École polytechnique fédérale de Lausanne

Semester Project Spring 2024

Bachelor in Mathematics

Rates of convergence for projection estimators with a focus on B-splines

Author:
Kilian WAN

Supervisor:
Victor PANARETOS
Almond STÖCKER

Abstract

This Bachelor project investigates B-splines, a type of piecewise polynomial function widely used in mathematical modelling and numerical approximation. The study focuses on understanding the mathematical properties of B-splines, such as their local support, continuity, and partition of unity, which contribute to their efficiency and flexibility in various applications. Additionally, the project explores a general theorem by Huang (1998) on the rate of convergence of functions to their approximations within orthogonal projections, and discusses the conditions and factors that influence this convergence. We provide detailed proofs of this result, which are not presented in Huang's paper, using various notions, such as functional analysis. Practical applications are demonstrated through implementing B-splines in R, showing their effectiveness in data smoothing and functional approximation. The results highlight the versatility and robustness of B-splines in solving complex mathematical and statistical problems.

Acknowledgements

First of all, I would like to thank my project supervisor Almond Stöcker for all the discussions we had on the topic, all the proofs we had to check, which often took a lot of time, but also for his feedback on the report and his guidance.
I would also like to thank Professor Victor Panaretos for giving me the opportunity to do my bachelor thesis in his department.
Finally, I would like to thank my family for their constant support, encouragement and understanding throughout this journey.

Contents

1 Introduction

This section provides an introduction and definitions of the B-splines. In the first subsection we will see the intuition behind the B-splines, more concrete definitions and properties that this basis satisfies. A significant portion of our study relies on Kagerer (2013)'s article, which provides an in-depth introduction to B-splines and their use in regression analysis. Kagerer discusses the construction of B-splines, their mathematical properties and applications in statistical modeling. Her work serves as a reference for understanding B-splines. In addition to our primary focus on B-splines, we also mention the work of Eilers and Marx (1996) in penalized splines, which are widely used due to their ability to handle overfitting and provide smoother estimates. Penalty splines add a penalty to term to the fitting process, balancing the fit and smoothness of the spline. However, in this report, we will only concentrate on unpenalized splines, emphasizing the theoretical aspects and applications of B-splines without additional penalties. In the second subsection, we will concentrate on the application of this basis, such as regression and combination of B-splines.

1.1 B-Splines

In this preliminary subsection, we will discuss the definition and intuition behind B-splines. In essence, the functions from the B-spline basis are piecewise polynomial functions of order k, connected in a special way. They are connected at the knots and have a small support. First, let's see what knots are. We say that $\boldsymbol{\kappa} := \{\kappa_i : 1 \leqslant i \leqslant m\}, m \in \mathbb{N}$ is a *knot sequence* if it is a non-decreasing sequence of real numbers, i.e. $\boldsymbol{\kappa} = \{\kappa_1 \leqslant \kappa_2 \leqslant \cdots \leqslant \kappa_m\}$, where the elements κ_i are called the *knots*. A characteristic feature of knots is that they can have a *multiplicity*, denoted as m_j, that means that $\kappa_j = \cdots = \kappa_{j+m_j-1}$, indicating that the knot repeats over some interval.

Once the knots are given, it is easy to compute the B-splines recursively, for any degree of the polynomial. To do so, we can use de Boor's relation. Before using this relation, let's get some intuition of these B-splines of small degree using Figure 1 from Eilers and Marx (1996).

We observe one B-spline of degree 1 on the left of Figure 1(a). It is composed of two linear pieces : the first one from x_1 to x_2, and the second one from x_2 to x_3. Moreover, to the left of x_1 and to the right of x_3 this B-spline is zero. In other terms, its support is $[x_1, x_3]$ and it connects the knot x_1 to the knot x_3. On the right of Figure 1(a), we have three more B-splines of degree 1.

Let's increase the degree of the B-spline. In the left part of Figure 1(b), a B-spline of degree 2 is shown. It is composed of three quadratics pieces, joined at two knots. At the joining points not only the ordinates of the polynomial pieces match, but also their first derivatives are equal — however not their second derivatives. The B-spline is based on four adjacent knots : $x_1, \ldots, x_4$. Furthermore, in the right of Figure 1(b), we have three some B-splines of degree 2.

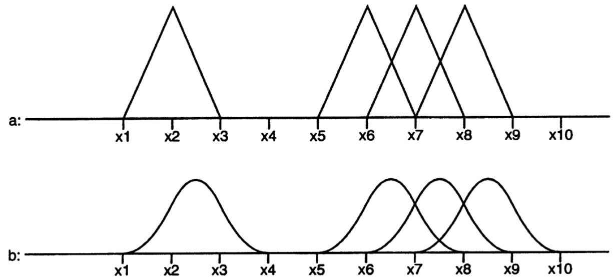

FIG. 1. Illustrations of one isolated B-spline and several overlapping ones (a) degree 1; (b) degree 2.

REMARK. *The B-splines overlap each other : first-degree B-splines overlap with two neighbors, second-degree B-splines with four neighbors and so on.*

Now we have all the definitions to derive the *B-spline basis of order k*. First, let $\boldsymbol{\kappa} := (\kappa_{-(k-1)}, \dots, \kappa_{m+k})$ be a knot sequence where each knot has at most multiplicity k (i.e. $\kappa_j \neq \kappa_{j+k}$). The two boundary knots κ_0 and κ_{m+1} define the interval of interest and the m knots $\kappa_1, \dots, \kappa_m$ are the *inner knots*. Moreover, the remaining $2(k-1)$ *exterior knots* — before κ_0 : $\kappa_{-(k-1)}, \dots, \kappa_{-1}$ and after κ_{m+1} : $\kappa_{m+2}, \dots, \kappa_{m+k}$ — are required to ensure a good behaviour on $[\kappa_0, \kappa_{m+1}]$. Now we define the B-spline using the recurrence relation from de Boor (1978, p.90) and the notation of Kagerer (2013):

DEFINITION (B-spline). The **B**-*spline basis functions of order* $k > 1$, denoted as $\mathbf{B}_j^{\boldsymbol{\kappa},k}$, are defined as follow:

$$\mathbf{B}_j^{\boldsymbol{\kappa},k}(x) = \frac{x-\kappa_j}{\kappa_{j+k-1}-\kappa_j}\mathbf{B}_j^{\boldsymbol{\kappa},k-1}(x) - \frac{x-\kappa_{j+k}}{\kappa_{j+k}-\kappa_{j+1}}\mathbf{B}_{j+1}^{\boldsymbol{\kappa},k-1}(x), \tag{1.1}$$

where

$$\mathbf{B}_j^{\boldsymbol{\kappa},1}(x) = (\kappa_{j+1}-x)_+^0 - (\kappa_j - x)_+^0 = \begin{cases} 1 & \text{for } \kappa_j \leqslant x < \kappa_{j+1} \\ 0 & \text{otherwise} \end{cases} \tag{1.2}$$

is the B-spline of order 1, and $j = -(k-1), \dots, m$.

Having defined B-splines, we now turn our attention to their key properties, which underpin their utility and versatility in computational and mathematical contexts. In the upcoming discussion, we will explore essential properties such as local support, continuity, and combinations of B-splines:

PROPOSITION 1 (Properties of B-splines).

(i) *They form a partition of unity on the interval $[\kappa_0, \kappa_{m+1}]$:*

$$\sum_{j=-(k-1)}^{m} \mathbf{B}_j^{\boldsymbol{\kappa},k}(x) = 1.$$

(ii) *The support of the function $\mathbf{B}_j^{\boldsymbol{\kappa},k}$ is the interval (κ_j, κ_{j+k}). Hence, for $|d| \geqslant k$, we have :*

$$\mathbf{B}_j^{\boldsymbol{\kappa},k}(x) \cdot \mathbf{B}_{j+d}^{\boldsymbol{\kappa},k}(x) = 0.$$

(iii) *B-splines are up to $(k - m_j - 1)$–times continuously differentiable at the knot κ_j. Moreover, the $(k - m_j)$–th derivative has a jump at κ_j, where m_j is the multiplicity of κ_j.*

(iv) *Linear combinations of the basis functions are also $(k-m_j-1)$–times continuously differentiable at the knot κ_j. They are of the form*

$$\mathbf{B}_{\boldsymbol{\alpha}}^{\boldsymbol{\kappa},k}(x) = \sum_{j=-(k-1)}^{m} \alpha_j \cdot \mathbf{B}_j^{\boldsymbol{\kappa},k}(x)$$

where $\boldsymbol{\alpha} = (\alpha_{-(k-1)}, \ldots, \alpha_m)^\top$.

Remark. *The linear combination of the B-spline basis functions of order k can generate all polynomial functions of degree smaller than k on $[\kappa_0, \kappa_{m+1}]$. See Figure 2.*

Now that we have established the fundamental properties of B-splines, we will take a better look at the third property of Proposition 1. We have seen that B-splines are continuously differentiable, so we will give the formula of Kagerer (2013) for the first order derivative:

Theorem 2 (First derivative of the B-spline functions). *The first order derivative of the B-spline functions is the following :*

$$\frac{\partial \mathbf{B}_j^{\boldsymbol{\kappa},k}(x)}{\partial x} = \frac{k-1}{\kappa_{j+k-1} - \kappa_j} \mathbf{B}_j^{\boldsymbol{\kappa},k-1}(x) - \frac{k-1}{\kappa_{j+k} - \kappa_{j+1}} \mathbf{B}_{j+1}^{\boldsymbol{\kappa},k-1}(x) \tag{1.3}$$

for $k > 1$. Using the expression (1.2), we see that for $k = 1$, the first order derivative is equal to 0. Moreover, as the derivative of a B-spline of order k is a linear combination of B-splines of order $k-1$, it is actually a B-spline of order $k-1$. In addition, from Equation (1.3), one can show that the first derivative of a spline as a linear combination of the B-spline functions is given by

$$\begin{aligned} \frac{\partial \mathbf{B}_{\boldsymbol{\alpha}}^{\boldsymbol{\kappa},j}(x)}{\partial x} &= \frac{\partial}{\partial x} \sum_{j=-(k-1)}^{m} \alpha_j \cdot \mathbf{B}_j^{\boldsymbol{\kappa},k}(x) \\ &= (k-1) \sum_{j=-(k-1)}^{m} \frac{\alpha_j - \alpha_{j-1}}{\kappa_{j+k-1} - \kappa_j} \mathbf{B}_j^{\boldsymbol{\kappa},k-1}(x), \end{aligned} \tag{1.4}$$

where $\alpha_{-(k-1)-1} := 0 =: \alpha_{m+1}$ by de Boor (1978, p.116).

For equidistant knot sequences, we can simplify Equations (1.3) and (1.4) as follows:

$$\frac{\partial \mathbf{B}_j^{\boldsymbol{\kappa},k}(x)}{\partial x} = \frac{1}{h} \mathbf{B}_j^{\boldsymbol{\kappa},k-1}(x) - \frac{1}{h} \mathbf{B}_{j+1}^{\boldsymbol{\kappa},k-1}(x)$$

and

$$\frac{\partial \mathbf{B}_{\boldsymbol{\alpha}}^{\boldsymbol{\kappa},j}(x)}{\partial x} = \frac{1}{h} \sum_{j=-(k-1)}^{m} (\alpha_j - \alpha_{j-1}) \mathbf{B}_j^{\boldsymbol{\kappa},k-1}(x)$$

respectively on $[\kappa_0, \kappa_{m+1}]$. This concludes this introduction to B-splines. We will now study some practical examples.

1.2 Some examples and plots of B-splines

First, we see some examples of the behaviour of the B-splines when the order is not the same. We also check some properties mentionned above. For example, the first line of Figure 2, we observe that, as seen in Proposition 1 (i), they form a partition of the unity on the interval $[\kappa_0, \kappa_{m+1}]$, that is the sum of the basis functions is one. We also look at some combinations of B-splines. The previous remark is also studied, showing the versatility of the B-splines.

Figure 2 shows examples of the B-splines with different orders k and with m equidistant inner knots.

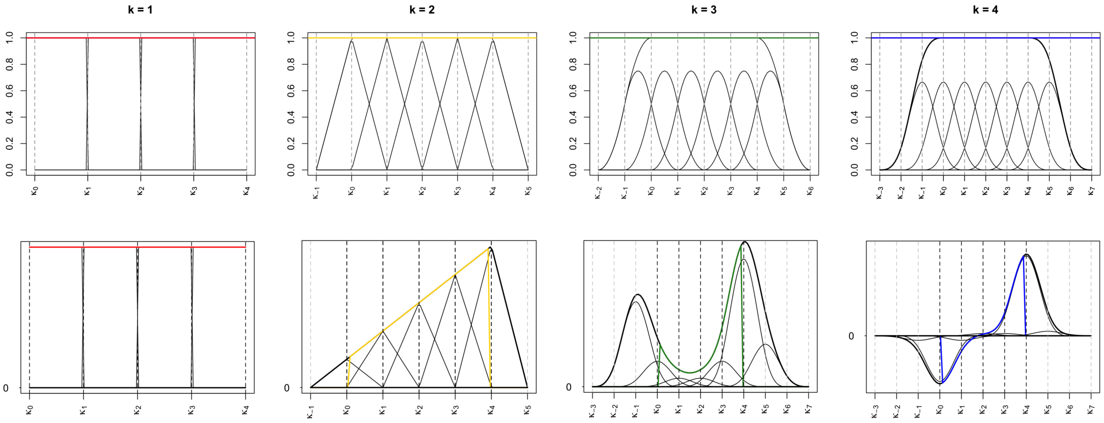

FIG. 2. B-spline basis and linear combinations for different orders k and $m = 3$. **First row**: B-spline basis functions of order $1, 2, 3$ and 4 with equidistant knots, and their sum on $[\kappa_0, \kappa_{m+1}]$. **Second row**: B-spline basis functions of order $1, 2, 3$ and 4 with equidistant knots, weighted such that their sum is a polynomial of degree $k - 1$ on $[\kappa_0, \kappa_{m+1}]$.

EXAMPLE. This first example uses the *motorcycle data*. The motorcycle data set with $n = 133$ observations has two variables describing a simulated motorcycle accident for a test of crash helmets. The two variables are `times` and `accel`. We plot the estimated regression curve of the following model in Figure 3:

$$\texttt{accel} = \sum_{j=-(k-1)}^{m} \alpha_j \mathbf{B}_j^{\boldsymbol{\kappa},k}(\texttt{times}) + \epsilon$$

where k is the degree, $m = 8$ is the number of inner knots in some knot sequence $\boldsymbol{\kappa}$ and ϵ the error term. We observe for $k = 1$, the regression curve is formed by horizontal straight lines, where we have jumps at some knots. For $k = 2$, the regression curve is also formed by straight lines, and not horizontal as in the first case. The quadratic and cubic cases ($k = 3$ and 4 respectively) are well observed. However, for k larger, we can't distinguish which curve represents each order.

Now we look at the problem differently. We fix some order, say $k = 4$, and increase the number of inner knots. Recall that for $k = 4$, we have a cubic curve.

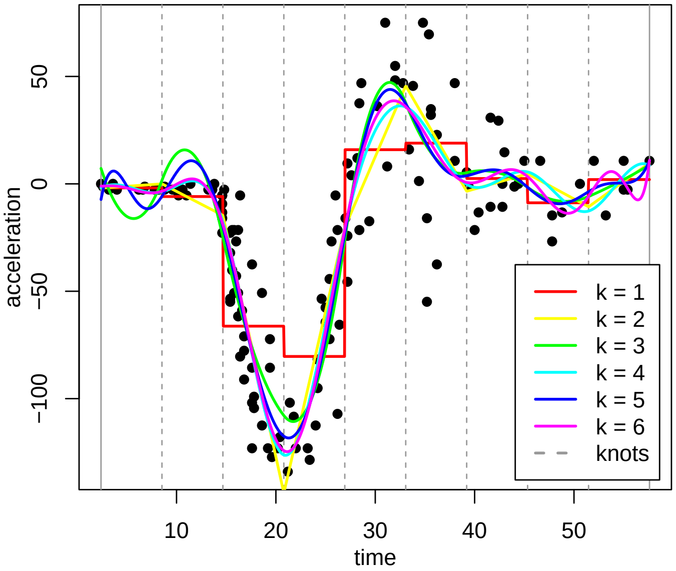

FIG. 3. B-spline regression functions with 8 equidistant inner knots and orders from 1 to 6.

For 0 knots, we observe a cubic function. When we increase the number of knots to 1, we observe two connected cubic polynomials. When we increase the number of knots, the number of cubic polynomials also increases. An observation is that, for $k = 4$ in Figure 3 and $m = 8$ in Figure 4, the same curve is showed.

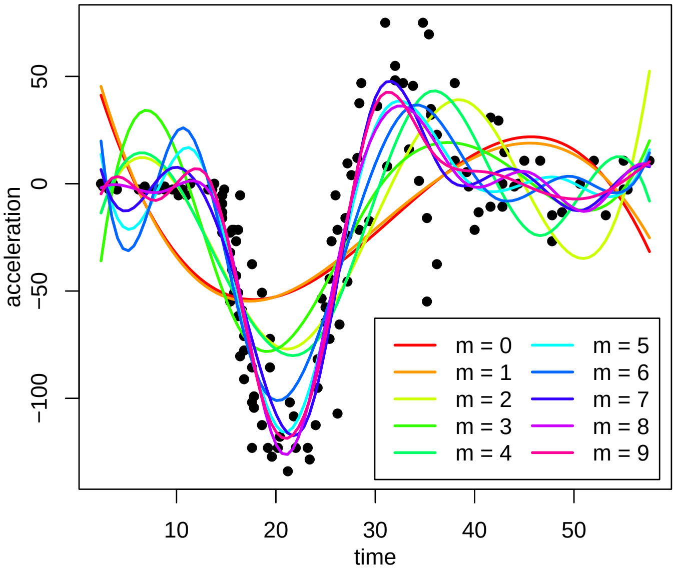

FIG. 4. B-spline regression functions with fixed order $k = 4$ and variant equidistant inner knots.

2 A theorem on rate of convergence

In this section, we provide a general result on the behaviour of the convergence of projection estimators. This result is Theorem 1 of Huang (1998). First, we lay down the fundamental of functional analysis, preparing the ground for a deeper exploration of the subject. In the second subsection, we give the context of the regression problem, where the notions of empirical and theoretical orthogonal projections come to the forefront. Moreover, we introduce important conditions that give us more information about the spaces we are using. In addition, we decompose orthogonally our problem, simplifying the analysis, and we finally give our main result. We also give some conditions to be able to apply this result with B-splines, but the main interest is the general theorem. Huang's paper provides a detailed analysis of the convergence rates of projection estimators in the context of functional data analysis.

2.1 Preliminaries on functional analysis

In this subsection of this second part, we will introduce key definitions and concepts of functional analysis to know with what we will be working for Huang's theorem. Some of these concepts will be distances, Hilbert spaces,...

Take any set M, we then can define the *distance* function d from $M \times M$ to the real numbers, which satisfies the following three properties. For x, y and z in M, we have *positivity* of the distance, that is, $\mathrm{d}(x,y) \geqslant 0$ and we have equality if and only if x and y are the same points. Moreover we have the *symmetry* of the distance, which means that $\mathrm{d}(x,y) = \mathrm{d}(y,x)$. Finally, we have the *triangle inequality*: $\mathrm{d}(x,y) + \mathrm{d}(y,z) \geqslant \mathrm{d}(x,z)$. Hence, if we have this distance d defined on M, the pair (M, d) is called a *metric space*.

Furthemore, in a metric space, say (M, d), if we take a sequence of points $x_1, x_2, \ldots$ in M, this sequence is called a *Cauchy sequence*, if when we take any radius $r > 0$, we can always find some range $N(r) \in \mathbb{N} - \{0\}$, depending on r, such that for any $m, n \geqslant N(r)$, we are sure that these two points x_m and x_n are at most at a distance r. This can be written as $\mathrm{d}(x_m, x_n) < r$ for any $m, n \geqslant N(r)$. If a metric space (M, d) has the property that any Cauchy sequence in M has a limit in M, the space is called *complete*.

But how can we actually find a distance? An easy case is when we have a vector space, as the distance can be induced by the inner product of this vector space. Take any inner product $\langle \cdot, \cdot \rangle$ on some vector space $\mathbb{V}$, then this gives us the norm $\|x\| := \langle x, x \rangle^{1/2}$ for any $x \in \mathbb{V}$. Thus, by setting $\mathrm{d}(x,y) := \|x - y\|$, and using properties of the scalar product and Cauchy-Schwarz, the three properties of the distance are verified. In addition, there is a special case of metric spaces: *Hilbert spaces*– a vector space with an inner product, where this last induces a distance for which the metric space is complete. Moreover, a finite dimensional subspace G of H, where H is a Hilbert space, is closed. This comes from the fact that G is closed if every convergent sequence in G has a limit in G. Recall that $\dim G = n < \infty$. Therefore G is isomorphic to $\mathbb{R}^n$ with some norm. By completeness of $\mathbb{R}^n$, every convergent sequence converges to a point in $\mathbb{R}^n$. Under isomorphism, every convergent sequence in G converges to a point in G. Moreover, in a Hilbert space, every bounded sequence has a convergent subsequence by Bolzano-Weierstrass. So if $\{v_n\} \subset G$ converges to some $v \in H$, then by continuity of the linear operations (addition and scalar multiplication), v must be in G. Hence G is closed.

In this paper we will be interesed on working with Hilbert spaces. An example of a Hilbert space that we will be using is the *Euclidean space*, that is a Hilbert space, where the vector space is defined on a subset of $\mathbb{R}^n$ for $n \geqslant 1$.

Before giving the context of the problem, we will give some definitions and notation that will be useful. We have for any function f on $\boldsymbol{\mathcal{X}}$, $\|f\|_\infty = \sup_{x\in\boldsymbol{\mathcal{X}}} |f(x)|$. And we write $a_n \prec b_n$ if a_n/b_n is bounded, where a_n and b_n are sequences of positive numbers for $n \geqslant 1$. And $a_n \asymp b_n$ if $a_n \prec b_n$ and $b_n \prec a_n$. Finally, we define the notion of $\mathcal{O}_P(\cdot)$ and $\mathbf{o}_P(\cdot)$. Let c_n be a sequence of positive numbers for $n \geqslant 1$. A set of random variables W_n for $n \geqslant 1$ is a $\mathcal{O}_P(c_n)$ if, for any $\epsilon > 0$, there exists a finite $N_\epsilon > 0$ and finite $\delta_\epsilon > 0$ such that

$$\mathbf{P}\left[\frac{|W_n|}{c_n} > \delta_\epsilon\right] < \epsilon, \quad \forall n > N_\epsilon,$$

and W_n is a $\mathbf{o}_P(c_n)$ if for every positive ϵ and δ, there exists $N_{\epsilon,\delta} > 0$ such that

$$\mathbf{P}\left[\frac{|W_n|}{c_n} > \delta\right] < \epsilon,$$

also known as convergence in probability. Recall that W_n is a $\mathcal{O}(c_n)$ if there exists a constant $K > 0$ such that $|W_n| \leqslant K \cdot c_n$ for all sufficiently large n. We show that $\mathcal{O}(\cdot)$ implies $\mathcal{O}_P(\cdot)$. By assumption, we have the existence of K such that $\forall n \geqslant N_0$, $|W_n|/c_n \leqslant K$. Therefore,

$$\mathbf{P}\left[\frac{|W_n|}{c_n} > K\right] = 0.$$

Now for any $\epsilon > 0$, choose $\delta_\epsilon = K$, and $N_\epsilon = N_0 > 0$,

$$\mathbf{P}\left[\frac{|W_n|}{c_n} > \delta_\epsilon\right] = 0 < \epsilon, \quad \forall n > N_\epsilon.$$

We conclude that W_n is $\mathcal{O}_P(c_n)$.

Now we give the preliminaries of functional analysis needed to understand the main result and proofs of this paper. First, a *linear operator* L on some normed vector space $\mathbb{V}$ over $\mathbb{K}$, where $\mathbb{K}$ is a field, is a map from $\mathbb{V}$ to itself that preserves the linear structure of $\mathbb{V}$, that is for any $\mathbf{v}, \mathbf{w} \in \mathbb{V}$ and $\lambda \in \mathbb{K}$, $L(\lambda\mathbf{v} + \mathbf{w}) = \lambda L(\mathbf{v}) + L(\mathbf{w})$. Then we define the *norm* of a linear operator as

$$\|L\| := \sup_{\mathbf{v}\neq 0} \frac{\|L\mathbf{v}\|}{\|\mathbf{v}\|} = \sup_{\|\mathbf{v}\|=1} \|L\mathbf{v}\|.$$

Now take two normed vectors spaces $\mathbb{V}_1, \mathbb{V}_2$ with norms $\|\cdot\|_i, i = 1, 2$. Then we say that the linear operator L from $\mathbb{V}_1$ to $\mathbb{V}_2$ is *bounded* if there exists a finite constant $C > 0$ such that

$$\|L\mathbf{v}\|_2 \leqslant C\|\mathbf{v}\|_1, \quad \forall \mathbf{v} \in \mathbb{V}_1.$$

We then define $\mathbf{B}(\mathbb{V}_1, \mathbb{V}_2)$, the *set of all bounded linear operators from* $\mathbb{V}_1$ *to* $\mathbb{V}_2$. Using the norm of the linear operator defined above, this becomes a normed space. Moreover, the elements of $\mathbf{B}(\mathbb{V}_1, \mathbb{V}_2)$ are called *bounded linear operators*, and if we have $\mathbb{V}_1 = \mathbb{V}_2 = \mathbb{V}$, we use $\mathbf{B}(\mathbb{V})$ for the set of all bounded operators on $\mathbb{V}$. In the case where $\mathbb{V}_1 = \mathbb{H}_1, \mathbb{V}_2 = \mathbb{H}_2$ are two Hilbert spaces with $\langle \cdot, \cdot \rangle_i, i = 1, 2$, and for any $L \in \mathbf{B}(\mathbb{H}_1, \mathbb{H}_2)$, the unique operator L^* of $\mathbf{B}(\mathbb{H}_2, \mathbb{H}_1)$ such that

$$\langle L\mathbf{v}_1, \mathbf{v}_2 \rangle_2 = \langle \mathbf{v}_1, L^*\mathbf{v}_2 \rangle_1, \quad \forall \mathbf{v}_1 \in \mathbb{H}_1, \mathbf{v}_2 \in \mathbb{H}_2,$$

is said to be the *ajdoint* of L. When $\mathbb{H}_1 = \mathbb{H}_2 = \mathbb{H}$, L is called *self-adjoint* when $L^* = L$. The existence of L^* comes from the Riesz representation theorem. See Theorem 3.3.1 of Hsing and Eubank (2015).

Now let $L \in \mathbf{B}(\mathbb{H})$ and suppose there are some $\lambda \in \mathbb{K}$ and a non-zero vector $\mathbf{e} \in \mathbb{H}$ such that

$$L\mathbf{e} = \lambda \mathbf{e},$$

then, we say that λ is an *eigenvalue* and $\mathbf{e}$ is a corresponding *eigenvector* of L.

Let's look at a special case of linear operator, the *linear functional*. A linear functional is a linear operator which maps the vector space $\mathbb{V}$ to the underlying field $\mathbb{K}$, that satisfies the linear property. We have the following useful proposition when the linear functional is a scalar product with some vector:

PROPOSITION 3. *Let $\langle \cdot, \cdot \rangle$ be any inner product and $Z \in \mathbb{V}$. Then, for the linear functional $A := \langle Z, \cdot \rangle$ we have*

$$\|A\| = \|Z\|,$$

where the norm on the left is the linear operator norm, and the norm on the right is the norm induced by the inner product.

PROOF. We prove it by double inequality. First, let $\mathbf{v} \neq 0$, then by Cauchy-Schwarz:

$$|\langle Z, \mathbf{v} \rangle| \leqslant \|Z\| \cdot \|\mathbf{v}\|.$$

as $\mathbf{v} \neq 0$, dividing both sides by $\|\mathbf{v}\|$:

$$\frac{|\langle Z, \mathbf{v} \rangle|}{\|\mathbf{v}\|} \leqslant \|Z\|.$$

Taking supremum over $\mathbf{v} \neq 0$, we obtain the first inequality. Now for the other direction, assume $Z \neq 0$, and let $\mathbf{v} = Z$, then we have

$$\|A\mathbf{v}\| = \|\langle Z, Z \rangle\| = \|Z\|^2,$$

therefore

$$\frac{\|A\mathbf{v}\|}{\|Z\|} = \frac{\|Z\|^2}{\|Z\|} = \|Z\|.$$

Thus for a particular choice of $\mathbf{v}$, we have $\|A\| \geqslant \|Z\|$ as $\|A\|$ is the supremum over all possible $\mathbf{v} \neq 0$. Combining both inequalities, we conclude that $\|A\| = \|Z\|$. ■

Finally, as we are working with projection, we will introduce some projection linear operator properties. We say that $\mathbf{P}$ is an *orthogonal projection operator onto a closed subspace* $\mathbb{M}$ of a Hilbert space $\mathbb{H}$ if it is a self-adjoint operator in $\mathbf{B}(\mathbb{H})$ that satisfies $\mathbf{P} = \mathbf{P}^2$ (idempotency). In this paper we will refer to orthogonal projections as projections. Moreover, if $\mathbb{M}$ is a closed subspace of a finite Hilbert space $\mathbb{H}$, then we have the following orthogonal decomposition: $\mathbb{H} = \mathbb{M} \oplus \mathbb{M}^{\perp}$, where $\mathbb{M}^{\perp}$ is the orthogonal complement of $\mathbb{M}$. Another property of projection operators is that $\|\mathbf{P}\| = 1$ as

$$\|\mathbf{Pv}\| = \|\mathbf{P}^2\mathbf{v}\| \leqslant \|\mathbf{P}\|\,\|\mathbf{Pv}\| \quad \text{and} \quad \|\mathbf{Pv}\| \leqslant \|\mathbf{v}\|.$$

Also we have that $\operatorname{im}(\mathbf{P}) = \ker(\mathrm{id}_{\mathbb{H}} - \mathbf{P})$ when the Hilbert space $\mathbb{H}$ is finite. This gives us that if $\mathbf{P}$ is the projection onto some space $\mathbb{M}$, then $\mathrm{id}_{\mathbb{H}} - \mathbf{P}$ is the orthogonal projection onto $\mathbb{M}^{\perp}$. Finally, we see that the only possible eigenvalues of $\mathbf{P}$ are 0 and 1. This is due to the following: take $\mathbf{e}$ a non-zero λ-eigenvector of $\mathbf{P}$, then by idempotency of $\mathbf{P}$:

$$\lambda\mathbf{e} = \mathbf{Pe} = \mathbf{P}^2\mathbf{e} = \mathbf{P}(\lambda\mathbf{e}) = \lambda^2\mathbf{e}.$$

Thus $\lambda(\lambda - 1) = 0$, so $\lambda \in \{0, 1\}$.

Now we have all definitions to start our regression problem, which will be the content of this next subsection.

2.2 Context

We consider the following regression problem. Suppose we have a predictor variable X and a real-valued response variable Y that we want to predict, where X might have an influence on Y. We only know that X and Y have a joint distribution. We assume that X ranges over $\boldsymbol{\mathcal{X}}$, which is a compact subset of some Euclidean space. Moreover we assume that X has a absolutely continuous distribution with respect to the Lebesgue measure, and its density $f_X(\cdot)$ is bounded in the following way: $\exists M_1, M_2 > 0 : 0 < M_1 \leqslant f_X(\cdot) \leqslant M_2 < \infty$, that is $f_X(\cdot)$ is bounded away from zero and infinity. We want to study the influence of X on Y. To do so, we first set $\mu(x) := \mathbf{E}[Y \mid X = x]$ and $\sigma^2(x) = \mathbf{var}[Y \mid X = x]$, where we suppose that these two functions are bounded on $\boldsymbol{\mathcal{X}}$. Also take a random sample of size n from the distribution of (X, Y), $(X_1, Y_1), \ldots, (X_n, Y_n)$. Moreover suppose that the residuals $\{\epsilon_i : 1 \leqslant i \leqslant n\}$ where $\epsilon_i := Y_i - \mu(X_i)$ are independent with each other, but also independent of $X_1, \ldots, X_n$. The main focus lies in estimating μ.

But first, we have to give some definitions to understand with what we are working. In this paragraph, we will introduce the necessary material to understand. Recall that $\mu(\cdot)$ and $\sigma^2(\cdot)$ are defined on $\boldsymbol{\mathcal{X}}$. Consider any function f that is integrable with respect to the probability measure of the random variable X, defined over its support $\boldsymbol{\mathcal{X}}$, and set $\mathbf{E}_n[f] := (1/n)\sum_{i=1}^{n} f(X_i)$. We also set $\mathbf{E}[f] := \mathbf{E}[f(X)]$. Using the following functions, we define the *empirical inner product* and *empirical norm* as follows:

$$\langle f_1, f_2 \rangle_n := \mathbf{E}_n[f_1 \cdot f_2], \quad \|f_1\|_n^2 := \langle f_1, f_1 \rangle_n \tag{2.1}$$

for any $f_1, f_2 \in \mathcal{L}^2(\boldsymbol{\mathcal{X}}, \mathbb{R})$ (square-integrable functions on $\boldsymbol{\mathcal{X}}$ with respect to the Lebesgue measure). But there are also the *theoretical versions*:

$$\langle f_1, f_2 \rangle = \mathbf{E}[f_1 \cdot f_2], \quad \|f_1\|^2 := \langle f_1, f_1 \rangle, \tag{2.2}$$

again for any $f_1, f_2 \in \mathcal{L}^2(\boldsymbol{\mathcal{X}}, \mathbb{R})$. We then have that the theoretical norm is equivalent to the $\mathcal{L}^2$ norm . This last equivalence comes from the fact that at the beginning of the paragraph, we supposed that X has a density that is bounded away from zero and infinity. So there are constants $M_1, M_2 > 0$ such that $f_X(\cdot)$ is bounded above by M_2 and below by M_1. Now take any $f \in \mathcal{L}^2(\boldsymbol{\mathcal{X}}, \mathbb{R})$, we want to show that there are $C_1, C_2 > 0$ such that

$$C_1 \|f\|_{\mathcal{L}^2} \leqslant \|f\| \leqslant C_2 \|f\|_{\mathcal{L}^2}.$$

We have $\|f\|^2 = \mathbf{E}[f \cdot f] = \int_{\boldsymbol{\mathcal{X}}} |f(x)|^2 f_X(x) \mathrm{d}x$ but then we can upper and lower bound this expression as follows:

$$M_1 \int_{\boldsymbol{\mathcal{X}}} |f(x)|^2 \mathrm{d}x \leqslant \|f\|^2 \leqslant M_2 \int_{\boldsymbol{\mathcal{X}}} |f(x)|^2 \mathrm{d}x.$$

Then set $C_1 = M_1^{1/2}$, and $C_2 = M_2^{1/2}$ and obtain the equivalence of the theoretical and the $\mathcal{L}^2$ norm.

Let's work on a more precise space. Take a closed subspace H of the vector space $\mathcal{L}^2(\boldsymbol{\mathcal{X}}, \mathbb{R})$ and suppose that $\mu \in H$. We call this subspace the *model space,* and it is a Hilbert space with the theoretical inner product defined in (2.2). This comes from the fact that the space $\mathcal{L}^2(\boldsymbol{\mathcal{X}}, \mathbb{R})$ is a Hilbert space with the $\mathcal{L}^2$ inner product. So any closed subspace is again a Hilbert space with the same inner product. The closedness of the subspace is essential to keep the completeness with respect to the distance induced by the inner product. Finally, since the theoretical and the $\mathcal{L}^2$ norms are equivalent, and any Cauchy sequence converges with the $\mathcal{L}^2$ norm, we also obtain the convergence with the theoretical norm. And clearly the theoretical inner product is an inner product, so this proves the claim.

Back to the main subject, the idea is to estimate μ, to do so we will use the *least-square estimate* of μ, where we will be minimizing the problem over a linear space $G \subset H$ of bounded functions such that $\mathrm{dim} G < \infty$ and the dimension depends on our sample. This space is called the *approximating space.* As G is a finite-dimensional subspace of the vector space H, then it must be closed. At the beginning of the paragraph, we took a sample of size n, thus G can vary with the sample size n. But the only thing we need is just to be sure that $\mathrm{dim} G = N_n > 0$ for any $n \geqslant 1$. Another property of this linear space G, is that we require it to be *theoretically identifiable*, that is if $g \in G$ is almost surely equal to zero with respect to the measure induced by the distribution of X, then $g = 0$ everywhere. Moreover, the space G is *empirically identifiable* relative to $X_1, \ldots, X_n$ if the only function g such that $g(X_i) = 0$ for any $1 \leqslant i \leqslant n$ is the null function. Finally, given a sample $X_1, \ldots, X_n$, if G is empirically identifiable, then it is a Hilbert space with the empirical inner product. This is as the three properties of the norm are verified. Without the empirical identifiability, we have that $\| \cdot \|_n$ is a semi-norm, that is we have symmetry and triangle inequality, but we don't have that if $\|g\|_n = 0$, then $g = 0$. This is why it is important for G to be empirically identifiable.

Define $\hat{\mu}$ to be the least square estimator of μ, where the minimization is on G. Then $\hat{\mu} = \operatorname{argmin}_{g \in G} \sum_{i=1}^{n} [g(X_i) - Y_i]^2$. Since X has a density with respect to the Lebesgue measure, then the probability distribution of X is absolutely continuous relative to the Lebesgue measure, which in turn suggests that there are no points of positive probability. Practically, this means that the probability of X taking on any single specific value (like $X_i = X_j$ for $i \neq j$) is zero. Hence, any two design points X_i and X_j are almost surely distinct because the probability that they would be exactly the same is zero. So we are sure that the points $X_1, \ldots, X_n$ are different, and we can find a function, say $\widetilde{Y} = Y(\cdot)$, defined on $\boldsymbol{\mathcal{X}}$, that interpolates $Y_1, \ldots, Y_n$ at these points. Then the interpretation of $\hat{\mu}$ is that it is the orthogonal projection of Y onto the linear space G. We could think that by choosing G in a certain way, when we increase the dimension of G, $\hat{\mu}$ converges to μ. However, μ has not to be necessary in H, so we could expect that $\hat{\mu}$ would converge to some μ^*, that is the orthogonal projection of μ onto H. Hence, in this paper we will discuss what are the condition to obtain the convergence of $\hat{\mu}$ to μ^*, where, as discussed just before, μ^* can or not be equal to μ. We will study the convergence with the following squared norms:

$$\|\hat{\mu} - \mu^*\|^2 \quad \text{or} \quad \|\hat{\mu} - \mu^*\|_n^2.$$

2.3 Decomposition of the problem

We denote Q the *empirical orthogonal projection* onto G, P the *theoretical orthogonal projection* onto G and finally P^* the *theoretical orthogonal projection* onto H. By using the same notation of the previous subsection, we see that $\hat{\mu} = Q\widetilde{Y}$ and $\mu^* = P^*\mu$. We define $\overline{\mu}$ to be the *best approximator* in G to μ relative to the theoretical norm. Then using our orthogonal projections, we have the following property:

$$\overline{\mu} = P\mu = P\mu^* \tag{2.3}$$

This property will help us to decompose orthogonally our minimization problem. We see that we can rewrite the problem as follows

$$\begin{aligned} \hat{\mu} - \mu^* &= (\hat{\mu} - \overline{\mu}) + (\overline{\mu} - \mu^*) \qquad (2.4) \\ &= (Q\widetilde{Y} - P\mu) + (P\mu - P^*\mu). \end{aligned}$$

Thanks to this relation we obtain the following formula:

Proposition 4.

$$\|\hat{\mu} - \mu^*\|^2 = \|\hat{\mu} - \overline{\mu}\|^2 + \|\overline{\mu} - \mu^*\|^2 \tag{2.5}$$

Proof. We must show that $\hat{\mu} - \overline{\mu}$ and $\overline{\mu} - \mu^*$ are orthogonal:

$$\langle \hat{\mu} - \overline{\mu}, \overline{\mu} - \mu^* \rangle = \langle \underbrace{Q\widetilde{Y} - P\mu}_{\in G}, \underbrace{P\mu^* - \mu^*}_{\in G^\perp} \rangle = 0.$$

We conclude using Pythagorean theorem. ■

More precisely, the component $\hat{\mu} - \overline{\mu}$ is called the *estimation error* and $\overline{\mu} - \mu^*$ is the *approximation error*. We now study the estimation error, and we decompose it into two parts that are orthogonal on the average relative to the empirical inner product, conditioned on the design points. To do so, we introduce the *best approximation* in G of μ relative to the empirical norm. Then $\tilde{\mu} = Q\mu$. Observe that by the self-adjointness of Q, $\langle \tilde{\mu}, g \rangle_n = \langle \mu, g \rangle_n$ for any $g \in G$. Now consider the following decomposition:

$$\begin{aligned} \hat{\mu} - \overline{\mu} &= (\hat{\mu} - \tilde{\mu}) + (\tilde{\mu} - \overline{\mu}) \\ &= (QY - Q\mu) + (Q\mu - P\mu) \end{aligned} \tag{2.6}$$

Again, since Q is its self-adjoint, $\langle \hat{\mu}, g \rangle_n = \langle \widetilde{Y}, g \rangle_n$ for every $g \in G$. Taking the conditional expectation given $X_1, \ldots, X_n$, and using the definition of μ, for every $1 \leqslant i \leqslant n$, we have $\mathbf{E}[\widetilde{Y} \mid X_1, \ldots, X_n](X_i) = \mu(X_i)$, and therefore

$$\begin{aligned} \langle \mathbf{E}[\hat{\mu} \mid X_1, \ldots, X_n], g \rangle_n &= \langle \mathbf{E}[Q\widetilde{Y} \mid X_1, \ldots, X_n], g \rangle_n \\ &= \langle Q\mathbf{E}[\widetilde{Y} \mid X_1, \ldots, X_n], g \rangle_n \\ &= \langle \mathbf{E}[\widetilde{Y} \mid X_1, \ldots, X_n], g \rangle_n \\ &= \langle \mu, g \rangle_n \\ &= \langle \tilde{\mu}, g \rangle_n. \end{aligned}$$

Now by definition of conditional expectation, we have $\mathbf{E}[\hat{\mu} \mid X_1, \ldots, X_n] \in G$, and by the empirical identifiability of G,

$$\langle \mathbf{E}[\hat{\mu} \mid X_1, \ldots, X_n] - \tilde{\mu}, g \rangle_n = 0,$$

and so $\mathbf{E}[\hat{\mu} \mid X_1, \ldots, X_n] = \tilde{\mu}$. We then refer to $\hat{\mu} - \tilde{\mu}$ as the *variance component*, and $\tilde{\mu} - \overline{\mu}$ as the *estimation bias*. But then

$$\mathbf{E}[\langle \hat{\mu} - \tilde{\mu}, \tilde{\mu} - \overline{\mu} \rangle_n \mid X_1, \ldots, X_n] = 0.$$

This last result gives us that the variance component and the estimation bias are orthogonal on the conditional expectation relative to the empirical norm. Therefore, and obtain by the Pythagorean theorem that

$$\begin{aligned} \mathbf{E}[\|\hat{\mu} - \overline{\mu}\|_n^2 \mid X_1, \ldots, X_n] &= \mathbf{E}[\|\hat{\mu} - \tilde{\mu}\|_n^2 \mid X_1, \ldots, X_n] + \mathbf{E}[\|\tilde{\mu} - \overline{\mu}\|_n^2 \mid X_1, \ldots, X_n] \\ &= \mathbf{E}[\|\hat{\mu} - \tilde{\mu}\|_n^2 \mid X_1, \ldots, X_n] + \|\tilde{\mu} - \overline{\mu}\|_n^2 \end{aligned}$$

since $\|\tilde{\mu} - \overline{\mu}\|_n^2$ is deterministic because both $\tilde{\mu}$ and $\overline{\mu}$ are fixed once the sample is fixed.

Combining (2.4) and (2.6) we obtain the decomposition:

$$\hat{\mu} - \mu^* = (\hat{\mu} - \tilde{\mu}) + (\tilde{\mu} - \overline{\mu}) + (\overline{\mu} - \mu^*).$$

Therefore we obtain the following proposition:

Proposition 5.

$$\|\hat{\mu} - \mu^*\| \leqslant \|\hat{\mu} - \tilde{\mu}\| + \|\tilde{\mu} - \overline{\mu}\| + \|\overline{\mu} - \mu^*\| \tag{2.7}$$

and

$$\|\hat{\mu} - \mu^*\|_n \leqslant \|\hat{\mu} - \tilde{\mu}\|_n + \|\tilde{\mu} - \overline{\mu}\|_n + \|\overline{\mu} - \mu^*\|_n \tag{2.8}$$

Proof. Apply the triangle inequality and the two decompositions seen on the discussion on top. ■

We have completed breaking down the problem and are now set to examine each term: the variance component, estimation bias, and approximation error. Prior to studying these components, it is essential to define some conditions on the approximating spaces. These preliminaries will constitute the focus of the next subsection.

2.4 Important conditions

The main result of the paper involves two important conditions A_n and ρ_n that we will establish. The first condition requires that the approximating spaces satisfy some stability constraint. The second one is about the approximation power of the approximating spaces. Recall that at the beginning of section 2.2, we saw that the dimension of G depends on n–the sample size. We define A_n as follows:

$$A_n := \sup_{g \in G} \left\{ \|g\|_\infty / \|g\| \right\}.$$

Since the dimension of G is positive, we have that $A_n \geqslant 1$. This is because G will contain at least one function that is not the null one, therefore A_n is well defined on G. As G is defined to be theoretically identifiable and is a linear subspace of bounded functions (see Section 2.2). Let $\{\phi_j : 1 \leqslant j \leqslant N_n\}$ be an orthonormal basis of G relative to the theoretical inner product. Then write g as follows:

$$g = \sum_{j=1}^{N_n} c_j \phi_j, \quad c_j \in \mathbb{R}.$$

Then by orthonormality, $\|g\|^2 = \|\sum_{j=1}^{N_n} c_j \phi_j\|^2 = \sum_{j=1}^{N_n} c_j^2$, and by the triangle inequality and Cauchy-Schwarz:

$$\begin{aligned} \|g\|_\infty = \|\sum_{j=1}^{N_n} c_j \phi_j\|_\infty &\leqslant \sum_{j=1}^{N_n} |c_j| \|\phi_j\|_\infty \\ &\leqslant \left(\sum_{j=1}^{N_n} c_j^2 \right)^{1/2} \left(\sum_{j=1}^{N_n} \|\phi_j\|_\infty^2 \right)^{1/2}. \end{aligned}$$

Combining both results we obtain $A_n \leqslant \left(\sum_{j=1}^{N_n} \|\phi_j\|_\infty^2 \right)^{1/2} < \infty$.

This number A_n gives us how peaked or oscillatory functions in G can be. For example if A_n is large, there could be functions in G that have very high peaks (when the $\mathcal{L}^\infty$ norm is large) relative to their overall size in the $\mathcal{L}^2$ sense. We talk about the $\mathcal{L}^2$ norm and not the theoretical as they are equivalent. See discussion above. Another example could be when A_n is close to one, as this would suggest that functions in G are more regular in the sense that their $\mathcal{L}^\infty$ norm is not that much larger than their $\mathcal{L}^2$ norm. Imagine the case where G has functions that are very irregular, then it might be able to approximate discontinuous or highly variable functions well. However, it could also mean that small changes in the function we are trying to estimate could lead to large changes on the approximation, which is inderisable in terms of stability. We saw early that A_n is finite, so even if G contains functions that might be complex, this complexity is controlled. Moreover, by finiteness of A_n we have that functions in G are not too peaked as to make the approximation process unstable or to sensitive to perturbation in the input function.

Is this constant A_n well defined for B-splines? First, fix some degree k, and let $\mathbf{Spl}_k(m_n)$ be the space of splines of degree k with m_n knots, using the same notations that in Section 1, and suppose that

$$\frac{\max_{0\leqslant i\leqslant m_n}(\kappa_{i+1}-\kappa_i)}{\min_{0\leqslant i\leqslant m_n}(\kappa_{i+1}-\kappa_i)} \leqslant \gamma \tag{2.9}$$

for some constant γ. Then we have the following result:

THEOREM 6. *If $G = \mathbf{Spl}_k(m_n)$, then*

$$A_n \asymp m_n^{1/2}$$

PROOF. See Theorem $5.1.2$ of DeVore and Lorentz (1993). ■

This was our first important condition to set, which gives us information about our set G. Now we look at our second condition. Set $\rho_n := \inf_{g\in G}\|g-\mu^*\|_\infty$. We see that $\rho_n < \infty$ if and only if μ^* is bounded. This comes from the triangle inequality and the fact that g is a function in G–linear space of bounded functions. Suppose this is the case, then

$$\rho_n \leqslant \inf_{g\in G}\|g\|_\infty + \|\mu^*\|_\infty = \|\mu^*\|_\infty$$

as G contains the null function. By assumption and the discussion in the context section, we know that G is finite-dimensional, therefore closed in H. Recall that μ^* is bounded. So define $r := \|\mu^*\|_\infty < \infty$ and consider the zero-centered closed ball $\mathrm{B}_{2r} = \{g \in G : \|g\|_\infty \leqslant 2r\} \subset G$. Clearly this ball bounded so by Heine-Borel it is compact. We conclude by using the continuity of the following function $g \mapsto \|g-\mu^*\|_\infty$. In other words, since it is a continuous function minimized on some compact set of G, it attains its minimum on this set. Therefore we have the existence of $g^* \in \mathrm{B}_{2r}$ such that $\rho_n = \|g^*-\mu^*\|_\infty$. We see that if there is some other $g' \in G - B_{2r}$ such that $\rho_n = \|g'-\mu^*\|_\infty$, then it's not a global minimizer. This constant can be seen as a measure of approximation quality. Since G depends on

n, the more data we have, the more G can change, and potentially allow a better approximation of μ^*. Thus we would expect ρ_n to decrease when n becomes large, reflecting improved approximation as more sample data are considered. ρ_n characterizes the target function μ^* in the context of approximability by G. A small value means that μ^* is well-represented by G. Conversely, a large value might indicate that μ^* has features that G cannot capture well, suggesting the need of a more diverse function space for approximation.

As in the case for A_n, we would like to know when is ρ_n well defined for B-splines. We need a condition to ensure that ρ_n is well defined for B-splines. To this end, we assume that $\boldsymbol{\mathcal{X}}$ is the Cartesian product of compact intervals $\boldsymbol{\mathcal{X}}_1 \times \ldots \times \boldsymbol{\mathcal{X}}_L$. Now let $0 < \beta \leqslant 1$, we say that a function f on $\boldsymbol{\mathcal{X}}$ satisfies the *β-Hölder condition* if there exists $C > 0$ such that

$$|f(x) - f(y)| \leqslant C|x - y|^{\beta}, \quad \forall x, y \in \boldsymbol{\mathcal{X}}.$$

In our case, we define $|x| = (\sum_{l=1}^{L} x_l^2)^{1/2}$ is the Euclidean norm for $x = (x_1, \ldots, x_L)$ in $\boldsymbol{\mathcal{X}}$. Moreover, given an L-tuple $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_L)$ where $\alpha_l \geqslant 0$ for $1 \leqslant l \leqslant L$, we set $[\boldsymbol{\alpha}] = \alpha_1 + \cdots + \alpha_L$ and we define the differential operator $D^{\boldsymbol{\alpha}}$ as follows:

$$D^{\boldsymbol{\alpha}} = \frac{\partial^{[\boldsymbol{\alpha}]}}{\partial x_1^{\alpha_1} \cdots \partial x_L^{\alpha_L}}.$$

Finally, let $k > 0$ be an integer and set $p = k + \beta$. Then we say that a function on $\boldsymbol{\mathcal{X}}$ is *p-smooth* if f is k-times continuously differentiable on $\boldsymbol{\mathcal{X}}$ and $D^{\boldsymbol{\alpha}}$ satisfies the β-Hölder condition for all $\boldsymbol{\alpha}$ with $[\boldsymbol{\alpha}] = k$.

Recall H is a closed subspace of $\mathcal{L}^2(\boldsymbol{\mathcal{X}}, \mathbb{R})$, and by boundedness of μ, we have $\mu = \mu^*$ (μ^* is the projection of μ onto H). We let G_l be a linear space of functions on $\boldsymbol{\mathcal{X}}_l$ for $1 \leqslant l \leqslant L$ and G the tensor product of these spaces. Lastly, we suppose that μ is p-smooth, then for $\boldsymbol{\mathcal{X}}_l = [0, 1]$ for $1 \leqslant l \leqslant L$, we obtain the following upper bound for ρ_n in the B-splines case admitting the following theorem:

THEOREM 7. *Suppose $k \geqslant p - 1$, for k the degree of the spline. If we have that each $G_l = \mathbf{Spl}_k(m_n)$ and (2.9) holds, then*

$$\rho_n \prec m_n^{-p}.$$

Now we are ready to look at the main result of this paper: Theorem 1 of Huang (1998).

2.5 Theorem

Theorem 8. *Suppose μ^* is bounded and that $\lim_n A_n^2 N_n/n = 0$. Then*

(i) ***(variance component)*** $\|\hat{\mu} - \tilde{\mu}\|^2 = \mathcal{O}_P(N_n/n)$ *and* $\|\hat{\mu} - \tilde{\mu}\|_n^2 = \mathcal{O}_P(N_n/n)$*;*

(ii) ***(estimation bias)*** $\|\tilde{\mu} - \overline{\mu}\|^2 = \mathcal{O}_P(N_n/n + \rho_n^2)$ *and* $\|\tilde{\mu} - \overline{\mu}\|_n^2 = \mathcal{O}_P(N_n/n + \rho_n^2)$*;*

(iii) ***(approximation error)*** $\|\overline{\mu} - \mu^*\|^2 = \mathcal{O}_P(\rho_n^2)$ *and* $\|\overline{\mu} - \mu^*\|_n^2 = \mathcal{O}_P(\rho_n^2)$*.*

Consequently,

$$\|\hat{\mu} - \mu^*\|^2 = \mathcal{O}_P(N_n/n + \rho_n^2) \quad \textit{and} \quad \|\hat{\mu} - \mu^*\|_n^2 = \mathcal{O}_P(N_n/n + \rho_n^2).$$

Remark. *When H is finite-dimensional, we can choose $G = H$, which does not depend on n, the sample size. Then A_n is as well independent of n, and we can set $\rho_n = 0$ for each n. Therefore $\hat{\mu}$ converges to μ^* with a rate $1/n$.*

3 Proof of Theorem 8

This section will be divided in three subsections, where in each subsection we will prove each point of Theorem 8. For the variance component and estimation bias we will prove it using the empirical norm. The approximation error will be proven using the theoretical norm. But first we need an important lemma that gives us the equivalence between both empirical and theoretical norms over G. We will also be giving proofs that are not given on Huang (1998)'s paper.

LEMMA 9. *Suppose that* $\lim_n A_n^2 N_n/n = 0$*, and let* $t > 0$*. Then, except on an event whose probability tends to zero as* $n \to \infty$*,*

$$|\langle f, g\rangle_n - \langle f, g\rangle| \leqslant t\|f\|\,\|g\|, \qquad f, g \in G.$$

PROOF. We do not give a proof, as it is beyond the scope of this thesis. ■

We have the following corollary that gives us the equivalence between the empirical norm and the theoretical norm on G.

COROLLARY 10. *Suppose that* $\lim_n A_n^2 N_n/n = 0$*. Then, except on an event whose probability tends to zero as* $n \to \infty$*,*

(i) *For* $g \in G$*,*

$$\frac{1}{2}\|g\|^2 \leqslant \|g\|_n^2 \leqslant 2\|g\|^2. \tag{2.8}$$

(ii) G *is empirically identifiable.*

(iii) *If* $\|g\|_n^2 = \mathcal{O}_P(c_n)$*, then* $\|g\|^2 = \mathcal{O}_P(c_n)$*. The converse is also true.*

PROOF.

(i) See that we apply Lemma 9 with $f = g$. First, we consider the upper bound for $\|g\|_n^2$:

$$\|g\|_n^2 = \langle g, g\rangle_n \leqslant \langle g, g\rangle + t\|g\|^2 = \|g\|^2 + t\|g\|^2.$$

Take without loss of generality $t = 1$ and obtain the desired upper bound. For the lower bound, we will use the inequality on the other direction:

$$\|g\|_n^2 = \langle g, g\rangle_n \geqslant \langle g, g\rangle - t\|g\|^2 = \|g\|^2 - t\|g\|^2.$$

Now take $t = 1/2$ and obtain the lower bound.

(ii) First suppose (2.8) holds. Let $g \in G$ be such that for all $1 \leqslant i \leqslant n$, $g(X_i) = 0$. Then $\|g\|_n^2 = 0$, so $\|g\|^2 = 0$ as well. Since $\|g\|_\infty \leqslant A_n\|g\|$, this implies that g is identically zero, as we require G to be theoretically identifiable. Therefore if (2.8) holds, then G is empirically identifiable. The desired result follows from (i).

(iii) Let $\|g\|_n^2 = \mathcal{O}_P(c_n)$, therefore for any $\epsilon > 0$, there exists $M > 0$ such that

$$\mathbf{P}\left[\|g\|_n^2 > M \cdot c_n\right] < \epsilon.$$

By (i):

$$\mathbf{P}\left[2\|g\|^2 > M \cdot c_n\right] \leqslant \mathbf{P}\left[\|g\|_n^2 > M \cdot c_n\right] < \epsilon.$$

Dividing by 2 we obtain that $\|g\|^2 = \mathcal{O}_P(c_n)$ since, for any $\epsilon > 0$, there is some $M' := M/2$ such that

$$\mathbf{P}\left[\|g\|^2 > M' \cdot c_n\right] < \epsilon.$$

For the converse, the proof is symmetric by exchanging the roles of $\|g\|_n^2$ and $\|g\|^2$, and instead of dividing by 2, we multiply by 2.

■

Remark. *The first two points of Corollary 10 gives us that G is a Hilbert space with the empirical inner product.*

Now we are ready to prove the three points of the main result:

3.1 Proof of variance component

First, we assume that G is empirically identifiable. This assumption is crucial since without it, our space G is not a Hilbert space – since we don't have a norm, but instead a semi-norm – and therefore the notion of orthonormality is not well defined. Recall $\dim G = N_n$, we let $\{\phi_j : 1 \leqslant j \leqslant N_n\}$ be an orthonormal basis of G relative to the empirical inner product. Since $\hat{\mu} = Q\widetilde{Y}$ and $\tilde{\mu} = Q\mu$ are both in G, we can decompose them on the orthonormal basis and obtain:

$$\begin{aligned}\hat{\mu} - \tilde{\mu} &= \sum_{j=1}^{N_n} \langle \hat{\mu} - \tilde{\mu}, \phi_j \rangle_n \phi_j \\ &= \sum_{j=1}^{N_n} \langle \widetilde{Y} - \mu, Q\phi_j \rangle_n \phi_j \\ &= \sum_{j=1}^{N_n} \langle \widetilde{Y} - \mu, \phi_j \rangle_n \phi_j\end{aligned}$$

where we used the fact that Q is a self-adjoint projection operator onto G, and since $\phi_j \in G$, then $Q\phi_j = \phi_j$ for any $1 \leqslant j \leqslant N_n$. Now by Parseval's identity, and the orthonormal decomposition we obtain:

$$\|\hat{\mu} - \tilde{\mu}\|_n^2 = \sum_{j=1}^{N_n} \langle \widetilde{Y} - \mu, \phi_j \rangle_n^2.$$

Observe that,

$$\mathbf{E}[\widetilde{Y} - \mu \mid X_1, \ldots, X_n](X_i) = \mathbf{E}[\widetilde{Y} \mid X_1, \ldots X_n](X_i) - \mu(X_i) = 0.$$

Therefore, by definition of the empirical inner product, we obtain that

$$\begin{aligned}\mathbf{E}[\langle \widetilde{Y} - \mu, \phi_j \rangle_n \mid X_1, \dots, X_n] &= \frac{1}{n} \sum_{i=1}^{n} \mathbf{E}\left[(\widetilde{Y} - \mu)(X_i) \cdot \phi_j(X_i) \mid X_1, \dots X_n \right] \\ &= \frac{1}{n} \sum_{i=1}^{n} \phi_j(X_i) \mathbf{E}[(\widetilde{Y} - \mu)(X_i) \mid X_1, \dots, X_n] = 0\end{aligned}$$

by the previous observation, where we have used the fact that the residuals are independent of X_i, so also independent of $\phi_j(X_i)$ for all i. Moreover, using again the independence of the residuals, we obtain, for δ_{ij} the Kronecker delta that

$$\mathbf{E}[(Y_i - \mu(X_i))(Y_j - \mu(X_j)) \mid X_1, \dots, X_n] = \delta_{ij} \sigma^2(X_i).$$

Recall that we assumed that $\sigma^2(\cdot)$ is a bounded function, so there exists a finite $M > 0$ such that $\sigma^2(x) \leqslant M$ for all $x \in \boldsymbol{\mathcal{X}}$. Putting all these results together:

$$\begin{aligned}\mathbf{E}\left[\langle Y - \mu, \phi_j \rangle_n^2 \mid X_1, \dots, X_n \right] &= \mathbf{E}\left[\left(\frac{1}{n} \sum_{i=1}^{n} (\widetilde{Y} - \mu)(X_i) \cdot \phi_j(X_i) \right)^2 \middle| X_1, \dots X_n \right] \\ &= \mathbf{E}\left[\frac{1}{n^2} \sum_{i=1}^{n} (Y_i - \mu(X_i))^2 \cdot \phi_j^2(X_i) \middle| X_1, \dots, X_n \right] \\ &= \frac{1}{n^2} \sum_{i=1}^{n} \phi_j^2(X_i) \mathbf{E}\left[(Y_i - \mu(X_i))^2 \mid X_1, \dots, X_n \right] \\ &= \frac{1}{n^2} \sum_{i=1}^{n} \phi_j^2(X_i) \sigma^2(X_i) \leqslant \frac{M}{n} \|\phi_j\|_n^2 = \frac{M}{n},\end{aligned}$$

where we used in the second equality that:

$$\begin{aligned}\left(\frac{1}{n} \sum_{i=1}^{n} (\widetilde{Y} - \mu)(X_i) \cdot \phi_j(X_i) \right)^2 &= \frac{1}{n^2} \sum_{i=1}^{n} \sum_{k=1}^{n} (Y_i - \mu(X_i))(Y_k - \mu(X_k)) \underbrace{\phi_j(X_i)\phi_j(X_k)}_{\delta_{ik}\phi_j^2(X_i)} \\ &= \frac{1}{n^2} \sum_{i=1}^{n} \phi_j^2(X_i)(Y_i - \mu(X_i))^2\end{aligned}$$

and by linearity of $\mathbf{E}$, we obtain the conditional variance. Finally, we obtain the desired result by applying this inequality to $\|\hat{\mu} - \tilde{\mu}\|_n^2$:

$$\begin{aligned}\mathbf{E}[\|\hat{\mu} - \tilde{\mu}\|_n^2 \mid X_1, \dots, X_n] &= \mathbf{E}\left[\sum_{j=1}^{N_n} \langle \widetilde{Y} - \mu, \phi_j \rangle_n^2 \mid X_1, \dots X_n \right] \\ &\leqslant \sum_{j=1}^{N_n} M/n = M(N_n/n).\end{aligned}$$

Let's show that this condition implies that $\|\hat{\mu} - \tilde{\mu}\|_n^2 = \mathcal{O}_P(N_n/n)$. Let $\delta > 0$. Then by Markov's inequality:

$$\mathbf{P}[\|\hat{\mu} - \tilde{\mu}\|_n^2 > \delta] \leqslant \frac{\mathbf{E}[\|\hat{\mu} - \tilde{\mu}\|_n^2 \mid X_1, \dots, X_n]}{\delta} \leqslant \frac{M(N_n/n)}{\delta}$$

As this holds for any $\delta > 0$, define $\delta := M'N_n/n$, where $M' > M$. Then:

$$\mathbf{P}[\|\hat{\mu} - \tilde{\mu}\|_n^2 > M'N_n/n] \leqslant M/M'.$$

Now for any $\epsilon > 0$, we can choose M' large enough so that $M/M' < \epsilon$, satisfying the definition of $\mathcal{O}_P(N_n/n)$ for the squared norm. Again by Corollary 10 we obtain that $\|\hat{\mu} - \tilde{\mu}\|^2 = \mathcal{O}_P(N_n/n)$. ■

3.2 Proof of estimation bias

Before starting the proof for the estimation bias, we will enunciate an important lemma:

Lemma 11. *Let $M > 0$. Let $\{h_n\}_{n\geqslant 1} \subset \boldsymbol{\mathcal{X}}$ such that $\|h_n\|_\infty \leqslant M$ for $n \geqslant 1$, then*

$$\sup_{g\in G} \left| \frac{\langle h_n, g\rangle_n - \langle h_n, g\rangle}{\|g\|} \right| = \mathcal{O}_P\left((N_n/n)^{1/2}\right).$$

First, recall that $\tilde{\mu} - \overline{\mu} = Q\mu - P\mu$. Thus, by Proposition 3 applied to the functional $A = \langle Q\mu - P\mu, \cdot\rangle_n$,

$$\begin{aligned}\|\tilde{\mu} - \overline{\mu}\|_n &= \|Q\mu - P\mu\|_n \\ &= \sup_{g\in G} \left| \frac{\langle Q\mu - P\mu, g\rangle_n}{\|g\|_n} \right| \\ &= \sup_{g\in G} \left| \frac{\langle \mu - P\mu, g\rangle_n - \langle \mu - P\mu, g\rangle}{\|g\|_n} \right| \end{aligned} \tag{3.1}$$

Here, the third equality uses the two following facts:

- Since Q is a self-adjoint operator, and it is the empirical orthogonal projection onto G, we have the following

$$\begin{aligned}\langle Q\mu - P\mu, g\rangle_n &= \langle Q\mu, g\rangle_n - \langle P\mu, g\rangle_n \\ &= \langle \mu, Qg\rangle_n - \langle P\mu, g\rangle_n \\ &= \langle \mu - P\mu, g\rangle_n,\end{aligned}$$

  where we used $Qg = g$ for any $g \in G$.

- $P\mu$ is the orthogonal projection of μ onto G, then $\mu - P\mu = (\mathrm{id}_G - P)\mu$ is the orthogonal projection onto $G^\perp$. So $\langle \mu - P\mu, g\rangle = 0$ for any $g \in G$.

We proved on the previous subsection the existence of some $g^* \in G$ such that $\rho_n = \|g^* - \mu^*\|_\infty$. We write differently the numerator of (3.1), using that for any $g \in G$:

$$\begin{aligned}\langle \mu - P\mu, g\rangle_n - \langle \mu - P\mu, g\rangle = &\left(\langle \mu - g^*, g\rangle_n - \langle \mu - g^*, g\rangle\right) \\ &+ \langle g^* - P\mu, g\rangle_n - \langle g^* - P\mu, g\rangle.\end{aligned}$$

Thus by triangle and $\sup$ inequalities we obtain:

$$\|\tilde{\mu}-\overline{\mu}\|_n \leqslant \underbrace{\sup_{g\in G}\left|\frac{\langle \mu-g^*,g\rangle_n-\langle \mu-g^*,g\rangle}{\|g\|_n}\right|}_{:=\mathrm{I}} + \underbrace{\sup_{g\in G}\left|\frac{\langle g^*-P\mu,g\rangle_n}{\|g\|_n}\right|}_{:=\mathrm{II}}$$
$$+\underbrace{\sup_{g\in G}\left|\frac{\langle g^*-P\mu,g\rangle}{\|g\|_n}\right|}_{:=\mathrm{III}}$$

Let's find some upper bounds for I,II and III. Recall that we supposed that μ was bounded, and it also ensures that ρ_n is finite. Then μ^* is also finite, and we obtain:

$$\begin{aligned}\sup_n \|g^*\|_\infty &\leqslant \|\mu^*\|_\infty + \sup_n \|g^*-\mu^*\|_\infty\\ &= \|\mu^*\|_\infty + \sup_n \rho_n < \infty.\end{aligned}$$

By applying Lemma 11 with $h_n=\mu-g^*$, we have $\|h_n\|_\infty \leqslant \|\mu\|_\infty + \|g^*\|_\infty < \infty$ and obtain that $\mathrm{I}=\mathcal{O}_P\left((N_n/n)^{1/2}\right)$. Now let's concentrate for an upper bound for II. We apply the Cauchy-Schwarz inequality and $P\mu = P\mu^*$ to obtain:

$$\begin{aligned}\mathrm{II} = \sup_{g\in G}\left|\frac{\langle g^*-P\mu,g\rangle_n}{\|g\|_n}\right| &\leqslant \sup_{g\in G}\|g^*-P\mu\|_n\\ &\leqslant 2\|g^*-P\mu^*\|\end{aligned}$$

where the second inequality comes from Corollary 10. Finally, we use the two following facts of linear operators:

$$\|A\| = \max\{|\lambda| : \lambda \text{ is an eigenvalue of } A\}$$

and that

$$\mathrm{Eigen}(\mathbf{P}) = \{0,1\}$$

where $\mathrm{Eigen}(\mathbf{P})$ is the set of eigenvalues of $\mathbf{P}$, where $\mathbf{P}$ is any projection operator. This was discussed on the subsection on the preliminaries of functional analysis. Thus, we obtain another proof using eigenvalues that $\|P\|=1$. We conclude with the following:

$$\begin{aligned}\mathrm{II} \leqslant 2\|g^*-P\mu^*\| &= 2\|Pg^*-P\mu^*\|\\ &\leqslant 2\|P\|\,\|g^*-\mu^*\|_\infty\\ &= 2\rho_n\end{aligned}$$

Hence $\mathrm{II}=\mathcal{O}_P(\rho_n)$. Finally, we find an upper bound for III. Again by Cauchy-Schwarz we obtain:

$$\begin{aligned}\mathrm{III} &= \sup_{g\in G}\left|\frac{\langle g^*-P\mu,g\rangle}{\|g\|_n}\right|\\ &\leqslant \|g^*-P\mu^*\|\sup_{g\in G}\frac{\|g\|}{\|g\|_n}\\ &\leqslant 2\rho_n\end{aligned}$$

where the second inequality comes from the lower bound of Corollary 10. Thus $\mathrm{III} = \mathcal{O}_P(\rho_n)$. Therefore we have

$$\|\tilde{\mu} - \overline{\mu}\|_n^2 = \mathcal{O}_P(N_n/n + \rho_n^2).$$

And by Corollary 10,

$$\|\tilde{\mu} - \overline{\mu}\|^2 = \mathcal{O}_P(N_n/n + \rho_n^2).$$

■

3.3 Proof of approximation error

Let $g^* \in G$ as before, that is such that $\rho_n = \|\mu^* - g^*\|_\infty$. Thus, we have for any $g \in G$:

$$\|\mu^* - g\| \leqslant \rho_n \quad \text{and} \quad \|\mu^* - g\|_n \leqslant \rho_n.$$

By definition of P—the theoretical projection onto G:

$$\|\overline{\mu} - g^*\|^2 = \|P\mu - g^*\|^2 \underset{(2.3)}{=} \|P(\mu^* - g^*)\|^2 \leqslant \|\mu^* - g^*\|^2. \tag{3.2}$$

Finally, by the triangle inequality and the inequalities above, we conclude:

$$\|\overline{\mu} - \mu^*\|^2 \leqslant 2\|\overline{\mu} - g^*\|^2 + 2\|g^* - \mu^*\|^2 \leqslant 4\|\mu^* - g^*\|^2 = \mathcal{O}_P(\rho_n^2).$$

To prove the result for the empirical norm, by Corollary 10 and (3.2), we obtain

$$\|\overline{\mu} - g^*\|_n^2 \leqslant 2\|\overline{\mu} - g^*\|^2 \leqslant 2\|\mu^* - g^*\|^2.$$

And by triangle inequality,

$$\|\overline{\mu} - \mu^*\|_n^2 \leqslant 2\|\overline{\mu} - g^*\|_n^2 + 2\|\mu^* - g^*\|_n^2 = \mathcal{O}_P(\rho_n^2).$$

■

4 Numerical application of Theorem 8

In this section we will apply Theorem 8 with a precise case, using B-splines as approximating space G.

4.1 Setting up the space and notations

First, the predictor variable X is assumed to be uniformly distributed on the interval $[0.01, 1]$. Formally, $X \sim \mathrm{Unif}(0.01, 1)$. Moreover, our response variable Y is given by the function $\mu(X)$ plus some Gaussian noise. Mathematically, we have :

$$Y_i = \mu(X_i) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2),$$

where $\mu(x) := x^2 \sin(x^{-3/2})$ represents the underlying true relationship between the predictor X and the response Y. Then we define the model space H to be the Hilbert space of square-integrable functions over $\boldsymbol{\mathcal{X}}$, where $\boldsymbol{\mathcal{X}}$ is the domain of X. Finally, our subspace G, the approximating space, is spanned by B-spline basis functions of degree k with a number of interior knots m_n, denoted as $\mathbf{Spl}_k(m_n)$. Here, we write our knot sequence $\boldsymbol{\kappa} = \{\kappa_i : 1 \leqslant i \leqslant m\}$ where $m = m_n + 2(k-1)$, where we add the $2(k-1)$ exterior knots to ensure the good behaviour on the boundaries of the B-splines. Using the same notations as in the first section, we let $\{\mathbf{B}_j^{\boldsymbol{\kappa},k}\}_{j=1}^{N_n}$ be the B-spline basis functions of degree k with N_n basis functions defined on the interval $\boldsymbol{\mathcal{X}}$ and $\boldsymbol{\kappa}$ our knot sequence.. Clearly G is a finite-dimensional closed subspace of H, where the dimension of G is equal to the number of B-spline basis functions $N_n := (k + m_n) > 0$.

4.2 Steps to obtain the projections

Recall the given function $\mu(x) = x^2 \sin(x^{-3/2})$ which represents the true relationship between X and Y. We first generate n samples of X uniformly distributed over the interval $[0.01, 1]$. We compute the true values $\mu := \mu(X)$ and we add the Gaussian noise as in the previous subsection. In this case we take $\sigma^2 = 0.1$. Then, we choose the degree k of the B-splines, and define the knot sequence $\boldsymbol{\kappa}$.

- **Least squares estimator** $\hat{\mu}$: we fit the B-spline model to the noisy data:

$$\hat{\mu}(x) = \sum_{j=1}^{N_n} \hat{\alpha}_j \mathbf{B}_j^{\boldsymbol{\kappa},k}(x)$$

where the coefficients $\hat{\alpha}_j$ are obtained by minimizing the following sum of squared residuals:

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\alpha} \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{N_n} \alpha_j \mathbf{B}_j^{\boldsymbol{\kappa},k}(X_i) \right)^2 .$$

This is equivalent to the following problem: finding $g \in G$ that minimizes :

$$\hat{\mu} \ = \arg\min_{g \in G} \sum_{i=1}^{n} (Y_i - g(X_i))^2$$

but since G is the subspace spanned by the B-splines, any $g \in G$ can be written as:

$$g(x) = \sum_{j=1}^{N_n} \alpha_j \mathbf{B}_j^{\boldsymbol{\kappa},k}(x).$$

Therefore both approaches are equivalent. We can also write this problem with a matrix form:

$$\arg\min_{\boldsymbol{\alpha}} \|\mathbf{Y} - \mathbf{B}\boldsymbol{\alpha}\|$$

where $\mathbf{Y}$ is the vector of observed values Y_i with size $n \times 1$, $\mathbf{B}$ the matrix of B-spline basis functions evaluated at each X_i with size $n \times N_n$, and $\boldsymbol{\alpha}$ with size $N_n \times 1$. We clearly obtain by the normal equations the solution

$$\hat{\boldsymbol{\alpha}} = (\mathbf{B}^\top \mathbf{B})^{-1} \mathbf{B}^\top \mathbf{Y}.$$

- **Best approximator μ^* in H**: Recall that μ^* is the projection relative to the theoretical norm of μ onto H, however in this context, μ is already in H. Therefore we take $\mu^* = \mu$.
- **Best approximator $\overline{\mu}$ in G**: The theoretical projection $\overline{\mu}$ is obtained by minimizing the following theoretical norm:

$$\overline{\mu} = \arg\min_{g \in G} \mathbf{E}\left[(\mu(X) - g(X))^2\right]$$

where, again we can write g as a combination of B-splines, and therefore we obtain:

$$\overline{\mu}(x) = \sum_{j=1}^{N_n} \overline{\alpha}_j \mathbf{B}_j^{\boldsymbol{\kappa},k}(x)$$

where the coefficients $\overline{\alpha}_j$ are obtain by minimizing:

$$\overline{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \int_{0.01}^{1} \left(\mu(x) - \sum_{j=1}^{N_n} \alpha_j \mathbf{B}_j^{\boldsymbol{\kappa},k}(x) \right)^2 p_X(x) \mathrm{d}x.$$

- **Empirical projection $\tilde{\mu}$**: We fit the B-spline model to the true values μ:

$$\tilde{\mu}(x) = \sum_{j=1}^{N_n} \tilde{\alpha}_j \mathbf{B}_j^{\boldsymbol{\kappa},k}(x)$$

where, again the coefficients $\tilde{\alpha}_j$ are obtain by minimizing the sum of squared differences between the true values $\mu(X_i)$ and the B-spline model:

$$\tilde{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \sum_{i=1}^{n} \left(\mu(X_i) - \sum_{j=1}^{N_n} \alpha_j \mathbf{B}_j^{\boldsymbol{\kappa},k}(X_i) \right)^2.$$

As in the case for $\hat{\mu}$, by writing $\boldsymbol{\mu}$ to be the vector of observed values $\mu(X_i)$ with size $n \times 1$, we obtain

$$\tilde{\boldsymbol{\alpha}} = (\mathbf{B}^\top \mathbf{B})^{-1} \mathbf{B}^\top \boldsymbol{\mu}.$$

4.3 Conditions for convergence

First, we check that, for our case, we have that

$$A_n \asymp m_n^{1/2}.$$

Numerically, we clearly see that this condition is always checked.

4.4 Numerical convergence

Now we plot the least squares estimator, the best approximator in G and the empirical projection, and see how well are they fitted when we use the B-splines of fixed degree $k = 3$ and with knots $m_n = \lfloor\sqrt{n}\rfloor - 2$. We choose this rule of thumb to ensure the first necessary condition to apply the theorem, that is the convergence to zero of the limit of $A_n^2 N_n/n$. We take three different sample sizes: $n = 30, 500$ and 10.000.

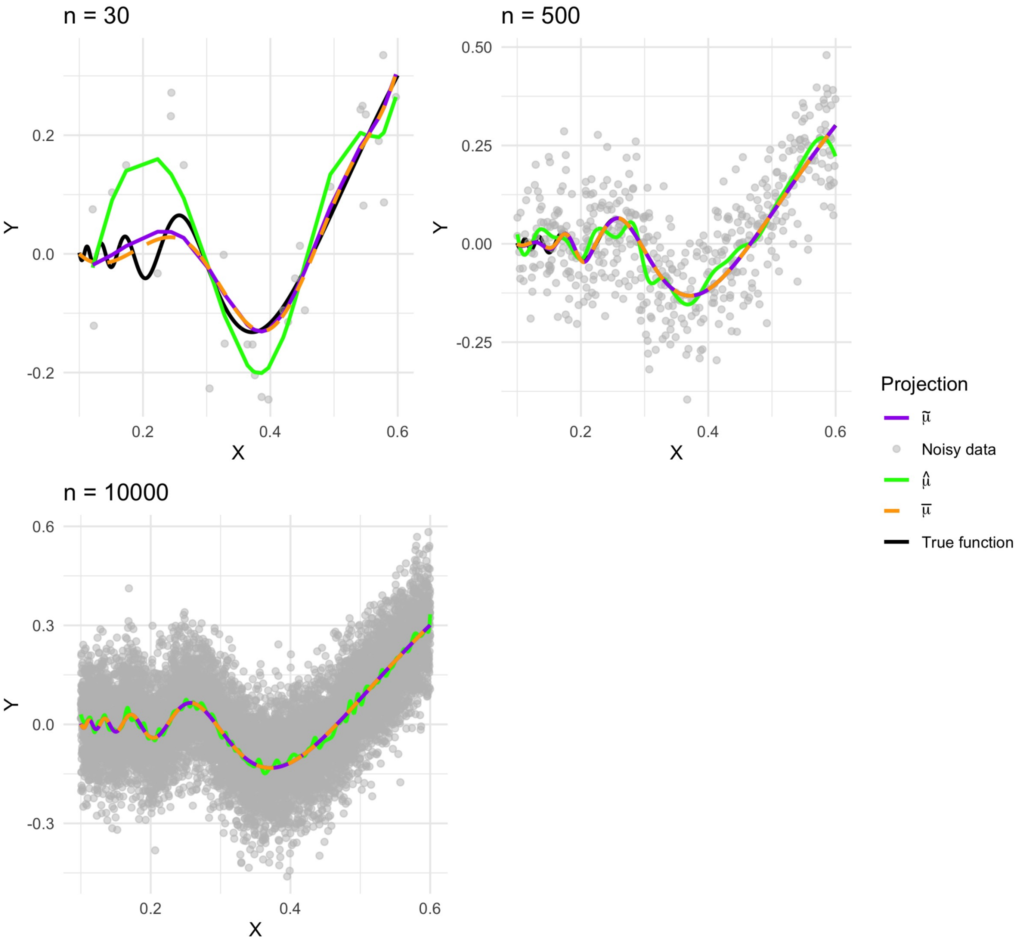

FIG. 1. Numerical approximation of the true function μ.

We observe that the least squares estimator $\hat{\mu}$ is sensitive to big changes ($n = 30$ in Figure 1). This estimator fits the B-spline model directly to the noisy observations Y. It minimizes the sum of the squared residuals between the observed data and the fitted model. As a result, $\hat{\mu}$ is directly influenced by the noise present in the data, making it more sensitive to fluctuations caused by the noise. Unlike $\hat{\mu}$, $\overline{\mu}$ is derived by projecting the true function values directly onto the B-spline space G. This projection minimizes the error in the theoretical norm, ensuring that $\overline{\mu}$ captures well the essential characteristics of μ as accurately as possible within the constraints of the chosen B-spline basis. However, the sensitivity of $\overline{\mu}$ depends on the degree of the B-splines and the number of inner knots, balancing the trade-off between flexibility and smoothness. Finally, the empirical projection $\tilde{\mu}$ is a combination of the empirical fit and the projection, therefore it tends to be less sensitive to noise compared to $\hat{\mu}$, especially when the sample is large as seen in Figure 1. $\tilde{\mu}$ is obtained by fitting the B-spline model to the true values μ rather than the noisy observations Y. This approach provides a more stable and accurate approximation of the true function because it depends on the structure of μ, without any influence of the noise.

Now we look at the behaviour of the three components of Theorem 8 with the empirical norm in Figure 2. Each subplot shows a different component and its corresponding theoretical convergence rate.

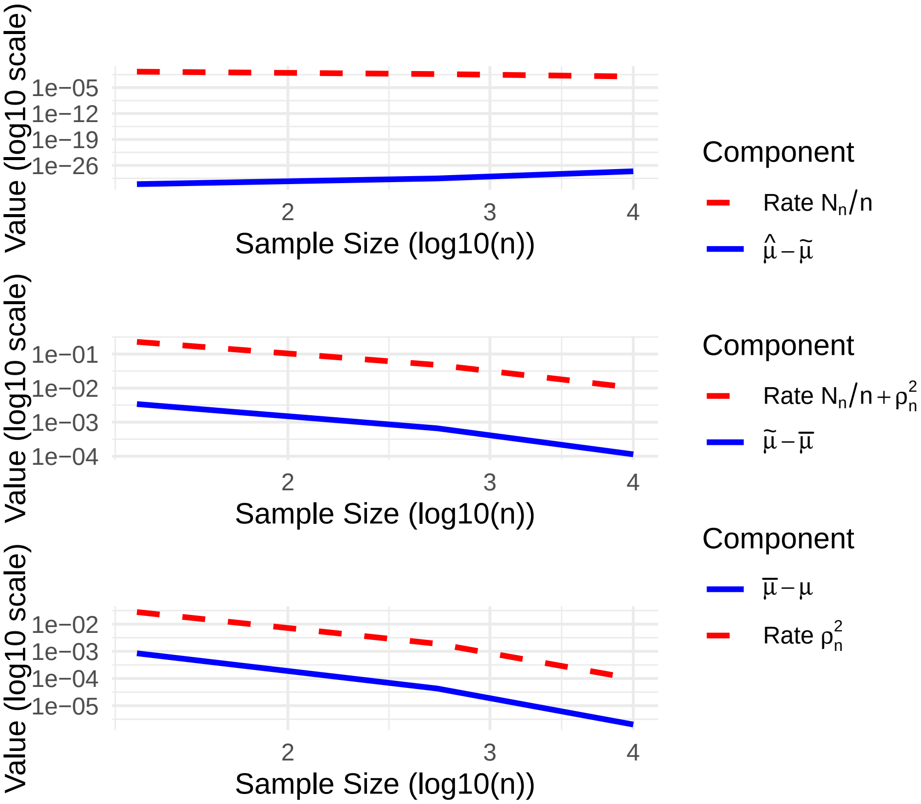

FIG. 2. Comparison of components and theoretical rates.

The top plot compares the variance component against the theoretical rate N_n/n. We observe that the empirical variance follows the theoretical rate closely, indicating that the variance component decreases with increasing sample size. The middle plot examines the estimation bias against its theoretical rate $N_n/n+\rho_n^2$. The empirical bias shows a trend consistent with the theoretical rate, indicating the expected convergence as the sample size grows. Finally, the bottom plot focuses on the approximation error compared to its theoretical rate ρ_n^2. The empirical approximation error aligns well with the theoretical rate, confirming that this component diminishes as the sample size increases. Overall, Figure 2 validates the theoretical convergence rates outlined in Theorem 8. These results demonstrate the robustness and reliability of the B-spline approximations in obtaining the desired convergence properties.

References

[1] Carl de Boor. *A Practical Guide to Spline*. **volume** Volume 27. **january** 1978.

[2] Ronald A DeVore and George G Lorentz. *Constructive approximation*. **volume** 303. Springer Science & Business Media, 1993.

[3] Paul HC Eilers and Brian D Marx. "Flexible smoothing with B-splines and penalties". **in***Statistical science*: 11.2 (1996), **pages** 89–121.

[4] Tailen Hsing and Randall Eubank. *Theoretical foundations of functional data analysis, with an introduction to linear operators*. **volume** 997. John Wiley **and** Sons, 2015.

[5] Jianhua Z Huang. "Projection estimation in multiple regression with application to functional ANOVA models". **in***The annals of statistics*: 26.1 (1998), **pages** 242–272.

[6] Kathrin Kagerer. "A short introduction to splines in least squares regression analysis". **in**(2013).