Infinitely deep neural networks as diffusion processes

Stefano Peluchetti
speluchetti@cogent.co.jp
Cogent Labs

Stefano Favaro
stefano.favaro@unito.it
Department ESOMAS
University of Torino
and Collegio Carlo Alberto

Abstract

When the parameters are independently and identically distributed (initialized) neural networks exhibit undesirable properties that emerge as the number of layers increases, e.g. a vanishing dependency on the input and a concentration on restrictive families of functions including constant functions. We consider parameter distributions that shrink as the number of layers increases in order to recover well-behaved stochastic processes in the limit of infinite depth. This leads to set forth a link between infinitely deep residual networks and solutions to stochastic differential equations, i.e. diffusion processes. We show that these limiting processes do not suffer from the aforementioned issues and investigate their properties.

1 Introduction

Modern neural networks (NN) models featuring a large number of layers (depth) and features per layer (width) have achieved a remarkable performance across many domains (LeCun et al., 2015). It is well known (Neal, 1995; Matthews et al., 2018) that in the limit of infinite width, NNs whose parameters are appropriately distributed converge to Gaussian processes. This connection helps to study properties of very wide NNs, and forms the basis of inferential algorithms directly targeting the infinite-dimensional setting (Lee et al., 2018; Garriga-Alonso et al., 2019; Lee et al., 2019; Arora et al., 2019). Based on this recent literature, it is natural to ask whether it is possible to set an analogous useful connection between infinitely deep neural networks (IDNN) and stochastic processes. At a first glance, this correspondence might prove elusive. To see why, we now look at the literature on initialization schemes. Indeed there is a duality between initialization schemes and Bayesian NNs: an initialization scheme can be seen as a prior on the model parameters, thus inducing a prior on the NN. A NN at initialization may thus be viewed as a stochastic process indexed by depth, whose distribution is defined by a sequence of conditional distributions mapping from each layer to the next. Early works focused on stabilizing the variance of key quantities of interest across the layers of deep NNs (Glorot and Bengio, 2010; He et al., 2015). More recent works (Poole et al., 2016; Schoenholz et al., 2017; Hayou et al., 2019a) consider the impact of initializations to the propagation of the input signal.

Even when initialized on the edge of chaos (EOC) for optimal signal propagation, feedforward NNs with fixed independent and identically distributed (i.i.d.) initialization exhibit some pathological properties as their total depth increases. In particular, the dependency on the input eventually vanishes for most activation functions. In addition to that, the layers seen as random functions on the input space eventually concentrate on restrictive families including constant functions. As an illustrative example, we show in Figure 1 function samples from the last layer of a feedforward deep NNs for two activation functions under EOC initialization. For a tanh activation, the input 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 smooth activation functions. For a ReLU activation, the input affects the variance of the output and the function samples are piece-wise linear. In both cases, the outputs of any two inputs end up perfectly correlated. While this study applies to feedforward NNs, very deep residual networks (ResNet) suffer from similar issues (Yang and Schoenholz, 2017), with the additional issue that the

Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

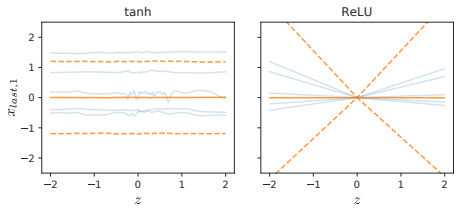

Figure 1: Function samples of a given pre-activation (number 1) of the last layer, $x_{last,1}$, of a fully connected feedforward NN with 500 layers of 500 units over a 1-dimensional input $z \in [-2, 2]$; tanh activation function and ReLU activation function, and parameters on the edge of chaos; 5 draws are displayed in blue in each figure; for each input the 5%, 50% and 95% quantiles are displayed in orange.

variance of the Gaussian-distributed pre-activations may grow unbounded over layers.

While it is possible to obtain a *well-defined* stochastic process corresponding to an IDNN, such a process is unexpressive: linear regression is a more flexible alternative. The difficulties discussed so far are determined by the fact that typical prior distributions on the model parameters introduce a constant level of randomness over each hidden layer. In this paper we consider prior distributions that depend on the number of layers, in such a way that they shrink as the number layers increases. This approach leads to our main result: as the number of layers increases, a class of ResNets converges, jointly over multiple inputs, to diffusion processes on a finite time interval. The conditions required for attaining convergence provide us with a general guideline for selecting compatible NN architectures, activation functions and parameters distributions. The limiting diffusion processes satisfy suitable stochastic differential equations (SDE) that describe the evolution of IDNN layers over time (depth). The limiting diffusion is *well-behaved* in the sense that: i) it retains dependency from the input; ii) it does not suffer from the perfect correlation constraint; iii) it does not collapse to a deterministic function nor does it diverge.

The paper is structured as follows. In Section 2 we recall some preliminary results on diffusion limits of discrete-time stochastic process. Section 3 contains our main result: the convergence of a class of ResNets to solutions of SDEs. Section 4 contains numerical experiments and Section 5 concludes. Proofs, additional experiments and plots, and additional discussions on related work are deferred to the Supplementary Material (SM).

Notation: for a matrix h, $h^\top$ is its transpose, and if h is square $\mathrm{diag}(h)$ is its diagonal vector and $\mathrm{Tr}(h)$ is its trace; $\|x\| = \sqrt{x^\top x}$ is the norm of the vector x; $\langle x, y\rangle = x^\top y$ is the inner product of vectors x and y; $\|h\| = \sqrt{\mathrm{Tr}(h^\top h)}$ is the norm of a matrix h; $\mathrm{vec}(u)$ is the vectorization the tensor u; I is the identity matrix and $\mathrm{1}$ is a vector of ones; for random variables z and w, $\mathrm{var}[z]$, $\mathrm{cov}[z, w]$ and $\rho[z, w]$ are the variance, covariance and correlation; for random vectors $x \in \mathbb{R}^r$ and $y \in \mathbb{R}^c$, $\mathbb{C}[x, y]_{i,j} = \mathrm{cov}[x_i, x_j]$ is the $r \times c$ cross-covariance matrix $\mathbb{C}[x, y]$; $\mathbb{V}[x] = \mathbb{C}[x, x]$ is the $r \times r$ covariance matrix of x; the expectation $\mathbb{E}[u]$ of a random tensor u is the tensor of the expectations of its elements; for two D-dimensional stochastic processes x_t, y_t, $[x]_t$ is the quadratic variation (a D-dimensional vector) and $[x, y]_t$ is the quadratic covariation (a $D \times D$-dimensional matrix); $\mathbb{1}$ is the indicator function.

2 Preliminaries

For $l = 1, \dots, L$ let $\boldsymbol{x}_l$ be the l-th layer of a NN with with L layers, and let $\boldsymbol{x}_0$ be the NN input. In this section we recall general results for diffusion approximations. The connection with NNs, i.e. defining what $\boldsymbol{x}_l$ exactly represents in a NN, is postponed to the next section. As we will be seeking a continuous time stochastic process limit we re-index $\boldsymbol{x}_0, \boldsymbol{x}_1, \dots, \boldsymbol{x}_L$ on a discrete time scale. Let $T > 0$ denote a terminal time, $\Delta t = T/L$, for each L we establish the correspondence between discrete indices $l \in \mathbb{Z}_+$ and discrete times $t \in \mathbb{R}_+$ by $l = 0, 1, \dots, L \leftrightarrow t = 0, \Delta t, 2\Delta t, \dots, T$. From now on we will consider without loss of generality a NN with input $\boldsymbol{x}_0$ and layers $\boldsymbol{x}_{\Delta t}, \dots, \boldsymbol{x}_T$, denoting a layer with $\boldsymbol{x}_t$.

Let $p(\boldsymbol{x}_T|\boldsymbol{x}_0)$ be the conditional distribution of the output given the input for a NN at initialization. Our strategy to enforce desirable properties on $p(\boldsymbol{x}_T|\boldsymbol{x}_0)$ consists in having a NN converge, as the number of layers L go to infinity ($\Delta t \downarrow 0$), to a continuous-time stochastic process on the time interval $[0, T]$. In this case, for L large enough, the distribution $p(\boldsymbol{x}_T|\boldsymbol{x}_0)$ will be close to the distribution of the limiting process at terminal time T given the same $\boldsymbol{x}_0$, and such limiting process should be chosen to make this transition density well behaved. In all NN architectures considered in this paper, each layer depends only on the previous one, hence $\boldsymbol{x}_t$ has the Markov property. These conditions identify a class of diffusion processes (Stroock and Varadhan, 2006), which are continuous-time Markov processes with continuous paths, as natural candidates for the limiting process. For simplicity we assume that the parameters of all layers follow the same distribution (extensions are discussed in Section 3.2), making $\boldsymbol{x}_t$ time-homogeneous.

Let $\boldsymbol{x}_t$ be a generic D-dimensional discrete-time Markov process and let $\Delta \boldsymbol{x}_t = \boldsymbol{x}_{t+\Delta t} - \boldsymbol{x}_t$ define the forward

increments. Hereafter we report a set of conditions that imply the convergence of $\boldsymbol{x}_t$ to the solution of a limiting SDE, and it is implicit that the distribution $p(\boldsymbol{x}_{t+\Delta t}|\boldsymbol{x}_t)$ depends on Δt for the limits to exist as required.

Assumption 2.1 (Convergence of instantaneous mean and covariance)**.** *There exist $\mu_x(x) : \mathbb{R}^D \to \mathbb{R}^D$ and $\sigma_x^2(x) : \mathbb{R}^D \to \mathbb{R}^{D\times D}$ such that:*

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

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

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

for some $\delta > 0$, where all convergences are uniform on compacts of $\mathbb{R}^D$ for each component, $\mu_x(x)$ and $\sigma_x^2(x)$ are continuous, and $\sigma_x^2(x)$ is positive semi-definite: $\sigma_x^2(x) = \sigma_x(x)\sigma_x(x)^\top$ for some $\sigma_x(x) : \mathbb{R}^D \to \mathbb{R}^{D\times D}$.

Assumptions (1) and (2) pinpoint the form of the limiting SDE, while assumption (3) is a technical condition that allows us to consider the limits (1) and (2) instead of their truncated version (Nelson, 1990). The next theorem establishes that, under additional assumptions, in the limit $\boldsymbol{x}_t$ can be embedded in the solution of a SDE.

Theorem 2.1. *Under Assumption 2.1, extend $\boldsymbol{x}_t$ to a continuous-time process $\overline{\boldsymbol{x}}_t$ on $t \in [0, T]$ by continuous-on-right step-wise-constant interpolation of $\boldsymbol{x}_t$:*

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

Consider the D-dimensional SDE on $[0, T]$ with initial value $x_0 = \boldsymbol{x}_0$, drift vector $\mu_x(x)$ given by (1), and diffusion matrix $\sigma_x(x)$ given by a square root of (2):

$$dx_t = \mu_x(x_t)dt + \sigma_x(x_t)dB_t \tag{5}$$

where B_t is a D-dimensional Brownian motion (BM) with independent components and (5) is short-hand notation for:

$$x_T = x_0 + \int_0^T \mu_x(x_t)dt + \int_0^T \sigma_x(x_t)dB_t$$

The first integral is a standard (Riemann) integral, and the second integral is an Ito integral. If SDE (5) admits a weak solution, and if this solution is unique in law and non-explosive, then the stochastic process defined by (4) converges in law to the solution of the SDE (5). This result still holds true for a random but 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 law 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).

We are dealing with three processes: the (discrete-time) NN $\boldsymbol{x}_t$, its continuous time interpolation $\overline{\boldsymbol{x}}_t$, and the limiting diffusion x_t (Øksendal, 2003). In Theorem 2.1, 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 one. The convergence established in Theorem 2.1 is strong in the sense that it concerns the convergence of the distribution of the stochastic process $(\overline{\boldsymbol{x}}_t)_{t\in[0,T]}$ as a stochastic object on the whole time interval $[0, T]$ to the diffusion limit $(x_t)_{t\in[0,T]}$ as $L \uparrow \infty$. We consider weak solutions, as opposed to a strong ones, where it suffices that a BM B_t can be found such that a solution can be obtained (Øksendal, 2003). The focus on weak solutions and uniqueness in law of such solutions (also called weak uniqueness) is justified by our interest in the distributional properties of the limiting behavior of $\boldsymbol{x}_t$, and it enables us to consider weaker requirements for attaining convergence of $\boldsymbol{x}_t$. Consider the discretization of SDE (5)

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

where ζ_t is a D-dimensional random vector whose components are i.i.d. as standard Gaussian (mean 0 and variance 1). Under suitable conditions (Kloeden and Platen, 1992), it can be proved that the discretized SDE (6) converges to the SDE (5), and we recognize the Euler discretization of an ordinary differential equation (ODE) in the deterministic part (6). In Theorem 2.1 we postulate the existence and uniqueness in law of the weak solution of the limiting SDE, and its non-explosive behavior. The following conditions suffice for our goals.

Assumption 2.2 (Existence of weak solution and uniqueness in law on compact sets)**.** *The functions $\mu_x(x)$ and $\sigma_x(x)$ are twice continuously differentiable.*

Assumption 2.3 (Non-explosive solution)**.** *There exist a finite $C > 0$ such that for each $x \in \mathbb{R}^D$: $\|\mu_x(x)\| + \|\sigma_x(x)\| \le C(1 + \|x\|)$.*

When Assumption 2.1 and Assumption 2.2 hold (as it will be the case in all the models considered), but Assumption 2.3 does not hold, we still obtain convergence to the solution of the SDE (5). However, the stochastic process x_t might diverge to infinity with positive probability on any time interval. We will return to this point more in detail.

3 Residual network diffusions

We focus on unmodified, albeit simplified, standard architectures. This is in line with the information propagation research (Poole et al., 2016; Schoenholz et al., 2017; Hayou et al., 2019a) but in contrast with

Chen et al. (2018), where the recursion is modified with an additional Δt term to achieve convergence to a limiting ODE.

In this section, we study the implications of Assumption 2.1, Assumption 2.2 and Assumption 2.3 in NNs. First of all, $\boldsymbol{x}_t$ needs to be of constant dimensionality, as otherwise $\Delta \boldsymbol{x}_t$ is undefined. Consistently with the previous section we assume $\boldsymbol{x}_t \in \mathbb{R}^D$. For Assumption 2.1 to hold we need $\Pr(\|\Delta \boldsymbol{x}_t\| > \varepsilon | \boldsymbol{x}_t) \downarrow 0$ as $\Delta t \downarrow 0$ for any $\varepsilon > 0$, i.e. we require the increments to vanish eventually. Intuitively this is due to the continuity of the paths of the limiting diffusion process. A fully connected feedforward NN is expressed by the relationship $\boldsymbol{x}_{t+\Delta t} = f_t(\boldsymbol{x}_t) = \phi(A_t \boldsymbol{x}_t + a_t)$ for a nonlinear activation $\phi : \mathbb{R} \to \mathbb{R}$ applied element-wise. As standard convention 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 reasoning rules out the ResNet originally introduced in the work of He et al. (2016a), where $\boldsymbol{x}_{t+\Delta t} = f_t(\boldsymbol{x}_t + r_t(\boldsymbol{x}_t))$. This leaves us with the identity ResNet of He et al. (2016b) where $\boldsymbol{x}_{t+\Delta t} = \boldsymbol{x}_t + r_t(\boldsymbol{x}_t)$ for some choice of r_t, the residual blocks, which we require to eventually vanish.

3.1 Shallow residual blocks

Each residual block r_t results from an interleaved application of affine transforms and non-linear activation functions. We consider the case of shallow residual blocks of the form:

$$\boldsymbol{x}_{t+\Delta t} = \boldsymbol{x}_t + \phi(A_t \psi(\boldsymbol{x}_t) + a_t) \tag{7}$$

for two activation functions $\phi : \mathbb{R} \to \mathbb{R}$, $\psi : \mathbb{R} \to \mathbb{R}$ which are applied element-wise. We point out that the non-standard use of 2 activation functions ϕ, ψ is to cover the case of shallow residual blocks in full generality.

3.2 Parameter distribution and activation functions

For a shallow residual block r_t, the vanishing increments requirement is satisfied by having the distributions of A_t and a_t concentrate around 0 provided that $\phi(0) = 0$. It proves advantageous to consider weights and biases given by increments of diffusions corresponding to solvable SDEs.

Assumption 3.1 (Parameters distribution and scaling)**.** *Let W_t and b_t be the diffusion processes respectively with values in $\mathbb{R}^{D \times D}$ and $\mathbb{R}^D$ solutions of:*

$$dW_t = \mu^W dt + d\widetilde{W}_t; \quad d\,\mathrm{vec}(\widetilde{W}_t) = \sigma^W d\,\mathrm{vec}(B_t^W) \tag{8}$$

$$db_t = \mu^b dt + \sigma^b dB_t^b \tag{9}$$

where B_t^W and B_t^b are independent BMs with independent components respectively with values in $\mathbb{R}^{D \times D}$ and $\mathbb{R}^D$, $\mu^W \in \mathbb{R}^{D \times D}, \mu^b \in \mathbb{R}^D$, $\sigma^W \in \mathbb{R}^{D^2 \times D^2}, \sigma^b \in \mathbb{R}^{D \times D}$, and $\Sigma^W = \sigma^W {\sigma^W}^\top$, $\Sigma^b = \sigma^b {\sigma^b}^\top$ are positive semi-definite.

Then the discretizations of W_t and b_t admit the (exact) representations:

$$\Delta W_t = \mu^W \Delta t + \varepsilon_t^W \sqrt{\Delta t}; \quad \Delta b_t = \mu^b \Delta t + \varepsilon_t^b \sqrt{\Delta t}$$

$$\mathrm{vec}(\varepsilon_t^W) \overset{i.i.d.}{\sim} \mathcal{N}_{D^2}\left(0, \Sigma^W\right); \; \varepsilon_t^b \overset{i.i.d.}{\sim} \mathcal{N}_D\left(0, \Sigma^b\right)$$

for $t = \Delta t, \ldots, T$ where $\mathcal{N}$ stands for the multivariate Gaussian distribution. We will consider residual blocks where $A_t = \Delta W_t$ and $a_t = \Delta b_t$:

$$\boldsymbol{x}_{t+\Delta t} = \boldsymbol{x}_t + \phi(\Delta W_t \psi(\boldsymbol{x}_t) + \Delta b_t) \tag{10}$$

Thus Assumption 3.1 covers the case where the parameters are independently and identically distributed across layers according to an arbitrary multivariate Gaussian distribution, up to the required scaling which is necessary to obtain the desired diffusion limit. 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, commonly 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 will require:

Assumption 3.2 (Activation functions regularity)**.** *The function $\phi : \mathbb{R} \to \mathbb{R}$ satisfies: $\phi(0) = 0$, ϕ is continuously differentiable three times on $\mathbb{R}$, its second and third derivatives have at most exponential tails growth, i.e. for some $k > 0$:*

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

The function $\psi : \mathbb{R} \to \mathbb{R}$ is locally bounded and continuously differentiable two times on $\mathbb{R}$.

3.3 Diffusion limits

The next theorem is the main result of the present paper, regarding the convergence of (10). Proofs are in SM A.

Theorem 3.1. *Under Assumption 3.1 and Assumption 3.2 the continuous-time interpolation $\overline{\boldsymbol{x}}_t$ of $\boldsymbol{x}_t$ converges in law to the solution on $[0,T]$ of*

$$\begin{aligned} dx_t &= \phi'(0)(\mathbb{V}[\varepsilon_t^W \psi(x_t) + \varepsilon_t^b | x_t])^{1/2} dB_t \\ &+ \phi'(0)(\mu^b + \mu^W \psi(x_t))dt \\ &+ \frac{1}{2}\phi''(0)\,\mathrm{diag}(\mathbb{V}[\varepsilon_t^W \psi(x_t) + \varepsilon_t^b | x_t])dt \end{aligned} \tag{11}$$

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

This result does not establish a direct connection between x_t and the driving sources of stochasticity W_t and b_t. As we are interested in the properties of deep ResNets in function space, i.e. over multiple inputs, a brute force approach would require us to establish diffusion limits as in Theorem 3.1 for an enlarged $\boldsymbol{x}_t = [\boldsymbol{x}_t^{(1)} \cdots \boldsymbol{x}_t^{(N)}] \in \mathbb{R}^{DN}$ corresponding to N initial values $\boldsymbol{x}_0 = [\boldsymbol{x}_0^{(1)} \cdots \boldsymbol{x}_0^{(N)}]$. Instead, we show that the limiting SDE is equivalent in law to the solution of another SDE which preserves the dependency on the driving sources of stochasticity. From here on $\boldsymbol{x}_t^{(i)}, \boldsymbol{x}_t^{(j)}$ denote ResNets corresponding to two initial values $\boldsymbol{x}_0^{(i)}, \boldsymbol{x}_0^{(j)}$, and $x_t^{(i)}, x_t^{(j)}$ denotes diffusion limits corresponding to the same two initial values (i.e. $x_0^{(i)} = \boldsymbol{x}_0^{(i)}, x_0^{(j)} = \boldsymbol{x}_0^{(j)}$). We will continue to use $\boldsymbol{x}_t$ for $\boldsymbol{x}_t^{(i)}$ and x_t for $x_t^{(i)}$ when no confusion arises.

Corollary 3.1. *Under the same assumptions of Theorem 3.1 the limiting process is also given by the solution on $[0,T]$ of:*

$$\begin{aligned} dx_t^{(i)} &= \phi'(0)(dW_t \psi(x_t^{(i)}) + db_t) \\ &+ \frac{1}{2}\phi''(0)(d[W\psi(x^{(i)})]_t + d[b]_t) \end{aligned} \tag{12}$$

where W_t and b_t are defined in Assumption 3.1 and over two initial values we have:

$$d[x^{(i)}, x^{(j)}]_t = \phi'(0)^2(d[W\psi(x^{(i)}), W\psi(x^{(j)})]_t + d[b,b]_t) \tag{13}$$

The results obtained so far 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. We thus consider more restrictive distribution assumptions with a more manageable $\mathcal{O}(D^2)$ parametrization cost.

Assumption 3.3 (Matrix normal weights)**.** *Let $b_t, \mu^b, \sigma_b, B_t^b, \mu^W, B_t^W$ be defined as in Assumption 3.1. Let W_t be the diffusion matrix with values in $\mathbb{R}^{D\times D}$ 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.

Under Assumption 3.3 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 $t = \Delta t, \dots, T$ where $\mathcal{MN}$ stands for the matrix normal distribution. This is an immediate consequence of the fact that if $\zeta \sim \mathcal{MN}(0, \mathrm{I}, \mathrm{I})$, then $A\zeta B \sim \mathcal{MN}(0, AA^\top, B^\top B)$. See Gupta and Nagar (1999). The main property of $\mathcal{MN}$ distributions is that the covariance factorizes as $\mathrm{cov}(\varepsilon_{o,i}^W, \varepsilon_{o',i'}^W) = \Sigma_{o,o'}^{W_O}\Sigma_{i,i'}^{W_I}$.

Corollary 3.2. *Under the same assumptions of Theorem 3.1, if W_t is distributed according to Assumption 3.3, (12) and (13) are given by:*

$$\begin{aligned} dx_t^{(i)} &= \phi'(0)\big((\mu^W \psi(x_t^{(i)}) + \mu^b)dt \\ &+ \sigma^{W_O} dB_t^W \sigma^{W_I}\psi(x_t^{(i)}) + \sigma^b dB_t^b\big) \\ &+ \frac{1}{2}\phi''(0)\,\mathrm{diag}\left(\Sigma^b + \Sigma^{W_O}(\psi(x_t^{(i)})^\top \Sigma^{W_I}\psi(x_t^{(i)}))\right)dt \end{aligned} \tag{14}$$

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

Finally, we consider the simplest "fully i.i.d." centered distribution assumptions for W_t, b_t. i.i.d. initializations are most commonly used in the training of NNs. We also introduce a scaling of the weights by $D^{-1/2}$ (which is the same scaling used to obtain Gaussian process limits in infinitely wide NNs). We will see in Section 4.2 that this scaling has a stabilizing effect on the dynamics of x_t.

Assumption 3.4 (Fully i.i.d. parameters)**.** *Let W_t and b_t be the diffusion processes respectively with values in $\mathbb{R}^{D\times D}$ and $\mathbb{R}^D$ solutions of:*

$$dW_t = \frac{\sigma_w}{\sqrt{D}} dB_t^W; \quad db_t = \sigma_b dB_t^b$$

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

Under Assumption 3.4 the discretizations of W_t, b_t satisfy:

$$\Delta W_t = \varepsilon_t^W \frac{\sigma_w}{\sqrt{D}}\sqrt{\Delta t}; \qquad \Delta b_t = \varepsilon_t^b \sigma_b \sqrt{\Delta t} \tag{15}$$

$$\varepsilon_t^W \overset{i.i.d.}{\sim} \mathcal{MN}_{D,D}\big(0, \mathrm{I}_D, \mathrm{I}_D\big); \quad \varepsilon_t^b \overset{i.i.d.}{\sim} \mathcal{N}_D\big(0, \mathrm{I}_D\big) \tag{16}$$

Corollary 3.3. *Under the same assumptions of Theorem 3.1, if W_t and b_t are distributed according to Assumption 3.4, (12) and (13) are given by:*

$$\begin{aligned} dx_t^{(i)} &= \phi'(0)\big(\frac{\sigma_w}{\sqrt{D}}\|\psi(x_t^{(i)})\| dB_t^W + \sigma_b dB_t^b\big) \\ &+ \frac{1}{2}\phi''(0)\big(\sigma_b^2 + \frac{\sigma_w^2}{D}\|\psi(x_t^{(i)})\|^2\big)\,\mathrm{I}_D\,dt \end{aligned} \tag{17}$$

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

3.4 Qualitative properties

Non-vanishing input dependency: a consequence of Theorem 3.1 is that the distribution of the ResNet output given the input $p(\boldsymbol{x}_T|\boldsymbol{x}_0)$ converges to the transition density $p(x_T|x_0)$ of the solution of (12). 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 parameter distributions and T.

Flexible output distributions: from (12)-(13) we see that the joint evolution of $x_t^{(i)}, x_t^{(j)}$ corresponding to $x_0^{(i)}, x_0^{(j)}$ is not perfectly correlated (unless there are no weight parameters, a not very relevant case). This remains true also in the parameterizations of Assumption 3.3 and Assumption 3.4. 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.

Role of integration time: a standard time-change result for SDEs (Revuz and Yor, 1999) implies that time-scaling a SDE is equivalent to multiplying the drift and diffusion coefficients respectively by the scaling constant and by the square root of the scaling constant, as can be intuitively seen from (6). From (11) we see that 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 3.1 to leave the dynamics of (11) invariant. This remains true also in the parameterizations of Assumption 3.3 and Assumption 3.4. Hence we can restrict $T = 1$ without loss of generality.

Matrix normal weights: in this case $\mathbb{V}[\varepsilon_t^W \psi(x_t) + \varepsilon_t^b | x_t]$ is given by $\Sigma^b + \Sigma^{W_O}(\psi(x_t)^\top \Sigma^{W_I} \psi(x_t))$. The dependency on the state x_t in (11) goes through a linear transformation and a weighted inner product. This sheds some light on the impact of introducing dependencies among row and columns of the weight parameters $A_t = \Delta W_t$. Specifically, Σ^{W_I} define the structure of the inner weighted product, while Σ^{W_O} defines how such transforms affect each dimension $d \in D$.

Fully i.i.d. parameters: in this case $\mathbb{V}[\varepsilon_t^W \psi(x_t) + \varepsilon_t^b | x_t]$ is given by $\sigma_b^2 + \frac{\sigma_w^2}{D}\|\psi(x_t)\|^2$. The dependency on the state x_t in (11) goes only through the norm of x_t which is permutation invariant in $d \in D$. Thus the law of the processes $x_{t,d}$ is exchangeable across $d \in D$ if the distribution of $x_{0,d}$ is so.

Explosive solutions: without further assumptions the solutions to the limiting SDEs can be explosive. From (11) we see that the potentially troublesome term is the variance matrix in the drift ((14) makes the issue easier to see in a more restricted setting). Assumption 2.3 is satisfied under all considered parameter distribution assumptions 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 $\phi = \tanh$.

Non-smooth activations: the diffusion limits are based on a sufficiently smooth activation ϕ per Assumption 3.2. We consider here the following case which includes the ReLU activation. If $\phi(a)$ 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 with (1) and (2), we see that unless $\mathbb{E}[\phi(h)] = 0$, choosing $\gamma = 1/2$ would result in the drift term blowing up. Choosing $\gamma = 1$ recovers a deterministic limit as in Chen et al. (2018).

3.5 Input and output layers

So far we have considered $\boldsymbol{x}_0 \in \mathbb{R}^D$ to be the input of the ResNet. A NN acts as a function approximator to be fitted to some dataset $\{(z^{(i)}, y^{(i)})\}_{i=1}^N$ where $z^{(i)} \in \mathbb{R}^Z$ represents an input and $y^{(i)} \in \mathbb{R}^Y$ represents the corresponding output. 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 NN prediction for $z^{(i)}$. As for $\boldsymbol{x}_t$, we will denote a single data-point $(z^{(i)}, y^{(i)})$ with (z, y) when no confusion arises.

4 Experiments

4.1 Sanity check

First of all we investigate numerically the correctness of the results obtained in Section 3.3. We consider the setting of Assumption 3.4 with $\phi = \tanh$, $\sigma_w^2 = \sigma_b^2 = 1$, $T = 1$, $L = D = 500$ and 1-dimensional inputs. In all the experiments ψ is set to the identity function. As noted in Section 3.5 we need to introduce an input layer mapping $z \in \mathbb{R} \mapsto \boldsymbol{x}_0 \in \mathbb{R}^D$. For this toy example we simply copy the input across all dimensions: $\boldsymbol{x}_{0,\bullet} = z$, i.e. $\boldsymbol{x}_{0,d} = z$ for each $d \in D$. We refer to this model as $\mathcal{SC}_{\tanh}$. We consider two 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 i) $\boldsymbol{x}_T^{(1)}$, $\boldsymbol{x}_T^{(2)}$ via the ResNet recursion (7); ii) $x_T^{(1)}$, $x_T^{(2)}$ via the discretization (6) of the limiting SDE (17). Our analysis imply that i) and ii) are equivalent in the limit $L \uparrow \infty$. We report the results in Figure 2 where good agreement is indeed observed. We replicate this experiment in SM B for $\mathcal{SC}_{\text{swish}}$, where the tanh activation in $\mathcal{SC}_{\tanh}$ is replaced by the swish activation ($\text{swish}(x) = x\,\text{sigmoid}(x)$) which has been shown empirically (Ramachandran et al., 2017) and theoreti-

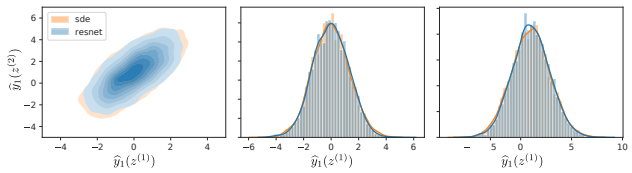

Figure 2: For model $\mathcal{SC}_{\tanh}$: 2D KDE plot for $(\widehat{y}_1(z^{(1)}), \widehat{y}_1(z^{(2)}))$ (left), 1D KDE and histogram plots for $\widehat{y}_1(z^{(1)})$ (center), $\widehat{y}_1(z^{(2)})$ (right) when $\widehat{y}_1$ is sampled from a ResNet and from the Euler discretization of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

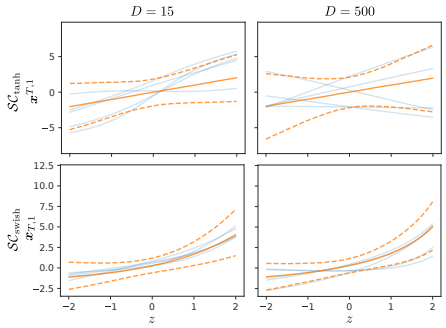

Figure 3: Function samples of $\boldsymbol{x}_{T,1}$ for $\mathcal{SC}_{\tanh}$ (top) and SC_{swish} (bottom), see Figure 1 for the description of the plotted quantities.

cally (Hayou et al., 2019a) to be competitive. In this case $\phi'(0) = \phi''(0) = 1/2$ and Assumption 2.3 is not satisfied.

4.2 Function space distributions

We show empirically that the dependency on the input is retained and the output distribution does not exhibit perfect correlation for very deep ResNet constructed as in the present paper. We consider the same model $\mathcal{SC}_{\tanh}$ of Section 4.1. First of all, from the center and right plots of Figure 2 we see that $\boldsymbol{x}_{T,1}^{(1)}$ and $\boldsymbol{x}_{T,1}^{(2)}$ are differently distributed, meaning the input dependency is retained, and from the left plot we see that they are not perfectly correlated, otherwise the 2D KDE would collapse to a straight line.

In Figure 3 (top) we visualize samples of $\boldsymbol{x}_{T,1}$ from $\mathcal{SC}_{\tanh}$ in function space for different combinations of L (more plots in SM B). More specifically, we approximate function draws by considering 400 inputs $z^{(i)}$ equally spaced on $[-2, 2]$. Using the ResNet recursion (7) we obtain 400 output values $\boldsymbol{x}_{T,1}^{(i)}$. We repeat this procedure to obtain 10.000 function draws. In Figure 3 (bottom) we repeat this experiment for $\mathcal{SC}_{\text{swish}}$. In this specific case we did not observe divergent trajectories for the 10.000 function draws. In Figure 3 we observe

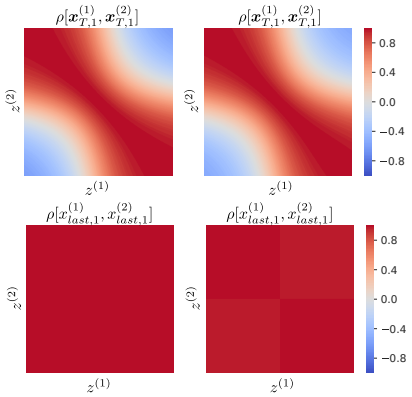

Figure 4: Output correlation heatmap for $\mathcal{SC}_{\tanh}$ (top-left), $\mathcal{SC}_{\text{swish}}$ (top-right), $\mathcal{EO}_{\tanh}$ (bottom-left), $\mathcal{EO}_{\text{ReLU}}$ (bottom-right).

similar distribution properties across different orders of magnitude for D, which suggests the existence of a stochastic limit in the doubly infinite setting where $L, D \uparrow \infty$.

In Figure 4 (top) we plot the correlations $\rho[\boldsymbol{x}_{T,1}^{(1)}, \boldsymbol{x}_{T,1}^{(2)}]$ for inputs $(z^{(1)}, z^{(2)})$ in the range $[-2, 2] \times [-2, 2]$ for the tanh and swish activations: for different inputs the output correlations are far from 1. Let us refer to the model of Figure 1 with tanh activation as $\mathcal{EO}_{\tanh}$, and to the model of Figure 1 with ReLU activation as $\mathcal{EO}_{\text{ReLU}}$. For comparison, we show in Figure 4 (bottom) the correlations $\rho[x_{last,1}^{(1)}, x_{last,1}^{(2)}]$ for pre-activation 1 for $\mathcal{EO}_{\tanh}$ and $\mathcal{EO}_{\text{ReLU}}$: all correlations are close to 1.

4.3 SGD training

In this experiment we consider the MNIST dataset (LeCun, 1998). Each observation (z, y) is composed of an image $z \in \mathbb{R}^{784}$ (we flatten to a vector) and a class $y \in \mathbb{R}^{10}$ (we use 1-hot encoding). We consider the setting of Assumption 3.4 with $\phi = \tanh$, $\sigma_w^2 = \sigma_b^2 = 1$, $T = 1$ and random input and output layers given by $\boldsymbol{x}_0 = W_I z, \widehat{y} = W_O \boldsymbol{x}_T$ where $W_I \in \mathbb{R}^{D \times 784}, W_O \in \mathbb{R}^{10 \times D}$ and $W_{I,d,i}, W_{O,c,o} \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$. We use the cross-entropy loss function and fit the model to the training dataset via SGD. Figure 5 (top) shows the evolution of the training losses over 1 epoch (mini-batches of 200 samples) when the gradients are taken with respect to $\{\varepsilon_t^W, \varepsilon_t^b\}_{t=0}^{T-\Delta t}$ ((16), reparametrized gradients) for a common learning rate. This choice results in stable loss decrease over all considered values for L and D. Moreover all average accuracies computed on the test dataset after 1 training epoch are in the range $[87.1\%, 90.6\%]$. In contrast, we were unable to obtain a test accuracy uniformly above 72.4% with a common (tuned via grid-search) learning rate when the gradients

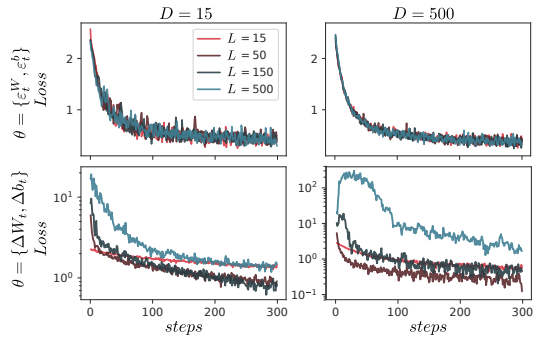

Figure 5: Averaged (over each batch) loss on MNIST training dataset for the model of Section 4.3, different L, D, for reparametrized gradients (top, shared linear-scale on y-axes) and for standard gradients (bottom, different log-scales on y-axes).

are computed with respect to $\{\Delta W_t, \Delta b_t\}_{t=0}^{T-\Delta t}$ ((15), standard gradients). Figure 5 (bottom) illustrates the issue: a common learning rate leads to either slow or divergent trajectories. Similar results (not shown) are obtained for commonly used initializations (Glorot and Bengio (2010); He et al. (2015)). Our experiment suggests the existence of results akin to Jacot et al. (2018); Hayou et al. (2019b) as both $L, D \uparrow \infty$.

Zhang et al. (2019) considers initializations for ResNets which are not encompassed yet by our analysis. Conversely, the residual blocks in Zhang et al. (2019) cannot be shallow. An analysis of gradient properties motivates initializing the residual block parameters so that the their variance shrinks as the ResNet gets deeper. However, the residual blocks are multiplied by parameters initialized at 0, hence our desiderata iii) (Section 1) is not satisfied. Moreover the gradients are not reparametrized as in the above experiment.

5 Discussion

We have established the convergence of identity ResNets He et al. (2016b) to solutions of SDEs as the number of layers goes to infinity. Our results rely on smooth activation functions and on model parameter distributions which shrink as total depth increases. Further conditions on the activation functions are obtained by restricting the limiting SDEs to be non explosive. As the infinitesimal evolution of SDEs is characterized by their instantaneous mean and covariance, it seemed natural to assume that model's parameters have Gaussian distributions. However, our results can be strengthened to hold for finite-variance parameter distributions.

Building on the connection between IDNN and diffusion processes we showed that, as the number of layers goes to infinity: the last layer does not collapse to a deterministic limit, nor does it diverge to infinity; the dependency of the last layer on the input does not vanish; the last layer, as stochastic function on input space, remains flexible without collapsing to restrictive families of distributions. We then investigated additional properties of the limiting diffusions. In contrast to the information propagation approach our analysis covers finitely-wide NNs and correlated parameters at the layer level.

While the limiting diffusions do not suffer from catastrophic limitations, to obtain competitive performance more attention needs to be paid to architectural choices, to parameters' distribution selection, and to input and output layers. Moreover, results on forward propagation do not trivially translate to corresponding results on gradient back-propagation. With this in mind, hereafter we list some promising future research directions. Firstly, we can consider more realistic residual blocks consisting of multiple convolutional layers as in Zhang et al. (2019). Extending the present work to convolutional NN does not require new theoretical developments as a convolutional transform (jointly over all positions) can be expressed via matrix multiplication. Deep residual blocks could be approached via fractional Brownian motions (Biagini et al., 2008) or via re-scaled Brownian motions. Secondly, the same techniques used to derive the evolution of IDNNs can be used to obtain the evolution of the input-output Jacobian. This would pave the way to an extensions of the neural tangent kernel (Jacot et al., 2018; Lee et al., 2019; Arora et al., 2019; Hayou et al., 2019b) to IDNNs. Thirdly, stable behavior has been observed with an appropriate scaling of the weight parameters as the wideness D increases. In particular, it would be instructive to characterize the distribution of NNs which are both infinitely deep and wide. This result could form the basis of Bayesian inference (Lee et al., 2018; Garriga-Alonso et al., 2019) for doubly infinite NNs and of data-dependent initializations.

6 Acknowledgements

We wish to thank the three anonymous reviewers and the meta reviewer for their valuable feedback. The authors acknowledge Thierry Sousbie and Tiago Ramalho for the many suggestions that greatly improved the presentation of the current work. 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

Arora, S., Du, S. S., Hu, W., Li, Z., Salakhutdinov, R., and Wang, R. (2019). On exact computation with an infinitely wide neural net. In *Advances in Neural Information Processing Systems 32*.

Biagini, F., Hu, Y., Øksendal, B., and Zhang, T. (2008). *Stochastic calculus for fractional Brownian motion and applications*. Springer Science & Business Media.

Billingsley, P. (1999). *Convergence of Probability Measures*. Wiley-Interscience, 2nd edition.

Chen, T. Q., Rubanova, Y., Bettencourt, J., and Duvenaud, D. K. (2018). Neural ordinary differential equations. In *Advances in Neural Information Processing Systems 31*, pages 6571–6583.

Garriga-Alonso, A., Rasmussen, C. E., and Aitchison, L. (2019). Deep convolutional networks as shallow gaussian processes. In *International Conference on Learning Representations*.

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

Gupta, A. K. and Nagar, D. K. (1999). *Matrix variate distributions*. Chapman and Hall/CRC, 1st edition.

Hayou, S., Doucet, A., and Rousseau, J. (2019a). 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.

Hayou, S., Doucet, A., and Rousseau, J. (2019b). Training dynamics of deep networks using stochastic gradient descent via neural tangent kernel. *arXiv preprint arXiv:1905.13654*.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034.

He, K., Zhang, X., Ren, S., and Sun, J. (2016a). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

He, K., Zhang, X., Ren, S., and Sun, J. (2016b). Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer.

Jacot, A., Gabriel, F., and Hongler, C. (2018). Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems 31*, pages 8571–8580.

Kloeden, P. E. and Platen, E. (1992). *Numerical Solution of Stochastic Differential Equations*. Springer, corrected edition.

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

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436.

Lee, J., Sohl-dickstein, J., Pennington, J., Novak, R., Schoenholz, S., and Bahri, Y. (2018). Deep neural networks as gaussian processes. In *International Conference on Learning Representations*.

Lee, J., Xiao, L., Schoenholz, S. S., Bahri, Y., Sohl-Dickstein, J., and Pennington, J. (2019). Wide neural networks of any depth evolve as linear models under gradient descent. In *Advances in Neural Information Processing Systems 32*.

Matthews, A. G. d. G., Rowland, M., Hron, J., Turner, R. E., and Ghahramani, Z. (2018). Gaussian process behaviour in wide deep neural networks. *arXiv preprint arXiv:1804.11271*.

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

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

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

Poole, B., Lahiri, S., Raghu, M., Sohl-Dickstein, J., and Ganguli, S. (2016). Exponential expressivity in deep neural networks through transient chaos. In *Advances in Neural Information Processing Systems 29*, pages 3360–3368.

Ramachandran, P., Zoph, B., and Le, Q. V. (2017). Searching for activation functions. *arXiv preprint arXiv:1710.05941*.

Revuz, D. and Yor, M. (1999). *Continuous Martingales and Brownian Motion*. Springer, 3rd edition.

Schoenholz, S. S., Gilmer, J., Ganguli, S., and Sohl-Dickstein, J. (2017). Deep information propagation. In *International Conference on Learning Representations*.

Stroock, D. W. and Varadhan, S. S. (2006). *Multidimensional diffusion processes*. Springer, 2006 edition.

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

Zhang, H., Dauphin, Y. N., and Ma, T. (2019). Residual learning without normalization via better initialization. In *International Conference on Learning Representations*.