Universality of Signatures in Rough Path Spaces

A Kernel-Theoretic Approach to Local and Global Approximations

Tomás Carrondo

TU Delft

Universality of Signatures in Rough Path Spaces

A Kernel-Theoretic Approach to Local and Global Approximations

by

Tomás Carrondo

Student Number
5832578

Supervisor:	Prof. Fenghui Yu
Co-supervisor:	Prof. Christa Cuchiero
Project Duration:	October 2023 - June 2024
Department:	Delft Institute of Applied Mathematics

Style: TU Delft Report, with modifications by Tomás Carrondo

Abstract

This thesis examines the approximation capabilities of *path signatures* within rough path spaces, focusing on both local and global universality. To this end, we provide a self-contained introduction to Rough Path theory, highlighting the interplay between additive and multiplicative functionals. This leads to the renowned Lyons' Extension theorem and the definition of rough path spaces. We also re-examine the concept of *universality* from a kernel-theoretic perspective, culminating in the classical universal approximation result for signatures over a compact subset of paths. To broaden the scope beyond compact domains, we introduce the framework of weighted spaces and elaborate on the notion of *global universality*. Specifically, we formally define *globally universal kernels* and prove sufficient conditions for their existence. The associated reproducing kernel Hilbert space is shown to approximate a wide range of functions over the entire domain, which may be non-(locally) compact. In particular, we apply these theoretical tools to rough path spaces, thereby cementing the global universality of signatures.

TU Delft

Contents

Preface

This thesis grew out of my dedication to explore the expanding field of *signature-based methods*. From the outset, delving into signature-based methodologies captivated me due to their profound theoretical foundations and potential for applications, particularly in Mathematical Finance. As this work aims to elucidate, theoretically, the signature of a path emerges as a somewhat intricate mathematical object with both analytical and algebraic aspects. However, in practical terms, it suffices to regard path signatures as a potent set of features offering high expressiveness, hence their success in applications.

Yet, even within this practical perspective, questions remain. In what sense exactly are signatures "potent features"? What does "high expressiveness" actually entail? As I continued my study of signature-based methods, I found myself increasingly drawn to these fundamental questions, shifting my focus from the methodologies themselves to the *underlying principles* enabling their efficacy. Furthermore, upon a thorough review of the recent literature on signature-based methods, particularly within Finance, ranging from portfolio optimisation to pricing and hedging, I observed a common thread: these underlying principles appeared to be pervasive across applications. Consequently, the crux of interest and a thesis-worthy direction lay within the theoretical backdrop. And, as is often the case, once theoretical proficiency is attained, applications become a matter of ingenuity.

Now, committed to determining a (theoretical!) thesis-worthy direction, I quickly recognised the necessity of learning about two previously unfamiliar subjects: *Rough Path theory* and *kernel theory*. Rough Path theory, essentially a pathwise theory of integration against highly irregular paths like Brownian motion, offers a re-examination of Itô's Stochastic Calculus from a purely analytical perspective. Consequently, it has recently laid the theoretical groundwork for many endeavours in Robust Finance, particularly within signature-based methods. Kernel theory, a broad theory within Functional Analysis, finds applications in a remarkable number of areas, including Machine Learning, justifying its involvement here. Additionally, it is through kernel-theoretic lenses that we analyse the concept of *universality*, a central notion in this work and a cornerstone of most, if not all, signature-based methods.

Hence, this thesis occupies a position at the intersection of these two theories — a position that I have endeavoured to elucidate clearly and hope to have achieved. Ultimately, this thesis follows a direction towards a deep understanding of the theory underpinning signature-based methods, with particular emphasis on the previously alluded concept of universality. It was this concept of universality that initially prompted me to pursue a more theoretical approach, and personally, I believe my primary contribution lies within its context.

Lastly, before proceeding to the actual thesis, I would like to express my gratitude to Professor Fenghui Yu, my daily supervisor, for introducing me to path signatures during the previous academic year and for her availability. I am indebted to Professor Christa Cuchiero, my co-supervisor, for her invaluable insights and clarifications during our online meetings. Above all, I am grateful for her inspiration, which has profoundly influenced my own work, directly linked to her expertise. I also wish to thank Professor Frank Redig and Professor Francesca Bartolucci for their careful reading of initial drafts pertaining to my contributions. Their feedback and corrections were instrumental in shaping the final outcome of this thesis. Additionally, I express my appreciation to Professor Andrew L. Allan for a brief yet fruitful conversation during his visit to Delft, which provided a valuable eye-opening moment. Each of these individuals has played a role in the development of this thesis, and their support and guidance are deeply appreciated.

Tomás Carrondo
Delft, June 2024

Introduction

This thesis is ultimately concerned with the approximation capabilities of *path signatures*. Given a path $X : [0, T] \to \mathbb{R}^d$, its signature $S(X)$ is formally defined as the infinite collection of all iterated integrals of its components X^i against each other, i.e., $S(X)$ comprises all elements of the form

$$\int_{0<t_1<\cdots<t_n<T} \mathrm{d}X_{t_1}^{i_1} \cdots \mathrm{d}X_{t_n}^{i_n},$$

with $i_1, \ldots, i_n \in \{1, \ldots, d\}$ and $n \geq 1$, forming a mathematically intriguing object with remarkable algebraic properties. Yet, as discussed below, defining these integrals poses difficulties when X is not of bounded variation. From a practical point of view, however, the signature of a path stands as a powerful and descriptive set of features, offering high expressiveness in the space of paths [32]. To elaborate, path signatures essentially encapsulate two key aspects: firstly, under mild conditions, the signature of a path X uniquely identifies its associated path. Secondly, path signatures possess the ability to approximate continuous functions of paths. This second aspect is precisely captured by the concept of *universality*.

Universality, in essence, forms the crux of this thesis and fundamentally concerns the ability to approximate functions defined over a given domain K. In the classical setting, where K is assumed to be compact, a collection of functions from K to $\mathbb{R}$ is commonly said to be *universal* if it is dense in the space of continuous functions with respect to the supremum norm. Notably, under this compactness assumption, universal approximation results are abundant, often deriving from the classical Stone-Weierstrass theorem [27]. One such result, pertains to the use of signatures to approximate functionals of paths.

As elucidated in this work, the signature of a path $S(X)$ yields a set of linear functionals $X \mapsto L(S(X))$, which uniformly approximate any continuous function defined over a fixed compact subset K of the path space under consideration ([48], Theorem 3.1). In other words, the set of linear functionals of the signature is universal over K. As an analogy, just as polynomials are dense in the space of continuous functions $f : [a, b] \to \mathbb{R}$, where $[a, b] \subset \mathbb{R}$, linear functionals of the signature exhibit density in the space of continuous functions $f : K \to \mathbb{R}$, where K is a compact subset of a predetermined path space. In this context, signatures stand as a natural selection for basis functions within the space of paths [9].

Here, it is worth noting that due to their notable properties as features, signatures have achieved numerous empirical successes in various applications, particularly within the domain of Mathematical Finance. Furthermore, while the origins of signatures can be traced back to the work of K. T. Chen [17], in today's theoretical landscape, they fall under Lyons' recently established theory of *rough paths* [54, 37]. Lyons' theory, alongside other contributions, offers a streamlined *pathwise* framework for differential equations driven by a broad spectrum of stochastic processes, or more precisely, rough paths. Consequently, path signatures within the framework of rough paths furnish us with versatile and exceptionally useful tools for robust and data-driven methodologies in Finance, a prominent area of application of Stochastic Analysis.

At the very end, this thesis will briefly elaborate on these financial applications and present the generic pipeline of signature-based methods. Nevertheless, even at this stage, the ongoing discussion already offers some insight on the efficacy of signatures, especially within Mathematical Finance. It is crucial to note that the vast majority of financial data can be conceptualised as paths, with quantities of interest such as payoff functions or trading strategies often seen as continuous functionals of these paths. Therefore, the capacity to extract meaningful features from paths efficiently, alongside the ability to approximate (continuous) functions of paths, holds immense value, both of which are provided by signatures.

For example, in [8] the authors tackle the issue of pricing American options, or more broadly, of solving an optimal stopping problem. Briefly put, they seek to compute $\sup_{\tau \in \mathcal{S}} \mathbb{E}[Y_{\tau \wedge T}]$, where Y represents a process adapted to the filtration generated by an underlying (rough) path process (typically the price process), and $\mathcal{S}$ denotes the set of all stopping times adapted to the same filtration. However, it is often the case that optimal stopping problems are discontinuous when viewed as a function of the underlying (price) path. To address this, the authors introduce *randomised stopping times* and reformulate the problem accordingly. This reformulation involves taking the supremum over the set of *continuous time policies* $\mathcal{T}$, representing real-valued continuous functions on the space of stopped rough paths. They then utilise the universality of signatures to uniformly approximate continuous time policies with linear functionals of the signature over a fixed compact subset of paths. This ultimately leads to a linearisation of the original optimal stopping problem, thus facilitating its solution.

In another work [5], the authors present *signature-payoffs*, a type of path-dependent derivative defined in terms of the signature of the underlying price path. Essentially, these signature-payoffs are linear functionals of the signature. They then establish an appropriate path space and utilise the universality of signatures to demonstrate that, within a compact set of (rough) paths, signature-payoffs offer a uniform approximation to any continuous payoff function. It is worth noting that the majority of financial derivatives, including vanilla and exotic options, stem from continuous payoff functions. Therefore, signatures provide a versatile derivative family capable of approximating an extensive range of derivatives effectively. The authors further demonstrate that signature payoffs can be efficiently priced, primarily due to their linearity, making them valuable for pricing derivatives that are computationally expensive.

These previous examples effectively demonstrate the diverse applications of path signatures within a financial context, particularly highlighting their universal property. However, various other instances showcase their versatility. In [43, 25, 7], the authors utilise signatures and their universal property to address optimal control problems. Similarly, in [6, 1, 28, 23], the authors apply these tools to directly model the dynamics of arbitrary financial assets, leading to the development of *Sig-SDE models*. Meanwhile, in [51, 52], signatures find specific application in pricing and hedging exotic derivatives, whereas [13] employs them for deriving optimal double-execution trading strategies. Additionally, in [38], signatures and universality prove instrumental in solving a portfolio optimisation problem within a path-dependent extension of the classical mean-variance framework. Lastly, in the context of Stochastic Portfolio Theory [33, 34], the authors of [25] introduce the class of *path-functional portfolios*, which can be approximated arbitrarily well by a subclass of linear path-functional portfolios, known as *signature portfolios*.

Let us also remark that signature-based methodologies are often model-agnostic. This means that there is no requirement to specify the dynamics of the market [3, 4]. Thanks to the theory of rough paths, they frequently accommodate very general price paths, including non-Markovian regimes and, in the realm of volatility modelling, non-semimartingale processes.

Now, setting financial applications momentarily aside, it becomes pertinent for the purposes of this thesis to re-examine universality within the framework of kernel theory — one of the many fields where this foundational concept is employed. In summary, a *kernel* on a nonempty set Y is a mapping $k : Y \times Y \to \mathbb{R}$ such that $k(y, y') = \langle \Phi(y), \Phi(y') \rangle_{\tilde{\mathcal{H}}}$, where $\Phi : Y \to \tilde{\mathcal{H}}$ is a map, commonly referred to as *feature map*, with values in a Hilbert space $\tilde{\mathcal{H}}$. Crucially, each kernel k corresponds uniquely to a function space $\mathcal{H}$, known as its *reproducing kernel Hilbert space*. If Y is assumed to be compact and the function space $\mathcal{H}$ is "large enough" to approximate any real-valued continuous function on Y, then, according to the discussion above, $\mathcal{H}$ is considered universal and, without ambiguity, the kernel k is also termed universal.

Ultimately, using this kernel-theoretic terminology, the signature of a path is shown to be a feature map, thus defining a *universal kernel* on a compact subset of the space of paths — the signature kernel [46]. More generally, we observe substantial work in providing sufficient and/or necessary conditions for the existence of universal kernels [57, 69, 70, 20]. Of particular significance to this thesis is the work of [20], where the authors provide a general technique to construct universal kernels defined on generic compact metric spaces, a method that we extensively discuss in the second half of the thesis.

At this juncture, however, we must underscore an assumption that has been pivotal in supporting universality thus far. Specifically, we recall that, in the *classical setting*, universality assumes a *compact* domain over which functions are approximated. In particular, all the aforementioned signature-based methods operate under this assumption, requiring a compact set of paths to invoke the universality of signatures. In other words, all the approximation results derived in the works above, whether in optimisation or pricing contexts, are *local* in nature. Recently, however, there has been a growing interest in the literature of signature-based methods to go beyond compact domains and establish *global approximations.* Notable works in this direction include [19, 27, 41, 9].

Beyond purely theoretical motivations, there are more practical reasons that justify this interest in global approximations, particularly in the context of signature-based methods. Firstly, most path spaces, where signatures are defined, are not even locally compact [19]. Secondly, practical applications frequently entail data resampling. As a result, while a sample may initially reside within a compact subset of the data space, this subset becomes invalid after resampling, making the assumption of a fixed compact set somewhat artificial [27]. The idea of approximating functions *globally*, i.e. across their entire domain, which may indeed be non-(locally) compact, has become known as *global universality*, contrasting with its local counterpart simply referred to as universality. We adhere to this terminology throughout this work.

Regarding signatures specifically, the authors in [27] demonstrate that signatures are *globally universal*, implying that linear functionals of the signature approximate a broad class of functions across the entire space of (rough) paths, and not only a compact subset. They achieve this by first establishing a *weighted version* of the classical Stone-Weierstrass theorem ([27], Theorem 3.6), a result contemplated and utilised in the present thesis. Here, the term "weighted" refers to the setting of *weighted spaces*, an ingenious concept introduced in the same work and also employed in this thesis, essentially corresponding to topological spaces accompanied by a predetermined function that controls the growth of maps therein. On another front, in [19], the authors also establish global universal approximation results for linear functionals of the signature. That said, they rely on the so-called *strict topology* which requires a normalised version of the signature, and not the "true" signature. Consequently, many properties of the signature are lost, resulting in less tractability [27].

Deeply inspired by the concept of global universality and the framework of weighted spaces, part of this thesis finds justification in the following observation: just as universality can be understood from a kernel-theoretic perspective, there is potential for global universality to be integrated into the realm of kernels. With this in mind, this thesis formally introduces *globally universal kernels* — a global analogue to universal kernels — which detain a general weighted space as domain rather than the typically considered compact domain. Furthermore, as suggested by the terminology, the reproducing kernel Hilbert space associated with a globally universal kernel approximates a large set of functions over their entire, possibly non-compact, weighted space domain. When this weighted space is a rough path space, we then obtain a family of *Taylor signature kernels* whose reproducing kernel Hilbert spaces globally approximate a wide range of path functionals, including all bounded and continuous functionals. Along with ([27], Theorem 5.4), this solidifies the global universality of path signatures.

The remainder of this thesis is divided into three chapters. Chapter 1 delves into path signatures and serves as an introduction to Rough Path theory. It is arguably the most intricate chapter of the thesis, and equally necessary, as it introduces the spaces of rough paths, typically used in signature-based methods. Meanwhile, Chapter 2 offers a primer on the general theory of reproducing kernels and acts as a bridge between Chapters 1 and 3 by systematically examining the notion of universality and reviewing the recently developed signature kernel. Moving forward, Chapter 3 picks up the thread on universality and presents global universality, offering a clear comparison between local and global approximations. To achieve this, the chapter initially explores weighted spaces and concludes by defining globally universal kernels and proving their existence. As a side remark, this thesis aims to be as self-contained as possible. Consequently, some sections may appear rather pedantic. For instance, Section 1.1 includes basic results in Analysis and Algebra, while Section 2.1 covers elementary kernel theory. Readers familiar with these topics may choose to skip these sections or use them as a reference when needed.

Another guiding principle adhered to throughout this thesis was to consistently strive to make pertinent observations and additions even to well-established bodies of work. Taking this into consideration, the present work makes the following contributions:

1. Within Chapter 1, we present an under-explored pathway leading to what is arguably the main result of Rough Path theory, *Lyons' Extension theorem*. The proof presented in this thesis deviates from the original, and heavily relies on a somewhat abstract result known as the *Sewing lemma*. Towards the chapter's end, we provide a simple proof demonstrating that the commonly used topology, which endows the domain of the signature map, is not initial. Although this observation builds on a technique frequently employed in the literature of rough paths, namely the compact embedding of (α-Hölder) rough path spaces, our formulation specifically addresses the continuity of the signature map, warranting its emphasis.

2. In Chapter 2, following a succinct review of what has become known as the signature kernel, we establish a basic yet significant property of this kernel. Specifically, we prove that the signature kernel is fully interpolating, or equivalently, strictly positive definite.

3. Chapter 3, in turn, encapsulates the most substantial contribution of this thesis. Here, we introduce the concept of *globally universal kernels* and outline sufficient conditions for their existence. More precisely, we provide a technique to construct globally universal kernels of *Taylor type*. We then apply the developed tools to the spaces of rough paths and to the signature kernel introduced in Chapters 1 and 2, respectively, thus consolidating the global universality of path signatures. Additionally, we provide a formal argument supporting the added benefits of global approximations when compared to local approximations.

Frequently Used Notation

Finite-dimensional objects

$\mathbb{R}^d$	Euclidean space with basis $\{e_1, \dots, e_d\}$
$(\mathbb{R}^d)^{\otimes n}$	n-fold tensor product of $\mathbb{R}^d$
$T^N(\mathbb{R}^d)$	Truncated tensor algebra
$G^N(\mathbb{R}^d)$	Free nilpotent group of step N over $\mathbb{R}^d$
$\pi_{\leq N}, \pi_N$	Projections onto $T^N(\mathbb{R}^d)$ and $(\mathbb{R}^d)^{\otimes n}$
$\mathcal{W}(A_d)$	Set of words over the alphabet $A_d = \{1, \dots, d\}$
$\lvert \cdot \rvert$	Euclidean norm on $\mathbb{R}^d$ or $(\mathbb{R}^d)^{\otimes n}$
$\lvert \cdot \rvert_{T^N(\mathbb{R}^d)}$	Banach space norm on $T^N(\mathbb{R}^d)$
$\lVert \cdot \rVert_{cc}$	Carnot-Carathéodory norm on $G^N(\mathbb{R}^d)$
$\Delta^n_{[s,t]}$	n-simplex $\{(t_1, \dots, t_n) \in \mathbb{R}^n : s < t_1 < \cdots < t_n < t\}$, $\Delta^2_T \equiv \Delta^2_{[0,T]}$

Path spaces and rough paths

$T((\mathbb{R}^d))$	Extended tensor algebra with elements $\mathbf{a} = (\mathbf{a}^i)_{i \geq 0}$
$T_1((\mathbb{R}^d))$	Elements of $\tilde{T}((\mathbb{R}^d))$ with finite Euclidean norm
$S(X)$	Signature of a path X
$C([0,T], E)$	Continuous paths $[0,T] \to E$ with norm $\lvert \cdot \rvert_{\infty;[0,T]}$
$C_o([0,T], E)$	Continuous paths such that $X_0 = o$
$C^{\alpha\text{-Höl}}([0,T], E)$	α-Hölder continuous paths with seminorm $\lvert \cdot \rvert_{\alpha\text{-Höl};[0,T]}$
$C^{p\text{-var}}([0,T], E)$	Continuous paths of finite p-variation with seminorm $\lvert \cdot \rvert_{p\text{-var};[0,T]}$
$\mathbb{X}$	Multiplicative functional $\Delta^2_T \to T^n(\mathbb{R}^d)$
$\mathbf{X}$	An element of $C\big([0,T], G^N(\mathbb{R}^d)\big)$
$\hat{\mathbf{X}}$	Rough path lift of time-augmented path, i.e., $\hat{\mathbf{X}}^{(1)} = \hat{X}_t = (t, X_t)$
$d_{\alpha\text{-Höl};[0,T]}$	Homogenous distance over $C^{\alpha\text{-Höl}}\big([0,T], G^N(\mathbb{R}^d)\big)$
$\rho_{\alpha\text{-Höl};[0,T]}$	Inhomogenous distance over $C^{\alpha\text{-Höl}}\big([0,T], G^N(\mathbb{R}^d)\big)$

Kernels and weighted spaces

k	Real-valued kernel function $k : X \times X \to \mathbb{R}$
$C(X)$	Space of real-valued continuous functions $f : X \to \mathbb{R}$
(X, ψ)	Weighted space X with admissible weight function $\psi : X \to (0, \infty)$
$\mathcal{B}_\psi(X)$	Weighted function space with norm $\lvert \cdot \rvert_{\mathcal{B}_\psi(X)}$

1 Path Signatures and Rough Paths

One of the most compelling motivations behind Rough Path theory and, in particular, path signatures, lies in the examination of *controlled differential equations*, represented as

$$\mathrm{d}Y_t = f(Y_t)\,\mathrm{d}X_t := \sum_{i=1}^{d} f_i(Y_t)\,\mathrm{d}X_t^i, \tag{1.1}$$

where $f = (f_1, \dots, f_d)$ represents a collection of continuous vector fields $f_i : \mathbb{R}^e \to \mathbb{R}^e$; $X : [0,T] \to \mathbb{R}^d$ with components X^i corresponds to an input signal, typically referred to as the *control*; and $Y : [0,T] \to \mathbb{R}^e$ denotes the output or solution.

Conceptually, at each instant, Y_t describes the state of a given system, subject to changes governed by f and dependent on infinitesimal variations of some external parameter described by X. In particular, there is interest in understanding the above equation for irregular signals X, i.e., "rough paths", as irregular paths are commonplace in real-world phenomena. Take, for instance, the field of Stochastic Analysis, where various rescaled models converge to rough objects.

A common way to interpret the differentials in (1.1) is through the integral equation

$$Y_t = Y_0 + \int_0^t f(Y_s)\,\mathrm{d}X_s, \tag{1.2}$$

where Y_0 denotes a given initial condition. In doing so, the challenge of making sense of a controlled differential equation (1.1) transforms into determining how to define the integral in (1.2). Additionally, it is desirable to achieve this in such a manner that the mapping $(Y_0, X) \mapsto Y$, i.e. the solution map, is continuous. This continuity is valuable as it enables us to handle complex dynamics through approximations. That said, it presents the highly nontrivial problem of determining a suitable topology in the path space under consideration. Rough Path theory introduces novel path spaces and topologies that are strong enough to ensure the continuity of the solution map.

As a means to further motivate the theory of rough paths, we now recap different attempts to assign meaning to the integral in (1.2) (see [2]) and observe how these attempts fall short when dealing with rougher signals. For our purposes, we restrict ourselves to continuous paths $X : [0,T] \to \mathbb{R}^d$ and $Y : [0,T] \to L(\mathbb{R}^d, \mathbb{R}^e)$, where $L(\mathbb{R}^d, \mathbb{R}^e)$ denotes the space of linear maps from $\mathbb{R}^d$ to $\mathbb{R}^e$ equipped with the usual operator norm. By varying the regularity of both X and Y, we will investigate the existence of the integral

$$\int_0^T Y_s \,\mathrm{d}X_s.$$

A natural initial approach is given by the Riemann-Stieltjes integral. Let us consider a sequence of partitions $P_n = \{0 = t_0^n < t_1^n < \ldots < t_{N_n}^n = T\}$, $n \geq 1$, with vanishing mesh size, i.e.

$$\max\{|t_{i+1}^n - t_i^n| : i = 0, 1, \ldots, N-1\} \to 0 \text{ as } n \to \infty.$$

Furthermore, for each $n \geq 1$ and $i \in \{0, 1, \ldots, N-1\}$, let u_i^n denote an arbitrary test point in the interval $[t_i^n, t_{i+1}^n]$. Then, the Riemann-Stieltjes integral of Y with respect to X, when it exists, is defined as

$$\int_0^T Y_s \,\mathrm{d}X_s := \lim_{n\to\infty} \sum_{i=0}^{N_n-1} Y_{u_i^n}(X_{t_{i+1}^n} - X_{t_i^n}). \tag{1.3}$$

To ensure existence, it suffices for X to be of bounded variation. Additionally, it is worth noting that the limit above is independent of the sequence of partitions or the choice of test points.

Theorem I ([37], Proposition 2.2)**:** Let $X \in C^{1\text{-var}}([0,T], \mathbb{R}^d)$ be a continuous path of bounded variation, and $Y : [0,T] \to L(\mathbb{R}^d, \mathbb{R}^e)$ continuous. Then, the Riemann-Stieltjes integral $\int_0^T Y \,\mathrm{d}X$ exists, is linear in both X and Y, and

$$\left| \int_0^T Y \,\mathrm{d}X \right| \leq |Y|_{\infty;[0,T]} \; |X|_{1\text{-var};[0,T]}.$$

Here, $C^{1\text{-var}}([0,T], \mathbb{R}^d)$ denotes the space of continuous paths with bounded variation from $[0,T]$ to $\mathbb{R}^d$, $|\cdot|_{1\text{-var};[0,T]}$ corresponds to the 1-variation norm and $|\cdot|_{\infty;[0,T]}$ denotes the usual supremum norm. This notation will be formally introduced in the subsequent section. For now, the main takeaway is that the integral varies continuously with both the integrand and the integrator.

Remarkably, if the regularity of Y remains unchanged, then the following result demonstrates that requiring bounded variation for X is not only sufficient but also necessary. We provide a proof of this result in Appendix B, given that it serves as a captivating application of the Banach-Steinhaus theorem (Appendix B), showcasing the efficacy of functional analytical arguments.

Theorem II ([63], Theorem 56)**:** If the sums in (1.3) converge for every continuous map $Y : [0,T] \to L(\mathbb{R}^d, \mathbb{R}^e)$, then X is of finite variation.

In many situations, however, the control X exhibits low regularity and is far from being of bounded variation. For example, $X : [0,T] \to \mathbb{R}^d$ may be an α-Hölder continuous path with $\alpha \in (0,1)$. This implies the existence of a constant $C > 0$ such that

$$|X_t - X_s| \leq C|t-s|^{\alpha},$$

for all $s,t \in [0,T]$ with $s < t$. It is worth noting that the lower the exponent α, the "rougher" the path can be. In such cases, Riemann-Stieltjes integration is inapplicable unless we change the regularity of the integrand, as demonstrated by Theorem II. This leads to the introduction of the Young integral. In essence, the idea is to offset the lower regularity of the integrator by imposing a higher regularity on the integrand.

Theorem III ([75])**:** Let $X : [0,T] \to \mathbb{R}^d$ be an α-Hölder continuous path and $Y : [0,T] \to L(\mathbb{R}^d, \mathbb{R}^e)$ a β-Hölder continuous path, such that $\alpha, \beta \in (0,1]$ and $\alpha + \beta > 1$. Moreover, let $P_n = \{0 = t_0^n < t_1^n < \ldots < t_{N_n}^n = T\}$, $n \geq 1$ be a sequence of partitions with vanishing mesh size, and let u_i^n denote an arbitrary test point in the interval $[t_i^n, t_{i+1}^n]$. Then, the integral

$$\int_0^T Y_s \, \mathrm{d}X_s := \lim_{n\to\infty} \sum_{i=0}^{N_n - 1} Y_{u_i^n}(X_{t_{i+1}^n} - X_{t_i^n})$$

exists and we call it the Young integral.

Similar to the Riemann-Stieltjes integral, we can establish that the Young integral is a continuous mapping

$$(X,Y) \mapsto \int_0^{\cdot} Y \, \mathrm{d}X$$

with respect to the respective topologies over the space of α-Hölder and β-Hölder continuous paths. This result can be proven using the Young-Lóeve estimate ([37], Proposition 6.4), or the so-called Sewing Lemma ([74], Theorem 3.3). Considering its significance to Rough Path theory, we explore the latter method in a subsequent section.

We note that if $\alpha = 1$, then the Young integral turns into a Riemann-Stieltjes integral. Furthermore, if $\alpha = \beta$, then, in order to satisfy $\alpha + \beta > 1$, α must exceed $1/2$. This observation specifically indicates that the Young integral lacks sufficient strength when considering highly irregular paths, namely α-Hölder continuous paths with $\alpha \leq 1/2$. Under the regime $\alpha = \beta$, which is often the case, we arrive at the conclusion that the Young integral cannot even be applied to Brownian motion, or, more generally, to fractional Brownian motion with Hurst parameter $H \leq 1/2$.

Theorem IV ([60])**:** Fractional Brownian motion B^H does not have α-Hölder continuous trajectories on $[0,T]$ for $\alpha \geq H$ almost surely.

With the aim of integrating against Brownian motion, we arrive at Itô Calculus and the concept of the stochastic integral. In a nutshell, recall that the construction of the Itô integral mirrors that of the Lebesgue integral in the sense that one defines the Itô integral as the $L^2(\mathbb{P})$ limit of simple (adapted) processes [68]. Subsequently, one can extend the definition of the stochastic integral to a broader class of processes: local martingales and semimartingales [56]. Given a continuous semimartingale X and a left-continuous locally bounded adapted process Y, it can be shown that the Itô integral of Y against X can be expressed as the limit in probability:

$$\sum_i Y_{t_i^n}(X_{t_{i+1}^n} - X_{t_i^n}) \xrightarrow{\mathbb{P}} \int_0^T Y_s \, \mathrm{d}X_s.$$

The crucial observation here is that the stochastic integral does not have a *pathwise* definition, and heavily depends on probability tools, such as martingales. Furthermore, it is a classical exercise in Stochastic Calculus to demonstrate that for the stochastic integral to be a martingale, we must select the left endpoint on each subinterval of the partitions as the test point. Opting for alternative test points leads to different definitions of the stochastic integral, such as the Stratonovich integral [45].

Overall, the regularity limitations of the integrals discussed above, combined with the non-analytical nature of the stochastic integral, illustrate that there is room for a pathwise theory of integration capable of handling highly irregular paths. This is Rough Path theory. In this chapter, we begin our exploration of Rough Path theory, with Section 1.3 specifically resuming the ongoing discussion on integration. Lastly, we note that this chapter serves solely as an introduction to Rough Path theory and intentionally omits the study of rough differential equations. Instead, the primary focus of Chapter 1 lies on the spaces of rough paths, which provide an appropriate class of integrators for subsequently defining rough integrals and rough differential equations for a suitable class of integrands.

1.1. Preliminaries to Rough Path Theory

We establish the prerequisites for path signatures and rough paths. Rather than providing a comprehensive overview, the aim is to provide a concise presentation of the concepts underpinning the subsequent sections, emphasising particularly useful results. All material covered in this section is classical, and hence most proofs are omitted. In instances where further elucidation is required, references to the appendices will be made for more detailed explanations and brief theoretical recaps.

Concurrently, this section introduces notation that will be utilised throughout. To enhance accessibility, a table summarising the notation is appended at the beginning of this work. We structure the presentation into two parts - one with a more analytical focus (Sections 1.1.1 and 1.1.2) and the other with a more algebraic emphasis (Section 1.1.3).

1.1.1. Path Spaces and Regularity

We begin by establishing the definition of a path and presenting pertinent results regarding the space of paths. Following this, we introduce the notions of regularity that will be utilised throughout, specifically defining spaces of finite p-variation and α-Hölder spaces. To conclude this section, we demonstrate the compact embedding between α-Hölder spaces. References are provided for readers interested in reviewing any of the results mentioned.

Definition 1.1: Let (E, d) be a metric space and consider some interval $[0, T] \subset \mathbb{R}$. We denote by $C([0, T], E)$ the set of all continuous functions $[0, T] \to E$. We refer to an element X of $C([0, T], E)$ as a path and denote $X(t)$ by X_t, for all $t \in [0, T]$. The path increment $X_t - X_s$ is denoted by $X_{s,t}$ for all $s < t$ in $[0, T]$. Additionally, we equip $C([0, T], E)$ with the supremum metric given by

$$d_{\infty;[0,T]}(X, Y) := \sup_{t\in[0,T]} d(X_t, Y_t).$$

Lastly, for a fixed $o \in E$, we agree that $C_o([0, T], E)$ denotes the subset of paths that start at o, i.e. all the $X \in C([0, T], E)$ such that $X(0) = o$.

Remark 1.1: The interval $[0, T]$ in Definition 1.1 can of course be replaced by any other interval $[a, b] \subset \mathbb{R}$. All the results below are readily adapted if we consider $[a, b]$ instead of $[0, T]$. Furthermore, as we shall observe, many results remain invariant under reparametrization.

It is straightforward to see that $d_{\infty;[0,T]}$ defines a metric on $C([0, T], E)$, thereby establishing a metric space. We refer to the topology induced by $d_{\infty;[0,T]}$ as the uniform or supremum topology. Moreover, we may use $d_{\infty;[0,T]}$ to define a norm. Indeed, given some element $o \in E$, we may identify it with the constant path and set

$$|X|_{\infty;[0,T]} := \sup_{t\in[0,T]} d(X_t, o).$$

Whenever E has a group structure, a common choice for o is the identity element. For instance, if $E = \mathbb{R}^d$, then o is taken to be the origin and $|X|_{\infty;[0,T]} = \sup_{t\in[0,T]} |X_t|$, where $|\cdot|$ denotes the usual Euclidean norm. It is pertinent to recall the following topological properties of $C([0, T], E)$.

Proposition 1.1 ([58], Theorem 21.6)**:** Any continuous mapping from $[0, T]$ to E is uniformly continuous.

Proposition 1.2 ([58], Theorem 43.6)**:** If (E, d) is a complete metric space, then $C([0, T], E)$ is complete under d_∞.

Considering its significance for future results, we also include the Arzelà-Ascoli theorem ([37], Theorem 1.4) that provides a criteria for identifying compact sets in $C([0,T],E)$, under mild assumptions on E.

Definition 1.2: A subset of paths $K \subset C([0,T],E)$ is said to be equicontinuous if for all $\varepsilon > 0$, there exists a $\delta > 0$ such that $|t-s| < \delta$ implies $d(X_t, X_s) < \varepsilon$ for all $X \in K$. Furthermore, it is said to be uniformly bounded if $\sup_{X \in K} |X|_{\infty;[0,T]} < \infty$.

Theorem 1.1 (Arzelà-Ascoli)**:** Let (E,d) be a complete metric space in which bounded subsets have compact closure. Then $K \in C([0,T],E)$ has compact closure, i.e. it is pre-compact (or relatively compact), if and only if K is equicontinuous and uniformly bounded.

Recall that a topological space is said to have the Heine-Borel property if each closed bounded set is compact, or, equivalently, if each bounded set is relatively compact. We now provide an overview of the various types of path regularity that will be used.

Definition 1.3: Let $[0,T] \subset \mathbb{R}$ be some interval. We denote by $\mathcal{P}([0,T])$ the set of partitions of $[0,T]$. A typical element $(t_i) \in \mathcal{P}([0,T])$ consists of n adjacent intervals $[t_i, t_{i+1}]$ that we usually write as $\{0 = t_0 < t_1 < \ldots < t_n = T\}$. We set $|(t_i)| := \max_{i=1,\ldots,n} |t_i - t_{i-1}|$ and refer to this quantity as the mesh of (t_i).

Definition 1.4: Let (E,d) be a metric space. Then, $X : [0,T] \to E$ is said to be:

1. α-Hölder continuous with exponent $\alpha \geq 0$ if

$$|X|_{\alpha\text{-Höl};[0,T]} := \sup_{0 \leq s \leq t \leq T} \frac{d(X_s, X_t)}{|t-s|^\alpha} < \infty$$

2. of finite p-variation for a fixed $p > 0$ if

$$|X|_{p\text{-var};[0,T]} := \left(\sup_{(t_i) \in \mathcal{P}([0,T])} \sum_i d(X_{t_i}, X_{t_{i+1}})^p \right)^{1/p} < \infty.$$

We denote the set of α-Hölder continuous paths by $C^{\alpha\text{-Höl}}([0,T],E)$, and the set of continuous paths of finite p-variation by $C^{p\text{-var}}([0,T],E)$. The subsets of α-Hölder continuous paths and continuous paths of finite p-variation that start at $o \in E$ are denoted by $C_o^{\alpha\text{-Höl}}([0,T],E)$ and $C_o^{p\text{-var}}([0,T],E)$, respectively.

Observe that the quantities above do not correspond to norms. Indeed, we have that $X = o$ for some $o \in E$ if and only if $|X|_{\alpha\text{-Höl};[0,T]} = 0$, and if and only if $|X|_{p\text{-var};[0,T]} = 0$. Moreover, it is worth noting that $C^{0\text{-Höl}}([0,T],E) \equiv C([0,T],E)$, and any $\alpha > 0$ can be expressed as $1/p$ for some $p > 0$, making apparent that any $1/p$-Hölder continuous path is a continuous path of finite p-variation. The next proposition shows to what extent the opposite claim is true. In addition, we include another result that justifies the common practice of considering $\alpha \in [0,1]$ and $p \geq 1$.

Proposition 1.3 ([37], Proposition 1.21)**:** Consider $X \in C([0,T],E)$. Then, X is of finite p-variation if and only if there exists a continuous increasing function $h : [0,T] \to [0,1]$ and a $1/p$-Hölder continuous path $Y : [0,1] \to E$ such that $X = Y \circ h$.

Proposition 1.4 ([37], Proposition 5.2)**:** Assume that X is α-Hölder continuous with $\alpha \in (1,\infty)$, or X is continuous of finite $p-$variation with $p \in (0,1)$. Then, X is a constant path.

Next, we highlight the existence of continuous embeddings for both Hölder and finite p-variation spaces. In the specific case of Hölder spaces, we prove the existence of compact embeddings, meaning that, whenever $0 < \beta < \alpha \leq 1$, any bounded subset with respect to $|\cdot|_{\alpha\text{-Höl};[0,T]}$ yields a relatively compact subset in $C^{\beta\text{-Höl}}([0,T],E)$. We conclude this section by establishing a condition for the completeness of these spaces.

Proposition 1.5 ([37], Proposition 5.3)**:** Consider $X \in C([0,T],E)$. If $1 \leq p \leq q < \infty$, then $|X|_{q\text{-var};[0,T]} \leq |X|_{p\text{-var};[0,T]}$. In particular, we have $C^{p\text{-var}}([0,T],E) \subset C^{q\text{-var}}([0,T],E)$. Analogously, if $0 \leq \beta \leq \alpha \leq 1$, then $C^{\alpha\text{-Höl}}([0,T],E) \subset C^{\beta\text{-Höl}}([0,T],E)$.

Proposition 1.6 ([2], Lemma 2.5)**:** Consider a Banach space $(E, |\cdot|_E)$ in which bounded subsets have a compact closure and $0 < \beta < \alpha \leq 1$. Let $(X^n)_{n\geq 1} \subset C^{\alpha\text{-Höl}}([0,T],E)$ be a sequence of α-Hölder continuous paths such that

$$\sup_{n\geq 1} \left(|X_0^n|_E + |X^n|_{\alpha\text{-Höl};[0,T]} \right) < \infty. \tag{1.4}$$

Then, there exists a path $X \in C^{\alpha\text{-Höl}}([0,T],E)$ and a subsequence $(X^{n_k})_{k\geq 1}$ such that $|X^{n_k} - X|_{\beta\text{-Höl};[0,T]} \to 0$ as $k \to \infty$.

Proof. From (1.4) we can directly conclude that $(X^n)_{n\geq 1}$ is uniformly bounded and equicontinuous. Therefore, by Arzelà-Ascoli (Theorem 1.1), there exists a continuous path X and a subsequence $(X^{n_k})_{k\geq 1}$ such that $X^{n_k} \to X$ uniformly. To prove convergence with respect to $|\cdot|_{\beta\text{-Höl};[0,T]}$, we observe that

$$\frac{|X_{s,t}|_E}{|t-s|^\beta} = \left(\frac{|X_{s,t}|_E}{|t-s|^\alpha} \right)^{\frac{\beta}{\alpha}} |X_{s,t}|_E^{1-\frac{\beta}{\alpha}},$$

which leads to the following interpolation estimate:

$$|X|_{\beta\text{-Höl};[0,T]} \leq |X|^{\frac{\beta}{\alpha}}_{\alpha\text{-Höl};[0,T]} \left(\sup_{0\leq s<t\leq T} |X_{s,t}|_E \right)^{1-\frac{\beta}{\alpha}}. \tag{1.5}$$

Furthermore, for all $0 \leq s < t \leq T$, we have

$$\frac{|X_{s,t}|_E}{|t-s|^\alpha} = \liminf_{k\to\infty} \frac{|X^{n_k}_{s,t}|_E}{|t-s|^\alpha} \leq \liminf_{k\to\infty} |X^{n_k}|_{\alpha\text{-Höl};[0,T]}.$$

Hence, $|X|_{\alpha\text{-Höl};[0,T]} \leq \liminf_{k\to\infty} |X^{n_k}|_{\alpha\text{-Höl};[0,T]}$, i.e. $|\cdot|_{\alpha\text{-Höl};[0,T]}$ is lower semi-continuous. Now, by (1.5), it follows that

$$|X^{n_k} - X|_{\beta\text{-Höl};[0,T]} \leq |X^{n_k} - X|^{\frac{\beta}{\alpha}}_{\alpha\text{-Höl};[0,T]} \left(\sup_{0\leq s<t\leq T} |X^{n_k}_{s,t} - X_{s,t}|_E \right)^{1-\frac{\beta}{\alpha}}. \tag{1.6}$$

Additionally, by lower semi-continuity and (1.4), we conclude that

$$|X|_{\alpha\text{-Höl};[0,T]} \leq \liminf_{k\to\infty} |X^{n_k}|_{\alpha\text{-Höl};[0,T]} \leq \sup_{k\geq 1} |X^{n_k}|_{\alpha\text{-Höl};[0,T]} < \infty,$$

and hence $\sup_{k\geq 1} |X^{n_k} - X|_{\alpha\text{-Höl};[0,T]} < \infty$. Since $X^{n_k} \to X$ uniformly, the RHS of (1.6) tends to zero as $k \to \infty$, and the result follows. □

Remark 1.2: We refer to Proposition 5.28 in [37] for the specific case $E = \mathbb{R}^d$. Also, recall that for metrizable spaces the usual notion of compactness via open coverings and the notion of sequentially compact coincide (see [58], Theorem 28.2).

Proposition 1.7: Consider $\alpha \in [0,1]$ and $p \geq 1$. If $(E, |\cdot|_E)$ is a Banach space, then $C^{\alpha\text{-Höl}}([0,T], E)$ is a complete metric space under the norm $X \mapsto |X_0|_E + |X|_{\alpha\text{-Höl};[0,T]}$. Likewise, $C^{p\text{-var}}([0,T], E)$ is a complete metric space under the norm $X \mapsto |X_0|_E + |X|_{p\text{-var};[0,T]}$.

For instance, both $C^{\alpha\text{-Höl}}([a,b], \mathbb{R}^d)$ and $C^{p\text{-var}}([a,b], \mathbb{R}^d)$ are Banach spaces with norms as in Proposition 1.7. However, these Banach spaces are not separable ([37], Example 5.26).

1.1.2. Bounded Variation Paths and Differential Equations

We now specialise to the case when $p = 1$. These paths hold a particular significance in what is to come. We explore the connection between Lipschitz paths ($\alpha = 1$) and continuous paths of bounded variation ($p = 1$). Additionally, we briefly digress into absolutely continuous paths and recall their integral representation, which is pertinent for our purposes as it relates closely to the length of a path. We end the section by stating a Picard-Lindelöf type result.

This section only serves a purpose in Section 1.2.2, where we establish basic properties of path signatures. As such, it is not highly relevant to the thesis as a whole. Nonetheless, it contributes to the self-contained nature of this work without becoming excessively pedantic.

Definition 1.5: Let (E,d) be a metric space. If $X \in C^{1\text{-var}}([0,T],E)$, i.e. X is continuous and

$$|X|_{1\text{-var};[0,T]} = \sup_{(t_i)\in\mathcal{P}([0,T])} \sum_i d(X_{t_i}, X_{t_{i+1}}) < \infty,$$

then we say that X is a path of bounded variation.

As mentioned earlier, any $1/p$-Hölder continuous path is a continuous path with finite p-variation. Specifically, any Lipschitz path qualifies as a continuous bounded variation path, owing to the estimate

$$|X|_{1\text{-var};[s,t]} \leq |X|_{1\text{-Höl};[s,t]}|t-s|.$$

We also note that during the proof of Proposition 1.6, it was demonstrated that α-Hölder norms exhibit lower semi-continuity. The same is true for bounded variation paths. Precisely, if $(X^n)_{n\geq 1}$ represents a sequence of finite 1-variation paths that converge pointwise to a path $X : [0,T] \to E$, then

$$|X|_{1\text{-var};[0,T]} \leq \liminf_{n\to\infty} |X^n|_{1\text{-var};[0,T]}.$$

Indeed, it is sufficient to observe that for any partition $(t_i) \in \mathcal{P}([0,T])$, we have

$$\sum_i d(X_{t_i}, X_{t_{i+1}}) = \liminf_{n\to\infty} \sum_i d(X^n_{t_i}, X^n_{t_{i+1}}) \leq \liminf_{n\to\infty} |X^n|_{1\text{-var};[0,T]}.$$

Under identical assumptions, we easily establish lower semi-continuity for the 1-Hölder norm.

Proposition 1.8 ([37], Lemma 1.23)**:** Let $(X^n)_{n\geq 1}$ be a sequence of finite 1-variation paths from $[0,T]$ to some metric space E. Assume $X^n \to X$ pointwise on $[0,T]$. Then, for all $s<t$ in $[0,T]$,

$$|X|_{1\text{-Höl};[0,T]} \leq \liminf_{n\to\infty} |X^n|_{1\text{-Höl};[0,T]}.$$

While not strictly essential for the present work, it is helpful to revisit the concept of absolute continuity for a better understanding of forthcoming results, particularly those in Section 1.2.3. Additionally, absolute continuous paths are closely related to the space $C^{1\text{-var}}([0,T],E)$.

Definition 1.6: Let (E,d) be a metric space. The path $X : [0,T] \to E$ is said to be absolutely continuous if for all $\varepsilon > 0$, there exists $\delta > 0$, such that for all $s_1 < t_1 \leq s_2 < t_2 \leq \ldots \leq s_n < t_n$ in $[0,T]$ with $\sum_i |t_i - s_i| < \delta$, we have $\sum_i d(X_{t_i}, X_{s_i}) < \varepsilon$.

Proposition 1.9 ([37], Proposition 1.18)**:** If $X : [0,T] \to E$ is absolutely continuous, then $X \in C^{1\text{-var}}([0,T],E)$.

Moreover, it is relatively straightforward to demonstrate that any Lipschitz path X is absolutely continuous. Specifically, for any $\varepsilon > 0$, one can simply select $\delta = \varepsilon/|X|_{1\text{-Höl};[0,T]}$.

For the remainder of this section, we specialise to the case where $E = \mathbb{R}^d$, focusing solely on paths with state-space $\mathbb{R}^d$.

Remark 1.3: Recall that the significance of absolutely continuous paths stems from their integral representation. Specifically, if $X : [0,T] \to \mathbb{R}^d$ is absolutely continuous, then

$$X_t = X_0 + \int_0^t \dot{X}_s \,\mathrm{d}s,$$

for a unique $\dot{X} \in L^1([0,T], \mathbb{R}^d)$ ([37], Proposition 1.32). This result is a basic consequence of the Radon–Nikodym theorem in Measure theory. Moreover, if $X \in C^{1\text{-Höl}}([0,T], \mathbb{R}^d)$, then $\dot{X}$ can be chosen (uniquely) from $L^\infty([0,T], \mathbb{R}^d)$ ([37], Proposition 1.37).

By Proposition 1.3, any continuous bounded variation path is a continuous time-change (i.e. reparametrization) of a Lipschitz path. Utilising the notation of Proposition 1.3, the reparametrization h is typically chosen as

$$h(t) = \frac{|X|_{1\text{-var};[0,t]}}{|X|_{1\text{-var};[0,T]}},$$

yielding a continuous increasing function from $[0,T]$ to $[0,1]$. This parametrization is commonly referred to as the "arc-length parametrization", a designation that becomes evident with the next result.

We also observe that the 1-variation of a path remains invariant under reparametrization. Therefore, if Y is the Lipschitz path such that $X = Y \circ h$, then

$$|X|_{1\text{-var};[0,T]} = |Y|_{1\text{-var};[0,1]}.$$

Furthermore, due to the additivity of $|\cdot|_{1\text{-var};[s,t]}$ as a function of s and t, we have $|Y|_{1\text{-Höl};[0,1]} \leq |X|_{1\text{-var};[0,T]}$ (see [37], Remark 1.22). Since trivially $|Y|_{1\text{-var};[0,1]} \leq |Y|_{1\text{-Höl};[0,1]}$, we obtain that

$$|Y|_{1\text{-Höl};[0,1]} = |X|_{1\text{-var};[0,T]}.$$

This implies that the 1-variation of X, i.e. the length of X, is equivalent to the Lipschitz norm of Y. The following proposition indicates that it is possible to reparametrize a continuous bounded variation path into a Lipschitz path in a manner such that the reparametrized path maintains constant speed. Moreover, this speed corresponds to the arc-length of the reparametrized path.

Proposition 1.10 ([37], Proposition 1.38)**:** Let $X \in C^{1\text{-var}}([0,T], \mathbb{R}^d)$ be a non-constant continuous path of bounded variation. Define $Y(\cdot)$ by $Y \circ h = X$, where

$$h(t) = \frac{|X|_{1\text{-var};[0,t]}}{|X|_{1\text{-var};[0,T]}}.$$

Then, $Y \in C^{1\text{-Höl}}([0,1],\mathbb{R}^d)$ and Y is the indefinite integral of some $\dot{Y} \in L^\infty([0,1],\mathbb{R}^d)$. Moreover,

$$|\dot{Y}(t)| = |X|_{1\text{-var};[0,T]} = |Y|_{1\text{-Höl};[0,1]}, \text{ for a.e. } t \in [0,1],$$

and,

$$|Y|_{1\text{-var};[0,h(t)]} = \int_0^{h(t)} |\dot{Y}(s)| \,\mathrm{d}s.$$

We conclude this section by presenting a Picard-Lindelöf type result that establishes sufficient conditions to obtain existence and uniqueness of the solution $Y : [0,T] \to \mathbb{R}^e$ for controlled differential equations of the form

$$\mathrm{d}Y_t = f(Y_t)\,\mathrm{d}X_t, \tag{1.7}$$

where $X : [0,T] \to \mathbb{R}^d$ is a continuous path of bounded variation and $f = (f_1,\ldots,f_d)$ is a collection of continuous linear vector fields on $\mathbb{R}^e$, i.e. $f \in C(\mathbb{R}^e, L(\mathbb{R}^d,\mathbb{R}^e))$.

As mentioned in the chapter's introduction, this work primarily introduces the space of rough paths rather than focusing extensively on solving rough differential equations. Thus, we do not provide an exhaustive treatment of the theory of differential equations. Interested readers are directed to Chapter 3 in [37].

Recall that, we define

$$\int_0^t f(Y_s)\,\mathrm{d}X_s := \sum_{i=1}^d \int_0^t f_i(Y_s)\,\mathrm{d}X_s^i,$$

implying that f is indeed a map taking values in $L(\mathbb{R}^d,\mathbb{R}^e)$, which we equip with the operator norm. Specifically, for a given $y \in \mathbb{R}^e$, we have

$$|f(y)|_{op} := \sup_{\substack{a\in\mathbb{R}^d \\ |a|\le 1}} \left| \sum_{i=1}^d f_i(y)a^i \right|.$$

It turns out that to achieve existence of a solution for equation (1.7), we only require mild assumptions on the (continuous) vector fields $f = (f_1,\ldots,f_d)$. Naturally, to obtain uniqueness of solution, we must impose stronger constraints on the regularity of f. These constraints are the content of the following definition.

Definition 1.7: Let $f = (f_1,\ldots,f_d)$ be a collection of vector fields viewed as a map $\mathbb{R}^e \to L(\mathbb{R}^d,\mathbb{R}^e)$. Then, f is said to be bounded if

$$|f|_{\infty,op} := \sup_{y\in\mathbb{R}^e} |f(y)|_{op} < \infty.$$

Additionally, for any $U \subset \mathbb{R}^e$, we define the 1-Lipschitz norm $|\cdot|_{\mathrm{Lip}^1(U)}$ of f by

$$|f|_{\mathrm{Lip}^1(U)} := \max\left\{ \sup_{y,z\in U: y\neq z} \frac{|f(y)-f(z)|_{op}}{|y-z|}, \sup_{y\in U} |f(y)|_{op} \right\}.$$

If $|f|_{\mathrm{Lip}^1} \equiv |f|_{\mathrm{Lip}^1(\mathbb{R}^e)} < \infty$, we say that $f \in \mathrm{Lip}^1(\mathbb{R}^e)$. If $|f|_{\mathrm{Lip}^1(U)} < \infty$ for all bounded subsets $U \subset \mathbb{R}^e$, we say that f is locally 1-Lipschitz.

We now state an existence and uniqueness result for the solution of (1.7), including only the strictly necessary implications for our purposes.

Theorem 1.2: Consider $X \in C^{1\text{-var}}([0,T],\mathbb{R}^d)$. Let $f = (f_1,\dots,f_d)$ be a locally 1-Lipschitz collection of vector fields on $\mathbb{R}^e$ of linear growth, i.e., for all $i \in \{1,\dots,d\}$, there exists $A > 0$ such that

$$|f_i(y)| \leq A(1+|y|) \text{ for all } y \in \mathbb{R}^e.$$

Then, for a given initial condition $y_0 \in \mathbb{R}^e$, there exists a unique solution to $\mathrm{d}Y_t = f(Y_t)\,\mathrm{d}X_t$ on $[0,T]$. Moreover, the solution map, i.e. $(y_0, X) \mapsto Y$, is continuous with respect to the 1-variation distance over $C^{1\text{-var}}([0,T],\mathbb{R}^d)$.

Proof. See Theorems 3.7, 3.8, and 3.18 in [37]. □

Remark 1.4: As a consequence of uniqueness, we observe that the solution of (1.7) commutes with time-changes. More precisely, if ϕ is a continuous non-decreasing surjection from $[0,T]$ to $[a,b]$ and Y is the unique solution to (1.7), then $Y \circ \phi$ coincides with the unique solution of

$$\mathrm{d}Y_{\phi(t)} = Y_{\phi(t)}\,\mathrm{d}X_{\phi(t)}.$$

This follows from the fact that Riemann-Stieltjes integrals are invariant under reparametrization (see [37], Proposition 3.10).

1.1.3. Tensor Algebras

This section presents the concept of extended tensor algebra, which will play a prominent role in subsequent sections, serving as the state space for various path spaces. For a concise review of the tensor product between vector spaces and its universal property, refer to Appendix A. For our purposes, it suffices to focus on tensor algebras over $\mathbb{R}^d$. In Section 1.2, once we delve into path signatures and grasp the ubiquity of iterated integrals, the significance of the tensor algebra will become apparent.

Definition 1.8: We define the extended tensor algebra over $\mathbb{R}^d$ as the direct product

$$T((\mathbb{R}^d)) := \prod_{n=0}^{\infty} (\mathbb{R}^d)^{\otimes n} = \left\{\mathbf{a} = (\mathbf{a}^n)_{n\geq 0} : \mathbf{a}^n \in (\mathbb{R}^d)^{\otimes n}\right\},$$

where $(\mathbb{R}^d)^{\otimes 0} := \mathbb{R}$. Similarly, we define the tensor algebra over $\mathbb{R}^d$ as the direct sum

$$T(\mathbb{R}^d) := \bigoplus_{n=0}^{\infty} (\mathbb{R}^d)^{\otimes n} = \left\{\mathbf{a} = (\mathbf{a}^n)_{n\geq 0} : \mathbf{a}^n \in (\mathbb{R}^d)^{\otimes n} \text{ and } \mathbf{a}^n \neq 0 \text{ for finitely many } n \in \mathbb{N}\right\}.$$

Remark 1.5: We note that we may define the (extended) tensor algebra over a generic vector space V. Indeed, those with some knowledge of Category theory might recognise the tensor algebra as being *functorial*, i.e. the tensor algebra construction can be seen as a functor $T : \mathbf{Vect}_{\mathbb{R}} \to \mathbb{R}\text{-}\mathbf{Mod}$, where $\mathbf{Vect}_{\mathbb{R}}$ denotes the category of real vector spaces and $\mathbb{R}\text{-}\mathbf{Mod}$ denotes the category of real modules.

It is also convenient to define the truncated tensor algebra $T^N(\mathbb{R}^d)$ given by

$$T^N(\mathbb{R}^d) := \bigoplus_{n=0}^{N} (\mathbb{R}^d)^{\otimes n}.$$

We observe the canonical embeddings $T^N(\mathbb{R}^d) \hookrightarrow T(\mathbb{R}^d) \hookrightarrow T((\mathbb{R}^d))$, and denote by $\pi_{\leq N}$ and π_N the projections $T((\mathbb{R}^d)) \to T^N(\mathbb{R}^d)$ and $T((\mathbb{R}^d)) \to (\mathbb{R}^d)^{\otimes N}$, respectively. We endow $T((\mathbb{R}^d))$ with the usual component-wise addition and scalar multiplication. Specifically, for all $\mathbf{a}, \mathbf{b} \in T((\mathbb{R}^d))$,

$$\mathbf{a} + \mathbf{b} = (\mathbf{a}^0 + \mathbf{b}^0, \mathbf{a}^1 + \mathbf{b}^1, \ldots) \quad \text{and} \quad \lambda \cdot \mathbf{a} = (\lambda \mathbf{a}^0, \lambda \mathbf{a}^1, \ldots), \text{ for all } \lambda \in \mathbb{R}.$$

Moreover, given $\mathbf{a}, \mathbf{b} \in T((\mathbb{R}^d))$, we define a multiplication operation $\mathbf{a} \otimes \mathbf{b}$ by setting

$$(\mathbf{a} \otimes \mathbf{b})^n = \sum_{i=0}^{n} \mathbf{a}^i \otimes \mathbf{b}^{n-i}. \tag{1.8}$$

Here, the product $\mathbf{a}^i \otimes \mathbf{b}^{n-i}$ is understood as an instance of the canonical isomorphism

$$(\mathbb{R}^d)^{\otimes i} \otimes (\mathbb{R}^d)^{\otimes (n-i)} \to (\mathbb{R}^d)^{\otimes n},$$

given by the usual tensor product (Appendix A). We note that the multiplication in $T^N(\mathbb{R}^d)$ is also given by (1.8), except that the higher-order terms with $n > N$ are omitted. The following straightforward result justifies the suggestive nomenclature of Definition 1.8.

Proposition 1.11 ([37], Proposition 7.4)**:** The space $(T((\mathbb{R}^d)), +, \cdot, \otimes)$ is a real associative algebra with unity $\mathbf{1} := (1, 0, 0, \ldots)$. Similarly, for all $N \geq 1$, $(T^N(\mathbb{R}^d), +, \cdot, \otimes)$ is a real associative algebra with neutral element $\pi_{\leq N}(\mathbf{1})$ and the product $\otimes$ truncated at level N.

We make a slight abuse of notation and denote the unity in $T^N(\mathbb{R}^d)$ by $\mathbf{1}$ as well. Now that we have an algebraic structure, it becomes pertinent to define a norm. To this end, we equip $(\mathbb{R}^d)^{\otimes n}$ with the usual Euclidean norm. Indeed, it is a well-established result that $\{e_{i_1} \otimes \cdots \otimes e_{i_n} : i_1, \ldots, i_n \in \{1, \ldots, d\}\}$ serves as a canonical basis for $(\mathbb{R}^d)^{\otimes n}$, where $(e_i)_{i=1}^d$ denotes the canonical basis of $\mathbb{R}^d$ (Appendix A). This way, any $\mathbf{g} \in (\mathbb{R}^d)^{\otimes n}$ can be written as

$$\mathbf{g} = \sum_{i_1, \ldots, i_n} g^{i_1, \ldots, i_n} e_{i_1} \otimes \cdots \otimes e_{i_n}, \quad \text{and so} \quad |\mathbf{g}|_{(\mathbb{R}^d)^{\otimes n}} := \sqrt{\sum_{i_1, \ldots, i_n} |g^{i_1, \ldots, i_n}|^2}$$

yields a norm on $(\mathbb{R}^d)^{\otimes n}$. Moreover, it becomes apparent that, for $0 \le i \le n$,

$$(\mathbf{g}, \mathbf{h}) \in (\mathbb{R}^d)^{\otimes i} \times (\mathbb{R}^d)^{\otimes (n-i)}, \quad |\mathbf{g} \otimes \mathbf{h}|_{(\mathbb{R}^d)^{\otimes n}} = |\mathbf{g}|_{(\mathbb{R}^d)^{\otimes i}} |\mathbf{h}|_{(\mathbb{R}^d)^{\otimes (n-i)}}.$$

This is typically referred to as "compatibility between tensor norms." Finally, for $\mathbf{g} \in T^N(\mathbb{R}^d)$, we set

$$|\mathbf{g}|_{T^N(\mathbb{R}^d)} := \max_{n=0,\dots,n} |\pi_n(\mathbf{g})|_{(\mathbb{R}^d)^{\otimes n}},$$

Equipped with $|\cdot|_{T^N(\mathbb{R}^d)}$, $T^N(\mathbb{R}^d)$ becomes a Banach space [37]. When no confusion is possible we simply write $|\mathbf{g}|$ instead of $|\mathbf{g}|_{(\mathbb{R}^d)^{\otimes n}}$.

Given that $\mathbb{R}^d$ is a Hilbert space and that the inner product naturally extends to $(\mathbb{R}^d)^{\otimes n}$ (see Appendix A), we can also establish an inner product on $T((\mathbb{R}^d))$, and extract a subset endowed with a Hilbert space structure by considering elements with finite norm.

Definition 1.9: We denote by $\tilde{T}((\mathbb{R}^d))$ the Hilbert space

$$\tilde{T}((\mathbb{R}^d)) := \left\{ \mathbf{a} \in T((\mathbb{R}^d)) : |\mathbf{a}|_{\tilde{T}(\mathbb{R}^d)} := \left(\sum_{n=0}^{\infty} |\mathbf{a}^n|^2 \right)^{1/2} < \infty \right\},$$

where $|\cdot|_{\tilde{T}((\mathbb{R}^d))}$ is the norm induced by the inner product $\langle \mathbf{a}, \mathbf{b} \rangle_{\tilde{T}((\mathbb{R}^d))} := \sum_{n \ge 0} \langle \mathbf{a}^n, \mathbf{b}^n \rangle_{(\mathbb{R}^d)^{\otimes n}}$.

Additionally, in the context of rough paths, it becomes essential to consider yet another subset of $T((\mathbb{R}^d))$. Namely, the linear-affine subspace

$$T_1((\mathbb{R}^d)) := \left\{ \mathbf{a} \in \tilde{T}((\mathbb{R}^d)) : \mathbf{a}^0 = 1 \right\}.$$

Proposition 1.12 ([37], Proposition 7.17)**:** The space $(T_1((\mathbb{R}^d)), \otimes)$ is a Lie group, i.e. $T_1((\mathbb{R}^d))$ is a group with a smooth manifold structure. Moreover, for all $\mathbf{a} \in T_1((\mathbb{R}^d))$, we have

$$\mathbf{a}^{-1} = \sum_{n=0}^{\infty} (\mathbf{1} - \mathbf{a})^{\otimes n}.$$

Although the Lie group structure of $T_1((\mathbb{R}^d))$ will not play a major role in what follows, it is still pertinent to be aware of it. We now identify the extended tensor algebra with the algebra of non-commutative formal power series in d indeterminates.

Definition 1.10: Let $e_1, \dots, e_d$ be d formal indeterminates. The algebra of non-commuting formal power series in d indeterminates, denoted by $\mathbb{R}[[e_1, \dots, e_d]]$, is the vector space of all series of the form

$$\sum_{n=0}^{\infty} \sum_{i_1,\dots,i_n \in \{1,\dots,d\}} \lambda_{i_1,\dots,i_n} e_{i_1} \dots e_{i_n},$$

where the second summation is over the set of multi-indices $(i_1, \dots, i_n) \in \{1, \dots, d\}^n$, and $\lambda_{i_1,\dots,i_n}$ are real coefficients. For $n = 0$ we simply consider $\lambda_0 \in \mathbb{R}$. The subset of formal power series for which only a finite number of coefficients $\lambda_{i_1,\dots,i_n}$ are non-zero is called the algebra of non-commuting formal polynomials, and is denoted by $\mathbb{R}[e_1, \dots, e_d]$. The multiplication operation is given by the usual Cauchy product of series.

Alternatively, we can express the second summation as spanning the set of words w of length n over the alphabet $\{1, \dots, d\}$. A word of length n is essentially a string of n numbers, denoted as $w = i_1 \dots i_n$. With this perspective, we can simplify the notation by representing the formal series as

$$\sum_{n=0}^{\infty} \sum_{|w|=n} \lambda_w e_w,$$

where w serves as shorthand for the multi-index $(i_1, \dots, i_n)$; $|w|$ indicates the length of the word w; $|w| = n$ denotes the set of length n words; and e_w is an abbreviation for $e_{i_1} \dots e_{i_n}$. For $n = 0$, we assign the empty word denoted by $\varnothing$. Considering that a generic element of $(\mathbb{R}^d)^{\otimes n}$ is given by

$$\sum_{i_1,\dots,i_n \in \{1,\dots,d\}} \alpha_{i_1,\dots,i_n} e_{i_1} \otimes \dots \otimes e_{i_n},$$

with e_i denoting an element of the canonical basis of $\mathbb{R}^d$, and that the product $\otimes$ emulates polynomial multiplication, it is apparent that we have the isomorphisms

$$T((\mathbb{R}^d)) \cong \mathbb{R}[[e_1, \dots, e_d]] \quad \text{and} \quad T(\mathbb{R}^d) \cong \mathbb{R}[e_1, \dots, e_d].$$

These identifications are particularly useful whenever we want to "unfold" the tensor notation. At the same time, it highlights the benefit of having a compact notation such as the tensor notation capable of concealing all the multi-indices. With the formal series notation, we see that

$$\begin{aligned} \mathbf{a} \otimes \mathbf{b} = & \left(\sum_{n=0}^{\infty} \sum_{|w|=n} a_w e_w \right) \otimes \left(\sum_{n=0}^{\infty} \sum_{|w|=n} b_w e_w \right) = \\ = & a_0 b_0 + \sum_{i=1}^{d} (a_0 b_i + a_i b_0) e_i + \sum_{i,j=1}^{d} (a_0 b_{i,j} + a_i b_j + a_{i,j} b_0) e_i e_j + \dots, \end{aligned}$$

for all $\mathbf{a}, \mathbf{b} \in T((\mathbb{R}^d))$.

Finally, using the word notation, we highlight the natural pairing between the extended tensor algebra $T((\mathbb{R}^d))$ and its dual space $T((\mathbb{R}^d)^*)$. Recall that $(\mathbb{R}^d)^* \cong \mathbb{R}^d$ denotes the (topological) dual of $\mathbb{R}^d$. Additionally, we define the shuffle product operation between words, an operation that will prove useful in the coming section.

Definition 1.11: The natural pairing $\langle\cdot,\cdot\rangle : T((\mathbb{R}^d)^*) \times T((\mathbb{R}^d)) \to \mathbb{R}$ of the extended tensor algebra with its dual is defined by

$$\langle \mathbf{l}, \mathbf{a} \rangle = \sum_{w \in \mathcal{W}(A_d)} l_w a_w,$$

where $\mathcal{W}(A_d) = \big\{ i_1 \dots i_n : n \in \mathbb{N}_0, i_1, \dots, i_n \in \{1, \dots, d\} \big\}$ denotes the set of all words over the alphabet $\{1, \dots, d\}$, and a_w denotes a component of $\mathbf{a}^{|w|}$. Precisely, $\mathbf{a}^i = \sum_{|w|=i} a_w e_w$ for all $i \in \mathbb{N}_0$.

In line with previous considerations, we can similarly identify basis elements of the dual of the extended tensor algebra with the space of words. Specifically, by letting $\{e_1^*, \dots, e_d^*\}$ denote the dual basis of $(\mathbb{R}^d)^*$, we establish the following correspondence:

$$e_{i_1}^* \otimes \dots \otimes e_{i_n}^* \in T((\mathbb{R}^d)^*) \leftrightarrow w = i_1 \dots i_n \in \mathcal{W}(A_d).$$

Consequently, any linear functional $L : T((\mathbb{R}^d)) \to \mathbb{R}$ can be identified via (formal) linear combinations of elements in $\mathcal{W}(A_d)$, a pertinent observation that will be used throughout in the latter chapters of this work. Moreover, considering the correspondence $e_w \leftrightarrow e_w^*$ for all words $w \in \mathcal{W}(A_d)$, we denote by $\langle e_w, \mathbf{a} \rangle$ the component of $\mathbf{a}$ corresponding to w, i.e. a_w. We now endow $\mathcal{W}(A_d)$ with two operations: concatenation and the shuffle product.

Definition 1.12: Let $w_1 = i_1 \dots i_n$ and $w_2 = j_1 \dots j_m$ be two words in $\mathcal{W}(A_d)$. We define the concatenation of w_1 and w_2 by $(w_1, w_2) \coloneqq i_1 \dots i_n j_1 \dots j_m$. In addition, for words w_1, w_2 and letters $i, j \in \{1, \dots, d\}$, we define the shuffle product $\sqcup\!\sqcup : \mathcal{W}(A_d) \times \mathcal{W}(A_d) \to \mathcal{W}(A_d)$ recursively by $(w_1, i) \sqcup\!\sqcup (w_2, j) \coloneqq ((w_1 \sqcup\!\sqcup (w_2, j)), i) + (((w_1, i) \sqcup\!\sqcup w_2), j)$, with $\varnothing \sqcup\!\sqcup w_1 = w_1 \sqcup\!\sqcup \varnothing = w_1$, and $\varnothing$ denoting the empty word.

In practice, it is more convenient to think of the shuffle product $w_1 \sqcup\!\sqcup w_2$ as the formal sum of all permutations of the letters in w_1 and w_2, while maintaining the order of the letters within each word. This operation resembles shuffling a deck of cards, hence the name. For instance, let $w_1 = i_1 i_2$ and $w_2 = j_1 j_2$. Their shuffle product is expressed as:

$$w_1 \sqcup\!\sqcup w_2 = i_1 i_2 j_1 j_2 + i_1 j_1 i_2 j_2 + i_1 j_1 j_2 i_2 + j_1 j_2 i_1 i_2 + j_1 i_1 j_2 i_2 + j_1 i_1 i_2 j_2.$$

More generally, for words $w_1 = i_1 \dots i_m$ and $w_2 = j_1 \dots j_n$, the shuffle product is given by

$$\sum_{\sigma \in S(m,n)} r_{\sigma(1)} \dots r_{\sigma(m+n)},$$

where $r_1 \dots r_n r_{n+1} \dots r_{m+n} = (w_1, w_2)$ and $S(m, n)$ denotes the set of all permutations σ of $\{1, \dots, n+m\}$ such that $\sigma^{-1}(1) < \dots < \sigma^{-1}(n)$ and $\sigma^{-1}(n+1) < \dots < \sigma^{-1}(n+m)$. The next section will clarify the importance of the shuffle product.

1.2. Path Signatures

We introduce path signatures, a central object in this thesis. In essence, the signature of a path is an infinite series composed of all its iterated integrals. Section 1.2.1 provides motivation for signatures as a natural object in the study of controlled differential equations and presents their definition. Subsequently, Section 1.2.2 outlines basic properties of signatures, and finally, Section 1.2.3 introduces a commonly used notion of distance between path signatures, namely, the Carnot-Carathéodory norm. Signatures will be used throughout in the remaining chapters of this work.

1.2.1. Motivation and Definition

We introduce and define signatures for paths of bounded variation. As emphasised in the chapter's introduction, a compelling motivation for path signatures lies in the study of controlled differential equations. In this section, we illustrate the non-commutativity of iterated integrals, thereby justifying the utilisation of extended tensor algebras.

Recall that, in the context of Rough Path theory, we aim to make sense of differential equations of the form

$$\mathrm{d}Y_t = f(Y_t)\,\mathrm{d}X_t,$$

where $f \in C(\mathbb{R}^e, L(\mathbb{R}^d, \mathbb{R}^e))$, $X \in C([0,T], \mathbb{R}^d)$ and $Y \in C([0,T], \mathbb{R}^e)$. To simplify the problem, we can start by considering

$$\mathrm{d}Y_t = Y_t\,\mathrm{d}X_t \coloneqq \sum_{i=1}^{d} Y_t\,\mathrm{d}X_t^i, \tag{1.9}$$

with some initial condition $Y_0 = y \in \mathbb{R}^e$. Formally, a solution is given by

$$Y_t = y + y\sum_{i=1}^{d}\int_0^t \mathrm{d}X_s^i + y\sum_{i,j=1}^{n}\int_0^t\int_0^s \mathrm{d}X_u^i\,\mathrm{d}X_s^j + \ldots + y\sum_{i_1,\ldots,i_n=1}^{d}\int_{\Delta^n_{[0,t]}} \mathrm{d}X_{t_1}^{i_1}\ldots\mathrm{d}X_{t_n}^{i_n} + \ldots$$

Thus, we observe that provided a proper meaning to the path integrals above, the solution of (1.9) is determined by the sequence of iterated integrals of the control (i.e. the driving signal).

Likewise, if we consider some linear map $f \in C(\mathbb{R}^e, L(\mathbb{R}^d, \mathbb{R}^e))$, we may employ the standard procedure of Picard iterations to seek a solution ([18], Section 1.2.3). To begin with, note that for $x \in \mathbb{R}^d$, each $y \in \mathbb{R}^e$ yields a map $\big(x \mapsto f(y)x\big) \in L(\mathbb{R}^d, \mathbb{R}^e)$. Equivalently, we may parameterise this map by x and consider $x \mapsto \big(y \mapsto f(y)x\big)$. Thus, f can be equivalently interpreted as a linear map $\mathbb{R}^d \to L(\mathbb{R}^e, \mathbb{R}^e)$. Denoting the identity operator in $L(\mathbb{R}^e, \mathbb{R}^e)$ by

I_e, the Picard iterations $(Y_t^n)_{n\geq 0}$ can be expressed as follows:

$$\begin{aligned}
Y_t^0 &= y,\\
Y_t^1 &= y + \int_0^t f(Y_s^0)\,\mathrm{d}X_s =: \left(\int_0^t f(\,\mathrm{d}X_s) + I_e\right)(y),\\
Y_t^2 &= y + \int_0^t f(Y_s^1)\,\mathrm{d}X_s = \left(\int_0^t \int_0^s f(\,\mathrm{d}X_u) f(\,\mathrm{d}X_s) + \int_0^t f(\,\mathrm{d}X_s) + I_e\right)(y),\\
&\vdots\\
Y_t^n &= y + \int_0^t f(Y_s^{n-1})\,\mathrm{d}X_s = \left(\sum_{k=1}^n \int_{\Delta^k_{[0,t]}} f(\,\mathrm{d}X_{t_1})\dots f(\,\mathrm{d}X_{t_k}) + I_e\right)(y).
\end{aligned}$$

Defining the tensor map $f^{\otimes n}(\,\mathrm{d}X_{t_1}\otimes\cdots\otimes\,\mathrm{d}X_{t_n}) \coloneqq f(\,\mathrm{d}X_{t_1})\dots f(\,\mathrm{d}X_{t_n})$ and extending it by linearity then yields

$$Y_t^n = y + \int_0^t f(Y_s^{n-1})\,\mathrm{d}X_s = \left(\sum_{k=1}^n f^{\otimes k}\left(\int_{\Delta^k_{[0,t]}} \mathrm{d}X_{t_1}\otimes\ldots\otimes\,\mathrm{d}X_{t_k}\right) + I_e\right)(y),$$

suggesting again that the solution Y_t is completely determined by the iterated integrals of the driving signal. We elaborate on the tensor notation below.

Remark 1.6: It is important to observe that iterated integrals are non-commutative. For instance, let us consider the path $X : [0,1] \to \mathbb{R}^2$ given by $X_t = (t, t^2)$. Observe that

$$\int_0^1 \int_0^s \mathrm{d}X_u^1\,\mathrm{d}X_s^2 = \frac{2}{3}, \quad \text{whereas} \quad \int_0^1 \int_0^s \mathrm{d}X_u^2\,\mathrm{d}X_s^1 = \frac{1}{3}.$$

Hence, the set of double-iterated integrals against X is in bijection with the set of words of length 2 over the alphabet $\{1, 2\}$.

Recognising the prevalence of iterated integrals, one may then consider collecting all the iterated integrals into a single object. With this objective in mind, we observe that the n-fold iterated integral can be seen as an element of $(\mathbb{R}^d)^{\otimes n}$. Indeed, owing to the non-commutativity of iterated integrals, there is a clear correspondence between the set of n-fold iterated integrals and elements in $(\mathbb{R}^d)^{\otimes n}$ of the form

$$\sum_{i_1,\dots,i_n=1}^d \left(\int_{\Delta^n_{[0,t]}} \mathrm{d}X_{t_1}^{i_1}\dots\,\mathrm{d}X_{t_n}^{i_n}\right)(e_{i_1}\otimes\ldots\otimes e_{i_n}). \tag{1.10}$$

To alleviate notation, henceforth, we denote (1.10) using the tensor notation

$$\int_{\Delta^n_{[0,t]}} \mathrm{d}X_{t_1}\otimes\ldots\otimes\,\mathrm{d}X_{t_n},$$

which, as previously discussed, has the advantage of suppressing numerous indices. Consequently, it becomes apparent that the natural space to represent the sequence of iterated integrals is the extended tensor algebra introduced in Section 1.1.3. Given that paths of bounded variation lead to the well-established Riemann-Stieltjes integral, we then arrive at the following definition.

Definition 1.13: Let $X \in C^{1\text{-var}}([0,T],\mathbb{R}^d)$ be a continuous path of bounded variation. The signature $S(X)_{s,t}$ of X is defined, for all $0 \le s < t \le T$, by

$$S(X)_{s,t} := \left(1, \int_s^t \mathrm{d}X_{t_1}, \int_{\Delta^2_{[s,t]}} \mathrm{d}X_{t_1} \otimes \mathrm{d}X_{t_2}, \dots, \int_{\Delta^n_{[s,t]}} \mathrm{d}X_{t_1} \otimes \dots \otimes \mathrm{d}X_{t_n}, \dots\right) \in T((\mathbb{R}^d)).$$

Additionally, we denote by $S_N(X)_{s,t}$ the projection $\pi_{\le N}\big(S(X)_{s,t}\big)$ onto $T^N(\mathbb{R}^d)$, and refer to it as the level-N truncated signature. Lastly, the n-th component of $S(X)_{s,t}$, i.e. $\pi_n\big(S(X)_{s,t}\big)$, is denoted by $S(X)^{(n)}_{s,t}$.

We observe that the signature can be interpreted as a mapping over Δ^2_T, or, if we fix the lower bound s of the integral, as a path over $[s,T]$. From now onwards, we agree that $S(X)_t$ denotes the path $t \mapsto S(X)_{0,t}$, and $S(X)$ denotes the evaluated signature $S(X)_{0,T} \in T((\mathbb{R}^d))$.

Remark 1.7: As per the notation in Section 1.1.3, it follows from the definition of signature that $\int_0^T \langle e_w, S(X)_t\rangle \,\mathrm{d}X^i_t = \langle e_w \otimes e_i, S(X)\rangle$, where w represents a word in $\mathcal{W}(A_d)$. Furthermore, $e_w \otimes e_i$ can be denoted as $e_{w'}$, where $w' = (w,i)$, and $\langle e_\varnothing, S(X)\rangle = 1$ for all paths. Lastly, it is worth noting that linear functionals $L : T((\mathbb{R}^d)) \to \mathbb{R}$ of the signature can be expressed as $L(S(X)_t) = \sum_{0\le|w|\le N} \alpha_w \langle e_w, S(X)_t\rangle$, where $N \in \mathbb{N}$ and $\alpha_w \in \mathbb{R}$ represent coefficients.

1.2.2. Basic Properties of Signatures

We explore the basic properties of path signatures. We begin by noting that the extended tensor algebra fully encapsulates the algebraic properties of iterated integrals.

Recall that one basic property of the integral is additivity, i.e. $\int_s^t = \int_s^u + \int_u^t$, for all $s < u < t$. For a double iterated integral, however, we observe that

$$\begin{aligned}
\int_s^t \int_s^v \mathrm{d}X^i_r \,\mathrm{d}X^j_v &= \int_s^t (X^i_v - X^i_s)\,\mathrm{d}X^j_v = \int_s^u (X^i_v - X^i_s)\,\mathrm{d}X^j_v + \int_u^t (X^i_v - X^i_s)\,\mathrm{d}X^j_v \\
&= \int_s^u (X^i_v - X^i_s)\,\mathrm{d}X^j_v + \int_u^t (X^i_v - X^i_u)\,\mathrm{d}X^j_v + \int_u^t (X^i_u - X^i_s)\,\mathrm{d}X^j_v \\
&= \int_s^u \int_s^v \mathrm{d}X^i_r \,\mathrm{d}X^j_v + \int_u^t \int_u^v \mathrm{d}X^i_r \,\mathrm{d}X^j_v + \int_s^u \mathrm{d}X^i_r \int_u^t \mathrm{d}X^j_v.
\end{aligned}$$

Hence, we conclude that, for all $s < u < t$,

$$S(X)^{(2)}_{s,t} = S(X)^{(2)}_{s,u} + S(X)^{(2)}_{u,t} + S(X)^{(1)}_{s,u} \otimes S(X)^{(1)}_{u,t},$$

illustrating that iterated integrals lack additivity. That said, iterated integrals exhibit a multiplicative relation, which resembles the product $\otimes$ of the extended tensor algebra. The following result formalises this idea.

Proposition 1.13 (Chen)**:** Let $X \in C^{1\text{-var}}([0,T],\mathbb{R}^d)$ be a continuous path of bounded variation and $S(X)$ its signature. Then, for all $0 \leq s < u < t \leq T$, we have

$$S(X)_{s,t} = S(X)_{s,u} \otimes S(X)_{u,t}. \tag{1.11}$$

Proof. We need to show that, for all $n \in \mathbb{N}$, we have

$$S(X)^{(n)}_{s,t} = \sum_{i=0}^{n} S(X)^{(i)}_{s,u} \otimes S(X)^{(n-i)}_{u,t}.$$

The proof is done by induction. For $n = 1$, the statement trivially follows from the usual additivity of the integral. Assuming the identity above holds for some arbitrary $n \geq 2$, we note that

$$\begin{aligned}
S(X)^{(n+1)}_{s,t} &= \int_s^t S(X)^{(n)}_{s,v} \otimes \mathrm{d}X_v = \int_s^u S(X)^{(n)}_{s,v} \otimes \mathrm{d}X_v + \int_u^t S(X)^{(n)}_{s,v} \otimes \mathrm{d}X_v \\
&= S(X)^{(n+1)}_{s,u} + \sum_{i=0}^{n} \int_u^t S(X)^{(i)}_{s,u} \otimes S(X)^{(n-i)}_{u,v} \otimes \mathrm{d}X_v \\
&= S(X)^{(n+1)}_{s,u} + \sum_{i=0}^{n} S(X)^{(i)}_{s,u} \otimes \int_u^t S(X)^{(n-i)}_{u,v} \otimes \mathrm{d}X_v \\
&= S(X)^{(n+1)}_{s,u} + \sum_{i=0}^{n} S(X)^{(i)}_{s,u} \otimes S(X)^{(n-i+1)}_{u,t} \\
&= \sum_{i=0}^{n+1} S(X)^{(i)}_{s,u} \otimes S(X)^{(n+1-i)}_{u,t},
\end{aligned}$$

thus proving the claim. □

Remark 1.8: The identity presented in Proposition 1.13 is commonly known as *Chen's identity*, and it will be employed consistently throughout this work. Signatures initially emerged within the realm of Cohomology and trace back to the 1950s in the seminal work [17] of Chen.

Definition 1.14: A map $\mathbb{X} : \Delta^2_T \to T((\mathbb{R}^d))$ satisfying Chen's identity (1.11) is said to be a multiplicative functional.

We note that Proposition 1.13 is equivalent to showing that the signature of the concatenation of two paths is equal to the tensor product of the respective signatures of the original paths. More precisely, given $X \in C^{1\text{-var}}([0,T],\mathbb{R}^d)$ and $Y \in C^{1\text{-var}}([T,2T],\mathbb{R}^d)$ we define the concatenation $X * Y$ of X with Y as the path

$$(X * Y)_t := \begin{cases} X_t, & t \in [0,T] \\ Y_t - Y_0 + X_T, & t \in [T,2T] \end{cases}.$$

Naturally, $(X * Y) \in C^{1\text{-var}}([0,2T],\mathbb{R}^d)$ and we obtain the following result.

Corollary 1.1 ([37], Theorem 7.11)**:** Given $X \in C^{1\text{-var}}([0,T],\mathbb{R}^d)$ and $Y \in C^{1\text{-var}}([T,2T],\mathbb{R}^d)$,

$$S(X * Y)_{0,2T} = S(X)_{0,T} \otimes S(Y)_{T,2T}.$$

Next, we demonstrate that polynomials of linear functionals of the signature can be represented as a new linear functional of the signature through the use of the shuffle product, thus elucidating its significance.

Proposition 1.14: Let $X : [0,T] \to \mathbb{R}^d$ be a continuous path of bounded variation. Consider the words $w_1 = i_1 \dots i_m$ and $w_2 = j_1 \dots j_n$. Then,

$$\langle e_{w_1}, S(X)\rangle\langle e_{w_2}, S(X)\rangle = \langle e_{w_1} \sqcup\!\sqcup\, e_{w_2}, S(X)\rangle.$$

Proof. This result follows by partitioning the domain of integration. Using the notation of Section 1.1.3, we have that

$$\begin{aligned} \langle e_{w_1}, S(X)\rangle\langle e_{w_2}, S(X)\rangle &= \int_{\Delta_T^m} \mathrm{d}X_{t_1}^{i_1} \dots \mathrm{d}X_{t_m}^{i_m} \int_{\Delta_T^n} \mathrm{d}X_{u_1}^{j_1} \dots \mathrm{d}X_{u_n}^{j_n} \\ &= \sum_{\sigma \in S(m,n)} \int_{\Delta_T^{m+n}} \mathrm{d}X_{v_1}^{r_{\sigma(1)}} \dots \mathrm{d}X_{v_{m+n}}^{r_{\sigma(m+n)}} = \langle e_{w_1} \sqcup\!\sqcup\, e_{w_2}, S(X)\rangle. \end{aligned} \tag{1.12}$$

To clarify, note that the product of integrals in (1.12) can be written as a single integral over $A = \{\overline{v} \in \mathbb{R}^{n+m} : 0 < \overline{v}_1 < \dots < \overline{v}_m < T, 0 < \overline{v}_{m+1} < \dots < \overline{v}_{n+m} < T\}$ and

$$A = \bigcup_{\sigma \in S(m,n)} \left\{\overline{v} \in \mathbb{R}^{m+n} : 0 < \overline{v}_{\sigma(1)} < \dots < \overline{v}_{\sigma(m+n)} < T\right\}.$$

The result then follows by setting $v_i = \overline{v}_{\sigma(i)}$ and $r = (w_1, w_2)$. □

As observed in the previous section, iterated integrals naturally appear in the solution of certain controlled differential equations. The next result demonstrates that the level-N truncated signature satisfies a differential equation on $T^N(\mathbb{R}^d) \cong \mathbb{R}^{1+d+\dots+d^N}$ controlled by a bounded variation path $X : [0,T] \to \mathbb{R}^d$. Furthermore, we deduce from this result the time-reversal property of signatures and their continuity.

Proposition 1.15 ([37], Proposition 7.8)**:** Let $X \in C^{1\text{-var}}([0,T],\mathbb{R}^d)$ be a continuous path of bounded variation. Then, for a fixed $s \in [0,T)$,

$$\begin{cases} \mathrm{d}S_N(X)_{s,t} = S_N(X)_{s,t} \otimes \mathrm{d}X_t \\ S_N(X)_{s,s} = \mathbf{1} \end{cases},$$

where $\mathbf{1}$ denotes the identity element in $T^N(\mathbb{R}^d)$.

Proof. The fact that $S_N(X)_{s,t}$ is a solution of the differential equation above follows almost by definition of signature. Observe that, for a fixed level $n \le N$, we have

$$\begin{aligned} S_N(X)^{(n)}_{s,t} &= \int_{\Delta^n_{[s,t]}} \mathrm{d}X_{t_1} \otimes \cdots \otimes \mathrm{d}X_{t_n} \\ &= \int_s^t \left(\int_{\Delta^{n-1}_{[s,t_n]}} \mathrm{d}X_{t_1} \otimes \cdots \otimes \mathrm{d}X_{t_{n-1}} \right) \otimes \mathrm{d}X_{t_n} = \int_s^t S_N(X)^{(n-1)}_{s,r} \otimes \mathrm{d}X_r \end{aligned}$$

Therefore,

$$S_N(X)_{s,t} = \mathbf{1} + \int_s^t S_N(X)_{s,r} \otimes \mathrm{d}X_r,$$

by considering multiplication over the truncated tensor algebra $T^N(\mathbb{R}^d)$. □

Remark 1.9: Observe that we may define the linear vector fields $f_i : T^N(\mathbb{R}^d) \to T^N(\mathbb{R}^d)$ given by $\mathbf{a} \mapsto \mathbf{a} \otimes e_i$, for $i \in \{1,\ldots,d\}$, and rewrite the differential equation in Proposition 1.15 in the more familiar form

$$\mathrm{d}S_N(X)_{s,t} = \sum_{i=1}^d f_i\big(S_N(X)_{s,t}\big)\,\mathrm{d}X^i_t.$$

This way, $f = (f_1,\ldots,f_d)$ can be viewed as a map

$$T^N(\mathbb{R}^d) \cong \mathbb{R}^E \ni \mathbf{a} \mapsto \left(x = (x^1,\ldots,x^d) \mapsto \sum_{i=1}^d f_i(\mathbf{a})x^i \right) \in L(\mathbb{R}^d,\mathbb{R}^E),$$

where $E = 1 + d + \ldots + d^N$ (see [37], Remark 7.9). Note that the vector fields in f satisfy the assumptions of Theorem 1.2.

Remark 1.9 together with Theorem 1.2 lead to the following result.

Corollary 1.2: The map $S_N : C^{1\text{-var}}([0,T],\mathbb{R}^d) \to T^N(\mathbb{R}^d)$ is continuous for all $N \ge 0$. Consequently, if $(X_n)_{n\ge 1} \subset C^{1\text{-var}}([0,T],\mathbb{R}^d)$ converges to some $X \in C^{1\text{-var}}([0,T],\mathbb{R}^d)$, then

$$\lim_{n\to\infty} S(X_n) = S(X).$$

Proposition 1.16 ([37], Proposition 7.10)**:** Let $X : [a,b] \to \mathbb{R}^d$ be a continuous path of bounded variation. Consider a continuous, non-decreasing surjection $\phi : [0,T] \to [a,b]$, and write $X_t^\phi \coloneqq X_{\phi(t)}$ for the reparametrization of X. Then, for all $s < t$ in $[0,T]$,

$$S(X)_{\phi(s),\phi(t)} = S(X^\phi)_{s,t}.$$

The same is true for the level-N truncated signature S_N.

Proof. This follows directly from the invariance under reparametrization of the Riemann-Stieltjes integral (Remark 1.4). □

Proposition 1.17 ([55], Proposition 2.14)**:** Let $X : [0,T] \to \mathbb{R}^d$ be continuous path of bounded variation. Consider the time-reversed path $\overleftarrow{X}_t \coloneqq X_{T-t}$, for $t \in [0,T]$. Then,

$$S(X) \otimes S(\overleftarrow{X}) = S(\overleftarrow{X}) \otimes S(X) = \mathbf{1}.$$

Proof. Parameterise $\overleftarrow{X}$ allowing for the concatenation $Z \coloneqq X * \overleftarrow{X} : [0,2T] \to \mathbb{R}^d$. Consider a collection of vector fields $f : \mathbb{R}^e \to L(\mathbb{R}^d, \mathbb{R}^e)$ satisfying the assumptions of Theorem 1.2. Then, by Remark 1.4, it is equivalent for a path $Y : [0,T] \to \mathbb{R}^e$ to satisfy

$$\forall t \in [0,T], \quad \mathrm{d}Y_t = f(Y_t)\,\mathrm{d}X_t, \quad Y_0 = \xi, Y_T = \eta,$$

or to satisfy

$$\forall t \in [0,T], \quad \mathrm{d}\overleftarrow{Y}_t = f(\overleftarrow{Y}_t)\,\mathrm{d}\overleftarrow{X}_t, \quad \overleftarrow{Y}_0 = \eta, \overleftarrow{Y}_T = \xi.$$

Consequently, the solution to

$$\mathrm{d}Y_t = f(Y_t)\,\mathrm{d}Z_t, \quad Y_0 = \xi \tag{1.13}$$

satisfies $Y_{2T} = \xi$. Now, take f to be given by the vector fields of Proposition 1.15 and observe that $\pi_N(\mathbf{1})$ is a solution to (1.13) for all $N \geq 0$. However, $S_N(Z)$ also satisfies (1.13). Therefore, by uniqueness of solution, $S_N(Z) = \pi_N(\mathbf{1})$ for all $N \geq 0$, i.e. $S(Z) = S(X * \overleftarrow{X}) = \mathbf{1}$. □

We conclude this section by establishing what is commonly known as the *factorial decay property* of signatures. In essence, we demonstrate that the tensor norms of the components of the signature decay factorially. This formalises the idea that the majority of information within the signature of a path is concentrated in the initial components.

Proposition 1.18 ([55], Proposition 2.2)**:** Let $X : [0,T] \to \mathbb{R}^d$ be a continuous path of bounded variation. Then, for all $n \geq 0$,

$$|S(X)^{(n)}| \leq \frac{1}{n!}|X|^n_{1\text{-var};[0,T]} \quad \text{and} \quad |S(X)|_{\tilde{T}((\mathbb{R}^d))} \leq \exp\left(|X|_{1\text{-var};[0,T]}\right) < \infty.$$

Proof. By Proposition 1.10, we may assume that X is Lipschitz continuous and hence almost everywhere differentiable with constant speed $|\dot{X}_t|$. Moreover, we can choose the reparametrization in such a way that $|\dot{X}_t| = 1$ ([37], Remark 1.39). This way, by the compatibility of tensor norms,

$$\left| \int_{\Delta_T^n} \mathrm{d}X_{t_1} \otimes \cdots \otimes \mathrm{d}X_{t_n} \right| = \left| \int_{\Delta_T^n} \dot{X}_{t_1} \otimes \cdots \otimes \dot{X}_{t_n} \, \mathrm{d}t_1 \ldots \mathrm{d}t_n \right| \leq \int_{\Delta_t^n} \mathrm{d}t_1 \ldots \mathrm{d}t_n = \frac{T^n}{n!}.$$

Since $|X|_{1\text{-var};[0,T]} = T$ the result follows. □

1.2.3. Free Nilpotent Groups and the Carnot-Carathéodory Norm

In Section 1.1.3, we introduced several subsets of the extended tensor algebra $T((\mathbb{R}^d))$. Specifically, we defined the subset $T_1((\mathbb{R}^d))$ consisting of all elements $\mathbf{a} \in T((\mathbb{R}^d))$ with $\mathbf{a}^0 = 1$ and finite Euclidean norm. By restricting the operation $\otimes$ to $T_1((\mathbb{R}^d))$, we obtained a Lie group (Proposition 1.12).

In this section, we narrow our focus further to a subgroup of the truncated tensor algebra $T^N(\mathbb{R}^d)$. This subgroup is commonly known as the *free nilpotent group of step N* over $\mathbb{R}^d$, and it will serve as the state-space for the paths defined in Section 1.3. Additionally, we define a norm over this subgroup, referred to as the Carnot-Carathéodory norm.

Definition 1.15: Let $\mathbf{1} + \mathfrak{t}^N := \{\mathbf{a} \in T^N(\mathbb{R}^d) : \mathbf{a}^0 = 1\}$ denote the Lie group (equipped with $\otimes$) contained in $T^N(\mathbb{R}^d)$. We denote by $G^N(\mathbb{R}^d)$ the set of all level-N truncated signatures of continuous paths of bounded variation, i.e.

$$G^N(\mathbb{R}^d) := \left\{ S_N(X)_{0,1} : X \in C^{1\text{-var}}([0,1], \mathbb{R}^d) \right\}.$$

Remark 1.10: The subset $G^N(\mathbb{R}^d)$ can be shown to be a closed sub-Lie group of $(\mathbf{1} + \mathfrak{t}^N, \otimes)$ ([37], Theorem 7.30). That said, proving this fact requires a significant amount of algebraic machinery, potentially leading to a substantial digression into Lie theory. For our predominantly analytical purposes, it suffices to acknowledge that $G^N(\mathbb{R}^d)$ has a group structure. A formal exploration of the algebraic intricacies behind rough paths, including an introduction to linear Lie groups, is deferred to a future work.

We proceed to define a norm over $G^N(\mathbb{R}^d)$, which, in turn, will lead to a metric, thereby enabling the definition of path spaces with values in $G^N(\mathbb{R}^d)$. Note that in Definition 1.15 we consider paths defined on $[0,1]$. We can, nevertheless, consider some other interval, owing to the invariance of signatures under reparametrization. In what follows, we denote the length of $X \in C^{1\text{-var}}([0,1], \mathbb{R}^d)$ by $\int_0^1 |\,\mathrm{d}X|$, where $\mathbb{R}^d$ is equipped with the usual Euclidean distance.

Definition 1.16: For every $\mathbf{g} \in G^N(\mathbb{R}^d)$, we define the Carnot-Carathéodory norm as

$$\|\mathbf{g}\|_{cc} := \inf\left\{\int_0^1 |\,\mathrm{d}X| : X \in C^{1\text{-var}}([0,1],\mathbb{R}^d) \text{ and } S_N(X)_{0,1} = \mathbf{g}\right\}.$$

The next result tells us that the Carnot-Carathéodory norm is finite and, most importantly, achieved at some minimising path.

Theorem 1.3 ([37], Theorem 7.32)**:** For every $\mathbf{g} \in G^N(\mathbb{R}^d)$, the Carnot-Carathéodory norm $\|\mathbf{g}\|_{cc}$ is finite and achieved. Precisely, there exists a minimising path X^* such that

$$\|\mathbf{g}\|_{cc} = \int_0^1 |\,\mathrm{d}X^*| \quad \text{and} \quad S_N(X^*)_{0,1} = \mathbf{g}.$$

Moreover, this minimising path can be reparametrized into a Lipschitz path of constant speed, i.e. $|\dot{X}^*(t)| = C > 0$, for a.e. $t \in [0,1]$.

Proof. Consider $\mathbf{g} \in G^N(\mathbb{R}^d)$. By Definition 1.15, it follows that the infimum is taken over a non-empty set. Hence, $\|\mathbf{g}\|_{cc} < \infty$. Moreover, by definition of infimum there exists a sequence $(X^n)_{n\geq 1}$ of continuous bounded variation paths with signature equal to $\mathbf{g}$, and whose lengths converge to $\|\mathbf{g}\|_{cc}$. Proposition 1.10, in turn, allows us to assume (by reparametrization) that each X^n has a.e. constant speed c_n, i.e.

$$|\dot{X}^n| \equiv |X^n|_{1\text{-Höl};[0,1]} = c_n \quad \text{and} \quad c_n \downarrow \|\mathbf{g}\|_{cc}.$$

Since the sequence of lengths is decreasing, we have that

$$\sup_n |X^n|_{1\text{-Höl};[0,1]} = \sup_n c_n < \infty,$$

and, by Arzelà-Ascoli (Theorem 1.1), there exists a subsequence $(X^{n_k})_{k\geq 1}$ such that X^{n_k} converges to some continuous path X^* uniformly. Additionally, by Proposition 1.8,

$$|X^*|_{1\text{-Höl};[0,1]} \leq \liminf_k |X^{n_k}|_{1\text{-Höl};[0,1]} < \infty,$$

showing that X^* itself is 1-Hölder continuous, and hence, absolutely continuous with

$$\int_0^1 |\,\mathrm{d}X^*| = \int_0^1 |\dot{X}^*_t|\,\mathrm{d}t.$$

Now, by continuity of the signature map (Corollary 1.2), $\mathbf{g} \equiv S_N(X^n) \to S_N(X^*)$, showing that $S_N(X^*) = \mathbf{g}$. Finally, it remains to see that

$$\|\mathbf{g}\|_{cc} = \int_0^1 |\dot{X}^*_t|\,\mathrm{d}t.$$

The inequality $\leq$ is trivial, since $c_n \downarrow \|\mathbf{g}\|_{cc}$. The opposite direction is obtained by observing that

$$\int_0^1 |\dot{X}_t^*| \, \mathrm{d}t = |X^*|_{1\text{-Höl};[0,1]} \leq \liminf_k c_{n_k} = \|\mathbf{g}\|_{cc}.$$

This finishes the proof. □

Remark 1.11: This result is commonly known as "geodesic existence." This terminology stems from the observation that $G^N(\mathbb{R}^d)$ transforms into a geodesic space under the metric induced by the Carnot-Carathéodory norm $\|\cdot\|_{cc}$. Intuitively, a geodesic between two points can be conceptualised as the shortest path joining these points, and a geodesic space is characterised by the existence of a geodesic between any pair of points within it. For more details refer to Sections 5.2 and 7.5.2 in [37].

We conclude this section by comparing the Carnot-Carathéodory norm with the norm $|\cdot|_{T^N(\mathbb{R}^d)}$ introduced in Section 1.1.3, specifically when restricted to $G^N(\mathbb{R}^d)$. Defining path spaces with the state-space given by $G^N(\mathbb{R}^d)$, it becomes convenient to establish estimates between these two norms to facilitate comparisons between the topologies induced by the respective different metric structures.

Given $\mathbf{g}, \mathbf{h} \in G^N(\mathbb{R}^d)$ let $d_{cc}(\mathbf{g}, \mathbf{h})$ denote the metric defined by $\|\mathbf{g}^{-1} \otimes \mathbf{h}\|_{cc}$, i.e. d_{cc} denotes the metric induced by $\|\cdot\|_{cc}$. It is straightforward to see that d_{cc} is indeed a genuine metric ([37], Proposition 7.40). Denote by $\rho(\mathbf{g}, \mathbf{h})$ the metric induced by the norm $|\cdot|_{T^N(\mathbb{R}^d)}$ such that $\rho(\mathbf{g}, \mathbf{h}) = \max_{i=1,\dots,N} |\pi_i(\mathbf{g}) - \pi_i(\mathbf{h})|$. In addition, for $\lambda \in \mathbb{R}$, define the dilation map $\delta_\lambda : T^N(\mathbb{R}^d) \to T^N(\mathbb{R}^d)$ as $\pi_k(\delta_\lambda(\mathbf{g})) = \lambda^k \pi_k(\mathbf{g})$.

Lemma 1.1: Let $\mathbf{g} \in G^N(\mathbb{R}^d)$. The following statements hold:

1. For all $\lambda \in \mathbb{R}$, we have that $\|\delta_\lambda \mathbf{g}\|_{cc} = |\lambda| \; \|\mathbf{g}\|_{cc}$. A norm on $T^N(\mathbb{R}^d)$ that scales with the dilation operator is said to be homogenous.
2. The Carnot-Carathéodory norm is symmetric, i.e. $\|\mathbf{g}^{-1}\|_{cc} = \|\mathbf{g}\|_{cc}$.
3. All homogenous norms on $G^N(\mathbb{R}^d)$ are equivalent. Precisely, given two homogenous norms $\|\cdot\|_1$ and $\|\cdot\|_2$, there exists a constant $C \geq 1$ such that

$$\frac{1}{C} \; \|\mathbf{g}\|_1 \leq \|\mathbf{g}\|_2 \leq C \; \|\mathbf{g}\|_1. \tag{1.14}$$

Proof. This lemma corresponds to Proposition 7.40 and Theorem 7.44 in [37]. To prove 1. consider some $\lambda \neq 0$, otherwise the claim is trivial. Using the notation from Theorem 1.3, we observe that $\lambda X_{\mathbf{g}}^*$ satisfies $S_N(\lambda X_{\mathbf{g}}^*) = \delta_\lambda \mathbf{g}$. Hence, $\|\delta_\lambda \mathbf{g}\|_{cc} \leq \int_0^1 |\, \mathrm{d}\lambda X_{\mathbf{g}}^*| = |\lambda| \; \|\mathbf{g}\|_{cc}$. To prove the other direction, repeat the reasoning replacing λ by $1/\lambda$, and $\mathbf{g}$ by $\delta_\lambda \mathbf{g}$.

Regarding the second claim, by Proposition 1.17 we have that $S_N(\overleftarrow{X^*_{\mathbf{g}}}) = \mathbf{g}^{-1}$. Hence,

$$\|\mathbf{g}^{-1}\|_{cc} \leq \int_0^1 |\,\mathrm{d}\overleftarrow{X^*}_{\mathbf{g}}| = \int_0^1 |\,\mathrm{d}X^*_{\mathbf{g}}| = \|\mathbf{g}\|_{cc}.$$

The opposite inequality follows by considering $\mathbf{g}^{-1}$ instead of $\mathbf{g}$. Finally, to prove the equivalence of homogenous norms, consider $\|\cdot\|_1$ to be $|||\mathbf{g}||| := \max_{i=1,\dots,N} |\pi_i(\mathbf{g})|^{1/i}$, which is evidently homogenous.

Consider the compact set $B = \{\mathbf{g} \in G^N(\mathbb{R}^d) : |||\mathbf{g}||| = 1\}$. By continuity of norms, $\|\cdot\|_2$ attains a minimum and a maximum in B. Specifically, there exist $m, M \in \mathbb{R}_0^+$ such that $m \leq \|\mathbf{g}\|_2 \leq M$. Since (1.14) is clearly satisfied by $\mathbf{g} = \mathbf{1}$, we assume that $\mathbf{g} \neq 1$. Now, define $\lambda = 1/|||\mathbf{g}|||$ so that $|||\delta_\lambda \mathbf{g}||| = 1$, and observe that $m \leq \|\delta_\lambda \mathbf{g}\|_2 \leq M$. Using the homogeneity of $\|\cdot\|_2$, we obtain $m \leq \|\mathbf{g}\|_2/|||\mathbf{g}||| \leq M$ and the result follows. □

Proposition 1.19 ([37], Proposition 7.45)**:** Let $|||\cdot|||$ denote any homogenous norm on $G^N(\mathbb{R}^d)$. Then, there exists a constant $C > 0$ such that, for all $\mathbf{g} \in G^N(\mathbb{R}^d)$,

$$\frac{1}{C}\ \min\left\{|||\mathbf{g}|||, |||\mathbf{g}|||^N\right\} \leq |\mathbf{g} - \mathbf{1}|_{T^N(\mathbb{R}^d)} \leq C\ \max\left\{|||\mathbf{g}|||, |||\mathbf{g}|||^N\right\},$$

$$\frac{1}{C}\ \min\left\{|\mathbf{g} - \mathbf{1}|_{T^N(\mathbb{R}^d)}, |\mathbf{g} - \mathbf{1}|^{1/N}_{T^N(\mathbb{R}^d)}\right\} \leq |||\mathbf{g}||| \leq C\ \max\left\{|\mathbf{g} - \mathbf{1}|_{T^N(\mathbb{R}^d)}, |\mathbf{g} - \mathbf{1}|^{1/N}_{T^N(\mathbb{R}^d)}\right\}.$$

Proof. Note that $|\mathbf{g} - \mathbf{1}|_{T^N(\mathbb{R}^d)} = \rho(\mathbf{g}, \mathbf{1}) = \max_{i=1,\dots,N} |\pi_i(\mathbf{g})|$. By equivalence of homogenous norms (Lemma 1.1), it suffices to consider $|||\mathbf{g}||| = \max_{i=1,\dots,N} |\pi_i(\mathbf{g})|^{1/i}$, making it apparent that

$$|||\mathbf{g}||| \leq \max\left\{|\mathbf{g} - \mathbf{1}|_{T^N(\mathbb{R}^d)}, |\mathbf{g} - \mathbf{1}|^{1/N}_{T^N(\mathbb{R}^d)}\right\}.$$

This implies that $|||\mathbf{g}|||^N \leq \max\left\{|\mathbf{g} - \mathbf{1}|^N_{T^N(\mathbb{R}^d)}, |\mathbf{g} - \mathbf{1}|_{T^N(\mathbb{R}^d)}\right\}$, and, analysing case by case, shows that

$$\min\left\{|||\mathbf{g}|||, |||\mathbf{g}|||^N\right\} \leq |\mathbf{g} - \mathbf{1}|_{T^N(\mathbb{R}^d)}.$$

Finally, it is clear that $|\mathbf{g} - \mathbf{1}|_{T^N(\mathbb{R}^d)} = \max_{i=1,\dots,N} |\pi_i(\mathbf{g})| \leq \max\left\{|||\mathbf{g}|||, |||\mathbf{g}|||^N\right\}$ □

Proposition 1.19, in particular, provides estimates between the Carnot-Carathéodory norm $\|\cdot\|_{cc}$ and the non-homogenous norm $|\cdot|_{T^N(\mathbb{R}^d)}$.

We now establish estimates between the metrics induced by these norms. However, before proceeding, we state a simple lemma whose proof we omit as it only involves basic algebra. Recall that, as per the notation in Section 1.1.3, $\mathbf{g}^i$ denotes $\pi_i(\mathbf{g})$.

Lemma 1.2 ([37], Lemma 7.48)**:** Consider $\mathbf{g}, \mathbf{h} \in G^N(\mathbb{R}^d)$. The following identities hold in $(\mathbb{R}^d)^{\otimes n}$ for $n = 1, \dots, N$:

$$(\mathbf{g}^{-1} \otimes \mathbf{h})^n = \sum_{i=1}^n (\mathbf{g}^{-1})^{n-i} \otimes (\mathbf{h}^i - \mathbf{g}^i) \quad \text{and} \quad \mathbf{h}^n - \mathbf{g}^n = \sum_{i=1}^n \mathbf{g}^{n-i} \otimes (\mathbf{g}^{-1} \otimes \mathbf{h})^i. \tag{1.15}$$

Proposition 1.20 ([37], Proposition 7.49)**:** Consider $\mathbf{g}, \mathbf{h} \in G^N(\mathbb{R}^d)$. There exists a constant $C \equiv C(N) > 0$ such that

$$d_{cc}(\mathbf{g}, \mathbf{h}) \le C \max\left\{\rho(\mathbf{g}, \mathbf{h}),\ \rho(\mathbf{g}, \mathbf{h})^{1/N} \max\left\{1, \|\mathbf{g}\|_{cc}^{1-\frac{1}{N}}\right\}\right\}, \text{ and} \tag{1.16}$$

$$\rho(\mathbf{g}, \mathbf{h}) \le C \max\left\{d_{cc}(\mathbf{g}, \mathbf{h}) \max\left\{1, \|\mathbf{g}\|_{cc}^{N-1}\right\},\ d_{cc}(\mathbf{g}, \mathbf{h})^N\right\}. \tag{1.17}$$

Particularly, we have that $\mathrm{Id} : \big(G^N(\mathbb{R}^d), d_{cc}\big) \leftrightarrow \big(G^N(\mathbb{R}^d), \rho\big)$ is Lipschitz on bounded sets in the $(\rightarrow)$ direction, and $1/N$-Hölder continuous on bounded sets in the $(\leftarrow)$ direction.

Proof. By (1.15) and the fact that all norms in $G^N(\mathbb{R}^d)$ are equivalent, we have that

$$\begin{aligned}
\left|(\mathbf{g}^{-1} \otimes \mathbf{h})^k\right| &\le \sum_{i=1}^k \left|(\mathbf{g}^{-1})^{k-i}\right| \left|\mathbf{h}^i - \mathbf{g}^i\right| \le \rho(\mathbf{g}, \mathbf{h}) \sum_{i=1}^k \left|(\mathbf{g}^{-1})^{k-i}\right| \\
&\le c_1\ \rho(\mathbf{g}, \mathbf{h}) \max\left\{1, \left|\pi_{\le k-1}(\mathbf{g}^{-1}) - \mathbf{1}\right|_{T^{k-1}(\mathbb{R}^d)}\right\} \le c_2\ \rho(\mathbf{g}, \mathbf{h}) \max\left\{1, \|\mathbf{g}\|_{cc}^{k-1}\right\},
\end{aligned}$$

where we used the symmetry in Lemma 1.1 for the final inequality. Therefore,

$$\begin{aligned}
\max_{k=1,\dots,N} \left|(\mathbf{g}^{-1} \otimes \mathbf{h})^k\right|^{1/k} &\le c_3 \max_{k=1,\dots,N} \left[\rho(\mathbf{g}, \mathbf{h})^{1/k} \max\left\{1, \|\mathbf{g}\|_{cc}^{1-\frac{1}{k}}\right\}\right] \\
&\le c_4 \max\left\{\rho(\mathbf{g}, \mathbf{h}), \rho(\mathbf{g}, \mathbf{h})^{1/N} \max\left\{1, \|\mathbf{g}\|_{cc}^{1-\frac{1}{N}}\right\}\right\}.
\end{aligned}$$

Conversely, again by (1.15) and the fact that all norms in $G^N(\mathbb{R}^d)$ are equivalent, we have

$$\begin{aligned}
\rho(\mathbf{g}, \mathbf{h}) = \max_{k=1,\dots,N} |\mathbf{g}^k - \mathbf{h}^k| &\le \max_{k=1,\dots,N} \sum_{i=1}^k |\mathbf{g}^{k-i}|\ |(\mathbf{g}^{-1} \otimes \mathbf{h})^i| = \sum_{i=1}^N |\mathbf{g}^{N-i}|\ |(\mathbf{g}^{-1} \otimes \mathbf{h})^i| \\
&\le \sum_{i=1}^N \max\left\{1, |\pi_{\le N-i}(\mathbf{g}) - \mathbf{1}|_{T^{N-i}(\mathbb{R}^d)}\right\} |(\mathbf{g}^{-1} \otimes \mathbf{h})^i| \\
&\le c_5 \sum_{i=1}^N \max\left\{1, \|\mathbf{g}\|_{cc}^{N-i}\right\} \max_{k=1,\dots,i} |(\mathbf{g}^{-1} \otimes \mathbf{h})^k|^{i/k} \\
&\le c_6 \sum_{i=1}^N \max\left\{1, \|\mathbf{g}\|_{cc}^{N-i}\right\} d_{cc}(\mathbf{g}, \mathbf{h})^i \\
&\le c_7 \max\left\{d_{cc}(\mathbf{g}, \mathbf{h}) \max\left\{1, \|\mathbf{g}\|_{cc}^{N-1}\right\}, d_{cc}(\mathbf{g}, \mathbf{h})^N\right\}.
\end{aligned}$$

This concludes the proof. $\square$

1.3. Introduction to Rough Path Theory

Section 1.3 constitutes the most technical and nuanced section of this thesis. Here, we offer a formal introduction to Rough Path theory per se, aiming to define the spaces of rough paths. Looking back on controlled differential equations, this section expounds upon the theory pertaining to the integrators, i.e., the controls. Further discussion of the integrands and equations driven by rough paths is deferred to future opportunities.

As highlighted in the Introduction, our exposition follows an under-explored path leading to the renowned Lyons' Extension theorem. After laying out motivation and intuition in Section 1.3.1, we present the Sewing lemma (Section 1.3.2), a somewhat abstract result fundamental to the theory of rough paths. Subsequently, in Section 1.3.3, building upon [35] and [22], we elaborate on the interplay between additive and multiplicative functionals, essentially presenting a deconstructed version of the original proof of the Lyons' Extension theorem.

Section 1.3.4 then elucidates the two primary properties of Lyons' extension, namely the factorial decay property and continuity. Notably, the proof of the latter appears to be original. Finally, Section 1.3.5 defines, following the style of [37], the spaces of rough paths, concluding with the pertinent observation that the commonly employed topology in the signature map is not initial. In what follows, we write $A = O(f(x))$ if there exists a constant C such that $|A| \leq C|f(x)|$. Additionally, we write $A = o(f(x))$ if the constant C can be made arbitrarily small as $x \to 0$, i.e. $A/f(x) \to 0$ as $x \to 0$.

1.3.1. Rough Integration: Motivation and Intuition

As discussed in the chapter's introduction, one of the primary objectives of Rough Path theory is to develop a pathwise notion of integration that remains applicable to highly irregular driving signals. Specifically, Rough Path theory aims to surmount the regularity threshold of $\alpha + \beta > 1$, or $\alpha > 1/2$ in the case where $\alpha = \beta$, which is necessary for Young integration. Intuitively, highly irregular signals fluctuate so rapidly that simple Riemann sums fail to adequately capture these variations. A key insight of Rough Path theory is the recognition of a certain "lack of information." This section will elucidate this idea further.

Let's begin our examination by considering an infinitely differentiable function $f : \mathbb{R}^d \to \mathbb{R}$, an α-Hölder continuous path $X : [0,1] \to \mathbb{R}^d$ with $\alpha \in (0,1]$, and the integral of $f(X)$ against X ([2], Section 1.4). Precisely, our goal is to give meaning to

$$\int_0^1 f(X_r)\,\mathrm{d}X_r. \tag{1.18}$$

Through a Taylor expansion, for a sufficiently small time interval $[s,t] \subset [0,1]$ and $r \in [s,t]$, we obtain that

$$f(X_r) = f(X_s) + \nabla f(X_s)(X_r - X_s) + \dots$$

Integrating with respect to X then yields

$$\int_s^t f(X_r)\,\mathrm{d}X_r = f(X_s)(X_t - X_s) + \nabla f(X_s)\int_s^t (X_r - X_s)\otimes\,\mathrm{d}X_r + \dots, \tag{1.19}$$

or, in component form, for $j \in \{1,\dots,d\}$,

$$\int_s^t f(X_r)\,\mathrm{d}X_r^j = f(X_s)(X_t^j - X_s^j) + \sum_{i=1}^d \partial_i f(X_s)\int_s^t (X_r^i - X_s^i)\,\mathrm{d}X_r^j + \dots$$

Now, consider a sequence of partitions $P_n = \{0 = t_0^n < t_1^n < \cdots < t_n^n = 1\}$, such that $|P_n| = 1/n$. Provided we have an adequate integral definition, this leads formally to

$$\begin{aligned}\int_0^1 f(X_r)\,\mathrm{d}X_r &= \lim_{n\to\infty}\sum_{i=0}^{n-1}\int_{t_i^n}^{t_{i+1}^n} f(X_r)\,\mathrm{d}X_r \\ &= \lim_{n\to\infty}\sum_{i=0}^{n-1}\left(f(X_{t_i})(X_{t_{i+1}} - X_{t_i}) + \nabla f(X_{t_i})\int_{t_i}^{t_{i+1}}(X_r - X_{t_i})\otimes\,\mathrm{d}X_r + \dots\right).\end{aligned}$$

Thus, we see that if $\alpha = 1$, then all the terms in the RHS of (1.19) apart from the first one vanish in the limit. Indeed, we have that

$$|X_{t_i} - X_{t_{i+1}}| = O(|t_i - t_{i+1}|) = O\left(\frac{1}{n}\right),$$

whereas

$$\left|\int_{t_i}^{t_{i+1}}(X_r - X_{t_i})\otimes\,\mathrm{d}X_r\right| = O(|t_i - t_{i+1}|^2) = O\left(\frac{1}{n^2}\right).$$

Hence, by summing over $i \in \{0,\dots,n-1\}$, the first term of (1.19) becomes $O(1)$, while the remaining terms become $o(n)$. Likewise, if $\alpha > 1/2$, we note that

$$\left|\int_{t_i}^{t_{i+1}}(X_r - X_{t_i})\otimes\,\mathrm{d}X_r\right| = O(|t_i - t_{i+1}|^{2\alpha}) = O\left(\frac{1}{n^{2\alpha}}\right),$$

which implies,

$$\lim_{n\to\infty}\sum_{i=0}^{n-1}\int_{t_i}^{t_{i+1}}(X_r - X_{t_i})\otimes\,\mathrm{d}X_r = 0,$$

together with all higher order terms.

However, if we now consider $\alpha \le 1/2$, the second term in (1.19) does not vanish. Formally, we have

$$\int_0^1 f(X_r)\,\mathrm{d}X_r = \lim_{n\to\infty}\sum_{i=0}^{n-1}\left(f(X_{t_i})(X_{t_{i+1}} - X_{t_i}) + \nabla f(X_{t_i})\int_{t_i}^{t_{i+1}}(X_r - X_{t_i})\otimes\,\mathrm{d}X_r\right),$$

for all $\alpha \in \left(\frac{1}{3}, \frac{1}{2}\right]$. This suggests that the information required to compute (1.18), assuming a regularity $\alpha \in \left(\frac{1}{3}, \frac{1}{2}\right]$, is contained not only in the increments $(X_t - X_s)$, but also in the double iterated integral $\int_s^t (X_r - X_s) \otimes \mathrm{d}X_r$. In other words, if we want to compute (1.18) in a pathwise fashion, we must consider as input both the path increments of X and the double iterated integral of X.

We thus arrive at perhaps the biggest conceptual leap of Rough Path theory: to achieve pathwise integration of highly irregular paths, sufficient knowledge of higher order iterated integrals must be *a priori* information, i.e. we must somehow enhance the original (highly irregular) path with features that emulate higher order iterated integrals. A "rough path" will then correspond to this "enhancement." Symbolically, we have

$$X \overset{\text{enhanced}}{\rightsquigarrow} (\mathbb{X}^{(1)}, \mathbb{X}^{(2)}, \ldots, \mathbb{X}^{(n)}),$$

where $\mathbb{X}^{(n)}$ is to be thought of as a "candidate" to the indefinite n-fold iterated integral

$$\int \cdots \int \mathrm{d}X_{t_1} \otimes \cdots \otimes \mathrm{d}X_{t_n}.$$

The forthcoming sections will formalise this heuristic approach, and, by the end of this chapter, the reader should ideally have a clear understanding of the integrators utilised in Rough Path theory.

1.3.2. Sewing Lemma and Abstract Integration

We provide a comprehensive overview of one of the cornerstones of Rough Path theory: the Sewing lemma. This theorem serves as a toolbox enabling the construction of an integral, or, more precisely, a functional with the characteristics of an integral. Although it is a rather abstract result, its interpretation will hopefully become apparent in this section. We conclude by offering a proof for the convergence of Young integrals (see Theorem III).

We begin by discussing "some sort of abstract Riemann integration," paraphrasing the authors in [36]. As elucidated in the previous section, our aim is to attribute meaning to

$$Z_t := \int_0^t Y_r \, \mathrm{d}X_r. \tag{1.20}$$

Under certain regularity restrictions, we have observed that the integral Z_t can be constructed via a limiting procedure. Specifically, we can define Z_t as the limit of Riemann sums, given by

$$Z_t := \lim_{|\mathcal{P}| \to 0} \sum_{[u,v] \in \mathcal{P}([0,t])} Y_u X_{u,v}.$$

Think of the Riemann-Stieltjes integral or the Young integral. Underlying this construction is the idea that $Y_s X_{s,t}$ serves as a good local approximation of $Z_{s,t}$. Symbolically,

$$Z_{s,t} = Y_s X_{s,t} + o(|t-s|),$$

where $\Xi_{s,t} := Y_s X_{s,t}$ is a good approximation in the sense that it fully determines the integral Z through the limit of Riemann sums. Consequently, we can reinterpret the integral Z as the image of Ξ under "some abstract integration map" $\mathcal{I}$, i.e., $Z \equiv \mathcal{I}(\Xi)$.

Furthermore, note that any reasonable notion of integral should exhibit additive increments, i.e., $Z_{s,t} = Z_{s,u} + Z_{u,t}$, for all $s < u < t$. However, $\Xi_{s,t}$ clearly lacks additivity. Nevertheless, again under certain regularity assumptions, the functional $\Xi_{s,t}$ is shown to be "almost additive."

For instance, assuming $X \in C^{\alpha\text{-Höl}}([0,T],\mathbb{R}^d)$, $Y \in C^{\beta\text{-Höl}}([0,T], L(\mathbb{R}^d,\mathbb{R}^e))$, and $\alpha+\beta > 1$ as in Theorem III, we observe that

$$\Xi_{s,t} - \Xi_{s,u} - \Xi_{u,t} = Y_s X_{s,t} - Y_s X_{s,u} - Y_u X_{u,t} = -(Y_u - Y_s)(X_t - X_u),$$

and thus,

$$|\Xi_{s,t} - \Xi_{s,u} - \Xi_{u,t}| = O(|t-s|^{\alpha+\beta}).$$

Definition 1.17: Let $(W, |\cdot|_W)$ be some Banach space. A continuous map $\Xi : \Delta_T^2 \to W$ is termed an almost additive functional if there exist constants $C, \varepsilon > 0$ such that

$$|\Xi_{s,t} - \Xi_{s,u} - \Xi_{u,t}|_W \leq C|t-s|^{1+\varepsilon}, \text{ for all } 0 \leq s < u < t \leq T.$$

With the Sewing lemma, we will demonstrate that any almost additive functional corresponds to a unique additive functional. This crucial insight was noted and demonstrated in [35], but the approach we adopt in this work is credited to Young. First, however, we introduce some more notation.

Definition 1.18: Let $(W, |\cdot|_W)$ be a Banach space. The space $C_2^{\alpha,\beta}([0,T],W)$ denotes the set of functions $\Xi : \Delta_T^2 \to W$ such that $\Xi_{t,t} = 0$ and

$$|\Xi|_{\alpha,\beta} := |\Xi|_\alpha + |\delta\Xi|_\beta < \infty,$$

$$|\Xi|_\alpha := \sup_{0\leq s<t\leq T} \frac{|\Xi_{s,t}|}{|t-s|^\alpha}, \quad \text{and} \quad \delta\Xi_{s,u,t} := \Xi_{s,t} - \Xi_{s,u} - \Xi_{u,t}, \quad |\delta\Xi|_\beta := \sup_{s<u<t} \frac{|\delta\Xi_{s,u,t}|}{|t-s|^\beta}.$$

To avoid confusion, it is important to note that if X is a path, then $X_{s,t}$ represents the increment $X_t - X_s$. However, when dealing with two-parameter functions Ξ defined on Δ_T^2, $\Xi_{s,t}$ denotes the value of Ξ at the point $(s,t) \in \Delta_T^2$.

Theorem 1.4 ([36], Sewing Lemma)**:** Let α and β be such that $0 < \alpha \leq 1 < \beta$. Then, there exists a unique continuous linear map $\mathcal{I} : C_2^{\alpha,\beta}([0,T],W) \to C^{\alpha\text{-Höl}}([0,T],W)$ such that $(\mathcal{I}\Xi)_0 = 0$ and

$$|(\mathcal{I}\Xi)_{s,t} - \Xi_{s,t}|_W \leq C|t-s|^\beta, \tag{1.21}$$

where $C = |\delta\Xi|_\beta\big(2^\beta(\zeta(\beta)-1)+1\big)$ and ζ denotes the Riemann zeta function. Additionally,

$$\mathcal{I}\Xi_{s,t} = \lim_{|\mathcal{P}|\to 0} \sum_{[u,v]\in\mathcal{P}([s,t])} \Xi_{u,v}. \tag{1.22}$$

Proof. **Uniqueness.** We first establish uniqueness. Assume that we have another continuous linear functional $\tilde{\mathcal{I}}$ such that (1.21) is satisfied. Then, by the triangle inequality, it holds that

$$|(\mathcal{I}\Xi - \tilde{\mathcal{I}}\Xi)_t - (\mathcal{I}\Xi - \tilde{\mathcal{I}}\Xi)_s|_W \leq C|t-s|^\beta.$$

Since $\beta > 1$ and $\mathcal{I}\Xi - \tilde{\mathcal{I}}\Xi$ is a path, we have by Proposition 1.4 that $\mathcal{I}\Xi - \tilde{\mathcal{I}}\Xi$ is constant. Lastly, given that $(\mathcal{I}\Xi - \tilde{\mathcal{I}}\Xi)_0 = 0$, uniqueness follows. Moreover, from uniqueness and (1.21) it follows that $\mathcal{I}\Xi_{s,t}$ is necessarily given as a Riemann-type limit. Indeed, fixing a partition $\mathcal{P}$ of $[s,t]$, we have that

$$\Big|\mathcal{I}\Xi_{s,t} - \sum_{[u,v]\in\mathcal{P}} \Xi_{u,v}\Big|_W = \Big|\sum_{[u,v]\in\mathcal{P}} \big(\mathcal{I}\Xi_{u,v} - \Xi_{u,v}\big)\Big|_W = O(|\mathcal{P}|^{\beta-1}),$$

and so (1.22) is fulfilled.

Existence. Let $[s,t]$ be a fixed interval and for a partition $\mathcal{P} = \{s = u_0 < \cdots < u_r = t\}$ set

$$\int_{\mathcal{P}} \Xi := \sum_{[u,v]\in\mathcal{P}} \Xi_{u,v}.$$

Note that $r \geq 1$ is the number of subintervals in $\mathcal{P}$. If $r \geq 2$, denote by u_- and u_+ the neighbouring points of u, i.e. $u_- < u < u_+ \in \mathcal{P}$. Observe that, in this case, we have

$$|u_+ - u_-| \leq \frac{2}{r-1}|t-s|. \tag{1.23}$$

Indeed, assuming otherwise yields the contradiction

$$2|t-s| \geq \sum_{u\in\mathcal{P}\setminus\{u_0,u_r\}} |u_+ - u_-| > 2|t-s|.$$

Thus, still assuming $r \geq 2$, we see that

$$\left|\int_{\mathcal{P}} \Xi - \int_{\mathcal{P}\setminus\{u\}} \Xi\right| = |\delta\Xi_{u_-,u,u_+}| \leq |\delta\Xi|_\beta |u_+ - u_-|^\beta \leq |\delta\Xi|_\beta \left(\frac{2}{r-1}\right)^\beta |t-s|^\beta,$$

where we used (1.23) in the final inequality. Subsequently, by successively removing points, we get the uniform bound

$$\sup_{\mathcal{P}} \left|\int_{\mathcal{P}} \Xi - \Xi_{s,t}\right| \leq 2^\beta |t-s|^\beta |\delta\Xi|_\beta \sum_{k=2}^{\infty} \frac{1}{k^\beta} + |t-s|^\beta |\delta\Xi|_\beta \tag{1.24}$$

$$\leq |\delta\Xi|_\beta \big(2^\beta(\zeta(\beta)-1)+1\big)|t-s|^\beta. \tag{1.25}$$

Note that the assumption $\beta > 1$ is essential here for the convergence of the infinite series above. It remains to show the existence of $\mathcal{I}\Xi$ as the limit $\lim_{|\mathcal{P}|\to 0} \int_{\mathcal{P}} \Xi$. It suffices to show

$$\sup_{|\mathcal{P}|\vee|\mathcal{P}'|\le\varepsilon} \left| \int_{\mathcal{P}} \Xi - \int_{\mathcal{P}'} \Xi \right| \to 0, \text{ as } \varepsilon \to 0.$$

Without loss of generality, we may assume that $\mathcal{P} \subset \mathcal{P}'$. Otherwise, we simply add and subtract $\int_{\mathcal{P}\cup\mathcal{P}'} \Xi$. In this case, we see that

$$\int_{\mathcal{P}} \Xi - \int_{\mathcal{P}'} \Xi = \sum_{[u,v]\in\mathcal{P}} \left(\Xi_{u,v} - \int_{\mathcal{P}'\cap[u,v]} \Xi \right).$$

Lastly, we use (1.24) to conclude that

$$\sup_{|\mathcal{P}|\le\varepsilon} \left| \int_{\mathcal{P}} \Xi - \int_{\mathcal{P}'} \Xi \right| \le |\delta\Xi|_\beta \big(2^\beta(\zeta(\beta) - 1) + 1\big) \sum_{[u,v]\in\mathcal{P}} |v - u|^\beta = O(|\mathcal{P}|^{\beta-1}) = O(\varepsilon^{\beta-1}),$$

and the result follows. □

Equipped with the Sewing lemma, establishing the well-definedness of the Young integral becomes relatively straightforward. For convenience, we recall Theorem III here and augment it with an estimate for completeness.

Theorem 1.5: Let $X : [0,T] \to \mathbb{R}^d$ be an α-Hölder continuous path and $Y : [0,T] \to L(\mathbb{R}^d, \mathbb{R}^e)$ a β-Hölder continuous path, such that $\alpha, \beta \in (0,1]$ and $\alpha + \beta > 1$. Moreover, let $P_n = \{0 = t_0^n < t_1^n < \ldots < t_{N_n}^n = T\}$, $n \ge 1$ be a sequence of partitions with vanishing mesh size, and let u_i^n denote an arbitrary test point in the interval $[t_i^n, t_{i+1}^n]$. Then, the integral

$$\int_0^T Y_s dX_s := \lim_{n\to\infty} \sum_{i=0}^{N_n-1} Y_{u_i^n}(X_{t_{i+1}^n} - X_{t_i^n})$$

exists and we call it the Young integral. Moreover, we have the estimate

$$\left| \int_s^t Y_r \, \mathrm{d}X_r - Y_s X_{s,t} \right| \le C |Y|_{\beta\text{-Höl};[0,T]} |X|_{\alpha\text{-Höl};[0,T]} |t - s|^{\alpha+\beta},$$

where $C > 0$ depends on $\alpha + \beta$.

Proof. Let $\Xi_{s,t} := Y_s X_{s,t}$. Then, for all $s < u < t$, we have that

$$\delta\Xi_{s,u,t} = Y_s X_{s,t} - Y_s X_{s,u} - Y_u X_{u,t} = -Y_{s,u} X_{u,t},$$

and so, $|\Xi|_\alpha \le |Y|_{\infty;[0,T]} \cdot |X|_{\alpha\text{-Höl};[0,T]} < \infty$ and $|\delta\Xi|_{\alpha+\beta} \le |Y|_{\beta\text{-Höl};[0,T]} \cdot |X|_{\alpha\text{-Höl};[0,T]} < \infty$. Since $\alpha + \beta > 1$, by the Sewing lemma (Theorem 1.4), we can set

$$\int_s^t Y_u \, \mathrm{d}X_u = \mathcal{I}\Xi_{s,t},$$

and the result follows. □

1.3.3. Lyons' Extension Theorem

The discussion in Section 1.3.1 shed light on the fact that the more irregular a path is, the more information is required to define an integral against it. Moreover, it suggested that this lack of information can be compensated for by incorporating higher-order iterated integrals. This insight led us to a key principle of Rough Path theory: irregular paths need to be enhanced or enriched with objects that encode the values of higher-order iterated integrals.

Subsequently, in Section 1.3.2, we explored a form of abstract integration and introduced a tool capable of constructing functionals that behave analogously to integrals. In this section, we delve into how the Sewing lemma can be readily applied to higher-order iterated integrals and establish what is arguably the main result in Rough Path theory: the Lyons' Extension theorem. As we examine this fundamental result, we uncover the interplay between additive and multiplicative functionals.

We start by defining the analogue of Definition 1.17 for multiplicative functionals. However, to facilitate our discussion, it is beneficial to introduce some notation first.

Definition 1.19: A map $\mathbb{X} : \Delta_T^2 \to T^N(\mathbb{R}^d)$ with $\mathbb{X} = (1, \mathbb{X}^{(1)}, \ldots, \mathbb{X}^{(N)})$ and components $\mathbb{X}^{(n)} \in (\mathbb{R}^d)^{\otimes n}$, for $n \in \{0, \ldots, N\}$, is said to be a multiplicative functional of degree N if, for all $0 \leq s \leq u \leq t \leq T$,

$$\mathbb{X}_{s,t} = \mathbb{X}_{s,u} \otimes \mathbb{X}_{u,t},$$

where the multiplication is taken in $T^N(\mathbb{R}^d)$.

Put differently, a multiplicative functional of degree N corresponds to a mapping over Δ_T^2 that adheres to Chen's identity (1.11) truncated at level N (Proposition 1.13). Conceptually, multiplicative functionals of degree N can be perceived as mappings that exhibit the algebraic traits of a sequence of iterated integrals up to level N.

Definition 1.20: A map $\mathbb{Y} : \Delta_T^2 \to T^N(\mathbb{R}^d)$, with $\mathbb{Y} = (1, \mathbb{Y}^{(1)}, \ldots, \mathbb{Y}^{(N)})$, is said to be an almost multiplicative functional of degree N, if there exist constants $C, \varepsilon > 0$ such that

$$|\mathbb{Y}_{s,t}^{(n)} - (\mathbb{Y}_{s,u} \otimes \mathbb{Y}_{u,t})^{(n)}| \leq C^n |t-s|^{1+\varepsilon}, \tag{1.26}$$

for all $0 \leq s \leq u \leq t \leq T$ and $n \in \{0, \ldots, N\}$.

Notice that a multiplicative functional inherently satisfies the conditions of being almost multiplicative, resulting in the LHS of (1.26) equating to zero. For clarification, let us consider the path signatures discussed in Section 1.2. Thanks to Proposition 1.13 we know that the signature (specifically, the level-N truncated signature) is a multiplicative functional of degree

N. Particularly, for any $X \in C^{\alpha\text{-Höl}}([0,T],\mathbb{R}^d)$, with $\alpha > 1/2$, the functional

$$\mathbb{X}_{s,t} := \left(1, \int_s^t \mathrm{d}X_u, \int_s^t \int_s^u \mathrm{d}X_r \otimes \mathrm{d}X_u\right) \tag{1.27}$$

is multiplicative of degree 2. It is worth noting that our selection of $\alpha > 1/2$ is justified by the well-defined nature of the Young integral, as demonstrated in the preceding section.

Now, aiming to find a concrete example of an almost multiplicative functional while still adhering to the idea of incorporating higher-order iterated integrals, let us consider

$$\mathbb{Y}_{s,t} := \left(1, \int_s^t \mathrm{d}X_u, \int_s^t \int_s^u \mathrm{d}X_r \otimes \mathrm{d}X_u, \mathbf{0}\right) \in T^3(\mathbb{R}^d).$$

It is apparent that $\mathbb{Y}$ is not multiplicative of degree 3, as Chen's identity breaks at level 3. Nevertheless, we note that $\mathbb{Y}$ is almost multiplicative of degree 3. Indeed, keeping in mind that for the Young integral we have

$$\int_s^t \mathrm{d}X_u = O(|t-s|^{\alpha}) \quad \text{and} \quad \int_s^t \int_s^u \mathrm{d}X_r \otimes \mathrm{d}X_u = O(|t-s|^{2\alpha}), \tag{1.28}$$

we ascertain that

$$\left|(\mathbb{Y}_{s,u} \otimes \mathbb{Y}_{u,t})^{(3)}\right| = \left|\mathbb{X}^{(1)}_{s,u} \otimes \mathbb{X}^{(2)}_{u,t} + \mathbb{X}^{(2)}_{s,u} \otimes \mathbb{X}^{(1)}_{u.t}\right| = O(|t-s|^{3\alpha}), \tag{1.29}$$

indicating that (1.26) holds true with $\varepsilon = 3\alpha - 1$. Moreover, similarly to the proof of Theorem 1.5, we define

$$\Xi_{s,t} := \mathbb{X}^{(2)}_{0,s} \otimes \mathbb{X}^{(1)}_{s,t} + \mathbb{X}^{(1)}_{0,s} \otimes \mathbb{X}^{(2)}_{s,t} \in (\mathbb{R}^d)^{\otimes 3},$$

and observe that, through the addition and subtraction of $\mathbb{X}^{(2)}_{0,s} \otimes \mathbb{X}^{(1)}_{u,t}$, and the application of Chen's identity, we obtain

$$\begin{aligned}
\delta\Xi_{s,u,t} &= \Xi_{s,t} - \Xi_{s,u} - \Xi_{u,t} \\
&= -\mathbb{X}^{(1)}_{0,u} \otimes \mathbb{X}^{(2)}_{u,t} + \mathbb{X}^{(1)}_{0,s} \otimes \left(\mathbb{X}^{(2)}_{s,t} - \mathbb{X}^{(2)}_{s,u}\right) + \left(\mathbb{X}^{(2)}_{0,s} - \mathbb{X}^{(2)}_{0,u}\right) \otimes \mathbb{X}^{(1)}_{u,t} \\
&= -\mathbb{X}^{(1)}_{0,u} \otimes \mathbb{X}^{(2)}_{u,t} + \mathbb{X}^{(1)}_{0,s} \otimes \left(\mathbb{X}^{(2)}_{u,t} + \mathbb{X}^{(1)}_{s,u} \otimes \mathbb{X}^{(1)}_{u,t}\right) - \left(\mathbb{X}^{(2)}_{s,u} + \mathbb{X}^{(1)}_{0,s} \otimes \mathbb{X}^{(1)}_{s,u}\right) \otimes \mathbb{X}^{(1)}_{u,t} \\
&= -\mathbb{X}^{(2)}_{s,u} \otimes \mathbb{X}^{(1)}_{u,t} - \mathbb{X}^{(1)}_{s,u} \otimes \mathbb{X}^{(2)}_{u,t}.
\end{aligned}$$

Therefore, $|\delta\Xi|_{3\alpha} < \infty$, showing Ξ to be an almost additive functional. Given that $|\Xi|_{\alpha} < \infty$ by (1.28) and Proposition 1.5, we deduce from the Sewing lemma (Theorem 1.4) that

$$\mathcal{I}\Xi_{s,t} = \lim_{|\mathcal{P}|\to 0} \sum_{[u,v]\in\mathcal{P}([s,t])} \mathbb{X}^{(2)}_{0,u} \otimes \mathbb{X}^{(1)}_{u,v} + \mathbb{X}^{(1)}_{0,u} \otimes \mathbb{X}^{(2)}_{u,v} =: \int_s^t \mathbb{X}^{(2)}_{0,u} \otimes \mathrm{d}X_u,$$

where the integral definition is justified by the fact that $|\mathbb{X}^{(2)}_{u,v}| = O(|u-v|^{2\alpha})$. Moreover,

$$\left|\int_{s<u_1<u_2<u_3<t} \mathrm{d}X_{u_1} \otimes \mathrm{d}X_{u_2} \otimes \mathrm{d}X_{u_3}\right| = O(|t-s|^{3\alpha}).$$

Indeed, by Chen's identity, $\mathbb{X}^{(2)}_{0,u} = \mathbb{X}^{(2)}_{0,s} + \mathbb{X}^{(2)}_{s,u} + \mathbb{X}^{(1)}_{0,s} \otimes \mathbb{X}^{(1)}_{s,u}$, which implies

$$\begin{aligned}\mathbb{X}^{(3)}_{s,t} &:= \int_s^t \mathbb{X}^{(2)}_{s,u} \otimes \mathrm{d}X_u = \int_s^t \left(\mathbb{X}^{(2)}_{0,u} - \mathbb{X}^{(2)}_{0,s} - \mathbb{X}^{(1)}_{0,s} \otimes \mathbb{X}^{(1)}_{s,u}\right) \otimes \mathrm{d}X_u \\ &= \int_s^t \mathbb{X}^{(2)}_{0,u} \otimes \mathrm{d}X_u - \mathbb{X}^{(2)}_{0,s} \otimes \mathbb{X}^{(1)}_{s,t} - \mathbb{X}^{(1)}_{0,s} \otimes \mathbb{X}^{(2)}_{s,t} = \mathcal{I}\Xi_{s,t} - \Xi_{s,t}.\end{aligned}$$

Several pertinent observations can now be made. Firstly, we notice that starting from a multiplicative functional $\mathbb{X}_{s,t}$ of degree 2, we defined an almost multiplicative functional $\mathbb{Y}_{s,t}$ of degree 3. This was driven by the underlying motivation of incorporating higher-order iterated integrals. Remarkably, from the almost multiplicative functional $\mathbb{Y}_{s,t}$, we inferred an almost additive functional $\Xi_{s,t}$. Subsequently, by applying the Sewing lemma, we obtained the additive functional $\mathcal{I}\Xi_{s,t}$, which allowed us to "postulate" a 3-fold iterated integral as the limit of Riemann sums. Finally, $\mathbb{X}^{(3)}_{s,t} = \mathcal{I}\Xi_{s,t} - \Xi_{s,t}$ leads to a multiplicative functional of degree 3, essentially by construction.

To see this, consider the additivity of $\mathcal{I}\Xi$, which implies

$$\mathbb{X}^{(3)}_{s,t} - \mathbb{X}^{(3)}_{s,u} - \mathbb{X}^{(3)}_{u,t} = -\delta\Xi_{s,u,t} = \mathbb{X}^{(2)}_{s,u} \otimes \mathbb{X}^{(1)}_{u,t} + \mathbb{X}^{(1)}_{s,u} \otimes \mathbb{X}^{(2)}_{u,t},$$

indicating that

$$\mathbb{X}_{s,t} := \left(1, \int_s^t \mathrm{d}X_u, \int_s^t \int_s^u \mathrm{d}X_r \otimes \mathrm{d}X_u, \int_s^t \int_s^u \int_s^v \mathrm{d}X_r \otimes \mathrm{d}X_v \otimes \mathrm{d}X_u\right) \in T^3(\mathbb{R}^d)$$

is multiplicative of degree 3. Furthermore, as noted above, the third level of $\mathbb{X}_{s,t}$ scales with $|t-s|^{3\alpha}$. Consequently, employing the notation from Definition 1.18, we have $|\mathbb{X}^{(3)}_{s,t}|_{3\alpha} < \infty$. Here, we make a deliberate abuse of notation and denote the newly obtained multiplicative functional of degree 3 by $\mathbb{X}$, highlighting that this functional is indeed an extension of (1.27).

In summary, by iterating the above reasoning, it is suggested that given a multiplicative functional of degree n with sufficiently strong regularity constraints on each level, we can uniquely extend it to a well-defined multiplicative functional of degree $N \geq n$ while preserving the regularity properties of the original functional. This encapsulates the essence of the Lyons' Extension theorem. We now proceed to formalise and generalise these observations into two auxiliary results, which ultimately lead to the Lyons' Extension theorem.

Proposition 1.21 ([35], Theorem 2.5)**:** Let $\mathbb{X} : \Delta^2_T \to T^n(\mathbb{R}^d)$ be a continuous multiplicative functional of degree n. Additionally, consider a continuous map $\mathbb{Y}^{(n+1)} : \Delta^2_T \to (\mathbb{R}^d)^{\otimes(n+1)}$ such that $\mathbb{Y} := \left(1, \mathbb{X}^{(1)}, \dots, \mathbb{X}^{(n)}, \mathbb{Y}^{(n+1)}\right)$ is an almost multiplicative functional of degree $n+1$. Then, $\Xi : \Delta^2_T \to (\mathbb{R}^d)^{\otimes(n+1)}$ given by

$$\Xi_{s,t} := \mathbb{Y}^{(n+1)}_{s,t} + \sum_{i=1}^{n} \mathbb{X}^{(i)}_{0,s} \otimes \mathbb{X}^{(n+1-i)}_{s,t}, \tag{1.30}$$

is an almost additive functional.

Proof. Since $\mathbb{Y}$ is an almost multiplicative functional, for all $0 \le s < u < t \le T$, we have

$$|\mathbb{Y}_{s,t}^{(n+1)} - (\mathbb{Y}_{s,u} \otimes \mathbb{Y}_{u,t})^{(n+1)}| \le C^{n+1}|t-s|^{1+\varepsilon}.$$

Now, using Chen's identity (1.11), we show that

$$-\sum_{i=1}^{n} \mathbb{X}_{s,u}^{(i)} \otimes \mathbb{X}_{u,t}^{(n+1-i)} = \sum_{i=1}^{n} \mathbb{X}_{0,s}^{(i)} \otimes \mathbb{X}_{s,t}^{(n+1-i)} - \sum_{i=1}^{n} \mathbb{X}_{0,s}^{(i)} \otimes \mathbb{X}_{s,u}^{(n+1-i)} - \sum_{i=1}^{n} \mathbb{X}_{0,u}^{(i)} \otimes \mathbb{X}_{u,t}^{(n+1-i)}.$$

To begin with, we notice that

$$\begin{aligned}
&\sum_{i=1}^{n} \mathbb{X}_{0,s}^{(i)} \otimes \mathbb{X}_{s,t}^{(n+1-i)} - \sum_{i=1}^{n} \mathbb{X}_{0,s}^{(i)} \otimes \mathbb{X}_{s,u}^{(n+1-i)} - \sum_{i=1}^{n} \mathbb{X}_{0,u}^{(i)} \otimes \mathbb{X}_{u,t}^{(n+1-i)} \\
&= \sum_{i=1}^{n} \Big(\mathbb{X}_{0,s}^{(i)} \otimes \big(\mathbb{X}_{s,t}^{(n+1-i)} - \mathbb{X}_{s,u}^{(n+1-i)} - \mathbb{X}_{u,t}^{(n+1-i)} \big) + \\
&\qquad + \big(\mathbb{X}_{0,s}^{(i)} - \mathbb{X}_{0,u}^{(i)} + \mathbb{X}_{s,u}^{(i)} \big) \otimes \mathbb{X}_{u,t}^{(n+1-i)} - \mathbb{X}_{s,u}^{(i)} \otimes \mathbb{X}_{u,t}^{(n+1-i)} \Big).
\end{aligned} \tag{1.31}$$

Then, by Chen's identity,

$$\mathbb{X}_{s,t}^{(n+1-i)} - \mathbb{X}_{s,u}^{(n+1-i)} - \mathbb{X}_{u,t}^{(n+1-i)} = \sum_{j=1}^{n-i} \mathbb{X}_{s,u}^{(j)} \otimes \mathbb{X}_{u,t}^{(n+1-i-j)}, \tag{1.32}$$

and, likewise,

$$\mathbb{X}_{0,s}^{(i)} - \mathbb{X}_{0,u}^{(i)} + \mathbb{X}_{s,u}^{(i)} = -\sum_{j=1}^{i-1} \mathbb{X}_{0,s}^{(j)} \otimes \mathbb{X}_{s,u}^{(i-j)}. \tag{1.33}$$

Here, we agree that a sum over an empty set of indices is identically zero. By plugging (1.32) and (1.33) into (1.31), changing the order of summation, and performing a change of variables, we get that

$$\begin{aligned}
&\sum_{i=1}^{n} \mathbb{X}_{0,s}^{(i)} \otimes \big(\mathbb{X}_{s,t}^{(n+1-i)} - \mathbb{X}_{s,u}^{(n+1-i)} - \mathbb{X}_{u,t}^{(n+1-i)} \big) + \sum_{i=1}^{n} \big(\mathbb{X}_{0,s}^{(i)} - \mathbb{X}_{0,u}^{(i)} - \mathbb{X}_{s,u}^{(i)} \big) \otimes \mathbb{X}_{u,t}^{(n+1-i)} \\
&= \sum_{i=1}^{n} \mathbb{X}_{0,s}^{(i)} \otimes \big(\sum_{j=1}^{n-i} \mathbb{X}_{s,u}^{(j)} \otimes \mathbb{X}_{u,t}^{(n+1-i-j)} \big) - \sum_{i=1}^{n} \big(\sum_{j=1}^{i-1} \mathbb{X}_{0,s}^{(j)} \otimes \mathbb{X}_{s,u}^{(i-j)} \big) \otimes \mathbb{X}_{u,t}^{(n+1-i)} \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n-i} \mathbb{X}_{0,s}^{(i)} \otimes \mathbb{X}_{s,u}^{(j)} \otimes \mathbb{X}_{u,t}^{(n+1-i-j)} - \sum_{i=1}^{n} \sum_{j=1}^{i-1} \mathbb{X}_{0,s}^{(j)} \otimes \mathbb{X}_{s,u}^{(i-j)} \otimes \mathbb{X}_{u,t}^{(n+1-i)} = 0.
\end{aligned}$$

Hence, we obtain

$$\delta\Xi_{s,u,t} = \mathbb{Y}_{s,t}^{(n+1)} - \mathbb{Y}_{s,u}^{(n+1)} - \mathbb{Y}_{u,t}^{(n+1)} - \sum_{i=1}^{n} \mathbb{X}_{s,u}^{(i)} \otimes \mathbb{X}_{u,t}^{(n+1-i)} = \big(\mathbb{Y}_{s,t} - \mathbb{Y}_{s,u} \otimes \mathbb{Y}_{u,t} \big)^{(n+1)}, \tag{1.34}$$

and so, $|\delta\Xi_{s,u,t}| \le C^{n+1}|t-s|^{1+\varepsilon}$, i.e. Ξ is almost additive. □

Proposition 1.22: Consider $\mathbb{X}$ and $\mathbb{Y}$ as in Proposition 1.21. Let Ξ denote the almost additive functional given by (1.30). Then, $\mathbb{X}^{(n+1)} := \mathcal{I}\Xi - \Xi + \mathbb{Y}^{(n+1)}$ is the unique functional $\Delta_T^2 \to (\mathbb{R}^d)^{\otimes(n+1)}$ such that $\mathbb{X} = \left(1, \mathbb{X}^{(1)}, \dots, \mathbb{X}^{(n)}, \mathbb{X}^{(n+1)}\right)$ is a multiplicative functional of degree $n+1$, and

$$\left|\mathbb{X}_{s,t}^{(n+1)} - \mathbb{Y}_{s,t}^{(n+1)}\right| = O(|t-s|^{1+\varepsilon}) \tag{1.35}$$

Proof. To begin with, by the Sewing lemma (Theorem 1.4), there exists a unique additive functional $\mathcal{I}\Xi : [0,T] \to (\mathbb{R}^d)^{\otimes(n+1)}$ such that

$$|\mathcal{I}\Xi_{s,t} - \Xi_{s,t}| \leq C|t-s|^{1+\varepsilon}.$$

Additionally, by setting

$$\mathbb{X}_{s,t}^{(n+1)} := \mathcal{I}\Xi_{s,t} - \sum_{i=1}^{n} \mathbb{X}_{0,s}^{(i)} \otimes \mathbb{X}_{s,t}^{(n+1-i)}, \tag{1.36}$$

we observe that,

$$\mathbb{X}_{s,t}^{(n+1)} - \mathbb{X}_{s,u}^{(n+1)} - \mathbb{X}_{u,t}^{(n+1)} = \sum_{i=1}^{n} \mathbb{X}_{s,u}^{(i)} \otimes \mathbb{X}_{u,t}^{(n+1-i)},$$

and

$$\left|\mathbb{X}_{s,t}^{(n+1)} - \mathbb{Y}_{s,t}^{(n+1)}\right| = |\mathcal{I}\Xi_{s,t} - \Xi_{s,t}| \leq C|t-s|^{1+\varepsilon}.$$

Hence, the extended $\mathbb{X}$ is multiplicative of degree $n+1$ and (1.35) follows. Finally, to prove uniqueness, assume there exists some other functional $\tilde{\mathbb{X}}^{(n+1)} : \Delta_T^2 \to (\mathbb{R}^d)^{\otimes(n+1)}$, turning $\tilde{\mathbb{X}} := \left(1, \mathbb{X}^{(1)}, \dots, \mathbb{X}^{(n)}, \tilde{\mathbb{X}}^{(n+1)}\right)$ into a degree $n+1$ multiplicative functional such that

$$\left|\tilde{\mathbb{X}}_{s,t}^{(n+1)} - \mathbb{Y}_{s,t}^{(n+1)}\right| = O(|t-s|^{1+\varepsilon}).$$

Set $\Psi_{s,t} := \mathbb{X}_{s,t}^{(n+1)} - \tilde{\mathbb{X}}_{s,t}^{(n+1)}$. Then, notice that

$$\begin{aligned}\Psi_{s,t} =& \pi_{n+1}(\tilde{\mathbb{X}}_{s,u} \otimes \tilde{\mathbb{X}}_{u,t}) - \pi_{n+1}(\mathbb{X}_{s,u} \otimes \mathbb{X}_{u,t}) \\ =& \tilde{\mathbb{X}}_{s,u}^{(n+1)} + \tilde{\mathbb{X}}_{u,t}^{(n+1)} - \mathbb{X}_{s,u}^{(n+1)} - \mathbb{X}_{u,t}^{(n+1)} = \Psi_{s,u} + \Psi_{u,t}.\end{aligned}$$

Therefore, $t \mapsto \Psi_{0,t}$ is a $(1+\varepsilon)$-Hölder continuous path, which, by Proposition 1.4, is constant. Since $\Psi_{0,0} = 0$, uniqueness follows. □

We are now prepared to state Lyons' Extension theorem. Assuming we begin with a multiplicative functional of degree n that satisfies sufficient regularity constraints on its components, the proof proceeds as follows:

1. From the degree n multiplicative functional, we derive a degree $n+1$ almost multiplicative functional.

2. From the degree $n+1$ almost multiplicative functional, we derive an almost additive functional (see Proposition 1.21).
3. Utilising the Sewing lemma (Theorem 1.4), we obtain a unique additive functional from the almost additive functional.
4. From the additive functional, we obtain a unique multiplicative functional of degree $n+1$ (see Proposition 1.22).

This reasoning underlies the original proof ([53, 54], Theorem 3.1.2 and Theorem 2.2.1, respectively), and was initially highlighted in ([35], Example 4).

Theorem 1.6 (Lyons' Extension)**:** Consider $\alpha \in (0,1]$. Let $\mathbb{X} : \Delta_T^2 \to T^n(\mathbb{R}^d)$ be a multiplicative functional of degree n such that

$$|\mathbb{X}^{(i)}| = O(|t-s|^{\alpha i}), \text{ for all } i \in \{1,\dots,n\}. \tag{1.37}$$

If $\alpha(n+1) > 1$, then we may uniquely extend $\mathbb{X}$ to be a multiplicative functional $\mathbb{X} : \Delta_T^2 \to T((\mathbb{R}^d))$ such that (1.37) remains true for all $i > n$.

Proof. Let $\mathbb{X} = (1, \mathbb{X}^{(1)}, \dots, \mathbb{X}^{(n)})$ denote a multiplicative functional of degree n such that (1.37) holds. Given the identity (1.34) and the assumption that $\alpha(n+1) > 1$, we establish that $(1, \mathbb{X}^{(1)}, \dots, \mathbb{X}^{(n)}, \mathbf{0}) \in T^{n+1}(\mathbb{R}^d)$ represents an almost multiplicative functional of degree $n+1$. Consequently, according to Proposition 1.21, we derive an almost additive functional Ξ defined by (1.30), such that $|\Xi|_\alpha, |\delta\Xi|_{\alpha(n+1)} < \infty$. Leveraging Proposition 1.22, we then obtain a unique functional $\mathbb{X}^{(n+1)} : \Delta_T^2 \to (\mathbb{R}^d)^{\otimes(n+1)}$ satisfying

$$|\mathbb{X}_{s,t}^{(n+1)}| \leq C|t-s|^{\alpha(n+1)},$$

and making $(1, \mathbb{X}^{(1)}, \dots, \mathbb{X}^{(n)}, \mathbb{X}^{(n+1)})$ multiplicative of degree $n+1$. Finally, by iterating this procedure infinitely many times, we obtain a multiplicative functional taking values in $T((\mathbb{R}^d))$ such that (1.37) holds. This concludes the proof. □

From Theorem 1.6, we conclude that a multiplicative functional $\mathbb{X} : \Delta_T^2 \to T^n(\mathbb{R}^d)$ satisfying (1.37) uniquely determines all levels higher than $\lfloor 1/\alpha \rfloor$ based on its components $\mathbb{X}^{(i)}$ for $i \leq \lfloor 1/\alpha \rfloor$. However, the first $\lfloor 1/\alpha \rfloor$ components are not uniquely determined, highlighting the importance of functionals whose state-space and regularity align (further elaboration in Section 1.3.5).

In Section 1.3.1, we established that the more irregular a path is, the more iterated integrals must be considered for integration. Now, we realise that once a sufficient number of iterated integrals (with appropriate regularity) are known, the entire set of iterated integrals can be determined.

1.3.4. Factorial Decay and Continuity

We demonstrate two significant properties of Lyons' extension. The first property indicates that, under reasonable assumptions, the norm of the components in the extension decays factorially ([53], Theorem 3.1.2). The second property establishes the continuity of the extension map ([53], Theorem 3.1.3). Additionally, we introduce our first "rough path metric," which will be further explored in the next section. We begin with the latter point.

One of the crucial assumptions of Theorem 1.6 is having the components $\mathbb{X}^{(i)}$ of the multiplicative functional be proportional to $|t-s|^{\alpha i}$. As such, we start this section by defining a convenient "metric" that aligns with this assumption. To do so, consider two multiplicative functionals $\mathbb{X}$ and $\mathbb{Y}$ of degree n. Then, for a given $\alpha \in (0,1]$, set

$$\rho_{\alpha\text{-Höl};[0,T]}(\mathbb{X},\mathbb{Y}) := \max_{i=1,\dots,n} \sup_{0\le s<t\le T} \frac{\left|\mathbb{X}^{(i)}_{s,t} - \mathbb{Y}^{(i)}_{s,t}\right|}{|t-s|^{\alpha i}}. \tag{1.38}$$

Note that assumption (1.37) is now equivalent to requiring $\rho_{\alpha\text{-Höl};[0,T]}(\mathbb{X},\mathbf{1}) < \infty$, where $\mathbf{1}$ denotes the unit of $T^n(\mathbb{R}^d)$. We observe that $\rho_{\alpha\text{-Höl};[0,T]}$ is non-negative, symmetric, and satisfies the triangle inequality. However, $\rho_{\alpha\text{-Höl};[0,T]}(\mathbb{X},\mathbb{Y}) = 0$ whenever $\mathbb{X}$ and $\mathbb{Y}$ differ by a constant, rendering $\rho_{\alpha\text{-Höl};[0,T]}$ not a genuine metric. This issue can be addressed by restricting our attention to paths with a fixed starting point or by adding $|\mathbb{X}_0 - \mathbb{Y}_0|$. Further discussion on this matter will be provided in the next section.

In Section 1.2.2, we established that the norms of the signature components decay factorially. Given that a multiplicative functional $\mathbb{X}$ of degree n represents the values of the first n iterated integrals, it is reasonable to impose a factorial upper bound on each component, similar to Proposition 1.18. We show that if the components of the initially considered $\mathbb{X}$ decay factorially, then the components of the extension to $T((\mathbb{R}^d))$ also decay factorially.

However, to establish this result we need a generalisation of the binomial formula, commonly known as the *neo-classical inequality*, which is surprisingly challenging to prove. To avoid a lengthy and tangential discussion, we refer the reader to ([40], Theorem 1.2) for a rigorous proof of the estimate. We simply state it here for completeness.

Theorem 1.7: Consider some $\alpha \in (0,1]$, $n \in \mathbb{N}$ and $s,t > 0$. Then,

$$\alpha \sum_{j=0}^{n} \frac{s^{\alpha j} t^{\alpha(n-j)}}{(\alpha j)!(\alpha(n-j))!} \le \frac{(t+s)^{\alpha n}}{(\alpha n)!},$$

where $(\alpha j)! := \Gamma(1+\alpha j)$, and Γ denotes the Gamma function.

Equipped with the neo-classical inequality, we now prove the factorial decay property of Lyons' extension ([74], Theorem 4.8).

Theorem 1.8 (Factorial decay)**:** Consider $\alpha \in (0,1]$. Let $\mathbb{X} : \Delta_T^2 \to T^n(\mathbb{R}^d)$ be a multiplicative functional of degree n such that, for some constants $\beta, M > 0$, we have

$$\sup_{0\le s<t\le T} \frac{|\mathbb{X}^{(i)}_{s,t}|}{|t-s|^{\alpha i}} \le \frac{M^i}{\beta(\alpha i)!}, \quad \text{for all } i \in \{1,\dots,n\}. \tag{1.39}$$

If $\alpha(n+1) > 1$, then the unique extension $\mathbb{X} : \Delta_T^2 \to T((\mathbb{R}^d))$ has components $\mathbb{X}^{(i)}$ that satisfy (1.39) for all $i > n$.

Proof. Note that assumption (1.39) implies $\rho_{\alpha\text{-Höl};[0,T]}(\mathbb{X}, \mathbf{1}) < \infty$. Consequently, the assumptions of Theorem 1.6 are satisfied, ensuring that $(1, \mathbb{X}^{(1)}, \dots, \mathbb{X}^{(n)})$ is uniquely extended to a multiplicative functional taking values in $T((\mathbb{R}^d))$. We demonstrate that (1.39) leads to the same upper bound for $\mathbb{X}^{(n+1)}$, and the result then follows through an iterative argument.

Recall that $\mathbb{X}^{(n+1)}$ is given by (1.36). Hence, by the Sewing lemma (Theorem 1.4), we have

$$|\mathbb{X}^{(n+1)}_{s,t}| \le (1 + 2^{\alpha(n+1)}(\zeta(\alpha(n+1)) - 1))|t-s|^{\alpha(n+1)}|\delta\Xi|_{\alpha(n+1)}, \tag{1.40}$$

where Ξ is the almost additive functional given by (1.30). Additionally, from (1.34) and the neo-classical inequality (Theorem 1.7), we observe that

$$\begin{aligned} |\delta\Xi_{s,u,t}| \le \sum_{i=1}^{n} |\mathbb{X}^{(i)}_{s,u}| \cdot |\mathbb{X}^{(n+1-i)}_{u,t}| &\le \frac{M^{n+1}}{\beta^2} \sum_{i=1}^{n} \frac{(u-s)^{\alpha(n+1-i)}(t-u)^{\alpha i}}{(\alpha i)!(\alpha(n+1-i))!} \\ &\le \frac{M^{n+1}}{\alpha\beta^2} \frac{(t-s)^{\alpha(n+1)}}{(\alpha(n+1))!}. \end{aligned} \tag{1.41}$$

Therefore, from (1.40) and (1.41), we conclude that

$$\begin{aligned} |\mathbb{X}^{(n+1)}_{s,t}| &\le \frac{2^{\alpha(n+1)}[\zeta(\alpha(n+1)) - 1] + 1}{\alpha\beta^2} \cdot \frac{M^{n+1}}{(\alpha(n+1))!}|t-s|^{\alpha(n+1)} \\ &\le \frac{2^{\alpha(\lfloor 1/\alpha \rfloor + 1)}[\zeta(\alpha(\lfloor 1/\alpha \rfloor + 1)) - 1] + 1}{\alpha\beta^2} \cdot \frac{M^{n+1}}{(\alpha(n+1))!}|t-s|^{\alpha(n+1)}. \end{aligned}$$

For β satisfying

$$\beta \ge \frac{2^{\alpha(\lfloor 1/\alpha \rfloor + 1)}[\zeta(\alpha(\lfloor 1/\alpha \rfloor + 1)) - 1] + 1}{\alpha^2},$$

and $M > 0$ satisfying (1.39) the result follows. □

Theorem 1.9 (Continuity)**:** Consider $\alpha \in (0,1]$. Let $\mathbb{X}$ and $\mathbb{Y}$ be multiplicative functionals of degree n such that $\alpha(n+1) > 1$ and, for some constants $\beta, M > 0$, we have

$$\sup_{0\le s<t\le T} \frac{|\mathbb{X}^{(i)}_{s,t}|}{|t-s|^{\alpha i}}, \sup_{0\le s<t\le T} \frac{|\mathbb{Y}^{(i)}_{s,t}|}{|t-s|^{\alpha i}} \le \frac{M^i}{\beta(\alpha i)!} \text{ for all } i \in \{1,\dots,n\}. \tag{1.42}$$

Suppose further that, for some $\varepsilon < 1$, we have

$$\big|\mathbb{X}^{(i)}_{s,t} - \mathbb{Y}^{(i)}_{s,t}\big| \leq \varepsilon \frac{M^i}{\beta(\alpha i)!}|t-s|^{\alpha i}, \text{ for all } i \in \{1,\dots,n\} \text{ and } (s,t) \in \Delta^2_T. \tag{1.43}$$

Then, (1.43) holds for all $i \in \mathbb{N}$ where $\mathbb{X}^{(i)}$ and $\mathbb{Y}^{(i)}$, for $i > n$, are the components of the respective (unique) multiplicative extensions taking values in $T((\mathbb{R}^d))$.

Proof. We establish that Lyons' extension is locally Hölder continuous, and consequently, continuous. The key observation is that all steps in the proof of Theorem 1.6 are continuous, ensuring the continuity of the extension. Furthermore, it is sufficient to show the continuity of the map

$$\big(1, \mathbb{X}^{(1)}, \dots, \mathbb{X}^{(n)}, \mathbf{0}\big) \mapsto \Xi_{s,t} \in C_2^{\alpha,\alpha(n+1)}([0,T], (\mathbb{R}^d)^{\otimes(n+1)}),$$

where Ξ, as usual, denotes the almost additive functional given by (1.30). Indeed, the initial augmentation by $\mathbf{0} \in (\mathbb{R}^d)^{\otimes(n+1)}$ is continuous, and the operator $\mathcal{I} - \mathrm{Id}$ that maps $\Xi_{s,t}$ to $\mathbb{X}^{(n+1)}_{s,t}$ is linear and bounded by Theorem 1.4, hence continuous. In addition, if one wishes to consider $T((\mathbb{R}^d))$ as the codomain of the extension, it is well-established that the evaluation functional, which maps $\big(1, \mathbb{X}^{(1)}, \dots, \mathbb{X}^{(n+1)}, \dots\big)$ to $\big(1, \mathbb{X}^{(1)}_{0,T}, \dots, \mathbb{X}^{(n+1)}_{0,T}, \dots\big)$ in $T((\mathbb{R}^d))$, is continuous.

As such, let $\Xi^{\mathbb{X}}$ and $\Xi^{\mathbb{Y}}$ denote the almost additive functionals associated to $\mathbb{X}$ and $\mathbb{Y}$, respectively. We first bound $|\delta(\Xi^{\mathbb{X}} - \Xi^{\mathbb{Y}})|_{\alpha(n+1)}$. Specifically, by (1.34) we have that

$$\big|\delta(\Xi^{\mathbb{X}} - \Xi^{\mathbb{Y}})_{s,u,t}\big| = \left|\sum_{i=1}^{n} \Big(\mathbb{X}^{(i)}_{s,u} \otimes \mathbb{X}^{(n+1-i)}_{u,t} - \mathbb{Y}^{(i)}_{s,u} \otimes \mathbb{Y}^{(n+1-i)}_{u,t}\Big)\right|. \quad (\dagger)$$

With the goal of using the estimates (1.42) and (1.43), we note that, for all $0 \leq s < u < t \leq T$,

$$\begin{aligned}\big(\mathbb{X}^{(i)}_{s,u} \otimes \mathbb{X}^{(n+1-i)}_{u,t} - \mathbb{Y}^{(i)}_{s,u} \otimes \mathbb{Y}^{(n+1-i)}_{u,t}\big) &= \big(\mathbb{X}^{(i)}_{s,u} - \mathbb{Y}^{(i)}_{s,u}\big) \otimes \big(\mathbb{X}^{(n+1-i)}_{u,t} - \mathbb{Y}^{(n+1-i)}_{u,t}\big) + \\ &+ \big(\mathbb{X}^{(i)}_{s,u} - \mathbb{Y}^{(i)}_{s,u}\big) \otimes \mathbb{Y}^{(n+1-i)}_{u,t} + \mathbb{Y}^{(i)}_{s,u} \otimes \big(\mathbb{X}^{(n+1-i)}_{u,t} - \mathbb{Y}^{(n+1-i)}_{u,t}\big). \end{aligned} \tag{1.44}$$

Hence, by using the compatibility between tensor norms, we see that

$$\begin{aligned}(\dagger) \leq \sum_{i=1}^{n} \Big(&\big|\mathbb{X}^{(i)}_{s,u} - \mathbb{Y}^{(i)}_{s,u}\big| \cdot \big|\mathbb{X}^{(n+1-i)}_{u,t} - \mathbb{Y}^{(n+1-i)}_{u,t}\big| + \big|\mathbb{X}^{(i)}_{s,u} - \mathbb{Y}^{(i)}_{s,u}\big| \cdot |\mathbb{Y}^{(n+1-i)}_{u,t}| + \\ &+ \big|\mathbb{Y}^{(i)}_{s,u}\big| \cdot \big|\mathbb{X}^{(n+1-i)}_{u,t} - \mathbb{Y}^{(n+1-i)}_{u,t}\big|\Big) \\ \leq \sum_{i=1}^{n} &\frac{M^{n+1}}{\beta^2(\alpha i)!(\alpha(n+1-i))!}(\varepsilon^2 + 2\varepsilon)|t-s|^{\alpha(n+1)} \leq (\varepsilon^2 + 2\varepsilon)\frac{M^{n+1}}{\alpha\beta^2(\alpha(n+1))!}|t-s|^{\alpha(n+1)},\end{aligned}$$

where we used Theorem 1.7 in the final inequality. Thus, we conclude that $|\delta(\Xi^{\mathbb{X}} - \Xi^{\mathbb{Y}})|_{\alpha(n+1)}$ can be made arbitrary small by sending $\varepsilon \to 0$. It remains to show that $|\Xi^{\mathbb{X}} - \Xi^{\mathbb{Y}}|_{\alpha} \to 0$ as $\varepsilon \to 0$. To this end, observe that

$$\big|(\Xi^{\mathbb{X}} - \Xi^{\mathbb{Y}})_{s,t}\big| = \left|\sum_{i=1}^{n} \Big(\mathbb{X}^{(i)}_{0,s} \otimes \mathbb{X}^{(n+1-i)}_{s,t} - \mathbb{Y}^{(i)}_{0,s} \otimes \mathbb{Y}^{(n+1-i)}_{s,t}\Big)\right|,$$

and that (1.44) holds true for $s = 0$ and with u replaced by s. Hence, we may repeat the previous reasoning and the claim follows by aplying the bound $|t-s|^{\alpha(n+1)} \leq T^{\alpha n}|t-s|^{\alpha}$ at the end. □

1.3.5. Spaces of Rough Paths

In Section 1.3.3, we proved that a multiplicative functional $\mathbb{X} : \Delta_T^2 \to T^n(\mathbb{R}^d)$, exhibiting a certain degree of regularity parameterised by $\alpha \in (0,1]$, and defined up to a sufficiently high level n, can be uniquely extended to any level $N > n$. Particularly, we observed that the regularity of $\mathbb{X}$ is in relation to the dimension of its state-space $T^n(\mathbb{R}^d)$, necessitating the level n to be at least $\lfloor 1/\alpha \rfloor$. To all effects, a rough path is then a functional that precisely fulfils these criteria ([55], Definition 3.11).

In alignment with [37], however, we do justice to the term "path" and, in this section, establish precise definitions of rough path spaces. These spaces consist of paths taking values in $G^N(\mathbb{R}^d)$, i.e. the free nilpotent group defined in Section 1.2.3, where the level N is in direct relation to the regularity of the path. This regularity, in turn, is characterised by imposing constraints, either through a finite α-Holder norm or finite p-variation. Alongside the distance (1.38) introduced in the preceding section, we define new distances by modifying the metric over the state-space $G^N(\mathbb{R}^d)$. Ultimately, we demonstrate that these distances induce the same topology in their respective spaces.

A few observations are now in order. Firstly, by Theorem 1.8, the Lyons lift $\mathbb{X}$ of an appropriately regular multiplicative functional with factorially bounded components showcases the factorial decay in (1.39). As such, recalling Definition 1.9, we see that

$$\left|\mathbb{X}_{0,T}\right|_{\tilde{T}((\mathbb{R}^d))} = \sqrt{\sum_{n=0}^{\infty} \left|\mathbb{X}_{0,T}^{(n)}\right|^2} \leq \sum_{n=0}^{\infty} \frac{(T^\alpha M)^n}{\beta(\alpha n)!} < \infty,$$

indicating that $\mathbb{X}_{0,T} \in \tilde{T}((\mathbb{R}^d))$. Moreover, by definition, we have $\mathbb{X}_{0,T}^{(0)} = 1$, and thus $\mathbb{X}_{0,T} \in T_1((\mathbb{R}^d))$, restricting the state-space of $\mathbb{X}$ to a subspace of $T((\mathbb{R}^d))$. In what follows, we further confine the state-space of rough paths to $G^N(\mathbb{R}^d)$. This restriction is justified when recalling that, by definition, $G^N(\mathbb{R}^d)$ corresponds to elements of $T^N(\mathbb{R}^d)$ representing the (evaluated) signature of some continuous 1-variation path. Additionally, as discussed in Section 1.3.1, $\mathbb{X}^{(n)}$ is to be interpreted as an n-fold iterated integral.

Secondly, once equipped with a (rough) path $\mathbf{X} : [0,T] \to G^N(\mathbb{R}^d)$, a multiplicative functional is readily obtained by considering $\mathbf{X}_{s,t} := \mathbf{X}_s^{-1} \otimes \mathbf{X}_t$ with $s < t$ in $[0,T]$. Regarding the regularity constraints on $\mathbf{X}_{s,t}$ to ensure a unique Lyons lift, these will directly follow from the regularity of $\mathbf{X}$

Thirdly, we observe that thus far we have utilised α-Hölder continuity to establish regularity constraints on the multiplicative functionals. However, the results from the preceding sections are not inherently dependent on α-Hölder continuity. Alternatively, we could impose regularity constraints using p-variation. Recall that according to Proposition 1.3, any continuous p-variation path can be considered a $1/p$-Hölder path up to a time-change.

Remark 1.12: In [53] and [37], the authors introduce regularity constraints through the concept of *controls*. In essence, controls provide a broad formalisation of the idea that path increments $X_{s,t}$ scale continuously based on some function of $|t-s|$. For details, refer to Section 1.2 in [37].

Finally, when considering a path $\mathbf{X} : [0,T] \to G^N(\mathbb{R}^d)$, irrespective of whether we employ α-Hölder continuity or p-variation, we encounter a choice: determining the metric structure of $G^N(\mathbb{R}^d)$. Different metrics within $G^N(\mathbb{R}^d)$ yield distinct notions of distance for a space of paths taking values in $G^N(\mathbb{R}^d)$. If the distances in the path spaces depend solely on the metric over $G^N(\mathbb{R}^d)$, then, according to Proposition 1.20, our expectation is that the path spaces share, at the very least, the same topology. All the above considerations motivate the following definitions.

Definition 1.21: Consider $p \geq 1$ and $\alpha \in (0,1]$. Let $\mathbf{X}, \mathbf{Y} \in C([0,T], G^N(\mathbb{R}^d))$ and set $\mathbf{X}_{s,t} := \mathbf{X}_s^{-1} \otimes \mathbf{X}_t$. We define the following distances:

1. The homogenous α-Hölder distance,

$$d_{\alpha\text{-Höl};[0,T]}(\mathbf{X}, \mathbf{Y}) := \sup_{0 \leq s < t \leq T} \frac{d_{cc}(\mathbf{X}_{s,t}, \mathbf{Y}_{s,t})}{|t-s|^\alpha};$$

2. The homogenous p-variation distance,

$$d_{p\text{-var};[0,T]}(\mathbf{X}, \mathbf{Y}) := \left(\sup_{(t_i) \in \mathcal{P}([0,T])} \sum_i d_{cc}\big(\mathbf{X}_{t_i,t_{i+1}}, \mathbf{Y}_{t_i,t_{i+1}}\big)^p \right)^{1/p};$$

3. The inhomogenous α-Hölder distance,

$$\rho_{\alpha\text{-Höl};[0,T]}(\mathbf{X}, \mathbf{Y}) := \max_{n=1,\ldots,N} \sup_{0 \leq s < t \leq T} \frac{\big|\mathbf{X}_{s,t}^{(n)} - \mathbf{Y}_{s,t}^{(n)}\big|}{|t-s|^{\alpha n}};$$

4. The inhomogenous p-variation distance,

$$\rho_{p\text{-var};[0,T]}(\mathbf{X}, \mathbf{Y}) := \max_{n=1,\ldots,N} \sup_{(t_i) \in \mathcal{P}([0,T])} \left(\sum_i \big|\mathbf{X}_{t_i,t_{i+1}}^{(n)} - \mathbf{Y}_{t_i,t_{i+1}}^{(n)}\big|^{p/n} \right)^{n/p};$$

These distances are commonly encountered in the literature, particularly in the context of signature-based methods. It is important to note that, similar to $\rho_{\alpha\text{-Höl};[0,T]}$, the distances mentioned above are not genuine metrics. For instance, $d_{\alpha\text{-Höl};[0,T]}(\mathbf{X},\mathbf{Y})=0$ if and only if $\mathbf{X}_t = c\otimes\mathbf{Y}_t$, where $c=\mathbf{X}_0\otimes\mathbf{Y}_0^{-1}$.

However, genuine metrics can be obtained if we focus on paths with a fixed starting point, as is the case with $C_o^{\alpha\text{-Höl}}([0,T],G^N(\mathbb{R}^d))$ or $C_o^{\text{p-var}}([0,T],G^N(\mathbb{R}^d))$ (recall Definition 1.4), or if we include the distance between the starting points and consider $\tilde{d}_{\alpha\text{-Höl};[0,T]}(\mathbf{X},\mathbf{Y}) \coloneqq d_{cc}(\mathbf{X}_0,\mathbf{Y}_0)+d_{\alpha\text{-Höl};[0,T]}(\mathbf{X},\mathbf{Y})$ instead. Similarly, following the notation of Section 1.2.3, we define $\tilde{\rho}_{\alpha\text{-Höl};[0,T]}(\mathbf{X},\mathbf{Y}) \coloneqq \rho(\mathbf{X}_0,\mathbf{Y}_0)+\rho_{\alpha\text{-Höl};[0,T]}(\mathbf{X},\mathbf{Y})$, as well as $\tilde{d}_{p\text{-var};[0,T]}$, and $\tilde{\rho}_{p\text{-var};[0,T]}$ in a completely analogous manner.

It is also noteworthy that all the aforementioned distances induce (semi-)norms by fixing $\mathbf{Y}$ to some constant path. Particularly, we have that $d_{\alpha\text{-Höl};[0,T]}(\mathbf{X},\mathbf{1}) \equiv |\mathbf{X}|_{\alpha\text{-Höl};[0,T]}$ and $d_{p\text{-var};[0,T]}(\mathbf{X},\mathbf{1}) \equiv |\mathbf{X}|_{p\text{-var};[0,T]}$, leading to the spaces in Definition 1.4 with $(E,d)=(G^N(\mathbb{R}^d),d_{cc})$. Moreover, since

$$|\mathbf{X}|_{\alpha\text{-Höl};[0,T]}<\infty \text{ iff } \rho_{\alpha\text{-Höl};[0,T]}(\mathbf{X},\mathbf{1})<\infty \quad\text{and}\quad |\mathbf{X}|_{p\text{-var};[0,T]}<\infty \text{ iff } \rho_{p\text{-var};[0,T]}(\mathbf{X},\mathbf{1})<\infty,$$

the sets $C^{\alpha\text{-Höl}}([0,T],G^N(\mathbb{R}^d))$ and $C^{p\text{-var}}([0,T],G^N(\mathbb{R}^d))$ coincide regardless of whether a homogenous or inhomogenous distance is used.

As previously suggested, within the context of Rough Path theory, paths whose regularity is in relation to the truncation level of the state-space are of particular significance. This forms the basis of the following definition.

Definition 1.22: Consider $p\geq 1$ and $\alpha\in(0,1]$. A weakly geometric α-Hölder rough path is an α-Hölder path with values in the free nilpotent group of step $\lfloor 1/\alpha\rfloor$, i.e. an element of $C^{\alpha\text{-Höl}}([0,T],G^{\lfloor 1/\alpha\rfloor}(\mathbb{R}^d))$. A weakly geometric p-rough path is a continuous path of finite p-variation with values in the free nilpotent group of step $\lfloor p\rfloor$, i.e. an element of $C^{p\text{-var}}([0,T],G^{\lfloor p\rfloor}(\mathbb{R}^d))$.

If $X:[0,T]\to\mathbb{R}^d$ is an α-Hölder path and $\mathbf{X}$ is a (weakly geometric) α-Hölder rough path such that $\mathbf{X}^{(1)}_{s,t}=X_{s,t}$, then we call $\mathbf{X}$ a *rough path lift* of X. Analogously for p-rough paths. As a way to naturally generalise the signatures of bounded variation paths discussed in Section 1.2, it is sensible to consider the closure of the "classical signatures." The term "weakly" serves to distinguish the paths of Definition 1.22 from those belonging to this closure.

Definition 1.23: We denote by $C_o^{0,\alpha\text{-Höl}}([0,T],G^{\lfloor 1/\alpha\rfloor}(\mathbb{R}^d))$ the set of continuous paths $\mathbf{X}:[0,T]\to G^{\lfloor 1/\alpha\rfloor}(\mathbb{R}^d)$ for which there exists a sequence of smooth $\mathbb{R}^d$-valued paths X_n such that

$$d_{\alpha\text{-Höl};[0,T]}(\mathbf{X},S_{\lfloor\frac{1}{\alpha}\rfloor}(X_n))\to 0,\ \text{ as } n\to\infty,$$

and define $C^{0,\alpha\text{-Höl}}([0,T],G^{\lfloor 1/\alpha\rfloor}(\mathbb{R}^d))$ as the set of paths $\mathbf{X}$ such that

$$\mathbf{X}_{0,\cdot}\coloneqq\mathbf{X}_0^{-1}\otimes\mathbf{X}_\cdot\in C_o^{0,\alpha\text{-Höl}}([0,T],G^{\lfloor 1/\alpha\rfloor}(\mathbb{R}^d)).$$

The elements of $C^{0,\alpha\text{-Höl}}([0,T], G^{\lfloor 1/\alpha \rfloor}(\mathbb{R}^d))$ are referred to as geometric α-Hölder rough paths. Analogously, $C_o^{0,p\text{-var}}([0,T], G^{\lfloor p \rfloor}(\mathbb{R}^d))$ denotes the set of continuous paths $\mathbf{X} : [0,T] \to G^{\lfloor p \rfloor}(\mathbb{R}^d)$ for which there exists a sequence of smooth $\mathbb{R}^d$-valued paths X_n such that

$$d_{p\text{-var};[0,T]}(\mathbf{X}, S_{\lfloor p \rfloor}(X_n)) \to 0, \text{ as } n \to \infty,$$

and $C^{0,p\text{-var}}([0,T], G^{\lfloor p \rfloor}(\mathbb{R}^d))$ is defined as the set of paths with $\mathbf{X}_{0,\cdot} \in C_o^{0,p\text{-var}}([0,T], G^{\lfloor p \rfloor}(\mathbb{R}^d))$. The elements of $C^{0,p\text{-var}}([0,T], G^{\lfloor p \rfloor}(\mathbb{R}^d))$ are referred to as geometric p-rough paths.

Remark 1.13: The term "geometric" signifies that solutions of differential equations driven by geometric rough paths satisfy the usual chain rule. As a result, the solution of a differential equation with values in a manifold transforms as expected under a change of coordinates, thus preserving the geometry of the underlying manifold. This remark simply underscores the rationale behind the terminology.

Clearly, $C^{0,\alpha\text{-Höl}}([0,T], G^{\lfloor 1/\alpha \rfloor}(\mathbb{R}^d)) \subset C^{\alpha\text{-Höl}}([0,T], G^{\lfloor 1/\alpha \rfloor}(\mathbb{R}^d))$, and the same holds true for p-rough paths. This inclusion can be shown to be strict, and the spaces of geometric rough paths are Polish concerning the appropriate distances, either homogenous or inhomogenous ([37], Proposition 8.25). We now show that homogenous and inhomogenous distances induce the same topology.

Theorem 1.10 ([37], Theorem 8.10)**:** Consider $p \geq 1$ and $\alpha \in (0,1]$. Each identity map:

$$\begin{aligned}
\text{Id} : \left(C^{\alpha\text{-Höl}}([0,T], G^N(\mathbb{R}^d)), \tilde{d}_{\alpha\text{-Höl};[0,T]}\right) &\leftrightarrow \left(C^{\alpha\text{-Höl}}([0,T], G^N(\mathbb{R}^d)), \tilde{\rho}_{\alpha\text{-Höl};[0,T]}\right) \\
\text{Id} : \left(C^{p\text{-var}}([0,T], G^N(\mathbb{R}^d)), \tilde{d}_{p\text{-var};[0,T]}\right) &\leftrightarrow \left(C^{p\text{-var}}([0,T], G^N(\mathbb{R}^d)), \tilde{\rho}_{p\text{-var};[0,T]}\right)
\end{aligned}$$

is Lipschitz on bounded sets in the $\to$ direction, and α-Hölder continuous on bounded sets in the $\leftarrow$ direction. In particular, this is true for α-Hölder rough paths ($N = \lfloor 1/\alpha \rfloor$), and p-rough paths ($N = \lfloor p \rfloor$).

Proof. We remain consistent with our preference to work primarily with α-Hölder spaces and establish the assertion for the first identity map. Note that the estimates between $d_{cc}(\mathbf{X}_0, \mathbf{Y}_0)$ and $\rho(\mathbf{X}_0, \mathbf{Y}_0)$ are directly derived from Proposition 1.20. Therefore, it suffices to analyse the distances in the path space without the tilde notation. Recall that, for $\lambda \in \mathbb{R}$, we denote by δ_λ the dilation operator. Additionally, observe that $\delta_\lambda \mathbf{a} \otimes \delta_\lambda \mathbf{b} = \delta_\lambda(\mathbf{a} \otimes \mathbf{b})$, for $\mathbf{a}, \mathbf{b} \in T^N(\mathbb{R}^d)$. Hence, we see that

$$\begin{aligned}
d_{\alpha\text{-Höl};[0,T]}(\mathbf{X}, \mathbf{Y}) &= \sup_{0 \leq s < t \leq T} d_{cc}\left(\delta_{\frac{1}{|t-s|^\alpha}} \mathbf{X}_{s,t}, \delta_{\frac{1}{|t-s|^\alpha}} \mathbf{Y}_{s,t}\right), \text{ and} \\
\rho_{\alpha\text{-Höl};[0,T]}(\mathbf{X}, \mathbf{Y}) &= \sup_{0 \leq s < t \leq T} \left|\delta_{\frac{1}{|t-s|^\alpha}} \mathbf{X}_{s,t} - \delta_{\frac{1}{|t-s|^\alpha}} \mathbf{Y}_{s,t}\right|.
\end{aligned}$$

This implies that these distances depend solely on the metric over $G^N(\mathbb{R}^d)$, for which Proposition 1.20 provides estimates. Therefore,

$$d_{\alpha\text{-Höl};[0,T]}(\mathbf{X},\mathbf{Y}) \le C \max\left\{\rho_{\alpha\text{-Höl};[0,T]}(\mathbf{X},\mathbf{Y}),\ \rho_{\alpha\text{-Höl};[0,T]}(\mathbf{X},\mathbf{Y})^{1/N} \max\left\{1, |\mathbf{X}|^{1-\frac{1}{N}}_{\alpha\text{-Höl};[0,T]}\right\}\right\},$$
$$\rho_{\alpha\text{-Höl};[0,T]}(\mathbf{X},\mathbf{Y}) \le C \max\left\{d_{\alpha\text{-Höl};[0,T]}(\mathbf{X},\mathbf{Y}) \max\left\{1, |\mathbf{X}|^{N-1}_{\alpha\text{-Höl};[0,T]}\right\},\ d_{\alpha\text{-Höl};[0,T]}(\mathbf{X},\mathbf{Y})^N\right\},$$

and since it suffices to examine the unit ball case, the result follows. □

Recall that for a given topological space Z, the identity map $\mathrm{Id} : (Z, \tau_1) \to (Z, \tau_2)$ is continuous if and only if $\tau_2 \subset \tau_1$. Hence, according to Theorem 1.10, we can equip $C^{\alpha\text{-Höl}}([0,T], G^N(\mathbb{R}^d))$ with either the homogenous or the inhomogenous distance, and the induced topologies are identical. Utilising the rough path spaces in Definition 1.22, we thus obtain, by Theorems 1.6 and 1.9, a continuous map

$$S^\alpha : C_o^{\alpha\text{-Höl}}([0,T], G^{\lfloor 1/\alpha \rfloor}(\mathbb{R}^d)) \to T_1((\mathbb{R}^d)),$$

regardless of whether $d_{\alpha\text{-Höl};[0,T]}$ or $\rho_{\alpha\text{-Höl};[0,T]}$ is chosen as the path distance. We deliberately denote this map by S^α to emphasise not only that the topology is the α-topology induced by the chosen metric, but also to underscore that this map is an analogue for lower regularity paths to the path signatures discussed in Section 1.2.1.

Definition 1.24: Consider $C_o^{\alpha\text{-Höl}}([0,T], G^{\lfloor 1/\alpha \rfloor}(\mathbb{R}^d))$ equipped with the α-topology induced by either $d_{\alpha\text{-Höl};[0,T]}$ or $\rho_{\alpha\text{-Höl};[0,T]}$. Let $\mathbf{X}$ be a weakly geometric α-Hölder rough path. We denote the mapping $\mathbf{X} \mapsto (1, \mathbf{X}^{(1)}_{0,T}, \dots, \mathbf{X}^{(\lfloor 1/\alpha \rfloor)}_{0,T}, \dots, \mathbf{X}^{(N)}_{0,T}, \dots)$ by S^α. In other words, S^α maps $\mathbf{X}$ to its (evaluated) Lyons' extension $\mathbf{X}_{0,T} \in T_1((\mathbb{R}^d))$. We refer to $S^\alpha(\mathbf{X})$ as the signature of $\mathbf{X}$.

We conclude this chapter with a rather trivial, but nonetheless pertinent observation that seems to be overlooked in the literature: the rough path space topology is not initial with respect to S^α. While the compact embedding of rough path spaces is a well-known technique in the literature, it is worth noting that the next result pertains specifically to the continuity of the signature map. This distinction is crucial for our subsequent discussions in Chapter 3.

Proposition 1.23: For any $\alpha \in (0,1]$, there exists $\beta < \alpha$ such that $C_o^{\alpha\text{-Höl}}([0,T], G^{\lfloor \frac{1}{\alpha} \rfloor}(\mathbb{R}^d))$ equipped with the β-topology induced by either $d_{\beta\text{-Höl};[0,T]}$ or $\rho_{\beta\text{-Höl};[0,T]}$ makes the signature map $C_o^{\alpha\text{-Höl}}([0,T], G^{\lfloor 1/\alpha \rfloor}(\mathbb{R}^d)) \to T_1((\mathbb{R}^d))$ continuous.

Proof. For all $\alpha \in (0,1]$ there exists $\varepsilon > 0$ such that $\lfloor 1/(\alpha - \varepsilon) \rfloor = \lfloor 1/\alpha \rfloor$. Set $\beta = \alpha - \varepsilon$. Then, by Theorem 1.9, we know that $S^\beta : C_o^{\beta\text{-Höl}}([0,T], G^{\lfloor 1/\beta \rfloor}(\mathbb{R}^d)) \to T_1((\mathbb{R}^d))$ is continuous, and, by Proposition 1.5, we have that $C_o^{\alpha\text{-Höl}}([0,T], G^{\lfloor 1/\alpha \rfloor}(\mathbb{R}^d))$ is continuously embedded in $C_o^{\beta\text{-Höl}}([0,T], G^{\lfloor 1/\beta \rfloor}(\mathbb{R}^d))$. Hence, the result follows. □

2 Kernel Theory

The general theory of reproducing kernels is vast, and, to paraphrase the authors in [66], "as important as the concept of Hilbert spaces." What is more, the theory plays a role in a remarkable number of areas, which include: Complex Analysis, operator theory, group representation theory, approximation theory, Statistics, Probability, and Machine Learning. It is within the context of the latter, or more broadly, within the realm of learning methods, that kernel theory assumes importance in the present work.

As is often the case in Mathematics, statistical and Machine Learning methods have been well-developed for the linear case, when there are linear dependencies in the data or when the data is linearly separable. However, in real-world applications, nonlinear methods are often necessary to effectively detect dependencies in the data that enable, for example, accurate predictions. With this in mind, kernels can be viewed as a means to address nonlinearities in a linear manner.

More precisely, the idea is to implicitly handle the nonlinearities by mapping our data/input space into a *feature space*, typically a high-dimensional space that furnishes us with more expressive features. This mapping is achieved via a *feature map*, and a kernel then corresponds to the inner product in the feature space. Remarkably, the kernel can often be computed without explicitly computing, or even knowing, the feature map. Thus, any learning method that we define is linear, as long as we formulate it in terms of kernel evaluations.

This chapter provides an elementary introduction to kernel theory, assuming only a basic understanding of Functional Analysis. The selection of topics is based on their relevance to kernel theory, as well as their common occurrence in the literature of signature-based methods. Section 2.1 outlines the fundamental concepts of kernel theory, while Section 2.2 explores the application of kernel theory tools to approximation problems. Additionally, Section 2.3 offers an interpretation of the signature as a feature map and reviews a specific kernel, the *signature kernel*, along with its key properties. Notably, within the discussion of the signature kernel, we highlight a property, albeit basic, that seems to have been overlooked in the literature.

It is worth noting that this chapter serves as a literal bridge between Chapters 1 and 3. Indeed, Section 2.3 utilises the tools introduced in Chapter 1 to define a kernel, while Section 2.2 introduces a central concept to this thesis: *universality*. Path signatures embody this notion of universality, explaining why they have found considerable success in various applications. Moreover, universality motivates much of the discussion in the upcoming Chapter 3.

2.1. Basics of Kernel Theory

We present the fundamentals of kernel theory, starting with key concepts like kernels, feature maps, and feature spaces. During our exploration, we emphasise that kernels can be defined in various equivalent ways. Additionally, we introduce reproducing kernel Hilbert spaces (RKHS) and elucidate their connection to kernels. We conclude this section by presenting some useful properties of RKHSs.

2.1.1. Kernels and Reproducing Kernel Hilbert Spaces

We begin by introducing a kernel definition that closely aligns with the motivation provided in the chapter's introduction. This definition will be used consistently throughout our discussion. As previously noted, there is often interest in mapping a given input or data space into a higher dimensional space equipped with an inner product. In this context, kernels can be viewed as functions that realise this inner product in the higher dimensional space. Keeping upcoming sections in view, we also introduce a specific type of kernels known as *Taylor kernels*. We focus solely on real-valued kernels.

Definition 2.1: Let X be a non-empty set. A function $k : X \times X \to \mathbb{R}$ is said to be a kernel if there exists a (real) Hilbert space $\mathcal{H}_0$ and a map $\Phi : X \to \mathcal{H}_0$ such that, for all $x, x' \in X$,

$$k(x, x') = \langle \Phi(x), \Phi(x') \rangle_{\mathcal{H}_0}.$$

We say that Φ is a feature map and $\mathcal{H}_0$ is a feature space of k.

A straightforward example of a kernel is the inner product of any real Hilbert space $\mathcal{H}$. Specifically, for $h, h' \in \mathcal{H}$, the inner product map $(h, h') \mapsto \langle h, h' \rangle_{\mathcal{H}}$ is a kernel, with the identity function $\mathrm{Id} : \mathcal{H} \to \mathcal{H}$ as the feature map. It is noteworthy to observe that feature maps and feature spaces are not unique. For instance, the kernel $(h, h') \mapsto \langle h, h' \rangle_{\mathcal{H}}$ also admits the function $\Phi' : \mathcal{H} \to \mathcal{H} \times \mathcal{H}$ defined by $\Phi'(h) := (h/\sqrt{2}, h/\sqrt{2})$ as a feature map. This can be verified by observing that

$$\langle \Phi'(h), \Phi'(h') \rangle_{\mathcal{H}\times\mathcal{H}} = \frac{1}{2}\langle h, h' \rangle_{\mathcal{H}} + \frac{1}{2}\langle h, h' \rangle_{\mathcal{H}} = k(h, h'),$$

where $\langle (h_1, h_2), (h_1', h_2') \rangle_{\mathcal{H}\times\mathcal{H}} := \langle h_1, h_1' \rangle_{\mathcal{H}} + \langle h_2, h_2' \rangle_{\mathcal{H}}$ is the inner product in $\mathcal{H} \times \mathcal{H}$. For additional examples of kernels, refer to Section 1.2 in [61] and Section 4.1 in [71]. We now specialise in Taylor type kernels, which will allow us to construct several kernels and will play a prominent role in the next chapter.

Proposition 2.1 ([71], Lemma 4.8)**:** Consider some $r \in (0, \infty]$ and let $\sqrt{r}B_{\mathbb{R}^d}$ denote the closed ball of radius $\sqrt{r}$ in $\mathbb{R}^d$. Moreover, consider a function $f : [-r, r] \to \mathbb{R}$ that can be expressed by its Taylor series, that is

$$f(t) = \sum_{n=0}^{\infty} a_n t^n, \text{ for all } t \in [-r, r].$$

If the Taylor coefficients a_n are non-negative, i.e. $a_n \geq 0$ for all $n \in \mathbb{N}_0$, then

$$k(x, x') := f(\langle x, x' \rangle_{\mathbb{R}^d}) = \sum_{n=0}^{\infty} a_n \langle x, x' \rangle_{\mathbb{R}^d}^n$$

defines a kernel on $\sqrt{r} B_{\mathbb{R}^d}$. We say that k is a Taylor type kernel.

Proof. For $x, x' \in \sqrt{r} B_{\mathbb{R}^d}$, we have by Cauchy-Schwarz that $|\langle x, x' \rangle_{\mathbb{R}^d}| \leq |x| \cdot |x'| \leq r$, ensuring that k is well-defined. Let x_i denote the i-th component of $x \in \mathbb{R}^d$. The claim follows by the multinomial formula. Note that

$$\begin{aligned} k(x, x') = \sum_{n=0}^{\infty} a_n \left(\sum_{i=1}^{d} x_i x_i' \right)^n &= \sum_{n=0}^{\infty} a_n \sum_{\substack{i_1, \ldots, i_d \geq 0 \\ i_1 + \cdots + i_d = n}} \frac{n!}{i_1! i_2! \cdots i_d!} \prod_{k=1}^{d} (x_i x_i')^{i_k} \\ &= \sum_{i_1, \ldots, i_d \geq 0} a_{i_1 + \cdots + i_d} \frac{(i_1 + \cdots + i_d)!}{i_1! i_2! \cdots i_d!} \prod_{k=1}^{d} (x_i x_i')^{i_k}. \end{aligned}$$

Now, set $c_{i_1, \ldots, i_d} := a_{i_1 + \cdots + i_d} \frac{n!}{\prod_{k=1}^{d} i_k!}$ with $n = i_1 + \cdots + i_d$, and consider the space of square summable sequences indexed by $\mathbb{N}_0^d$, denoted by $l_2(\mathbb{N}_0^d)$. The map $\Phi : \sqrt{r} B_{\mathbb{R}^d} \to l_2(\mathbb{N}_0^d)$ defined by

$$\Phi(x) := \left(\sqrt{c_{i_1, \ldots, i_d}} \prod_{k=1}^{d} x_i^{i_k} \right)_{i_1, \ldots, i_d \geq 0},$$

constitutes a feature map, i.e. $k(x, x') = \langle \Phi(x), \Phi(x') \rangle_{l_2(\mathbb{N}_0^d)}$, and hence, k is a kernel. □

We proceed to examine the more functional analytical aspects of kernels. In the next definition, $\mathcal{F}(X, \mathbb{R})$ denotes the set of functions from X to $\mathbb{R}$. The set $\mathcal{F}(X, \mathbb{R})$ is evidently a real vector space with the operations of addition defined as $(f + g)(x) := f(x) + g(x)$, and scalar multiplication given by $(\lambda f)(x) := \lambda f(x)$.

Definition 2.2: Let X be a non-empty set. Consider $\mathcal{H}$ to be a Hilbert function space over X, i.e. $\mathcal{H}$ is a Hilbert space and $\mathcal{H} \subset \mathcal{F}(X, \mathbb{R})$.

1. The space $\mathcal{H}$ is said to be a reproducing kernel Hilbert space (RKHS) over X if for all $x \in X$ the evaluation functional $\delta_x : \mathcal{H} \to \mathbb{R}$ defined by $\delta_x(f) := f(x)$ is bounded.
2. A map $K : X \times X \to \mathbb{R}$ is said to be a reproducing kernel for $\mathcal{H}$ if for all $x \in X$ we have $K(\cdot, x) \in \mathcal{H}$, and the so-called *reproducing property*,

$$\langle f, K(\cdot, x) \rangle_{\mathcal{H}} = f(x),$$

   holds for all $f \in \mathcal{H}$ and $x \in X$. We shall often denote $K(\cdot, x)$ by $K_x(\cdot)$, or simply K_x.

Remark 2.1: It directly follows from Definition 2.2 that within a given RKHS $\mathcal{H}$, convergence in norm implies pointwise convergence. Specifically, if $(f_n) \subset \mathcal{H}$ is a sequence such that $|f - f_n|_{\mathcal{H}} \to 0$ as $n \to \infty$ for some $f \in \mathcal{H}$, then owing to the assumed continuity of evaluation functionals, we have that, for all $x \in X$,

$$\lim_{n\to\infty} f_n(x) = \lim_{n\to\infty} \delta_x(f_n) = \delta_x(f) = f(x).$$

Observe that if $\mathcal{H}$ represents a RKHS, then we can derive a reproducing kernel for $\mathcal{H}$ using the Riesz representation theorem. Indeed, if $\delta_x : \mathcal{H} \to \mathbb{R}$ is a bounded (and consequently continuous) linear functional, then the Riesz representation theorem (see Appendix B) guarantees the existence of a unique vector $K_x \in \mathcal{H}$ such that $f(x) = \delta_x(f) = \langle f, K_x \rangle_{\mathcal{H}}$. By defining $K : X \times X \to \mathbb{R}$ as $K(x, x') := K_{x'}(x)$, we establish a reproducing kernel for $\mathcal{H}$. The subsequent result shows that K is in reality the unique reproducing kernel for $\mathcal{H}$.

Proposition 2.2 ([71], Theorem 4.20)**:** Let $\mathcal{H}$ denote a RKHS over X. Then, $K : X \times X \to \mathbb{R}$ defined by

$$K(x, x') := \langle K_{x'}, K_x \rangle_{\mathcal{H}} = \langle K_x, K_{x'} \rangle_{\mathcal{H}}, \text{ for all } x, x' \in X,$$

where K_x denotes the Riesz representer of the evaluation functional $\delta_x \in \mathcal{H}^*$, is the only reproducing kernel for $\mathcal{H}$.

Proof. The fact that K is a reproducing kernel for $\mathcal{H}$ follows essentially by construction. As discussed above, K_x denotes the unique Riesz representer of $\delta_x : \mathcal{H} \to \mathbb{R}$. Consequently,

$$K(x, x') = \langle K_{x'}, K_x \rangle_{\mathcal{H}} = \delta_x(K_{x'}) = K_{x'}(x)$$

which shows that, for all $f \in \mathcal{H}$ and $x' \in X$,

$$f(x') = \delta_{x'}(f) = \langle f, K_{x'} \rangle_{\mathcal{H}} = \langle f, K(\cdot, x') \rangle_{\mathcal{H}},$$

i.e., K has the reproducing property. To show the symmetry of K, let $\mathcal{H}^*$ denote the dual of $\mathcal{H}$ and recall that the map corresponding each dual element in $\mathcal{H}^*$ to its Riesz representer is an isometry (Appendix B). Hence, $\langle K_{x'}, K_x \rangle_{\mathcal{H}} = \langle \delta_{x'}, \delta_x \rangle_{\mathcal{H}^*} = \langle \delta_x, \delta_{x'} \rangle_{\mathcal{H}^*} = \langle K_x, K_{x'} \rangle_{\mathcal{H}}$.

Lastly, assume that $\tilde{K}$ is another reproducing kernel for $\mathcal{H}$. Then, for all $x \in X$,

$$\begin{aligned} |K_x - \tilde{K}_x|^2_{\mathcal{H}} =& \langle K_x, K_x \rangle_{\mathcal{H}} + \langle \tilde{K}_x, \tilde{K}_x \rangle_{\mathcal{H}} - \langle K_x, \tilde{K}_x \rangle_{\mathcal{H}} - \langle \tilde{K}_x, K_x \rangle_{\mathcal{H}} \\ =& K(x, x) + \tilde{K}(x, x) - K_x(x) - \tilde{K}_x(x) = 0, \end{aligned}$$

and uniqueness follows. □

The following two results demonstrate the relationship between reproducing kernels (as defined in Definition 2.2) and kernels (as defined in Definition 2.1). First, we show that reproducing kernels are kernels in the sense of Definition 2.1. Subsequently, we show that every kernel has a unique RKHS. The latter result, as inferred from Proposition 2.2, implies that each kernel corresponds to a unique reproducing kernel. Ultimately, this shows the existence of a one-to-one correspondence between kernels and RKHSs.

Proposition 2.3 ([71], Lemma 4.19)**:** Let $\mathcal{H}$ be a Hilbert function space with reproducing kernel K. Then, $\mathcal{H}$ is a RKHS and a feature space of K. Specifically, the map $\Phi_c : X \to \mathcal{H}$ given by

$$\Phi_c(x) := K(\cdot, x), \text{ for all } x \in X,$$

is a feature map, rendering K a kernel. We call Φ_c the canonical feature map.

Proof. By the reproducing property, every evaluation functional can be represented by $K(\cdot, x)$. Thus, by Cauchy-Schwarz, we obtain

$$|\delta_x(f)| = |f(x)| = |\langle f, K(\cdot, x)\rangle_{\mathcal{H}}| \leq |K(\cdot, x)|_{\mathcal{H}} \cdot |f|_{\mathcal{H}},$$

for all $x \in X$ and $f \in \mathcal{H}$, implying that all evaluation functionals are bounded. Lastly, to show that Φ_c is a feature map, simply notice that

$$\langle \Phi_c(x'), \Phi_c(x)\rangle_{\mathcal{H}} = \langle K(\cdot, x'), K(\cdot, x)\rangle_{\mathcal{H}} = K_{x'}(x) = K(x, x'),$$

by the reproducing property. This concludes the proof. □

Theorem 2.1 ([71], Theorem 4.21)**:** Consider a non-empty set X. Let k be a kernel over X with feature space $\mathcal{H}_0$ and feature map $\Phi_0 : X \to \mathcal{H}_0$. Then,

$$\mathcal{H} := \left\{ f : X \to \mathbb{R} : \exists h \in \mathcal{H}_0 \text{ with } f(x) = \langle h, \Phi_0(x)\rangle_{\mathcal{H}_0}, \text{ for all } x \in X \right\}, \tag{2.1}$$

equipped with the norm

$$|f|_{\mathcal{H}} := \inf \left\{ |h|_{\mathcal{H}_0} : h \in \mathcal{H}_0 \text{ with } f = \langle h, \Phi_0(\cdot)\rangle_{\mathcal{H}_0} \right\}$$

is the only RKHS for which k is a reproducing kernel.

Proof. We commence by showing that $\mathcal{H}$ is a Hilbert function space over X. The fact that $\mathcal{H}$ is a vector space of functions from X to $\mathbb{R}$ is clear from (2.1). Consider the operator $V : \mathcal{H}_0 \to \mathcal{H}$ defined by $V(h) := \langle h, \Phi_0(\cdot)\rangle_{\mathcal{H}_0}$. By definition, V is a surjective linear operator, and we have that

$$|f|_{\mathcal{H}} \equiv \inf_{h \in V^{-1}(f)} |h|_{\mathcal{H}_0}.$$

We will show that $|\cdot|_{\mathcal{H}}$ is a Hilbert space norm on $\mathcal{H}$ by showing that $\mathcal{H}$ is isometrically isomorphic to a Hilbert space. To this end, let $(h_n)_{n\geq 1} \subset \ker V$ be a convergent sequence in the null space of V. Denote its limit by h. Given that $0 = \langle h_n, \Phi(x)\rangle_{\mathcal{H}_0} \to \langle h, \Phi(x)\rangle_{\mathcal{H}_0}$ for all $x \in X$ and $n \in \mathbb{N}$, we conclude that $h \in \ker V$, and hence, the null space of V is closed. Subsequently, this allows us to have the orthogonal decomposition $\mathcal{H}_0 = \ker V \oplus (\ker V)^\perp$, where $(\ker V)^\perp$ denotes the orthogonal complement of $\ker V$ (Appendix B). By construction, $V|_{(\ker V)^\perp}$ is injective. We show that $V|_{(\ker V)^\perp}$ is also surjective. Consider $f \in \mathcal{H}$ and $h \in \mathcal{H}_0$ such that $V(h) = f$. Note that $f = V(h) = V(h_0 + h_0^\perp) = V(h_0^\perp) = V|_{(\ker V)^\perp}(h_0^\perp)$. This proves surjectivity. Similarly,

$$|f|_{\mathcal{H}}^2 = \inf_{h_0 + h_0^\perp \in V^{-1}(f)} \left|h_0 + h_0^\perp\right|_{\mathcal{H}_0}^2 = \inf_{h_0 + h_0^\perp \in V^{-1}(f)} \left|h_0\right|_{\mathcal{H}_0}^2 + \left|h_0^\perp\right|_{\mathcal{H}_0}^2 = \left|V|_{(\ker V)^\perp}^{-1}(f)\right|_{(\ker V)^\perp}^2,$$

showing that $V|_{(\ker V)^\perp} : (\ker V)^\perp \to \mathcal{H}$ is an isometric isomorphism. Given that $(\ker V)^\perp$ is a Hilbert space, it follows that $|\cdot|_{\mathcal{H}}$ is a Hilbert space norm on $\mathcal{H}$.

Now, let us show that k is a reproducing kernel for $\mathcal{H}$. Since k has feature map Φ_0, we observe that $k(\cdot, x) = \langle \Phi_0(x), \Phi_0(\cdot)\rangle_{\mathcal{H}_0} = V(\Phi_0(x)) \in \mathcal{H}$. Moreover, for all $h_0 \in \ker V$, $\langle h_0, \Phi_0(x)\rangle_{\mathcal{H}_0} = 0$, hence $\Phi_0(x) \in (\ker V)^\perp$ and

$$f(x) = \left\langle V|_{(\ker V)^\perp}^{-1}(f), \Phi_0(x)\right\rangle_{\mathcal{H}_0} = \left\langle f, V|_{(\ker V)^\perp}\Phi_0(x)\right\rangle_{\mathcal{H}} = \langle f, k(\cdot, x)\rangle_{\mathcal{H}}.$$

Therefore, k has the reproducing property, and by Proposition 2.3 $\mathcal{H}$ is a RKHS. It remains only to show uniqueness. To this end, consider the set

$$\mathcal{H}_{\text{pre}} := \left\{\sum_{i=1}^{n} \alpha_i k(\cdot, x_i) : n \in \mathbb{N}, \alpha_1, \ldots, \alpha_n \in \mathbb{R}, \text{ and } x_1, \ldots, x_n \in X\right\}. \tag{2.2}$$

It is evident that $\mathcal{H}_{\text{pre}} \subset \mathcal{H}$. We show that $\mathcal{H}_{\text{pre}}$ is dense in $\mathcal{H}$. Assume otherwise. This assumption implies that $(\mathcal{H}_{\text{pre}})^\perp \neq \{0\}$, and hence, there exists an $f \in (\mathcal{H}_{\text{pre}})^\perp$ and an $x \in X$ such that $f(x) \neq 0$. Consequently,

$$0 = \langle f, k(\cdot, x)\rangle_{\mathcal{H}} = f(x) \neq 0,$$

which yields a contradiction. Thus, $\mathcal{H}_{\text{pre}}$ is dense in $\mathcal{H}$. From the density of $\mathcal{H}_{\text{pre}}$ and Remark 2.1, uniqueness follows. □

We conclude this section with a complete characterisation of kernel functions. Recall that a map $k : X \times X \to R$ is said to be positive definite if, for all $n \in \mathbb{N}$, $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$, and $x_1, \ldots, x_n \in X$, we have

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j k(x_j, x_i) \geq 0.$$

Additionally, we say that k is strictly positive definite if the inequality above is strict, unless $\alpha_1 = \cdots = \alpha_n = 0$. Lastly, k is said to be symmetric if $k(x, x') = k(x', x)$ for all $x, x' \in X$.

Theorem 2.2 ([71], Theorem 4.16)**:** A map $k : X \times X \to \mathbb{R}$ is a kernel if and only if it is symmetric and positive definite.

Proof. The direction $(\rightarrow)$ is relatively straightforward, so we omit its proof. Conversely, we show that any symmetric and positive definite map k is a kernel. As in Theorem 2.1, let us consider the set $\mathcal{H}_{\text{pre}}$ given by (2.2). Consider $f, g \in \mathcal{H}_{\text{pre}}$ such that

$$f := \sum_{i=1}^{n} \alpha_i k(\cdot, x_i) \quad \text{and} \quad g := \sum_{j=1}^{m} \beta_j k(\cdot, x'_j).$$

Define $\langle \cdot, \cdot \rangle$ as

$$\langle f, g \rangle := \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \beta_j k(x'_j, x_i).$$

Since k is symmetric and positive definite, it follows that $\langle \cdot, \cdot \rangle$ is symmetric and $\langle f, f \rangle \geq 0$. Additionally, $\langle \cdot, \cdot \rangle$ is clearly bilinear. Now, let us assume that $\langle f, f \rangle = 0$. Then, by the Cauchy-Schwarz inequality, we have that

$$|f(x)|^2 = \left| \sum_{i=1}^{n} \alpha_i k(x, x_i) \right|^2 = \left| \langle f, k(\cdot, x) \rangle \right|^2 \leq \langle k(\cdot, x), k(\cdot, x) \rangle \cdot \langle f, f \rangle = 0,$$

for all $x \in X$. Hence, $f = 0$ and we conclude that $\langle \cdot, \cdot \rangle$ denotes an actual inner product. The result now follows by considering the completion of $\mathcal{H}_{\text{pre}}$. Specifically, let $\mathcal{H}$ be the completion of $\mathcal{H}_{\text{pre}}$ and let $I : \mathcal{H}_{\text{pre}} \to \mathcal{H}$ denote the corresponding isometric embedding (Appendix B). Then, $\mathcal{H}$ is a Hilbert space and

$$\langle Ik(\cdot, x'), Ik(\cdot, x) \rangle_{\mathcal{H}} = \langle k(\cdot, x'), k(\cdot, x) \rangle_{\mathcal{H}_{\text{pre}}} = k(x, x'),$$

for all $x, x' \in X$. In other words, $x \mapsto Ik(\cdot, x)$ defines a feature map and k is a kernel. □

2.1.2. Basic Properties of RKHSs

In this concise section, we compile a few properties of RKHSs that are relevant for what follows. The first result has already been proven.

Lemma 2.1: Let $\mathcal{H}$ be a RKHS over X with kernel k. Then, the linear span of functions $k_x(\cdot) \equiv k(\cdot, x)$ is dense in $\mathcal{H}$.

Proof. See the proof of Theorem 2.1. □

Lemma 2.2 ([71], Lemma 4.29)**:** Let (X, τ) be some topological space and k a kernel on X with feature space $\mathcal{H}$ and feature map $\Phi : X \to \mathcal{H}$. Then, the following claims are equivalent:

1. k is continuous.
2. k is continuous on each variable and $x \mapsto k(x, x)$ is continuous.
3. Φ is continuous.

Proof. Note that a feature map induces a pseudo-metric on X. Specifically, for all $x, x' \in X$, we set $d_k(x, x') := |\Phi(x) - \Phi(x')|_{\mathcal{H}}$. This pseudo-metric is commonly known as *kernel metric*. Remarkably, d_k is actually independent of Φ. Indeed,

$$d_k(x, x') = \sqrt{k(x, x) - 2k(x, x') + k(x', x')}.$$

Now, to begin with, the implication $1 \to 2$ is trivial. Subsequently, assuming 2), we conclude that $d_k(x, \cdot) : (X, \tau) \to \mathbb{R}$ is continuous for every $x \in X$. Hence, $\mathrm{id} : (X, \tau) \to (X, d_k)$ is continuous. Given that $\Phi : (X, d_k) \to \mathcal{H}$ is clearly continuous, $\Phi \circ \mathrm{id} : (X, \tau) \to \mathcal{H}$ is continuous and 3) follows, thus establishing $2 \to 3$. Lastly, $3 \to 1$ follows by noticing that

$$\begin{aligned} |k(x_1, x_1') - k(x_2, x_2')| \\ &= |\langle \Phi(x_1), \Phi(x_1') \rangle_{\mathcal{H}} - \langle \Phi(x_1'), \Phi(x_2) \rangle_{\mathcal{H}} + \langle \Phi(x_1'), \Phi(x_2) \rangle_{\mathcal{H}} - \langle \Phi(x_2), \Phi(x_2') \rangle_{\mathcal{H}}| \\ &\leq |\Phi(x_1')|_{\mathcal{H}} |\Phi(x_1) - \Phi(x_2)|_{\mathcal{H}} + |\Phi(x_2)|_{\mathcal{H}} |\Phi(x_1') - \Phi(x_2')|_{\mathcal{H}}, \end{aligned}$$

for all $x_1, x_1', x_2, x_2' \in X$. This concludes the proof. □

Lemma 2.3 ([61], Theorem 2.17)**:** Let X be a topological space and k a kernel on X with RKHS $\mathcal{H}$. If k is continuous with respect to the product topology, then every function in $\mathcal{H}$ is continuous.

Proof. See [61] for an "$\varepsilon - \delta$" proof. That said, the result follows directly from Lemma 2.2 and (2.1). □

Lemma 2.4 ([71], Lemma 4.33)**:** Let X be a separable topological space and k a continuous kernel on X. Then, the RKHS of k is separable.

Proof. By Lemma 2.2 the canonical feature map Φ_c is continuous. Since the continuous image of a separable set is again separable, $\Phi_c(X)$ is separable. Consequently, the set $\mathcal{H}_{\text{pre}}$ is separable and the result follows by Lemma 2.1. □

2.2. Approximations and RKHSs

This section delves into the approximation capabilities of reproducing kernel Hilbert spaces $\mathcal{H}$ over X. It is divided into two parts based on the problem at hand. The first part focuses on interpolation: Given a finite set of coordinates $\{(x_1, \lambda_1), \dots, (x_n, \lambda_n)\}$, where $x_i \in X$ and $\lambda_i \in \mathbb{R}$ for all $i \in \{1, \dots, n\}$, we explore whether there exists a function in $\mathcal{H}$ that interpolates all these points. The second part examines the approximation of real-valued continuous functions over X. Specifically, let $C(X)$ denote the space of continuous functions from X to $\mathbb{R}$. We aim to determine whether $\mathcal{H}$ is sufficiently expressive to provide arbitrarily accurate approximations of any function in $C(X)$. Naturally, such a problem is greatly dependent on the domain X and its topological properties.

2.2.1. Interpolation

This section begins by presenting the interpolation problem. Following that, we establish the uniqueness of the interpolating function in $\mathcal{H}$ with minimal norm and offer necessary and sufficient conditions for the existence of interpolating functions. Finally, we explore the relationship between strictly positive kernels and the interpolation problem.

Definition 2.3: Let X and Y be arbitrary non-empty sets. Let $\{x_1, \dots, x_n\} \subset X$ be a set of distinct points, and let $\{\lambda_1, \dots, \lambda_n\} \subset Y$ be some subset. We say that a function $g : X \to Y$ interpolates the given set of points if $g(x_i) = \lambda_i$, for all $i \in \{1, \dots, n\}$. We call g the interpolating function.

In what follows, we consider a RKHS $\mathcal{H}$ over a non-empty set X with reproducing kernel k. We take a finite set of distinct points $F = \{x_1, \dots, x_n\} \subset X$, and a set of values $\{\lambda_1, \dots, \lambda_n\} \subset \mathbb{R}$. We denote by $\mathcal{H}_F \subset \mathcal{H}$ the subspace spanned by the maps $\{k_{x_1}, \dots, k_{x_n}\}$.

Observe that $\dim(\mathcal{H}_F) \leq n$. Additionally, note that $\dim(\mathcal{H}_F) < n$ if and only if there exists a linear dependence for every $f \in \mathcal{H}$ evaluated at the points in F. More precisely, for arbitrary $\alpha_i \in \mathbb{R}$, if $\sum_{i=1}^n \alpha_i k_{x_i} = 0$, then, for every $f \in \mathcal{H}$,

$$\langle f, \sum_{i=1}^n \alpha_i k_{x_i} \rangle_{\mathcal{H}} = \sum_{i=1}^n \alpha_i f(x_i) = 0.$$

In this scenario, certain sets of values $\{\lambda_1, \dots, \lambda_n\}$ cannot be interpolated by a function in $\mathcal{H}$. The subsequent result shows that whenever an interpolating function for a set of coordinates exists, the interpolating function of minimal norm is unique.

Proposition 2.4 ([61], Proposition 3.2)**:** Let P_F denote the projection of $\mathcal{H}$ onto $\mathcal{H}_F$. If there exists an interpolating function g in $\mathcal{H}$ for a given set of coordinates, then $P_F(g)$ is the unique function of minimal norm that interpolates these values.

Proof. For a brief recap of orthogonal projections in Hilbert space refer to Appendix B. Consider a fixed set of coordinates $\{(x_1, \lambda_1), \dots, (x_n, \lambda_n)\}$ and let $\mathcal{H}_F^\perp$ denote the orthogonal complement of $\mathcal{H}_F$. Note that $h \in \mathcal{H}_F^\perp$ if and only if $h(x_i) = \langle h, k_{x_i} \rangle_{\mathcal{H}} = 0$ for all $i \in \{1, \dots, n\}$. Hence, for any $h \in \mathcal{H}$, we have that

$$h(x_i) = P_F(h)(x_i) \text{ for all } i \in \{1, \dots, n\},$$

i.e., if h is an interpolating function, then its projection onto $\mathcal{H}_F$ is also interpolating. Furthermore, if we have two interpolating functions g_1 and g_2, then by the aforementioned equivalence we have that $g_1 - g_2 \in \mathcal{H}_F^\perp$, implying that any solution of the interpolation problem is of the form $g + h$ with $h \in \mathcal{H}_F^\perp$. Finally, observe that, for any $h \in \mathcal{H}_F^\perp$,

$$|P_F(g)|_{\mathcal{H}} = |P_F(g + h)|_{\mathcal{H}} \leq |g + h|_{\mathcal{H}},$$

meaning that $P_F(g)$ is the unique interpolating function of minimal norm. □

We now provide sufficient and necessary conditions for the existence of an interpolating function.

Theorem 2.3 ([61], Theorem 3.4)**:** Let $F = \{x_1, \ldots, x_n\} \subset X$ be a set of distinct points and consider $\{\lambda_1, \ldots, \lambda_n\} \subset \mathbb{R}$. Then, there exists an interpolating function $g \in \mathcal{H}$ for this set of points if and only if the vector $\lambda := (\lambda_1, \ldots, \lambda_n)^T$ is in the range of the matrix $K := [k(x_i, x_j)]_{i,j}$. Additionally, in case $\alpha := (\alpha_1, \ldots, \alpha_n)^T$ is a vector whose image is λ, i.e. $K\alpha = \lambda$, then the function $h := \sum_i \alpha_i k_{x_i}$ is the unique interpolating function of minimal norm in $\mathcal{H}$. Lastly, we have that $|h|^2_{\mathcal{H}} = \langle \alpha, \lambda \rangle$.

Proof. We start with ($\rightarrow$). Assume that there exists an interpolating function $g \in \mathcal{H}$. Then, by Proposition 2.4, the (unique) solution of minimal norm is $P_F(g) = \sum_i \beta_i k_{x_i}$, for some scalars $\beta_1, \ldots, \beta_n$. Now, simply notice that

$$g(x_j) = P_F(g)(x_j) = \sum_i \beta_i k_{x_i}(x_j) = \lambda_j,$$

is equivalent to having $K\beta = \lambda$, where $\beta := (\beta_1, \ldots, \beta_n)^T$. Hence, λ is in the range of K. To prove ($\leftarrow$), assume that α satisfies $K\alpha = \lambda$ and set $h = \sum_i \alpha_i k_{x_i}$. Then, h is interpolating. Finally, to see that h is the unique interpolating function of minimal norm, we show that h coincides with $P_F(g)$. Note that $\alpha - \beta$ is in the nullspace of K. Hence,

$$|P_F(g) - h|^2_{\mathcal{H}} = \sum_{i=1}^{n} \sum_{j=1}^{n} (\alpha_i - \beta_i)(\alpha_j - \beta_j) k(x_i, x_j) = \langle K(\alpha - \beta), \alpha - \beta \rangle_{\mathbb{R}^n} = 0,$$

which means that $P_F(g) - h$ is identically zero and uniqueness follows. □

Theorem 2.3 has the following immediate corollary.

Corollary 2.1: Let $F = \{x_1, \ldots, x_n\} \subset X$ be a set of distinct points. If the matrix $K = [k(x_i, x_j)]_{i,j}$ is invertible, then, for any set of values $\{\lambda_1, \ldots, \lambda_n\} \subset \mathbb{R}$, there exists an interpolating function in $\mathcal{H}$. Moreover, the unique interpolating function of minimal norm is given by $g = \sum_i \alpha_i k_{x_i}$, where $\alpha = K^{-1}\lambda$.

We conclude this section by establishing a connection between strictly positive definite kernels and interpolation. As per the preceding results, having an invertible matrix $K = [k(x_i, x_j)]_{i,j}$ is both necessary and sufficient to ensure the existence of an interpolating function. Drawing from Linear Algebra, we recall that positive definite matrices are invertible if and only if they are strictly positive definite. Hence, the relation between the two concepts should not come as a surprise.

Theorem 2.4 ([61], Theorem 3.6)**:** Let X be a non-empty set, $k : X \times X \to \mathbb{R}$ a kernel and $\mathcal{H}$ the respective RKHS. Then, the following claims are equivalent:

1. The kernel k is strictly positive definite.
2. For $n \in \mathbb{N}$ and any set of distinct points $\{x_1, \ldots, x_n\} \subset X$, the functions $k_{x_1}, \ldots, k_{x_n}$ are linearly independent.
3. For $n \in \mathbb{N}$, any set of distinct points $\{x_1, \ldots, x_n\} \subset X$, and any set of values $\{\alpha_1, \ldots, \alpha_n\} \subset \mathbb{R}$ that are not all 0, there exists $f \in \mathcal{H}$ such that
$$\alpha_1 f(x_1) + \cdots \alpha_n f(x_n) \neq 0.$$
4. For $n \in \mathbb{N}$ and any set of distinct points $\{x_1, \ldots, x_n\} \subset X$, there exist functions $g_1, \ldots, g_n \in \mathcal{H}$ such that
$$g_i(x_j) = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases} . \tag{2.3}$$

Proof. The equivalence (1) $\Leftrightarrow$ (2) is easily obtained by recalling that $\sum_{i,j=1}^{n} \alpha_i \alpha_j k(x_i, x_j) = |\sum_{i=1}^{n} \alpha_i k_{x_i}|_{\mathcal{H}}^2$. To get (2) $\Leftrightarrow$ (3), notice that $\sum_i \alpha_i k_{x_i} = 0$, if and only if $\langle f, \sum_i \alpha_i k_{x_i} \rangle_{\mathcal{H}} = 0$ for all $f \in \mathcal{H}$, if and only if $\alpha_1 f(x_1) + \cdots + \alpha_n f(x_n) = 0$ for all $f \in \mathcal{H}$. Lastly, (4) $\implies$ (3) since if $\alpha_i \neq 0$, then we can simply take $f = g_i$. And the implication (1) $\implies$ (4) follows by Corollary 2.1. □

Definition 2.4: A RKHS $\mathcal{H}$ satisfying any of the equivalent conditions above is said to be fully interpolating.

2.2.2. Universal Kernels

We introduce one of the central themes of this thesis: *universality*. In essence, universality relates to the capability of approximating continuous (real-valued) functions over a given domain X, i.e. elements of $C(X)$. As discussed in Section 2.1, a kernel function k corresponds uniquely to a RKHS $\mathcal{H}$. Furthermore, Lemma 2.3 demonstrates that when k is continuous, every function in $\mathcal{H}$ is continuous as well. Consequently, it raises the question of whether $\mathcal{H}$ possesses sufficient expressiveness to approximate elements of $C(X)$ with arbitrary accuracy. This inquiry leads us to the concept of a universal kernel.

Unsurprisingly, the density of $\mathcal{H}$ in $C(X)$ hinges significantly on the topological characteristics of the domain X. Particularly, in the classical setting under examination, X is assumed to be a compact metric space. In hindsight, we observe that this compactness assumption stems from the Stone-Weierstrass theorem (Appendix C), which offers sufficient conditions for the existence of universal kernels, necessitating a compact domain. In Chapter 3, we will delve deeper into the implications of the compactness assumption on X, and explore ways to relax it. As mentioned in the Introduction, this will culminate in the notion of *global universality.*

Definition 2.5: Let X be a compact metric space. A continuous kernel $k : X \times X \to \mathbb{R}$ is said to be universal if the corresponding RKHS $\mathcal{H}$ is dense in $C(X)$, i.e., for every $g \in C(X)$ and $\varepsilon > 0$, there exists $f \in \mathcal{H}$ such that

$$|f - g|_\infty := \sup_{x \in X} |f(x) - g(x)| \leq \varepsilon.$$

Remark 2.2: One can modify the definition of universal kernel and consider an arbitrary feature space $\mathcal{H}_0$ of k. Indeed, if $\Phi_0 : X \to \mathcal{H}_0$ denotes the corresponding feature map, then by (2.1) we observe that k is universal if and only if, for every $\varepsilon > 0$, there exists $h \in \mathcal{H}_0$ such that $|\langle h, \Phi_0(\cdot)\rangle_{\mathcal{H}_0} - g|_\infty \leq \varepsilon$.

We now prove that universal kernels do indeed exist and state what is commonly referred to as a "test for universality." The idea behind the following result is to simply restate the assumptions of the Stone-Weierstrass theorem (Appendix C) in the context of kernels. For convenience, we recall here the required terminology in order to apply the Stone-Weierstrass theorem.

Definition 2.6: Given some topological space X, we say that $\mathcal{A} \subset C(X)$ is a subalgebra if $\mathcal{A}$ is a vector subspace and is closed under multiplication, i.e. if $f, g \in \mathcal{A}$ then $fg \in \mathcal{A}$. Moreover, a subset of functions $\mathcal{C} \subset C(X)$ is said to be point-separating if, for any two distinct points $x_1, x_2 \in X$, there exists $f \in \mathcal{C}$ such that $f(x_1) \neq f(x_2)$. Lastly, we say $\mathcal{C} \subset C(X)$ vanishes nowhere if, for all $x \in X$, there exists at least one $f \in \mathcal{C}$ such that $f(x) \neq 0$.

Theorem 2.5 ([71], Theorem 4.56)**:** Let X be a compact metric space and $k : X \times X \to \mathbb{R}$ a continuous kernel such that $k(x, x) > 0$ for all $x \in X$. Suppose there is an injective feature map $\Phi : X \to l_2(\mathbb{N})$ of k. We denote by $\Phi_n : X \to \mathbb{R}$ the components of Φ, i.e. $\Phi(x) \equiv (\Phi_n(x))_{n \in \mathbb{N}}$. If $\mathcal{A} := \text{span}\{\Phi_n : n \in \mathbb{N}\}$ is a subalgebra, then k is universal.

Proof. We just need to verify the assumptions of the Stone-Weierstrass theorem. Firstly, the algebra $\mathcal{A}$ vanishes nowhere since $|\Phi(x)|^2_{l_2(\mathbb{N})} = k(x, x) > 0$ by assumption. Moreover, given that k is continuous, it follows by Lemma 2.2 that every Φ_n is continuous. Hence, $\mathcal{A} \subset C(X)$. Additionally, since Φ is injective, $\mathcal{A}$ is point-separating. As such, by Stone-Weierstrass we conclude that $\mathcal{A}$ is dense in $C(X)$, i.e. for all $\varepsilon > 0$ and $g \in C(X)$ there exists an $f \in \mathcal{A}$ such that $|f - g|_\infty < \varepsilon$. Since f is necessarily a linear combination of functions Φ_n, it follows that $f = \langle w, \Phi(\cdot)\rangle_{l_2(\mathbb{N})}$ for some $w \in l_2(\mathbb{N})$, and, by (2.1) and Remark 2.2, we conclude that k is universal. □

The Stone-Weierstrass theorem can be interpreted as a generalisation of the Weierstrass theorem (Appendix C), given that polynomials are a particular example of an algebra. Proposition 2.1, in turn, establishes the existence of Taylor-type kernels by constructing an explicit feature map composed of polynomial functions. Consequently, we readily deduce that Taylor-type kernels are universal, in the sense of Definition 2.5.

Corollary 2.2 ([71], Corollary 4.57)**:** Consider $r \in (0, \infty]$ and let $f : [-r, r] \to \mathbb{R}$ be a continuous function that can be expressed by its Taylor series, i.e.

$$f(t) = \sum_{n=0}^{\infty} a_n t^n, \text{ for all } t \in [-r, r].$$

If the Taylor coefficients are such that $a_n > 0$ for all $n \in \mathbb{N}_0$, then the kernel k given by $k(x, x') := f(\langle x, x' \rangle_{\mathbb{R}^d})$ is universal on every compact subset of the closed ball $\sqrt{r} B_{\mathbb{R}^d}$.

Corollary 2.2 demonstrates the universality of certain classical examples of kernels. Readers interested in further exploration can refer to Corollary 4.58 in [71].

2.3. Signature as a Feature Map

As discussed earlier, kernels serve as means to map a low-dimensional data space X into a higher-dimensional feature space $\mathcal{H}$. This mapping allows for increased expressivity, potentially enabling the approximation of a wide range of functions using the associated RKHS. This was demonstrated with the Taylor kernels in Section 2.1.1, which were proven to be universal (Corollary 2.2) under the conditions of $X \subset \mathbb{R}^d$ being compact and the Taylor coefficients being strictly positive.

Here, we note that developing kernels suited for sequential data is of great interest, given its prevalence in various applications such as time series analysis. Instead of considering data in Euclidean space like $X \subset \mathbb{R}^d$, it becomes relevant to explore $X \subset \mathbb{R}^d_{\text{seq}}$, where $\mathbb{R}^d_{\text{seq}}$ represents the set of sequences of arbitrary length in $\mathbb{R}^d$. That said, this poses some challenges. Note that $\mathbb{R}^d_{\text{seq}}$ is not even a linear space due to the absence of a natural addition operation between sequences of different lengths. One approach, however, is to consider the piecewise linear interpolation of the data points, embedding $\mathbb{R}^d_{\text{seq}}$ into the class of bounded variation paths [46, 50].

Recognising that streams of data can often be viewed as paths suggests seeing the signature of a path as a feature map. This section elaborates on this idea. Specifically, Section 2.3.1 defines the so-called *signature kernel* and establishes some of its relevant properties. Subsequently, further exploration in Section 2.3.2 establishes the signature kernel as being universal in the sense of Definition 2.5.

2.3.1. The Signature Kernel

We briefly review the signature kernel and demonstrate that it satisfies a hyperbolic PDE belonging to a class of differential equations known as Goursat problems [67], thus providing a "kernel trick."

To start, let us recall that the signature of a path takes values in $T_1((\mathbb{R}^d))$, a subset of the extended tensor algebra with a Hilbert space structure (Section 1.1.3 and Appendix A). Given $\mathbf{a}, \mathbf{b} \in T_1((\mathbb{R}^d))$, the inner product is defined by

$$\langle \mathbf{a}, \mathbf{b} \rangle_{T_1((\mathbb{R}^d))} := \sum_{n=0}^{\infty} \langle \mathbf{a}^n, \mathbf{b}^n \rangle_{(\mathbb{R}^d)^{\otimes n}} \equiv \sum_{n=0}^{\infty} \sum_{|w|=n} a_w b_w.$$

Consequently, viewed as a mapping from $C^{1\text{-var}}([0,T], \mathbb{R}^d)$ to $T_1((\mathbb{R}^d))$, the signature is indeed a feature map. It is worth noting that, due to the discrete nature of data in applications, it is usually sufficient to consider paths of bounded variation, or even restrict our attention to piecewise linear paths.

Definition 2.7: Let $I = [u, u']$ and $J = [v, v']$ be two compact intervals, and consider two paths x and y with values on $\mathbb{R}^d$, continuously differentiable over I and J, respectively. The signature kernel $k_{x,y} : I \times J \to \mathbb{R}$ is defined as

$$k_{x,y}(s,t) := \langle S(x)_{u,s}, S(y)_{v,t} \rangle_{T_1((\mathbb{R}^d))}.$$

Lemma 2.5 ([44], Lemma 4.3)**:** The signature kernel is well-defined, i.e. $k_{x,y}$ is a kernel and $k_{x,y} < \infty$.

Proof. The fact that $k_{x,y}$ is a kernel follows immediately by definition. The finiteness of the inner product follows by Proposition 1.18 and the Cauchy-Schwarz inequality. □

Now, we demonstrate that the signature kernel constitutes the solution to a Goursat PDE. This observation is particularly significant as it presents a "kernel trick." Typically, the term "kernel trick" refers to any method that bypasses the computation of the inner product in the feature space. In other words, a kernel trick allows us to avoid the explicit calculation of embeddings and directly compute inner products in potentially infinite-dimensional spaces. In the context of the signature kernel, rather than computing an inner product in $T_1((\mathbb{R}^d))$, we instead address a relatively simple hyperbolic PDE.

Remark 2.3: For our purposes of establishing the signature as a feature map and showing that the signature kernel solves a Goursat PDE, it is convenient to consider continuously differentiable paths x and y. We note, however, that one can lower this regularity assumption to paths of bounded variation, and even extend it to a class of rough paths. We refer to the original work [67, 15] for details. Additionally, it is worth mentioning that by fixing s and t, the signature kernel can be interpreted as a kernel over a path space.

Theorem 2.6 ([67], Theorem 2.5)**:** Let $I = [u, u']$ and $J = [v, v']$ be two compact intervals, and consider two paths x and y continuously differentiable over I and J, respectively. The signature kernel $k_{x,y} : I \times J \to \mathbb{R}$ is a solution of the following linear, second order, hyperbolic PDE

$$\frac{\partial^2 k_{x,y}}{\partial s \partial t} = \langle \dot{x}_s, \dot{y}_t \rangle_{\mathbb{R}^d} k_{x,y}(s,t), \quad k_{x,y}(u,\cdot) = k_{x,y}(\cdot, v) = 1,$$

where $\dot{x}_s$ and $\dot{y}_t$ denote $\frac{dx_p}{dp}\Big|_s$ and $\frac{dy_q}{dq}\Big|_t$, respectively.

Proof. Since $S(x)_{u,u} = S(y)_{v,v} = \mathbf{1}$, the initial conditions are clearly satisfied. To derive the PDE, recall that, by Proposition 1.15,

$$S(x)_{u,s} = \mathbf{1} + \int_{p=u}^{s} S(x)_p \otimes \mathrm{d}x_p,$$

which also holds true for $S(y)_{v,t}$. Hence, we note that

$$\begin{aligned}
k_{x,y}(s,t) &= \langle S(x)_{u,s}, S(y)_{v,t} \rangle_{T_1} \\
&= \Big\langle \mathbf{1} + \int_{p=u}^{s} S(x)_p \otimes \mathrm{d}x_p, \mathbf{1} + \int_{q=v}^{t} S(x)_q \otimes \mathrm{d}y_q \Big\rangle_{T_1} \\
&= 1 + \Big\langle \int_{p=u}^{s} S(x)_p \otimes \dot{x}_p \,\mathrm{d}p, \int_{q=v}^{t} S(x)_q \otimes \dot{y}_q \,\mathrm{d}q \Big\rangle_{T_1} \\
&= 1 + \int_{p=u}^{s} \int_{q=v}^{t} \langle S(x)_p \otimes \dot{x}_p, S(x)_q \otimes \dot{y}_q \rangle_{T_1} \,\mathrm{d}q \,\mathrm{d}p \\
&= 1 + \int_{p=u}^{s} \int_{q=v}^{t} \langle S(x)_p, S(x)_q \rangle_{T_1} \langle \dot{x}_p, \dot{y}_q \rangle_{\mathbb{R}^d} \,\mathrm{d}q \,\mathrm{d}p \\
&= 1 + \int_{p=u}^{s} \int_{q=v}^{t} k_{x,y}(p,q) \langle \dot{x}_p, \dot{y}_q \rangle_{\mathbb{R}^d} \,\mathrm{d}q \,\mathrm{d}p.
\end{aligned}$$

Observe that the exchange between the integrals and the inner product is justified by linearity and continuity. Additionally, the second-last equality follows from the coproduct property of the inner product in $T_1((\mathbb{R}^d))$ (Appendix A). Finally, by applying the fundamental theorem of calculus twice, we obtain the desired PDE. □

2.3.2. Universality of Signatures

We establish the universality of the signature kernel on the space $\hat{C}^{1\text{-var}}([0,T], \mathbb{R}^{d+1})$ of continuous time-augmented bounded variation paths. Specifically, within a compact set $K \subset \hat{C}^{1\text{-var}}([0,T], \mathbb{R}^{d+1})$, we demonstrate that the set of linear functionals of the signature can uniformly and accurately approximate any real-valued continuous map over K. Extending this universality property to rough paths is then fairly straightforward. We note that the effectiveness of path signatures as features in various applications largely stems from this property of universality.

Furthermore, while demonstrating universality, we establish the uniqueness of the signature for time-augmented paths. This serves as a sufficient condition for uniqueness, although it is not necessary. The topic of signature uniqueness is notably complex, and we intentionally deferred its exploration until now. Indeed, fully characterising the uniqueness of signatures extends beyond the scope of this work and delves into sophisticated territory. We provide some remarks on uniqueness at the end of the section and conclude the chapter by establishing that the signature kernel is fully interpolating according to Definition 2.4. This shows that the signature kernel is strictly positive definite, a basic but fundamental property that appears to have been overlooked in previous literature.

Definition 2.8: Let $x \in C^{1\text{-var}}([0,T],\mathbb{R}^d)$ be a continuous path of bounded variation. We define the time-augmented path $\hat{x} : [0,T] \to \mathbb{R}^{d+1}$ by $\hat{x}_t = (t, x_t)$, for all $t \in [0,T]$. We denote by $\hat{C}^{1\text{-var}}([0,T],\mathbb{R}^{d+1})$ the subspace of time-augmented paths in $C^{1\text{-var}}([0,T],\mathbb{R}^{d+1})$.

In what follows, the components of x_t are still denoted by x_t^i for $i \in \{1,\dots,d\}$, and we introduce a 0-th component to $\hat{x}_t$ such that $\hat{x}_t^0 = t$. Employing the natural pairing notation, we find that $\langle e_i, S(\hat{x})_t\rangle = \hat{x}_t^i = x_t^i$ for $i \in \{1,\dots,d\}$, and $\langle e_0, S(\hat{x})_t\rangle = \hat{x}_t^0 = t$. It is worth recalling that $\langle e_\varnothing, S(\hat{x})_t\rangle = 1$ for any path.

Theorem 2.7 ([26], Theorem 3.6)**:** Let K be a compact subset of $\hat{C}^{1\text{-var}}([0,T],\mathbb{R}^{d+1})$. Consider a continuous function $f : K \to \mathbb{R}$, i.e. an element of $C(K)$. Then, for every $\varepsilon > 0$, there exists a linear functional $\hat{x} \mapsto L(S(\hat{x})) := \sum_{0\le|w|\le N} \alpha_w \langle e_w, S(\hat{x})\rangle$, for some $N \in \mathbb{N}_0$ and $\alpha_w \in \mathbb{R}$, such that

$$\sup_{\hat{x}\in K} \big| f(\hat{x}) - L(S(\hat{x}))\big| < \varepsilon.$$

Proof. The result follows by the Stone-Weierstrass theorem (Appendix C) applied to the set

$$\mathcal{A} := \operatorname{span}\Big\{\hat{x} \mapsto \langle e_w, S(\hat{x})\rangle : w \in \{0,1,\dots,d\}^N, N \in \mathbb{N}_0\Big\}.$$

Hence, we must prove that $\mathcal{A}$ satisfies the following conditions:

1) It is a linear subspace of continuous functions from K to $\mathbb{R}$;
2) It is a subalgebra that vanishes nowhere (Definition 2.6);
3) It is point-separating;

Point 1) follows directly from Corollary 1.2 and the fact that the linear functionals $\langle e_w, \cdot \rangle : T((\mathbb{R}^d)) \to \mathbb{R}$ are bounded. Point 2), in turn, is established by Proposition 1.14 and the fact that $\langle e_\varnothing, S(\hat{x}) \rangle = 1$, for all $\hat{x} \in K$. It remains to show that $\mathcal{A}$ is point-separating. To this end, consider functionals of the form

$$\hat{x} \mapsto \langle (e_i \sqcup\!\sqcup e_0^{\otimes k}) \otimes e_0, S(\hat{x}) \rangle, \tag{2.4}$$

for $k \in \mathbb{N}_0$ and $i \in \{0, 1, \ldots, d\}$. Observe that, by Proposition 1.14 and Remark 1.7,

$$\begin{aligned} \langle (e_i \sqcup\!\sqcup e_0^{\otimes k}) \otimes e_0, S(\hat{x}) \rangle &= \int_0^T \langle e_i \sqcup\!\sqcup e_0^{\otimes k}, S(\hat{x})_t \rangle \, \mathrm{d}t \\ &= \int_0^T \langle e_i, S(\hat{x})_t \rangle \langle e_0^{\otimes k}, S(\hat{x})_t \rangle \, \mathrm{d}t = \int_0^T \hat{x}_t^i \, \frac{t^k}{k!} \, \mathrm{d}t. \end{aligned} \tag{2.5}$$

Now, let $\hat{x}, \hat{y} \in \hat{C}^{1\text{-var}}([0,T], \mathbb{R}^{d+1})$ be distinct time-augmented paths. Assuming that, for all $k \in \mathbb{N}_0$ and $i \in \{0, 1, \ldots, d\}$ it holds $\langle (e_i \sqcup\!\sqcup e_0^{\otimes k}) \otimes e_0, S(\hat{x}) \rangle = \langle (e_i \sqcup\!\sqcup e_0^{\otimes k}) \otimes e_0, S(\hat{y}) \rangle$, then we have

$$\int_0^T \left(\hat{x}_t^i - \hat{y}_t^i \right) \frac{t^k}{k!} \, \mathrm{d}t = 0.$$

However, by Theorem C.10 in Appendix C, this implies $\hat{x}_t^i - \hat{y}_t^i = 0$, contradicting the assumption that $\hat{x}$ and $\hat{y}$ are distinct paths. Hence, $\mathcal{A}$ is point-separating, and, in particular, functionals of the form (2.4) are enough to separate paths. By Stone-Weierstrass we thus conclude that $\mathcal{A}$ is dense in $C(K)$. □

Corollary 2.3: Let K be a compact subset of $\hat{C}^{1\text{-var}}([0,T], \mathbb{R}^{d+1})$. The signature kernel is universal on K.

Proof. By the Riesz representation theorem (Appendix B), linear functionals of the signature $\hat{x} \mapsto L(S(\hat{x}))$ can be written as $\hat{x} \mapsto \langle \mathbf{a}, S(\hat{x}) \rangle_{T_1((\mathbb{R}^d))}$ for some $\mathbf{a} \in T_1((R^d))$. Additionally, it follows from (2.1) that maps of the form $\hat{x} \mapsto \langle \mathbf{a}, S(\hat{x}) \rangle_{T_1((\mathbb{R}^d))}$ belong to the RKHS of the signature kernel. Hence, by Theorem 2.7, the signature kernel is universal in the sense of Definition 2.5. □

We observe that (2.5) implies that $S(\hat{x})$ uniquely determines $\hat{x}_t$ for every $t \in [0,T]$. Therefore, we have inadvertently established the following uniqueness result.

Proposition 2.5 ([24], Lemma 2.6)**:** Consider $\hat{x}, \hat{y} \in \hat{C}^{1\text{-var}}([0,T], \mathbb{R}^{d+1})$. Then, $S(\hat{x}) = S(\hat{y})$ if and only if $\hat{x}_t = \hat{y}_t$ for every $t \in [0,T]$.

Proposition 2.5 is typically stated under the assumption of continuous bounded variation paths in $\mathbb{R}^d$, where at least one component is monotone. In our scenario, this monotone component is represented by time. This of course aligns precisely with the condition of having at least one monotone component, and hence, there is no loss of generality in considering time-augmented paths.

Remark 2.4: We observe that Proposition 2.5 holds for rough paths. In a completely analogous way, we can consider the subset of paths $\hat{\mathbf{x}} \in C^{\alpha\text{-Höl}}([0,T], G^{\lfloor 1/\alpha \rfloor}(\mathbb{R}^{d+1}))$ such that $\hat{\mathbf{x}}_t^{(1)} = (t, x_t)$ for some $x \in C^{\alpha\text{-Höl}}([0,T], \mathbb{R}^d)$. Then, the point separation argument in the proof of Theorem 2.7 can be replicated entirely ([27], Theorem 5.4), showing that the signature map of Definition 1.24 is injective for time-augmented rough paths.

To gain additional insight into why paths with a monotone component are uniquely identified by their signature, let us recall Proposition 1.17. This proposition states that the signature of a path, when concatenated with its time-reversal, results in a trivial signature. Consequently, this suggests that the signature map fails to distinguish paths that retract back onto themselves. Now, by assuming that at least one component of the path is monotone, we necessarily exclude the possibility of having such a retracting path.

In [39], the authors introduce the concept of *tree-like* paths, formalising the notion of paths retracting back onto themselves. They prove, for bounded variation paths, that $S(x) = \mathbf{1}$ if and only if x is tree-like. In other words, the signature $S(x)$ is unique up to tree-like equivalence. Furthermore, this result indicates that the space of bounded variation paths, quotiented by the space of tree-like paths, forms a group under the concatenation operation. Subsequently, in [11], the authors extend the concept of tree-like paths and generalise the uniqueness result for weakly geometric rough paths.

We end this chapter by proving that the signature kernel over $\hat{C}^{1\text{-var}}([0,T], \mathbb{R}^{d+1})$ is fully interpolating and, consequently, strictly positive definite.

Proposition 2.6: The signature kernel over $\hat{C}^{1\text{-var}}([0,T], \mathbb{R}^{d+1})$ is a fully interpolating kernel. Consequently, the signature kernel is strictly positive definite.

Proof. The result follows by the last point in Theorem 2.4. Specifically, consider n distinct paths $\{\hat{x}_1, \dots, \hat{x}_n\} \subset \hat{C}^{1\text{-var}}([0,T], \mathbb{R}^{d+1})$ and a set of non-empty words $w_{i,j} \in \mathcal{W}(A_d)$ such that $\langle e_{w_{i,j}}, S(\hat{x}_j)\rangle_{T_1} \neq \langle e_{w_{i,j}}, S(\hat{x}_i)\rangle_{T_1}$ for $i \neq j$. Note that the existence of such a set of words $w_{i,j}$ is guaranteed by Proposition 2.5. Define the following *Lagrange signature polynomials*,

$$L_j(\hat{x}) \coloneqq \prod_{\substack{1 \leq i \leq n \\ i \neq j}} \frac{\langle e_{w_{i,j}}, S(\hat{x})\rangle_{T_1} - \langle e_{w_{i,j}}, S(\hat{x}_i)\rangle_{T_1}}{\langle e_{w_{i,j}}, S(\hat{x}_j)\rangle_{T_1} - \langle e_{w_{i,j}}, S(\hat{x}_i)\rangle_{T_1}},$$

for $j \in \{1, \dots, n\}$. Observe that $\langle e_{w_{i,j}}, S(\hat{x}_j)\rangle_{T_1} - \langle e_{w_{i,j}}, S(\hat{x}_i)\rangle_{T_1} \neq 0$ for $i \neq j$, hence the Lagrange polynomials are well-defined. Moreover, it is clear that $L_j(\hat{x}_j) = 1$ and $L_j(\hat{x}_i) = 0$ for $i \neq j$.

It remains to show that $L_j(\hat{x})$ belongs to the RKHS of the signature kernel. However, this follows directly from Proposition 1.14, which guarantees that any polynomial function of $\langle e_{w_{i,j}}, S(\hat{x})\rangle_{T_1}$ is again a linear functional of $S(\hat{x})$. This concludes the proof.

To write $L_j(\hat{x})$ in the form of a linear functional of the signature explicitly, let c_i denote the constant $\langle e_{w_{i,j}}, S(\hat{x}_i)\rangle_{T_1}$ and $\mathcal{C}_k^n$ be the set of k-combinations in $A_n = \{1, \dots, n\}$, e.g., $\mathcal{C}_2^3 = \{\{1,2\},\{1,3\},\{2,3\}\}$. Additionally, for $B_k = \{i_1, \dots, i_k\} \subset A_n$, let c_{B_k} denote the product $c_{i_1} \cdots c_{i_k}$ and, similarly, define e_{B_k} as $e_{w_{i_1,n}} \sqcup\!\sqcup \dots \sqcup\!\sqcup\, e_{w_{i_k,n}}$. Using this notation,

$$L_j(\hat{x}) = \prod_{\substack{1\le i\le n \\ i\ne j}} \frac{\langle e_{w_{i,j}}, S(\hat{x})\rangle_{T_1} - c_i}{c_j - c_i} = C_n \prod_{\substack{1\le i\le n \\ i\ne j}} \Big(\langle e_{w_{i,j}}, S(\hat{x})\rangle_{T_1} - c_i\Big),$$

where $C_n \coloneqq \prod_{i\ne j}(c_j - c_i)^{-1}$. By relabelling the paths $\{\hat{x}_1, \dots, \hat{x}_n\}$ if necessary, we assume without loss of generality that $j = n$ and derive the following explicit linear form:

$$\frac{L_n(\hat{x})}{C_n} \equiv \prod_{1\le i\le n-1} \Big(\langle e_{w_{i,n}}, S(\hat{x})\rangle_{T_1} - c_i\Big) = \sum_{k=0}^{n-1}(-1)^k \sum_{B_k\in\mathcal{C}_k^{n-1}} c_{B_k}\langle e_{A_{n-1}\setminus B_k}, S(\hat{x})\rangle_{T_1}.$$

We proceed by induction. For $n = 1$, there is nothing to prove. For $n = 2$, we see that

$$\frac{L_2(\hat{x})}{(c_2 - c_1)^{-1}} \equiv \Big(\langle e_{w_{1,2}}, S(\hat{x})\rangle_{T_1} - c_1\Big) = (-1)^0 c_\emptyset \big\langle e_{A_1\setminus\emptyset}, S(\hat{x})\big\rangle_{T_1} + (-1)^1 c_{A_1}\big\langle e_\varnothing, S(\hat{x})\big\rangle_{T_1},$$

where we agree that $c_\emptyset = 1$ and $e_\emptyset = e_\varnothing$. Assuming the expression above holds for $n - 1 \ge 2$, we deduce that:

$$\begin{aligned}
\frac{L_{n+1}(\hat{x})}{C_{n+1}} &\equiv \prod_{1\le i\le n} \Big(\langle e_{w_{i,n+1}}, S(\hat{x})\rangle_{T_1} - c_i\Big) = \frac{L_n(\hat{x})}{C_n} \cdot \big(\langle e_{w_{n,n+1}}, S(\hat{x})\rangle_{T_1} - c_n\big) \\
&= \sum_{k=0}^{n-1}(-1)^k \sum_{B_k\in\mathcal{C}_k^{n-1}} c_{B_k}\langle e_{A_{n-1}\setminus B_k}, S(\hat{x})\rangle_{T_1} \cdot \big(\langle e_{w_{n,n+1}}, S(\hat{x})\rangle_{T_1} - c_n\big) \\
&= \sum_{k=0}^{n-1}(-1)^k \sum_{B_k\in\mathcal{C}_k^{n-1}} c_{B_k}\langle e_{A_{n-1}\setminus B_k}, S(\hat{x})\rangle_{T_1}\langle e_{w_{n,n+1}}, S(\hat{x})\rangle_{T_1} + \\
&\qquad + \sum_{k=0}^{n-1}(-1)^{k+1} \sum_{B_k\in\mathcal{C}_k^{n-1}} c_{B_k} c_n \langle e_{A_{n-1}\setminus B_k}, S(\hat{x})\rangle_{T_1} \\
&= \sum_{k=0}^{n-1}(-1)^k \sum_{B_k\in\mathcal{C}_k^{n-1}} \Big(c_{B_k}\big\langle e_{A_{n-1}\setminus B_k\cup\{n\}}, S(\hat{x})\big\rangle_{T_1} - c_{B_k}c_n\big\langle e_{A_{n-1}\setminus B_k}, S(\hat{x})\big\rangle_{T_1}\Big),
\end{aligned}$$

where $\langle e_{A_{n-1}\setminus B_k}, S(\hat{x})\rangle_{T_1}\langle e_{w_{n,n+1}}, S(\hat{x})\rangle_{T_1} = \big\langle e_{A_{n-1}\setminus B_k\cup\{n\}}, S(\hat{x})\big\rangle_{T_1}$ by Proposition 1.14. Lastly, note that $c_{B_k}c_n = c_{B_k\cup\{n\}}$ and

$$\left(\bigcup_{k=0}^{n-1}\mathcal{C}_k^{n-1}\right) \cup \left(\bigcup_{k=0}^{n-1}\bigcup_{B_k\in\mathcal{C}_k^{n-1}} (B_k\cup\{n\})\right) = \bigcup_{k=0}^{n}\mathcal{C}_k^n.$$

Moreover, $B_k \cup (A_{n-1} \setminus B_k) \cup \{n\} = A_n$ for all $k \in \{0, \dots, n-1\}$. Hence, by reordering the terms

$$\sum_{k=0}^{n-1} (-1)^k \sum_{B_k \in \mathcal{C}_k^{n-1}} \Big(c_{B_k} \big\langle e_{A_{n-1} \setminus B_k \cup \{n\}}, S(\hat{x}) \big\rangle_{T_1} - c_{B_k} c_n \big\langle e_{A_{n-1} \setminus B_k}, S(\hat{x}) \big\rangle_{T_1} \Big)$$

$$= \sum_{k=0}^{n} (-1)^k \sum_{B_k \in \mathcal{C}_k^{n}} c_{B_k} \langle e_{A_n \setminus B_k}, S(\hat{x}) \rangle_{T_1},$$

and the claim follows. □

Remark 2.5: In [69], the authors extend the definitions of universal and strictly positive kernels and show that these are essentially equivalent. Specifically, Theorem 6 in [69] demonstrates that universal kernels, in the sense of Definition 2.5, are equivalent to strictly positive kernels, as defined in Section 2.1.1. This seems to clash with Proposition 2.6. Indeed, we have proved that the signature kernel is strictly positive definite over $\hat{C}^{1\text{-var}}([0,T], \mathbb{R}^{d+1})$. Therefore, Theorem 6 in [69] would imply that the signature kernel is universal over $\hat{C}^{1\text{-var}}([0,T], \mathbb{R}^{d+1})$, a dubious proposition considering the non-compactness of $\hat{C}^{1\text{-var}}([0,T], \mathbb{R}^{d+1})$. However, we note that all results in [69] are only applicable to kernels defined over locally compact domains X, and $\hat{C}^{1\text{-var}}([0,T], \mathbb{R}^{d+1})$ is not locally compact ([64], Theorem 1.22). Additionally, we note that the authors in [69] examine *characteristic kernels*, a concept deeply related to universality which we defer to future investigation.[1]

[1]Theorem 6 in [69] was only noticed when, almost at the conclusion of the present work, the lecture notes [16] were published.

3
Weighted Spaces and Global Universal Approximations

Chapter 3 is ultimately concerned with global approximations. For a given compact Hausdorff space X, the classical Stone-Weierstrass theorem offers necessary and sufficient conditions for the uniform approximation of real-valued continuous functions over X, thereby yielding several universal approximation results. Notably, in Chapter 2, we leveraged this theorem to demonstrate that linear functionals of the signature have the capacity to uniformly approximate continuous functions over a fixed compact subset of paths (Theorem 2.7).

That said, as previously noted in the Introduction, assuming a fixed compact subset of paths may be unreasonable due to the lack of local compactness in most path spaces. Additionally, in practical applications, data resampling is often necessary, further complicating the requirement of a fixed compact domain. Hence, there is a considerable interest in establishing global (universal) approximation results, capable of going beyond compact sets.

Not considering a compact domain, however, poses a challenge. Indeed, the compactness assumption ultimately serves as a means to control the growth of continuous functions. Without it, one may encounter continuous functions that grow uncontrollably, making approximation difficult, if not impossible. It is in response to this issue that *weighted spaces* are introduced. In essence, a weighted space comprises a possibly non-compact topological space X, satisfying certain separation axioms, alongside with an *admissible weight function* that regulates the growth of functions outside of compact sets in X.

This chapter draws heavily from [27], where weighted spaces and the concept of global universality are extensively explored. Section 3.1 introduces the framework of weighted spaces and defines the global analogue of continuous functions, i.e., those admitting a global approximation. Moving forward, Section 3.2 presents the weighted version of the Stone-Weierstrass theorem ([27], Theorem 3.6) and provides a clear comparison between local and global approximations. Finally, Section 3.3 introduces globally universal kernels, which are the global counterpart of universal kernels and represent a significant contribution in this thesis. The chapter concludes by presenting a family of globally universal kernels termed as *Taylor signature kernels*.

3.1. Weighted Spaces and Weighted Function Spaces

We present the concept of a weighted space and define the function space for which we approximate functions across the entire domain. Specifically, given a completely regular Hausdorff space X and a Banach space $(Y, |\cdot|_Y)$, our aim, as previously discussed, is to approximate functions $f : X \to Y$ over the entire domain X, rather than solely within compact subsets as in the classical Stone-Weierstrass theorem. The fundamental concept involves introducing a new function $\psi : X \to (0, \infty)$ and considering functions f satisfying

$$\sup_{x \in X} \frac{|f(x)|_Y}{\psi(x)} < \infty.$$

In other words, we examine maps $f : X \to Y$ whose growth is controlled by a fixed auxiliary function ψ. Additionally, we require ψ to identify compact subsets of X, thus materialising the idea of approximating functions beyond compact sets. This section formalises these concepts.

3.1.1. Weighted Spaces: Definition and Examples

Throughout this section, (X, τ_X) denotes a completely regular Hausdorff space, which is also known as a Tychonoff space or a $T_{3^1/2}$ space. These conditions are related to the separation axioms and are not essential for the subsequent discussion. For our purposes, it suffices to remember that any metric space is Tychonoff ([58], Theorem 32.2).

Definition 3.1: A function $\psi : X \to (0, \infty)$ is said to be an admissible weight function if, for all $R > 0$, the pre-image $K_R := \psi^{-1}((0, R])$ is compact with respect to τ_X. The pair (X, ψ) is called a weighted space.

By definition, we observe that any weighted space (X, ψ) is σ-compact, in the sense that

$$X = \bigcup_{R \in \mathbb{N}} K_R = \bigcup_{R \in \mathbb{N}} \{x \in X : \psi(x) \leq R\}.$$

We also note that to establish a weighted space, constructing an admissible weight function is essential, and the choice of topology for X significantly influences this process. It is worth noting that a weaker topology increases the likelihood of having compact pre-images K_R. Therefore, defining a weighted space frequently involves considering a weaker topology.

Remark 3.1: The realisation that a weaker topology might be necessary raises an interesting point. Suppose we are dealing with a space X and, with some application in mind, we define a continuous map f over X, such as a feature map (see Definition 2.1). If we require X to be a weighted space, we will likely need to weaken its topology to accommodate an admissible weight function. However, this adjustment may disrupt the continuity of our previously defined map f. Therefore, working with weighted spaces often demands some finesse: the chosen topology must be weak enough to yield compact sets, yet strong enough to preserve the continuity of the maps we require to be continuous.

To ascertain the compactness of the subsets K_R, we must employ a suitable criterion, such as the Banach-Alaoglu theorem (see Appendix B) or the Arzelà-Ascoli theorem (Theorem 1.1). For clarification, we now present a series of examples. Following [27], we consider examples where $(X, |\cdot|_X)$ is a normed space and $\psi : X \to (0, \infty)$ is a function of the form $\psi(x) = \eta(|x|_X)$, where $\eta : [0, \infty) \to (0, \infty)$ denotes a continuous increasing function.

Example 3.1 ([27], Example 2.3 (i))**:** Consider $X = \mathbb{R}^d$ equipped with the usual Euclidean norm. By the Heine-Borel theorem (Appendix C), the pre-image $K_R = \psi^{-1}((0, R])$ is compact, as it is closed and bounded. Therefore, $(\mathbb{R}^d, \psi)$ constitutes a weighted space. It is worth noting that the same rationale applies to any space with the Heine-Borel property, i.e., any space where closed and bounded sets are compact. Additionally, in this scenario, the topology considered is the one induced by the norm, removing the need for a weaker alternative.

Example 3.2 ([27], Example 2.3 (ii))**:** Let $(X, |\cdot|_X)$ denote a dual space endowed with the weak-$*$-topology (Appendix B). Specifically, we mean that there exists some other Banach space $(V, |\cdot|_V)$ and an isometric isomorphism $X \to V^*$, where V^* denotes the dual of V. As before, we consider the function $\psi(x) = \eta(|x|_X)$. Then, it follows immediately by the Banach-Alaoglu theorem (Appendix B) that the pre-images K_R are compact with respect to the weak-$*$-topology, for all $R > 0$. Hence, $(X, \eta(|\cdot|_X))$ is a weighted space.

Example 3.3 ([27], Example 2.3 (iv))**:** Consider some $\alpha \in (0, 1]$ and $X = C_o^{\alpha\text{-Höl}}([0, T], E)$ (Definition 1.4), where $(E, |\cdot|_E)$ denotes a Banach space with the Heine-Borel property. Moreover, equip X with the β-Hölder norm $|\cdot|_{\beta\text{-Höl};[0,T]}$, for $\beta < \alpha$. In other words, consider $C_o^{\alpha\text{-Höl}}([0, T], E)$ endowed with some β-Hölder topology for $\beta < \alpha$. Set $\psi(x) = \eta(|x|_{\alpha\text{-Höl};[0,T]})$. Observe that, for all $R > 0$, the pre-image K_R is bounded with respect to $|\cdot|_{\alpha\text{-Höl};[0,T]}$. Hence, by Proposition 1.6, K_R is compact. This shows $(C_o^{\alpha\text{-Höl}}([0, T], E), \psi)$ to be a weighted space.

Example 3.4 ([27], Example 2.3 (v))**:** Take $\alpha \in (0, 1]$ and, for $x \in C_o^{\alpha\text{-Höl}}([0, T], \mathbb{R}^d)$, let x^t denote the stopped path $[0, T] \ni s \mapsto x_{s\wedge t}$. Consider $X = \Lambda_T^\alpha$ to be the space of stopped α-Hölder continuous paths, i.e.

$$\Lambda_T^\alpha := \left\{(t, x^t) : t \in [0, T],\ x \in C_o^{\alpha\text{-Höl}}([0, T], \mathbb{R}^d)\right\} \cong [0, T] \times C_o^{\alpha\text{-Höl}}([0, T], \mathbb{R}^d)/\sim,$$

where $\sim$ is defined as $(t, x) \sim (s, y)$ if and only if $t = s$ and $x^t = y^s$.

Equip Λ_T^α with the metric

$$d_\Lambda((t,x),(s,y)) := |t-s| + \sup_{u\in[0,T]} |x^t(u) - y^s(u)|.$$

Observe that the set $A_R := \{x \in C_o^{\alpha\text{-Höl}}([0,T],\mathbb{R}^d) : (t,x) \in K_R\}$ is equicontinuous and uniformly bounded. Hence, by Arzelà-Ascoli (Theorem 1.1), A_R is compact with respect to $|\cdot|_{\infty;[0,T]}$. Subsequently, by Tychonoff's theorem (Appendix C), $[0,T]\times A_R$ is compact in the product space $[0,T]\times C_o^{\alpha\text{-Höl}}([0,T],\mathbb{R}^d)$. Since quotient maps are continuous, K_R is compact with respect to the quotient topology. Finally, by recalling that the quotient topology is a final topology, and noting that the pre-image of open balls in Λ_T^α with respect to d_Λ are open in $[0,T]\times C_o^{\alpha\text{-Höl}}([0,T],\mathbb{R}^d)$, we conclude that K_R is compact with respect to the topology induced by d_Λ.

Example 3.5 ([27], Example 2.3 (viii))**:** As a final example, consider a σ-finite measure space $(\Omega,\mathcal{F},\mu)$ and, for $p\in(1,\infty)$, let $X = L^p(\Omega,\mathcal{F},\mu)$ be the usual space of $\mathcal{F}$-measurable functions $x:\Omega\to\mathbb{R}$ such that $|x|_{L^p(\Omega)} = \left(\int_\Omega |x(\omega)|^p d\mu(\omega)\right)^{1/p} < \infty$. Then, for q such that $1/p+1/q=1$, we have $L^p(\Omega)\cong L^q(\Omega)^*$ with $L^q(\Omega)^*$ equipped with the weak-$*$-topology. By the Banach-Alaoglu theorem, we conclude that $\psi(x) = \eta(|x|_{L^p(\Omega)})$ is an admissible weight function, making $L^p(\Omega)$ a weighted space.

3.1.2. Weighted Function Spaces

In this section, given a weighted space (X,ψ), our objective is to define an appropriate function space where *global approximation* becomes feasible. This entails establishing a space of (real-valued) functions defined over X, allowing for approximation throughout the entire domain. Following the approach outlined in the section's introduction, we consider a Banach space $(Y,|\cdot|_Y)$ as co-domain, and define the vector space

$$B_\psi(X,Y) = \left\{ f : X \to Y : \sup_{x\in X} \frac{|f(x)|_Y}{\psi(x)} < \infty \right\},$$

which we equip with the norm

$$|f|_{\mathcal{B}_\psi(X,Y)} := \sup_{x\in X} \frac{|f(x)|_Y}{\psi(x)}.$$

Subsequently, we note that the space of continuous bounded functions from X to Y, denoted by $C_b(X,Y)$, is continuously embedded in $B_\psi(X,Y)$. This observation leads to the following definition.

Definition 3.2: Given a weighted space (X,ψ) and a Banach space $(Y,|\cdot|_Y)$, we define the weighted function space $\mathcal{B}_\psi(X,Y)$ as the $|\cdot|_{\mathcal{B}_\psi(X,Y)}$-closure of $C_b(X,Y)$ in $B_\psi(X,Y)$. Whenever $Y=\mathbb{R}$, we simply write $\mathcal{B}_\psi(X)$.

As mentioned earlier, the space $\mathcal{B}_\psi(X,Y)$ intuitively represents functions whose growth is controlled by an auxiliary admissible weight function $\psi : X \to (0,\infty)$. In particular, this set can include numerous unbounded functions. The following result offers a complete characterisation of $\mathcal{B}_\psi(X)$, which will be extensively used in the upcoming sections.

Theorem 3.1 (Theorem 2.7, [29])**:** Let (X,ψ) be a weighted space. Then, $f : X \to \mathbb{R}$ is in $\mathcal{B}_\psi(X)$ if and only if $f|_{K_R} \in C(K_R,\mathbb{R})$ for all $R > 0$, and

$$\lim_{R\to\infty} \sup_{x\in X\setminus K_R} \frac{|f(x)|}{\psi(x)} = 0. \tag{3.1}$$

Proof. We start with the direction $(\rightarrow)$. Let $f \in \mathcal{B}_\psi(X)$. By definition of weighted function space, there exists a $g \in C_b(X,\mathbb{R})$ such that $|f - g|_{\mathcal{B}_\psi(X)} < \frac{\varepsilon}{2}$. Equivalently, we have that

$$\frac{|f(x)|}{\psi(x)} \leq \frac{\varepsilon}{2} + \frac{|g(x)|}{\psi(x)} \text{ for all } x \in X,$$

the last term being bounded by $\frac{\varepsilon}{2}$ for all $x \in X \setminus K_R$ with $R := 2\varepsilon^{-1}|g|_\infty$. Hence,

$$\sup_{x\in X\setminus K_R} \frac{|f(x)|}{\psi(x)} \leq \varepsilon,$$

and (3.1) follows. Regarding the continuity of $f|_{K_R}$, for all $R > 0$, we observe that

$$\sup_{x\in K_R} |f(x) - g(x)| \leq R \sup_{x\in K_R} \frac{|f(x) - g(x)|}{\psi(x)} \leq \frac{\varepsilon}{2}R.$$

This implies that $f|_{K_R}$ is the uniform limit of continuous functions, and hence continuous ([58], Theorem 21.6).

To prove $(\leftarrow)$, assume that f satisfies (3.1) and $f|_{K_R} \in C(K_R,\mathbb{R})$ for all $R > 0$. For $n \in \mathbb{N}$, set $f_n := \min(\max(f(\cdot), -n), n)$. Note that $f \in C(K_R,\mathbb{R})$ implies $f_n \in C(K_R,\mathbb{R})$. We would like to show that $f_n \in \mathcal{B}_\psi(X)$ for all $n \in \mathbb{N}$ and $R > 0$. The idea is to find a sequence of continuous bounded functions over X that converge to f_n with respect to $|\cdot|_{\mathcal{B}_\psi(X)}$. We note, however, that the ambient space X is only assumed to be a Hausdorff completely regular space. Hence, X is not necessarily compact and continuous functions are not immediately bounded.

To address this issue we use the fact that all completely regular spaces can be embedded into a compact space Z. Specifically, X is homeomorphic to a subspace of $[0,1]^J$ for some indexing set J (Appendix C). This way, we observe that K_R is a compact, and hence closed, subset of $[0,1]^J$. Subsequently, by Tietze Extension theorem (Appendix C), we obtain the existence of $g_{n,R} \in C_b(X,\mathbb{R})$ such that $g_{n,R}|_{K_R} \equiv f_n|_{K_R}$ and $\sup_{x\in X} |g_{n,R}| \leq n$ for all $x \in X$.

By (3.1), we then obtain

$$|f_n - g_{n,R}|_{\mathcal{B}_\psi(X)} \leq \sup_{x \in X \setminus K_R} \frac{|f_n(x) - g_{n,R}(x)|}{\psi(x)} \leq \frac{2n}{R},$$

showing that $f_n \in \mathcal{B}_\psi(X)$. Lastly, choose $R > 0$ such that $\sup_{x \in X \setminus K_R} \psi(x)^{-1}|f(x)| < \varepsilon$ and pick $n > \sup_{x \in K_R} |f(x)|$ so that $f(x) = f_n(x)$ on K_R. This way we see that,

$$|f - f_n|_{\mathcal{B}_\psi(X)} \leq \sup_{x \in X \setminus K_R} \frac{|f(x) - f_n(x)|}{\psi(x)} \leq \varepsilon + \frac{n}{R}.$$

Since n/R can be made arbitrarily small, we conclude that $f \in \mathcal{B}_\psi(X)$. □

3.2. Weighted Stone-Weierstrass

The classical Stone-Weierstrass theorem provides sufficient and necessary conditions for the existence of a dense subset of functions in the space of real-valued continuous functions over a compact domain. Since we are now interested in achieving global approximations, it is pertinent to establish a more general version of the Stone-Weierstrass theorem. This is the purpose of this section.

Section 3.2.1 presents the weighted analogue of the Stone-Weierstrass theorem, while Section 3.2.2 provides a comparison between local and global approximations. In particular, we prove under mild assumptions that global approximations can achieve everything that local approximations allowed for and more.

3.2.1. Weighted Real-valued Stone-Weierstrass Theorem

We formulate and analyse the proof of the weighted variant of the Stone-Weierstrass theorem. This section draws heavily from Section 3 of [27], where the proof of the weighted Stone-Weierstrass theorem can be found. We include it here for the sake of self-containment. This result is at the core of the present work, as it is the theorem that enables the approximation of functions beyond compact domains. Additionally, we revisit the classical Stone-Weierstrass theorem, which was previously employed in Section 2.2, to facilitate the comparison with the weighted variant. We introduce some supplementary terminology.

Theorem 3.2 (Stone-Weierstrass on $C(X)$)**:** Let X denote a compact Hausdorff space and assume that $\mathcal{A} \subset C(X)$ is a subalgebra. Then, $\mathcal{A}$ is dense in $C(X)$ if and only if $\mathcal{A}$ is point-separating and vanishes nowhere.

This is a classical result in Analysis, and we refer to [73] for the original proof. Hereafter, let (X, ψ) denote a weighted space. We proceed to present the weighted variant of Theorem

3.2, necessitating the definition of the weighted analogue of a point-separating subalgebra of functions.

Definition 3.3: A subalgebra $\mathcal{A} \subset \mathcal{B}_\psi(X)$ is said to be point-separating of ψ-moderate growth if there exists a point-separating vector subspace $\tilde{\mathcal{A}} \subset \mathcal{A}$ such that $x \mapsto \exp(|\tilde{a}(x)|) \in \mathcal{B}_\psi(X)$, for all $\tilde{a} \in \tilde{\mathcal{A}}$.

Theorem 3.3 ([27], Theorem 3.6)**:** Let $\mathcal{A} \subset \mathcal{B}_\psi(X)$ be a point-separating subalgebra of ψ-moderate growth that vanishes nowhere. Then, $\mathcal{A}$ is dense in $\mathcal{B}_\psi(X)$.

Proof. Note that it suffices to show that $\mathcal{A}$ can approximate any element of $C_b(X)$ to arbitrary precision. By definition, $C_b(X)$ is dense in $\mathcal{B}_\psi(X)$, hence the final assertion can be established through a triangle inequality argument. Firstly, let us assume that $\mathcal{A}$ consists only of bounded maps. In this setting, the requirement for $\mathcal{A}$ to be point-separating of ψ-moderate growth simplifies to $\mathcal{A}$ being point-separating. Consider $f \in C_b(X)$, some $\varepsilon > 0$, and define the following constants:

$$M := (\inf_{x\in X} \psi(x))^{-1} > 0, \quad \text{and} \quad b := \sup_{x\in X} |f(x)| + \frac{\varepsilon}{4M}.$$

Note that $\mathcal{A}|_{K_R}$, referring to the set of functions in $\mathcal{A}$ restricted to K_R, forms a point-separating subalgebra of $C(K_R)$ that vanishes nowhere. Therefore, by the classical Stone-Weierstrass (Theorem 3.2), there exists $a \in \mathcal{A}$ such that

$$\sup_{x\in K_R} |f(x) - a(x)| \leq \frac{\varepsilon}{4M} \quad \text{and so,} \quad |a(x)| \leq \frac{\varepsilon}{4M} + \sup_{x\in X} |f(x)| = b,$$

for all $x \in K_R$. Moreover, to control the growth of a outside K_R, let $g \in C_b(\mathbb{R})$ be the function defined as $g(s) = \max(\min(s, b), -b)$ for $s \in \mathbb{R}$, and consider $g(a(x))$. Note that $g(a(x)) = a(x)$ for all $x \in K_R$. Therefore,

$$\begin{aligned} |f - g \circ a|_{\mathcal{B}_\psi(X)} \leq & M \sup_{x\in K_R} |f(x) - a(x)| + \sup_{x\in X\setminus K_R} \frac{|f(x)|}{\psi(x)} + \sup_{x\in X\setminus K_R} \frac{g(a(x))}{\psi(x)} \\ < & M\frac{\varepsilon}{4M} + \frac{b}{R} + \frac{b}{R} \leq \frac{3\varepsilon}{4}, \quad \text{for } R \geq \frac{4b}{\varepsilon}. \end{aligned} \tag{3.2}$$

Next, let us set $c = \sup_{x\in X} |a(x)|$, and utilise the Weierstrass theorem (Appendix C) to acquire a polynomial p such that $\sup_{|s|\leq c} |g(s) - p(s)| < \varepsilon/(4M)$. Observe that $c < \infty$, since a is assumed to be bounded. It follows that,

$$|g \circ a - p \circ a|_{\mathcal{B}_\psi(X)} \leq M \sup_{x\in X} |g(a(x)) - p(a(x))| \leq \sup_{|s|\leq c} |g(s) - p(s)| \leq \frac{\varepsilon}{4}. \tag{3.3}$$

Hence, by combining (3.2) and (3.3), we conclude that

$$|f - p \circ a|_{\mathcal{B}_\psi(X)} \leq |f - g \circ a|_{\mathcal{B}_\psi(X)} + |g \circ a - p \circ a|_{\mathcal{B}_\psi(X)} \leq \varepsilon.$$

Given that $\mathcal{A}$ is a subalgebra, we have that $p \circ a \in \mathcal{A}$. Moreover, $\varepsilon > 0$ and $f \in C_b(X)$ were chosen arbitrarily, hence $\mathcal{A} \subset C_b(X)$ is dense in $\mathcal{B}_\psi(X)$.

Now, we consider the general case of a point-separating subalgebra $\mathcal{A} \subset \mathcal{B}_\psi(X)$ of ψ-moderate growth with a point-separating vector subspace $\tilde{\mathcal{A}} \subset \mathcal{A}$ such that $x \mapsto \exp(|\tilde{a}(x)|) \in \mathcal{B}_\psi(X)$, for all $\tilde{a} \in \tilde{\mathcal{A}}$. We begin by showing that the maps $x \mapsto \cos(\tilde{a}(x))$ and $x \mapsto \sin(\tilde{a}(x))$, with $\tilde{a} \in \tilde{\mathcal{A}}$, belong to the $|\cdot|_{\mathcal{B}_\psi(X)}$-closure of $\mathcal{A}$. Fix $\varepsilon > 0$. Then, by Theorem 3.1, there exists $R > 4/\varepsilon$ such that

$$\sup_{x \in X \setminus K_R} \frac{\exp(|\tilde{a}(x)|)}{\psi(x)} < \frac{\varepsilon}{4}.$$

Set $c = \sup_{x \in K_R} |\tilde{a}(x)|$, and consider the Taylor polynomial $p_n(s) = \sum_{k=0}^n \frac{(-1)^k}{(2k)!} s^{2k}$ of the cosine, which satisfies $\sup_{|s| \leq c} |\cos(s) - p_n(s)| \leq \frac{c^{2n+1}}{(2n+1)!}$. Choose $n \in \mathbb{N}$ large enough such that $\frac{c^{2n+1}}{(2n+1)!} \leq \frac{\varepsilon}{2M}$. Then, by using that $|p_n(s)| \leq \exp(|s|)$ for any $s \in \mathbb{R}$, we obtain that

$$\begin{aligned}
|\cos \circ\, \tilde{a} - &p_n \circ \tilde{a}|_{\mathcal{B}_\psi(X)} \\
&\leq M \sup_{x \in K_R} |\cos(\tilde{a}(x)) - p_n(\tilde{a}(x))| + \sup_{x \in X \setminus K_R} \frac{|\cos(\tilde{a}(x))|}{\psi(x)} + \sup_{x \in X \setminus K_R} \frac{|p_n(\tilde{a}(x))|}{\psi(x)} \\
&< M \sup_{|s| \leq c} |\cos(s) - p_n(s)| + \frac{1}{R} + \sup_{x \in X \setminus K_R} \frac{\exp(|\tilde{a}(x)|)}{\psi(x)} \\
&< M \frac{\varepsilon}{2M} + \frac{1}{R} + \frac{\varepsilon}{4} < \varepsilon.
\end{aligned}$$

Since $\varepsilon > 0$ was chosen arbitrarily, we conclude that $x \mapsto \cos(\tilde{a}(x))$ is in the $|\cdot|_{\mathcal{B}_\psi(X)}$-closure of $\mathcal{A}$. Analogously, the same claim holds true for $x \mapsto \sin(\tilde{a}(x))$. Hence, the subalgebra

$$\mathcal{A}_{\text{trig}} := \left\{ x \mapsto \lambda_1 \cos(\tilde{a}_1(x)) + \lambda_2 \sin(\tilde{a}_2(x)) : \lambda_1, \lambda_2 \in \mathbb{R}, \tilde{a}_1, \tilde{a}_2 \in \tilde{A} \right\}$$

of $\mathcal{B}_\psi(X)$ is contained in the $|\cdot|_{\mathcal{B}_\psi(X)}$-closure of $\mathcal{A}$. Thus, by applying the previous reasoning to the point-separating subalgebra $\mathcal{A}_{\text{trig}} \subset \mathcal{B}_\psi(X)$, which vanishes nowhere and is composed of bounded maps, we conclude that $\mathcal{A}_{\text{trig}}$ is dense in $\mathcal{B}_\psi(X)$. Since $\mathcal{A}_{\text{trig}} \subset \overline{\mathcal{A}}$, it follows that $\mathcal{A}$ is dense in $\mathcal{B}_\psi(x)$. □

3.2.2. Local and Global Approximations

We provide a comparison between local and global approximations, highlighting the advantages of the latter. In particular, building on [27], we prove under mild assumptions that global approximations lead to uniform approximations across all compact subsets that could be considered in the classical setting. We conclude by presenting a global universal approximation result over the space of (time-augmented) rough paths, thereby establishing the global universality of signatures.

Global approximations, as the term implies, have the advantage of approximating functions $f : X \to \mathbb{R}$ across their entire domain, extending beyond compact sets. This capability is particularly significant when X is not locally compact, as is often the case with path spaces.

Furthermore, achieving density in $\mathcal{B}_\psi(X)$, regardless of the chosen admissible weight function ψ, provides a notion of closeness for a broad set of real-valued functions defined over X. As elucidated in Section 3.1.2, the weighted function space $\mathcal{B}_\psi(X)$ encompasses not only all continuous and bounded functions $X \to \mathbb{R}$, but also many unbounded functions whose growth is controlled by ψ. Note that the faster ψ grows, the easier it gets to control the growth of unbounded functions. More precisely, we note that if ψ_1 and ψ_2 are two admissible functions such that $\psi_2(x) \geq \psi_1(x)$ for all $x \in X$, then $\mathcal{B}_{\psi_1}(X) \subset \mathcal{B}_{\psi_2}(X)$.

However, one might object that the admissible weight function interferes with the actual approximation between two functions. Specifically, having $|f - g|_{\mathcal{B}_\psi(X)} \leq \varepsilon$ does not necessarily imply that f and g are close in a pointwise sense. Indeed, the uniform approximation offered by $|\cdot|_{\mathcal{B}_\psi(X)}$ has $|f(x) - g(x)|$ scaled by a factor of $1/\psi(x)$, which may be significant if x is outside a sufficiently large compact set K_R. That said, we argue that the effect of the admissible weight function is often innocuous, especially when we consider local approximations.

To clarify, let us revisit the examples in Section 3.1.1 and consider a normed space $(X, |\cdot|_X)$. This setting already covers many cases of interest. Typically, X is endowed with the norm topology, and hence, given a continuous increasing function $\eta : [0, \infty) \to (0, \infty)$, the map $\psi(x) = \eta(|x|_X)$ is usually not admissible. Instead, as previously discussed, to make ψ admissible and turn X into a weighted space we must weaken its topology.

Remark 3.2: Recall that two topologies are not necessarily comparable. However, in the present context, whenever we weaken a topology, it is implied that the weaker topology is coarser than the original one. Consequently, if we consider a compact subset $K \subset X$ before weakening the topology, then K retains its status as a compact subset once X is turned into a weighted space.

The next result demonstrates that for a fixed compact subset $K \subset X$ with respect to the norm topology — the usual setting of (local) universal approximation results — we obtain the same uniform approximation over K without the factor $1/\psi(x)$ if we view X as a weighted space and approximate functions in $\mathcal{B}_\psi(X)$ instead.

Proposition 3.1: Let $(X, |\cdot|_X)$ be a normed space, and consider any compact subset $K \subset X$ with respect to the usual norm topology. Additionally, set $\psi(x) = \eta(|x|_X)$, where $\eta : [0, \infty) \to (0, \infty)$ is a continuous increasing function. Provided we can turn (X, ψ) into a weighted space and assuming $\mathcal{A}$ to be a dense subset of $\mathcal{B}_\psi(X)$, then for every $f \in \mathcal{B}_\psi(X)$ and $\varepsilon > 0$, there exists a $g \in \mathcal{A}$ such that $\sup_{x \in K} |f(x) - g(x)| \leq \varepsilon$.

Proof. Let $K \subset X$ be a compact subset with respect to the norm topology. Note that K is necessarily bounded with respect to $|\cdot|_X$. Indeed, for every $x \in K$, let $U_1(x)$ denote the open ball of radius 1 centred at x. Then, $K \subset \bigcup_{x\in K} U_1(x)$ and, by compactness, there exists a finite subcovering $\{U_1(x_i)\}_{i=1}^n$ such that $K \subset \bigcup_{i=1}^n U_1(x_i)$. Hence, K is bounded.

Now, by considering a coarser topology if needed, let (X, ψ) be a weighted space. Since K is bounded with respect to $|\cdot|_X$, there exists an $R > 0$ such that $K \subset K_R = \psi^{-1}((0, R])$. Moreover, for every $f \in \mathcal{B}_\psi(X)$ and $\varepsilon > 0$, there exists $g \in \mathcal{A}$ such that $|f - g|_{\mathcal{B}_\psi(X)} \leq \varepsilon/R$. Hence, we see that

$$\sup_{x\in K} |f(x) - g(x)| \leq \sup_{x\in K_R} |f(x) - g(x)| \leq R \sup_{x\in K_R} \frac{|f(x) - g(x)|}{\psi(x)} \leq R|f - g|_{\mathcal{B}_\psi(X)} \leq \varepsilon.$$

□

Under the assumptions of Proposition 3.1, the advantage of global approximations becomes clear: not only can we consider the entire space X as the domain of approximation, but we can also achieve uniform approximations across all the compact sets considered in the local classical setting, resulting in a more streamlined framework. Note that for a fixed precision $\varepsilon > 0$, the choice of g is dependent on K, implying that the uniform approximation occurs inside a given compact K, and not across all compacts simultaneously. However, this is also the case in local approximations: whenever we change the compact domain, the approximating function changes as well. In this sense, global approximations encapsulate local approximations entirely.

We conclude this section by stating a global approximation result for signatures, which also serves as an example of application of Proposition 3.1. First, however, we define the rough analogue of the time-augmented paths in Definition 2.8. As in Theorem 2.7, time augmentation guarantees that the set of linear functionals of the signature is point-separating. We will also make use of time-augmented rough paths in the upcoming section.

Definition 3.4: We define the subset $\hat{C}_T^\alpha$ of time-augmented α-Hölder rough paths by

$$\hat{C}_T^\alpha := \left\{ \hat{\mathbf{X}} \in C_o^{\alpha\text{-Höl}}([0,T], G^{\lfloor 1/\alpha \rfloor}(\mathbb{R}^{d+1})) : \hat{X}_t = (t, X_t), \text{ for all } t \in [0,T] \right\},$$

where $X \in C_o^{\alpha\text{-Höl}}([0,T], \mathbb{R}^d)$ and time corresponds to the 0-th coordinate.

Following [27], we turn $\hat{C}_T^\alpha$ into a weighted space by equipping $C_o^{\alpha\text{-Höl}}([0,T], G^{\lfloor 1/\alpha \rfloor}(\mathbb{R}^{d+1}))$ with a β-topology for $\beta < \alpha$, i.e., the topology induced by the homogenous β-Hölder distance $d_{\beta\text{-Höl};[0,T]}$, for instance (see Definition 1.21 and Theorem 1.10). By doing so, any function of the form $\psi(\hat{\mathbf{X}}) = \eta(|\hat{\mathbf{X}}|_{\alpha\text{-Höl};[0,T]})$ becomes admissible.

Remark 3.3: In [27], the authors show that endowing $\hat{C}_T^\alpha$ with a topology induced by $d_{\beta\text{-Höl};[0,T]}$ for $0 \leq \beta < \alpha$, or even with the weak-$*$-topology, leads to the same weighted function space $\mathcal{B}_\psi(\hat{C}_T^\alpha)$. We refer to the original work for more details.

Theorem 3.4 ([27], Theorem 5.4)**:** Consider the weighed space $(\hat{C}_T^\alpha, \psi)$ with admissible weight function $\psi(\hat{\mathbf{X}}) = \exp\left(\lambda |\hat{\mathbf{X}}|^\kappa_{\alpha\text{-Höl};[0,T]}\right)$ for $\lambda > 0$ and $\kappa > \lfloor 1/\alpha \rfloor$. Then, the set

$$\operatorname{span}\left\{ \hat{\mathbf{X}} \mapsto \left\langle e_w, S(\hat{\mathbf{X}}) \right\rangle : w \in \{0, 1, \ldots, d\}^N, N \in \mathbb{N}_0 \right\}$$

is dense in $\mathcal{B}_\psi(\hat{C}_T^\alpha)$. Precisely, for every function $f \in \mathcal{B}_\psi(\hat{C}_T^\alpha)$ and $\varepsilon > 0$, there exists a linear functional of the form $L(S(\hat{\mathbf{X}})) = \sum_{0 \leq |w| \leq N} a_w \left\langle e_w, S(\hat{\mathbf{X}}) \right\rangle$, with $N \in \mathbb{N}_0$ and $a_w \in \mathbb{R}$, such that

$$\sup_{\mathbf{X} \in \hat{C}_T^\alpha} \frac{\left| f(\hat{\mathbf{X}}) - L(S(\hat{\mathbf{X}})) \right|}{\psi(\hat{\mathbf{X}})} \leq \varepsilon.$$

We refrain from including the proof as it would deviate from the current discussion. However, the result above ultimately follows from a direct application of Theorem 3.3. More importantly for our purposes is to note that $(\hat{C}_T^\alpha, |\cdot|_{\alpha\text{-Höl};[0,T]})$ is a normed space which we can turn into a weighted space by replacing the topology induced by $|\cdot|_{\alpha\text{-Höl};[0,T]}$ with the topology induced by $|\cdot|_{\beta\text{-Höl};[0,T]}$ for some $\beta < \alpha$, and considering the usual admissible function $\psi(\hat{\mathbf{X}}) = \eta(|\hat{\mathbf{X}}|_{\alpha\text{-Höl};[0,t]})$. Thanks to Proposition 1.5, the β-topology is coarser than the α-topology, and the conditions of Proposition 3.1 are met, showing that Theorem 3.4 encapsulates most local approximation results in the literature of signature-based methods.

3.3. Globally Universal Kernels

As discussed in Section 2.2.2, a continuous kernel k defined on a compact metric space X is deemed universal if its RKHS is dense in $C(X)$. In addition, the assessment of universality, as inferred from Theorem 2.5, involves an application of the Stone-Weierstrass theorem (Appendix C). Now, equipped with the weighted Stone-Weierstrass result (Theorem 3.3), one may contemplate how this resonates with the notion of universal kernels. This prompts the introduction of what we term *globally universal kernels.*

Section 3.3.1 delves into the definition and existence of these kernels. In particular, it offers a method to construct globally universal kernels on weighted spaces. Section 3.3.2 provides examples of globally universal kernels. Both sections draw inspiration from [20], where a method for constructing universal (Taylor) kernels defined on compact metric spaces is presented.

3.3.1. Globally Universal Kernels on Weighted Spaces

In [20], the authors begin by constructing an explicit universal (Taylor) kernel defined on a compact set of $l_2(\mathbb{N})$, thus extending Corollary 2.2. Subsequently, by leveraging the fact that every separable Hilbert space is isometrically isomorphic to $l_2(\mathbb{N})$ (Appendix B), the authors proceed to devise universal kernels (in the sense of Definition 2.5) defined on a generic compact metric space X, provided there exists a separable Hilbert space $\mathcal{H}$ and a continuous injective map $\rho : X \to \mathcal{H}$. This prompts the question of whether a similar endeavour is conceivable in the realm of weighted spaces. We begin this exploration by defining the analogue of universal kernels within the context of weighted spaces.

Definition 3.5: Let (X, ψ) be a weighted space. A kernel $k : X \times X \to \mathbb{R}$ is called globally universal if the RKHS $\mathcal{H}$ of k is dense in $\mathcal{B}_\psi(X)$, i.e. for all $f \in \mathcal{B}_\psi(X)$ and $\varepsilon > 0$ there exists a $g \in \mathcal{H}$ such that

$$\sup_{x \in X} \frac{|f(x) - g(x)|}{\psi(x)} \leq \varepsilon.$$

Remark 3.4: According to Definition 2.5, for a kernel k to be universal it has to be continuous. In our context, however, we refrain from making this assumption. Briefly put, the justification lies in the observation that the topology accompanying the weighted space is frequently too weak to permit k to be jointly continuous.

We now move forward to establish a test for global universality, which serves as a method to ascertain whether a kernel k is globally universal. As discussed in Section 2.2.2, the typical approach to proving the universality of a kernel involves the utilisation of the classical version of the Stone-Weierstrass theorem. However, for our specific objectives, we require the weighted version of this theorem (Theorem 3.3).

Let J denote a non-empty countable set. We keep denoting the space of square summable sequences indexed by J by $l_2(J)$, but we abbreviate $l_2(\mathbb{N})$ to l_2. Furthermore, we use $\mathbb{N}_0^{\mathbb{N}}$ to represent the set of all sequences $j = (j_i)_{i \in \mathbb{N}}$ with values in $\mathbb{N}_0$, and define $|j|$ to be $|j| := \sum_{i=1}^{\infty} j_i$. It is worth noting that $|j| < \infty$ if and only if j has only finitely many nonzero components. The next result is an analogue to Theorem 2.5.

Theorem 3.5: Let (X, ψ) be a weighted space and k be a kernel on X with $k(x, x) > 0$ for all $x \in X$. Assume that we have an injective feature map $\Phi : X \to l_2(J)$ of k, where J is some countable set. Denote by Φ_j its j-th component, i.e. $\Phi(x) = (\Phi_j(x))_{j \in J}$, for all $x \in X$. If $\mathcal{A} := \operatorname{span}\{\Phi_j : j \in J\}$ is a $\psi-$moderate growth subalgebra of $\mathcal{B}_\psi(X)$, then k is globally universal.

Proof. We apply Theorem 3.3. First, observe that $\mathcal{A}$ necessarily does not vanish. Indeed, for any $x \in X$ we have $|\Phi(x)|^2_{l_2(J)} = k(x,x) > 0$. Moreover, the injectivity of Φ implies that $\mathcal{A}$ is point separating. Hence, by the weighted Stone-Weierstrass theorem, $\mathcal{A}$ is dense in $\mathcal{B}_\psi(X)$, i.e. for all $f \in \mathcal{B}_\psi(X)$ and $\varepsilon > 0$, there exists $g \in \mathcal{A}$ of the form

$$g(x) = \sum_{i=1}^{m} \alpha_i \Phi_{j_i}(x),$$

such that $|f - g|_{\mathcal{B}_\psi(X)} \leq \varepsilon$. Moreover, there exists $h \in l_2(J)$ such that $g = \langle h, \Phi(\cdot)\rangle_{l_2(J)}$. Simply take $h_j = \alpha_i$ if $j = j_i$ for $i \in \{1, ..., m\}$, and $h_j = 0$ otherwise. Since the RKHS of k is composed precisely of maps of the form $\langle h, \Phi(\cdot)\rangle_{l_2(J)}$ for some h (see (2.1)), we conclude that k is globally universal. □

With a test for global universality in hand, we can now begin to investigate methods for constructing globally universal kernels, thereby establishing their existence. Following [20], we initially define a kernel over l_2. The significance of l_2 arises from its status as the archetype of separable Hilbert spaces: every separable Hilbert space is homeomorphic to l_2 (Appendix B). Subsequently, we leverage this property to endeavour a construction of globally universal kernels on a generic weighted space. The next proposition generalises Proposition 2.1.

Lemma 3.1 ([20], Lemma 4.2)**:** Assume that $n \in \mathbb{N}$ is fixed. Then, for all $j \in \mathbb{N}_0^{\mathbb{N}}$ with $|j| = n$, there exists a constant $c_j \in (0, \infty)$ such that for all summable sequences $(b_i)_{i \in \mathbb{N}} \subset [0, \infty)$ we have

$$\left(\sum_{i=1}^{\infty} b_i\right)^n = \sum_{j \in \mathbb{N}_0^{\mathbb{N}} : |j| = n} c_j \prod_{i=1}^{\infty} b_i^{j_i}.$$

Proposition 3.2 ([20], Proposition 4.3)**:** Let $f : \mathbb{R} \to \mathbb{R}$ be a function globally expressible by its Taylor series expanded at zero, i.e. $f(t) = \sum_{n \geq 0} a_n t^n$ for all $t \in \mathbb{R}$. Consider $J := \{j \in \mathbb{N}_0^{\mathbb{N}} : |j| < \infty\}$. If $a_n \geq 0$ for all $n \geq 0$, then $k : l_2 \times l_2 \to \mathbb{R}$ defined by

$$k(z, z') = f\big(\langle z, z'\rangle_{l_2}\big) = \sum_{n=0}^{\infty} a_n \langle z, z'\rangle_{l_2}^n,$$

is a kernel. Moreover, for all $j \in J$ there exits a $c_j \in (0, \infty)$ such that $\Phi : l_2 \to l_2(J)$ given by

$$\Phi(z) = \left(c_j \prod_{i=1}^{\infty} z_i^{j_i}\right)_{j \in J},$$

is a feature map of k, where we use the convention $0^0 = 1$.

Proof. Since f is defined over $\mathbb{R}$, the kernel k is well-defined. As $f(t)$ is absolutely convergent, Lemma 3.1 shows that, for all $j \in J$, there exists a constant $\tilde{c}_j \in (0,\infty)$ such that

$$k(z,z') = \sum_{n=0}^{\infty} a_n \langle z, z'\rangle_{l_2}^n = \sum_{n=0}^{\infty} a_n \left(\sum_{i=1}^{\infty} z_i z_i'\right)^n = \sum_{n=0}^{\infty} a_n \sum_{j\in\mathbb{N}_0^{\mathbb{N}}:|j|=n} \tilde{c}_j \prod_{i=1}^{\infty} (z_i z_i')^{j_i} = \\ = \sum_{j\in J} a_{|j|}\tilde{c}_j \prod_{i=1}^{\infty} z_i^{j_i} \prod_{i=1}^{\infty} z_i'^{j_i}.$$

Thus, by setting $c_j = \sqrt{a_{|j|}\tilde{c}_j}$ we obtain that Φ is a feature map of k, i.e. $k(z,z') = \langle \Phi(z), \Phi(z')\rangle_{l_2}$. This proves that k is a kernel. □

With a kernel defined on l_2, our expectation is that by demonstrating its global universality, we can exploit the isomorphism between l_2 and some feature space $\mathcal{H}$ to construct a kernel on a given weighted space X, assuming X can be embedded into $\mathcal{H}$. We now proceed to establish the main result of this section. However, before doing so, we state two auxiliary lemmas.

Lemma 3.2: Let (X, ψ_X) be a weighted space and consider a completely regular Hausdorff space (Z, τ_Z). Assume that there exists a continuous bijection $h : X \to Z$. Then, the map $\psi_Z := \psi_X \circ h^{-1}$ is an admissible weight function, rendering Z a weighted space.

Proof. Consider an arbitrary $R > 0$. We need to show that the pre-image $K_R^Z := \psi_Z^{-1}((0,R])$ is a compact subset with respect to τ_Z. By assumption, we know that $K_R^X := \psi_X^{-1}((0,R])$ is compact. Now, simply observe that $K_R^Z = h \circ \psi_X^{-1}((0,R]) = h(K_R^X)$. Given the continuity of h, it follows that K_R^Z is compact (Appendix C). □

Lemma 3.3: Let k_Z denote a kernel defined on a non-empty set Z. Assume there exists a continuous bijection $h : X \to Z$. Then k_Z induces a kernel k_X on X such that there is a one-to-one correspondence between the respective RKHSs $\mathcal{H}_Z$ and $\mathcal{H}_X$. Precisely, we have that $\mathcal{H}_X = \{f \circ h : f \in \mathcal{H}_Z\}$.

Proof. Let $\Phi_Z^c : Z \to \mathcal{H}_Z$ denote the canonical feature map of k_Z. Define $k_X : X \times X \to \mathbb{R}$ by setting

$$k_X(x,x') := k_Z(h(x), h(x')) = \langle \Phi_Z^c \circ h(x), \Phi_Z^c \circ h(x')\rangle_{\mathcal{H}_Z}.$$

Then, it is apparent that k_X is a kernel with feature map $\Phi_X := \Phi_Z^c \circ h$. Furthermore, by (2.1), we have that $\mathcal{H}_X = \{\langle f, \Phi_Z^c \circ h(\cdot)\rangle_{\mathcal{H}_Z} : f \in \mathcal{H}_Z\}$ and, by the reproducing property, we note that $\langle f, \Phi_Z^c \circ h(x)\rangle_{H_Z} = f \circ h(x)$. Hence,

$$\mathcal{H}_X = \{f \circ h : f \in \mathcal{H}_Z\},$$

and $f \in \mathcal{H}_Z$ if and only if $f \circ h \in \mathcal{H}_X$, i.e. there exists a bijection between $\mathcal{H}_X$ and $\mathcal{H}_Z$. □

Remark 3.5: The construction described in Lemma 3.3 can be applied to a generic map $h : X \to Z$, and $\mathcal{H}_X$ is commonly known as the pull-back of $\mathcal{H}_Z$ ([61], Section 5.4).

Theorem 3.6: Let (X, ψ_X) be a weighted space and $\mathcal{H}$ be a separable infinite-dimensional Hilbert space such that there exists a continuous injective map $\rho : X \to \mathcal{H}$. Set $Z := I(\rho(X))$, where I denotes the homeomorphism between $\mathcal{H}$ and l_2. Let $f : \mathbb{R} \to \mathbb{R}$ be a function globally expressible by its Taylor series expanded at zero, i.e. $f(t) = \sum_{n\geq 0} a_n t^n$ for all $t \in \mathbb{R}$. Lastly, define $\psi_Z := \psi_X \circ \rho^{-1} \circ I^{-1}$ over Z and assume that $\psi_Z(z) \geq \exp(|z|_{l_2}^{\gamma})$, with $\gamma > 1$. Then, the following statements hold:

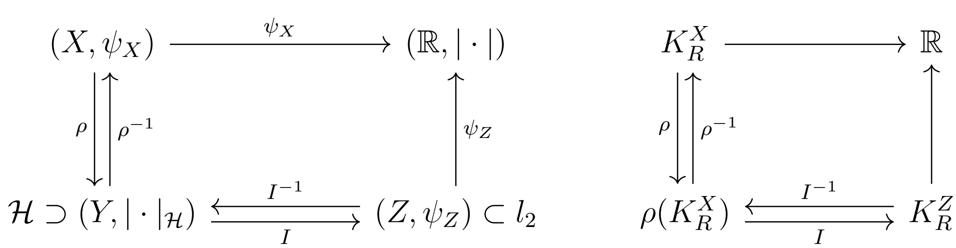
- If $a_n \geq 0$ for all $n \in \mathbb{N}_0$, then $k : X \times X \to \mathbb{R}$ given by

$$k(x, x') := f(\langle \rho(x), \rho(x') \rangle_{\mathcal{H}}) = \sum_{n=0}^{\infty} a_n \langle \rho(x), \rho(x') \rangle_{\mathcal{H}}^n \tag{3.4}$$

  defines a kernel on X.

- If $a_n \geq 0$ for all $n \geq 1$, and $a_0 > 0$, then k is globally universal.

Proof. **Step 1: Z is a weighted space.** First, recall that I is not only a homeomorphism, but an isometric isomorphism. Set $Y := \rho(X)$ so that $Z = I(Y)$. Observe that $I \circ \rho$, as a map from X to Z, is a continuous bijection. Hence, by Lemma 3.2, ψ_Z constitutes an admissible weight function on Z, and there is a bijection between the compact sets $K_R^Z := \psi_Z^{-1}((0, R])$ and $K_R^X := \psi_X^{-1}((0, R])$. Note that l_2 needs to have the norm topology, otherwise l_2 would not be homeomorphic to $\mathcal{H}$. Consequently, Z has the norm topology inherited from l_2. Nevertheless, we are able to define an admissible weight function on Z, which yields the weighted space (Z, ψ_Z). Schematically,



Step 2: Bijection between $\mathcal{B}_{\psi_X}(X)$ and $\mathcal{B}_{\psi_Z}(Z)$. We show that there exists a bijection between the weighted function spaces $\mathcal{B}_{\psi_X}(X)$ and $\mathcal{B}_{\psi_Z}(Z)$. Given that $(Y, |\cdot|_{\mathcal{H}})$ is a Hausdorff space, and that ρ restricted to K_R^X is a continuous bijection onto $\rho(K_R^X)$, we have that both I and ρ are homeomorphisms in the right diagram. Consequently, for all $R > 0$, $g \in C(K_R^Z)$ if and only if $g \circ I \circ \rho \in C(K_R^X)$. Additionally,

$$\lim_{R\to\infty} \sup_{z \in Z \setminus K_R^Z} \frac{|g(z)|}{\psi_Z(z)} = 0 \quad \text{if and only if} \quad \lim_{R\to\infty} \sup_{x \in X \setminus K_R^X} \frac{|g(I(\rho(x)))|}{\psi_X(x)} = 0,$$

since

$$\lim_{R\to\infty} \sup_{x\in X\setminus K_R^X} \frac{|g(I(\rho(x)))|}{\psi_Z(I(\rho(x)))} = \lim_{R\to\infty} \sup_{z\in Z\setminus K_R^Z} \frac{|g(z)|}{\psi_Z(z)}.$$

Therefore, by Theorem 3.1, $g \in \mathcal{B}_{\psi_Z}(Z)$ if and only if $g \circ I \circ \rho \in \mathcal{B}_{\psi_X}(X)$.

Step 3: Globally universal kernel on Z. Using the test for global universality in Theorem 3.5, we prove that the kernel of Proposition 3.2 restricted to Z is globally universal under an additional mild assumption. Precisely, let $k_Z : Z \times Z \to \mathbb{R}$ denote the kernel defined by

$$k_Z(z, z') = \sum_{n=0}^{\infty} a_n \langle z, z' \rangle_{l_2}^n,$$

assuming that $a_n \geq 0$ for all $n \geq 1$, and $a_0 > 0$. As inferred from Proposition 3.2, k_Z has a feature map $\Phi : l_2 \to l_2(J)$ given by

$$\Phi(z) = \left(c_j \prod_{i=1}^{\infty} z_i^{j_i} \right)_{j\in J},$$

where $J = \{j \in \mathbb{N}_0^{\mathbb{N}} : |j| < \infty\}$ and $c_j > 0$ for all $j \in J$. We equip Z with the admissible weight function ψ_Z, and consider the weighted space of Step 1. We proceed to verify the assumptions of Theorem 3.5. First, note that

$$k_Z(z, z) = \sum_{n=0}^{\infty} a_n |z|_{l_2}^{2n} \geq a_0 > 0.$$

Regarding the feature map, if $z \neq z'$, then there exists an $i \in \mathbb{N}$ such that $z_i \neq z'_i$. Thus, for the multi-index $j \in J$ such that $j_i = 1$ and vanishes everywhere else, we have that $\Phi_j(z) = c_j z_i \neq c_j z'_i = \Phi_j(z')$. Hence, Φ is injective. We are only left with proving the conditions for the weighted Stone-Weierstrass (Theorem 3.3).

Following the notation of Theorem 3.5, we have that

$$\mathcal{A} := \operatorname{span} \left\{ c_j \prod_{i=1}^{\infty} z_i^{j_i} : j \in J \right\},$$

is an algebra. Note that since $|j| < \infty$, we have for all $j \in J$ that $\Phi_j(z) = c_j z_{i_1}^{j_{i_1}} \dots z_{i_n}^{j_{i_n}}$, where n corresponds to the number of nonzero components of j.

Additionally, each $\Phi_j \in \mathcal{A}$ is continuous over l_2 with respect to the norm topology, and hence continuous over K_R^Z, for all $R > 0$. This follows from the fact that the projection mapping $\pi_i(z) = z_i$ is norm continuous.

Moreover, by the assumption that $\psi_Z(z) \geq \exp(|z|_{l_2}^{\gamma})$ with $\gamma > 1$, we observe that

$$\begin{aligned}
\lim_{R\to\infty} \sup_{z\in Z\setminus K_R^Z} \frac{|\Phi_j(z)|}{\psi_Z(z)} &\leq \lim_{R\to\infty} \sup_{z\in Z\setminus K_R^Z} \left| c_j \prod_{i=1}^{n} z_i^{j_i} \right| \Big/ \exp\left(|z|_{l_2}^{\gamma}\right) \\
&\leq \lim_{R\to\infty} \sup_{z\in Z\setminus K_R^Z} c_j \left(\max_{1\leq k\leq n} |z_{i_k}| \right)^{|j|} \Big/ \exp\left(|z|_{l_2}^{\gamma}\right) \\
&\leq \lim_{R\to\infty} \sup_{z\in Z\setminus K_R^Z} c_j \left(\sum_{k=1}^{n} z_{i_k}^2 \right)^{|j|/2} \Big/ \exp\left(|z|_{l_2}^{\gamma}\right) \\
&\leq \lim_{R\to\infty} \sup_{z\in Z\setminus K_R^Z} \frac{c_j \, |z|_{l_2}^{|j|}}{\exp\left(|z|_{l_2}^{\gamma}\right)} = 0
\end{aligned}$$

Therefore, $\mathcal{A} \subset \mathcal{B}_{\psi_Z}(Z)$ by Theorem 3.1. Now, set $J_1 = \{j \in J \mid \exists i \in \mathbb{N} : j_i = 1 \text{ and } j_k = 0 \text{ for } k \neq i\}$ and consider $\tilde{\mathcal{A}} \subset \mathcal{A}$ given by

$$\tilde{\mathcal{A}} := \operatorname{span}\left\{ c_j \prod_{i=1}^{\infty} z_i^{j_i} : j \in J_1 \right\} = \operatorname{span}\left\{ c_j z_i : i \in \mathbb{N},\ j \in J_1 \right\},$$

which is a point separating vector subspace. Note that, for $\tilde{\Phi} \in \tilde{\mathcal{A}}$,

$$\begin{aligned}
\lim_{R\to\infty} \sup_{z\in Z\setminus K_R^Z} \frac{\exp|\tilde{\Phi}(z)|}{\psi_Z(z)} &\leq \lim_{R\to\infty} \sup_{z\in Z\setminus K_R^Z} \exp\left(\left| \sum_{k=1}^{n} \alpha_k z_{i_k} \right| - |z|_{l_2}^{\gamma} \right) \\
&\leq \lim_{R\to\infty} \sup_{z\in Z\setminus K_R^Z} \exp\left(\max_{1\leq k\leq n} |\alpha_k| \sum_{k=1}^{n} |z_{i_k}| - |z|_{l_2}^{\gamma} \right) \\
&\leq \lim_{R\to\infty} \sup_{z\in Z\setminus K_R^Z} \exp\left(\max_{1\leq k\leq n} |\alpha_k| \sqrt{n} \cdot |z|_{l_2} - |z|_{l_2}^{\gamma} \right) = 0,
\end{aligned}$$

since $\gamma > 1$ by assumption. Note that $\max_{1\leq k\leq n} |\alpha_k|\sqrt{n}$ depends only on $\tilde{\Phi}$ and not on z, hence the constants are not affected by the supremum. Moreover, the exponential function preserves the continuity over K_R^Z. Therefore, $\mathcal{A}$ is a $\psi-$moderate growth subalgebra and we conclude, by Theorem 3.5, that k_Z is globally universal.

Step 4: k_Z induces a globally universal kernel in X. By Lemma 3.3, k_Z induces a kernel k_X on X, with $h \equiv I \circ \rho$ and $\mathcal{H}_X = \{g \circ I \circ \rho : g \in \mathcal{H}_Z\}$. Furthermore, considering that I is an isometry, we have that

$$k_X(x, x') = k_Z(I \circ \rho(x), I \circ \rho(x')) = \sum_{n=0}^{\infty} a_n \langle I(\rho(x)), I(\rho(x')) \rangle_{l_2}^{n} = \sum_{n=0}^{\infty} a_n \langle \rho(x), \rho(x') \rangle_{\mathcal{H}}^{n},$$

which is precisely the kernel in (3.4). Finally, we show that k_X is globally universal. Specifically, consider some $g \in \mathcal{B}_{\psi_X}(X)$. Then $\tilde{g} := g \circ \rho^{-1} \circ I^{-1} \in \mathcal{B}_{\psi_Z}(Z)$ can be approximated by some

$h \in \mathcal{H}_Z$, i.e. $|\tilde{g} - h|_{\mathcal{B}_{\psi_Z}(Z)} \leq \varepsilon$, for an arbitrary $\varepsilon > 0$. Now, simply note that

$$\sup_{z\in Z} \frac{|\tilde{g}(z) - h(z)|}{\psi_Z(z)} = \sup_{z\in Z} \frac{|g \circ \rho^{-1} \circ I^{-1}(z) - h(z)|}{\psi_X \circ \rho^{-1} \circ I^{-1}(z)} \leq \varepsilon \implies \sup_{x\in X} \frac{|g(x) - h \circ I \circ \rho(x)|}{\psi_X(x)} \leq \varepsilon.$$

Since $h \circ I \circ \rho \in \mathcal{H}_X$, we conclude that k_X is globally universal, thus finishing the proof. □

Remark 3.6: We observe that although we assume ρ to be continuous, it suffices to assume continuity over all compact sets $K_R := \psi_X^{-1}((0, R])$, where $R > 0$. Additionally, it is worth highlighting once again that the map ψ_Z will inherently have compact pre-images whenever Z is equipped with the subspace topology, which aligns with the topology induced by the norm in l_2. This may come as a surprise considering our discussion in Section 3.1.1, where we pointed out that typically one needs to consider a weaker topology to obtain a weighted space. That said, Z being a weighted space with respect to the norm topology does not come for free, and relies on the additional structure provided by Lemma 3.2.

Remark 3.7: Initially, fulfilling the final assumption of Theorem 3.6, i.e., $\psi_Z(z) \geq \exp(|z|_{l_2}^{\gamma})$ with $\gamma > 1$, might appear somewhat complicated. However, it is worth noting that there is typically considerable flexibility in selecting ψ_X. Consequently, to meet this criterion, one can opt for an admissible weight function ψ_X over X, ensuring that $\psi_X(\rho^{-1}(y))$ dominates $\exp(|y|_{\mathcal{H}}^{\gamma})$ for all $y \in I^{-1}(Z)$. It is important to recall that I represents an isometry.

3.3.2. Examples of Globally Universal Kernels

We provide examples of globally universal kernels. In particular, we consider the weighted rough path space in Definition 3.4 (see [27] for details) and define a family of globally universal kernels that we refer to as *Taylor signature kernels*. These kernels provide a collection of functionals of the signature capable of approximating functions over the entire space of rough paths.

Let $X = \mathbb{R}^d$. Various (Taylor) kernels can be defined over $\mathbb{R}^d$, including:

- The exponential kernel: $k_1(x, x') = \exp(\langle x, x' \rangle_{\mathbb{R}^d})$
- The polynomial kernel: $k_2(x, x') = \big(b + c\langle x, x' \rangle_{\mathbb{R}^d}\big)^d$, with $b, c, d > 0$
- The hyperbolic cosine kernel: $k_3(x, x') = \cosh\big(\langle x, x' \rangle_{\mathbb{R}^d}\big)$

All kernels can be shown to be universal in the classical sense when the domain is restricted to a fixed compact set $K \subset \mathbb{R}^d$ ([20], Theorem 2.2). Next, we demonstrate that these kernels are globally universal in the sense of Definition 3.5. It is important to emphasise once again that we now consider the entire domain of any of the kernels mentioned above, rather than being limited to a compact subset. The domain of a given Taylor kernel is determined by the radius of convergence of its Taylor expansion. The series for the kernels above all converge in $\mathbb{R}$, hence all kernels have $\mathbb{R}^d$ as their domain.

Example 3.6: We begin with the exponential kernel. Consider the function $\psi_X(x) = \eta(|x|)$, where $\eta : [0,\infty) \to (0,\infty)$ is a continuous, increasing function. Recall that $(\mathbb{R}^d, \psi_X)$ is a weighted space (Example 3.1). We need to verify the assumptions of Theorem 3.6. To this end, consider $\rho : \mathbb{R}^d \to l_2$ defined by $x \mapsto (x^1, x^2, \ldots, x^d, 0, 0, \ldots)$. Clearly, l_2 is a separable infinite-dimensional Hilbert space, and ρ is continuous and injective since it is an isometry onto $\rho(\mathbb{R}^d)$. It remains to verify that $\psi_Z(z) := \psi_X \circ \rho^{-1} \circ I^{-1}(z) \geq \exp(|z|_{l_2}^\gamma)$, for $\gamma > 1$ and an appropriate choice of ψ_X. Note that $I : l_2 \to l_2$ is taken to be the identity map. Set $\eta(t) = \exp(t^\gamma)$ and observe that, for all $z \in Z \equiv I(\rho(\mathbb{R}^d))$,

$$\psi_Z(z) = \psi_X\left(\rho^{-1} \circ I^{-1}(z)\right) = \exp\left(|\rho^{-1} \circ I^{-1}(z)|^\gamma\right) = \exp(|z|_{l_2}^\gamma),$$

since both ρ and I are isometries over $\mathbb{R}^d$ and $\rho(\mathbb{R}^d)$, respectively. Hence, by Theorem 3.6, the exponential kernel k_1 is globally universal.

Example 3.7: The remaining kernels k_2 and k_3 follow easily from the previous example. Indeed, using the notation of Theorem 3.6 only the function f changes. For k_2, we have

$$f(t) = (b + ct)^d = \sum_{k=0}^{d} \binom{d}{k} b^{d-k} c^k t^k,$$

for all $t \in \mathbb{R}$, whereas for k_3,

$$f(t) = \sum_{k=0}^{\infty} \frac{t^{2k}}{(2k)!} = 1 + \frac{t^2}{2!} + \frac{t^4}{4!} + \cdots$$

Given that both expansions have strictly positive coefficients, it follows by Theorem 3.6 that k_2 and k_3 are globally universal kernels.

We now consider a more intricate example that utilises the tools from the previous chapters. This example will lead to a family of globally universal kernels on a space of rough paths. Following [27], we consider a subset of $\hat{C}_T^\alpha$ (Definition 3.4), which we now define.

Definition 3.6: Consider $p \geq 1$ and $\alpha \in (0, 1]$ such that $p\alpha < 1$. We say $\mathbf{X} : [0, T] \to G^{\lfloor 1/\alpha \rfloor}(\mathbb{R}^d)$ with $\mathbf{X}_0 = \mathbf{1} \in G^{\lfloor 1/\alpha \rfloor}(\mathbb{R}^d)$ is a weakly geometric (p, α)-rough path if the (p, α)-norm

$$|\mathbf{X}|_{cc,p,\alpha} := \sup_{0 \leq s < t \leq T} \frac{d_{cc}(\mathbf{X}_s, \mathbf{X}_t)}{|s - t|^\alpha} + \left(\sup_{(t_i) \in \mathcal{P}([0,T])} \sum_i d_{cc}(\mathbf{X}_{t_i}, \mathbf{X}_{t_{i+1}})^p \right)^{\frac{1}{p}}$$

is finite. We denote by $C_o^{p,\alpha}\big([0,T],G^{\lfloor 1/\alpha\rfloor}(\mathbb{R}^d)\big)$ the space of weakly geometric (p,α)-rough paths that preserve the origin. Lastly, we endow this space with a metric given by

$$d_{cc,q,\beta}(\mathbf{X},\mathbf{Y}) \coloneqq \sup_{0\le s<t\le T}\frac{d_{cc}(\mathbf{X}_{s,t},\mathbf{Y}_{s,t})}{|s-t|^{\beta}} + \left(\sup_{(t_i)\in\mathcal{P}([0,T])}\sum_i d_{cc}(\mathbf{X}_{t_i,t_{i+1}},\mathbf{Y}_{t_i,t_{i+1}})^q\right)^{\frac{1}{q}},$$

for $\mathbf{X},\mathbf{Y}\in C_o^{p,\alpha}\big([0,T],G^{\lfloor 1/\alpha\rfloor}(\mathbb{R}^d)\big)$ and $(q,\beta)\in[p,\infty)\times(0,\alpha]$ with $q\beta<1$.

Similarly to $C_o^{\alpha\text{-Höl}}\big([0,T],G^{\lfloor 1/\alpha\rfloor}(\mathbb{R}^d)\big)$, by equipping $C_o^{p,\alpha}\big([0,T],G^{\lfloor 1/\alpha\rfloor}(\mathbb{R}^d)\big)$ with the metric $d_{cc,q,\beta}$ such that $(q,\beta)\in(p,\infty)\times(0,\alpha)$ and $q\beta<1$, we obtain a weighted space due to the compact embedding

$$C_o^{p,\alpha}\big([0,T],G^{\lfloor 1/\alpha\rfloor}(\mathbb{R}^d)\big)\hookrightarrow C_o^{q,\beta}\big([0,T],G^{\lfloor 1/\alpha\rfloor}(\mathbb{R}^d)\big).$$

This embedding follows essentially from Proposition 1.5 and a reasoning analogous to the one in Proposition 1.6. We refer to ([27], Theorem A.8) for a precise statement. In the spirit of Definition 3.4, we consider the subset of time-augmented (p,α)-rough paths

$$\hat{C}_T^{p,\alpha}\coloneqq\left\{\hat{\mathbf{X}}\in C_o^{p,\alpha}\big([0,T],G^{\lfloor 1/\alpha\rfloor}(\mathbb{R}^{d+1})\big) : \hat{X}_t=(t,X_t),\ \text{for all } t\in[0,T]\right\}.$$

Observe that $\hat{C}_T^{p,\alpha}\subset\hat{C}_T^{\alpha}$.

Example 3.8: Consider $X=\hat{C}_T^{p,\alpha}$. We equip $C_o^{p,\alpha}\big([0,T],G^{\lfloor 1/\alpha\rfloor}(\mathbb{R}^{d+1})\big)$ with a metric $d_{cc,q,\beta}$ such that $(q,\beta)\in(p,\infty)\times(0,\alpha)$, $q\beta<1$ and $\lfloor 1/\beta\rfloor=\lfloor 1/\alpha\rfloor$. Then, we have that:

1. For any continuous increasing function η, the map $\psi_X(\hat{\mathbf{X}})=\eta(|\hat{\mathbf{X}}|_{cc,p,\alpha})$ is admissible, turning $(\hat{C}_T^{p,\alpha},\psi_X)$ into a weighted space ([27], Example 2.3 (vii)).
2. Since $\lfloor 1/\beta\rfloor=\lfloor 1/\alpha\rfloor$, it follows by Proposition 1.23 that the signature map, i.e., the Lyons' extension map, $S:\hat{C}_T^{p,\alpha}\to T_1((\mathbb{R}^d))$, is continuous when $\hat{C}_T^{p,\alpha}$ is endowed with the metric $d_{\beta\text{-Höl};[0,T]}$. Additionally, it is clear that $d_{\beta\text{-Höl};[0,T]}\le d_{cc,q,\beta}$, and hence S is continuous with respect to the topology induced by $d_{cc,q,\beta}$, as the larger metric always induces a stronger topology.

Moreover, in Section 2.3.2, we proved that time-augmented paths ensure point separation. Therefore, $\rho\coloneqq S$ is a continuous injective map into a separable infinite-dimensional Hilbert space (Appendix A). It remains to choose a suitable function η such that $\psi_Z(z)\ge\exp(|z|_{l_2}^{\gamma})$, for all $z\in I(\rho(\hat{C}_T^{p,\alpha}))$ and $\gamma>1$. To this end, we observe that

$$|S(\hat{\mathbf{X}})|_{T_1}=\sqrt{\sum_{n=0}^{\infty}|\hat{\mathbf{X}}_T^{(n)}|^2}\le\sum_{n=0}^{\infty}|\hat{\mathbf{X}}_T^{(n)}|\le C_p\sum_{n=0}^{\infty}\frac{|\hat{\mathbf{X}}|_{cc,p,\alpha}^n}{n!}=C_p\exp\big(|\hat{\mathbf{X}}|_{cc,p,\alpha}\big),$$

where the last inequality follows from the estimate $|\hat{\mathbf{X}}_T^{(n)}| \leq C_p |\hat{\mathbf{X}}|^n_{p\text{-var};[0,T]}/n!$ in Theorem 3.7 of [55].[1] Finally, by setting $\eta(t) = \exp(C_p^\gamma \exp(\gamma t))$ and $\hat{\mathbf{X}}_z := \rho^{-1} \circ I^{-1}(z)$, we see that

$$\psi_Z(z) = \psi_X(\hat{\mathbf{X}}_z) = \exp\left(C_p^\gamma \exp(\gamma |\hat{\mathbf{X}}_z|_{cc,p,\alpha})\right) \geq \exp\left(|S(\hat{\mathbf{X}}_z)|_{T_1}^\gamma\right) = \exp(|z|_{l_2}^\gamma).$$

And so, by Theorem 3.6, any Taylor kernel over $\hat{C}_T^{p,\alpha}$, with non-negative coefficients $(a_n)_{n\geq 1}$ and $a_0 > 0$, is globally universal.

Definition 3.7: Consider the weighted space $\hat{C}_T^{p,\alpha}$ as in Example 3.8. We refer to kernels $k : \hat{C}_T^{p,\alpha} \times \hat{C}_T^{p,\alpha} \to \mathbb{R}$ of the form $k(\hat{\mathbf{X}}, \hat{\mathbf{X}}') = \sum_{n\geq 0} a_n \langle S(\hat{\mathbf{X}}, \hat{\mathbf{X}}')\rangle_{T_1}^n$, with $a_n \geq 0$ for all $n \geq 1$ and $a_0 > 0$, as Taylor signature kernels.

[1] This inequality holds for paths of finite p-variation, which is why we consider $\hat{C}_T^{p,\alpha}$ instead of $\hat{C}_T^{\alpha}$.

Conclusion and Future Research

In conclusion, this thesis provided a comprehensive examination of the approximation capabilities of path signatures within rough path spaces. It addressed both the classical universality setting, rigorously demonstrating that linear functionals of the signature approximate continuous functions over compact sets of paths, and the more recent framework of global universality, showing that this approximating capacity extends beyond compact sets to the entire path space. For completeness, the thesis offered a thorough introduction to Rough Path theory, emphasising the interplay between additive and multiplicative functionals, and highlighting relevant topological considerations for signature-based methodologies.

In a subsequent stage, the thesis delved into kernel theory and re-examined the concept of universality from this perspective. This approach is particularly relevant since the signature map can be interpreted as a feature map, thereby defining a universal kernel. As a tangential contribution, this thesis proved that the signature kernel is strictly positive definite, or, equivalently, fully interpolating. The final part of the thesis focused on the setting of weighted spaces, which underpin global universality. In this context, it was demonstrated, under mild assumptions, that global approximations formally encompass local approximations over fixed compact sets, even when the domain's topology is weakened to form a weighted space. Lastly, this thesis projected the concept of global universality into the realm of kernels by defining globally universal kernels and establishing their existence.

Having explored the universality of signatures within rough path spaces in detail, this thesis reaches a fitting conclusion. That said, the completion of this study solely paves the way for further research in related directions. Indeed, various research paths naturally extend and build upon the concepts presented here. For example, the theoretical frameworks established in this thesis underpin many recent signature-based methods in Mathematical Finance. This highlights the inherent complexity of signature-based methodologies, even within highly specific real-world contexts, necessitating a solid theoretical foundation. By providing a comprehensive account of said foundation, the present work now facilitates practical applications, especially within Finance.

In summary, most financial applications that rely on signatures can be encapsulated in the following pipeline:

1. Many quantities of interest — such as payoff functions or trading strategies — can be understood as continuous functionals of some price path. More precisely, given some price path $X : [0, T] \to \mathbb{R}^d$, potentially augmented, we are interested in quantities of the form $\theta(\mathbf{X}|_{[0,t]})$, where θ is a continuous function defined over some (rough) path space

Ω and $\mathbf{X}|_{[0,t]}$ denotes the lift of X restricted to $[0,t]$, signalling the non-anticipative nature of the problem, i.e., we can only price or trade based on past information.

2. Taking advantage of the universality of signatures, we may approximate $\theta(\mathbf{X}|_{[0,t]})$ by linear functionals of the signature, i.e., $\theta(\mathbf{X}|_{[0,t]}) \approx L(S(\mathbf{X}|_{[0,t]}))$. At this step, it is usually essential to assume a fixed compact subset of paths K as the working domain; otherwise, universality, at least in the classical sense, is not applicable.

3. Inevitably, the market contains a random component, which is usually captured by equipping the space of (market) paths Ω with a probability space $(\Omega, \mathcal{B}(\Omega), \mathbb{P})$. With this in mind, the problem at hand is typically formulated as the expected value of our quantity of interest. For instance, if θ is a payoff function, then ultimately we wish to compute the price of some derivative, that is $\mathbb{E}^{\mathbb{Q}}[\theta(\mathbf{X}|_{[0,t]})]$, where $\mathbb{Q}$ denotes the risk-neutral measure. Based on the previous point, we then have $\mathbb{E}^{\mathbb{Q}}[\theta(\mathbf{X}|_{[0,t]})] \approx \mathbb{E}^{\mathbb{Q}}[L(S(\mathbf{X}|_{[0,t]}))]$.

4. By linearity, $\mathbb{E}[L(S(\mathbf{X}|_{[0,t]}))] = L(\mathbb{E}[S(\mathbf{X}|_{[0,t]})])$, which reduces the problem to the computation of the expected signature with respect to an appropriate measure. Here, it is important to note that the linear functional L can be precomputed, as it only depends on θ and not on market conditions. Moreover, if one assumes a model for the price dynamics, then $\mathbb{E}[S(\mathbf{X}|_{[0,t]})]$ may be computed explicitly.

This general philosophy underlies most signature-based methods in Finance and can be adapted to one's needs depending on the specific application — pricing, hedging, portfolio optimisation, and others.

As noted several times in this work, there is a growing interest in replacing the local approximation in Step 2 with a global approximation across the entire path space. Notable contributions in this direction include [41] and [9], both addressing optimal stopping problems. However, these works focus exclusively on approximating continuous and bounded functions of paths. As a result, they cannot utilise standard path signatures and must rely instead on a normalised version of the signature [19]. This normalisation leads to a loss in tractability; for example, the expected signature, which has a known explicit form in many cases, is not available for normalised signatures.

In contrast, the global universality of signatures discussed in this thesis and introduced in [27] enables the approximation of a potentially larger set of functions — the weighted function space (Definition 3.2) — while using standard signatures. This approach preserves the tractability of the expected signature, offering a significant advantage in practical applications and presenting opportunities for future exploration.

To integrate global universality into the pipeline, we need to ensure in Step 1 that Ω is a weighted space. Subsequently, in Step 2, we can drop the assumption of a fixed compact domain K and instead establish density over functions θ in $\mathcal{B}_\psi(\Omega)$. Depending on the application, we can use Theorem 3.4 or Theorem 3.6 to achieve density.

Table 1 below presents a simple numerical exercise designed to stimulate further inquiry and raise questions. Inspired by the work in [9], this table addresses the task of optimally stopping a fractional Brownian motion — a non-Markovian optimal stopping problem. The method involves solving a (discrete) optimal stopping problem through backward induction [62] and approximating the continuation values [49] using kernel ridge regression. This approach allows for the use of the linear signature kernel (Definition 2.7) as well as other signature-based kernels, such as those in Definition 3.7.

As is often the case, obtaining sensible empirical results is easier than rigorously justifying why such results are possible. We emphasise that the numerical results presented below do not constitute a formal experiment; rather, they are intended to instigate future research.

For instance, the main result in [9] (Theorem 2.6) demonstrates that L^p-functionals over the space of stopped rough paths — a rough analogue of Example 3.4 — can be approximated in the L^p-norm by linear functionals of the signature. Functionals over a space of stopped paths are known in the literature of *functional Itô calculus* as *non-anticipative functionals* [31, 21], and they frequently appear in signature-based applications [43, 27]. Notably, [25] endows the space of stopped rough paths with a weighted space structure and shows that linear functionals of the signature are globally universal for (weighted) non-anticipative functionals ([25], Theorem 2.18). This suggests that a density result for L^p-functionals, similar to the one in [9], can be established within the framework of weighted spaces and global universality.

On another front, working with L^p spaces and weighted spaces points towards the integrability of admissible weight functions. Concretely, by considering (Λ, ψ) to be a space of stopped rough paths with a weighted space structure (see, for instance, [25], Lemma 2.17), working with non-anticipative functionals $f \in L^p(\Lambda, \mu)$ inevitably leads to questioning the finiteness of $\int_\Lambda \psi^p \, \mathrm{d}\mu$. In particular, with applications in mind, one may wonder whether it is reasonable to only consider measures μ that guarantee the integrability of ψ^p.

Additionally, this inquiry indicates yet another potential research direction. In [29] the authors prove that, for any weighted space (Ω, ψ_ω), we have the isomorphism $\mathcal{B}_{\psi_\Omega}(\Omega)^* \cong \mathcal{M}_{\psi_\Omega}(\Omega)$, where $\mathcal{M}_{\psi_\Omega}(\Omega)$ denotes the Banach space of signed Radon measures μ fulfilling $\int_\Omega \psi_\Omega(\omega) \, \mathrm{d}\mu(\omega) < \infty$. This, in turn, suggests the formalisation of a weighted/global counterpart of *characteristic kernels* ([27], Remark 5.5).

In the classical setting, assuming X to be a compact metric space, a continuous kernel $k : X \times X \to \mathbb{R}$ is said to be characteristic if the map $\mathbb{P} \mapsto \int_X k(\cdot, x) \, \mathrm{d}\mathbb{P}(x)$ is injective over the set of Borel probability measures — see [20, 69, 70] for further details and several generalisations. Remarkably, it can be shown that a kernel is universal if and only if it is characteristic [69]. There appears to be margin for a similar result specialised to globally universal kernels. More interestingly, this discussion may lead to methods that leverage kernel scoring rules [72], which are essentially a way to assess the quality of a probability forecast.

We conclude our exposition with two final proposals. First, while substantial work has been done regarding the signature kernel [14, 44, 46, 67], there remains significant scope for a comprehensive study of its RKHS ([27], Section 6). Second, given the expanding interest in rough paths and signature-based methods, it is pertinent to formalise their mathematics within a proof assistant, such as *Lean*. Although this proposal may seem unorthodox, the importance of *computer-assisted proofs* is undeniable, especially when sophisticated mathematics is being used in applications.

H	Linear	Polynomial	[9], J=100	[9], J=500	[10]
0.9	0.333	0.338	0.331	0.337	0.335
0.8	0.274	0.274	0.275	0.281	0.276
0.7	0.201	0.203	0.203	0.205	0.206
0.6	0.112	0.113	0.112	0.112	0.115
0.5	0.001	-0.001	-0.001	-0.002	0
0.4	0.153	0.151	0.153	0.155	0.154
0.3	0.362	0.363	0.363	0.371	0.368
0.2	0.651	0.649	0.654	0.662	0.657
0.1	1.044	1.041	1.045	1.065	1.048

Table 1: Optimal stopping of fractional Brownian motion.

Table 1 contains the estimated lower bounds for the optimal stopping values y_0^H of a fractional Brownian motion with varying Hurst parameter H. Specifically, it addresses the task of solving the optimal stopping problem $y_0^H = \sup_{\tau\in\mathcal{S}_0} \mathbb{E}[X_\tau^H]$, where $(X_t^H)_{t\in[0,T]}$ denotes the fractional Brownian motion with Hurst parameter $H \in (0,1)$ and $\mathcal{S}_0$ the set of $(\mathcal{F}_t^{\hat{\mathbf{X}}^H})$-adapted stopping times.

As mentioned above, to estimate y_0^H we first discretise the problem and then apply the *Longstaff-Schwartz algorithm*, using kernel ridge regression to approximate the continuation values, similar to the approach in [42]. This allows us to choose from different kernels, contrasting with [9], where solely the standard signature kernel is used. Particularly, for Table 1 we use the (linear) signature kernel k_{sig} (Definition 2.7) and the polynomial signature kernel, corresponding to the composition $f \circ k_{sig}$, where $f(t) = (b + ct)^d$, $b, c, d > 0$. We observe that the exponential signature kernel, with $f(t) = e^t$, also produced reasonable results, but frequently led to ill-conditioned matrices during the kernel ridge regression step, especially for more irregular paths, i.e., when $H < 0.5$.

We compare our results with [9], which approximates the continuation values with signature-based linear regression, and with [10], which approximates the optimal stopping decisions using neural networks. In our experiments, we truncate the signature at level 5, discretise the interval $[0, 1]$ with $J = 100$ grid points, and consider a ridge parameter

$\alpha \in \{0.1, 0.01, 0.001, 0.0001, 0.00001\}$. We note that the influence of the ridge parameter α is more pronounced for the lowest values of H, namely $H = 0.2$ and $H = 0.1$. For higher values of H, different choices of α produce similar results. For further details, we refer to [9], as this experiment is largely based on the methodology presented therein. We also provide the code below for interested readers.[2] This experiment was performed for the sake of curiosity and Table 1 solely serves as an addendum to the conclusion of this work.

[2]Table code

References

[1] Eduardo Abi Jaber and Louis-Amand Gérard. "Signature volatility models: pricing and hedging with Fourier". *Available at SSRN* (2024).

[2] Andrew L Allan. "Rough Path Theory" (2021).

[3] Andrew L Allan, Christa Cuchiero, Chong Liu, and David J Prömel. "Model-free portfolio theory: A rough path approach". *Mathematical Finance* 33.3 (2023), 709–765.

[4] Andrew L Allan, Chong Liu, and David J Prömel. "A càdlàg rough path foundation for robust finance". *Finance and Stochastics* 28.1 (2024), 215–257.

[5] Imanol Perez Arribas. "Derivatives pricing using signature payoffs". *arXiv preprint arXiv:1809.09466* (2018).

[6] Imanol Perez Arribas, Cristopher Salvi, and Lukasz Szpruch. "Sig-SDEs model for quantitative finance". *Proceedings of the First ACM International Conference on AI in Finance.* 2020, 1–8.

[7] P Bank, C Bayer, PP Hager, S Riedel, and T Nauen. "Stochastic Control with Signatures". *arXiv preprint arXiv:2406.01585* (2024).

[8] Christian Bayer, Paul P Hager, Sebastian Riedel, and John Schoenmakers. "Optimal stopping with signatures". *The Annals of Applied Probability* 33.1 (2023), 238–273.

[9] Christian Bayer, Luca Pelizzari, and John Schoenmakers. "Primal and dual optimal stopping with signatures". *arXiv preprint arXiv:2312.03444* (2023).

[10] Sebastian Becker, Patrick Cheridito, and Arnulf Jentzen. "Deep optimal stopping". *Journal of Machine Learning Research* 20.74 (2019), 1–25.

[11] Horatio Boedihardjo, Xi Geng, Terry Lyons, and Danyu Yang. "The signature of a rough path: uniqueness". *Advances in Mathematics* 293 (2016), 720–737.

[12] Haim Brezis and Haim Brézis. *Functional analysis, Sobolev spaces and partial differential equations.* Vol. 2. 3. Springer, 2011.

[13] Álvaro Cartea, Imanol Perez Arribas, and Leandro Sánchez-Betancourt. "Optimal execution of foreign securities: A double-execution problem". *Available at SSRN 3562251* (2020).

[14] Thomas Cass, Terry Lyons, and Xingcheng Xu. "General signature kernels". *arXiv preprint arXiv:2107.00447* (2021).

[15] Thomas Cass and Jeffrey Pei. "A Fubini type theorem for rough integration." *Revista Mathematica Iberoamericana* 39.2 (2023).

[16] Thomas Cass and Cristopher Salvi. "Lecture notes on rough paths and applications to machine learning". *arXiv preprint arXiv:2404.06583* (2024).

[17] Kuo-Tsai Chen. "Integration of paths–A faithful representation of paths by noncommutative formal power series". *Transactions of the American Mathematical Society* 89.2 (1958), 395–407.

[18] Ilya Chevyrev and Andrey Kormilitzin. "A primer on the signature method in machine learning". *arXiv preprint arXiv:1603.03788* (2016).

[19] Ilya Chevyrev and Harald Oberhauser. "Signature moments to characterize laws of stochastic processes". *arXiv preprint arXiv:1810.10971* (2018).

[20] Andreas Christmann and Ingo Steinwart. "Universal kernels on non-standard input spaces". *Advances in neural information processing systems* 23 (2010).

[21] Rama Cont and David-Antoine Fournié. "Functional Itô calculus and stochastic integral representation of martingales" (2013).

[22] Laure Coutin. "Rough paths via sewing lemma". *ESAIM: Probability and Statistics* 16 (2012), 479–526.

[23] Christa Cuchiero, Guido Gazzani, Janka Möller, and Sara Svaluto-Ferro. "Joint calibration to SPX and VIX options with signature-based models". *arXiv preprint arXiv:2301.-13235* (2023).

[24] Christa Cuchiero, Guido Gazzani, and Sara Svaluto-Ferro. "Signature-based models: Theory and calibration". *SIAM journal on financial mathematics* 14.3 (2023), 910–957.

[25] Christa Cuchiero and Janka Möller. "Signature Methods in Stochastic Portfolio Theory". *arXiv preprint arXiv:2310.02322* (2023).

[26] Christa Cuchiero, Francesca Primavera, and Sara Svaluto-Ferro. "Universal approximation theorems for continuous functions of càdlàg paths and Lévy-type signature models". *arXiv preprint arXiv:2208.02293* (2022).

[27] Christa Cuchiero, Philipp Schmocker, and Josef Teichmann. "Global universal approximation of functional input maps on weighted spaces". *arXiv preprint arXiv:2306.03303* (2023).

[28] Christa Cuchiero, Sara Svaluto-Ferro, and Josef Teichmann. "Signature SDEs from an affine and polynomial perspective". *arXiv preprint arXiv:2302.01362* (2023).

[29] Philipp Dörsek and Josef Teichmann. "A semigroup point of view on splitting schemes for stochastic (partial) differential equations". *arXiv preprint arXiv:1011.2651* (2010).

[30] David Steven Dummit and Richard M Foote. *Abstract algebra.* Vol. 3. Wiley Hoboken, 2004.

[31] Bruno Dupire. "Functional Itô Calculus". *Quantitative Finance* 19.5 (2019), 721–729.

[32] Adeline Fermanian. “Embedding and learning with signatures”. *Computational Statistics & Data Analysis* 157 (2021), 107148.

[33] E Robert Fernholz and E Robert Fernholz. *Stochastic portfolio theory*. Springer, 2002.

[34] Robert Fernholz and Ioannis Karatzas. “Stochastic portfolio theory: an overview”. *Handbook of numerical analysis* 15.89-167 (2009), 1180–91267.

[35] Denis Feyel and Arnaud de La Pradelle. “Curvilinear integrals along enriched paths” (2006).

[36] Peter K Friz and Martin Hairer. *A course on rough paths*. Springer, 2020.

[37] Peter K Friz and Nicolas B Victoir. *Multidimensional stochastic processes as rough paths: theory and applications*. Vol. 120. Cambridge University Press, 2010.

[38] Owen Futter, Blanka Horvath, and Magnus Wiese. “Signature trading: A path-dependent extension of the mean-variance framework with exogenous signals”. *arXiv preprint arXiv:2308.15135* (2023).

[39] Ben Hambly and Terry Lyons. “Uniqueness for the signature of a path of bounded variation and the reduced path group”. *Annals of Mathematics* (2010), 109–167.

[40] Keisuke Hara and Masanori Hino. “Fractional order Taylor’s series and the neo-classical inequality”. *Bulletin of the London Mathematical Society* 42.3 (2010), 467–477.

[41] Blanka Horvath, Maud Lemercier, Chong Liu, Terry Lyons, and Cristopher Salvi. “Optimal stopping via distribution regression: a higher rank signature approach”. *arXiv preprint arXiv:2304.01479* (2023).

[42] Wenbin Hu and Tomasz Zastawniak. “Pricing high-dimensional American options by kernel ridge regression”. *Quantitative Finance* 20.5 (2020), 851–865.

[43] Jasdeep Kalsi, Terry Lyons, and Imanol Perez Arribas. “Optimal execution with rough path signatures”. *SIAM Journal on Financial Mathematics* 11.2 (2020), 470–493.

[44] Franz J Király and Harald Oberhauser. “Kernels for sequentially ordered data”. *Journal of Machine Learning Research* 20.31 (2019), 1–45.

[45] H.H. Kuo. *Introduction to Stochastic Integration*. Universitext. Springer New York, 2006.

[46] Darrick Lee and Harald Oberhauser. “The Signature Kernel”. *arXiv preprint arXiv:2305.-04625* (2023).

[47] J.M. Lee. *Introduction to Smooth Manifolds*. Graduate Texts in Mathematics. Springer, 2012.

[48] Daniel Levin, Terry Lyons, and Hao Ni. “Learning from the past, predicting the statistics for the future, learning an evolving system”. *arXiv preprint arXiv:1309.0260* (2013).

[49] Francis A Longstaff and Eduardo S Schwartz. “Valuing American options by simulation: a simple least-squares approach”. *The review of financial studies* 14.1 (2001), 113–147.

[50] Terry Lyons. “Rough paths, signatures and the modelling of functions on streams”. *arXiv preprint arXiv:1405.4537* (2014).

[51] Terry Lyons, Sina Nejad, and Imanol Perez Arribas. “Numerical method for model-free pricing of exotic derivatives using rough path signatures”. *arXiv preprint arXiv:1905.01720* (2019).

[52] Terry Lyons, Sina Nejad, and Imanol Perez Arribas. “Non-parametric pricing and hedging of exotic derivatives”. *Applied Mathematical Finance* 27.6 (2020), 457–494.

[53] Terry Lyons and Zhongmin Qian. *System control and rough paths*. Oxford University Press, 2002.

[54] Terry J Lyons. “Differential equations driven by rough signals”. *Revista Matemática Iberoamericana* 14.2 (1998), 215–310.

[55] Terry J Lyons, Michael Caruana, and Thierry Lévy. *Differential equations driven by rough paths*. Springer, 2007.

[56] Péter Medvegyev. *Stochastic integration theory*. Vol. 14. OUP Oxford, 2007.

[57] Charles A Micchelli, Yuesheng Xu, and Haizhang Zhang. “Universal Kernels.” *Journal of Machine Learning Research* 7.12 (2006).

[58] James R Munkres. *Topology: Pearson New International Edition*. Pearson Higher Ed., 2013.

[59] Jan van Neerven. *Functional analysis*. Vol. 201. Cambridge University Press, 2022.

[60] Ivan Nourdin et al. *Selected aspects of fractional Brownian motion*. Vol. 4. Springer, 2012.

[61] Vern I Paulsen and Mrinal Raghupathi. *An introduction to the theory of reproducing kernel Hilbert spaces*. Vol. 152. Cambridge university press, 2016.

[62] Goran Peskir and Albert Shiryaev. *Optimal stopping and free-boundary problems*. Springer, 2006.

[63] Philip E Protter and Philip E Protter. *Stochastic differential equations*. Springer, 2005.

[64] Walter Rudin. “Functional analysis 2nd ed”. *International Series in Pure and Applied Mathematics. McGraw-Hill, Inc., New York* (1991).

[65] Walter Rudin et al. *Principles of mathematical analysis*. Vol. 3. McGraw-hill New York, 1964.

[66] Saburou Saitoh, Yoshihiro Sawano, et al. *Theory of reproducing kernels and applications*. Springer, 2016.

[67] 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 (2021), 873–899.

[68] René L Schilling and Lothar Partzsch. *Brownian motion: an introduction to stochastic processes*. Walter de Gruyter GmbH & Co KG, 2014.

[69] Carl-Johann Simon-Gabriel and Bernhard Schölkopf. "Kernel distribution embeddings: Universal kernels, characteristic kernels and kernel metrics on distributions". *Journal of Machine Learning Research* 19.44 (2018), 1–29.

[70] Bharath K Sriperumbudur, Kenji Fukumizu, and Gert RG Lanckriet. "Universality, Characteristic Kernels and RKHS Embedding of Measures." *Journal of Machine Learning Research* 12.7 (2011).

[71] Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

[72] Ingo Steinwart and Johanna F Ziegel. "Strictly proper kernel scores and characteristic kernels on compact spaces". *Applied and Computational Harmonic Analysis* 51 (2021), 510–542.

[73] Marshall H Stone. "The generalized Weierstrass approximation theorem". *Mathematics Magazine* 21.5 (1948), 237–254.

[74] Mazyar Ghani Varzaneh and Sebastian Riedel. "Introduction to rough paths theory". *arXiv preprint arXiv:2302.04653* (2023).

[75] Laurence C Young. "An inequality of the Hölder type, connected with Stieltjes integration" (1936).

Appendix A

We briefly recall the definition of the tensor product between two vector spaces and outline some basic properties, including its characteristic property. Additionally, we clarify the Hilbert space structure on the tensor algebra. This appendix provides an overview of the tensor product between vector spaces. Readers who are completely unfamiliar with tensor products are referred to Chapter 12 in [47] or Chapter 10 in [30] for more detailed explanations.

A.I. Tensor products and the characteristic property. Let V and W denote finite-dimensional real vector spaces. The tensor product $V \otimes W$ is typically defined in a constructive way as the quotient space of the *free vector space* on $V \times W$ that makes the equivalence class of (v, w) depend bilinearly on v and w. Specifically, let $\mathbb{R}\langle V \times W\rangle$ denote the set of all formal linear combinations of elements in $V \times W$, i.e.,

$$\mathbb{R}\langle V \times W\rangle \equiv \{f : V \times W \to \mathbb{R} : f(v, w) = 0 \text{ for all but finitely many } (v, w) \in V \times W\}.$$

Under pointwise addition and scalar multiplication, $\mathbb{R}\langle V \times W\rangle$ becomes a real vector space, known as the free vector space on $V \times W$. Observe that $\mathbb{R}\langle V \times W\rangle$ corresponds to an infinite-dimensional vector space with basis given by the set $\{(v, w) : v \in V \text{ and } w \in W\}$. Subsequently, consider the subspace $\mathcal{R} \subset \mathbb{R}\langle V \times W\rangle$ spanned by all elements of the following forms:

$$\begin{aligned} \alpha(v, w) - (\alpha v, w),\\ \alpha(v, w) - (v, \alpha w),\\ (v, w) + (v', w) - (v + v', w),\\ (v, w) + (v, w') - (v, w + w'), \end{aligned}$$

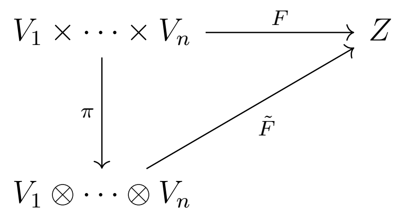
with $\alpha \in \mathbb{R}$, $v, v' \in V$ and $w, w' \in W$. Then, the *tensor product* of V and W is defined as the quotient space $V \otimes W := \mathbb{R}\langle V \times W\rangle/\mathcal{R}$. Furthermore, we denote by $v \otimes w$ the equivalence class associated with (v, w) and observe that by definition the *elementary tensors* $v \otimes w$ satisfy

$$\begin{aligned} \alpha(v \otimes w) = \alpha v \otimes w = v \otimes \alpha w,\\ v \otimes w + v' \otimes w = (v + v') \otimes w,\\ v \otimes w + v \otimes w' = v \otimes (w + w'). \end{aligned}$$

To better understand the tensor product, it is beneficial to examine its characteristic property. Considering a third vector space Z, the characteristic property of the tensor product establishes a one-to-one correspondence between bilinear maps $V \times W \to Z$ and linear maps $V \otimes W \to Z$. In other words, any bilinear map $V \times W \to Z$ factors uniquely through $V \times W \to V \otimes W$.

The same holds true for multilinear maps, for which we provide the precise statement below. Moreover, we note that the characteristic property of the tensor product is *universal* in the categorical sense, i.e., any space satisfying the characteristic property of the tensor product is isomorphic to the tensor product itself ([47], Exercise 12.3).

Theorem A.1 ([47], Proposition 12.7)**:** Let $V_1, \dots, V_n$ be finite-dimensional real vector spaces. Additionally, let $\pi : V_1 \times \cdots \times V_n \to V_1 \otimes \cdots \otimes V_n$ denote the quotient map $(v_1, \dots, v_n) \mapsto v_1 \otimes \cdots \otimes v_n$. For any vector space Z, if $F : V_1 \times \cdots \times V_n \to Z$ is a multilinear map, then there exists a unique linear map $\tilde{F} : V_1 \otimes \cdots \otimes V_n \to Z$ such that the following diagram commutes:



We compile some useful properties of the tensor product that may provide further clarification to Section 1.1.3. These correspond to Propositions 12.8, 12.9 and 12.10 in [47].

Theorem A.2: Let $V_1, \dots, V_n$ be real vector spaces of dimensions $d_1, \dots, d_n$, respectively. Then, the following properties hold:

1. For each $i \in \{1, \dots, n\}$, let $\{e_1^{(i)}, \dots, e_{d_i}^{(i)}\}$ be a basis for V_i. Then, the set
$$\left\{e_{i_1}^{(1)} \otimes \cdots \otimes e_{i_n}^{(n)} : 1 \leq i_1 \leq d_1, \dots, 1 \leq i_n \leq d_n\right\}$$
forms a basis for $V_1 \otimes \cdots \otimes V_n$. Moreover, it becomes apparent that $V_1 \otimes \cdots \otimes V_n$ has dimension $d_1 \cdots d_n$.

2. The tensor product construction is associative, i.e., there are unique isomorphisms
$$V_1 \otimes (V_2 \otimes V_3) \cong V_1 \otimes V_2 \otimes V_3 \cong (V_1 \otimes V_2) \otimes V_3,$$
under which elements of the form $v_1 \otimes (v_2 \otimes v_3)$, $v_1 \otimes v_2 \otimes v_3$ and $(v_1 \otimes v_2) \otimes v_3$ are all identified. The same holds true for n-fold tensor products.

3. Let $L(V_1, \dots, V_n; \mathbb{R})$ denote the set of multilinear maps $V_1 \times \cdots \times V_n \to \mathbb{R}$. There exists a canonical isomorphism $V_1^* \otimes \cdots \otimes V_n^* \cong L(V_1, \dots, V_n; \mathbb{R})$.

As per the notation in Section 1.1.3, Theorem A.2 clarifies the canonical isomorphism $(\mathbb{R}^d)^{\otimes k} \otimes (\mathbb{R}^d)^{\otimes (n-k)} \cong (\mathbb{R}^d)^{\otimes n}$, which maps $(e_{i_1} \otimes \cdots \otimes e_{i_k}) \otimes (e_{i_{k+1}} \otimes \cdots \otimes e_{i_n})$ to $e_{i_1} \otimes \cdots \otimes e_{i_n}$, where $\{e_1, \dots, e_d\}$ denotes the canonical basis of $\mathbb{R}^d$. Additionally, Theorem A.2 justifies why we write a generic element $\mathbf{g} \in (\mathbb{R}^d)^{\otimes n}$ as $\sum_{i_1,\dots,i_n} g^{i_1,\dots,i_n} e_{i_1} \otimes \cdots \otimes e_{i_n}$.

A.II. Hilbert space structure on the extended tensor algebra. The construction of the tensor product as a quotient space is not restricted to finite-dimensional vector spaces. We may consider arbitrary vector spaces, or, more generally, modules over a ring. With this in mind, let $\mathcal{H}_1, \ldots, \mathcal{H}_N$ denote arbitrary Hilbert spaces and consider their tensor product $\mathcal{H}_1 \otimes \cdots \otimes \mathcal{H}_N$. The *Hilbert space tensor product of* $\mathcal{H}_1, \ldots, \mathcal{H}_N$ is defined as the completion of the (algebraic) tensor product $\mathcal{H}_1 \otimes \cdots \otimes \mathcal{H}_N$ with respect to the norm induced by the inner product

$$\Big\langle \sum_{i=1}^{k} h_1^{(i)} \otimes \cdots \otimes h_N^{(i)}, \sum_{j=1}^{l} \tilde{h}_1^{(j)} \otimes \cdots \otimes \tilde{h}_N^{(j)} \Big\rangle \coloneqq \sum_{i=1}^{k} \sum_{j=1}^{l} \prod_{n=1}^{N} \langle h_n^{(i)}, h_n^{(j)} \rangle_{\mathcal{H}_n}.$$

It is common to use a slight abuse of notation by denoting the Hilbert space tensor product as $\mathcal{H}_1 \otimes \cdots \otimes \mathcal{H}_N$ again. If the spaces $\mathcal{H}_1, \ldots, \mathcal{H}_N$ are separable with orthonormal basis $(h_j^{(n)})_{j\geq 1}$ for $n = 1, \ldots, N$, then the tensors $h_{i_1}^{(1)} \otimes \cdots \otimes h_{i_N}^{(N)}$ form an orthonormal basis for $\mathcal{H}_1 \otimes \cdots \otimes \mathcal{H}_N$ ([59], Section 14.4), rendering it separable as well.

As a specific example, we may consider $\mathcal{H}_1 = \cdots = \mathcal{H}_N = \mathbb{R}^d$, which yields $(\mathbb{R}^d)^{\otimes N}$. This is the only Hilbert space tensor product used in the present work. In this case, the inner product on elementary tensors is given by

$$\big\langle a_1 \otimes \cdots \otimes a_N, b_1 \otimes \cdots \otimes b_N \big\rangle_{(\mathbb{R}^d)^{\otimes N}} = \prod_{n=1}^{N} \langle a_n, b_n \rangle_{\mathbb{R}^d},$$

and extended by linearity to $(\mathbb{R}^d)^{\otimes N}$. In particular, using the basis representation of $(\mathbb{R}^d)^{\otimes N}$, we observe that

$$\begin{aligned}
&\Big\langle \sum_{i_1,\ldots,i_N} a^{i_1,\ldots,i_N} e_{i_1} \otimes \cdots \otimes e_{i_N}, \sum_{j_1,\ldots,j_N} b^{j_1,\ldots,j_N} e_{j_1} \otimes \cdots \otimes e_{j_N} \Big\rangle_{(\mathbb{R}^d)^{\otimes N}} \\
&= \sum_{i_1,\ldots,i_N} \sum_{j_1,\ldots,j_N} a^{i_1,\ldots,i_N} b^{j_1,\ldots,j_N} \prod_{n=1}^{N} \langle e_{i_n}, e_{j_n} \rangle_{\mathbb{R}^d} \; = \sum_{i_1,\ldots,i_N} a^{i_1,\ldots,i_N} b^{i_1,\ldots,i_N},
\end{aligned}$$

where we used the fact that $\prod_{n=1}^{N} \langle e_{i_n}, e_{j_n} \rangle_{\mathbb{R}^d} \neq 0$ if and only if $i_n = j_n$ for all $n \in \{1, \ldots, N\}$. Put differently, the inner product on $(\mathbb{R}^d)^{\otimes N}$ is equivalent to summing over the set of words of length N. Lastly, by considering the elements $\mathbf{a} = (\mathbf{a}^0, \mathbf{a}^1, \ldots)$ of the extended tensor algebra $T((\mathbb{R}^d))$ with finite Euclidean norm, i.e., the set

$$\tilde{T}((\mathbb{R}^d)) \coloneqq \left\{ \mathbf{a} \in T((\mathbb{R}^d)) : |\mathbf{a}|_{\tilde{T}((\mathbb{R}^d))} = \left(\sum_{n=0}^{\infty} |\mathbf{a}^n|^2 \right)^{1/2} < \infty \right\},$$

we define an inner product therein by setting $\langle \mathbf{a}, \mathbf{b} \rangle_{\tilde{T}((\mathbb{R}^d))} \coloneqq \sum_{n\geq 0} \langle \mathbf{a}^n, \mathbf{b}^n \rangle_{(\mathbb{R}^d)^{\otimes n}}$. Using the word notation, this inner product can be perceived as a summation over all words, that is

$$\langle \mathbf{a}, \mathbf{b} \rangle_{\tilde{T}((\mathbb{R}^d))} = \sum_{n=0}^{\infty} \sum_{|w|=n} a_w b_w.$$

Appendix B

This appendix collects some basic results from Functional Analysis used throughout the thesis. For a thorough exposition of these facts, refer to [59].

B.I. Hilbert spaces. With the additional structure provided by an inner product, Hilbert spaces acquire two essential aspects when compared to Banach spaces: a geometric structure derived from the orthogonal complementation of closed subspaces and the property of self-duality.

Theorem B.1 (Completion) ([59], Theorem 1.5, Proposition 3.9)**:** Let X be a normed space. Then,

1. There exists a Banach space $\overline{X}$ that contains X isometrically as a dense subspace. Moreover, the space $\overline{X}$ is unique up to isometry.
2. If X is equipped with an inner product, then its completion $\overline{X}$ as a normed space has a well-defined inner product given by

$$\langle x, x' \rangle_{\overline{X}} := \lim_{n\to\infty} \langle x_n, x'_n \rangle_X, \quad x, x' \in \overline{X},$$

whenever $x_n, x'_n \in X$ satisfy $x_n \to x$ and $x'_n \to x'$. Additionally, the norm induced by $\langle \cdot, \cdot \rangle_{\overline{X}}$ coincides with the norm of $\overline{X}$ obtained by completion.

In a Hilbert space $\mathcal{H}$, two elements $h, h' \in \mathcal{H}$ are said to be orthogonal if $\langle h, h' \rangle_{\mathcal{H}} = 0$. Furthermore, given some subset $A \subset \mathcal{H}$, its orthogonal complement is defined by the set

$$A^{\perp} := \{h \in \mathcal{H} : \langle h, a \rangle_{\mathcal{H}} = 0 \text{ for all } a \in A\}.$$

Remarkably, a Hilbert space can always be decomposed into the direct sum of a fixed linear subspace and its orthogonal complement.

Theorem B.2 (Orthogonal complement) ([59], Theorem 3.13)**:** Let $\mathcal{H}$ be a Hilbert space. If Y is closed (linear) subspace of $\mathcal{H}$, then we have an orthogonal direct sum decomposition, i.e.,

$$\mathcal{H} = Y \oplus Y^{\perp}.$$

More precisely, this implies that $Y \cap Y^{\perp} = \{0\}$ and for every $h \in \mathcal{H}$ there exist $y \in Y$ and $y^{\perp} \in Y^{\perp}$ such that $h = y + y^{\perp}$.

With this decomposition, we may refer to orthogonal projections onto closed subspaces unambiguously. Specifically, the projection π_Y onto Y along $Y^\perp$ is defined by $\pi_Y(y+y^\perp) := y$.

One of the most notable properties of Hilbert spaces is that they are isomorphic to their own duals. In this sense, Hilbert spaces are sometimes referred to as *self-dual*.

Theorem B.3 (Riesz Representation theorem) ([59], Theorem 3.15)**:** If $\phi : \mathcal{H} \to \mathbb{R}$ is a bounded linear functional, there exists a unique element $h_\phi \in \mathcal{H}$ such that $\phi(h) = \langle h, h_\phi \rangle_{\mathcal{H}}$ for all $h \in \mathcal{H}$. The element h_ϕ is said to be the Riesz representer of ϕ. In addition, the map $\phi \mapsto h_\phi$ that associates each linear bounded functional with its Riesz representer is an isometric isomorphism.

A final relevant property specific to *separable* Hilbert spaces is that they are all isometrically isomorphic to $l_2(\mathbb{N})$.

Theorem B.4 ([59], Corollary 3.23)**:** Any two infinite-dimensional separable Hilbert spaces are isometrically isomorphic.

B.II. Duality and bounded operators. At the core of Functional Analysis is the concept of duality. The dual of a Banach space X is defined as the set of all bounded linear functionals from X to $\mathbb{R}$, and denoted by X^*. Furthermore, recall that for any two vector spaces V and W, and any bilinear mapping $f : V \times W \to \mathbb{R}$, the *weak topology of V induced by W* is the smallest topology on V such that the linear mapping $v \mapsto f(v, w)$ is continuous for all $w \in W$. In particular, if $V = X^*$ is set to be the dual of $W = X$, then the weak topology induced by X is referred to as the *weak-$*$-topology*.

Theorem B.5 (Banach-Alaoglu theorem) ([59], Theorem 4.51)**:** The closed unit ball of every dual Banach space is compact with respect to the weak-$*$-topology.

We conclude this appendix by stating the *Uniform boundedness theorem*, also known as the Banach-Steinhaus theorem. We showcase how this theorem may be applied by proving Theorem II.

Theorem B.6 ([59], Theorem 5.2)**:** Consider a Banach space V and a normed space W. Let $(T_i)_{i\in I}$ be an arbitrary family of bounded linear operators from V to W. If for every $v \in V$,

$$\sup_{i\in I} |T_i(v)|_W < \infty,$$

then

$$\sup_{i\in I} |T_i|_{op} \equiv \sup_{i\in I} \sup_{|v|_V \le 1} |T_i(v)|_W < \infty.$$

In other words, pointwise boundedness for all operators T_i implies uniform boundedness with respect to the usual operator norm. Although Banach spaces are typically denoted by the letter X, in the theorem above we use V to reserve X for paths, as in Theorem II.

proof of Theorem II. To begin with, observe that it suffices to consider $\mathbb{R}$-valued paths, since the integral of Y against X is defined component-wise.

Taking this into account, let V be the Banach space of continuous functions $C([0,T],\mathbb{R})$ equipped with the supremum norm, and let W be $\mathbb{R}$ equipped with the usual absolute value norm. Consider a sequence of partitions $P_n = \{0 < t_0^n < t_1^n < \cdots < t_{N_n}^n = T\}$ with vanishing mesh size and, for $Y \in V$, define the operators

$$T_n(Y) \coloneqq \sum_{i=0}^{N_n-1} Y_{t_i^n}\big(X_{t_{i+1}^n} - X_{t_i^n}\big),$$

where X is some $\mathbb{R}$-valued path. For a fixed n, note that we may construct a map $Y \in V$ such that $Y_{t_i^n} = \operatorname{sign}\big\{X_{t_{i+1}^n} - X_{t_i^n}\big\}$ and $|Y|_{\infty;[0,T]} = 1$. For such a path Y we have that

$$T_n(Y) = \sum_{i=0}^{N_n-1} |X_{t_{i+1}^n} - X_{t_i^n}|,$$

and hence,

$$|T_n|_{op} \geq \sum_{i=0}^{N_n-1} |X_{t_{i+1}^n} - X_{t_i^n}|,$$

for all $n \in \mathbb{N}$. Consequently, $\sup_{n\in\mathbb{N}} |T_n|_{op} \geq |X|_{1\text{-var};[0,T]}$. On the other hand, by assumption we know that $\lim_{n\to\infty} T_n(Y)$ exists. Therefore, $\sup_n |T_n(Y)| < \infty$ and, by Theorem B.6, we conclude that $\sup_n |T_n|_{op}$ is finite, implying $|X|_{1\text{-var};[0,T]} < \infty$.

Appendix C

This appendix compiles some relevant results from Topology used in this work.

C.I. Basic topological results. We recall certain topological properties for the reader's convenience. Most of these results pertain to compact spaces and can be found in any topology textbook. We refer to [58] for more details.

Theorem C.1 (Basic properties of compact sets) ([58], Theorems 26.2, 26.3 and 26.5)**:**

1. Every closed subset of a compact space is compact.
2. Every compact subset of a Hausdorff space is closed.
3. The image of a compact set under a continuous map is compact.

Theorem C.2 (Tychonoff's theorem) ([58], Theorem 37.3)**:** An arbitrary product of compact spaces is compact in the product topology.

Theorem C.3 ([58], Theorem 26.6)**:** Let X be a compact space and Y Hausdorff. If $f : X \to Y$ is a continuous bijection, then f is a homeomorphism. More precisely, f^{-1} is a continuous bijection as well.

Recall that a space X is said to be *sequentially compact* if every sequence of points in X has a convergent subsequence. Under the assumption that X is metrizable, all notions of compactness coincide.

Theorem C.4 ([58], Theorem 28.2)**:** If X is a metrizable space, then X is compact if and only if X is sequentially compact.

Theorem C.5 (Heine-Borel theorem) ([58], Theorem 27.3)**:** A subset A of $\mathbb{R}^d$ is compact if and only if A is closed and bounded.

We also use the following two well-established results related to the separation axioms.

Theorem C.6 ([58], Theorem 34.3)**:** A space X is completely regular if and only if it is homeomorphic to a subset of $[0,1]^J$ for some indexing set J.

Theorem C.7 (Tietze Extension theorem) ([58], Theorem 35.1)**:** Consider a normal space X and let A be a closed subset of X. Then, any continuous function from A to $\mathbb{R}$ can be extended to a continuous map from X to $\mathbb{R}$.

C.II. Function spaces. We list some classical results pertaining to spaces of functions, starting with the Stone-Weierstrass theorem, which underpins much of the discussion in the second half of the thesis.

Theorem C.8 (Stone-Weierstrass theorem) ([65], Theorem 7.32)**:** Let X denote a compact Hausdorff space and assume that $\mathcal{A} \subset C(X)$ is a subalgebra. Then, $\mathcal{A}$ is dense in $C(X)$ if and only if $\mathcal{A}$ is point-separating and vanishes nowhere.

Theorem C.9 (Weierstrass theorem) ([65], Theorem 7.26)**:** Consider $f \in C([a,b],\mathbb{R})$. For every $\varepsilon > 0$ there exists a polynomial p such that $|f - p|_{\infty;[0,T]} \leq \varepsilon$.

Finally, the following result is used to guarantee that linear functionals of the signature are point-separating over the space of time-augmented paths. In what follows, the space $C_c^\infty(\Omega)$ denotes the set of all infinitely differentiable functions $\varphi : \Omega \to \mathbb{R}$ with compact support contained in Ω.

Theorem C.10 ([12], Corollary 4.24)**:** Let $\Omega \subset \mathbb{R}^d$ be an open set and consider $f \in L^1_{\mathrm{loc}}(\Omega)$, i.e., f is integrable on all compact subsets of Ω. If

$$\int_\Omega f\varphi \,\mathrm{d}\mu, \quad \text{for all } \varphi \in C_c^\infty(\Omega),$$

then $f = 0$ almost everywhere on Ω.