$\mathbb{L}_2$-BOOSTING ON GENERALIZED HOEFFDING DECOMPOSITION FOR DEPENDENT VARIABLES APPLICATION TO SENSITIVITY ANALYSIS

Magali Champion[1,3], Gaelle Chastaing[1,2], Sébastien Gadat[1],Clémentine Prieur[2]

[1] *Institut de Mathématiques de Toulouse*

[2] *Université Joseph Fourier, LJK/MOISE*

[3] *Institut National de la Recherche Agronomique, MIA*

Abstract:

This paper is dedicated to the study of an estimator of the generalized Hoeffding decomposition. We build such an estimator using an empirical Gram-Schmidt approach and derive a consistency rate in a large dimensional settings. Then, we apply a greedy algorithm with these previous estimators to Sensitivity Analysis. We also establish the consistency of this $\mathbb{L}_2$-boosting up to sparsity assumptions on the signal to analyse. We end the paper with numerical experiments, which demonstrates the low computational cost of our method as well as its efficiency on standard benchmark of Sensitivity Analysis.

Key words and phrases: $\mathbb{L}_2$-boosting, convergence, dependent variables, generalized ANOVA decomposition, sensitivity analysis.

1. Introduction

In many scientific fields, it is desirable to extend a multivariate regression model as a specific sum of increasing dimension functions. Functional ANOVA decomposition or High Dimensional Representation Model (HDMR) given by Hooker (2007); Li, Rabitz, Yelvington, Oluwole, Bacon and Schoendorf (2010) are well known expansions that allow for understanding the model behaviour, and for detecting how inputs interact to each other. For high dimensional models, the HDMR is also a good way to deal with the curse of dimensionality. Indeed, a model function may be well approximated by some first order functional components, making easier the study of a complex model. However, the existence and uniqueness of the functional ANOVA components is of major importance to valid a study. Thus, some identifiability constraints need to be imposed to make the ANOVA decomposition unique.

When input variables are independent, Hoeffding establishes the uniqueness of the decomposition provided that the summands are mutually orthogonal (see *e.g.* Hoeffding (1948)). Further, as pointed by Sobol (1993), the analytical expression of these components can be recursively obtained in terms of conditional expectations. Thus, their estimation can be deduced by numerical approximation of integrals (see *e.g* Sobol (2001); Saltelli, Ratto, Andres, Campolongo, Cariboni, Gatelli, Saisana and Tarantola (2008)).

Nevertheless, the independence assumption is often unrealistic for some real-world phenomena. In this paper, we are interested in the ANOVA expansion of some models that depend on not necessarily independent input variables. Following the work of Stone (1994), later exploited in machine learning by Hooker (2007), and in sensitivity analysis by Chastaing, Gamboa and Prieur (2012), we focus on a generalized Hoeffding decomposition under general assumptions on the inputs distribution. That is, any model function can be uniquely decomposed as a sum of hierarchically orthogonal component functions. Two summands are called *hierarchically orthogonal* whenever all variables included in one of them are also involved in the other. For a better understanding of the paper, this generalized ANOVA expansion will be called a Hierarchically Orthogonal Functional Decomposition (HOFD), as done in Chastaing, Gamboa and Prieur (2012).

Since analytical formulation for HOFD is rarely available, it is of great importance to develop estimation procedures. In this paper, we focus on an alternative method proposed in Chastaing, Gamboa and Prieur (2013) to estimate the HOFD components. It consists of constructing a hierarchically orthogonal basis from a suitable Hilbert orthonormal basis. Inspired by the usual Gram-Schmidt algorithm, the procedure recursively builds for each component a multidimensional basis that satisfies the identifiability constraints imposed to this summand. Then, each component is well approximated on a truncated basis, where the unknown coefficients are deduced by solving an ordinary least-squares. Nevertheless, in a high-dimensional paradigm, this procedure suffers from a curse of dimensionality. Moreover, it is numerically observed that only a few of coefficients are not close to zero, meaning that only a small number of predictors restore the major part of the information contained in the components. Thus, it is important to be able

to select the most relevant representative functions, and next identify the HOFD with a limited computational budget.
In this view, we suggest in this article to transform the ordinary least-squares into a penalized regression as it has been proposed in Chastaing, Gamboa and Prieur (2013). In the present paper, we focus here on the $\mathbb{L}_2$-boosting to deal with the ℓ_0 penalization, developped by Friedman (2001). The $\mathbb{L}_2$-boosting is a greedy strategy that performs variable selection and shrinkage. The choice of such an algorithm is motivated by the fact that the $\mathbb{L}_2$-boosting is very intuitive and easy to implement. It is also closely related (in some practical sense) to the LARS algorithm, proposed by Efron, Hastie, Johnstone and Tibshirani (2004), which solves the Lasso regression with a ℓ_1 penalization (see *e.g.* Bühlmann and van de Geer (2011); Tibshirani (1996)). The $\mathbb{L}_2$-boosting and the LARS both select predictors using the maximal correlation with the current residuals. The question that naturally arises now is the following: provided that the theoretical procedure of components reconstruction is well tailored, do the estimators obtained by the $\mathbb{L}_2$-boosting converge to the theoretical true sparse parameters when the number of observations tends to infinity ?
The goal of this paper is to extend the work of Chastaing, Gamboa and Prieur (2013) by addressing this question. More precisely, the aim is to determine sufficient conditions for which the consistency of the estimators is satisfied. Further, we discuss these conditions and give some numerical examples where such conditiones are fulfilled. One interesting application of the general theory is the global sensitivity analysis (SA). We apply the $\mathbb{L}_2$-boosting to estimate the generalized sensitivity indices defined in Chastaing, Gamboa and Prieur (2012, 2013). After reminding the form of these indices, we numerically compare the $\mathbb{L}_2$-boosting performance with the LARS technique and the Forward-Backward algorithm, proposed by Zhang (2011).

The article is organized as follows. Paragraph 2.1 aims at introducing the notation of the paper.We also remind the HOFD representation of the model function in Paragraph 2.2. In Paragraph 2.3, we recall the procedure detailed in Chastaing, Gamboa and Prieur (2013) that consists in constructing well tailored hierarchically orthogonal basis to represent the components of the HOFD. At last, we highlight the curse of dimensionality we are exposed to, and present the $\mathbb{L}_2$-boosting. Section 3 gathers our main theoretical results on the proposed algorithms. Section 4 presents a numerical study of our method. We finally conclude this work in Section 5, and we provide the proofs of the two main theorems in an Appendix.

Acknowledgment Authors are indebted to Fabrice Gamboa for motivating discussions and numerous suggestions on the subject.

2. Estimation of the generalized Hoeffding decomposition components

2.1 Notation

We consider a measurable function f of a random real vector $\mathbf{X} = (X_1, \cdots, X_p)$ of $\mathbb{R}^p$, $p \geq 1$. The response variable Y is a real-valued random variable defined as

$$Y = f(\mathbf{X}) + \varepsilon, \tag{2.1}$$

where ε stands for a centered random variable independent of $\mathbf{X}$ and models the variability of the response around its theoretical unknown value f. We denote by $P_{\mathbf{X}}$ the distribution law of $\mathbf{X}$, which is unknown in our setting, and we assume that $\mathbf{X}$ admits a density function $p_{\mathbf{X}}$ with respect to the Lebesgue measure on $\mathbb{R}^p$. Note that $P_{\mathbf{X}}$ is not necessarily a tensor product of univariate distributions since the components of X may be correlated.

Further, we suppose that $f \in L^2_{\mathbb{R}}(\mathbb{R}^p, \mathcal{B}(\mathbb{R}^p), P_{\mathbf{X}})$, where $\mathcal{B}(\mathbb{R}^p)$ denotes the Borel set of $\mathbb{R}^p$. The Hilbert space $L^2_{\mathbb{R}}(\mathbb{R}^p, \mathcal{B}(\mathbb{R}^p), P_{\mathbf{X}})$ is denoted by $L^2_{\mathbb{R}}$, for which we use the inner product $\langle \cdot, \cdot \rangle$, and the norm $\|\cdot\|$ as follows,

$$\langle h, g \rangle = \int h(\mathbf{x}) g(\mathbf{x}) p_{\mathbf{X}} d\mathbf{x} = \mathbb{E}(h(\mathbf{X}) g(\mathbf{X}))$$
$$\|h\|^2 = \langle h, h \rangle = \mathbb{E}(h(\mathbf{X})^2), \quad \forall h, g \in L^2_{\mathbb{R}}.$$

Here, $\mathbb{E}(\cdot)$ stands for the expected value. Further, $V(\cdot) = \mathbb{E}[(\cdot - \mathbb{E}(\cdot))^2]$ denotes the variance, and $\mathrm{Cov}(\cdot, *) = \mathbb{E}[(\cdot - \mathbb{E}(\cdot))(* - \mathbb{E}(*))]$ the covariance.

For any $1 \leq i \leq p$, we denote by $P_{\mathbf{X}_i}$ the marginal distribution of X_i and extend naturally the former notation to $L^2_{\mathbb{R}}(\mathbb{R}, \mathcal{B}(\mathbb{R}), P_{\mathbf{X}_i}) := L^2_{\mathbb{R},i}$.

2.2 The generalized Hoeffding decomposition

Let us denote $[1:k] := \{1, 2, \cdots, k\}$, with $k \in \mathbb{N}^*$, and let S be the collection of all subsets of $[1:p]$. We also define $S^* := S \setminus \{\emptyset\}$. For $u \in S$, the subvector $\mathbf{X}_u$ of $\mathbf{X}$ is defined as $\mathbf{X}_u := (X_i)_{i \in u}$. Conventionally, for $u = \emptyset$, $\mathbf{X}_u = 1$. The marginal distribution (*resp.* density) of $\mathbf{X}_u$ is denoted $P_{\mathbf{X}_u}$ (*resp.* $p_{\mathbf{X}_u}$).

A functional ANOVA decomposition consists in expanding f as a sum of increasing dimension functions,

$$\begin{aligned} f(\mathbf{X}) &= f_\emptyset + \textstyle\sum_{i=1}^p f_i(X_i) + \sum_{1 \le i < j \le p} f_{ij}(X_i, X_j) + \cdots + f_{1,\cdots,p}(\mathbf{X}) \\ &= \textstyle\sum_{u \in S} f_u(\mathbf{X}_u), \end{aligned} \tag{2.2}$$

where $f_\emptyset$ is a constant term, f_i, $i \in [1:p]$ are the main effects, $f_{ij}, f_{ijk}, \cdots$, $i, j, k \in [1:p]$ are the interaction effects, and the last component $f_{1,\cdots,p}$ is the residual.

Decomposition (2.2) is generally not unique. However, under mild assumptions on the joint density $p_{\mathbf{X}}$ (see Assumptions (C.1) and (C.2) in Chastaing, Gamboa and Prieur (2012)), the decomposition is unique under some additional orthogonality assumptions.

More precisely, let us introduce $H_\emptyset = H_\emptyset^0$ the set of constant functions, and for all $u \in S^*$, $H_u := L^2_{\mathbb{R}}(\mathbb{R}^u, \mathcal{B}(\mathbb{R}^u), P_{\mathbf{X}_u})$. Then we define H_u^0, $u \in S \setminus \emptyset$ as follows:

$$H_u^0 = \left\{ h_u \in H_u,\ \langle h_u, h_v \rangle = 0, \forall\ v \subset u, \forall\ h_v \in H_v^0 \right\},$$

where $\subset$ denotes the strict inclusion.

Definition 1 (Hierarchical Orthogonal Functional Decomposition - HOFD)**.** *Under Assumption (C.1) and (C.2) in Chastaing, Gamboa and Prieur (2012), the decomposition (2.2) is unique as soon as we assume $f_u \in H_u^0$ for all $u \in S$.*

Remark 1. *The components of the HOFD* (2.2) *are referred as hierarchically orthogonal, that is $\langle f_u, f_v \rangle = 0\ \forall v \subset u$.*

To get more details on the HOFD, the reader is referred to Hooker (2007); Chastaing, Gamboa and Prieur (2012). In this paper, we are interested in estimating the summands in (2.2). As underlined in Huang (1998), estimating all components of (2.2) suffers from a curse of dimensionality, leading to an intractable problem in practice. To bypass this issue, we assume further along the article (without loss of generality) that f is centered, so that $f_\emptyset = 0$ and suppose that f is well approximated by

$$f(\mathbf{X}) \simeq \sum_{\substack{u \in S^* \\ |u| \leq d}} f_u(\mathbf{X}_u), \quad d \ll p \tag{2.3}$$

We thus assume that interactions of order $\geq d+1$ can be neglected. But even by choosing $d = 2$, the number of components in (2.3) can become prohibitive if the number of inputs p is high. We therefore are interested by estimation procedures under sparse assumptions when the number of variables p is large.
In the next section, we remind the procedure to identify components of (2.3). Through this strategy, we highlight the curse of dimensionality when p is getting large, and we propose to use a greedy $\mathbb{L}_2$-boosting to tackle this issue.

2.3 Practical determination of the Sparse HOFD

General description of the procedure

We propose in this section a Two-Steps estimation procedure to identify the components in (2.3): the first one is a simplified version of the Hierarchical Orthogonal Gram-Schmidt (HOGS) procedure developed in Chastaing, Gamboa and Prieur (2013), and the second consists of a $\mathbb{L}_2$-boosting algorithm (see *e.g.* Friedman (2001); Bühlmann (2006)). The specificity of our new $\mathbb{L}_2$-boosting algorithm is that it is based on a random dictionary and then falls into the framework of sparse recovery problem with error in the variables.

To lead this two-steps procedure, we assume that we observe two independent and identically distributed samples $(y^r, \mathbf{x}^r)_{r=1,\cdots,n_1}$ and $(y^s, \mathbf{x}^s)_{s=1,\cdots,n_2}$ from the distribution of $(Y, \mathbf{X})$ (the initial sample can be splitted in such two samples). We define the empirical inner product $\langle \cdot, \cdot \rangle_n$ and the empirical norm $\|\cdot\|_n$ associated to a n-sample as

$$\langle h, g \rangle_n = \frac{1}{n} \sum_{s=1}^{n} h(\mathbf{x}^s) g(\mathbf{x}^s), \quad \|h\|_n = \langle h, h \rangle_n.$$

Also, for $u = (u_1, \cdots, u_t) \in S$, we define the multi-index $\boldsymbol{l_u} = (l_{u_1}, \cdots, l_{u_t}) \in \mathbb{N}^t$. We use the notation Span$\{B\}$ to define the set of all finite linear combination of elements of B, also called the linear span of B.

Step 1 and Step 2 of our sparse HOFD procedure will be described in details further below.

Remark 2. *In the following, we assume that $d = 2$ in (2.3). The procedure could be extended to any higher order approximation, but we think that the description of the methodology for $d = 2$ helps for a better understanding. We thus have chosen to only describe this situation for the sake of clarity.*

Step 1: Hierarchically Orthogonal Gram-Schmidt procedure

For each $i \in [1 : p]$, let $\{\Psi^i_{l_i},\ l_i \in \mathbb{N}\}$ denote an orthonormal basis of $H_i := L^2(\mathbb{R}, \mathcal{B}(\mathbb{R}), P_{X_i})$. For $L \in \mathbb{N}^*$, for $i \neq j \in [1 : p]$, we set

$$H^L_\emptyset = \text{Span}\{1\} \quad \text{and} \quad H^L_i = \text{Span}\left\{1, \psi^i_1, \cdots, \psi^i_L\right\},$$

as well as

$$H^L_{ij} = \text{Span}\left\{1, \psi^i_1, \cdots, \psi^i_L, \psi^j_1, \cdots, \psi^j_L, \psi^i_1 \otimes \psi^j_1, \cdots, \psi^i_L \otimes \psi^j_L\right\}.$$

We define $H^{L,0}_u$, the approximation of H^0_u, as

$$H^{L,0}_u = \left\{h_u \in H^L_u,\ \langle h_u, h_v\rangle = 0, \forall\ v \subset u, \forall\ h_v \in H^{L,0}_v\right\},$$

The recursive procedure below aims at constructing a basis of $H^{L,0}_i$ and a basis of $H^{L,0}_{ij}$ for any $i \neq j \in [1 : p]$.

Initialization For any $1 \leq i \leq p$, define $\phi^i_{l_i} := \Psi^i_{l_i}$, $l_i \in [1 : L]$. Then, thanks to the orthogonality of $\{\Psi^i_{l_i},\ l_i \in \mathbb{N}\}$, we get $H^{L,0}_i := \text{Span}\left\{\phi^i_1, \cdots \phi^i_L\right\}$.

Second order interactions Let $u = \{i, j\}$, with $i \neq j \in [1 : p]$. As the dimension of H^L_{ij} is equal to $L^2 + 2L + 1$, and that the approximation space $H^{L,0}_{ij}$ is subject to $2L + 1$ constraints, its dimension is then equal to L^2. We want to construct a basis for $H^{L,0}_{ij}$, which satisfies the hierarchical orthogonal constraints. We are looking for such a basis of the form:

$$\begin{aligned} \phi^{ij}_{\boldsymbol{l_{ij}}}(X_i, X_j) \quad = \quad & \phi^i_{l_i}(X_i) \times \phi^j_{l_j}(X_j) + \textstyle\sum_{k=1}^L \lambda^i_{k,\boldsymbol{l_{ij}}} \phi^i_k(X_i) \\ & + \textstyle\sum_{k=1}^L \lambda^j_{k,\boldsymbol{l_{ij}}} \phi^j_k(X_j) + C_{\boldsymbol{l_{ij}}}, \end{aligned} \tag{2.4}$$

with $\boldsymbol{l_{ij}} = (l_i, l_j) \in [1 : L]^2$.

The constants $(C_{\boldsymbol{l_{ij}}}, (\lambda^i_{k,\boldsymbol{l_{ij}}})_{k=1}^L, (\lambda^j_{k,\boldsymbol{l_{ij}}})_{k=1}^L)$ are determined by resolving the following constraints:

$$\begin{array}{ll} \langle \phi^{ij}_{\boldsymbol{l_{ij}}}, \phi^i_k \rangle = 0, & \forall\ k \in [1:L] \\ \langle \phi^{ij}_{\boldsymbol{l_{ij}}}, \phi^j_k \rangle = 0, & \forall\ k \in [1:L] \\ \langle \phi^{ij}_{\boldsymbol{l_{ij}}}, 1 \rangle = 0. & \end{array} \tag{2.5}$$

We first solve the linear system:

$$A^{ij} \boldsymbol{\lambda}^{\boldsymbol{l_{ij}}} = D^{\boldsymbol{l_{ij}}}, \tag{2.6}$$

where $A^{ij} = \begin{pmatrix} \mathbb{E}(\Phi_i{}^t\Phi_i) & \mathbb{E}(\Phi_i{}^t\Phi_j) \\ \mathbb{E}(\Phi_j{}^t\Phi_i) & \mathbb{E}(\Phi_j{}^t\Phi_j) \end{pmatrix}$, with $(\Phi_i)_k = \phi^i_k$, and $(\Phi_j)_k = \phi^j_k$ for $k \in [1:L]$. Also, $\boldsymbol{\lambda}^{\boldsymbol{l_{ij}}} = \begin{pmatrix} \lambda^i_{1,\boldsymbol{l_{ij}}} & \cdots & \lambda^i_{L,\boldsymbol{l_{ij}}} & \lambda^j_{1,\boldsymbol{l_{ij}}} & \cdots & \lambda^j_{L,\boldsymbol{l_{ij}}} \end{pmatrix}^t$,
$D^{\boldsymbol{l_{ij}}} = -\begin{pmatrix} \langle \phi^i_{l_i} \times \phi^j_{l_j}, \phi^i_1 \rangle & \cdots & \langle \phi^i_{l_i} \times \phi^j_{l_j}, \phi^i_L \rangle & \langle \phi^i_{l_i} \times \phi^j_{l_j}, \phi^j_1 \rangle & \cdots & \langle \phi^i_{l_i} \times \phi^j_{l_j}, \phi^j_L \rangle \end{pmatrix}^t$.
As shown in Chastaing, Gamboa and Prieur (2013), $A^{\boldsymbol{l_{ij}}}$ is a definite positive Gramian matrix and (2.6) admits a unique solution in $\boldsymbol{\lambda}^{\boldsymbol{l_{ij}}}$. Next, $C_{\boldsymbol{l_{ij}}}$ is deduced with

$$C_{\boldsymbol{l_{ij}}} = -\mathbb{E}\Bigg[\phi^i_{l_i} \otimes \phi^j_{l_j}(X_i, X_j) + \sum_{k=1}^{L} \lambda^i_{k,\boldsymbol{l_{ij}}} \phi^i_k(X_i) + \sum_{k=1}^{L} \lambda^j_{k,\boldsymbol{l_{ij}}} \phi^j_k(X_j)\Bigg]. \tag{2.7}$$

Higher interactions This construction can be extended to any $|u| \geq 3$. We refer the interested reader to Chastaing, Gamboa and Prieur (2013). Just note that the dimension of the approximation space $H_u^{L,0}$ is given by $L_u = L^{|u|}$, where $|u|$ denotes the cardinality of u.

Empirical procedure Algorithm 1 below proposes an empirical version of the HOGS procedure. It consists in substituting the inner product $\langle \cdot, \cdot \rangle$ by its empirical version $\langle \cdot, \cdot \rangle_{n_1}$ obtained with the first data set $(y^r, \mathbf{x}^r)_{r=1,\cdots,n_1}$.

Algorithm 1: Empirical HOFD (EHOFD)

```
Input: Orthonormal system $(\phi^i_{l_i})_{l_i=0}^L$ of $H_i$, $i \in [1:p]$, i.i.d. observations
       $\mathcal{O}_1 := (y^r, \mathbf{x}^r)_{r=1,\cdots,n_1}$ of (2.1), threshold $|u_{max}|$
Initialization: for any $i \in [1:p]$ and $l_i \in [1:L]$, define first $\hat{\phi}^i_{l_i,n_1} = \phi^i_{l_i}$.
    • For any $u$ such that $2 \leq |u| \leq |u_{max}|$, write the matrix $(\hat{A}^{ij}_{n_1})$ as well as
      $(\hat{D}^{\boldsymbol{l_{ij}}}_{n_1})$ obtained using the former expressions with $\langle \cdot, \cdot \rangle_{n_1}$.
    • Solve (2.6) with the empirical inner product $\langle \cdot, \cdot \rangle_{n_1}$ and compute $(\hat{\boldsymbol{\lambda}}^{\boldsymbol{l_{ij}}}_{n_1})$.
    • Compute $\hat{C}^{n_1}_{\boldsymbol{l_{ij}}}$ by using Equation (2.7) and $(\hat{\boldsymbol{\lambda}}^{\boldsymbol{l_{ij}}}_{n_1})$.
    • The empirical version of the basis given by (2.4) is then:
      $\forall u \in [2 : |u_{max}|] \quad \hat{H}^{L,0,n_1}_u = \text{Span}\left\{\hat{\phi}^u_{1,n_1}, \cdots, \hat{\phi}^u_{L_u,n_1}\right\}$, where $L_u = L^{|u|}$.
```

Step 2: Greedy selection of Sparse HOFD

Each component f_u of the HOFD defined in Definition 1 is a projection onto H^0_u. Since, for $u \in S^*$, the space $\hat{H}^{L,0,n_1}_u$ well approximates H^0_u, it is then natural to approximate f by:

$$f(\mathbf{x}) \simeq \bar{f}(\mathbf{x}) = \sum_{\substack{u \in S^* \\ |u| \leq d}} \bar{f}_u(\mathbf{x}_u), \text{ with } \bar{f}_u(\mathbf{x}_u) = \sum_{\boldsymbol{l_u}} \beta^u_{\boldsymbol{l_u}} \hat{\phi}^u_{\boldsymbol{l_u},n_1}(\mathbf{x}_u),$$

where $\boldsymbol{l_u}$ is the multi-index $\boldsymbol{l_u} = (l_i)_{i \in u} \in [1:L]^{|u|}$. For the sake of clarity (since there is no ambiguity), we will omit the summation support of $\boldsymbol{l_u}$ in the sequel.

Now, we consider the second sample $(y^s, \mathbf{x}^s)_{s=1,\cdots,n_2}$ and we aim to recover the unknown coefficients $(\beta^u_{\boldsymbol{l_u}})_{\boldsymbol{l_u},|u|\leq d}$ on the regression problem,

$$y^s = \bar{f}(\mathbf{x}^s) + \varepsilon^s, \quad s = 1, \cdots, n_2.$$

However, the number of coefficients is equal to $\sum_{k=1}^d \binom{p}{k} L^k$. When p gets large, the usual least-squares estimator is not adapted to estimate the coefficients $(\beta^u_{\boldsymbol{l_u}})_{\boldsymbol{l_u},u}$. We then use the penalized regression,

$$(\hat{\beta}^u_{\boldsymbol{l_u}}) \in \underset{\beta^u_{\boldsymbol{l_u}} \in \mathbb{R}}{\text{Argmin}} \frac{1}{n_2} \sum_{s=1}^{n_2} \left[y^s - \sum_{\substack{u \in S^* \\ |u| \leq d}} \sum_{\boldsymbol{l_u}} \beta^u_{\boldsymbol{l_u}} \hat{\phi}^u_{\boldsymbol{l_u},n_1}(\mathbf{x}^s_u) \right]^2 + \lambda J(\beta^1_1, \cdots, \beta^u_{\boldsymbol{l_u}}, \cdots),$$

where $J(\cdot)$ is the ℓ_0-penalty, i.e.

$$J(\beta_1^1, \cdots, \beta_{\boldsymbol{l_u}}^u, \cdots) = \sum_{\substack{u \in S^* \\ |u| \leq d}} \sum_{\boldsymbol{l_u}} \mathbb{1}(\beta_{\boldsymbol{l_u}}^u \neq 0).$$

Of course, such an optimisation procedure is not tractable and we instead consider the relaxed $\mathbb{L}_2$-boosting (see *e.g.* Friedman (2001)) to solve this penalized problem. Mimicking the notation of Temlyakov (2000); Champion, Cierco-Ayrolles, Gadat and Vignes (2013), we define the dictionary $\mathcal{D}$ of functions as

$$\mathcal{D} = \{\hat{\phi}_{1,n_1}^1, \cdots \hat{\phi}_{L,n_1}^1, \cdots, \hat{\phi}_{1,n_1}^u, \cdots, \hat{\phi}_{L_u,n_1}^u, \cdots\}.$$

The quantity $G_k(\bar{f})$ denotes the approximation of $\bar{f}$ at step k, as a linear combination of elements of $\mathcal{D}$. At the end of the algorithm, the estimation of $\bar{f}$ is denoted $\hat{f}$. The $\mathbb{L}_2$-boosting is described in Algorithm 2.

Algorithm 2: The $\mathbb{L}_2$-boosting

```
Input: Observations $\mathcal{O}_2 := (y^s, \mathbf{x}^s)_{s=1,\cdots,n_2}$, shrinkage parameters
       $\gamma \in ]0,1]$ and number of iterations $k_{up} \in \mathbb{N}^*$.
Initialization: $G_0(\bar{f}) = 0$.
for $k = 1$ to $k_{up}$ do
    1. Select $\hat{\phi}_{\boldsymbol{l_{u_k}},n_1}^{u_k} \in \mathcal{D}$ such that
       $|\langle Y - G_{k-1}(\bar{f}), \hat{\phi}_{\boldsymbol{l_{u_k}},n_1}^{u_k}\rangle_{n_2}| = \max_{\hat{\phi}_{\boldsymbol{l_u},n_1}^u \in \mathcal{D}} |\langle Y - G_{k-1}(\bar{f}), \hat{\phi}_{\boldsymbol{l_u},n_1}^u\rangle_{n_2}|.$   (2.8)
    2. Compute the new approximation of $\bar{f}$ as
       $G_k(\bar{f}) = G_{k-1}(\bar{f}) + \gamma\langle Y - G_{k-1}(\bar{f}), \hat{\phi}_{\boldsymbol{l_{u_k}},n_1}^{u_k}\rangle_{n_2} \cdot \hat{\phi}_{\boldsymbol{l_{u_k}},n_1}^{u_k}.$   (2.9)
end
Output: $\hat{f} = G_{k_{up}}(\bar{f})$.
```

For any step k, Algorithm 2 selects a function from $\mathcal{D}$ wich provides a sufficient information on the residual $Y - G_{k-1}(\bar{f})$. The shrinkage parameter γ is the standard step-length parameter of the boosting algorithm. It actually smoothly inserts the next predictor in the model, making possible a refinement of the greedy algorithm, and may statistically guarantees its convergence rate.

Remark 3. *In a deterministic setting, the shrinkage parameter is not really useful and may be set to* 1 *(see Temlyakov (2000) for further details). It is indeed useful from a practical point of view to smooth the boosting iterations.*

An algorithm for our new sparse HOFD procedure

Algorithm 3 below provides now a simplified description of our sparse HOFD procedure, whose steps have been described further above.

```
Algorithm 3: Greedy Hierarchically Orthogonal Functional Decomposition

Input: Orthonormal system $(\Psi^i_{l_i})_{l_i=0}^L$ of $L^2(\mathbb{R},\mathcal{B}(\mathbb{R}),P_{X_i})$, $i \in [1:p]$, i.i.d.
       observations $\mathcal{O} := (y^j, \mathbf{x}^j)_{j=1\dots n}$ of (2.1)
Initialization: Split $\mathcal{O}$ in a partition $\mathcal{O}_1 \cup \mathcal{O}_2$ of size $(n_1, n_2)$.
    • For any $u \in S$, use Step 1 with observations $\mathcal{O}_1$ to construct the
      approximation $\hat{H}_u^{L,0,n_1} := \text{Span}\left\{\hat{\phi}^u_{1,n_1}, \cdots, \hat{\phi}^u_{L_u,n_1}\right\}$ of $H_u^{L,0}$ (see Algorithm
      1).
    • Use an $\mathbb{L}_2$-boosting algorithm on $\mathcal{O}_2$ with the random dictionary
      $\mathcal{D} = \{\hat{\phi}^1_{1,n_1}, \cdots \hat{\phi}^1_{L,n_1}, \cdots, \hat{\phi}^u_{1,n_1}, \cdots, \hat{\phi}^u_{L_u,n_1}, \cdots\}$ to obtain the Sparse
      Hierarchically Orthogonal Decomposition (see Algorithm 2).
```

We now obtain a strategy to estimate the components of the decomposition (2.3) in a high-dimensional paradigm. We aim to show that the obtained estimators are consistent, and that the Two-Steps procedure (summarized in Algorithm 3) is numerically convincing. The next section is devoted to the asymptotic properties of the estimators.

3. Consistency of the estimator

In this section, we study the asymptotic properties of the estimator $\hat{f}$ obtained from the Algorithm 3 described in Section 2. To this end, we restrict our study to the case of $d = 2$ and assume that f is well approximated by first and second order interaction components. Hence, the observed signal Y may be represented as

$$Y = \sum_{\substack{u \in S^* \\ |u| \leq 2}} \sum_{\boldsymbol{l_u}} \beta^{u,0}_{\boldsymbol{l_u}} \phi^u_{\boldsymbol{l_u}}(\mathbf{X}_u) + \varepsilon, \quad \mathbb{E}(\varepsilon) = 0,\ \mathbb{E}(\varepsilon^2) = \sigma^2,$$

where $\boldsymbol{\beta}^0 = (\beta^{u,0}_{\boldsymbol{l_u}})_{\boldsymbol{l_u},u}$ is the true parameter, and the functions $(\phi^u_{\boldsymbol{l_u}})_{\boldsymbol{l_u}}$, $|u| \leq 2$ are constructed according to the HOFD described in the paragraph . We assume that we have in hand a n-sample of observations, divided into two samples $\mathcal{O}_1$ and $\mathcal{O}_2$. Samples in $\mathcal{O}_1$ (*resp.* in $\mathcal{O}_2$) of size $n_1 = n/2$ (*resp.* of size $n_2 = n/2$) are used for the construction of $(\hat{\phi}^u_{\boldsymbol{l_u},n_1})_{\boldsymbol{l_u},u}$ described in Algorithm 1 (*resp.* for the $\mathbb{L}_2$-boosting Algorithm 2 to estimate $(\beta^u_{\boldsymbol{l_u}})_{\boldsymbol{l_u},u}$).

The goal of this section is to study the consistency of $\hat{f} = G_{k_n}(\bar{f})$ when the sample size n tends to infinity. Its objective is also to determine an optimal number of steps k_n to get a consistent estimator from Algorithm 2.

3.1 Assumptions

We first briefly recall some notation: for any sequences $(a_n)_{n\geq 0}$, $(b_n)_{n\geq 0}$, we write $a_n = \underset{n\to+\infty}{\mathcal{O}}(b_n)$ when a_n/b_n is a bounded sequence for n large enough. Now, for any random sequence $(X_n)_{n\geq 0}$, $X_n = \mathcal{O}_P(a_n)$ means that $|X_n/a_n|$ is bounded in probability.

We have chosen to present our assumptions in three parts to deal with the dimension, the noise and the sparseness of the entries.

Bounded Assumptions $(\mathbf{H_b})$ The first set of hypotheses matches with the *bounded case* and is adapted to the special situation of bounded support for the random variable X, for instance when each X_j follows a uniform law on a compact set $\mathcal{K}_j \subset K$ where K is a compact set of $\mathbb{R}$ independent of $j \in [1:p]$. It is refered as $(\mathbf{H_b})$ in the sequel and corresponds to the following three conditions.

$(\mathbf{H_b^1})$ $M := \sup_{\substack{i\in[1:p]\\ l_i\in[1:L]}} \left\|\phi^i_{l_i}(X_i)\right\|_\infty < +\infty$,

$(\mathbf{H_b^2})$ The number of variables p_n satisfies

$$p_n = \underset{n\to+\infty}{\mathcal{O}}(\exp(Cn^{1-\xi})), \text{ where } 0 < \xi \leq 1 \text{ and } C > 0.$$

$(\mathbf{H_b^{3,\vartheta}})$ The Gram matrices A^{ij} introduced in (2.6) satisfies:

$$\exists C > 0 \quad \forall (i,j) \in [1:p_n]^2 \qquad det(A^{ij}) \geq Cn^{-\vartheta},$$

where *det* denotes the determinant of a matrix.

Roughly speaking, this will be the favorable situation from a technical point of view since it will be possible to apply a Matrix Hoeffding's type Inequality. It may be possible to slightly relax such an hypothesis using a sub-exponential tail argument. For the sake of simplicity, we have chosen to only restrict our work to the settings of $(\mathbf{H_b})$.

Whatever the joint law of the random variables $(X_1, \ldots, X_p)$ is, it is always possible to build an orthonormal basis $(\phi^i_{l_i})_{1 \leq l_i \leq L}$ from a bounded (frequency truncated) Fourier basis and thus $(\mathbf{H^1_b})$ is not so restrictive in practice.

Assumption $(\mathbf{H^2_b})$ copes with the high dimensional situation. The number of variables p_n can grow exponentially fast with the number of observations n.

Note that Hypothesis $(\mathbf{H^{3,\vartheta}_b})$ stands for a lower bound of the determinant of the Gram matrices involved in the HOFD. It is shown in Chastaing, Gamboa and Prieur (2013) that each of these Gram matrices are invertible and thus each $\det(A^{ij})$ are positive. Nevertheless, if $\vartheta = 0$, this hypothesis assume that such an invertibility is *uniform* over all choices of tensor (i, j). This hypothesis may be too strong for a large number of variables $p_n \to +\infty$ when $\vartheta = 0$. However, when $\vartheta > 0$, Hypothesis $(\mathbf{H^{3,\vartheta}_b})$ drastically relax the case $\vartheta = 0$ and becomes very weak. It will be satisfied in many of our numerical examples. In the sequel, the parameters ϑ and ξ will be related each other and we will obtain a consistency result of the sparse HOFD up to the condition $\vartheta < \xi/2$. This constraint implicitely limits the size of p_n since $\log p_n = \underset{n \to +\infty}{\mathcal{O}}(n^{1-\xi})$.

Noise Assumption $(\mathbf{H_{\varepsilon,q}})$ We will assume the noise measurement ε to get some bounded moments of sufficiently high order, which is true for Gaussian or bounded noise. This assumption is given by

$$(\mathbf{H_{\varepsilon,q}})\ \mathbb{E}(|\varepsilon|^q) < \infty, \quad \text{for one } q \in \mathbb{R}_+.$$

Sparsity Assumption $(\mathbf{H_s})$ The last assumption concerns the sparse representation of the unknown signal described by Y in the basis $(\phi^u_{\boldsymbol{l_u}}(\mathbf{X}_u))_u$. Such an hypothesis will be usefull to assess the statistical performance of the $\mathbb{L}_2$-boosting and will be refered as $(\mathbf{H_s})$ in the sequel. It is legitimate by our high dimension setting and our motivation to identify the main interactions $\mathbf{X}_u$.

$(\mathbf{H_s})$ The true parameter $\boldsymbol{\beta}^0$ satisfies uniformly with n

$$\|\beta^0\|_{L^1} := \sum_{\substack{u\in S^*\\ |u|\leq d}} \sum_{\boldsymbol{l_u}} \left|\beta^{u,0}_{\boldsymbol{l_u}}\right| < \infty.$$

It is possible to relax this former condition and let $\|\boldsymbol{\beta}^0\|_{L^1}$ growing to $+\infty$ as $n \to +\infty$. The price to pay to face such a situation is then a more restrictive condition on the number of variables p_n. We refer to Bühlmann (2006) for a short discussion on a related problem and will only consider the situation described by $(\mathbf{H_s})$ for the sake of simplicity.

3.2 Main results

We first provide our main result on the efficiency of the EHOFD (Algorithm 1).

Theorem 1. *Assume that* $(\mathbf{H_b})$ *holds with* ξ *(resp.* ϑ*) given by* $(\mathbf{H_b^2})$ *(resp.* $(\mathbf{H_b^{3,\vartheta}})$*). Then, if* $\vartheta < \xi/2$*, the sequence of estimators* $(\hat{\phi}^u_{\boldsymbol{l_u},n_1})_u$ *satisfies:*

$$\sup_{\substack{u\in S^*,|u|\leq d\\ \boldsymbol{l_u}}} \left\|\hat{\phi}^u_{\boldsymbol{l_u},n_1} - \phi^u_{\boldsymbol{l_u}}\right\| = \zeta_{n,0} = \mathcal{O}_P(n^{\vartheta-\xi/2}).$$

The proof of this Theorem is deferred to the Appendix section. Our second main result concerns the $\mathbb{L}_2$-boosting which recovers the unknown $\tilde{f}$ up to a preprocessing estimation of $(\hat{\phi}^u_{\boldsymbol{l_u},n_1})_{\boldsymbol{l_u},u}$ on a first sample $\mathcal{O}_1$. Such a result is satisfied provided the sparsity Assumptions $(\mathbf{H_s})$. We assume that

$$Y = \tilde{f}(\mathbf{X}) + \varepsilon, \quad \tilde{f}(\mathbf{X}) = \sum_{\substack{u\in S^*\\ |u|\leq d}} \sum_{\boldsymbol{l_u}} \beta^{u,0}_{\boldsymbol{l_u}} \phi^u_{\boldsymbol{l_u}}(\mathbf{X}_u) \in H^L_u,$$

where $\boldsymbol{\beta}^0 = (\beta^{u,0}_{\boldsymbol{l_u}})_{\boldsymbol{l_u},u}$ is the true parameter that expands $\tilde{f}$.

Theorem 2 (Consistency of the $\mathbb{L}_2$-boosting)**.** *Consider an estimation* $\hat{f}$ *of* $\tilde{f}$ *from an i.i.d.* n*-sample broken up into* $\mathcal{O}_1 \cup \mathcal{O}_2$*. Assume that functions* $(\hat{\phi}^u_{\boldsymbol{l_u},n_1})_{\boldsymbol{l_u},u}$ *are estimated from the first sample* $\mathcal{O}_1$ *under* $(\mathbf{H_b})$ *with* $\vartheta < \xi/2$*. Then,* $\hat{f}$ *is defined by (6.13) of Algorithm 2 on* $\mathcal{O}_2$ *as*

$$\hat{f}(\mathbf{X}) = G_{k_n}(\bar{f}), \quad \textit{with} \;\; \bar{f} = \sum_{\substack{u\in S^*\\ |u|\leq d}} \sum_{\boldsymbol{l_u}} \beta^{u,0}_{\boldsymbol{l_u}} \hat{\phi}^u_{\boldsymbol{l_u},n_1}(\mathbf{X}_u).$$

If we assume that $(\mathbf{H_s})$ *and* $(\mathbf{H_{\varepsilon,q}})$ *are satisfied with* $q > 4/\xi$*, then there exists a sequence* $k_n := C \log n$*, with* $C < (\xi/2 - \vartheta)/(2 \cdot \log 3)$ *such that*

$$\|\hat{f} - \tilde{f}\| \xrightarrow{\mathbb{P}} 0, \textit{when} \quad n \to +\infty.$$

We briefly describe the proof and postpone the technical details to the Appendix section.

Sketch of Proof of Theorem 2. Mimicking the scheme of Bühlmann (2006) and Champion, Cierco-Ayrolles, Gadat and Vignes (2013), the proof first consists in defining the theoretical residual of Algorithm 2 at step k as

$$\begin{array}{rcl} R_k(\bar{f}) & = & \bar{f} - G_k(\bar{f}) \\ & = & \bar{f} - G_{k-1}(\bar{f}) - \gamma\langle Y - G_{k-1}(\bar{f}), \hat{\phi}^{u_k}_{\boldsymbol{l_{u_k}},n_1}\rangle_{n_2} \cdot \hat{\phi}^{u_k}_{\boldsymbol{l_{u_k}},n_1} \end{array} \tag{3.1}$$

Further, following the work of Champion, Cierco-Ayrolles, Gadat and Vignes (2013), we introduce a *phantom* residual in order to reproduce the behaviour of a deterministic boosting, studied in Temlyakov (2000). This *phantom* algorithm is the theoretical $\mathbb{L}_2$-boosting, performed using the randomly chosen elements of the dictionary by Equations (2.8) and (6.13), but updated using the deterministic inner product. The *phantom* residuals $\tilde{R}_k(\bar{f})$, $k \geq 0$, are defined as follows,

$$\left\{ \begin{array}{l} \tilde{R}_0(\bar{f}) = \bar{f} \\ \tilde{R}_k(\bar{f}) = \tilde{R}_{k-1}(\bar{f}) - \gamma\langle \tilde{R}_{k-1}(\bar{f}), \hat{\phi}^{u_k}_{\boldsymbol{l_{u_k}},n_1}\rangle \hat{\phi}^{u_k}_{\boldsymbol{l_{u_k}},n_1}, \end{array} \right. \tag{3.2}$$

where $\hat{\phi}^{u_k}_{\boldsymbol{l_{u_k}},n_1}$ has been selected with Equation (2.8) of Algorithm 2. The aim is to decompose the quantity $\left\|\hat{f} - \tilde{f}\right\|$ to introduce the theoretical residuals and the *phantom* ones,

$$\left\|\hat{f} - \tilde{f}\right\| = \left\|G_{k_n}(\bar{f}) - \tilde{f}\right\| \leq \left\|\bar{f} - \tilde{f}\right\| + \left\|R_{k_n}(\bar{f}) - \tilde{R}_{k_n}(\bar{f})\right\| + \left\|\tilde{R}_{k_n}(\bar{f})\right\|. \tag{3.3}$$

We then have to show that each term of the right-hand side of (3.3) converges towards zero in probability. □

4. Numerical Applications

In this section, we are interested by the numerical efficiency of the Two-Steps procedure given in Section 2, and we primarily focus on the practical use of the HOFD through sensitivity analysis (SA). The goal of SA is to identify and to rank the input variables that drive the uncertainty of the model output. For further details, the reader may refer to Saltelli, Chan and Scott (2000); Cacuci, Ionescu-Bujor and Navon (2005). Therefore, the HOFD presented in Paragraph 2.2 is of great interest, because it may be used to decompose the global variance of the model. Here, as each HOFD is subject to hierarchical orthogonality constraints given in Definition 1, we obtain that

$$V(Y) = \sum_{u \in S^*} \left[V(f_u(\mathbf{X}_u)) + \sum_{u \cap v \neq u,v} \mathrm{Cov}(f_u(\mathbf{X}_u), f_v(\mathbf{X}_v)) \right]$$

Therefore, to measure the contribution of $\mathbf{X}_u$, for $|u| \geq 1$, in terms of variability in the model, it is then quite natural to define a sensitivity index S_u as follows,

$$S_u = \frac{V(f_u(\mathbf{X}_u)) + \sum_{u \cap v \neq u,v} \mathrm{Cov}(f_u(\mathbf{X}_u), f_v(\mathbf{X}_v))}{V(Y)}.$$

This definition is given and discussed in Chastaing, Gamboa and Prieur (2012). In practice, once we have applied the procedure described in Algorithm 3 to get $(\hat{f}_u, \hat{f}_v, u \cap v \neq u, v)$, it is straightforward to deduce the empirical estimation of S_u, for all u. In the following, we are mostly interested by the estimation of the first and second order sensitivity indices (i.e. S_i and S_{ij}, $i, j \in [1:p]$).

4.1 Description

We end the work with a short simulation study and we are primarily interested by the performance of the greedy selection algorithm for the prediction of generalized sensitivity indices. As the estimation of these indices consists in estimating the summands of the generalized functional ANOVA decomposition (called HOFD), we start by constructing a hierarchically orthogonal system of functions to approximate the components. As pointed above (see Assumption $(\mathbf{H}_{\mathbf{b}}^{\mathbf{3},\vartheta})$ in Theorem 1 and 2), the invertibility of each linear system plays an important role in our theoretical study. We hence have measured for each situation the degeneracy of involved matrices given by

$$d(A) = \inf_{i,j \in [1:p]} \det(A^{ij}).$$

Then, we use a variable selection method to select a sparse number of predictors. The goal is to numerically compare three variable selection methods: the $\mathbb{L}_2$-boosting, the Forward-Backward greedy algorithm (refered as FoBa in the sequel), and the Lasso estimator. As pointed above, we have in hand a n-sample of i.i.d. observations $(y^s, \mathbf{x}^s)_{s=1,\cdots,n}$ broken up into two samples of size $n_1 = n_2 = n/2$. The first sample is used to construct the system of functions according to Algorithm 1. Let us now briefly describe how we use the Lasso and the FoBa. Each of the three selection methods aims to solve a generic minimization problem

$$(\hat{\beta}^u_{\boldsymbol{l_u}})_{\boldsymbol{l_u},u} \in \underset{\beta^u_{\boldsymbol{l_u}} \in \mathbb{R}}{\operatorname{Argmin}} \frac{1}{n_2} \sum_{s=1}^{n_2} \Big[y^s - \sum_{\substack{u \in S \\ |u| \leq d}} \sum_{\boldsymbol{l_u}} \beta^u_{\boldsymbol{l_u}} \hat{\phi}^u_{\boldsymbol{l_u},n_1}(\mathbf{x}^s_u) \Big]^2 + \lambda J(\beta^1_1, \cdots, \beta^u_{\boldsymbol{l_u}}, \cdots),$$

4.2 Feature selection Algorithms

FoBa procedure The FoBa algorithm, as well as the $\mathbb{L}_2$-boosting, uses a greedy exploration to minimize the previous criterion when $J(\cdot)$ is a ℓ_0 penalty, i.e.

$$J(\beta^1_1, \cdots, \beta^u_{\boldsymbol{l_u}}, \cdots) = \sum_{\substack{u \in S^* \\ |u| \leq d}} \sum_{\boldsymbol{l_u}} \mathbb{1}(\beta^u_{\boldsymbol{l_u}} \neq 0).$$

This algorithm is an iterative scheme that sequentially selects or deletes an element of $\mathcal{D}$ that has the least impact on the fit, i.e. that significantly reduces the model residual. This algorithm is described in Zhang (2011), and exploited for HOFD in Chastaing, Gamboa and Prieur (2013). We refer to these references for a deeper description of this algorithm. This procedure depends on two shrinkage parameters ϵ and δ. The parameter ϵ is the stopping criterion, that predefines if a large number of predictors is going to be introduced in the model. The second parameter, $\delta \in]0,1]$ offers a flexibility in the *backward* step, as it allows the algorithm to smoothly eliminate at each step a predictor.

In our numerical experiments, we have found a well suited behaviour of the FoBa procedure with $\epsilon = 10^{-2}$ and $\delta = 1/2$.

Calibration of the Boosting We have set $\gamma = 0.7$ since it has been previously reported in Champion, Cierco-Ayrolles, Gadat and Vignes (2013) that it was a suitable value for high dimensional regression. As we do not know a priori the optimal value for k_{up}, we use a C_p-Mallows type criterion to fix the optimal number of iterations. We follow the recommendations of Efron, Hastie, Johnstone and Tibshirani (2004) to select the best solution in the LARS algorithm. First, we define a large number of iterations, say K. For each step $k \in \{1, \cdots, K\}$, the boosting algorithm computes an estimation of the solution $\hat{\boldsymbol{\beta}}(k)$. From this, we compute the following quantity,

$$E_k^{\text{Boost}} = \frac{1}{n} \sum_{s=1}^{n_2} \left[y^s - \sum_{\hat{\phi}^u_{\boldsymbol{l_u},n_1} \in \mathcal{D}} \hat{\beta}^u_{\boldsymbol{l_u}}(k) \hat{\phi}^u_{\boldsymbol{l_u},n_1}(\mathbf{x}^s_u) \right]^2 - n_2 + 2k,$$

where the implied set of functions $\hat{\phi}^u_{\boldsymbol{l_u},n_1}$ have been selected through the first k steps of the algorithm. At last, we choose the optimal number of selected functions $\hat{k}_{\text{up}}$ such that

$$\hat{k}_{\text{up}} = \operatorname*{Argmin}_{k=1,\cdots,K} E_k^{\text{Boost}}.$$

Lasso algorithm As the ℓ_0 strategy is very difficult to handle and may suffer from a lack of robustness, the ℓ_0 penalty is often replaced by the $\lambda \times \ell_1$ one, that yield to the Lasso estimator for a given penalization parameter $\lambda > 0$. A numerical way to solve it is to use the LARS regression, described in Efron, Hastie, Johnstone and Tibshirani (2004) and we refer to this standard reference for a sharp description of this procedure.

Admitting that for a given $\lambda > 0$, the Lasso regression admits a unique solution, as described in Tibshirani (1996), Efron, Hastie, Johnstone and Tibshirani (2004) show that the estimated solution with LARS coincide with the theoretical regularization path $\hat{\boldsymbol{\beta}}(\lambda)$. The LARS algorithm performs the Lasso regression by offering a set of solutions $\{\hat{\boldsymbol{\beta}}(\lambda),\ \lambda \in \mathbb{R}^+\}$. However, the "best" λ must be determined to only obtain one solution. In this view, we consider here the criterion defined in Efron, Hastie, Johnstone and Tibshirani (2004). At each step k of the algorithm, the following quantity is computed,

$$E_k^{\text{Lars}} = \left\| \mathbb{Y} - \mathbb{X}\hat{\boldsymbol{\beta}}(\lambda_k) \right\|_{n_2}^2 - n_2 + 2k$$

where λ_k is the regularization parameter of the kth step. The optimal $\hat{\lambda} = \lambda(\hat{k})$ is selected such that $\hat{k} = \operatorname{Argmin}_k E_k^{\text{Lars}}$ and we keep for the Lasso estimator $\hat{\beta}(\hat{\lambda})$.

4.3 Data sets

Each experiment on each data set has been randomly reproduced 50 times to compute the Monte-Carlo errors.

First Data set: the Ishigami function Well known in sensitivity analysis, the analytical form of the Ishigami model is given by,

$$Y = \sin(X_1) + a\sin^2(X_2) + bX_3^4\sin(X_1),$$

where we set $a = 7$ and $b = 0.1$, and where it is assumed that the inputs are independent. In the numerical experience, we consider the following cases.

1. For all $i = 1, 2, 3$, the inputs are uniformly distributed on $[-\pi, \pi]$. We choose $n = 300$ observations, with the first 8 Legendre basis functions ($L = 8$).
2. For all $i = 1, 2, 3$, the inputs are uniformly distributed on $[-\pi, \pi]$. We choose $n = 300$ observations, with the first 8 Fourier basis functions.

Each time, the number of predictors is $m_n = pL + \binom{p}{2}L^2 = 408 \geq n$.

Second Data set: the g-Sobol function This function is referred in Saltelli, Chan and Scott (2000), and is given by

$$Y = \prod_{i=1}^{p} \frac{|4X_i - 2| + a_i}{1 + a_i}, \quad a_i \geq 0,$$

where the inputs X_i are independent and uniformly distributed over $[0, 1]$. The analytical Sobol indices are given by

$$S_u = \frac{1}{D}\prod_{i \in u} D_i, \quad D_i = \frac{1}{3(1 + a_i)^2}, \ \ D = \prod_{i=1}^{p}(D_i + 1) - 1, \ \forall\ u \subseteq [1:p].$$

Here, we give $a = (0, 1, 4.5, 9, 99, 99, 99, 99, 99, 99)$. For the construction of the hierarchical basis functions, we choose the first 5 Legendre polynomials ($L = 5$).

The ANOVA representation is approximated by first and second order interaction effects, i.e. $d = 2$. We use $n = 700$ evaluations of the model and the number of predictors $m_n = pL + \binom{p}{2}L^2 = 1175$, which clearly exceeds the sample size n.

4.4 The tank pressure model

This real case study concerns a shell closed by a cap and subject to an internal pressure. Figure 4.1 illustrates a simulation of tank distortion. We are interested in the von Mises stress, detailed in von Mises (1913) on the point y labelled in Figure 4.1. The von Mises stress allows for predicting material yielding which occurs when it reaches the material yield strength. The selected point y corresponds to the point for which the von Mises stress is maximal in the tank. Therefore, we want to prevent the tank from material damage induced by plastic deformations. To offer a large panel of tanks able to resist to the internal pressure, a manufacturer wants to know the most contributive parameters to the von Mises criterion variability. In the model we propose, the von Mises criterion depends on three geometrical parameters: the shell internal radius (R_{int}), the shell thickness (T_{shell}), and the cap thickness (T_{cap}). It also depends on five physical parameters concerning the Young's modulus (E_{shell} and E_{cap}) and the yield strength ($\sigma_{y,shell}$ and $\sigma_{y,cap}$) of the shell and the cap. The last parameter is the internal pressure (P_{int}) applied to the shell. The system is modelized by a 2D finite elements code ASTER. In table 4.1, we give the input distributions.

Inputs	Distribution
R_{int}	$\mathcal{U}([1800;2200])$, $\gamma(R_{int}, T_{shell}) = 0.85$
T_{shell}	$\mathcal{U}([360;440])$, $\gamma(T_{shell}, T_{cap}) = 0.3$
T_{cap}	$\mathcal{U}([180;220])$, $\gamma(T_{cap}, R_{int}) = 0.3$
E_{cap} $\sigma_{y,cap}$	$\alpha N(\mu,\Sigma) + (1-\alpha)N(\mu,\Omega)$ $\alpha = 0.02$, $\mu = \begin{pmatrix} 210 \\ 500 \end{pmatrix}$, $\Sigma = \begin{pmatrix} 350 & 0 \\ 0 & 29 \end{pmatrix}$, $\Omega = \begin{pmatrix} 175 & 81 \\ 81 & 417 \end{pmatrix}$
E_{shell} $\sigma_{y,shell}$	$\alpha N(\mu,\Sigma) + (1-\alpha)N(\mu,\Omega)$ $\alpha = 0.02$, $\mu = \begin{pmatrix} 70 \\ 300 \end{pmatrix}$, $\Sigma = \begin{pmatrix} 117 & 0 \\ 0 & 500 \end{pmatrix}$, $\Omega = \begin{pmatrix} 58 & 37 \\ 37 & 250 \end{pmatrix}$
P_{int}	$N(80,10)$

Table 4.1: Description of inputs of the shell model

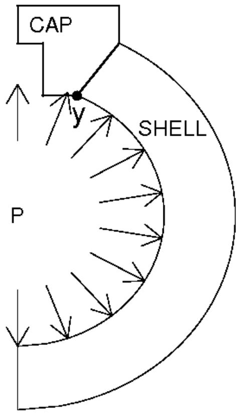

Figure 4.1: Tank distortion at point y

The geometrical parameters are uniformly distributed because of the large choice left for the tank building. The correlation γ between the geometrical pa-

rameters is induced by the constraints of manufacturing processes. The physical inputs are normally distributed and their uncertainty are due to the manufacturing process and the properties of the elementary constituents variabilities. The large variability of P_{int} in the model corresponds to the different internal pressure values which could be applied to the shell by the user.
To measure the contribution of the correlated inputs to the output variability, we estimate the generalized sensitivity indices. We proceed to $n = 1000$ simulations. We use the first Hermite basis functions whose maximum degree is 5 for every parameters.

4.5 Results

We consider both the estimation of the sensitivity indices, the ability to select the good representation of the different signals, and the computation time needed to obtain the sparse representation. "Greedy" refers to the Foba procedure as well as "LARS" refers to the Lasso resolution, and we refer to our method as "Boosting".

Sensitivity estimation Figures 4.2 and 4.3 provide the dispersion of the sensitivity indices estimated by our three methods on the Ishigami function. We can see that the three methods behave well with the two basis. Note that handling the Fourier basis is, as expected, more suitable for the Ishigami function than the Legendre basis (see the sensitivity index S_3 in Figures 4.2 and 4.3). We can also draw similar conclusions with Figure 4.4, where the three methods yields the same conclusion. Note also that the standard deviations of each method seem quite equivalent.

At last, as pointed by Figure 4.5, the most contributive parameter to the von Mises criterion variability is the internal pressure P_{int}, which is not surprising. Concerning now the geometric characteristics, the three methods exhibit as main parameters the cap thickness T_{cap} and the shell thickness T_{shell} using their expensive code although the shell internal radius does not seem so important.

Computation time and accuracy We enumerate in Table 4.2 the performances of the three methods, according to their computational cost, and accuracy of the

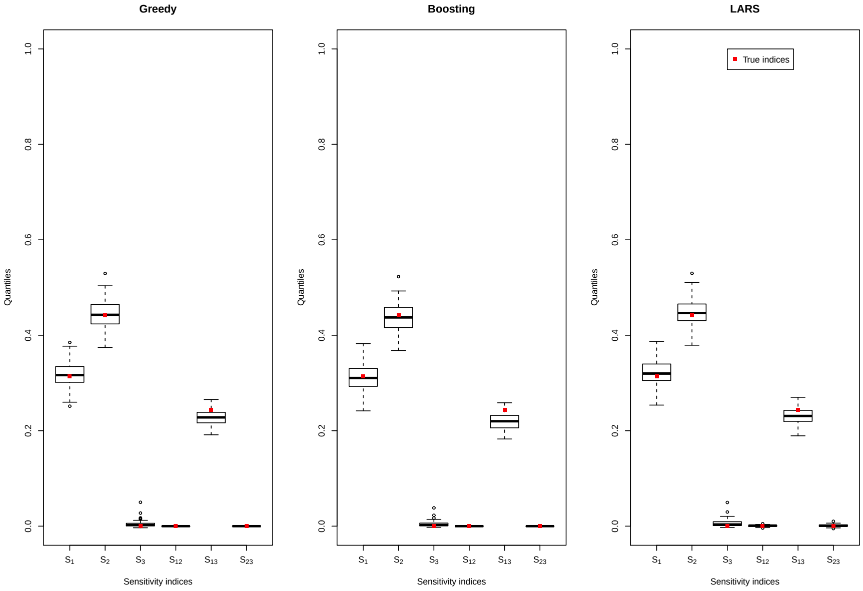

Figure 4.2: Representation of the first-order components on the First Data set (Ishigami function) described through the Fourier basis

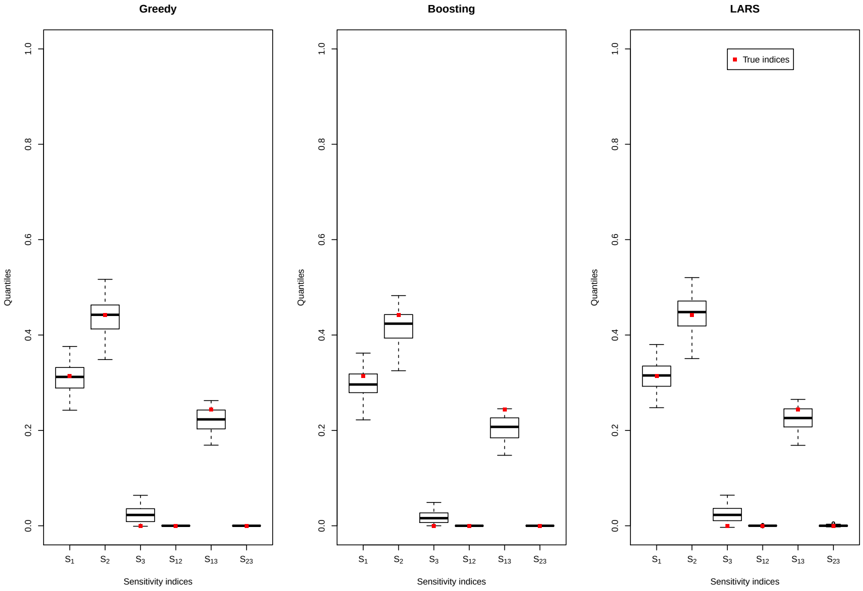

Figure 4.3: Representation of the first-order components on the First Data set (Ishigami function) described through the Legendre basis

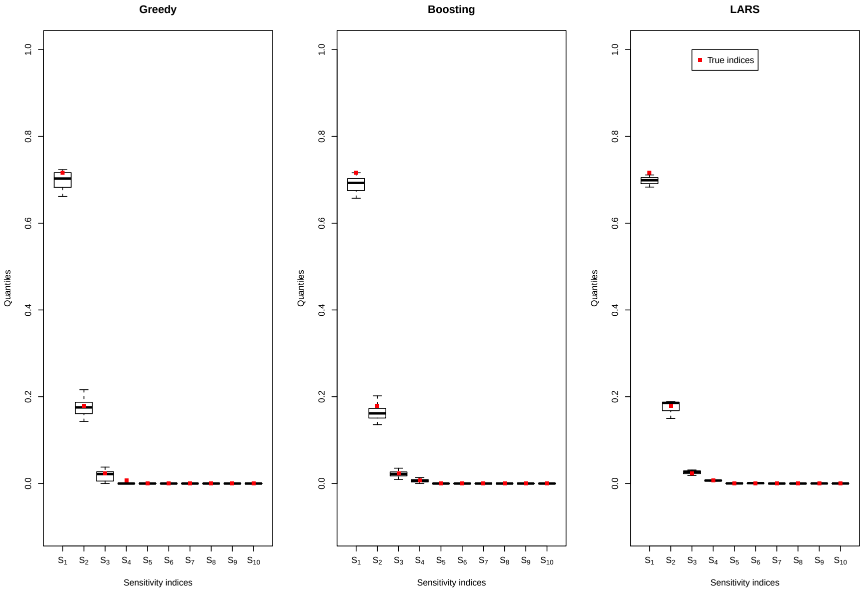

Figure 4.4: Representation of the first-order components on the Second Data set (g-Sobol function)

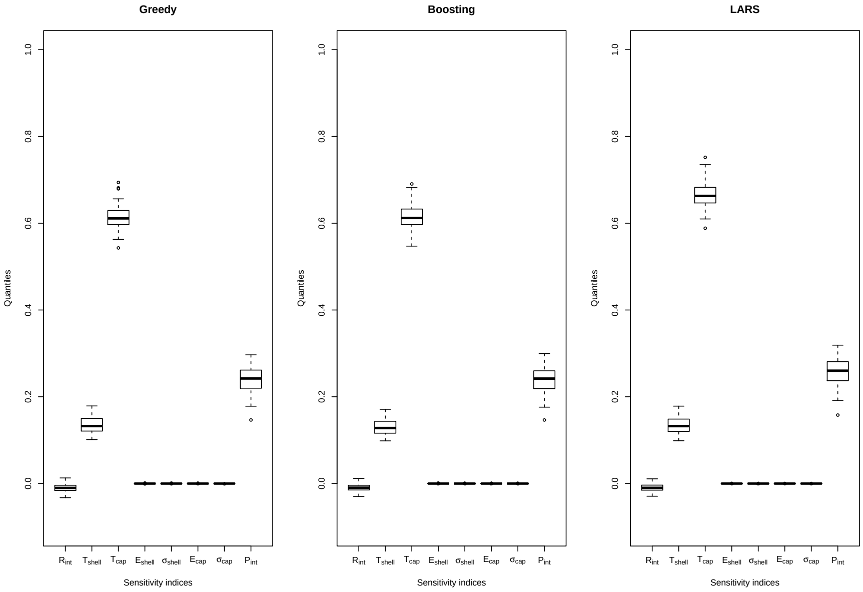

Figure 4.5: Dispersion of the first order sensitivity indices of the tank model parameters

feature selection.

Data set	Procedure	$\left\|\hat{\boldsymbol{\beta}}\right\|_0$	Elapsed Time (in sec.)
Ishigami	$\mathbb{L}_2$-boosting	19	0.0941
function	FoBa	21	2.2917
Case 1	LARS	50	53.03
Ishigami	$\mathbb{L}_2$-boosting	15	0.0884
function	FoBa	12	1.0752
Case 2	LARS	45	23.2062
g-Sobol function	$\mathbb{L}_2$-boosting	7.4	1.0620
	FoBa	4.7	2.9195
	LARS		10^3
Tank	$\mathbb{L}_2$-boosting	10	0.0266
pressure	FoBa	22	0.3741
model	LARS	10	0.1756

Table 4.2: Features of the three algorithms

It clearly appears in Table 4.2 that our proposed $\mathbb{L}_2$-boosting is the fastest method. Also, although we do not have access to the theoretical support recovery $\|\boldsymbol{\beta}\|_0$, we notice that the $\mathbb{L}_2$-boosting selects a small number of predictors, and yet performs quite well through the applications. This presumes that the $\mathbb{L}_2$-boosting is more accurate, as it seems to make a good support recovery. The FoBa procedure performances are also very good regarding their ability to obtain a sparse representation and the fraction of additional time required by this last algorithm in comparison with the $\mathbb{L}_2$-boosting oscillates between two and about ten, or so. At last, the LARS algorithm possesses a somewhat larger computational cost although its performances on our several data sets were quite disappointing.

Note that we have computed the maximal "degeneracy" which is involved in the resolution of the linear systems and quantified by Assumption ($\mathbf{H}_{\mathbf{b}}^{\mathbf{3},\vartheta}$) in the column 2 of Table 4.3. In many cases, we obtain a significantly larger value than 0. The third column of Table 4.3 shows the admissible size of the parameter ϑ and we can check that the number of variables p_n allowed by ($\mathbf{H}_{\mathbf{b}}^{\mathbf{2}}$) and the balance between ξ and ϑ (ξ should be greater than 2ϑ in our theoretical results) is not restrictive since $n^{1-2\vartheta}$ is always significantly greater than $log(m_n)$ in Table

4.3.

Data set	Degeneracy $d(A)$	$\vartheta \geq \frac{\log(1/d(A))}{log(n)}$	$n^{1-2\vartheta}$	$\log(m_n)$
Ishigami function Case1	0.6388	$[0.0786, +\infty[$	191.6106	6.0113
Ishigami function Case1	0.76	$[0.0481, +\infty[$	228.0194	6.0113
g-Sobol function	0.9410	$[0.0093, +\infty[$	619.6967	7.0690
Polynomial function	0.2736	$[0.2446, +\infty[$	14.9750	5.7991

Table 4.3: Degeneracy of the linear systems and admissible size of p_n

5. Conclusions and Perspectives

This paper brings a rigorous framework for the hierarchically orthogonal Gram-Schmidt procedure in a high-dimensional paradigm, when the greedy $\mathbb{L}_2$-boosting is used. It also appears that we obtain satisfying numerical results through our three Data sets with a very low computational cost. From a mathematical point of view, assumption $(\mathbf{H_b^1})$ presents a restrictive condition, and to relax it would open a wider class of basis functions for applications. We let this development open for a future work, which may rely either on a development of a concentration inequality for unbounded random matrices or on a truncating argument.

6 Appendix

6.1 Notation and reminder

Let us first recall some standard notation on matricial norms. For any square matrix M, its spectral radius $\rho(M)$ will refer to the largest absolute value of the elements of its spectrum:

$$\rho(M) := \max_{\alpha \in Sp(M)} |\alpha|.$$

Moreover, $\|M\|_2$ is the euclidean endomorphism norm and is given by

$$\|M\|_2 := \sqrt{\rho(M^t M)},$$

where M^t is the transpose of M. Note that for self-adjoint matrices, $\|M\|_2 = \rho(M)$. At last, the Frobenius norm of M is given by

$$\|M\|_F := \left(Tr(M^t M)\right)^{1/2}.$$

6.2 Hoeffding 's type Inequality for random bounded matrices For sake of completeness, we quote here Theorem 1.3 of Tropp (2012).

Theorem 3 (Matrix Hoeffding: bounded case)**.** *Consider a finite sequence $(X_k)_{1\leq k\leq n}$ of independent random self-adjoint matrices with dimension d, and let $(A_k)_{1\leq k\leq n}$ a deterministic sequence of self-adjoint matrices. Assume that*

$$\forall 1 \leq k \leq n \qquad \mathbb{E}X_k = 0 \qquad \textit{and} \qquad X_k^2 \preceq A_k^2 \quad a.s.$$

Then, for all $t \geq 0$

$$P\left(\lambda_{max}\left(\sum_{k=1}^n X_k\right) \geq t\right) \leq de^{-t^2/8\sigma^2}, \qquad \textit{where} \qquad \sigma^2 = \|\sum_{k=1}^n A_k^2\|.$$

In our work, it is useless to use a more precise concentration inequality such as the Bernstein one (see Theorem 6.1 of Tropp (2012)) since we do not consider any asymptotic on L (the number of basis functions for each variables X^j). Such asymptotic setting is far beyond the scope of the paper and we let this problem open for a future work.

6.3 Proof of Theorem 1

Consider any subset $u = (u_1, ..., u_t) \in S^*$ with $t \geq 1$ and remark that if $u = \{i\}$, i.e. $t = 1$, and $L \geq 1$, we have seen in the *Initialization* of Algorithm 1 that

$$\hat{\phi}^i_{l_i,n_1} = \phi^i_{l_i}, \quad \forall\ l_i \in [1:L],$$

Therefore, we obviously have that $\sup_{\substack{i\in[1:p] \\ l_i\in[1:L]}} \left\|\hat{\phi}^i_{l_i,n_1} - \phi^i_{l_i}\right\| = 0$.

Now, for $t = 2$, let $u = \{i, j\}$, with $i \neq j \in [1 : p]$, and $\boldsymbol{l_{ij}} = (l_i, l_j) \in [1 : L]^2$, remind that $\phi^{ij}_{\boldsymbol{l_{ij}}}$ is defined as:

$$\phi^{ij}_{\boldsymbol{l_{ij}}}(x_i, x_j) = \phi^i_{l_i}(x_i) \times \phi^j_{l_j}(x_j) + \sum_{k=1}^{L} \lambda^i_{k,\boldsymbol{l_{ij}}} \phi^i_k(x_i) + \sum_{k=1}^{L} \lambda^j_{k,\boldsymbol{l_{ij}}} \phi^j_k(x_j) + C_{\boldsymbol{l_{ij}}},$$

where $(C_{\boldsymbol{l_{ij}}}, (\lambda^i_{k,\boldsymbol{l_{ij}}})_k, (\lambda^j_{k,\boldsymbol{l_{ij}}})_k)$ are given as the solutions of:

$$\begin{array}{l} \langle \phi^{ij}_{\boldsymbol{l_{ij}}}, \phi^i_k \rangle = 0, \quad \forall\ k \in [1 : L] \\ \langle \phi^{ij}_{\boldsymbol{l_{ij}}}, \phi^j_k \rangle = 0, \quad \forall\ k \in [1 : L] \\ \langle \phi^{ij}_{\boldsymbol{l_{ij}}}, 1 \rangle = 0. \end{array} \tag{6.1}$$

When removing $C_{\boldsymbol{l_{ij}}}$, the resolution of (6.1) leads to the resolution of a linear system of the type:

$$A^{ij} \boldsymbol{\lambda}^{\boldsymbol{l_{ij}}} = D^{\boldsymbol{l_{ij}}}, \tag{6.2}$$

with $\boldsymbol{\lambda}^{\boldsymbol{l_{ij}}} = \left(\lambda^i_{1,\boldsymbol{l_{ij}}} \cdots \lambda^i_{L,\boldsymbol{l_{ij}}} \lambda^j_{1,\boldsymbol{l_{ij}}} \cdots \lambda^j_{L,\boldsymbol{l_{ij}}}\right)^t$ and

$$A^{ij} = \begin{pmatrix} B^{ii} & B^{ij} \\ {}^tB^{ij} & B^{jj} \end{pmatrix}, \quad B^{ij} = \begin{pmatrix} \langle \phi^i_1, \phi^j_1 \rangle & \cdots & \langle \phi^i_1, \phi^j_L \rangle \\ \vdots & & \\ \langle \phi^i_L, \phi^j_1 \rangle & \cdots & \langle \phi^i_L, \phi^j_L \rangle \end{pmatrix}, \quad D^{\boldsymbol{l_{ij}}} = -\begin{pmatrix} \langle \phi^i_{l_i} \times \phi^j_{l_j}, \phi^i_1 \rangle \\ \vdots \\ \langle \phi^i_{l_i} \times \phi^j_{l_j}, \phi^i_L \rangle \\ \langle \phi^i_{l_i} \times \phi^j_{l_j}, \phi^j_1 \rangle \\ \vdots \\ \langle \phi^i_{l_i} \times \phi^j_{l_j}, \phi^j_L \rangle \end{pmatrix}.$$

Consider now $\hat{\phi}^{ij}_{\boldsymbol{l_{ij}},n_1}$ which is decomposed on the dictionary as follows:

$$\hat{\phi}^{ij}_{\boldsymbol{l_{ij}},n_1}(x_i, x_j) \quad = \quad \phi^i_{l_i}(x_i) \times \phi^j_{l_j}(x_j) + \textstyle\sum_{k=1}^{L} \hat{\lambda}^i_{k,\boldsymbol{l_{ij}},n_1} \phi^i_k(x_i) + \textstyle\sum_{k=1}^{L} \hat{\lambda}^j_{k,\boldsymbol{l_{ij}},n_1} \phi^j_k(x_j) + \hat{C}^{n_1}_{\boldsymbol{l_{ij}}},$$

where $(\hat{C}^{n_1}_{\boldsymbol{l_{ij}}}, (\hat{\lambda}^i_{k,\boldsymbol{l_{ij}},n_1})_k, (\hat{\lambda}^j_{k,\boldsymbol{l_{ij}},n_1})_k)$ are given as solutions of the following *random* equalities:

$$\begin{array}{l} \langle \hat{\phi}^{ij}_{\boldsymbol{l_{ij}},n_1}, \phi^i_k \rangle_{n_1} = 0, \quad \forall\ k \in [1 : L] \\ \langle \hat{\phi}^{ij}_{\boldsymbol{l_{ij}},n_1}, \phi^j_k \rangle_{n_1} = 0, \quad \forall\ k \in [1 : L] \\ \langle \hat{\phi}^{ij}_{\boldsymbol{l_{ij}},n_1}, 1 \rangle_{n_1} = 0. \end{array} \tag{6.3}$$

When removing $\hat{C}^{n_1}_{\boldsymbol{l_{ij}}}$, the resolution of (6.3) can also lead to the resolution of a linear system of the type:

$$\hat{A}^{ij}_{n_1}\hat{\boldsymbol{\lambda}}^{\boldsymbol{l_{ij}}}_{n_1} = \hat{D}^{\boldsymbol{l_{ij}}}_{n_1}, \tag{6.4}$$

where $\hat{\boldsymbol{\lambda}}^{\boldsymbol{l_{ij}}}_{n_1} = \left(\hat{\lambda}^i_{1,\boldsymbol{l_{ij}},n_1} \cdots \hat{\lambda}^i_{L,\boldsymbol{l_{ij}},n_1} \hat{\lambda}^j_{1,\boldsymbol{l_{ij}},n_1} \cdots \hat{\lambda}^j_{L,\boldsymbol{l_{ij}},n_1}\right)^t$ and $\hat{A}^{ij}_{n_1}$ (*resp.* $\hat{D}^{\boldsymbol{l_{ij}}}_{n_1}$) are obtained from A^{ij} (*resp.* $D^{\boldsymbol{l_{ij}}}$) by changing the theoretical inner product by its empirical version.

Remark 4. *Remark that A^{ij} depends on (i,j) as well as $\boldsymbol{\lambda}^{\boldsymbol{l_{ij}}}$ and $D^{\boldsymbol{l_{ij}}}$ depend on (i,j) and $\boldsymbol{l_{ij}}$, but we will deliberately omit these indexes in the sequel for sake of convenience when no confusion is possible. For instance, when a couple (i,j) is handled, we will frequently use the notation $A, \boldsymbol{\lambda}, D, C, \lambda^i_k, \lambda^j_k$ instead of $A^{ij}, \boldsymbol{\lambda}^{\boldsymbol{l_{ij}}}, D^{\boldsymbol{l_{ij}}}, C_{\boldsymbol{l_{ij}}}, \lambda^i_{k,\boldsymbol{l_{ij}}}$ and $\lambda^j_{k,\boldsymbol{l_{ij}}}$. This will be also the case for the estimators $\hat{A}_{n_1}, \hat{\boldsymbol{\lambda}}_{n_1}, \hat{D}_{n_1}, \hat{C}^{n_1}, \hat{\lambda}^i_{k,n_1}$ and $\hat{\lambda}^j_{k,n_1}$.*

Then, the following useful lemma compares the two matrices $\hat{A}_{n_1}$ and A.

Lemma 1. *Under Assumption ($\mathbf{H_b}$), and for any ξ given by ($\mathbf{H^2_b}$), one has*

$$\sup_{1\leq i,j\leq p_n} \left|\!\left|\!\left| \hat{A}_{n_1} - A \right|\!\right|\!\right|_2 = \mathcal{O}_P(n^{-\xi/2}).$$

Proof. First consider one couple (i,j) and note that $\left|\!\left|\!\left| \hat{A}_{n_1} - A \right|\!\right|\!\right|_2 = \rho(\hat{A}_{n_1} - A)$, since $\hat{A}_{n_1} - A$ is self-adjoint. To obtain a concentration inequality on the matricial norm $\left|\!\left|\!\left| \hat{A}_{n_1} - A \right|\!\right|\!\right|_2$, we mainly use the results of Tropp (2012), which give concentration inequalities for the largest eigenvalue of self-adjoint matrices (see section). Denote $\preceq$ the semi-definite order on self-adjoint matrices, which is defined for all self-adjoint matrices M_1 and M_2 of size q as:

$$M_1 \preceq M_2 \text{ iff } \forall u \in \mathbb{R}^q, \quad u^t M_1 u \leq u^t M_2 u.$$

Remark that $\hat{A}_{n_1} - A$ could be written as follows:

$$\hat{A}_{n_1} - A = \frac{1}{n_1}\sum_{r=1}^{n_1} \Theta_{r,ij}, \quad \Theta_{r,ij} = \begin{pmatrix} \Theta^{ii}_r & \Theta^{ij}_r \\ {}^t\Theta^{ij}_r & \Theta^{jj}_r \end{pmatrix}, \quad \forall\ r \in [1:n_1],$$

where, for all $k,m \in [1:L]$, $(\Theta^{i_1 i_2}_r)_{k,m} = \phi^{i_1}_k(x^r_{i_1})\phi^{i_2}_m(x^r_{i_2}) - \mathbb{E}[\phi^{i_1}_k(X_{i_1})\phi^{i_2}_m(X_{i_2})]$ with $i_1, i_2 \in \{i,j\}$. Since the observations $(\mathbf{x}^r)_{r=1,\cdots,n_1}$ are supposed to be independent, $\Theta_{1,ij}, \cdots, \Theta_{n_1,ij}$ is a sequence of independent, random, centered, self-adjoint matrices. Moreover, for all $u \in \mathbb{R}^{2L}$, all $r \in [1 : n_1]$,

$$u^t \Theta_{r,ij}^2 u = \|\Theta_{r,ij} u\|_2^2 \leq \|u\|_2^2 \, \|\Theta_{r,ij}\|_F^2,$$

where

$$\begin{aligned} \|\Theta_{r,ij}\|_F^2 &\leq (2L)^2 \left(\max_{k,m\in[1:L]} |(\Theta_{r,ij})_{k,m}|\right)^2 \\ &\leq (2L)^2 \left(\max_{\substack{k,m\in[1:L] \\ i_1,i_2\in\{i,j\}}} |\phi_k^{i_1}(x_{i_1}^r)\phi_m^{i_2}(x_{i_2}^r) - \mathbb{E}[\phi_k^{i_1}(X_{i_1})\phi_m^{i_2}(X_{i_2})]| \right)^2 \\ &\leq 16L^2M^4 \quad \text{by } (\mathbf{H_b^1}). \end{aligned}$$

We then deduce that each element of the sum satisfies $X_{l,ij}^2 \preceq 16L^2M^4 \mathrm{I}_{L^2}$, where I_{L^2} denotes the identity matrix of size L^2.

Applying now the Hoeffding's type Inequality stated in Theorem 1.3 of Tropp (2012) to our sequence $\Theta_{1,ij}, \cdots, \Theta_{n_1,ij}$, with $\sigma^2 = 16n_1L^2M^4$, we then obtain that

$$\forall t \geq 0 \qquad P\left(\rho\left(\frac{1}{n_1}\sum_{r=1}^{n_1}\Theta_{r,ij}\right) \geq t\right) \leq 2Le^{-\frac{(n_1t)^2}{8\sigma^2}},$$

Considering now the whole set of estimators $\hat{A}_{n_1}$, we obtain

$$\forall t \geq 0 \qquad P\left(\sup_{1\leq i,j\leq p_n} \rho\left(\frac{1}{n_1}\sum_{r=1}^{n_1}\Theta_{r,ij}\right) \geq t\right) \leq 2Lp_n^2e^{-\frac{(n_1t)^2}{8\sigma^2}},$$

Now, we take $t = \gamma n^{-\xi/2}$, where $\gamma > 0$, and $0 < \xi \leq 1$ given in $(\mathbf{H_b^2})$. Then, the following inequality holds:

$$P\left(\sup_{1\leq i,j\leq p_n} \rho\left(\hat{A}_{n_1} - A\right) \geq \gamma n^{-\xi/2}\right) \leq 2Lp_n^2 e^{-\frac{{n_1}^{1-\xi}\gamma^2}{128L^2M^4}}. \tag{6.5}$$

Since $n_1 = n/2$, and $p_n = \underset{n\to+\infty}{\mathcal{O}}(\exp(Cn^{1-\xi}))$ by Assumption $(\mathbf{H_b^2})$, the right-hand side of the previous inequality becomes arbitrarily small for n sufficiently large and $\gamma > 0$ large enough. The end of the proof follows using Inequality (6.5). □

Similarly, we can show that the estimated quantity $\hat{D}_{n_1}$ is not so far from the theoretical D with high probability.

Lemma 2. *Under Assumptions* $(\mathbf{H_b})$, *and for any* ξ *given by* $(\mathbf{H_b^2})$, *one has*

$$\sup_{i,j,\boldsymbol{l_{ij}}} \left\|\hat{D}_{n_1} - D\right\|_2 = \mathcal{O}_P(n^{-\xi/2}).$$

Proof. First consider one couple (i,j). We aim to apply another concentration inequality on $\left\|\hat{D}_{n_1}^{\boldsymbol{l_{ij}}} - D^{\boldsymbol{l_{ij}}}\right\|_2$. Remark that $\left\|\hat{D}_{n_1} - D\right\|_2$ can be written as:

$$\begin{aligned}
\left\|\hat{D}_{n_1} - D\right\|_2 &= \Big(\textstyle\sum_{k=1}^{L}\Big(\langle\phi_{l_i}^i \times \phi_{l_j}^j, \phi_k^i\rangle_{n_1} - \langle\phi_{l_i}^i \times \phi_{l_j}^j, \phi_k^i\rangle\Big)^2 + \\
&\quad \textstyle\sum_{k=1}^{L}\Big(\langle\phi_{l_i}^i \times \phi_{l_j}^j, \phi_k^j\rangle_{n_1} - \langle\phi_{l_i}^i \times \phi_{l_j}^j, \phi_k^j\rangle\Big)^2\Big)^{1/2} \\
&\le \textstyle\sum_{k=1}^{L}\left|\frac{1}{n_1}\sum_{r=1}^{n_1}\phi_{l_i}^i({x_i}^r)\phi_{l_j}^j({x_j}^r)\phi_k^i({x_i}^r) - \langle\phi_{l_i}^i \times \phi_{l_j}^j, \phi_k^i\rangle\right| + \\
&\quad \textstyle\sum_{k=1}^{L}\left|\frac{1}{n_1}\sum_{r=1}^{n_1}\phi_{l_i}^i({x_i}^r)\phi_{l_j}^j({x_j}^r)\phi_k^j({x_j}^r) - \langle\phi_{l_i}^i \times \phi_{l_j}^j, \phi_k^j\rangle\right|.
\end{aligned}$$

Now, Bernstein's Inequality (see Birgé and Massart (1998) for instance) implies that, for all $\gamma > 0$,

$$\begin{aligned}
P\left(n_1^{\xi/2}\left\|\hat{D}_{n_1} - D\right\|_2 \ge \gamma\right) &\le P\left(n_1^{\xi/2}\textstyle\sum_{k=1}^{L}\left|\frac{1}{n_1}\sum_{r=1}^{n_1}\phi_{l_i}^i(x_i^r)\phi_{l_j}^j(x_j^r)\phi_k^i(x_i^r) - \langle\phi_{l_i}^i \times \phi_{l_j}^j, \phi_k^i\rangle\right| > \gamma/2\right) \\
&+ P\left(n_1^{\xi/2}\textstyle\sum_{k=1}^{L}\left|\frac{1}{n_1}\sum_{r=1}^{n_1}\phi_{l_i}^i(x_i^r)\phi_{l_j}^j(x_j^r)\phi_k^i(x_i^r) - \langle\phi_{l_i}^i \times \phi_{l_j}^j, \phi_k^i\rangle\right| > \gamma/2\right) \\
&\le 4L\exp\left(-\frac{1}{8}\frac{\gamma^2 n_1^{1-\xi}}{M^6 + M^3\gamma/6n_1^{-\xi/2}}\right),
\end{aligned}$$

which gives:

$$P\left(\sup_{i,j,\boldsymbol{l_{ij}}}\left\|\hat{D}_{n_1} - D\right\|_2 \ge \gamma n_1^{-\xi/2}\right) \le 4L \times L^2 p_n^2 \exp\left(-\frac{1}{8}\frac{\gamma^2 {n_1}^{1-\xi}}{M^6 + M^3\gamma/6{n_1}^{-\xi/2}}\right). \tag{6.6}$$

Now, since $n_1 = n/2$, Assumption $(\mathbf{H_b^2})$ implies that the right-hand side of Inequality (6.6) can also become arbitrarily small for n sufficiently large, which concludes the proof. □

The next lemma then compares the estimated $\hat{\boldsymbol{\lambda}}_{n_1}$ with $\boldsymbol{\lambda}$.

Lemma 3. *Under Assumptions* $(\mathbf{H_b})$, *we have when* $\vartheta < \xi/2$,

$$\sup_{i,j,\boldsymbol{l_{ij}}}\left\|\hat{\boldsymbol{\lambda}}_{n_1} - \boldsymbol{\lambda}\right\|_2 = \mathcal{O}_P(n^{\vartheta-\xi/2}).$$

Proof. Fix any couple (i,j), $\boldsymbol{\lambda}$ and $\hat{\boldsymbol{\lambda}}_{n_1}$ satisfy Equations (6.2) and (6.4). Hence,

$$\begin{array}{rcl} A(\hat{\boldsymbol{\lambda}}_{n_1}-\boldsymbol{\lambda})-A\hat{\boldsymbol{\lambda}}_{n_1} & = & -D=\hat{D}_{n_1}-D-\hat{D}_{n_1} \\ & = & (\hat{D}_{n_1}-D)-\hat{A}_{n_1}\hat{\boldsymbol{\lambda}}_{n_1} \\ \Leftrightarrow \qquad A(\hat{\boldsymbol{\lambda}}_{n_1}-\boldsymbol{\lambda}) & = & (\hat{D}_{n_1}-D)+(A-\hat{A}_{n_1})\hat{\boldsymbol{\lambda}}_{n_1} \\ \Leftrightarrow \qquad \hat{\boldsymbol{\lambda}}_{n_1}-\boldsymbol{\lambda} & = & A^{-1}[(A-\hat{A}_{n_1})\hat{\boldsymbol{\lambda}}_{n_1}]+A^{-1}(\hat{D}_{n_1}-D), \end{array}$$

since the matrix A is positive definite. It follows that

$$\hat{\boldsymbol{\lambda}}_{n_1}-\boldsymbol{\lambda}=A^{-1}(A-\hat{A}_{n_1})(\hat{\boldsymbol{\lambda}}_{n_1}-\boldsymbol{\lambda})+A^{-1}(A-\hat{A}_{n_1})\boldsymbol{\lambda}+A^{-1}(\hat{D}_{n_1}-D),$$

and

$$\left(\mathrm{I}-A^{-1}(A-\hat{A}_{n_1})\right)(\hat{\boldsymbol{\lambda}}_{n_1}-\boldsymbol{\lambda})=A^{-1}(A-\hat{A}_{n_1})\boldsymbol{\lambda}+A^{-1}(\hat{D}_{n_1}-D), \qquad (6.7)$$

Remark that $\left|\!\left|\!\left|\hat{A}_{n_1}-A\right|\!\right|\!\right|_2=\mathcal{O}_P(n^{-\xi/2})$ by Lemma 1. Hence, with high probability and for n large enough $\mathrm{I}-A^{-1}(A-\hat{A}_{n_1})$ is invertible, and Inequality (6.7) can be rewritten as:

$$\hat{\boldsymbol{\lambda}}_{n_1}-\boldsymbol{\lambda}=\left(\mathrm{I}-A^{-1}(A-\hat{A}_{n_1})\right)^{-1}\left(A^{-1}(A-\hat{A}_{n_1})\boldsymbol{\lambda}+A^{-1}(\hat{D}_{n_1}-D)\right).$$

We then deduce that,

$$\begin{array}{rcl} \left\|\hat{\boldsymbol{\lambda}}_{n_1}-\boldsymbol{\lambda}\right\|_2 & \leq & \left|\!\left|\!\left|\left(\mathrm{I}-A^{-1}(A-\hat{A}_{n_1})\right)^{-1}\right|\!\right|\!\right|_2 \\ & & \times\left(\left|\!\left|\!\left|A^{-1}[A-\hat{A}_{n_1}]\right|\!\right|\!\right|_2\|\boldsymbol{\lambda}\|_2+\left\|A^{-1}(\hat{D}_{n_1}-D)\right\|_2\right) \\ & \leq & \left|\!\left|\!\left|\left(\mathrm{I}-A^{-1}(A-\hat{A}_{n_1})\right)^{-1}\right|\!\right|\!\right|_2 \\ & & \times\left(|\!|\!|A^{-1}|\!|\!|_2\left|\!\left|\!\left|A-\hat{A}_{n_1}\right|\!\right|\!\right|_2\|\boldsymbol{\lambda}\|_2+|\!|\!|A^{-1}|\!|\!|_2\left\|\hat{D}_{n_1}-D\right\|_2\right). \end{array} \qquad (6.8)$$

A uniform bound for $|\!|\!|A^{-1}|\!|\!|_2$ (over all couples (i,j)) can be easily obtain since A (and obviously A^{-1}) is Hermitian.

$$|\!|\!|A^{-1}|\!|\!|_2\leq\max_{(i',j')\in[1:p_n]^2}\rho\left(\left(A^{i'j'}\right)^{-1}\right)$$

Simple algebra then yields

$$\rho\left(\left(A^{i'j'}\right)^{-1}\right)\leq Tr\left(\left(A^{i'j'}\right)^{-1}\right)=\frac{Tr\left(Com(A^{i'j'})^t\right)}{det(A^{i'j'})}=\frac{1}{det(A^{i'j'})}\sum_{k=1:2L}Com(A^{i'j'})_{k,k}$$

where $Com(A^{ij})$ is the cofactor matrix associated to A^{ij}. Now, recall the classical inequality (that can be found in Bullen (1998)): for any symetric definite positive matrix squared S of size $Q \times Q$

$$\det(S) \leq \prod_{\ell=1}^{Q} |S_{\ell\ell}|.$$

This last inequality applied to the determinant involved in $Com(A^{i'j'})_{k,k}$ associated with $(\mathbf{H_b^1})$ implies

$$\forall k \in [1:2L] \qquad \left|Com(A^{i'j'})_{k,k}\right| \leq \{M^2\}^{2L-1}.$$

We then deduce from $(\mathbf{H_b^{3,\vartheta}})$ that there exists a constant $C > 0$ such that:

$$\begin{aligned} \left\|\left\|A^{-1}\right\|\right\|_2 &\leq \max_{(i,j)\in[1:p_n]^2} \tfrac{2LM^{4L-2}}{det(A^{i'j'})} \\ &\leq 2C^{-1}LM^{4L-2}n^{\vartheta}. \end{aligned} \tag{6.9}$$

Similarly, if we denote $\Delta_{n_1} = A - \hat{A}_{n_1}$, we have

$$\begin{aligned} \left\|\left\|\left(\mathrm{I} - A^{-1}(A - \hat{A}_{n_1})\right)^{-1}\right\|\right\|_2 &= \rho\left(\left(I - A^{-1}\Delta_{n_1}\right)^{-1}\right) \\ &= \max_{\alpha\in Sp(A^{-1}\Delta_{n_1})} \frac{1}{|1-\alpha|}, \end{aligned}$$

using the fact that $A - \hat{A}_{n_1}$ is self-adjoint. We have seen that $\rho(A^{-1}) \leq 2C^{-1}LM^{4L-2}n^{\vartheta}$ and Lemma 1 yields $\rho\left(\Delta_{n_1}\right) = \mathcal{O}_P(n^{-\xi/2})$. As a consequence, we have

$$\max_{\alpha\in Sp(A^{-1}\Delta_{n_1})} |\alpha| \leq \rho(A^{-1})\rho\left(\Delta_{n_1}\right) = \mathcal{O}_P(n^{\vartheta-\xi/2}).$$

At last, remark that

$$\max_{\alpha\in Sp(A^{-1}\Delta_{n_1})} \frac{1}{|1-\alpha|} - 1 = \max_{\alpha\in Sp(A^{-1}\Delta_{n_1})} \frac{1-|1-\alpha|}{|1-\alpha|}$$

We know that for n large enough, each absolute value of $\alpha \in Sp(A^{-1}\Delta_{n_1})$ becomes smaller than 1/2 with a probability tending to one. Hence, we have with probability tending to one

$$\max_{\alpha\in Sp(A^{-1}\Delta_{n_1})} \left|\frac{1-|1-\alpha|}{|1-\alpha|}\right| \leq \max_{\alpha\in Sp(A^{-1}\Delta_{n_1})} \frac{|\alpha|}{1-\alpha} \leq 2\rho(A^{-1}\Delta_{n_1}).$$

Since $\rho(A^{-1}\Delta_{n_1}) = \mathcal{O}_P(n^{\vartheta-\xi/2})$, we deduce

$$\sup_{i,j,\boldsymbol{l_{ij}}} \left\| \left(\mathrm{I} - A^{-1}(A - \hat{A}_{n_1}) \right)^{-1} \right\|_2 \leq 1 + 2LM^{4L-2}C^{-1}\mathcal{O}_P(n^{\vartheta-\xi/2}). \tag{6.10}$$

To conclude the proof, we can now apply the same argument as the one used in Lemmas 1 and 2 with Bernstein's Inequality, using Equations (6.9) and (6.10). □

The last lemma finally compares the constant $\hat{C}^{n_1}$ with C.

Lemma 4. *Under Assumptions* $(\mathbf{H_b})$*, we have:*

$$\sup_{i,j,\boldsymbol{l_{ij}}} \left| \hat{C}^{n_1} - C \right| = \mathcal{O}_P(n^{-\xi/2}).$$

Proof. For any couple (i,j), remark that constants $\hat{C}^{n_1}$ and C satisfy:

$$C = -\langle \phi^i_{l_i} \times \phi^j_{l_j}, 1 \rangle \quad \text{and} \quad \hat{C}^{n_1} = -\langle \phi^i_{l_i} \times \phi^j_{l_j}, 1 \rangle_{n_1}.$$

If we denote

$$\Delta_{i,j,\boldsymbol{l_{ij}}} := \frac{1}{n_1} \sum_{r=1}^{n_1} \phi^i_{l_i}({x_i}^r)\phi^j_{l_j}({x_j}^r) - \mathbb{E}(\phi^i_{l_i}(X_i)\phi^j_{l_j}(X_j)),$$

we can apply again Bernstein's Inequality on $(\phi^i_{l_i}({x_i}^r)\phi^j_{l_j}({x_j}^r))_{r=1,\cdots,n_1}$. From $(\mathbf{H_b^1})$, these independent random variables are bounded by M^2 and

$$\begin{aligned} P\left(\sup_{i,j,\boldsymbol{l_{ij}}} \left| \Delta_{i,j,\boldsymbol{l_{ij}}} \right| \geq \gamma n_1^{-\xi/2} \right) &\leq \sum_{i,j,\boldsymbol{l_{ij}}} P\left(\left| \Delta_{i,j,\boldsymbol{l_{ij}}} \right| \geq \gamma n_1^{-\xi/2} \right) \\ &\leq \sum_{i,j,\boldsymbol{l_{ij}}} 2\exp\left(-\frac{1}{2} \frac{\gamma^2 n_1^{1-\xi}}{M^4 + M^2\gamma/3n_1^{-\xi/2}} \right) \\ &\leq 2L^2p_n^2 \exp\left(-\frac{1}{2} \frac{\gamma^2 n_1^{1-\xi}}{M^4 + M^2\gamma/3n_1^{-\xi/2}} \right). \end{aligned}$$

Under Assumption $(\mathbf{H_b^2})$, the right-hand side of this inequality can be arbitrarly small for n large enough, which ends the proof. □

To finish the proof of Theorem 1, remark that:

$$\begin{aligned}
\left\|\hat{\phi}^{ij}_{\boldsymbol{l_{ij}},n_1} - \phi^{ij}_{\boldsymbol{l_{ij}}}\right\| &= \left\|\textstyle\sum_{k=1}^{L}(\hat{\lambda}^i_{k,n_1} - \lambda^i_k)\phi^i_k + \sum_{k=1}^{L}(\hat{\lambda}^j_{k,n_1} - \lambda^j_k)\phi^j_k + (\hat{C}^{n_1} - C)\right\| \\
&\leq \underbrace{\left\|\sum_{k=1}^{L}(\hat{\lambda}^i_{k,n_1} - \lambda^i_k)\phi^i_k + \sum_{k=1}^{L}(\hat{\lambda}^j_{k,n_1} - \lambda^j_k)\phi^j_k\right\|}_{I} + \left|\hat{C}^{n_1} - C\right|.
\end{aligned}$$

Moreover,

$$\begin{aligned}
I^2 &= \textstyle\int \left(\sum_{k=1}^{L}(\hat{\lambda}^i_{k,n_1} - \lambda^i_k)\phi^i_k + \sum_{k=1}^{L}(\hat{\lambda}^j_{k,n_1} - \lambda^j_k)\phi^j_k\right)^2 p_{X_i,X_j}(x_i,x_j)dx_idx_j \\
&= \underbrace{\int\left(\sum_{k=1}^{L}(\hat{\lambda}^i_{k,n_1} - \lambda^i_k)\phi^i_k\right)^2 p_{X_i}(x_i)dx_i}_{I_1} + \underbrace{\int\left(\sum_{k=1}^{L}(\hat{\lambda}^j_{k,n_1} - \lambda^j_k)\phi^j_k\right)^2 p_{X_j}(x_j)dx_j}_{I_2} \\
&\quad \underbrace{+2\int\left(\sum_{k=1}^{L}(\hat{\lambda}^i_{k,n_1} - \lambda^i_k)\phi^i_k\right)\left(\sum_{k=1}^{L}(\hat{\lambda}^i_{k,n_1} - \lambda^i_k)\phi^i_k\right) p_{X_i,X_j}(x_i,x_j)dx_idx_j}_{I_3}.
\end{aligned}$$

Using the inequality $2ab \leq a^2 + b^2$, we thus deduce that $I_3 \leq I_1 + I_2$, and

$$\begin{aligned}
I_1 &= \textstyle\int \sum_{k=1}^{L}\sum_{m=1}^{L}(\hat{\lambda}^i_{k,n_1} - \lambda^i_k)(\hat{\lambda}^i_{m,n_1} - \lambda^i_m)\phi^i_k(x_i)\phi^i_m(x_i)p_{X_i}(x_i)dx_i \\
&= \textstyle\sum_{k=1}^{L}(\hat{\lambda}^i_{k,n_1} - \lambda^i_k)^2 \quad \text{by orthonormality.}
\end{aligned}$$

And the same equality is satisfied for I_2: $I_2 = \sum_{k=1}^{L}(\hat{\lambda}^j_{k,n_1} - \lambda^j_k)^2$. Consequently, we obtain

$$\begin{aligned}
\left\|\hat{\phi}^{ij}_{\boldsymbol{l_{ij}},n_1} - \phi^{ij}_{\boldsymbol{l_{ij}}}\right\| &\leq \sqrt{2\left[\textstyle\sum_{k=1}^{L}(\hat{\lambda}^i_{k,n_1} - \lambda^i_k)^2 + \sum_{k=1}^{L}(\hat{\lambda}^j_{k,n_1} - \lambda^j_k)^2\right]} + \left|\hat{C}^{n_1} - C\right| \\
&= \sqrt{2}\left\|\hat{\boldsymbol{\lambda}}_{n_1} - \boldsymbol{\lambda}\right\|_2 + \left|\hat{C}^{n_1} - C\right|.
\end{aligned} \tag{6.11}$$

The end of the proof follows with Lemmas 3 and 4.

□

6.4 Proof of Theorem 2

We recall first that $\langle,\rangle$ denotes the theoretical inner product based on the law $P_{\mathbf{X}}$ (and $\|\|$ is the derived Hilbertian norm). A careful inspection of the

Gram-Schmidt procedure used to build the HOFD shows that

$$M^* := \sup_{u,\boldsymbol{l_u}} \left\| \phi^u_{\boldsymbol{l_u}}(\mathbf{X}_u) \right\|_\infty < \infty,$$

provided that $(\mathbf{H_b^1})$ holds.

Now, remark that the EHOFD is obtained through the first sample $\mathcal{O}_1$ which determines the first empirical inner product $\langle , \rangle_{n_1}$ although the $\mathbb{L}^2$-boosting depends on the second sample $\mathcal{O}_2$. Indeed, $\mathcal{O}_2$ determines the second empirical inner product $\langle , \rangle_{n_2}$. Hence, $\langle , \rangle_{n_2}$ uses observations which are *independent* to the ones used to build the HOFD.

We begin this section with a lemma which establishes that the estimated functions $\hat{\phi}^u_{\boldsymbol{l_u},n_1}$ (which result in the EHOFD) are bounded.

Lemma 5. *Under Assumption* $(\mathbf{H_b})$*, define*

$$N_{n_1} := \sup_{u,\boldsymbol{l_u}} \left\| \hat{\phi}^u_{\boldsymbol{l_u},n_1}(\mathbf{X}_u) \right\|_\infty .$$

Then, we have:

$$N_{n_1} - M^* = \mathcal{O}_P(n^{\vartheta - \xi/2}).$$

Proof. Using the decomposition of $\hat{\phi}^u_{\boldsymbol{l_u},n_1}$ on the dictionary, Assumption $(\mathbf{H_b^2})$ and Cauchy-Schwarz Inequality, there exists a fixed constant $C > 0$ such that for all $u \in S$, $\boldsymbol{l_u}$:

$$\forall x \in \mathbb{R}^p \qquad |\hat{\phi}^u_{\boldsymbol{l_u},n_1}(x) - \phi^u_{\boldsymbol{l_u}}(x)| \le CM\sqrt{L}\sqrt{\left\| \hat{\boldsymbol{\lambda}}_{n_1} - \boldsymbol{\lambda} \right\|_2 + \left\| \hat{C}^{n_1}_{\boldsymbol{l_u}} - C_{\boldsymbol{l_u}} \right\|}.$$

The conclusion then follows using Lemmas 3 and 4. □

We now present a key lemma which compares the elements $(\phi^u_{\boldsymbol{l_u}})_{\boldsymbol{l_u},u}$ with its estimated version $(\hat{\phi}^u_{\boldsymbol{l_u},n_1})_{\boldsymbol{l_u},u}$.

Lemma 6. *Assume that* $(\mathbf{H_b})$ *holds with* $\xi \in (0,1)$*, that the noise* ε *satisfies* $(\mathbf{H_{\varepsilon,q}})$ *with* $q > 4/\xi$ *and that* $(\mathbf{H_s})$ *is fullfilled. Then, the following equalities hold,*

(i)

$$\sup_{u,v,\boldsymbol{l_u},\boldsymbol{l_v}} |\langle \hat{\phi}^u_{\boldsymbol{l_u},n_1}, \hat{\phi}^v_{\boldsymbol{l_v},n_1} \rangle - \langle \phi^u_{\boldsymbol{l_u}}, \phi^v_{\boldsymbol{l_v}} \rangle| = \zeta_{n,1} = \mathcal{O}_P(n^{\vartheta - \xi/2})$$

(ii)

$$\sup_{u,v,\boldsymbol{l_u},\boldsymbol{l_v}} |\langle \hat{\phi}^u_{\boldsymbol{l_u},n_1}, \hat{\phi}^v_{\boldsymbol{l_v},n_1}\rangle_{n_2} - \langle \phi^u_{\boldsymbol{l_u}}, \phi^v_{\boldsymbol{l_v}}\rangle| = \zeta_{n,2} = \mathcal{O}_P(n^{\vartheta-\xi/2})$$

(iii)

$$\sup_{u,v,\boldsymbol{l_u},\boldsymbol{l_v}} |\langle \varepsilon, \hat{\phi}^u_{\boldsymbol{l_u},n_1}\rangle_{n_2}| = \zeta_{n,3} = \mathcal{O}_P(n^{-\xi/2})$$

(iv)

$$\sup_{u,\boldsymbol{l_u}} \left|\langle \tilde{f}, \hat{\phi}^u_{\boldsymbol{l_u},n_1}\rangle_{n_2} - \langle \tilde{f}, \hat{\phi}^u_{\boldsymbol{l_u},n_1}\rangle\right| = \zeta_{n,4} = \mathcal{O}_P(n^{-\xi/2})$$

In the sequel, we will denote $\zeta_n := \max_{i\in[0:4]}\{\zeta_{n,i}\}$.

Proof. **Assertion** (i) Let $u, v \in S$, $\boldsymbol{l_u} \in [1:L]^{|u|}$ and $\boldsymbol{l_v} \in [1:L]^{|v|}$. Then, we have

$$\begin{aligned} \left|\langle \hat{\phi}^u_{\boldsymbol{l_u},n_1}, \hat{\phi}^v_{\boldsymbol{l_v},n_1}\rangle - \langle \phi^u_{\boldsymbol{l_u}}, \phi^v_{\boldsymbol{l_v}}\rangle\right| &\leq \left|\langle \hat{\phi}^u_{\boldsymbol{l_u},n_1} - \phi^u_{\boldsymbol{l_u}}, \hat{\phi}^v_{\boldsymbol{l_v},n_1}\rangle - \langle \phi^u_{\boldsymbol{l_u}}, \phi^v_{\boldsymbol{l_v}} - \hat{\phi}^v_{\boldsymbol{l_v},n_1}\rangle\right| \\ &\leq \left\|\hat{\phi}^u_{\boldsymbol{l_u},n_1} - \phi^u_{\boldsymbol{l_u}}\right\| \left\|\hat{\phi}^v_{\boldsymbol{l_v},n_1}\right\| + \left\|\phi^u_{\boldsymbol{l_u}}\right\| \left\|\hat{\phi}^v_{\boldsymbol{l_v},n_1} - \phi^v_{\boldsymbol{l_v}}\right\| \\ &\leq \left\|\hat{\phi}^u_{\boldsymbol{l_u},n_1} - \phi^u_{\boldsymbol{l_u}}\right\| \left(\left\|\hat{\phi}^v_{\boldsymbol{l_v},n_1} - \phi^v_{\boldsymbol{l_v}}\right\| + 1\right) + \left\|\hat{\phi}^v_{\boldsymbol{l_v},n_1} - \phi^v_{\boldsymbol{l_v}}\right\|, \end{aligned}$$

and the conclusion holds applying Theorem 1.

Assertion (ii) We breakdown it in two parts:

$$\begin{aligned} \left|\langle \hat{\phi}^u_{\boldsymbol{l_u},n_1}, \hat{\phi}^v_{\boldsymbol{l_v},n_1}\rangle_{n_2} - \langle \phi^u_{\boldsymbol{l_u}}, \phi^v_{\boldsymbol{l_v}}\rangle\right| \leq\ & \underbrace{\left|\langle \hat{\phi}^u_{\boldsymbol{l_u},n_1}, \hat{\phi}^v_{\boldsymbol{l_v},n_1}\rangle_{n_2} - \langle \hat{\phi}^u_{\boldsymbol{l_u},n_1}, \hat{\phi}^v_{\boldsymbol{l_v},n_1}\rangle\right|}_{I} \\ & + \underbrace{\left|\langle \hat{\phi}^u_{\boldsymbol{l_u},n_1}, \hat{\phi}^v_{\boldsymbol{l_v},n_1}\rangle - \langle \phi^u_{\boldsymbol{l_u}}, \phi^v_{\boldsymbol{l_v}}\rangle\right|}_{II}. \end{aligned}$$

Assertion (i) implies that,

$$\sup_{u,v,\boldsymbol{l_u},\boldsymbol{l_v}} |II| = \mathcal{O}_P(n^{\vartheta-\xi/2}).$$

To control $\sup\limits_{u,v,\boldsymbol{l_u},\boldsymbol{l_v}} |I|$, we use Bernstein's inequality to the family of independent random variables $\left(\hat{\phi}^u_{\boldsymbol{l_u},n_1}(\mathbf{x}^s_u)\hat{\phi}^v_{\boldsymbol{l_v},n_1}(\mathbf{x}^s_v)\right)_{s=1\dots n_2}$ and we denote

$$\Delta_{u,v,\boldsymbol{l_u},\boldsymbol{l_v}} = \left|\frac{1}{n_2}\sum_{s=1}^{n_2} \hat{\phi}^u_{\boldsymbol{l_u},n_1}(\mathbf{x}^s_u)\hat{\phi}^v_{\boldsymbol{l_v},n_1}(\mathbf{x}^s_v) - \mathbb{E}(\hat{\phi}^u_{\boldsymbol{l_u},n_1}(\mathbf{X}_u)\hat{\phi}^v_{\boldsymbol{l_v},n_1}(\mathbf{X}_v))\right|$$

Then, Bernstein's inequality implies that

$$\begin{aligned} P\left(\sup_{u,v,\boldsymbol{l_u},\boldsymbol{l_v}} \Delta_{u,v,\boldsymbol{l_u},\boldsymbol{l_v}} \geq \gamma n_2^{-\xi/2}\right) &\leq P\left(\sup_{u,v,\boldsymbol{l_u},\boldsymbol{l_v}} \Delta_{u,v,\boldsymbol{l_u},\boldsymbol{l_v}} \geq \gamma n_2^{-\xi/2} \& N_{n_1} < M^* + 1\right) \\ &+P\left(\sup_{u,v,\boldsymbol{l_u},\boldsymbol{l_v}} \Delta_{u,v,\boldsymbol{l_u},\boldsymbol{l_v}} \geq \gamma n_2^{-\xi/2} \& N_{n_1} > M^* + 1\right) \\ &\leq 64L^4 p_n^4 \exp\left(-\frac{1}{2}\frac{\gamma^2 {n_2}^{1-\xi}}{(M^*+1)^4 + (M^*+1)^2\gamma/3{n_2}^{-\xi/2})}\right) \\ &+P\left(N_{n_1} > M^* + 1\right) \end{aligned}$$

Lemma 5 and Assumption $(\mathbf{H_b^2})$ allows for deducing (ii).

Assertion (iii) The proof follows the roadmap of (ii) of Lemma 1 of Bühlmann (2006). We thus define the truncated variable ε_t for all $s \in [1 : n_2]$,

$$\varepsilon_t^s = \begin{cases} \varepsilon^s & \text{if } |\varepsilon^s| \leq K_n \\ sg(\varepsilon^s)K_n & \text{if } |\varepsilon^s| > K_n \end{cases}$$

where $sg(\varepsilon)$ denotes the sign of ε. Then, for $\gamma > 0$, we have:

$$\begin{aligned} P\left(n_2^{\xi/2}\sup_{u,\boldsymbol{l_u}}\left|\langle\hat{\phi}^u_{\boldsymbol{l_u},n_1}, \varepsilon\rangle_{n_2}\right| > \gamma\right) &\leq P\left(n_2^{\xi/2}\sup_{u,\boldsymbol{l_u}}\left|\langle\hat{\phi}^u_{\boldsymbol{l_u},n_1}, \varepsilon_t\rangle_{n_2} - \langle\hat{\phi}^u_{\boldsymbol{l_u},n_1}, \varepsilon_t\rangle\right| > \gamma/3\right) \\ &+P\left(n_2^{\xi/2}\sup_{u,\boldsymbol{l_u}}\left|\langle\hat{\phi}^u_{\boldsymbol{l_u},n_1}, \varepsilon - \varepsilon_t\rangle_{n_2}\right| > \gamma/3\right) \\ &+P\left(n_2^{\xi/2}\sup_{u,\boldsymbol{l_u}}\left|\langle\hat{\phi}^u_{\boldsymbol{l_u},n_1}, \varepsilon_t\rangle\right| > \gamma/3\right) \\ &= I + II + III \end{aligned}$$

<u>Term II:</u> We can bound II using the following simple inclusion:

$$\begin{aligned} \left\{n_2^{\xi/2}\sup_{u,\boldsymbol{l_u}}\left|\langle\hat{\phi}^u_{\boldsymbol{l_u},n_1}, \varepsilon_t\rangle_{n_2} - \langle\hat{\phi}^u_{\boldsymbol{l_u},n_1}, \varepsilon_t\rangle\right| > \gamma/3\right\} &\subset \{\text{there exists s such that}\, \varepsilon^s - \varepsilon_t^s \neq 0\} \\ &= \{\text{there exists s such that}\, |\varepsilon^s| > K_n\} \end{aligned}$$

Hence,

$$\begin{aligned} II &\leq P(\text{some } |\varepsilon^s| > K_n) \\ &\leq n_2 P(|\varepsilon| > K_n) \leq n_2 K_n^{-q}\mathbb{E}(|\varepsilon|^q) = \underset{n\to+\infty}{\mathcal{O}}(n^{1-q\xi/4}), \end{aligned}$$

where $n_2 = n/2$ and we have chosen $K_n := n^{\xi/4}$ since $q > 4/\xi$ by Assumption of the Lemma. Hence, II can become arbitrarily small.

Term I: Using again Bernstein's Inequality to the family of independent random variables $(\hat{\phi}^u_{\boldsymbol{l_u},n_1}(\mathbf{x}^s_u)\varepsilon^s_t)_{s=1,\cdots,n_2}$ and considering the two events $\{N_{n_1} > M^* + 1\}$ and $\{N_{n_1} < M^* + 1\}$, we can also show that:

$$I \leq 2Lp_n \exp\left(-\frac{1}{2}\frac{(\gamma^2/9){n_2}^{1-\xi}}{(M^*+1)^4\sigma^2 + (M^*+1)K_n\gamma/9{n_2}^{-\xi/2}}\right) + P(N_{n_1} > M^* + 1),$$

where $\sigma^2 := \mathbb{E}(|\varepsilon|^2)$. We can then make the right-hand side of the previous inequality arbitrarily small owing to $(\mathbf{H^2_b})$ with $K_n = n^{\xi/2}$.

Term III: by assumption, $\mathbb{E}(\phi^u_{\boldsymbol{l_u}}(\mathbf{X}_u)\varepsilon) = 0$. We then have:

$$\begin{aligned} III &\leq P\left(n_2^{\xi/2}\sup_{u,\boldsymbol{l_u}}\left|\mathbb{E}[(\hat{\phi}^u_{\boldsymbol{l_u},n_1} - \phi^u_{\boldsymbol{l_u}})(\mathbf{X}_u)\varepsilon_t]\right| > \gamma/6\right) + P\left(n_2^{\xi/2}\sup_{u,\boldsymbol{l_u}}\left|\mathbb{E}[\phi^u_{\boldsymbol{l_u}}(\mathbf{X}_u)(\varepsilon - \varepsilon_t)]\right| > \gamma/6\right) \\ &= III_1 + III_2, \end{aligned}$$

with,

$$\begin{aligned} III_1 &= P\left(n_2^{\xi/2}\sup_{u,\boldsymbol{l_u}}\left|\mathbb{E}[(\hat{\phi}^u_{\boldsymbol{l_u},n_1} - \phi^u_{\boldsymbol{l_u}})(\mathbf{X}_u)]\right| |\mathbb{E}(\varepsilon_t)| > \gamma/6\right) \\ &\leq P\left(n_2^{\xi/2}\sup_{u,\boldsymbol{l_u}}\left|\mathbb{E}[(\hat{\phi}^u_{\boldsymbol{l_u},n_1} - \phi^u_{\boldsymbol{l_u}})(\mathbf{X}_u)]\right| |\mathbb{E}(\varepsilon_t)| > \gamma/6\right) \\ &\leq \mathbb{1}_{\{n_2^{\xi/2}\sup_{u,\boldsymbol{l_u}}\left|\mathbb{E}[(\hat{\phi}^u_{\boldsymbol{l_u},n_1} - \phi^u_{\boldsymbol{l_u}})(\mathbf{X}_u)]\right||\mathbb{E}(\varepsilon_t)|>\gamma/6\}} \end{aligned}$$

Moreover, one has

$$\begin{aligned} |\mathbb{E}(\varepsilon_t)| &= \left|\int_{|x|\leq K_n} x\mathrm{d}P_\varepsilon(x) + \int_{|x|>K_n} sg(x)K_n\mathrm{d}P_\varepsilon(x)\right| = \left|\int_{|x|>K_n}(sg(x)K_n - x)\mathrm{d}P_\varepsilon(x)\right| \\ &\leq \int \mathbb{1}_{|x|>K_n}(K_n + |x|)\mathrm{d}P_\varepsilon(x) \\ &\leq K_nP_\varepsilon(|\varepsilon| > K_n) + \int |x|\,\mathbb{1}_{|x|>K_n}\mathrm{d}P_\varepsilon(x) \\ &\leq K_n^{1-t}\mathbb{E}(|\varepsilon|^t) + \mathbb{E}(\varepsilon^2)^{1/2}K_n^{-t/2}\mathbb{E}(|\varepsilon|^t)^{1/2} \quad \text{by the Tchebychev Inequality} \\ &\leq O(K_n^{1-t}) + O(K_n^{-t/2}) = o(K_n^{-2}) \end{aligned} \tag{6.12}$$

since $0 < \xi < 1$ and $t > 4/\xi > 4$. Then, set $K_n = n^{\xi/4}$, we obtain:

$$n_2^{\xi/2}\left\|\hat{\phi}^u_{\boldsymbol{l_u},n_1} - \phi^u_{\boldsymbol{l_u}}\right\| |\mathbb{E}(\varepsilon_t)| \leq n_2^{\xi/2}o(1)o(n^{-\xi/2}) = o(1),$$

when o is the usual Landau notation of relative insignificance.

Hence, $III_1 = 0$ for n large enough. For III_2, one has

$$III_2 \leq \mathbb{1}_{\{n_2^{\xi/2} \sup_{u,\boldsymbol{l_u}} \left|\mathbb{E}[\phi^u_{\boldsymbol{l_u}}(\mathbf{X}_u)(\varepsilon - \varepsilon_t)]\right| > \gamma/6\}},$$

and, by independance,

$$\left|\mathbb{E}[\phi^u_{\boldsymbol{l_u}}(\mathbf{X}_u)(\varepsilon - \varepsilon_t)]\right| = \left|\mathbb{E}[\phi^u_{\boldsymbol{l_u}}(\mathbf{X}_u)]\right| \left|\mathbb{E}(\varepsilon - \varepsilon_t)\right| \leq M^* \left|\mathbb{E}(\varepsilon - \varepsilon_t)\right|.$$

Equation (6.12) then implies,

$$|\mathbb{E}(\varepsilon - \varepsilon_t)| = \left|\int_{|x|>K_n} (sg(x)K_n - x)\mathrm{d}P_\varepsilon(x)\right| \leq o(K_n^{-2}) = o(n^{-\xi/2})$$

Thus, III is arbitrarily small for n and γ large enough and (iii) holds.

Assertion (iv) Remark that,

$$\sup_{u,\boldsymbol{l_u}} \left|\langle \tilde{f}, \hat{\phi}^u_{\boldsymbol{l_u},n_1}\rangle_{n_2} - \langle \tilde{f}, \hat{\phi}^u_{\boldsymbol{l_u},n_1}\rangle\right| \leq \|\boldsymbol{\beta^0}\|_{L^1} \sup_{u,\boldsymbol{l_u}} \left|\langle \phi^v_{\boldsymbol{l_v}}, \hat{\phi}^u_{\boldsymbol{l_u},n_1}\rangle_{n_2} - \langle \phi^v_{\boldsymbol{l_v}}, \hat{\phi}^u_{\boldsymbol{l_u},n_1}\rangle\right|.$$

Now, $(\mathbf{H_s})$ and Bernstein's Inequality implies

$$\begin{aligned} P\left(\sup_{u,\boldsymbol{l_u}} \left|\langle \phi^v_{\boldsymbol{l_v}}, \hat{\phi}^u_{\boldsymbol{l_u},n_1}\rangle_{n_2} - \langle \phi^v_{\boldsymbol{l_v}}, \hat{\phi}^u_{\boldsymbol{l_u},n_1}\rangle\right| \geq \gamma n_2^{-\xi/2}\right) &\leq P(N_{n_1} > M^* + 1) \\ +2Lp_n \exp\left(-\frac{1}{2}\frac{\gamma^2 {n_2}^{1-\xi}}{(M^*+1)^4 + (M^*+1)^2\gamma/3{n_2}^{-\xi/2}}\right), \end{aligned}$$

which implies with Assumption $(\mathbf{H_b^2})$ that:

$$\sup_{u,\boldsymbol{l_u}} \left|\langle \phi^v_{\boldsymbol{l_v}}, \hat{\phi}^u_{\boldsymbol{l_u},n_1}\rangle_{n_2} - \langle \phi^v_{\boldsymbol{l_v}}, \hat{\phi}^u_{\boldsymbol{l_u},n_1}\rangle\right| = \mathcal{O}_P(n^{-\xi/2}).$$

□

The following lemma, similar to Lemma 2 from Bühlmann (2006), then holds:

Lemma 7. *Under Assumptions* $(\mathbf{H_b})$, $(\mathbf{H}_{\varepsilon,\mathbf{q}})$ *with* $q > 4/\xi$ *and* $(\mathbf{H_s})$, *there exists a constant* $C > 0$ *such that, on the set* $\Omega_n = \{\omega, |\zeta_n(\omega)| < 1/2\}$*:*

$$\sup_{u,\boldsymbol{l_u}} |\langle Y - G_k(\bar{f}), \hat{\phi}^u_{\boldsymbol{l_u},n_1}\rangle_{n_2} - \langle \tilde{R}_k(\bar{f}), \phi^u_{\boldsymbol{l_u}}\rangle| \leq C\left(\frac{5}{2}\right)^k \zeta_n.$$

Proof. Denote $A_n(k,u) = \langle Y - G_k(\bar{f}), \hat{\phi}^u_{\boldsymbol{l_u},n_1}\rangle_{n_2} - \langle \tilde{R}_k(\bar{f}), \phi^u_{\boldsymbol{l_u}}\rangle$. Assume first that $k=0$,

$$\begin{array}{rcl} \sup\limits_{u,\boldsymbol{l_u}}|A_n(0,u)| & = & \sup\limits_{u}|\langle Y, \hat{\phi}^u_{\boldsymbol{l_u},n_1}\rangle_{n_2} - \langle \bar{f}, \phi^u_{\boldsymbol{l_u}}\rangle| \\ & \leq & \sup\limits_{u,\boldsymbol{l_u}}\left\{\left|\langle \tilde{f}, \hat{\phi}^u_{\boldsymbol{l_u},n_1}\rangle_{n_2} - \langle \tilde{f}, \hat{\phi}^u_{\boldsymbol{l_u},n_1}\rangle\right| + \left|\langle \tilde{f} - \bar{f}, \hat{\phi}^u_{\boldsymbol{l_u},n_1}\rangle\right| + \left|\langle \bar{f}, \hat{\phi}^u_{\boldsymbol{l_u},n_1} - \phi^u_{\boldsymbol{l_u}}\rangle\right|\right\} \\ & & +\sup\limits_{u,\boldsymbol{l_u}}\left|\langle \varepsilon, \hat{\phi}^u_{\boldsymbol{l_u},n_1}\rangle_{n_2}\right| \\ & \leq & (3 + \|\bar{f}\|)\zeta_n \quad \text{by (iii)-(iv) of Lemma 6 and Theorem 1} \end{array}$$

From the main document, we remind that

$$G_k(\bar{f}) = G_{k-1}(\bar{f}) + \gamma\langle Y - G_{k-1}(\bar{f}), \hat{\phi}^{u_k}_{\boldsymbol{l_{u_k}},n_1}\rangle_{n_2} \cdot \hat{\phi}^{u_k}_{\boldsymbol{l_{u_k}},n_1}, \tag{6.13}$$

$$\begin{array}{rcl} R_k(\bar{f}) & = & \bar{f} - G_k(\bar{f}) \\ & = & \bar{f} - G_{k-1}(\bar{f}) - \gamma\langle Y - G_{k-1}(\bar{f}), \hat{\phi}^{u_k}_{\boldsymbol{l_{u_k}},n_1}\rangle_{n_2} \cdot \hat{\phi}^{u_k}_{\boldsymbol{l_{u_k}},n_1} \end{array} \tag{6.14}$$

and

$$\left\{\begin{array}{l} \tilde{R}_0(\bar{f}) = \bar{f} \\ \tilde{R}_k(\bar{f}) = \tilde{R}_{k-1}(\bar{f}) - \gamma\langle \tilde{R}_{k-1}(\bar{f}), \hat{\phi}^{u_k}_{\boldsymbol{l_{u_k}},n_1}\rangle \hat{\phi}^{u_k}_{\boldsymbol{l_{u_k}},n_1}. \end{array}\right. \tag{6.15}$$

From the recursive relations (6.13) and (6.15), for any $k \geq 0$, we obtain:

$$\begin{array}{rcl} A_n(k,u) & = & \langle Y - G_{k-1}(\bar{f}) - \gamma\langle Y - G_{k-1}(\bar{f}), \hat{\phi}^{u_k}_{\boldsymbol{l_{u_k}},n_1}\rangle_{n_2} \cdot \hat{\phi}^{u_k}_{\boldsymbol{l_{u_k}},n_1}, \hat{\phi}^{u}_{\boldsymbol{l_u},n_1}\rangle_n \\ & & -\langle \tilde{R}_{k-1}(\bar{f}) - \gamma\langle \tilde{R}_{k-1}(\bar{f}), \hat{\phi}^{u_k}_{\boldsymbol{l_{u_k}},n_1}\rangle \hat{\phi}^{u_k}_{\boldsymbol{l_{u_k}},n_1}, \phi^u_{\boldsymbol{l_u}}\rangle \\ & \leq & A_n(k-1,u) \\ & & -\gamma\underbrace{\left(\langle Y - G_{k-1}(\bar{f}), \hat{\phi}^{u_k}_{\boldsymbol{l_{u_k}},n_1}\rangle_{n_2} - \langle \tilde{R}_{k-1}(\bar{f}), \phi^{u_k}_{\boldsymbol{l_{u_k}}}\rangle\right)\langle \hat{\phi}^{u_k}_{\boldsymbol{l_{u_k}},n_1}, \hat{\phi}^{u}_{\boldsymbol{l_u},n_1}\rangle_{n_2}}_{I} \\ & & +\gamma\underbrace{\langle \tilde{R}_{k-1}(\bar{f}), \phi^{u_k}_{\boldsymbol{l_{u_k}}}\rangle\left(\langle \hat{\phi}^{u_k}_{\boldsymbol{l_{u_k}},n_1}, \phi^u_{\boldsymbol{l_u}}\rangle - \langle \hat{\phi}^{u_k}_{\boldsymbol{l_{u_k}},n_1}, \hat{\phi}^{u}_{\boldsymbol{l_u},n_1}\rangle_{n_2}\right)}_{II} \\ & & +\gamma\underbrace{\langle \tilde{R}_{k-1}(\bar{f}), \hat{\phi}^{u_k}_{\boldsymbol{l_{u_k}},n_1} - \phi^{u_k}_{\boldsymbol{l_{u_k}}}\rangle\langle \hat{\phi}^{u_k}_{\boldsymbol{l_{u_k}},n_1}, \phi^u_{\boldsymbol{l_u}}\rangle}_{III}. \end{array}$$

On the one hand, using assertion (ii) of Lemma 6, and the Cauchy-Schwarz inequality (with $\left\|\phi^u_{\boldsymbol{l_u}}\right\| = 1$), it comes

$$\begin{aligned}
\sup_{u,\boldsymbol{l_u}}|I| &\leq \sup_{u,\boldsymbol{l_u}}|\langle\hat{\phi}^{u_k}_{\boldsymbol{l_{u_k}},n_1}, \hat{\phi}^u_{\boldsymbol{l_u},n_1}\rangle_{n_2}|\sup_{u,\boldsymbol{l_u}}|A_n(k-1,u)| \\
&\leq (\sup_{u,\boldsymbol{l_u}}|\langle\phi^{u_k}_{\boldsymbol{l_{u_k}}}, \phi^u_{\boldsymbol{l_u}}\rangle| + \zeta_n)\sup_{u,\boldsymbol{l_u}}|A_n(k-1,u)| \\
&\leq (1+\zeta_n)\sup_{u,\boldsymbol{l_u}}|A_n(k-1,u)|.
\end{aligned}$$

Consider now the phantom residual, from its recursive relation, we can show that $\left\|\tilde{R}_k(\bar{f})\right\|^2 = \left\|\tilde{R}_{k-1}(\bar{f})\right\|^2 - \gamma(2-\gamma)\langle\tilde{R}_{k-1}(\bar{f}), \hat{\phi}^{u_k}_{\boldsymbol{l_{u_k}},n_1}\rangle^2 \leq \left\|\tilde{R}_{k-1}(\bar{f})\right\|^2$ and we deduce

$$\left\|\tilde{R}_k(\bar{f})\right\|^2 \leq \left\|\bar{f}\right\|^2. \tag{6.16}$$

Then,

$$\begin{aligned}
\sup_{u,\boldsymbol{l_u}}|II| &\leq \left\|\tilde{R}_{k-1}(\bar{f})\right\| \left\|\phi^{u_k}_{\boldsymbol{l_{u_k}}}\right\| \sup_{u,\boldsymbol{l_u}}|\langle\hat{\phi}^{u_k}_{\boldsymbol{l_{u_k}},n_1}, \phi^u_{\boldsymbol{l_u}}\rangle - \langle\hat{\phi}^{u_k}_{\boldsymbol{l_{u_k}},n_1}, \hat{\phi}^u_{\boldsymbol{l_u},n_1}\rangle_{n_2}| \\
&\leq \left\|\bar{f}\right\| \sup_{u,\boldsymbol{l_u}}|\langle\hat{\phi}^{u_k}_{\boldsymbol{l_{u_k}},n_1}, \phi^u_{\boldsymbol{l_u}}\rangle - \langle\hat{\phi}^{u_k}_{\boldsymbol{l_{u_k}},n_1}, \hat{\phi}^u_{\boldsymbol{l_u},n_1}\rangle_{n_2}|,
\end{aligned}$$

with

$$\begin{aligned}
|\langle\hat{\phi}^{u_k}_{\boldsymbol{l_{u_k}},n_1}, \phi^u_{\boldsymbol{l_u}}\rangle - \langle\hat{\phi}^{u_k}_{\boldsymbol{l_{u_k}},n_1}, \hat{\phi}^u_{\boldsymbol{l_u},n_1}\rangle_{n_2}| &\leq |\langle\hat{\phi}^{u_k}_{\boldsymbol{l_{u_k}},n_1}, \hat{\phi}^u_{\boldsymbol{l_u},n_1}\rangle_{n_2} - \langle\phi^{u_k}_{\boldsymbol{l_{u_k}}}, \phi^u_{\boldsymbol{l_u}}\rangle| \\
&\quad +|\langle\phi^{u_k}_{\boldsymbol{l_{u_k}}} - \hat{\phi}^{u_k}_{\boldsymbol{l_{u_k}},n_1}, \phi^u_{\boldsymbol{l_u}}\rangle|.
\end{aligned}$$

Using again assertion (ii) from Lemma 6 and Theorem 1, we obtain the following bound for II,

$$\begin{aligned}
\sup_{u,\boldsymbol{l_u}}|II| &\leq \left\|\bar{f}\right\|(\zeta_n + \sup_{u,\boldsymbol{l_u}}\left\|\phi^u_{\boldsymbol{l_u}} - \hat{\phi}^u_{\boldsymbol{l_u},n_1}\right\|) \\
&\leq 2\zeta_n\left\|\bar{f}\right\|.
\end{aligned}$$

Finally, Theorem 1 gives

$$\begin{aligned}
\sup_{u,\boldsymbol{l_u}}|III| &\leq \sup_{u,\boldsymbol{l_u}}\left\|\tilde{R}_{k-1}(\bar{f})\right\| \left\|\hat{\phi}^{u_k}_{\boldsymbol{l_{u_k}},n_1} - \phi^{u_k}_{\boldsymbol{l_{u_k}}}\right\| \left\|\hat{\phi}^{u_k}_{\boldsymbol{l_{u_k}},n_1}\right\| \left\|\phi^u_{\boldsymbol{l_u}}\right\| \\
&\leq \left\|\bar{f}\right\|\zeta_n.
\end{aligned}$$

Our bounds on I, II and III, and $\gamma < 1$ yields on $\Omega_n = \{\zeta_n < 1/2\}$ that

$$\begin{aligned}
\sup_{u,\boldsymbol{l_u}}|A_n(k,u)| &\leq \sup_{u,\boldsymbol{l_u}}|A_n(k-1,u)| + (1+\zeta_n)\sup_{u,\boldsymbol{l_u}}|A_n(k-1,u)| + 3\zeta_n\left\|\bar{f}\right\| \\
&\leq \frac{5}{2}\sup_{u,\boldsymbol{l_u}}|A_n(k-1,u)| + 3\zeta_n\left\|\bar{f}\right\|.
\end{aligned}$$

A simple induction yields:

$$\begin{aligned}
\sup_{u,\boldsymbol{l_u}}|A_n(k,u)| &\leq \left(\frac{5}{2}\right)^k \underbrace{\sup_{u,\boldsymbol{l_u}}|A_n(0,u)|}_{\leq(3+\left\|\bar{f}\right\|)\zeta_n} + 3\zeta_n \left\|\bar{f}\right\| \sum_{\ell=0}^{k-1}\left(\frac{5}{2}\right)^\ell \\
&\leq \left(\frac{5}{2}\right)^k \zeta_n \left(3 + \|\boldsymbol{\beta^0}\|_{L^1}\left(1 + 3\sum_{\ell=1}^{\infty}\left(\frac{5}{2}\right)^{-\ell}\right)\right),
\end{aligned}$$

which ends the proof with $C = 3 + \|\boldsymbol{\beta^0}\|_{L^1}\left(1 + 3\sum_{\ell=1}^{\infty}\left(\frac{5}{2}\right)^{-\ell}\right)$.

□

We then aim at applying Theorem 2.1 from Champion, Cierco-Ayrolles, Gadat and Vignes (2013) to the phantom residuals $(\tilde{R}_k(\bar{f}))_k$. Using the notation of Champion, Cierco-Ayrolles, Gadat and Vignes (2013), this will be possible if we can show that the phantom residuals follows a theoretical boosting with a shrinkage parameter $\nu \in [0,1]$. Thanks to Lemma 7 and by definiton of $\hat{\phi}^{u_k}_{\boldsymbol{l_{u_k}},n_1}$, one has

$$\begin{aligned}
|\langle Y - G_{k-1}(\bar{f}), \hat{\phi}^{u_k}_{\boldsymbol{l_{u_k}},n_1}\rangle_{n_2}| &= \sup_{u,\boldsymbol{l_u}}|\langle Y - G_{k-1}(\bar{f}), \hat{\phi}^{u}_{\boldsymbol{l_u},n_1}\rangle_{n_2}| \\
&\geq \sup_{u,\boldsymbol{l_u}}\left\{|\langle \tilde{R}_{k-1}(\bar{f}), \phi^u_{\boldsymbol{l_u}}\rangle| - C\left(\frac{5}{2}\right)^{k-1}\zeta_n\right\}. \qquad (6.17)
\end{aligned}$$

Applying again Lemma 7 on the set Ω_n, we obtain:

$$\begin{aligned}
|\langle \tilde{R}_{k-1}(\bar{f}), \phi^{u_k}_{\boldsymbol{l_{u_k}}}\rangle| &\geq |\langle Y - G_{k-1}(\bar{f}), \hat{\phi}^{u_k}_{\boldsymbol{l_{u_k}},n_1}\rangle_{n_2}| - C\left(\frac{5}{2}\right)^{k-1}\zeta_n \\
&\geq \sup_{u,\boldsymbol{l_u}}|\langle \tilde{R}_{k-1}(\bar{f}), \phi^u_{\boldsymbol{l_u}}\rangle| - 2C\left(\frac{5}{2}\right)^{k-1}\zeta_n. \qquad (6.18)
\end{aligned}$$

Consider now the set $\tilde{\Omega}_n = \left\{\omega, \quad \forall k \leq k_n, \quad \sup_{u,\boldsymbol{l_u}}|\langle \tilde{R}_{k-1}(\bar{f}), \phi^u_{\boldsymbol{l_u}}\rangle| > 4C\left(\frac{5}{2}\right)^{k-1}\zeta_n\right\}$. We deduce from Equation (6.18) the following inequality on $\Omega_n \cap \tilde{\Omega}_n$:

$$|\langle \tilde{R}_{k-1}(\bar{f}), \phi^{u_k}_{\boldsymbol{l_{u_k}}}\rangle| \geq \frac{1}{2}\sup_{u,\boldsymbol{l_u}}|\langle \tilde{R}_{k-1}(\bar{f}), \phi^u_{\boldsymbol{l_u}}\rangle|. \qquad (6.19)$$

Consequently, on $\Omega_n \cap \tilde{\Omega}_n$, the family $(\tilde{R}_k(\bar{f}))_k$ satisfies a theoretical boosting, given by Algorithm 1 of Champion, Cierco-Ayrolles, Gadat and Vignes (2013), with constant $\nu = 1/2$ and we have:

$$\left\|\tilde{R}_k(\bar{f})\right\| \leq C' \left(1 + \frac{1}{4}\gamma(2-\gamma)k\right)^{-\frac{2-\gamma}{2(6-\gamma)}}. \tag{6.20}$$

Consider now the complementary set

$$\tilde{\Omega}_n^C = \left\{\omega, \quad \exists\, k \leq k_n \quad \sup_{u,\boldsymbol{l_u}} |\langle \tilde{R}_{k-1}(\bar{f}), \phi^u_{\boldsymbol{l_u}}\rangle| \leq 4C\left(\frac{5}{2}\right)^{k-1} \zeta_n\right\}.$$

Remark that

$$\begin{aligned}
\left\|\tilde{R}_k(\bar{f})\right\|^2 &= \langle \tilde{R}_k(\bar{f}), \bar{f} - \gamma \textstyle\sum_{j=0}^{k-1} \langle \tilde{R}_j(\bar{f}), \hat{\phi}^{u_j}_{\boldsymbol{l_{u_j}},n_1}\rangle \hat{\phi}^{u_j}_{\boldsymbol{l_{u_j}},n_1}\rangle \\
&\leq \|\boldsymbol{\beta^0}\|_{L^1} \sup_{u,\boldsymbol{l_u}} \left|\langle \tilde{R}_k(\bar{f}), \hat{\phi}^u_{\boldsymbol{l_u},n_1}\rangle\right| + \gamma \textstyle\sum_{j=0}^{k-1} \left|\langle \tilde{R}_j(\bar{f}), \hat{\phi}^{u_j}_{\boldsymbol{l_{u_j}},n_1}\rangle\right| \displaystyle\sup_{u,\boldsymbol{l_u}} \left|\langle \tilde{R}_k(\bar{f}), \hat{\phi}^u_{\boldsymbol{l_u},n_1}\rangle\right|.
\end{aligned}$$

Moreover,

$$\begin{aligned}
\sup_{u,\boldsymbol{l_u}} \left|\langle \tilde{R}_k(\bar{f}), \hat{\phi}^u_{\boldsymbol{l_u},n_1}\rangle\right| &\leq \sup_{u,\boldsymbol{l_u}} \left|\langle \tilde{R}_k(\bar{f}), \phi^u_{\boldsymbol{l_u}}\rangle\right| + \sup_{u,\boldsymbol{l_u}} \left|\langle \tilde{R}_k(\bar{f}), \hat{\phi}^u_{\boldsymbol{l_u},n_1} - \phi^u_{\boldsymbol{l_u}}\rangle\right| \\
&\leq \sup_{u,\boldsymbol{l_u}} \left|\langle \tilde{R}_k(\bar{f}), \phi^u_{\boldsymbol{l_u}}\rangle\right| + \left\|\bar{f}\right\| \zeta_n \quad \text{by Theorem 1 and (6.16)}
\end{aligned}$$

We hence have

$$\begin{aligned}
\left\|\tilde{R}_k(\bar{f})\right\|^2 &\leq \left(\|\boldsymbol{\beta^0}\|_{L^1} + \gamma \textstyle\sum_{j=0}^{k-1} \left|\langle \tilde{R}_j(\bar{f}), \hat{\phi}^{u_j}_{\boldsymbol{l_{u_j}},n_1}\rangle\right|\right) \left(\displaystyle\sup_{u,\boldsymbol{l_u}} \left|\langle \tilde{R}_k(\bar{f}), \phi^u_{\boldsymbol{l_u}}\rangle\right| + \left\|\bar{f}\right\| \zeta_n\right) \\
&\leq \left(\|\boldsymbol{\beta^0}\|_{L^1} + \gamma k \left\|\bar{f}\right\|\right) \left(\sup_{u,\boldsymbol{l_u}} \left|\langle \tilde{R}_k(\bar{f}), \phi^u_{\boldsymbol{l_u}}\rangle\right| + \left\|\bar{f}\right\| \zeta_n\right) \\
&\leq \left(\|\boldsymbol{\beta^0}\|_{L^1} + \gamma k \left\|\bar{f}\right\|\right) \left(4C\left(\tfrac{5}{2}\right)^k \zeta_n + \left\|\bar{f}\right\| \zeta_n\right) \quad \text{on } \tilde{\Omega}_n^C
\end{aligned} \tag{6.21}$$

Finally, on the set $(\Omega_n \cap \tilde{\Omega}_n) \cup \tilde{\Omega}_n^C$, by Equations (6.20) and (6.21),

$$\left\|\tilde{R}_k(\bar{f})\right\|^2 \leq C'^2 \left(1 + \frac{1}{4}\gamma(2-\gamma)k\right)^{-\frac{2-\gamma}{6-\gamma}} + \left(\|\boldsymbol{\beta^0}\|_{L^1} + \gamma k \left\|\bar{f}\right\|\right) (4C\left(\frac{5}{2}\right)^k \zeta_n + \left\|\bar{f}\right\| \zeta_n) \tag{6.22}$$

To conclude the first part of the proof, remark that

$$P\left((\Omega_n \cap \tilde{\Omega}_n) \cup \tilde{\Omega}_n^C\right) \geq P(\Omega_n) \underset{n\to+\infty}{\longrightarrow} 1.$$

Now, by Assumption $(\mathbf{H_s})$ and by Lemma 5, we have,

$$\left\|\bar{f}\right\| \zeta_n \leq \|\boldsymbol{\beta^0}\|_{L^1} N_{n_1} \zeta_n \leq \|\boldsymbol{\beta^0}\|_{L^1} (M^* + \mathcal{O}_P(n^{\vartheta - \xi/2}))\zeta_n \to 0.$$

Thus, Inequality (6.22) holds almost surely, and for $k_n < (\xi/2-\vartheta)/2\log(3)\log(n)$, which grows sufficiently slowly, we get

$$\left\|\tilde{R}_{k_n}(\bar{f})\right\| \xrightarrow[n\to+\infty]{\mathbb{P}} 0. \tag{6.23}$$

Consider now $A_k := \left\|R_k(\bar{f}) - \tilde{R}_k(\bar{f})\right\|$ for $k \geq 1$. By definitions reminded in (6.14)-(6.15), we have:

$$\begin{aligned} A_k &\leq A_{k-1} + \gamma|\langle Y - G_{k-1}(\bar{f}), \hat{\phi}^{u_k}_{\boldsymbol{l_{u_k}},n_1}\rangle_{n_2} - \langle \tilde{R}_{k-1}(\bar{f}), \hat{\phi}^{u_k}_{\boldsymbol{l_{u_k}},n_1}\rangle| \\ &\leq A_{k-1} + \gamma|\langle Y - G_{k-1}(\bar{f}), \hat{\phi}^{u_k}_{\boldsymbol{l_{u_k}},n_1}\rangle_{n_2} - \langle \tilde{R}_{k-1}(\bar{f}), \phi^{u_k}_{\boldsymbol{l_{u_k}}}\rangle| \qquad (6.24)\\ &\quad +\gamma|\langle \tilde{R}_{k-1}(\bar{f}), \hat{\phi}^{u_k}_{\boldsymbol{l_{u_k}},n_1} - \phi^{u_k}_{\boldsymbol{l_{u_k}}}\rangle|. \end{aligned}$$

By Lemma 7, we then deduce the following inequality on Ω_n:

$$A_k \leq A_{k-1} + \gamma\left(C\left(\frac{5}{2}\right)^{k-1} + 1\right)\zeta_n + \gamma\left\|\bar{f}\right\|\zeta_n. \tag{6.25}$$

Since $A_0 = 0$, we deduce recursively from Equation (6.25) that, on Ω_n,

$$A_{k_n} \xrightarrow[n\to+\infty]{\mathbb{P}} 0.$$

Finally, as

$$\left\|\hat{f} - \tilde{f}\right\| = \left\|G_{k_n}(\bar{f}) - \tilde{f}\right\| \leq \left\|\bar{f} - \tilde{f}\right\| + \left\|R_{k_n}(\bar{f}) - \tilde{R}_{k_n}(\bar{f})\right\| + \left\|\tilde{R}_{k_n}(\bar{f})\right\|,$$

it remains to treat the term $\left\|\bar{f} - \tilde{f}\right\|$. As,

$$\left\|\bar{f} - \tilde{f}\right\| \leq \|\boldsymbol{\beta^0}\|_{L^1}\left\|\phi^u_{\boldsymbol{l_u}} - \hat{\phi}^u_{\boldsymbol{l_u},n_1}\right\|,$$

and the end of the proof follows using Assumption ($\mathbf{H_s}$) and Theorem 1. □

Bibliography

Birgé, L. and Massart, P. (1998). Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, 4(3):329–375.

Bühlmann, P. (2006). Boosting for high-dimensional linear models. *The Annals of Statistics*, 34(2):559–583.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for high-dimensional data*. Springer, Berlin.

Bullen, P. (1998). *A dictionary of Inequalities*. Addison-Wesley Longman.

Cacuci, D., Ionescu-Bujor, M., and Navon, I. (2005). *Sensitivity and Uncertainty Analysis, Volume II: Applications to Large-Scale Systems*, volume 2. Chapman & Hall/CRC.

Champion, M., Cierco-Ayrolles, C., Gadat, S., and Vignes, M. (2013). Sparse regression and support recovery with $\mathbb{L}_2$-boosting algorithm, *Preprint*.

Chastaing, G., Gamboa, F., and Prieur, C. (2012). Generalized hoeffding-sobol decomposition for dependent variables -Application to sensitivity analysis. *Electronic Journal of Statistics*, 6:2420–2448.

Chastaing, G., Gamboa, F., and Prieur, C. (2013). Generalized sobol sensitivity indices for dependent variables: Numerical methods. Available at `http://arxiv.org/abs/1303.4372`.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, 32(2):407–451.

Friedman, J. (2001). Greedy function approximation: a gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(3):293–325.

Hooker, G. (2007). Generalized functional anova diagnostics for high-dimensional functions of dependent variables. *Journal of Computational and Graphical Statistics*, 16(3):709–732.

Huang, J. (1998). Projection estimation in multiple regression with application to functional anova models. *The Annals of Statistics*, 26(1):242–272.

Li, G., Rabitz, H., Yelvington, P., Oluwole, O., Bacon, F., C.E., K., and Schoendorf, J. (2010). Global sensitivity analysis with independent and/or correlated inputs. *Journal of Physical Chemistry A*, 114:6022–6032.

Saltelli, A., Chan, K., and Scott, E. (2000). *Sensitivity Analysis*. Wiley, West Sussex.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., and Tarantola, S. (2008). *Global sensitivity analysis: The primer*. Wiley-Interscience, West Sussex.

Sobol, I. M. (1993). Sensitivity estimates for nonlinear mathematical models. *Mathematical Modeling and Computational Experiment*, 1(4):407–414.

Sobol, I. M. (2001). Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates. *Mathematics and Computers in Simulations*, 55:271–280.

Stone, C. (1994). The use of polynomial splines and their tensor products in multivariate function estimation. *The Annals of Statistics*, 22(1):118–171.

Temlyakov, V. N. (2000). Weak Greedy Algorithms. *Advances in Computational Mathematics*, 12(2,3):213–227.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 58(1):267–288.

Tropp, J. (2012). User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434.

von Mises, R. (1913). Mechanik der festen körper im plastisch deformablen zustand. *Göttin. Nachr. Math. Phys.*, 1:582–592.

Zhang, T. (2011). Adaptive forward-backward algorithm for learning sparse representations. *IEEE transactions on information theory*, 57(7):4689–4708.

Institut de Mathématiques de Toulouse, 118, route de Narbonne F-31062 Toulouse Cedex 9, France

magali.champion@math.univ-toulouse.fr

Institut de Mathématiques de Toulouse, 118, route de Narbonne F-31062 Toulouse Cedex 9, France

gaelle.chastaing@math.univ-toulouse.fr

Institut de Mathématiques de Toulouse, 118, route de Narbonne F-31062 Toulouse Cedex 9, France

sebastien.gadat@math.univ-toulouse.fr

Université Joseph Fourier, LJK/MOISE BP 53, 38041 Grenoble Cedex, France

clementine.prieur@imag.fr