Signature Regression and an Application to Credit Cycle Forecasting

Master's Thesis

Niklas Weber

Supervisor: Prof. L. Gonon

Fakultät für Mathematik, Informatik und Statistik
Ludwig-Maximilians-Universität München

This thesis is submitted for the degree
Master of Science

February 2022

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this thesis are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This thesis is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This thesis contains fewer than 300,000 symbols including appendices, bibliography, footnotes, tables, and equations.

Niklas Weber
February 2022

Acknowledgements

Writing this thesis would not have been possible without the remarkable support I received.

First, I would like to thank my supervisor, Professor Lukas Gonon. His determined support alongside his impressive expertise encouraged me throughout the entire process. Simultaneously he allowed me a momentous amount of freedom to pursue various directions and approaches.

Furthermore, I would like to acknowledge Dr. Oleg Reichmann at ECB and Dr. Giuseppe Bonavolontà at EIB, who both played a significant role in the selection of the topics and gave me many valuable insights into the applications of these topics at their institutions.

I would also like to thank my parents for their wise counsel and patient support. Finally, I wish to thank Henry Kleineidam and Yannick Limmer for helpful comments and their patient assistance.

Table of contents

List of figures

Chapter 1

Introduction

Would it not be nice to be able to predict the future? Will this thesis be interesting? Will you learn something new, or is it a waste of time? Should you take an umbrella with you, when leaving the house tomorrow and what are the numbers for the lottery jackpot on Saturday?

The task of predicting the future in its most general sense has been present ever since the dawn of time. It is fair to assume that prophecies of priests or oracles, like the famous oracle of Delphi, caused uncountable many wars. And even besides this area where the line between history and mythology becomes blurry, our ancestors faced many questions whose answers would make the difference between surviving, or not: "How long will the winter last? Will the harvest be good enough to feed the whole tribe? Can this lion run faster than I can?".

Other than appeasing the gods, a very powerful tool to find answers to this questions is to make use of the experiences and knowledge – either personal or traditional – from similar situations in the past. A reason why this can work is that often there exist observable indicators which might have some connection to the specific outcome in question.

Nowadays, the field of estimating the relationship between a dependent variable (an outcome) and independent/explanatory variables on the other side (also called predictors, covariates or features) is the field of *regression analysis*. The name *regression* for this kind of problem was popularized by Francis Galton in the 19th century, when he discovered the phenomenon that the heights of descendants of tall ancestors would tend to *regress* down towards an average [7].

In the remainder of this introduction we will present two types of regression techniques, i.e. the linear regression and functional regression. Both regression types utilize different forms of explanatory variables: vector valued data and functional data

respectively. After this we will introduce the signature, we will see that the signature has potential to naturally connect both regression types and we will outline aim and contents of the following chapters of this thesis.

1.1 Linear regression

There exist several types of regression techniques, but one of the most common ones is the so called linear regression: Starting from a data set containing $n \in \mathbb{N}$ observations of the $p \in \mathbb{N}$ dimensional vector of predictors $x_i = (x_1^i, ..., x_p^i) \in \mathbb{R}^p$, $i = 1, ..., n$ and the corresponding observed outcomes $y_i \in \mathbb{R}$, $i = 1, ..., n$, we assume a linear relation

$$y_i = (1, x_1^i, ..., x_p^i) \cdot (\beta_0, ..., \beta_m) + \epsilon_i, \quad i = 1, ..., n \tag{1.1}$$

where $\cdot$ denotes the inner product, or in matrix notation

$$y = X\beta + \epsilon$$

where $\beta = (\beta_0, \beta_1, ..., \beta_p) \in \mathbb{R}^{p+1}$ are unobservable coefficients (β_0 is called intercept) and X is a matrix where every row contains one observation x_i preceded by a one in order to handle the intercept. The ϵ_i are random error terms incorporating either actual errors, e.g. measurement inaccuracies, or additional influences of other predictors not explicitly considered in the model, which add noise to the "expected" responses $X\beta$ resulting in the outcomes y_i we can observe.[1]

Since β is not observable the goal is to obtain a good estimate $\hat{\beta}$ and use this estimated $\hat{\beta}$ to make predictions about unseen data, i.e. data where we only observed the predictors x_j, but not yet the response y_j, hoping that the linear regression model generalizes well to the new data.

Assuming one was successful selecting the predictors and hence incorporating all major drivers of the dependent variable in the linear regression model, there exist central limit theorems like the Lindeberg-Feller theorem [4, Thm. 3.4.5] that suggest the error terms ϵ_i to be distributed normally. In prose the theorem states that the infinite sum of many small and independent random variables is distributed normally under some assumptions.

[1] Throughout this thesis we will shorten the notation where it seems appropriate and deliberately skip declaring indexing variables like i or j if the dimension of the object was already clarified earlier, it is clear from the context or it is not important. For example "one observation x_i" relates to one arbitrary observation from the set $\{x_i\}_{i=1,...,n}$. In definitions, theorems, etc., however, we will rigorously declare every variable.

In such a setting where the errors are normal and X^TX is invertible, i.e. no predictor variable is redundant, the ordinary least squares (OLS) estimator is the "best" linear unbiased estimator. This means that

$$\hat{\beta} = (X^TX)^{-1}X^Ty$$

is unbiased, i.e. $\mathbb{E}[\hat{\beta}] = \beta$, and the square error is minimal in the sense that

$$\mathbb{E}\left[\left(\sum_{i=0}^{p} \lambda_i(\beta_i - \hat{\beta}_i)\right)^2\right]$$

is minimal for all $\lambda \in \mathbb{R}^{p+1}$. This result is known as the Gauß-Markov theorem [10, Chapter 7] and in fact already in the early 1800s Gauss and Legendre used the method of least squares to calculate the orbits of celestial bodies around the sun.

1.2 Functional regression

Sometimes, unfortunately, the setting is just not suitable for a linear regression. For example if the predictors are not simple data points x_i but continuous paths X_i, or paths sampled with such high frequency that they are "almost" continuous. In this case it can be unfeasible or impossible to treat every point of this path as a predictor for the regression and one has to come up with a different approach. An example would be predicting the yearly overall precipitation from daily sampled average temperatures over one year, which is discussed by A. Fermanian in [6].

One possible solution is to approximate the input path X_i mapping from some arbitrary domain[2] D into the real numbers $X_i : D \to \mathbb{R}$ by a set of orthonormal basis functions $\{\phi_i\}_{i=1}^{\infty}$, e.g. normed polynomials. Then instead of

$$y_i = (1, x_1^i, ..., x_p^i) \cdot (\beta_0, ..., \beta_m) + \epsilon_i = \beta_0 + \langle x^i, \beta \rangle$$

with the Euclidean scalar product in $\mathbb{R}^n$, one could switch to the Scalar product in L^2, i.e. the space of measurable functions for which the second power of the absolute value is Lebesgue integrable and functions which agree almost everywhere are identified, and consider the model

$$y_i = \beta_0 + \langle X_i, \beta \rangle + \epsilon_i = \beta_0 + \int_D X_i(t)\beta(t)dt + \epsilon_i, \quad i = 1, ..., n \tag{1.2}$$

[2] We will usually consider time intervals $[a, b] \subset \mathbb{R}$.

where β is not a vector of coefficients anymore but a weighting function $\beta : D \mapsto \mathbb{R}$. If β is approximated by the same orthonormal basis such that

$$X_i(\cdot) = \sum_{i=0}^{\infty} x_i \phi_i(\cdot)$$

and

$$\beta(\cdot) = \sum_{i=0}^{\infty} \beta_i \phi_i(\cdot),$$

we can plug this into formula (1.2) and the orthonormality of the basis function leaves us with

$$y_i = \beta_0 + \sum_{i,j=0}^{\infty} \beta_j x_i \int_D \phi_i(t)\phi_j(t)dt = \beta_0 + \sum_{i=0}^{\infty} \beta_i x_i.$$

Considering only a finite number $N \in \mathbb{N}$ of the basis functions we can approximate the x_i, $i = 1, ..., N$, which brings us back to a linear regression setting where we can estimate the coefficients β_i.

A major drawback of this approach is that one has to choose a finite set of basis functions, which are meant to span the function space in which β and all X_i are projected into. During this step we could already inherit inaccuracies affecting the whole model. Moreover, this is the 1-dimensional case only, i.e. the input path X_i maps to $\mathbb{R}$.

If the path is multidimensional $X_i : D \to \mathbb{R}^n$ the choice of possible basis functions becomes even more complicated. One possible approach is to decide which basis functions should represent every single dimension. However, by doing so we impose that it is indeed possible for our data to evolve in every dimension independently of the other dimensions. This is a very strong assumption on the dependency structure of the underlying "true" process and can be wrong. Furthermore, it seems unlikely to successfully select a good set of basis functions, which is close to the corresponding "true" dependence structure of the data generating process, for every single regression task.

Consequently, it is a valid question to ask if there is any general method to transform a possibly multidimensional path into something that can serve as an input for a linear regression. Indeed it seems that the so called signature of a path is capable of doing so.

1.3 Signature

The signature of a path is a sequence of iterated integrals. For a $d \in \mathbb{N}$ dimensional path $X : D \to \mathbb{R}^d$, where D is a time interval $[a, b] \subset \mathbb{R}$ we can define one signature term for every $k \in \mathbb{N}$-tuple of components $\mathbb{I} = (i_1, ...i_k) \subseteq \{1, ..., d\}^k$. The signature term of path X for tuple $\mathbb{I} = (i_1, ...i_k)$ is defined as the iterated integral

$$S(X)^{\mathbb{I}} := \int_{a \leq u_1 \leq ... \leq u_k \leq b} \cdots \int dX_{u_1}^{i_1} ... dX_{u_k}^{i_k}.$$

Subsequently the complete signature of such a path is defined as the sequence of all the signature terms of all possible multi-indices

$$S(X) = (1, S(X)^1, ..., S(X)^d, S(X)^{(1,1)}, S(X)^{(1,2)}, ..., S(X)^{(d,d)}, S(X)^{(1,1,1)}, ...).$$

This signature object consisting of iterated integrals is an interesting feature set for continuous paths that captures deep geometric properties as we will see in the next chapter. Selecting a finite number of signature terms by truncating the signature reduces the functional regression setting to a linear regression in a more general manner than choosing basis functions. In fact A. Fermanian showed in [6] that linear regression on the truncated signature (it is a infinite sequence, so it must be truncated for applications) can outperform or match functional regression techniques.

In this thesis we proceed as follows: The next chapter intends to define the signature of bounded variation paths rigorously and investigate the properties of the signature closely. We will also provide examples for a better understanding of the signature object and mention some computational aspects.

Afterwards we will introduce a model for estimating the truncation order and subsequently performing linear regression with the truncated signature, i.e. signature regression.

In Chapter 4 we prove error bounds, that were originally proven by A. Fermanian [6] and make the model from Chapter 3 even more appealing.

In the last chapter we will try to answer the questions if and when signature regression is a good alternative to linear regression. We consider different types of synthetically generated data and also investigate one real-world application in the context of Credit Cycle Forecasting. The source code implementing signature regression and performing all simulation can be found in the GitHub repository: https://github.com/NiklasMWeber/CreditCycleForecasting.

Chapter 2

Properties of the Signature

2.1 The Path Signature

In this chapter we will introduce the path signature and inspect some interesting properties. Since the signature is defined via iterated integrals, it is crucial to make sure those integrals are well defined. To guarantee this we choose the Riemann-Stieltjes integral type for paths of bounded variation, which is defined as follows.

Let X be a path mapping from some time interval $D = [a, b]$ into $\mathbb{R}^d$:

$$\begin{aligned} X : D &\to \mathbb{R}^d \\ t &\mapsto X_t. \end{aligned}$$

Definition 2.1 (Total variation)**.** *The total variation of the path X is defined as*

$$\|X\|_{TV} := \sup_{\pi \in \mathbf{P}} \sum_{i=0}^{k-1} \|X_{t_{i+1}} - X_{t_i}\|,$$

where the supremum is taken over the set of all partitions of $[a, b]$*, i.e.* $\pi : 0 = t_0 < ... < t_k = 1$ *and* $\|\cdot\|$ *denotes the Euclidean norm.*

Definition 2.2 (Bounded Variation paths)**.** *Furthermore, we say a path X is of bounded variation, if its total variation is finite. Hence the set of bounded variation paths (on D) is denoted by*

$$BV(\mathbb{R}^d) = \{X : D \to \mathbb{R}^d | \|X\|_{TV} < \infty\}.$$

The Riemann-Stieltjes integral finally is defined by:

Definition 2.3 (Riemann-Stieltjes integral)**.** *Let $f : [a, b] \to \mathbb{R}$ be a real valued function on some real interval and $g \in BV(\mathbb{R})$ on the same interval. The Riemann-Stieltjes integral is defined as*

$$\int_a^b f(x)dg(x) := \lim_{\|\pi\| \to 0} \sum_{i=0}^{k-1} f(c_i) \left(g(t_{i+1}) - g(t_i)\right),$$

where $c_i \in [t_i, t_{i+1}]$ and the limit is taken over a sequence of partitions π of $[a, b]$ with decreasing norm $\|\pi\|$.

This notion of an integral matches our setting, because a path $X : D \to \mathbb{R}^d, t \mapsto X_t = (X_t^1, ..., X_t^d)$ consists component-wise of exactly what we have in the definition, i.e. real-valued functions on some real interval. Accordingly we can define the signature of a path:

Definition 2.4 (Signature)**.** *The signature of a path is a sequence of iterated integrals of the following form: For a $d \in \mathbb{N}$ dimensional continuous path $X : D \to \mathbb{R}^d$, where D is a time interval $[a, b] \in \mathbb{R}$ and the path X is of bounded variation the signature term corresponding to the multi-index $\mathbb{I}$ of length $k \in \mathbb{N}$, $\mathbb{I} = (i_1, ...i_k) \subseteq \{1, ..., d\}^k$ is defined as the iterated Riemann-Stieltjes integral*

$$S(X)^{\mathbb{I}} := \int_{a \le u_1 \le ... \le u_k \le b} \cdots \int dX_{u_1}^{i_1} ... dX_{u_k}^{i_k}. \tag{2.1}$$

Subsequently the signature of a path is defined as the sequence of all the signatures terms corresponding to all possible multi-indices, where the first term (level zero term) is 1 by convention:

$$S(X) = (1, S(X)^1, ..., S(X)^d, S(X)^{(1,1)}, S(X)^{(1,2)}, ..., S(X)^{(d,d)}, S(X)^{(1,1,1)}, ...) \tag{2.2}$$

The signature truncated at level $m \in \mathbb{N}$ is the finite sequence only considering multi-indices $\mathbb{I}$ of length m or less, $\mathbb{I} \in \{(i_1, ..., i_l) | l \in \mathbb{N}, l \le m, (i_1, ..., i_l) \in \{1, ..., d\}^l\}$:

$$S(X)^m := (1, S(X)^1, ..., S(X)^{\overbrace{(d, ..., d)}^{length\ m}})$$

Remark 1. A multi-index $\mathbb{I}$ of length k, $\mathbb{I} = (i_1, ...i_k) \in \{1, ..., d\}^k$ can also be called a *word* of length k of the *alphabet* {1,...,d}.

For applications, of course, we have to work with the truncated signature. It is obvious that we cannot compute an infinite amount of signature terms, but even

for high, finite truncation orders we would encounter problems very quickly, as the number of signature terms grows exponentially. From simple combinatorics we know the number of m-th order terms of a d-dimensional path is d^m.
Therefore, the total number of entries of a $d \geq 2$ dimensional path's signature truncated at level m corresponds to

$$s_d(m) := \sum_{k=0}^{m} d^k = \frac{d^{m+1} - 1}{d - 1}. \tag{2.3}$$

This number of terms can explode very easily as the following table with typical couples of d and m shows (2.1).

	$d = 2$	$d = 3$	$d = 6$
$m = 1$	2	3	6
$m = 2$	6	12	42
$m = 5$	62	363	9330
$m = 7$	254	3279	335922

Fig. 2.1 Table of signature length for typical values of m and d
Note. From "Linear functional regression with truncated signatures," by A. Fermanian, 2021, Preprint, p. 7 (https://arxiv.org/abs/2006.08442v2). Copyright 2021 by A. Fermanian.

However, truncating the signature does not necessarily lead to a huge loss of information. Indeed we can justify theoretically that the information loss from truncating the signature is in some sense limited. The following proposition shows that the norm of higher order terms decays quickly and in fact for every truncation order the signature is bounded by the exponential of the path's total variation. We borrow the notation from A. Fermanian [6]. The proof by B. Hambly and T. Lyons can be found in [8] in a more general setting.

Proposition 2.5 (Decaying norm of signature terms)**.** *Let $X : [0, 1] \to \mathbb{R}^d$ be a bounded variation path. Then for any $m \geq 0$,*

$$\|S^m(X)\| \leq \sum_{k=0}^{m} \frac{\|X\|_{TV}^k}{k!} \leq e^{\|X\|_{TV}}.$$

Before moving on to the next section, we will now formulate some examples from I. Chevyrev and A. Kormilitzin's "Primer on the Signature Method in Machine Learning"

[3] to enhance the reader's intuition on signatures. We already know that the order zero term is one by convention. Furthermore, it is easy to see that the first order terms will always consist of the increment in the corresponding coordinate:

$$\int_a^b dX_t^i = X_b^i - X_a^i.$$

Example 1. The simplest example we can consider for a path is the one dimensional path $X : [a, b] \mapsto \mathbb{R}$, $X_t = t$. The zeroth and first order terms are clear:

$$\begin{aligned} S(X)^0 &= 1, \\ S(X)^1 &= X_b - X_a = b - a, \end{aligned}$$

In order to calculate the second order term we do two things. First, it will be helpful to also denote on which interval a signature is being calculated as index, e.g. $S(X)^{\mathbb{I}}_{a,b}$. This reveals an iterative structure, which we can use for calculating higher order terms. Furthermore, we note that X_t is differentiable in time and therefore, by considering some other path Y_t, we can write $\int Y_t dX_t = \int Y_t \dot{X}_t d_t = \int Y_t \frac{\partial}{\partial t} X_t dt$. With this observations we get the following:

$$\begin{aligned} S(X)^{(1,1)}_{a,b} &= \iint\limits_{a \le u_1 \le u_2 \le b} dX^1_{u_1} dX^1_{u_2} = \int_{a \le u_2 \le b} \underbrace{\int_{a \le u_1 \le u_2} dX^1_{u_1}}_{= S(X)^1_{a,u_2}} dX^1_{u_2} \\ &= \int_{a \le u_2 \le b} S(X)^1_{a,u_2} \, dX^1_{u_2} = \int_{a \le u_2 \le b} (u_2 - a) \, dX^1_{u_2} \\ &= \int_{a \le u_2 \le b} (u_2 - a) \underbrace{\dot{X}_t}_{=1} \, du_2 = \frac{(b-a)^2}{2!}. \end{aligned}$$

In the same manner we get for the third and all following orders

$$\begin{aligned} S(X)^{(1,1,1)}_{a,b} &= \frac{(b-a)^3}{3!} \\ S(X)^{\overbrace{(1,...,1)}^{\text{length } m}}_{a,b} &= \frac{(b-a)^m}{m!}, \quad m \in \mathbb{N}. \end{aligned}$$

This illustrates that all signature terms of the one dimensional path $X_t = t$ depend only on the increment in this dimension. Actually it can be shown that this is also true for any one dimensional path which will become more intuitive to us later after

discussing some properties of the signature. For this reason the signature is more interesting for paths of higher dimensions, which brings us to the next, more involved example:

Example 2. We define the path

$$\begin{aligned} X_t &= (X_t^1, X_t^2) = (3+t, (3+t)^2) \quad t \in [0,5],\ (a=0, b=5) \\ dX_t &= (dX_t^1, dX_t^2) = (dt, 2(3+t)dt). \end{aligned}$$

For the signature terms we calculate

$$\begin{aligned} S(X)_{0,5}^{(1)} &= \int_0^5 dX_t^1 = X_5^1 - X_0^1 = 8 - 3 = 5 \\ S(X)_{0,5}^{(2)} &= \int_0^5 dX_t^2 = X_5^2 - X_0^2 = 64 - 9 = 55, \\ S(X)_{0,5}^{(1,1)} &= \iint_{0\le t_1\le t_2\le 5} dX_{t_1}^1 dX_{t_2}^1 = \int_0^5 \int_0^{t_2} dt_1 dt_2 = \int_0^5 t_2 dt_2 = \frac{25}{2}, \\ S(X)_{0,5}^{(1,2)} &= \iint_{0\le t_1\le t_2\le 5} dX_{t_1}^1 dX_{t_2}^2 = \int_0^5 \int_0^{t_2} dt_1\, 2(3+t_2)\, dt_2 = \int_0^5 6t_2 + 2t_2^2\, dt_2 = \frac{475}{3}, \\ S(X)_{0,5}^{(2,1)} &= \iint_{0\le t_1\le t_2\le 5} dX_{t_1}^2 dX_{t_2}^1 = \int_0^5 \int_0^{t_2} 2(3+t_1) dt_1\, dt_2 = \int_0^5 6t_2 + t_2^2\, dt_2 = \frac{350}{3}, \\ S(X)_{0,5}^{(2,2)} &= \iint_{0\le t_1\le t_2\le 5} dX_{t_1}^2 dX_{t_2}^2 = \int_0^5 \int_0^{t_2} 2(3+t_1) dt_1\, 2(3+t_2)\, dt_2 \\ &= \int_0^5 (6t_2 + t_2^2)\, 2(3+t_2)\, dt_2 = \int_0^5 36t_2 + 18t_2^2 + 2t_2^3\, dt_2 = \frac{3025}{2}, \\ S(X)_{0,5}^{(1,1,1)} &= \iiint_{0\le t_1\le t_2\le t_3\le 5} dX_{t_1}^1 dX_{t_2}^1 dX_{t_3}^1 = \int_0^5 \int_0^{t_3} \int_0^{t_2} dt_1\, dt_2\, dt_3 = \int_0^5 \frac{t_3^2}{2} dt_3 = \frac{125}{6}. \end{aligned}$$

We see that in this case the signature terms are no longer driven by the increments in such an obvious way, as it was the case in the previous example. We could calculate even more values, but instead let us think about one peculiarity of this signature for a minute: We can clearly see that the only information about the path entering the iterated integrals is the information encoded by the differential dX, but never actually X. The signature only cares for the change in the path X, not its absolute position. Remember that the first coordinate of the path was 3+t. The offset 3 was completely irrelevant for calculating of the signature and might as well have been any other number instead.

2.2 Invariances

Having seen two examples of the signature in the last section now we might wonder which information about the original path the signature can capture and what properties it has. This will be investigated in this section starting with three basic properties of the path that the signature cannot capture. From now on without loss of generality we will assume $[a, b] = [0, 1]$.

- We start with the translation invariance that was already motivated in Example 2.

Proposition 2.6 (Translation invariance)**.** *Let $X : [0, 1] \to \mathbb{R}^d$ be a path of bounded variation. Then for any $a \in \mathbb{R}^d$ and $\tilde{X} := X + a$ it holds that*

$$S(X) = S(\tilde{X}).$$

Proof. The signature is defined via iterated integrals. This integrals only consider dX and not X. Since $dX = d(X + a)$ for any constant a, shifting the path X by a has no influence on the signature. □

- The second property that the signature cannot capture is "speed". Figure 2.2 shows two examples of one dimensional reparametrized paths. The first example are two reparametrizations of monotonous paths from 0 to 1. The second example are paths traveling from 0 to 1.2, descending to -0.2 and finally ending at 1. We can only distinguish the paths because of the additional time dimension plotted in the figure. Merely handed a list of points visited and their order without time stamps, in both cases the two paths would look identical and we would be as blind as the signature is. We call this property invariance under time-reparametrization.

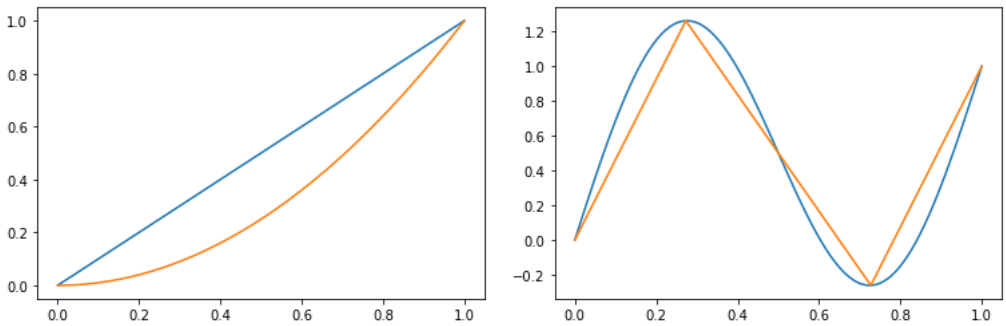

Fig. 2.2 Time reparametrisations of 1-dimensional paths

Proposition 2.7 (Invariance under time reparametrization)**.** *Let $X : [0,1] \to \mathbb{R}^d$ be a bounded variation path and $\phi : [0,1] \mapsto [0,1]$ a reparametrization, i.e. a surjective, continuous, non-decreasing function. Then for the path $\tilde{X} : [0,1] \to \mathbb{R}^d$ defined by $\tilde{X}_t = X_{\phi(t)}$ it is true that:*

$$S(\tilde{X}) = S(X).$$

Proof. A proof for smooth reparametrizations is given in [3]. We will skip the proof here and restrict ourselves to the intuition that the signature is only capable of capturing the visited points and their order, but not the speed of traversal. □

- A third interesting property of the signature is that the signature of a constant path and that of a path starting and ending in the same point is identical. This is a result of Chen's identity, which we will see later, after we defined the concatenation of paths. The intuition for us is that the signature cannot detect loops that start and end in the same point. This is also the reason why the signature of every one dimensional path is strictly driven by its increment. Every one dimensional path can be decomposed into a monotonous bit moving from start- to end-point, possibly interrupted by loops that end where they started. With only one dimension available it is simply not possible to move back and forth without visiting the same points again, hence creating loops, which cannot be detected by the signature. This inability to detect loops is made rigorous with the notion of "tree-like equivalence" by B. Hambly and T. Lyons in [8].

Remark 2. With this insight we realize that not only the two monotonous paths and the two non-monotonous paths from Figure 2.2 have the same signature, but actually that all four of them share the same signature, i.e. the signature we computed in Example 1.

For us the moral of the story is whenever we are working with the signature of paths it might be a good idea to add a monotonous dimension to the paths first (usually the time) to make sure there are no time reparametrizations or loops. Furthermore, it can also be helpful to add a common base-point to all paths in order to distinguish between paths and their translations.

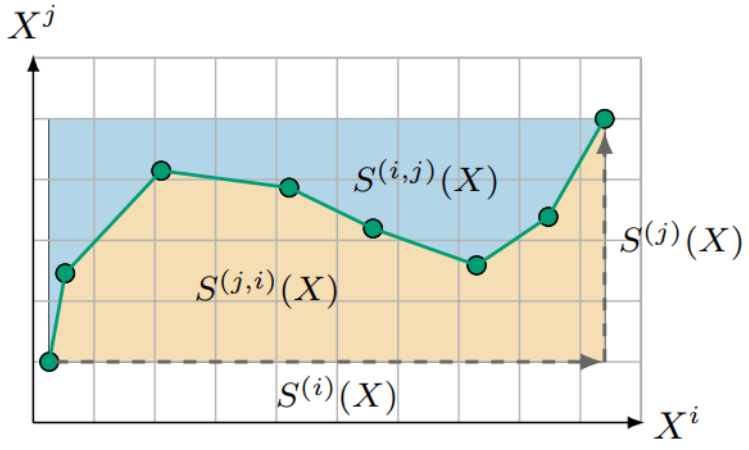

Fig. 2.3 Illustration of second order signature terms
Note. From "Linear functional regression with truncated signatures," by A. Fermanian, 2021, Preprint, p. 6 (https://arxiv.org/abs/2006.08442v2). Copyright 2021 by A. Fermanian.

2.3 Fundamental properties

The last paragraphs painted a very pessimistic picture of the signature because of its insufficiencies. But the signature also captures useful geometric information.

- We already discussed that the first order terms of the signature correspond to the increments in the respective coordinates.
- There is also a nice visualization of the information captured by the second order terms. In Figure 2.3 we see the X^i and X^j component of some path plotted against each other. This illustrates that the path divides the rectangle spanned by $S(X)^i$ and $S(X)^j$, i.e. the increments, into two different areas. The area underneath the path corresponds to $S(X)^{(j,i)}$, the area above the path to $S(X)^{(i,j)}$.
- Higher order signature terms in general contain information about the joint evolution of tuples of coordinates.

Figure 2.3 also shows that the signature terms are connected by the following formula:

$$S(X)^{(i)}S(X)^{(j)} = S(X)^{(i,j)} + S(X)^{(j,i)}.$$

This example indicates that the signature terms are in some sense redundant. We could calculate any of the four values with the three remaining ones. There is a version of the signature that excludes some of these redundancies, the "log-signature". The log-signature also has interesting properties including some advantages but also some

Fig. 2.4 A riffle shuffle

disadvantages, which is the reason why it is not presented in this thesis. The interested reader can be referred to [3].

Indeed the redundancy above is only one specific example of the so called "shuffle product identity" that was originally shown by R. Ree [13]. This fundamental property states that we can always calculate the *product* of two signature terms $S(X)^{(i_1,...,i_k)}S(X)^{(j_1,...,j_m)}$ by a *sum* of signature terms $S(X)^{\mathbb{I}}$ whose indices $\mathbb{I}$ only depend on the multi-indices $(i_1, ..., i_k)$ and $(j_1, ..., j_m)$. We have to introduce the shuffle product of two multi-indices, which is a set of permutations that we can imagine as the set of all possible riffle shuffles one might know from playing Poker or watching magicians perform a card trick. The riffle shuffle is a technique to shuffle two piles of cards into one pile as shown in Figure 2.4. The important property of the riffle shuffle is, that the order of cards from each pile is maintained in the resulting pile. The cards of pile one might be interrupted by one or multiple cards from pile two, but they will be found in the same order they were before.

Following this intuition we call a permutation σ of the set {1,...,k+m} a $(k, m)-$shuffle, if $\sigma^{-1}(1) \leq ... \leq \sigma^{-1}(k)$ and $\sigma^{-1}(k+1) \leq ... \leq \sigma^{-1}(k+m)$, i.e. the first k and the last m cards (the "piles") maintain their order. We denote the set of all possible $(k, m)-$shuffles by Shuffles(k, m). Formally we write:

Definition 2.8 (Shuffle product)**.** *Consider two multi-indexes* $\mathbb{I} = (i_1, ..., i_k)$ *and* $\mathbb{J} = (j_1, ..., j_m)$ *with* $i_1, ..., i_k, j_1, ..., j_m \in \{1, ..., d\}$*. Define the multi-index*

$$(r_1, ..., r_k, r_k + 1, ..., r_k + m) = (i_1, ..., i_k, j_1, ..., j_m).$$

The shuffle product of $\mathbb{I}$ *and* $\mathbb{J}$*, denoted by* $\mathbb{I} \sqcup\!\sqcup \mathbb{J}$*, is a finite set of multi-indices of length* $k + m$ *defined by*

$$\mathbb{I} \sqcup\!\sqcup \mathbb{J} = \{(r_{\sigma(1)}, ..., r_{\sigma(k+m)}) | \sigma \in \mathit{Shuffles}(k, m)\}.$$

Proposition 2.9 (Shuffle product identity)**.** *Let* $X : [0, 1] \to \mathbb{R}^d$ *be a path and* $\mathbb{I} = (i_1, ..., i_k)$*,* $\mathbb{J} = (j_1, ..., j_m)$ *two multi-indices with* $i_1, ..., i_k, j_1, ..., j_m \in \{1, ..., d\}$*. It holds that*

$$S(X)^{\mathbb{I}} S(X)^{\mathbb{J}} = \sum_{K \in \mathbb{I} \sqcup\!\sqcup \mathbb{J}} S(X)^K.$$

Example 3. We already demonstrated one example of this identity, which was

$$S(X)^{(1)} S(X)^{(2)} = S(X)^{(1,2)} + S(X)^{(2,1)}$$

if we set $i, j = 1, 2$. We know the reason for this equation is the shuffle product identity and the fact that $(1) \sqcup\!\sqcup (2) = \{(1, 2), (2, 1)\}$. Another example is

$$S(X)^{(1,2)} S(X)^{1} = 2S(X)^{(1,1,2)} + S(X)^{(1,2,1)},$$

since $(1,2) \sqcup\!\sqcup (1) = \{(1, 1, 2), (1, 1, 2), (1, 2, 1)\}$.

Remark 3. The shuffle product identity implies that products of two signature terms can be expressed as a linear combination of higher order signature terms. This is one reason why the signature is thought of as a powerful feature transform for path valued data in the machine learning context. Like polynomials in the vector-valued case the signature is the analogue of polynomials in the path-space. We will further strengthen this relationship at the end of this chapter with Proposition 2.13.

2.4 Computational aspects

As one of the last topics we will now introduce K.-T. Chen's identity [2] as it can be found in [5]. Theoretically, Chen's identity has a much deeper meaning if the signature is interpreted as an element in the space of formal power series, but we are only interested in the relationship between the signature terms of two paths and those of their concatenation, that follows from Chen's identity. This will help us to understand how to compute the signature algorithmically from a discrete data stream. We define the concatenation of two paths as follows.

Definition 2.10 (Concatenation of paths)**.** *Let $X : [s,t] \mapsto \mathbb{R}^d$ and $Y : [t,u] \mapsto \mathbb{R}^d$ be two paths, $0 \leq s < t < u \leq 1$. The concatenation of X and Y, denoted by $X * Y$, is defined as the path from $[s,u]$ to $\mathbb{R}^d$, such that for any $v \in [s,u]$*

$$(X * Y)_v = \begin{cases} X_v & \text{if } v \in [s,t] \\ X_t + Y_v - Y_t & \text{if } v \in [t,u]. \end{cases}$$

Proposition 2.11 (Chen's identity)**.** *Let $X : [s,t] \to \mathbb{R}^d$ and $Y : [t,u] \to \mathbb{R}^d$ be two paths of bounded variation. Then for any multi-index $(i_1, ..., i_k) \subset \{1, ..., d\}^k$, we have*

$$S(X * Y)^{(i_1,...,i_k)} = \sum_{l=0}^{k} S(X)^{(i_1,...,i_l)} S(Y)^{(i_{l+1},...,i_k)}.$$

In particular Chen's identity tells us, that every k-th order signature term of a concatenation of paths can be calculated considering signature terms of the original paths of order only up to k. This is practical, because in order to calculate the signature truncated at level k of a concatenation, apparently we only need the signatures truncated at level k of the original paths and no additional terms. The next example from [5] shows how to calculate the signature terms of any linear path. Afterwards we will be able to formulate an algorithm.

Example 4. Let X be a linear path $X_t = (X_t^1, X_t^2) = (a_1 + b_1 t, a_2 + b_2 t), [s,t] \subset \mathbb{R}$.

$$\begin{aligned} S(X)_{s,t}^{(i)} &= \int_s^t dX_u^i = b_i(t-s) \\ S(X)_{s,t}^{(1,1)} &= \iint_{s \leq u_1 \leq u_2 \leq t} dX_{u_1}^1 dX_{u_2}^1 = \int_s^t \int_s^{u_2} b_1^2 \, du_1 \, du_2 \\ &= b_1^2 \int_s^t (u_2 - s) \, du_2 = \frac{b_1^2 (t-s)^2}{2} \end{aligned}$$

$$S(X)_{s,t}^{(i_1,...,i_k)} = \frac{b_{i_1} ... b_{i_k} (t-s)^k}{k!}.$$

With Chen's identity and Example 4 the following algorithm would give us the signature of a discrete data stream, i.e. a sequence of points.

1. Interpolate data linearly to obtain a continuous path.
2. Compute piecewise signatures of linear sections of the path.
3. Compute the signature of the concatenated paths using Chen's identity.

The benefit of this procedure is, that we never have to calculate any integral by methods involving numerical approximations. An implementation of this procedure can be found in the Python library *iisignature* by J. Reizenstein and B. Graham [14].

2.5 Path uniqueness

In this chapter we have extensively studied the path signature and its properties. Especially in light of the invariances it is clear that we can never find one unique path that generates one particular signature. But apart from those invariances (time reparametrization, translation) it can be shown that the signature is actually able to completely describe and determine a path that never crosses itself, i.e. does not contain any loops. This was proven by B. Hambly and T. Lyons [8] and we summarize this result in the following theorem as it is formulated by A. Fermanian in [6]:

Theorem 2.12 (Path uniqueness)**.** *Assuming $X \in BV(\mathbb{R}^d)$ contains one monotone coordinate, then $S(X)$ characterizes X up to translations and reparametrizations.*

Remark 4 (Rough paths). Throughout this chapter we assumed our paths to be of bounded variation. The reason for this was to ensure existence of the Riemann-Stieltjes integral in the definition of the signature. Indeed the signature can also be defined for rougher paths that are not of bounded variation, e.g. Brownian Motion. However, this requires a different definition of iterated integrals. Further information can be found in [12].

2.6 Approximation of continuous functions

Eventually, we will present the theorem that motivates the next chapter of this thesis where we want to perform linear regression on the truncated signature. It also explains why the signature can be called the equivalent to polynomials, only in the path-space. We know that polynomials can approximate continuous functions arbitrarily well on compact sets. A similar result found by F.J. Király and H. Oberhauser in [11] is true for continuous functions of the path approximated by linear functions of the signature.

Proposition 2.13. *Let $D \subset BV(\mathbb{R}^d)$ be a compact set of paths that have at least one monotone coordinate and such that for any $X \in D$, $X_0 = 0$. Let $f : D \to \mathbb{R}$ be continuous. Then for every $\epsilon > 0$, there exists $m^* \in \mathbb{N}$, $\beta^* \in R^{s_d(m^*)}$, such that for any*

$X \in D$

$$|f(X) - \langle \beta^*, S^{m^*}(X) \rangle| \leq \epsilon, \tag{2.4}$$

where $\langle \cdot, \cdot \rangle$ *denotes the Euclidean scalar product on* $\mathbb{R}^{s_d(m^*)}$ *and* $s_d(m^*)$ *is the length of the truncated signature defined in (2.3).*

Remark 5. Of course, by now we understand that requiring one monotone component is just a sufficient condition to ensure that the path is not crossing itself. The condition $X_0 = 0$, on the other hand, is our way to deal with the translation invariance.

This proposition demonstrates that we can approximate any real valued continuous function on a compact set of paths arbitrarily well by linear functions of the signature, if we consider a high enough order of truncation. With this knowledge in mind we are well prepared to move to the next chapter, where the regression model will be explained.

Chapter 3

Signature regression model

After having presented the path signature and some of its properties, in this chapter the aim will be to present the model at the core of performing linear regression on the signature – the signature linear model [6].

3.1 Presentation of the model

We want to approximate the relationship between a random variable $Y \in \mathbb{R}$ and random path $X \in BV(\mathbb{R}^d)$. In the setting of a classical linear regression we would assume that the response is the sum of some linear combination of the (scalar) regressors and some random disturbance term ϵ, i.e.

$$Y = X\beta + \epsilon.$$

In our model the regressors are not scalar but functional covariates therefore it is natural to consider not only linear functions, but even continuous functions:

$$Y = f(X) + \epsilon$$

or

$$\mathbb{E}\left[Y \middle| X\right] = f(X)$$

for some continuous function $f : BV(\mathbb{R}^d) \to \mathbb{R}$.

Motivated by Proposition 2.13 we can bring this very general model back to a form where linear regression can be applied. Proposition 2.13 states that on a compact set

continuous functions of paths can be approximated arbitrarily well by linear functions of the truncated signature. Therefore in our model, we will assume that there exists indeed some true truncation level m^* and some true parameter $\beta^* \in R^{s_d(m^*)}$, such that

$$\mathbb{E}\left[Y|\,X\right] = f(X) = \langle \beta^*, S^{m^*}(X) \rangle \quad \text{and} \quad \text{Var}(Y|X) \leq \sigma^2 < \infty, \tag{3.1}$$

for some $\sigma^2 > 0$.

Remark 6. Of course there is a difference between the statement of Proposition 2.13, i.e. "approximate arbitrarily well" and imposing equation (3.1). However, for our purposes this subtlety will not be crucial. Since we are estimating the relation between Y and X, we might as well ignore the approximation error $\tilde{\epsilon}$ between $f(X)$ and $\langle \beta^*, S^{m^*}(X) \rangle$ and regard it as one part of the disturbance term ϵ of the regression model:

$$\begin{aligned} Y &= f(X) + \epsilon_{old} = \langle \beta^*, S^{m^*}(X) \rangle + \tilde{\epsilon} + \epsilon_{old} \\ &= \langle \beta^*, S^{m^*}(X) \rangle + \epsilon_{new}. \end{aligned}$$

This model is what A. Fermanian calls the "signature linear model" [6]. Since the signature's first entry is always 1, this model naturally contains an intercept. Furthermore it contains two parameters, i.e. the truncation order of the signature m^* and the coefficients for the signature entries β^*. The truncation order m^* is a key quantity, because it not only specifies the dimension of $\beta^* \in R^{s_d(m^*)}$, it also determines the computational costs of computing the path signature whose length increases exponentially as m^* grows.

For this reason we are interested in some methodology to estimate the truncation order in a sensible way, before searching for the β coefficient. A. Fermanian [6] imposes the technique of penalized empirical risk minimization. The results in the next chapter explain, why this might be an appropriate choice, but let us have a look at the procedure first.

3.2 Choosing the truncation order

Given a set of $n \in \mathbb{N}$ i.i.d. observations of our signal and response pair (X,Y), $D = \{(X_1, Y_1), ..., (X_n, Y_n)\}$ the questions is, for which truncation order we want to estimate the model in (3.1). To formulate an estimator we need to introduce some quantities first. Let $m \in \mathbb{N}$ be some fixed truncation order and let $\alpha > 0$ be a positive

real number. We denote the closed $s_d(m)$ dimensional α ball around 0 by

$$B_{m,\alpha} = \{\beta \in \mathbb{R}^{s_d(m)} | \|\beta\| \leq \alpha\},$$

where $\|\cdot\|$ is the Euclidean norm in the suitable dimension. Embedding $\mathbb{R}^{s_d(m)}$ into $\mathbb{R}^{s_d(m+1)}$ by setting all unused components to 0, we see that the ascending order of α balls is a nested sequence of sets:

$$B_{0,\alpha} \subset B_{1,\alpha} \subset ... \subset B_{m,\alpha} \subset B_{m+1,\alpha} \subset ...$$

Let us now fix some $\alpha > 0$. The first assumption that we make is that the true coefficient $\beta^*_{m^*}$ lies in one of those balls. This is one of two assumptions that we will need to justify this procedure later in Chapter 4 with a pleasant theoretical result. We write

$$(H_\alpha) : \beta^*_{m^*} \in B_{m^*,\alpha}. \tag{3.2}$$

Furthermore, the theoretical risk of some truncation order m and coefficient β is defined as the expected squared deviation of Y from the prediction obtained by the truncated path signature of X and coefficient β:

$$R_m(\beta) := \mathbb{E}\left[(Y - \langle \beta, S^m(X)\rangle)^2\right],$$

where $\langle \cdot, \cdot \rangle$ is the Euclidean scalar product in $\mathbb{R}^{s_d(m)}$.

Subsequently, the minimal theoretical risk $L(m)$ is the minimal value across all β's in some previously fixed α-ball:

$$L(m) := \inf_{\beta \in B_{m,\alpha}} R_m(\beta) = R_m(\beta^*_m),$$

where $\beta^*_m := \text{argmin}_{\beta \in B_{m,\alpha}} R_m(\beta)$.

Finally, for both quantities we denote their empirical counterparts obtained from $D = \{(X_1, Y_1), ..., (X_n, Y_n)\}$ as

$$\hat{R}_{m,n}(\beta) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \langle \beta, S^m(X_i)\rangle)^2$$

and

$$\hat{L}_n(m) = \min_{\beta \in B_{m,\alpha}} \hat{R}_{m,n}(\beta) = \hat{R}_{m,n}(\hat{\beta}),$$

where $\hat{\beta}$ is a point in $B_{m,\alpha}$ where the minimum is attained.

Following A. Fermanian's approach from [6] we can minimize $\hat{R}_{m,n}(\beta)$ over $B_{m,\alpha}$ by performing a Ridge regression

$$\hat{\beta}^{Ridge} = \operatorname*{argmin}_{\beta} \left\{ \sum_{i=1}^{n} (Y_i - \langle \beta, S^m(X_i) \rangle)^2 + \lambda \|\beta\|^2 \right\} \tag{3.3}$$

with a certain regularization parameter $\lambda > 0$ that corresponds to α. This is very useful for us, because the regularization term in the objective function of the Ridge regression will additionally protect us from overfitting when the number of observations is small compared to the number of coefficients to be estimated. A problem that we are likely to face in practice since the number of entries of the signature grows exponentially as the truncation order increases.

This means that for every truncation order m we are able to find the best estimator $\hat{\beta}$ for estimating Y by a linear function of the signature of X. In order to find a good estimator $\hat{m}$ for the true truncation order m^* we will consider the sequence of minimal empirical risks $(\hat{L}(m))_{m \in \mathbb{N}}$. Since the α-balls are nested, the sequence will be decreasing, due to the increasing number of parameters to fit the data to, as the truncation order increases. By adding a penalization term that makes models with higher numbers of coefficients less attractive for us, we can formulate an estimator for m^* as the trade-off between minimal empirical risk and penalization term

$$\hat{m} = \min \left(\operatorname*{argmin}_{m \in \mathbb{N}} (\hat{L}_n(m) + pen_n(m)) \right), \tag{3.4}$$

where the concrete form of the increasing penalization function $pen_n(m)$ will be specified in Chapter 4.

With this estimator for the truncation order at hand, it will be straight forward to choose $\hat{\beta}_{\hat{m}}$ for the respective truncation level by Ridge regression. In the next chapter we will discuss two results by A. Fermanian [6], which make this procedure of choosing $\hat{m}$ particularly attractive.

Chapter 4

Performance bounds

In this chapter we want to investigate two performance bounds of the signature linear model that were originally established by A. Fermanian in [6]. It is our aim to rewrite the proofs of the original, rather dense work in greater detail, with some helpful comments whilst correcting some inaccuracies.

4.1 Theorem: Probability of choosing a wrong truncation order

As formulated in the previous chapter, the model contains one meta-parameter m, which is the truncation order of the signature. The underlying assumption from the last chapter, see (3.4), was that there exists a smallest true truncation order m^* such that

$$\mathbb{E}\left[Y \middle| X\right] = f(X) = \langle \beta^*, S^{m^*}(X) \rangle. \tag{4.1}$$

This means on one hand, we assume that the expected value of Y conditional on all information we have about X can be written as some continuous function f depending only on X. For the second equation on the other hand, we rely on the findings of F.J. Király and H. Oberhauser [11] that a continuous function of a path can be approximated arbitrarily well by a linear function of the truncated signature, see Proposition 2.13.

As already explained in Remark 6 our assumption is indeed stronger, because we assume that there exists m^* and a corresponding linear function of the truncated signature $\langle \beta^*, S^{m^*}(X) \rangle$ to actually match $f(X)$ and not only approximate arbitrarily

well. In order to estimate the coefficients β^* of this linear function, we need to choose the truncation order beforehand, as described in Chapter 3, see (3.4).

Now the question arises, what the probability is that we choose the right or a wrong truncation order for our regression model in (3.1). We will see that under certain circumstances this probability decreases exponentially as the sample size increases, i.e. for some $0 \leq \rho \leq \frac{1}{2}$

$$\mathbb{P}(\hat{m} \neq m^*) \leq C_1 \exp(-C_2 n^{1-2\rho}). \tag{4.2}$$

Besides the assumption, that the β lie inside some α-Ball (H_α) we need the additional assumption (H_K) that there exist two real numbers $K_Y > 0$ and $K_X > 0$, such that almost surely $|Y| \leq K_Y$ and $||X||_{TV} \leq K_X$.

Although this might not be very satisfying from a theoretical point of view, in practice when dealing with real world data points, (H_K) isn't a restrictive assumption because we would almost certainly be able to figure out some bounds that should not be exceeded by any stretch of the imagination, e.g. $K_Y = 100$ for the earth's surface temperature measured in °C.

For further convenience we define the constant K to be

$$K = 2(K_Y + \alpha e^{K_X}) e^{K_X}. \tag{4.3}$$

The statement we want to prove can then be formulated as the following theorem and is similar to [6, Thm. 4.1] in the original work by Fermanian. The differences are mentioned in Remark 7.

Theorem 4.1. *Let $K_{\text{pen}} > 0$, $0 < p < \frac{1}{2}$ and*

$$\text{pen}_n(m) = K_{\text{pen}} n^{-p} \sqrt{s_d(m)}. \tag{4.4}$$

Let n_0 be the smallest integer satisfying

$$n_0^{\tilde{\rho}} \geq \Bigg((864 K \alpha \sqrt{\pi} + K_{\text{pen}}) \times \left(\frac{2\sqrt{s_d(m^*+1)}}{L(m^*-1) - L(m^*)} + \frac{\sqrt{s_d(m^*+1)}}{K_{pen}\left(\sqrt{s_d(m^*+1} - \sqrt{s_d(m^*)}\right)} \right) \Bigg) \tag{4.5}$$

where $\tilde{\rho} = \min(\rho, \frac{1}{2} - \rho)$. *Then under the assumptions* (H_α) *and* (H_K), *for any* $n \geq n_0$,

$$\mathbb{P}(\hat{m} \neq m^*) \leq C_1 \exp(-C_2 n^{1-2\rho}), \tag{4.6}$$

where the constants C_1 *and* C_2 *are defined by*

$$C_1 = 74 \sum_{m>0} e^{-C_3 s_d(m)} + 148 m^*, \tag{4.7}$$

where

$$C_3 = \frac{K_{pen}^2 d^{m^*+1}}{128 s_d(m^*+1)(72K^2\alpha^2 + K_Y^2)}, \tag{4.8}$$

and

$$C_2 = \frac{1}{16(1152K^2\alpha^2 + K_Y^2} \min\left(\frac{K_{pen}^2 d^{m^*+1}}{8 s_d(m^*+1)}, L(m^*-1) - L(m^*) \right). \tag{4.9}$$

This theorem proves that the probability of choosing a wrong truncation order can be bounded from above for every sample size of n_0 or more points, when choosing the parameters accordingly and estimating the truncation order as described in the previous chapter. Indeed taking the limit $n \to \infty$ the theorem assures that the probability of choosing a wrong truncation order vanishes exponentially.

At the same time this bound depends on many different quantities and it makes sense to stop here for a minute and think about their meaning and the interactions between each other before proceeding with the actual proof.

- K_{pen} is an arbitrary regularization parameter in the penalization function. It is easy to see that there exists an optimal K_{pen}^* that minimizes the parameter n_0 (or equivalently $n_0^{\tilde{p}}$ because of monotonicity). Simplifying the notation with $a, b, c > 0$ such that

$$a = 432K\alpha\sqrt{\pi}, \quad b = \frac{2\sqrt{s_d(m^*+1)}}{L(m^*-1) - \sigma^2}, \quad c = \frac{\sqrt{2s_d(m^*+1)}}{K_{pen}\sqrt{d^{m^*+1}}}$$

and taking the derivative gives

$$\frac{\partial}{\partial K_{pen}} n_0^{\tilde{p}} = 0$$

$$\begin{aligned}
\frac{\partial}{\partial K_{pen}}(a + K_{pen})(b + \frac{c}{K_{pen}}) &= 0 \\
b - \frac{ac}{K_{pen}^2} &= 0 \\
b &= \frac{ac}{K_{pen}^2} \\
K_{pen} &= \sqrt{\frac{ac}{b}} = \\
&= \sqrt{\frac{432K\alpha\sqrt{\pi}\sqrt{2s_d(m^*+1)}(L(m^*-1)-\sigma^2)}{2\sqrt{s_d(m^*+1)}\sqrt{d^{m^*+1}}}} > 0
\end{aligned}$$

which is indeed a minimum since

$$\begin{aligned}
\frac{\partial^2}{\partial K_{pen}^2} n_0^{\tilde{p}}\Big|_{K_{pen}=\sqrt{\frac{ac}{b}}} &= \frac{\partial^2}{\partial K_{pen}^2}\left(b - \frac{ac}{K_{pen}^2}\right)\Big|_{K_{pen}=\sqrt{\frac{ac}{b}}} \\
&= 2\frac{ac}{K_{pen}^3}\Big|_{K_{pen}=\sqrt{\frac{ac}{b}}} \\
&= 2\frac{acb^{\frac{3}{2}}}{(ac)^{\frac{3}{2}}} = 2\sqrt{\frac{b^3}{ac}} > 0.
\end{aligned}$$

Unfortunately a,b and c contain many theoretical quantities that are not know in a real-world application like σ^2 or m^* (remember this is the value we need to estimate). In [6], A. Fermanian suggests to make use of the slope heuristics method of L. Birgé and P. Massart [1], a strategy that allows to find good penalization parameters for model selection tasks with a penalized least-squares type criterion, which corresponds to our setting.
Furthermore C_2, and C_3 increase linearly with K_{pen}^2. Therefore for the error bound larger values of K_{pen} would be desirable, if getting enough data is not a problem.

- p is an arbitrary parameter between 0 and $\frac{1}{2}$ that appears in the upper bound of $\mathbb{P}(\hat{m} \neq m^*)$ where it determines the speed of convergence as n increases. The closer p is to 0, the faster the probability vanishes. But at the same time p interacts with n_0 through $\tilde{p} = \min(p, \frac{1}{2} - p)$. Here on the other hand n_0 would be minimal, if p is close to $\frac{1}{4}$. So one has to carefully find a balance when choosing p. Furthermore n^{-p} appears in the penalization function and hence influences the penalization for models chosen from larger or smaller sample sizes.

- $pen_n(m)$ is the penalization function used for estimating $\hat{m}$. In (4.4) we see that it scales linearly with K_{pen}. A larger sample size is rewarded with the factor n^{-p}. As m

grows the penalization function grows proportional to the square root of the dimension of a d-dimensional signature truncated at order m, i.e. $\sqrt{s_d(m)} \sim \mathcal{O}(d^{\frac{m}{2}})$.

- d is the dimension of the path X. From Chapter 2 we already know that $s_d(m) \sim \mathcal{O}(d^m)$ for large d or m. This means that all others equal the constants C_2 and C_3 stay of same order as d increases, while in $C_1 = 74 \sum_{m>0} e^{-C_3 s_d(m)} + 148m^*$ the first summand decreases. Therefore the error bound does not directly suffer from increasing the path's dimension, if indeed all other quantities stay the same. However n_0 will increase proportional to $\mathcal{O}(d^{\frac{m^*+1}{2\tilde{p}}})$ as d increases, such that exponentially more data is needed to estimate $\hat{m}$.
- m^* is the true truncation order in (4.1). As m^* increases the order of C_2 and C_3 stays the same again, but this time C_1 increases linearly $C \sim 148m^*$. n_0 grows at a rate of $\mathcal{O}(d^{\frac{m^*+1}{2\tilde{p}}})$. This means that estimating the truncation order becomes increasingly difficult and more data is needed, if the true truncation order m^* is large and therefore the relationship between $f(X)$ and $S^{m^*}(X)$ is more complex (i.e. the dimension of β^* is higher).
- α is the radius of the ball in which we suspect β^* to lie. Increasing α increases n_0 and C_1 (C_1 is affected since C_3 decreases), while C_2 decreases, which means the error bound becomes weaker. This is not surprising, since a larger α enlarges the space of possible β^*. Therefore the relationship between $f(X)$ and $S^{m^*}(X)$ is more complex, m^* is harder to estimate and it requires more data.
- $L(m^* - 1) - \sigma^2$: L is the minimal theoretical risk for a certain truncation order and σ^2 (the upper bound of) $= Var(Y|X)$ both presented in Chapter 3, see (3.1). The term $L(m^* - 1) - \sigma^2 = L(m^* - 1) - L(m^*)$ can be interpreted as the difference of risk, between a model truncated at m^* and a smaller model. This quantity is strictly positive. If it gets small, which means that a model truncated at $m^* - 1$ performs similar to a model truncated at m^*, then C_2 decreases loosening the error bound and at the same time n_0 increases. If the difference is big, i.e. the models differ more in terms of theoretical risk, then n_0 decreases, the error bound becomes stricter and estimation is easier.
- K is driven by K_Y, K_X and α from assumptions (H_K) and (H_α). A large K increases both n_0 and the error bound, therefore one would typically try to keep K_Y, K_X, α and hence K as small as possible, while still fulfilling the corresponding assumptions.

Remark 7 (Differences between the theorems in this work and [6]). The general result of the theorem is the same in both cases, but there are some differences in the constants, i.e. the lower bound of n_0, C_2 and C_3:

- Considering an incorrect ϵ (see Remark 12) yields 432 in n_0. We use a correct ϵ and get 864.
- In this thesis we write $\sqrt{s_d(m^*+1)} - \sqrt{s_d(m^*)}$ instead of $\frac{\sqrt{d^{m^*+1}}}{\sqrt{2}}$. The reason is, that Fermanian used a faulty inequality to get the latter expression (see Remark 10), so we stick to the former one. This change concerns n_0, C_2 and C_3.
- Instead of $L(m^*-1) - L(m^*)$ Fermanian writes $L(m^*-1) - \sigma^2$. It is true that $L(m^*) \leq \sigma^2$ and therefore $L(m^*-1) - \sigma^2 \leq L(m^*-1) - L(m^*)$. But we can not ensure that $L(m^*-1) - \sigma^2$ is positive. However, if $L(m^*-1) - \sigma^2$ is indeed negative, we run into trouble, see Remark 11. This affects n_0 and C_2.
- In C_2 the term $L(m^*-1) - L(m^*)$ is squared and some numbers deviate from the C_2 in the original work. The difference stems from an incorrect application of Proposition 4.17 in the proof of Theorem 4.1, see Remark 13.
- In the original work the term K_Y often appears squared. In fact K_Y should indeed be taken to the fourth power. Probably Fermanian applied Hoeffding's inequality not correct twice, see Remark 9. The adjusted values can be found in C_2 and C_3.

4.2 Proof of the Theorem 4.1

Having dealt with the meaning of the quantities appearing in the theorem, in the following we want to investigate the actual proof proposed by A. Fermanian [6]. The main idea is the following:

- We want to bound the probability of choosing a wrong truncation order from above by some quantity that vanishes as $n \to \infty$:

$$\mathbb{P}(\hat{m} \neq m^*) \leq \dots \tag{4.10}$$

- In order to do this, we split the probability in two sums. One sum representing the probabilities of choosing $\hat{m} > m^*$ and the other sum choosing $\hat{m} < m^*$,

$$\mathbb{P}(\hat{m} \neq m^*) = \sum_{m>m^*} \mathbb{P}(\hat{m} = m) + \sum_{m<m^*} \mathbb{P}(\hat{m} = m). \tag{4.11}$$

- To handle both sums we will find upper bounds of the form

$$\mathbb{P}(\hat{m} = m) \leq ... \tag{4.12}$$

for $m < m^*$ as well as $m > m^*$.

- They will be derived by establishing a relation between

$$\mathbb{P}(\hat{m} = m) \quad \text{and} \quad \mathbb{P}(\sup_{\beta \in B_{m,\alpha}} |\hat{R}_{m,n}(\beta) - R_m(\beta)| \geq \dots) \tag{4.13}$$

  and showing that

$$Z_{m,n}(\beta) := \hat{R}_{m,n}(\beta) - R_m(\beta) \tag{4.14}$$

  is a subgaussian process, which enables us to use an inequality of the form

$$\mathbb{P}\left(\sup_{\beta \in B_{m,\alpha}} (Z_{m,n}(\beta) - Z_{m,n}(\beta_0)) \geq \dots\right) \leq \dots \quad \text{for any } \beta_0. \tag{4.15}$$

To study this proof with rigor, we need to introduce the following tools that can also be found in [9].

Definition 4.2 (**Subgaussian random variables**). *A random variable X is called* ***subgaussian*** *if its moment generating function ϕ fulfills*

$$\phi(\lambda) \leq \exp(\lambda^2 \sigma^2 / 2) \quad \textit{for all } \lambda \in \mathbb{R}.$$

The constant σ^2 is called variance-proxy and X can also be called σ^2-subgaussian.

One helpful means to show that bounded random variables must always be subgaussian is Hoeffding's lemma:

Lemma 4.3 (**Hoeffding Lemma**). *Let $a \leq X \leq b$ a.s. for some $a, b \in \mathbb{R}$. Then*

$$\mathbb{E}[e^{\lambda(X - \mathbb{E}[X])}] \leq e^{\lambda^2 (b-a)^2 / 8},$$

i.e. by Definition 4.2 X is $(b-a)^2/4$-subgaussian.

Proof. See [9, Lem. 3.6].

This already gives some intuition, why assumption (H_α) and (H_K) will be helpful: They ensure some random variables to be bounded by assumption and, therefore, also subgaussian. Using subgaussian random variables, we will be able to show that a process is subgaussian:

Definition 4.4 (**Subgaussian process**). *A random process* $\{X_t\}_{t\in T}$ *on the metric space* (T,d) *is called* ***subgaussian*** *if* $\mathbb{E}[X_t]=0$ *and*

$$\mathbb{E}[e^{\lambda(X_t-X_s)}] \leq e^{\lambda^2 d(s,t)^2/2} \quad \textit{for all } t,s\in T, \lambda\geq 0$$

This subgaussian property of a process is interesting for us, because following the explanations of R. v. Handel in [9] this property is - up to some constants - equivalent to an inequality of the form

$$\mathbb{P}\left(|X_t-X_s|\geq xd(t,s)\right)\leq Ce^{-x^2/C}.$$

which can heuristically be interpreted as a probabilistic version of being Lipschitz, i.e. $\exists C>0 : |X_t-X_s|\leq Cd(t,s)$ for all $t,s\in T$.

We will use of a special version of this inequality borrowed from R. v. Handel [9]. But in order to introduce the theorem we need to understand the concept of ϵ-net, covering number, diameter of a metric space, as well as the concept of separable random processes.

Definition 4.5 (**Separable process**). *A random process* $\{X_t\}_{t\in T}$ *is called separable if there is a countable set* $T_0\subseteq T$ *such that*

$$X_t\in \lim_{\substack{s\to t\\ s\in T_0}} X_s \quad \textit{for all } t\in T \quad a.s.$$

Definition 4.6 (**Diameter**). *The diamater of a metric space* (T,d) *is defined as*

$$diam(T):=\sup_{t,s\in T} d(t,s)$$

Definition 4.7 (ϵ**-net and covering number**). *A set* N *is called an* ϵ*-net for* (T,d) *if for every* $t\in T$ *there exists* $\pi(t)\in N$ *such that* $d(t,\pi(t))\leq\epsilon$.
The smallest cardinality of an ϵ*-net for* (T,d) *is called the covering number*

$$N(T,d,\epsilon):=\inf\{|N| : N \textit{ is an } \epsilon\textit{-net for } (T,d)\}.$$

Having in mind the m-dimensional α-ball $B_{m,\alpha}$ as space T, equipped with the metric d induced by the Euclidean norm, the ϵ-covering number is the smallest number of balls with radius ϵ needed to cover T. The diameter of T would obviously be $diam(T)=2\alpha$.

Now we have all definitions ready to introduce the chaining tail inequality [9, Thm. 5.29] that we will be using in the proof:

Theorem 4.8 (**Chaining tail inequality**). *Let $\{X_t\}_{t\in T}$ be a separable subgaussian process on the metric space (T, d). Then for all $t_0 \in T$ and $x \geq 0$*

$$\mathbb{P}\left(\sup_{t\in T}(X_t - X_{t_0}) \geq C\int_0^\infty \sqrt{\log N(T,d,\epsilon)}d\epsilon + x\right) \leq Ce^{-x^2/(C\,\mathrm{diam}(T)^2)},$$

where $C < \infty$ is a constant.

Proof. See [9, Thm. 5.29].

For fixed (X, Y), their identically distributed copies $(X_i, Y_i)_{i=1,\dots,n}$, a fixed truncation order m and $\beta \in B_{m,\alpha}$ recall from Chapter 3 the definitions of the theoretical risk

$$R_m(\beta) := \mathbb{E}\left[(Y - \langle\beta, S^m(X)\rangle)^2\right], \tag{4.16}$$

the minimal theoretical risk

$$L(m) := \inf_{\beta\in B_{m,\alpha}} R_m(\beta) = R_m(\beta_m^*) \tag{4.17}$$

where $\beta_m^* = \mathrm{argmin}_{\beta\in B_{m,\alpha}} R_m(\beta)$, as well as their empirical counterparts

$$\hat{R}_{m,n}(\beta) = \frac{1}{n}\sum_{i=1}^n (Y_i - \langle\beta, S^m(X_i)\rangle)^2 \tag{4.18}$$

and

$$\hat{L}_n(m) = \min_{\beta\in B_{m,\alpha}} \hat{R}_{m,n}(\beta) = \hat{R}_{m,n}(\hat{\beta}), \tag{4.19}$$

where $\hat{\beta}$ is a point in $B_{m,\alpha}$ where the minimum is attained.

The proof starts with two lemmata. The first one showing that for any truncation order $m \in \mathbb{N}$ the absolute difference between empirical and theoretical **minimal risk** can be bounded by the maximal absolute difference of empirical and theoretical **risk** over $B_{m,\alpha}$. The second lemma shows that this bound can be used to restrict $\mathbb{P}(\hat{m} = m)$ in the case $m > m^*$.

Lemma 4.9. *For any $m \in \mathbb{N}$,*

$$|\hat{L}_n(m) - L(m)| \leq \sup_{\beta \in B_{m,\alpha}} |\hat{R}_{m,n}(\beta) - R_m(\beta)|.$$

Proof. Plugging in the definitions and adding $\pm\hat{R}_{m,n}(\beta_m^*)$ yields

$$\begin{aligned}\hat{L}_n(m) - L(m) &= \hat{R}_{m,n}(\hat{\beta}_m) - R_m(\beta_m^*) \\ &= \hat{R}_{m,n}(\hat{\beta}_m) - \hat{R}_{m,n}(\beta_m^*) + \hat{R}_{m,n}(\beta_m^*) - R_m(\beta_m^*).\end{aligned}$$

Further since $\hat{\beta}_m$ minimizes $\hat{R}_{m,n}$ over $B_{m,\alpha}$, we know that $\hat{R}_{m,n}(\hat{\beta}_m) - \hat{R}_{m,n}(\beta_m^*) \leq 0$. By dropping this term we get

$$\begin{aligned}\hat{L}_n(m) - L(m) &\leq \hat{R}_{m,n}(\beta_m^*) - R_m(\beta_m^*) \\ &\leq \sup_{\beta \in B_{m,\alpha}} |\hat{R}_{m,n}(\beta) - R_m(\beta)|.\end{aligned}$$

For $L(m) - \hat{L}_n(m)$ we extend with $\pm R_m(\hat{\beta}_m)$ and since β_m^* minimizes R_m we can drop $R_m(\beta_m^*) - R_m(\hat{\beta}_m) \leq 0$:

$$\begin{aligned}L(m) - \hat{L}_n(m) &= -\hat{R}_{m,n}(\hat{\beta}_m) + R_m(\beta_m^*) \\ &= -\hat{R}_{m,n}(\hat{\beta}_m) - \mathbb{R}_m(\hat{\beta}_m) + \mathbb{R}_m(\hat{\beta}_m) + R_m(\beta_m^*) \\ &\leq -\hat{R}_{m,n}(\hat{\beta}_m) + R_m(\hat{\beta}_m) \\ &\leq \sup_{\beta \in B_{m,\alpha}} |\hat{R}_{m,n}(\beta) - R_m(\beta)|,\end{aligned}$$

which proves the lemma. □

Lemma 4.10. *For any $m > m^*$, it holds that*

$$\mathbb{P}(\hat{m} = m) \leq \mathbb{P}\left(2 \sup_{\beta \in B_{m,\alpha}} |\hat{R}_{m,n}(\beta) - R_m(\beta)| \geq pen_n(m) - pen_n(m^*)\right).$$

Proof. Since according to Chapter 3 $\hat{m}$ is chosen such that

$$\hat{m} = \min\left(\operatorname*{argmin}_{m \in \mathbb{N}}(\hat{L}_n(m) + pen_n(m))\right),$$

it is clear that $\hat{L}_n(m) + pen_n(m) \leq \hat{L}_n(m^*) + pen_n(m^*)$ is a necessary condition for $\hat{m} = m$ and therefore

$$\begin{aligned}\mathbb{P}(\hat{m} = m) &\leq \mathbb{P}\left(\hat{L}_n(m) + pen_n(m) \leq \hat{L}_n(m^*) + pen_n(m^*)\right) \\ &= \mathbb{P}\left(\hat{L}_n(m^*) - \hat{L}_n(m) \geq pen_n(m^*) - pen_n(m)\right).\end{aligned}$$

Furthermore, we already established that $m \mapsto L(m)$ is a decreasing function whose minimum is attained at $m = m^*$, hence $L(m^*) - L(m) \leq 0$ and

$$\begin{aligned}\hat{L}_n(m^*) - \hat{L}_n(m) &= \hat{L}_n(m^*) - L(m^*) + L(m^*) - L(m) + L(m) - \hat{L}_n(m) \\ &\leq \hat{L}_n(m^*) - L(m^*) + L(m) - \hat{L}_n(m) \\ &\leq |\hat{L}_n(m^*) - L(m^*)| + |L(m) - \hat{L}_n(m)| \\ &\leq \sup_{\beta \in B_{m^*,\alpha}} |\hat{R}_{m^*,n}(\beta) - R_{m^*}(\beta)| + \sup_{\beta \in B_{m,\alpha}} |\hat{R}_{m,n}(\beta) - R_m(\beta)|,\end{aligned}$$

where we used Lemma 4.9 in the last line.
Finally, since the α-balls of growing dimensions are nested, we know that for all $m > m^*$ $B_{m^*,\alpha}$ is a subset of $B_{m,\alpha}$ and we can further simplify the expression

$$\hat{L}_n(m^*) - \hat{L}_n(m) \leq \sup_{\beta \in B_{m,\alpha}} |\hat{R}_{m,n}(\beta) - R_m(\beta)|,$$

completing the proof. □

As already mentioned previously, we now introduce the centered empirical risk for signatures truncated at m:

$$Z_{m,n}(\beta) = \hat{R}_{m,n}(\beta) - R_m(\beta) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \langle \beta, S^m(X_i)\rangle)^2 - \mathbb{E}[(Y - \langle \beta, S^m(X)\rangle)^2] \tag{4.20}$$

We will prove in the following lemma that this process is subgaussian for $\beta \in B_{m,\alpha}$ and some appropriate distance.

Lemma 4.11. *Under the assumptions (H_α) and (H_K), for any $m \in \mathbb{N}$, the process $(Z_{m,n}(\beta))_{\beta \in B_{m,\alpha}}$ is subgaussian for the distance*

$$D(\beta, \gamma) = \frac{K}{\sqrt{n}}\|\beta - \gamma\|, \tag{4.21}$$

where K is a constant defined by equation (4.3).

Proof. Since Y and Y_i as well as X and X_i are distributed identically it is clear that $\mathbb{E}[Z_{m,n}(\beta)] = 0$, for any $\beta \in B_{m,\alpha}$. For the random variables X, Y we define an auxiliary function

$$\begin{aligned}\ell_{(X,Y)} &: B_{m,\alpha} \to \mathbb{R}\\ \ell_{(X,Y)}(\beta) &= (Y - \langle \beta, S^m(X)\rangle)^2.\end{aligned}$$

With $K = 2(K_Y + \alpha e^{K_X})e^{K_X}$ from (4.3) we will show that this function is K-Lipschitz. Let $\beta, \gamma \in B_{m,\alpha}$,

$$\begin{aligned}|\ell_{(X,Y)}(\beta) - \ell_{(X,Y)}(\gamma)| &= |(Y - \langle \beta, S^m(X)\rangle)^2 - (Y - \langle \gamma, S^m(X)\rangle)^2|\\ &\leq 2\max(|(Y - \langle \beta, S^m(X)\rangle)^2|, |(Y - \langle \gamma, S^m(X)\rangle)^2|)\\ &\quad \times |\langle \beta - \gamma, S^m(X)\rangle|\\ &\leq 2\max(|(Y - \langle \beta, S^m(X)\rangle)^2|, |(Y - \langle \gamma, S^m(X)\rangle)^2|)\\ &\quad \times \|\beta - \gamma\|\|S^m(X)\|.\end{aligned}$$

The first inequality holds since $|a^2 - b^2| = |(a+b)(a-b)| \leq 2\max(|a|, |b|)|a-b|$, for the second inequality we applied the Cauchy-Schwartz inequality.

Further we can also bound both expressions inside the maximum. We use the triangle inequality, the Cauchy-Schwartz inequality and the fact that $|Y|$ is bounded by K_Y and $\|\beta\|$ (or $\|\gamma\|$) by α,

$$|Y - \langle \beta, S^m(X)\rangle| \leq |Y| + \|\beta\|\|S^m(X)\| \leq K_Y + \alpha\|S^m(X)\|.$$

By Proposition (2.5) we know that the truncated signature can be bounded too,

$$\|S^m(X)\| \leq e^{\|X\|_{TV}} \leq e^{K_X}.$$

Putting everything together we get,

$$|\ell_{(X,Y)}(\beta) - \ell_{(X,Y)}(\gamma)| \leq 2(K_Y + \alpha e^{K_X})e^{K_X}\|\beta - \gamma\| = K\|\beta - \gamma\|.$$

Now for fixed β and γ this random variable is bounded and we can use Hoeffding's lemma (4.3) to prove $\ell_{(X,Y)}(\beta) - \ell_{(X,Y)}(\gamma)$ is subgaussian with a variance proxy $K^2\|\beta - \gamma\|^2$.

For $\lambda \geq 0$ this yields

$$\mathbb{E}\left[e^{\lambda\left(\ell_{(X,Y)}(\beta)-\ell_{(X,Y)}(\gamma)-\mathbb{E}\left[\ell_{(X,Y)}(\beta)-\ell_{(X,Y)}(\gamma)\right]\right)}\right] \leq \exp\left(\frac{\lambda^2 K^2 \|\beta-\gamma\|^2}{2}\right).$$

We use this relationship to show that $\left(Z_{m,n}(\beta)\right)_{\beta\in B_{m,\alpha}}$ is a subgaussian process. We already established $\mathbb{E}\left[Z_{m,n}(\beta)\right] = 0$. According to the definition of a subgaussian process (4.4) it is left to show that

$$\mathbb{E}\left[e^{\lambda\left(Z_{m,n}(\beta)-Z_{m,n}(\gamma)\right)}\right] \leq e^{\lambda^2 D(\beta,\gamma)^2/2},$$

for $\beta, \gamma \in B_{m,\alpha}$, $\lambda \geq 0$ and a metric D on $B_{m,\alpha}$.

We see that $Z_{m,n}(\beta) - Z_{m,n}(\gamma)$ can be written as a function of $\ell_{(X_i,Y_i)}(\beta) - \ell_{(X_i,Y_i)}(\gamma)$,

$$\begin{aligned}
\mathbb{E}\left[e^{\lambda\left(Z_{m,n}(\beta)-Z_{m,n}(\gamma)\right)}\right] &= \mathbb{E}\Bigg[\exp\Bigg(\frac{\lambda}{n}\sum_{i=1}^{n}\Big(Y_i - \langle\beta, S^m(X_i)\rangle\Big)^2 - \mathbb{E}\Big[\Big(Y - \langle\beta, S^m(X)\Big)^2\Big] \\
&\quad - \Big(Y_i - \langle\gamma, S^m(X_i)\rangle\Big)^2 - \mathbb{E}\Big[(Y - \langle\gamma, S^m(X))^2\Big]\Bigg)\Bigg] \\
&= \mathbb{E}\Bigg[\exp\Bigg(\frac{\lambda}{n}\sum_{i=1}^{n}\ell_{(X_i,Y_i)}(\beta) - \ell_{(X_i,Y_i)}(\gamma) \\
&\quad - \mathbb{E}\Big[\Big(Y_i - \langle\beta, S^m(X_i)\Big)^2\Big] + \mathbb{E}\Big[\Big(Y_i - \langle\gamma, S^m(X_i)\Big)^2\Big]\Bigg)\Bigg] \\
&= \mathbb{E}\Bigg[\exp\Bigg(\frac{\lambda}{n}\sum_{i=1}^{n}\ell_{(X_i,Y_i)}(\beta) - \ell_{(X_i,Y_i)}(\gamma) \\
&\quad - \mathbb{E}\Big[\ell_{(X_i,Y_i)}(\beta) - \ell_{(X_i,Y_i)}(\gamma)\Big]\Bigg)\Bigg].
\end{aligned}$$

Using that the (X_i, Y_i) are i.i.d. and in combination with the previous results we get,

$$\begin{aligned}
&= \prod_{i=1}^{n}\mathbb{E}\Bigg[\exp\Bigg(\frac{\lambda}{n}\ell_{(X_i,Y_i)}(\beta) - \ell_{(X_i,Y_i)}(\gamma) \\
&\quad - \mathbb{E}\Big[\ell_{(X_i,Y_i)}(\beta) - \ell_{(X_i,Y_i)}(\gamma)\Big]\Bigg)\Bigg]
\end{aligned}$$

$$\leq \prod_{i=1}^{n} \exp\left(\frac{\lambda^2 K^2 \|\beta-\gamma\|^2}{2n^2}\right) = \exp\left(n\frac{\lambda^2 K^2 \|\beta-\gamma\|^2}{2n^2}\right)$$
$$= \exp\left(\frac{\lambda^2 K^2 \|\beta-\gamma\|^2}{2n}\right) = \exp\left(\frac{\lambda^2 D(\beta,\gamma)^2}{2}\right),$$

where we defined the distance D as $D(\beta,\gamma) = \frac{K\|\beta-\gamma\|}{\sqrt{n}}$. Since this is merely a scaled version of the Euclidean norm it is clear that $(B_{m,\alpha}, D)$ is a metric space and thus the proof is complete. □

Remark 8. Fermanian does not mention that the process $Z_{m,n}(\cdot)$ on $(B_{m,\alpha}, D)$ is separable, even though we will use this fact in the next part of the proof. However, it is not difficult to see that this is indeed the case:

Lemma 4.12. *The process $Z_{m,n}(\cdot)$ on $(B_{m,\alpha}, D)$ is separable.*

Proof. $B_{m,\alpha}$ is a subset of $\mathbb{R}^{s_d(m)}$ and we know that $\mathbb{Q}^{s_d(m)}$ is countable and dense in $\mathbb{R}^{s_d(m)}$. Therefore, using $B_{m,\alpha} \cap \mathbb{Q}^{s_d(m)}$ as a countable subset, we see that every $\beta \in B_{m,\alpha}$ can be approximated arbitrarily well by some sequence $(\beta_i)_{i\in\mathbb{N}} \subset B_{m,\alpha} \cap \mathbb{Q}^{s_d(m)}$. Since $e^x + e^{-x} = \sum_{n\in\mathbb{N}} \frac{x^n}{n!} + \sum_{n\in\mathbb{N}} \frac{(-x)^n}{n!} \geq \sum_{n \text{ even}} 2\frac{x^n}{n!} \geq x^2$, we can calculate

$$\begin{aligned}
\mathbb{E}\left[\left(Z_{m,n}(\beta) - Z_{m,n}(\beta_i)\right)^2\right] &\leq \mathbb{E}\left[e^{Z_{m,n}(\beta)-Z_{m,n}(\beta_i)} + e^{-Z_{m,n}(\beta)+Z_{m,n}(\beta_i)}\right] \\
&\leq \mathbb{E}\left[e^{Z_{m,n}(\beta)-Z_{m,n}(\beta_i)}\right] + \mathbb{E}\left[e^{-Z_{m,n}(\beta)+Z_{m,n}(\beta_i)}\right] \\
&\leq e^{\frac{K^2\|\beta-\beta_i\|}{2n}} + e^{\frac{K^2\|\beta_i-\beta\|}{2n}} \xrightarrow{i\to\infty} 0,
\end{aligned}$$

where we used the subgaussian property of $Z_{m,n}$ to show L^2 convergence. The L^2 convergence implies convergence in probability, and convergence in probability implies that there exists a subset of indices $i_n \in \mathbb{N}$ such that $Z_{m,n}(\beta_{i_n}) \xrightarrow{n\to\infty} Z_{m,n}(\beta)$ a.s., which is what we need to prove separability, see Definition 4.5. □

Now that we have shown that $(Z_{m,n}(\beta))_{\beta\in B_{m,\alpha}}$ is a separable subgaussian process on the metric space $(B_{m,\alpha}, D)$ we are in the position to apply Theorem (4.8) in order to bound the probability of $\sup_{\beta\in B_{m,\alpha}} Z_{m,n}(\beta)$ being larger then some arbitrary value, that will be carefully chosen later on.

Proposition 4.13. *Under the assumptions (H_α) and (H_K), for any $m \in \mathbb{N}$, $x > 0$, $\beta_0 \in B_{m,\alpha}$,*

$$\mathbb{P}\left(\sup_{\beta\in B_{m,\alpha}} Z_{m,n}(\beta) \geq 108\sqrt{\pi}K\alpha\sqrt{\frac{s_d(m)}{n}} + Z_{m,n}(\beta_0) + x\right) \leq 36\exp\left(-\frac{x^2 n}{144K^2\alpha^2}\right),$$

where the constant K is defined by (4.3).

Proof. As already outlined we know that $Z_{m,n}$ is subgaussian on $(B_{m,\alpha}, D)$ by Lemma (4.11) and separable (4.12). Applying Theorem (4.8) we get

$$\begin{aligned}\mathbb{P}\left(\sup_{\beta\in B_{m,\alpha}} Z_{m,n}(\beta) - Z_{m,n}(\beta_0) \geq 36\int_0^\infty \sqrt{\log(N(B_{m,\alpha}, D, \epsilon))}d\epsilon + x\right)\\ \leq 36\exp\left(-\frac{x^2}{36\,\mathrm{diam}(B_{m,\alpha})^2}\right) = 36\exp\left(-\frac{x^2 n}{36\times 4K^2\alpha^2}\right),\end{aligned}$$

where we used that

$$\mathrm{diam}(B_{m,\alpha}) = \frac{2K\alpha}{\sqrt{n}}$$

and $N(B_{m,\alpha}, D, \epsilon)$ is the covering number defined in (4.7). Since the distance D is just a scaled version of the Euclidean norm it holds that $N(B_{m,\alpha}, D, \epsilon) = N(B_{m,\alpha}, \|\cdot\|, \frac{\sqrt{n}}{K}\epsilon)$ and further we can use a result from R. v. Handel [9, Lemma 5.13] that lets us bound the covering number from above, more details can be found in the appendix (A.1):

$$\begin{aligned}&N(B_{m,\alpha}, D, \epsilon) \leq \left(\frac{3K\alpha}{\sqrt{n}\epsilon}\right)^{s_d(m)} && \text{if } \epsilon \leq \frac{K\alpha}{\sqrt{n}},\\ &N(B_{m,\alpha}, D, \epsilon) = 1 && \text{else.}\end{aligned}$$

This property of the covering number enables us to simplify the integral,

$$\begin{aligned}\int_0^\infty \sqrt{\log(N(B_{m,\alpha}, D, \epsilon))}d\epsilon &= \int_0^{\frac{K\alpha}{\sqrt{n}}} \sqrt{\log(N(B_{m,\alpha}, D, \epsilon))}d\epsilon\\ &\leq \int_0^{\frac{K\alpha}{\sqrt{n}}} \sqrt{s_d(m)\log\left(\frac{3K\alpha}{\sqrt{n}\epsilon}\right)}d\epsilon.\end{aligned}$$

We proceed with the substitution $x = \sqrt{\log(3K\alpha/\sqrt{n}\epsilon)}$. Switching from $d\epsilon$ to dx requires multiplying with a factor of $-2x\frac{3K\alpha}{\sqrt{n}}e^{-x^2}$ and the new integral bounds are ∞ and $\sqrt{\log(3)}$:

$$\begin{aligned}\int_0^{\frac{K\alpha}{\sqrt{n}}} \sqrt{s_d(m)\log\left(\frac{3K\alpha}{\sqrt{n}\epsilon}\right)}d\epsilon &= \sqrt{s_d(m)}\int_\infty^{\sqrt{\log(3)}} -2x^2\frac{3K\alpha}{\sqrt{n}}e^{-x^2}dx\\ &= 3K\alpha\sqrt{\frac{s_d(m)}{n}}\int_{\sqrt{\log(3)}}^\infty 2x^2e^{-x^2}dx\end{aligned}$$

$$\begin{aligned}
&\leq 3K\alpha\sqrt{\frac{s_d(m)}{n}}\int_0^\infty 2x^2e^{-x^2}dx \\
&= 3K\alpha\sqrt{\frac{s_d(m)}{n}}\int_{-\infty}^\infty x^2e^{-x^2}dx \\
&= 3K\alpha\sqrt{\frac{s_d(m)}{n}}\sqrt{\pi}.
\end{aligned}$$

This would finish the proof, but of course we have yet to show that the last integral indeed equals $\sqrt{\pi}$. We use partial integration and in the end identify the density function of a normal distribution whose integral over $\mathbb{R}$ is well known to equal 1:

$$\begin{aligned}
\int_{-\infty}^\infty x^2e^{-x^2}dx &= -\frac{1}{2}\int_{-\infty}^\infty x(-2xe^{-x^2})dx \\
&= \underbrace{-\frac{1}{2}\left[xe^{-x^2}\right]_{-\infty}^\infty}_{=0} + \frac{1}{2}\int_{-\infty}^\infty e^{-x^2}dx \\
&= \sqrt{\pi}\underbrace{\int_{-\infty}^\infty \frac{1}{2\sqrt{\pi}}e^{-x^2}dx}_{=1} = \sqrt{\pi}.
\end{aligned}$$

□

In the next proposition we establish an upper bound for $\mathbb{P}(\hat{m} = m)$ for the case $m \geq m^*$. Remember this will be helpful because

$$\mathbb{P}(m \neq m^*) = \sum_{m>m^*}\mathbb{P}(\hat{m} = m) + \sum_{m<m^*}\mathbb{P}(\hat{m} = m).$$

Proposition 4.14. *Let $0 < \rho < \frac{1}{2}$, and $\mathrm{pen}_n(m)$ be defined by (4.4):*

$$\mathrm{pen}_n(m) = K_{pen}n^{-\rho}\sqrt{s_d(m)}.$$

Let n_1 be the smallest integer satisfying

$$n_1 \geq \left(\frac{432\sqrt{\pi}K\alpha\sqrt{s_d(m^*+1)}}{K_{pen}\left(\sqrt{s_d(m^*+1} - \sqrt{s_d(m^*)}\right)}\right)^{1/(\frac{1}{2}-\rho)}. \tag{4.22}$$

Then, under the assumptions (H_α) and (H_K), for any $m > m^*$, $n \geq n_1$,

$$\mathbb{P}(\hat{m} = m) \leq 74\exp\left(-C_3(n^{1-2\rho} + s_d(m))\right),$$

where the constant C_3 is defined by

$$C_3 = \frac{1}{2} \times \frac{K_{\text{pen}}^2 \left(\sqrt{s_d(m^*+1)} - \sqrt{s_d(m^*)}\right)^2}{32 s_d(m^*+1)(72K^2\alpha^2 + K_Y^4)} \tag{4.23}$$

Proof. To proof this statement let us introduce an abbreviation for the difference of the penalization functions of m and m^*:

$$u_{m,n} = \frac{1}{2}(\text{pen}_n(m) - \text{pen}_n(m^*)) = \frac{K_{\text{pen}}}{2} n^{-\rho} \left(\sqrt{s_d(m)} - \sqrt{s_d(m^*)}\right).$$

Since the penalizing function is increasing in m, it is clear that in the case $m > m^*$ the value of $u_{m,n} > 0$. Additionally, the expression for $\mathbb{P}(\hat{m} = m)$ from Lemma 4.10 can be simplified to

$$\begin{aligned}\mathbb{P}(\hat{m} = m) &\leq \mathbb{P}\left(\sup_{\beta \in B_{m,\alpha}} |Z_{m,n}(\beta)| \geq u_{m,n}\right) \\ &= \mathbb{P}\left(\sup_{\beta \in B_{m,\alpha}} Z_{m,n}(\beta) \geq u_{m,n}\right) + \mathbb{P}\left(\sup_{\beta \in B_{m,\alpha}} (-Z_{m,n}(\beta)) \geq u_{m,n}\right). \end{aligned} \tag{4.24}$$

We start with the first term and use Proposition 4.13 to find a suitable bound. If $Z_{m,n}$ is subgaussian, of course also $-Z_{m,n}$ is subgaussian, therefore proposition 4.13 also holds for $-Z_{m,n}(\beta)$ and $-Z_{m,n}(\beta_0)$, implying that the second term can be handled similarly. Let β_0 be arbitrary but fixed in $B_{m,\alpha}$, where the exact value will be determined later,

$$\begin{aligned}\mathbb{P}\left(\sup_{\beta \in B_{m,\alpha}} (Z_{m,n}(\beta)) > u_{m,n}\right) &= \mathbb{P}\left(\sup_{\beta \in B_{m,\alpha}} (Z_{m,n}(\beta)) > u_{m,n}, Z_{m,n}(\beta_0) \leq \frac{u_{m,n}}{2}\right) \\ &\quad + \mathbb{P}\left(\sup_{\beta \in B_{m,\alpha}} (Z_{m,n}(\beta)) > u_{m,n}, Z_{m,n}(\beta_0) > \frac{u_{m,n}}{2}\right) \\ &\leq \mathbb{P}\left(\sup_{\beta \in B_{m,\alpha}} (Z_{m,n}(\beta)) > \frac{u_{m,n}}{2} + Z_{m,n}(\beta_0)\right) \\ &\quad + \mathbb{P}\left(\sup_{\beta \in B_{m,\alpha}} (Z_{m,n}(\beta)) > \frac{u_{m,n}}{2}\right). \end{aligned} \tag{4.25}$$

We want to restrict the first expression in (4.25) by using an upper bound for

$$\mathbb{P}\left(\sup_{\beta \in B_{m,\alpha}} Z_{m,n}(\beta) \geq 108\sqrt{\pi} K \alpha \sqrt{\frac{s_d(m)}{n}} + Z_{m,n}(\beta_0) + x\right)$$

from Proposition 4.13. By choosing $x = u_{m,n}/2 - 108K\alpha\sqrt{\pi s_d(m)/n}$ we can find such a bound, but we must ensure that $x > 0$:

$$\begin{aligned}
&\frac{u_{m,n}}{2} - 108K\alpha\sqrt{\frac{\pi s_d(m)}{n}} \\
&\quad = \frac{K_{\text{pen}}}{2} n^{-\rho}\left(\sqrt{s_d(m)} - \sqrt{s_d(m^*)}\right) - 108K\alpha\sqrt{\frac{\pi s_d(m)}{n}} \\
&\quad = \sqrt{s_d(m)} n^{-\rho}\frac{K_{\text{pen}}}{2}\left(1 - \sqrt{\frac{s_d(m^*)}{s_d(m)}} - \frac{2\times 108\sqrt{\pi}K\alpha}{K_{\text{pen}}} n^{\rho-\frac{1}{2}}\right) \\
&\quad \geq \sqrt{s_d(m)} n^{-\rho}\frac{K_{\text{pen}}}{2}\left(1 - \sqrt{\frac{s_d(m^*)}{s_d(m^*+1)}} - \frac{2\times 108\sqrt{\pi}K\alpha}{K_{\text{pen}}} n^{\rho-\frac{1}{2}}\right),
\end{aligned}$$

where the last inequality stems from $s_d(m^*+1) \leq s_d(m)$ for every $m > m^*$.

Choosing $n_1 \in \mathbb{N}$ such that

$$\begin{aligned}
& 1 - \sqrt{\frac{s_d(m^*)}{s_d(m^*+1)}} - \frac{216\sqrt{\pi}K\alpha}{K_{\text{pen}}} n_1^{\rho-\frac{1}{2}} > \frac{1}{2}\left(1 - \sqrt{\frac{s_d(m^*)}{s_d(m^*+1)}}\right) \\
\iff \quad & \frac{1}{2}\left(1 - \sqrt{\frac{s_d(m^*)}{s_d(m^*+1)}}\right) > \frac{216\sqrt{\pi}K\alpha}{K_{\text{pen}}} n_1^{\rho-\frac{1}{2}} \\
\iff \quad & \frac{K_{\text{pen}}}{216\sqrt{\pi}K\alpha} n_1^{\frac{1}{2}-\rho} > 2\frac{\sqrt{s_d(m^*+1)}}{\sqrt{s_d(m^*+1)} - \sqrt{s_d(m^*)}} \\
\iff \quad & n_1 > \left(\frac{432\sqrt{\pi}K\alpha\sqrt{s_d(m^*+1)}}{K_{\text{pen}}\left(\sqrt{s_d(m^*+1)} - \sqrt{s_d(m^*)}\right)}\right)^{1/(\frac{1}{2}-\rho)}
\end{aligned}$$

ensures that for any $n > n_1$ and any $m > m^*$,

$$x = \frac{u_{m,n}}{2} - 108K\alpha\sqrt{\frac{\pi s_d(m)}{n}} \geq \sqrt{s_d(m)} n^{-\rho}\frac{K_{\text{pen}}}{4}\left(1 - \sqrt{\frac{s_d(m^*)}{s_d(m^*+1)}}\right) > 0.$$

Applying Proposition 4.13 to $x = u_{m,n}/2 - 108K\alpha\sqrt{\pi s_d(m)/n}$ yields, for any $n \geq n_1$,

$$\mathbb{P}\left(\sup_{\beta\in B_{m,\alpha}} (Z_{m,n}(\beta)) > \frac{u_{m,n}}{2} + Z_{m,n}(\beta_0)\right)$$

$$\leq 36 \exp\left(-\frac{n}{144K^2\alpha^2}\left(\frac{u_{m,n}}{2} - 108K\alpha\sqrt{\frac{\pi s_d(m)}{n}}\right)^2\right)$$
$$\leq 36 \exp\left(-\frac{s_d(m)n^{1-2\rho}K_{\text{pen}}^2}{144K^2\alpha^2 \times 16}\left(1 - \sqrt{\frac{s_d(m^*)}{s_d(m^*+1)}}\right)^2\right)$$
$$= 36 \exp\left(-\kappa_1 s_d(m) n^{1-2\rho}\right), \tag{4.26}$$

where

$$\kappa_1 = \frac{K_{\text{pen}}^2}{2304K^2\alpha^2}\left(1 - \sqrt{\frac{s_d(m^*)}{s_d(m^*+1)}}\right)^2.$$

In order to bound the second term in (4.25) Hoeffding's inequality will be useful:

Lemma 4.15 (Hoeffding's inequality)**.** *Let $X_1, \ldots, X_N$ be independent random variables. Assume that $X_i - \mathbb{E}[X_i] \in [a_i, b_i]$ for every i. Then, for the empirical expectation $\hat{S}_N = \frac{1}{N}\sum_{i=1}^N X_i$ and any $\epsilon > 0$, we have*

$$\mathbb{P}\left(\hat{S}_N - \mathbb{E}[\hat{S}_N] \geq \epsilon\right) \leq \exp\left(-\frac{2\epsilon^2 N^2}{\sum_{i=1}^N (b_i - a_i)^2}\right).$$

A proof can be found in the appendix, see Lemma A.2.

We want to apply Hoeffding's inequality to

$$\mathbb{P}\left(Z_{m,n}(\beta_0) > \frac{u_{m,n}}{2}\right).$$

This is possible because

$$Z_{m,n}(\beta_0) = \frac{1}{n}\sum_{i=1}^n (Y_i - \langle \beta_0, S^m(X_i)\rangle)^2 - \mathbb{E}[(Y_i - \langle \beta_0, S^m(X_i)\rangle)^2],$$

and the fact that

$$|Y_i - \langle \beta_0, S^m(X_i)\rangle|^2 \leq (K_Y + \|\beta_o\| e^{K_X})^2$$

implies

$$-(K_Y + \|\beta_o\| e^{K_X})^2 \leq (Y_i - \langle \beta_0, S^m(X_i)\rangle)^2 - \mathbb{E}[(Y_i - \langle \beta_0, S^m(X_i)\rangle)^2] \leq (K_Y + \|\beta_o\| e^{K_X})^2.$$

Applying the inequality reveals that for any $n \geq n_1$,

$$\begin{aligned}
\mathbb{P}\left(Z_{m,n}(\beta_0) > \frac{u_{m,n}}{2}\right) &= \mathbb{P}\left(\sum_{i=1}^{n}(Y_i - \langle \beta_0, S^m(X_i)\rangle)^2 - \mathbb{E}[(Y_i - \langle \beta_0, S^m(X_i)\rangle)^2] > \frac{nu_{m,n}}{2}\right) \\
&\leq \exp\left(-\frac{2n^2\frac{u_{m,n}^2}{4}}{\sum_{i=1}^{n}(2(K_Y + \|\beta_0\|e^{K_X})^2)^2}\right) \\
&= \exp\left(-\frac{nu_{m,n}^2}{8(K_Y + \|\beta_0\|e^{K_X})^4}\right) \\
&\leq \exp\left(-\frac{n^{1-2\rho}K_{\text{pen}}^2\left(\sqrt{s_d(m)} - \sqrt{s_d(m^*)}\right)^2}{32(K_Y + \|\beta_0\|e^{K_X})^4}\right) \\
&= \exp\left(-\frac{n^{1-2\rho}K_{\text{pen}}^2}{32(K_Y + \|\beta_0\|e^{K_X})^4}\left(\sqrt{s_d(m)}\left(1 - \frac{\sqrt{s_d(m^*)}}{\sqrt{s_d(m)}}\right)\right)^2\right) \\
&\leq \exp\left(-\frac{n^{1-2\rho}K_{\text{pen}}^2 s_d(m)}{32(K_Y + \|\beta_0\|e^{K_X})^4}\left(1 - \sqrt{\frac{s_d(m^*)}{s_d(m^*+1)}}\right)^2\right) \\
&= \exp(-\kappa_2 n^{1-2\rho}s_d(m)), \qquad (4.27)
\end{aligned}$$

where

$$\kappa_2 = \frac{K_{\text{pen}}^2}{32(K_Y + \|\beta_0\|e^{K_X})^4}\left(1 - \sqrt{\frac{s_d(m^*)}{s_d(m^*+1)}}\right)^2.$$

With bounds for both terms in (4.25) at hand, we can combine (4.26) and (4.27) and obtain

$$\begin{aligned}
\mathbb{P}\left(\sup_{\beta \in B_{m,\alpha}} Z_{m,n} > u_{m,n}\right) &\leq 36\exp\left(-\kappa_1 s_d(m)n^{1-2\rho}\right) + \exp(-\kappa_2 n^{1-2\rho}s_d(m)) \\
&\leq 37\exp\left(-\kappa_3 n^{1-2\rho}s_d(m)\right) \\
&\leq 37\exp\left(-\frac{\kappa_3}{2}(n^{1-2\rho} + s_d(m)\right).
\end{aligned}$$

Since we also need to handle the second term in (4.24), the next challenge is to ensure that the same procedure also works for $-Z_{m,n}$. The negative counterpart to (4.25) reads

$$\mathbb{P}\left(\sup_{\beta \in B_{m,\alpha}}(-Z_{m,n}(\beta)) > u_{m,n}\right) = \mathbb{P}\left(\sup_{\beta \in B_{m,\alpha}}(-Z_{m,n}(\beta)) > u_{m,n}, -Z_{m,n}(\beta_0) \leq \frac{u_{m,n}}{2}\right)$$

$$+ \mathbb{P}\left(\sup_{\beta\in B_{m,\alpha}}(-Z_{m,n}(\beta)) > u_{m,n}, -Z_{m,n}(\beta_0) > \frac{u_{m,n}}{2}\right)$$
$$\leq \mathbb{P}\left(\sup_{\beta\in B_{m,\alpha}}(-Z_{m,n}(\beta)) > \frac{u_{m,n}}{2} - Z_{m,n}(\beta_0)\right)$$
$$+ \mathbb{P}\left(\sup_{\beta\in B_{m,\alpha}}(-Z_{m,n}(\beta)) > \frac{u_{m,n}}{2}\right). \tag{4.28}$$

Now, we benefit form the fact that we used Proposition 4.13 to handle the first term and Hoeffding's inequality 4.15 to handle the second term.

Of course if $Z_{m,n}$ is a subgaussian process, analogously $-Z_{m,n}$ will be a subgaussian process as well. This is the only property we need, in order to apply Proposition 4.13. This means we can follow exactly the same proof as for the first term of (4.28).

To control the second term in (4.28) we used Hoeffding's inequality, but again Hoeffding's inequality 4.15 only needs a sum of bounded random variables. We established that

$$-(K_Y + \|\beta_o\|e^{K_X})^2 \leq (Y_i - \langle\beta_0, S^m(X_i)\rangle)^2 - \mathbb{E}[(Y_i - \langle\beta_0, S^m(X_i)\rangle)^2] \leq (K_Y + \|\beta_o\|e^{K_X})^2.$$

Similarly we can multiply this chain of inequalities with -1 and obtain

$$(K_Y + \|\beta_o\|e^{K_X})^2 \geq -(Y_i - \langle\beta_0, S^m(X_i)\rangle)^2 + \mathbb{E}[(Y_i - \langle\beta_0, S^m(X_i)\rangle)^2] \geq -(K_Y + \|\beta_o\|e^{K_X})^2,$$

which means we can apply Hoeffding's inequality in the same manner as before.

Altogether the proofs for $Z_{m,n}$ and $-Z_{m,n}$ work analogously and consequently

$$\mathbb{P}(\hat{m} = m) \leq 2 \times 37 \exp\left(-\frac{\kappa_3}{2}(n^{1-2\rho} + s_d(m)\right).$$

The only thing left to do is choose β_0 optimal. κ_3 drives $P(\hat{m} = m)$ small exponentially, the bigger it gets. Therefore, the goal is to make κ_3 as large as possible. Since

$$\kappa_3 = \min(\kappa_1, \kappa_2) = \frac{K_{\text{pen}}^2}{32}\left(1 - \sqrt{\frac{s_d(m^*)}{s_d(m^*+1)}}\right)^2 \min\left(\frac{1}{72K^2\alpha^2}, \frac{1}{(K_Y + \|\beta_0\|e^{K_X})^4}\right),$$

we choose $\beta_0 = 0$ to minimise $\|\beta_0\|$ and maximise κ_3, which yields

$$\begin{aligned}\kappa_3 &= \min(\kappa_1, \kappa_2) \\ &= \frac{K_{\text{pen}}^2}{32}\left(1 - \sqrt{\frac{s_d(m^*)}{s_d(m^*+1)}}\right)^2 \min\left(\frac{1}{72K^2\alpha^2}, \frac{1}{K_Y^4}\right) \\ &= \frac{K_{\text{pen}}^2}{32}\left(\frac{\sqrt{s_d(m^*+1)} - \sqrt{s_d(m^*)}}{\sqrt{s_d(m^*+1)}}\right)^2 \min\left(\frac{1}{72K^2\alpha^2}, \frac{1}{K_Y^4}\right).\end{aligned}$$

Choosing

$$C_3 = \frac{1}{2} \times \frac{K_{\text{pen}}^2\left(\sqrt{s_d(m^*+1)} - \sqrt{s_d(m^*)}\right)^2}{32 s_d(m^*+1)(72K^2\alpha^2 + K_Y^4)} \leq \frac{\kappa_3}{2},$$

we obtain

$$\mathbb{P}(\hat{m} = m) \leq 2 \times 37 \exp\left(-\frac{\kappa_3}{2}(n^{1-2\rho} + s_d(m)\right) \leq 74 \exp\left(-C_3(n^{1-2\rho} + s_d(m))\right)$$

and the proof is complete. □

Remark 9. In line (4.27) there is a difference to the original work. Fermanian uses Hoeffding's inequality and obtains $(K_Y + \|\beta_0\|e^{K_X})^2$. Common versions of Hoeffding's inequality, however, yield $(K_Y + \|\beta_0\|e^{K_X})^4$. The same procedure can be observed in (4.32).

This difference will change the final values of C_2 and C_3, compared to the constants in [6], see Remark 7.

Remark 10. In her original work Fermanian simplifies

$$\sqrt{s_d(m^*+1)} - \sqrt{s_d(m^*)} = \sqrt{d^{m^*+1} + s_d(m^*)} - \sqrt{s_d(m^*)} \geq \sqrt{\frac{d^{m*+1}}{2}},$$

and justifies this with the fact that for $a, b \geq 0$, $\sqrt{a} + \sqrt{b} \geq \sqrt{2}\sqrt{a+b}$. Unfortunately the latter inequality is only correct the other way round, i.e.

$$\begin{aligned}\sqrt{a} + \sqrt{b} \leq \sqrt{2}\sqrt{a+b} &\iff a + 2\sqrt{ab} + b \leq 2(a+b) \\ &\iff 0 \leq a - 2\sqrt{ab} + b = (\sqrt{a} - \sqrt{b})^2.\end{aligned}$$

However, even if we consider the correct inequality, applying it only yields

$$\sqrt{d^{m^*+1} + s_d(m^*)} - \sqrt{\frac{s_d(m^*)}{2}} \geq \sqrt{\frac{d^{m*+1}}{2}}$$

and not the above statement. Indeed choosing $s_d(m^*) = 5$ and plotting

$$\sqrt{d^{m^*+1} + s_d(m^*)} - \sqrt{s_d(m^*)} - \sqrt{\frac{d^{m*+1}}{2}}$$

shows that this quantity is not strictly positive. Therefore, the inequality does not hold in general without further restrictions. We stick to the not simplified expression and therefore n_0, C_2 and C_3 deviate from those in the original work, see Remark 7.

Establishing a bound for $\mathbb{P}(\hat{m} = m)$ in the case $m > m^*$ finishes the first of three parts on the way to proving Theorem 4.1. The second part is devoted to finding a similar bound in the case that $m < m^*$. The tools we use for this latter case are mostly the same as in the first one, but Lemma 4.10 stating that

$$\mathbb{P}(\hat{m} = m) \leq \mathbb{P}\left(2 \sup_{\beta \in B_{m,\alpha}} |\hat{R}_{m,n}(\beta) - R_m(\beta)| \geq pen_n(m) - pen_n(m^*)\right)$$

is only valid for $m > m^*$, which is not the case we want to elaborate on.

Therefore, in the following the strategy will be slightly different and we introduce a new bound for the expression

$$\mathbb{P}\left(|\hat{L}_n(m) - L(m)| > \epsilon\right)$$

in the next Proposition. This will help us to handle $\mathbb{P}(\hat{m} = m)$ in the $m < m^*$ case.

Proposition 4.16. *For any $\epsilon > 0$, $m \in \mathbb{N}$, let $n_2 \in \mathbb{N}$ be the smallest integer such that*

$$n_2 \geq \frac{432^2 K^2 \pi \alpha^2 s_d(m)}{\epsilon^2}. \tag{4.29}$$

Then for any $n \geq n_2$,

$$\mathbb{P}\left(|\hat{L}_n(m) - L(m)| > \epsilon\right) \leq 74(-C_4 n\epsilon^2),$$

where C_4 is defined by

$$C_4 = \frac{1}{2(1152K^2\alpha^2 + K_Y^4)}. \tag{4.30}$$

Proof. Recall Lemma 4.9 stating that

$$|\hat{L}_n(m) - L(m)| \leq \sup_{\beta \in B_{m,\alpha}} |\hat{R}_{m,n}(\beta) - R_m(\beta)| = \sup_{\beta \in B_{m,\alpha}} |Z_{m,n}(\beta)|.$$

Hence

$$\begin{aligned}\mathbb{P}\left(|\hat{L}_n(m) - L(m)| > \epsilon\right) &\leq \mathbb{P}\left(\sup_{\beta \in B_{m,\alpha}} |Z_{m,n}(\beta)| > \epsilon\right) \\ &= \mathbb{P}\left(\sup_{\beta \in B_{m,\alpha}} Z_{m,n}(\beta) > \epsilon\right) + \mathbb{P}\left(\sup_{\beta \in B_{m,\alpha}} (-Z_{m,n}(\beta)) > \epsilon\right).\end{aligned}$$

Now we fix some $\beta_0 \in B_{m,\alpha}$ and proceed as in Proposition 4.14, i.e. we first make sure to choose n such that

$$\frac{\epsilon}{2} - 108K\alpha\sqrt{\frac{\pi s_d(m)}{n}} > 0,$$

and then we use Hoeffding's inequality 4.15 and Proposition 4.13 to bound $\mathbb{P}\left(\sup_{\beta \in B_{m,\alpha}} Z_{m,n}(\beta) > \epsilon\right)$ and $\mathbb{P}\left(\sup_{\beta \in B_{m,\alpha}} (-Z_{m,n}(\beta)) > \epsilon\right)$.

Indeed n_2 is chosen exactly so, that $n \geq n_2$ fulfills

$$\begin{aligned} & n \geq n_2 \geq \frac{432^2 K^2 \pi \alpha^2 s_d(m)}{\epsilon^2} \\ \iff \quad & \sqrt{n_2} \geq \frac{432K\alpha\sqrt{\pi s_d(m)}}{\epsilon} \\ \iff \quad & \frac{\epsilon}{432K\alpha} \geq \sqrt{\frac{\pi s_d(m)}{n_2}} \\ \iff \quad & \frac{\epsilon}{4} - 108K\alpha\sqrt{\frac{\pi s_d(m)}{n_2}} + \frac{\epsilon}{4} \geq \frac{\epsilon}{4} \\ \iff \quad & \frac{\epsilon}{2} - 108K\alpha\sqrt{\frac{\pi s_d(m)}{n_2}} \geq \frac{\epsilon}{4} > 0. \end{aligned}$$

The following arguments should look familiar, because they are indeed exactly the same that we already used in (4.25):

$$\mathbb{P}\left(\sup_{\beta\in B_{m,\alpha}}(Z_{m,n}(\beta))>\epsilon\right)\leq\mathbb{P}\left(\sup_{\beta\in B_{m,\alpha}}(Z_{m,n}(\beta))>\frac{\epsilon}{2}+Z_{m,n}(\beta_0)\right)+\mathbb{P}\left(\sup_{\beta\in B_{m,\alpha}}(Z_{m,n}(\beta))>\frac{\epsilon}{2}\right).$$

Then we use Proposition 4.13 and Hoeffding's inequality 4.15 to obtain

$$\begin{aligned}\mathbb{P}\left(\sup_{\beta\in B_{m,\alpha}}(Z_{m,n}(\beta))>\epsilon\right)&\leq\mathbb{P}\left(\sup_{\beta\in B_{m,\alpha}}(Z_{m,n}(\beta))>\frac{\epsilon}{2}+Z_{m,n}(\beta_0)\right)+\\&\quad\mathbb{P}\left(\sup_{\beta\in B_{m,\alpha}}(Z_{m,n}(\beta))>\frac{\epsilon}{2}\right)\\&\leq 36\exp\left(-\frac{n}{144K^2\alpha^2}\left(\underbrace{\frac{\epsilon}{2}-108K\alpha\sqrt{\frac{\pi s_d(m)}{n}}}_{\geq\epsilon/4}\right)^2\right)\\&\quad+\exp\left(-\frac{n\epsilon^2}{2(K_Y+\|\beta_0\|e^{K_X})^4}\right)\\&\leq 36\exp\left(-\frac{n\epsilon^2}{2304K^2\alpha^2}\right)+\exp\left(-\frac{n\epsilon^2}{2(K_Y+\|\beta_0\|e^{K_X})^4}\right)\\&\leq 37\exp(-\kappa_4 n\epsilon^2),\end{aligned}\tag{4.31}$$

(4.32)

where

$$\kappa_4=\min\left(\frac{1}{2304K^2\alpha^2},\frac{1}{2(K_Y+\|\beta_0\|e^{K_X})^4}\right).$$

We justified in the previous Proposition that the same procedure is valid for $-Z_{m,n}$, hence

$$\mathbb{P}\left(|\hat{L}_n(m)-L(m)|>\epsilon\right)\leq 74\exp(-\kappa_4 n\epsilon^2).$$

Choosing $\beta_0=0$ maximizes κ_4 and with

$$C_4=\frac{1}{2(1152K^2\alpha^2+K_Y^4)}\leq\min\left(\frac{1}{2304K^2\alpha^2},\frac{1}{2(K_Y+\|\beta_0\|e^{K_X})^4}\right)=\kappa_4,\tag{4.33}$$

the proof is complete. □

With this bound for $\mathbb{P}\left(|\hat{L}_n(m) - L(m)| > \epsilon\right)$ we are well prepared for finding an upper bound for $\mathbb{P}(\hat{m} = m)$ in the case $m < m^*$.

Proposition 4.17. *Let* $0 < \rho < \frac{1}{2}$ *and* $\mathrm{pen}_n(m)$ *be defined by (4.4). Let* n_3 *be the smallest integer satisfying*

$$n_3 \geq \left(\frac{2\sqrt{s_d(m^*)}}{L(m^* - 1) - L(m^*)}(432K\alpha\sqrt{\pi} + K_{\mathrm{pen}})\right)^{1/\rho}. \tag{4.34}$$

Then, under the assumptions (H_α) *and* (H_K)*, for any* $m < m^*$,$n \geq n_3$,

$$\mathbb{P}(\hat{m} = m) \leq 148(-n\frac{C_4}{4}(L(m) - L(m^*) - \mathrm{pen}_n(m^*) + \mathrm{pen}_n(m))^2),$$

where C_4 *is defined by (4.30).*

Proof. Recall that $\hat{m}$ is chosen by,

$$\hat{m} = \min\left(\underset{m\in\mathbb{N}}{\mathrm{argmin}}(\hat{L}_n(m) + pen_n(m))\right).$$

Hence it is clear that $\hat{L}_n(m) + pen_n(m) \leq \hat{L}_n(m^*) + pen_n(m^*)$ is a necessary condition for $\hat{m} = m$ and therefore

$$\begin{aligned}
\mathbb{P}(\hat{m} = m) &\leq \mathbb{P}\left(\hat{L}_n(m) + \mathrm{pen}_n(m) \leq \hat{L}_n(m^*) + \mathrm{pen}_n(m^*)\right) \\
&= \mathbb{P}\left(\hat{L}_n(m^*) - \hat{L}_n(m) \geq \mathrm{pen}_n(m) - \mathrm{pen}_n(m^*)\right) \\
&= \mathbb{P}\left(\hat{L}_n(m^*) - L(m^*) + L(m) - \hat{L}_n(m) \leq L(m) - L(m^*) + \mathrm{pen}_n(m) - \mathrm{pen}_n(m^*)\right) \\
&\leq \mathbb{P}\left(|\hat{L}_n(m) - L(m)| \leq \frac{1}{2}(L(m) - L(m^*) + \mathrm{pen}_n(m) - \mathrm{pen}_n(m^*))\right) \\
&\quad + \mathbb{P}\left(|\hat{L}_n(m^*) - L(m^*)| \leq \frac{1}{2}(L(m) - L(m^*) + \mathrm{pen}_n(m) - \mathrm{pen}_n(m^*))\right).
\end{aligned}$$

This already looks like the required setting to apply Proposition 4.16. However, we have to ensure that $L(m) - L(m^*) + \mathrm{pen}_n(m) - \mathrm{pen}_n(m^*)$ is positive. Since $m \mapsto L(m)$ is decreasing and minimal for m^*, we know that for $m < m^*$

$$L(m) \geq L(m^* - 1) > L(m^*).$$

Since $m \mapsto \mathrm{pen}_n(m)$ is strictly increasing,

$$\mathrm{pen}_n(m) - \mathrm{pen}_n(m^*) > -\mathrm{pen}_n(m^*) = -K_{\mathrm{pen}}n^{-\rho}\sqrt{s_d(m^*)},$$

and together

$$L(m) - L(m^*) + \text{pen}_n(m) - \text{pen}_n(m^*) > L(m^* - 1) - L(m^*) - K_{\text{pen}} n^{-\rho} \sqrt{s_d(m^*)}. \tag{4.35}$$

We use that

$$L(m^* - 1) - L(m^*) - K_{\text{pen}} n^{-\rho} \sqrt{s_d(m^*)} > \frac{1}{2}(L(m^* - 1) - L(m^*)), \tag{4.36}$$

is sufficient for $L(m^* - 1) - L(m^*) - K_{\text{pen}} n^{-} \rho \sqrt{s_d(m^*)} > 0$. Rearranging reveals that the above inequality is fulfilled for all $n \geq n_3$, if

$$n_3 \geq \left(\frac{2K_{\text{pen}} \sqrt{s_d(m^*)}}{L(m^{*-}) - L(m^*)} \right)^{\frac{1}{\rho}}. \tag{4.37}$$

Additionally, in order to apply Proposition 4.16 it is necessary that n also fulfills

$$n \geq \frac{432^2 K^2 \pi \alpha^2 s_d(m)}{\epsilon^2},$$

where in our case

$$\epsilon = \frac{1}{2}(L(m) - L(m^*) + \text{pen}_n(m) - \text{pen}_n(m^*)).$$

This imposes a second condition on n_3, i.e.

$$n_3 \geq \frac{4 \times 432^2 K^2 \pi \alpha^2 s_d(m)}{(L(m) - L(m^*) + \text{pen}_n(m) - \text{pen}_n(m^*))^2}, \tag{4.38}$$

which can be bounded in m using equation (4.35) and (4.36),

$$\begin{aligned} \frac{4 \times 432^2 K^2 \pi \alpha^2 s_d(m)}{(L(m) - L(m^*) + \text{pen}_n(m) - \text{pen}_n(m^*))^2} &\leq \frac{4 \times 4 \times 432^2 K^2 \pi \alpha^2 s_d(m)}{(L(m^* - 1) - L(m^*))^2} \\ &= \left(\frac{4 \times 432 K \alpha \sqrt{\pi s_d(m^*)}}{L(m^* - 1) - L(m^*)} \right)^2. \end{aligned}$$

This means that

$$n_3 \geq \max \left(\left(\frac{2K_{\text{pen}} \sqrt{s_d(m^*)}}{L(m^* - 1) - L(m^*)} \right)^{\frac{1}{\rho}}, \left(\frac{4 \times 432 K \alpha \sqrt{\pi s_d(m^*)}}{L(m^* - 1) - L(m^*)} \right)^2 \right),$$

would fulfill all necessary conditions. But we can go even one step further and change the exponent of the second component from 2 to $1/\rho$. This is reasonable, because if the quantity is smaller than 1, the condition will automatically be satisfied, since $n \in \mathbb{N}$. Thus, we can exponentiate outside the maximum and summing both terms instead of taking the maximum, we obtain a more compact form reading

$$n_3 \geq \left(\frac{2(K_{\text{pen}} + 864K\alpha\sqrt{\pi})\sqrt{s_d(m^*)}}{L(m^* - 1) - L(m^*)} \right)^{1/\rho}.$$

Applying Proposition 4.16 we conclude the proof:

$$\begin{aligned}
\mathbb{P}(\hat{m} = m) &\leq \mathbb{P}\left(|\hat{L}_n(m) - L(m)| \leq \frac{1}{2}(L(m) - L(m^*) + \text{pen}_n(m) - \text{pen}_n(m^*)) \right) \\
&\quad + \mathbb{P}\left(|\hat{L}_n(m^*) - L(m^*)| \leq \frac{1}{2}(L(m) - L(m^*) + \text{pen}_n(m) - \text{pen}_n(m^*)) \right) \\
&\leq 2 \times 74(-n\frac{C_4}{4}(L(m) - L(m^*) - \text{pen}_n(m^*) + \text{pen}_n(m))^2).
\end{aligned}$$

□

Remark 11. In the original work the expressions in (4.35) and (4.36) are further bounded by plugging in $\sigma^2 \geq L(m^*)$. Even though this is true, in contrast to $L(m^* - 1) - L(m^*)$ we cannot ensure $L(m^* - 1) - \sigma^2$ to be positive. This leads to the unsatisfying situation that for certain values equation (4.36) can never be fulfilled, which is a crucial part because it ensures $L(m) - L(m^*) + \text{pen}_n(m) - \text{pen}_n(m^*)$ to be positive. In this thesis, we stick to $L(m^*)$ and do not use σ^2. This deviation from the original work will change the constants n_0 and C_2 in Theorem 4.1, see Remark 7.

Remark 12. When we apply Proposition 4.16 in the above proof we choose

$$\epsilon = \frac{1}{2}(L(m) - L(m^*) + \text{pen}_n(m) - \text{pen}_n(m^*)).$$

This, of course, has to correspond to our choice of n_3. In [6] A. Fermanian made a small mistake: In (4.38) she drops the 1/2 and only considers

$$\epsilon = L(m) - L(m^*) + \text{pen}_n(m) - \text{pen}_n(m^*)$$

and consequently

$$n_3 \geq \frac{432^2 K^2 \pi \alpha^2 s_d(m)}{(L(m) - L(m^*) + \text{pen}_n(m) - \text{pen}_n(m^*))^2}.$$

instead of

$$n_3 \geq \frac{4 \times 432^2 K^2 \pi \alpha^2 s_d(m)}{(L(m) - L(m^*) + \text{pen}_n(m) - \text{pen}_n(m^*))^2},$$

which would be the correct term. This mistake is carried through the rest of the proof and eventually changes a value in n_0, see Remark 7.

Now that we have upper bounds for the expression $\mathbb{P}(\hat{m} = m)$ for both, $m < m^*$ and $m > m^*$ it is time to combine both parts and proof the theorem.

Proof of Theorem 4.1. The proof is now straightforward. We just need to assemble the bounds from Proposition 4.14 and Proposition 4.17. Both propositions impose conditions on n, so we need to ensure that n suffices conditions (4.22) and (4.34). To this end, we introduce

$$M = \max \left(\left(\frac{432\sqrt{\pi} K \alpha \sqrt{s_d(m^*+1)}}{K_{pen}\left(\sqrt{s_d(m^*+1} - \sqrt{s_d(m^*)}\right)} \right)^{1/(\frac{1}{2}-\rho)}, \right.$$
$$\left. \left(\frac{2(K_{\text{pen}} + 864 K \alpha \sqrt{\pi}) \sqrt{s_d(m^*)}}{L(m^*-1) - L(m^*)} \right)^{1/\rho} \right).$$

Letting $\tilde{\rho} = \min(\rho, \frac{1}{2} - \rho)$, M can be bounded by

$$M \leq \max \left(\left(\frac{(K_{\text{pen}} + 864\sqrt{\pi} K \alpha) \sqrt{s_d(m^*+1)}}{K_{pen}\left(\sqrt{s_d(m^*+1} - \sqrt{s_d(m^*)}\right)} \right)^{1/\tilde{\rho}}, \right.$$
$$\left. \left(\frac{2(K_{\text{pen}} + 864 K \alpha \sqrt{\pi}) \sqrt{s_d(m^*+1)}}{L(m^*-1) - L(m^*)} \right)^{1/\tilde{\rho}} \right).$$
$$= \left((864 K \alpha \sqrt{\pi} + K_{\text{pen}}) \right.$$
$$\left. \times \left(\frac{2\sqrt{s_d(m^*+1)}}{L(m^*-1) - L(m^*)} + \frac{\sqrt{s_d(m^*+1)}}{K_{pen}\left(\sqrt{s_d(m^*+1} - \sqrt{s_d(m^*)}\right)} \right) \right)^{1/\tilde{\rho}} \leq n_0.$$

This corresponds exactly to condition (4.5) from Theorem 4.1 and every $n \geq n_0$ will satisfy the requirements (4.22) and (4.34).

With both conditions fulfilled the only thing left is to combine the results from both propositions in a reasonable manner. We have

$$\mathbb{P}(\hat{m} \neq m^*) = \mathbb{P}(\hat{m} > m^*) + \mathbb{P}(\hat{m} < m^*) = \sum_{m>m^*} \mathbb{P}(\hat{m} = m) + \sum_{m<m^*} \mathbb{P}(\hat{m} = m).$$

Proposition 4.14 guarantees for $n \geq n_0$,

$$\sum_{m>m^*} \mathbb{P}(\hat{m} = m) \leq 74 e^{-C_3 n^{1-2\rho}} \sum_{m>m^*} e^{-C_3 s_d(m)},$$

and on the other hand Proposition 4.17 gives

$$\sum_{m<m^*} \mathbb{P}(\hat{m} = m) \leq 148 \sum_{m=0}^{m^*-1} \exp\left(-\frac{C_4}{4} n (L(m) - L(m^*) - \mathrm{pen}_n(m^*) + \mathrm{pen}_n(m))^2\right) \tag{4.39}$$

$$\leq 148 m^* \exp\left(-\frac{C_4}{16} n \left(L(m^* - 1) - L(m^*)\right)^2\right), \tag{4.40}$$

where we used (4.35) and (4.36), i.e. for $n \geq n_0 \geq n_3$,

$$L(m) - L(m^*) - \mathrm{pen}_n(m^*) + \mathrm{pen}_n(m) \geq \frac{1}{2} \left(L(m^* - 1) - L(m^*)\right.\,.$$

Defining

$$\kappa_5 = \min\left(C_3, \frac{C_4 \left(L(m^* - 1) - L(m^*)\right)^2}{16}\right)$$

yields

$$\mathbb{P}(\hat{m} \neq m^*) \leq 74 e^{-\kappa_5 n^{1-2\rho}} + 148 m^* e^{-\kappa_5 n} \leq C_1 e^{-\kappa_5 n^{1-2\rho}},$$

where

$$C_1 = 74 \sum_{m>0} e^{-C_3 s_d(m)} + 148 m^*.$$

In a last step we simplify the expression by finding a lower bound for κ_5:

$$\kappa_5 = \min\left(C_3, \frac{C_4 \left(L(m^* - 1) - L(m^*)\right)^2}{16}\right)$$

$$
\begin{aligned}
&= \min\left(\frac{K_{\text{pen}}^2\left(\sqrt{s_d(m^*+1)}-\sqrt{s_d(m^*)}\right)^2}{64 s_d(m^*+1)(72K^2\alpha^2+K_Y^4)}, \frac{(L(m^*-1)-L(m^*))^2}{32(1152K^2\alpha^2+K_Y^4)}\right)\\
&\geq \frac{1}{32(1152K^2\alpha^2+K_Y^4)} \min\left(\frac{K_{\text{pen}}^2\left(\sqrt{s_d(m^*+1)}-\sqrt{s_d(m^*)}\right)^2}{2 s_d(m^*+1)}, (L(m^*-1)-L(m^*))^2\right)\\
&= C_2. \qquad (4.41)
\end{aligned}
$$

We showed that for

$$\text{pen}_n(m) = K_{\text{pen}} n^{-p}\sqrt{s_d(m)}$$

and $n \geq n_0$, n_0 being the smallest integer, such that

$$
\begin{aligned}
n_0^{\tilde{\rho}} \geq \Bigg(&(864K\alpha\sqrt{\pi} + K_{\text{pen}})\\
&\times \left(\frac{2\sqrt{s_d(m^*+1)}}{L(m^*-1)-L(m^*)} + \frac{\sqrt{s_d(m^*+1)}}{K_{pen}\left(\sqrt{s_d(m^*+1}-\sqrt{s_d(m^*)}\right)}\right)\Bigg)
\end{aligned}
$$

and $\tilde{\rho} = \min(\rho, 1-2\rho)$, we have that: Under the conditions (H_K) and (H_α) the estimated truncation order $\hat{m}$ chosen by

$$\hat{m} = \min\left(\underset{m\in\mathbb{N}}{\text{argmin}}(\hat{L}_n(m) + pen_n(m))\right)$$

fulfills

$$\mathbb{P}(\hat{m} \neq m^*) \leq C_1 e^{-C_2 n^{1-2\rho}},$$

where

$$C_1 = 74\sum_{m>0} e^{-C_3 s_d(m)} + 148m^*,$$

$$C_2 = \frac{\min\left(\frac{K_{\text{pen}}^2\left(\sqrt{s_d(m^*+1)}-\sqrt{s_d(m^*)}\right)^2}{2s_d(m^*+1)}, (L(m^*-1)-L(m^*))^2\right)}{32(1152K^2\alpha^2+K_Y^4)}$$

and finally

$$C_3 = \frac{K_{\text{pen}}^2 \left(\sqrt{s_d(m^*+1)} - \sqrt{s_d(m^*)}\right)^2}{64 s_d(m^*+1)(72K^2\alpha^2 + K_Y^4)}.$$

This corresponds to Theorem 4.1 and the proof is complete. □

Remark 13. When A. Fermanian applies Proposition 4.17 in (4.39) she only considers

$$-\frac{C_4}{4} n(L(m) - L(m^*) - \text{pen}_n(m^*) + \text{pen}_n(m))$$

when actually the Proposition yields

$$-\frac{C_4}{4} n \left(L(m) - L(m^*) - \text{pen}_n(m^*) + \text{pen}_n(m)\right)^2.$$

As a result in [6] the next line contains

$$\frac{C_4}{4 \times 2} n(L(m^* - 1) - L(m^*))$$

and not

$$\frac{C_4}{4 \times 2^2} n \left(L(m^* - 1) - L(m^*)\right)^2$$

which should be the case. This changes C_2 compared to the original theorem, see Remark 7.

4.3 Corollary: Convergence rate of the mean square error

With the estimator of $\hat{m}$ we consequently are going to estimate $\beta^*_{m^*}$ by $\hat{\beta}_{\hat{m}}$ and hereby get an estimator for our regression model (3.1). With the tools we developed in order to proof Theorem 4.1 we can also proof that the mean square error decays with a rate of convergence $\mathcal{O}(n^{-1/2})$:

Corollary 4.18. *Under the assumptions of Theorem 4.1,*

$$\mathbb{E}\left[\left(\langle \hat{\beta}_{\hat{m}}, S^{\hat{m}}(X)\rangle - \langle \beta^*_{m^*}, S^{m^*}(X)\rangle\right)^2\right] = \mathcal{O}(n^{-1/2}).$$

A. Fermanian [6] points out that this convergence rate of $\mathcal{O}(n^{-1/2})$ is similar to the ones usually obtained for functional linear models when $d = 1$, except that less assumptions on the path X and its regularity are needed. While, indeed, we managed to obtain this rate of convergence for any $d \in \mathbb{N}$, we have to keep in mind that we did so by bounding the total variation of X in assumption (H_K), which is an assumption on the regularity of X.

Concerning the proof most of the work has already been done. We will use the results we have already established so far and rely heavily on the properties of the conditional expectation and the fact that per our assumption

$$\langle \beta^*_{m^*}, S^{m^*}(X)\rangle = \mathbb{E}[Y|X].$$

As a first step recall

$$R_m(\beta) = \mathbb{E}[(Y - \langle \beta, S^m(X)\rangle)^2].$$

Then, we note that we can rewrite the expectation in question as

$$\mathbb{E}\left[\left(\langle \hat{\beta}_{\hat{m}}, S^{\hat{m}}(X)\rangle - \langle \beta^*_{m^*}, S^{m^*}(X)\rangle\right)^2\right] = \mathbb{E}[R_{\hat{m}}(\hat{\beta}_{\hat{m}}) - R_{m^*}(\beta^*_{m^*})], \tag{4.42}$$

because

$$\begin{aligned}
&\mathbb{E}\left[\left(\langle \hat{\beta}_{\hat{m}}, S^{\hat{m}}(X)\rangle - \langle \beta^*_{m^*}, S^{m^*}(X)\rangle\right)^2\right] \\
&= \mathbb{E}\left[\left((Y - \langle \beta^*_{m^*}, S^{m^*}(X)\rangle) - (Y - \langle \hat{\beta}_{\hat{m}}, S^{\hat{m}}(X)\rangle)\right)^2\right] \\
&= \mathbb{E}\left[\left((Y - \langle \beta^*_{m^*}, S^{m^*}(X)\rangle)\right)^2\right] + \mathbb{E}\left[\left((Y - \langle \hat{\beta}_{\hat{m}}, S^{\hat{m}}(X)\rangle)\right)^2\right] \\
&\quad - 2\mathbb{E}\left[(Y - \langle \beta^*_{m^*}, S^{m^*}(X)\rangle)(Y - \langle \hat{\beta}_{\hat{m}}, S^{\hat{m}}(X)\rangle)\right] \\
&= R_{\hat{m}}(\hat{\beta}_{\hat{m}}) + R_{m^*}(\beta^*_{m^*}) - 2R_{m^*}(\beta^*_{m^*}) \\
&= \mathbb{E}[R_{\hat{m}}(\hat{\beta}_{\hat{m}}) - R_{m^*}(\beta^*_{m^*})],
\end{aligned}$$

where the last equality is true since

$$\begin{aligned}
&\mathbb{E}\left[(Y - \langle \beta^*_{m^*}, S^{m^*}(X)\rangle)(Y - \langle \hat{\beta}_{\hat{m}}, S^{\hat{m}}(X)\rangle)\right] \\
&= \mathbb{E}\left[Y^2 - Y\langle \beta^*_{m^*}, S^{m^*}(X)\rangle - Y\langle \hat{\beta}_{\hat{m}}, S^{\hat{m}}(X)\rangle + \langle \beta^*_{m^*}, S^{m^*}(X)\rangle\langle \hat{\beta}_{\hat{m}}, S^{\hat{m}}(X)\rangle\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[Y^2 - Y\langle \beta^*_{m^*}, S^{m^*}(X)\rangle - Y\langle \hat{\beta}_{\hat{m}}, S^{\hat{m}}(X)\rangle + \langle \beta^*_{m^*}, S^{m^*}(X)\rangle\langle \hat{\beta}_{\hat{m}}, S^{\hat{m}}(X)\rangle \middle| X\right]\right]
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}\left[\mathbb{E}\left[Y^2\middle| X\right] - \mathbb{E}\left[Y\middle| X\right]^2 - \langle\hat{\beta}_{\hat{m}}, S^{\hat{m}}(X)\rangle\mathbb{E}\left[Y\middle| X\right] + \mathbb{E}\left[Y\middle| X\right]\langle\hat{\beta}_{\hat{m}}, S^{\hat{m}}(X)\rangle\mathbb{E}\left[1\middle| X\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[Y^2\middle| X\right] - \mathbb{E}\left[Y\middle| X\right]^2\right] = \mathbb{E}\left[\mathrm{Var}(Y|X)\right] = \mathbb{E}\left[\mathbb{E}\left[\left(Y - \mathbb{E}\left[Y\middle| X\right]\right)^2\middle| X\right]\right] \\
&= \mathbb{E}\left[\left(Y - \langle\beta^*_{m^*}, S^{m^*}(X)\rangle\right)^2\right] = R_{m^*}(\beta^*_{m^*}).
\end{aligned}
$$

As $\hat{\beta}_{\hat{m}}$ minimizes the empirical risk $R_{\hat{m},n}$ it is clear that $R_{\hat{m},n}(\hat{\beta}_{\hat{m}}) - R_{\hat{m},n}(\beta^*_{\hat{m}}) \leq 0$ and we rearrange

$$
R_{\hat{m}}(\hat{\beta}_{\hat{m}}) - R_{m^*}(\beta^*_{m^*}) = R_{\hat{m}}(\hat{\beta}_{\hat{m}}) - R_{\hat{m}}(\beta^*_{\hat{m}}) + R_{\hat{m}}(\beta^*_{\hat{m}}) - R_{m^*}(\beta^*_{m^*}), \tag{4.43}
$$

which is true because:

$$
\begin{aligned}
R_{\hat{m}}(\hat{\beta}_{\hat{m}}) - R_{m^*}(\beta^*_{m^*}) &= R_{\hat{m}}(\hat{\beta}_{\hat{m}}) - R_{\hat{m}}(\beta^*_{\hat{m}}) + R_{\hat{m}}(\beta^*_{\hat{m}}) - R_{m^*}(\beta^*_{m^*}) \\
&= R_{\hat{m}}(\hat{\beta}_{\hat{m}}) - \hat{R}_{\hat{m},n}(\hat{\beta}_{\hat{m}}) + \hat{R}_{\hat{m},n}(\hat{\beta}_{\hat{m}}) - \hat{R}_{\hat{m},n}(\beta^*_{\hat{m}}) \\
&\quad + \hat{R}_{\hat{m},n}(\beta^*_{\hat{m}}) - R_{\hat{m}}(\beta^*_{\hat{m}}) + R_{\hat{m}}(\beta^*_{\hat{m}}) - R_{m^*}(\beta^*_{m^*}) \\
&\leq R_{\hat{m}}(\hat{\beta}_{\hat{m}}) - \hat{R}_{\hat{m},n}(\hat{\beta}_{\hat{m}}) + \hat{R}_{\hat{m},n}(\beta^*_{\hat{m}}) - R_{\hat{m}}(\beta^*_{\hat{m}}) \\
&\quad + R_{\hat{m}}(\beta^*_{\hat{m}}) - R_{m^*}(\beta^*_{m^*}) \\
&\leq 2 \sup_{\beta\in B_{\hat{m},\alpha}} |\hat{R}_{\hat{m}}(\beta) - R_{\hat{m}}(\beta)| \\
&\quad + R_{\hat{m}}(\beta^*_{\hat{m}}) - R_{m^*}(\beta^*_{m^*}).
\end{aligned}
$$

We will analyze the expectation of each summand separately in the following two lemmas.

Lemma 4.19. *Under the standing assumptions,*

$$
\mathbb{E}\left[\sup_{\beta\in B_{\hat{m},\alpha}} |\hat{R}_{\hat{m}}(\beta) - R_{\hat{m}}(\beta)|\right] = \mathcal{O}\left(\frac{1}{\sqrt{n}}\right).
$$

Proof. Corollary A.3 gives us an expression for this expectation for any $m \in \mathbb{N}$, because we know from Lemma 4.11 that $Z_{m,n}$ is subgaussian on $(B_{m,\alpha}, D)$. We further simplify the integral and the ϵ-covering number $N(B_{m,\alpha}, D, \epsilon)$ since we already found a bound for it in Proposition 4.13: For any fixed $m \in \mathbb{N}$,

$$
\begin{aligned}
\mathbb{E}\left[\sup_{\beta\in B_{m,\alpha}} |\hat{R}_m(\beta) - R_m(\beta)|\right] &\leq 12\int_0^\infty \sqrt{\log(N(B_{m,\alpha}, D, \epsilon)}d\epsilon \\
&= 36K\alpha\sqrt{s_d(m)}\sqrt{\frac{\pi}{n}}.
\end{aligned}
$$

For $m = \hat{m}$ (see remark 14) we obtain

$$\mathbb{E}\left[\sup_{\beta \in B_{\hat{m},\alpha}} |\hat{R}_{\hat{m}}(\beta) - R_{\hat{m}}(\beta)|\right] \leq 36K\alpha\sqrt{\frac{\pi}{n}}\mathbb{E}\left[\sqrt{s_d(\hat{m})}\right].$$

We compute this expectation splitting it in two sums and using Proposition 4.14:

$$\begin{aligned}
\mathbb{E}\left[\sqrt{s_d(\hat{m})}\right] &= \sum_{m \leq m^*} \sqrt{s_d(m)}\mathbb{P}(\hat{m} = m) + \sum_{m > m^*} \sqrt{s_d(m)}\mathbb{P}(\hat{m} = m) \\
&\leq (m^* + 1)\sqrt{s_d(m^*)} + \sum_{m > m^*} \sqrt{s_d(m)} 74 \exp\left(-C_3(n^{1-2\rho} + s_d(m))\right) \\
&\leq (m^* + 1)\sqrt{s_d(m^*)} + e^{-C_3 n^{1-2\rho}} \sum_{m > m^*} \sqrt{s_d(m)} 74 \exp\left(-C_3 s_d(m)\right) \\
&= \mathcal{O}(1).
\end{aligned}$$

□

Remark 14. It is not obvious that this equation should hold for $m = \hat{m}$. If $\hat{R}_{\hat{m}}(\beta)$ was independent from $\hat{m}$, the equation would be true. But this can hardly be the case since

$$\hat{m} = \min\left(\operatorname*{argmin}_{m \in \mathbb{N}}(\hat{L}_n(m) + pen_n(m))\right) \quad \text{with} \quad \hat{L}_n(m) = \min_{\beta \in B_{m,\alpha}} \hat{R}_{m,n}(\beta),$$

and, hence, both random variables depend on the realizations of (X_i, Y_i), $i = 1, ..., n$. But, we can show that the equation holds anyway if $Z_{\hat{m},n}$ is subgaussian. We already showed this for $Z_{m,n}$ and in general it would not be trivial to substitute a fixed m by the random variable $\hat{m}$. However, if we go back to Lemma 4.11 we see that we established this result mainly by using our assumptions (H_K) and in particular the only crucial appearance of m is in $S^m(X)$ which we subsequently bound by $e^{\|X\|_{TV}}$ according to proposition (2.5). Hence we actually get rid of the random variable for the rest of the proof and the above result holds true also for $\hat{m}$.

Lemma 4.20. *Using the notation introduced above,*

$$\mathbb{E}\left[R_{\hat{m}}(\beta^*_{\hat{m}}) - R_{m^*}(\beta^*_{m^*})\right] = \mathcal{O}\left(e^{-C_2 n^{1-2\rho}}\right),$$

where C_2 is defined in (4.41).

Proof. From assumptions (H_K), (H_α) and the Cauchy-Schwartz inequality we know that for every $m \in \mathbb{N}$,

$$\langle \beta^*_m, S^m(X)\rangle^2 \leq \|\beta\|^2 \|S^m(X)\|^2 \leq \alpha^2 e^{K_X}.$$

This will help us bound $\mathbb{E}\left[R_{\hat{m}}(\beta^*_{\hat{m}}) - R_{m^*}(\beta^*_{m^*})\right]$ because

$$\begin{aligned}
\mathbb{E}\left[R_{\hat{m}}(\beta^*_{\hat{m}}) - R_{m^*}(\beta^*_{m^*})\right] &= \mathbb{E}\left[\left(Y - \langle\beta^*_{\hat{m}}, S^{\hat{m}}(X)\rangle\right)^2 - \left(\underbrace{Y - \langle\beta^*_{m^*}, S^{m^*}(X)\rangle}_{:=\epsilon}\right)^2\right] \\
&= \mathbb{E}\left[\left(\langle\beta^*_{m^*}, S^{m^*}(X)\rangle + \epsilon - \langle\beta^*_{\hat{m}}, S^{\hat{m}}(X)\rangle\right)^2 - \epsilon^2\right] \\
&= \mathbb{E}\left[\left(\langle\beta^*_{m^*}, S^{m^*}(X)\rangle - \langle\beta^*_{\hat{m}}, S^{\hat{m}}(X)\rangle\right)^2 + \epsilon^2 - \epsilon^2\right. \\
&\quad \left. -2\epsilon\langle\beta^*_{m^*}, S^{m^*}(X)\rangle\langle\beta^*_{\hat{m}}, S^{\hat{m}}(X)\rangle\right] \\
&= \mathbb{E}\left[\left(\langle\beta^*_{m^*}, S^{m^*}(X)\rangle - \langle\beta^*_{\hat{m}}, S^{\hat{m}}(X)\rangle\right)^2\right] \\
&\quad - 2\mathbb{E}\left[\mathbb{E}\left[\epsilon\langle\beta^*_{m^*}, S^{m^*}(X)\rangle\langle\beta^*_{\hat{m}}, S^{\hat{m}}(X)\rangle\middle| X\right]\right] \\
&= \mathbb{E}\left[\left(\langle\beta^*_{m^*}, S^{m^*}(X)\rangle - \langle\beta^*_{\hat{m}}, S^{\hat{m}}(X)\rangle\right)^2\right] \\
&\quad - 2\mathbb{E}\left[\underbrace{\mathbb{E}\left[\epsilon\middle| X\right]}_{=\mathbb{E}[Y|X]-\mathbb{E}[Y|X]=0}\langle\beta^*_{m^*}, S^{m^*}(X)\rangle\langle\beta^*_{\hat{m}}, S^{\hat{m}}(X)\rangle\right] \\
&= \mathbb{E}\left[\left(\langle\beta^*_{m^*}, S^{m^*}(X)\rangle - \langle\beta^*_{\hat{m}}, S^{\hat{m}}(X)\rangle\right)^2\right] \\
&\leq 2\alpha^2 e^{K_X}\mathbb{P}(\hat{m} \neq m^*).
\end{aligned}$$

Of course, Theorem 4.1 gives us a bound for $\mathbb{P}(\hat{m} \neq m^*)$ and finally we have

$$\begin{aligned}
\mathbb{E}\left[R_{\hat{m}}(\beta^*_{\hat{m}}) - R_{m^*}(\beta^*_{m^*})\right] &\leq 2\alpha^2 e^{K_X}\mathbb{P}(\hat{m} \neq m^*) \\
&\leq 2\alpha^2 e^{K_X} C_1 e^{-C_2 n^{1-2\rho}} = \mathcal{O}\left(e^{-C_2 n^{1-2\rho}}\right).
\end{aligned}$$

□

By utilizing (4.42) and (4.43) we can apply both of the above lemmata and prove Corollary 4.18:

Proof of Corollary 4.18. With (4.42), (4.43), Lemma 4.19 and Lemma 4.20 we conclude

$$\mathbb{E}\left[\left(\langle\hat{\beta}_{\hat{m}}, S^{\hat{m}}(X)\rangle - \langle\beta^*_{m^*}, S^{m^*}(X)\rangle\right)^2\right] = \mathcal{O}\left(\frac{1}{\sqrt{n}}\right) + \mathcal{O}\left(e^{-C_2 n^{1-2\rho}}\right) = \mathcal{O}\left(\frac{1}{\sqrt{n}}\right).$$

□

Chapter 5

Numerical Experiments

In this chapter we conduct our numerical experiments. We are wondering if the signature regression, as it was introduced in (3.1) can reach the performance of the classical linear regression presented in (1.1). In [6] A. Fermanian found evidence that the signature regression can match and even beat the performance of various functional regression techniques without the need of making strong assumption like number and type of basis functions. She also demonstrated that the signature regression copes better with an increasing dimension of the explanatory paths than the functional regression techniques do. We will use the same synthetically generated data that Fermanian used in her work and compare the signature regression to linear regression. For the signature regression we will transform the generated data stream into a piecewise linear path and calculate its signature. For the linear regression on the other hand we will simply concatenate the d-dimensional stream into one long vector.

Our hope that the signature regression might be a appropriate alternative mainly stems from two observations. First, the signature can describe the evolution of a path. It captures a time component that raw data points can not. This additional information could be an advantage for the regression task. Second, as the number of data points increases the feature set of the linear regression increases too, because for every point we need to estimate one coefficient β_j. This makes the isolation of the "true" parameter harder as the number increases. The signature regression, on the other hand, only needs one coefficient for every signature term. Therefore, if we assume a fixed truncation order, the number of parameters of the signature regression stays the same no matter how fine the path is. This could be an indicator that the signature regression performs better for longer or finer paths, i.e. paths consisting of more data points.

To compare the performance of the two competing regression types, we will fix the dimension of the generated explanatory path at three and restrict ourselves to the path on the time interval $[0, 1]$, hence $X : [0, 1] \mapsto \mathbb{R}^3$. For our analysis we vary two parameters: The number of points (`nPoints`) sampled from the underlying process and the number of paths (`nPaths`) generated. The motive of varying the number of points was already mentioned above. Whilst varying the number of paths we expect both regressions to stabilize and increase their performance when the number of paths available for training increases if there is indeed some "true" relation that can be found.

For both types of regression, the number of features can be high compared to the number of observations. Either because of a high number of points in the linear regression case, or due to a high truncation order in the signature regression case. We might even obtain underdetermined systems for some combinations of (`nPaths`, `nPoints`). Therefore, it will be necessary to choose a regularization type for the task of finding the coefficient. In Chapter 3 we followed A. Fermanian [6] and decided to use a Ridge regression (see 3.3) implementing Thikonov-regularization, hence we will also use this approach for the linear regression. Since we do not know, what an appropriate choice for the radius of the α-balls in (H_α) would be (see 3.2), nor which λ parameter of the Ridge regression corresponds to this choice, we consider a range of possible λ between zero and 1000 and simply choose the best candidate for every regression by cross-validation (CV)[1]

In order to accomplish the best possible signature regression we will augment the paths with a time dimension before regressing. This ensures that our input data does not contain any time-reparametrisations or paths with loops. Additionally, we will add the base point zero to every path. Since the first order terms contain the increment of the respective coordinate, the otherwise translation invariant signature contains information about the absolute location of the path.

Although Chapter 3 provides an approach to choose an appropriate truncation order that is even backed theoretically by the theorems in Chapter 4, in the experiments we cannot use this particular method. It would involve specifying hyper-parameters like the penalization parameter K_{pen} for every combination (`nPaths`, `nPoints`), which is not feasible. Instead it will be easier to select the truncation order by 5-fold cross-validation (CV) on the training set, a typical machine learning procedure to avoid overfitting.

[1] In the implementation the parameter α does not appear, only the corresponding λ does. Therefore, Fermanian identifies λ and α in the implementation. We also follow this approach in the implementation and the upcoming explanations, so from now on the Ridge regression parameter λ will be denoted by α, which is not the radius of the α-ball anymore, but the corresponding Ridge parameter.

Considering everything stated so far, Algorithm 5.1 summarizes how we compare the performance of a signature- and linear regression on a particular data set.

Get or generate **Data** $\{(X_1, Y_1), ..., (X_n, Y_n)\}$
Augment X_i in **Data** with time dimension and base point
Split **Data** into **Train** and **Test** set
procedure SELECT $\hat{m}$ VIA CV(**Train**)
    Calculate m_{max} ▷ Such that $s_d(m) \leq 10.000$
    for $0 \leq m \leq m_{max}$ **do**
        Ridge Regression on $\{S^m(X_i), Y_i\}$ in **Train** ▷ Ridge-α by CV
        Measure performance by 5-fold cross-validation
    end for
    Choose best performing m as $\hat{m}$
end procedure
procedure COMPARE REGRESSION TYPES(**Train**, **Test**, $\hat{m}$)
    Fit Ridge Regression to $\{S^{\hat{m}}(X_i), Y_i\}$ in **Train** ▷ Here we get get $\hat{\beta}^{\hat{m}}$
    With $\hat{m}$ and $\hat{\beta}^{\hat{m}}$ predict $\{\hat{Y}_i\}$ from $\{S^{\hat{m}}(X_i)\}$ in **Test**
    Measure prediction performance of Signature Regression (e.g. MSE, R^2,...)
    For linear regression reshape X_i in **Train** and **Test** into 1-dim vectors $\tilde{X}_i$
    Fit Ridge Regression to $\{\tilde{X}_i, Y_i\}$ in **Train**
    Predict $\{\hat{Y}_i\}$ from $\{\tilde{X}_i\}$ in **Test**
    Measure prediction performance of Linear Regression (e.g. MSE, R^2,...)
end procedure

Fig. 5.1 Algorithm to compare regression types

Remark 15. Due to the decaying norm of signature terms (2.5) it seems appropriate to apply feature normalization in all the Ridge regressions on the signature in the above algorithm. Exploiting this typical machine learning procedure avoids inheriting a bias towards lower order terms.

In the following experiments with synthetically generated data we will repeat Algorithm 5.1 twenty times for every combination of (`nPaths`, `nPoints`). Hence we do not only get one single performance value per combination but even multiple ones. This allows us to compute mean and variance of the performance.

The analysis will cover two aspects in particular. On one hand, it will be interesting to see, which truncation order $\hat{m}$ is chosen for the signature regression. On the other hand, we want to judge how good the regression performed. To do so the measure

of choice is the R^2-value. R^2 is calculated as $1 - \frac{u^2}{\sigma^2}$, where u^2 is the mean squared error resulting from the predictions on the test set, while σ^2 is the variance of the true responses in the test set. Compared to the mean squared error the R^2 value enjoys the benefit of putting the mean squared error into perspective to the variance of the responses. A mean squared error of 100 for example would be very high for responses in the range [-1,1], whereas it would be small for responses in the range $[-1000, 1000]$. The following interpretation of the R^2 value could be appropriate:

- $R^2 = 1$: perfect prediction
- $R^2 > 0$: the mean squared error is lower than the variance of the responses. We are better than always guessing the average response.
- $R^2 < 0$: the predictions are worse than always guessing the average response. Our model might not add valuable information to the prediction task.

Remark 16. The formulation "guessing the average response" is in some sense misleading. The only average response we could calculate is the average response on the training set, while σ^2 is the variance of the responses in the validation set, which we do not know at the time of fitting our model. This is the reason why interpreting the R^2-value as a value comparing one model with the constant model, which is predicting the average response, is not accurate. It rather tells us in percentage how much of the variance of the responses we were able to reduce by exploiting our model.

5.1 Smooth paths, signature model response

In the first experiment we generate smooth paths $X_i : [0, 1] \mapsto \mathbb{R}^d$ and more importantly we will generate the response using the signature linear model (3.1). This means that the responses Y_i will be calculated as a (random) linear function of the signature truncated at some fixed level. With this experiment we can check if selecting $\hat{m}$ by cross-validation manages to find the true truncation level. Furthermore, we will get an impression of how the linear regression performs if the response actually follows the signature linear model.

For $1 \leq i \leq n$, the smooth paths $X_i : [0, 1] \to \mathbb{R}^d$, $X_{i,t} = (X^1_{i,t}, ..., X^d_{i,t})$ are defined by

$$X^k_{i,t} = \alpha^k_{i,1} + 10\alpha^k_{i,2} \sin\left(\frac{2\pi t}{\alpha^k_{i,3}}\right) + 10(t - \alpha^k_{i,4})^3, \quad 1 \leq k \leq d, \tag{5.1}$$

where $\alpha^k_{i,l}$, $1 \leq l \leq 4$ are sampled uniformly on $[0, 1]$.

For the responses we choose the truncation level $m^* = 5$. Then

$$Y_i = \langle \beta, S^{m^*}(X_i) \rangle + \epsilon_i, \tag{5.2}$$

where ϵ_i uniformly on [-100,100], and β is given by

$$\beta_j = \frac{1}{1000} u_j, \quad 1 \leq j \leq s_d(m^*), \tag{5.3}$$

with u_j sampled uniformly on [0,1].

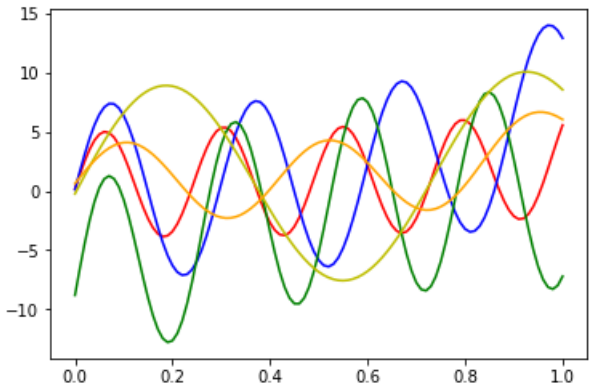

Fig. 5.2 One instance of a 5-dimensional smooth path

In Figure 5.2 we can see all components of one instance of a $d = 5$-dimensional path. The number of points in the chart is 100 and indeed the paths seem to be "smooth". However, this is not surprising, since the paths are the sum of a scaled sine function and an order three polynomial term. We will consider noisier paths later on.

5.1.1 Results

As stated earlier we choose three dimensional paths ($d = 3$) for the experiments. We consider `nPoints` $\in [3, 5, 10, 20, 50, 100]$ and `nPaths` $\in [33, 50, 100, 200, 500, 1000]$. The ratio for the train-test-split is 50:50.

In Figure 5.3 we can see that the on average selected truncation level seems to increase both as `nPoints` and `nPaths` increase. In the right bottom corner of the table we are already close to the true truncation order of five. The standard deviation in Figure 5.4 seems to decrease as `nPaths` increases, which makes sense intuitively, but it also seems to increase from left to right, i.e. as `nPoints` increases. One possible

explanation is that the signature is a much more complex object if it is driven by a path consisting of a higher number of points. Following this intuition we can understand, why for only three or five points (first two columns) the selected truncation order even for many paths is three and not five with a relatively small standard deviation. It seems evident that a signature from a path consisting of merely three or five points does not contain enough information to produce higher order signature terms that are helpful for the regression task.

nPaths \ nPoints	3	5	10	20	50	100
33	1.20	1.45	2.50	2.15	2.70	2.30
50	2.25	2.25	2.10	2.25	2.25	2.30
100	2.35	2.60	2.95	2.40	2.75	2.60
200	3.05	2.90	3.35	2.50	3.65	3.35
500	3.25	3.25	3.35	4.10	3.95	4.80
1000	3.15	3.10	4.20	4.95	5.05	4.85

Fig. 5.3 $\hat{m}$ average for smooth paths and signature model response

nPaths \ nPoints	3	5	10	20	50	100
33	1.36	1.99	1.94	1.98	2.30	2.08
50	1.58	1.18	1.84	2.28	2.21	2.05
100	1.11	1.50	1.99	2.01	1.79	1.91
200	0.86	0.89	1.42	1.94	1.82	1.85
500	0.43	0.43	1.46	1.37	1.69	1.25
1000	0.36	0.44	1.08	0.38	0.38	0.73

Fig. 5.4 $\hat{m}$ std for smooth paths and signature model response

The figures containing average R^2 (5.5) and its standard deviation (5.6) in the linear regression case clearly suggest that this regression only works (= R^2 value is positive with relatively small std.) in the left bottom corner, where `nPoints` is small and `nPaths` high.

On the other hand, the average R^2 values for the signature regression (5.7, 5.8) do not seem to depend on `nPoints` to the same extend. Moreover, the signature regression appears to require a higher `nPaths` compared to the linear regression. This would explain why the best results can be found in the last two rows corresponding to `nPaths` equals 500 or 1000, and further why for good results the standard deviation of the signature regression still tends to be higher compared to linear regression.

nPaths \ nPoints	3	5	10	20	50	100
33	-0.62	-4.98	-0.63	-0.88	-1.04	-1.17
50	-0.07	-0.84	-0.76	-0.80	-1.49	-2.19
100	0.02	-0.23	-0.58	-0.69	-1.59	-3.19
200	0.09	0.04	-0.12	-0.89	-1.36	-1.09
500	0.21	0.17	0.04	-0.23	-0.70	-0.73
1000	0.23	0.19	0.09	-0.01	-0.29	-1.75

Fig. 5.5 R^2 average of linear regression for smooth paths and signature model response

nPaths \ nPoints	3	5	10	20	50	100
33	0.95	15.39	0.70	0.86	1.74	1.57
50	0.30	1.15	1.32	1.09	3.06	6.56
100	0.21	0.28	0.58	0.48	2.35	5.38
200	0.15	0.15	0.19	0.68	1.89	1.05
500	0.06	0.06	0.09	0.14	0.73	0.73
1000	0.04	0.05	0.05	0.07	0.42	4.43

Fig. 5.6 R^2 std of linear regression for smooth paths and signature model response

nPaths \ nPoints	3	5	10	20	50	100
33	-0.16	-0.15	-0.04	-0.40	-0.28	-0.18
50	-0.18	-0.27	-0.04	-0.30	-0.22	-0.17
100	0.04	-0.10	-0.06	-0.11	-0.22	-0.48
200	0.16	0.09	0.09	0.04	-0.05	0.08
500	0.33	0.27	0.12	0.11	0.15	0.19
1000	0.42	0.33	0.25	0.41	0.48	0.33

Fig. 5.7 R^2 average of signature regression for smooth paths and signature model response

nPaths \ nPoints	3	5	10	20	50	100
33	0.33	0.23	0.27	0.53	0.51	0.44
50	0.44	0.75	0.22	0.59	0.52	0.35
100	0.22	0.34	0.29	0.28	0.61	1.09
200	0.18	0.27	0.16	0.12	0.27	0.24
500	0.12	0.10	0.13	0.25	0.41	0.38
1000	0.07	0.09	0.13	0.18	0.15	0.48

Fig. 5.8 R^2 std of signature regression for smooth paths and signature model response

It is not surprising that in general the signature regression works better in this example, since we generated the response following exactly the model that a signature regression tries to find. The fact that the linear regression works at all for small `nPoints` suggests that the signature and the raw data points of a path consisting of only few points capture similar information or put differently, linear functions of a small number of data points are not too different from linear functions of their signature (if we create a path from them by linear interpolation).

5.2 Smooth paths, maximum response

In the next example we will generate the paths in exactly the same manner as before, but the response will not follow the signature model anymore. Instead, we sample one additional time point for $X_{i,T+1}, 1 \leq i \leq n$, and set the response for every path to the maximum component at this additional time point:

$$Y_i = \max(X^1_{i,T+1}, ..., X^d_{i,T+1}).$$

In other words, we try to predict the maximum component of the path at time $T+1$ from the path up to time T.

5.2.1 Results

First, we have a look at the selected truncation level in Figure 5.9. Of course, this time we do not know the true value. It is notable that for enough paths the most favourable value with a small standard deviation (5.10) seems to be three, while for a small amount of data the results are still more erratic, i.e. also higher and lower values are considered with higher deviation.

The R^2 values of the linear regression (5.11) indicate that this regression requires 500 or more paths to consistently contribute meaningful information to the prediction task (last two rows). At the same time the signature regression already works for 200 paths (5.13), however, only if the number of points is high (20 or more). This result encourages our conjecture that the signature regression works better for finer paths. Looking at the milieu where both regressions do not work properly (100 paths or less) we can clearly see that both mean and standard deviation (5.12, 5.14) of the signature regression are worse. If the signature regression goes wrong, it tends to do that in a much more exaggerated way than the linear regression.

nPaths \ nPoints	3	5	10	20	50	100
33	3.45	2.35	3.60	2.45	4.65	4.70
50	4.70	4.85	4.25	2.60	5.50	5.70
100	4.45	4.95	4.75	2.05	4.85	3.70
200	3.25	3.40	3.65	2.15	3.30	3.10
500	3.00	3.00	3.00	2.50	3.00	3.00
1000	3.05	3.00	3.00	2.75	3.00	3.00

Fig. 5.9 $\hat{m}$ average for smooth paths and maximum response

nPaths \ nPoints	3	5	10	20	50	100
33	1.36	1.99	1.94	1.98	2.30	2.08
50	1.58	1.18	1.84	2.28	2.21	2.05
100	1.11	1.50	1.99	2.01	1.79	1.91
200	0.86	0.89	1.42	1.94	1.82	1.85
500	0.43	0.43	1.46	1.37	1.69	1.25
1000	0.36	0.44	1.08	0.38	0.38	0.73

Fig. 5.10 $\hat{m}$ std for smooth paths and maximum response

5.3 Gaussian paths, trend slope response

As a last example with synthetically generated data, we consider noisy data to see how both regression type deal with this challenge. The paths will component-wise consist of a drift term and a Gaussian term. Formally, for $1 \leq i \leq n$, let $X_i : [0,1] \to \mathbb{R}^d$, $X_{i,t} = (X^1_{i,t}, ..., X^d_{i,t})$ be defined by

$$X^k_{i,t} = \alpha^k_i t + \epsilon^k_{i,t} \tag{5.4}$$

where α^k_i, is sampled uniformly on $[-3,3]$ and ϵ^k_i is a Gaussian process with exponential covariance matrix.

An example of such a path can be found in Figure 5.15. The response we want to extrapolate is the norm of the drift vector

$$Y_i = \|a_i\|. \tag{5.5}$$

nPaths \ nPoints	3	5	10	20	50	100
33	-0.28	-2.61	-0.33	-0.40	-0.41	-0.13
50	-0.21	-0.41	-0.60	-0.30	-0.23	-0.41
100	-0.06	-0.47	-0.40	-0.26	-0.21	-0.41
200	0.08	-0.02	-0.06	-0.21	0.06	-0.03
500	0.09	0.08	0.07	0.09	0.22	0.23
1000	0.13	0.11	0.16	0.23	0.20	0.36

Fig. 5.11 R^2 average of linear regression for smooth paths and maximum response

nPaths \ nPoints	3	5	10	20	50	100
33	0.41	8.23	0.54	0.81	0.81	0.63
50	0.32	0.43	0.84	0.34	0.38	0.49
100	0.29	1.00	0.33	0.29	0.42	0.43
200	0.07	0.10	0.17	0.21	0.16	0.23
500	0.06	0.06	0.10	0.11	0.11	0.13
1000	0.03	0.04	0.04	0.05	0.11	0.09

Fig. 5.12 R^2 std of linear regression for smooth paths and maximum response

nPaths \ nPoints	3	5	10	20	50	100
33	-4.40	-4.54	-3.87	-2.29	-3.17	-3.73
50	-2.67	-2.35	-2.88	-0.65	-2.32	-2.13
100	-2.61	-1.67	-1.66	-0.04	-0.96	-1.16
200	-4.53	-0.63	-0.66	0.15	0.24	0.25
500	-0.18	-0.10	0.00	0.27	0.52	0.66
1000	0.01	0.01	0.10	0.33	0.58	0.72

Fig. 5.13 R^2 average of signature regression for smooth paths and maximum response

nPaths \ nPoints	3	5	10	20	50	100
33	1.73	1.86	1.47	2.06	1.89	1.82
50	1.17	1.00	1.26	0.77	1.49	1.48
100	3.43	1.38	1.12	0.32	1.13	1.75
200	15.00	0.35	1.05	0.11	0.30	0.24
500	0.10	0.12	0.10	0.07	0.05	0.06
1000	0.06	0.05	0.06	0.04	0.05	0.04

Fig. 5.14 R^2 std of signature regression for smooth paths and maximum response

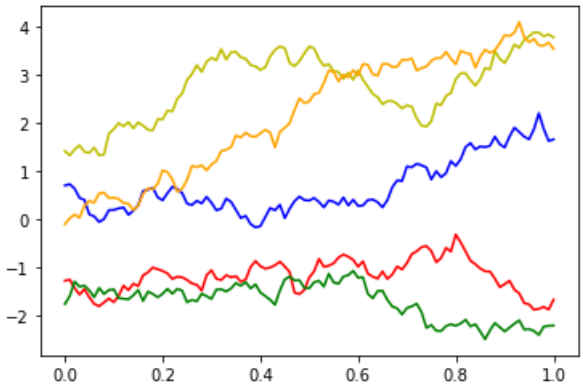

Fig. 5.15 One instance of a 5-dimensional Gaussian path

5.3.1 Results

For this particular task the truncation order selected by cross-validation was small again (5.16). For a sufficiently large number of paths the best results were obtained by truncating at level three or two (depending on `nPoints`) mostly without any deviation (5.17).

nPaths \ nPoints	3	5	10	20	50	100
33	1.95	4.10	5.30	5.10	2.50	2.20
50	5.50	5.60	6.00	5.40	2.40	2.30
100	4.55	4.75	4.70	4.40	2.00	2.05
200	3.00	3.00	3.00	3.90	2.00	2.00
500	3.00	3.00	3.00	3.00	2.00	2.00
1000	3.00	3.00	3.00	3.00	2.05	2.00

Fig. 5.16 $\hat{m}$ average for Gaussian paths and drift response

nPaths \ nPoints	3	5	10	20	50	100
33	2.73	2.30	1.52	2.14	1.24	0.87
50	0.87	0.66	0.00	0.92	1.20	0.90
100	1.43	1.44	1.42	1.46	0.00	0.22
200	0.00	0.00	0.00	1.37	0.00	0.00
500	0.00	0.00	0.00	0.00	0.00	0.00
1000	0.00	0.00	0.00	0.00	0.22	0.00

Fig. 5.17 $\hat{m}$ std for for Gaussian paths and drift response

The R^2 values of the linear regression (5.18) draw a bleak picture. The values attained – even though with very small standard deviation – are all negative. This might be an indicator that linear regression is not able to find the response in this case. Meanwhile the signature regression performs better. R^2 values are positive for 500 or 1000 paths. If the number of points is high enough, e.g. for 200 paths, the regression is successful. These results are backed by relatively small deviations illustrated in Figure 5.21.

This last experiment suggests that linear regression does not cope well with noisy paths, while the signature is still able to extract valuable information.

5.4 Credit Cycle Forecasting

In this last section we will investigate an example with real-world data. What we are going to do is called "Credit Cycle Forecasting". In the context of IFRS 9 reporting standards, a lending financial institution is required to forecast the "credit cycle" over a forward time horizon up to the lifespan of its credit products. Multiple time series are typically used (GDP, unemployement, interest rates, etc.) as leading variables in order to estimate and predict the "credit cycle" by the use of standard econometric models (e.g. ARCH, ARMA, GARCH,...). The aim of this part of the thesis is, to apply signature regression in this specific context, where the "credit cycle" is the response variable to be regressed against the truncated signature of the multidimensional path of explanatory variables.

One possible proxy of this "credit cycle" is the probability of default for a region or sector. We select as response variable the quarterly probability of default for North American firms from 1990 until 2021 as published by the Credit Research Initiative of the National University of Singapore [15]. As explanatory variables we consider US GDP growth, US unemployment rate, S&P 500 growth and US interest rate spread, i.e. long term minus short term interest rate.[2] Prior to calculating the signature we transform the data as follows:

[2]Data sources accessed on January 31, 2022:
- GDP: https://fred.stlouisfed.org/series/GDP
- Unemployment: https://fred.stlouisfed.org/series/LRUN64TTUSQ156S
- S&P500: https://finance.yahoo.com/quote/%5EGSPC/history
- Short term IR: https://data.oecd.org/interest/short-term-interest-rates.htm#
- Long term IR: https://data.oecd.org/interest/long-term-interest-rates.htm#indicator-chart
- PDs: https://nuscri.org/en/data/cdsaggregatedata/e503s0/0/

nPaths \ nPoints	3	5	10	20	50	100
33	-0.68	-0.55	-0.48	-0.88	-1.32	-1.13
50	-0.18	-0.24	-0.31	-0.34	-0.39	-0.52
100	-0.14	-0.12	-0.13	-0.12	-0.27	-0.25
200	-0.07	-0.04	-0.09	-0.09	-0.14	-0.11
500	-0.02	-0.03	-0.04	-0.04	-0.08	-0.08
1000	-0.02	-0.02	-0.02	-0.03	-0.04	-0.06

Fig. 5.18 R^2 average of linear regression for Gaussian paths and drift response

nPaths \ nPoints	3	5	10	20	50	100
33	1.78	0.76	0.81	1.78	2.11	1.41
50	0.16	0.23	0.33	0.42	0.49	0.48
100	0.16	0.13	0.12	0.13	0.29	0.21
200	0.06	0.04	0.09	0.09	0.10	0.10
500	0.02	0.02	0.03	0.03	0.03	0.04
1000	0.01	0.01	0.02	0.01	0.01	0.02

Fig. 5.19 R^2 std of linear regression for Gaussian paths and drift response

nPaths \ nPoints	3	5	10	20	50	100
33	-14.66	-8.37	-8.28	-7.64	-9.34	-5.75
50	-6.95	-5.45	-4.57	-5.76	-0.99	-1.75
100	-3.65	-2.72	-2.44	-3.06	-0.09	-0.15
200	-0.37	-0.42	-0.55	-0.75	0.13	0.16
500	0.05	0.04	0.02	0.03	0.23	0.21
1000	0.17	0.17	0.18	0.17	0.25	0.25

Fig. 5.20 R^2 average of signature regression for Gaussian paths and drift response

nPaths \ nPoints	3	5	10	20	50	100
33	9.32	4.54	5.38	2.46	16.02	6.60
50	2.67	2.07	1.61	2.54	1.63	2.66
100	3.55	1.34	1.18	1.84	0.37	0.31
200	0.34	0.29	0.54	0.64	0.11	0.11
500	0.08	0.10	0.08	0.08	0.04	0.05
1000	0.05	0.07	0.05	0.05	0.04	0.03

Fig. 5.21 R^2 std of signature regression for Gaussian paths and drift response

- The explanatory variables are normalized to have zero mean and unit variance.
- The probabilities of default are centered and then scaled such that the highest absolute value is 1. This range [-1,1] fits our interpretation of the PD as a "credit cycle" indicator ranging from a "very favourable" to "very adverse" environment to lend money.

We split the quarterly sampled explanatory variables into rolling windows of varying length, ranging from window size 3 to 16 quarters (4 years). From this windows we try to predict future PDs on a forecast horizon from one quarter up 24 quarters (6 years). Additional to the explanatory variables mentioned above for our prediction task, the past probabilities of default are available, too. Hence, we include the lagged variable in the explanatory variables. One example of the explanatory variables with window size 12 can be found in Figure 5.22.

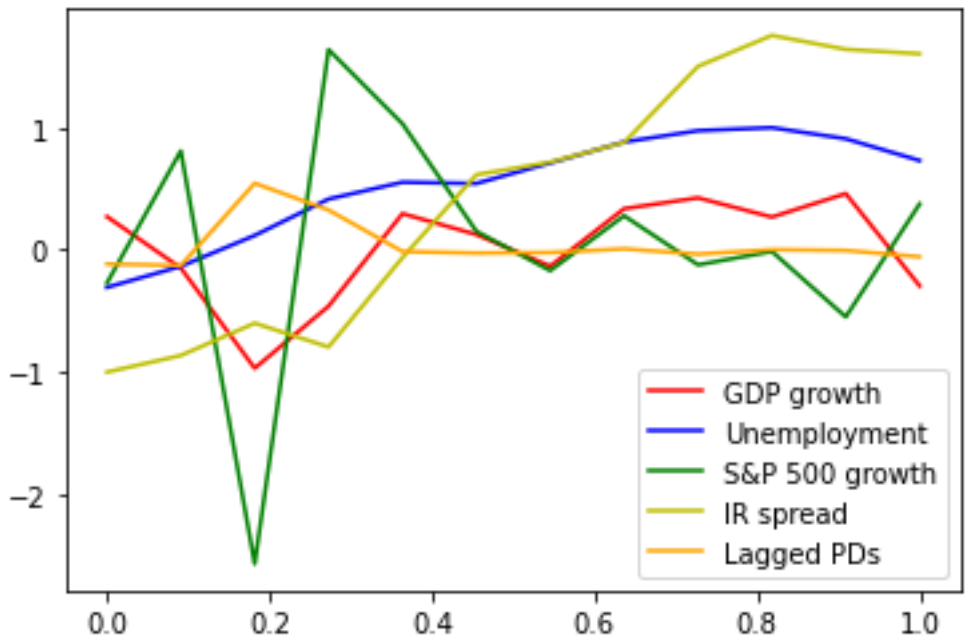

Fig. 5.22 One instance of the four transformed explanatory variables with window size 12 scaled to time $[0, 1]$

In this real-world example we obviously do not have to worry about choosing the number of paths or the number of points for generating our paths. The window size loosely corresponds to the number of points from previous examples, even though making the path longer is not exactly the same as making the path finer. Depending on forecasting horizon and window size, the quarterly data between 1990 and 2021 gives us roughly 120 observations to train and test on. Regarding the time augmentation of

the paths we will not worry about the actual dates and just map the data into the time interval [0,1]. Instead of generating new data for every iteration we randomly select differing train-test splits of the same data in order to obtain mean and variance of the performance.

5.4.1 Results

Analyzing the results after iterating Algorithm 5.1 twenty times, the first thing to notice is that the truncation orders seem to be small again. For short term forecasts of only one or two quarters, values are close to one. For longer forecast horizons values increase up to two, but never above. We know this since all appearances of two in Figure 5.23 correspond to zero standard deviation in (5.24). Therefore, the value is indeed always two and not only two on average.

Forecasting Horizon

Window Size	1	2	3	4	5	6	8	10	12	16	20	24
3	1.05	1.20	1.15	1.25	1.40	1.40	1.85	1.85	1.85	1.95	1.95	1.85
4	1.10	1.20	1.10	1.20	1.20	1.65	2.00	2.00	1.95	1.80	1.90	1.80
6	1.00	1.15	1.20	1.50	1.80	1.80	1.90	1.95	1.85	1.95	1.90	1.85
8	1.10	1.30	1.25	1.40	1.55	1.80	2.00	1.95	1.70	1.95	2.00	1.70
12	1.00	1.15	1.50	1.60	1.50	1.60	1.90	1.75	1.90	1.90	2.00	1.80
16	1.00	1.40	1.70	1.60	1.55	1.65	1.95	2.00	1.85	2.00	1.95	1.80

Fig. 5.23 $\hat{m}$ average for "credit-cycle" forecasting

Forecasting Horizon

Window Size	1	2	3	4	5	6	8	10	12	16	20	24
3	0.22	0.51	0.36	0.43	0.49	0.49	0.36	0.36	0.36	0.22	0.22	0.36
4	0.30	0.40	0.30	0.40	0.40	0.48	0.00	0.00	0.22	0.40	0.30	0.40
6	0.00	0.36	0.40	0.50	0.40	0.40	0.30	0.22	0.36	0.22	0.30	0.36
8	0.30	0.46	0.43	0.49	0.59	0.40	0.00	0.22	0.46	0.22	0.00	0.46
12	0.00	0.36	0.50	0.49	0.50	0.49	0.30	0.43	0.30	0.30	0.00	0.40
16	0.00	0.58	0.46	0.49	0.50	0.48	0.22	0.00	0.36	0.00	0.22	0.51

Fig. 5.24 $\hat{m}$ std for "credit-cycle" forecasting

The performance of the linear regression is good. R^2 values (5.25) are positive with only one exception, which could be the result of a couple of disadvantageous train-test splits, as the standard deviation is very high (5.24). The best performances (≥ 0.4) are achieved for relatively long windows of 12 and 16 combined with forecast horizons

of 10 or smaller. For long term forecasts (12 or more quarters), the performance seems to decline, even for large windows.

The signature regression delivers a different result. There is a visible border between R^2 values of forecast horizons ≥ 8 and ≤ 6 (see 5.27). The shorter horizon forecasts sometimes perform very well ($R^2 \geq 0.5$). However, the results on this half of the table are corrupted by multiple faulty regressions with negative R^2 scores below minus one and extraordinary large standard deviations of five or more (5.28). These values suggest that short-term forecasts based on the signature are erratic and depend a lot on the respective train-test split. For long-term forecasts on the other hand, the signature seems to be a more useful tool. Across all window sizes the signature regression outperforms the linear regression for forecast horizons of eight quarters (2 years) or longer. The stability of longer forecasts is supported by relatively small standard deviations.

It makes sense that the explanatory variables can not only explain the upcoming PDs, but also PDs further in the future. After all, the default of a firm is usually not a spontaneous event, but rather the result of multiple years of economic descent or mismanagement. However, it is surprising that only the signature regression struggles so much with short-term predictions and that the border where the signature regression starts working reliably is so prominent between six and eight quarters independent of the window size. Meanwhile, the linear regression manages the short-term forecasts, but is not able perform similarly successful regarding long-term forecasts.

We conclude this analysis by investigating the regression coefficients to gain insights which signature terms are selected in order to predict the "credit cycle". To this end we compare two of the best performances of the signature regression that coincidentally share the window size of 16, namely the one quarter forecast and the eight quarter (2 years) forecast.

Figures 5.23 and 5.24 reveal that for the window-16-horizon-1 forecast the regression selected truncation level 1 for all 20 iterations of the experiment. The truncation level for the window-16-horizon-8 forecast was on average 1.95, which is attained by a single order one choice and 19 order two choices. We do not want to mix the average values for signatures of different sizes (as we will see shortly they behave differently), hence we discard the single order one choice and calculate the average coefficients of the 19

Forecasting Horizon

Window Size	1	2	3	4	5	6	8	10	12	16	20	24
3	0.38	0.16	0.22	0.13	0.12	0.32	0.35	0.36	0.34	0.12	0.08	0.16
4	0.23	0.28	0.23	0.16	0.20	0.25	0.40	0.29	0.29	0.14	0.11	0.24
6	0.23	0.31	0.25	0.30	0.30	0.38	0.42	0.41	0.28	0.17	0.26	0.13
8	0.07	0.33	0.29	0.33	0.37	0.38	0.39	0.34	0.27	0.19	0.29	0.12
12	0.46	0.36	0.44	0.44	0.42	0.45	0.48	0.37	0.20	0.24	0.11	0.15
16	0.45	0.46	0.42	0.43	0.36	0.38	0.54	0.41	0.27	0.19	-0.40	0.07

Fig. 5.25 R^2 average of linear regression for "credit-cycle" forecasting

Forecasting Horizon

Window Size	1	2	3	4	5	6	8	10	12	16	20	24
3	0.47	0.33	0.10	0.15	0.33	0.16	0.12	0.07	0.09	0.20	0.16	0.18
4	0.72	0.15	0.14	0.25	0.27	0.36	0.10	0.19	0.09	0.11	0.22	0.09
6	0.39	0.17	0.17	0.09	0.12	0.11	0.11	0.11	0.08	0.09	0.19	0.29
8	0.68	0.18	0.22	0.10	0.19	0.20	0.13	0.19	0.19	0.25	0.27	0.22
12	0.14	0.25	0.08	0.10	0.12	0.08	0.07	0.10	0.28	0.31	0.38	0.16
16	0.10	0.14	0.17	0.08	0.23	0.32	0.12	0.22	0.39	0.31	1.87	0.52

Fig. 5.26 R^2 std of linear regression for "credit-cycle" forecasting

Forecasting Horizon

Window Size	1	2	3	4	5	6	8	10	12	16	20	24
3	0.40	-1.52	0.29	0.16	0.05	-0.22	0.39	0.42	0.42	0.26	0.13	0.19
4	-1.31	0.27	0.30	0.17	0.21	0.09	0.49	0.32	0.42	0.30	0.25	0.37
6	0.60	-0.51	0.20	0.18	0.28	0.33	0.50	0.53	0.40	0.35	0.35	0.19
8	-0.29	-1.60	0.12	0.24	0.26	0.07	0.43	0.44	0.37	0.41	0.50	0.16
12	0.54	-0.08	-0.90	0.25	0.24	-0.02	0.53	0.42	0.34	0.31	0.38	0.26
16	0.59	0.16	-0.44	0.29	0.13	0.11	0.60	0.35	0.39	0.39	0.39	0.17

Fig. 5.27 R^2 average of signature regression for "credit-cycle" forecasting

Forecasting Horizon

Window Size	1	2	3	4	5	6	8	10	12	16	20	24
3	0.91	5.54	0.17	0.23	0.34	1.39	0.13	0.09	0.09	0.15	0.44	0.30
4	8.48	0.56	0.13	0.48	0.13	0.55	0.09	0.39	0.15	0.13	0.17	0.10
6	0.09	3.79	0.28	0.43	0.08	0.11	0.11	0.14	0.10	0.10	0.16	0.45
8	2.80	5.85	0.70	0.11	0.20	1.11	0.24	0.11	0.10	0.12	0.07	0.35
12	0.37	1.62	4.43	0.29	0.21	0.97	0.09	0.11	0.17	0.28	0.12	0.24
16	0.13	0.89	1.51	0.14	0.73	0.80	0.10	0.36	0.16	0.13	0.13	0.39

Fig. 5.28 R^2 std of signature regression for "credit-cycle" forecasting

order two signatures in the window-16-horizon-8 case and the average coefficient of the 20 order one signatures in the window-16-horizon-1 case. The resulting coefficients can be compared in figure 5.29.

To better understand the coefficients let us recall from Chapter 2 which information the signature contains in both cases. After time and base-point (=0) augmentation the paths are six dimensional:

1. GDP growth,
2. Unemployment,
3. S&P 500 growth,
4. Interest-rate spread,
5. Lagged PDs,
6. Time.

Hence the first order terms correspond to

$$S(X)^i_{0,1} = X^i_1 - \underbrace{X^i_0}_{=0} = X^i_1, \tag{5.6}$$

which is simply the last point of each of the six components. The second order terms can be calculated making use of the iterative structure of the signature and we get

$$\begin{aligned} S(X)^{(i,j)}_{0,1} &= \iint\limits_{0\le u_1\le u_2\le 1} dX^i_{u_1} dX^j_{u_2} = \int_{0\le u_2\le 1} \underbrace{\int_{0\le u_1\le u_2} dX^i_{u_1}}_{=S(X)^i_{0,u_2}} dX^j_{u_2} \\ &= \int_{0\le u_2\le 1} S(X)^i_{0,u_2}\, dX^j_{u_2} = \int_{0\le u_2\le 1} X^i_{u_2}\, dX^j_{u_2}. \end{aligned}$$

In particular for j=6 the terms equal

$$S(X)^{(i,6)}_{0,1} = \int_{0\le u_2\le 1} X^i_{u_2}\, dt. \tag{5.7}$$

In figure 5.29 we first look at the six coefficients of the horizon-8 forecast. It is not surprising that the signature term corresponding to index (5) has a significant positive coefficient, in fact the biggest one. Term (5) corresponds to the lagged probability of default and it seems intuitive that the following probability of default should be close to this value. The terms (3) and (4) have comparably high negative coefficients, (1) and (2) seem not to play an important role. Term (6) corresponding to the time dimension

is zero, which makes sense. The constant final time $t = 1$ should not contain important information for the prediction task, since we already have an intercept corresponding to the constant signature term of level zero, which omit in the present analysis.

Actually, this is one of only few similarities between the coefficients of the horizon-1 and the horizon-8 forecast. Apart from term (6) being zero, we see that term (4) has a similar important negative coefficient and again (2) does not seem to be important. All other level 1 term coefficients are different. Especially the lagged PD has a small negative coefficient for the long-term forecast, opposed to a big positive for the short-term forecast.

Searching for the second order terms that are important for the long-term forecast it stands out that (1,6), (3,6), (4,4), (4,6) and (5,6) have a deep impact. According to (5.7) the (i,6) terms contain information about the evolution of variable X^i over time. If these values contain important information for the long-term forecast as suggested by the coefficients, it makes sense that the linear regression does not perform equally well in this task. It simply has no access to this type of information using only linear functions of the data points. Finally, that the term (6,6), i.e. the evolution of time over time, is irrelevant appears to be intuitive again. Furthermore, it can be noted that term (2,6) is the only remaining (i,6)-term that seems to not play an important role, while (4,4) is one of the few non-(i,6) second order terms that seems to do so.

A last word on the unpredictable performance of the signature regression for short-term forecasts. As we can see in Figure 5.23 the truncation orders for short-term forecasts are comparably small. This suggests that the second order terms of the signature do not play an important role on the short run. But if we restrict ourselves to truncation order one, according to (5.6) the signature will only contain the end points of the explanatory variables. However, looking at the expected linear regression coefficients of the explanatory variables for the window-16-horizon-1 forecast in Figure 5.30 we can clearly see that, although the absolute values of the end points play an important role, the previous values are not negligible. This explains, to some extend, the erratic behaviour of the signature regression compared to the linear regression for short-term forecasts. For truncation level one, the signature contains only the end points, while the linear regression has access to all values of the time series and hence strictly more information. Therefore, the variability of the performances might be a symptom of overfitting to the end points of the time series. In particular, the forecasts with comparably high performance (around 0.6) can be explained by favourable train-test-splits, where the end points of the time series tend to be meaningful. Likewise, the

bad performances with values under -1 can be explained by adverse train-test splits, where the earlier values of the time series were not negligible.

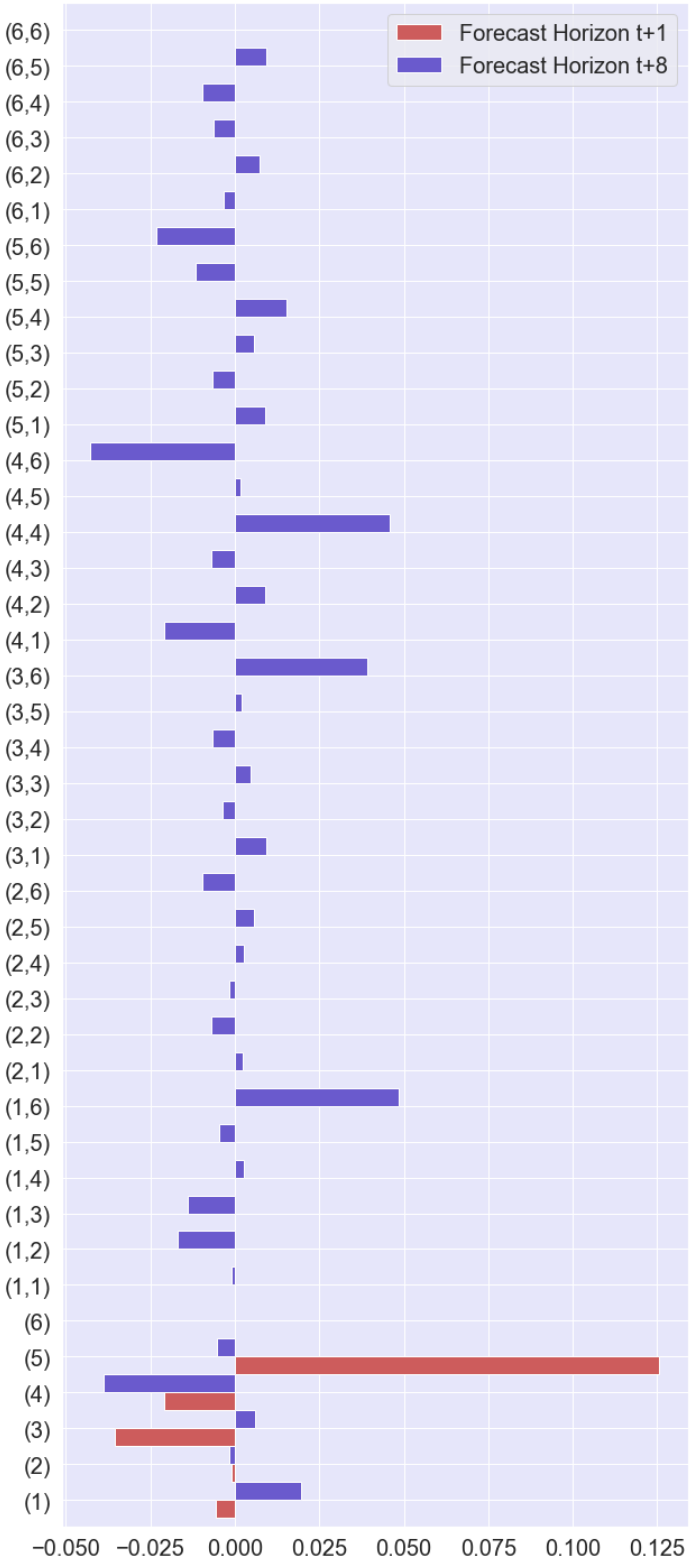

Fig. 5.29 Average signature regression coefficients for t+1 and t+8 forecast (window size = 16)

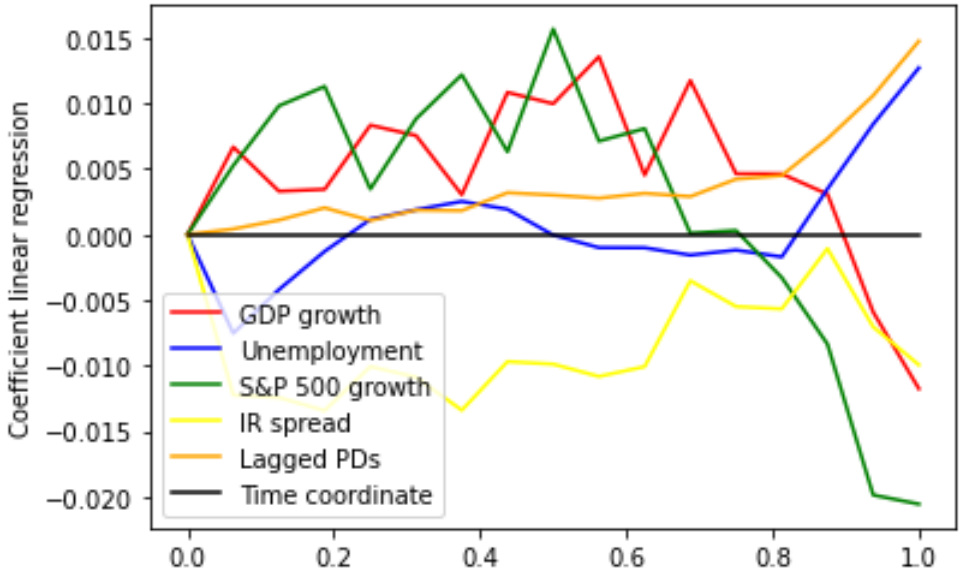

Fig. 5.30 Evolution of linear regression coefficients of the explanatory variables along time in window scaled to $[0, 1]$

Chapter 6

Conclusion

This thesis provided an overview over the path signature and some of its most important properties. We introduced the signature linear model utilizing the truncated signature for linear regression tasks as proposed by A. Fermanian [6]. In order to justify this procedure we rolled up the proof of two propositions assuring beneficial performance bounds of this model. We filled some gaps in the original proof and found evidence that some constants should be changed, but could confirm the general results otherwise. Based on the work of A. Fermanian [6] we implemented a framework to compare linear regression and signature regression on a variety of synthetically generated data sets and real-world data in the context of "Credit Cycle Forecasting".

On one hand, we found evidence that the signature regression needs more observations for training to filter out meaningful information compared to linear regression on the generated data. On the other hand, signature regression seemed to be more capable of dealing with fine paths and noisy data.

In the context of "Credit-Cycle-Forecasting" with quarterly real-world data we experienced that only small truncation levels of one or two were selected by our algorithm. Subsequently the signature regression did not perform well for short-term forecasts, where second order signature terms do not seem to play an important role, while linear regression obtained a more stable result. For long-term forecasts, however, our experiments suggest that in particular the second order terms containing information about the evolution of the path over time are significant. This explains why the linear regression, which has no access to this type of information, was outperformed by the signature regression in this regard. Furthermore, it underlines the importance of augmenting paths with time dimension and base point prior to calculating signatures, which makes sense in the context of time-reparametrization- and translation-invariance of the signature.

Appendix A

Foundations

In this appendix we outsource some mathematical results and proofs, that are used in this work, but whose proof or detailed explanation would distract from the actual focus of the thesis at the point of usage.

The following lemma borrowed from R. v. Handel [9, Lem. 5.13] provides an upper bound for the covering number of the n-dimensional Euclidean unit ball $B^n = \{x \in \mathbb{R}^n | \|x\| \leq 1\}$ with respect to the Euclidean distance.

Lemma A.1. *It holds* $N(B^n, \|\cdot\|, \epsilon) = 1$, *for* $\epsilon \geq 1$ *and*

$$\left(\frac{1}{\epsilon}\right)^n \leq N(B^n, \|\cdot\|, \epsilon) \leq \left(\frac{3}{\epsilon}\right)^n \quad \textit{for } 0 < \epsilon < 1.$$

Proof. The proof is based on a duality between covering number and packing number and uses a volume argument. An ϵ-packing of (T, d) is a set $N \subseteq T$, where $d(t, t') > \epsilon$ for every $t, t' \in T$, $t \neq t'$. The largest possible cardinality of such a packing is then called the packing number

$$D(T, d, \epsilon) := \sup\{|N| : N \text{ is an } \epsilon\text{-packing of } (T, d)\}.$$

The duality of packing- and covering number reads, for every $\epsilon > 0$

$$D(T, d, 2\epsilon) \leq N(T, d, \epsilon) \leq D(T, d, \epsilon).$$

A proof for this duality can be found in R. v. Handel [9, Lem. 5.12].

The volume argument can be summarised as follows: First let's consider the unit ball B and an ϵ-packing N of said ball. Every member $t \in N$ of this packing is inside a radius of 1 from the origin, therefore every point within the radius ϵ of any t must be inside the $(1+\epsilon)$-ball $B_{1+\epsilon}$. Consequently the ϵ-packing N of B is completely contained by $B_{1+\epsilon}$. It is clear that a ball of Lebesgue-measure $\lambda(B_{1+\epsilon})$ can at most contain $\lambda(B_{1+\epsilon})/\lambda(B_\epsilon)$ disjoint ϵ-balls B_ϵ, which bounds the packing number from above.

Meanwhile the ϵ-net of B must cover B completely an therefore contains at least $\lambda(B)/\lambda(B_\epsilon)$ possibly overlapping ϵ-balls, which bounds the covering number from below.

This is rolled out as a precise proof in [9, Lem. 5.13]. □

Lemma A.2 (Hoeffding's Inequality)**.** *Let $X_1, \dots, X_N$ be independent random variables. Assume that $X_i - \mathbb{E}[X_i] \in [a_i, b_i]$ for every i. Then, for the empirical expectation $\hat{S}_N = \frac{1}{N}\sum_{i=1}^N X_i$ and any $\epsilon > 0$, we have*

$$\mathbb{P}\left(\hat{S}_N - \mathbb{E}[\hat{S}_N] \geq \epsilon\right) \leq \exp\left(-\frac{2\epsilon^2 N^2}{\sum_{i=1}^N (b_i - a_i)^2}\right).$$

Proof. The following proof was presented by Prof. Dirk-André Deckert in his 2020 Mathematical Statistics lecture at Ludwigs-Maximilians-Universität in Munich.

Let $\epsilon > 0$. First we use that $Z \mapsto \exp(tZ)$ is increasing for random variables Z and use Markov's inequality to find that

$$\begin{aligned}
\mathbb{P}\left(\hat{S}_N - \mathbb{E}[\hat{S}_N] \geq \epsilon\right) &= \mathbb{P}\left(\exp(t(\hat{S}_N - \mathbb{E}[\hat{S}_N])) \geq e^{t\epsilon}\right) \\
&\leq e^{-t\epsilon}\mathbb{E}[\exp(t(\hat{S}_N - \mathbb{E}[\hat{S}_N]))] \\
&= e^{-t\epsilon}\mathbb{E}\left[\prod_{i=1}^N \exp(\frac{t}{N}(X_i - \mathbb{E}[X_i]))\right] \\
&= e^{-t\epsilon}\prod_{i=1}^N \mathbb{E}\left[\exp(\frac{t}{N}(X_i - \mathbb{E}[X_i]))\right],
\end{aligned}$$

where we used independence of the X_i in the last step.
Next up, we bound the factors $\mathbb{E}[\exp(tZ)]$ for random variables Z having $\mathbb{E}[Z] = 0$ and $\text{range}(Z) \subseteq [a, b]$, where we of course intend to define $Z = X_i - \mathbb{E}[X_i]$ later on:

The exp is convex, therefore $\forall \lambda \in [0, 1]$ we have

$$\exp(\lambda z_1 + (1-\lambda)z_2) \leq \lambda e^{z_1} + (1-\lambda)e^{z_2}.$$

Setting $z_1 = a$, $z_2 = b$, $\lambda = \frac{b-Z}{b-a}$ gives

$$\begin{aligned}\exp(tZ) &= \exp\left(t\left(\frac{b-Z}{b-a}a + \frac{Z-a}{b-a}b\right)\right)\\ &\leq \frac{b-Z}{b-a}e^{ta} + \frac{Z-a}{b-a}e^{tb}.\end{aligned}$$

Furthermore, $\mathbb{E}[Z] = 0$ so that

$$\mathbb{E}\left[e^{tZ}\right] \leq \frac{b}{b-a}e^{ta} + \frac{-a}{b-a}e^{tb} =: e^{\phi(t)}.$$

Let us bound the right-hand side

$$\begin{aligned}\phi(t) &= \log\left(\underbrace{\frac{b}{b-a}}_{=1-\alpha}e^{ta} + \underbrace{\frac{-a}{b-a}}_{=\alpha}e^{tb}\right)\\ &= ta + \log\left((1-\alpha) + \alpha e^{t(b-a)}\right),\end{aligned}$$

for which the first two derivatives are given by

$$\begin{aligned}\phi'(t) &= a + \frac{\alpha(b-a)e^{t(b-a)}}{(1-\alpha) + \alpha e^{t(b-a)}}\\ &= \alpha + \frac{\alpha(b-a)}{(1-\alpha)e^{-t(b-a)} + \alpha}\end{aligned}$$

$$\begin{aligned}\phi''(t) &= -\frac{\alpha(b-a)(1-\alpha)(-1)(b-a)}{\left((1-\alpha)e^{-t(b-a)} + \alpha\right)^2}e^{-t(b-a)}\\ &= -\frac{\alpha(1-\alpha)e^{-t(b-a)}}{\left((1-\alpha)e^{-t(b-a)} + \alpha\right)^2}(b-a)^2\\ &= u(1-u)(b-a)^2 \quad \text{for } u = \frac{\alpha}{(1-\alpha)e^{-t(b-a)} + \alpha} \in [0,1].\end{aligned}$$

By Taylor's theorem $\exists\theta \in [0,t]$ such that

$$\begin{aligned}\phi(t) &= \underbrace{\phi(0)}_{=0} + \underbrace{t\phi'(0)}_{=0} + \frac{t^2}{2}\phi''(\theta)\\ \text{and} \quad \phi''(\theta) &\leq \sup_{u\in[0,1]} u(1-u)(b-a)^2 \leq \frac{(b-a)^2}{4},\end{aligned}$$

which implies

$$\phi(t) \leq \frac{(b-a)^2}{8}t^2.$$

Because of our earlier inequality

$$\mathbb{E}\left[e^{tZ}\right] \leq e^{\phi(t)} \leq \exp\left(\frac{(b-a)^2}{8}t^2\right)$$

and we may resume to our original term to bound:

$$\begin{aligned}
\mathbb{P}\left(\hat{S}_N - \mathbb{E}[\hat{S}_N] \geq \epsilon\right) &\leq e^{-t\epsilon N} \prod_{i=1}^{N} \mathbb{E}\left[\exp(t(X_i - \mathbb{E}[X_i]))\right] \\
&\leq e^{-t\epsilon N} \prod_{i=1}^{N} \exp\left(\frac{(b_i-a_i)^2}{8}t^2\right) \\
&= \exp\left(t^2\frac{(b_i-a_i)^2}{8} - t\epsilon N\right).
\end{aligned}$$

The argument of the exp. can be optimized over t:
$f(t) = t^2x - ty$ attains a minimum at $f'(t) = 2tx - y = 0 \implies t = \frac{y}{2x}$ because $f''(t) = 2x > 0$ for our case, and

$$\begin{aligned}
f(\frac{y}{2x}) &= \frac{y^2}{4x} - \frac{y^2}{2x} = -\frac{y^2}{4x} \\
&= -\frac{2\epsilon^2N^2}{\sum_{i=1}^{N}(b_i-a_i)^2}
\end{aligned}$$

so that we can conclude

$$\mathbb{P}\left(\hat{S}_N - \mathbb{E}[\hat{S}_N] \geq \epsilon\right) \leq \exp\left(-\frac{2\epsilon^2N^2}{\sum_{i=1}^{N}(b_i-a_i)^2}\right).$$

□

With the next lemma we obtain an upper bound for the expectation of $\sup_{t\in T} X_t$ if X is a subgaussian process on the metric space (T, d). The proof is straightforward, because the lemma is actually a corollary of Dudley's Theorem [9, Thm. 5.24].

Lemma A.3 (Entropy integral)**.** *Let $(X_t)_{t\in T}$ be a separable subgaussian process on the metric space (T,d). Then we have the following estimate:*

$$\mathbb{E}\left[\sup_{t\in T} X_t\right] \leq 12\int_0^\infty \sqrt{\log(N(T,d,\epsilon)}d\epsilon.$$

Proof. According to Dudley's Theorem [9, Thm. 5.24] we can use that

$$\mathbb{E}\left[\sup_{t\in T} X_t\right] \leq 6\sum_{k\in\mathbb{Z}} 2^{-k}\sqrt{\log(N(T,d,2^{-k}))}.$$

We further estimate

$$\begin{aligned}\sum_{k\in\mathbb{Z}} 2^{-k}\sqrt{\log(N(T,d,2^{-k}))} &= 2\sum_{k\in\mathbb{Z}}\int_{2^{-k-1}}^{2^{-k}} \sqrt{\log(N(T,d,2^{-k}))}d\epsilon \\ &\leq 2\sum_{k\in\mathbb{Z}}\int_{2^{-k-1}}^{2^{-k}} \sqrt{\log(N(T,d,\epsilon))}d\epsilon \\ &= 2\int_0^\infty \sqrt{\log(N(T,d,\epsilon))}d\epsilon,\end{aligned}$$

which is possible since the covering number $N(T,d,\epsilon)$ is decreasing in ϵ: The bigger ϵ is chosen, the smaller is the amount of ϵ-Balls required to cover the metric space T. □

Appendix B

Implementation

The source code of the python implementation to obtain all results presented in the thesis can be found in the following public GitHub repository:
https://github.com/NiklasMWeber/CreditCycleForecasting

Bibliography

[1] Birgé, L. and Massart, P. (2007). Minimal penalties for Gaussian model selection. *Probability theory and related fields* 138.1, pp. 33–73.

[2] Chen, K.-T. (1958). Integration of paths — a faithfuhl representation of paths by noncommutative formal power series. *Transactions of the American Mathematical Society* 89, pp. 395–407.

[3] Chevyrev, I. and Kormilitzin, A. (2016). A primer on the signature method in machine learning. arXiv: 1603.03788.

[4] Durrett, R. (2019). *Probability: Theory and examples*. Vol. 49. Cambridge university press.

[5] Fermanian, A. (2020). *Embedding and learning with signatures*. arXiv: 1911.13211 [stat.ML].

[6] Fermanian, A. (2021). Linear functional regression with truncated signatures. arXiv: 2006.08442.

[7] Galton, F. (1886). Regression Towards Mediocrity in Hereditary Stature. *The Journal of the Anthropological Institute of Great Britain and Ireland* 15, pp. 246–263.

[8] Hambly, B. and Lyons, T. (Mar. 2010). Uniqueness for the signature of a path of bounded variation and the reduced path group. *Annals of Mathematics* 171.1, pp. 109–167. DOI: 10.4007/annals.2010.171.109.

[9] Handel, R. van (n.d.). Probability in high dimension, December 2016. APC.

[10] Johnson, R. and Wichern, D. (2002). Prentice hall Englewood Cliffs. *Applied multivariate statistical analysis. 5th ed. New Jersy: Prentice hall Englewood Cliffs*.

[11] Király, F. J. and Oberhauser, H. (2019). Kernels for sequentially ordered data. *Journal of Machine Learning Research* 20.31, pp. 1–45.

[12] Lyons, T. J., Caruana, M., and Lévy, T. (2007). *Differential equations driven by rough paths*. Springer.

[13] Ree, R. (1958). Lie Elements and an Algebra Associated With Shuffles. *Annals of Mathematics* 68, p. 210.

[14] Reizenstein, J. and Graham, B. (2018). *The iisignature library: efficient calculation of iterated-integral signatures and log signatures*. arXiv: 1802.08252 [cs.DS].

[15] The Credit Research Initiative - NUS Asian Institute of Digital Finance (2022). *Aggregate PD & AS*. URL: https://nuscri.org/en/data/cdsaggregatedata/e501s0/0/ (visited on 02/02/2022).