Imperial College
London

Imperial College London

Department of Mathematics

Truncated order decision for Signature least square regression model under the point view of empirical process theory

Author: Chenhao Jin (CID: 01788758)

A thesis submitted for the degree of

MSc in Mathematics and Finance, 2019-2020

Abstract

Feature extraction for time sequential data is not a new topic in machine learning. The Signature transform which is derived from the theory of controlled differential equations is one of the powerful methods in this domain. In practice, we usually use the truncated Signature to describe the pattern of the time sequential data. Large truncated order can provide a more detailed description, but it also brings the risk of over-fitting. Thus, a theoretical explanation for the selection of truncated order can be interesting.

Empirical processes theory comes from the need for generalizations for Glivenko-Cantelli theorem and Donsker theorem. Many problems in statistics and machine learning can be regarded as an empirical risk minimization problem. Study the asymptotic performance of the estimator is a commonly recurring theme in statistics. Some theorems in this domain provide us a powerful way to study the rate of convergence for the empirical risk minimizer.

In this thesis, we would like to learn the solution of some controlled differential equations by the least square regression method. The main result of this thesis provides a theoretical formula to describe the selection of truncated order as a function of the observation size using empirical process theory. **To the best knowledge of the author, this theoretical formula is first proposed in this thesis. The whole theoretical proof is the completely original work of the author**

Keywords: Machine Learning, Signature Transform, Truncated Order, Empirical Process, Rate of Convergence

Acknowledgements

I would first like to thank my supervisor, Dr. Thomas Cass, for introducing me to this interesting and challenging topic. Dr. Cass provided continued support and guidance during the running of this thesis. He consistently allowed this project to be my own work, but steered me in the right direction whenever he thought I needed it.

I would also like to thank the Mathematics and Finance department at Imperial College London for the opportunity to undertake this MSc programme which has been a challenging but amazing journey.

Last but not least, I would like to thank my family, especially my parents, for their unwavering love and support they have given me. Their encouragement and support were worth more than I can express on paper.

List of symbols

$\mathbb{P}_n$	Empirical measure
$\mathbb{G}_n$	Empirical process
$\mathbb{P}_n^o$	Symmetrized empirical measure
$\mathbb{G}_n^o$	Symmetrized empirical process
$B(\cdot)$	Standard Brownian bridge
$N(\varepsilon, \Theta, d)$	ε-covering numbers for the semi-metric space (Θ, d)
$D(\varepsilon, \Theta, d)$	ε-packing numbers for the semi-metric space (Θ, d)
$N_{[\ ]}(\varepsilon, \mathcal{F}, \\|\cdot\\|)$	ε-bracketing numbers for the normed space of real functions $(\mathcal{F}, \\|\cdot\\|)$
$\\|\cdot\\|_{L_2(\mathbb{P}_n)}$ or $\\|\cdot\\|_n$	L_2 semi-norm generated by empirical measure $\mathbb{P}_n$
$O_{\mathbb{P}}(1)$	Stochastic boundedness
$BV_c\left([a,b];\mathbb{R}^d\right)$	Class of continuous bounded variation path $[a,b] \mapsto \mathbb{R}^d$
$S(\gamma)_{[a,b]}$	Signature for the continuous bounded variation path $\gamma : [a,b] \mapsto \mathbb{R}^d$
$S^N(\gamma)_{[a,b]}$	N order truncated Signature for the continuous bounded variation path γ
$\tilde{S}^N(\gamma)_{[a,b]}$	Flattened N order truncated Signature for the continuous bounded variation path γ
$\shuffle$	Shuffle product operation
Θ^N	Class of linear regressors with respect to N-truncated Signature
$I : \mathbb{N}_+ \mapsto \mathbb{N}_+$	Decision function witch decides the truncated order through data size

Contents

Chapter 1

Introduction

Analysis of time series data is an important subject in finance. One popular topic is time series forecasting due to its large number of practical applications in finance. From the perspective of machine learning, the main challenge for feature engineering is how to efficiently extract historical information and associated patterns to describe the time series data. The feature engineering we would like to discuss in this thesis is path Signature. Daniel Levin, Terry Lyons, and Hao Ni have used truncated Signature features to learn the solution of an SDE [1]. However, there is no theoretical suggestion for the selection of truncated order for a general machine learning model. In this thesis, we would like to find some theoretical explanations for truncated order selection in the context of least square regression.

From a statistical point of view, least square regression problem can be classified as an empirical risk minimization problem. This inspires us to use empirical process theory to study this problem. The main theorem we use to solve our problem is the extensional rate of convergence theorem [2, Page 57, Theorem 6.1]. Another important tool required by our solution is the maximal inequalities which is one of the core parts of empirical process theory.

We will start with the basics of the empirical process in Chapter 2. We first define the main object of study (i.e., the empirical process, Glivenko-Cantelli class, and Donsker class [3][4][5]) and associated applications. Glivenko-Cantelli class and Donsker class are two important classes in the empirical process theory that answer the two essential convergence problems [3][4] in empirical process theory. A fundamental finding of empirical process theory shows that the complexity of an underlying function class is the key for it to be Glivenko-Cantelli class or Donsker class. The Different measures(covering numbers, packing numbers and bracketing numbers) for the complexity of a class are also defined in this Chapter.

In Chapter 3, we will introduce some maximal inequalities which act as a bridge between an empirical process and its underlying index class. To build the maximal inequalities we need for this thesis, we will first introduce Dudley's metric entropy bound in detail (Section 3.2). This result will be used to prove a maximal inequality (with uniform entropy) which will be used in deriving rates of convergence theorem (Chapter 4).

For the last Chapter (which is the original work of the author), we will start with the basics of rough path theory. We would like to learn a controlled ODE using Signature features. We expect that the solution of the ODE can be expressed as a continuous function of Signature, then rough path theory tells us that this continuous function can be arbitrary well approximated by a linear functional. This inspires us to build a least square regression model with truncated Signature as an explanatory variable. This thesis aims to find some mapping $I : \mathbb{N}_+ \mapsto \mathbb{N}_+$ to decide the truncated order N as a function of input data size n. We will carefully explain the solution by several steps in the section 5.4. Generally speaking, the main finding of this thesis shows that the rate of convergence with a decision function I is $O_{\mathbb{P}}\left(\sqrt{\frac{d^{I(n)}}{n}}\right)$ where d is the dimension of input path and $O_{\mathbb{P}}(\cdot)$ denotes the stochastic boundedness. This derives that the suggested truncated order can be chosen by rules like $I : n \mapsto \lceil \log \log_d n \rceil$ or $I : n \mapsto \lceil \sqrt{\log_d n} \rceil$.

Chapter 2

Basics of Empirical Process

2.1 Introduction to empirical processes

We first start with the history of the empirical process theory. This theory first developed in the 1930's and 1940's motivated by the study of the empirical distribution function and its extension. In this section, we introduce some important concepts and results in the empirical process theory. We mainly use the notations and conceptions of [2, Chapter 1, page 4-9].

We first start with some classical statistic concepts. Let $X_1, ...X_n$ are i.i.d. random variables in $\mathbb{R}$ with common cumulative distribution function (c.d.f.) F then the empirical distribution function (e.d.f.) F_n is defined as

$$F_n(x) := \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{(-\infty,x]}\left(X_i\right), \quad x \in \mathbb{R} \tag{2.1.1}$$

Then we can define the corresponding empirical process G_n as :

$$G_n(x) := \sqrt{n}\left(F_n(x) - F(x)\right), \quad x \in \mathbb{R} \tag{2.1.2}$$

It is clear that the strong law of large numbers and the central limit theorem give us the following two classical results:

$$F_n(x) \overset{\text{a.s.}}{\to} F(x)$$

$$\mathbb{G}_n(x) \overset{d}{\to} \mathcal{N}(0, F(x)(1 - F(x)))$$

Furthermore, empirical process theory gives two other further results concerning F_n and G_n.

Theorem 2.1.1 (Glivenko-Cantelli 1933 [3][6])**.**

$$\|F_n - F\|_\infty := \sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \overset{a.s.}{\to} 0$$

Theorem 2.1.2 (Donsker 1952 [4][5])**.**

$$\mathbb{G}_n \overset{d}{\to} B(F) \qquad in\ D\left(\mathbb{R}, \|\cdot\|_\infty\right)$$

Where where $B(\cdot)$ is the standard Brownian bridge process 2.4.8 on $[0,1]$ and $D\left(\mathbb{R}, \|\cdot\|_\infty\right)$ is the space of cadlag functions on $\mathbb{R}$

In the 1950's and 1960's, the need for generalizations of theorems 2.1.1 and 2.1.2 appeared. More precisely, it became apparent that when the observations take values in more general space (such as $\mathbb{R}^d$, or some space of functions, etc.). The e.d.f can no longer be defined as equation 2.1.1. Instead, we can define empirical measure as an extension of e.d.f.

Now, lets consider $X_1, ...X_n$ are i.i.d. random variables take values in a general space $\mathcal{X}$. Then we can define the empirical measure $\mathbb{P}_n$ indexed by some class of real-valued measurable functions $\mathcal{F}$ defined on $\mathcal{X}$:

$$\mathbb{P}_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i} \tag{2.1.3}$$

where δ_x denotes the Dirac measure at x. Then, for any Borel set $A \subset \mathcal{X}$, we write:

$$\mathbb{P}_n(A) := \frac{1}{n}\sum_{i=1}^{n} 1_A\left(X_i\right) = \frac{|\{i \leq n : X_i \in A\}|}{n}$$

Similarly, for any real-valued function f on $\mathcal{X}$, we define:

$$\mathbb{P}_n(f) := \int f d\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^{n} f\left(X_i\right)$$

Moreover, we use the following operator notation for the integral of each function $f \in \mathcal{F}$ with respect to $\mathbb{P}$ (we assume the integrability holds):

$$Pf := \int f d\mathbb{P}$$

Then it is nature to define the **empirical process** $\mathbb{G}_n$ by:

$$\mathbb{G}_n := \sqrt{n}\left(\mathbb{P}_n - P\right) \tag{2.1.4}$$

The collection of random variables $\{\mathbb{G}_n(f) : f \in \mathcal{F}\}$ is called the **empirical process** indexed by $\mathcal{F}$ [2, Chapter 1, page 4-9].

Remark 2.1.3. It is not difficult to check that the classical e.d.f. and empirical process for real-valued random variables can be obtained by setting $\mathcal{X} = \mathbb{R}$ and $\mathcal{F} = \left\{\mathbf{1}_{(-\infty,x]}(\cdot) : x \in \mathbb{R}\right\}$

With the above extension, we can briefly conclude that the main subject of empirical process theory is to analyse the approximation of Pf by $\mathbb{P}_n(f)$. Mainly, we focus on the convergence of following two objects:

- The supremum of the approximation error: $\|\mathbb{P}_n - P\|_{\mathcal{F}} := \sup_{f\in\mathcal{F}} |\mathbb{P}_n f - Pf|$
- The probabilistic limit theorems for the process: $\sqrt{n}\left(\mathbb{P}_n f - Pf\right), \quad f \in \mathcal{F}$

We can easily find out that these two objects are the extension problems for theorems 2.1.1 and 2.1.2. In order to study these problems, we need to introduce two kinds of function class. Let $\mathcal{F} := \{f : \mathcal{X} \to \mathbb{R} : \quad P|f| < \infty\}$ a class of measurable functions.

- We say $\mathcal{F}$ is a P-**Glivenko-Cantelli** class [3] if: $\|\mathbb{P}_n - P\|_{\mathcal{F}} \xrightarrow[n\uparrow\infty]{a.s.} 0$
- We say $\mathcal{F}$ is a P-**Donsker** class [4] if: $\{\mathbb{G}_n(f) : f \in \mathcal{F}\}$ converge to some limiting object in distribution in the space $l^\infty(\mathcal{F})$.

A natural question is that what are the conditions to make a function class to be a P-Glivenko-Cantelli class or a P-Donsker class? This is one of the most important subjects in empirical process theory, it will be discussed in the following sections.

2.2 Empirical Process and Machine Learning

Some regression problems in statistics and machine learning can be generalized as the optimisation problems with M-estimator [7] of the form

$$\hat{\theta}_n := \arg\max_{\theta\in\Theta} \mathbb{P}_n [m_\theta] = \arg\max_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} m_\theta (X_i) \tag{2.2.1}$$

where $X_1, ..., X_n$ are i.i.d variables which take values in a general space $\mathcal{X}$, Θ is the class of candidates for estimator (i.e parameter space), and m_θ is a real valued function (loss function) on $\mathcal{X}$ which evaluates the performance of corresponding estimator $\theta \in \Theta$. In particular, some well known methods like maximum likelihood estimation or least square regression are just special cases with some choices of function m_θ.

Example 2.2.1. *Let $m_\theta = \log p_\theta$, where p_θ is the density of the observations. Then we have*

$$\hat{\theta}_n := \arg\max_{\theta\in\Theta} \mathbb{P}_n [m_\theta] = \arg\max_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \log_\theta p (X_i)$$

Which is the form of maximum likelihood estimator.
More over, let $\{X_i\}_{i=1}^n = \{(Z_i, Y_i)\}_{i=1}^n$ be the observations of a regression model. We choose the square error function for m_θ, i.e. $m_\theta (x) = -(y - \theta(z))^2$ where θ is a regression function in some class of regressor Θ. Then we build the least square regression estimator with the form of

$$\hat{\theta}_n := \arg\max_{\theta\in\Theta} \mathbb{P}_n [m_\theta] = \arg\max_{\theta\in\Theta} -\frac{1}{n}\sum_{i=1}^{n} (Y_i - \theta(Z_i))^2$$

In chapter 5, we will build our main problem of this thesis with the form of least square regression estimation like above example.

In M-estimator problems with form 2.2.1, we are interested in the “true parameter”

$$\theta_0 := \arg\max_{\theta\in\Theta} P [m_\theta]$$

By the law of large numbers, it is clear chat we can approximate the $P [m_\theta]$ with a fixed parameter θ by the empirical risk $\mathbb{P}_n [m_\theta]$ which depends only on the data. Moreover, if the class of all possible m_θ is P-Glivenko Cantelli (i.e. $\mathcal{F} := \{m_\theta(\cdot) : \theta \in \Theta\}$ is P-Glivenko Cantelli), then by its definition [3], $P [m_\theta]$ and $\mathbb{P}_n [m_\theta]$ are uniformly close as n increasing. However, we don't know if their argmax (i.e. θ_0 and$\hat{\theta}_n$) are close. This problem is what we called consistency of M-estimator. The following simple lemma shows that M-estimator is consistent with some assumptions.

Lemma 2.2.1. *[2, Page 30] Let (Θ, d) is a metric space, let $\hat{\theta}_n$ be a M-estimator with the form of 2.2.1. We assume that $\mathcal{F} := \{m_\theta(\cdot) : \theta \in \Theta\}$ is P-Glivenko Cantelli and $\theta_0 := \arg\max_{\theta\in\Theta} P [m_\theta]$ is a well-separated maximizer i.e. $P [m_{\theta_0}] > \sup_{\theta\in\Theta:d(\theta,\theta_0)\geq\delta} P [m_\theta]$, for every $\delta > 0$. Then we have $d\left(\hat{\theta}_n, \theta_0\right) \xrightarrow{\mathbb{P}} 0$.*

Proof. For a fixed $\delta > 0$, let

$$\phi(\delta) := P [m_{\theta_0}] - \sup_{\theta\in\Theta:d(\theta,\theta_0)\geq\delta} P [m_\theta]$$

$d(\hat{\theta}_n, \theta_0) > \delta$ can imply that

$$\begin{aligned}
&P\left[m_{\hat{\theta}_n}\right] \leq \sup_{\theta\in\Theta: d(\theta,\theta_0)\geq\delta} P\left[m_\theta\right] \\
\Leftrightarrow &P\left[m_{\hat{\theta}_n}\right] - P\left[m_{\theta_0}\right] \leq -\phi(\delta) \\
\Rightarrow &P\left[m_{\hat{\theta}_n}\right] - P\left[m_{\theta_0}\right] + \left(\mathbb{P}_n\left[m_{\theta_0}\right] - \mathbb{P}_n\left[m_{\hat{\theta}_n}\right]\right) \leq -\phi(\delta) \\
\Rightarrow &2\sup_{\theta\in\Theta}\left|\mathbb{P}_n\left[m_\theta\right] - P\left[m_\theta\right]\right| \geq \phi(\delta)
\end{aligned}$$

As we assumed that $\mathcal{F}$ is P-Glivenko Cantelli, we have

$$\mathbb{P}\left(d\left(\hat{\theta}_n, \theta_0\right) \geq \delta\right) \leq \mathbb{P}\left(2\sup_{\theta\in\Theta}\left|\mathbb{P}_n\left[m_\theta\right] - P\left[m_\theta\right]\right| \geq \phi(\delta)\right) \to 0 \quad \text{as } n \to \infty$$

□

The conditions of the above lemma maybe too strict sometimes, a more general theorem called the Argmax Theorem is discussed in [8, Pages 285-289, Section 3.2.1]. Assume that we have the consistency for an M-estimator, another problem we are interested in is that how fast the $\hat{\theta}_n$ approach to θ_0 as n increasing. We call this problem the convergence rate of M-estimator. The convergence rate theorems are the most important tools to solve the main problem of this thesis. These theorems will be discussed with details in Chapter 4.

2.3 Complexity of a function class

We have seen that the P-Glivenko Cantelli class and P-Donsker class give some interesting statistical properties, but we still have no idea how to check a class is P-Glivenko Cantelli or P-Donsker. A part of the main findings of the empirical process shows that the complexity of the underlying function class is closely relevant to the convergence problems we have discussed in section 2.1. More precisely, consider $\mathcal{F}$ some class of measurable real-valued functions defined on a general space $\mathcal{X}$, the complexity of the class determines whether $\mathcal{F}$ is P-Glivenko Cantelli class or P-Donsker class. We can define the following quantities [2, Chapter 2, page 14-17] to help us describe how complex the class $\mathcal{F}$ is.

2.3.1 Covering numbers and Packing numbers

Covering numbers and packing numbers are two relatively simple ways to measure the complexity of space. For any semi-metric space (Θ, d), we can first define its ε-cover

Definition 2.3.1 (ε-cover)**.** [2, Page 14-16, Section 2.1] We say the set $\{\theta_1, ..., \theta_N\}$ is a ε-cover of the set Θ with respect to the semi-metric d if any $\theta \in \Theta$ can be covered by a ε-ball with respect to the semi-metric d such that its centre is in the set $\{\theta_1, ..., \theta_N\}$, i.e. for any $\theta \in \Theta$, there exists $i \in \{1, ..., N\}$ such that $d(\theta, \theta_i) < \varepsilon$.

With the above conception, we can define the ε-covering numbers of the set Θ

Definition 2.3.2 (Covering numbers)**.** [2, Page 14-16, Section 2.1] We note the ε-covering numbers of Θ as $N(\varepsilon, \Theta, d)$ which is the minimum number of ε-balls (with respect to the semi-metric d) we need to cover the set Θ. More precisely,

$$N(\varepsilon, \Theta, d) := \inf\left\{N \in \mathbb{N} : \exists \text{ a } \varepsilon\text{-cover } \{\theta_1, \ldots, \theta_N\} \text{ of } \Theta\right\}$$

As the above definition stated, the basic idea of covering numbers is to find how many balls of radius $\varepsilon > 0$ we need to cover the set. On the contrary, we can also define another related measure of complexity that represents how many disjoint balls of radius $\varepsilon > 0$ we can place into the set. In order to do that, we can first define the ε-packing of set Θ.

Definition 2.3.3 (ε-packing)**.** [2, Page 14-16, Section 2.1] We say the set $\{\theta_1, ..., \theta_N\}$ is a ε-packing of the set Θ with respect to the semi-metric d if all the ε-ball with centre $\theta_1, ..., \theta_N$ are disjoint, i.e. for all $i, j \in \{1, ..., N\}$, we have $d(\theta_i, \theta_j) > \varepsilon$.

Like covering number, we can now define the ε-packing number of the set Θ by ε-packing:

Definition 2.3.4 (Packing number)**.** [2, Page 14-16, Section 2.1] We note the ε-packing numbers of Θ as $D(\varepsilon, \Theta, d)$ which is defined by the maximum number of disjoint ε-balls(with respect to the semi-metric d) can be placed in the set Θ, i.e.

$$D(\varepsilon, \Theta, d) := \sup\{N \in \mathbb{N} : \exists \text{ a } \varepsilon\text{-packing } \{\theta_1, \dots, \theta_N\} \text{ of } \Theta\}$$

Remark 2.3.5. There are also equivalent way to define ε-covering and ε-packing. Let $B(\theta_i, \varepsilon)$ be a ε-ball of the semi-metric space (Θ, d). Then we can also say the set $\{\theta_1, ..., \theta_N\}$ is a ε-cover if $\Theta \subset \cup_{i=1}^{N} B(\theta_i, \epsilon)$. We apply the same idea to ε-packing, we say the set $\{\theta_1, ..., \theta_N\}$ is a ε-packing if $\cap_{i=1}^{N} B(\theta_i, \epsilon/2) = \emptyset$

Following the above definitions, there is an important fact that covering numbers and packing numbers are closely related as a consequence of their constructions. More precisely, the following lemma shows that these two measure have the same scaling with the radius ε.

Lemma 2.3.6. *[2, Page 14-16, Section 2.1] For any $\varepsilon > 0$, we have:*

$$D(2\varepsilon, \Theta, d) \le N(\varepsilon, \Theta, d) \le D(\varepsilon, \Theta, d)$$

Proof. We first consider the second inequality. Let $\Delta \subset \Theta$ be a maximal ε-packing. Then by the construction of maximal ε-packing, for every $\theta \in \Theta \setminus \Delta$, there always exists $i \in \{1, ..., |\Delta|\}$ such that $d(\theta, \theta_i) \le \varepsilon$ ($|\cdot|$ notes the cardinal number of the set). Thus, Δ satisfies the definition of ε-covering which means that Δ is also a ε-covering. By the definition of ε-covering numbers, $N(\varepsilon, \Theta, d)$ is the minimal cardinal number of all possible ε-covering sets. Hence, we have the inequality: $N(\varepsilon, \Theta, d) \le D(\varepsilon, \Theta, d)$.

Now we move to the first inequality. Let $\{\alpha_1, ..., \alpha_D\}$ a 2ε-packing and $\{\beta_1, ..., \beta_N\}$ an ε-covering such that $D \ge N + 1$. By pigeonhole principle, there must exists α_i, α_j and a ε-ball $B(\beta_k, \varepsilon)$ such that $\alpha_i, \alpha_j \in B(\beta_k, \varepsilon)$ for some $i \neq j$ and k. Hence, the distance between α_i and α_j can not be larger than the diameter of the ε-ball (β_k, ε). Thus we get $d(\alpha_i, \alpha_j) \le 2\epsilon$, which is a contradiction due to the fact that $\{\alpha_1, ..., \alpha_D\}$ is a 2ε-packing implies that $d(\alpha_i, \alpha_j) > 2\epsilon$. As a consequence, we can conclude that the cardinal number of any 2ε-packing is less or equal to the cardinal number of any ε-covering. □

The above lemma can be useful when we would like to construct some inequalities with complexity (which is actually very common in empirical process theory). For example, assume that we have an upper bound with ε-covering numbers, but it is not easy to be calculated. However, it is not difficult to construct an ε-packing, then we can quickly have the upper bound with ε-packing numbers.

Now let us see an example based on the metric space $(\mathcal{L}, \|\cdot\|_\infty)$ where $\mathcal{L}$ is a Lipschitz ball defined by

$$\mathcal{L} := \{f : [0,1] \to [0,1] | f \text{ is } 1 - \text{Lipschitz}\} \tag{2.3.1}$$

As we showed in lemma 2.3.6, covering numbers and packing numbers are closely related and they are in the same scale, we just consider the ε-covering number of this class in the following example.

Example 2.3.1. *[2, Page 14-16, Section 2.1] Let us consider the ε-covering number of the metric space $(\mathcal{L}, \|\cdot\|_\infty)$ we defined in 2.3.1. We will show an upper bound for $N(\varepsilon, \mathcal{L}, \|\cdot\|_\infty)$ for all $\varepsilon > 0$.*

Firstly, let $\varepsilon \geq 1$, it is nature to choose $f_0 \equiv 0$ and we have $\|f - f_0\|_\infty \leq 1 \leq \varepsilon$ for any $f \in \mathcal{L}$. As a consequence, $N(\varepsilon, \mathcal{L}, \|\cdot\|_\infty) \equiv 1$ for all $\varepsilon \geq 1$.

Now let us set $\varepsilon < 1$. We construct a partition of interval $[0,1]$: $0 = t_0 < t_1 < ... < t_N = 1$ where $t_k = k \cdot \varepsilon$ for $k = 0, 1, 2, ..., N-1$. We define the intervals $A_1 := [t_0, t_1]$ and $A_k := (t_{k-1}, t_k]$ for $k = 2, 3, ...N$. Then we approach any $f \in \mathcal{L}$ by $\tilde{f}$ which is defined by the linear interpolation:

$$\tilde{f}(x) = \sum_{k=1}^{N} \left\{ \frac{f(t_k) - f(t_{k-1})}{t_k - t_{k-1}} (x - t_{k-1}) + f(t_{k-1}) \right\} \mathbf{1}_{A_k}(x)$$

It is clear that $\tilde{f} \in \mathcal{L}$ and $\tilde{f}$ can only take values of form $k \cdot \varepsilon$. By this construction, we have

$$|f(x) - \tilde{f}(x)| \leq |f(x) - f(t_{k-1})| + \left| f(t_{k-1}) - \tilde{f}(x) \right| \leq 2\varepsilon \quad \forall x \in A_k \quad \forall k \in \{1, 2, ..., N\}$$

This is equivalent to $\|f - \tilde{f}\|_\infty \leq 2\varepsilon$ which implies that the collection of all $\tilde{f}$ is a 2ε-covering of $\mathcal{L}$. Now, we can consider the number of distinct $\tilde{f}$ can be constructed as f varies over $\mathcal{L}$. For the first point t_1, due to the construction of $\tilde{f}$, there are at most $\lfloor 1/\varepsilon \rfloor + 1$ possible choices of $\tilde{f}(t_1)$. Moreover, considering the fact that

$$\left| \tilde{f}(t_k) - \tilde{f}(t_{k-1}) \right| = |f(t_k) - f(t_{k-1})| \leq \varepsilon$$

Thus, there are 3 choices for the next value $\tilde{f}(t_k)$ once $\tilde{f}(t_{k-1})$ is fixed. As a conclusion, we can conclude that we have the inequality

$$N(2\varepsilon, \mathcal{L}, \|\cdot\|_\infty) \leq \left(\left\lfloor \frac{1}{\varepsilon} \right\rfloor + 1 \right) 3^{\lfloor 1/\varepsilon \rfloor + 1}$$

which provides an upper bound for the covering numbers of $\mathcal{L}$.

2.3.2 Bracketing numbers

We have defined the covering numbers and packing numbers for any semi-metric space (Θ, d). Now let us move to $(\mathcal{F}, ||\cdot||)$ which is some subset of some normed space of real functions $(\{f : \mathcal{X} \to \mathbb{R}\}, ||\cdot||)$. We would like to introduce the third quantity which measures the complexity of the subset $\mathcal{F}$. We first need the conception of the ε-bracket.

Definition 2.3.7 (ε-bracket)**.** [2, Page 17-18, Section 2.2] Let $l(\cdot)$ and $u(\cdot)$ two real valued functions on $\mathcal{X}$. We define the bracket $[l, u] := \{f : \quad l(x) \leq f(x) \leq u(x), \quad \forall x \in \mathcal{X}\}$. We say $[l, u]$ is an ε-bracket if $||l - u|| < \varepsilon$.

Definition 2.3.8 (Bracketing numbers)**.** [2, Page 17-18, Section 2.2] We define the ε-bracketing number as the minimal number of ε-brackets we need to cover $\mathcal{F}$ with respect to the notation $N_{[\,]}(\varepsilon, \mathcal{F}, \|\cdot\|)$, i.e.

$$N_{[\,]}(\varepsilon, \mathcal{F}, \|\cdot\|) := \inf \left\{ N : \exists \varepsilon\text{-brackets}\, [l_i, u_i] \text{ such that } \mathcal{F} \subset \cup_{i=1}^{N} [l_i, u_i] \right\}$$

One remark of this definition is that the two bound functions $l(\cdot)$ and $u(\cdot)$ of any bracket $[l, u]$ are not necessary belong to $\mathcal{F}$, but we assume that they have finite norms. Another important concept we will need is the envelop function of class $\mathcal{F}$:

Definition 2.3.9 (Envelope function)**.** [2, Page 17-18, Section 2.2] We say $F(\cdot)$ is an envelope function of a class of functions $\mathcal{F}$ if $|f(x)| \leq F(x)$ for every $x \in \mathcal{X}$ and $f \in \mathcal{F}$. Moreover, it is nature to define the minimal envelope function as $x \mapsto \sup_{f\in\mathcal{F}} |f(x)|$.

Now let us consider the class $\mathcal{F} := \left\{f_t := \mathbf{1}_{(-\infty,t]}(\cdot) \middle| t \in \bar{\mathbb{R}}\right\}$ which is the collection of all indicate functions. We have mentioned in Remark 2.1.3 that the empirical process $\mathbb{G}_n$ indexed by this collection is the classical empirical process:

$$\mathbb{G}_n(f_t) = \sqrt{n}\left(\mathbb{F}_n(t) - F(t)\right)$$

with $X_1, ..., X_n$ i.i.d. random variables with c.d.f F under measure P. We will show an upper bound of bracketing number for this class.

Example 2.3.2. *[2, Page 17-18, Section 2.2] Let $(\mathcal{F}, \|\cdot\|_r)$ be a metric space where $\mathcal{F} := \left\{f_t := \mathbf{1}_{(-\infty,t]}(\cdot) \middle| t \in \bar{\mathbb{R}}\right\}$ is the collection of indicate functions and $\|\cdot\|_r$ is $L_r(P)$ norm. We would like to consider the bracketing number of this space. Let $-\infty = t_0 < t_1 < ... < t_k = \infty$ be a sequence of grid point. Then we can construct a sequence of brackets: $\left[\mathbf{1}_{(-\infty,t_{i-1}]}, \mathbf{1}_{(-\infty,t_i]}\right], i = 1, 2, ...k$. It is clear that this sequence of brackets cover the $\mathcal{F}$ due to the construction. Now we control the size of brackets by choosing the grids points such that*

$$\|\mathbf{1}_{(-\infty,t_{i-1}]} - \mathbf{1}_{(-\infty,t_i]}\|_r = \left[F(t_i^-) - F(t_{i-1})\right]^{\frac{1}{r}} \leq \varepsilon$$

Thus, we need at most $\left\lfloor \frac{1}{\varepsilon^r} \right\rfloor + 1$ grid points. As a consequence, we can conclude that

$$N_{[\ ]}\left(\varepsilon, \mathcal{F}, L_r(P)\right) \leq \left\lfloor \frac{1}{\varepsilon^r} \right\rfloor + 1$$

We have build a connection for the ε-covering numbers and the ε-packing numbers by lemma 2.3.6. It is nature to ask if we can build a relation which involves the ε-bracketing numbers. In order to build a connection for these three measures of complexity, we now present a inequality between the ε-covering number and the ε-bracketing numbers.

Theorem 2.3.10. *[2, Page 17-18, Section 2.2] Let $(\mathcal{F}, \|\cdot\|)$ be a metric space where $\mathcal{F}$ is an arbitrary class of function and $\|\cdot\|$ is an arbitrary norm on $\mathcal{F}$. Then we have:*

$$N(\varepsilon, \mathcal{F}, \|\cdot\|) \leq N_{[\ ]}(\varepsilon, \mathcal{F}, \|\cdot\|) \quad \forall \varepsilon > 0$$

Proof. We first fix the $\varepsilon > 0$. We simply note the ε-bracketing number $N_{[\ ]}(\varepsilon, \mathcal{F}, \|\cdot\|)$ of as N. Let $B_1, B_2, ...B_N$ be a sequence of brackets which cover the $\mathcal{F}$. For each bracket B_i, we choose one function g_i such that $g_i \in B_i \cap \mathcal{F}$. Then we define the collection $\mathcal{G} := \cup_{i=1}^N \{g_i\}$. We remark that the cardinal number of $\mathcal{G}$ is N due to its construction. Now we will show that $\mathcal{G}$ is a ε-covering. Consider any $f \in \mathcal{F}$, there exists a bracket B_i such that $f \in B_i$ since the sequence $\{B_i\}_{i=1...N}$ covers the $\mathcal{F}$. Hence, $\|f - g_i\| \leq \varepsilon$ due to the definition of ε-bracket which implies that $\mathcal{G}$ is an ε-covering. Since the ε-covering number is the minimal cardinal number of all possible ε-covering sets, we have $N(\varepsilon, \mathcal{F}, \|\cdot\|) \leq N$ which is the desired result. □

2.4 Glivenko-Cantelli classes and Donsker classes

As we have mentioned in the previous section, the complexity of a class determines whether it is a Glivenko-Cantelli class or a Donsker class. We have also defined three different measures for the complexity of a class. In this section, we will show how the complexities determine a class by some important theorems in the empirical process theory. However, the sufficient or necessary conditions for a class to be Glivenko-Cantelli or Donsker are not the key for this thesis and the proofs of most of these theorems are very long. For simplicity, we only give a brief introduction for this part, more details can be found in [8, Section 2.4, 2.5 and 2.13].

2.4.1 Glivenko-Cantelli classes

Let us first start with the Glivenko-Cantelli classes. We would like to introduce two Glivenko-Cantelli theorems. The first one is relatively simple which is based on the ε-bracketing numbers. We will simply show its proof. The second theorem is based on the ε-covering numbers which is much more difficult to prove. We will only introduce the theorem without proof.

Theorem 2.4.1. *[2, Page 19-20, Section 3.1] Let $\mathcal{F}$ be a class of measurable functions with finite ε-bracketing numbers with respect to $L_1(P)$ norm for any $\varepsilon > 0$ (i.e. $N_{[\,]}(\varepsilon, \mathcal{F}, L_1(P)) < \infty, \quad \forall \varepsilon > 0$). Then $\mathcal{F}$ is Glivenko-Cantelli.*

Proof. The proof of this theorem is just an application of bracketing. Since we assume that any ε-bracketing number is finite, we fix some $\varepsilon > 0$ and we can choose finitely many ε-brackets such that their union cover the $\mathcal{F}$. By this construction, for any $f \in \mathcal{F}$, there exists a ε-bracket $[l_i, u_i]$ such that $l_i(x) \leq f(x) \leq u_i(x)$. Then we have $P(u_i - f) \leq P(u_i - l_i) \leq \varepsilon$ and $P(l_i - f) \geq P(l_i - u_i) \geq -\varepsilon$. Hence we have:

$$(\mathbb{P}_n - P) f \leq (\mathbb{P}_n - P) u_i + P(u_i - f) \leq (\mathbb{P}_n - P) u_i + \varepsilon$$

which implies:

$$\sup_{f \in \mathcal{F}} (\mathbb{P}_n - P) f \leq \max_i (\mathbb{P}_n - P) u_i + \varepsilon \tag{2.4.1}$$

On the other side, we can also build the inequality:

$$(\mathbb{P}_n - P) f \geq (\mathbb{P}_n - P) l_i + P(l_i - f) \geq (\mathbb{P}_n - P) l_i - \varepsilon$$

which implies:

$$\inf_{f \in \mathcal{F}} (\mathbb{P}_n - P) f \geq \min_i (\mathbb{P}_n - P) l_i - \varepsilon \tag{2.4.2}$$

By strong law of large numbers, the right side of inequalities 2.4.1 and 2.4.2 almost surly converge to ε and $-\varepsilon$ respectively. Moreover,

$$\|\mathbb{P}_n - P\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} |(\mathbb{P}_n - P) f| = \max \left\{ \sup_{f \in \mathcal{F}} (\mathbb{P}_n - P) f, - \inf_{f \in \mathcal{F}} (\mathbb{P}_n - P) f \right\} \tag{2.4.3}$$

Then we can conclude that $\limsup \|\mathbb{P}_n - P\|_{\mathcal{F}} \leq \varepsilon$ almost surely, for every $\varepsilon > 0$. Finally, taking a sequence of ε_n such that $\varepsilon_n \downarrow 0$ completes the proof. □

Following theorem gives us another way to identify Glivenko-Cantelli by its covering numbers. Both the statement and the proof of this theorem are more complicated than the previous theorem based on the ε-bracketing numbers. However, the result gives a necessary and sufficient characterization for a class of functions to be Glivenko-Cantelli.

Theorem 2.4.2. *[2, Page 21-22, Section 3.2] Consider $\mathcal{F}$ a class of measurable functions with an integrable envelope F. Let $\mathcal{F}_M$ be a class of functions $\mathcal{F}_M := \{ f \mathbf{1}_{\{F \leq M\}} : f \in \mathcal{F} \}$. Then $\mathcal{F}$ is Glivenko-Cantelli if and only if:*

$$\frac{1}{n} \log N(\varepsilon, \mathcal{F}_M, L_1(\mathbb{P}_n)) \xrightarrow{\mathbb{P}} 0, \quad \forall \varepsilon > 0 \text{ and } \forall M > 0$$

2.4.2 Donsker classes

Now we move to Donsker class. We will also show two theorems for checking Donsker classes. The first one shows a sufficient condition for a class to be Donsker which is related to the grow speed of ε-bracketing number as $\varepsilon \downarrow 0$. In order to introduce this theorem, we first need to define the conception of bracketing entropy integral.

Definition 2.4.3 (Bracketing entropy integral)**.** [2, Page 126-127, Section 11.1] Let $\mathcal{F}$ be a class of measurable functions equipped with L_2 norm. We note $J_{[\,]}(\delta, \mathcal{F}, L_2(P))$ the bracketing entropy integral which is defined as:

$$J_{[\,]}(\delta, \mathcal{F}, L_2(P)) := \int_0^\delta \sqrt{\log N_{[\,]}(\varepsilon, \mathcal{F} \cup \{0\}, L_2(P))} d\varepsilon$$

A nature question for the above definition is if the bracketing entropy integral is finite. This question is actually the key for the following theorem.

Theorem 2.4.4 (Donsker theorem)**.** *[2, Page 126-127, Section 11.1] If $\mathcal{F}$ is a class of measurable functions with square-integrable and measurable envelope F such that the bracketing entropy integral $J_{[\,]}(\delta, \mathcal{F}, L_2(P)) < \infty$. Then $\mathcal{F}$ is P-Donsker.*

Instead of the proof, let us see a simple example of the application for the above theorem. Let us continue with example 2.3.2.

Example 2.4.1. *Let $(\mathcal{F}, \|\cdot\|_1)$ be a metric space where $\mathcal{F} := \{f_t := \mathbf{1}_{(-\infty,t]}(\cdot) \big| t \in \bar{\mathbb{R}}\}$ is the collection of indicate functions and $\|\cdot\|_1$ is $L_1(P)$ norm. Remark that $0 \in \mathcal{F}$ for this example. With the result in example 2.3.2, we have*

$$N_{[\,]}(\varepsilon, \mathcal{F}, L_1(P)) \leq \left\lfloor \frac{1}{\varepsilon^1} \right\rfloor + 1$$

Then we have

$$\begin{aligned} J_{[\,]}(\delta, \mathcal{F}, L_1(P)) &\leq \int_0^\delta \sqrt{\left\lfloor \frac{1}{\varepsilon} \right\rfloor + 1}\; d\varepsilon \\ &\leq \delta + \int_0^\delta \sqrt{\frac{1}{\varepsilon}}\; d\varepsilon \\ &\leq \delta + 2\sqrt{\delta} \end{aligned}$$

Finally, by theorem 2.4.4, the class $\mathcal{F} := \{f_t := \mathbf{1}_{(-\infty,t]}(\cdot) \big| t \in \bar{\mathbb{R}}\}$ is P-Donsker with respect to $L_1(P)$ norm.

The second theorem indicates that the ε-bracketing numbers term in the above Donsker theorem can be replaced by the uniform covering numbers which are defined as $\sup_Q N(\varepsilon\|F\|_{Q,2}, \mathcal{F}, L_2(Q))$ where Q runs over all probability measures such that $Q[F^2] > 0$. Then, we can first define the uniform entropy integral as following.

Definition 2.4.5 (Uniform entropy integral)**.** [2, Page 127-130, Section 11.2] Let $\mathcal{F}$ a class measurable functions with square-integrable and measurable envelope function F, we note $J(\delta, \mathcal{F}, F)$ the uniform entropy integral which is defined by:

$$J(\delta, \mathcal{F}, F) := \int_0^\delta \sup_Q \sqrt{\log N(\varepsilon\|F\|_{Q,2}, \mathcal{F} \cup \{0\}, L_2(Q))} d\varepsilon, \quad \delta > 0$$

Then the following theorem shows another sufficient condition for a class to be Donsker.

Theorem 2.4.6 (Donsker theorem)**.** *[2, Page 127-130, Section 11.2] If $\mathcal{F}$ is a pointwise-measurable class of measurable functions with square-integrable (with respect to $L^2(P)$) and measurable envelope F such that $J(1, \mathcal{F}, F) < \infty$, then $\mathcal{F}$ is P-Donsker.*

In fact, the second theorem can be transferred by the first theorem using a maximal inequality (theorem 3.3.4) which we will discuss in chapter 3. The details are discussed in [2, Page 127-129, Section 11.2].

2.4.3 Functional central limit theorem

The Donsker class plays a key role in weak convergence, some results can derive some useful properties in rough path theory. We would like to explore more properties of Donsker class by simplifying the index set of the empirical process which we consider. For the simplicity, here we assume that the index set is some class of real valued functions in one dimension.

We first back to the classical case that $X_1, ...X_n$ are i.i.d random variables with c.d.f F and the empirical process $\mathbb{G}_n$ is indexed by the real-valued function $\mathbf{1}_{(-\infty,x]}(\cdot)$ where $x \in \mathbb{R}$, then we are in the case of the classical empirical process:

$$\mathbb{G}_n(x) = \sqrt{n}\left(\mathbb{F}_n(x) - F(x)\right), \quad x \in \mathbb{R}$$

where $\mathbb{F}_n$ is the e.d.f of X_i. Then by the classic central limit theorem, we have

$$\mathbb{G}_n(x) \xrightarrow[n\uparrow\infty]{d} \mathcal{N}(0, F(x)(1 - F(x)))$$

Now, let us focus on the classical case, i.e. regard the process $\mathbb{G}_n$ as a stochastic process indexed by the real line $x \in \mathbb{R}$. Functional central limit theorem (also known as Donsker's invariance principle [4] [5]) shows that the sequence $\mathbb{G}_n$, as random elements of the Skorokhod space (the collection of càdlàg functions) converges in distribution to a Gaussian process [9]. More precisely:

Theorem 2.4.7 (Functional central limit theorem)**.** *[5] Let $X_1, ..., X_n$ i.i.d random variables with c.d.f F. The classical empirical process $\mathbb{G}_n(x) := \sqrt{n}\left(\mathbb{F}_n(x) - F(x)\right)$, $x \in \mathbb{R}$ converges in distribution to a Gaussian process $\mathbb{G}$ such that:*

$$\mathbb{E}[\mathbb{G}(x)] = 0 \quad \forall x \in \mathbb{R}$$

$$\operatorname{cov}[\mathbb{G}(s), \mathbb{G}(t)] = F(s) \wedge F(t) - F(s)F(t) \quad \forall s, t \in \mathbb{R}$$

In fact, we can rewrite the above theorem with the form of standard Brownian bridge. We fist recall the definition of Brownian bridge process.

Definition 2.4.8 (Brownian bridge)**.** [10] Let W_t be a Wiener process, we say a continuous-time stochastic process $\{B_t\}_{t\in[0,T]}$ is a Brownian bridge for $t \in [0, T]$ if

$$B_t := \left(W_t \mid W_T = 0\right), t \in [0, T]$$

We remark that the expected value of the bridge is zero and the covariance of $B(s)$ and $B(t)$ is $(s \wedge t)(T - t)/T$. Then we can see that $\mathbb{G}(\cdot)$ can be represented as a Brownian bridge $B(F(\cdot)))$ on the unit interval (i.e. $T = 1$) If we rephrase the above result implies, we have immediately the following lemma [11, Page 14, Section 1]:

Lemma 2.4.9. *[11, Page 14, Section 1] Let $U_{(1)}, U_{(2)}, \ldots, U_{(n)}$ be the order statistics of family on n independent uniform random variables on $[0, 1]$. Define a stochastic process $U^n_{i/(n+1)} := U_{(i)}$*

for $i = 1, 2, \ldots, n$ with the setting $U_0^n := 0$ and $U_1^n := 1$. Taking the piece-wise linear interpolation for all other points t in $[0, 1]$. Define the process $\hat{U}_t^n := \sqrt{n}\,(U_t^n - t)$, then $\hat{U}^n$ converges to the Brownian bridge in distribution as $n \uparrow \infty$.

Proof. This lemma is actually a direct application of theorem 2.4.7. Let $X_1, ...X_n$ i.i.d random variables with c.d.f F. We the fact that $F(X_i) \overset{d}{=} U$ where U is a uniform random variable on $[0, 1]$. We note $F^{-1}(\cdot)$ as the inverse distribution function of F which is defined by

$$F^{-1}(p) := \inf\{x \in \mathbb{R} : F(x) \geq p\}$$

Then we consider the classical empirical process $\mathbb{G}_n(x) = \sqrt{n}\,(\mathbb{F}_n(x) - F(x))$, we note $F(x) = t$ and we have:

$$\begin{aligned}
\mathbb{G}_n\left(F^{-1}(t)\right) &= \sqrt{n}\left(\mathbb{F}_n\left(F^{-1}(t)\right) - t\right) \\
&= \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{(-\infty, F^{-1}(t)]}(X_i) - t\right) \\
&= \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{(0,t)]}(U_i) - t\right) \\
&= \sqrt{n}\,(U_t^n - t) \\
&= \hat{U}_t^n
\end{aligned} \tag{2.4.4}$$

By the theorem 2.4.7, we know that $\mathbb{G}_n(\cdot) \xrightarrow[n\uparrow\infty]{d} \mathbb{G}(\cdot) \overset{d}{=} B(F(\cdot))$. We apply this result to the equation 2.4.4 we get $\hat{U}_n(\cdot) \xrightarrow[n\uparrow\infty]{d} B(\cdot)$ where B is a Brownian bridge on $[0, 1]$, which is the desired result. □

The lemma 2.4.9 is also useful in rough path theory, with this property, we can obtain an asymptotic results for the Signature. Another interesting result is about the convergence rate. Komlós, Major and Tusnády[12][13] established a sharp bound for the speed of the weak convergence for theorem 2.4.7.

Theorem 2.4.10 (Komlós–Major–Tusnády approximation)**.** *[12][13] Let $X_1, ..., X_n$ i.i.d random variables with c.d.f F. The classical empirical process $\mathbb{G}_n(x) := \sqrt{n}\,(\mathbb{F}_n(x) - F(x))$ can be approximated by a sequence of Brownian bridges $B_n(F(x))$ on $[0, 1]$ such that*

$$\mathbb{P}\left[\sup_{x\in\mathbb{R}} |\mathbb{G}_n(x) - B_n(F(x))| > \frac{1}{\sqrt{n}}(a \log n + y)\right] \leqslant be^{-cy} \tag{2.4.5}$$

for all positive integers n and all $y > 0$, where a, b, c are positive constants.

In the next chapter, we will introduce the maximal inequalities which is one of the important tools we need for the main problem of this thesis.

Chapter 3

Maximal Inequalities

3.1 Preliminary

As we have mentioned, maximal inequalities play an important role in empirical process theory. In this chapter, we will introduce some useful maximal inequalities which serve the main problem of this thesis. Before we start the inequalities, we first recall the concept of Sub-Gaussian processes which is the base of maximal inequalities. There are several equivalent definitions, here we show the two most commonly used.

Definition 3.1.1. [2, Page 22-27, Section 3.3] A zero-mean process indexed by T : $\{X_t : t \in T\}$ is a sub-Gaussian process with respect to a metric d on T if:

$$\mathbb{E}\left[e^{\lambda(X_t - X_s)}\right] \leq \exp\left(\frac{\lambda^2 d^2(s,t)}{2}\right), \quad \forall s,t \in T, \quad \forall \lambda \in \mathbb{R} \tag{3.1.1}$$

or equivalently:

$$\mathbb{P}\left[|X_t - X_s| \geq u\right] \leq 2\exp\left(\frac{-u^2}{2d^2(s,t)}\right), \quad \forall s,t \in T, \quad \forall u \geq 0 \tag{3.1.2}$$

An elementary bound for sub-Gaussian random variables indexed by a finite set is given by the following lemma. It is based on the [2, Page 36, Proposition 4.2].

Lemma 3.1.2. *[2, Page 22-27, Section 3.3] Let $\{X_t, t \in T\}$ be a stochastic process indexed by a finite set T such that*

$$\mathbb{P}\left(|X_t| \geq u\right) \leq 2\exp\left(-\frac{u^2}{2\sigma^2}\right), \quad \forall t \in T, \quad \forall u \geq 0 \tag{3.1.3}$$

Where σ is a fixed constant. Then there exists a universal positive constant such that

$$\mathbb{E}\left[\max_{t\in T}|X_t|\right] \leq C\sigma\sqrt{\log(2|T|)} \tag{3.1.4}$$

Moreover, if $|T| > 1$ which is the normal case, we also have

$$\mathbb{E}\left[\max_{t\in T}|X_t|\right] \leq C\sigma\sqrt{\log(|T|)} \tag{3.1.5}$$

Proof. By integration by parts, $\mathbb{E}\left[\max_{t\in T}|X_t|\right]$ can be computed by

$$\mathbb{E}\left[\max_{t\in T}|X_t|\right] = \int_0^\infty \mathbb{P}\left[\max_{t\in T}|X_t| \geq u\right] du$$

Then we can bound $\mathbb{E}\left[\max_{t\in T}|X_t|\right]$ by bounding the tail probability of $\mathbb{P}\left[\max_{t\in T}|X_t|\geq u\right]$:

$$\mathbb{P}\left[\max_{t\in T}|X_t|\geq u\right]=\mathbb{P}\left[\cup_{t\in T}\left\{|X_t|\geq u\right\}\right]\leq\sum_{t\in T}\mathbb{P}\left[|X_t|\geq u\right]\leq 2|T|\exp\left(\frac{-u^2}{2\sigma^2}\right)$$

This bounding method is not good for small u (e.g. u=0). Hence we split the integral by some $u_0>0$:

$$\begin{aligned}\mathbb{E}\left[\max_{t\in T}|X_t|\right]&=\int_0^\infty\mathbb{P}\left[\max_{t\in T}|X_t|\geq u\right]du\\&=\int_0^{u_0}\mathbb{P}\left[\max_{t\in T}|X_t|\geq u\right]du+\int_{u_0}^\infty\mathbb{P}\left[\max_{t\in T}|X_t|\geq u\right]du\\&\leq u_0+\int_{u_0}^\infty 2|T|\exp\left(\frac{-u^2}{2\sigma^2}\right)du\\&\leq u_0+\int_{u_0}^\infty 2|T|\frac{u}{u_0}\exp\left(\frac{-u^2}{2\sigma^2}\right)du\\&\leq u_0+\frac{2|T|}{u_0}\sigma^2\exp\left(\frac{-u_0^2}{2\sigma^2}\right)\end{aligned}$$

Now we can chose u_0 to minimize the above inequality, notice that if $|T|>1$, one simple solution is to set

$$u_0=\sqrt{2}\sigma\sqrt{\log(|T|)}$$

Then we have:

$$\mathbb{E}\left[\max_{t\in T}|X_t|\right]\leq\sqrt{2}\sigma\left(\sqrt{\log\left(|T|\right)}+\frac{1}{\sqrt{\log\left(|T|\right)}}\right)\leq C\sigma\sqrt{\log(|T|)}$$

We already get the second maximal inequality we want, we notice that the this inequality is not true if $|T|=1$, hence we may set $u_0=\sqrt{2}\sigma\sqrt{\log(2|T|)}$ to avoid this problem(since $\log(2|T|)>0$, $\forall T\neq\emptyset$). Then we can build the first result due to the fact that there exists a universal constant C such that

$$\mathbb{E}\left[\max_{t\in T}|X_t|\right]\leq\sqrt{2}\sigma\left(\sqrt{\log\left(2|T|\right)}+\frac{1}{\sqrt{\log\left(2|T|\right)}}\right)\leq C\sigma\sqrt{\log(2|T|)}$$

□

3.2 Dudley's entropy bound

In this section, we would like to extend lemma 3.1.2 to a maximal inequality called Dudley's integral entropy bound. Suppose (T,d) is a metric space and $\{X_t\}_{t\in T}$ a stochastic process indexed by T with zero means. We would like to find upper bonds of $\mathbb{E}\left[\sup_{t\in T}|X_t-X_{t'}|\right]$ for any fixed t'. It is clear that the upper bonds only depend on the structure of T. In order to get a general result, we first constrain the index set T to be finite and try to extend it to infinite case.

3.2.1 Dudley's entropy bound for finite index set

Theorem 3.2.1 (Dudley's integral entropy bound for finite space). *[2, Page 22-27, Section 3.3] Let (T,d) be a finite metric space and $\{X_t:t\in T\}$ be a sub-Gaussian process. Then we have the inequality:*

$$\mathbb{E}\left[\max_{t\in T}|X_t-X_{t_0}|\right]\leq C\int_0^\infty\sqrt{\log N(\epsilon,T,d)}\,d\epsilon,\quad\forall t_0\in T\qquad(3.2.1)$$

Proof. The main methodology of the proof is an idea called chaining. We use the fact that T is finite, then the diameter of T is well defined, let us note it as D which is clearly a finite number. For each $m = 1, 2, ...$ define $\varepsilon_m = D \cdot 2^{-m}$. Let $T_m := \left\{t_1, ..., t_{N(\varepsilon_m, T, d)}\right\} \subset T$ be a minimal ε_m-covering of T. It is clear that ε_m is decreasing and the cardinal number of T_m is increasing and there always exists some ε_m small enough such that $|T_m| = |T|$. In this case, we can chose $T_m = T$. We define M is the minimal number such that T_M can be chosen as T, i.e.

$$M := \min \{m \geq 1 : |T_m| = |T|\}$$

We define the mapping $\pi_m : T \to T_m$ as:

$$\pi_m(t) = \underset{s \in T_m}{\operatorname{argmin}}\ d(t, s)$$

i.e. map each point $t \in T$ to a point in T_m which is the closest to t. For the simplicity of notation, we define $\pi_0(t) := t_0$. By the construction, we have $\pi_M(t) = t$ for every $t \in T$, then we can decompose $X_t - X_{t_0}$ by the technique called chaining:

$$X_t - X_{t_0} = \sum_{k=1}^{M} \left(X_{\pi_k(t)} - X_{\pi_{k-1}(t)}\right), \quad \forall t \in T$$

Then it is clear that:

$$\begin{aligned}
\max_{t \in T} |X_t - X_{t_0}| &= \max_{t \in T} \left|\sum_{k=1}^{M} X_{\pi_k(t)} - X_{\pi_{k-1}(t)}\right| \\
&\leq \max_{t \in T} \sum_{k=1}^{M} \left|X_{\pi_k(t)} - X_{\pi_{k-1}(t)}\right| \\
&\leq \sum_{k=1}^{M} \max_{t \in T} \left|X_{\pi_k(t)} - X_{\pi_{k-1}(t)}\right|
\end{aligned}$$

Taking the expectation, we obtain

$$\mathbb{E}\left[\max_{t \in T} |X_t - X_{t_0}|\right] \leq \sum_{k=1}^{M} \mathbb{E}\left[\max_{t \in T} \left|X_{\pi_k(t)} - X_{\pi_{k-1}(t)}\right|\right] \tag{3.2.2}$$

Now we would like to bound the terms $\mathbb{E}\left[\max_{t \in T} \left|X_{\pi_k(t)} - X_{\pi_{k-1}(t)}\right|\right]$. Since $\{X_t\}_{t \in T}$ is a sub-Gaussian process, by inequality 3.1.2 we have:

$$\mathbb{P}\left[\left|X_{\pi_n(t)} - X_{\pi_{n-1}(t)}\right| \geq u\right] \leq 2 \exp\left(\frac{-u^2}{2\sigma^2}\right)$$

where σ is chosen as

$$\sigma = d\left(\pi_n(t), \pi_{n-1}(t)\right) \leq d\left(\pi_n(t), t\right) + d\left(\pi_{n-1}(t), t\right) \leq D2^{-n} + D2^{-(n-1)} = 3D2^{-n}$$

Then we can apply the maximal inequality 3.1.5 of Lemma 3.1.2 to find an upper bound for

$\mathbb{E}\left[\max_{t\in T}\left|X_{\pi_k(t)} - X_{\pi_{k-1}(t)}\right|\right]$ by

$$\begin{aligned}\mathbb{E}\left[\max_{t\in T}\left|X_{\pi_k(t)} - X_{\pi_{k-1}(t)}\right|\right] &\leq 2\cdot C\frac{3D}{2^k}\sqrt{\log|T_k|}\\ &\leq 3CD2^{-(k+1)}\sqrt{\log N\left(D2^{-k},T,d\right)} \qquad (3.2.3)\\ &\leq 3C\int_{D\cdot 2^{-(k+1)}}^{D\cdot 2^{-k}}\sqrt{\log N(\varepsilon,T,d)}\;d\varepsilon\end{aligned}$$

However, we should be attention to the case $k=1$ since the maximal inequality 3.1.5 cannot be applied for this situation. However, this inequality still holds for our situation. When $k=1$, there is only one element in T_1, which implies that $\mathbb{E}\left[\max_{t\in T}\left|X_{\pi_1(t)} - X_{t_0}\right|\right] = 0$ while the integral term is always non negative. As a consequence the above inequality holds for every k. Then we have

$$\begin{aligned}\mathbb{E}\left[\max_{t\in T}|X_t - X_{t_0}|\right] &= \sum_{k=1}^{|T|}3C\int_{D\cdot 2^{-(k+1)}}^{D\cdot 2^{-k}}\sqrt{\log N(\varepsilon,T,d)}\;d\varepsilon\\ &\leq 3C\int_0^{D/2}\sqrt{\log N(\varepsilon,T,d)}d\varepsilon\\ &\leq 3C\int_0^{D}\sqrt{\log N(\varepsilon,T,d)}d\varepsilon\end{aligned}$$

The proof will be completed the fact of covering numbers that:

$$\int_0^\infty\sqrt{\log N(\varepsilon,T,d)}d\varepsilon = \int_0^D\sqrt{\log N(\varepsilon,T,d)}d\varepsilon \qquad (3.2.4)$$

□

3.2.2 Dudley's entropy bound for infinite index set

Now, we would like to extend the theorem 3.2.1 to the case where the index set T is infinite. In order to do that, we need to make an assumption of separability for the process $\{X_t\}_{t\in T}$. We first recall the definition of the separable stochastic process.

Definition 3.2.2 (Separable stochastic process)**.** [2, Page 22-27, Section 3.3] Suppose (T,d) is a metric space and $\{X_t, t\in T\}$ a stochastic process indexed by T. We say $\{X_t\}$ is separable if there exists a null set Ω_0 and a countable dense subset $\tilde{T}$ of T such that for all $\omega\notin\Omega_0$ and $t\in T$, there exists a sequence of $\{t_n\}$ in $\tilde{T}$ with $\lim_{n\to\infty} d(t_n,t) = 0$ and $\lim_{n\to\infty} X_{t_n}(\omega) = X_t(\omega)$.

With the assumption of separability, we can easily build the following lemma by the separability and continuity.

Lemma 3.2.3. *[2, Page 22-27, Section 3.3] If $\{X_t, t\in T\}$ is a separable stochastic process, then we have*

$$\sup_{t\in T}|X_t - X_{t_0}| = \sup_{t\in\tilde{T}}|X_t - X_{t_0}|, \quad \forall t_0\in T \quad \textit{almost surly}$$

where $\tilde{T}$ is a countable dense subset of T.

Now we can state Dudley's bound for the separable process.

Theorem 3.2.4. *Suppose (T,d) is a separable metric space and $\{X_t, t\in T\}$ a separable stochastic process indexed by T. If $\{X_t\}$ is also a sub-Gaussain process, then for any $t_0\in T$, we have*

$$\mathbb{E}\left[\max_{t\in T}|X_t - X_{t_0}|\right] \leq C\int_0^\infty\sqrt{\log N(\varepsilon,T,d)}d\varepsilon, \quad \forall t_0\in T \qquad (3.2.5)$$

Proof. Let $\tilde{T}$ be a countable subset of T such that lemma 3.2.3 holds. Moreover, we can also add t_0 into $\tilde{T}$. Then for each $n \geq 1$, we define $\tilde{T_n}$ be the finite subset of $\tilde{T}$ by taking the first n elements of $\tilde{T}$ in some enumeration of $\tilde{T}$ such that $t_0 \in \tilde{T_n}$ for every n. Then we can apply the Dudley's bound for finite index set (Theorem 3.2.1) and the equation 3.2.4.

$$\begin{aligned}
\mathbb{E}\left[\max_{t\in\tilde{T}_n}|X_t - X_{t_0}|\right] &\leq C\int_0^\infty \sqrt{\log N\left(\varepsilon, \tilde{T_n}, d\right)}d\varepsilon \\
&\leq C\int_0^{\operatorname{diam}\left(\tilde{T}_n\right)} \sqrt{\log N\left(\varepsilon, \tilde{T_n}, d\right)}d\varepsilon \\
&\leq C\int_0^{D} \sqrt{\log N(\varepsilon, T, d)}d\varepsilon \\
&\leq C\int_0^{\infty} \sqrt{\log N(\varepsilon, T, d)}d\varepsilon
\end{aligned}$$

It is clear that the right side is independent of n. Letting $n \to \infty$, we apply the Monotone Convergence Theorem to the left side, then we can obtain the desired result.

$$\mathbb{E}\left[\max_{t\in T}|X_t - X_{t_0}|\right] \leq C\int_0^\infty \sqrt{\log N(\varepsilon, T, d)}d\varepsilon$$

□

3.3 Maximal inequality with uniform entropy

With the results we obtained in previous sections, we can finally build the most important maximal inequality for this thesis. The maximal inequality we will state in this section serves to drive the rate of convergence in the next chapter. In piratical application, we usually consider the need to bound the uniform entropy of empirical process $\mathbb{G}_n$ indexed by some function space $\mathcal{F}$ when we study the convergence problem. In order to construct this inequality, we first need to introduce a technique called symmetrization.

3.3.1 Symmetrization

Let us consider an arbitrary empirical process $\mathbb{G}_n$ indexed by some class of functions $\mathcal{F}$. The main idea of symmetrization is to build a symmetric empirical process to approximate the process $\sum_{i=1}^n (f(X_i) - Pf)$ for some $f \in \mathcal{F}$. With this construction, we can also build a symmetric empirical process to approximate $\mathbb{G}_n$. The symmetric process we use here is called the Rademacher process.

Definition 3.3.1 (Rademacher process)**.** [2, Page 22-27, Section 3.3] Let $\varepsilon_1, ...\varepsilon_n$ i.i.d Rademacher variables (i.e. $\mathbb{P}(\varepsilon_i = 1) = \mathbb{P}(\varepsilon_i = -1) = \frac{1}{2}$). The following process X_n is called a Rademacher process:

$$X_a := \sum_{i=1}^n a_i\varepsilon_i, \quad a := (a_1, \ldots, a_n) \in \mathbb{R}^n$$

A very important fact of Rademacher process is that a Rademacher process is also a sub-Gaussian process. The following lemma will show this fact by checking the second definition of the sub-Gaussian process (equation 3.1.1).

Lemma 3.3.2 (Hoeffding's inequality)**.** *[2, Page 22-27, Section 3.3] Let $a = (a_1, ..., a_n)$ be a real* *constant vector, $\varepsilon_1, ..., \varepsilon_n$ be Rademacher random variables. Then we have*

$$\mathbb{P}\left[\left|\sum_{i=1}^{n} a_i\varepsilon_i\right| \geq x\right] \leq 2e^{-x^2/\left(2\|a\|_2^2\right)}$$

which implies that $\{\sum_{i=1}^{n} a_i\varepsilon_i\}_{n=1,2,...}$ is a sub-Gaussian process.

Proof. It is not difficult to show that for any $\beta \in R$, we have the inequality

$$\mathbb{E}\left[e^{\beta\varepsilon}\right] = \left(e^{\beta} + e^{-\beta}\right)/2 \leq e^{\beta^2/2}$$

Then by Markov's inequality

$$\begin{aligned}\mathbb{P}\left[\left|\sum_{i=1}^{n} a_i\varepsilon_i\right| \geq x\right] &= 2 \cdot \mathbb{P}\left[\sum_{i=1}^{n} a_i\varepsilon_i \geq x\right] \\ &= 2 \cdot \mathbb{P}\left[e^{\beta\sum_{i=1}^{n} a_i\varepsilon_i} \geq e^{\beta x}\right] \\ &\leq 2 \cdot e^{-\beta x} \cdot \mathbb{E}\left[e^{\beta\sum_{i=1}^{n} a_i\varepsilon_i}\right] \\ &\leq 2 \cdot e^{-\beta x} \cdot e^{\left(\frac{\beta^2}{2}\right)\cdot||a||_2^2} \\ &\leq 2e^{-x^2/\left(2\|a\|_2^2\right)}\end{aligned}$$

□

As we mentioned at the beginning, the idea of symmetrization is to replace $\sum_{i=1}^{n}\left(f\left(X_i\right) - Pf\right)$ by $\sum_{i=1}^{n}\varepsilon_i f\left(X_i\right)$. By this construction, $\sum_{i=1}^{n}\varepsilon_i f\left(X_i\right)$ is a sub-Gaussian process conditionally on $X_1, ...X_n$. We define the symmetrized empirical measure and the symmetrized empirical process by:

$$f \mapsto \mathbb{P}_n^o f := \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f\left(X_i\right), \quad f \mapsto \mathbb{G}_n^o f := \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i f\left(X_i\right) \tag{3.3.1}$$

The reason we use approximation is that we can find an upper bound for $\mathbb{E}\left[\|\mathbb{P}_n - P\|_{\mathcal{F}}\right]$ by the upper bound for the expectation of its corresponding symmetrized empirical process. Moreover, the symmetrized empirical process is a sub-Gaussian process (due to the lemma 3.3.2), we can then use the maximal inequalities like 3.2.5 to express the upper bond. This fact is stated by the following theorem which is also discussed in [2, Page 27-28, Theorem 3.14].

Theorem 3.3.3. *[2, Page 22-27, Section 3.3] Let $X_1, ..., X_n$ i.i.d random variables, $\varepsilon_1, ...\varepsilon_n$ i.i.d Rademacher variables inpedent to X_i and $\mathcal{F}$ be a class of measurable function, then*

$$\mathbb{E}\left[\|\mathbb{P}_n - P\|_{\mathcal{F}}\right] \leq 2\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f\left(X_i\right)\right\|_{\mathcal{F}}\right] \tag{3.3.2}$$

Proof. Let $Y_1, ..., Y_n$ independent copies of $X_1, ..., X_n$. For fixed values of $X_1, ..., X_n$, by Jensen's inequality we have

$$\|\mathbb{P}_n - P\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}}\frac{1}{n}\left|\sum_{i=1}^{n}\left[f\left(X_i\right) - \mathbb{E}\left[f\left(Y_i\right)\right]\right]\right| \leq \mathbb{E}_Y\left[\sup_{f\in\mathcal{F}}\frac{1}{n} \mid \sum_{i=1}^{n}\left[f\left(X_i\right) - f\left(Y_i\right)\right]\right]$$

Where $\mathbb{E}_Y$ denotes the expectation with respect to $Y_1, ..., Y_n$. Then we take the expectation with respect to $X_1, ..., X_n$, we obtain

$$\mathbb{E}\left[\|\mathbb{P}_n - P\|_{\mathcal{F}}\right] \leq \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\left[f\left(X_i\right) - f\left(Y_i\right)\right]\right\|_{\mathcal{F}}\right]$$

We notice that adding a minus sign in front of a term $[f(X_i) - f(Y_i)]$ has the effect of exchanging X_i and Y_i because of the independent copy construction. Hence we can build the following equality

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}[f(X_i) - f(Y_i)]\right\|_{\mathcal{F}}\right] = \mathbb{E}_{\varepsilon}\left[\mathbb{E}_{X,Y}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i[f(X_i) - f(Y_i)]\right\|_{\mathcal{F}}\right]\right]$$

where $\varepsilon_1, ...\varepsilon_n$ are i.i.d Rademacher variables. Hence, with triangle inequality we have

$$\begin{aligned}\mathbb{E}[\|\mathbb{P}_n - P\|_{\mathcal{F}}] &\leq \mathbb{E}_{\varepsilon}\left[\mathbb{E}_X\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i[f(X_i) - f(Y_i)]\right\|_{\mathcal{F}}\right]\right] \\ &\leq 2\mathbb{E}_{\varepsilon}\left[\mathbb{E}_{X,Y}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(X_i)\right\|_{\mathcal{F}}\right]\right] \qquad (3.3.3) \\ &\leq 2\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(X_i)\right\|_{\mathcal{F}}\right]\end{aligned}$$

□

As a consequence of the above theorem, we can always use $\mathbb{E}[\|\mathbb{P}_n^o\|_{\mathcal{F}}]$ as an upper bond for $\mathbb{E}[\|\mathbb{P}_n - P\|_{\mathcal{F}}]$.

3.3.2 Uniform entropy inequality

In last section, we already showed that we have $\mathbb{E}[\|\mathbb{P}_n - P\|_{\mathcal{F}}] \leq \mathbb{E}[\|\mathbb{P}_n^o\|_{\mathcal{F}}]$. With inequality 3.2.5, we can directly build the maximal inequality with covering numbers 3.3.5. In this section, we would like to develop this results to find more maximal inequalities. We first recall definition of the uniform entropy for a measurable functions class $\mathcal{F}$ with a square integrable and measurable envelope function F

$$J(\delta, \mathcal{F}, F) := \int_0^{\delta}\sup_Q\sqrt{\log N(\varepsilon\|F\|_{Q,2}, \mathcal{F}\cup\{0\}, L_2(Q))}\,d\varepsilon, \quad \delta > 0$$

The following theorem states a maximal inequality with uniform entropy.

Theorem 3.3.4 (Uniform entropy inequality)**.** *[2, Page 22-27, Section 3.3] Let $\mathcal{F}$ be a class of measurable functions with a square integrable and measurable envelope function F, then*

$$\mathbb{E}[\|\mathbb{G}_n\|_{\mathcal{F}}] \lesssim \mathbb{E}[J(\theta_n, \mathcal{F}, F)\|F\|_n] \lesssim J(1, \mathcal{F}, F)\|F\|_{P,2} \qquad (3.3.4)$$

where $\theta_n := \sup_{f\in\mathcal{F}}\|f\|_n/\|F\|_n$, $\|f\|_n^2 := \frac{1}{n}\sum_{i=1}^{n} f^2(X_i)$

Here we use the idea of the proof for [2, Page 43-44, Theorem 4.8].

Proof. We use the idea of the proof of the theorem 3.3.3 (inequality 3.3.3), we can bound $\mathbb{E}[\|\mathbb{G}_n\|_{\mathcal{F}}]$ by bounding $\mathbb{E}[\|\mathbb{G}_n^o\|_{\mathcal{F}}]$ where $\mathbb{G}_n^o f := \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i f(X_i)$. We remark that $\mathbb{G}_n^o$ is a sub-Gaussian process due to the lemma 3.3.2. Then we have

$$\mathbb{P}\left(\left|\sum_{i=1}^{n}\varepsilon_i\frac{f(X_i)}{\sqrt{n}} - \sum_{i=1}^{n}\varepsilon_i\frac{g(X_i)}{\sqrt{n}}\right| \geq u \mid X_1, \ldots, X_n\right) \leq 2e^{-u^2/\left(2\|f-g\|_n^2\right)}, \forall f, g \in \mathcal{F}, \forall u \geq 0$$

We note $\sigma_{n,2}^2 := \sup_{f\in\mathcal{F}}\mathbb{P}_n f^2 = \sup_{f\in\mathcal{F}}\|f\|_n$ the upper bound for the squared radius of $\mathcal{F}\cup\{0\}$ with respect to this norm. We add the function $f \equiv 0$ to $\mathcal{F}$, so that the symmetrized process is zero at some parameter. Then we can apply the maximal inequality 3.2.5 with $X_{t_0} = 0$

$$\mathbb{E}_\varepsilon \left[\|\mathbb{G}_n^o\|_{\mathcal{F}}\right] \lesssim \int_0^{\sigma_{n,2}} \sqrt{\log N\left(\varepsilon, \mathcal{F} \cup \{0\}, L_2\left(\mathbb{P}_n\right)\right)}\, d\varepsilon \tag{3.3.5}$$

where $\mathbb{E}_\varepsilon$ is the expectation with respect to the Rademacher variables, given fixed $X_1, ...X_n$ and $L_2\left(\mathbb{P}_n\right)$ is the semi-norm generated by the empirical measure $\mathbb{P}_n$ (also denoted as $\|\cdot\|_n$ for the simplification of notation), in another word

$$\|f\|_{L_2(\mathbb{P}_n)}^2 := \|f\|_n^2 = \frac{1}{n}\sum_{i=1}^n f^2\left(X_i\right). \tag{3.3.6}$$

Making a change of variable and bounding the random entropy by a supremum we see that the right side is bounded by

$$\int_0^{\sigma_{n,2}/\|F\|_n} \sqrt{\log N\left(\varepsilon\|F\|_n, \mathcal{F} \cup \{0\}, L_2\left(\mathbb{P}_n\right)\right)}\, d\varepsilon \|F\|_n \leq J\left(\theta_n, \mathcal{F}, F\right)\|F\|_n$$

Taking the expectation over $X_1, ..., X_n$, we obtain the first inequality we need. It is clear that $\theta_n \leq 1$, so we have $J\left(\theta_n, \mathcal{F}, F\right) \leq J(1, \mathcal{F}, F)$. Then we apply the Jensen's inequality to the root function $\mathbb{E}\left[\|F\|_n\right] \leq \sqrt{\mathbb{E}\left[n^{-1}\sum_{i=1}^n F^2\left(X_i\right)\right]} = \|F\|_{P,2}$ which completes the proof. □

The maximal inequality 3.3.4 and 3.3.5 are two important results that we will use in Chapter 4 and Chapter 5.

Chapter 4

Rate of Convergence for M-estimator

As we mentioned in section 2.2, the rate of convergence for M-estimator is an important topic in statistical and machine learning problems. We first introduce the abstract result stated by the key theorem [8, Page 322, Theorem 3.2.5] in the following section.

4.1 Rate of convergence theorem

We first state the setting for the problem. Let $X_1, ...X_n$ be i.i.d observations for M-estimation model, we would like to estimate the true predictor θ_0 by an M-estimator $\hat{\theta}_n$ in a class of candidates Θ. We assume that there is a semi-metric d, and (Θ, d) is a semi-metric space. Let $\{\mathbb{M}_n(\theta) : \theta \in \Theta\}$ denote a stochastic process indexed by Θ and let $\{M_n(\theta) : \theta \in \Theta\}$ deterministic function such that

$$\hat{\theta}_n = \underset{\theta\in\Theta}{\mathrm{argmax}} \mathbb{M}_n(\theta) =: \arg\max_{\theta\in\Theta} \mathbb{P}\left[m_\theta\right]$$

and

$$\theta_0 = \underset{\theta\in\Theta}{\mathrm{argmax}} M(\theta) =: \arg\max_{\theta\in\Theta} P\left[m_\theta\right]$$

where m_θ is a function depends on the statistical model we use.

Then we assume that the metric d is appropriately chosen such that we may expect that the asymptotic difference decreases quadratically when θ moves away from θ_0, i.e. for every θ in a neighborhood of θ_0, there exists a $c_1 > 0$ such that

$$M(\theta) - M\left(\theta_0\right) \leq -c_1 d^2\left(\theta, \theta_0\right) \tag{4.1.1}$$

We are interested in the situation that our estimator can δ closely approach to the real predictor θ_0 when n is large, i.e. we only need to search the estimator in the candidates space $\{\theta : d\left(\theta, \theta_0\right) < \delta\}$. We assume that for a small $\delta > 0$ and large n , the centred process $\mathbb{M}_n - M_n$ satisfies

$$\mathbb{E}\left[\sup_{d(\theta,\theta_0)<\delta} \sqrt{n}\left|\left(\mathbb{M}_n - M\right)(\theta) - \left(\mathbb{M}_n - M\right)\left(\theta_0\right)\right|\right] \leq c_2\phi_n(\delta) \tag{4.1.2}$$

for some $c_2 < \infty$ and function ϕ_n such that $\delta \mapsto \phi_n(\delta)/\delta^\alpha$ is decreasing for some $\alpha < 2$ but not depending on n. Moreover, if we can found a series δ_n such that

$$\phi_n\left(\delta_n\right) \leq c_3\sqrt{n}\delta_n^2 \tag{4.1.3}$$

for every n and some $c_3 < \infty$. Then δ_n is the rate of convergence for the M-estimator $\hat{\theta}_n$. Before we rephrase the above statement as the rate of convergence theorem, we first define the notation for stochastic boundedness $O_{\mathbb{P}}(1)$.

Definition 4.1.1 (Stochastic Boundedness)**.** [14] Let $\{X_n\}_{n\in\mathbb{N}}$ be a stochastic process, then we define the notation

$$X_n = O_{\mathbb{P}}(1) \Longleftrightarrow \forall\varepsilon \quad \exists N_\varepsilon, \delta_\varepsilon \quad \text{ such that } \mathbb{P}\left(\left|X_n\right| \geq \delta_\varepsilon\right) \leq \varepsilon \quad \forall n > N_\varepsilon \tag{4.1.4}$$

Theorem 4.1.2 (Rate of convergence)**.** *[2, Page 49, Theorem 5.2] Let $\{\mathbb{M}_n(\theta) : \theta \in \Theta\}$ be a process indexed by Θ and let $\{M_n(\theta) : \theta \in \Theta\}$ be a deterministic function such that the assumptions 4.1.1, 4.1.2 and 4.1.3 are satisfied. If the sequence $\hat{\theta}_n$ satisfies*

$$\mathbb{M}_n\left(\hat{\theta}_n\right) \geq \mathbb{M}_n\left(\theta_0\right) - O_{\mathbb{P}}\left(\delta_n^2\right) \tag{4.1.5}$$

and converges in probability to θ_0, then we have the convergence rate:

$$\delta_n^{-1} d\left(\hat{\theta}_n, \theta_0\right) = O_{\mathbb{P}}(1) \tag{4.1.6}$$

Proof. [proof of Theorem 4.1.2] Here we restate the proof in a more detailed way based on the idea of [15, Page 259-261].

We first fix some $\varepsilon > 0$. By the definition of stochastic boundedness, there exists some K such that

$$\mathbb{P}\left[\left(M_n\left(\hat{\theta}_n\right) - \mathbb{M}_n\left(\theta_0\right)\right) \leqslant -K \cdot \delta_n^2\right] \leqslant \varepsilon.$$

For each n, we can decompose the range of $\delta_n^{-1} d\left(\hat{\theta}_n, \theta_0\right)$ into "peels" $S_{j,n}$ which is defined by

$$S_{j,n} := \left\{\theta : 2^{j-1} < \delta_n^{-1} d\left(\theta, \theta_0\right) \leq 2^j\right\}$$

with j running over the integers.

For any $\eta > 0$, we have:

$$\begin{aligned}
\mathbb{P}\left[d\left(\hat{\theta}_n, \theta_0\right) > 2^M\delta_n\right] &= \sum_{j>M, 2^j\delta_n\leq\eta} \mathbb{P}\left[\hat{\theta}_n \in S_{j,n}\right] + \mathbb{P}\left(2d\left(\hat{\theta}_n, \theta_0\right) > \eta\right) \\
&\leq \sum_{j\geq M, 2^j\delta_n\leq\eta} \mathbb{P}\left[\sup_{\theta\in S_{j,n}} \left[\mathbb{M}_n(\theta) - \mathbb{M}_n\left(\theta_0\right) + K\delta_n^2\right] \geq 0\right] \\
&\quad + \mathbb{P}\left[2d\left(\hat{\theta}_n, \theta_0\right) \geq \eta\right] + \mathbb{P}\left[M_n\left(\hat{\theta}_n\right) - M_n\left(\theta_0\right) < -K\delta_n^2\right]
\end{aligned} \tag{4.1.7}$$

Taking $n \to \infty$, the sum of the last two terms of 4.1.7 is smaller than 2ε due to the choose of K and the consistency of $\hat{\theta}_n$. Considering the centered process $U_n(\theta) := \mathbb{M}_n(\theta) - M(\theta)$ for $\theta \in \Theta$. Using the fact that for every $\theta \in S_{j,n}$, we have $M(\theta) - M\left(\theta_0\right) \leq -c_1 d\left(\theta, \theta_0\right) \leq -c_1 2^{2j-2}\delta_n^2$, the rest term of right side of 4.1.7 can be written as:

$$\begin{aligned}
&\sum_{j\geq M, 2^j\delta_n\leq\eta} \mathbb{P}\left[\sup_{\theta\in S_{j,n}}\left[\mathbb{M}_n(\theta)-\mathbb{M}_n(\theta_0)+K\delta_n^2\right]\geq 0\right] \\
&=\sum_{j\geq M, 2^j\delta_n\leq\eta} \mathbb{P}\left[\|U_n(\theta)-U_n(\theta_0)\|_{S_{j,n}}\geq c_1(2^{2j-2}-K)\delta_n^2\right] \\
&\leq \sum_{j\geq M}\frac{c_2\phi_n\left(2^j\delta_n\right)}{\sqrt{n}\left(c_1 2^{2j-2}-K\right)\delta_n^2} \qquad (4.1.8)\\
&\leq \sum_{j\geq M}\frac{c_2c_3 2^{j\alpha}}{(c_1 2^{2j-2}-K)} \\
&\lesssim C\sum_{j\geq M}2^{j(\alpha-2)}
\end{aligned}$$

It is clear that the above summation goes to zero as $M\to\infty$ since $\alpha<2$. As a conclusion, we showed that $\mathbb{P}\left[d\left(\hat{\theta}_n,\theta_0\right)>2^M\delta_n\right]<3\varepsilon$ by some choose of M and K, which is the desired result. □

Remark 4.1.3. At the beginning, we assumed that $\hat{\theta}_n$ is a M-estimator which maximizes $\mathbb{M}_n(\theta)$. The above theorem works for a more general case that we just request $\mathbb{M}_n\left(\hat{\theta}_n\right)\geq\mathbb{M}_n(\theta_0)-O_{\mathbb{P}}\left(\delta_n^2\right)$ instead of being a global maximizer.

4.2 Existence for upper bound function

Let us back to a practical situation with $X_1,...X_n$ are i.i.d observations. Let the statistical model be an M-estimation with $\mathbb{M}_n(\theta)=\mathbb{P}[m_\theta]$ and $M(\theta)=P[m_\theta]$. The centered and scaled process

$$\sqrt{n}\left(\mathbb{M}_n-M\right)(\theta)=:\mathbb{G}_n\left[m_\theta\right]$$

is exactly the empirical process at m_θ. Let $\mathbb{G}_n$ indexed by the class of function which satisfies the condition 4.1.1:

$$\mathcal{M}_u:=\{m_\theta-m_{\theta_0}:d(\theta,\theta_0)\leq u\}$$

In order to apply the rate of convergence theorem, we first need to find a bound function $\phi_n(\cdot)$ which satisfies 4.1.2, (i.e. $\mathbb{E}\left[\|\mathbb{G}_n\|_{\mathcal{M}_u}\right]\leq\phi_n(u)$) and then construct a series δ which satisfies the condition 4.1.3. If the model is well setted i.e. the estimator $\hat{\theta}_n$ converges to θ_0 in probability, the rate of convergence should only depend on the statistical model we used i.e. the class of estimation functions with constraint candidates $\mathcal{M}_u$. Thus, $\phi_n(\cdot)$ should be a function witch depends on $\mathcal{M}_u$. Then, the problem becomes to find a function of $\mathcal{M}_u$ as an upper bound for $\mathbb{E}\left[\|\mathbb{G}_n\|_{\mathcal{M}_u}\right]$. The maximal inequalities 3.3.4 and 3.3.5 we showed in section 3 give us a direct solution.

Let $M_{n,u}$ be a square integarable and measurable envelop function of class $\mathcal{M}_u$. By applying theorem 3.3.4, then we can set $\phi_n(\cdot)$ as:

$$\mathbb{E}\left[\|\mathbb{G}_n\|_{\mathcal{M}_{n,u}}\right]\lesssim J\left(1,\mathcal{M}_u,M_u\right)\left[P\left(M_{n,u}^2\right)\right]^{1/2}=:\phi_n(u) \qquad (4.2.1)$$

where

$$J\left(1,\mathcal{M}_u,M_{n,u}\right)=\int_0^1\sup_Q\sqrt{\log N\left(\varepsilon\left\|M_{n,u}\right\|_{Q,2},\mathcal{M}_u,L_2(Q)\right)}\,d\varepsilon$$

and

$$P\left(M_{n,u}^2\right) = \mathbb{E}\left[n^{-1}\sum_{i=1}^{n} M_{n,u}^2\left(X_i\right)\right]$$

By applying maximal inequality 3.3.5, we have

$$\mathbb{E}\left[\|\mathbb{G}_n\|_{\mathcal{M}_{n,u}}\right] \lesssim \int_0^u \sqrt{\log N\left(\varepsilon, \mathcal{M}_{n,u}, L_2\left(\mathbb{P}_n\right)\right)}\, d\varepsilon =: \phi_n(u) \tag{4.2.2}$$

For our main problem in Chapter 5, we will use the inequality 4.2.2 to build the upper bound function $\phi_n(\cdot)$.

Chapter 5

Signature Method in Least Square Regression

Feature extraction for time series is one of the preliminary steps for some machine learning problems especially when we work on a financial data set. Various techniques have been developed in recent years. The technique we are interested in this thesis is called path Signature transform which allows describing a path by some tensor object in through iterated integration. We first introduce the idea of path Signature in the following section.

5.1 Introduction to path Signature

Path Signature is first defined in rough path theory. Rough path theory is built to construct to find the solution for controlled differential equations driven by classically irregular signals under the form

$$dY_t = f\left(Y_t\right) dX_t, \quad Y_0 = y_0 \tag{5.1.1}$$

which emphasises the linear dependence of the right-hand side with respect to X_t. This kind of controlled differential equations can also be applied in financial modeling. For example, we may consider a liquidating problem, let X_t be a cash process and Y_t is the price of the underlying asset. We may model the impact on price caused by the selling process as a linear dependence like equation 5.1.1.

The rough path theory was developed in the 1990s by Terry Lyons [16][17][18]. The theory was initially developed to capture and make precise the interactions between highly oscillatory and non-linear systems. It allows a deterministic treatment of SDEs or even controlled differential equations driven by much rougher signals than semi-martingales. The path Signature is a fundamental object in rough path theory. The Signature of a path X_t arises naturally when solving a linear differential equation driven by X_t and it is also a basis to represent a solution to a general controlled differential equation 5.1.1 with smooth vector fields as an analogy to Taylor expansion [1]. Before we define the path Signature, we first recall the definition of bounded variation path.

Definition 5.1.1 (Bounded variation path)**.** [19] Let f be a mapping $f : [a, b] \mapsto \mathbb{R}^d$ where $[a, b] \subset \mathbb{R}$ and d is some positive integer. We define the total variation of f by

$$V_a^b(f) = \sup_{P \in \mathcal{P}} \sum_{i=0}^{n_P - 1} \left| f\left(x_{i+1}\right) - f\left(x_i\right) \right|$$

where $\mathcal{P} = \{P = \{x_0, \ldots, x_{n_P}\} : P$ is a partition of $[a, b]$ satisfying $x_i \leq x_{i+1}$ for $0 \leq i \leq n_P - 1\}$ is the collection of all partitions of the interval $[a, b]$. We say f is bounded variation with the notation $f \in BV\left([a, b]; \mathbb{R}^d\right)$ if its total variation $V_a^b(f)$ is finite.

We use the notation $BV_c\left([a, b]; \mathbb{R}^d\right) = BV\left([a, b]; \mathbb{R}^d\right) \cap C\left([a, b]; \mathbb{R}^d\right)$ for the class of continuous bounded variation functions.

Definition 5.1.2 (Path Signature)**.** [16] Let $\gamma \in BV_c\left([a, b]; \mathbb{R}^d\right)$, the path Signature of γ is defined as a collection of tensors

$$S(\gamma)_{[a,b]} = \left(1, S(\gamma)^1_{a,b}, \ldots, S(\gamma)^k_{a,b}, \ldots\right) \in \prod_{k=0}^{\infty} \left(\mathbb{R}^d\right)^{\otimes k} \tag{5.1.2}$$

where

$$S(\gamma)^k_{[a,b]} := \int_{a<t_1<\cdots<t_k<b} d\gamma_{t_1} \otimes \cdots \otimes d\gamma_{t_k} \in \left(\mathbb{R}^d\right)^{\otimes k} \tag{5.1.3}$$

More precisely, let $i_1, ..., i_k \in \{1, 2, ..., d\}$, then

$$S(\gamma)^{k;i_1 i_2 \ldots i_k}_{[a,b]} = \int_{a\leq t_1\leq t_2\leq \ldots \leq t_k\leq b} d\gamma^{i_1}_{t_1} d\gamma^{i_2}_{t_2} \ldots d\gamma^{i_k}_{t_{k_k}} \tag{5.1.4}$$

In piratical problem, another important object that we use is the truncated Signature.

Definition 5.1.3 (Truncated Signature)**.** Let $\gamma \in BV_c\left([a, b]; \mathbb{R}^d\right)$, for a positive integer N, we denote the N order truncated Signature $S^N(\gamma)_{[a,b]}$ which is defined as

$$S^N(\gamma)_{[a,b]} = \left(1, S(\gamma)^1_{a,b}, \ldots, S(\gamma)^N_{a,b})\right) \in \prod_{k=0}^{N} \left(\mathbb{R}^d\right)^{\otimes N} \tag{5.1.5}$$

We can see that this definition make sense since $\gamma \in BV_c\left([a, b]; \mathbb{R}^d\right)$ so that the iterated integral 5.1.4 is always well defined. The path Signature may seem a complex object, a nature problem is that why we describe a path by its Signature? A remarkable fact of Signature discovered by Ben Hambly and Terry Lyons in 2010 [20] is that the mapping $\gamma \mapsto S(\gamma)_{[a,b]}$ is an injection under the sense of tree-like equivalence. More precisely, we have following two theorems.

Theorem 5.1.4 (Uniqueness of Signature)**.** *Let $\gamma \in BV_c\left([a, b]; \mathbb{R}^d\right)$ Then $S(\gamma)_{[a,b]}$ determines γ up to the tree-like equivalence. [20]*

Definition 5.1.5 (Stratonovich Signature of Brownian motion)**.** [18] Let $B = \{B_t\}_{t\in[0,T]}$ be a Brownian motion in $\mathbb{R}^d$. Let $\circ$ denote the Stratonovich integration. Then the Stratonovich Signature of Brownian motion is defined as

$$S(B)_{[0,T]} := \left(1, \int_{0<t_1<T} \circ \mathrm{d}B_{t_1}, \int_{0<t_1<t_2<T} \circ \mathrm{d}B_{t_1} \otimes \circ \mathrm{d}B_{t_2}, \cdots\right) \tag{5.1.6}$$

Theorem 5.1.6 (Uniqueness of Signature of Brownian motion)**.** *Let B denote a standard d-dimensional Brownian motion and $S(B)_{[0,T]}$ denote the Stratonovich Signatures of B up to time T, where $T > 0$ as we defined above. Then all Brownian motion sample paths up to time T are determined by their Signature $S(B)_{[0,T]}$ up to time reparameterization almost surely. [20]*

The above results imply that the countable infinite collections of component of $S(\gamma)_{[a,b]}$ describe the uncountably infinite stream $(\gamma_t)_{t\in[a,b]}$. This inspires us to use the Signature of a path as features set to describe the path itself. However, the path Signature is still an infinite object, we cannot use the whole collection as features for a path in piratical problems. The choice of truncated order is an important problem when we use path Signature as features for a path. In practice, we usually truncate the collection until some order N as a feature set. It is clear that the larger N is, the more information we can extract for a path. However, the space required by Signature increases exponentially with the growth of N. Another potential problem is over-fitting, let us consider a linear regression model that takes truncated Signature as the explanatory variable. If we do not have enough data and we use a large N, then we may actually build an interpolation model instead of a regression model.

5.2 Basic properties of path Signature

The main problem of this thesis is under the framework of least square regression on truncated Signature. In this section, we will introduce some properties of path Signature which supports our model in the next section. We first introduce the concept of reparametrisation for a path Signature.

Definition 5.2.1 (Reparametrisation)**.** [21] Let $\lambda : [a,b] \mapsto [c,d]$ be a continuous increasing function. We call λ a reparametrisation of $[a,b]$ onto $[c,d]$. Let $\gamma \in BV_c\left([a,b];\mathbb{R}^d\right)$ be a continues bounded variation path, the path $\rho := \gamma \circ \lambda \in BV_c\left([c,d];\mathbb{R}^d\right)$ called a reparametrisation of γ.

The first property we want to introduce is the reparametrisation invariant of Signature transform.

Lemma 5.2.2. *[21] Let $\gamma \in BV_c\left([a,b];\mathbb{R}^d\right)$ a continues bounded variation path and $\rho \in BV_c\left([c,d];\mathbb{R}^d\right)$ be any reparametrisation of γ. Then we have*

$$S(\gamma)_{[a,b]} = S(\rho)_{[c,d]}$$

Proof. We use induction to prove the above lemma. Let $t \in [a,b]$, $\lambda : [a,b] \mapsto [c,d]$ be a continuous increasing function, the first order iterated integral satisfies

$$\int_a^t d\gamma_u = \gamma_t - \gamma_a = \rho_{\lambda(t)} - \rho_c = \int_c^{\lambda(t)} d\rho_u, \quad \forall t \in [a,b]$$

Make the induction hypothesis that for any word $w = i_1 i_2 ... i_m$ of length $m \leq k$ with $i_j \in \{1,...,d\}$, the following equation holds.

$$S(\gamma)^{j;w}_{[a,t]} = S(\rho)^{j;w}_{[c,\lambda(t)]} \text{ for all } t \text{ in } [a,b]$$

Let us consider the word $w' = i_1 i_2 ... i_k i = wi$ with length $k+1$ and $i \in \{1,...,d\}$, we have

$$S(\gamma)^{k+1;w'}_{[a,t]} = \int_a^t S(\gamma)^{k;w}_{[a,u]} d\gamma^i_u = \int_a^t S(\rho)^{k;w}_{[c,\lambda(u)]} d\rho^i_{\lambda(u)} = \int_c^{\lambda(t)} S(\rho)^{k;w}_{[c,u]} d\rho^i_u = S(\rho)^{k+1;wi}_{[c,\lambda(t)]} = S(\rho)^{k+1;w'}_{[c,\lambda(t)]}$$

which completes the proof. □

In some piratical machine learning problems (e.g. handwriting recognition), we may need to consider the operation of the concatenation for paths. The next result is Chen's formula which tells us that the Signature of the concatenation path is the tensor product of the respective Signatures. We first define the concatenation of two paths in a mathematical way.

Definition 5.2.3 (Concatenation)**.** [21] If we have two path $\gamma \in BV_c\left([a,b];\mathbb{R}^d\right)$ and $\rho \in BV_c\left([b,c];\mathbb{R}^d\right)$ then we can define the concatenating path with the notation $\gamma * \rho \in BV_c\left([a,c];\mathbb{R}^d\right)$

$$(\gamma * \rho)(t) := \begin{cases} \gamma(t) & \text{if } t \in [a,b] \\ \rho(t) - \rho(b) + \gamma(b) & \text{if } t \in [b,c] \end{cases} \tag{5.2.1}$$

Then we introduce Chen's theorem

Theorem 5.2.4 (Chen's theorem)**.** *[22] Let $\gamma \in BV_c\left([a,b];\mathbb{R}^d\right)$ and $\rho \in BV_c\left([b,c];\mathbb{R}^d\right)$ be two continuous and bounded variation paths, then the following identity holds*

$$S(\gamma * \rho)_{[a,c]} = S(\gamma)_{[a,b]} \otimes S(\rho)_{[b,c]} \tag{5.2.2}$$

Proof. The proof is a direct application of Fubini's theorem. Let us denote $\zeta = \gamma * \rho$, $t_0 = a$, $t_{k+1} = b$, then we split the iterated integral by $[a,b]$ and $[b,c]$ where $b \in [t_j, t_{j+1}]$ for $j = 0, 1, ..., k$

$$\begin{aligned} S(\zeta)^k_{[a,c]} &= \int_{a \le t_1 \le t_2 \le \ldots \le t_k \le c} \dot{\zeta}_{t_1} \otimes \ldots \otimes \dot{\zeta}_{t_k} dt_1 \ldots dt_k \\ &= \sum_{j=0}^{k} \int_{a \le t_1 \le t_2 \le \ldots \le t_j \le b \le t_{j+1} \le \ldots \le t_k \le c} \dot{\gamma}_{t_1} \otimes \ldots \otimes \dot{\gamma}_{t_j} \otimes \dot{\rho}_{t_{j+1}} \otimes \ldots \otimes \dot{\rho}_{t_k} dt_1 \ldots dt_j dt_{j+1} \ldots dt_k \\ &= \sum_{j=0}^{k} \int_{a \le t_1 \le t_2 \le \ldots \le t_j \le b} \dot{\gamma}_{t_1} \otimes \ldots \otimes \dot{\gamma}_{t_j} dt_1 \ldots dt_j \otimes \int_{b \le t_{j+1} \le \ldots \le t_k \le c} \dot{\rho}_{t_{j+1}} \otimes \ldots \otimes \dot{\rho}_{t_k} dt_{j+1} \ldots dt_k \\ &= \sum_{j=0}^{k} S(\gamma)^j_{[a,b]} \otimes S(\rho)^{k-j}_{[b,c]} \end{aligned} \tag{5.2.3}$$

□

Remark 5.2.5. The above theorem gives us a multiplicative property of the Signature. Let us consider a continuous bounded variation path γ defined on $[a,b]$, for any $c \in [a,b]$, the path itself we can be regarded as the concatenation $\gamma = \gamma|_{[a,c]} * \gamma|_{[c,b]}$ for any $c \in [a,b]$. Then by Chen's theorem, we have $S(\gamma)_{[a,b]} = S(\gamma)_{[a,c]} \otimes S(\gamma)_{[c,b]}$.

Now we would like to discuss another important algebraic property of Signature which is called shuffle products. We first define the (n,m)-shuffle.

Definition 5.2.6. An (n,m)-shuffle is a permutation σ of the set $\{1, 2, ..., n+m\}$ such that

$$\sigma(1) < \sigma(2) < \ldots < \sigma(n) \text{ and } \sigma(n+1) < \sigma(n+2) < \ldots < \sigma(n+m)$$

we use the notation $Sh(n,m)$ as the set of all (n,m)-shuffles.

Lemma 5.2.7. *[21] Let $\gamma \in BV_c\left([a,b];\mathbb{R}^d\right)$ be a continuous bounded path, let $u = i_1...i_n$ and $v = j_1...j_n$ be two words,then*

$$S(\gamma)^{n;u}_{[a,b]} S(\gamma)^{m;v}_{[a,b]} = \sum_{\sigma \in Sh(n,m)} S(\gamma)^{n+m;\sigma(uv)}_{[a,b]} \tag{5.2.4}$$

Proof. We just need develop the left hand side by Fubini's theorem

$$\begin{aligned}
S(\gamma)^{n;u}_{a,b}S(\gamma)^{m;v}_{a,b} &= \int_{a\leq t_1\leq\cdots\leq t_n\leq b} d\gamma^{i_1}_{t_1}d\gamma^{i_2}_{t_2}\dots d\gamma^{i_n}_{t_n} \cdot \int_{a\leq s_1\leq\ldots\leq s_m b} d\gamma^{j_1}_{s_1}d\gamma^{j_2}_{s_2}\dots d\gamma^{j_m}_{s_m} \\
&= \int_{a\leq t_1\leq\ldots\leq t_n\leq s_1\leq\ldots\leq s_m\leq b} d\gamma^{i_1}_{t_1}d\gamma^{i_2}_{t_2}\dots d\gamma^{i_n}_{t_n}d\gamma^{j_1}_{s_1}\dots d\gamma^{j_m}_{s_m} \\
&+ \int_{a\leq t_1\leq\ldots\leq t_{n-1}\leq s_1\leq t_n\ldots\leq s_m\leq b} d\gamma^{i_1}_{t_1}d\gamma^{i_2}_{t_2}\dots d\gamma^{j_1}_{s_1}d\gamma^{i_n}_{t_n}\dots d\gamma^{j_m}_{s_m} \\
&+\cdots \\
&+ \int_{a\leq s_1\leq\ldots\leq s_m\leq t_1\leq\ldots\leq t_n\leq b} d\gamma^{j_1}_{s_1}d\gamma^{j_2}_{s_2}\dots d\gamma^{j_m}_{s_m}d\gamma^{i_1}_{t_1}\dots d\gamma^{i_n}_{t_n} \\
&= \sum_{\sigma\in Sh(n,m)} S(\gamma)^{n+m;\sigma(uv)}_{[a,b]}
\end{aligned}$$

□

The shuffle products property shows that the product of two truncated Signatures can be represented as a summation of higher order Signatures. The above property also motivates the definition of the following shuffle product on the dual tensor space $(T(\mathbb{R}^d))^* = T((\mathbb{R}^d)^*)$. Let $u = i_1, ..., i_n$ and $v = j_1, ..., j_m$ be two words and $e^*_u,\ e^*_v \in T((\mathbb{R}^d)^*)$, then we define the following notation:

$$e^*_u \sqcup\!\sqcup\, e^*_v = \sum_{\sigma\in Sh(n,m)} e^*_{\sigma^{-1}(uv)} \tag{5.2.5}$$

Remark 5.2.8. The usage of σ^{-1} may seem odd at first glance, let us see an example to see how it works. Let $\sigma \in Sh(n,m)$, denote $k := n+m$. Let $u = i_1...i_k$ where $i_j \in \{1,2,...,d\}$ for $j \in \{1,2,...,k\}$ be a word and $x = x_1 \otimes ... \otimes x_k$ be a tensor in $\left(\mathbb{R}^d\right)^{\otimes k}$, then we have

$$\langle e^*_{\sigma^{-1}u}, x\rangle = x_1^{i_{\sigma^{-1}(1)}}\dots x_n^{i_{\sigma^{-1}(n)}} = x^{i_1}_{\sigma(1)}\dots x^{i_n}_{\sigma(n)} = (e^*_u, \sigma x)$$

where $\sigma x := x_{\sigma(1)} \otimes \ldots \otimes x_{\sigma(n)}$.

With the above notation, we can express the product of two truncated Signatures (i.e. the left hand side of equation 5.2.4) as a linear functional of the Signature.

Proposition 5.2.9. *[21] Let u and v be two finite words, then we have*

$$\left\langle e^*_u, S(\gamma)_{[a,b]}\right\rangle \cdot \left\langle e^*_v, S(\gamma)_{[a,b]}\right\rangle = \left\langle e^*_u \sqcup\!\sqcup\, e^*_v, S(\gamma)_{[a,b]}\right\rangle \tag{5.2.6}$$

Proof. The proof is just a direct computation using lemma 5.2.7.

$$\begin{aligned}
\left\langle e^*_u, S(\gamma)_{[a,b]}\right\rangle \cdot \left\langle e^*_v, S(\gamma)_{[a,b]}\right\rangle &= S(\gamma)^{n;u}_{[a,b]}S(\gamma)^{m;v}_{[a,b]} \\
&= \sum_{\sigma\in \mathrm{Sh}(n,m)} S(\gamma)^{\sigma(uv)}_{[a,b]} \\
&= \left\langle e^*_u \sqcup\!\sqcup\, e^*_v, S(\gamma)_{[a,b]}\right\rangle
\end{aligned} \tag{5.2.7}$$

□

The last property we would like to introduce ensures the consistence of Signature approximation. We first need the concept of group-like elements.

Definition 5.2.10 (Group-like elements)**.** [21] Let $x \in T\left(\mathbb{R}^d\right)$ be a tensor, we say x is group-like if for every n, the canonical projection of x to $T^n\left(\mathbb{R}^d\right)$ belongs to $G^n\left(\mathbb{R}^d\right)$ where $G^n\left(\mathbb{R}^d\right)$ is step-n nilpotent Lie group with d generators. We denote the collection of group-like elements as G^*.

The above definition is given by a very theoretical way with Lie group, an equivalent definition can be stated as follows. We say a continuous bounded variation path γ in $\mathbb{R}^d$ is group like if for any truncated order N, there exist another path γ' such that at least the first N terms of its Signature are equal to γ's.
With the Stone-Weierstrass Theorem, we have the following Signature approximation theorem:

Theorem 5.2.11. *[21] Let $K \subset G^*$ be a compact subset, let $A = \left\{ L|_{G^*} : L \in T\left(\mathbb{R}^d\right)^* \right\}$ be a set of linear functionals. Then A is dense in $C(K)$ with respect to the uniform topology. Equivalently, for any continuous function f and any $\varepsilon > 0$, there exists $L \in T\left(\mathbb{R}^d\right)^*$ such that*

$$\sup_{x \in K} |L(x) - f(x)| < \epsilon$$

In order to prove this result, we first recall the Stone-Weierstrass Theorem.

Theorem 5.2.12 (Stone-Weierstrass Theorem)**.** *[23] Suppose X is a compact Hausdorff space and A is a subalgebra of $C(X)$ which contains a non-zero constant function. Then A is dense in $C(X)$ with the uniform topology if and only if it separates points.*

proof of Theorem 5.2.11. We first show that A is an algebra. Let L_1, $L_2 \in T\left(\mathbb{R}^d\right)^*$ with the form:

$$L_i = \sum_{u \in F_{L_i}} \lambda_u^i e_u^*$$

where F_{L_i} is a finite set of words. It is clear that for any λ, $\nu \in \mathbb{R}$, we have $\lambda L_1 + \nu L_2 \in T\left(\mathbb{R}^d\right)^*$. Let $N = \max\{|u| + |v| : u \in F_{L_1}, v \in F_{L_2}\}$, let $x \in G^*$, then there exists some paths $\gamma \in BV_c\left([a,b]; \mathbb{R}^d\right)$ such that at least the first N terms in the tensor series of $S(\gamma)_{[a,b]}$ agree with those of x. We apply the shuffle products property on the product of L_1 and L_2:

$$\begin{aligned}
L_1(x)L_2(x) &= L_1(S(\gamma)_{[a,b]})L_2(S(\gamma)_{[a,b]}) \\
&= \sum_{u \in F_{L_1}, v \in F_{L_2}} \lambda_u^1 \lambda_v^2 \left\langle e_u^*, S(\gamma)_{[a,b]} \right\rangle \cdot \left\langle e_v^*, S(\gamma)_{[a,b]} \right\rangle \\
&= \sum_{u \in F_{L_1}, v \in F_{L_2}} \lambda_u^1 \lambda_v^2 \left\langle e_u^* \sqcup\!\sqcup\, e_v^*, S(\gamma)_{[a,b]} \right\rangle \\
&= \sum_{u \in F_{L_1}, v \in F_{L_2}} \lambda_u^1 \lambda_v^2 \left\langle e_u^* \sqcup\!\sqcup\, e_v^*, S(\gamma)_{[a,b]} \right\rangle \\
&= L_1 \sqcup\!\sqcup\, L_2(x)
\end{aligned} \tag{5.2.8}$$

The above equation shows that the product of L_1 and L_2 is still a linear functional i.e. $L_1 L_1 \in A$. Hence, A is an algebra. It is also clear that A contains constant functionals (e.g. $\mathbf{1} : x \mapsto x^0$) , we still have one condition to verify. In order to see that it separates points we suppose that x and y are two distinct tensor of G^*, then it is clear that there exist some n such that $0 \neq x^n \neq y^n$. We assume that $\|y^n\| \leq \|x^n\|$, we define a linear functional $L \in A$ such that $L(\cdot) := \langle \pi_n(\cdot), x^n \rangle$, we notice that

$$|L(y)| = |\langle y^n, x^n \rangle| \leq \langle x^n, x^n \rangle = |L(x^n)|$$

This equality is strict unless $y^n = \pm x^n$, since we have assumed $x^n \neq y^n$, we either have strict inequality or $x^n = -y^n$. It is clear that L separates x and y in both two cases. Then by applying theorem 5.2.12, A is dense in $C(K)$ with respect to uniform topology. □

The above result is remarkable since it indicates that any real-valued continuous function on a compact subspace of the range of the Signature can be uniformly well-approximated by a continuous linear functional. This result is also the theoretical support for our modelling in the next section.

5.3 Learn ODE with truncated Signature

In this section, we would like to state the setting for the main problem of this thesis. We would like to study the controlled differential equations driven by Lipschitz's continuous control from the perspective of least square regression. More precisely, we would like to find the approximation value for the endpoint $Y_{t=1}$ for the equation

$$dY_t = V\left(Y_t\right) dX_t, \quad Y_{t=0} = y_0, \quad \left\{X_t\right\}_{t\in[0,1]} \in Lip\left([0,1];\mathbb{R}^d\right) \tag{5.3.1}$$

We make the assumption that $V(\cdot)$ is a Lipschitz vector field, then the equation 5.3.1 has a unique solution on $[0,1]$. Let $\tilde{f} : Lip([0,1];\mathbb{R}^d) \mapsto \mathbb{R}^e$ be the mapping that describes the relation between the input control process $\{X_t\}_{t\in[0,1]}$ and the final state of the output $Y_{t=1}$ (i.e. the functional $\tilde{f}$ solves the equation 5.3.1). By the theorem 5.1.4, we have the uniqueness of Signature under the sense of tree-like equivalence, then it is reasonable to expect we can express Y by the Signature of X. More precisely, we expect there exists some mapping $f : \prod_{k=0}^{\infty}\left(\mathbb{R}^d\right)^{\otimes k} \mapsto \mathbb{R}^e$ such that

$$Y_{t=1} = \tilde{f}\left(\left\{X_t\right\}_{t\in[0,1]}\right) = f\left(S(X)_{[0,1]}\right) \tag{5.3.2}$$

Moreover, we assume that $f(\cdot)$ is a continuous function on a compact subspace of the range of the Signature, then by the theorem 5.2.11, ic can be uniformly well-approximated by a continuous linear functional. Thus, we would like to approximate the final state $Y_{t=1}$ by:

$$Y_{t=1} = \tilde{f}\left(\left\{X_t\right\}_{t\in[0,1]}\right) = f\left(S(X)_{[0,1]}\right) \simeq \theta\left(S\left(X\right)_{[0,1]}\right) \tag{5.3.3}$$

Where $S\left(X\right)_{[0,1]}$ is the N order truncated Signature of $\{X_t\}_{t\in[0,1]}$ and θ is a linear functional which is belong to the candidates space

$$\Theta^N := \left\{\theta \in \Theta | \theta(X) = a_0 + \sum_{k=1}^{N}\sum_{i=1}^{e}\left\langle a_k^i, S(X)_{[0,1]}^k\right\rangle \nu_i, \quad a_0 \in \mathbb{R}^e, \ a_k^i \in \left(\left(\mathbb{R}^d\right)^{\otimes k}\right)^*\right\} \tag{5.3.4}$$

Where $(\nu_i)_{i=1,\dots,e}$ is the basis of $\mathbb{R}^e$.
By theorem 5.2.11, it is clear that when $N \to \infty$, the limit of Θ^N is dense in Θ which is defined as

$$\Theta := \left\{\theta | \theta : Lip([0,1];\mathbb{R}^d) \mapsto \mathbb{R}^e\right\}$$

which is the class contains all candidate functions for $\tilde{f}$.
In practice, we may need to choose one set Θ^N in the increasing sequence $\Theta^i \subset \Theta^{i+1} \subset \cdots \subset \Theta$ such that we can find a good predictor $\theta \in \Theta^N$ which approximates well the exact solution $\tilde{f}$. In another word, we may need a proper N such that we can approximate $Y_{t=1}$ by the linear combination for all components in N truncated Signature of $\{X_t\}_{t\in[0,1]}$

$$Y_{t=1} \simeq \hat{Y}_{t=1} := \sum_{k=1}^{N}\sum_{i=1}^{e}\sum_{j_1\cdots j_n\in\{1,\cdots,e\}} a_{j_1,\cdots,j_n}^{k,i} S(X)_{[0,1]}^{k;j_1,\cdots,j_n}\nu_i, \quad a_{j_1,\cdots,j_n}^{k,i} \in \mathbb{R} \tag{5.3.5}$$

In order to build the model, we assume that we have collected a training dataset $(X_i;Y_i)_{i=1,\dots,n}$ where the input $X_i \in Lip\left([0,1];\mathbb{R}^d\right)$ represents some Lipschitz continues control and the output $Y_i \in \mathbb{R}^e$ is the final state under the control X_i. With a fixed N, we first computed the N order truncated Signatures for every x_i in the dataset. Then we would like to fit the coefficients $a_{j_1,\cdots,j_n}^{k,i} \in \mathbb{R}$ in equation 5.3.5 by least square method with explanatory variable $\left(S^N(x_i)\right)_{i=1,2,\dots,n}$ and target variable $(Y_i)_{i=1,\dots,n}$. We have simply analysed the effect of truncated order N in previous sections. Here, with the point of view of Theorem 5.2.11, it is clear that the larger N can provide a more accurate approximation. For our problem, we assume that we have enough space to store all the truncated Signature. However, it does not mean that a large N is a good idea. From equation 5.3.5, we can observe that we need to fit $\frac{d^{N+1}-1}{d-1}$ parameters for a fixed N which increases exponentially with the growth of N. If we do not have enough data to fit those parameters, then we will get an over fitting model. In a word, the challenge here is to get a balance between the accuracy of the model and over fitting risk.

The goal of this thesis to give a theoretical result which explains how to find a proper truncated order N for a given dataset with size n, i.e. to find some mappings $I : \mathbb{N}_+ \mapsto \mathbb{N}_+$ which help us decide the truncated order N by $N = I(n)$.

5.4 Truncated order decision by empirical process theory

In this section, we will introduce the solution for the main problem of this thesis which is the original work of the author. Before we state the solution, we need an extended version for the rate of convergence theorem 4.1.2.

5.4.1 Extensional rate of convergence theorem

In this section, we would like to extend the rate of convergence theorem. For the following theorem, we will consider an increasing sequence of sets $\{\Theta_n\}_{n=1,2,...}$ as the candidates set. More precisely, we now consider the stochastic process $\{\mathbb{M}_n(\theta) : \theta \in \Theta_n\}$ rather than the process indexed by a fixed index set Θ. We suppose there exists a "true predictor" $\theta_{n,0}$ for any fixed n. For the generality, we do not need assume that Θ_n is a metric space, but we assume some mapping to $d_n(\cdot, \theta_{n,0}) : \Theta_n \mapsto [0, \infty)$ to measure the "discrepancy" between $\theta \in \Theta_n$ and the "true predictor" $\theta_{n,0}$. The extensional rate of convergence theorem [2, Page 57, Theorem 6.1] states the following idea.
Like theorem 4.1.2, we assume that our model is properly constructed, i.e. the stochastic process $\{\mathbb{M}_n(\theta) : \theta \in \Theta_n\}$, the determined process $\{M_n(\theta) : \theta \in \Theta_n\}$ and the "discrepancy" function $d_n(\cdot, \theta_{n,0})$ is appropriately chosen such that for any fixed n, we have some $\tilde{\delta_n}$ which represents the smallest approximation error, for some $\delta > \tilde{\delta_n}$ we may expect $M_n(\theta)$ and $M_n(\theta_{n,0})$ are δ-close when the "discrepancy" between θ and $\theta_{n,0}$ are in $(\delta/2, \delta)$. In the other word

$$\sup_{\theta \in \Theta_n : \delta/2 < d_n(\theta, \theta_{n,0}) \leq \delta} [M_n(\theta) - M_n(\theta_{n,0})] \leq -c_1 \delta^2$$

. Then, we construct the upper bond function $\phi_n(\cdot)$ like equation 4.1.2 with the dynamic candidates set Θ_n and "true predictor" $\theta_{n,0}$. Moreover, we try to find the sequence δ which satisfied the conditions we have in theorem 4.1.2, then we may build the convergence rate

$$d_n\left(\hat{\theta}_n, \theta_{n,0}\right) = O_{\mathbb{P}}(\delta_n).$$

We show the complete theorem as following

Theorem 5.4.1 (Rate of convergence). *[2, Page 57, Theorem 6.1] We keep the background and notation of theorem 4.1.2. For each n, let $\mathbb{M}_n$ be a stochastic processes indexed by a set $\Theta_n \cup \theta_{n,0}$, and M_n a deterministic process indexed by the same set. Let $d_n(\cdot, \theta_{n,0}) : \Theta_n \mapsto [0, \infty)$ be a mapping to measure the difference between θ and $\theta_{n,0}$. Let $\tilde{\delta_n} > 0$ and suppose that for every n and $\delta > \tilde{\delta_n}$*

$$\sup_{\theta \in \Theta_n : \delta/2 < d_n(\theta, \theta_{n,0}) \leq \delta} [M_n(\theta) - M_n(\theta_{n,0})] \leq -c_1 \delta^2 \tag{5.4.1}$$

and

$$\mathbb{E}\left[\sup_{\theta\in\Theta_n:d_n(\theta,\theta_{n,0})\leq\delta}\sqrt{n}\left|\left(\mathbb{M}_n-M_n\right)(\theta)-\left(\mathbb{M}_n-M_n\right)\left(\theta_{n,0}\right)\right|\right]\leq c_2\phi_n(\delta) \tag{5.4.2}$$

for increasing functions $\phi_n:\left[\tilde{\delta}_n,\infty\right)\to\mathbb{R}$ such that $\delta\mapsto\phi_n(\delta)/\delta^{\alpha}$ is decreasing for some $\alpha<2$. Let $\theta_n\in\Theta_n$ and let δ_n satisfy

$$\phi_n\left(\delta_n\right)\leq\sqrt{n}\delta_n^2,\quad \delta_n^2\geq M_n\left(\theta_{n,0}\right)-M_n\left(\theta_n\right),\quad \delta_n\geq\tilde{\delta}_n \tag{5.4.3}$$

If the sequence $\hat{\theta}_n$ takes values in Θ_n and satisfies $\mathbb{M}_n\left(\hat{\theta}_n\right)\geq\mathbb{M}_n\left(\theta_n\right)-O_{\mathbb{P}}\left(\delta_n^2\right)$, then we have

$$d_n\left(\hat{\theta}_n,\theta_{n,0}\right)=O_{\mathbb{P}}\left(\delta_n\right) \tag{5.4.4}$$

The proof of this theorem is just a notational change for the proof 4.1. Roughly speaking, the philosophy of the theorem can be concluded as following. We assume that the model is properly constructed under the sense that condition 5.4.1 is satisfied. More precisely, for a fixed n, let $\tilde{\delta}_n$ be the "best approximation error", our model has at least the ability to get an estimator δ-close to the "true" predictor $\theta_{n,0}$ where δ is some constant larger than $\tilde{\delta}_n$. With the increasing of n, we expect that our estimator $\hat{\theta}_n$ can be closer to the "true" predictor $\theta_{n,0}$ under the sense that $d_n\left(\hat{\theta}_n,\theta_{n,0}\right)=O_{\mathbb{P}}\left(\delta_n\right)$ where δ_n is a decreasing sequence which is smaller than δ but large than $\tilde{\delta}_n$ ($\delta>\delta_n\geq\tilde{\delta}_n$) and towards 0 as $n\to\infty$. In order to get this convergence rate δ_n we first need to build the upper bound function $\phi_n(\cdot)$ with the form of 5.4.2. Then, the sequence can be constructed in a way that satisfies condition 5.4.3.

5.4.2 Suggested truncated order for least square regression model

In this section, we will state our solution for the truncated order problem that we raised at the begging of this Chapter. **To the best knowledge of the author, this solution is first proposed in this thesis. The whole solution is the completely original work of the author**. The solution is divided into 5 steps. Step 1 aims to model the problem in a way that corresponds to the framework of the Theorem 5.4.1. In step 2, we will show that our model is properly constructed under the sense of condition 5.4.1. For step 3, we will apply the maximal inequalities we have shown in Chapter 3 to build the upper bound function $\phi_n(\cdot)$ which satisfies the condition 5.4.2. Step 4, the convergence rate sequence δ_n will be found by verifying the 5.4.3. In the end, we will analyze the relation between the convergence rate and the truncated order and build the decision function $I:\mathbb{N}_+\mapsto\mathbb{N}_+$.

Step 1

First, for a fixed dataset size n, we would like to build a least square regression model with N truncated Signature as explanatory variable. With the assumption we made in section 5.3, we assume that we have regression model (for the simplicity, we denote $S(X)$ for $S(X)_{[0,1]}$):

$$Y_i=\theta_0\left(S\left(X_i\right)\right)+\epsilon_i,\qquad \text{for}\qquad i=1,\cdots,n$$

where $Y_i\in R^e$ is the observed response variable (i.e. solution computed by finite increments method), $X_i\in R^{d\times m}$ is the corresponding control path where m is the number of the grid we make on $[0,1]$, and ϵ is the error of approximation. The errors are assumed to be independent identical distributed sub-Gaussian random variables with zero expectation and finite variation. The function θ_0 is unknown, but we assume that $\theta_0\in\Theta:=\left\{\theta|\theta:Lip([0,1];\mathbb{R}^d)\mapsto\mathbb{R}^e\right\}$. This assumption is reasonable because we expect that the solution of equation 5.3.1 can be expressed

as some continuous function of Signature, then Theorem 5.2.11 indicates that it can be arbitrarily approximated by a continuous linear functional with uniform topology.
As we mentioned in previous sections, we expect there exists some mapping $I : \mathbb{N}_+ \mapsto \mathbb{N}_+$ which logically decides the truncated order $N = I(n)$. Hence, for a fixed n, we would like to find the "true" predictor in the class Θ_n which is determined by the decision function I, i.e.

$$\Theta_n := \Theta^{I(n)} = \Theta^N \tag{5.4.5}$$

where Θ^N is the class of regressor defined as equation 5.3.4.
Then, we may determine the process M_n and build the stochastic process $\mathbb{M}_n$ by the least square model. More precisely,

$$\begin{aligned}
\hat{\theta}_n &= \arg\min_{\theta\in\Theta_n} \frac{1}{n}\sum_{i=1}^{n} \|Y_i - \theta\left(S\left(X_i\right)\right)\|_2^2 \\
&= \arg\min_{\theta\in\Theta_n} \frac{1}{n}\sum_{i=1}^{n} \|\left(\theta_{n,0}(S\left(X_i\right)) - \theta(S\left(X_i\right)) + \epsilon_i\right\|_2^2 \\
&= \arg\min_{\theta\in\Theta_n} \frac{1}{n}\sum_{i=1}^{n} \left\{\|\left(\theta_{n,0} - \theta\right)\left(S\left(X_i\right)\right)\|_2^2 + 2\left\langle\left(\theta_{n,0} - \theta\right)\left(S\left(X_i\right)\right), \epsilon_i\right\rangle + \|\epsilon_i\|_2^2\right\}
\end{aligned} \tag{5.4.6}$$

which is equivalent to:

$$\hat{\theta}_n = \arg\max_{\theta\in\Theta_n} \frac{2}{n}\sum_{i=1}^{n} \left\langle\left(\theta - \theta_{n,0}\right)\left(S\left(X_i\right)\right), \epsilon_i\right\rangle - \mathbb{P}_n\left(\|\theta - \theta_{n,0}\|_2^2\right) =: \arg\max_{\theta\in\Theta_n} \mathbb{M}_n(\theta) \tag{5.4.7}$$

where $\mathbb{P}_n$ is the empirical measure we defined as 2.1.3. We have assumed that $\mathbb{E}[\epsilon_i] = 0$, then we consider the process $M_n(\theta)$ which is defined as the expectation of $\mathbb{M}_n(\theta)$, i.e.

$$M_n\left(\theta\right) := \mathbb{E}\left[\mathbb{M}_n\left(\theta\right)\right] = -\mathbb{P}_n\left(\|\theta - \theta_{n,0}\|_2^2\right) \tag{5.4.8}$$

Moreover, we may choose the empirical measure as the "discrepancy" between two estimators. In the other word, for any $\theta \in \Theta_n$, we define

$$M_n\left(\theta\right) - M_n\left(\theta_{n,0}\right) = -\mathbb{P}_n\left(\|\theta - \theta_{n,0}\|_2^2\right) =: -d_n^2\left(\theta, \theta_{n,0}\right) \tag{5.4.9}$$

With the above construction, it is clear that we have

$$\sup_{\theta\in\Theta_n:\delta/2<d_n(\theta,\theta_{n,0})\leq\delta} \left[M_n(\theta) - M_n\left(\theta_{n,0}\right)\right] = \sup_{\theta\in\Theta_n:\delta/2<d_n(\theta,\theta_{n,0})\leq\delta} -d_n^2\left(\theta, \theta_{n,0}\right) \leq -\frac{1}{4}\delta^2 \tag{5.4.10}$$

which implies that the condition 5.4.1 is satisfied.

Step 2

With our choice of $\mathbb{M}_n$ and M_n, we have

$$\left|\left(\mathbb{M}_n - M_n\right)\left(\theta\right) - \left(\mathbb{M}_n - M_n\right)\left(\theta_{n,0}\right)\right| = \frac{2}{\sqrt{n}}\left|\sum_{i=1}^{n} \left\langle\left(\theta - \theta_{n,0}\right)\left(S\left(X_i\right)\right), \epsilon_i\right\rangle\right|$$

Follow the Theorem 5.4.1, we would like to find some bounded function $\phi_n(\cdot)$ such that

$$\phi_n(\delta) \geq \mathbb{E}\left[\sup_{\mathbb{P}_n(\theta-\theta_{n,0})^2\leq\delta^2,\theta\in\Theta_n} \left|\frac{1}{\sqrt{n}}\left\langle\left(\theta - \theta_{n,0}\right)\left(S\left(X_i\right)\right), \epsilon_i\right\rangle\right|\right] \tag{5.4.11}$$

As we have assumed that the error ϵ_i is some sub-Gaussian random variable, then the process $G_n(\theta) := \frac{1}{\sqrt{n}} \left\langle (\theta - \theta_{n,0}) \left(S\left(X_i\right)\right), \epsilon_i \right\rangle$ is also a sub-Gaussian process. By maximal inequality 3.3.5 and lemma 2.3.6, we can chose $\phi_n(\cdot)$ by the following inequality:

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\mathbb{P}_n(\|\theta-\theta_{n,0}\|)_2^2 \leq \delta^2, \theta \in \Theta_n} \left| \frac{1}{\sqrt{n}} \left\langle (\theta - \theta_{n,0}) \left(S\left(X_i\right)\right), \epsilon_i \right\rangle \right| \right] \\
&\leq \int_0^\delta \sqrt{\log N\left(\varepsilon, \Theta_n \cap \left\{\theta : \mathbb{P}_n\left(\|\theta - \theta_{n,0}\|_2^2\right) \leq \delta^2\right\}, L_2\left(\mathbb{P}_n\right)\right)} d\varepsilon \\
&\leq \int_0^\delta \sqrt{\log D\left(\varepsilon, \Theta_n \cap \left\{\theta : \mathbb{P}_n\left(\|\theta - \theta_{n,0}\|_2^2\right) \leq \delta^2\right\}, L_2\left(\mathbb{P}_n\right)\right)} d\varepsilon
\end{aligned} \tag{5.4.12}$$

Where $L_2\left(\mathbb{P}_n\right)$ is the semi-norm generated by the empirical measure $\mathbb{P}_n$ which is defined as equation 3.3.6.

In order to compute the above integration, we first need an explicit expression (or an upper bound) for the packing numbers of the set $\Theta_n \cap \left\{\theta : \mathbb{P}_n\left(\|\theta - \theta_{n,0}\|^2\right) \leq \delta^2\right\}$. To solve this problem, we may change our point view of the problem. As Θ_N is defined as a linear function space in 5.3.4, we may flatten the N order truncated tensor to a vector in $\mathbb{R}^{\frac{d^{N+1}-1}{d-1}}$, we denote $\tilde{S}^N\left(X_i\right) \in \mathbb{R}^{\frac{d^{N+1}-1}{d-1}}$ as the flattened vector for N order truncated Signature of X_i. Under this simplification, for any $\theta \in \Theta^N$, we can rearrange the regression coefficients of θ to a matrix $\mathbf{W}_\theta \in \mathbb{R}^{e \times \frac{d^{N+1}-1}{d-1}}$ such that for any input path X_i, we have $\theta\left(S\left(X_i\right)\right) = \mathbf{W}_\theta \cdot \tilde{S}^N(X_i) = \mathbf{W}_\theta \cdot \tilde{S}^{I(n)}(X_i)$. Thus, to determine the ε-packing numbers of the set $\Theta_n \cap \left\{\theta : \mathbb{P}_n\left(\|\theta - \theta_{n,0}\|^2\right) \leq \delta^2\right\}$ is actually equivalent to determine ε-packing numbers of the set

$$\Omega_n := \left\{\mathbf{W}_\theta : \|\mathbf{W}_\theta - \mathbf{W}_{\theta_{n,0}}\|^2_{\tilde{L}^2(n)} \leq \delta^2\right\} \tag{5.4.13}$$

Where $\mathbf{W}_{\theta_{n,0}}$ is the matrix corresponds to $\theta_{n,0}$ and the norm $\|\cdot\|_{\tilde{L}^2(n)}$ is defined as

$$\begin{aligned}
\|\mathbf{W}_\theta\|^2_{\tilde{L}^2(n)} &:= \frac{1}{n}\sum_{i=1}^n \|\mathbf{W}_\theta \cdot \tilde{S}^{I(n)}(X_i)\|_2^2 \\
&= \frac{1}{n}\sum_{i=1}^n \|\theta(S^{I(n)}(X_i))\|_2^2 \\
&= \mathbb{P}_n\left(\|\theta(S^{I(n)}(\cdot))\|_2^2\right) \\
&= \|\theta\|^2_{L_2(\mathbb{P}_n)}
\end{aligned} \tag{5.4.14}$$

for $\theta \in \Theta_n$ and its corresponding matrix $\mathbf{W}_\theta \in \mathbb{R}^{e \times \frac{d^{I(n)+1}-1}{d-1}}$

Remark 5.4.2. In our model, the explanatory variable is not the input $\left\{X_i\right\}_{i=1,2,\ldots,n}$ but its truncated Signature transform $\left\{S^{I(n)}(X_i)\right\}_{i=1,2,\ldots,n}$ which makes $\|\theta\|^2_{L_2(\mathbb{P}_n)} = \mathbb{P}_n\left(\|\theta(S^{I(n)}(\cdot))\|_2^2\right)$

With above construction, we have transferred the packing number problem

$$D\left(\varepsilon, \Theta_n \cap \left\{\theta : \mathbb{P}_n\left(\|\theta - \theta_{n,0}\|_2^2\right) \leq \delta^2\right\}, L_2\left(\mathbb{P}_n\right)\right)$$

by the equivalent problem

$$D\left(\varepsilon, \Omega_n, \|\cdot\|_{\tilde{L}^2(n)}\right)$$

Before we solve this problem, we first check that $\left(D, \|\cdot\|_{\tilde{L}^2(n)}\right)$ is a metric space.

- $\forall\ \mathbf{W}_\theta \in \Omega_n$ it is clear that $\|\mathbf{W}_\theta\|^2_{\tilde{L}^2(n)} = \frac{1}{n}\sum_{i=1}^n \|\mathbf{W}_\theta \cdot \tilde{S}^{I(n)}(X_i)\|_2^2 \geq 0$

- $\forall\ \mathbf{W}_{\theta_1}, \mathbf{W}_{\theta_2} \in \Omega_n$ we have

$$\begin{aligned}
\|\mathbf{W}_{\theta_1} - \mathbf{W}_{\theta_2}\|^2_{\tilde{L}^2(n)} &= \frac{1}{n}\sum_{i=1}^{n} \|(\mathbf{W}_{\theta_1} - \mathbf{W}_{\theta_2}) \cdot \tilde{S}^{I(n)}(X_i)\|_2^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} \|(\mathbf{W}_{\theta_2} - \mathbf{W}_{\theta_1}) \cdot \tilde{S}^{I(n)}(X_i)\|_2^2 \\
&= \|\mathbf{W}_{\theta_2} - \mathbf{W}_{\theta_1}\|^2_{\tilde{L}^2(n)}
\end{aligned}$$

- $\forall\ \mathbf{W}_{\theta_1}, \mathbf{W}_{\theta_2}, \mathbf{W}_{\theta_2} \in \Omega_n$ we have

$$\begin{aligned}
\|\mathbf{W}_{\theta_1} - \mathbf{W}_{\theta_2}\|^2_{\tilde{L}^2(n)} &= \frac{1}{n}\sum_{i=1}^{n} \|(\mathbf{W}_{\theta_1} - \mathbf{W}_{\theta_2}) \cdot \tilde{S}^{I(n)}(X_i)\|_2^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} \|(\mathbf{W}_{\theta_1} - \mathbf{W}_{\theta_3} + \mathbf{W}_{\theta_3} - \mathbf{W}_{\theta_1}) \cdot \tilde{S}^{I(n)}(X_i)\|_2^2 \\
&\le \frac{1}{n}\sum_{i=1}^{n} \|(\mathbf{W}_{\theta_1} - \mathbf{W}_{\theta_3}) \cdot \tilde{S}^{I(n)}(X_i)\|_2^2 + \frac{1}{n}\sum_{i=1}^{n} \|(\mathbf{W}_{\theta_3} - \mathbf{W}_{\theta_2}) \cdot \tilde{S}^{I(n)}(X_i)\|_2^2 \\
&\le \|\mathbf{W}_{\theta_1} - \mathbf{W}_{\theta_3}\|^2_{\tilde{L}^2(n)} + \|\mathbf{W}_{\theta_3} - \mathbf{W}_{\theta_2}\|^2_{\tilde{L}^2(n)}
\end{aligned}$$

Now, we move to the ε-packing numbers. It is clear that Ω_n is a δ-ball with centre $\mathbf{W}_{\theta_{n,0}}$ with respect to the semi-norm $\|\cdot\|_{\tilde{L}^2(n)}$. Let $\{\mathbf{W}_1, ..., \mathbf{W}_D\}$ be a ε-packing of Ω, then the balls of radius $\frac{\varepsilon}{2}$ around the $\mathbf{W}_i$ are disjoint, and their union is contained in the set Ω'_n:

$$\Omega'_n := \left\{\mathbf{W} \in \mathbb{R}^{e\times\frac{d^{N+1}-1}{d-1}} : \|\mathbf{W} - \Omega\|_{\tilde{L}^2(n)} < \varepsilon/2\right\} \tag{5.4.15}$$

which is a ball with radius $\delta + \frac{\varepsilon}{2}$ and centre $\mathbf{W}^{n,0}$ with respect to $\|\cdot\|_{\tilde{L}^2(n)}$. Hence, the sum of the volumes of these balls is bounded by the volume of Ω'. Let V_u be the volume of unit ball with respect to the semi-norm $\|\cdot\|_{\tilde{L}^2(n)}$, we can build the following inequality

$$D\left(\varepsilon, \Omega_n, \|\cdot\|_{\tilde{L}^2(n)}\right) \le D\left(\varepsilon, \Omega'_n, \|\cdot\|_{\tilde{L}^2(n)}\right) \le \frac{V_u \cdot (\delta + \frac{\varepsilon}{2})^{e\cdot\frac{d^{I(n)+1}-1}{d-1}}}{V_u \cdot \left(\frac{\varepsilon}{2}\right)^{e\cdot\frac{d^{I(n)+1}-1}{d-1}}} = \left(1 + \frac{2\delta}{\varepsilon}\right)^{e\cdot\frac{d^{I(n)+1}-1}{d-1}} \tag{5.4.16}$$

Combine the results of 5.4.12, 5.4.13 and 5.4.16, we have

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\mathbb{P}_n(\|\theta-\theta_{n,0}\|)_2^2\le\delta^2, \theta\in\Theta_n}\left|\frac{1}{\sqrt{n}}\left\langle(\theta-\theta_{n,0})\left(S\left(X_i\right)\right), \epsilon_i\right\rangle\right|\right] \\
&\le \int_0^\delta \sqrt{\log D\left(\varepsilon, \Omega', L_2\left(\mathbb{P}_n\right)\right)} d\varepsilon \\
&\le \int_0^\delta \sqrt{e\cdot\frac{d^{I(n)+1}-1}{d-1}\log\left(1+\frac{2\delta}{\varepsilon}\right)} d\varepsilon \\
&\le 2\sqrt{2e\cdot\frac{d^{I(n)+1}-1}{d-1}}\cdot\delta
\end{aligned} \tag{5.4.17}$$

We can choose the last term of 5.4.17 as the upper bound function $\phi_n(\cdot)$, i.e. setting

$$\phi_n(\delta) := 2\sqrt{2e\cdot\frac{d^{I(n)+1}-1}{d-1}}\cdot\delta \tag{5.4.18}$$

which is clearly an increasing function and $\delta \mapsto \phi_n(\delta)/\delta^\alpha$ is decreasing for some $\alpha < 2$ (e.g. $\alpha = \frac{3}{2}$). Thus, the condition 5.4.2 is satisfied.

Step 3

We move to construct the a sequence of δ_n which satisfies the condition 5.4.3. Firstly, we need

$$\phi_n\left(\delta_n\right) \leq \sqrt{n}\delta_n^2 \Leftrightarrow 2\sqrt{2e \cdot \frac{d^{I(n)+1} - 1}{n\left(d-1\right)}} \leq \delta_n \tag{5.4.19}$$

Then, as we mentioned, we assume that the predictor sequence $\hat{\theta}_n$ has the basic approximation ability which makes them δ-close to "true predictor" $\theta_{n,0}$ (i.e. we only need to find $\hat{\theta}_n$ in $\left\{\theta : \mathbb{P}_n\left(\|\theta - \theta_{n,0}\|^2\right) \leq \delta^2\right\}$). By setting $\delta_n \geq \delta$, we directly have $\delta_n^2 \geq \delta \geq M_n\left(\theta_{n,0}\right) - M_n\left(\theta_n\right)$ holds for any $\theta_n \in \Theta_n \cap \left\{\theta : \mathbb{P}_n\left(\|\theta - \theta_{n,0}\|_2^2\right) \leq \delta^2\right\} \subset \Theta_n$ due to our construction. Hence, we can chose δ_n as the maximum between δ and $2\sqrt{2e \cdot \frac{d^{I(n)+1}-1}{n(d-1)}}$. As the we are considering the case that our sequence of regressor $\hat{\theta}_n$ is close to the "true predictor", it is reasonable to simply the problem by setting:

$$\delta_n = 2\sqrt{2e \cdot \frac{d^{I(n)+1} - 1}{n\left(d-1\right)}} \tag{5.4.20}$$

As a consequence, the condition 5.4.3 is also satisfied. For the last condition, it is straightforward that θ_n takes value in Θ_n and $\mathbb{M}_n\left(\hat{\theta}_n\right) \geq \mathbb{M}_n\left(\theta_n\right) - O_{\mathbb{P}}\left(\delta_n^2\right)$ due to the problem setting. By the theorem 5.4.1, we have the convergence rate:

$$d_n\left(\hat{\theta}_n, \theta_{n,0}\right) = O_{\mathbb{P}}\left(2\sqrt{2e \cdot \frac{d^{I(n)+1} - 1}{n\left(d-1\right)}}\right) = O_{\mathbb{P}}\left(\sqrt{\cdot\frac{d^{I(n)}}{n}}\right) \tag{5.4.21}$$

Step 4

The above result 5.4.21 shows that the choice of truncated order can directly affect the convergence rate. It is clear that when $I(\cdot)$ is a constant (i.e ordinary linear regression case), the above result degenerates to the $O_{\mathbb{P}}(n^{-1/2})$ which is exactly the convergence rate for linear regression.

As for the truncated Signature regression problem, we may choose the truncated order by rules like $I : n \mapsto \lceil \log\log_d n \rceil$ or $I : n \mapsto \lceil \sqrt{\log_d n} \rceil$ which gives convergence rates $O_{\mathbb{P}}\left(\sqrt{\frac{\lceil \log n \rceil}{n}}\right)$ and $O_{\mathbb{P}}\left(n^{-1/4}\right)$ respectively. In the other words, with the increasing of dataset size n, the increasing order for truncated order could be choose like $o(\lceil \log\log_d n \rceil)$ or $o(\lceil \sqrt{\log_d n} \rceil)$ and the rate of convergence rate for the model can be easily determined by equation 5.4.21.

In the other side, if we take truncated order $N = \lceil \log_d n \rceil$, then we will have the corresponding convergence rate $O_{\mathbb{P}}\left(1\right)$, which implies that we will always hold the same "distance" between the least square estimator $\hat{\theta}_n$ and the "true" predictor $\theta_{n,0}$. In other words, the performance of the model will not be improved as the input dataset size n increases. This is actually a natural result, we have mentioned that we need to fit $\frac{d^{N+1}-1}{d-1}$ parameters with truncated order N. If we take $N = \lceil \log_d n \rceil$, then we will actually have the situation that the number of parameters $\simeq n$ which almost an interpolation situation which is not desired for regression problem.

Chapter 6

Conclusion and further research

In this thesis, we used truncated Signature features to build a least square regression model to learn the solution for a controlled ODE with forms 5.3.1. Applying results in empirical process theory, a theoretical explanation for the effect of truncated order is discovered. The main result 5.4.21 of this thesis tells us the relationship between the truncated order decision and the rate of convergence. One can design any decision rules $I : \mathbb{N}_+ \mapsto \mathbb{N}_+$ and easily get the corresponding rate of convergence by this results. Some suggested truncated order examples are also given at the end of the last section.

This result is not limited to the learning controlled ODE example we have discussed in this thesis. In fact, for problems that we can model with form $y = f(\{X_t\}_{t\in[a,b]})$ where y is the target, $\{X_t\}_{t\in[a,b]}$ is some continuous bounded variation path and f is an ideal model, then we may expect that there exists a continuous function $\tilde{f}$ such that $y = \tilde{f}\left(S(X)_{[a,b]}\right)$. Theorem 5.2.11 tells us that with some weak assumptions, $\tilde{f}$ can be uniformly approximated by a linear functional. Thus, we can model this problem like section 5.3 and our final result still holds.

However, this result only holds for the regression models with form 5.3.5. Further research can be expanded in other regression models. If we still choose square error as the loss function, then we can still build an empirical process in the same way as this thesis. We will see that the only thing that will be changed is the candidate space Θ_n. We may meet new challenge will be build the upper bond function $\phi_n(\cdot)$ using maximal inequalities since the ε-covering number $N\left(\varepsilon, \Theta_n \cap \left\{\theta : \mathbb{P}_n\left(\|\theta - \theta_{n,0}\|_2^2\right) \leq \delta^2\right\}, L_2\left(\mathbb{P}_n\right)\right)$ for the new candidates space may not be trivial.

Another further work can be expanded on log Signature. The space in which the log Signature of a path in $\mathbb{R}^d$ up to level N is equivalent to the free N-nilpotent Lie algebra. The log Signature is like a compressed version of the Signature up to the same level [21]. In another word, one may use the log Signature as an efficient representation of a path. Thus, the truncated order for log Signature regression problem may be interesting to discuss.

Bibliography

[1] D. Levin, T. Lyons, and H. Ni. Learning from the past, predicting the statistics for the future, learning an evolving system. *arXiv preprint arXiv:1309.0260*, 2016.

[2] B. Sen. *A Gentle Introduction to Empirical Process Theory and Applications.* `http://www.stat.columbia.edu/~bodhi/Talks/Emp-Proc-Lecture-Notes.pdf`, 2018.

[3] V. Glivenko and F. P. Cantelli. Sulla determinazione empirica delle leggi di probabilità. *Giorn. Ist. Ital. Attuari*, 4:92–99, 1993.

[4] M. D. Donsker. Justification and extension of doob's heuristic approach to the komogorov-smirnov theorems. *Ann. Math. Statistics*, 23:642–669, 1952.

[5] M. D. Donsker. An invariance principle for certain probability limit theorems. *Memoirs of the American Mathematical Society*, 1951.

[6] F. P. Cantelli. Sulla determinazione empirica delle leggi di probabilità. *Giorn. Ist. Ital. Attuari*, 1933.

[7] Vidyadhar P. Godambe. *Estimating functions, volume 7 of Oxford Statistical Science Series.* The Clarendon Press Oxford University Press, New York, 1991.

[8] A. W. Van Der Vaart and Jon A. Wellner. *Weak convergence and empirical processes.* Springer Series in Statistics. Springer-Verlag, New York. With applications to statistics, 1996.

[9] Wikipedia. Donsker's theorem, 2020.

[10] P. Glasserman. *Monte Carlo Methods in Financial Engineering.* New York: Springer-Verlag, 2004.

[11] Galen R Shorack and Jon A Wellner. *Empirical processes with applications to statistics.* SIAM, 2009.

[12] J. Komlos, P. Major, and G. Tusnady. An approximation of partial sums of independent rv's and the sample df. i. *Wahrsch verw Gebiete/Probability Theory and Related Fields*, 32:111–131, 1975.

[13] J. Komlos, P. Major, and G. Tusnady. An approximation of partial sums of independent rv's and the sample df. ii. *Wahrsch verw Gebiete/Probability Theory and Related Fields*, 34:33–58, 1976.

[14] P. Bartlettl. *Theoretical Statistics, Lecture 2.* UC Berkeley Statistics `https://www.stat.berkeley.edu/~bartlett/courses/2013spring-stat210b/notes/2notes.pdf`.

[15] M. R. Kosorok. *Introduction to empirical processes and semiparametric inference.* Springer Science & Business Media, 2007.

[16] T. Lyons. Differential equations driven by rough signals. *Revista Matemática Iberoamericana*, pages 215–310, 1998.

[17] T. Lyons and Z. Qian. System control and rough paths. *Oxford Mathematical Monographs. Oxford*, 2002.

[18] T. Lyons, M. Caruana, and T. Lévy. *Differential Equations Driven by Rough Paths.* Springer-Verlag Berlin Heidelberg, 2004.

[19] Luigi Ambrosio, Nicola Fusco, and Diego Pallara. *Functions of bounded variation and free discontinuity problems*, volume 254. Clarendon Press Oxford, 2000.

[20] B. Hambly and T. Lyons. Uniqueness for the signature of a path of bounded variation and the reduced path group. *Annals of Mathematics*, 141:109–167, 2010.

[21] Jeremy Reizenstein and Benjamin Graham. The iisignature library: efficient calculation of iterated-integral signatures and log signatures. *arXiv preprint arXiv:1802.08252*, 2018.

[22] Ilya Chevyrev and Andrey Kormilitzin. A primer on the signature method in machine learning. *arXiv preprint arXiv:1603.03788*, 2016.

[23] Matt Young. The stone-weierstrass theorem. In *MATH 328 Notes*. Queen's University at Kingston, 2006.