Journal of Machine Learning Research 22 (2021) 1-48 Submitted 7/20; Published 8/21

Doubly infinite residual neural networks: a diffusion process approach

Stefano Peluchetti SPELUCHETTI@COGENT.CO.JP
Cogent Labs
106-0032, Tokyo, Japan.

Stefano Favaro STEFANO.FAVARO@UNITO.IT
Department of Economics and Statistics
University of Torino and Collegio Carlo Alberto
10122, Torino, Italy.

Editor: Philipp Hennig

Abstract

Modern neural networks featuring a large number of layers (depth) and units per layer (width) have achieved a remarkable performance across many domains. While there exists a vast literature on the interplay between infinitely wide neural networks and Gaussian processes, a little is known about analogous interplays with respect to infinitely deep neural networks. Neural networks with independent and identically distributed (i.i.d.) initializations exhibit undesirable forward and backward propagation properties as the number of layers increases, e.g., vanishing dependency on the input, and perfectly correlated outputs for any two inputs. To overcome these drawbacks, Peluchetti and Favaro (2020) considered fully-connected residual networks (ResNets) with network's parameters initialized by means of distributions that shrink as the number of layers increases, thus establishing an interplay between infinitely deep ResNets and solutions to stochastic differential equations, i.e. diffusion processes, and showing that infinitely deep ResNets does not suffer from undesirable forward-propagation properties. In this paper, we review the results of Peluchetti and Favaro (2020), extending them to convolutional ResNets, and we establish analogous backward-propagation results, which directly relate to the problem of training fully-connected deep ResNets. Then, we investigate the more general setting of doubly infinite neural networks, where both network's width and network's depth grow unboundedly. We focus on doubly infinite fully-connected ResNets, for which we consider i.i.d. initializations. Under this setting, we show that the dynamics of quantities of interest converge, at initialization, to deterministic limits. This allow us to provide analytical expressions for inference, both in the case of weakly trained and fully trained ResNets. Our results highlight a limited expressive power of doubly infinite ResNets when the unscaled network's parameters are i.i.d. and the residual blocks are shallow.

Keywords: convolutional neural network; deep neural network; diffusion process; doubly infinite neural network; neural tangent kernel; residual neural network; stochastic differential equation.

1. Introduction

Modern neural networks featuring a large number of layers (depth) and units per layer (width) have achieved a remarkable performance across many domains (LeCun et al., 2015).

©2021 Stefano Peluchetti and Stefano Favaro.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v22/20-706.html.

Under suitable distributional assumptions for network's parameters, the large-width limit of a neural network is a Gaussian process (Neal, 1995; Matthews et al., 2018). Such an interplay between infinitely wide neural networks and Gaussian processes has contributed to the study of properties of neural networks, and most recently it has become a critical tool for introducing inferential algorithms that directly target the infinite-dimensional setting (Lee et al., 2018; Garriga-Alonso et al., 2019; Lee et al., 2019b; Arora et al., 2019). According to recent studies on infinitely wide neural networks, it seems natural to ask whether there exists an interplay between infinitely deep neural networks and classes of stochastic processes. At a first glance, this interplay might prove elusive. Because of the duality between initialization schemes and Bayesian neural networks, it is well-known that any initialization scheme for neural networks may be interpreted as a prior distribution on network's parameters, and in turns the initialization induces a prior on the network. Therefore, any neural network at initialization may be viewed as a suitable stochastic process indexed by depth, whose distribution is defined through a sequence of conditional distributions mapping from each layer to the next layer. Early works have focused on the stabilization of the variance of key quantities of interest across the layers of neural networks (Glorot and Bengio, 2010; He et al., 2015). More recent works (Poole et al., 2016; Schoenholz et al., 2017; Hayou et al., 2019a) considered the impact of initialization schemes to the forward-propagation of the input signal.

Even when initialized on the edge of chaos (EOC) for optimal forward-propagation of the signal (Hayou et al., 2019a), neural networks with an independent and identically distributed (i.i.d.) initialization exhibit pathological properties as their total depth increases. Arguably, the two most common pathological properties are: i) the dependency on the input signal eventually vanishes for most activation functions (Neal, 1995; Poole et al., 2016; Schoenholz et al., 2017); ii) the distributions of the layers, when viewed as random functions on the input space, eventually concentrate on restrictive families including constant functions (Hayou et al., 2019a). As an illustrative example, in Figure 1 we show functions sampled from the last layer of a feedforward neural network for two activation functions under EOC initialization. For the hyperbolic tangent (tanh) activation function, i.e. $\phi(x) = \tanh(x)$, the input signal has no discernible impact on the output, as can be seen by the constant marginal distributions, and the sampled functions are almost constant. This behavior is representative of most classes of smooth activation functions used in practice (Hayou et al., 2019a). For the rectified linear unit (ReLU) activation function, i.e. $\phi(x) = \max(0, x)$, the input signal affects the variance of the output and the sampled functions are piece-wise linear functions. In both cases, the outputs corresponding to any two input signals end up perfectly correlated. While this analysis applies to feedforward neural networks, residual neural networks (ResNets) suffer from analogous issues (Yang and Schoenholz, 2017). In addition, for the ResNets it is known that the variance of the Gaussian-distributed pre-activations may grow unbounded over the layers.

The above-mentioned pathological properties, as well as other critical difficulties, are determined by the fact that common prior distributions on network's parameters introduce a constant level of randomness over each network's layer. To overcome these difficulties, Peluchetti and Favaro (2020) introduced a prior distribution that depends on the number of layers, in such a way that the distribution of network's parameters shrinks as the number layers increases. Under this novel prior construction, Peluchetti and Favaro (2020) showed

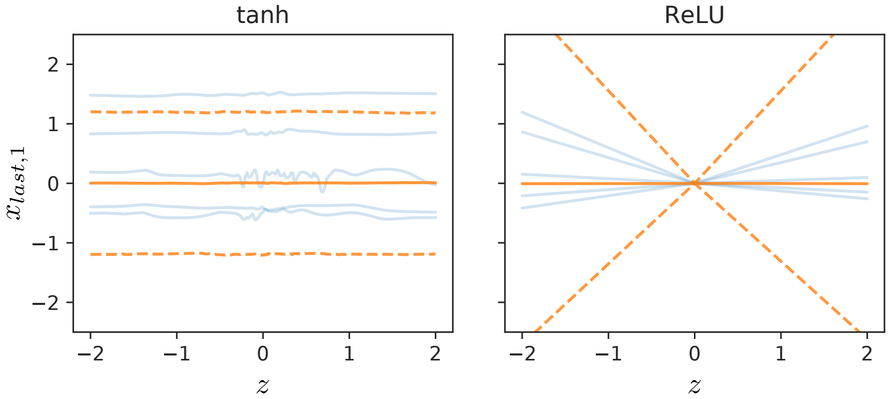

Figure 1: Samples of a given pre-activation from the last layer $x_{last,1}$ of a feedforward neural network with 500 layers of 500 units over an input signal $z \in [-2, 2]$, and tanh and ReLU activation functions. Initialization on the EOC. Samples are displayed in blue, and for each input signal the 5%, 50% and 95% quantiles are displayed in orange.

that fully-connected ResNets converge to certain classes of diffusion processes on a finite time interval, jointly over multiple input signals, as the number of layers increases. The conditions required for attaining the convergence to diffusion processes provide with a concrete guideline for selecting compatible neural network architectures, activation functions and distributions for network's parameters. The resulting limiting diffusion processes satisfy stochastic differential equations (SDE) that describe the evolution of infinitely deep neural networks over time (depth). The connection with SDEs sheds light on properties of very deep neural networks in a general framework, which includes finitely wide neural networks and correlated distributions for network's parameters. In particular, Peluchetti and Favaro (2020) showed that the limiting diffusion process is a well-behaved stochastic process in the sense that: i) it retains the dependency from the input signal; ii) it does not suffer from the perfect correlation constraint; iii) it does not collapse to a deterministic function nor it diverges.

1.1 Our contributions

In this paper, we review the forward-propagation results of Peluchetti and Favaro (2020), and we extend them to convolutional ResNets. Then, we study analogous backward-propagation results, which relate to the fundamental problem of training ResNets. Stochastic gradient descent (SGD) is arguably the most common paradigm for training neural networks (Robbins and Monro, 1951; Bottou et al., 2018). Focusing on the gradient backward-propagation, in neural networks we face a large number of Jacobian matrix multiplications for computing the gradients with respect to the networks' parameters of the lowest layers. This may result in a vanishing (or exploding) gradient, that is the gradient's magnitude in the the lowest layers goes to zero (or grows unbounded) as the number of layers increases. As SGD relies on gradients to perform the updates of the parameters, a vanishing (or exploding) gradient is detrimental to the training performance. The information propagation literature covers this setting, with results qualitatively similar to the forward-propagation analysis (Schoenholz et al., 2017; Hayou et al., 2019a; Yang and Schoenholz, 2017). Our study is critical to the performance of the training. We show that the Jacobian matrix of any layer with respect to the input layer converges to a matrix-valued diffusion process, which is the solution of a matrix SDE. Moreover, under appropriate non-explosivity conditions such a matrix-valued limiting diffusion process is shown to be invertible with dynamics given by a related matrix-valeud SDE. These results imply that in large-depth limit the Jacobian of the final layer with respect to any layer is again well-behaved and that exploding gradients are not possible.

We conclude our study by extending the results of Peluchetti and Favaro (2020), as well as the corresponding backward-propagation results presented in this paper, to the general setting of infinitely deep and infinitely wide ResNets. For short, these neural networks are referred to as doubly infinite ResNets. In such a setting, we assume that network's parameters are i.i.d., and we focus on doubly infinite ResNets defined through a restricted class of activation functions. With regards to the forward-propagation analysis of doubly infinite ResNets, we show that the dynamic of the neural network simplifies, and quantities of interest are either analytically available or can be efficiently approximated by numerical computations. Moreover, we show that the distribution in function space of a doubly infinite ResNet converges to the distribution of a Gaussian process with an affine kernel. With regards to the backward-propagation analysis, the recent literature on the Neural Tangent Kernel (NTK) investigates problem of training a neural network by means of gradient descent with infinitesimally small learning rate and quadratic loss for infinitely wide neural networks (Jacot et al., 2018; Arora et al., 2019). We establish a connection with this line of research, showing that for doubly infinite ResNets the NTK at initialization converges to an affine kernel. Under suitable assumptions for our class of doubly infinite ResNets, this result implies that weakly and fully trained neural networks with a large depth and width collapse to linear regression.

1.2 Organization of the paper

The paper is structured as follows. Section 2 contains some preliminary results on diffusion process approximations of discrete-time stochastic processes, and the notation to be used throughout the paper. In Section 3 we review the forward-propagation results of Peluchetti and Favaro (2020), and we extend them to convolutional ResNets. Section 4 contains the backward-propagation analysis for infinitely deep ResNets, whereas Section 5 contains forward-propagation and backward-propagation analyses for doubly infinite ResNets. In Section 6 we present numerical experiments, and in Section 7 we discuss our work and directions for future work. Proofs of our results and additional plots are deferred to Appendix A and Appendix B, respectively.

2. Preliminaries on diffusion process approximations, and notation

We recall assumptions and results for diffusion process approximations of discrete-time stochastic processes, which are at the basis of the work of Peluchetti and Favaro (2020). Let

$\boldsymbol{x}_l$, for $l = 1, \ldots, L$, denote the l-th layer of a neural network with L layers, and let $\boldsymbol{x}_0$ be the network's input; we refer to the next section for a precise description of $\boldsymbol{x}_l$ in the context of neural networks, and in particular for ResNets. As we consider limiting continuous-time stochastic process, we re-index $\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_L$ on a discrete-time scale. In particular, let $T > 0$ be a terminal time, $\Delta t = T/L$, for each L we establish a correspondence between discrete indices $l \in \mathbb{Z}_+$ and discrete times $t \in \mathbb{R}_+$ by $l = 0, 1, \ldots, L \leftrightarrow t = 0,\ \Delta t,\ 2\Delta t,\ \ldots,\ T$. Without loss of generality, we consider a neural network with input $\boldsymbol{x}_0$ and layers $\boldsymbol{x}_{\Delta t}, \ldots, \boldsymbol{x}_T$, with $\boldsymbol{x}_t$ being a generic layer. Let $p(\boldsymbol{x}_T \,|\, \boldsymbol{x}_0)$ denote the conditional distribution of the network's output given the input for a neural network at initialization. To enforce desirable properties on $p(\boldsymbol{x}_T \,|\, \boldsymbol{x}_0)$, the strategy of Peluchetti and Favaro (2020) consists in having the neural networks to converge, as L goes to infinity, i.e. $\Delta t \downarrow 0$, to a continuous-time stochastic process on the finite time interval $[0, T]$. In this case, for L large enough, the conditional distribution $p(\boldsymbol{x}_T \,|\, \boldsymbol{x}_0)$ is close to the distribution of the limiting process at T given the same $\boldsymbol{x}_0$, with the limiting process being chosen in such a way to make the transition density function well behaved. Among possible constructions of continuous-time stochastic processes as limits of discrete-time processes, here we consider the case where the limiting stochastic process has continuous paths. In all the neural network architectures considered in the present paper, each network's layer depends exclusively on the previous one, and hence $\boldsymbol{x}_t$ has the Markov property. These conditions identify diffusion processes (Stroock and Varadhan, 2006) that are continuous-time Markov processes with continuous paths as natural candidates for the limiting process.

Let $\boldsymbol{x}_t$ denote a generic D-dimensional discrete-time Markov process, and let $\Delta\boldsymbol{x}_t = \boldsymbol{x}_{t+\Delta t} - \boldsymbol{x}_t$ be the corresponding forward increments of the stochastic process. We recall assumptions that imply the convergence of $\boldsymbol{x}_t$ to a limiting diffusion process solution of a SDE. Such a SDE is referred to as the limiting SDE. In particular, it is implicit that the conditional distribution $p(\boldsymbol{x}_{t+\Delta t} \,|\, \boldsymbol{x}_t)$ depends on Δt for the limiting diffusion process to exist as required.

Assumption 1 (instantaneous mean function and covariance function) *There exist a function $\mu(x) : \mathbb{R}^D \to \mathbb{R}^D$ and a function $\sigma^2(x) : \mathbb{R}^D \to \mathbb{R}^{D\times D}$ such that for some $\delta > 0$ it holds*

$$\lim_{\Delta t \downarrow 0} \frac{\mathbb{E}[\Delta\boldsymbol{x}_t \,|\, \boldsymbol{x}_t]}{\Delta t} = \mu(\boldsymbol{x}_t), \tag{1}$$

$$\lim_{\Delta t \downarrow 0} \frac{\mathbb{V}[\Delta\boldsymbol{x}_t \,|\, \boldsymbol{x}_t]}{\Delta t} = \sigma^2(\boldsymbol{x}_t) \tag{2}$$

and

$$\lim_{\Delta t \downarrow 0} \frac{\mathbb{E}[(\Delta\boldsymbol{x}_t)^{2+\delta} \,|\, \boldsymbol{x}_t]}{\Delta t} = 0 \tag{3}$$

uniformly on compact sets of $\mathbb{R}^D$ for each component, where $\mu(x)$ and $\sigma^2(x)$ are continuous functions, and $\sigma^2(x)$ is positive semi-definite, i.e. $\sigma^2(x) = \sigma(x)\sigma(x)^\top$ for $\sigma(x) : \mathbb{R}^D \to \mathbb{R}^{D\times D}$.

The infinitesimal evolution of the diffusion processes considered in the present work is characterized by their instantaneous mean vector (1) and instantaneous covariance matrix

(2). That is, the first two limits in Assumption 1 pinpoint the form of the limiting SDE. The condition (3) represents a technical condition, in the sense that it allows us to consider the limits (1) and (2) instead of their truncated version. We refer to Nelson (1990) for additional details on Assumption 1 and related assumptions. The next theorem establishes that, under Assumption 1 and additional assumptions, in the limit of $\boldsymbol{x}_t$ can be embedded in the solution of an SDE.

Theorem 1 *Under Assumption 1, extend the discrete-time stochastic process $\boldsymbol{x}_t$ to the stochastic process $\overline{\boldsymbol{x}}_t$ on $t \in [0, T]$ by continuous-on-right step-wise-constant interpolation of $\boldsymbol{x}_t$, i.e.*

$$\overline{\boldsymbol{x}}_t = \boldsymbol{x}_u \mathbb{1}_{u \leq t < u+\Delta t} \qquad u \in \{0, \Delta t, 2\Delta t, \dots, T\}. \tag{4}$$

According to the construction (4), consider the D-dimensional SDE on $[0, T]$ with initial value $x_0 = \boldsymbol{x}_0$, drift vector $\mu(x)$ given by (1), and diffusion matrix $\sigma(x)$ given by a square root of (2):

$$dx_t = \mu(x_t)dt + \sigma(x_t)dB_t, \tag{5}$$

where B_t is a D-dimensional Brownian motion (BM) with independent components, and (5) means that

$$x_T = x_0 + \int_0^T \mu(x_t)dt + \int_0^T \sigma(x_t)dB_t,$$

with the first and the second integral being a Riemann integral and an Ito integral, respectively. If the SDE (5) admits a weak solution, and if this solution is unique in law and non-explosive, then the stochastic process $\overline{\boldsymbol{x}}_t$ defined in (4) converges in distribution to the solution of the SDE (5). This result still holds true for an independent and square integrable random variable $\boldsymbol{x}_0 \sim p(\boldsymbol{x}_0)$, provided that the driving BM is independent of $\boldsymbol{x}_0$. In both cases the convergence in distribution is on $\mathcal{D}([0, \infty), \mathbb{R}^D)$, the space of $\mathbb{R}^D$-valued processes on $[0, \infty)$ which are continuous from the right with finite left limits, endowed with the Skorohod metric (Billingsley, 1999, Chapter 3).

We are dealing with three stochastic processes: i) the discrete-time stochastic process $\boldsymbol{x}_t$; ii) the continuous-time interpolation process $\overline{\boldsymbol{x}}_t$ of $\boldsymbol{x}_t$; iii) the limiting diffusion process x_t of $\boldsymbol{x}_t$. In Theorem 1, the continuous-time interpolation $\overline{\boldsymbol{x}}_t$ of $\boldsymbol{x}_t$ is introduced because we are seeking a continuous-time limiting process from a discrete-time stochastic process. The convergence established in Theorem 1 is a strong convergence in the sense that it concerns with the convergence of the distribution of the stochastic process $(\overline{\boldsymbol{x}}_t)_{t\in[0,T]}$ as a stochastic object on the finite interval $[0, T]$ to the limiting diffusion process $(x_t)_{t\in[0,T]}$ as $L \uparrow \infty$. We consider weak solutions, as opposed to a strong solution, where it suffices that a BM B_t can be found such that a solution can be obtained (Øksendal, 2003, Section 5.3). The focus on weak solutions and uniqueness in law of such a solutions, also known as weak uniqueness, is justified by our interest in the distributional properties of the limiting behavior of $\boldsymbol{x}_t$. In particular, weak solutions and uniqueness enable us to consider weaker requirements for attaining convergence of $\boldsymbol{x}_t$. In particular, consider the following discretization of the SDE (5):

$$x_{t+\Delta t} = x_t + \mu(x_t)\Delta t + \sigma(x_t)\zeta_t\sqrt{\Delta t}, \tag{6}$$

where ζ_t is a D-dimensional random vector whose components are i.i.d. as standard Gaussian random variables (mean 0 and variance 1). Under suitable conditions, and in an appropriate sense (Kloeden and Platen, 1992, Sections 10.2 and 14.1), it can be proved that the SDE (6) converges to the SDE (5). In the deterministic part of the SDE (6) we recognize the so-called Euler discretization of an ODE. In particular, Theorem 1 postulates the existence and uniqueness in law of the weak solution of the limiting SDE, and its non-explosive behavior. This following assumptions state sufficient conditions for the postulated solution of the limiting SDE.

Assumption 2 (weak solution and uniqueness in law on compact sets) *The mean function $\mu(x)$ and the covariance function $\sigma(x)$ are twice continuous and differentiable functions.*

Assumption 3 (non-explosive solution) *There exists a finite $C > 0$ such that for each $x \in \mathbb{R}^D$:*

$$\|\mu(x)\| + \|\sigma(x)\| \leq C(1 + \|x\|).$$

If Assumption 1 and Assumption 2 hold true and Assumption 3 does not hold, then we still obtain convergence to the solution x_t of the SDE (5). However, x_t might diverge to infinity with positive probability on any time interval. We will return to this problem in the next section.

We conclude by introducing the main notation to be used throughout the paper: tensors (matrices, vectors) are indexed via subscripts (u_i, $u_{i,j}$, $\dots$), and we make use of $\bullet$ to index all elements and of : to index ranges; we make no distinction between vectors and $n \times 1$ matrices, i.e. vectors are assumed to be column vectors; for a matrix u, $u^\top$ is its transpose, and if u is square $\mathrm{diag}(u)$ is its diagonal vector and $\mathrm{Tr}(u)$ is its trace; the norm of a vector u is $\|u\| = \sqrt{u^\top u}$; if v is another vector their inner product is $\langle u, v\rangle = u^\top v$; the norm of a matrix u is $\|u\| = \sqrt{\mathrm{Tr}(u^\top u)}$; for two matrices u and v, uv stands for the matrix multiplication, $u \otimes v$ for Kronecker's tensor product and $u \odot v$ for the element-wise product; we assume that matrix multiplication has higher precedence than element-wise product; for a tensor u, $\mathrm{vec}(u)$ is its vectorization (row-wise for matrices, with elements being traversed starting from the last dimension); we make use of I to denote the identity matrix and of 1 to denote a vector of ones; for random variables z and w, let $\mathrm{var}[z]$ be the variance of z, $\mathrm{cov}[z, w]$ be the covariance between z and w and $\rho[z, w]$ be their correlation; for random vectors $z \in \mathbb{R}^r$ and $w \in \mathbb{R}^c$ the $r \times c$ cross-covariance matrix $\mathbb{C}[z, w]$ is given by $\mathbb{C}[z, w]_{i,j} = \mathrm{cov}[z_i, w_j]$; the $r \times r$ covariance matrix of z is thus $\mathbb{V}[z] = \mathbb{C}[z, z]$; the expectation $\mathbb{E}[z]$ of a random tensor z is defined as the tensor of the expectations of its elements; for D-dimensional stochastic processes z_t and w_t, we make use of $[z]_t$ to denote the quadratic variation, which is a D-dimensional vector, and of $[z, w]_t$ to denote the quadratic covariation, which is a $D \times D$-dimensional matrix; for a differentiable function $f : \mathbb{R}^k \to \mathbb{R}$, $\nabla_u f(u) \in \mathbb{R}^k$ is its corresponding gradient vector; if $f : \mathbb{R}^k \to \mathbb{R}^m$, $J(f(u), u) \in \mathbb{R}^{m \times k}$ is its Jacobian matrix. We refer to Table 1 for the complete notation.

3. Infinitely deep ResNets

Peluchetti and Favaro (2020) investigated the implications of Assumption 1, Assumption 2 and Assumption 3 in the context of forward-propagation analysis of fully-connected ResNets, thus establishing an interplay between infinitely deep ResNets and solutions of SDEs, that is

Table 1: Notation: symbols and variables

Symbol	Description
u_i, $u_{i,j}$, $u_{i,j,l}$, $\ldots$	vector, matrix, tensor indexing
	we use : for ranges, $\bullet$ for all elements, $\star$ for convolutional patches
$u^\top$	matrix transpose
$\mathrm{diag}(u)$	matrix diagonal
$\mathrm{Tr}(u)$	matrix trace
$\lVert u \rVert$	vector or matrix norm
$\langle u, v \rangle$	inner product of 2 vectors
uv	matrix multiplication
$u \otimes v$	Kronecker's tensor product
$u \odot v$	element-wise tensor product
$\mathrm{vec}(u)$	tensor vectorization
I	identity matrix
1	vector of ones
$\mathbb{1}$	indicator function
$\mathrm{var}[z]$	variance of a random variable
$\mathrm{cov}[z, w]$	covariance between 2 random variables
$\rho[z, w]$	correlation between 2 random variables
$\mathbb{C}[z, w]$	cross-covariance matrix between 2 random vectors
$\mathbb{V}[z]$	covariance matrix of a random vector
$\mathbb{E}[z]$	expectation of a random tensor
$[z]_t$	quadratic variation of a stochastic process
$[z, w]_t$	quadratic covariation between 2 stochastic processes
$\nabla_u f(u)$	function gradient
$J(f(u), u)$	function jacobian
$\mathcal{N}, \mathcal{MN}, \mathcal{TN}$	Gaussian, matrix-Gaussian, tensor-Gaussian distributions
Variable	Description
D	width of each layer, indexed by d
L	number or layers, indexed by l
T	integration time, indexed by t
Δt	time discretization interval
$\boldsymbol{x}_l$	neural network layers
$\boldsymbol{x}_t$	neural network layers over time
$\overline{\boldsymbol{x}}_t$	interpolated neural network layers over time
x_t	limit diffusion of $\overline{\boldsymbol{x}}_t$ (as $\Delta t \downarrow 0$)
$\boldsymbol{g}_t$	neural network output-input Jacobian over time
$\overline{\boldsymbol{g}}_t$	interpolated neural network output-input Jacobian over time
g_t	limit diffusion of $\overline{\boldsymbol{g}}_t$ (as $\Delta t \downarrow 0$)
ϕ, ψ	activation functions
A_t, a_t	neural network weights and biases
W_t, b_t	neural network weights and biases diffusions

diffusion processes. In this section, we review and discuss the main results established in Peluchetti and Favaro (2020), and we extend them to the more general setting of convolutional ResNets.

3.1 Fully-connected ResNets

We consider unmodified, albeit simplified, standard neural network architectures, which is in line with the research area of forward-propagation analysis (Poole et al., 2016; Schoenholz et al., 2017; Hayou et al., 2019a). Consistently with Section 2, we consider a discrete-time stochastic process $\boldsymbol{x}_t \in \mathbb{R}^D$, which is assumed to be of constant dimensionality in order that $\Delta \boldsymbol{x}_t$ is well-defined. For Assumption 1 to hold, we need $\Pr(\|\Delta \boldsymbol{x}_t\| > \varepsilon \,|\, \boldsymbol{x}_t) \downarrow 0$ as $\Delta t \downarrow 0$ for any $\varepsilon > 0$, that is we require the increments of the stochastic process $\boldsymbol{x}_t$ to vanish eventually. Intuitively, this is because the continuity of the paths of the limiting diffusion process. According to the stochastic process $\boldsymbol{x}_t$, a fully-connected feedforward neural network is expressed by

$$\boldsymbol{x}_{t+\Delta t} = f_t(\boldsymbol{x}_t) = \phi(A_t \boldsymbol{x}_t + a_t),$$

for a nonlinear activation function $\phi : \mathbb{R} \to \mathbb{R}$ applied element-wise. We refer to $A_t \in \mathbb{R}^{D \times D}$ as weights and to $a_t \in \mathbb{R}^D$ as biases. Hence $\Delta \boldsymbol{x}_t = \phi(A_t \boldsymbol{x}_t + a_t) - \boldsymbol{x}_t$. Shrinking increments would imply that for all x, $\phi(A_t x + a_t)$ can be made arbitrarily concentrated around x with a suitable choice of distributions for (A_t, a_t). This cannot be achieved unless ϕ is linear or the distribution of (A_t, a_t) depends on x. Indeed, fixing x determines the values around which (A_t, a_t) need to concentrate for the increments to vanish (if any), hence the increments will not vanish for a different $x' \neq x$, a fact that is most easily seen in the specific case where (A_t, a_t) are scalars.

The same lines of reasoning rules out the residual network architecture (ResNet), originally introduced in the work of He et al. (2016a). In particular, in the ResNet architecture we write $\boldsymbol{x}_{t+\Delta t} = f_t(\boldsymbol{x}_t + r_t(\boldsymbol{x}_t))$. This leaves us with the identity ResNet of He et al. (2016b) where we write

$$\boldsymbol{x}_{t+\Delta t} = \boldsymbol{x}_t + r_t(\boldsymbol{x}_t) \tag{7}$$

for some choice of r_t, the residual blocks, which we require to eventually vanish. Each r_t results from an interleaved application of affine transforms and non-linear activation functions. Peluchetti and Favaro (2020) considered the case of shallow residual blocks, such that (7) becomes

$$\boldsymbol{x}_{t+\Delta t} = \boldsymbol{x}_t + \phi(A_t \psi(\boldsymbol{x}_t) + a_t) \tag{8}$$

for two activation functions $\phi : \mathbb{R} \to \mathbb{R}$, $\psi : \mathbb{R} \to \mathbb{R}$ which are applied element-wise. We remark that the non-standard approach of using of 2 activation functions, i.e. ϕ, ψ, is applied to cover the case of shallow residual blocks in full generality. For a shallow residual block r_t, the vanishing increments requirement is satisfied by having the distributions of weights A_t and biases a_t both concentrate around 0 provided that $\phi(0) = 0$. Furthermore, it proves to be advantageous to consider weights and biases given by increments of diffusion processes corresponding to solvable SDEs. Notice that the use of increments implies independence across layers, and the simplest parametrization corresponds to typical fully i.i.d. initializations used in practice.

Assumption 4 (distributions of network's parameters and scaling) *Let W_t be a diffusion process in $\mathbb{R}^{D\times D}$ and let b_t be a diffusion process in $\mathbb{R}^{D}$, which are defined as the solutions of*

$$dW_t = \mu^W dt + d\widetilde{W}_t \tag{9}$$

with $d\,\mathrm{vec}(\widetilde{W}_t) = \sigma^W d\,\mathrm{vec}(B_t^W)$, and

$$db_t = \mu^b dt + \sigma^b dB_t^b, \tag{10}$$

respectively, where $B_t^W \in \mathbb{R}^{D\times D}$ and $B_t^b \in \mathbb{R}^D$ are independent BMs with independent components, $\mu^b \in \mathbb{R}^D$, $\sigma^W \in \mathbb{R}^{D^2\times D^2}$, $\sigma^b \in \mathbb{R}^{D\times D}$, and $\Sigma^W = \sigma^W {\sigma^W}^\top$, $\Sigma^b = \sigma^b {\sigma^b}^\top$ are positive semi-definite. That is, W_t and b_t are matrix-valued and vector-valued diffusion processes, respectively, solutions of SDEs with deterministic time-homogeneous drift and diffusion coefficients.

Now, we consider the setting of Assumption 4. In particular, the discretizations of the diffusion processes W_t and b_t displayed in (9) and (10), respectively, admit exact representations as

$$\Delta W_t = \mu^W \Delta t + \varepsilon_t^W \sqrt{\Delta t}$$

and

$$\Delta b_t = \mu^b \Delta t + \varepsilon_t^b \sqrt{\Delta t},$$

where $\mathrm{vec}(\varepsilon_t^W) \overset{i.i.d.}{\sim} \mathcal{N}_{D^2}\big(0, \Sigma^W\big)$ and $\varepsilon_t^b \overset{i.i.d.}{\sim} \mathcal{N}_D\big(0, \Sigma^b\big)$ for $t = \Delta t, \ldots, T$, with $\mathcal{N}$ being the multivariate Gaussian distribution. According to the above discretizations of W_t and b_t, we consider residual blocks where $A_t = \Delta W_t$ and $a_t = \Delta b_t$, and then we write the ResNet as follows

$$\boldsymbol{x}_{t+\Delta t} = \boldsymbol{x}_t + \phi(\Delta W_t \psi(\boldsymbol{x}_t) + \Delta b_t). \tag{11}$$

Assumption 4 covers the case where network's parameters are i.i.d. across layers according to a multivariate Gaussian distribution, up to the required scaling which is necessary to obtain the desired limiting diffusion process. By considering deterministic but time-dependent $\mu_t^W, \mu_t^b, \Sigma_t^W, \Sigma_t^b$ the extension to layer-dependent distributions is immediate. More generally, we can consider W_t and b_t driven by arbitrary SDEs. Moreover, dependencies across the parameters of different layers can be accommodated by introducing additional SDE-driven processes, driving the evolution of W_t and b_t. We do not pursue further these directions in the present work. As for the activation functions, we require the following regularity assumptions.

Assumption 5 (regularity of activation functions) *The activation function $\psi : \mathbb{R} \to \mathbb{R}$ is continuously differentiable two times on $\mathbb{R}$. The activation function $\phi : \mathbb{R} \to \mathbb{R}$ is such that $\phi(0) = 0$ and, moreover, ϕ is continuously differentiable three times on $\mathbb{R}$ and its second and third derivatives have at most exponential tails growth, that is for some $k > 0$ it holds true*

$$\lim_{|x|\uparrow\infty} \frac{|\phi''(x)|}{e^{k|x|}} + \lim_{|x|\uparrow\infty} \frac{|\phi'''(x)|}{e^{k|x|}} < \infty.$$

The assumption $\phi(0) = 0$ and the smoothness assumptions on ϕ have been shown to be key requirements to achieve good signal propagation (Hayou et al., 2019a,b). On the basis of Assumption 4 and Assumption 5, we now report the main result of Peluchetti and Favaro (2020).

Theorem 2 *Under Assumption 4 and Assumption 5, Assumption 1 with $\delta = 2$ holds true for the ResNet $\boldsymbol{x}_t$ defined in (11). The infinitesimal mean $\mu(x)$and covariance $\sigma^2(x)$ functions are such that*

$$\mu(x) = \phi'(0)(\mu^b + \mu^W \psi(x)) + \frac{1}{2}\phi''(0)\,\mathrm{diag}(\mathbb{V}[\varepsilon_t^W \psi(x) + \varepsilon_t^b \,|\, x])$$

and

$$\sigma^2(x) = \phi'(0)^2\,\mathbb{V}[\varepsilon_t^W \psi(x) + \varepsilon_t^b \,\|, x].$$

Furthermore, Assumption 2 is satisfied and, by means of Theorem 1, the continuous-time interpolation $\overline{\boldsymbol{x}}_t$ of the ResNet $\boldsymbol{x}_t$ converges in distribution to the solution on $[0, T]$ of the SDE

$$\begin{aligned} dx_t = {} & \phi'(0)\big(\,\mathbb{V}[\varepsilon_t^W \psi(x_t) + \varepsilon_t^b \,|\, x_t]\big)^{1/2} dB_t \\ & + \big(\phi'(0)(\mu^b + \mu^W \psi(x_t)) + \frac{1}{2}\phi''(0)\,\mathrm{diag}(\mathbb{V}[\varepsilon_t^W \psi(x_t) + \varepsilon_t^b \,|\, x_t])\big) dt \end{aligned} \tag{12}$$

with initial value $x_0 = \boldsymbol{x}_0$ where B_t is a D-dimensional BM vector with independent components.

Theorem 2 does not establish a direct connection between the limiting diffusion process x_t and the driving sources of randomness provided by the diffusion processes W_t and b_t. Since we are interested in the study of properties of ResNets in the function space, that is over multiple input signals, a brute force approach would require to establish limiting diffusion processes as in Theorem 2 for an enlarged discrete-time stochastic process $\boldsymbol{x}_t = [\boldsymbol{x}_t^{(1)} \cdots \boldsymbol{x}_t^{(N)}] \in \mathbb{R}^{DN}$ corresponding to a collection of N initial values $\boldsymbol{x}_0 = [\boldsymbol{x}_0^{(1)} \cdots \boldsymbol{x}_0^{(N)}]$. Instead, Peluchetti and Favaro (2020) showed that the limiting SDE is equivalent in law to the solution of another SDE which preserves the dependency on the driving sources of randomness. From here on, let $\boldsymbol{x}_t^{(i)}$ and $\boldsymbol{x}_t^{(j)}$ denote ResNets corresponding to two initial values $\boldsymbol{x}_0^{(i)}$ and $\boldsymbol{x}_0^{(j)}$, respectively. Moreover, let $x_t^{(i)}$ and $x_t^{(j)}$ denote limiting diffusion processes corresponding to the same two initial values, i.e. $x_0^{(i)} = \boldsymbol{x}_0^{(i)}$ and $x_0^{(j)} = \boldsymbol{x}_0^{(j)}$ respectively. Hereafter, we continue to make use of $\boldsymbol{x}_t$ for $\boldsymbol{x}_t^{(i)}$ and x_t for $x_t^{(i)}$ when no confusion arises. Under Assumption 4 and Assumption 5, the next corollary characterizes the limiting diffusion process of the ResNet $\boldsymbol{x}_t^{(i)}$

Corollary 3 *Let $\boldsymbol{x}_t^{(i)}$ be the ResNet corresponding to the initial value $\boldsymbol{x}_0^{(i)}$. Under Assumption 4 and Assumption 5, the limiting diffusion process $x_t^{(i)}$ of $\boldsymbol{x}_t^{(i)}$ is the solution on $[0, T]$ of the SDE*

$$dx_t^{(i)} = \phi'(0)(dW_t \psi(x_t^{(i)}) + db_t) + \frac{1}{2}\phi''(0)(d[W\psi(x^{(i)})]_t + d[b]_t), \tag{13}$$

and

$$d[x^{(i)}, x^{(j)}]_t = \phi'(0)^2(d[W\psi(x^{(i)}), W\psi(x^{(j)})]_t + d[b, b]_t). \tag{14}$$

A direct consequence of Theorem 2 is that the distribution of the ResNet output given the input, i.e. $p(\boldsymbol{x}_T \,|\, \boldsymbol{x}_0)$, converges to the transition density $p(x_T \,|\, x_0)$ of the solution of (13). In particular, as T is finite, the dependency on the input does not vanish in the limit of infinite total depth L and can be controlled via the distributions of network's parameters and the integration time T. The representations (12) and (13) are complementary: depending on the situation it will prove advantageous to use one or the other. Theorem 2 and Corollary 3 are general in the sense that we allow for an arbitrary covariance structure between the elements of ε_t^W, i.e. an arbitrary (constant and deterministic) quadratic covariation for W_t. This makes it difficult to derive more explicit results, and is also an impractical approach as the parametrization requires $\mathcal{O}(D^4)$ elements. Peluchetti and Favaro (2020) then considered more restrictive distribution assumptions with a more manageable $\mathcal{O}(D^2)$ parametrization cost.

Assumption 6 (matrix Gaussian network's parameters) *Let $b_t, \mu^b, \sigma_b, B_t^b, \mu^W, B_t^W$ be defined as in Assumption 4, and let W_t be a diffusion process in $\mathbb{R}^{D\times D}$, which is defined as the solution of*

$$dW_t = \mu^W dt + \sigma^{W_O} dB_t^W \sigma^{W_I},$$

where $\sigma^{W_O}, \sigma^{W_I} \in \mathbb{R}^{D\times D}$ and $\Sigma^{W_O} = \sigma^{W_O}{\sigma^{W_O}}^\top$, $\Sigma^{W_I} = {\sigma^{W_I}}^\top \sigma^{W_I}$ are positive semi-definite matrices.

We consider the setting of Assumption 6. Under this setting the discretization of W_t satisfies

$$\varepsilon_t^W \overset{i.i.d.}{\sim} \mathcal{MN}_{D,D}\big(0, \Sigma^{W_O}, \Sigma^{W_I}\big)$$

for $t = \Delta t, \ldots, T$, where $\mathcal{MN}$ stands for the matrix Gaussian distribution. This is a direct consequence of the following fact: if $\zeta \sim \mathcal{MN}(0, \mathrm{I}, \mathrm{I})$ then $A\zeta B \sim \mathcal{MN}(0, AA^\top, B^\top B)$. The fundamental property of the $\mathcal{MN}$ distributions is that the covariance factorizes as $\mathrm{cov}(\varepsilon_{o,i}^W, \varepsilon_{o',i'}^W) = \Sigma_{o,o'}^{W_O}\Sigma_{i,i'}^{W_I}$. The reader is referred to Gupta and Nagar (1999) for a comprehensive treatment of matrix Gaussian distributions and their properties. Under Assumption 5 and Assumption 6, the next corollary characterizes the limiting diffusion process of the ResNet $\boldsymbol{x}_t^{(i)}$.

Corollary 4 *Let $\boldsymbol{x}_t^{(i)}$ be the ResNet corresponding to the initial value $\boldsymbol{x}_0^{(i)}$. Under Assumption 5 and Assumption 6, the limiting diffusion process $x_t^{(i)}$ of $\boldsymbol{x}_t^{(i)}$ is the solution on $[0, T]$ of the SDE*

$$\begin{aligned} dx_t^{(i)} = {} & \phi'(0)\big((\mu^W\psi(x_t^{(i)}) + \mu^b)dt + \sigma^{W_O} dB_t^W \sigma^{W_I}\psi(x_t^{(i)}) + \sigma^b dB_t^b\big) \\ & + \frac{1}{2}\phi''(0)\,\mathrm{diag}\,\big(\Sigma^b + \Sigma^{W_O}(\psi(x_t^{(i)})^\top \Sigma^{W_I}\psi(x_t^{(i)}))\big)dt, \end{aligned} \tag{15}$$

and

$$d[x^{(i)}, x^{(j)}]_t = \phi'(0)^2\big(\Sigma^b + \Sigma^{W_O}\psi(x_t^{(i)})^\top \Sigma^{W_I}\psi(x_t^{(j)})\big)dt.$$

Under Assumption 6 we have that the covariance function $\mathbb{V}[\varepsilon_t^W\psi(x_t) + \varepsilon_t^b \,|\, x_t]$ is given by $\Sigma^b + \Sigma^{W_O}(\psi(x_t)^\top\Sigma^{W_I}\psi(x_t))$. In particular, the dependency on the state x_t in Equation (12) goes through a linear transformation and a weighted inner product. This fact sheds light on the impact of introducing dependencies among row and columns of network's parameters $A_t = \Delta W_t$. Specifically, the matrix Σ^{W_I} defines the structure of the inner weighted product, while the matrix Σ^{W_O} defines how such transforms affect each dimension $d \in D$. Peluchetti and Favaro (2020) completed their study by considering the simplest fully i.i.d. setting with the assumption of centered distributions for W_t and b_t. Fully i.i.d. initializations are commonly used in training of neural networks. A scaling of the weights by $D^{-1/2}$ is also introduced; this is the same scaling used to obtain Gaussian process limits in infinitely wide networks (Neal, 1995; Lee et al., 2018). In Section 5 we show that such a scaling allows to study $D \uparrow \infty$.

Assumption 7 (fully i.i.d. network's parameters) *Let W_t be a diffusion process in $\mathbb{R}^{D\times D}$ and let b_t be a diffusion process in $\mathbb{R}^D$, which are defined as the solutions of the SDSs*

$$dW_t = \frac{\sigma_w}{\sqrt{D}} dB^W_t \tag{16}$$

and

$$db_t = \sigma_b dB^b_t, \tag{17}$$

respectively, where $B^W_t \in \mathbb{R}^{D\times D}$ and $B^b_t \in \mathbb{R}^D$ are independent BMs and $\sigma_w > 0, \sigma_b > 0$ are scalars.

Now, we consider the setting of Assumption 7. In particular, the discretizations of the diffusion processes W_t and b_t displayed in (16) and (17), respectively, admit exact representations as

$$\Delta W_t = \varepsilon^W_t \frac{\sigma_w}{\sqrt{D}} \sqrt{\Delta t} \tag{18}$$

and

$$\Delta b_t = \varepsilon^b_t \sigma_b \sqrt{\Delta t} \tag{19}$$

where $\varepsilon^W_t \overset{i.i.d.}{\sim} \mathcal{MN}_{D,D}\big(0, \mathrm{I}_D, \mathrm{I}_D\big)$ and $\varepsilon^b_t \overset{i.i.d.}{\sim} \mathcal{N}_D\big(0, \mathrm{I}_D\big)$. Under Assumption 5 and Assumption 7, the next corollary characterizes the limiting diffusion process of the ResNet $\boldsymbol{x}^{(i)}_t$.

Corollary 5 *et $\boldsymbol{x}^{(i)}_t$ be the ResNet corresponding to the initial value $\boldsymbol{x}^{(i)}_0$. Under Assumption 5 and Assumption 7, the limiting diffusion process $x^{(i)}_t$ of $\boldsymbol{x}^{(i)}_t$ is the solution on $[0, T]$ of the SDE*

$$\begin{aligned} dx^{(i)}_t = {} & \phi'(0)\big(\frac{\sigma_w}{\sqrt{D}} dB^W_t \psi(x^{(i)}_t) + \sigma_b dB^b_t\big) \\ & + \frac{1}{2}\phi''(0)\big(\sigma_b^2 + \frac{\sigma_w^2}{D}\|\psi(x^{(i)}_t)\|^2)\big)\, \mathrm{I}_D\, dt, \end{aligned} \tag{20}$$

and

$$d[x^{(i)}, x^{(j)}]_t = \phi'(0)^2\big(\sigma_b^2 + \frac{\sigma_w^2}{D}\langle\psi(x^{(i)}_t), \psi(x^{(j)}_t)\rangle\big)\, \mathrm{I}_D\, dt.$$

In all cases the activation function ϕ only impacts the dynamics through its local behavior at the origin, while this is not the case for the activation ψ. Under the setting of Assumption 7, i.e. fully i.i.d. network's parameters, we have that $\mathbb{V}[\varepsilon_t^W \psi(x_t) + \varepsilon_t^b \,|\, x_t]$ is given by $\sigma_b^2 + \frac{\sigma_w^2}{D}\|\psi(x_t)\|^2$. In particular, the dependency on the state x_t in (12) goes only through the norm of x_t, which is permutation invariant in $d \in D$. Accordingly, the distribution of the stochastic processes $x_{t,d}$ is exchangeable across $d \in D$ if the distribution of $x_{0,d}$ is so. We will show in Section 5.3 that, under Assumption 7, as $D \uparrow \infty$ $x_{t,d}$ will become i.i.d. over d if $x_{0,d}$ is so.

Remark 6 *According to Equation (13) and Equation (14), the joint evolution of the diffusion processes $x_t^{(i)}$ and $x_t^{(j)}$ corresponding to two inputs $x_0^{(i)}$ and $x_0^{(j)}$, respectively, is not perfectly correlated. This remains true also in the parameterizations of Assumption 6 and Assumption 7. Thus in the limit of infinite total depth L the distribution in function space does not suffer from the perfect correlation problem. The joint distribution $p(x_T^{(i)}, x_T^{(j)} \,|\, x_0^{(i)}, x_0^{(j)})$ is not Gaussian. We show in Section 5.3 that we recover the Gaussian case as $D \uparrow \infty$ under the parametrization of Assumption 7.*

Remark 7 *A standard time-change result for SDEs (Øksendal, 2003, Theorem 8.5.7) implies that the time-scaling of an SDE is equivalent to multiplying the drift and the diffusion coefficients by the scaling constant and by the square root of the scaling constant, respectively. Furthermore, according to Equation (12), it is possible to compensate changes in the integration time T with changes in the "hyper-parameters" $\mu^b, \mu^W, \Sigma^b, \Sigma^W$ in Assumption 4 to leave the dynamics of (12) invariant. These observations remain true also in the parameterizations of Assumption 6 and Assumption 7. Therefore, without loss of generality, we can restrict to $T = 1$.*

Remark 8 *Without further assumptions, the solutions to the limiting SDEs can be explosive solutions. Assumption 3 is satisfied under all considered distributional assumptions for network's parameters if either: i) ψ exhibits at most square-root growth, in particular ψ is bounded; or ii) ψ exhibits at most linear growth, in particular ψ is the identity function, and $\phi''(0) = 0$, in particular $\phi = \tanh$. We will show in Section 5.3 that, under Assumption 7, as $D \uparrow \infty$ in the case of $\phi''(0) \neq 0$ with ψ the identity function, the explosion time becomes deterministic.*

Remark 9 *The limiting diffusion processes that we have obtained are based on smoothness assumptions for ϕ. Given the popularity of the* ReLU *activation function $\phi(a) = \max(0, a)$, we consider here a brief analysis which includes it. In particular, we assume that $\phi(a)$ is positively homogeneous, i.e. $\phi(\alpha a) = \alpha\phi(a)$ for $\alpha > 0$, h is random variable, and $\gamma > 0$ then: $\mathbb{E}[\phi(h\Delta t^\gamma)/\Delta t] = \mathbb{E}[\phi(h)]\,\Delta t^{\gamma-1}$ and $\mathbb{E}[\phi(h\Delta t^\gamma)^2/\Delta t] = \mathbb{E}\left[\phi(h)^2\right]\Delta t^{2\gamma-1}$. Comparing these results with (1) and (2), we see that unless $\mathbb{E}[\phi(h)] = 0$, the choice of $\gamma = 1/2$ would result in the drift term blowing up. The alternative of choosing $\gamma = 1$ recovers a non stochastic limit which can be interpreted as a particular form of Chen et al. (2018). The positive homogeneity of* ReLU *activations makes equivalent to modify the recursion or reparameterize the parameter.*

We have considered $\boldsymbol{x}_0 \in \mathbb{R}^D$ to be the input of the ResNet. A neural network acts as a function approximator to be fitted to some dataset $\mathcal{D} = (\mathcal{Z}, \mathcal{Y}) = \{(z^{(i)}, y^{(i)})\}_{i=1}^N$ of size

N where $z^{(i)} \in \mathbb{R}^Z$ represents an input and $y^{(i)} \in \mathbb{R}^Y$ represents the corresponding output. Classification problems can be framed in this setting if we use a one-hot representation for $y^{(i)}$. In general, there can be a mismatch between D, Z and Y, making it is necessary to introduce adaptation layers $z^{(i)} \mapsto \boldsymbol{x}_0^{(i)}$ and $\boldsymbol{x}_T^{(i)} \mapsto \widehat{y}^{(i)}$ where $\widehat{y}^{(i)}$ is the network prediction for $z^{(i)}$. As for $\boldsymbol{x}_t$, we denote a single data-point $(z^{(i)}, y^{(i)})$ with (z, y) when no confusion arises.

3.2 Convolutional ResNets

We extend the main results of Peluchetti and Favaro (2020) to the more general setting of convolutional neural networks (CNNs). Such extension relies on the equivalence between convolutional transformations (either at a given position, or over all positions) and specific forms of matrix multiplication. For the sake of simplicity and clarity in exposition, we present our results for 2D convolutions and square filters. Analogous results follow in the more general setting. CNNs are best described by keeping the features, height and width dimensions separated, in which case $\boldsymbol{x}_t$ is a three-dimensional tensor. This does not cause issues to our analysis, since we can consider the vectorization $\mathrm{vec}(\boldsymbol{x}_t)$ which allows us to refer to definitions and results of Section 2. We denote the input image to the convolutional neural network and its layers with $\boldsymbol{x}_t$, $t = 0, \Delta t, \dots, T$. As before $\boldsymbol{x}_t$ needs to be of fixed dimensionality: $\boldsymbol{x}_t \in \mathbb{R}^{U \times V \times D}$, D being the number of channels, and U and V being respectively the height and the width.

We consider square filters of spatial length K, with K being odd, in which case the off-center range of the filter is $E = (K-1)/2$. Assuming unitary strides in both height and width dimensions, constant dimensionality is achieved by padding the width and height dimensions of $\boldsymbol{x}_t$, $t > 0$, with E pixels borders. The padding can be performed arbitrarily here: typically the values next to the boarder are copied or paddings have the same value of a background reference level. We enumerate the set of $P = UV$ positions, where positions are ordered in row-wise manner (the ordering does not affect the results as long as it is the same everywhere). A convolutional transform $x \in \mathbb{R}^{(U \times V \times D)} \mapsto y \in \mathbb{R}^{(U \times V \times D)}$ is obtained by applying (convolving) the same filter $W \in \mathbb{R}^{D \times (U \times V \times D)}$ to the extracted patches $x_{\star p} \in \mathbb{R}^{(U \times V \times D)}$ by matrix multiplication: $y_p = W x_{\star p}$, $y_p \in \mathbb{R}^D$ for each $p = 1, \dots, P$. Parentheses indicate how the dimensions are flattened (vectorized), and each patch is given by $x_{\star p} = x_{\star p, \bullet} = x_{u-E:u+E, v-E:v+E, \bullet}$ for position $p = (u, v)$, $u = 1, \dots, U$, $v = 1, \dots, V$. We incorporate the padding in the patch extraction operation: indexing outside the allowed ranges (which happens for positions at the boarders) returns the padded values. More generally a bias term $b \in \mathbb{R}^D$ can be included resulting in $y_p = W x_{\star p} + b$. See Dumoulin and Visin (2016) and references therein for a comprehensive account. For convenience let $F = UVD$ denote the extracted patch size. We begin with the most generic parametrization for CNNs covered in this work, which corresponds to Assumption 4 for the fully-connected case.

Assumption 8 (distributions of network's parameters and scaling) *Let W_t be a diffusion process in $\mathbb{R}^{D \times K \times K \times D}$ and let b_t be the diffusion process in $\mathbb{R}^D$, which are defined as the solutions of*

$$dW_t = \mu^W dt + d\widetilde{W}_t \tag{21}$$

with $d\,\mathrm{vec}(\widetilde{W}_t) = \sigma^W d\,\mathrm{vec}(B_t^W)$*, and*

$$db_t = \mu^b dt + \sigma^b dB_t^b \tag{22}$$

where $B_t^W \in \mathbb{R}^{D\times K\times K\times D}$ *and* $B_t^b \in \mathbb{R}^D$ *are independent BMs with independent components,* $\mu^W \in \mathbb{R}^{D\times K\times K\times D}, \mu^b \in \mathbb{R}^D$*,* $\sigma^W \in \mathbb{R}^{DF\times DF}, \sigma^b \in \mathbb{R}^{D\times D}$*, and* $\Sigma^W = \sigma^W {\sigma^W}^\top$*,* $\Sigma^b = \sigma^b {\sigma^b}^\top$ *are positive semi-definite. That is,* W_t *and* b_t *are tensor-valued and vector-valued diffusion processes, respectively, solutions of SDEs with deterministic time-homogeneous drift and diffusion coefficients.*

Now, we consider the setting of Assumption 8. In particular, the discretizations of the diffusion processes W_t and b_t displayed in (21) and (22), respectively, admit exact representations as

$$\Delta W_t = \mu^W \Delta t + \varepsilon_t^W \sqrt{\Delta t}$$

and

$$\Delta b_t = \mu^b \Delta t + \varepsilon_t^b \sqrt{\Delta t},$$

where $\mathrm{vec}(\varepsilon_t^W) \overset{i.i.d.}{\sim} \mathcal{N}_{DF}\big(0, \Sigma^W\big)$ and $\varepsilon_t^b \overset{i.i.d.}{\sim} \mathcal{N}_D\big(0, \Sigma^b\big)$ for $t = \Delta t, \dots, T$ where $\mathcal{N}$ stands for the multivariate Gaussian distribution. As in Peluchetti and Favaro (2020), we consider shallow residual blocks and two activation functions. This leads to write the ResNet as follows

$$\boldsymbol{x}_{t+\Delta t,p} = \boldsymbol{x}_{t,p} + \phi(\Delta W_t \psi(\boldsymbol{x}_{t,\star p}) + \Delta b_t) \qquad (p = 1, \dots, P) \tag{23}$$

where $\Delta W_t \psi(\boldsymbol{x}_{t,\star p})$ is computed by means of matrix multiplication as we have explained above. The next theorem states our main convergence result in the setting of convolutional ResNets. In particular, the next theorem provides the convolutional counterpart of Theorem 2.

Theorem 10 *Under Assumption 8 and Assumption 5, Assumption 1 with* $\delta = 2$ *holds true for the ResNet* $\boldsymbol{x}_t$ *defined in (23). The infinitesimal mean* $\mu(x)$ *and covariance* $\sigma^2(x)$ *functions are such that*

$$\mu(x) = \begin{bmatrix} \mu_\star(x_{\star 1}) \\ \vdots \\ \mu_\star(x_{\star P}) \end{bmatrix}$$

where

$$\mu_\star(x_{\star p}) = \phi'(0)(\mu^b + \mu^W \psi(x_{\star p})) + \frac{1}{2}\phi''(0)\,\mathrm{diag}(\mathbb{V}[\varepsilon_t^W \psi(x_{\star p}) + \varepsilon_t^b \,|\, x_{\star p}]),$$

and

$$\sigma^2(x) = \phi'(0)^2 \begin{bmatrix} \sigma_\star^2(x_{\star 1}, x_{\star 1}) & \cdots & \sigma_\star^2(x_{\star 1}, x_{\star P}) \\ \vdots & \vdots & \vdots \\ \sigma_\star^2(x_{\star P}, x_{\star 1}) & \cdots & \sigma_\star^2(x_{\star P}, x_{\star P}) \end{bmatrix}$$

where

$$\sigma_\star^2(x_{\star p}, x_{\star p'}) = \mathbb{C}[\varepsilon_t^W \psi(x_{\star p}) + \varepsilon_t^b, \varepsilon_t^W \psi(x_{\star p'}) + \varepsilon_t^b \,|\, x_{\star p}, x_{\star p'}].$$

Furthermore, Assumption 2 is satisfied and, by means of Theorem 1, the continuous-time interpolation $\mathrm{vec}(\overline{\boldsymbol{x}_t})$ *of the ResNet* $\mathrm{vec}(\boldsymbol{x}_t)$ *converges in distribution to the solution on* $[0,T]$ *of the SDE*

$$\begin{aligned} & d\,\mathrm{vec}(x_t) \\ &= \phi'(0)\begin{bmatrix} \sigma^2_\star(x_{t,\star 1},x_{t,\star 1}) & \cdots & \sigma^2_\star(x_{t,\star 1},x_{t,\star P}) \\ \vdots & \vdots & \vdots \\ \sigma^2_\star(x_{t,\star P},x_{t,\star 1}) & \cdots & \sigma^2_\star(x_{t,\star P},x_{t,\star P}) \end{bmatrix}^{1/2} dB_t + \begin{bmatrix} \mu_\star(x_{t,\star 1}) \\ \vdots \\ \mu_\star(x_{t,\star P}) \end{bmatrix} dt \end{aligned} \tag{24}$$

with initial value $x_0 = \boldsymbol{x}_0$ *where* B_t *is a* PD*-dimensional BM vector with independent components.*

The proof of Theorem 10 is omitted. This is because the proof is obtained along lines similar to the proof of Theorem 2, while being more cumbersome due to the extra spacial dimensions. Notice that the dimensionality of the driving Brownian motion depends on U, V. As in Section 3.1 we can restate Theorem 10 by making explicit the dependency on the driving sources of randomness. In particular, this allows us to formulate the dynamics of $\boldsymbol{x}_t$ as integration with respect to Brownian motions whose dimensionality does not depend on the number of inputs, nor their spatial sizes U, V. Under Assumption 8 and Assumption 5, the next corollary characterizes the limiting diffusion process of the convolutional RenNet at the initial value $\boldsymbol{x}_0^{(i)}$.

Corollary 11 *Let* $\boldsymbol{x}_{t,p}^{(i)}$ *be the ResNet corresponding to the initial value* $\boldsymbol{x}_0^{(i)}$*. Under Assumption 8 and Assumption 5, the limiting diffusion process* $x_{t,p}^{(i)}$ *of* $\boldsymbol{x}_{t,p}^{(i)}$ *is the solution on* $[0,T]$ *of the SDE*

$$dx_{t,p}^{(i)} = \phi'(0)(dW_t\psi(x_{t,\star p}^{(i)}) + db_t) + \frac{1}{2}\phi''(0)(d[W\psi(x_{\star p}^{(i)})]_t + d[b]_t) \tag{25}$$

for $p = 1,\ldots,P$*, and*

$$d[x_p^{(i)}, x_{p'}^{(j)}]_t = \phi'(0)^2(d[W\psi(x_{\star p}^{(i)}), W\psi(x_{\star p'}^{(j)})]_t + d[b,b]_t) \tag{26}$$

The parametrization of Assumption 8 is $\mathcal{O}(D^2F^2)$. Hereafter we introduce a more parsimonious parameterization which is based on tensor Gaussian distributions; this is a natural generalization of the matrix Gaussian distribution in Gupta and Nagar (1999). The use of Kronecker products allows us to cover this parametrization with a compact notation. We also introduce a fully i.i.d. initialization with the same scaling with D as in the fully-connected case.

Assumption 9 (tensor Gaussian network's parameters) *Let* $b_t, \mu^b, \sigma_b, B_t^b, \mu^W, B_t^W$ *be defined as in Assumption 8, and let* W_t *be a diffusion process in* $\mathbb{R}^{D\times(K\times K\times D)}$*, which is defined as the solution of*

$$dW_t = \mu^W dt + \sigma^{W_O} dB_t^W(\sigma^{W_U} \otimes \sigma^{W_V} \otimes \sigma^{W_I}),$$

where $\sigma^{W_O}, \sigma^{W_I} \in \mathbb{R}^{D\times D}$*,* $\sigma^{W_U}, \sigma^{W_V} \in \mathbb{R}^{K\times K}$ *and* $\Sigma^{W_O} = \sigma^{W_O}\sigma^{W_O\top}$*,* $\Sigma^{W_U} = \sigma^{W_U}\sigma^{W_U\top}$*,* $\Sigma^{W_V} = \sigma^{W_V}\sigma^{W_V\top}$*,* $\Sigma^{W_I} = \sigma^{W_I\top}\sigma^{W_I}$ *are positive semi-definite matrices*

We consider the setting of Assumption 9. Under this setting the discretization of W_t satisfies:

$$\varepsilon_t^W \overset{i.i.d.}{\sim} \mathcal{TN}_{D,K,K,D}\big(0, \Sigma^{W_O}, \Sigma^{W_U}, \Sigma^{W_V}, \Sigma^{W_I}\big)$$

for $t = \Delta t, \dots, T$, where $\mathcal{TN}$ stands for the tensor Gaussian distribution, and we have $\operatorname{cov}(\varepsilon^W_{o,u,v,i}, \varepsilon^W_{o',u',v',i'}) = \Sigma^{W_O}_{o,o'}\Sigma^{W_U}_{u,u'}\Sigma^{W_V}_{v,v'}\Sigma^{W_I}_{i,i'}$. Under Assumption 9 and Assumption 5, the next corollary characterizes the limiting diffusion process of the convolutional ResNet at the initial value $\boldsymbol{x}_0^{(i)}$. In particular, the next corollary provides the convolutional counterpart of Corollary 4.

Corollary 12 *Let $\boldsymbol{x}_{t,p}^{(i)}$ be the ResNet corresponding to the initial value $\boldsymbol{x}_0^{(i)}$. Under Assumption 9 and Assumption 5, the limiting diffusion process $x_{t,p}^{(i)}$ of $\boldsymbol{x}_{t,p}^{(i)}$ is the solution on $[0, T]$ of the SDE*

$$\begin{aligned} dx_{t,p}^{(i)} = \phi'(0)\big((&\mu^W \psi(x_{t,\star p}^{(i)}) + \mu^b)dt \\ &+ \sigma^{W_O} dB_t^W (\sigma^{W_U} \otimes \sigma^{W_V} \otimes \sigma^{W_I})\psi(x_{t,\star p}^{(i)}) + \sigma^b dB_t^b\big) \\ &+ \frac{1}{2}\phi''(0) \operatorname{diag}\big(\Sigma^b + \Sigma^{W_O}(\psi(x_{t,\star p}^{(i)})^\top (\Sigma^{W_U} \otimes \Sigma^{W_V} \otimes \Sigma^{W_I})\psi(x_{t,\star p}^{(i)}))\big)dt \end{aligned} \tag{27}$$

and

$$d[x_p^{(i)}, x_{p'}^{(j)}]_t = \phi'(0)^2\big(\Sigma^b + \Sigma^{W_O}\psi(x_{t,\star p}^{(i)})^\top (\Sigma^{W_U} \otimes \Sigma^{W_V} \otimes \Sigma^{W_I})\psi(x_{t,\star p'}^{(j)})\big)dt.$$

Assumption 10 (fully i.i.d. network's parameters) *Let W_t be a diffusion process in $\mathbb{R}^{D\times K\times K\times D}$ and let b_t be the diffusion process in $\mathbb{R}^D$, which are defined as solutions of the SDEs*

$$dW_t = \frac{\sigma_w}{\sqrt{D}} dB_t^W \tag{28}$$

and

$$db_t = \sigma_b dB_t^b, \tag{29}$$

respectively, where $B_t^W \in \mathbb{R}^{D\times D}$ and $B_t^b \in \mathbb{R}^D$ are independent BMs and $\sigma_w > 0, \sigma_b > 0$ are scalars.

Now, we consider the setting of Assumption 10. In particular, the discretizations of the diffusion processes W_t and b_t displayed in (28) and (29), respectively, admit exact representations as

$$\Delta W_t = \zeta_t^W \frac{\sigma_w}{\sqrt{D}}\sqrt{\Delta t}$$

and

$$\Delta b_t = \zeta_t^b \sigma_b \sqrt{\Delta t},$$

where $\zeta_t^W \overset{i.i.d.}{\sim} \mathcal{TN}_{D,K,K,D}\big(0, \mathrm{I}_D, \mathrm{I}_K, \mathrm{I}_K, \mathrm{I}_D\big)$ and $\zeta_t^b \overset{i.i.d.}{\sim} \mathcal{N}_D\big(0, \mathrm{I}_D\big)$. As for the setting of fully-connected neural networks, fully i.i.d. initializations are commonly used in the context of the training of convolutional neural networks. Under Assumption 10 and Assumption 5, the next corollary characterizes the limiting diffusion process of the convolutional ResNet at the initial value $\boldsymbol{x}_0^{(i)}$. In particular, the next corollary provides the convolutional counterpart of Corollary 5.

Corollary 13 *Let $\boldsymbol{x}_{t,p}^{(i)}$ be the ResNet corresponding to the initial value $\boldsymbol{x}_0^{(i)}$. Under Assumption 10 and Assumption 5, the limiting diffusion process $x_{t,p}^{(i)}$ of $\boldsymbol{x}_{t,p}^{(i)}$ is the solution on $[0,T]$ of the SDE*

$$\begin{aligned} dx_{t,p}^{(i)} &= \phi'(0)\big(\frac{\sigma_w}{\sqrt{D}} dB_t^W \psi(x_{t,\star p}^{(i)}) + \sigma_b dB_t^b\big) \\ &\quad + \frac{1}{2}\phi''(0)\big(\sigma_b^2 + \frac{\sigma_w^2}{D}\|\psi(x_{t,\star p}^{(i)})\|^2\big)\big)\,\mathrm{I}_D\,dt \end{aligned} \tag{30}$$

and

$$d[x_p^{(i)}, x_{p'}^{(j)}]_t = \phi'(0)^2\big(\sigma_b^2 + \frac{\sigma_w^2}{D}\langle\psi(x_{t,\star p}^{(i)}), \psi(x_{t,\star p'}^{(j)})\rangle\big)\,\mathrm{I}_D\,dt$$

In view of the results obtained in this section, all the remarks of Section 3.1 have a corresponding remark that applies to infinitely deep convolutional ResNets. Namely, the main qualitative conclusions continue to hold. That is, the stochastic process limit is well-behaved and perfect-correlation problems are avoided, explosive solutions are possible whenever $\phi''(0) \neq 0$.

4. Infinitely deep ResNets' gradient

We consider the problem of trainability at initialization of very deep ResNets which are finitely wide. In a generic setting, gradient descent iterations with a fixed learning rate η are of the form

$$\theta(b+1) = \theta(b) - \eta\nabla\mathcal{R}(\theta(b)))$$

for $b = 0, 1, \dots,$ where $\theta(b) \in \mathbb{R}^\Theta$ is the generic iteration of network's parameters of interest, and $\mathcal{R}(\theta)$ is a smooth real-valued loss function to be minimized. Differently from the gradient descent, the SGD relies on unbiased estimates of the gradient of the loss function of interest. In particular, for $\mathbb{E}[\nabla\mathcal{R}_b(\theta)] = \nabla\mathcal{R}(\theta)$, SGD iterations with a fixed learning rate η are of the form

$$\theta(b+1) = \theta(b) - \eta\nabla\mathcal{R}_b(\theta(b)).$$

Both $\mathcal{R}(\theta)$ and $\mathcal{R}_b(\theta)$ are obtained by summing or averaging terms of the form $R(\widehat{y}(z), y)$ with $R : \mathbb{R}^Y \times \mathbb{R}^Y \to \mathbb{R}$ being the loss function for 1 data-point (z, y) and $\widehat{y}(z)$ being the prediction of the neural network for z. For the rest of this section we consider a single data point and smooth R. A key difficulty in training very deep neural networks is that the gradients with respect to lower layers, i.e. small t for large L in our setting, might vanish or explode. This phenomenon results in negligible or diverging network's parameter updates and ultimately in bad training performance. This intuition can be made rigorous by linking the norm of the gradients, or their expectations, to loss function decrements (Bottou et al., 2018).

For each t, let θ_t denote the weight, or the bias, at layer t, either in the "standard" form $(\Delta W_t, \Delta b_t)$ or in the "reparametrized" form $(\varepsilon_t^W, \varepsilon_t^b)$. Then, we can write the following equations

$$(\nabla_{\theta_{t-\Delta t}} R)^\top = J(R, \boldsymbol{x}_T) J(\boldsymbol{x}_T, \boldsymbol{x}_t) J(\boldsymbol{x}_t, \theta_{t-\Delta t})$$

and

$$J(\boldsymbol{x}_T, \boldsymbol{x}_t) = J(\boldsymbol{x}_T, \boldsymbol{x}_{T-\Delta t}) J(\boldsymbol{x}_{T-\Delta t}, \boldsymbol{x}_{T-2\Delta t}) \cdots J(\boldsymbol{x}_{t+\Delta t}, \boldsymbol{x}_t).$$

The problematic term is represented by the Jacobian matrix $J(\boldsymbol{x}_T, \boldsymbol{x}_t)$. Indeed the matrix $J(\boldsymbol{x}_T, \boldsymbol{x}_t)$ involves a large (infinite in the limit $L \uparrow \infty$) number of matrix multiplications for the lower layers of a neural network, where $t \approx 0$. Note that $J(\boldsymbol{x}_T, \boldsymbol{x}_t)$ is closely related to $J(\boldsymbol{x}_t, \boldsymbol{x}_0)$ as, provided that $J(\boldsymbol{x}_t, \boldsymbol{x}_0)$ is invertible, $J(\boldsymbol{x}_T, \boldsymbol{x}_t)$ can be obtained as $J(\boldsymbol{x}_T, \boldsymbol{x}_t) = J(\boldsymbol{x}_T, \boldsymbol{x}_0)J(\boldsymbol{x}_t, \boldsymbol{x}_0)^{-1}$. In any case, the properties of $J(\boldsymbol{x}_t, \boldsymbol{x}_0)$ are most closely related to the problem of a vanishing/exploding gradient. Hereafter, we show that for infinitely deep ResNets, under suitable assumptions on the activation functions ϕ and ψ, the problem of an exploding gradient is avoided. Moreover, we show that under the same assumptions on ϕ and ψ the limiting process is invertible. Construction of invertible neural networks is the main focus of recent research (Behrmann et al., 2019), and the invertibility of ResNets has been empirically shown to be related to model robustness (Engstrom et al., 2019).

Let $\boldsymbol{x}_t$ follow the ResNet (11) with the activation functions ψ and ϕ satisfying Assumption 5. Let $\boldsymbol{g}_t = J(\boldsymbol{x}_t, \boldsymbol{x}_0)$, hence $\boldsymbol{g}_{t+\Delta t} = J(\boldsymbol{x}_{t+\Delta t}, \boldsymbol{x}_t)\boldsymbol{g}_t$, and by direct computation we can write

$$\Delta \boldsymbol{g}_t = \big(\phi'(\Delta W_t \psi(\boldsymbol{x}_t) + \Delta b_t) 1_D{}^\top \odot \Delta W_t \odot 1_D \psi'(\boldsymbol{x}_t)^\top\big)\boldsymbol{g}_t.$$

Now, we show that the Jacobian matrix $J(\boldsymbol{x}_t, \boldsymbol{x}_0)$ is well behaved in the sense that it converges to the solution $J(x_t, x_0)$ of a matrix SDE as $L \uparrow \infty$. As in the case of $\boldsymbol{x}_t$, we can derive a limiting SDE to which $\boldsymbol{g}_t$ converges, as $L \uparrow \infty$, by establishing the convergence of the corresponding instantaneous mean and covariance of $\boldsymbol{g}_t$. We denote this limiting SDE with $g_t = J(x_t, x_0)$. Subsequently we can link g_t with W_t and b_t by showing the equivalence in law between g_t and the solution to another SDE. The next theorem states directly the final result.

Theorem 14 *Let $\boldsymbol{g}_t = J(\boldsymbol{x}_t, \boldsymbol{x}_0)$ and let $\overline{\boldsymbol{g}}_t$ denote the continuous-time interpolation of $\boldsymbol{g}_t$. Under Assumption 4 and Assumption 5, $\overline{\boldsymbol{g}}_t$ converges in distribution to the solution of the matrix SDE*

$$dg_t = \Big(\Big(\phi'(0) dW_t + \phi''(0) d[W\psi(x)1_D{}^\top \odot W]_t\Big) \odot 1_D \psi'(x_t)^\top\Big)\, g_t \tag{31}$$

We require the matrix SDE (31) to be non-explosive. This is for instance the case when $\phi''(0)$ and ψ is the identity function. See also Remark 8 in Section 3.1. As long as (31) is not explosive, $J(x_t, x_0)$ is invertible and we can find the SDE determining the evolution of its inverse.

Corollary 15 *Let g_t be the matrix-valued diffusion process satisfying the matrix SDE (31). Under Assumption 4 and Assumption 5, g_t is invertible and its inverse satisfies the matrix SDE*

$$dg_t^{-1} = g_t^{-1}\big(-\Big(\phi'(0)dW_t + \phi''(0)d[W\psi(x)1_D{}^\top \odot W]_t\Big) \odot 1_D\psi'(x_t)^\top + \phi'(0)^2 d[W \odot 1_D \psi'(x)^\top]_t\big). \tag{32}$$

Hence, $J(x_T, x_t)$ can be obtained as $J(x_T, x_t) = g_T g_t^{-1}$ by means of integrating (31) and (32), which are driven by the same process W. Theorem 14 and Corollary 15 have two fundamental consequences: i) as g_t is the Jacobian of the last layer with respect of the first layer of the limiting process x_t, it follows that the exploding gradient problem is avoided as long as (31) is not explosive; ii) by the inverse function theorem the limiting process x_t is invertible. Note that the results of this section hold for all the parametrizations discussed in Section 3.1.

5. Doubly infinite ResNets

We study the more general setting of infinitely deep and infinitely wide ResNets, that is ResNets where both the depth L and the dimension D grow unboundedly. In particular, it is assumed that first $L \uparrow \infty$, and then $D \uparrow \infty$. Most of the analysis that follows assumes that ψ is the identity function. Although we only provide heuristics for the convergence results of Proposition 18 and Proposition 19, numerical experiments reported in Section 6 support their correctness. Moreover, the numerical experiments of Section 6 support the conjecture that analogous results hold when D and L grow unbounded jointly when ψ is the identity function, and ϕ is suitably smooth as assumed thorough this work. A more detailed discussion on interchanging the width and depth limits under these assumptions is contained in Section 5.6. Hereafter, we consider the setting of Corollary 5, i.e. fully i.i.d. network's parameters. In this case, we can rewrite (20) as

$$dx_t^{(i)} = \phi'(0)(\frac{\sigma_w}{\sqrt{D}} dB_t^W \psi(x_t^{(i)}) + \sigma_b dB_t^b) + \frac{1}{2}\phi''(0)(\sigma_b^2 + \sigma_w^2 q_t^{(i)}) 1_D dt \qquad (33)$$

and

$$d[x^{(i)}, x^{(j)}]_t = \phi'(0)^2(\sigma_b^2 + \sigma_w^2 \lambda_t^{(i,j)}) \, \mathrm{I}_D \, dt$$

with $\lambda_t^{(i,j)} = \langle \psi(x_t^{(i)}), \psi(x_t^{(j)}) \rangle / D$ and $q_t^{(i)} = \lambda_t^{(i,i)} = \|\psi(x_t^{(i)})\|^2/D$. As a starting point, we need to ensure the well-posedness of (33) for small $t > 0$ as $D \uparrow \infty$. Therefore, we assume that the following limits exist and are finite: $q_0^{(i),\infty} = \lim_{D\uparrow\infty} q_0^{(i)}$, and $\lambda_0^{(i,j),\infty} = \lim_{D\uparrow\infty} \lambda_0^{(i,j)}$. Note that the notation does not convey explicitly the dependence of x_t, and hence of q_t, λ_t on D.

5.1 Weakly and fully trained ResNets

The connection between Gaussian processes and infinitely wide neural networks is well-known (Neal, 1995; Lee et al., 2018; Garriga-Alonso et al., 2019). In Section 5.3 we show that similar results hold true for infinitely deep and infinitely wide ResNets, thus obtaining convergence to a Gaussian process. For infinitely wide neural networks, the NTK of Jacot et al. (2018); Arora et al. (2019); Lee et al. (2019b) allows for computing the solution obtained by fully training a neural network according to continuous-time, i.e. infinitesimal learning rate, gradient descent under the assumption of a quadratic loss. More in detail, let $\widehat{y}_\theta^{(i)} \in \mathbb{R}$ be the output of a neural network with parameters $\theta \in \mathbb{R}^\Theta$ for its i-th input, $i = 1, \ldots, N$. Let $\mathcal{R}(\theta) = \frac{1}{2}\sum_{i=1}^N (\widehat{y}_\theta^{(i)} - y^{(i)})^2$ be the squared loss over training data, where $y^{(i)}$ denotes the i-th target. We report (Arora et al., 2019, Proposition 3.1) according to our notation:

Proposition 16 *Consider minimizing the squared loss $\mathcal{R}(\theta)$ by gradient descent with infinitesimally small learning rate: $\frac{d\theta(t)}{dt} = -\nabla_\theta \mathcal{R}(\theta(t))$. Let $\widehat{y}(t) = \{\widehat{y}_{\theta(t)}^{(i)}\}_{i=1}^N \in \mathbb{R}^N$ be the network outputs on all inputs at time t, and $y = \{y^{(i)}\}_{i=1}^N$ be the desired outputs. Then $\widehat{y}(t)$*

follows the following evolution, where $\mathcal{K}(t)$ *is an* $N \times N$ *positive semidefinite matrix whose* (i,j)*-th entry is* $\mathcal{K}(t)^{(i,j)} = \langle \nabla_\theta \widehat{y}^{(i)}_{\theta(t)}, \nabla_\theta \widehat{y}^{(j)}_{\theta(t)} \rangle$*:*

$$\frac{d\widehat{y}(t)}{dt} = -\mathcal{K}(t)(\widehat{y}(t) - y). \tag{34}$$

According to Equation (34), the empirical NTK corresponding to inputs i, j is then defined as follows:

$$\mathcal{K}^{(i,j)} = \langle \nabla_\theta \widehat{y}^{(i)}_{\theta(0)}, \nabla_\theta \widehat{y}^{(j)}_{\theta(0)} \rangle, \tag{35}$$

i.e. it is $\mathcal{K}(t)$ for $t = 0$. As the width of a neural network goes to infinity, $\mathcal{K}^{(i,j)}$ converges to a deterministic limit $\mathcal{K}^{(i,j,\infty)}$ for each pair of points under the considered assumptions. Moreover, it is possible to bound the fluctuations of $\mathcal{K}(t)$ around its initial value $\mathcal{K}(0)$ as the width increases. Building on these results, it is possible to establish the equivalence between the solution obtained by fully training a neural network via continuous-time gradient descent and kernel regression via the $\mathcal{K}^{(i,j,\infty)}$ kernel. Indeed, within the aforementioned setting, the time-independent $\mathcal{K}^{(i,j,\infty)}$ can be substituted for $\mathcal{K}(t)^{(i,j)}$ in (34) yielding the constant coefficients ODE

$$\frac{d\widehat{y}(t)}{dt} = -\mathcal{K}^{(\infty)}(\widehat{y}(t) - y). \tag{36}$$

The stationary solution of the ODE (36), with initial condition $\widehat{y}(0) = 0$, is given by the standard kernel regression formula associated to the kernel $\mathcal{K}^{(\infty)}$. In particular, the desired initial condition can be enforced either by scaling the neural network output by an appropriately small constant (Arora et al., 2019) or by subtracting the output of an independent and untrained copy of the considered neural network at initialization (Lee et al., 2019b).

We show in Section 5.4 that similar results hold for the case of infinitely deep and infinitely wide ResNets: (35) at initialization converges to a deterministic limit. Also, it is known (Arora et al., 2019) that in the aforementioned setting, training only the last output of a neural network under the same conditions corresponds to performing Bayesian inference under the Gaussian process prior arising in the infinite wide limit. Hence, we talk equivalently of Bayesian inference and weak training, and we refer to the standard NTK setting as full training. All the results of Section 5.3 and Section 5.4 concern with a ResNet $\boldsymbol{x}_{0:T}$ with input $\boldsymbol{x}_0$. As previously mentioned, it is necessary to complete the ResNet with an input layer adapting the infinitely wide ResNet to finite-dimensional inputs. Moreover, to more closely resemble neural networks used in practice, an output adaptation layer is commonly introduced as well. In Section 5.5 we study the implications to the training of completing the ResNet with input and output layers. Moreover, in line with the NTK literature, it is necessary to consider an appropriate parametrization in order to obtained the desired NTK convergence results. The reparametrized gradients used by gradient descent are computed with respect to network's parameters which are i.i.d. distributed as standard Gaussian distributions and any scaling is expressed via multiplication, not via the Gaussian distribution's variance. In our context this corresponds to gradients with respect to $(\varepsilon^W_t, \varepsilon^b_t)$ in (18) and (19).

5.2 Weakly trained ResNets - general case

Observe that the evolution of (33) is directly governed by $q_t^{(i)}$ and $\lambda_t^{(i,j)}$. The following results forms the basis of our analysis. Its proof is obtained by a straightforward but tedious application of multi-dimensional Ito's formula (Øksendal, 2003, Section 4.2) and is thus omitted.

Proposition 17 *Let $x_t = x_t^{(i)}$ evolve according to the SDE (33), and let $f(x) = \langle g(x), h(x)\rangle/D$ with g and h being twice continuously differentiable scalar functions applied element-wise. Then*

$$\begin{aligned} df(x_t) = \phi'(0)\frac{1}{D}(g'(x_t)h(x_t) + g(x_t)h'(x_t))^\top(\frac{\sigma_w}{\sqrt{D}}dB_t^W\psi(x_t) + \sigma_b dB_t^b) \\ + \frac{1}{2}(\sigma_b^2 + \sigma_w^2 q_t^{(i)})\phi'(0)^2(\langle g''(x_t), h(x_t)\rangle/D + \langle g(x_t), h''(x_t)\rangle/D + 2\langle g'(x_t), h'(x_t)\rangle/D)dt \\ + \frac{1}{2}(\sigma_b^2 + \sigma_w^2 q_t^{(i)})\phi''(0)(\langle g'(x_t), h(x_t)\rangle/D + \langle g(x_t), h'(x_t)\rangle/D)dt \end{aligned}$$

Moreover, Let $x_t = x_t^{(i)}$ and $y_t = x_t^{(j)}$ evolve according to (33), and let $F(x, y) = \langle G(x), H(y)\rangle/D$ with G and H being twice continuously differentiable scalar functions applied element-wise. Then

$$\begin{aligned} dF(x_t, y_t) = \phi'(0)\frac{1}{D}(G'(x_t)H(y_t))^\top(\frac{\sigma_w}{\sqrt{D}}dB_t^W\psi(x_t) + \sigma_b dB_t^b) \\ + \phi'(0)\frac{1}{D}(G(x_t)H'(y_t))^\top(\frac{\sigma_w}{\sqrt{D}}dB_t^W\psi(y_t) + \sigma_b dB_t^b) \\ + \frac{1}{2D}(\sigma_b^2 + \sigma_w^2 q_t^{(i)})\phi''(0)\langle G'(x_t), H(y_t)\rangle dt \\ + \frac{1}{2D}(\sigma_b^2 + \sigma_w^2 q_t^{(j)})\phi''(0)\langle G(x_t), H'(y_t)\rangle dt \\ + \frac{1}{2D}(\sigma_b^2 + \sigma_w^2 q_t^{(i)})\langle G''(x_t), H(y_t)\rangle dt \\ + \frac{1}{2D}(\sigma_b^2 + \sigma_w^2 q_t^{(j)})\langle G(x_t), H''(y_t)\rangle dt \\ + \frac{1}{D}(\sigma_b^2 + \sigma_w^2 \lambda_t^{(i,j)})\langle G'(x_t), H'(y_t)\rangle dt \end{aligned}$$

The evolution of $q_t^{(i)}$ and $\lambda_t^{(i,j)}$ is obtained by an application of Proposition 17 with $h := g := H := G := \psi$, resulting in $dq_t^{(i)} = df(x_t^{(i)})$ and $d\lambda_t^{(i,j)} = dF(x_t^{(i)}, x_t^{(j)})$. Inspecting the corresponding dynamics highlights the difficulties in obtaining closed-form solutions for the evolution of $q_t^{(i)}$ and $\lambda_t^{(i,j)}$. The drift and diffusion coefficients describing the evolution of $q_t^{(i)}$ involves inner-products of the form $\langle\psi^{(m)}(x_t^{(i)}), \psi^{(n)}(x_t^{(i)})\rangle$ where $m, n = 0, 1, 2$ denote the degrees of differentiation. By relying on Proposition 17 to determine the evolution of $\langle\psi^{(m)}(x_t^{(i)}), \psi^{(n)}(x_t^{(i)})\rangle$ results in drift and diffusion coefficients depending on other terms of the form $\langle\psi^{(m)}(x_t^{(i)}), \psi^{(n)}(x_t^{(i)})\rangle$ involving higher degrees of differentiation. Similar considerations hold true for $\lambda_t^{(i,j)}$. Ultimately, determining the evolution of $q_t^{(i)}$ and $\lambda_t^{(i,j)}$ separately from $x_t^{(i)}$ requires solving an infinite recursion of SDEs. Similarly, solving (33) directly requires solving an infinite dimensional SDE as $D \uparrow \infty$. In either cases it doesn't seem possible to solve (33) exactly in the infinite-width limit with a finite amount of computational effort. In the special case where ψ is the identity function the dynamics of $q_t^{(i)}$ and $\lambda_t^{(i,j)}$ simplify considerably. We explore the implications of this modeling assumption in the remaining sections of this work.

5.3 Weakly trained ResNets - identity ψ case

Hereafter, we assume the activation function ψ to be the identity function, and hence only the activation function ϕ affects the neural network. In Lemma 21 in Appendix A we derive the evolution of $q_t^{(i)}$ and $\lambda_t^{(i,j)}$ in this specific setting. Now, we show that as D increases $q_t^{(i)}$ and $\lambda_t^{(i,j)}$ converge to deterministic limits which are obtained as solutions of a finite-dimensional ODE system.

Proposition 18 (Heuristic) *As $D \uparrow \infty$, the quantities $m_t^{(i)}$, $q_t^{(i)}$, $\lambda_t^{(i,j)}$ converge to the solutions of the ODEs*

$$dm_t^{(i),\infty} = \frac{1}{2}\phi''(0)\big(\sigma_b^2 + \sigma_w^2 q_t^{(i),\infty}\big)dt,$$

$$dq_t^{(i),\infty} = \big(\phi''(0)m_t^{(i),\infty} + \phi'(0)^2\big)\big(\sigma_b^2 + \sigma_w^2 q_t^{(i),\infty}\big)dt$$

and

$$d\lambda_t^{(i,j),\infty} = \big(\frac{1}{2}\phi''(0)((\sigma_b^2 + \sigma_w^2 q_t^{(i),\infty})m_t^{j,\infty} + (\sigma_b^2 + \sigma_w^2 q_t^{(j),\infty})m_t^{(i),\infty}) \\ + \phi'(0)^2\big(\sigma_b^2 + \sigma_w^2 \lambda_t^{(i,j),\infty}\big)\big)dt,$$

respectively. Moreover, under the assumption $\phi''(0) = 0$, the solutions for $m_T^{(i),\infty}$, $q_T^{(i),\infty}$ and $\lambda_T^{(i,j),\infty}$ are

$$m_T^{(i),\infty} = m_0^{(i),\infty},$$

$$q_T^{(i),\infty} = q_0^{(i),\infty} + \left(q_0^{(i),\infty} + \frac{\sigma_b^2}{\sigma_w^2}\right)\left(e^{\phi'(0)^2\sigma_w^2 T} - 1\right).$$

and

$$\lambda_T^{(i,j),\infty} = \lambda_0^{(i,j),\infty} + \left(\lambda_0^{(i,j),\infty} + \frac{\sigma_b^2}{\sigma_w^2}\right)\left(e^{\phi'(0)^2\sigma_w^2 T} - 1\right),$$

respectively. Moreover, under the assumption $\phi''(0) \neq 0$, the solutions for $m_T^{(i),\infty}$ and $q_T^{(i),\infty}$ are

$$m_T^{(i),\infty} = \frac{1}{\phi''(0)}\left\{-\phi'(0)^2 + \frac{1}{\sigma_w}\sqrt{C}\tan\left(\frac{1}{2}\sigma_w\sqrt{C}(T + 2c_2)\right)\right\} \tag{37}$$

and

$$q_T^{(i),\infty} = \frac{1}{\phi''(0)^2\sigma_w^2}\left\{-\phi''(0)^2\sigma_b^2 + C\sec\left(\frac{1}{2}\sigma_w\sqrt{C}(T + 2c_2)\right)^2\right\}, \tag{38}$$

respectively, where c_1 and c_2 are two constants that depend on the initial conditions of Equations (37) and (38), and $C = -\phi'(0)^4\sigma_w^2 + \phi''(0)^2(\sigma_b^2 + \sigma_w^2 c_1)$.

In Proposition 18, $m_t^{(i)} = D^{-1}\sum_{1\leq d\leq D} x_{t,d}^{(i)}$ and we further require that $m_0^{(i),\infty} = \lim_{D\uparrow\infty} m_0^{(i)}$ exists and it is finite. Crucially, when ψ is the identity function, it is not possible to decouple the evolution of $q_t^{(i)}$ and $\lambda_t^{(i,j)}$ from the evolution of the driving stochastic process x_t when $D \uparrow \infty$. The drift, diffusion, and correlation coefficients driving the SDE (33) converge to deterministic limit too which results in i.i.d. stochastic processes across the dimensions of the neural network.

Proposition 19 (Heuristic) *As $D \uparrow \infty$ each $x_t^{(i)}$ converges to $x_t^{(i),\infty}$, with the $x_t^{(i),\infty}$'s being i.i.d. across the dimensions of the neural network. Moreover, for $x_t^{(i),\infty} = x_{t,1}^{(i),\infty}, x_{t,2}^{(i),\infty}, \dots,$ and $d, u \geq 1$ it holds*

$$dx_{t,d}^{(i),\infty} = \phi'(0)(\sigma_b^2 + \sigma_w^2 q_t^{(i),\infty})^{1/2} dB_{t,d}^{(i),\infty} \tag{39}$$

and

$$d[B_d^{(i),\infty}, B_u^{(j),\infty}]_t = \frac{\sigma_b^2 + \sigma_w^2 \lambda_t^{(i,j),\infty}}{\left((\sigma_b^2 + \sigma_w^2 q_t^{(i),\infty})(\sigma_b^2 + \sigma_w^2 q_t^{(j),\infty})\right)^{1/2}} \delta_{d,u} dt,$$

where $B_{t,1}^{(i),\infty}, B_{t,2}^{(i),\infty}, \dots$ are scalar BMs dependent over i and $q_t^{(i),\infty}, \lambda_t^{(i,j),\infty}$ are obtained by solving the ODEs in Proposition 18. Over the two data-points indexed by i, j this is a 2-dimensional SDE with time-dependent and deterministic drift and diffusion coefficients, and such that

$$p(x_{T,d}^{(i),\infty}, x_{T,d}^{(j),\infty} \,|\, x_{0,d}^{(i)}, x_{0,d}^{(j)}) = \mathcal{N}_2\left(\begin{bmatrix} x_{0,d}^{(i)} + m_T^{(i),\infty} - m_0^{(i),\infty} \\ x_{0,d}^{(j)} + m_T^{(j),\infty} - m_0^{(j),\infty} \end{bmatrix}, \right. \\ \left. \begin{bmatrix} v_T^{(i),\infty} - v_0^{(i),\infty} & c_T^{(i,j),\infty} - c_0^{(i,j),\infty} \\ c_T^{(i,j),\infty} - c_0^{(i,j),\infty} & v_T^{(j),\infty} - v_0^{(j),\infty} \end{bmatrix} \right), \tag{40}$$

where $v_t^{(i),\infty} = q_t^{(i),\infty} - (m_t^{(i),\infty})^2$, $c_t^{(i,j),\infty} = \lambda_t^{(i,j),\infty} - m_t^{(i),\infty} m_t^{(j),\infty}$.

According to Proposition 19, doubly infinite ResNets are non-centered Gaussian processes with covariance kernel $K^{(i,j),\infty} = c_T^{(i,j),\infty} - c_0^{(i,j),\infty}$ and mean function $M_d^{(i),\infty} = x_{0,d}^{(i)} + m_T^{(i),\infty} - m_0^{(i),\infty}$. That is: i) when $\phi''(0) = 0$, we have $M_d^{(i),\infty} = x_{0,d}^{(i)}$ and $K^{(i,j),\infty} = (\lambda_0^{(i,j),\infty} + \sigma_b^2/\sigma_w^2)(e^{\phi'(0)^2\sigma_w^2 T} - 1)$; ii) when $\phi''(0) \neq 0$, from Equation (37) and Equation (38) we obtain the deterministic explosion time of $x_{t,d}^{(i),\infty}$ by solving $2^{-1}\sigma_w\sqrt{C}(T + 2c_2) = \pi/2$ in T; in particular, the constants c_1, c_2 depend on $m_0^{(i),\infty}$, $q_0^{(i),\infty}$ and have to be determined numerically.

5.4 Fully trained ResNets

Let $\theta = \{\varepsilon_t^W, \varepsilon_t^b\}_{t=0}^{T-1}$ denote the "reparametrized" collection of network's parameters with respect to which we compute the NTK. We establish the convergence of $\mathcal{K}^{(i,j)}$ to a deterministic limit as $L \uparrow \infty$ and then $D \uparrow \infty$. We operate under the following assumptions: i) $\boldsymbol{x}_t$ follows (20) with ψ being the identity function; ii) network's parameters follow Assumption 7; iii)

$\phi''(0) = 0$; iv) $\widehat{y} = \boldsymbol{x}_{T,1}$. We have $\mathcal{K}^{(i,j)} = \mathcal{K}_W^{(i,j)} + \mathcal{K}_b^{(i,j)}$, where $\mathcal{K}_W^{(i,j)} = \sum_{t=\Delta t}^T \mathcal{K}_{W,t}^{(i,j)}$, $\mathcal{K}_b^{(i,j)} = \sum_{t=\Delta t}^T \mathcal{K}_{b,t}^{(i,j)}$, and

$$\begin{aligned}\mathcal{K}_{W,t}^{(i,j)} &= J(\widehat{y}^{(i)}, \boldsymbol{x}_T^{(i)}) J(\boldsymbol{x}_T^{(i)}, \boldsymbol{x}_t^{(i)}) J(\boldsymbol{x}_t^{(i)}, \Delta W_{t-\Delta t}) \\ &\quad \times \left(J(\widehat{y}^{(j)}, \boldsymbol{x}_T^{(j)}) J(\boldsymbol{x}_T^{(j)}, \boldsymbol{x}_t^{(j)}) J(\boldsymbol{x}_t^{(j)}, \Delta W_{t-\Delta t})\right)^\top \sigma_w^2 \Delta t / D\end{aligned}$$

and

$$\begin{aligned}\mathcal{K}_{b,t}^{(i,j)} &= J(\widehat{y}^{(i)}, \boldsymbol{x}_T^{(i)}) J(\boldsymbol{x}_T^{(i)}, \boldsymbol{x}_t^{(i)}) J(\boldsymbol{x}_t^{(i)}, \Delta b_{t-\Delta t}) \\ &\quad \times \left(J(\widehat{y}^{(j)}, \boldsymbol{x}_T^{(j)}) J(\boldsymbol{x}_T^{(j)}, \boldsymbol{x}_t^{(j)}) J(\boldsymbol{x}_t^{(j)}, \Delta b_{t-\Delta t})\right)^\top \sigma_b^2 \Delta t,\end{aligned}$$

as

$$J(\boldsymbol{x}_t, \zeta_{t-\Delta t}^W) = J(\boldsymbol{x}_t, \Delta W_{t-\Delta t}) \sigma_w \sqrt{\Delta t} / \sqrt{D}$$

and

$$J(\boldsymbol{x}_t, \zeta_{t-\Delta t}^b) = J(\boldsymbol{x}_t, \Delta b_{t-\Delta t}) \sigma_b \sqrt{\Delta t}.$$

Recall from our study in Section 4 that, as $L \to \infty$, we have that $J(\boldsymbol{x}_T^{(i)}, \boldsymbol{x}_t^{(i)}) \to g_T g_t^{-1}$ and $J(\boldsymbol{x}_T^{(j)}, \boldsymbol{x}_t^{(j)}) \to g_T g_t^{-1}$, as the evolution of g_t does not depend on x_t when $\phi''(0) = 0$. Furthermore, note that $J(\boldsymbol{x}_t, \Delta W_{t-\Delta t})_{d,i,j} \to \phi'(0) \delta_{d,i} x_{t,j}$ and $J(\boldsymbol{x}_t, \Delta b_{t-\Delta t})_{d,i} \to \phi'(0) \delta_{d,i}$. By combining these results, and by assuming that the interchange of limits is justified, we write

$$\mathcal{K}_W^{(i,j)} \to \phi'(0)^2 \sigma_w^2 J(\widehat{y}^{(i)}, x_T^{(i)}) g_T \left[\int_0^T \frac{\langle x_t^{(i)}, x_t^{(j)} \rangle}{D} g_t^{-1} g_t^{-1\top} dt\right] g_T^\top J(\widehat{y}^{(j)}, x_T^{(j)})^\top$$

and

$$\mathcal{K}_b^{(i,j)} \to \phi'(0)^2 \sigma_b^2 J(\widehat{y}^{(i)}, x_T^{(i)}) g_T \left[\int_0^T g_t^{-1} g_t^{-1\top} dt\right] g_T^\top J(\widehat{y}^{(j)}, x_T^{(j)})^\top.$$

Now, $g_t^{-1} g_t^{-1\top} = (g_t^\top g_t)^{-1}$. Accordingly, by an application of Ito's formula for matrix SDE products (Protter, 2005, Chapter V, Theorem 47), for $U_t = g_t^\top g_t$ we obtain the following SDE

$$dU_t = \phi'(0) \frac{\sigma_w}{\sqrt{D}} g_t^\top \left(dB_t^W + dB_t^{W\top}\right) g_t + \phi'(0)^2 \sigma_w^2 U_t dt,$$

where $U_0 = \mathrm{I}_D$, and whose quadratic variation (a matrix, in this particular case) is of the following form

$$d[U]_t = \phi'(0)^2 \frac{\sigma_w^2}{D} \left(g_t{}^\top \odot g_t{}^\top\right)\left(g_t \odot g_t\right) dt,$$

vanishing as $D \to \infty$. Therefore, $U_t \to U_t^\infty$ where $dU_t^\infty = \phi'(0)^2 \sigma_w^2 U_t^\infty dt$. Thus, as $D \to \infty$ the term $g_t^\top g_t$ is an infinite dimensional diagonal matrix with constant element u_t^∞ computable by solving the ODE $du_t^\infty = \phi'(0)^2 \sigma_w^2 u_t^\infty dt$ with initial value $u_0^\infty = 1$, i.e. $u_t^\infty = \exp(\phi'(0)^2 \sigma_w^2 t)$.

Note that: i) the matrix $\int_0^T (g_t^\top g_t)^{-1} dt$ is asymptotically diagonal with constant element $(1 - \exp(\phi'(0)^2 \sigma_w^2 T)) / (\phi'(0)^2 \sigma_w^2)$; ii) the matrix $g_T g_T^\top$ is asymptotically diagonal with constant element $\exp(\phi'(0)^2\sigma_w^2 T)/(\phi'(0)^2\sigma_w^2)$. Therefore, one has that the matrix $g_T\left[\int_0^T g_t^{-1}{g_t^{-1}}^\top dt\right] g_T^\top$ is asymptotically diagonal with value $(\exp(\phi'(0)^2\sigma_w^2 T)-1)/(\phi'(0)^2\sigma_w^2)$. Note that we rely on the assumption that the approximation errors due to considering each term separately vanish as $D \uparrow \infty$, or at least the approximation errors cancel out. Finally, $\widehat{y} = \boldsymbol{x}_{T,1}$ corresponds to selecting the first element of this diagonal matrix. If $E = e^{\phi'(0)^2\sigma_w^2 T}$ then

$$\mathcal{K}_b^{(i,j)} \to \mathcal{K}_b^{(i,j),\infty} = \frac{\sigma_b^2}{\sigma_w^2}(E-1).$$

Along similar lines we obtain the deterministic limit to which $K_{\mathcal{NT},W,t}^{(i,j)}$ converges as $D \uparrow \infty$, i.e.,

$$\mathcal{K}_W^{(i,j)} \to \mathcal{K}_W^{(i,j),\infty} = \lambda_0^{(i,j),\infty}\phi'(0)^2\sigma_w^2 TE + \frac{\sigma_b^2}{\sigma_w^2}\left[\phi'(0)^2\sigma_w^2 TE - (E-1)\right]$$

hence obtaining

$$\mathcal{K}^{(i,j)} \to \mathcal{K}^{(i,j),\infty} = \lambda_0^{(i,j),\infty} CE + \frac{\sigma_b^2}{\sigma_w^2} CE.$$

where $E = \exp(C)$ and $C = \phi'(0)^2\sigma_w^2 T$. This can be contrasted with the main result of Section 5.3, where we have shown that the (standard) kernel corresponding to $D \uparrow \infty$ for $\phi''(0) = 0$ is given by

$$K^{(i,j)} \to K^{(i,j),\infty} = \lambda_0^{(i,j),\infty}(E-1) + \frac{\sigma_b^2}{\sigma_w^2}(E-1).$$

Note that the two kernels correspond to two different training regimes: i) training all layers of the neural network; ii) training only the output layer of the neural network. However, the two kernels are qualitatively similar. In particular, both kernels depend linearly on $\lambda_0^{(i,j),\infty}$. The only difference is with respect to the behavior of $(E-1)$ compared to CE as a function of C.

5.5 Training of completed ResNets

Results presented in Section 5.3 and Section 5.4 entail a neural network with an infinite-dimensional input. Let $z, z' \in \mathbb{R}^Z$[1] be two inputs of the neural network. We consider a linear adaptation layer, i.e. an embedding, of the form $\boldsymbol{x}_0 = Az$ where, in line with Section 5.3 and Section 5.4, the elements of $A \in \mathbb{R}^{D\times Z}$ are i.i.d. as $\mathcal{N}(0, \sigma_Z^2)$. It follows that across d we have

$$(\boldsymbol{x}_{0,d}, \boldsymbol{x}'_{0,d}) \overset{i.i.d.}{\sim} \mathcal{N}_2(0, \Sigma^Z(z, z')),$$

where

$$\Sigma^Z(z, z') = \sigma_Z^2 \begin{bmatrix} \|z\|^2 & \langle z, z'\rangle \\ \langle z', z\rangle & \|z'\|^2 \end{bmatrix}.$$

By the strong law of large numbers $\lambda_0 = \frac{1}{D}\langle \boldsymbol{x}_0, \boldsymbol{x}'_0\rangle \to \lambda_0^\infty = \mathbb{E}[\boldsymbol{x}_{0,1}\boldsymbol{x}'_{0,1}] = \sigma_Z^2\langle z, z'\rangle$ as $D \uparrow \infty$, hence $q_0^\infty = \sigma_Z^2\|z\|^2$ and ${q'}_0^\infty = \sigma_Z^2\|z'\|^2$. In the weakly training setting, which is

[1] for convenience we use in this section the z, z' notation instead of $z^{(i)}, z^{(j)}$, and proceed in the same way for all other quantities depending on i, j

equivalent to Bayesian inference with a Gaussian process prior, we know from Section 5.3 that across d we have

$$(x^{\infty}_{T,d}, {x'}^{\infty}_{T,d} \,|\, \boldsymbol{x}_{0,d}, \boldsymbol{x'}_{0,d}) \overset{i.i.d.}{\sim} \mathcal{N}_2\left(\begin{bmatrix} \boldsymbol{x}_{0,d} \\ \boldsymbol{x'}_{0,d} \end{bmatrix}, \Sigma^{\text{weak}}(z,z') \right),$$

where

$$\Sigma^{\text{weak}}(z,z') = \sigma_Z^2(E-1) \begin{bmatrix} \|z\|^2 & \langle z, z' \rangle \\ \langle z', z \rangle & \|z'\|^2 \end{bmatrix} + \frac{\sigma_b^2}{\sigma_w^2}(E-1)$$

with $E = e^{\phi'(0)^2 \sigma_w^2 T}$. Then, by direct computation, we obtain the following Gaussian distribution

$$\begin{aligned}(x^{\infty}_{T,d}, {x'}^{\infty}_{T,d}) &\overset{i.i.d.}{\sim} \mathcal{N}_2\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma^Z(z,z') + \Sigma^{\text{weak}}(z,z') \right) \\ &= \mathcal{N}_2\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \sigma_Z^2 E \begin{bmatrix} \|z\|^2 & \langle z, z' \rangle \\ \langle z', z \rangle & \|z'\|^2 \end{bmatrix} + \frac{\sigma_b^2}{\sigma_w^2}(E-1) \right).\end{aligned}$$

That is, the prior distribution induced by a doubly infinite ResNet with the input adaptation layer is i.i.d. across the dimensions d, and distributed as a centered Gaussian process with kernel

$$\overline{K}(z,z') = \sigma_Z^2 E \langle z, z' \rangle + \frac{\sigma_b^2}{\sigma_w^2}(E-1). \tag{41}$$

We also augment the neural network with an output adaptation layer $\widehat{y} = \boldsymbol{G}\boldsymbol{x}_T$, where the elements of $G \in \mathbb{R}^{1 \times D}$ are i.i.d. as $\mathcal{N}(0, \sigma_Y^2/D)$. Then, it follows that the doubly infinite ResNet with both input and output adaption layers still follows a Gaussian process whose kernel is

$$\overline{\overline{K}}(z,z') = \sigma_Y^2 \overline{K}(z,z'). \tag{42}$$

Now, consider the Bayesian noiseless linear model with fully independent prior distributions formulated by $\widehat{y} = \alpha + \beta z$ where $\alpha \in \mathbb{R}$, $\alpha \sim \mathcal{N}(0, \sigma_\alpha^2)$, $\beta \in \mathbb{R}^Z$, $\beta_i \sim \mathcal{N}(0, \sigma_\beta^2)$ for $i = 1, \ldots, Z$, then:

$$(\widehat{y}, \widehat{y}') \sim \mathcal{N}_2\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \sigma_\beta^2 \begin{bmatrix} \|z\|^2 & \langle z, z' \rangle \\ \langle z', z \rangle & \|z'\|^2 \end{bmatrix} + \sigma_\alpha^2 \right).$$

Thus, according to (41) it follows that, within the doubly infinite limit, the completed ResNet prior model collapses to a noiseless Bayesian linear regression prior where $\sigma_\alpha^2 = \frac{\sigma_b^2}{\sigma_w^2}\sigma_Y^2(E-1)$ and $\sigma_\beta^2 = \sigma_Y^2 \sigma_Z^2 E$.

Under the fully trained setting, in Section 5.4 we have established the convergence of the NTK. Now, we consider directly the doubly infinite ResNet augmented with both input and output layers as previously defined. Recall that in the NTK literature the input layer is sometimes not trained, and the output layer is sometimes omitted (Arora et al., 2019). Hereafter, we report only the results for the special case in which all layers are present and trained as it most closely resembles standard practice for finitely-sized networks. In particular,

$$\overline{\overline{\mathcal{K}}}(z,z') = \sigma_Y^2 \left(\sigma_Z^2 (C+1) E \langle z, z' \rangle + \frac{\sigma_b^2}{\sigma_w^2} CE \right) + \overline{\overline{K}}(z,z')$$

$$= \sigma_Y^2 \left(\sigma_Z^2 (C+2) E \langle z, z' \rangle + \frac{\sigma_b^2}{\sigma_w^2} (CE + E - 1) \right). \tag{43}$$

For more general cases, the results follow along lines similar to the steps detailed in Section 5.4. In particular, in all cases the kernel remains affine in $\langle z, z' \rangle$, and only the coefficients are affected.

The problem of establishing the equivalence between kernel regression and fully trained neural networks requires two steps: i) establishing the NTK convergence at initialization, as we have proved in Section 5.4; ii) bounding the NTK fluctuations during training (Arora et al., 2019; Lee et al., 2019a). To the best of our knowledge, step ii) has not been formally established for architectures which are not feed-forward, e.g. for ResNets. Assuming such a result, Equation (43) shows that in the doubly infinite limit fully trained ResNet correspond to noiseless (kernel) linear regression. Kernel regression is equivalent to the posterior predictive mean of a Gaussian process with the same kernel. Hence, relatively to point predictions, both weakly and fully trained doubly infinite ResNets of the considered class collapse to a linear model.

5.6 Concluding remarks

We considered doubly infinite ResNets by first establishing a diffusion limit ($L \uparrow \infty$) and then by considering increasing dimensionality of the diffusion process ($D \uparrow \infty$). Here, we briefly present the alternative approach where we invert the order to taking limits, that is first the width tends to infinity, then the depth tends to infinity. We assume fully i.i.d. network's parameters according to a time-discretization of Assumption 7, i.e. (18) and (19), with $\sigma_b^2 = 0, \sigma_w^2 = 1$ for simplicity of exposition. We consider two inputs $x, y \in \mathbb{R}^I$ and the same input adaption layer of Section 5.5, with $\sigma_I^2 = 1$. Once again for simplicity of exposition, we consider in this section the ResNet (11) without the ϕ activation function. In such a setting, the case $L \uparrow \infty$ is thoroughly studied, starting with the seminal work of Neal (1995) and proceeding with the more recent developments of Lee et al. (2018), Jacot et al. (2018) and Lee et al. (2019b). By means of the standard approach we obtain the following recursion over the layers of an infinitely wide ResNet:

$$(x_{0,d}, y_{0,d}) \overset{i.i.d.}{\sim} \mathcal{N}_2(0, \Sigma_0)$$

with

$$\Sigma_0^{xy} = \langle x, y \rangle,$$

and

$$(x_{t+\Delta t,d}, y_{t+\Delta t,d}) \overset{i.i.d.}{\sim} \mathcal{N}_2(0, \Sigma_{t+\Delta t} = \Sigma_t + \Sigma_{t:t+\Delta t})$$

with

$$\Sigma_{t:t+\Delta t}^{xy} = \mathop{\mathbb{E}}_{\varepsilon,\gamma \sim \Sigma_t^{xy}} [\psi(\varepsilon)\psi(\gamma)] \Delta t,$$

where $d = 1, \dots$ are the neural network units, which are i.i.d. in the infinitely wide limit, and we used the notation Σ^{xy} to denote upper-right element of a 2×2 covariance matrix (the remaining elements are obtained setting $x = y$ or $y = x$). The recursion over layer depth (time):

$$\Sigma_{t+\Delta t}^{xy} = \Sigma_t^{xy} + \mathop{\mathbb{E}}_{\varepsilon,\gamma \sim \Sigma_t^{xy}} [\psi(\varepsilon)\psi(\gamma)] \Delta t$$

provides an easy way to establish the infinitely deep limit ($D \uparrow \infty$) via the ODE limit of the recursion over

$$\dot{\Sigma}_t^{xy} = \mathop{\mathbb{E}}_{\varepsilon,\gamma \sim \Sigma_t^{xy}}[\psi(\varepsilon)\psi(\gamma)]. \tag{44}$$

The extension to an arbitrary number of inputs is immediate. In particular, the resulting continuum of Gaussian processes, one at each $t \geq 0$, requires to solve the ODE (44). The expectation defining the drift of (44) has a closed-form solution for some specific choices of the activation functions, and requires numerical approximations in the general case. The ODE (44) also requires numerical integration aside from specific cases, such as ψ being the identity function. It is easy to check that in this case one obtains the same results of Section 5.3 for $\phi'(0) = 1$ (for instance when $\phi = \tanh$). Thus, in this setting, the order in which the limits of depth and width are taken does not matter. When $\phi'(0) \neq 1$, the results of this Section only differ by the multiplicative constant $\phi'(0)$ which can be absorbed into σ_b and σ_w, resulting again in an equivalent model. Moreover, the derivations of this Section can be extended to the cover the case of a suitably smooth activation ϕ, in which case the constant $\phi'(0)$ would be recovered as well. Ultimately, the ϕ activation plays a very limited role for large L, and when the ψ activation is missing (i.e. it is the identity) the ResNet tends to a linear neural network in the doubly infinite limit.

6. Numerical results

We start by introducing all the neural network models considered in this section. In all the experiments we set ψ to the identity function and, without loss of generality, we assume $T = 1$.

Regarding fully-connected networks, we consider the fully i.i.d. parametrization of Assumption 7. When $Z = 1$, i.e. for 1-dimensional inputs, we can opt for copying the input across all dimensions: $\boldsymbol{x}_{0,\bullet} = z$ for an input z, i.e. $\boldsymbol{x}_{0,d} = z$ for each $d \in D$. We refer to this model as $\mathcal{F}_{\tanh}$ when $\phi = \tanh$ and as $\mathcal{F}_{\text{swish}}$ when $\phi = \text{swish}$. The swish activation function ($\text{swish}(x) = x\,\text{sigmoid}(x)$) has been shown empirically (Ramachandran et al., 2017) and theoretically (Hayou et al., 2019a) to be competitive. More in general, for any input dimension Z, we complete the model with input and output adaptation layers as defined in Section 5.5. We choose to use $\sigma_Z^2 = Z/I$ and $\sigma_Y = 1/D$. We will refer to such completed models as $\overline{\overline{\mathcal{F}}}_{\tanh}$ and $\overline{\overline{\mathcal{F}}}_{\text{swish}}$.

Regarding convolutional networks, we consider the fully i.i.d. parameterization of Assumption 10. A generic input is here of dimension $U \times V \times C$, with U, V, C representing the input height, width and number of channels. The adaptation layer is here an 1-by-1 convolution adapting the number of channels to the model dimension D. More precisely: for each p $\boldsymbol{x}_{0,p} = Az_p$ where p index the UV positions and the elements of $A \in \mathbb{R}^{D \times Z}$ are i.i.d. as $\mathcal{N}(0, 1/C)$. The output layer is composed again of a 1-by-1 convolution which is followed by global space averaging. That is: $\widehat{y} = \frac{1}{UV}\sum_{p=1}^{UV} G\boldsymbol{x}_{T,p}$, here again p index the UV positions and the elements of $G \in \mathbb{R}^{Y \times D}$ are i.i.d. as $\mathcal{N}(0, 1/D)$. We refer to this convolutional model with $\phi = \tanh$ as $\overline{\overline{\mathcal{C}}}_{\tanh}$.

6.1 Correctness checks

We start with a numerical study of the correctness of the results of Section 3, Section 4 and Section 5. We consider $\mathcal{F}_{\text{tanh}}$ with $\sigma_w^2 = \sigma_b^2 = 1$ and two 1-dimensional inputs $z^{(1)} = 0$, $z^{(2)} = 1$, hence $\boldsymbol{x}_{0,\bullet}^{(1)} = z^{(1)}, \boldsymbol{x}_{0,\bullet}^{(2)} = z^{(2)}$, and simulate 10.000 draws of the first dimension $(d = 1)$ of:

a) $\boldsymbol{x}_T^{(1)}$, $\boldsymbol{x}_T^{(2)}$ via the ResNet recursion (8);

b) $x_T^{(1)}$, $x_T^{(2)}$ via the discretization (6) of the limiting SDE (33);

c) $x_T^{(1),\infty}$, $x_T^{(2),\infty}$ via the analytical transition density (40).

for $L = D = 500$. We only consider the first dimension because, as observed in Section 5.3, in the limit of $D \uparrow \infty$ the dimensions are i.i.d. Our analysis imply that a) and b) are equivalent when $L \uparrow \infty$, and c) is equivalent to b) when additionally $D \uparrow \infty$. As both D and L are large we expect good agreement between the distributions corresponding to a) b) and c). Numerical results are reported in Figure 2 where indeed a good agreement with the theory is observed.

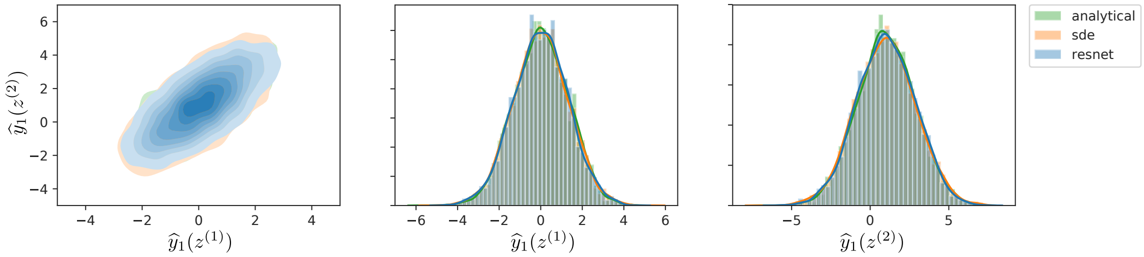

Figure 2: For model $\mathcal{F}_{\text{tanh}}$: 2D KDE (kernel density estimator) plot for $(\widehat{y}_1(z^{(1)}), \widehat{y}_1(z^{(2)}))$ (left), 1D KDE and histogram plots for $\widehat{y}_1(z^{(1)})$ (center), $\widehat{y}_1(z^{(2)})$ (right) when $\widehat{y}_1$ is sampled from a ResNet (resnet), from the Euler discretization of its limiting SDE for $L \uparrow \infty$ (sde) and from the analytical SDE transition density for $L, D \uparrow \infty$ (analytical); $\widehat{y}$ denotes a generic model output, hence $\widehat{y}_1$ is its first dimension.

For the same neural network model $\mathcal{F}_{\text{tanh}}$, Figure 3 displays the convergence of the NTKs $\mathcal{K}_W^{(1,2)}, \mathcal{K}_b^{(1,2)}$ to their limits $\mathcal{K}_W^{(1,2),\infty}, \mathcal{K}_b^{(1,2),\infty}$ for $z^{(1)} = 1$, $z^{(2)} = 2$. The convergence is assessed in the setting where both the depth L and the dimension D grow unbounded jointly. Results displayed in Figure 3 support the numerical analysis of Section 5.4. Results also support the conjecture that the order in which the limits are taken does not impact the results for the smooth activation functions considered in the present work.

6.2 Function space distributions

We show empirically that the dependency on the input is retained, and that the output distribution does not exhibit perfect correlation for very deep residual networks constructed as in the present paper. Again, we consider the neural network model $\mathcal{F}_{\text{tanh}}$ with $\sigma_w^2 = \sigma_b^2 = 1$.

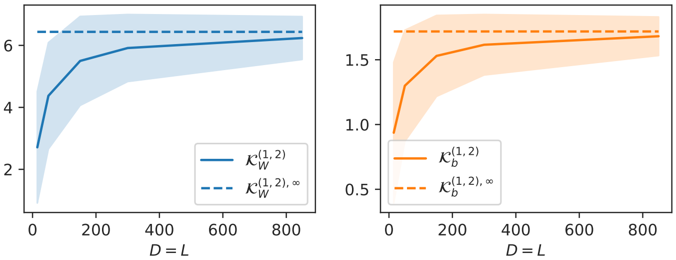

Figure 3: For model $\mathcal{F}_{\text{tanh}}$: NTKs $\mathcal{K}_W^{(1,2)}, \mathcal{K}_b^{(1,2)}$ as function of $L = D$ and their analytical limits $\mathcal{K}_W^{(1,2),\infty}, \mathcal{K}_b^{(1,2),\infty}$ corresponding to $D, L \uparrow \infty$; empirical average plotted with solid line, shaded areas correspond to ± 2 empirical standard deviations.

Figure 2 shows that $\boldsymbol{x}_{T,1}^{(1)}$ and $\boldsymbol{x}_{T,1}^{(2)}$ have different distributions. This means that the input dependency is retained in the neural network. Furthermore, from the left plot we see that $\boldsymbol{x}_{T,1}^{(1)}$ and $\boldsymbol{x}_{T,1}^{(2)}$ are not perfectly correlated, otherwise the 2D KDE would collapse to a straight line.

Figure 4 (top panels), which can be contrasted with Figure 1, displays samples of $\boldsymbol{x}_{T,1}$ from $\mathcal{F}_{\text{tanh}}$ in function space for different combinations of L and D. More specifically, we approximate function draws by considering 400 inputs $z^{(i)}$ equally spaced on $[-2, 2]$. Using the ResNet recursion (8) we obtain 400 output values $\boldsymbol{x}_{T,1}^{(i)}$. We repeat this procedure to obtain 10.000 function draws and report the results in Figure 4 (top). For $L = D = 500$, i.e. for jointly large width and depth, the function draws are close to linear in agreement with Section 5.4. We then replicate this experiment for $\mathcal{F}_{\text{swish}}$ and we report the results in Figure 4 (bottom panels). In this case $\phi'(0) = \phi''(0) = 1/2$ and Assumption 3 is not satisfied, but in this specific instance we did not observe divergent trajectories for the 10.000 function draws. The impact of adding an input adaptation layer is limited to symmetrizing the function space distributions around the origin, while $\mathcal{F}_{\text{tanh}}$ and $\mathcal{F}_{\text{swish}}$ trend upward with z. Hence, we do not include additional plots for this additional case as they add little information. Appendix B contains additional 2D plots of samples of $\boldsymbol{x}_{T,1}$ for both $\mathcal{F}_{\text{tanh}}$ and $\mathcal{F}_{\text{swish}}$.

Finally, Figure 5 (top panels) displays the correlations $\rho[\boldsymbol{x}_{T,1}^{(1)}, \boldsymbol{x}_{T,1}^{(2)}]$ for the neural network's inputs $(z^{(1)}, z^{(2)})$ in the range $[-2, 2] \times [-2, 2]$, for the tanh and swish activation functions: for different inputs, the corresponding output correlations are far from 1. We refer to the model of Figure 1 with tanh activation as $\mathcal{E}_{\text{tanh}}$, and to the model of Figure 1 with ReLU activation as $\mathcal{E}_{\text{ReLU}}$. For the sake of comparison, we show in Figure 5 (bottom panels) the correlations $\rho[x_{last,1}^{(1)}, x_{last,1}^{(2)}]$ for pre-activation 1 for $\mathcal{E}_{\text{tanh}}$ and $\mathcal{E}_{\text{ReLU}}$: all correlations are close to 1.

6.3 Doubly infinite fully trained fully-connected ResNets

We consider the MNIST dataset (LeCun, 1998). In particular, each observation (z, k) is composed of an image z and a target k among 10 classes representing the numbers 0 to 9. We flatten the images obtaining $z \in \mathbb{R}^{784}$ and, as common, we rescale each z as $z/255$

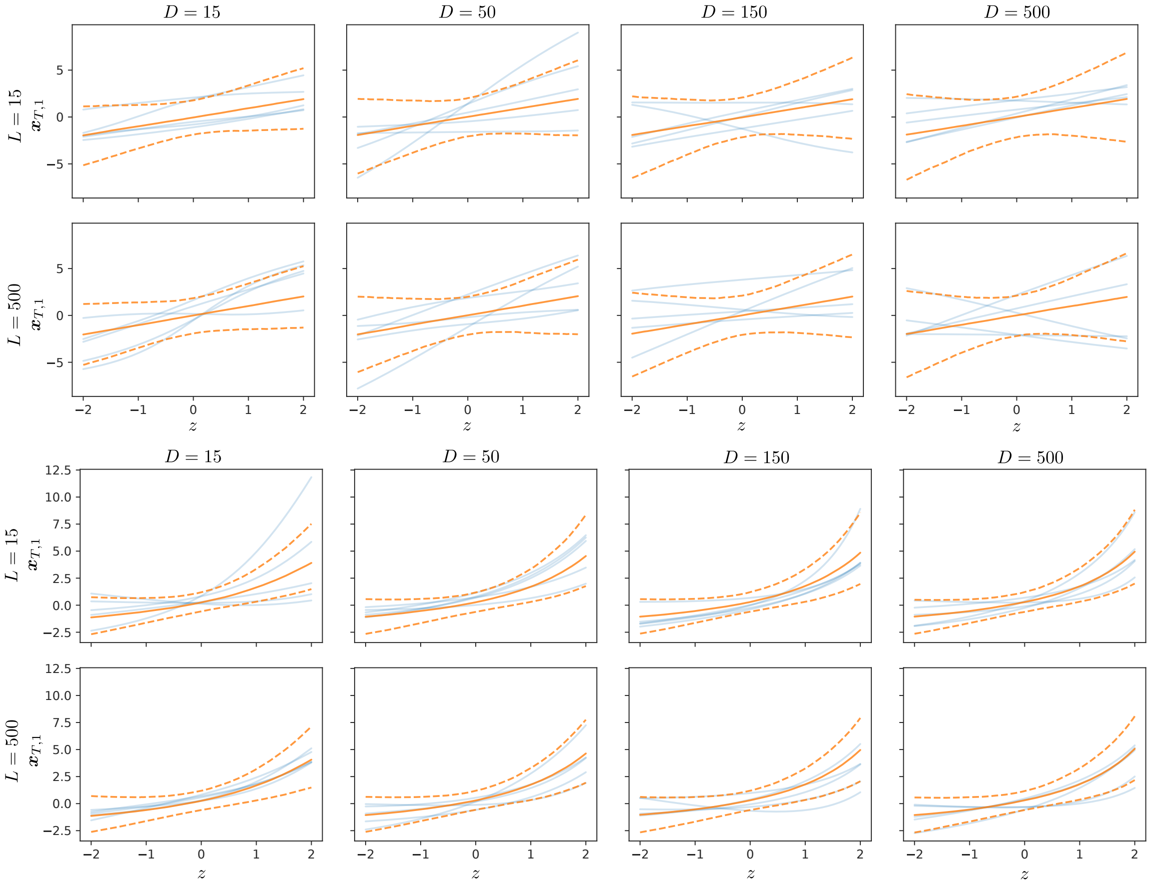

Figure 4: Function samples of $\boldsymbol{x}_{T.1}$ for $\mathcal{F}_{\text{tanh}}$ (top) and F_{swish} (bottom), see Figure 1 for the description of the plotted quantities.

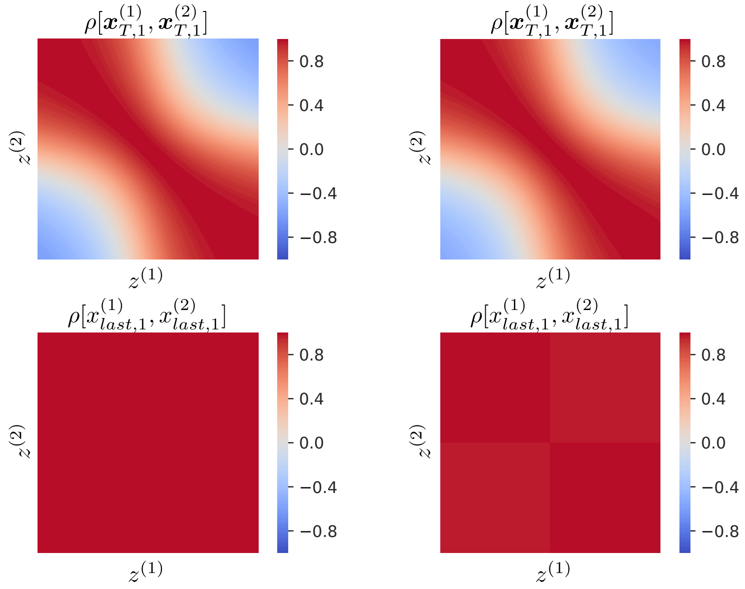

Figure 5: Output correlation heatmap for $\mathcal{F}_{\text{tanh}}$ (top-left), $\mathcal{F}_{\text{swish}}$ (top-right), $\mathcal{E}_{\text{tanh}}$ (bottom-left), $\mathcal{E}_{\text{ReLU}}$ (bottom-right).

to bound the inputs on $[0, 1]$. We consider the neural network $\overline{\overline{\overline{\mathcal{F}}}}_{\text{tanh}}$ trained via full-batch GD training and average MSE loss. In order to frame classification as a regression problem, we use 1-hot encoding: each class $k = 1, \ldots, 10$ is encoded as $y_k \in \mathbb{R}^{10}$ which has the k-th component equal to 1 and all other components equal to 0. The gradients are computed with respect to $(\varepsilon_t^W, \varepsilon_t^b)$ in (18) and (19).

We consider $\overline{\overline{\overline{\mathcal{F}}}}_{\text{tanh}}$ with $\sigma_w^2 = 1$ and $\sigma_b^2 = 0.1^2$. The use of a smaller bias variance is common in the NTK literature (Arora et al., 2019). From Section 5.5 we know that as L and D increase the fully trained $\overline{\overline{\overline{\mathcal{F}}}}_{\text{tanh}}$ collapses to noiseless Bayesian linear regression. We consider 20.000 randomly sampled observations from the training portion of the MNIST dataset, and we compute the test accuracy on the test portion of the MNIST dataset, which is composed of 10.000 observations. Using 1-hot encoding we perform kernel regression using kernel (43) via standard kernel regression (Williams and Rasmussen, 2006) for the predictive posterior mean of Gaussian processes. For numerical stability the model is augmented with a small noise variance equal to 1/20.000, and we obtain a test accuracy of 85.36%. We compare this accuracy with test accuracies computed for $\overline{\overline{\overline{\mathcal{F}}}}_{\text{tanh}}$ under different values of $D = L$, which is fully-trained for 120 epochs. We use a single learning rate tuned to optimize final test accuracy.

In practice, the training of neural networks typically is performed via SGD, or via other stochastic variants of GD, as full-batch training is prohibitively expensive for large datasets. Accordingly, here we perform SGD training of $\overline{\overline{\overline{\mathcal{F}}}}_{\text{tanh}}$, with batches of 200 observations each. Again, we consider 120 epochs and different different values of $D = L$. The same learning rate is used. In both experiments no further adjustments are performed, such as gradient clipping. The results are reported in Figure 6 (left). We observe that there is strong agreement between

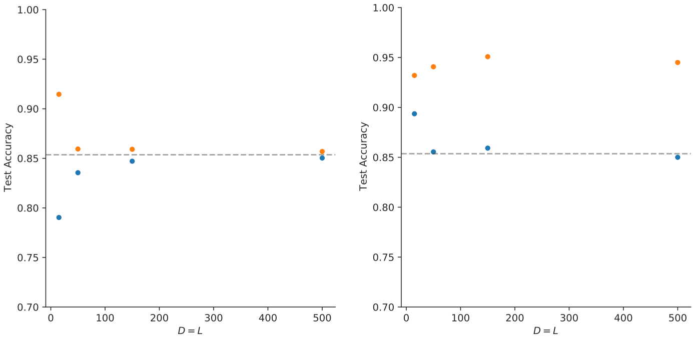

Figure 6: Final MNIST test accuracy after 120 epochs of training $\overline{\overline{\overline{\mathcal{F}}}}_{\mathrm{tanh}}$ via GD (left, blue), SGD (left, orange), full-batch Adam (right, blue), mini-batch Adam (right, orange) compared with the theoretical limiting value corresponding to Bayesian linear regression (dashed gray).

the limiting theoretical test accuracy and the final test accuracy of $\overline{\overline{\overline{\mathcal{F}}}}_{\mathrm{tanh}}$ fully trained with GD, which is the case covered by our theory. Moreover this result empirically extends to SGD.

For completeness, we consider the same training setting with Adam (Kingma and Ba, 2015), a popular adaptive stochastic optimizer, and we report the results in Figure 6 (right) for both full-batch and mini-batch variants. While there is no strong consensus on whether Adam outperforms or underperforms SGD when a carefully tuned learning rate is used (Wilson et al., 2017; Choi et al., 2019), Adam is known to be less sensitive to learning rate specifications and exhibits more robust behavior in difficult optimization problems. In particular, the proposed experiment provides an alternative viewpoint: Adam (with mini-batching, as standard) is able to "escape" the domain of attraction of linear model solutions, at least up to the largest model size here considered. We suspect that more complex neural network architectures might exhibit analogous pathologies at initialization when the number of network's parameters is very large, and Adam seems more robust to these issues. In any case, a formal investigation would require new results in the NTK literature to cover adaptive optimizers.

6.4 Doubly infinite fully trained convolutional ResNets

While a theoretical investigation of the backward properties of CNNs is beyond the scope of the present paper, in this section we empirically investigate to what extent the observations of Section 6.3 extends to convolutional neural networks. In particular, we consider $\overline{\overline{\mathcal{C}}}_{\mathrm{tanh}}$ with $\sigma_w^2 = 1$ and $\sigma_b^2 = 0.1^2$. The setting is the same of Section 6.3, with the exception that the input images are not flattened. We consider training under MSE loss for 120 epochs with both SGD and Adam. For computation reasons we restrict the maximum model size to $D = L = 150$ and do not investigate full-batch training. We report the results of this experiment in Figure 7. These results suggests convergence as $D = L$ which is reminiscent of

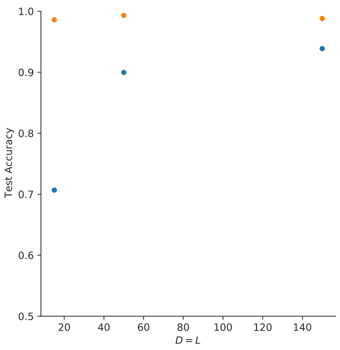

Figure 7: Final MNIST test accuracy after 120 epochs of training $\overline{\overline{\mathcal{C}}}_{\text{tanh}}$ with SGD (blue) and Adam (orange).

what observed in Figure 6. Similarly to Figure 6 Adam exhibits superior performance for large $D = L$.

6.5 Empirical Bayes for wide and deep fully-connected ResNets

In this section we present a proof of concept experiment where hyperparameter optimization is performed via empirical Bayes on a surrogate Bayesian linear model (BLM). In Section 5.5 we established that the prior of a ResNet with fully i.i.d. parameterization and the identity ψ activation converges to a BLM as width and depths grows unboundedly. We can thus consider the marginal likelihood of the limiting BLN as a proxy for the marginal likelihood of a wide and deep ResNet and leverage on it to conduct approximate empirical Bayes inference for the finite ResNet. More in detail, we consider a fully connected ResNet with the input adaptation layer detailed in Section 5.5. The corresponding limiting kernel in the doubly infinite case is given by (41). We consider a simple classification task consisting of discriminating between the digits '3' and '7' on the MNIST dataset. We randomly sample 100 images corresponding to the classes '3' and '7', which we will use to maximize the marginal likelihood. Each greyscale image z is mapped to the $[-1, +1]$ interval via the transform `z -> z/255*2-1`. The '3' class is mapped to -1 while '7' is mapped to $+1$. We assume a small i.i.d. Gaussian observation noise ($\sigma_e = 0.01$) to improve numerical stability. The marginal likelihood of the limiting BLM is optimized via a differential evolution algorithm (Price, 2013) over the positive hyperparameters $\sigma_z, \sigma_w, \sigma_b$. We obtain an average, i.e. per sample, negative marginal log-likelihood (NLL) of 0.65 which suggests a good fit to the data: for comparison the "default" hyperparameters $\sigma_z = \sigma_w = \sigma_b = 1.0$ result in a NLL of 3.25. In Figure 8 we plot the optimal and "default" NLL of the BLM jointly with the NLL of the ResNet for different widths and depths. The NLL of the finite ResNet is naïvely estimated via Monte Carlo integration over the hidden layers' units, with the last layer integrated out analytically. This estimation can be considered reliable only for large width, where the stochastic dependency of each layer on the previous layer becomes deterministic (Lee et al., 2018). Nonetheless, this simple approach suffices to illustrate the convergence of the finite width and depth NLLs to their limiting counterparts.

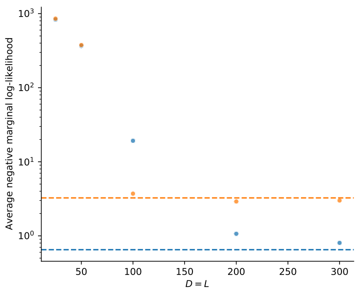

Figure 8: NLL of the limiting BLM (dashed line) and of the corresponding ResNet for different widths and depths (dots); orange: "default" hyperparameters, blue: optimal hyperparameters.

7. Discussion

We investigated the large-depth limit of identity ResNets (He et al., 2016b), establishing convergence to solutions of SDEs. Our results rely on smooth activation functions and on distributions for network's parameters which shrink as the total depth increases; further conditions on activation functions are obtained by restricting the limiting SDEs to be non explosive. Building upon our connection between infinitely deep fully-connected ResNets and diffusion processes, we showed that both forward and backward dynamics are well-behaved. With regards to forward-propagation, we showed that as the total depth grows unboundedly: i) the dependency of the last layer on the input does not vanish; ii) the last layer, as a stochastic function on input space, remains flexible without collapsing to restrictive families of distributions, iii) the last layer does not collapse to a deterministic limit, nor it diverges. Moreover, we established conditions on the activation functions in order to avoid explosive dynamics over the layers. With regards to backward-propagation, under non-explosivity conditions, we showed that the Jacobian of the final layer with respect to any layer can be expressed as the multiplication of two matrix diffusions which satisfy the same desiderata i), ii) and iii) in the large-depth limit, and hence it is well-behaved. Moreover, we addressed the problem of the trainability at initialization of such a neural network, showing that exploding gradients are not possible in the large-width limit, and that the ResNet is invertible. In contrast to the information propagation approach, our analysis covers finitely-wide neural networks and correlated network's parameters. While we focussed on fully-connected ResNets, there are no theoretical issues to extend our results to the more general convolutional architectures, as we have shown for the forward-propagation analysis of convolutional ResNets.

Limited to fully i.i.d. network's parameters and fully-connected neural networks without the second activation ψ, we investigated the case of the doubly infinite ResNets where both the network's depth and the network's width grow unboundedly. The attractiveness of the doubly infinite setting is mainly related to the potential of obtaining analytical results. We showed that doubly infinite fully-connected ResNets converge to Gaussian processes whose kernels can be computed by solving corresponding systems of ODEs. In particular, when

$\phi''(0) = 0$ and the model is completed with a fully i.i.d. input adaptation, we showed that the doubly infinite ResNet collapses to a Bayesian linear model with a fully factorial prior distribution. To conclude, we obtained the form of the NTK that corresponds to full training with continuous time gradient descent and quadratic loss of doubly infinite ResNets. In particular, we observed that such a kernel is qualitatively identical to the kernel of the Gaussian process arising in this doubly infinite limit, thus implying that fully trained doubly infinite networks of the considered class are again equivalent to performing linear regression. Although our results on the doubly infinite setting are not completely mathematically rigorous, numerical experiments support the validity of the proposed derivations. Our study on doubly infinite ResNets illustrate the pitfalls that must be overcome in order to derive non-trivial limits. Architectures, network's parameters and activation functions need to satisfy precise conditions. However, still under these conditions, the resulting limiting behavior can be very unexpressive. While this is an undesirable result if inference via the limiting process is the final goal, the connection to simple models allows to perform hyper-parameter optimization on the finitely-sized neural network by means of empirical Bayes procedures on the corresponding linear model.

The fields of diffusion processes and SDEs are mature and rich fields (Øksendal, 2003; Karatzas and Shreve, 1999; Revuz and Yor, 1999; Kloeden and Platen, 1992; Stroock and Varadhan, 2006), with a vast range of theoretical results and simulation methods. We envision that examining neural networks properties from the point of view of SDEs will bring further insights. Our study suggests two main research directions of future work. Firstly, to overcome modeling limitations one could narrow the gap between theory and practice by considering more realistic residual blocks consisting of multiple layers. This may be approached either via fractional Brownian motions (Biagini et al., 2008) or via re-scaled Brownian motions; such an extension would allow to consider neural networks which are infinitely wide only in the residual blocks internal dimension. Secondly, a mathematically rigorous treatment of of doubly infinite setting of Section 5 could be developed. This, as a main difficulty, involves to work with infinite-dimensional stochastic processes (Prato and Zabczyk, 2014). Indeed, the standard approach to establish large-width limits consists in postulating both the limiting and the converging processes on the same infinite space, while limiting the connectivity of the converging processes (Matthews et al., 2018). This approach allows to establish convergence on a space of fixed (infinite) dimension. While the standard theory of diffusion limits is well established (Stroock and Varadhan, 2006), it covers only the case of diffusion processes of finite-dimensions.

Appendices

A. Proofs

This appendix contains all the proofs of the theorems stated in the main text and the lemmas required to prove them.

Proof [Proof of Theorem 1] This is (Nelson, 1990, Theorem 2.2): Assumption 1 and the postulated weakly unique and non-explosive weak solution satisfy all the conditions required for the application of (Nelson, 1990, Theorem 2.2). Note that we use a stronger non-explosivity condition (Øksendal (2003)). Alternatively, for this standard result the reader can refer to the monograph Stroock and Varadhan (2006) on which Nelson (1990) is based; yet another reference is Ethier and Kurtz (2009). ■

Lemma 20 *If ϕ satisfies Assumption 5, $\epsilon \sim \mathcal{N}(0, \sigma^2)$ with $\sigma^2 \leq \sigma_*^2$, $\alpha > 0$, then we can find $M_2(\alpha, \sigma_*^2) < \infty$ and $M_3(\alpha, \sigma_*^2) < \infty$ such that:*

$$\begin{aligned}\mathbb{E}\left[|\phi''(\epsilon)|^\alpha\right] &\leq M_2(\alpha, \sigma_*^2)\\ \mathbb{E}\left[|\phi'''(\epsilon)|^\alpha\right] &\leq M_3(\alpha, \sigma_*^2)\end{aligned}$$

Proof We prove the result only for $\phi''(\epsilon)$, the case for $\phi'''(\epsilon)$ being identical. Let L large enough such that $|\phi''(x)| \leq K_1 e^{K_2|x|}$ for $|x| \geq L$ then:

$$\begin{aligned}\mathbb{E}\left[|\phi''(\epsilon)|^\alpha\right] &= \mathbb{E}\left[|\phi''(\epsilon)|^\alpha \mathbb{1}_{|\epsilon|\leq L}\right] + \mathbb{E}\left[|\phi''(\epsilon)|^\alpha \mathbb{1}_{|\epsilon|> L}\right]\\ &\leq \sup_{|x|\leq L} |\phi''(x)|^\alpha + K_1^\alpha\, \mathbb{E}[e^{K_2\alpha|\epsilon|}]\end{aligned}$$

The first term is finite, that the second one can be bounded by a finite and increasing function in σ^2 follows from the symmetry in law of ϵ and the form of its movement generating function. ■

Proof [Proof of Theorem 2] We suppress the dependency on t of vector and matrices and the conditioning in expectations and covariances in this proof to ease the notation. We also drop the boldness of x_t as no confusion arises in this setting. We instead reserve subscripts for indexing: for example x_d denotes the d-th element of a vector x.

Let $h = (\mu^W\sqrt{\Delta t} + \varepsilon^W)\psi(x) + (\mu^b\sqrt{\Delta t} + \varepsilon^b)$ so that $h\sqrt{\Delta t} = \Delta W \psi(x) + \Delta b$. By second order Taylor expansion of ϕ around 0 we have for $d = 1, \dots, D$

$$\frac{\Delta x_d}{\Delta t} = \frac{\phi(h_d\sqrt{\Delta t})}{\Delta t} = \phi'(0)h_d\Delta t^{-1/2} + \frac{1}{2}\phi''(0)h_d^2 + \frac{1}{6}\phi'''(\vartheta_d)h_d^3\Delta t^{1/2}$$

with $\vartheta_d \in (-h_d\sqrt{\Delta t}, h_d\sqrt{\Delta t})$. To prove (1) we want to show that $\forall R > 0$

$$\lim_{\Delta t \downarrow 0} \sup_{\|x\|<R} \left|\mu(x)_d - \mathbb{E}\left[\phi'(0)h_d\Delta t^{-1/2} + \frac{1}{2}\phi''(0)h_d^2 + \frac{1}{6}\phi'''(\vartheta)h_d^3\Delta t^{1/2}\right]\right| = 0.$$

Now, $h_d = (\mu^W_d \sqrt{\Delta t} + \varepsilon^W_d)\psi(x) + \mu^b_d\sqrt{\Delta t} + \varepsilon^b_d$ and the distribution assumptions on ε^W and ε^b lead to

$$\begin{aligned}
&\mathbb{E}\left[\phi'(0)h_d\Delta t^{-1/2} + \frac{1}{2}\phi''(0)h_d^2\right]\\
&= \phi'(0)(\mu^b_d + \mu^W_d\psi(x))\\
&\quad + \frac{1}{2}\phi''(0)\,\mathbb{V}[\varepsilon^W\psi(x) + \varepsilon^b]_{d,d}\\
&\quad + \frac{1}{2}\phi''(0)\left(\mu^b_d + \mu^W_d\psi(x)\right)^2\Delta t\\
&= \mu(x)_d + \frac{1}{2}\phi''(0)\left(\mu^b_d + \mu^W_d\psi(x)\right)^2\Delta t.
\end{aligned}$$

It remains to show that

$$\lim_{\Delta t\downarrow 0}\sup_{\|x\|<R}\left|\left(\mu^b_d + \mu^W_d\psi(x)\right)^2\right|\Delta t = 0,$$

which holds as ψ is locally bounded, and that

$$\lim_{\Delta t\downarrow 0}\sup_{\|x\|<R}\left|\mathbb{E}\left[\phi'''(\vartheta_d)h_d^3\right]\right|\Delta t^{1/2} = 0,$$

for which it suffices to show that $\sup_{\|x\|<R}\left|\mathbb{E}\left[\phi'''(\vartheta_d)h_d^3\right]\right|$ can be bounded by $M(R) < \infty$ uniformly in Δt. By Cauchy–Schwarz

$$\sup_{\|x\|<R}\left|\mathbb{E}\left[\phi'''(\vartheta_d)h_d^3\right]\right| \le \sup_{\|x\|<R}\mathbb{E}\left[\phi'''(\vartheta_d)^2\right]^{1/2}\sup_{\|x\|<R}\mathbb{E}\left[h_d^6\right]^{1/2}. \tag{45}$$

Again, as ψ is locally bounded the constraint $\sup_{\|x\|<R}$ corresponds to a constraint on the variance of h_d hence the second sup is finite. By Lemma 20 the first sup is finite too and not increasing in Δt as $|\vartheta_d| \le \sqrt{\Delta t}|h_d|$ which allows us to produce the desired bound $M(R)$.

Regarding (3), by first order Taylor expansion of ϕ around 0 we need to show that for $d = 1,\dots,D$ and $R > 0$

$$\lim_{\Delta t\downarrow 0}\sup_{\|x\|<R}\left|\mathbb{E}\left[\frac{\left(\phi'(0)h_d\Delta t^{1/2} + \frac{1}{2}\phi''(\vartheta_d)h_d^2\Delta t\right)^4}{\Delta t}\right]\right| = 0$$

with $\vartheta_d \in (-h_d\sqrt{\Delta t}, h_d\sqrt{\Delta t})$. Note that the term inside the expectation is composed of a sum of terms of the form $kh_d^n\phi''(\vartheta_d)^m\Delta t^\alpha$ for integers $n, m \ge 0$ and reals $\alpha > 0$, $k \in \mathbb{R}$. This results from repeated applications of the Cauchy–Schwarz inequality and Lemma 20 as we did previously to prove (1).

Regarding (2), we can compute $\mathbb{E}[\Delta x(\Delta x)^\top]/\Delta t$ instead of $\mathbb{V}[\Delta x]/\Delta t$ as in the infinitesimal limit of $\Delta t \downarrow 0$ the two quantities have to agree due to the convergence of the infinitesimal mean that we have already established. Hence by first order Taylor expansion of ϕ around 0 we need to show that for $d, u = 1,\dots,D$ and $R > 0$:

$$\lim_{\Delta t\downarrow 0}\sup_{\|x\|<R}\Bigg|\sigma^2(x)_{d,u}$$

$$- \mathbb{E}\left[\frac{\left(\phi'(0)h_d\Delta t^{1/2}+\frac{1}{2}\phi''(\vartheta_d)h_d^2\Delta t\right)\left(\phi'(0)h_u\Delta t^{1/2}+\frac{1}{2}\phi''(\vartheta_u)h_u^2\Delta t\right)}{\Delta t}\right]\Bigg| = 0$$

with $\vartheta_d \in (-h_d\sqrt{\Delta t}, h_d\sqrt{\Delta t})$, $\vartheta_u \in (-h_u\sqrt{\Delta t}, h_u\sqrt{\Delta t})$. The only term inside the expectation not vanishing in Δt is

$$\begin{aligned}
&\mathbb{E}[\phi'(0)^2 h_d h_u] \\
&= \phi'(0)^2\, \mathbb{V}[\varepsilon^W\psi(x)+\varepsilon^b]_{d,u} + \phi'(0)^2\left(\mu_d^b+\mu_d^W\psi(x)\right)\left(\mu_u^b+\mu_u^W\psi(x)\right)\Delta t \\
&= \sigma^2(x)_{d,u} + \phi'(0)^2\left(\mu_d^b+\mu_d^W\psi(x)\right)\left(\mu_u^b+\mu_u^W\psi(x)\right)\Delta t.
\end{aligned}$$

The (uniform on compacts) convergence of all terms aside from $\sigma^2(x)_{d,u}$ to 0 once again follows from repeated applications of the Cauchy–Schwarz inequality and Lemma 20.

Now, the continuity of $\mu(x)$ and $\sigma(x)$ are a consequence of the continuity of the conditional covariance $\mathbb{V}[\varepsilon^W\psi(x)+\varepsilon^b]$, and as $\mathbb{V}[\varepsilon^W\psi(x)+\varepsilon^b]$ is positive semi-definite so is $\sigma^2(x)$. Hence all the conditions of Assumption Assumption 1 hold true.

Finally, as ψ is differentiable two times with continuity, it follows from the dependency of μ and σ^2 on x only through $\mathbb{V}[\varepsilon^W\psi(x)+\varepsilon^b]$ that Assumption 2 is satisfied too. The application of Theorem 1 completes the proof. ■

Proof [Proof of Corollary 3] Notice that

$$d[W\psi(x)]_t + d[b]_t = d[W\psi(x)+b]_t = \operatorname{diag}(\mathbb{V}[\varepsilon_t^W\psi(x_t)+\varepsilon_t^b|x_t])dt$$

Then expanding dW_t and db_t in (13) shows that the drift terms are matched between (12) and (13). The quadratic variation of (12) is

$$\phi'(0)^2\operatorname{diag}(\mathbb{V}[\varepsilon_t^W\psi(x_t)+\varepsilon_t^b|x_t])dt$$

which is equal to the quadratic variation of (13) as it is computed as

$$d[x]_t = d[\phi'(0)(W\psi(x)+b)]_t = \phi'(0)^2 d[W\psi(x)+b]_t$$

This shows the equivalence in law between the solution of (12) and the solution of (13). Then (14) immediately follows by direct computation. ■

Proof [Proof of Corollary 4 and Corollary 5] Notice that

$$\begin{aligned}
&d[W\psi(x^{(i)})+b, W\psi(x^{(j)})+b]_t \\
&= \mathbb{C}[\varepsilon_t^W\psi(x_t^{(i)})+\varepsilon_t^b, \varepsilon_t^W\psi(x_t^{(j)})+\varepsilon_t^b|x_t^{(i)}, x_t^{(j)}]dt \\
&= \left(\Sigma^b + \mathbb{C}[\varepsilon_t^W\psi(x_t^{(i)}), \varepsilon_t^W\psi(x_t^{(j)})|x_t^{(i)}, x_t^{(j)}]\right)dt
\end{aligned}$$

and

$$\mathbb{C}[\varepsilon_t^W\psi(x_t^{(i)}), \varepsilon_t^W\psi(x_t^{(j)})|x_t^{(i)}, x_t^{(j)}]_{r,c}$$

$$
\begin{aligned}
&= \mathbb{E}[(\varepsilon^W_{t,r,\bullet}\psi(x^{(i)}_t))(\varepsilon^W_{t,c,\bullet}\psi(x^{(j)}_t))|x^{(i)}_t, x^{(j)}_t] \\
&= \sum_{d,u=1}^{D} \psi(x^{(i)}_{t,d})\psi(x^{(j)}_{t,u})\,\mathbb{E}[W_{r,d}W_{c,u}] \\
&= \Sigma^{W_O}_{r,c} \sum_{d,u=1}^{D} \psi(x^{(i)}_{t,d})\psi(x^{(j)}_{t,u})\Sigma^{W_I}_{d,u} \\
&= \Sigma^{W_O}_{r,c}(\psi(x^{(i)}_t)^\top \Sigma^{W_I}\psi(x^{(j)}_t)).
\end{aligned}
$$

This proves Corollary 4. Corollary 5 follows by setting $\sigma^b = \sigma_b \,\mathrm{I}_D$, $\sigma^{W_I} = \mathrm{I}_D$ and $\sigma^{W_O} = \sigma_w D^{-1/2}\,\mathrm{I}_D$. ■

Proof [Proof of Theorem 14] Here g_t is a $D \times D$ matrix-valued SDE instead of standard D-dimensional (vector) SDEs. All the theory presented in Section 2 continues to hold with the obvious modifications by working on the vectorization of matrix-valued processes. When establishing the limits for g_t in Assumption 1 the conditioning is both on g_t and on x_t, indeed the convergence to the limiting process is obtained jointly in x_t and g_t. This proof follows the exact same path of the proofs of Theorem 2 and Corollary 3 so we highlight the main steps only. And once again we suppress the dependency on t of vector and matrices, the conditioning in expectations and covariances, and the boldness of x_t and g_t as no confusion arises in this setting:

$$\Delta g = \big(\phi'(\Delta W\psi(x) + \Delta b)1_D{}^\top \odot \Delta W \odot 1_D\psi'(x)^\top\big)g$$

Let $h = (\mu^W\sqrt{\Delta t} + \varepsilon^W)\psi(x) + (\mu^b\sqrt{\Delta t} + \varepsilon^b)$ so that $h\sqrt{\Delta t} = \Delta W\psi(x) + \Delta b$. By second order Taylor expansion of ϕ' around 0 we have for $d = 1, \ldots, D$

$$\phi'(h_d\sqrt{\Delta t}) = \phi'(0) + \phi''(0)h_d\sqrt{\Delta t} + \frac{1}{2}\phi'''(\vartheta_d)h_d^2\Delta t$$

with $\vartheta_d \in (-h_d\sqrt{\Delta t}, h_d\sqrt{\Delta t})$. Then with $\vartheta = \big[\vartheta_1 \cdots \vartheta_D\big]$

$$\Delta g = \big((\phi'(0)1_D\sqrt{\Delta t} + \phi''(0)h\Delta t + \frac{1}{2}\phi'''(\vartheta)h^2\Delta t^{3/2})1_D{}^\top \odot (\mu^W\sqrt{\Delta t} + \varepsilon^W) \odot 1_D\psi'(x)^\top\big)g$$

In order to obtain the instantaneous mean of g we need to compute

$$
\begin{aligned}
\mathbb{E}\left[\frac{\Delta g}{\Delta t}\right] &= \phi'(0)(\mu^W \odot 1_D\psi'(x)^\top)g + \phi''(0)(\mathbb{E}[\varepsilon^W\psi(x)1_D^\top \odot \varepsilon^W] \odot 1_D\psi'(x)^\top)g + r_{g,\mu}(g, x, \Delta t) \\
&= \mu_g(g, x) + r_{g,\mu}(g, x, \Delta t)
\end{aligned}
$$

where $r_\mu(g, x, \Delta t)$ is a vector-valued remainder term and we want to show that for each $R > 0$

$$\lim_{\Delta t \downarrow 0} \sup_{\|g\|+\|x\|<R} \|r_{g,\mu}(g, x, \Delta t)\| = 0$$

By first order Taylor expansion of ϕ' around 0 we have for $d, d' = 1, \ldots, D$

$$\Delta g_{d,d'} = \big((\phi'(0) + \phi''(\vartheta_d)h_d\sqrt{\Delta t})1_D^\top \odot \Delta W_{d,\bullet} \odot \psi'(x)\big)g_{\bullet,d'}$$

with $\vartheta_d \in (-h_d\sqrt{\Delta t}, h_d\sqrt{\Delta t})$. In order to obtain the instantaneous covariance of g we need to compute for $d, d', u, u' = 1, \dots, D$

$$\begin{aligned}\mathbb{E}\left[\frac{\Delta g_{d,d'}\Delta g_{u,u'}}{\Delta t}\right] &= \phi'(0)^2\,\mathbb{E}[(\varepsilon^W_{d,\bullet} \odot \psi'(x))g_{\bullet,d'}(\varepsilon^W_{u,\bullet} \odot \psi'(x))g_{\bullet,u'}] + r_{g,\sigma}(g, x, \Delta t)_{d,d',u,u'}\\ &= \sigma^2_g(g, x)_{d,d',u,u'} + r_{g,\sigma^2}(g, x, \Delta t)_{d,d',u,u'}\end{aligned}$$

where $r_{g,\sigma^2}(g, x, \Delta t)$ is a remainder term (a 4 dimensional tensor, as $\sigma^2_g(g, x)$) and we want to show that for each $R > 0$

$$\lim_{\Delta t\downarrow 0}\ \sup_{\|g\|+\|x\|<R} \|\mathrm{vec}(r_{g,\sigma^2}(g, x, \Delta t))\| = 0$$

Again by first order Taylor expansion of ϕ' around 0 we want to prove for $d, d' = 1, \dots, D$

$$\lim_{\Delta t\downarrow 0}\ \sup_{\|g\|+\|x\|<R} \|c_g(g, x, \Delta t)\| = 0$$

where

$$c_g(g, x, \Delta t) = \mathbb{E}\left[\frac{\Big(\big((\phi'(0) + \phi''(\vartheta)h\sqrt{\Delta t})1_D^\top \odot \Delta W \odot 1_D\psi'(x)\big)g\Big)^4}{\Delta t}\right]$$

(here the fourth power is element-wise) with $\vartheta = \begin{bmatrix}\vartheta_1 \cdots \vartheta_D\end{bmatrix}$ and $\vartheta_d \in (-h_d\sqrt{\Delta t}, h_d\sqrt{\Delta t})$ to satisfy the continuity in probability requirement.

Then the limit

$$\lim_{\Delta t\downarrow 0}\ \sup_{\|g\|+\|x\|<R} \big(\|r_{g,\mu}(g, x, \Delta t)\| + \|\mathrm{vec}(r_{g,\sigma^2}(g, x, \Delta t))\| + \|c_g(g, x, \Delta t)\|\big) = 0$$

again follows from repeated applications of the Cauchy–Schwarz inequality and Lemma 20. Now, Assumption 2 follows from the linearity of the expectation operator and the positive semi-definiteness of $\sigma^2_g(g, x)$ is easily checked. The equivalence of (31) to the matrix-SDE defined by $\mu_g(g, x)$ and $\sigma^2_g(g, x)$ is established by comparing the drift and quadratic covariation terms. This completes the proof. ■

Proof [Proof of Corollary 15] Let $dZ_t = \big(\phi'(0)dW_t + \phi''(0)d[W\psi(x)1_D{}^\top \odot W]_t\big) \odot 1_D\psi'(x_t)$. Then g_t given by $dg_t = dZ_tg_t$ is the right stochastic exponential of Z_t which we denote, following Protter (2005), as $g = \mathcal{E}^R(Z)$. Let define the (left) stochastic exponential $u = \mathcal{E}(Z)$ of Z_t by $du_t = u_tdZ_t$. From (Protter, 2005, Chapter V, Theorem 48) we know that g_t is invertible and that

$$\mathcal{E}(Z)\mathcal{E}^R(-Z + [Z, Z]) = \mathrm{I}_D$$

It follows that

$$\mathcal{E}^R(Z)\mathcal{E}(-Z + [Z, Z]) = \mathrm{I}_D$$

hence $g^{-1} = \mathcal{E}(-Z + [Z, Z])$ which completes the proof. ■

Lemma 21 *When ψ is the identity function SDE (20) is equivalent, in distribution, to the representation where each data point i has an associated D-dimensional BM $B_t^{(i)}$, $\{B_t^{(i)}\}_{i=1}^N$ are dependent over i, and each $B_t^{(i)}$ corresponds to both the weights and biases sources of randomness. That is, the SDE (20) is equivalent to the following*

$$dx_t^{(i)} = \phi'(0)(\sigma_b^2 + \sigma_w^2 q_t^{(i)})^{1/2} dB_t^{(i)} + \frac{1}{2}\phi''(0)(\sigma_b^2 + \sigma_w^2 q_t^{(i)}) 1_D dt$$

$$d[B^{(i)}, B^{(j)}]_t = \frac{\sigma_b^2 + \sigma_w^2 \lambda_t^{(i,j)}}{\big((\sigma_b^2 + \sigma_w^2 q_t^{(i)})(\sigma_b^2 + \sigma_w^2 q_t^{(j)})\big)^{1/2}} \mathrm{I}_D \, dt$$

Let $\widetilde{\mu}_x(x) = \frac{1}{2}\phi''(0)(\sigma_b^2 + \sigma_w^2 q(x))$, $\widetilde{\sigma}_x(x) = \phi'(0)(\sigma_b^2 + \sigma_w^2 q(x))^{1/2}$, $\widetilde{\sigma}_{xx}(x,y) = \phi'(0)(\sigma_b^2 + \sigma_w^2 \lambda(x,y))^{1/2}$. Then the processes $m_t^{(i)}$, $q_t^{(i)}$, $\lambda_t^{(i,j)}$ follow the SDEs:

$$dm_t^{(i)} = \widetilde{\mu}_x(x_t^{(i)})dt + \widetilde{\sigma}_x(x_t^{(i)})\frac{1}{D}\sum_{d=1}^{D} dB_{t,d}^{(i)}$$

$$dq_t^{(i)} = \left(2\widetilde{\mu}_x(x_t^{(i)})m_t^{(i)} + \widetilde{\sigma}_x^2(x_t^{(i)})\right)dt + 2\widetilde{\sigma}_x(x_t^{(i)})\frac{1}{D}\sum_{d=1}^{D} x_{t,d}^{(i)} dB_{t,d}^{(i)}$$

$$\begin{aligned} d\lambda_t^{(i,j)} &= \left(\widetilde{\mu}_x(x_t^{(i)})m_t^{(j)} + \widetilde{\mu}_x(x_t^{(j)})m_t^{(i)} + \widetilde{\sigma}_{xx}^2(x_t^{(i)}, x_t^{(j)})\right)dt \\ &+ \frac{1}{D}\widetilde{\sigma}_x(x_t^{(i)})\sum_{d=1}^{D} x_{t,d}^{(j)} dB_{t,d}^{(i)} + \frac{1}{D}\widetilde{\sigma}_x(x_t^{(j)})\sum_{d=1}^{D} x_{t,d}^{(i)} dB_{t,d}^{(j)} \end{aligned}$$

where $m_t^{(i)} = \frac{1}{D}\sum_{d=1}^{D} x_{t,d}^{(i)}$.

Proof This is again a direct consequence of multi-dimensional Ito's formula (Øksendal, 2003, Section 4.2). ■

Proof [Heuristic for Proposition 18] From Lemma 21 we have

$$[m^{(i)}]_T = \frac{1}{D}\int_0^T \widetilde{\sigma}_x^2(x_t^{(i)})dt$$

$$[q^{(i)}]_T = \frac{1}{D}\int_0^T 4\widetilde{\sigma}_x^2(x_t^{(i)})q_t^{(i)}dt$$

$$[\lambda^{(i,j)}]_T = \frac{1}{D}\int_0^T \widetilde{\sigma}_x^2(x_t^{(i)})q_t^{(j)} + \widetilde{\sigma}_x^2(x_t^{(j)})q_t^{(i)}dt$$

where $\widetilde{\sigma}_x^2(x) = \phi'(0)^2(\sigma_b^2 + \sigma_w^2 q(x))$. Assuming that $q_t^{(i)}$ can be controlled (for instance bounds on SDE solutions can be used to bound $\mathbb{E}[\sup_{0\leq t\leq T} q_t^{(i)}]$ when $\phi''(0) = 0$) all the quadratic variations can be shown to converge to 0 leaving out only the deterministic component. The rest of Proposition 18 follows by assuming that the small noise limit of the SDEs is given by the corresponding ODEs, and by computing the ODEs solutions. ■

Proof [Heuristic for Proposition 19] We know from Proposition 18 that

$$
\begin{aligned}
&\phi'(0)(\sigma_b^2 + \sigma_w^2 q_t^{(i)}) \to \phi'(0)(\sigma_b^2 + \sigma_w^2 q_t^{(i),\infty}) \\
&\frac{1}{2}\phi''(0)(\sigma_b^2 + \sigma_w^2 q_t^{(i)}) \to \frac{1}{2}\phi''(0)(\sigma_b^2 + \sigma_w^2 q_t^{(i),\infty}) \\
&\frac{\sigma_b^2 + \sigma_w^2 \lambda_t^{(i,j)}}{\big((\sigma_b^2 + \sigma_w^2 q_t^{(i)})(\sigma_b^2 + \sigma_w^2 q_t^{(j)})\big)^{1/2}} \to \frac{\sigma_b^2 + \sigma_w^2 \lambda_t^{(i,j),\infty}}{\big((\sigma_b^2 + \sigma_w^2 q_t^{(i),\infty})(\sigma_b^2 + \sigma_w^2 q_t^{(j),\infty})\big)^{1/2}}
\end{aligned}
$$

as $D \uparrow \infty$. Then Proposition 19 follows by assuming that the solution of (33) converges to the solution of (39) as $D \uparrow \infty$ and by computing the transition density of (39). ■

B. Additional plots

In Figure 9 we plot 2D function samples of $\boldsymbol{x}_{T,1}$ for $\mathcal{F}_{\text{tanh}}$ and F_{swish} to complement the visualizations of Section 4.2.

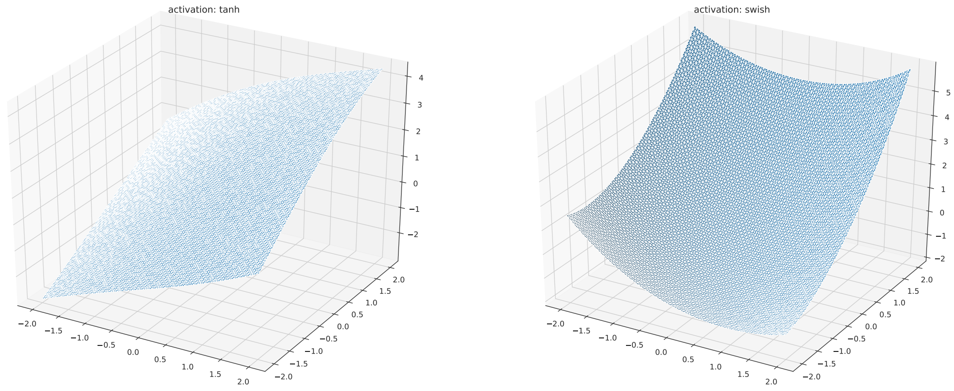

Figure 9: Function samples of $\boldsymbol{x}_{T,1}$ for $\mathcal{F}_{\text{tanh}}$ (left) and F_{swish} (right) for $L = 100$ and $D = 100$ on the rectangle $[-2, 2] \times [-2, 2]$.

Acknowledgements

The authors are grateful to an Associate Editor and three anonymous Referees for all their constructive comments, corrections, and suggestions which improved remarkably the paper. Stefano Favaro is also affiliated to IMATI-CNR “Enrico Magenes" (Milan, Italy). Stefano Favaro received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme under grant agreement No 817257. Stefano Favaro gratefully acknowledge the financial support from the Italian Ministry of Education, University and Research (MIUR), “Dipartimenti di Eccellenza” grant 2018-2022.

References

Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Ruslan Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. In *Advances in Neural Information Processing Systems 32*, 2019.

Jens Behrmann, Will Grathwohl, Ricky TQ Chen, David Duvenaud, and Joern-Henrik Jacobsen. Invertible residual networks. In *International Conference on Machine Learning*, pages 573–582, 2019.

Francesca Biagini, Yaozhong Hu, Bernt Øksendal, and Tusheng Zhang. *Stochastic calculus for fractional Brownian motion and applications*. Springer Science & Business Media, 2008.

Patrick Billingsley. *Convergence of Probability Measures*. Wiley-Interscience, 2nd edition, 1999.

Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018.

Tian Qi Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems 31*, pages 6571–6583, 2018.

Dami Choi, Christopher J Shallue, Zachary Nado, Jaehoon Lee, Chris J Maddison, and George E Dahl. On empirical comparisons of optimizers for deep learning. *arXiv preprint arXiv:1910.05446*, 2019.

Vincent Dumoulin and Francesco Visin. A guide to convolution arithmetic for deep learning. *arXiv preprint arXiv:1603.07285*, 2016.

Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Brandon Tran, and Aleksander Madry. Adversarial robustness as a prior for learned representations, 2019.

Stewart N Ethier and Thomas G Kurtz. *Markov processes: characterization and convergence*. Wiley-Interscience, 2009.

Adrià Garriga-Alonso, Carl Edward Rasmussen, and Laurence Aitchison. Deep convolutional networks as shallow gaussian processes. In *International Conference on Learning Representations*, 2019.

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

Arjun K Gupta and Daya K Nagar. *Matrix variate distributions*. Chapman and Hall/CRC, 1st edition, 1999.

Soufiane Hayou, Arnaud Doucet, and Judith Rousseau. On the impact of the activation function on deep neural networks training. In *Proceedings of the 36th International Conference on Machine Learning*, pages 2672–2680, 2019a.

Soufiane Hayou, Arnaud Doucet, and Judith Rousseau. Training dynamics of deep networks using stochastic gradient descent via neural tangent kernel. *arXiv preprint arXiv:1905.13654*, 2019b.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016a.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016b.

Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems 31*, pages 8571–8580, 2018.

Ioannis Karatzas and Steven Shreve. *Brownian Motion and Stochastic Calculus*. Springer, 2nd edition, 1999.

Diederick P Kingma and Jimmy 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, corrected edition, 1992.

Yann LeCun. The mnist database of handwritten digits. *http://yann.lecun.com/exdb/mnist/*, 1998.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.

Jaehoon Lee, Jascha Sohl-dickstein, Jeffrey Pennington, Roman Novak, Sam Schoenholz, and Yasaman Bahri. Deep neural networks as gaussian processes. In *International Conference on Learning Representations*, 2018.

Jaehoon Lee, Lechao Xiao, Samuel S. Schoenholz, Yasaman Bahri, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent, 2019a.

Jaehoon Lee, Lechao Xiao, Samuel S Schoenholz, Yasaman Bahri, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. In *Advances in Neural Information Processing Systems 32*, 2019b.

Alexander G. de G. Matthews, Mark Rowland, Jiri Hron, Richard E Turner, and Zoubin Ghahramani. Gaussian process behaviour in wide deep neural networks. *arXiv preprint arXiv:1804.11271*, 2018.

Radford M Neal. *Bayesian Learning for Neural Networks.* PhD thesis, University of Toronto, 1995.

Daniel B Nelson. Arch models as diffusion approximations. *Journal of econometrics*, 45(1-2): 7–38, 1990.

B. Øksendal. *Stochastic Differential Equations: An Introduction with Applications.* Springer, 6th edition, 2003.

Stefano Peluchetti and Stefano Favaro. Infinitely deep neural networks as diffusion processes. In *Proceedings of the twenty-third international conference on artificial intelligence and statistics*, 2020.

Ben Poole, Subhaneil Lahiri, Maithra Raghu, Jascha Sohl-Dickstein, and Surya Ganguli. Exponential expressivity in deep neural networks through transient chaos. In *Advances in Neural Information Processing Systems 29*, pages 3360–3368, 2016.

Giuseppe Da Prato and Jerzy Zabczyk. *Stochastic Equations in Infinite Dimensions.* Second edition edition, 2014.

Kenneth V Price. Differential evolution. In *Handbook of optimization*, pages 187–214. Springer, 2013.

Philip E Protter. *Stochastic integration and differential equations.* Springer, 2005.

Prajit Ramachandran, Barret Zoph, and Quoc V Le. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017.

Daniel Revuz and Marc Yor. *Continuous Martingales and Brownian Motion.* Springer, 3rd edition, 1999.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

Samuel S Schoenholz, Justin Gilmer, Surya Ganguli, and Jascha Sohl-Dickstein. Deep information propagation. In *International Conference on Learning Representations*, 2017.

Daniel W Stroock and SR Srinivasa Varadhan. *Multidimensional diffusion processes.* Springer, 2006 edition, 2006.

Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning.* MIT press Cambridge, MA, 2006.

Ashia C Wilson, Rebecca Roelofs, Mitchell Stern, Nati Srebro, and Benjamin Recht. The marginal value of adaptive gradient methods in machine learning. In *Advances in Neural Information Processing Systems*, pages 4148–4158, 2017.

Ge Yang and Samuel Schoenholz. Mean field residual networks: On the edge of chaos. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 7103–7114. Curran Associates, Inc., 2017.