Neural signature kernels as infinite-width-depth-limits of controlled ResNets

Nicola Muça Cirone [1]   **Maud Lemercier** [2]   **Cristopher Salvi** [1]

Abstract

Motivated by the paradigm of reservoir computing, we consider randomly initialized *controlled ResNets* defined as Euler-discretizations of *neural controlled differential equations* (Neural CDEs), a unified architecture which enconpasses both RNNs and ResNets. We show that in the infinite-width-depth limit and under proper scaling, these architectures converge weakly to Gaussian processes indexed on some spaces of continuous paths and with kernels satisfying certain partial differential equations (PDEs) varying according to the choice of activation function φ, extending the results of Hayou (2022); Hayou & Yang (2023) to the controlled and homogeneous case. In the special, homogeneous, case where φ is the identity, we show that the equation reduces to a linear PDE and the limiting kernel agrees with the *signature kernel* of Salvi et al. (2021a). We name this new family of limiting kernels *neural signature kernels*. Finally, we show that in the infinite-depth regime, finite-width controlled ResNets converge in distribution to Neural CDEs with random vector fields which, depending on whether the weights are shared across layers, are either time-independent and Gaussian or behave like a matrix-valued Brownian motion.

1. Introduction

The symbiosis between differential equations and deep learning has become an active research area in recent years, notably through the introduction of hybrid models named *neural differential equations* (Kidger, 2022). In fact, many standard neural network architectures may be interpreted as approximations to some differential equations.

This approximation has been treated rigorously for finite-width ResNets, which in the infinite-depth limit converge in distribution to zero-drift *neural stochastic differential equations* (Neural SDEs) with diffusion depending on the choice of activation function (Cohen et al., 2021; Hayou, 2022; Marion et al., 2022; Cont et al., 2022).

The "dual" scenario of finite-depth and infinite-width neural networks has also been the object of many recent studies (Neal, 2012; Matthews et al., 2018; Novak et al., 2018). Notably, through the unifying algorithmic language of *Tensor Programs* designed by Yang (2019), many standard feedforward, convolutional and recurrent architectures of finite-depth can be shown to converge to Gaussian processes (GPs) in the infinite-width limit.

In the context of deep learning for sequential data, finite-width RNNs have been informally identified as approximations to *neural controlled differential equations* (Neural CDEs) introduced by Kidger et al. (2020); Morrill et al. (2021) and inspired from the homonymous class of dynamical systems studied in *rough analysis*, a branch of stochastic analysis providing a robust solution theory for differential equations driven by irregular signals (Lyons, 1998; Lyons et al., 2007; Friz & Hairer, 2020; Friz & Victoir, 2010).

However, contrarily to this widespread interpretation, the Euler discretization of a Neural CDE with vector fields[1] f produces a recursive relation for the hidden state h in the form of (1), where the increments of the input signal x enter the recursion in a multiplicative manner rather than via an additive interaction typically assumed in RNNs:

$$h_{k+1} = h_k + f(h_k)(x_{k+1} - x_k). \qquad (1)$$

Furthermore, the addition of the previous hidden state h_k on the right-hand side of (1), commonly referred to as a skip connection, is characteristic of ResNets and absent in classical RNNs. We will refer to architectures defined by (1) as *homogeneous controlled ResNets*. We will also consider their *inhomogenous* counterparts where the map $f = f(k, h_k)$ depends on the iteration k.

Dynamical systems in the form of (1) are often called *reservoirs* in the paradigm of *reservoir computing* (Tanaka et al.,

---

[1]Department of Mathematics, Imperial College London, London, United Kingdom [2]Department of Mathematics, University of Oxford, Oxford, United Kingdom. Correspondence to: Nicola Muça Cirone <nm2322@ic.ac.uk>.

[1]Typically f is taken to be a randomly initialized feedforward neural network with Gaussian weights and biases.

2019; Lukoševičius & Jaeger, 2009; Verstraeten et al., 2007). Contrarily to deep learning, in reservoir computing, only the final readout linear map is trained, while the function f is randomly sampled but remains untrained.

It is worth noting that Neural CDEs are deep learning models that map between infinite dimensional spaces of continuous paths. Therefore, if these continuous models converge, in the infinite-width limit, to some limiting GPs, the latter should be equipped with kernel functions indexed on the same spaces of continuous paths. In the sequel we will demonstrate that controlled ResNet indeed behave like such GPs in the large width-depth regime.

1.1. Contributions

Our objective here is to provide a rigorous mathematical analysis of the behavior of controlled ResNets that are randomly initialised with Gaussian weights and biases in the large width and depth regimes. More specifically:

- We prove that both in the infinite-width-depth limit these architectures converge weakly to GPs with limiting kernels satisfying certain (possibly non-linear) partial differential equations varying according to the (in)homogeneity of the network and to the choice of activation function φ (see Table 1). Moreover, we show that under some further conditions on the regularity of the driving paths the limits commute, i.e. the limiting GP is unchanged upon reversing the order of the limits. We name this new class of kernel *neural signature kernels*.

- In the case where the system is homogeneous and φ is the identity, we show that the equation reduces to a linear PDE and the limiting kernel is proportional to the *signature kernel* introduced in (Salvi et al., 2021a).

- We then prove that in the infinite-depth regime, finite-width controlled ResNets converge in distribution to Neural CDEs with random vector fields. In the inhomogeneous case, these fields behave as a matrix-valued Brownian motion, while for homogeneous networks they are time-independent and Gaussian.

1.2. Notation

Since we are mainly interested in studying multivariate time-series, our data space will be a space of continuous paths on the interval $[0, 1]^2$ and with values in $\mathbb{R}^d$, for some $d \in \mathbb{N}$. More specifically, we consider the space

$$\mathbb{X} := \{x \in C^0([0, 1]; \mathbb{R}^d) : x(0) = 0, \exists \dot{x} \in L^2([0, 1]; \mathbb{R}^d)\}$$

[2]The choice of the interval $[0, 1]$ is not at all restrictive and has been made to ease the notation.

of continuous paths with a square integrable derivative.

We denote by $x^j(t) \in \mathbb{R}$ the jth coordinate of a path $x \in \mathbb{X}$ for $j \in \{1, ..., d\}$. Given n paths $\mathcal{X} = \{x_1, \ldots, x_n\} \subset \mathbb{X}$ and functions $f : \mathbb{X} \to \mathbb{R}$, $G : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ we will write $f(\mathcal{X}) \in \mathbb{R}^n$ for the vector $[f(\mathcal{X})]_\alpha = f(x_\alpha)$ and $G(\mathcal{X}, \mathcal{X}) \in \mathbb{R}^{n \times n}$ for the matrix $[G(\mathcal{X}, \mathcal{X})]_\alpha^\beta = G(x_\alpha, x_\beta)$ for any $\alpha, \beta \in \{1, .., n\}$.

We will consider partitions $\mathcal{D} = \{0 = t_0 < \cdots < t_M = 1\}$ of the interval $[0, 1]$ and write their length as $\|\mathcal{D}\| := M$ and their mesh size as $|\mathcal{D}| := \max_{i=1,\ldots,\|\mathcal{D}\|} |t_i - t_{i-1}|$.

For an activation function $\varphi : \mathbb{R} \to \mathbb{R}$ and a positive semidefinite matrix $\Sigma \in \mathbb{R}^{2 \times 2}$, in the paper we will repeatedly make use of the following function

$$V_\varphi(\Sigma) = \mathbb{E}_{\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \Sigma)}[\varphi(z_1)\varphi(z_2)].$$

The explicit form of V_φ changes significantly depending on the activation function. We list it for a restricted class of them in Proposition A.7 in the appendix.

Henceforth, we fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

The paper is organized as follows: in Section 2 we discuss some related work, in Section 3 we study the *inhomogeneous* version and in Section 4 we analyse the *homogeneous* version of controlled ResNets. We conclude in Section 5 with numerical results validating our claims. All proofs can be found in the appendix.

2. Related Work

Results relating infinite-width limits of neural networks to GPs have been extended from shallow networks (Neal, 2012) to richer architectures of feedforward (Lee et al., 2018; Matthews et al., 2018), convolutional (Novak et al., 2018; Garriga-Alonso et al., 2018) and recurrent (Alemohammad et al., 2020) type. This line of work culminated with the framework of *Tensor Programs* formulated in Yang (2019) which offers an algorithmic procedure to systematically compute the limiting GP kernels for a wide range of different architectures. One major advantage of this formalism is that it makes it possible to consider weight sharing between layers, something mostly avoided in previous literature but central in the types of systems we consider here.

The reverse scenario of infinite-depth limit of finite-width architectures has mainly been explored for ResNets. In this regime, appropriately rescaled ResNets have been shown to behave like stochastic differential equations (SDEs) (Chen et al., 2018; Cohen et al., 2021; Marion et al., 2022). In particular, Hayou (2022) considers the simpler setting where the architecture does not exhibit weight sharing (we refer to this setting as inhomogeneous) and single out limiting kernels of exponential type. In what follows, we will show

that the exponential nature of the limiting kernels is somewhat kept intact even when the ResNets architectures are controlled by an external stream of information, but that the limiting kernels have more structure, particularly in the homogeneous case.

It is natural to investigate the behavior of neural networks when both the width and the depth are very large. The literature on the topic has particularly developed around the infinite-width-then-depth limit, most notably with the derivation of the Edge of Chaos (Poole et al., 2016; Schoenholz et al., 2017) which has shown how feedforward networks suffer from a kind of information dispersal which limits the propagation of the input signals through their depth. Noteworthy are also the results in Li et al. (2021), where both limits are taken together for some fixed depth-to-width ratio and non-Gaussian behaviour at the limit of a specific kind of ResNets is discovered, and in Li et al. (2022) where, in the feedforward setting, stochastic dynamics are given for the covariance between output layers in the same regime. In the existing literature, the study most closely aligned with our work is (Hayou & Yang, 2023) where the commutativity of the limits is shown for residual Networks corresponding to the simplest example of controlled ResNets: *inhomogeneous* ones, driven by linear controls. [3]

As anticipated in the introduction, ResNets that are controlled by sequential data streams are generalised forms of RNNs and correspond to Euler discretizations of Neural CDEs and variants (Kidger et al., 2020; Morrill et al., 2021; Salvi et al., 2022; Fermanian et al., 2021). These models offer a memory-efficient way to model functions of potentially irregular signals in continuous-time and have achieved state-of-the art performance on a wide range of time series tasks (Singh et al., 2022; Bellot & Van Der Schaar, 2021; Morrill et al., 2021). They stem from the well-understood mathematics of *controlled differential equations*, which are the central objects studied in rough analysis.

Rough path theory introduced by Lyons (1998) is a modern mathematical framework focused on making precise the interactions between highly oscillatory signals and non-linear dynamical systems. The theory provides a deterministic toolbox to recover many classical results in stochastic analysis without resorting to specific probabilistic arguments. Notably, it extends Itô's theory of SDEs far beyond the semimartingale setting and it has had a significant impact in the development of the theory of *regularity structures* by Hairer (2014), providing a mathematically rigorous description of many stochastic PDEs arising in physics.

More recently, interest has grown rapidly to develop machine learning algorithms based on rough path theoretical tools, particularly in the context of time series analysis (Kidger et al., 2019; Arribas et al., 2020; Lemercier et al., 2021b). The *signature*, a centrepiece of the theory, provides a top-down description of a stream; it captures crucial information such as the order of different events occurring across different channels, and filters out potentially superfluous information, such as the sampling rate of the signal.

In reservoir computing, the trajectory of a dynamical system is described through its interaction with a random dynamical system that is capable of storing information. In rough path theory the random system is replaced by a deterministic system given by the signature. Recently (Cuchiero et al., 2021a;b) have investigated empirically the idea of a continuous-time reservoir through the randomization of the signature yielding controlled residual architectures similar to the ones of interest to us.

A significant effort has been made to scale methods based on the signature to high dimensional signals. *Signature kernels* are defined as inner products of signatures and provide an elegant solution to this challenge thanks to the recent development of specific kernel tricks (Király & Oberhauser, 2019). Notably, Salvi et al. (2021a) establish that the signature kernel can be computed efficiently by solving a linear PDE. Algorithms based on signature kernels have been used in a wide range of applications including hypothesis testing (Salvi et al., 2021b), cybersecurity (Cochrane et al., 2021), and probabilistic forecasting (Toth & Oberhauser, 2020; Lemercier et al., 2021a) among others.

3. Inhomogeneous controlled ResNets

We begin by considering the case of inhomogeneous controlled ResNets. Contrarily to what one might expect, although in this setting the residual map changes at each iteration, the limiting kernels will be governed by simpler differential equations than their homogeneous counterparts, as it can be observed in Table 1. At an intuitive level, this fact can be justified by noting that sharing common random weights and biases throughout all iterations introduces a more intricate dependence structure on the dynamics of the system than if the weights and biases were independently sampled at each iteration.

3.1. The model

Let $\mathcal{D}_M = \{0 = t_0 < \cdots < t_M = 1\}$ be a partition, $N \in \mathbb{N}$ be the width, and $\varphi : \mathbb{R} \to \mathbb{R}$ an activation function. Define a *randomly initialized, 1-layer inhomogeneous controlled ResNet* $\Psi^{M,N} : \mathbb{X} \to \mathbb{R}$ as follows

$$\Psi_\varphi^{M,N}(x) := \left\langle \psi, \mathcal{S}_{t_M}^{M,N}(x) \right\rangle_{\mathbb{R}^N}$$

---

[3]To be precise one has to first expand our framework to include different initial values, of the form $W_{in}x_0$, where $x_0 \in \mathbb{R}^d$ is the "classical" model input. The corresponding control is then $x_t = x_0 + t\boldsymbol{e}_1$. Such richer models will be studied in future work.

Table 1. PDEs satisfied by the limiting kernels in the infinite-width-depth limit of controlled ResNets. The **first row** is for the *inhomogeneous case* while the **second row** for the *homogeneous* one. The kernel k_{sig} is the *signature kernel* (Salvi et al., 2021a).

Activation function	
General case	Identity case

$$\partial_t \kappa_\varphi^{x,y}(t) = \left[\sigma_A^2 V_\varphi\left(\begin{pmatrix} \kappa_\varphi^{x,x}(t) & \kappa_\varphi^{x,y}(t) \\ \kappa_\varphi^{x,y}(t) & \kappa_\varphi^{y,y}(t) \end{pmatrix}\right) + \sigma_b^2\right]\langle \dot{x}_t, \dot{y}_t\rangle \qquad \kappa_{id}^{x,y}(t) = \left(\sigma_a^2 + \frac{\sigma_b^2}{\sigma_A^2}\right)e^{\int_0^t \langle \sigma_A \dot{x}_s, \sigma_A \dot{y}_s\rangle ds} - \frac{\sigma_b^2}{\sigma_A^2}$$

$$\partial_t \partial_s \mathcal{K}_\varphi^{x,y}(s,t) = \left[\sigma_A^2 V_\varphi\left(\begin{pmatrix} \mathcal{K}_\varphi^{x,x}(s,s) & \mathcal{K}_\varphi^{x,y}(s,t) \\ \mathcal{K}_\varphi^{x,y}(s,t) & \mathcal{K}_\varphi^{y,y}(t,t) \end{pmatrix}\right) + \sigma_b^2\right]\langle \dot{x}_s, \dot{y}_t\rangle \qquad \mathcal{K}_{id}^{x,y}(s,t) = \left(\sigma_a^2 + \frac{\sigma_b^2}{\sigma_A^2}\right)k_{sig}^{\sigma_A x, \sigma_A y}(s,t) - \frac{\sigma_b^2}{\sigma_A^2}$$

where $\langle \cdot, \cdot\rangle_{\mathbb{R}^N}$ is the Euclidean inner product on $\mathbb{R}^N$, $\psi \in \mathbb{R}^N$ is a random vector with entries $[\psi]_\alpha \overset{iid}{\sim} \mathcal{N}(0, \frac{1}{N})$, and where the random functions $\mathcal{S}_{t_i}^{M,N} : \mathbb{X} \to \mathbb{R}^N$ satisfy the following recursive relation

$$\mathcal{S}_{t_{i+1}}^{M,N} = \mathcal{S}_{t_i}^{M,N} + \sum_{j=1}^{d}\left(A_{j,i}\varphi(\mathcal{S}_{t_i}^{M,N}) + b_{j,i}\right)\Delta x_{t_{i+1}}^j$$

$$\mathcal{S}_{t_0}^{M,N} = a \quad \text{and} \quad \Delta x_{t_i}^j = (x_{t_i}^j - x_{t_{i-1}}^j)$$

for $i = 0, ..., M$, with initial condition $[a]_\alpha \overset{iid}{\sim} \mathcal{N}(0, \sigma_a^2)$, and Gaussian weights $A_{k,l} \in \mathbb{R}^{N \times N}$ and biases $b_{k,l} \in \mathbb{R}^N$ sampled independently according to

$$[A_{j,i}]_\alpha^\beta \overset{iid}{\sim} \mathcal{N}\left(0, \frac{\sigma_A^2}{N\Delta t_i}\right), \quad [b_{j,i}]_\alpha \overset{iid}{\sim} \mathcal{N}\left(0, \frac{\sigma_b^2}{\Delta t_i}\right)$$

with time step $\Delta t_i = (t_i - t_{i-1}) > 0$.

Here $\sigma_a, \sigma_A > 0$ and $\sigma_b \geq 0$ are all model hyperparameters.

Remark. The time scaling $\frac{1}{\Delta t_i}$ in the random weights and biases is crucial as it is exactly the scaling one needs to get an Itô diffusion in the distributional infinite-depth limit, as we will prove in Theorem 3.3 below.

3.2. The infinite-width-depth regime

The first problem we are interested in studying is that of characterizing the limiting behavior of these neural networks in the infinite-width-then-depth regime.

Theorem 3.1 states that in this regime, these architectures converge weakly to GPs indexed on the path space $\mathbb{X}$ with kernels satisfying a one-parameter differential equation.

Theorem 3.1. *Let $\{\mathcal{D}_M\}_{M \in \mathbb{N}}$ be a sequence of partitions of $[0,1]$ such that $|\mathcal{D}_M| \to 0$. Let the activation function $\varphi : \mathbb{R} \to \mathbb{R}$ be linearly bounded, absolutely continuous and with exponentially bounded derivative. Then the following*

weak convergence[4] holds

$$\lim_{M \to \infty} \lim_{N \to \infty} \Psi_\varphi^{M,N} = \mathcal{GP}(0, \kappa_\varphi), \tag{2}$$

where the positive semidefinite kernel $\kappa_\varphi : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ is defined for any two paths $x, y \in \mathbb{X}$ as $\kappa_\varphi(x,y) = \kappa_\varphi^{x,y}(1)$, where $\kappa_\varphi^{x,y} : [0,1] \to \mathbb{R}$ is the unique solution of the following differential equation

$$\partial_t \kappa_\varphi^{x,y} = \left[\sigma_A^2 V_\varphi\left(\begin{pmatrix} \kappa_\varphi^{x,x} & \kappa_\varphi^{x,y} \\ \kappa_\varphi^{x,y} & \kappa_\varphi^{y,y} \end{pmatrix}\right) + \sigma_b^2\right]\langle \dot{x}_t, \dot{y}_t\rangle_{\mathbb{R}^d} \tag{3}$$

with initial condition $\kappa_\varphi^{x,y}(0) = \sigma_a^2$.

If moreover φ is Lipschitz with $\varphi(0) = 0$ and $x \in \mathbb{X} \cap C^{1,\frac{1}{2}}$, where $C^{1,\frac{1}{2}}$ denotes the set of C^1 paths with $\frac{1}{2}$-Hölder derivative, the limits can be exchanged and

$$\lim_{N \to \infty} \lim_{M \to \infty} \Psi_\varphi^{M,N} = \lim_{M \to \infty} \lim_{N \to \infty} \Psi_\varphi^{M,N} = \mathcal{GP}(0, \kappa_\varphi).$$

Idea of proof. We prove the weak convergence (2) in Appendix B.1. The first step consists in showing, for a fixed depth M, the existence of an infinite-width distributional limit using the techniques established in (Yang, 2019); this limit will be shown to be Gaussian and with covariance kernels $\kappa_{\mathcal{D}_M}^{x,y} : \mathcal{D}_M \to \mathbb{R}$ satisfying a difference equation. The second step amounts to prove that given any sequence of partitions $\mathcal{D}_M$ with $|\mathcal{D}_M| \to 0$, the sequence $\{\kappa_{\mathcal{D}_M}^{x,y}\}_M$ is uniformly bounded and uniformly equicontinuous so that by the Ascoli-Arzelà theorem the sequence admits a uniformly convergent subsequence. Finally, we prove that the limit of this subsequence is a solution of the differential equation (3) and that this solution is actually unique.

The statement about commutativity of limits is proved in Appendix B.3, after a characterization of the infinite-depth limit under these more stringent regularity assumptions, by proving that the distributional limit in depth is uniform in

---

[4]By weak convergence we mean that for any subset of paths $\mathcal{X} = \{x_1, \ldots, x_n\} \subset \mathbb{X}$ the random vector $\Psi_\varphi^{M,N}(\mathcal{X})$ converges in distribution to corresponding evaluations of the RHS limit.

width. This generalizes the results of (Hayou & Yang, 2023) in our more complex case.

□

In some cases we can explicitly characterize the limiting kernels by solving analytically the differential equation (3), as stated in the following corollary [5].

Corollary 3.2. *With the same notation and assumptions as in Theorem 3.1, upon taking $\varphi = id$ the limiting kernel admits the following explicit expression*

$$\kappa_{id}(x, y) = \left(\sigma_a^2 + \frac{\sigma_b^2}{\sigma_A^2}\right) \exp\left\{\sigma_A^2 \int_0^1 \langle \dot{x}_t, \dot{y}_t \rangle_{\mathbb{R}^d} dt\right\} - \frac{\sigma_b^2}{\sigma_A^2}.$$

If $\varphi = ReLU$ and $x = y$, then the limiting kernel satisfies

$$\kappa_{\varphi}(x, x) = \left(\sigma_a^2 + \frac{2\sigma_b^2}{\sigma_A^2}\right) \exp\left\{\frac{\sigma_A^2}{2} \int_0^1 \|\dot{x}_t\|_{\mathbb{R}^d}^2 dt\right\} - \frac{2\sigma_b^2}{\sigma_A^2}$$

Remark. In Lemma B.12 in the appendix we show that in the kernels governed by the dynamics (3), the parameters σ_A and σ_b satisfy the following path-rescaling symmetry

$$\kappa_{\varphi}^{x,y}(t; \sigma_A, \sigma_b) = \kappa_{\varphi}^{\sigma_A x, \sigma_A y}\left(t; 1, \frac{\sigma_b}{\sigma_A}\right).$$

Next we show that infinite-depth, finite-width networks are solutions of SDEs where the vector fields are controlled by the input stream. We will then specialise to the case $\varphi = id$ and identify the limiting kernel with κ_{id} from Corollary 3.2

3.3. The finite-width, infinite-depth regime

Our next result states that when their width N is fixed, these networks converge to a well defined distributional limit as their depth M tends to infinity. In particular, in this limit, the random weights behave like white noise, and thanks to the careful choice of time scaling we have made, the limit is in fact a zero-drift Itô diffusion with diffusion coefficient depending on the driving path.

Theorem 3.3. *Let $\{\mathcal{D}_M\}_{M \in \mathbb{N}}$ be a sequence of partitions of $[0, 1]$ such that $|\mathcal{D}_M| \to 0$ as $M \to \infty$. Assume the activation function φ is Lipschitz and linearly bounded. Let $\rho_M(t) := \sup\{s \in \mathcal{D}_M : s \leq t\}$. For any path $x \in \mathbb{X} \cap C^{1,\frac{1}{2}}$, where $C^{1,\frac{1}{2}}$ denotes the set of C^1 paths with $\frac{1}{2}$-Hölder derivative, the $\mathbb{R}^N$-valued process $t \mapsto \mathcal{S}_{\rho_M(t)}^{M,N}(x)$ converges in distribution, as $M \to \infty$, to the solution $\mathcal{S}^N(x)$ of the following SDE*

$$d\mathcal{S}_t^N(x) = \sum_{j=1}^d \frac{\sigma_A}{\sqrt{N}} \dot{x}_t^j dW_t^j \varphi(\mathcal{S}_t^N(x)) + \sigma_b \dot{x}_t^j dB_t^j \quad (4)$$

with $\mathcal{S}_0^N(x) = a$ and where $W^j \in \mathbb{R}^{N \times N}$ and $B^j \in \mathbb{R}^N$ are independent Brownian motions for $j \in \{1, ..., d\}$.

---

[5]We note that these characterizations expressed by means of an exponential are consistent with the results of (Hayou, 2022).

Idea of proof. The idea is proving that the finite difference scheme defining the inhomogeneous architecture gets closer and closer, as the mesh size of the partition becomes finer, to a Euler discretization of Equation (4). One then concludes with standard results which guarantee the convergence of Euler discretizations to the relative SDE's solution. □

Remark. Equation (4) can be easily rewritten in more standard SDE form as follows (see Appendix for more details)

$$d\mathcal{S}_t^N(x) = \sigma_x(t, \mathcal{S}_t^N(x))dZ_t$$

where $Z_t \in \mathbb{R}^{dN(N+1)}$ is a standard Brownian motion, independent from a and $\sigma_x : [0, 1] \times \mathbb{R}^N \to \mathbb{R}^{N \times dN(N+1)}$ is an input-dependent matrix valued function.

Passing directly to the infinite-width limit is not as easy as it could seem, Tensor Program arguments do not apply any longer since they are built for discrete layers and "collapse" in the continuous case we have to now work with. In simpler cases the limit can be found using McKean-Vlasov arguments as in (Hayou, 2022) and we conjecture that similar results can be found in this more general setting. We leave such a study to future work.

In any case it is possible to directly prove this in the simplest case, when $\varphi = id$. The result is proved in Appendix B.2.2.

4. Homogeneous controlled ResNets

In this section we consider the more complex setting of networks in which the weights are shared across layers. We will see that this weight-sharing feature will yield limiting kernels governed by two-parameter, non-local partial differential differential equations. We will follow a similar structure as in the previous section, commenting on the crucial differences along the way.

4.1. The Model

Define a *randomly initialized, 1-layer homogeneous controlled ResNet* $\Phi_{\varphi}^{M,N} : \mathbb{X} \to \mathbb{R}$ as follows

$$\Phi_{\varphi}^{M,N}(x) := \left\langle \phi, S_{t_M}^{M,N}(x) \right\rangle_{\mathbb{R}^N}$$

where $\phi \in \mathbb{R}^N$ is the random vector $[\phi]_\alpha \overset{\text{iid}}{\sim} \mathcal{N}(0, \frac{1}{N})$, and where the random functions $S_{t_i}^{M,N} : \mathbb{X} \to \mathbb{R}^N$ satisfy the following recursive relation

$$S_{t_{i+1}}^{M,N} = S_{t_i}^{M,N} + \sum_{k=1}^d \left(A_k \varphi(S_{t_i}^{M,N}) + b_k\right) \Delta x_{t_{i+1}}^k$$

with initial condition $S_{t_0} = a$ with $[a]_\alpha \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_a^2)$, and Gaussian weights $A_k \in \mathbb{R}^{N \times N}$ and biases $b_k \in \mathbb{R}^N$ sampled independently according to

$$[A_k]_\alpha^\beta \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \frac{\sigma_A^2}{N}\right), \quad [b_k]_\alpha \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \sigma_b^2\right).$$

As done in the homogeneous case, we now study the limiting behavior of homogeneous controlled ResNets in the infinite-width-depth limit; as in the in-homogeneous case of the previous section, we will show that the limits commute.

4.2. The infinite-width-depth regime

Theorem 4.1 states that in this regime, these architectures converge weakly to GPs indexed on the path space $\mathbb{X}$ with kernels satisfying a two-parameters differential equation.

Theorem 4.1. *Let $\{\mathcal{D}_M\}_{M \in \mathbb{N}}$ be a sequence of partitions of $[0,1]$ such that $|\mathcal{D}_M| \to 0$ as $M \to \infty$. Let the activation function φ be linearly bounded, absolutely continuous and with exponentially bounded derivative. Then the following weak convergence holds*

$$\lim_{M \to \infty} \lim_{N \to \infty} \Phi_\varphi^{M,N} = \mathcal{GP}(0, \mathcal{K}_\varphi) \quad (5)$$

where the positive semidefinite kernel $\mathcal{K}_\varphi : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ is defined for any two paths $x, y \in \mathbb{X}$ as $\mathcal{K}_\varphi(x,y) = \mathcal{K}_\varphi^{x,y}(1,1)$ where the function $\mathcal{K}_\varphi^{x,y} : [0,1] \times [0,1] \to \mathbb{R}$ is the unique solution of the following differential equation

$$\partial_s \partial_t \mathcal{K}_\varphi^{x,y} = \left[\sigma_A^2 V_\varphi \left(\Sigma_\varphi^{x,y}(s,t) \right) + \sigma_b^2 \right] \langle \dot{x}_s, \dot{y}_t \rangle \quad (6)$$

where

$$\Sigma_\varphi^{x,y}(s,t) = \begin{pmatrix} \mathcal{K}_\varphi^{x,x}(s,s) & \mathcal{K}_\varphi^{x,y}(s,t) \\ \mathcal{K}_\varphi^{x,y}(s,t) & \mathcal{K}_\varphi^{y,y}(t,t) \end{pmatrix}$$

and with initial conditions for any $s,t \in [0,1]$

$$\mathcal{K}_\varphi^{x,y}(0,0) = \mathcal{K}_\varphi^{x,y}(s,0) = \mathcal{K}_\varphi^{x,y}(0,t) = \sigma_a^2.$$

If moreover φ is Lipschitz the limits can be exchanged and

$$\lim_{M \to \infty} \lim_{N \to \infty} \Phi_\varphi^{M,N} = \lim_{N \to \infty} \lim_{M \to \infty} \Phi_\varphi^{M,N} = \mathcal{GP}(0, \kappa_\varphi).$$

Remark. It is non-trivial to show not only that the problem is well-posed but even that equation (6) is well defined because the "instantaneous rate of change" $\partial_s \partial_t \mathcal{K}_\varphi^{x,y}(s,t)$ at times $s < t$ depends on the "past" values $\mathcal{K}_\varphi^{x,x}(s,s)$, on the "present" values $\mathcal{K}_\varphi^{x,y}(s,t)$ and on the "future" values $\mathcal{K}_\varphi^{y,y}(t,t)$. The nonlocal nature of these dynamics is such that it is *a priori* not clear that the RHS of (6) even has meaning since the matrix $\Sigma_\varphi^{x,y}(s,t)$ could be not positive semidefinite and V_φ is only defined on PSD matrices.

Idea of proof. Similarly to Theorem 3.1, due to the complexity of the arguments, this result is proved in several steps. In in Appendix C.1. The first step consists of showing that the infinite width limit is well defined for any choice of $\mathcal{D}_M$. This will be a GP defined by a kernel $\mathcal{K}_{\mathcal{D}_M \times \mathcal{D}_M}$ found as the terminal value of a finite difference scheme having the

same form as that of a Euler discretization, on $\mathcal{D}_M \times \mathcal{D}_M$, of equation (6). The second step consists in proving that the kernels $\{\mathcal{K}_{\mathcal{D}_M \times \mathcal{D}_M}\}_M$ constitute, in a suitable metric, a Cauchy sequence as $|\mathcal{D}_M| \to 0$, that the limit is independent from the chosen sequence of partitions and that it does indeed uniquely solve equation (6). The final step, proved in Appendix C.3, concerns the exchange of limits. After a characterization of the infinite-depth limits, we will prove that the distributional limit as depth goes to infinity is uniform in the width, thus we will be able to use the classical Moore-Osgood theorem to justify the exchange. $\square$

When the activation function φ is the identity, equation (6) reduces to a linear hyperbolic PDE. Upon inspection, we unveil a surprising link with the *signature kernel*, a well-studied object in rough analysis corresponding to an inner product between two *path-signatures*, and that was shown by Salvi et al. (2021a) to satisfy a similar PDE. This is the content of the next corollary.

Corollary 4.2. *Using the same notation and assumptions as in Theorem 4.1, choosing $\varphi = id$ the limiting kernel satisfies the following identity*

$$\mathcal{K}_{id}^{x,y}(s,t) = \left(\sigma_a^2 + \frac{\sigma_b^2}{\sigma_A^2} \right) k_{sig}^{\sigma_A x, \sigma_A y}(s,t) - \frac{\sigma_b^2}{\sigma_A^2}$$

where $k_{sig}^{x,y}$ is the signature kernel from (Salvi et al., 2021a) which for any two paths $x, y \in \mathbb{X}$ and $s, t \in [0,1]$ satisfies the following linear hyperbolic PDE

$$\partial_s \partial_t k_{sig}^{x,y} = \langle \dot{x}_s, \dot{y}_t \rangle k_{sig}^{x,y} \quad (7)$$

with initial conditions $k_{sig}^{x,y}(s,0) = k_{sig}^{x,y}(0,t) = 1$.

In other words we have unveiled a novel family of kernels indexed on continuous paths which generalizes the signature kernel in (Salvi et al., 2021a). We name this new class of kernels *neural signature kernels*. We note that this generalization is done directly at the level of the driving PDE unlike the extensions studied in (Cass et al., 2021) which use a different inner product structure on the space where signatures live.

Remark. Analogously to the inhomogeneous case, the parameters σ_A and σ_b defining the neural signature kernels governed by the dynamics in equation (6) satisfy the following path-rescaling symmetry

$$\mathcal{K}_\varphi^{x,y}(s,t; \sigma_A, \sigma_b) = \mathcal{K}_\varphi^{\sigma_A x, \sigma_A y}\left(s,t; 1, \frac{\sigma_b}{\sigma_A} \right)$$

as shown in Lemma C.15 in the appendix.

Remark. For non-linear activation functions φ, the neural signature kernel non-linear PDE (6) and the id-neural signature kernel linear PDE (7) might in principle admit the same solution, which would mean essentially that linear and
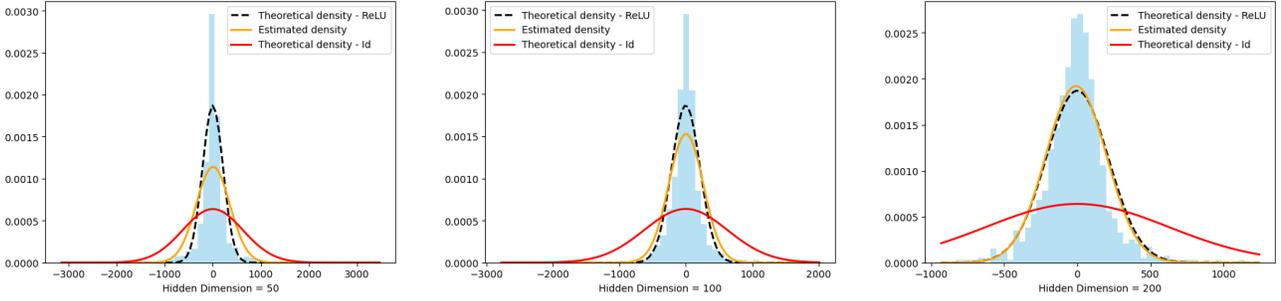
Figure 1. Fixed input $y(t) := cos(15t) + 3e^t$. Histogram of 700 independent realizations of $N^{-1} \langle S_1(y), S_1(y) \rangle_{\mathbb{R}^N}$ for *ReLU*-RandomizedSignatures $S(y) \in \mathbb{R}^N$, $N \in \{50, 100, 200\}$, plotted against Closest Gaussian Fit, ReLU-NeuralSigKer and id-NeuralSigKer.

non-linear controlled ResNets behave in the same way in the infinite-width-depth regime. In in Figure 1 we show empirically that this is not the case in general, by comparing the limiting empirical distributions for $\varphi = id$ and $\varphi = ReLU$.

4.3. The finite-width, infinite-depth regime

Our next result states that when their width N is fixed, homogeneous controlled ResNets converge in distribution, in $[0, 1]$, to a Neural CDE with random vector fields.

Theorem 4.3. *Let $\{\mathcal{D}_M\}_{M \in \mathbb{N}}$ be a sequence of partitions of $[0, 1]$ such that $|\mathcal{D}_M| \to 0$ as $M \to \infty$. Assume the activation function φ is Lipschitz and linearly bounded. Let $x \in \mathbb{X}$ and let $\rho_M(t) := \sup\{s \in \mathcal{D}_M : s \leq t\}$. Then, the $\mathbb{R}^N$-valued process $t \mapsto S_{\rho_M(t)}^{M,N}(x)$ converges in distribution[6], as $M \to \infty$, to the solution $S^N(x)$ of the following Neural CDE*

$$dS_t^N(x) = \sum_{j=1}^d \left(A_j \varphi(S_t^N(x)) + b_j \right) dx_t^j \qquad (8)$$

where $A_j \in \mathbb{R}^{N \times N}$ and $b_j \in \mathbb{R}^N$ are sampled according in the definition of the homogeneous controlled ResNet.

Idea of proof. If we fix a, the A_ks and the b_ks to be the same for all $\mathcal{D}_M$ then we have uniform convergence by classical results. The rate of convergence can be bounded with some constants depending on the entries of a, A_k, b_k and which, thanks to Gaussianity, have finite expectation. It is just a matter of applying the classical *portmanteau* lemma to conclude. $\square$

Remark. This can be naturally extended in order to take into consideration the joint distribution for different input choices.

The solutions to equation (8) have been informally introduced in (Cuchiero et al., 2021b; Akyildirim et al., 2022) as

---

[6]As random variables with values in $L^\infty([0, 1]; \mathbb{R}^N)$.

lower dimensional approximations of path-signatures, and have been dubbed by the authors *randomized signatures*.

Taking directly the infinite width limit is once again far from trivial, reasoning *à la* Tensor Program quickly collapse and there is no clear possible future path corresponding to the McKean-Vlasov ideas for the inhomogeneous case. The problem is that the randomness is in the vector fields themselves and not in the driving paths, courtesy of the cross-layer dependencies in the homogeneous networks. This is why it's necessary to sidestep the problem by proving the existence of uniform convergence bounds.

4.4. The infinite-depth-then-width regime: $\varphi = id$

As anticipated, contrary to the inhomogeneous case, in the current homogeneous setting, when φ is the identity, we are able to prove directly that the limits in Equation (5) commute as well as explicit convergence bounds.

Theorem 4.4. *If $\varphi = id$ and for any $x, y \in \mathbb{X}$*

$$\frac{1}{N} \left\langle S_s^N(x), S_t^N(y) \right\rangle_{\mathbb{R}^N} \xrightarrow[N \to \infty]{\mathbb{L}^2} \mathcal{K}_{id}^{x,y}(s, t)$$

on $[0, 1]^2$. Moreover the convergence is of order $\mathcal{O}(\frac{1}{N})$.

5. Numerics

In this section, we first illustrate theoretical results established in Section 4 and then outline numerical considerations to scale the computation of signature kernels.

5.1. Convergence of homogeneous controlled ResNets

We start by illustrating the convergence in distribution of a homogeneous controlled ResNet to a GP endowed with neural signature kernel as per Theorem 4.1. To this aim, we consider a homogeneous ResNet $\Phi_\varphi^{M,N}$ with activation function $\varphi = $ ReLU, and $(\sigma_a, \sigma_A, \sigma_b) = (0.5, 1., 1.2)$. For $R = 250$ realizations of the weights and biases, we run the model on a 2-dimensional path $x : t \mapsto (\sin(15t), \cos(30t) + 3e^t)$ observed at 100 regularly spaced time points in $[0, 1]$. We
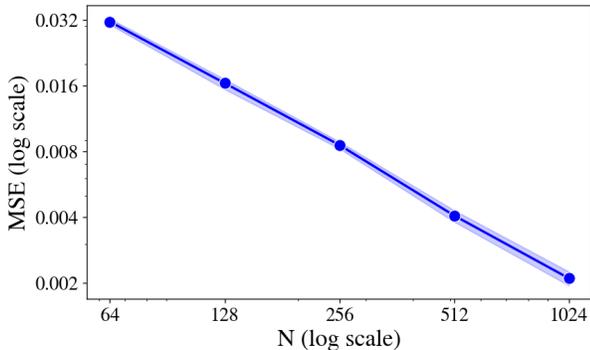
Figure 2. Mean squared error of $\frac{1}{N}\left\langle S_1^N(x), S_1^N(y)\right\rangle_{\mathbb{R}^N}$, the estimator of $\mathcal{K}_{id}^{x,y}(1,1)$, as a function of the width N on a logarithmic scale. Standard deviations were obtained by repeating the experiment 5 times.

then verify that, as N increases, $\Phi_\varphi^{M,N}(x)$ converges to a Gaussian random variable with mean zero and variance $\mathcal{K}_\varphi(x,x)$. This limiting variance is computed by solving Equation (6) on a fine discretization grid. As it can be observed on Figure 3 the Gaussian fit for this one-dimensional marginal gets better as N increases. Further results can be found in the appendix.

We then provide empirical evidence for the order of convergence provided in Theorem 4.4. Here, we consider a linear homogeneous ResNet, controlled by x and y, two sample paths from a zero-mean GP with RBF kernel $r_{RBF}(s,t) = \exp\left(-5(s-t)^2\right)$ with 50 observation points in $[-2,2]$. Similarly to the previous setup, we run the model with $M = 250$ different random initializations to estimate the mean squared error $\mathbb{E}[(\frac{1}{N}\langle S_1^N(x), S_1^N(y)\rangle - \mathcal{K}_{id}(x,y))^2]$ increasing the width N. Our empirical results, as displayed on Figure 2, align with the theoretical convergence rate.

5.2. Scaling signature kernels

The signature kernel of two paths is typically computed by approximating the solution of the PDE in (7) on a 2-dimensional time grid, which scales quadratically with the discretization step of the solver. Although an efficient numerical scheme leveraging GPU computations to update the solution at multiple time points on the grid in parallel has been proposed in Salvi et al. (2021a), the maximum number of threads in a GPU block imposes a hard limit on the discretization step of the solver, limiting the applicability of signature kernel methods to long time series. Theorem 4.4 offers a new way to compute the signature kernel by solving two CDEs linearly in time instead of one PDE quadratically in time; one would first run a wide and infinite-depth ResNet on the two control paths of interest, and then com-

pute the (rescaled) dot-product between the outputs of the penultimate layer. This approach allows for more flexibility regarding the choice of path interpolations and numerical solvers, as several options are made readily available in dedicated python packages such as torchcde (Kidger et al., 2020). Next, we describe possible ways to increase further the scalability of this approach.

Log-ODE method To further improve scalability of Neural CDEs for long time series Morrill et al. (2021) made use of the so-called *log-ODE scheme* to forward-solve the differential equation on much larger time intervals than the ones that would be expected given the sampling rate or length of the data. We leave the investigation of this numerical scheme for computing signature kernels as future work.

Sparse random matrices The forward pass of a ResNet involves several ($M \times d$ where M is the number of time steps, and d the dimension of the input path) matrix-vector multiplications where the entries of each N-by-N matrix are Gaussian distributed. As remarked in (Dong et al., 2020), in the context of random RNNs, to speed-up these computations, the dense weight matrices can be replaced by structured random matrices given by the products of random (binary) diagonal matrices and Walsh-Hadamard matrices. The complexity of the matrix-vector product can be reduced to $\mathcal{O}(N \log N)$ leveraging the fast Hadamard transform algorithm (without sampling the Walsh-Hadamard matrices).

Random Fourier features Several machine learning use cases of the signature kernel have provided empirical evidence that embedding the input paths pointwise in time in a feature space can be beneficial to increase the performance of kernel methods on sequential data. In particular, when the paths evolve in a Euclidean space, the RBF kernel often turns out to be a good choice. Although this embedding is infinite-dimensional, random Fourier features (Rahimi & Recht, 2007) make it possible to approximate it by a finite-dimensional one. One could then investigate randomly initialized ResNets, controlled by sequences of such approximate embeddings.

6. Conclusion and future work

In this paper we considered controlled ResNets defined as Euler-discretizations of Neural CDEs. We showed that in both the infinite-depth-then-width and in the infinite-width-then-depth limit, these converge weakly to the same GP indexed on path space endowed with neural signature kernels satisfying certain (possibly non-linear) PDEs varying according to the choice of activation function φ. In the special case where φ is the identity, we showed that the equation reduces to a linear PDE and the limiting kernel agrees with the signature kernel. In this setting, we also pro-
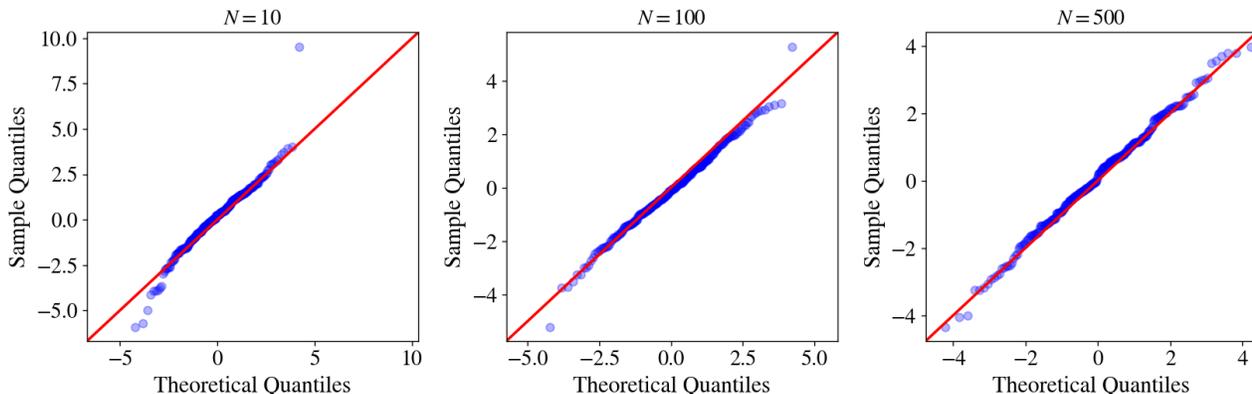
Figure 3. Empirical quantiles against the theoretical quantiles of $\mathcal{N}(0, \mathcal{K}_{\varphi}^{x,x}(1))$ for $N \in \{10, 100, 500\}$ with $\varphi = \text{ReLU}$

vided explicit convergence rates. Finally, we showed that in the infinite-depth regime, finite-width controlled ResNets converge in distribution to Neural CDEs with random vector fields which are either time-independent and Gaussian, if the system is homogeneous, or behave like a matrix-valued Brownian motion, if the system is inhomogeneous.

We believe that a rigorous investigation of the functional analytic properties of the *reproducing kernel Hilbert spaces* (RKHSs) associated to the new family of neural signature kernels is also a compelling future research direction. In particular, it would allow to build an understanding of the expressivity and generalization properties of these kernels.

In the homogeneous setting, the vector fields are constant functions while in the in-homogeneous setting they are described by white noise. Investigating the intermediate regularity cases is an interesting avenue for future research; for example considering matrices and biases sampled from of Fractional Brownian Motion increments with Hurst exponent $H \in [0, 1]$ (the inhomogeneous case corresponds to the case $H = 0.5$ while the homogeneous one to $H = 1$).

Last but not least, establishing expressions and analyzing the associated *Neural Tangent Kernels* (NTK) (Jacot et al., 2018; Yang, 2020) would provide quantitative insights on the training mechanism of Neural CDEs by gradient descent.

All the experiments presented in this paper are reproducible following the code at https://github.com/MucaCirone/NeuralSignatureKernels

7. Aknowledgements

References

Akyildirim, E., Gambara, M., Teichmann, J., and Zhou, S. Applications of signature methods to market anomaly detection, 2022. URL https://arxiv.org/abs/2201.02441.

Alemohammad, S., Wang, Z., Balestriero, R., and Baraniuk, R. The recurrent neural tangent kernel. In *International Conference on Learning Representations*, 2020.

Arribas, I. P., Salvi, C., and Szpruch, L. Sig-sdes model for quantitative finance. In *Proceedings of the First ACM International Conference on AI in Finance*, pp. 1–8, 2020.

Baudoin, F. and Zhang, X. Taylor expansion for the solution of a stochastic differential equation driven by fractional Brownian motions. *Electronic Journal of Probability*, 17 (none):1 – 21, 2012. doi: 10.1214/EJP.v17-2136. URL https://doi.org/10.1214/EJP.v17-2136.

Bellot, A. and Van Der Schaar, M. Policy analysis using synthetic controls in continuous-time. In *International Conference on Machine Learning*, pp. 759–768. PMLR, 2021.

Cass, T., Lyons, T., and Xu, X. General signature kernels. *arXiv preprint arXiv:2107.00447*, 2021.

Chen, R. T., Rubanova, Y., Bettencourt, J., and Duvenaud, D. K. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

Cochrane, T., Foster, P., Chhabra, V., Lemercier, M., Lyons, T., and Salvi, C. Sk-tree: a systematic malware detection algorithm on streaming trees via the signature kernel. In *2021 IEEE International Conference on Cyber Security and Resilience (CSR)*, pp. 35–40. IEEE, 2021.

Cohen, A.-S., Cont, R., Rossier, A., and Xu, R. Scaling properties of deep residual networks. In *International Conference on Machine Learning*, pp. 2039–2048. PMLR, 2021.

Cont, R., Rossier, A., and Xu, R. Asymptotic analysis of deep residual networks. *arXiv preprint arXiv:2212.08199*, 2022.

Cuchiero, C., Gonon, L., Grigoryeva, L., Ortega, J.-P., and Teichmann, J. Discrete-time signatures and randomness in reservoir computing. *IEEE Transactions on Neural Networks and Learning Systems*, 2021a.

Cuchiero, C., Gonon, L., Grigoryeva, L., Ortega, J.-P., and Teichmann, J. Expressive power of randomized signature. In *The Symbiosis of Deep Learning and Differential Equations*, 2021b. URL https://openreview.net/forum?id=KWWFPULvmVw.

Dong, J., Ohana, R., Rafayelyan, M., and Krzakala, F. Reservoir computing meets recurrent kernels and structured transforms. *Advances in Neural Information Processing Systems*, 33:16785–16796, 2020.

Fermanian, A., Marion, P., Vert, J.-P., and Biau, G. Framing rnn as a kernel method: A neural ode approach. *Advances in Neural Information Processing Systems*, 34:3121–3134, 2021.

Friz, P. K. and Hairer, M. *A course on rough paths*. Springer, 2020.

Friz, P. K. and Victoir, N. B. *Multidimensional Stochastic Processes as Rough Paths: Theory and Applications*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2010. ISBN 9781139487214. URL https://books.google.co.uk/books?id=CVgwLatxfGsC.

Garriga-Alonso, A., Rasmussen, C. E., and Aitchison, L. Deep convolutional networks as shallow gaussian processes. *arXiv preprint arXiv:1808.05587*, 2018.

Geman, S. A limit theorem for the norm of random matrices. *The Annals of Probability*, 8(2):252–261, 1980. ISSN 00911798. URL http://www.jstor.org/stable/2243269.

Hairer, M. A theory of regularity structures. *Inventiones mathematicae*, 198(2):269–504, 2014.

Hayou, S. On the infinite-depth limit of finite-width neural networks. *arXiv preprint arXiv:2210.00688*, 2022.

Hayou, S. and Yang, G. Width and depth limits commute in residual networks, 2023.

Huang, Y. and Tsai, C.-Y. Euler scheme for a stochastic goursat problem. *Stochastic Analysis and Applications*, 22(2):275–287, 2004. doi: 10.1081/SAP-120028590. URL https://doi.org/10.1081/SAP-120028590.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Johansson, K. Shape fluctuations and random matrices. *Communications in Mathematical Physics*, 209(2):437–476, 2000. doi: 10.1007/s002200050027. URL https://doi.org/10.1007/s002200050027.

Johnstone, I. M. On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics*, 29(2):295–327, 2001. ISSN 00905364. URL http://www.jstor.org/stable/2674106.

Kidger, P. On neural differential equations. *arXiv preprint arXiv:2202.02435*, 2022.

Kidger, P., Bonnier, P., Perez Arribas, I., Salvi, C., and Lyons, T. Deep signature transforms. *Advances in Neural Information Processing Systems*, 32, 2019.

Kidger, P., Morrill, J., Foster, J., and Lyons, T. Neural controlled differential equations for irregular time series. *Advances in Neural Information Processing Systems*, 33:6696–6707, 2020.

Király, F. J. and Oberhauser, H. Kernels for sequentially ordered data. *Journal of Machine Learning Research*, 20, 2019.

Kloeden, P. and Platen, E. *Numerical Solution of Stochastic Differential Equations*. Applications of mathematics : stochastic modelling and applied probability. Springer, 1992. ISBN 9783540540625. URL https://books.google.co.uk/books?id=7bkZAQAAIAAJ.

Lee, J., Sohl-dickstein, J., Pennington, J., Novak, R., Schoenholz, S., and Bahri, Y. Deep neural networks as gaussian processes. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=B1EA-M-0Z.

Lemercier, M., Salvi, C., Cass, T., Bonilla, E. V., Damoulas, T., and Lyons, T. J. Siggpde: Scaling sparse gaussian processes on sequential data. In *International Conference on Machine Learning*, pp. 6233–6242. PMLR, 2021a.

Lemercier, M., Salvi, C., Damoulas, T., Bonilla, E., and Lyons, T. Distribution regression for sequential data. In *International Conference on Artificial Intelligence and Statistics*, pp. 3754–3762. PMLR, 2021b.

Li, M., Nica, M., and Roy, D. The future is log-gaussian: Resnets and their infinite-depth-and-width limit at initialization. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 7852–7864. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/412758d043dd247bddea07c7ec558c31-Paper.pdf.

Li, M. B., Nica, M., and Roy, D. M. The neural covariance sde: Shaped infinite depth-and-width networks at initialization, 2022. URL https://arxiv.org/abs/2206.02768.

Lukoševičius, M. and Jaeger, H. Reservoir computing approaches to recurrent neural network training. *Computer Science Review*, 3(3):127–149, 2009.

Lyons, T. J. Differential equations driven by rough signals. *Revista Matemática Iberoamericana*, 14(2):215–310, 1998.

Lyons, T. J., Caruana, M., and Lévy, T. *Differential equations driven by rough paths*. Springer, 2007.

Majumdar, S. N. and Vergassola, M. Large deviations of the maximum eigenvalue for wishart and gaussian random matrices. *Phys. Rev. Lett.*, 102: 060601, Feb 2009. doi: 10.1103/PhysRevLett.102. 060601. URL https://link.aps.org/doi/10.1103/PhysRevLett.102.060601.

Marion, P., Fermanian, A., Biau, G., and Vert, J.-P. Scaling resnets in the large-depth regime. *arXiv preprint arXiv:2206.06929*, 2022.

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

Morrill, J., Kidger, P., Yang, L., and Lyons, T. Neural controlled differential equations for online prediction tasks. *arXiv preprint arXiv:2106.11028*, 2021.

Neal, R. M. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

Novak, R., Xiao, L., Lee, J., Bahri, Y., Yang, G., Hron, J., Abolafia, D. A., Pennington, J., and Sohl-Dickstein, J. Bayesian deep convolutional networks with many channels are gaussian processes. *arXiv preprint arXiv:1810.05148*, 2018.

Poole, B., Lahiri, S., Raghu, M., Sohl-Dickstein, J., and Ganguli, S. Exponential expressivity in deep neural networks through transient chaos, 2016. URL https://arxiv.org/abs/1606.05340.

Rahimi, A. and Recht, B. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007.

Salvi, C., Cass, T., Foster, J., Lyons, T., and Yang, W. The signature kernel is the solution of a goursat pde. *SIAM Journal on Mathematics of Data Science*, 3(3):873–899, 2021a.

Salvi, C., Lemercier, M., Liu, C., Horvath, B., Damoulas, T., and Lyons, T. Higher order kernel mean embeddings to capture filtrations of stochastic processes. *Advances in Neural Information Processing Systems*, 34:16635–16647, 2021b.

Salvi, C., Lemercier, M., and Gerasimovics, A. Neural stochastic pdes: Resolution-invariant learning of continuous spatiotemporal dynamics. In *Advances in Neural Information Processing Systems*, 2022.

Schoenholz, S. S., Gilmer, J., Ganguli, S., and Sohl-Dickstein, J. Deep information propagation. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=H1W1UN9gg.

Singh, S., Ramirez, F. M., Varley, J., Zeng, A., and Sindhwani, V. Multiscale sensor fusion and continuous control with neural cdes. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 10897–10904. IEEE, 2022.

Tanaka, G., Yamane, T., Héroux, J. B., Nakane, R., Kanazawa, N., Takeda, S., Numata, H., Nakano, D., and Hirose, A. Recent advances in physical reservoir computing: A review. *Neural Networks*, 115:100–123, 2019. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2019.03.005. URL https://www.sciencedirect.com/science/article/pii/S0893608019300784.

Tao, T. *Analysis I: Third Edition*. Texts and Readings in Mathematics. Springer Nature Singapore, 2016. ISBN 9789811017896. URL https://books.google.co.uk/books?id=ecTsDAAAQBAJ.

Toth, C. and Oberhauser, H. Bayesian learning from sequential data using gaussian processes with signature covariances. In *International Conference on Machine Learning*, pp. 9548–9560. PMLR, 2020.

Verstraeten, D., Schrauwen, B., d'Haene, M., and Stroobandt, D. An experimental unification of reservoir computing methods. *Neural networks*, 20(3):391–403, 2007.

Yang, G. Wide feedforward or recurrent neural networks of any architecture are gaussian processes. *Advances in Neural Information Processing Systems*, 32, 2019.

Yang, G. Tensor programs ii: Neural tangent kernel for any architecture. *arXiv preprint arXiv:2006.14548*, 2020.

A. Preliminaries

In this section we are going to state some preliminary results and considerations which we are going to refer to through the entirety of the text.

Throughout the paper we fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

A.1. Assumptions on path regularity

Since we are mainly interested in studying time-series, our data space will be a space of paths, more specifically we are going to consider the space

$$\mathbb{X} := \{x \in C^0([0,1]; \mathbb{R}^d) : x(0) = 0, \exists \dot{x} \in L^2([0,1]; \mathbb{R}^d)\}$$

i.e. $\mathbb{X}$ is the space of continous paths which have a square integrable derivative. $\mathbb{X}$ is the closed subset of the Sobolev space $\left(W^{1,2}([0,1])\right)^d$ made of those functions starting at the origin.

Note that every path $x \in \mathbb{X}$ can be uniquely written as

$$x_t = \int_0^t \dot{x}_s ds$$

thus it naturally corresponds to the space $L^2([0,1]; \mathbb{R}^d)$ trough the identification $x \mapsto \dot{x}$. This identification gives the space the natural norm $\|x\|_{\mathbb{X}} = \|\dot{x}\|_{L^2}$.

This norm is equivalent to the induced norm from $\left(W^{1,2}([0,1])\right)^d$ since

$$\|\dot{x}\|_{L^2}^2 \leq \|x\|_{W^{1,2}}^2 = \int_0^1 |x_s|^2 ds + \|\dot{x}\|_{L^2}^2 = \int_0^1 |\int_0^s \dot{x}_r dr|^2 ds + \|\dot{x}\|_{L^2}^2$$

$$\leq \int_0^1 \int_0^1 |\dot{x}_r|^2 dr ds + \|\dot{x}\|_{L^2}^2 = 2 \|\dot{x}\|_{L^2}^2$$

for $x \in \mathbb{X}$, hence $(\mathbb{X}, \|\cdot\|_{\mathbb{X}})$ is a Banach space. Moreover one can easily see that every $x \in \mathbb{X}$ has bounded variation and

$$\|x\|_{1-var,[0,1]} = \int_0^1 |\dot{x}_t| dt \leq \sqrt{\int_0^1 |\dot{x}_t|^2 dt} = \|x\|_{\mathbb{X}}$$

Given n paths $\mathcal{X} = \{x_1, \ldots, x_n\} \subset \mathbb{X}$ and functions $f : \mathbb{X} \to \mathbb{R}$, $G : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ we will write $f(\mathcal{X}) \in \mathbb{R}^n$ for the vector $[f(\mathcal{X})]_\alpha = f(x_\alpha)$ and $G(\mathcal{X}, \mathcal{X}) \in \mathbb{R}^{n \times n}$ for the matrix $[G(\mathcal{X}, \mathcal{X})]_\alpha^\beta = G(x_\alpha, x_\beta)$ for any $\alpha, \beta \in \{1, .., n\}$.

A.2. Assumptions on the activation function

Now we are going to state the main assumptions on the activation function $\varphi : \mathbb{R} \to \mathbb{R}$ and prove some important technical results of which we will frequently make use in the following sections. We will particularly be interested in how regularity assumptions made on φ influence the regularity of the expectations of Equation (6) and (3).

Here are the crucial assumptions we make on the activation:

Assumption A.1. The activation function $\varphi : \mathbb{R} \to \mathbb{R}$ is linearly bounded *i.e.* such that there exist some $M > 0$ such that $|\varphi(x)| \leq M(1 + |x|)$.

Assumption A.2. The activation function $\varphi : \mathbb{R} \to \mathbb{R}$ is absolutely continuous and with exponentially bounded derivative.

Lemma A.3. *If the activation function $\varphi : \mathbb{R} \to \mathbb{R}$ is K-Lipschitz then its componentwise extension $\varphi : \mathbb{R}^N \to \mathbb{R}^N$ is K-Lipschitz too. If the activation function $\varphi : \mathbb{R} \to \mathbb{R}$ is M-linealy-bounded then its componentwise extension is $\sqrt{2N}M$-linealy-bounded.*

Proof. Regarding the first proposition, for $x, y \in \mathbb{R}^n$ we have

$$|\varphi(x) - \varphi(y)|_{\mathbb{R}^N}^2 = \sum_{i=1}^N |\varphi(x_i) - \varphi(y_i)|^2 \leq \sum_{i=1}^N K^2 |x_i - y_i|^2 = K^2 |x - y|_{\mathbb{R}^N}^2$$

Concerning the second, notice how

$$|\varphi(x)|_{\mathbb{R}^N}^2 = \sum_{i=1}^N |\varphi(x_i)|^2 \leq \sum_{i=1}^N M^2(1 + |x_i|)^2 \leq 2M^2 \sum_{i=1}^N (1 + |x_i|^2)$$
$$= 2M^2(N + |x|_{\mathbb{R}^n}^2) \leq 2NM^2(1 + |x|_{\mathbb{R}^N}^2)$$

thus $|\varphi(x)|_{\mathbb{R}^N} \leq \sqrt{2N}M(1 + |x|_{\mathbb{R}^N})$ since $\sqrt{1 + \epsilon} \leq 1 + \sqrt{\epsilon}$ for all $\epsilon \geq 0$. $\qquad\square$

Remark. Note how we have proved also that, under the linear boundedness assumption, using the same final bound

$$|\varphi(x)|_{\mathbb{R}^N} \leq \sqrt{2}M(\sqrt{N} + |x|_{\mathbb{R}^N}) \tag{9}$$

A.3. Positive semidefinite matrices and the map V_φ

Definition A.4. Let PSD_2 denote the set of 2×2 positive semidefinite matrices

$$PSD_2 := \{\Sigma \in \mathbb{R}^{2 \times 2} : \Sigma = \Sigma^T; ([\Sigma]_1^2)^2 \leq [\Sigma]_1^1 [\Sigma]_2^2; 0 \leq [\Sigma]_1^1 \wedge [\Sigma]_2^2\}$$

For a fixed $R > 0$ we define the space

$$PSD_2(R) := \{\Sigma \in PSD_2 : \frac{1}{R} \leq [\Sigma]_1^1, [\Sigma]_2^2 \leq R\}$$

Lemma A.5. *Under **Assumption A.2** and for any $R > 0$ the function $V_\varphi : \mathrm{PSD}_2(R) \to \mathbb{R}$ defined for any $\Sigma \in \mathrm{PSD}_2(R)$ as*

$$V_\varphi(\Sigma) = \mathbb{E}_{(Z_x, Z_y) \sim \mathcal{N}(0, \Sigma)}[\varphi(Z_x)\varphi(Z_y)]$$

is κ_R-Lipschitz for some $\kappa_R > 0$, i.e.

$$|V_\varphi(\Sigma) - V_\varphi(\tilde{\Sigma})| \leq \kappa_R \|\Sigma - \tilde{\Sigma}\|_\infty$$

Proof. This is the content of Theorem F.4 in (Novak et al., 2018). $\qquad\square$

Proposition A.6. *Under **Assumption A.1**, there exists a positive constant $\tilde{M} > 0$ such that*

$$|V_\varphi(\Sigma)| \leq \tilde{M} \left(1 + \sqrt{[\Sigma]_1^1}\right)\left(1 + \sqrt{[\Sigma]_2^2}\right)$$

Proof. In fact given a PSD matrix $\Sigma \in \mathbb{R}^{2 \times 2}$ we have

$$\Sigma = AA^T$$

where

$$A = \begin{pmatrix} \alpha, 0 \\ \beta\gamma, \beta\sqrt{1 - \gamma^2} \end{pmatrix}$$

with

$$\alpha = \sqrt{[\Sigma]_1^1}, \beta = \sqrt{[\Sigma]_2^2} \text{ and } \gamma = \frac{[\Sigma]_1^2}{\sqrt{[\Sigma]_1^1 [\Sigma]_2^2}}$$

Then it can be easily observed that

$$\mathbb{E}_{(Z_x, Z_y) \sim \mathcal{N}(0, \Sigma)}[\varphi(Z_x)\varphi(Z_y)] = \mathbb{E}_{Z \sim \mathcal{N}(0, \mathrm{Id})}\left[\varphi(\alpha Z_1)\varphi\left(\beta\left(\gamma Z_1 + \sqrt{1 - \gamma^2}Z_2\right)\right)\right]$$

so that

$$
\begin{aligned}
|V_\varphi(\Sigma)| &= \left| \mathbb{E}_{Z\sim\mathcal{N}(0,\mathrm{Id})}\left[\varphi(\alpha Z_1)\varphi\left(\beta\left(\gamma Z_1 + \sqrt{1-\gamma^2}Z_2\right)\right)\right]\right| \\
&\leq \mathbb{E}_{Z\sim\mathcal{N}(0,\mathrm{Id})}\left[|\varphi(\alpha Z_1)|\left|\varphi\left(\beta\left(\gamma Z_1 + \sqrt{1-\gamma^2}Z_2\right)\right)\right|\right] \\
&\leq M^2\mathbb{E}_{Z\sim\mathcal{N}(0,\mathrm{Id})}\left[(1+|(\alpha Z_1)|)\left(1+\left|\left(\beta\left(\gamma Z_1 + \sqrt{1-\gamma^2}Z_2\right)\right)\right|\right)\right] \\
&= M^2\Big[1 + \alpha\mathbb{E}_{Z\sim\mathcal{N}(0,\mathrm{Id})}[|Z_1|] + \beta\mathbb{E}_{Z\sim\mathcal{N}(0,\mathrm{Id})}\left[\left|\gamma Z_1 + \sqrt{1-\gamma^2}Z_2\right|\right] \\
&\qquad + \alpha\beta\mathbb{E}_{Z\sim\mathcal{N}(0,\mathrm{Id})}\left[|Z_1|\left|\gamma Z_1 + \sqrt{1-\gamma^2}Z_2\right|\right]\Big] \\
&\leq M^2\Big[1 + (\alpha + \beta|\gamma|)\mathbb{E}_{Z\sim\mathcal{N}(0,\mathrm{Id})}[|Z_1|] + \beta\sqrt{1-\gamma^2}\mathbb{E}_{Z\sim\mathcal{N}(0,\mathrm{Id})}[|Z_2|] \\
&\qquad + \alpha\beta|\gamma|\mathbb{E}_{Z\sim\mathcal{N}(0,\mathrm{Id})}[|Z_1|^2] + \alpha\beta\sqrt{1-\gamma^2}\mathbb{E}_{Z\sim\mathcal{N}(0,\mathrm{Id})}[|Z_1 Z_2|]\Big] \\
&\leq \bar{M}\left(1 + \alpha + \beta|\gamma| + \beta\sqrt{1-\gamma^2} + \alpha\beta|\gamma| + \alpha\beta\sqrt{1-\gamma^2}\right)
\end{aligned}
$$

where the first inequality follows from Jensen's inequality, the second from Assumption A.1, the third from the triangle inequality, and the fourth from the fact that $\mathcal{N}(0, Id)$ has finite moments with $\bar{M}$ a constant incorporating M^2 and these bounds.

Using $\gamma^2 \leq 1$, for some constant $\tilde{M}$ one has

$$
|V_\varphi(\Sigma)| \quad \leq \quad \tilde{M}(1 + \alpha + \beta + \alpha\beta) = \tilde{M}(1+\alpha)(1+\beta)
$$

$\square$

We end the section showing the explicit characterization of the maps V_φ for some selected[7] activation functions. We write V_φ with the obvious meaning.

Proposition A.7. *Defining* $\gamma(\Sigma) := \frac{[\Sigma]_1^2}{\sqrt{[\Sigma]_1^1[\Sigma]_2^2}}$ *we have*

$$
\begin{aligned}
V_{id}(\Sigma) &= [\Sigma]_1^2 = [\Sigma]_2^1 \\
V_{ReLU}(\Sigma) &= \frac{1}{2\pi}\left(\pi + \frac{\sqrt{1-\gamma(\Sigma)^2}}{\gamma(\Sigma)} - arccos(\gamma(\Sigma))\right)[\Sigma]_1^2 \\
V_{erf}(\Sigma) &= \frac{2}{\pi}arcsin\left(\frac{[\Sigma]_1^2}{\sqrt{(0.5+[\Sigma]_1^1)(0.5+[\Sigma]_2^2)}}\right)
\end{aligned}
$$

(10)

Proof. See (Yang, 2019)[Facts B.2, B.3]. $\square$

B. Proofs for inhomogeneous controlled ResNets

In this section of the appendix we are going to prove all the results stated for the inhomogeneous case. The section will be subdivided in three main parts: in the first we consider the infinite-width-then-depth limit, in the second the infinite-depth-then-width one, in the final one we prove the commutativity of the integrals.

We start by recalling the defintion of the model.

Definition B.1 (Inhomogeneous controlled ResNets). Let $\mathcal{D}_M = \{0 = t_0 < \cdots < t_M = 1\}$ be a partition, $N \in \mathbb{N}$ be the width, and $\varphi : \mathbb{R} \to \mathbb{R}$ an activation function. Define a *randomly initialised, 1-layer inhomogeneous controlled ResNet* $\Psi^{M,N} : \mathbb{X} \to \mathbb{R}$ as follows

$$
\Psi_\varphi^{M,N}(x) := \left\langle \psi, \mathcal{S}_{t_M}^{M,N}(x)\right\rangle_{\mathbb{R}^N}
$$

---
[7]by the availability in the literature.

where $\langle \cdot, \cdot \rangle_{\mathbb{R}^N}$ is the L^2 inner product on $\mathbb{R}^N$, $\psi \in \mathbb{R}^N$ is a random vector with entries $[\psi]_\alpha \overset{\text{iid}}{\sim} \mathcal{N}(0, \frac{1}{N})$, and where the random functions $\mathcal{S}_{t_i}^{M,N} : \mathbb{X} \to \mathbb{R}^N$ satisfy the following recursive relation

$$\mathcal{S}_{t_{i+1}}^{M,N} = \mathcal{S}_{t_i}^{M,N} + \sum_{j=1}^{d} \left(A_{j,i} \varphi(\mathcal{S}_{t_i}^{M,N}) + b_{j,i} \right) \Delta x_{t_{i+1}}^{j}$$

$$\mathcal{S}_{t_0}^{M,N} = a \quad \text{and} \quad \Delta x_{t_i}^{j} = (x_{t_i}^{j} - x_{t_{i-1}}^{j})$$

for $i = 0, ..., M$, with initial condition $[a]_\alpha \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_a^2)$, and Gaussian weights $A_{k,l} \in \mathbb{R}^{N \times N}$ and biases $b_{k,l} \in \mathbb{R}^N$ sampled independently according to

$$[A_{j,i}]_\alpha^\beta \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \frac{\sigma_A^2}{N \Delta t_i} \right), \quad [b_{j,i}]_\alpha \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \frac{\sigma_b^2}{\Delta t_i} \right)$$

with time step $\Delta t_i = (t_i - t_{i-1}) > 0$ and parameters $\sigma_a, \sigma_A > 0$ and $\sigma_b \geq 0$.

B.1. The infinite-width-then-depth regime

The main goal in this subsection is to prove the first part of Theorem 3.1, which we restate here:

Theorem B.2. *Let $\{\mathcal{D}_M\}_{M \in \mathbb{N}}$ be a sequence of partitions of $[0,1]$ such that $|\mathcal{D}_M| \downarrow 0$. Let the activation function $\varphi : \mathbb{R} \to \mathbb{R}$ be linearly bounded, absolutely continuous and with exponentially bounded derivative. For any subset of paths $\mathcal{X} = \{x_1, \ldots, x_n\} \subset \mathbb{X}$ the following convergence in distribution holds*

$$\lim_{M \to \infty} \lim_{N \to \infty} \Psi_\varphi^{M,N}(\mathcal{X}) = \mathcal{N}(0, \kappa_\varphi(\mathcal{X}, \mathcal{X})) \tag{11}$$

where the positive semidefinite kernel $\kappa_\varphi : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ is defined for any two paths $x, y \in \mathbb{X}$ as $\kappa_\varphi(x, y) = \kappa_\varphi^{x,y}(1)$, where $\kappa_\varphi^{x,y} : [0,1] \to \mathbb{R}$ is the unique solution of the following differential equation

$$\partial_t \kappa_\varphi^{x,y} = \left[\sigma_A^2 V_\varphi \left(\begin{pmatrix} \kappa_\varphi^{x,x} & \kappa_\varphi^{x,y} \\ \kappa_\varphi^{x,y} & \kappa_\varphi^{y,y} \end{pmatrix} \right) + \sigma_b^2 \right] \langle \dot{x}_t, \dot{y}_t \rangle_{\mathbb{R}^d} \tag{12}$$

with initial condition $\kappa_\varphi^{x,y}(0) = \sigma_a^2$.

Remark. As mentioned in the paper, this convergence is equivalent to say that the sequence of random functions $\Psi_\varphi^{M,N} : \mathbb{X} \to \mathbb{R}$ convergence weakly to a GP with zero mean function and with kernel κ_φ.

We split the lengthy proof in two parts:

1. The infinite width-convergence of the finite dimensional distributions

$$\Psi_\varphi^{M,N}(\mathcal{X}) \xrightarrow{N \to \infty} \mathcal{N}(0, \kappa_{\mathcal{D}_M}(\mathcal{X}, \mathcal{X}))$$

for a fixed depth M to those of a Gaussian process $\mathcal{GP}(0, \kappa_{\mathcal{D}_M})$ defined by a kernel computed as the final value of a finite difference scheme on the partition $\mathcal{D}_M$. This will be done using Tensor Programs (Yang, 2019), and is the content of subsection B.1.1.

2. The infinite-depth (uniform) convergence of the discrete kernels $\kappa_{\mathcal{D}_M}$ to a limiting kernel κ_φ which solves the differential equation (12). This will be done in subsection B.1.2.

B.1.1. INFINITE-WIDTH LIMIT WITH FIXED DEPTH

In the next theorem, we show that finite width inhomogeneous ResNets converge in distribution to GPs with discrete kernels satisfying some difference equations.

Theorem B.3. *Let $\mathcal{D}_M = \{0 = t_0, \ldots, t_i, \ldots, t_M = 1\}$ be a fixed partition of $[0,1]$. Let the activation function $\varphi : \mathbb{R} \to \mathbb{R}$ be linearly bounded [8]. For any subset $\mathcal{X} = \{x_1, \ldots, x_n\} \subset \mathbb{X}$ the following convergence in distribution holds*

$$\lim_{N \to \infty} \Psi_\varphi^{M,N}(\mathcal{X}) = \mathcal{N}(0, \kappa_{\mathcal{D}_M}(\mathcal{X}, \mathcal{X}))$$

---

[8]in the sense that $\exists C > 0$. such that $|\phi(x)| \leq C(1 + |x|)$.

where for any two paths $x, y \in \mathbb{X}$ the discrete kernel $\kappa_{\mathcal{D}_M}(x,y) := \kappa_{\mathcal{D}_M}^{x,y}(1)$ where $\kappa_{\mathcal{D}_M}^{x,y}$ satisfies the following difference equation

$$\kappa_{\mathcal{D}_M}^{x,y}(t_i) = \kappa_{\mathcal{D}_M}^{x,y}(t_{i-1}) + \left(\sigma_A^2 V_\varphi(\Sigma_{\mathcal{D}_M}^{x,y}(t_{i-1})) + \sigma_b^2\right)\frac{\langle \Delta x_{t_i}, \Delta y_{t_i}\rangle_{\mathbb{R}^d}}{\Delta t_i} \tag{13}$$

with $\kappa_{\mathcal{D}_M}^{x,y}(0) = \sigma_a^2$ and where

$$\Sigma_{\mathcal{D}_M}^{x,y}(t) = \begin{pmatrix} \kappa_{\mathcal{D}_M}^{x,x}(t) & \kappa_{\mathcal{D}_M}^{x,y}(t) \\ \kappa_{\mathcal{D}_M}^{x,y}(t) & \kappa_{\mathcal{D}_M}^{y,y}(t) \end{pmatrix} \tag{14}$$

Proof. We use (Yang, 2019)[Corollary 5.5] applied to the Tensor Program of Algorithm 1 where the input variables are independently sampled according to

$$[A_{j,i}]_\alpha^\beta \sim \mathcal{N}(0, \frac{\sigma_A^2}{N\Delta t_i}), \quad [v]_\alpha \sim \mathcal{N}(0,1), \quad [a]_\alpha \sim \mathcal{N}(0, \sigma_a^2), \quad [b_{j,i}]_\alpha \sim \mathcal{N}(0, \frac{\sigma_b^2}{\Delta t_i})$$

The above sampling scheme follows (Yang, 2019)[Assumption 5.1]. Furthermore, linearly bounded functions are controlled in the sense of (Yang, 2019)[Definition 5.3] since for all $x \in \mathbb{R}$ one has $|\phi(x)| \leq C(1 + |x|) \leq e^{|x| + log(C)}$. Thus, we are under the needed assumptions to apply (Yang, 2019)[Corollary 5.5]. This result states that the output vector of the discrete controlled ResNet in Algorithm 1, on the partition $\mathcal{D}_M$ converges in law, as $N \to \infty$, to a Gaussian distribution $\mathcal{N}(0, K)$ where for $i, j = 1, ..., n$

$$[K]_i^j = \mathbb{E}_{Z \sim \mathcal{N}(\mu, \Sigma)}\left[Z^{\mathcal{S}_M^{x_i}} Z^{\mathcal{S}_M^{x_j}}\right] = \Sigma(\mathcal{S}_M^{x_i}, \mathcal{S}_M^{x_j})$$

with μ, Σ computed according to (Yang, 2019)[Definition 5.2] and defined on the set of all G-vars in the program *i.e.*

$$\mu(g) = \begin{cases} \mu^{in}(g) & \text{if } g \text{ is \textbf{Input} G-var} \\ \sum_k a_k \mu(g_k) & \text{if } g \text{ is introduced as } \sum_k a_k g_k \text{ via \textbf{LinComb}} \\ 0 & \text{otherwise} \end{cases}$$

$$\Sigma(g, g') = \begin{cases} \Sigma^{in}(g, g') & \text{if both } g \text{ and } g' \text{ are \textbf{Input} G-var} \\ \sum_k a_k \Sigma(g_k, g') & \text{if } g \text{ is introduced as } \sum_k a_k g_k \text{ via \textbf{LinComb}} \\ \sum_k a_k \Sigma(g, g_k') & \text{if } g' \text{ is introduced as } \sum_k a_k g_k' \text{ via \textbf{LinComb}} \\ \sigma_W^2 \mathbb{E}_{Z \sim \mathcal{N}(\mu, \Sigma)}[\phi(Z)\phi'(Z)] & \text{if } g = Wh, g' = Wh' \text{ via \textbf{MatMul} w/ same } W \\ 0 & \text{otherwise} \end{cases}$$

where $h = \phi((g_k)_{k=1}^m)$ for some function ϕ and $\phi(Z) := \phi((Z^{g_k})_{k=1}^m)$, similarly for g'.

In our setting $\mu^{in} \equiv 0$ since all **Input** variables are independent, from which $\mu \equiv 0$; furthermore $\Sigma^{in}(g, g') = 0$ except if $g = g'$ when it takes values in $\{\sigma_a^2, \frac{\sigma_b^2}{t_l - t_{l-1}}, 1\}$ accordingly.

Following the rules of Σ, assuming $l_i, l_j \in \{1, \ldots, M\}$, we obtain

$$
\begin{aligned}
\Sigma(\mathcal{S}_{l_i}^{x_i}, \mathcal{S}_{l_j}^{x_j}) &= \Sigma(\mathcal{S}_{l_i-1}^{x_i}, \mathcal{S}_{l_j}^{x_j}) + \Sigma\left(\sum_{k=1}^d \gamma_{k,l_i}^i \Delta(x_i)_{t_{l_i}}^k, \mathcal{S}_{l_j}^{x_j}\right) \\
&= \Sigma(\mathcal{S}_{l_i-1}^{x_i}, \mathcal{S}_{l_j}^{x_j}) + \Sigma\left(\sum_{k=1}^d \gamma_{k,l_i}^i \Delta(x_i)_{t_{l_i}}^k, \mathcal{S}_{l_j-1}^{x_j}\right) \\
&\quad + \Sigma\left(\sum_{k=1}^d \gamma_{k,l_i}^i \Delta(x_i)_{t_{l_i}}^k, \sum_{l=1}^d \gamma_{m,l_j}^j \Delta(x_j)_{t_{l_j}}^l\right) \\
&= \Sigma(\mathcal{S}_{l_i-1}^{x_i}, \mathcal{S}_{l_j}^{x_j}) + \Sigma(\mathcal{S}_{l_i}^{x_i}, \mathcal{S}_{l_j-1}^{x_j}) - \Sigma(\mathcal{S}_{l_i-1}^{x_i}, \mathcal{S}_{l_j-1}^{x_j}) \\
&\quad + \sum_{k,m=1}^d \Sigma(\gamma_{k,l_i}^i, \gamma_{m,l_j}^j)\Delta(x_i)_{t_{l_i}}^k \Delta(x_j)_{t_{l_j}}^m
\end{aligned}
$$

Now

$$\Sigma(\gamma_{k,l_i}^i, \gamma_{m,l_j}^j) = \delta_{k,m}\delta_{i,j}\frac{\sigma_A^2}{t_{l_i} - t_{l_i-1}}\mathbb{E}[\varphi(Z_1)\varphi(Z_2)] + \Sigma(b_{k,l_i}, b_{m,l_j})$$

$$= \frac{\delta_{k,m}\delta_{i,j}}{t_{l_i} - t_{l_i-1}}\left[\sigma_A^2\mathbb{E}[\varphi(Z_1)\varphi(Z_2)] + \sigma_b^2\right]$$

where $[Z_1, Z_2]^\top \sim \mathcal{N}(0, \tilde{\Sigma}_{l_i-1}(x_i, x_j))$ with

$$\tilde{\Sigma}_l(x_i, x_j) = \begin{pmatrix} \Sigma(\mathcal{S}_l^{x_i}, \mathcal{S}_l^{x_i}) & \Sigma(\mathcal{S}_l^{x_i}, \mathcal{S}_l^{x_j}) \\ \Sigma(\mathcal{S}_l^{x_i}, \mathcal{S}_l^{x_j}) & \Sigma(\mathcal{S}_l^{x_j}, \mathcal{S}_l^{x_j}) \end{pmatrix}$$

In particular we see that if $l_i \neq l_j$ then $\Sigma(\mathcal{S}_{l_i}^{x_i}, \mathcal{S}_{l_j}^{x_j}) = \Sigma(\mathcal{S}_{l_{i\wedge j}}^{x_i}, \mathcal{S}_{l_{i\wedge j}}^{x_j})$. Thus if we set, for $t_{l_i} \in \mathcal{D}_M$,

$$\kappa_{\mathcal{D}_M}^{x_i, x_j}(t_{l_i}) := \Sigma(\mathcal{S}_{l_i}^{x_i}, \mathcal{S}_{l_i}^{x_j})$$

we get

$$\kappa_{\mathcal{D}_M}^{x_i, x_j}(t_{l_i}) = \kappa_{\mathcal{D}_M}^{x_i, x_j}(t_{l_i-1})$$

$$+ \sum_{k=1}^d \left(\sigma_A^2\mathbb{E}_{(Z_x, Z_y)\sim\mathcal{N}(0,\tilde{\Sigma}_{l_i-1}(x_i,x_j))}[\varphi(Z_x)\varphi(Z_y)] + \sigma_b^2\right)\frac{\Delta(x_i)_{t_{l_i}}^k \Delta(x_j)_{t_{l_i}}^k}{t_{l_i} - t_{l_i-1}}$$

which is exactly what Equation 13 states. Then note how

$$\Sigma(\mathcal{S}_0^{x_i}, \mathcal{S}_0^{x_j}) = \sigma_a^2$$

Thus finally we can conclude and write the entries of the matrix $\tilde{K}$ as

$$[K]_i^j = \kappa_{\mathcal{D}_M}^{x_i, x_j}(1)$$

$\square$

Algorithm 1 $\mathcal{S}_1^{M,N}$ as Nestor program

Input: $\mathcal{S}_0 : \mathsf{G}(N)$ ⊳ *initial value*
Input: $(b_1, \ldots, b_d) : \mathsf{G}(N)$ ⊳ *biases*
Input: $(A_{1,l}, \ldots, A_{d,l})_{l=1,\cdots,M} : \mathsf{A}(N, N)$ ⊳ *matrices*
Input: $v : \mathsf{G}(N)$ ⊳ *readout layer weights*
for $i = 1, \ldots, n$ **do**
  // Compute $\mathcal{S}_1^{M,N}(x_i)$ (here $\mathcal{S}_0^{x_i}$ is to be read as $\mathcal{S}_0$)
  for $l = 1, \ldots, M$ **do**
    for $k = 1, \ldots, d$ **do**
      $\alpha_{k,l}^i := \varphi(\mathcal{S}_{l-1}^{x_i}) : \mathsf{H}(N)$ ⊳ *by Nonlin;*
      $\beta_{k,l}^i := A_{k,l}\alpha_{k,l}^i : \mathsf{G}(N)$ ⊳ *by Matmul;*
      $\gamma_{k,l}^i := \beta_{k,l}^i + b_{k,l} : \mathsf{G}(N)$ ⊳ *by LinComb;*
    end for
    $\mathcal{S}_l^{x_i} := \mathcal{S}_{l-1}^{x_i} + \sum_{k=1}^d \gamma_{k,l}^i[(x_i)_{t_l}^k - (x_i)_{t_{l-1}}^k] : \mathsf{G}(N)$ ⊳ *by LinComb;*
  end for
end for
Output: $(v^T\mathcal{S}_M^{x_i}/\sqrt{N})_{i=1,\ldots,n}$

Remark. There are two things to notice, done in the above proof in order to satisfy the required formalism:

- In the program the output projector v is sampled according to $\mathcal{N}(0,1)$ while the original $\phi \sim \mathcal{N}(0, \frac{1}{N})$. This does not pose any problems since the output of the formal programs uses $v/\sqrt{N} \sim \mathcal{N}(0, \frac{1}{N})$.

- The input paths x_i enter program 1 not as *Inputs* but as coefficients of *LinComb*, this means that for any choice of input paths we must formally consider different algorithms. In any case, for any possible choice, the result has always the same functional form; hence *a posteriori* it is legitimate to think about *one* algorithm.

Actually we have proved the even stronger statement, in the sense that the previous result holds for intermediate times too:

Corollary B.4. *For all $t_m, t_n \in \mathcal{D}_M$ one has the following distributional limit*

$$\left\langle \psi^N, \mathcal{S}_{t_m}^{N,M}(\mathcal{X}) \right\rangle_{\mathbb{R}^N} \left\langle \psi^N, \mathcal{S}_{t_n}^{N,M}(\mathcal{X}) \right\rangle_{\mathbb{R}^N} \xrightarrow[N\to\infty]{} \mathcal{N}(0, \kappa_{\mathcal{D}_M}^{\mathcal{X},\mathcal{X}}(t_m \wedge t_n))$$

and the matrices $\Sigma_{\mathcal{D}_M}^{x,y}(t_n)$ are always in PSD_2.

B.1.2. UNIFORM CONVERGENCE OF DISCRETE KERNELS

We now prove the convergence of the kernels $\kappa_{\mathcal{D}_M}^{x,y} : \mathcal{D}_M \to \mathbb{R}$ established in the previous section to a unique limiting kernel $\kappa_\varphi^{x,y} : [0,1] \to \mathbb{R}$ as $|\mathcal{D}_M| \to 0$. We first extend the discrete kernels $\kappa_{\mathcal{D}_M}^{x,y}$ to maps defined on $[0,1]$ in two ways.

Definition B.5. We extend the map $\kappa_{\mathcal{D}_M}^{x,y} : \mathcal{D}_M \to \mathbb{R}$ to the whole interval $[0,1]$ in two ways: for any $t \in [t_m, t_{m+1})$

1. (piecewise linear interpolation) using a slight abuse of notation that overwrites the previous one, define the map $\kappa_{\mathcal{D}_M}^{x,y} : [0,1] \to \mathbb{R}$ as

$$\kappa_{\mathcal{D}_M}^{x,y}(t) = \kappa_{\mathcal{D}_M}^{x,y}(t_m) + \left(\sigma_A^2 V_\varphi \left(\Sigma_{\mathcal{D}_M}^{x,y}(t_m) \right) + \sigma_b^2 \right) \left\langle \frac{x_t - x_{t_m}}{t - t_m}, \frac{y_t - y_{t_m}}{t - t_m} \right\rangle_{\mathbb{R}^d} (t - t_m)$$

We extend in a similar way the matrix $\Sigma_{\mathcal{D}_M}^{x,y}$ defined in equation (14).

2. (piecewise constant interpolation) define the map $\tilde{\kappa}_{\mathcal{D}_M}^{x,y} : [0,1] \to \mathbb{R}$ as

$$\tilde{\kappa}_{\mathcal{D}_M}^{x,y}(t) = \kappa_{\mathcal{D}_M}^{x,y}(t_m).$$

and similarly for the matrix $\tilde{\Sigma}_{\mathcal{D}_M}^{x,y}$.

Remark. It is important to consider both these types of extensions. The *piecewise linear*, being continuous, is used to prove uniform convergence to a limiting map in the space of continuous functions $C^0([0,1];\mathbb{R})$. The *piecewise constant* is proved to converge to the same object, this time in $L^\infty([0,1];\mathbb{R})$ since it's not continuous, and is well suited to prove how the positive semidefiniteness properties of the discrete kernels pass to the limit.

Theorem B.6. *Fix a sequence $\{\mathcal{D}_M\}_{M\in\mathbb{N}}$ of partitions of $[0,1]$ with $|\mathcal{D}_M| \to 0$ as $M \to \infty$. Then, for any two paths $x, y \in \mathbb{X}$, the sequence of functions $\{\kappa_{\mathcal{D}_M}^{x,y}\}_M$ converges uniformly on $C^0([0,1];\mathbb{R})$ to the unique solution $\kappa_\varphi^{x,y} : [0,1] \to \mathbb{R}$ of the following differential equation*

$$\kappa_\varphi^{x,y}(t) = \sigma_a^2 + \int_0^t (\sigma_A^2 V_\varphi(\Sigma_\varphi^{x,y}(s)) + \sigma_b^2) \langle \dot{x}_s, \dot{y}_s \rangle_{\mathbb{R}^d} \, ds \tag{15}$$

with

$$\Sigma_\varphi^{x,y}(t) = \begin{pmatrix} \kappa_\varphi^{x,x}(t), \kappa_\varphi^{x,y}(t) \\ \kappa_\varphi^{x,y}(t), \kappa_\varphi^{y,y}(t) \end{pmatrix}$$

One of the main difficulties when dealing with Equation (15) is making sure that the matrix $\Sigma_\varphi^{x,y}$ stays in PSD_2 for all times $t \in [0,1]$, in order to have $V_\varphi(\Sigma^{x,y}(t))$ well defined. This is clear if $x = y$ when $\Sigma_\varphi^{x,x}(t) = \kappa_\varphi^{x,x}(t)\mathbf{1}$, and one can just use (Friz & Victoir, 2010)[Theorem 3.7] to conclude, but it is not in the general case. One possible way to tackle this problem would be to consider the triplet $(\kappa_\varphi^{x,x}, \kappa_\varphi^{x,y}, \kappa_\varphi^{y,y}) : [0,1] \to \mathbb{R}^3$ and find it as the solution of (15) on a submanifold of $\mathbb{R}^3$. We decided to employ a more elementary technique which, unlike this "PDE on manifold" one, extends to the homogeneous case too, and we will find $\Sigma_\varphi^{x,y} : [0,1] \to \mathbb{R}^4$ as uniform limit of matrices in PSD_2.

The main idea for proving Theorem B.6 will be to show that the sequence $\{\kappa_{\mathcal{D}_M}^{x,y}\}_M$ is uniformly bounded and uniformly equicontinuous so that Ascoli-Arzelà theorem applies. One then proves that the limit of the resulting subsequence is the unique solution of Equation (15) and that the whole sequence converges to it. Before proving Theorem B.6 we need several lemmas.

The first step is to establish a uniform lower bound.

Lemma B.7 (Uniform lower bounds). *For any $x \in \mathbb{X}$, any partition $\mathcal{D}$*

$$\|\kappa_{\mathcal{D}}^{x,x}\|_{\infty,[0,1]} := \sup_{t \in [0,1]} |\kappa_{\mathcal{D}}^{x,x}(t)| \geq \sigma_a^2$$

and similarly for $\tilde{\kappa}_{\mathcal{D}}^{x,x}$.

Proof. Setting, for any $t_m \in \mathcal{D}$ by definition of the kernel $\kappa_{\mathcal{D}}^{x,x}$

$$\kappa_{\mathcal{D}}^{x,x}(t_m) = \sigma_a^2 + \sum_{0 \leq l < m} \left(\sigma_A^2 V_\varphi\left(\Sigma_{\mathcal{D}}^{x,x}(t_l)\right) + \sigma_b^2\right) \left|\frac{\Delta x_{t_{l+1}}}{\Delta t_{l+1}}\right|^2 \Delta t_{l+1}$$

Recall the definition of the map V_φ in Lemma A.5. For any $t_l \in \mathcal{D}$ we note that $V_\varphi(\Sigma_{\mathcal{D}}^{x,x}(t_l)) = \mathbb{E}[\varphi(\sqrt{\kappa_{\mathcal{D}}^{x,x}(t_l)})^2]$ for $Z \sim \mathcal{N}(0,1)$ since we have that $\Sigma_{\mathcal{D}}^{x,x}(t_l) = \kappa_{\mathcal{D}}^{x,x}(t_l)\mathbf{1}$ with $\mathbf{1} \in R^{2 \times 2}$ the matrix with all entries equal to 1. Thus $V_\varphi(\Sigma_{\mathcal{D}}^{x,x}(t_l)) \geq 0$ and we can conclude that $\tilde{\kappa}_{\mathcal{D}}^{x,x}(t) \geq \sigma_a^2$ since we are summing to σ_a^2 only the non-negative terms

$$\left(\sigma_A^2 V_\varphi\left(\Sigma_{\mathcal{D}}^{x,x}(t_l)\right) + \sigma_b^2\right) \left|\frac{\Delta x_{t_{l+1}}}{\Delta t_{l+1}}\right|^2 \Delta t_{l+1}$$

For any $t \in [0,1]$, by definition, we can write

$$\kappa_{\mathcal{D}}^{x,x}(t) = \kappa_{\mathcal{D}}^{x,x}(t_m) + \left(\sigma_A^2 V_\varphi\left(\tilde{\Sigma}_{\mathcal{D}}^{x,x}(t_m)\right) + \sigma_b^2\right) \left|\frac{x_t - x_{t_m}}{t - t_m}\right|_{\mathbb{R}^d}^2 (t - t_m)$$

thus is the sum of $\kappa_{\mathcal{D}}^{x,x}(t_m) = \tilde{\kappa}_{\mathcal{D}}^{x,x}(t) \geq \sigma_a^2$ and the quantity

$$\left(\sigma_A^2 V_\varphi\left(\tilde{\Sigma}_{\mathcal{D}}^{x,x}(t_m)\right) + \sigma_b^2\right) \left|\frac{x_t - x_{t_m}}{t - t_m}\right|_{\mathbb{R}^d}^2 (t - t_m)$$

which is ≥ 0 by the same arguments as above. $\qquad\square$

The second step is to establish a uniform upper bound.

Lemma B.8 (Uniform upper bounds). *For any $x \in \mathbb{X}$ and any partition $\mathcal{D}$ there exists a constant $C_x > 0$ independent of $\mathcal{D}$ such that*

$$\|\tilde{\kappa}_{\mathcal{D}}^{x,x}\|_{\infty,[0,1]} \leq \|\kappa_{\mathcal{D}}^{x,x}\|_{\infty,[0,1]} \leq C_x$$

Proof. Note how all the values taken by $\tilde{\kappa}_{\mathcal{D}}^{x,x}$ are also taken by $\kappa_{\mathcal{D}}^{x,x}$ in the corresponding partition points, thus the first inequality is trivial. Let us find an upper bound $\tilde{C}_x$ for $\tilde{\kappa}_{\mathcal{D}}^{x,x}$ first, intuitively since $\kappa_{\mathcal{D}}^{x,x}$ cannot be far from $\tilde{\kappa}_{\mathcal{D}}^{x,x}$ the constant $\tilde{C}_x$ should help find the bound C_x.

Let $t \in [t_m, t_{m+1})$. Recall the definition of $\tilde{\kappa}_{\mathcal{D}}^{x,x}$

$$\begin{aligned}
\tilde{\kappa}_{\mathcal{D}}^{x,x}(t) &= \kappa_{\mathcal{D}}^{x,x}(t_m) \\
&= \sigma_a^2 + \sum_{0 \leq l < m} \left(\sigma_A^2 V_\varphi(\Sigma_{\mathcal{D}}^{x,x}(t_l)) + \sigma_b^2\right) \left\langle \frac{\Delta x_{t_{l+1}}}{\Delta t_{l+1}}, \frac{\Delta x_{t_{l+1}}}{\Delta t_{l+1}} \right\rangle \Delta t_{l+1}
\end{aligned}$$

We then have

$$
\begin{aligned}
\left\langle \frac{\Delta x_{t_{l+1}}}{\Delta t_{l+1}}, \frac{\Delta x_{t_{l+1}}}{\Delta t_{l+1}} \right\rangle \Delta t_{l+1} &= \left| \frac{\Delta x_{t_{l+1}}}{\Delta t_{l+1}} \right|^2_{\mathbb{R}^d} \Delta t_{l+1} \\
&= \left| \frac{1}{\Delta t_{l+1}} \int_{t_l}^{t_{l+1}} \dot{x}_t dt \right|^2_{\mathbb{R}^d} \Delta t_{l+1} \\
&\leq \left(\frac{1}{\Delta t_{l+1}} \int_{t_l}^{t_{l+1}} |\dot{x}_t|_{\mathbb{R}^d} dt \right)^2 \Delta t_{l+1} \\
&\leq \frac{1}{\Delta t_{l+1}} \int_{t_l}^{t_{l+1}} |\dot{x}_t|^2_{\mathbb{R}^d} dt \Delta t_{l+1} \\
&= \int_{t_l}^{t_{l+1}} |\dot{x}_t|^2_{\mathbb{R}^d} dt
\end{aligned}
$$

where the last inequality is by Jensen's inequality. In addition, by Lemma A.5 there exists a positive constant $\tilde{M}$ such that

$$
|V_\varphi(\Sigma_{\mathcal{D}}^{x,x}(t_l))| \leq \tilde{M}(1 + \sqrt{\kappa_{\mathcal{D}}^{x,x}(t_l)})^2 \leq 2\tilde{M}(1 + \kappa_{\mathcal{D}}^{x,x}(t_l)).
$$

Thus

$$
\begin{aligned}
\tilde{\kappa}_{\mathcal{D}}^{x,x}(t) &\leq \sigma_a^2 + \sum_{0 \leq l < m} (2\sigma_A^2 \tilde{M}(1 + \kappa_{\mathcal{D}}^{x,x}(t_l)) + \sigma_b^2) \int_{t_l}^{t_{l+1}} |\dot{x}_t|^2_{\mathbb{R}^d} dt \\
&= \sigma_a^2 + \int_0^{t_m} (2\sigma_A^2 \tilde{M}(1 + \tilde{\kappa}_{\mathcal{D}}^{x,x}(t)) + \sigma_b^2) |\dot{x}_t|^2_{\mathbb{R}^d} dt
\end{aligned}
$$

By Gronwall inequality ((Friz & Victoir, 2010), Lemma 3.2) we have

$$
1 + \tilde{\kappa}_{\mathcal{D}}^{x,x}(t) \leq (1 + \sigma_a^2 + \sigma_b^2 \|x\|_{\mathbb{X}}^2) \exp\{2\sigma_A^2 \tilde{M} \|x\|_{\mathbb{X}}^2\}
$$

hence the statement of this lemma holds for $\tilde{\kappa}_{\mathcal{D}}^{x,x}$ with the constant

$$
\tilde{C}_x = (1 + \sigma_a^2 + \sigma_b^2 \|x\|_{\mathbb{X}}^2) e^{2\sigma_A^2 \tilde{M} \|x\|_{\mathbb{X}}^2} - 1
$$

To prove a similar inequality for $\kappa_{\mathcal{D}}^{x,x}$, consider

$$
\kappa_{\mathcal{D}}^{x,x}(t) = \kappa_{\mathcal{D}}^{x,x}(t_m) + \left(\sigma_A^2 V_\varphi \left(\Sigma_{\mathcal{D}}^{x,x}(t_m) \right) + \sigma_b^2 \right) \left| \frac{x_t - x_{t_m}}{t - t_m} \right|^2 (t - t_m)
$$

Then we have

$$
\begin{aligned}
|\kappa_{\mathcal{D}}^{x,x}(t)| &\leq |\kappa_{\mathcal{D}}^{x,x}(t_m)| + (2\sigma_A^2 \tilde{M}(1 + \tilde{C}_x) + \sigma_b^2) \left| \frac{x_t - x_{t_m}}{t - t_m} \right|^2 (t - t_m) \\
&\leq \tilde{C}_x + (2\sigma_A^2 \tilde{M}(1 + \tilde{C}_x) + \sigma_b^2) \|x\|_{\mathbb{X}}^2
\end{aligned}
$$

Hence the statement follows by setting

$$
C_x = \tilde{C}_x + (2\sigma_A^2 \tilde{M}(1 + \tilde{C}_x) + \sigma_b^2) \|x\|_{\mathbb{X}}^2
$$

$\square$

Remark. Note how C_x only depends on $\|x\|_{\mathbb{X}}$ and is increasing in it, thus this bound is uniform on bounded subsets of $\mathbb{X}$.

The following lemma shows that the kernels are in fact elements of $PSD(R)$.

Lemma B.9. *There exists a constant $R = R_x \in \mathbb{R}$ such that $\tilde{\Sigma}_{\mathcal{D}}^{x,x}(t), \Sigma_{\mathcal{D}}^{x,x}(t) \in \mathrm{PSD}_2(R)$ for every $\mathcal{D}$ and $t \in [0,1]$. Moreover, as before, this R_x only depends on $\|x\|_{\mathbb{X}}$ and is increasing in it.*

Proof. $\tilde{\Sigma}_{\mathcal{D}}^{x,x}(t)$ and $\Sigma_{\mathcal{D}}^{x,x}(t)$ are PSD_2 since they are of form $a\mathbf{1}$ for some $a > 0$. The diagonal elements $\tilde{\kappa}_{\mathcal{D}}^{x,x}(t), \kappa_{\mathcal{D}}^{x,x}(t)$ are bounded above by a common constant C_x and below by σ_a^2. We thus can simply choose $R_x = C_x \vee \sigma_a^{-2}$. $\qquad\square$

We now extend the results of Lemma B.8 and Lemma B.9 to the case $x \neq y$.

Lemma B.10. *Fix a $\alpha > 0$. There exist a constant C_α such that for all $x, y \in \mathbb{X}$ with $\|x\|_{\mathbb{X}}, \|y\|_{\mathbb{X}} \leq \alpha$ and all partitions $\mathcal{D}$ it holds*

$$\|\kappa_{\mathcal{D}}^{x,y}\|_{\infty,[0,1]} \leq C_\alpha$$

Moreover for $R_\alpha := C_\alpha \vee \sigma_a^{-2}$ we get

$$\tilde{\Sigma}_{\mathcal{D}}^{x,y}(t) \in \mathrm{PSD}_2(R_\alpha)$$

Proof. Remember how the maps $\tilde{\Sigma}_{\mathcal{D}}^{x,y}(t)$ are PSD_2 since their values are found as covariance matrices of Gaussian random variables with Tensor Program arguments in Theorem B.3 , in particular by positive semidefinitiveness and the previous bounds

$$|\tilde{\kappa}_{\mathcal{D}}^{x,y}(t)| \leq \sqrt{\tilde{\kappa}_{\mathcal{D}}^{x,x}(t)\tilde{\kappa}_{\mathcal{D}}^{y,y}(t)} \leq \sqrt{\tilde{C}_x\tilde{C}_y} \leq \tilde{C}_x \vee \tilde{C}_y$$

For $\kappa_{\mathcal{D}}^{x,y}(t)$ we proceed similarly to before: when $t \in [t_m, t_{m+1})$ one has

$$\begin{aligned}
|\kappa_{\mathcal{D}}^{x,y}(t)| &\leq |\kappa_{\mathcal{D}}^{x,y}(t_m)| \\
&\quad + \left|(\sigma_A^2\tilde{M}(1 + \sqrt{\tilde{\kappa}_{\mathcal{D}}^{x,x}(t)})(1 + \sqrt{\tilde{\kappa}_{\mathcal{D}}^{y,y}(t)}) + \sigma_b^2)\left\langle \frac{x_t - x_{t_m}}{t - t_m}, \frac{y_t - y_{t_m}}{t - t_m} \right\rangle_{\mathbb{R}^d} (t - t_m)\right| \\
&\leq \tilde{C}_x \vee \tilde{C}_y \\
&\quad + (2\sigma_A^2\tilde{M}(1 + \tilde{C}_x \vee \tilde{C}_y) + \sigma_b^2)|\left\langle \frac{x_t - x_{t_m}}{t - t_m}, \frac{y_t - y_{t_m}}{t - t_m} \right\rangle_{\mathbb{R}^d}|(t - t_m) \\
&\leq \tilde{C}_x \vee \tilde{C}_y + (2\sigma_A^2\tilde{M}(1 + \tilde{C}_x \vee \tilde{C}_y) + \sigma_b^2)(|\frac{x_t - x_{t_m}}{t - t_m}|^2 + |\frac{y_t - y_{t_m}}{t - t_m}|^2)(t - t_m) \\
&\leq \tilde{C}_x \vee \tilde{C}_y + (2\sigma_A^2\tilde{M}(1 + \tilde{C}_x \vee \tilde{C}_y) + \sigma_b^2)(\|x\|_{\mathbb{X}} + \|y\|_{\mathbb{X}}) \leq C_\alpha
\end{aligned}$$

where we have used $|\langle a, b\rangle| \leq |a||b| \leq (|a|^2 + |b|^2)$ for $a, b \in \mathbb{R}^d$. The second part follows from the definition of $\mathrm{PSD}_2(R_\alpha)$. $\qquad\square$

We are now ready to prove Theorem B.6.

Proof of Theorem B.6.

Part I (Convergence $\sup_{[0,1]\times[0,1]} |\kappa_{\mathcal{D}_M}^{x,y}(t) - \tilde{\kappa}_{\mathcal{D}_M}^{x,y}(t)| \to 0$ as $|\mathcal{D}| \to 0$) The set $\{x, y\}$ is bounded in $\mathbb{X}$, we can thus fix two constants $C_{x,y}, R_{x,y}$ with the properties given in Lemma B.10. We will often, for ease of notation, refer to them as C and R. We have, for $t \in [t_m, t_{m+1})$, that

$$\begin{aligned}
&|\kappa_{\mathcal{D}_M}^{x,y}(t) - \tilde{\kappa}_{\mathcal{D}_M}^{x,y}(t)| \\
&\leq |(\sigma_A^2\tilde{M}(1 + \sqrt{\tilde{\kappa}_{\mathcal{D}}^{x,x}(t)})(1 + \sqrt{\tilde{\kappa}_{\mathcal{D}_M}^{y,y}(t)}) + \sigma_b^2)\left\langle \frac{x_t - x_{t_m}}{t - t_m}, \frac{y_t - y_{t_m}}{t - t_m} \right\rangle_{\mathbb{R}^d} (t - t_m)| \\
&\leq (2\sigma_A^2\tilde{M}(1 + C_{x,y}) + \sigma_b^2)(|\frac{x_t - x_{t_m}}{t - t_m}|^2 + |\frac{y_t - y_{t_m}}{t - t_m}|^2)(t - t_m) \\
&\leq (2\sigma_A^2\tilde{M}(1 + C_{x,y}) + \sigma_b^2)\int_{t_m}^t |\dot{x}_s|^2 + |\dot{y}_s|^2 ds
\end{aligned}$$

In particular as $|\mathcal{D}| \to 0$ we have

$$\sup_{[0,1]\times[0,1]} |\kappa_{\mathcal{D}_M}^{x,y}(t) - \tilde{\kappa}_{\mathcal{D}_M}^{x,y}(t)| \to 0$$

since, by dominated convergence, one has

$$\int_{t_m}^t |\dot{x}_s|^2 + |\dot{y}_s|^2 ds = \int_0^1 \mathbb{I}_{[t_m,t)}(|\dot{x}_s|^2 + |\dot{y}_s|^2)ds \to 0$$

Part II (Ascoli-Arzelà) By Lemma B.10 the sequence of functions $\{\kappa_{\mathcal{D}_M}^{x,y}\}_M$ is uniformly bounded. Assume $s < t$ then one has, with the same bounds just used, that

$$|\kappa_{\mathcal{D}_M}^{x,y}(t) - \kappa_{\mathcal{D}_M}^{x,y}(s)| \leq 2(2\sigma_A^2 \tilde{M}(1 + C_{x,y}) + \sigma_b^2) \int_s^t |\dot{x}_r|^2 + |\dot{y}_r|^2 ds$$

thus the sequence of functions $\{\kappa_{\mathcal{D}_M}^{x,y}\}_M$ is also uniformly equicontinuous, in fact the bound is independent from the partition. We can apply Ascoli-Arzelà to conclude that there exist a subsequence $\{\mathcal{D}_{M_k}\}$ with $\kappa_{\mathcal{D}_M}^{x,y}$ uniformly converging to a limiting $\kappa^{x,y} \in C^0([0,1];\mathbb{R})$. Since this must be the uniform limit of $\tilde{\kappa}_{\mathcal{D}_M}^{x,y}$ too, by the result of *Part* I, we obtain that the corresponding limiting matrices $\Sigma_\varphi^{x,y}(t) \in PSD_2(R_{x,y})$ for all $t \in [0,1]$ since the $\tilde{\Sigma}_\varphi^{x,y}(t)$ are and $PSD_2(R_{x,y})$ is closed [9].

Part III (Uniqueness of solutions to equation (15)) Here we prove that if the PDE (15) admits a solution this must be unique. Assume the existence of different solutions $K = (K_{x,x}, K_{x,y}, K_{y,y})$ and $G = (G_{x,x}, G_{x,y}, G_{y,y})$ with all the Σ^K and Σ^G in PSD_2. From Eq (15) it is clear that $K_{x,x}, K_{y,y}, G_{x,x}, G_{y,y} \geq \sigma_a^2$ and, by continuity, that they are bounded by some constant; thus all the Σ_K and Σ_G are in some in $PSD_2(\bar{R})$. Then by the Lipschitz property of V_φ one sees that

$$\|\Sigma_K^{x,y}(t) - \Sigma_G^{x,y}(t)\|_\infty \leq \int_0^t \sigma_A^2 k_{\bar{R}} \|\Sigma_K^{x,y}(t) - \Sigma_G^{x,y}(t)\|_\infty (|\langle \dot{x}_r, \dot{x}_r \rangle| + |\langle \dot{x}_r, \dot{y}_r \rangle| + |\langle \dot{y}_r, \dot{y}_r \rangle|)dr$$

thus $\|\Sigma_K^{x,y}(t) - \Sigma_G^{x,y}(t)\|_\infty = 0$ by Gronwall for all $t \in [0,1]$ *i.e* $K = G$.

Part IV (Limiting kernel solves equation (15)) We now need to prove that the limit $\kappa_\varphi^{x,y}$ of the subsequence $\{\kappa_{\mathcal{D}_M}^{x,y}\}_k$ solves the PDE, it will then follow that any sub-sequence $\{\kappa_{\mathcal{D}_M}^{x,y}\}$ admits a further sub-sequence converging to the same map $\kappa_\varphi^{x,y}$, giving us the convergence of the whole sequence. Thus without loss of generality we can assume in the sequel that the whole sequence converges.

Let us prove that the limit $\kappa_\varphi^{x,y}$ is, in fact, a solution of the PDE. Let $t \in [t_m, t_{m+1})$ for some fixed $\mathcal{D}$, then

$$\left| \kappa_\varphi^{x,y}(t) - \sigma_a^2 + \int_0^t (\sigma_A^2 V_\varphi(\Sigma_\varphi^{x,y}(s)) + \sigma_b^2) \langle \dot{x}_s, \dot{y}_s \rangle_{\mathbb{R}^d} ds \right|$$

$$\leq \left| \kappa_\varphi^{x,y}(t) - \kappa_{\mathcal{D}_M}^{x,y}(t) \right| + \sigma_A^2 \int_0^t |V_\varphi(\Sigma_\varphi^{x,y}(s)) - V_\varphi(\tilde{\Sigma}_{\mathcal{D}_M}^{x,y}(s))| |\langle \dot{x}_s, \dot{y}_s \rangle_{\mathbb{R}^d}| ds$$

$$+ \sum_{0 \leq l < m} \left| (\sigma_A^2 V_\varphi (\Sigma_{\mathcal{D}_M}^{x,x}(t_l)) + \sigma_b^2) \right| \left| \int_{t_l}^{t_{l+1}} \langle \dot{x}_s, \dot{y}_s \rangle_{\mathbb{R}^d} ds - \left\langle \frac{\Delta x_{t_{l+1}}}{\Delta t_{l+1}}, \frac{\Delta y_{t_{l+1}}}{\Delta t_{l+1}} \right\rangle \Delta t_{l+1} \right|$$

$$+ \left| (\sigma_A^2 V_\varphi (\Sigma_{\mathcal{D}_M}^{x,x}(t_m)) + \sigma_b^2) \right| \left| \int_{t_m}^t \langle \dot{x}_s, \dot{y}_s \rangle_{\mathbb{R}^d} ds - \left\langle \frac{x_t - x_{t_m}}{t - t_m}, \frac{y_t - y_{t_m}}{t - t_m} \right\rangle (t - t_m) \right|$$

$$\leq |\kappa_\varphi^{x,y}(t) - \kappa_{\mathcal{D}_M}^{x,y}(t)| + \sigma_A^2 \int_0^t |V_\varphi(\Sigma_\varphi^{x,y}(s)) - V_\varphi(\tilde{\Sigma}_{\mathcal{D}_M}^{x,y}(s))| |\langle \dot{x}_s, \dot{y}_s \rangle_{\mathbb{R}^d}| ds$$

$$+ \left(2\sigma_A^2 \tilde{M}(1 + C_{x,y}) + \sigma_b^2 \right) \left\{ \sum_{0 \leq l < m} \left| \int_{t_l}^{t_{l+1}} \langle \dot{x}_s, \dot{y}_s \rangle_{\mathbb{R}^d} ds - \left\langle \frac{\Delta x_{t_{l+1}}}{\Delta t_{l+1}}, \frac{\Delta y_{t_{l+1}}}{\Delta t_{l+1}} \right\rangle \Delta t_{l+1} \right| \right.$$

$$\left. + \left| \int_{t_m}^t \langle \dot{x}_s, \dot{y}_s \rangle_{\mathbb{R}^d} ds - \left\langle \frac{x_t - x_{t_m}}{t - t_m}, \frac{y_t - y_{t_m}}{t - t_m} \right\rangle (t - t_m) \right| \right\}$$

but

---

[9]This is important to have a candidate solution to Equation (15) which otherwise would not be well defined.

$$\left| \int_a^b \langle \dot{x}_s, \dot{y}_s \rangle_{\mathbb{R}^d} \, ds - \left\langle \frac{x_b - x_a}{b-a}, \frac{y_b - y_a}{b-a} \right\rangle (b-a) \right| = \left| \int_a^b \langle \dot{x}_s, \dot{y}_s \rangle_{\mathbb{R}^d} - \left\langle \frac{x_b - x_a}{b-a}, \dot{y}_s \right\rangle_{\mathbb{R}^d} ds \right|$$

$$= \left| \int_a^b \left\langle \dot{x}_s - \frac{x_b - x_a}{b-a}, \dot{y}_s \right\rangle_{\mathbb{R}^d} ds \right|$$

$$\leq \int_a^b \left| \dot{x}_s - \frac{x_b - x_a}{b-a} \right| |\dot{y}_s| ds$$

hence

$$\left| \kappa_\varphi^{x,y}(t) - \sigma_a^2 + \int_0^t (\sigma_A^2 V_\varphi(\Sigma_\varphi^{x,y}(s)) + \sigma_b^2) \langle \dot{x}_s, \dot{y}_s \rangle_{\mathbb{R}^d} \, ds \right|$$

$$\leq |\kappa_\varphi^{x,y}(t) - \kappa_{\mathcal{D}_M}^{x,y}(t)| + \sigma_A^2 \int_0^t |V_\varphi(\Sigma_\varphi^{x,y}(s)) - V_\varphi(\tilde{\Sigma}_{\mathcal{D}_M}^{x,y}(s))| |\langle \dot{x}_s, \dot{y}_s \rangle_{\mathbb{R}^d}| ds$$

$$+ \left(2\sigma_A^2 \tilde{M}(1 + C_{x,y}) + \sigma_b^2 \right) \left\{ \int_0^t \left(\mathbb{I}_{[t_m,t)} |\dot{x}_s - \frac{x_t - x_{t_m}}{t - t_m}| + \sum_{0 \leq l < m} \mathbb{I}_{[t_l, t_{l+1})} |\dot{x}_s - \frac{\Delta x_{t_{l+1}}}{\Delta t_{l+1}}| \right) |\dot{y}_s| ds \right\}$$

Now, considering the sequence $\mathcal{D}_M$, by convergence

$$|\kappa_\varphi^{x,y}(t) - \kappa_{\mathcal{D}_M}^{x,y}(t)| = o(1)$$

and

$$\int_0^t |V_\varphi(\Sigma^{x,y}(s)) - V_\varphi(\tilde{\Sigma}_{\mathcal{D}_M}^{x,y}(s))| |\langle \dot{x}_s, \dot{y}_s \rangle_{\mathbb{R}^d}| ds \leq \int_0^t k_R |\Sigma_\varphi^{x,y}(s) - \tilde{\Sigma}_{\mathcal{D}_M}^{x,y}(s)|_\infty |\langle \dot{x}_s, \dot{y}_s \rangle_{\mathbb{R}^d}| ds = o(1)$$

Moreover, setting m_M to be such that $t \in [t_{m_M}, t_{m_M+1})$ in $\mathcal{D}_M$, we have that by Lebesgue differentiation theorem

$$\left(\mathbb{I}_{[t_{m_M}^{\mathcal{D}_M}, t)}(s) |\dot{x}_s - \frac{x_t - x_{t_{m_M}^{\mathcal{D}_M}}}{t - t_{m_M}^{\mathcal{D}_M}}| + \sum_{0 \leq l < m_M} \mathbb{I}_{[t_l^{\mathcal{D}_M}, t_{l+1}^{\mathcal{D}_M})}(s) |\dot{x}_s - \frac{\Delta x_{t_{l+1}^{\mathcal{D}_M}}}{\Delta t_{l+1}^{\mathcal{D}_M}}| \right) \xrightarrow{M \to \infty} 0$$

almost surely as a function of s, thus using Dominated convergence we conclude that

$$\int_0^t \left(\mathbb{I}_{[t_{m_M}^{\mathcal{D}_M}, t)}(s) |\dot{x}_s - \frac{x_t - x_{t_{m_M}^{\mathcal{D}_M}}}{t - t_{m_M}^{\mathcal{D}_M}}| + \sum_{0 \leq l < m_M} \mathbb{I}_{[t_l^{\mathcal{D}_M}, t_{l+1}^{\mathcal{D}_M})}(s) |\dot{x}_s - \frac{\Delta x_{t_{l+1}^{\mathcal{D}_M}}}{\Delta t_{l+1}^{\mathcal{D}_M}}| \right) |\dot{y}_s| ds = o(1)$$

thus

$$\left| \kappa_\varphi^{x,y}(t) - \sigma_a^2 + \int_0^t (\sigma_A^2 V_\varphi(\Sigma_\varphi^{x,y}(s)) + \sigma_b^2) \langle \dot{x}_s, \dot{y}_s \rangle_{\mathbb{R}^d} \, ds \right| = 0$$

i.e

$$\kappa_\varphi^{x,y}(t) = \sigma_a^2 + \sigma_a^2 + \int_0^t (\sigma_A^2 V_\varphi(\Sigma_\varphi^{x,y}(s)) + \sigma_b^2) \langle \dot{x}_s, \dot{y}_s \rangle_{\mathbb{R}^d} \, ds$$

$\square$

B.1.3. Proof of Theorem 3.1: Part 1

It is finally time to prove the first part Theorem 3.1 in the main body of the paper, which we restated in the appendix as Theorem B.2.

Proof of Theorem B.2. The proof is now just a matter of combining Theorem B.3 and Theorem B.6. Under our hypotheses Theorem B.3 tells us that, for any subset $\mathcal{X} = \{x_1, \ldots, x_n\} \subset \mathbb{X}$, we have in distribution

$$\lim_{N \to \infty} \Psi_\varphi^{M,N}(\mathcal{X}) = \mathcal{N}(0, \kappa_{\mathcal{D}_M}(\mathcal{X}, \mathcal{X}))$$

where $\kappa_{\mathcal{D}_M}(x_\alpha, x_\beta) = \kappa_{\mathcal{D}_M}^{x_\alpha, x_\beta}(1)$ for all $\alpha, \beta = 1, \ldots, n$. Thus to conclude we just have to prove that, still in distribution, it holds

$$\lim_{M \to \infty} \mathcal{N}(0, \kappa_{\mathcal{D}_M}(\mathcal{X}, \mathcal{X})) = \mathcal{N}(0, \kappa_\varphi(\mathcal{X}, \mathcal{X}))$$

or equivalently that

$$\lim_{M \to \infty} \kappa_{\mathcal{D}_M}^{x_\alpha, x_\beta}(1) = \kappa_\varphi^{x_\alpha, x_\beta}(1)$$

This last needed limit follows from Theorem B.6 with the sequence $\{\mathcal{D}_M\}_{M \in \mathbb{N}}$. □

We can now study some particular cases.

Corollary B.11. *If $\varphi = id$, then*

$$\kappa_{id}^{x,y}(t) = \left(\sigma_a^2 + \frac{\sigma_b^2}{\sigma_A^2}\right) \exp\left\{\sigma_A^2 \int_{\eta=0}^s \langle \dot{x}_\eta, \dot{y}_\eta \rangle_{\mathbb{R}^d} \, d\eta\right\} - \frac{\sigma_b^2}{\sigma_A^2} \tag{16}$$

If $\varphi = ReLu$ and $x = y$, then

$$\kappa_{ReLu}^{x,x}(t) = \left(\sigma_a^2 + \frac{2\sigma_b^2}{\sigma_A^2}\right) \exp\left\{\frac{\sigma_A^2}{2} \int_{\eta=0}^s \|\dot{x}_\eta\|_{\mathbb{R}^d}^2 \, d\eta\right\} - \frac{2\sigma_b^2}{\sigma_A^2} \tag{17}$$

Proof. This is just a matter of computation. For the case $\varphi = id$ one has

$$\mathbb{E}_{Z \sim \mathcal{N}(0,\Sigma)}[\varphi(Z_1)\varphi(Z_2)] = \mathbb{E}_{Z \sim \mathcal{N}(0,\Sigma)}[Z_1 Z_2] = [\Sigma]_1^2$$

hence

$$\kappa_{id}^{x,y}(t) = \sigma_a^2 + \int_0^t \left[\sigma_A^2 \kappa_{id}^{x,y}(s) + \sigma_b^2\right] \langle \dot{x}_s, \dot{y}_s \rangle_{\mathbb{R}^d} \, ds$$

Notice then that substituting (16) for $K_s^{id}(x,y)$ in the integral leads to

$$\sigma_a^2 + \int_0^t \left[\sigma_A^2 \left\{(\sigma_a^2 + \frac{\sigma_b^2}{\sigma_A^2}) \exp\left\{\sigma_A^2 \int_{\eta=0}^s \langle \dot{x}_\eta, \dot{y}_\eta \rangle_{\mathbb{R}^d} \, d\eta\right\} - \frac{\sigma_b^2}{\sigma_A^2}\right\} + \sigma_b^2 \right] \langle \dot{x}_s, \dot{y}_s \rangle_{\mathbb{R}^d} \, ds$$

$$= \sigma_a^2 + \int_0^t \sigma_A^2 (\sigma_a^2 + \frac{\sigma_b^2}{\sigma_A^2}) \exp\left\{\sigma_A^2 \int_{\eta=0}^s \langle \dot{x}_\eta, \dot{y}_\eta \rangle_{\mathbb{R}^d} \, d\eta\right\} \langle \dot{x}_s, \dot{y}_s \rangle_{\mathbb{R}^d} \, ds$$

$$= \sigma_a^2 + \int_0^t (\sigma_a^2 + \frac{\sigma_b^2}{\sigma_A^2}) \partial_T \left(\exp\left\{\sigma_A^2 \int_{\eta=0}^T \langle \dot{x}_\eta, \dot{y}_\eta \rangle_{\mathbb{R}^d} \, d\eta\right\}\right)\big|_{T=s} ds$$

$$= \sigma_a^2 + (\sigma_a^2 + \frac{\sigma_b^2}{\sigma_A^2}) \left[\exp\left\{\sigma_A^2 \int_{\eta=0}^s \langle \dot{x}_\eta, \dot{y}_\eta \rangle_{\mathbb{R}^d} \, d\eta\right\} - 1 \right]$$

$$= (\sigma_a^2 + \frac{\sigma_b^2}{\sigma_A^2}) \exp\left\{\sigma_A^2 \int_{\eta=0}^s \langle \dot{x}_\eta, \dot{y}_\eta \rangle_{\mathbb{R}^d} \, d\eta\right\} - \frac{\sigma_b^2}{\sigma_A^2}$$

which means, by uniqueness of solutions, that the thesis holds.

For what concerns the case $\varphi = ReLU$ notice that for one dimensional Gaussian variables centered in the origin one has

$$\mathbb{E}_{Z \sim \mathcal{N}(0,\sigma^2)}[ReLU(Z)^2] = \frac{1}{2}\mathbb{E}_{Z \sim \mathcal{N}(0,\sigma^2)}[Z^2] = \frac{1}{2}\sigma^2$$

hence since

$$\mathbb{E}_{Z \sim \mathcal{N}(0, \kappa_{ReLU}^{x,x}(s)\mathbf{1})}[ReLU(Z_1)ReLU(Z_2)] = \mathbb{E}_{Z \sim \mathcal{N}(0, \kappa_{ReLU}^{x,x}(s))}[ReLU(Z)^2]$$

one has

$$\kappa_{ReLu}^{x,x}(t) = \sigma_a^2 + \int_0^t \left[\frac{\sigma_A^2}{2}\kappa_{ReLu}^{x,x}(s) + \sigma_b^2\right]\langle \dot{x}_s, \dot{x}_s\rangle_{\mathbb{R}^d} \, ds$$

which equals $\kappa_{id}^{x,x}(t; \sigma_a, \frac{\sigma_A}{\sqrt{2}}, \sigma_b)$. □

We conclude this section by proving Remark 4.2 about the path scaling symmetry mentioned in the main paper.

Lemma B.12. *For all choices $(\sigma_a, \sigma_A, \sigma_b)$ and for all φ as in Theorem 3.1 we have, with abuse of notation and the obvious meaning, that*

$$\kappa_\varphi^{x,y}(t; \sigma_a, \sigma_A, \sigma_b) = \kappa_\varphi^{\sigma_A x, \sigma_A y}(t; \sigma_a, 1, \sigma_b\sigma_A^{-1})$$

Proof. We have

$$\kappa_\varphi^{x,y}(t; \sigma_a, \sigma_A, \sigma_b)$$
$$= \sigma_a^2 + \int_{\eta=0}^s \left[\sigma_A^2 \mathbb{E}_{Z\sim\mathcal{N}(0,\Sigma_\varphi^{x,y}(\eta;\sigma_a,\sigma_A,\sigma_b))}[\varphi(Z_1)\varphi(Z_2)] + \sigma_b^2\right]\langle \dot{x}_\eta, \dot{y}_\eta\rangle_{\mathbb{R}^d} \, d\eta$$
$$= \sigma_a^2 + \int_{\eta=0}^s \sigma_A^2\left[\mathbb{E}_{Z\sim\mathcal{N}(0,\Sigma_\varphi^{x,y}(\eta;\sigma_a,\sigma_A,\sigma_b))}[\varphi(Z_1)\varphi(Z_2)] + \frac{\sigma_b^2}{\sigma_A^2}\right]\langle \dot{x}_\eta, \dot{y}_\eta\rangle_{\mathbb{R}^d} \, d\eta$$
$$= \sigma_a^2 + \int_{\eta=0}^s \left[\mathbb{E}_{Z\sim\mathcal{N}(0,\Sigma_\varphi^{x,y}(\eta;\sigma_a,\sigma_A,\sigma_b))}[\varphi(Z_1)\varphi(Z_2)] + \frac{\sigma_b^2}{\sigma_A^2}\right]\langle \sigma_A\dot{x}_\eta, \sigma_A\dot{y}_\eta\rangle_{\mathbb{R}^d} \, d\eta$$

and

$$\kappa_\varphi^{\sigma_A x, \sigma_A y}(t; \sigma_a, 1, \sigma_b\sigma_A^{-1})$$
$$= \sigma_a^2 + \int_{\eta=0}^s \left[\mathbb{E}_{Z\sim\mathcal{N}(0,\Sigma_\varphi^{\sigma_A x, \sigma_A y}(\eta;\sigma_a,1,\frac{\sigma_b}{\sigma_A}))}[\varphi(Z_1)\varphi(Z_2)] + \frac{\sigma_b^2}{\sigma_A^2}\right]\langle \sigma_A\dot{x}_\eta, \sigma_A\dot{y}_\eta\rangle_{\mathbb{R}^d} \, d\eta$$

Thus the respective triplets solve the same equation and we can conclude by uniqueness. □

B.2. The infinite-depth-then-width regime

As mentioned in the paper, it is natural to ask what happens if the order of the width-depth limits in is reversed.

B.2.1. PROOF OF THEOREM 3.3

We begin by proving Theorem 3.3 which we restate for the reader's convenience. We will follow arguments used in (Hayou, 2022), extending the results obtained therein.

Theorem B.13. *Let $\{\mathcal{D}_M\}_{M\in\mathbb{N}}$ be a sequence of partitions of $[0,1]$ such that $|\mathcal{D}_M| \downarrow 0$ as $M \to \infty$. Assume the activation function φ is Lipschitz and linearly bounded. Let $\rho_M(t) := \sup\{s \in \mathcal{D}_M : s \leq t\}$. For any path $x \in \mathbb{X} \cap C^{1,\frac{1}{2}}$, where $C^{1,\frac{1}{2}}$ denotes the set of C^1 paths with $\frac{1}{2}$-Hölder derivative, the $\mathbb{R}^N$-valued process $t \mapsto \mathcal{S}_{\rho_M(t)}^{M,N}(x)$ converges in distribution, as $M \to \infty$, to the solution $S^N(x)$ of the following SDE*

$$d\mathcal{S}_t^N(x) = \sum_{j=1}^d \frac{\sigma_A}{\sqrt{N}}\dot{x}_t^j dW_t^j \varphi(\mathcal{S}_t^N(x)) + \sigma_b\dot{x}_t^j dB_t^j \tag{18}$$

with $\mathcal{S}_0^N(x) = a$ and where $W^j \in \mathbb{R}^{N\times N}$ and $B^j \in \mathbb{R}^N$ are independent Brownian motions for $j \in \{1, ..., d\}$.

Proof. We will transform Equation (18) in a usual SDE form to then use the classical result (Kloeden & Platen, 1992)[Theorem 10.2.2] to prove the convergence of Euler Discretizations to the unique solution. We first show that Equation (18) can be re-written as follows

$$\mathcal{S}_t^N(x) = \sigma_a^2 + \int_0^t \sigma_x(s, \mathcal{S}_s^N(x)) dZ_s$$

for a Brownian Motion $Z_t \in \mathbb{R}^{\Gamma_{d,N}}$ with $\Gamma_{d,N} := dN(N+1)$.

Take in fact

$$Z_s := [([W_s^1]^1)^T, ([W_s^1]^2)^T, \ldots, ([W_s^1]^N)^T, (B_s^1)^T, ([W_s^2]^1)^T, \ldots, (B_s^d)^T]^T$$

or equivalently

$$[Z_s]_i = \begin{cases} [W_s^k]^m & \text{if } m = 1, \ldots, N \\ B_s^k & \text{if } m = N+1 \end{cases}$$

for $i = N(N+1)(k-1) + N(m-1) + 1, \ldots, N(N+1)(k-1) + Nm$ [10]. In addition, define $\sigma_x : [0,1] \times \mathbb{R}^N \to \mathbb{R}^{N \times \Gamma_{d,N}}$ as

$$[\sigma_x(s,y)]_i^j = \begin{cases} \frac{\sigma_A}{\sqrt{N}} \dot{x}_s^k [\varphi(y)]_m Id_{N \times N} & \text{if } m = 1, \ldots, N \\ \sigma_b \dot{x}_s^k Id_{N \times N} & \text{if } m = N+1 \end{cases}$$

for $i = 1, \ldots, N$ and $j = N(N+1)(k-1) + N(m-1) + 1, \ldots, N(N+1)(k-1) + Nm$. It is then just a matter of checking the required conditions to apply (Kloeden & Platen, 1992)[Theorem 10.2.2]:

1. There is a $K > 0$ such that $\forall t \in [0,1].\forall y, y' \in \mathbb{R}^N$ one has

$$\|\sigma(t,y) - \sigma(t,y')\|_F \leq K \|y - y'\|_{\mathbb{R}^N}$$

2. There is a $K' > 0$ such that $\forall t \in [0,1].\forall y \in \mathbb{R}^N$ one has

$$\|\sigma(t,y)\|_F \leq K'(1 + \|y\|_{\mathbb{R}^N})$$

3. There is a $K'' > 0$ such that $\forall t, s \in [0,1].\forall y \in \mathbb{R}^N$ one has

$$\|\sigma(t,y) - \sigma(s,y)\|_F \leq K''(1 + \|y\|_{\mathbb{R}^N})|t - s|^{\frac{1}{2}}$$

As a first step note how

$$\|\sigma(t,y)\|_F^2 = N \sum_{k=1}^d (\dot{x}_t^k)^2 \left(\sigma_b^2 + \sum_{m=1}^N \frac{\sigma_A^2}{N}([\varphi(y)]_m)^2\right) = N \|\dot{x}_t\|_{\mathbb{R}^d}^2 \left(\sigma_b^2 + \sigma_A^2 \frac{\|\varphi(y)\|_{\mathbb{R}^N}^2}{N}\right)$$

thus using sublinearity of φ for some $M > 0$ one has

$$\|\sigma(t,y)\|_F^2 \leq N \|\dot{x}_t\|_{\mathbb{R}^d}^2 \left(\sigma_b^2 + \frac{\sigma_A^2}{N} M(1 + \|y\|_{\mathbb{R}^N})^2\right)$$

from which we easily get condition (ii).

For condition (i) note that

$$\|\sigma(t,y) - \sigma(t,y')\|_F^2 = N \|\dot{x}_t\|_{\mathbb{R}^d}^2 \sigma_A^2 \frac{\|\varphi(y) - \varphi(y')\|_{\mathbb{R}^N}^2}{N}$$

thus one just uses Lipschitz property of φ with Lemma A.3.

Finally for (iii) one computes just as before

$$\|\sigma(t,y) - \sigma(s,y)\|_F^2 = N \|\dot{x}_t - \dot{x}_s\|_{\mathbb{R}^d}^2 \left(\sigma_b^2 + \sigma_A^2 \frac{\|\varphi(y)\|_{\mathbb{R}^N}^2}{N}\right)$$

and using Hölder property of x and linear bound on φ one has, for some $\tilde{M} > 0$

$$\|\sigma(t,y) - \sigma(s,y)\|_F^2 \leq N(\sigma_b^2 + \frac{\sigma_A^2}{N} M(1 + \|y\|_{\mathbb{R}^N})^2)\tilde{M}^2|t - s|$$

hence the concluding inequality

$$\|\sigma(t,y) - \sigma(s,y)\|_F \leq (N\tilde{M}(\sigma_b^2 + \frac{\sigma_A^2}{N} M))(1 + \|y\|_{\mathbb{R}^N})|t - s|^{\frac{1}{2}}$$

---

[10]meaning that $[Z_s]_i$ is considered as the element of $\mathbb{R}^N$ corresponding to these indices.

Because in general

$$\frac{\Delta x^k_{t_l^{\mathcal{D}_M}}}{\Delta t_l^{\mathcal{D}_M}} \neq \dot{x}^k_{t_{l-1}^{\mathcal{D}_M}}$$

the recursive equation

$$\mathcal{S}^{M,N}_{t_l^{\mathcal{D}_M}}(x) = \mathcal{S}^{M,N}_{t_{l-1}^{\mathcal{D}_M}}(x) + \sum_{k=1}^d \frac{\Delta x^k_{t_l^{\mathcal{D}_M}}}{\Delta t_l^{\mathcal{D}_M}} \left(\frac{\sigma_A}{\sqrt{N}} \sqrt{\Delta t_l^{\mathcal{D}_M}} W_{k,l} \varphi(\mathcal{S}^{M,N}_{t_{l-1}^{\mathcal{D}_M}}(x)) + \sigma_b \sqrt{\Delta t_l^M} B_{k,l} \right)$$

is not the Euler discretization of the above SDE. However, after classical though tedious calculations it is easy to show that

$$\lim_{M\to\infty} \sup_{t\in[0,1]} \mathbb{E}[|\tilde{\mathcal{S}}^{M,N}_t(x) - \mathcal{S}^{M,N}_t(x)|^2] = 0$$

where $\tilde{\mathcal{S}}^{M,N}_t(x)$ is defined as $\mathcal{S}^{M,N}_t(x)$ but substituting the difference quotients with the actual derivatives.

$\square$

Remark. By considering the concatenated system

$$[\mathcal{S}^N_t(x_i)]_{x_i\in\mathcal{X}} = [a]_{x_i\in\mathcal{X}} + \int_0^t [\sigma_{x_i}(s, \mathcal{S}^N_s(x_i))]_{x_i\in\mathcal{X}} dZ_s$$

we straightforwardly extend the previous result to the case with multiple inputs considered at the same time.

B.2.2. INFINITE-DEPTH-THEN-WIDTH LIMIT: $\varphi = id$

Here we directly prove that, in the case $\varphi = id$, the covariances of the infinite-depth networks converge to κ_{id} as the width increases without bounds.

Proposition B.14. *Let $\varphi = id$. For any subset $\mathcal{X} = \{x_1, \ldots, x_n\} \subset \mathbb{X}$. Then for all integers $N \geq 1$ and for all $k, m = 1, \ldots, d$ that the following convergence holds*

$$\lim_{M\to\infty} \mathbb{E}\left[\Psi^{M,N}_{id}(\mathcal{X}) \Psi^{M,N}_{id}(\mathcal{X}) \right] = \kappa_{id}(\mathcal{X}, \mathcal{X})$$

Proof. To start notice that

$$\mathbb{E}[\Psi^{M,N}_{id}(x_k)\Psi^{M,N}_{id}(x_m)] = \frac{1}{N}\mathbb{E}\left[\left\langle \mathcal{S}^{M,N}_1(x_k), \mathcal{S}^{M,N}_1(x_m) \right\rangle \right]$$

Thanks to the extension of Theorem B.13 to the multi-input case we have

$$\lim_{M\to\infty} \frac{1}{N}\mathbb{E}\left[\left\langle \mathcal{S}^{M,N}_1(x_k), \mathcal{S}^{M,N}_1(x_m) \right\rangle \right] = \frac{1}{N}\mathbb{E}\left[\left\langle \mathcal{S}^N_1(x_k), \mathcal{S}^N_1(x_m) \right\rangle \right]$$

and for $x, y \in \mathcal{X}$

$$\frac{1}{N}\mathbb{E}\left[\langle \mathcal{S}_1^N(x), \mathcal{S}_1^N(y)\rangle\right]$$

$$=\frac{1}{N}\mathbb{E}\left[\left\langle a + \sum_{i=1}^{d}\int_0^1 \frac{\sigma_A}{\sqrt{N}}\dot{x}_t^i dW_t^i \mathcal{S}_t^N(x) + \sigma_b \dot{x}_t^i dB_t^i, \mathcal{S}_1^N(y)\right\rangle\right]$$

$$=\frac{1}{N}\mathbb{E}\left[\|a\|_{\mathbb{R}^N}^2\right] + \frac{1}{N}\sum_{i,j=1}^{d}\sigma_b^2\mathbb{E}\left[\left\langle \int_0^1 \dot{x}_t^i dB_t^i, \int_0^1 \dot{y}_t^j dB_t^j\right\rangle\right]$$

$$+\frac{1}{N}\sum_{i,j=1}^{d}\frac{\sigma_A^2}{N}\mathbb{E}\left[\left\langle \int_0^1 \dot{x}_t^i dW_t^i \mathcal{S}_t^N(x), \int_0^1 \dot{y}_t^j dW_t^j \mathcal{S}_t^N(y)\right\rangle\right]$$

$$=\frac{1}{N}(N\sigma_a^2) + \sum_{i,j=1}^{d}\frac{\sigma_b^2}{N}\sum_{\alpha=1}^{N}\mathbb{E}\left[\int_0^1 \dot{x}_t^i d[B_t^i]_\alpha \cdot \int_0^1 \dot{y}_t^j d[B_t^j]_\alpha\right]$$

$$+\frac{1}{N}\sum_{i,j=1}^{d}\frac{\sigma_A^2}{N}\sum_{\alpha,\beta,\gamma=1}^{N}\mathbb{E}\left[\int_0^1 \dot{x}_t^i[\mathcal{S}_t^N(x)]_\alpha d[W_t^i]_\gamma^\alpha \cdot \int_0^1 \dot{y}_t^j[\mathcal{S}_t^N(y)]_\beta d[W_t^j]_\gamma^\beta\right]$$

$$=\sigma_a^2 + \frac{\sigma_b^2}{N}\sum_{l=1}^{N}\sum_{i=1}^{d}\mathbb{E}\left[\int_0^1 \dot{x}_t^i\dot{y}_t^i dt\right]$$

$$+\frac{\sigma_A^2}{N^2}\sum_{\alpha,\gamma=1}^{N}\sum_{i=1}^{d}\mathbb{E}\left[\int_0^1 \dot{x}_t^i[\mathcal{S}_t^N(x)]_\alpha \dot{y}_t^i[\mathcal{S}_t^N(y)]_\alpha dt\right]$$

$$=\sigma_a^2 + \sigma_b^2\int_0^1 \langle\dot{x}_t, \dot{y}_t\rangle\, dt + \sigma_A^2\int_0^1 \mathbb{E}\left[\frac{1}{N}\langle\mathcal{S}_t^N(x), \mathcal{S}_t^N(y)\rangle_{\mathbb{R}^N}\right]\langle\dot{x}_t, \dot{y}_t\rangle\, dt$$

where the penultimate equality follows from Itô's isometry. Hence

$$\frac{1}{N}\mathbb{E}\left[\langle\mathcal{S}_1^N(x_k), \mathcal{S}_1^N(x_m)\rangle\right] = \sigma_a^2 + \int_0^1 \left(\frac{\sigma_A^2}{N}\mathbb{E}[\langle\mathcal{S}_1^N(x_k), \mathcal{S}_1^N(x_m)\rangle] + \sigma_b^2\right)\langle(\dot{x}_k)_s, (\dot{x}_m)_s\rangle_{\mathbb{R}^d}\, ds$$

Defining $K_t^N(x_k, x_m) := \frac{1}{N}\mathbb{E}[\langle\mathcal{S}_t^N(x_k), \mathcal{S}_t^N(x_m)\rangle$ and noticing that we can repeat the previous arguments for all $t \in [0, 1]$, not only for $t = 1$, by considering for example piecewise constant extensions of the $\mathcal{S}^{M,N}$ we see that

$$K_t^N(x_k, x_m) = \sigma_a^2 + \int_0^t \left(\sigma_A^2 K_s^N(x_k, x_m) + \sigma_b^2\right)\langle(\dot{x}_k)_s, (\dot{x}_m)_s\rangle_{\mathbb{R}^d}\, ds$$

meaning that $K_t^N(x_k, x_m)$ solves Equation (15).

Since we showed in the proof of Theorem B.6 that the unique solution to the equation is $\kappa_{id}^{x_k, x_m}(1)$ we conclude.

$\square$

Directly proving the convergence in distribution of the infinite-width networks to centered Gaussians with covariance function κ_{id} is difficult since the networks are not Gaussian processes. However we believe that it is possible to employ McKean-Vlasov arguments to prove that they are at the limit, to then obtain a commmutativity of the limits; we leave the exploration of this direction to future work.

B.3. Commutativity of Limits

In this section we are going to prove the commutativity of limits in the inhomogeneous case. To do this we are going to proceed similarly to (Hayou & Yang, 2023) which proves the same result in a much restricted case.

Note that if φ is K-Lip *i.e.* $|\varphi(x) - \varphi(y)| \leq K|x - y|$ and $\varphi(0) = 0$ then

$$\|\varphi(x)\|^2 = \sum |\varphi(x_i)|^2 \leq \sum K^2|x_i|^2 \leq K^2\|x\|^2$$

hence
$$\|\varphi(x)\| \le K\,\|x\|$$

Assumption B.15. In this subsection φ is considered Lipschitz and with $\varphi(0) = 0$.

Recall Theorem B.13 and let $\tilde{\mathcal{S}}_t^{M,N}(x)$ be the Euler discretization of (18). Then one has (see the proof of (Kloeden & Platen, 1992)[10.2.2])

$$\sup_{N \ge 1} \frac{1}{N} \mathbb{E}\left[\sup_{t \in [0,1]} \left\|\tilde{\mathcal{S}}_t^{M,N}(x) - \mathcal{S}_t^N(x)\right\|^2\right] \le C|\mathcal{D}_M| \tag{19}$$

We can say the same for $\mathcal{S}_t^{M,N}(x)$ instead of $\tilde{\mathcal{S}}_t^{M,N}(x)$ *i.e* when the derivatives are replaced by difference quotients on the successive interval.

Proposition B.16. *The following inequality holds*

$$\sup_{N \ge 1} \frac{1}{N} \sup_{t \in [0,1]} \mathbb{E}\left[\left\|\mathcal{S}_t^{M,N}(x) - \mathcal{S}_t^N(x)\right\|^2\right] \le \tilde{C}|\mathcal{D}_M|$$

Proof. The bounding constant in equation (19) depends on the path x only trough $\|\dot{x}\|_{\infty,[0,1]}$. In particular the bound is uniform on bounded sets, with respect to this norm.

In particular we can interpolate $\{(t_m, x_{t_m})\}_{m=0,\dots,|\mathcal{D}_M|}$ in such a way to have the resulting map $\bar{x}$ with

$$\dot{\bar{x}}_{t_m} = \frac{x_{t_{m+1}} - x_{t_m}}{t_{m+1} - t_m} = \frac{\Delta x_{t_{m+1}}}{\Delta t_{m+1}}$$

such that definitely in M $\dot{x}$ is $\frac{1}{2}$-Holder continuous and

$$\|\dot{\bar{x}} - \dot{x}\|_{\infty,[0,1]}^2 \lesssim |\mathcal{D}_M|$$

One can for example interpolate the points with a polynomial close enough to $\dot{x}$, being the polynomial defined on the compact $[0,1]$ it is Lipschitz on $[0,1]$ hence $\frac{1}{2}$-Holder continuous. The existence of such a polynomial follows from

$$\left|\frac{\Delta x_{t_{m+1}}}{\Delta t_{m+1}} - \dot{x}_{t_m}\right| \le \frac{1}{\Delta t_{m+1}} \int_{t_m}^{t_{m+1}} |\dot{x}_s - \dot{x}_{t_m}| ds \le \frac{2}{3} C_{\frac{1}{2}-Ho}\sqrt{\Delta t_{m+1}}$$

Then $\mathcal{S}^{M,N}(x)$ will be exactly the Euler discretisation of $\mathcal{S}^N(\bar{x})$ on $\mathcal{D}_M$, hence

$$\sup_{N \ge 1} \frac{1}{N} \mathbb{E}\left[\sup_{t \in [0,1]} \left\|\mathcal{S}_t^{M,N}(x) - \mathcal{S}_t^N(\bar{x})\right\|^2\right] \le C|\mathcal{D}_M|$$

To conclude we just need to prove

$$\sup_{N \ge 1} \frac{1}{N} \sup_{t \in [0,1]} \mathbb{E}\left[\left\|\mathcal{S}_t^N(\bar{x}) - \mathcal{S}_t^N(x)\right\|^2\right] \le \bar{C}|\mathcal{D}_M|$$

since

$$\sup_{t \in [0,1]} \mathbb{E}\left[\left\|\mathcal{S}_t^{M,N}(x) - \mathcal{S}_t^N(x)\right\|^2\right] \le$$

$$2\left(\sup_{t \in [0,1]} \mathbb{E}\left[\left\|\mathcal{S}_t^{M,N}(x) - \mathcal{S}_t^N(\bar{x})\right\|^2\right] + \sup_{t \in [0,1]} \mathbb{E}\left[\left\|\mathcal{S}_t^N(\bar{x}) - \mathcal{S}_t^N(x)\right\|^2\right]\right) \le$$

$$2\left(\mathbb{E}\left[\sup_{t \in [0,1]} \left\|\mathcal{S}_t^{M,N}(x) - \mathcal{S}_t^N(\bar{x})\right\|^2\right] + \sup_{t \in [0,1]} \mathbb{E}\left[\left\|\mathcal{S}_t^N(\bar{x}) - \mathcal{S}_t^N(x)\right\|^2\right]\right) \le$$

$$2(C + \bar{C})|\mathcal{D}_M|$$

Then note how

$$\mathcal{S}_t^N(x) = \mathcal{S}_0^N(x) + \sum_{k=1}^{d} \int_0^t \frac{\sigma_A}{\sqrt{N}} \dot{x}_s^k dW_s^k \varphi(\mathcal{S}_s^N(x)) + \sigma_b \dot{x}_s^k dB_s^k$$

$$= \mathcal{S}_0^N(x) + \sum_{k=1}^{d} \sum_{j=1}^{N} \int_0^t \dot{x}_s^k \left(\frac{\sigma_A}{\sqrt{N}} \varphi([\mathcal{S}_s^N(x)]_j) d[W^k]_s^j + \sigma_b dB_s^k \right)$$

thus

$$\mathcal{S}_t^N(\bar{x}) - \mathcal{S}_t^N(x) = \sum_{k=1}^{d} \sum_{j=1}^{N} \int_0^t (\dot{\bar{x}}_s^k - \dot{x}_s^k) \left(\frac{\sigma_A}{\sqrt{N}} \varphi([\mathcal{S}_s^N(\bar{x})]_j) d[W^k]_s^j + \sigma_b dB_s^k \right)$$

$$+ \sum_{k=1}^{d} \sum_{j=1}^{N} \int_0^t \dot{x}_s^k \frac{\sigma_A}{\sqrt{N}} \left(\varphi([\mathcal{S}_s^N(\bar{x})]_j) - \varphi([\mathcal{S}_s^N(x)]_j) \right) d[W^k]_s^j$$

which leads to

$$\mathbb{E}\left[\left\| \mathcal{S}_t^N(\bar{x}) - \mathcal{S}_t^N(x) \right\|^2 \right] = N \sum_{k=1}^{d} \int_0^t (\dot{\bar{x}}_s^k - \dot{x}_s^k)^2 \left(\frac{\sigma_A^2}{N} \mathbb{E}[\|\varphi(\mathcal{S}_s^N(\bar{x}))\|^2] + \sigma_b^2 \right) ds$$

$$+ N \sum_{k=1}^{d} \int_0^t (\dot{x}_s^k)^2 \frac{\sigma_A^2}{N} \left\| \mathbb{E}[\varphi(\mathcal{S}_s^N(\bar{x})) - \varphi(\mathcal{S}_s^N(x))] \right\|^2 ds$$

$$\le NK \sum_{k=1}^{d} \int_0^t (\dot{\bar{x}}_s^k - \dot{x}_s^k)^2 \left(\frac{\sigma_A^2}{N} \mathbb{E}[\|\mathcal{S}_s^N(\bar{x})\|^2] + \sigma_b^2 \right) ds$$

$$+ NK \sum_{k=1}^{d} \int_0^t (\dot{x}_s^k)^2 \frac{\sigma_A^2}{N} \mathbb{E}[\|\mathcal{S}_s^N(\bar{x}) - \mathcal{S}_s^N(x)\|^2] ds$$

$$\le NK \|\dot{\bar{x}} - \dot{x}\|_{\infty,[0,1]}^2 \sum_{k=1}^{d} \int_0^t \left(\frac{\sigma_A^2}{N} \mathbb{E}[\|\mathcal{S}_s^N(\bar{x})\|^2] + \sigma_b^2 \right) ds$$

$$+ NK \|\dot{x}\|_{\infty,[0,1]}^2 \sum_{k=1}^{d} \int_0^t \frac{\sigma_A^2}{N} \mathbb{E}[\|\mathcal{S}_s^N(\bar{x}) - \mathcal{S}_s^N(x)\|^2] ds$$

Using the fact that, again (Kloeden & Platen, 1992)[10.2.2],

$$\frac{1}{N} \mathbb{E}[\|\mathcal{S}_s^N(\bar{x})\|^2] \le \tilde{K}_{\bar{x}}$$

with $\tilde{K}_{\bar{x}}$ only depending on $\|\dot{\bar{x}}\|_\infty$ we obtain, via Gronwall, a bound of type

$$\mathbb{E}\left[\left\| \mathcal{S}_t^N(\bar{x}) - \mathcal{S}_t^N(x) \right\|^2 \right] \le \|\dot{\bar{x}} - \dot{x}\|_{\infty,[0,1]}^2 \bar{K} \tag{20}$$

where $\bar{K}$ only depends on $\|\dot{\bar{x}}\|_\infty + \|\dot{x}\|_\infty$. By our choice of $\bar{x}$ we conclude. $\square$

We can finally state the main result:

Theorem B.17. *If the activation function is Lipschitz and $\varphi(0) = 0$ then there is a constant C depending only on $\|\dot{x}\|_{\infty,[0,1]}$ in an increasing fashion such that:*

$$\sup_{N \ge 1} \sup_{t \in [0,1]} \mathcal{W}_1(\mu_t^{M,N}(x), \mu_t^N(x)) \le \bar{C} \sqrt{|\mathcal{D}_M|}$$

where $\mu_t^{M,N}$ is the distribution of any coordinate of $\mathcal{S}_t^{M,N}(x)$ and $\mu_t^N(x)$ that of any coordinate of $\mathcal{S}_t^N(x)$ (they are identically distributed).

Proof. Let $G : \mathbb{R} \to \mathbb{R}$ be 1-Lipschitz, we have

$$\mathbb{E}[|G([\mathcal{S}_t^{M,N}(x)]_1) - G([\mathcal{S}_t^N(x)]_1)|] \leq \mathbb{E}\left[|[\mathcal{S}_t^{M,N}(x)]_1 - [\mathcal{S}_t^N(x)]_1|\right] \leq$$

$$\left(\mathbb{E}\left[|[\mathcal{S}_t^{M,N}(x)]_1 - [\tilde{\mathcal{S}}_t^{M,N}(x)]_1|^2\right]\right)^{\frac{1}{2}} =$$

$$\left(\frac{1}{N}\sum_{\alpha=1}^N \mathbb{E}\left[|[\mathcal{S}_t^{M,N}(x)]_\alpha - [\mathcal{S}_t^N(x)]_\alpha|^2\right]\right)^{\frac{1}{2}} \leq$$

$$\left(\frac{1}{N}\mathbb{E}\left[\left\|\mathcal{S}_t^{M,N}(x) - \mathcal{S}_t^N(x)\right\|^2\right]\right)^{\frac{1}{2}} \leq$$

$$\left(\frac{1}{N}\sup_{t\in[0,1]}\mathbb{E}\left[\left\|\mathcal{S}_t^{M,N}(x) - \mathcal{S}_t^N(x)\right\|^2\right]\right)^{\frac{1}{2}} \leq \sqrt{C|\mathcal{D}_M|}$$

hence

$$\mathcal{W}_1(\mu_t^{M,N}(x), \mu_t^N(x)) \leq \sqrt{C|\mathcal{D}_M|}$$

$\square$

Note that the entries have the same distribution as the normal projections since for $t \in \mathbb{R}$

$$\mathbb{E}[\exp\{it\langle v^N, \mathcal{S}^{M,N}\rangle\}] = \mathbb{E}\left[\mathbb{E}[\exp\{it\langle v^N, \mathcal{S}^{M,N}\rangle\}|\mathcal{S}^{M,N}]\right]$$

$$= \mathbb{E}\left[\mathbb{E}[\exp\{i\langle v^N, t\mathcal{S}^{M,N}\rangle\}|\mathcal{S}^{M,N}]\right] = \mathbb{E}\left[\exp\{-\frac{t^2}{2N}\left\|\mathcal{S}^{M,N}\right\|^2\}\right]$$

and

$$\mathbb{E}[\exp\{it[\mathcal{S}^{M,N}]_1\}] = \mathbb{E}\left[\mathbb{E}[\exp\{it\langle v^N, \mathcal{S}^{M,N}\rangle\}|\mathcal{S}^{M,N}]\right]$$

$$= \mathbb{E}\left[\mathbb{E}[\exp\{i\langle v^N, t\mathcal{S}^{M,N}\rangle\}|\mathcal{S}^{M,N}]\right] = \mathbb{E}\left[\exp\{-\frac{t^2}{2N}\left\|\mathcal{S}^{M,N}\right\|^2\}\right]$$

thus giving us

Theorem B.18. *If the activation function is Lipschitz and $\varphi(0) = 0$ then there is a constant C depending only on $\|\dot{x}\|_{\infty,[0,1]}$ in an increasing fashion such that:*

$$\sup_{N\geq1}\sup_{t\in[0,1]}\mathcal{W}_1(\mu_t^{M,N}(x), \mu_t^N(x)) \leq \bar{C}\sqrt{|\mathcal{D}_M|}$$

where $\mu_t^{M,N}$ is the distribution of $\langle v^N, \mathcal{S}_t^{M,N}(x)\rangle$ and $\mu_t^N(x)$ that of $\langle v^N, \mathcal{S}_t^N(x)\rangle$ for some independent vector with iid entries $[v^N]_\alpha \sim \mathcal{N}(0, \frac{1}{N})$.

Remark. With the same proof we have also the convergence of the laws of the rescaled processes $\frac{1}{\sqrt{N}}\mathcal{S}_t^{M,N}(x)$.

Remark. Note that the exact arguments, being of L^2 type, can be repeated for the "stacked" vector $(\mathcal{S}^{N,M}(x_1), \ldots, \mathcal{S}^{N,M}(x_N))$ extending (qualitatively) the bounds to the multi-input case.

B.3.1. PROOF OF THEOREM 3.1: PART 2

Corollary B.19 (Thm. 3.1)**.** *If the activation function is Lipschitz and $\varphi(0) = 0$ then the limits in Thm. B.2 commute.*

Proof. By the classical Moore-Osgood theorem we need to prove that one of the two limits is uniform in the other, for example that the limit in distribution as $M \to \infty$ is uniform in N in some metric which describes convergence in distribution. But this is just the content of the previous results, extended to the multi-input case. $\square$

C. Proofs for homogeneous controlled ResNets

In this section we prove all the results for the homogeneous case. Recall that a *randomly initialised, 1-layer homogeneous controlled ResNet* $\Phi_\varphi^{M,N} : \mathbb{X} \to \mathbb{R}$ is defined as follows

$$\Phi_\varphi^{M,N}(x) := \left\langle \phi, S_{t_M}^{M,N}(x) \right\rangle_{\mathbb{R}^N}$$

where $\phi \in \mathbb{R}^N$ is the random vector $[\phi]_\alpha \overset{\text{iid}}{\sim} \mathcal{N}(0, \frac{1}{N})$, and where the random functions $S_{t_i}^{M,N} : \mathbb{X} \to \mathbb{R}^N$ satisfy the following recursive relation

$$S_{t_{i+1}}^{M,N} = S_{t_i}^{M,N} + \sum_{k=1}^{d} \left(A_k \varphi(S_{t_i}^{M,N}) + b_k \right) \Delta x_{t_{i+1}}^k$$

with initial condition $S_{t_0} = a$ with $[a]_\alpha \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_a^2)$, and Gaussian weights $A_k \in \mathbb{R}^{N \times N}$ and biases $b_k \in \mathbb{R}^N$ sampled independently according to

$$[A_k]_\alpha^\beta \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \frac{\sigma_A^2}{N}\right), \quad [b_k]_\alpha \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \sigma_b^2\right)$$

This section too will be subdivided in three main parts: in the first we consider the infinite-width-then-depth limit, in the second we reverse the order and consider the infinite-depth-then-width limit and in the third one we prove that the limits can be exchanged.

C.1. The infinite-width-then-depth regime

The main goal is that of proving the first part of Theorem 4.1, which we restate here:

Theorem C.1. *Let $\{\mathcal{D}_M\}_{M \in \mathbb{N}}$ be a sequence of partitions of $[0,1]$ such that $|\mathcal{D}_M| \to 0$ as $M \to \infty$. Let the activation function φ be linearly bounded, absolutely continuous and with exponentially bounded derivative. For any subset $\mathcal{X} = \{x_1, \ldots, x_n\} \subset \mathbb{X}$ of paths the following convergence in distribution holds*

$$\lim_{M \to \infty} \lim_{N \to \infty} \Phi_\varphi^{M,N}(\mathcal{X}) = \mathcal{N}(0, \mathcal{K}_\varphi(\mathcal{X}, \mathcal{X})) \tag{21}$$

where the positive semidefinite kernel $\mathcal{K}_\varphi : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ is defined for any two paths $x, y \in \mathbb{X}$ as $\mathcal{K}_\varphi(x,y) = \mathcal{K}_\varphi^{x,y}(1,1)$, where $\mathcal{K}_\varphi^{x,y} : [0,1]^2 \to \mathbb{R}$ is the unique solution of the following differential equation

$$\partial_s \partial_t \mathcal{K}_\varphi^{x,y} = \left[\sigma_A^2 V_\varphi \left(\Sigma_\varphi^{x,y} \right) + \sigma_b^2 \right] \langle \dot{x}_s, \dot{y}_t \rangle \tag{22}$$

where

$$\Sigma_\varphi^{x,y}(s,t) = \begin{pmatrix} \mathcal{K}_\varphi^{x,x}(s,s), \mathcal{K}_\varphi^{x,y}(s,t) \\ \mathcal{K}_\varphi^{x,y}(s,t), \mathcal{K}_\varphi^{y,y}(t,t) \end{pmatrix}$$

and with initial conditions for any $s,t \in [0,1]$

$$\mathcal{K}_\varphi^{x,y}(0,0) = \mathcal{K}_\varphi^{x,y}(s,0) = \mathcal{K}_\varphi^{x,y}(0,t) = \sigma_a^2$$

Once again, it is clearer to subdivide the proof of this result in two parts:

1. The infinite width-convergence of the finite dimensional distributions

$$\Phi_\varphi^{M,N}(\mathcal{X}) \xrightarrow{N \to \infty} \mathcal{N}(0, \mathcal{K}_{\mathcal{D}_M \times \mathcal{D}_M}(\mathcal{X}, \mathcal{X}))$$

   to those of a Gaussian process $\mathcal{GP}(0, \mathcal{K}_{\mathcal{D}_M \times \mathcal{D}_M})$ defined by a Kernel computed as the final value of a difference equation on the partition $\mathcal{D}_M \times \mathcal{D}_M$ of $[0,1] \times [0,1]$. This will be done using Tensor Programs (Yang, 2019) in C.1.1.

2. The infinite-depth convergence of the discrete kernels $\mathcal{K}_{\mathcal{D}_M \times \mathcal{D}_M}$ to a limiting Kernel $\mathcal{K}_\varphi$ which solves the differential equation 22. This will be done in C.1.2.

C.1.1. INFINITE-WIDTH LIMIT WITH FIXED-DEPTH

We prove convergence in the finite-depth, infinite-width limit to a GP endowed with discrete kernels. using the formalism of *Tensor Programs*.

Theorem C.2. *Let $\varphi : \mathbb{R} \to \mathbb{R}$ be linearly bounded [11]. For any subset $\mathcal{X} = \{x_1, \ldots, x_n\} \subset \mathbb{X}$ the following convergence in distribution holds*

$$\lim_{N \to \infty} \Phi_\varphi^{M,N}(\mathcal{X}) = \mathcal{N}(0, \mathcal{K}_{\mathcal{D}_M \times \mathcal{D}_M}(\mathcal{X}, \mathcal{X}))$$

where the map $\mathcal{K}_{\mathcal{D} \times \mathcal{D}'} : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ is, given any two partitions $\mathcal{D}, \mathcal{D}'$ of $[0, 1]$, defined for any two paths $x, y \in \mathbb{X}$ as $\mathcal{K}_{\mathcal{D} \times \mathcal{D}'}(x, y) = \mathcal{K}_{\mathcal{D} \times \mathcal{D}'}^{x,y}(1, 1)$ where $\mathcal{K}_{\mathcal{D} \times \mathcal{D}'}^{x,y} : \mathcal{D} \times \mathcal{D}' \to \mathbb{R}$ solves the following difference equation

$$\mathcal{K}_{\mathcal{D} \times \mathcal{D}'}(s_m, t_n) = \mathcal{K}_{\mathcal{D} \times \mathcal{D}'}(s_{m-1}, t_n) + \mathcal{K}_{\mathcal{D} \times \mathcal{D}'}(s_m, t_{n-1}) - \mathcal{K}_{\mathcal{D} \times \mathcal{D}'}(s_{m-1}, t_{n-1})$$

$$+ \left(\sigma_A^2 V_\varphi \left(\Sigma_{\mathcal{D} \times \mathcal{D}'}^{x,y}(s_{m-1}, t_{n-1}) \right) + \sigma_b^2 \right) \langle \Delta x_{s_m}, \Delta y_{t_n} \rangle_{\mathbb{R}^d} \tag{23}$$

with initial conditions

$$\mathcal{K}_{\mathcal{D} \times \mathcal{D}'}^{x,y}(0, 0) = \mathcal{K}_{\mathcal{D} \times \mathcal{D}'}^{x,y}(0, t_n) = \mathcal{K}_{\mathcal{D} \times \mathcal{D}'}^{x,y}(s_m, 0) = \sigma_a^2$$

and where

$$\Sigma_{\mathcal{D} \times \mathcal{D}'}^{x,y}(s, t) = \begin{pmatrix} \mathcal{K}_{\mathcal{D} \times \mathcal{D}}^{x,x}(s, s) & \mathcal{K}_{\mathcal{D} \times \mathcal{D}'}^{x,y}(s, t) \\ \mathcal{K}_{\mathcal{D} \times \mathcal{D}'}^{x,y}(s, t) & \mathcal{K}_{\mathcal{D}' \times \mathcal{D}'}^{y,y}(s, s) \end{pmatrix}$$

Proof. We use (Yang, 2019)[Corollary 5.5] applied to the Tensor Program of Algorithm 2 where the input variables are independently sampled according to

$$[A_k]_\alpha^\beta \sim \mathcal{N}(0, \frac{\sigma_A^2}{N}), [v]_\alpha \sim \mathcal{N}(0, 1), [S_0]_\alpha \sim \mathcal{N}(0, \sigma_a^2) \ \text{and} \ [b_k]_\alpha \sim \mathcal{N}(0, \sigma_b^2)$$

Note how the sampling scheme follows (Yang, 2019)[Assumption 5.1] and how linearly bounded functions are *controlled* in the sense of (Yang, 2019)[Definition 5.3] since for all $x \in \mathbb{R}$ one has $|\phi(x)| \le C(1 + |x|) \le e^{|x| + log(C)}$. Thus we are under the needed assumptions to apply (Yang, 2019)[Corollary 5.5].

There are two things to notice, done in order to satisfy the required formalism:

- In the program the output projector v is sampled according to $\mathcal{N}(0, 1)$ while the original $\phi \sim \mathcal{N}(0, \frac{1}{N})$. This does not pose any problems since the output of the formal programs uses $v/\sqrt{N} \sim \mathcal{N}(0, \frac{1}{N})$.

- The input paths x_i enter program 2 not as *Inputs* but as coefficients of *LinComb*, this means that for any choice of input paths we must formally consider different algorithms. In any case, for any possible choice, the result has always the same functional form; hence *a posteriori* it is legitimate to think about *one* algorithm.

Note that S_m^x stands for the formal variable in program 2, not for $S_{t_m^{\mathcal{D}_M}}^{M,N}(x)$ even thogh this is the value which it "stores" for a fixed hidden dimension N.

The result in (Yang, 2019)[Corollary 5.5] tells us that the output vector converges in law, as $N \to \infty$, to a Gaussian distribution $\mathcal{N}(0, \tilde{K}^M)$ where

$$[\tilde{K}^M]_i^j = \mathbb{E}_{Z \sim \mathcal{N}(\mu, \Sigma)} \left[Z^{S_{\|\mathcal{D}_M\|}^{x_i}} Z^{S_{\|\mathcal{D}_M\|}^{x_j}} \right] = \Sigma(S_{\|\mathcal{D}_M\|}^{x_i}, S_{\|\mathcal{D}_M\|}^{x_j})$$

with μ, Σ computed according to (Yang, 2019)[Definition 5.2] and defined on the set of all G-vars in the program *i.e.*

$$\mu(g) = \begin{cases} \mu^{in}(g) & \text{if } g \text{ is } \textbf{Input} \text{ G-var} \\ \sum_k a_k \mu(g_k) & \text{if } g \text{ is introduced as } \sum_k a_k g_k \text{ via } \textbf{LinComb} \\ 0 & \text{otherwise} \end{cases}$$

---

[11] in the sense that $\exists C > 0$. such that $|\phi(x)| \le C(1 + |x|)$.

$$\Sigma(g,g') = \begin{cases} \Sigma^{in}(g,g') & \text{if both } g \text{ and } g' \text{ are \textbf{Input} G-var} \\ \sum_k a_k \Sigma(g_k, g') & \text{if } g \text{ is introduced as } \sum_k a_k g_k \text{ via \textbf{LinComb}} \\ \sum_k a_k \Sigma(g, g'_k) & \text{if } g' \text{ is introduced as } \sum_k a_k g'_k \text{ via \textbf{LinComb}} \\ \sigma_W^2 \mathbb{E}_{Z \sim \mathcal{N}(\mu,\Sigma)}[\phi(Z)\phi'(Z)] & \text{if } g = Wh, \ g' = Wh' \text{ via \textbf{MatMul} w/ same } W \\ 0 & \text{otherwise} \end{cases}$$

where $h = \phi((g_k)_{k=1}^m)$ for some function ϕ and $\phi(Z) := \phi((Z^{g_k})_{k=1}^m)$, similarly for g'.

In our setting $\mu^{in} \equiv 0$ since all **Input** variables are independent, from which $\mu \equiv 0$; furthermore $\Sigma^{in}(g,g') = 0$ except if $g = g'$ when it takes values in $\{\sigma_a^2, \sigma_b^2, \sigma_A^2\}$ accordingly.

Following the rules of Σ, assuming $m_i, m_j \in \{1, \ldots, \|\mathcal{D}_M\|\}$, we obtain

$$\begin{aligned} \Sigma(S_{m_i}^{x_i}, S_{m_j}^{x_j}) &= \Sigma(S_{m_i-1}^{x_i}, S_{m_j}^{x_j}) + \Sigma\left(\sum_{k=1}^d \gamma_{m_i,k}^i \Delta(x_i)_{t_{m_i}}^k, S_{m_j}^{x_j}\right) \\ &= \Sigma(S_{m_i-1}^{x_i}, S_{m_j}^{x_j}) + \Sigma\left(\sum_{k=1}^d \gamma_{m_i,k}^i \Delta(x_i)_{t_{m_i}}^k, S_{m_j-1}^{x_j}\right) \\ &\quad + \Sigma\left(\sum_{k=1}^d \gamma_{m_i,k}^i \Delta(x_i)_{t_{m_i}}^k, \sum_{l=1}^d \gamma_{m_j,l}^j \Delta(x_j)_{t_{m_j}}^l\right) \\ &= \Sigma(S_{m_i-1}^{x_i}, S_{m_j}^{x_j}) + \Sigma(S_{m_i}^{x_i}, S_{m_j-1}^{x_j}) - \Sigma(S_{m_i-1}^{x_i}, S_{m_j-1}^{x_j}) \\ &\quad + \sum_{k,l=1}^d \Sigma(\gamma_{m_i,k}^i, \gamma_{m_j,l}^j) \Delta(x_i)_{t_{m_i}}^k \Delta(x_j)_{t_{m_j}}^l \end{aligned}$$

Now

$$\Sigma(\gamma_{m_i,k}^i, \gamma_{m_j,l}^j) = \delta_{k,l}\sigma_A^2 \mathbb{E}[\varphi(Z_1)\varphi(Z_2)] + \Sigma(b_k, b_l) = \delta_{kl}\left[\sigma_A^2 \mathbb{E}[\varphi(Z_1)\varphi(Z_2)] + \sigma_b^2\right]$$

where $[Z_1, Z_2]^\top \sim \mathcal{N}(0, \tilde{\Sigma}_{m_i-1,m_j-1}(x_i, x_j))$ with

$$\tilde{\Sigma}_{m,m'}(x_i, x_j) = \begin{pmatrix} \Sigma(S_m^{x_i}, S_m^{x_i}) & \Sigma(S_m^{x_i}, S_{m'}^{x_j}) \\ \Sigma(S_m^{x_i}, S_m^{x_j}) & \Sigma(S_{m'}^{x_j}, S_{m'}^{x_j}) \end{pmatrix}$$

thus if we set, for $t_{m_i}, t_{m_j} \in \mathcal{D}_M$, $\mathcal{K}_{\mathcal{D}_M \times \mathcal{D}_M}^{x_i,x_j}(t_{m_i}, t_{m_j}) := \Sigma(S_{m_i}^{x_i}, S_{m_j}^{x_j})$ then we get

$$\begin{aligned} \mathcal{K}_{\mathcal{D}_M \times \mathcal{D}_M}^{x_i,x_j}(t_{m_i}, t_{m_j}) &= \mathcal{K}_{\mathcal{D}_M \times \mathcal{D}_M}^{x_i,x_j}(t_{m_i-1}, t_{m_j}) + \mathcal{K}_{\mathcal{D}_M \times \mathcal{D}_M}^{x_i,x_j}((t_{m_i}, t_{m_j-1}) \\ &\quad - \mathcal{K}_{\mathcal{D}_M \times \mathcal{D}_M}^{x_i,x_j}((t_{m_i-1}, t_{m_j-1}) \\ &\quad + \sum_{k=1}^d \left(\sigma_A^2 V_\varphi\left(\tilde{\Sigma}_{m_i-1,m_j-1}(x_i, x_j)\right) + \sigma_b^2\right)\Delta(x_i)_{t_{m_i}}^k \Delta(x_j)_{t_{m_j}}^k \end{aligned}$$

which is exactly what Equation (23) states. Then note how

$$\Sigma(S_{m_i}^{x_i}, S_0^{x_j}) = \Sigma(S_{m_i-1}^{x_i}, S_0^{x_j}) = \cdots = \Sigma(S_0^{x_i}, S_0^{x_j}) = \sigma_a^2$$

Thus finally we can conclude and write the entries of the matrix $\tilde{K}^M$ as

$$[\tilde{K}^M]_i^j = \mathcal{K}_{\mathcal{D}_M \times \mathcal{D}_M}^{x_i,x_j}(1,1)$$

$\square$

Algorithm 2 $S_1^{M,N}$ as Nestor program

Input: $S_0 : \mathsf{G}(N)$ ▷ *initial value*
Input: $b_1, \ldots, b_d : \mathsf{G}(N)$ ▷ *biases*
Input: $A_1, \ldots, A_d : \mathsf{A}(N, N)$ ▷ *matrices*
Input: $v : \mathsf{G}(N)$ ▷ *readout layer weights*
for $i = 1, \ldots, n$ **do**
   // *Compute* $S_1^{\mathcal{D}_M, N}(x_i)$ *(here* $S_0^{x_i}$ *is to be read as* S_0*)*
   for $m = 1, \ldots, \|\mathcal{D}_M\|$ **do**
     for $k = 1, \ldots, d$ **do**
       $\alpha_{m,k}^i := \varphi(S_{m-1}^{x_i}) : \mathsf{H}(N)$ ▷ *by Nonlin;*
       $\beta_{m,k}^i := A_k \alpha_{m,k}^i : \mathsf{G}(N)$ ▷ *by Matmul;*
       $\gamma_{m,k}^i := \beta_{m,k}^i + b_k : \mathsf{G}(N)$ ▷ *by LinComb;*
     end for
     $S_m^{x_i} := S_{m-1}^{x_i} + \sum_{k=1}^d \gamma_{m,k}^i [(x_i)_{t_m}^k - (x_i)_{t_{m-1}}^k] : \mathsf{G}(N)$ ▷ *by LinComb;*
   end for
end for
Output: $(v^T S_{\|\mathcal{D}_M\|}^{x_i} / \sqrt{N})_{i=1,\ldots,n}$

Actually we have proved the even stronger statement that

Corollary C.3. *For all $t, s \in \mathcal{D}_M$ one has the following distributional limit*

$$\left\langle \phi^N, S_t^{M,N}(\mathcal{X}) \right\rangle \left\langle \phi^N, S_s^{M,N}(\mathcal{X}) \right\rangle \xrightarrow[N\to\infty]{} \mathcal{N}(0, K_{\mathcal{D}_M \times \mathcal{D}_M}^{\mathcal{X}, \mathcal{X}}(t, s))$$

where $[\phi^N]_\alpha \sim \mathcal{N}(0, \frac{1}{N})$ are independently sampled.

Moreover with the same proof but considering different partitions, $\mathcal{D}$ and $\mathcal{D}'$, for different paths $x, y \in \mathbb{X}$ in the algorithm we see that:

Proposition C.4. *For all $t \in \mathcal{D}$, for all $s \in \mathcal{D}'$, with a clear abouse of notation, one has the following distributional limit*

$$\left\langle \phi^N, S_t^{\mathcal{D}, N}(x) \right\rangle \left\langle \phi^N, S_s^{\mathcal{D}', N}(y) \right\rangle \xrightarrow[N\to\infty]{} \mathcal{N}(0, K_{\mathcal{D} \times \mathcal{D}'}^{x, y}(t, s))$$

where $[\phi^N]_\alpha \sim \mathcal{N}(0, \frac{1}{N})$ are independently sampled and the matrices $\Sigma_{\mathcal{D} \times \mathcal{D}'}^{x, y}(t, s)$ are always in PSD_2.

Remark. The fact that the matrices $\Sigma_{\mathcal{D} \times \mathcal{D}'}^{x, y}(t, s)$ are always in PSD_2 will be crucial in the following.

C.1.2. UNIFORM CONVERGENCE TO NEURAL SIGNATURE KERNELS

We now prove that the sequence of discrete kernels convergence uniformly to our neural signature kernels. First we need definitions for extending the discrete kernels to $[0, 1]^2$ similarly to the previous section.

Definition C.5. Fix two partitions $\mathcal{D} := \{0 = s_0 < \cdots < s_M = 1\}$ and $\mathcal{D}' := \{0 = t_0 < \cdots < t_{M'} = 1\}$. For any $x, y \in \mathbb{X}$ define

$$\mathcal{K}_{\mathcal{D} \times \mathcal{D}'}^{x, y} : \mathcal{D} \times \mathcal{D}' \to \mathbb{R}$$

as the map satisfying the following recursion

$$\begin{aligned}
&\mathcal{K}_{\mathcal{D} \times \mathcal{D}'}^{x, y}(s_m, t_n) \\
=&\mathcal{K}_{\mathcal{D} \times \mathcal{D}'}^{x, y}(s_{m-1}, t_n) + \mathcal{K}_{\mathcal{D} \times \mathcal{D}'}^{x, y}(s_m, t_{n-1}) - \mathcal{K}_{\mathcal{D} \times \mathcal{D}'}^{x, y}(s_m, t_{n-1}) \\
&+ \sum_{k=1}^d \left(\sigma_A^2 V_\varphi \left(\Sigma_{\mathcal{D} \times \mathcal{D}'}^{x, y}(s_{m-1}, t_{n-1}) \right) + \sigma_b^2 \right) \Delta x_{s_m}^k \Delta y_{t_n}^k \\
=&\sigma_a^2 + \sum_{\substack{0 \le k_1 < m \\ 0 \le k_2 < n}} \sum_{k=1}^d \left(\sigma_A^2 V_\varphi \left(\Sigma_{\mathcal{D} \times \mathcal{D}'}^{x, y}(s_{k_1}, t_{k_2}) \right) + \sigma_b^2 \right) \Delta x_{s_{k_1+1}}^k \Delta y_{t_{k_2+1}}^k
\end{aligned}$$

where

$$\mathcal{K}^{x,y}_{\mathcal{D}\times\mathcal{D}'}(0, t_n) = \mathcal{K}_{\mathcal{D}\times\mathcal{D}'}(s_m, 0) = \sigma_a^2$$

and

$$\Sigma^{x,y}_{\mathcal{D}\times\mathcal{D}'}(s_m, t_n) = \begin{pmatrix} \mathcal{K}^{x,x}_{\mathcal{D}\times\mathcal{D}'}(s_m, s_m), \mathcal{K}^{x,y}_{\mathcal{D}\times\mathcal{D}'}(s_m, t_n) \\ \mathcal{K}^{x,y}_{\mathcal{D}\times\mathcal{D}'}(s_m, t_n), \mathcal{K}^{y,y}_{\mathcal{D}\times\mathcal{D}'}(t_n, t_n) \end{pmatrix}$$

Definition C.6. We extend the map $\mathcal{K}^{x,y}_{\mathcal{D}\times\mathcal{D}'} : \mathcal{D} \times \mathcal{D}' \to \mathbb{R}$ to the whole square $[0, 1] \times [0, 1]$ in two ways: for any $(s, t) \in [s_{m-1}, s_m) \times [t_{n-1}, t_n)$

1. (integral interpolation) using a slight abuse of notation that overwrites the previous one, define the map $\mathcal{K}^{x,y}_{\mathcal{D}\times\mathcal{D}'} : [0, 1] \times [0, 1] \to \mathbb{R}$ as

$$\mathcal{K}^{x,y}_{\mathcal{D}\times\mathcal{D}'}(s, t) = \mathcal{K}^{x,y}_{\mathcal{D}\times\mathcal{D}'}(s_{m-1}, t_n) + \mathcal{K}^{x,y}_{\mathcal{D}\times\mathcal{D}'}(s_m, t_{n-1}) - \mathcal{K}^{x,y}_{\mathcal{D}\times\mathcal{D}'}(s_m, t_{n-1})$$
$$+ \int_{\eta=s_{m-1}}^{s} \int_{\tau=t_{n-1}}^{t} \left(\sigma_A^2 V_\varphi \left(\Sigma^{x,y}_{\mathcal{D}\times\mathcal{D}'}(s_{m-1}, t_{n-1}) \right) + \sigma_b^2 \right) \langle \dot{x}_\eta, \dot{y}_\tau \rangle_{\mathbb{R}^d} \, d\eta d\tau$$

We extend in a similar way the matrix $\Sigma^{x,y}_{\mathcal{D}\times\mathcal{D}'}$.

2. (piecewise constant interpolation) define the map $\tilde{\mathcal{K}}^{x,y}_{\mathcal{D}\times\mathcal{D}'} : [0, 1] \times [0, 1] \to \mathbb{R}$ as

$$\tilde{\mathcal{K}}^{x,y}_{\mathcal{D}\times\mathcal{D}'}(s, t) = \begin{cases} \mathcal{K}^{x,y}_{\mathcal{D}\times\mathcal{D}'}(s_m, t_n) & \text{if} \quad (s, t) \in [s_m, s_{m+1}) \times [t_n, t_{n+1}) \\ \mathcal{K}^{x,y}_{\mathcal{D}\times\mathcal{D}'}(1, t_n) & \text{if} \quad (s, t) \in \{1\} \times [t_n, t_{n+1}) \\ \mathcal{K}^{x,y}_{\mathcal{D}\times\mathcal{D}'}(s_m, 1) & \text{if} \quad (s, t) \in [s_m, s_{m+1}) \times \{1\} \\ \mathcal{K}^{x,y}_{\mathcal{D}\times\mathcal{D}'}(1, 1) & \text{if} \quad (s, t) = (1, 1) \end{cases}$$

and similarly for the matrix $\tilde{\Sigma}^{x,y}_{\mathcal{D}\times\mathcal{D}'}$.

Next we prove that the discrete kernels converge in distribution to a limiting kernel that satisfies a two-parameters differential equation.

Theorem C.7. *Let $\{\mathcal{D}_M\}_{M\in\mathbb{N}}$ be a sequence of partitions $[0, 1]$ with $|\mathcal{D}_M| \to 0$. Then we have the following convergence in $(C^0([0, 1] \times [0, 1]; \mathbb{R}), \|\cdot\|_\infty)$*

$$\mathcal{K}^{x,y}_{\mathcal{D}_M \times \mathcal{D}_M}(s, t) \to \mathcal{K}^{x,y}_\varphi(s, t)$$

where $\mathcal{K}^{x,y}(s, t)$ satisfies the following differential equation

$$\partial_s \partial_t \mathcal{K}^{x,y}_\varphi = \left[\sigma_A^2 V_\varphi \left(\Sigma^{x,y}_\varphi \right) + \sigma_b^2 \right] \langle \dot{x}_s, \dot{y}_t \rangle \tag{24}$$

where

$$\Sigma^{x,y}_\varphi(s, t) = \begin{pmatrix} \mathcal{K}^{x,x}_\varphi(s, s), \mathcal{K}^{x,y}_\varphi(s, t) \\ \mathcal{K}^{x,y}_\varphi(s, t), \mathcal{K}^{y,y}_\varphi(t, t) \end{pmatrix}$$

and with initial conditions for any $s, t \in [0, 1]$

$$\mathcal{K}^{x,y}_\varphi(0, 0) = \mathcal{K}^{x,y}_\varphi(s, 0) = \mathcal{K}^{x,y}_\varphi(0, t) = \sigma_a^2$$

The limit is independent of the chosen sequence and the convergence is, in x, y, uniform on bounded sets of $\mathbb{X}$.

The proof will rely on classic arguments used to prove convergence of Euler schemes to CDE solutions (see (Friz & Victoir, 2010) and (Huang & Tsai, 2004)). However, our setting is more complex than the ones we found in the literature; this is due to the presence of two "independent" driving signals and on the unique nature of the driving fields, which not only depend on the "present" but also on "past" and "future". This complex dependency structure of the driving fields makes them completely non-local and requires more involved arguments for the solution to be made sense of. Note that due to their definition as limiting kernels the $\tilde{\Sigma}^{x,y}_{\mathcal{D}_M \times \mathcal{D}_M}$ are always PSD. However we are going to need a more quantitative result, in the form of a uniform membership in some $PSD(R)$, in order to later leverage Lemma A.5. We first prove such a result for the case $x = y$ and $\mathcal{D} = \mathcal{D}'$, the general case will follow easily. The first "ingredient" is a uniform, in $x \in \mathbb{X}$, bound from below of the kernels for $t = s$.

Lemma C.8. $\forall x \in \mathbb{X}. \forall t \in \mathcal{D}.$ $\quad \tilde{\mathcal{K}}^{x,x}_{\mathcal{D} \times \mathcal{D}}(t, t) \geq \sigma_a^2$ *for any choice of* $\mathcal{D}$.

Proof. We have

$$\mathcal{K}^{x,x}_{\mathcal{D} \times \mathcal{D}}(t_m, t_m) = \sigma_a^2 + \sum_{\substack{0 \leq k_1 < m \\ 0 \leq k_2 < m}} \sum_{k=1}^{d} \left(\sigma_A^2 V_\varphi \left(\Sigma^{x,x}_{\mathcal{D} \times \mathcal{D}}(t_{k_1}, t_{k_2}) \right) + \sigma_b^2 \right) \Delta x_{t_{k_1+1}}^k \Delta x_{t_{k_2+1}}^k$$

and by Tensor Programs we know that

$$\sum_{\substack{0 \leq k_1 < m \\ 0 \leq k_2 < m}} \sum_{k=1}^{d} \left(\sigma_A^2 V_\varphi \left(\Sigma^{x,x}_{\mathcal{D} \times \mathcal{D}}(t_{k_1}, t_{k_2}) \right) + \sigma_b^2 \right) \Delta x_{t_{k_1+1}}^k \Delta x_{t_{k_2+1}}^k = \lim_{N \to \infty} \mathbb{E} \left[\frac{\langle S_{t_m}^{\mathcal{D},N}(x) - S_0^N, S_{t_m}^{\mathcal{D},N}(x) - S_0^N \rangle}{N} \right] \geq 0$$

$\square$

The second step is another bound, this time from above, which will be uniform on bounded subsets of $\mathbb{X}$.

Lemma C.9. *There is a constant* $C_x > 0$ *independent of* $\mathcal{D}$ *such that*

$$\|\mathcal{K}^{x,x}_{\mathcal{D} \times \mathcal{D}}\|_{\infty, [0,1] \times [0,1]} := \sup_{(s,t) \in [0,1]^2} |\mathcal{K}^{x,x}_{\mathcal{D} \times \mathcal{D}}(s, t)| \leq C_x$$

Moreover C_x *only depends on* $\|x\|_{1-var}$ *and is increasing in it, thus there is a uniform bound on bounded subsets of* $\mathbb{X}$.

Proof. Let us prove this result for $\tilde{\mathcal{K}}^{x,x}$ first. Wlog $(s, t) \in [s_m, s_{m+1}) \times [t_n, t_{n+1})$. Remember how

$$\begin{aligned}
\tilde{\mathcal{K}}^{x,x}_{\mathcal{D} \times \mathcal{D}}(s, t) &= \mathcal{K}^{x,x}_{\mathcal{D} \times \mathcal{D}}(s_m, t_n) \\
&= \sigma_a^2 + \int_{\eta=0}^{s_m} \int_{\tau=0}^{t_n} \left(\sigma_A^2 V_\varphi \left(\tilde{\Sigma}^{x,x}_{\mathcal{D} \times \mathcal{D}}(\eta, \tau) \right) + \sigma_b^2 \right) \langle \dot{x}_\eta, \dot{x}_\tau \rangle_{\mathbb{R}^d} \, d\eta d\tau
\end{aligned}$$

thus

$$\begin{aligned}
|\tilde{\mathcal{K}}_{s,t}| &\leq \sigma_a^2 + \int_{\eta=0}^{s_m} \int_{\tau=0}^{t_n} |\sigma_A^2 V_\varphi(\tilde{\Sigma}_{\eta,\tau})| |\langle \dot{x}_\eta, \dot{x}_\tau \rangle_{\mathbb{R}^d}| d\eta d\tau + \int_{\eta=0}^{s_m} \int_{\tau=0}^{t_n} |\sigma_b^2| |\langle \dot{x}_\eta, \dot{x}_\tau \rangle_{\mathbb{R}^d}| d\eta d\tau \\
&= \sigma_a^2 + \sigma_b^2 \|x\|_{1-var,[0,s_m]} \|x\|_{1-var,[0,s_n]} + \sigma_A^2 \int_{\eta=0}^{s_m} \int_{\tau=0}^{t_n} |V_\varphi(\tilde{\Sigma}_{\eta,\tau})| |\langle \dot{x}_\eta, \dot{x}_\tau \rangle_{\mathbb{R}^d}| d\eta d\tau \\
&\leq \sigma_a^2 + \sigma_b^2 \|x\|_{1-var,[0,1]}^2 + \sigma_A^2 \int_{\eta=0}^{s_m} \int_{\tau=0}^{t_n} \tilde{M} \left(1 + \sqrt{\tilde{\mathcal{K}}_{\eta,\eta}} \right) \left(1 + \sqrt{\tilde{\mathcal{K}}_{\tau,\tau}} \right) |\langle \dot{x}_\eta, \dot{x}_\tau \rangle_{\mathbb{R}^d}| d\eta d\tau \\
&= \sigma_a^2 + \sigma_b^2 \|x\|_{1-var,[0,1]}^2 + \sigma_A^2 \tilde{M} \left(\int_{\eta=0}^{s_m} \left(1 + \sqrt{\tilde{\mathcal{K}}_{\eta,\eta}} \right) |\dot{x}_\eta| d\eta \right) \left(\int_{\tau=0}^{t_n} \left(1 + \sqrt{\tilde{\mathcal{K}}_{\tau,\tau}} \right) |\dot{x}_\tau| d\tau \right)
\end{aligned}$$

In particular

$$\begin{aligned}
\tilde{\mathcal{K}}_{s,s} = |\tilde{\mathcal{K}}_{s,s}| &\leq \sigma_a^2 + \sigma_b^2 \|x\|_{1-var,[0,1]}^2 + \sigma_A^2 \tilde{M} \left| \int_{\eta=0}^{s_m} \left(1 + \sqrt{\tilde{\mathcal{K}}_{\eta,\eta}} \right) |\dot{x}_\eta| d\eta \right|^2 \\
&\leq \sigma_a^2 + \sigma_b^2 \|x\|_{1-var,[0,1]}^2 + \sigma_A^2 \tilde{M} \|x\|_{1-var,[0,t_m]} \left(\int_{\eta=0}^{s_m} \left(1 + \sqrt{\tilde{\mathcal{K}}_{\eta,\eta}} \right)^2 |\dot{x}_\eta| d\eta \right) \\
&\leq \sigma_a^2 + \sigma_b^2 \|x\|_{1-var,[0,1]}^2 + 2\sigma_A^2 \tilde{M} \|x\|_{1-var,[0,1]} \int_{\eta=0}^{s_m} (1 + \tilde{\mathcal{K}}_{\eta,\eta}) |\dot{x}_\eta| d\eta \\
&\leq \sigma_a^2 + \sigma_b^2 \|x\|_{1-var,[0,1]}^2 + 2\sigma_A^2 \tilde{M} \|x\|_{1-var,[0,1]} \int_{\eta=0}^{s} (1 + \tilde{\mathcal{K}}_{\eta,\eta}) |\dot{x}_\eta| d\eta
\end{aligned}$$

and we can use Gronwall Inequality ((Friz & Victoir, 2010), Lemma 3.2) to obtain

$$1 + \tilde{\mathcal{K}}_{s,s} \quad \leq \quad (1 + \sigma_a^2 + \sigma_b^2 \|x\|_{1-\mathrm{var},[0,1]}^2) \exp\{2\sigma_A^2 \tilde{M}\|x\|_{1-\mathrm{var},[0,1]}^2\}$$

hence the thesis with

$$\tilde{C}_x \quad = \quad (1 + \sigma_a^2 + \sigma_b^2\|x\|_{1-\mathrm{var},[0,1]}^2)e^{2\sigma_A^2 \tilde{M}\|x\|_{1-\mathrm{var},[0,1]}^2} - 1$$

In order to extend the bound to $\mathcal{K}^{x,x}$ notice that

$$|\mathcal{K}_{\mathcal{D}\times\mathcal{D}}^{x,x}(s,t)| = |\sigma_a^2 + \int_{\eta=0}^{s}\int_{\tau=0}^{t}\left(\sigma_A^2 V_\varphi\left(\tilde{\Sigma}_{\mathcal{D}\times\mathcal{D}}^{x,\,x}(\eta,\tau)\right) + \sigma_b^2\right)\langle \dot{x}_\eta, \dot{x}_\tau\rangle_{\mathbb{R}^d}\,d\eta d\tau|$$

$$\leq \sigma_a^2 + \left(2\sigma_A^2\tilde{M}(1+\tilde{C}_x) + \sigma_b^2\right)\int_{\eta=0}^{s}\int_{\tau=0}^{t}|\langle \dot{x}_\eta, \dot{x}_\tau\rangle_{\mathbb{R}^d}|d\eta d\tau$$

$$\leq \sigma_a^2 + \left(2\sigma_A^2\tilde{M}(1+\tilde{C}_x) + \sigma_b^2\right)\|x\|_{1-var,[0,1]}^2 =: C_x$$

$\square$

Thus finally we can conclude, as promised, that

Lemma C.10. *There is an $R = R_x \in \mathbb{R}$ such that $\tilde{\Sigma}_{\mathcal{D}\times\mathcal{D}}^{x,x} \in \mathrm{PSD}_2(R)$ for every $\mathcal{D}$. Moreover, as before, this R_x only depends on $\|x\|_{1-var}$ and is increasing in it.*

Proof. The diagonal elements $\tilde{\mathcal{K}}_{\mathcal{D}\times\mathcal{D}}^{x,x}(s,s)$ are bounded above by a common constant C_x and below by σ_a^2 We thus can simply choose $R_x = C_x \vee \sigma_a^{-2}$. $\square$

We are ready to extend the result to the general case.

Proposition C.11. *Fix a $\alpha > 0$. There exist a constant C_α such that for all $x, y \in \mathbb{X}$ with $\|x\|_{1-var}, \|y\|_{1-var} \leq \alpha$ and partitions $\mathcal{D}, \mathcal{D}'$ it holds*

$$\sup_{0\leq t,s\leq 1}|\mathcal{K}_{\mathcal{D}\times\mathcal{D}'}^{x,y}(s,t)| \leq C_\alpha$$

Moreover for $R_\alpha := C_\alpha \vee \sigma_a^{-2}$ we get

$$\tilde{\Sigma}_{\mathcal{D}\times\mathcal{D}'}^{x,y}(s,t) \in \mathrm{PSD}(R_\alpha)$$

Proof. Remember how the maps $\tilde{\Sigma}_{\mathcal{D}\times\mathcal{D}'}^{x,y}$ are PSD_2 since they arise as such from Tensor limiting arguments, in particular by positive semidefinitiveness and the previous bounds

$$|\tilde{\mathcal{K}}_{\mathcal{D}\times\mathcal{D}'}^{x,y}(s,t)| \leq \sqrt{\tilde{\mathcal{K}}_{\mathcal{D}\times\mathcal{D}}^{x,x}(s,s)\tilde{\mathcal{K}}_{\mathcal{D}'\times\mathcal{D}'}^{y,y}(t,t)} \leq \sqrt{\tilde{C}_x\tilde{C}_y} \leq \tilde{C}_x \vee \tilde{C}_y$$

For $K_{x,y}(t)$ we proceed similarly to before:

$$|\mathcal{K}_{\mathcal{D}\times\mathcal{D}'}^{x,y}(s,t)| = |\sigma_a^2 + \int_{\eta=0}^{s}\int_{\tau=0}^{t}\left(\sigma_A^2 V_\varphi\left(\tilde{\Sigma}_{\mathcal{D}\times\mathcal{D}'}^{x,\,y}(\eta,\tau)\right) + \sigma_b^2\right)\langle \dot{x}_\eta, \dot{y}_\tau\rangle_{\mathbb{R}^d}\,d\eta d\tau|$$

$$\leq \sigma_a^2 + \left(2\sigma_A^2\tilde{M}(1+\tilde{C}_x \vee \tilde{C}_y) + \sigma_b^2\right)\int_{\eta=0}^{s}\int_{\tau=0}^{t}|\langle \dot{x}_\eta, \dot{y}_\tau\rangle_{\mathbb{R}^d}|d\eta d\tau$$

$$\leq \sigma_a^2 + \left(2\sigma_A^2\tilde{M}(1+\tilde{C}_x \vee \tilde{C}_y) + \sigma_b^2\right)\|x\|_{1-var,[0,1]}\|y\|_{1-var,[0,1]} \leq C_\alpha$$

The second part follows from the definition of $\mathrm{PSD}(R_\alpha)$. $\square$

Proof of Theorem C.7.

The central idea will be proving that, given any sequence $\mathcal{D}_n \times \mathcal{D}'_n$ with $|\mathcal{D}_n| \vee |\mathcal{D}'_n| \downarrow 0$, the maps $\{\mathcal{K}^{x,y}_{\mathcal{D}_M \times \mathcal{D}'_n}\}_n$ form a Cauchy sequence in the Banach space $(C^0([0,1] \times [0,1]; \mathbb{R}), \|\cdot\|_\infty)$. By completeness of this space the sequence will admit a limit $\mathcal{K}^{x,y}$, which we will prove to be independent from the chosen sequence. Finally we will prove that this limit is the sought after kernel.

Note that trivially the set $\{x, y\}$ is bounded in $\mathbb{X}$, we can thus fix two constants $C_{x,y}, R_{x,y}$ with the properties given in Proposition C.11. We will often, for ease of notation, refer to them as C and R.

Part I: how close are $\mathcal{K}_{\mathcal{D} \times \mathcal{D}'}$ and $\tilde{\mathcal{K}}_{\mathcal{D} \times \mathcal{D}'}$?

We have

$$
|\mathcal{K}^{x,y}_{\mathcal{D} \times \mathcal{D}'}(s,t) - \tilde{\mathcal{K}}^{x,y}_{\mathcal{D} \times \mathcal{D}'}(s,t)|
$$
$$
= | \int_{\eta=0}^{s} \int_{\tau=0}^{t} (\sigma_A^2 V_\varphi(\widetilde{\Sigma}^{x,y}_{\mathcal{D} \times \mathcal{D}'}(\eta, \tau)) + \sigma_b^2) \langle \dot{x}_\eta, \dot{y}_\tau \rangle_{\mathbb{R}^d} \, d\eta d\tau
$$
$$
- \int_{\eta=0}^{s_m} \int_{\tau=0}^{t_n} (\sigma_A^2 V_\varphi(\widetilde{\Sigma}^{x,y}_{\mathcal{D} \times \mathcal{D}'}(\eta, \tau)) + \sigma_b^2) \langle \dot{x}_\eta, \dot{y}_\tau \rangle_{\mathbb{R}^d} \, d\eta d\tau |
$$
$$
= \left| \int_{\Omega_{s,t}} (\sigma_A^2 V_\varphi(\widetilde{\Sigma}^{x,y}_{\mathcal{D} \times \mathcal{D}'}(\eta, \tau)) + \sigma_b^2) \langle \dot{x}_\eta, \dot{y}_\tau \rangle_{\mathbb{R}^d} \, d\eta d\tau \right|
$$

with

$$
\Omega_{s,t} := [0,s] \times [0,t] \setminus [0, s_m] \times [0, t_n]
$$

Note that

$$
\Omega_{s,t} \subseteq ([s_m, s_{m+1}] \times [0,1]) \cup ([0,1] \times [t_n, t_{n+1}])
$$

thus

$$
\mathcal{L}^2(\Omega_{s,t}) \le |\mathcal{D}| + |\mathcal{D}'| \le 2(|\mathcal{D}| \vee |\mathcal{D}'|)
$$

hence

$$
|\mathcal{K}^{x,y}_{\mathcal{D} \times \mathcal{D}'}(s,t) - \tilde{\mathcal{K}}^{x,y}_{\mathcal{D} \times \mathcal{D}'}(s,t)| \le \int_{\Omega_{s,t}} |\sigma_A^2 V_\varphi(\widetilde{\Sigma}^{x,y}_{\mathcal{D} \times \mathcal{D}'}(\eta, \tau)) + \sigma_b^2| |\langle \dot{x}_\eta, \dot{x}_\tau \rangle_{\mathbb{R}^d}| \, d\eta d\tau
$$
$$
\le \sigma_A^2 \int_{\Omega_{s,t}} \tilde{M} \left(1 + \sqrt{\mathcal{K}^{x,x}_{\mathcal{D} \times \mathcal{D}}(\eta, \eta)}\right) \left(1 + \sqrt{\mathcal{K}^{y,y}_{\mathcal{D}' \times \mathcal{D}'}(\tau, \tau)}\right) |\langle \dot{x}_\eta, \dot{y}_\tau \rangle_{\mathbb{R}^d}| \, d\eta d\tau
$$
$$
+ \sigma_b^2 \int_{\Omega_{s,t}} |\langle \dot{x}_\eta, \dot{y}_\tau \rangle_{\mathbb{R}^d}| \, d\eta d\tau
$$
$$
\le \sigma_A^2 \tilde{M} \left(1 + \sqrt{C_{x,y}}\right)^2 \int_{\Omega_{s,t}} |\langle \dot{x}_\eta, \dot{y}_\tau \rangle_{\mathbb{R}^d}| \, d\eta d\tau + \sigma_b^2 \int_{\Omega_{s,t}} |\langle \dot{x}_\eta, \dot{y}_\tau \rangle_{\mathbb{R}^d}| \, d\eta d\tau
$$
$$
\le (2\sigma_A^2 \tilde{M}(1 + C_{x,y}) + \sigma_b^2) \int_{[0,1] \times [0,1]} \mathbb{I}_{\Omega_{s,t}}(\eta, \tau) |\langle \dot{x}_\eta, \dot{y}_\tau \rangle_{\mathbb{R}^d}| \, d\eta d\tau
$$
$$
\le (2\sigma_A^2 \tilde{M}(1 + C_{x,y}) + \sigma_b^2) \sqrt{\mathcal{L}^2(\Omega_{s,t})} \sqrt{\int_{[0,1] \times [0,1]} |\langle \dot{x}_\eta, \dot{y}_\tau \rangle_{\mathbb{R}^d}|^2 d\eta d\tau}
$$
$$
\le (2\sigma_A^2 \tilde{M}(1 + C_{x,y}) + \sigma_b^2) \sqrt{\mathcal{L}^2(\Omega_{s,t})} \sqrt{\int_{[0,1] \times [0,1]} |\dot{x}_\eta|^2 |\dot{y}_\tau|^2 d\eta d\tau}
$$
$$
\le (2\sigma_A^2 \tilde{M}(1 + C_{x,y}) + \sigma_b^2) \sqrt{\mathcal{L}^2(\Omega_{s,t})} \sqrt{\|x\|_{\mathbb{X}}^2 \|y\|_{\mathbb{X}}^2}
$$
$$
\le 2(2\sigma_A^2 \tilde{M}(1 + C_{x,y}) + \sigma_b^2) \sqrt{|\mathcal{D}| \vee |\mathcal{D}'|} \sqrt{\|x\|_{\mathbb{X}}^2 \|y\|_{\mathbb{X}}^2}
$$

In particular as $(|\mathcal{D}| \vee |\mathcal{D}'|) \to 0$ we have

$$\sup_{[0,1]\times[0,1]} |\mathcal{K}^{x,y}_{\mathcal{D}\times\mathcal{D}'}(s,t) - \tilde{\mathcal{K}}^{x,y}_{\mathcal{D}\times\mathcal{D}'}(s,t)| \leq$$

$$2(2\sigma_A^2 \tilde{M}(1 + C_{x,y}) + \sigma_b^2)\sqrt{|\mathcal{D}| \vee |\mathcal{D}'|}\sqrt{\|x\|_{\mathbb{X}}^2 \|y\|_{\mathbb{X}}^2} \to 0$$

Notice that, since we can repeat the argument for $\mathcal{K}^{x,x}_{\mathcal{D}\times\mathcal{D}}$ and $\mathcal{K}^{y,y}_{\mathcal{D}'\times\mathcal{D}'}$, we have

$$\sup_{[0,1]\times[0,1]} |\Sigma^{x,y}_{\mathcal{D}\times\mathcal{D}'}(s,t) - \tilde{\Sigma}^{x,y}_{\mathcal{D}\times\mathcal{D}'}(s,t)|_\infty \leq$$

$$2(2\sigma_A^2 \tilde{M}(1 + C_{x,y}) + \sigma_b^2)\sqrt{|\mathcal{D}| \vee |\mathcal{D}'|}(\|x\|_{\mathbb{X}}^2 \vee \|y\|_{\mathbb{X}}^2) =: \Gamma_{x,y}\sqrt{|\mathcal{D}| \vee |\mathcal{D}'|}$$

Part II : Cauchy Bounds
Consider now another partition $\check{\mathcal{D}} \times \check{\mathcal{D}}'$, we have

$$|\mathcal{K}^{x,y}_{\check{\mathcal{D}}\times\check{\mathcal{D}}'}(s,t) - \mathcal{K}^{x,y}_{\mathcal{D}\times\mathcal{D}'}(s,t)|$$

$$= \left|\sigma_A^2 \int_{\eta=0}^{s} \int_{\tau=0}^{t} [V_\varphi(\tilde{\Sigma}^{x,y}_{\check{\mathcal{D}}\times\check{\mathcal{D}}'}(\eta,\tau)) - V_\varphi(\tilde{\Sigma}^{x,y}_{\mathcal{D}\times\mathcal{D}'}(\eta,\tau))] \langle \dot{x}_\eta, \dot{y}_\tau \rangle_{\mathbb{R}^d} \, d\eta d\tau \right|$$

$$\leq \sigma_A^2 \int_{\eta=0}^{s} \int_{\tau=0}^{t} |V_\varphi(\tilde{\Sigma}^{x,y}_{\check{\mathcal{D}}\times\check{\mathcal{D}}'}(\eta,\tau)) - V_\varphi(\tilde{\Sigma}^{x,y}_{\mathcal{D}\times\mathcal{D}'}(\eta,\tau))|| \langle \dot{x}_\eta, \dot{y}_\tau \rangle_{\mathbb{R}^d} | \, d\eta d\tau$$

$$\leq \sigma_A^2 \int_{\eta=0}^{s} \int_{\tau=0}^{t} k_R |\tilde{\Sigma}^{x,y}_{\check{\mathcal{D}}\times\check{\mathcal{D}}'}(\eta,\tau) - \tilde{\Sigma}^{x,y}_{\mathcal{D}\times\mathcal{D}'}(\eta,\tau)|_\infty | \langle \dot{x}_\eta, \dot{y}_\tau \rangle_{\mathbb{R}^d} | \, d\eta d\tau$$

$$\leq \sigma_A^2 k_R \int_{\eta=0}^{s} \int_{\tau=0}^{t} |\tilde{\Sigma}_{\check{\mathcal{D}}\times\check{\mathcal{D}}'}(\eta,\tau) - \Sigma_{\check{\mathcal{D}}\times\check{\mathcal{D}}'}(\eta,\tau)|_\infty | \langle \dot{x}_\eta, \dot{y}_\tau \rangle_{\mathbb{R}^d} | \, d\eta d\tau$$

$$+ \sigma_A^2 k_R \int_{\eta=0}^{s} \int_{\tau=0}^{t} |\Sigma_{\check{\mathcal{D}}\times\check{\mathcal{D}}'}(\eta,\tau) - \Sigma_{\mathcal{D}\times\mathcal{D}'}(\eta,\tau)|_\infty | \langle \dot{x}_\eta, \dot{y}_\tau \rangle_{\mathbb{R}^d} | \, d\eta d\tau$$

$$+ \sigma_A^2 k_R \int_{\eta=0}^{s} \int_{\tau=0}^{t} |\Sigma_{\mathcal{D}\times\mathcal{D}'}(\eta,\tau) - \tilde{\Sigma}_{\mathcal{D}\times\mathcal{D}'}(\eta,\tau)|_\infty | \langle \dot{x}_\eta, \dot{y}_\tau \rangle_{\mathbb{R}^d} | \, d\eta d\tau$$

$$\leq \sigma_A^2 \Gamma_{x,y}\sqrt{|\check{\mathcal{D}}| \vee |\check{\mathcal{D}}'|} \int_{\eta=0}^{s} \int_{\tau=0}^{t} | \langle \dot{x}_\eta, \dot{y}_\tau \rangle_{\mathbb{R}^d} | \, d\eta d\tau$$

$$+ \sigma_A^2 k_R \int_{\eta=0}^{s} \int_{\tau=0}^{t} |\Sigma_{\check{\mathcal{D}}\times\check{\mathcal{D}}'}(\eta,\tau) - \Sigma_{\mathcal{D}\times\mathcal{D}'}(\eta,\tau)|_\infty | \langle \dot{x}_\eta, \dot{y}_\tau \rangle_{\mathbb{R}^d} | \, d\eta d\tau$$

$$+ \sigma_A^2 \Gamma_{x,y}\sqrt{|\mathcal{D}| \vee |\mathcal{D}'|} \int_{\eta=0}^{s} \int_{\tau=0}^{t} | \langle \dot{x}_\eta, \dot{y}_\tau \rangle_{\mathbb{R}^d} | \, d\eta d\tau$$

$$\leq \sigma_A^2 2\Gamma_{x,y}\sqrt{|\mathcal{D}| \vee |\mathcal{D}'| \vee |\check{\mathcal{D}}| \vee |\check{\mathcal{D}}'|}(\|x\|_{1-\text{var},[0,1]}^2 \vee \|y\|_{1-\text{var},[0,1]}^2)$$

$$+ \sigma_A^2 k_R \int_{\eta=0}^{s} \int_{\tau=0}^{t} |\Sigma^{x,y}_{\check{\mathcal{D}}\times\check{\mathcal{D}}'}(\eta,\tau) - \Sigma^{x,y}_{\mathcal{D}\times\mathcal{D}'}(\eta,\tau)|_\infty | \langle \dot{x}_\eta, \dot{y}_\tau \rangle_{\mathbb{R}^d} | \, d\eta d\tau$$

Assume to be in the case $x = y$, $\mathcal{D} = \mathcal{D}'$, $\check{\mathcal{D}} = \check{\mathcal{D}}'$ and define the following quantity:

$$\Xi_t := \sup_{0\leq\eta,\tau\leq t} |\Sigma^{x,x}_{\check{\mathcal{D}}\times\check{\mathcal{D}}}(\eta,\tau) - \Sigma^{x,x}_{\mathcal{D}\times\mathcal{D}}(\eta,\tau)|_\infty = \sup_{0\leq\eta,\tau\leq t} |\mathcal{K}^{x,x}_{\check{\mathcal{D}}\times\check{\mathcal{D}}}(\eta,\tau) - \mathcal{K}^{x,x}_{\mathcal{D}\times\mathcal{D}}(\eta,\tau)|$$

One has, from the previous inequality, that

$$\Xi_t \leq \sigma_A^2 2\Gamma_{x,x}\sqrt{|\check{\mathcal{D}}| \vee |\mathcal{D}|}\|x\|_{1-\text{var},[0,1]}^2$$

$$+ \sigma_A^2 k_R \int_{\eta=0}^{s} \int_{\tau=0}^{t} |\Sigma_{\check{\mathcal{D}}\times\check{\mathcal{D}}}(\eta,\tau) - \Sigma_{\mathcal{D}\times\mathcal{D}}(\eta,\tau)|_\infty | \langle \dot{x}_\eta, \dot{x}_\tau \rangle_{\mathbb{R}^d} | \, d\eta d\tau$$

and since

$$\int_{\eta=0}^{s}\int_{\tau=0}^{t}|\Sigma_{\check{\mathcal{D}}\times\check{\mathcal{D}}}(\eta,\tau)-\Sigma_{\mathcal{D}\times\mathcal{D}}(\eta,\tau)|_{\infty}|\langle\dot{x}_{\eta},\dot{x}_{\tau}\rangle_{\mathbb{R}^d}|d\eta d\tau$$

$$\leq\int_{\eta=0}^{s}\int_{\tau=0}^{t}\Xi_{\eta\vee\tau}|\langle\dot{x}_{\eta},\dot{x}_{\tau}\rangle_{\mathbb{R}^d}|d\eta d\tau$$

$$=\int_{\eta=0}^{s}\int_{\tau=0}^{\eta}\Xi_{\eta}|\langle\dot{x}_{\eta},\dot{x}_{\tau}\rangle_{\mathbb{R}^d}|d\eta d\tau+\int_{\eta=0}^{s}\int_{\tau=\eta}^{t}\Xi_{\tau}|\langle\dot{x}_{\eta},\dot{x}_{\tau}\rangle_{\mathbb{R}^d}|d\eta d\tau$$

$$\leq\int_{\eta=0}^{s}\|x\|_{1-\mathrm{var},[0,\eta]}\Xi_{\eta}|\dot{x}_{\eta}|d\eta+\int_{\tau=0}^{t}\int_{\eta=0}^{\sigma}\Xi_{\tau}|\dot{x}_{\eta}||\dot{x}_{\tau}|d\eta d\tau$$

$$=\int_{\eta=0}^{s}\|x\|_{1-\mathrm{var},[0,\eta]}\Xi_{\eta}|\dot{x}_{\eta}|d\eta+\int_{\tau=0}^{t}\|x\|_{1-\mathrm{var},[0,\tau]}\Xi_{\tau}|\dot{x}_{\tau}|d\tau$$

$$\leq 2\|x\|_{1-\mathrm{var},[0,1]}\int_{\eta=0}^{s}\Xi_{\eta}|\dot{x}_{\eta}|d\eta$$

we get

$$\Xi_t \quad \leq \quad \sigma_A^2 2\Gamma_{x,x}\|x\|_{1-\mathrm{var},[0,1]}^2\sqrt{|\check{\mathcal{D}}|\vee|\mathcal{D}|}$$

$$+2\sigma_A^2 k_R\|x\|_{1-\mathrm{var},[0,1]}\int_{\eta=0}^{s}\Xi_{\eta}|\dot{x}_{\eta}|d\eta$$

thus, by Gronwall,

$$\Xi_t \leq 2\sigma_A^2\Gamma_{x,x}\|x\|_{1-\mathrm{var},[0,1]}^2\sqrt{|\check{\mathcal{D}}|\vee|\mathcal{D}|}\cdot e^{2\sigma_A^2 k_R\|x\|_{1-\mathrm{var},[0,1]}^2}$$

Coming back to the general case and setting

$$\Lambda_{x,y}:=\|x\|_{1-\mathrm{var},[0,1]}^2\vee\|y\|_{1-\mathrm{var},[0,1]}^2$$

we can now say that

$$\Xi_t^{x,y}:=\sup_{0\leq\eta,\tau\leq t}|\Sigma_{\check{\mathcal{D}}\times\check{\mathcal{D}}'}(\eta,\tau)-\Sigma_{\mathcal{D}\times\mathcal{D}'}(\eta,\tau)|_{\infty}$$

$$\leq 2\sigma_A^2\Gamma_{x,y}\sqrt{|\check{\mathcal{D}}|\vee|\check{\mathcal{D}}'|\vee|\mathcal{D}|\vee|\mathcal{D}'|}\Lambda_{x,y}(1+e^{2\sigma_A^2 k_R\Lambda_{x,y}})$$

$$+\sigma_A^2 k_R\int_{\eta=0}^{s}\int_{\tau=0}^{t}|\Sigma_{\check{\mathcal{D}}\times\check{\mathcal{D}}'}(\eta,\tau)-\Sigma_{\mathcal{D}\times\mathcal{D}'}(\eta,\tau)|_{\infty}|\langle\dot{x}_{\eta},\dot{y}_{\tau}\rangle_{\mathbb{R}^d}|d\eta d\tau$$

and since

$$\int_{\eta=0}^{s}\int_{\tau=0}^{t}|\Sigma_{\check{\mathcal{D}}\times\check{\mathcal{D}}'}(\eta,\tau)-\Sigma_{\mathcal{D}\times\mathcal{D}'}(\eta,\tau)|_{\infty}|\langle\dot{x}_{\eta},\dot{y}_{\tau}\rangle_{\mathbb{R}^d}|d\eta d\tau$$

$$\leq\int_{\eta=0}^{s}\int_{\tau=0}^{t}\Xi_{\eta\vee\tau}^{x,y}|\langle\dot{x}_{\eta},\dot{y}_{\tau}\rangle_{\mathbb{R}^d}|d\eta d\tau$$

$$=\int_{\eta=0}^{s}\int_{\tau=0}^{\eta}\Xi_{\eta}^{x,y}|\langle\dot{x}_{\eta},\dot{y}_{\tau}\rangle_{\mathbb{R}^d}|d\eta d\tau+\int_{\eta=0}^{s}\int_{\tau=\eta}^{t}\Xi_{\tau}^{x,y}|\langle\dot{x}_{\eta},\dot{y}_{\tau}\rangle_{\mathbb{R}^d}|d\eta d\tau$$

$$\leq\int_{\eta=0}^{s}\|y\|_{1-\mathrm{var},[0,\eta]}\Xi_{\eta}^{x,y}|\dot{x}_{\eta}|d\eta+\int_{\tau=0}^{t}\int_{\eta=0}^{\sigma}\Xi_{\tau}^{x,y}|\dot{x}_{\eta}||\dot{y}_{\tau}|d\eta d\tau$$

$$=\int_{\eta=0}^{s}\|y\|_{1-\mathrm{var},[0,\eta]}\Xi_{\eta}^{x,y}|\dot{x}_{\eta}|d\eta+\int_{\tau=0}^{t}\|x\|_{1-\mathrm{var},[0,\tau]}\Xi_{\tau}^{x,y}|\dot{y}_{\tau}|d\tau$$

$$\leq(\|x\|_{1-\mathrm{var},[0,1]}\vee\|y\|_{1-\mathrm{var},[0,1]})\int_{\eta=0}^{s}\Xi_{\eta}^{x,y}(|\dot{x}_{\eta}|+|\dot{y}_{\eta}|)d\eta$$

we get

$$\Xi_t^{x,y} \leq 2\sigma_A^2 \Gamma_{x,y} \Lambda_{x,y} \sqrt{|\check{\mathcal{D}}| \vee |\check{\mathcal{D}}'| \vee |\mathcal{D}| \vee |\mathcal{D}'|}(1 + e^{2\sigma_A^2 k_R \Lambda_{x,y}})$$

$$+ 2\sigma_A^2 k_R(\|x\|_{1-\text{var},[0,1]} \vee \|y\|_{1-\text{var},[0,1]}) \int_{\eta=0}^{s} \Xi_\eta^{x,y}(|\dot{x}_\eta| + |\dot{y}_\eta|)d\eta$$

thus, by Gronwall,

$$\Xi_t \leq 2\sigma_A^2 \Gamma_{x,y} \Lambda_{x,y} \sqrt{|\check{\mathcal{D}}| \vee |\check{\mathcal{D}}'| \vee |\mathcal{D}| \vee |\mathcal{D}'|}(1 + e^{2\sigma_A^2 k_R \Lambda_{x,y}})$$

$$\cdot e^{2\sigma_A^2 k_R(\|x\|_{1-\text{var},[0,1]} \vee \|y\|_{1-\text{var},[0,1]})(\|x\|_{1-\text{var},[0,t]} + \|y\|_{1-\text{var},[0,t]})}$$

hence

$$\|\mathcal{K}_{\check{\mathcal{D}} \times \check{\mathcal{D}}'}^{x,y}(s,t) - \mathcal{K}_{\mathcal{D} \times \mathcal{D}'}^{x,y}(s,t)\|_{\infty,[0,1]^2} \leq$$

$$2\sigma_A^2 \Gamma_{x,y} \Lambda_{x,y} \sqrt{|\check{\mathcal{D}}| \vee |\check{\mathcal{D}}'| \vee |\mathcal{D}| \vee |\mathcal{D}'|} \cdot (1 + e^{2\sigma_A^2 k_R \Lambda_{x,y}}) \cdot e^{2\sigma_A^2 k_R \Lambda_{x,y}}$$

Part III : Existence and Uniqueness of limit Given any sequence of partitions $\{\mathcal{D}_n \times \mathcal{D}'_n\}$ with $|\mathcal{D}_n| \vee |\mathcal{D}'_n| \to 0$, due to the bounds we have just proven we have

$$\{\mathcal{K}_{\mathcal{D}_M \times \mathcal{D}'_n}^{x,y}\} \text{is a Cauchy sequence in } (C^0([0,1] \times [0,1]; \mathbb{R}), \|\cdot\|_\infty)$$

and the limit $\mathcal{K}_\varphi^{x,y}$ does not depend on the sequence (i.e. the limit exists and is unique).

The limit is indeed unique: assume $K^{x,y}$ and $G^{x,y}$ are limits along two different sequences of partitions $\{\mathcal{D}_n \times \mathcal{D}'_n\}$ and $\{\mathcal{G}_n \times \mathcal{G}'_n\}$; then the sequence $\{\mathcal{P}_n \times \mathcal{P}'_n\}$ such that $\mathcal{P}_{2n} \times \mathcal{P}'_{2n} := \mathcal{D}_n \times \mathcal{D}'_n$ and $\mathcal{P}_{2n+1} \times \mathcal{P}'_{2n+1} := \mathcal{G}_n \times \mathcal{G}'_n$ is still such that $|\mathcal{P}_n| \vee |\mathcal{P}'_n| \to 0$ hence the associated kernels have a limit which must be equal to both $K^{x,y}$ and $G^{x,y}$.

Since PSD matrices form a closed set we moreover have that the matrices $\Sigma_\varphi^{x,y}(s,t)$ obtained as limits using the previous result are all PSD. We can actually say more: they belong to $\text{PSD}(R)$.

Part IV : Limit Kernel solves Equation (22) We can finally conclude by proving that $\mathcal{K}_\varphi^{x,y}$ is, in fact, a solution of the PDE:

$$\left| \mathcal{K}_\varphi^{x,y}(s,t) - \sigma_a^2 + \int_{\eta=0}^{s} \int_{\tau=0}^{t} (\sigma_A^2 V_\varphi(\Sigma_\varphi^{x,y}(\eta,\tau)) + \sigma_b^2) \langle \dot{x}_\eta, \dot{y}_\tau \rangle_{\mathbb{R}^d} d\eta d\tau \right|$$

$$\leq |\mathcal{K}_\varphi^{x,y}(s,t) - \mathcal{K}_{\mathcal{D} \times \mathcal{D}'}^{x,y}(s,t)|$$

$$+ \sigma_A^2 \int_{\eta=0}^{s} \int_{\tau=0}^{t} |V_\varphi(\tilde{\Sigma}_{\mathcal{D} \times \mathcal{D}'}^{x,y}(\eta,\tau)) - V_\varphi(\Sigma_\varphi^{x,y}(\eta,\tau))| |\langle \dot{x}_\eta, \dot{y}_\tau \rangle_{\mathbb{R}^d}| d\eta d\tau$$

$$\leq o(1) + \sigma_A^2 k_R \int_{\eta=0}^{s} \int_{\tau=0}^{t} |\tilde{\Sigma}_{\mathcal{D} \times \mathcal{D}'}^{x,y}(\eta,\tau) - \Sigma_\varphi^{x,y}(\eta,\tau)|_\infty |\langle \dot{x}_\eta, \dot{y}_\tau \rangle_{\mathbb{R}^d}| d\eta d\tau$$

$$\leq o(1) + \sigma_A^2 k_R \int_{\eta=0}^{s} \int_{\tau=0}^{t} |\tilde{\Sigma}_{\mathcal{D} \times \mathcal{D}'}^{x,y}(\eta,\tau) - \Sigma_{\mathcal{D} \times \mathcal{D}'}^{x,y}(\eta,\tau)|_\infty |\langle \dot{x}_\eta, \dot{y}_\tau \rangle_{\mathbb{R}^d}| d\eta d\tau$$

$$+ \sigma_A^2 k_R \int_{\eta=0}^{s} \int_{\tau=0}^{t} |\Sigma_{\mathcal{D} \times \mathcal{D}'}^{x,y}(\eta,\tau) - \Sigma_\varphi^{x,y}(\eta,\tau)|_\infty |\langle \dot{x}_\eta, \dot{y}_\tau \rangle_{\mathbb{R}^d}| d\eta d\tau$$

$$= o(1) + o(1) + o(1) = o(1)$$

thus

$$\left| \mathcal{K}_\varphi^{x,y}(s,t) - \sigma_a^2 + \int_{\eta=0}^{s} \int_{\tau=0}^{s} (\sigma_A^2 V_\varphi(\Sigma_\varphi^{x,y}(\eta,\tau)) + \sigma_b^2) \langle \dot{x}_\eta, \dot{y}_\tau \rangle_{\mathbb{R}^d} d\eta d\tau \right| = 0$$

i.e

$$\mathcal{K}_\varphi^{x,y}(s,t) = \sigma_a^2 + \int_{\eta=0}^{s} \int_{\tau=0}^{t} (\sigma_A^2 V_\varphi(\Sigma_\varphi^{x,y}(\eta,\tau)) + \sigma_b^2) \langle \dot{x}_\eta, \dot{y}_\tau \rangle_{\mathbb{R}^d} d\eta d\tau$$

- Part V : Uniformity in x, y on bounded sets

Note that, once again, the bounds we have just proven only depend on the norms $\|\cdot\|_{1-var,[0,1]}$ and $\|\cdot\|_{\mathbb{X}}$ of the paths, and they do so in an increasing manner. Since $\|\cdot\|_{1-var,[0,1]} \leq \|\cdot\|_{\mathbb{X}}$ this means that we have uniform convergence rates on bounded sets on $\mathbb{X}$. $\qquad\square$

Remark. We could have stated this result for $x \in W^{1,1}([0,1];\mathbb{R}^d)$ thus requiring the derivative to only be in L^1, this is done in an analogous way to the proof of Theorem (3.1) using Ascoli-Arzela. The cost to pay for the decreased regularity of the driving path is the loss of uniform convergence bounds.

Remark. Being careful one could maintain uniform bounds, at the cost of slower convergence, and state the result for $x \in W^{1,1+\epsilon}([0,1];\mathbb{R}^d)$ for any $\epsilon > 0$. More specifically the bound would be proportional to $|\mathcal{D}|^{\frac{\epsilon}{1+\epsilon}}$.

Proposition C.12 (Uniqueness). *Under the previous assumptions on the activation function, fix $x, y \in \mathbb{X}$. Then any two triples*

$$K(s,t) := (K^{x,x}(s,t), K^{x,y}(s,t), K^{y,y}(s,t))$$
$$G(s,t) := (G^{x,x}(s,t), G^{x,y}(s,t), G^{y,y}(s,t))$$

defined on $[0,1] \times [0,1]$, satisfying Equation (22) and such that

$$K^{x,x}(s,s), K^{y,y}(t,t), G^{x,x}(s,s), G^{y,y}(t,t) > 0$$

for all $t \in [0,1]$ must be equal.

Proof. We will write $|K|_\infty := |K^{x,x}| \vee |K^{x,y}| \vee |K^{y,y}|$. To satisfy equation (22) the associated covariance matrices $\Sigma_K(\eta,\tau), \Sigma_G(\eta,\tau)$ must be always PSD_2. Using the assumed bound from below the one from above given by continuity we can assume that they uniformly in time are contained in some $PSD_2(\bar{R})$. This true for all choices $(x,x), (x,y), (y,y)$.

$$|K^{x,y}(s,t) - G^{x,y}(s,t)|$$
$$= |\int_{\eta=0}^{s} \int_{\tau=0}^{t} \sigma_A^2 (V_\varphi(\Sigma_K^{x,y}(\eta,\tau)) - V_\varphi(\Sigma_G^{x,y}(\eta,\tau))) \langle \dot{x}_\eta, \dot{y}_\tau \rangle_{\mathbb{R}^d} \, d\eta d\tau|$$
$$\leq \int_{\eta=0}^{s} \int_{\tau=0}^{t} \sigma_A^2 |V_\varphi(\Sigma_K^{x,y}(\eta,\tau)) - V_\varphi(\Sigma_G^{x,y}(\eta,\tau))| |\langle \dot{x}_\eta, \dot{y}_\tau \rangle_{\mathbb{R}^d}| d\eta d\tau$$
$$\leq \int_{\eta=0}^{s} \int_{\tau=0}^{t} \sigma_A^2 k_{\bar{R}} |\Sigma_K^{x,y}(\eta,\tau) - \Sigma_G^{x,y}(\eta,\tau)|_\infty |\langle \dot{x}_\eta, \dot{y}_\tau \rangle_{\mathbb{R}^d}| d\eta d\tau$$
$$\leq \sigma_A^2 k_{\bar{R}} \int_{\eta=0}^{s} \int_{\tau=0}^{t} (\sup_{0 \leq t_1, s_1 \leq \eta \vee \tau} |K(t_1,s_1) - G(t_1,s_1)|_\infty) |\langle \dot{x}_\eta, \dot{y}_\tau \rangle_{\mathbb{R}^d}| d\eta d\tau$$

Moreover this holds substituting (x,y) with (x,x) and (y,y).

Let $\Xi_t := \sup_{0 \leq t_1, s_1 \leq \eta t} |K(t_1,s_1) - G(t_1,s_1)|_\infty)$ then

$$|K(s,t) - G(s,t)|_\infty \leq \sigma_A^2 k_{\bar{R}} \int_{\eta=0}^{s} \int_{\tau=0}^{t} \Xi_{\eta \vee \tau} |\langle \dot{x}_\eta, \dot{y}_\tau \rangle_{\mathbb{R}^d}| d\eta d\tau$$
$$\leq \int_{\eta=0}^{s} \int_{\tau=0}^{t} \Xi_{\eta \vee \tau} |\dot{x}_\eta||\dot{y}_\tau| d\eta d\tau$$

thus

$$\Xi_t \leq \int_{\eta=0}^{s} \int_{\tau=0}^{t} \Xi_{\eta \vee \tau} |\dot{x}_\eta||\dot{y}_\tau| d\eta d\tau$$
$$\leq \int_{\eta=0}^{s} \int_{\tau=0}^{\eta} \Xi_\eta |\dot{x}_\eta||\dot{y}_\tau| d\eta d\tau + \int_{\eta=0}^{s} \int_{\tau=\eta}^{t} \Xi_\tau |\dot{x}_\eta||\dot{y}_\tau| d\eta d\tau$$
$$\leq \|y\|_{1-var,[0,1]} \int_{\eta=0}^{s} \Xi_\eta |\dot{x}_\eta| d\eta + \|x\|_{1-var,[0,1]} \int_{\tau=0}^{t} \Xi_\tau |\dot{y}_\tau| d\tau$$
$$\leq (\|x\|_{1-var,[0,1]} \vee \|y\|_{1-var,[0,1]}) \int_{\eta=0}^{s} \Xi_\eta (|\dot{x}_\eta| + |\dot{y}_\eta|) d\eta$$

44

By Gronwall we finally conclude that $\Xi_t = 0$ for all t, which concludes the proof.z $\qquad\square$

C.1.3. PROOF OF THEOREM 4.1: PART 1

It is finally time to prove the first part of the main result of the paper, which we restate below for the reader's convenience.

Fix partitions $\{\mathcal{D}_M\}_{M \in \mathbb{N}}$ of $[0,1]$ with $|\mathcal{D}_M| \downarrow 0$. Write $\Phi_\varphi^{\mathcal{D}_M,N}(\cdot) = \left\langle v^N, S_1^{\mathcal{D}_M,N}(\cdot) \right\rangle$ for the ResNet initialized with the *time-homogeneous* scheme.

Theorem C.13. *Let the activation function* $\varphi : \mathbb{R} \to \mathbb{R}$ *be a linearly bounded, absolutely continuous map with exponentially bounded derivative. For any subset* $\mathcal{X} = \{x_1, \dots, x_n\} \subset \mathbb{X}$ *the following convergence in distribution holds*

$$\lim_{M \to \infty} \lim_{N \to \infty} \Phi_\varphi^{M,N}(\mathcal{X}) = \mathcal{N}(0, \mathcal{K}_\varphi(\mathcal{X}, \mathcal{X}))$$

where the map $\mathcal{K}_\varphi : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ *is given by the unique final values* $\mathcal{K}_\varphi^{x,y}(1,1)$ *of the following integro-differential equation*

$$\mathcal{K}_\varphi^{x,y}(s,t) = \sigma_a^2 + \int_{\eta=0}^s \int_{\tau=0}^t \left[\sigma_A^2 \mathbb{E}_{Z \sim \mathcal{N}(0, \Sigma_{\eta,\tau}^{x,y})} [\varphi(Z_1)\varphi(Z_2)] + \sigma_b^2 \right] \langle \dot{x}_\eta, \dot{y}_\tau \rangle_{\mathbb{R}^d} \, d\eta d\tau \qquad (25)$$

with

$$\Sigma_{s,t}^{x,y} = \begin{pmatrix} \mathcal{K}_\varphi^{x,x}(s,s), \mathcal{K}_\varphi^{x,y}(s,t) \\ \mathcal{K}_\varphi^{x,y}(s,t), \mathcal{K}_\varphi^{y,y}(t,t) \end{pmatrix}$$

Proof. The proof is now just a matter of combining Theorem C.2 and Theorem C.7.

Under our hypotheses Theorem C.2 tells us that, for any subset $\mathcal{X} = \{x_1, \dots, x_n\} \subset \mathbb{X}$, we have in distribution

$$\lim_{N \to \infty} \Phi_\varphi^{\mathcal{D}_M,N}(\mathcal{X}) = \mathcal{N}(0, \mathcal{K}_{\mathcal{D}_M \times \mathcal{D}_M}(\mathcal{X}, \mathcal{X}))$$

thus to conclude we just have to prove that, still in distribution, it holds

$$\lim_{M \to \infty} \mathcal{N}(0, \mathcal{K}_{\mathcal{D}_M \times \mathcal{D}_M}(\mathcal{X}, \mathcal{X})) = \mathcal{N}(0, \mathcal{K}_\varphi(\mathcal{X}, \mathcal{X}))$$

or, equivalently, that in $R^{n \times n}$ it holds $\lim_{M \to \infty} \mathcal{K}_{\mathcal{D}_M \times \mathcal{D}_M}(\mathcal{X}, \mathcal{X}) = \mathcal{K}_\varphi(\mathcal{X}, \mathcal{X})$.

This last needed limit follows from Theorem C.7 with the sequence $\{\mathcal{D}_M \times \mathcal{D}_M\}_{M \in \mathbb{N}}$. The uniqueness follows from Proposition C.12. $\qquad\square$

Like in the inhomogeneous case we can explicitly write the Kernel for the simplest case:

Corollary C.14. *If* $\sigma = id$, *then*

$$\mathcal{K}_{id}^{x,y}(s,t) = \left(\sigma_a^2 + \frac{\sigma_b^2}{\sigma_A^2} \right) k_{sig}^{\sigma_A x, \sigma_A y}(s,t) - \frac{\sigma_b^2}{\sigma_A^2} \qquad (26)$$

Proof. We readily see that

$$\mathcal{K}_{id}^{x,y}(s,t) = \sigma_a^2 + \int_{\eta=0}^s \int_{\tau=0}^t \left[\sigma_A^2 \mathcal{K}_{id}^{x,y}(\eta,\tau) + \sigma_b^2 \right] \langle \dot{x}_\eta, \dot{y}_\tau \rangle_{\mathbb{R}^d} \, d\eta d\tau \qquad (27)$$

By substituting (26) for $\mathcal{K}_{id}^{x,y}(\eta,\tau)$ in the integral and using (Salvi et al., 2021a)[Theorem 2.5] we have

$$\sigma_a^2 + \int_{\eta=0}^s \int_{\tau=0}^t \left[\sigma_A^2 \left\{ (\sigma_a^2 + \frac{\sigma_b^2}{\sigma_A^2}) k_{sig}^{\sigma_A x, \sigma_A y}(\eta,\sigma) - \frac{\sigma_b^2}{\sigma_A^2} \right\} + \sigma_b^2 \right] \langle \dot{x}_\eta, \dot{y}_\tau \rangle_{\mathbb{R}^d} \, d\eta d\tau$$

$$= \sigma_a^2 + \int_{\eta=0}^s \int_{\tau=0}^t \sigma_A^2 (\sigma_a^2 + \frac{\sigma_b^2}{\sigma_A^2}) k_{sig}^{\sigma_A x, \sigma_A y}(\eta,\sigma) \langle \dot{x}_\eta, \dot{y}_\tau \rangle_{\mathbb{R}^d} \, d\eta d\tau$$

$$= \sigma_a^2 + (\sigma_a^2 + \frac{\sigma_b^2}{\sigma_A^2}) \int_{\eta=0}^s \int_{\tau=0}^t k_{sig}^{\sigma_A x, \sigma_A y}(\eta,\sigma) \langle \sigma_A \dot{x}_\eta, \sigma_A \dot{y}_\tau \rangle_{\mathbb{R}^d} \, d\eta d\tau$$

$$= \sigma_a^2 + (\sigma_a^2 + \frac{\sigma_b^2}{\sigma_A^2})(k_{sig}^{\sigma_A x, \sigma_A y}(s,t) - 1)$$

$$= (\sigma_a^2 + \frac{\sigma_b^2}{\sigma_A^2}) k_{sig}^{\sigma_A x, \sigma_A y}(s,t) - \frac{\sigma_b^2}{\sigma_A^2}$$

which means, by uniqueness of solutions (Proposition C.12), that the thesis holds. $\qquad\square$

Remark. A proof similar to that given in the inhomogeneous case for $\varphi = ReLU$ does not work now since the covariance matrix is not, in general, degenerate for $x = y$ as in that case.

Lemma C.15. *For all choices $(\sigma_a, \sigma_A, \sigma_b)$ and for all φ as in Theorem 3.1 we have, with abuse of notation and the obvious meaning, that*

$$\mathcal{K}_\varphi^{x,y}(s,t;\sigma_a, \sigma_A, \sigma_b) = \mathcal{K}_\varphi^{\sigma_A x, \sigma_A y}(s,t;\sigma_a, 1, \sigma_b \sigma_A^{-1})$$

Proof. Follow the exact same steps and arguments of B.12. $\qquad\square$

C.2. The infinite-depth-then-width regime

C.2.1. THE INFINITE-DEPTH LIMIT WITH FINITE-WIDTH

Proposition C.16. *Let $\{\mathcal{D}_M\}_{M \in \mathbb{N}}$ be a sequence of partitions of $[0,1]$ such that $|\mathcal{D}_M| \to 0$ as $M \to \infty$. Assume the activation function φ is Lipschitz and linearly bounded. Let $x \in \mathbb{X}$ and let $\rho_M(t) := \sup\{s \in \mathcal{D}_M : s \leq t\}$. Then, the the $\mathbb{R}^N$-valued process $t \mapsto S_{\rho_M(t)}^{M,N}(x)$ converges in distribution, as $M \to \infty$, to the solution $S^N(x)$ of the following Neural CDE*

$$S_t^N(x) = a + \int_0^t \sum_{j=1}^d \left(A_j \varphi(S_s^N(x)) + b_j \right) dx_s^j \tag{28}$$

where $A_j \in \mathbb{R}^{N \times N}$ and $b_j \in \mathbb{R}^N$ are sampled according in the definition of the homogeneous controlled ResNet.

Proof. Assume the A_k and b_k to be fixed for all choices of $\mathcal{D}_M$. The system has unique solution by (Friz & Victoir, 2010)[Theorems 3.7, 3.8] with

$$V_i(x) = A_i \varphi(x) + b_k$$

noting that these are Lipschitz since composition of Lipschitz and linearly bounded.

By reasoning as in the proof of Theorem 4.1 one gets that

$$\left\| S^{M,N}(x) - S^N(x) \right\|_{\infty,[0,1]} \leq \sqrt{|\mathcal{D}_M|} \exp \left\{ K_{\|x\|_\mathbb{X}} (\|a\|_{\mathbb{R}^N} + \sum_{k=1}^d \|A_k\|_F + \|b_k\|_F) \right\}$$

for some constant $K_{\|x\|_\mathbb{X}}$ depending in an increasing fashion on the norm of the input path.

Taking the expectation over A_k, b_k and the initial condition leads to

$$\mathbb{E}[\left\| S^{M,N}(x) - S^N(x) \right\|_{\infty,[0,1]}] \leq \tilde{K}_{\|x\|_\mathbb{X}} \sqrt{|\mathcal{D}_M|}$$

for some other constant $\tilde{K}_{\|x\|_\mathbb{X}}$ since the matrices and vectors are all distributed as Gaussians. One then concludes by portmanteau lemma considering Lipschitz functions on $C^0([0,1];\mathbb{R}^N)$.

$\qquad\square$

Remark. Since the bounds $\tilde{K}_{\|x\|_{\mathbb{X}}}$ depend increasingly on the norm of the input path they are uniform on bounded sets of $\mathbb{X}$ and this result can be extended to arbitrary finite sets of paths $\mathcal{X} \subseteq \mathbb{X}$ with exactly the same portmanteau arguments.

Definition C.17 (Randomized Signatures). We call *Randomized Signatures* the solutions to

$$S_t^N(x) = a + \sum_{k=1}^{d} \int_0^t \left(A_k[\varphi(S_\tau^N(x))] + b_k \right) dx_\tau^k$$

These are the same objects defined in (Cuchiero et al., 2021b).

C.2.2. INFINITE-DEPTH-THEN-WIDTH LIMIT: $\varphi = id$

Theorem C.18. *In the case $\varphi = id$*

$$\frac{1}{N} \left\langle S_s^N(\mathcal{X}), S_t^N(\mathcal{X}) \right\rangle_{\mathbb{R}^N} \xrightarrow[N\to\infty]{\mathbb{L}^2} \mathcal{K}_{id}^{\mathcal{X},\mathcal{X}}(s,t)$$

moreover the convergence is of order $\mathcal{O}(\frac{1}{N})$

Proof. This is proved in Appendix C.4. □

C.3. Commutativity of Limits

In this section we are going to prove the commutativity of limits in the homogeneous case. The core arguments are the same as those employed for the inhomogeneous counterpart, this time however we won't be able to take advantage of *ready-made* results from stochastic analysis, thus we are going to carefully obtain bounds in more direct ways.

Let $\{\mathcal{D}_M\}_{M\in\mathbb{N}}$ be a sequence of partitions of $[0,1]$ such that $|\mathcal{D}_M| \to 0$ as $M \to \infty$. Fix N and the matrices S_0, A_k, b_k for all partitions. For any $x \in \mathbb{X}$ let $S^{M,N}(x) : \mathcal{D}_M \to \mathbb{R}^N$ be the homogeneous cResNet corresponding to the above quantities.

Definition C.19. Given $x \in \mathbb{X}$ and a partition $\mathcal{D}_M$ define

$$\rho_M(t) := \sup\{s \in \mathcal{D}_M : s \leq t\}$$

Then define the piecewise constant extension $Z^{M,N}(x) : [0,1] \to \mathbb{R}^N$

$$Z_t^{M,N}(x) := S_{\rho_M(t)}^{M,N}(x) \tag{29}$$

and the integral extension $S^{M,N}(x) : [0,1] \to \mathbb{R}^N$

$$S^{M,N}(x) = S_0^N + \sum_{k=1}^{d} \int_0^t \left(A_k \varphi(Z_s^{M,N}(x)) + b_k \right) dx_s^k \tag{30}$$

Note how the two extensions coincide on $\mathcal{D}_M$.

Proposition C.20. *Assume the activation function φ is Lipschitz and linearly bounded. There is a constant $K_x > 0$ independent of N, M and increasing in $\|x\|_{1-var}$ such that*

$$\mathbb{E}\left[\left\| Z_t^{N,M}(x) \right\|_{\infty,[0,1]}^2 \right] \leq NK_x$$

where the expectation is taken over the joint distribution of $S_0, \{A_k, b_k\}_{k=1,\dots,d}$

Proof. Let $t \in [t_m, t_{m+1})$ then

$$|Z_t^{N,M}(x)| = |S_{t_m}^{N,M}(x)| \leq$$

$$|S_0^N| + \sum_{i=1}^d |\int_0^t A_i^N \varphi(Z_r^{N,M}(x)) + b_i^N dx_r^i| \leq$$

$$|S_0^N| + \sum_{i=1}^d \int_0^t |A_i^N \varphi(Z_r^{N,M}(x)) + b_i^N||dx_r^i| \leq$$

$$|S_0^N| + \sum_{i=1}^d \int_0^t \left(\left\| A_i^N \right\|_{op} |\varphi(Z_r^{N,M}(x))| + |b_i^N| \right)|dx_r^i| \leq$$

$$|S_0^N| + \sum_{i=1}^d \int_0^t \left(\left\| A_i^N \right\|_{op} \sqrt{2}C(\sqrt{N} + \left\| Z_r^{N,M}(x) \right\|) + |b_i^N| \right)|dx_r^i| \leq$$

$$\left[|S_0^N| + (\sum_{i=1}^d \sqrt{2N}C \left\| A_i^N \right\|_{op} + |b_i^N|) \left\| x \right\|_{1-var,[0,1]} \right] + \sqrt{2}C(\sum_{i=1}^d \left\| A_i^N \right\|_{op}) \sum_{i=1}^d \int_0^t |Z_r^{N,M}(x)||dx_r^i|$$

thus by *Lemma 3.2* of (Friz & Victoir, 2010) we obtain

$$|Z_t^{N,M}(x)| \leq \left[|S_0^N| + (\sum_{i=1}^d \sqrt{2N}C \left\| A_i^N \right\|_{op} + |b_i^N|) \left\| x \right\|_{1-var,[0,1]} \right] \exp \left\{ \sqrt{2}C(\sum_{i=1}^d \left\| A_i^N \right\|_{op}) \left\| x \right\|_{1-var,[0,t]} \right\}$$

$$\leq \left[|S_0^N| + (\sqrt{2N}C + \sum_{i=1}^d |b_i^N|) \left\| x \right\|_{1-var,[0,1]} \right] \exp \left\{ (\sum_{i=1}^d \left\| A_i^N \right\|_{op})(1 + \sqrt{2}C \left\| x \right\|_{1-var,[0,t]}) \right\}$$

$$\leq \left[|S_0^N| + (\sqrt{2N}C + \sum_{i=1}^d |b_i^N|) \left\| x \right\|_{1-var,[0,1]} \right] \prod_{i=1}^d \exp \left\{ (1 + \sqrt{2}C \left\| x \right\|_{1-var,[0,1]}) \left\| A_i^N \right\|_{op} \right\}$$

Hence using independence

$$\mathbb{E}[\sup_{t \in [0,1]} |Z_t^{N,M}(x)|^2] \leq$$

$$2\mathbb{E}\left[|S_0^N|^2 + (2NC^2 + \sum_{i=1}^d |b_i^N|^2) \left\| x \right\|_{1-var,[0,1]}^2 \right] \prod_{i=1}^d \mathbb{E}\left[\exp \left\{ 2(1 + \sqrt{2}C \left\| x \right\|_{1-var,[0,1]}) \left\| A_i^N \right\|_{op} \right\} \right]$$

Note that $\left\| A_i^N \right\|_{op} = \sqrt{\rho((A_i^N)^T A_i^N)}$ is the square root of the biggest eigenvalue l^N of $(A_i^N)^T A_i^N$. The distribution of eigenvalues $\frac{1}{\sigma_A^2}(A_i^N)^T A_i^N$ is well understood as converging to the Marchenko–Pastur distribution (a classical distribution supported on the interval $[0, 4]$), and l^N converges almost surely to $4\sigma_A^2$ (Geman, 1980; Johnstone, 2001) with fluctuations around this limit having exponential tails (Majumdar & Vergassola, 2009; Johansson, 2000). It follows that $\mathbb{E}[\exp\{\lambda \left\| A_i^N \right\|_{op}\}]$ is uniformly bounded in N, with the bound depending on σ_A and λ.

Moreover by Hölder

$$\mathbb{E}[|S_0^N|^2] = \mathbb{E}[\sum_{i=1}^N |[S_0^N]_i|^2] = \sigma_a^2 N$$

and similarly $\mathbb{E}[|b_i^N|^2] \leq \sigma_{b_i}^2 N$.

Hence

$$\mathbb{E}[\sup_{t \in [0,1]} |Z_t^{N,M}(x)|^2] \leq$$

$$2N \left(\sigma_a^2 + (2C^2 + \sum_{i=1}^d \sigma_{b_i}^2) \left\| x \right\|_{1-var,[0,1]}^2 \right) \mathbb{E}\left[\exp \left\{ 2(1 + \sqrt{2}C \left\| x \right\|_{1-var,[0,1]}) \left\| A_1^N \right\|_{op} \right\} \right]^d$$

$$=: NK_x$$

□

Proposition C.21. *Assume the activation function φ is Lipschitz and linearly bounded.*

$$\mathbb{E}\left[\left\|S_t^{N,M}(x) - Z_t^{N,M}(x)\right\|_{\infty,[0,1]}^2\right] \leq N|\mathcal{D}_M|\tilde{K}_x$$

where $\tilde{K}_x$ is a universal constant depending only on $\|x\|_{\mathbb{X}}$ in an increasing manner.

Proof. Let $t \in [t_m, t_{m+1})$ then, using the bound in Remark A.2,

$$|S_t^{N,M}(x) - Z_t^{N,M}(x)| = |\sum_{i=1}^d \int_{t_m}^t \left(A_i^N \varphi(S_{t_m}^{N,M}(x)) + b_i^N\right)dx_r^i| \leq$$

$$\sum_{i=1}^d \int_{t_m}^t \left(\|A_i^N\|_{op}|\varphi(S_{t_m}^{N,M}(x))| + |b_i^N|\right)|dx_r^i| \leq$$

$$\sum_{i=1}^d \int_{t_m}^t \left(\|A_i^N\|_{op}\sqrt{2}C(\sqrt{N} + |S_{t_m}^{N,M}(x)|) + |b_i^N|\right)|dx_r^i| \leq$$

$$[\sum_{i=1}^d \|A_i^N\|_{op}\sqrt{2}C(\sqrt{N} + |Z_{t_m}^{N,M}(x)|) + |b_i^N|] \|x\|_{1-var,[t_m,t]} \leq$$

$$[\sum_{i=1}^d \|A_i^N\|_{op}\sqrt{2}C(\sqrt{N} + |Z_{t_m}^{N,M}(x)|) + |b_i^N|] \|x\|_{\mathbb{X}} \sqrt{|\mathcal{D}_M|}$$

where we have used

$$\|x\|_{1-var,[s,t]} = \int_s^t |\dot{x}_r|dr \leq \sqrt{\int_0^1 |\dot{x}_r||t-s|dr} \leq \|x\|_{\mathbb{X}}\sqrt{t-s}$$

thus

$$\mathbb{E}\left[\sup_{t\in[0,1]}|S_t^{N,M}(x) - Z_t^{N,M}(x)|^2\right] \leq$$

$$2\mathbb{E}\left[\sup_{t_m\in\mathcal{D}_M}[\sum_{i=1}^d \|A_i^N\|_{op}^2 2C^2(N + |Z_{t_m}^{N,M}(x)|^2) + |b_i^N|^2] \|x\|_{\mathbb{X}}^2 |\mathcal{D}_M|\right]$$

$$\leq N\bar{K}_x \|x\|_{\mathbb{X}}^2 |\mathcal{D}_M|$$

□

Proposition C.22. *Assume the activation function φ is Lipschitz and linearly bounded.*

$$\mathbb{E}\left[\left\|S_t^{N,M}(x) - S_t^{N,M'}(x)\right\|_{\infty,[0,1]}^2\right] \leq N\tilde{K}_x(|D_M| \vee |\mathcal{D}_{M'}|)$$

where $\tilde{K}_x$ is a universal constant depending only on $\|x\|_{\mathbb{X}}$ in an increasing manner.

Proof. Let $t \in [t_m, t_{m+1})$ then

$$|S_t^{N,M}(x) - S_t^{N,M'}(x)| \le |\sum_{i=1}^d \int_0^t |A_i^N(\varphi(Z_t^{N,M}(x)) - \varphi(Z_t^{N,M'}(x)))||dx_r^i| \le$$

$$\sum_{i=1}^d \int_0^t \left(\|A_i^N\|_{op} |\varphi(Z_t^{N,M}(x)) - \varphi(Z_t^{N,M'}(x))| \right)|dx_r^i| \le$$

$$\sum_{i=1}^d \int_0^t \|A_i^N\|_{op} K|Z_t^{N,M}(x) - Z_t^{N,M'}(x)||dx_r^i| =$$

$$K \sum_{i=1}^d \|A_i^N\|_{op} \int_0^t |Z_t^{N,M}(x) - Z_t^{N,M'}(x)||dx_r^i|$$

Using triangle inequality we obtain

$$|S_t^{N,M}(x) - S_t^{N,M'}(x)| \le$$

$$K \sum_{i=1}^d \|A_i^N\|_{op} \int_0^t |Z_t^{N,M}(x) - S_t^{N,M}(x)||dx_r^i| +$$

$$K \sum_{i=1}^d \|A_i^N\|_{op} \int_0^t |S_t^{N,M}(x) - S_t^{N,M'}(x)||dx_r^i| +$$

$$K \sum_{i=1}^d \|A_i^N\|_{op} \int_0^t |S_t^{N,M'}(x) - Z_t^{N,M'}(x)||dx_r^i|$$

Note how

$$K \sum_{i=1}^d \|A_i^N\|_{op} \int_0^t |Z_t^{N,M}(x) - S_t^{N,M}(x)||dx_r^i| \le$$

$$K \sum_{i=1}^d \|A_i^N\|_{op} \int_0^t \left[\sum_{k=1}^d \|A_k^N\|_{op} \sqrt{2}C(\sqrt{N} + |Z_{t_m}^{N,M}(x)|) + |b_k^N| \right] \|x\|_{\mathbb{X}} \sqrt{|\mathcal{D}_M|}|dx_r^i| \le$$

$$K(\sum_{i=1}^d \|A_i^N\|_{op})\left[\sum_{i=k}^d \|A_k^N\|_{op} \sqrt{2}C(\sqrt{N} + \Gamma_x) + |b_k^N| \right] \|x\|_{1-var} \|x\|_{\mathbb{X}} \sqrt{|\mathcal{D}_M|} =: \Lambda_x \sqrt{|\mathcal{D}_M|}$$

with

$$\Gamma_x = \left[|S_0^N| + (\sqrt{2N}C + \sum_{i=1}^d |b_i^N|) \|x\|_{1-var,[0,1]} \right] \prod_{i=1}^d \exp\left\{ (1 + \sqrt{2}C \|x\|_{1-var,[0,1]}) \|A_i^N\|_{op} \right\}$$

Hence it follows that

$$|S_t^{N,M}(x) - S_t^{N,M'}(x)| \le$$

$$(\sqrt{|\mathcal{D}_M|} + \sqrt{|\mathcal{D}_{M'}|})\Lambda_x + K \sum_{i=1}^d \|A_i^N\|_{op} \int_0^t |S_t^{N,M}(x) - S_t^{N,M'}(x)||dx_r^i|$$

and with Gronwall

$$\left\| S_t^{N,M}(x) - S_t^{N,M'}(x) \right\|_\infty \le 2\sqrt{|\mathcal{D}_M| \vee |\mathcal{D}_{M'}|}\Lambda_x \exp\{K\left(\sum_{i=1}^d \|A_i^N\|_{op} \right) \|x\|_{1-var}\}$$

Taking expectations of the squares and proceeding as before we have the thesis. □

We can finally prove the main bound which will allow the exchange of limits:

Theorem C.23. *Let $\{\mathcal{D}_M\}_{M\in\mathbb{N}}$ be a sequence of partitions of $[0,1]$ such that $|\mathcal{D}_M| \to 0$ as $M \to \infty$. Assume the activation function φ is Lipschitz and linearly bounded. Fix the matrices S_0, A_k, b_k for all partitions. Let $S^N(x)$ be the solution of the following Neural CDE*

$$S_t^N(x) = S_0^N + \sum_{k=1}^{d} \int_0^t \left(A_k \varphi(S_s^N(x)) + b_k\right) dx_s^k \tag{31}$$

Then there exist a constant $\mathbf{K}_x$ independent of N, M and increasing function of $\|x\|_{\mathbb{X}}$ such that

$$\mathbb{E}[\sup_{t\in[0,1]} \left\|S_t^{M,N}(x) - S_t^N(x)\right\|_{\mathbb{R}^N}^2] \le N|D_M|\mathbf{K}_x \tag{32}$$

Proof. Note first that $S^N(x)$ is well defined since the system has unique solution by (Friz & Victoir, 2010)[Theorems 3.7, 3.8] with

$$V_k(x) = A_k \varphi(x) + b_k$$

noting that these are Lipschitz since composition of Lipschitz and linearly bounded.

Moreover the proof tells us that $S_t^{N,M}(x)$ is a cauchy sequence in $C^0([0,1];\mathbb{R}^N)$ hence for any $M > 0$, eventually in M' it holds

$$\sup_{t\in[0,1]} \left\|S_t^{M,N}(x) - S_t^N(x)\right\|_{\mathbb{R}^N} \le 2\sup_{t\in[0,1]} \left\|S_t^{M,N}(x) - S_t^{M',N}(x)\right\|_{\mathbb{R}^N}$$

Proposition (C.22) then gives the sought after bound.

$\square$

Theorem C.24. *Assume the activation function φ is Lipschitz and linearly bounded. There is a constant C depending only on $\|x\|_{\mathbb{X}}$ in an increasing fashion such that:*

$$\sup_{N\ge 1}\sup_{t\in[0,1]} \mathcal{W}_1(\mu_t^{M,N}(x), \mu_t^N(x)) \le \bar{\mathbf{K}}_x\sqrt{|\mathcal{D}_M|}$$

where $\mu_t^{M,N}$ is the distribution of $\left\langle v^N, S_t^{M,N}(x)\right\rangle$ and $\mu_t^N(x)$ that of $\left\langle v^N, S_t^N(x)\right\rangle$ for some independent vector with iid entries $[v^N]_\alpha \sim \mathcal{N}(0, \frac{1}{N})$.

Proof. Let $G : \mathbb{R} \to \mathbb{R}$ be 1-Lipschitz, we have

$$\mathbb{E}[|G(\left\langle v^N, S_t^{M,N}(x)\right\rangle) - G(\left\langle v^N, S_t^N(x)\right\rangle)|] \le \mathbb{E}\left[|\left\langle v^N, S_t^{M,N}(x)\right\rangle - \left\langle v^N, S_t^N(x)\right\rangle|\right]$$

which leads to

$$\begin{aligned}
\mathbb{E}\left[|\left\langle v^N, S_t^{M,N}(x)\right\rangle - \left\langle v^N, S_t^N(x)\right\rangle|\right] &= \mathbb{E}\left[|\left\langle v^N, S_t^{M,N}(x) - S_t^N(x)\right\rangle|\right] \\
&\le \left(\mathbb{E}\left[|\left\langle v^N, S_t^{M,N}(x) - S_t^N(x)\right\rangle|^2\right]\right)^{\frac{1}{2}} \\
&= \left(\frac{1}{N}\mathbb{E}\left[\left\|S_t^{M,N}(x) - S_t^N(x)\right\|^2\right]\right)^{\frac{1}{2}} \\
&\le \left(\frac{1}{N}\mathbb{E}\left[\sup_{t\in[0,1]}\left\|S_t^N(x) - S_t^{M,N}(x)\right\|^2\right]\right)^{\frac{1}{2}} \\
&\le \sqrt{\mathbf{K}_x|\mathcal{D}_M|}
\end{aligned}$$

hence

$$\mathcal{W}_1(\mu_t^{M,N}(x), \mu_t^N(x)) \le \sqrt{\mathbf{K}_x|\mathcal{D}_M|}$$

$\square$

Remark. Note that the exact arguments, being of L^2 type, can be repeated for the "stacked" vector $(S^{N,M}(x_1), \ldots, S^{N,M}(x_N))$ extending (qualitatively) the bounds to the multi-input case.

Corollary C.25 (Thm. 4.1)**.** *The limits in Thm. C.7 commute.*

Proof. By the classical Moore-Osgood theorem we need to prove that one of the two limits is uniform in the other, for example that the limit in distribution as $M \to \infty$ is uniform in N in some metric which describes convergence in distribution. But this is just the content of the previous result, extended to the multi-input case. $\qquad\square$

C.4. An alternative proof for the case $\varphi = id$

In this last section we prove Theorem C.18. between *Randomized Signature Kernels* and the original *Signature Kernel*.

We are going to consider words $I = (i_1, \cdots, i_k) \in \mathcal{A}_d^k$ in the alphabet $\mathcal{A}_d := \{1, \ldots, d\}$; the space of all words will be denoted by $\mathbb{W}_d$, the length of a word by $|I| = k$ and the sum of its elements by $\|I\| = \sum_{m=1}^{k} i_m$.

Recall we consider randomized Signatures

$$S_t^N(x) = S_0 + \sum_{k=1}^{d} \int_0^t \left(A_k S_\tau^N(x) + b_k \right) dx_\tau^k$$

where $x \in \mathbb{X}$ and

$$[A_k]_\alpha^\beta \sim \mathcal{N}(0, \tfrac{\sigma_A^2}{N}) \qquad [S_0]_\alpha \sim \mathcal{N}(0, \sigma_{S_0}^2) \qquad [b_k]_\alpha \sim \mathcal{N}(0, \sigma_b^2)$$

Our goal is that of proving the following result:

Theorem C.26. *In the assumptions stated above*

$$\lim_{N \to \infty} \mathbb{E}\left[\frac{1}{N} \langle S_s^N(x), S_t^N(y) \rangle_{\mathbb{R}^N} \right] = \left(\sigma_{S_0}^2 + \frac{\sigma_b^2}{\sigma_A^2} \right) k_{sig}^{\sigma_A x, \sigma_A y}(s, t) - \frac{\sigma_b^2}{\sigma_A^2}$$

and the variance around the limit is of order $O(\frac{1}{N})$.

We know, see (Baudoin & Zhang, 2012)[Remark 2.10], that it is possible to write a closed form for $S_t^N(x)$ which *decouples* the effects of the vector fields and those of the driving control using the *Signature*:

$$S_t^N(x) = \sum_{I \in \mathbb{W}_d} V_I(S_0^N) \, Sig_{0,s}^I(x) \tag{33}$$

where $\mathbb{W}_d$ is the set of words in the alphabet $\{1, \ldots, d\}$ and if $I = (i_1, \ldots, i_k)$ then $[V_I(z)]_\alpha := V_{i_1} \cdots V_{i_k} \langle e_\alpha, \cdot \rangle(z)$ with $V_i(z) := A_i z + b_i$ and $V f(z) =:= df_z[V_j(z)]$ for $f \in C^\infty(\mathbb{R}^N; \mathbb{R})$. Notice that $V_I(S_0^N) \in \mathbb{R}^N$ and $Sig_{0,s}^I(x) \in \mathbb{R}$.

In fact the "Taylor expansion" with respect of the signature of $f(S_t^N(x))$ for $f \in C^\infty(\mathbb{R}^N; \mathbb{R})$ is

$$f(S_t^N(x)) = \sum_{I \in \mathbb{W}_d} V_I f(S_0^N) \, Sig_{0,s}^I(x)$$

where, with I as above, $V_I f(x) := V_{i_1}(V_{i_2} \cdots (V_{i_k} f) \cdots)(x)$ with

$$V_j f(x) := df_x[V_j(x)]$$

From Equation 33 we also get

$$\langle S_s^N(x), S_t^N(y) \rangle_{\mathbb{R}^N} = \sum_{I \in \mathbb{W}_d} \sum_{J \in \mathbb{W}_d} \langle V_I f(S_0^N), V_J f(S_0^N) \rangle_{\mathbb{R}^N} \, Sig_{0,s}^I(x) Sig_{0,t}^J(y)$$

where y is another control. If we could exchange expectation with the series we would thus get

$$\mathbb{E}[\langle S_s^N(x), S_t^N(y)\rangle_{\mathbb{R}^N}] = \sum_{I\in\mathbb{W}_d}\sum_{J\in\mathbb{W}_d}\mathbb{E}[\langle V_I f(S_0^N), V_J f(S_0^N)\rangle_{\mathbb{R}^N}]\, Sig_{0,s}^I(x)Sig_{0,t}^J(y)$$

Thus we have to study the expectations $\mathbb{E}[\langle V_I f(S_0^N), V_J f(S_0^N)\rangle_{\mathbb{R}^N}]$.

Note that if $I = ()$, the empty word, then $V_I f(S_0^N) = S_0^N$.

C.4.1. PRODUCTS OF GAUSSIAN MATRICES

The most important result which will make our plan succeed is the following classical theorem:

Theorem C.27 (Isserlis). *Let $(X_1, \ldots, X_N)$ be a zero mean multivariate normal vector, then*

$$\mathbb{E}[X_1 \cdots X_N] = \sum_{p\in P_n^2}\prod_{\{i,j\}\in p}\mathbb{E}[X_i X_j]$$

where the sum is over all distinct ways of partitioning $\{1, \ldots, N\}$ into pairs $\{i, j\}$, and the product is over the pairs contained in p.

Proposition C.28. *Assume $I, J \in \mathbb{W}_d$ such that $|I| + |J| > 0$. In the hypotheses stated above*

$$\frac{1}{N}\mathbb{E}[\langle V_I f(S_0^N), V_J f(S_0^N)\rangle_{\mathbb{R}^N}] = (\sigma_{S_0}^2 + \frac{\sigma_b^2}{\sigma_A^2})\sigma_A^{|I|+|J|}[\delta_I^J + O(\frac{1}{N})]$$

Proof. Let us first consider the case $\sigma_b = 0$.

Taking $f(z) = \langle v, z\rangle_{\mathbb{R}^N}$ for some $v \in \mathbb{R}^N$ and $V_j(z) := A_j z$ we get

$$V_j f(x) := d(\langle v, \cdot\rangle_{\mathbb{R}^N})_x[A_j x] = v^T A_j x = \langle A_j^T v, x\rangle_{\mathbb{R}^N}$$

hence by induction

$$V_I f(x) = \langle A_{i_1}^T \cdots A_{i_k}^T v, x\rangle_{\mathbb{R}^N} = \langle v, A_{i_k} \cdots A_{i_1} x\rangle_{\mathbb{R}^N}$$

and in our notation

$$V_I(x) = A_I x = A_{i_k} \cdots A_{i_1} x$$

We are thus interested in the quantities $\mathbb{E}[(S_0^N)^T A_I^T A_J S_0^N]$.

Let $\alpha, \beta \in \{1, \ldots, N\}$ and given a matrix $M \in \mathbb{R}^{N\times N}$ write $[M]_\alpha^\beta$ for its component in row α and column β. Remember how

$$[M_1 M_2]_\alpha^\beta = \sum_{\gamma=1}^N [M_1]_\alpha^\gamma [M_2]_\gamma^\beta$$

Thus we have

$$\mathbb{E}\left[(S_0^N)^T A_I^T A_J S_0^N\right] = \sum_{n=1}^N\sum_{m=1}^N \mathbb{E}\left[[S_0^N]_n[S_0^N]_m[A_I^T A_J]_n^m\right]$$

$$= \sum_{n=1}^N\sum_{m=1}^N \mathbb{E}\left[[S_0^N]_n[S_0^N]_m\right]\mathbb{E}\left[[A_I^T A_J]_n^m\right]$$

Moreover, for $I = (i_1, \ldots, i_k)$, we obtain

$$
\begin{aligned}
[A_I]_\alpha^\beta &= \sum_{\delta_1=1}^N [A_{(i_2,\ldots,i_k)}]_\alpha^{\delta_1} [A_{i_1}]_{\delta_1}^\beta \\
&= \sum_{\delta_{k-1}=1}^N \cdots \sum_{\delta_1=1}^N [A_{i_k}]_\alpha^{\delta_{k-1}} [A_{i_{l-1}}]_{\delta_{k-1}}^{\delta_{k-2}} \cdots [A_{i_2}]_{\delta_2}^{\delta_1} [A_{i_1}]_{\delta_1}^\beta \\
&= \sum_{\bar\delta \in \Lambda_{\alpha,\beta}^{N,k}} \prod_{n=1}^k [A_{i_n}]_{\delta_n}^{\delta_{n-1}}
\end{aligned}
$$

where $\Lambda_{\alpha,\beta}^{N,k} := \{(\delta_0,\ldots,\delta_k) \in \{1,\ldots,N\}^{k+1} : \delta_k = \alpha \text{ and } \delta_0 = \beta\}$. With this notation we can write

$$
\begin{aligned}
[A_I^T A_J]_\alpha^\beta &= \sum_{\gamma=1}^N [A_I^T]_\alpha^\gamma [A_J]_\gamma^\beta = \sum_{\gamma=1}^N [A_I]_\gamma^\alpha [A_J]_\gamma^\beta \\
&= \sum_{\gamma=1}^N \sum_{\bar\delta \in \Lambda_{\gamma,\alpha}^{N,|I|}} \sum_{\bar\epsilon \in \Lambda_{\gamma,\beta}^{N,|J|}} \prod_{n=1}^{|I|} \prod_{m=1}^{|J|} [A_{i_n}]_{\delta_n}^{\delta_{n-1}} [A_{j_m}]_{\epsilon_m}^{\epsilon_{m-1}}
\end{aligned}
$$

thus

$$
\mathbb{E}\Big[[A_I^T A_J]_\alpha^\beta \Big] = \sum_{\gamma=1}^N \sum_{\bar\delta \in \Lambda_{\gamma,\alpha}^{N,|I|}} \sum_{\bar\epsilon \in \Lambda_{\gamma,\beta}^{N,|J|}} \mathbb{E}\Big[\prod_{n=1}^{|I|} \prod_{m=1}^{|J|} [A_{i_n}]_{\delta_n}^{\delta_{n-1}} [A_{j_m}]_{\epsilon_m}^{\epsilon_{m-1}} \Big] \tag{34}
$$

The time is ripe for the application of *Isserlis' Theorem*.

First of all notice how the sum in Isserlis runs over the possible pairings of the index set which in our case is the set of elements of the concatenation

$$
I * J = (i * j_1, \ldots, i * j_{|I|+|J|}) = (i_1, \ldots, i_{|I|}, j_1, \ldots, j_{|J|})
$$

of I and J. Then

$$
\begin{aligned}
&\mathbb{E}\Big[[A_I^T A_J]_\alpha^\beta \Big] \\
&= \sum_{\gamma=1}^N \sum_{\bar\delta \in \Lambda_{\gamma,\alpha}^{N,|I|}} \sum_{\bar\epsilon \in \Lambda_{\gamma,\beta}^{N,|J|}} \sum_{p \in P_{|I|+|J|}^2} \prod_{\{a,b\} \in p} \mathbb{E}\Big[[A_{(i*j)_a}]_{(\delta*\epsilon)_a}^{(\delta*\epsilon)_a'} [A_{(i*j)_b}]_{(\delta*\epsilon)_b}^{(\delta*\epsilon)_b'} \Big] \\
&= \sum_{p \in P_{|I|+|J|}^2} \sum_{\gamma=1}^N \sum_{\bar\delta \in \Lambda_{\gamma,\alpha}^{N,|I|}} \sum_{\bar\epsilon \in \Lambda_{\gamma,\beta}^{N,|J|}} \prod_{\{a,b\} \in p} \mathbb{E}\Big[[A_{(i*j)_a}]_{(\delta*\epsilon)_a}^{(\delta*\epsilon)_a'} [A_{(i*j)_b}]_{(\delta*\epsilon)_b}^{(\delta*\epsilon)_b'} \Big]
\end{aligned}
$$

where

$$
(\delta * \epsilon) = (\delta_1, \ldots, \delta_{|I|}, \epsilon_1, \ldots, \epsilon_{|J|})
$$

and

$$
(\delta * \epsilon)' = (\delta_0, \ldots, \delta_{|I|-1}, \epsilon_0, \ldots, \epsilon_{|J|-1})
$$

In particular this sum is 0 by default if $I * J$ has an odd number of elements. This means that

$$\mathbb{E}\left[[A_I^T A_J]_\alpha^\beta\right] = 0$$

whenever $|I| + |J|$ is odd.

Moreover, even if $|I| + |J|$ is even, the pairings must be in such a way that no factor of the product vanishes; since we are working with matrices with independent normal entries this is equivalent to requiring for each $\{a, b\} \in p \in P_{|I|+|J|}^2$ that $(i * j)_a = (i * j)_b$, $(\delta * \epsilon)_a = (\delta * \epsilon)_b$, and $(\delta * \epsilon)'_a = (\delta * \epsilon)'_b$.

Evidently then $I * J$ must be at least "pairable" with pairs of identical indices! In this case we will say that $I * J$ is *twinable* and we will get a factor $\sigma_A^2 N^{-1}$ out of every $\mathbb{E}\left[[A_{(i*j)_a}]_{(\delta*\epsilon)_a}^{(\delta*\epsilon)'_a}[A_{(i*j)_b}]_{(\delta*\epsilon)_b}^{(\delta*\epsilon)'_b}\right]$ thus

$$\mathbb{E}\left[[A_I^T A_J]_\alpha^\beta\right] = \left(\frac{\sigma_A^2}{N}\right)^{\frac{|I|+|J|}{2}} \Xi(I, J, N, \alpha, \beta)$$

where

$$\Xi(I, J, N, \alpha, \beta)$$
$$= |\{(p, \gamma, \bar{\delta}, \bar{\epsilon}) : \forall \{a, b\} \in p. \ \mathbb{E}\left[[A_{(i*j)_a}]_{(\delta*\epsilon)_a}^{(\delta*\epsilon)'_a}[A_{(i*j)_b}]_{(\delta*\epsilon)_b}^{(\delta*\epsilon)'_b}\right] \neq 0\}|$$

Define, given $p \in P_{|I|+|J|}^2$,

$$\omega(p, I, J, N, \alpha, \beta)$$
$$= |\{(\gamma, \bar{\delta}, \bar{\epsilon}) : \forall \{a, b\} \in p. \ \mathbb{E}\left[[A_{(i*j)_a}]_{(\delta*\epsilon)_a}^{(\delta*\epsilon)'_a}[A_{(i*j)_b}]_{(\delta*\epsilon)_b}^{(\delta*\epsilon)'_b}\right] \neq 0\}|$$

so that

$$\Xi(I, J, N, \alpha, \beta) = \sum_{p \in P_{|I|+|J|}^2} \omega(p, I, J, N, \alpha, \beta)$$

First of all notice how

$$\omega(p, I, J, N, \alpha, \beta) \leq N^{\frac{|I|+|J|}{2}}$$

in fact considering only the constraints given by α and β we have $N^{|I|}$ ways to choose $\bar{\delta} \in \{1, \ldots, N\}^{|I|+1}$ and $N^{|J|}$ ways to choose $\bar{\epsilon} \in \{1, \ldots, N\}^{|J|+1}$, but since they must come in pairs as dictated by p we actually have $N^{\frac{|I|+|J|}{2}}$ possible choices *i.e.* N per pair.

Notice however how we have equality if and only if $I = J$, $\alpha = \beta$ and the pairings are such that $p \ni \{a, b\} = \{a, |I| + a\}$ for $a \in \{1, \ldots, |I|\}$ *i.e.* every element in I is paired to the corresponding one in $J = I$.

The *if* part is easy to see: the full constraints are just $\delta_{|I|} = \epsilon_{|I|} = \gamma$, $\delta_a = \epsilon_a$ for every $1 < a \leq |I|$ and $\alpha = \delta_0 = \epsilon_0 = \beta$ thus there are

$$N \times N^{|I|-1} \times 1 = N^{|I|} = N^{\frac{|I|+|J|}{2}}$$

choices to make.

The *only if* is more complicated and follows from the constraint $\delta_{|I|} = \epsilon_{|J|}$. Fix γ and assume $i_{|I|}$ is not paired with $j_{|J|}$, then the choices for $(\delta * \epsilon)_a = (\delta * \epsilon)_b$ for 2 out of the $\frac{|I|+|J|}{2}$ pairs are constrained to be γ, all in all we have N choices for γ and at most $N^{\frac{|I|+|J|}{2}-2}$ for the other entries, thus at most $N^{\frac{|I|+|J|}{2}-1}$ in total. But then we must require $i_{|I|}$ and $j_{|J|}$ to be paired. The same argument can now be repeated with $i_{|I|-1}$ and $j_{|J|-1}$ since we have established that $\{|I|, |I|+|J|\} \in p$, thus not only $\delta_{|I|} = (\delta * \epsilon)_{|I|} = (\delta * \epsilon)_{|I|+|J|} = \epsilon_{|J|}$ but also $\delta_{|I|-1} = (\delta * \epsilon)'_{|I|} = (\delta * \epsilon)'_{|I|+|J|} = \epsilon_{|J|-1}$. This goes on until, without loss of generality, we run out of elements in I. If the same happens for J (*i.e.* $|I| = |J|$) we

are done since we have proved that $\forall a \in 1, \ldots, |I|$ we have $i_a = j_a$ and $\alpha = (\delta * \epsilon)'_1 = (\delta * \epsilon)'_{|I|+1} = \beta$. Otherwise $|J| \geq 2 + |I|$ and $(j_{|J|-|I|}, \ldots, j_1)$ are paired between themselves. But since $\{1, |I| + (|J| - |I| + 1)\} \in p$ we have $\alpha = (\delta * \epsilon)'_1 = (\delta * \epsilon)'_{|I|+(|J|-|I|+1)} = (\delta * \epsilon)_{|I|+(|J|-|I|+1)-1} = (\delta * \epsilon)_{|I|+(|J|-|I|)} = \epsilon_{|J|-|I|}$. Which means that there is no free choice one of the remaining couples (*i.e.* the one containing $|I| + (|J| - |I|)$ corresponding to $j_{|J|-|I|}$) thus, reasoning just as before, we cut the number of choices of at least a factor N.

We have just shown that, given a pairing p,

$$\left(\frac{\sigma_A^2}{N}\right)^{\frac{|I|+|J|}{2}} \omega(p, I, J, N, \alpha, \beta) \leq \left(\frac{\sigma_A^2}{N}\right)^{\frac{|I|+|J|}{2}} N^{\frac{|I|+|J|}{2}-1} = \frac{\sigma_A^{|I|+|J|}}{N}$$

except when $I = J$, $\alpha = \beta$ and the pairings are such that $p \ni \{a, b\} = \{a, |I| + a\}$ for $a \in \{1, \ldots, |I|\}$, in which case

$$\left(\frac{\sigma_A^2}{N}\right)^{\frac{|I|+|J|}{2}} \omega(p, I, J, N, \alpha, \beta) = \left(\frac{\sigma_A^2}{N}\right)^{\frac{|I|+|J|}{2}} N^{\frac{|I|+|J|}{2}} = \sigma_A^{|I|+|J|}$$

This means that

$$\mathbb{E}\left[[A_I^T A_J]_\alpha^\beta\right] = \left(\frac{\sigma_A^2}{N}\right)^{\frac{|I|+|J|}{2}} \Xi(I, J, N, \alpha, \beta) = \sigma_A^{|I|+|J|}\left[\delta_\alpha^\beta \delta_I^J + \frac{1}{N}\psi(I, J, N, \alpha, \beta)\right]$$

where $0 < \psi(I, J, N, \alpha, \beta) \leq (|I| + |J|)!!$ i.e. $\psi(I, J, N)$ is a positive constant bounded above by the maximal number of pairings (which occur only when I and J are made up of the same one index). Notice how this bound depends only on $|I|$ and $|J|$ and not on N!

Finally, if S_0^N is normally distributed then

$$\mathbb{E}\left[(S_0^N)^T A_I^T A_J S_0^N\right] = \sum_{n=1}^{N} \sum_{m=1}^{N} \mathbb{E}\left[[S_0^N]_n [S_0^N]_m\right] \mathbb{E}\left[[A_I^T A_J]_n^m\right]$$

$$= \sum_{n=1}^{N} \sigma_{S_0}^2 \mathbb{E}\left[[A_I^T A_J]_n^n\right] = N\sigma_{S_0}^2 \sigma_A^{|I|+|J|}\left[\delta_I^J + \frac{1}{N}\psi(I, J, N)\right] = N\sigma_{S_0}^2 \sigma_A^{|I|+|J|}\left[\delta_I^J + O(\frac{1}{N})\right]$$

with $0 \leq \psi(I, J, N) := \frac{1}{N}\sum_{n=1}^{N} \psi(I, J, N, n, n) \leq (|I| + |J|)!!$.

Let us now look at the case with $\sigma_b > 0$.

Let $V_j(x) := A_j x + b_j$ Take $f(x) = \langle v, x\rangle_{\mathbb{R}^N}$ for some $v \in \mathbb{R}^N$. Then

$$V_j f(x) = d(\langle v, \cdot\rangle_{\mathbb{R}^N})_x[A_j x + b_j] = v^T(A_j x + b_j) = \langle A_j^T v, x\rangle_{\mathbb{R}^N} + \langle v, b_j\rangle_{\mathbb{R}^N}$$

hence by induction

$$V_I f(x) = \langle v, A_{i_k} \cdots A_{i_2}(A_{i_1} x + b_{i_1})\rangle_{\mathbb{R}^N}$$

now we have to study, with $\hat{I} := (i_2, \ldots, i_{|I|})$, the terms

$$\mathbb{E}\left[\langle A_I S_0^N + A_{\hat{I}} b_{i_1}, A_J S_0^N + A_{\hat{J}} b_{j_1}\rangle_{\mathbb{R}^N}\right]$$

$$= \mathbb{E}\left[\langle A_I S_0^N, A_J S_0^N\rangle_{\mathbb{R}^N}\right] + \mathbb{E}\left[\langle A_{\hat{I}} b_{i_1}, A_{\hat{J}} b_{j_1}\rangle_{\mathbb{R}^N}\right]$$

$$+ \mathbb{E}\left[\langle A_{\hat{I}} b_{i_1}, A_J S_0^N\rangle_{\mathbb{R}^N}\right] + \mathbb{E}\left[\langle A_I S_0^N, A_{\hat{J}} b_{j_1}\rangle_{\mathbb{R}^N}\right]$$

$$= \mathbb{E}\left[(S_0^N)^T A_I^T A_J S_0^N\right] + \mathbb{E}\left[b_{i_1}^T A_{\hat{I}}^T A_{\hat{J}} b_{j_1}\right]$$

$$+ \mathbb{E}\left[b_{i_1}^T A_{\hat{I}}^T A_J S_0^N\right] + \mathbb{E}\left[(S_0^N)^T A_I^T A_{\hat{J}} b_{j_1}\right]$$

$$= \mathbb{E}\left[(S_0^N)^T A_I^T A_J S_0^N\right] + \mathbb{E}\left[b_{i_1}^T A_{\hat{I}}^T A_{\hat{J}} b_{j_1}\right]$$

where in the last equality we have used the independence of the terms and their 0 mean.

We already know that the first term is

$$\mathbb{E}\Big[(S_0^N)^T A_I^T A_J S_0^N\Big] = N\sigma_{S_0}^2 \sigma_A^{|I|+|J|}\big[\delta_I^J + O(\frac{1}{N})\big]$$

Concerning the second term: using independence we readily see how we must have $i_1 = j_1$, then

$$\mathbb{E}\Big[b_{i_1}^T A_{\hat{I}}^T A_{\hat{j}} b_{j_1}\Big] = \sigma_b^2 \delta_{i_1}^{j_1} \cdot \sum_{n=1}^{N} \mathbb{E}\Big[[A_{\hat{I}}^T A_{\hat{j}}]_n^n\Big]$$

$$= \sigma_b^2 \delta_{i_1}^{j_1} \cdot \sum_{n=1}^{N} \sigma_A^{|\hat{I}|+|\hat{J}|}\big(\delta_{\hat{I}}^{\hat{j}} + \frac{1}{N}\psi(I,J,N,n,n)\big) = N\sigma_b^2 \sigma_A^{|\hat{I}|+|\hat{J}|}\big[\delta_I^J + O(\frac{1}{N})\big]$$

Hence

$$\mathbb{E}\Big[\langle A_I S_0^N + A_{\hat{I}} b_{i_1}, A_J S_0^N + A_{\hat{j}} b_{j_1}\rangle_{\mathbb{R}^N}\Big] = N(\sigma_{S_0}^2 + \frac{\sigma_b^2}{\sigma_A^2})\sigma_A^{|I|+|J|}\big[\delta_I^J + O(\frac{1}{N})\big]$$

$\square$

C.4.2. CONVERGENCE TO THE SIGNATURE KERNEL

Assume now that the exchange of series with limits and expectation are justified, which we will prove later, then we would like to study the variance of the *expected signature kernels* around their limits.

Proposition C.29. *The coefficients in the expansion of the variance*

$$\mathbb{E}\Big[\Big(\frac{\langle S_s^N(x), S_t^N(y)\rangle_{\mathbb{R}^N}}{N} - \sigma_{S_0}^2\langle Sig(\sigma_A x)_{0,s}, Sig(\sigma_A y)_{0,t}\rangle_{T((\mathbb{R}^d))}\Big)^2\Big]$$

are all $O(\frac{1}{N})$ when $\sigma_b = 0$.

Proof. We have

$$\frac{\langle S_s^N(x), S_t^N(y)\rangle_{\mathbb{R}^N}}{N} - \sigma_{S_0}^2\langle Sig(\sigma_A x)_{0,s}, Sig(\sigma_A y)_{0,t}\rangle_{T((\mathbb{R}^d))}$$

$$= \sum_{I,J\in\mathbb{W}_d} \big[\frac{1}{N}(S_0^N)^T A_I^T A_J S_0^N - \sigma_{S_0}^2 \sigma_A^{|I|+|J|}\delta_I^J\big]\, Sig_{0,s}^I(x)Sig_{0,t}^J(y)$$

thus

$$\Big(\frac{\langle S_s^N(x), S_t^N(y)\rangle_{\mathbb{R}^N}}{N} - \sigma_{S_0}^2\langle Sig(\sigma_A x)_{0,s}, Sig(\sigma_A y)_{0,t}\rangle_{T((\mathbb{R}^d))}\Big)^2 =$$

$$\sum_{I,J,K,L\in\mathbb{W}_d} \big[\frac{1}{N}(S_0^N)^T A_I^T A_J S_0^N - \sigma_{S_0}^2 \sigma_A^{|I|+|J|}\delta_I^J\big]\cdot$$

$$\cdot\big[\frac{1}{N}(S_0^N)^T A_K^T A_L S_0^N - \sigma_{S_0}^2 \sigma_A^{|L|+|K|}\delta_K^L\big]\, Sig_{0,s}^I(x)Sig_{0,t}^J(y)Sig_{0,s}^K(x)Sig_{0,t}^K(y)$$

Let us study

$$
\mathbb{E}\Big[\big[\frac{1}{N}(S_0^N)^T A_I^T A_J S_0^N - \sigma_{S_0}^2 \sigma_A^{|I|+|J|}\delta_I^J\big]\big[\frac{1}{N}(S_0^N)^T A_K^T A_L S_0^N - \sigma_{S_0}^2 \sigma_A^{|L|+|K|}\delta_K^L\big]\Big]
$$

$$
= \frac{1}{N^2}\mathbb{E}\Big[(S_0^N)^T A_I^T A_J S_0^N (S_0^N)^T A_K^T A_L S_0^N\Big]
$$

$$
-\frac{\sigma_{S_0}^2 \sigma_A^{2|I|}}{N}\delta_I^J \mathbb{E}\Big[(S_0^N)^T A_K^T A_L S_0^N\Big] - \frac{\sigma_{S_0}^2 \sigma_A^{2|L|}}{N}\delta_K^L \mathbb{E}\Big[(S_0^N)^T A_I^T A_J S_0^N\Big] + (\sigma_{S_0}^2 \sigma_A^{|L|+|I|})^2 \delta_I^J \delta_K^L
$$

$$
= \frac{1}{N^2}\mathbb{E}\Big[(S_0^N)^T A_I^T A_J S_0^N (S_0^N)^T A_K^T A_L S_0^N\Big]
$$

$$
-\frac{\sigma_{S_0}^2 \sigma_A^{2|I|}}{N}\delta_I^J N\sigma_{S_0}^2 \sigma_A^{|K|+|L|}\big[\delta_K^L + O(\frac{1}{N})\big] - \frac{\sigma_{S_0}^2 \sigma_A^{2|L|}}{N}\delta_J^K N\sigma_{S_0}^2 \sigma_A^{|I|+|J|}\big[\delta_I^J + O(\frac{1}{N})\big] + (\sigma_{S_0}^2 \sigma_A^{|L|+|I|})^2 \delta_I^J \delta_K^L
$$

$$
= \frac{1}{N^2}\mathbb{E}\Big[(S_0^N)^T A_I^T A_J S_0^N (S_0^N)^T A_K^T A_L S_0^N\Big] - (\sigma_{S_0}^2 \sigma_A^{|L|+|I|})^2 \delta_I^J \delta_K^L + O(\frac{1}{N})
$$

from our previous results.

We have

$$
(S_0^N)^T A_I^T A_J S_0^N (S_0^N)^T A_K^T A_L S_0^N =
$$

$$
\sum_{\alpha=1}^N [S_0^N]_\alpha [A_I^T A_J S_0^N (S_0^N)^T A_K^T A_L S_0^N]_\alpha =
$$

$$
\sum_{\alpha,\delta=1}^N [S_0^N]_\alpha [A_I^T A_J S_0^N (S_0^N)^T A_K^T A_L]_\alpha^\delta [S_0^N]_\delta =
$$

$$
\sum_{\alpha,\beta,\gamma,\delta=1}^N [S_0^N]_\alpha [A_I^T A_J]_\alpha^\beta [S_0^N (S_0^N)^T]_\beta^\gamma [A_K^T A_L]_\gamma^\delta [S_0^N]_\delta =
$$

$$
\sum_{\alpha,\beta,\gamma,\delta=1}^N [S_0^N]_\alpha [A_I^T A_J]_\alpha^\beta [S_0^N]_\beta [S_0^N]_\gamma [A_K^T A_L]_\gamma^\delta [S_0^N]_\delta =
$$

$$
\sum_{\alpha,\beta,\gamma,\delta=1}^N [S_0^N]_\alpha [S_0^N]_\beta [S_0^N]_\gamma [S_0^N]_\delta [A_I^T A_J]_\alpha^\beta [A_K^T A_L]_\gamma^\delta
$$

hence by independence

$$
\mathbb{E}\Big[(S_0^N)^T A_I^T A_J S_0^N (S_0^N)^T A_K^T A_L S_0^N\Big] =
$$

$$
\sum_{\alpha,\beta,\gamma,\delta=1}^N \mathbb{E}\Big[[S_0^N]_\alpha [S_0^N]_\beta [S_0^N]_\gamma [S_0^N]_\delta\Big] \mathbb{E}\Big[[A_I^T A_J]_\alpha^\beta [A_K^T A_L]_\gamma^\delta\Big]
$$

If S_0^N is sampled from a Normal distribution as before then

$$
\mathbb{E}\Big[[S_0^N]_\alpha [S_0^N]_\beta [S_0^N]_\gamma [S_0^N]_\delta\Big]
$$

is equal to $\sigma_{S_0}^4$ if $|\{\alpha,\beta,\gamma,\delta\}| = 2$, to $3\sigma_{S_0}^4$ if $|\{\alpha,\beta,\gamma,\delta\}| = 1$ and to 0 otherwise.

Remember how

$$
[A_I^T A_J]_\alpha^\beta = \sum_{\gamma=1}^N \sum_{\bar\delta \in \Lambda_{\gamma,\alpha}^{N,|I|}} \sum_{\bar\epsilon \in \Lambda_{\gamma,\beta}^{N,|J|}} \prod_{i=1}^{|I|}\prod_{j=1}^{|J|} [A_{I_i}]_{\delta_i}^{\delta_{i-1}} [A_{J_j}]_{\epsilon_j}^{\epsilon_{j-1}}
$$

thus

$$[A_I^T A_J]_\alpha^\beta [A_K^T A_L]_\gamma^\delta =$$

$$\sum_{\epsilon,\phi=1}^{N} \sum_{\bar\iota\in\Lambda_{\epsilon,\alpha}^{N,|I|}} \sum_{\bar\zeta\in\Lambda_{\epsilon,\beta}^{N,|J|}} \sum_{\bar\kappa\in\Lambda_{\phi,\gamma}^{N,|K|}} \sum_{\bar\lambda\in\Lambda_{\phi,\delta}^{N,|L|}} \prod_{i=1}^{|I|}\prod_{j=1}^{|J|}\prod_{k=1}^{|K|}\prod_{l=1}^{|L|} [A_{I_i}]_{\iota_i}^{\iota_{i-1}} [A_{J_j}]_{\zeta_j}^{\zeta_{j-1}} [A_{K_k}]_{\kappa_k}^{\kappa_{k-1}} [A_{L_l}]_{\lambda_l}^{\lambda_{l-1}}$$

Let us then define $T = I * J * K * L$, $\theta = (\iota * \zeta * \kappa * \lambda)$ and $\theta' = (\iota * \zeta * \kappa * \lambda)'$ just like before, setting $\mathcal{P} := P_{|I|+|J|+|K|+|L|}^2$ we get

$$\mathbb{E}\Big[[A_I^T A_J]_\alpha^\beta [A_K^T A_L]_\gamma^\delta\Big] =$$

$$\sum_{p\in\mathcal{P}} \sum_{\epsilon,\phi=1}^{N} \sum_{\bar\iota\in\Lambda_{\epsilon,\alpha}^{N,|I|}} \sum_{\bar\zeta\in\Lambda_{\epsilon,\beta}^{N,|J|}} \sum_{\bar\kappa\in\Lambda_{\phi,\gamma}^{N,|K|}} \sum_{\bar\lambda\in\Lambda_{\phi,\delta}^{N,|L|}} \prod_{a,b\in p} \mathbb{E}\Big[[A_{T_a}]_{\theta_a}^{\theta_a'} [A_{T_b}]_{\theta_b}^{\theta_b'}\Big]$$

$$= \Big(\frac{\sigma_A^2}{N}\Big)^{\frac{|I|+|J|+|K|+|L|}{2}} \sum_{p\in\mathcal{P}} \omega(p,I,J,K,L,\alpha,\beta,\gamma,\delta)$$

and once again we need to analyze these ωs which, just as before, must satisfy the constraint

$$\omega(p,I,J,K,L,\alpha,\beta,\gamma,\delta) \leq N^{\frac{|I|+|J|+|K|+|L|}{2}}$$

Since we are interested in the behavior for $N\to\infty$ and $|\mathcal{P}|$ is independent from N we just need to discover when the previous inequality is an equality. This happens, as we have previously discovered, when we the pairing does not add to the possible choices of $\bar\iota,\bar\zeta,\bar\kappa,\bar\lambda$ any more constraints than the unavoidable ones *i.e.* $\theta_{|I|} = \theta_{|I|+|J|}$, $\theta_{|I|+|J|+|K|} = \theta_{|I|+|J|+|K|+|L|}$, $\theta_1' = \alpha$, $\theta_{|I|+1}' = \beta$, $\theta_{|I|+|J|+1}' = \gamma$ and $\theta_{|I|+|J|+|K|+1}' = \delta$.

Reasoning exactly as before this can happen if and only if $I = J$, $\alpha = \beta$, $K = L$, $\gamma = \delta$, $\theta_a = \theta_{|I|+a}$ for $a = 1,\ldots,|I|$ and $\theta_{2|I|+a} = \theta_{2|I|+|K|+a}$ for $a = 1,\ldots,|K|$.

This means that

$$\mathbb{E}\Big[[A_I^T A_J]_\alpha^\beta [A_K^T A_L]_\gamma^\delta\Big] = (\sigma_A)^{|I|+|J|+|K|+|L|}\big[\delta_I^J \delta_\alpha^\beta \delta_K^L \delta_\gamma^\delta + O(\frac{1}{N})\big] \leq$$

$$(\sigma_A)^{|I|+|J|+|K|+|L|}\big[\delta_I^J \delta_\alpha^\beta \delta_K^L \delta_\gamma^\delta + \frac{1}{N}\sigma(I,J,K,L)\big]$$

where $\sigma(I,J,K,L)$ is a positive constant corresponding to the maximal number of non zero pairings.

Since

$$\mathbb{E}\Big[(S_0^N)^T A_I^T A_J S_0^N (S_0^N)^T A_K^T A_L S_0^N\Big]$$

$$= \sum_{\alpha,\beta,\gamma,\delta=1}^{N} \mathbb{E}\Big[[S_0^N]_\alpha [S_0^N]_\beta [S_0^N]_\gamma [S_0^N]_\delta\Big] \mathbb{E}\Big[[A_I^T A_J]_\alpha^\beta [A_K^T A_L]_\gamma^\delta\Big]$$

$$= \sum_{\alpha=1}^{N} 3\sigma_{S_0}^4 \mathbb{E}\Big[[A_I^T A_J]_\alpha^\alpha [A_K^T A_L]_\alpha^\alpha\Big] + \sum_{\substack{\alpha,\beta=1\\\alpha\neq\beta}}^{N} \sigma_{S_0}^4 \mathbb{E}\Big[[A_I^T A_J]_\alpha^\alpha [A_K^T A_L]_\beta^\beta\Big]$$

$$+ \sum_{\substack{\alpha,\beta=1\\\alpha\neq\beta}}^{N} \sigma_{S_0}^4 \mathbb{E}\Big[[A_I^T A_J]_\alpha^\beta [A_K^T A_L]_\alpha^\beta\Big] + \sum_{\substack{\alpha,\beta=1\\\alpha\neq\beta}}^{N} \sigma_{S_0}^4 \mathbb{E}\Big[[A_I^T A_J]_\alpha^\beta [A_K^T A_L]_\beta^\alpha\Big]$$

we end up with

$$\mathbb{E}\Big[(S_0^N)^T A_I^T A_J S_0^N (S_0^N)^T A_K^T A_L S_0^N\Big] =$$

$$\sigma_{S_0}^4 (\sigma_A)^{|I|+|J|+|K|+|L|}\big\{ N3(\delta_I^J \delta_K^L + O(\tfrac{1}{N}))+$$

$$+(N^2-N)(\delta_I^J \delta_K^L + O(\tfrac{1}{N})) + (N^2-N)O(\tfrac{1}{N}) + (N^2-N)O(\tfrac{1}{N})\big\}$$

$$= \sigma_{S_0}^4 (\sigma_A)^{|I|+|J|+|K|+|L|} N^2 \big[\delta_I^J \delta_K^L + O(\tfrac{1}{N})\big] = (\sigma_{S_0}^2 \sigma_A^{|L|+|I|})^2 N^2 \big[\delta_I^J \delta_K^L + O(\tfrac{1}{N})\big]$$

In the end

$$\mathbb{E}\Big[\big[\tfrac{1}{N}(S_0^N)^T A_I^T A_J S_0^N - \sigma_{S_0}^2 \sigma_A^{|I|+|J|}\delta_I^J\big]\big[\tfrac{1}{N}(S_0^N)^T A_K^T A_L S_0^N - \sigma_{S_0}^2 \sigma_A^{|L|+|K|}\delta_K^L\big]\Big] =$$

$$\tfrac{1}{N^2}\mathbb{E}\Big[(S_0^N)^T A_I^T A_J S_0^N (S_0^N)^T A_K^T A_L S_0^N\Big] - (\sigma_{S_0}^2 \sigma_A^{|L|+|I|})^2 \delta_I^J \delta_K^L + O(\tfrac{1}{N}) =$$

$$(\sigma_{S_0}^2 \sigma_A^{|L|+|I|})^2 \big[\delta_I^J \delta_K^L + O(\tfrac{1}{N})\big] - (\sigma_{S_0}^2 \sigma_A^{|L|+|I|})^2 \delta_I^J \delta_K^L + O(\tfrac{1}{N}) = O(\tfrac{1}{N})$$

$\square$

Proposition C.30. *The coefficients in the expansion of the variance*

$$\mathbb{E}\Big[\Big(\frac{\langle S_s^N(x), S_t^N(y)\rangle_{\mathbb{R}^N}}{N} - \big(\sigma_{S_0}^2 + \frac{\sigma_b^2}{\sigma_A^2}\big)\langle Sig(\sigma_A x)_{0,s}, Sig(\sigma_A y)_{0,t}\rangle_{T((\mathbb{R}^d))} + \frac{\sigma_b^2}{\sigma_A^2}\Big)^2\Big]$$

are all $O(\tfrac{1}{N})$ with $\sigma_b > 0$.

Proof. Now

$$\langle S_s^N(x), S_t^N(y)\rangle_{\mathbb{R}^N} =$$

$$\sum_{I,J\in\mathbb{W}_d} \big[(S_0^N)^T A_I^T A_J S_0^N + (b_{I_1})^T A_{\hat{I}}^T A_J S_0^N + (S_0^N)^T A_I^T A_{\hat{J}} b_{J_1} + (b_{I_1})^T A_{\hat{I}}^T A_{\hat{J}} b_{J_1}\big]\cdot$$

$$\cdot Sig_{0,s}^I(x) Sig_{0,t}^J(y)$$

Thus to study

$$\mathbb{E}\Big[\Big(\frac{\langle S_s^N(x), S_t^N(y)\rangle_{\mathbb{R}^N}}{N} - \big(\sigma_{S_0}^2 + \frac{\sigma_b^2}{\sigma_A^2}\big)\langle Sig(\sigma_A x)_{0,s}, Sig(\sigma_A y)_{0,t}\rangle_{T((\mathbb{R}^d))} + \frac{\sigma_b^2}{\sigma_A^2}\Big)^2\Big]$$

we need to study the terms

$$\frac{1}{N^2}\mathbb{E}\Big[\big[(S_0^N)^T A_I^T A_J S_0^N + (b_{I_1})^T A_{\hat{I}}^T A_J S_0^N + (S_0^N)^T A_I^T A_{\hat{J}} b_{J_1} + (b_{I_1})^T A_{\hat{I}}^T A_{\hat{J}} b_{J_1} - N\big(\sigma_{S_0}^2 + \frac{\sigma_b^2}{\sigma_A^2}\big)\sigma_A^{|I|+|J|}\delta_I^J\big]\cdot$$

$$\cdot\big[(S_0^N)^T A_K^T A_L S_0^N + (b_{K_1})^T A_{\hat{K}}^T A_L S_0^N + (S_0^N)^T A_K^T A_{\hat{L}} b_{L_1} + (b_{K_1})^T A_{\hat{K}}^T A_{\hat{L}} b_{L_1} - N\big(\sigma_{S_0}^2 + \frac{\sigma_b^2}{\sigma_A^2}\big)\sigma_A^{|K|+|L|}\delta_K^L\big]\Big]$$

To ease the notation let us write i, j, k and l instead of, respectively, I_1, J_1, K_1 and L_1.

We have

$$\mathbb{E}\Big\{\big[(S_0^N)^T A_I^T A_J S_0^N\big]\cdot$$

$$\cdot\big[(S_0^N)^T A_K^T A_L S_0^N + (b_k)^T A_{\hat{K}}^T A_L S_0^N + (S_0^N)^T A_K^T A_{\hat{L}} b_l + (b_k)^T A_{\hat{K}}^T A_{\hat{L}} b_l - N\big(\sigma_{S_0}^2 + \frac{\sigma_b^2}{\sigma_A^2}\big)\sigma_A^{|K|+|L|}\delta_K^L\big]\Big\}$$

$$= \mathbb{E}\Big[(S_0^N)^T A_I^T A_J S_0^N (S_0^N)^T A_K^T A_L S_0^N\Big]$$

$$+\mathbb{E}\Big[(S_0^N)^T A_I^T A_J S_0^N (b_k)^T A_{\hat{K}}^T A_L S_0^N\Big]$$

$$+\mathbb{E}\Big[(S_0^N)^T A_I^T A_J S_0^N (S_0^N)^T A_K^T A_{\hat{L}} b_l\Big]$$

$$+\mathbb{E}\Big[(S_0^N)^T A_I^T A_J S_0^N (b_k)^T A_{\hat{K}}^T A_{\hat{L}} b_l\Big]$$

$$-N\big(\sigma_{S_0}^2 + \frac{\sigma_b^2}{\sigma_A^2}\big)\sigma_A^{|K|+|L|}\delta_K^L \mathbb{E}\Big[(S_0^N)^T A_I^T A_J S_0^N\Big]$$

which by previous results is equal to

$$(\sigma_{S_0}^2\sigma_A^{|L|+|I|})^2 N^2\big[\delta_I^J\delta_K^L + O(\frac{1}{N})\big]$$

$$+\mathbb{E}\Big[(S_0^N)^T A_I^T A_J S_0^N (b_k)^T A_{\hat{K}}^T A_L S_0^N\Big]$$

$$+\mathbb{E}\Big[(S_0^N)^T A_I^T A_J S_0^N (S_0^N)^T A_K^T A_{\hat{L}} b_l\Big]$$

$$+\mathbb{E}\Big[(S_0^N)^T A_I^T A_J S_0^N (b_k)^T A_{\hat{K}}^T A_{\hat{L}} b_l\Big]$$

$$-\big(\sigma_{S_0}^2 + \frac{\sigma_b^2}{\sigma_A^2}\big)\sigma_A^{|K|+|L|}\delta_K^L \sigma_{S_0}^2\sigma_A^{|I|+|J|} N^2\big[\delta_I^J + O(\frac{1}{N})\big]$$

Note then how

$$\mathbb{E}\Big[(S_0^N)^T A_I^T A_J S_0^N (b_k)^T A_{\hat{K}}^T A_L S_0^N\Big]$$

$$= \sum_{\alpha,\beta,\gamma,\delta=1}^{N} \mathbb{E}\Big[[S_0^N]_\alpha[S_0^N]_\beta[b_k]_\gamma[S_0^N]_\delta\Big]\mathbb{E}\Big[[A_I^T A_J]_\alpha^\beta[A_K^T A_L]_\gamma^\delta\Big]$$

$$= \sum_{\alpha,\beta,\gamma,\delta=1}^{N} \mathbb{E}\Big[[S_0^N]_\alpha[S_0^N]_\beta[S_0^N]_\delta\Big]\mathbb{E}\Big[[b_k]_\gamma\Big]\mathbb{E}\Big[[A_I^T A_J]_\alpha^\beta[A_K^T A_L]_\gamma^\delta\Big] = 0$$

given that $\mathbb{E}\Big[[b_k]_\gamma\Big] = 0$. Analogously we have

$$\mathbb{E}\Big[(S_0^N)^T A_I^T A_J S_0^N (S_0^N)^T A_K^T A_{\hat{L}} b_l\Big] = 0$$

Finally

$$\mathbb{E}\Big[(S_0^N)^T A_I^T A_J S_0^N (b_k)^T A_{\hat{K}}^T A_{\hat{L}} b_l\Big]$$

$$= \sum_{\alpha,\beta,\gamma,\delta=1}^{N} \mathbb{E}\Big[[S_0^N]_\alpha [S_0^N]_\beta [b_k]_\gamma [b_l]_\delta\Big] \mathbb{E}\Big[[A_I^T A_J]_\alpha^\beta [A_{\hat{K}}^T A_{\hat{L}}]_\gamma^\delta\Big]$$

$$= \sum_{\alpha,\beta,\gamma,\delta=1}^{N} \mathbb{E}\Big[[S_0^N]_\alpha [S_0^N]_\beta\Big] \mathbb{E}\Big[[b_k]_\gamma [b_l]_\delta\Big] \mathbb{E}\Big[[A_I^T A_J]_\alpha^\beta [A_{\hat{K}}^T A_{\hat{L}}]_\gamma^\delta\Big]$$

$$= \sigma_{S_0}^2 \sigma_b^2 \sigma_A^{|I|+|J|+|\hat{K}|+|\hat{L}|} \sum_{\alpha,\beta,\gamma,\delta=1}^{N} \delta_\alpha^\beta \delta_\gamma^\delta \delta_k^l \big(\delta_I^J \delta_{\hat{K}}^{\hat{L}} \delta_\alpha^\beta \delta_\gamma^\delta + O(\tfrac{1}{N})\big)$$

$$= \sigma_{S_0}^2 \sigma_b^2 \sigma_A^{|I|+|J|+|\hat{K}|+|\hat{L}|} \sum_{\alpha,\beta,\gamma,\delta=1}^{N} \delta_\alpha^\beta \delta_\gamma^\delta \big(\delta_k^l \delta_I^J \delta_{\hat{K}}^{\hat{L}} + O(\tfrac{1}{N})\big)$$

$$= \sigma_{S_0}^2 \sigma_b^2 \sigma_A^{|I|+|J|+|\hat{K}|+|\hat{L}|} \sum_{\alpha,\beta,\gamma,\delta=1}^{N} \delta_\alpha^\beta \delta_\gamma^\delta \big(\delta_I^J \delta_K^L + O(\tfrac{1}{N})\big)$$

$$= \sigma_{S_0}^2 \sigma_b^2 \sigma_A^{|I|+|J|+|\hat{K}|+|\hat{L}|} \sum_{\alpha,\gamma=1}^{N} \big(\delta_I^J \delta_K^L + O(\tfrac{1}{N})\big) = \sigma_{S_0}^2 \sigma_b^2 \sigma_A^{|I|+|J|+|\hat{K}|+|\hat{L}|} N^2 \big[\delta_I^J \delta_K^L + O(\tfrac{1}{N})\big]$$

where once again we use the fact that all the $O(\frac{1}{N})$ are uniformly bounded above by some $O(\frac{1}{N})$.

Putting everything together

$$\mathbb{E}\Big\{\big[(S_0^N)^T A_I^T A_J S_0^N\big] \cdot$$

$$\cdot \big[(S_0^N)^T A_K^T A_L S_0^N + (b_k)^T A_{\hat{K}}^T A_L S_0^N + (S_0^N)^T A_K^T A_{\hat{L}} b_l + (b_k)^T A_{\hat{K}}^T A_{\hat{L}} b_l - N\big(\sigma_{S_0}^2 + \tfrac{\sigma_b^2}{\sigma_A^2}\big)\sigma_A^{|K|+|L|}\delta_K^L\big]\Big\}$$

$$= (\sigma_{S_0}^2 \sigma_A^{|L|+|I|})^2 N^2 \big[\delta_I^J \delta_K^L + O(\tfrac{1}{N})\big] + 0 + 0 + \sigma_{S_0}^2 \sigma_b^2 \sigma_A^{|I|+|J|+|\hat{K}|+|\hat{L}|} N^2 \big[\delta_I^J \delta_K^L + O(\tfrac{1}{N})\big]$$

$$- \big(\sigma_{S_0}^2 + \tfrac{\sigma_b^2}{\sigma_A^2}\big)\sigma_A^{|K|+|L|}\delta_K^L \sigma_{S_0}^2 \sigma_A^{|I|+|J|} N^2 \big[\delta_I^J + O(\tfrac{1}{N})\big] = N^2 O(\tfrac{1}{N})$$

With analogous arguments we obtain

$$\mathbb{E}\Big[\big[(b_i)^T A_{\hat{I}}^T A_J S_0^N\big] \cdot$$

$$\cdot \big[(S_0^N)^T A_K^T A_L S_0^N + (b_k)^T A_{\hat{K}}^T A_L S_0^N + (S_0^N)^T A_K^T A_{\hat{L}} b_l + (b_k)^T A_{\hat{K}}^T A_{\hat{L}} b_l - N\big(\sigma_{S_0}^2 + \tfrac{\sigma_b^2}{\sigma_A^2}\big)\sigma_A^{|K|+|L|}\delta_K^L\big]\Big]$$

$$= 0 + \mathbb{E}\Big[\big((b_i)^T A_{\hat{I}}^T A_J S_0^N (b_k)^T A_{\hat{K}}^T A_L S_0^N\big] + \mathbb{E}\Big[(b_i)^T A_{\hat{I}}^T A_J S_0^N (S_0^N)^T A_K^T A_{\hat{L}} b_l\Big] + 0 - N\big(\sigma_{S_0}^2 + \tfrac{\sigma_b^2}{\sigma_A^2}\big)\sigma_A^{|K|+|L|}\delta_K^L \cdot 0$$

and since

$$\mathbb{E}\Big[(b_i)^T A_{\hat{I}}^T A_J S_0^N (b_k)^T A_{\hat{K}}^T A_L S_0^N\Big]$$

$$= \sum_{\alpha,\beta,\gamma,\delta=1}^{N} \mathbb{E}\Big[[S_0^N]_\beta [S_0^N]_\delta\Big] \mathbb{E}\Big[[b_i]_\alpha [b_k]_\gamma\Big] \mathbb{E}\Big[[A_{\hat{I}}^T A_J]_\alpha^\beta [A_{\hat{K}}^T A_L]_\gamma^\delta\Big]$$

$$= \sigma_{S_0}^2 \sigma_b^2 \sigma_A^{|\hat{I}|+|J|+|\hat{K}|+|L|} \sum_{\alpha,\beta,\gamma,\delta=1}^{N} \delta_\beta^\delta \delta_\alpha^\gamma \delta_i^k (\delta_{\hat{I}}^J \delta_{\hat{K}}^L \delta_\alpha^\beta \delta_\gamma^\delta + O(\frac{1}{N}))$$

$$= \sigma_{S_0}^2 \sigma_b^2 \sigma_A^{|\hat{I}|+|J|+|\hat{K}|+|L|} \Big[\sum_{\alpha,\beta,\gamma,\delta=1}^{N} \delta_\beta^\delta \delta_\alpha^\gamma \delta_i^k \delta_{\hat{I}}^J \delta_{\hat{K}}^L \delta_\alpha^\beta \delta_\gamma^\delta + \sum_{\alpha,\beta,\gamma,\delta=1}^{N} \delta_\beta^\delta \delta_\alpha^\gamma \delta_i^k O(\frac{1}{N})\Big]$$

$$= \sigma_{S_0}^2 \sigma_b^2 \sigma_A^{|\hat{I}|+|J|+|\hat{K}|+|L|} \Big[\sum_{\alpha=1}^{N} \delta_i^k \delta_{\hat{I}}^J \delta_{\hat{K}}^L + \sum_{\alpha,\beta=1}^{N} \delta_i^k O(\frac{1}{N})\Big]$$

$$\sigma_{S_0}^2 \sigma_b^2 \sigma_A^{|\hat{I}|+|J|+|\hat{K}|+|L|} \Big[N\delta_i^k \delta_{\hat{I}}^J \delta_{\hat{K}}^L + N^2 \delta_i^k O(\frac{1}{N})\Big] = N^2 O(\frac{1}{N})$$

and similarly

$$\mathbb{E}\Big[(b_i)^T A_{\hat{I}}^T A_J S_0^N (S_0^N)^T A_K^T A_{\hat{L}} b_l\Big] = N^2 O(\frac{1}{N})$$

we finally obtain

$$\mathbb{E}\Big\{ \big[(b_i)^T A_{\hat{I}}^T A_J S_0^N\big] \cdot$$

$$\cdot \big[(S_0^N)^T A_K^T A_L S_0^N + (b_k)^T A_{\hat{K}}^T A_L S_0^N + (S_0^N)^T A_K^T A_{\hat{L}} b_l + (b_k)^T A_{\hat{K}}^T A_{\hat{L}} b_l - N(\sigma_{S_0}^2 + \frac{\sigma_b^2}{\sigma_A^2})\sigma_A^{|K|+|L|}\delta_K^L\big]\Big\}$$

$$= N^2 O(\frac{1}{N})$$

Using the same arguments developed up to here all other terms in the product end up as being $N^2 O(\frac{1}{N})$ thus dividing finally by N^2 we have the thesis. $\qquad\square$

We will just do the case $(\sigma_{S_0}, \sigma_A, \sigma_B) = (1, 1, 0)$. The arguments for the general case are the same.

Proposition C.31.

$$\lim_{N\to\infty} \sum_{I\in\mathbb{W}_d} \sum_{J\in\mathbb{W}_d} \frac{1}{N} \mathbb{E}[(S_0^N)^T A_I^T A_J S_0^N] Sig_{0,s}^I(x) Sig_{0,t}^J(y)$$

$$= \sum_{I\in\mathbb{W}_d} \sum_{J\in\mathbb{W}_d} \lim_{N\to\infty} \frac{1}{N} \mathbb{E}[(S_0^N)^T A_I^T A_J S_0^N] Sig_{0,s}^I(x) Sig_{0,t}^J(y)$$

Proof. First of all, to be thoroughly rigorous, we need to define a probability space over which we take all the expectations, everything is numerable thus there is no issues with this.

To justify the exchange of sum and limit we want to use Lebesgue dominated convergence, we thus need to bound the

$$|\frac{1}{N} \mathbb{E}[(S_0^N)^T A_I^T A_J S_0^N] Sig_{0,s}^I(x) Sig_{0,t}^J(y)|$$

from above.

We know, from previous considerations, that

$$\frac{1}{N} \mathbb{E}[(S_0^N)^T A_I^T A_J S_0^N] = \sum_{n=1}^{N} \frac{1}{N} \mathbb{E}\Big[[A_I^T A_J]_n^n\Big]$$

and that

$$\mathbb{E}\left[[A_I^T A_J]_n^n\right] = \left(\frac{1}{N}\right)^{\frac{|I|+|J|}{2}} \sum_{p \in P_{|I|+|J|}^2} \omega(p, I, J, N, n, n)$$

thus, since $\omega(p, I, J, N, \alpha, \beta) \leq N^{\frac{|I|+|J|}{2}}$, we obtain

$$\mathbb{E}\left[[A_I^T A_J]_n^n\right] \leq |P_{|I|+|J|}^2| \leq (|I| + |J|)!!$$

We have also seen that $\mathbb{E}\left[[A_I^T A_J]_n^n\right] = 0$ if $|I| + |J|$ is odd, thus we can consider $|I| + |J|$ to be even; then, writing i, j instead of $|I|, |J|$ for ease of reading, we get

$$(|I| + |J|)!! = \left(2\left(\frac{i+j}{2}\right)\right)!! = 2^{\frac{i+j}{2}}\left(\frac{i+j}{2}\right)!$$

Putting everything together we have found

$$|\frac{1}{N}\mathbb{E}[(S_0^N)^T A_I^T A_J S_0^N]| \leq 2^{\frac{i+j}{2}}\left(\frac{i+j}{2}\right)!$$

and

$$2 \nmid i + j \implies \frac{1}{N}\mathbb{E}[(S_0^N)^T A_I^T A_J S_0^N] = 0$$

Finally remember how, by factorial decay,

$$|Sig_{0,s}^I(x)| \leq \frac{\|x\|_{1-var}^{|I|}}{|I|!}$$

We have now to prove, by (Tao, 2016)[8.2.1 and 8.2.2], that

$$\sum_{I \in \mathbb{W}_d} \sum_{J \in \mathbb{W}_d} 2^{\frac{|I|+|J|}{2}}\left(\frac{|I|+|J|}{2}\right)! \ \mathbb{I}_{2||I|+|J|} \ \frac{\|x\|^{|I|}}{|I|!} \frac{\|y\|^{|J|}}{|J|!}$$

$$= \sum_{i \in \mathbb{N}} \sum_{j \in \mathbb{N}} d^i d^j 2^{\frac{i+j}{2}}\left(\frac{i+j}{2}\right)! \ \mathbb{I}_{2|i+j} \ \frac{\|x\|^i}{i!} \frac{\|y\|^j}{j!} < \infty$$

Once again by (Tao, 2016)[8.2.1 and 8.2.2] we have to find a bijection $\phi : \mathbb{N} \to \mathbb{N} \times \mathbb{N}$ such that, if

$$f(i, j) = d^i d^j 2^{\frac{i+j}{2}}\left(\frac{i+j}{2}\right)! \ \mathbb{I}_{2|i+j} \ \frac{\|x\|^i}{i!} \frac{\|y\|^j}{j!}$$

then

$$\sum_{k \in \mathbb{N}} f(\phi(k)) < \infty$$

As a first step assume $2|i + j$ and, writing $i \wedge j := \min\{i, j\}$ and $i \vee j := \max\{i, j\}$, note how

$$f(i, j) = f(i \wedge j, i \vee j)$$

$$\leq \frac{(d\|x\|)^i (d\|y\|)^j 2^{\frac{i+j}{2}}\left(\frac{i+j}{2}\right)!}{(i \wedge j)!(i \vee j)!} = \frac{(d\|x\|)^i (d\|y\|)^j 2^{\frac{i+j}{2}}\left(\frac{i+j}{2}\right) \cdots (i \wedge j + 1)}{(i \vee j)!} = \frac{(d\|x\|)^i (d\|y\|)^j 2^{\frac{i+j}{2}}}{(i \vee j) \cdots \left(\frac{i+j}{2}+1\right) \cdot (i \wedge j)!}$$

$$\leq \frac{d^{i+j} \|x\|^i \|y\|^j 2^{\frac{i+j}{2}}}{\left(\frac{i+j}{2}\right)!} \leq \frac{d^{i+j}\left[(1 + \|x\|)(1 + \|y\|)\right]^{i+j} 2^{\frac{i+j}{2}}}{\left(\frac{i+j}{2}\right)!}$$

Consider then ϕ as the inverse of the map $(i,j) \mapsto \frac{1}{2}(i+j)(i+j+1) + j$ *i.e.* ϕ is the map enumerating pairs (i,j) starting from $(0,0)$ and proceeding with diagonal motions of the form

$$(m,0) \to (m-1,1) \to \cdots \to (0,m)$$

and then $(0,m) \to (m+1,0)$. Notice that such diagonal strides have length $m+1$ and comprise all couples (i,j) such that $i+j=m$.

Since there are exactly $2k+1$ choices of (i,j) such that $\frac{i+j}{2} = k$, corresponding to the couples $(i, 2k-i)$ for $i = 0, \ldots, 2k$, using the above ϕ it suffices to prove

$$\sum_{k \in \mathbb{N}} (2k+1) \frac{\left[d(1+\|x\|)(1+\|y\|)\right]^{2k} 2^k}{k} = \sum_{k \in \mathbb{N}} (2k+1) \frac{\left[2d^2(1+\|x\|)^2(1+\|y\|)^2\right]^k}{k!} < \infty$$

Since

$$\sum_{k \in \mathbb{N}} (2k+1) \frac{\left[2d^2(1+\|x\|)^2(1+\|y\|)^2\right]^k}{k!}$$

$$\leq \sum_{k \in \mathbb{N}} 2^{k+1} \frac{\left[2d^2(1+\|x\|)^2(1+\|y\|)^2\right]^k}{k!} = 2e^{\left[2d(1+\|x\|)(1+\|y\|)\right]^2} < \infty$$

we are done.

$\square$

Proposition C.32. *The exchange of sums and expectations has always been justified.*

Proof. For the case with just two indices I, J we need to use Fubini-Tonelli and prove that

$$\mathbb{E}\left[\sum_{I \in \mathbb{W}_d} \sum_{J \in \mathbb{W}_d} \frac{1}{N} |(S_0^N)^T A_I^T A_J S_0^N Sig_{0,s}^I(x) Sig_{0,t}^J(y)| \right]$$

$$= \sum_{I \in \mathbb{W}_d} \sum_{J \in \mathbb{W}_d} \mathbb{E}[\frac{1}{N} |(S_0^N)^T A_I^T A_J S_0^N|] |Sig_{0,s}^I(x) Sig_{0,t}^J(y)| < \infty$$

We have

$$\frac{1}{N} \mathbb{E}[|(S_0^N)^T A_I^T A_J S_0^N|] \leq \sqrt{\frac{1}{N^2} \mathbb{E}[((S_0^N)^T A_I^T A_J S_0^N)^2]}$$

Remember how

$$\frac{1}{N^2} \mathbb{E}[(S_0^N)^T A_I^T A_J S_0^N (S_0^N)^T A_I^T A_J S_0^N]$$

$$= \sum_{\alpha=1}^{N} \frac{3}{N^2} \mathbb{E}\left[[A_I^T A_J]_\alpha^\alpha [A_I^T A_J]_\alpha^\alpha \right] + \sum_{\substack{\alpha,\beta=1 \\ \alpha \neq \beta}}^{N} \frac{1}{N^2} \mathbb{E}\left[[A_I^T A_J]_\alpha^\alpha [A_I^T A_J]_\beta^\beta \right]$$

$$+ \sum_{\substack{\alpha,\beta=1 \\ \alpha \neq \beta}}^{N} \frac{1}{N^2} \mathbb{E}\left[[A_I^T A_J]_\alpha^\beta [A_I^T A_J]_\alpha^\beta \right] + \sum_{\substack{\alpha,\beta=1 \\ \alpha \neq \beta}}^{N} \frac{1}{N^2} \mathbb{E}\left[[A_I^T A_J]_\alpha^\beta [A_I^T A_J]_\beta^\alpha \right]$$

and how

$$\mathbb{E}\left[[A_I^T A_J]_\alpha^\beta [A_K^T A_L]_\gamma^\delta \right] =$$

$$= \frac{1}{N^{\frac{|I|+|J|+|K|+|L|}{2}}} \sum_{p \in \mathcal{P}} \omega(p, I, J, K, L, \alpha, \beta, \gamma, \delta)$$

65

where

$$\omega(p, I, J, K, L, \alpha, \beta, \gamma, \delta) \leq N^{\frac{|I|+|J|+|K|+|L|}{2}}$$

thus

$$\mathbb{E}\left[[A_I^T A_J]_\alpha^\beta [A_K^T A_L]_\gamma^\delta\right] \leq |P_{|I|+|J|+|K|+|L|}^2| \leq (|I|+|J|+|K|+|L|)!!$$

In our case $K = I$ and $J = L$, hence

$$(|I|+|J|+|K|+|L|)!! = (2|I|+2|J|)!! \leq 2^{|I|+|J|}(|I|+|J|)!$$

Putting all together

$$\frac{1}{N}\mathbb{E}[|(S_0^N)^T A_I^T A_J S_0^N|] \leq \sqrt{\frac{1}{N^2}\mathbb{E}[((S_0^N)^T A_I^T A_J S_0^N)^2]}$$
$$\leq \sqrt{6 \cdot 2^{|I|+|J|}(|I|+|J|)!} \leq \sqrt{6} \cdot 2^{\frac{|I|+|J|}{2}}(|I|+|J|)!!$$

where we have used

$$(|I|+|J|)! = (|I|+|J|)!!(|I|+|J|-1)!! \leq [(|I|+|J|)!!]^2$$

Finally

$$\sum_{I \in \mathbb{W}_d} \sum_{J \in \mathbb{W}_d} \frac{1}{N}\mathbb{E}[|(S_0^N)^T A_I^T A_J S_0^N|]|Sig_{0,s}^I(x)Sig_{0,t}^J(y)|$$
$$\leq \sum_{I \in \mathbb{W}_d} \sum_{J \in \mathbb{W}_d} \sqrt{6} \cdot 2^{\frac{|I|+|J|}{2}}(|I|+|J|)!!\frac{\|x\|^{|I|}}{|I|!}\frac{\|y\|^{|J|}}{|J|!}$$

which is proved to be $< \infty$ exactly as in the previous proof, this time taking care to consider also the case $|I|+|J|$ *not* even.

The case with 4 words, *i.e.* the variance case, goes similarly. $\qquad\square$

Putting all of this together we have finally proved the theorem:

Theorem C.33. *Consider randomized Signatures of the type*

$$S_t^N(x) = S_0 + \sum_{k=1}^d \int_0^t \left(A_k S_\tau^N(x) + b_k\right)dx_\tau^k$$

where $x \in \mathbb{X}$ and

$$[A_k]_\alpha^\beta \sim \mathcal{N}(0, \frac{\sigma_A^2}{N}) \qquad [Y_0]_\alpha \sim \mathcal{N}(0, \sigma_{S_0}^2) \qquad [b_k]_\alpha \sim \mathcal{N}(0, \sigma_b^2)$$

Then

$$\lim_{N \to \infty} \mathbb{E}\left[\frac{1}{N}\langle S_s^N(x), S_t^N(y)\rangle_{\mathbb{R}^N}\right] = \left(\sigma_{S_0}^2 + \frac{\sigma_b^2}{\sigma_A^2}\right)k_{sig}^{\sigma_A x, \sigma_A y}(s, t) - \frac{\sigma_b^2}{\sigma_A^2}$$

and the variance around the limit is of order $O(\frac{1}{N})$.

C.4.3. CONVERGENCE TO A GAUSSIAN PROCESS

Fix $\mathcal{X} := \{x_1, \ldots, x_m\} \subseteq \mathbb{X}$. Let moreover $\phi \in \mathbb{R}^N$ be sampled from $\mathcal{N}(0, \frac{1}{N})$. Define the vectors

$$\Phi_{\mathcal{X}}^N := [\langle \phi, S_1^N(x_j)\rangle_{\mathbb{R}^N}]_{j=1,\ldots,m}$$

Proposition C.34. $\Phi_{\mathcal{X}}^N$ *converge in distribution to a $\mathcal{N}(0, \mathcal{K}_{id}^{\mathcal{X}})$.*

Proof. By *Lévy's continuity theorem* it suffices to study the limiting behavior of the characteristic functions

$$\varphi_N(u) := \mathbb{E}[\exp\{iu^T \Phi_{\mathcal{X}}^N\}]$$

$$=\mathbb{E}[\exp\{i\sum_{j=1}^m u_j[\Phi_{\mathcal{X}}^N]_j\}] = \mathbb{E}[\exp\{i\sum_{j=1}^m u_j \langle \phi, S_1^N(x_j)\rangle\}]$$

$$=\mathbb{E}[\exp\{i\left\langle \sum_{j=1}^m u_j S_1^N(x_j), \phi \right\rangle\}]$$

$$=\mathbb{E}[\mathbb{E}[\exp\{i\left\langle \sum_{j=1}^m u_j S_1^N(x_j), \phi \right\rangle\}|A_1, b_1, \ldots, A_d, b_d, S_0]]$$

But if we know S_0 and the A_k, b_k the randomized signatures are deterministic objects, and ϕ is normally distributed; thus

$$\varphi_N(u) = \mathbb{E}[\mathbb{E}[\exp\{i\left\langle \sum_{j=1}^m u_j S_1^N(x_j), \phi \right\rangle\}|A_1, b_1, \ldots, A_d, b_d, S_0]]$$

$$=\mathbb{E}[\exp\{-\frac{1}{2N}\left\langle \sum_{j=1}^m u_j S_1^N(x_j), \sum_{j=1}^m u_j S_1^N(x_j) \right\rangle\}]$$

$$=\mathbb{E}[\exp\{-\frac{1}{2}\sum_{i,j=1}^m u_i u_j \frac{1}{N}\left\langle S_1^N(x_i), S_1^N(x_j)\right\rangle\}]$$

$$=E[\exp\{-\frac{1}{2}\left\langle u, \frac{1}{N}G_{\mathcal{X}}^N u\right\rangle\}]$$

where

$$[G_{\mathcal{X}}^N]_i^j := \left\langle S_1^N(x_i), S_1^N(x_j)\right\rangle$$

Now, since we have proven that

$$\frac{1}{N}\left\langle S_1^N(x_i), S_1^N(x_j)\right\rangle \xrightarrow[N\to\infty]{\mathbb{L}^2} [\mathcal{K}_{id}^{\mathcal{X}}]_i^j$$

we have

$$\frac{1}{N}G_{\mathcal{X}}^N \xrightarrow[N\to\infty]{\mathbb{L}^2} \mathcal{K}_{id}^{\mathcal{X}} := \Sigma$$

Thus, since $\mathbb{L}^2$ convergence implies convergence in distribution, by the classical *Portmanteau theorem* we must have

$$\mathbb{E}[h(\frac{1}{N}G_{\mathcal{X}}^N)] \xrightarrow[N\to\infty]{} \mathbb{E}[h(\Sigma)] = h(\Sigma)$$

for every continuous and bounded $h : \mathbb{R}^{m\times m} \to \mathbb{R}$

Fix $u \in \mathbb{R}^m$ and consider

$$f_u : \mathbb{R}^{m\times m} \to \mathbb{R} \text{ s.t. } A \mapsto \exp\{-\frac{1}{2}\langle u, Au\rangle\}$$

We would like to take $h = f_u$, unfortunately if $u \neq 0$ then f_u is *not* bounded: taking $A = -\mathbb{I}_{m\times m}$ we have

$$f(kA) = \exp\{\frac{k}{2}\|u\|_2^2\} \xrightarrow[k\to\infty]{} +\infty$$

Fortunately we are only interested in evaluating f_u on the $G_{\mathcal{X}}^N$ which are all positive semidefinite matrices:

$$G_{\mathcal{X}}^N = (W_{\mathcal{X}}^N)^T W_{\mathcal{X}}^N$$

where $W_{\mathcal{X}}^N \in \mathbb{R}^{N \times m}$ is defined by

$$[W_{\mathcal{X}}^N]_i^j := [S^N(x_j)]_i$$

thus for any $u \in \mathbb{R}^m$ it holds that

$$u^T \frac{1}{N} G_{\mathcal{X}}^N u = u^T \frac{1}{N} (W_{\mathcal{X}}^N)^T W_{\mathcal{X}}^N u = (\frac{1}{\sqrt{N}} W_{\mathcal{X}}^N u)^T \frac{1}{\sqrt{N}} W_{\mathcal{X}}^N u = \left\| \frac{1}{\sqrt{N}} W_{\mathcal{X}}^N u \right\|_2^2 \geq 0$$

Since the $G_{\mathcal{X}}^N$ are semidefinite we have

$$0 < f_u(G_{\mathcal{X}}^N) \leq 1$$

for any u, hence we always have

$$f_u(G_{\mathcal{X}}^N) = \psi \circ f_u(G_{\mathcal{X}}^N)$$

where $\psi : \mathbb{R} \to \mathbb{R}$ is defined by $\psi(x) := \min\{x, 2\}$.

To end notice how $\psi \circ f_u : \mathbb{R}^{m \times m} \to \mathbb{R}$ is continuous, being composition of continuous functions, and bounded, since

$$0 < \psi \circ f_u(A) \leq 2$$

Then we have, for every $u \in \mathbb{R}^m$, that

$$\mathbb{E}[f_u(G_{\mathcal{X}}^N)] = \mathbb{E}[(\psi \circ f_u)(\frac{1}{N} G_{\mathcal{X}}^N)] \xrightarrow[N \to \infty]{} (\psi \circ f_u)(\Sigma) = f_u(\Sigma)$$

where we have used the semidefinitiveness of Σ. With this we finally conclude that

$$\varphi_N(u) \xrightarrow[N \to \infty]{} \exp\left\{ -\frac{1}{2} \langle u, \Sigma u \rangle \right\}$$

$\square$

We have just proved

Proposition C.35. *Let $\varphi = id$. For any subset $\mathcal{X} = \{x_1, \ldots, x_n\} \subset \mathbb{X}$ the following convergence in distribution holds*

$$\lim_{N \to \infty} \lim_{M \to \infty} \Psi_{id}^{M,N}(\mathcal{X}) = \lim_{M \to \infty} \lim_{N \to \infty} \Psi_{id}^{M,N}(\mathcal{X})$$