Metamodel construction for sensitivity analysis

Sylvie Huet, Marie-Luce Taupin

ESAIM: PROCEEDINGS AND SURVEYS, Vol. ?, 2019, 1-10
Editors: Will be set by the publisher

METAMODEL CONSTRUCTION FOR SENSITIVITY ANALYSIS

Sylvie Huet[1] and Marie-Luce Taupin[2,1]

Abstract. We propose to estimate a metamodel and the sensitivity indices of a complex model m in the Gaussian regression framework. Our approach combines methods for sensitivity analysis of complex models and statistical tools for sparse non-parametric estimation in multivariate Gaussian regression model. It rests on the construction of a metamodel for aproximating the Hoeffding-Sobol decomposition of m. This metamodel belongs to a reproducing kernel Hilbert space constructed as a direct sum of Hilbert spaces leading to a functional ANOVA decomposition. The estimation of the metamodel is carried out via a penalized least-squares minimization allowing to select the subsets of variables that contribute to predict the output. It allows to estimate the sensitivity indices of m. We establish an oracle-type inequality for the risk of the estimator, describe the procedure for estimating the metamodel and the sensitivity indices, and assess the performances of the procedure via a simulation study.

Résumé. Nous considérons l'estimation d'un méta-modèle d'un modèle complexe m à partir des observations d'un n-échantillon dans un modèle de régression Gaussien. Nous en déduisons une estimation des indices de sensibilité de m. Notre approche combine les méthodes d'analyse de sensibilité de modèles complexes et les outils statistiques de l'estimation non-paramétrique en régression multivariée. Elle repose sur la construction d'un méta-modèle qui approche la décomposition de Hoeffding-Sobol de m. Ce méta-modèle appartient à un espace de Hilbert à noyau reproduisant qui est lui-même la somme directe d'espaces de Hilbert, permettant ainsi une décomposition de type ANOVA. On en déduit des estimateurs des indices de sensibilité de m. Nous établissons une inégalité de type oracle pour le risque de l'estimateur, nous décrivons la procédure pour estimer le méta-modèle et les indices de sensibilité, et évaluons les performances de notre méthode à l'aide d'une étude de simulations.

1. Introduction

We consider a Gaussian regression model

$$Y = m(\mathbf{X}) + \sigma\varepsilon, \tag{1}$$

where $\mathbf{X}$ is a d random vector with a known distribution $P_{\mathbf{X}} = P_1 \times \ldots \times P_d$ on $\mathcal{X}$ a compact subset of $\mathbb{R}^d$, and ε is independent of $\mathbf{X}$, and distributed as a $\mathcal{N}(0,1)$ variable. The variance σ^2 is unknown and the number of variables d may be large. The function m is a complex and unknown function from $\mathbb{R}^d$ to $\mathbb{R}$. It may present strong non-linearities and high order interaction effects between its coordinates. On the basis of a n-sample $(Y_i, \mathbf{X}_i), i = 1, \ldots, n$, we aim to construct metamodels and perform sensitivity analysis in order to determine the infuence of each variable or group of variables on the output.

[1] Unit MaIAGE, INRA Jouy-en-Josas, France; e-mail: sylvie.huet@inra.fr
[2] Laboratoire LaMME, UMR CNRS 8071- USC INRA, Université d'Evry Val d'Essonne, France;
e-mail: marie-luce.taupin@univ-evry.fr

Our approach combines methods for sensitivity analysis of a complex model and statistical tools for sparse non-parametric estimation in multivariate Gaussian regression model. It rests on the construction of a meta-model for aproximating the Hoeffding-Sobol decomposition of the function m. This metamodel belongs to a reproducing kernel Hilbert space constructed as a direct sum of Hilbert spaces leading to a functional ANOVA decomposition involving variables and interactions between them. The estimation of the metamodel is carried out via a penalized least-squares minimization allowing to select the subsets of variables $\mathbf{X}$ that contribute to predict the output Y. Finally, the estimated metamodel allows to estimate the sensitivity indices of m.

A lot of work has been done around meta-modelling and sensitivity indices estimation.

For a complete account on global sensitivity analysis, see for example the book by Saltelli et al. [35]. Let us briefly present the context of usual global sensitivity analysis. Suppose that we are able to calculate the ouputs z of a model m for n realizations of the input vector $\mathbf{X}$, such that $z_i = m(\mathbf{X}_i)$ for $i = 1, \dots, n$. Starting from the values $(z_i, \mathbf{X}_i), i = 1, \dots n$, the objectives of meta-modelling and global sensitivity analysis are to approximate the function m by what is called a metamode or are to quantify the influence of some subsets of the variables $\mathbf{X}$ on the output z. This metamodel helps to understand the behavior of the model, or allows to speed up future calculation using it in place of the original model m.

In particular when the inputs variables $\mathbf{X}$ are independent, if m is square integrable, one may consider the classical Hoeffding-Sobol decomposition [36, 41] that leads to write m according to its ANOVA functional expansion:

$$m(\mathbf{x}) = m_0 + \sum_{v \in \mathcal{P}} m_v(\mathbf{x}_v), \tag{2}$$

where $\mathcal{P}$ denotes the set of parts of $\{1, \dots, d\}$ with dimension 1 to d and where for all $\mathbf{x} \in \mathbb{R}^d$, $\mathbf{x}_v$ denotes the vector with components x_j for $j \in v$. The functions m_v are centered and orthogonal in $\mathbb{L}^2(P_{\mathbf{X}})$ leading to the following decomposition of the variance of m: $\mathrm{Var}\left(m(\mathbf{x})\right) = \sum_v \mathrm{Var}\left(m_v(\mathbf{x}_v)\right)$. The Sobol sensitivity indices, introduced by Sobol [36], are defined for any group of variables $\mathbf{x}_v$, $v \in \mathcal{P}$ by

$$S_v = \frac{\mathrm{Var}\left(m_v(\mathbf{x}_v)\right)}{\mathrm{Var}\left(m(\mathbf{x})\right)}.$$

They quantify the contribution of a subset of variables $\mathbf{x}$ to the output $m(\mathbf{x})$. Several approaches are available for estimating these sensitivity indices, see for example Iooss and Lemaître [17] for a recent review. Among all of them, let us consider the one based on metamodel construction that allows to directly obtain the sensitivity indices. Generally one consider metamodels that correspond to an ANOVA-type decomposition, and that are candidate to approximate the Hoeffding decomposition of m. The ANOVA-type decomposition leads to consider functions defined as follows:

$$f : \mathcal{X} \to \mathbb{R}, f(\mathbf{x}) = f_0 + \sum_{v \in \mathcal{P}} f_v(\mathbf{x}_v),\ \mathbb{E}_{P_{\mathbf{X}}} f_v(\mathbf{x}_v) = \mathbb{E}_{P_{\mathbf{X}}} f_v(\mathbf{x}_v) f_{v'}(\mathbf{x}_{v'}) = 0\ \forall v, v' \in \mathcal{P} \tag{3}$$

for functions f_v that are chosen to belong to some functionnal spaces. The polynomial Chaos construction, see for example Ghanem and Spanos [14], Soize and Ghanem [38], can be used to approximate the Hoeffding decomposition of m. This approach was considered by Blatman and Sudret [5] who propose a method for truncating (such that to keep polynomials with degree less than some integer) the polynomial Chaos expansion and then an algorithm based on least angle regression for selecting the terms in the expansion. For the same purpose, Gu and Wu [15] propose an algorithm based on the hierarchy principle (lower order effects are more likely to be important than higher order effects) and on the heredity principle (interaction can be active only if one or all of its parent effects are also active). This approach joins the one proposed by Bach [2] for variable selection based on hierarchical kernel learning.

Inspired by Touzani [39], Durrande et al. [12] propose to approximate m by functions belonging to a reproducing kernel Hilbert space (RKHS). The RKHS is constructed as a direct sum of Hilbert spaces leading to

a functional ANOVA decomposition (see Equation (3)), such that the projection of m onto the RKHS is an approximation of the Hoeffding decomposition of m.

Following Lin and Zhang [25], Touzani and Busby [40] propose an algorithm to calculate the penalized least-square estimator of m on the RKHS space, where the least-square criteria is penalized by the sum of the norms of m on each Hilbert subspace. This group-lasso type procedure allows both to select and calculate the non-zero terms in the functional ANOVA decomposition.

Our objective is to propose an estimator of a metamodel which will approximate the Hoeffding decomposition of m considering a Gaussian regression model defined at Equation (1) and to deduce from this estimated metamodel, estimators for the sensitivity indices of m. Contrary to the usual setting of sensitivity analysis where $m(\mathbf{X}_i)$ is available, only the observations Y are available, which leads us to consider the nonparametric multivariate regression setting.

Let us briefly describe the methods related to this regression setting and review their theoretical properties, starting with papers assuming an univariate additive decomposition for the function m in the context of high-dimensional sparse additive models. Precisely, denote by $\mathcal{F}^{\text{1-add}}$, the set of functions f defined on $\mathcal{X}$ such that $f(\mathbf{x}) = f_0 + \sum_{a=1}^{d} f_a(x_a)$ where f_0 is a constant, and where for $a = 1, \ldots, d$, the functions f_a are centered and square-integrable with respect to P_a. For each function f, the set S_f of indices $a \in \{1, \ldots d\}$ such that f_a is not identically zero is called the actice set of f.

Ravikumar et al. [32] propose a group-lasso procedure where each function f_a is approximated by its truncated decomposition on a basis of functions. They provide an algorithm and, assuming that the function m belongs to the set $\mathcal{F}^{\text{1-add}}$ and that S_m is sparse, prove the consistency of the active set and of the risk of the estimator of m.

Meier et al. [27] propose to combine a sparsity penalty (group-lasso) and a smoothness penalty (ridge) for estimating $m(\mathbf{x})$. Considering the fixed design framework, they establish some oracle properties of the empirical risk for estimating the projection of m onto the set of univariate additive functions $\mathcal{F}^{\text{1-add}}$.

Raskutti et al. [31] consider the case where each univariate function f_a belongs to a RKHS and as Meier et al. combine a sparsity and a smoothness penalty. Assuming that the d variables $\mathbf{X}$ are independent, they derive upper bounds for the integrated and the empirical risks, as well as a lower bound for the integrated risk over spaces of sparse additive models whose each component is bounded with respect to the RKHS norm.

Additive sparse modelling is too restrictive in practical settings because it does not take into account interactions between variables that may affect the relationship between Y and $\mathbf{X}$. The generalization of additive smoothing splines to interaction smoothing splines leading to an ANOVA-type decomposition (see Equation (3)) was proposed by several authors (see for example Wahba [42], Friedman [13], Wahba et al. [43]).

To control smoothness and to enforce sparsity in the ANOVA-type decomposition, Gunn and Kandola [16] propose to consider the ANOVA decomposition as a weighted linear sum of kernels and to use a lasso penalty on the weights to select the terms in the decomposition as well as a ridge penalty to ensure smoothness of the kernel expansion. The COSSO proposed by Lin and Zhang [25] is based on smoothness penalty defined as the sum of the RKHS-norms of the functions f_v. The authors study existence and rate of convergence of the estimator. In a more general framework, where the function m is written as a linear span of a large number of kernels, Koltchinskii and Yuan [21] established oracle inequalities on the excess risk assuming that the function m has a sparse representation (the set of $v \in \mathcal{P}$ such that f_v^* is non zero is sparse). The authors generalized their results to a penalty function that combines sparsity and smoothness [22], as proposed by Meier et al. [27] and Raskutti et al. [31].

Recently Kandasamy and Wu [19] proposed an estimator called SALSA, based on a ridge penalty, where the ANOVA-type decomposition is restricted to set $v \in \mathcal{P}$ such that $|v| \leq D_{\max}$. The authors propose to choose $D_{\max}$ via a cross-validation procedure.

Our contributions

Using the functionnal ANOVA-type decomposition as proposed by Durrande et al. [12], we propose an estimator of a metamodel which approximates the Hoeffding decomposition of m. Following the most recent

works in the framework of nonparametric estimation of sparse additive models, we propose a penalized least-square estimator where the penalty function enforces both the sparsity and the smoothness of the terms in the decomposition. We show that our estimator satisfies an oracle inequality with respect to the empirical and integrated risks.

Our procedure allows both to select and estimate the terms in the ANOVA decomposition, and therefore, to select the sensitivity indices that are non-zero and estimate them. In particular it makes possible to estimate Sobol indices of high order, a point known to be difficult in practice.

Finally, using convex optimization tools, we develop an algorithm (available on request) in *R* [30], for calculating the estimator. A simulation study shows the good performances of our estimator in practice.

The paper is organised as follows: The RKHS construction based on ANOVA kernels and the procedure for estimating a metamodel are presented in Section 2. The estimators of the Sobol indices are given in Section 3. The theoretical properties of the metamodel estimator are stated in Theorem 4.1 and Corollaries 4.1 and 4.2 whose proofs are postponed in Sections 7 and 8. Section 5 is devoted to the calculation of the estimator and Section 6 to the simulation study.

2. Meta-modelling

We start from the Hoeffding decomposition (see Sobol [37] and Van der Vaart [41], p. 157) of the function m that consists in writting m as in Equation (2)

$$m(\mathbf{x}) = m_0 + \sum_{v \in \mathcal{P}} m_v(\mathbf{x}_v),$$

where $\mathcal{P}$ denotes the set of parts of $\{1, \ldots, d\}$ with dimension 1 to d and where for all $\mathbf{x} \in \mathbb{R}^d$, $\mathbf{x}_v$ denotes the vector with components x_j for $j \in v$. For all v, v' in $\mathcal{P}$,

$$E_{\mathbf{X}}\left(m_v(\mathbf{X}_v)\right) = E_{\mathbf{X}}\left(m_v(\mathbf{X}_v) m_{v'}(\mathbf{X}_{v'})\right) = 0.$$

We propose to consider a functionnal space based on the tensorial product of Reproducing Kernel Hilbert spaces (RKHS), and to approximate the unknown function m by its projection denoted f^* on such such RKHS space. One of the key point is to construct the space $\mathcal{H}$ such that the terms of the decomposition of a function f in $\mathcal{H}$ correspond to its Hoeffding-Sobol decomposition.

2.1. **RKHS construction**

Let us describe the construction of spaces $\mathcal{H}$, based on ANOVA kernels, construction which was given by Durrande et al. [12].

Let $\mathcal{X} = \mathcal{X}_1 \times \ldots \times \mathcal{X}_d$ be a compact subset of $\mathbb{R}^d$. For each coordinate $a \in \{1, \cdots, d\}$, we choose a RKHS $\mathcal{H}_a$ and its associated kernel k_a defined on the set $\mathcal{X}_a \subset \mathbb{R}$ such that the two following properties are satisfied

(1) $k_a : \mathcal{X}_a \times \mathcal{X}_a \to \mathbb{R}$ is $P_a \times P_a$ measurable
(2) $\mathbb{E}_{P_a} \sqrt{k_a(X_a, X_a)} < \infty$

The RKHS $\mathcal{H}_a$ may be decomposed as $\mathcal{H}_a = \mathcal{H}_{0a} \overset{\perp}{\oplus} \mathcal{H}_{1a}$, where

$$\begin{aligned} \mathcal{H}_{0a} &= \{f_a \in \mathcal{H}_a, \mathbb{E}_{P_a}(f_a(X_a)) = 0\} \\ \mathcal{H}_{1a} &= \{f_a \in \mathcal{H}_a, f_a(X_a) = C\}, \end{aligned}$$

the kernel k_{0a} associated to the RKHS $\mathcal{H}_{0a}$ being defined as follows (see Berlinet et Thomas-Agnan [4]):

$$k_{0a}(x_a, x'_a) = k_a(x_a, x'_a) - \frac{\mathbb{E}_{U \sim P_a}(k_a(x_a, U)) \mathbb{E}_{U \sim P_a}(k_a(x'_a, U))}{\mathbb{E}_{(U,V) \sim P_a \times P_a} k_a(U, V)}.$$

The ANOVA kernel is defined as

$$k(\mathbf{x},\mathbf{x}') = \prod_{a=1}^{d}\left(1+k_{0a}(\mathbf{x}_a,\mathbf{x}'_a)\right) = 1+\sum_{v\in\mathcal{P}} k_v(\mathbf{x}_v,\mathbf{x}'_v), \text{ where } k_v(\mathbf{x}_v,\mathbf{x}'_v) = \prod_{a\in v} k_{0a}(x_a,x'_a),$$

and the corresponding RKHS

$$\mathcal{H} = \otimes_{a=1}^{d}\left(1 \overset{\perp}{\oplus} \mathcal{H}_{0a}\right) = 1+\sum_{v\in\mathcal{P}}\mathcal{H}_v,$$

where the RKHS $\mathcal{H}_v$ is associated with kernel k_v. According to this construction, any function $f\in\mathcal{H}$ satisfies

$$f(\mathbf{x}) = \langle f, k(\mathbf{x},\cdot)\rangle_{\mathcal{H}} = f_0 + \sum_{v\in\mathcal{P}} f_v(\mathbf{x}),$$

where $f_v(\mathbf{x}) = \langle f, k_v(\mathbf{x},\cdot)\rangle_{\mathcal{H}}$ depends on $\mathbf{x}_v$ only. For all $v\in\mathcal{P}$, $f_v(\mathbf{x}_v)$ is centered and for all $v'\neq v$, $f_v(\mathbf{x}_v)$ and $f_{v'}(\mathbf{x}_{v'})$ are uncorrelated. We get thus the Hoeffding decomposition of f.

2.2. **Approximating the Hoeffding decomposition of** m

Let $f^* = f_0^* + \sum_{v\in\mathcal{P}} f_v^*$ which minimizes

$$\|m-f\|^2_{\mathbb{L}^2(P_{\mathbf{X}})} = E_{\mathbf{X}}\left(m(\mathbf{X})-f(\mathbf{X})\right)^2$$

over functions $f\in\mathcal{H}$. This f^* can be viewed as an approximation of m and more specifically his Hoeffding decomposition is an approximation of the Hoeffding decomposition of m. Therefore if the Hoeffding decomposition of m is written as in Equation (2), each function f_v^* approximates the function m_v.

The idea is propose an estimator of f^* as estimator of m.

2.3. **Selection step**

Since $\mathcal{P}$ is the set of parts of $\{1,\dots,d\}$, the number of functionsf_v^* is related to the cardinality of $\mathcal{P} = 2^d-1$ that may be huge. Our construction is thus associated to a selection strategy.

The selection of f_v^* in f^* is based on a *ridge-group-sparse* type procedure which minimizes the penalized least-squares criteria over functions $f\in\mathcal{H}$. The least-squares criteria is penalized in order to both select few terms in the additive decomposition of f over sets $v\in\mathcal{P}$, and to favour smoothness of the estimated f_v. The ridge regularization is ensured by controling the norm of f_v in the Hilbert space $\mathcal{H}_v$ for all v, and the group-sparse regularization is strengthened by controling the empirical norm of f_v, defined as

$$\|f\|_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n} f_v^2(\mathbf{X}_{v,i})}.$$

For any $f\in\mathcal{H}$ such that $f = f_0+\sum_{v\in\mathcal{P}} f_v$, and for some tuning parameters $(\mu_v,\gamma_v, v\in\mathcal{P})$, let $\mathcal{L}(f)$ be defined as

$$\mathcal{L}(f) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - f_0 - \sum_{v\in\mathcal{P}} f_v(\mathbf{X}_{v,i})\right)^2 + \sum_{v\in\mathcal{P}}\mu_v\|f_v\|_{\mathcal{H}_v} + \sum_{v\in\mathcal{P}}\gamma_v\|f_v\|_n. \tag{4}$$

Let us define the set of functions $\mathcal{F}$

$$\mathcal{F} = \left\{ f \text{ such that } f = f_0 + \sum_{v\in\mathcal{P}} f_v, f_v\in\mathcal{H}_v, \|f_v\|_{\mathcal{H}_v}\leq 1\right\}. \tag{5}$$

Then the estimator $\widehat{f}$ is defined as

$$\widehat{f} = \operatorname{argmin} \{\mathcal{L}(f), f \in \mathcal{F}\}. \tag{6}$$

Remark 2.1. *The construction of the RKHS spaces described above, allows to consider functionnal spaces that suit well to the smoothness of the function m, irrespectively of the distribution of $\mathbf{X}$. Indeed, the kernels $k_{0,a}$ depend on the distribution of $\mathbf{X}$ only for calculating the projection onto the space of constant functions. In comparison, the decomposition based on the truncated polynomial Chaos expansion, used for sensitivity analysis (see Blatman and Sudret [5]), is based on the distribution of $\mathbf{X}$, and only the choice of the truncation handles the smoothness of the approximation.*

3. Sensitivity analysis

3.1. **Sobol indices**

Let us go back to the Hoeffding decomposition Equation (2). The orthogonality between two terms in this decomposition leads to the additive decomposition of the variance of $m(\mathbf{x})$:

$$\operatorname{Var}(m(\mathbf{x})) = \sum_{v \in \mathcal{P}} \operatorname{Var}(m_v(\mathbf{x}_v)).$$

Each of these variance terms are related to Sobol indices [36]. For example, the Sobol indice linked with the interaction between variables $\mathbf{x}_v$ is defined as

$$S_v = \frac{\operatorname{Var}(m_v(\mathbf{x}_v))}{\operatorname{Var}(m(\mathbf{x}))},$$

or the global Sobol indices for the variable x_a, $a \in \{1, \cdots, d\}$, is

$$G_a = \frac{\sum_{v \supseteq \{a\}} \operatorname{Var}(m_v(\mathbf{x}_v))}{\operatorname{Var}(m(\mathbf{x}))}.$$

Those Sobol indices and global Sobol indices quantify the contribution of a subset of variables $\mathbf{x}$ to the output $m(\mathbf{x})$. As it is said in the introduction direct estimation of these Sobol indices may require lot of calculations. We consider here methods based on metamodels to directly obtain Sensitivity indices.

3.2. **Estimation of Sobol indices**

Thanks to the orthogonality property of functions in $\mathcal{H}$, the variance of $m(\mathbf{x})$ will be estimated by

$$\widehat{\operatorname{Var}}(m(\mathbf{x})) = \sum_{v \in \mathcal{P}} \widehat{\operatorname{Var}}(m_v(\mathbf{x}_v)), \text{ where } \widehat{\operatorname{Var}}(m_v(\mathbf{x}_v)) = \mathbb{E}_{\mathbf{X}}\left(\widehat{f}_v^2(\mathbf{X}_v)\right) = \|\widehat{f}_v\|^2_{\mathbb{L}^2(P_{\mathbf{X}})}. \tag{7}$$

In practice, in order to avoid calculating the variance of $\widehat{f}_v(\mathbf{X}_v)$, one may use an estimator based on the empirical variances of functions $\widehat{f}_v$. Precisely, if $\widehat{f}_{v,\cdot}$ is the mean of the $\widehat{f}_v(\mathbf{X}_{v,i})$, for $i = 1, \ldots, n$, then

$$\widehat{\operatorname{Var}}^{\text{emp}}(m_v(\mathbf{x}_v)) = \frac{1}{n-1} \sum_{i=1}^{n} \left(\widehat{f}_v(\mathbf{X}_{v,i}) - \widehat{f}_{v,\cdot}\right)^2. \tag{8}$$

One of the main contribution of this approach is to allow the estimation of Sobol indices of any order, whereas classical methods only deal with small order, generally less than two.

4. Theoretical result: oracle inequality for metamodel

In this section we state the oracle inequality for the estimated metamodel $\widehat{f}$ which approximates the Hoeffding decomposition of the unknown function m.

4.1. Notations and Assumptions

For a function $f \in \mathcal{H}$, $f = f_0 + \sum_{v\in\mathcal{P}} f_v$, we denote by S_f its support and $|S_f|$ its cardinality. More precisely

$$S_f = \{v \in \mathcal{P}, f_v \neq 0\}. \tag{9}$$

We consider RKHS spaces $\mathcal{H}_v, v \in \mathcal{P}$ satisfying the following assumptions:

- for all $f_v \in \mathcal{H}_v$, $E_{\mathbf{X}} f_v^2(\mathbf{X}) < \infty$,
- for all $f_v \in \mathcal{H}_v$, $f_{v'} \in \mathcal{H}_{v'}$, $E_{\mathbf{X}} f_v(\mathbf{X}) = 0$ and $E_{\mathbf{X}} f_v(\mathbf{X}) f_{v'}(\mathbf{X}) = 0$,
- there exists $R' > 0$ such that

$$\forall f_v \in \mathcal{H}_v \;\; \|f_v\|_\infty = \sup\{|f_v(\mathbf{X})|, \mathbf{X} \in \mathcal{X}\} \leq R'. \tag{10}$$

For each $v \in \mathcal{P}$, let $\omega_{v,k}$, for $k \geq 1$ be the eigenvalues of the operator associated to the self reproducing kernel k_v, arranged in the decreasing order. Let us define the function $Q_{n,v}(t)$, for positive t, as follows:

$$Q_{n,v}(t) = \sqrt{\frac{5}{n}\sum_{k\geq 1} \min(t^2, \omega_{v,k})}, \tag{11}$$

and for some $\Delta > 0$ let $\nu_{n,v}$ be defined as follows

$$\nu_{n,v} = \inf\left\{t \text{ such that } Q_{n,v}(t) \leq \Delta t^2\right\}. \tag{12}$$

For each $v \in \mathcal{P}$, $\nu_{n,v}$ refers to the so-called critical univariate rate, the minimax-optimal rate for $\mathbb{L}^2(P_{\mathbf{X}})$-estimation of a single univariate function in the hilbert space $\mathcal{H}_v$ (e.g. Mendelson [28]).

Our choices of regularization parameters and rates are specified in terms of the quantities:

$$\lambda_{n,v} \quad = \quad \max\left\{\nu_{n,v}, \sqrt{d/n}\right\}. \tag{13}$$

Theorem 4.1. *Let us consider the regression model defined at Equation (1). Let $(Y_i, \mathbf{X}_i)$, $i = 1, \ldots, n$ be a n-sample with the same law as $(Y, \mathbf{X})$. Let $\widehat{f}$ be defined by (6).*

If there exist constants $C_l, l = 1, 2, 3$, $C_1 \geq 1$, and $0 < \eta < 1$ such that the following conditions are satisfied:

$$\text{for all } v \in \mathcal{P}, \lambda_{n,v} \leq \min\left\{\frac{\gamma_v}{C_1}, \sqrt{\frac{\mu_v}{C_1}}\right\}, \; n\lambda_{n,v}^2 \geq -C_2 \log \lambda_{n,v}, \tag{14}$$

and

$$\text{for all } f \in \mathcal{F}, \; \max\left(\sum_{v\in S_f} \gamma_v^2, \sum_{v\in S_f} \mu_v\right) \leq 1, \tag{15}$$

then, with probability greater than $1 - \eta$,

$$\|m - \widehat{f}\|_n^2 \leq C_3 \inf_{f\in\mathcal{F}} \left\{\|m - f\|_n^2 + \sigma^2 \sum_{v\in S_f} \left(\mu_v + \gamma_v^2\right)\right\}.$$

The result can be easily generalized to the minimization of $\mathcal{L}(f)$ over the space $\mathcal{G}$ defined as follows: For some positive constants $r_v, v \in \mathcal{P}$,

$$\mathcal{F}_r = \left\{ f \text{ such that } f = f_0 + \sum_{v \in \mathcal{P}} f_v, f_v \in \mathcal{H}_v, \|f_v\|_{\mathcal{H}_v} \leq r_v \right\}. \tag{16}$$

Indeed, we just have to consider the RKHS $\mathcal{H}_v^{'}$ associated with the kernel $k_v^{'}(\mathbf{x}_v, \mathbf{x}_v^{'}) = r_v^2 k_v(\mathbf{x}_v, \mathbf{x}_v^{'})$ in place of the RKHS $\mathcal{H}_v$. Then minimizing

$$\frac{1}{n}\sum_i \left(Y_i - f_0 - \sum_v f_v(X_{v,i}) \right)^2 + \sum_v \mu_v r_v \|f_v\|_{\mathcal{H}_v'} + \sum_v \gamma_v \|f_v\|_n$$

over the set

$$\mathcal{F}^{'} = \left\{ f \text{ such that } f = f_0 + \sum_v f_v, f_v \in \mathcal{H}_v^{'}, \|f_v\|_{\mathcal{H}_v'} \leq 1 \right\}.$$

is equivalent to minimizing $\mathcal{L}(f)$ over the set $\mathcal{F}_r$.

Let us now comment on the theorem.

- The term $\|m - f\|_n^2$ refers to the usual bias term quantifying the approximation properties of the Hilbert space $\mathcal{H}$ as a distance between the true m and f, its approximation into $\mathcal{H}$.
- Koltchinskii et Yuan [22] considered what is called the multiple kernel learning problem, where the functions in $\mathcal{H}$ have an additive representation over kernel spaces. They do not assume that the variable $\mathbf{X}$ are independent, nor that the kernel spaces satisfy an orthogonality condition. In return, they assume that some decomposability type properties are satisfied, and they introduce some characteristics related to the degree of "dependence" of the kernel spaces.
- In the particular case when the decomposition is limited to the main effects of the variables, then the problem comes back to the classical nonparametric additive model. The theoretical properties of the estimator based on a *ridge-group-sparse* type procedure have already been established (see for example Meier et al. [27], and Raskutti et al. [31]).
- Weights in the penalty terms may be of interest for applications. The theoretical result highlights that the tuning parameters (μ_v, γ_v) should depend on the decreasing of the eigenvalues of the kernel defining $\mathcal{H}_v$. Besides, we may be interested by introducing weigths that favor small order interaction terms.
- Because we aim to approximate the Hoeffding decomposition of m, we need to have orthogonality between the spaces $\mathcal{H}_v, v \in \mathcal{P}$. This condition, required by our objective, is also a key point in the proof of the Theorem, when the problem is to compare the euclidean norm of functions in $\mathcal{H}$ with the norm in $\mathbb{L}^2(P_{\mathbf{X}})$. At this step we need to assume that Assumption (15) holds to conclude.
- We do not assume that the functions in $\mathcal{F}$ are uniformly globally bounded, that is that the $\sup\{|f(\mathbf{x})|, \mathbf{x} \in \mathcal{X}\}$ is bounded by a constant that does not depend on f. Instead we assume that each function within the unit ball of the Hilbert space $\mathcal{H}_v$ is uniformly bounded by a constant multiple of its Hilbert norm. In fact, the functions f in the space $\mathcal{F}$, written as $f_0 + \sum_v f_v$ satisfy that for each group v, f_v is uniformly bounded. This assumption is easily satisfied as soon as the kernel k_v is bounded on the compact set $\mathcal{X}$. Indeed, $\|f_v\|_\infty \leq \sup_{\mathbf{X} \in \mathcal{X}} \sqrt{k_v(\mathbf{X}_v, \mathbf{X}_v)} \|f_v\|_{\mathcal{H}_v}$. We refer to [31] for a discussion on that subject and a comparison with the work of Koltchinskii et Yuan [22].
- Assuming that $n\lambda_{n,v}^2 \geq -C_2 \log \lambda_{n,v}$ allows to control the probability of the union of $|\mathcal{P}|$ events. This is a mild condition, satisfied for $\lambda_{n,v} = K_n/\sqrt{n}$ for K_n of the order $\sqrt{\log n}$.

The following corollary gives an upper bound of the risk with respect to the $\mathbb{L}^2(P_{\mathbf{X}})$ norm. It is mainly a consequence of Theorem 4.1 (its proof is given Section 8.2 page 21).

Corollary 4.1. *Under the assumptions of Theorem 4.1, we have that, with probability greater than* $1-\eta$*, for some constant* C_4*,*

$$\|m-\widehat{f}\|^2_{\mathbb{L}^2(P_{\mathbf{X}})} \leq C_4 \inf_{f\in\mathcal{F}} \left\{ \|m-f\|_n^2 + \|m-f\|^2_{\mathbb{L}^2(P_{\mathbf{X}})} + \sigma^2 \sum_{v\in S_f} \left(\mu_v + \gamma_v^2\right) \right\}.$$

From this corollary we can compare the $\widehat{\mathrm{Var}}\left(m_v(\mathbf{x}_v)\right)$ (see Equation (7)) with the variance of $m_v(\mathbf{x}_v)$. Thanks to the following inequality

$$\left| \|\widehat{f}_v\|_{\mathbb{L}^2(P_{\mathbf{X}})} - \|m_v\|_{\mathbb{L}^2(P_{\mathbf{X}})} \right| \leq \|\widehat{f}_v - m_v\|_{\mathbb{L}^2(P_{\mathbf{X}})} \leq \|\widehat{f}-m\|_{\mathbb{L}^2(P_{\mathbf{X}})},$$

and to Corollary 4.1, we get the following result

$$\begin{aligned} \text{if } m_v \equiv 0 \quad \text{then} \quad & \widehat{\mathrm{Var}}\left(m_v(\mathbf{x}_v)\right) \leq \|\widehat{f}-m\|^2_{\mathbb{L}^2(P_{\mathbf{X}})} \\ \text{if } \|m_v(x_v)\|_{\mathbb{L}^2(P_{\mathbf{X}})} \geq c>0 \quad \text{then} \quad & \left| \frac{\widehat{\mathrm{Var}}\left(m_v(\mathbf{x}_v)\right)}{\mathrm{Var}\left(m_v(\mathbf{x}_v)\right)} - 1 \right| \leq \|\widehat{f}-m\|^2_{\mathbb{L}^2(P_{\mathbf{X}})}/c. \end{aligned}$$

4.2. **Rate of convergence**

Corollary 4.2. *Under the same condition as Theorem 4.1, if* $\gamma_v = c\lambda_{n,v}$ *and* $\mu_v = c\lambda^2_{n,v}$ *for* $c \geq C1$*, then*

$$\|m-\widehat{f}\|_n^2 \leq C_3 \inf_{f\in\mathcal{F}} \left\{ \|m-f\|_n^2 + \left(\sum_{v\in S_f} \nu^2_{n,v} + \frac{d|S_f|}{n} \right) \sigma^2 \right\}.$$

The result is non asymptotic in the sense that it is shown for any (n,d). Nevertheless, the upper bound is relevant when the infimum is reached for functions f whose decomposition in $\mathcal{H}$ is sparse, and when d is small face to n. In fact, the coefficient d occuring in the rate $d|S_f|/n$ comes from the logarithm of the cardinality of $\mathcal{P}$ equal to $\log(2^d-1)$. When d is large, it may be judicious to limit the decomposition of functions in $\mathcal{H}$, to interactions of limited order, so that the number of terms in the decomposition stays of the order $\log(d)$.

Let us discuss the rate of convergence given by $\sum_{v\in S_f} \nu^2_{n,v}$. For the sake of simplicity let us consider the case where the variables $X_1,\ldots,X_d$ have the same distribution P_1 on $\mathcal{X}_1 \subset \mathbb{R}$, and where the unidimensionnal kernels k_{0a} are all identical, such that $k_v(\mathbf{x}_v,\mathbf{x}'_v) = \prod_{a\in v} k_0(x_a,x'_a)$. The kernel k_0 admits an eigen expansion given by

$$k_0(x,x') = \sum_{\ell\geq 1} \omega_{0,\ell}\zeta_\ell(x)\zeta_\ell(x')$$

where the eigenvalues $\omega_{0,\ell}$ are non negative and ranged in the decreasing order, and where the ζ_ℓ are the associated eigen functions, orthonormal with respect to $\mathbb{L}^2(P_1)$. Therefore the kernel k_v admits the following expansion

$$k_v(\mathbf{x}_v,\mathbf{x}_v^{'}) = \sum_{\boldsymbol{\ell}=(\ell_1\ldots\ell_{|v|})} \underbrace{\prod_{a=1}^{|v|} \omega_{0,\ell_a}}_{\omega_{v,\boldsymbol{\ell}}} \underbrace{\prod_{a=1}^{|v|} \zeta_{\ell_a}(x_a)}_{\zeta_{v,\boldsymbol{\ell}}(\mathbf{x}_v)} \underbrace{\prod_{a=1}^{|v|} \zeta_{\ell_a}(x_a^{'})}_{\zeta_{v,\boldsymbol{\ell}}(\mathbf{x}'_v)}.$$

Consider now the case where the eigenvalues $\omega_{0,\ell}$ are decreasing at a rate $\ell^{-2\alpha}$ for some $\alpha > 1/2$. It can be shown, see Section 8.3, that the rate $\nu_{n,v}$ defined at Equation (12) is bounded above by a term of order $n^{-\alpha/(2\alpha+1)}(\log n)^\gamma$, where $\gamma \geq (|v|-1)\alpha/(2\alpha-1)$. Note that in this particular case, the rate of convergence depends on $|v|$ through the logarithmic term, and that up to this logarihmic term the rate of convergence has the same order than the usual nonparametric rate for unidimensionnal functions. It follows that the RKHS

space $\mathcal{H}$ should be chosen such that the unknown function m is well approximated by sparse functions in $\mathcal{H}$ with low order of interactions.

5. Calculation of the estimator

The functional minimization problem described at Equation (6) is equivalent to a parametric minimization problem. Indeed, we know that if $\mathcal{H}$ is a RKHS associated with a kernel $k : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$, then for all $(\mathbf{x}_1, \dots, \mathbf{x}_n) \in \mathcal{X}^n$, and for all $(\alpha_1, \dots, \alpha_n) \in \mathbb{R}^n$, the function $f(\cdot) = \sum_{i=1}^n \alpha_i k(\mathbf{x}_i, \cdot)$ is in $\mathcal{H}$ and $\|f\|^2_{\mathcal{H}} = \sum_{i,i'=1}^n \alpha_i\alpha_{i'} k(\mathbf{x}_i, \mathbf{x}'_{i'})$. In particular, it can be shown that the solution to our minimization problem is written as $f = f_0 + \sum_{v\in\mathcal{P}} f_v$ where, according to the representer Theorem (see Kimeldorf and Whahba [20]), $f_v(\cdot) = \sum_{i=1}^n \theta_{vi} k_v(\mathbf{X}_{vi}, \cdot)$ for some parameter $\boldsymbol{\theta} \in \mathbb{R}^{n|\mathcal{P}|}$ with components $(\theta_{v,i}, i = 1, \dots, n, v = 1, \dots, |\mathcal{P}|)$.

Let $\|\cdot\|$ denotes the usual euclidean norm in $\mathbb{R}^n$. For each $v \in \mathcal{P}$, let K_v be the $n \times n$ matrix with components $(K_v)_{i,i'} = k_v(\mathbf{X}_{vi}, \mathbf{X}_{vi'})$ that satisfies $t(K^{1/2})K^{1/2} = K$. Let $\widehat{f}_0$ and $\widehat{\boldsymbol{\theta}}$ be the minimizer of the following penalized least-squares criteria:

$$C(f_0, \boldsymbol{\theta}) = \frac{1}{n}\|\mathbf{Y} - f_0 1_n - \sum_{v\in\mathcal{P}} K_v\boldsymbol{\theta}_v\|^2 + \frac{1}{\sqrt{n}}\sum_{v\in\mathcal{P}} \gamma_v\|K_v\boldsymbol{\theta}_v\| + \sum_{v\in\mathcal{P}} \mu_v\|K_v^{1/2}\boldsymbol{\theta}_v\|. \tag{17}$$

Then the estimator $\widehat{f}$ defined at Equation (6) satisfies

$$\widehat{f}(\mathbf{x}) = \widehat{f}_0 + \sum_{v\in\mathcal{P}} \widehat{f}_v(\mathbf{x}_v) \text{ with } \widehat{f}_v(\mathbf{x}_v) = \sum_{i=1}^n \widehat{\theta}_{v,i} k_v(\mathbf{X}_{v,i}, \mathbf{x}_v).$$

Because $C(f_0, \boldsymbol{\theta})$ is a convex and separable criteria, we propose to calculate $\widehat{\boldsymbol{\theta}}$ using a block coordinate descent algorithm described in the following section.

Note that the estimator $\widehat{f}$ defined at Equation (6) should satisfy $\|f_v\|_{\mathcal{H}_v} = \|K_v^{1/2}\boldsymbol{\theta}_v\| \leq 1$, or generally $\|f_v\|_{\mathcal{H}_v} \leq r_v$ for some positive r_v, see (16). Usually we have no idea of the value of this upper-bound in practice, and we propose to remove this contraint in the optimization procedure. Nevertheless, if one wants to consider such an additional constraint, the problem can be solved at the price of some additional complication, considering a Lagrangian method for example, see Section 8.7 for more details.

5.1. **Algorithm**

We will assume that for all $v \in \mathcal{P}$ the matrices K_v are strictly definite positive. If it is not the case, one modifies K_v by $K_v + \xi I_n$ where ξ is a small positive value, in order to ensure positive definiteness.

Using a coordinate descent procedure, we minimize the criteria $C(f_0, \boldsymbol{\theta})$ along each group v at a time. At each step of the algorithm, the criteria is minimized as a function of the current block's parameters, while the parameters values for the other blocks are fixed to their current values. The procedure is repeated until convergence, considering for example that the convergence is obtained if the norm of the difference between two consecutive solutions is small enough. See for exemple Boyd et al. [8] for optimization in such context.

For the sake of simplicity, we consider the minimization of the following criteria:

$$C'(f_0, \boldsymbol{\theta}) = \|\mathbf{Y} - f_0 - \sum_{v\in\mathcal{P}} K_v\boldsymbol{\theta}_v\|^2 + \sum_{v\in\mathcal{P}} \gamma'_v\|K_v\boldsymbol{\theta}_v\| + \sum_{v\in\mathcal{P}} \mu'_v\|K_v^{1/2}\boldsymbol{\theta}_v\|. \tag{18}$$

Taking $\gamma'_v = \sqrt{n}\gamma_v$ and $\mu'_v = n\mu_v$, this is exactly the same criteria as the one defined at Equation (17).

Let us begin with the constant term f_0. Because the penalty function does not depend on f_0, minimizing $C'(f_0, \boldsymbol{\theta})$ with respect to f_0 for fixed values of $\boldsymbol{\theta}$ leads to

$$f_0 = Y_\cdot - \sum_v \sum_{i=1}^{n} \left(K_v \boldsymbol{\theta}_v\right)_i / n, \tag{19}$$

where $Y_\cdot$ denotes the mean of $\mathbf{Y}$ and $\left(K_v \boldsymbol{\theta}_v\right)_i$ denotes the i-th component of $K_v \boldsymbol{\theta}_v$. In what follows, we consider a group v, and fix the values of the parameters for all the other groups. We describe the algorithm and postpone the proofs in Section 8.7.

Let us first consider the case where both μ'_v and γ'_v are non zero. If $\partial C'_v$ denotes the subdifferential of $C'(f_0, \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}_v$, we need to solve $0 \in \partial C'_v$, which is equivalent to

$$-2K_v \mathbf{R}_v + 2K_v^2 \boldsymbol{\theta}_v + \gamma'_v s_v + \mu'_v t_v = 0, \tag{20}$$

where

$$\mathbf{R}_v = \mathbf{Y} - f_0 - \sum_{w \neq v} K_w \boldsymbol{\theta}_w$$

and where s_v and t_v satisfy:

$$\begin{aligned} &\text{if } \boldsymbol{\theta}_v = 0 \qquad \|K_v^{-1} s_v\| \leq 1, \text{ and } \|K_v^{-1/2} t_v\| \leq 1 \\ &\text{if } \boldsymbol{\theta}_v \neq 0 \qquad s_v = \frac{K_v^2 \boldsymbol{\theta}_v}{\|K_v \boldsymbol{\theta}_v\|}, \text{ and } t_v = \frac{K_v \boldsymbol{\theta}_v}{\|K_v^{1/2} \boldsymbol{\theta}_v\|}. \end{aligned}$$

The first task is to obtain necessary and sufficient conditions for which the solution $\boldsymbol{\theta}_v = 0$ is the optimal one. Let

$$J(t) = \|2\mathbf{R}_v - \mu'_v K_v^{-1} t\|^2, \text{ and } J^* = \operatorname{argmin}\left\{ J(t), \text{ for } t \in \mathbb{R}^n \text{ such that } \|K_v^{-1/2} t\| \leq 1 \right\}. \tag{21}$$

Then the solution to Equation (20) is zero if and only if $J^* \leq \gamma_v'^2$. Calculating J^* is a ridge regression problem that can be easily solved (see Propositions 8.2 and 8.3 in Section 8.7).

If the solution to Equation (20) is not $\boldsymbol{\theta}_v = 0$, the problem is to solve the subgradient equation:

$$\boldsymbol{\theta}_v = \left(\frac{\mu'_v}{2\|K_v^{1/2} \boldsymbol{\theta}_v\|} I_n + K_v + \frac{\gamma'_v}{2\|K_v \boldsymbol{\theta}_v\|} K_v \right)^{-1} \mathbf{R}_v. \tag{22}$$

Because $\boldsymbol{\theta}_v$ appears in both sides of the equation, a numerical procedure is needed (see Proposition 8.4).

Other cases.

- If all the μ'_v are equal to 0, the parameters $\boldsymbol{\theta}_v$ are not identifiable.
- If all the γ'_v are equal to 0, then we have to solve a classical group-lasso problem with respect to the parameters $\boldsymbol{\theta}'_v$ defined as $\boldsymbol{\theta}'_v = K_v^{1/2} \boldsymbol{\theta}_v$ for all $v \in \mathcal{P}$.
- Let v such that $\mu'_v = 0$, $\gamma'_v \neq 0$ and assume that at least one of the μ'_w is non zero for $w \in \mathcal{P}$, $w \neq v$. Then it is shown in Proposition 8.1 that $\boldsymbol{\theta}_v = 0$ if and only if $2\|R_v\| \leq \gamma'_v$.
- In the same way, if $\gamma'_v = 0$, and $\mu'_v \neq 0$, then $\boldsymbol{\theta}_v = 0$ if and only if $2\|K_v^{1/2} R_v\| \leq \mu'_v$.

Finally the algorithm is the following:

(1) Start with an initial value $\boldsymbol{\theta} = \boldsymbol{\theta}_0$.
(2) Calculate f_0 using Equation (19). For the group v, calculate $\mathbf{R}_v$ and determine if the group v should be excluded or not either by solving the problem defined at Equation (21) if $\mu'_v \neq 0$ and $\gamma'_v \neq 0$, or directly if one of them equals 0. If it is the case, set $\boldsymbol{\theta}_v = 0$. If not, solve Equation (22) to obtain $\boldsymbol{\theta}_v$.
(3) Iterate step 2. over all the groups v.
(4) Iterate step 2. and 3. until convergence.

5.2. **Choice of the tuning parameters**

For each value of the tuning parameters $(\mu'_v, \gamma'_v), v \in \mathcal{P}$, the algorithm provides an estimate of m and of the Sobol indices. The problem for choosing these parameters values is crucial. We propose to restrict this choice by considering tuning parameters proportionnal to known weights: for all $v \in \mathcal{P}$, $\mu'_v = \mu\omega_v$ and $\gamma'_v = \gamma\zeta_v$, where the weights ω_v and ζ_v are fixed. For example, one can take weights that increase with the cardinal of v in order to favour effects with small interaction order between variables. Or, according to the theoretical result given at Theorem 4.1, we can choose $\omega_v = \widehat{\nu}^2_{n,v}$ and $\zeta_v = \widehat{\nu}_{n,v}$, where $\widehat{\nu}_{n,v}$ is an estimate of $\nu_{n,v}$ based on the eigenvalues of the matrix K_v. Any other choice, depending on the problem of interest, may be relevant.

Once the weights are chosen, we estimate m, on the basis of a learning data set $(Y_i, \mathbf{X}_i), i = 1, \dots n$, for a grid of values of (μ, γ). We first set $\gamma = 0$, and calculate $\mu_{\max}$ the smallest value of μ such that the solution to the minimization of

$$\|\mathbf{Y} - f_0 - \sum_{v\in\mathcal{P}} K_v\boldsymbol{\theta}_v\|^2 + \mu \sum_{v\in\mathcal{P}} \omega_v \|K_v^{1/2}\boldsymbol{\theta}_v\|,$$

is $\boldsymbol{\theta}_v = 0$ for all $v \in \mathcal{P}$. Then we can consider $\mu_\ell = \mu_{\max} 2^{-\ell}$ for $\ell \in \{1, \dots, \ell_{\max}\}$, as a grid of values for μ. The grid of values for γ may be chosen after few attempts.

For choosing the final estimator, say $\widehat{f}$, we suppose that we have at our disposal a testing data set $(Y_i^T, \mathbf{X}_i^T), i = 1, \dots n^T$, and we propose two procedures.

Proc. GS: The first one uses the testing data set for estimating the prediction error. Precisely, for each value of (μ, γ) in the grid, let $\widehat{f}_{(\mu,\gamma)}(\cdot)$ be the estimation of m obtained with the learning data set. Then

$$\mathrm{PE}(\mu, \gamma) = \frac{1}{n^T} \sum_{i=1}^{n^T} \left(Y_i^T - \widehat{f}_{(\mu,\gamma)}(\mathbf{X}_i^T) \right)^2$$

estimates the prediction error, and we propose to choose the pair (μ, γ) that minimizes $\mathrm{PE}(\mu, \gamma)$, say $(\widehat{\mu}, \widehat{\gamma})$. Finally the estimator, denoted $\widehat{f}^{\mathrm{GS}}$ is defined as $\widehat{f}^{\mathrm{GS}} = \widehat{f}_{(\widehat{\mu},\widehat{\gamma})}$. In the following, we will refer to this procedure as the Group-Sparse procedure.

Proc. rdg: Doing the parallel with the inconsistency of the lasso for estimating the support of the parameters in the classical regression problem, we propose to choose the tuning parameter that minimizes the risk of the ridge estimator over the support estimated by the ridge-group-sparse procedure. Indeed, if the tuning parameter is chosen to minimize the prediction error, the lasso is not consistent for support estimation (see [24] for example). One idea to overcome this problem, is to choose the tuning parameter that minimizes the risk of the Gauss-lasso estimator which is calculated in two steps: For a given value of the tuning parameter, the estimation of the support of the parameter is estimated using a lasso procedure, then the least square estimator over this support is calculated. When the objective is support estimation, some numerical simulations [33] and theoretical results [18] suggest that it may be more advisable not to apply the selection schemes based on prediction risk to the lasso estimators, but rather to the Gauss-lasso estimators. Our procedure, called the ridge procedure, applies the same idea in the framework of sparse nonparametric estimation. Precisely it considers the collection of supports composed of the different $S_{\widehat{f}_{(\mu,\gamma)}}$ when (μ, γ) belongs to the grid. For each support S in this collection, we estimate f using a ridge procedure assuming that the support of f is S: for a given λ, $f^{\mathrm{rdg}}_{\lambda,S}$ is defined as follows:

$$f^{\mathrm{rdg}}_{\lambda,S} = \mathrm{argmin} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - f_0 - \sum_{v\in S} f_v(\mathbf{X}_{v,i}) \right)^2 + \lambda \sum_{v\in S} \|f_v\|^2_{\mathcal{H}_v}, f = f_0 + \sum_{v\in S} f_v, f_v \in \mathcal{H}_v \right\}.$$

We choose a grid of values for λ, and for each S in the collection and λ in the grid, we estimate the prediction error $\mathrm{PE}(\lambda, S)$ defined as follows:

$$\mathrm{PE}(\lambda, S) = \frac{1}{n^T} \sum_{i=1}^{n^T} \left(Y_i^T - f_{\lambda,S}^{\mathrm{rdg}}(\mathbf{X}_i^T) \right)^2.$$

For each S in the collection, let $\widehat{\lambda}(S)$ be the minimizer of $\mathrm{PE}(\lambda, S)$ when λ varies in the grid, and let $\widehat{S}$ be the minimizer of $\mathrm{PE}(\widehat{\lambda}(S), S)$, then the estimator denoted $\widehat{f}^{\mathrm{rdg}}$ is defined as $\widehat{f}^{\mathrm{rdg}} = f_{\widehat{\lambda}(\widehat{S}),\widehat{S}}^{\mathrm{rdg}}$.

If a testing data set is not available, we can use the classical V-fold cross validation (see [1] for example) either to estimate $\mathrm{PE}(\mu, \gamma)$ or to estimate $\mathrm{PE}(\lambda, S)$.

6. Simulation study

In order to evaluate the performances of our method for estimating a meta-model and sensitivity indices of a function m we carried out a simulation study. We consider the g-function of Sobol defined on $[0,1]^d$ as

$$m(\mathbf{x}) = \prod_{a=1}^{d} \frac{|4x_a - 2| + c_a}{1 + c_a}, c_a > 0,$$

whose Sobol indices can be expressed analytically (see Saltelli et al. [34]). Following the simulation experiment proposed by Durrande et al. [12], we take $d = 5$ and $(c_1, c_2, c_3, c_4, c_5) = (0.2, 0.6, 0.8, 100, 100)$. The lower the value of c_a, the more significant the variable x_a. The variables $X_a, a = 1, \ldots, d$ are independent and uniformly distributed on $[0,1]$. We consider the regression model $Y_i = m(\mathbf{X}_i) + \sigma\varepsilon_i$, for $i = 1, \ldots, n$, with $\mathcal{N}(0,1)$ independent error terms ε_i.

Simulation design. We present the results for $n \in \{50, 100, 200\}$, $\sigma \in \{0, 0.2\}$. For all $a = 1, \ldots, d$, the kernels k_a are the same: we considered the Brownian kernel, $k^b(x, x') = 1 + \min\{x, x'\}$, the Matérn kernel, $k^m(x, x') = (1 + 2|x - x'|)\exp(-2|x - x'|)$, and the Gaussian kernel, $k^g(x, x') = \exp(x - x')^2$.

For each simulation, we generate three independent data sets as follows : a Latin Hypercube Sample of the inputs is simulated to give the matrix $\mathbf{X}$ with n rows and $d = 5$ columns, and a n-sample of independent centered Gaussian variable with variance 1 is simulated. This operation is repeated three times in order to obtain the learning and testing data sets and a third data set for estimating the estimators performances. As explained in Section 5.2, we choose optimal values of the tuning parameters, (μ, γ) by minimizing a prediction error PE, and get an estimator $\widehat{f}$ of m, as well as estimates of the Sobol indices:

$$\widehat{S}_v = \frac{\widehat{\mathrm{Var}}\left(m_v(\mathbf{x}_v)\right)}{\sum_{w \in \mathcal{P}} \widehat{\mathrm{Var}}\left(m_w(\mathbf{x}_w)\right)}.$$

Let Ω_v be the matrix whose components satisfy

$$(\Omega_v)_{i,i'} = \prod_{a \in v} \mathbb{E}_{U \sim P_a}\left(k_{0a}(U, X_{a,i}) k_{0a}(U, X_{a,i'})\right) \text{ for } i, i' = 1, \ldots, n.$$

The estimator $\widehat{\mathrm{Var}}\left(m_v(\mathbf{x}_v)\right)$ is calculated as follows:

$$\widehat{\mathrm{Var}}\left(m_v(\mathbf{x}_v)\right) = \widehat{\alpha}_v^T \Omega_v \widehat{\alpha}_v.$$

We also propose to estimate these quantities by their empirical variances as in Equation (8).

	$\sigma = 0$			$\sigma = 0.2$		
	$n = 50$	$n = 100$	$n = 200$	$n = 50$	$n = 100$	$n = 200$
Proc. GS	0.814	0.920	0.959	0.737	0.835	0.889
Proc. rdg	0.874	0.976	0.989	0.763	0.854	0.892

TABLE 1. Estimated coefficient of determination R^2 for different values of n and σ, with the Matérn kernel.

	$\sigma = 0$			$\sigma = 0.2$		
	$n = 50$	$n = 100$	$n = 200$	$n = 50$	$n = 100$	$n = 200$
Proc. GS	0.033	0.0137	0.0139	0.051	0.028	0.020
Proc. rdg	0.011	0.0009	0.0007	0.042	0.022	0.013

TABLE 2. Estimated empirical risk ER for different values of n and σ with the Matérn kernel.

Performance indicators. To evaluate the performances of our method for estimating a meta-model, we use the classical coefficient of determination R^2 estimated using the third data set $(Y_i^P, \mathbf{X}_i^P), i = 1, \ldots, n$:

$$R^2 = 1 - \frac{\sum_{i=1}^n \left(Y_i^P - \widehat{f}(\mathbf{X}_i^P)\right)^2}{\sum_{i=1}^n \left(Y_i^P - Y_\cdot^P\right)^2}.$$

Moreover we calculate the empirical risk $\mathrm{ER} = \|m - \widehat{f}\|_n^2$. For each simulation s, we get R_s^2 and ER_s and we report the means of these quantities over all simulations.

Similarly, for each v, and each simulation s, we get $\widehat{S}_{v,s}$ and we report its mean, $\widehat{S}_{v,\cdot}$, its estimated standard-error, and to sum up the behaviour of our procedure for estimating the sensitivity indices, we estimate the global error, denoted GE, defined as follows

$$\mathrm{GE} \quad = \quad \sum_v (\widehat{S}_{v,s} - S_v)^2.$$

In order to assess the performances of our procedure for selecting the greatest Sobol Indices, precisely those that are greater than some small quantity as $\rho = 10^{-4}$, we calculate for each group $v \in \mathcal{P}$, the percentage of simulations for which v is in the support of the estimator $\widehat{f}$. Then we average these quantities, on one hand over groups v such that $S_v > \rho$, and on the other hand, over groups v such that $S_v \leq \rho$. Let us denote these quantities $\mathrm{pSel}_{S_v > \rho}$ and $\mathrm{pSel}_{S_v \leq \rho}$ respectively.

For each (μ, γ) in a grid of values, the estimator $\widehat{f}_{\mu,\gamma}$ is defined as the minimizer of the criteria given at Equation (18) taking $\omega_v = \zeta_v = 1$ for all $v \in \mathcal{P}$. In order to save computation time, we restrict the optimisation to sets v such that $|v| \leq 3$. Some preliminary simulations showed that the terms coresponding to $|v| \geq 4$ are nearly always equal to 0.

Choosing the tuning parameters. Let us begin with the comparison of the two methods proposed for choosing the final estimator, see Section 5.2. The results are given in Tables 1 and 2. It appears that the procedure based on the ridge estimator after selection of the groups outperforms the method based on the group-sparse estimator. As expected both methods perform better when n increases, and when $\sigma = 0$.

Similarly, the Sobol indices are better estimated, in the sense of the global error, with the procedure based on the ridge estimate of the metamodel, see Table 3. The means of the estimators for the Sobol indices greater than $\rho = 10^{-4}$ are given in Tables 4 and 5. It appears than $S_{\{1\}}$ is over-estimated using the procedure based on the group-sparse estimator, leading to under-estimate the Sobol indices associated with interactions of order 2. This tendancy is much less pronounced with the procedure based on the ridge estimator.

	$\sigma = 0$			$\sigma = 0.2$		
	$n = 50$	$n = 100$	$n = 200$	$n = 50$	$n = 100$	$n = 200$
Proc. GS	1.91	0.79	0.45	2.41	1.16	0.54
Proc. rdg	0.80	0.10	0.03	1.50	0.47	0.15

TABLE 3. Estimated global error GE $\times$ 100 for different values of n and σ with the Matérn kernel.

	$v = \{1\}$	$v = \{2\}$	$v = \{3\}$	$v = \{1,2\}$	$v = \{1,3\}$	$v = \{2,3\}$	$v = \{1,2,3\}$	sum
S.I.	43.3	24.3	19.2	5.63	4.45	2.50	0.579	99.98
Proc. GS	50.1 (6.2)	26.5 (5.4)	20.9 (4.9)	0.69 (0.8)	0.63 (1.1)	0.51 (0.9)	0.02 (0.07)	99.29
Proc. rdg	45.4 (4.3)	25.3 (3.4)	20.3 (3.5)	3.08 (2.5)	2.18 (2.1)	1.44 (1.8)	0.09 (0.5)	98.66

TABLE 4. The first line of the table gives the true values of the Sobol indices $\times 100$ greater than 10^{-2}, as well as their sum in the last columns. The following lines give the mean of the estimators as well as their standard-error (in parenthesis), calculated over 100 simulations, for $n = 50$, and $\sigma = 0$ with the Matérn kernel.

	$v = \{1\}$	$v = \{2\}$	$v = \{3\}$	$v = \{1,2\}$	$v = \{1,3\}$	$v = \{2,3\}$	$v = \{1,2,3\}$	sum
S.I.	43.3	24.3	19.2	5.63	4.45	2.50	0.579	99.98
Proc. GS	47.5 (5.4)	26.2 (4.9)	19.8 (3.7)	2.35 (1.5)	1.45 (1.2)	0.84 (0.8)	0.03 (0.1)	98.95
Proc. rdg	43.5 (3.9)	25.0 (3.6)	19.7 (2.8)	4.85 (1.7)	3.39 (1.5)	2.02 (1.3)	0.05 (0.3)	99.02

TABLE 5. The first line of the table gives the true values of the Sobol indices $\times 100$ greater than 10^{-2}, as well as their sum in the last columns. The following lines give the mean of the estimators as well as their standard-error (in parenthesis), calculated over 100 simulations, for $n = 100$, and $\sigma = 0.2$ with the Matérn kernel.

	$\text{SI} < \rho$	$\text{SI} \geq \rho$	$v = \{1\}$	$v = \{2\}$	$v = \{3\}$	$v = \{1,2\}$	$v = \{1,3\}$	$v = \{2,3\}$	$v = \{1,2,3\}$
Proc. GS	17.9	68	100	100	100	72	66	67	9
Proc. rdg	6.3	51	100	100	100	72	62	47	4

TABLE 6. The first two columns give respectively $\text{p}Sel_{S_v \leq \rho}$ and $\text{p}Sel_{S_v > \rho}$. The last columns give the values of $\text{p}Sel_v$ for each group v such that $S_v > \rho$. Results for $n = 50$ and $\sigma = 0$ with the Matérn kernel.

Let us now consider the performances of the procedure for selecting the non zero Sobol indices. In Tables 6 and 7 we report the percentages of simulations for which the Sobol indices smaller (respectively greater) than ρ are selected, and for which each of Sobol index greater than ρ is selected. From these results, we conclude that the procedure based on the ridge estimator is more strict for selecting non-zero Sobol indices.

Comparing different kernels. Finally we compare the performances of the procedures for different kernels, see Tables 8 and 9. The means of the estimated empirical risk, ER, and of the global error for estimating the sensitivity indices, GE, are calculated for each kernel. It appears that the Matérn kernel gives the best results, except for the case $n = 50$ and $\sigma = 0$ where the empirical risk of $\widehat{f}^{\text{GS}}$ is smaller for the Brownian kernel.

In practice, one may want to choose the kernel according to the smallest prediction error. For that purpose, we propose to calculate the estimators $\widehat{f}^{\text{GS}}$ and/or $\widehat{f}^{\text{rdg}}$ for each kernel, as described in Section 5.2, as well as their associated prediction errors. Then we choose the kernel for which the prediction error is minimized. The results are reported under the column "mixed" in Tables 8 and 9. It appears that the estimated empirical risks for this "mixed" procedure are nearly equal to the minimum estimated empirical risks over the different kernels.

	SI $< \rho$	SI $\geq \rho$	$v = \{1\}$	$v = \{2\}$	$v = \{3\}$	$v = \{1,2\}$	$v = \{1,3\}$	$v = \{2,3\}$	$v = \{1,2,3\}$
Proc. GS	38	85	100	100	100	100	99	98	18
Proc. rdg	6	58	100	100	100	98	92	77	3

TABLE 7. blablaThe first two columns give respectively p$Sel_{S_v \leq \rho}$ and p$Sel_{S_v > \rho}$. The last columns give the values of pSel_v for each group v such that $S_v > \rho$. Results for $n = 100$ and $\sigma = 0.2$ with the Matérn kernel.

	$n = 50$, $\sigma = 0$				$n = 100$, $\sigma = 0.2$			
	Matérn	Gaussian	Brownian	mixed	Matérn	Gaussian	Brownian	mixed
Proc. GS	0.033	0.054	0.027	0.028	0.033	0.054	0.027	0.028
Proc. rdg	0.011	0.024	0.025	0.011	0.011	0.024	0.025	0.023

TABLE 8. Estimated empirical risk ER: Performances of the procedures according to the kernel choice.

	$n = 50$, $\sigma = 0$				$n = 100$, $\sigma = 0.2$			
	Matérn	Gaussian	Brownian	mixed	Matérn	Gaussian	Brownian	mixed
Proc. GS	1.91	2.49	2.19	1.88	1.16	1.90	1.69	1.16
Proc. rdg	0.80	1.39	1.40	0.85	0.48	0.83	0.73	0.51

TABLE 9. Estimated global error GE $\times$ 100: Performances of the procedures according to the kernel choice.

7. Sketch of proof of Theorem 4.1

We give here a sketch of the proof and we postpone to Section 8 for complete statements. In particular, we denote by C constants that vary from an equation to the other, and we assume that $\sigma = 1$.

The proof of Theorem 4.1 starts in the same way as the proof of Theorem 1 in Raskutti et al. [31]. Nevertheless it differs in several points, in particular because the terms occuring in the decomposition of functions in $\mathcal{H}$ depend on several variables and thus are not independent. Indeed, $f_v(\mathbf{X}_v)$ and $f'_v(\mathbf{X}_{v'})$ are not independent as soon as the groups v and v' share some of the variables $X_a, a = 1, \ldots, d$. Moreover, we do not assume that the function m is in $\mathcal{F}$.

Starting from the definition of $\widehat{f}$, some simple calculation (see Equation (28)) give that for all $f \in \mathcal{F}$

$$C\|m - \widehat{f}\|_n^2 \leq \|m - f\|_n^2 + \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i (\widehat{f}(\mathbf{X}_i) - f(\mathbf{X}_i)) \right| + \sum_{v \in S_f} \left(\gamma_v \|\widehat{f}_v - f_v\|_n + \mu_v \|\widehat{f}_v - f_v\|_{\mathcal{H}_v} \right).$$

If we set $g = \widehat{f} - f$, then $g \in \mathcal{H}$, $g = g_0 + \sum_v g_v$, with $g_v = \widehat{f}_v - f_v$, and for each v, $\|g_v\|_{\mathcal{H}_v} \leq 2$.

The main problem is now to control the empirical process. For each v, letting $\lambda_{n,v}$ as in (13), we state (see Lemma 8.1, page 18) that, with high probability,

$$\text{if } \|g_v\|_n \leq \lambda_{n,v} \|g_v\|_{\mathcal{H}_v} \quad \text{then} \quad \left| \sum_{i=1}^n \varepsilon_i g_v(\mathbf{X}_{v,i}) \right| \leq C n \lambda_{n,v}^2 \|g_v\|_{\mathcal{H}_v}$$

$$\text{if } \|g_v\|_n > \lambda_{n,v} \|g_v\|_{\mathcal{H}_v} \quad \text{then} \quad \left| \sum_{i=1}^n \varepsilon_i g_v(\mathbf{X}_{v,i}) \right| \leq C n \lambda_{n,v} \|g_v\|_n.$$

Therefore, if for all v, μ_v and γ_v satisfy Equation (14), we deduce that with high probability (setting $g = \widehat{f} - f$)

$$C\|m - \widehat{f}\|_n^2 \leq \|m - f\|_n^2 \sum_{v \in S_f} (\gamma_v \|g_v\|_n + \mu_v \|g_v\|_{\mathcal{H}_v}) + \sum_{v \notin S_f} \left(\gamma_v \|\widehat{f}_v\|_n + \mu_v \|\widehat{f}_v\|_{\mathcal{H}_v}\right).$$

Besides we can express the decomposability property of the penalty as follows (see lemma 8.2, page 19): with high probability (in the set where the empirical process is controled as stated above),

$$\sum_{v \notin S_f} \left(\gamma_v \|\widehat{f}_v\|_n + \mu_v \|\widehat{f}_v\|_{\mathcal{H}_v}\right) \leq C \sum_{v \in S_f} (\gamma_v \|g_v\|_n + \mu_v \|g_v\|_{\mathcal{H}_v}).$$

Putting the things together, and noting again that $\|g_v\|_{\mathcal{H}_v} \leq 2$, we obtain the following upper bound

$$C\|m - \widehat{f}\|_n^2 \leq \|m - f\|_n^2 + \sum_{v \in S_f} (\ \mu_v + \gamma_v \|g_v\|_n).$$

The last important step consists in comparing $\sum_{v \in S_f} \|g_v\|_n$ to $\|\sum_{v \in S_f} g_v\|_n$. More precisely, it can be shown (see lemma 8.3, page 19) that for all $v \in \mathcal{P}$, with high probability, we have

$$\|g_v\|_n \leq 2\|g_v\|_{\mathbb{L}^2(P_{\mathbf{X}})} + \gamma_v.$$

Using the orthogonality assumption between the spaces $\mathcal{H}_v$, we have $\sum_{v \in S_f} \|g_v\|_{\mathbb{L}^2(P_{\mathbf{X}})}^2 = \|\sum_v g_v\|_{\mathbb{L}^2(P_{\mathbf{X}})}^2$, and thus we get

$$C\|m - \widehat{f}\|_n^2 \leq \|m - f\|_n^2 + \sum_{v \in S_f} \mu_v + \sum_{v \in S_f} \gamma_v^2 + \|\sum_{v \in S_f} (\widehat{f}_v - f_v)^2\|_{\mathbb{L}^2(P_{\mathbf{X}})}^2.$$

Finally it remains to consider different cases according to the rankings of $\|\widehat{f} - f\|_{\mathbb{L}^2(P_{\mathbf{X}})}^2$, $\|\widehat{f} - f\|_n^2$ and $\sum_{v \in S_f} \mu_v + \gamma_v^2$ to get the result of Theorem 4.1.

8. Proofs

Recall that we cconsider the regression model defined at Equation (1), where $\mathbf{X}$ has distribution $P_{\mathbf{X}} = P_1 \times \ldots \times P_d$ defined on $\mathcal{X}$ a compact subset of $\mathbb{R}^d$, and ε is distributed as $\mathcal{N}(0, \sigma^2)$. We denote by $P_{\mathbf{X},\boldsymbol{\varepsilon}}$ the distribution of $(\mathbf{X}, \boldsymbol{\varepsilon})$. We observe a n sample $(Y_i, \mathbf{X}_i), i = 1, \ldots, n$ with law $P_{\mathbf{X},\boldsymbol{\varepsilon}}$.

The notation and the procedure are given in Sections 2 and 4.

Let us add on few notations that will be used along the proofs.

For $v \in \mathcal{P}$ we denote $|v|$ the cardinal of v. For a function $\phi : \mathbb{R}^{|v|} \mapsto \mathbb{R}$, we denote $V_{n,\varepsilon}$ the empirical process defined by

$$V_{n,\varepsilon}(\phi) = \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \phi(\mathbf{X}_{v,i}). \tag{23}$$

For the sake of simplicity we assume $\sigma = 1$. Moreover, we set $R' = 1$, see (10). Consequently, for any function $f \in \mathcal{H}$, $\|f_v\|_{\mathcal{H}_v} \leq 1$, and $\|f_v\|_\infty \leq \|f\|_{\mathcal{H}_v}$. The proofs can be done exactly in the same way by considering the general case. In the proofs, the $\|\cdot\|_{\mathbb{L}^2(P_{\mathbf{X}})}$ norm will be denoted by $\|\cdot\|_2$.

8.1. **Proof of Theorem 4.1**

The proof is based on four main lemmas proved in Section 8.5. In Section 8.4 other lemmas used all along the proof are stated. Their proof are postponed to Section 8.6.

Let us first establish inequalities that will be used in the following. Let $f \in \mathcal{H}$ and $v \in S_f$ (see (9)).

Using that for any $v \in S_f$, and any norm $\|\cdot\|$ in $\mathcal{H}_v$, $\|f_v\| - \|\widehat{f}_v\| \leq \|f_v - \widehat{f}_v\|$ and that for any $v \notin S_f$, $\|f_v\| = 0$, we get that

$$\sum_{v\in\mathcal{P}} \mu_v \|f_v\|_{\mathcal{H}_v} - \sum_{v\in\mathcal{P}} \mu_v \|\widehat{f}_v\|_{\mathcal{H}_v} \leq \sum_{v\in S_f} \mu_v \|f_v - \widehat{f}_v\|_{\mathcal{H}_v} - \sum_{v\in S_f^c} \mu_v \|\widehat{f}_v\|_{\mathcal{H}_v}, \tag{24}$$

$$\sum_{v\in\mathcal{P}} \gamma_v \|f_v\|_n - \sum_{v\in\mathcal{P}} \gamma_v \|\widehat{f}_v\|_n \leq \sum_{v\in S_f} \gamma_v \|f_v - \widehat{f}_v\|_n - \sum_{v\in S_f^c} \gamma_v \|\widehat{f}_v\|_n. \tag{25}$$

Combining (24), and (25), to the fact that for any function $f \in \mathcal{H}$, $\mathcal{L}(\widehat{f}) \leq \mathcal{L}(f)$, we obtain that

$$\|m - \widehat{f}\|_n^2 \leq \|m - f\|_n^2 + B$$

with

$$B = 2V_{n,\varepsilon}(\widehat{f} - f) + \sum_{v\in S_f} \left[\mu_v \|\widehat{f}_v - f_v\|_{\mathcal{H}_v} + \gamma_v \|\widehat{f}_v - f_v\|_n\right] - \sum_{v\in S_f^c} \left[\mu_v \|\widehat{f}_v\|_{\mathcal{H}_v} + \gamma_v \|\widehat{f}_v\|_n\right]. \tag{26}$$

If $\|m - f\|_n^2 \geq B$, we immediately get the result since in that case

$$\|m - \widehat{f}\|_n^2 \leq 2\|m - f\|_n^2 \leq 2\|m - f\|_n^2 + \sum_{v\in S_f} \mu_v + \sum_{v\in S_f} \gamma_v^2.$$

If $\|m - f\|_n^2 < B$, we get that

$$\|\widehat{f} - m\|_n^2 \leq 2B \tag{27}$$

$$\leq 4|V_{n,\varepsilon}(\widehat{f} - f)| + 2\sum_{v\in S_f} \left[\mu_v \|\widehat{f}_v - f_v\|_{\mathcal{H}_v} + \gamma_v \|\widehat{f}_v - f_v\|_n\right]. \tag{28}$$

The control of the empirical process $|V_{n,\varepsilon}(\widehat{f} - f)|$ is given by the following lemma (proved in Section 8.5.1, page 26).

Lemma 8.1. *Let $V_{n,\varepsilon}$ be defined in (23). For any f in $\mathcal{F}$, we consider the event $\mathcal{T}$ defined as*

$$\mathcal{T} = \left\{\forall f \in \mathcal{F}, \forall v \in \mathcal{P}, |V_{n,\varepsilon}(\widehat{f}_v - f_v)| \leq \kappa\lambda_{n,v}^2 \|\widehat{f}_v - f_v\|_{\mathcal{H}_v} + \kappa\lambda_{n,v} \|\widehat{f}_v - f_v\|_n\right\}, \tag{29}$$

where the quantities $\lambda_{n,v}$ are defined by Equation (13) and where $\kappa = 10 + 4\Delta$. Then, for some positive constants c_1, c_2,

$$P_{\mathbf{X},\boldsymbol{\varepsilon}}(\mathcal{T}) \geq 1 - c_1 \sum_{v\in\mathcal{P}} \exp(-nc_2\lambda_{n,v}^2).$$

Conditionning on $\mathcal{T}$, Inequality (28) becomes

$$\|\widehat{f} - m\|_n^2 \leq 4\kappa \sum_{v\in\mathcal{P}} \left[\lambda_{n,v}^2 \|\widehat{f}_v - f_v\|_{\mathcal{H}_v} + \lambda_{n,v} \|\widehat{f}_v - f_v\|_n\right] + 2\sum_{v\in S_f} \left[\mu_v \|\widehat{f}_v - f_v\|_{\mathcal{H}_v} + \gamma_v \|\widehat{f}_v - f_v\|_n\right],$$

which may be decomposed as follows

$$\|\widehat{f} - m\|_n^2 \leq \sum_{v\in S_f} \left[4\kappa\lambda_{n,v}^2 + 2\mu_v\right] \|\widehat{f}_v - f_v\|_{\mathcal{H}_v} + \sum_{v\in S_f} \left[4\kappa\lambda_{n,v} + 2\gamma_v\right] \|\widehat{f}_v - f_v\|_n +$$
$$4\sum_{v\in S_f^c} \kappa\lambda_{n,v}^2 \|\widehat{f}_v - f_v\|_{\mathcal{H}_v} + 4\sum_{v\in S_f^c} \kappa\lambda_{n,v} \|\widehat{f}_v - f_v\|_n.$$

If we choose $C_1 \geq \kappa$ in Theorem 4.1, then $\kappa\lambda_{n,v}^2 \leq \mu_v/2$ and $\kappa\lambda_{n,v} \leq \gamma_v/2$ and the previous inequality becomes

$$\|\widehat{f} - m\|_n^2 \leq 6 \sum_{v \in S_f} \left[\mu_v \|\widehat{f}_v - f_v\|_{\mathcal{H}_v} + \gamma_v \|\widehat{f}_v - f_v\|_n\right] + 4 \sum_{v \in S_f^c} \left[\mu_v \|\widehat{f}_v\|_{\mathcal{H}_v} + \gamma_v \|\widehat{f}_v\|_n\right]. \tag{30}$$

Next we use the decomposability property of the penalty expressed in the following lemma (proved in Section 8.5.2 page 28).

Lemma 8.2. *For any $f \in \mathcal{F}$, under the assumptions of Theorem 4.1 with $C_1 \geq \kappa$, conditionnally on $\mathcal{T}$, see (29), we have*

$$\sum_{v \in S_f^c} \mu_v \|\widehat{f}_v\|_{\mathcal{H}_v} + \sum_{v \in S_f^c} \gamma_v \|\widehat{f}_v\|_n \leq 3 \sum_{v \in S_f} \mu_v \|\widehat{f}_v - f_v\|_{\mathcal{H}_v} + 3 \sum_{v \in S_f} \gamma_v \|\widehat{f}_v - f_v\|_n.$$

Hence, by combining (30) and Lemma (8.2) we obtain

$$\|\widehat{f} - m\|_n^2 \quad \leq \quad 18 \sum_{v \in S_f} \left[\mu_v \|\widehat{f}_v - f_v\|_{\mathcal{H}_v} + \gamma_v \|\widehat{f}_v - f_v\|_n\right].$$

For each v, $\|\widehat{f}_v - f_v\|_{\mathcal{H}_v} \leq 2$ (because the functions $\widehat{f}_v$ et f_v belong to the class $\mathcal{F}$, see (5)), and consequently, for some constant C,

$$\|\widehat{f} - m\|_n^2 \leq C \left\{ \sum_{v \in S_f} \mu_v + \sum_{v \in S_f} \gamma_v \|\widehat{f}_v - f_v\|_n \right\}. \tag{31}$$

To finish the proof it remains to compare the two quantities $\sum_{v \in S_f} \|\widehat{f}_v - f_v\|_n^2$ and $\|\sum_{v \in S_f} \widehat{f}_v - f_v\|_n^2$. For that purpose we show that $\|\sum_{v \in S_f} \widehat{f}_v - f_v\|_n$ is less than $\|\sum_{v \in S_f} \widehat{f}_v - f_v\|_2^2$ plus an additive term coming from concentration results (see the Lemma given below). Next, thanks to the orthogonality of the spaces $\mathcal{H}_v$ with respect to $\mathbb{L}^2(P_{\mathbf{X}})$, $\|\sum_{v \in S_f} \widehat{f}_v - f_v\|_2^2 = \sum_{v \in S_f} \|\widehat{f}_v - f_v\|_2^2$. To conclude, it remains to consider several cases, according to the rankings of $\|\sum_{v \in S_f} \widehat{f}_v - f_v\|_2^2$, $\|\sum_{v \in S_f} \widehat{f}_v - f_v\|_n^2$, and $d^2(f)$. This is the subject of the following lemma whose proof is given in Section 8.5.3, page 29.

Lemma 8.3. *For $f \in \mathcal{H}$, let $\mathcal{A}$ be the event*

$$\mathcal{A} = \left\{ \forall f \in \mathcal{F}, \forall v \in \mathcal{P},\ \|\widehat{f}_v - f_v\|_n \leq 2\|\widehat{f}_v - f_v\|_2 + \gamma_v \right\}. \tag{32}$$

Then, for some positive constant c_2,

$$P_{\mathbf{X},\boldsymbol{\varepsilon}}(\mathcal{A}) \geq 1 - \sum_v \exp(-nc_2\gamma_v^2).$$

On the set $\mathcal{A}$, Inequality (31) provides that, for all $K > 0$

$$\begin{aligned}
\frac{1}{C}\|\widehat{f} - m\|_n^2 &\leq \sum_{v\in S_f}\left(\mu_v + 2\gamma_v\|\widehat{f}_v - f_v\|_2 + \gamma_v^2\right) &(33)\\
&\leq \sum_{v\in S_f}\left(\mu_v + (1+K)\gamma_v^2 + \frac{1}{K}\|\widehat{f}_v - f_v\|_2^2\right), &(34)\\
&\leq \sum_{v\in S_f}\left(\mu_v + (1+K)\gamma_v^2\right) + \frac{1}{K}\sum_{v\in\mathcal{P}}\|\widehat{f}_v - f_v\|_2^2\\
&\leq \sum_{v\in S_f}\left(\mu_v + (1+K)\gamma_v^2\right) + \frac{1}{K}\|\sum_{v\in\mathcal{P}}\widehat{f}_v - f_v\|_2^2. &(35)
\end{aligned}$$

Inequality (34) uses the inequality $2ab \leq \frac{1}{K}a^2 + Kb^2$ for all positive K, and Inequality (35) uses the orthogonality with respect to $\mathbb{L}^2(P_{\mathbf{X}})$.

In the following we have to consider several cases, according to the rankings of $\|\sum_{v\in\mathcal{P}}\widehat{f}_v - f_v\|_2$, $\|\sum_{v\in\mathcal{P}}\widehat{f}_v - f_v\|_n$ and $d(f)$ defined as follows

$$d^2(f) = \max\left(\sum_{v\in S_f}\gamma_v^2, \sum_{v\in S_f}\mu_v\right). \tag{36}$$

More precisely, we consider three cases

Case 1: $\|\sum_{v\in\mathcal{P}}\widehat{f}_v - f_v\|_2 \leq \|\sum_{v\in\mathcal{P}}\widehat{f}_v - f_v\|_n$.
Case 2: $\|\sum_{v\in\mathcal{P}}\widehat{f}_v - f_v\|_n \leq \|\sum_{v\in\mathcal{P}}\widehat{f}_v - f_v\|_2 \leq d(f)$
Case 3: $\|\sum_{v\in\mathcal{P}}\widehat{f}_v - f_v\|_n \leq \|\sum_{v\in\mathcal{P}}\widehat{f}_v - f_v\|_2$ and $d(f) \leq \|\sum_{v\in\mathcal{P}}\widehat{f}_v - f_v\|_2$.

Case 1: From (35), for any $f \in \mathcal{H}$, we get

$$\frac{1}{C}\|\widehat{f} - m\|_n^2 \leq \sum_{v\in S_f}\left(\mu_v + (1+K)\gamma_v^2\right) + \frac{1}{K}\|\widehat{f} - f\|_n^2.$$

Hence, using that for all $K' > 0$,

$$\|\widehat{f} - f\|_n^2 \leq (1+K')\|\widehat{f} - m\|_n^2 + (1+1/K')\|f - m\|_n^2, \tag{37}$$

we obtain for a suitable choice of K', say $1 + K' < K/C$, that, for some positive constant C',

$$\|\widehat{f} - m\|_n^2 \leq C'\left[\|f - m\|_n^2 + \sum_{v\in S_f}\mu_v + \sum_{v\in S_f}\gamma_v^2\right].$$

This shows the result in Case 1.

Case 2: Inequality (35) becomes

$$\frac{1}{C}\|\widehat{f} - m\|_n^2 \leq \sum_{v\in S_f}\left(\mu_v + (1+K)\gamma_v^2\right) + \frac{1}{K}d^2(f),$$

which gives the expected result since $d^2(f) = \max\left\{\sum_{v\in S_f}\mu_v, \sum_{v\in S_f}\gamma_v^2\right\}$.

Case 3: Recall that in this case, $\|\sum_{v\in\mathcal{P}}\widehat{f}_v - f_v\|_n \le \|\sum_{v\in\mathcal{P}}\widehat{f}_v - f_v\|_2$ and $d(f) \le \|\sum_{v\in\mathcal{P}}\widehat{f}_v - f_v\|_2$. This case is solved by applying the following Lemma (proved in Section 8.5.4, page 29), which states that with high probability, $\|\widehat{f} - f\|_2 \le \sqrt{2}\|\widehat{f} - f\|_n$.

Lemma 8.4. *Let $f = \sum_v f_v \in \mathcal{F}$ with support S_f, $d(f)$ be defined by (36), and let $\mathcal{G}(f)$ be the class of functions written as $g = \sum_{v\in\mathcal{P}} g_v$, such that $\|g_v\|_{\mathcal{H}_v} \le 2$ satisfying for all $f \in \mathcal{F}$*

$$\textbf{C1} \qquad \sum_{v\in\mathcal{P}} \mu_v\|g_v\|_{\mathcal{H}_v} + \sum_{v\in\mathcal{P}} \gamma_v\|g_v\|_n \le \sum_{v\in S_f} 4\mu_v\|g_v\|_{\mathcal{H}_v} + \sum_{v\in S_f} 4\gamma_v\|g_v\|_n$$

$$\textbf{C2} \qquad \sum_{v\in S_f} \gamma_v\|g_v\|_n \le 2\sum_{v\in S_f} \gamma_v\|g_v\|_2 + \sum_{v\in S_f} \gamma_v^2$$

$$\textbf{C3} \qquad \|g\|_n \le \|g\|_2.$$

Then the event

$$\left\{\|g\|_n^2 \ge \frac{\|g\|_2^2}{2}, \quad \|g\|_2 \ge d(f)\right\}$$

has probability greater than $1 - \exp(-nc_3\sum_{v\in S_f}\lambda_{n,v}^2)$ for some constant c_3.

Note that Assumption $n\lambda_{n,v}^2 \ge -C_2\log(\lambda_{n,v})$ implies that $\lambda_{n,v} = K_{n,v}/\sqrt{n}$ with $K_{n,v} \to \infty$. Then, if f is such that $|S_f| \ge 1$, $\exp(-nc_3\sum_{v\in S_f}\lambda_{n,v}^2) \le \exp(-c_3\min_{v\in\mathcal{P}} K_{n,v})$. If f is such that $|S_f| = 0$, then Condition **C1** is not satisfied except if $g_v = 0$ for all $v \in \mathcal{P}$. Because we will apply Lemma 8.4 to $g_v = \widehat{f}_v - f_v$, this event has probability 0. Therefore the event

$$\mathcal{C} = \left\{\forall f \in \mathcal{F}, \text{ such that } g = \sum_{v\in\mathcal{P}}(\widehat{f}_v - f_v) \in \mathcal{G}(f), \text{ and } \|g\|_n^2 \ge \frac{\|g\|_2^2}{2}, \ \|g\|_2^2 \ge d(f)\right\} \tag{38}$$

has probability greater than $1 - \eta/3$ for some $0 < \eta < 1$.

Conditionning on the events $\mathcal{T}$ and $\mathcal{A}$ (defined by (29) and (32)), and according to Lemma 8.2, $\sum_{v\in\mathcal{P}}(\widehat{f}_v - f_v)$ belongs to the set $\mathcal{G}(f)$. According to (35), we conclude in the same way as in Case 1.

Finally, it remains to quantify $P_{\mathbf{X},\boldsymbol{\varepsilon}}(\mathcal{T}\cap\mathcal{A}\cap\mathcal{C})$. Following Lemma 8.1, and Lemma 8.3, $\mathcal{T}$, respectively $\mathcal{A}$, has probability greater than $1 - c_1\sum_{v\in\mathcal{P}}\exp(-nc_2\lambda_{n,v}^2)$, respectively $1 - \sum_v \exp(-n\gamma_v^2)$. Each of these probabilities is greater $1 - \eta/3$ thanks to the assumption $n\lambda_{n,v}^2 \ge -C_2\log\lambda_{n,v}$.

8.2. **Proof of Corrolary 4.1**

We start from Theorem 4.1 which states that with high probability,

$$\|\widehat{f} - m\|_n^2 \le C\left\{\|f - m\|_n^2 + \sum_{v\in S_f}\mu_v + \sum_{v\in S_f}\gamma_v^2\right\},$$

and use that for all $\theta > 0$,

$$\|\widehat{f} - m\|_2^2 \le (1+\theta)\|\widehat{f} - f\|_2^2 + (1 + \frac{1}{\theta})\|m - f\|_2^2. \tag{39}$$

For d defined by (36), we consider once again the three cases defined page 20.

Case 1: According to (37) and (39), we get the result since

$$\|\widehat{f} - m\|_2^2 \le (1+\theta)\ \|\ \widehat{f} - f\ \|_n^2 + (1 + \frac{1}{\theta})\|m - f\|_2^2.$$

Case 2: We directly obtain that

$$\|\widehat{f} - m\|_2^2 \leq (1+\theta)d^2(f) + (1+\frac{1}{\theta})\|m-f\|_2^2.$$

Case 3: Recall that in this case, $\|\sum_{v\in\mathcal{P}}\widehat{f}_v - f_v\|_n \leq \|\sum_{v\in\mathcal{P}}\widehat{f}_v - f_v\|_2$ and $d(f) \leq \|\sum_{v\in S}\widehat{f}_v - f_v\|_2$. Apply Lemma 8.4 (page 21) and conclude that conditionning on the events $\mathcal{T}$ and $\mathcal{A}$, defined by (29) and (32), then $\sum_{v\in\mathcal{P}}\widehat{f}_v - f_v$ belongs to $\mathcal{G}(f)$ defined in Lemma 8.4. Now, conditionning on the event $\mathcal{C}$ we get the result since

$$\|\sum_{v\in\mathcal{P}}\widehat{f}_v - f_v\|_2^2 \leq 2\|\sum_{v\in\mathcal{P}}\widehat{f}_v - f_v\|_n^2.$$

□

8.3. **Rate of convergence**

Recall that we consider the case where the variables $X_1,\ldots,X_d$ have the same distribution P_1 on $\mathcal{X}_1 \subset \mathbb{R}$, and where the unidimensionnal kernels k_{0a} are all identical.

In this context, our goal is to show that the rate $\nu_{n,v}$ defined at Equation (12) is bounded above by a term of order $n^{-\alpha/(2\alpha+1)}(\log n)^\gamma$, where $\gamma \geq (|v|-1)\alpha/(2\alpha-1)$.

We start from the fact that, $k_v(\mathbf{x}_v,\mathbf{x}'_v) = \prod_{a\in v} k_0(x_a,x'_a)$, with a kernel k_0 admitting an eigen expansion given by

$$k_0(x,x') = \sum_{\ell\geq 1}\omega_{0,\ell}\zeta_\ell(x)\zeta_\ell(x')$$

where the eigenvalues $\omega_{0,\ell}$ are non negative and ranged in the decreasing order at the rate $\ell^{-2\alpha}$ for some $\alpha > 1/2$, and where the ζ_ℓ are the associated eigen functions, orthonormal with respect to $\mathbb{L}^2(P_1)$.

Therefore the kernel k_v admits the following expansion

$$k_v(\mathbf{x}_v,\mathbf{x}_v^{'}) = \sum_{\boldsymbol{\ell}=(\ell_1\ldots\ell_{|v|})} \underbrace{\prod_{a=1}^{|v|}\omega_{0,\ell_a}}_{\omega_{v,\boldsymbol{\ell}}} \underbrace{\prod_{a=1}^{|v|}\zeta_{\ell_a}(x_a)}_{\zeta_{v,\boldsymbol{\ell}}(\mathbf{x}_v)} \underbrace{\prod_{a=1}^{|v|}\zeta_{\ell_a}(x_a^{'})}_{\zeta_{v,\boldsymbol{\ell}}(\mathbf{x}'_v)}.$$

According to this expansion the $\omega_{v,\boldsymbol{\ell}}$ are of order $(\prod_{a=1}^{|v|}\ell_a)^{-2\alpha}$.

In order to control the rate $\nu_{n,v}$ defined as

$$\nu_{n,v} = \inf\left\{t \text{ such that } Q_{n,v}(t) \leq \Delta t^2\right\}$$

, we have to calculate an upper bound for $Q_{n,v}^2(t)$ We start with the following inequalities that hold up to some constant, for $t^{-1/\alpha} > 1$

$$\begin{aligned}
Q_{n,v}^2(t) &= \frac{5}{n}\sum_{\boldsymbol{\ell}}\min(t^2,\omega_{v,\boldsymbol{\ell}}) \\
&\lesssim \frac{1}{n}t^2\sum_{\boldsymbol{\ell}=(\ell_1\ldots\ell_{|v|})} I(\ell_1^{-2\alpha}\times\ldots\times\ell_{|v|}^{-2\alpha}\geq t^2) + \frac{1}{n}\sum_{\boldsymbol{\ell}=(\ell_1\ldots\ell_{|v|})}\prod_{a=1}^{|v|}\ell_a^{-2\alpha} \\
&\lesssim \frac{1}{n}t^2\sum_{\boldsymbol{\ell}=(\ell_1\ldots\ell_{|v|})} I(\ell_1\times\ldots\times\ell_{|v|}\leq t^{-1/\alpha}) + \frac{1}{n}\left(\sum_{j\geq 1}\frac{1}{j^{2\alpha}}\right)^{|v|} \qquad (40)
\end{aligned}$$

Now let us mention that $\alpha > 1/2$, $\sum_{j\geq 1} \frac{1}{j^{2\alpha}}$ is a constant that depends on α. We thus focus on the first term in the right hand side of Equation (40).

Let $u = t^{-1/\alpha} \geq 1$ and let $B_{|v|}$ be defined as follows:

$$B_{|v|} = \sum_{\boldsymbol{\ell}=(\ell_1 \dots \ell_{|v|})} I(\ell_1 \leq u, \dots, \ell_{|v|} \leq u).$$

Let us prove that

$$B_{|v|} \leq u \left(1 + \log(u)\right)^{|v|-1}\Big). \tag{41}$$

Proof of Equation (41): First note that

$$B_1 = \sum_{\ell \geq 1} I(\ell \leq u) \leq u I(u \geq 1).$$

In the same way,

$$\begin{aligned} B_2 = \sum_{\ell_1 \geq 1, \ell_2 \geq 1} I(\ell_1 \ell_2 \leq u) &= \sum_{\ell_1 \geq 1, \ell_2 \geq 1} I(\ell_1 \leq u) I(\ell_2 \leq u/\ell_1) = \sum_{\ell_1 \geq 1} I(\ell_1 \leq u) \sum_{\ell_2 \geq 1} I(\ell_2 \leq u/\ell_1) \\ &\leq \sum_{\ell_1 \geq 1} I(\ell_1 \leq u) \frac{u}{\ell_1} I(\frac{u}{\ell_1} \geq 1) = u \sum_{\ell_1 \geq 1} \frac{1}{\ell_1} I(\ell_1 \leq u) \\ &\leq u \left(1 + \log(u)\right). \end{aligned}$$

More generally,

$$\begin{aligned} B_{|v|} &= \sum_{\ell_1 \geq 1, \dots \ell_{|v|} \geq 1} I(\ell_1 \dots \ell_{|v|} \leq u) \\ &= \sum_{\ell_1 \geq 1, \dots, \ell_{|v|-1} \geq 1} I(\ell_1 \dots \ell_{|v|-1} \leq u) \sum_{\ell_{|v|} \geq 1} I(\ell_{|v|} \leq u/\ell_1 \dots \ell_{|v|-1}) \\ &\leq u \sum_{\ell_1 \geq 1, \dots, \ell_{|v|-1} \geq 1} \frac{1}{\ell_1 \dots \ell_{|v|-1}} I(\ell_1 \dots \ell_{|v|-1} \leq u) \end{aligned}$$

Let $A_{|v|}$ be defined as follows:

$$A_{|v|} = \sum_{\ell_1 \geq 1, \dots \ell_{|v|} \geq 1} \frac{1}{\ell_1 \dots \ell_{|v|}} I(\ell_1 \dots \ell_{|v|} \leq u),$$

then we get $B_{|v|} \leq u A_{|v|-1}$. If we show that

$$A_{|v|} \leq \left(1 + \log(u)\right)^{|v|}, \tag{42}$$

then Inequality (41) is proved.

Proof of Equation (42) :

$$A_1 = \sum_{\ell \geq 1} \frac{1}{\ell} I(\ell \leq u) = 1 + \sum_{\ell \geq 2} \frac{1}{\ell} I(\ell \leq u) \leq 1 + \int_1^u \frac{1}{v} dv = 1 + \log(u)$$

$$\begin{aligned} A_2 = \sum_{\ell_1\geq 1,\ell_2\geq 1} \frac{1}{\ell_1\ell_2} I(\ell_1\ell_2 \leq u) &= \sum_{\ell_1\geq 1} \frac{1}{\ell_1} I(\ell_1 \leq u) \sum_{\ell_2\geq 1} \frac{1}{\ell_2} I(\ell_2 \leq u/\ell_1) \\ &= \sum_{\ell_1\geq 1} \frac{1}{\ell_1} I(\ell_1 \leq u) \left(1+\log(u/\ell_1)\right) \\ &= (1+\log(u))^2 - \sum_{\ell_1\geq 1} \frac{\log(\ell_1)}{\ell_1} I(\ell_1 \leq u) \\ &\leq (1+\log(u))^2 . \end{aligned}$$

In the same way we have

$$\begin{aligned} A_{|v|} &= \sum_{\ell_1\geq 1,\ldots\ell_{|v|}\geq 1} \frac{1}{\ell_1\ldots\ell_{|v|}} I(\ell_1\ldots\ell_{|v|} \leq u) \\ &= \sum_{\ell_1\geq 1,\ldots,\ell_{|v|-1}\geq 1} \frac{1}{\ell_1\ldots\ell_{|v|-1}} I(\ell_1\ldots\ell_{|v|-1} \leq u) \sum_{\ell_{|v|}\geq 1} \frac{1}{\ell_{|v|}} I(\ell_{|v|} \leq u/\ell_1\ldots\ell_{|v|-1}) \\ &\leq \sum_{\ell_1\geq 1,\ldots,\ell_{|v|-1}\geq 1} \frac{1}{\ell_1\ldots\ell_{|v|-1}} I(\ell_1\ldots\ell_{|v|-1} \leq u) \left(1+\log(u/\ell_1\ldots\ell_{|v|-1})\right) \\ &\leq (1+\log(u))^{|v|} . \end{aligned}$$

And Bound (42) is proved.

Rate of convergence. Let us come back to the control of the rate $\nu_{n,v} = \inf\left\{t \text{ such that } Q_{n,v}(t) \leq \Delta t^2\right\}$. Thanks to (41) we obtain that, up to some constant that depends on $|v|$ and α,

$$\begin{aligned} Q^2_{n,v}(t) &\lesssim \frac{1}{n} t^2 \sum_{\boldsymbol{\ell}} I(\ell_1 \times \ldots \times \ell_{|v|} \leq t^{-1/\alpha}) + \frac{1}{n} \left(\sum_{j\geq 1} \frac{1}{j^{2\alpha}}\right)^{|v|} \\ &\lesssim \frac{1}{n} t^{2-1/\alpha} \left(1 - \frac{1}{\alpha}\log(t)\right)^{|v|-1} + \frac{1}{n} \end{aligned}$$

It remains now to find t such that, up to constant

$$\frac{1}{\sqrt{n}} t^{1-1/2\alpha} \left(1 - \frac{1}{\alpha}\log(t)\right)^{(|v|-1)/2} \leq t^2$$

If $t = n^{-\beta}(\log(n))^\gamma$ with $\beta = \alpha/(1+2\alpha)$, $\gamma > 0$, $\alpha > 1/2$, then

$$\begin{aligned} 1 - \frac{1}{\alpha}\log(t) &= 1 - \frac{1}{\alpha}\log\left(n^{-\alpha/(1+2\alpha)}(\log(n))^\gamma\right) \\ &= 1 - \frac{1}{\alpha}\left(-\frac{\alpha}{1+2\alpha}\log(n) + \gamma\log\log(n)\right) \\ &= 1 + \frac{1}{1+2\alpha}\log(n) - \frac{\gamma}{\alpha}\log\log(n) \\ &\leq \log(n) \text{ as soon as } \log(n) > 1 + \frac{1}{2\alpha}. \end{aligned}$$

Therefore $\nu_{n,v}$ will be smaller than the infimum of t such that

$$\frac{1}{\sqrt{n}} t^{1-1/2\alpha} \left(\log(n)\right)^{(|v|-1)/2} \leq t^2,$$

which is satisfied if $\gamma \geq (|v|-1)\alpha/(2\alpha-1)$.

8.4. **Intermediate Lemmas**

For $v \in \mathcal{P}$, let $\mathcal{H}_v$ be the RKHS associated to the self reproducing kernel k_v. Let $Q_{n,v}$ and $\nu_{n,v}$ and be defined by Equations (11) and (12). For any function $g_v \in \mathcal{H}_v$, let $V_{n,\varepsilon}$ be defined at Equation (23) and consider the following processes

$$W_{n,2,v}(t) = \sup\left\{|V_{n,\varepsilon}(g_v)|,\ \|g_v\|_{\mathcal{H}_v} \leq 2, \|g_v\|_2 \leq t\right\} \quad (43)$$

$$W_{n,n,v}(t) = \sup\left\{|V_{n,\varepsilon}(g_v)|, \|g_v\|_{\mathcal{H}_v} \leq 2,\ \|g_v\|_n \leq t\right\}. \quad (44)$$

Lemma 8.5. *If $E_{\mathbf{X},\boldsymbol{\varepsilon}}$ denotes the expectation with respect to the distribution of $(\mathbf{X},\varepsilon)$, we have for all $t>0$,*

$$E_{\mathbf{X},\boldsymbol{\varepsilon}} W_{n,2,v}(t) \leq Q_{n,v}(t).$$

Its proof is given in Section 8.6.1 page 33.

Lemma 8.6. *Let $b>0$ and let $\mathcal{G}(t)$ be the following class of functions:*

$$\mathcal{G}(t) = \left\{g_v \in \mathcal{H}_v, \|g_v\|_{\mathcal{H}_v} \leq 2, \|g_v\|_2 \leq t, \|g_v\|_\infty \leq b\right\}. \quad (45)$$

Let $\Omega_{v,t}$ be the event defined as

$$\Omega_{v,t} = \left\{\sup\left\{|\|g_v\|_2 - \|g_v\|_n|,\ g_v \in \mathcal{G}(t)\right\} \leq \frac{bt}{2}\right\}. \quad (46)$$

Then for any $t \geq \nu_{n,v}$, the event $\Omega_{v,t}$ has probability greater than $1-\exp(-c_2 n t^2)$, for some positive constant c_2.

Its proof is given in Section 8.6.2, page 34.

Lemma 8.7. *For any function $g_v \in \mathcal{H}_v$ satisfying $\|g_v\|_{\mathcal{H}_v} \leq 2$, $\|g_v\|_\infty \leq b$ and $\|g_v\|_2 \geq t$, for all $t \geq \nu_{n,v}$ and $b \geq 1$, the event*

$$\left(1-\frac{b}{2}\right)\|g_v\|_2 \leq \|g_v\|_n \leq \left(1+\frac{b}{2}\right)\|g_v\|_2$$

has probabilty greater than $1-\exp(-c_2 n t^2)$ for some positive constant c_2.

Its proof is given in Section 8.6.3, page 35.

Lemma 8.8. *If $E_{\boldsymbol{\varepsilon}}$ denotes the expectation with respect to the distribution of ε, we have*

$$P_{\mathbf{X},\boldsymbol{\varepsilon}}\left\{\left|W_{n,n,v}(t) - E_{\boldsymbol{\varepsilon}}\left(W_{n,n,v}(t)\right)\right| \geq \delta t\right\} \leq 4\exp\left(-\frac{n\delta^2}{2}\right). \quad (47)$$

Its proof is given in Section 8.6.4, page 36.

Lemma 8.9. *Conditionnaly on the space $\Omega_{v,t}$ defined by (46), we have the two following inequalities:*

$$P_{\mathbf{X},\boldsymbol{\varepsilon}}\left\{\left|W_{n,2,v}(t)-E_{\boldsymbol{\varepsilon}}\big(W_{n,2,v}(t)\big)\right|\geq \delta t\right\} \leq 4\exp\left(-\frac{n\delta^2}{8}\right) \tag{48}$$

$$P_{\mathbf{X}}\left\{E_{\boldsymbol{\varepsilon}}W_{n,2,v}(t)-E_{\mathbf{X},\boldsymbol{\varepsilon}}\big(W_{n,2,v}(t)\big)\geq x\right\} \leq \exp\left(-\frac{nx^2}{Q_{n,v}(t)}\right). \tag{49}$$

Its proof is given in Section 8.6.5, page 36.

Lemma 8.10. *Let $\lambda_{n,v}$ be defined at Equation (13), Δ at Equation (12) and $\kappa = 10+4\Delta$. Conditionnaly on the space $\Omega_{v,\lambda_{n,v}}$ defined at Equation (46), for some positive constants c_1, c_2, with probability greater than $1-c_1\exp(-c_2 n\lambda_{n,v}^2)$, we have*

$$W_{n,n,v}(\lambda_{n,v})\leq \kappa\lambda_{n,v}^2 \quad \textit{and} \quad E_{\boldsymbol{\varepsilon}}W_{n,n,v}(\lambda_{n,v})\leq \kappa\lambda_{n,v}^2. \tag{50}$$

Its proof is given in Section 8.6.6, page 37.

8.5. **Proofs of Lemma 8.1 to 8.4:**

8.5.1. *Proof of Lemma 8.1 (page 18)*

For $f\in\mathcal{F}$ and $v\in\mathcal{P}$, let $g_v=\widehat{f}_v-f_v$. Note that $\|g_v\|_{\mathcal{H}_v}\leq 2$. Let us show that

$$|V_{n,\varepsilon}(g_v)|\leq \kappa\left[\lambda_{n,v}^2\|g_v\|_{\mathcal{H}_v}+\lambda_{n,v}\|g_v\|_n\right]. \tag{51}$$

We start by writing that

$$|V_{n,\varepsilon}(g_v)| = \|g_v\|_{\mathcal{H}_v}\left|V_{n,\varepsilon}\left(\frac{g_v}{\|g_v\|_{\mathcal{H}_v}}\right)\right| \leq \|g_v\|_{\mathcal{H}_v}W_{n,n,v}\Big(\frac{\|g_v\|_n}{\|g_v\|_{\mathcal{H}_v}}\Big). \tag{52}$$

Consider the two following cases:

Case A: $\|g_v\|_n\leq\lambda_{n,v}\|g_v\|_{\mathcal{H}_v}$
Case B: $\|g_v\|_n>\lambda_{n,v}\|g_v\|_{\mathcal{H}_v}$

Case A: Since $\|g_v\|_n\leq\lambda_{n,v}\|g_v\|_{\mathcal{H}_v}$, we have

$$W_{n,n,v}\Big(\frac{\|g_v\|_n}{\|g_v\|_{\mathcal{H}_v}}\Big)\leq W_{n,n,v}(\lambda_{n,v}).$$

We then apply Lemma 8.10, page 26, and conclude that (51) holds in Case A for each $v\in\mathcal{P}$ since, with high probability

$$|V_{n,\varepsilon}(g_v)| \leq \kappa\lambda_{n,v}^2\|g_v\|_{\mathcal{H}_v}\leq\kappa\lambda_{n,v}^2\|g_v\|_{\mathcal{H}_v}+\kappa\lambda_{n,v}\|g_v\|_n. \tag{53}$$

Case B: Consider now the case $\|g_v\|_n>\lambda_{n,v}\|g_v\|_{\mathcal{H}_v}$ and let us show that for any $v\in\mathcal{P}$,

$$W_{n,n,v}\left(\frac{\|g_v\|_n}{\|g_v\|_{\mathcal{H}_v}}\right)\leq\kappa\lambda_{n,v}\|g_v\|_n.$$

Let r_v be a deterministic number such that $r_v>\lambda_{n,v}$. Our first step relies on the study of the process $W_{n,n,v}(r_v)$, for $r_v>\lambda_{n,v}$. In that case we state two results:

R1 For any deterministic $r_v \geq \lambda_{n,v}$, with probability greater than $1 - c_1 \exp(-c_2 n \lambda_{n,v}^2)$,

$$W_{n,n,v}(r_v) \leq \kappa r_v \lambda_{n,v}. \tag{54}$$

R2 Inequality (54) continues to hold for random r_v of the form

$$r_v = \frac{\|g_v\|_n}{\|g_v\|_{\mathcal{H}_v}}.$$

Combining these two points implies that, with probability greater than $1 - c_1 \exp(-c_2 n \lambda_{n,v}^2)$,

$$\|g_v\|_{\mathcal{H}_v} W_{n,n,v}\Big(\frac{\|g_v\|_n}{\|g_v\|_{\mathcal{H}_v}}\Big) \leq \kappa \|g_v\|_n \lambda_{n,v}.$$

Consequently, in Case B, according to (52), for each v, Inequality (51) holds because

$$|V_{n,\varepsilon}(g_v)| \leq \kappa \|g_v\|_n \lambda_{n,v} \leq \kappa \lambda_{n,v}^2 \|g_v\|_{\mathcal{H}_v} + \kappa \lambda_{n,v} \|g_v\|_n.$$

This ends up the proof of Lemma 8.1.

Proof of **R1**. Taking $t = r_v$ and $\delta = \lambda_{n,v}$ in (47), with probability greater than $1 - 4\exp(-n\lambda_{n,v}^2)$, we have

$$W_{n,n,v}(r_v) \leq E_{\boldsymbol{\varepsilon}}[W_{n,n,v}(r_v)] + r_v \lambda_{n,v}.$$

Next we prove that for some positive r_v, with probability greater than $1 - \exp(-nc\lambda_{n,v}^2)$, we have

$$E_{\boldsymbol{\varepsilon}} W_{n,n,v}(r_v) \leq \kappa r_v \lambda_{n,v}. \tag{55}$$

Let $\widehat{\nu}_{n,v}$ defined as the smallest solution of $E_{\boldsymbol{\varepsilon}}[W_{n,n,v}(t)] \leq \kappa t^2$. For $W_{n,n,v}$, defined by (43), we write

$$\begin{aligned} E_{\boldsymbol{\varepsilon}} W_{n,n,v}(r_v) &= \frac{r_v}{\widehat{\nu}_{n,v}} E_{\boldsymbol{\varepsilon}} \sup\{|V_{n,\varepsilon}(g_v)|,\ \|g_v\|_{\mathcal{H}_v} \leq \widehat{\nu}_{n,v}/r_v,\ \|g_v\|_n \leq \widehat{\nu}_{n,v}\} \\ &\leq \frac{r_v}{\widehat{\nu}_{n,v}} E_{\boldsymbol{\varepsilon}} W_{n,n,v}(\widehat{\nu}_{n,v}) \leq \frac{r_v}{\widehat{\nu}_{n,v}} \kappa \widehat{\nu}_{n,v}^2 = \kappa r_v \widehat{\nu}_{n,v}. \end{aligned}$$

Besides, Lemma 8.10 stated that on the event $\Omega_{v,\lambda_{n,v}}$, $E_{\boldsymbol{\varepsilon}} W_{n,n,v}(\lambda_{n,v}) \leq \kappa \lambda_{n,v}^2$. It follows from the definition of $\widehat{\nu}_{n,v}$, and Lemma 8.6, that $\widehat{\nu}_{n,v} \leq \lambda_{n,v}$ for all $v \in \mathcal{P}$ with probability greater than $1 - \exp(-nc_2 \sum_{v\in\mathcal{P}} \lambda_{n,v}^2)$. Consequently, for any deterministic r_v such that $r_v \geq \lambda_{n,v}$, (54) is satisfied with high probability.

Proof of **R2**. Let us prove **R2** by using a peeling-type argument. Our aim is to prove that (54) holds for any r_v of the form

$$r_v = \frac{\|g_v\|_n}{\|g_v\|_{\mathcal{H}_v}}.$$

Since $\|g_v\|_\infty / \|g_v\|_{\mathcal{H}_v} \leq 1$, we have $\|g_v\|_n / \|g_v\|_{\mathcal{H}_v} \leq 1$. We thus restrict ourselves to r_v satisfying $r_v = \|g_v\|_n / \|g_v\|_{\mathcal{H}_v}$ with $\|g_v\|_n / \|g_v\|_{\mathcal{H}_v} \in (\lambda_{n,v}, 1]$.

We start by splitting the interval $(\lambda_{n,v}, 1]$ into M disjoint intervals such that $(\lambda_{n,v}, 1] = \cup_{k=1}^M (2^{k-1}\lambda_{n,v}, 2^k \lambda_{n,v}]$, for some M that will be chosen later. Consider the event $\mathcal{D}^c$ defined as follows:

$$\mathcal{D}^c = \{\exists v \in \mathcal{P} \text{ and } \exists \overline{g}_v, \text{ such that } |V_{n,\varepsilon}(\overline{g}_v)| \geq \kappa \lambda_{n,v} \|\overline{g}_v\|_n, \text{ with } \frac{\|\overline{g}_v\|_n}{\|\overline{g}_v\|_{\mathcal{H}}} \in (\lambda_{n,v}, 1]\}.$$

We prove that, for some positive constants c_1, c_2,

$$P[\mathcal{D}^c] \leq c_1 \exp(-c_2 n \lambda_{n,v}^2).$$

For $\overline{g}_v \in \mathcal{D}^c$, let $\overline{k}$ be the integer in $\{1, \cdots, M\}$, such that

$$2^{\overline{k}-1}\lambda_{n,v} \leq \frac{\|\overline{g}_v\|_n}{\|\overline{g}_v\|_{\mathcal{H}_\sqsubseteq}} \leq 2^{\overline{k}}\lambda_{n,v}.$$

This $\overline{k}$ satisfies

$$\|\overline{g}_v\|_{\mathcal{H}_v} W_{n,n,v}\Big(2^{\overline{k}}\lambda_{n,v}\Big) \quad \geq \quad \|\overline{g}_v\|_{\mathcal{H}_v} W_{n,n,v}\Big(\frac{\|g_v\|_n}{\|g_v\|_{\mathcal{H}_v}}\Big) \geq |V_{n,\varepsilon}(\overline{g}_v)| \geq \kappa\lambda_{n,v}\|\overline{g}_v\|_n.$$

Therefore, we get

$$W_{n,n,v}(2^{\overline{k}}\lambda_{n,v}) \geq \kappa\lambda_{n,v}\frac{\|\overline{g}_v\|_n}{\|\overline{g}_v\|_{\mathcal{H}_v}} \geq \kappa\lambda_{n,v}^2 2^{\overline{k}-1} \geq \kappa\frac{\lambda_{n,v}}{2}2^{\overline{k}}\lambda_{n,v}.$$

By taking $r_v = 2^{\overline{k}}\lambda_{n,v}$ in (54), we have

$$\mathcal{P}\left[W_{n,n,v}(2^{\overline{k}}\lambda_{n,v}) \geq \kappa\frac{\lambda_{n,v}}{2}2^{\overline{k}}\lambda_{n,v}\right] \leq c_1\exp(-c_2 n\lambda_{n,v}^2).$$

Now let us write $\mathcal{D}^c$ as follows:

$$\mathcal{D}^c = \bigcup_{k=1,\cdots,M}\left\{\exists v, \exists\, \overline{g}_v \text{ such that} |V_{n,\varepsilon}(\overline{g}_v)| \geq \kappa\lambda_{n,v}\frac{\|\overline{g}_v\|_n}{\|\overline{g}_v\|_{\mathcal{H}_v}} \text{ with } \frac{\|\overline{g}_v\|_n}{\|\overline{g}_v\|_{\mathcal{H}}} \in (2^{k-1}\lambda_{n,v}, 2^k\lambda_{n,v})\right\}.$$

The set $\mathcal{D}^c$ has probability smaller than $c_1 M\exp(-c_2 n\lambda_{n,v}^2)$. If we choose M such that $\log M \leq (c_2/2)n\lambda_{n,v}^2$, then the probability of the set $\mathcal{T}$ is greater than

$$1 - \sum_{v\in\mathcal{P}} c_1\exp -\frac{c_2}{2}n\lambda_{n,v}^2.$$

It follows that **R2** is proved which ends up the proof of Lemma 8.1. □

8.5.2. *Proof of Lemma 8.2 (page 19).*

Starting from (27) with B defined by Equation (26), we write

$$\begin{aligned}\frac{1}{2}\|\widehat{f} - m\|_n^2 \quad &\leq \quad 2|V_{n,\varepsilon}(\widehat{f} - fv)| + \\ &\sum_{v\in S_f}[\mu_v\|\widehat{f}_v - f_v\|_{\mathcal{H}_v} + \gamma_v\|\widehat{f}_v - f_v\|_n] - \sum_{v\in S^c}[\mu_v\|\widehat{f}_v\|_{\mathcal{H}_v} + \gamma_v\|\widehat{f}_v\|_n].\end{aligned}$$

On the event $\mathcal{T}$ defined in (29) we have

$$\begin{aligned}\frac{1}{2}\|\widehat{f} - m\|_n^2 \quad &\leq \quad 2\kappa\sum_{v\in\mathcal{P}}\lambda_{n,v}^2\|\widehat{f}_v - f_v\|_{\mathcal{H}_v} + 2\kappa\sum_{j\in\mathcal{P}}\lambda_{n,v}\|\widehat{f}_v - f_v\|_n + \\ &\sum_{v\in S_f}[\mu_v\|\widehat{f}_v - f_v\|_{\mathcal{H}_v} + \gamma_v\|\widehat{f}_v - f_v\|_n] - \sum_{v\in S^c}[\mu_v\|\widehat{f}_v\|_{\mathcal{H}_v} + \gamma_v\|\widehat{f}_v\|_n].\end{aligned}$$

Rearranging the terms we obtain that

$$\begin{aligned}\frac{1}{2}\|\widehat{f}-m\|_n^2 &\leq \sum_{v\in S_f}(2\kappa\lambda_{n,v}^2+\mu_v)\|\widehat{f}_v-f_v\|_{\mathcal{H}_v}+\sum_{v\in S_f}(2\kappa\lambda_{n,v}+\gamma_v)\|\widehat{f}_v-f_v\|_n+\\ &\quad\sum_{v\in S_f^c}(2\kappa\lambda_{n,v}^2-\mu_v)\|\widehat{f}_v\|_{\mathcal{H}_v}+\sum_{v\in S_f^c}(2\kappa\lambda_{n,v}-\gamma_v)\|\widehat{f}_v\|_n.\end{aligned}$$

Now, thanks to Assumption (14) with $C_1 \geq \kappa$ we have $\kappa\lambda_{n,v}^2 \leq \mu_v$ and $2\kappa\lambda_{n,v} \leq \gamma_v$ and Lemma 8.2 is shown since

$$\begin{aligned}0 \leq \frac{1}{2}\|\widehat{f}-m\|_n^2 &\leq 3\sum_{v\in S_f}\mu_v\|\widehat{f}_v-f_v\|_{\mathcal{H}_v}+3\sum_{v\in S_f}\|\widehat{f}_v-f_v\|_n-\\ &\quad\sum_{v\in S_f^c}\mu_v\|\widehat{f}_v\|_{\mathcal{H}_v}-\sum_{v\in S_f^c}\gamma_v\|\widehat{f}_v\|_n.\end{aligned}$$

□

8.5.3. *Proof of Lemma 8.3 (page 19):*

Let us consider the following two cases:

- $\|\widehat{f}_v - f_v\|_2 \leq \gamma_v$. We apply Lemma 8.6 (page 25) to the function $g_v = \widehat{f}_v - f_v$. It satisfies $g_v \in \mathcal{G}(\gamma_v)$ with $b = 2$ (recall that $\|\cdot\|_\infty \leq \|\cdot\|_{\mathcal{H}_v}$). Moreover, $\gamma_v \geq C_1\lambda_{n,v} \geq C_1\nu_{nv} \geq \nu_{n,v}$ as soon as $C1 \geq 1$.
  It follows that, for some positive c_2, with probability greater than $1-\exp(-nc_2\gamma_v^2)$

$$\|\widehat{f}_v - f_v\|_n \leq \|\widehat{f}_v - f_v\|_2 + \gamma_v$$

- $\|\widehat{f}_v - f_v\|_2 \geq \gamma_v$. We apply Lemma 8.7 (page 25) to the function $g_v = \widehat{f}_v - f_v$ with $b = 2$. It follows that, for some positive c_2, with probability greater than $1-\exp(-nc_2\gamma_v^2)$,

$$\|\widehat{f}_v - f_v\|_n \leq 2\|\widehat{f}_v - f_v\|_2.$$

□

8.5.4. *Proof of Lemma 8.4 (page 21):*

Let $d(f)$ be defined by (36), and let $\mathcal{G}(f)$ and $\mathcal{G}'(f)$ be the following sets

$$\begin{aligned}\mathcal{G}(f) &= \left\{ g=\sum_{v\in\mathcal{P}} g_v, \text{ satisfying } \|g_v\|_{\mathcal{H}_v}\leq 2, \text{ and Conditions } \mathbf{C1, C2, C3} \right\},\\ \mathcal{G}'(f) &= \{g\in\mathcal{G}(f), \text{ such that } \|g\|_2 = d(f)\}.\end{aligned}$$

Let us consider the two events $\mathcal{B}$ and $\mathcal{B}'$ defined as follows:

$$\mathcal{B}' = \{\forall h\in\mathcal{G}',\ \|h\|_n^2 \geq d(f)^2/2\} \text{ and } \mathcal{B} = \{\forall h\in\mathcal{G},\ \|h\|_n^2 \geq \|h\|_2^2/2, \text{ and } \|h\|_2 \geq d(f)\}.$$

Let us first remark that $\mathcal{B}'$ is included into $\mathcal{B}$: if $h \in \mathcal{B}'$, then $h \in \mathcal{G}$, $\|h\|_2 = d(f)$ and $\|h\|_n^2 \geq d^(f)2/2$. It follows that $\|h\|_n^2 \geq \|h\|_2^2/2$ and $\|h\|_2 \geq d(f)$. Therefore Lemma 8.4 is proved if $\mathcal{B}'$ holds with high probability. Consider

$$Z_n(\mathcal{G}') = \sup_{g\in\mathcal{G}'}\{d(f)^2 - \|g\|_n^2\}.$$

We show that the event $Z_n(\mathcal{G}') \leq d(f)^2/2$ has probability greater than $1 - c_1 \exp(-nc_2 d(f)^2)$.

Let us briefly recall the notion of covering numbers for a totally bounded metric space $(\mathbb{G}, \rho)$, consisting of a set $\mathbb{G}$ and a metric ρ defined from $\mathbb{G} \times \mathbb{G}$ into $\mathbb{R}_+$. A δ-covering set of $\mathbb{G}$ is a collection of functions $f^1, \cdots, f^N$ such that for all $f \in \mathbb{G}$ there exists $k \in \{1, 2, \cdots, N\}$ such that $\rho(f, f^k) \leq \delta$.

The δ-covering number $N(\delta, \mathbb{G}, \rho)$ is the cardinality of the smallest δ-covering set. A propercovering restricts the covering to use only elements in the set $\mathbb{G}$. The proper covering number denoted $N_{\rm pr}(\delta, \mathbb{G}, \rho)$ satisfies

$$N(\delta, \mathbb{G}, \rho) \leq N_{\rm pr}(\delta, \mathbb{G}, \rho) \leq N(\delta/2, \mathbb{G}, \rho). \tag{56}$$

Let us now consider a $d(f)/8$-covering of $(\mathcal{G}', \|\cdot\|_n)$, so that, for all g in $\mathcal{G}'$ there exists g^k such that $\|g - g^k\|_n \leq d(f)/8$. The associated proper covering number is

$$N_{\rm pr} = N_{\rm pr}(d(f)/8, \mathcal{G}', \|\cdot\|_n). \tag{57}$$

Now, for all $g \in \mathcal{G}'$, $T_1 = \|g^k\|_n^2 - \|g\|_n^2$, and $T_2 = d^2(f) - \|g^k\|_n^2$, we write

$$d(f)^2 - \|g\|_n^2 = T_1 + T_2.$$

The proof is splitted into four steps:

Step 1 The first step consists in showing that

$$T_1 = \|g^k\|_n^2 - \|g\|_n^2 \leq \frac{d(f)^2}{4}. \tag{58}$$

Step 2 The second step consists in proving that, for $N_{\rm pr}$ given at Equation (57) and for some constant C,

$$P_{\mathbf{X}}\left[\max_{k \in \{1, \cdots, N_{\rm pr}\}} [d(f)^2 - \|g^k\|_n^2] \geq d^2/4\right] \leq \exp\Big(\log N_{\rm pr} - Cnd(f)^2\Big).$$

Step 3 The third step concerns the control of $N_{\rm pr}$: we show the following result

$$\log N_{\rm pr} \leq n\left(\frac{64}{d(f)} \mathbb{E}_\varepsilon \sup_{g \in \mathcal{G}'} |V_{n,\varepsilon}(g)|\right)^2.$$

Step 4 The last step consists in bounding from above the Gaussian complexity:

$$\mathbb{E}_\varepsilon \sup_{g \in \mathcal{G}'} |V_{n,\varepsilon}(g)| \leq \frac{20\kappa}{C_1} d(f)^2.$$

Let us conclude the proof of the lemma before proving these four steps.

Putting together Steps 3 and 4, for $c_3 < C$ and C_1 large enough, then Step 2 states that

$$P_{\mathbf{X}}\left(T_2 \geq \frac{d(f)^2}{4}\right) \leq P_{\mathbf{X}}\left[\max_{k \in \{1, \cdots, N_{\rm pr}\}} [d(f)^2 - \|g^k\|_n^2] \geq d(f)^2/4\right] \leq \exp\Big(-c_3 nd(f)^2\Big).$$

Now, we have

$$P_{\mathbf{X}}\left[Z_n(\mathcal{G}') \leq d(f)^2/2\right] = P_{\mathbf{X}}\left[\max_{g^1, \cdots, g^N} \{d(f)^2 - \|g^k\|_n^2\} \geq \frac{d(f)^2}{4}\right] \leq \exp\Big(-c_3 nd(f)^2\Big).$$

We conclude the proof of the lemma, by noting that $d(f)^2 \geq C_1^2 \sum_v \lambda_{n,v}^2$ (see (13), (36) and (14)).

Proof of Step 1: We start by writing that

$$\begin{aligned}\|g^k\|_n^2 - \|g\|_n^2 &= \frac{1}{n}\sum_{i=1}^{n}\left[(g^k(\mathbf{X}_i))^2 - (g(\mathbf{X}_i))^2\right] \\ &\leq \|g^k - g\|_n \sqrt{\frac{1}{n}\sum_{i=1}^{n}[g^k(\mathbf{X}_i) + g(\mathbf{X}_i)]^2}.\end{aligned}$$

Using that $(a+b)^2 \leq 2a^2 + 2b^2$, $g \in \mathcal{G}'$, and g satisfies Condition **C3**, we get

$$\frac{1}{n}\sum_{i=1}^{n}[g^k(X^{(i)}) + g(X^{(i)})]^2 \quad \leq 2\|g^k\|_n^2 + 2\|g\|_n^2 \leq 4d(f)^2.$$

Besides, the covering set is constructed such that $\|g^k - g\|_n \leq d(f)/8$. It follows that Step 1 is proved.

Proof of Step 2: We prove that for some constant C,

$$P_{\mathbf{X}}\left[T_2 \geq \frac{d^2}{4}\right] \leq P_{\mathbf{X}}\left[\max_{1\leq k\leq N_{\rm pr}}\{d(f)^2 - \|g^k\|_n^2\} \geq \frac{d(f)^2}{4}\right] \leq \exp\left(\log N_{\rm pr} - C\frac{nd(f)^2}{1 + d(f) + d(g)^2}\right).$$

As $g^k \in \mathcal{G}'$, $d = \|g^k\|_2$. Then

$$\max_{1\leq k\leq N_{\rm pr}}\left\{d(f)^2 - \|g^k\|_n^2\right\} = \max_{1\leq k\leq N_{\rm pr}}[\|g^k\|_2^2 - \|g^k\|_n^2].$$

Applying Theorem 3.5 in Chung and Lu [10] we have that for all positive λ

$$P_{\mathbf{X}}\left[\sum_{i=1}^{n}(g^k(\mathbf{X}_i))^2 \leq n\mathbb{E}(g^k(\mathbf{X}_i))^2 - \lambda\right] \leq \exp\left(-\frac{\lambda^2}{2n\mathbb{E}(g^k(\mathbf{X}))^4}\right).$$

Taking $\lambda = nd(f)^2/4$ and using that $\|g^k\|_2^2 = d(f)^2$ we get

$$P_{\mathbf{X}}\left[\{d^2 - \|g^k\|_n^2\} \geq \frac{d(f)^2}{4}\right] \leq \exp\Big(-\frac{nd(f)^4}{32\mathbb{E}(g^k(\mathbf{X}))^4}\Big).$$

It follows that

$$P_{\mathbf{X}}\left[\max_{1\leq k\leq N_{\rm pr}} d(f)^2 - \|g^k\|_n^2 \geq \frac{d(f)^2}{4}\right] \leq \sum_{k=1}^{N_{\rm pr}}\exp\Big(-\frac{nd(f)^4}{32\mathbb{E}(g^k(\mathbf{X}))^4}\Big) \leq \exp\Big(\log N_{\rm pr} - \frac{nd(f)^4}{32\max_k \mathbb{E}(g^k(\mathbf{X}))^4}\Big). \quad (59)$$

It remains to calculate $E_{\mathbf{X}}g^4(\mathbf{X})$ for $g \in \mathcal{G}'$. Precisely we show the following result:

$$\mathbb{E}g^4(\mathbf{X}) \leq 2d(f)^2\left(2 + 11d(f)^2 + 4d\right).$$

This result comes from the property of the RKHS $\mathcal{H}$: indeed $g \in \mathcal{H}$ is written $g = \sum_{v\in\mathcal{P}} g_v$ where the functions g_v are centered and orthogonal in $\mathbb{L}^2(P_{\mathbf{X}})$. Therefore $\mathbb{E}g^4(\mathbf{X})$ is the sum of the following terms:

$$\begin{aligned}
A_1 &= \sum_{v\in\mathcal{P}} E_{\mathbf{X}} g_v^4(\mathbf{X}_v), \qquad A_2 = \binom{4}{2} \sum_{v\neq v'} E_{\mathbf{X}} g_v^2(\mathbf{X}_v) g_{v'}^2(\mathbf{X}_{v'}),\\
A_3 &= \binom{4}{3} \sum_{v_1\neq v_2\neq v_3} E_{\mathbf{X}} g_{v_1}^2(\mathbf{X}_{v_1}) g_{v_2}(\mathbf{X}_{v_2}) g_{v_3}(\mathbf{X}_{v_3}), \quad A_4 = \binom{4}{3} \sum_{v_1\neq v_2} E_{\mathbf{X}} g_{v_1}^3(\mathbf{X}_{v_1}) g_{v_2}(\mathbf{X}_{v_2})\\
A_5 &= \binom{4}{1} \sum_{v_1\neq v_2\neq v_3\neq v_4} E_{\mathbf{X}} g_{v_1}(\mathbf{X}_{v_1}) g_{v_2}(\mathbf{X}_{v_2}) g_{v_3}(\mathbf{X}_{v_3}) g_{v_4}(\mathbf{X}_{v_4}).
\end{aligned}$$

Using the Cauchy-Schwartz inequality and the fact that $\|g_v\|_\infty \leq \|g_v\|_{\mathcal{H}_v} \leq 2$, and $\|g\|_2 = d(f)$ (because $g \in \mathcal{G}'$), we get that A_1 is proportionnal to $d(f)^2$, A_2, A_3, A_5 to $d(f)^4$, and A_4 to $d(f)^3$. For example

$$A_1 = \sum_{v\in\mathcal{P}} E_{\mathbf{X}} g_v^4(X_v) \leq \|g\|_\infty^2 \sum_{v\in\mathcal{P}} \|g_v\|_2^2 = \|g\|_\infty^2 \| \sum_{v\in\mathcal{P}} g_v\|_2^2 \leq 4d(f)^2.$$

After calculation of the terms A_i, since $d(f)^2$ is assumed to be smaller than one, we get that

$$\max_k E_{\mathbf{X}}(g^k(\mathbf{X}))^4 \leq 4d(f)^2(1 + d(f) + (11/2)d(f)^2) \leq 34d(f)^2. \tag{60}$$

Step 2 is proved by combining (59) and (60).

Proof of Step 3: Let N_{pr} be defined at Equation (57). We prove that

$$\sqrt{\frac{\log N_{\mathrm{pr}}}{n}} \leq \frac{64}{d(f)} E_{\boldsymbol{\varepsilon}} \sup_{g\in\mathcal{G}'} |V_{n,\varepsilon}(g)|.$$

We start from (56) and write that

$$\log N_{\mathrm{pr}}(d(f)/8, \mathcal{G}', \|\cdot\|_n) \leq \log N(d(f)/16, \mathcal{G}', \|\cdot\|_n).$$

Using the Sudakov minoration (see Pisier [29]) we have that for all positive ω

$$\sqrt{\log N(\omega, \mathcal{G}', \|\cdot\|_n)} \leq \frac{4\sqrt{n}}{\omega} E_{\boldsymbol{\varepsilon}} \left[\sup_{g\in\mathcal{G}'} |V_{n,\varepsilon}(g)| \right].$$

Hence by taking $\omega = d(f)/16$, Step 3 is proved.

Proof of Step 4: The last step consists in bounding from above the Gaussian complexity $\mathbb{E}_\varepsilon \sup_{g\in\mathcal{G}'} |V_{n,\varepsilon}(g)|$. This control is performed by using Lemma 8.5 (page 25). According to Inequality (51),

$$|V_{n,\varepsilon}(g)| \leq \kappa \left[\sum_{v\in\mathcal{P}} \lambda_{n,v}^2 \|g_v\|_{\mathcal{H}_v} + \sum_{v\in\mathcal{P}} \lambda_{n,v} \|g_v\|_n \right],$$

with $\lambda_{n,v}$ defined by Equation (13) satisfying $C_1\lambda_{n,v} \leq \gamma_v$ and $C_1\lambda_{n,v}^2 \leq \mu_v$ for all $v \in \mathcal{P}$. It follows

$$\begin{aligned}\sup_{g\in\mathcal{G}'}\sum_{v\in\mathcal{P}}|V_{n,\varepsilon}(g_v)| &\leq \kappa\sup_{g\in\mathcal{G}'}\left[\sum_{v\in\mathcal{P}}\lambda_{n,v}^2\|g_v\|_{\mathcal{H}_v}+\sum_{v\in\mathcal{P}}\lambda_{n,v}\|g_v\|_n\right]\\ &\leq \frac{\kappa}{C_1}\sup_{g\in\mathcal{G}'}\left[\sum_{v\in\mathcal{P}}\mu_v\|g_v\|_{\mathcal{H}_v}+\sum_{v\in\mathcal{P}}\gamma_v\|g_v\|_n\right]\end{aligned}$$

and according to Condition **C1**,

$$\begin{aligned}\sup_{g\in\mathcal{G}'}\sum_{v\in\mathcal{P}}|V_{n,\varepsilon}(g_v)| &\leq \frac{4\kappa}{C_1}\left(\sup_{g\in\mathcal{G}'}\sum_{v\in\mathcal{S}}\mu_v\|g_v\|_{\mathcal{H}_v}+\sup_{g\in\mathcal{G}'}\sum_{v\in\mathcal{S}}\gamma_v\|g_v\|_n\right),\\ &\leq \frac{4\kappa}{C_1}\left(2\sum_{v\in\mathcal{S}}\mu_v+\sup_{g\in\mathcal{G}'}\sum_{v\in\mathcal{S}}\gamma_v\|g_v\|_n\right),\end{aligned}$$

because $\|g_v\|_{\mathcal{H}_v} \leq 2$. Now, according to Condition **C2**, and using that $2ab \leq a^2+b^2$, we get

$$\begin{aligned}\sup_{g\in\mathcal{G}'}\sum_{v\in\mathcal{P}}|V_{n,\varepsilon}(g_v)| &\leq \frac{4\kappa}{C_1}\left[2\sum_{v\in S}\mu_v+2\sum_{v\in S}\gamma_v^2+\sup_{g\in\mathcal{G}'}\sum_{v\in S}\|g_v\|_2^2\right]\\ &\leq \frac{4\kappa}{C_1}\left[2\sum_{v\in S}(\mu_v+\gamma_v^2)+d(f)^2\right],\end{aligned}$$

the last inequality coming from the fact for all $g \in \mathcal{G}'$, $\|g\|_2^2 = d(f)^2 \geq \sum_{v\in S}\|g_v\|_2^2$.

Finally, thanks to (36), we get

$$\sup_{g\in\mathcal{G}'}\sum_{v\in\mathcal{P}}|V_{n,\varepsilon}(g_v)| \leq \frac{20\kappa}{C_1}d(f)^2.$$

□

8.6. **Proofs of intermediate Lemmas.**

8.6.1. *Proof of Lemma 8.5 (page 25):*

Let us write that the kernel k_v is written as :

$$k_v(\mathbf{x}_v,\mathbf{y}_v) = \sum_k \omega_{v,k}\phi_{v,k}(\mathbf{x}_v)\phi_{v,k}(\mathbf{y}_v)$$

where $(\phi_{v,k})_{k=1}^{\infty}$ is an orthonormal basis of $\mathbb{L}^2(P_v)$, where $P_v = \prod_{a\in v}P_a$.

Let us consider the class of functions $\mathcal{K}(t)$ defined as

$$\mathcal{K}(t) = \{g_v \in \mathcal{H}_v, \|g_v\|_{\mathcal{H}_v} \leq 2, \|g_v\|_2 \leq t\}.$$

It comes that

$$g_v = \sum_i a_i\phi_{v,i}, \quad \text{with } \|g_v\|_{\mathcal{H}_v}^2 = \sum_i \frac{a_i^2}{\omega_{v,i}} \leq 4, \text{ and } \|g_v\|_2^2 = \sum_i a_i^2 \leq t^2$$

In the following, we set $\mu_{k,v}(t) = \min\left\{t^2, \omega_{k,v}\right\}$. Hence

$$\sum_k \frac{a_k^2}{\mu_{k,v}(t)} \leq \frac{1}{t^2}\sum_k a_k^2 + \sum_k \frac{a_k^2}{\omega_{k,v}} = \frac{1}{t^2}\|g_v\|_2^2 + \|g_v\|_{\mathcal{H}_v}^2 \leq 5, \tag{61}$$

as soon as $g_v \in \mathcal{K}(t)$.

Now, let us prove the lemma:

$$\begin{aligned} E_{\mathbf{X},\boldsymbol{\varepsilon}} W_{n,2,v}(t) &= E_{\mathbf{X},\boldsymbol{\varepsilon}} \sup_{g\in\mathcal{K}(t)} \left| \frac{1}{n}\sum_{i=1}^n \varepsilon_i \sum_\ell a_\ell \phi_{v,\ell}(x_{v,i}) \right| \\ &= E_{\mathbf{X},\boldsymbol{\varepsilon}} \sup_{g\in\mathcal{K}(t)} \left| \frac{1}{n}\sum_\ell \frac{a_\ell}{\sqrt{\mu_{v,\ell}(t)}} \sum_{i=1}^n \varepsilon_i \sqrt{\mu_{v,\ell}(t)}\phi_{v,\ell}(x_{v,i}) \right| \\ &\leq \sqrt{5}\sqrt{E_{\mathbf{X},\boldsymbol{\varepsilon}} \sum_\ell \left(\frac{1}{n}\sum_{i=1}^n \varepsilon_i \sqrt{\mu_{v,\ell}(t)}\phi_{v,l}(x_{v,i}) \right)^2}. \end{aligned}$$

The last inequality follows from the Cauchy-Schwartz inequality and Inequality (61). Now, simple calculation leads to

$$E_{\mathbf{X},\boldsymbol{\varepsilon}} W_{n,2,v}(t) \leq \sqrt{5}\sqrt{\frac{1}{n}\sum_\ell \mu_{v,\ell}(t)},$$

□

8.6.2. *Proof of Lemma 8.6 (page 25):*

Using that $\left|\sqrt{a}-\sqrt{b}\right| \leq \sqrt{|a-b|}$, we get

$$|\|g_v\|_2 - \|g_v\|_n| \leq \sqrt{|\|g_v\|_2^2 - \|g_v\|_n^2|}.$$

Hence

$$\left\{ \|g_v\|_\infty \leq b,\ |\|g_v\|_2 - \|g_v\|_n| \geq \frac{bt}{2} \right\} \subset \left\{ \left|\|g_v\|_2^2 - \|g_v\|_n^2\right| \geq \frac{b^2t^2}{4} \right\}.$$

The centered process

$$\left|\|g_v\|_2^2 - \|g_v\|_n^2\right| = \left| \frac{1}{n}\sum_{i=1}^n g_v^2(\mathbf{X}_{v,i}) - \mathbb{E}(g_v^2(\mathbf{X}_v)) \right|,$$

satisfies a concentration inequality given, for example, by Theorem 2.1 in Bartlett *et al.* [3] : if $\mathcal{C}$ is a class of functions f such that $\|f\|_\infty \leq B$ and $Ef(\mathbf{X}) = 0$, and if there exists $\gamma > 0$ such that for every $f \in \mathcal{C}$, $\operatorname{Var} f(\mathbf{X}) \leq \gamma^2$. Then for every $x > 0$, with probability at least $1 - e^{-x}$,

$$\sup_{f\in\mathcal{C}} \frac{1}{n}\left|\sum_{j=1}^n f(\mathbf{X}_j)\right| \leq \inf_{\alpha>0} \left\{ 2(1+\alpha)E\left(\sup_{f\in\mathcal{C}} \frac{1}{n}\left|\sum_{j=1}^n f(\mathbf{X}_j)\right| \right) + \sqrt{\frac{2x}{n}}\gamma + B\left(\frac{1}{3}+\frac{1}{\alpha}\right)\frac{x}{n} \right\}. \tag{62}$$

For any $t > 0$, for $\mathcal{G}(t)$ defined by (45), let us consider the class of functions $\mathcal{C}(t)$ defined as follows

$$\mathcal{C}(t) = \left\{ f \text{ such that } f = g_v^2 - \mathbb{E}(g_v^2),\ \text{with } g_v \in \mathcal{G}(t) \right\}.$$

Note that if $f \in \mathcal{C}(t)$, $E_{\mathbf{X}} f(\mathbf{X}_v) = 0$ and $\|f\|_\infty \leq b^2$. We have to study

$$\gamma^2(t) = \sup_{g_v\in\mathcal{G}(t)} E_{\mathbf{X}}\left(g_v^2(\mathbf{X}) - \|g_v\|_2^2\right)^2 \text{ and } \Gamma(t) = E_{\mathbf{X}}\left(\sup_{g_v\in\mathcal{G}(t)} \left|\|g_v\|_n^2 - \|g_v\|_2^2\right| \right).$$

It is easy to see that

$$\gamma^2(t) \quad \leq \quad b^2 \sup_{g_v \in \mathcal{G}(t)} E_{\mathbf{X}} \left(g_v(\mathbf{X}) + \|g_v\|_2 \right)^2 \leq 4b^2 t^2.$$

Let ζ_i be independent and identically random variables Rademacher distributed and let $E_{\mathbf{X},\zeta}$ denotes the expectation with respect to the law of $(\mathbf{X}, \zeta)$. By a symmetrization argument,

$$\Gamma(t) \leq 2E_{\mathbf{X},\zeta} \sup_{g_v \in \mathcal{G}(t)} \left| \frac{1}{n} \sum_{i=1}^{n} \zeta_i g_v^2(\mathbf{X}_i) \right|.$$

Since $\|g_v\|_\infty \leq b$, applying the contraction principal (see Ledoux-Talagrand [23]) we get that, for $Q_{n,v}(t)$ defined by (11),

$$\begin{aligned} E_{\mathbf{X},\zeta} \sup_{g_v \in \mathcal{G}(t)} \left| \frac{1}{n} \sum_{i=1}^{n} \zeta_i g_v^2(\mathbf{X}_i) \right| &\leq 4bE_{\mathbf{X},\zeta} \sup_{g_v \in \mathcal{G}(t)} \left| \frac{1}{n} \sum_{i=1}^{n} \zeta_i g_v(\mathbf{X}_i) \right| \\ &\leq 4bQ_{n,v}(t). \end{aligned}$$

The last inequality was proved by Mendelson [28], Theorem 41 (see the proof of Lemma 8.5). Now, thanks to (62) we get that for all $x > 0$, with probability greater than $1 - e^{-x}$

$$\sup_{g_v \in \mathcal{G}(t)} \left| \|g_v\|_n^2 - \|g_v\|_2^2 \right| \leq \inf_{\alpha > 0} \left\{ 16(1+\alpha) bQ_{n,v}(t) + \sqrt{\frac{2x}{n}} 2bt + b^2 \left(\frac{1}{3} + \frac{1}{\alpha} \right) \frac{x}{n} \right\}.$$

Taking $x = c_2 n t^2$, $t \geq \nu_n$, we have that with probability greater than $1 - e^{-c_2 n t^2}$

$$\sup_{g_v \in \mathcal{G}(t)} \left| \|g_v\|_n^2 - \|g_v\|_2^2 \right| \leq \inf_{\alpha > 0} t^2 \left\{ 16(1+\alpha) b\Delta + \sqrt{2c_2} 4b + b^2 \left(\frac{1}{3} + \frac{1}{\alpha} \right) c_2 \right\}.$$

The infimum of the right hand side is reached in $\alpha = \sqrt{c_2 b / 16\Delta}$, and equals

$$\frac{b^2 c_2}{3} + 8\sqrt{\Delta c_2} b^{3/2} + 4(4\Delta + \sqrt{2c_2}) b.$$

The constants Δ and c_2 should satisfy that this infimum is strictly smaller than $b^2/4$. For example, if $16\Delta < b/8$, it remains to choose c_2 small enough such that

$$b \left(\frac{c_2}{3} + \frac{\sqrt{2c_2}}{2} \right) + 4\sqrt{2c_2} < \frac{b}{8}.$$

□

8.6.3. *Proof of Lemma 8.7 (page 25):*

Let $t > \nu_{n,v}$ and h be defined as $h = t g_v / \|g_v\|_2$. If g_v satisfies the assumptions of the lemma, then h satisfies $\|h\|_2 = t$, $\|h\|_{\mathcal{H}} \leq 2$ and $\|h\|_\infty \leq b$. Applying Lemma 8.6 (page 25) to the function h, we obtain that for all $t \geq \nu_{n,v}$, with probability greater than $1 - \exp(-c_2 n t^2)$, we have $|t - \|h\|_n| \leq bt/2$ for all $h \in \mathcal{G}(t)$. This concludes the proof of the lemma.

8.6.4. *Proof of Lemma 8.8 (page 25):*

The proof of Lemma 8.8 is based on an isoperimetric inequality for Gaussian processes (Borell [6] or Cirel'son *et al.* [11]) as it is stated in Theorem (3.8), page 61 in Massart [26]. Let us recall this inequality:

Lemma 8.11. *Let P be the Gaussian probability measure on $\mathbb{R}^n$ and let ϕ be a function from $\mathbb{R}^n$ to $\mathbb{R}$, and $\|\phi\|_L$ its Lipschitz semi-norm:*

$$\|\phi\|_L = \sup_{x \neq y} \frac{|\phi(x) - \phi(y)|}{\sqrt{n}\|x - y\|_n}.$$

Let $\overline{\Phi}$ be the cumulative distribution of the standard Gaussian distribution. Then for any u,

$$P(|f - E_P f| \geq u\ ) \leq 4\overline{\Phi}\left(\frac{u}{\|\phi\|_L}\right). \tag{63}$$

We apply Lemma 8.11 to $\phi(\varepsilon_1, \ldots, \varepsilon_n) = W_{n,n,v}(t)$. By Cauchy-Schwarz Inequality, $\|\phi\|_L = t/\sqrt{n}$. It follows that Lemma 8.8 is proved since

$$P_{\mathbf{X},\boldsymbol{\varepsilon}}\left(|W_{n,n,v}(t) - \mathbb{E}_\varepsilon W_{n,n,v}(t)| \geq \delta t\right) \leq 4\exp\left\{-\frac{(\delta t)^2}{2\left(\frac{t}{\sqrt{n}}\right)^2}\right\} \leq 4\exp\left(-\frac{n\delta^2}{2}\right).$$

□

8.6.5. *Proof of Lemma 8.9 (page 26):*

We start with the proof of (48) in Lemma 8.9 by applying once again Lemma 8.11 given above, to the function $\phi(\boldsymbol{\varepsilon}) = \phi(\varepsilon_1, \ldots, \varepsilon_n) = W_{n,2,v}(t)$. On the event $\Omega_{v,t}$ defined by (46), we have $\|g_v\|_n \leq bt/2 + \|g_v\|_2$. Besides if $\|g_v\|_{\mathcal{H}_v} \leq 2$, then $\|g_v\|_\infty \leq 2$. Therefore applying Lemma 8.6 with $b = 2$, we get that if $\|g_v\|_2 \leq t$,

$$|\phi(\boldsymbol{\varepsilon}) - \phi(\boldsymbol{\varepsilon}')| \leq \sup_{\|g_v\|_n \leq 2t} \|g_v\|_n \|\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}'\|_n \leq 2t\|\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}'\|_n,$$

leading to $\|\phi\|_L = 2t/\sqrt{n}$. It follows that (48) in Lemma 8.9 is proved since

$$P_{\mathbf{X},\boldsymbol{\varepsilon}}\left[\{|W_{n,2,v}(t) - \mathbb{E}_\varepsilon W_{n,2,v}(t)| \geq \delta t\} \cap \Omega_{v,t}^c\right] \leq 4\exp\left\{-\frac{(\delta t)^2}{2\left(\frac{2t}{\sqrt{n}}\right)^2}\right\} \leq 4\exp\left(-\frac{n\delta^2}{8}\right).$$

We now come to the proof of (49) in Lemma 8.9 using a Poissonian inequality for self-bounded processes (see Boucheron *et al.* [7]) and Theorem 5.6, p 158 in Massart [26]). Let us recall it in the particular case we are interested in:

Theorem 8.1. *Let $X_1, \cdots, X_n$ be n iid variables. For $i \in \{1, \cdots, n\}$ let $X_{(-i)} = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$. Let Z be a nonnegative and bounded measurable function of $X = (X_1, \cdots, X_n)$. Assume that for all $i \in \{1, \cdots, n\}$, there exists a measurable function Z_i of $X_{(-i)}$ such that $0 < Z - Z_i \leq 1$, and $\sum_{i=1}^n (Z - Z_i) \leq Z$. Then, for all $x > 0$, we have $P\{Z \geq E(Z) + x\} \leq \exp\left(-x^2/2E(Z)\right)$.*

We apply this result to Z defined as

$$Z = Z(\mathbf{X}_1, \cdots, \mathbf{X}_n) = nE_{\boldsymbol{\varepsilon}} W_{n,2,v}(t) = nE_{\boldsymbol{\varepsilon}} \sup\left\{|V_{n,\varepsilon}(g_v)|, \|g_v\|_2 \leq t, \|g_v\|_{\mathcal{H}_v} \leq 2\right\}.$$

The variable Z is positive, and because the distribution of $(\varepsilon_1, \ldots, \varepsilon_n)$ is symmetric, we have that

$$Z = E_{\boldsymbol{\varepsilon}} \sup\left\{nV_{n,\varepsilon}(g_v), \|g_v\|_2 \leq t, \|g_v\|_{\mathcal{H}_v} \leq 2\right\}.$$

Let τ be the function in $\mathcal{H}_v$ such that $Z = E_{\boldsymbol{\varepsilon}} n V_{n,\varepsilon}(\tau)$ (note that τ depends on $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$ and on $(\varepsilon_1, \ldots, \varepsilon_n)$), and let

$$Z_i = E_{\boldsymbol{\varepsilon}} \sup_{g_v} \sum_{j \neq i} \varepsilon_j g_v(\mathbf{X}_j).$$

We show that Z and Z_i satisfy the assumptions of Theorem 8.1:

$$\begin{aligned} Z - Z_i &= E_{\boldsymbol{\varepsilon}} \left(\varepsilon_i \tau(\mathbf{X}_i) + \sum_{j \neq i} \varepsilon_j \tau(\mathbf{X}_j) - \sup_{g_v} \sum_{j \neq i} \varepsilon_j g_v(\mathbf{X}_j) \right) \\ &\leq E_{\boldsymbol{\varepsilon}} \varepsilon_i \tau(\mathbf{X}_i) \leq \frac{1}{\sqrt{2\pi}} E_{\boldsymbol{\varepsilon}} \sup_{x \in \mathcal{X}} |\tau(\mathbf{X})| \leq \sqrt{\frac{2}{\pi}}, \end{aligned}$$

where the last inequality comes from the fact that $\sup_{x \in \mathcal{X}} |\tau(\mathbf{X})| \leq \|\tau\|_{\mathcal{H}_v} \leq 2$. Moreover $Z - Z_i \geq 0$ since

$$\begin{aligned} Z &= E_{\boldsymbol{\varepsilon}} \sup_{g_v} \sum_{j=1}^{n} \varepsilon_j g_v(\mathbf{X}_j) = E_{\boldsymbol{\varepsilon}} \left(E_{\varepsilon_i} \sup_{g_v} \sum_{j=1}^{n} \varepsilon_j g_v(\mathbf{X}_j) \right) \\ &\geq E_{\boldsymbol{\varepsilon}} \left(\sup_{g_v} E_{\varepsilon_i} \sum_{j=1}^{n} \varepsilon_j g_v(\mathbf{X}_j) \right) = Z_i. \end{aligned}$$

Finally we have:

$$\sum_i (Z - Z_i) = \sum_{i=1}^{n} E_{\boldsymbol{\varepsilon}} \left(\varepsilon_i \tau(X_i) + \sum_{j \neq i}^{n} \varepsilon_j \tau(X_j) - \sup_{g_{-v}} \sum_{j \neq i}^{n} \varepsilon_j g_v(X_j) \right) \leq \sum_{i=1}^{n} E_{\boldsymbol{\varepsilon}} \varepsilon_i \tau(X_i) = Z.$$

Therefore, following Theorem 8.1, we get that for all postive u

$$P_{\mathbf{X},\boldsymbol{\varepsilon}} \left[E_{\boldsymbol{\varepsilon}} W_{n,2,v}(t) - E_{\mathbf{X},\boldsymbol{\varepsilon}} W_{n,2,v}(t) \leq \frac{u}{n} \right] \leq \exp\left[-\frac{u^2}{E_{\mathbf{X},\boldsymbol{\varepsilon}} W_{n,2,v}(t)} \right].$$

As $E_{\mathbf{X},\boldsymbol{\varepsilon}} W_{n,2,v}(t) \leq Q_{n,v}(t)$, see Lemma 8.5 page 25, we get the expected result since for all positive x

$$P_{\mathbf{X}} \left[\mathbb{E}_{\varepsilon} W_{n,2,v}(t) \geq E_{\mathbf{X},\boldsymbol{\varepsilon}} W_{n,2,v}(t) + x \right] \leq \exp\left[-\frac{nx^2}{Q_{n,v}(t)} \right].$$

□

8.6.6. *Proof of Lemma 8.10 (page 26):*

From Lemma 8.8, page 25 with $t = \lambda_{n,v} = \delta$, we get that with probability greater than $1 - 4\exp(-n\lambda_{n,v}^2/2)$,

$$E_{\boldsymbol{\varepsilon}} W_{n,n,v}(\lambda_{n,v}) \leq \lambda_{n,v}^2 + W_{n,n,v}(\lambda_{n,v}).$$

The next step consists in comparing $W_{n,n,v}(\lambda_{n,v})$ and $W_{n,2,v}(2\lambda_{n,v})$. Recall that $\lambda_{n,v} \geq \nu_{n,v}$, see (13). Let g_v such that $\|g_v\|_n \leq \lambda_{n,v}$.

- When $\|g_v\|_2 \leq \lambda_{n,v}$, according to Lemma 8.6 (page 25), taking $b = 2$, since since $\|g_v\|_n \leq \lambda_{n,v}$, we get that with probability greater than $1 - \exp(-c_2 n\lambda_{n,v}^2)$,

$$\|g_v\|_n - \lambda_{n,v} \leq \|g_v\|_2 \leq \|g_v\|_n + \lambda_{n,v} \leq 2\lambda_{n,v}.$$

- When $\|g_v\|_2 \geq t$, we apply Lemma 8.7 (page 25) with $b = 2$. For any function g_v such that $\|g_v\|_\infty \leq 2$, and $\|g_v\|_2 \geq \lambda_{n,v}$, we have $\|g_v\|_2 \leq 2\|g_v\|_n \leq 2\lambda_{n,v}$.

This implies that, with probability greater than $1 - \exp(-c_2 n\lambda_{n,v}^2)$ we have

$$W_{n,n,v}(\lambda_{n,v}) \leq W_{n,2,v}(2\lambda_{n,v}).$$

We now study the process $W_{n,2,v}(\lambda_{n,v})$. By applying (48) in Lemma 8.9, page 26, with $\delta = t = \lambda_{n,v}$ we get that with probability greater than $1 - 4\exp(-n\lambda_{n,v}^2/8)$

$$W_{n,2,v}(\lambda_{n,v}) \leq \lambda_{n,v}^2 + E_{\boldsymbol{\varepsilon}}(W_{n,2,v}(\lambda_{n,v})).$$

It follows that

$$\begin{aligned} E_{\boldsymbol{\varepsilon}} W_{n,n,v}(\lambda_{n,v}) &\leq \lambda_{n,v}^2 + W_{n,n,v}(\lambda_{n,v})) \\ &\leq \lambda_{n,v}^2 + W_{n,2,v}(2\lambda_{n,v})) \\ &\leq 5\lambda_{n,v}^2 + E_{\boldsymbol{\varepsilon}}(W_{n,2,v}(2\lambda_{n,v})). \end{aligned}$$

Next, we apply (49) in Lemma 8.9, with $t = 2\lambda_{n,v}$ and $x = 4\lambda_{n,v}^2$. We get that

$$E_{\boldsymbol{\varepsilon}} W_{n,2,v}(2\lambda_{n,v}) \leq 4\lambda_{n,v}^2 + E_{\mathbf{X},\boldsymbol{\varepsilon}}(W_{n,2,v}(2\lambda_{n,v})),$$

with probability greater than

$$1 - 2\exp\left(-16\frac{n\lambda_{n,v}^4}{Q_{n,v}(2\lambda_{n,v})}\right) \geq 1 - 2\exp\left(-\frac{4n\lambda_{n,v}^2}{\Delta}\right).$$

The last inequality comes from the definition of $\nu_{n,v}$, see (12), and from the fact that $\lambda_{n,v} \geq \nu_{n,v}$, see (13).

Putting everything together, we get that with probability greater than $1 - c_1\exp(-c_2 n\lambda_{n,v}^2)$ for some positive constants c_1, c_2,

$$\begin{aligned} E_{\boldsymbol{\varepsilon}} W_{n,n,v}(\lambda_{n,v}) &\leq 9\lambda_{n,v}^2 + E_{\mathbf{X},\boldsymbol{\varepsilon}}(W_{n,2,v}(2\lambda_{n,v})) \\ &\leq 9\lambda_{n,v}^2 + Q_{n,v}(2\lambda_{n,v}), \text{ thanks to Lemma 8.5, page 25,} \\ &\leq 9\lambda_{n,v}^2 + 4\Delta\lambda_{n,v}^2. \end{aligned}$$

Applying once again Lemma 8.8, page 25, we get that

$$W_{n,n,v}(\lambda_{n,v}) \leq E_{\boldsymbol{\varepsilon}} W_{n,n,v}(\lambda_{n,v}) + \lambda_{n,v}^2 \leq \Big(10 + 4\Delta\Big)\lambda_{n,v}^2.$$

This ends the proof of the lemma by taking $\kappa = 10 + 4\Delta$.

□

8.7. **Algorithm: Propositions 8.1-8.4**

We consider the minimization of $C'(f_0, \boldsymbol{\theta})$ given at Equation (18). Because $C'(f_0, \boldsymbol{\theta})$ is convex and separable, we use a block coordinate descent algorithm, group v by group v. We refer to Bubeck [9] for a review on convex optimization.

In what follows, the group v is fixed, and for given values of f_0 and $\boldsymbol{\theta}_w, w \neq v$, we look for the minimizer of C' with respect to $\boldsymbol{\theta}_v$: Setting

$$\mathbf{R}_v = \mathbf{Y} - f_0 - \sum_{w \neq v} K_w \boldsymbol{\theta}_w,$$

we aim at minimizing with respect to $\boldsymbol{\theta}_v$

$$Q(\boldsymbol{\theta}_v) = \|\mathbf{R}_v - K_v\boldsymbol{\theta}_v\|^2 + \gamma_v^{'}\|K_v\boldsymbol{\theta}_v\| + \mu_v^{'}\|K_v^{1/2}\boldsymbol{\theta}_v\|.$$

If $\gamma_v' = \mu_v' = 0$, then $\boldsymbol{\theta}_v = K_v^{-1}\mathbf{R}_v$ is the solution. In what follows we consider the case where at least one of both is non zero.

If ∂Q_v denotes the subdifferential of $Q(\boldsymbol{\theta}_v)$ with respect to $\boldsymbol{\theta}_v$, we need to solve $0 \in \partial Q_v$. Let us recall that for all $v \in \mathcal{P}$ the matrices K_v are symmetric and strictly definite positive.

Let us begin with the calculation of the subdifferential of $\|K_v\theta_v\|$ with respect to θ_v: if $\theta_v \neq 0$, we have

$$\frac{\partial\|K_v\theta_v\|}{\partial\theta_v} = \frac{K_v^2\theta_v}{\|K_v\theta_v\|},$$

and if $\theta_v = 0$ the subdifferential is the set of $x \in \mathbb{R}^n$ such that $\|K_v^{-1}x\| \leq 1$.

Therefore if $\theta_v \neq 0$,

$$\partial Q_v \quad = \quad -2K_v\mathbf{R}_v + 2K_v^2\theta_v + \gamma_v'\frac{K_v^2\theta_v}{\|K_v\theta_v\|} + \mu_v'\frac{K_v\theta_v}{\|K_v^{1/2}\theta_v\|},$$

while if $\theta_v = 0$,

$$\partial Q_v = \left\{-2K_v\mathbf{R}_v + \gamma_v' t + \mu_v' s \text{ where } t, s \in \mathbb{R}^n \text{ such that } \|K_v^{-1}t\| \leq 1, \|K_v^{-1/2}s\| \leq 1\right\}.$$

Let us begin with the case $\mu_v' = 0$.

Proposition 8.1. *Let $(\rho)_+$ denotes the positive part of $\rho \in \mathbb{R}$. If $\mu_v' = 0$,*

$$\theta_v = \left(1 - \frac{\gamma_v'}{\|2R_v\|}\right)_+ K_v^{-1}\mathbf{R}_v.$$

Proof of Proposition 8.1. The problem comes to minimize

$$U(\boldsymbol{\beta}_v) = \|\mathbf{R}_v - \boldsymbol{\beta}_v\|^2 + \gamma_v^{'}\|\boldsymbol{\beta}_v\|$$

and to take $\boldsymbol{\theta}_v = K_v^{-1}\boldsymbol{\beta}_v$. The subdifferential of U is given by

$$\begin{aligned} \partial U_v(\boldsymbol{\beta}_v) &= -2\mathbf{R}_v + 2\boldsymbol{\beta}_v + \gamma_v'\frac{\boldsymbol{\beta}_v}{\|\boldsymbol{\beta}_v\|} \text{ if } \boldsymbol{\beta}_v \neq 0 \\ \partial U_v(0) &= \{-2\mathbf{R}_v + \gamma_v' t \text{ where } \|t\| \leq 1\}. \end{aligned}$$

We get

$$\boldsymbol{\beta}_v = \left(1 - \frac{\gamma_v'}{\|2R_v\|}\right)_+ \mathbf{R}_v.$$

□

Let $\mu_v' > 0$, the following proposition gives a necessary and sufficient condition for $\boldsymbol{\theta}_v = 0$ to be the minimizer of Q. For the sake of clarity let us recall the definitions of J and J^* given in Section 5.1: For all $x \in \mathbb{R}^n$

$$J(x) = \|2\mathbf{R}_v - \mu_v' K_v^{-1}x\|^2, \text{ and } J^* = \min\left\{J(x), \text{ for } x \text{ such that } \|K_v^{-1/2}x\| \leq 1\right\}.$$

Proposition 8.2. *The minimizer of $Q(\boldsymbol{\theta}_v)$ is $\boldsymbol{\theta}_v = 0$ if and only if $J^* \leq \gamma_v'^2$.*

Proof of Proposition 8.2.

(1) Let us assume that $\boldsymbol{\theta}_v = 0$ is the minimizer of $Q(\boldsymbol{\theta}_v)$. Then $0 \in \partial Q$ implies that there exists t^* and s^* such that $\|K_v^{-1}t^*\| \leq 1$ and $\|K_v^{-1/2}s^*\| \leq 1$ and such that $\gamma_v' t^* + \mu_v' s^* = 2K_v\mathbf{R}_v$.
If $\gamma_v' > 0$, then

$$K_v^{-1}t^* = \frac{1}{\gamma_v'}\left(2\mathbf{R}_v - \mu_v' K_v^{-1}s^*\right), \text{ and } \|K_v^{-1}t^*\| = \frac{1}{\gamma_v'}\sqrt{J(s^*)}.$$

Because $J^* \leq J(s^*)$ and $\|K_v^{-1}t^*\| \leq 1$, we get that $J^* \leq \gamma_v'^2$.
If $\gamma_v' = 0$, then $\mu_v' s^* = 2K_v\mathbf{R}_v$ and $J(s^*) = J^* = 0$.

(2) Let us now assume that $J^* \leq \gamma_v'^2$. Note that minimizing the convex function $J(s)$ over the convex set $\|K_v^{-1/2}s\| \leq 1$ has always a solution. Let us denote this solution by s^*.
If $\gamma_v' > 0$, let $t^* = (2K_v\mathbf{R}_v - \mu_v' s^*)/\gamma_v'$. Then $-2K_v\mathbf{R}_v + \gamma_v' t^* + \mu_v' s^* = 0$, and $-2K_v\mathbf{R}_v + \gamma_v' t^* + \mu_v' s^* \in \partial Q(0)$ since $\|K_v^{-1/2}s^*\| \leq 1$, and $\|K_v^{-1}t^*\| = J(s^*)/\gamma_v' \leq 1$. Therefore $\boldsymbol{\theta}_v = 0$ is the minimizer of Q.
If $\gamma_v' = 0$, and $J^* = 0$, then $-2K_v\mathbf{R}_v + \mu_v' s^* = 0$, and $\|K_v^{-1/2}s^*\| \leq 1$. Therefore $\boldsymbol{\theta}_v = 0$ is the minimizer of Q.

□

Proposition 8.3. *Let $\mu_v' > 0$ and $\boldsymbol{\theta}_v$ be the minimizer of Q.*

(1) *If $\|2K_v^{1/2}\mathbf{R}_v\| \leq \mu_v'$, then $\boldsymbol{\theta}_v = 0$*
(2) *If not, for $\rho > 0$, let*

$$\boldsymbol{\theta}(\rho) = 2\mu_v'\left(\mu_v'^2 K_v^{-1} + \rho I_n\right)^{-1} K_v^{-1/2}R_v,$$

and let ρ^* defined as $\|\boldsymbol{\theta}(\rho^*)\| = 1$. Then $J(\theta(\rho^*)) \leq \gamma_v'^2$ if and only if $\boldsymbol{\theta}_v = 0$,

Proof of Proposition 8.3. Minimizing $J(x)$ under the constraint $\|K_v^{-1/2}x\| \leq 1$ is equivalent to minimizing $K(\boldsymbol{\beta}) = \|2\mathbf{R}_v - \mu_v' K_v^{-1/2}\boldsymbol{\beta}\|^2$ under the constraint $\|\boldsymbol{\beta}\|^2 \leq 1$. Let $\boldsymbol{\beta}^* = 2K_v^{1/2}\mathbf{R}_v/\mu_v'$. Then $K(\boldsymbol{\beta}^*) = 0$, which is smaller than $\gamma_v'^2$, and if $\|\boldsymbol{\beta}^*\| \leq 1$, following Proposition 8.2, we get $\boldsymbol{\theta}_v = 0$.

If $\|\boldsymbol{\beta}^*\| > 1$, we have to solve a ridge regression problem by minimizing $K(\boldsymbol{\beta}) + \rho\|\boldsymbol{\beta}\|^2$ for some positive ρ. The solution is given by $\boldsymbol{\theta}(\rho)$. Let us note that $\|\boldsymbol{\theta}(\rho)\|$ decreases to 0 when ρ tends to infinity and that its maximum is $\|\boldsymbol{\theta}(0)\| = \|\boldsymbol{\beta}^*\|$. Therefore if $\|\boldsymbol{\beta}^*\| > 1$, there exists ρ^* such that $\|\boldsymbol{\theta}(\rho^*)\| = 1$. Following Proposition 8.2, Proposition 8.3 is proved. A numerical procedure can be used for calculating this ρ^*. □

Let us now consider the case where $\boldsymbol{\theta}_v$ is non zero. It should satisfy the subgradient condition $\partial Q_v = 0$ which leads to

$$\theta_v = \left(2K_v^2 + \gamma_v' \frac{K_v^2}{\|K_v\theta_v\|} + \mu_v' \frac{K_v}{\|K_v^{1/2}\theta_v\|}\right)^{-1} 2K_v\mathbf{R}_v$$

which is equivalent to Equation (22).

Proposition 8.4. *For all $\rho_1, \rho_2 > 0$ let*

$$\boldsymbol{\theta}(\rho_1, \rho_2) = \left(K_v + \rho_1 K_v + \rho_2 I_n\right)^{-1}\mathbf{R}_v.$$

If $\mu_v' > 0$, there exists a non zero solution to Equation (22) if and only if there exists $\rho_1, \rho_2 > 0$ such that

$$\left.\begin{array}{rcl} \gamma_v' & = & 2\rho_1\|K_v\boldsymbol{\theta}(\rho_1,\rho_2)\| \\ \mu_v' & = & 2\rho_2\|K_v^{1/2}\boldsymbol{\theta}(\rho_1,\rho_2)\| \end{array}\right\} \tag{64}$$

Then $\boldsymbol{\theta}_v = \boldsymbol{\theta}(\rho_1, \rho_2)$.

Proof of Proposition 8.4. If there exists a non zero solution to Equation (22) then $\|K_v\boldsymbol{\theta}_v\|$ and $\|K_v^{1/2}\boldsymbol{\theta}_v\|$ are non zero because K_v is definite positive. Let $\rho_1 = \gamma_v'/2\|K_v\boldsymbol{\theta}_v\|$ and $\rho_2 = \mu_v'/2\|K_v^{1/2}\boldsymbol{\theta}_v\|$, then

$$\boldsymbol{\theta}(\rho_1,\rho_2) = \left(K_v + \frac{\gamma_v'}{2\|K_v\boldsymbol{\theta}_v\|}K_v + \frac{\mu_v'}{2\|K_v^{1/2}\boldsymbol{\theta}_v\|}I_n\right)^{-1}\mathbf{R}_v = \boldsymbol{\theta}_v,$$

and, for such ρ_1, ρ_2, Equation (64) is satisfied.

Conversely, if there exist ρ_1, ρ_2 such that Equation (64) is satisfied, then necessarily $\|K_v\boldsymbol{\theta}(\rho_1,\rho_{2)})\|$ and $\|K_v^{1/2}\boldsymbol{\theta}(\rho_1,\rho_{2)})\|$ are non zero and $\rho_1 = \gamma_v'/2\|K_v\boldsymbol{\theta}(\rho_1,\rho_2)\|$ and $\rho_2 = \mu_v'/2\|K_v^{1/2}\boldsymbol{\theta}(\rho_1,\rho_2)\|$. Then

$$\boldsymbol{\theta}(\rho_1,\rho_2) = \left(K_v + \frac{\gamma_v'}{2\|K_v\boldsymbol{\theta}(\rho_1,\rho_2)\|}K_v + \frac{\mu_v'}{2\|K_v^{1/2}\boldsymbol{\theta}(\rho_1,\rho_2)\|}I_n\right)^{-1}\mathbf{R}_v,$$

which is exactly Equation (22) calculated in $\boldsymbol{\theta}_v = \boldsymbol{\theta}(\rho_1,\rho_2)$. □

Taking into account that $\|K_{v^{1/2}}\boldsymbol{\theta}_v\| \leq r_v$. As already mentionned in Section 5, one may want to minimize $C(f_0,\boldsymbol{\theta})$ under the additional constraint that $\|K_{v^{1/2}}\boldsymbol{\theta}_v\| \leq r_v$ for some positive constant $r_v, v \in \mathcal{P}$.

For each group v, we have thus to minimize $Q(\boldsymbol{\theta}_v)$ under the constraint $\|K_{v^{1/2}}\boldsymbol{\theta}_v\| \leq r_v$. We know that this problem is equivalent to minimize

$$Q(\boldsymbol{\theta}_v) + \lambda\|K_v^{1/2}\boldsymbol{\theta}_v\| = \|\mathbf{R}_v - K_v\boldsymbol{\theta}_v\|^2 + \gamma_v^{'}\|K_v\boldsymbol{\theta}_v\| + (\mu_v^{'} + \lambda)\|K_v^{1/2}\boldsymbol{\theta}_v\|.$$

for some λ that depends on r_v.

Let us first remark that, for a fixed $\gamma_v^{'}$, and $\lambda \geq 0$, if $\widehat{\boldsymbol{\theta}}_v(\mu_v^{'} + \lambda)$ minimizes $Q(\boldsymbol{\theta}_v) + \lambda\|K_v^{1/2}\boldsymbol{\theta}_v\|$ with respect to $\boldsymbol{\theta}_v$, then $\|K_v^{1/2}\widehat{\boldsymbol{\theta}}_v(\mu_v^{'} + \lambda)\| \leq \|K_v^{1/2}\widehat{\boldsymbol{\theta}}_v(\mu_v^{'})\|$. It can be easily proved by writing

$$\begin{aligned} Q\left(\widehat{\boldsymbol{\theta}}_v(\mu_v^{'} + \lambda)\right) + \lambda\|K_v^{1/2}\widehat{\boldsymbol{\theta}}_v(\mu_v^{'} + \lambda)\| &\leq Q\left(\widehat{\boldsymbol{\theta}}_v(\mu_v^{'})\right) + \lambda\|K_v^{1/2}\widehat{\boldsymbol{\theta}}_v(\mu_v^{'})\| \\ &\leq Q\left(\widehat{\boldsymbol{\theta}}_v(\mu_v^{'} + \lambda)\right) + \lambda\|K_v^{1/2}\widehat{\boldsymbol{\theta}}_v(\mu_v^{'})\|. \end{aligned}$$

Therefore, one can proceed as follows: calculate $\widehat{\boldsymbol{\theta}}_v$ for $\lambda = 0$. If $\|K_v^{1/2}\widehat{\boldsymbol{\theta}}_v\| \leq r_v$, then one go to the next step of the algorithm. If $\|K_v^{1/2}\widehat{\boldsymbol{\theta}}_v\| > r_v$, one increases λ untill $\|K_v^{1/2}\widehat{\boldsymbol{\theta}}_v\| = r_v$.

References

[1] Sylvain Arlot, Alain Celisse, et al. A survey of cross-validation procedures for model selection. Statistics surveys, 4:40–79, 2010.

[2] Francis Bach. High-Dimensional Non-Linear Variable Selection through Hierarchical Kernel Learning. working paper or preprint, September 2009.

[3] Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local Rademacher complexities. Ann. Statist., 33(4):1497–1537, 2005.

[4] A. Berlinet and C. Thomas-Agnan. Reproducing Kernel Hilbert Spaces in Probability and Statistics. Springer US, 2003.

[5] G. Blatman and B. Sudret. Adaptive sparse polynomial chaos expansion based on least angle regression. Journal of computational Physics, 230:2345–2367, 2011.

[6] Christer Borell. The Brunn-Minkowski inequality in Gauss space. Invent. Math., 30:207–216, 1975.

[7] Stéphane Boucheron, Gabor Lugosi, and Pascal Massart. A sharp concentration inequality with applications. Random Structures & Algorithms, 16(3):277–292, 2000.

[8] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. Found. Trends Mach. Learn., 3(1):1–122, 2011.

[9] S. Bubeck. Convex optimization: Algorithms and complexity. volume 8 of Foundation and trends in Machine Learning, pages 231–357. now, 2015.

[10] Fan Chung and Linyuan Lu. Concentration inequalities and martingale inequalities: a survey. Internet Math., 3(1):79–127, 2006.
[11] B.S. Cirel'son, I.A. Ibragimov, and V.N. Sudakov. Norms of Gaussian sample functions. Proc. 3rd Japan-USSR Symp. Probab. Theory, Taschkent 1975, Lect. Notes Math. 550, 20-41 (1976)., 1976.
[12] N. Durrande, D. Ginsbourger, O. Roustant, and L. Carraro. {ANOVA} kernels and {RKHS} of zero mean functions for model-based sensitivity analysis. Journal of Multivariate Analysis, 115(0):57 – 67, 2013.
[13] Jerome H. Friedman. Multivariate adaptive regression splines. Ann. Statist., 19(1):1–141, 1991. With discussion and a rejoinder by the author.
[14] Roger G. Ghanem and Pol D. Spanos. Stochastic Finite Elements: A Spectral Approach. Springer-Verlag New York, Inc., New York, NY, USA, 1991.
[15] L. Gu and F.J. Wu. A unified framework for uncertainty and sensitivity analysis of computational models with many input. In SIMUL 2014 : The Sixth International Conference on Advances in System Simulation. IARIA XPS Press, 2014.
[16] SR Gunn and JS Kandola. Structural modelling with sparse kernels. MACHINE LEARNING, 48(1-3):137–163, 2002.
[17] Bertrand Iooss and Paul Lemaître. A review on global sensitivity analysis methods. In C. Meloni and G. Dellino, editors, Uncertainty management in Simulation-Optimization of Complex Systems: Algorithms and Applications. Springer, 2015.
[18] Adel Javanmard and Andrea Montanari. Model selection for high-dimensional regression under the generalized irrepresentability condition. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, Advances in Neural Information Processing Systems 26, pages 3012–3020. Curran Associates, Inc., 2013.
[19] K. Kandasamy and Y. Yu. Additive approximations in high dimensional nonparametric regression via the salsa. In Proceedings of the 33rd International Conference on Machine Learning, New-York, volume 48. JMLR: W & CP, 2016.
[20] George S. Kimeldorf and Grace Wahba. A correspondence between bayesian estimation on stochastic processes and smoothing by splines. Ann. Math. Statist., 41(2):495–502, 04 1970.
[21] V. Koltchinskii and M. Yuan. Sparse recovery in large ensembles of kernel machines. In Proceeding of the 21st Annual Conference on Learning Theory, Helsinki, pages 229–238. 2008.
[22] Vladimir Koltchinskii and Ming Yuan. Sparsity in multiple kernel learning. The Annals of Statistics, 38(6):3660–3695, 12 2010.
[23] M. Ledoux and M. Talagrand. Probability in Banach Spaces: Isoperimetry and Processes. Ergebnisse der Mathematik und ihrer Grenzgebiete. U.S. Government Printing Office, 1991.
[24] Chenlei Leng, Yi Lin, and Grace Wahba. A note on the Lasso and related procedures in model selection. STATISTICA SINICA, 16(4):1273–1284, OCT 2006.
[25] Yi Lin and Hao Helen Zhang. Component selection and smoothing in multivariate nonparametric regression. Ann. Statist., 34(5):2272–2297, 2006.
[26] P. Massart. Concentration inequalities and model selection, volume 1896 of Lecture Notes in Mathematics. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.
[27] Lukas Meier, Sara van de Geer, and Peter Bühlmann. High-dimensional additive modeling. The Annals of Statistics, 37(6B):3779–3821, 12 2009.
[28] Shahar Mendelson. Geometric parameters of kernel machines. In Computational learning theory (Sydney, 2002), volume 2375 of Lecture Notes in Comput. Sci., pages 29–43. Springer, Berlin, 2002.
[29] Gilles Pisier. The volume of convex bodies and Banach space geometry, volume 94 of Cambridge Tracts in Mathematics. Cambridge University Press, Cambridge, 1989.
[30] R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0.
[31] Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Minimax-optimal rates for sparse additive models over kernel classes via convex programming. J. Mach. Learn. Res., 13:389–427, 2012.
[32] Pradeep Ravikumar, John Lafferty, Han Liu, and Larry Wasserman. Sparse additive models. J. R. Stat. Soc. Ser. B Stat. Methodol., 71(5):1009–1030, 2009.
[33] Philippe Rigollet and Alexandre B. Tsybakov. Sparse Estimation by Exponential Weighting. STATISTICAL SCIENCE, 27(4):558–575, NOV 2012.
[34] A. Saltelli, K. Chan, and E.M. (eds) Scott. Sensitivity analysis. J. Wiley & Sons, 2000.
[35] Andrea Saltelli, Marco Ratto, Terry Andres, Francesca Campolongo, Jessica Cariboni, Debora Gatelli, Michaela Saisana, and Stefano Tarantola. Global sensitivity analysis: The primer. Chichester: John Wiley & Sons. , 2008.
[36] I. M. Sobol. Sensitivity estimates for nonlinear mathematical models. Math. Model. Comput. Exp., 1(4):407–414, 1993.
[37] I.M Sobol. Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates. Mathematics and Computers in Simulation, 55(13):271 – 280, 2001. The Second {IMACS} Seminar on Monte Carlo Methods.
[38] Christian Soize and Roger Ghanem. Physical systems with random uncertainties: chaos representations with arbitrary probability measure. SIAM Journal on Scientific Computing, 26(2):395–410, 2004.
[39] Samir Touzani. Response surface methods based on analysis of variance expansion for sensitivity analysis. PhD thesis, Université Grenoble, 2011.

[40] Samir Touzani and Daniel Busby. Smoothing spline analysis of variance approach for global sensitivity analysis of computer codes. RELIABILITY ENGINEERING & SYSTEM SAFETY, 112:67–81, APR 2013.

[41] A. W. van der Vaart. Asymptotic statistics. Cambridge series in statistical and probabilistic mathematics. Cambridge University Press, Cambridge (UK), New York (N.Y.), 1998. Autre tirage : 2000 (dition broche), 2005, 2006, 2007.

[42] Grace Wahba. Spline models for observational data, volume 59 of CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1990.

[43] Grace Wahba, Yuedong Wang, Chong Gu, Ronald Klein, and Barbara Klein. Smoothing spline ANOVA for exponential families, with application to the Wisconsin Epidemiological Study of Diabetic Retinopathy. Ann. Statist., 23(6):1865–1895, 1995.