A Proofs

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

Proof of Theorem 2.1. This is (Nelson, 1990, Theorem 2.2): Assumption 2.1 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 A.1. *If ϕ satisfies Assumption 3.2, $\epsilon \sim \mathcal{N}(0, \sigma^2)$ 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:*

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

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 of Theorem 3.1. 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 $\boldsymbol{x}_t$ 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 $\forall R > 0$

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

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

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

It remains to show that

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

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

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

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 A.1 the first sup is finite too and not increasing in Δt as $|\vartheta_d| \leq \sqrt{\Delta t}|h_d|$ 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 $d = 1, \ldots, D$ 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 $k h_d^n \phi''(\vartheta_d)^m \Delta t^\alpha$ for integers $n, m \geq 0$ and reals $\alpha > 0$, $k \in \mathbb{R}$. This results from repeated applications of the Cauchy–Schwarz inequality and Lemma A.1 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 $d, u = 1, \ldots, D$ and $R > 0$:

$$\lim_{\Delta t \downarrow 0} \sup_{\|x\|<R} \left| \sigma_x^2(x)_{d,u} - \mathbb{E}\left[\frac{\left(\phi'(0) h_d \Delta t^{1/2} + \frac{1}{2}\phi''(\vartheta_d) h_d^2 \Delta t\right)\left(\phi'(0) h_u \Delta t^{1/2} + \frac{1}{2}\phi''(\vartheta_u) h_u^2 \Delta t\right)}{\Delta t} \right] \right| = 0$$

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

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

The (uniform on compacts) convergence of all terms aside from $\sigma_x^2(x)_{d,u}$ to 0 once again follows from repeated applications of the Cauchy–Schwarz inequality and Lemma A.1.

Now, the continuity of $\mu_x(x)$ and $\sigma_x(x)$ 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 $\sigma_x^2(x)$. Hence all the conditions of Assumption 2.1 hold true.

Finally, as ψ is differentiable two times with continuity, it follows from the dependency of μ_x and σ_x^2 on x only through $\mathbb{V}[\varepsilon^W \psi(x) + \varepsilon^b]$ that Assumption 2.2 is satisfied too. The application of Theorem 2.1 completes the proof. □

Proof of Corollary 3.1. 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 (12) shows that the drift terms are matched between (11) and (12). The quadratic variation of (11) 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 (12) 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 (11) and the solution of (12). Then (13) immediately follows by direct computation. □

Proof of Corollary 3.2 and Corollary 3.3. Notice that

$$\begin{aligned}
&d[W\psi(x^{(i)})+b, W\psi(x^{(j)})+b]_t \\
&\quad= \mathbb{C}[\varepsilon^W_t\psi(x^{(i)}_t)+\varepsilon^b_t, \varepsilon^W_t\psi(x^{(j)}_t)+\varepsilon^b_t | x^{(i)}_t, x^{(j)}_t]dt \\
&\quad= \big(\Sigma^b + \mathbb{C}[\varepsilon^W_t\psi(x^{(i)}_t), \varepsilon^W_t\psi(x^{(j)}_t) | x^{(i)}_t, x^{(j)}_t]\big)dt
\end{aligned}$$

and

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

This proves Corollary 3.2. Corollary 3.3 follows by setting $\sigma^b = \sigma_b\,\mathrm{I}_D$, $\sigma^{W_I} = \mathrm{I}_D$ and $\sigma^{W_O} = \sigma_w D^{-1/2}\,\mathrm{I}_D$. □

B Additional experiments and plots

B.1 Bayesian inference

In this toy experiment we perform approximate Bayesian inference via Approximate Bayesian Computation (ABC, (Sisson et al., 2018)) rejection sampling for function regression. We consider the setting of Assumption 3.4 with $\phi = \tanh$, $\sigma^2_w = \sigma^2_b = 10$, $T = 1$ and $L = D = 500$. For this experiment we use a random input layer given by $\boldsymbol{x}_0 = W_I z$ where $W_I \in \mathbb{R}^{D\times 1}$ and $W_{I,d,1} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ which makes the distribution of $\boldsymbol{x}_{T,1}$ symmetric around 0. We refer to this model as $\mathcal{SR}_{\text{tanh}}$. We fix a computational budget of 10.000 simulations and compute 10 approximate posteriors samples by selecting the 10 function draws with smallest l_2 distance from a synthetic dataset $\mathcal{D}$ consisting of 3 data points. The results are reported in Figure 1 where we also compare with the results obtained by applying the same ABC algorithm to the first pre-activation of the last layer of $\mathcal{EO}_{\text{tanh}}$. We observe that the more flexible prior results in significantly improved efficiency with ABC: as the prior draws are almost constant in $\mathcal{EO}_{\text{tanh}}$ we are constrained to finding the line with minimum distance from the 3 points. The use of $\sigma^2_w = \sigma^2_b = 10$ in $\mathcal{SR}_{\text{tanh}}$ compared to $\mathcal{SC}_{\text{tanh}}$ allows for increased range and variability of $\boldsymbol{x}_{T,1}$. While it's possible to similarly increase weight and bias variances in $\mathcal{EO}_{\text{tanh}}$ while remaining on the edge of chaos, this does not solve the underlying issue that the model a priori (hence a posteriori) assigns all probability mass to constant functions in the limit of $L \uparrow \infty$. Simulations (not shown) confirms that this modification does not improve the posterior inference efficiency for $\mathcal{EO}_{\text{tanh}}$. It should be noted that we do not advocate the use of ABC rejection sampling as a realistic solution for this inference setting. Nonetheless this toy experiment exemplifies how a prior-data conflict typically frustrates inference algorithms.

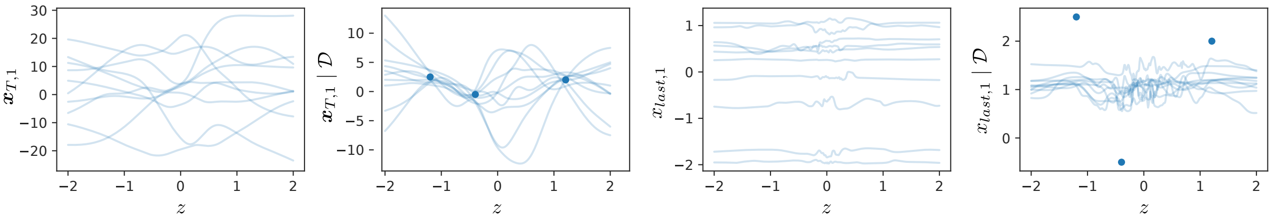

Figure 1: 10 function samples in light blue over $z \in [-2, 2]$ for: prior $\boldsymbol{x}_{T,1}$ and ABC-posterior $\boldsymbol{x}_{T,1}|\mathcal{D}$ for $\mathcal{SR}_{\text{tanh}}$ (2 leftmost); prior $x_{last,1}$ and ABC-posterior $x_{last,1}|\mathcal{D}$ for $\mathcal{EO}_{\text{tanh}}$ (2 rightmost); 3 data points of $\mathcal{D}$ in blue.

B.2 Additional experiment for Section 4.1

We replicate the experiment of Section 4.1 for the swish activation and plot the results in Figure 2 where again good agreement is observed.

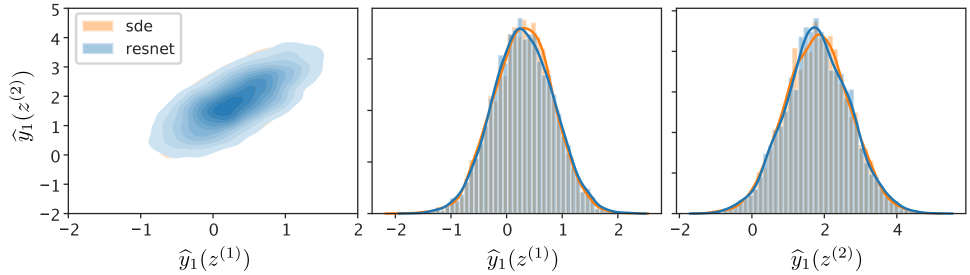

Figure 2: For model $\mathcal{SC}_{\text{swish}}$: 2D KDE plot for $(\widehat{y}_1(z^{(1)}), \widehat{y}_1(z^{(2)}))$ (left), 1D KDE and histogram plots for $\widehat{y}_1(z^{(1)})$ (center), $\widehat{y}_1(z^{(2)})$ (right) when $\widehat{y}_1$ is sampled from a ResNet (resnet) and from the Euler discretization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

B.3 Additional plots for Section 4.2

In Figure 3 we plot additional function samples for the models $\mathcal{SC}_{\text{tanh}}$ and SC_{swish} of Section 4.2 corresponding to different combinations of L and D. We observe similar dynamics across different orders of magnitude for both L and D.

B.4 Additional 2D plots

In Figure 4 we plot 2D function samples of $\boldsymbol{x}_{T,1}$ for $\mathcal{SC}_{\text{tanh}}$ and SC_{swish} to complement the visualizations of Section 4.2.

C Related work

In this section we discuss more in detail further connections with related work. Chen et al. (2018) investigates the connection between infinite ResNets and ordinary differential equations (ODE), with a focus on potential computational advantages from gradient-descent training perspective. A difference between our work and that of Chen et al. (2018) is that we only operate on the distribution of the model parameters, while Chen et al. (2018) modifies the ResNet recursion with a Δt multiplicative term. The two approaches are equivalent for ReLU activations (see Section 3.4), where the SDE limit collapses to a deterministic ODE. While the focus of Chen et al. (2018) is on the training of neural networks in the present work the emphasis is on the priori distribution of neural networks induced by a class of priors on the model parameters.

Pennington et al. (2017, 2018) investigates the properties of the spectrum of the input-output Jacobian of deep neural networks at initialization which affects gradient-descent training speed. Of particular interest to the setting of the present work are the orthogonal initializations, proposed in Pennington et al. (2017) to achieve dynamical isometry, which seem amenable to a modification of the approach presented in our paper.

While Pennington et al. (2017, 2018) relies on mean-field assumptions, i.e. infinite width and iid finite variance initializations, Burkholz and Dubatovka (2019) consider the setting of finite width and ReLU activations to derive exact results. Unfortunately, due to the use of ReLU activations the proposed initialization scheme is not of particular interest to our setting (see our discussion above regarding Chen et al. (2018)).

Hanin and Rolnick (2018); Hanin (2018) study ReLU networks when the width and depth are comparable and both large. As hinted in Section 4.2 and in the Discussion it would be interesting to investigate the behavior of the limiting SDEs when both depth and width grow unbounded. In this setting it is possible for the order of the limits and for the relative speed of convergence of L and D to affect the limiting dynamics.

Finally there has been recent interest in using heavy-tailed distributions for gradient noise and for trained parameter distributions, see for instance Simsekli et al. (2019); Martin and Mahoney (2019). The present

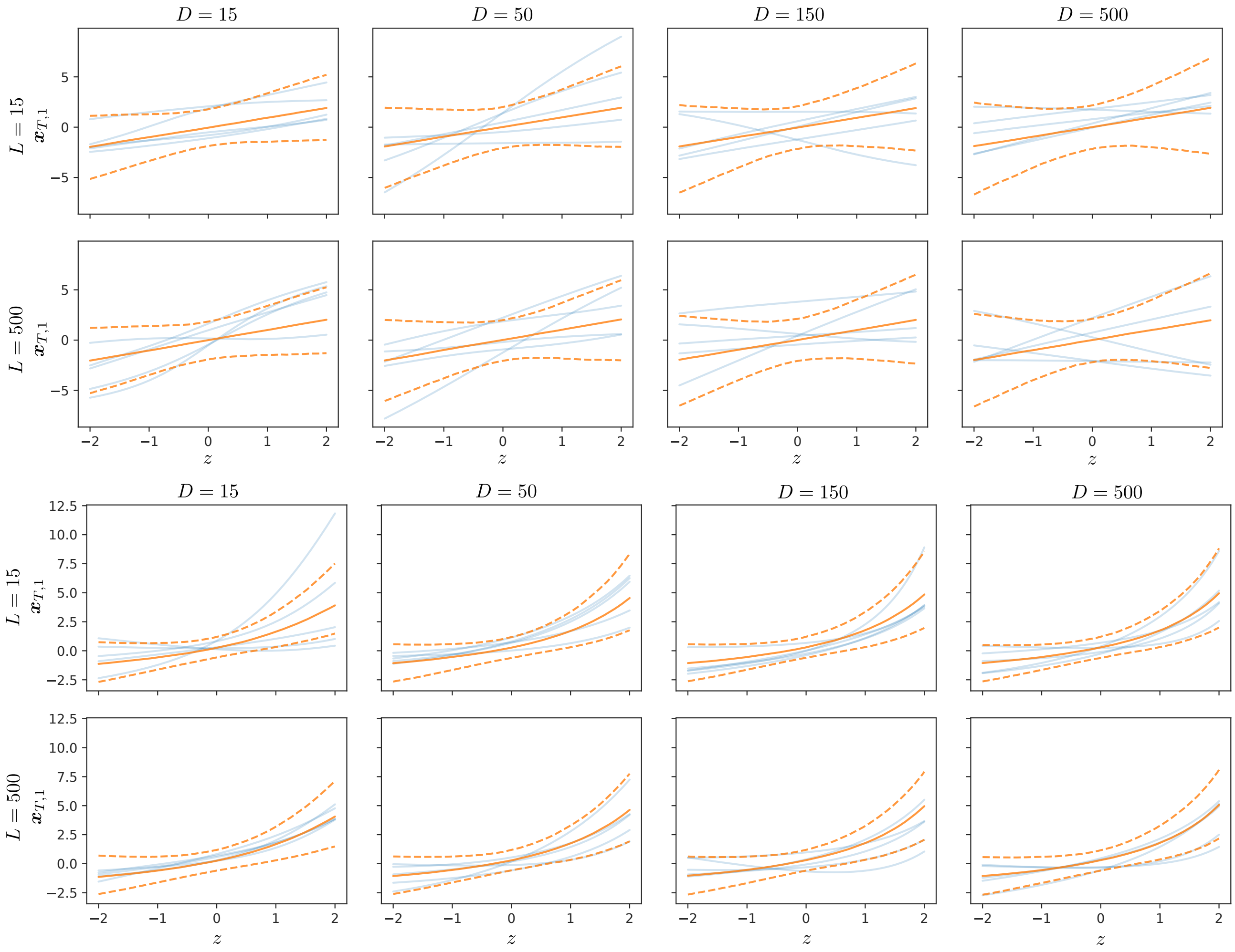

Figure 3: Function samples of $\boldsymbol{x}_{T,1}$ for $\mathcal{SC}_{\text{tanh}}$ (top) and SC_{swish} (bottom) for different values of L and D.

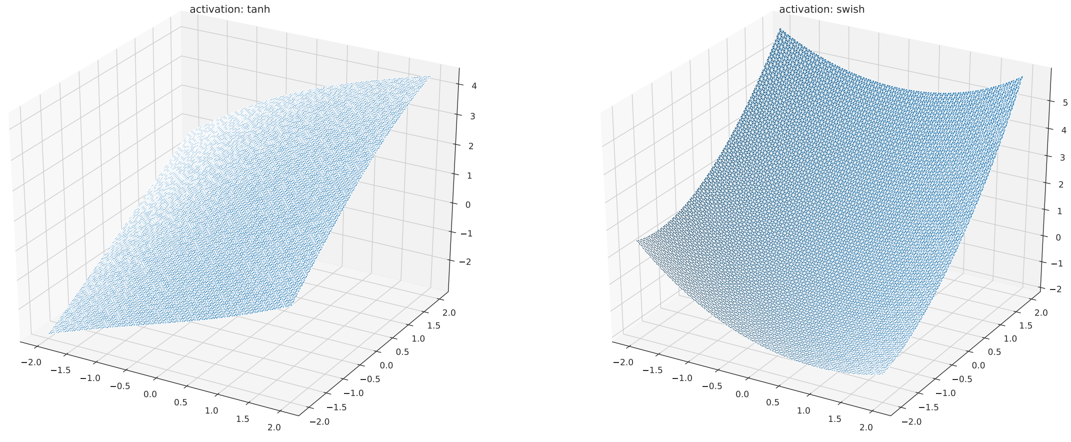

Figure 4: Function samples of $\boldsymbol{x}_{T,1}$ for $\mathcal{SC}_{\text{tanh}}$ (left) and SC_{swish} (right) for $L = 100$ and $D = 100$ on the bounded interval $[-2, 2] \times [-2, 2]$.

work covers exclusively Gaussian initializations and could be extended with some effort to cover finite-variance ones. Extensions to heavy tailed distributions would me more involved and likely resulting in less tractable Semimartingales (instead of SDEs) due to the presence of finite and infinite activity jump components. On the other hand the added flexibility could prove beneficial in bridging the gap between the performance of finitely trained neural networks and their limiting stochastic processes counterparts at least in some settings.

References

Burkholz, R. and Dubatovka, A. (2019). Initialization of relus for dynamical isometry. In *Advances in Neural Information Processing Systems*, pages 2382–2392.

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.

Ethier, S. N. and Kurtz, T. G. (2009). *Markov processes: characterization and convergence*. Wiley-Interscience.

Hanin, B. (2018). Which neural net architectures give rise to exploding and vanishing gradients? In *Advances in Neural Information Processing Systems*, pages 582–591.

Hanin, B. and Rolnick, D. (2018). How to start training: The effect of initialization and architecture. In *Advances in Neural Information Processing Systems*, pages 571–581.

Martin, C. H. and Mahoney, M. W. (2019). Traditional and heavy-tailed self regularization in neural network models. *arXiv preprint arXiv:1901.08276*.

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.

Pennington, J., Schoenholz, S., and Ganguli, S. (2017). Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice. In *Advances in neural information processing systems*, pages 4785–4795.

Pennington, J., Schoenholz, S. S., and Ganguli, S. (2018). The emergence of spectral universality in deep networks. *arXiv preprint arXiv:1802.09979*.

Simsekli, U., Sagun, L., and Gurbuzbalaban, M. (2019). A tail-index analysis of stochastic gradient noise in deep neural networks. *arXiv preprint arXiv:1901.06053*.

Sisson, S. A., Fan, Y., and Beaumont, M. (2018). *Handbook of approximate Bayesian computation*. Chapman and Hall/CRC.

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