$\mathbb{L}_2$-Boosting on a 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 setting. We then apply a greedy algorithm with these previous estimators to a sensitivity analysis. We also establish the consistency of this $\mathbb{L}_2$-boosting under sparsity assumptions of the signal to be analyzed. The paper concludes with numerical experiments, that demonstrate the low computational cost of our method, as well as its efficiency on the standard benchmark of sensitivity analysis.

Keywords 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 decompositions and High Dimensional Representation Models (HDMR) (Hooker, 2007; Li et al., 2010) are well known expansions that make it possible to understand model behavior and to detect how inputs interact with each other.

When input variables are independent, Hoeffding establishes the uniqueness of the decomposition provided that the summands are mutually orthogonal (Hoeffding, 1948). However, in practice, this assumption is sometimes difficult to justify, or may even be wrong (see Li and Rabitz (2010) for an application to correlated ionosonde data, or Jacques et al. (2006), who studied an adjusted

neutron spectrum inferred from a correlated dependent nuclear dataset).

When inputs are correlated, the orthogonality properties of the classical Sobol decomposition (Sobol, 1993) are no longer satisfied. As pointed out by several authors (Hooker, 2007; Da Veiga et al., 2009), a global sensitivity analysis based on this decomposition may lead to erroneous conclusions. Following the work of Stone (1994), later applied in Machine Learning by Hooker (2007), and to sensitivity analysis by Chastaing et al. (2012), we consider a hierarchically orthogonal decomposition in this paper, whose uniqueness has been proven under mild conditions on the dependence structure of the inputs (Chastaing et al., 2012). In other words, any model function can be uniquely decomposed as a sum of *hierarchically orthogonal* component functions. Two summands are considered to be *hierarchically orthogonal* whenever all of the 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 referred to as a Hierarchically Orthogonal Functional Decomposition (HOFD).

It is of great importance to develop estimation procedures since the analytical formulation for HOFD is rarely available. In this paper, we focus on an alternative method proposed in Chastaing et al. (2013) to estimate the HOFD components. It consists in constructing a hierarchically orthogonal basis from a suitable Hilbert orthonormal basis. Inspired by the usual Gram-Schmidt algorithm, the procedure recursively builds a multidimensional basis for each component that satisfies the identifiability constraints imposed on this summand. Each component is then well approximated on a truncated basis, where the unknown coefficients are deduced by solving an ordinary least-squares regression. Nevertheless, in a high-dimensional paradigm, this procedure suffers from a curse of dimensionality. Moreover, it is numerically observed that only a few of the 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 to then identify the HOFD with a limited computational budget.

With this in mind, we suggest in this article to transform the ordinary least-squares regression into a penalized regression, as has been proposed in Chastaing

et al. (2013). In the present paper, we focus on the $\mathbb{L}_2$-boosting to deal with the ℓ_0 penalization, developed 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 (from the practical point of view) to the LARS algorithm proposed by Efron et al. (2004), which solves the Lasso regression (Tibshirani, 1996; Bühlmann and van de Geer, 2011). 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 component 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 provide an overall consistent estimation of a signal spanned into a large dimensional dictionary derived from a HOFD. Hence, our work significantly improves the results of Chastaing et al. (2013): we first address the convergence rate of the empirical HOFD and then use this result to obtain a sparse estimator of the unknown signal. We will need to manage sufficient theoretical conditions to ensure the consistency of our estimator. In addition, we discuss these conditions and provide some numerical examples in which such conditions are fulfilled.

The article is organized as follows. The notation used in the paper is presented in Section 2.1. Section 2.2 provides the HOFD representation of the model function. In Section 2.3, we review the procedure detailed in Chastaing et al. (2013) that consists in constructing well-tailored hierarchically orthogonal bases to represent the components of the HOFD. Finally, we highlight the curse of dimensionality that we are exposed to, and present the $\mathbb{L}_2$-boosting. Section 3 describes our main theoretical results on the proposed algorithms. One interesting application of the general theory is the global sensitivity analysis (SA). In Section 4, we apply the $\mathbb{L}_2$-boosting to estimate the generalized sensitivity indices defined in Chastaing et al. (2012, 2013). After recalling the form of these indices, we numerically compare the $\mathbb{L}_2$-boosting performance with a Lasso strategy and the Forward-Backward algorithm proposed by Zhang (2011). Section 5

contains the conclusion, and the proofs of the two main theorems are given in the Appendix.

Acknowledgment The authors are indebted to Fabrice Gamboa for his stimulating discussions and his 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 the distribution law of $\mathbf{X}$ by $P_{\mathbf{X}}$, which is unknown in our setting, and we assume that $P_{\mathbf{X}}$ 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 $\mathbf{X}$ may be correlated.

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

$$\langle h, g \rangle = \int h(\mathbf{x}) g(\mathbf{x}) p_{\mathbf{X}} d\mathbf{x} = \mathbb{E}(h(\mathbf{X}) g(\mathbf{X}))$$

$$\|h\|^2 = \langle h, h \rangle = \mathbb{E}\left(h(\mathbf{X})^2\right), \quad \forall h, g \in L^2_{\mathbb{R}},$$

where $\mathbb{E}(\cdot)$ stands for the expected value, $V(\cdot) = \mathbb{E}[(\cdot - \mathbb{E}(\cdot))^2]$ 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 the marginal distribution of X_i by $P_{\mathbf{X}_i}$ and naturally extend 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 by $P_{\mathbf{X}_u}$ (*resp.* $p_{\mathbf{X}_u}$).

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

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

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 et al. (2012)), the decomposition is unique under some additional orthogonality assumptions.

More precisely, let us introduce $H_\emptyset = H_\emptyset^0$ as the set of constant functions, and for all $u \in S^*$, $H_u := L^2_{\mathbb{R}}(\mathbb{R}^u, \mathcal{B}(\mathbb{R}^u), P_{\mathbf{X}_u})$. We then 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 Assumptions (C.1) and (C.2) in Chastaing et al. (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 assumed to be hierarchically orthogonal, that is, $\langle f_u, f_v \rangle = 0\ \forall v \subset u$.*

To obtain more information about the HOFD, the reader is referred to Hooker (2007) and Chastaing et al. (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 will assume (without loss of generality) that f is centered, so that $f_\emptyset = 0$. Furthermore, most of the models are only governed by low-order interaction effects, as pointed out in Crestaux et al. (2009), Blatman (2009) and Li et al. (2010). We thus suppose that f is

well approximated by:

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

so that interactions of order $\geq d+1$ can be neglected. The choice of d, which is directly related to the notion of effective dimension in the superposition sense (see Definition 1 in Wang and Fang (2003)), is addressed in Zuniga et al. (2013), but is not of great interest in the present article, so that it is assumed to be fixed by the user. Even by choosing a small d, the number of components in (2.3) can become prohibitive if the number of inputs p is high. We therefore are interested in estimation procedures under sparse assumptions when the number of variables p is large.

In the next section, we recall the procedure to identify components of (2.3). As a result of this strategy, we highlight the curse of dimensionality when p becomes 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 a two-step estimation procedure in this section 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 et al. (2013), and the second consists of a sparse estimation in the dictionary learned by the empirical HOGS.

In the following, we have chosen to use the so-called $\mathbb{L}_2$-boosting procedure instead of the widely used Lasso estimator. This choice is motivated by two reasons.

- First, from a technical point of view, the empirical HOGS will produce a noisy estimation of the theoretical dictionary, in which the true signal f is expanded. Hence, the arguments produced for the Lasso estimation would have to be completely adjusted to this situation with errors in the variables. Moreover, as an M-estimator, such a modification is far from being trivial (see Cavalier and Hengartner (2005) for an example of oracle inequalities derived from M estimators with noise in the variables). In contrast, the

approximation obtained in the empirical HOGS can be easily handled with the boosting algorithm since we just have to quantify how the empirical inner products built with noisy variables are close to theoretical ones. Our proofs rely precisely on this strategy: we obtain a uniform bound on our statistical estimation of the HOGS dictionary, and then take advantage of the sequential description of the boosting with empirical inner products.

- Second, in order to obtain consistent estimation with the boosting procedure, we do not need to make any coherence assumption on the dictionary (such as the RIP assumption of Candes and Tao (2007) or the weakest RE(s, c_0) assumption of Bickel et al. (2009)). Such assumptions are generally necessary to assert some consistency results for the Dantzig and Lasso procedures, such as Sparse Oracle Inequalities (SOI), for example. Nevertheless, it would be only reasonable here to impose these latter assumptions on the *theoretical* version of the HOGS although it seems difficult to deduce coherence results on the *empirical* HOGS from coherence results on the *theoretical* version of the HOGS. Our Theorem 2 below will not produce a SOI in expectation and our results will instead be expressed in probability. We will discuss the asymptotics involved in our Theorem 2 after its statement, and underline the differences with the state of the art results on the Lasso estimator.

To carry out this two-step 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 split into such two samples). We define the empirical inner product $\langle \cdot, \cdot \rangle_n$ and the empirical norm $\|\cdot\|_n$ associated with an 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 combinations of elements of B, also referred to as the linear span of B.

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

Remark 2. *The procedure could be extended to any higher order approximation, but we think that the description of the methodology for $d = 2$ provides a better understanding. We have thus 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 $\{1,\ \psi_{l_i}^i,\ l_i \in \mathbb{N}^*\}$ 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:

$$H_\emptyset^L = \mathrm{Span}\,\{1\} \quad \text{and} \quad H_i^L = \mathrm{Span}\left\{1, \psi_1^i, \cdots, \psi_L^i\right\},$$

as well as:

$$H_{ij}^L = \mathrm{Span}\left\{1, \psi_1^i, \cdots, \psi_L^i, \psi_1^j, \cdots, \psi_L^j, \psi_1^i \otimes \psi_1^j, \cdots, \psi_L^i \otimes \psi_L^j\right\},$$

where $\otimes$ denotes the tensor product between two elements of the basis. We define $H_u^{L,0}$, the approximation of H_u^0, as:

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

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

Initialization For any $1 \leq i \leq p$, define $\phi_{l_i}^i := \psi_{l_i}^i$, $l_i \in [1:L]$. Then, as a result of the orthogonality of $\{\psi_{l_i}^i,\ l_i \in \mathbb{N}\}$, we obtain $H_i^{L,0} := \mathrm{Span}\left\{\phi_1^i, \cdots \phi_L^i\right\}$. For this step, we just need the orthogonality of the constant function equal to 1 with each of the $\psi_{l_i}^i$, $l_i \in \mathbb{N}^*$. However, orthogonality is needed for the proof of the consistency of the $\mathbb{L}_2$-boosting procedure (see the proof of Lemma 4 in the Appendix).

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

$$\begin{array}{rcl} \phi_{\boldsymbol{l_{ij}}}^{ij}(X_i, X_j) & = & \phi_{l_i}^i(X_i) \times \phi_{l_j}^j(X_j) + \sum_{k=1}^L \lambda_{k,\boldsymbol{l_{ij}}}^i \phi_k^i(X_i) \\ & & + \sum_{k=1}^L \lambda_{k,\boldsymbol{l_{ij}}}^j \phi_k^j(X_j) + C_{\boldsymbol{l_{ij}}}, \end{array} \tag{2.4}$$

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:

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

We first solve the linear system:

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

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

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

As shown in Chastaing et al. (2013), $A^{\boldsymbol{l_{ij}}}$ is a definite positive Gramian matrix and (2.6) provides a unique solution in $\boldsymbol{\lambda}^{\boldsymbol{l_{ij}}}$. $C_{\boldsymbol{l_{ij}}}$ is then deduced with:

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

Higher interactions This construction can be extended to any $|u| \geq 3$. We refer the interested reader to Chastaing et al. (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 dataset $(y^r, \mathbf{x}^r)_{r=1,\cdots,n_1}$.

Algorithm 1: Empirical HOFD (EHOFD)

Input: Orthonormal system $(\psi^i_{l_i})_{l_i=0}^L$ 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 $l_i \in [1:L]$, first define $\hat{\phi}^i_{l_i,n_1} = \psi^i_{l_i}$.

- For any u such that $2 \leq |u| \leq |u_{max}|$, write the matrix $\left(\hat{A}^u_{n_1}\right)$ as well as $\left(\hat{D}^{\boldsymbol{l_u}}_{n_1}\right)$ obtained using the former expressions with $\langle \cdot, \cdot \rangle_{n_1}$.
- Solve (2.6) with the empirical inner product $\langle \cdot, \cdot \rangle_{n_1}$. Compute $\left(\hat{\boldsymbol{\lambda}}^{\boldsymbol{l_{ij}}}_{n_1}\right)$ and $\hat{C}^{n_1}_{\boldsymbol{l_{ij}}}$ with (2.7).
- The empirical version of the basis given by (2.4) is then:

$$\forall u \in [2:|u_{max}|] \quad \hat{H}^{L,0,n_1}_u = \text{Span}\left\{\hat{\phi}^u_{1,n_1}, \cdots, \hat{\phi}^u_{L^{|u|},n_1}\right\}.$$

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.

We now consider the second sample $(y^s, \mathbf{x}^s)_{s=1,\cdots,n_2}$ and we attempt to recover the unknown coefficients $\left(\beta^u_{\boldsymbol{l_u}}\right)_{\boldsymbol{l_u}, |u| \leq d}$ 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 becomes 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:

$$(\hat{\beta}^u_{\boldsymbol{l_u}}) \in \underset{\beta^u_{\boldsymbol{l_u}} \in \mathbb{R}}{\text{Argmin}} \frac{1}{n_2} \sum_{s=1}^{n_2} \left[y^s - \sum_{\substack{u \in S^* \\ |u| \leq d}} \sum_{\boldsymbol{l_u}} \beta^u_{\boldsymbol{l_u}} \hat{\phi}^u_{\boldsymbol{l_u},n_1}(\mathbf{x}^s_u) \right]^2 + \lambda J(\beta^1_1, \cdots, \beta^u_{\boldsymbol{l_u}}, \cdots), \tag{2.8}$$

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

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

Of course, such an optimization procedure is not tractable and we instead consider the relaxed $\mathbb{L}_2$-boosting (Friedman, 2001) to solve this penalized problem. Mimicking the notation of Temlyakov (2000) and Champion et al. (2013), we define the dictionary D of functions as:

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

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 by $\hat{f}$. 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 parameter $\gamma \in]0, 1]$ and number of iterations $k_{up} \in \mathbb{N}^*$.

Initialization: $G_0(\bar{f}) = 0$.

for $k = 1$ *to* k_{up} **do**

1. Select $\hat{\phi}^{u_k}_{\boldsymbol{l_{u_k}},n_1} \in \mathcal{D}$ such that

$$\left| \langle Y - G_{k-1}(\bar{f}), \hat{\phi}^{u_k}_{\boldsymbol{l_{u_k}},n_1} \rangle_{n_2} \right| = \max_{\hat{\phi}^u_{\boldsymbol{l_u},n_1} \in \mathcal{D}} \left| \langle Y - G_{k-1}(\bar{f}), \hat{\phi}^u_{\boldsymbol{l_u},n_1} \rangle_{n_2} \right|. \tag{2.9}$$

2. Compute the new approximation of $\bar{f}$ as

$$G_k(\bar{f}) = G_{k-1}(\bar{f}) + \gamma \langle Y - G_{k-1}(\bar{f}), \hat{\phi}^{u_k}_{\boldsymbol{l_{u_k}},n_1} \rangle_{n_2} \cdot \hat{\phi}^{u_k}_{\boldsymbol{l_{u_k}},n_1}. \tag{2.10}$$

end

Output: $\hat{f} = G_{k_{up}}(\bar{f})$.

For any step k, Algorithm 2 selects a function from $\mathcal{D}$ that provides sufficient information about 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 into the model, making a refinement of the greedy algorithm possible, and 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 particularly useful from a practical point of view to smooth the boosting iterations.*

An algorithm for our new sparse HOFD procedure

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

Algorithm 3: Greedy Hierarchically Orthogonal Functional Decomposition

Input: Orthonormal system $(\psi_{l_i}^i)_{l_i=0}^L$ 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 $\hat{H}_u^{L,0,n_1} := \text{Span}\left\{\hat{\phi}_{1,n_1}^u, \cdots, \hat{\phi}_{L_u,n_1}^u\right\}$ of $H_u^{L,0}$ (see Algorithm 1).

- Use an $\mathbb{L}_2$-boosting algorithm on $\mathcal{O}_2$ with the random dictionary $\mathcal{D} = \{\hat{\phi}_{1,n_1}^1, \cdots \hat{\phi}_{L,n_1}^1, \cdots, \hat{\phi}_{1,n_1}^u, \cdots, \hat{\phi}_{L_u,n_1}^u, \cdots\}$ 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-step 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 Algorithm 3 described in Section 2. To do this, we restrict our study to the case of $d = 2$ and assume that f is well approximated by first and second

order interaction components (see Remark 4 below). 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 $\boldsymbol{\beta}^0 = \left(\beta^{u,0}_{\boldsymbol{l_u}}\right)_{\boldsymbol{l_u},u}$ is the true parameter, and the functions $\left(\phi^u_{\boldsymbol{l_u}}\right)_{\boldsymbol{l_u}}$, $|u| \leq 2$ are constructed according to the HOFD described in Section 2.3. We assume that we have an 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 $\left(\hat{\phi}^u_{\boldsymbol{l_u},n_1}\right)_{\boldsymbol{l_u},u}$ described in Algorithm 1 (*resp.* for the $\mathbb{L}_2$-boosting Algorithm 2 to estimate $\left(\beta^u_{\boldsymbol{l_u}}\right)_{\boldsymbol{l_u},u}$).

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 necessary to obtain a consistent estimator from Algorithm 2.

Remark 4. *We choose the truncature order $d = 2$ in order to simplify the presentation, but it may be extended to arbitrary larger thresholds independent of the sample size n. This choice is legitimate as soon as the function f is well approximated by low interaction components and this assumption is well suited for many practical situations (Rabitz et al., 1999; Sobol, 2001). Indeed, a data-dependent choice of d_n (with $d_n \longrightarrow +\infty$ as $n \to +\infty$) would rely on a smoothness assumption on the signal f with respect to the order of the considered interactions by exploiting the size of the bias term induced by the truncature given in Theorem 5 of Sobol (2001). However, this challenging task is far beyond the scope of this paper and we have chosen to leave this problem open.*

3.1 Assumptions

We first briefly recall some notation: for all 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 large enough n. 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 the *bounded case* and is adapted to the special situation of bounded support for the random variable X, for example, 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 referred to as ($\mathbf{H_b}$) in the sequel and corresponds to the following three conditions:

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

($\mathbf{H_b^2}$) 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 matricial Hoeffding type inequality. It may be possible to slightly relax such a hypothesis using a sub-exponential tail argument. For the sake of simplicity, we have chosen to only restrict our work to the settings of ($\mathbf{H_b}$).

Regardless of the joint law of the random variables $(X_1, \ldots, X_p)$, it is always possible to build an orthonormal basis $(\phi_{l_i}^i)_{1\leq l_i\leq L}$ from a bounded (frequency truncated) Fourier basis and, therefore, ($\mathbf{H_b^1}$) is not as restrictive in practice.

Assumption ($\mathbf{H_b^2}$) deals with the high dimensional situation. We are in fact interested in practical situations where the number of variables can be much larger than the number of observations n. Hence, in our mathematical study, the number of variables p_n can grow exponentially fast with the number of observations n. This obviously implies that the collection of subsets u also depends on n and will now be denoted S_n^*. As a consequence, S_n^* also increases rapidly and is much larger than n.

Note that Hypothesis ($\mathbf{H_b^{3,\vartheta}}$) stands for a lower bound of the determinant of the Gram matrices involved in the HOFD. It is shown in Chastaing et al. (2013) that each of these Gram matrices is invertible and, as a result, each $\det(A^{ij})$ is

positive. Nevertheless, if $\vartheta = 0$, this hypothesis assumes 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 $(\mathbf{H}_{\mathbf{b}}^{\mathbf{3},\vartheta})$ drastically relaxes the case $\vartheta = 0$ and becomes very weak. The verification of $(\mathbf{H}_{\mathbf{b}}^{\mathbf{3},\vartheta})$ requires the computation of an order of p_n^2 determinants of size $L^2 \times L^2$. We have checked this assumption in our experiments. However, for very large values of n, this may become impossible from a numerical point of view.

In the following, the parameters ϑ and ξ will be related to each other and we will obtain a consistency result of the sparse HOFD up to the condition $\vartheta < \xi/2$. This constraint implicitly limits the size of p_n since $\log p_n = \underset{n\to+\infty}{\mathcal{O}}(n^{1-\xi})$.

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

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

Sparsity Assumption $(\mathbf{H}_{\mathbf{s},\alpha})$ The last assumption concerns the sparse representation of the unknown signal described by Y in the basis $(\phi_{\boldsymbol{l_u}}^u(\mathbf{X}_u))_u$. Such a hypothesis will be useful to assess the statistical performance of the $\mathbb{L}_2$-boosting and will be referred to as $(\mathbf{H}_{\mathbf{s},\alpha})$ below. It is legitimate due to our high dimension setting and our motivation to identify the main interactions $\mathbf{X}_u$.

$(\mathbf{H}_{\mathbf{s},\alpha})$ There exists $\alpha > 0$ such that the parameter $\boldsymbol{\beta}^0$ satisfies:

$$\|\beta^0\|_{\ell_1} := \sum_{\substack{u\in S_n^* \\ |u|\le d}} \sum_{\boldsymbol{l_u}} \left|\beta_{\boldsymbol{l_u}}^{u,0}\right| = \underset{n\to+\infty}{\mathcal{O}}(n^\alpha).$$

3.2 Main results

We recall below that $\|.\|$ is the $\mathbb{L}_2$ norm on functions decomposed in the orthonormal basis $(\phi_{\boldsymbol{l_u}}^u)_u$. We first provide our main result on the efficiency of the EHOFD (Algorithm 1).

Theorem 1. *Assume that* $(\mathbf{H}_{\mathbf{b}})$ *holds with* ξ *(resp.* ϑ*) given by* $(\mathbf{H}_{\mathbf{b}}^{\mathbf{2}})$ *(resp.* $(\mathbf{H}_{\mathbf{b}}^{\mathbf{3},\vartheta})$*) and that there exists a constant* Λ *such that* $\left\|\boldsymbol{\lambda}^{\boldsymbol{l_{ij}}}\right\|_2 \le \Lambda$ *for any couple*

(i,j). *Then, if* $\vartheta < \xi/2$, *the sequence of estimators* $\left(\hat{\phi}^u_{\boldsymbol{l_u},n_1}\right)_u$ *satisfies:*

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

The proof of this theorem can be found in the Appendix.

Let us mention the contribution of Theorem 1 compared to the results obtained in Chastaing et al. (2013). Proposition 5.1 of Chastaing et al. (2013) leads to an almost sure convergence of their estimator without any quantitative rate when the number of functions in the HOFD is kept fixed and does not grow with the number of observations n. In contrast, in our high dimensional paradigm, we allow S_n^* to grow with n and also obtain an almost sure result associated with a convergence rate. This will be essential for the derivation of our next result.

Our second main result concerns the $\mathbb{L}_2$-boosting that 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 assumption $(\mathbf{H}_{\mathbf{s},\alpha})$ holds. We assume that

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

where $\boldsymbol{\beta}^0 = (\beta^{u,0}_{\boldsymbol{l_u}})_{\boldsymbol{l_u},u}$ is the true parameter that expands $\tilde{f}$. To the best of our knowledge, such a high dimensional inference with noise in the variables appears to be novel. As already pointed out above, the greedy boosting seems to be a well tailored approach to handle noisy dictionaries in comparison to a penalized regression strategy, which relies on a somewhat unverifiable "RIP-type" hypothesis on the learned dictionary.

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 down into* $\mathcal{O}_1 \cup \mathcal{O}_2$. *Assume that functions* $\left(\hat{\phi}^u_{\boldsymbol{l_u},n_1}\right)_{\boldsymbol{l_u},u}$ *are estimated from the first sample* $\mathcal{O}_1$ *under* $(\mathbf{H_b})$ *with* $\vartheta < \xi/2$, *and that there exists a constant* Λ *such that* $\left\|\boldsymbol{\lambda}^{\boldsymbol{l_{ij}}}\right\|_2 \leq \Lambda$ *for any couple* (i,j).
$\hat{f}$ *is then defined by (2.10) of Algorithm 2 on* $\mathcal{O}_2$ *as:*

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

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

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

In particular, for Gaussian noises that possess arbitrary large moments, the constraint on q disappears and Theorem 2 can be applied as soon as $\xi < 1$.

Let us discuss the asymptotic setting involved in our Theorem. First, our result is a result in probability, rather than in expectation. It is a frequently encountered fact that SOI in expectation are derived with additional assumptions on the coherence of the dictionary; some detailed discussions can be found in Bickel et al. (2009) and Rigollet and Tsybakov (2011). With some coherence and boundedness assumptions, Bickel et al. (2009) deduced convergence rates of the Lasso estimator in expectation as soon as:

$$\|\beta^0\|_{\ell_0} \frac{\log(p)}{n} \longrightarrow 0. \tag{3.1}$$

Later, Rigollet and Tsybakov (2011) extended the study of the Lasso behavior with a result on the Lasso estimator on bounded variables without any coherence assumption and showed a consistency result in probability when:

$$\|\beta^0\|_{\ell_1} \sqrt{\frac{\log(p)}{n}} \longrightarrow 0. \tag{3.2}$$

Hence, the rate is damaged by the appearance of the $\sqrt{\cdot}$ in (3.2) in comparison with (3.1). Concerning the boosting algorithm, Champion et al. (2013) also obtained consistency results in probability under the asymptotic setting given in (3.2) without a coherence assumption. It should be observed that our results with a noisy dictionary requires that

$$\left(\inf_{i,j} \det(A^{ij}) \right)^{-1} \|\beta^0\|_{\ell_1}^2 \sqrt{\frac{\log p}{n}} \longrightarrow 0 \,\text{as}\, n \longrightarrow +\infty, \tag{3.3}$$

which is a stronger assumption in comparison with (3.2). From a technical point of view, the asymptotic setting is due to inequality (S4.10) where $\|\beta^0\|_{\ell_1}^2 \zeta_n$ appears instead of $\|\beta^0\|_{\ell_1} \zeta_n$ for boosting algorithms without noise on the variables (see the proof of Theorem 2 in the Appendix).

In favorable cases where all linear systems defined through the Gram matrices A^{ij} are well conditioned, $\vartheta = 0$ and the condition becomes $\|\beta^0\|_{\ell_1}^2 \sqrt{\frac{\log p}{n}} \longrightarrow 0$, and there is still a price to pay for the preliminary estimation of the elements of the HOGS. Theorem 2 can be applied only for sequences of coefficients such that $\|\beta^{u,0}_{\boldsymbol{l_u}}\|_{L_1} \lesssim n^{1/4}$. Note also that the degeneracy of the Gram determinants must be strictly larger than $n^{-1/2}$. For example, when $\vartheta = 1/4$, the norm $\|\beta^{u,0}_{\boldsymbol{l_u}}\|_{L_1}$ cannot be larger than $n^{1/8}$.

We briefly describe the proof below and provide the technical details in the Appendix.

Sketch of Proof of Theorem 2. Mimicking the scheme of Bühlmann (2006) and Champion et al. (2013), the proof first consists in defining the theoretical residual of Algorithm 2 at step k as:

$$\begin{aligned} R_k(\bar{f}) &= \bar{f} - G_k(\bar{f}) \\ &= \bar{f} - G_{k-1}(\bar{f}) - \gamma \langle Y - G_{k-1}(\bar{f}), \hat{\phi}^{u_k}_{\boldsymbol{l_{u_k}},n_1} \rangle_{n_2} \cdot \hat{\phi}^{u_k}_{\boldsymbol{l_{u_k}},n_1} \end{aligned} \tag{3.4}$$

Furthermore, following the work of Champion et al. (2013), we introduce a *phantom* residual in order to reproduce the behavior 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 (2.9) and (2.10), but updated using the deterministic inner product. The *phantom* residuals $\tilde{R}_k(\bar{f})$, $k \geq 0$, are defined as follows:

$$\left\{ \begin{array}{l} \tilde{R}_0(\bar{f}) = \bar{f} \\ \tilde{R}_k(\bar{f}) = \tilde{R}_{k-1}(\bar{f}) - \gamma \langle \tilde{R}_{k-1}(\bar{f}), \hat{\phi}^{u_k}_{\boldsymbol{l_{u_k}},n_1} \rangle \hat{\phi}^{u_k}_{\boldsymbol{l_{u_k}},n_1}, \end{array} \right. \tag{3.5}$$

where $\hat{\phi}^{u_k}_{\boldsymbol{l_{u_k}},n_1}$ has been selected with Equation (2.9) 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:

$$\left\| \hat{f} - \tilde{f} \right\| = \left\| G_{k_n}(\bar{f}) - \tilde{f} \right\| \leq \left\| \bar{f} - \tilde{f} \right\| + \left\| R_{k_n}(\bar{f}) - \tilde{R}_{k_n}(\bar{f}) \right\| + \left\| \tilde{R}_{k_n}(\bar{f}) \right\|. \tag{3.6}$$

We then have to show that each term on the right-hand side of (3.6) converges towards zero in probability. □

4. Numerical Applications

This section is devoted to the numerical efficiency of the two-step procedure given in Section 2, and primarily focuses on the practical use of the HOFD through sensitivity analysis (SA). SA aims to identify the most contributive variables to the variability of a regression model (Saltelli et al., 2000; Cacuci et al., 2005). The most common quantification is a variance-based index, known as the Sobol index (Sobol, 1993). This measure relies on the Hoeffding decomposition that provides an elegant and meaningful theoretical framework when inputs are known to be independent. However, as mentioned in the introduction, the interpretation of such indices may be irrelevant when strong dependencies arise. The HOFD presented in Section 2.2 is of great interest in this situation because it provides a general and rigorous multivariate regression extension that can be used to define sensitivity indices well-tailored to dependent inputs. As detailed in Chastaing et al. (2012), the model variance can be expanded as follows:

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

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

$$S_u = \frac{V(f_u(\mathbf{X}_u)) + \sum_{u \cap v \neq u,v} \mathrm{Cov}(f_u(\mathbf{X}_u), f_v(\mathbf{X}_v))}{V(Y)}. \tag{4.1}$$

Furthermore, we deduce the empirical estimation of (4.1) once we have applied the procedure described in Algorithm 3 to obtain $(\hat{f}_u, \hat{f}_v, u \cap v \neq u, v)$.

4.1 Description

We end this paper with a short simulation study, focused primarily on the performance of the greedy selection algorithm for the prediction of generalized sensitivity indices. Since the estimation of these indices consists in estimating the summands of the generalized functional ANOVA decomposition (referred to as HOFD), we begin by constructing a hierarchically orthogonal system of functions to approximate the components. As pointed out 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. For each situation, we therefore measured the degeneracy of the matrices involved, given by:

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

We then 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 (referred to as FoBa below), and the Lasso estimator. As pointed out above, we have an n-sample of i.i.d. observations $(y^s, \mathbf{x}^s)_{s=1,\cdots,n}$ broken down 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. The second sample is used to solve the penalized regression problem given by (2.8) and illustrated here:

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

We will now briefly describe how we use the Lasso, the FoBa and the Boosting.

4.2 Feature selection Algorithms

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

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

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 used for HOFD in Chastaing et al. (2013). We refer to these references for a more in-depth description of this algorithm. This procedure depends on two shrinkage parameters, ϵ and δ. The parameter ϵ is the stopping criterion that predetermines if a large number of predictors is going to be introduced into the model. The second parameter, $\delta \in]0,1]$, offers a flexibility in the *backward* step since it allows the algorithm to smoothly eliminate a predictor at each step.

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

Calibration of the Boosting As previously reported by Champion et al. (2013), we fixed the shrinkage parameter to $\gamma = 0.7$ since it provides a suitable value for high dimensional regression, even though we did not find any extreme differences when γ varies in $[0.5; 1[$. Since the optimal value for k_{up} is unknown in practice, we use a C_p Mallows-type criterion to fix the optimal number of iterations. This stopping criterion is much more important than the choice of the shrinkage parameter. It is, of course, induced by γ since it depends on the sequence of the boosting iterations.

Like in the LARS algorithm, we follow the recommendations of Efron et al. (2004) to select the best solution. 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 $\hat{\boldsymbol{\beta}}(k)$. On the basis of 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 $\hat{\phi}^u_{\boldsymbol{l_u}, n_1}$ has been selected through the first k steps of the algorithm. Finally, we choose the optimal number of selected functions $\hat{k}_{\text{up}}$ such that:

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

Lasso algorithm Since 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 $\lambda \times \ell_1$ strategy that yields the Lasso estimator for a given penalization parameter $\lambda > 0$, *i.e.*,

$$J(\beta_1^1, \cdots, \beta^u_{\boldsymbol{l_u}}, \cdots) = \sum_{\substack{u \in S_n^* \\ |u| \leq d}} \sum_{\boldsymbol{l_u}} \left| \beta^u_{\boldsymbol{l_u}} \neq 0 \right|.$$

Several algorithms have been proposed in the literature to solve the Lasso regression. One of the most popular is the LARS method, described in Efron et al. (2004), because it performs a solution that coincides with the theoretical regularization path $\{\hat{\boldsymbol{\beta}}(\lambda),\ \lambda \in \mathbb{R}^+\}$. However, the LARS strategy is very expensive in large Lasso problems. To make a good numerical comparison with the greedy algorithms, we choose to perform a coordinate descent algorithm proposed by Fu

(1998), and Friedman et al. (2007) because of its low computational cost compared to the LARS implementation. The tuning parameter λ is first selected by generalized cross-validation, and the Lasso Coordinate Descent (LCD) algorithm is performed with the R `lassoshooting` package.

4.3 Datasets

Each experiment on each dataset was randomly reproduced 50 times to compute the Monte-Carlo errors. Since each dataset has very few instances, the size L of the initial orthonormal systems has to be small. Here, we arbitrarily choose $5 \leq L \leq 8$ and the approximation performance do not suffer from the sensitivity of L in these models.

First Dataset: 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 experiment, 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 eight 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 eight Fourier basis functions.

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

Second Dataset: the g-Sobol function This function is referred to in Saltelli et al. (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

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

Here, we take $p = 25$ and $a = (0, 0, 0, 1, 1, 2, 3, 4.5, 4.5, 4.5, 9, 9, 9, 9, 9, 99, \cdots, 99)$. For the construction of the hierarchical basis functions, we choose the first five Legendre polynomials ($L = 5$). We use $n = 2000$ evaluations of the model and the number of predictors $m_n = pL + \binom{p}{2}L^2 = 7625$, which clearly exceeds the sample size n.

Third dataset: dependent inputs The third data set stands for a rarely investigated situation, where the inputs are correlated. As proposed by Mara and Tarantola (2012), we generate a sample set according to the following distribution: X_1 and X_2 are uniformly sampled in the set $\mathcal{S}$:

$$\mathcal{S} := \left\{(x_1, x_2) \in [-1, 1]^2 \,|\, 2x_1^2 - 1 \leq x_2 \leq 2x_1^2\right\}.$$

Furthermore, X_3 is also sampled uniformly in $[-1; 1]$. Then, Y is built following

$$Y = X_1 + X_2 + X_3.$$

The inputs X_1 and X_2 are clearly not independent and we do not exactly know the analytical Sobol indices. We choose $n = 100$ observations, with the first six Legendre basis functions ($L = 6$).

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 indicated in Figure 4.1. The von Mises stress makes it possible to predict material yielding that occurs when the material yield strength is reached. 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. In order to provide a large panel of tanks able to resist the internal pressure, a manufacturer wants to know the parameters that contribute the most to the von Mises criterion variability. In the model that 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 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. There exists some strong correlations between some of the inputs of the system owing to the constraints of manufacturing processes, for instance between the shell radius and its thickness. The system is modeled by a 2D finite element ASTER code. Input distributions are provided in Table 4.1.

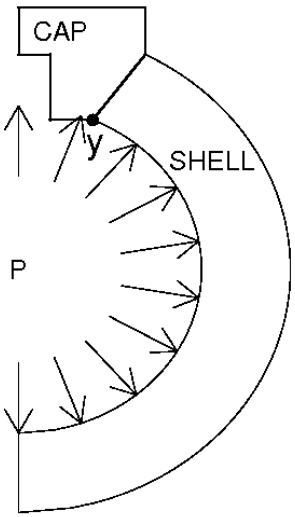


Figure 4.1: Tank distortion at point y

The geometrical parameters are uniformly distributed because of the large choice left for tank construction. The correlation γ between the geometrical parameters is induced by the constraints linked to manufacturing processes. The physical inputs are normally distributed and their uncertainty is due to the manufacturing process and the properties of the elementary constituent variabilities. The large variability of P_{int} in the model corresponds to the different internal pressure values that 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 do $n = 1000$ simulations. We use the first Hermite basis functions, whose maximum degree is 5 for every parameter.

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}	$\alpha N(\mu, \Sigma) + (1-\alpha)N(\mu, \Omega)$
$\sigma_{y,cap}$	$\alpha = 0.02$, $\mu = \begin{pmatrix} 210 \\ 500 \end{pmatrix}$, $\Sigma = \begin{pmatrix} 350 & 0 \\ 0 & 29 \end{pmatrix}$, $\Omega = \begin{pmatrix} 175 & 81 \\ 81 & 417 \end{pmatrix}$
E_{shell}	$\alpha N(\mu, \Sigma) + (1-\alpha)N(\mu, \Omega)$
$\sigma_{y,shell}$	$\alpha = 0.02$, $\mu = \begin{pmatrix} 70 \\ 300 \end{pmatrix}$, $\Sigma = \begin{pmatrix} 117 & 0 \\ 0 & 500 \end{pmatrix}$, $\Omega = \begin{pmatrix} 58 & 37 \\ 37 & 250 \end{pmatrix}$
P_{int}	$N(80, 10)$

Table 4.1: Description of inputs of the shell model

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 and "LCD" refers to the Lasso coordinate descent method. Our method is, of course, referred to 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 functions. 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). For the sake of clarity, Figure 4.4 only represents the first ten sensitivity indices. We can also draw similar conclusions with Figure 4.4, where the three methods lead to the same conclusion. It should also be noted that the standard deviations of each method seem to be relatively equivalent. Figure 4.5 represents the estimated sensitivity indices when the inputs are correlated. The analytical results

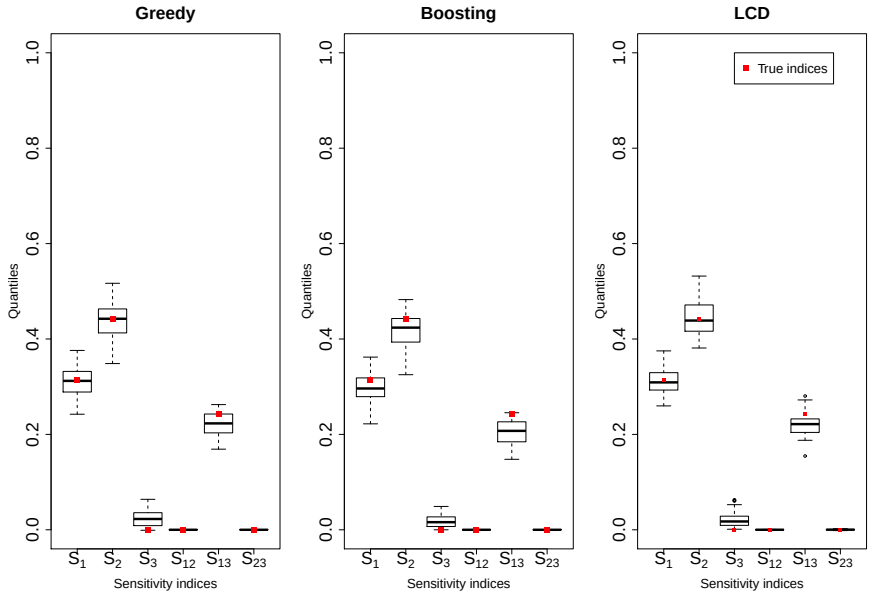


Figure 4.2: Representation of the first-order components on the First dataset (Ishigami function) described through the Legendre basis.

are obviously unknown, but we obtain similar results for the three methods.

Finally, as illustrated in Figure 4.6, the most contributive parameter to the von Mises criterion variability is the internal pressure P_{int}, which is not surprising. Concerning the geometric characteristics, the main parameters of the three methods are cap thickness, T_{cap}, and shell thickness, T_{shell}, using their expensive code, although the shell internal radius does not seem to be that important.

Computation time and accuracy The performances of the three methods are illustrated in Table 4.2, on the basis of their computational cost and the accuracy of the feature selection.

Regarding the statistical accuracy, it should be noted that each estimator of high dimensional regression possesses a comparable dispersion on all the datasets and performs quite similarly on the first dataset. The Lasso estimator seems a little bit unprecise in the third data-set in comparison with the FoBa and Boosting methods. At last, the LCD method is also outperformed on the third data-set (with dependent inputs): it selects a significantly larger number of sensitivity in-

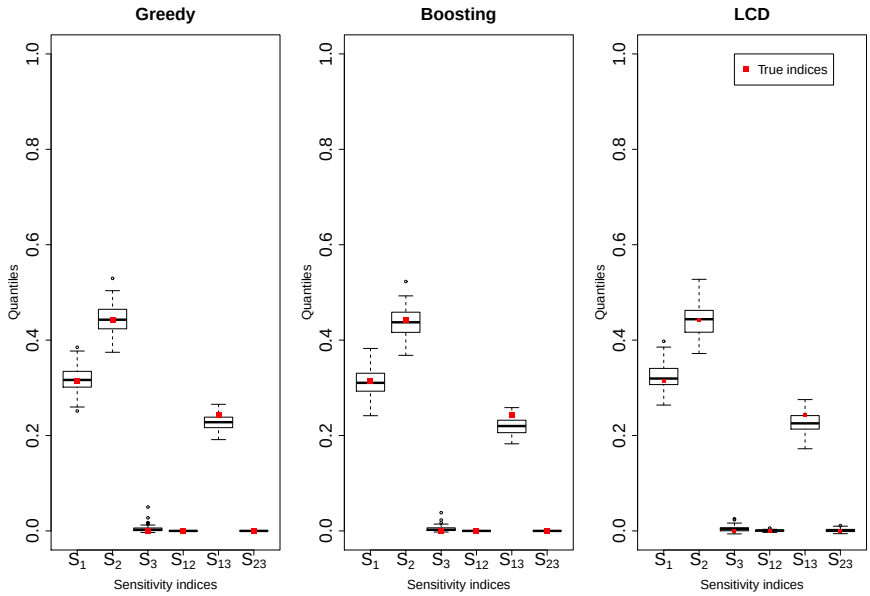


Figure 4.3: Representation of the first-order components on the First dataset (Ishigami function) described through the Fourier basis.

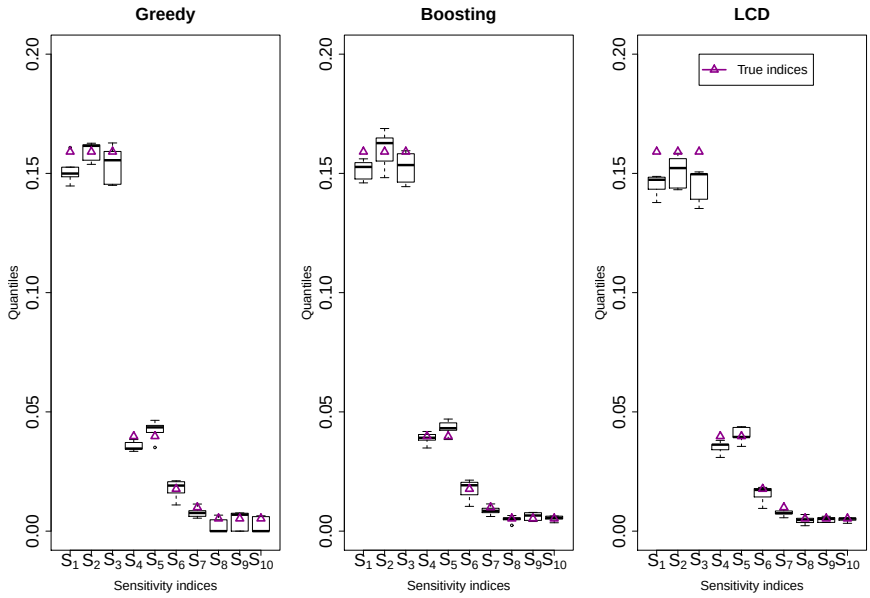


Figure 4.4: Representation of the first-order components on the Second dataset (g-Sobol function).

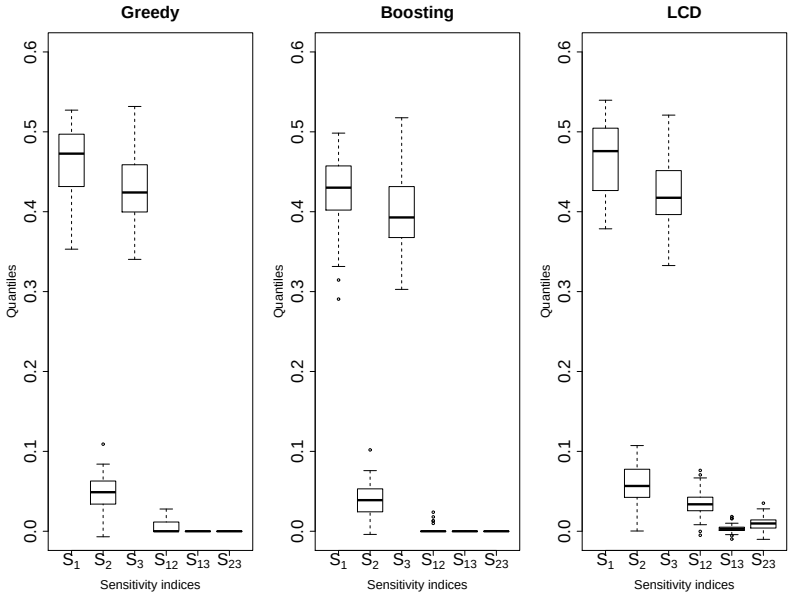
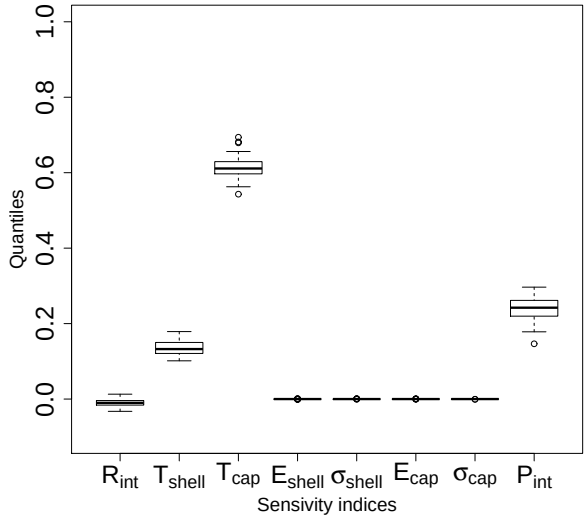


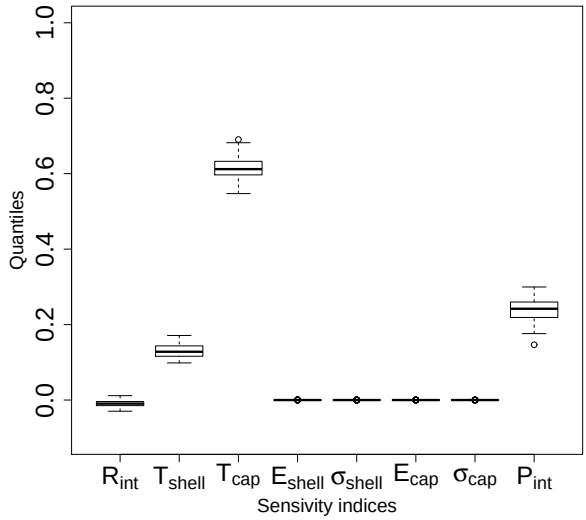
Figure 4.5: Representation of the first-order components on the third dataset (dependent inputs).

dices in comparison with Boosting and FoBa methods (for instance, the indices S_{13} and S_{23} are certainly equals to 0 owing to the definition of Y). This may be due to the influence of the dependency among the inputs X_1 and X_2 in this data-set on the Lasso estimator.

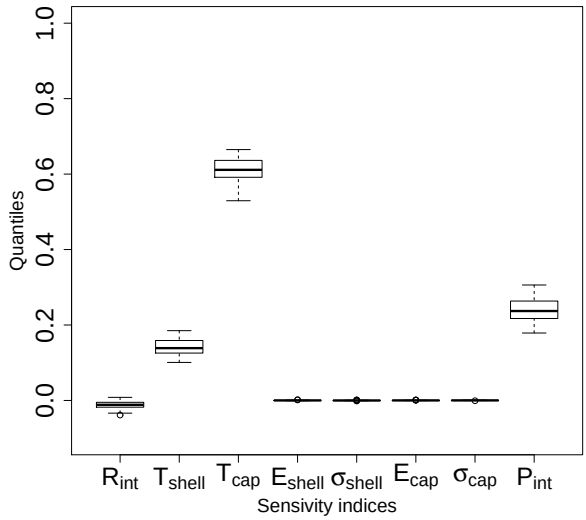
Furthermore, it clearly appears in Table 4.2 that our proposed $\mathbb{L}_2$-boosting is the fastest method. This is particularly true on the 25-dimension g-Sobol function where the fraction of additional time required by the LCD algorithm in comparison to the $\mathbb{L}_2$-boosting is about 100. Although we do not have access to the theoretical support recovery $\|\boldsymbol{\beta}\|_0$, we can observe that the results of the $\mathbb{L}_2$-boosting are equivalent to those of other algorithms in terms of its feature selection ability. Hence, for the same degree of accuracy, our method seems to be much faster.



(a) Greedy



(b) Boosting



(c) LCD

Figure 4.6: Dispersion of the first order sensitivity indices of the tank model parameters for the 3 methods.

Dataset	Procedure	$\left\|\hat{\boldsymbol{\beta}}\right\|_0$	Elapsed Time (in sec.)
Ishigami function Case 1	$\mathbb{L}_2$-boosting	19	0.0941
	FoBa	21	2.2917
	LCD	20	2.25
Ishigami function Case 2	$\mathbb{L}_2$-boosting	15	0.0884
	FoBa	12	1.0752
	LCD	13.9	0.41
g-Sobol function	$\mathbb{L}_2$-boosting	99	49.8
	FoBa	22.4	827.9
	LCD	91.8	5047.4
Dependent inputs	$\mathbb{L}_2$-boosting	4.14	0.028
	FoBa	4.76	0.1056
	LCD	24.1	0.061
Tank pressure model	$\mathbb{L}_2$-boosting	10	0.0266
	FoBa	22	0.3741
	LCD	23	0.15

Table 4.2: Features of the three algorithms

Note that we have computed the maximal "degeneracy" that is involved in the resolution of the linear systems and quantified by Assumption ($\mathbf{H_b^{3,\vartheta}}$) in 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 ($\mathbf{H_b^2}$) 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.

Dataset	Degeneracy $d(A)$	$\vartheta \geq \frac{\log(1/d(A))}{log(n)}$	$n^{1-2\vartheta}$	$\log(m_n)$
Ishigami function Case1	0.6388	$[0.0786, +\infty[$	122.3821	6.0113
Ishigami function Case1	0.76	$[0.0481, +\infty[$	173.3094	6.0113
g-Sobol function	0.9745	$[0.0034, +\infty[$	1899	8.9392
Dependent inputs	0.628	$[0.101, +\infty[$	39.4457	4.8363

Table 4.3: Degeneracy of the linear systems and admissible size of m_n ($n^{1-2\vartheta}$ should be greater than $\log(m_n)$.

5. Conclusions and Perspectives

This paper provides a rigorous framework for the hierarchically orthogonal Gram-Schmidt procedure in a high-dimensional paradigm, with the use of the greedy $\mathbb{L}_2$-boosting. Overall, the procedure falls into the category of sparse estimation with a noisy dictionary, and we demonstrate its consistency up to some mild assumptions on the structure of the real underlying basis. 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 leave this development open for a future study, which could be based either on the development of a concentration inequality for unbounded random matrices or on a truncating argument. It also appears that our algorithm produces very satisfactory numerical results through our three datasets as a result of its very low computational cost. It can also be extended with some further numerical work to a larger truncation order of $d \geq 3$. Such an improvement may also be of interest from a theoretical point of view when dealing with a function that smoothly depends on the interaction order. In particular, a data-driven adaptive choice of d may be of practical interest in the future.

Bibliography

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.*, 37(4):1705–1732.

Blatman, G. (2009). *Adaptive sparse polynomial chaos expansions for uncertainty propagation and sensitivity analysis*. PhD thesis, Université BLAISE PASCAL - Clermont II.

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.

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.

Candes, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.*, 35(6):2313–2351.

Cavalier, L. and Hengartner, N. W. (2005). Adaptive estimation for inverse problems with noisy operators. *Inverse Problems*, 21(4):1345–1361.

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

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`.

Crestaux, T., Le Maître, O., and Martinez, J. (2009). Polynomial chaos expansion for sensitivity analysis. *Reliability Engineering & System Safety*, 94(7):1161–1172.

Da Veiga, S., Wahl, F., and Gamboa, F. (2009). Local polynomial estimation for sensitivity analysis on models with correlated inputs. *Technometrics*, 51(4):452–463.

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. *Annals of Statistics*, 29(5):1189–1232.

Friedman, J., Hastie, T., Höfling, H., and Tibshirani, R. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332.

Fu, W. (1998). Penalized regressions: the bridge versus the lasso. *Journal of computational and graphical statistics*, 7(3):397–416.

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.

Jacques, J., Lavergne, C., and Devictor, N. (2006). Sensitivity analysis in presence of model uncertainty and correlated inputs. *Reliability Engineering and System Safety*, (91):1126–1134.

Li, G. and Rabitz, H. (2010). D-morph regression: application to modeling with unknown parameters more than observation data. *Journal of mathematical chemistry*, 48(4):1010–1035.

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.

Mara, T. and Tarantola, S. (2012). Variance-based sensitivity analysis of computer models with dependent inputs. *Reliability Engineering and System Safety*, 107:115–121.

Rabitz, H., Ali, O., Shorter, J., and Shim, K. (1999). Efficient input-output model representations. *Computer Physics Communications*, 117(1):11–20.

Rigollet, P. and Tsybakov, A. (2011). Exponential screening and optimal rates of sparse estimation. *Ann. Statist.*, 39(2):731–771.

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

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

Sobol, I. (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.

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

Wang, X. and Fang, K.-T. (2003). The effective dimension and quasi-Monte Carlo integration. *J. Complexity*, 19(2):101–124.

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

Zuniga, M., Kucherenko, S., and Shah, N. (2013). Metamodelling with independent and dependent inputs. *Computer Physics Communications*, 184(6):1570–1580.

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

gal.chastaing@gmail.com

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

$\mathbb{L}_2$-Boosting on a 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*

Supplementary Material

We present here the proofs of Theorem 1 and Theorem 2 of the main document. Section S1 sets the notation that will be used all along the document. Section S2 quotes a concentration inequality on random matrices that will be exploited in the rest of the work. We develop the proofs of Theorem 1 and 2 in Section S3 and Section S4.

S1 Notation and reminder

Let us first recall some standard notations on matricial norms. For any square matrix M, its spectral radius $\rho(M)$ will refer to as 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

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

where ${}^tM$ 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

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

where $Tr(M)$ is the trace of the matrix M.

S2 Hoeffding 's type Inequality for random bounded matrices

For the sake of completeness, we quote here Theorem 1.3 of Tropp (2012). 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:

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

Theorem 1 (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*

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

Then, for all $t \geq 0$

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

In our work, a more precise concentration inequality such as the Bernstein one (see Theorem 6.1 of Tropp (2012)) is useless 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.

S3 Proof of Theorem 1

Consider any subset $u = (u_1, ..., u_t) \in S_n^*$ with $t \geq 1$ and remark that if $u = \{i\}$, i.e. $t = 1$, the *Initialization* of Algorithm 1 is such 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 $\boldsymbol{l_{ij}} = (l_i, l_j) \in [1:L]^2$, and 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:

$$\begin{array}{l} \langle \phi^{ij}_{\boldsymbol{l_{ij}}}, \phi^i_k \rangle = 0, \quad \forall\, k \in [1:L] \\ \langle \phi^{ij}_{\boldsymbol{l_{ij}}}, \phi^j_k \rangle = 0, \quad \forall\, k \in [1:L] \\ \langle \phi^{ij}_{\boldsymbol{l_{ij}}}, 1 \rangle = 0. \end{array} \tag{S3.1}$$

When removing $C_{\boldsymbol{l_{ij}}}$, the resolution of (S3.1) leads to the resolution of a linear system of the type:

$$A^{ij}\boldsymbol{\lambda}^{\boldsymbol{l_{ij}}} = D^{\boldsymbol{l_{ij}}}, \tag{S3.2}$$

with $\boldsymbol{\lambda}^{\boldsymbol{l_{ij}}} = {}^t\left(\lambda^i_{1,\boldsymbol{l_{ij}}} \cdots \lambda^i_{L,\boldsymbol{l_{ij}}} \lambda^j_{1,\boldsymbol{l_{ij}}} \cdots \lambda^j_{L,\boldsymbol{l_{ij}}}\right)$ 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 $\hat{\phi}^{ij}_{\boldsymbol{l_{ij}},n_1}$ that is decomposed on the dictionary as follows:

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

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:

$$\begin{array}{ll} \langle \hat{\phi}^{ij}_{\boldsymbol{l_{ij}},n_1}, \phi^i_k\rangle_{n_1} = 0, & \forall\ k \in [1:L] \\ \langle \hat{\phi}^{ij}_{\boldsymbol{l_{ij}},n_1}, \phi^j_k\rangle_{n_1} = 0, & \forall\ k \in [1:L] \\ \langle \hat{\phi}^{ij}_{\boldsymbol{l_{ij}},n_1}, 1\rangle_{n_1} = 0. & \end{array} \tag{S3.3}$$

When removing $\hat{C}^{n_1}_{\boldsymbol{l_{ij}}}$, the resolution of (S3.3) can also lead to the resolution of a linear system of the type:

$$\hat{A}^{ij}_{n_1}\hat{\boldsymbol{\lambda}}^{\boldsymbol{l_{ij}}}_{n_1} = \hat{D}^{\boldsymbol{l_{ij}}}_{n_1}, \tag{S3.4}$$

where $\hat{\boldsymbol{\lambda}}^{\boldsymbol{l_{ij}}}_{n_1} = {}^t\left(\hat{\lambda}^i_{1,\boldsymbol{l_{ij}},n_1} \cdots \hat{\lambda}^i_{L,\boldsymbol{l_{ij}},n_1} \hat{\lambda}^j_{1,\boldsymbol{l_{ij}},n_1} \cdots \hat{\lambda}^j_{L,\boldsymbol{l_{ij}},n_1}\right)$ 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 1. *Remark that each 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 the 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* $(\mathbf{H_b^2})$*, one has*

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

Proof. First consider a couple (i,j) and note that $\left\| \hat{A}_{n_1} - A \right\|_2 = \rho(\hat{A}_{n_1} - A)$, since $\hat{A}_{n_1} - A$ is self-adjoint. To obtain a concentration inequality on the matricial norm $\left\| \hat{A}_{n_1} - A \right\|_2$, we use the result of Tropp (2012) (see Theorem 1), which give concentration inequalities for the largest eigenvalue of self-adjoint matrices (see section 6.2).

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

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

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

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

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 as Theorem 1.3 of Tropp (2012) to our sequence $\Theta_{1,ij}, \cdots, \Theta_{n_1,ij}$, with $\sigma^2 = 16 n_1 L^2 M^4$, we then obtain that

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

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

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

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

$$P\left(\sup_{1\leq i,j\leq p_n} \rho\left(\hat{A}_{n_1} - A\right) \geq \gamma n^{-\xi/2}\right) \leq 2Lp_n^2 e^{-\frac{{n_1}^{1-\xi}\gamma^2}{128L^2M^4}}. \tag{S3.5}$$

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 (S3.5). □

Similarly, we can show that the estimated quantity $\hat{D}_{n_1}$ is not 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 $\left\|\hat{D}_{n_1} - D\right\|_2$. Remark that $\left\|\hat{D}_{n_1} - D\right\|_2$ can be written as:

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

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

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

which gives:

$$P\left(\sup_{i,j,\boldsymbol{l_{ij}}} \left\|\hat{D}_{n_1} - D\right\|_2 \geq \gamma {n_1}^{-\xi/2}\right) \leq 4L \times L^2 p_n^2 \exp\left(-\frac{1}{8}\frac{\gamma^2 {n_1}^{1-\xi}}{M^6 + M^3\gamma/6{n_1}^{-\xi/2}}\right). \tag{S3.6}$$

Now, since $n_1 = n/2$, Assumption $(\mathbf{H_b^2})$ implies that the right-hand side of Inequality (S3.6) 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* $(\mathbf{H_b})$ *with* $\vartheta < \xi/2$*, we have*

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

Proof. Fix any couple (i,j), $\boldsymbol{\lambda}$ and $\hat{\boldsymbol{\lambda}}_{n_1}$ satisfy Equations (S3.2) and (S3.4). Hence,

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

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

$$\left(\mathrm{I} - A^{-1}(A - \hat{A}_{n_1})\right)(\hat{\boldsymbol{\lambda}}_{n_1} - \boldsymbol{\lambda}) = A^{-1}(A - \hat{A}_{n_1})\boldsymbol{\lambda} + A^{-1}(\hat{D}_{n_1} - D), \tag{S3.7}$$

Remark that $\left\| \hat{A}_{n_1} - A \right\|_2 = \mathcal{O}_P(n^{-\xi/2})$ 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 (S3.7) 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,

$$\begin{array}{rcl} \left\| \hat{\boldsymbol{\lambda}}_{n_1} - \boldsymbol{\lambda} \right\|_2 &\leq& \left\| \left(\mathrm{I} - A^{-1}(A - \hat{A}_{n_1})\right)^{-1} \right\|_2 \\ && \times \left(\left\| A^{-1}[A - \hat{A}_{n_1}] \right\|_2 \|\boldsymbol{\lambda}\|_2 + \left\| A^{-1}(\hat{D}_{n_1} - D) \right\|_2 \right) \\ &\leq& \left\| \left(\mathrm{I} - A^{-1}(A - \hat{A}_{n_1})\right)^{-1} \right\|_2 \\ && \times \left(\left\| A^{-1} \right\|_2 \left\| A - \hat{A}_{n_1} \right\|_2 \|\boldsymbol{\lambda}\|_2 + \left\| A^{-1} \right\|_2 \left\| \hat{D}_{n_1} - D \right\|_2 \right). \end{array} \tag{S3.8}$$

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

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

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:

$$\begin{array}{rcl} \left|\left|\left|A^{-1}\right|\right|\right|_2 & \leq & \max_{(i,j)\in[1:p_n]^2} \frac{2LM^{4L-2}}{det(A^{i'j'})} \\ & \leq & 2C^{-1}LM^{4L-2}n^{\vartheta}. \end{array} \tag{S3.9}$$

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

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

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 that

$$\sup_{i,j,\boldsymbol{l_{ij}}} \left|\left|\left|\left(\mathrm{I} - A^{-1}(A - \hat{A}_{n_1})\right)^{-1}\right|\right|\right|_2 \leq 1 + 2LM^{4L-2}C^{-1}\mathcal{O}_P(n^{\vartheta-\xi/2}). \tag{S3.10}$$

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 (S3.9), (S3.10) and the assumption on the uniform bound $\|\boldsymbol{\lambda}\|_2 < \Lambda$ over all the couples (i,j) for the norm $\left\|\boldsymbol{\lambda}^{l_{ij}}\right\|_2$. □

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 $(\mathbf{H^1_b})$, these independent random variables are bounded by M^2 and

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

Under Assumption $(\mathbf{H^2_b})$, 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,

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

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

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

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

$$\begin{aligned} \left\|\hat{\phi}^{ij}_{\boldsymbol{l_{ij}},n_1} - \phi^{ij}_{\boldsymbol{l_{ij}}}\right\| &\leq \sqrt{2\left[\sum_{k=1}^{L}(\hat{\lambda}^i_{k,n_1}-\lambda^i_k)^2 + \sum_{k=1}^{L}(\hat{\lambda}^j_{k,n_1}-\lambda^j_k)^2\right]} + \left|\hat{C}^{n_1}-C\right| \\ &= \sqrt{2}\left\|\hat{\boldsymbol{\lambda}}_{n_1}-\boldsymbol{\lambda}\right\|_2 + \left|\hat{C}^{n_1}-C\right|. \end{aligned} \tag{S3.11}$$

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

□

S4 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 ($\mathbf{H^1_b}$) 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 $(\mathbf{H^2_b})$ and Cauchy-Schwarz Inequality, there exists a fixed constant $C > 0$ such that for all $u \in S$, $\boldsymbol{l_u}$:

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

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 their 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* $(\mathbf{H}_{\varepsilon,\mathbf{q}})$ *with* $q > 4/\xi$ *and that* $(\mathbf{H}_{\mathbf{s},\alpha})$ *is fullfilled. Then, the following inequalities 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)

$$\sup_{u,\boldsymbol{l_u}} \left| \langle \tilde{f}, \hat{\phi}^u_{\boldsymbol{l_u},n_1} \rangle_{n_2} - \langle \tilde{f}, \hat{\phi}^u_{\boldsymbol{l_u},n_1} \rangle \right| = \|\boldsymbol{\beta^0}\|_{L^1} \mathcal{O}_P(n^{-\xi/2})$$

In the sequel, we will denote $\zeta_n := \max_{i \in [1:3]} \{\zeta_{n,i}\}$.

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

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

Lemma 5 and Assumption $(\mathbf{H^2_b})$ yields (ii).

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

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

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

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

Term II: We can bound II using the following simple inclusion:

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

Hence,

$$\begin{aligned} II &\leq P(\text{some } |\varepsilon^s|>K_n)\\ &\leq n_2P(|\varepsilon|>K_n)\leq n_2K_n^{-q}\mathbb{E}(|\varepsilon|^q)=\underset{n\to+\infty}{\mathcal{O}}(n^{1-q\xi/4}), \end{aligned}$$

where $n_2 = n/2$ with the choice $K_n := n^{\xi/4}$, 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 have 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

$$\begin{aligned} |\mathbb{E}(\varepsilon_t)| &= \left| \int_{|x|\leq K_n} x \mathrm{d}P_\varepsilon(x) + \int_{|x|>K_n} sg(x)K_n \mathrm{d}P_\varepsilon(x) \right| = \left| \int_{|x|>K_n} (sg(x)K_n - x)\mathrm{d}P_\varepsilon(x) \right| \\ &\leq \int \mathbb{1}_{|x|>K_n}(K_n + |x|)\mathrm{d}P_\varepsilon(x) \\ &\leq K_n P_\varepsilon(|\varepsilon| > K_n) + \int |x| \, \mathbb{1}_{|x|>K_n} \mathrm{d}P_\varepsilon(x) \\ &\leq K_n^{1-t}\mathbb{E}(|\varepsilon|^t) + \mathbb{E}(\varepsilon^2)^{1/2} K_n^{-t/2} \mathbb{E}(|\varepsilon|^t)^{1/2} \quad \text{by the Tchebychev Inequality} \\ &\leq \mathcal{O}(K_n^{1-t}) + \mathcal{O}(K_n^{-t/2}) = o(K_n^{-2}) \end{aligned} \tag{S4.1}$$

since $0 < \xi < 1$ and $t > 4/\xi > 4$. With the choice $K_n = n^{\xi/4}$, 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,

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

Equation (S4.1) 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, $(\mathbf{H}_{\mathbf{s},\alpha})$ and Bernstein's Inequality implies

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

which implies with Assumption $(\mathbf{H^2_b})$ 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 of Bühlmann (2006), holds:

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

$$\sup_{u,\boldsymbol{l_u}}|\langle Y-G_k(\bar{f}),\hat{\phi}^u_{\boldsymbol{l_u},n_1}\rangle_{n_2}-\langle\tilde{R}_k(\bar{f}),\phi^u_{\boldsymbol{l_u}}\rangle|\leq\left(\frac{5}{2}\right)^k(1+C\|\boldsymbol{\beta^0}\|_{L^1})\zeta_n.$$

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 $k=0$,

$$\begin{aligned}
\sup_{u,\boldsymbol{l_u}}|A_n(0,u)| &= \sup_u|\langle Y,\hat{\phi}^u_{\boldsymbol{l_u},n_1}\rangle_{n_2}-\langle\bar{f},\phi^u_{\boldsymbol{l_u}}\rangle|\\
&\leq \sup_{u,\boldsymbol{l_u}}\left\{\left|\langle\tilde{f},\hat{\phi}^u_{\boldsymbol{l_u},n_1}\rangle_{n_2}-\langle\tilde{f},\hat{\phi}^u_{\boldsymbol{l_u},n_1}\rangle\right|+\left|\langle\tilde{f}-\bar{f},\hat{\phi}^u_{\boldsymbol{l_u},n_1}\rangle\right|+\left|\langle\bar{f},\hat{\phi}^u_{\boldsymbol{l_u},n_1}-\phi^u_{\boldsymbol{l_u}}\rangle\right|\right\}\\
&\quad+\sup_{u,\boldsymbol{l_u}}\left|\langle\varepsilon,\hat{\phi}^u_{\boldsymbol{l_u},n_1}\rangle_{n_2}\right|\\
&\leq (1+4\|\boldsymbol{\beta^0}\|_{L^1})\zeta_n \quad \text{by } (iii)-(iv) \text{ of Lemma 6 and Theorem 1}
\end{aligned}$$

From the main document, we remind that

$$G_k(\bar{f})=G_{k-1}(\bar{f})+\gamma\langle Y-G_{k-1}(\bar{f}),\hat{\phi}^{u_k}_{\boldsymbol{l_{u_k}},n_1}\rangle_{n_2}\cdot\hat{\phi}^{u_k}_{\boldsymbol{l_{u_k}},n_1}, \tag{S4.2}$$

$$\begin{aligned}
R_k(\bar{f}) &= \bar{f}-G_k(\bar{f})\\
&= \bar{f}-G_{k-1}(\bar{f})-\gamma\langle Y-G_{k-1}(\bar{f}),\hat{\phi}^{u_k}_{\boldsymbol{l_{u_k}},n_1}\rangle_{n_2}\cdot\hat{\phi}^{u_k}_{\boldsymbol{l_{u_k}},n_1}
\end{aligned} \tag{S4.3}$$

and

$$\left\{\begin{array}{l}\tilde{R}_0(\bar{f})=\bar{f}\\ \tilde{R}_k(\bar{f})=\tilde{R}_{k-1}(\bar{f})-\gamma\langle\tilde{R}_{k-1}(\bar{f}),\hat{\phi}^{u_k}_{\boldsymbol{l_{u_k}},n_1}\rangle\hat{\phi}^{u_k}_{\boldsymbol{l_{u_k}},n_1}.\end{array}\right. \tag{S4.4}$$

The recursive relations (S4.2) and (S4.4) leads to, for any $k \geq 0$:

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

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

$$\left\|\tilde{R}_k(\bar{f})\right\|^2 \leq \left\|\bar{f}\right\|^2. \tag{S4.5}$$

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:

$$\begin{aligned}\sup_{u,\boldsymbol{l_u}}|A_n(k,u)| &\leq \left(\frac{5}{2}\right)^k \underbrace{\sup_{u,\boldsymbol{l_u}}|A_n(0,u)|}_{\leq(1+4\|\boldsymbol{\beta^0}\|_{L^1})\zeta_n}+3\zeta_n\left\|\bar{f}\right\|\sum_{\ell=0}^{k-1}\left(\frac{5}{2}\right)^{\ell} \\ &\leq \left(\frac{5}{2}\right)^k\zeta_n\left(1+\|\boldsymbol{\beta^0}\|_{L^1}\left(4+6\sum_{\ell=1}^{\infty}\left(\frac{5}{2}\right)^{-\ell}\right)\right),\end{aligned}$$

which ends the proof with $C=14$. □

We then aim at applying Theorem 2.1 from Champion et al. (2013) to the phantom residuals $(\tilde{R}_k(\bar{f}))_k$. Using the notation of Champion et al. (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

$$\begin{aligned}|\langle Y-G_{k-1}(\bar{f}),\hat{\phi}^{u_k}_{\boldsymbol{l_{u_k}},n_1}\rangle_{n_2}| &= \sup_{u,\boldsymbol{l_u}}|\langle Y-G_{k-1}(\bar{f}),\hat{\phi}^{u}_{\boldsymbol{l_u},n_1}\rangle_{n_2}| \\ &\geq \sup_{u,\boldsymbol{l_u}}\left\{|\langle\tilde{R}_{k-1}(\bar{f}),\phi^u_{\boldsymbol{l_u}}\rangle|-C\left(\frac{5}{2}\right)^{k-1}\zeta_n\|\boldsymbol{\beta^0}\|_{L^1}\right\}. \quad \text{(S4.6)}\end{aligned}$$

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

$$\begin{aligned}|\langle\tilde{R}_{k-1}(\bar{f}),\phi^{u_k}_{\boldsymbol{l_{u_k}}}\rangle| &\geq |\langle Y-G_{k-1}(\bar{f}),\hat{\phi}^{u_k}_{\boldsymbol{l_{u_k}},n_1}\rangle_{n_2}|-C\left(\frac{5}{2}\right)^{k-1}\zeta_n\|\boldsymbol{\beta^0}\|_{L^1} \\ &\geq \sup_{u,\boldsymbol{l_u}}|\langle\tilde{R}_{k-1}(\bar{f}),\phi^u_{\boldsymbol{l_u}}\rangle|-2C\left(\frac{5}{2}\right)^{k-1}\zeta_n\|\boldsymbol{\beta^0}\|_{L^1}. \quad \text{(S4.7)}\end{aligned}$$

Consider now the set

$$\tilde{\Omega}_n=\left\{\omega,\quad\forall k\leq k_n,\quad\sup_{u,\boldsymbol{l_u}}|\langle\tilde{R}_{k-1}(\bar{f}),\phi^u_{\boldsymbol{l_u}}\rangle|>4C\left(\frac{5}{2}\right)^{k-1}\zeta_n\|\boldsymbol{\beta^0}\|_{L^1}\right\}.$$

We deduce from Equation (S4.7) the following inequality on $\Omega_n \cap \tilde{\Omega}_n$:

$$|\langle \tilde{R}_{k-1}(\bar{f}), \phi^{u_k}_{\boldsymbol{l}_{u_k}} \rangle| \geq \frac{1}{2} \sup_{u,\boldsymbol{l}_u} |\langle \tilde{R}_{k-1}(\bar{f}), \phi^{u}_{\boldsymbol{l}_u} \rangle|. \tag{S4.8}$$

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 et al. (2013), with constant $\nu = 1/2$ and we have:

$$\left\| \tilde{R}_k(\bar{f}) \right\| \leq C' \left(1 + \frac{1}{4}\gamma(2-\gamma)k \right)^{-\frac{2-\gamma}{2(6-\gamma)}}. \tag{S4.9}$$

Consider now the complementary set

$$\tilde{\Omega}_n^C = \left\{ \omega, \quad \exists\, k \leq k_n \quad \sup_{u,\boldsymbol{l}_u} |\langle \tilde{R}_{k-1}(\bar{f}), \phi^{u}_{\boldsymbol{l}_u} \rangle| \leq 4C \left(\frac{5}{2} \right)^{k-1} \zeta_n \|\boldsymbol{\beta^0}\|_{L^1} \right\}.$$

Remark that

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

Moreover,

$$\begin{aligned} \sup_{u,\boldsymbol{l}_u} \left| \langle \tilde{R}_k(\bar{f}), \hat{\phi}^{u}_{\boldsymbol{l}_u,n_1} \rangle \right| &\leq \sup_{u,\boldsymbol{l}_u} \left| \langle \tilde{R}_k(\bar{f}), \phi^{u}_{\boldsymbol{l}_u} \rangle \right| + \sup_{u,\boldsymbol{l}_u} \left| \langle \tilde{R}_k(\bar{f}), \hat{\phi}^{u}_{\boldsymbol{l}_u,n_1} - \phi^{u}_{\boldsymbol{l}_u} \rangle \right| \\ &\leq \sup_{u,\boldsymbol{l}_u} \left| \langle \tilde{R}_k(\bar{f}), \phi^{u}_{\boldsymbol{l}_u} \rangle \right| + 2\|\boldsymbol{\beta^0}\|_{L^1} \zeta_n \quad \text{by Theorem 1 and (S4.5)} \end{aligned}$$

We hence have

$$\begin{aligned} \left\| \tilde{R}_k(\bar{f}) \right\|^2 &\leq \left(\|\boldsymbol{\beta^0}\|_{L^1} + \gamma \sum_{j=0}^{k-1} \left| \langle \tilde{R}_j(\bar{f}), \hat{\phi}^{u_j}_{\boldsymbol{l}_{u_j},n_1} \rangle \right| \right) \left(\sup_{u,\boldsymbol{l}_u} \left| \langle \tilde{R}_k(\bar{f}), \phi^{u}_{\boldsymbol{l}_u} \rangle \right| + 2\|\boldsymbol{\beta^0}\|_{L^1} \zeta_n \right) \\ &\leq \|\boldsymbol{\beta^0}\|_{L^1} (1 + 2\gamma k) \left(\sup_{u,\boldsymbol{l}_u} \left| \langle \tilde{R}_k(\bar{f}), \phi^{u}_{\boldsymbol{l}_u} \rangle \right| + 2\|\boldsymbol{\beta^0}\|_{L^1} \zeta_n \right) \\ &\leq 4C \|\boldsymbol{\beta^0}\|^2_{L^1} \zeta_n (1 + 2\gamma k) \left(\frac{5}{2} \right)^k \quad \text{on} \quad \tilde{\Omega}_n^C \end{aligned} \tag{S4.10}$$

Finally, on the set $(\Omega_n \cap \tilde{\Omega}_n) \cup \tilde{\Omega}_n^C$, by Equations (S4.9) and (S4.10),

$$\left\| \tilde{R}_k(\bar{f}) \right\|^2 \leq C'^2 \left(1 + \frac{1}{4}\gamma(2-\gamma)k \right)^{-\frac{2-\gamma}{6-\gamma}} + 4C \|\boldsymbol{\beta^0}\|^2_{L^1} \zeta_n (1 + 2\gamma k) \left(\frac{5}{2} \right)^k \tag{S4.11}$$

To conclude the first part of the proof, remark that

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

On this set, Inequality (S4.11) holds almost surely, and for $k_n < c\log(n)$ with $c < \frac{\xi/2-\vartheta-2\alpha}{2\log(3)}$, we get

$$\left\|\tilde{R}_{k_n}(\bar{f})\right\| \xrightarrow[n\to+\infty]{P} 0. \tag{S4.12}$$

Consider now $A_k := \left\|R_k(\bar{f}) - \tilde{R}_k(\bar{f})\right\|$ for $k \geq 1$. By definitions reminded in (S4.3)-(S4.4), we have:

$$\begin{aligned} A_k &\leq A_{k-1} + \gamma|\langle Y - G_{k-1}(\bar{f}), \hat{\phi}^{u_k}_{\boldsymbol{l}_{\boldsymbol{u_k}},n_1}\rangle_{n_2} - \langle \tilde{R}_{k-1}(\bar{f}), \hat{\phi}^{u_k}_{\boldsymbol{l}_{\boldsymbol{u_k}},n_1}\rangle| \\ &\leq A_{k-1} + \gamma|\langle Y - G_{k-1}(\bar{f}), \hat{\phi}^{u_k}_{\boldsymbol{l}_{\boldsymbol{u_k}},n_1}\rangle_{n_2} - \langle \tilde{R}_{k-1}(\bar{f}), \phi^{u_k}_{\boldsymbol{l}_{\boldsymbol{u_k}}}\rangle| \\ &\quad +\gamma|\langle \tilde{R}_{k-1}(\bar{f}), \hat{\phi}^{u_k}_{\boldsymbol{l}_{\boldsymbol{u_k}},n_1} - \phi^{u_k}_{\boldsymbol{l}_{\boldsymbol{u_k}}}\rangle|. \end{aligned} \tag{S4.13}$$

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

$$A_k \leq A_{k-1} + \gamma\left(\frac{5}{2}\right)^{k-1}\left(1 + C\|\boldsymbol{\beta^0}\|_{L^1}\right)\zeta_n + 2\gamma\|\boldsymbol{\beta^0}\|_{L^1}\zeta_n. \tag{S4.14}$$

Since $A_0 = 0$, we deduce recursively from Equation (S4.14) that, on Ω_n,

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

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 deal with the term $\left\|\bar{f} - \tilde{f}\right\|$. But remark that

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

and the proof follows using $(\mathbf{H}_{\mathbf{s},\alpha})$ with $\alpha < \xi/4 - \vartheta/2$ 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.

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

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

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