Imperial College London

Department of Mathematics

Graph Expansions of Neural Networks

Theory and Applications for Time-Series Analysis

A Dissertation Submitted in Partial Fulfillment
of the Requirements for the Degree of

Doctor of Philosophy

by

Nicola Muça Cirone

Supervised by
Dr. Cristopher Salvi

September 13, 2025

Copyright Declaration

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial 4.0 International Licence (CC BY-NC). Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose. When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes. Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

©2024 Nicola Muça Cirone

Statement of Originality

I certify that this thesis and the research to which it refers are the product of my own work, and that any ideas or quotations from the work of other people, published or otherwise, are fully acknowledged in accordance with the standard referencing practices of the discipline.

Nicola Muça Cirone
September 2025

To Gigia and Ninuccia.

Abstract

The study of scaling limits in neural networks, particularly as the number of parameters grows large, provides crucial insights into their generalization properties, training stability, and information processing efficiency, making it an area of significant empirical interest. However, existing approaches often lack mathematical rigor and are specific to certain architectures and hyperparameters. Our research introduces a novel mathematical framework for analyzing the scaling limits of arbitrary architectures. By utilizing genus expansion techniques from random matrix theory, we represent neural networks as Taylor-like series of operator graphs, separating the effects of activation functions from the network's random weights.

We then leverage these expansions to study the statistical properties of Feed Forward Neural Networks, State Space Models, and Neural Controlled Differential Equations, discovering new expressive kernels on path spaces.

Acknowledgements

I would like to thank my supervisor, Cristopher Salvi, for allowing me to embark on this extraordinary journey of intellectual growth and for his unwavering support of my ideas, even when these were only germs in my mind, or incomprehensible scribbles on the blackboard. This work would not have been possible without his patient guidance, mathematical vision, and encouragement.

Then I must thank my colleague, co-author, dear friend Jad Hamdan with whose help the graphs, protagonists of this thesis, took their polished form, reached the mathematical maturity needed to be shared to the world. Thank you for bearing with these ideas, with wacky springs and "pairings in pairs", for seeing the beauty underneath the initial confusion, for working hard together in order to make it emerge.

A deep thank you to Josef Teichmann, for his great support and example, for the ever-interseting discussions about Mathematics, Literature, Art.

Thank you to Enrico, Luca, Ruben, Yun, Owen, Henry, Joseph, Lorenzo, Jean and everyone else in WCM, for having been such an energizing force, for leading the path.

Thank you to Pietro, Matteo, Gabor and Mateo for invaluable intellectual enrichment, for giving salt to my student life.

Thank you to Filippo and Alessandro for embarking on this academic journey in England together, and to Prachi and XingYun for patiently listening to (and apparently enjoying) my mathematical musings.

Finally, thank you Virginia, Çlirim and Ilir for your ever-present encouragement and support.

Nicola Muca Cirone
November 14, 2025

Contents

Acknowledgements **v**

Introduction **1**

0.1 Motivation . . . 1
0.2 Contributions . . . 2
0.3 Thesis Outline . . . 3

1 Genus Expansion **7**

1.1 Introduction . . . 7
1.1.1 Overview of our method. . . . 8
1.1.2 Notation and nomenclature . . . 9
1.2 Graphical Descriptions of Analytic Operations . . . 10
1.2.1 Operator graphs and their associated linear map . . . 12
1.2.2 Operations on graphs. . . . 14
1.3 Genus Expansion and Limit Theorems . . . 21
1.3.1 Wick expansion for product graphs with random inputs . . . 21
1.3.2 Genus expansion . . . 25

2 Graph Expansions of Neural Networks **31**

2.1 Introduction . . . 31
2.1.1 Overview of our method. . . . 35
2.1.2 Possible generalizations . . . 36
2.2 Tree Expansion of Feed Forward Neural Networks . . . 37
2.2.1 Expansions for related quantities . . . 41
2.2.2 Atomic pairings of trees . . . 44
2.3 Gaussian Process Limit . . . 45
2.4 Neural Tangent Kernel . . . 50
2.5 Jacobian Moments . . . 55
2.6 Observations on the Algebra of Trees . . . 64

3 State Space Models **67**

3.1 Introduction . . . 67
3.2 State Space Models and Linear CDEs . . . 68

3.2.1 Linear CDEs . . . 69
3.2.2 Taylor expansion for Linear CDEs . . . 71
3.3 Signature Kernel as Limit of SSMs . . . 73
3.3.1 L^2 convergence to the signature kernel . . . 75
3.3.2 Convergence to a Gaussian process . . . 77
3.4 Expressive Power of SSMs . . . 79
3.4.1 Characterization of the closure: Dense Case . . . 80
3.4.2 Characterization of the closure: Diagonal Case . . . 89
3.5 Chaining of Diagonal Systems . . . 93
3.6 Extensions to Structured Matrices . . . 96
3.6.1 Sparse Matrices . . . 97
3.6.2 Walsh-Hadamard Matrices . . . 97
3.6.3 Block Diagonal . . . 100
3.7 Experiments . . . 101

4 Neural Signature Kernels **105**
4.1 Introduction . . . 105
4.2 Neural Signature Kernels . . . 106
4.2.1 Notation . . . 106
4.2.2 The Model . . . 107
4.2.3 The infinite-width-depth regime . . . 107
4.2.4 The finite-width, infinite-depth regime . . . 110
4.2.5 Assumptions on path regularity . . . 111
4.2.6 Assumptions on the activation function . . . 111
4.2.7 Positive semidefinite matrices and the map V_φ . . . 112
4.3 Proofs for homogeneous controlled ResNets . . . 114
4.3.1 The infinite-width-then-depth regime . . . 114
4.3.2 Proof Main Theorem, part I . . . 128
4.3.3 The infinite-depth-then-width regime . . . 130
4.3.4 Commutativity of Limits . . . 131
4.4 Numerics . . . 137

A Genus Expansion for non-Gaussian, sparse and complex weights. **139**
A.1 Non-Gaussian Weights . . . 139
A.2 Sparse Weights . . . 142
A.3 Complex Weights . . . 144

B Technical Combinatorial Lemmas **147**

C Background on Rough Paths Theory **153**

C.1 Basic Definitions . . . 153
C.2 Rough Paths Primer . . . 154
C.3 Signature . . . 158
C.3.1 Notable Results . . . 158
C.4 Wronskian Theory . . . 161
C.4.1 ZOH and Exact Solutions . . . 165

D Graph Expansion for Neural CDEs **167**
D.1 Preliminaries . . . 167
D.1.1 Tree expansion for MLPs . . . 170
D.2 Graph Expansion . . . 172
D.3 Assumptions . . . 174
D.4 GP limits . . . 175
D.5 NTK . . . 177

E Scaling Limits of Training Dynamics **183**
E.0.1 The Model . . . 183
E.0.2 Stochastic Gradient Descent . . . 184
E.1 The Linear Case . . . 185
E.1.1 Graphical expansion . . . 185
E.1.2 Training dynamics . . . 186
E.1.3 The L^2 limit . . . 194
E.2 The Non-Linear Case . . . 203

Introduction

"It should be one's sole endeavour to see everything afresh and create it anew"
- Gustav Mahler

0.1 Motivation

As deep learning models increase in size and complexity, they demand exponentially more computational resources [122, 126], making efficiency in resource utilization a critical concern [11]. By understanding the limiting behavior as the number of parameters grows without bounds, practitioners can optimize resource allocation [132], ensuring that models are trained efficiently and without unnecessary computational overhead—an essential consideration in environments where computational power is limited or costly [116].

Traditionally, this regime has been studied in the context of networks at initialization, such as the Gaussian Process (GP) limit of Neal [91]. However, recent research has begun to address training dynamics in the infinite-width regime [58, 117], revealing that different parameterizations yield distinct behaviors during gradient descent. Notably, recent work on μP parameterization [132] has highlighted a clear connection with hyperparameter transfer, a process that significantly reduces the resource-intensive task of hyperparameter tuning by enabling the reuse of optimized parameters across similar tasks or models; particularly advantageous for large-scale models, where tuning new instances from scratch would be prohibitively expensive.

Despite these advances, the field is still hindered by the lack of a unified treatment and broad applicability of these results, which are derived through diverse and often fragmented mathematical techniques [24, 39, 103, 130]. This fragmentation makes existing results and techniques obscure, requiring highly specialized knowledge to study different architectures. In particular, existing techniques do not extend to the controlled differential equation case [65], encompassing modern architectures such as linear transformers [133] and state space models [88]. The need for a unified, theoretically sound framework built on first principles is therefore evident.

0.2 Contributions

This thesis introduces a unified mathematical framework for analyzing neural network scaling limits that directly addresses the theoretical fragmentation identified above. Our approach centers on genus expansion techniques from random matrix theory [28, 51], through which we represent neural networks as Taylor-like series indexed over decorated graphs. The summands of these series systematically decompose the effects of inputs, activation functions, and network weights according to the topological properties of their corresponding graphs, creating a natural separation of concerns that has been absent in previous approaches.

The core innovation lies in establishing a functorial correspondence between analytic operations on neural networks and combinatorial operations on their graphical representations. This correspondence transforms seemingly intractable probabilistic scaling limit problems into highly structured combinatorial ones, leveraging the rich topological properties of the underlying graphical objects. Through this transformation, we achieve what previous approaches could not: a unified treatment that recovers classical results while enabling fundamentally new computations.

Our framework provides the first rigorous unified derivation of foundational results in neural network theory. We recover both the Gaussian Process limit of Neal [91] and the Neural Tangent Kernel limit of Jacot et al. [58] as special cases of our general combinatorial theory, eliminating the need for the disparate analytical techniques traditionally required for each result [24, 130]. More significantly, our approach enables previously impossible computations, such as explicit combinatorial formulas for the moments of network Jacobians [103], providing new quantitative tools for understanding network behavior that extend far beyond existing characterizations. Furthermore we believe our theory to be well suited to study training dynamics, as of now fully understood for the linear case [13], in the general case.

The power of our unified approach becomes particularly evident when applied to modern architectures that have resisted traditional analysis. We provide the first scaling limit results for neural networks defined as controlled differential equations [88, 89], establishing convergence theorems for Neural Controlled Differential Equations and deriving scaling limits for State Space Models, including recent architectures such as linear transformers and Mamba models [44, 63]. Through our combinatorial lens, we precisely characterize the expressive power of these modern sequence-to-sequence models [89], revealing fundamental connections between their architectural choices and their limiting behaviors.

Our theoretical analysis reveals the existence of previously unknown mathematical structures: a new class of signature kernels on path space with enhanced expressive properties compared to existing methods. The discovery of these kernels represents

not merely a theoretical curiosity but a genuine advance in our mathematical toolkit for analyzing sequential data.

0.3 Thesis Outline

The thesis proceeds systematically from theoretical foundations to practical applications. Here we present an outline of the contents of each chapter.

In chapter 1, we establish the mathematical groundwork by presenting genus expansion techniques and demonstrating the connection between graphs and analytic operators. The main character of this chapter, and arguably of the whole work, are *operator graphs*: these are graphs where each vertex is associated to a vector, each edge to a matrix, and whose topological structure defines a scalar value as shown in fig. 1.

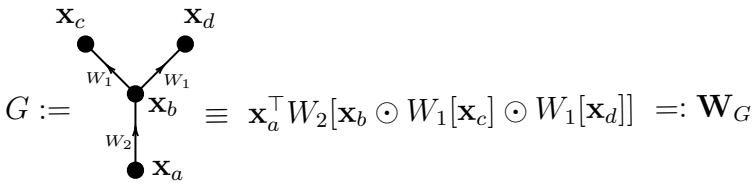


Figure 1: *An operator graph G and its corresponding scalar value $\mathbf{W}_G$ ($\odot$ stands for componentwise multiplication).*

Throughout the chapter, we will establish both a remarkable correspondence between graphical manipulations of such graphs and analytic operations on their corresponding scalar values, and the elegant *genus expansion* which relates the expectations of random graphs with "random edges" to deterministic graphs obtained through a grapho-combinatorial procedure taking the simple form

$$\mathbb{E}[\mathbf{W}_G] = \sum_{\phi} \mathbf{W}_{G_\phi}.$$

This differs from, while generalising, similar results in the literature [28] in that the objects in the right and left hand sides are of the same type and can all be analitycally manipulated with graphical operations, providing a new method for studying nonlinear random matrix theory that underlies all subsequent developments.

Following, chapter 2 applies these techniques to construct Taylor expansions of neural networks

$$\mathcal{NN}(\mathbf{x}) = \sum_{G \in \mathcal{F}} c(G)\mathbf{W}_G \tag{1}$$

not indexed on natural numbers, but in terms of decorated graphs which, in this case, are all elements of a family of "nicely-structured" trees.

We then employ our framework by graphically manipulating these trees to obtain dot-products and differentials, and subsequently applying the genus expansion to com-

pute expectations in a combinatorial fashion. Through this approach, we recover the Gaussian Process and Neural Tangent Kernel limits and compute Jacobian moments, demonstrating both the consistency of our method with existing results and its capacity to enable new calculations previously beyond reach. In chapter E we moreover present how to derive the results of Chizat et al. [13] (studying the μP dynamics of *linear* networks) within our framework, commenting on how to extend to the general non-linear case.

We then proceed to extend the framework to State Space Models in chapter 3, identifying these *state of the art* [44] models, whose core mechanism we recognize as being a linear controlled differential equation of type

$$dZ_t^x = \sum_i A_i Z_t^x \, dx_t^i$$

with series (unsurprisingly) indexed over graphs, simple ladders in this case, whose terms are the corresponding values weighted by *signature* entries:

$$Z_t^x = \sum_{\tau \in \mathbb{L}} \mathbf{W}_\tau \mathrm{Sig}(x)_{0,t}^\tau \in \mathbb{R}^N. \tag{2}$$

As in the feed-forward case studied in the preceeding chapter, using our methods we characterise their limiting kernels and using these their expressive power, providing as a result the first rigorous analysis of these modern sequence-to-sequence architectures.

Two chapters are then devoted to the extension of results to the more general Neural Controlled Differential Equation case for which the driving vector fields become feed-forward networks instead of just linear maps:

$$A_i Z_t \mapsto \mathcal{NN}_i(Z_t)$$

In chapter 4 (historically preceding the graph expansions) we consider the shallow case in which $\mathcal{NN}_i(z) = A_i \varphi(z)$ using the formalism of *tensor programs* [131] and uncover as a result a new class of path-valued kernels, *neural signature kernels*, extending the signature kernel of Salvi et al. [112] (which corresponds to the case $\varphi = id$) and likewise solving the partial differential equation

$$\mathcal{K}_\varphi^{x,y}(1,1) = \sum_{i=1}^d \int_0^1 \int_0^1 \left(\sigma_A^2 V_\varphi(\Sigma_\varphi^{x,y}(s,t)) + \sigma_b^2 \right) \mathrm{d}x_s^i \mathrm{d}y_t^i$$

with border conditions given by $\mathcal{K}_\varphi^{x,y}(0,0) = \mathcal{K}_\varphi^{x,y}(s,0) = \mathcal{K}_\varphi^{x,y}(0,t) = 1$, and where

$$V_\varphi(\Sigma) = \mathbb{E}_{\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \Sigma)}[\varphi(z_1)\varphi(z_2)], \quad \Sigma_\varphi^{x,y}(s,t) = \begin{pmatrix} \mathcal{K}_\varphi^{x,x}(s,s) & \mathcal{K}_\varphi^{x,y}(s,t) \\ \mathcal{K}_\varphi^{x,y}(s,t) & \mathcal{K}_\varphi^{y,y}(t,t) \end{pmatrix}.$$

In chapter D we *formally* use the graphical techniques to extend results of the literature to the general case of the networks $\mathcal{NN}_i$ having arbitrary depth, thus for the first time computing the respective GP and NTK limits which, albeit being formal derivations, nonetheless generalise naturally the corresponding results in simpler cases.

Author's note: Further acknowledgements

The principal part of this thesis emerged from several published collaborations, and I am grateful to the colleagues who contributed to the ideas and experiments reported here. The Genus Expansion was developed in its full form with the help of Jad Hamdan, particularly with regards to the study of the Jacobian moments and connections to Random Matrix theory. For the material on State Space Models I'm indebted to Ben Walker and Maud Lemercier for their experimental contributions, and to Antonio Orvieto for his exceptional knowledge of the relevant literature. I also wish to thank my supervisor, Cristopher Salvi, for his guidance on the conception of the presented works.

Where contributions were joint, they are acknowledged explicitly in the corresponding chapters by citing the relevant works; any remaining errors are my own.

Chapter 1

Genus Expansion

Both this chapter and chapter 2 closely follow Cirone et al. [16], a joint work with *Jad Hamdan* and *Cristopher Salvi*.

1.1 Introduction

The *genus expansion* technique, the main charachter of the present chapter, has its roots in the connection between matrix integrals and the enumeration of (combinatorial) maps, which was first discovered in the context of quantum field theory (see [7, 34, 54], as well as [137] for an accessible introduction to the subject).

The link to random matrix theory was later established by Harer and Zagier [51] in a seminal work investigating moduli spaces of curves, and has since been used to study various matrix ensembles and their asymptotic first and second-order freeness (e.g. [85, 93, 107, 108], as well as [28] and the references therein). Roughly speaking, the approach consists of expanding traces of random matrix products and evaluating their expectation using Wick's principle. This yields a sum whose terms are in bijection with a set of graphs, and one determines which terms are of leading order by embedding their corresponding graphs into surfaces and computing their Euler characteristic.

In a recent work, Dubach and Peled [28] used this approach to study products of non-Hermitian random matrices, and in particular showed that it can be adapted to non-Gaussian and sparse matrices with relative ease. This would, at least in principle, make it ideal for the study of multilayer neural networks, but the presence of non-linear activations in that setting prohibits one from being able to apply Wick's formula.

We circumvent this problem by developing a graphical language to express a large class of matrix/vector products, deriving with it, in chapter 2, an expansion for neural networks in terms of this language. This linearises the effect of the activation functions, allowing for the use of Wick's formula and the connection to the enumeration of maps to be made. To the best of our knowledge, this is the first use of the genus expansion technique in deep learning and non-linear RMT.

1.1.1 Overview of our method.

A graphical language for neural network computations.

The idea of using a graphical language to simplify computations involving multilinear maps is not new, dating back to at least the 1970s with the introduction of Penrose diagrams [104], which have more recently been applied in the context of machine learning (see [5, 15]). Graphs have also been used to evaluate expectations of products of Gaussian variables, and this forms the basis of the genus expansion technique.

The graphs that we introduce are novel and accomplish both of these tasks at once. On the one hand, they can be used to express deterministic products and operations involving multilinear maps. On the other, when dealing with tensors with Gaussian entries, the expectation of these operations can once again be expressed in terms of graphs (in the sense of eq. (1.1)). We explain this briefly below, deferring to section 1.2 for more details.

In what we call a *product* graph $G = (V, E)$, edges will correspond to matrices and vertices to vectors, which we call the inputs of their respective edge/vertex. The graph's structure then dictates a well-defined product involving these inputs, the result of which we call the *value* of the graph and denote by $\mathbf{W}_G$. For instance, a path of length k can be used to express an (ordinary) product of k matrices, while trees can be used to express Hadamard (entrywise) products (this is depicted in figs. 1.1 and 1.2).

If we omit inputs for some vertices and edges of the graph and view them as variables, then the resulting graph can be associated to a (multi)linear map and we call it an *operator graph*. Differentiation, composition and other operations involving these maps then turn out to be expressible using simple manipulations of their corresponding graphs (composition, for instance, reduces to attaching graphs by a vertex), as explained in section 1.2.1 and the figures therein.

Remark. *After completing this work, we were made aware of the theory of traffic probability [79], with which the graphical formalism developed here overlaps significantly. We leave the task of reconciling both theories to future work, but point out similarities between the two across section 1.2.*

Wick's principle and genus expansion.

When G is a product graph whose edge inputs have Gaussian entries, our main tool to evaluate $\mathbb{E}\mathbf{W}_G$ is Wick's principle, which reduces the expectation of products of Gaussian variables to the sum of their pairwise covariances. Applied to $\mathbf{W}_G$, it yields the following simple identity

$$\mathbb{E}\mathbf{W}_G = \sum_{\phi} \mathbf{W}_{G_\phi}, \tag{1.1}$$

(see theorem 1.15), where the sum is taken over *admissible* pairings ϕ of the edges of G (see theorem 1.14), and G_ϕ is the graph obtained from G after identifying edges paired by ϕ (meaning that we consider such edges to be the same edge in G_ϕ). Under additional assumptions on G (c.f. theorem 1.19), we find that $\mathbf{W}_{G_\phi} = \sigma_G N^{|V(G_\phi)|}$ for every ϕ, where $|V(G_\phi)|$ is the number of vertices in G_ϕ and σ_G is a variance parameter. The leading order of $\mathbf{W}_G$ is thus determined by the pairings for which $|V(G_\phi)|$ is maximised.

Instead of counting this quantity directly, it turns out to be much simpler to embed the graph onto a surface S_ϕ (as defined in eq. (1.12)) and to then compute $|V(G_\phi)|$ using the *Euler characteristic formula*

$$|V(G_\phi)| - |E(G_\phi)| + f(G_\phi : S_\phi) = 2 - 2g(S_\phi),$$

where $|E(G_\phi)|$ is the number of edges of G_ϕ, $f(G_\phi : S_\phi)$ the number of faces of G_ϕ in S_ϕ and $g(S_\phi)$ the genus of S_ϕ. This formula allows us to identify which ϕ give rise to leading and sub-leading order terms in eq. (1.1), which we call *atomic* and *bi-atomic* pairings, respectively, following [28]. We use this to give a more explicit version of equation (1.1), and to extend it to centred mixed moments $\mathbb{E}\{\prod_G(\mathbf{W}_G - \mathbb{E}\mathbf{W}_G)\}$ as well (see theorem 1.17). Lastly, we combine these results to obtain a limit theorem for the joint moments of product graphs (theorem 1.22), reminiscent of a celebrated result of Diaconis and Shahshahani [25] for traces of powers of random unitary matrices and its recent extension generalisation in Dubach and Peled [28][Theorem 1.2].

When the edge inputs in G are complex, non-Gaussian or sparse matrices (or any combination of the three), we show that the expressions for $\mathbb{E}\mathbf{W}_G$ and $\mathbb{E}\{\prod_G(\mathbf{W}_G - \mathbb{E}\mathbf{W}_G)\}$ hold up to an $o(1)$ error term (see sections A.1 to A.3, respectively).

1.1.2 Notation and nomenclature

Given a matrix $A \in \mathbb{R}^{N\times M}$ and a vector $v \in \mathbb{R}^N$ we write $[A]_{ij}$ and $[v]_j$ for their i,j and j-th coordinate, respectively. More generally, we will use square brackets with subscripts to denote coordinates of tensors. We use the notation $\mathbf{I}_{M,N} \in \mathbb{R}^{M\times N}, \mathbf{1}_N \in \mathbb{R}^N$ to denote the matrix and vector having all entries equal to 1 (omitting the subscripts whenever it does not hurt comprehension, and $\mathbf{I}_M = \mathbf{I}_{M,M}$), $\mathbf{E}_{ij} \in \mathbb{R}^{M\times N}$ and $\mathbf{e}_i \in \mathbb{R}^N$ to denote the canonical basis matrices/vectors in $\mathbb{R}^N$, and Id_N to denote the identity matrix in $\mathbb{R}^{N\times N}$. $\langle , \rangle$ will denote the standard inner product, with the space in subscript when it is not clear from the context. If A is a matrix with complex entries, we use $\bar{A}$ to denote its conjugate and A^* its Hermitian transpose. $\mathcal{N}(\mu, \sigma^2)$ will denote a Gaussian with mean μ and variance σ^2, and similarly $\mathcal{N}_{\mathbb{C}}(0,1)$ will denote a standard complex Gaussian (with real and imaginary parts each having variance $1/2$). We will use standard asymptotic notation, writing $f(T) = o(g(T))$ to mean that $|f(T)/g(T)| \to_{T\to\infty} 0$

and $f(T) = \mathcal{O}(g(T))$ to mean that $\limsup_{T\to\infty} |f(T)/g(T)|$ is bounded. For any positive integer N, we will use $[N]$ to denote the set $\{1, \dots, N\}$. Whenever G is a directed graph, we let $V(G), E(G)$ denote its set of vertices and edges respectively, and omit the dependence on G whenever the graph is clear from the context. For any edge $e = (u, v)$ we call u the *head* and v the *tail* of e, and say that e is *adjacent* to u, v and vice versa.

Symbol	Meaning
$\mathfrak{d}$	Dimension function associated to cells of a graph G (Sec. 1.2).
$\mathbf{X}_c \in \mathbb{R}^{\mathfrak{d}(c)}$	Input of cell c in a graph G (Sec. 1.2).
$\mathfrak{C}$	Set of cell inputs of a product graph (Def. 1.1).
$\mathcal{F} = (\mathcal{F}_{\text{in}}, \mathcal{F}_{\text{out}})$	Free cells (Sec. 1.2.1) of an operator graph, comprised of a sequence of in-cells and a sequence of out-cells.
$\mathbf{W}_G$	Operator associated to a graph G (Def. 1.3), noting that if G is a product graph, $\mathbf{W}_G$ is its value.
$\bullet_{\mathbf{X}_v}$, $\xrightarrow{\mathbf{X}_e}$	Cell fixed to $\mathbf{X}_c$ (Sec. 1.2.1).
$\bullet$, $\to$	Cell fixed to $\mathbf{1}$ or $\mathbf{I}$ (Sec. 1.2.1).
$\circ$, $\dashrightarrow$	In-cell (Sec. 1.2.1).
$\circ$, $\dashrightarrow$	Out-cell (Sec. 1.2.1).
$\bullet_{\mathbf{x}} \mapsto \bullet_{\mathbf{y}}$	Switching the input of a cell (Def. 2.5).
$G_1 \wedge G_2$	Out-vertex identification (corresponds to a Hadamard product) (Def. 1.5).
$s(\cdot)$	Symmetry factor of trees (Sec. 2.2).
$\mathcal{P}(G)$	Admissible pairings of edges of G with random inputs (Def. 1.14).
$\mathcal{P}_{\text{A}}(G), \mathcal{P}_{\text{B}}(G), \mathcal{P}_{\text{AF}}(G)$	Atomic, bi-atomic and atom-free pairings of G (Def. 1.16).
G_ϕ	Graph resulting from identification of G's edges according to ϕ (Sec. 1.3).

Table 1.1: *Notation and symbols used throughout this chapter.*

1.2 Graphical Descriptions of Analytic Operations

In this section we explain how various matrix-vector products can be expressed by means of directed graphs.

Consider a directed graph $G = (V, E)$. We associate a vector $\mathbf{X}_v \in \mathbb{R}^{\mathfrak{d}(v)}$ to each vertex $v \in V$ and a matrix $\mathbf{X}_e \in \mathbb{R}^{\mathfrak{d}(u)\times\mathfrak{d}(v)}$ to each edge $e = (u, v) \in E$, calling $\mathbf{X}_e$ and $\mathbf{X}_v$ the *inputs* of the edge e and vertex v respectively. Here, $\mathfrak{d}(v)$ is a positive integer which we call the *dimension* of the vertex v, and we extend $\mathfrak{d}$ to edges by using the shorthand $\mathfrak{d}(e) = (\mathfrak{d}(u), \mathfrak{d}(v))$ for $e = (u, v) \in E$ (in which case $\mathbb{R}^{\mathfrak{d}(e)} := \mathbb{R}^{\mathfrak{d}(u)\times\mathfrak{d}(v)}$).

As we are often going to talk about vertices and edges simultaneously, it will be useful to have a single term to refer to both types of objects. For this purpose we will use the term *cells*. Every cell in $C := V \cup E$ will thus have a dimension assigned to it by $\mathfrak{d}$, and this map is entirely determined by the values that it takes on vertices by definition.

As we will soon make precise, G can be seen as describing a type of product of its inputs. This motivates the naming in the following definition, which summarizes what we have introduced so far.

Definition 1.1 (Product graph)**.** *A* product graph *is a triple* $(G, \mathfrak{d}, \mathfrak{C})$*, where*

- $G = (V, E)$ *is a directed graph with vertex set* V *and edge set* $E \subseteq V \times V$*.*
- $\mathfrak{d}$ *assigns a dimension to each vertex in* V*, and thus to each cell of* G*.*
- $\mathfrak{C} = \{\mathbf{X}_c \in \mathbb{R}^{\mathfrak{d}(c)} : c \in C = V \cup E\}$ *is a collection of inputs (vectors/matrices) associated to the cells* C *of* G*.*

When $\mathfrak{d}$, $\mathfrak{C}$ are clear or implied from the context, we will omit them and simply use G to denote the product graph $(G, \mathfrak{d}, \mathfrak{C})$. To a product graph G we can uniquely associate the following *value*.

Definition 1.2. *The value* $\mathbf{W}_G \in \mathbb{R}$ *of a product graph* $(G, \mathfrak{d}, \mathfrak{C})$ *is defined as the following scalar*

$$\mathbf{W}_G := \sum_{(i_v)_{v\in V}} \prod_{(u,v)\in E} [\mathbf{X}_{(u,v)}]_{i_u i_v} \prod_{w\in V} [\mathbf{X}_w]_{i_w} = \sum_{\mathbf{i}_V} \prod_{c\in C} [\mathbf{X}_c]_{i_c} \in \mathbb{R} \tag{1.2}$$

where the sum on the right hand side is taken over all indexations $\mathbf{i}_V = (i_v)_{v\in V} \in \prod_{v\in V}\{1, \dots, \mathfrak{d}(v)\}$ *and where for* $c = (u, v) \in E$ *we define* $i_c := i_u i_v$*.*

Remark. *The constraints imposed by* $\mathfrak{d}$ *on the dimensions of the elements of* $\mathfrak{C}$ *in theorem 1.1 ensure that product (1.2) is well-defined. Furthermore, if* G *is disconnected with connected components* $\{G_i\}_i$*, then* $\mathbf{W}_G = \prod_i \mathbf{W}_{G_i}$*. In the event that some of its inputs are random,* $\mathbf{W}_G$ *then becomes a random variable.*

Remark. *Note how theorem 1.2 inherently depends on the bases relative to which the coordinates of the* $\mathbf{X}_c$ *are being taken. This will always be the canonical basis. In general, taking* $A_v \in \mathrm{GL}(\mathbb{R}^{\mathfrak{d}(v)})$ *for each* $v \in V$ *and substituting*

$$\mathbf{X}_v \mapsto A_v^{-1}\mathbf{X}_v, \forall v \in V. \quad \mathbf{X}_{(u,v)} \mapsto A_u^{-1}\mathbf{X}_{(u,v)}A_v, \forall (u, v) \in E, \tag{1.3}$$

can change the value of $\mathbf{W}_G$*. On the other hand, the latter is invariant to (1.3) if each* A_v *is a permutation matrix: this mimics Lemma 1.4 in [79], with their *-graph monomials corresponding to (1.2) without vertex inputs.*

We illustrate this definition with some examples.

Example. *If* G *is the product graph in fig. 1.1, then* $\mathbf{W}_G = \mathbf{x}_a^\top W_3 W_2 W_1 \mathbf{x}_d$ *since*

$$\mathbf{x}_a^\top W_3 W_2 W_1 \mathbf{x}_d = \sum_{\alpha=1}^{\mathfrak{d}(a)} \sum_{\beta=1}^{\mathfrak{d}(b)} \sum_{\gamma=1}^{\mathfrak{d}(c)} \sum_{\delta=1}^{\mathfrak{d}(d)} [W_3]_{\alpha,\beta}[W_2]_{\beta,\gamma}[W_1]_{\gamma,\delta}[\mathbf{x}_a]_\alpha[\mathbf{x}_d]_\delta.$$

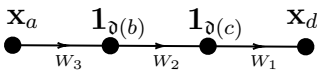


Figure 1.1: *A product graph giving rise to a word of matrices multiplied by vectors on either side.* $\mathbf{1}_N$ *is the* $N \times 1$ *vector of ones.*

Example. *If* G *is a tree, then the corresponding value* $\mathbf{W}_G$ *is defined by means of Hadamard products of matrices. For instance, taking* G *to be the product graph in fig. 1.2 gives*

$$\mathbf{W}_G = \mathbf{x}_a^\top W_2\Big(\mathbf{x}_b \odot (W_1\mathbf{x}_d)\odot(W_1\mathbf{x}_c)\Big)$$

where $A\odot B$ *denotes the Hadamard product of* A *and* B*, since*

$$\mathbf{W}_G := \sum_{\alpha=1}^{\mathfrak{d}(a)}\sum_{\beta=1}^{\mathfrak{d}(b)}\sum_{\gamma=1}^{\mathfrak{d}(c)}\sum_{\delta=1}^{\mathfrak{d}(d)}[W_2]_{\alpha,\beta}[W_1]_{\beta,\gamma}[W_1]_{\beta,\delta}[\mathbf{x}_a]_\alpha[\mathbf{x}_b]_\beta[\mathbf{x}_c]_\gamma[\mathbf{x}_d]_\delta.$$

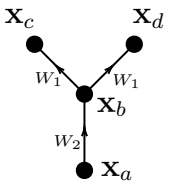


Figure 1.2: *Trees give rise to Hadamard products.*

Note that while the previous two examples can be described in terms of ordinary matrix/vector multiplication and entry-wise products, it isn't necessarily so straightforward in general. For instance, taking G to be the product graph in fig. 1.3 gives

$$z\mathbf{W}_G = \mathrm{Tr}\left\{\left[\mathbf{x}_a\mathbf{1}^\top \odot W_1\right]\left[(\mathbf{x}_b \odot W_3\mathbf{x}_c)\mathbf{1}^\top \odot W_2\right]\right\}$$

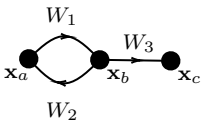


Figure 1.3: *A simple product graph leading to a more involved analytical expression for its value.*

1.2.1 Operator graphs and their associated linear map

So far, we have associated a vector/matrix to each cell in our graphs. By *freeing* some of these cells-meaning that we consider their inputs as variables-we can use graphs to define more general linear maps between tensors.

Consider for example the product graph $(G, \mathfrak{d}, \mathfrak{C})$ in fig. 1.2, noting how the product defining $\mathbf{W}_G$ is linear in $\mathbf{x}_a$. If we consider $\mathbf{x}_a$ to be a variable, this yields a well-defined linear map $\mathbb{R}^{\mathfrak{d}(a)} \to \mathbb{R}$, or equivalently a vector in $\mathbb{R}^{\mathfrak{d}(a)}$. In this particular case, the linear map in question would be $\mathbf{x} \mapsto \mathbf{x}^\top W_2\Big(\mathbf{x}_b \odot (W_1\mathbf{x}_d) \odot (W_1\mathbf{x}_c)\Big)$, and its vector representation is $W_2\Big(\mathbf{x}_b \odot (W_1\mathbf{x}_d) \odot (W_1\mathbf{x}_c)\Big)$.

More generally, let G be a directed graph with dimensions assigned by $\mathfrak{d}$, and let $\mathcal{F}$ be an (ordered) sequence of free cells of G (meaning vertices and edges which are considered as variables), assuming that the remaining cells c which do not belong to $\mathcal{F}$ are each *fixed* to some input $\mathbf{X}_c$. Then for any input sequence $(\mathbf{X}_c)_{c\in\mathcal{F}}$, the product defining $\mathbf{W}_{(G,\mathfrak{d},\{\mathbf{X}_c\}_{c\in C})} = \mathbf{W}_G$ is $|\mathcal{F}|-$linear in these inputs.

Any map defined this way will be scalar-valued, but we can generalize to vector-valued maps through partial evaluation. Namely, given a partition of $\mathcal{F}$ into two subsequences $\mathcal{F}_{\text{in}}$ and $\mathcal{F}_{\text{out}}$, we can consider the multilinear map

$$(\mathbf{X}_c)_{c\in\mathcal{F}_{\text{in}}} \mapsto \Big(\otimes_{c\in\mathcal{F}_{\text{out}}} \mathbf{X}_c \mapsto \mathbf{W}_G \Big). \tag{1.4}$$

This is equivalent to considering the map

$$(\mathbf{X}_c)_{c\in\mathcal{F}_{\text{in}}} \mapsto \mathbf{Y}, \quad \langle \mathbf{Y}, \otimes_{c\in\mathcal{F}_{\text{out}}} \mathbf{X}_c \rangle = \mathbf{W}_G \text{ for any } (\mathbf{X}_c)_{c\in\mathcal{F}_{\text{out}}} \tag{1.5}$$

where $\mathbf{Y} \in \bigotimes_{c\in\mathcal{F}_{\text{out}}} \mathbb{R}^{\mathfrak{d}(c)}$ is the Riesz representation of $\Big(\otimes_{c\in\mathcal{F}_{\text{out}}} \mathbf{X}_c \mapsto \mathbf{W}_G \Big)$. Note that the input and output dimensions of this map depend on the partition of $\mathcal{F}$ that is being taken. We refer to it as the *operator associated to* G, which we define below. In what follows, we abuse notation slightly and write $c \notin \mathcal{F}$ to mean $c \in C \setminus (\mathcal{F}_{\text{in}} \sqcup \mathcal{F}_{\text{out}})$.

Definition 1.3 (Operator associated to a graph)**.** *Let G and $\mathfrak{d}$ be as in theorem 1.1. Let $\mathcal{F} = (\mathcal{F}_{\text{in}}, \mathcal{F}_{\text{out}})$ be a sequence of free cells of C. Then:*

- *The operator associated to $(G, \mathfrak{d}, \mathcal{F}, \{\mathbf{X}_c\}_{c\notin\mathcal{F}})$ is defined as*

$$\mathbf{W}_G : \prod_{c\in\mathcal{F}_{\text{in}}} \mathbb{R}^{\mathfrak{d}(c)} \to \bigotimes_{c\in\mathcal{F}_{\text{out}}} \mathbb{R}^{\mathfrak{d}(c)}, \quad \mathbf{W}_G\Big((\mathbf{X}_c)_{c\in\mathcal{F}_{\text{in}}}\Big) = \mathbf{Y}$$

  where $\mathbf{Y} \in \bigotimes_{c\in\mathcal{F}_{\text{out}}} \mathbb{R}^{\mathfrak{d}(c)}$ is the unique vector satisfying

$$\langle \mathbf{Y}, \otimes_{c\in\mathcal{F}_{\text{out}}} \mathbf{X}_c \rangle = \mathbf{W}_{(G,\mathfrak{d},\{\mathbf{X}_c\}_{c\in C})},$$

  namely the Riesz representation of $\otimes_{c\in\mathcal{F}_{\text{out}}} \mathbf{X}_c \mapsto \mathbf{W}_{(G,\mathfrak{d},\{\mathbf{X}_c\}_{c\in C})}$.

Note that if $\mathcal{F}_{\text{in}}$ is empty but $\mathcal{F}_{\text{out}}$ isn't, then $\mathbf{W}_G$ is a constant map from $\{0\}$ to $\bigotimes_{c\in\mathcal{F}_{\text{out}}} \mathbb{R}^{\mathfrak{d}(c)}$ which we then identify with its value (see the top right example in fig. 1.4).

Just as we did for product graphs, we will abuse notation and just use G to denote an operator graph when $\mathcal{F}, \{\mathbf{X}_c\}_{c\notin\mathcal{F}}$ and $\mathfrak{d}$ are clear from the context. As a result, $\mathbf{W}_G$ can denote both an operator (when G is an operator graph) and a scalar (when G is a product graph).

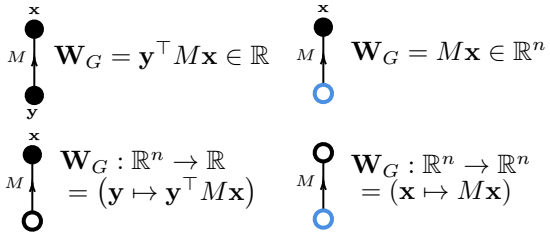


Figure 1.4: *A simple product graph G (top left), from which we can free vertices to make $\mathbf{W}_G$ either a vector, a scalar-valued map, or a vector-valued map (top right, bottom left, and bottom right respectively).*

Some examples to illustrate these definitions might be in order, and are given in fig. 1.4 and in fig. 1.5. When depicting a graph, a free vertex is drawn as a circle (e.g. $\circ$) and is depicted in blue if it is in the set of outputs $\mathcal{F}_{\text{out}}$, while a vertex fixed to $\mathbf{x}$ is drawn as a black dot labeled by said vector (e.g. $\bullet_{\mathbf{x}}$). Similarly, we draw free edges with dotted directed lines (e.g. $\dashrightarrow$) and in blue if they are in the set $\mathcal{F}_{\text{out}}$ of output cells, while fixed edges are drawn as solid lines labeled by their input matrix (e.g. $\xrightarrow{W}$). For simplicity, we will omit labels from vertices (resp. edges) whose input is $\mathbf{1}$ (resp. $\mathbf{I}$) when drawing them. We will also not indicate dimensions given by $\mathfrak{d}$ nor the order of the sequence $\mathcal{F}$ on the graph itself.

Example. *Let G be the graph in fig. 1.5 (left). To turn this into an operator graph, let a, b, c, d refer to its vertices from bottom to top/ left to right and let $\mathfrak{d} : \{a, b, c, d\} \to \mathbb{N}$ be any fixed dimension map. Then if we consider $\mathcal{F}_{\text{in}} = (d, (a, b))$, $\mathcal{F}_{\text{out}} = (a)$ and fix the remaining cells as*

$$b \mapsto \mathbf{x}_b,\ \ c \mapsto \mathbf{1},\ \ (b, d) \mapsto \mathbf{I},\ \ (b, c) \mapsto W_1$$

(c.f. fig. 1.5), then $\mathbf{W}_G$ is the map $(\mathbf{x}_d, W_{(a,b)}) \mapsto W_{(a,b)}\Big(\mathbf{x}_b \odot (W_1\mathbf{1}) \odot (\mathbf{I}\ \mathbf{x}_d)\Big)$.

1.2.2 Operations on graphs.

Having outlined the correspondence between graphs and operators, we can now go further and show how various analytic operations can be expressed as binary operations on graphs.

Proposition 1.4 (Evaluation)**.** *Let $(G, \mathfrak{d}, (\mathcal{F}_{\text{in}}, \mathcal{F}_{\text{out}}), \{\mathbf{X}_c\}_{c\notin\mathcal{F}})$ be an operator graph, $\mathbf{W}_G$ its associated map and $c_0 \in \mathcal{F}_{\text{in}}$ be a free cell of G. Let $G[\mathbf{X}_{c_0}]$ denote the graph*

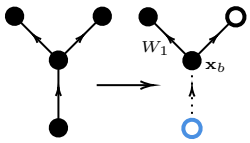


Figure 1.5: *Example of an operator graph (right) obtained from a graph (left) by a choice of $\mathcal{F}_{\text{in}}$, $\mathcal{F}_{\text{out}}$, $\mathfrak{d}$ and $(\mathbf{X}_c)_{c\notin\mathcal{F}}$. Note how on the right, we omit labels from vertices and edges whose inputs are $\mathbf{1}$ or $\mathbf{I}$ respectively.*

obtained by fixing c_0's input to some $\mathbf{X}_{c_0} \in \mathbb{R}^{\mathfrak{d}(c_0)}$. Then if $|\mathcal{F}_{\text{in}}| = 1$, we have $\mathbf{W}_{G[\mathbf{X}_{c_0}]} = \mathbf{W}_G(\mathbf{X}_{c_0})$. Otherwise,

$$\mathbf{W}_{G[\mathbf{X}_{c_0}]} : \prod_{c\in\mathcal{F}_{\text{in}}\setminus c_0} \mathbb{R}^{\mathfrak{d}(c)} \to \bigotimes_{c\in\mathcal{F}_{\text{out}}} \mathbb{R}^{\mathfrak{d}(c)}$$

$$\mathbf{W}_{G[\mathbf{X}_{c_0}]}\Big((\mathbf{X}_c)_{c\in\mathcal{F}_{\text{in}}\setminus c_0}\Big) = \mathbf{W}_G\Big((\mathbf{X}_c)_{c\in\mathcal{F}_{\text{in}}}\Big).$$

In particular,

$$\mathbf{W}_{G[(\mathbf{X}_c)_{c\in\mathcal{F}_{\text{in}}}]} = \mathbf{W}_G\Big((\mathbf{X}_c)_{c\in\mathcal{F}_{\text{in}}}\Big).$$

Proof. This is an immediate consequence of theorem 1.3. □

For example, let G be the operator graph on the left in fig. 1.6 below and denote its top-right vertex by v. Note that $\mathcal{F}_{\text{in}} = (v)$ and $\mathcal{F}_{\text{out}} = ()$. Ignoring dimensions for the sake of this example,

$$\mathbf{W}_G : \mathbf{x} \mapsto W_2\Big((W_1\mathbf{1}) \odot (W_1\mathbf{x})\Big),$$

and fixing the unique in-vertex's input to $\mathbf{X}_v$ (c.f. fig. 1.6, right) gives $W_2((W_1\mathbf{1}) \odot (W_1\mathbf{X}_v))$.

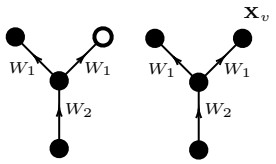


Figure 1.6: *Fixing a free in-vertex (left) with the vector $\mathbf{x}_v$ (right).*

Recall that by definition, an operator graph $(G, \mathfrak{d}, ((), \mathcal{F}_{\text{out}}), \{\mathbf{X}_c\}_{c\notin\mathcal{F}})$ gives rise to $\mathbf{W}_G \in \bigotimes_{c\in\mathcal{F}_{\text{out}}} \mathbb{R}^{\mathfrak{d}(c)}$ (see fig. 1.4). For two such graphs G_1 and G_2, it therefore makes sense to talk about the Hadamard (entrywise) product of $\mathbf{W}_{G_1}$ and $\mathbf{W}_{G_2}$.

Proposition 1.5 (Hadamard product)**.** *Let*

$$G_1 = (G_1, \mathfrak{d}_1, ((), (v_1)), \{\mathbf{X}_c\}_{c\in C_1\setminus\{v_1\}}),\ G_2 = (G_2, \mathfrak{d}_2, ((), (v_2)), \{\mathbf{X}_c\}_{c\in C_2\setminus\{v_2\}})$$

be operator graphs with unique free out-vertices $v_1 \in V(G_1)$ and $v_2 \in V(G_2)$ respectively. Assume that $\mathfrak{d}_1(v_1) = \mathfrak{d}_2(v_2) = d > 0$.

Define $G_1 \wedge G_2$ to be the operator graph obtained by identifying v_1 and v_2 (meaning that we take the union of the two graphs and consider $v_1 = v_2$ to be the same vertex). Then

$$\mathbf{W}_{G_1 \wedge G_2} = \mathbf{W}_{G_1} \odot \mathbf{W}_{G_2} \in \mathbb{R}^d,$$

where $\odot$ denotes the Hadamard product.

Proof. Let $v_1 \wedge v_2$ be the unified vertex in $G_1 \wedge G_2$, and let $V_i^- = V(G_i) \setminus \{v_i\}$ for $i \in \{1, 2\}$. Then for any $\mathbf{X} \in \mathbb{R}^d$, theorem 1.4 yields $\langle \mathbf{W}_{G_1 \wedge G_2}, \mathbf{X} \rangle = \mathbf{W}_{(G_1 \wedge G_2)[\mathbf{X}]}$, which by definition equals

$$\begin{aligned}
&\sum_{\mathbf{i}_{V(G_1)}, \mathbf{i}_{V(G_2)}} \delta_{i_{v_1} i_{v_2}} \left(\prod_{c_1 \in C_1 \setminus \{v_1\}} [\mathbf{X}_{c_1}]_{i_{c_1}} \right) \left(\prod_{c_2 \in C_2 \setminus \{v_2\}} [\mathbf{X}_{c_2}]_{i_{c_2}} \right) [\mathbf{X}]_{i_{v_1}} \\
&= \sum_{i_{v_1}, i_{v_2} = 1}^{d} \delta_{i_{v_1} i_{v_2}} \left(\sum_{\mathbf{i}_{V_1^-}} \prod_{c_1 \in C_1 \setminus \{v_1\}} [\mathbf{X}_{c_1}]_{i_{c_1}} \right) \left(\sum_{\mathbf{i}_{V_2^-}} \prod_{c_2 \in C_2 \setminus \{v_2\}} [\mathbf{X}_{c_2}]_{i_{c_2}} \right) [\mathbf{X}]_{i_{v_1}} \\
&= \sum_{k=1}^{d} [\mathbf{W}_{G_1}]_k [\mathbf{W}_{G_2}]_k [\mathbf{X}]_k = \sum_{k=1}^{d} [\mathbf{W}_{G_1} \odot \mathbf{W}_{G_2}]_k [\mathbf{X}]_k \\
&= \langle \mathbf{W}_{G_1} \odot \mathbf{W}_{G_2}, \mathbf{X} \rangle,
\end{aligned}$$

where the first sum is over indexations $\mathbf{i}_{V(G_j)} \in \prod_{v \in V(G_j)} [\mathfrak{d}(v)]$, $j \in \{1, 2\}$ and δ_{ij} is the Kronecker delta. The proposition then follows by uniqueness of the Riesz representation. □

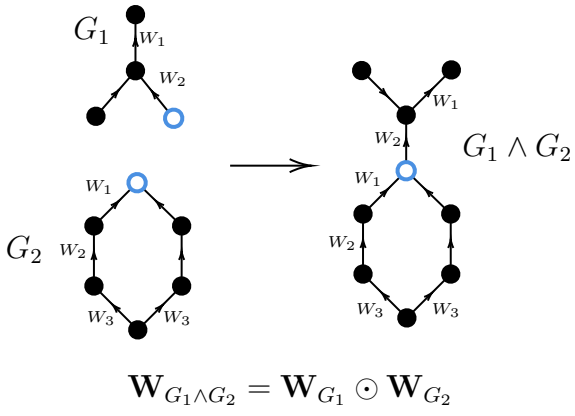


Figure 1.7: *Identifying two graphs by their unique free (out-)vertex gives rise to a Hadamard product.*

Note that the graph operation $\wedge$ is associative and symmetric much like its analytic counterpart $\odot$, meaning that there is no confusion in writing $G_1 \wedge \cdots \wedge G_n = \bigwedge_{i \in [n]} G_i$.

Proposition 1.6 (Transpose)**.** *Let* $G = (G, \mathfrak{d}, (\mathcal{F}_{\text{in}}, \mathcal{F}_{\text{out}}, \{\mathbf{X}_c\}_{c\notin\mathcal{F}}))$ *be such that* $\mathcal{F}_{\text{in}}$ *and* $\mathcal{F}_{\text{out}}$ *have at most one element, but are not both empty. Define*

$$G^\top := (G, \mathfrak{d}, (\mathcal{F}_{\text{out}}, \mathcal{F}_{\text{in}}, \{\mathbf{X}_c\}_{c\notin\mathcal{F}}).$$

Then $\mathbf{W}_{G^\top}$ *is the adjoint of* $\mathbf{W}_G$*.*

Proof. Assume that $\mathcal{F}_{\text{in}} = (c_1)$ and $\mathcal{F}_{\text{out}} = (c_2)$. Then by definition, for any fixed $\mathbf{X}_{c_1}, \mathbf{X}_{c_2}$,

$$\left\langle \mathbf{W}_G\Big(\mathbf{X}_{c_1}\Big), \mathbf{X}_{c_2}\right\rangle = \mathbf{W}_{(G,\mathfrak{d},(\mathbf{X}_c)_{c\in C})} = \left\langle \mathbf{X}_{c_1}, \mathbf{W}_{G^\top}\Big(\mathbf{X}_{c_2}\Big)\right\rangle.$$

The proof is essentially identical if one of $\mathcal{F}_{\text{in}}$ or $\mathcal{F}_{\text{out}}$ is $(())$. □

Proposition 1.7 (Composition)**.** *Let* $G_1 = (G_1, \mathfrak{d}_1, \mathcal{F}_1 = ((a_1), (b_1)), \{\mathbf{X}_c\}_{c\notin\mathcal{F}_1})$ *and* $G_2 = (G_2, \mathfrak{d}_2, \mathcal{F}_2 = ((b_2), (c_2)), \{\mathbf{X}_c\}_{c\notin\mathcal{F}_2})$ *be such that* $\mathfrak{d}_1(b_1) = \mathfrak{d}_2(b_2)$*. Define* $G_2 \circ G_1$ *to be the graph obtained by identifying* b_1 *and* b_2 *and fixing the resulting cell's input to* $\mathbf{1}$ *if the* b_i *are vertices, and* $\mathbf{I}$ *if they are edges. In the event that vertices* v_1, v_2 *get identified as a result (e.g. when* b_1, b_2 *are edges), the resulting vertex is given input* $\mathbf{X}_{v_1} \odot \mathbf{X}_{v_2}$*. Then we have*

$$\mathbf{W}_{G_2\circ G_1} = \mathbf{W}_{G_2} \circ \mathbf{W}_{G_1}$$

where on the right hand side, $\circ$ *denotes composition.*

Proof. Assume that b_1, b_2 are vertices (the proof is essentially identical if they are edges) for which $\mathfrak{d}(b_1) = \mathfrak{d}(b_2) = d > 0$. Fix a pair of inputs $\mathbf{X}_{a_1}, \mathbf{X}_{c_2}$ for a_1 and c_2. Then on the one hand, theorem 1.4 yields

$$\left\langle \mathbf{W}_{G_2\circ G_1}\Big(\mathbf{X}_{a_1}\Big), \mathbf{X}_{c_2}\right\rangle = \left\langle \mathbf{W}_{G_2\circ G_1[\mathbf{X}_{a_1}]}, \mathbf{X}_{c_2}\right\rangle,$$

which is by definition equal to the value of the product graph obtained from $G_2 \circ G_1$ by fixing a_1 and c_2 to have input $\mathbf{X}_{a_1}$ and $\mathbf{X}_{c_2}$ respectively. Letting $V_i^- = V(G_i) \setminus \{b_i\}$ and $C_i^- = C_i \setminus \{b_i\}$ for simplicity, this equals

$$\begin{aligned}
&\sum_{\mathbf{i}_{V(G_1)}, \mathbf{i}_{V(G_2)}} \delta_{i_{b_1} i_{b_2}} \prod_{s_1\in C_1^-} [\mathbf{X}_{s_1}]_{i_{s_1}} \prod_{s_1\in C_2^-} [\mathbf{X}_{s_2}]_{i_{s_2}} \\
&= \sum_{i_{b_1}, i_{b_2}=1}^{d} \delta_{i_{b_1} i_{b_2}} \left(\sum_{\mathbf{i}_{V_1^-}} \prod_{s_1\in C_1^-} [\mathbf{X}_{s_1}]_{i_{s_1}}\right)\left(\sum_{\mathbf{i}_{V_2^-}} \prod_{s_2\in C_2^-} [\mathbf{X}_{s_2}]_{i_{s_2}}\right). \qquad (1.6)
\end{aligned}$$

On the other hand, theorems 1.5 and 1.6 yield

$$\begin{aligned}\left\langle \mathbf{W}_{G_2}\big(\mathbf{W}_{G_1}(\mathbf{X}_{a_1})\big), \mathbf{X}_{c_2}\right\rangle &= \left\langle \mathbf{W}_{G_1}\big(\mathbf{X}_{a_1}\big), \mathbf{W}_{G_2^\top}\big(\mathbf{X}_{c_2}\big)\right\rangle \\ &= \left\langle \mathbf{W}_{G_1}\big(\mathbf{X}_{a_1}\big) \odot \mathbf{W}_{G_2^\top}\big(\mathbf{X}_{c_2}\big), \mathbf{1}\right\rangle_{\mathbb{R}^d}\end{aligned}$$

which is equal to the right-hand side of (1.6). □

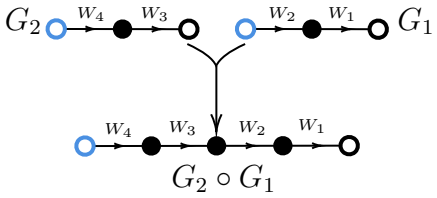


Figure 1.8: *Two graphs G_1, G_2 for which $\mathbf{W}_{G_1}(\mathbf{x}) = W_2W_1\mathbf{x}$ and $\mathbf{W}_{G_2}(\mathbf{x}) = W_4W_3\mathbf{x}$. Composing them yields the map $\mathbf{W}_{G_2\circ G_1}(\mathbf{x}) = W_4W_3W_2W_1\mathbf{x}$.*

This operation is depicted in fig. 1.8 when c_1, c_2 are vertices. A concrete example of composition by identifying edges will be given in section 2.4 when computing the neural tangent kernel.

Proposition 1.8 (Trace)**.** *Let $G = (G, \mathfrak{d}, ((c_1), (c_2)), \{\mathbf{X}_c\}_{c\in C\setminus\{c_1,c_2\}})$ be such that $\mathfrak{d}(c_1) = \mathfrak{d}(c_2)$. Define the* trace $\mathrm{Tr}(G)$ *of G to be the graph obtained by identifying c_1 with c_2 and fixing the resulting cell to $\mathbf{1}$ or $\mathbf{I}$ accordingly. Again, if any vertices v_1, v_2 are identified as a result (e.g., when c_1, c_2 are edges), the resulting vertex is given input $\mathbf{X}_1 \odot \mathbf{X}_2$ where $\mathbf{X}_1, \mathbf{X}_2$ are the respective inputs of v_1 and v_2. Then we have*

$$\mathrm{Tr}(\mathbf{W}_G) = \mathbf{W}_{\mathrm{Tr}(G)}.$$

Proof. Assume that c_1, c_2 are vertices (the argument is essentially identical if they are edges). Let $c_1 \wedge c_2$ denote the unified vertex in $\mathrm{Tr}(G)$, and C' be the set of cells of $\mathrm{Tr}(G)$. Then there is a canonical map $\phi : C \to C'$ which restricts to a bijection between $C\setminus\{c_1, c_2\}$ and $C'\setminus\{c_1\wedge c_2\}$, and such that $\phi(c_1) = \phi(c_2) = c_1\wedge c_2$. Moreover, this map induces a bijection between the indexations of $V(G)$ for which $i_{c_1} = i_{c_2}$ and the indexations $\mathbf{i}_{V(\mathrm{Tr}(G))}$ of $V(\mathrm{Tr}(G))$, which is given by $i_v \mapsto i_{\phi(v)}$ and under which $[\mathbf{X}_c]_{i_c} = [\mathbf{X}_{\phi(c)}]_{i_{\phi(c)}}$ for all $c \in C \setminus \{c_1, c_2\}$. It follows that

$$\begin{aligned}\mathrm{Tr}(\mathbf{W}_G) &= \sum_{\mathbf{i}_{V(G)}} \delta_{i_{c_1} i_{c_2}} \prod_{c\in C\setminus\{c_1,c_2\}} [\mathbf{X}_{\phi(c)}]_{i_{\phi(c)}} = \sum_{\mathbf{i}_{V(\mathrm{Tr}(G))}} \prod_{c'\in C'\setminus\{c_1\wedge c_2\}} [\mathbf{X}_{c'}]_{i_{c'}} \\ &= \sum_{\mathbf{i}_{V(\mathrm{Tr}(G))}} \prod_{c'\in C'\setminus\{c_1\wedge c_2\}} [\mathbf{X}_{c'}]_{i_{c'}} [\mathbf{1}]_{i_{c_1\wedge c_2}} = \mathbf{W}_{\mathrm{Tr}(G)}.\end{aligned}$$

□

Note that the cyclic property of the trace translates into the formation of a cycle in $\mathrm{Tr}(G)$, as depicted in fig. 1.9).

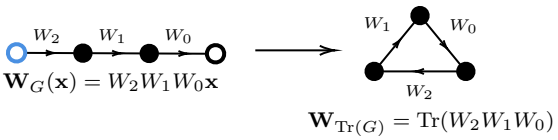


Figure 1.9: *The trace of a path graph.*

Remark. *Note how theorems 1.5 to 1.7 are analogous to the rules for entry-wise products, transposition and composition in Male [79] (namely (7), (4) and (3) in Example 1.3, respectively).*

We close this section by showing how differentiation can be performed graphically. To that end, it will be convenient to introduce the notion of sums of graphs.

Let $\{G_i\}_i$ be a set of operator graphs for which the associated $\{\mathbf{W}_{G_i}\}_i$ all belong to the same space. Let $\mathfrak{F} = (\mathfrak{F}(\{G_i\}_i), +)$ denote the free abelian group generated by the G_i. Then the map $G \mapsto \mathbf{W}_G$ can be uniquely extended to $\mathfrak{F}$ by

$$\mathbf{W}_{aG_1+bG_2} := a\mathbf{W}_{G_1} + b\mathbf{W}_{G_2}, \quad \forall a, b \in \mathbb{N}, aG_1 + bG_2 \in \mathfrak{F}.$$

In what follows, linear combinations of operator graphs are always understood as elements of an underlying free abelian group (which we will not specify).

Definition 1.9 (Differential)**.** *Let* $G = (G, \mathfrak{d}, (\mathcal{F}_{\mathrm{in}}, \mathcal{F}_{\mathrm{out}}), \{\mathbf{X}_c\}_{c\notin\mathcal{F}})$ *for which* $\mathcal{F}_{\mathrm{in}} = (c_1, \ldots, c_d)$ *for some* $d > 0$*. Fix a sequence* $(\mathbf{X}_{c_j})_{j=1}^d$ *of inputs and a sequence of* increments $(\mathbf{H}_{c_j})_{j=1}^d$ [1]*, and define*

$$(\mathrm{d}_i G)_{(\mathbf{X}_{c_j})_{j\in[d]}}[\mathbf{H}_{c_i}] := (G, \mathfrak{d}, ((), \mathcal{F}_{\mathrm{out}}), \{\mathbf{X}_c\}_{c\in C\setminus(\mathcal{F}_{\mathrm{out}}\cup\{c_i\})} \cup \{\mathbf{H}_{c_i}\}),$$

namely the graph obtained from G *by inserting* $\mathbf{X}_{c_j}$ *in* c_j *for all* $j \neq i$*, and* $\mathbf{H}_{c_i}$ *in* c_i*. Then the* differential *of* G *at* $(\mathbf{X}_{c_j})_{j\in[d]}$ *computed in the increment* $(\mathbf{H}_{c_j})_{j=1}^d$ *is defined as*

$$(\mathrm{d}G)_{(\mathbf{X}_{c_j})_{j\in[d]}}[(\mathbf{H}_{c_j})_{j=1}^d] := \sum_{i=1}^d (\mathrm{d}_i G)_{(\mathbf{X}_{c_j})_{j\in[d]}}[\mathbf{H}_{c_i}].$$

Proposition 1.10. *Let* $(G, \mathfrak{d}, (\mathcal{F}_{\mathrm{in}}, \mathcal{F}_{\mathrm{out}}), \{\mathbf{X}_c\}_{c\notin\mathcal{F}})$ *be such that* $|\mathcal{F}_{\mathrm{in}}| = d > 0$*. Denote*

[1] These are the increments of the Frechét derivative, meaning that each $\mathbf{H}_{c_j}$ has the same dimensionality of the corresponding $\mathbf{X}_{c_j}$, and that these are the inputs to the Frechét differential at the point $(\mathbf{X}_{c_j})_{j=1}^d$.

its i-th partial derivative by $\mathrm{d}_i\mathbf{W}_G$ and its total differential by $\mathrm{d}\mathbf{W}_G$. Then

$$(\mathrm{d}\mathbf{W}_G)_{(\mathbf{X}_{c_j})_{j\in[d]}}[(\mathbf{H}_{c_j})_{j=1}^d] = \sum_{i=1}^d (\mathrm{d}_i\mathbf{W}_G)_{(\mathbf{X}_{c_j})_{j\in[d]}}[\mathbf{H}_{c_i}]$$
$$= \sum_{i=1}^d \mathbf{W}_{(\mathrm{d}_i G)_{(\mathbf{X}_{c_j})_{j\in[d]}}[\mathbf{H}_{c_i}]} = \mathbf{W}_{(\mathrm{d}G)_{(\mathbf{X}_{c_j})_{j\in[d]}}[(\mathbf{H}_{c_j})_{j=1}^d]}$$

Proof. It suffices to show that $(\mathrm{d}_1\mathbf{W}_G)_{(\mathbf{X}_{c_j})_{j\in[d]}}[\mathbf{H}_{c_1}] = \mathbf{W}_{(\mathrm{d}_1 G)_{(\mathbf{X}_{c_j})_{j\in[d]}}[\mathbf{H}_{c_1}]}$. The claim then follows by definition of the total Frechét differential.

Let $\mathcal{F}_{\mathrm{in}} = (c_1, \dots, c_d)$, so that $\mathrm{d}_1\mathbf{W}_G$ denotes the partial derivative with respect to the cell c_1's input. Then since

$$\left\langle \mathbf{W}_G\big((\mathbf{X}_{c_i})_{i\in[d]}\big), \otimes_{c\in\mathcal{F}_{\mathrm{out}}}\mathbf{X}_c \right\rangle = \sum_{\mathbf{i}_{V(G)}} \left(\prod_{c\in C\setminus\mathcal{F}_{\mathrm{in}}} [\mathbf{X}_c]_{i_c} \right) \left(\prod_{j=2}^d [\mathbf{X}_{c_j}]_{i_{c_j}} \right) [\mathbf{X}_{c_1}]_{i_{c_1}},$$

and $(\mathrm{d}_1\mathbf{W}_G)_{(\mathbf{X}_{c_j})_{j\in[d]}}$ is linear (by definition), we have that

$$\begin{aligned}
\left\langle (\mathrm{d}_1\mathbf{W}_G)_{(\mathbf{X}_{c_j})_{j\in[d]}}[\mathbf{H}_{c_1}], \otimes_{c\in\mathcal{F}_{\mathrm{out}}}\mathbf{X}_c \right\rangle &= \sum_{\mathbf{i}_{V(G)}} \left(\prod_{c\in C\setminus\mathcal{F}_{\mathrm{in}}} [\mathbf{X}_c]_{i_c} \right) \left(\prod_{j=2}^d [\mathbf{X}_{c_j}]_{i_{c_j}} \right) [\mathbf{H}_{c_1}]_{i_{c_1}} \\
&= \sum_{\mathbf{i}_{V(G)}} \left(\prod_{c\in C\setminus\{c_1\}} [\mathbf{X}_c]_{i_c} \right) [\mathbf{H}_{c_1}]_{i_{c_1}} \\
&= \sum_{\mathbf{i}_{V((\mathrm{d}_1 G)_{(\mathbf{X}_{c_j})_{j\in[d]}})}} \left(\prod_{c\in C\setminus\{c_1\}} [\mathbf{X}_c]_{i_c} \right) [\mathbf{H}_{c_1}]_{i_{c_1}} \\
&=: \left\langle \mathbf{W}_{(\mathrm{d}_1 G)_{(\mathbf{X}_{c_j})_{j\in[d]}}}[\mathbf{H}_{c_1}], \otimes_{c\in\mathcal{F}_{\mathrm{out}}}\mathbf{X}_c \right\rangle
\end{aligned}$$

for any increment $\mathbf{H}_{c_1}$, where $C = C(G)$ denotes the set of cells of G. □

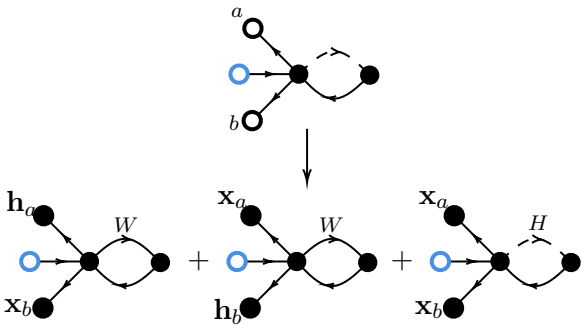


Figure 1.10: *An operator graph (top) and its total differential at $(\mathbf{x}_a, \mathbf{x}_b, W)$ (right) evaluated in the increment vector $(\mathbf{h}_a, \mathbf{h}_b, H)$.*

1.3 Genus Expansion and Limit Theorems

In this section, we develop tools to compute quantities of the form $\mathbb{E}\{\mathbf{W}_G\}$, where G is a product graph whose edge inputs are matrices with Gaussian entries. We will extend our methods to non-Gaussian and sparse inputs in section A.1. These results are all consequences of the classical Wick's Theorem, which we now recall.

For any set S of even size k, define a *pairing* ϕ of S to be a partition of S into pairs $S_j = \{\phi_{j,1}, \phi_{j,2}\}$ for $1 \le j \le k/2$. Then Wick's theorem reduces the task of computing expectations of a product of Gaussian random variables to that of computing pairwise covariances.

Theorem 1.11 (Wick's theorem)**.** *Let $(Z_i)_{i\ge 1}$ be a real centered Gaussian vector. Then for any sequence of integers $i_1, \dots, i_k$,*

$$\mathbb{E}\Big\{\prod_{j=1}^{k} Z_{i_j}\Big\} = \sum_{\phi}\prod_{j=1}^{k/2}\mathbb{E}\Big\{Z_{i_{\phi_{j,1}}} Z_{i_{\phi_{j,2}}}\Big\}.$$

where the sum is taken over all pairings of $\{1, \dots, k\}$.

The arguments below are inspired by those of Dubach and Peled [28], albeit with substantial modifications to accommodate our more general setting.

1.3.1 Wick expansion for product graphs with random inputs

Throughout this section, we will let $\mathcal{W}$ denote the following sequence of random matrices.

Definition 1.12. *Let $\mathcal{W} := (W_i)_{i\in\mathbb{N}}$, where $W_i \in \mathbb{R}^{N_{r_i}\times N_{c_i}}$ are independent Gaussian random matrices with i.i.d. entries distributed as $\mathcal{N}_{\mathbb{R}}(0, \sigma_i^2)$.*

Assumption 1.13. *We assume that any graph G considered throughout the section is a product graph whose inputs are either deterministic or matrices in $\mathcal{W}$.*

For any product graph G satisfying theorem 1.13 and edge $e \in E$, we will say that e is an "ℓ-edge" if its input is fixed to $W_\ell \in \mathcal{W}$. For simplicity, we label such edges by ℓ instead of W_ℓ when depicting the graph, as the latter can quickly become cumbersome.

We will consider pairings of edges of G whose inputs are matrices in $\mathcal{W}$, denoting the set of such edges by $E_{\mathcal{W}}$. Such a pairing is said to be *admissible* if it only pairs $\ell-$edges with $\ell-$edges.

Definition 1.14 (Admissible pairing)**.** *Let G be a product graph, and $E_{\mathcal{W}} = \{e \in E : \mathbf{X}_e = W_\ell$ for some $\ell(e) := \ell \in \mathbb{N}\}$. Then a pairing ϕ of $E_{\mathcal{W}}$ is said to be* admissible *if any two paired edges $\{e, e'\} \in \phi$ satisfy $\ell(e) = \ell(e')$. We denote the set of all admissible pairings of $E_{\mathcal{W}}$ by $\mathcal{P}(G)$.*

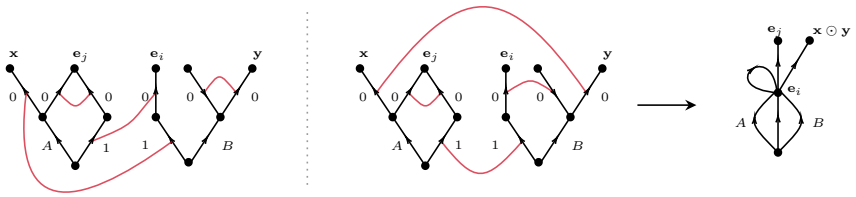


Figure 1.11: *Here A and B are determinitstic matrices. Left: A pairing which is* not *admissible (in red), since a* 1*-edge is paired to a* 0*-edge. Right: An admissible pairing (in red) and the corresponding G_ϕ (note that not all edges are necessarily paired).*

For any product graph G and admissible pairing $\phi \in \mathcal{P}(G)$, let G_ϕ be the product graph constructed by identifying the edges of G paired by ϕ coherently with their orientation[2] (see, e.g., fig. 1.11). The inputs of the cells in G_ϕ are assigned as follows:

- When identifying two edges fixed to W_ℓ, the resulting edge will be fixed to $\mathbb{E}[W_\ell \odot W_\ell] = \sigma_\ell^2 \mathbf{I}$.
- When identifying two vertices u, v (e.g. by pairing edges which have u and v as their respective endpoints), the resulting vertex will have input $\mathbf{x}_u \odot \mathbf{x}_v$.

Examples of pairings which are and aren't admissible are depicted in fig. 1.11, along with the graph G_ϕ that arises when admissibility holds.

In the event that the graph G is not connected, we write it as a disjoint union $G_\phi = \sqcup_{i=1}^{c(G_\phi)} (G_i)_\phi$ of its connected components $(G_i)_\phi$, where $c(G_\phi)$ denotes the number of connected components of G_ϕ.

Theorem 1.15 (Wick expansion)**.** *Let G be a (possibly disconnected) product graph satisfying theorem 1.13. Then*

$$\mathbb{E}\{\mathbf{W}_G\} = \sum_{\phi \in \mathcal{P}(G)} \mathbf{W}_{G_\phi}. \tag{1.7}$$

where ϕ runs over all of G's admissible *pairings.*

Proof. Let $G = (G, \mathfrak{d}, \{\mathbf{X}\}_{c \in C})$ be a product graph and consider the set of indexations $\mathcal{I}^G := \prod_{v \in V(G)} [N_v]$. Recall that for each $e = (u, v) \in E$, $i_e = i_u i_v$. Then for any $\mathbf{i}_V \in \mathcal{I}^G$,

$$\mathbb{E}\Big\{ \prod_{c \in C} [\mathbf{X}_c]_{i_c} \Big\} = \prod_{c \in C \setminus E_\mathcal{W}} [\mathbf{X}_c]_{i_c} \mathbb{E}\Big\{ \prod_{e \in E_\mathcal{W}} [\mathbf{X}_e]_{i_e} \Big\}.$$

By Wick's theorem, we have

$$\mathbb{E}\Big\{ \prod_{e \in E_\mathcal{W}} [\mathbf{X}_e]_{i_e} \Big\} = \sum_{\phi} \prod_{\{e,e'\} \in \phi} \mathbb{E}\Big\{ [\mathbf{X}_e]_{i_e} [\mathbf{X}_{e'}]_{i_{e'}} \Big\}$$

[2] As will become clear from theorem 1.15's proof, paired edges must be identified in the same orientation since by Wick's theorem the matrices' ordered index has to coincide to obtain a non-vanishing term.

where the sum is over all pairings ϕ of $E_{\mathcal{W}}$. Note that if $\phi \notin \mathcal{P}(G)$ then

$$\prod_{\{e,e'\}\in\phi} \mathbb{E}\Big\{[\mathbf{X}_e]_{i_e}[\mathbf{X}_{e'}]_{i_{e'}}\Big\} = 0$$

and we can therefore write

$$\begin{aligned}\mathbb{E}\{\mathbf{W}_G\} &= \sum_{\mathbf{i}_V}\Bigg(\prod_{c\in C\setminus E_{\mathcal{W}}}[\mathbf{X}_c]_{i_c}\Bigg)\sum_{\phi\in\mathcal{P}(G)}\prod_{\{e,e'\}\in\phi}\mathbb{E}\Big\{[\mathbf{X}_e]_{i_e}[\mathbf{X}_{e'}]_{i_{e'}}\Big\}\\ &= \sum_{\phi\in\mathcal{P}(G)}\sum_{\mathbf{i}_V}\Bigg(\prod_{c\in C\setminus E_{\mathcal{W}}}[\mathbf{X}_c]_{i_c}\Bigg)\prod_{\{e,e'\}\in\phi}\sigma^2_{\ell(e)}\delta_{i_e i_{e'}}, \end{aligned} \tag{1.8}$$

where $\delta_{ij} := \mathbf{1}(i=j)$ is the Kronecker delta function.

For any vertices $u, v \in V$, we write $u \sim_\phi v$ if they are identified after identifying edges according to ϕ. Similarly, we write $e \sim_\phi e'$ if $\{e, e'\} \in \phi$. If we let $V_{\mathcal{W}}$ be the set of vertices adjacent to edges in $E_{\mathcal{W}}$, we then have canonical bijections

$$V(G_\phi) \simeq V^c_{\mathcal{W}} \sqcup (V_{\mathcal{W}}/\sim_\phi), \quad E(G_\phi) \simeq E^c_{\mathcal{W}} \sqcup (E_{\mathcal{W}}/\sim_\phi)$$

Note that each element in $\varepsilon \in E_{\mathcal{W}}/\sim_\phi$ corresponds to a block $\varphi(\varepsilon) \in \phi$, while each element of $\nu \in V_{\mathcal{W}}/\sim_\phi$ corresponds to a subset $\varphi(\nu) \subset V$ (these are the pre-images under the quotient maps). Then for any $\phi \in \mathcal{P}(G)$ we can write the product of G_ϕ in terms of indexations of G as

$$\begin{aligned}\mathbf{W}_{G_\phi} &= \sum_{\mathbf{i}\in\mathcal{I}^{G_\phi}}\prod_{c\in E^c_{\mathcal{W}}\cup V^c_{\mathcal{W}}}[\mathbf{X}_c]_{i_c}\Bigg(\prod_{\varepsilon\in(E_{\mathcal{W}}/\sim_\phi)}\sigma^2_{\ell(\varepsilon)}\Bigg)\Bigg(\prod_{\nu\in(V_{\mathcal{W}}/\sim_\phi)}\prod_{u\in\varphi(\nu)}[\mathbf{X}_u]_{i_\nu}\Bigg)\\ &= \sum_{\mathbf{i}\in\mathcal{I}^{G}}\prod_{c\in E^c_{\mathcal{W}}\cup V^c_{\mathcal{W}}}[\mathbf{X}_c]_{i_c}\Bigg(\prod_{\{e,e'\}\in\phi}\sigma^2_{\ell(e)}\delta_{i_e i_{e'}}\Bigg)\Bigg(\prod_{\nu\in(V_{\mathcal{W}}/\sim_\phi)}\delta_{\mathbf{i}|_\nu}\prod_{u\in\varphi(\nu)}[\mathbf{X}_u]_{i_u}\Bigg)\end{aligned}$$

Here $\delta_{\mathbf{i}|_\nu} := \mathbf{1}\left(\forall u, u' \in \varphi(\nu),\ i_u = i_{u'}\right)$ and $\sigma_{\ell(\varepsilon)} = \sigma_{\ell(e)}$ for any choice of $e \in \varphi(\varepsilon)$, since they all yield the same $\sigma_{\ell(e)}$. By definition, there exists a $\nu \in (V_{\mathcal{W}}/\sim_\phi)$ such that $u, u' \in \varphi(\nu)$ if and only if we have $\{e, e'\} \in \phi$ such that u, u' are either the respective heads or or respective tails of e, e', it follows that

$$\prod_{\{e,e'\}\in\phi}\delta_{i_e i_{e'}} = \prod_{\nu\in(V_{\mathcal{W}}/\sim_\phi)}\delta_{\mathbf{i}|_\nu}.$$

Using this, the claim then follows from the equality

$$\prod_{c\in C\setminus E_{\mathcal{W}}}[\mathbf{X}_c]_{i_c} = \prod_{c\in E^c_{\mathcal{W}}\cup V^c_{\mathcal{W}}}[\mathbf{X}_c]_{i_c}\prod_{\nu\in(V_{\mathcal{W}}/\sim_\phi)}\prod_{u\in\varphi(\nu)}[\mathbf{X}_u]_{i_u}.$$

□

We now derive a centered version of theorem 1.15 (theorem 1.17), which we will need to derive limit theorems for $\mathbf{W}_G$. To that end, we need to single out the admissible pairings which give rise to leading and subleading order terms in the limit. The nomenclature below is borrowed from [28], the following is essentially Definition 2.3.

Definition 1.16. *Let $G = \sqcup_{i=1}^{c(G)} G_i$ be a product graph, given as a disjoint union of its connected components, and fix an admissible pairing $\phi \in \mathcal{P}(G)$ of its edges.*

1. *ϕ is said to be* atomic *if and only if for every $1 \leq i \leq c(G)$, the edges in $E(G_i)$ are paired between themselves and $(G_i)_{\phi|_{G_i}}$ is a tree, where $\phi|_{G_i}$ denotes the pairing induced by ϕ on the edges of G_i.*
2. *ϕ is* bi-atomic *if and only if for every $1 \leq i \leq c(G)$, there exists a $j \neq i$ such that ϕ pairs edges of $E(G_i) \sqcup E(G_j)$ between themselves and $(G_i \sqcup G_j)_{\phi|_{G_i \sqcup G_j}}$ is a tree.*
3. *ϕ is* atom-free *if and only if there is no G_i for which ϕ pairs the edges in $E(G_i)$ between themselves and $(G_i)_\phi$ is a tree.*

We denote the set of (admissible) atomic, bi-atomic and atom-free pairings of G by $\mathcal{P}_{\mathrm{A}}(G)$, $\mathcal{P}_{\mathrm{B}}(G)$, and $\mathcal{P}_{\mathrm{AF}}(G)$ respectively, note that these pairings do not fully cover the set of admissible pairings. With these definitions, we can now state and prove the following centered version of theorem 1.15.

Proposition 1.17. *Let $G = \sqcup_{i=1}^{c(G)} G_i$ be a product graph satisfying theorem 1.13, with disjoint connected components G_i. Then*

$$\mathbb{E}\left\{ \prod_{i=1}^{c(G)} \left(\mathbf{W}_{G_i} - \sum_{\psi \in \mathcal{P}_{\mathrm{A}}(G_i)} \mathbf{W}_{(G_i)_\psi} \right) \right\} = \sum_{\phi \in \mathcal{P}_{\mathrm{AF}}(G)} \mathbf{W}_{G_\phi} \tag{1.9}$$

Proof. Let $[c(G)] := \{1, \ldots, c(G)\}$, and for any $T \subseteq [c(G)]$ define $G_T := \sqcup_{i \in T} G_i$. Then by expanding the expectation in (1.9), we get

$$\begin{aligned} &\mathbb{E}\left\{ \prod_i \left(\mathbf{W}_{G_i} - \sum_{\psi \in \mathcal{P}_{\mathrm{A}}(G_i)} \mathbf{W}_{(G_i)_\psi} \right) \right\} \\ &\qquad = \sum_{T \subseteq [c(G)]} (-1)^{|T^c|} \mathbb{E}\left\{ \prod_{i \in T} \mathbf{W}_{G_i} \right\} \left(\prod_{i \notin T} \sum_{\psi \in \mathcal{P}_{\mathrm{A}}(G_i)} \mathbf{W}_{(G_i)_\psi} \right) \end{aligned}$$

Using theorem 1.15, we have

$$\mathbb{E}\left\{ \prod_{i \in T} \mathbf{W}_{G_i} \right\} = \sum_{\phi \in \mathcal{P}(G_T)} \mathbf{W}_{(G_T)_\phi}$$

and in turn

$$= \sum_{T\subseteq[c(G)]} \sum_{\phi\in\mathcal{P}(G_T)} (-1)^{|T^c|}\mathbf{W}_{(G_T)_\phi} \prod_{i\notin T} \sum_{\psi\in\mathcal{P}_{\mathrm{A}}(G_i)} \mathbf{W}_{(G_i)_\psi}$$
$$= \sum_{T\subseteq[c(G)]} (-1)^{|T^c|} \sum_{\phi\in\mathcal{P}(G_T)\times(\prod_{i\in T^c}\mathcal{P}_{\mathrm{A}}(G_i))} \mathbf{W}_{G_\phi}$$

and the conclusion follows from the inclusion-exclusion principle since

$$\mathcal{P}_{\mathrm{AF}}(G) = \bigcap_{i=1}^{c(G)} (\mathcal{P}(G_{[c(G)]\setminus\{i\}}) \times \mathcal{P}_{\mathrm{A}}(G_i))^c = \mathcal{P}(G) \setminus \bigcup_{i=1}^{c(G)} \Big(\mathcal{P}(G_{[c(G)]\setminus\{i\}}) \times \mathcal{P}_{\mathrm{A}}(G_i)\Big)$$

and

$$\bigcap_{i\in T^c} \Big(\mathcal{P}(G_{[c(G)]\setminus\{i\}}) \times \mathcal{P}_{\mathrm{A}}(G_i)\Big) = \mathcal{P}(G_T) \times \prod_{i\in T^c} \mathcal{P}_{\mathrm{A}}(G_i).$$

□

More generally, the same proof gives the following.

Proposition 1.18. *Let $G = \sqcup_{i=1}^{c(G)} G_i$ be a product graph satisfying theorem 1.13, with disjoint connected components G_i. Given any sequence $\mathcal{P}_i \subseteq \mathcal{P}(G_i)$ and setting $\mathcal{P}_\star := \bigcap_i [\mathcal{P}(G_{[c(G)]\setminus\{i\}}) \sqcup \mathcal{P}_i]^c$ one has*

$$\mathbb{E}\Big\{ \prod_{i=1}^{c(G)} \Big(\mathbf{W}_{G_i} - \sum_{\psi\in\mathcal{P}_i} \mathbf{W}_{(G_i)_\psi} \Big) \Big\} = \sum_{\phi\in\mathcal{P}_\star} \mathbf{W}_{G_\phi}. \tag{1.10}$$

1.3.2 Genus expansion

We now study a particular setting in which it will be possible to explicitly express $\mathbf{W}_{G_\phi}$ for each ϕ. Recalling that $\mathcal{W}$ is the sequence of random matrices serving as inputs to edges in $E_{\mathcal{W}}(G) \subset E(G)$, we assume the following.

Assumption 1.19. *$\mathcal{W}$ is such that all $N_{c_i} = N_{r_i} = N$ for some integer $N > 0$, and G is such that $\mathfrak{d}(v) = N$, $\mathbf{X}_v = \mathbf{1}_N$ for all $v \in V$ and $\mathbf{X}_e = \mathbf{I}$ for all $e \in E \setminus E_{\mathcal{W}}(G)$.*

Under theorem 1.19, the definition of $\mathbf{W}_{G_\phi}$ then directly yields

$$\mathbf{W}_{G_\phi} = \sigma_G N^{|V(G_\phi)|}, \quad \sigma_G = \prod_{e\in E_{\mathcal{W}}} \sigma_{\ell(e)}, \tag{1.11}$$

given that all the edges of G_ϕ are labeled by $\mathbf{I}_{N\times N}$. The order of $\mathbf{W}_{G_\phi}$ is thus determined by the number of vertices in G_ϕ.

Instead of counting the number of vertices $|V(G_\phi)|$ directly, it will be simpler to associate a surface to G_ϕ and use its Euler characteristic to express $|V(G_\phi)|$ in terms of this surface's genus, number of edges and connected components. This forms the basis of the *genus expansion* technique.

To use it, we will need some basic notions from topological graph theory (see [43] for a more in-depth introduction). Any connected graph $G = (V, E)$ can be viewed as a topological space, by viewing vertices as distinct points and edges as subspaces homeomorphic to $[0, 1]$ joining their ends, such that edges only meet at their endpoints. An *embedding* ρ of G on some orientable, compact surface S without boundary is then defined as a homeomorphism from G (viewed as a space) to S, and is said to be *cellular* if the resulting faces are all homeomorphic to the open disk. For our purposes, all embeddings are assumed to be cellular.

The connected components of $S \setminus \rho(G)$ are referred to as the *faces* of the embedding, and we denote the total number of faces by $f(G : S)$. Lastly, we denote the genus of S is denoted by $g(S)$. For any cellular embedding, we have the following formula.

Proposition 1.20 (Euler Characteristic formula)**.** *Let ρ be a (cellular) embedding of a connected $G = (V, E)$ on an orientable surface S. Then we have*

$$|V| - |E| + f(G : S) = 2 - 2g(S).$$

For instance, if G is a planar graph with F faces, then its drawing on the plane is a cellular embedding and this reduces to the famous identity $|V| - |E| + F = 2$.
Further, it is well known that all compact, orientable surfaces without boundary of genus g are equivalent up to homeomorphism (as a representative of their equivalence class, one can for instance take the surface obtained by adding g "handles" to a 2-sphere, an idea going back to [6]). For any $g \geq 0$, it thus makes sense to talk about *the* surface of genus g, which we henceforth denote by S_g.

Lastly, we can define embeddings of disconnected graphs by considering each connected component separately. Letting $c(G)$ denote the number of components and $G = \sqcup_{i=1}^{c(G)} G_i$, an embedding of G is given by embedding each G_i on an oriented surface S_i. We can also extend the definitions of f and g by writing $S = \sqcup_{i=1}^{c(G)} S_i$,

$$g(S) := \sum_{i=1}^{c(G)} g(S_i), \quad \text{and } f(G : S) := \sum_{i=1}^{c(G)} f(G_i : S_i),$$

and we note that the Euler characteristic formula still holds with these definitions.

Going back the setting of theorem 1.19, let G be such a product graph and $\phi \in \mathcal{P}(G)$ be an admissible pairing of its edges. Recalling that G_ϕ is the graph obtained from G after identifying edges paired by ϕ, we let

$$S_\phi := \min_{g \geq 0} \{S_g : \text{ there exists an embedding } \rho : G_\phi \to S_g\}. \tag{1.12}$$

if G is connected. Otherwise, let $(S_i)_\phi$ be the surface obtained this way from each connected component G_i of G, and define $S_\phi = \sqcup_{i=1}^{c(G)} (S_i)_\phi$.

Remark. *This minimum in (1.12) is always attained, and the value of g for which $S_g = S_\phi$ is often referred to as the minimum orientable genus of G in the literature (see [43]).*

We do not dwell on the details of these concepts and definitions, as we essentially only care about the case when G_ϕ is a tree in our arguments. In that case, $g(S_\phi) = 0$ and we concern ourselves with $f(G_\phi : S_\phi)$ and $c(G_\phi)$ instead.

Note also that we obtained S_ϕ from G_ϕ in a different way than in [28]: in their case, G is always a cycle, which naturally gives rise to a closed surface after identifying edges. No such construction seems to exist in our case.

For any G satisfying theorem 1.19 and $\phi \in \mathcal{P}(G)$ we note that G_ϕ always has the same number of edges, which we denote by

$$\check{e}(G) := |E(G_\phi)| = |E_{\mathcal{W}}(G)|/2 + |E \setminus E_{\mathcal{W}}(G)|, \tag{1.13}$$

since the random edges are paired up by ϕ and the rest are left untouched. Then theorem 1.15 yields the following:

Corollary (Genus expansion)**.** *Let G be a product graph satisfying theorem 1.13 and theorem 1.19. Then for any $\phi \in \mathcal{P}(G)$,*

$$\mathbf{W}_{G_\phi} = \sigma_G N^{2(c(G_\phi)-g(S_\phi))-f(G_\phi:S_\phi)+\check{e}(G)}, \tag{1.14}$$

and it follows that

$$\mathbb{E}\{\mathbf{W}_G\} = \sigma_G N^{\check{e}(G)} \sum_{\phi\in\mathcal{P}(G)} N^{2(c(G_\phi)-g(S_\phi))-f(G_\phi:S_\phi)}. \tag{1.15}$$

Proof. This follows directly from eq. (1.11) and the Euler characteristic formula, which gives

$$|V(G_\phi)| = 2(c(G_\phi) - g(S_\phi)) + |E(G_\phi)| - f(G_\phi : S_\phi). \tag{1.16}$$

□

Let us demonstrate how this result can be used to study words of random matrices, as defined in [28]. As explained in the following example, this amounts to computing the values of polygonal product graphs.

Example. *Under theorem 1.19, let $w = W_1 W_4^T W_2^T W_5 W_3^T$ be a word of matrices in $\mathcal{W}$. Then taking G to be the product graph in fig. 1.12 below gives $\mathbf{W}_G = \mathrm{Tr}(w)$. Products of such traces (the subject of Theorem 2.2 in Dubach and Peled [28]) can then be evaluated using section 1.3.2 by considering disjoint unions of such graphs.*

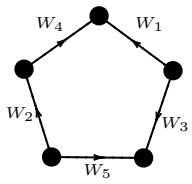


Figure 1.12: *A product graph whose value is the trace of a word of Ginibre random matrices.*

In this specific setting, we can also refine our result on centered mixed moments.

Lemma 1.21. *Let G satisfy theorem 1.13 and theorem 1.19. Then for every $\phi \in \mathcal{P}_{\mathrm{AF}}(G)$, we have*

$$2(c(G_\phi) - g(S_\phi)) - f(G_\phi : S_\phi) \leq \frac{c(G)}{2}.$$

Further, equality holds if and only if ϕ is bi-atomic, giving

$$\mathbb{E}\left\{\prod_{i=1}^{c(G)} \left(\mathbf{W}_{G_i} - \sigma_{G_i} N^{\check{e}(G_i)+1} |\mathcal{P}_{\mathrm{A}}(G_i)|\right)\right\} = \sigma_G N^{\check{e}(G)+\frac{c(G)}{2}} \left(|\mathcal{P}_{\mathrm{B}}(G)| + \mathcal{O}\left(\frac{1}{N}\right)\right).$$

Proof. Assume that $c(G_\phi) \leq c(G)/2$. Then it suffices to prove that for each connected component G_i of G and its corresponding surface S_i,

$$2(1 - g((S_i)_\phi)) - f\Big((G_i)_\phi : (S_i)_\phi\Big) \leq 1$$

and that the equality holds if and only if $(G_i)_\phi$ is a tree.

Since $f\Big((G_i)_\phi : (S_i)_\phi\Big) \geq 1$ and $g\Big((S_i)_\phi\Big) \geq 0$, we have

$$2(1 - g\Big((S_i)_\phi\Big)) - f\Big((G_i)_\phi : (S_i)_\phi\Big) \leq 2(1-0) - 1 = 1$$

with equality if and only if $f\Big((G_i)_\phi : (S_i)_\phi\Big) = 1$ and $g\Big((S_i)_\phi\Big) = 0$ (which is equivalent to $(G_i)_\phi$ being a tree). Now if $c(G_\phi) > c(G)/2$ instead, then there must be at least

$$2\left(c(G_\phi) - \frac{c(G)}{2}\right)$$

components of G_ϕ of the form $(G_j)\Big|_\phi$ for some $1 \leq j \leq c(G)$. Letting k be the number of components obtained this way, then we must have that $k \geq 2(c(G_\phi) - c(G)/2)$: assuming otherwise, the inequality

$$c(G) \geq k + 2(c(G_\phi) - k) = 2\left(c(G_\phi) - \frac{k}{2}\right)$$

quickly gives a contradiction. Since ϕ is atom-free,

$$f((G_i)_\phi : (S_i)_\phi) = 1 \Rightarrow g((S_i)_\phi) > 0$$

for these components, and it follows that

$$2(1 - g\Big((S_i)_\phi\Big)) - f\Big((G_i)_\phi : (S_i)_\phi\Big) \leq 0.$$

As for the remaining components (of which there are strictly less than $c(G)/2$) this quantity is ≤ 1 as proved above, and the lemma follows. □

Corollary. *Let G be a product graph satisfying theorem 1.13 and theorem 1.19, and $\lambda_G := (\sigma_G N^{\check{e}(G)+1})^{-1}$*

$$\mathbb{E}\left\{(\lambda_G \mathbf{W}_G - |\mathcal{P}_{\mathrm{A}}(G)|)^2\right\} = \frac{1}{N}\left(|\mathcal{P}_{\mathrm{B}}(G)| + \mathcal{O}\left(\frac{1}{N}\right)\right). \tag{1.17}$$

Lastly, we note that by combining the previous two results, we can derive the following limit theorem for the joint distribution of product graphs.

Theorem 1.22. *Let $(G_i)_{i\geq 1}$ be a sequence of connected product graphs satisfying theorem 1.13 and theorem 1.19. Then for any $m \geq 1$,*

$$\Big((\sigma_{G_i} N^{\check{e}(G_i)+\frac{1}{2}})^{-1}\mathbf{W}_{G_i} - |\mathcal{P}_{\mathrm{A}}(G_i)|\sqrt{N}\Big)_{i=1}^{m} \xrightarrow{d} \mathbf{Z} = \Big(Z_1, \ldots, Z_m\Big)^\top$$

where the limit is a centered Gaussian vector with covariance function $C := \mathbb{E}[\mathbf{Z}\mathbf{Z}^\top]$ having entries $[C]_{i,j} = |\mathcal{P}_{\mathrm{B}}(G_i \sqcup G_j)|$.

Proof. We use the method of moments. Setting

$$\tilde{Z}_i := (\sigma_{G_i} N^{\check{e}(G_i)+\frac{1}{2}})^{-1}\mathbf{W}_{G_i} - |\mathcal{P}_{\mathrm{A}}(G_i)|\sqrt{N}$$

we want to compute the expectations

$$\mathbb{E}\left\{\prod_{i=1}^{m}(\tilde{Z}_i)^{k_i}\right\}$$

for non-negative integers k_i. We are in the setting of theorem 1.21 and the leading term of the expectation is the number of bi-atomic pairings of the induced graph. Given a possible pairing of its connected components, for each of its pairs $(\tilde{G}, \tilde{G}')$ this contributes a factor $|\mathcal{P}_{\mathrm{A}}(\tilde{G} \sqcup \tilde{G}')|$ to the total sum. But this is exactly the result of computing

$$\mathbb{E}\left\{\prod_{i=1}^{m}(Z_i)^{k_i}\right\}$$

by applying Wick's theorem. □

Chapter 2

Graph Expansions of Neural Networks

2.1 Introduction

Randomly initialised neural networks and non-linear random matrix theory

Deep neural networks (NNs) whose weights' and biases' entries are initialised as appropriately rescaled, independent and identically distributed (i.i.d.) Gaussian random variables converge to Gaussian processes (GPs) as their width tends to infinity. This well-known fact was originally observed by Neal [91] for shallow feedforward networks and more recently by de G. Matthews et al. [24] for multi-layer feedforward networks, by Novak et al. [94] and Garriga-Alonso et al. [39] for deep convolutional networks, and by Yang [129] for more general architectures, including recurrent and attention-based networks. The study of such random initializations is of practical interest, as a better understanding of their properties can help make design choices that improve training. For instance, the concentration of the spectrum of the network's input-output Jacobian can help prevent vanishing and exploding gradients [41, 103, 115], a crucial ingredient for the stability of training dynamics. Other relevant objects which have been studied include the Gram [100] and Fisher information [101] matrices, as well as the *neural tangent kernel* (NTK) [58]. From a theoretical viewpoint, these offer natural non-linear generalizations of the classical questions of random matrix theory (RMT) and free probability, and as such have attracted the attention of both communities [3, 29, 97, 98, 128].

Albeit of a similar nature, these results have been derived using a wide array of tools which do not always offer a clear path towards universality (meaning initialisations with non-Gaussian random weights). Our goal is to propose a unifying framework that is naturally suited to this task, and is built on the *genus expansion* techniques developed in chapter 1.

Main results of this chapter

Fix sequences $(\varphi_\ell : \mathbb{R} \to \mathbb{R} \mid \ell \in \mathbb{N}_{>0})$ of polynomial *activation functions*, $(N_\ell \in \mathbb{N}_{>0} \mid \ell \in \mathbb{N})$ of *layer dimensions* and $(W_\ell \in \mathbb{R}^{N_{\ell+1} \times N_\ell} \mid \ell \in \mathbb{N})$ of weight matrices. We define a *feed-forward neural network* Φ_L of depth L by the recursion

$$\Phi_0(\mathbf{x}) = W_0\mathbf{x}, \quad \Phi_{\ell+1}(\mathbf{x}) = W_{\ell+1}\varphi_{\ell+1}(\Phi_\ell(\mathbf{x})), \tag{2.1}$$

where each φ_ℓ is applied entry-wise. We omit bias terms and restrict ourselves to polynomial activations for simplicity here (generalizations are discussed in section 2.1.2).

To illustrate the proposed method, we prove three results on NNs initialised with sparse random weights, beginning with the following Gaussian process limit, a crucial result in the theory of expressivity of neural networks.

Theorem 2.1 (Gaussian process limit for sparse neural networks)**.** *Let $N_\ell = N$ when $\ell > 0$, and assume that each W_ℓ has i.i.d. entries drawn from a symmetric, centred distribution with finite moments and variance $\frac{1}{N}\mathbf{1}(\ell > 0) + \mathbf{1}(\ell = 0)$.*

Then for any $M, L \geq 1$ we have

$$([\Phi_L]_1, \ldots, [\Phi_L]_M) \xrightarrow[N\to\infty]{d} \mathcal{GP}(0, K_L \otimes \mathrm{Id}_M) \tag{2.2}$$

where the right-hand side is a Gaussian Process indexed on $\mathbb{R}^{N_0}$, with diagonal covariance function defined by

$$K_0(\mathbf{x}, \mathbf{y}) = \langle \mathbf{x}, \mathbf{y} \rangle_{\mathbb{R}^{N_0}}, \; K_{\ell+1}(\mathbf{x}, \mathbf{y}) = \mathbb{E}\left[\varphi_{\ell+1}(X_\ell)\varphi_{\ell+1}(Y_\ell)\right] \tag{2.3}$$

$$(X_\ell, Y_\ell) \sim \mathcal{N}\left(0, \begin{bmatrix} K_\ell(\mathbf{x}, \mathbf{x}) & K_\ell(\mathbf{x}, \mathbf{y}) \\ K_\ell(\mathbf{y}, \mathbf{x}) & K_\ell(\mathbf{y}, \mathbf{y}) \end{bmatrix}\right), \tag{2.4}$$

and Id_M is the $M \times M$ identity matrix. Furthermore, the same result holds if the weight matrices in (2.1) are replaced by $\tilde{W}_\ell := W_\ell \odot \frac{1}{\sqrt{p_N}} B_\ell$, where W_ℓ are as above and the B_ℓ are independent matrices with i.i.d., Bernoulli distributed entries with parameter p_N satisfying $Np_N \to \infty$.

Proof. The claim for W_ℓ with Gaussian entries is theorem 2.9, and the corollaries in chapter A generalize this to the non-Gaussian and sparse cases. □

This adds to the growing list of generalisations of the result of de G. Matthews et al. [24] to non-Gaussian settings, such as that of Huang et al. [55] to orthogonal weights, and Hanin [50] to weights with i.i.d. entries satisfying finite moment assumptions. More recently, Saada et al. [111] encompassed both of these results by showing that one can relax the i.i.d. assumption to a class of weights which they call Pseudo-IID. In particular, this class includes *structured sparse* weights, making their work the first

to rigorously show that the Gaussian process limit holds in a sparse setting. That said, while their result holds for more general activations than the ones considered here, it only deals with sparsification using a fixed binary mask B, whereas theorem 2.9 allows for masks B_ℓ whose expected proportion of ones decreases as N_ℓ tends to infinity.

Our second result concerns the NTK ([27, 58]), which is defined as the matrix representation $\Theta_L(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{N_{L+1} \times N_{L+1}}$ of

$$\sum_{\ell=0}^{L} \lambda_\ell^2 (\mathrm{d}\Phi_L(\mathbf{x}))_{W_\ell} \circ (\mathrm{d}\Phi_L(\mathbf{y}))_{W_\ell}^\top, \tag{2.5}$$

where $(\mathrm{d}\Phi_L(\mathbf{y}))_{W_\ell}^\top$ denotes the adjoint of $(\mathrm{d}\Phi_L(\mathbf{y}))_{W_\ell}$, and the $(\lambda_\ell)_\ell$ are the so-called *layer-wise learning rates*. We show that at initialisation and under "NTK parametrisation", Θ_L converges in L^2 to a deterministic kernel matrix (in the space of $N_{L+1} \times N_{L+1}$ matrices equipped with Frobenius norm), at a rate that is inversely proportional to the layer width. Note that the implicit constant in this estimate depends on the depth L.

The importance of NTK theory is in its offering a theoretical explanation for the empirical success of overparameterized deep neural networks, which are able to generalise well to unseen data even when having many more parameters than training data points. In the infinite-width limit, these networks behave as kernel machines during training and converge to a global minimum on the training loss landscape, much like other well-understood kernel methods (see Jacot et al. [58]). This contrasts with traditional statistical learning theory which often predicts overfitting in such scenarios.[1]

Theorem 2.2 (Convergence in L^2 of the NTK at intialisation)**.** *For each $0 < \ell \leq L$, let $N_\ell = N$ and assume that W_ℓ has i.i.d. entries drawn from a symmetric, centered distribution with finite moments and variance $\sigma_\ell^2 = \frac{1}{N}\mathbf{1}(\ell > 0) + \mathbf{1}(\ell = 0)$. Moreover, assume that the layer-wise learning rates are $\lambda_\ell = \sigma_\ell^2$.*

Then

$$\mathbb{E}\left\{ \left\| \Theta_L(\mathbf{x}, \mathbf{y}) - \Theta_L^\infty(\mathbf{x}, \mathbf{y}) \otimes \mathrm{Id}_{N_{L+1}} \right\|^2 \right\} = \mathcal{O}\left(\frac{N_{L+1}^2}{N} \right) \tag{2.6}$$

where $\|\cdot\|$ denotes the Frobenius norm, and $\Theta_L^\infty : \mathbb{R}^{N_0} \times \mathbb{R}^{N_0} \to \mathbb{R}$ is a scalar-valued kernel defined recursively as

$$\Theta_0^\infty(\mathbf{x}, \mathbf{y}) = \langle \mathbf{x}, \mathbf{y} \rangle_{\mathbb{R}^{N_0}}, \quad \Theta_L^\infty(\mathbf{x}, \mathbf{y}) = K_L(\mathbf{x}, \mathbf{y}) + \dot{K}_L(\mathbf{x}, \mathbf{y}) \Theta_{L-1}^\infty(\mathbf{x}, \mathbf{y}). \tag{2.7}$$

($\dot{K}_\ell$ is defined in the same way as K_ℓ only substituting φ_ℓ for φ_ℓ' in (2.3).)

[1] The story is more nuanced: in the NTK limit, the features learned by networks remain "frozen" at their initialization values and cannot evolve. This limitation hinders performance in practice, leading to recent proposals of new parametrizations that enable controlled "feature learning" regimes (see Chizat et al. [13], Yang and Hu [132]). See chapter E for a derivation of Chizat et al. [13]'s results within our framework.

The same result holds with a rate of $\mathcal{O}(N_{L+1}^2/(Np_N))$ *if the weight matrices in* (2.1) *are replaced by* $\tilde{W}_\ell := W_\ell \odot \frac{1}{\sqrt{p_N}} B_\ell$*, where* W_ℓ *are as above and the* B_ℓ *are independent matrices with i.i.d., Bernoulli distributed entries with parameter* p_N *satisfying* $Np_N \to \infty$*.*

Proof. The claim for W_ℓ with Gaussian entries is theorem 2.14, and the corollaries in chapter A generalize this to the non-Gaussian and sparse cases. □

Previous results regarding the NTK at initialization have only been shown for Gaussian and orthogonal weights [55, 58]. With the caveat of only holding for polynomial activations, our result is one of the few quantitative results on the NTK for deep networks, and to the best of our knowledge, the only result to hold with universal and sparse weights.

We then move on to the problem of analysing the Jacobian spectrum of (the post-activations of) Φ_L. Defining $\mathbf{J}_{L,\mathbf{x}} := \mathrm{d}(\varphi_L \circ \Phi_{L-1})_\mathbf{x}$, we're interested in the macroscopic behaviour of the squared singular values of $\mathbf{J}_{L,\mathbf{x}}$, and study the empirical spectral distribution of $\mathbf{J}_{L,\mathbf{x}}\mathbf{J}_{L,\mathbf{x}}^\top$, defined as

$$\rho_L := \frac{1}{N}\sum_{I=1}^{N} \delta_{\xi_i} \tag{2.8}$$

where $\{\xi_1, \dots, \xi_N\}$ are the eigenvalues of $\mathbf{J}_{L,\mathbf{x}}\mathbf{J}_{L,\mathbf{x}}^\top$ and δ_{ξ_i} denotes a Dirac mass on ξ_i.

This distribution plays a crucial role in network stability, as shown by Glorot and Bengio [41], who demonstrated that when its second moment is close to 1, it prevents gradient explosion and vanishing, thereby maintaining stability and leading to significantly faster learning in practice ([102, 103]). The measure is known to converge to a deterministic limiting measure as $N \to \infty$, following the works of Collins and Hayase [18] and Pastur [97] for orthogonal and Gaussian weights, respectively. These works justify an asymptotic freeness assumption, under which Pennington et al. [103] first identified the limit of ρ_L using the analytic machinery of free probability. The result of Pastur [97] was then extended to non-Gaussian weights in a recent work of Pastur and Slavin [98].

All of these works characterise the limiting measure through an implicit functional equation for its Stieltjes transform. This allows one to compute the first two moments $m_{1,L}$ and $m_{2,L}$ by expanding and solving for coefficients (see Equation (21) in Pennington et al. [103]), but this approach breaks down for higher moments.

Our main result (theorem 2.3) computes the entire sequence of moments for the first time, by expressing them as weighted sums over non-crossing partitions. As with our previous two results, we also show that these formulae hold for non-Gaussian and sparse weights.

Theorem 2.3 (Moments of the limiting spectral distribution of the Jacobian)**.** *For each $\ell \geq 0$, assume that $N_\ell = N$ and that W_ℓ has i.i.d. entries drawn from a symmetric, centered distribution with finite moments and variance $1/N$. Let $\mathbf{x} \in \mathbb{R}^N$ be such that $\frac{1}{N}\langle \mathbf{x}, \mathbf{x}\rangle \to x^2 \in \mathbb{R}$, and define ρ_L as in eq. (2.8).*

Then for every $k \geq 0$,

$$\int t^k \mathrm{d}\rho_L(t) \underset{N\to\infty}{\longrightarrow} m_{k,L}(x)$$

where $\{m_{k,L}(x)\}_{k,L}$ is defined by the recursion in eq. (2.53).

Furthermore, the same result holds if the weight matrices in (2.1) are replaced by $\tilde{W}_\ell := W_\ell \odot \frac{1}{\sqrt{p_N}} B_\ell$, where W_ℓ are as above and the B_ℓ are independent matrices with i.i.d., Bernoulli distributed entries with parameter p_N satisfying $Np_N \to \infty$.

Proof. The claim for W_ℓ with Gaussian entries is theorem 2.19, and the corollaries in chapter A generalize this to the non-Gaussian and sparse cases. □

The recursion of eq. (2.53) sheds light on the relationship between the moments $m_{k,L}$ and the Fuss-Catalan numbers $\mathrm{FC}_L^{(k)}$ (see [93]): they are obtained by inserting activation-dependent coefficients in the latter's defining recursion (see section 2.5)[2]. The limit of ρ_L can thus be seen as a non-linear generalisation the Fuss-Catalan or Raney distribution, which is known to be the universal first-order limit of squared singular values for products of Ginibre matrices (in the language of free probability, it is the L-fold free multiplicative convolution of the Marchenko-Pastur law).

2.1.1 Overview of our method.

Graph expansions of neural networks.

Having outlined in chapter 1 our correspondence between analytic and graphical operations, we now show how to express neural networks and related quantities using these graphs. The connection is made by expanding the output of neural networks at a given input $\mathbf{x}$ as a linear combination of product graphs

$$\Phi(\mathbf{x}) = \sum_{G \in \mathcal{F}} c(G) \mathbf{W}_G \tag{2.9}$$

for some family of graphs $\mathcal{F}$ and combinatorial factors $c(G)$.

This is achieved in theorem 2.6, which generalizes Faà di Bruno's formula (see [19, 61]) to the case of an arbitrary number of compositions. In similar tasks, trees have been shown to be a natural combinatorial tool to keep track of terms (see the literature on Butcher series [47, 83], and, more generally, on Runge-Kutta methods for ordinary differential equations [48]), and this is reflected here in the fact that $\mathcal{F}$ (in eq. (2.9)) turns out to be a set of rooted trees.

[2] when expressed as a sum over non-crossing partitions of k elements.

By applying the graphical rules of chapter 1 to each term in this sum, we derive similar expansions for various related quantities, namely the k-th coordinate of $\Phi(\mathbf{x})$, the neural tangent kernel, and the trace of the input-output Jacobian of Φ times its transpose, raised to an arbitrary power. The moment method then reduces the study of the asymptotic distribution of these quantities to computing $\mathbb{E}\mathbf{W}_G$ for various operator graphs G.

2.1.2 Possible generalizations

We imposed two simplifying assumptions in order to better illustrate our method (polynomial activations and the absence of biases in (2.1)). We briefly discuss how these can be lifted below.

Random biases

Typically, one defines Φ_L as

$$\Phi_0(\mathbf{x}) = W_0\mathbf{x}, \quad \Phi_{\ell+1}(\mathbf{x}) = W_{\ell+1}\varphi_{\ell+1}(\Phi_\ell(\mathbf{x})) + b_{\ell+1}$$

for some vectors $\{b_\ell\}_\ell$ (the *biases*) which can also be initialized at random. The method presented here can be adapted to this setting, albeit at the cost of additional combinatorial bookkeeping (namely a more complicated set of graphs $\mathcal{F}$ in eq. (2.9)). When applying Wick's theorem to compute the resulting $\mathbb{E}\mathbf{W}_G$, one would also have to pair vertices first, before being able to pair edges as explained in section 1.1.1.

Non-polynomial activations

Restricting to polynomial activations will allow us to ignore issues of convergence in theorem 2.6. That said, the expansion in the latter formally holds for any (reasonably well-behaved) function, and extending the results here to more commonly used activations (e.g. Sigmoid, ReLU) should thus be feasible by an approximation argument as was done by Benigni and Péché [3].

Symbol	Meaning
Φ_L	Neural network of depth L (Eqn. 2.10).
$\mathbf{J}_{L,\mathbf{x}}$	Neural network Jacobian $\mathbf{J}_{L,\mathbf{x}} = \mathrm{d}(\varphi_L \circ \Phi_{L-1})_{\mathbf{x}}$ (Eqn. 2.17).
$\mathbb{T}_L$	Input-independent operator trees space for NN expansion (Sec. 2.2).
$\mathbb{T}_L(\mathbf{x})$	Input-dependent operator trees space for NN expansion (Def. 2.5).
$(\mathbb{T}_L(\mathbf{x}))^*$	Input-dependent operator trees with pruned root (Sec. 2.2.1).
$\partial_{\mathbf{x}}\mathbb{T}_L(\mathbf{x})$	Space of derivative operator trees with respect to input $\mathbf{x}$ (Sec. 2.2.1).
$\partial_\ell\mathbb{T}_L(\mathbf{x})$	Space of derivative operator trees with respect to matrix W_ℓ (Sec. 2.2.1).

Table 2.1: *Notation and symbols used throughout this chapter.*

2.2 Tree Expansion of Feed Forward Neural Networks

Fix sequences $(\varphi_\ell : \mathbb{R} \to \mathbb{R} \mid \ell \in \mathbb{N}_{>0})$ of polynomial *activation functions*, $(N_\ell \in \mathbb{N}_{>0} \mid \ell \in \mathbb{N})$ of *layer dimensions* and $(W_\ell \in \mathbb{R}^{N_{\ell+1}\times N_\ell} \mid \ell \in \mathbb{N})$ of weight matrices. Recall that a *feed-forward neural network* Φ_L of depth L with no bias term is defined by the recursion

$$\Phi_0(\mathbf{x}) = W_0 x, \quad \Phi_{\ell+1}(\mathbf{x}) = W_{\ell+1}\varphi_{\ell+1}(\Phi_\ell(\mathbf{x})). \tag{2.10}$$

As explained in the introduction, we aim to derive an expansion for $\Phi_L(\mathbf{x})$ that linearizes the effect of the activation functions.

This is achieved in the main result of the section (theorem 2.6), in which we expand Φ_L as a sum of operator graphs which are rooted trees (see theorem 2.4). Some examples of the trees arising in this expansion are plotted in fig. 2.1: they all have weight matrices as edge inputs, basis vectors as leaf inputs and their internal nodes all have input $\mathbf{1}$. Each of these trees has a unique free vertex (its root) which is an out-vertex, and the latter always has out-degree one.

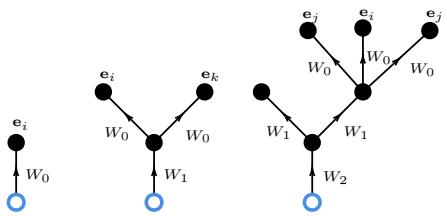


Figure 2.1: *Examples of trees arising in the expansion of* Φ_3

Consider k such trees $\tau_1, \ldots, \tau_k$ (along with the data required to make them operator graphs), assuming that their respective root edges are all fixed to the same $W_{\ell-1}$ for

some $\ell > 1$ (which implies that the dimension assigned to each of the root vertices is N_ℓ). We define $[\tau_1 \cdots \tau_k]_\ell$ to be the tree $(\circ\xrightarrow{W_\ell}\bullet)\circ(\wedge_{i=1}^k \tau_i)$, where $\circ$ denotes composition as defined in section 1.2.2 and where for $\circ\xrightarrow{W_\ell}\bullet$ the dimension of the vertices is given by $\mathfrak{d}(\bullet) = N_\ell, \mathfrak{d}(\circ) = N_{\ell+1}$ (recalling that empty products in $\mathbb{R}^N$ are valued as $\mathbf{1}_N$, we have $[\ ]_\ell = \circ\xrightarrow{W_\ell}\bullet$). Then using this bracket operation, any tree like those in fig. 2.1 can be constructed from the following family of operator graphs:

$$\{[\bullet_{\mathbf{e}_i}]_0\}_{i\in[N_0]}, \text{ where } [\bullet_{\mathbf{e}_i}]_0 = \circ\xrightarrow{W_0}\bullet_{\mathbf{e}_i} \text{ so that } \mathfrak{d}(\bullet) = N_0, \mathfrak{d}(\circ) = N_1,$$

where $\mathfrak{d}$ is the dimension function for $[\bullet_{\mathbf{e}_i}]_0$. For instance, the first and second tree in fig. 2.1 can be written as $[\bullet_{\mathbf{e}_i}]_0$ and $[[\bullet_{\mathbf{e}_i}]_0[\bullet_{\mathbf{e}_j}]_0]_1$.

Definition 2.4. *Fix $L > 0$, $\mathbb{T}_L$ is the set of operator graphs defined inductively by*

$$\mathbb{T}_0 = \{[\bullet_{\mathbf{e}_1}]_0, \ldots, [\bullet_{\mathbf{e}_{N_0}}]_0\}, \quad \mathbb{T}_{\ell+1} := \{[\tau_1 \cdots \tau_M]_{\ell+1} \mid M \geq 0,\ \forall i \in [M]\ \tau_i \in \mathbb{T}_\ell\}.$$

Remark. *Note that for any $\tau \in \mathbb{T}_\ell$, there is a unique dimension function $\mathfrak{d}$ that is compatible with its inputs: if an edge (u, v) is labeled by W_k then we must have $\mathfrak{d}(u) = N_{k+1}$ and $\mathfrak{d}(v) = N_k$. In what follows, we will implicitly assume this choice of dimension function for all of the trees being considered.*

By definition of the bracket operation $[\,\cdot\,]$ and the graph operations $\wedge$ and $\circ$ (see theorems 1.5 and 1.7), we have that

$$\mathbf{W}_{[\tau_1\cdots\tau_M]_{\ell+1}} = W_{\ell+1}(\odot_{i=1}^M \mathbf{W}_{\tau_i})$$

for every $\tau = [\tau_1 \cdots \tau_M]_{\ell+1} \in \mathbb{T}_{\ell+1}$.

Note that by construction, the trees in $\mathbb{T}_L$ are *non-plane*, meaning that the order of the edges coming out of each vertex is ignored. For any $\tau \in \mathbb{T}_\ell$, it will thus be useful to keep track in how many ways it can be obtained from plane trees (where the order of edges is not ignored). Following Gubinelli [47], we call this quantity the *symmetry factor* $s(\tau)$ of τ, and define it recursively as

$$s([\bullet_{\mathbf{e}_i}]_0) = 1, \quad s([(\tau_1)^{k_1} \cdots (\tau_m)^{k_m}]_\ell) = \prod_i (k_i!) s(\tau_i)^{k_i}, \tag{2.11}$$

where $\tau_1, \ldots, \tau_m$ are assumed to be distinct and $\tau_1^{k_1}$ denotes the repetition of τ_1, k_1 times. Lastly, we introduce the following bit of notation: for any vector $\mathbf{x}$, we let

$$\mathbf{x}_{[\bullet_{\mathbf{e}_i}]_0} = [\mathbf{x}]_i, \quad \mathbf{x}_{[\tau_1\cdots\tau_M]_\ell} = \prod_{i=1}^M \mathbf{x}_{\tau_i} \in \mathbb{R}, \tag{2.12}$$

and for any activation function φ,

$$\varphi_{[\bullet_{\mathbf{e}_i}]_0} = 1, \quad \varphi_{[\tau_1 \cdots \tau_M]_{\ell+1}} = \varphi_{\ell+1}^{(M)}(0) \cdot \prod_{i=1}^{M} \varphi_{\tau_i} \in \mathbb{R}, \tag{2.13}$$

where $\varphi^{(M)}$ is the M-th derivative of φ.

Definition 2.5. *Given $\mathbf{x} \in \mathbb{R}^{N_0}$ we define the sets $\mathbb{T}_\ell(\mathbf{x})$ of operator graphs obtained from the trees in $\mathbb{T}_\ell$ by changing all the leaves' inputs from $\bullet_{\mathbf{e}_i}$ to $\mathbf{x}$. We will henceforth denote this operation by $\bullet_{\mathbf{e}_i} \mapsto \bullet_{\mathbf{x}}$.*

If we extend $s(\eta)$ and φ_η to $\mathbb{T}_\ell(\mathbf{x})$ setting the base cases as $s([\bullet_{\mathbf{x}}]_0) = \varphi_{[\bullet_{\mathbf{x}}]_0} = 1$, then clearly $\varphi_\tau = \varphi_{\tau(\mathbf{x})}$ for all $\tau \in \mathbb{T}_\ell$ since the definition of φ only depends on the out-degree of the vertices, which is left unchanged. While it is not generally true that $s(\tau) = s(\tau(\mathbf{x}))$, we nonetheless have the useful equality for $\eta \in \mathbb{T}_\ell(\mathbf{x})$

$$\frac{1}{s(\eta)} \mathbf{W}_\eta = \sum_{\tau \in \mathbb{T}_\ell : \tau(\mathbf{x}) = \eta} \frac{\mathbf{x}_\tau}{s(\tau)} \mathbf{W}_\tau. \tag{2.14}$$

This follows from the fact that $s(\tau(\mathbf{x}))/s(\tau)$ counts the number of ways in which the non-plane tree τ is obtained from $\tau(\mathbf{x})$ by applying the operation $\bullet_{\mathbf{x}} \mapsto \bullet_{\mathbf{e}_i}$.

Example. *Consider the trees $\eta := [[\bullet_{\mathbf{x}}]_0\ [\bullet_{\mathbf{x}}]_0]_1$ and $\tau := [[\bullet_1]_0\ [\bullet_2]_0]_1$. Then $\eta = \tau(\mathbf{x})$, but $s(\eta) = 2! s([\bullet_{\mathbf{x}}]_0) = 2$ while $s(\tau) = s([\bullet_1]_0) s([\bullet_2]_0) = 1$. Notice that τ can be obtained both by $[[\bullet_{\mathbf{x}}]_0\ [\bullet_{\mathbf{x}}]_0]_1 \mapsto [[\bullet_1]_0\ [\bullet_2]_0]_1$ and $[[\bullet_{\mathbf{x}}]_0\ [\bullet_{\mathbf{x}}]_0]_1 \mapsto [[\bullet_2]_0\ [\bullet_1]_0]_1$, so in $2 = s(\eta)/s(\tau)$ ways.*

We can now state the main result of this section.

Theorem 2.6 (Tree expansion)**.** *Let Φ_L be as defined in eq. (2.10). Then for any $\mathbf{x} \in \mathbb{R}^{N_0}$, $\Phi_L(\mathbf{x})$ admits the following expansion over $\mathbb{T}_\ell$:*

$$\Phi_\ell(\mathbf{x}) = \sum_{\tau \in \mathbb{T}_\ell} \frac{\varphi_\tau}{s(\tau)} \mathbf{x}_\tau \mathbf{W}_\tau = \sum_{\eta \in \mathbb{T}_\ell(\mathbf{x})} \frac{\varphi_\eta}{s(\eta)} \mathbf{W}_\eta \in \mathbb{R}^{N_{\ell+1}}. \tag{2.15}$$

Remark. *For $\ell = 1$ this expansion reduces to the classical (vector valued) Taylor expansion since $s([[\bullet_{\mathbf{e}_1}]_0^{k_1} \cdots [\bullet_{\mathbf{e}_d}]_0^{k_d}]_1) = \prod_i (k_i)!$.*

Remark. *Despite the space $\mathbb{T}_\ell$ being infinite, this sum is always well-defined since only finitely many φ_τ are non-zero (having assumed that the φ are polynomials).*

Proof. We proceed by induction on depth ℓ for the first equality. If $\ell = 0$ then $\Phi_0(\mathbf{x}) = W_0 \mathbf{x}$ and

$$\sum_{\tau \in \mathbb{T}_0} \frac{\varphi_\tau \mathbf{x}_\tau}{s(\tau)} \mathbf{W}_\tau = \sum_{i=1}^{N_0} \frac{1 \cdot [\mathbf{x}]_i}{1} W_0 \mathbf{e}_i = W_0 \left(\sum_{i=1}^{N_0} [\mathbf{x}]_i \mathbf{e}_i \right) = W_0 \mathbf{x} \in \mathbb{R}^{N_1}$$

as needed.

Now assume then the claim holds for ℓ. For $\mathbf{y} \in \mathbb{R}^{N_\ell}$, we write $\varphi_{\ell+1}(\mathbf{y})$ to mean $\varphi_{\ell+1}$ applied componentwise to $\mathbf{y}$. Then by Taylor expansion around the origin,

$$\varphi_{\ell+1}(\mathbf{y}) = \sum_{M=0}^{\infty} \frac{\varphi_{\ell+1}^{(M)}(0)}{M!} \mathbf{y}^{\odot M}$$

where $\odot$ denotes the Hadamard product. Note that this sum is actually finite, since $\varphi_{\ell+1}$ is assumed to be a polynomial.

If $y = \sum_{\tau \in \mathbb{T}_\ell} \frac{\varphi_\tau}{s(\tau)} \mathbf{W}_\tau \mathbf{x}_\tau$ then theorem B.3 yields

$$y^{\odot M} = \sum_{\bar{\tau} = [\![\tau_1 \cdots \tau_M]\!] \in \mathbb{X}_{\mathbb{T}_\ell}^M} \frac{M!}{\mathfrak{s}(\bar{\tau})} \odot_{i=1}^M \left(\frac{\varphi_{\tau_i}}{s(\tau_i)} \mathbf{W}_{\tau_i} \mathbf{x}_\tau \right).$$

Here, $\mathbb{X}_{\mathbb{T}_\ell}^M$ denotes the set of cardinality-M multisets of elements of $\mathbb{T}_\ell$, and the $\mathfrak{s}(\bar{\tau})$ are symmetry factors defined as

$$\mathfrak{s}([\![]\!]) = 1, \quad \mathfrak{s}([\![(\eta_1)^{k_1} \cdots (\eta_m)^{k_m}]\!]) := \prod_{i=1}^m (k_i)!.$$

($[\![(\eta_1)^{k_1} \cdots (\eta_m)^{k_m}]\!]$ is the multiset containing $\eta_i \in \mathbb{T}_\ell$ with multiplicity k_i.) Letting $\mathbb{X}_{\mathbb{T}_\ell} = \cup_{M \geq 0} \mathbb{X}_{\mathbb{T}_\ell}^M$, we have a natural bijection between $\mathbb{T}_{\ell+1}$ and $\mathbb{X}_{\mathbb{T}_\ell}$ given by

$$\tau := [\tau_1 \cdots \tau_m]_{\ell+1} \mapsto \bar{\tau} := [\![\tau_1 \cdots \tau_m]\!].$$

Put simply, we identify a rooted non-plane tree in $\mathbb{T}_{\ell+1}$ with the multiset of subtrees ($\in \mathbb{T}_\ell$) rooted at its depth-1 vertices. Under this identification, $s(\tau) = \mathfrak{s}(\bar{\tau}) \prod_i s(\tau_i)$ and we conclude that

$$\begin{aligned}
W_{\ell+1} \varphi_{\ell+1}(\mathbf{y}) &= W_{\ell+1} \sum_{M=0}^{\infty} \frac{\varphi^{(M)}(0)}{M!} \mathbf{y}^{\odot M} \\
&= \sum_{M=0}^{\infty} \varphi^{(M)}(0) \sum_{\bar{\tau} = [\![\tau_1 \cdots \tau_M]\!] \in \mathbb{X}_{\mathbb{T}_\ell}^M} \frac{1}{\mathfrak{s}(\bar{\tau})} W_{\ell+1} \left[\odot_{i=1}^M \left(\frac{\varphi_{\tau_i}}{s(\tau_i)} \mathbf{W}_{\tau_i} \mathbf{x}_{\tau_i} \right) \right] \\
&= \sum_{M=0}^{\infty} \sum_{\bar{\tau} = [\![\tau_1 \cdots \tau_M]\!] \in \mathbb{X}_{\mathbb{T}_\ell}^M} \frac{\varphi^{(M)}(0) \prod_i \varphi_{\tau_i}}{\mathfrak{s}(\bar{\tau}) \prod_i s(\tau_i)} \left(W_{\ell+1} \left[\odot_{i=1}^M \mathbf{W}_{\tau_i} \right] \right) \left(\prod_i \mathbf{x}_{\tau_i} \right) \\
&= \sum_{\tau \in \mathbb{T}_{\ell+1}} \frac{\varphi_\tau \mathbf{x}_\tau}{s(\tau)} \mathbf{W}_\tau \in \mathbb{R}^{N_{\ell+1}}.
\end{aligned}$$

The second equality in the theorem statement readily follows from

$$\sum_{\eta \in \mathbb{T}_\ell(\mathbf{x})} \frac{\varphi_\eta}{s(\eta)} \mathbf{W}_\eta = \sum_{\eta \in \mathbb{T}_\ell(\mathbf{x})} \sum_{\tau \in \mathbb{T}_\ell : \tau(\mathbf{x}) = \eta} \frac{\varphi_\tau}{s(\tau)} \mathbf{W}_\tau \mathbf{x}_\tau = \sum_{\tau \in \mathbb{T}_\ell} \frac{\varphi_\tau}{s(\tau)} \mathbf{W}_\tau \mathbf{x}_\tau.$$

□

2.2.1 Expansions for related quantities

Combined with the dictionary in section 1.2.2, this expansion allows us to express the coordinates of Φ_L and various related quantities as linear combinations of values of product graphs. Note that in the figures below, we label an edge with input W_ℓ by ℓ for simplicity.

Entry-wise expansion

We use theorem 2.6 to derive an expansion for the k-th entry of $\Phi_L(\mathbf{x})$ (in the standard basis). To that end, let $\mathbb{T}_{\ell,k}$ be the set of trees obtained from $\mathbb{T}_\ell$ by fixing the root to $\mathbf{e}_k$ (*i.e.* applying $\circ \mapsto \bullet_{\mathbf{e}_k}$), noting that this makes the tree a product graph. Then we can write the k-th entry of Φ_L as

$$[\Phi_L(\mathbf{x})]_k = \sum_{\tau \in \mathbb{T}_{L,k}} \frac{\varphi_\tau}{s(\tau)} \mathbf{W}_\tau \mathbf{x}_\tau = \sum_{\eta \in \mathbb{T}_{L,k}(\mathbf{x})} \frac{\varphi_\eta}{s(\eta)} \mathbf{W}_\eta \in \mathbb{R}. \tag{2.16}$$

where $s(\tau), \varphi_\tau$ and $\mathbf{x}_\tau$ are defined as in equations (2.11),(2.12) and (2.13), as these definitions only depend on τ through its set of vertices and edges.

Input-output Jacobian

Consider the (post-activation) Jacobian of Φ_L with respect to an input $\mathbf{x}$, which we define as the matrix representation $\mathbf{J}_{L,\mathbf{x}} \in \mathbb{R}^{N_L \times N_0}$ of

$$\mathrm{d}(\varphi_L \circ \Phi_{L-1})_{\mathbf{x}}. \tag{2.17}$$

Note that, following the same arguments as in the proof of theorem 2.6, we can write

$$(\varphi_L \circ \Phi_{L-1})(\mathbf{x}) = \sum_{\eta \in \mathbb{T}_L(\mathbf{x})} \frac{\varphi_{\eta^*}}{s(\eta^*)} \mathbf{W}_{\eta^*} \in \mathbb{R}^{N_{\ell+1}} \tag{2.18}$$

where η^* denotes the tree obtained from η by deleting the root and the edge which stems from it, and where φ_{η^*} and $s(\eta^*)$ are equal to φ_η and $s(\eta)$, respectively. This operation is depicted in fig. 2.2, and let $(\mathbb{T}_L(\mathbf{x}))^* = \{\eta^* : \eta \in \mathbb{T}_L(\mathbf{x})\}$.

We now take the derivative. For any tree $\eta \in (\mathbb{T}(\mathbf{x}))^*$, we want to see it as a multi-linear map in the $\mathbf{x}$'s and compute the total derivative of said map at $(\mathbf{x}, \dots, \mathbf{x})$. To be precise, let k denote the number of vertices in η with input $\mathbf{x}$, and η_i denote the tree obtained from η by applying $\bullet_{\mathbf{x}} \mapsto \circ$ to the i-th such vertex (respectively). Then

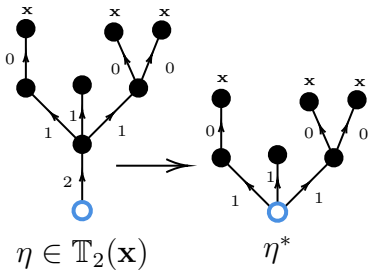


Figure 2.2: *A tree $\eta \in \mathbb{T}_2(\mathbf{x})$ (left) and η^* (right).*

by theorem 1.10,

$$(\mathrm{d}\eta)_{\mathbf{x}} = \sum_{i=1}^{k} \eta_i$$

and, arguing similarly as we did to get (2.14) yields

$$\frac{1}{s(\eta)} \mathbf{W}_{(\mathrm{d}\eta)_{\mathbf{x}}} = \sum_{\eta^* : \eta^*(\mathbf{x}) = \eta} \frac{1}{s(\eta^*)} \mathbf{W}_{\eta^*}. \tag{2.19}$$

This equality holds since $s(\eta^*)/s(\eta)$ is equal to the number of ways in which η^* can be obtained from a tree in $(\partial_{\mathbf{x}} \mathbb{T}_L(\mathbf{x}))^*$ as defined below, which is equal to the number of times it appears in the total derivative of η.

We let $\partial_{\mathbf{x}} \mathbb{T}_L(\mathbf{x})$ denote the space of trees which can be obtained from trees in $\mathbb{T}_L(\mathbf{x})$ by applying $\bullet_{\mathbf{x}} \mapsto \circ$ to exactly one of their leaves. Summarizing what we have said so far, we have

$$\mathrm{d}(\varphi_L \circ \Phi_{L-1})_{\mathbf{x}} = \sum_{\eta \in (\partial_{\mathbf{x}} \mathbb{T}_L(\mathbf{x}))^*} \frac{\varphi_\eta}{s(\eta)} \mathbf{W}_\eta, \tag{2.20}$$

noting that φ_η is invariant under $\bullet_{\mathbf{x}} \mapsto \circ$.

Recall that $\mathbf{J}_{L,\mathbf{x}}$ denotes the matrix representation of $\mathrm{d}(\varphi_L \circ \Phi_{L-1})_{\mathbf{x}}$. Later on in the chapter, we will study the distribution of the singular values of $\mathbf{J}_{L,\mathbf{x}}$ when the weight matrices of Φ_L are chosen at random. This will amount to estimating the elements of the sequence

$$\mathrm{Tr}\Big((\mathbf{J}_{L,\mathbf{x}} \mathbf{J}_{L,\mathbf{x}}^\top)^k\Big), \quad k \in \mathbb{N}.$$

We note that every η in the expansion of $\mathbf{J}_{L,\mathbf{x}}$ has exactly two free vertices: its root ($\circ$) and one of its leaves ($\circ$). The transpose of $\mathbf{J}_{L,\mathbf{x}}$ can thus be obtained from the same expansion by taking $\mathbf{W}_{\eta^\top}$ instead of $\mathbf{W}_\eta$ (noting that φ_η and $s(\eta)$ are invariant under $\eta \mapsto \eta^\top$). We can thus write

$$\mathbf{J}_{L,\mathbf{x}} \mathbf{J}_{L,\mathbf{x}}^\top = \sum_{\eta_1, \eta_2 \in (\partial_{\mathbf{x}} \mathbb{T}_L(\mathbf{x}))^*} \frac{\varphi_{\eta_1} \varphi_{\eta_2}}{s(\eta_1) s(\eta_2)} \mathbf{W}_{\eta_1 \circ \eta_2^\top} \tag{2.21}$$

which by linearity of the trace gives

$$\mathrm{Tr}(\mathbf{J}_{L,\mathbf{x}}\mathbf{J}_{L,\mathbf{x}}^\top) = \sum_{\eta_1,\eta_2\in(\partial_{\mathbf{x}}\mathbb{T}_L(\mathbf{x}))^*} \frac{\varphi_{\eta_1}\varphi_{\eta_2}}{s(\eta_1)s(\eta_2)}\mathbf{W}_{\mathrm{Tr}(\eta_1\circ\eta_2^\top)}. \tag{2.22}$$

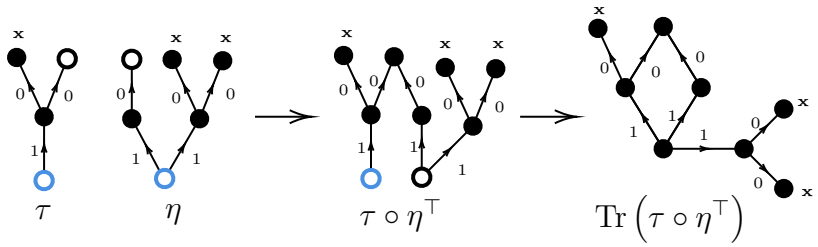


Figure 2.3: *Operator graphs arising in the expansion of* $\mathrm{Tr}(\mathbf{J}_{2,\mathbf{x}}\mathbf{J}_{2,\mathbf{x}}^\top)$.

Similarly, for $k = 2$ we can repeat the above procedure to obtain

$$\sum_{\eta_1^{(1)},\eta_2^{(1)},\eta_1^{(2)},\eta_2^{(2)}\in(\partial_{\mathbf{x}}\mathbb{T}_L(\mathbf{x}))^*} \left(\prod_{i=1}^{2}\prod_{j=1}^{2}\frac{\varphi_{\eta_j^{(i)}}}{s(\eta_j^{(i)})}\right)\mathbf{W}_{\mathrm{Tr}\left(\eta_1^{(1)}\circ(\eta_2^{(1)})^\top\circ\eta_1^{(2)}\circ(\eta_2^{(2)})^\top\right)}, \tag{2.23}$$

and the case for general k follows analogously, with the final expression involving product graphs of the form $\mathrm{Tr}\left(\eta_1^{(1)}\circ(\eta_2^{(1)})^\top\circ\cdots\circ\eta_1^{(k)}\circ(\eta_2^{(k)})^\top\right)$. This will be discussed in detail in section 2.5, and is depicted in fig. 2.8.

Neural tangent kernel

The same procedure can be used to compute the NTK $\Theta_L(\mathbf{x},\mathbf{y}) \in \mathbb{R}^{N_{L+1}\times N_{L+1}}$, which we recall is the matrix representation of

$$\sum_{\ell=0}^{L}\lambda_\ell^2(\mathrm{d}\Phi_L(\mathbf{x}))_{W_\ell}\circ(\mathrm{d}\Phi_L(\mathbf{y}))_{W_\ell}^\top, \tag{2.24}$$

for layer-wise learning rates $(\lambda_\ell)_\ell \in \mathbb{R}$.

To find an expansion for Θ_L, let $\partial_\ell\mathbb{T}_L(\mathbf{x})$ denote the space of trees which can be obtained from trees in $\mathbb{T}_L(\mathbf{x})$ by freeing exactly one of its height ℓ edges (i.e one edge with input W_ℓ), making it an in-edge. We can then write

$$(\mathrm{d}\Phi_L(\mathbf{x}))_{W_\ell} = \sum_{\tau\in\partial_\ell\mathbb{T}_L(\mathbf{x})}\frac{\varphi_\tau}{s(\tau)}\mathbf{W}_\tau, \tag{2.25}$$

where $\mathbf{W}_\tau : \mathbb{R}^{N_{\ell+1}\times N_\ell} \to \mathbb{R}^{N_{L+1}}$, noting that the $s(\tau)$ take care of any overcounting. Arguing as we did above, we then find that

$$[(\mathrm{d}\Phi_L(\mathbf{x}))_{W_\ell}\circ(\mathrm{d}\Phi_L(\mathbf{y}))_{W_\ell}^\top]_{ij} = \sum_{\tau\in\partial_\ell\mathbb{T}_{L,i}(\mathbf{x})}\sum_{\eta\in\partial_\ell\mathbb{T}_{L,j}(\mathbf{y})}\frac{\varphi_\tau\varphi_\eta}{s(\tau)s(\eta)}\mathbf{W}_{\tau\circ\eta^\top}. \tag{2.26}$$

2.2.2 Atomic pairings of trees

We begin with the following useful lemma, which highlights important features of atomic pairings of edges of trees. It will be used in the proof of theorems 2.12 and 2.16 below.

In what follows, for any tree τ, we refer to edges whose head is the root as *root edges* and denote them by $r(\tau)$. Given any $e \in E(\tau)$, we also let $\tau|_e$ denote the sub-tree of τ rooted at e's head. This new tree $\tau|_e$ then has a unique root edge (which is e), and a sub-tree rooted at e's tail. Lastly, recall that admissible pairings only pair edges in $E_{\mathcal{W}}$ (namely edges with random matrices as inputs), and that an $\ell-$edge in $E_{\mathcal{W}}$ is one that has been fixed to the random matrix $W_\ell \in \mathcal{W}$.

Lemma 2.7. *Let τ be a connected, directed graph with all vertices having in-degree 1, except for the root $r_\tau \in V(\tau)$ which has in-degree 0 (so that τ is a directed tree). Then for any $\phi \in \mathcal{P}_{\mathrm{A}}(\tau)$, the following holds:*

1. *ϕ only pairs edges which are equidistant from the root.*
2. *Let e, e' be arbitrary edges in τ. If $\{e, e'\} \in \phi$, then ϕ is an atomic pairing of the edges in $\tau|_e \wedge \tau|_{e'}$ (the tree formed by identifying the roots of $\tau|_e$ and $\tau|_{e'}$).*
3. *For any $k \geq 0$, let $e_0, \dots, e_k$ and $e'_0, \dots, e'_k$ be two sequences of edges forming paths of length $k+1$ which start at r_τ (so that e_i, e'_i are at distance i from the root and e_0, e'_0 are root edges). Then*

$$\{e_k, e'_k\} \in \phi \Rightarrow \{e_i, e'_i\} \in \phi \text{ or } e_i = e'_i \text{ for all } 0 \leq i \leq k.$$

Proof. We begin by proving that root edges have to be paired with other root edges: consider a root edge $e = (r_\tau, v) \in E_{\mathcal{W}}(\tau)$ and assume that it is paired to a non-root edge $e' = (u', v')$ by ϕ.

Assume first that $e' \in E_{\mathcal{W}}(\tau|_e)$ with $u' \neq r_\tau$. Then consider the image in $(\tau)_\phi$ of the path from r_τ to u' in τ: it starts and ends at the same vertex of $(\tau)_\phi$, while only traversing the (undirected) edge corresponding to $\{e, e'\}$ once (*cf.* fig. 2.4 left). This is a cycle, which contradicts the fact that $\phi \in \mathcal{P}_A(\tau)$. Now assume that $e' \in E_{\mathcal{W}}(\tau) \setminus E_{\mathcal{W}}(\tau|_e)$ with $u' \neq r_\tau$. By what we have just shown, there must be another root edge $\tilde{e} = (r_\tau, \tilde{v})$ such that $e' \in E_{\mathcal{W}}(\tau|_{\tilde{e}})$, and which is not paired to any edge in $E_{\mathcal{W}}(\tau|_{\tilde{e}})$. Arguing as above, the image of the path from r_τ to u' is a cycle in $(\tau)_\phi$ traversing the image of $\tilde{e}$ only once (*cf.* fig. 2.4 (middle)), giving the desired contradiction.

We now prove (2). Assume that $\{e, e'\} \in \phi$, and for contradiction that $\{\varepsilon, \varepsilon'\} \in \phi$ for some $\varepsilon \in E_{\mathcal{W}}(\tau|_e)$ (which is $\neq e$) and $\varepsilon' \in E_{\mathcal{W}}(\tau) \setminus E(\tau|_e \sqcup \tau|_{e'})$. Then consider the path which starts at the head of ε', goes through r_τ without using edges in $\tau|_e \sqcup \tau|_{e'}$ and ends at the head of ε using the edge e (*cf.* fig. 2.4 right). This path's image in

$(\tau)|_\phi$ is then once again a cycle since it traverses $\{e, e'\}$ only once, we conclude that $\phi|_{\tau|_e \sqcup \tau|_{e'}} \in \mathcal{P}_A(\tau|_e \wedge \tau|_{e'})$

To prove (3) we proceed by induction. Assume that $\{e_k, e'_k\} \in \phi$. If $e_{k-1} = e'_{k-1}$, there is nothing to prove. Otherwise, assume that $e_{k-1} \notin E_{\mathcal{W}}(\tau)$, so that it is not paired by ϕ. Then the path which starts at the head of e_k, goes to r_τ through e_{k-1} and ends at the head of e'_k via e'_{k-1} is a cycle in $(\tau)_\phi$. We must therefore have $e_{k-1}, e'_{k-1} \in E_{\mathcal{W}}(\tau)$.

Now if these edges were not paired together by ϕ, then $\{e_k, e'_k\} \notin \phi$ by (2), giving us a contradiction. This proves (3) by induction on k.

Finally (1) follows from (3) by induction on the distance from the root, using the fact that root edges have to be paired with other root edges. □

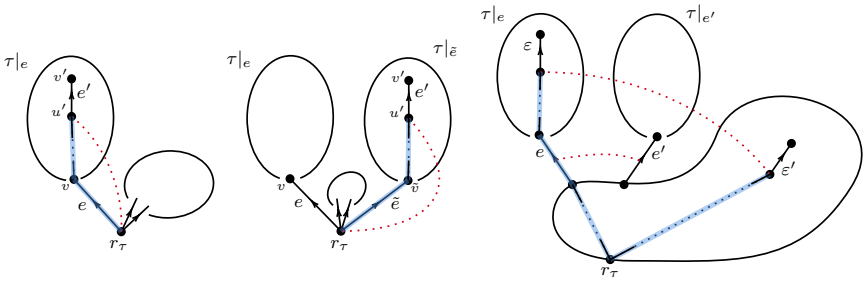


Figure 2.4: *Paths in τ which are cycles in $(\tau)_\phi$ in the proof of theorem 2.7. Red dotted lines denote vertices/edges which are the same in $(\tau)_\phi$, and paths highlighted in blue are cycles in the latter.*

2.3 Gaussian Process Limit

Assumption 2.8. *The input dimension N_0 of Φ_L is fixed, and $N_\ell = N$ for all other layers $\ell > 0$. Assume also that $\sigma_0 = 1$ and $\sigma_\ell = N_\ell^{-1/2}$ for $\ell \geq 1$,*

Theorem 2.9. *Under theorem 2.8, for any $M, L \geq 1$ we have*

$$([\Phi_L]_1, \ldots, [\Phi_L]_M) \xrightarrow[N\to\infty]{d} \mathcal{GP}(0, K_L \otimes \mathrm{Id}_M) \tag{2.27}$$

where the right hand side is a Gaussian Process indexed on $\mathbb{R}^{N_0}$, with diagonal covariance function defined by

$$K_0(\mathbf{x}, \mathbf{y}) = \langle \mathbf{x}, \mathbf{y} \rangle_{\mathbb{R}^{N_0}}, \quad K_{\ell+1}(\mathbf{x}, \mathbf{y}) = \mathbb{E}\left[\varphi_{\ell+1}(X_\ell)\varphi_{\ell+1}(Y_\ell)\right] \tag{2.28}$$

$$(X_\ell, Y_\ell) \sim \mathcal{N}\left(0, \begin{bmatrix} K_\ell(\mathbf{x}, \mathbf{x}) & K_\ell(\mathbf{x}, \mathbf{y}) \\ K_\ell(\mathbf{y}, \mathbf{x}) & K_\ell(\mathbf{y}, \mathbf{y}) \end{bmatrix}\right). \tag{2.29}$$

To prove this, we first consider the coordinate-wise tree expansion of Φ_L given in eq. (2.16):

$$[\Phi_L(\mathbf{x})]_k = \sum_{\eta \in \mathbb{T}_{L,k}(\mathbf{x})} \frac{\varphi_\eta}{s(\eta)} \mathbf{W}_\eta \in \mathbb{R}.$$

Finding a distributional limit of Φ_L using this expansion thus amounts to finding the limiting distribution of

$$\Big(\mathbf{W}_\eta \mid \eta \in \cup_{i=1}^M \mathbb{T}_{L,k_i}(\mathbf{x}_i)\Big)$$

as $N \to \infty$, for any fixed sequences $(\mathbf{x}_i \in \mathbb{R}^{N_0} \mid i \in [M])$ and $(k_i \in [N_{L+1}] \mid i \in [M])$.

The mixed moments of the $\mathbf{W}_\eta$ can be computed using theorem 1.17, which yields

$$\mathbb{E}\Big\{\prod_{i=1}^m \Big(\mathbf{W}_{\eta_i} - \sum_{\phi \in \mathcal{P}_{\mathrm{A}}(\eta_i)} \mathbf{W}_{(\eta_i)_\phi}\Big)\Big\} = \sum_{\phi \in \mathcal{P}_{\mathrm{AF}}(\sqcup_i \eta_i)} \mathbf{W}_{(\sqcup_i \eta_i)_\phi}. \tag{2.30}$$

To identify the leading order terms on the right-hand side, we must therefore identify the pairings $\phi \in \mathcal{P}_{\mathrm{AF}}(\sqcup_i \eta_i)$ which maximize $\mathbf{W}_{(\sqcup_i \eta_i)_\phi}$. This is done in the following lemma.

Lemma 2.10. *Under theorem 2.8, for any $L \geq 0$ and fixed sequences $(\mathbf{x}_i \in \mathbb{R}^{N_0} \mid i \in [M])$, $(k_i \in [N_{L+1}] \mid i \in [M])$ and $(\eta_i \in \mathbb{T}_{L,k_i}(\mathbf{x}_i) \mid i \in [M])$, we have*

$$\forall i \in [M], \quad \mathcal{P}_{\mathrm{A}}(\eta_i) = \emptyset, \text{ and thus } \sum_{\psi \in \mathcal{P}_{\mathrm{A}}(\eta_i)} \mathbf{W}_{(\eta_i)_\psi} = 0, \tag{2.31}$$

$$\sum_{\phi \in \mathcal{P}_{\mathrm{AF}}(\sqcup_i \eta_i)} \mathbf{W}_{(\sqcup_i \eta_i)_\phi} = \sigma_{\sqcup_i \eta_i} N^{\frac{1}{2}(|E(\sqcup_i \eta_i)| - |\mathcal{L}_0(\sqcup_i \eta_i)|)} \Big(\alpha(\sqcup_i \eta_i) + \mathcal{O}\Big(\frac{1}{N}\Big)\Big), \tag{2.32}$$

where $\mathcal{L}_0(\sqcup_i \eta_i)$ denotes the set of leaves of $\sqcup_i \eta_i$ which are endpoints of 0-labeled edges,

$$\alpha(\sqcup_i \eta_i) := \sum_{\phi \in \mathcal{P}_{\mathrm{B}}(\sqcup_i \eta_i)} \delta_\phi \langle \mathbf{x}, \mathbf{x} \rangle_\phi$$

where δ_ϕ is equal to one if roots paired by ϕ are fixed to the same $\mathbf{e}_i$ and zero otherwise, and

$$\langle \mathbf{x}, \mathbf{x} \rangle_\phi := \prod_{\{\bullet_{\mathbf{x}_i}, \bullet_{\mathbf{x}_j}\} \in (\mathcal{L}_0(\sqcup_i \eta_i)/\sim_\phi)} \langle \mathbf{x}_i, \mathbf{x}_j \rangle_{\mathbb{R}^{N_0}}.$$

Proof. For the first assertion, we just have to prove that given any $\eta \in \mathbb{T}_{\ell,k}(\mathbf{x})$ one has $\mathcal{P}_{\mathrm{A}}(\eta) = \emptyset$. This is clear, since the tree η only has one $\ell-$edge (namely, the one stemming from its root), which then cannot be paired with any other edge.

For the second assertion, recall that the elements of $\mathcal{L}_0(\sqcup_i \eta_i)$ are the only cells in $C(\sqcup_i \eta_i)$ which are tails of 0-labeled edges. This means that any $\phi \in \mathcal{P}(\sqcup_i \eta_i)$ identifies these vertices in pairs as shown in fig. 2.5. It follows that for any pair of vertices $\bullet_{\mathbf{x}_i}, \bullet_{\mathbf{x}_j} \in \mathcal{L}_0(\sqcup_i \eta_i)$, if we let G denote the product graph obtained from $\sqcup_i \eta_i$ by deleting

these vertices along with their adjacent edges (see fig. 2.5), then

$$\begin{aligned}\mathbf{W}_{(\sqcup_i \eta_i)_\phi} &= \sum_{\alpha,\beta} [\mathbf{W}_{G_\phi}]_\alpha [\sigma_0^2 \mathbf{I}]_{\alpha,\beta} [\mathbf{x}_i \odot \mathbf{x}_j]_\beta \\ &= \sigma_0^2 \sum_{\alpha,\beta} [\mathbf{W}_{G_\phi}]_\alpha [\mathbf{x}_i]_\beta [\mathbf{x}_j]_\beta = \sigma_0^2 \langle \mathbf{x}_i, \mathbf{x}_j \rangle_{\mathbb{R}^{N_0}} \mathbf{W}_{G_\phi}. \end{aligned} \tag{2.33}$$

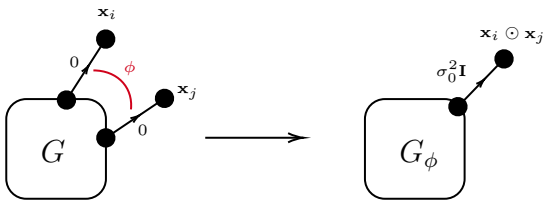


Figure 2.5: *Any admissible edge pairing induces a pairing of leaves.*

A similar argument holds for every pair of roots (heads of $\ell-$edges) in $\sqcup_i \eta_i$, of which there are $c(\sqcup_i \eta_i)$ many. Repeating (2.33) for each pair of leaves and roots then yields the following equality for any $\phi \in \mathcal{P}(\sqcup_i \eta_i)$:

$$\mathbf{W}_{(\sqcup_i \eta_i)_\phi} = \sigma_0^{|\mathcal{L}_0(\sqcup_i \eta_i)|} \sigma_\ell^{c(\sqcup_i \eta_i)} \langle \mathbf{x}, \mathbf{x} \rangle_\phi \delta_\phi \mathbf{W}_{G_\phi^1} \tag{2.34}$$

where G^1 is obtained from $\sqcup_i \eta_i$ by removing all leaves, roots and their adjacent edges, and

$$\langle \mathbf{x}, \mathbf{x} \rangle_\phi := \prod_{\{\bullet_{\mathbf{x}_i}, \bullet_{\mathbf{x}_j}\} \in (\mathcal{L}_0(\sqcup_i \eta_i)/\sim_\phi)} \langle \mathbf{x}_i, \mathbf{x}_j \rangle_{\mathbb{R}^{N_0}}.$$

Note that here, δ_ϕ arises from the roots in the same way as $\langle \mathbf{x}, \mathbf{x} \rangle_\phi$ did from the leaves, except that since the roots all have inputs $\mathbf{e}_i$ for some i, the resulting inner product must either equal zero or one.

Further, note that under theorem 2.8, G^1 is a product graph satisfying theorem 1.19, in particular a product graph with all vertices having dimension $\mathfrak{d} = N$, fixed to $\mathbf{1}_N$ and with all edges fixed to random Gaussian matrices ($E \setminus E_{\mathcal{W}}(G^1) = \emptyset$). We can thus leverage the genus expansion results of section 1.3.2 to compute the values $\mathbf{W}_{G_\phi}$ explicitly.

Note that any $\phi \in \mathcal{P}_{\mathrm{AF}}(\sqcup_i \eta_i)$ must induce an atom-free partition of the edges of G^1, since each of the removed edges shares a vertex with G^1. By theorem 1.21 we know that $\mathbf{W}_{G_\phi^1}$ is maximized when this induced partition is bi-atomic, in which case

$$\mathbf{W}_{G_\phi^1} = \sigma_{G^1} N^{(|E(G^1)| + c(G^1))/2}.$$

This yields the claim, since $\sigma_{\sqcup_i \eta_i} = \sigma_0^{|\mathcal{L}_0(\sqcup_i \eta_i)|} \sigma_\ell^{c(\sqcup_i \eta_i)} \sigma_{G^1}$,

$$|E(G^1)| = |E(\sqcup_i \eta_i)| - |\mathcal{L}_0(\sqcup_i \eta_i)| - c(\sqcup_i \eta_i), \quad c(G^1) = c(\sqcup_i \eta_i),$$

and

$$\{\psi \in \mathcal{P}_{\mathrm{B}}(G^1) : \psi \text{ is induced by some } \phi = \mathcal{P}_{\mathrm{B}}(\sqcup_i \eta_i)\} = \mathcal{P}_{\mathrm{B}}(G^1).$$

□

Here, $\alpha(\sqcup_i \eta_i)$ is essentially a weighted version of the cardinality of $\mathcal{P}_{\mathrm{B}}(\sqcup_i \eta_i)$. Applying this lemma to (2.30) gives

$$\mathbb{E}\left\{\prod_{i=1}^{m} \mathbf{W}_{\eta_i}\right\} = \sigma_{\sqcup_i \eta_i} N^{\frac{1}{2}(|E(\sqcup_i \eta_i)| - |\mathcal{L}_0(\sqcup_i \eta_i)|)} \left(\alpha(\sqcup_i \eta_i) + \mathcal{O}\Big(\frac{1}{N}\Big)\right). \tag{2.35}$$

Remark. *To have a well-defined distributional limit for Φ_L, the σ_ℓ must be chosen in a way that makes the above quantity $\mathcal{O}(1)$ for all $\eta_i \in \mathbb{T}_{L,k_i}$. This is equivalent to requiring that $\sigma_\eta N^{\frac{1}{2}(|E(\eta)| - |\mathcal{L}_0(\eta)|)} = \mathcal{O}(1)$ for all $L, k \geq 1$ and $\eta \in \mathbb{T}_{L,k}$, which is in turn equivalent to*

$$\sigma_0 = \mathcal{O}(1) \text{ and } \sigma_\ell = \mathcal{O}(N_\ell^{-1/2}) \ \forall \ell \geq 1. \tag{2.36}$$

The canonical choice would then be the one for which $\sigma_\eta N^{\frac{1}{2}(|E(\eta)| - |\mathcal{L}_0(\eta)|)} = 1$, which explains our theorem 2.8 and is known as the "GP limit" parametrisation (see [68]).

The proof of theorem 1.22 (replacing $|\mathcal{P}_{\mathrm{B}}(\sqcup_i \eta_i)|$ with $\alpha(\sqcup_i \eta_i)$) then gives the following limit.

Proposition 2.11. *Assume that theorem 2.8 holds. Then for any $M, L \geq 1$ and fixed sequences $(\mathbf{x}_i \in \mathbb{R}^{N_0} \mid i \in [M])$, $(k_i \in [N_{L+1}] \mid i \in [M])$ and $(\eta_i \in \mathbb{T}_{L,k_i}(\mathbf{x}_i) \mid i \in [M])$ we have*

$$\left(\mathbf{W}_{\eta_i} \;\middle|\; i \in [M]\right) \xrightarrow[N\to\infty]{d} \left(Z_{\eta_i} \;\middle|\; i \in [M]\right) \overset{d}{=} \mathcal{GP}(0, C) \tag{2.37}$$

where the right-hand side denotes a Gaussian process indexed on $(\eta_i \in \mathbb{T}_{L,k_i}(\mathbf{x}_i) \mid i \in [M])$ with covariance $C(\eta, \nu) = \alpha(\eta \sqcup \nu)$, $\overset{d}{=}$ denotes equality in distribution and α is defined as in the statement of theorem 2.10.

Lastly, we will need the following decomposition of the covariances $\alpha(\eta \sqcup \nu)$.

Lemma 2.12. *Let α be as in theorem 2.10 and fix $\ell \geq 1$. For any $\tau = [\tau_1 \cdots \tau_n]_\ell \in \mathbb{T}_\ell(\mathbf{x}_1)$ and $\eta = [\eta_1 \cdots \eta_m]_\ell \in \mathbb{T}_\ell(\mathbf{x}_2)$, let $\tau^{(i)}$ denote the tree obtained from τ by fixing its unique out-vertex to $\mathbf{e}_i$, and similarly for $\eta^{(j)}$. Then we have that*

$$\alpha(\tau^{(i)} \sqcup \eta^{(j)}) = \delta_{ij} \sum_{\pi} \prod_{\{\pi_1, \pi_2\} \in \pi} \alpha(\pi_1 \sqcup \pi_2)$$

where the sum is over all pairings π of the elements of $(\tau_1^{(i)}, \ldots, \tau_n^{(i)}, \eta_1^{(j)}, \ldots, \eta_m^{(j)})$.

Proof. Let $\phi \in \mathcal{P}_{\mathrm{B}}(\tau^{(i)} \sqcup \eta^{(j)})$ be arbitrary. Denote by $r(\tau)$ (resp. $r(\eta)$) denote the unique edge whose head is the root of τ (resp. η). Then $r(\tau)$ and $r(\eta)$ are the only

$\ell-$edges in their respective trees, and must thus be paired by ϕ. Since ϕ must then pair $\ell-1$-labeled edges with each other (which are precisely the edges $\{r(\tau_k^{(i)})\}_{k=1}^n$, $\{r(\eta_k^{(j)})\}_{k=1}^m$) this induces a pairing π of the trees in

$$(\tau_1^{(i)}, \dots, \tau_n^{(i)}, \eta_1^{(j)}, \dots, \eta_m^{(j)})$$

via the bijection $\tau_k^{(i)} \mapsto r(\tau_k^{(i)})$ (and similarly for the $\eta_k^{(j)}$).

By theorem 2.7, (1) and (2), for each $\{\pi_1, \pi_2\} \in \pi$, ϕ contains a bi-atomic pairing of the edges of $\pi_1 \sqcup \pi_2$, which we denote by $\phi\Big|_{\pi_1 \sqcup \pi_2}$ $(\in \mathcal{P}_{\mathrm{B}}(\pi_1 \sqcup \pi_2))$. In other words, we can decompose ϕ as

$$\phi = \Big\{\{r(\tau), r(\eta)\}\Big\} \bigsqcup_{\{\pi_1,\pi_2\}\in\pi} \phi|_{\pi_1\sqcup\pi_2}, \tag{2.38}$$

and the lemma then follows since we can write

$$\langle \mathbf{x}, \mathbf{x}\rangle_\phi = \prod_{\{\pi_1,\pi_2\}\in\pi} \langle \mathbf{x}, \mathbf{x}\rangle_{\phi|_{\pi_1\sqcup\pi_2}}.$$

□

We are now ready to prove theorem 2.9.

Proof of theorem 2.9. It follows directly from theorem 2.11 and eq. (2.16) that

$$([\Phi_L]_1, \dots, [\Phi_L]_M) \xrightarrow{d} \mathcal{GP}(0, K_L \otimes \mathrm{Id}_M)$$

where for any $\ell \geq 0$, K_ℓ is defined by

$$K_\ell(\mathbf{x}, \mathbf{y}) = \sum_{\eta_{\mathbf{x}}\in\mathbb{T}_{\ell,1}(\mathbf{x})} \sum_{\eta_{\mathbf{y}}\in\mathbb{T}_{\ell,1}(\mathbf{y})} \frac{\varphi_{\eta_{\mathbf{x}}}\varphi_{\eta_{\mathbf{y}}}}{s(\eta_{\mathbf{x}})s(\eta_{\mathbf{y}})} \alpha(\eta_{\mathbf{x}} \sqcup \eta_{\mathbf{y}}) \in \mathbb{R}. \tag{2.39}$$

What's left is to show that this definition of K_ℓ coincides with the one in the theorem statement. We proceed by induction on ℓ. The base case $\ell = 0$ follows from $\mathbb{T}_{0,i}(\mathbf{x}) = \{ \mathrm{e}_i \overset{W_0}{\bullet\!\!-\!\!\bullet} \mathbf{x} \}$ and $\alpha(\mathrm{e}_i \overset{W_0}{\bullet\!\!-\!\!\bullet} \mathbf{x} \sqcup \mathrm{e}_i \overset{W_0}{\bullet\!\!-\!\!\bullet} \mathbf{y}) = \langle \mathbf{x}, \mathbf{y}\rangle_{\mathbb{R}^{N_0}}$.

For the inductive step, we assume that the claim holds for ℓ and apply theorem B.4 with $\mathcal{A}_1 = \mathbb{T}_{\ell,1}(\mathbf{x})$, $\mathcal{A}_2 = \mathbb{T}_{\ell,1}(\mathbf{y})$ and

$$\lambda(\eta, \nu) = \frac{\varphi_\eta\varphi_\nu}{s(\eta)s(\nu)}\alpha(\eta \sqcup \nu).$$

Note that $\mathbb{T}_{\ell+1,1}(\mathbf{x})$ and $\mathbb{X}_{\mathcal{A}_1}$ are in bijection via $\eta := [\eta_1 \cdots \eta_m]_{\ell+1} \mapsto \bar{\eta} := [\![\eta_1 \cdots \eta_m]\!]$ and under this identification we have

$$s(\eta) = \mathfrak{s}(\bar{\eta}) \prod_i s(\eta_i), \quad \varphi_\eta = \varphi_{\ell+1}^{(m)} \prod_i \varphi_{\eta_i}.$$

theorem 2.12 thus gives

$$\alpha(\eta \sqcup \nu) = \sum_{\pi \in \mathcal{P}(\bar{\eta} \sqcup \bar{\nu})} \prod_{\{\pi_1, \pi_2\} \in \pi} \alpha(\pi_1 \sqcup \pi_2),$$

and this completes the proof since

$$\frac{\varphi_\eta \varphi_\nu}{s(\eta) s(\nu)} \alpha(\eta \sqcup \nu) = \frac{\varphi_{\bar{\eta}} \varphi_{\bar{\nu}}}{\mathfrak{s}(\bar{\eta}) \mathfrak{s}(\bar{\nu})} \sum_{\pi \in \mathcal{P}(\bar{\eta} \sqcup \bar{\nu})} \prod_{\{\pi_1, \pi_2\} \in \pi} \lambda(\pi_1, \pi_2).$$

□

2.4 Neural Tangent Kernel

Recall that $\Theta_L(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{N_{L+1} \times N_{L+1}}$ is the matrix representation of

$$\sum_{\ell=0}^{L} \lambda_\ell^2 (\mathrm{d}\Phi_L(\mathbf{x}))_{W_\ell} \circ (\mathrm{d}\Phi_L(\mathbf{y}))_{W_\ell}^\top, \tag{2.40}$$

for $(\lambda_\ell^2)_{\ell \geq 0}$. We prove convergence of $\Theta_L(\mathbf{x}, \mathbf{y})$ to its limit under the following conditions.

Assumption 2.13. *Assume that the input and output dimensions N_0 and N_{L+1} of Φ_L are fixed, and that $N_\ell = N$ for all other layers $0 < \ell \leq L$. Assume also that $\sigma_0 = 1$ and $\sigma_\ell = N^{-\frac{1}{2}}$ for $\ell > 0$ (the "GP limit" initialization).*

Theorem 2.14. *Under theorem 2.13 and if $\lambda_\ell^2 = \sigma_\ell^2 \mathbf{1}(\ell > 0) + \mathbf{1}(\ell = 0)$, we have*

$$\mathbb{E}\left\{ \left\| \Theta_L(\mathbf{x}, \mathbf{y}) - \Theta_L^\infty(\mathbf{x}, \mathbf{y}) \otimes \mathrm{Id}_{N_{L+1}} \right\|^2 \right\} = \mathcal{O}\left(\frac{N_{L+1}^2}{N} \right), \tag{2.41}$$

where

$$\Theta_0^\infty(\mathbf{x}, \mathbf{y}) = \langle \mathbf{x}, \mathbf{y} \rangle_{\mathbb{R}^{N_0}}, \quad \Theta_L^\infty(\mathbf{x}, \mathbf{y}) = K_L(\mathbf{x}, \mathbf{y}) + \dot{K}_L(\mathbf{x}, \mathbf{y}) \Theta_{L-1}^\infty(\mathbf{x}, \mathbf{y}) \tag{2.42}$$

and $\dot{K}_\ell$ is defined in the same way as K_ℓ but substituting φ_ℓ for φ_ℓ' in (2.28).

Recall that the i, j-th entry of the NTK can be expressed by the following graph expansion (see eq. (D.19))

$$\sum_{\ell=0}^{L} \lambda_\ell^2 [(\mathrm{d}\Phi_L(\mathbf{x}))_{W_\ell} \circ (\mathrm{d}\Phi_L(\mathbf{y}))_{W_\ell}^\top]_{ij} = \sum_{\ell=0}^{L} \sum_{\substack{\tau \in \partial_\ell \mathbb{T}_{L,i}(\mathbf{x}) \\ \eta \in \partial_\ell \mathbb{T}_{L,j}(\mathbf{y})}} \frac{\varphi_\tau \varphi_\eta}{s(\tau) s(\eta)} \lambda_\ell^2 \mathbf{W}_{\tau \circ \eta^\top}. \tag{2.43}$$

We once again use theorem 1.17 to study the joint distribution of the product graphs on the right-hand side, from which the following lemma follows.

Lemma 2.15. *Under theorem 2.13, for any $\tau \in \partial_\ell \mathbb{T}_{L,i_1}(\mathbf{x}_1)$ and $\eta \in \partial_\ell \mathbb{T}_{L,i_2}(\mathbf{x}_2)$, we have*

$$\mathbb{E}\left\{\left(\sigma_\ell^2 \mathbf{W}_{\tau\circ\eta^\top} - a(\tau\circ\eta^\top)\right)^2\right\} = \mathcal{O}\left(\frac{1}{N}\right), \tag{2.44}$$

where[3]

$$a(\tau\circ\eta^\top) := \sum_{\psi\in\mathcal{P}_{\mathrm{A}}(\tau\circ\eta^\top)} \delta_{i_1 i_2}\langle \mathbf{x},\mathbf{x}\rangle_\psi. \tag{2.45}$$

Proof. By theorem 1.17 we know, setting $G := \tau\circ\eta^\top$, that

$$\mathbb{E}\left\{\left(\mathbf{W}_G - \sum_{\psi\in\mathcal{P}_{\mathrm{A}}(G)} \mathbf{W}_{G_\psi}\right)^2\right\} = \sum_{\phi\in\mathcal{P}_{\mathrm{AF}}(G\sqcup G)} \mathbf{W}_{(G\sqcup G)_\phi}. \tag{2.46}$$

We begin by proving the claim for $0 < \ell < L \neq 0$. We will be reasoning as we did to prove theorem 2.10, namely by analyzing the structure of a pairing in $\mathcal{P}_{\mathrm{A}}(G)$. To that end, note that the elements of $\mathcal{L}_0(G)$ are the only cells of G which are tails of 0-labeled edges, while the roots of τ and η are the only heads of L-labeled edges.

It follows that for any $\psi \in \mathcal{P}_{\mathrm{A}}(G)$, these vertices will necessarily all get paired, yielding

$$\mathbf{W}_{G_\psi} = \sigma_0^{|\mathcal{L}_0(G)|}\sigma_L^2\langle \mathbf{x},\mathbf{x}\rangle_\psi \delta_{i_1,i_2}\mathbf{W}_{G^1_\psi}$$

where the product graph G^1 is obtained from G by removing all leaves, roots and their adjacent vertices. Under theorem 2.13, G^1 is then a product graph satisfying theorem 1.19: all of its vertices have dimension $\mathfrak{d} = N$ and are fixed to $\mathbf{1}_N$, all edges except one are fixed to random Gaussian matrices, the remaining one is fixed to $\mathbf{I}$. We thus apply the genus expansion formula (section 1.3.2) with

$$|E_{\mathcal{W}}(G^1)| = |E(G^1)| - 1 = |E(G)| - (|\mathcal{L}_0(G)| + 2) - 1,$$

$$|E \setminus E_{\mathcal{W}}(G^1)| = 1,$$

and since any atomic partition of G induces an atomic one on G^1, we conclude that

$$\mathbf{W}_{G^1_\psi} = \sigma_{G^1} N^{\frac{1}{2}(|E(G)|-|\mathcal{L}_0(G)|-3)+2},$$
$$\sum_{\psi\in\mathcal{P}_{\mathrm{A}}(G)} \mathbf{W}_{G_\psi} = \sigma_G N^{\frac{1}{2}(|E(G)|-|\mathcal{L}_0(G)|+1)} a(G).$$

The same reasoning applied to $\mathcal{P}_{\mathrm{AF}}(G\sqcup G)$ gives

[3] Recall how i_1 and i_2 are fixed when choosing τ and η, and stand for the $\mathbf{e}_i$s at their roots.

$$\mathbf{W}_{(G\sqcup G)_\phi} = \sigma_G^2 N^{|E(G)|-|\mathcal{L}_0(G)|}\langle \mathbf{x}, \mathbf{x}\rangle_\phi \delta_\phi \qquad (\text{for } \phi \in \mathcal{P}_{\mathrm{B}}(G \sqcup G)),$$
$$\sum_{\phi \in \mathcal{P}_{\mathrm{AF}}(G\sqcup G)} \mathbf{W}_{(G\sqcup G)_\phi} = \sigma_G^2 N^{|E(G)|-|\mathcal{L}_0(G)|}\left(\alpha(G \sqcup G) + \mathcal{O}\Big(\frac{1}{N}\Big)\right).$$

We now turn to the cases when $\ell = 0$ or $\ell = L$, noting that the pruned graphs do not contain any $\mathbf{I}$ this time. In the $\ell = L$ case, we get

$$\mathbf{W}_{G_\psi} \asymp \sigma_G N^{\frac{1}{2}(|E(G)|-(|\mathcal{L}_0(G)|+1))+1} = \sigma_G N^{\frac{1}{2}(|E(G)|-|\mathcal{L}_0(G)|+1)},$$
$$\mathbf{W}_{(G\sqcup G)_\phi} \asymp \sigma_G^2 N^{\frac{1}{2}(2|E(G)|-2(|\mathcal{L}_0(G)|+1))+1} = \sigma_G^2 N^{|E(G)|-|\mathcal{L}_0(G)|},$$

respectively, while for $\ell = 0$ we get

$$\mathbf{W}_{G_\psi} \asymp \sigma_G N^{\frac{1}{2}(|E(G)|-(|\mathcal{L}_0(G)|+2))+1} = \sigma_G N^{\frac{1}{2}(|E(G)|-|\mathcal{L}_0(G)|)},$$
$$\mathbf{W}_{(G\sqcup G)_\phi} \asymp \sigma_G^2 N^{\frac{1}{2}(2|E(G)|-2(|\mathcal{L}_0(G)|+2))+1} = \sigma_G^2 N^{|E(G)|-|\mathcal{L}_0(G)|-1}.$$

Putting everything together, for any ℓ and under theorem 2.8 (in which case $\sigma_G = N^{-\frac{1}{2}(|E(G)|-|\mathcal{L}_0(G)|-\mathbf{1}(\ell>0))}$), we always obtain

$$\mathbb{E}\left\{\Big(\sigma_\ell^2 \mathbf{W}_G - a(G)\Big)^2\right\} = \frac{1}{N}\left(\alpha(G \sqcup G) + \mathcal{O}\Big(\frac{1}{N}\Big)\right). \tag{2.47}$$

□

Remark. *This implies that we need to take $\lambda_\ell^2 = \sigma_\ell^2$ to get a well-defined limit, as assumed in theorem 2.14. This is known as the "NTK parametrization" (see Jacot et al. [58]).*

As with theorem 2.12 in the previous section, we will need a combinatorial decomposition of the limiting covariances $a(\tau \circ \eta^\top)$ in order to prove theorem 2.14 by induction.

Lemma 2.16. *For $L \geq 1$, there is a canonical bijection*

$$\partial_\ell \mathbb{T}_{L,i}(\mathbf{x}) \to \mathbb{T}_{L,i}(\mathbf{x}) \times \partial_\ell \mathbb{T}_{L-1,1}(\mathbf{x}), \quad \eta \mapsto (\eta^+, \eta^-)$$

under which

$$a(\tau \circ \eta^\top) = \alpha(\tau^+ \sqcup \eta^+)a(\tau^- \circ \eta^{-\top}), \quad s(\tau^+)s(\tau^-) = s(\tau), \quad \varphi_\tau = \dot{\varphi}_{\tau^+}\varphi_{\tau^-}$$

where $\dot{\varphi}_{[\tau_1 \cdots \tau_n]_L} := \varphi_L^{(n+1)} \prod_i \varphi_{\tau_i}$ and a is defined as in the statement of theorem 2.15. Further, if $\ell = L$, we have $\partial_\ell \mathbb{T}_{L-1,1}(\mathbf{x}) := \{\bullet\}$ so that $\partial_\ell \mathbb{T}_{L,i}(\mathbf{x}) \simeq \mathbb{T}_{L,i}(\mathbf{x})$ and $a(\tau \circ \eta^\top) = \alpha(\tau^+ \sqcup \eta^+)$.

Proof. Each tree in $\mathbb{T}_{L,i}(\mathbf{x})$ has exactly one $L-$edge, namely the unique edge adjacent to its root. If $\ell = L$, the canonical bijection in the lemma's statement is the map which assigns to $\eta \in \partial_L \mathbb{T}_{L,i}(\mathbf{x})$ the tree η^+ obtained by fixing η's root edge to W_L. Since each bi-atomic pairing of $\tau^+ \sqcup \eta^+$ has to necessarily pair the root edges, it follows that $\mathcal{P}_{\mathrm{B}}(\tau^+ \sqcup \eta^+)$ and $\mathcal{P}_{\mathrm{A}}(\tau \circ \eta^\top)$ are in canonical bijection, and in particular that $a(\tau \circ \eta^\top) = \alpha(\tau^+ \sqcup \eta^+)$.

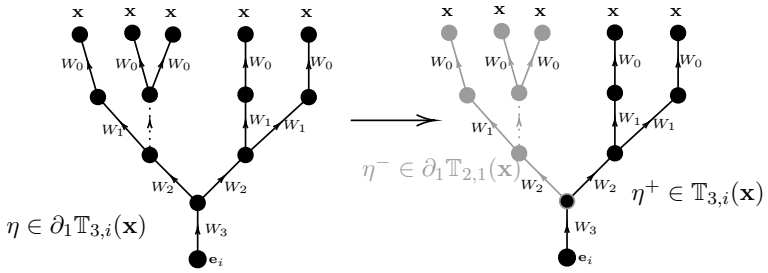


Figure 2.6: *Decomposition of a tree in* $\partial_1 \mathbb{T}_{3,i}$

If $\ell < L$, then any $\eta \in \partial_\ell \mathbb{T}_{L,i}(\mathbf{x})$ can be viewed as a tree $\tau^- \in \partial_\ell \mathbb{T}_{L-1,1}(\mathbf{x})$, which is attached by its root to a tree $\tau^+ \in \mathbb{T}_{L,i}(\mathbf{x})$ as in the example of fig. 2.6. Note that this decomposition is unique. It thus remains to show that $\mathcal{P}_{\mathrm{A}}(\tau \circ \eta^\top)$ is in bijection with $\mathcal{P}_{\mathrm{B}}(\tau^+ \sqcup \eta^+) \times \mathcal{P}_{\mathrm{A}}(\tau^- \circ (\eta^-)^\top)$.

Let $M \in \mathbb{R}^{N \times N}$ be a random matrix which is independent from the $\{W_\ell\}_\ell$, and is populated with i.i.d. standard Gaussian random variables. Consider the trees $\tau[M]$, $\eta[M]$, which we recall are obtained from τ and η by inserting M into their unique free in-edge, respectively.

Clearly, $\mathcal{P}_{\mathrm{A}}(\tau \circ \eta^\top)$ is then in bijection with $\mathcal{P}_{\mathrm{A}}(\tau[M] \wedge \eta[M])$, since any pairing in the latter must necessarily pair the edges given M as input, as well as the root edges $r(\tau[M])$ and $r(\eta[M])$ (since they are the only L-labeled edges).

We can now use theorem 2.7, (3). Fix any pairing $\phi \in \mathcal{P}_{\mathrm{A}}(\tau[M] \wedge \eta[M])$. In $\tau[M]$, the path starting at the root and ending at the head of the unique M-labeled edge must be paired by ϕ with the analogous path in $\eta[M]$, meaning that edges which are equidistant from the respective roots are paired with each other. In particular, this will end up pairing $r(\tau^-[M])$ with $r(\eta^-[M])$.

Furthermore, by theorem 2.7, (2), ϕ will only pair edges of $\tau^-[M]$ with edges of $\eta^-[M]$ (and similarly for τ^+, η^+ as a consequence). This implies that ϕ can be decomposed into the union of some $\phi^- \in \mathcal{P}_{\mathrm{A}}(\tau^-[M] \wedge \eta^-[M])$ with some $\phi^+ \in \mathcal{P}_{\mathrm{A}}(\tau^+ \wedge \eta^+)$.

Repeating this argument for the unique $\tilde{\phi} \in \mathcal{P}_{\mathrm{A}}(\tau \circ \eta^\top)$ corresponding to ϕ, we conclude that $\mathcal{P}_{\mathrm{A}}(\tau \circ \eta^\top)$ is in bijection with

$$\mathcal{P}_{\mathrm{A}}(\tau^+ \sqcup \eta^+) \times \mathcal{P}_{\mathrm{A}}(\tau^- \circ (\eta^-)^\top),$$

and it follows that

$$a(\tau \circ \eta^\top) = \alpha(\tau^+ \sqcup \eta^+) a(\tau^- \circ \eta^{-\top}).$$

□

We can now derive the rate of convergence of the NTK to its limit.

Proof of theorem 2.14. Let $i, j \in \{1, \dots, N_{L+1}\}$ be arbitrary, and recall from eq. (2.43) that

$$\lambda_\ell^2 [(\mathrm{d}\Phi_L(\mathbf{x}))_{W_\ell} \circ (\mathrm{d}\Phi_L(\mathbf{y}))_{W_\ell}^\top]_{ij} = \sum_{\tau \in \partial_\ell \mathbb{T}_{L,i}(\mathbf{x})} \sum_{\eta \in \partial_\ell \mathbb{T}_{L,j}(\mathbf{y})} \frac{\varphi_\tau \varphi_\eta}{s(\tau) s(\eta)} \lambda_\ell^2 \mathbf{W}_{\tau \circ \eta^\top}$$

for any $0 \leq \ell \leq L$. Note that the number of terms in this sum is finite, and crucially does not depend on N nor N_0, N_{L+1}. The same is true for the factors $\varphi_\tau, \varphi_\eta$ and $s(\tau), s(\eta)$. Thus theorem 2.15 and repeated applications of the elementary inequality $(a+b)^2 \leq 2(a^2+b^2)$ yield

$$\mathbb{E}\left\{ \left(\sum_{\tau,\eta} \frac{\varphi_\tau \varphi_\eta}{s(\tau) s(\eta)} a(\tau \circ \eta^\top) - \sum_{\tau,\eta} \frac{\varphi_\tau \varphi_\eta}{s(\tau) s(\eta)} \lambda_\ell^2 \mathbf{W}_{\tau \circ \eta^\top} \right)^2 \right\} = \mathcal{O}\left(\frac{1}{N} \right), \tag{2.48}$$

where the implicit constant depends on L.

For $\ell = L$, theorem 2.16 then gives

$$\sum_{\tau,\eta} \frac{\varphi_\tau \varphi_\eta}{s(\tau) s(\eta)} a(\tau \circ \eta^\top) = \sum_{\tau^+, \eta^+} \frac{\varphi_{\tau^+} \varphi_{\eta^+}}{s(\tau^+) s(\eta^+)} \alpha(\tau^+ \sqcup \eta^+)$$

which is in turn equal to $\delta_{ij} K_L(\mathbf{x}, \mathbf{y})$ by the proof of theorem 2.9. Otherwise, noting that $s(\tau) = s(\tau^+) s(\tau^-)$ and $\varphi_\tau = \dot{\varphi}_{\tau^+} \varphi_{\tau^-}$, we instead have

$$\begin{aligned} &\sum_{\tau,\eta} \frac{\varphi_\tau \varphi_\eta}{s(\tau) s(\eta)} a(\tau \circ \eta^\top) \\ &= \sum_{\tau^+, \eta^+} \sum_{\tau^-, \tilde{\eta}} \frac{\dot{\varphi}_{\tau^+} \dot{\varphi}_{\eta^+}}{s(\tau^+) s(\eta^+)} \frac{\varphi_{\tau^-} \varphi_{\eta^-}}{s(\tau^-) s(\eta^-)} \alpha(\tau^+ \sqcup \eta^+) a(\tau^- \circ \eta^{-\top}) \end{aligned}$$

which is equal to

$$\delta_{ij} \dot{K}_L(\mathbf{x}, \mathbf{y}) \lim_{N \to \infty} \sigma_\ell^2 [(\mathrm{d}\Phi_{L-1}(\mathbf{x}))_{W_\ell} (\mathrm{d}\Phi_{L-1}(\mathbf{y}))_{W_\ell}^\top]_{1,1},$$

(once again by the proof of theorem 2.9).

It follows that $[\Theta_L(\mathbf{x}, \mathbf{y})]_{ij}$ converges in L^2 to $[\Theta_L^\infty(\mathbf{x}, \mathbf{y}) \otimes \mathrm{Id}_{N_{L+1}}]_{ij}$ at a rate of $\mathcal{O}(1/N)$ for each $i, j \in \{1, \dots, N_{L+1}\}$. The theorem then follows since N_{L+1} is finite. □

2.5 Jacobian Moments

Consider the input-output Jacobian $\mathbf{J}_{L,\mathbf{x}}$, which we defined in eq. (2.17) as the matrix representation of

$$\mathrm{d}(\varphi_L \circ \Phi_{L-1})_{\mathbf{x}} \in \mathbb{R}^{N_L \times N_0}. \tag{2.49}$$

Letting $\{\xi_1, \ldots, \xi_{N_L}\}$ be the eigenvalues of $\mathbf{J}_{L,\mathbf{x}}\mathbf{J}_{L,\mathbf{x}}^\top$, recall that the empirical spectral distribution (ESD) of this matrix is defined as the random measure

$$\rho_L := \frac{1}{N_L}\sum_{i=1}^{N_L} \delta_{\xi_i} \tag{2.50}$$

where δ_{ξ_i} denotes a Dirac mass at ξ_i. Fix $x = (\lim_{N\to\infty} \frac{1}{N}\langle \mathbf{x}, \mathbf{x}\rangle)^{1/2}$ and let $m_{k,L}^{(N)}(x)$ denote the k-th moment of ρ_L, noting that

$$m_{k,L}^{(N)}(x) = \frac{1}{N}\mathrm{Tr}\left((\mathbf{J}_{L,\mathbf{x}}\mathbf{J}_{L,\mathbf{x}}^\top)^k\right). \tag{2.51}$$

We will assume the following for the remainder of this section.

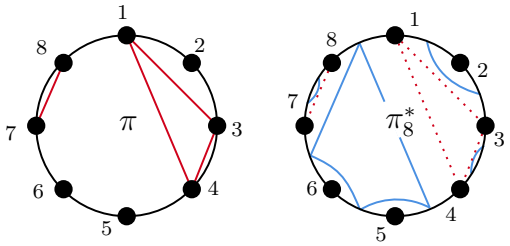
Assumption 2.17. *For all $\ell \geq 0$, $N_\ell = N$ and $\sigma_\ell^2 = 1/N$. Assume also that the $\mathbf{x} \in \mathbb{R}^N$ in (2.49) satisfies $\frac{1}{N}\langle \mathbf{x}, \mathbf{x}\rangle_{\mathbb{R}^N} \to x^2 \in \mathbb{R}$ as $N \to \infty$.*

Our goal is to show that under this assumption, $m_{k,L}^{(N)}(x) \to m_{k,L}(x)$, where $m_{k,L}$ is defined by the recursion in eq. (2.53). This is the content of theorem 2.19. In order to state the latter, we must first introduce the notion of a non-crossing partition and its Kreweras complement.

Definition 2.18. *A partition π of $\{1, \ldots, n\}$ is said to be non-crossing if, whenever a, b belong to a block (disjoint subset) of π and c, d to another, we cannot have $a < c < b < d$. We denote the set of all such partitions by* NC_n.

To visualize the non-crossing property, plot the sequence $(1, \ldots, n)$ in clockwise order as evenly spaced points on a circle. For every block $\{a_1, \ldots, a_j\}$ of $\pi \in \mathrm{NC}_n$, connect the points corresponding to each a_i to form polygons with these points as vertices, as depicted in fig. 2.7 (if $j = 2$, simply form a line segment). If π is non-crossing, the resulting polygons and line segments will never intersect.

Further, interpreting the circle as a cycle graph whose vertices are the same points $1, \ldots, n$ and edges are the arcs connecting these points, such a $\pi \in \mathrm{NC}_8$ induces a partition on the edges as follows. Label by i the edge connecting v_i and v_{i+1} and by n the edge connecting v_n and v_1. Then two edges are put in the same block if we can draw a line inside the circle connecting them without intersecting a polygon (or line segment) corresponding to a block of π. This resulting edge partition is known as the *Kreweras complement of* π (see Nica and Speicher [93]), and denote it by π^*. Note that this is also an element of NC_n.



$$\pi = \{\{1,3,4\},\{2\},\{5\},\{6\},\{7,8\}\} \in \mathrm{NC}_8$$
$$\pi^* = \{\{1,2\},\{3\},\{4,5,6,8\},\{7\}\}$$

Figure 2.7: *A non-crossing partition π of eight elements (red), and its Kreweras complement π^* (blue).*

Theorem 2.19. *Define*

$$\mu_{k,\ell+1}(x) := \mathbb{E}[\varphi'_{\ell+1}(X_\ell)^{2k}] \text{ for } \quad X_\ell \sim \mathcal{N}(0, K_\ell(x)), \tag{2.52}$$

where K_ℓ is defined recursively by

$$K_0(x) = x^2, \quad K_{\ell+1}(x) = \mathbb{E}[\varphi_{\ell+1}(X_\ell)^2].$$

Then under theorem 2.17,

$$m_{k,L}^{(N)}(x) \underset{N\to\infty}{\longrightarrow} m_{k,L}(x),$$

where $m_{k,L}$ is defined recursively as follows:

$$m_{k,0}(x) = 1, \quad m_{1,L}(x) = \mu_{1,L}(x) m_{1,L-1}(x),$$

$$m_{k,L}(x) = \sum_{\pi\in\mathrm{NC}_k} \mu_{\pi,L}(x) m_{\pi^*,L-1}(x), \tag{2.53}$$

where $\mu_{\pi,L}(x) := \prod_{B\in\pi} \mu_{|B|,L}(x)$ and similarly for $m_{\pi^*,L-1}(x)$.

Remark. *The moments $m_{k,L}$ can be seen as a non-linear (activation-dependent) generalisation of the Fuss-Catalan numbers (see, e.g., Nica and Speicher [93])*

$$\mathrm{FC}_L^{(k)} = \frac{1}{kL+1}\binom{(L+1)k}{k}, \quad \mathrm{FC}_L^{(k)} = \sum_{\pi\in\mathrm{NC}_k} \prod_{B\in\pi} \mathrm{FC}_{L-1}^{(|B|)}.$$

Indeed, if we consider the linear case with $\varphi = \mathrm{Id}$ in the above, then $\mu_{\pi,\ell} = 1$ for any π, ℓ and we recover $m_{k,L} = \sum_{\pi\in\mathrm{NC}_k} m_{\pi,L-1}$ (noting that $\pi \mapsto \pi^*$ is a bijection).

Since x is fixed a priori by theorem 2.17, we omit it from the notation in what follows.

Proof of theorem 2.19

Recall that we have derived a graph expansion for $m_{k,L}^{(N)}$ in eqs. (2.22) and (2.23). For general k, the latter takes the form

$$\frac{1}{N}\sum_{\eta_j^{(i)}}\left(\prod_{i=1}^{2}\prod_{j=1}^{k}\frac{\varphi_{\eta_j^{(i)}}}{s(\eta_j^{(i)})}\right)\mathbf{W}_{\mathrm{Tr}\left(\eta_1^{(1)}\circ(\eta_2^{(1)})^\top\circ\cdots\circ\eta_1^{(k)}\circ(\eta_2^{(k)})^\top\right)}, \tag{2.54}$$

where the sum if taken over all trees $\eta_j^{(i)}$ in $(\partial_{\mathbf{x}}\mathbb{T}_L(\mathbf{x}))^*$, for $i\in[2]$ and $j\in[k]$.

We call the graphs

$$\mathrm{Tr}\left(\eta_1^{(1)}\circ(\eta_2^{(1)})^\top\circ\cdots\circ\eta_1^{(k)}\circ(\eta_2^{(k)})^\top\right),$$

arising in this expansion *decorated cycles*, owing to their particular form: they each consist of a unique undirected cycle (meaning that one obtains a cycle after ignoring edge directions in the graph) which in this case is of length $2kL$, and trees rooted at vertices belonging to this cycle. An example for $L=2$ is depicted in fig. 2.8.

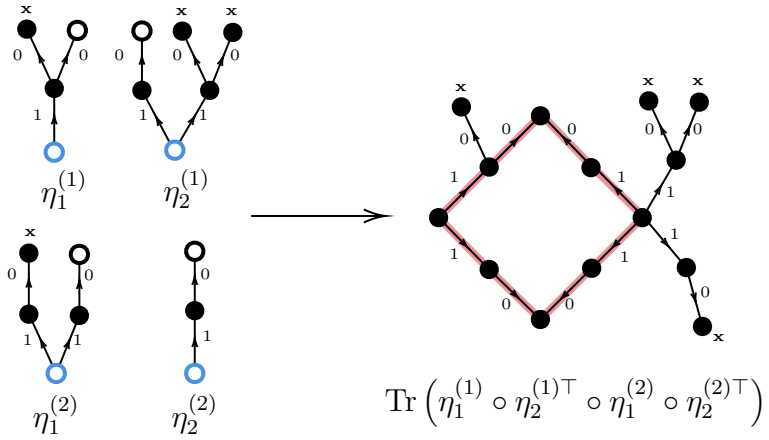


Figure 2.8: *A decorated cycle (right) formed by four trees (left), with the latter's unique (undirected) cycle highlighted in red.*

It will turn out to be convenient to work directly with the trees rooted at at the cycle instead of the initial $\eta_i^{(j)}$. To that end, note that in any $\eta_i^{(j)}\in(\partial_{\mathbf{x}}\mathbb{T}_L(\mathbf{x}))^*$, we have a unique path of length L from the root to the (unique) free leaf, which we call the *trunk*. The tree $\eta_i^{(j)}$ then consists of this trunk, together with L subtrees $\eta_{i,1}^{(j)},\ldots,\eta_{i,L}^{(j)}$ belonging to $(\mathbb{T}_1(\mathbf{x}))^*,\ldots,(\mathbb{T}_L(\mathbf{x}))^*$, respectively which come out of the trunk. This bijection is depicted in fig. 2.9 below.

Using this bijection, any graph of the form

$$\mathrm{Tr}\left(\eta_1^{(1)}\circ(\eta_2^{(1)})^\top\circ\cdots\circ\eta_1^{(k)}\circ(\eta_2^{(k)})^\top\right), \tag{2.55}$$

thus corresponds to a unique sequence of trees

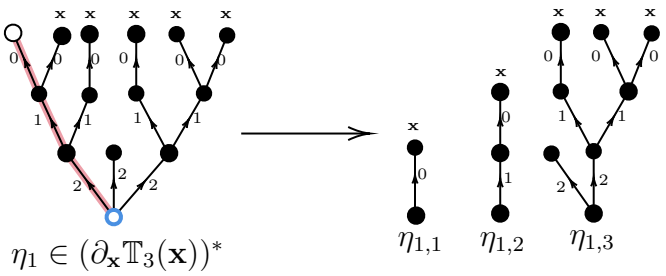


Figure 2.9: *A tree in $(\partial_{\mathbf{x}} \mathbb{T}(\mathbf{x}))^*$ is in bijection with the sequence of trees rooted at its trunk (highlighted in red).*

$$\boldsymbol{\eta} = (\eta_{i,l}^{(j)})_{j\in[k],i\in[2],l\in[L]} \in \prod_{i=1}^{k}((\mathbb{T}_0(\mathbf{x}))^* \times \cdots \times (\mathbb{T}_L(\mathbf{x}))^*)^2, \tag{2.56}$$

(which are product graphs) and so we adopt the notation $G(\boldsymbol{\eta})$ to denote the graph, as well as $C(\boldsymbol{\eta})$ to denote its cycle. Lastly, we use $\mathrm{DC}(k, L)$ to denote the space of all such graphs. With this notation, we can rewrite the expansion in eq. (2.54) as follows.

Lemma 2.20.

$$\mathrm{Tr}\Big((\mathbf{J}_{L,\mathbf{x}}\mathbf{J}_{L,\mathbf{x}}^\top)^k\Big) = \sum_{G(\boldsymbol{\eta})} \frac{\dot\varphi_{\boldsymbol{\eta}}}{s(\boldsymbol{\eta})}\mathbf{W}_{G(\boldsymbol{\eta})},$$

where the sum is taken over all $G(\boldsymbol{\eta}) \in \mathrm{DC}(k, L)$ (or equivalently, all $\boldsymbol{\eta} = (\eta_{i,l}^{(j)})_{i,j,l}$, c.f. eq. (2.56)), and

$$\frac{\dot\varphi_{\boldsymbol{\eta}}}{s(\boldsymbol{\eta})} := \prod_{i,j,l} \frac{\dot\varphi_{\eta_{i,l}^{(j)}}}{s(\eta_{i,l}^{(j)})}.$$

Proof. This follows from the aforementioned observation, and the fact that

$$\left(\prod_{i=1}^{2}\prod_{j=1}^{k} \frac{\varphi_{\eta_i^{(j)}}}{s(\eta_i^{(j)})}\right) = \prod_{i,j,l} \frac{\dot\varphi_{\eta_{i,l}^{(j)}}}{s(\eta_{i,l}^{(j)})}$$

where $\dot\varphi$ is defined as in theorem 2.16. □

We thus need to determine the limits of (values of) decorated cycles, which is achieved in the following proposition.

Proposition 2.21. *Let $k, L \geq 0$ and $G = G(\boldsymbol{\eta}) \in \mathrm{DC}(k, L)$ be a decorated cycle. Then under theorem 2.17, we have that*

$$\mathbb{E}\left\{(\lambda_G \mathbf{W}_G - a(G))^2\right\} = \mathcal{O}\left(\frac{1}{N}\right), \tag{2.57}$$

where

$$\lambda_G = \frac{1}{\sigma_G N^{\check e(G)+1}}, \text{ and } a(G) := \sum_{\psi\in\mathcal{P}_{\mathrm{A}}(G)} \prod_{\{\bullet_{\mathbf{x}},\bullet_{\mathbf{x}}\}\in(\mathcal{L}_0(G)/\sim_\psi)} \frac{\langle \mathbf{x}, \mathbf{x}\rangle}{N}. \tag{2.58}$$

Proof of theorem 2.21. Let $G = G(\boldsymbol{\eta})$ be a decorated cycle. We first show that for any admissible pairing $\phi \in \mathcal{P}(G)$, $\mathbf{W}_{G_\phi} = \mathcal{O}(\sigma_G N^{|V(G_\phi)|})$. Indeed, for any such ϕ, we have

$$\mathbf{W}_{G_\phi} = \sigma_G \prod_{\pi \in V(G)/\sim_\phi} \sum_{\alpha=1}^{N} [\odot_{\bullet_{\mathbf{y}} \in \pi} \mathbf{y}]_\alpha = \sigma_G N^{|V(G_\phi)|} \prod_{\pi \in V(G)/\sim_\phi} \frac{\sum_{\alpha=1}^{N} [\odot_{\bullet_{\mathbf{y}} \in \pi} \mathbf{y}]_\alpha}{N},$$

and as argued in the proof of theorem 2.10 (see fig. 2.5), the only vertices with input $\mathbf{x}$ in G are of in-degree 1. All other vertices have input $\mathbf{1}$.

For any $\pi \in V(G)/\sim_\phi$, we therefore have

$$\odot_{\bullet_{\mathbf{y}} \in \pi} \mathbf{y} \in \{\mathbf{1}, \mathbf{x}, \mathbf{x} \odot \mathbf{x}\},$$

so that

$$\frac{\sum_{\alpha=1}^{N} [\odot_{\bullet_{\mathbf{y}} \in \pi} \mathbf{y}]_\alpha}{N} \in \left\{1, \frac{\sum_{\alpha=1}^{N} [\mathbf{x}]_\alpha}{N}, \frac{\langle \mathbf{x}, \mathbf{x} \rangle}{N}\right\}.$$

Since $\langle \mathbf{x}, \mathbf{x} \rangle / N \to_{N \to \infty} x$ and

$$\frac{\sum_{\alpha=1}^{N} [\mathbf{x}]_\alpha}{N} \leq \frac{\|\mathbf{x}\|_1}{N} \leq \frac{\sqrt{N}\, \|\mathbf{x}\|_2}{N} = \frac{\|\mathbf{x}\|_2}{\sqrt{N}} \to_{N \to \infty} \sqrt{x},$$

we can conclude that

$$\mathbf{W}_{G_\phi} \leq C_x \sigma_G N^{|V(G_\phi)|}$$

for some constant C_x depending on x. The order of $\mathbf{W}_{G_\phi}$ is therefore determined by the number of vertices in G_ϕ.

Following section 1.3.2, it suffices to prove that

$$\sum_{\psi \in \mathcal{P}_{\mathrm{A}}(G)} \mathbf{W}_{G_\psi} = \sigma_G N^{|E(G_\psi)|+1} a(G),$$

which is true in the event that

$$\prod_{\pi \in V(G)/\sim_\psi} \frac{\sum_{\alpha=1}^{N} [\odot_{\bullet_{\mathbf{y}} \in \pi} \mathbf{y}]_\alpha}{N} = \prod_{\{\bullet_{\mathbf{x}}, \bullet_{\mathbf{x}}\} \in (\mathcal{L}(G)/\sim_\psi)} \frac{\langle \mathbf{x}, \mathbf{x} \rangle}{N}$$

holds for every $\psi \in \mathcal{P}_{\mathrm{A}}(G)$.

The latter is equivalent to proving that the edges whose input is equal to $\mathbf{x}$ are paired between themselves.

Note that on the one hand, these are 0-labeled edges. On the other, 0-labeled edges arise in one of two ways. They are either:

1. Edges whose tails are leaves of G (meaning vertices for which the sum of the in and out-degree equals one), and which therefore have input $\mathbf{x}$, or
2. Edges belonging to the cycle $C(\boldsymbol{\eta})$.

Since ψ is atomic, the result then follows by theorem 2.22, as it prohibits ψ form pairing edges inside of $C(\boldsymbol{\eta})$ with edges outside of it in G. □

Remark. *Note that by theorem 2.17, $\lambda_G = \frac{1}{N}$ for all decorated cycles G since $\sigma_\ell = \frac{1}{\sqrt{N}}$ for all $\ell \geq 0$.*

Since $\sigma_\ell = \frac{1}{\sqrt{N}}$, we then have that

$$\frac{1}{N}\mathrm{Tr}\Big((\mathbf{J}_{L,\mathbf{x}}\mathbf{J}_{L,\mathbf{x}}^\top)^k\Big) = \sum_{G(\boldsymbol{\eta})\in \mathrm{DC}(k,L)} \frac{\dot\varphi_{\boldsymbol{\eta}}}{s(\boldsymbol{\eta})}\lambda_{G(\boldsymbol{\eta})}\mathbf{W}_{G(\boldsymbol{\eta})} \xrightarrow{L^2} \sum_{G(\boldsymbol{\eta})\in \mathrm{DC}(k,L)} \frac{\dot\varphi_{\boldsymbol{\eta}}}{s(\boldsymbol{\eta})} a(G(\boldsymbol{\eta}))$$

with rate $\mathcal{O}\big(\frac{1}{N}\big)$, since these are finite sums.

Now letting

$$\tilde m_{k,L} = \sum_{G(\boldsymbol{\eta})\in \mathrm{DC}(k,L)} \frac{\dot\varphi_{\boldsymbol{\eta}}}{s(\boldsymbol{\eta})} a(G(\boldsymbol{\eta})),$$

it suffices to show that $\tilde m_{k,L}$ satisfies the recursion in (2.53).

To derive a recursive formula for the coefficients $m_{k,L}$, we analyze the structure of atomic pairings for decorated cycles, beginning with the following result.

Lemma 2.22. *Let $G(\boldsymbol{\eta}) \in \mathrm{DC}(k, L)$, $C(\boldsymbol{\eta})$ be this graph's cycle (*i.e. *the subgraph which is a cycle after forgetting about edge orientations), and $\phi \in \mathcal{P}_{\mathrm{A}}(G(\boldsymbol{\eta}))$. Then $\phi = \phi_C \sqcup \phi_{G\setminus C}$ for some $\phi_C \in \mathcal{P}(C(\boldsymbol{\eta}))$. In particular, all of the L-edges of $C(\boldsymbol{\eta})$ are paired between themselves by some pairing $\phi_L \subseteq \phi_C$.*

Proof. For ϕ to be atomic, it must pair edges in a way that collapses the undirected cycle $C(\boldsymbol{\eta})$ without creating any new cycles in the process. But this is only possible if edges of $C(\boldsymbol{\eta})$ are paired between themselves, as any edge that isn't paired this way would end up in a cycle as depicted in fig. 2.10 below.

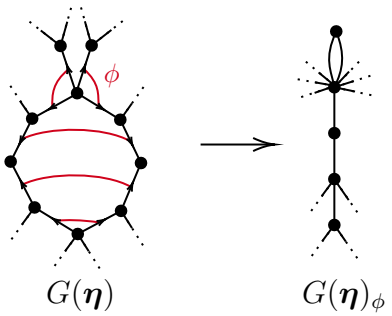


Figure 2.10: *A cycle formed by pairing edges that belong to $C(\boldsymbol{\eta})$ with edges that do not.*

Further, since edges can only be paired with other edges with the same input, ϕ must contain a pairing ϕ_L of the L-edges in $C(\boldsymbol{\eta})$. □

Now let $\phi \in \mathcal{P}_{\mathrm{A}}(G(\boldsymbol{\eta}))$, and ϕ_L be as in the lemma above. Note that in $G(\boldsymbol{\eta})$, the trees $\eta_{1,L}^{(j)}, \eta_{2,L}^{(j)}$ share a root for each j, which belongs to $C(\boldsymbol{\eta})$. For each $j \in [k]$, we let

$v_j := \mathrm{root}(\eta_{1,L}^{(j)} \wedge \eta_{2,L}^{(j)})$ denote said roots, and note that each v_j is the head of exactly two L-edges in $C(\boldsymbol{\eta})$. All other vertices in the latter are heads of edges with inputs W_i with $i < L$.

Said differently, the k pairs of L-edges of $C(\boldsymbol{\eta})$ are in bijection with $v_1, \ldots, v_k$, and through this bijection, $\phi_L \subseteq \phi$ induces a unique partition π of $\{1, \ldots, k\}$ defined by letting a and b be in the same block of π if they label endpoints of edges which have been paired by ϕ_L. This is depicted in fig. 2.11, and we write $\phi_L \sim \pi$ to denote the fact that ϕ_L induces π.

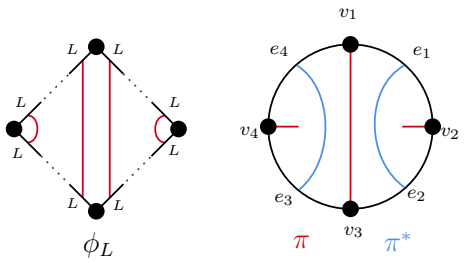


Figure 2.11: *A pairing ϕ_L of the L-edges of a cycle $C(\boldsymbol{\eta})$ (left), the partition $\pi \in \mathrm{NC}_4$ that is induced by ϕ_L (right, in red), and its complement π^* (in blue).*

For atomic ϕ's, we notice that the only π that arise this way must be non-crossing. Indeed, if $\phi_L \sim \pi$ and π admits a crossing, then a cycle is formed as depicted in fig. 2.12 below.

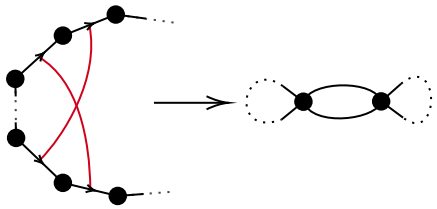


Figure 2.12: *If $\phi_L \sim \pi$ and π has a crossing, then identifying edges according to ϕ_L creates a cycle.*

Next, recall that by definition, v_j is the root of $\eta_{1,L}^{(j)} \wedge \eta_{2,L}^{(j)}$ for every $j \in [k]$. Upon identifying L-edges of $C(\boldsymbol{\eta})$ by ϕ_L, such trees get joined at their roots. More precisely, if $\phi_L \sim \pi$, then for every block $B = \{j_1, \ldots, j_l\}$ of π we obtain the tree

$$(\eta_{1,L}^{(j_1)} \wedge \eta_{2,L}^{(j_1)} \wedge \cdots \wedge \eta_{1,L}^{(j_l)} \wedge \eta_{2,L}^{(j_l)}) := \wedge_{i=1}^{l} (\eta_{1,L}^{(j_i)} \wedge \eta_{2,L}^{(j_i)}).$$

Furthermore, for ϕ to be acyclic, it must at the very least pair edges of such trees internally within each tree (otherwise, a cycle would be formed at their roots). We therefore get, for each block $B = \{j_1, .., j_l\}$ of π, a pairing $\phi_B \in \mathcal{P}_{\mathrm{A}}(\wedge_{i=1}^{l}(\eta_{1,L}^{(j_i)} \wedge \eta_{2,L}^{(j_i)}))$

Having now paired L-edges belonging to $C(\boldsymbol{\eta})$ (using ϕ_L), and those belonging to trees of height L decorating this cycle (using the ϕ_B's), we must describe how ϕ pairs

the remaining edges. Crucially, observe that the unpaired edges have inputs W_i for $i \in \{1, \ldots, L-1\}$ and form disjoint decorated cycles themselves. To be precise, we will need the following definition.

Definition 2.23. *Let $G(\boldsymbol{\eta}) \in \mathrm{DC}(k, L), \phi_L$ be as above and $\phi_L \sim \pi$ for some $\pi \in \mathrm{NC}_k$. Let $\tilde{B}$ be a block of $\pi^* \in \mathrm{NC}_k$. Then the* restriction of $G(\boldsymbol{\eta})_{\phi_L}$ to $\tilde{B}$ *is the decorated cycle in* $\mathrm{DC}(|\tilde{B}|, L-1)$ *determined by the trees*

$$\boldsymbol{\eta}|_{\tilde{B}} = (\eta_{i,l}^{(j)})_{i,j,l}, \quad i \in [2], l \in [L-1], j \in B.$$

Returning to our previous point, the leftover edges (that were not paired by ϕ_L or ϕ_B) form one decorated cycle per block of π^*, which are given by

$$G(\boldsymbol{\eta}|_{\tilde{B}}), \quad \tilde{B} \in \pi^*.$$

For ϕ to be atomic, we claim that the edges of each of these $G(\boldsymbol{\eta}|_{\tilde{B}})$ must once again be paired internally. This is because the $G(\boldsymbol{\eta}|_{\tilde{B}})$ are connected to one another in $G(\boldsymbol{\eta})_{\phi_L \sqcup \{\phi_B\}_{B \in \pi}}$ by paths whose edges that have already been paired by ϕ_B or ϕ_L; pairing some $e \in G(\boldsymbol{\eta}|_{\tilde{B}})$ to another $e' \in G(\boldsymbol{\eta}|_{\tilde{B}'})$ would then create a cycle which we cannot collapse (as in fig. 2.12). Put differently, this implies that any pairing ϕ' of the remaining edges can be decomposed into a disjoint union as follows:

$$\phi' = \bigsqcup_{\tilde{B} \in \pi^*} \phi'_{\tilde{B}}.$$

Summarizing everything we have said so far, any pairing $\phi \in \mathcal{P}_{\mathrm{A}}(G(\boldsymbol{\eta}))$ consists of the following:

1. A pairing ϕ_L of the L-edges of $C(\boldsymbol{\eta})$, such that $\phi_L \sim \pi \in \mathrm{NC}_k$.
2. For every block $B = \{j_1, \ldots, j_l\}$ of π, a pairing $\phi_B \in \mathcal{P}_{\mathrm{A}}(\wedge_{i=1}^{l}(\eta_{1,L}^{(j_i)} \wedge \eta_{2,L}^{(j_i)}))$.
3. For every block $\tilde{B}$ of π_k^*, a pairing $\phi_{\tilde{B}} \in \mathcal{P}_{\mathrm{A}}(G(\boldsymbol{\eta})|_{\tilde{B}})$.

Noting how the quantity

$$\prod_{\{\bullet_{\mathbf{x}}, \bullet_{\mathbf{x}}\} \in (\mathcal{L}_0(G(\eta))/\sim_\phi)} \frac{\langle \mathbf{x}, \mathbf{x} \rangle}{N}$$

then factors into

$$\left(\prod_{B \in \pi} \prod_{\{\bullet_{\mathbf{x}}, \bullet_{\mathbf{x}}\} \in (\mathcal{L}_0(G(\eta))/\sim_{\phi_B})} \frac{\langle \mathbf{x}, \mathbf{x} \rangle}{N} \right) \left(\prod_{\tilde{B} \in \pi^*} \prod_{\{\bullet_{\mathbf{x}}, \bullet_{\mathbf{x}}\} \in (\mathcal{L}_0(G(\eta))/\sim_{\phi_{\tilde{B}}})} \frac{\langle \mathbf{x}, \mathbf{x} \rangle}{N} \right),$$

the decomposition of $\phi \in \mathcal{P}_{\mathrm{A}}(G(\boldsymbol{\eta}))$ given by (1), (2), (3) yields the following result.

Lemma 2.24.

$$a(G(\boldsymbol{\eta})) = \sum_{\pi \in \mathrm{NC}_k} \left(\prod_{B=\{j_i\}_{i=1}^{l}} a(\wedge_{i=1}^{l}(\eta_{1,L}^{(j_i)} \wedge \eta_{2,L}^{(j_i)})) \right) \left(\prod_{\tilde{B} \in \pi_k^*} a(G(\boldsymbol{\eta}|_{\tilde{B}})) \right). \tag{2.59}$$

Before finishing the proof of theorem 2.19, an example illustrating this lemma and the heavy notation that we have introduced for it is well in order. Consider the case when we're computing $m_{4,2}$, for which we end up studying decorated cycles $G(\boldsymbol{\eta}) \in \mathrm{DC}(4,2)$. Then the decomposition of $a(G(\boldsymbol{\eta}))$ given in theorem 2.24 is depicted in full detail in the fig. 2.13 below.

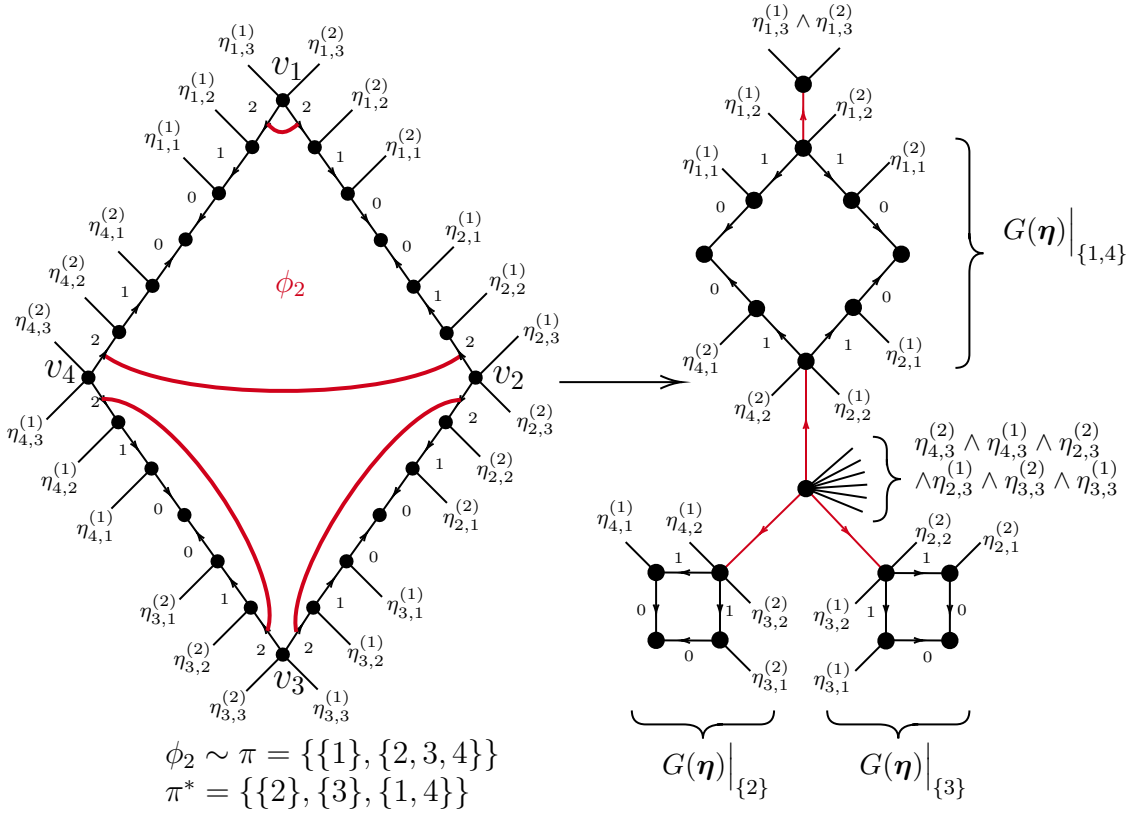


Figure 2.13: *Let $\phi \in \mathcal{P}_{\mathrm{A}}(G(\boldsymbol{\eta}))$ and $\phi_2 \subset \phi$ be as in theorem 2.22 (depicted in red on the left in this figure). Then ϕ_2 corresponds to a non-crossing partition π: here, π only has two blocks, which gives rise to the trees $\eta_{1,3}^{(1)} \wedge \eta_{1,3}^{(2)}$ and $\eta_{4,3}^{(1)} \wedge \eta_{4,3}^{(2)} \wedge \eta_{2,3}^{(1)} \wedge \eta_{2,3}^{(2)} \wedge \eta_{3,3}^{(1)} \wedge \eta_{3,3}^{(2)}$. Its complement has three blocks, giving rise to three decorated cycles: $G(\boldsymbol{\eta})|_{\{1,4\}} \in \mathrm{DC}(2,1)$, $G(\boldsymbol{\eta})|_{\{2\}} \in \mathrm{DC}(1,1)$ and $G(\boldsymbol{\eta})|_{\{3\}} \in \mathrm{DC}(1,1)$.*

We now have the tools to prove theorem 2.19.

Proof of theorem 2.19. Recall that

$$\tilde{m}_{k,L} = \sum_{G(\boldsymbol{\eta}) \in \mathrm{DC}(k,L)} \frac{\dot{\varphi}_{\boldsymbol{\eta}}}{s(\boldsymbol{\eta})} a(G(\boldsymbol{\eta}))$$

by theorem 2.21. We can then use the decomposition given in theorem 2.24 for each $a(G(\boldsymbol{\eta}))$ in this sum to get

$$\sum_{G(\boldsymbol{\eta})\in \mathrm{DC}(k,L)} \frac{\dot{\varphi}_{\boldsymbol{\eta}}}{s(\boldsymbol{\eta})} \sum_{\pi\in \mathrm{NC}_k} \left(\prod_{B=\{j_i\}_{i=1}^{l}\in\pi} a(\wedge_{i=1}^{l}(\eta_{1,L}^{(j_i)} \wedge \eta_{2,L}^{(j_i)})) \right) \left(\prod_{\tilde{B}\in\pi_k^*} a(G(\boldsymbol{\eta}|_{\tilde{B}})) \right).$$

The next step will be to interchange the order of summation.

To that end, we must first distribute the factor $\frac{\dot{\varphi}_{\boldsymbol{\eta}}}{s(\boldsymbol{\eta})}$. The same arguments of theorem B.4 used to prove theorem 2.9 show that

$$\mu_{k,L} = \sum_{(\eta_{i,L}^{(j)})_{j\in[k],i\in[2]}\in((\mathbb{T}_L(\mathbf{x}))^*)^{2k}} \left(\prod_{j=1}^{k} \frac{\dot{\varphi}_{\eta_{2,L}^{(j)}} \dot{\varphi}_{\eta_{1,L}^{(j)}}}{s(\eta_{1,L}^{(j)}) s(\eta_{2,L}^{(j)})} \right) a(\wedge_{j=1}^{k}(\eta_{1,L}^{(j)} \wedge \eta_{2,L}^{(j)})).$$

Notice that the sum over NC_k does not depend on a specific $G(\boldsymbol{\eta})$. Splitting the sum over $G(\eta)$ into two sums, the first over $(\eta_{i,L}^{(j)})_{j\in[k],i\in[2]} \in ((\mathbb{T}_L(\mathbf{x}))^*)^{2k}$ and the second over the remaining trees in $\boldsymbol{\eta}$, interchanging the sums then gives

$$\sum_{\pi\in\mathrm{NC}_k} \left(\prod_{B\in\pi} \mu_{|B|,L} \right) \left(\prod_{\tilde{B}\in\pi^*} \sum_{G(\boldsymbol{\eta}|_{\tilde{B}})} \frac{\dot{\varphi}_{\boldsymbol{\eta}|_{\tilde{B}}}}{s(\boldsymbol{\eta}|_{\tilde{B}})} a(G(\boldsymbol{\eta}|_{\tilde{B}})) \right).$$

Lastly, we note that

$$\sum_{G(\boldsymbol{\eta}|_{\tilde{B}})} \frac{\dot{\varphi}_{\boldsymbol{\eta}|_{\tilde{B}}}}{s(\boldsymbol{\eta}|_{\tilde{B}})} a(G(\boldsymbol{\eta}|_{\tilde{B}})) = \sum_{G'\in\mathrm{DC}(|\tilde{B}|,L-1)} \frac{\dot{\varphi}_{G'}}{s(G')} a(G') = \tilde{m}_{|\tilde{B}|,L-1}.$$

It follows that

$$\tilde{m}_{k,L} = \sum_{\pi\in\mathrm{NC}_k} \mu_{\pi,L} \tilde{m}_{\pi^*,L-1},$$

and we conclude that $\tilde{m}_{k,L} = m_{k,L}$ for every k, L, where $m_{k,L}$ is as in the Proposition's statement.

□

2.6 Observations on the Algebra of Trees

The trees which emerge from our expansions are remarkably similar to those appearing in the context of Butcher series [83], which constitute a related type of Taylor-like expansion. The context of Butcher series has been extensively studied, and a rich *Hopf*-algebraic structure has emerged as playing a crucial role [35, 90].

It becomes natural to ask whether a similar structure is expressed by our objects, which, as we have shown in this chapter, do exhibit a complex combinatorial structure. We leave this as a promising avenue for further research.

Note that the trees of chapter D, which can fundamentally be seen as Butcher trees where each branch is a tree of the type studied in this chapter, seem amenable to a natural extension of the algebraic structure found in the literature.

Before concluding, note that thanks to theorem 2.11, it is possible to imbue the vector space generated by trees with a Hilbert structure given by $\langle \tau, \eta \rangle := \alpha(\tau \sqcup \eta)$, fundamentally identifying τ with the Gaussian variable Z_τ in L^2. This corresponds to taking the dot product of the trees with independent Gaussian weights as in the proposition, and then taking the limit as $N \to \infty$. Interestingly, as we present in chapter E, it is possible to construct a (limiting) orthonormal basis of this space combinatorially by working on trees, in the same way as Chizat et al. [13] construct their orthonormal basis in the linear case: by restricting the sum over indexations in theorem 1.2 to those indexations where each index appears only once. These objects are orthonormal and each tree can be decomposed into them, with a methodology expressible in a recursive way. These observations might open a way to study the training regime, extending the rigorous arguments of Chizat et al. [13] to the non-linear case.

Chapter 3

State Space Models

In this chapter we study how the genus techniques of chapter 1 can be applied to the analysis of State Space Models (SSMs), a modern class of machine learning models for processing sequential data that have emerged as a compelling alternative to Transformer architectures [44, 123]. SSMs have gained prominence in language modeling and other sequential tasks due to their unique computational profile, which addresses key limitations of attention-based models.

The appeal of SSMs lies in their dual computational advantages. During inference, their stateful nature enables extremely efficient autoregressive generation with $O(1)$ complexity per step, in stark contrast to the $O(L)$ complexity of Transformers, where L denotes sequence length. This efficiency stems from the model's ability to maintain a fixed-size hidden state that is updated incrementally. Simultaneously, during training and prefill phases, SSMs retain the parallelizability that makes modern deep learning tractable, with operations that are rich in matrix multiplications and thus well-suited to contemporary hardware accelerators [21].

This combination of efficient inference and parallelizable training has positioned SSMs at the forefront of recent breakthroughs in language modeling, where they demonstrate remarkable performance across diverse tasks while offering significant computational advantages [44, 134].

This chapter follows the theoretical foundations laid out in Muca Cirone et al. [89], while extending the analysis in section 3.3 with the linear case proofs from Muca Cirone et al. [88] using genus expansion techniques. The final section 3.6, extending results to structured matrices, appeared in Movahedi et al. [87], Walker et al. [125].

3.1 Introduction

Sequence-to-sequence blocks are fundamental components of modern deep learning models for language, images, video, audio, time series, and genomics. For the last five years, attention [26, 123] has been the dominant mechanism powering these architec-

tures. However, competitive results have recently been achieved without attention, by using state-space models (SSMs): GPU-efficient linear recurrent sequence-to-sequence blocks stemming from S4 [46]. SSMs achieve state-of-the-art results on long-range-reasoning benchmarks [121] and show outstanding performance in various domains including vision [92], audio [42], biological signals [46], reinforcement learning [73] and online learning [136]. SSMs recently have gained significant interest in the community since their computational complexity scales linearly in sequence length, while attention scales quadratically; moreover, unlike other recurrent mechanisms such as LSTMs [52] and GRUs [14], they can be efficiently parallelized on GPUs during training using parallel scans [81, 118].

While standard SSMs were shown to be particularly powerful on signal processing tasks, their computation power is limited: the core sequential mechanism of S4 is equivalent to a convolution (filtering) [71]. This represents a drawback in challenging domains such as text and genetics, where the ability to select data efficiently in an input-dependent manner *i.e.*, perform content-based reasoning is crucial (see [2, 38, 127]). Towards reaching this goal with recurrent models, various adaptations of S4 have been proposed in the last few months. Notably, Mamba [44] implements simple and efficient gating mechanisms on the S4 recurrence, unlocking input selectivity in the memory update. Mamba achieved state-of-the-art performance in various language modeling tasks while greatly improving the inference throughput. Similar ideas can be found in recent developments inspired by attention, such as RWKV [99], RetNet [119], Gateloop [64], Gated Linear Attention (GLA) [133], and HGRN2 [105]. Very recently, [23] surpassed the performance of Mamba with a gated RNN architecture - Griffin - based on an improved version of the LRU [95], and [30] introduced minimal versions of GRU and LSTM as gated SSMs.

Contributions At the core of the models discussed above is a time-varying dynamical system, where reasoning is performed through an efficient and parallelizable update *linear* in the hidden state. We generalize the structure of such models, drawing a direct link to *controlled differential equations (CDEs)* [32, 53, 67, 75, 86, 114, 124, 135] and use tools from *rough path theory* [77] to study expressivity. Our framework not only provides significant theoretical insight regarding some recently proposed SSM architectures, but we also envision it to be a useful tool in analysing, and perhaps developing, future architectural advances.

3.2 State Space Models and Linear CDEs

As sequence to sequence models, SSMs process a d-dimensional input sequence $x_1, \dots, x_L \in \mathbb{R}^{d_{in}}$ and produce a output sequence $y_1, \dots, y_L \in \mathbb{R}^{d_{out}}$, to do this these models use a

hidden state $z_1, \dots, z_L \in \mathbb{R}^{d_{hidden}}$ produced by a recurrence of the form

$$z_\ell = z_{\ell-1} + A(x_\ell) \cdot z_{\ell-1} + B(x_\ell), \tag{3.1}$$

where $A : \mathbb{R}^{d_{in}} \to \mathbb{R}^{d_{hidden} \times d_{hidden}}$ and $B : \mathbb{R}^{d_{in}} \to \mathbb{R}^{d_{hidden}}$ are learnable functions mapping the input to the transition matrices defining the recurrence. The input dependecy, refered to as *gating* and introduced in the seminal work of Gu and Dao [44], is a key feature that allows SSMs to process sequential data in an expressive way as we will see in the next section. The output is then computed as

$$y_\ell = F(z_\ell, x_\ell), \tag{3.2}$$

where $F : \mathbb{R}^{d_{hidden}} \times \mathbb{R}^{d_{in}} \to \mathbb{R}^{d_{out}}$ is a learnable function mapping the hidden state and the input at a given time step to the corresponding output.

3.2.1 Linear CDEs

Most State Space Models can be viewed as discretized versions of Linear Controlled Differential Equations (Linear CDEs). Considering them in this continuous-time setting allows us to study their expressivity in a unified way leveraging the rich theory of *Rough Paths* (*c.f.* chapter C). According to their continuous-time formulation, SSMs process input data sampled from a continuous path $X : [0, 1] \to \mathbb{R}^d$. By X_t^i we denote the input channel i, evaluated at time $t \in [0, 1]$. More formally, we consider input trajectories in the separable Banach space $\mathbb{X} = C_{1,0}([0, 1]; \mathbb{R}^d)$ of absolutely continuous $\mathbb{R}^d$ dimensional paths starting in the origin. In this space, one can write $X = \int_0^t \dot{X}_s \, ds$ since $\dot{X} \in L^1([0, 1]; \mathbb{R}^d)$; and the norm is $\|X\|_{1;[0,1]} := \int_0^1 |\dot{X}_s| \, ds$.

To model the input dependency of transition matrices, we introduce three maps transforming the path X: two path-valued maps with output trajectories $\omega^{\mathrm{X}}, \xi^{\mathrm{X}}$ living in potentially different dimensions, and one vector-valued map for the initial condition

$$(\cdot)_0 : \mathbb{X} \to \mathbb{R}^{d_0}; \quad \omega : \mathbb{X} \to C_{1,0}([0, 1]; \mathbb{R}^{d_\omega}), \quad \xi : \mathbb{X} \to C_{1,0}([0, 1]; \mathbb{R}^{d_\xi})$$

where we used the shorthand notation $\omega^{\mathrm{X}} := \omega(X), \xi^{\mathrm{X}} = \xi(X)$. Akin the notation for the i-the channel of X (i.e. X^i), we denote by $\omega^{\mathrm{X},i}$ the i-th channel in ω^{X}.

Definition 3.1 (Linear CDE)**.** *Fix $N \in \mathbb{N}$ (hidden-state dimension), transition matrices $A_1, ..., A_{d_\omega} \in \mathbb{R}^{N \times N}$ and $B \in \mathbb{R}^{N \times d_\xi}$. The hidden state $Z^X := Z(X) : [0, 1] \to \mathbb{R}^N$ is computed by the Linear CDE through a linear (in the hidden state) differential equation driven by ω^X, ξ^X:*

$$dZ_t^X = \sum_{i=1}^{d_\omega} A_i Z_t^X \, d\omega_t^{X,i} + B d\xi_t^X, \quad Z_0^X = (X)_0 \in \mathbb{R}^N \tag{3.3}$$

Notable examples

We show that both S4 [46] and Mamba [44], two of the most popular SSMs, can be written in continuous-time as systems of parallel Linear CDEs. The key ingredient setting the two models apart is the choice of drivers ω and ξ.

S4 as a Linear CDE In S4 each channel $X^i : [0,1] \to \mathbb{R}$ is processed independently to generate a hidden state $Z^i : [0,1] \to \mathbb{R}^{N_i}$, these hidden states are then linearly combined to produce the output. Thus it suffices to consider the setting $d = 1$, $N = N_i$.

The original formulation of S4 defines the hidden state as

$$\frac{dZ_t}{dt} = AZ_t + BX_t, \quad Z_0 = Z_0 \in \mathbb{R}^N \tag{3.4}$$

where $A \in \mathbb{R}^{N \times N}$ is diagonal matrix, often structured i.e. initialized deterministically through HiPPO theory [45] in diagonal form, and $B, Z_0 \in \mathbb{R}^N$ are fixed vectors.

The discrete version is obtained by sampling the continuous-time path X at a fixed stepsize Δ and using the Zero-Order Hold discretization[1]. In continuous-time, this can be equivalently written as a Linear CDE with $\omega_t^{\mathrm{X}} = t \in \mathbb{R}$ and $\xi_t^{\mathrm{X}} = \int_0^t X_s ds \in \mathbb{R}$.

$$(X)_0 = Z_0, \quad \omega_t^{\mathrm{X}} = t \in \mathbb{R}, \quad \xi_t^{\mathrm{X}} = \int_0^t X_s ds \in \mathbb{R}, \tag{3.5}$$

since $d\omega_t^{\mathrm{X}} = dt$ and $d\xi_t^{\mathrm{X}} = X_t dt$. As the reader can promptly notice, here ω^{X} is not a function of X – this will have a crucial role in expressivity (see section 3.4.1).

Mamba as a Linear CDE For Mamba, to each channel i is again associated a hidden state $Z^i : [0,1] \to \mathbb{R}^{N_i}$, but instead of being independent these are now coupled through the input-dependency of the transition matrices A_i and B_i. In particular, the recurrence can be written as

$$\frac{dZ_t^i}{dt} = A_i(X_t) Z_t^i + B_i(X_t), \quad Z_0^i = Z_0^i \in \mathbb{R}^{N_i} \tag{3.6}$$

where $A_i : \mathbb{R}^d \to \mathbb{R}^{N_i \times N_i}$ and $B_i : \mathbb{R}^d \to \mathbb{R}^{N_i}$ depend on all input channels at once as

$$A_i(X_t) = \Delta_i(X_t) A_i, \quad B_i(X_t) = X_t^i \Delta_i(X_t)(B \cdot X_t)$$

[1] This corresponds to: (i). considering the continuous underlying signal X_t to be piecewise constant, (ii). solving exactly ODE (3.4) and finally (iii). sampling Z_t at the sample times of X_t. See section C.4.1

where

$$\Delta_i(x) := \text{softplus}(\alpha_i \cdot x + \beta_i) \in \mathbb{R}, \quad \alpha_i \in \mathbb{R}^{d_{in}}, \quad \beta_i \in \mathbb{R}$$

is a non-linear function of the input channels, B is a channel-independent linear function and the A_i are *diagonal* matrices. Once again, the discrete version is obtained by sampling the continuous-time path X at a fixed stepsize Δ and using the Zero-Order Hold discretization. This is a system of Linear CDEs with channel-dependent gates $(X)_{i;0} = Z_0^i$ and drivers ω_i, ξ_i chosen as:

$$\omega_{i;t}^X = \int_0^t \text{softplus}(\alpha_i \cdot X_s + \beta_i) ds \in \mathbb{R}, \quad \xi_{i;t}^X = \int_0^t X_s \; X_s^i \text{softplus}(\alpha_i \cdot X_s + \beta_i) ds \in \mathbb{R}^d.$$

Remark (Are hidden state components recurrently mixed?)**.** *In the Linear CDE defined by* $dZ_t^X = [\sum_{j=1}^{d_\omega} A_j d\omega_t^{X,j}] \cdot Z_t^X + B \cdot d\xi_t^X$ *the channels of the transformed input path* ω^X *are mixed in a linear way in the recurrent step and stored in a shared hidden state* Z^X *. This allows our framework to be more general compared to S4 and Mamba, where each channel of the input is processed individually, and hidden states are later combined. We know already from Merrill et al. [84] that this distinction between hidden state mixing strategies is crucial for expressivity. Our discussion in section 3.4.2 provides an in-depth look at the effects of separate channel processing. This can be achieved in our framework by choosing the* A_i*s to be diagonal and non-zero only around a channel-specific portion.*

3.2.2 Taylor expansion for Linear CDEs

A great advantage of the Linear CDE formalism is that it allows to leverage the theory of Rough Paths. A crucial tool in this context is the Taylor expansion of the solution of the Linear CDE, a result which allows to completely decouple the effect on the hidden state of the gates (the time and input dependent part), and of the transition matrices (the proper parameters of the system).

To present this result, we briefly introduce the *signature transform* [33, 66, 77], a fundamental tool from stochastic analysis that transforms paths into sequences of iterated integrals. The signature transform will play a central role in studying the expressivity of Linear CDEs - much like monomials form building blocks for approximating functions on Euclidean spaces, linear functionals on the signature provide the essential components for approximating continuous functions on path space, a fact that follows from the Stone-Weierstrass theorem.

Consider then a path $\gamma \in C_{1,0}([0,1]; \mathbb{R}^{d_\gamma})$ and define as $\mathbb{W}_{d_\gamma}$ the set of *words* (*i.e.* ordered sequences) in $\{1, \dots, d_\gamma\}$ [2]. The signature transform is the following infinite

[2] So when we write $I \in \mathbb{W}_{d_\gamma}$ we mean $I = i_1 i_2 \cdots i_M$ for some integer $M \geq 0$ and letters $i_m \in \{1, \dots, d_\gamma\}$, then Ii will be the word $i_1 i_2 \cdots i_M i$ for $i \in \{1, \dots, d_\gamma\}$.

collection of scalar iterated integrals

$$\mathrm{Sig}(\gamma)_{s,t} := \left(\mathrm{Sig}(\gamma)^I_{s,t}\right)_{I\in\mathbb{W}_{d_\gamma}}, \quad \mathrm{Sig}(\gamma)^I_{s,t} := \int_{s<u_1<\ldots<u_n<t} \dot\gamma^{(i_1)}_{u_1}\ldots\dot\gamma^{(i_n)}_{u_n} du_1\ldots du_n.$$

We refer to chapter C for additional details and references on the signature.

Proposition 3.2 (Taylor expansion for Linear CDEs, theorem C.18)**.** *For any choice of matrices $A_1, \ldots, A_{d_\omega}$, $B = [B_1|\ldots|B_{d_\xi}]$ and $C = [C_1|\ldots|C_{d_{in}}] \in \mathbb{R}^{N\times d_{in}}$, the unique solution of the Linear CDE (3.3) with $(\cdot)_0$ gate $X \mapsto CX_0$ is, using the notation $A_I := A_{i_n}\ldots A_{i_1}$, given by the absolutely convergent series*

$$Z^X_t = \sum_{i=1}^{d_{in}} \sum_{I\in\mathbb{W}_{d_\omega}} A_I C_i \; X^i_0 Sig(\omega^X)^I_{0,t} + \sum_{j=1}^{d_\xi} \sum_{I\in\mathbb{W}_{d_\omega}} A_I B_j \int_0^t Sig(\omega^X)^I_{s,t} d\xi^{X,j}_s \in \mathbb{R}^N. \tag{3.7}$$

Notice that this result *does not* rely on any assumptions on the nature of C, A_i, and B; for any such choice the result is a *time-independent linear map* on a *feature vector* given, at time t, by

$$T(X)_{0,t} := \left(X^i_0 \mathrm{Sig}(\omega^{\mathrm{X}})^I_{0,t}, \int_0^t \mathrm{Sig}(\omega^{\mathrm{X}})^I_{s,t} d\xi^{\mathrm{X},j}_s : i \in [d_{in}], I \in \mathbb{W}_{d_\omega}, j \in [d_\xi]\right) \tag{3.8}$$

Note also that the $A_I C_i$ and $A_I B_j$ terms can be expressed as *operator graphs*, which we call *ladders*, of the form

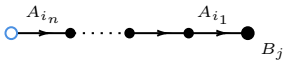


Figure 3.1: *Operator graph of $A_I B_j$. The operator graph of $A_I C_i$ is obtained by replacing the B_j node with the C_i node.*

In fact if we define $\tau_{I,i,0}$ to be the ladder of $A_I C_i$, $\tau_{I,0,j}$ to be the ladder of $A_I B_j$, and $\mathbb{L} \equiv \mathbb{L}(d_{in}, d_\omega, d_\xi)$ to be the set of all ladders of this form for $I \in \mathbb{W}_{d_\omega}$, $i \in \{1, \ldots, d_{in}\}$ and $j \in \{1, \ldots, d_\xi\}$, then the solution of the Linear CDE can be written as

$$Z^{\mathrm{X}}_t = \sum_{\tau\in\mathbb{L}} \mathbf{W}_\tau T(X)^\tau_{0,t} \in \mathbb{R}^N \tag{3.9}$$

with clear abuse of notation for the entries of the feature vector $T(X)_{0,t}$.

Remark. *A property highlighted by this expansion is the interpretability of these models. After training the CDEs one can compute the matrix multiplications and observe which entries of the signature the model chooses to take into consideration, to attend.*

3.3 Signature Kernel as Limit of SSMs

We have just seen how Linear CDEs can always be written in terms of the signature transform, which is a universal feature map on path space. It follows that to study the expressiveness of these systems, a crucial question is how well the signature information is stored and to what extent it can be recovered from the Linear CDEs.

We now proceed to study how signature kernels can be obtained as kernel limits, in an L^2 sense, of randomly initialised Linear CDEs. These systems therefore act as *random Fourier feature expansions* [106] for signature kernels and, in the limit $N \to \infty$, capture all the information contained in the signature kernel.

By writing the solution of the Linear CDE as a sum over ladders, we can leverage the results of chapter 1 to compute the distributional properties of these objects.

Assumption 3.3. *We assume the entries of all the matrices $A_i \in \mathbb{R}^{N\times N}$ to be independent Gaussian random variables with mean 0 and variance $\frac{1}{N}$, while, for any i, j, the entries of the vectors $C_i, B_j \in \mathbb{R}^N$ to be iid Gaussians with mean 0 and variance 1.*

Proposition 3.4. *Under theorem 3.3, for any $\tau, \eta \in \mathbb{L}$ we have that*

$$\left\| \frac{1}{N}\langle \mathbf{W}_\tau, \mathbf{W}_\eta \rangle_{\mathbb{R}^N} - \delta_{\tau,\eta} \right\|_{L^2} \leq \frac{2^{\frac{|E(\tau)|+|E(\eta)|}{2}}}{\sqrt{N}} (|E(\tau)| + |E(\eta)|)!! \tag{3.10}$$

Proof. This is an application of theorem 1.17 to the graphs $G = (\tau^\top \eta) \sqcup (\tau^\top \eta)$ since $\langle \mathbf{W}_\tau, \mathbf{W}_\eta \rangle_{\mathbb{R}^N} = \mathbf{W}_{\tau^\top \eta}$, the arguments will be the same used to prove theorems 2.10 and 2.15. Recall, theorem 1.17 gives us that

$$\mathbb{E}\left\{ \left(\mathbf{W}_{\tau^\top \eta} - \sum_{\psi \in \mathcal{P}_{\mathrm{A}}(\tau^\top \eta)} \mathbf{W}_{(\tau^\top \eta)_\psi} \right)^2 \right\} = \sum_{\phi \in \mathcal{P}_{\mathrm{AF}}(G)} \mathbf{W}_{G_\phi}. \tag{3.11}$$

For a pairing $\psi \in \mathcal{P}(\tau^\top \eta)$ to be atomic, the terminal leaves of τ and η must clearly be fixed to the same C_i or B_j. The pairing of these terminal leaves gives us a cycle with $|E(\tau)| + |E(\eta)|$ edges needing to be paired, and we know that the only way for this pairing to lead to a tree is if it produces a ladder. Note that the cycle is split in two halves, corresponding to τ and η, with edges running in opposite directions. The arguments employed in proving theorem 2.22 then show how the pairing leads to a ladder if and only if each edge of τ is paired with the one in η at the same depth, meaning that atomic pairings exist if and only if $\tau = \eta$ and in that case the only one is the "identity" pairing. Such identity pairing ψ_{id} leads to

$$\mathbf{W}_{(\tau^\top \tau)_\psi} = \frac{1}{N^{|E(\tau)|}} N^{|V(\tau)|} = \frac{1}{N^{|E(\tau)|}} N^{|E(\tau)|+1} = N.$$

Consider the pairings $\psi \in \mathcal{P}_{AF}(\tau^\top \eta \sqcup \tau^\top \eta)$, since trivially $\mathcal{P}_{AF}(G) \subseteq \mathcal{P}(G)$ we have

$$|\mathcal{P}_{AF}(G)| \leq |\mathcal{P}(G)| \leq (2 \cdot E(\tau^\top \eta))!! = (2 \cdot (E(\tau) + E(\eta)))!!$$

and also $(2 \cdot (E(\tau) + E(\eta)))!! = 2^{|E(\tau)|+|E(\eta)|}(|E(\tau)| + |E(\eta)|)!$ Moreover for any $\psi \in \mathcal{P}_{AF}(G)$ it must hold under our assumptions

$$|\mathbf{W}_{G_\psi}| \leq \frac{1}{N^{\frac{E(G)}{2}}} N^{\frac{V(G)}{2}} = \frac{1}{N^{E(\tau)+E(\eta)}} N^{E(\tau)+E(\eta)+1} = N$$

so that

$$|\sum_{\phi \in \mathcal{P}_{\mathrm{AF}}(G)} \mathbf{W}_{G_\phi}| \leq N 2^{|E(\tau)|+|E(\eta)|}(|E(\tau)| + |E(\eta)|)!$$

Putting all together we have

$$\mathbb{E}\left\{ \left(\mathbf{W}_{\tau^\top \eta} - N\delta_{\tau,\eta} \right)^2 \right\} \leq N 2^{|E(\tau)|+|E(\eta)|}(|E(\tau)| + |E(\eta)|)!$$

which, multiplying by $\frac{1}{N^2}$, proves that

$$\left\| \frac{1}{N} \langle \mathbf{W}_\tau, \mathbf{W}_\eta \rangle_{\mathbb{R}^N} - \delta_{\tau,\eta} \right\|_{L^2}^2 \leq \frac{2^{|E(\tau)|+|E(\eta)|}}{N}(|E(\tau)| + |E(\eta)|)! \tag{3.12}$$

to conclude note that by definition of double factorial

$$(|E(\tau)| + |E(\eta)|)! = (|E(\tau)| + |E(\eta)|)!! \; (|E(\tau)| + |E(\eta)| - 1)!!$$

so we have that

$$\sqrt{(|E(\tau)| + |E(\eta)|)!} \leq (|E(\tau)| + |E(\eta)|)!!$$

□

Remark. *We have found that the vectors* $\frac{1}{\sqrt{N}}\mathbf{W}_\tau$ *are* almost orthogonal, *becoming increasingly so as the dimension* N *increases. The quality of this orthogonality is controlled by the length of the ladders, and for any fixed length* L *there is an exponential number* $\mathcal{O}(d^L)$ *of ladders. Analogous to the Johnson-Lindenstrauss lemma [22, 60], we thus find, for a fixed dimension* $N \gg 1$*, an exponential number (in* d*) of almost orthogonal vectors. Intuitively, as* N *increases, the randomized signature has more freedom to accurately mimic the dynamic behavior of the true signature. The dynamics of the signature make it continuously move into orthogonal directions; the presence of more dimensions allows its randomized counterpart to act similarly, at least over small time frames.*

3.3.1 L^2 convergence to the signature kernel

Theorem 3.5. *For $N \in \mathbb{N}$ let $A_i^N \in \mathbb{R}^{N\times N}$ and $C_i^N, B_i^N \in \mathbb{R}^N$ satisfy theorem 3.3, let $Z^N(X) : [0,1] \to \mathbb{R}^N$ be the solution to (3.3) for these matrix choices. For any $X, Y \in \mathbb{X}$ we have that*

$$\frac{1}{N}\langle Z^N(X)_t, Z^N(Y)_s\rangle_{\mathbb{R}^N} \xrightarrow[N\to\infty]{L^2} \langle T(X)_{0,t}, T(Y)_{0,s}\rangle := \sum_{\tau\in\mathbb{L}} T(X)^\tau_{0,t} T(Y)^\tau_{0,s} \tag{3.13}$$

this is of order $\mathcal{O}(\frac{1}{\sqrt{N}})$, uniformly in $s,t \in [0,T]$ and on bounded sets of $X, Y \in \mathbb{X}$.

Proof. Note how

$$\begin{aligned}
&\frac{1}{N}\langle Z^N(X)_t, Z^N(Y)_s\rangle_{\mathbb{R}^N} - \langle T(X)_{0,t}, T(Y)_{0,s}\rangle \\
&= \sum_{\tau,\eta\in\mathbb{L}} \frac{1}{N}\langle \mathbf{W}_\tau, \mathbf{W}_\eta\rangle_{\mathbb{R}^N} T(X)^\tau_{0,t} T(Y)^\eta_{0,s} - \sum_{\tau\in\mathbb{L}} T(X)^\tau_{0,t} T(Y)^\tau_{0,s} \\
&= \sum_{\tau,\eta\in\mathbb{L}} \left(\frac{1}{N}\langle \mathbf{W}_\tau, \mathbf{W}_\eta\rangle_{\mathbb{R}^N} - \delta_{\tau,\eta}\right) T(X)^\tau_{0,t} T(Y)^\eta_{0,s}
\end{aligned}$$

so that

$$\begin{aligned}
&\left\| \frac{1}{N}\langle Z^N(X)_t, Z^N(Y)_s\rangle_{\mathbb{R}^N} - \langle T(X)_{0,t}, T(Y)_{0,s}\rangle \right\|_{L^2} \\
&\le \sum_{\tau,\eta\in\mathbb{L}} \left\| \frac{1}{N}\langle \mathbf{W}_\tau, \mathbf{W}_\eta\rangle_{\mathbb{R}^N} - \delta_{\tau,\eta} \right\|_{L^2} |T(X)^\tau_{0,t} T(Y)^\eta_{0,s}| \\
&\le \frac{1}{\sqrt{N}} \sum_{\tau,\eta\in\mathbb{L}} 2^{\frac{|E(\tau)|+|E(\eta)|}{2}} \, (|E(\tau)| + |E(\eta)|)!! \, |T(X)^\tau_{0,t} T(Y)^\eta_{0,s}| \\
&\le \frac{1}{\sqrt{N}} \sum_{\tau,\eta\in\mathbb{L}} 2^{\frac{|E(\tau)|+|E(\eta)|}{2}} \, (|E(\tau)| + |E(\eta)|)!! \, \frac{\|X\|^{|E(\tau)|} \|Y\|^{|E(\eta)|}}{|E(\tau)|!|E(\eta)|!}
\end{aligned}$$

where we have used both theorem 3.4 and the factorial decay of the signature.

Setting $D = d_{in} + d_\omega + d_\xi$ we have that for all $k \in \mathbb{N}$

$$|\{\tau \in \mathbb{L} : |E(\tau)| = k\}| \le D^{k+1}$$

so that

$$\begin{aligned}
&\frac{1}{\sqrt{N}} \sum_{\tau,\eta\in\mathbb{L}} 2^{\frac{|E(\tau)|+|E(\eta)|}{2}} \, (|E(\tau)| + |E(\eta)|)!! \, \frac{\|X\|^{|E(\tau)|} \|Y\|^{|E(\eta)|}}{|E(\tau)|!|E(\eta)|!} \\
&\le \frac{D}{\sqrt{N}} \sum_{i,j\in\mathbb{N}} D^{i+j} 2^{\frac{i+j}{2}} \, (i+j)!! \, \frac{\|X\|^i \|Y\|^j}{i!j!} \\
&\le \frac{D}{\sqrt{N}} \sum_{i,j\in\mathbb{N}} M^{i+j} \, \frac{(i+j)!!}{i!j!}
\end{aligned}$$

for some finite M. Note now that

$$\frac{(i+j)!!}{i!j!} \leq \frac{4^{i+j}}{(i+j)!!} \tag{3.14}$$

In fact, this is equivalent to

$$((i+j)!!)^2 \leq 4^{i+j} i!j!$$

and we have that

$$\begin{aligned}((i+j)!!)^2 &\leq (i+j+1)! = (i+j+1)(i+j)! \\ &\leq (i+j+1)2^{i+j} i!j! \leq 4^{i+j} i!j!\end{aligned}$$

where we have used the fact that $(i+j+1) \leq 2^{i+j}$ and the inequality

$$\frac{(i+j)!}{i!j!} = \binom{i+j}{i} \leq 2^{i+j}.$$

Thus we have that

$$\frac{D}{\sqrt{N}} \sum_{i,j \in \mathbb{N}} M^{i+j} \frac{(i+j)!!}{i!j!} \leq \frac{1}{\sqrt{N}} \sum_{i,j \in \mathbb{N}} \frac{\bar{M}^{i+j}}{(i+j)!!}$$

for some finite $\bar{M}$.

We want to show that this sum is convergent, to this end by [120][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

$$f(i,j) = \frac{\bar{M}^{i+j}}{(i+j)!!}$$

then

$$\sum_{k \in \mathbb{N}} f(\phi(k)) < \infty$$

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 $m+1$ choices of (i,j) such that $i+j=m$, corresponding

to the couples $(i, m-i)$ for $i = 0, \dots, m$, using the above ϕ it suffices to prove

$$\sum_{m\in\mathbb{N}} (m+1)\frac{\bar{M}^m}{m!!} \leq \sum_{m\in\mathbb{N}} 2^m \frac{\bar{M}^m}{m!!} = \sum_{m\in\mathbb{N}} \frac{(2\bar{M})^m}{m!!} \leq \sum_{m\in\mathbb{N}} \frac{(2\bar{M})^m}{\sqrt{m!}} < \infty$$

This is the case by the ratio test: set $a_m = \frac{(2\bar{M})^m}{\sqrt{m!}}$ and note that

$$\frac{a_{m+1}}{a_m} = \frac{2\bar{M}}{\sqrt{m+1}}$$

so that

$$\lim_{m\to\infty} \left|\frac{a_{m+1}}{a_m}\right| = 0$$

thus the series is absolutely convergent and we can conclude that

$$\left\| \frac{1}{N}\langle Z^N(X)_t, Z^N(Y)_s\rangle_{\mathbb{R}^N} - \langle T(X)_{0,t}, T(Y)_{0,s}\rangle \right\|_{L^2} = \mathcal{O}(\frac{1}{\sqrt{N}})$$

as desired. □

Remark. *Note how, given paths $X, Y \in \mathbb{X}$ such that $X_0^1, Y_0^1 = 1$, for the choices of $C_i = 0$ for $i > 1$, $\xi \equiv 0$ and $\omega \equiv Id$, the limit is given by the signature kernel i.e.*

$$\frac{1}{N}\langle Z^N(X)_t, Z^N(Y)_s\rangle_{\mathbb{R}^N} \xrightarrow[N\to\infty]{L^2} SigKer(X,Y)_{s,t} := \sum_{I\in\mathbb{W}_{d_{in}}} Sig(X)^I_{0,t} Sig(Y)^I_{0,s} \tag{3.15}$$

3.3.2 Convergence to a Gaussian process

We can use the L^2 limit to prove a distributional result on the readouts of the randomly initialised Linear CDEs we are considering, showing how the outputs converge to a Gaussian process indexed on path space with covariance structure given by the limiting kernel.

In the setting of theorem 3.5, under theorem 3.3, let $\mathcal{X} := \{X_1, \dots, X_m\} \subseteq \mathbb{X}$ and let $\mathbf{Z}^N_{\mathcal{X}} := [Z^N(X_1)_1, \dots, Z^N(X_m)_1] \in \mathbb{R}^{N\times m}$.

Proposition 3.6. *Let $v^N \in \mathbb{R}^N$ be sampled from $\mathcal{N}(0, \frac{1}{N})$. Then the vectors*

$$\Phi^N_{\mathcal{X}} := [\langle v^N, Z^N(X_j)_1\rangle_{\mathbb{R}^N}]_{j=1,\dots,m} = (\mathbf{Z}^N_{\mathcal{X}})^\top v^N \in \mathbb{R}^m$$

converge in distribution to a $\mathcal{N}(0, K_T(\mathcal{X},\mathcal{X})_{1,1})$ where

$$[K_T(\mathcal{X},\mathcal{X})_{1,1}]_{i,j} := \langle T(X_i)_{0,1}, T(X_j)_{0,1}\rangle_{L^2}$$

Proof. By *Lévy's continuity theorem* it suffices to study the limiting behavior of the

characteristic functions

$$\varphi_N(u) := \mathbb{E}[\exp\left\{iu^T\Phi_{\mathcal{X}}^N\right\}].$$

In particular we want to show that they converge pointwise to the characteristic function of a $\mathcal{N}(0, \mathrm{SigKer}(\mathcal{X},\mathcal{X})_{1,1})$, which would imply convergence in distribution.

Note how

$$\begin{aligned}\varphi_N(u) &= \mathbb{E}[\exp\left\{iu^T(\mathbf{Z}_{\mathcal{X}}^N)^\top v^N\right\}]\\ &= \mathbb{E}[\exp\left\{i(\mathbf{Z}_{\mathcal{X}}^N u)^\top v^N\right\}]\\ &= \mathbb{E}[\mathbb{E}[\exp\left\{i\langle \mathbf{Z}_{\mathcal{X}}^N u, v^N\rangle\right\}|\{A_i^N, C_j^N, B_k^N\}_{i,j,k}]]\end{aligned}$$

But if we know all the A_i^N, C_j^N, B_k^N then the $Z^N(X_j)_1$ are deterministic objects, and v^N is normally distributed; thus

$$\begin{aligned}\varphi_N(u) &= \mathbb{E}[\mathbb{E}[\exp\left\{i\langle \mathbf{Z}_{\mathcal{X}}^N u, v^N\rangle\right\}|\{A_i^N, C_j^N, B_k^N\}_{i,j,k}]]\\ &= \mathbb{E}[\exp\Big\{-\frac{1}{2N}\langle \mathbf{Z}_{\mathcal{X}}^N u, \mathbf{Z}_{\mathcal{X}}^N u\rangle\Big\}]\\ &= \mathbb{E}[\exp\Big\{-\frac{1}{2N}\langle u, (\mathbf{Z}_{\mathcal{X}}^N)^\top\mathbf{Z}_{\mathcal{X}}^N u\rangle\Big\}]\\ &= E[\exp\Big\{-\frac{1}{2}\langle u, G_{\mathcal{X}}^N u\rangle\Big\}]\end{aligned}$$

where

$$G_{\mathcal{X}}^N = \frac{1}{N}(\mathbf{Z}_{\mathcal{X}}^N)^\top\mathbf{Z}_{\mathcal{X}}^N \in \mathbb{R}^{m\times m} \text{ i.e. } [G_{\mathcal{X}}^N]_{i,j} := \frac{1}{N}\langle Z^N(X_i)_1, Z^N(X_j)_1\rangle$$

Now, since we have proven that

$$\frac{1}{N}\langle Z^N(X_i)_1, Z^N(X_j)_1\rangle \xrightarrow[N\to\infty]{\mathbb{L}^2} [K_T(\mathcal{X},\mathcal{X})_{1,1}]_{i,j}$$

we have

$$G_{\mathcal{X}}^N \xrightarrow[N\to\infty]{\mathbb{L}^2} K_T(\mathcal{X},\mathcal{X})_{1,1} =: \Sigma$$

We want to show that this implies for all $u \in \mathbb{R}^m$ the convergence

$$\mathbb{E}[\exp\Big\{-\frac{1}{2}\langle u, G_{\mathcal{X}}^N u\rangle\Big\}] \xrightarrow[N\to\infty]{} \exp\Big\{-\frac{1}{2}\langle u, \Sigma u\rangle\Big\}$$

which would allow us to conclude. To this end note how $\mathbb{L}^2$ convergence implies convergence in distribution, by the classical *Portmanteau theorem* we must have

$$\mathbb{E}[h(G_{\mathcal{X}}^N)] \xrightarrow[N\to\infty]{} \mathbb{E}[h(\Sigma)] = h(\Sigma)$$

for every continuous and bounded $h : \mathbb{R}^{m\times m} \to \mathbb{R}$

Fix $u \in \mathbb{R}^m$ and consider

$$f_u : \mathbb{R}^{m\times m} \to \mathbb{R} \text{ s.t. } A \mapsto \exp\Big\{-\frac{1}{2}\langle u, Au\rangle\Big\}$$

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:

$$G_{\mathcal{X}}^N = \frac{1}{N}(\mathbf{Z}_{\mathcal{X}}^N)^\top \mathbf{Z}_{\mathcal{X}}^N$$

thus for any $u \in \mathbb{R}^m$ it holds that

$$u^T G_{\mathcal{X}}^N u = u^T \frac{1}{N}(\mathbf{Z}_{\mathcal{X}}^N)^\top \mathbf{Z}_{\mathcal{X}}^N u = (\frac{1}{\sqrt{N}}\mathbf{Z}_{\mathcal{X}}^N u)^T \frac{1}{\sqrt{N}}\mathbf{Z}_{\mathcal{X}}^N u = \left\|\frac{1}{\sqrt{N}}\mathbf{Z}_{\mathcal{X}}^N u\right\|_2^2 \geq 0$$

Since the $G_{\mathcal{X}}^N \in \mathbb{R}^{m\times m}$ are always 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

$$\mathbb{E}[f_u(G_{\mathcal{X}}^N)] = \mathbb{E}[(\psi \circ f_u)(G_{\mathcal{X}}^N)] \xrightarrow[N\to\infty]{} (\psi \circ f_u)(\Sigma) = f_u(\Sigma)$$

where we have used the semidefinitiveness of Σ. With this we finally conclude that

$$\varphi_N(u) \xrightarrow[N\to\infty]{} \exp\Big\{-\frac{1}{2}\langle u, \Sigma u\rangle\Big\}.$$

□

3.4 Expressive Power of SSMs

Equipped with the previous results, we now embark on answering the following question:

"*What is the most one can achieve using a recurrence which is linear in the hidden state and potentially non-linear in the input?*"

3.4.1 Characterization of the closure: Dense Case

For any path $\gamma \in C_{1,0}([0,1], \mathbb{R}^{d_\gamma})$ and any sub-interval $[s,t] \subset [0,1]$, we denote by $\gamma_{[s,t]} \in C_{1,0}([0,1], \mathbb{R}^{d_\gamma})$ the path $\gamma_{[s,t]}(u) = \gamma_{t\wedge u} - \gamma_{s\wedge u}$, where $a \wedge b = \min(a,b)$.

Statement of Main Result

Here we present the unified expressivity result in its most general form:

Theorem 3.7. *Fix a compact set $\mathbb{K} \subseteq \mathbb{X}$ and continuous gates $(\cdot)_0, \omega, \xi$ with* [3] $\omega_t^{X,1} \equiv t$ *and $\omega_t^{X,2} \equiv t^2$. For any $\epsilon > 0$ and any*

$$F \in \left\{ (X,t) \mapsto \Psi(\omega^X_{[0,t]}) \cdot X_0 + \int_0^t \Phi(\omega^X_{[s,t]}) \cdot d\xi^X_s \right\} \tag{3.16}$$

where $\Psi \in C^0(C_{1,0}; \mathbb{R}^{d_{in}})$ and $\Phi \in C^0(C_{1,0}; \mathbb{R}^{d_\xi})$, there exist a choice of hidden dimension $N \geq 1$ and parameters $v \in \mathbb{R}^N, A_i \in \mathbb{R}^{N\times N}, B \in \mathbb{R}^{N\times d_\xi}, C \in \mathbb{R}^{N\times d_{in}}$ such that

$$\sup_{(X,t)\in\mathbb{K}\times[0,1]} |F(X,t) - \langle v, Z_t^X \rangle| \leq \epsilon \tag{3.17}$$

Moreover generic parameters suffice with high probability in the sense that under theorem 3.3 the following holds:

$$\lim_{N\to\infty} \mathbb{P}\Big[\exists v \in \mathbb{R}^N : (3.17) \text{ holds}\Big] = 1$$

Moreover the "reverse" also holds in the sense that, given any choice of matrices A_i, B, C there is an ϵ-close map F in the family.

Intuition on the closure result: role of gates

We note how theorem 3.7 shows that dense Linear CDEs have drastically superior expressive power compared to dense linear RNNs, for which $\omega_t^{\mathrm{X}} = t$ and $\xi_t^{\mathrm{X}} = \int_0^t X_s ds$. This contrast is to be attributed completely to the gate ω.

Linear RNNs. We can see theorem 3.7 as a generalization of the *Universal Approximation for Linear RNNs* presented by Li et al. [72][Thm. 7] for generic gates ω, ξ. In fact, their setting is restricted to $\omega_t^{\mathrm{X}} = t$ and $\xi_t^{\mathrm{X}} = \int_0^t X_s ds$. This is also the case for S4, S5 and the LRU [46, 95, 118]– which are linear RNNs at test time. Then the only information contained in $\omega_{[s,t]}$ is the increment $t - s$ so that family (3.16) reduces to

$$\left\{ (X,t) \mapsto \psi(t) + \int_0^t \phi(t-s) \cdot X_s ds \right\}.$$

[3] These assumptions are of technical nature and can be relaxed at the cost of talking about *tree-like* equivalence of paths *cf.* [49]. Intuitively $\omega_t^{\mathrm{X},1} = t$ ensures general $\mathbb{X}$-uniformity while $\omega_t^{\mathrm{X},2} = t^2$ ensures $[0,1]$-uniformity.

This is, fundamentally, the set of linear filters on the input. As shown in [44], such processing is unable to modulate information flow in-context.

How does a Linear CDE process inputs? Take without loss in generality[4] $\omega_t^{\mathrm{X}} = X_t$. The first term in the function class $\{\Psi(X_{[0,t]}) + \int_0^t \Phi(X_{[s,t]}) \cdot d\xi_s^{\mathrm{X}}\}$ is already enough to establish that the output CZ_t^{X} is a nonlinear function of all previously seen inputs $X_{[0,t]}$ – interestingly, ξ could be set to zero. The term $\Psi(X_{[0,t]})$ is however non-trivial only in the $Z_0 \neq 0$ case: picking $\xi_t^{\mathrm{X}} = 0 \in \mathbb{R}^N$ for all t indeed leads to $dZ_t^{\mathrm{X}} = \left[\sum_{i=1}^{d_\omega} A_i d\omega_t^{\mathrm{X},i}\right] Z_t^{\mathrm{X}}$, which is evolving only if $Z_0 \neq 0$. While this setting is interesting and suggests a clear direction for future research, a case more similar to Mamba (see section 3.4.2) is $Z_0 = 0$ and (ignoring input gating) $\xi_t^{\mathrm{X}} = \int_0^t X_s ds$. Here, we can approximate arbitrarily well outputs of the form

$$\left\{(X,t) \mapsto \int_0^t \Phi(X_{[s,t]}) \cdot X_s ds\right\} \tag{3.18}$$

where Φ is any continuous function of the input path, restricted to the portion $[s,t]$. This clearly shows that dense Linear CDEs are capable of *context-dependent filtering*: the output is again a linear combination of previously seen inputs, but weights are not predetermined as in Linear RNNs (S4) – they are a function of the context. A similar yet less powerful processing is happening in the diagonal setting, which we explore next.

Remark. *The specific restriction of ω to subsets of $[0,1]$ is a crucial part of (3.16). The family of approximable maps* does not *include* all *path-to-path causal*[5] *functions $t \mapsto Y_t^X$ but a subset of them, of type $t \mapsto Y_t^X := \Psi(\omega_{[0,t]}^X)$, satisfying the specific* time-homogeneity *specified by the form of the restriction, akin to that in Li et al. [72].*

The feature map T and its RKHS

Definition 3.8. *Let $\mathbb{W}_{d_{in},d_\omega,d_\xi}$ be the set of words in the alphabet*

$$\mathcal{A}_{d_{in},d_\omega,d_\xi} := \{\boldsymbol{e}_i\}_{i=1,\dots,d_{in}} \cup \{\boldsymbol{\epsilon}_j^\xi\}_{j=1,\dots,d_\xi} \cup \{\boldsymbol{\epsilon}_k^\omega\}_{k=1,\dots,d_\omega}$$

Fix the gates $(\cdot)_0, \omega, \xi$, we define $T(X) : [0,1] \to l^2(\mathbb{W}_{d_{in},d_\omega,d_\xi}) \subseteq T((\mathcal{A}_{d_{in},d_\omega,d_\xi}))$ as the unique solution to:

$$T(X)_{0,t} = \sum_{i=1}^{d} X_0^i \boldsymbol{e}_i + \sum_{j=1}^{d_\xi} \xi_t^{X,j} \boldsymbol{\epsilon}_j^\xi + \sum_{k=1}^{d_\omega} \int_0^t T(X)_{0,s} \, d\omega_s^{X,k} \otimes \boldsymbol{\epsilon}_k^\omega \tag{3.19}$$

[4] Let us ignore the terms t, t^2 at this stage; their role is purely technical due to possible path equivalences.

[5] A *causal* map is one which does not "look in the future".

In fact one readily sees that the only non-zero terms of $T(X)_{0,t}$ are

$$\begin{aligned}
&\langle T(X)_{0,t}, \boldsymbol{e}_i \rangle = X_0^i \\
&\langle T(X)_{0,t}, \boldsymbol{e}_i \otimes \boldsymbol{\epsilon}^{\omega}_{Ik} \rangle = \int_0^t \langle T(X)_{0,s}, \boldsymbol{e}_i \otimes \boldsymbol{\epsilon}^{\omega}_{I} \rangle d\omega_s^{\mathrm{X},k} = X_0^i \ \mathrm{Sig}(\omega^{\mathrm{X}})^{Ik}_{0,t} \\
&\langle T(X)_{0,t}, \boldsymbol{\epsilon}^{\xi}_{j} \rangle = \xi_t^{\mathrm{X},j} = \int_0^t d\xi_s^{\mathrm{X},j} \\
&\langle T(X)_{0,t}, \boldsymbol{\epsilon}^{\xi}_{j} \otimes \boldsymbol{\epsilon}^{\omega}_{Ik} \rangle = \int_0^t \langle T(X)_{0,s}, \boldsymbol{\epsilon}^{\xi}_{j} \otimes \boldsymbol{\epsilon}^{\omega}_{I} \rangle d\omega_s^{\mathrm{X},k} = \int_{s=0}^t \int_{r=0}^s \mathrm{Sig}(\omega^{\mathrm{X}})^{I}_{r,s} d\xi_r^{\mathrm{X},j} d\omega_s^{\mathrm{X},k} \\
&\quad = \int_{r=0}^t \int_{s=r}^t \mathrm{Sig}(\omega^{\mathrm{X}})^{I}_{r,s} d\omega_s^{\mathrm{X},k} d\xi_r^{\mathrm{X},j} = \int_0^t \mathrm{Sig}(\omega^{\mathrm{X}})^{Ik}_{r,t} \ d\xi_r^{\mathrm{X},j}
\end{aligned}$$

This is similar to the tensor-valued CDE defining the signature as a tensor *i.e.* [112]

$$\mathrm{Sig}(\omega)_{0,t} = () + \int_0^t \mathrm{Sig}(\omega)_{0,s} \otimes d\omega_s$$

with the addition of two terms to track X_0 and ξ^{X}. One could also understand $T(X)_{s,t}$ as a sub-tensor of

$$X_0 \otimes \mathrm{Sig}((\omega^{\mathrm{X}}, \xi^{\mathrm{X}}))_{s,t}$$

but in doing this one would have to explicitly ignore most of the terms of this vector; the CDE (3.19) does exactly this, but implicitly. In any case the subtensor view shows that $T : \mathbb{X} \times [0,1]^2 \to l^2(\mathbb{W}_{d_{in},d_\omega,d_\xi})$ is well defined and continuous.

To the feature map $T(\cdot)_{0,t}$ with values in the Hilbert space $l^2(\mathbb{W}_{d_{in},d_\omega,d_\xi})$ is then associated a Reproducing Kernel Hilbert Space [4], where the Kernel is the one induced by the l^2 product, which we denote by

$$\mathcal{H}_t^{(\cdot)_0,\omega,\eta} \subseteq C^0(\mathbb{X}; \mathbb{R}) \tag{3.20}$$

Classical RKHS theory tells us that we can characterize its elements as:

Proposition 3.9. *A map $F(\cdot)_t : \mathbb{X} \to \mathbb{R}$ is an element of $\mathcal{H}_t^{(\cdot)_0,\omega,\eta}$ if and only if it is of the form*

$$F(x)_t = \sum_{i=1}^{d_{in}} X_0^i \langle \alpha_i, Sig(\omega^X)_{0,t} \rangle + \sum_{j=1}^{d_\xi} \int_0^t \langle \beta_j, Sig(\omega^X)_{s,t} \rangle d\xi_s^{X,j} \tag{3.21}$$

for $\alpha_i, \beta_j \in l^2(\mathbb{W}_{d_\omega})$. Moreover $\|F(\cdot)_t\|^2_{\mathcal{H}_t^{(\cdot)_0,\omega,\eta}}$ is equal to the minimal value of

$$\sum_{i=1}^{d_{in}} \|\alpha_i\|^2_{l^2(\mathbb{W}_{d_\omega})} + \sum_{j=1}^{d_\xi} \|\beta_j\|^2_{l^2(\mathbb{W}_{d_\omega})}$$

taken over those γ, β for which the above equality holds.

Signature kernels [112] are a class of *universal kernels* on sequential data which have

received attention in recent years thanks to their efficiency in handling path-dependent problems [17, 57, 69, 70, 80, 88, 96, 113]. Just as signature kernels, the kernel associated to $T(X)_{0,t}$ can be explicitly written as the solution of a two-parameter CDE:

Lemma 3.10. *Let* $\mathcal{K}^{X,Y}(s,t) := \langle T(X)_{0,s}, T(Y)_{0,t}\rangle_{l^2}$ *then*

$$\mathcal{K}^{X,Y}(s,t) = \langle X_0, Y_0\rangle + \langle \xi_s^X, \xi_t^Y\rangle + \int_{\eta=0}^{s}\int_{\tau=0}^{t} \mathcal{K}^{X,Y}(\eta,\tau)\langle d\omega_\eta^X, d\omega_\tau^Y\rangle \tag{3.22}$$

or, directly in terms of Signature, also

$$\mathcal{K}^{X,Y}(s,t) = \langle X_0, Y_0\rangle\langle Sig(\omega^X)_{0,s}, Sig(\omega^Y)_{0,t}\rangle + \int_{\eta=0}^{s}\int_{\tau=0}^{t} \langle Sig(\omega^X)_{\eta,s}, Sig(\omega^Y)_{\tau,t}\rangle\langle d\xi_\eta^X, d\xi_\tau^Y\rangle \tag{3.23}$$

Proof. The first expression follows immediately from (3.19) the second one by summing the products of $T(X)_{0,s}$'s and $T(Y)_{0,t}$'s entries given above. □

Definition 3.11. *Define* $\mathcal{H}_{[0,1]}^{(\cdot)_0,\omega,\eta} \subseteq C^0(\mathbb{X}\times[0,1];\mathbb{R})$ *as the space of functions of form*

$$(X,t) \mapsto \sum_{i=1}^{d_{in}} X_0^i\langle \alpha_i, Sig(\omega^X)_{0,t}\rangle + \sum_{j=1}^{d_\xi}\int_0^t \langle \beta_j, Sig(\omega^X)_{s,t}\rangle d\xi_s^{X,j} \tag{3.24}$$

for $\alpha_i, \beta_j \in l^2(\mathbb{W}_{d_\omega})$*. Thus for all* $t\in[0,1]$ *and* $F \in \mathcal{H}_{[0,1]}^{(\cdot)_0,\omega,\eta}$ *it holds* $F(\cdot,t) \in \mathcal{H}_t^{(\cdot)_0,\omega,\eta}$*.*

Linear maps on Z_t^X are close to the RKHS

The following proposition will show how linear maps on Z_t^X cannot be more expressive than elements of the RKHS $\mathcal{H}_{[0,1]}^{(\cdot)_0,\omega,\eta}$ since their closure is in the closure of $\mathcal{H}_{[0,1]}^{(\cdot)_0,\omega,\eta}$. In this precise sense these spaces act like upper bounds to expressiveness.

Proposition 3.12. *Assume* $\mathbb{X}$ *compact. Consider fixed the gates and* $A_i \in \mathbb{R}^{N\times N}, B \in \mathbb{R}^{N\times d_\xi}$ *and* $C \in \mathbb{R}^{N\times d_{in}}$*. Consider a linear readout* $v \in \mathbb{R}^N$*. For any* $\epsilon > 0$ *there exist choices of* $\alpha_i, \beta_j \in l^2(\mathbb{W}_{d_\omega})$ *defining a* $F \in \mathcal{H}_{[0,1]}^{(\cdot)_0,\omega,\eta}$*, such that*

$$\sup_{(X,t)\in\mathbb{X}\times[0,1]} |\langle v, Z_t^X\rangle - F(X,t)| \le \epsilon. \tag{3.25}$$

In other words, linear maps on the Z_t^X *are in the uniform closure of* $\mathcal{H}_{[0,1]}^{(\cdot)_0,\omega,\eta}$*.*

Proof. By (3.2) we see that $\langle v, Z_t^X\rangle$ is a linear map on $T(X)_{0,t}$ with coefficients

$$v^\top A_I B_j \quad v^\top A_I C_i$$

using Cauchy-Schwartz it's moreover easy to see the existence of a constant $\lambda \ge 0$ such

that for all I, i, j one has

$$|v^\top A_I B_j| \leq \lambda^{|I|} \quad |v^\top A_I C_i| \leq \lambda^{|I|}.$$

Since $|\mathrm{Sig}(\omega)^I_{s,t}| \leq \frac{1}{|I|!} \|\omega\|^{|I|}_{1-var,[s,t]}$ we have that, given an integer $M \geq 0$, the bound

$$\begin{aligned} R_M(t) &:= \left| \sum_{i=1}^{d_{in}} \sum_{|I|\geq M} v^\top A_I C_i \, X_0^i \mathrm{Sig}(\omega^{\mathrm{X}})^I_{0,t} + \sum_{j=1}^{d_\xi} \sum_{|I|\geq M} v^\top A_I B_j \int_0^t \mathrm{Sig}(\omega^{\mathrm{X}})^I_{s,t} d\xi^{\mathrm{X},j}_s \right| \\ &\leq (\|X_0\|_1 + \|\xi^{\mathrm{X}}_t\|_1) \sum_{m=M}^{\infty} \frac{\lambda^m d_\omega^m \|\omega^{\mathrm{X}}\|^m_{1-var,[0,t]}}{m!} \\ &\leq (\|X_0\|_1 + \|\xi^{\mathrm{X}}_t\|_1) \sum_{m=M}^{\infty} \frac{\lambda^m d_\omega^m \|\omega^{\mathrm{X}}\|^m_{1-var,[0,1]}}{m!} \leq K \sum_{m=M}^{\infty} \frac{(\lambda d_\omega K)^m}{m!} \end{aligned}$$

where $K \geq 0$ is a constant which must exist by compactness of $\mathbb{X}$ and continuity of the gates. Since $K \sum_{m=M}^{\infty} \frac{(\lambda d_\omega K)^m}{m!}$ is just the tail of the taylor expansion of $K e^{\lambda d_\omega K}$ there must be an M such that $\sup_{t\in[0,1]} R_M(t) \leq \epsilon$. But then the choice

$$\alpha^I_i := v^\top A_I C_i \, \mathbb{I}(|I| < M) \quad \beta^I_j := v^\top A_I B_j \, \mathbb{I}(|I| < M)$$

suffices for the required bound.

□

Uniform closure of the RKHS

Now that we have established the theoretical interest of the $\mathcal{H}^{(\cdot)_0,\omega,\eta}_{[0,1]}$ we proceed to characterize which maps $(X,t) \to \mathbb{R}$ can be uniformly approximated through them.

Proposition 3.13. *Fix a compact input set $\mathbb{X}$ and continuous gates $(\cdot)_0, \omega, \xi$ with $\omega^{X,1}_t \equiv t$ and $\omega^{X,2}_t \equiv t^2$. For any $\epsilon > 0$ and any*

$$F \in \left\{ (X,t) \mapsto \Psi(\omega^X_{[0,t]}) \cdot X_0 + \int_0^t \Phi(\omega^X_{[s,t]}) \cdot d\xi^X_s \right\} \tag{3.26}$$

where $\Psi \in C^0(C_{1,0}; \mathbb{R}^{d_{in}})$ and $\Phi \in C^0(C_{1,0}; \mathbb{R}^{d_\xi})$, there exist a $G \in \mathcal{H}^{(\cdot)_0,\omega,\eta}_{[0,1]}$ such that

$$\sup_{(X,t)\in\mathbb{X}\times[0,1]} |F(X,t) - G(X,t)| \leq \epsilon \tag{3.27}$$

Proof. Note first that the map

$$\mathbb{X} \times [0,1] \times [0,1] \to C_{1,0}([0,1]; \mathbb{R}^{d_\omega}) \quad (X,s,t) \mapsto \omega^{\mathrm{X}}_{[s,t]}$$

is a continuous map from a compact space, thus the image must be compact too. Moreover by theorem C.12 the Signature separates the points in this image. Since any

G as above has form

$$G(X,t) = \sum_{i=1}^{d_{in}} X_0^i \langle \alpha_i, \mathrm{Sig}(\omega^{\mathrm{X}})_{0,t} \rangle + \sum_{j=1}^{d_\xi} \int_0^t \langle \beta_j, \mathrm{Sig}(\omega^{\mathrm{X}})_{s,t} \rangle d\xi_s^{\mathrm{X},j}$$
$$= \sum_{i=1}^{d_{in}} X_0^i \langle \alpha_i, \mathrm{Sig}(\omega^{\mathrm{X}}_{[0,t]})_{0,1} \rangle + \sum_{j=1}^{d_\xi} \int_0^t \langle \beta_j, \mathrm{Sig}(\omega^{\mathrm{X}}_{[s,t]})_{0,1} \rangle d\xi_s^{\mathrm{X},j}$$

the proof follows from the uniform density on compact sets of linear functionals on the (truncated) Signature (theorem C.10), by also uniformly bounding thanks to compactness and continuity the norms of X_0 and ξ_1^{X}. □

Remark. *The specific restriction of ω to subsets of $[0,1]$ is a crucial part of the result. The family of approximable maps* does not *include* all *path-to-path causal*[6] *functions $t \mapsto Y_t^X$ but a subset of them, of type $t \mapsto Y_t^X := \Psi(\omega^X_{[0,t]})$, satisfying the specific* time-homogeneity *specified by the form of the restriction, akin to that in Li et al. [72].*

Generic Weights are fully expressive

We have seen how linear maps on Z_t^{X} are in the uniform closure of $\mathcal{H}_{[0,1]}^{(\cdot)_0,\omega,\eta}$, and we have explicitly characterized this closure. It is then natural to ask "how much" of this closure the Z_t^{X} are able to "explore". The present section not only shows that the Z_t^{X} "explore" all the closure, but also that a generic choice of weights is enough to eventually do this with high probability.

The fact that these maps are "universal" in the above sense is not surprising, since it is well known that Linear CDEs are universal for path-to-point tasks *cf.* [65], what is surprising is that this universality can be achieved probabilistically with one of the *standard* parametrizations used in ML practice (LeCun)[7].

Theorem 3.14. *Fix $\mathbb{X}$ compact and $\epsilon > 0$. For all $F \in \mathcal{H}_{[0,1]}^{(\cdot)_0,\omega,\eta}$ there exist a choice of hidden dimension $N \geq 1$ and parameters $v \in \mathbb{R}^N, A_i \in \mathbb{R}^{N\times N}, B \in \mathbb{R}^{N\times d_\xi}, C \in \mathbb{R}^{N\times d_{in}}$ such that*

$$\sup_{(X,t)\in\mathbb{X}\times[0,1]} |F(X,t) - \langle v, Z_t^X \rangle| \leq \epsilon \tag{3.28}$$

Moreover generic weight choices suffice with high probability, in the sense that under theorem 3.3 the following holds

$$\lim_{N\to\infty} \mathbb{P}\Big[\exists v \in \mathbb{R}^N : (3.28) \text{ holds } \Big] = 1$$

We propose two proofs, the first one of a deterministic character concerns the first claim in the theorem, the second one is probabilistic and concerns the whole result.

[6] A *causal* map is one which does not "look in the future".

[7] It can be proved, using the results of [28], that the sampling measure does not have to be Gaussian if it satisfies certain moment requirements.

The deterministic proof follows the same arguments employed by Kidger [65] and is included to highlight the main idea of the probabilistic result, which reduces to a "spin" on the central argument of this proof.

Deterministic Proof. Any $F \in \mathcal{H}_{[0,1]}^{(\cdot)_0,\omega,\eta}$ has form

$$F(X,t) = \sum_{i=1}^{d_{in}} X_0^i \langle \alpha_i, \mathrm{Sig}(\omega^{\mathrm{X}})_{0,t} \rangle + \sum_{j=1}^{d_\xi} \int_0^t \langle \beta_j, \mathrm{Sig}(\omega^{\mathrm{X}})_{s,t} \rangle d\xi_s^{\mathrm{X},j} \tag{3.29}$$

for fixed $\alpha_i, \beta_j \in l^2(\mathbb{W}_{d_\omega})$. Consider an integer $M \geq 0$ such that

$$\sup_{(x,t)\in\mathbb{X}\times[0,1]} |F(X,t) - \sum_{i=1}^{d_{in}} X_0^i \langle \pi_M \alpha_i, \mathrm{Sig}(\omega^{\mathrm{X}})_{0,t} \rangle - \sum_{j=1}^{d_\xi} \int_0^t \langle \pi_M \beta_j, \mathrm{Sig}(\omega^{\mathrm{X}})_{s,t} \rangle d\xi_s^{\mathrm{X},j}| \leq \epsilon \tag{3.30}$$

where π_M is the truncation at length M.

Fix $d = d_{in} + d_\omega + d_\xi$. Consider $\mu(M,d) \in \mathbb{N}$ such that $\mathbb{R}^{\mu(M,d)} \simeq T^M(\mathbb{R}^d)$. We are going to write $e_I \in \mathbb{R}^{\mu(M,d)}$ to mean the image of $e_I \in T^M(\mathbb{R}^d)$ through this identification. Note that $(\cdot) \otimes_M e_k : T^M(\mathbb{R}^d) \to T^M(\mathbb{R}^d)$ is a linear map, it does then correspond to a matrix $\Lambda_k \in \mathbb{R}^{\mu(M,d)\times\mu(M,d)}$. Write $\varepsilon_j^\xi := e_{d_{in}+j}$ for $j = 1, \ldots, d_\xi$ and $\varepsilon_k^\omega := e_{d_{in}+d_\xi+k}$ for $k = 1, \ldots, d_\omega$. Then the solution to

$$\mathbb{R}^{\mu(M,d)} \ni \tilde{Z}_t = \sum_{i=1}^{d_{in}} X_0^i e_i + \sum_{j=1}^{d_\xi} \xi_t^{\mathrm{X},j} \varepsilon_j^B + \sum_{k=1}^{d_\omega} \int_0^t \Lambda_{d_{in}+d_\xi+k} \tilde{Z}_s d\omega_s^{\mathrm{X},k} \tag{3.31}$$

is the object in $\mathbb{R}^{\mu(M,d)}$ corresponding to the truncated tensor $\pi_M(T(X)_{0,t})$.

This $\tilde{Z}_t$ is of the form Z_t^{X} with $N = \mu(M,d)$, $A_k = \Lambda_{d_{in}+d_\xi+k}$, $B = \left[\varepsilon_1^\xi | \cdots | \varepsilon_{d_\xi}^\xi\right]$ and $C = [e_1 | \cdots | e_{d_{in}}]$.

In particular note how these matrices are such that

$$e_J^T A_I C_i = \mathbb{I}(e_J = e_i \otimes \varepsilon_I^\omega), \quad e_J^T A_I B_j = \mathbb{I}(e_J = \varepsilon_j^\xi \otimes \varepsilon_I^\omega), \tag{3.32}$$

since it holds that

$$A_I C_i = e_i \otimes \varepsilon_I^\omega, \quad A_I B_j = \varepsilon_j^\xi \otimes \varepsilon_I^\omega, \tag{3.33}$$

and for all $|I| > M$ one has necessarily $A_I = 0$.

Our strategy is that of using these equaities to create a vector $v \in \mathbb{R}^N$ corresponding to the $\pi_M \alpha_i$ and $\pi_M \beta_j$. Define the vector

$$v := \sum_{i=1}^{d_{in}} \sum_{|I|\leq M} \alpha_i^I \; e_i \otimes \varepsilon_I^\omega + \sum_{j=1}^{d_\omega} \sum_{|I|\leq M} \beta_j^I \; \varepsilon_j^\xi \otimes \varepsilon_I^\omega \in \mathbb{R}^{\mu(M,d)} \tag{3.34}$$

Then expanding Z_t^{X} and using the equalities above one has

$$\begin{aligned}\langle v, Z_t^{\mathrm{X}}\rangle &= \sum_{i=1}^{d_{in}} \sum_I v^\top A_I C_i \; X_0^i \mathrm{Sig}(\omega^{\mathrm{X}})_{0,t} + \sum_{j=1}^{d_\omega} \sum_I v^\top A_I B_j \; \int_0^t \mathrm{Sig}(\omega^{\mathrm{X}})_{s,t} d\xi_s^{\mathrm{X},j} \\ &= \sum_{i=1}^{d_{in}} \sum_{|I|\le M} \alpha_i^I X_0^i \mathrm{Sig}(\omega^{\mathrm{X}})_{0,t} + \sum_{j=1}^{d_\xi} \sum_{|I|\le M} \beta_j^I \int_0^t \mathrm{Sig}(\omega^{\mathrm{X}})_{s,t} d\xi_s^{\mathrm{X},j} \\ &= \sum_{i=1}^{d_{in}} X_0^i \langle \pi_M \alpha_i, \mathrm{Sig}(\omega^{\mathrm{X}})_{0,t}\rangle + \sum_{j=1}^{d_\xi} \int_0^t \langle \pi_M \beta_j, \mathrm{Sig}(\omega^{\mathrm{X}})_{s,t}\rangle d\xi_s^{\mathrm{X},j}\end{aligned}$$

proving that for such v and Z^{X} it holds

$$\sup_{(x,t)\in \mathbb{X}\times[0,1]} |F(X,t) - \langle v, Z_t^{\mathrm{X}}\rangle| \le \epsilon \tag{3.35}$$

□

The crucial ingredient for the success of this proof is the possibility to recreate the space $T^M(\mathbb{R}^{d_{in}+d_\omega+d_\xi})$ as an euclidean space. To do this one needs $\mu(M, d_{in}+d_\omega+d_\xi) \sim (d_{in}+d_\omega+d_\xi)^M$ orthogonal vectors and a way to express them using the matrices A_i, B and C, the essential equations which capture this are given by (3.32).

The core idea of the following probabilistic proof of this same result is that of allowing for some error in (3.32), so the idea is that of exhibiting only *approximately* orthogonal vectors. At the cost of losing exactness, one can leverage results of the Johnson-Lindenstrauss [22] type to find on the order of $\sim e^{\varepsilon^2 N}$ vectors in $\mathbb{R}^N$ orthogonal up to an ε error, using random projections. This idea in the context of Signature goes back to Cuchiero et al. [20], and allows for much smaller hidden dimensions.

Probabilistic Proof. Any $F \in \mathcal{H}_{[0,1]}^{(\cdot)_0,\omega,\eta}$ has form

$$F(X,t) = \sum_{i=1}^{d_{in}} X_0^i \langle \alpha_i, \mathrm{Sig}(\omega^{\mathrm{X}})_{0,t}\rangle + \sum_{j=1}^{d_\xi} \int_0^t \langle \beta_j, \mathrm{Sig}(\omega^{\mathrm{X}})_{s,t}\rangle d\xi_s^{\mathrm{X},j} \tag{3.36}$$

for fixed $\alpha_i, \beta_j \in l^2(\mathbb{W}_{d_\omega})$. Consider an integer $M \ge 0$ such that

$$\sup_{(x,t)\in \mathbb{X}\times[0,1]} |F(X,t) - \sum_{i=1}^{d_{in}} X_0^i \langle \pi_M \alpha_i, \mathrm{Sig}(\omega^{\mathrm{X}})_{0,t}\rangle - \sum_{j=1}^{d_\xi} \int_0^t \langle \pi_M \beta_j, \mathrm{Sig}(\omega^{\mathrm{X}})_{s,t}\rangle d\xi_s^{\mathrm{X},j}| \le \epsilon \tag{3.37}$$

where π_M is the truncation at length M.

From theorem 3.4 we know that

$$\left\|\frac{1}{N}C_i^\top A_I^\top A_J C_j - \delta_{iI}^{jJ}\right\|_{L^2} = \mathcal{O}(\frac{1}{\sqrt{N}})2^{\frac{|I|+|J|}{2}}(|I|+|J|)!! \tag{3.38}$$

$$\left\|\frac{1}{N}C_i^\top A_I^\top A_J B_j\right\|_{L^2} = \mathcal{O}(\frac{1}{\sqrt{N}})2^{\frac{|I|+|J|}{2}}(|I|+|J|)!! \tag{3.39}$$

$$\left\|\frac{1}{N}B_i^\top A_I^\top A_J B_j - \delta_{iI}^{jJ}\right\|_{L^2} = \mathcal{O}(\frac{1}{\sqrt{N}})2^{\frac{|I|+|J|}{2}}(|I|+|J|)!! \tag{3.40}$$

Our strategy is that of using these bounds to create a vector $v \in \mathbb{R}^N$ "acting" like the $\pi_M\alpha_i$ and $\pi_M\beta_j$. Define, noting that the A_i, C_i, B_j depend on N, the vector

$$v^{\mathrm{N}} := \frac{1}{N}\left(\sum_{i=1}^{d_{in}}\sum_{|I|\le M}\alpha_i^I\ A_I C_i + \sum_{j=1}^{d_\omega}\sum_{|I|\le M}\beta_j^I\ A_I B_j\right) \tag{3.41}$$

Then expanding Z_t^{X}

$$\begin{aligned}
R_M :=& \left\|\sup_{(X,t)\in\mathbb{X}\times[0,1]}\left|\langle v^{\mathrm{N}}, Z_t^{\mathrm{X}}\rangle - \sum_{i=1}^{d_{in}} X_0^i\langle\pi_M\alpha_i, \mathrm{Sig}(\omega^{\mathrm{X}})_{0,t}\rangle - \sum_{j=1}^{d_\xi}\int_0^t\langle\pi_M\beta_j, \mathrm{Sig}(\omega^{\mathrm{X}})_{s,t}\rangle d\xi_s^{\mathrm{X},j}\right|\right\|_{L^2} \\
\le& \sum_{i=1}^{d_{in}}\sum_{|I|\le M}\left\|(v^{\mathrm{N}})^\top A_I C_i - \alpha_i^I\right\|_{L^2}\sup_{(X,t)\in\mathbb{X}\times[0,1]}|X_0^i\mathrm{Sig}(\omega^{\mathrm{X}})_{0,t}^I| \\
&+ \sum_{i=1}^{d_{in}}\sum_{|I|>M}\left\|(v^{\mathrm{N}})^\top A_I C_i\right\|_{L^2}\sup_{(X,t)\in\mathbb{X}\times[0,1]}|X_0^i\mathrm{Sig}(\omega^{\mathrm{X}})_{0,t}^I| \\
&+ \sum_{j=1}^{d_\omega}\sum_{|I|\le M}\left\|(v^{\mathrm{N}})^\top A_I B_j - \beta_j^I\right\|_{L^2}\sup_{(X,t)\in\mathbb{X}\times[0,1]}|\int_0^t Sig(\omega^{\mathrm{X}})_{s,t}^I d\xi_s^{\mathrm{X},j}| \\
&+ \sum_{j=1}^{d_\omega}\sum_{|I|>M}\left\|(v^{\mathrm{N}})^\top A_I B_j\right\|_{L^2}\sup_{(X,t)\in\mathbb{X}\times[0,1]}|\int_0^t Sig(\omega^{\mathrm{X}})_{s,t}^I d\xi_s^{\mathrm{X},j}|
\end{aligned}$$

Note how for $|I| \le M$ one has

$$\left\|(v^{\mathrm{N}})^\top A_I C_i - \alpha_i^I\right\|_{L^2} \le \mathcal{O}_M(\frac{1}{\sqrt{N}})$$

and that similarly for $|I| > M$

$$\left\|(v^{\mathrm{N}})^\top A_I C_i\right\|_{L^2} \le \mathcal{O}_M(\frac{1}{\sqrt{N}})2^{\frac{|I|}{2}}(M+|I|)!!$$

Which leads, thanks to the same bounds of [88][Appendix C], to

$$R_M = \frac{1}{\sqrt{N}}\mathcal{O}_{M,\mathbb{X}}(1) \tag{3.42}$$

But then by Markov's inequality it holds that

$$\mathbb{P}\left[\sup_{(X,t)\in\mathbb{X}\times[0,1]}\left|\langle v^{\mathrm{N}}, Z_t^{\mathrm{X}}\rangle - \sum_{i=1}^{d_{in}} X_0^i\langle \pi_M\alpha_i, \mathrm{Sig}(\omega^{\mathrm{X}})_{0,t}\rangle - \sum_{j=1}^{d_\xi}\int_0^t \langle \pi_M\beta_j, \mathrm{Sig}(\omega^{\mathrm{X}})_{s,t}\rangle d\xi_s^{\mathrm{X},j}\right| \le \epsilon\right] \to 1 \tag{3.43}$$

and thus there must be a choice of $N, \{A_i\}, B, C$ such that the inequality holds, and we thus obtain using (3.37)

$$\sup_{(X,t)\in\mathbb{X}\times[0,1]} |F(X,t) - \langle v^{\mathrm{N}}, Z_t^{\mathrm{X}}\rangle| \le 2\epsilon$$

and we conclude by arbitrariness of ϵ.

□

3.4.2 Characterization of the closure: Diagonal Case

In practice using dense transition matrices A_i is computationally infeasible, thus all main SSMs models employ some type of diagonal transition. Unfortunately this choice of diagonal weights[8] considerably restricts the family of learnable functionals. Intuitively the diagonal choice corresponds to running N independent 1-dimensional systems, this absence of mixing between the different hidden dimensions is the main culprit for the loss of expressivity. The full power can however be recovered by *chaining* the diagonal schemes (*cf.* theorem 3.18).

What we'll discover is that the Z_t^{X} cannot differentiate between $\mathrm{Sig}(\omega^{\mathrm{X}})_{s,t}^I$ and other $\mathrm{Sig}(\omega^{\mathrm{X}})_{s,t}^{\sigma(I)}$ for any permutation σ of the letters in the word I.

Diagonal Expansion for $Z_t^{\mathbf{x}}$

Proposition 3.15. *For any choice of $V \in \mathbb{R}^{N\times d_\omega}, B \in \mathbb{R}^{N\times d_\xi}$ and $C \in \mathbb{R}^{N\times d_{in}}$, writing $A_i := \mathrm{diag}(V_i)$, the unique solution to*

$$\begin{aligned} dZ_t^X &= \sum_{i=1}^{d_\omega} A_i Z_t^X d\omega_t^{X,i} + B d\xi_t^X \\ Z_0^X &= CX_0 \in \mathbb{R}^N \end{aligned} \tag{3.44}$$

is given by

$$Z_t^X = e^{\mathrm{diag}(V\omega_t^X)} CX_0 + \int_0^t e^{\mathrm{diag}(V(\omega_t^X - \omega_s^X))} B d\xi_s^X \tag{3.45}$$

[8] It is equivalent to ask for them to be commuting.

which can be expanded as

$$Z_t^X = \sum_{i=1}^{d_{in}} \sum_{I \in \mathbb{W}_{d_\omega}} A_I^{sym} C_i \; X_0^i Sig(\omega^X)_{0,t}^{sym,I} + \sum_{j=1}^{d_\xi} \sum_{I \in \mathbb{W}_{d_\omega}} A_I^{sym} B_j \int_0^t Sig(\omega^X)_{s,t}^{sym,I} d\xi_s^{X,j} \in \mathbb{R}^N \tag{3.46}$$

where

$$A_I^{sym} := \frac{1}{|I|!} \sum_{\sigma \in S_k} A_{\sigma(I)} = A_I \quad Sig(\omega^X)_{s,t}^{sym,I} := \frac{1}{|I|!} \sum_{\sigma \in S_k} Sig(\omega^X)_{s,t}^{\sigma(I)}. \tag{3.47}$$

Proof. By theorem C.13 and theorem C.18 we know that the solution of

$$Z_t^{\mathrm{X}} = Z_0^{\mathrm{X}} + \sum_{i=1}^{d_\omega} \int_0^t A_i Z_t^{\mathrm{X}} d\omega_t^{\mathrm{X},i} + \int_0^t B d\xi_t^{\mathrm{X}} \tag{3.48}$$

is explicitly given by

$$Z_t^{\mathrm{X}} = W_{0,t}^{\mathrm{X}} Z_0^{\mathrm{X}} + \int_0^t W_{s,t}^{\mathrm{X}} B d\xi_s^{\mathrm{X}} \tag{3.49}$$

where $W_{s,t}$ is the unique solution to

$$W_{s,t}^{\mathrm{X}} = Id + \sum_{i=1}^{d_\omega} \int_s^t A_i W_{s,r}^{\mathrm{X}} d\omega_r^{\mathrm{X},i} \tag{3.50}$$

In case the A_is are commuting matrices one can explicitly write the solution as

$$W_{s,t}^{\mathrm{X}} = \exp\left(\sum_{i=1}^{d_\omega} \int_s^t A_i d\omega_r^{\mathrm{X},i} \right) = \exp\left(\mathrm{diag}(V(\omega_t^{\mathrm{X}} - \omega_s^{\mathrm{X}})) \right) \tag{3.51}$$

since for fixed s one has, using commutativity, that

$$dW_{s,t}^{\mathrm{X}} = W_{s,t}^{\mathrm{X}} \left(\sum_{i=1}^{d_\omega} A_i d\omega_t^{\mathrm{X},i} \right) = \sum_{i=1}^{d_\omega} A_i W_{s,t}^{\mathrm{X}} d\omega_t^{\mathrm{X},i}$$

On the other hand we know, theorem C.14, that

$$W_{s,t}^{\mathrm{X}} = \sum_{I \in \mathbb{W}_{d_\omega}} A_I \mathrm{Sig}(\omega^{\mathrm{X}})_{s,t}^I = \sum_{k=0}^{\infty} \sum_{I \in \mathbb{W}_{d_\omega}^k} A_I \mathrm{Sig}(\omega^{\mathrm{X}})_{s,t}^I \tag{3.52}$$

The two views are reconciled by noticing that the symmetric group S_k acts on $\mathbb{W}_{d_\omega}^k$, the space of words of length k, by permuting the letters and, by commutativity,

$$\forall \sigma \in S_k . \forall I \in \mathbb{W}_{d_\omega}^k . \; A_I = A_{\sigma(I)}$$

Then we have

$$\sum_{I\in\mathbb{W}^k_{d_\omega}} A_I \mathrm{Sig}(\omega^{\mathrm{X}})^I_{s,t} = \sum_{I\in\mathbb{W}^k_{d_\omega}} \frac{1}{k!} \sum_{\sigma\in S_k} A_{\sigma(I)} \mathrm{Sig}(\omega^{\mathrm{X}})^{\sigma(I)}_{s,t} = \sum_{I\in\mathbb{W}^k_{d_\omega}} \frac{A_I}{k!} \sum_{\sigma\in S_k} \mathrm{Sig}(\omega^{\mathrm{X}})^{\sigma(I)}_{s,t}$$

recalling then how $e_{I_1} \sqcup\!\sqcup \cdots \sqcup\!\sqcup e_{I_k} = \sum_{\sigma\in S_k} e_{\sigma(I)}$ (see theorem C.3) we get to

$$\begin{aligned}
&\sum_{I\in\mathbb{W}^k_{d_\omega}} A_I \mathrm{Sig}(\omega^{\mathrm{X}})^I_{s,t} \\
&= \sum_{I\in\mathbb{W}^k_{d_\omega}} \frac{A_I}{k!} \sum_{\sigma\in S_k} \mathrm{Sig}(\omega^{\mathrm{X}})^{\sigma(I)}_{s,t} = \sum_{I\in\mathbb{W}^k_{d_\omega}} \frac{A_I}{k!} \prod_{i=1}^k \mathrm{Sig}(\omega^{\mathrm{X}})^{I_i}_{s,t} \\
&= \sum_{I\in\mathbb{W}^k_{d_\omega}} \frac{1}{k!} \prod_{i=1}^k A_{I_i} \mathrm{Sig}(\omega^{\mathrm{X}})^{I_i}_{s,t} = \frac{1}{k!} \left(\sum_{i=1}^{d_\omega} A_i \mathrm{Sig}(\omega^{\mathrm{X}})^i_{s,t} \right)^k = \frac{1}{k!} \left(\sum_{i=1}^{d_\omega} \int_s^t A_i d\omega_r^{\mathrm{X},i} \right)^k
\end{aligned}$$

In particular we see how in the commuting case

$$W^{\mathrm{X}}_{s,t} = \sum_{I\in\mathbb{W}_{d_\omega}} A^{sym}_I \mathrm{Sig}(\omega^{\mathrm{X}})^{sym,I}_{s,t} \tag{3.53}$$

where

$$A^{sym}_I := \frac{1}{|I|!} \sum_{\sigma\in S_k} A_{\sigma(I)} = A_I \quad \mathrm{Sig}(\omega^{\mathrm{X}})^{sym,I}_{s,t} := \frac{1}{|I|!} \sum_{\sigma\in S_k} \mathrm{Sig}(\omega^{\mathrm{X}})^{\sigma(I)}_{s,t}.$$

□

Diagonal Expressiveness

Theorem 3.16. *Fix a compact input set $\mathbb{X}$ and continuous gates $(\cdot)_0, \omega, \xi$. For any $\epsilon > 0$ and any*

$$F \in \left\{ (X,t) \mapsto \psi(\omega^X_t) \cdot X_0 + \int_0^t \phi(\omega^X_t - \omega^X_s) \cdot d\xi^X_s \right\} \tag{3.54}$$

for $\psi \in C^0(\mathbb{R}^{d_\omega}; \mathbb{R}^{d_{in}})$ and $\phi \in C^0(\mathbb{R}^{d_\omega}; \mathbb{R}^{d_\xi})$, there exist a choice of hidden dimension $N \geq 1$ and parameters $v \in \mathbb{R}^N, B \in \mathbb{R}^{N\times d_\xi}, C \in \mathbb{R}^{N\times d_{in}}$ and diagonal $A_i \in \mathbb{R}^{N\times N}$ such that

$$\sup_{(X,t)\in\mathbb{X}\times[0,1]} |F(X,t) - \langle v, Z^X_t \rangle| \leq \epsilon \tag{3.55}$$

Moreover the "reverse" also holds i.e. *given any choice of matrices A_i, B, C there is an ϵ-close map F in the family.*

Proof. This is just a repetition of the arguments used for the dense case with little more care to get the uniformity in time.

One defines the subset $Sym(\mathcal{H}^{(\cdot)_0,\omega,\eta}_{[0,1]}) \subset \mathcal{H}^{(\cdot)_0,\omega,\eta}_{[0,1]}$ of those F of type (3.24) defined by $\alpha_i, \beta_j \in l^2(\mathbb{W}_{d_\omega})$ such that for any word I and any permutation $\sigma(I)$ of it

$$\alpha_i^I = \alpha_i^{\sigma(I)} \quad \beta_j^I = \beta_j^{\sigma(I)} \tag{3.56}$$

The same argument of theorem 3.12 shows that the uniform closure of the space of linear maps on the Z_t^{X} is contained in the uniform closure of $Sym(\mathcal{H}^{(\cdot)_0,\omega,\eta}_{[0,1]})$, and the same bounds show that this latter closure is the same as that of its subset composed of those $F \in Sym(\mathcal{H}^{(\cdot)_0,\omega,\eta}_{[0,1]})$ having entries eventually equal to 0.

Since

$$\mathrm{Sig}(\omega^{\mathrm{X}})^{sym,I}_{s,t} := \frac{1}{|I|!}\sum_{\sigma\in S_k}\mathrm{Sig}(\omega^{\mathrm{X}})^{\sigma(I)}_{s,t} = \frac{1}{|I|!}\prod_{i=1}^{|I|}(\omega_t^{\mathrm{X},I_i} - \omega_s^{\mathrm{X},I_i}),$$

such maps can be expressed exactly in the form

$$P(\omega_t^{\mathrm{X}})\cdot X_0 + \int_0^t Q(\omega_t^{\mathrm{X}} - \omega_s^{\mathrm{X}})\cdot d\xi_s^{\mathrm{X}}$$

for polynomial maps P, Q fixed in time. The usual compactness and continuity argument, together with an application of Stone-Weiestrass, thus proves that the uniform closure of $Sym(\mathcal{H}^{(\cdot)_0,\omega,\eta}_{[0,1]})$ has the form needed.

The final ingredient is the density of the space of linear maps on the Z_t^{X} in $Sym(\mathcal{H}^{(\cdot)_0,\omega,\eta}_{[0,1]})$; this is another consequence of Stone-Weiestrass as seen from theorem 3.17. □

Remark. *Notice how here there is no need to augment the paths in creative ways in order to ensure separability of the points. The map $(\omega, s, t) \mapsto \omega_{[s,t]} \in C_{1,0}([0,1];\mathbb{R}^{d_\omega})$ is replaced by $(\omega, s, t) \mapsto \omega_t - \omega_s \in \mathbb{R}^{d_\omega}$ and the space of polynomials always separates points in $\mathbb{R}^{d_\omega}$.*

Remark. *It is not necessary to pass through $Sym(\mathcal{H}^{(\cdot)_0,\omega,\eta}_{[0,1]})$ to prove the previous result, since it directly follows from theorem 3.17. This choice of presentation has been motivated by the conviction of the usefulness of drawing parallels and comparisons.*

Proposition 3.17. *Fix a compact set $\mathbb{K} \subset \mathbb{R}^d$ and a d-dimensional convex cone C containing the origin. The space*

$$\mathcal{E} := Span\left(\mathbb{K} \ni x \mapsto e^{\langle \alpha, x\rangle_{\mathbb{R}^d}} \in \mathbb{R} : \alpha \in C\right)$$

is uniformly dense in $C^0(\mathbb{K};\mathbb{R})$.

Proof. This is an application of Stone-Weiestrass: $\mathcal{E}$ is a sub-algebra since

$$e^{\langle \alpha, x\rangle} e^{\langle \beta, x\rangle_{\mathbb{R}^d}} = e^{\langle \alpha+\beta, x\rangle_{\mathbb{R}^d}}$$

and $\alpha, \beta \in C \Rightarrow \alpha + \beta \in C$ by convexity of the cone; $\mathcal{E}$ contains the constant function $e^{\langle 0,x \rangle} = 1$ and is clearly point separating since the cone, being d-dimensional, it contains a basis of the whole space. □

Remark. *The usefulness of stating the previous result in such a general setting is the following: with this formalism we can, for example, restrict to $\alpha \leq 0$, in this way we would have a method to control the stability of the Linear CDEs by choosing the gate with a.s. $\dot{\omega}^X \geq 0$.*

Corollary (Mamba Case)**.** *In the Mamba setting, the closure reduces to*

$$\left\{ (X,t) \mapsto \sum_{i=1}^{d_\omega} \psi_i(\omega_t^{X,i}) + \sum_{i=1}^{d_\omega} \int_0^t \phi_i(\omega_t^{X,i} - \omega_s^{X,i}) \; d\xi_s^{X,i} \right\} \tag{3.57}$$

for continuous $\psi_i : \mathbb{R} \to \mathbb{R}$ and $\phi_i : \mathbb{R} \to \mathbb{R}$.

Proof. In this setting one runs in parallel d_ω diagonal systems and then takes a linear combination of the stacked hidden state. The maps in the closure of the whole system are then just the sums of maps in the closure of the subsystems. □

3.5 Chaining of Diagonal Systems

Fortunately it is possible to re-gain expressivity without sacrificing the computational advantages of diagonal schemes through *chaining*. This means driving a new Linear CDE by the solution of a previous Linear CDE, and repeating this procedure K times (*cf.* section 3.5).

Proposition 3.18. *Assume a compact $\mathbb{X} \subset C_{1,0}([0,1];\mathbb{R}^d)$. For any functional $F : \mathbb{X} \times [0,1] \to \mathbb{R}$ of the form (3.16) and $\epsilon > 0$ there exists a sequence of linear maps $W_k \in \mathbb{R}^{M_k \times N_k}$,* diagonal *weights $A_i^{(k)}$ and $B^{(k)}$ for the following family of* chained *diagonal Linear CDEs*

$$Z_t^{0,X} \equiv 0, \quad dZ_t^{k+1,X} = \sum_{i=1}^{d+M_k} A_i^{(k+1)} Z_t^{k+1,X} d \begin{bmatrix} W_k Z^{k,X} \\ X \end{bmatrix}_t^i + B^{(k+1)} dX_t \in \mathbb{R}^{N_{k+1}}, \tag{3.58}$$

such that as $k \to \infty$, there exists $C_k \in \mathbb{R}^{1 \times N_k}$ with $\sup_{(X,t) \in \mathbb{X} \times [0,1]} |F(X,t) - C_k Z_t^{k,X}| \leq \epsilon$.

Intuitively, with chaining one recovers the mixing between the input dimensions which was so important for the expressiveness of *dense* Linear CDEs. This does not happen immediately but crucially depends on the length of the chain, the result in fact tells us that the recovery holds for *long enough* chains (and big enough hidden states).

The diagonal case differs from the general one not only in the fact that the class of approximable functions is much weaker but also in the necessity for the presence of ξ^{X} in order to obtain any path-dependence. The term

$$\int_0^t \phi(\omega_t^{\mathrm{X}} - \omega_s^{\mathrm{X}}) \cdot d\xi_s^{\mathrm{X}}$$

becomes then a crucial component. At first sight one might think that such a term allows to recover at least level two components of the Signature of $(\omega^{\mathrm{X}}, \xi^{\mathrm{X}})$, unfortunately things are not as easy as they may seem. Notice how inside of the integral time is "going backwards" from the perspective of ω^{X}, thus we can in general approximate terms of type

$$\int_0^t \int_s^t d\omega_r^{\mathrm{X},i} d\xi_s^{\mathrm{X},j} = \int_{1-t}^1 \int_{1-t}^r d\overleftarrow{\omega}_r^{\mathrm{X},i} d\overleftarrow{\xi}_s^{\mathrm{X},j} = \mathrm{Sig}((\overleftarrow{\omega}^{\mathrm{X}}, \overleftarrow{\xi}^{\mathrm{X}}))_{1-t,1}^{i\omega j_\xi}$$

which are indeed terms of the Signature, but of the reverse paths $\overleftarrow{\omega}_r^{\mathrm{X}} = \omega_{1-r}^{\mathrm{X}}$ and $\overleftarrow{\xi}_s = \xi_{1-s}^{\mathrm{X}}$.

Proposition 3.19. *Fix a compact input set $\mathbb{X}$, continuous gates $(\cdot)_0, \omega, \xi$ and $X_0^1 = 1$. If the components of ξ^X are linear combinations of those of ω^X, with time-independent weights, then linear functionals on Z_t^X can, uniformly in $\mathbb{X} \times [0,1]$, approximate arbitrarily well the following level 2 terms of $Sig((\omega^X, \xi^X))_{0,t}$:*

$$\int_0^t \int_0^s d\omega_r^{X,i} d\xi_s^{X,j} = Sig((\omega^X, \xi^X))_{0,t}^{i\omega j_\xi}$$

Proof. Under these hypotheses we know that linear functionals on Z_t^{X} are uniformly dense, for continuous ψ, ϕ, in

$$\left\{ (X,t) \mapsto \psi(\omega_t^{\mathrm{X}}) \cdot X_0 + \int_0^t \phi(\omega_t^{\mathrm{X}} - \omega_s^{\mathrm{X}}) \cdot d\xi_s^{\mathrm{X}} \right\}.$$

Assume $\xi_s^{\mathrm{X},j} = \langle \alpha_j, \omega_t^{\mathrm{X}} \rangle$ and consider the choices

$$\psi(x) = (x^i \langle \alpha_j, x \rangle, 0, \ldots, 0)^\top, \quad \phi(x) = -(0, \ldots, 0, x^i, 0, \ldots, 0). \tag{3.59}$$

so that

$$\psi(\omega_t^{\mathrm{X}}) \cdot X_0 = \omega_t^{\mathrm{X},i} \xi_t^{\mathrm{X},j} \quad \phi(\omega_t^{\mathrm{X}} - \omega_s^{\mathrm{X}}) \cdot d\xi_s^{\mathrm{X}} = -(\omega_t^{\mathrm{X},i} - \omega_s^{\mathrm{X},i}) d\xi_s^{\mathrm{X},j}. \tag{3.60}$$

To conclude note that

$$\begin{aligned} \omega_t^{\mathrm{X},i} \xi_t^{\mathrm{X},j} &= \int_{s=0}^t \int_{r=0}^t d\omega_r^{\mathrm{X},i} d\xi_s^{\mathrm{X},j} = \int_{s=0}^t \int_{r=0}^s d\omega_r^{\mathrm{X},i} d\xi_s^{\mathrm{X},j} + \int_{s=0}^t \int_{r=s}^t d\omega_r^{\mathrm{X},i} d\xi_s^{\mathrm{X},j} \\ &= \int_0^t \int_0^s d\omega_r^{\mathrm{X},i} d\xi_s^{\mathrm{X},j} + \int_{s=0}^t (\omega_t^{\mathrm{X},i} - \omega_s^{\mathrm{X},i}) d\xi_s^{\mathrm{X},j} \end{aligned}$$

hence

$$\int_0^t \int_0^s d\omega_r^{\mathrm{x},i} d\xi_s^{\mathrm{x},j} = \omega_t^{\mathrm{x},i}\xi_t^{\mathrm{x},j} - \int_{s=0}^t (\omega_t^{\mathrm{x},i} - \omega_s^{\mathrm{x},i}) d\xi_s^{\mathrm{x},j} = \psi(\omega_t^{\mathrm{x}}) \cdot X_0 + \int_0^t \phi(\omega_t^{\mathrm{x}} - \omega_s^{\mathrm{x}}) \cdot d\xi_s^{\mathrm{x}}$$

□

If $X \in C_{1,0}([0,1];\mathbb{R}^d)$ we can use the previous result to compute its Signature entries by chaining *diagonal* Linear CDEs.

Theorem 3.20. *Assume a compact input set $\mathbb{X} \subset C_{1,0}([0,1];\mathbb{R}^d)$. For any $I \in \mathbb{W}_d$ with $|I| \geq 2$ and $\epsilon > 0$ there is a sequence of linear maps $W_k \in \mathbb{R}^{N_k \times 1}$ and weights for the following family of* chained *Linear CDEs*

$$dZ_t^{1,X} = \sum_{i=1}^{d} A_i^{(1)} Z_t^{1,X} dX_t^i + B^{(1)} dX_t \in \mathbb{R}^{N_1}, \quad Z_0^{1,X} = Z_0^1, \tag{3.61}$$

$$dZ_t^{k+1,X} = \sum_{i=1}^{d+1} A_i^{(k+1)} Z_t^{k+1,X} d \begin{bmatrix} W_k Z^{k,X} \\ X \end{bmatrix}_t^i + B^{(k+1)} dX_t \in \mathbb{R}^{N_{k+1}}, \quad Z_0^{k+1,X} = Z_0^{k+1}, \tag{3.62}$$

such that for some $v \in \mathbb{R}^{N_{|I|-1}}$ one has

$$\sup_{(X,t)\in\mathbb{X}\times[0,1]} |Sig(X)_{0,t}^I - \langle v, Z_t^{|I|-1,X}\rangle| \leq \epsilon \tag{3.63}$$

Proof. For $|I| = 2$ we can apply theorem 3.19. Assume the theorem holds for $|I| \leq k$ and let $M := \sup_{X\in\mathbb{X}} \|X\|_{1-var}$. Fix $|Ij| = k+1$ and $W_{k-1} \in \mathbb{R}^{N_{k-1}\times 1}$ such that

$$\sup_{(X,t)\in\mathbb{X}\times[0,1]} |Sig(X)_{0,t}^I - W_{k-1} Z_t^{k-1,\mathrm{x}}| \leq \frac{\epsilon}{M}.$$

Again by theorem 3.19 there are a N_k and $v \in \mathbb{R}^{N_k}$ such that

$$\sup_{(X,t)\in\mathbb{X}\times[0,1]} |\int_0^t W_{k-1} Z_s^{k-1,\mathrm{x}} dX_s^j - \langle v, Z_t^{k,\mathrm{x}}\rangle| \leq \epsilon.$$

Then

$$\begin{aligned}
|\mathrm{Sig}(X)_{0,t}^{Ij} - \langle v, Z_t^{k,\mathrm{x}}\rangle| \leq & |\int_0^t \mathrm{Sig}(X)_{0,s}^I dX_s^j - \int_0^t W_{k-1} Z_s^{k-1,\mathrm{x}} dX_s^j| \\
& + |\int_0^t W_{k-1} Z_s^{k-1,\mathrm{x}} dX_s^j - \langle v, Z_t^{k,\mathrm{x}}\rangle| \\
\leq & \int_0^t |\mathrm{Sig}(X)_{0,s}^I - W_{k-1} Z_s^{k-1,\mathrm{x}}||dX_s^j| + \epsilon \\
\leq & \int_0^t \frac{\epsilon}{M} |dX_s^j| + \epsilon \leq 2\epsilon
\end{aligned}$$

thus concluding the proof. □

Then theorem 3.18 follows as a corollary by running in parallel the systems above to recover simultaneously multiple Signature entries.

3.6 Extensions to Structured Matrices

Here we extend some of the previous results to the case of more general assumptions on the matrices A_i, B, C. We will consider the simpler case

$$dZ_t = \sum_{i=1}^{d_\omega} A_i Z_t d\omega_t^i, \quad Z_0 = Z_0 \in \mathbb{R}^N$$

but the similar results hold for the more general setting. We write $Z((A_i)_i, Z_0, \omega)_t$ for the solution of the above equation.

Recall how theorem 3.2 shows that any linear readout of Z_t^{x} is expressed as a series in signature terms, independently on the nature of the matrices A_i, B, C as

$$Z_t^{\mathrm{x}} = \sum_{I \in \mathbb{W}_{d_\omega}} (A_I Z_0)\ \mathrm{Sig}(\omega^{\mathrm{x}})_{0,t}^{(I)}$$

Consequently, such systems are inherently limited to learning functions that are close to these (uniformly convergent) series. *Maximal expressivity* is then achieved when *any* finite linear combination in signature terms can be approximated by a linear readout on Z_t^{x} through appropriate choices of the matrices A_i, B, C.

Definition 3.21. *Fix a set of paths $\mathcal{X} \subseteq C^{1-var}([0,1];\mathbb{R}^d)$. We say that a sequence $(\mathcal{A}_N, \mathcal{H}_N)_{N\in\mathbb{N}}$, where $\mathcal{H}_N \subseteq \mathbb{R}^N$ and $\mathcal{A}_N \subseteq \mathbb{R}^{N\times N}$, achieves* maximal expressivity *for $\mathcal{X}$ whenever for any positive tolerance $\epsilon > 0$ and any finite linear combination coefficients $\alpha \in T(\mathbb{R}^d)$, there exist a choice of parameters $v, (A_i), Z_0$ in $\mathbb{R}^N, \mathcal{A}_N, \mathcal{H}_N$ in the sequence, such that $v^\top Z((A_i), Z_0, \omega)$. is uniformly close to $\langle \alpha, S(\omega)_{[0,\cdot]}\rangle$ up to an error of ϵ,* i.e.

$$\forall \epsilon > 0,\ \forall \alpha \in T(\mathbb{R}^d),\ \exists N \geq 0,\ \exists (v, (A_i), Z_0) \in \mathbb{R}^N \times \mathcal{A}_N^d \times \mathcal{H}_N \ s.t.$$
$$\sup_{(\omega,t)\in\mathcal{X}\times[0,1]} |\langle \alpha, S(\omega)_{[0,t]}\rangle - v^\top Z((A_i), Z_0, \omega)_t| < \epsilon.$$

If we are given a sequence of probabilities $\mathbb{P}_N$ on $\mathcal{A}_N^d \times \mathcal{H}_N$ such that $\forall \epsilon > 0,\ \forall \alpha \in T(\mathbb{R}^d)$, it holds that

$$\lim_{N\to\infty} \mathbb{P}_N \left\{ \exists v \in \mathbb{R}^N \ s.t. \sup_{(\omega,t)\in\mathcal{X}\times[0,1]} |\langle \alpha, S(\omega)_{[0,t]}\rangle - v^\top Z((A_i), Z_0, \omega)_t| < \epsilon \right\} = 1, \quad (3.64)$$

then we say that $(\mathcal{A}_N, \mathcal{H}_N, \mathbb{P}_N)_{N\in\mathbb{N}}$ achieves maximal probabilistic expressivity *for $\mathcal{X}$.*

The deterministic argument by Kidger [65] reproposed above in proving theorem 3.14 demonstrates the existence of a specific choice of $\mathcal{A}_N, \mathcal{H}_N$ mimicking the algebraic structure of tensors and provides *maximal expressivity* for compact sets of paths. Furthermore, in the same proof we established that matrices (almost) replicating the algebraic structure of tensors are, in fact, abundant, showing that the triplet $(\mathbb{R}^{N\times N}, \mathbb{R}^N, \mathbb{P}_N)$, where $\mathbb{P}_N$ is a Gaussian measure achieves *maximal probabilistic expressivity* for compact sets.

This result implies that for dense matrices A_i, if the hidden dimension N is sufficiently large, there is a significant abundance of parameters capable of achieving uniformly and arbitrarily low error rates. These parameters should therefore be readily discoverable through classical optimisation methods. Unfortunately using dense matrices is infeasible in practice due to computational constraints. In this section, we present three alternative choices of parameters that lead to *maximal probabilistic expressivity* for compact sets. These alternatives offer better computational properties compared to the naive use of dense matrices, their empirical performance is studied in Walker et al. [125].

3.6.1 Sparse Matrices

Proposition 3.22. *The sequence of triplets* $(\mathbb{R}^{N\times N}, \mathbb{R}^N, \mathbb{P}_N)$ *where* $\mathbb{P}_N$ *is such that*

- *the initial value has independent standard Gaussian entries* $[Z_0]_\alpha \overset{iid}{\sim} \mathcal{N}(0,1)$,
- *the weight matrices are distributed as* $A_i \overset{iid}{\sim} \frac{1}{\sqrt{Np_N}} W \odot B$ *with* W *and* B *independent matrices having entries* $[W]_{\alpha,\beta} \overset{iid}{\sim} \mathcal{N}(0,1)$ *and* $[B]_{\alpha,\beta} \overset{iid}{\sim} Ber(p_N)$,
- *the sparsity parameter* p_N *satisfies* $Np_N \to \infty$ *as* $N \to \infty$,

achieves maximal probabilistic expressivity *for compact sets.*

Proof. Following the proof of theorem 3.14, we only need to prove a bound of type

$$\left\| \frac{1}{N} \langle A_I Z_0, A_J Z_0 \rangle_{\mathbb{R}^N} - \delta_{I,J} \right\|_{L^2(\mathbb{P}_N)} \leq (\kappa(|I| + |J|))!! \, \mathcal{O}(\frac{1}{\sqrt{N}}) \tag{3.65}$$

as in the dense Gaussian case. That such sparse matrices present the same bounds as dense Gaussian ones follows from section A.2, where it is shown that the bounds can only differ by a correction term of order $\mathcal{O}_{I,J}(\frac{1}{\sqrt{N}})$ where the constants are bounded by the number of pairings of $I \cup J \cup I \cup J$. □

3.6.2 Walsh-Hadamard Matrices

Proposition 3.23. *The sequence of triplets* $(\mathcal{A}_N, \mathbb{R}^N, \mathbb{P}_N)$ *where* $\mathcal{A}_N$ *and* $\mathbb{P}_N$ *are such that*

- $\mathcal{A}_N := \{W \, diag(\Delta) : \Delta \in \mathbb{R}^N, \; W \in \mathbb{R}^{N\times N}, \; WW^\top = I_N\}$,

- *the initial value has independent standard Gaussian entries* $[Z_0]_\alpha \overset{iid}{\sim} \mathcal{N}(0,1)$,
- *the weight matrices are distributed as* $A_i \overset{iid}{\sim} \frac{1}{\sqrt{N}} H\,diag(\Delta)$ *for a fixed* $H \in \mathbb{R}^{N\times N}$ *satisfying* $HH^\top = NI_N$ *and having entries bounded uniformly in* N *by a constant* C, *and* $\Delta \in \mathbb{R}^N$ *having entries* $[\Delta]_\alpha \overset{iid}{\sim} \mathcal{N}(0,1)$,

achieves maximal probabilistic expressivity *for compact sets. In particular one can choose* H *to be a Walsh-Hadamard matrix of order* N *for computational efficiency.*

Proof. Following the proof of theorem 3.14, we only need to prove a bound of type

$$\left\| \frac{1}{N}\langle A_I Z_0, A_J Z_0\rangle_{\mathbb{R}^N} - \delta_{I,J} \right\|_{L^2(\mathbb{P}_N)} \leq (\kappa(|I|+|J|))!!\ \mathcal{O}(\frac{1}{\sqrt{N}}). \tag{3.66}$$

Notice that for $x \in \mathbb{R}^N$ one has the equivalence $W\,\mathrm{diag}(\Delta)\cdot x = W\cdot(\Delta \odot x)$ which can be represented graphically as in fig. 3.2.

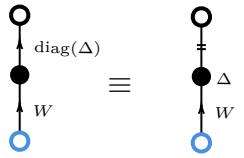


Figure 3.2: *Graphical representations of the matrix* $W\,diag(\Delta)$. *Here* $\ddagger$ *edges correspond to identity matrices.*

This leads to the *product graph* representation $G_{I,J}$ for $\frac{1}{N}\langle A_I Z_0, A_J Z_0\rangle_{\mathbb{R}^N}$, where $I = i_1 \dots i_n$ and $J = j_1 \dots j_m$, given in fig. 3.3, which allow us to use the genus expansion of chapter 1 to obtain the bounds.

$$\frac{1}{N}\langle A_I h_0,\ A_J h_0\ \rangle \quad \equiv \quad \frac{1}{N}\frac{1}{N^{\frac{n+m}{2}}}$$

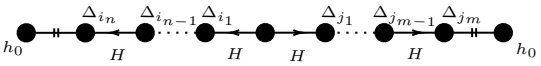


Figure 3.3: *The product* $\frac{1}{N}\langle A_I Z_0, A_J Z_0\rangle_{\mathbb{R}^N}$ *as a* product graph G.

In the present setting, the vertices labelled Z_0 must be identified, and the remaining decorated vertices must be paired such that Δ_a is identified with Δ_b only if $a = b$. Each of these *admissible pairings* ϕ produces a graph $(G_{I,J})_\phi$ in which each vertex is assigned the vector of ones. Note that since $|[H]_{\alpha,\beta}| \leq C$, one has $|[H^{\odot k}]_{\alpha,\beta}| \leq C^k$. The procedure is shown in fig. 3.4 for all pairings when $I = J = 11$.

Under these boundedness assumptions, note how leading-order pairings of $G_{I,J}$ are the ones having the maximum number of surviving vertices, meaning having $\frac{|I|+|J|}{2}+1$ vertices. This can happen iff $I = J$, and even in this case, there exists only one pairing (the middle one in fig. 3.4) for which it holds that $(G_{I,I})_\phi \equiv \frac{1}{N}\frac{1}{N^{\frac{2|I|}{2}}}N^{|I|+1} = 1$. To see this, note that only the "middle" vertex of $G_{I,J}$ is not paired, so to get $\frac{|I|+|J|}{2}+1$, all other vertices have to be identified in couples. This implies that the left and right adjacent vertices (call them v_l and v_r) must be paired together: in fact the sub-graph

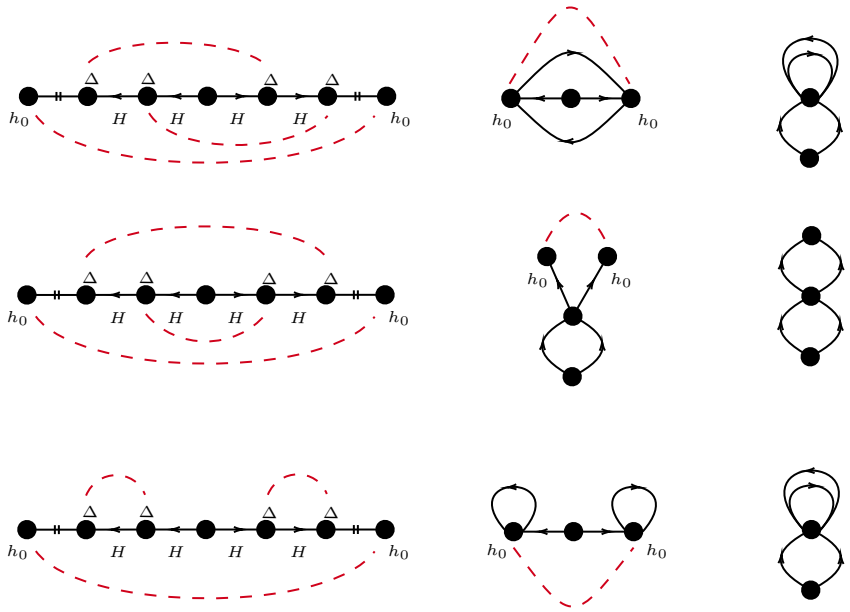


Figure 3.4: *Construction of $(G_{I,J})_\phi$. Here, we display all the pairings, represented by the red dashed lines, for $I = J = 11$ along with their intermediate stages. For simplicity, we omit the H labels from edges with arrows.*

comprising the middle vertex and the two adjacent edges corresponds to the matrix $HH^\top = NI_N$, hence we can remove this sub-graph, identify the adjacent vertices and take the factor N in front without changing the value of the graph; but then if v_l and v_r were paired with other vertices we would identify at least 4 vertices. Hence i_1 has to be equal to j_1 and v_l and v_r must be identified. Proceeding by induction, from the middle out, we see that only the identity pairing has non-vanishing value.

To conclude we are left to prove that any pairing ψ of $G_{I,J} \sqcup G_{I,J}$ not inducing the identity one on any of the two copies produces $|\,(G_{I,J} \sqcup G_{I,J})_\psi\,| \leq \frac{1}{N}$, since the number of these pairings is less than the total number of pairings of the Δ-labelled vertices of $G_{I,J} \sqcup G_{I,J}$, which is $(2(|I| + |J|))!!$. Then the inequality would hold with $\kappa = 2C^4$ since from the definition of *product-graph*, we see that the bound

$$|\,(G_{I,J} \sqcup G_{I,J})_\psi\,| \leq \frac{1}{N^2} \frac{1}{N^{2\frac{|I|+|J|}{2}}} N^{V_\psi} C^{2(|I|+|J|)} \tag{3.67}$$

must hold, where V_ψ is the number of vertices in $(G_{I,J} \sqcup G_{I,J})_\psi$. Thus it suffices to show $V_\psi \leq |I| + |J| + 1$.

To see this notice that in any case $V_\psi \leq |I| + |J| + 2$ and that $V_\psi = |I| + |J| + 2$ iff ψ identifies the random vertices in pairs. Once again there are special vertices which are not paired, the "middle" ones, the sub-graphs containing them correspond to the matrix $HH^\top = NI_N$ and can thus be removed by identifying the adjacent vertices and taking the N factor out, so these vertices must be paired between themselves, and so

on. This shows that ψ has to separately pair both copies of $G_{I,J}$ in $G_{I,J} \sqcup G_{I,J}$ with identity pairings, but then such a ψ would not be an *atom-free* pairing! Hence such a ψ cannot exist and $V_\psi \leq |I| + |J| + 1$ always holds. □

Remark. *Just as in the sparse case, and following, it is possible to prove that the Δ matrices can be taken to have i.i.d. entries drawn from a centered, symmetric, but possibly heavy-tailed distribution, provided that the even moments are finite. This distributional adjustment is a useful technique for controlling the eigenvalue distribution of $\frac{1}{\sqrt{N}} H\, diag(\Delta)$, ensuring it has favourable computational properties while providing a theoretical guarantee of preserving expressive power.*

3.6.3 Block Diagonal

Proposition 3.24. *The sequence of triplets $(\mathcal{A}_N, \mathbb{R}^N, \mathbb{P}_N)$ where $\mathcal{A}_N$ and $\mathbb{P}_N$ are such that*

- $\mathcal{A}_N := \{\text{BlockDiag}(B_1, B_2, \ldots, B_{k_N}) : b_N = \lceil \log(N) \rceil, k_N = \lceil N/b_N \rceil, B_h \in \mathbb{R}^{k_N b_N}\}$,
- *the initial value has independent standard Gaussian entries $[Z_0]_\alpha \overset{iid}{\sim} \mathcal{N}(0,1)$,*
- *the weight matrices are distributed as $[B_i]_{\alpha,\beta} \overset{iid}{\sim} \frac{1}{\sqrt{b_N}}\mathcal{N}(0,1)$.*

achieves maximal probabilistic expressivity *for compact sets.*

Proof. Once again, we only need to prove a bound of type

$$\left\| \frac{1}{N} \langle A_I Z_0, A_J Z_0 \rangle_{\mathbb{R}^N} - \delta_{I,J} \right\|_{L^2(\mathbb{P}_N)} \leq (\kappa(|I| + |J|))!!\ \mathcal{O}(1). \tag{3.68}$$

It suffices to notice that

$$\frac{1}{N} \langle A_I Z_0, A_J Z_0 \rangle_{\mathbb{R}^N} = \frac{1}{k_N} \sum_{l=1}^{k_N} \frac{1}{b_N} \langle B_l^I h_{0;l}, B_l^J h_{0;l} \rangle_{\mathbb{R}^{b_N}}$$

where $h_{0;l} := [Z_0]_{(l-1)b_N+1,\ldots,lb_N}$, since we then know that

$$\left\| \frac{1}{b_N} \langle B_l^I h_{0;l}, B_l^J h_{0;l} \rangle_{\mathbb{R}^{b_N}} - \delta_{I,J} \right\|_{L^2(\mathbb{P}_N)} \leq (\kappa(|I| + |J|))!!\ \mathcal{O}(\frac{1}{\sqrt{b_N}}). \tag{3.69}$$

From the sum and the initial factor $\frac{1}{k_N}$ we obtain

$$\left\| \frac{1}{N} \langle A_I Z_0, A_J Z_0 \rangle_{\mathbb{R}^N} - \delta_{I,J} \right\|_{L^2(\mathbb{P}_N)} \leq (\kappa(|I| + |J|))!!\ \mathcal{O}(\frac{1}{\sqrt{b_N}}). \tag{3.70}$$

□

3.7 Experiments

Datasets. The first task is based on a dataset from Walker et al. [124] where the aim is to predict terms in the anti-symmetric part of the input path's signature. The dataset's objective aligns with the proofs of theorem 3.7, which characterise the closure using the path's signature. We created two datasets with dimensions 2 and 3 respectively. The increment in each channel at each step is an integer-rounded sample from a standard Normal distribution. The 2D dataset's target is an area integral $\int_0^1 \int_0^v dX_u^1 dX_v^2$, and the 3D dataset's target is a volume integral $\int_0^1 \int_0^w \int_0^v dX_u^1 dX_v^2 dX_w^3$.

The second task is the A_5 benchmark from Merrill et al. [84]. It tests models on state-tracking, a crucial ability for tasks involving permutation composition, such as chess. The dataset comprises sequences from the group of even permutations on five elements, A_5, where the target is the cumulative composition of all preceding permutations. Datasets vary by sequence length, ranging from 3 to 20.

Models. On the anti-symmetric signature task, we considered seven models: (i-ii) S4 or Mamba recurrence with linear readout, (iii-iv) two stacked S4 or Mamba recurrences with a linear mixing layer in-between and a linear readout, (v-vi) two stacked S4 or Mamba recurrences with a linear mixing layer + GeLU in-between, and a linear readout, (vii) a linear CDE with gates $\omega_t^{\mathrm{X}} = \xi_t^{\mathrm{X}} = (t, X_t)$ and a linear readout. All state space models have **trainable** matrices in their recurrences, whereas the linear CDE is using **fixed** matrices. All models use a hidden dimension of 256, with the state space models using a state dimension of 256. The state space models are trained using gradient descent with a batch size of 32 and Adam with a learning rate of 10^{-4}. The output from the linear CDE's recurrence is obtained using the Tsit5 adaptive ODE solver, with an absolute and relative tolerance of 10^{-2}. The linear CDEs linear readout is optimised via ordinary least squares.

On the A_5 benchmark, we consider five models: a linear CDE, a RNN, a transformer, and S4 and Mamba recurrences. All models use an embedding layer followed by a series of blocks that combine the sequence-to-sequence model, linear mixing with a non-linear activation function, layer normalization, and a residual connection. Furthermore, all models have **trainable** matrices in their recurrences and are trained using batch gradient descent with a batch size of 32 and AdamW with a weight decay of 0.01. The RNN, transformer, S4, and Mamba use a hidden dimension of 1024, with the state space models using a state dimension of 64 and the transformer using 64 heads. Due to memory constraints, the linear CDE uses a hidden dimension of 256. Note that the IDS4 recurrence introduced by Merrill et al. [84] corresponds directly to a linear CDE with dense transition matrices, hence we did not include the model as a baseline.

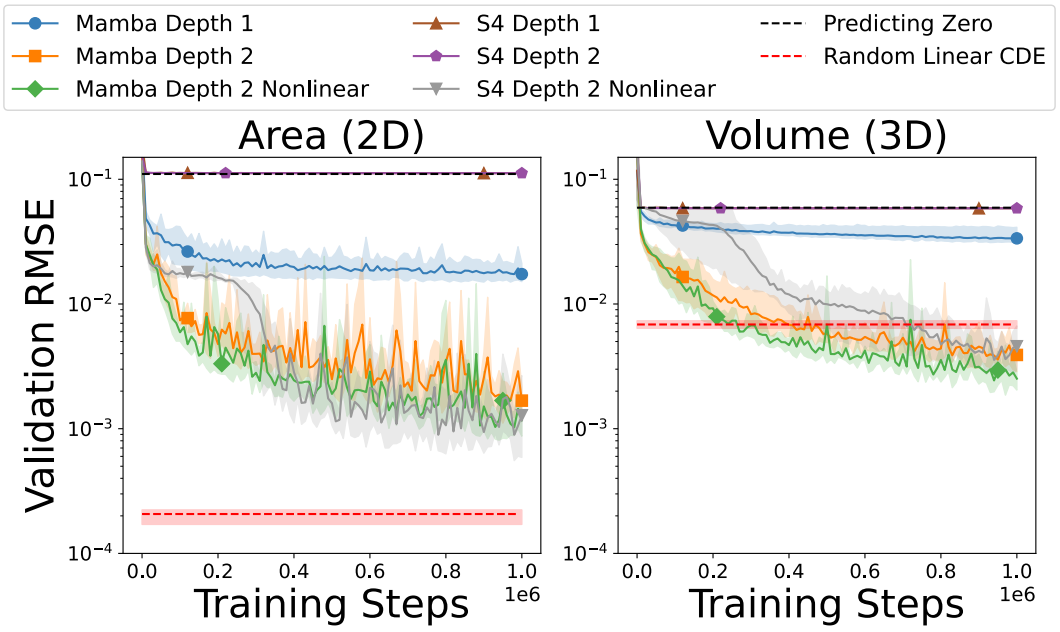


Figure 3.5: *Comparison of the Linear CDE, Mamba, and S5 on the anti-symmetric signature prediction tasks. For each model, we plotted the mean and range of the validation accuracy over 5 independent runs.*

Results. Results on the anti-symmetric signature prediction task (fig. 3.5) empirically demonstrate a number of the theoretical results presented in this paper. Firstly, as discussed in section 3.4.1, recurrences which are linear in the input, such as S4, require a non-linearity in-between the layers to perform well. Furthermore, as stated in theorem 3.16 and theorem 3.18, even if the recurrence is non-linear in the input, such as Mamba, the expressivity of models with diagonal matrices is improved by stacking. Additionally, the inclusion of the non-linearity in-between the Mamba layers does not improve performance, as the recurrences themselves are expressive enough. Finally, as stated in theorem 3.7, dense matrices can achieve strong expressivity with random initialization, no stacking, and only a trainable linear readout.

In fig. 3.6 we plot the results on the A_5 benchmark. The figure shows that the number of blocks S4, Mamba, and the transformer require to achieve greater than 90% validation accuracy grows with the sequence length. In fact, given our model architecture, none of these models could achieve greater than 90% validation accuracy for sequences of length twenty[9]. On the other hand, the RNN and Linear CDE are able to achieve greater than 90% validation accuracy for all lengths considered using only one block. These empirical results further validate theorem 3.16 and theorem 3.18:

[9]Merrill et al. [84] show that with tailored architectures, S4, Mamba, and a Transformer can achieve over 90% validation accuracy on length-20 sequences using 4 blocks. However, in our experiments we wished to keep the architecture consistent between models, only changing the sequence-to-sequence model.

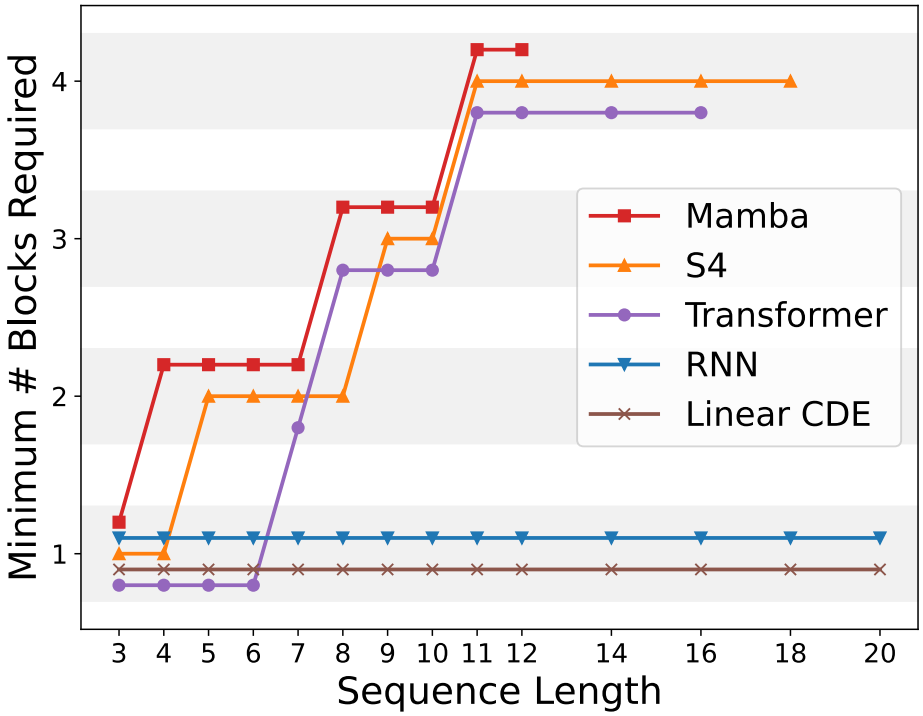


Figure 3.6: *For each sequence length, the plot shows the minimum number of blocks required to achieve at least* 90% *validation accuracy, with each grey band corresponding to a number of blocks. Missing points mean the model did not achieve at least* 90% *validation accuracy with* 4 *blocks or less.*

there exists a gap in expressivity between diagonal and dense transition matrices, and stacking is required to recover the expressivity. Furthermore, they provide empirical evidence that even for simple state-tracking problems, the number of blocks required can grow quickly with sequence length.

Chapter 4

Neural Signature Kernels

In chapter 3 we introduced Linear Controlled Differential Equations (Linear CDEs) and studied their expressive power, showing how they naturally recover signature kernels as a special case of the associated kernels (*c.f.* theorem 3.5 and following remark).

Building on this foundation, and following Muca Cirone et al. [88], this chapter considers Neural CDEs [67], a broader and more flexible class of models for processing continuous-time data. These models extend Linear CDEs by substituting the linear operator defining their transition dynamics with a neural network. Our key contribution, presented in this chapter, is showing that Neural CDEs give rise to a novel family of path-based kernels. These kernels, which we dub *beural signature kernels*, naturally generalize the classical signature kernel through their dependence on neural network activation functions.

4.1 Introduction

Definition 4.1. *A Neural CDE is a Controlled Differential Equation of the form*

$$Z_t^\omega = Z_0 + \sum_{i=1}^{d} \int_0^t \Phi_{i;L_i}(Z_t^\omega) \mathrm{d}\omega_t^i \tag{4.1}$$

defined by Feed Forward Networks (FFNs) $\Phi_{1;L_1}, \dots, \Phi_{d;L_d} : \mathbb{R}^N \to \mathbb{R}^N$.

Note how this definition extends that of linear CDEs, recovered by taking the FFNs to be affine functions. In particular, in this chapter we will consider the case where the FFNs are of form

$$A_i \varphi(z) + b_i$$

for some $A_i \in \mathbb{R}^{N \times N}$, $b_i \in \mathbb{R}^N$ and shared nonlinearity $\varphi : \mathbb{R} \to \mathbb{R}$. These we will call *homogeneous controlled ResNets*, since they take the form of a controlled residual network where the transition matrices are homogeneous in time. We will prove that,

under suitable choices of weight matrices, such systems converge in distribution to Gaussian processes on path space with covariance structure defined by the kernel

$$(x, y) \mapsto \mathcal{K}_\varphi^{x,y}(1,1) = \sum_{i=1}^{d} \int_0^1 \int_0^1 \left(\sigma_A^2 V_\varphi(\Sigma_\varphi^{x,y}(s,t)) + \sigma_b^2\right) \mathrm{d}x_s^i \mathrm{d}y_t^i$$

where

$$V_\varphi(\Sigma) = \mathbb{E}_{\mathbf{z}\sim\mathcal{N}(\mathbf{0},\Sigma)}[\varphi(z_1)\varphi(z_2)], \quad \Sigma_\varphi^{x,y}(s,t) = \begin{pmatrix} \mathcal{K}_\varphi^{x,x}(s,s) & \mathcal{K}_\varphi^{x,y}(s,t) \\ \mathcal{K}_\varphi^{x,y}(s,t) & \mathcal{K}_\varphi^{y,y}(t,t) \end{pmatrix}$$

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

$$\mathcal{K}_\varphi^{x,y}(0,0) = \mathcal{K}_\varphi^{x,y}(s,0) = \mathcal{K}_\varphi^{x,y}(0,t) = 1$$

The proof strategy diverges from the linear case, where we relied on the genus techniques developed in chapters 1 and 2. Here, we adopt a fundamentally different approach for two principal reasons. First, the theory presented in this chapter was developed independently and predates the genus framework. Second, while the genus techniques can indeed be extended to this setting, their application remains primarily at a formal level. As we detail in chapter D, at this formal level the approach yields precisely the expected results, recovering and generalising those obtained via the method presented in this chapter.

Our approach will therefore be different, inherently discrete in nature, employing careful limit arguments to extend discrete-time results to the continuous setting through delicate (due to the need of absolute convergence) commutativity considerations. We note that the graph expansion techniques of chapter 2 were, in fact, originally conceived within the Neural CDE framework specifically to avoid such discrete-time approximations, representing a more direct path to the continuous-time results presented in this chapter.

4.2 Neural Signature Kernels

4.2.1 Notation

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

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

$d \in \mathbb{N}$. More specifically, we consider the space[2]

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

of continuous paths with a square integrable derivative.

We denote by $x^j(t) \in \mathbb{R}$ the j^{th} coordinate of a path $x \in \mathbb{X}$ for $j \in \{1,\dots,d\}$. Given n paths $\mathcal{X} = \{x_1,\dots,x_n\} \subset \mathbb{X}$ 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 $[G(\mathcal{X},\mathcal{X})]^\beta_\alpha = G(x_\alpha, x_\beta)$ 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 $|\mathcal{D}| := \max_{i=1,\dots,\|\mathcal{D}\|} |t_i - t_{i-1}|$.

For an activation function $\varphi : \mathbb{R} \to \mathbb{R}$ and a positive semidefinite matrix $\Sigma \in \mathbb{R}^{2\times 2}$, in the chapter 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 theorem 4.10 in the appendix.

4.2.2 The Model

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

$$O^{M,N}_\varphi(x) := \langle \mathbf{o}, Z^{M,N}_{t_M}(x)\rangle_{\mathbb{R}^N}$$

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

$$Z^{M,N}_{t_{i+1}} = Z^{M,N}_{t_i} + \sum_{k=1}^{d} \Big(A_k \varphi(Z^{M,N}_{t_i}) + b_k\Big)\Delta x^k_{t_{i+1}}$$

with initial condition $Z_{t_0} = a$ 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

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

4.2.3 The infinite-width-depth regime

In this regime, as theorem 4.2 states, these architectures converge weakly to GPs indexed on the path space $\mathbb{X}$ with kernels satisfying a two-parameters differential equa-

[2]This is the Sobolev space H^1_0.

tion.

Theorem 4.2. *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*

$$\lim_{M\to\infty}\lim_{N\to\infty} O_\varphi^{M,N} = \mathcal{GP}(0, \mathcal{K}_\varphi) \tag{4.2}$$

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

$$\partial_s\partial_t\mathcal{K}_\varphi^{x,y} = \left[\sigma_A^2 V_\varphi\left(\Sigma_\varphi^{x,y}(s,t)\right) + \sigma_b^2\right]\langle \dot{x}_s, \dot{y}_t\rangle \tag{4.3}$$

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

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

Remark. *It is non-trivial to show not only that the problem is well-posed but even that eq. (4.3) 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 (4.3) 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. This result is proved in several steps in section 4.3.1. The first step consists of showing that the infinite width limit is well defined for any choice of $\mathcal{D}_M$. This will be done using the *tensor program* formalism of Yang [131], and 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 an Euler discretization, on $\mathcal{D}_M\times\mathcal{D}_M$, of (4.3). 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 eq. (4.3). The final step, proved in section 4.3.4, 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. □

When the activation function φ is the identity, eq. (4.3) 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. [112] to satisfy a similar PDE. This is the content of the next corollary.

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

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

where $k_{sig}^{x,y}$ is the signature kernel from [112] which for any two paths $x, y \in \mathbb{X}$ and $s, t \in [0,1]$ satisfies the following linear hyperbolic PDE

$$\partial_s \partial_t k_{sig}^{x,y} = \langle \dot{x}_s, \dot{y}_t \rangle k_{sig}^{x,y} \tag{4.4}$$

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. [112]. 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. [9] which use a different inner product structure on the space where signatures live.

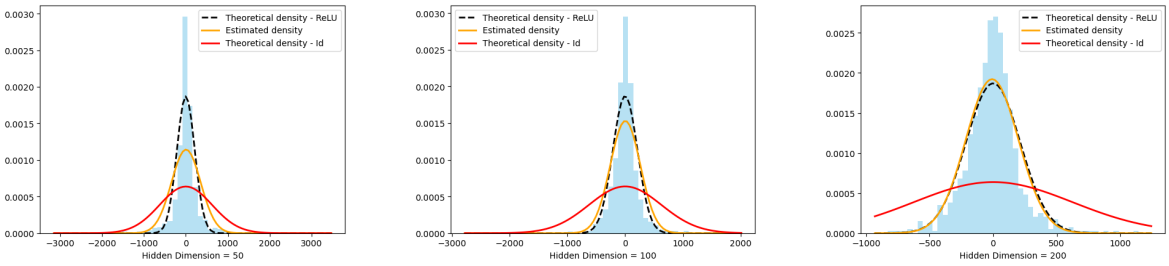


Figure 4.1: *Fixed input $y(t) := cos(15t) + 3e^t$. Histogram of 700 independent realizations of $N^{-1}\langle Z_1(y), Z_1(y)\rangle_{\mathbb{R}^N}$ for ReLU-RandomizedSignatures $S(y) \in \mathbb{R}^N$, $N \in \{50, 100, 200\}$, plotted against Closest Gaussian Fit, ReLU-NeuralSigKer and id-NeuralSigKer.*

Remark. *For non-linear activation functions φ, the neural signature kernel non-linear PDE (4.3) and the id-neural signature kernel linear PDE (4.4) might in principle admit*

the same solution, which would mean essentially that linear and non-linear controlled ResNets behave in the same way in the infinite-width-depth regime. In in fig. 4.1 we show empirically that this is not the case in general, by comparing the limiting empirical distributions for $\varphi = id$ and $\varphi = ReLU$.

4.2.4 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 $t \mapsto Z^{M,N}_{\rho_M(t)}(x)$ converges in distribution[3], as $M \to \infty$, to the solution $Z^N(x)$ of the following Neural CDE*

$$dZ^N_t(x) = \sum_{j=1}^{d} \Big(A_j\varphi(Z^N_t(x)) + b_j\Big)dx^j_t \tag{4.5}$$

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. □

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

The solutions to eq. (4.5) have been informally introduced in Akyildirim et al. [1], Cuchiero et al. [20] as lower dimensional approximations of path-signatures, and have been dubbed by the authors *randomized signatures*.

Taking directly the infinite width limit is far from trivial, reasoning *à la* Tensor Program [131] quickly collapses and there is no clear possible future path corresponding to the McKean-Vlasov ideas for the inhomogeneous case studied in Muca Cirone et al. [88]. 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.

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

4.2.5 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

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

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 $\alpha, \beta \in \{1, .., n\}$.

4.2.6 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 eq. (4.3).

Here are the crucial assumptions we make on the activation:

Assumption 4.4. *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 4.5. *The activation function $\varphi : \mathbb{R} \to \mathbb{R}$ is absolutely continuous and with exponentially bounded derivative.*

Lemma 4.6. *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

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

Concerning the second, notice how

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

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 $\epsilon \geq 0$. □

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

$$\|\varphi(x)\|_{\mathbb{R}^N} \leq \sqrt{2}M(\sqrt{N} + \|x\|_{\mathbb{R}^N}) \tag{4.6}$$

4.2.7 Positive semidefinite matrices and the map V_φ

Definition 4.7. *Let PSD_2 denote the set of 2×2 positive semidefinite matrices*

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

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 4.8. *Under theorem 4.5 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*

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

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

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

Proof. This is the content of theorem F.4 in [94]. □

Proposition 4.9. *Under theorem 4.4, 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. By *Cauchy-Schwartz* one has

$$|V_\varphi(\Sigma)|^2 \leq \mathbb{E}[\varphi(\mathbf{z}_x)^2]\cdot\mathbb{E}[\varphi(\mathbf{z}_y)^2]$$

and since for positive reals a, b it holds $(a+b)^2 \leq 2(a^2+b^2)$ we obtain

$$\mathbb{E}[\varphi(z)^2] \leq M^2\mathbb{E}[(1+|z|)^2] \leq 2M^2(1+\mathbb{E}[|z|^2])$$

hence

$$|V_\varphi(\Sigma)|^2 \leq 4M^4(1+[\Sigma]_1^1)(1+[\Sigma]_2^2).$$

We conclude by noticing that for $a \geq 0$ one has $\sqrt{1+a} \leq 1+\sqrt{a}$ since

$$1+a \leq 1+a+2\sqrt{a} = (1+\sqrt{a})^2.$$

□

We end the section showing the explicit characterization of the maps V_φ for some selected[4] activation functions. We write V_φ with the obvious meaning.

Proposition 4.10. *Defining* $\gamma(\Sigma) := \frac{[\Sigma]_1^2}{\sqrt{[\Sigma]_1^1[\Sigma]_2^2}}$ *we have*

$$\begin{aligned} V_{id}(\Sigma) &= [\Sigma]_1^2 = [\Sigma]_2^1 \\ V_{ReLU}(\Sigma) &= \frac{1}{2\pi}\left(\pi + \frac{\sqrt{1-\gamma(\Sigma)^2}}{\gamma(\Sigma)} - arccos(\gamma(\Sigma))\right)[\Sigma]_1^2 \\ V_{erf}(\Sigma) &= \frac{2}{\pi} arcsin\left(\frac{[\Sigma]_1^2}{\sqrt{(0.5+[\Sigma]_1^1)(0.5+[\Sigma]_2^2)}}\right) \end{aligned} \tag{4.7}$$

Proof. See [129][Facts B.2, B.3]. □

[4] by the availability in the literature.

4.3 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* $O_\varphi^{M,N} : \mathbb{X} \to \mathbb{R}$ is defined as follows

$$O_\varphi^{M,N}(x) := \langle \mathbf{o}, Z_{t_M}^{M,N}(x) \rangle_{\mathbb{R}^N}$$

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

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

with initial condition $Z_{t_0} = a$ 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

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

The section too will be subdivided in two main parts: in the first we consider the infinite-width-then-depth limit and in the second we reverse the order and consider the infinite-depth-then-width limit.

4.3.1 The infinite-width-then-depth regime

The main goal is that of proving theorem 4.2, which we restate here:

Theorem 4.11. *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 finite subset $\mathcal{X} = \{x_1, \dots, x_n\} \subset \mathbb{X}$ of paths the following convergence in distribution holds*

$$\lim_{M \to \infty} \lim_{N \to \infty} O_\varphi^{M,N}(\mathcal{X}) = \mathcal{N}(0, \mathcal{K}_\varphi(\mathcal{X}, \mathcal{X})) \tag{4.8}$$

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

$$\partial_s \partial_t \mathcal{K}_\varphi^{x,y} = \left[\sigma_A^2 V_\varphi \left(\Sigma_\varphi^{x,y} \right) + \sigma_b^2 \right] \langle \dot{x}_s, \dot{y}_t \rangle \tag{4.9}$$

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

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

   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 [129] in section 4.3.1.
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 eq. (4.9). This will be done in section 4.3.1.

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*.

Definition 4.12. *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

$$\begin{aligned}\mathcal{K}_{\mathcal{D}\times\mathcal{D}'}(s_m,t_n) &= \mathcal{K}_{\mathcal{D}\times\mathcal{D}'}(s_{m-1},t_n) + \mathcal{K}_{\mathcal{D}\times\mathcal{D}'}(s_m,t_{n-1}) - \mathcal{K}_{\mathcal{D}\times\mathcal{D}'}(s_{m-1},t_{n-1}) \\ &+ \left(\sigma_A^2 V_\varphi\left(\Sigma_{\mathcal{D}\times\mathcal{D}'}^{x,y}(s_{m-1},t_{n-1})\right) + \sigma_b^2\right)\langle \Delta x_{s_m}, \Delta y_{t_n}\rangle_{\mathbb{R}^d} \end{aligned} \tag{4.10}$$

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

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

Theorem 4.13. *Let* $\varphi : \mathbb{R}\to\mathbb{R}$ *be linearly bounded*[5]. *For any subset* $\mathcal{X} = \{x_1,\ldots,x_n\} \subset$

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

$\mathbb{X}$ *the following convergence in distribution holds*

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

Proof. We use Yang [129][Corollary 5.5] applied to the Tensor Program of Algorithm algorithm 1 where the input variables are independently sampled according to

$$[A_k]_\alpha^\beta \sim \mathcal{N}(0, \frac{\sigma_A^2}{N}), [v]_\alpha \sim \mathcal{N}(0,1), [Z_0]_\alpha \sim \mathcal{N}(0,\sigma_a^2) \text{ and } [b_k]_\alpha \sim \mathcal{N}(0,\sigma_b^2)$$

Note how the sampling scheme follows [129][Assumption 5.1] and how linearly bounded functions are *controlled* in the sense of [129][Definition 5.3] since for all $x \in \mathbb{R}$ one has $|\varphi(x)| \leq C(1+|x|) \leq e^{|x|+log(C)}$. Thus we are under the needed assumptions to apply [129][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 $\mathbf{o} \sim \mathcal{N}(0, \frac{1}{N})$. 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 algorithm 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.

Note that Z_m^x stands for the formal variable in program algorithm 1, not for $Z_{t_m^{\mathcal{D}_M}}^{M,N}(x)$ even thogh this is the value which it "stores" for a fixed hidden dimension N.

The result in [129][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

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

with μ, Σ computed according to [129][Definition 5.2] and defined on the set of all G-vars in the program *i.e.*

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

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

where $h = \phi((g_k)_{k=1}^m)$ for some function ϕ and $\phi(\mathbf{z}) := \phi((\mathbf{z}^{g_k})_{k=1}^m)$, 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 $m_i, m_j \in \{1, \dots, \|\mathcal{D}_M\|\}$, we obtain

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

Now

$$\Sigma(\gamma_{m_i,k}^i, \gamma_{m_j,l}^j) = \delta_{k,l}\sigma_A^2 \mathbb{E}[\varphi(\mathbf{z}_1)\varphi(\mathbf{z}_2)] + \Sigma(b_k, b_l) = \delta_{kl}\left[\sigma_A^2 \mathbb{E}[\varphi(\mathbf{z}_1)\varphi(\mathbf{z}_2)] + \sigma_b^2\right]$$

where $[\mathbf{z}_1, \mathbf{z}_2]^\top \sim \mathcal{N}(0, \tilde{\Sigma}_{m_i-1,m_j-1}(x_i, x_j))$ with

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

thus if we set, for $t_{m_i}, t_{m_j} \in \mathcal{D}_M$, $\mathcal{K}_{\mathcal{D}_M\times\mathcal{D}_M}^{x_i,x_j}(t_{m_i}, t_{m_j}) := \Sigma(Z_{m_i}^{x_i}, Z_{m_j}^{x_j})$ 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 eq. (4.10) states. Then note how

$$\Sigma(Z_{m_i}^{x_i}, Z_0^{x_j}) = \Sigma(Z_{m_i-1}^{x_i}, Z_0^{x_j}) = \cdots = \Sigma(Z_0^{x_i}, Z_0^{x_j}) = \sigma_a^2$$

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

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

□

Algorithm 1 $Z_1^{M,N}$ as Nestor program

Input: $Z_0 : \mathsf{G}(N)$ ▷ *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 $Z_1^{\mathcal{D}_M,N}(x_i)$ (here $Z_0^{x_i}$ is to be read as Z_0)
  for $m = 1, \ldots, \|\mathcal{D}_M\|$ **do**
    for $k = 1, \ldots, d$ **do**
      $\alpha_{m,k}^i := \varphi(Z_{m-1}^{x_i}) : \mathsf{H}(N)$ ▷ *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
    $Z_m^{x_i} := Z_{m-1}^{x_i} + \sum_{k=1}^d \gamma_{m,k}^i [(x_i)_{t_m}^k - (x_i)_{t_{m-1}}^k] : \mathsf{G}(N)$ ▷ *by LinComb;*
  end for
end for
Return: $(v^T Z_{\|\mathcal{D}_M\|}^{x_i} / \sqrt{N})_{i=1,\ldots,n}$

Actually we have proved the even stronger statement that

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

$$\langle \mathbf{o}^N, Z_t^{M,N}(\mathcal{X}) \rangle \langle \mathbf{o}^N, Z_s^{M,N}(\mathcal{X}) \rangle \xrightarrow[N\to\infty]{} \mathcal{N}(0, K^{\mathcal{X},\mathcal{X}}_{\mathcal{D}_M \times \mathcal{D}_M}(t,s))$$

where $[\mathbf{o}^N]_\alpha \sim \mathcal{N}(0, \frac{1}{N})$ 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 4.14. *For all $t \in \mathcal{D}$, for all $s \in \mathcal{D}'$, with a clear abouse of notation, one has the following distributional limit*

$$\langle \mathbf{o}^N, Z_t^{\mathcal{D},N}(x) \rangle \langle \mathbf{o}^N, Z_s^{\mathcal{D}',N}(y) \rangle \xrightarrow[N\to\infty]{} \mathcal{N}(0, K^{x,y}_{\mathcal{D}\times\mathcal{D}'}(t,s))$$

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

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

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 $[0,1]^2$:

Definition 4.15. *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*

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

   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 4.16. *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*

$$\partial_s\partial_t \mathcal{K}^{x,y}_\varphi = \left[\sigma_A^2 V_\varphi\left(\Sigma^{x,y}_\varphi\right) + \sigma_b^2\right]\langle \dot{x}_s, \dot{y}_t\rangle \tag{4.11}$$

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 and Victoir [37] and Huang and Tsai [56]). 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 theorem 4.8. 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 4.17. $\forall x \in \mathbb{X}. \forall t \in \mathcal{D}. \quad \tilde{\mathcal{K}}^{x,x}_{\mathcal{D}\times\mathcal{D}}(t,t) \geq \sigma_a^2 \quad$ *for any choice of* $\mathcal{D}$*.*

Proof. We have

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

and by Tensor Programs we know that

$$\sum_{\substack{0 \leq k_1 < m \\ 0 \leq k_2 < m}} \sum_{k=1}^{d} \left(\sigma_A^2 V_\varphi\left(\Sigma^{x,x}_{\mathcal{D}\times\mathcal{D}}(t_{k_1}, t_{k_2})\right) + \sigma_b^2\right) \Delta x^k_{t_{k_1+1}} \Delta x^k_{t_{k_2+1}} = \lim_{N\to\infty} \mathbb{E}\left[\frac{\langle Z^{\mathcal{D}N}_{t_m}(x) - Z^N_0, Z^{\mathcal{D}N}_{t_m}(x) - Z^N_0\rangle}{N}\right] \geq 0$$

□

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

Lemma 4.18. *There is a constant* $C_x > 0$ *independent of* $\mathcal{D}$ *such that*

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

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

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

In particular

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

and we can use Gronwall Inequality ([37], lemma 3.2) to obtain

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

hence the thesis with

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

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

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

□

Thus finally we can conclude, as promised, that

Lemma 4.19. *There is an $R = R_x \in \mathbb{R}$ such that $\tilde{\Sigma}^{x,x}_{\mathcal{D}\times\mathcal{D}} \in \mathrm{PSD}_2(R)$ 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 $\tilde{\mathcal{K}}^{x,x}_{\mathcal{D}\times\mathcal{D}}(s,s)$ are bounded above by a common constant C_x and below by σ_a^2 We thus can simply choose $R_x = C_x \vee \sigma_a^{-2}$. □

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

Proposition 4.20. *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*

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

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

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

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

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

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

$$\begin{aligned}
|\mathcal{K}^{x,y}_{\mathcal{D}\times\mathcal{D}'}(s,t)| &= |\sigma_a^2 + \int_{\eta=0}^{s}\int_{\tau=0}^{t} \left(\sigma_A^2 V_\varphi\left(\tilde{\Sigma}^{x,y}_{\mathcal{D}\times\mathcal{D}'}(\eta,\tau)\right) + \sigma_b^2\right)\langle \dot{x}_\eta, \dot{y}_\tau\rangle_{\mathbb{R}^d} d\eta d\tau| \\
&\leq \sigma_a^2 + \left(2\sigma_A^2\tilde{M}(1+\tilde{C}_x\vee\tilde{C}_y) + \sigma_b^2\right)\int_{\eta=0}^{s}\int_{\tau=0}^{t} |\langle \dot{x}_\eta, \dot{y}_\tau\rangle_{\mathbb{R}^d}| d\eta d\tau \\
&\leq \sigma_a^2 + \left(2\sigma_A^2\tilde{M}(1+\tilde{C}_x\vee\tilde{C}_y) + \sigma_b^2\right)\|x\|_{1-var,[0,1]}\|y\|_{1-var,[0,1]} \leq C_\alpha
\end{aligned}$$

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

Proof of theorem 4.16.

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 theorem 4.20. 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

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

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

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

hence

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

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

$$\sup_{[0,1]\times[0,1]} |\mathcal{K}^{x,y}_{\mathcal{D}\times\mathcal{D}'}(s,t) - \tilde{\mathcal{K}}^{x,y}_{\mathcal{D}\times\mathcal{D}'}(s,t)| \leq$$
$$2(2\sigma_A^2\tilde{M}(1+C_{x,y})+\sigma_b^2)\sqrt{|\mathcal{D}| \vee |\mathcal{D}'|}\sqrt{\|x\|^2_{\mathbb{X}}\|y\|^2_{\mathbb{X}}} \to 0$$

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

$$\sup_{[0,1]\times[0,1]} \left\|\Sigma^{x,y}_{\mathcal{D}\times\mathcal{D}'}(s,t) - \tilde{\Sigma}^{x,y}_{\mathcal{D}\times\mathcal{D}'}(s,t)\right\|_\infty \leq$$
$$2(2\sigma_A^2\tilde{M}(1+C_{x,y})+\sigma_b^2)\sqrt{|\mathcal{D}| \vee |\mathcal{D}'|}(\|x\|^2_{\mathbb{X}} \vee \|y\|^2_{\mathbb{X}}) =: \Gamma_{x,y}\sqrt{|\mathcal{D}| \vee |\mathcal{D}'|}$$

Part II : Cauchy Bounds

Consider now another partition $\check{\mathcal{D}} \times \check{\mathcal{D}}'$, we have

$$\begin{aligned}
&|\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\left\|\tilde{\Sigma}^{x,y}_{\check{\mathcal{D}}\times\check{\mathcal{D}}'}(\eta,\tau) - \tilde{\Sigma}^{x,y}_{\mathcal{D}\times\mathcal{D}'}(\eta,\tau)\right\|_\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 \left\|\tilde{\Sigma}_{\check{\mathcal{D}}\times\check{\mathcal{D}}'}(\eta,\tau) - \Sigma_{\check{\mathcal{D}}\times\check{\mathcal{D}}'}(\eta,\tau)\right\|_\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 \left\|\Sigma_{\check{\mathcal{D}}\times\check{\mathcal{D}}'}(\eta,\tau) - \Sigma_{\mathcal{D}\times\mathcal{D}'}(\eta,\tau)\right\|_\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 \left\|\Sigma_{\mathcal{D}\times\mathcal{D}'}(\eta,\tau) - \tilde{\Sigma}_{\mathcal{D}\times\mathcal{D}'}(\eta,\tau)\right\|_\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 \left\|\Sigma_{\check{\mathcal{D}}\times\check{\mathcal{D}}'}(\eta,\tau) - \Sigma_{\mathcal{D}\times\mathcal{D}'}(\eta,\tau)\right\|_\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\|^2_{1-\mathrm{var},[0,1]} \vee \|y\|^2_{1-\mathrm{var},[0,1]}) \\
&+\sigma_A^2 k_R\int_{\eta=0}^s\int_{\tau=0}^t \left\|\Sigma^{x,y}_{\check{\mathcal{D}}\times\check{\mathcal{D}}'}(\eta,\tau) - \Sigma^{x,y}_{\mathcal{D}\times\mathcal{D}'}(\eta,\tau)\right\|_\infty|\langle\dot{x}_\eta,\dot{y}_\tau\rangle_{\mathbb{R}^d}|d\eta d\tau
\end{aligned}$$

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

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

One has, from the previous inequality, that

$$\begin{aligned}\Xi_t &\leq \sigma_A^2 2\Gamma_{x,x}\sqrt{|\check{\mathcal{D}}| \vee |\mathcal{D}|}\,\|x\|^2_{1-\mathrm{var},[0,1]} \\ &\quad +\sigma_A^2 k_R \int_{\eta=0}^t\int_{\tau=0}^t \left\|\Sigma_{\check{\mathcal{D}}\times\check{\mathcal{D}}}(\eta,\tau) - \Sigma_{\mathcal{D}\times\mathcal{D}}(\eta,\tau)\right\|_\infty |\langle \dot{x}_\eta, \dot{x}_\tau\rangle_{\mathbb{R}^d}| d\eta d\tau\end{aligned}$$

and since

$$\begin{aligned}&\int_{\eta=0}^t\int_{\tau=0}^t \left\|\Sigma_{\check{\mathcal{D}}\times\check{\mathcal{D}}}(\eta,\tau) - \Sigma_{\mathcal{D}\times\mathcal{D}}(\eta,\tau)\right\|_\infty |\langle \dot{x}_\eta, \dot{x}_\tau\rangle_{\mathbb{R}^d}| d\eta d\tau \\ \leq &\int_{\eta=0}^t\int_{\tau=0}^t \Xi_{\eta\vee\tau}|\langle \dot{x}_\eta, \dot{x}_\tau\rangle_{\mathbb{R}^d}| d\eta d\tau \\ = &\int_{\eta=0}^t\int_{\tau=0}^\eta \Xi_\eta|\langle \dot{x}_\eta, \dot{x}_\tau\rangle_{\mathbb{R}^d}| d\eta d\tau + \int_{\eta=0}^t\int_{\tau=\eta}^t \Xi_\tau|\langle \dot{x}_\eta, \dot{x}_\tau\rangle_{\mathbb{R}^d}| d\eta d\tau \\ \leq &\int_{\eta=0}^t \|x\|_{1-\mathrm{var},[0,\eta]}\,\Xi_\eta|\dot{x}_\eta| d\eta + \int_{\tau=0}^t\int_{\eta=0}^\tau \Xi_\tau|\dot{x}_\eta||\dot{x}_\tau| d\eta d\tau \\ = &\int_{\eta=0}^t \|x\|_{1-\mathrm{var},[0,\eta]}\,\Xi_\eta|\dot{x}_\eta| d\eta + \int_{\tau=0}^t \|x\|_{1-\mathrm{var},[0,\tau]}\,\Xi_\tau|\dot{x}_\tau| d\tau \\ \leq &2\,\|x\|_{1-\mathrm{var},[0,1]}\int_{\eta=0}^t \Xi_\eta|\dot{x}_\eta| d\eta\end{aligned}$$

we get

$$\begin{aligned}\Xi_t &\leq \sigma_A^2 2\Gamma_{x,x}\,\|x\|^2_{1-\mathrm{var},[0,1]}\sqrt{|\check{\mathcal{D}}| \vee |\mathcal{D}|} \\ &\quad +2\sigma_A^2 k_R\,\|x\|_{1-\mathrm{var},[0,1]}\int_{\eta=0}^t \Xi_\eta|\dot{x}_\eta| d\eta\end{aligned}$$

thus, by Gronwall,

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

Coming back to the general case and setting

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

we can now say that

$$\begin{aligned}\Xi_t^{x,y} &:= \sup_{0\leq\eta,\tau\leq t}\left\|\Sigma_{\check{\mathcal{D}}\times\check{\mathcal{D}}'}(\eta,\tau) - \Sigma_{\mathcal{D}\times\mathcal{D}'}(\eta,\tau)\right\|_\infty \\ &\leq 2\sigma_A^2\Gamma_{x,y}\sqrt{|\check{\mathcal{D}}| \vee |\check{\mathcal{D}}'| \vee |\mathcal{D}| \vee |\mathcal{D}'|}\Lambda_{x,y}(1+e^{2\sigma_A^2 k_R\Lambda_{x,y}}) \\ &\quad +\sigma_A^2 k_R\int_{\eta=0}^t\int_{\tau=0}^t \left\|\Sigma_{\check{\mathcal{D}}\times\check{\mathcal{D}}'}(\eta,\tau) - \Sigma_{\mathcal{D}\times\mathcal{D}'}(\eta,\tau)\right\|_\infty |\langle \dot{x}_\eta, \dot{y}_\tau\rangle_{\mathbb{R}^d}| d\eta d\tau\end{aligned}$$

and since

$$\begin{aligned}
&\int_{\eta=0}^{t}\int_{\tau=0}^{t} \left\|\Sigma_{\check{\mathcal{D}}\times\check{\mathcal{D}}'}(\eta,\tau) - \Sigma_{\mathcal{D}\times\mathcal{D}'}(\eta,\tau)\right\|_\infty |\langle \dot{x}_\eta, \dot{y}_\tau\rangle_{\mathbb{R}^d}| d\eta d\tau \\
\leq &\int_{\eta=0}^{t}\int_{\tau=0}^{t} \Xi^{x,y}_{\eta\vee\tau} |\langle \dot{x}_\eta, \dot{y}_\tau\rangle_{\mathbb{R}^d}| d\eta d\tau \\
= &\int_{\eta=0}^{t}\int_{\tau=0}^{\eta} \Xi^{x,y}_{\eta} |\langle \dot{x}_\eta, \dot{y}_\tau\rangle_{\mathbb{R}^d}| d\eta d\tau + \int_{\eta=0}^{t}\int_{\tau=\eta}^{t} \Xi^{x,y}_{\tau} |\langle \dot{x}_\eta, \dot{y}_\tau\rangle_{\mathbb{R}^d}| d\eta d\tau \\
\leq &\int_{\eta=0}^{t} \|y\|_{1-\mathrm{var},[0,\eta]} \Xi^{x,y}_{\eta} |\dot{x}_\eta| d\eta + \int_{\tau=0}^{t}\int_{\eta=0}^{\tau} \Xi^{x,y}_{\tau} |\dot{x}_\eta||\dot{y}_\tau| d\eta d\tau \\
= &\int_{\eta=0}^{t} \|y\|_{1-\mathrm{var},[0,\eta]} \Xi^{x,y}_{\eta} |\dot{x}_\eta| d\eta + \int_{\tau=0}^{t} \|x\|_{1-\mathrm{var},[0,\tau]} \Xi^{x,y}_{\tau} |\dot{y}_\tau| d\tau \\
\leq &(\|x\|_{1-\mathrm{var},[0,1]} \vee \|y\|_{1-\mathrm{var},[0,1]}) \int_{\eta=0}^{t} \Xi^{x,y}_{\eta} (|\dot{x}_\eta| + |\dot{y}_\eta|) d\eta
\end{aligned}$$

we get

$$\begin{aligned}
\Xi^{x,y}_t \leq\ & 2\sigma_A^2 \Gamma_{x,y}\Lambda_{x,y}\sqrt{|\check{\mathcal{D}}| \vee |\check{\mathcal{D}}'| \vee |\mathcal{D}| \vee |\mathcal{D}'|}(1 + e^{2\sigma_A^2 k_R \Lambda_{x,y}}) \\
&+ 2\sigma_A^2 k_R (\|x\|_{1-\mathrm{var},[0,1]} \vee \|y\|_{1-\mathrm{var},[0,1]}) \int_{\eta=0}^{s} \Xi^{x,y}_{\eta} (|\dot{x}_\eta| + |\dot{y}_\eta|) d\eta
\end{aligned}$$

thus, by Gronwall,

$$\begin{aligned}
\Xi_t \leq\ & 2\sigma_A^2 \Gamma_{x,y}\Lambda_{x,y}\sqrt{|\check{\mathcal{D}}| \vee |\check{\mathcal{D}}'| \vee |\mathcal{D}| \vee |\mathcal{D}'|}(1 + e^{2\sigma_A^2 k_R \Lambda_{x,y}}) \\
&\cdot e^{2\sigma_A^2 k_R (\|x\|_{1-\mathrm{var},[0,1]} \vee \|y\|_{1-\mathrm{var},[0,1]})(\|x\|_{1-\mathrm{var},[0,t]} + \|y\|_{1-\mathrm{var},[0,t]})}
\end{aligned}$$

hence

$$\begin{aligned}
&\left\|\mathcal{K}^{x,y}_{\check{\mathcal{D}}\times\check{\mathcal{D}}'} - \mathcal{K}^{x,y}_{\mathcal{D}\times\mathcal{D}'}\right\|_{\infty,[0,1]^2} \leq \\
&2\sigma_A^2 \Gamma_{x,y}\Lambda_{x,y}\sqrt{|\check{\mathcal{D}}| \vee |\check{\mathcal{D}}'| \vee |\mathcal{D}| \vee |\mathcal{D}'|} \cdot (1 + e^{2\sigma_A^2 k_R \Lambda_{x,y}}) \cdot e^{2\sigma_A^2 k_R \Lambda_{x,y}}
\end{aligned}$$

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

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

and the limit $\mathcal{K}^{x,y}_\varphi$ 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 $\Sigma^{x,y}_\varphi(s,t)$ obtained as limits using the previous result are all PSD. We can actually say more: they belong to $\mathrm{PSD}(R)$.

Part IV : Limit Kernel solves equation (4.9) We can finally conclude by proving that $\mathcal{K}_\varphi^{x,y}$ is, in fact, a solution of the PDE:

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

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 $\mathbb{X}$. □

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 4.21 (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 (4.9)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 (4.9) 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 $(x,x),(x,y),(y,y)$.

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

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

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

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

thus

$$
\begin{aligned}
&\Xi_t \leq \sigma_A^2 k_{\bar{R}} \int_{\eta=0}^{t}\int_{\tau=0}^{t} \Xi_{\eta\vee\tau}|\dot{x}_\eta||\dot{y}_\tau| d\eta d\tau \\
\leq &\sigma_A^2 k_{\bar{R}} \left(\int_{\eta=0}^{t}\int_{\tau=0}^{\eta} \Xi_\eta|\dot{x}_\eta||\dot{y}_\tau| d\eta d\tau + \int_{\eta=0}^{t}\int_{\tau=\eta}^{t} \Xi_\tau|\dot{x}_\eta||\dot{y}_\tau| d\eta d\tau\right) \\
\leq &\sigma_A^2 k_{\bar{R}} \left(\|y\|_{1-var,[0,1]} \int_{\eta=0}^{t} \Xi_\eta|\dot{x}_\eta| d\eta + \|x\|_{1-var,[0,1]} \int_{\tau=0}^{t} \Xi_\tau|\dot{y}_\tau| d\tau\right) \\
\leq &\sigma_A^2 k_{\bar{R}} (\|x\|_{1-var,[0,1]} \vee \|y\|_{1-var,[0,1]}) \int_{\eta=0}^{t} \Xi_\eta(|\dot{x}_\eta| + |\dot{y}_\eta|) d\eta
\end{aligned}
$$

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

4.3.2 Proof of theorem 4.2: Part I

It is finally time to prove the first part of main result of the chapter, 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 $O_\varphi^{\mathcal{D}_M,N}(\cdot) = \langle v^N, Z_1^{\mathcal{D}_M,N}(\cdot)\rangle$.

Theorem 4.22. *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*

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

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

$$\mathcal{K}_\varphi^{x,y}(s,t) = \sigma_a^2 + \int_{\eta=0}^{s}\int_{\tau=0}^{t} \left[\sigma_A^2 \mathbb{E}_{\mathbf{z}\sim\mathcal{N}(0,\Sigma_{\eta,\tau}^{x,y})}[\varphi(\mathbf{z}_1)\varphi(\mathbf{z}_2)] + \sigma_b^2\right] \langle \dot{x}_\eta, \dot{y}_\tau\rangle_{\mathbb{R}^d} d\eta d\tau \tag{4.12}$$

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 4.13 and theorem 4.16.

Under our hypotheses theorem 4.13 tells us that, for any subset $\mathcal{X} = \{x_1, \dots, x_n\} \subset \mathbb{X}$, we have in distribution

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

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 4.16 with the sequence $\{\mathcal{D}_M \times \mathcal{D}_M\}_{M\in\mathbb{N}}$. The uniqueness follows from theorem 4.21. □

We can explicitly write the Kernel for the simplest case:

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

$$\mathcal{K}_{id}^{x,y}(s,t) = \Big(\sigma_a^2 + \frac{\sigma_b^2}{\sigma_A^2}\Big) k_{sig}^{\sigma_A x, \sigma_A y}(s,t) - \frac{\sigma_b^2}{\sigma_A^2} \tag{4.13}$$

Proof. We readily see that

$$\mathcal{K}_{id}^{x,y}(s,t) = \sigma_a^2 + \int_{\eta=0}^{s}\int_{\tau=0}^{t} \left[\sigma_A^2 \mathcal{K}_{id}^{x,y}(\eta,\tau) + \sigma_b^2\right] \langle \dot{x}_\eta, \dot{y}_\tau\rangle_{\mathbb{R}^d} d\eta d\tau \tag{4.14}$$

By substituting (4.13) for $\mathcal{K}_{id}^{x,y}(\eta,\tau)$ in the integral and using [112][Theorem 2.5]

we have

$$\begin{aligned}
&\sigma_a^2 + \int_{\eta=0}^{s}\int_{\tau=0}^{t}\Big[\sigma_A^2\Big\{\Big(\sigma_a^2 + \frac{\sigma_b^2}{\sigma_A^2}\Big)k_{sig}^{\sigma_A x,\sigma_A y}(\eta,\sigma) - \frac{\sigma_b^2}{\sigma_A^2}\Big\} + \sigma_b^2\Big]\langle \dot{x}_\eta, \dot{y}_\tau\rangle_{\mathbb{R}^d} d\eta d\tau \\
=&\sigma_a^2 + \int_{\eta=0}^{s}\int_{\tau=0}^{t}\sigma_A^2\Big(\sigma_a^2 + \frac{\sigma_b^2}{\sigma_A^2}\Big)k_{sig}^{\sigma_A x,\sigma_A y}(\eta,\sigma)\langle \dot{x}_\eta, \dot{y}_\tau\rangle_{\mathbb{R}^d} d\eta d\tau \\
=&\sigma_a^2 + \Big(\sigma_a^2 + \frac{\sigma_b^2}{\sigma_A^2}\Big)\int_{\eta=0}^{s}\int_{\tau=0}^{t} k_{sig}^{\sigma_A x,\sigma_A y}(\eta,\sigma)\langle \sigma_A\dot{x}_\eta, \sigma_A\dot{y}_\tau\rangle_{\mathbb{R}^d} d\eta d\tau \\
=&\sigma_a^2 + \Big(\sigma_a^2 + \frac{\sigma_b^2}{\sigma_A^2}\Big)(k_{sig}^{\sigma_A x,\sigma_A y}(s,t) - 1) \\
=&\Big(\sigma_a^2 + \frac{\sigma_b^2}{\sigma_A^2}\Big)k_{sig}^{\sigma_A x,\sigma_A y}(s,t) - \frac{\sigma_b^2}{\sigma_A^2}
\end{aligned}$$

which means, by uniqueness of solutions (theorem 4.21), that the thesis holds. □

4.3.3 The infinite-depth-then-width regime

The infinite-depth limit with finite-width

Proposition 4.23. *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. Let $x \in \mathbb{X}$ and let $\rho_M(t) := \sup\{s \in \mathcal{D}_M : s \le t\}$. Then, the the $\mathbb{R}^N$-valued process $t \mapsto Z^{M,N}_{\rho_M(t)}(x)$ converges in distribution, as $M \to \infty$, to the solution $Z^N(x)$ of the following Neural CDE*

$$Z_t^N(x) = a + \int_0^t \sum_{j=1}^{d}\Big(A_j\varphi(Z_s^N(x)) + b_j\Big)dx_s^j \tag{4.15}$$

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 and Victoir [37][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 4.2 one gets that

$$\big\|Z^{M,N}(x) - Z^N(x)\big\|_{\infty,[0,1]} \le \sqrt{|\mathcal{D}_M|}\exp\left\{K_{\|x\|_{\mathbb{X}}}(\|a\|_{\mathbb{R}^N} + \sum_{k=1}^{d}\|A_k\|_F + \|b_k\|_F)\right\}$$

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

$$\mathbb{E}[\left\| Z^{M,N}(x) - Z^N(x) \right\|_{\infty,[0,1]}] \leq \tilde{K}_{\|x\|_{\mathbb{X}}} \sqrt{|\mathcal{D}_M|}$$

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 $C^0([0,1];\mathbb{R}^N)$.

□

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 4.24 (Randomized Signatures)**.** *We call* Randomized Signatures *the solutions to*

$$Z_t^N(x) = a + \sum_{k=1}^{d} \int_0^t \Big(A_k[\varphi(Z_\tau^N(x))] + b_k \Big) dx_\tau^k$$

These are the same objects defined in Cuchiero et al. [20].

4.3.4 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 Z_0, A_k, b_k for all partitions. For any $x \in \mathbb{X}$ let $Z^{M,N}(x) : \mathcal{D}_M \to \mathbb{R}^N$ be the homogeneous cResNet corresponding to the above quantities.

Definition 4.25. *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 $\tilde{Z}^{M,N}(x) : [0,1] \to \mathbb{R}^N$

$$\tilde{Z}_t^{M,N}(x) := Z_{\rho_M(t)}^{M,N}(x) \tag{4.16}$$

and the integral extension $Z^{M,N}(x) : [0,1] \to \mathbb{R}^N$

$$Z_t^{M,N}(x) = Z_0^N + \sum_{k=1}^{d} \int_0^t \Big(A_k \varphi(\tilde{Z}_s^{M,N}(x)) + b_k \Big) dx_s^k \tag{4.17}$$

Note how the two extensions coincide on $\mathcal{D}_M$.

Proposition 4.26. *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*

$$\mathbb{E}\left[\left\|\tilde{Z}_t^{N,M}(x)\right\|_{\infty,[0,1]}^2\right] \leq NK_x$$

where the expectation is taken over the joint distribution of $Z_0, \{A_k, b_k\}_{k=1,\dots,d}$

Proof.

$$\begin{aligned}
\left\|\tilde{Z}_t^{N,M}(x)\right\| = \left\|Z_{t_m}^{N,M}(x)\right\| &\leq \left\|Z_0^N\right\| + \sum_{i=1}^d \left\|\int_0^t A_i^N \varphi(\tilde{Z}_r^{N,M}(x)) + b_i^N dx_r^i\right\| \\
&\leq \left\|Z_0^N\right\| + \sum_{i=1}^d \int_0^t \left\|A_i^N \varphi(\tilde{Z}_r^{N,M}(x)) + b_i^N\right\| |dx_r^i| \\
&\leq \left\|Z_0^N\right\| + \sum_{i=1}^d \int_0^t \left(\left\|A_i^N\right\|_{op} \left\|\varphi(\tilde{Z}_r^{N,M}(x))\right\| + \left\|b_i^N\right\| \right) |dx_r^i| \\
&\leq \left\|Z_0^N\right\| + \sum_{i=1}^d \int_0^t \left(\left\|A_i^N\right\|_{op} \sqrt{2}C(\sqrt{N} + \left\|\tilde{Z}_r^{N,M}(x)\right\|) + \left\|b_i^N\right\| \right) |dx_r^i| \\
&\leq \left[\left\|Z_0^N\right\| + (\sum_{i=1}^d \sqrt{2N}C \left\|A_i^N\right\|_{op} + \left\|b_i^N\right\|) \|x\|_{1-var,[0,1]}\right] \\
&\quad + \sqrt{2}C(\sum_{i=1}^d \left\|A_i^N\right\|_{op}) \sum_{i=1}^d \int_0^t \left\|\tilde{Z}_r^{N,M}(x)\right\| |dx_r^i|
\end{aligned}$$

thus by *Lemma 3.2* of [37] we obtain

$$\begin{aligned}
\left\|\tilde{Z}_t^{N,M}(x)\right\| &\leq \left[\left\|Z_0^N\right\| + (\sum_{i=1}^d \sqrt{2N}C \left\|A_i^N\right\|_{op} + \left\|b_i^N\right\|) \|x\|_{1-var,[0,1]}\right] \\
&\quad \cdot \exp\left\{\sqrt{2}C(\sum_{i=1}^d \left\|A_i^N\right\|_{op}) \|x\|_{1-var,[0,t]}\right\} \\
&\leq \left[\left\|Z_0^N\right\| + (\sqrt{2N}C + \sum_{i=1}^d \left\|b_i^N\right\|) \|x\|_{1-var,[0,1]}\right] \\
&\quad \cdot \exp\left\{(\sum_{i=1}^d \left\|A_i^N\right\|_{op})(1 + \sqrt{2}C \|x\|_{1-var,[0,t]})\right\} \\
&\leq \left[\left\|Z_0^N\right\| + (\sqrt{2N}C + \sum_{i=1}^d \left\|b_i^N\right\|) \|x\|_{1-var,[0,1]}\right] \\
&\quad \cdot \prod_{i=1}^d \exp\left\{(1 + \sqrt{2}C \|x\|_{1-var,[0,1]}) \left\|A_i^N\right\|_{op}\right\}
\end{aligned}$$

Hence using independence

$$\mathbb{E}[\sup_{t\in[0,1]} \left\|\tilde{Z}_t^{N,M}(x)\right\|^2] \leq$$

$$2\mathbb{E}\left[\left\|Z_0^N\right\|^2 + (2NC^2 + \sum_{i=1}^{d}\left\|b_i^N\right\|^2)\left\|x\right\|_{1-var,[0,1]}^2\right]\prod_{i=1}^{d}\mathbb{E}\left[\exp\left\{2(1+\sqrt{2}C\left\|x\right\|_{1-var,[0,1]})\left\|A_i^N\right\|_{op}\right\}\right]$$

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 $4\sigma_A^2$ [40, 62] with fluctuations around this limit having exponential tails [59, 78]. 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

$$\mathbb{E}[\left\|Z_0^N\right\|^2] = \mathbb{E}[\sum_{i=1}^{N}|[Z_0^N]_i|^2] = \sigma_a^2 N$$

and similarly $\mathbb{E}[\left\|b_i^N\right\|^2] \leq \sigma_{b_i}^2 N$.

Hence

$$\mathbb{E}[\sup_{t\in[0,1]} \left\|\tilde{Z}_t^{N,M}(x)\right\|^2] \leq$$

$$2N\left(\sigma_a^2 + (2C^2 + \sum_{i=1}^{d}\sigma_{b_i}^2)\left\|x\right\|_{1-var,[0,1]}^2\right)\mathbb{E}\left[\exp\left\{2(1+\sqrt{2}C\left\|x\right\|_{1-var,[0,1]})\left\|A_1^N\right\|_{op}\right\}\right]^d$$

$$=: NK_x$$

□

Proposition 4.27. *Assume the activation function φ is Lipschitz and linearly bounded.*

$$\mathbb{E}\left[\left\|Z_t^{N,M}(x) - \tilde{Z}_t^{N,M}(x)\right\|_{\infty,[0,1]}^2\right] \leq N|\mathcal{D}_M|\tilde{K}_x$$

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 section 4.2.6,

$$\left\|Z_t^{N,M}(x) - \tilde{Z}_t^{N,M}(x)\right\| = \left\|\sum_{i=1}^{d}\int_{t_m}^{t}\left(A_i^N\varphi(Z_{t_m}^{N,M}(x)) + b_i^N\right)dx_r^i\right\| \leq$$

$$\sum_{i=1}^{d}\int_{t_m}^{t}\left(\left\|A_i^N\right\|_{op}\left\|\varphi(Z_{t_m}^{N,M}(x))\right\| + \left\|b_i^N\right\|\right)|dx_r^i| \leq$$

$$\sum_{i=1}^{d}\int_{t_m}^{t}\left(\left\|A_i^N\right\|_{op}\sqrt{2}C(\sqrt{N} + \left\|Z_{t_m}^{N,M}(x)\right\|) + \left\|b_i^N\right\|\right)|dx_r^i| \leq$$

$$\left[\sum_{i=1}^{d}\left\|A_i^N\right\|_{op}\sqrt{2}C(\sqrt{N} + \left\|\tilde{Z}_{t_m}^{N,M}(x)\right\|) + \left\|b_i^N\right\|\right]\|x\|_{1-var,[t_m,t]} \leq$$

$$\left[\sum_{i=1}^{d}\left\|A_i^N\right\|_{op}\sqrt{2}C(\sqrt{N} + \left\|\tilde{Z}_{t_m}^{N,M}(x)\right\|) + \left\|b_i^N\right\|\right]\|x\|_{\mathbb{X}}\sqrt{|\mathcal{D}_M|}$$

where we have used

$$\|x\|_{1-var,[s,t]} = \int_s^t |\dot{x}_r|dr \leq \sqrt{\int_0^1 |\dot{x}_r|\mathbb{I}(r \in |t-s|)dr} \leq \|x\|_{\mathbb{X}}\sqrt{t-s}$$

thus

$$\mathbb{E}\left[\sup_{t\in[0,1]}\left\|Z_t^{N,M}(x) - \tilde{Z}_t^{N,M}(x)\right\|^2\right] \leq$$

$$2\mathbb{E}\left[\sup_{t_m\in\mathcal{D}_M}\left[\sum_{i=1}^{d}\left\|A_i^N\right\|_{op}^2 2C^2(N + \left\|\tilde{Z}_{t_m}^{N,M}(x)\right\|^2) + \left\|b_i^N\right\|^2\right]\|x\|_{\mathbb{X}}^2|\mathcal{D}_M|\right]$$

$$\leq N\bar{K}_x\|x\|_{\mathbb{X}}^2|\mathcal{D}_M|$$

□

Proposition 4.28. *Assume the activation function φ is Lipschitz and linearly bounded.*

$$\mathbb{E}\left[\left\|Z_t^{N,M}(x) - Z_t^{N,M'}(x)\right\|_{\infty,[0,1]}^2\right] \leq N\tilde{K}_x(|\mathcal{D}_M| \vee |\mathcal{D}_{M'}|)$$

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

$$\left\|Z_t^{N,M}(x) - Z_t^{N,M'}(x)\right\| \leq \sum_{i=1}^{d}\int_0^t\left\|A_i^N\left(\varphi(\tilde{Z}_t^{N,M}(x)) - \varphi(\tilde{Z}_t^{N,M'}(x))\right)\right\||dx_r^i| \leq$$

$$\sum_{i=1}^{d}\int_0^t\left(\left\|A_i^N\right\|_{op}\left\|\varphi(\tilde{Z}_t^{N,M}(x)) - \varphi(\tilde{Z}_t^{N,M'}(x))\right\|\right)|dx_r^i| \leq$$

$$\sum_{i=1}^{d}\int_0^t\left\|A_i^N\right\|_{op}K\left\|\tilde{Z}_t^{N,M}(x) - \tilde{Z}_t^{N,M'}(x)\right\||dx_r^i| =$$

$$K\sum_{i=1}^{d}\left\|A_i^N\right\|_{op}\int_0^t\left\|\tilde{Z}_t^{N,M}(x) - \tilde{Z}_t^{N,M'}(x)\right\||dx_r^i|$$

Using triangle inequality we obtain

$$\left\|Z_t^{N,M}(x) - Z_t^{N,M'}(x)\right\| \leq$$
$$K\sum_{i=1}^{d}\left\|A_i^N\right\|_{op}\int_0^t\left\|\tilde{Z}_t^{N,M}(x) - Z_t^{N,M}(x)|\right\|dx_r^i| +$$
$$K\sum_{i=1}^{d}\left\|A_i^N\right\|_{op}\int_0^t\left\|Z_t^{N,M}(x) - Z_t^{N,M'}(x)\right\||dx_r^i| +$$
$$K\sum_{i=1}^{d}\left\|A_i^N\right\|_{op}\int_0^t\left\|Z_t^{N,M'}(x) - \tilde{Z}_t^{N,M'}(x)\right\||dx_r^i|$$

Note how

$$K\sum_{i=1}^{d}\left\|A_i^N\right\|_{op}\int_0^t\left\|\tilde{Z}_t^{N,M}(x) - Z_t^{N,M}(x)\right\||dx_r^i| \leq$$
$$K\sum_{i=1}^{d}\left\|A_i^N\right\|_{op}\int_0^t\Big[\sum_{k=1}^{d}\left\|A_k^N\right\|_{op}\sqrt{2}C(\sqrt{N} + \left\|\tilde{Z}_{t_m}^{N,M}(x)\right\|) + \left\|b_k^N\right\|\Big]\|x\|_{\mathbb{X}}\sqrt{|\mathcal{D}_M|}|dx_r^i| \leq$$
$$K(\sum_{i=1}^{d}\left\|A_i^N\right\|_{op})\Big[\sum_{i=k}^{d}\left\|A_k^N\right\|_{op}\sqrt{2}C(\sqrt{N} + \Gamma_x) + \left\|b_k^N\right\|\Big]\|x\|_{1-var}\|x\|_{\mathbb{X}}\sqrt{|\mathcal{D}_M|} =: \Lambda_x\sqrt{|\mathcal{D}_M|}$$

with

$$\Gamma_x = \left[\left\|Z_0^N\right\| + (\sqrt{2N}C + \sum_{i=1}^{d}\left\|b_i^N\right\|)\|x\|_{1-var,[0,1]}\right]\prod_{i=1}^{d}\exp\left\{(1+\sqrt{2}C\|x\|_{1-var,[0,1]})\left\|A_i^N\right\|_{op}\right\}$$

Hence it follows that

$$\left\|Z_t^{N,M}(x) - Z_t^{N,M'}(x)\right\| \leq$$
$$(\sqrt{|\mathcal{D}_M|} + \sqrt{|\mathcal{D}_{M'}|})\Lambda_x + K\sum_{i=1}^{d}\left\|A_i^N\right\|_{op}\int_0^t\left\|Z_t^{N,M}(x) - Z_t^{N,M'}(x)\right\||dx_r^i|$$

and with Gronwall

$$\left\|Z_t^{N,M}(x) - Z_t^{N,M'}(x)\right\|_\infty \leq 2\sqrt{|\mathcal{D}_M| \vee |\mathcal{D}_{M'}|}\Lambda_x\exp\{K\Big(\sum_{i=1}^{d}\left\|A_i^N\right\|_{op}\Big)\|x\|_{1-var}\}$$

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 4.29. *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 Z_0, A_k, b_k for all partitions. Let $Z^N(x)$ be the solution of the following Neural CDE*

$$Z_t^N(x) = Z_0^N + \sum_{k=1}^{d}\int_0^t\Big(A_k\varphi(Z_s^N(x)) + b_k\Big)dx_s^k \tag{4.18}$$

Then there exist a constant $\mathbf{K}_x$ *independent of* N, M *and increasing function of* $\|x\|_{\mathbb{X}}$ *such that*

$$\mathbb{E}[\sup_{t\in[0,1]} \left\| Z_t^{M,N}(x) - Z_t^N(x) \right\|_{\mathbb{R}^N}^2] \leq N|\mathcal{D}_M|\mathbf{K}_x \tag{4.19}$$

Proof. Note first that $Z^N(x)$ is well defined since the system has unique solution by Friz and Victoir [37][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 $Z_t^{N,M}(x)$ is a cauchy sequence in $C^0([0,1];\mathbb{R}^N)$ hence for any $M > 0$, eventually in M' it holds

$$\sup_{t\in[0,1]} \left\| Z_t^{M,N}(x) - Z_t^N(x) \right\|_{\mathbb{R}^N} \leq 2 \sup_{t\in[0,1]} \left\| Z_t^{M,N}(x) - Z_t^{M',N}(x) \right\|_{\mathbb{R}^N}$$

Then theorem 4.28 gives the sought after bound.

□

Theorem 4.30. *Assume the activation function* φ *is Lipschitz and linearly bounded. There is a constant* C *depending only on* $\|x\|_{\mathbb{X}}$ *in an increasing way such that:*

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

where $\mu_t^{M,N}$ *is the distribution of* $\langle v^N, Z_t^{M,N}(x)\rangle$ *and* $\mu_t^N(x)$ *that of* $\langle v^N, Z_t^N(x)\rangle$ *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

$$\mathbb{E}[|G(\langle v^N, Z_t^{M,N}(x)\rangle) - G(\langle v^N, Z_t^N(x)\rangle)|] \leq \mathbb{E}\left[|\langle v^N, Z_t^{M,N}(x)\rangle - \langle v^N, Z_t^N(x)\rangle|\right]$$

which leads to

$$\begin{aligned}
\mathbb{E}\left[|\langle v^N, Z_t^{M,N}(x)\rangle - \langle v^N, Z_t^N(x)\rangle|\right] &= \mathbb{E}\left[|\langle v^N, Z_t^{M,N}(x) - Z_t^N(x)\rangle|\right] \\
&\leq \left(\mathbb{E}\left[|\langle v^N, Z_t^{M,N}(x) - Z_t^N(x)\rangle|^2\right]\right)^{\frac{1}{2}} \\
&= \left(\frac{1}{N}\mathbb{E}\left[\left\| Z_t^{M,N}(x) - Z_t^N(x)\right\|^2\right]\right)^{\frac{1}{2}} \\
&\leq \left(\frac{1}{N}\mathbb{E}\left[\sup_{t\in[0,1]}\left\| Z_t^N(x) - Z_t^{M,N}(x)\right\|^2\right]\right)^{\frac{1}{2}} \\
&\leq \sqrt{\mathbf{K}_x|\mathcal{D}_M|}
\end{aligned}$$

hence

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

□

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

Corollary (theorem 4.2)**.** *The limits in theorem 4.16 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. □

4.4 Numerics

We start by illustrating the convergence in distribution of a homogeneous controlled ResNet to a GP endowed with neural signature kernel as per Theorem 4.2. To this aim, we consider a homogeneous ResNet $\Phi_\varphi^{M,N}$ with activation function $\varphi = \text{ReLU}$, 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 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 eq. (4.3) on a fine discretization grid. As it can be observed on Figure 4.2 the Gaussian fit for this one-dimensional marginal gets better as N increases.

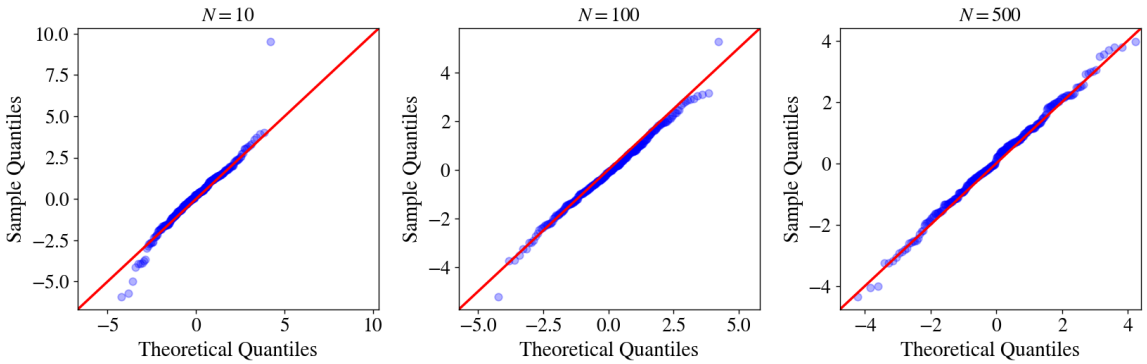


Figure 4.2: *Empirical quantiles against the theoretical quantiles of $\mathcal{N}(0, \mathcal{K}_\varphi^{x,x}(1))$ for $N \in \{10, 100, 500\}$ with $\varphi = ReLU$*

We then provide empirical evidence for the order of convergence provided in theorem 3.5, corresponding to the case $\varphi = id$. Here, we consider a linear homogeneous

ResNet, controlled by x and y, two sample paths from a zero-mean GP with RBF kernel $r_{\text{RBF}}(s,t) = \exp(-5(s-t)^2)$ 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 $\mathbb{E}[(\frac{1}{N}\langle Z_1^N(x), Z_1^N(y)\rangle - \mathcal{K}_{id}(x,y))^2]$ increasing the width N. Our empirical results, as displayed on Figure 4.3, align with the theoretical convergence rate.

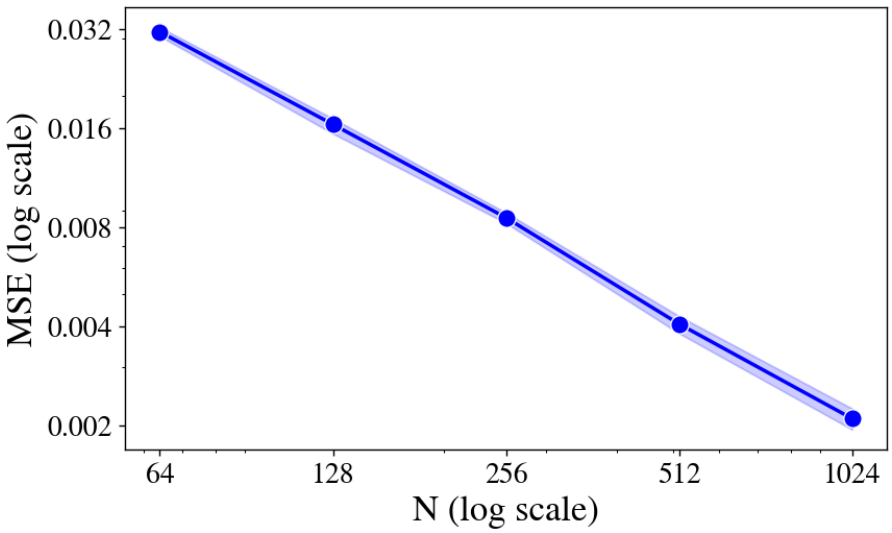


Figure 4.3: *Mean squared error of* $\frac{1}{N}\langle Z_1^N(x), Z_1^N(y)\rangle_{\mathbb{R}^N}$*, the estimator of* $\mathcal{K}_{id}^{x,y}(1,1)$*, as a function of the width* N *on a logarithmic scale. Standard deviations were obtained by repeating the experiment* 5 *times.*

Appendix A

Genus Expansion for non-Gaussian, sparse and complex weights.

A.1 Non-Gaussian Weights

All of the results obtained in the chapter 2 can be extended to neural networks Φ with non-Gaussian random weights, under moment assumptions. Consider the following more general sequence $\mathcal{W}$.

Definition A.1. *Let $N > 0$ and $\mathcal{W} := (W_i \mid i \in \mathbb{N})$, where the $W_i \in \mathbb{R}^{N\times N}$ are independent matrices with entries $[W_i]_{j,k} \sim \sigma_i Z_i$, where $\{Z_i\}_{i\geq 0}$ is a family of i.i.d. random variables satisfying*

$$\mathbb{E}\{Z_i^2\} = 1, \quad \mathbb{E}\{Z_i^{2k+1}\} = 0, \quad \mathbb{E}\{Z_i^{2k}\} < \infty \quad \forall k \geq 0.$$

Note that we take the matrices in this sequence to be square for simplicity only, the results in this section carrying over to rectangular matrices as in chapter 1 by straightforward modifications. The graphs considered in this section are all assumed to satisfy the following condition.

Assumption A.2. *G is a product graph with inputs which are either elements of $\mathcal{W}$, or deterministic vector/matrices with entries uniformly bounded in N.*

We define $E_{\mathcal{W}}$ and the set of admissible pairings of $E_{\mathcal{W}}$ in exactly the same way as in the Gaussian case. Recall also that for any G, $\check{e}(G) := |E(G_{\phi_0})|$ for any fixed admissible pairing ϕ_0 of its edges (the choice of ϕ_0 is irrelevant, as the value of $|E(G_{\phi_0})|$ is constant over pairings).

Theorem A.3 (Approximate Wick expansion)**.** *For any (possibly disconnected) product graph G*

$$\mathbb{E}\{\mathbf{W}_G\} = \sum_{\phi\in\mathcal{P}(G)} \mathbf{W}_{G_\phi} + \sigma_G \mathcal{O}(N^{\check{e}(G)+c(G)-1}) \tag{A.1}$$

where ϕ runs over all of G's admissible *pairings.*

Proof. Consider the set of indexations $\mathcal{I}^G := \prod_{v\in V(G)}\{1,\dots,N_v\}$. Recall that for each $e=(u,v)\in E$, $i_e = i_u i_v$. Then

$$\mathbf{W}_G = \sum_{\mathbf{i}\in\mathcal{I}^G}\prod_{c\in C}[\mathbf{X}_c]_{i_c}, \quad \mathbb{E}\{\mathbf{W}_G\} = \sum_{\mathbf{i}}\left(\prod_{c\in C\setminus E_{\mathcal{W}}}[\mathbf{X}_c]_{i_c}\right)\mathbb{E}\left\{\prod_{e\in E_{\mathcal{W}}}[\mathbf{X}_e]_{i_e}\right\}$$

Following [28] we write $\mathbf{i}\rightsquigarrow\phi$ whenever

$$\forall\{e,e'\}\in\phi,\ i_e = i_{e'}$$

and say that an indexation $\mathbf{i}$ is *balanced* whenever there exists a $\phi\in\mathcal{P}(G)$ for which $\mathbf{i}\rightsquigarrow\phi$. Note how for *unbalanced* $\mathbf{i}$, due to the vanishing of odd moments, one has $\mathbb{E}\left\{\prod_{e\in E_{\mathcal{W}}}[\mathbf{X}_e]_{i_e}\right\}=0$ so that we can write the expectation as

$$\mathbb{E}\{\mathbf{W}_G\} = \sum_{\phi\in\mathcal{P}(G)}\sum_{\mathbf{i}\rightsquigarrow\phi}\frac{1}{|\{\phi\in\mathcal{P}(G)\mid\mathbf{i}\rightsquigarrow\phi\}|}\left(\prod_{c\in C\setminus E_{\mathcal{W}}}[\mathbf{X}_c]_{i_c}\right)\mathbb{E}\left\{\prod_{e\in E_{\mathcal{W}}}[\mathbf{X}_e]_{i_e}\right\}$$

For simplicity write $\Gamma_\phi := \sum_{\mathbf{i}\rightsquigarrow\phi}\frac{1}{|\{\phi\in\mathcal{P}(G)\mid\mathbf{i}\rightsquigarrow\phi\}|}\left(\prod_{c\in C\setminus E_{\mathcal{W}}}[\mathbf{X}_c]_{i_c}\right)\mathbb{E}\left\{\prod_{e\in E_{\mathcal{W}}}[\mathbf{X}_e]_{i_e}\right\}$ so that

$$\mathbb{E}\{\mathbf{W}_G\} = \sum_{\phi\in\mathcal{P}(G)}\Gamma_\phi.$$

Note that for all $\phi\in\mathcal{P}(G)$ it holds, for constants depending on G but not on N, that

$$\Gamma_\phi = \sum_{\mathbf{i}\rightsquigarrow\phi}\sigma_G\mathcal{O}(1) = \sigma_G\mathcal{O}(|\{\mathbf{i}\mid\mathbf{i}\rightsquigarrow\phi\}|) = \sigma_G\mathcal{O}(N^{|V(G_\phi)|})$$

Notice that the maximum value of $|V(G_\phi)|$ is reached only when ϕ is fully atomic, in which case one has

$$|V(G_\phi)| = \check{e}(G) + c(G)$$

where we recall that $\check{e} = |E(G_\phi)|$ (otherwise, this is a strict inequality).

Furthermore, for such ϕ, we have

$$\Gamma_\phi = \sum_{\mathcal{I}^G_{(1)}\ni\mathbf{i}\rightsquigarrow\phi}\left(\prod_{c\in C\setminus E_{\mathcal{W}}}[\mathbf{X}_c]_{i_c}\right)\mathbb{E}\left\{\prod_{e\in E_{\mathcal{W}}}[\mathbf{X}_e]_{i_e}\right\} + \sigma_G\mathcal{O}(N^{\check{e}(G)+c(G)-1}) \tag{A.2}$$

where

$$\mathcal{I}^G_{(1)} := \{\mathbf{i}\in\mathcal{I}^G\mid\forall e,e'\in E_{\mathcal{W}}, i_e = i_{e'}\Rightarrow\mathbf{X}_e=\mathbf{X}_{e'}\text{ and }i_e\neq i_{e_0}\quad\forall e_0\in E_{\mathcal{W}}\setminus\{e,e'\}\}, \tag{A.3}$$

noting that if $\mathbf{i} \in \mathcal{I}^G_{(1)}$, there exists exactly one $\phi \in \mathcal{P}(G)$ for which $\mathbf{i} \rightsquigarrow \phi$.

But in this case, we also have

$$\mathbf{W}_{G_\phi} = \sum_{\mathcal{I}^G_{(1)} \ni \mathbf{i} \rightsquigarrow \phi} \left(\prod_{c \in C \setminus E_{\mathcal{W}}} [\mathbf{X}_c]_{i_c} \right) \mathbb{E} \left\{ \prod_{e \in E_{\mathcal{W}}} [\mathbf{X}_e]_{i_e} \right\} + \sigma_G \mathcal{O}(N^{\check{e}(G)+c(G)-1}),$$

so that we can write

$$\begin{aligned}
\mathbb{E}\{\mathbf{W}_G\} &= \sum_{\phi \in \mathcal{P}_{\mathrm{A}}(G)} \Gamma_\phi + \sum_{\phi \in \mathcal{P}(G) \setminus \mathcal{P}_{\mathrm{A}}(G)} \Gamma_\phi \\
&= \sum_{\phi \in \mathcal{P}_{\mathrm{A}}(G)} \Gamma_\phi + \sigma_G \mathcal{O}(N^{\check{e}(G)+c(G)-1}) \\
&= \sum_{\phi \in \mathcal{P}_{\mathrm{A}}(G)} \left(\mathbf{W}_{G_\phi} + \sigma_G \mathcal{O}(N^{\check{e}(G)+c(G)-1}) \right) + \sigma_G \mathcal{O}(N^{\check{e}(G)+c(G)-1}) \\
&= \sum_{\phi \in \mathcal{P}_{\mathrm{A}}(G)} \mathbf{W}_{G_\phi} + \sigma_G \mathcal{O}(N^{\check{e}(G)+c(G)-1}) \\
&= \sum_{\phi \in \mathcal{P}(G)} \mathbf{W}_{G_\phi} + \sigma_G \mathcal{O}(N^{\check{e}(G)+c(G)-1})
\end{aligned}$$

where in the last equality we have used the fact that

$$\forall \phi \in \mathcal{P}(G) \setminus \mathcal{P}_{\mathrm{A}}(G),\ \mathbf{W}_{G_\phi} = \sigma_G \mathcal{O}(N^{\check{e}(G)+c(G_\phi)-1}) = \sigma_G \mathcal{O}(N^{\check{e}(G)+c(G)-1}).$$

□

Proposition A.4. *Let $G = \sqcup_{i=1}^{c(G)} G_i$ be a product graph with disjoint connected components G_i. Then*

$$\begin{aligned}
\mathbb{E}\left\{ \prod_{i=1}^{c(G)} \left(\mathbf{W}_{G_i} - \sum_{\psi \in \mathcal{P}_{\mathrm{A}}(G_i)} \mathbf{W}_{(G_i)_\psi} + \sigma_{G_i} \mathcal{O}(N^{\check{e}(G_i)}) \right) \right\} \\
= \sum_{\phi \in \mathcal{P}_{\mathrm{B}}(G)} \mathbf{W}_{G_\phi} + \sigma_G \mathcal{O}(N^{\check{e}(G)+c(G)/2-1}).
\end{aligned}$$

Proof. For any $\phi \in \mathcal{P}_{\mathrm{A}}(G)$, let $\Gamma_\phi = \mathbf{W}_{G_\phi} + \sigma_G \mathcal{O}(N^{\check{e}(G)+c(G_\phi)-1})$ and define Γ_{ϕ_i} similarly for each $\phi_i \in \mathcal{P}_{\mathrm{A}}(G_i)$. Proceeding as in theorem 1.17 we obtain

$$\mathbb{E}\left\{ \prod_{i=1}^{c(G)} \left(\mathbf{W}_{G_i} - \sum_{\phi \in \mathcal{P}(G_i)} \Gamma_\phi \right) \right\} = \sum_{T \subseteq [c(G)]} (-1)^{|T^c|} \sum_{\phi_T \times_{i \in T^c} \phi_i} \left(\Gamma_{\phi_T} \prod_{i \in T^c} \Gamma_{\phi_i} \right)$$

where the last sum is over $\mathcal{P}(G_T) \times (\prod_{i \in T^c} \mathcal{P}_{\mathrm{A}}(G_i))$.

Arguing as in the proof of the previous lemma, we find that

$$\Gamma_{\phi_T} \prod_{i \in T^c} \Gamma_{\phi_i} = \Gamma_{\phi_T \times_{i \in T^c} \phi_i} + \mathcal{O}(N^{|V(G_{\phi_T \times_{i \in T^c} \phi_i})|-1}) \tag{A.4}$$

where the implicit constants do not depend from N.

But then using inclusion exclusion, we obtain on the one hand

$$\sum_{T\subseteq[c(G)]}(-1)^{|T^c|}\sum_{\phi_T\times_{i\in T^c}\phi_i}\Gamma_{\phi_T\times_{i\in T^c}\phi_i}=\sum_{\phi\in\tilde{\mathcal{P}}(G)}\Gamma_\phi$$

where $\tilde{\mathcal{P}}(G)\subseteq\mathcal{P}(G)$ is the set of pairings of G such that none of its connected components are paired within themselves, and on the other hand,

$$\sum_{T\subseteq[c(G)]}(-1)^{|T^c|}\sum_{\phi_T\times_{i\in T^c}\phi_i}\mathcal{O}(N^{|V(G_{\phi_T\times_{i\in T^c}\phi_i})|-1})=\mathcal{O}\Big(\sum_{\phi\in\tilde{\mathcal{P}}(G)}N^{|V(G_\phi)|-1}\Big). \tag{A.5}$$

We conclude by arguing as before that the maximal order of the Γ_ϕ is reached when the partition is bi-atomic, which is the only case when $V(G_\phi)=\check{e}(G)+\frac{c(G)}{2}$ and when

$$\Gamma_\phi=\mathbf{W}_{G_\phi}+\mathcal{O}\Big(N^{\check{e}(G)+c(G)/2-1}\Big).$$

□

Reiterating the arguments in chapter 2 then yields the following.

Corollary. *Let Φ be a neural network with weights W_ℓ with entries as in theorem A.1. Then the conclusions of theorems 2.9, 2.14 and 2.19 still hold.*

A.2 Sparse Weights

Following [28], we can also go further and consider sparse matrices with i.i.d. entries satisfying the same moment assumptions as in theorem A.1.

Consider the sequence $\tilde{\mathcal{W}}$ of matrices $\tilde{W}_i:=W_i\odot B_i$, where $W_i\in\mathcal{W}$ as defined above and the B_i are independent matrices with i.i.d., Bernoulli distributed entries with parameter p_N satisfying $Np_N\to\infty$. Note that

$$\mathbb{E}\{[\tilde{W}_i]_{\alpha,\beta}^{2k}\}=p_N\sigma_i^{2k}\mathbb{E}\{Z_i^{2k}\},\quad \mathbb{E}\{[\tilde{W}_i]_{\alpha,\beta}^{2k+1}\}=0,\quad \forall k\geq 0.$$

If G is a product graph whose random inputs are from $\tilde{\mathcal{W}}$, we will use $\tilde{\mathbf{X}}_e=\mathbf{X}_e\odot B_i$ to denote such inputs.

Assumption A.5. *G is a product graph with inputs which are either elements of $\tilde{\mathcal{W}}$, or deterministic vector/matrices with entries uniformly bounded in N.*

Proposition A.6. *Let G be a product graph satisfying theorem A.5. Then*

$$\mathbb{E}\{\mathbf{W}_G\}=\sum_{\phi\in\mathcal{P}_{\mathrm{A}}(G)}\mathbf{W}_{G_\phi}+\sigma_G p_N^{|E_{\mathcal{W}}|/2-1}\mathcal{O}(N^{\check{e}(G)+c(G)-1}) \tag{A.6}$$

Proof. Consider the set of indexations $\mathcal{I}^G := \prod_{v\in V(G)}\{1,\dots,N_v\}$, and define $\mathcal{I}^G_{(1)}$, balanced and unbalanced indexations as in the proof of theorem A.3.

If $\mathbf{i}\in\mathcal{I}^G$ is unbalanced, one has $\mathbb{E}\left\{\prod_{e\in E_{\mathcal{W}}}[\tilde{\mathbf{X}}_e]_{i_e}\right\}=0$ while if $\mathbf{i}\in\mathcal{I}^G_{(1)}$,

$$\mathbb{E}\left\{\prod_{e\in E_{\mathcal{W}}}[\tilde{\mathbf{X}}_e]_{i_e}\right\}=\sigma_G p_N^{|E_{\mathcal{W}}|/2}.$$

If $\mathbf{i}\notin\mathcal{I}^G_{(1)}$ is balanced, then

$$\mathbb{E}\left\{\prod_{e\in E_{\mathcal{W}}}[\tilde{\mathbf{X}}_e]_{i_e}\right\}=\sigma_G p_N^{l}\mathcal{O}(1)$$

for some $1\leq l<|E_{\mathcal{W}}|/2$. We can thus partition the set of balanced indexations as $\sqcup_{l=1}^{|E_{\mathcal{W}}|/2}\mathcal{J}^G_l$ where $\mathcal{J}^G_l$ is the subset of indexations which result in a factor p_N^l. Noting that $\mathcal{J}^G_{|E_{\mathcal{W}}|/2}=\mathcal{I}^G_{(1)}$, we can write

$$\begin{aligned}\mathbb{E}\{\mathbf{W}_G\}=&\sigma_G p_N^{|E_{\mathcal{W}}|/2}\sum_{\mathbf{i}\in\mathcal{I}^G_{(1)}}\left(\prod_{c\in C\setminus E_{\mathcal{W}}}[\mathbf{X}_c]_{i_c}\right)+\sum_{l=1}^{|E_{\mathcal{W}}|/2-1}\sum_{\mathbf{i}\in\mathcal{J}^G_l}\left(\prod_{c\in C\setminus E_{\mathcal{W}}}[\mathbf{X}_c]_{i_c}\right)\sigma_G p_N^l\mathcal{O}(1)\\ =&\sigma_G p_N^{|E_{\mathcal{W}}|/2}\sum_{\mathbf{i}\in\mathcal{I}^G_{(1)}}\left(\prod_{c\in C\setminus E_{\mathcal{W}}}[\mathbf{X}_c]_{i_c}\right)+\sum_{l=1}^{|E_{\mathcal{W}}|/2-1}\sigma_G p_N^l\mathcal{O}\Big(|\mathcal{J}^G_l|\Big)\end{aligned}$$

Note also that $|\mathcal{J}^G_l|=\mathcal{O}\Big(N^{|E\setminus E_{\mathcal{W}}|+c(G)+l}\Big)$.

In fact, each $\mathbf{i}$ is compatible to some $\phi\in\mathcal{P}(G)$ by virtue of being balanced, thus any $\mathbf{i}$ takes at most $c(G)+|E\setminus E_{\mathcal{W}}|+\frac{|E_{\mathcal{W}}|}{2}$ different values, this being the maximal number of vertices in any G_ϕ. Note though that, by definition of $\mathcal{J}^G_l$, the indices corresponding to the $|E_{\mathcal{W}}|/2$ pairs take exactly l different values. Each $\mathbf{i}\in\mathcal{J}^G_l$ thus takes at most $c(G)+|E\setminus E_{\mathcal{W}}|+l$ values, ranging from 1 to N.

Hence we can write

$$\begin{aligned}\sum_{l=1}^{|E_{\mathcal{W}}|/2-1}\sigma_G p_N^l\mathcal{O}\Big(|\mathcal{J}^G_l|\Big)&=\sigma_G\mathcal{O}\Big(N^{|E\setminus E_{\mathcal{W}}|+c(G)}\max\{Np_N,(Np_N)^{|E_{\mathcal{W}}|/2-1}\}\Big)\\ &=\sigma_G p_N^{|E_{\mathcal{W}}|/2-1}\mathcal{O}(N^{\check{e}(G)+c(G)-1}).\end{aligned}$$

The claim then follows from the fact that

$$\begin{aligned}\sum_{\phi\in\mathcal{P}_{\mathrm{A}}(G)}\mathbf{W}_{G_\phi}&=\sigma_G p_N^{|E_{\mathcal{W}}|/2}\left(\sum_{\mathbf{i}\in\mathcal{I}^G_{(1)}}\left(\prod_{c\in C\setminus E_{\mathcal{W}}}[\mathbf{X}_c]_{i_c}\right)+\mathcal{O}(N^{\check{e}(G)+c(G)-1})\right)\\ &=\sigma_G p_N^{|E_{\mathcal{W}}|/2}\sum_{\mathbf{i}\in\mathcal{I}^G_{(1)}}\left(\prod_{c\in C\setminus E_{\mathcal{W}}}[\mathbf{X}_c]_{i_c}\right)+p_N^{|E_{\mathcal{W}}|/2}\mathcal{O}(N^{\check{e}(G)+c(G)-1}),\end{aligned}$$

and that $p_N^{|E_{\mathcal{W}}|/2}\mathcal{O}(N^{\check{e}(G)+c(G)-1}) = p_N^{|E_{\mathcal{W}}|/2-1}\mathcal{O}(N^{\check{e}(G)+c(G)-1})$ (since $p_N \leq 1$). □

Similar arguments prove the centered version of this result, giving

$$\begin{aligned}\mathbb{E}\Big\{\prod_{i=1}^{c(G)}\Big(\mathbf{W}_{G_i} - \sum_{\psi\in\mathcal{P}_{\mathrm{A}}(G_i)}\mathbf{W}_{(G_i)_\psi} + \sigma_{G_i}p_N^{|E_{\mathcal{W}}(G_i)|/2-1}\mathcal{O}(N^{\check{e}(G_i)})\Big)\Big\} \\ = \sum_{\phi\in\mathcal{P}_{\mathrm{B}}(G)}\mathbf{W}_{G_\phi} + \sigma_G p_N^{|E_{\mathcal{W}}(G)|/2-1}\mathcal{O}(N^{\check{e}(G)+c(G)/2-1}).\end{aligned} \tag{A.7}$$

We may then conclude the following.

Corollary. *Let Φ be a neural network with weights $\tilde{W}_\ell$ (as defined above) and parameters. Then after rescaling standard deviations as $\sigma_\ell \mapsto \frac{1}{\sqrt{p_N}}\sigma_\ell$, the conclusions of theorems 2.9, 2.14 and 2.19 still hold.*

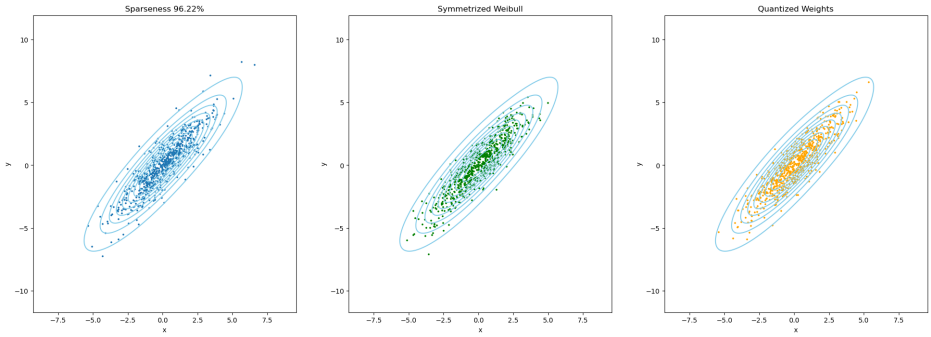


Figure A.1: *We consider a ReLU network Φ_2 (meaning that the activations are all $x \mapsto \max(x, 0)$), where $(N_0, N_1, N_2, N_3) = (2, 700, 700, 1)$. We fix two inputs $\mathbf{x} = (\pi, 1), \mathbf{y} = (e, 3) \in \mathbb{R}^2$, and sample the weight matrices W_0, W_1 and W_2 700 times from 4 different ensembles and plot $(\Phi_2(\mathbf{x}), \Phi_2(\mathbf{y}))$. The contour plots in light blue in all three figures serve as a reference, and correspond to W_0, W_1, W_2 being Gaussian. The leftmost scatter plot (dark blue) shows the result of multiplying W_1 and W_2 entry-wise by $p_N^{-1/2}Ber(p_N)$ for $p_N = \frac{1}{\sqrt{300}}$ (so that roughly 96.22% of their entries are set to 0). In the middle (green), the W_i have Weibull entries multiplied by a random sign, while on the right (orange) the entries are uniformly distributed in $\{-1, 0, 1\}$.*

A.3 Complex Weights

We end this section by showing how all of our results (including the extensions to sparse matrices in the previous section) hold for complex-valued matrices. Noting that the definition of product graphs and the extensions in chapter 2 hold *mutatis mutandis* in the complex case, we will once again only need to adapt the arguments of chapter 1.

Definition A.7. *Let $\mathcal{W}_{\mathbb{C}} := (W_i \mid i \in \mathbb{Z}\setminus\{0\})$, where for $i > 0$, the $W_i \in \mathbb{C}^{N_{r_i}\times N_{c_i}}$ are independent matrices with i.i.d. entries equal to $\sigma_i Z$ where $Z \sim \mathcal{N}_{\mathbb{C}}(0,1)$. For $i < 0$, we set $W_{-i} := \overline{W}_i$.*

Assumption A.8. *G is a product graph whose inputs are either deterministic, or matrices in $\mathcal{W}_{\mathbb{C}}$.*

Recall how a complex random variable $Z \sim \mathcal{N}_{\mathbb{C}}(0,1)$ satisfies the following moment assumptions

$$\mathbb{E}[|Z|] = 1, \quad n \neq m \Rightarrow \mathbb{E}[Z^n \bar{Z}^m] = 0$$

where $\bar{Z}$ denotes the conjugate of Z.

This means that, in view of Wick's theorem (which takes the same form as in the real case), the correct notion of admissible pairing is given as follows.

Definition A.9 (Complex admissible pairing)**.** *Let G be a product graph, and $E_{\mathcal{W}} = \{e \in E : W_e = W_\ell$ for some $\ell(e) := \ell \in \mathbb{Z}\setminus\{0\}\}$. Then a pairing ϕ of $E_{\mathcal{W}}$ is said to be* admissible *if any two paired edges $\{e, e'\} \in \phi$ satisfy $\ell(e) = -\ell(e')$. We denote the set of all admissible pairings of $E_{\mathcal{W}}$ by $\mathcal{P}(G)$.*

At this point notice that the arguments of chapter 1 follow verbatim, with the exception of theorem 1.22 which now takes the following form.

Theorem A.10. *Let $(G_i)_{i\geq 1}$ be a sequence of connected product graphs satisfying theorem A.8 and theorem 1.19 (for $\mathcal{W}_{\mathbb{C}}$). Then for any $m \geq 1$,*

$$\Big((\sigma_{G_i} N^{\check{e}(G_i)+\frac{1}{2}}\Big)^{-1}\mathbf{W}_{G_i} - |\mathcal{P}_{\mathrm{A}}(G_i)|\sqrt{N})_{i=1}^m \xrightarrow{d} \mathbf{Z} = \Big(Z_1, \dots, Z_m\Big)^\top$$

where the limit is a centered complex Gaussian vector with covariance function $C := \mathbb{E}[\mathbf{Z}\mathbf{Z}^\top]$ having entries $[C]_{i,j} = |\mathcal{P}_{\mathrm{A}}(G_i \sqcup G_j)|$ and pseudo covariance $\Gamma := \mathbb{E}[\mathbf{Z}\mathbf{Z}^H]$ with entries $[\Gamma]_{i,j} = |\mathcal{P}_{\mathrm{A}}(G_i \sqcup \bar{G}_j)|$ where $\bar{G}$ is obtained from G by conjugating all the random matrices..

As for non-Gaussian complex-valued matrices, we replace the entries of W_i (for $i > 0$) with $\sigma_i Z_i$, where $\{Z_i\}_{i\geq 0}$ is now a family of i.i.d. random variables satisfying

$$\mathbb{E}[|Z|] = 1, \quad n \neq m \Rightarrow \mathbb{E}[Z^n \bar{Z}^m] = 0, \quad \mathbb{E}[|Z|^k] < \infty.$$

Complex analogues of section A.1 and section A.2 then follow straightforwardly.

Appendix B

Technical Combinatorial Lemmas

The goal of this section is to prove theorem B.4, which is a key ingredient in the proof of theorems 2.6 and 2.9. We begin with the following technical lemma.

Lemma B.1. *Consider a sequence of sets* $(\mathcal{A}_i \mid i \in \mathbb{N}_{>0})$ *and absolutely summable sequences of vectors* $(v_a \mid a \in \sqcup_i \mathcal{A}_i) \subseteq \mathbb{R}^d$ *indexed by them. Let* $v_{(a_1,\dots,a_M)} := \otimes_{i=1}^M v_{a_i}$ *with the convention that empty products have value* 1 *and* $a_i \in \mathcal{A}_i$*. Then for* $M \in \mathbb{N}$*,*

$$(\mathbb{R}^d)^{\otimes M} \ni \otimes_{i=1}^M \left(\sum_{a_i \in \mathcal{A}_i} v_{a_i} \right) = \sum_{\boldsymbol{a}_M \in \times_{i=1}^M \mathcal{A}_i} v_{\boldsymbol{a}_M}. \tag{B.1}$$

Proof. We proceed by induction on M. If $M = 0$ then the result holds by convention. For the inductive case assume the lemma true up to M, then

$$\begin{aligned} \otimes_{i=1}^{M+1} \sum_{a_i \in \mathcal{A}_i} v_{a_i} &= \left(\otimes_{i=1}^{M} \sum_{a_i \in \mathcal{A}_i} v_{a_i} \right) \otimes \left(\sum_{a_{M+1} \in \mathcal{A}_{M+1}} v_{a_{M+1}} \right) \\ &= \left(\sum_{\boldsymbol{a}_M \in \times_{i=1}^M \mathcal{A}_i} v_{\boldsymbol{a}_M} \right) \otimes \left(\sum_{a_{M+1} \in \mathcal{A}_{M+1}} v_{a_{M+1}} \right) \\ &= \sum_{a_{M+1} \in \mathcal{A}_{M+1}} \sum_{\boldsymbol{a}_M \in \times_{i=1}^M \mathcal{A}_i} v_{\boldsymbol{a}_M} \otimes v_{a_{M+1}} \end{aligned}$$

Letting $\boldsymbol{a}_M = (a_1, \dots, a_M)$, we have $v_{\boldsymbol{a}_M} \otimes v_a = v_{(a_1,\dots,a_{M+1})}$ and the claim follows. □

The next step will be to prove an analogous result with Hadamard products $\odot$ instead of tensor products $\otimes$. Given that the Hadamard product is commutative, it will be convenient to write the results in terms of multisets (meaning sets where elements can appear more than once) since they are unordered.

Definition B.2. *Given a set* $\mathcal{A}$ *we define the set of its* symmetric M*-tuples as the quotient* $\mathcal{A}^M / S_M$ *where* a_M *is the symmetric group of* M *elements acting on* $\mathcal{A}^M$ *by*

permuting the elements as

$$\sigma \cdot (a_1, \dots, a_M) \mapsto (a_{\sigma(1)}, \dots, a_{\sigma(M)}). \tag{B.2}$$

We denote this set and its elements as

$$\mathbb{X}_{\mathcal{A}}^{M} := \mathcal{A}^M / S_M = \{[\![a_1 \cdots a_M]\!] : \forall i \in [M].\, a_i \in \mathcal{A}\}, \tag{B.3}$$

We also set $\mathbb{X}_{\mathcal{A}} = \bigcup_{M=0}^{\infty} \mathbb{X}_{\mathcal{A}}^{M}$. *For a* symmetric *tuple* $\boldsymbol{\tau} = [\![(a_1)^{k_1} \cdots (a_N)^{k_N}]\!] \in \mathbb{X}_{\mathcal{A}}$, *where* $a_1, \dots, a_N$ *are distinct elements of* $\mathcal{A}$, *we define the* symmetry factor *of* $\boldsymbol{\tau}$ *as*

$$\mathfrak{s}([\![\ ]\!]) = 1, \quad \mathfrak{s}(\boldsymbol{\tau}) := \prod_{i=1}^{N} (k_i)! \tag{B.4}$$

Informally a *symmetric* tuple $[\![a_1 \cdots a_M]\!]$ is just the tuple $(a_1, \dots, a_M)$ where we have forgotten about the order of its elements. Note that this is *not* the same thing as the set $\{a_1, \dots, a_M\}$, we want to keep track of multiple appearances of the same objects but while $[\![a\ a\ b]\!] \neq [\![a\ b]\!]$ for sets $\{a, a, b\} = \{a, b\}$.

Given a set $\mathcal{A}$, let $\mathbb{X}_{\mathcal{A}}^{M}$ be the set of multisets of elements of $\mathcal{A}$ of cardinality M, and $\mathbb{X}_{\mathcal{A}} = \bigcup_{M=0}^{\infty} \mathbb{X}_{\mathcal{A}}^{M}$. For any multiset $\tau = [\![(a_1)^{k_1} \cdots (a_N)^{k_N}]\!] \in \mathbb{X}_{\mathcal{A}}$, where $a_1, \dots, a_N$ are distinct elements of $\mathcal{A}$, its *symmetry factor* is defined recursively as

$$\mathfrak{s}([\![\ ]\!]) = 1, \quad \mathfrak{s}(\tau) := \prod_{i=1}^{N} (k_i)!. \tag{B.5}$$

Using these definitions, we can express Hadamard products of sums of vectors in the following compact form.

Proposition B.3. *Consider a set* $\mathcal{A}$ *and a summable sequence* $(v_a)_{a \in \mathcal{A}} \subseteq \mathbb{R}^d$ *indexed by* $\mathcal{A}$. *We have for any* $M \in \mathbb{N}$

$$(\mathbb{R}^d)^{\otimes M} \ni \left(\sum_{a \in \mathcal{A}} v_a \right)^{\odot M} = \sum_{\tau \in \mathbb{X}_{\mathcal{A}}^{M}} \frac{M!}{\mathfrak{s}(\tau)} v_\tau \tag{B.6}$$

where for $\tau := [\![a_1 \cdots a_M]\!]$ *we define* $v_\tau := v_{a_1} \odot \cdots \odot v_{a_M}$.

Proof. Considering M copies of $\mathcal{A}$ as different sets and using theorem B.1, we have on the one hand that

$$\otimes_{i=1}^{M} \left(\sum_{a \in \mathcal{A}} v_a \right) = \sum_{\boldsymbol{a} \in \times_{i=1}^{M} \mathcal{A}} \otimes_{i=1}^{M} (v_{a_i}).$$

On the other hand if we define linearly

$$\bigodot : (\mathbb{R}^d)^{\otimes M} \to \mathbb{R}^d, \quad v_1 \otimes \cdots \otimes v_M \mapsto v_1 \odot \cdots \odot v_M$$

we have

$$\left(\sum_{a\in\mathcal{A}} v_a\right)^{\odot M} = \bigodot \left(\sum_{a\in\mathcal{A}} v_a\right)^{\otimes M} = \bigodot \sum_{\boldsymbol{a}\in\mathcal{A}^M} \otimes_{i=1}^{M}(v_{a_i}) = \sum_{\boldsymbol{a}\in\mathcal{A}^M} \odot_{i=1}^{M}(v_{a_i})$$

by definition. Note that for any $\boldsymbol{a} = (a_1, \ldots, a_M), \boldsymbol{a'} = (a'_1, \ldots, a'_M) \in \mathcal{A}^M$,

$$\odot_{i=1}^{M}(v_{a_i}) = \odot_{i=1}^{M}(v_{a'_i})$$

whenever

$$(a_1, \ldots, a_M) = (a'_{\sigma(1)}, \ldots, a'_{\sigma(M)}).$$

for some $\sigma \in S_M$ *i.e.* $\odot$ is invariant to the action of S_M. Therefore we can write

$$\left(\sum_{a\in\mathcal{A}} v_a\right)^{\odot M} = \sum_{\tau\in\mathbb{X}_{\mathcal{A}}^M} \frac{M!}{\mathfrak{s}(\tau)} v_\tau.$$

since the cardinality of orbit τ is exactly $\frac{M!}{\mathfrak{s}(\tau)}$. □

We are now equipped to prove theorem B.4, which we do using the previous proposition and Wick's theorem. This proposition gives an expansion for covariances of type $\mathbb{E}[\varphi(X)\psi(Y)]$, where φ, ψ are polynomials and (X, Y) is a Gaussian vector with covariance of the form (B.7).

Proposition B.4. *Let $\varphi, \psi : \mathbb{R} \to \mathbb{R}$ be polynomial functions. Fix index sets $\mathcal{A}_1, \mathcal{A}_2$ and $\lambda : (\mathcal{A}_1 \cup \mathcal{A}_2) \times (\mathcal{A}_1 \cup \mathcal{A}_2) \to \mathbb{R}$ a symmetric and* summable *function. Then if*

$$(X, Y) \sim \mathcal{N}\left(0, \begin{bmatrix} \sum_{(a,a')\in\mathcal{A}_1\times\mathcal{A}_1} \lambda(a, a') & \sum_{(a,b)\in\mathcal{A}_1\times\mathcal{A}_2} \lambda(a, b) \\ \sum_{(b,a)\in\mathcal{A}_2\times\mathcal{A}_1} \lambda(b, a) & \sum_{(b,b')\in\mathcal{A}_2\times\mathcal{A}_2} \lambda(b, b') \end{bmatrix}\right) \tag{B.7}$$

*is well defined (*i.e. *the covariance matrix is positive definite), the following expansion holds*

$$\mathbb{E}\{\varphi(X)\psi(Y)\} = \sum_{\tau\in\mathbb{X}_{\mathcal{A}_1}} \sum_{\eta\in\mathbb{X}_{\mathcal{A}_2}} \frac{\varphi_\tau \psi_\eta}{\mathfrak{s}(\tau)\mathfrak{s}(\eta)} \sum_{\pi\in\mathcal{P}_{\mathbb{X}}(\tau\sqcup\eta)} \prod_{\{\pi_1,\pi_2\}\in\pi} \lambda(\pi_1, \pi_2) \tag{B.8}$$

where $\varphi_{[\![\alpha_1\cdots\alpha_m]\!]} := \varphi^{(m)}(0)$ and $\mathcal{P}_{\mathbb{X}}(\tau \sqcup \eta)$ is the set of pairings of the entries of τ and η.

Proof. Using Wick's theorem we have

$$\begin{aligned}\mathbb{E}\{\varphi(X)\psi(Y)\} &= \sum_{M,N=0}^{\infty} \frac{\varphi^{(M)}(0)\psi^{(N)}(0)}{M!N!}\mathbb{E}\{X^M Y^N\} \\ &= \sum_{M,N=0}^{\infty} \frac{\varphi^{(M)}(0)\psi^{(N)}(0)}{M!N!} \sum_{\pi\in\mathcal{P}([M]\sqcup[N])} \prod_{\substack{\{\pi_1,\pi_2\}\in\pi \\ \pi_1<\pi_2}} \mathbb{E}\{Z_{\pi_1}Z_{\pi_2}\} \\ &= \sum_{M,N=0}^{\infty} \frac{\varphi^{(M)}(0)\psi^{(N)}(0)}{M!N!} \sum_{\pi\in\mathcal{P}([M]\sqcup[N])} \prod_{\substack{\{\pi_1,\pi_2\}\in\pi \\ \pi_1<\pi_2}} \sum_{\gamma_1,\gamma_2\in\mathcal{A}_1\sqcup\mathcal{A}_2} \lambda_{\pi_1,\pi_2}(\gamma_1,\gamma_2).\end{aligned}$$

Here $[M]\sqcup[N] := ((1,1),\dots,(1,M),(2,1),\dots,(2,N))$ is ordered lexicographically, where with a slight abuse of notation $\mathcal{P}([M]\sqcup[N])$ is the set of pairings of elements in $[M]\sqcup[N]$ and

$$\lambda_{(i,*),(j,*)}(c,c') = \lambda(c,c')\mathbf{1}(c\in\mathcal{A}_i)\mathbf{1}(c'\in\mathcal{A}_j).$$

The claim would thus follow from the equality

$$\begin{aligned}&\frac{1}{M!N!} \sum_{\pi\in\mathcal{P}([M]\sqcup[N])} \prod_{\substack{\{\pi_1,\pi_2\}\in\pi \\ \pi_1<\pi_2}} \sum_{\gamma_1,\gamma_2\in\mathcal{A}_1\sqcup\mathcal{A}_2} \lambda_{\pi_1,\pi_2}(\gamma_1,\gamma_2) \\ &= \sum_{\tau\in\mathbb{X}^M_{\mathcal{A}_1}} \sum_{\eta\in\mathbb{X}^N_{\mathcal{A}_2}} \frac{1}{\mathfrak{s}(\tau)\mathfrak{s}(\eta)} \sum_{\pi\in\mathcal{P}_{\mathbb{X}}(\tau\sqcup\eta)} \prod_{\{\gamma_1,\gamma_2\}\in\pi} \lambda(\gamma_1,\gamma_2),\end{aligned}$$

which we now prove. Begin by noticing that

$$\sum_{\pi\in\mathcal{P}([M]\sqcup[N])} \prod_{\substack{\{\pi_1,\pi_2\}\in\pi \\ \pi_1<\pi_2}} \sum_{\gamma_1,\gamma_2\in\mathcal{A}_1\sqcup\mathcal{A}_2} \lambda_{\pi_1,\pi_2}(\gamma_1,\gamma_2) \tag{B.9}$$

$$= \sum_{\Gamma_1:[M]\to\mathcal{A}_1} \sum_{\Gamma_2:[N]\to\mathcal{A}_2} \sum_{\pi\in\mathcal{P}([M]\sqcup[N])} \prod_{\substack{\{\pi_1,\pi_2\}\in\pi \\ \pi_1<\pi_2}} \lambda(\Gamma(\pi_1),\Gamma(\pi_2)) \tag{B.10}$$

where $\Gamma((i,k)) = \Gamma_i(k)$. We are simply double counting: the left-hand side in (B.9) is obtained by labeling the elements of $[M]\sqcup[N]$ after having paired them, while the right-hand side is obtained by reversing the order of the operations with the functions Γ_1,Γ_2 specifying the labels.

Then for any fixed $\tau\in\mathbb{X}^M_{\mathcal{A}_1}$ and $\eta\in\mathbb{X}^N_{\mathcal{A}_2}$, we have

$$\sum_{\pi\in\mathcal{P}_{\mathbb{X}}(\tau\sqcup\eta)} \prod_{\{\gamma_1,\gamma_2\}\in\pi} \lambda(\gamma_1,\gamma_2) = \sum_{\pi\in\mathcal{P}([M]\sqcup[N])} \prod_{\substack{\{\pi_1,\pi_2\}\in\pi \\ \pi_1<\pi_2}} \lambda(\Gamma(\pi_1),\Gamma(\pi_2)) \tag{B.11}$$

whenever $\Gamma_1:[M]\to\mathcal{A}_1$ and $\Gamma_2:[M]\to\mathcal{A}_2$ satisfy

$$\tau = [\![\Gamma_1(1)\cdots\Gamma_1(M)]\!] \text{ and } \eta = [\![\Gamma_2(1)\cdots\Gamma_2(N)]\!].$$

Letting $\Omega_{\tau,\eta}$ denote the set of all such pairs (Γ_1, Γ_2), we note that

$$|\Omega_{\tau,\eta}| = \frac{M!}{\mathfrak{s}(\tau)} \frac{N!}{\mathfrak{s}(\eta)}$$

and, in turn, that

$$\bigsqcup_{\tau \in \mathbb{X}^M_{\mathcal{A}_1}} \bigsqcup_{\eta \in \mathbb{X}^N_{\mathcal{A}_2}} \Omega_{\tau,\eta} = \{\Gamma_1 : [M] \to \mathcal{A}_1\} \times \{\Gamma_2 : [N] \to \mathcal{A}_2\}. \tag{B.12}$$

This allows us to conclude that

$$\begin{aligned}
&\sum_{\tau \in \mathbb{X}^M_{\mathcal{A}_1}} \sum_{\eta \in \mathbb{X}^N_{\mathcal{A}_2}} \frac{1}{\mathfrak{s}(\tau)\mathfrak{s}(\eta)} \sum_{\pi \in \mathcal{P}_{\mathbb{X}}(\tau \sqcup \eta)} \prod_{\{\gamma_1,\gamma_2\} \in \pi} \lambda(\gamma_1, \gamma_2) \\
&= \sum_{\tau \in \mathbb{X}^M_{\mathcal{A}_1}} \sum_{\eta \in \mathbb{X}^N_{\mathcal{A}_2}} \frac{1}{\mathfrak{s}(\tau)\mathfrak{s}(\eta)} \frac{1}{|\Omega_{\tau,\eta}|} \sum_{(\Gamma_1,\Gamma_2) \in \Omega_{\tau,\eta}} \sum_{\pi \in \mathcal{P}([M] \sqcup [N])} \prod_{\substack{\{\pi_1,\pi_2\} \in \pi \\ \pi_1 < \pi_2}} \lambda(\Gamma(\pi_1), \Gamma(\pi_2)) \\
&= \frac{1}{M!N!} \sum_{\tau \in \mathbb{X}^M_{\mathcal{A}_1}} \sum_{\eta \in \mathbb{X}^N_{\mathcal{A}_2}} \sum_{(\Gamma_1,\Gamma_2) \in \Omega_{\tau,\eta}} \sum_{\pi \in \mathcal{P}([M] \sqcup [N])} \prod_{\substack{\{\pi_1,\pi_2\} \in \pi \\ \pi_1 < \pi_2}} \lambda(\Gamma(\pi_1), \Gamma(\pi_2)) \\
&= \frac{1}{M!N!} \sum_{\Gamma_1 : [M] \to \mathcal{A}_1} \sum_{\Gamma_2 : [M] \to \mathcal{A}_2} \sum_{\pi \in \mathcal{P}([M] \sqcup [N])} \prod_{\substack{\{\pi_1,\pi_2\} \in \pi \\ \pi_1 < \pi_2}} \lambda(\Gamma(\pi_1), \Gamma(\pi_2)) \\
&= \frac{1}{M!N!} \sum_{\pi \in \mathcal{P}([M] \sqcup [N])} \prod_{\substack{\{\pi_1,\pi_2\} \in \pi \\ \pi_1 < \pi_2}} \sum_{\gamma_1,\gamma_2 \in \mathcal{A}_1 \sqcup \mathcal{A}_2} \lambda_{\pi_1,\pi_2}(\gamma_1, \gamma_2)
\end{aligned}$$

where the first equality follows from (B.11), the second from $|\Omega_{\tau,\eta}| = \frac{M!}{\mathfrak{s}(\tau)} \frac{N!}{\mathfrak{s}(\eta)}$, the third is (B.12) and the last (B.9). □

Appendix C

Background on Rough Paths Theory

We here give a brief overview of Rough Paths theory, citing without proof the results we are using for our proofs. We follow Chevyrev [12], Friz and Hairer [36], Lyons and Qian [76], to which we refer the reader for further details. In addition to the above references, we refer to Cass and Salvi [8] for the reader focused on applications to Machine Learning.

C.1 Basic Definitions

Denote by $(\mathbb{R}^d)^{\otimes n} := \mathbb{R}^d \otimes \cdots \otimes \mathbb{R}^d$ the tensor product of n copies $\mathbb{R}^d$, set $(\mathbb{R}^d)^{\otimes 0} := \mathbb{R}$. Let $T(\mathbb{R}^d) := \bigoplus_{k=0}^{\infty} (\mathbb{R}^d)^{\otimes k}$ be the tensor algebra equipped with sum and tensor product. We will denote by $T^{(M)}(\mathbb{R}^d) := \bigoplus_{k=0}^{M} (\mathbb{R}^d)^{\otimes k}$ the *truncated* tensor algebra.

Definition C.1. *Let $\{e_1, \dots, e_d\}$ be the canonical basis of $\mathbb{R}^d$, then*

$$\{e_{i_1} \otimes \cdots \otimes e_{i_k} : (i_1, \dots, i_k) \in [d]^k\}$$

is a basis of $(\mathbb{R}^d)^{\otimes k}$. We equip $(\mathbb{R}^d)^{\otimes k}$ with the inner product $\langle \cdot, \cdot \rangle_{(\mathbb{R}^d)^{\otimes k}}$ defined on basis elements as

$$\langle e_{i_1} \otimes \cdots \otimes e_{i_k}, e_{j_1} \otimes \cdots \otimes e_{j_k} \rangle_{(\mathbb{R}^d)^{\otimes k}} = \delta^{(j_1,\dots,j_k)}_{(i_1,\dots,i_k)}$$

and we extend it to $T(\mathbb{R}^d)$ by

$$\langle A, B \rangle_{T(\mathbb{R}^d)} := \sum_{k=0}^{\infty} \langle a_k, b_k \rangle_{(\mathbb{R}^d)^{\otimes k}}$$

where $A = (a_0, a_1, \dots)$ and $B = (b_0, b_1, \dots)$

Note how for any $A, B \in T(\mathbb{R}^d)$ we have $\langle A \otimes e_i, B \otimes e_j \rangle_{T(\mathbb{R}^d)} = \langle A, B \rangle_{T(\mathbb{R}^d)} \langle e_i, e_j \rangle_{\mathbb{R}^d}$.

Definition C.2 (Infinite Tensor Algebra)**.** *The infinite tensor algebra is defined as the space* $T((\mathbb{R}^d)) := \prod_{k=0}^{\infty}(\mathbb{R}^d)^{\otimes k}$ *equipped with the operations* $+$ *and* $\otimes$ *which act in the natural algebraic way; its elements are called* tensor series*.*

It is easily seen that $(T((\mathbb{R}^d)), +, \otimes)$ is an algebra with unit $\mathbf{1} = (1, 0, 0, \cdots)$ and we can endow it with a natural product .

Another point of view could be taken on the definitions of these spaces, one that we will prefer later on. If we define $\mathbb{W}_d$ to be the set of words in d letters then $T((\mathbb{R}^d)) \sim \mathbb{R}^{\mathbb{W}_d}$, $T(\mathbb{R}^d)$ is the subset of such functions with finite support and

$$\langle A, B\rangle_{T(\mathbb{R}^d)} = \sum_{I\in\mathbb{W}_d} A_I B_I = \sum_{k=0}^{\infty}\sum_{|I|=k} A_I B_I$$

where $|I|$ is the length of the word I. The empty word, the only one with length 0, is denoted by () and corresponds to the basis element of $(\mathbb{R}^d)^{\otimes 0}$. In this view the tensor product coincides with concatenation of words accordingly distributed and the closure of $T(\mathbb{R}^d)$ with respect to its product is just the l^2 space $l^2(\mathbb{W}_d)$.

The product $\otimes$ is not the only one available on the space $T(\mathbb{R}^d)$, in fact one can define on it what is known as the *shuffle product* $\sqcup\!\sqcup$, which

Definition C.3. *The* shuffle product $\sqcup\!\sqcup : T(\mathbb{R}^d)\times T(\mathbb{R}^d) \to T(\mathbb{R}^d)$ *is defined inductively on basis elements as*

$$\forall I \in \mathbb{W}_d, \quad e_I \sqcup\!\sqcup e_{()} = e_{()} \sqcup\!\sqcup e_I = e_I$$

$$\forall I, J \in \mathbb{W}_d, \forall a, b \in \{1, \dots, d\}, \quad e_{Ia} \sqcup\!\sqcup e_{Jb} = (e_I \sqcup\!\sqcup e_{Jb}) \otimes e_a + (e_{Ia} \sqcup\!\sqcup e_J) \otimes e_b$$

and extended to the whole space by linearity. Here for $I = i_1 \dots i_{|I|}$ *we write* $e_I := e_{i_1} \otimes \cdots \otimes e_{i_{|I|}}$.

The notion of shuffle product is important in the context of *geometric rough paths*.

C.2 Rough Paths Primer

Rough paths theory was developed by Lyons in the 1990's to make path-wise sense of equations of type

$$Y_t = \int_0^t F(Y_s) dX_s \in \mathbb{R}^N \tag{C.1}$$

where $X : [0, 1] \to \mathbb{R}^d$ is a α-Holder path with $\alpha \leq \frac{1}{2}$, for some $F : \mathbb{R}^N \to \mathbb{R}^{N\times d}$.

This class of paths includes Brownian motion and semi-martingales. While their integration theory, *stochastic calculus*, is extensively studied (see [109]), it is always understood in a distributional, probabilistic sense.

In fact, in the semi-martingale case, while the map $X \mapsto Y$ is measurable, it is not continuous. This means that there is no canonical way to construct the solution

Y given a realization of X. It is not possible to extend the theory from smooth paths to the entire path space in a unique way, as evidenced by the fact that Itô and Stratonovich integration, while coinciding on smooth paths, generally yield different results for semi-martingales.

The situation is even more problematic: Lyons [74] showed that there exists no separable Banach space $V \subset C^0([0,1];\mathbb{R}^d)$ containing Brownian trajectories almost surely and such that the bilinear map $(X,Y) \mapsto \int_0^\cdot Y_s dX_s$, defined on smooth elements of V, extends to the entire space $V \times V$.

Rough paths theory, in contrast, provides a pathwise understanding of such integrals. The main insight of the theory is that, in order to guarantee existence and uniqueness of solutions to eq. (C.1), the information contained in the path X alone is not sufficient. Additional higher-order information is required, corresponding to sufficiently many (up to $M := \lfloor \frac{1}{\alpha} \rfloor$) iterated integrals

$$\int_{s \le r_1 \le \dots r_m \le t} dX_{r_1} \otimes \cdots \otimes dX_{r_m} \in (\mathbb{R}^d)^{\otimes m}. \tag{C.2}$$

More precisely, this means the collection of integrals for all combinations of coordinate indices:

$$\left(\int_{s \le r_1 \le \dots r_m \le t} dX^{i_1}_{r_1} \dots dX^{i_m}_{r_m} \right)_{1 \le i_1, \dots, i_m \le d}. \tag{C.3}$$

To develop intuition for why iterated integrals are essential, consider the case $\alpha > \frac{1}{3}$ and, for some smooth $F : \mathbb{R}^d \to \mathbb{R}^{d \times d}$, the simpler case of an integral

$$\int_0^t F(X_s) dX_s.$$

Using a first-order Taylor expansion, we have locally $F(X_t) \simeq F(X_s) + dF_{X_s}[X_{s,t}]$, where we wrote as it's custumary $X_{s,t} = X_t - X_s = \int_s^t dX_r$, so that for some partition $\mathcal{P}$ of $[0,1]$ one has

$$\int_0^1 F(X_r) dX_r \simeq \sum_{[s,t] \in \mathcal{P}} \left(F(X_s) \int_s^t dX_r + dF_{X_s} \int_s^t \int_s^r dX_u \otimes dX_r \right) + \text{higher order terms.}$$

The second term involves the second-order iterated integral $\int_s^t \int_s^r dX_u \otimes dX_r$, demonstrating why such higher-order information might be necessary for a well-defined theory of integration. In fact, iterated integrals play, on path space, a similar role to that played by monomials on $\mathbb{R}^d$ — they provide the fundamental building blocks for representing more complex functions.

This necessity is also evident in the stochastic setting: when X is a semi-martingale, equation (C.1) can be understood in either the Itô or Stratonovich sense. These two

notions of stochastic integration differ, and this difference is precisely quantified by a second-order object: the quadratic variation of X. Thus, regardless of whether we approach the problem from a deterministic or stochastic perspective, higher-order information beyond the path itself is essential for a coherent theory of integration.

It turns out that the extra information needed is encoded in continuous (according to the normed structure of the truncated tensor algebra $T^{(M)}$ (*c.f.* Friz and Hairer [36, 2.4])) maps

$$\mathbf{X} : \{(s,t) \in [0,1]^2 \mid s \leq t\} =: \Delta \to T^{(M)}(\mathbb{R}^d) := \bigoplus_{m=0}^{M} (\mathbb{R}^d)^{\otimes m}$$

extending X in the sense that $\mathbf{X}^1_{s,t} = X_t - X_s$ and behaving like iterated integrals both algebraically, meaning that *Chen's relation*

$$\mathbf{X}_{s,t} = \mathbf{X}_{s,u} \otimes \mathbf{X}_{u,t} \tag{C.4}$$

holds for any $0 \leq s \leq u \leq t \leq 1$, and analytically, meaning that there exist $C > 0$ such that

$$\forall\ |I| \leq M,\ \forall s \leq t,\ |\langle \mathbf{X}_{s,t}, e_I \rangle| \leq C|s-t|^{|I|\alpha}. \tag{C.5}$$

Definition C.4. *The space of functions* $\mathbf{X} : \Delta \to T^{(M)}(\mathbb{R}^d)$ *satisfying Chen's relation* (C.4) *and the analytic requirement* (C.5) *is the space of* α-Hölder rough paths, *denoted by* $\mathscr{C}^\alpha([0,1];\mathbb{R}^d)$.

While the set of α-Hölder rough paths does not have a vector-space structure, it can be endowed with a distance, and hence constitutes a metric space (*c.f.* Friz and Hairer [36, 2.4]). Importantly, every smooth path $X \in C^\infty([0,1];\mathbb{R}^d)$ can be canonically lifted to an element of $\mathscr{C}^\alpha([0,1];\mathbb{R}^d)$ for any $\alpha \in (0,1)$ by defining $\mathbf{X}^m_{s,t}$ as the iterated integrals (C.2), which are always well-defined under the smoothness assumption. The space of such lifts is denoted $\mathscr{L}^\infty(C^\infty) \subset \mathscr{C}^\alpha$.

With this lifting procedure, one can extend the solution map $\mathbf{X} \mapsto Y$ from $\mathscr{L}^\infty(C^\infty)$ to the entire space $\mathscr{C}^\alpha$, in the sense that the classical solution map $X \mapsto Y$ on smooth paths factors through both the canonical lift and the rough path solution map. Crucially, this extended map is continuous under the rough path metric, which accounts for the higher-order data.

Returning to the stochastic setting, consider Brownian motion $B : [0,1] \to \mathbb{R}$. One can probabilistically define its second-order integrals in two ways: the Stratonovich integral $\mathbb{B}^S_{s,t} := \int_s^t B_r \circ dB_r$ and the Itô integral $\mathbb{B}^I_{s,t} := \int_s^t B_r dB_r$. Then, almost surely, both $(B, \mathbb{B}^S)$ and $(B, \mathbb{B}^I)$ define elements of $\mathscr{C}^\alpha$ for $\alpha \in (\frac{1}{3}, \frac{1}{2})$. The corresponding solution maps $B \mapsto (B, \mathbb{B}^S) \mapsto Y$ and $B \mapsto (B, \mathbb{B}^I) \mapsto Y$ yield the solutions of eq. (C.1) in the Stratonovich and Itô senses, respectively.

The generality of rough path theory makes it particularly attractive: it extends beyond semi-martingales to rougher settings, such as fractional Brownian motion, which is crucial in financial applications. This broad applicability motivates our choice to work directly in the rough path framework.

For our purposes, we need to give meaning to simpler integrals than (C.1), namely "linear" integrals of the form

$$\int_0^t Y_s dX_s.$$

To accomplish this, we introduce the key notion of *controlled rough paths*:

Definition C.5 (Def. 4.6 in Friz and Hairer [36]). *Let $\alpha \in (0,1)$, $M = \lfloor \frac{1}{\alpha} \rfloor$ and W a Banach Space. Given $\mathbf{X} \in \mathscr{C}^\alpha([0,1];\mathbb{R}^d)$ we define the space of* controlled (by $\mathbf{X}$) W-valued rough paths *as the set*

$$\mathscr{D}_{\mathbf{X}}^{M\alpha}([0,1];W) \subseteq \left\{[0,1] \to \mathcal{L}(T^{(M-1)}(\mathbb{R}^d);W)\right\}$$

of functions $\mathbf{Y}$ satisfying

$$\forall |I| \leq M-1, \ \|\langle \mathbf{Y}_t, e_I\rangle - \langle \mathbf{Y}_s, \mathbf{X}_{s,t} \otimes e_I\rangle\|_W \leq C|t-s|^{(M-|I|)\alpha} \tag{C.6}$$

We say that $\mathbf{Y}$ lifts $Y_t = \mathbf{Y}_t^{()} \in W$.

With this framework in place, we can now rigorously define the desired integrals. Given a controlled rough path $\mathbf{Y}$ driven by $\mathbf{X}$, the following proposition shows that the integral is well-defined and extends classical integration theory:

Proposition C.6 (Ex. 4.18 in Friz and Hairer [36]). *Given $\mathbf{X} \in \mathscr{C}^\alpha([0,1];\mathbb{R}^d)$ and $\mathbf{Y} \in \mathscr{D}_{\mathbf{X}}^{M\alpha}([0,1];W)$ the integral*[1]

$$\int_0^t Y_s \, \mathrm{d}\mathbf{X}_s^i := \lim_{|\mathcal{P}|\to 0} \sum_{[s,t]\in\mathcal{P}} \sum_{|I|<M} \mathbf{Y}_s^I \mathbf{X}_{s,t}^{Ii} \in W$$

is well defined and exists, for all $i \in 1,\dots,d$.

Given $W = \mathcal{L}(\mathbb{R}^d;V)$, using the identification

$$\mathcal{L}((\mathbb{R}^d)^{\otimes m};W) \simeq \mathcal{L}((\mathbb{R}^d)^{\otimes m+1};V) \text{ as } \phi(\mathbf{x})[y] \sim \phi[\mathbf{x}\otimes y],$$

one can then define

$$\int_0^t Y_s[\,\mathrm{d}\mathbf{X}_s] = \sum_{i=1}^d \int_0^t Y_s[e_i] \ \mathrm{d}\mathbf{X}_s^i := \lim_{|\mathcal{P}|\to 0} \sum_{[s,t]\in\mathcal{P}} \mathbf{Y}_s[\mathbf{X}_{s,t}] \in V$$

[1] Following standard convention in the literature (*c.f.* Friz and Hairer [36][eq. 4.35]), we write $\int_0^t Y_s \, \mathrm{d}\mathbf{X}_s^i$ rather than $\int_0^t \mathbf{Y}_s \, \mathrm{d}\mathbf{X}_s^i$ when the lift $\mathbf{Y}$ is clear from context.

C.3 Signature

We have seen that to define a well-posed integration theory for α-Hölder paths, it suffices to consider iterated-integral-like objects (rough paths) up to level $\lfloor \frac{1}{\alpha} \rfloor$. In fact, given the data up to this level, there exists a unique and canonical way to obtain all remaining iterated integrals:

Proposition C.7 (Thm. 3.7 Lyons et al. [77]). *Given $\mathbf{X} \in \mathscr{C}^\alpha([0,1];\mathbb{R}^d)$ there exists a unique element $Sig(\mathbf{X})$ of $\Delta \to T((\mathbb{R}^d))$ extending $\mathbf{X} : \Delta \to T^{\lfloor \frac{1}{\alpha} \rfloor}(\mathbb{R}^d)$, satisfying Chen's relation*

$$Sig(\mathbf{X})_{s,u} \otimes Sig(\mathbf{X})_{u,t} = Sig(\mathbf{X})_{s,t}$$

and such that

$$\forall I,\ \exists C_{|I|} \geq 0,\ \forall s \leq t,\ |\langle Sig(\mathbf{X})_{s,t}, e_I \rangle| \leq C_{|I|} |t-s|^{|I|\alpha}$$

This object, the collection of *all* iterated integrals, has recently emerged as a topic of significant interest in machine learning, independent of its origins in rough paths theory, due to its remarkable properties.

This section presents the key aspects of this object, in the context of machine learning, so considering the space of paths with finite variation $C_{1,0}([0,1] : \mathbb{R}^d)$, part of the $\alpha > \frac{1}{2}$ case. For a comprehensive treatment, we refer to Cass and Salvi [8].

Definition C.8 (Signature). *Given $\gamma \in C_{1,0}([0,1];\mathbb{R}^d)$ and $s,t \in [0,1]$ s.t. $s \leq t$, the signature $Sig(\gamma)_{s,t} \in T((\mathbb{R}^d))$ of the path γ over $[s,t]$ is defined as*

$$Sig(\gamma)_{s,t} := (1, \int_{s<u_1<t} d\gamma_{u_1}, \cdots, \int\cdots\int_{s<u_1<\cdots<u_k<t} d\gamma_{u_1} \otimes \cdots \otimes d\gamma_{u_k}, \cdots) \tag{C.7}$$

where the integral is intended in the Riemann-Stjeltes sense. Equivalently $Sig(\gamma)_{s,t}$ is that element of $l^2(\mathbb{W}_d)$ defined recursively on words as

$$Sig(\gamma)^{()}_{s,t} = 1, \quad Sig(\gamma)^{Ij}_{s,t} = \int_s^t Sig(\gamma)^I_{s,r} d\gamma^j_r. \tag{C.8}$$

C.3.1 Notable Results

Here we present some notable results of which we will make use through the paper. We omit the proofs if they can be easily found in the suggested references.

The first result is about bounding the norm of Signature entries:

Proposition C.9 (Factorial Decay Rate). *Given $\gamma \in C_{1,0}([0,1];\mathbb{R}^d)$, for all $k \geq 1$ and*

$s, t \in [0,1]$ s.t. $s \leq t$ one has

$$\left\| \int_{s<u_1<\cdots<u_k<t} \cdots \int d\gamma_{u_1} \otimes \cdots \otimes d\gamma_{u_k} \right\|_{(\mathbb{R}^d)^{\otimes k}} \leq \frac{\|\gamma\|^k_{1-var,[s,t]}}{k!} \tag{C.9}$$

The most important fact about Signature is that it acts as the basis for a Taylor expansion in path space. In fact just as finite linear combinations of monomials are dense in the continuous functions with a compact input set, finite linear combinations of Signature entries are dense in continuous functions from compact path-spaces:

Theorem C.10 (Universal Approximation [31])**.** *Fix $K \subset C_{1,0}([0,1];\mathbb{R}^{d+1})$ compact such that for any $\gamma \in K$ it holds $\gamma^1_t = t$. For any $F \in C^0(K;\mathbb{R})$ and $\epsilon > 0$ there is an integer $N \geq 0$ such that*

$$\sup_{\gamma \in K} |F(\gamma) - \sum_{|I| \leq N} \alpha_I Sig(\gamma)^I_{0,1}| \leq \epsilon \tag{C.10}$$

for some finite sequence $(\alpha_I)_{|I|\leq N}$ of real numbers.

Remark. *There is no magic in this result, it is just an application of Stone-Weiestrass enabled by the rich algebraic structure of iterated integrals, that of the shuffle algebra, studied originally in Chen [10].*

We will need to restrict some paths to sub-intervals of $[0,1]$ in such a way to still be able to consider them meaningfully as elements of $C_{1,0}([0,1];\mathbb{R}^d)$, this is done in the following way:

Definition C.11. *Given any path $\gamma \in C_{1,0}([0,1];\mathbb{R}^d)$ we define its restriction on a sub-interval $[s,t] \subseteq [0,1]$ as the path $\gamma_{[s,t]} \in C_{1,0}([0,1];\mathbb{R}^d)$ with values*

$$\gamma_{[s,t]}(r) := \begin{cases} 0 & \text{if } r < s \\ \gamma_r - \gamma_s & \text{if } s \leq r \leq t \\ \gamma_t - \gamma_s & \text{if } r > t \end{cases} \tag{C.11}$$

This definition is such that the following important equation holds

$$\mathrm{Sig}(\gamma)_{s,t} = \mathrm{Sig}(\gamma_{[s,t]})_{0,1}. \tag{C.12}$$

With the right augmentation of the paths one can see that the Signature distinguishes between different sections of paths, this will be crucial for some of the original results presented in this work.

Lemma C.12. *Assume* $\omega, \gamma \in C_{1,0}([0,1];\mathbb{R}^{d+2})$ *with* $\omega_t^1 = \gamma_t^1 \equiv t$ *and* $\omega_t^2 = \gamma_t^2 \equiv t^2$. *Then*

$$Sig(\omega)_{s,t} = Sig(\gamma)_{s',t'} \Leftrightarrow \omega_{[s,t]} = \gamma_{[s',t']} \tag{C.13}$$

Proof. The *if* part follows from $\mathrm{Sig}(\gamma)_{s,t} = \mathrm{Sig}(\gamma_{[s,t]})_{0,1}$. For the *only if* part, If $s = s'$ and $t = t'$ the statement holds; this is because if the signatures over the time interval $[s,t]$ of two time-augmented paths are equal, then the two paths must be equal on $[s,t]$. We now show that augmenting the path with t^2 and imposing equality of signatures, implies $s = s'$ and $t = t'$, which will in turn allow us to conclude the proof by the previous remark. Assume $\mathrm{Sig}(\omega)_{s,t} = \mathrm{Sig}(\gamma)_{s',t'}$, in particular we must have

$$\int_s^t d(r^2) = t^2 - s^2 = (t')^2 - (s')^2 = \int_{s'}^{t'} d(r^2) \tag{C.14}$$

$$\int_s^t d(r) = t - s = t' - s' = \int_{s'}^{t'} d(r) \tag{C.15}$$

which reduces to the system

$$\begin{cases} t^2 - s^2 = (t')^2 - (s')^2 \\ t - s = t' - s' \end{cases} \quad \begin{cases} t + s = t' + s' \\ t - s = t' - s' \end{cases} \quad \begin{cases} 2t = 2t' \\ 2s = 2s' \end{cases}$$

Hence it must be true that $t = t'$ and $s = s'$. □

C.4 Wronskian Theory

In this section we obtain a unified theory studying the solutions to general Linear CDEs. The results presented here are not new and can be found in different terms in the literature [37], despite this we have decided to reproduce them from scratch for completeness, notational reasons and to present a self-contained theory.

Theorem C.13. *For any choice* $\{A^1, \dots, A^d\} \subseteq C^0([0,1]; \mathbb{R}^{N\times N})$ *and* $\omega \in C_1([0,1]; \mathbb{R}^d)$ *there exist a unique map* $W \in C^0([0,1]\times[0,1]; \mathbb{R}^{N\times N})$ *solving the following CDE*

$$W_{s,t} = Id_N + \sum_{i=1}^{d} \int_{\tau=s}^{t} A^i_\tau W_{s,\tau} d\omega^i_\tau \tag{C.16}$$

Proof. We will use Banach fixed point theorem leveraging the completeness of the space $\Omega := C^0([0,1]\times[0,1]; \mathbb{R}^{N\times N})$ with the uniform norm

$$\|X\|_\infty := \sup_{s,t\in[0,1]} \|X_{s,t}\|_{op}$$

Define the map $\Gamma : \Omega \to \Omega$ as

$$\Gamma(X)_{s,t} = Id_N + \sum_{i=1}^{d} \int_{\tau=s}^{t} A^i_\tau X_{s,\tau} d\omega^i_\tau$$

One has, for $X, Y \in \Omega$ and $k \in \mathbb{N}$ setting $\Gamma^0 = Id_\Omega$, that

$$\Gamma^{k+1}(X)_{s,t} - \Gamma^{k+1}(Y)_{s,t} = \sum_{i=1}^{d} \int_{\tau=s}^{t} A^i_\tau \left(\Gamma^k(X)_{s,\tau} - \Gamma^k(Y)_{s,\tau}\right) d\omega^i_\tau$$

which iterated gives

$$\begin{aligned}\Gamma^{k+1}(X)_{s,t} - \Gamma^{k+1}(Y)_{s,t} &= \sum_{\substack{I\in\mathbf{W}_d \\ |I|=k+1}} \int_{\tau_{k+1}=s}^{t} \cdots \int_{\tau_1=s}^{\tau_2} \left(\prod_{j=k+1}^{1} A^{I_j}_{\tau_j}\right) (X_{s,\tau_1} - Y_{s,\tau_1}) \prod_{j=1}^{k+1} d\omega^{I_j}_{\tau_j} \\ &= \sum_{\substack{I\in\mathbf{W}_d \\ |I|=k+1}} \int_{\tau\in\Delta^{k+1}_{[s,t]}} A^I_\tau \left(X_{s,\tau_1} - Y_{s,\tau_1}\right) d\omega^I_\tau\end{aligned}$$

where $\mathbf{W}_d$ is the set of words in the alphabet $\{1, \dots, d\}$ and

$$\Delta^k_{[s,t]} := \{(\tau_1, \dots, \tau_k) \in [0,1]^k : \forall j \in 1, \dots, k-1.\ \tau_j \le \tau_{j+1}\}$$

$$A^I_\tau := \prod_{j=k+1}^{1} A^{I_j}_{\tau_j} \quad d\omega^I_\tau := \prod_{j=1}^{k+1} d\omega^{I_j}_{\tau_j}$$

By defining $M = \max\{\|A^i\|_\infty : i \in \{1, \dots, d\}\}$ then one clearly has

$$\left\|\Gamma^k(X) - \Gamma^k(Y)\right\|_\infty \leq \frac{(dM\,\|\omega\|_{1-var})^k}{k!}\,\|X - Y\|_\infty \tag{C.17}$$

thus definitely (in k) the map Γ^k is a contraction. By Banach fixed point there exist a unique fixed point $W \in \Omega$. □

Theorem C.14. *Under the assumptions of the previous theorem one can write $W_{s,t}$ explicitly as*

$$W_{s,t} = \sum_{I \in \mathbf{W}_d} \int_{\tau \in \Delta^{|I|}_{[s,t]}} A^I_\tau d\omega^I_\tau \tag{C.18}$$

moreover if for all i the matrix-valued maps are constant on all $[0,1]$ i.e. *$A^i_t \equiv A_i$ then*

$$W_{s,t} = \sum_{I \in \mathbf{W}_d} A_I Sig(\omega)^I_{s,t} \tag{C.19}$$

where $Sig(\omega)^I_{s,t}$ is the Signature of the path ω.

Proof. The second assertion follows from the first by definition of the Signature of a path.

Regarding the first notice how the series is absolutely convergent in $\mathbb{R}^{N\times N}$, uniformly in s, t since

$$\begin{aligned}\sum_{I \in \mathbf{W}_d} \left\| \int_{\tau \in \Delta^{|I|}_{[s,t]}} A^I_\tau d\omega^I_\tau \right\|_{op} \leq & \sum_{k=0}^{\infty} d^k M^k \frac{\|\omega\|^k_{1-var,[s,t]}}{k!} \\ = & e^{dM\|\omega\|_{1-var,[s,t]}} \leq e^{dM\|\omega\|_{1-var,[0,1]}}\end{aligned}$$

thus for any $s, t \in [0,1]$ the series defines an element of $\tilde{W}_{s,t} \in \mathbb{R}^{N\times N}$.

Using the uniformity of this bound and the fact that for all $I \in \mathbf{W}_d$ one has

$$\tilde{W}^I_{s,t} := \int_{\tau \in \Delta^{|I|}_{[s,t]}} A^I_\tau d\omega^I_\tau \in \Omega$$

as a function of (s,t), which moreover is uniformly continuous

$$\begin{aligned}\left\|\tilde{W}^I_{s_1,t_1} - \tilde{W}^I_{s_2,t_2}\right\|_{op} = & \left\| \int_{\tau \in \Delta^{|I|}_{[s_1\wedge s_2, t_1 \vee t_2]}} (\delta_{\tau \in \Delta^{|I|}_{[s_1,t_1]}} - \delta_{\tau \in \Delta^{|I|}_{[s_2,t_2]}}) A^I_\tau d\omega^I_\tau \right\|_{op} \\ \leq & M^{|I|} \int_{\tau \in \Delta^{|I|}_{[s_1\wedge s_2, t_1\vee t_2]}} \left| \delta_{\tau \in \Delta^{|I|}_{[s_1,t_1]}} - \delta_{\tau \in \Delta^{|I|}_{[s_2,t_2]}} \right| |d\omega^I_\tau| \\ \leq & M^{|I|} \|\omega\|^{|I|}_{[s_1\wedge s_2, s_1\vee s_2]\cup[t_1\wedge t_2, t_1\vee t_2]},\end{aligned}$$

one concludes that $\tilde{W}_{s,t} \in \Omega$. Finally notice that $\tilde{W}_{s,t}$ is a fixed point of Γ

$$\begin{aligned}\Gamma(\tilde{W}_{s,t}) =& Id_N + \sum_{i=1}^{d} \int_{\tau=s}^{t} A_\tau^i \left(\sum_{I \in \mathbf{W}_d} \int_{\tau \in \Delta_{[0,1]}^{|I|}} A_\tau^I d\omega_\tau^I \right) d\omega_\tau^i \\ =& Id_N + \sum_{\substack{I \in \mathbf{W}_d \\ |I| \geq 1}} \int_{\tau \in \Delta_{[0,1]}^{|I|}} A_\tau^I d\omega_\tau^I d\omega_\tau^i \\ =& \sum_{I \in \mathbf{W}_d} \int_{\tau \in \Delta_{[0,1]}^{|I|}} A_\tau^I d\omega_\tau^I = \tilde{W}_{s,t}\end{aligned}$$

and conclude by uniqueness. □

Proposition C.15. *Under the previous conditions, the unique solution of the N-dimensional CDE*

$$X_t = X_0 + \sum_{i=1}^{d} \int_{\tau=0}^{t} A_\tau^i X_\tau d\omega_\tau^i \tag{C.20}$$

is given by

$$X_t = W_{0,t} X_0 \tag{C.21}$$

Proof. The solutions are unique by standard results [37][Thm. 3.7], moreover

$$W_{0,t}X_0 = \left(Id_N + \sum_{i=1}^{d} \int_{\tau=0}^{t} A_\tau^i W_{0,\tau} d\omega_\tau^i \right) X_0 = X_0 + \sum_{i=1}^{d} \int_{\tau=0}^{t} A_\tau^i (W_{0,\tau} X_0) d\omega_\tau^i$$

□

Proposition C.16. *The Wronskian matrix has the following properties:*

1. $\forall r, s, t \in [0,1]. \quad W_{r,t} = W_{s,t} W_{r,s}$
2. $\forall s, t \in [0,1]. \quad W_{s,t}^{-1} = W_{t,s}$
3. $\forall s, t \in [0,1]. \quad W_{s,t} = Id_N + \sum_{i=1}^{d} \int_{\sigma=s}^{t} W_{\sigma,t} A_\sigma^i d\omega_\sigma^i$

Proof. Regarding the first statement notice that for all $X_0 \in \mathbb{R}^N$ one has

$$\begin{aligned}\tilde{X}_t :=& W_{s,t} W_{r,s} X_0 = \left(Id_N + \sum_{i=1}^{d} \int_{\tau=s}^{t} A_\tau^i W_{s,\tau} d\omega_\tau^i \right) W_{r,s} X_0 \\ =& W_{r,s} X_0 + \sum_{i=1}^{d} \int_{\tau=s}^{t} A_\tau^i (W_{s,\tau} W_{r,s} X_0) d\omega_\tau^i \\ =& W_{r,s} X_0 + \sum_{i=1}^{d} \int_{\tau=s}^{t} A_\tau^i \tilde{X}_\tau d\omega_\tau^i\end{aligned}$$

and by the previous proposition also

$$X_t := W_{r,t} X_0 = W_{r,t} X_0 + \sum_{i=1}^{d} \int_{\tau=r}^{t} A_\tau^i X_\tau d\omega_\tau^i$$

thus X_t and $\tilde{X}_t$ solve the same CDE and coincide at time $t = s$. This means, by uniqueness, that X_t and $\tilde{X}_t$ coincide for all times; hence $W_{s,t}W_{r,s}$ and $W_{r,t}$ coincide for all times too since for any choice of X_0 one has $W_{s,t}W_{r,s}X_0 = W_{r,t}X_0$.

The second statement follows from the previous one setting first $r = t$ and subsequently exchanging s and t.

To prove the third equality note that

$$0 = d_s(W_{s,t}W_{t,s}) = (d_sW_{s,t})W_{t,s} + W_{s,t}(d_sW_{t,s})$$

hence

$$d_sW_{s,t} = -W_{s,t}(d_sW_{t,s})W_{t,s}^{-1} = -W_{s,t}(\sum_{i=1}^{d} A_s^i W_{t,s} d\omega_s^i)W_{t,s}^{-1} = -\sum_{i=1}^{d} W_{s,t}A_s^i d\omega_s^i$$

□

Proposition C.17 (Liouville's Formula)**.** *Under the assumptions of the previous theorems, if* $\omega \in C^1([0,1];\mathbb{R}^d)$ *then*

$$\det(W_{s,t}) = 1 + \sum_{i=1}^{d} \int_{\tau=s}^{t} tr(A_\tau^i)\det(W_{s,\tau})d\omega_\tau^i = \exp\left(\sum_{i=1}^{d}\int_{\tau=s}^{t} tr(A_\tau^i)d\omega_\tau^i\right) \tag{C.22}$$

Proof. This just follows from the classical case since we can write

$$\sum_{i=1}^{d}\int_{\tau=s}^{t} A_\tau^i W_{s,\tau} d\omega_\tau^i = \int_{\tau=s}^{t}\left(\sum_{i=1}^{d} A_\tau^i \dot{\omega}_\tau^i\right) W_{s,\tau} d\tau$$

thus Liouville's Formula gives us

$$\det(W_{s,t}) = \det(W_{s,s})\exp\left(\int_{\tau=s}^{t} tr(\sum_{i=1}^{d} A_\tau^i \dot{\omega}_\tau^i)d\tau\right).$$

□

We can now state the main result of the section:

Theorem C.18. *Under the assumptions of the previous theorems, given continuous functions* $\{B^1,\dots,B^t\} \in (\mathbb{R}^d)^{[0,1]}$ *the unique solution of the N-dimensional CDE*

$$X_t = X_0 + \sum_{i=1}^{d}\int_{\tau=0}^{t}\left(A_\tau^i X_\tau + B_\tau^i\right)d\omega_\tau^i \tag{C.23}$$

is given explicitly *by*

$$X_t = W_{0,t}X_0 + \sum_{i=1}^{d}\int_0^t W_{s,t}B_s^i d\omega_s^i \tag{C.24}$$

where $W_{s,t} \in C^0([0,1]\times[0,1];\mathbb{R}^{N\times N})$ *is the Wronskian matrix defined by*

$$W_{s,t} = \sum_{I\in\mathbf{W}_d} \int_{\tau\in\Delta^{|I|}_{[s,t]}} A^I_\tau d\omega^I_\tau \tag{C.25}$$

Proof. Given the unique solution X_t one has

$$\begin{aligned} d_s(W_{s,t}X_s) =& d_s(W_{s,t})X_s + W_{s,t}d_s(X_s) \\ =& \sum_{i=1}^{d} \left(-W_{s,t}A^i_s X_s + W_{s,t}A^i_s X_s + W_{s,t}B^i_s\right) d\omega^i_s \\ =& \sum_{i=1}^{d} W_{s,t}B^i_s d\omega^i_s \end{aligned}$$

hence

$$X_t - W_{0,t}X_0 = W_{t,t}X_t - W_{0,t}X_0 = \sum_{i=1}^{d}\int_{s=0}^{t} W_{s,t}B^i_s d\omega^i_s$$

□

C.4.1 ZOH and Exact Solutions

Consider the linear CDE

$$dZ_t = \sum_{i=1}^{d_\omega} A_i Z_t d\omega^i_t + B d\xi_t$$

and recall how the solution can be explicitly written, for times $s<t$, as

$$Z_t = W_{s,t}Z_s + \int_s^t W_{r,t}B d\xi_r$$

Assume moreover that in the interval $[s,t]$ both drivers have constant derivative

$$\omega_r = \omega_s + \boldsymbol{w}(r-s) \quad \xi_r = \xi_s + \boldsymbol{v}(r-s)$$

Then if $\mathbb{A}_{\boldsymbol{w}} := \sum_{i=1}^{d_\omega} A_i \boldsymbol{w}^i$ we get that $W_{r,t} = e^{\mathbb{A}_{\boldsymbol{w}}(t-r)}$ thus

$$Z_t = e^{\mathbb{A}_{\boldsymbol{w}}(t-s)}Z_s + \int_s^t e^{\mathbb{A}_{\boldsymbol{w}}(t-r)}B\boldsymbol{v}dr = e^{\mathbb{A}_{\boldsymbol{w}}(t-s)}Z_s + \left(\int_s^t e^{\mathbb{A}_{\boldsymbol{w}}(t-r)}dr\right)B\boldsymbol{v} \tag{C.26}$$

But the integral can be explicitly solved as

$$\int_s^t e^{\mathbb{A}_{\boldsymbol{w}}(t-r)}dr = \left(-\mathbb{A}^{-1}_{\boldsymbol{w}}e^{\mathbb{A}_{\boldsymbol{w}}(t-r)}\Big|^t_{r=s} = \mathbb{A}^{-1}_{\boldsymbol{w}}\left(e^{\mathbb{A}_{\boldsymbol{w}}(t-s)} - \mathbb{I}\right)\right. \tag{C.27}$$

leaving us with

$$Z_t = e^{\mathbb{A}_{\boldsymbol{w}}(t-s)}Z_s + \mathbb{A}^{-1}_{\boldsymbol{w}}\left(e^{\mathbb{A}_{\boldsymbol{w}}(t-s)} - \mathbb{I}\right)B\boldsymbol{v} \tag{C.28}$$

which, setting $\Delta = t - s$, can be rewritten as

$$Z_t = e^{\mathbb{A}_{\boldsymbol{w}}\Delta} Z_s + (\mathbb{A}_{\boldsymbol{w}}\Delta)^{-1} \left(e^{\mathbb{A}_{\boldsymbol{w}}\Delta} - \mathbb{I}\right) (B\Delta)\boldsymbol{v} \tag{C.29}$$

i.e. what is known as the ZOH scheme.

Appendix D

Graph Expansion for Neural CDEs

In this chapter we will give *formal* proofs of the GP limit and NTK results for the case of Neural Controlled Differential equations with arbitrary MLPs as vector fields. The reason for the formal nature of such results is the presence of infinite sums coming from the expansion in terms of *branched* signatures. Similar expansions have been studied by Gubinelli [47] in a simpler setting, rigorously proving their absolute convergence remains a challenging open question.

D.1 Preliminaries

We are now interested in using the genus expansion technique to study NCDEs of the form

$$Z_t^\omega = \mathbf{z}_0 + \sum_{i=1}^{d} \int_0^t \Phi_{i;L_i}(Z_s^\omega) \mathrm{d}\omega_s^i \tag{D.1}$$

where $\Phi_{1;L_1}, \dots, \Phi_{d;L_d}$ are generic (polynomial) MLPs of possibly different depths. As done in chapter 2 for the MLP case we want to expand this into a sum over graphs splitting the parameter-dependent from the input dependent part.

Imagine then to have an expansion for $Z_s^\omega = \sum_{g \in \mathcal{G}} \alpha_g(\omega|_{0,s}) \mathbf{W}_g$ at each time s, note how this has to be consistent with eq. (D.1) and how when expanding the NNs gives $\Phi_L(\mathbf{x}) = \sum_{\tau \in \mathbb{T}_L(\bullet_\mathbf{x})} \frac{\varphi_\tau}{s(\tau)} \mathbf{W}_\tau$. Substituting $\sum_g \alpha_g(\omega|_{0,s}) \mathbf{W}_g$ for $\mathbf{x}$ in this latter expansion we conclude then that the graphs of $\mathcal{G}$ must have a nested structure: each of them must be written as a $\tau \in \mathbb{T}_L(\bullet_\mathbf{x})$ where the vertices fixed to $\mathbf{x}$ are replaced by arbitrary graphs $g_1, \dots, g_k$ coming from $\mathcal{G}$ itself!

This nested structure will play a crucial role, this preliminary section will present combinatorial results in order to later meaningfully and conveniently talk about it.

Recall how by theorem B.4

Proposition D.1. *Let $\varphi, \psi : \mathbb{R} \to \mathbb{R}$ be polynomials. Fix index sets $\mathcal{A}_1, \mathcal{A}_2$ and*

$\lambda : (\mathcal{A}_1 \cup \mathcal{A}_2) \times (\mathcal{A}_1 \cup \mathcal{A}_2) \to \mathbb{R}$ *a symmetric and* summable *function. Then if*

$$(X, Y) \sim \mathcal{N}\left(0, \begin{bmatrix} \sum_{(a,a') \in \mathcal{A}_1 \times \mathcal{A}_1} \lambda(a, a') & \sum_{(a,b) \in \mathcal{A}_1 \times \mathcal{A}_2} \lambda(a, b) \\ \sum_{(b,a) \in \mathcal{A}_2 \times \mathcal{A}_1} \lambda(b, a) & \sum_{(b,b') \in \mathcal{A}_2 \times \mathcal{A}_2} \lambda(b, b') \end{bmatrix}\right)$$

the following expansion holds

$$\mathbb{E}\{\varphi(X)\psi(Y)\} = \sum_{\boldsymbol{\tau} \in \mathbb{X}_{\mathcal{A}_1}} \sum_{\boldsymbol{\eta} \in \mathbb{X}_{\mathcal{A}_2}} \frac{\varphi_{\boldsymbol{\tau}} \psi_{\boldsymbol{\eta}}}{\mathfrak{s}(\boldsymbol{\tau})\mathfrak{s}(\boldsymbol{\eta})} \sum_{\pi \in \mathcal{P}(\boldsymbol{\tau} \sqcup \boldsymbol{\eta})} \prod_{\{\alpha,\beta\} \in \pi} \lambda(\alpha, \beta) \tag{D.2}$$

where $\varphi_{[\![\alpha_1 \cdots \alpha_m]\!]} := \varphi^{(m)}(0)$ *and* $\mathcal{P}(\boldsymbol{\tau} \sqcup \boldsymbol{\eta})$ *is the set of pairings of entries of* $\boldsymbol{\tau}$ *and* $\boldsymbol{\eta}$*.*

Our immediate goal is that of extending this result to a nested expectation setting. It will be thus important to consider nested sequences of symmetric tuples, this structure is naturally encoded in *non-planar* rooted trees. We define these as

Definition D.2. *Given a set* $\mathcal{A}$ *we define the set of depth-*ℓ *non-planar rooted trees with* $\mathcal{A}$*-decorated terminal leaves* $\mathbf{T}_\ell(\mathcal{A})$ *recursively as*

$$\mathbf{T}_0(\mathcal{A}) = \mathcal{A}, \quad \mathbf{T}_{\ell+1}(\mathcal{A}) = \mathbb{X}_{\mathbf{T}_\ell(\mathcal{A})} \tag{D.3}$$

Finally we define the set of non-planar rooted trees with $\mathcal{A}$*-decorated terminal leaves as the union* $\mathbf{T}(\mathcal{A}) := \cup_{\ell=0}^{\infty} \mathbf{T}_\ell(\mathcal{A})$*. For any tree* $\boldsymbol{\tau} = [\![(\boldsymbol{\tau}_1)^{k_1} \cdots (\boldsymbol{\tau}_N)^{k_N}]\!] \in \mathbf{T}_{\ell+1}(\mathcal{A})$*, where* $\boldsymbol{\tau}_1, \ldots, \boldsymbol{\tau}_N$ *are distinct elements of* $\mathbf{T}_\ell(\mathcal{A})$*, we define the* symmetry factor $\mathbf{s}$ *of* $\boldsymbol{\tau}$ *as*

$$\mathbf{s}([\![\ ]\!]) = 1, \quad \mathbf{s}(\boldsymbol{\tau}) := \mathfrak{s}(\boldsymbol{\tau}) \prod_{i=1}^{N} \mathbf{s}(\boldsymbol{\tau}_i)^{k_i} \tag{D.4}$$

where for all $a \in \mathcal{A}$ *we set* $\mathbf{s}(a) = 1$*. We also write* $\boldsymbol{\tau} = \boldsymbol{\tau}[\![a_1, \ldots, a_m]\!]$ *to show the fact that* $\boldsymbol{\tau}$ *has exactly* m *leaves from* $\mathcal{A}$ *and that these are* $a_1, \ldots, a_m$*.*

The choice of nomenclature becomes clear after noticing how naturally nested symmetric tuples are graphically represented by non-planar trees *c.f.* fig. D.1.

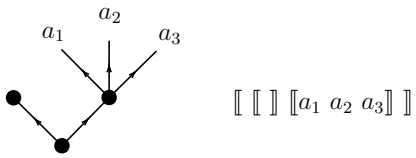


Figure D.1: *A simple example of the correspondance between trees and nested symmetric tuples.*

Definition D.3. *Given sets* $\mathcal{A}_1, \mathcal{A}_2$ *and* $\ell \in \mathbb{N}_0$ *we define the* nested *pairings of the terminal leaves of two trees* $\boldsymbol{\tau} \in \mathbf{T}_\ell(\mathcal{A}_1)$ *and* $\boldsymbol{\eta} \in \mathbf{T}_\ell(\mathcal{A}_2)$ *as the set* $\mathcal{P}_{\text{nest}}(\boldsymbol{\tau} \sqcup \boldsymbol{\eta})$ *recursively defined as*

$$\forall (a, a') \in \mathcal{A}_1 \times \mathcal{A}_2.\ \mathcal{P}_{\text{nest}}(a \sqcup a') = \{\ \{\{a, a'\}\}\ \} \tag{D.5}$$

and for $\boldsymbol{\tau} = [\![\boldsymbol{\tau}_1 \cdots \boldsymbol{\tau}_n]\!] \in \mathbf{T}_{\ell+1}(\mathcal{A}_1)$ *and* $\boldsymbol{\eta} = [\![\boldsymbol{\eta}_1 \cdots \boldsymbol{\eta}_m]\!] \in \mathbf{T}_{\ell+1}(\mathcal{A}_2)$ *as*

$$\mathcal{P}_{\text{nest}}(\boldsymbol{\tau} \sqcup \boldsymbol{\eta}) = \bigsqcup_{\pi \in \mathcal{P}(\boldsymbol{\tau} \sqcup \boldsymbol{\eta})} \bigotimes_{\{\boldsymbol{\pi}_1, \boldsymbol{\pi}_2\} \in \pi} \mathcal{P}_{\text{nest}}(\boldsymbol{\pi}_1 \sqcup \boldsymbol{\pi}_2) \tag{D.6}$$

where $\mathcal{P}(\boldsymbol{\tau} \sqcup \boldsymbol{\eta})$ *is the set of pairings of the entries of* $\boldsymbol{\tau}$ *and* $\boldsymbol{\eta}$ *as symmetric tuples, and where for sets* A, B

$$A \otimes B := \{a \sqcup b \mid \forall a \in A, \forall b \in B\}$$

Remark. *The definition of* $\mathcal{P}_{\text{nest}}(\boldsymbol{\tau} \sqcup \boldsymbol{\eta})$ *implies the following combinatorial equality*

$$\sum_{\phi \in \mathcal{P}_{\text{nest}}(\boldsymbol{\tau} \sqcup \boldsymbol{\eta})} \alpha_\phi = \sum_{\pi \in \mathcal{P}(\boldsymbol{\tau} \sqcup \boldsymbol{\eta})} \prod_{\{\boldsymbol{\pi}_1, \boldsymbol{\pi}_2\} \in \pi} \sum_{\phi' \in \mathcal{P}_{\text{nest}}(\boldsymbol{\pi}_1 \sqcup \boldsymbol{\pi}_2)} \alpha_{\phi'} \tag{D.7}$$

where $\{\alpha_\phi\}_{\phi \in \mathcal{P}_{\text{nest}}(\boldsymbol{\tau} \sqcup \boldsymbol{\eta})}$ *is a subset of some arbitrary vector space.*

Remark. *Note how if* $\boldsymbol{\tau} = \boldsymbol{\tau}[\![a_1, \ldots, a_m]\!]$ *and* $\boldsymbol{\eta} = \boldsymbol{\eta}[\![a'_1, \ldots, a'_n]\!]$ *then any element of* $\mathcal{P}_{\text{nest}}(\boldsymbol{\tau} \sqcup \boldsymbol{\eta})$ *is a pairing of* $(a_1, \ldots, a_m, a'_1, \ldots, a'_n)$*. On the other hand, a pairing of* $(a_1, \ldots, a_m, a'_1, \ldots, a'_n)$ *might appear "multiple times" in* $\mathcal{P}_{\text{nest}}(\boldsymbol{\tau} \sqcup \boldsymbol{\eta})$ e.g.

$$\mathcal{P}_{\text{nest}}([\![\ [\![\ ]\!]\ [\![\ ]\!]\ [\![a]\!]\ ]\!] \sqcup [\![\ [\![\ ]\!]\ [\![\ ]\!]\ [\![b]\!]\ ]\!]) = \{\ \{\{a, b\}\}\ \} \sqcup \{\ \{\{a, b\}\}\ \} \sqcup \{\ \{\{a, b\}\}\ \}$$

In fact $\mathcal{P}_{\text{nest}}$ *is encoding information about the pairing of* all *subtrees of* $\boldsymbol{\tau} \sqcup \boldsymbol{\eta}$*, not only of the labelled leaves.*

Definition D.4. *Let* $(\varphi_\ell : \mathbb{R} \to \mathbb{R} \mid \ell \in \mathbb{N})$ *be a sequence of polynomials and* $\mathcal{A}$ *a set. We define the tree-derivatives of the* φ *on* $\mathbf{T}(\mathcal{A})$ *as*

$$\forall a \in \mathcal{A}.\ \varphi_a := 1, \quad \forall \boldsymbol{\tau} = [\![\boldsymbol{\tau}_1 \cdots \boldsymbol{\tau}_m]\!] \in \mathbf{T}_{\ell+1}(\mathcal{A}).\ \varphi_{\boldsymbol{\tau}} := \varphi^{(m)}_{\ell+1}(0) \prod_{i=1}^{m} \varphi_{\boldsymbol{\tau}_i} \tag{D.8}$$

Theorem D.5. *Let* $((\varphi_\ell, \psi_\ell) \mid \ell \in \mathbb{N})$ *be a sequence of tuples of polynomials. Fix index sets* $\mathcal{A}_x, \mathcal{A}_y$ *and* $\lambda : (\mathcal{A}_x \cup \mathcal{A}_y) \times (\mathcal{A}_x \cup \mathcal{A}_y) \to \mathbb{R}$ *a symmetric and* summable *function. Then if*

$$\begin{bmatrix} K_0(x,x) & K_0(x,y) \\ K_0(y,x) & K_0(y,y) \end{bmatrix} = \begin{bmatrix} \sum_{(a,a') \in \mathcal{A}_x \times \mathcal{A}_x} \lambda(a,a') & \sum_{(a,b) \in \mathcal{A}_x \times \mathcal{A}_y} \lambda(a,b) \\ \sum_{(b,a) \in \mathcal{A}_y \times \mathcal{A}_x} \lambda(b,a) & \sum_{(b,b') \in \mathcal{A}_y \times \mathcal{A}_y} \lambda(b,b') \end{bmatrix}$$

and we define recursively

$$K_{\ell+1}(x,y) := \mathbb{E}\{\varphi_{\ell+1}(X_\ell)\psi_{\ell+1}(Y_\ell)\}, \quad (X_\ell, Y_\ell) \sim \mathcal{N}\left(0, \begin{bmatrix} K_\ell(x,x) & K_\ell(x,y) \\ K_\ell(y,x) & K_\ell(y,y) \end{bmatrix}\right)$$

the following expansion holds

$$K_\ell(x,y) = \sum_{\boldsymbol{\tau}\in\mathbf{T}_\ell(\mathcal{A}_x)} \sum_{\boldsymbol{\eta}\in\mathbf{T}_\ell(\mathcal{A}_y)} \frac{\varphi_{\boldsymbol{\tau}}\psi_{\boldsymbol{\eta}}}{\mathbf{s}(\boldsymbol{\tau})\mathbf{s}(\boldsymbol{\eta})} \sum_{\phi\in\mathcal{P}_{\mathrm{nest}}(\boldsymbol{\tau}\sqcup\boldsymbol{\eta})} \prod_{\{\alpha,\beta\}\in\phi} \lambda(\alpha,\beta) \tag{D.9}$$

Proof. By induction on ℓ. If $\ell = 0$ the thesis is clear by definition of the constituent objects. Assume then expansion true for ℓ and let's show it for $\ell+1$. By theorem D.1 we have

$$\mathbb{E}\{\varphi_{\ell+1}(X_\ell)\psi_{\ell+1}(Y_\ell)\} = \sum_{\boldsymbol{\tau}\in\mathbb{X}_{\mathbf{T}_\ell(\mathcal{A}_x)}} \sum_{\boldsymbol{\eta}\in\mathbb{X}_{\mathbf{T}_\ell(\mathcal{A}_y)}} \frac{(\varphi_{\ell+1})_{\boldsymbol{\tau}}(\psi_{\ell+1})_{\boldsymbol{\eta}}}{\mathfrak{s}(\boldsymbol{\tau})\mathfrak{s}(\boldsymbol{\eta})} \sum_{\pi\in\mathcal{P}(\boldsymbol{\tau}\sqcup\boldsymbol{\eta})} \prod_{\{\pi_1,\pi_2\}\in\pi} \lambda_\ell(\pi_1,\pi_2)$$

where for $\boldsymbol{\tau}\in\mathbf{T}_\ell(\mathcal{A}_x)$ and $\boldsymbol{\eta}\in\mathbf{T}_\ell(\mathcal{A}_y)$ one defines

$$\lambda_\ell(\boldsymbol{\tau},\boldsymbol{\eta}) := \frac{\varphi_{\boldsymbol{\tau}}\psi_{\boldsymbol{\eta}}}{\mathbf{s}(\boldsymbol{\tau})\mathbf{s}(\boldsymbol{\eta})} \sum_{\phi\in\mathcal{P}_{\mathrm{nest}}(\boldsymbol{\tau}\sqcup\boldsymbol{\eta})} \prod_{\{\alpha,\beta\}\in\phi} \lambda(\alpha,\beta)$$

and similarly if $\boldsymbol{\tau},\boldsymbol{\eta}$ live in the same space then in this definition either $\varphi\mapsto\psi$ or vice-versa.

Putting things together we notice, from the definitions of $\mathbf{T}_\ell(\mathcal{A})$, s and $\varphi_{\boldsymbol{\tau}}$, that

$$K_{\ell+1}(x,y) = \sum_{\boldsymbol{\tau}\in\mathbf{T}_{\ell+1}(\mathcal{A}_x)} \sum_{\boldsymbol{\eta}\in\mathbf{T}_{\ell+1}(\mathcal{A}_y)} \frac{\varphi_{\boldsymbol{\tau}}\psi_{\boldsymbol{\eta}}}{\mathbf{s}(\boldsymbol{\tau})\mathbf{s}(\boldsymbol{\eta})} \sum_{\pi\in\mathcal{P}(\boldsymbol{\tau}\sqcup\boldsymbol{\eta})} \prod_{\{\pi_1,\pi_2\}\in\pi} \sum_{\phi\in\mathcal{P}_{\mathrm{nest}}(\pi_1\sqcup\pi_2)} \prod_{\{\alpha,\beta\}\in\phi} \lambda(\alpha,\beta)$$

and the conclusion follows using the equality

$$\sum_{\phi\in\mathcal{P}_{\mathrm{nest}}(\boldsymbol{\tau}\sqcup\boldsymbol{\eta})} \prod_{\{\alpha,\beta\}\in\phi} \lambda(\alpha,\beta) = \sum_{\pi\in\mathcal{P}(\boldsymbol{\tau}\sqcup\boldsymbol{\eta})} \prod_{\{\pi_1,\pi_2\}\in\pi} \sum_{\phi\in\mathcal{P}_{\mathrm{nest}}(\pi_1\sqcup\pi_2)} \prod_{\{\alpha,\beta\}\in\phi} \lambda(\alpha,\beta)$$

□

D.1.1 Tree expansion for MLPs

Since the vector fields defining the NCDE are by definition MLPs it will be useful to first have an expansion for these, we recall here the results of chapter 2, contextualized in the richer language just developed.

Remark. *Given sequences* $(W_\ell \in \mathbb{R}^{N_{\ell+1}\times N_\ell} \mid \ell\in\mathbb{N}_0)$ *and* $(\mathbf{x}_a \mid a\in\mathcal{A})$ *there is a natural way to associate to each* $\boldsymbol{\tau}\in\mathbf{T}(\mathcal{A})$ *an operator graph* τ*:*

$$\forall a\in\mathcal{A}.\ a\mapsto \circ\, -_{W_0}\, \bullet_{\mathbf{x}_a}, \quad \forall \boldsymbol{\tau} = [\![\boldsymbol{\tau}_1\cdots\boldsymbol{\tau}_m]\!]\in\mathbf{T}_{\ell+1}(\mathcal{A}).\ \boldsymbol{\tau}\mapsto (\circ\, -_{W_{\ell+1}}\, \circ)\circ(\wedge_i\boldsymbol{\tau}_i)$$

We will use the notation $\boldsymbol{\tau}\mapsto\tau$ (i.e. *un-bolding the greek letter) for this map.*

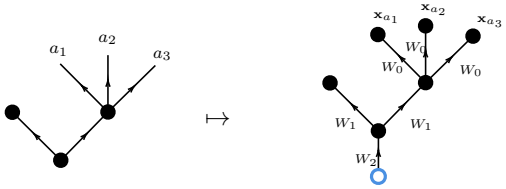


Figure D.2: *The mapping between the tree* $[\![\, [\![\,]\!]\, [\![a_1\ a_2\ a_3]\!]\,]\!] \in \mathbf{T}_2(\mathcal{A})$ *and its operator-graph image.*

Proposition D.6. *Given a (polynomial) NNs* $\Phi_L : \mathbb{R}^N \to \mathbb{R}^N$ *and* $\mathbf{x} = \sum_{a\in\mathcal{A}} \lambda_a \mathbf{x}_a$ *for some summable sequence* $(\mathbf{x}_a \in \mathbb{R}^N \mid a \in \mathcal{A})$

$$\Phi_L(\mathbf{x}) = \sum_{\boldsymbol{\tau}[\![a_1,\dots,a_m]\!]\in\mathbf{T}_L(\mathcal{A})} \frac{\varphi_{\boldsymbol{\tau}}}{\mathbf{s}(\boldsymbol{\tau})} \left(\prod_i \lambda_{a_i}\right) \mathbf{W}_\tau$$

Proof. We proceed by induction on L. For $L = 0$ by definition we have $\Phi_0(\mathbf{x}) = W_0\mathbf{x}$, $\mathbf{T}_L(\mathcal{A}) = \mathcal{A}$ and for each $a \in \mathcal{A}$ also $\varphi_a = 1, \boldsymbol{s}(a) = 1$ and $\mathbf{W}_a = W_0[\mathbf{x}_a]$. Hence

$$\sum_{a\in\mathbf{T}_0(\mathcal{A})} \frac{\varphi_a}{\mathbf{s}(a)} \lambda_a \mathbf{W}_a = \sum_{a\in\mathbf{T}_0(\mathcal{A})} W_0[\lambda_a \mathbf{x}_a] = W_0\mathbf{x} = \Phi_0(\mathbf{x}).$$

Reasoning by induction and using the Taylor expansion in 0 we have

$$\begin{aligned}
\Phi_{L+1}(\mathbf{x}) &= W_{L+1}[\varphi_{L+1}(\Phi_L(\mathbf{x}))] \\
&= \sum_{m=0}^{\infty} \frac{\varphi_{L+1}^{(m)}(0)}{M!} W_{L+1}[\left(\sum_{\boldsymbol{\tau}[\![a_1,\dots,a_m]\!]\in\mathbf{T}_L(\mathcal{A})} \frac{\varphi_{\boldsymbol{\tau}}}{\mathbf{s}(\boldsymbol{\tau})} \left(\prod_i \lambda_{a_i}\right) \mathbf{W}_{\boldsymbol{\tau}}\right)^{\odot m}] \\
&= \sum_{[\![\boldsymbol{\tau}_1\cdots\boldsymbol{\tau}_m]\!]\in\mathbb{X}_{\mathbf{T}_L(\mathcal{A})}} \frac{(\varphi_{L+1})_{[\![\boldsymbol{\tau}_1\cdots\boldsymbol{\tau}_m]\!]}}{\mathfrak{s}([\![\boldsymbol{\tau}_1\cdots\boldsymbol{\tau}_m]\!])} W_{L+1}[\odot_{i=1}^m \frac{\varphi_{\boldsymbol{\tau}_i}}{\mathbf{s}(\boldsymbol{\tau}_i)} \left(\prod_j \lambda_{a_{i,j}}\right) \mathbf{W}_{\boldsymbol{\tau}_i}]
\end{aligned}$$

where we have used the equality

$$\left(\sum_{a\in\mathcal{A}} v_a\right)^{\odot M} = \sum_{\boldsymbol{\tau}\in\mathbb{X}_{\mathcal{A}}^M} \frac{M!}{\mathfrak{s}(\boldsymbol{\tau})} v_{\boldsymbol{\tau}}$$

from theorem B.3. We can continue the equality using the definition of $\mathbf{T}_{L+1}(\mathcal{A})$ obtaining

$$\begin{aligned}
&= \sum_{\boldsymbol{\tau}=[\![\boldsymbol{\tau}_1\cdots\boldsymbol{\tau}_m]\!]\in\mathbf{T}_{L+1}(\mathcal{A})} \frac{\varphi_{\boldsymbol{\tau}}}{\mathbf{s}(\boldsymbol{\tau})} W_{L+1}[\odot_{i=1}^m \left(\prod_j \lambda_{a_{i,j}}\right) \mathbf{W}_{\tau_i}] \\
&= \sum_{\boldsymbol{\tau}[\![a_1\cdots a_m]\!]\in\mathbf{T}_{L+1}(\mathcal{A})} \frac{\varphi_{\boldsymbol{\tau}}}{\mathbf{s}(\boldsymbol{\tau})} \left(\prod_i \lambda_{a_i}\right) \mathbf{W}_\tau
\end{aligned}$$

□

D.2 Graph Expansion

In this section we expand in terms of operator-graphs a generic Neural Controlled Differential Equation with initial value $\mathbf{z}_0 \in \mathbb{R}^N$

$$Z_t^\omega = \mathbf{z}_0 + \sum_{i=1}^{d} \int_0^t \Phi_{i;L_i}(Z_t^\omega) \mathrm{d}\omega_t^i \tag{D.10}$$

defined by (polynomial) MLPs $\Phi_{1;L_1}, \dots, \Phi_{d;L_d} : \mathbb{R}^N \to \mathbb{R}^N$.

For each $\Phi_{i;L_i}$ let $\mathbb{T}_{i;L_i}(\bullet_{\mathbf{x}})$ be the set of trees appearing in its graph expansion, then for all $m \in \mathbb{N}$ fix $\mathbb{T}_{i;L_i}[\circ_1, \dots, \circ_m]$ to be the set of operator-graphs obtained from the trees in $\mathbb{T}_{i;L_i}(\bullet_{\mathbf{x}})$ with exactly m $\bullet_{\mathbf{x}}$-labelled leaves by replacing each $\bullet_{\mathbf{x}}$-vertex with a unique $\circ_i$. Note that $\bigcup_{m\geq 0} \mathbb{T}_{i;L_i}[\circ_1, \dots, \circ_m] \twoheadrightarrow \mathbb{T}_{i;L_i}(\bullet_{\mathbf{x}})$.

Definition D.7 (CDE Trees). *Given (polynomial) NNs $\Phi_{1;L_1}, \dots, \Phi_{d;L_d} : \mathbb{R}^N \to \mathbb{R}^N$, define $\mathbb{T}^{\mathrm{CDE}}(\bullet_{\mathbf{z}_0})$ as the smallest set of operator graphs containing $\circ -_{Id} \bullet_{\mathbf{z}_0}$ and closed by* full *composition with trees in $\bigcup_{i,m} \mathbb{T}_{i;L_i}[\circ_1, \dots, \circ_m]$* i.e. *if $\eta_1, \dots, \eta_m \in \mathbb{T}^{\mathrm{CDE}}(\bullet_{\mathbf{z}_0})$ and $\hat{\eta} \in \mathbb{T}_{i;L_i}[\circ_1, \dots, \circ_m]$ then $\hat{\eta}[\eta_1, \dots, \eta_m] \in \mathbb{T}^{\mathrm{CDE}}$ where $\hat{\eta}[\eta_1, \dots, \eta_m]$ is produced by, for each $k = 1, \dots, m$, composing the (out-)root of η_i with $\hat{\eta}$ at its in-vertex $\circ_i$.*

Note how there is a natural bijection

$$\mathbb{T}^{\mathrm{CDE}}(\bullet_{\mathbf{z}_0}) \leftrightarrow \{\circ -_{Id} \bullet_{\mathbf{z}_0}\} \bigsqcup_{i=1}^{d} \mathbf{T}_{L_i}(\mathbb{T}^{\mathrm{CDE}}(\bullet_{\mathbf{z}_0})) \tag{D.11}$$

By this we mean that in case $\tau \in \mathbb{T}^{\mathrm{CDE}}(\bullet_{\mathbf{z}_0})$ is written as $\hat{\tau}[\tau_1, \dots, \tau_m]$ for $\tau_1, \dots, \tau_m \in \mathbb{T}^{\mathrm{CDE}}(\bullet_{\mathbf{z}_0})$ and $\hat{\tau} \in \mathbb{T}_{i;L_i}[\circ_1, \dots, \circ_m]$, then $\tau \simeq \boldsymbol{\tau} = \boldsymbol{\tau}[\![\tau_1 \cdots \tau_m]\!] \in \mathbf{T}_{i;L_i}(\mathbb{T}^{\mathrm{CDE}}(\bullet_{\mathbf{z}_0}))$. For simplicity we write $\mathbf{T}_{i;L_i} := \mathbf{T}_{L_i}(\mathbb{T}^{\mathrm{CDE}}(\bullet_{\mathbf{z}_0}))$.

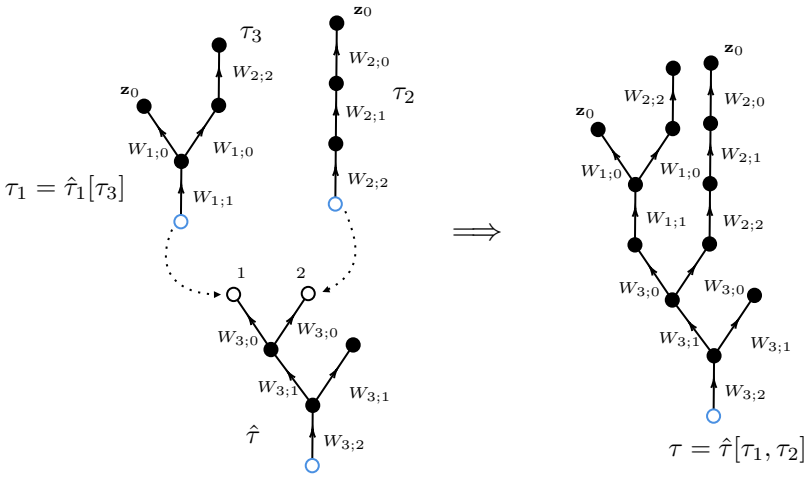


Figure D.3: *An example of tree $\tau \in \mathbb{T}^{\mathrm{CDE}}(\bullet_{\mathbf{z}_0})$ and its construction as composition of trees.*

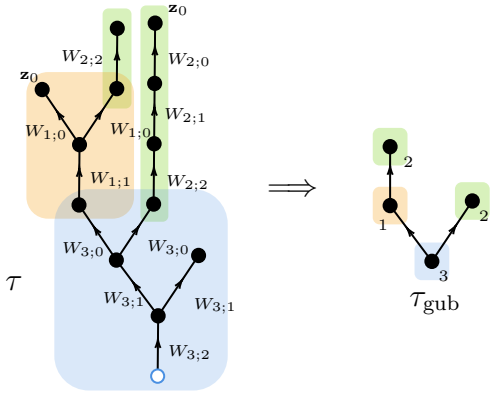
Moreover, under such a bijection, for $\tau \simeq \boldsymbol{\tau} = \boldsymbol{\tau}[\![\tau_1, \dots, \tau_m]\!] \in \mathbf{T}_{i;L_i}$ one has

$$\varphi_\tau = \varphi_{\boldsymbol{\tau}} \prod_{i=1}^{m} \varphi_{\tau_i}, \quad s(\tau) = \mathbf{s}(\boldsymbol{\tau}) \prod_i s(\tau_i)$$

Note finally how we can associate to each $\tau \in \mathbb{T}^{\mathrm{CDE}}(\bullet_{\mathbf{z}_0})$ a tree $\tau_{\mathrm{gub}} \in \mathcal{T}_d$ (*cf.* [47]) corresponding to the "tree-of-trees" structure of τ inherited from the definition. This is defined recursively for $\tau \simeq \boldsymbol{\tau} = \boldsymbol{\tau}[\![\tau_1 \cdots \tau_m]\!] \in \mathbf{T}_{i;L_i}$ as $\tau_{\mathrm{gub}} = [(\tau_1)_{\mathrm{gub}} \cdots (\tau_m)_{\mathrm{gub}}]$ where $(\circ -_{Id} \bullet_{\mathbf{z}_0})_{\mathrm{gub}}$ is the empty tree. Then if we define

$$\mathrm{Sig}(\omega)^{\tau}_{0,t} := \int_0^t \prod_{j=1}^{m} \mathrm{Sig}(\omega)^{\tau_j}_{0,s} \mathrm{d}\omega^i_s, \quad \mathrm{Sig}(\omega)^{\circ -_{Id} \bullet_{\mathbf{z}_0}}_{0,t} := 1$$

we recover the *branched signature* entries of ω corresponding to $\tau_{\mathrm{gub}} \in \mathcal{T}_d$.



$$\mathrm{Sig}(\omega)^{\tau}_{0,t} = \int_0^t \left(\int_0^s \left(\int_0^r \mathrm{d}\omega^2_u\right) \mathrm{d}\omega^1_r\right) \left(\int_0^s \mathrm{d}\omega^2_r\right) \mathrm{d}\omega^3_s$$

Figure D.4: *A color-coded example of how a tree τ, taken from fig. D.3, gives rise to τ_{gub} and to the corresponding $\mathrm{Sig}(\omega)^{\tau}_{0,t}$*

Theorem D.8 (Formal NCDE Expansion)**.** *Given (polynomial) NNs $\Phi_{1;L_1}, \dots, \Phi_{d;L_d} : \mathbb{R}^N \to \mathbb{R}^N$, consider the NCDE*

$$Z^\omega_t = \mathbf{z}_0 + \sum_{i=1}^{d} \int_0^t \Phi_{i;L_i}(Z^\omega_t) \mathrm{d}\omega^i_t \tag{D.12}$$

then we can write explicitly

$$Z^\omega_t = \sum_{\tau \in \mathbb{T}^{\mathrm{CDE}}(\bullet_{\mathbf{z}_0})} \frac{\varphi_\tau}{s(\tau)} \mathbf{W}_\tau \mathrm{Sig}(\omega)^{\tau}_{0,t} \tag{D.13}$$

Proof. We reason formally, only absolute summability of the series is needed to make the arguments rigorous. Let's prove that Z^ω_t thus defined solves the CDE. Using the

bijection of $\mathbb{T}^{\mathrm{CDE}}(\bullet_{\mathbf{z}_0})$ given above

$$\begin{aligned} Z_t^\omega &= \sum_{\tau \in \mathbb{T}^{\mathrm{CDE}}(\bullet_{\mathbf{z}_0})} \frac{\varphi_\tau}{s(\tau)} \mathbf{W}_\tau \mathrm{Sig}(\omega)_{0,t}^\tau \\ &= \mathbf{z}_0 + \sum_{i=1}^d \sum_{\boldsymbol{\tau}[\![\tau_1,\ldots,\tau_m]\!] \in \mathbf{T}_{i;L_i}} \frac{\varphi_\tau}{s(\tau)} \mathbf{W}_\tau \mathrm{Sig}(\omega)_{0,t}^\tau \\ &= \mathbf{z}_0 + \sum_{i=1}^d \int_0^t \sum_{\boldsymbol{\tau}[\![\tau_1,\ldots,\tau_m]\!] \in \mathbf{T}_{i;L_i}} \frac{\varphi_{\boldsymbol{\tau}}}{s(\boldsymbol{\tau})} \left(\prod_{j=1}^m \frac{\varphi_{\tau_j}}{s(\tau_j)} \mathrm{Sig}(\omega)_{0,s}^{\tau_j} \right) \mathbf{W}_\tau \mathrm{d}\omega_s^i. \end{aligned}$$

By theorem D.6 we see that

$$\sum_{\boldsymbol{\tau}[\![\tau_1,\ldots,\tau_m]\!] \in \mathbf{T}_{i;L_i}} \frac{\varphi_{\boldsymbol{\tau}}}{s(\boldsymbol{\tau})} \left(\prod_{j=1}^m \frac{\varphi_{\tau_j}}{s(\tau_j)} \mathrm{Sig}(\omega)_{0,s}^{\tau_j} \right) \mathbf{W}_\tau = \Phi_{i;L_i}(\mathbf{x})$$

where

$$\mathbf{x} = \sum_{\tau \in \mathbb{T}^{\mathrm{CDE}}(\bullet_{\mathbf{z}_0})} \frac{\varphi_\tau}{s(\tau)} \mathbf{W}_\tau \mathrm{Sig}(\omega)_{0,s}^\tau = Z_s^\omega.$$

Thus, as needed, we can write

$$Z_t^\omega = \mathbf{z}_0 + \sum_{i=1}^d \int_0^t \Phi_{i;L_i}(Z_s^\omega) \mathrm{d}\omega_s^i.$$

□

D.3 Assumptions

In what follows we will assume the following architecture for our Neural CDE:

Assumption D.9. *Assume given d sequences $\mathcal{W}_i = (W_{i;n} \mid n \geq 0)$ of independent $N \times N$ matrices with* real *Gaussian $\mathcal{N}(0, \frac{1}{\sqrt{N}})$ entries, a vector $\mathbf{z}_0 \in \mathbb{R}^N$, independent from the matrices, having iid Gaussian $\mathcal{N}(0,1)$ entries and a matrix $V \in \mathbb{R}^{N_{out} \times N}$ with* real *Gaussian $\mathcal{N}(0, \frac{1}{\sqrt{N}})$ entries.*

The sequences $\mathcal{W}_i$ define d (polynomial) NNs $\Phi_{1;L_1}, \ldots, \Phi_{d;L_d} : \mathbb{R}^N \to \mathbb{R}^N$ with all dimensions equal to N. Given a path $\omega \in C_1([0,1]; \mathbb{R}^d)$ we recall how the hidden state is defined as

$$Z_t^\omega = \mathbf{z}_0 + \sum_{i=1}^d \int_0^t \Phi_{i;L_i}(Z_t^\omega) d\omega_t^i. \tag{D.14}$$

D.4 GP limits

Lemma D.10. *Given $\tau, \eta \in \mathbb{T}^{\mathrm{CDE}}(\bullet_{\mathbf{z}_0})$, with $\tau \simeq \boldsymbol{\tau}[\![\tau_1, \dots, \tau_m]\!] \in \mathbf{T}_{i;L_i}$ and $\eta \simeq \boldsymbol{\eta}[\![\eta_1, \dots, \eta_n]\!] \in \mathbf{T}_{j;L_j}$ we have a bijection*

$$\mathcal{P}_{\mathrm{A}}(\langle \tau, \eta \rangle) \leftrightarrow \delta_{i,j} \bigsqcup_{\boldsymbol{\phi} \in \mathcal{P}_{\mathrm{nest}}(\boldsymbol{\tau} \sqcup \boldsymbol{\eta})} \bigotimes_{\{\pi_1, \pi_2\} \in \boldsymbol{\phi}} \mathcal{P}_{\mathrm{A}}(\langle \pi_1, \pi_2 \rangle)$$

which we write as $\phi \mapsto (\boldsymbol{\phi}, (\phi_{\boldsymbol{\pi}})_{\boldsymbol{\pi}})$, where $\{\pi_1, \pi_2\}$ is a pair in $(\tau_1, \dots, \eta_n)$ and using which one defines

$$a(\langle \tau, \eta \rangle) := |\mathcal{P}_{\mathrm{A}}(\langle \tau, \eta \rangle)| = \delta_{i,j} \sum_{\boldsymbol{\phi} \in \mathcal{P}_{\mathrm{nest}}(\boldsymbol{\tau} \sqcup \boldsymbol{\eta})} \prod_{\{\pi_1, \pi_2\} \in \boldsymbol{\phi}} a(\langle \pi_1, \pi_2 \rangle)$$

Proof. With the same arguments of theorems 2.10 and 2.15 one sees that any atomic pairing of $\langle \tau, \eta \rangle$ must pair edges at the same distance from the joined root. This means that any atomic pairing ϕ of $\langle \tau, \eta \rangle$ induces an atomic pairing $\hat{\phi}$ of the "base tree" $\langle \hat{\tau}, \hat{\eta} \rangle$ such that whenever the bases of two trees in $(\tau_1, \dots, \eta_n)$ are paired by $\hat{\phi}$ then ϕ induces an atomic pairing of those two trees. Vice versa, any choice of $\hat{\phi}$ and atomic pairings of the paired $(\tau_1, \dots, \eta_n)$ can be combined to form a unique atomic pairing ϕ of $\langle \tau, \eta \rangle$. The bijection then follows from the definition of $\mathcal{P}_{\mathrm{nest}}(\boldsymbol{\tau} \sqcup \boldsymbol{\eta})$, which is natural in bijection with $\mathcal{P}_{\mathrm{A}}(\langle \hat{\tau}, \hat{\eta} \rangle)$ giving the pairings of roots of $(\tau_1, \dots, \eta_n)$ induced by the atomic $\hat{\phi}$.

□

Lemma D.11. *Given any $\tau, \eta \in \mathbb{T}^{\mathrm{CDE}}(\bullet_{\mathbf{z}_0})$ one has*

$$\left\| \frac{1}{N} \mathbf{W}_{\langle \tau, \eta \rangle} - a(\langle \tau, \eta \rangle) \right\|_{L^2} = \mathcal{O}(\frac{1}{\sqrt{N}})$$

Proof. The proof is a repetition of the arguments used to prove theorem 3.4, where instead of one cycle to collapse generated from connecting only two terminal leaves one has to collapse the cycles generated from connecting *all* terminal leaves. □

Proposition D.12 (Formal Limiting Kernel)**.** *Under theorem D.9*

$$\frac{1}{N} \langle Z^{\omega}_s, Z^{\tilde{\omega}}_t \rangle \xrightarrow[N \to \infty]{L^2} K^{\omega, \tilde{\omega}}_{\mathrm{CDE}}(s, t) = 1 + \sum_{i=1}^{d} \int_{\sigma=0}^{s} \int_{\tau=0}^{t} K^{\omega, \tilde{\omega}}_{i;L_i}(\sigma, \tau) d\omega^i_\sigma d\tilde{\omega}^i_\tau \tag{D.15}$$

where

$$K^{\omega, \tilde{\omega}}_{i;0}(\sigma, \tau) := K^{\omega, \tilde{\omega}}_{\mathrm{CDE}}(\sigma, \tau), \quad K^{\omega, \tilde{\omega}}_{i;\ell+1}(\sigma, \tau) = \mathbb{E}\left[\varphi_{i;\ell+1}(X_\ell) \varphi_{i;\ell+1}(Y_\ell)\right]$$

$$(X_\ell, Y_\ell) \sim \mathcal{N}\left(0, \begin{bmatrix} K^{\omega,\omega}_{i;\ell}(\sigma, \sigma) & K^{\omega,\tilde{\omega}}_{i;\ell}(\sigma, \tau) \\ K^{\omega,\tilde{\omega}}_{i;\ell}(\sigma, \tau) & K^{\tilde{\omega},\tilde{\omega}}_{i;\ell}(\tau, \tau) \end{bmatrix}\right)$$

Proof. From eq. (D.13) we can write

$$\frac{1}{N}\langle Z_s^\omega, Z_t^{\tilde\omega}\rangle = \frac{1}{N}\sum_{\tau\in\mathbb{T}^{\mathrm{CDE}}(\bullet_{\mathbf{z}_0})}\sum_{\eta\in\mathbb{T}^{\mathrm{CDE}}(\bullet_{\mathbf{z}_0})}\frac{\varphi_\tau\varphi_\eta}{s(\tau)s(\eta)}\mathbf{W}_{\langle\tau,\eta\rangle}\mathrm{Sig}(\omega)^\tau_{0,s}\mathrm{Sig}(\tilde\omega)^\eta_{0,t}$$

Using theorem D.11 and noticing that if $\tau = \circ\,{}_{-Id}\bullet_{\mathbf{z}_0}$ then $\mathcal{P}_{\mathrm{A}}(\langle\tau,\eta\rangle)\neq\emptyset$ iff $\eta=\tau$ in which case the L^2 limit is $\sum_{\phi\in\mathcal{P}_{\mathrm{A}}(\langle\tau,\eta\rangle)}\mathbf{W}_{\langle\tau,\eta\rangle_\phi} = N$ we have that the scaled product *formally* converges to

$$\begin{aligned}
&\sum_{\tau\in\mathbb{T}^{\mathrm{CDE}}(\bullet_{\mathbf{z}_0})}\sum_{\eta\in\mathbb{T}^{\mathrm{CDE}}(\bullet_{\mathbf{z}_0})}\frac{\varphi_\tau\varphi_\eta}{s(\tau)s(\eta)}a(\langle\tau,\eta\rangle)\mathrm{Sig}(\omega)^\tau_{0,s}\mathrm{Sig}(\tilde\omega)^\eta_{0,t}\\
&= 1+\sum_{i=1}^{d}\sum_{\boldsymbol{\tau}[\![\tau_1,\dots,\tau_m]\!]\in\mathbf{T}_{i;L_i}}\sum_{\boldsymbol{\eta}[\![\eta_1,\dots,\eta_n]\!]\in\mathbf{T}_{i;L_i}}\frac{\varphi_\tau\varphi_\eta}{s(\tau)s(\eta)}a(\langle\tau,\eta\rangle)\mathrm{Sig}(\omega)^\tau_{0,s}\mathrm{Sig}(\tilde\omega)^\eta_{0,t}
\end{aligned}$$

Now, for $\tau\simeq\boldsymbol{\tau}[\![\tau_1,\dots,\tau_m]\!]$ and $\eta\simeq\boldsymbol{\eta}[\![\eta_1,\dots,\eta_n]\!]\in\mathbf{T}_{i;L_i}$, by theorem D.10 and with a small abuse of notation writing ω_π instead of $\tilde\omega_\pi$ below whenever $\pi\in\{\eta_1,\dots,\eta_n\}$:

$$\begin{aligned}
&\frac{\varphi_\tau\varphi_\eta}{s(\tau)s(\eta)}a(\langle\tau,\eta\rangle)\mathrm{Sig}(\omega)^\tau_{0,s}\mathrm{Sig}(\tilde\omega)^\eta_{0,t}\\
&=\sum_{\phi\in\mathcal{P}_{\mathrm{nest}}(\boldsymbol{\tau}\sqcup\boldsymbol{\eta})}\frac{\varphi_\tau\varphi_\eta}{s(\tau)s(\eta)}\left(\prod_{\boldsymbol{\pi}\in\boldsymbol{\phi}}a(\langle\pi_1,\pi_2\rangle)\right)\mathrm{Sig}(\omega)^\tau_{0,s}\mathrm{Sig}(\tilde\omega)^\eta_{0,t}\\
&=\int_0^s\int_0^t\frac{\varphi_{\boldsymbol{\tau}}\varphi_{\boldsymbol{\eta}}}{\mathbf{s}(\boldsymbol{\tau})\mathbf{s}(\boldsymbol{\eta})}\sum_{\boldsymbol{\phi}\in\mathcal{P}_{\mathrm{nest}}(\boldsymbol{\tau}\sqcup\boldsymbol{\eta})}\prod_{\boldsymbol{\pi}\in\boldsymbol{\phi}}\left(\frac{\varphi_{\pi_1}\varphi_{\pi_2}}{s(\pi_1)s(\pi_2)}a(\langle\pi_1,\pi_2\rangle)\mathrm{Sig}(\omega_{\pi_1})^{\pi_1}_{0,s_{\pi_1}}\mathrm{Sig}(\omega_{\pi_2})^{\pi_2}_{0,s_{\pi_2}}\right)\mathrm{d}\omega^i_{s_1}\mathrm{d}\tilde\omega^i_{t_1}
\end{aligned}$$

where we have used that for all $\boldsymbol{\phi}\in\mathcal{P}_{\mathrm{nest}}(\boldsymbol{\tau}\sqcup\boldsymbol{\eta})$ one has

$$\frac{\varphi_\tau\varphi_\eta}{s(\tau)s(\eta)} = \frac{\varphi_{\boldsymbol{\tau}}\varphi_{\boldsymbol{\eta}}}{\mathbf{s}(\boldsymbol{\tau})\mathbf{s}(\boldsymbol{\eta})}\prod_{\boldsymbol{\pi}\in\boldsymbol{\phi}}\frac{\varphi_{\pi_1}\varphi_{\pi_2}}{s(\pi_1)s(\pi_2)},$$

and

$$\mathrm{Sig}(\omega)^\tau_{0,s}\mathrm{Sig}(\tilde\omega)^\eta_{0,t} = \int_0^s\int_0^t\prod_{\boldsymbol{\pi}\in\boldsymbol{\phi}}\left(\mathrm{Sig}(\omega_{\pi_1})^{\pi_1}_{0,s_{\pi_1}}\mathrm{Sig}(\omega_{\pi_2})^{\pi_2}_{0,s_{\pi_2}}\right)\mathrm{d}\omega^i_{s_1}\mathrm{d}\tilde\omega^i_{t_1}$$

Now using theorem D.5 with

$$\lambda(\pi_1,\pi_2) = \frac{\varphi_{\pi_1}\varphi_{\pi_2}}{s(\pi_1)s(\pi_2)}a(\langle\pi_1,\pi_2\rangle)\mathrm{Sig}(\omega_{\pi_1})^{\pi_1}_{0,s_{\pi_1}}\mathrm{Sig}(\omega_{\pi_2})^{\pi_2}_{0,s_{\pi_2}}$$

we obtain

$$\begin{aligned}
&\sum_{\boldsymbol{\tau}[\![\tau_1,\dots,\tau_m]\!]\in\mathbf{T}_{i;L_i}}\sum_{\boldsymbol{\eta}[\![\eta_1,\dots,\eta_n]\!]\in\mathbf{T}_{i;L_i}}\frac{\varphi_\tau\varphi_\eta}{s(\tau)s(\eta)}a(\langle\tau,\eta\rangle)\mathrm{Sig}(\omega)^\tau_{0,s}\mathrm{Sig}(\tilde\omega)^\eta_{0,t}\\
&=\int_0^s\int_0^t K^{\omega,\tilde\omega}_{i;L_i}(s_1,t_1)\mathrm{d}\omega^i_{s_1}\mathrm{d}\tilde\omega^i_{t_1}
\end{aligned}$$

which allows to conclude. □

Corollary. *For $K^{\omega,\tilde{\omega}}_{\mathrm{CDE}}(s,t)$ the following expansion holds formally*

$$K^{\omega,\tilde{\omega}}_{\mathrm{CDE}}(s,t) = \sum_{\tau,\eta\in\mathbb{T}^{\mathrm{CDE}}(\bullet_{\mathbf{z}_0})} \frac{\varphi_\tau\varphi_\eta}{s(\tau)s(\eta)} a(\langle\tau,\eta\rangle)\mathrm{Sig}(\omega)^\tau_{0,s}\mathrm{Sig}(\tilde{\omega})^\eta_{0,t} \tag{D.16}$$

D.5 NTK

Recall how the NTK is defined as

$$\Theta^{\omega,\tilde{\omega}}_N(s,t) := \sum_{i=1}^{d}\sum_{\ell=0}^{L_i} \lambda_{i;\ell}(d_{W_{i;\ell}}v\cdot Z^\omega_s)(d_{W_{i;\ell}}v\cdot Z^{\tilde{\omega}}_t)^\top \in \mathbb{R} \tag{D.17}$$

where the λ_ℓ denote the layer-wise learning rates.

To find an expansion for Θ, let $\partial_{i;\ell}\mathbb{T}^{\mathrm{CDE}}(\bullet_{\mathbf{z}_0})$ denote the space of trees which can be obtained from trees in $\mathbb{T}^{\mathrm{CDE}}(\bullet_{\mathbf{z}_0})$ by freeing exactly one of its $(i;\ell)$-labeled edges (i.e. one edge with input $W_{i;\ell}$), making it an in-edge. We can then write

$$(\mathrm{d}\langle v, Z^\omega_t\rangle)_{W_{i;\ell}} = \sum_{\tau\in\partial_{i;\ell}\mathbb{T}^{\mathrm{CDE}}(\bullet_{\mathbf{z}_0})} \frac{\varphi_\tau}{s(\tau)}\mathbf{W}_{\langle v,\tau\rangle}\mathrm{Sig}(\omega)^\tau_{0,t}, \tag{D.18}$$

where $\langle v,\tau\rangle$ denotes τ with out-vertex fixed to v and thus $\mathbf{W}_\tau : \mathbb{R}^{N_{\ell+1}\times N_\ell}\to\mathbb{R}$. Arguing as we above, we find that

$$\lambda_{i;\ell}(d_{W_{i;\ell}}\langle v, Z^\omega_s\rangle)(d_{W_{i;\ell}}\langle v, Z^{\tilde{\omega}}_t\rangle)^\top \tag{D.19}$$

$$= \sum_{\tau\in\partial_{i;\ell}\mathbb{T}^{\mathrm{CDE}}(\bullet_{\mathbf{z}_0})}\sum_{\eta\in\partial_{i;\ell}\mathbb{T}^{\mathrm{CDE}}(\bullet_{\mathbf{z}_0})} \frac{\varphi_\tau\varphi_\eta}{s(\tau)s(\eta)}\lambda_{i;\ell}\mathbf{W}_{\langle v,\tau\rangle\circ\langle v,\eta\rangle^\top}\mathrm{Sig}(\omega)^\tau_{0,s}\mathrm{Sig}(\tilde{\omega})^\eta_{0,t}. \tag{D.20}$$

Assumption D.13. *Assume $\lambda_{i;n} = \frac{1}{N}$.*

Proposition D.14. *Under theorem D.13*

$$\mathbb{E}\left\{\left(\lambda_{i;\ell}\mathbf{W}_{\langle v,\tau\rangle\circ\langle v,\eta\rangle^\top} - |\mathcal{P}_{\mathrm{A}}(\langle v,\tau\rangle\circ\langle v,\eta\rangle^\top)|\right)^2\right\} = \mathcal{O}(\frac{1}{N}) \tag{D.21}$$

Proof. The proof is a repetition of the arguments used to prove theorem 2.15, noticing that given $\tau,\eta\in\partial_{i;\ell}\mathbb{T}^{\mathrm{CDE}}(\bullet_{\mathbf{z}_0})$, if their open edges are not at the same distance from the root then $|\mathcal{P}_{\mathrm{A}}(\langle v,\tau\rangle\circ\langle v,\eta\rangle^\top)| = \emptyset$, and that in any $\phi\in\mathcal{P}_{\mathrm{A}}(\langle v,\tau\rangle\circ\langle v,\eta\rangle^\top)$ the two root edges are paired together. □

Theorem D.15 (Formal NTK Kernel Limit)**.** *Under theorem D.13*

$$\Theta^{\omega,\tilde{\omega}}_N(s,t) \xrightarrow[N\to\infty]{L^2} \Theta^{\omega,\tilde{\omega}}_{\mathrm{CDE}}(s,t) = \sum_{i=1}^{d}\int_{s_1=0}^{s}\int_{t_1=0}^{t}\Theta^{\omega,\tilde{\omega}}_{i;L_i}(s_1,t_1)d\omega^i_{s_1}d\tilde{\omega}^i_{t_1} \tag{D.22}$$

where

$$\Theta^{\omega,\tilde{\omega}}_{i;0}(s,t) = K^{\omega,\tilde{\omega}}_{\mathrm{CDE}}(s,t) + \Theta^{\omega,\tilde{\omega}}_{\mathrm{CDE}}(s,t), \quad K^{\omega,\tilde{\omega}}_{i;0}(s,t) = K^{\omega,\tilde{\omega}}_{\mathrm{CDE}}(s,t), \tag{D.23}$$

$$\Theta^{\omega,\tilde{\omega}}_{i;\ell}(s,t) = K^{\omega,\tilde{\omega}}_{i;\ell}(s,t) + \dot{K}^{\omega,\tilde{\omega}}_{i;\ell}(s,t)\Theta^{\omega,\tilde{\omega}}_{i;\ell-1}(s,t) \tag{D.24}$$

with

$$K^{\omega,\tilde{\omega}}_{i;\ell+1}(s,t) := \mathbb{E}\left[\varphi_{i;\ell+1}(X_\ell)\varphi_{i;\ell+1}(Y_\ell)\right], \quad \dot{K}^{\omega,\tilde{\omega}}_{i;\ell+1}(s,t) := \mathbb{E}\left[\dot{\varphi}_{i;\ell+1}(X_\ell)\dot{\varphi}_{i;\ell+1}(Y_\ell)\right]$$

$$(X_\ell, Y_\ell) \sim \mathcal{N}\left(0, \begin{bmatrix} K^{\omega,\omega}_{i;\ell}(s,s) & K^{\omega,\tilde{\omega}}_{i;\ell}(s,t) \\ K^{\omega,\tilde{\omega}}_{i;\ell}(s,t) & K^{\tilde{\omega},\tilde{\omega}}_{i;\ell}(t,t) \end{bmatrix}\right)$$

Proof. We assume $L_1 = \cdots = L_d = L$, the general case following. Using the tree expansion and the L^2 convergence we obtain

$$\Theta^{\omega,\tilde{\omega}}_N(s,t) \xrightarrow[N\to\infty]{L^2} \sum_{i=1}^{d}\sum_{\ell=0}^{L}\sum_{\tau,\eta\in\partial_{i;\ell}\mathbb{T}^{\mathrm{CDE}}(\bullet_{\mathbf{z}_0})} \frac{\varphi_\tau\varphi_\eta}{s(\tau)s(\eta)}|\mathcal{P}_{\mathrm{A}}(\langle v,\tau\rangle\circ\langle v,\eta\rangle^\top)|\mathrm{Sig}(\omega)^\tau_{0,s}\mathrm{Sig}(\tilde{\omega})^\eta_{0,t}.$$

Define $\Theta^{\omega,\tilde{\omega}}_{\mathrm{CDE}}(s,t)$ as the right hand side of the limit.

Any tree $\tau \in \mathbb{T}^{\mathrm{CDE}}(\bullet_{\mathbf{z}_0})$ is of form $\tau = \hat{\tau}[\tau_1,\ldots,\tau_m]$, we will call $\hat{\tau}$ its base (note this might or might not contain an open edge) and write $\mathbb{T}^{\mathrm{CDE}}_{i;L} := \mathbb{T}_{i;L}(\mathbb{T}^{\mathrm{CDE}}(\bullet_{\mathbf{z}_0}))$ for the set of trees with i-labelled base. More generally given two sets of trees $\mathbb{T}_1$ with open in-terminal leaves and $\mathbb{T}_2$ with open out-root we write $\mathbb{T}_1(\mathbb{T}_2)$ for the set of trees created by composing $\mathbb{T}_2$ in the open terminal leaves of trees in $\mathbb{T}_1$.

Now we reason as in theorem 2.16, looking for convenient bijections of the spaces $\partial_{i;\ell}\mathbb{T}^{\mathrm{CDE}}(\bullet_{\mathbf{z}_0}) = \sqcup_j\partial_{i;\ell}\left(\mathbb{T}_{j;L}(\mathbb{T}^{\mathrm{CDE}})\right)$, where we have dropped the $(\bullet_{\mathbf{z}_0})$ from $\mathbb{T}^{\mathrm{CDE}}(\bullet_{\mathbf{z}_0})$ for ease of notation. It will be necessary to study the more general spaces

$$\sqcup_j\partial_{i;\ell}\left(\mathbb{T}_{j;\ell'}(\mathbb{T}^{\mathrm{CDE}})\right)$$

for $0 \le \ell, \ell' \le L$.

Each tree $\tau \in \partial_{i;\ell}\left(\mathbb{T}_{j;\ell'}(\mathbb{T}^{\mathrm{CDE}})\right)$ can be of one of two kinds: either the root edge is open (thus also $(j,\ell') = (i,\ell)$) or this is not the case. In the second case we can however, just as in theorem 2.16, uniquely identify the tree as a pair $(\tau^+,\tau^-) \in \mathbb{T}_{j;\ell'}(\mathbb{T}^{\mathrm{CDE}}) \times \partial_{i;\ell}\mathbb{T}_{j;\ell'-1}(\mathbb{T}^{\mathrm{CDE}})$ where τ^- is the subtree attached to τ's base containing the open edge, and τ^+ is its complement. Note that vice versa each couplet $(\tau^+,\tau^-) \in \mathbb{T}_{j;\ell'}(\mathbb{T}^{\mathrm{CDE}}) \times \partial_{i;\ell}\mathbb{T}_{j;\ell'-1}(\mathbb{T}^{\mathrm{CDE}})$ defines a tree of the second kind, while any tree of the first kind can be identified uniquely with the tree $\mathbb{T}_{i;\ell}(\mathbb{T}^{\mathrm{CDE}})$ produced by fixing the root edge to $W_{i,\ell}$.

We thus have the bijections

$$\Psi_{i;\ell}^{j;\ell'} : \partial_{i;\ell}\left(\mathbb{T}_{j;\ell'}(\mathbb{T}^{\mathrm{CDE}})\right) \to \delta_{i,j}\delta_{\ell',\ell}\mathbb{T}_{i;\ell}(\mathbb{T}^{\mathrm{CDE}}) \sqcup \left(\mathbb{T}_{j;\ell'}(\mathbb{T}^{\mathrm{CDE}}) \times \partial_{i;\ell}\mathbb{T}_{j;\ell'-1}(\mathbb{T}^{\mathrm{CDE}})\right) \tag{D.25}$$

for $\ell' > 0$ and

$$\Psi_{i;\ell}^{j;0} : \partial_{i;\ell}\left(\mathbb{T}_{j;0}(\mathbb{T}^{\mathrm{CDE}})\right) \to \delta_{i,j}\delta_{\ell',\ell}\mathbb{T}_{i;0}(\mathbb{T}^{\mathrm{CDE}}) \sqcup \mathbb{T}_{i;0}(\partial_{i;\ell}\mathbb{T}^{\mathrm{CDE}}) \tag{D.26}$$

Let $i = j$, $\ell = \ell'$, $\Psi = \Psi_{i;\ell}^{j;\ell'}$ and τ, η have open root edges. Then

$$|\mathcal{P}_{\mathrm{A}}(\langle v, \tau\rangle \circ \langle v, \eta\rangle^{\top})| = |\mathcal{P}_{\mathrm{A}}(\langle \Psi(\tau), \Psi(\eta)\rangle)| \tag{D.27}$$

since the v-fixed vertices must be paired just as the roots $\Psi(\tau)$ and $\Psi(\eta)$ in any atomic pairing. Moreover it is clear that

$$\varphi_\tau = \varphi_{\Psi(\tau)}, \quad s(\tau) = s(\Psi(\tau)), \quad \mathrm{Sig}(\omega)_{0,s}^{\tau} = \mathrm{Sig}(\omega)_{0,s}^{\Psi(\tau)}.$$

In the generic case with $\ell' > 0$, let $\Psi = \Psi_{i;\ell}^{j;\ell'}$ and τ, η be trees without open roots so that $\Psi(\tau) = (\tau^+, \tau^-)$ and similarly for η. Arguing as in theorem 2.16 the roots of τ^- and η^- must be paired by any atomic pairing of $\langle v, \tau\rangle \circ \langle v, \eta\rangle^{\top}$, so that we can split

$$\mathcal{P}_{\mathrm{A}}(\langle v, \tau\rangle \circ \langle v, \eta\rangle^{\top}) = \mathcal{P}_{\mathrm{A}}(\langle \tau^+, \eta^+\rangle) \times \mathcal{P}_{\mathrm{A}}(\langle v, \tau^-\rangle \circ \langle v, \eta^-\rangle^{\top})$$

and thus obtain

$$|\mathcal{P}_{\mathrm{A}}(\langle v, \tau\rangle \circ \langle v, \eta\rangle^{\top})| = |\mathcal{P}_{\mathrm{A}}(\langle \tau^+, \eta^+\rangle)||\mathcal{P}_{\mathrm{A}}(\langle v, \tau^-\rangle \circ \langle v, \eta^-\rangle^{\top})| \tag{D.28}$$

Moreover, just as in theorem 2.14's proof, we have in this setting that

$$\varphi_\tau = \dot{\varphi}_{\tau^+}\varphi_{\tau^-}, \quad s(\tau) = s(\tau^+)s(\tau^-).$$

Finally, recall how for $\tau = \hat{\tau}[\tau_1, \dots, \tau_m] \in \mathbb{T}_{j;\ell}^{\mathrm{CDE}}$, and similarly for trees in $\partial_{i;\ell}\mathbb{T}_{j;\ell'}^{\mathrm{CDE}}$, one defines recursively

$$\mathrm{Sig}(\omega)_{0,s}^{\tau} = \int_0^s \prod_{k=1}^{m} \mathrm{Sig}(\omega)_{0,s'}^{\tau_k} \mathrm{d}\omega_{s'}^{j},$$

if we define $\widehat{\mathrm{Sig}}(\omega)_{0,s}^{\tau} := \prod_{k=1}^{m} \mathrm{Sig}(\omega)_{0,s}^{\tau_k}$ then $\mathrm{Sig}(\omega)_{0,s}^{\tau} = \int_0^s \widehat{\mathrm{Sig}}(\omega)_{0,s'}^{\tau} \mathrm{d}\omega_{s'}^{j}$ and for $\tau \in \partial_{i;\ell}\left(\mathbb{T}_{j;\ell'}(\mathbb{T}^{\mathrm{CDE}})\right)$ without open root

$$\mathrm{Sig}(\omega)_{0,s}^{\tau} = \int_0^s \widehat{\mathrm{Sig}}(\omega)_{0,s'}^{\tau^+} \widehat{\mathrm{Sig}}(\omega)_{0,s'}^{\tau^-} \, \mathrm{d}\omega_{s'}^{j}.$$

We thus obtain, writing $\mathbb{T}^{\mathrm{CDE}}_{j;\ell} := \mathbb{T}_{j;\ell}(\mathbb{T}^{\mathrm{CDE}})$, in case $\ell' > 0$ that

$$\begin{aligned}
&\sum_{i=1}^{d}\sum_{\ell=0}^{L}\sum_{\tau,\eta\in\partial_{i;\ell}\left(\mathbb{T}_{j;\ell'}(\mathbb{T}^{\mathrm{CDE}})\right)}\frac{\varphi_\tau\varphi_\eta}{s(\tau)s(\eta)}|\mathcal{P}_{\mathrm{A}}(\langle v,\tau\rangle\circ\langle v,\eta\rangle^\top)|\mathrm{Sig}(\omega)^\tau_{0,s}\mathrm{Sig}(\tilde\omega)^\eta_{0,t}\\
&=\sum_{\tau,\eta\in\mathbb{T}^{\mathrm{CDE}}_{j;\ell'}}\frac{\varphi_\tau\varphi_\eta}{s(\tau)s(\eta)}|\mathcal{P}_{\mathrm{A}}(\langle\tau,\eta\rangle)|\mathrm{Sig}(\omega)^\tau_{0,s}\mathrm{Sig}(\tilde\omega)^\eta_{0,t}\\
&\quad+\sum_{i=1}^{d}\sum_{\ell=0}^{L}\sum_{\tau^+,\eta^+\in\mathbb{T}^{\mathrm{CDE}}_{j;\ell'}}\sum_{\tau^-,\eta^-\in\partial_{i;\ell}\mathbb{T}^{\mathrm{CDE}}_{j;\ell'-1}}\frac{\varphi_\tau\varphi_\eta}{s(\tau)s(\eta)}|\mathcal{P}_{\mathrm{A}}(\langle v,\tau\rangle\circ\langle v,\eta\rangle^\top)|\mathrm{Sig}(\omega)^\tau_{0,s}\mathrm{Sig}(\tilde\omega)^\eta_{0,t},
\end{aligned}$$

and since, in the context of the last equation above,

$$\begin{aligned}
&\frac{\varphi_\tau\varphi_\eta}{s(\tau)s(\eta)}|\mathcal{P}_{\mathrm{A}}(\langle v,\tau\rangle\circ\langle v,\eta\rangle^\top)|\mathrm{Sig}(\omega)^\tau_{0,s}\mathrm{Sig}(\tilde\omega)^\eta_{0,t}\\
&=\int_0^s\int_0^t\frac{\dot\varphi_{\tau^+}\dot\varphi_{\eta^+}}{s(\tau^+)s(\eta^+)}|\mathcal{P}_{\mathrm{A}}(\langle\tau^+,\eta^+\rangle)|\widehat{\mathrm{Sig}}(\omega)^{\tau^+}_{0,s'}\widehat{\mathrm{Sig}}(\omega)^{\eta^+}_{0,t'}\cdot\\
&\qquad\cdot\frac{\varphi_{\tau^-}\varphi_{\eta^-}}{s(\tau^-)s(\eta^-)}|\mathcal{P}_{\mathrm{A}}(\langle v,\tau^-\rangle\circ\langle v,\eta^-\rangle^\top)|\widehat{\mathrm{Sig}}(\omega)^{\tau^-}_{0,s'}\widehat{\mathrm{Sig}}(\omega)^{\eta^-}_{0,t'}\,\mathrm{d}\omega^j_{s'}\mathrm{d}\tilde\omega^j_{t'}
\end{aligned}$$

we can write

$$\begin{aligned}
&\sum_{i=1}^{d}\sum_{\ell=0}^{L}\sum_{\tau^+,\eta^+\in\mathbb{T}^{\mathrm{CDE}}_{j;\ell'}}\sum_{\tau^-,\eta^-\in\partial_{i;\ell}\mathbb{T}^{\mathrm{CDE}}_{j;\ell'-1}}\frac{\varphi_\tau\varphi_\eta}{s(\tau)s(\eta)}|\mathcal{P}_{\mathrm{A}}(\langle v,\tau\rangle\circ\langle v,\eta\rangle^\top)|\mathrm{Sig}(\omega)^\tau_{0,s}\mathrm{Sig}(\tilde\omega)^\eta_{0,t}\\
&=\sum_{i=1}^{d}\int_0^s\int_0^t\sum_{\ell=0}^{L}\left(\sum_{\tau^+,\eta^+\in\mathbb{T}^{\mathrm{CDE}}_{j;\ell'}}\frac{\dot\varphi_{\tau^+}\dot\varphi_{\eta^+}}{s(\tau^+)s(\eta^+)}|\mathcal{P}_{\mathrm{A}}(\langle\tau^+,\eta^+\rangle)|\widehat{\mathrm{Sig}}(\omega)^{\tau^+}_{0,s'}\widehat{\mathrm{Sig}}(\tilde\omega)^{\eta^+}_{0,t'}\right)\cdot\\
&\quad\cdot\left(\sum_{\tau^-,\eta^-\in\partial_{i;\ell}\mathbb{T}^{\mathrm{CDE}}_{j;\ell'-1}}\frac{\varphi_{\tau^-}\varphi_{\eta^-}}{s(\tau^-)s(\eta^-)}|\mathcal{P}_{\mathrm{A}}(\langle v,\tau^-\rangle\circ\langle v,\eta^-\rangle^\top)|\widehat{\mathrm{Sig}}(\omega)^{\tau^-}_{0,s'}\widehat{\mathrm{Sig}}(\tilde\omega)^{\eta^-}_{0,t'}\right)\,\mathrm{d}\omega^j_{s'}\mathrm{d}\tilde\omega^j_{t'}.
\end{aligned}$$

The case $\ell' = 0$ is similar but simpler and one gets

$$\begin{aligned}
&\sum_{i=1}^{d}\sum_{\ell=0}^{L}\sum_{\tau,\eta\in\partial_{i;\ell}(\mathbb{T}_{j;0}(\mathbb{T}^{\mathrm{CDE}}))}\frac{\varphi_\tau\varphi_\eta}{s(\tau)s(\eta)}|\mathcal{P}_{\mathrm{A}}(\langle v,\tau\rangle\circ\langle v,\eta\rangle^\top)|\mathrm{Sig}(\omega)^\tau_{0,s}\mathrm{Sig}(\tilde\omega)^\eta_{0,t}\\
&=\sum_{\tau,\eta\in\mathbb{T}^{\mathrm{CDE}}_{j;0}}\frac{\varphi_\tau\varphi_\eta}{s(\tau)s(\eta)}|\mathcal{P}_{\mathrm{A}}(\langle\tau,\eta\rangle)|\mathrm{Sig}(\omega)^\tau_{0,s}\mathrm{Sig}(\tilde\omega)^\eta_{0,t}\\
&\quad+\sum_{i=1}^{d}\sum_{\ell=0}^{L}\sum_{\tau^-,\eta^-\in\mathbb{T}_{j;0}(\partial_{i;\ell}\mathbb{T}^{\mathrm{CDE}})}\frac{\varphi_\tau\varphi_\eta}{s(\tau)s(\eta)}|\mathcal{P}_{\mathrm{A}}(\langle v,\tau\rangle\circ\langle v,\eta\rangle^\top)|\mathrm{Sig}(\omega)^{\tau^-}_{0,s}\mathrm{Sig}(\tilde\omega)^{\eta^-}_{0,t},
\end{aligned}$$

and

$$\sum_{i=1}^{d}\sum_{\ell=0}^{L}\sum_{\tau^-,\eta^-\in\mathbb{T}_{j;0}(\partial_{i;\ell}\mathbb{T}^{\mathrm{CDE}})}\frac{\varphi_\tau\varphi_\eta}{s(\tau)s(\eta)}|\mathcal{P}_{\mathrm{A}}(\langle v,\tau\rangle\circ\langle v,\eta\rangle^\top)|\mathrm{Sig}(\omega)^\tau_{0,s}\mathrm{Sig}(\tilde{\omega})^\eta_{0,t}$$
$$=\sum_{i=1}^{d}\int_0^s\int_0^t\sum_{\ell=0}^{L}\sum_{\tau,\eta\in\partial_{i;\ell}\mathbb{T}^{\mathrm{CDE}}}\frac{\varphi_\tau\varphi_\eta}{s(\tau)s(\eta)}|\mathcal{P}_{\mathrm{A}}(\langle v,\tau\rangle\circ\langle v,\eta\rangle^\top)|\mathrm{Sig}(\omega)^\tau_{0,s'}\mathrm{Sig}(\tilde{\omega})^\eta_{0,t'}\,\mathrm{d}\omega^j_{s'}\mathrm{d}\tilde{\omega}^j_{t'}$$

from which we recognise

$$\sum_{i=1}^{d}\sum_{\ell=0}^{L}\sum_{\tau,\eta\in\partial_{i;\ell}\mathbb{T}^{\mathrm{CDE}}}\frac{\varphi_\tau\varphi_\eta}{s(\tau)s(\eta)}|\mathcal{P}_{\mathrm{A}}(\langle v,\tau\rangle\circ\langle v,\eta\rangle^\top)|\mathrm{Sig}(\omega)^\tau_{0,s'}\mathrm{Sig}(\tilde{\omega})^\eta_{0,t'}=\Theta_{\mathrm{CDE}}(s',t')$$

If we define, omitting the dependence on $\omega,\tilde{\omega}$,

$$\Theta^{i;\ell}_{j;\ell'}(s,t):=\sum_{\tau,\eta\in\partial_{i;\ell}\mathbb{T}^{\mathrm{CDE}}_{j;\ell'}}\frac{\varphi_\tau\varphi_\eta}{s(\tau)s(\eta)}|\mathcal{P}_{\mathrm{A}}(\langle v,\tau\rangle\circ\langle v,\eta\rangle^\top)|\widehat{\mathrm{Sig}}(\omega)^\tau_{0,s}\widehat{\mathrm{Sig}}(\tilde{\omega})^\eta_{0,t}\tag{D.29}$$

$$\Theta^{\mathrm{CDE}}_{j;\ell'}(s,t):=\sum_{i=1}^{d}\sum_{\ell=0}^{L}\Theta^{i;\ell}_{j;\ell'}(s,t)\tag{D.30}$$

$$K_{i;\ell}(s,t):=\sum_{\tau,\eta\in\mathbb{T}^{\mathrm{CDE}}_{i;\ell}}\frac{\varphi_\tau\varphi_\eta}{s(\tau)s(\eta)}|\mathcal{P}_{\mathrm{A}}(\langle\tau,\eta\rangle)|\widehat{\mathrm{Sig}}(\omega)^\tau_{0,s}\widehat{\mathrm{Sig}}(\tilde{\omega})^\eta_{0,t}\tag{D.31}$$

and

$$\dot{K}_{i;\ell}(s,t):=\sum_{\tau,\eta\in\mathbb{T}^{\mathrm{CDE}}_{i;\ell}}\frac{\dot{\varphi}_\tau\dot{\varphi}_\eta}{s(\tau)s(\eta)}|\mathcal{P}_{\mathrm{A}}(\langle\tau,\eta\rangle)|\widehat{\mathrm{Sig}}(\omega)^\tau_{0,s}\widehat{\mathrm{Sig}}(\tilde{\omega})^\eta_{0,t}\tag{D.32}$$

we obtain from the previous computations that for $\ell'>0$ it holds

$$\sum_{i=1}^{d}\sum_{\ell=0}^{L}\Theta^{i;\ell}_{j;\ell'}(s,t)=K_{j;\ell'}(s,t)+\dot{K}_{j;\ell'}(s,t)\cdot\left(\sum_{i=1}^{d}\sum_{\ell=0}^{L}\Theta^{i;\ell}_{j;\ell'-1}(s,t)\right)$$

i.e.

$$\Theta^{\mathrm{CDE}}_{j;\ell'}(s,t)=K_{j;\ell'}(s,t)+\dot{K}_{j;\ell'}(s,t)\cdot\Theta^{\mathrm{CDE}}_{j;\ell'-1}(s,t)\tag{D.33}$$

and for $\ell'=0$

$$K_{j;0}(s,t)=K_{\mathrm{CDE}}(s,t),\tag{D.34}$$

$$\Theta^{\mathrm{CDE}}_{j;0}(s,t)=K_{\mathrm{CDE}}(s,t)+\Theta_{\mathrm{CDE}}(s,t)\tag{D.35}$$

To conclude note how

$$\Theta_{\mathrm{CDE}}(s,t)=\sum_{j=1}^{d}\int_0^s\int_0^t\sum_{i=1}^{d}\sum_{\ell=0}^{L}\Theta^{i;\ell}_{j;L}(s',t')\,\mathrm{d}\omega^j_{s'}\mathrm{d}\tilde{\omega}^j_{t'}\tag{D.36}$$

so that

$$\Theta_{\mathrm{CDE}}(s,t) = \sum_{j=1}^{d} \int_0^s \int_0^t \Theta_{j;L}^{\mathrm{CDE}}(s',t') \, \mathrm{d}\omega_{s'}^j \mathrm{d}\tilde{\omega}_{t'}^j. \tag{D.37}$$

and how, from the proof of theorem D.12 one has

$$K_{i;\ell+1}(s,t) = K_{i;\ell+1}^{\omega,\tilde{\omega}}(s,t), \quad \dot{K}_{i;\ell+1}(s,t) = \dot{K}_{i;\ell+1}^{\omega,\tilde{\omega}}(s,t).$$

□

Appendix E

Scaling Limits of Training Dynamics

An obvious question motivates this chapter: can the tools developed in the main body of this work be used to analyse the training dynamics of neural networks under Stochastic Gradient Descent (SGD) in the feature-learning (non-kernel) regime? This regime is substantially richer than the Neural Tangent Kernel (NTK) setting, where training is described by a fixed kernel obtained by linearising the network around its initialization and representation learning is effectively suppressed [132]. To date, a satisfactory theory exists primarily in the linear case—where a network reduces to a composition of linear maps—as developed by Chizat et al. [13].

In this chapter we show how the constructions and results of Chizat et al. [13] emerge naturally from our framework. Our focus will be on interpreting their objects within our graphical formalism and clarifying the correspondences between the two viewpoints; for complete proofs we refer the reader back to Chizat et al. [13]. For these reasons the chapter is placed in the Appendix, serving chiefly as a guidepost, shedding light on the underlying combinatorial structure, toward a full extension to the general, genuinely nonlinear setting.

We conclude by presenting a new result that identifies the SGD-driven neural dynamics with a flow in an appropriate space of polynomials (*c.f.* Theorem E.4).

E.0.1 The Model

To fix notation, consider the MLP model Φ_L for $L \geq 1$ defined as

$$\Phi_0(\mathbf{x}) = W_0\mathbf{x}, \quad \Phi_{\ell+1}(\mathbf{x}) = W_{\ell+1}[\varphi_{\ell+1}(\Phi_\ell(\mathbf{x}))] \tag{E.1}$$

for matrices $W_k \in \mathbb{R}^{N_k \times N_{k+1}}$. If $\theta = (W_0, \dots, W_L)$ is the sequence of weights we will also write

$$\Phi_L(\mathbf{x}) = \Phi_L(\mathbf{x}; \theta) = \Phi(\mathbf{x}; \theta)$$

E.0.2 Stochastic Gradient Descent

We consider a supervised learning probelm with SGD, meaning that we fix a finite-dimensional data space $\mathbb{X} \times \mathbb{Y} \subseteq \mathbb{R}^{N_0} \times \mathbb{R}^{N_L}$ of input-output pairs, a sequence of probability measures

$$(\rho_k \in \mathcal{M}(\mathbb{X} \times \mathbb{Y}) \mid k \in \mathbb{N})$$

having uniformly bounded second moments in the first variable

$$\sup_k \int |\mathbf{x}|^2 \mathrm{d}\rho_k(\mathbf{x}, y),$$

and we update the model weights W_ℓ as[1]

$$W_\ell(k+1) = W_\ell(k) + \lambda_\ell \nabla_{W_\ell} \mathcal{L}^k(\theta(k)) \tag{E.2}$$

where $\theta(k) := (W_0(k), \dots, W_L(k))$ and

$$\mathcal{L}^k(\theta) := \int \mathbb{L}(\Phi_L(\mathbf{x}; \theta), y) \; d\rho_k(\mathbf{x}, y) \tag{E.3}$$

for some *smooth* loss function $\mathbb{L} : \mathbb{Y} \times \mathbb{Y} \to \mathbb{R}$. Note that

$$\begin{aligned}
\mathrm{d}_{W_\ell} \mathcal{L}^k_\theta[H] &= \int \mathrm{d}_1 \mathbb{L}_{(\Phi(\mathbf{x};\theta),y)}[\mathrm{d}_{W_\ell} \Phi_{(x;\theta)}[H]] \; d\rho_k(\mathbf{x}, y) \\
&= \int \langle \; \nabla_1 \mathbb{L}_{(\Phi(\mathbf{x};\theta),y)} \; , \; \mathrm{d}_{W_\ell} \Phi_{(x;\theta)}[H] \; \rangle \; d\rho_k(\mathbf{x}, y) \\
&= \int \langle \; (\mathrm{d}_{W_\ell} \Phi_{(x;\theta)})^\top [\nabla_1 \mathbb{L}(\Phi(\mathbf{x}; \theta), y)] \; , \; H \; \rangle \; d\rho_k(\mathbf{x}, y) \\
&= \langle \; \int (\mathrm{d}_{W_\ell} \Phi_{(x;\theta)})^\top [\nabla_1 \mathbb{L}(\Phi(\mathbf{x}; \theta), y)] \; d\rho_k(\mathbf{x}, y) \; , \; H \; \rangle
\end{aligned}$$

so that

$$\nabla_{W_\ell} \mathcal{L}^k(\theta) = \int (\mathrm{d}_{W_\ell} \Phi_{(x;\theta)})^\top [\; \nabla_1 \mathbb{L}(\Phi(\mathbf{x}; \theta), y) \;] \; d\rho_k(\mathbf{x}, y) \in \mathbb{R}^{N_\ell \times N_{\ell+1}}. \tag{E.4}$$

We note that $\mathrm{d}_{W_\ell} \Phi_{(x;\theta)} : \mathbb{R}^{N_\ell \times N_{\ell+1}} \to \mathbb{R}^{N_{L+1}}$ and $\nabla_1 \mathbb{L}(\Phi(\mathbf{x}; \theta), y) \in \mathbb{R}^{N_{L+1}}$.

[1] Here $\lambda_\ell < 0$, we write it like this for simplicity, not to keep track of the sign explicitly.

E.1 The Linear Case

In what follows we analyze the linear case of Chizat et al. [13], taking $\varphi_\ell = \mathrm{id}$ for all ℓ. We recast their constructions in our graphical framework, which, we argue, affords a clearer view of the underlying algebraic mechanisms of the dynamics.

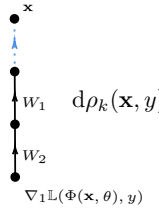
For clarity we restrict the exposition with the depth-2 model, $\Phi(\mathbf{x};\theta) = W_2W_1W_0\mathbf{x}$.

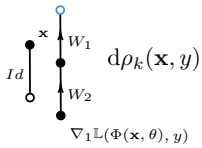
E.1.1 Graphical expansion

Naturally, the starting point is writing eq. (E.2) and eq. (E.4) in terms of operator graphs, only then we will be able to follow the update rule. Recall how as an operator graph our linear network Φ is represented by

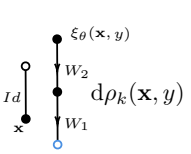
$\Phi(\mathbf{x},\theta) \;=$

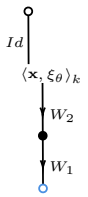
so that, as an operator graph, the gradient of the loss (eq. (E.4)) becomes

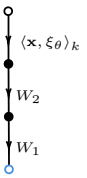
$\nabla_{W_0}\mathcal{L}^k_\theta \;=\; \int$ 

$\mathrm{d}\rho_k(\mathbf{x},y) \;=\; \int$ 

$\mathrm{d}\rho_k(\mathbf{x},y)$

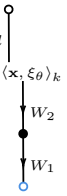
$= \int$ 

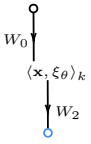
$\mathrm{d}\rho_k(\mathbf{x},y) \;=:$ 

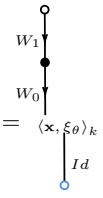
$=:$ 

where we write $\langle \mathbf{x},\xi_\theta\rangle_k := \int \mathbf{x}\otimes\xi_\theta(\mathbf{x},y)\; d\rho_k(\mathbf{x},y) \in \mathbb{R}^{d_{in}} \times \mathbb{R}^{d_{out}}$.

Similarly for other ℓs we obtain

$\nabla_{W_0}\mathcal{L}^k_\theta \;=$ 

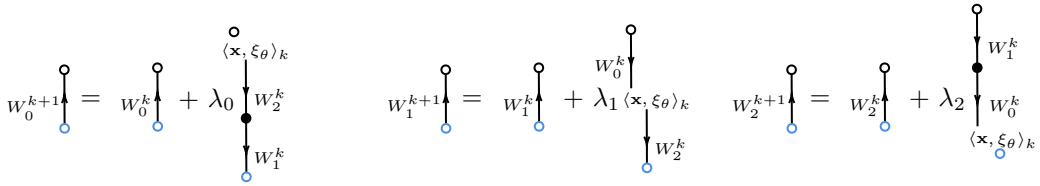
$\nabla_{W_1}\mathcal{L}^k_\theta \;=$ 

$\nabla_{W_2}\mathcal{L}^k_\theta \;=$ 

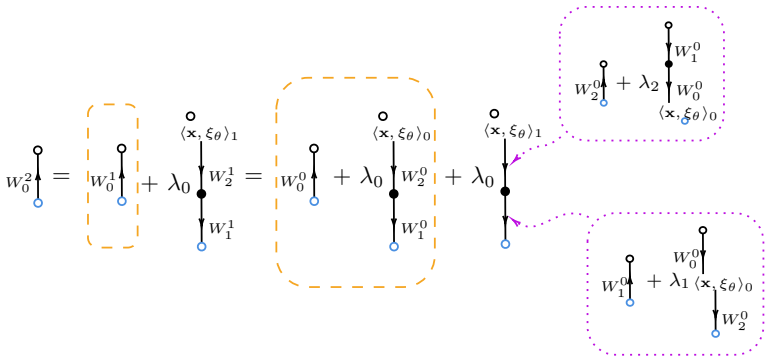
Observe how this operation corresponds to substituting in $\Phi(\cdot)$ the matrix with respect to which we are differentiating with $\langle \mathbf{x},\xi_\theta\rangle_k$, and then transposing the graphs attached to this edge's in and out-vertices.

E.1.2 Training dynamics

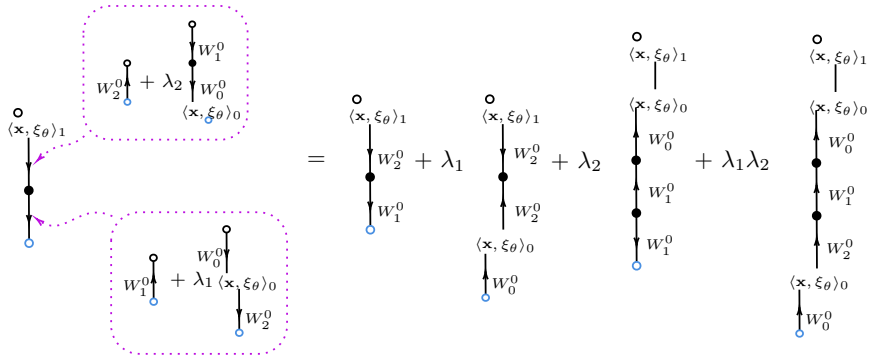
Let us now look at the *graphical training dynamics*: in terms of operator graphs the update rules (E.2) dictate the following equalities



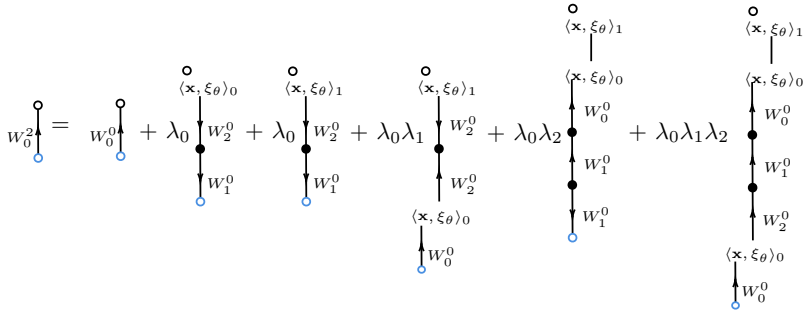
Consider then the matrix W_0 after two updates, the operator graph representation of W_0^2 is obtained as



where in orange we have highlighted the application of the update rule to W_0^1 and in violet the implied one for W_1^1 and W_2^1. Since we can expand the last term as

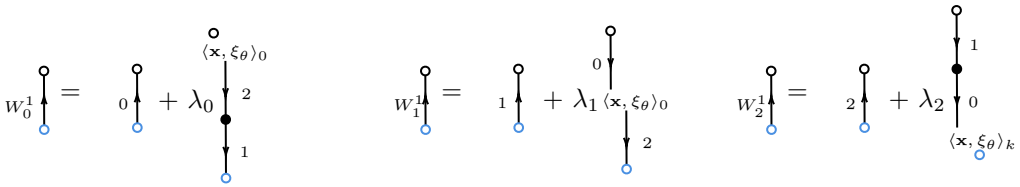


we finally get to the complete expansion of W_0^2 in terms of the original W_i^0:

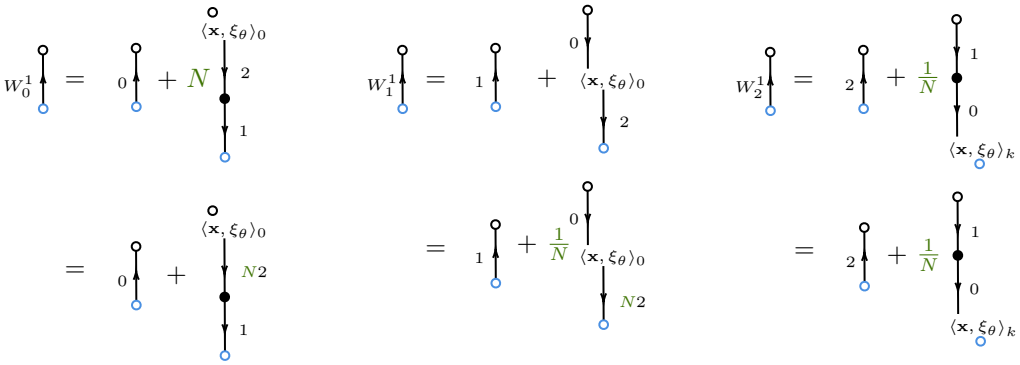


The usefulness of this reduction in terms of the original matrices lies in the fact that the distributional properties of the initial weights W_i^0 are completely understood, which enables a systematic study of operator graphs through the tools introduced earlier. Although this procedure can be iterated arbitrarily, it does so at the cost of rapidly increasing combinatorial complexity. The central insight of Chizat et al. [13] is to circumvent this difficulty by introducing a simpler system with the same asymptotic properties—in particular, one reproducing the same *bi-atomic* pairings.

Let us continue the study of the training dynamics, now using our machinery to study the asymptotic properties of these objects, starting from the first gradient step:



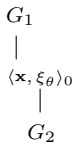
where we have decorated edges with $\ell \equiv W_\ell^0$ since we will write everything in terms of these starting matrices. For simplicity we place ourselves in the μP case, but this is the point where one can reason for general parametrizations as well. Under μP one sets $(\lambda_0, \lambda_1, \lambda_2) = (N, 1, \frac{1}{N})$ and $(\sigma_0, \sigma_1, \sigma_2) = (1, \frac{1}{\sqrt{N}}, \frac{1}{N})$ so that



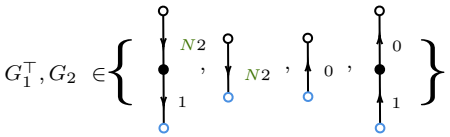
This rewriting is useful since both NW_2^0 and W_0^0 have variance 1 and are the matrices adjacent to the *bounded*-dimensional vertices labelled by $\mathbf{x}$ and ξ, think of them as collections of a fixed size of standard multivariate Gaussians of arbitrary dimension. Moreover the $\frac{1}{N}$ factors in ΔW_1^1 and ΔW_2^1 will force a *dot-product* as we will shortly see.

Remark. *Before proceeding a quick word on the terms $\langle \mathbf{x}, \xi_\theta \rangle_k$: these are random 2-tensors, dependent on W_0, W_1, W_2; fortunately by carefully leveraging bounds on the loss function one can prove that the part which does* not *vanish asymptoticallu is actually independent from them (*c.f. *Chizat et al. [13][Step 3]). As such we will treat these terms as deterministic in this simplified exposition.*

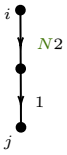
After the first update, all the increments we are dealing with have form

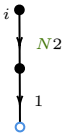
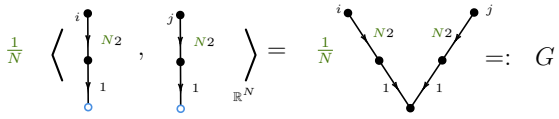


for ladders G_1, G_2 such that



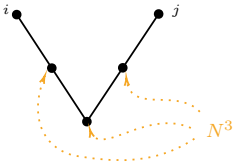
Let's use our graphical tools to study the entries of these matrices in L^2 sense:

The entries of the first matrix  all have mean 0 since there are no *atomic* partitions of itself and variance $\sigma_{N2}^2 \sigma_1^2 N = 1\frac{1}{N}N$. We can prove that these are actually iid $\mathcal{N}(0,1)$ with method of moments.

Consider the columns $\in \mathbb{R}^N$  and consider the scaled L^2 products 

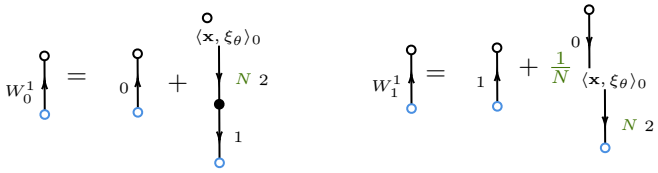
We evince that $\mathbb{E}[\mathbf{W}_G] = \delta_{i,j}$ since partitions exist iff $i = j$, in which case only one exists and for it $\mathbf{W}_{G_\phi} = \frac{1}{N}\sigma_{N2}^2 \sigma_1^2 N^2 = \frac{1}{N} 1 \frac{1}{N} N^2 = 1$. Since this is also the only atomic partition we have that

$$\mathrm{Var}(\mathbf{W}_G) = |\{\phi \text{ bi-atomic for } G\}| \cdot \frac{1}{N^2}\sigma_{2N}^4 \sigma_1^4 \cdot N^3 \propto \frac{1}{N^4} \cdot N^3 = \frac{1}{N} \to 0$$

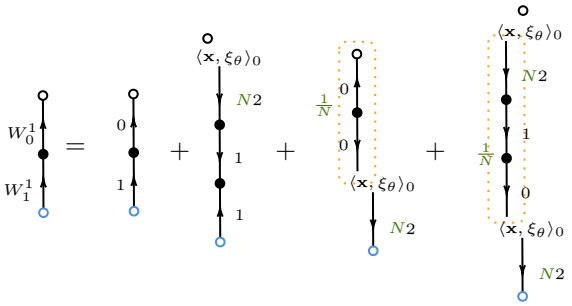
In fact all *bi-atomic* partitions result in $\mathbf{W}_{G_\phi} = \frac{1}{N^2}\sigma_{N2}^4 \sigma_1^4$ 

We have just proved that the columns of $(W_1^0)^\top (N W_2^0)^\top$ are orthonormal in a limiting L^2 sense. In this way we can prove that all the matrices corresponding to the ladders above have iid standard Gaussian entries, and that their columns are orthonormal in this limiting L^2 sense.

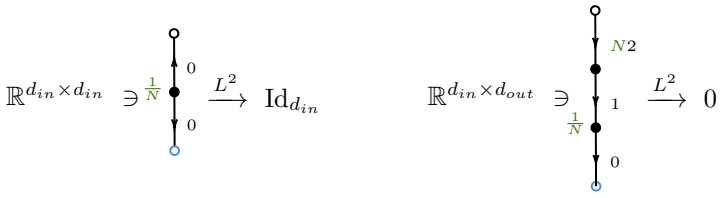
We are now ready to compute the full network after 1 gradient step, one matrix multiplication at the time. Recall first the expansions of W_0^1 and W_1^1



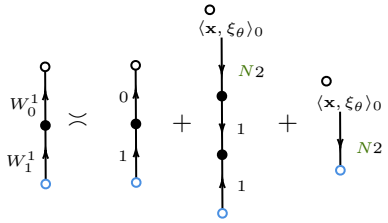
so that (where we have highlighted the components we are about to consider in more detail) their composition is



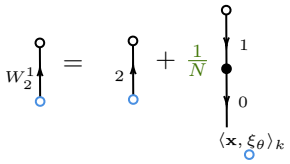
From our study above we know that



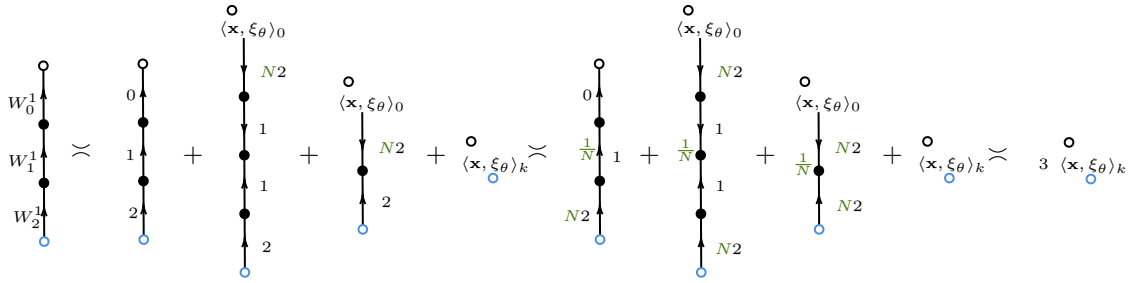
hence we can conclude that $W_1^1 W_0^1$ behaves, at the limit, just as



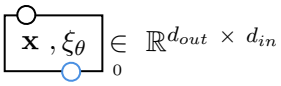
In the same vein, since



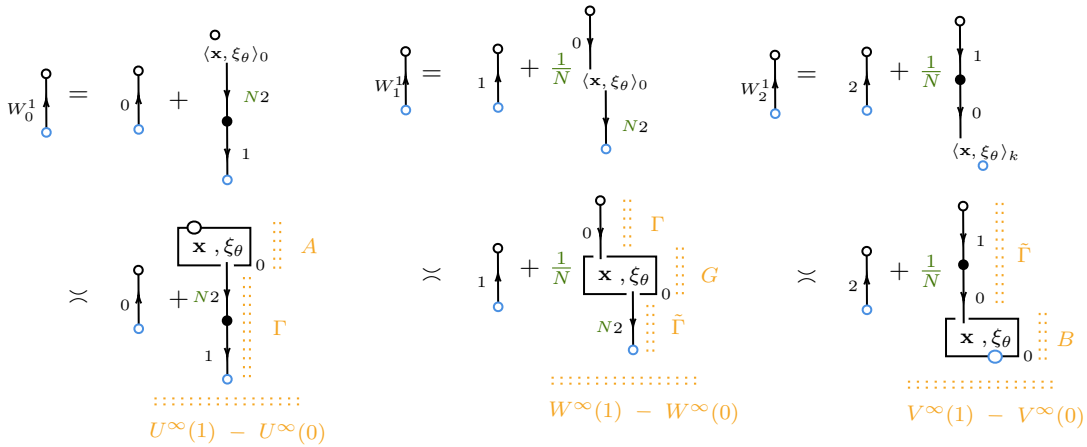
we obtain



Note how $W_2^0 W_1^0 W_0^0 \xrightarrow{L^2} 0$, and how we are able to compose the limiting operations as the presence of deterministic matrices rules out the possibility of new,un-accounted for, non-vanishing pairings. Bounding the loss function we can prove that $\langle \mathbf{x}, \xi \rangle_0$ converges in L^2 to a deterministic quantity (what Chizat et al. [13] call $\xi_{0,\tau}^{\top}$ in (2.10))

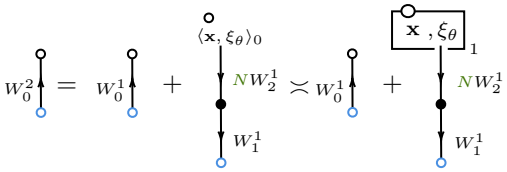


Using this we can write (in orange the corresponding objects in Chizat et al. [13])

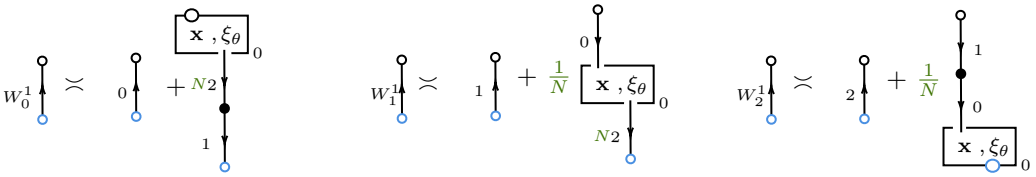


Remark. *The entries of these matrices behave, in the limit, as $Z_1^{\top} M Z_2$ for standard Gaussians Z_1, Z_2 and deterministic matrices M.*

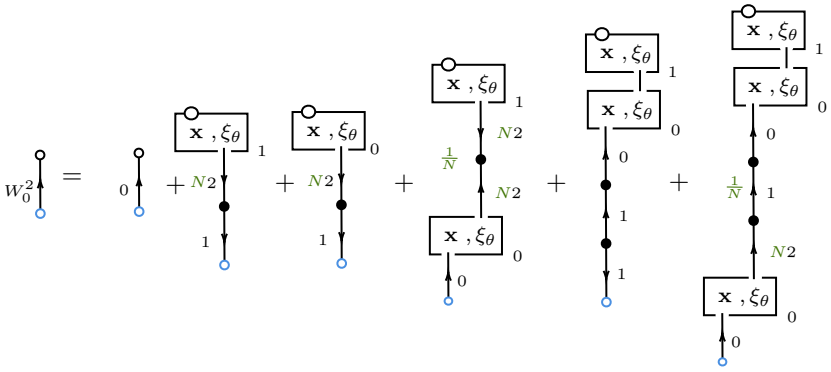
With the first gradient step under our belt, let's consider the successive stages:



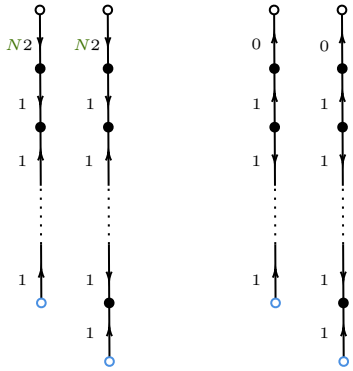
so that, since we already know that



we obtain



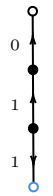
Once again we see how we can write the entries in terms of deterministic matrices multiplying operators corresponding to the matrices represented by



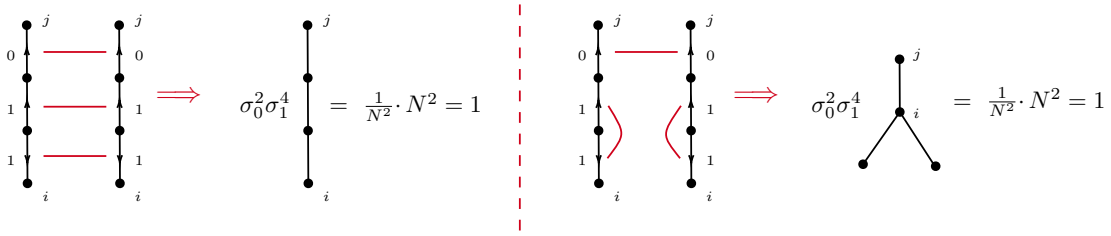
By induction one proves this happens for all layers and steps.

In the $k = 1$ case the columns of the ladders above were orthonormal in the limiting L^2 sense, behaving as vectors with standard normal entries. This is not the case anymore:

take as example the following term appearing in W_0^2:

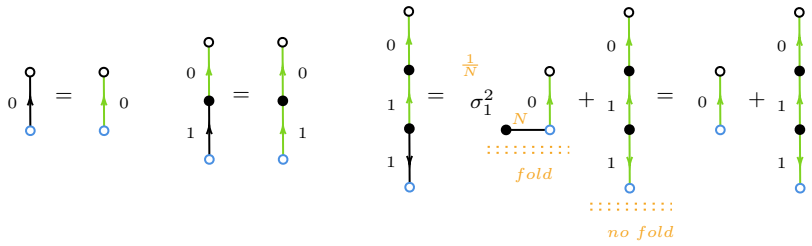


Its entries have mean 0 since this graph admits no admissible pairings, but the variance is 2 since it admits the following *bi-atomic* partitions

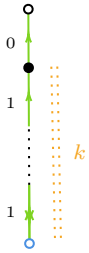


Still, the convergence is to a Gaussian by moments method. Note that the added complexity arises from the terms corresponding to $(W_1^0)^\top(W_1^0)$, since partitions can *fold* such pieces! Chizat et al. [13] control this problem by providing a decomposition into terms which cannot fold by contruction. This is done by enforcing the indices in the definition of product graphs to be different,

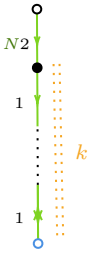
Graphically we use the color green to stand for parts of graphs unable to pair to other green pieces of themselves. Then



Chizat et al. [13] call the



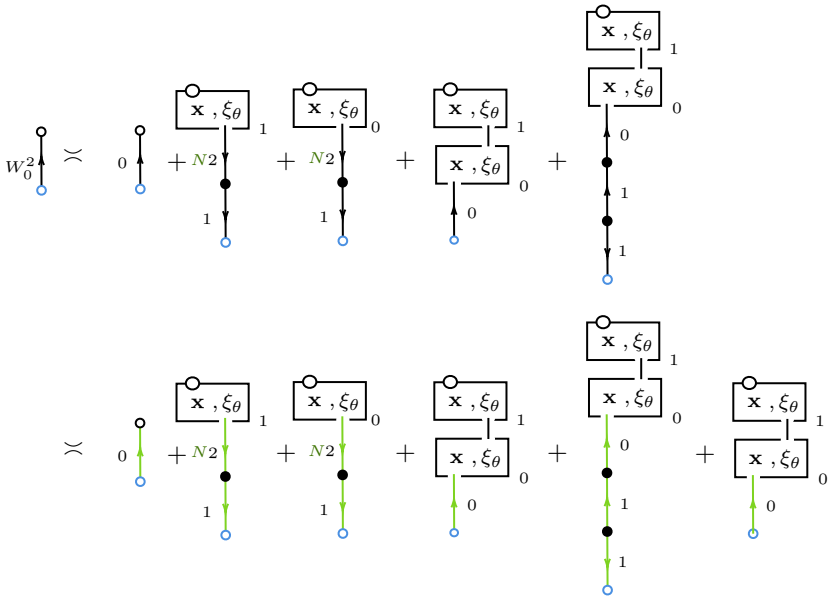
and the



respectively J_k and $K_k^\top$.

These J_k and $K_k^\top$ are orthogonal, in the limit, as the only *bi-atomic* pairings possible are the ones with exact copies of themselves since no self-pairing, like the folds singled out above, is allowed. This result, which readily follows from our theory, is the subject of Chizat et al. [13, Theorem 3.2, Proposition 3.6]

Then, for example,

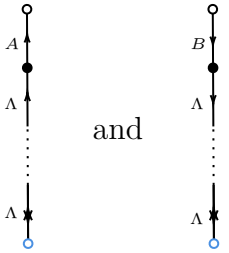


Chizat et al. [13] write everything in terms of these orthonormal objects, which are easy to manage and recursively handle. Note also that at every step the limiting Φ is a deterministic object, since the J_k and K_k disappear in the dot-products. Moreover as in the $t = 1$ case, the limiting distributions of the matrices are of the form $\sum_{i,j} Z_i^\top M_{ij} Z_j$ for Gaussian Zs and deterministic matrices M, deterministic matrices which graphically constitute the building blocks for the limiting Φ.

Finally, the last step of Chizat et al. [13] can be interpreted as that of finding fixed linear operators

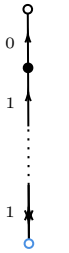
$$A : \mathbb{R}^{d_{in}} \to \mathcal{H}_1, \quad B : \mathcal{H}_1 \to \mathcal{H}_2, \quad C : \mathcal{H}_2 \to \mathbb{R}^{d_{out}}$$

between Hilbert spaces $\mathcal{H}_1, \mathcal{H}_2 \simeq \ell^2$ such that the operators

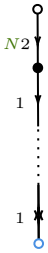


and

have the same behavior (as detailed in the next section) as the limiting L^2 one of the matrices



and



Running gradient descent with rates $(\lambda_A, \lambda_\Lambda, \lambda_B) = (1, \frac{1}{N}, \frac{1}{N})$ on

$$\chi : \mathbb{R}^{d_{in}} \to \mathbb{R}^{d_{in}}, \quad \mathbf{x} \mapsto (B \circ \Lambda \circ A)(\mathbf{x})$$

would then produce the L^2-limiting training dynamics of Φ as $N \to \infty$, just as in Chizat et al. [13, Theorem 2.2 and Eq. 2.16].

E.1.3 The L^2 limit

Let us then look for such operators mimicking the limiting dynamics. Using our graphical tools we have written these dynamics in terms of the matrices J_k, K_k defined as[2]

$$J_{2k} = ((W_1)^\top W_1)^k W_0 \in \mathbb{R}^{N_1 \times d_{in}}, \quad J_{2k+1} = W_1((W_1)^\top W_1)^k W_0 \in \mathbb{R}^{N_2 \times d_{in}}$$

$$K_{2k} = (W_1(W_1)^\top)^k (W_2)^\top \in \mathbb{R}^{N_2 \times d_{out}}, \quad K_{2k+1} = (W_1)^\top (W_1(W_1)^\top)^k (W_2)^\top \in \mathbb{R}^{N_1 \times d_{out}}$$

Proposition E.1 (Chizat et al. [13], Proposition 3.3)**.** *The matrices J_m, K_m satisfy the following equalities*

$$\frac{1}{N}\langle J_k[\mathbf{e}_i], J_{k'}[\mathbf{e}_j]\rangle \asymp \lim_{N\to\infty} \frac{1}{N}\langle K_k[\mathbf{e}_i], K_{k'}[\mathbf{e}_j]\rangle \asymp \delta_{i,j}\delta_{k,k'}, \tag{E.5}$$

$$\frac{1}{N}\langle J_k[\mathbf{e}_i], K_{k'}[\mathbf{e}_j]\rangle \asymp 0 \tag{E.6}$$

and also, setting $J_{-1} = 0$ and $K_{-1} = 0$,

$$W_1 J_{2k} \asymp J_{2k+1} + J_{2k-1}, \quad (W_1)^\top J_{2k+1} \asymp J_{2k+2} + J_{2k} \tag{E.7}$$

$$(W_1)^\top K_{2k} \asymp K_{2k+1} + K_{2k-1}, \quad W_1 K_{2k+1} \asymp K_{2k+2} + K_{2k} \tag{E.8}$$

Proof. This is an immediate consequence of the definition of these objects and theorem 1.17. □

Direct Systems

Consider then the Hilbert Spaces

$$\mathcal{H}_1^N := (Span\{J_{2k}[e_i], K_{2k+1}[e_j] \mid k \in \mathbb{N},\ i, j\},\ \frac{1}{N}\langle\cdot,\cdot\rangle_{N_1}) \tag{E.9}$$

$$\mathcal{H}_2^N := (Span\{J_{2k+1}[e_i], K_{2k}[e_j] \mid k \in \mathbb{N},\ i, j\},\ \frac{1}{N}\langle\cdot,\cdot\rangle_{N_2}) \tag{E.10}$$

If we consider the maps $\phi^1_{n,m} : \mathcal{H}_1^n \to \mathcal{H}_1^m$ given by $J_{2k}, K_{2k+1} \mapsto J_{2k}, K_{2k+1}$ and similarly $\phi^2_{n,m} : \mathcal{H}_2^n \to \mathcal{H}_2^m$ then

$$(\mathcal{H}_1^n, \phi^1_{n,m}), \quad (\mathcal{H}_2^n, \phi^2_{n,m})$$

define *direct systems* over $\mathbb{N}$ thus producing *direct limits* $\mathcal{H}_1^\infty$ and $\mathcal{H}_2^\infty$.

Definition E.2 (Direct Limit)**.** *A directed system of vector spaces over $\mathbb{N}$ is the pair of sequences $((V_n)_{n\in\mathbb{N}}, (f_{n,m})_{n\le m\in\mathbb{N}})$ where*

[2] Recall here $d_{in} := N_0$ and $d_{out} = N_3$ are fixed dimensions, while N_1 and N_2 scale as $\Omega(N)$, a dependence which is hidden in the notation.

- *Each V_n is a vector space and the $f_{n,m} : V_n \to V_m$ are linear maps.*
- *For all $n \in \mathbb{N}$ holds $f_{n,n} = Id_{V_n}$.*
- *For all $n \leq k \leq m$ holds $f_{n,m} = f_{k,m} \circ f_{n,k}$.*

The direct limit of the directed system $((V_n)_{n\in\mathbb{N}}, (f_{n,m})_{n\leq m\in\mathbb{N}})$ is defined as the quotient space

$$V_\infty := \left(\bigsqcup_{n\in\mathbb{N}} V_n\right) / \sim \tag{E.11}$$

whith the equivalence relation defined as

$$\forall n \leq m.\ \forall x \in V_n.\ x \sim f_{n,m}(x)$$

i.e. *each element of the product is equivalent to all its images under the $f_{n,m}$. In this case we write f_n for the quotient maps induced on each V_n.*

Since the scalar products of the spaces converge pointwise their limit defines a scalar prouct on the corresponding direct limit, leading to Hilbert Spaces

$$(\mathcal{H}_1^\infty, \langle\cdot,\cdot\rangle_1) \text{ and } (\mathcal{H}_2^\infty, \langle\cdot,\cdot\rangle_2) \tag{E.12}$$

Remark. *Being direct limits these spaces are* universal targets. *This means, taking $\mathcal{H}_1^\infty$ as example, that for any vector space Y and linear maps $\psi_n : \mathcal{H}_1^n \to Y$ satisfying $\forall n \leq m.\ \psi_m \circ \phi_{n,m}^1 = \psi_n$ there exist a map $u : \mathcal{H}_1^\infty \to Y$ such that $\forall n.\ \psi_n = u \circ \phi_n^1$.*

Note that this pair of spaces is not unrelated thanks to (E.7) and (E.8), we would like to find a "transition" map $W_1^\infty : \mathcal{H}_1^\infty \to \mathcal{H}_2^\infty$ corresponding to W_1^n "at the limit" but we cannot apply the *universality property* of direct limits right away since (E.7) and (E.8) are not equalities. Fortunately the following result proves that the space Y has to be a target only in the limit:

Proposition E.3. *Fix a direct system of Hilbert spaces $(\mathcal{H}^n, \phi_{n,m})$ with direct limit $(\mathcal{H}^\infty, \phi_n)$ and uniformly bounded $\|\phi_{n,m}\|_{op}$. Given a Hilbert space Y and linear functions $\psi_n : \mathcal{H}^n \to Y$ satisfying $\forall n \leq m.\ \|\psi_m \circ \phi_{n,m} - \psi_n\|_{op} = o_n(1)$ uniformly in m, there exists a map $u : \mathcal{H}^\infty \to Y$ such that $\forall n.\ \|u \circ \phi_n - \psi_n\|_{op} = o_n(1)$.*

Proof. Note how for any $n \leq m \leq m'$ one has

$$\begin{aligned}
\psi_{m'} \circ \phi_{n,m'} - \psi_m \circ \phi_{n,m} &= \psi_{m'} \circ \phi_{m,m'} \circ \phi_{n,m} - \psi_m \circ \phi_{n,m} \\
&= (\psi_{m'} \circ \phi_{m,m'} - \psi_m) \circ \phi_{n,m} \\
&= o_m(1) \circ \phi_{n,m} = o_m(1)
\end{aligned}$$

so that for all $x \in \mathcal{H}^n$. $\lim_{m\to\infty} \psi_m \circ \phi_{n,m}(x) \in Y$ is well defined by completeness of

Hilbert spaces. Define then $u : \mathcal{H}^\infty \to Y$ as

$$\forall n. \forall x \in \mathcal{H}^n.\ u(\phi_n(x)) := \lim_{m\to\infty} \psi_m \circ \phi_{n,m}(x) \tag{E.13}$$

This is a well definition since

$$\begin{aligned} u(\phi_n(x)) &= \lim_{m'\to\infty} \psi_{m'} \circ \phi^1_{n,m'}(x) \\ &= \lim_{m'\to\infty} \psi_{m'} \circ \phi^1_{m,m'} \circ \phi^1_{n,m}(x) \\ &= \lim_{m'\to\infty} \psi_{m'} \circ \phi^1_{m,m'} \circ \phi^1_{n,m}(x) = u(\phi_m \circ \phi_{n,m}(x)) \end{aligned}$$

moreover, and this concludes the proof, for all n

$$\left\| u \circ \phi^1_n - \psi_n \right\|_{op} \leq \sup_{m\geq n} \left\| \psi_m \circ \phi^1_{n,m} - \psi_n \right\|_{op} = o_n(1) \tag{E.14}$$

□

Considering diagram E.1 and theorem E.3 we obtain the maps $W_1^\infty : \mathcal{H}_1^\infty \to \mathcal{H}_2^\infty$ and $(W_1^\infty)^\top : \mathcal{H}_2^\infty \to \mathcal{H}_1^\infty$ relating the two spaces.

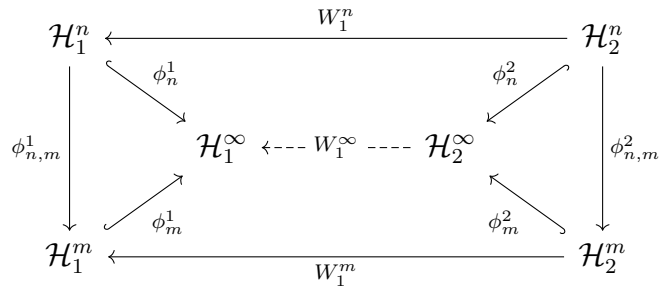


Figure E.1: *The "approximatively commutative" diagram defining W_1^∞.*

If we define $J_{2k}^\infty := \phi_n^1(J_{2k})$ and similarly for the other basis elements we see that these are orthonormal basis elements and

$$W_1^\infty J_{2k}^\infty = J_{2k+1}^\infty + J_{2k-1}^\infty, \quad (W_1^\infty)^\top J_{2k+1}^\infty = J_{2k+2}^\infty + J_{2k}^\infty \tag{E.15}$$

$$(W_1^\infty)^\top K_{2k}^\infty = K_{2k+1}^\infty + K_{2k-1}^\infty, \quad W_1^\infty K_{2k+1}^\infty = K_{2k+2}^\infty + K_{2k}^\infty \tag{E.16}$$

Similarly we are able to define the maps in diagram E.2 and we see how

$$W_0^\infty = J_0^\infty, \quad W_2^\infty = (K_0^\infty)^\top \tag{E.17}$$

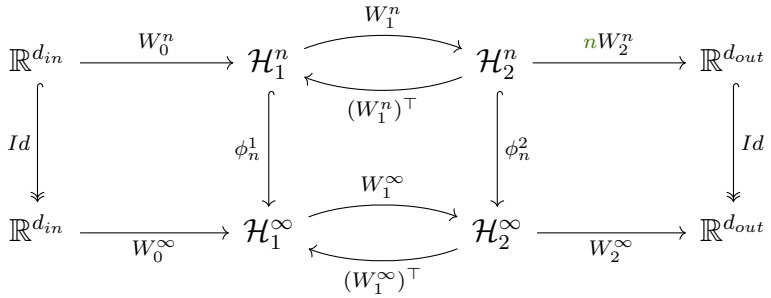


Figure E.2: *The limiting diagram for our directed systems.*

Explicit limiting spaces

Working with these direct limits is not practical, therefore we look for representations of these spaces in more practical terms. In particular it would useful to have *isometric embeddings* ι_1, ι_2 of these spaces into some L^2 spaces $\mathbb{H}_1, \mathbb{H}_2$.

We show that this is possible in the linear case, where the isometric embedding is in a space of polynomials.

Theorem E.4. *Consider independent random variables*

$$Z \sim \mathcal{N}(0, Id_{d_{in}}), \quad \tilde{Z} \sim \mathcal{N}(0, Id_{d_{out}})$$

and let $\rho(\mathbf{z}, \tilde{\mathbf{z}})$ be the density of $(Z, \tilde{Z})$ on $\mathbb{R}^{d_{in}+d_{out}}$. Let $\mathbb{H}_1 = \mathbb{H}_2$ be the closure of

$$Span\{\begin{bmatrix} W_n(x)Z_i \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ U_n(x)\tilde{Z}_j \end{bmatrix}, \mid n \in \mathbb{N}, i \in [d_{in}], j \in [d_{out}]\} \tag{E.18}$$

in the space $L^2([-1,1] \times \mathbb{R}^{d_{in}+d_{out}}; \mu_W \otimes \rho) \oplus L^2([-1,1] \times \mathbb{R}^{d_{in}+d_{out}}; \mu_W \otimes \rho)$ for

$$W_n(x) := \frac{\sin\left((n+\frac{1}{2})\arccos(x)\right)}{\sin\left(\frac{1}{2}\arccos(x)\right)}, \quad U_n(x) := \frac{\sin\left((n+1)\arccos(x)\right)}{\sin\left(\arccos(x)\right)} \tag{E.19}$$

and $d\mu_W(x) := \frac{1}{\pi}\sqrt{\frac{1-x}{1+x}}dx$ and $d\mu_U(x) := \frac{2}{\pi}\sqrt{(1-x)(1+x)}dx$.

Then the maps ι_1, ι_2 defined on basis elements by

$$\iota_1 J_{2n}[e_i] := \begin{bmatrix} W_n(x)z_i \\ 0 \end{bmatrix}, \quad \iota_1 K_{2n+1}[e_j] := \begin{bmatrix} 0 \\ U_n(x)\tilde{z}_j \end{bmatrix} \tag{E.20}$$

$$\iota_2 J_{2n+1}[e_i] := \begin{bmatrix} 0 \\ U_n(x)z_i \end{bmatrix}, \quad \iota_2 K_{2n}[e_j] := \begin{bmatrix} W_n(x)\tilde{z}_j \\ 0 \end{bmatrix} \tag{E.21}$$

define isometric embeddings.

Proof. Start noting how from (E.15) we obtain

$$\begin{aligned}(W_1^\infty)^\top W_1^\infty J_{2k}^\infty &= (W_1^\infty)^\top [J_{2k+1}^\infty + J_{2k-1}^\infty] \\ &= J_{2k+2}^\infty + (2-\delta_{k,0}) J_{2k}^\infty + J_{2k-2}^\infty\end{aligned}$$

so that setting $\Sigma_1 := (W_1^\infty)^\top W_1^\infty$

$$J_2^\infty = (\Sigma_1 - 1) J_0^\infty, \quad J_{2(k+1)}^\infty = (\Sigma_1 - 2) J_{2k}^\infty - J_{2(k-1)}^\infty \tag{E.22}$$

Similarly

$$J_3^\infty = (\Sigma_1^\top - 2) J_1^\infty, \quad J_{2(k+1)+1}^\infty = (\Sigma_1^\top - 2) J_{2k+1}^\infty - J_{2(k-1)+1}^\infty \tag{E.23}$$

$$K_2^\infty = (\Sigma_1^\top - 1) K_0^\infty, \quad K_{2(k+1)}^\infty = (\Sigma_1^\top - 2) K_{2k}^\infty - K_{2(k-1)}^\infty \tag{E.24}$$

$$K_3^\infty = (\Sigma_1 - 2) K_1^\infty, \quad K_{2(k+1)+1}^\infty = (\Sigma_1 - 2) K_{2k+1}^\infty - K_{2(k-1)+1}^\infty \tag{E.25}$$

Consider $(J_{2k} \mid k \in \mathbb{N})$. To find the images $\iota_1 J_{2k}[e_i]$ we look for a space of orthogonal polynomials $(P_k \mid k \in \mathbb{N})$ satisfying the recurrence relation

$$P_1(x) = (\tilde{\Sigma}_1 - 1) P_0(x), \quad P_{k+1}(x) = (\tilde{\Sigma}_1 - 2) P_k(x) - P_{k-1}(x)$$

for the operator $\tilde{\Sigma}_1 := \iota_1 \Sigma_1 \iota_1^{-1}$. These properties are satisfied by the *Chebyshev Polynomials of the fourth kind* [82], also known as *airfoil polynomials* [110], defined by

$$W_n(x) := \frac{\sin\left((n+\frac{1}{2})\arccos(x)\right)}{\sin\left(\frac{1}{2}\arccos(x)\right)} \tag{E.26}$$

with the choice $\tilde{\Sigma}_1 p(x) := 2(x+1)p(x)$ given that

$$W_0(x) = 1, \quad W_1(x) = 2x + 1 = 2(x+1) - 1, \tag{E.27}$$

$$W_{k+1}(x) = 2xW_k(x) - W_{k-1}(x) = (2(x+1) - 2)W_k(x) - W_{k-1}(x). \tag{E.28}$$

Since

$$\int_{-1}^{1} W_n(x) W_m(x) \frac{1}{\pi} \sqrt{\frac{1-x}{1+x}} dx = \delta_{n,m}$$

the polynomials we look for can be defined as

$$P_n(x) := W_n(x) \tag{E.29}$$

orthonormal under $d\mu_W(x) := \frac{1}{\pi}\sqrt{\frac{1-x}{1+x}}dx$ on $[-1, 1]$.

Similarly for $(J_{2k+1} \mid k \in \mathbb{N})$ we look for a space of orthogonal polynomials $(Q_k \mid k \in$

$\mathbb{N}$) satisfying the recurrence relation

$$Q_1(x) = (\tilde{\Sigma}_1^\top - 2)Q_0(x), \quad Q_{k+1}(x) = (\tilde{\Sigma}_1^\top - 2)Q_k(x) - Q_{k-1}(x).$$

Chebyshev Polynomials of the second kind defined by

$$U_n(x) := \frac{\sin\left((n+1)\arccos(x)\right)}{\sin\left(\arccos(x)\right)} \tag{E.30}$$

fit the bill with the choice $\tilde{\Sigma}_1^\top p(x) := 2(x+1)p(x)$ given that

$$U_0(x) = 1, \quad U_1(x) = 2x = 2(x+1) - 2, \tag{E.31}$$

$$U_{k+1}(x) = 2xU_k(x) - U_{k-1}(x) = (2(x+1) - 2)U_k(x) - U_{k-1}(x) \tag{E.32}$$

Since

$$\int_{-1}^{1} U_n(x)U_m(x)\frac{2}{\pi}\sqrt{(1-x)(1+x)}dx = \delta_{n,m}$$

the polynomials we look for can be defined as

$$Q_n(x) := U_n(x) \tag{E.33}$$

orthonormal under $d\mu_U(x) := \frac{2}{\pi}\sqrt{(1-x)(1+x)}dx$ on $[-1, 1]$.

Consider defining the spaces $\mathbb{H}_1$ and $\mathbb{H}_2$ by

$$\iota_1 J_{2n}[e_i] := W_n(x), \quad \iota_1 K_{2n+1}[e_j] := U_n(x) \tag{E.34}$$

$$\iota_2 J_{2n+1}[e_i] := U_n(x), \quad \iota_2 K_{2n}[e_j] := W_n(x). \tag{E.35}$$

Doing so would satisfy the images of eqs. (E.15) and (E.16) hold in these polynomial spaces. In fact these two families of polynomials are related by the following equalities [110]:

$$W_n(x) = U_n(x) + U_{n-1}(x) \tag{E.36}$$

$$2(x+1)U_n(x) = W_{n+1}(x) + W_n(x). \tag{E.37}$$

This implies, using eq. (E.15), both

$$\iota_2 W_1^\infty \iota_1^{-1} : W_n(x) \to W_n(x) = U_n(x) + U_{n-1}(x)$$

and also

$$\iota_1 (W_1^\infty)^\top \iota_2^{-1} = (\iota_2 W_1^\infty \iota_1^{-1})^\top : U_n(x) \mapsto 2(x+1)U_n(x) = W_{n+1}(x) + W_n(x),$$

this last one since

$$\begin{aligned}\langle(\iota_2 W_1^\infty \iota_1^{-1})W_n, U_m\rangle_{\mu_U} &= \langle W_n, U_m\rangle_{\mu_U} \\ &= \langle U_n + U_{n-1}, U_m\rangle_{\mu_U} \\ &= \delta_{n,m} + \delta_{n-1,m} \\ &= \delta_{n,m} + \delta_{n,m+1} \\ &= \langle W_n, W_m + W_{m+1}\rangle_{\mu_W} \\ &= \langle W_n, 2(x+1)U_m\rangle_{\mu_W}\end{aligned}$$

Completely analogous arguments hold for K using eq. (E.16).

But such a definition of $\mathbb{H}_1$ and $\mathbb{H}_2$ lacks one last crucial property: in order to prove that ι_1, ι_2 are isometries is then left to considered the orthogonality condition

$$\langle \iota_1 J_{2n}[e_i], \iota_1 J_{2m}[e_j]\rangle_{\mathbb{H}_1} = \delta_{m,n}\delta_{i,j}$$

across different choices of e_i, e_j.

This is why in the theorem statement we make use of independent Gaussians. Consider in fact independent r.v.

$$Z \sim \mathcal{N}(0, Id_{d_{in}}), \quad \tilde{Z} \sim \mathcal{N}(0, Id_{d_{out}})$$

and let $\rho(\mathbf{z}, \tilde{\mathbf{z}})$ be the density of $(Z, \tilde{Z})$ on $\mathbb{R}^{d_{in}+d_{out}}$. Define finally $\mathbb{H}_1 = \mathbb{H}_2$ as the closure of

$$Span\{\begin{bmatrix} W_n(x)z_i \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ U_n(x)\tilde{z}_j \end{bmatrix}, \mid n \in \mathbb{N}, i \in [d_{in}], j \in [d_{out}]\} \tag{E.38}$$

in the space $L^2([-1,1] \times \mathbb{R}^{d_{in}+d_{out}}; \mu_W \otimes \rho) \oplus L^2([-1,1] \times \mathbb{R}^{d_{in}+d_{out}}; \mu_W \otimes \rho)$ and define ι_1, ι_2 by

$$\iota_1 J_{2n}[e_i] := \begin{bmatrix} W_n(x)z_i \\ 0 \end{bmatrix}, \quad \iota_1 K_{2n+1}[e_j] := \begin{bmatrix} 0 \\ U_n(x)\tilde{z}_j \end{bmatrix} \tag{E.39}$$

$$\iota_2 J_{2n+1}[e_i] := \begin{bmatrix} 0 \\ U_n(x)z_i \end{bmatrix}, \quad \iota_2 K_{2n}[e_j] := \begin{bmatrix} W_n(x)\tilde{z}_j \\ 0 \end{bmatrix} \tag{E.40}$$

to obtain the two isomorphisms. □

Proposition E.5. *Under these isomorphisms*

$$\Lambda_1 := \iota_2 W_1^\infty \iota_1^{-1} := \begin{bmatrix} 0 & 2(x+1) \\ 1 & 0 \end{bmatrix} = \Lambda_1^\top = \iota_1 (W_1^\infty)^\top \iota_2^{-1} \tag{E.41}$$

Proof. On one hand $\iota_2 W_1^\infty \iota_1^{-1} := \begin{bmatrix} 0 & 2(x+1) \\ 1 & 0 \end{bmatrix}$ follows from eq. (E.36) and eq. (E.37); on the other $\Lambda_1 = \Lambda_1^\top$ follows from the following chain of equalities:

$$\begin{aligned}
\langle \Lambda_1 \begin{bmatrix} W_n \\ U_m \end{bmatrix}, \begin{bmatrix} W_{n'} \\ U_{m'} \end{bmatrix} \rangle_{\mathbb{H}} &= \langle \begin{bmatrix} 2(x+1)U_m \\ W_n \end{bmatrix}, \begin{bmatrix} W_{n'} \\ U_{m'} \end{bmatrix} \rangle_{\mathbb{H}} \\
&= \langle \begin{bmatrix} W_{m+1} + W_m \\ U_n + U_{n-1} \end{bmatrix}, \begin{bmatrix} W_{n'} \\ U_{m'} \end{bmatrix} \rangle_{\mathbb{H}} \\
&= \langle W_{m+1} + W_m, W_{n'} \rangle_{\mu_W} \langle U_n + U_{n-1}, U_{m'} \rangle_{\mu_U} \\
&= (\delta_{m+1,n'} + \delta_{m,n'})(\delta_{n,m'} + \delta_{n-1,m'}) \\
&= (\delta_{m,n'-1} + \delta_{m,n'})(\delta_{n,m'} + \delta_{n,m'+1}) \\
&= \langle U_m, U_{n'} + U_{n'-1} \rangle_{\mu_U} \langle W_n, W_{m'+1} + W_{m'} \rangle_{\mu_W} \\
&= \langle W_n, 2(x+1)U_{m'} \rangle_{\mu_W} \langle U_m, W_{n'} \rangle_{\mu_U} \\
&= \langle \begin{bmatrix} W_n \\ U_m \end{bmatrix}, \Lambda_1 \begin{bmatrix} W_{n'} \\ U_{m'} \end{bmatrix} \rangle_{\mathbb{H}}
\end{aligned}$$

□

Under these identifications

$$A := \iota_1 W_0^\infty = \begin{bmatrix} 1 \\ 0 \end{bmatrix} z^\top, \quad B := W_2^\infty \iota_2^{-1} = \begin{bmatrix} \sum_j \tilde{z}_j \\ 0 \end{bmatrix}^\top$$

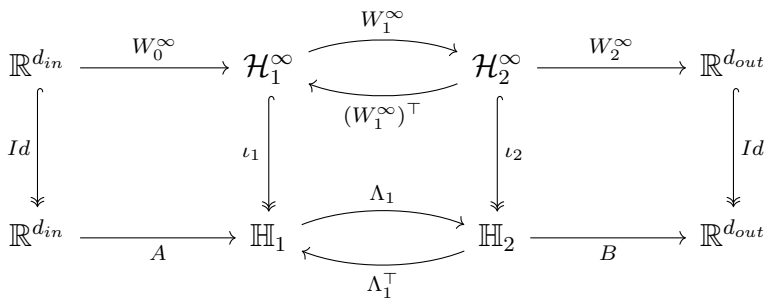


Figure E.3: *The commutative diagram between limiting spaces.*

Note how, as expected, we obtain

$$\begin{aligned}\Phi^{\infty}(\mathbf{x}) :&= W_2^{\infty}W_1^{\infty}W_0^{\infty}\mathbf{x} = B\Lambda_1 A\mathbf{x} \\ &= B\Lambda_1 \begin{bmatrix} \langle z, \mathbf{x}\rangle_{\mathbb{R}^{d_{in}}} \\ 0 \end{bmatrix} = B \begin{bmatrix} 0 \\ \langle z, \mathbf{x}\rangle_{\mathbb{R}^{d_{in}}} \end{bmatrix} \\ &= \langle \sum_j \tilde{z}_j, 0\rangle_{\mu_W\otimes\rho}\langle 0, \langle z, \mathbf{x}\rangle_{\mathbb{R}^{d_{in}}}\rangle_{\mu_U\otimes\rho} = 0\end{aligned}$$

Moreover

$$\Lambda_1\Lambda_1^{\top} = \Lambda_1^{\top}\Lambda_1 = \begin{bmatrix} 2(x+1) & 0 \\ 0 & 2(x+1) \end{bmatrix} = 2(x+1)Id_{\mathbb{H}}$$

An equivalent characterization

With the change of variable $x \mapsto t := \frac{x+1}{2}$ we observe how we can equivalently define the polynomials as

$$P_k(t) = 1, \quad P_k(t) = 4t - 1 \tag{E.42}$$

$$P_{k+1}(t) = (4t-2)P_k(t) - P_{k-1}(t) \tag{E.43}$$

now satisfying

$$\int_0^1 P_n(t)P_m(t)\frac{2}{\pi}\sqrt{\frac{1-t}{t}}dt = \delta_{n,m}$$

thus orthonormal under $d\mu_P(t) = \frac{2}{\pi}\sqrt{\frac{1-t}{t}}dt$ which is the density of a $Beta(\frac{1}{2}, \frac{3}{2})$ distribution. Analogously

$$Q_k(t) = 1, \quad Q_k(t) = 4t - 2 \tag{E.44}$$

$$Q_{k+1}(t) = (4t-2)Q_k(t) - Q_{k-1}(t) \tag{E.45}$$

now satisfying

$$\int_0^1 Q_n(t)Q_m(t)\frac{2}{\pi}\sqrt{t(1-t)}dt = \delta_{n,m}$$

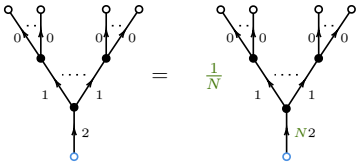
thus orthonormal under $d\mu_Q(t) = \frac{2}{\pi}\sqrt{t(1-t)}dt$ which is the density of a $Beta(\frac{3}{2}, \frac{3}{2})$.

Finally in this case

$$\Lambda_1 = \begin{bmatrix} 0 & 4t \\ 1 & 0 \end{bmatrix}.$$

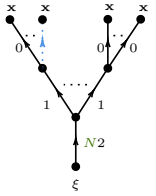
E.2 The Non-Linear Case

Let us try to consider the non-linear case, as to observe the limitations of our approach. The operator graphs now appearing in the graphical expansion of Φ_2 are trees of form

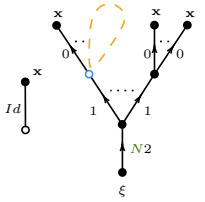


Training dynamics follow a completely analogous pattern and we easily see how

$$W_0^1 = W_0^0 + \frac{\lambda_0}{N} \sum_{\tau \in \mathbb{T}} \frac{\varphi_\tau}{s(\tau)} \sum_{\overleftarrow{0}} \int \quad \mathrm{d}\rho_0(\mathbf{x}, y)$$

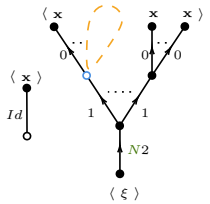


$$= W_0^0 + \frac{\lambda_0}{N} \sum_{\tilde{\tau} \in \mathrm{d}_0\mathbb{T}} \frac{\dot{\varphi}_{\tilde{\tau}}}{s(\tilde{\tau})} \int \quad \mathrm{d}\rho_0(\mathbf{x}, y)$$

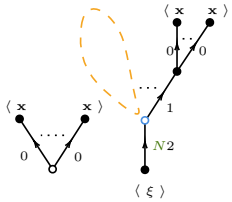


where in the first equality the fist sum is over the trees defining the graph expansion of the network, the second over 0-labelled edges, and where in the second equality we highlight the sub-graph excised for the equivalent representation. Using the braket notation to indicate integration, proceeding similarly for W_1^1 and W_2^1, we get to

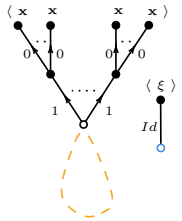
$$W_0^1 = W_0^0 + \frac{\lambda_0}{N} \sum_{\tilde{\tau} \in \mathrm{d}_0\mathbb{T}} \frac{\dot{\varphi}_{\tilde{\tau}}}{s(\tilde{\tau})}$$



$$W_1^1 = W_1^0 + \frac{\lambda_1}{N} \sum_{\tilde{\tau} \in \mathrm{d}_1\mathbb{T}} \frac{\dot{\varphi}_{\tilde{\tau}}}{s(\tilde{\tau})}$$

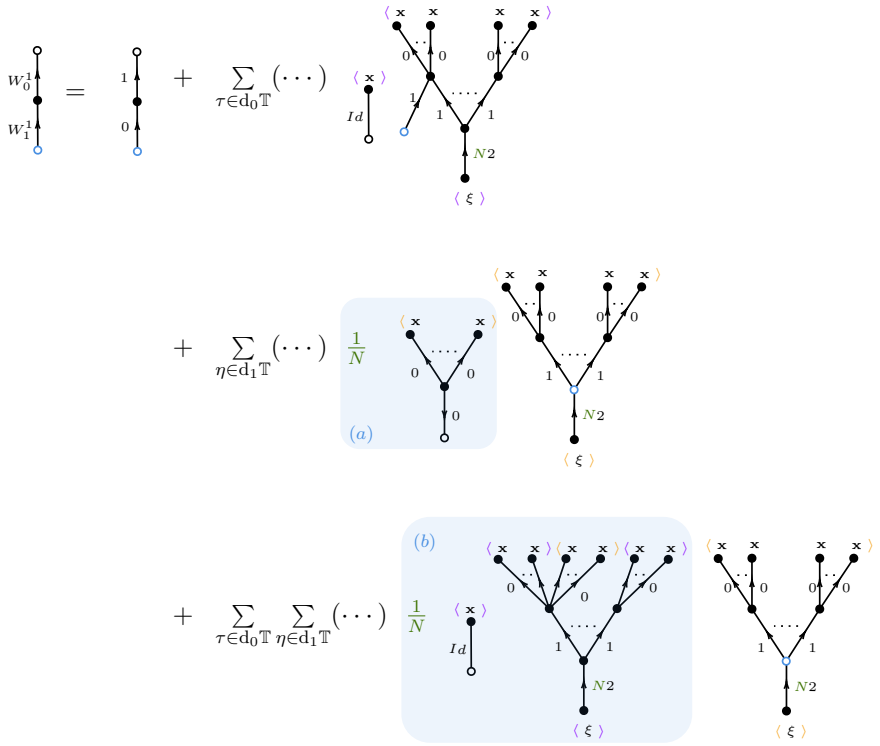


$$W_2^1 = W_2^0 + \lambda_2 \sum_{\tilde{\tau} \in \mathrm{d}_2\mathbb{T}} \frac{\dot{\varphi}_{\tilde{\tau}}}{s(\tilde{\tau})}$$



Note that in these representations the bracketed objects, even if disjoint, are integrated jointly under the same measure.

Consider what happens when computing $W_1^1\varphi(W_0^1 x)$: in each tree we have to substitute the new matrices $W_\star^1$ with sums over trees. For example in te case of the simple term $W_1^1 W_0^1$, under μP where $(\frac{\lambda_0}{N}, \frac{\lambda_1}{N}, \lambda_2) = (1, \frac{1}{N}, \frac{1}{N})$ we obtain [3]



Since the variance of (a) and (b) is of order $\frac{1}{N}$ these objects *do converge* in L^2 to their expectations, mimicking what happens in the linear case. However now expectations depend on the graphs' topology and might not vanish, as in the case of (a).

One sees that our theory is not confined to the linear regime: it is, at least in principle, possible to track gradient updates and obtain limiting dynamics for non-linear networks as well. Analogues of the Js and Ks could be introduced to form an orthonormal basis, realized as graphs with a recursive tree-like structure that do not fold onto themselves. The challenge, however, lies in the combinatorial complexity inherent in the non-linear case, in these recursive compositions of arbitrarily wide trees. To move forward, we therefore posit the necessity of an *algebraic description*, one which we do not yet possess. The purpose of this appendix has been precisely to lay some preliminary groundwork: by recasting the linear case in our language, we aim to show that the extension to non-linear dynamics, while daunting, remains approachable in principle. It is our hope that future work will build upon these observations to fully develop the non-linear case.

[3] In the last term of the sum we colored brackets in two different ways to keep track of terms integrated jointly, *c.f.* previous remark.

Bibliography

[1] Erdinc Akyildirim, Matteo Gambara, Josef Teichmann, and Syang Zhou. Applications of Signature Methods to Market Anomaly Detection, 2022. URL https://arxiv.org/abs/2201.02441.

[2] Simran Arora, Sabri Eyuboglu, Aman Timalsina, Isys Johnson, Michael Poli, James Zou, Atri Rudra, and Christopher Re. Zoology: Measuring and improving recall in efficient Language Models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=LY3ukUANko.

[3] Lucas Benigni and Sandrine Péché. Eigenvalue distribution of some nonlinear models of random matrices. *Electron. J. Probab.*, 26:Paper No. 150, 37, 2021. ISSN 1083-6489. doi: 10.1214/21-ejp699. URL https://doi.org/10.1214/21-ejp699.

[4] Alain Berlinet and Christine Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer US, 2011. ISBN 9781441990969. URL https://books.google.co.uk/books?id=bX3TBwAAQBAJ.

[5] Jacob Biamonte and Ville Bergholm. Tensor Networks in a nutshell, 2017. URL https://arxiv.org/abs/1708.00006.

[6] Henry Roy Brahana. Systems of circuits on two-dimensional manifolds. *Annals of Mathematics*, 23:144, 1921. URL https://api.semanticscholar.org/CorpusID:16854121.

[7] Edouard Brézin, Claude Itzykson, Giorgio Parisi, and Jean-Bernard Zuber. Planar diagrams. *Communications in Mathematical Physics*, 59(1):35 – 51, 1978.

[8] Thomas Cass and Cristopher Salvi. Lecture notes on Rough Paths and applications to machine learning, 2024. URL https://arxiv.org/abs/2404.06583.

[9] Thomas Cass, Terry Lyons, and Xingcheng Xu. Weighted signature kernels. *The Annals of Applied Probability*, 34(1A):585 – 626, 2024. doi: 10.1214/23-AAP1973. URL https://doi.org/10.1214/23-AAP1973.

[10] Kuo-Tsai Chen. Integration of paths–a faithful representation of paths by non-commutative formal power series. *Transactions of the American Mathematical Society*, 89(2):395–407, 1958. ISSN 00029947. URL http://www.jstor.org/stable/1993193.

[11] Lihu Chen and Gaël Varoquaux. What is the role of Small Models in the LLM Era: A survey, 2025. URL https://arxiv.org/abs/2409.06857.

[12] Ilya Chevyrev. Rough path theory. In Richard Szabo and Martin Bojowald, editors, *Encyclopedia of Mathematical Physics (Second Edition)*, pages 1–15. Academic Press, Oxford, second edition edition, 2025. ISBN 978-0-323-95706-9. doi: https://doi.org/10.1016/B978-0-323-95703-8.00027-6. URL https://www.sciencedirect.com/science/article/pii/B9780323957038000276.

[13] Lénaïc Chizat, Maria Colombo, Xavier Fernández-Real, and Alessio Figalli. Infinite-width limit of deep linear neural networks. *Communications on Pure and Applied Mathematics*, pages 1–50, 2024. doi: https://doi.org/10.1002/cpa.22200. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.22200.

[14] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1179. URL https://aclanthology.org/D14-1179/.

[15] Andrzej Cichocki, Namgil Lee, Ivan Oseledets, Anh-Huy Phan, Qibin Zhao, and Danilo P. Mandic. Tensor Networks for dimensionality reduction and large-scale optimization: Part 1 low-rank tensor decompositions. *Foundations and Trends in Machine Learning*, 9(4–5):249–429, 2016. ISSN 1935-8245. doi: 10.1561/2200000059. URL http://dx.doi.org/10.1561/2200000059.

[16] Nicola Muca Cirone, Jad Hamdan, and Cristopher Salvi. Genus expansion for non-linear random matrix ensembles with applications to neural networks, 2025. URL https://arxiv.org/abs/2407.08459.

[17] Thomas Cochrane, Peter Foster, Varun Chhabra, Maud Lemercier, Terry Lyons, and Cristopher Salvi. 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)*, pages 35–40. IEEE, 2021.

[18] Benoît Collins and Tomohiro Hayase. Asymptotic freeness of layerwise Jacobians caused by invariance of multilayer perceptron: the Haar orthogonal case. *Comm. Math. Phys.*, 397(1):85–109, 2023. ISSN 0010-3616,1432-0916. doi: 10.1007/s00220-022-04441-7. URL https://doi.org/10.1007/s00220-022-04441-7.

[19] Gregory M. Constantine and Thomas H. Savits. A multivariate Faa di Bruno formula with applications. *Transactions of the American Mathematical Society*, 348: 503–520, 1996. URL https://api.semanticscholar.org/CorpusID:18986788.

[20] Christa Cuchiero, Lukas Gonon, Lyudmila Grigoryeva, Juan-Pablo Ortega, and Josef Teichmann. Expressive power of Randomized Signature. In *The Symbiosis of Deep Learning and Differential Equations*, 2021. URL https://openreview.net/forum?id=KWWFPULvmVw.

[21] Tri Dao and Albert Gu. Transformers are SSMs: generalized models and efficient algorithms through structured state space duality. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

[22] Sanjoy Dasgupta and Anupam Gupta. An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Structures & Algorithms*, 22, 2003. URL https://api.semanticscholar.org/CorpusID:10327785.

[23] Soham De, Samuel L. Smith, Anushan Fernando, Aleksandar Botev, George Cristian-Muraru, Albert Gu, Ruba Haroun, Leonard Berrada, Yutian Chen, Srivatsan Srinivasan, Guillaume Desjardins, Arnaud Doucet, David Budden, Yee Whye Teh, Razvan Pascanu, Nando De Freitas, and Caglar Gulcehre. Griffin: Mixing gated linear recurrences with local Attention for efficient language models, 2024. URL https://arxiv.org/abs/2402.19427.

[24] Alexander G. de G. Matthews, Jiri Hron, Mark Rowland, Richard E. Turner, and Zoubin Ghahramani. Gaussian process behaviour in wide deep neural networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=H1-nGgWC-.

[25] Persi Diaconis and Mehrdad Shahshahani. On the eigenvalues of random matrices. *Journal of Applied Probability*, 31:49–62, 1994. ISSN 00219002. URL http://www.jstor.org/stable/3214948.

[26] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth

16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

[27] Simon S. Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=S1eK3i09YQ.

[28] Guillaume Dubach and Yuval Peled. On words of non-Hermitian random matrices. *Ann. Probab.*, 49(4):1886–1916, 2021. ISSN 0091-1798,2168-894X. doi: 10.1214/20-aop1496. URL https://doi.org/10.1214/20-aop1496.

[29] Zhou Fan and Zhichao Wang. Spectra of the Conjugate Kernel and Neural Tangent Kernel for linear-width neural networks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 7710–7721. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/572201a4497b0b9f02d4f279b09ec30d-Paper.pdf.

[30] Leo Feng, Frederick Tung, Mohamed Osama Ahmed, Yoshua Bengio, and Hossein Hajimirsadeghi. Were RNNs all we needed?, 2024. URL https://arxiv.org/abs/2410.01201.

[31] Adeline Fermanian. Embedding and learning with signatures. *Comput. Statist. Data Anal.*, 157:Paper No. 107148, 23, 2021. ISSN 0167-9473,1872-7352. doi: 10.1016/j.csda.2020.107148. URL https://doi.org/10.1016/j.csda.2020.107148.

[32] Adeline Fermanian, Pierre Marion, Jean-Philippe Vert, and Gérard Biau. Framing rnn as a kernel method: A neural ode approach. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 3121–3134. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/18a9042b3fc5b02fe3d57fea87d6992f-Paper.pdf.

[33] Adeline Fermanian, Terry Lyons, James Morrill, and Cristopher Salvi. New directions in the applications of rough path theory. *IEEE BITS the Information Theory Magazine*, 3(2):41–53, 2023. doi: 10.1109/MBITS.2023.3243885.

[34] Richard Phillips Feynman. Space-Time approach to Non-Relativistic Quantum Mechanics. *Reviews of Modern Physics*, 20:367–387, 1948. URL https://api.semanticscholar.org/CorpusID:268744472.

[35] Loic Foissy. An introduction to Hopf algebras of trees. URL https://www2.mathematik.hu-berlin.de/.

[36] Peter K. Friz and Martin Hairer. *A course on rough paths*. Universitext. Springer, Cham, second edition, 2020. ISBN 978-3-030-41556-3; 978-3-030-41555-6. doi: 10.1007/978-3-030-41556-3. URL https://doi.org/10.1007/978-3-030-41556-3. With an introduction to regularity structures.

[37] Peter K. Friz and Nicolas B. Victoir. *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.

[38] Daniel Y. Fu, Tri Dao, Khaled Kamal Saab, Armin W. Thomas, Atri Rudra, and Christopher Re. Hungry Hungry Hippos: Towards language modeling with State Space Models. In *The Eleventh International Conference on Learning Representations*, 2022.

[39] Adrià Garriga-Alonso, Carl Edward Rasmussen, and Laurence Aitchison. Deep convolutional networks as shallow Gaussian processes. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bklfsi0cKm.

[40] Stuart Geman. 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.

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

[42] Karan Goel, Albert Gu, Chris Donahue, and Christopher Ré. It's raw! audio generation with state-space models. *International Conference on Machine Learning*, 2022.

[43] Jonathan L. Gross and Thomas W. Tucker. *Topological graph theory*. Wiley-Interscience Series in Discrete Mathematics and Optimization. John Wiley & Sons Inc., New York, 1987. ISBN 0-471-04926-3.

[44] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=tEYskw1VY2.

[45] Albert Gu, Tri Dao, Stefano Ermon, Atri Rudra, and Christopher Ré. HiPPO: Recurrent Memory with Optimal Polynomial Projections. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1474–1487. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/102f0bb6efb3a6128a3c750dd16729be-Paper.pdf.

[46] Albert Gu, Karan Goel, and Christopher Re. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*, 2021.

[47] Massimiliano Gubinelli. Ramification of rough paths. *J. Differential Equations*, 248(4):693–721, 2010. ISSN 0022-0396,1090-2732. doi: 10.1016/j.jde.2009.11.015. URL https://doi.org/10.1016/j.jde.2009.11.015.

[48] Ernst Hairer, Syvert P. Nørsett, and Gerhard Wanner. *Solving ordinary differential equations I (2nd revised. ed.): nonstiff problems*. Springer-Verlag, Berlin, Heidelberg, 1993. ISBN 0387566708.

[49] Ben Hambly and Terry Lyons. Uniqueness for the signature of a path of bounded variation and the reduced path group. *Annals of Mathematics*, pages 109–167, 2010.

[50] Boris Hanin. Random neural networks in the infinite width limit as Gaussian processes. *The Annals of Applied Probability*, 33(6A):4798 – 4819, 2023. doi: 10.1214/23-AAP1933. URL https://doi.org/10.1214/23-AAP1933.

[51] John Harer and Don Zagier. The Euler characteristic of the moduli space of curves. *Inventiones mathematicae*, 85:457–485, 1986.

[52] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735.

[53] Melker Hoglund, Emilio Ferrucci, Camilo Hernandez, Aitor Muguruza Gonzalez, Cristopher Salvi, Leandro Sanchez-Betancourt, and Yufei Zhang. A Neural RDE approach for continuous-time non-Markovian stochastic control problems, 2023. URL https://arxiv.org/abs/2306.14258.

[54] Gerard't Hooft. A planar diagram theory for strong interactions. *Nuclear Physics*, 72:461–473, 1974. URL https://api.semanticscholar.org/CorpusID:122597626.

[55] Wei Huang, Weitao Du, and Richard Yi Da Xu. On the Neural Tangent Kernel of deep networks with orthogonal initialization. In Zhi-Hua Zhou, editor,

Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21, pages 2577–2583. International Joint Conferences on Artificial Intelligence Organization, 8 2021. doi: 10.24963/ijcai.2021/355. URL https://doi.org/10.24963/ijcai.2021/355. Main Track.

[56] Yenkun Huang and Che-Yuan Tsai. 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.

[57] Zacharia Issa, Blanka Horvath, Maud Lemercier, and Cristopher Salvi. Non-adversarial training of Neural SDEs with signature kernel scores. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 11102–11126. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/2460396f2d0d421885997dd1612ac56b-Paper-Conference.pdf.

[58] Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural Tangent Kernel: Convergence and generalization in neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/5a4be1fa34e62bb8a6ec6b91d2462f5a-Paper.pdf.

[59] Kurt Johansson. 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.

[60] William B. Johnson and Joram Lindenstrauss. Extensions of Lipschitz mappings into hilbert space. *Contemporary mathematics*, 26:189–206, 1984. URL https://api.semanticscholar.org/CorpusID:117819162.

[61] Samuel G.G. Johnston and Joscha Prochno. Faà di Bruno's formula and inversion of power series. *Advances in Mathematics*, 395:108080, 2022. ISSN 0001-8708. doi: https://doi.org/10.1016/j.aim.2021.108080. URL https://www.sciencedirect.com/science/article/pii/S0001870821005193.

[62] Iain M. Johnstone. 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.

[63] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast autoregressive Transformers with linear Atten-

tion. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5156–5165. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/katharopoulos20a.html.

[64] Tobias Katsch. Gateloop: Fully data-controlled linear recurrence for sequence modeling, 2024. URL https://arxiv.org/abs/2311.01927.

[65] Patrick Kidger. On Neural Differential Equations, 2022. URL https://arxiv.org/abs/2202.02435.

[66] Patrick Kidger, Patric Bonnier, Imanol Perez Arribas, Cristopher Salvi, and Terry Lyons. Deep Signature Transforms. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/d2cdf047a6674cef251d56544a3cf029-Paper.pdf.

[67] Patrick Kidger, James Morrill, James Foster, and Terry Lyons. Neural controlled differential equations for irregular time series. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6696–6707. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/4a5876b450b45371f6cfe5047ac8cd45-Paper.pdf.

[68] Jaehoon Lee, Jascha Sohl-dickstein, Jeffrey Pennington, Roman Novak, Sam Schoenholz, and Yasaman Bahri. Deep neural networks as Gaussian processes. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=B1EA-M-0Z.

[69] Maud Lemercier, Cristopher Salvi, Thomas Cass, Edwin V. Bonilla, Theodoros Damoulas, and Terry Lyons. Siggpde: Scaling sparse Gaussian processes on sequential data. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 6233–6242. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/lemercier21a.html.

[70] Maud Lemercier, Cristopher Salvi, Theodoros Damoulas, Edwin Bonilla, and Terry Lyons. Distribution regression for sequential data. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 3754–3762. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/lemercier21a.html.

[71] Yuhong Li, Tianle Cai, Yi Zhang, Deming Chen, and Debadeepta Dey. What makes convolutional models great on long sequence modeling? In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=TGJSPbRpJX-.

[72] Zhong Li, Jiequn Han, Weinan E, and Qianxiao Li. Approximation and optimization theory for linear continuous-time recurrent neural networks. *Journal of Machine Learning Research*, 23(42):1–85, 2022. URL http://jmlr.org/papers/v23/21-0368.html.

[73] Chris Lu, Yannick Schroecker, Albert Gu, Emilio Parisotto, Jakob Foerster, Satinder Singh, and Feryal Behbahani. Structured state space models for in-context reinforcement learning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 47016–47031. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/92d3d2a9801211ca3693ccb2faa1316f-Paper-Conference.pdf.

[74] Terry Lyons. On the non-existence of path integrals. *Proceedings: Mathematical and Physical Sciences*, 432(1885):281–290, 1991. ISSN 09628444. URL http://www.jstor.org/stable/51856.

[75] Terry Lyons. Differential equations driven by rough signals (i): An extension of an inequality of L.C. young. *Mathematical Research Letters*, 1(4):451–464, 1994.

[76] Terry Lyons and Zhongmin Qian. *System control and rough paths*. Oxford University Press, 2002.

[77] Terry Lyons, Michael Caruana, and Thierry Lévy. *Differential equations driven by rough paths*. Springer, 2007.

[78] Satya N. Majumdar and Massimo Vergassola. 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.

[79] Camille Male. Traffic distributions and independence: permutation invariant random matrices and the three notions of independence. *Mem. Amer. Math. Soc.*, 267(1300):v+88, 2020. ISSN 0065-9266,1947-6221. doi: 10.1090/memo/1300. URL https://doi.org/10.1090/memo/1300.

[80] Georg Manten, Cecilia Casolo, Emilio Ferrucci, Søren Wengel Mogensen, Cristopher Salvi, and Niki Kilbertus. Signature kernel conditional independence tests in

causal discovery for stochastic processes. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Nx4PMtJ1ER.

[81] Eric Martin and Chris Cundy. Parallelizing linear recurrent neural nets over sequence length. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=HyUNwulC-.

[82] John C. Mason and G. H. Elliott. Near-minimax complex approximation by four kinds of Chebyshev polynomial expansion. volume 46, pages 291–300. 1993. doi: 10.1016/0377-0427(93)90303-S. URL https://doi.org/10.1016/0377-0427(93)90303-S. Computational complex analysis.

[83] Robert I McLachlan, Klas Modin, Hans Munthe-Kaas, and Olivier Verdier. Butcher series: A story of rooted trees and numerical methods for evolution equations, 2017. URL https://arxiv.org/abs/1512.00906.

[84] William Merrill, Jackson Petty, and Ashish Sabharwal. The illusion of state in state-space models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

[85] James A. Mingo and Roland Speicher. *Free probability and random matrices*, volume 35 of *Fields Institute Monographs*. Springer, New York; Fields Institute for Research in Mathematical Sciences, Toronto, ON, 2017. ISBN 978-1-4939-6941-8; 978-1-4939-6942-5. doi: 10.1007/978-1-4939-6942-5. URL https://doi.org/10.1007/978-1-4939-6942-5.

[86] James Morrill, Patrick Kidger, Lingyi Yang, and Terry Lyons. Neural controlled differential equations for online prediction tasks, 2021. URL https://arxiv.org/abs/2106.11028.

[87] Sajad Movahedi, Felix Sarnthein, Nicola Muca Cirone, and Antonio Orvieto. Fixed-point RNNs: Interpolating from diagonal to dense, 2025. URL https://arxiv.org/abs/2503.10799.

[88] Nicola Muca Cirone, Maud Lemercier, and Cristopher Salvi. Neural signature kernels as infinite-width-depth-limits of controlled ResNets. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 25358–25425. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/muca-cirone23a.html.

[89] Nicola Muca Cirone, Antonio Orvieto, Benjamin Walker, Cristopher Salvi, and Terry Lyons. Theoretical foundations of deep selective state-space models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 127226–127272. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/e6231c5f46598cfd09ff1970524e0436-Paper-Conference.pdf.

[90] Hans Z. Munthe-Kaas and Kristoffer K. Fø llesdal. Lie-Butcher series, geometry, algebra and computation. In *Discrete mechanics, geometric integration and Lie-Butcher series*, volume 267 of *Springer Proc. Math. Stat.*, pages 71–113. Springer, Cham, 2018. ISBN 978-3-030-01397-4; 978-3-030-01396-7. doi: 10.1007/978-3-030-01397-4_3. URL https://doi.org/10.1007/978-3-030-01397-4_3.

[91] Radford M. Neal. *Bayesian Learning for Neural Networks*. Springer-Verlag, Berlin, Heidelberg, 1996. ISBN 0387947248.

[92] Eric Nguyen, Karan Goel, Albert Gu, Gordon W. Downs, Preey Shah, Tri Dao, Stephen A. Baccus, and Christopher Ré. S4ND: modeling images and videos as multidimensional signals using state spaces. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

[93] Alexandru Nica and Roland Speicher. *Lectures on the Combinatorics of Free Probability*. London Mathematical Society Lecture Note Series. Cambridge University Press, 2006.

[94] Roman Novak, Lechao Xiao, Jaehoon Lee, Yasaman Bahri, Greg Yang, Jiri Hron, Daniel A. Abolafia, Jeffrey Pennington, and Jascha Sohl-Dickstein. Bayesian deep convolutional networks with many channels are Gaussian processes, 2020. URL https://arxiv.org/abs/1810.05148.

[95] Antonio Orvieto, Samuel L Smith, Albert Gu, Anushan Fernando, Caglar Gulcehre, Razvan Pascanu, and Soham De. Resurrecting recurrent neural networks for long sequences. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 26670–26698. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/orvieto23a.html.

[96] Alexandre Pannier and Cristopher Salvi. A path-dependent PDE solver based on signature kernels, 2024. URL https://arxiv.org/abs/2403.11738.

[97] Leonid Pastur. On random matrices arising in deep neural networks. Gaussian case, 2020. URL https://arxiv.org/abs/2001.06188.

[98] Leonid Pastur and Victor Slavin. On random matrices arising in deep neural networks: General i.i.d. case. *Random Matrices: Theory and Applications*, 12 (01):2250046, 2023. doi: 10.1142/S2010326322500460. URL https://doi.org/10.1142/S2010326322500460.

[99] Bo Peng, Eric Alcaide, Quentin Gregory Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, Huanqi Cao, Xin Cheng, Michael Nguyen Chung, Leon Derczynski, Xingjian Du, Matteo Grella, Kranthi Kiran GV, Xuzheng He, Haowen Hou, Przemyslaw Kazienko, Jan Kocon, Jiaming Kong, Bartłomiej Koptyra, Hayden Lau, Jiaju Lin, Krishna Sri Ipsit Mantri, Ferdinand Mom, Atsushi Saito, Guangyu Song, Xiangru Tang, Johan S. Wind, Stanisław Woźniak, Zhenyuan Zhang, Qinghua Zhou, Jian Zhu, and Rui-Jie Zhu. RWKV: Reinventing RNNs for the transformer era. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://openreview.net/forum?id=7SaXczaBpG.

[100] Jeffrey Pennington and Pratik Worah. Nonlinear random matrix theory for deep learning. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/0f3d014eead934bbdbacb62a01dc4831-Paper.pdf.

[101] Jeffrey Pennington and Pratik Worah. The spectrum of the Fisher information matrix of a single-hidden-layer neural network. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/18bb68e2b38e4a8ce7cf4f6b2625768c-Paper.pdf.

[102] Jeffrey Pennington, Samuel S. Schoenholz, and Surya Ganguli. Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4788–4798, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

[103] Jeffrey Pennington, Samuel Schoenholz, and Surya Ganguli. The emergence of spectral universality in deep networks. In Amos Storkey and Fernando Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference*

on Artificial Intelligence and Statistics, volume 84 of *Proceedings of Machine Learning Research*, pages 1924–1932. PMLR, 09–11 Apr 2018. URL https://proceedings.mlr.press/v84/pennington18a.html.

[104] Roger Penrose. Applications of negative dimensional tensors. *Combinatorial Mathematics and its Applications*, 1971.

[105] Zhen Qin, Songlin Yang, Weixuan Sun, Xuyang Shen, Dong Li, Weigao Sun, and Yiran Zhong. HGRN2: Gated linear RNNs with state expansion. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=y6SqbJfCSk.

[106] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007. URL https://proceedings.neurips.cc/paper_files/paper/2007/file/013a006f03dbc5392effeb8f18fda755-Paper.pdf.

[107] C. Emily I. Redelmeier. Genus expansion for real Wishart matrices. *J. Theoret. Probab.*, 24(4):1044–1062, 2011. ISSN 0894-9840,1572-9230. doi: 10.1007/s10959-010-0278-7. URL https://doi.org/10.1007/s10959-010-0278-7.

[108] C. Emily I. Redelmeier. Real second-order freeness and the asymptotic real second-order freeness of several real matrix models. *Int. Math. Res. Not. IMRN*, (12):3353–3395, 2014. ISSN 1073-7928,1687-0247. doi: 10.1093/imrn/rnt043. URL https://doi.org/10.1093/imrn/rnt043.

[109] Daniel Revuz and Marc Yor. *Continuous Martingales and Brownian Motion*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2013. ISBN 9783662217269.

[110] Desmarais Robert N. and Bland Samuel R. Tables of properties of airfoil polynomials. Technical report, 1995. URL https://ntrs.nasa.gov/citations/19960001864.

[111] Thiziri Nait Saada, Alireza Naderi, and Jared Tanner. Beyond IID weights: sparse and low-rank deep neural networks are also Gaussian processes. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=1Wi0Ys33Nm.

[112] Cristopher Salvi, Thomas Cass, James Foster, Terry Lyons, and Weixin Yang. The signature kernel is the solution of a Goursat pde. *SIAM Journal on Mathematics of Data Science*, 3(3):873–899, 2021. doi: 10.1137/20M1366794. URL https://doi.org/10.1137/20M1366794.

[113] Cristopher Salvi, Maud Lemercier, Chong Liu, Blanka Horvath, Theodoros Damoulas, and Terry Lyons. Higher order kernel mean embeddings to capture filtrations of stochastic processes. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

[114] Cristopher Salvi, Maud Lemercier, and Andris Gerasimovičs. Neural stochastic PDEs: resolution-invariant learning of continuous spatiotemporal dynamics. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

[115] Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks, 2014. URL https://arxiv.org/abs/1312.6120.

[116] Raghavendra Selvan, Julian Schön, and Erik B. Dam. Operating critical machine learning models in resource constrained regimes. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 325–335. Springer, 2023.

[117] Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: A law of large numbers. *SIAM J. Appl. Math.*, 80(2):725–752, January 2020. ISSN 0036-1399. doi: 10.1137/18M1192184. URL https://doi.org/10.1137/18M1192184.

[118] Jimmy T.H. Smith, Andrew Warrington, and Scott Linderman. Simplified state space layers for sequence modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Ai8Hw3AXqks.

[119] Yutao Sun, Li Dong, Shaohan Huang, Shuming Ma, Yuqing Xia, Jilong Xue, Jianyong Wang, and Furu Wei. Retentive Network: A successor to Transformer for Large Language Models, 2023. URL https://arxiv.org/abs/2307.08621.

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

[121] Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. Long range arena : A benchmark for efficient transformers. In *International Conference*

on Learning Representations, 2021. URL https://openreview.net/forum?id=qVyeW-grC2k.

[122] Neil C. Thompson, Kristjan Greenewald, Keeheon Lee, and Gabriel F. Manso. The computational limits of deep learning, 2022. URL https://arxiv.org/abs/2007.05558.

[123] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

[124] Benjamin Walker, Andrew D. McLeod, Tiexin Qin, Yichuan Cheng, Haoliang Li, and Terry Lyons. Log neural controlled differential equations: the Lie brackets make a difference. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

[125] Benjamin Walker, Lingyi Yang, Nicola Muca Cirone, Cristopher Salvi, and Terry Lyons. Structured linear CDEs: Maximally expressive and parallel-in-time sequence models, 2025. URL https://arxiv.org/abs/2505.17761.

[126] Zhongwei Wan, Xin Wang, Che Liu, Samiul Alam, Yu Zheng, Jiachen Liu, Zhongnan Qu, Shen Yan, Yi Zhu, Quanlu Zhang, Mosharaf Chowdhury, and Mi Zhang. Efficient large language models: A survey. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=bsCCJHbO8A. Survey Certification.

[127] Junxiong Wang, Jing Nathan Yan, Albert Gu, and Alexander Rush. Pretraining without Attention. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 58–69, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.5. URL https://aclanthology.org/2023.findings-emnlp.5/.

[128] Zhichao Wang and Yizhe Zhu. Deformed semicircle law and concentration of nonlinear random matrices for ultra-wide neural networks. *Ann. Appl. Probab.*, 34(2):1896–1947, 2024. ISSN 1050-5164,2168-8737. doi: 10.1214/23-aap2010. URL https://doi.org/10.1214/23-aap2010.

[129] Greg Yang. Wide feedforward or recurrent neural networks of any architecture are Gaussian processes. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neu-*

ral Information Processing Systems, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/5e69fda38cda2060819766569fd93aa5-Paper.pdf.

[130] Greg Yang. Scaling limits of wide neural networks with weight sharing: Gaussian process behavior, gradient independence, and Neural Tangent Kernel derivation, 2020. URL https://arxiv.org/abs/1902.04760.

[131] Greg Yang. Tensor programs II: Neural Tangent Kernel for any architecture, 2020. URL https://arxiv.org/abs/2006.14548.

[132] Greg Yang and Edward J. Hu. Tensor programs IV: Feature learning in infinite-width neural networks. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 11727–11737. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/yang21c.html.

[133] Songlin Yang, Bailin Wang, Yikang Shen, Rameswar Panda, and Yoon Kim. Gated linear Attention transformers with hardware-efficient training. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

[134] Songlin Yang, Bailin Wang, Yu Zhang, Yikang Shen, and Yoon Kim. Parallelizing linear transformers with the delta rule over sequence length. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

[135] Laurence C. Young. An inequality of the Hölder type, connected with Stieltjes integration. *Acta Mathematica*, 67(none):251 – 282, 1936. doi: 10.1007/BF02401743. URL https://doi.org/10.1007/BF02401743.

[136] Nicolas Zucchet, Robert Meier, Simon Schug, Asier Mujika, and João Sacramento. Online learning of long-range dependencies. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

[137] Alexander Zvonkin. Matrix integrals and map enumeration: An accessible introduction. *Mathematical and Computer Modelling*, 26(8):281–304, 1997. ISSN 0895-7177. doi: https://doi.org/10.1016/S0895-7177(97)00210-0. URL https://www.sciencedirect.com/science/article/pii/S0895717797002100.