_k - \mathbb{E}(Y_k|h_k)$. By (E₁) and (E₂),

$$\sum_{k=0}^{L-1} \mathbb{E}(Y_k|h_k) \leq L\alpha_L^2.$$

So, for $c > \exp(L\alpha_L^2)$,

$$\begin{aligned}
 \mathbb{P}\left(\frac{\|h_L\|^2}{\|h_0\|^2} \geq c\right) &\leq \mathbb{P}\left(S \geq \ln(c) - \sum_{k=0}^{L-1} \mathbb{E}(Y_k|h_k)\right) \\
 &\leq \mathbb{P}(S \geq \ln(c) - L\alpha_L^2) \\
 &\leq \mathbb{P}(S^2 \geq (\ln(c) - L\alpha_L^2)^2) \\
 &\leq \frac{\mathbb{E}(S^2)}{(\ln(c) - L\alpha_L^2)^2} \tag{17} \\
 &\quad (\text{by Markov's inequality.})
 \end{aligned}$$

It remains to upper bound $\mathbb{E}(S^2)$. To this aim, note that

$$\begin{aligned}
 \mathbb{E}(S^2) &= \sum_{k=0}^{L-1} \mathbb{E}\left((Y_k - \mathbb{E}(Y_k|h_k))^2\right) \leq \sum_{k=0}^{L-1} \mathbb{E}(Y_k^2) \\
 &\leq 4 \frac{L\alpha_L^2}{d} + 128\sqrt{C} \frac{L\alpha_L^4 s^4}{d} + \sqrt{C} L\alpha_L^4 \\
 &\quad (\text{by (E}_3\text{), (E}_4\text{), and (E}_7\text{)}) \\
 &\leq 5 \frac{L\alpha_L^2}{d}.
 \end{aligned}$$

The last inequality is true for $\alpha_L^2 \leq \frac{1}{\sqrt{C}(d+128s^4)}$. Therefore, by inequality (17), we obtain, for $c > \exp(L\alpha_L^2)$,

$$\mathbb{P}\left(\frac{\|h_L\|^2}{\|h_0\|^2} \geq c\right) \leq \frac{5L\alpha_L^2}{d(\ln(c) - L\alpha_L^2)^2}.$$

We conclude that, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$\frac{\|h_L\|^2}{\|h_0\|^2} < \exp\left(L\alpha_L^2 + \sqrt{\frac{5L\alpha_L^2}{d\delta}}\right).$$

This shows statement (ii) of the proposition.

Next, to prove statement (i), observe that $c > 0$,

$$\begin{aligned} \mathbb{P}\left(\frac{\|h_L\|^2}{\|h_0\|^2} \leq c\right) &= \mathbb{P}\left(\ln(\|h_L\|^2) - \ln(\|h_0\|^2) \leq \ln(c)\right) \\ &= \mathbb{P}\left(\sum_{k=0}^{L-1} \ln(1 + Y_k) \leq \ln(c)\right) \\ &= \mathbb{P}\left(\sum_{k=0}^{L-1} \ln(1 + Y_k) \leq \ln(c) \text{ and } \forall k, Y_k \geq -\frac{1}{2}\right) \\ &\quad + \mathbb{P}\left(\sum_{k=0}^{L-1} \ln(1 + Y_k) \leq \ln(c) \text{ and } \exists k, Y_k < -\frac{1}{2}\right). \end{aligned}$$

Using the inequality $\ln(1 + x) \geq x - x^2$ for $x \geq -1/2$, we obtain

$$\begin{aligned} \mathbb{P}\left(\frac{\|h_L\|^2}{\|h_0\|^2} \leq c\right) &\leq \mathbb{P}\left(\sum_{k=0}^{L-1} Y_k - Y_k^2 \leq \ln(c) \text{ and } \forall k, Y_k \geq -\frac{1}{2}\right) \\ &\quad + \mathbb{P}\left(\sum_{k=0}^{L-1} \ln(1 + Y_k) \leq \ln(c) \text{ and } \exists k, Y_k < -\frac{1}{2}\right). \end{aligned}$$

Thus,

$$\mathbb{P}\left(\frac{\|h_L\|^2}{\|h_0\|^2} \leq c\right) \leq \mathbb{P}\left(\sum_{k=0}^{L-1} Y_k - Y_k^2 \leq \ln(c)\right) + \sum_{k=0}^{L-1} \mathbb{P}\left(Y_{k,2} < -\frac{1}{2}\right). \quad (18)$$

We handle the two terms above on the right-hand side separately. For the first term, let $Z_k = Y_k - Y_k^2$ and $S = \sum_{k=0}^{L-1} Z_k - \mathbb{E}(Z_k|h_k)$. Observe that, by (E1)-(E4) and (E7),

$$\sum_{k=0}^{L-1} \mathbb{E}(Z_k|h_k) \geq \frac{L\alpha_L^2}{2} - 4\frac{L\alpha_L^2}{d} - 128\sqrt{C}\frac{L\alpha_L^4 s^4}{d} - \sqrt{C}L\alpha_L^4 \geq \frac{3}{8}L\alpha_L^2, \quad (19)$$

where the last inequality is valid for $d \geq 64$ and $\alpha_L^2 \leq \frac{1}{16\sqrt{C}(2s^4+1)}$. Hence, for $0 < c < \exp(3L\alpha_L^2/8)$,

$$\begin{aligned} \mathbb{P}\left(\sum_{k=0}^{L-1} Y_k - Y_k^2 \leq \ln c\right) &= \mathbb{P}\left(S \leq \ln(c) - \sum_{k=0}^{L-1} \mathbb{E}(Z_k|h_k)\right) \\ &\leq \mathbb{P}\left(S \leq \ln(c) - \frac{3L\alpha_L^2}{8}\right) \\ &\leq \mathbb{P}\left(S^2 \geq \left(\ln(c) - \frac{3L\alpha_L^2}{8}\right)^2\right) \\ &\leq \frac{\mathbb{E}(S^2)}{\left(\ln(c) - \frac{3L\alpha_L^2}{8}\right)^2} \\ &\quad \text{(by Markov's inequality.)} \end{aligned}$$

Using the c_r -inequality $(a+b)^n \leq 2^{n-1}(a^n + b^n)$ respectively for $n=2$ and $n=4$, we see that

$$\begin{aligned} \mathbb{E}(S^2) &= \sum_{k=0}^{L-1} \mathbb{E}\left((Z_k - \mathbb{E}(Z_k|h_k))^2\right) \leq \sum_{k=0}^{L-1} \mathbb{E}(Z_k^2) \leq 2 \sum_{k=0}^{L-1} \mathbb{E}(Y_k^2) + \mathbb{E}(Y_k^4) \\ &\leq 2 \sum_{k=0}^{L-1} \mathbb{E}(Y_{k,1}^2) + \mathbb{E}(Y_{k,2}^2) + 2\mathbb{E}(Y_{k,1}Y_{k,2}) + 8\mathbb{E}(Y_{k,1}^4) + 8\mathbb{E}(Y_{k,2}^4). \end{aligned}$$

By (E_3) - (E_7) , it is easy to verify that, for $d \geq 64$ and $\alpha_L^2 \leq \frac{1}{(\sqrt{C}s^4/16+2\sqrt{C}+8s^4)d}$,

$$\mathbb{E}(S^2) \leq 10 \frac{L\alpha_L^2}{d}.$$

This shows that, for $c < \exp(3L\alpha_L^2/8)$,

$$\mathbb{P}\left(\sum_{k=0}^{L-1} Y_k - Y_k^2 \leq \ln(c)\right) \leq \frac{10L\alpha_L^2}{d\left(\ln(c) - \frac{3L\alpha_L^2}{8}\right)^2}.$$

To conclude the proof, it remains to upper bound the second term of inequality (18). According to inequality (22) in the proof of Lemma 15 (with $t=1/2$), one has

$$\sum_{k=0}^{L-1} \mathbb{P}\left(Y_{k,2} < -\frac{1}{2}\right) \leq 2L \exp\left(-\frac{d}{64\alpha_L^2 s^2}\right).$$

Putting everything together, we are led to

$$\mathbb{P}\left(\frac{\|h_L\|^2}{\|h_0\|^2} \leq c\right) \leq \frac{10L\alpha_L^2}{d\left(\ln(c) - \frac{3L\alpha_L^2}{8}\right)^2} + 2L \exp\left(-\frac{d}{64\alpha_L^2 s^2}\right).$$

Take $\delta \in (0, 1)$. Then, if $2L \exp\left(-\frac{d}{64\alpha_L^2 s^2}\right) \leq \frac{\delta}{11}$, with probability at least $1 - \delta$,

$$\frac{\|h_L\|^2}{\|h_0\|^2} > \exp\left(\frac{3L\alpha_L^2}{8} - \sqrt{\frac{11L\alpha_L^2}{d\delta}}\right).$$

Notice that this inequality is valid under the assumption $\alpha_L^2 \leq \frac{2}{(\sqrt{C}s^4+4\sqrt{C}+16s^4)d}$.

A.4 Proof of Corollary 4

Statement (i) is a consequence of Proposition 2, whereas (ii) is a consequence of Proposition 3 (i). The latter is valid under the conditions $d \geq 64$ and $\alpha_L \leq \frac{2}{(\sqrt{C}s^4 + 4\sqrt{C} + 16s^4)d}$, which is automatically satisfied for all L large enough. Furthermore, an inspection of the proof of Proposition 3 reveals that the divergence in high probability of $\|h_L\|$ can be proved under the relaxed assumption $d \geq 9$. Indeed, the main constraint on d comes from the lower bound (19), where one needs to make sure that $\frac{L\alpha_L^2}{2} - 4\frac{L\alpha_L^2}{d} > 0$, which is the case for $d = 9$.

To prove (iii), we use a union bound on both statements of Proposition 3.

A.5 Proof of Proposition 5

The first claim follows from the observation that

$$\frac{\partial g(h_k, \theta_{k+1})}{\partial h} q_k = \begin{pmatrix} \sigma'(h_{k,1}) & 0 & \dots & 0 \\ 0 & \sigma'(h_{k,2}) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma'(h_{k,d}) \end{pmatrix} q_k,$$

from (A₁), and from the assumption on σ' .

Let us now prove (ii). In the rest of the proof, the subscript k is ignored to lighten the notation. Observe that

$$\frac{\partial g(h, \theta)}{\partial h} q = V \begin{pmatrix} \sigma'(\langle W_1, h \rangle) & 0 & \dots & 0 \\ 0 & \sigma'(\langle W_2, h \rangle) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma'(\langle W_d, h \rangle) \end{pmatrix} Wq.$$

Denote by D the matrix in the middle of the right-hand side. Then

$$\mathbb{E} \left(\left\| \frac{\partial g(h, \theta)}{\partial h} q \right\|^2 \middle| h, q \right) = \mathbb{E} (\|VDWq\|^2 | h, q) = \mathbb{E} (\|DWq\|^2 | h, q) \\ \text{(by (A}_1\text{))}$$

For model **res-2**, we have

$$\mathbb{E} \left(\left\| \frac{\partial g(h, \theta)}{\partial h} q \right\|^2 \middle| h, q \right) = \mathbb{E} \left(\sum_{i=1}^d \left(\sum_{j=1}^d W_{ij} q_j \right)^2 \sigma'(\langle W_i, h \rangle) \middle| h, q \right).$$

The conclusion follows from the hypothesis that $a \leq \sigma' \leq b$ and $\mathbb{E}(\|Wq\|^2 | q) = \|q\|^2$. For model **res-3**, we have

$$\mathbb{E} \left(\left\| \frac{\partial g(h, \theta)}{\partial h} q \right\|^2 \middle| h, q \right) = \mathbb{E} \left(\sum_{i=1}^d \left(\sum_{j=1}^d W_{ij} q_j \right)^2 \mathbf{1}_{\sum_{j=1}^d W_{ij} h_j \geq 0} \middle| h, q \right).$$

Since the $(W_{ij})_{1 \leq i, j \leq d}$ are symmetric random variables, we conclude that

$$\mathbb{E} \left(\left\| \frac{\partial g(h, \theta)}{\partial h} q \right\|^2 \middle| h, q \right) = \frac{1}{2} \mathbb{E} \left(\sum_{i=1}^d \left(\sum_{j=1}^d W_{ij} q_j \right)^2 \middle| q \right) = \frac{1}{2} \mathbb{E} (\|Wq\|^2 | q) = \frac{\|q\|^2}{2}.$$

A.6 Proof of Proposition 6

Letting $b = p_L / \|p_L\|$, as in Assumption (A₃), and taking expectation in (10), we obtain

$$\mathbb{E}\left(\frac{\|p_0\|^2}{\|p_L\|^2}\right) = \mathbb{E}(|b^\top q_L(z)|^2) = \frac{1}{d}\mathbb{E}(\|q_L(z)\|^2) \quad (20)$$

(by (A₃)).

The rest of the proof is similar to the proof of Proposition 2. From (9), we have

$$\|q_{k+1}(z)\|^2 = \|q_k(z)\|^2 + \alpha_L^2 \left\| V_{k+1} \frac{\partial g(h_k, \theta_{k+1})}{\partial h} q_k(z) \right\|^2 + 2\alpha_L \left\langle q_k(z), V_{k+1} \frac{\partial g(h_k, \theta_{k+1})}{\partial h} q_k(z) \right\rangle.$$

By independence of V_{k+1} from $q_k(z)$ and $\frac{\partial g(h_k, \theta_{k+1})}{\partial h}$,

$$\mathbb{E}\left(\left\langle q_k(z), V_{k+1} \frac{\partial g(h_k, \theta_{k+1})}{\partial h} q_k(z) \right\rangle\right) = 0.$$

Next,

$$\begin{aligned} \mathbb{E}\left(\left\| V_{k+1} \frac{\partial g(h_k, \theta_{k+1})}{\partial h} q_k(z) \right\|^2\right) &= \mathbb{E}\left(\mathbb{E}\left(\left\| V_{k+1} \frac{\partial g(h_k, \theta_{k+1})}{\partial h} q_k(z) \right\|^2 \middle| h_k, \theta_{k+1}, q_k(z)\right)\right) \\ &= \mathbb{E}\left(\left\| \frac{\partial g(h_k, \theta_{k+1})}{\partial h} q_k(z) \right\|^2\right) \\ &\quad \text{(by (A₁))} \\ &= \mathbb{E}\left(\mathbb{E}\left(\left\| \frac{\partial g(h_k, \theta_{k+1})}{\partial h} q_k(z) \right\|^2 \middle| h_k, q_k(z)\right)\right). \end{aligned}$$

By Assumption (A₃), we are led to

$$\left(1 + \frac{1}{2}\alpha_L^2\right)\mathbb{E}(\|q_k(z)\|^2) \leq \mathbb{E}(\|q_{k+1}(z)\|^2) \leq (1 + \alpha_L^2)\mathbb{E}(\|q_k(z)\|^2),$$

and thus, by induction, since $q_0(z) = z$ and $\mathbb{E}(\|z\|^2) = d$,

$$d\left(1 + \frac{1}{2}\alpha_L^2\right)^k \leq \mathbb{E}(\|q_k(z)\|^2) \leq d(1 + \alpha_L^2)^k.$$

In particular, for $k = L$,

$$d\left(1 + \frac{1}{2}\alpha_L^2\right)^L \leq \mathbb{E}(\|q_L(z)\|^2) \leq d(1 + \alpha_L^2)^L.$$

Therefore, by (20),

$$\left(1 + \frac{1}{2}\alpha_L^2\right)^L \leq \mathbb{E}\left(\frac{\|p_0\|^2}{\|p_L\|^2}\right) \leq (1 + \alpha_L^2)^L.$$

To finish the proof, observe that

$$\frac{1}{\|p_L\|}(p_0 - p_L)^\top z = b^\top (q_L(z) - z).$$

Using arguments similar to (20), we may write

$$\mathbb{E}\left(\frac{\|p_0 - p_L\|^2}{\|p_L\|^2}\right) = \frac{1}{d}\mathbb{E}\left(\frac{\|q_L(z) - z\|^2}{\|z\|^2}\right).$$

Now, upon noting that $q_L(z) - z = q_L(z) - q_0(z) = \alpha_L \sum_{k=0}^{L-1} V_{k+1} \frac{\partial g(h_k, \theta_{k+1})}{\partial h} q_k(z)$,

$$\begin{aligned} \mathbb{E}(\|q_L(z) - z\|^2) &= \alpha_L^2 \sum_{k,k'=0}^{L-1} \mathbb{E}\left(q_k(z)^\top \frac{\partial g(h_k, \theta_{k+1})}{\partial h} V_{k+1}^\top V_{k'+1} \frac{\partial g(h_{k'}, \theta_{k'+1})}{\partial h} q_{k'}(z)\right) \\ &= \alpha_L^2 \sum_{k=0}^{L-1} \mathbb{E}\left(\left\|V_{k+1} \frac{\partial g(h_k, \theta_{k+1})}{\partial h} q_k(z)\right\|^2\right) \\ &\leq d\alpha_L^2 \sum_{k=0}^{L-1} (1 + \alpha_L^2)^k \\ &= d((1 + \alpha_L^2)^L - 1) \leq d(\exp(L\alpha_L^2) - 1) \leq 2dL\alpha_L^2, \end{aligned}$$

for $L\alpha_L^2 \leq 1$. Note that the second equality is obtained by conditioning on every random variable except $V_{k'+1}$ for $k < k'$ (and V_{k+1} for $k > k'$). Finally, by using Markov's inequality, we conclude that, for any $\varepsilon > 0$,

$$\mathbb{P}(\|p_0 - p_L\|^2 \geq \varepsilon \|p_L\|^2) \leq \frac{2L\alpha_L^2}{\varepsilon}.$$

A.7 Proof of Proposition 7

The proof of Proposition 6 reveals that

$$\mathbb{E}\left(\frac{\|p_0 - p_L\|^2}{\|p_L\|^2}\right) \leq (1 + \alpha_L^2)^L - 1.$$

Using similar arguments, one has

$$\mathbb{E}\left(\frac{\|p_0 - p_L\|^2}{\|p_L\|^2}\right) = \frac{1}{d}\mathbb{E}\left(\frac{\|q_L(z) - z\|^2}{\|z\|^2}\right) \geq \alpha_L^2 \sum_{k=0}^{L-1} \left(1 + \frac{1}{2}\alpha_L^2\right)^k = \left(1 + \frac{1}{2}\alpha_L^2\right)^L - 1.$$

A.8 Proof of Corollary 8

The first statement is an immediate consequence of Proposition 6. The second one is a consequence of Proposition 7 and the fact that, for $\beta < 1$,

$$\left(1 + \frac{1}{L^\beta}\right)^L = \exp\left(L \ln\left(1 + \frac{1}{L^\beta}\right)\right) \sim \exp(L^{1-\beta}) \rightarrow \infty.$$

Finally, (iii) follows from Proposition 7.

A.9 Proof of Proposition 10

The proposition is a consequence of Kloeden and Platen (1992, Theorems 4.5.3 and 10.2.2) for the SDE

$$dH_t^\top = \sqrt{\frac{d}{2}} \sigma(H_t^\top) dB_t.$$

Letting $a(h, t) = 0$ and $b(h, t) = \sqrt{\frac{d}{2}} \sigma(h)$, we need to check the following assumptions:

(H₁) The functions $a(\cdot, \cdot)$ and $b(\cdot, \cdot)$ are jointly measurable on $\mathbb{R}^d \times [0, 1]$.

(H₂) There exists a constant $C_1 > 0$ such that, for any $x, y \in \mathbb{R}^d$, $t \in [0, 1]$,

$$\|a(x, t) - a(y, t)\| + \|b(x, t) - b(y, t)\| \leq C_1 \|x - y\|.$$

(H₃) There exists a constant $C_2 > 0$ such that, for any $x \in \mathbb{R}^d$, $t \in [0, 1]$,

$$\|a(x, t)\| + \|b(x, t)\| \leq C_2(1 + \|x\|).$$

(H₄) $\mathbb{E}(\|H_0\|^2) < \infty$.

(H₅) There exists a constant $C_3 > 0$ such that, for any $x \in \mathbb{R}^d$, $s, t \in [0, 1]$,

$$\|a(x, t) - a(x, s)\| + \|b(x, t) - b(x, s)\| \leq C_3(1 + \|x\|)|t - s|^{1/2}.$$

Assumptions (H₁), (H₄), and (H₅) readily follow from the definitions. Assumption (H₂) is true since σ is Lipschitz continuous, and (H₃) follows from

$$\|\sigma(x)\| \leq b\|x\| \leq \|x\| \leq 1 + \|x\|.$$

A.10 Proof of Proposition 11

Let $\psi : \mathbb{R}^d \times [0, 1] \rightarrow \mathbb{R}^d$ be defined for any $h \in \mathbb{R}^d$, $t \in [0, 1]$, by $\psi(h, t) = \mathcal{V}_t g(h, \Theta_t)$. With this notation, the ODE (14) is equivalent to the initial value problem

$$dH_t = \psi(H_t, t)dt, \quad H_0 = Ax.$$

By Assumptions (A₅) and (A₆), ψ is Lipschitz continuous in its first argument, in the sense that there exists $K > 0$ such that, for all $h, h' \in \mathbb{R}^d$, $t \in [0, 1]$,

$$\|\psi(h, t) - \psi(h', t)\| \leq K\|h - h'\|.$$

In addition, it is continuous in its second one. Thus, according to the Picard-Lindelöf theorem (Theorem 22 in Appendix D), this is enough to show that the neural ODE (14) has a unique solution on $[0, 1]$. Note that the solution H is continuous on $[0, 1]$ and is therefore bounded by a constant $M > 0$.

In order to prove the approximation bound of Proposition 11, we start by proving that both ψ and H are Lipschitz continuous in t . Under (A_5) and (A_6) , this is clear for ψ since H is bounded. Moreover, for any $[s, t] \subset [0, 1]$, we have

$$\begin{aligned} \|H_t - H_s\| &= \left\| \int_s^t \psi(H_u, u) du \right\| \leq \int_s^t \|\psi(H_u, u)\| du \\ &\leq (t - s) \sup_{\substack{u \in [0, 1] \\ h \in \mathbb{R}^d, \|h\| \leq M}} \|\psi(h, u)\|. \end{aligned}$$

Now, let K_1 and K_2 denote the Lipschitz constants of ψ (in both arguments) and H respectively, and, for any $0 \leq k \leq L$, let $t_k = k/L$. Then we have, for $k \geq 1$,

$$\begin{aligned} &\|H_{t_k} - h_k\| \\ &= \left\| H_{t_{k-1}} + \int_{t_{k-1}}^{t_k} \psi(H_u, u) du - h_{k-1} - \frac{1}{L} \psi(h_{k-1}, t_{k-1}) \right\| \\ &\leq \|H_{t_{k-1}} - h_{k-1}\| + \int_{t_{k-1}}^{t_k} \|\psi(H_u, u) - \psi(h_{k-1}, t_{k-1})\| du \\ &\leq \|H_{t_{k-1}} - h_{k-1}\| + K_1 \int_{t_{k-1}}^{t_k} \|H_u - h_{k-1}\| du + K_1 \int_{t_{k-1}}^{t_k} |u - t_{k-1}| du \\ &\leq \left(1 + \frac{K_1}{L}\right) \|H_{t_{k-1}} - h_{k-1}\| + K_1 \int_{t_{k-1}}^{t_k} \|H_u - H_{t_{k-1}}\| du + K_1 \int_{t_{k-1}}^{t_k} |u - t_{k-1}| du \\ &\leq \left(1 + \frac{K_1}{L}\right) \|H_{t_{k-1}} - h_{k-1}\| + (K_2 + 1) K_1 \int_{t_{k-1}}^{t_k} |u - t_{k-1}| du \\ &= \left(1 + \frac{K_1}{L}\right) \|H_{t_{k-1}} - h_{k-1}\| + \frac{(K_2 + 1) K_1}{2L^2}. \end{aligned}$$

By recurrence, we obtain

$$\begin{aligned} \|H_{t_k} - h_k\| &\leq \sum_{j=0}^{k-1} \left(1 + \frac{K_1}{L}\right)^j \times \frac{(K_2 + 1) K_1}{2L^2} \leq L \left(1 + \frac{K_1}{L}\right)^L \times \frac{(K_2 + 1) K_1}{2L^2} \\ &\leq e^{K_1} \frac{(K_2 + 1) K_1}{2L}, \end{aligned}$$

which concludes the proof.

A.11 Proof of Proposition 12

Starting from (4) and using Assumption (A_6) , one easily obtains the existence of C_1 and C_2 (whose values depend on the realization of \mathcal{V} and Θ) such that

$$\|h_{k+1}\| \leq (1 + C_1 \alpha_L) \|h_k\| + C_2 \alpha_L.$$

By recurrence,

$$\|h_{k+1}\| \leq (1 + C_1 \alpha_L)^k \left(\|h_0\| + \frac{C_2}{C_1} \right).$$

Hence, using $\alpha_L \leq 1/L$,

$$\|h_{k+1}\| \leq \exp(C_1) \left(\|h_0\| + \frac{C_2}{C_1} \right).$$

Since g is Lipschitz continuous on compact sets, it is bounded on every ball of $\mathbb{R}^d \times \mathbb{R}^p$. The result is then a consequence of the identity

$$h_L - h_0 = \alpha_L \sum_{k=0}^{L-1} V_{k+1} g(h_k, \theta_{k+1}),$$

since we showed that each term in the sum is bounded by some constant $C_3 > 0$, independent of L and k . Hence we have that

$$\|h_L - h_0\| \leq C_3 L \alpha_L = C_3 L^{1-\beta},$$

yielding the results depending on the value of β .

A.12 Proof of Proposition 13

In the linear case, (4) can be written

$$h_{k+1} = h_k + \alpha_L V_{k+1} h_k, \quad 0 \leq k \leq L-1.$$

Take y a unit-norm eigenvector of \mathcal{V}_0^\top with associated eigenvalue $\lambda > 0$. Then

$$\begin{aligned} \langle h_{k+1}, y \rangle &= \langle h_k + \alpha_L V_{k+1} h_k, y \rangle \\ &= \langle h_k, y \rangle + \alpha_L \langle h_k, V_{k+1}^\top y \rangle \\ &= \langle h_k, y \rangle + \lambda \alpha_L \langle h_k, y \rangle + \alpha_L \langle h_k, (V_{k+1} - \mathcal{V}_0)^\top y \rangle. \end{aligned}$$

Since \mathcal{V} is Lipschitz and $V_{k+1} = \mathcal{V}_{k+1/L}$, there exists c such that $\|V_{k+1} - \mathcal{V}_0\| \leq c \frac{k+1}{L}$. Hence

$$|\langle h_{k+1}, y \rangle| \geq (1 + \lambda \alpha_L) |\langle h_k, y \rangle| - c \alpha_L \frac{k+1}{L} \|h_k\|.$$

Then, by recurrence,

$$\begin{aligned} |\langle h_L, y \rangle| &\geq (1 + \lambda \alpha_L)^L |\langle h_0, y \rangle| - c \frac{\alpha_L}{L} \sum_{k=0}^{L-1} (k+1) (1 + \lambda \alpha_L)^k \|h_k\| \\ &\geq (1 + \lambda \alpha_L)^L |\langle h_0, y \rangle| - c \alpha_L (1 + \lambda \alpha_L)^L \max_k \|h_k\|. \end{aligned}$$

Let $M = \frac{|\langle h_0, y \rangle|}{2c\alpha_L}$, and suppose that $\|h_k\| \leq M$ for all $0 \leq k \leq L$. Then

$$\begin{aligned} \|h_L\| &\geq |\langle h_L, y \rangle| \\ &\quad (\text{by the Cauchy-Schwartz inequality}) \\ &\geq (1 + \lambda \alpha_L)^L (|\langle h_0, y \rangle| - cM\alpha_L). \end{aligned}$$

Then, for $\lambda\alpha_L \leq 1$,

$$\|h_L\| \geq \frac{1}{2}(1 + \lambda\alpha_L)^L |\langle h_0, y \rangle| \geq \frac{1}{2} \exp\left(\frac{\lambda L \alpha_L}{2}\right) |\langle h_0, y \rangle|.$$

Thus, since $L\alpha_L = L^{1-\beta}$, we have that $\|h_L\| \rightarrow \infty$, which contradicts our assumption that $\|h_k\| \leq M$ for all $0 \leq k \leq L$. We deduce that, for all L large enough,

$$\max_k \|h_k\| > \frac{|\langle h_0, y \rangle|}{2c\alpha_L} \xrightarrow{L \rightarrow \infty} \infty.$$

Furthermore,

$$\max_k \frac{\|h_k - h_0\|}{\|h_0\|} > \frac{|\langle h_0, y \rangle|}{2c\|h_0\|\alpha_L} - 1 \xrightarrow{L \rightarrow \infty} \infty.$$

B. Technical results

B.1 Lemmas 16 and 17

Lemma 16 *Let $W \in \mathbb{R}^{d \times d}$ be a matrix whose entries are symmetric i.i.d. random variables, with finite variance, and let σ be an activation function such that, for all $x \in \mathbb{R}$, $a|x| \leq |\sigma(x)| \leq b|x|$, $1/\sqrt{2} \leq a < b \leq 1$. Then, for any $x \in \mathbb{R}^d$,*

$$\frac{1}{2}\mathbb{E}(\|Wx\|^2) \leq \mathbb{E}(\|\sigma(Wx)\|^2) \leq \mathbb{E}(\|Wx\|^2) \quad \text{and} \quad \mathbb{E}(\|\text{ReLU}(Wx)\|^2) = \frac{1}{2}\mathbb{E}(\|Wx\|^2).$$

Proof The first part is a consequence of the assumption on σ . To prove the equality, let $X_i = \sum_{j=1}^d W_{ij}x_j$. Then

$$\mathbb{E}(\|\text{ReLU}(Wx)\|^2) = \mathbb{E}\left(\sum_{i=1}^d \left(\sum_{j=1}^d W_{ij}x_j\right)^2 \mathbf{1}_{\sum_{j=1}^d W_{ij}x_j \geq 0}\right) = \mathbb{E}\left(\sum_{i=1}^d X_i^2 \mathbf{1}_{X_i \geq 0}\right).$$

Since the $(W_{ij})_{1 \leq j \leq d}$ are symmetric and independent random variables, X_i is also symmetric. Hence $\mathbb{E}(X_i^2 \mathbf{1}_{X_i \geq 0}) = 1/2\mathbb{E}(X_i^2)$, which concludes the proof. \blacksquare

Lemma 17 *Let $W \in \mathbb{R}^{d \times d}$ be a matrix whose entries are centered i.i.d. random variables, with finite variance s^2 . Then, for any $x \in \mathbb{R}^d$, $\mathbb{E}(\|Wx\|^2) = s^2 d \|x\|^2$.*

Proof For any $1 \leq i \leq d$,

$$|Wx|_i^2 = \left(\sum_{j=1}^d W_{ij}x_j\right)^2 = \sum_{j,j'=1}^d W_{ij}W_{ij'}x_jx_{j'}.$$

Thus, by independence,

$$\mathbb{E}(|Wx|_i^2) = \mathbb{E}\left(\sum_{j,j'=1}^d W_{ij}W_{ij'}x_jx_{j'}\right) = \sum_{j=1}^d \mathbb{E}(W_{ij}^2)x_j^2 = s^2 \|x\|^2. \quad (21)$$

The result follows by summing over all $i \in \{1, \dots, d\}$. \blacksquare

B.2 Proof of Lemma 15

(E_1) and (E_2) are simple consequences of Assumptions (A_1) and (A_2).

To prove (E_3), let $f(h_k, \theta_{k+1}) = V_{k+1}g(h_k, \theta_{k+1})$. Then

$$\begin{aligned}\mathbb{E}(Y_{k,2}Y_{k,1}|h_k) &= \frac{1}{\|h_k\|^4}\mathbb{E}(\|f(h_k, \theta_{k+1})\|^2\langle h_k, f(h_k, \theta_{k+1}) \rangle|h_k) \\ &= \mathbb{E}\left(\sum_{i=1}^d\sum_{j=1}^df(h_k, \theta_{k+1})_i^2(h_k)_j f(h_k, \theta_{k+1})_j|h_k\right).\end{aligned}$$

It is easy to verify that, under Assumption (A_1), each term of the sum above has zero expectation. This shows (E_3).

To establish (E_4), we start by noting that

$$\begin{aligned}\mathbb{E}\left(\left\langle\frac{h_k}{\|h_k\|}, \frac{f(h_k, \theta_{k+1})}{\|h_k\|}\right\rangle^2|h_k\right) &= \frac{1}{\|h_k\|^4}\mathbb{E}(h_k^\top f(h_k, \theta_{k+1})f(h_k, \theta_{k+1})^\top h_k|h_k) \\ &= \frac{1}{\|h_k\|^4}h_k^\top\mathbb{E}(f(h_k, \theta_{k+1})f(h_k, \theta_{k+1})^\top|h_k)h_k.\end{aligned}$$

Clearly, $\mathbb{E}(f(h_k, \theta_{k+1})_i f(h_k, \theta_{k+1})_j) = 0$ for $i \neq j$. Since, furthermore, the coordinates of $f(h_k, \theta_{k+1})$ are identically distributed conditionally on h_k , we obtain

$$\mathbb{E}(f(h_k, \theta_{k+1})f(h_k, \theta_{k+1})^\top|h_k) = \frac{1}{d}\mathbb{E}(\|f(h_k, \theta_{k+1})\|^2|h_k)I_d.$$

Thus,

$$\mathbb{E}\left(\left\langle\frac{h_k}{\|h_k\|}, \frac{f(h_k, \theta_{k+1})}{\|h_k\|}\right\rangle^2|h_k\right) = \frac{1}{d\|h_k\|^4}\mathbb{E}(\|f(h_k, \theta_{k+1})\|^2|h_k)h_k^\top h_k \leq \frac{1}{d},$$

by Assumptions (A_1) and (A_2).

To prove (E_5), let $\varphi = \frac{\langle V_{k+1}g(h_k, \theta_{k+1}), h_k \rangle}{\|g(h_k, \theta_{k+1})\|\|h_k\|}$. Then, for any $t > 0$,

$$\begin{aligned}\mathbb{P}(|Y_{k,2}| > t) &= \mathbb{P}\left(|\varphi| > \frac{t\|h_k\|}{2\alpha_L\|g(h_k, \theta_{k+1})\|}\right) \\ &= \mathbb{E}\left(\mathbb{P}\left(|\varphi| > \frac{t\|h_k\|}{2\alpha_L\|g(h_k, \theta_{k+1})\|}\middle|h_k, \theta_{k+1}\right)\right).\end{aligned}$$

So, by Lemma 20 in Appendix C,

$$\begin{aligned}\mathbb{P}(|Y_{k,2}| > t) &\leq \mathbb{E}\left(2\exp\left(-\frac{dt^2\|h_k\|^2}{16\alpha_L^2s^2\|g(h_k, \theta_{k+1})\|^2}\right)\right) \\ &= \mathbb{E}\left(\mathbb{E}\left(2\exp\left(-\frac{dt^2\|h_k\|^2}{16\alpha_L^2s^2\|g(h_k, \theta_{k+1})\|^2}\right)\middle|h_k\right)\right) \\ &\leq \mathbb{E}\left(2\exp\left(-\frac{dt^2\|h_k\|^2}{16\alpha_L^2s^2\mathbb{E}(\|g(h_k, \theta_{k+1})\|^2|h_k)}\right)\right),\end{aligned}$$

by Jensen's inequality. Finally, using Assumption (A_2) , we deduce that

$$\mathbb{P}(|Y_{k,2}| > t) \leq \mathbb{E}\left(2 \exp\left(-\frac{dt^2}{16\alpha_L^2 s^2}\right)\right) = 2 \exp\left(-\frac{dt^2}{16\alpha_L^2 s^2}\right). \quad (22)$$

In particular, for all $q \geq 1$ (see, e.g., [Pauwels, 2020](#)),

$$\mathbb{E}(Y_{k,2}^{2q}) \leq q! \left(\frac{32s^2\alpha_L^2}{d}\right)^q.$$

The result is obtained by taking $q = 2$.

Finally, (E_6) and (E_7) are consequences of Lemma 21 in Appendix C.

C. Concentration of sub-Gaussian random matrices

In this appendix, we are interested in the concentration of linear and quadratic forms of sub-Gaussian matrices (Lemma 20 and Lemma 21). These two propositions are byproducts of the main result of [Kontorovich \(2014\)](#), which generalizes McDiarmid's inequality to sub-Gaussian variables. We start by a technical result regarding the sub-Gaussian diameter introduced by [Kontorovich \(2014\)](#), whose definition is recalled below.

Definition 18 *Let X be a real-valued random variable, X' an independent copy of X , and ε a Rademacher random variable, independent of X and X' . Then the sub-Gaussian diameter of X is defined as the smallest t such that $\varepsilon|X - X'|$ is t^2 sub-Gaussian.*

Lemma 19 *Let X be a s^2 sub-Gaussian symmetric random variable. Then the sub-Gaussian diameter of X is less than $\sqrt{2}s$.*

Proof Let $\lambda \in \mathbb{R}$. Then, using the notation of Definition 18, one has

$$\begin{aligned} \mathbb{E}(\exp^{\lambda\varepsilon|X-X'|}) &= \mathbb{E}(\exp^{\lambda(X-X')} \mathbf{1}_{\varepsilon=1}) + \mathbb{E}(\exp^{\lambda\varepsilon(X'-X)} \mathbf{1}_{\varepsilon=-1}) \\ &= \mathbb{E}(\exp^{\lambda(X-X')}) \\ &= \mathbb{E}(\exp^{\lambda X})^2 \\ &\leq \exp^{2\lambda^2 s^2}, \end{aligned}$$

where the last equality is a consequence of the symmetry of X . ■

We are now ready to prove the two main results of this appendix.

Lemma 20 (Bound on the deviation of linear forms) *Let V be a $\mathbb{R}^{d \times d}$ matrix whose entries are i.i.d s^2/d sub-Gaussian random variables. Then, for any $x, y \in \mathbb{R}^d$, $x, y \neq 0$,*

$$\mathbb{P}\left(\frac{\langle Vx, y \rangle}{\|x\| \|y\|} \geq t\right) \leq 2 \exp\left(-\frac{dt^2}{4s^2}\right).$$

Proof For any $1 \leq i, j \leq d$, set $X_{ij} = \frac{x_i V_{ij} y_j}{\|x\| \|y\|}$. Let $\mathcal{X} = \mathbb{R}^{d^2}$ endowed with the ℓ_1 norm, let X be the vector in \mathcal{X} whose $(id + j)$ -th coordinate is X_{ij} , and let the function φ be defined by

$$\varphi : \mathcal{X} \ni Y \mapsto \sum_{i=1}^{d^2} Y_i.$$

By the triangle inequality, φ is a Lipschitz continuous function, with Lipschitz constant equal to 1. Observe also that X_{ij} is a $x_i^2 s^2 y_j^2 / d \|x\|^2 \|y\|^2$ sub-Gaussian. Thus, according to Lemma 19, the sub-Gaussian diameter of X_{ij} is less than $\sqrt{2} x_i s y_j / \sqrt{d} \|x\| \|y\|$. By Kontorovich (2014, Theorem 1), for any $t > 0$, one has

$$\mathbb{P}(\varphi(X) \geq t) \leq 2 \exp\left(-\frac{t^2}{2 \sum_{i,j=1}^d \frac{2s^2 x_i^2 y_j^2}{d \|x\|^2 \|y\|^2}}\right),$$

that is

$$\mathbb{P}\left(\frac{\langle Vx, y \rangle}{\|x\| \|y\|} \geq t\right) \leq 2 \exp\left(-\frac{dt^2}{4s^2}\right).$$

■

Lemma 21 (Bound of moments of quadratic forms) *Let V be a $\mathbb{R}^{d \times d}$ matrix whose entries are i.i.d s^2/d sub-Gaussian random variables, with variance $1/d$. Then, for any $x \in \mathbb{R}^d$, $x \neq 0$,*

$$\mathbb{E}\left(\frac{\|Vx\|^4}{\|x\|^4}\right) \leq 1 + \frac{128s^4}{d} \quad \text{and} \quad \mathbb{E}\left(\frac{\|Vx\|^8}{\|x\|^8}\right) \leq 1 + \frac{384s^4}{d} + \frac{3072s^6}{d^2}.$$

Proof The proof is similar to the one of Lemma 20, with $X_{ij} = \frac{V_{ij} x_j}{\|x\|}$, $\mathcal{X} = \mathbb{R}^d$, and

$$\varphi_i : \mathcal{X} \ni X \mapsto \sum_{j=1}^d X_{ij}.$$

Each function φ_i is a Lipschitz continuous function, with Lipschitz constant equal to 1. Observe now that the random variable X_{ij} is $x_j^2 s^2 / d \|x\|^2$ sub-Gaussian. Thus, according to Lemma 19, the sub-Gaussian diameter of X_{ij} is less than $\sqrt{2} x_j s / \sqrt{d} \|x\|$. Therefore, according to Kontorovich (2014, Theorem 1), for any $t > 0$,

$$\mathbb{P}(\varphi_i(X) \geq t) \leq 2 \exp\left(-\frac{t^2}{2 \sum_{j=1}^d \frac{2s^2 x_j^2}{d \|x\|^2}}\right),$$

that is

$$\mathbb{P}\left(\frac{|\langle V_i, x \rangle|}{\|x\|} \geq t\right) \leq 2 \exp\left(-\frac{dt^2}{4s^2}\right).$$

Hence (see, e.g., [Pauwels, 2020](#)),

$$\mathbb{E}\left(\left(\frac{\langle V_i, x \rangle}{\|x\|}\right)^{2q}\right) \leq q! \left(\frac{8s^2}{d}\right)^q. \quad (23)$$

From identity (21) in the proof of technical Lemma 17, given in Appendix B, we obtain that, for $q = 1$,

$$\mathbb{E}\left(\left(\frac{\langle V_i, x \rangle}{\|x\|}\right)^2\right) = \frac{1}{d}, \quad (24)$$

which is an improvement by a factor $8s^2$ over the previous upper bound. To conclude, it remains to conclude $\|Vx\|^4$ and $\|Vx\|^8$ with the $\langle V_i, x \rangle$. To do so, observe that

$$\|Vx\|^4 = \left(\sum_{i=1}^d \langle V_i, x \rangle^2\right)^2 = \sum_{\substack{i,j=1 \\ i \neq j}}^d \langle V_i, x \rangle^2 \langle V_j, x \rangle^2 + \sum_{i=1}^d \langle V_i, x \rangle^4.$$

Hence, by independence of the $(V_i)_{1 \leq i \leq d}$,

$$\begin{aligned} \mathbb{E}\left(\frac{\|Vx\|^4}{\|x\|^4}\right) &= \sum_{\substack{i,j=1 \\ i \neq j}}^d \mathbb{E}\left(\frac{\langle V_i, x \rangle^2}{\|x\|^2}\right) \mathbb{E}\left(\frac{\langle V_j, x \rangle^2}{\|x\|^2}\right) + \sum_{i=1}^d \mathbb{E}\left(\frac{\langle V_i, x \rangle^4}{\|x\|^4}\right) \\ &= d(d-1) \frac{1}{d^2} + d \frac{2(8s^2)^2}{d^2} \leq 1 + \frac{128s^4}{d} \\ &\quad \text{(by (23) and (24))} \end{aligned}$$

Similarly,

$$\|Vx\|^8 = \left(\sum_{i=1}^d \langle V_i, x \rangle^2\right)^3 = \sum_{\substack{i,j,k=1 \\ i \neq j \neq k}}^d \langle V_i, x \rangle^2 \langle V_j, x \rangle^2 \langle V_k, x \rangle^2 + \sum_{\substack{i,j=1 \\ i \neq j}}^d \langle V_i, x \rangle^2 \langle V_j, x \rangle^4 + \sum_{i=1}^d \langle V_i, x \rangle^8.$$

Hence,

$$\begin{aligned} \mathbb{E}\left(\frac{\|Vx\|^8}{\|x\|^8}\right) &= \sum_{\substack{i,j,k=1 \\ i \neq j \neq k}}^d \mathbb{E}\left(\frac{\langle V_i, x \rangle^2}{\|x\|^2}\right) \mathbb{E}\left(\frac{\langle V_j, x \rangle^2}{\|x\|^2}\right) \mathbb{E}\left(\frac{\langle V_k, x \rangle^2}{\|x\|^2}\right) \\ &\quad + \sum_{\substack{i,j=1 \\ i \neq j}}^d \mathbb{E}\left(\frac{\langle V_i, x \rangle^4}{\|x\|^4}\right) \mathbb{E}\left(\frac{\langle V_j, x \rangle^2}{\|x\|^2}\right) + \sum_{i=1}^d \mathbb{E}\left(\frac{\langle V_i, x \rangle^8}{\|x\|^8}\right) \\ &= d(d-1)(d-2) \frac{1}{d^3} + 3d(d-1) \frac{2(8s^2)^2}{d^3} + d \frac{6(8s^2)^3}{d^3} \\ &\leq 1 + \frac{384s^4}{d} + \frac{3072s^6}{d^2}. \end{aligned}$$

■

D. A version of the Picard-Lindelöf theorem

Theorem 22 *Assume that f is Lipschitz continuous in its first argument and continuous in its second one. Then, for any $z \in \mathbb{R}^d$, the initial value problem*

$$dH_t = f(H_t, t)dt, \quad H_0 = z, \quad (25)$$

admits a unique solution $H : [0, 1] \rightarrow \mathbb{R}^d$.

Proof Let $\mathcal{C}([s, t], \mathbb{R}^d)$ be the set of continuous functions from $[s, t]$ to \mathbb{R}^d . For any $[s, t] \subset [0, 1]$, $\zeta \in \mathbb{R}^d$, let Ψ be the function

$$\begin{aligned} \Psi : \mathcal{C}([s, t], \mathbb{R}^d) &\rightarrow \mathcal{C}([s, t], \mathbb{R}^d) \\ Y &\mapsto (v \mapsto \zeta + \int_s^v f(Y_u, u)du). \end{aligned}$$

For any $Y, Y' \in \mathcal{C}([s, t], \mathbb{R}^d)$, $v \in [s, t]$, one has, denoting by K_f the Lipschitz constant of f in its first argument,

$$\begin{aligned} \|\Psi(Y)_v - \Psi(Y')_v\| &\leq \int_s^v \|(f(Y_u, u) - f(Y'_u, u))\| du \\ &\leq \int_s^v K_f \|Y_u - Y'_u\| du \\ &\leq K_f \int_s^v \|Y - Y'\|_\infty du \\ &\leq K_f \|Y - Y'\|_\infty (t - s). \end{aligned}$$

This yields

$$\|\Psi(Y) - \Psi(Y')\|_\infty \leq K_f (t - s) \|Y - Y'\|_\infty,$$

which means that the function Ψ is Lipschitz continuous on $\mathcal{C}([s, t], \mathbb{R}^d)$ endowed with the supremum norm, with Lipschitz constant $K_f(t - s)$. So, on any interval $[s, t]$ of length smaller than $\delta = 1/2K_f$, the function Ψ is a contraction. Thus, by the Banach fixed-point theorem, for any initial value ζ , Ψ has a unique fixed point. Hence, there exists a unique solution to (25) on any interval of length δ with any initial condition. To obtain a solution on $[0, 1]$ it is sufficient to concatenate these solutions. \blacksquare

E. Detailed experimental setting

Our code is available at <https://github.com/PierreMarion23/scaling-resnets>.

To obtain Figures 1 to 3, we initialize ResNets from `res-3` with the hyper-parameters of Table 2.

Each experiment is repeated 50 times, with independent data and weight sampling.

For Figures 4 and 5, we take the same hyper-parameters except for β , which now takes values in $\{0.5, 1, 2\}$, and for the weight distribution. The weights are now initialized as

Name	Value
d	40
n_{in}	64
n_{out}	1
L	10 to 1000
β	0.25, 0.5, 1
weight distribution	$\mathcal{U}(-\sqrt{3/d}, \sqrt{3/d})$
data distribution	standard Gaussian

Table 2: Hyperparameters of Figures 1 to 3

Name	Value
d	40
n_{in}	64
n_{out}	1
L	1000
β	0.2 to 1.3
weight distribution	fractional Brownian motion with Hurst index from 0.05 to 0.97
data distribution	standard Gaussian

Table 3: Hyperparameters of Figure 7

discretizations of a Gaussian process. More precisely, each entry of \mathcal{V} and Θ is an independent Gaussian process with zero mean and an RBF kernel of variance 10^{-2} .

To obtain Figure 7, we take the hyper-parameters of Table 3.

More precisely, for each $1 \leq i, j \leq d$, we let $(V_{k+1,i,j})_{0 \leq k \leq L-1}$ be the increments of a fractional Brownian motion (fBm), where the various fBm involved are independent. The procedure is the same for θ .

In Figure 9, we use `res-1`, with the hyper-parameters of Table 4.

Name	Value
d	30
L	1000
β	0.2 to 1.3
weight distribution	fractional Brownian motion with Hurst index from 0.05 to 0.97
learning rate grid	$10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1$

Table 4: Hyperparameters of Figure 9

We train on MNIST¹ and CIFAR-10² using the Adam optimizer (Kingma and Ba, 2015) for 10 epochs. The learning rate is divided by 10 after 5 epochs. The best performance on the learning rate grid is reported in the figure.

Figure 8 is obtained by plotting a random coordinate of θ_k , after training on MNIST.

References

- D. Arpit, V. Campos, and Y. Bengio. How to initialize your network? Robust initialization for WeightNorm & ResNets. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, pages 10902–10911. Curran Associates, Inc., 2019.
- T. Bachlechner, B.P. Majumder, H. Mao, G. Cottrell, and J. Auley. ReZero is all you need: Fast convergence at large depth. In C. de Campos and M.H. Maathuis, editors, *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pages 1352–1361. PMLR, 2021.
- C. Bayer, P.K. Friz, and N. Tapia. Stability of deep neural networks via discrete rough paths. *arXiv:2201.07566*, 2022.
- A. Brock, S. De, and S.L. Smith. Characterizing signal propagation to close the performance gap in unnormalized ResNets. In *International Conference on Learning Representations*, 2021.
- B. Chang, M. Chen, E. Haber, and E.H. Chi. AntisymmetricRNN: A dynamical system view on recurrent neural networks. In *International Conference on Learning Representations*, 2019.
- R.T.Q. Chen, Y. Rubanova, J. Bettencourt, and D.K. Duvenaud. Neural ordinary differential equations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31, pages 6572–6583. Curran Associates, Inc., 2018.
- F. Chollet et al. Keras. <https://github.com/fchollet/keras>, 2015.
- A.-S. Cohen, R. Cont, A. Rossier, and R. Xu. Scaling properties of deep residual networks. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 2039–2048. PMLR, 2021.
- S. De and S. Smith. Batch normalization biases residual blocks towards the identity function in deep networks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 19964–19975. Curran Associates, Inc., 2020.
- L. Deng. The MNIST database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29:141–142, 2012.

1. <http://yann.lecun.com/exdb/mnist>

2. <https://www.cs.toronto.edu/~kriz/cifar.html>

- W. E, J. Han, and Q. Li. A mean-field optimal control formulation of deep learning. *Research in the Mathematical Sciences*, 6:10, 2019.
- A. Fermanian, P. Marion, J.-P. Vert, and G. Biau. Framing RNN as a kernel method: A neural ODE approach. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 3121–3134. Curran Associates, Inc., 2021.
- X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In Y.W. Teh and M. Titterton, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9, pages 249–256. PMLR, 2010.
- I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.
- W. Grathwohl, R.T.Q. Chen, J. Bettencourt, I. Sutskever, and D. Duvenaud. FFJORD: Free-form continuous dynamics for scalable reversible generative models. In *International Conference on Learning Representations*, 2019.
- F. Götze and J. Jalowy. Rate of convergence to the circular law via smoothing inequalities for log-potentials. *Random Matrices: Theory and Applications*, 10:2150026, 2021.
- B. Hanin and D. Rolnick. How to start training: The effect of initialization and architecture. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31, pages 569–579. Curran Associates, Inc., 2018.
- K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034. IEEE Computer Society, 2015.
- K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016a.
- K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, *Computer Vision – ECCV 2016*, pages 630–645. Springer International Publishing, 2016b.
- S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In F. Bach and D. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pages 448–456. PMLR, 2015.
- S. Jastrzebski, Z. Kenton, D. Arpit, N. Ballas, A. Fischer, Y. Bengio, and A. Storkey. Three factors influencing minima in SGD. *arXiv:1711.04623*, 2017.
- P. Kidger, J. Foster, X. Li, and T. Lyons. Efficient and accurate gradients for neural SDEs. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 18747–18761. Curran Associates, Inc., 2021.

- D.P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.
- P.E. Kloeden and E. Platen. *Numerical Solution of Stochastic Differential Equations*. Springer, Berlin, 1992.
- A. Kontorovich. Concentration in unbounded metric spaces and algorithmic stability. In E.P. Xing and T. Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32, pages 28–36. PMLR, 2014.
- A. Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.
- X. Li, T.-K. L. Wong, R.T.Q. Chen, and D.K. Duvenaud. Scalable gradients and variational inference for stochastic differential equations. In C. Zhang, F. Ruiz, T. Bui, A.B. Dieng, and D. Liang, editors, *Proceedings of the 2nd Symposium on Advances in Approximate Bayesian Inference*, volume 118, pages 1–28. PMLR, 2020.
- E. Pauwels. *Statistics, Optimization and Algorithms in High Dimension*. Lecture Notes, Toulouse 3 Paul Sabatier University, 2020.
- S. Peluchetti and S. Favaro. Infinitely deep neural networks as diffusion processes. In S. Chiappa and R. Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108, pages 1126–1136. PMLR, 2020.
- J. Shao, K. Hu, C. Wang, X. Xue, and B. Raj. Is normalization indispensable for training deep neural network? In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 13434–13444. Curran Associates, Inc., 2020.
- M. Thorpe and Y. van Gennip. Deep limits of residual neural networks. *arXiv:1810.11741*, 2018.
- R. van Handel. *Probability in High Dimension*. APC 550 Lecture Notes, Princeton University, 2016.
- A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc., 2017.
- H. Wang, S. Ma, L. Dong, S. Huang, D. Zhang, and F. Wei. Deepnet: Scaling transformers to 1,000 layers. *arXiv:2203.00555*, 2022.
- G. Yang and S. Schoenholz. Mean field residual networks: On the edge of chaos. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 2865–2873. Curran Associates, Inc., 2017.

- H. Zhang, Y.N. Dauphin, and T. Ma. Fixup initialization: Residual learning without normalization. In *International Conference on Learning Representations*, 2019a.
- H. Zhang, D. Yu, M. Yi, W. Chen, and T.-Y. Liu. Convergence theory of learning over-parameterized ResNet: A full characterization. *arXiv:1903.07120*, 2019b.
- J. Zhang, B. Han, L. Wynter, B.K.H. Low, and M. Kankanhalli. Towards robust ResNet: A small step but a giant leap. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 4285–4291. International Joint Conferences on Artificial Intelligence Organization, 2019c.