



Optimal upper and lower bounds for the true and empirical excess risks in heteroscedastic least-squares regression

Adrien Saumard

► To cite this version:

Adrien Saumard. Optimal upper and lower bounds for the true and empirical excess risks in heteroscedastic least-squares regression. 2010. hal-00512304v1

HAL Id: hal-00512304

<https://hal.science/hal-00512304v1>

Preprint submitted on 13 Sep 2010 (v1), last revised 26 Jun 2015 (v3)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Optimal upper and lower bounds for the true and empirical excess risks in heteroscedastic least-squares regression

A. Saumard
University Rennes 1, IRMAR
adrien.saumard@univ-rennes1.fr

August 29, 2010

Abstract

We consider the estimation of a bounded regression function with nonparametric heteroscedastic noise. We are interested by the true and empirical excess risks of the least-squares estimator on a finite-dimensional vector space. For these quantities, we give upper and lower bounds in probability that are optimal at the first order. Moreover, these bounds show the equivalence between the true and empirical excess risks when, among other things, the least-squares estimator is consistent in sup-norm towards the projection of the regression function onto the considered model. Consistency in sup-norm is then proved for suitable histogram models and more general models of piecewise polynomials that are endowed with a localized basis structure.

keywords: Least-squares regression, Heteroscedasticity, Excess risk, Lower bounds, Empirical process, M-estimation.

1 Introduction

This article is devoted to least-squares estimation of a regression function on a finite dimensional linear model. We derive sharp upper and lower bounds in probability for the true and empirical excess risks of the least-squares estimator. We only focus on the “stochastic” parts of the excess risks and we do not discuss on the possible behaviors of the bias of the model, neither on the trade-off that can be achieved between the bias and the variance terms. However, our framework is closely related to the method of sieves and particularly to the work of Birgé and Massart [6]. The leading idea of the sieve method is to replace a complicated set of parameters by a more tractable one having good approximation properties, an idea that goes back to Cencov [9], considering orthogonal series for density estimation, and to Le Cam [17] where the author investigate the relationship between the metric structure of the parameter space and the rate of optimal estimators, see also Le Cam [18] Section 16.5 and Le Cam and Yang [19] Section 6.5. Since the formalization of the sieve method by Grenander [12], many authors have considered this method for MLEs or more general M-estimators. Inspired by a work of van de Geer [31] in regression, Birgé and Massart [5] proposed to study minimum of contrast estimation on general parameter spaces under entropy with bracketing conditions, and proved that sub-optimality of M-estimators can happen when the parameter space is too large. The entropy with bracketing covering property has then been a central tool for studying minimum contrast estimation on general sieves in Shen and Wong [25], Wong and Shen [36] and van de Geer [32]. Van de Geer [33] more recently considers M-estimation with convex loss functions, a situation that allows to “localize” the problem to a small neighborhood in the parameter space. In a series of papers that started with Stone [27], Stone extensively studied log-spline density estimation and spline regression, see [26], [28], [29] and Stone and Kooperberg [16].

Birgé and Massart [6] introduced metric properties on the sieves relating the L_2 -structure to the L_∞ -structure, and which involve covering numbers related to both L_2 and L_∞ norms. These metric conditions are satisfied for linear sieves commonly used in practice, such as Fourier expansions, piecewise polynomials and wavelet expansions, but also for non-linear sieves, which can have better approximation properties, and that include finite linear combinations of D sigmoidal functions related to neural networks, see also Barron

[2], and histograms generated by any partition on $[0, 1]$ into D subintervals. Birgé and Massart [6] pointed out that the use of covering numbers, even in the case of linear sieves, is quite natural since linearity is lost on the contrasted functions for a non-linear contrast such as in the regression and maximum likelihood estimation contexts. This allows them to derive sharp exponential bounds and rates of convergence for the excess risk on such sieves, using in particular a Talagrand’s concentration inequality for the supremum of the empirical process.

The starting point of our method is to remark that the least-squares contrast in regression can be expanded to the sum of a linear part and a quadratic part. This allows us to recover some linearity on the contrasted functions and avoid the use of entropy methods to control the empirical process on a linear model. The gain is that we achieve optimal rates of convergence for the true and empirical excess risks with exact constants, for models of reasonable dimension. In our study, the metric properties defined by Birgé and Massart in [6] play a center role, in particular the notion of localized basis. In addition, we point out the importance of the behavior in sup-norm of the least-squares estimator and we have to assume its consistency in sup-norm towards the linear projection of the regression function onto the model. We show that such a condition is satisfied by histograms and piecewise polynomial models when they are endowed with a localized basis structure, which corresponds in that case to a lower regularity assumption on the considered partition. By doing so, we recover some recent results of Arlot and Massart [1] on the empirical and true excess risk for least-squares estimator on histogram models, and extend them to the case of piecewise polynomials.

Although we do not make an explicit use of the margin conditions that can hold in the context of bounded regression, this property also connects our work with the statistical learning theory. The margin conditions were first introduced by Mammen and Tsybakov [22] in the context of discrimination analysis. They allow to get faster rates of convergence than the pioneering bounds of Vapnik and Červonenkis, see [35] and [34], using “localization” techniques. Under entropy with bracketing conditions, Tsybakov [30] shows some fast rates in the binary classification setting, and these results have been recovered and extended by Massart and Nédélec [24], Koltchinskii [15] and by Giné and Koltchinskii [11], where the authors also give asymptotic results for ratio type empirical processes. The obtained bounds are proved to be optimal in a minimax sense in [24], up to a logarithmic factor shown by Massart and Nédélec to be unavoidable for “rich” VC-classes. This analysis is refined in [11] by the use of localized $L_2(P)$ -envelopes of the models, allowing to remove the logarithmic factor in good cases.

The main tools in [24], [15] and [11] are Talagrand’s type concentration inequalities for the supremum of the empirical process and the *slicing* or *peeling* technique through the use of ratio type empirical processes. The slicing technique consists in considering subsets of the model, called the slices, and that are localized in terms of excess risk, a quantity that is related to the variance of the empirical process through margin conditions. Our method of proof may be viewed as a variant of the technique of slicing that allows to avoid the use of ratio type empirical processes, where in general sharp constants are lost due to the use of chaining techniques. The very first lines of our proofs differ from those of [24], [15] and [11], and permit in particular to relate both upper and lower bounds for the excess risks of the M-estimator to the behavior of the empirical process indexed by contrasted functions on localized slices of excess risk. This rewriting of the problem of upper and lower bounds for the excess risks is closely related to the work of Bartlett and Mendelson [3], where a “direct” approach of the empirical minimization algorithm is proposed, and proved to lead to more accurate bounds than the traditional “structural” approach developed in [24], [11] or [15].

Finally, it should be said that, in a quite unusual manner, we not only focus on the (true) excess risk of the least-squares estimator, which corresponds to the quadratic loss, but we also give attention on the empirical excess risk. While the true excess risk was extensively studied in the last decade as mentioned above, much less is known about the empirical excess risk. In fact, the study of the latter quantity is motivated by some recent advances in model selection theory that are due to Birgé and Massart [4], followed by Arlot and Massart [1]. More precisely, Birgé and Massart [4] have focused on optimality of penalization methods in a generalized linear Gaussian framework and have proposed the so-called *slope heuristics*, which is a practical method of data-driven and efficient calibration of penalties. Then, Arlot and Massart [1] extended this method to more general bounded M-estimation problems, proved its efficiency on heteroscedastic regression, and pointed out that the empirical excess risk on a fixed model was a fundamental quantity in this problem, in the sense that its equivalence with the true excess risk is the keystone to prove the slope phenomenon. Hopefully, the main result of the present paper shows that this equivalence is a quite general fact in regression, when using linear models. We also notice that we avoid the use of Boucheron and Massart results, recently exposed in [7], that

establish concentration inequalities for the empirical excess risk, in some general bounded M-estimation setting with generalized margin conditions that they call low noise conditions,

The article is organized as follows. We present the statistical framework in Section 2 where we show in particular the existence of an expansion of the least-squares regression contrast into the sum of a linear and a quadratic part. We then derive general results for models of reasonable dimensions and also for small models in Section 3. General results are then applied in the case of histograms and piecewise polynomials in Section 4 and 5 respectively, where explicit rates of convergence in sup-norm are derived. Finally, the proofs are postponed to the end of the article.

2 Regression framework and notations

2.1 Least-squares estimator

Let $(\mathcal{X}, \mathcal{T}_{\mathcal{X}})$ be a measurable space and set $\mathcal{Z} = \mathcal{X} \times \mathbb{R}$. We assume that $\xi_i = (X_i, Y_i) \in \mathcal{X} \times \mathbb{R}$, $i \in \{1, \dots, n\}$ are n i.i.d. observations with law P . The marginal law of X_i is denoted by P^X . We assume that the data satisfy the following relation

$$Y_i = s_*(X_i) + \sigma(X_i) \varepsilon_i, \quad (1)$$

where $s_* \in L_2(P^X)$, ε_i are i.i.d. random variables with mean 0 and variance 1 conditionally to X_i and $\sigma : \mathcal{X} \rightarrow \mathbb{R}$ is an heteroscedastic noise level. A generic random variable of law P , independent of (ξ_1, \dots, ξ_n) , is denoted by $\xi = (X, Y)$.

Hence, s_* is the regression function of Y with respect to X , that we want to estimate. Given a finite dimensional linear vector space M , we denote by s_M the linear projection of s_* onto M in $L^2(P^X)$ and by D the linear dimension of the model M .

We consider on M a least-squares estimator s_n (possibly non unique), defined as follows

$$s_n \in \arg \min_{s \in M} \left\{ \frac{1}{n} \sum_{i=1}^n (Y_i - s(X_i))^2 \right\}. \quad (2)$$

So, if we denote by

$$P_n = \frac{1}{n} \sum_{i=1}^n \delta_{(X_i, Y_i)}$$

the empirical distribution of the data and by $K : L_2(P^X) \rightarrow L_1(P)$ the least-squares contrast, defined by

$$K(s) = (x, y) \in \mathcal{Z} \rightarrow (y - s(x))^2, \quad s \in L_2(P^X)$$

we then remark that s_n belongs to the general class of M-estimators, as it satisfies

$$s_n \in \arg \min_{s \in M} \{P_n(K(s))\}. \quad (3)$$

2.2 Excess risk and contrast

As defined in (3), s_n is the empirical risk minimizer of the least-squares contrast. The regression function s_* can be defined as the minimizer in $L_2(P^X)$ of the mean of the contrast over the unknown law P ,

$$s_* = \arg \min_{s \in L_2(P^X)} PK(s),$$

where

$$PK(s) = P(Ks) = PKs = \mathbb{E}[K(s)(X, Y)] = \mathbb{E}[(Y - s(X))^2]$$

is called the risk of the function s . In particular we have $PKs_* = \mathbb{E}[\sigma^2(X)]$. We first notice that for any $s \in L^2(P^X)$, if we denote by

$$\|s\|_2 = \left(\int_{\mathcal{X}} s^2 dP^X \right)^{1/2}$$

its quadratic norm, then we have, by (1) above,

$$\begin{aligned}
PKs - PKs_* &= P(Ks - Ks_*) \\
&= \mathbb{E} \left[(Y - s(X))^2 - (Y - s_*(X))^2 \right] \\
&= \mathbb{E} [(s_* - s)(X) (2(Y - s_*(X)) + (s_* - s)(X))] \\
&= \mathbb{E} [(s_* - s)^2(X)] + 2\mathbb{E} \left[(s_* - s)(X) \underbrace{\mathbb{E}[Y - s_*(X) | X]}_{=0} \right] \\
&= \|s - s_*\|_2^2 \geq 0,
\end{aligned}$$

and $PKs - PKs_*$ is called the excess risk of s . So if we denote by s_M the linear projection of s_* onto M in $L^2(P^X)$, we have

$$PKs_M - PKs_* = \inf_{s \in M} \{PKs - PKs_*\}, \quad (4)$$

and for all $s \in M$

$$P^X(s \cdot (s_M - s_*)) = 0. \quad (5)$$

From (4), we deduce that

$$s_M = \arg \min_{s \in M} PK(s).$$

Our goal is to study the performance of the least-squares estimator, that we measure by its excess risk. So we are mainly interested by the random quantity $P(Ks_n(M) - Ks_*)$. Moreover, as we can write

$$P(Ks_n(M) - Ks_*) = P(Ks_n(M) - Ks_M) + P(Ks_M - Ks_*)$$

we naturally focus on the quantity

$$P(Ks_n(M) - Ks_M) \geq 0$$

that we want to upper and lower bound in probability. Abusively we will often call this last quantity the excess risk of the estimator on M or the true excess risk of $s_n(M)$, in opposition to the empirical excess risk for which the expectation is taken over the empirical measure,

$$P_n(Ks_M - Ks_n(M)) \geq 0.$$

The following lemma establishes the key expansion of the regression contrast around s_M on M . This expansion exhibits a linear part and a quadratic part.

Lemma 1 *We have, for every $z = (x, y) \in \mathcal{Z}$,*

$$(Ks)(z) - (Ks_M)(z) = \psi_{1,M}(z)(s - s_M)(x) + \psi_2((s - s_M)(x)) \quad (6)$$

with $\psi_{1,M}(z) = -2(y - s_M(x))$ and $\psi_2(t) = t^2$, for all $t \in \mathbb{R}$. Moreover, for all $s \in M$,

$$P(\psi_{1,M} \cdot s) = 0. \quad (7)$$

Proof. Start with

$$\begin{aligned}
(Ks)(z) - (Ks_M)(z) &= (y - s(x))^2 - (y - s_M(x))^2 \\
&= ((s - s_M)(x))((s - s_M)(x) - 2(y - s_M(x))) \\
&= -2(y - s_M(x))((s - s_M)(x)) + ((s - s_M)(x))^2,
\end{aligned}$$

which gives (6). Moreover, observe that for any $s \in M$,

$$P(\psi_{1,M} \cdot s) = -2\mathbb{E}[(Y - s_*(X))s(X)] + 2\mathbb{E}[s(X)(s_M - s_*)(X)]. \quad (8)$$

We have

$$\mathbb{E}[(Y - s_*(X)) s(X)] = \mathbb{E} \left[\underbrace{\mathbb{E}[(Y - s_*(X)) | X]}_{=0} s(X) \right] = 0 . \quad (9)$$

and, by (5),

$$\mathbb{E}[s(X) (s_M - s_*)(X)] = P^X(s \cdot (s_M - s_*)) = 0 . \quad (10)$$

Combining (8), (9) and (10) we get that for any $s \in M$, $P(\psi_{1,M} \cdot s) = 0$. This concludes the proof. ■

3 True and empirical excess risk bounds

In this section, we show that under assumptions that extend a previous work of Arlot and Massart [1], the true excess risk is equivalent to the empirical one for models of reasonable dimension.

More precisely, we assume that M is a linear model with a localized basis in $L_2(P)$ and that the least-squares estimator is consistent in sup-norm towards the linear projection s_M on M of the target s_* when the dimension of the model is not too heavy. This is a natural generalization of the case of histograms studied by Arlot and Massart in [1], since the assumption of lower regularity of the partitions made in their work indeed provides the histograms with a structure of localized basis in $L_2(P)$, see Lemma 5. We further show in Lemma 6 that the assumption of consistency is satisfied for histograms.

3.1 Main assumptions

We turn now to the statement of some assumptions that will be needed to derive our results in Section 3.2. These assumptions will be further discussed in Section 3.3.

Boundedness assumptions :

- **(H1)** The data and the linear projection of the target onto M are bounded : a positive finite constant A exists such that

$$|Y_i| \leq A \text{ a.s.} \quad (11)$$

and

$$\|s_M\|_\infty \leq A . \quad (12)$$

Hence, from **(H1)** we deduce that

$$\|s_*\|_\infty = \|\mathbb{E}[Y | X = \cdot]\|_\infty \leq A \quad (13)$$

and that there exists a constant $\sigma_{\max} > 0$ such that

$$\sigma^2(X_i) \leq \sigma_{\max}^2 \leq A^2 \text{ a.s.} \quad (14)$$

Moreover, as $\psi_{1,M}(z) = -2(y - s_M(x))$ for all $z = (x, y) \in \mathcal{Z}$, we also deduce that

$$|\psi_{1,M}(X_i, Y_i)| \leq 4A \text{ a.s.} \quad (15)$$

- **(H2)** The heteroscedastic noise level σ is uniformly bounded from below : a positive finite constant σ_{\min} exists such that

$$0 < \sigma_{\min} \leq \sigma(X_i) \text{ a.s.}$$

Models with localized basis in $L_2(P^X)$:

Let us define a function Ψ_M on \mathcal{X} , that we call the unit envelope of M , such that

$$\Psi_M(x) = \frac{1}{\sqrt{D}} \sup_{s \in M, \|s\|_2 \leq 1} |s(x)| . \quad (16)$$

As M is a finite dimensional real vector space, the supremum in (16) can also be taken over a countable subset of M , so Ψ_M is a measurable function.

- **(H3)** The unit envelope of M is uniformly bounded on \mathcal{X} : a positive constant $A_{3,M}$ exists such that

$$\|\Psi_M\|_\infty \leq A_{3,M} < \infty .$$

The following assumption is stronger than **(H3)**.

- **(H4)** Existence of a localized basis in $(M, \|\cdot\|_2)$: there exists an orthonormal basis $\varphi = (\varphi_k)_{k=1}^D$ in $(M, \|\cdot\|_2)$ that satisfies, for a positive constant $r_M(\varphi)$ and all $\beta = (\beta_k)_{k=1}^D \in \mathbb{R}^D$,

$$\left\| \sum_{k=1}^D \beta_k \varphi_k \right\|_\infty \leq r_M(\varphi) \sqrt{D} |\beta|_\infty ,$$

where $|\beta|_\infty = \max \{|\beta_k|; k \in \{1, \dots, D\}\}$ is the sup-norm of the D -dimensional vector β .

Remark 2 *(H4) implies (H3) and in that case $A_{3,M} = r_M(\varphi)$ is convenient.*

The assumption of consistency in sup-norm :

In order to handle second order terms in the expansion of the contrast (6) we assume that the least-squares estimator is consistent for the sup-norm on the space \mathcal{X} . More precisely, this requirement can be stated as follows.

- **(H5)** Assumption of consistency in sup-norm : for any $A_+ > 0$, if M is a model of dimension D satisfying

$$D \leq A_+ \frac{n}{(\ln n)^2} ,$$

then for every $\alpha > 0$, we can find a positive integer n_1 and a positive constant A_{cons} satisfying the following property : there exists $R_{n,D,\alpha} > 0$ depending on D , n and α , such that

$$R_{n,D,\alpha} \leq \frac{A_{cons}}{\sqrt{\ln n}} \tag{17}$$

and by setting

$$\Omega_{\infty,\alpha} = \{\|s_n - s_M\|_\infty \leq R_{n,D,\alpha}\} , \tag{18}$$

it holds for all $n \geq n_1$,

$$\mathbb{P}[\Omega_{\infty,\alpha}] \geq 1 - n^{-\alpha} . \tag{19}$$

3.2 Theorems

We state here the general results of this article, that will be applied in Section 4 and 5 in the case of piecewise constant functions and piecewise polynomials respectively.

Theorem 3 *Let $A_+, A_-, \alpha > 0$ and let M be a linear model of finite dimension D . Assume that **(H1)**, **(H2)**, **(H4)** and **(H5)** hold and take $\varphi = (\varphi_k)_{k=1}^D$ an orthonormal basis of $(M, \|\cdot\|_2)$ satisfying **(H4)**. If it holds*

$$A_- (\ln n)^2 \leq D \leq A_+ \frac{n}{(\ln n)^2} , \tag{20}$$

*then a positive finite constant A_0 exists, only depending on α, A_- and on the constants $A, \sigma_{\min}, r_M(\varphi)$ defined in the assumptions **(H1)**, **(H2)** and **(H4)** respectively, such that by setting*

$$\varepsilon_n = A_0 \max \left\{ \left(\frac{\ln n}{D} \right)^{1/4} , \left(\frac{D \ln n}{n} \right)^{1/4} , \sqrt{R_{n,D,\alpha}} \right\} , \tag{21}$$

we have for all $n \geq n_0(A_-, A_+, A, A_{\text{cons}}, r_M(\varphi), \sigma_{\min}, n_1, \alpha)$,

$$\mathbb{P} \left[P(Ks_n - Ks_M) \geq (1 - \varepsilon_n) \frac{1}{4} \frac{D}{n} \mathcal{K}_{1,M}^2 \right] \geq 1 - 5n^{-\alpha}, \quad (22)$$

$$\mathbb{P} \left[P(Ks_n - Ks_M) \leq (1 + \varepsilon_n) \frac{1}{4} \frac{D}{n} \mathcal{K}_{1,M}^2 \right] \geq 1 - 5n^{-\alpha}, \quad (23)$$

$$\mathbb{P} \left[P_n(Ks_M - Ks_n) \geq (1 - \varepsilon_n^2) \frac{1}{4} \frac{D}{n} \mathcal{K}_{1,M}^2 \right] \geq 1 - 2n^{-\alpha}, \quad (24)$$

$$\mathbb{P} \left[P_n(Ks_M - Ks_n) \leq (1 + \varepsilon_n^2) \frac{1}{4} \frac{D}{n} \mathcal{K}_{1,M}^2 \right] \geq 1 - 3n^{-\alpha}, \quad (25)$$

where $\mathcal{K}_{1,M}^2 = \frac{1}{D} \sum_{k=1}^D \text{Var}(\psi_{1,M} \cdot \varphi_k)$. In addition, when **(H5)** does not hold, but **(H1)**, **(H2)** and **(H4)** hold, we still have for all $n \geq n_0(A_-, A_+, A, r_M(\varphi), \sigma_{\min}, \alpha)$,

$$\mathbb{P} \left(P_n(Ks_M - Ks_n) \geq \left(1 - A_0 \max \left\{ \sqrt{\frac{\ln n}{D}}, \sqrt{\frac{D \ln n}{n}} \right\} \right) \frac{D}{4n} \mathcal{K}_{1,M}^2 \right) \geq 1 - 2n^{-\alpha}. \quad (26)$$

In Theorem 3 above, we achieve sharp upper and lower bounds for the true and empirical excess risks on M . They are optimal at the first order since the leading constants are equal for upper and lower bounds. Moreover, Theorem 3 states the equivalence with high probability of the true and empirical excess risks for models of reasonable dimensions. We notice that second orders are smaller for the empirical excess risk than for the true one. Indeed, when normalized by the first order, the deviations of the empirical excess risk are square of the deviations of the true one. Our bounds also give another evidence of the concentration phenomenon of the empirical excess risk exhibited by Boucheron and Massart [7] in the slightly different context of M-estimation with bounded contrast where some margin condition hold. Notice that considering the lower bound of the empirical excess risk given in (26), we do not need to assume the consistency of the least-squares estimator s_n towards the linear projection s_M .

We turn now to upper bounds in probability for the true and empirical excess risks on models with possibly small dimensions. In this context, we do not achieve sharp or explicit constants in the rates of convergence.

Theorem 4 *Let $\alpha, A_+ > 0$ be fixed and let M be a linear model of finite dimension*

$$1 \leq D \leq A_+ \frac{n}{(\ln n)^2}.$$

*Assume that assumptions **(H1)**, **(H3)** and **(H5)** hold. Then a positive constant A_u exists, only depending on $A, A_{\text{cons}}, A_{3,M}$ and α , such that for all $n \geq n_0(A_{\text{cons}}, n_1)$,*

$$\mathbb{P} \left[P(Ks_n - Ks_M) \geq A_u \frac{D \vee \ln n}{n} \right] \leq 3n^{-\alpha} \quad (27)$$

and

$$\mathbb{P} \left[P_n(Ks_M - Ks_n) \geq A_u \frac{D \vee \ln n}{n} \right] \leq 3n^{-\alpha}. \quad (28)$$

Notice that on contrary to the situation of Theorem 3, we do not assume that **(H2)** hold. This assumption states that the noise level is uniformly bounded away from zero over the space \mathcal{X} , and allows in Theorem 3 to derive lower bounds for the true and empirical excess risks, as well as to achieve sharp constants in the deviation bounds for models of reasonable dimensions. In Theorem 4, we just derive upper bounds and assumption **(H2)** is not needed. The price to pay is that constants in the rates of convergence derived in (27) and (28) are possibly larger than the corresponding ones of Theorem 3, but our results still hold true for small models. Moreover, in the case of models with reasonable dimensions, that is dimensions satisfying assumption (20) of Theorem 3, the rate of decay is preserved compared to Theorem 3 and is proportional to D/n .

The proofs of the above theorems can be found in Section 6.3.

3.3 Some additional comments

Let us first comment on the assumptions given in Section 3.1. Assumptions (11) and **(H2)** are rather mild and can also be found in the work of Arlot and Massart [1] related to the case of histograms, where they are respectively denoted by **(Ab)** and **(An)**. The histogram case will be further commented in Section 4.3.

In assumption **(H4)** we require that the model M is provided with an orthonormal localized basis in $L_2(P^X)$. This property is convenient when dealing with the L_∞ -structure on the model, and this allows us to control the sup-norm of the functions in the model by the sup-norm of the vector of their coordinates in the localized basis. For examples of models with localized basis, and their use in a model selection framework, we refer for instance to Section 7.4.2 of Massart [23], where it is shown that models of histograms, piecewise polynomials and compactly supported wavelets are typical examples of models with localized basis for the $L_2(\text{Leb})$ structure, considering that $\mathcal{X} \subset \mathbb{R}^k$. In Sections 4 and 5, we show that models of piecewise constant and piecewise polynomials respectively can also have a localized basis for the $L_2(P^X)$ structure, under rather mild assumptions on P^X . Assumption **(H4)** is needed in Theorem 3, whereas in Theorem 4 we only use the weaker assumption **(H3)** on the unit envelope of the model M , relating the L_2 -structure of the model to the L_∞ -structure. In fact, assumption **(H4)** allows us in the proof of Theorem 3 to achieve sharp lower bounds for the quantities of interest, whereas in Theorem 4 we only give upper bounds in the case of small models.

We ask in assumption **(H5)** that the M-estimator is consistent towards the linear projection s_M of s_* onto the model M , at a rate at least better than $(\ln n)^{-1/2}$. This can be considered as a rather strong assumption, but it is essential for our methodology. Moreover, we show in Sections 4 and 5 that this assumption is satisfied under mild conditions for histogram models and models of piecewise polynomials respectively, both at the rate

$$R_{n,D,\alpha} \propto \sqrt{\frac{D \ln n}{n}}.$$

Secondly, let us comment on the rates of convergence given in Theorem 3 for models of reasonable dimensions. As we can see in Theorem 3, the rate of estimation in a fixed model M of reasonable dimension is determined at the first order by a key quantity that relates the structure of the model to the unknown law P of data. We call this quantity the **complexity** of the model M and we denote it by \mathcal{C}_M . More precisely, let us define

$$\mathcal{C}_M = \frac{1}{4}D \times \mathcal{K}_{1,M}^2$$

where

$$\mathcal{K}_{1,M} = \sqrt{\frac{1}{D} \sum_{k=1}^D \text{Var}(\psi_{1,M} \cdot \varphi_k)}$$

for a localized orthonormal basis $(\varphi_k)_{k=1}^D$ of $(M, \|\cdot\|_2)$. Notice that $\mathcal{K}_{1,M}$ is well defined as it does not depend on the choice of the basis $(\varphi_k)_{k=1}^D$. Indeed, since we have $P(\psi_{1,M} \cdot \varphi_k) = 0$, we deduce that

$$\mathcal{K}_{1,M}^2 = P\left(\psi_{1,M}^2 \cdot \left(\frac{1}{D} \sum_{k=1}^D \varphi_k^2\right)\right).$$

Now observe that, by using Cauchy-Schwarz inequality in Definition (16), as pointed out by Birgé and Massart [6], we get

$$\Psi_M^2 = \frac{1}{D} \sum_{k=1}^D \varphi_k^2 \quad (29)$$

and so

$$\begin{aligned} \mathcal{K}_{1,M}^2 &= P(\psi_{1,M}^2 \Psi_M^2) \\ &= 4\mathbb{E}\left[\mathbb{E}\left[(Y - s_M(X))^2 | X\right] \Psi_M^2(X)\right] \\ &= 4\left(\mathbb{E}[\sigma^2(X) \Psi_M^2(X)] + \mathbb{E}\left[(s_M - s_*)^2(X) \Psi_M^2(X)\right]\right). \end{aligned} \quad (30)$$

On the one hand, if we assume **(H1)** then we obtain by elementary computations

$$\mathcal{K}_{1,M} \leq 2\sigma_{\max} + 4A \leq 6A . \quad (31)$$

On the other hand, **(H2)** implies

$$\mathcal{K}_{1,M} \geq 2\sigma_{\min} > 0 . \quad (32)$$

To fix ideas, let us explicitly compute $\mathcal{K}_{1,M}^2$ in a simple case. Consider homoscedastic regression on a histogram model M , in which the homoscedastic noise level σ is such that

$$\sigma^2(X) = \sigma^2 \quad a.s. ,$$

so that we have

$$\mathbb{E} [\sigma^2(X) \Psi_M^2(X)] = \sigma^2 \mathbb{E} [\Psi_M^2(X)] = \sigma^2 .$$

Now, under notations of Lemma 5 below,

$$s_M = \sum_{I \in \mathcal{P}} \mathbb{E}[Y \varphi_I(X)] \varphi_I = \sum_{I \in \mathcal{P}} \mathbb{E}[Y | X \in I] \mathbf{1}_I ,$$

thus we deduce, by (29) and the previous equality, that

$$\begin{aligned} \mathbb{E} [(s_M - s_*)^2(X) \Psi_M^2(X)] &= \frac{1}{|\mathcal{P}|} \sum_{I \in \mathcal{P}} \mathbb{E} [(s_M - s_*)^2(X) \varphi_I^2(X)] \\ &= \frac{1}{|\mathcal{P}|} \sum_{I \in \mathcal{P}} \mathbb{E} [(\mathbb{E}[Y | X \in I] - \mathbb{E}[Y | X])^2 \frac{\mathbf{1}_{X \in I}}{P^X(I)}] \\ &= \frac{1}{|\mathcal{P}|} \sum_{I \in \mathcal{P}} \mathbb{E} [(\mathbb{E}[Y | X \in I] - \mathbb{E}[Y | X])^2 | X \in I] \\ &= \frac{1}{|\mathcal{P}|} \sum_{I \in \mathcal{P}} \mathbb{V}[\mathbb{E}[Y | X] | X \in I] , \end{aligned}$$

where the conditional variance $\mathbb{V}[U | \mathcal{A}]$ of a variable U with respect to the event \mathcal{A} is defined to be

$$\mathbb{V}[U | \mathcal{A}] := \mathbb{E} [(U - \mathbb{E}[U | \mathcal{A}])^2 | \mathcal{A}] = \mathbb{E}[U^2 | \mathcal{A}] - (\mathbb{E}[U | \mathcal{A}])^2 .$$

By (30), we explicitly get

$$\mathcal{K}_{1,M}^2 = 4 \left(\sigma^2 + \frac{1}{|\mathcal{P}|} \sum_{I \in \mathcal{P}} \mathbb{V}[\mathbb{E}[Y | X] | X \in I] \right) . \quad (33)$$

A careful look at the proof of Theorem 3 given in Section 6.3 show that condition **(H2)** is only used through the lower bound (32), and thus **(H2)** can be replaced by the following slightly more general assumption :

(H2bis) Lower bound on the normalized complexity $\mathcal{K}_{1,M}$: a positive constant A_{\min} exists such that

$$\mathcal{K}_{1,M} \geq A_{\min} > 0 .$$

When **(H2)** holds, we see from Inequality 32 that **(H2bis)** is satisfied with $A_{\min} = 2\sigma_{\min}$. For suitable models we can have for a positive constant A_{Ψ}^- and for all $x \in \mathcal{X}$,

$$\Psi_M(x) \geq A_{\Psi}^- > 0 , \quad (34)$$

and this allows to consider vanishing noise level, as we then have by (30),

$$\mathcal{K}_{1,M} \geq 2A_{\Psi}^- \sqrt{\mathbb{E}[\sigma^2(X)]} = 2A_{\Psi}^- \|\sigma\|_2 > 0 .$$

As we will see in Sections 4 and 5, Inequality (34) can be satisfied for histogram and piecewise polynomial models on a partition achieving some upper regularity assumption with respect to the law P^X .

4 The histogram case

In this section, we particularize the results stated in Section 3 to the case of piecewise constant functions. We show that under a lower regularity assumption on the considered partition, the assumption **(H4)** of existence of a localized basis in $L_2(P^X)$ and **(H5)** of consistency in sup-norm of the M-estimator towards the linear projection s_M are satisfied.

4.1 Existence of a localized basis

The following lemma states the existence of an orthonormal localized basis for piecewise constant functions in $L_2(P^X)$, on a partition which is lower-regular for the law P^X .

Lemma 5 *Let consider a linear model M of histograms defined on a finite partition \mathcal{P} on \mathcal{X} , and write $|\mathcal{P}| = D$ the dimension of M . Moreover, assume that for a positive finite constant $c_{M,P}$,*

$$\sqrt{|\mathcal{P}| \inf_{I \in \mathcal{P}} P^X(I)} \geq c_{M,P} > 0. \quad (35)$$

Set, for $I \in \mathcal{P}$,

$$\varphi_I = (P^X(I))^{-1/2} \mathbf{1}_I.$$

Then the family $(\varphi_I)_{I \in \Lambda_M}$ is an orthonormal basis in $L_2(P^X)$ and we have,

$$\text{for all } \beta = (\beta_I)_{I \in \mathcal{P}} \in \mathbb{R}^D, \quad \left\| \sum_{I \in \mathcal{P}} \beta_I \varphi_I \right\|_{\infty} \leq c_{M,P}^{-1} \sqrt{D} |\beta|_{\infty}. \quad (36)$$

Condition (35) can also be found in Arlot and Massart [1] and is named lower regularity of the partition \mathcal{P} for the law P^X . It is easy to see that the lower regularity of the partition is equivalent to the property of localized basis in the case of histograms, i.e. (35) is equivalent to (36). The proof of Lemma 5 is straightforward and can be found in Section 6.1.

4.2 Rates of convergence in sup-norm

The following lemma allows to derive property **(H5)** for histogram models.

Lemma 6 *Consider a linear model M of histograms defined on a finite partition \mathcal{P} of \mathcal{X} , and denote by $|\mathcal{P}| = D$ the dimension of M . Assume that Inequality (11) holds, that is, a positive constant A exists such that $|Y| \leq A$ a.s. Moreover, assume that for some positive finite constant $c_{M,P}$,*

$$\sqrt{|\mathcal{P}| \inf_{I \in \mathcal{P}} P^X(I)} \geq c_{M,P} > 0 \quad (37)$$

and that $D \leq A_+ n (\ln n)^{-2} \leq n$ for some positive finite constant A_+ . Then, for any $\alpha > 0$ and for all $n \geq n_0(\alpha, c_{M,P}, A_+)$, there exists an event of probability at least $1 - n^{-\alpha}$ on which s_n exists, is unique and it holds,

$$\|s_n - s_M\|_{\infty} \leq L_{A_+, A, c_{M,P}, \alpha} \sqrt{\frac{D \ln n}{n}}. \quad (38)$$

In Lemma 6 we thus achieve the convergence in sup-norm of the regressogram s_n towards the linear projection s_M at the rate $\sqrt{D \ln(n)/n}$. It is worth noticing that for a model of histograms satisfying the assumptions of Lemma 6, if we set

$$A_{cons} = L_{A, c_{M,P}, \alpha} \sqrt{A_+}, \quad n_1 = n_0(\alpha, c_{M,P}, A_+) \quad \text{and} \quad R_{n,D,\alpha} = L_{A_+, A, c_{M,P}, \alpha} \sqrt{\frac{D \ln n}{n}},$$

then Assumption **(H5)** is satisfied. To derive Inequality (38), we need to assume that the response variable Y is almost surely bounded and that the considered partition is lower-regular for the law P^X . Hence, we fit again with the framework of [1] and we can thus view the general set of assumptions exposed in Section 3.1 as a natural generalization for linear models of the framework developed in [1] in the case of histograms. The proof of Lemma 6 can be found in Section 6.1.

4.3 Bounds for the excess risks

The next theorem is a straightforward application of Lemmas 5, 6 and Theorems 3, 4. Indeed, we recover results of Theorems 3 and 4 for models of histograms, under the lower regularity assumption on the considered partition of the space \mathcal{X} with respect to the unknown law P^X . As seen in Section 4.2, we have in that case

$$R_{n,D,\alpha} \propto \sqrt{\frac{D \ln n}{n}}.$$

Theorem 7 *Given $A_+, A_-, \alpha > 0$, consider a linear model M of histograms defined on a finite partition \mathcal{P} of \mathcal{X} , and write $|\mathcal{P}| = D$ the dimension of M . Assume that for some positive finite constant $c_{M,P}$, it holds*

$$\sqrt{|\mathcal{P}| \inf_{I \in \mathcal{P}} P^X(I)} \geq c_{M,P} > 0. \quad (39)$$

If **(H1)** and **(H2)** of Section 3.1 are satisfied and if

$$A_- (\ln n)^2 \leq D \leq A_+ \frac{n}{(\ln n)^2},$$

then there exists a positive finite constant A_0 , only depending on $\alpha, A, \sigma_{\min}, A_-, A_+, c_{M,P}$ such that, by setting

$$\varepsilon_n = A_0 \max \left\{ \left(\frac{\ln n}{D} \right)^{1/4}, \left(\frac{D \ln n}{n} \right)^{1/4} \right\}$$

we have, for all $n \geq n_0(A_-, A_+, A, \sigma_{\min}, c_{M,P}, \alpha)$,

$$\mathbb{P} \left[(1 + \varepsilon_n) \frac{1}{4} \frac{D}{n} \mathcal{K}_{1,M}^2 \geq P(Ks_n - Ks_M) \geq (1 - \varepsilon_n) \frac{1}{4} \frac{D}{n} \mathcal{K}_{1,M}^2 \right] \geq 1 - 10n^{-\alpha} \quad (40)$$

and

$$\mathbb{P} \left[(1 + \varepsilon_n^2) \frac{1}{4} \frac{D}{n} \mathcal{K}_{1,M}^2 \geq P_n(Ks_M - Ks_n) \geq (1 - \varepsilon_n^2) \frac{1}{4} \frac{D}{n} \mathcal{K}_{1,M}^2 \right] \geq 1 - 5n^{-\alpha}. \quad (41)$$

If (39) holds together with **(H1)** and if we assume that

$$1 \leq D \leq A_+ \frac{n}{(\ln n)^2},$$

then a positive constant A_u exists, only depending on $A, c_{M,P}, A_+$ and α , such that for all $n \geq n_0(A, c_{M,P}, A_+, \alpha)$,

$$\mathbb{P} \left[P(Ks_n - Ks_M) \geq A_u \frac{D \vee \ln n}{n} \right] \leq 3n^{-\alpha}$$

and

$$\mathbb{P} \left[P_n(Ks_M - Ks_n) \geq A_u \frac{D \vee \ln n}{n} \right] \leq 3n^{-\alpha}.$$

As announced before, we recover in Theorem 7 the general results of Section 3.2 for the case of histograms on a lower-regular partition. Moreover, in the case of histograms, assumption (12) which is part of **(H1)** is a straightforward consequence of (11). Indeed, we easily see that the projection s_M of the regression function s_* onto the model of piecewise constant functions with respect to \mathcal{P} can be written

$$s_M = \sum_{I \in \mathcal{P}} \mathbb{E}[Y | X \in I] \mathbf{1}_I. \quad (42)$$

Under (11), we have $|\mathbb{E}[Y | X \in I]| \leq \|Y\|_\infty \leq A$ for every $I \in \mathcal{P}$ and we deduce by (42) that $\|s_M\|_\infty \leq A$.

4.4 Comments

Our bounds in Theorem 7 are obtained by following a general methodology that consists, among other things, in expanding the contrast and to take advantage of explicit computations that can be derived on the linear part of the contrast - for more details, see the proofs in Section 6.3 below. It is then instructive to compare them to the best available results in this special case. Let us compare them to the bounds obtained by Arlot and Massart in [1], in the case of a fixed model. Such results can be found in Proposition 10, 11 and 12 of [1].

The strategy adopted by the authors in this case is as follows. By remarking that easy bounds are available for the mean of the empirical excess risk on histograms since it holds

$$\mathbb{E}[P_n(Ks_M - Ks_n)] = \frac{D}{4n} \kappa_{1,M}^2,$$

they derive concentration inequalities for the true excess risk and its empirical counterpart to their mean. They further give upper and lower bounds in terms of $\mathbb{E}[P_n(Ks_M - Ks_n)]$ for the mean of the true excess risk. The deviations in all these inequalities are made of sums of quantities that can not be compared to ours in a concise manner, as some of them are loose compared to our results and some of them gain.

Nevertheless, using our notations, Inequality (34) of Proposition 10 in [1] states that for every $x \geq 0$ there exists an event of probability at least $1 - e^{1-x}$ on which,

$$\begin{aligned} & |P_n(Ks_M - Ks_n) - \mathbb{E}[P_n(Ks_M - Ks_n)]| \\ & \leq \frac{L}{\sqrt{D_M}} \left[P(Ks_M - Ks_*) + \frac{A^2 \mathbb{E}[P_n(Ks_M - Ks_n)]}{\sigma_{\min}^2} (\sqrt{x} + x) \right], \end{aligned} \quad (43)$$

for some absolute constant L . We can notice that Inequality (43), which is a special case of general concentration inequalities given by Boucheron and Massart [7], involves the bias of the model $P(Ks_M - Ks_*)$. By pointing out that the bias term arises from the use of some margin conditions that are satisfied for bounded regression, we believe that it can be removed from Proposition 10 of [1], since in the case of histograms models for bounded regression, some margin-like conditions hold, that are directly pointed at the linear projection s_M . Apart from the bias term, the deviations of the empirical excess risk are then of the order

$$\frac{\ln(n) \sqrt{D_M}}{n},$$

considering the same probability of event than ours, so it becomes significantly better than Inequality (41) for large models.

Concentration inequalities for the true excess risk given in Proposition 11 of [1] give a magnitude of deviations that is again smaller than ours for sufficiently large models and that is in fact closer to ε_n^2 than ε_n , where ε_n is defined in Theorem 7. But the mean of the true excess risk has to be compared to the mean of the empirical excess risk and it is remarkable that in Proposition 12 of [1] where such a result is given in a way that seems very sharp, there is a term lower bounded by

$$\left(n \times \inf_{I \in \mathcal{P}} P^X(I) \right)^{-1/4} \propto \left(\frac{D}{n} \right)^{1/4},$$

due to the lower regularity assumption on the partition. This allows us to conjecture that up to a logarithmic factor, the term proportional to $\left(\frac{D \ln n}{n} \right)^{1/4}$ appearing in ε_n and also in the deviations of the true excess risk in Theorem 3 is not improvable in general, and that the empirical excess risk concentrates better around its mean than the true excess risk in general.

We can conclude that the bounds given in Proposition 10, 11 and 12 of [1] are better than ours, apart from the bias term involved in concentration inequalities of Proposition 10, but this term could be removed as explained above. Furthermore, concentration inequalities for the empirical excess risk are significantly better than ours for large models.

Arlot and Massart [1] also propose generalizations in the case of unbounded noise and when the noise level vanishes. The unbounded case seems to be beyond the reach of our strategy, due to our repeated use of Bousquet and Klein-Rio's inequalities along the proofs. However, we recover the case of vanishing noise level

for histogram models, when the partition is upper regular with respect to the law P^X , a condition also needed in [1] in this case. Indeed, we have noticed in Section 3.3 that assumption **(H2)** can be weakened by **(H2bis)** where we assume that

$$\mathcal{K}_{1,M} \geq A_{\min} > 0$$

for some positive constant A_{\min} . So, if we assume the upper regularity of the partition \mathcal{P} with respect to P^X , that is

$$|\mathcal{P}| \sup_{I \in \mathcal{P}} P^X(I) \leq c_{M,P}^+ < +\infty \quad (44)$$

for a positive constant $c_{M,P}^+$, we then have from identity (30)

$$\mathcal{K}_{1,M}^2 \geq 4\mathbb{E} [\sigma^2(X) \Psi_M^2(X)] ,$$

where from identity (29), we have in the case of histograms,

$$\Psi_M^2(x) = \frac{1}{|\mathcal{P}|} \sum_{I \in \mathcal{P}} \frac{\mathbf{1}_{x \in I}}{P^X(I)} , \text{ for all } x \in \mathcal{X} .$$

Now from inequality (44) we have

$$\Psi_M^2(x) \geq \left(c_{M,P}^+\right)^{-1} > 0 , \text{ for all } x \in \mathcal{X} ,$$

and so $A_{\min} = 2 \left(c_{M,P}^+\right)^{-1/2} \|\sigma\|_2 > 0$ is convenient in **(H2bis)**.

5 The case of piecewise polynomials

In this Section, we generalize the results given in Section 4 for models of piecewise constant functions to models of piecewise polynomials uniformly bounded in their degree.

5.1 Existence of a localized basis

The following lemma states the existence of a localized orthonormal basis in $(M, \|\cdot\|_2)$ where M is a model of piecewise polynomials and $\mathcal{X} = [0, 1]$ is the unit interval.

Lemma 8 *Let Leb denote the Lebesgue measure on $[0, 1]$. Let assume that $\mathcal{X} = [0, 1]$ and that P^X has a density f with respect to Leb satisfying, for a positive constant c_{\min} ,*

$$f(x) \geq c_{\min} > 0, \quad x \in [0, 1] .$$

Consider a linear model M of piecewise polynomials on $[0, 1]$ with degree r or smaller, defined on a finite partition \mathcal{P} made of intervals. Then there exists an orthonormal basis $\{\varphi_{I,j}, I \in \mathcal{P}, j \in \{0, \dots, r\}\}$ of $(M, \|\cdot\|_2)$ such that,

$$\text{for all } j \in \{0, \dots, r\} \quad \varphi_{I,j} \text{ is supported by the element } I \text{ of } \mathcal{P},$$

and a constant $L_{r,c_{\min}}$ depending only on r, c_{\min} exists, satisfying for all $I \in \mathcal{P}$,

$$\max_{j \in \{0, \dots, r\}} \|\varphi_{I,j}\|_{\infty} \leq L_{r,c_{\min}} \frac{1}{\sqrt{\text{Leb}(I)}} . \quad (45)$$

As a consequence, if it holds

$$\sqrt{|\mathcal{P}| \inf_{I \in \mathcal{P}} \text{Leb}(I)} \geq c_{M,\text{Leb}} \quad (46)$$

a constant $L_{r,c_{\min},c_{M,\text{Leb}}}$ depending only on r, c_{\min} and $c_{M,\text{Leb}}$ exists, such that for all $\beta = (\beta_{I,j})_{I \in \mathcal{P}, j \in \{0, \dots, r\}} \in \mathbb{R}^D$,

$$\left\| \sum_{I,j} \beta_{I,j} \varphi_{I,j} \right\|_{\infty} \leq L_{r,c_{\min},c_{M,\text{Leb}}} \sqrt{D} |\beta|_{\infty} \quad (47)$$

where $D = (r+1) |\mathcal{P}|$ is the dimension of M .

Lemma 8 states that if $\mathcal{X} = [0, 1]$ is the unit interval and P^X has a density with respect to the Lebesgue measure Leb on \mathcal{X} uniformly bounded away from zero, then there exists an orthonormal basis in $L_2(P^X)$ of piecewise polynomials where the sup-norm of its elements are suitably controlled by (45). Moreover, if we assume the lower regularity of the partition with respect to Leb then the orthonormal basis is localized, where the constant of localization in (47) depend on the maximal degree r . We notice that in the case of piecewise constant functions we do not need to assume the existence of a density for P^X or to restrict ourselves to the unit interval. The proof of Lemma 8 can be found in Section 6.2.

5.2 Rates of convergence in sup-norm

The following lemma allows to derive property **(H5)** for piecewise polynomials.

Lemma 9 *Assume that Inequality (11) holds, that is a positive constant A exists such that $|Y| \leq A$ a.s. Denote by Leb the Lebesgue measure on $[0, 1]$. Assume that $\mathcal{X} = [0, 1]$ and that P^X has a density f with respect to Leb , satisfying for positive constants c_{\min} and c_{\max} ,*

$$0 < c_{\min} \leq f(x) \leq c_{\max} < +\infty, \quad x \in [0, 1] . \quad (48)$$

Consider a linear model M of piecewise polynomials on $[0, 1]$ with degree less than r , defined on a finite partition \mathcal{P} made of intervals, that satisfies for some finite positive constants $c_{M,\text{Leb}}$

$$\sqrt{|\mathcal{P}| \inf_{I \in \mathcal{P}} \text{Leb}(I)} \geq c_{M,\text{Leb}} > 0 . \quad (49)$$

Assume moreover that $D \leq A_+ n (\ln n)^{-2}$ for a positive finite constant A_+ . Then, for any $\alpha > 0$, there exists an event of probability at least $1 - n^{-\alpha}$ such that s_n exists, is unique on this event and it holds, for all $n \geq n_0(r, A_+, c_{\min}, c_{M,\text{Leb}}, \alpha)$,

$$\|s_n - s_M\|_{\infty} \leq L_{A,r,A_+,c_{\min},c_{\max},c_{M,\text{Leb}},\alpha} \sqrt{\frac{D \ln n}{n}} . \quad (50)$$

In Lemma 6 we thus obtain the convergence in sup-norm of the M-estimator s_n towards the linear projection s_M at the rate $\sqrt{\frac{D \ln n}{n}}$. It is worth noticing that for a model of piecewise polynomials satisfying the assumptions of Lemma 6, if we set

$$A_{\text{cons}} = L_{A,r,A_+,c_{\min},c_{\max},c_{M,\text{Leb}},\alpha} \sqrt{A_+} , \quad R_{n,D,\alpha} = L_{A,r,A_+,c_{\min},c_{\max},c_{M,\text{Leb}},\alpha} \sqrt{\frac{D \ln n}{n}} , \\ n_1 = n_0(r, A_+, c_{\min}, c_{M,\text{Leb}}, \alpha)$$

then Assumption **(H5)** is satisfied. To derive Inequality (38), we need to assume that the response variable Y is almost surely bounded, we give the conditions to ensure that the model is provided with a localized basis and also we assume that the density of P^X with respect to the Lebesgue measure on the unit interval is uniformly bounded from above. The proof of Lemma 9 can be found in Section 6.2.

5.3 Bounds for the excess risks

The forthcoming result is a straightforward application of Lemmas 8, 9 and Theorems 3, 4.

Theorem 10 *Denote by Leb the Lebesgue measure on $[0, 1]$ and fix some positive finite constant α . Assume that $\mathcal{X} = [0, 1]$ and that P^X has a density f with respect to Leb satisfying, for some positive finite constants c_{\min} and c_{\max} ,*

$$0 < c_{\min} \leq f(x) \leq c_{\max} < +\infty, \quad x \in [0, 1] . \quad (51)$$

Consider a linear model M of piecewise polynomials on $[0, 1]$ with degree less than r , defined on a finite partition \mathcal{P} made of intervals, that satisfy for a finite constant $c_{M,\text{Leb}}$,

$$\sqrt{|\mathcal{P}| \inf_{I \in \mathcal{P}} \text{Leb}(I)} \geq c_{M,\text{Leb}} > 0 . \quad (52)$$

Assume that **(H1)** and **(H2)** hold. Then, if there exist some positive finite constants A_- and A_+ such that

$$A_- (\ln n)^2 \leq D \leq A_+ \frac{n}{(\ln n)^2} ,$$

then there exists a positive finite constant A_0 , depending on $\alpha, A, \sigma_{\min}, A_-, A_+, r, c_{M,\text{Leb}}, c_{\min}$ and c_{\max} such that, by setting

$$\varepsilon_n = A_0 \max \left\{ \left(\frac{\ln n}{D} \right)^{1/4}, \left(\frac{D \ln n}{n} \right)^{1/4} \right\}$$

we have, for all $n \geq n_0(A_-, A_+, A, r, \sigma_{\min}, c_{M,\text{Leb}}, c_{\min}, c_{\max}, \alpha)$,

$$\mathbb{P} \left[(1 + \varepsilon_n) \frac{1}{4} \frac{D}{n} \mathcal{K}_{1,M}^2 \geq P(Ks_n - Ks_M) \geq (1 - \varepsilon_n) \frac{1}{4} \frac{D}{n} \mathcal{K}_{1,M}^2 \right] \geq 1 - 10n^{-\alpha}$$

and

$$\mathbb{P} \left[(1 + \varepsilon_n^2) \frac{1}{4} \frac{D}{n} \mathcal{K}_{1,M}^2 \geq P_n(Ks_M - Ks_n) \geq (1 - \varepsilon_n^2) \frac{1}{4} \frac{D}{n} \mathcal{K}_{1,M}^2 \right] \geq 1 - 5n^{-\alpha} .$$

Moreover, if (51) and (52) hold together with **(H1)** and if we assume that

$$1 \leq D \leq A_+ \frac{n}{(\ln n)^2} ,$$

then a positive constant A_u exists, only depending on $A_+, A, r, c_{M,\text{Leb}}, c_{\min}$ and α , such that for all $n \geq n_0(A_+, A, r, c_{\min}, c_{\max}, c_{M,\text{Leb}}, \alpha)$,

$$\mathbb{P} \left[P(Ks_n - Ks_M) \geq A_u \frac{D \vee \ln n}{n} \right] \leq 3n^{-\alpha}$$

and

$$\mathbb{P} \left[P_n(Ks_M - Ks_n) \geq A_u \frac{D \vee \ln n}{n} \right] \leq 3n^{-\alpha} .$$

We derive in Theorem 10 optimal upper and lower bounds for the excess risk and its empirical counterpart in the case of models of piecewise polynomials uniformly bounded in their degree with reasonable dimension. We give also upper bounds for models of possibly small dimension, without assumption **(H2)**. Notice that we need stronger assumptions than in the case of histograms. Namely, we require the existence of a density uniformly bounded from above and from below for the unknown law P^X , with respect to the Lebesgue measure on the unit interval. However we recover the bounds of Theorem 7 yet with different constants, since by Lemma 9 we still have $R_{n,D,\alpha} \propto \sqrt{\frac{D \ln n}{n}}$ as in the histogram case.

Moreover, as in the case of histograms, assumption (12) which is part of **(H1)** is a straightforward consequence of (11). Indeed, we easily see that the projection s_M of the regression function s_* onto the model of piecewise polynomials with respect to \mathcal{P} can be written

$$s_M = \sum_{(I,j) \in \mathcal{P} \times \{0, \dots, r\}} P(Y \varphi_{I,j}) \varphi_{I,j}$$

where $\varphi_{I,j}$ is the orthonormal basis given in Lemma 8. It is then easy to show, using (45) of Lemma 8 and (11), that $\|s_M\|_\infty \leq L_{A,r,c_{\min},c_{\max}}$.

Again, we can consider vanishing noise at the prize to ask that the partition is upper regular with respect to Leb. By **(H2bis)** of Section 3.3, if we show that

$$\mathcal{K}_{1,M} \geq A_{\min} > 0$$

for a positive constant A_{\min} instead of **(H2)**, then the conclusions of Theorem 10 still hold. Now, from identity (30) we have

$$\mathcal{K}_{1,M}^2 \geq 4\mathbb{E} [\sigma^2(X) \Psi_M^2(X)]$$

where from identity (29), it holds in the case of piecewise polynomials, for all $x \in \mathcal{X}$,

$$\Psi_M^2(x) = \frac{1}{(r+1)|\mathcal{P}|} \sum_{(I,j) \in \mathcal{P} \times \{0, \dots, r\}} \varphi_{I,j}^2 \geq \frac{1}{(r+1)|\mathcal{P}|} \sum_{I \in \mathcal{P}} \frac{\mathbf{1}_{x \in I}}{P^X(I)} . \quad (53)$$

Furthermore, if we ask that

$$|\mathcal{P}| \sup_{I \in \mathcal{P}} \text{Leb}(I) \leq c_{M,P}^+ < +\infty \quad (54)$$

for a positive constant $c_{M,P}^+$, then by using (51), (53) and (54), we obtain for all $x \in \mathcal{X}$,

$$\Psi_M^2(x) \geq \left(c_{\max} \times c_{M,P}^+ \times (r+1) \right)^{-1} > 0 ,$$

and so $A_{\min} = 2 \left(c_{\max} \times c_{M,P}^+ \times (r+1) \right)^{-1/2} \sqrt{\mathbb{E}[\sigma^2(X)]} > 0$ is convenient in **(H2bis)**.

6 Proofs

We begin with the simpler proofs of Sections 4 and 5, in Sections 6.1 and 6.2 respectively. The proofs of Theorems 3 and 4 of Section 3.2 can be found in Section 6.3.

6.1 Proofs of Section 4

Proof of Lemma 5. It suffices to observe that

$$\begin{aligned} \left\| \sum_{I \in \mathcal{P}} \beta_I \varphi_I \right\|_{\infty} &\leq |\beta|_{\infty} \sup_{I \in \mathcal{P}} \|\varphi_I\|_{\infty} \\ &= |\beta|_{\infty} \sup_{I \in \mathcal{P}} (P^X(I))^{-1/2} \\ &\leq c_{M,P}^{-1} \sqrt{D} |\beta|_{\infty} . \end{aligned}$$

■

We now intend to prove (38) under the assumptions of Lemma 6.

Proof of Lemma 6. Along the proof, we denote by misuse of notation, for any $I \in \mathcal{P}$,

$$P(I) := P(I \times \mathbb{R}) = P^X(I) \text{ and } P_n(I) := P_n(I \times \mathbb{R}) .$$

Let $\alpha > 0$ be fixed and let $\beta > 0$ to be chosen later. We first show that, since we have $D \leq A_+ n (\ln n)^{-2}$, it holds with large probability and for all n sufficiently large,

$$\inf_{I \in \mathcal{P}} P_n(I) > 0 .$$

Since

$$\|\mathbf{1}_I\|_{\infty} \leq 1 \quad \text{and} \quad \mathbb{E}[\mathbf{1}_I^2] = P(I)$$

we get by Bernstein's inequality (213), for any $x > 0$ and $I \in \mathcal{P}$,

$$\mathbb{P} \left[|(P_n - P)(I)| \geq \sqrt{\frac{2P(I)x}{n}} + \frac{x}{3n} \right] \leq 2 \exp(-x) . \quad (55)$$

Further note that by (37), $D \geq c_{M,P}^2 P(I)^{-1} > 0$ for any $I \in \mathcal{P}$, and thus by taking $x = \beta \ln n$, we easily deduce from inequality (55) that there exists a positive constant $L_{\beta, c_{M,P}}^{(1)}$ only depending on $c_{M,P}$ and β such that, for any $I \in \mathcal{P}$,

$$\mathbb{P} \left[\frac{|(P_n - P)(I)|}{P(I)} \geq L_{\beta, c_{M,P}}^{(1)} \sqrt{\frac{D \ln n}{n}} \right] \leq 2n^{-\beta} . \quad (56)$$

Now, as $D \leq A_+ n (\ln n)^{-2}$ for some positive constant A_+ , a positive integer $n_0(\beta, c_{M,P}, A_+)$ exists such that

$$L_{\beta, c_{M,P}}^{(1)} \sqrt{\frac{D \ln n}{n}} \leq \frac{1}{2}, \text{ for all } n \geq n_0(\beta, c_{M,P}, A_+) . \quad (57)$$

Therefore we get, for all $n \geq n_0(\beta, c_{M,P}, A_+)$,

$$\begin{aligned} & \mathbb{P}[\forall I \in \mathcal{P}, P_n(I) > 0] \\ & \geq \mathbb{P} \left[\forall I \in \mathcal{P}, \frac{P(I)}{2} > |(P_n - P)(I)| \right] \\ & \geq \mathbb{P} \left[\forall I \in \mathcal{P}, \frac{|(P_n - P)(I)|}{P(I)} < L_{\beta, c_{M,P}}^{(1)} \sqrt{\frac{D \ln n}{n}} \right] \quad \text{by (57)} \\ & \geq 1 - 2Dn^{-\beta} . \end{aligned}$$

Introduce the event

$$\Omega_+ = \{\forall I \in \mathcal{P}, P_n(I) > 0\} .$$

We have shown that

$$\mathbb{P}[\Omega_+] \geq 1 - 2Dn^{-\beta} . \quad (58)$$

Moreover, on the event Ω_+ , the least-squares estimator s_n exists, is unique and it holds

$$s_n = \sum_{I \in \mathcal{P}} \frac{P_n(y \mathbf{1}_{x \in I})}{P_n(I)} \mathbf{1}_I .$$

We also have

$$s_M = \sum_{I \in \mathcal{P}} \frac{P(y \mathbf{1}_{x \in I})}{P(I)} \mathbf{1}_I .$$

Hence it holds on Ω_+ ,

$$\begin{aligned} \|s_n - s_M\|_\infty &= \sup_{I \in \mathcal{P}} \left| \frac{P_n(y \mathbf{1}_{x \in I})}{P_n(I)} - \frac{P(y \mathbf{1}_{x \in I})}{P(I)} \right| \\ &= \sup_{I \in \mathcal{P}} \left| \frac{P_n(y \mathbf{1}_{x \in I})}{P(I) \left(1 + \frac{(P_n - P)(I)}{P(I)}\right)} - \frac{P(y \mathbf{1}_{x \in I})}{P(I)} \right| \\ &\leq \sup_{I \in \mathcal{P}} \left| \frac{(P_n - P)(y \mathbf{1}_{x \in I})}{P(I) \left(1 + \frac{(P_n - P)(I)}{P(I)}\right)} \right| \\ &\quad + \sup_{I \in \mathcal{P}} \left| \frac{P(y \mathbf{1}_{x \in I})}{P(I)} \right| \times \sup_{I \in \mathcal{P}} \left| 1 - \frac{1}{1 + \frac{(P_n - P)(I)}{P(I)}} \right| . \end{aligned} \quad (59)$$

Moreover, by Bernstein's inequality (213), as

$$\|y \mathbf{1}_{x \in I}\|_\infty \leq A \quad \text{and} \quad \mathbb{E}[(Y \mathbf{1}_{X \in I})^2] \leq A^2 P(I)$$

we get for all $I \in \mathcal{P}$,

$$\mathbb{P} \left[|(P_n - P)(y \mathbf{1}_{x \in I})| \geq \sqrt{\frac{2A^2 P(I) x}{n}} + \frac{Ax}{3n} \right] \leq 2 \exp(-x) .$$

By putting $x = \beta \ln n$ in the latter inequality and using the fact that $D \geq c_{M,P}^2 P(I)^{-1}$ it follows that there exists a positive constant $L_{A,c_{M,P},\beta}^{(2)}$ only depending on A , $c_{M,P}$ and β such that

$$\mathbb{P} \left[\frac{|(P_n - P)(y \mathbf{1}_{x \in I})|}{P(I)} \geq L_{A,c_{M,P},\beta}^{(2)} \sqrt{\frac{D \ln n}{n}} \right] \leq 2n^{-\beta} . \quad (60)$$

Now define

$$\Omega_{1,2} = \bigcap_{I \in \mathcal{P}} \left\{ \left\{ \frac{|(P_n - P)(I)|}{P(I)} < L_{\beta,c_{M,P}}^{(1)} \sqrt{\frac{D \ln n}{n}} \right\} \cap \left\{ \frac{|(P_n - P)(y \mathbf{1}_{x \in I})|}{P(I)} < L_{A,c_{M,P},\beta}^{(2)} \sqrt{\frac{D \ln n}{n}} \right\} \right\} .$$

Clearly, since $D \leq n$ we have, by (56) and (60),

$$\mathbb{P} [\Omega_{1,2}^c] \leq 4n^{-\beta+1} . \quad (61)$$

Moreover, for all $n \geq n_0(\beta, c_{M,P}, A_+)$, we get by (57) that

$$\frac{|(P_n - P)(I)|}{P(I)} < \frac{1}{2}$$

on the event $\Omega_{1,2}$, and so, for all $n \geq n_0(\beta, c_{M,P}, A_+)$, $\Omega_{1,2} \subset \Omega_+$. Hence, we get that

$$\begin{aligned} & \sup_{I \in \mathcal{P}} \left| \frac{(P_n - P)(y \mathbf{1}_{x \in I})}{P(I) \left(1 + \frac{(P_n - P)(I)}{P(I)}\right)} \right| + \sup_{I \in \mathcal{P}} \left| \frac{P(y \mathbf{1}_{x \in I})}{P(I)} \right| \times \sup_{I \in \mathcal{P}} \left| 1 - \frac{1}{1 + \frac{(P_n - P)(I)}{P(I)}} \right| \\ & \leq 2 \sup_{I \in \mathcal{P}} \left| \frac{(P_n - P)(y \mathbf{1}_{x \in I})}{P(I)} \right| + 2 \sup_{I \in \mathcal{P}} \left| \frac{P(y \mathbf{1}_{x \in I})}{P(I)} \right| \times \sup_{I \in \mathcal{P}} \left| \frac{(P_n - P)(I)}{P(I)} \right| \\ & \leq 2L_{A,c_{M,P},\beta}^{(2)} \sqrt{\frac{D \ln n}{n}} + 2L_{\beta,c_{M,P}}^{(1)} \sqrt{\frac{D \ln n}{n}} \times \sup_{I \in \mathcal{P}} \left| \frac{P(y \mathbf{1}_{x \in I})}{P(I)} \right| . \end{aligned} \quad (62)$$

Finally we have, for any $I \in \mathcal{P}$,

$$|P(y \mathbf{1}_{x \in I})| \leq P(|y| \mathbf{1}_{x \in I}) \leq AP(I) , \quad (63)$$

so by (59), (62) and (63) we finally get, on the event $\Omega_{1,2}$ and for all $n \geq n_0(\beta, c_{M,P}, A_+)$,

$$\|s_n - s_M\|_\infty \leq \left(2L_{A,c_{M,P},\beta}^{(2)} + 2AL_{\beta,c_{M,P}}^{(1)} \right) \sqrt{\frac{D \ln n}{n}} .$$

Taking $\beta = \alpha + 3$, we get by (61) for all $n \geq 2$, $\mathbb{P} [\Omega_{1,2}^c] \leq n^{-\alpha}$ which implies (38). ■

6.2 Proofs of Section 5

Under the assumptions of Lemma 8, we intend to establish (47).

Proof of Lemma 8. Let I be any interval of $[0, 1]$ and w a positive measurable function on I . Denote by $L_2(I, \text{Leb})$ the space of square integrable functions on I with respect to the Lebesgue measure Leb and set

$$L_2(I, w) = \{g : I \longrightarrow \mathbb{R} ; g\sqrt{w} \in L_2(I, \text{Leb})\} .$$

This space is equipped with the natural inner product

$$\langle g, h \rangle_{I, w} = \int_{x \in I} g(x) h(x) w(x) dx .$$

Write $\|\cdot\|_{I, w}$ its associated norm.

Now, consider an interval I of \mathcal{P} with bounds a and b , $a < b$. Also denote by $f|_I : x \in I \longmapsto f(x)$ the restriction of the density f to the interval I . We readily have for $g, h \in L_2(I, f|_I)$,

$$\begin{aligned} & \int_{x \in I} g(x) h(x) f|_I(x) \frac{dx}{\text{Leb}(I)} \\ &= \int_{y \in [0, 1]} g((b-a)y + a) h((b-a)y + a) f|_I((b-a)y + a) dy . \end{aligned} \quad (64)$$

Define the function f^I from $[0, 1]$ to \mathbb{R}_+ by

$$f^I(y) = f|_I((b-a)y + a), \quad y \in [0, 1] .$$

If $(p_{I,0}, p_{I,1}, \dots, p_{I,r})$ is an orthonormal family of polynomials in $L_2([0, 1], f^I)$ then by setting, for all $x \in I$, $j \in \{0, \dots, r\}$,

$$\tilde{\varphi}_{I,j}(x) = p_{I,j} \left(\frac{x-a}{b-a} \right) \frac{1}{\sqrt{\text{Leb}(I)}} ,$$

we deduce from equality (64) that $(\tilde{\varphi}_{I,j})_{j=0}^r$ is an orthonormal family of polynomials in $L_2(I, f|_I)$ such that $\deg(\tilde{\varphi}_{I,j}) = \deg(p_{I,j})$.

Now, it is a classical fact of orthogonal polynomials theory (see for example Theorems 1.11 and 1.12 of [10]) that there exists a unique family $(q_{I,0}, q_{I,1}, \dots, q_{I,r})$ of orthogonal polynomials on $[0, 1]$ such that $\deg(q_{I,j}) = j$ and the coefficient of the highest monomial x^j of $q_{I,j}$ is equal to 1. Moreover, each $q_{I,j}$ has j distinct real roots belonging to $]0, 1[$. Thus, we can write

$$q_{I,j}(x) = \prod_{k=1}^j (x - \alpha_{I,j}^k), \quad \alpha_{I,j}^k \in]0, 1[\text{ and } \alpha_{I,j}^k \neq \alpha_{I,j}^l \text{ for } k \neq l . \quad (65)$$

Clearly, $\|q_{I,j}\|_\infty \leq 1$. Moreover,

$$\begin{aligned} \|q_{I,j}\|_{[0,1], f^I}^2 &= \int_{[0,1]} (q_{I,j})^2 f^I dx \\ &\geq c_{\min} \int_{[0,1]} (q_{I,j})^2 dx . \end{aligned}$$

Now we set $B(\alpha, r) =]\alpha - r, \alpha + r[$ for $\alpha \in \mathbb{R}$, so that by (65) we get

$$\forall x \in [0, 1] \setminus \cup_{k=1}^j B(\alpha_{I,j}^k, (4j)^{-1}), \quad |q_{I,j}(x)| \geq (4j)^{-j} ,$$

and

$$\text{Leb}([0, 1] \setminus \cup_{k=1}^j B(\alpha_{I,j}^k, (4j)^{-1})) \geq \frac{1}{2} .$$

Therefore,

$$\begin{aligned}
\|q_{I,j}\|_{[0,1],f_I}^2 &\geq c_{\min} \int_{[0,1]} (q_{I,j})^2 dx \\
&\geq c_{\min} \int_{[0,1] \setminus \cup_{k=1}^j B(\alpha_{I,j}^k, (4j)^{-1})} (q_{I,j})^2 dx \\
&\geq \frac{c_{\min}}{2} (4j)^{-2j} .
\end{aligned}$$

Finally, introduce $p_{I,j} = \|q_{I,j}\|_{[0,1],f_I}^{-1} q_{I,j}$ and denote by $\varphi_{I,j}$ its associated orthonormal family of $L_2(I, f_I)$. Then, by considering the extension $\varphi_{I,j}$ of $\tilde{\varphi}_{I,j}$ to $[0, 1]$ by adding null values, it is readily checked that the family

$$\{\varphi_{I,j}, I \in \mathcal{P}, j \in \{0, \dots, r\}\}$$

is an orthonormal basis of $(M, \|\cdot\|_2)$. In addition,

$$\begin{aligned}
\|\varphi_{I,j}\|_{\infty} &= \|\tilde{\varphi}_{I,j}\|_{\infty} \\
&= \|q_{I,j}\|_{[0,1],f_I}^{-1} \|q_{I,j}\|_{\infty} \text{Leb}(I)^{-1/2} \\
&\leq \sqrt{2} c_{\min}^{-1/2} (4r)^r \text{Leb}(I)^{-1/2} \tag{66} \\
&\leq \sqrt{2} c_{M,\text{Leb}}^{-1} c_{\min}^{-1/2} (4r)^r (r+1)^{-1/2} \sqrt{D} \tag{67}
\end{aligned}$$

where in the last inequality we used the fact that

$$\sqrt{|\mathcal{P}| \inf_{I \in \mathcal{P}} \text{Leb}(I)} \geq c_{M,\text{Leb}} \quad \text{and} \quad D = (r+1) |\mathcal{P}| .$$

For all $j \in \{0, \dots, r\}$, $\varphi_{I,j}$ is supported by the element I of \mathcal{P} , hence we deduce from (66) that the orthonormal basis $\{\varphi_{I,j}, I \in \mathcal{P}, j \in \{0, \dots, r\}\}$ of $(M, \|\cdot\|_2)$ satisfies (45) with

$$L_{r,c_{\min}} = \sqrt{2} c_{\min}^{-1/2} (4r)^r .$$

To conclude, observe that

$$\begin{aligned}
\left\| \sum_{I,j} \beta_{I,j} \varphi_{I,j} \right\|_{\infty} &= \max_{I \in \mathcal{P}} \left\{ \left\| \sum_{j=0}^r \beta_{I,j} \varphi_{I,j} \right\|_{\infty} \right\} \\
&\leq |\beta|_{\infty} \max_{I \in \mathcal{P}} \left\{ \sum_{j=0}^r \|\varphi_{I,j}\|_{\infty} \right\} \\
&\leq (r+1) |\beta|_{\infty} \max_{I \in \mathcal{P}} \max_{j \in \{0, \dots, r\}} \left\{ \|\varphi_{I,j}\|_{\infty} \right\}
\end{aligned}$$

and thus, by plugging (67) into the right-hand side of the last inequality, we finally obtain that the value

$$L_{r,c_{\min},c_{M,\text{Leb}}} = \sqrt{2} c_{M,\text{Leb}}^{-1} c_{\min}^{-1/2} (4r)^r (r+1)^{1/2}$$

gives the desired bound (47). \blacksquare

We now turn to the proof of (50) under the assumptions of Lemma 9. The proof is based on concentration inequalities recalled in Section 6.5 and on inequality (45) of Lemma 8, that allows us to control the sup-norm of elements of an orthonormal basis for a model of piecewise polynomials.

Proof of Lemma 9. Let $\alpha > 0$ be fixed and $\gamma > 0$ to be chosen later. The partition \mathcal{P} associated to M will be denoted by

$$\mathcal{P} = \{I_0, \dots, I_{m-1}\} ,$$

so that $|\mathcal{P}| = m$ and $D = (r+1)m$ where D is the dimension of the model M . By (45) of Lemma 8 there exists an orthonormal basis $\{\varphi_{I_k, j}; k \in \{0, \dots, m-1\}, j \in \{0, \dots, r\}\}$ of $(M, L^2(P^X))$ such that,

$$\varphi_{I_k, j} \text{ is supported by the element } I_k \text{ of } \mathcal{P}, \text{ for all } j \in \{0, \dots, r\}$$

and a constant $L_{r, c_{\min}}$ depending only on r, c_{\min} and satisfying

$$\max_{j \in \{0, \dots, r\}} \|\varphi_{I_k, j}\|_{\infty} \leq L_{r, c_{\min}} \frac{1}{\sqrt{\text{Leb}(I_k)}}, \text{ for all } k \in \{0, \dots, m-1\}. \quad (68)$$

In order to avoid cumbersome notation, we define a total ordering \preceq on the set

$$\mathcal{I} = \{(I_k, j); k \in \{0, \dots, m-1\}, j \in \{0, \dots, r\}\} ,$$

as follows. Let \prec be a binary relation on $\mathcal{I} \times \mathcal{I}$ such that

$$(I_k, j) \prec (I_l, i) \text{ if } (k < l \text{ or } (k = l \text{ and } j < i)),$$

and consider the total ordering \preceq defined to be

$$(I_k, j) \preceq (I_l, i) \text{ if } ((I_k, j) = (I_l, i) \text{ or } (I_k, j) \prec (I_l, i)) .$$

So, from the definition of \preceq , the vector $\beta = (\beta_{I_k, j})_{(I_k, j) \in \mathcal{I}} \in \mathbb{R}^D$ has coordinate $\beta_{I_k, j}$ at position $(r+1)k + j + 1$ and when the matrix

$$A = (A_{(I_k, j), (I_l, i)})_{(I_k, j), (I_l, i) \in \mathcal{I} \times \mathcal{I}} \in \mathbb{R}^{D \times D} ,$$

has coefficient $A_{(I_k, j), (I_l, i)}$ at line $(r+1)k + j + 1$ and column $(r+1)l + i + 1$.

Now, for some $s = \sum_{(I_k, j) \in \mathcal{I}} \beta_{I_k, j} \varphi_{I_k, j} \in M$, we have

$$\begin{aligned} P_n(K(s)) &= P_n \left[\left(y - \left(\sum_{(I_k, j) \in \mathcal{I}} \beta_{I_k, j} \varphi_{I_k, j}(x) \right) \right)^2 \right] \\ &= P_n y^2 - 2 \sum_{(I_k, j) \in \mathcal{I}} \beta_{I_k, j} P_n(y \varphi_{I_k, j}(x)) + \sum_{(I_k, j), (I_l, i) \in \mathcal{I} \times \mathcal{I}} \beta_{I_k, j} \beta_{I_l, i} P_n(\varphi_{I_k, j} \varphi_{I_l, i}) . \end{aligned}$$

Hence, by taking the derivative with respect to $\beta_{I_k, j}$ in the last quantity,

$$\begin{aligned} &\frac{1}{2} \frac{\partial}{\partial \beta_{I_k, j}} P_n \left[\left(y - \left(\sum_{(I_k, j) \in \mathcal{I}} \beta_{I_k, j} \varphi_{I_k, j}(x) \right) \right)^2 \right] \\ &= -P_n(y \varphi_{I_k, j}(x)) + \sum_{(I_l, i) \in \mathcal{I}} \beta_{I_l, i} P_n(\varphi_{I_k, j} \varphi_{I_l, i}) . \end{aligned} \quad (69)$$

We see that if $\beta^{(n)} = (\beta_{I_k, j}^{(n)})_{(I_k, j) \in \mathcal{I}} \in \mathbb{R}^D$ is a critical point of

$$P_n \left[\left(y - \left(\sum_{(I_k, j) \in \mathcal{I}} \beta_{I_k, j} \varphi_{I_k, j}(x) \right) \right)^2 \right] ,$$

it holds

$$\left(\frac{\partial}{\partial \beta_{I_k,j}} P_n \left[\left(y - \left(\sum_{(I_k,j) \in \mathcal{I}} \beta_{I_k,j} \varphi_{I_k,j}(x) \right) \right)^2 \right] \right) (\beta^{(n)}) = 0$$

and by combining (69) with the fact that

$$P(\varphi_{I_k,j})^2 = 1, \text{ for all } (I_k,j) \in \mathcal{I} \quad \text{and} \quad P(\varphi_{I_k,j} \varphi_{I_l,i}) = 0 \text{ if } (I_k,j) \neq (I_l,i),$$

we deduce that $\beta^{(n)}$ satisfies the following random linear system,

$$(I_D + L_{n,D}) \beta^{(n)} = X_{y,n} \tag{70}$$

where $X_{y,n} = (P_n(y \varphi_{I_k,j}(x)))_{(I_k,j) \in \mathcal{I}} \in \mathbb{R}^D$, I_D is the identity matrix of dimension D and $L_{n,D} = ((L_{n,D})_{(I_k,j),(I_l,i)})_{(I_k,j),(I_l,i)}$ is a $D \times D$ matrix satisfying

$$(L_{n,D})_{(I_k,j),(I_l,i)} = (P_n - P)(\varphi_{I_k,j} \varphi_{I_l,i}).$$

Now, by inequality (82) in Lemma 11 below, one can find a positive integer $n_0(r, A_+, c_{\min}, c_{M,\text{Leb}}, \gamma)$ such that on an event Ω_n of probability at least $1 - 3Dn^{-\gamma}$, we have

$$\|L_{n,D}\| \leq \frac{1}{2}, \tag{71}$$

where for a $D \times D$ matrix L , the operator norm $\|\cdot\|$ associated to the sup-norm on vectors is

$$\|L\| = \sup_{x \neq 0} \frac{|Lx|_\infty}{|x|_\infty}.$$

Then we deduce from (71) that $(I_D + L_{n,D})$ is a non-singular $D \times D$ matrix and, as a consequence, that the linear system (70) admits a unique solution $\beta^{(n)}$ on Ω_n for all $n_0(r, A_+, c_{\min}, c_{M,\text{Leb}}, \gamma)$. Moreover, since $P_n \left(y - \left(\sum_{(I_k,j) \in \mathcal{I}} \beta_{I_k,j} \varphi_{I_k,j}(x) \right) \right)^2$ is a nonnegative quadratic functional with respect to $(\beta_{I_k,j})_{(I_k,j) \in \mathcal{I}} \in \mathbb{R}^D$ we can easily deduce that on Ω_n , $\beta^{(n)}$ achieves the unique minimum of $P_n \left(y - \left(\sum_{(I_k,j) \in \mathcal{I}} \beta_{I_k,j} \varphi_{I_k,j}(x) \right) \right)^2$ on \mathbb{R}^D . In other words,

$$s_n = \sum_{(I_k,j) \in \mathcal{I}} \beta_{I_k,j}^{(n)} \varphi_{I_k,j}$$

is the unique least-squares estimator on M , and by (70) it holds,

$$\beta_{I_k,j}^{(n)} \left(1 + \sum_{(I_l,i) \in \mathcal{I}} (P_n - P)(\varphi_{I_k,j} \varphi_{I_l,i}) \right) = P_n(y \varphi_{I_k,j}(x)), \text{ for all } (I_k,j) \in \mathcal{I}. \tag{72}$$

Now, as $\varphi_{I_k,j}$ and $\varphi_{I_l,i}$ have disjoint supports when $k \neq l$, it holds $\varphi_{I_k,j} \varphi_{I_l,i} = 0$ whenever $k \neq l$, and so equation (72) reduces to

$$\beta_{I_k,j}^{(n)} \times \left(1 + \sum_{i=0}^r (P_n - P)(\varphi_{I_k,j} \varphi_{I_k,i}) \right) = P_n(y \varphi_{I_k,j}(x)), \text{ for all } (I_k,j) \in \mathcal{I}. \tag{73}$$

Moreover, recalling that $s_M = \sum_{(I_k, j) \in \mathcal{I}} P(y\varphi_{I_k, j}(x)) \varphi_{I_k, j}$, it holds

$$\begin{aligned}
\|s_n - s_M\|_\infty &= \left\| \sum_{(I_k, j) \in \mathcal{I}} \left(\beta_{I_k, j}^{(n)} - P(y\varphi_{I_k, j}(x)) \right) \varphi_{I_k, j} \right\|_\infty \\
&\leq \max_{k \in \{0, \dots, m-1\}} \left\| \sum_{j=0}^r \left(\beta_{I_k, j}^{(n)} - P(y\varphi_{I_k, j}(x)) \right) \varphi_{I_k, j} \right\|_\infty \\
&\leq (r+1) \max_{k \in \{0, \dots, m-1\}} \left\{ \left(\max_{j \in \{0, \dots, r\}} \left| \beta_{I_k, j}^{(n)} - P(y\varphi_{I_k, j}(x)) \right| \right) \right. \\
&\quad \left. \times \max_{j \in \{0, \dots, r\}} \|\varphi_{I_k, j}\|_\infty \right\}
\end{aligned} \tag{74}$$

where the first inequality comes from the fact that $\varphi_{I_k, j}$ and $\varphi_{I_l, i}$ have disjoint supports when $k \neq l$. We next turn to the control of the right-hand side of (74). Let the index (I_k, j) be fixed. By subtracting the quantity $(1 + \sum_{i=0}^r (P_n - P)(\varphi_{I_k, j} \varphi_{I_k, i})) \times P(y\varphi_{I_k, j}(x))$ in each side of equation (73), we get

$$\begin{aligned}
&\left(\beta_{I_k, j}^{(n)} - P(y\varphi_{I_k, j}(x)) \right) \times \left(1 + \sum_{i=0}^r (P_n - P)(\varphi_{I_k, j} \varphi_{I_k, i}) \right) \\
&= (P_n - P)(y\varphi_{I_k, j}(x)) - \left(\sum_{i=0}^r (P_n - P)(\varphi_{I_k, j} \varphi_{I_k, i}) \right) \times P(y\varphi_{I_k, j}(x)) .
\end{aligned} \tag{75}$$

Moreover, by Inequality (83) of Lemma 11, we have for all $n \geq n_0(r, A_+, c_{\min}, c_{M, \text{Leb}}, \gamma)$,

$$\sum_{i=0}^r |(P_n - P)(\varphi_{I_k, j} \varphi_{I_k, i})| \leq L_{r, A_+, c_{\min}, c_{M, \text{Leb}}, \gamma} \sqrt{\frac{\ln n}{n \text{Leb}(I_k)}} \leq \frac{1}{2} \tag{76}$$

on the event Ω_n . We thus deduce that

$$\left| \left(\beta_{I_k, j}^{(n)} - P(y\varphi_{I_k, j}(x)) \right) \times \left(1 + \sum_{i=0}^r (P_n - P)(\varphi_{I_k, j} \varphi_{I_k, i}) \right) \right| \geq \frac{1}{2} \left| \beta_{I_k, j}^{(n)} - P(y\varphi_{I_k, j}(x)) \right| \tag{77}$$

and

$$\left| \left(\sum_{i=0}^r (P_n - P)(\varphi_{I_k, j} \varphi_{I_k, i}) \right) \times P(y\varphi_{I_k, j}(x)) \right| \leq L_{r, A_+, c_{\min}, c_{M, \text{Leb}}, \gamma} \sqrt{\frac{\ln n}{n \text{Leb}(I_k)}} \times |P(y\varphi_{I_k, j}(x))| . \tag{78}$$

Moreover, by (11), (48) and (68) we have

$$\begin{aligned}
|P(y\varphi_{I_k, j}(x))| &\leq A \|\varphi_{I_k, j}\|_\infty P(I_k) \\
&\leq A c_{\max} \|\varphi_{I_k, j}\|_\infty \text{Leb}(I_k) \\
&\leq A c_{\max} L_{r, c_{\min}} \sqrt{\text{Leb}(I_k)} \\
&\leq L_{A, r, c_{\min}, c_{\max}} \sqrt{\text{Leb}(I_k)} .
\end{aligned} \tag{79}$$

Putting inequality (79) in (78) we obtain

$$\left| \left(\sum_{i=0}^r (P_n - P)(\varphi_{I_k, j} \varphi_{I_k, i}) \right) \times P(y\varphi_{I_k, j}(x)) \right| \leq L_{r, A_+, c_{\min}, c_{\max}, c_{M, \text{Leb}}, \gamma} \sqrt{\frac{\ln n}{n}} . \tag{80}$$

Hence, using inequalities (77), (80) and inequality (84) of Lemma 11 in equation (75), we obtain that

$$\left| \beta_{I_k, j}^{(n)} - P(y\varphi_{I_k, j}(x)) \right| \leq L_{A, r, A_+, c_{\min}, c_{\max}, c_{M, \text{Leb}}, \gamma} \sqrt{\frac{\ln n}{n}}$$

on Ω_n . Since the constant $L_{A,r,A_+,c_{\min},c_{\max},c_{M,\text{Leb}},\gamma}$ does not depend on the index (I_k, j) we deduce by (68) that

$$\begin{aligned} & \left(\max_{j \in \{0, \dots, r\}} \left| \beta_{I_k, j}^{(n)} - P(y\varphi_{I_k, j}(x)) \right| \right) \times \max_{j \in \{0, \dots, r\}} \|\varphi_{I_k, j}\|_\infty \\ & \leq L_{A,r,A_+,c_{\min},c_{\max},c_{M,\text{Leb}},\gamma} \sqrt{\frac{\ln n}{n}} \times \max_{j \in \{0, \dots, r\}} \|\varphi_{I_k, j}\|_\infty \\ & \leq L_{A,r,A_+,c_{\min},c_{\max},c_{M,\text{Leb}},\gamma} \sqrt{\frac{\ln n}{n \text{Leb}(I_k)}}. \end{aligned} \quad (81)$$

Finally, by using (49) and (81) in (74), we get for all $n \geq n_0(r, A_+, c_{\min}, c_{M,\text{Leb}}, \gamma)$, on the event Ω_n of probability at least $1 - 3Dn^{-\gamma}$,

$$\begin{aligned} \|s_n - s_M\|_\infty & \leq (r+1) \max_{k \in \{0, \dots, m-1\}} \left\{ \left(\max_{j \in \{0, \dots, r\}} \left| \beta_{I_k, j}^{(n)} - P(y\varphi_{I_k, j}(x)) \right| \right) \times \max_{j \in \{0, \dots, r\}} \|\varphi_{I_k, j}\|_\infty \right\} \\ & \leq L_{A,r,A_+,c_{\min},c_{M,\text{Leb}},\gamma} \sqrt{\frac{\ln n}{n}} \max_{k \in \{0, \dots, m-1\}} \frac{1}{\sqrt{\text{Leb}(I_k)}} \\ & \leq L_{A,r,A_+,c_{\min},c_{M,\text{Leb}},\gamma} \sqrt{\frac{|\mathcal{P}| \ln n}{n}} \\ & \leq L_{A,r,A_+,c_{\min},c_{M,\text{Leb}},\gamma} \sqrt{\frac{D \ln n}{n}}. \end{aligned}$$

To conclude, simply take $\gamma = \frac{\ln 3}{\ln 2} + \alpha + 1$, so that it holds for $n \geq 2$, $\mathbb{P}[\Omega_n^c] \leq n^{-\alpha}$ which implies (50). It remains to prove the following lemma that has been used all along the proof.

Lemma 11 *Recall that $L_{n,D} = \left((L_{n,D})_{(I_k, j), (I_l, i)} \right)_{(I_k, j), (I_l, i) \in \mathcal{I} \times \mathcal{I}}$ is a $D \times D$ matrix such that for all $(k, l) \in \{0, \dots, m-1\}^2$, $(j, i) \in \{0, \dots, r\}^2$,*

$$(L_{n,D})_{(I_k, j), (I_l, i)} = (P_n - P)(\varphi_{I_k, j} \varphi_{I_l, i}).$$

Also recall that for a $D \times D$ matrix L , the operator norm $\|\cdot\|$ associated to the sup-norm on the vectors is

$$\|L\| = \sup_{x \neq 0} \frac{|Lx|_\infty}{|x|_\infty}.$$

Then, under the assumptions of Lemma 9, a positive integer $n_0(r, A_+, c_{\min}, c_{M,\text{Leb}}, \gamma)$ exists such that, for all $n \geq n_0(r, A_+, c_{\min}, c_{M,\text{Leb}}, \gamma)$, the following inequalities hold on an event Ω_n of probability at least $1 - 3Dn^{-\gamma}$,

$$\|L_{n,D}\| \leq L_{r,A_+,c_{\min},c_{M,\text{Leb}},\gamma} \sqrt{\frac{D \ln n}{n}} \leq \frac{1}{2} \quad (82)$$

and for all $k \in \{0, \dots, m-1\}$,

$$\max_{j \in \{0, \dots, r\}} \left\{ \sum_{i=0}^r |(P_n - P)(\varphi_{I_k, j} \varphi_{I_k, i})| \right\} \leq L_{r,A_+,c_{\min},c_{M,\text{Leb}},\gamma} \sqrt{\frac{\ln n}{n \text{Leb}(I_k)}} \leq \frac{1}{2}, \quad (83)$$

$$\max_{j \in \{0, \dots, r\}} |(P_n - P)(y\varphi_{I_k, j}(x))| \leq L_{A,A_+,r,c_{\min},c_{M,\text{Leb}},\gamma} \sqrt{\frac{\ln n}{n}}. \quad (84)$$

Proof of Lemma 11. Let us begin with the proof of inequality (84). Let the index $(I_k, j) \in \mathcal{I}$ be fixed. By using Bernstein's inequality (213) and observing that, by (11),

$$\text{Var}(y\varphi_{I_k, j}(x)) \leq P\left[(y\varphi_{I_k, j}(x))^2\right] \leq \|Y\|_\infty^2 \leq A^2$$

and, by (11), (68) and (49),

$$\begin{aligned}
\|Y\varphi_{I_k,j}(X)\|_\infty &\leq A\|\varphi_{I_k,j}(X)\|_\infty \\
&\leq AL_{r,c_{\min}} \frac{1}{\sqrt{\text{Leb}(I_k)}} \\
&\leq L_{A,r,c_{\min},c_{M,\text{Leb}}} \sqrt{|\mathcal{P}|} \\
&\leq L_{A,r,c_{\min},c_{M,\text{Leb}}} \sqrt{D} ,
\end{aligned}$$

we get

$$\mathbb{P} \left[|(P_n - P)(y\varphi_{I_k,j}(x))| \geq \sqrt{2A^2 \frac{x}{n}} + \frac{L_{A,r,c_{\min},c_{M,\text{Leb}}} \sqrt{D}}{3n} x \right] \leq 2 \exp(-x) . \quad (85)$$

By taking $x = \gamma \ln n$ in inequality (85), we obtain that

$$\mathbb{P} \left[|(P_n - P)(y\varphi_{I_k,j}(x))| \geq \sqrt{2A^2 \gamma \frac{\ln n}{n}} + \frac{L_{A,r,c_{\min},c_{M,\text{Leb}}} \sqrt{D} \gamma \ln n}{3n} \right] \leq 2n^{-\gamma} . \quad (86)$$

Now, as $D \leq A_+ n (\ln n)^{-2}$, we deduce from (86) that for some well chosen positive constant $L_{A,A_+,r,c_{\min},c_{M,\text{Leb}},\gamma}$, we have

$$\mathbb{P} \left[|(P_n - P)(y\varphi_{I_k,j}(x))| \geq L_{A,A_+,r,c_{\min},c_{M,\text{Leb}},\gamma} \sqrt{\frac{\ln n}{n}} \right] \leq 2n^{-\gamma}$$

and by setting

$$\Omega_n^{(1)} = \bigcap_{(I_k,j) \in \mathcal{I}} \left\{ |(P_n - P)(y\varphi_{I_k,j}(x))| \leq L_{A,A_+,r,c_{\min},c_{M,\text{Leb}},\gamma} \sqrt{\frac{\ln n}{n}} \right\}$$

we deduce that

$$\mathbb{P}(\Omega_n^{(1)}) \geq 1 - 2Dn^{-\gamma} . \quad (87)$$

Hence the expected bound (84) holds on $\Omega_n^{(1)}$, for all $n \geq 1$.

We turn now to the proof of inequality (83). Let the index $(I_k, j) \in \mathcal{I}$ be fixed. By Cauchy-Schwarz inequality, we have

$$\sum_{i=0}^r |(P_n - P)(\varphi_{I_k,j}\varphi_{I_k,i})| \leq \sqrt{r+1} \sqrt{\sum_{i=0}^r ((P_n - P)(\varphi_{I_k,j}\varphi_{I_k,i}))^2} . \quad (88)$$

Let write

$$\chi_{I_k,j} = \sqrt{\sum_{i=0}^r ((P_n - P)(\varphi_{I_k,j}\varphi_{I_k,i}))^2} \quad \text{and} \quad B_{I_k} = \left\{ \sum_{i=0}^r \beta_{I_k,i} \varphi_{I_k,i} ; (\beta_{I_k,i})_{i=0}^r \in \mathbb{R}^{r+1} \text{ and } \sum_{i=0}^r \beta_{I_k,i}^2 \leq 1 \right\} .$$

By Cauchy-Schwarz inequality again, it holds

$$\chi_{I_k,j} = \sup_{s \in B_{I_k}} |(P_n - P)(\varphi_{I_k,j}s)| .$$

Then, Bousquet's inequality (214), applied with $\varepsilon = 1$ and $\mathcal{F} = B_{I_k}$, implies that

$$\mathbb{P} \left[\chi_{I_k,j} - \mathbb{E}[\chi_{I_k,j}] \geq \sqrt{2\sigma_{I_k,j}^2 \frac{x}{n}} + \mathbb{E}[\chi_{I_k,j}] + \frac{4}{3} \frac{b_{I_k,j} x}{n} \right] \leq \exp(-x) \quad (89)$$

where, by (68),

$$\sigma_{I_k,j}^2 = \sup_{s \in B_{I_k}} \text{Var}(\varphi_{I_k,j}s) \leq \|\varphi_{I_k,j}\|_\infty^2 \leq \frac{L_{r,c_{\min}}}{\text{Leb}(I_k)} \quad (90)$$

and

$$b_{I_k,j} \leq 2 \sup_{s \in B_{I_k}} \|\varphi_{I_k,j} s\|_\infty \leq 2 \|\varphi_{I_k,j}\|_\infty \sup_{s \in B_{I_k}} \|s\|_\infty . \quad (91)$$

Moreover, for $s = \sum_{i=0}^r \beta_{I_k,i} \varphi_{I_k,i} \in B_{I_k}$, we have $\max_i |\beta_{I_k,i}| \leq \sqrt{\sum_{i=0}^r \beta_{I_k,i}^2} \leq 1$, so by (68),

$$\sup_{s \in B_{I_k}} \|s\|_\infty \leq \sum_{i=0}^r \|\varphi_{I_k,i}\|_\infty \leq \frac{L_{r,c_{\min}}}{\sqrt{\text{Leb}(I_k)}}$$

and injecting the last bound in (91) we get

$$b_{I_k,j} \leq \|\varphi_{I_k,j}\|_\infty \frac{L_{r,c_{\min}}}{\sqrt{\text{Leb}(I_k)}} \leq \frac{L_{r,c_{\min}}}{\text{Leb}(I_k)} . \quad (92)$$

In addition, we have

$$\begin{aligned} \mathbb{E} [\chi_{I_k,j}] &\leq \sqrt{\mathbb{E} [\chi_{I_k,j}^2]} = \sqrt{\frac{\sum_{i=0}^r \text{Var}(\varphi_{I_k,j} \varphi_{I_k,i})}{n}} \\ &\leq \|\varphi_{I_k,j}\|_\infty \sqrt{\frac{\sum_{i=0}^r P(\varphi_{I_k,i}^2)}{n}} \\ &= \|\varphi_{I_k,j}\|_\infty \sqrt{\frac{r+1}{n}} \\ &\leq L_{r,c_{\min}} \sqrt{\frac{1}{n \text{Leb}(I_k)}} . \end{aligned} \quad (93)$$

Therefore, combining (90), (92), (93) and (89) while taking $x = \gamma \ln n$, we get

$$\mathbb{P} \left[\chi_{I_k,j} \geq L_{r,c_{\min},\gamma} \left(\sqrt{\frac{1}{n \text{Leb}(I_k)}} + \sqrt{\frac{\ln n}{n \text{Leb}(I_k)}} + \frac{\ln n}{n \text{Leb}(I_k)} \right) \right] \leq n^{-\gamma} . \quad (94)$$

Now, since by (49) and the fact that $D \leq A_+ n (\ln n)^{-2}$ we have

$$\frac{1}{\text{Leb}(I_k)} \leq c_{M,\text{Leb}}^{-2} D \leq c_{M,\text{Leb}}^{-2} A_+ \frac{n}{(\ln n)^2} ,$$

we obtain from (94) that a positive constant $L_{r,A_+,c_{\min},c_{M,\text{Leb}},\gamma}$ exists, depending only on γ, r, A_+, c_{\min} and $c_{M,\text{Leb}}$ such that

$$\mathbb{P} \left[\chi_{I_k,j} \geq L_{r,A_+,c_{\min},c_{M,\text{Leb}},\gamma} \sqrt{\frac{\ln n}{n \text{Leb}(I_k)}} \right] \leq n^{-\gamma} . \quad (95)$$

Finally, define

$$\Omega_n^{(2)} = \bigcap_{(I_k,j) \in \mathcal{I}} \left\{ \chi_{I_k,j} \leq L_{r,A_+,c_{\min},c_{M,\text{Leb}},\gamma} \sqrt{\frac{\ln n}{n \text{Leb}(I_k)}} \right\} .$$

For all $n \geq n_0(r, A_+, c_{\min}, c_{M,\text{Leb}}, \gamma)$, we have

$$\begin{aligned} &\sqrt{r+1} \times L_{r,A_+,c_{\min},c_{M,\text{Leb}},\gamma} \sqrt{\frac{\ln n}{n \text{Leb}(I_k)}} \\ &\leq L_{r,A_+,c_{\min},c_{M,\text{Leb}},\gamma} \sqrt{\frac{D \ln n}{n}} \\ &\leq L_{r,A_+,c_{\min},c_{M,\text{Leb}},\gamma} \frac{1}{\sqrt{\ln n}} \leq \frac{1}{2} . \end{aligned} \quad (96)$$

Moreover by (95) it holds

$$\mathbb{P}\left(\Omega_n^{(2)}\right) \geq 1 - Dn^{-\gamma} \quad (97)$$

and, by (88), the expected bound (83) holds on $\Omega_n^{(2)}$, for all $n \geq n_0(r, A_+, c_{\min}, c_{M, \text{Leb}}, \gamma)$.

Next, notice that for a $D \times D$ matrix $L = \left(L_{(I_k, j), (I_l, i)}\right)_{(I_k, j), (I_l, i) \in \mathcal{I} \times \mathcal{I}}$ we have the following classical formula,

$$\|L\| = \max_{(I_k, j) \in \mathcal{I}} \sum_{(I_l, i) \in \mathcal{I}} |L_{(I_k, j), (I_l, i)}|.$$

Applied to the matrix of interest $L_{n, D}$, this gives

$$\begin{aligned} \|L_{n, D}\| &= \max_{(I_k, j) \in \mathcal{I}} \sum_{(I_l, i) \in \mathcal{I}} |(P_n - P)(\varphi_{I_k, j} \varphi_{I_l, i})| \\ &= \max_{k \in \{0, \dots, m-1\}} \max_{j \in \{0, \dots, r\}} \left\{ \sum_{(I_l, i) \in \mathcal{I}} |(P_n - P)(\varphi_{I_k, j} \varphi_{I_l, i})| \right\}. \end{aligned} \quad (98)$$

Thus, using formula (98), inequalities (83), (49) and (96) give that for all $n \geq n_0(r, A_+, c_{\min}, c_{M, \text{Leb}}, \gamma)$, we have on $\Omega_n^{(2)}$,

$$\|L_{n, D}\| \leq L_{r, A_+, c_{\min}, c_{M, \text{Leb}}, \gamma} \sqrt{\frac{D \ln n}{n}} \leq \frac{1}{2}.$$

Finally, by setting $\Omega_n = \Omega_n^{(1)} \cap \Omega_n^{(2)}$, we have $\mathbb{P}(\Omega_n) \geq 1 - 3Dn^{-\gamma}$, and inequalities (83), (82) and (84) are satisfied on Ω_n for all $n \geq n_0(r, A_+, c_{\min}, c_{M, \text{Leb}}, \gamma)$, which completes the proof of Lemma 11. ■

6.3 Proofs of Section 3

In order to express the quantities of interest in the proofs of Theorems 3 and 4, we need preliminary definitions. Let $\alpha > 0$ be fixed and for $R_{n, D, \alpha}$ defined in (H5), see Section 3.1, we set

$$\tilde{R}_{n, D, \alpha} = \max \left\{ R_{n, D, \alpha} ; A_\infty \sqrt{\frac{D \ln n}{n}} \right\} \quad (99)$$

where A_∞ is a positive constant to be chosen later. Moreover, we set

$$\nu_n = \max \left\{ \sqrt{\frac{\ln n}{D}} ; \sqrt{\frac{D \ln n}{n}} ; R_{n, D, \alpha} \right\}. \quad (100)$$

Thanks to the assumption of consistency in sup-norm (H5), our analysis will be localized in the subset

$$B_{(M, L_\infty)}(s_M, \tilde{R}_{n, D, \alpha}) = \left\{ s \in M, \|s - s_M\|_\infty \leq \tilde{R}_{n, D, \alpha} \right\}$$

of M .

Let us define several slices of excess risk on the model M : for any $C \geq 0$,

$$\begin{aligned} \mathcal{F}_C &= \{s \in M, P(Ks - Ks_M) \leq C\} \cap B_{(M, L_\infty)}(s_M, \tilde{R}_{n, D, \alpha}) \\ \mathcal{F}_{>C} &= \{s \in M, P(Ks - Ks_M) > C\} \cap B_{(M, L_\infty)}(s_M, \tilde{R}_{n, D, \alpha}) \end{aligned}$$

and for any interval $J \subset \mathbb{R}$,

$$\mathcal{F}_J = \{s \in M, P(Ks - Ks_M) \in J\} \cap B_{(M, L_\infty)}(s_M, \tilde{R}_{n, D, \alpha}).$$

We also define, for all $L \geq 0$,

$$D_L = \{s \in M, P(Ks - Ks_M) = L\} \cap B_{(M, L_\infty)}(s_M, \tilde{R}_{n, D, \alpha}) .$$

Recall that, by Lemma 1 of Section 2.2, the contrasted functions satisfy, for every $s \in M$ and $z = (x, y) \in \mathcal{X} \times \mathbb{R}$,

$$(Ks)(z) - (Ks_M)(z) = \psi_{1, M}(z)(s - s_M)(x) + \psi_2((s - s_M)(x))$$

where $\psi_{1, M}(z) = -2(y - s_M(x))$ and $\psi_2(t) = t^2$, for all $t \in \mathbb{R}$. For convenience, we will use the following notation, for any $s \in M$,

$$\psi_2 \circ (s - s_M) : x \in \mathcal{X} \mapsto \psi_2((s - s_M)(x)) .$$

Note that, for all $s \in M$,

$$P(\psi_{1, M} \cdot s) = 0 \tag{101}$$

and by **(H1)** inequality (15) holds true, that is

$$\|\psi_{1, M}\|_\infty \leq 4A . \tag{102}$$

Also, for $\mathcal{K}_{1, M}$ defined in Section 3.3, we have

$$\mathcal{K}_{1, M} = \sqrt{\frac{1}{D} \sum_{k=1}^D \text{Var}(\psi_{1, M} \cdot \varphi_k)}$$

for any orthonormal basis $(\varphi_k)_{k=1}^D$ of $(M, \|\cdot\|_2)$. Moreover, inequality (31) holds under **(H1)** and we have

$$\mathcal{K}_{1, M} \leq 2\sigma_{\max} + 4A \leq 6A . \tag{103}$$

Assuming **(H2)**, we have from (32)

$$0 < 2\sigma_{\min} \leq \mathcal{K}_{1, M} . \tag{104}$$

Finally, when **(H3)** holds (it is the case when **(H4)** holds), we have by (16),

$$\sup_{s \in M, \|s\|_2 \leq 1} \|s\|_\infty \leq A_{3, M} \sqrt{D} \tag{105}$$

and so, for any orthonormal basis $(\varphi_k)_{k=1}^D$ of $(M, \|\cdot\|_2)$, it holds for all $k \in \{1, \dots, D\}$, as $P(\varphi_k^2) = 1$,

$$\|\varphi_k\|_\infty \leq A_{3, M} \sqrt{D} . \tag{106}$$

6.3.1 Proofs of the theorems

The proof of Theorem 3 relies on Lemmas 18, 19 and 20 stated in Section 6.4, and that give sharp estimates of suprema of the empirical process on the contrasted functions over slices of interest.

Proof of Theorem 3. Let $\alpha > 0$ be fixed and let $\varphi = (\varphi_k)_{k=1}^D$ be an orthonormal basis of $(M, \|\cdot\|_2)$ satisfying **(H4)**. We divide the proof of Theorem 3 into four parts, corresponding to the four Inequalities (22), (23), (24) and (25). The values of A_0 and A_∞ , respectively defined in (21) and (99), will then be chosen at the end of the proof.

Proof of Inequality (22). Let $r \in (1, 2]$ to be chosen later and $C > 0$ such that

$$rC = \frac{D}{4n} \mathcal{K}_{1,M}^2. \quad (107)$$

By **(H5)** there exists a positive integer n_1 such that it holds, for all $n \geq n_1$,

$$\mathbb{P}(P(Ks_n - Ks_M) \leq C) \leq \mathbb{P}\left(\{P(Ks_n - Ks_M) \leq C\} \cap \Omega_{\infty, \alpha}\right) + n^{-\alpha} \quad (108)$$

and also

$$\begin{aligned} & \mathbb{P}\left(\{P(Ks_n - Ks_M) \leq C\} \cap \Omega_{\infty, \alpha}\right) \\ & \leq \mathbb{P}\left(\inf_{s \in \mathcal{F}_C} P_n(Ks - Ks_M) \leq \inf_{s \in \mathcal{F}_{>C}} P_n(Ks - Ks_M)\right) \\ & \leq \mathbb{P}\left(\inf_{s \in \mathcal{F}_C} P_n(Ks - Ks_M) \leq \inf_{s \in \mathcal{F}_{(C, rC]}} P_n(Ks - Ks_M)\right) \\ & = \mathbb{P}\left(\sup_{s \in \mathcal{F}_C} P_n(Ks_M - Ks) \geq \sup_{s \in \mathcal{F}_{(C, rC]}} P_n(Ks_M - Ks)\right). \end{aligned} \quad (109)$$

Now, by (107) and (104) we have

$$\frac{D}{2n} \sigma_{\min}^2 \leq C \leq (1 + A_4 \nu_n)^2 \frac{D}{4n} \mathcal{K}_{1,M}^2$$

where A_4 is defined in Lemma 18. Hence we can apply Lemma 18 with $\alpha = \beta$, $A_l = \sigma_{\min}^2/2$ and $A_{3,M} = r_M(\varphi)$, by Remark 2. Therefore it holds, for all $n \geq n_0(A_{\infty}, A_{cons}, A_+, \sigma_{\min}, \alpha)$,

$$\mathbb{P}\left[\sup_{s \in \mathcal{F}_C} P_n(Ks_M - Ks) \geq (1 + L_{A_{\infty}, A, r_M(\varphi), \sigma_{\min}, A_-, \alpha} \times \nu_n) \sqrt{\frac{CD}{n}} \mathcal{K}_{1,M} - C\right] \leq 2n^{-\alpha}. \quad (110)$$

Moreover, by using (104) and (103) in (107) we get

$$\frac{D}{n} \sigma_{\min}^2 \leq rC \leq \frac{D}{n} (\sigma_{\max} + 2A)^2.$$

We then apply Lemma 20 with

$$\alpha = \beta, \quad A_l = \sigma_{\min}^2, \quad A_u = (\sigma_{\max} + 2A)^2$$

and

$$A_{\infty} \geq 64\sqrt{2}B_2A(\sigma_{\max} + 2A)\sigma_{\min}^{-1}r_M(\varphi), \quad (111)$$

so it holds for all $n \geq n_0(A_-, A_+, A, A_{\infty}, A_{cons}, B_2, r_M(\varphi), \sigma_{\max}, \sigma_{\min}, \alpha)$,

$$\mathbb{P}\left(\sup_{s \in \mathcal{F}_{(C, rC]}} P_n(Ks_M - Ks) \leq (1 - L_{A_-, A, A_{\infty}, \sigma_{\max}, \sigma_{\min}, r_M(\varphi), \alpha} \times \nu_n) \sqrt{\frac{rCD}{n}} \mathcal{K}_{1,M} - rC\right) \leq 2n^{-\alpha}. \quad (112)$$

Now, from (110) and (112) we can find a positive constant \tilde{A}_0 , only depending on $A_-, A, A_{\infty}, \sigma_{\max}, \sigma_{\min}, r_M(\varphi)$ and α , such that for all $n \geq n_0(A_-, A_+, A, A_{\infty}, A_{cons}, B_2, r_M(\varphi), \sigma_{\max}, \sigma_{\min}, \alpha)$, there exists an event of probability at least $1 - 4n^{-\alpha}$ on which

$$\sup_{s \in \mathcal{F}_C} P_n(Ks_M - Ks) \leq (1 + \tilde{A}_0 \nu_n) \sqrt{\frac{CD}{n}} \mathcal{K}_{1,M} - C \quad (113)$$

and

$$\sup_{s \in \mathcal{F}_{(C, rC)}} P_n(Ks_M - Ks) \geq \left(1 - \tilde{A}_0 \nu_n\right) \sqrt{\frac{rCD}{n}} \mathcal{K}_{1,M} - rC. \quad (114)$$

Hence, from (113) and (114) we deduce, using (108) and (109), that if we choose $r \in (1, 2]$ such that

$$\left(1 + \tilde{A}_0 \nu_n\right) \sqrt{\frac{CD}{n}} \mathcal{K}_{1,M} - C < \left(1 - \tilde{A}_0 \nu_n\right) \sqrt{\frac{rCD}{n}} \mathcal{K}_{1,M} - rC \quad (115)$$

then, for all $n \geq n_0(A_-, A_+, A, A_\infty, A_{cons}, B_2, r_M(\varphi), \sigma_{\max}, \sigma_{\min}, n_1, \alpha)$ we have

$$P(Ks_n - Ks_M) \geq C$$

with probability at least $1 - 5n^{-\alpha}$. Now, by (107) it holds

$$\sqrt{\frac{rCD}{n}} \mathcal{K}_{1,M} = 2rC = \frac{1}{2} \frac{D}{n} \mathcal{K}_{1,M}^2,$$

and as a consequence Inequality (115) is equivalent to

$$\left(1 - 2\tilde{A}_0 \nu_n\right) r - 2\left(1 + \tilde{A}_0 \nu_n\right) \sqrt{r} + 1 > 0. \quad (116)$$

Moreover, we have by (100) and **(H5)**, for all $n \geq n_0(A_+, A_-, A_{cons}, \tilde{A}_0, \alpha)$,

$$\tilde{A}_0 \nu_n \leq \frac{1}{4} \quad (117)$$

and so, for all $n \geq n_0(A_+, A_-, A_{cons}, \tilde{A}_0, \alpha)$, simple computations involving (117) show that by taking

$$r = 1 + 48\sqrt{\tilde{A}_0 \nu_n} \quad (118)$$

inequality (116) is satisfied. Notice that, for all $n \geq n_0(A_+, A_-, A_{cons}, \tilde{A}_0, \alpha)$ we have $0 < 48\sqrt{\tilde{A}_0 \nu_n} < 1$, so that $r \in (1, 2)$. Finally, we compute C by (107) and (118), in such a way that for all $n \geq n_0(A_+, A_-, A_{cons}, \tilde{A}_0, \alpha)$,

$$C = \frac{rC}{r} = \frac{1}{1 + 48\sqrt{\tilde{A}_0 \nu_n}} \frac{1}{4} \frac{D}{n} \mathcal{K}_{1,M}^2 \geq \left(1 - 48\sqrt{\tilde{A}_0 \nu_n}\right) \frac{1}{4} \frac{D}{n} \mathcal{K}_{1,M}^2 > 0 \quad (119)$$

which yields the result by noticing that the dependence on σ_{\max} can be released in n_0 and \tilde{A}_0 since by **(H1)** we have $\sigma_{\max} \leq A$.

Proof of Inequality (23). Let $C > 0$ and $\delta \in (0, \frac{1}{2})$ to be chosen later in such a way that

$$(1 - \delta)C = \frac{D}{4n} \mathcal{K}_{1,M}^2 \quad (120)$$

and

$$C \geq \frac{1}{4} (1 + A_5 \nu_n)^2 \frac{D}{n} \mathcal{K}_{1,M}^2, \quad (121)$$

where A_5 is defined in Lemma 19. We have by **(H5)**, for all $n \geq n_1$,

$$\mathbb{P}(P(Ks_n - Ks_M) > C) \leq \mathbb{P}\left(\{P(Ks_n - Ks_M) > C\} \cap \bigcap \Omega_{\infty, \alpha}\right) + n^{-\alpha} \quad (122)$$

and also

$$\begin{aligned}
& \mathbb{P} \left(\{P(Ks_n - Ks_M) > C\} \cap \Omega_{\infty, \alpha} \right) \\
& \leq \mathbb{P} \left(\inf_{s \in \mathcal{F}_C} P_n(Ks - Ks_M) \geq \inf_{s \in \mathcal{F}_{>C}} P_n(Ks - Ks_M) \right) \\
& = \mathbb{P} \left(\sup_{s \in \mathcal{F}_C} P_n(Ks_M - Ks) \leq \sup_{s \in \mathcal{F}_{>C}} P_n(Ks_M - Ks) \right) \\
& \leq \mathbb{P} \left(\sup_{s \in \mathcal{F}_{\left(\frac{C}{2}, (1-\delta)C\right]}} P_n(Ks_M - Ks) \leq \sup_{s \in \mathcal{F}_{>C}} P_n(Ks_M - Ks) \right). \tag{123}
\end{aligned}$$

Now by (121) we can apply Lemma 19 with $\alpha = \beta$ and we obtain, for all $n \geq n_0(A_\infty, A_{cons}, A_+, \alpha)$,

$$\mathbb{P} \left[\sup_{s \in \mathcal{F}_{>C}} P_n(Ks_M - Ks) \geq (1 + A_5 \nu_n) \sqrt{\frac{CD}{n}} \mathcal{K}_{1,M} - C \right] \leq 2n^{-\alpha} \tag{124}$$

where A_5 only depends on $A, A_{3,M}, A_\infty, \sigma_{\min}, A_-$ and α . Moreover, we can take $A_{3,M} = r_M(\varphi)$ by Remark 2. Also, by (120), (104) and (103) we can apply Lemma 20 with the quantity C in Lemma 20 replaced by $C/2$, $\alpha = \beta$, $r = 2(1-\delta)$, $A_u = (\sigma_{\max} + 2A)^2$, $A_l = \sigma_{\min}^2$ and the constant A_∞ satisfying

$$A_\infty \geq 64\sqrt{2}B_2A(\sigma_{\max} + 2A)\sigma_{\min}^{-1}r_M(\varphi), \tag{125}$$

and so it holds, for all $n \geq n_0(A_-, A_+, A, A_\infty, A_{cons}, B_2, r_M(\varphi), \sigma_{\max}, \sigma_{\min}, \alpha)$,

$$\mathbb{P} \left(\begin{aligned} & \sup_{s \in \mathcal{F}_{\left(\frac{C}{2}, (1-\delta)C\right]}} P_n(Ks_M - Ks) \\ & \leq (1 - L_{A_-, A, A_\infty, \sigma_{\max}, \sigma_{\min}, r_M(\varphi), \alpha} \times \nu_n) \sqrt{\frac{(1-\delta)CD}{n}} \mathcal{K}_{1,M} - (1-\delta)C \end{aligned} \right) \leq 2n^{-\alpha}. \tag{126}$$

Hence from (124) and (126), we deduce that a positive constant \check{A}_0 exists, only depending on $A_-, A, A_\infty, \sigma_{\max}, \sigma_{\min}, r_M(\varphi)$ and α , such that

for all $n \geq n_0(A_-, A_+, A, A_\infty, A_{cons}, B_2, r_M(\varphi), \sigma_{\max}, \sigma_{\min}, \alpha)$ it holds on an event of probability at least $1 - 4n^{-\alpha}$,

$$\sup_{s \in \mathcal{F}_{\left(\frac{C}{2}, (1-\delta)C\right]}} P_n(Ks_M - Ks) \geq (1 - \check{A}_0 \nu_n) \sqrt{\frac{(1-\delta)CD}{n}} \mathcal{K}_{1,M} - (1-\delta)C \tag{127}$$

and

$$\sup_{s \in \mathcal{F}_{>C}} P_n(Ks_M - Ks) \leq (1 + \check{A}_0 \nu_n) \sqrt{\frac{CD}{n}} \mathcal{K}_{1,M} - C. \tag{128}$$

Now, from (127) and (128) we deduce, using (122) and (123), that if we choose $\delta \in (0, \frac{1}{2})$ such that (121) and

$$(1 + \check{A}_0 \nu_n) \sqrt{\frac{CD}{n}} \mathcal{K}_{1,M} - C < (1 - \check{A}_0 \nu_n) \sqrt{\frac{(1-\delta)CD}{n}} \mathcal{K}_{1,M} - (1-\delta)C \tag{129}$$

are satisfied then, for all $n \geq n_0(A_-, A_+, A, A_\infty, A_{cons}, B_2, r_M(\varphi), \sigma_{\max}, \sigma_{\min}, n_1, \alpha)$,

$$P(Ks_n - Ks_M) \leq C,$$

with probability at least $1 - 5n^{-\alpha}$. By (120) it holds

$$\sqrt{\frac{(1-\delta)CD}{n}} \mathcal{K}_{1,M} = 2(1-\delta)C = \frac{1}{2} \frac{D}{n} \mathcal{K}_{1,M}^2,$$

and by consequence, inequality (129) is equivalent to

$$(1 - 2\check{A}_0 \nu_n)(1-\delta) - 2(1 + \check{A}_0 \nu_n)\sqrt{1-\delta} + 1 > 0. \tag{130}$$

Moreover, we have by (100) and (H5), for all $n \geq n_0 (A_+, A_-, A_{cons}, \check{A}_0, A_5, \alpha)$,

$$(\check{A}_0 \vee A_5) \nu_n < \frac{1}{72} \quad (131)$$

and so, for all $n \geq n_0 (A_+, A_-, A_{cons}, \check{A}_0, \alpha)$, simple computations involving (131) show that by taking

$$\delta = 6 \left(\sqrt{\check{A}_0} \vee \sqrt{A_5} \right) \sqrt{\nu_n} , \quad (132)$$

inequalities (130) and (121) are satisfied and $\delta \in (0, \frac{1}{2})$. Finally, we can compute C by (120) and (132), in such a way that for all $n \geq n_0 (A_+, A_-, A_{cons}, \check{A}_0, \alpha)$

$$0 < C = \frac{(1-\delta)C}{(1-\delta)} = \frac{1}{(1-\delta)} \frac{1}{4} \frac{D}{n} \mathcal{K}_{1,M}^2 \leq \left(1 + 12 \left(\sqrt{\check{A}_0} \vee \sqrt{A_5} \right) \sqrt{\nu_n} \right) \frac{1}{4} \frac{D}{n} \mathcal{K}_{1,M}^2 , \quad (133)$$

which yields the result by noticing that the dependence on σ_{\max} can be released from n_0 and \check{A}_0 since by (H1) we have $\sigma_{\max} \leq A$.

Proof of Inequality (24). Let $C = \frac{D}{8n} \mathcal{K}_{1,M}^2 > 0$ and let $r = 2$. By (103) and (104) we have

$$\frac{D}{n} \sigma_{\min}^2 \leq rC = \frac{D}{4n} \mathcal{K}_{1,M}^2 \leq \frac{D}{n} (\sigma_{\max} + 2A)^2$$

so we can apply Lemma 20 with $\alpha = \beta$, $A_l = \sigma_{\min}^2$ and $A_u = (\sigma_{\max} + 2A)^2$. So if

$$A_{\infty} \geq 64\sqrt{2}B_2A(\sigma_{\max} + 2A)\sigma_{\min}^{-1}r_M(\varphi) , \quad (134)$$

it holds, for all $n \geq n_0 (A_-, A_+, A, A_{\infty}, A_{cons}, B_2, r_M(\varphi), \sigma_{\max}, \sigma_{\min}, \alpha)$,

$$\mathbb{P} \left(\sup_{s \in \mathcal{F}_{(C, rC]}} P_n(Ks_M - Ks) \leq (1 - L_{A_-, A, A_{\infty}, \sigma_{\max}, \sigma_{\min}, r_M(\varphi), \alpha} \times \nu_n) \sqrt{\frac{rCD}{n}} \mathcal{K}_{1,M} - rC \right) \leq 2n^{-\alpha} . \quad (135)$$

Since $rC = \frac{D}{4n} \mathcal{K}_{1,M}^2$, if we set $\hat{A}_0 = 2L_{A_-, A, A_{\infty}, \sigma_{\max}, \sigma_{\min}, r_M(\varphi), \alpha}$ with $L_{A_-, A, A_{\infty}, \sigma_{\max}, \sigma_{\min}, r_M(\varphi), \alpha}$ the constant in (135), we get

$$\mathbb{P} \left(\sup_{s \in \mathcal{F}_{(C, rC]}} P_n(Ks_M - Ks) \leq (1 - \hat{A}_0 \nu_n) \frac{D}{4n} \mathcal{K}_{1,M}^2 \right) \leq 2n^{-\alpha} . \quad (136)$$

Notice that

$$P_n(Ks_M - Ks_n) = \sup_{s \in M} P_n(Ks_M - Ks) \geq \sup_{s \in \mathcal{F}_{(C, rC]}} P_n(Ks_M - Ks)$$

so from (136) we deduce that

$$\mathbb{P} \left(P_n(Ks_M - Ks_n) \geq (1 - \hat{A}_0 \nu_n) \frac{D}{4n} \mathcal{K}_{1,M}^2 \right) \geq 1 - 2n^{-\alpha} . \quad (137)$$

Remark 12 Notice that in the proof of inequality (24), we do not need to assume the consistency of the least-squares estimator s_n towards the projection s_M . Straightforward adaptations of Lemma 20 allow to take

$$\tilde{\nu}_n = \max \left\{ \sqrt{\frac{\ln n}{D}}, \sqrt{\frac{D \ln n}{n}} \right\}$$

instead of the quantity ν_n defined in (100). This readily gives the expected bound (26) of Theorem 3.

Proof of Inequality (25). Let

$$C = \frac{1}{4} (1 + A_5 \nu_n)^2 \frac{D}{n} \mathcal{K}_{1,M}^2 > 0 \quad (138)$$

where A_5 is defined in Lemma 19 applied with $\beta = \alpha$. By **(H5)** we have

$$\mathbb{P}(P_n(Ks_M - Ks_n) > C) \leq \mathbb{P}\left(\{P_n(Ks_M - Ks_n) > C\} \cap \Omega_{\infty, \alpha}\right) + n^{-\alpha}. \quad (139)$$

Moreover, on $\Omega_{\infty, \alpha}$, we have

$$\begin{aligned} P_n(Ks_M - Ks_n) &= \sup_{s \in B_{(M, L_\infty)}(s_M, \tilde{R}_{n,D,\alpha})} P_n(Ks_M - Ks) \\ &= \sup_{s \in \mathcal{F}_{>0}} P_n(Ks_M - Ks) \end{aligned} \quad (140)$$

and by (198) of Lemma 19 applied with $\alpha = \beta$ it holds, for all $n \geq n_0(A_\infty, A_{cons}, A_+, \alpha)$,

$$\mathbb{P}\left(\sup_{s \in \mathcal{F}_{>0}} P_n(Ks_M - Ks) > C\right) \leq 2n^{-\alpha}. \quad (141)$$

Finally, using (140) and (141) in (139) we get, for all $n \geq n_0(A_\infty, A_{cons}, n_1, A_+, \alpha)$,

$$\mathbb{P}(P_n(Ks_M - Ks_n) > C) \leq 3n^{-\alpha}.$$

Conclusion. To complete the proof of Theorem 3, just notice that by (111), (125) and (134) we can take

$$A_\infty = 64\sqrt{2}B_2A(\sigma_{\max} + 2A)\sigma_{\min}^{-1}r_M(\varphi)$$

and by (119), (133), (137) and (138),

$$A_0 = \max \left\{ 48\sqrt{\tilde{A}_0}, 12 \left(\sqrt{\tilde{A}_0} \vee \sqrt{A_5} \right), \sqrt{\hat{A}_0}, \sqrt{A_5} \right\}$$

is convenient. ■

Proof of Theorem 4. We localize our analysis in the subset

$$B_{(M, L_\infty)}(s_M, R_{n,D,\alpha}) = \{s \in M, \|s - s_M\|_\infty \leq R_{n,D,\alpha}\} \subset M.$$

Unlike in the proof of Theorem 3, see (99), we need not to consider the quantity $\tilde{R}_{n,D,\alpha}$, a radius possibly larger than $R_{n,D,\alpha}$. Indeed, the use of $\tilde{R}_{n,D,\alpha}$ rather than $R_{n,D,\alpha}$ in the proof of Theorem 3 is only needed in Lemma 14, where we derive a sharp lower bound for the mean of the supremum of the empirical process indexed by the contrasted functions centered by the contrasted projection over a slice of interest. To prove Theorem 4, we just need upper bounds, and Lemma 14 is avoided as well as the use of $\tilde{R}_{n,D,\alpha}$. Let us define several slices of excess risk on the model M : for any $C \geq 0$,

$$\begin{aligned}\mathcal{G}_C &= \{s \in M, P(Ks - Ks_M) \leq C\} \cap B_{(M, L_\infty)}(s_M, R_{n,D,\alpha}) , \\ \mathcal{G}_{>C} &= \{s \in M, P(Ks - Ks_M) > C\} \cap B_{(M, L_\infty)}(s_M, R_{n,D,\alpha}) .\end{aligned}$$

We also define, for all $U \geq 0$,

$$\mathcal{D}_U = \{s \in M, P(Ks - Ks_M) = U\} \cap B_{(M, L_\infty)}(s_M, R_{n,D,\alpha}) .$$

I. Proof of Inequality (27). Let $C_1 > 0$ to fixed later, satisfying

$$C_1 \geq \frac{D}{n} =: C_- > 0 . \quad (142)$$

We have by **(H5)**, for all $n \geq n_1$,

$$\mathbb{P}(P(Ks_n - Ks_M) > C_1) \leq \mathbb{P}\left(\{P(Ks_n - Ks_M) > C_1\} \cap \Omega_{\infty,\alpha}\right) + n^{-\alpha} \quad (143)$$

and also

$$\begin{aligned}& \mathbb{P}\left(\{P(Ks_n - Ks_M) > C_1\} \cap \Omega_{\infty,\alpha}\right) \\ & \leq \mathbb{P}\left(\inf_{s \in \mathcal{G}_{C_1}} P_n(Ks - Ks_M) \geq \inf_{s \in \mathcal{G}_{>C_1}} P_n(Ks - Ks_M)\right) \\ & = \mathbb{P}\left(\sup_{s \in \mathcal{G}_{C_1}} P_n(Ks_M - Ks) \leq \sup_{s \in \mathcal{G}_{>C_1}} P_n(Ks_M - Ks)\right) \\ & \leq \mathbb{P}\left(0 \leq \sup_{s \in \mathcal{G}_{>C_1}} P_n(Ks_M - Ks)\right) .\end{aligned} \quad (144)$$

Moreover, it holds

$$\begin{aligned}& \sup_{s \in \mathcal{G}_{>C_1}} P_n(Ks_M - Ks) \\ & = \sup_{s \in \mathcal{G}_{>C_1}} \{P_n(\psi_{1,M} \cdot (s_M - s) - \psi_2 \circ (s - s_M))\} \\ & = \sup_{s \in \mathcal{G}_{>C_1}} \{(P_n - P)(\psi_{1,M} \cdot (s_M - s)) - (P_n - P)(\psi_2 \circ (s - s_M)) - P(Ks - Ks_M)\} \\ & = \sup_{s \in \mathcal{G}_{>C_1}} \{(P_n - P)(\psi_{1,M} \cdot (s_M - s)) - P(Ks - Ks_M) - (P_n - P)(\psi_2 \circ (s - s_M))\} \\ & = \sup_{U > C_1} \sup_{s \in \mathcal{D}_U} \{(P_n - P)(\psi_{1,M} \cdot (s_M - s)) - U - (P_n - P)(\psi_2 \circ (s - s_M))\} \\ & \leq \sup_{U > C_1} \left\{ \sqrt{U} \sqrt{\sum_{k=1}^D (P_n - P)^2(\psi_{1,M} \cdot \varphi_k)} - U + \sup_{s \in \mathcal{G}_U} |(P_n - P)(\psi_2 \circ (s - s_M))| \right\} .\end{aligned} \quad (145)$$

Now, from inequality (164) of Lemma 13 applied with $\beta = \alpha$, we get

$$\mathbb{P}\left[\sqrt{\sum_{k=1}^D (P_n - P)^2(\psi_{1,M} \cdot \varphi_k)} \geq L_{A,A_{3,M},\alpha} \sqrt{\frac{D \vee \ln n}{n}}\right] \leq n^{-\alpha} . \quad (146)$$

In addition, we handle the empirical process indexed by the second order terms by straightforward modifications of Lemmas 16 and 17 as well as their proofs. It thus holds, by the same type of arguments as those given in Lemma 16,

$$\mathbb{E} \left[\sup_{s \in \mathcal{G}_{C_1}} |(P_n - P)(\psi_{2,M}^s \cdot (s - s_M))| \right] \leq 8 \sqrt{\frac{CD}{n}} R_{n,D,\alpha} . \quad (147)$$

Moreover, using (147), the same type of arguments as those leading to inequality (191) of Lemma 17, allow to show that for any $q \geq 1$ and $j \in \mathbb{N}^*$, for all $x > 0$,

$$\begin{aligned} & \mathbb{P} \left[\sup_{s \in \mathcal{G}_{q^j C_-}} |(P_n - P)(\psi_2 \circ (s - s_M))| \geq 16 \sqrt{\frac{q^j C_- D}{n}} R_{n,D,\alpha} + \sqrt{\frac{2R_{n,D,\alpha}^2 q^j C_- x}{n}} + \frac{8}{3} \frac{R_{n,D,\alpha}^2 x}{n} \right] \\ & \leq \exp(-x) . \end{aligned} \quad (148)$$

Hence, taking $x = \gamma \ln n$ in (148) and using the fact that $C_- = Dn^{-1} \geq n^{-1}$, we get

$$\mathbb{P} \left[\sup_{s \in \mathcal{G}_{q^j C_-}} |(P_n - P)(\psi_2 \circ (s - s_M))| \geq L_{A_{cons}, \gamma} R_{n,D,\alpha} \sqrt{\frac{q^j C_- (D \vee \ln n)}{n}} \right] \leq n^{-\gamma} . \quad (149)$$

Now, by straightforward modifications of the proof of Lemma 17, we get that for all $n \geq n_0(A_{cons})$,

$$\mathbb{P} \left[\forall U > C_-, \sup_{s \in \mathcal{G}_U} |(P_n - P)(\psi_2 \circ (s - s_M))| \leq L_{A_{cons}, \alpha} R_{n,D,\alpha} \sqrt{\frac{U (D \vee \ln n)}{n}} \right] \geq 1 - n^{-\alpha} . \quad (150)$$

Combining (145), (146) and (150), we have on an event of probability at least $1 - 2n^{-\alpha}$, for all $n \geq n_0(A_{cons})$,

$$\begin{aligned} \sup_{s \in \mathcal{G}_{>C_1}} P_n(Ks_M - Ks) & \leq \sup_{U > C_1} \left\{ L_{A, A_{3,M}, \alpha} \sqrt{\frac{U (D \vee \ln n)}{n}} - U + L_{A_{cons}, \alpha} R_{n,D,\alpha} \sqrt{\frac{U (D \vee \ln n)}{n}} \right\} \\ & \leq \sup_{U > C_1} \left\{ L_{A, A_{cons}, A_{3,M}, \alpha} (1 + R_{n,D,\alpha}) \sqrt{\frac{U (D \vee \ln n)}{n}} - U \right\} . \end{aligned} \quad (151)$$

Now, as $R_{n,D,\alpha} \leq A_{cons} (\ln n)^{-1/2}$, we deduce from (151) that for

$$C_1 = L_{A, A_{cons}, A_{3,M}, \alpha} \frac{D \vee \ln(n)}{n} > C_- \quad (152)$$

with $L_{A, A_{cons}, A_{3,M}, \alpha}$ large enough, it holds with probability at least $1 - 2n^{-\alpha}$ and for all $n \geq n_0(A_{cons})$,

$$\sup_{s \in \mathcal{G}_{>C_1}} P_n(Ks_M - Ks) < 0 ,$$

and so by using (143) and (144), this yields inequality (27).

II. Proof of Inequality (28). Let $C_2 > 0$ to fixed later, satisfying

$$C_2 \geq \frac{D}{n} = C_- > 0 . \quad (153)$$

We have by **(H5)**, for all $n \geq n_1$,

$$\mathbb{P}(P_n(Ks_M - Ks_n) > C_2) \leq \mathbb{P} \left(\{P_n(Ks_M - Ks_n) > C_2\} \cap \Omega_{\infty, \alpha} \right) + n^{-\alpha} . \quad (154)$$

Moreover, we have on $\Omega_{\infty, \alpha}$,

$$\begin{aligned} P_n(Ks_M - Ks_n) & = \sup_{s \in B_{(M, L_\infty)}(s_M, R_{n,D,\alpha})} P_n(Ks_M - Ks) \\ & = \max \left\{ \sup_{s \in \mathcal{G}_{C_1}} P_n(Ks_M - Ks) ; \sup_{s \in \mathcal{G}_{>C_1}} P_n(Ks_M - Ks) \right\} , \end{aligned} \quad (155)$$

where C_1 is defined in the first part of the proof dedicated to the establishment of inequality (27). Moreover, let us recall that in the first part of the proof, we have proved that an event of probability at least $1 - 2n^{-\alpha}$ exists, that we call Ω_1 , such that it holds on this event, for all $n \geq n_0(A_{cons})$,

$$\sqrt{\sum_{k=1}^D (P_n - P)^2 (\psi_{1,M} \cdot \varphi_k)} \leq L_{A,A_3,M,\alpha} \sqrt{\frac{D \vee \ln n}{n}}, \quad (156)$$

$$\forall U > C_-, \quad \sup_{s \in \mathcal{G}_U} |(P_n - P)(\psi_2 \circ (s - s_M))| \leq L_{A_{cons},\alpha} R_{n,D,\alpha} \sqrt{\frac{U(D \vee \ln n)}{n}}, \quad (157)$$

and

$$\sup_{s \in \mathcal{G}_{>C_1}} P_n(Ks_M - Ks) < 0. \quad (158)$$

By (155) and (158), we thus have on $\Omega_{\infty,\alpha} \cap \Omega_1$, for all $n \geq n_0(A_{cons})$,

$$0 \leq P_n(Ks_M - Ks_n) = \sup_{s \in \mathcal{G}_{C_1}} P_n(Ks_M - Ks). \quad (159)$$

In addition, it holds

$$\begin{aligned} & \sup_{s \in \mathcal{G}_{C_1}} P_n(Ks_M - Ks) \\ &= \sup_{s \in \mathcal{G}_{C_1}} \{P_n(\psi_{1,M} \cdot (s_M - s) - \psi_2 \circ (s - s_M))\} \\ &= \sup_{s \in \mathcal{G}_{C_1}} \{(P_n - P)(\psi_{1,M} \cdot (s_M - s)) - (P_n - P)(\psi_2 \circ (s - s_M)) - P(Ks - Ks_M)\} \\ &\leq \sup_{s \in \mathcal{G}_{C_1}} \{(P_n - P)(\psi_{1,M} \cdot (s_M - s))\} + \sup_{s \in \mathcal{G}_{C_1}} |(P_n - P)(\psi_2 \circ (s - s_M))|. \end{aligned} \quad (160)$$

Now, we have on Ω_1 , for all $n \geq n_0(A_{cons})$,

$$\begin{aligned} \sup_{s \in \mathcal{G}_{C_1}} \{(P_n - P)(\psi_{1,M} \cdot (s_M - s))\} &\leq \sqrt{C_1} \sqrt{\sum_{k=1}^D (P_n - P)^2 (\psi_{1,M} \cdot \varphi_k)} \\ &\leq L_{A,A_3,M,\alpha} \sqrt{\frac{C_1(D \vee \ln n)}{n}} \quad \text{by (156)} \\ &= L_{A,A_{cons},A_3,M,\alpha} \frac{D \vee \ln(n)}{n} \quad \text{by (152)}, \end{aligned} \quad (161)$$

and also, by (157) and (152),

$$\begin{aligned} \sup_{s \in \mathcal{G}_{C_1}} |(P_n - P)(\psi_2 \circ (s - s_M))| &\leq L_{A_{cons},\alpha} R_{n,D,\alpha} \sqrt{\frac{C_1(D \vee \ln n)}{n}} \\ &\leq L_{A,A_{cons},A_3,M,\alpha} R_{n,D,\alpha} \frac{D \vee \ln(n)}{n}. \end{aligned} \quad (162)$$

Finally, as $R_{n,D,\alpha} \leq A_{cons}(\ln n)^{-1/2}$, we deduce from (159), (160), (161) and (162), that it holds on $\Omega_{\infty,\alpha} \cap \Omega_1$, for all $n \geq n_0(A_{cons})$,

$$P_n(Ks_M - Ks_n) \leq L_{A,A_{cons},A_3,M,\alpha} \frac{D \vee \ln(n)}{n},$$

and so, this yields to inequality (28) by using (154) and this concludes the proof of Theorem 4. ■

6.4 Technical Lemmas

We state here some lemmas needed in the proofs of Section 6.3. First, in Lemmas 13, 14 and 15, we derive some controls, from above and from below, of the empirical process indexed by the “linear parts” of the contrasted functions over slices of interest. Secondly, we give upper bounds in Lemmas 16 and 17 for the empirical process indexed by the “quadratic parts” of the contrasted functions over slices of interest. And finally, we use all these results in Lemmas 18, 19 and 20 to derive upper and lower bounds for the empirical process indexed by the contrasted functions over slices of interest.

Lemma 13 *Assume that (H1), (H2) and (H3) hold. Then for any $\beta > 0$, by setting*

$$\tau_n = L_{A,A_3,M,\sigma_{\min},\beta} \left(\sqrt{\frac{\ln n}{D}} \vee \frac{\sqrt{\ln n}}{n^{1/4}} \right),$$

It holds, for any orthonormal basis $(\varphi_k)_{k=1}^D$ of $(M, \|\cdot\|_2)$,

$$\mathbb{P} \left[\sqrt{\sum_{k=1}^D (P_n - P)^2 (\psi_{1,M} \cdot \varphi_k)} \geq (1 + \tau_n) \sqrt{\frac{D}{n}} \mathcal{K}_{1,M} \right] \leq n^{-\beta}. \quad (163)$$

If (H1) and (H3) hold, then for any $\beta > 0$, it holds

$$\mathbb{P} \left[\sqrt{\sum_{k=1}^D (P_n - P)^2 (\psi_{1,M} \cdot \varphi_k)} \geq L_{A,A_3,M,\beta} \sqrt{\frac{D \vee \ln n}{n}} \right] \leq n^{-\beta}. \quad (164)$$

Proof. By Cauchy-Schwarz inequality we have

$$\chi_M := \sqrt{\sum_{k=1}^D (P_n - P)^2 (\psi_{1,M} \cdot \varphi_k)} = \sup_{s \in M, \|s\|_2 \leq 1} \{ |(P_n - P)(\psi_{1,M} \cdot s)| \}.$$

Hence, we get by Bousquet’s inequality (215) applied with $\mathcal{F} = \{\psi_{1,M} \cdot s ; s \in M, \|s\|_2 \leq 1\}$, for all $x > 0$, $\delta > 0$,

$$\mathbb{P} \left[\chi_M \geq \sqrt{2\sigma^2 \frac{x}{n}} + (1 + \delta) \mathbb{E}[\chi_M] + \left(\frac{1}{3} + \frac{1}{\delta} \right) \frac{bx}{n} \right] \leq \exp(-x) \quad (165)$$

where

$$\sigma^2 \leq \sup_{s \in M, \|s\|_2 \leq 1} P \left[(\psi_{1,M} \cdot s)^2 \right] \leq \|\psi_{1,M}\|_\infty^2 \leq 16A^2 \quad \text{by (102)}$$

and

$$b \leq \sup_{s \in M, \|s\|_2 \leq 1} \|\psi_{1,M} \cdot s - P(\psi_{1,M} \cdot s)\|_\infty \leq 4A\sqrt{D}A_{3,M} \quad \text{by (101), (102) and (105).}$$

Moreover,

$$\mathbb{E}[\chi_M] \leq \sqrt{\mathbb{E}[\chi_M^2]} = \sqrt{\frac{D}{n}} \mathcal{K}_{1,M}.$$

So, from (165) it follows that, for all $x > 0$, $\delta > 0$,

$$\mathbb{P} \left[\chi_M \geq \sqrt{32A^2 \frac{x}{n}} + (1 + \delta) \sqrt{\frac{D}{n}} \mathcal{K}_{1,M} + \left(\frac{1}{3} + \frac{1}{\delta} \right) \frac{4A\sqrt{D}A_{3,M}x}{n} \right] \leq \exp(-x). \quad (166)$$

Hence, taking $x = \beta \ln n$, $\delta = \frac{\sqrt{\ln n}}{n^{1/4}}$ in (166), we derive by (104) that a positive constant $L_{A,A_3,M,\sigma_{\min},\beta}$ exists such that

$$\mathbb{P} \left[\chi_M \geq \left(1 + L_{A,A_3,M,\sigma_{\min},\beta} \left(\sqrt{\frac{\ln n}{D}} \vee \frac{\sqrt{\ln n}}{n^{1/4}} \right) \right) \sqrt{\frac{D}{n}} \mathcal{K}_{1,M} \right] \leq n^{-\beta},$$

which yields inequality (163). By (103) we have $\mathcal{K}_{1,M} \leq 6A$, and by taking again $x = \beta \ln n$ and $\delta = \frac{\sqrt{\ln n}}{n^{1/4}}$ in (166), simple computations give

$$\mathbb{P} \left[\sqrt{\sum_{k=1}^D (P_n - P)^2 (\psi_{1,M} \cdot \varphi_k)} \geq L_{A,A_3,M,\beta} \left(\sqrt{\frac{D}{n}} \vee \sqrt{\frac{\ln n}{n}} \vee \sqrt{\frac{D \ln n}{n^{3/2}}} \right) \right] \leq n^{-\beta},$$

and by consequence, (164) follows. ■

In the next lemma, we state sharp lower bounds for the mean of the supremum of the empirical process on the linear parts of constraisted functions of M belonging to a slice of excess risk. This is done for a model of reasonable dimension.

Lemma 14 *Let $r > 1$ and $C > 0$. Assume that **(H1)**, **(H2)**, **(H4)** and (17) hold and let $\varphi = (\varphi_k)_{k=1}^D$ be an orthonormal basis of $(M, \|\cdot\|_2)$ satisfying **(H4)**. If positive constants A_-, A_+, A_l, A_u exist such that*

$$A_+ \frac{n}{(\ln n)^2} \geq D \geq A_- (\ln n)^2 \quad \text{and} \quad A_l \frac{D}{n} \leq rC \leq A_u \frac{D}{n},$$

and if the constant A_∞ defined in (99) satisfies

$$A_\infty \geq 64B_2A\sqrt{2A_u}\sigma_{\min}^{-1}r_M(\varphi), \quad (167)$$

then a positive constant $L_{A,A_l,A_u,\sigma_{\min}}$ exists such that, for all $n \geq n_0(A_-, A_+, A_u, A_l, A, B_2, r_M(\varphi), \sigma_{\min})$,

$$\mathbb{E} \left[\sup_{s \in \mathcal{F}_{(C,rC)}} (P_n - P) (\psi_{1,M} \cdot (s_M - s)) \right] \geq \left(1 - \frac{L_{A,A_l,A_u,\sigma_{\min}}}{\sqrt{D}} \right) \sqrt{\frac{rCD}{n}} \mathcal{K}_{1,M}. \quad (168)$$

Our argument leading to Lemma 14 shows that we have to assume that the constant A_∞ introduced in (99) is large enough. In order to prove Lemma 14 the following result is needed.

Lemma 15 *Let $r > 1$, $\beta > 0$ and $C \geq 0$. Assume that **(H1)**, **(H2)**, **(H4)** and (17) hold and let $\varphi = (\varphi_k)_{k=1}^D$ be an orthonormal basis of $(M, \|\cdot\|_2)$ satisfying **(H4)**. If positive constants A_+, A_- and A_u exist such that*

$$A_+ \frac{n}{(\ln n)^2} \geq D \geq A_- (\ln n)^2, \quad rC \leq A_u \frac{D}{n},$$

and if

$$A_\infty \geq 32B_2A\sqrt{2A_u}\beta\sigma_{\min}^{-1}r_M(\varphi)$$

then for all $n \geq n_0(A_-, A_+, A, B_2, r_M(\varphi), \sigma_{\min}, \beta)$, it holds

$$\mathbb{P} \left[\max_{k \in \{1, \dots, D\}} \left| \frac{\sqrt{rC} (P_n - P) (\psi_{1,M} \cdot \varphi_k)}{\sqrt{\sum_{j=1}^D (P_n - P)^2 (\psi_{1,M} \cdot \varphi_j)}} \right| \geq \frac{\tilde{R}_{n,D,\alpha}}{r_M(\varphi) \sqrt{D}} \right] \leq \frac{2D+1}{n^\beta}.$$

Proof of Lemma 15. By Cauchy-Schwarz inequality, we get

$$\chi_M = \sqrt{\sum_{k=1}^D (P_n - P)^2 (\psi_{1,M} \cdot \varphi_k)} = \sup_{s \in S_M} |(P_n - P) (\psi_{1,M} \cdot s)|,$$

where S_M is the unit sphere of M , that is

$$S_M = \left\{ s \in M, \ s = \sum_{k=1}^D \beta_k \varphi_k \text{ and } \sqrt{\sum_{k=1}^D \beta_k^2} = 1 \right\}.$$

Thus we can apply Klein-Rio's inequality (217) to χ_M by taking $\mathcal{F} = S_M$ and use the fact that

$$\sup_{s \in S_M} \|\psi_{1,M} \cdot s - P(\psi_{1,M} \cdot s)\|_\infty \leq 4A\sqrt{D}r_M(\varphi) \quad \text{by (101), (102) and (H4)}. \quad (169)$$

$$\sup_{s \in S_M} \text{Var}(\psi_{1,M} \cdot s) = \sup_{s \in S_M} P(\psi_{1,M} \cdot s)^2 \leq 16A^2 \quad \text{by (101), (102)}$$

and also, by using (169) in Inequality (212) applied to χ_M , we get that

$$\begin{aligned} \mathbb{E}[\chi_M] &\geq B_2^{-1} \sqrt{\mathbb{E}[\chi_M^2]} - \frac{4A\sqrt{D}r_M(\varphi)}{n} \\ &= B_2^{-1} \sqrt{\frac{D}{n} \mathcal{K}_{1,M}} - \frac{4A\sqrt{D}r_M(\varphi)}{n}. \end{aligned}$$

We thus obtain by (217), for all $\varepsilon, x > 0$,

$$\mathbb{P}\left(\chi_M \leq (1-\varepsilon) B_2^{-1} \sqrt{\frac{D}{n} \mathcal{K}_{1,M}} - \sqrt{32A^2 \frac{x}{n}} - \left(1-\varepsilon + \left(1 + \frac{1}{\varepsilon}\right)x\right) \frac{4A\sqrt{D}r_M(\varphi)}{n}\right) \leq \exp(-x). \quad (170)$$

So, by taking $\varepsilon = \frac{1}{2}$ and $x = \beta \ln n$ in (170), and by observing that $D \geq A_- (\ln n)^2$ and $\mathcal{K}_{1,M} \geq 2\sigma_{\min}$, we conclude that, for all $n \geq n_0(A_-, A, B_2, r_M(\varphi), \sigma_{\min}, \beta)$,

$$\mathbb{P}\left[\chi_M \leq \frac{B_2^{-1}}{8} \sqrt{\frac{D}{n} \mathcal{K}_{1,M}}\right] \leq n^{-\beta}. \quad (171)$$

Furthermore, combining Bernstein's inequality (213), with the observation that we have, for every $k \in \{1, \dots, D\}$,

$$\begin{aligned} \|\psi_{1,M} \cdot \varphi_k\|_\infty &\leq 4A\sqrt{D}r_M(\varphi) \quad \text{by (102) and (H4)} \\ P(\psi_{1,M} \cdot \varphi_k)^2 &\leq \|\psi_{1,M}\|_\infty^2 \leq 16A^2 \quad \text{by (102)} \end{aligned}$$

we get that, for every $x > 0$ and every $k \in \{1, \dots, D\}$,

$$\mathbb{P}\left[|(P_n - P)(\psi_{1,M} \cdot \varphi_k)| \geq \sqrt{32A^2 \frac{x}{n}} + \frac{4A\sqrt{D}r_M(\varphi)}{3} \frac{x}{n}\right] \leq 2 \exp(-x)$$

and so

$$\mathbb{P}\left[\max_{k \in \{1, \dots, D\}} |(P_n - P)(\psi_{1,M} \cdot \varphi_k)| \geq \sqrt{32A^2 \frac{x}{n}} + \frac{4A\sqrt{D}r_M(\varphi)}{3} \frac{x}{n}\right] \leq 2D \exp(-x). \quad (172)$$

Hence, taking $x = \beta \ln n$ in (172), it comes

$$\mathbb{P}\left[\max_{k \in \{1, \dots, D\}} |(P_n - P)(\psi_{1,M} \cdot \varphi_k)| \geq \sqrt{\frac{32A^2 \beta \ln n}{n}} + \frac{4A\sqrt{D}r_M(\varphi) \beta \ln n}{3n}\right] \leq \frac{2D}{n^\beta}, \quad (173)$$

then, by using (171) and (173), we get for all $n \geq n_0(A_-, A, B_2, r_M(\varphi), \sigma_{\min}, \beta)$,

$$\mathbb{P}\left[\max_{k \in \{1, \dots, D\}} \left| \frac{\sqrt{rC} (P_n - P)(\psi_{1,M} \cdot \varphi_k)}{\chi_M} \right| \geq \frac{8B_2 \sqrt{rC}}{\sqrt{\frac{D}{n} \mathcal{K}_{1,M}}} \left(\sqrt{\frac{32A^2 \beta \ln n}{n}} + \frac{4A\sqrt{D}r_M(\varphi) \beta \ln n}{3n} \right)\right] \leq \frac{2D+1}{n^\beta}.$$

Finally, as $A_+ \frac{n}{(\ln n)^2} \geq D$ we have, for all $n \geq n_0(A, A_+, r_M(\varphi), \beta)$,

$$\frac{4A\sqrt{D}r_M(\varphi) \beta \ln n}{3n} \leq \sqrt{\frac{32A^2 \beta \ln n}{n}}$$

and we can check that, since $rC \leq A_u \frac{D}{n}$ and $\mathcal{K}_{1,M} \geq 2\sigma_{\min}$, if

$$A_\infty \geq 32B_2\sqrt{2A_uA^2\beta\sigma_{\min}^{-1}}r_M(\varphi)$$

then, for all $n \geq n_0(A_-, A_+, A, B_2, r_M(\varphi), \sigma_{\min}, \beta)$,

$$\mathbb{P} \left[\max_{k \in \{1, \dots, D\}} \left| \frac{\sqrt{rC}(P_n - P)(\psi_{1,M} \cdot \varphi_k)}{\chi_M} \right| \geq \frac{A_\infty}{r_M(\varphi)} \sqrt{\frac{\ln n}{n}} \right] \leq \frac{2D+1}{n^\beta}$$

which readily gives the result. ■

We are now ready to prove the lower bound (168) for the expected value of the largest increment of the empirical process over $\mathcal{F}_{(C, rC]}$.

Proof of Lemma 14. Let us begin with the lower bound of

$$\mathbb{E}^{\frac{1}{2}} \left(\sup_{s \in \mathcal{F}_{(C, rC]}} (P_n - P)(\psi_{1,M} \cdot (s_M - s)) \right)^2,$$

a result that will be need further in the proof. Introduce for all $k \in \{1, \dots, D\}$,

$$\beta_{k,n} = \frac{\sqrt{rC}(P_n - P)(\psi_{1,M} \cdot \varphi_k)}{\sqrt{\sum_{j=1}^D (P_n - P)^2(\psi_{1,M} \cdot \varphi_j)}},$$

and observe that the excess risk on M of $(\sum_{k=1}^D \beta_{k,n} \varphi_k + s_M) \in M$ is equal to rC . We also set

$$\tilde{\Omega} = \left\{ \max_{k \in \{1, \dots, D\}} |\beta_{k,n}| \leq \frac{\tilde{R}_{n,D,\alpha}}{r_M(\varphi)\sqrt{D}} \right\}.$$

By Lemma 15 we have for all $\beta > 0$, if $A_\infty \geq 32B_2\sqrt{2A_uA^2\beta\sigma_{\min}^{-1}}r_M(\varphi)$ then, for all $n \geq n_0(A_-, A_+, A, B_2, r_M(\varphi), \sigma_{\min}, \beta)$,

$$\mathbb{P}(\tilde{\Omega}) \geq 1 - \frac{2D+1}{n^\beta}. \quad (174)$$

Moreover, by **(H4)**, we get on the event $\tilde{\Omega}$,

$$\left\| \sum_{k=1}^D \beta_{k,n} \varphi_k \right\|_\infty \leq \tilde{R}_{n,D,\alpha},$$

and so, on $\tilde{\Omega}$,

$$\left(s_M + \sum_{k=1}^D \beta_{k,n} \varphi_k \right) \in \mathcal{F}_{(C, rC]}. \quad (175)$$

As a consequence, by (175) it holds

$$\begin{aligned} & \mathbb{E}^{\frac{1}{2}} \left(\sup_{s \in \mathcal{F}_{(C, rC]}} (P_n - P)(\psi_{1,M} \cdot (s_M - s)) \right)^2 \\ & \geq \mathbb{E}^{\frac{1}{2}} \left[\left((P_n - P) \left(\psi_{1,M} \cdot \left(\sum_{k=1}^D \beta_{k,n} \varphi_k \right) \right) \right)^2 \mathbf{1}_{\tilde{\Omega}} \right] \\ & = \sqrt{rC} \sqrt{\mathbb{E} \left[\left(\sum_{k=1}^D (P_n - P)^2(\psi_{1,M} \cdot \varphi_k) \right) \mathbf{1}_{\tilde{\Omega}} \right]}. \end{aligned} \quad (176)$$

Furthermore, since by (101) $P(\psi_{1,M} \cdot \varphi_k) = 0$ and by **(H4)** $\|\varphi_k\|_\infty \leq \sqrt{D}r_M(\varphi)$ for all $k \in \{1, \dots, D\}$, we have

$$\begin{aligned} \left| \sum_{k=1}^D (P_n - P)^2(\psi_{1,M} \cdot \varphi_k) \right| &\leq D \max_{k=1, \dots, D} \left| (P_n - P)^2(\psi_{1,M} \cdot \varphi_k) \right| \\ &= D \max_{k=1, \dots, D} |P_n^2(\psi_{1,M} \cdot \varphi_k)| \\ &\leq D \max_{k=1, \dots, D} \|\psi_{1,M} \cdot \varphi_k\|_\infty^2 \\ &\leq 16A^2 D^2 r_M^2(\varphi) \end{aligned}$$

and it ensures

$$\mathbb{E} \left[\left(\sum_{k=1}^D (P_n - P)^2(\psi_{1,M} \cdot \varphi_k) \right) 1_{\tilde{\Omega}} \right] \geq \mathbb{E} \left[\left(\sum_{k=1}^D (P_n - P)^2(\psi_{1,M} \cdot \varphi_k) \right) \right] - 16A^2 D^2 r_M^2(\varphi) \mathbb{P} \left[\left(\tilde{\Omega} \right)^c \right]. \quad (177)$$

Comparing inequality (177) with (176) and using (174), we obtain the following lower bound for all $n \geq n_0(A_-, A_+, A, B_2, r_M(\varphi), \sigma_{\min}, \beta)$,

$$\begin{aligned} \mathbb{E}^{\frac{1}{2}} \left(\sup_{s \in \mathcal{F}_{(C, rC)}} (P_n - P)(\psi_{1,M} \cdot (s_M - s)) \right)^2 &\geq \sqrt{rC} \sqrt{\mathbb{E} \left[\left(\sum_{k=1}^D (P_n - P)^2(\psi_{1,M} \cdot \varphi_k) \right) \right]} \\ &\quad - 4Ar_M(\varphi) D \sqrt{rC} \sqrt{\mathbb{P} \left[\left(\tilde{\Omega} \right)^c \right]} \\ &\geq \sqrt{\frac{rCD}{n}} \mathcal{K}_{1,M} - 4Ar_M(\varphi) D \sqrt{rC} \sqrt{\frac{2D+1}{n^\beta}}. \end{aligned} \quad (178)$$

We take $\beta = 4$, and we must have

$$A_\infty \geq 64AB_2 \sqrt{2A_u} \sigma_{\min}^{-1} r_M(\varphi).$$

Since $D \leq A_+ n (\ln n)^{-2}$ and $\mathcal{K}_{1,M} \geq 2\sigma_{\min}$ under **(H2)**, we get, for all $n \geq n_0(A, A_+, r_M(\varphi), \sigma_{\min})$,

$$4Ar_M(\varphi) D \sqrt{rC} \sqrt{\frac{2D+1}{n^\beta}} \leq \frac{1}{\sqrt{D}} \times \sqrt{\frac{rCD}{n}} \mathcal{K}_{1,M} \quad (179)$$

and so, by combining (178) and (179), for all $n \geq n_0(A_-, A_+, A, B_2, r_M(\varphi), \sigma_{\min})$, it holds

$$\mathbb{E}^{\frac{1}{2}} \left(\sup_{s \in \mathcal{F}_{(C, rC)}} (P_n - P)(\psi_{1,M} \cdot (s_M - s)) \right)^2 \geq \left(1 - \frac{1}{\sqrt{D}} \right) \sqrt{\frac{rCD}{n}} \mathcal{K}_{1,M}. \quad (180)$$

Now, as $D \geq A_- (\ln n)^2$ we have for all $n \geq n_0(A_-)$, $D^{-1/2} \leq 1/2$. Moreover, we have $\mathcal{K}_{1,M} \geq 2\sigma_{\min}$ by **(H2)** and $rC \geq A_l D n^{-1}$, so we finally deduce from (180) that, for all $n \geq n_0(A_-, A_+, A, B_2, A_l, r_M(\varphi), \sigma_{\min})$,

$$\mathbb{E}^{\frac{1}{2}} \left(\sup_{s \in \mathcal{F}_{(C, rC)}} (P_n - P)(\psi_{1,M} \cdot (s_M - s)) \right)^2 \geq \sigma_{\min} \sqrt{A_l} \frac{D}{n}. \quad (181)$$

We turn now to the lower bound of $\mathbb{E} \left[\sup_{s \in \mathcal{F}_{(C, rC)}} (P_n - P)(\psi_{1,M} \cdot (s_M - s)) \right]$. First observe that $s \in \mathcal{F}_{(C, rC)}$ implies that $(2s_M - s) \in \mathcal{F}_{(C, rC)}$, so that

$$\mathbb{E} \left[\sup_{s \in \mathcal{F}_{(C, rC)}} (P_n - P)(\psi_{1,M} \cdot (s_M - s)) \right] = \mathbb{E} \left[\sup_{s \in \mathcal{F}_{(C, rC)}} |(P_n - P)(\psi_{1,M} \cdot (s_M - s))| \right]. \quad (182)$$

In the next step, we apply Corollary 27. More precisely, using notations of Corollary 27, we set

$$\mathcal{F} = \{ \psi_{1,M} \cdot (s_M - s) ; s \in \mathcal{F}_{(C,rC)} \}$$

and

$$Z = \sup_{s \in \mathcal{F}_{(C,rC)}} |(P_n - P)(\psi_{1,M} \cdot (s_M - s))| .$$

Now, since for all $n \geq n_0(A_+, A_-, A_\infty, A_{cons})$ we have $\tilde{R}_{n,D,\alpha} \leq 1$, we get by (101) and (102), for all $n \geq n_0(A_+, A_-, A_\infty, A_{cons})$,

$$\sup_{f \in \mathcal{F}} \|f - Pf\|_\infty = \sup_{s \in \mathcal{F}_{(C,rC)}} \|\psi_{1,M} \cdot (s_M - s)\|_\infty \leq 4A\tilde{R}_{n,D,\alpha} \leq 4A$$

we set $b = 4A$. Since we assume that $rC \leq A_u \frac{D}{n}$, it moreover holds by (102),

$$\sup_{f \in \mathcal{F}} \text{Var}(f) \leq \sup_{s \in \mathcal{F}_{(C,rC)}} P(\psi_{1,M} \cdot (s_M - s))^2 \leq 16A^2 rC \leq 16A^2 A_u \frac{D}{n}$$

and so we set $\sigma^2 = 16A^2 A_u \frac{D}{n}$. Now, by (181) we have, for all $n \geq n_0(A_-, A_+, A, B_2, A_l, r_M(\varphi), \sigma_{\min})$,

$$\sqrt{\mathbb{E}[Z^2]} \geq \sigma_{\min} \sqrt{A_l} \frac{D}{n} . \quad (183)$$

Hence, a positive constant $L_{A,A_l,A_u,\sigma_{\min}}(\max(4A\sqrt{A_u}A_l^{-1/2}\sigma_{\min}^{-1}; 2\sqrt{A}A_l^{-1/4}\sigma_{\min}^{-1/2}))$ holds) exists such that, by setting

$$\kappa_n = \frac{L_{A,A_l,A_u,\sigma_{\min}}}{\sqrt{D}}$$

we get, using (183), that, for all $n \geq n_0(A_-, A_+, A_l, A_u, A, B_2, r_M(\varphi), A_{cons}, \sigma_{\min})$,

$$\begin{aligned} \kappa_n^2 \mathbb{E}[Z^2] &\geq \frac{\sigma^2}{n} , \\ \kappa_n^2 \sqrt{\mathbb{E}[Z^2]} &\geq \frac{b}{n} . \end{aligned}$$

Furthermore, since $D \geq A_-(\ln n)^2$, we have for all $n \geq n_0(A_-, A, A_u, A_l, \sigma_{\min})$,

$$\kappa_n \in (0, 1) .$$

So, using (182) and Corollary 27, it holds for all $n \geq n_0(A_-, A_+, A_l, A_u, A, B_2, r_M(\varphi), \sigma_{\min})$,

$$\begin{aligned} &\mathbb{E} \left[\sup_{s \in \mathcal{F}_{(C,rC)}} (P_n - P)(\psi_{1,M} \cdot (s_M - s)) \right] \\ &\geq \left(1 - \frac{L_{A,A_l,A_u,\sigma_{\min}}}{\sqrt{D}} \right) \mathbb{E}^{\frac{1}{2}} \left(\sup_{s \in \mathcal{F}_{(C,rC)}} (P_n - P)(\psi_{1,M} \cdot (s_M - s)) \right)^2 . \end{aligned} \quad (184)$$

Finally, by comparing (180) and (184), we deduce that for all $n \geq n_0(A_-, A_+, A_l, A_u, A, B_2, r_M(\varphi), \sigma_{\min})$,

$$\mathbb{E} \left[\sup_{s \in \mathcal{F}_{(C,rC)}} (P_n - P)(\psi_{1,M} \cdot (s_M - s)) \right] \geq \left(1 - \frac{L_{A,A_l,A_u,\sigma_{\min}}}{\sqrt{D}} \right) \sqrt{\frac{rCD}{n}} \mathcal{K}_{1,M}$$

and so (168) is proved. ■

Let us now turn to the control of second order terms appearing in the expansion of the least-squares contrast, see (6). Let us define

$$\Omega_C(x) = \sup_{s \in \mathcal{F}_{(C,rC)}} \left\{ \frac{|\psi_2((s - s_M)(x)) - \psi_2((t - s_M)(x))|}{|s(x) - t(x)|} ; (s, t) \in \mathcal{F}_C, s(x) \neq t(x) \right\} .$$

After straightforward computations using that $\psi_2(t) = t^2$ for all $t \in \mathbb{R}$ and assuming **(H3)**, we get that, for all $x \in \mathcal{X}$,

$$\Omega_C(x) = 2 \sup_{s \in \mathcal{F}_C} \{|s(x) - s_M(x)|\} \quad (185)$$

$$\leq 2 \left(\tilde{R}_{n,D,\alpha} \wedge \sqrt{CD} A_{3,M} \right) . \quad (186)$$

Lemma 16 *Let $C \geq 0$. Under **(H3)**, it holds*

$$\mathbb{E} \left[\sup_{s \in \mathcal{F}_C} |(P_n - P)(\psi_2 \circ (s - s_M))| \right] \leq 8 \sqrt{\frac{CD}{n}} \left(\tilde{R}_{n,D,\alpha} \wedge \sqrt{CD} A_{3,M} \right) .$$

Proof. We define the Rademacher process \mathcal{R}_n on a class \mathcal{F} of measurable functions from \mathcal{X} to \mathbb{R} , to be

$$\mathcal{R}_n(f) = \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i) , \quad f \in \mathcal{F}$$

where ε_i are independent Rademacher random variables also independent from the X_i . By the usual symmetrization argument we have

$$\mathbb{E} \left[\sup_{s \in \mathcal{F}_C} |(P_n - P)(\psi_2 \circ (s - s_M))| \right] \leq 2 \mathbb{E} \left[\sup_{s \in \mathcal{F}_C} |\mathcal{R}_n(\psi_2 \circ (s - s_M))| \right] .$$

Taking the expectation with respect to the Rademacher variables, we get

$$\begin{aligned} & \mathbb{E}_\varepsilon \left[\sup_{s \in \mathcal{F}_C} |\mathcal{R}_n(\psi_2 \circ (s - s_M))| \right] \\ &= \mathbb{E}_\varepsilon \left[\sup_{s \in \mathcal{F}_C} \left| \mathcal{R}_n \left((s - s_M)^2 \right) \right| \right] \\ &\leq \left(\max_{1 \leq i \leq n} \Omega_C(X_i) \right) \mathbb{E}_\varepsilon \left[\sup_{s \in \mathcal{F}_C} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \varphi_i((s - s_M)(X_i)) \right| \right] \end{aligned} \quad (187)$$

where the functions $\varphi_i : \mathbb{R} \rightarrow \mathbb{R}$ are defined by

$$\varphi_i(t) = \begin{cases} (\Omega_C(X_i))^{-1} t^2 & \text{for } |t| \leq \sup_{s \in \mathcal{F}_C} \{|s(X_i) - s_M(X_i)|\} = \frac{\Omega_C(X_i)}{2} \\ \frac{1}{4} \Omega_C(X_i) & \text{otherwise} \end{cases}$$

Then by (185) we deduce that φ_i is a contraction mapping with $\varphi_i(0) = 0$. We thus apply Theorem 23 to get

$$\begin{aligned} \mathbb{E}_\varepsilon \left[\sup_{s \in \mathcal{F}_C} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \varphi_i((s - s_M)(X_i)) \right| \right] &\leq 2 \mathbb{E}_\varepsilon \left[\sup_{s \in \mathcal{F}_C} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i (s - s_M)(X_i) \right| \right] \\ &= 2 \mathbb{E}_\varepsilon \left[\sup_{s \in \mathcal{F}_C} |\mathcal{R}_n(s - s_M)| \right] \end{aligned} \quad (188)$$

and so we derive successively the following upper bounds in mean,

$$\begin{aligned}
& \mathbb{E} \left[\sup_{s \in \mathcal{F}_C} |\mathcal{R}_n(\psi_2 \circ (s - s_M))| \right] = \mathbb{E} \left[\mathbb{E}_\varepsilon \left[\sup_{s \in \mathcal{F}_C} |\mathcal{R}_n(\psi_2 \circ (s - s_M))| \right] \right] \\
& \leq \mathbb{E} \left[\left(\max_{1 \leq i \leq n} \Omega_C(X_i) \right) \mathbb{E}_\varepsilon \left[\sup_{s \in \mathcal{F}_C} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \varphi_i((s - s_M)(X_i)) \right| \right] \right] \quad \text{by (187)} \\
& \leq 2 \mathbb{E} \left[\left(\max_{1 \leq i \leq n} \Omega_C(X_i) \right) \mathbb{E}_\varepsilon \left[\sup_{s \in \mathcal{F}_C} |\mathcal{R}_n(s - s_M)| \right] \right] \quad \text{by (188)} \\
& = 2 \mathbb{E} \left[\left(\max_{1 \leq i \leq n} \Omega_C(X_i) \right) \sup_{s \in \mathcal{F}_C} |\mathcal{R}_n(s - s_M)| \right] \\
& \leq 2 \sqrt{\mathbb{E} \left[\max_{1 \leq i \leq n} \Omega_C^2(X_i) \right]} \sqrt{\mathbb{E} \left[\left(\sup_{s \in \mathcal{F}_C} |\mathcal{R}_n(s - s_M)| \right)^2 \right]}
\end{aligned}$$

We consider now an orthonormal basis of $(M, \|\cdot\|_2)$ and denote it by $(\varphi_k)_{k=1}^D$. Whence

$$\begin{aligned}
& \sqrt{\mathbb{E} \left[\left(\sup_{s \in \mathcal{F}_C} |\mathcal{R}_n(s - s_M)| \right)^2 \right]} \\
& \leq \sqrt{\mathbb{E} \left[\left(\sup \left\{ \left| \sum_{k=1}^D a_k \mathcal{R}_n(\varphi_k) \right| ; \sum_{k=1}^D a_k^2 \leq C \right\} \right)^2 \right]} \\
& = \sqrt{C} \sqrt{\mathbb{E} \left[\sum_{k=1}^D (\mathcal{R}_n(\varphi_k))^2 \right]} = \sqrt{\frac{CD}{n}},
\end{aligned}$$

to complete the proof, it remains to observe that, by (186),

$$\sqrt{\mathbb{E} \left[\max_{1 \leq i \leq n} \Omega_C^2(X_i) \right]} \leq 2 \left(\tilde{R}_{n,D,\alpha} \wedge \sqrt{CD} A_{3,M} \right).$$

■

In the following Lemma, we provide uniform upper bounds for the supremum of the empirical process of second order terms in the contrast expansion when the considered slices are not too small.

Lemma 17 *Let $A_+, A_-, A_l, \beta, C_- > 0$, and assume **(H3)** and (17). If $C_- \geq A_l \frac{D}{n}$ and $A_+ n (\ln n)^{-2} \geq D \geq A_- (\ln n)^2$, then a positive constant $L_{A_-, A_l, \beta}$ exists such that, for all $n \geq n_0(A_\infty, A_{cons}, A_+, A_l)$,*

$$\mathbb{P} \left[\forall C > C_-, \quad \sup_{s \in \mathcal{F}_C} |(P_n - P)(\psi_2 \circ (s - s_M))| \leq L_{A_-, A_l, \beta} \sqrt{\frac{CD}{n}} \tilde{R}_{n,D,\alpha} \right] \geq 1 - n^{-\beta}.$$

Proof. First notice that, as $A_+ n (\ln n)^{-2} \geq D$, we have by (17),

$$\tilde{R}_{n,D,\alpha} \leq \frac{\max \{A_{cons} ; A_\infty \sqrt{A_+}\}}{\sqrt{\ln n}}.$$

By consequence, for all $n \geq n_0(A_\infty, A_{cons}, A_+)$,

$$\tilde{R}_{n,D,\alpha} \leq 1. \tag{189}$$

Now, since $\cup_{C>C_-} \mathcal{F}_C \subset B_{(M,L_\infty)}(s_M, \tilde{R}_{n,D,\alpha})$ where

$$B_{(M,L_\infty)}(s_M, \tilde{R}_{n,D,\alpha}) = \left\{ s \in M, \|s - s_M\|_\infty \leq \tilde{R}_{n,D,\alpha} \right\},$$

we have by (189), for all $s \in \cup_{C>C_-} \mathcal{F}_C$ and for all $n \geq n_0(A_\infty, A_{cons}, A_+)$,

$$\begin{aligned} P(Ks - Ks_M) &= P\left[(s - s_M)^2\right] \\ &\leq \|s - s_M\|_\infty^2 \\ &\leq \tilde{R}_{n,D,\alpha}^2 \leq 1. \end{aligned}$$

We thus have, for all $n \geq n_0(A_\infty, A_{cons}, A_+)$,

$$\bigcup_{C>C_-} \mathcal{F}_C = \bigcup_{C_- \wedge 1 < C \leq 1} \mathcal{F}_C$$

and by monotonicity of the collection \mathcal{F}_C , for some $q > 1$ and $J = \left\lfloor \frac{|\ln(C_- \wedge 1)|}{\ln q} \right\rfloor + 1$, it holds

$$\bigcup_{C_- \wedge 1 < C \leq 1} \mathcal{F}_C \subset \bigcup_{j=0}^J \mathcal{F}_{q^j C_-}.$$

Simple computations show that, since $D \geq 1$ and $C_- \geq A_l \frac{D}{n} \geq \frac{A_l}{n}$, one can find a constant $L_{A_l, q}$ such that

$$J \leq L_{A_l, q} \ln n.$$

Moreover, by monotonicity of $C \mapsto \sup_{s \in \mathcal{F}_C} |(P_n - P)(\psi_2 \circ (s - s_M))|$, we have uniformly in $C \in (q^{j-1}C_-, q^j C_-]$,

$$\sup_{s \in \mathcal{F}_C} |(P_n - P)(\psi_2 \circ (s - s_M))| \leq \sup_{s \in \mathcal{F}_{q^{j+1}C_-}} |(P_n - P)(\psi_2 \circ (s - s_M))|.$$

Hence, taking the convention $\sup_{s \in \emptyset} |(P_n - P)(\psi_2 \circ (s - s_M))| = 0$, we get for all $n \geq n_0(A_\infty, A_{cons}, A_+)$ and any $L > 0$,

$$\begin{aligned} &\mathbb{P} \left[\forall C > C_-, \sup_{s \in \mathcal{F}_C} |(P_n - P)(\psi_2 \circ (s - s_M))| \leq L \sqrt{\frac{CD}{n}} \tilde{R}_{n,D,\alpha} \right] \\ &\geq \mathbb{P} \left[\forall j \in \{1, \dots, J\}, \sup_{s \in \mathcal{F}_{q^j C_-}} |(P_n - P)(\psi_2 \circ (s - s_M))| \leq L \sqrt{\frac{q^j C_- D}{n}} \tilde{R}_{n,D,\alpha} \right]. \end{aligned}$$

Now, for any $L > 0$,

$$\begin{aligned} &\mathbb{P} \left[\forall j \in \{1, \dots, J\}, \sup_{s \in \mathcal{F}_{q^j C_-}} |(P_n - P)(\psi_2 \circ (s - s_M))| \leq L \sqrt{\frac{q^j C_- D}{n}} \tilde{R}_{n,D,\alpha} \right] \\ &= 1 - \mathbb{P} \left[\exists j \in \{1, \dots, J\}, \sup_{s \in \mathcal{F}_{q^j C_-}} |(P_n - P)(\psi_2 \circ (s - s_M))| > L \sqrt{\frac{q^j C_- D}{n}} \tilde{R}_{n,D,\alpha} \right] \\ &\geq 1 - \sum_{j=1}^J \mathbb{P} \left[\sup_{s \in \mathcal{F}_{q^j C_-}} |(P_n - P)(\psi_2 \circ (s - s_M))| > L \sqrt{\frac{q^j C_- D}{n}} \tilde{R}_{n,D,\alpha} \right]. \end{aligned} \tag{190}$$

Given $j \in \{1, \dots, J\}$, Lemma 16 yields

$$\mathbb{E} \left[\sup_{s \in \mathcal{F}_{q^j C_-}} |(P_n - P)(\psi_2 \circ (s - s_M))| \right] \leq 8 \sqrt{\frac{q^j C_- D}{n}} \tilde{R}_{n,D,\alpha},$$

and next, we apply Bousquet's inequality (215) to handle the deviations around the mean. We have

$$\begin{aligned} & \sup_{s \in \mathcal{F}_{q^j C_-}} \|\psi_2 \circ (s - s_M) - P(\psi_2 \circ (s - s_M))\|_\infty \\ & \leq 2 \sup_{s \in \mathcal{F}_{q^j C_-}} \left\| (s - s_M)^2 \right\|_\infty \leq 2\tilde{R}_{n,D,\alpha}^2 \end{aligned}$$

and, for all $s \in \mathcal{F}_{q^j C_-}$,

$$\begin{aligned} & \text{Var}(\psi_2 \circ (s - s_M)) \\ & \leq P\left[(s - s_M)^4\right] \\ & \leq \|s - s_M\|_\infty^2 P\left[(s - s_M)^2\right] \\ & \leq \tilde{R}_{n,D,\alpha}^2 q^j C_- . \end{aligned}$$

It follows that, for $\varepsilon = 1$ and all $x > 0$,

$$\mathbb{P} \left[\sup_{s \in \mathcal{F}_{q^j C_-}} |(P_n - P)(\psi_2 \circ (s - s_M))| \geq 16\sqrt{\frac{q^j C_- D}{n}} \tilde{R}_{n,D,\alpha} + \sqrt{\frac{2\tilde{R}_{n,D,\alpha}^2 q^j C_- x}{n}} + \frac{8}{3} \frac{\tilde{R}_{n,D,\alpha}^2 x}{n} \right] \leq \exp(-x) . \quad (191)$$

By consequence, as $D \geq A_- (\ln n)^2$ and as $\tilde{R}_{n,D,\alpha} \leq 1$ for all $n \geq n_0(A_\infty, A_{\text{cons}}, A_+)$, taking $x = \gamma \ln n$ in (191) for some $\gamma > 0$, easy computations show that a positive constant $L_{A_-, A_l, \gamma}$ independent of j exists such that for all $n \geq n_0(A_\infty, A_{\text{cons}}, A_+)$,

$$\mathbb{P} \left[\sup_{s \in \mathcal{F}_{q^j C_-}} |(P_n - P)(\psi_2 \circ (s - s_M))| \geq L_{A_-, A_l, \gamma} \sqrt{\frac{q^j C_- D}{n}} \tilde{R}_{n,D,\alpha} \right] \leq \frac{1}{n^\gamma} .$$

Hence, using (190), we get for all $n \geq n_0(A_\infty, A_{\text{cons}}, A_+)$,

$$\begin{aligned} & \mathbb{P} \left[\forall C > C_-, \sup_{s \in \mathcal{F}_C} |(P_n - P)(\psi_2 \circ (s - s_M))| \leq L_{A_-, A_l, \gamma} \sqrt{\frac{CD}{n}} \tilde{R}_{n,D,\alpha} \right] \\ & \geq 1 - \frac{J}{n^\gamma} . \end{aligned}$$

And finally, as $J \leq L_{A_l, q} \ln n$, taking $\gamma = \beta + 1$ and $q = 2$ gives the result for all $n \geq n_0(A_\infty, A_{\text{cons}}, A_+, A_l)$. \blacksquare

Having controlled the residual empirical process driven by the remainder terms in the expansion of the contrast, and having proved sharp bounds for the expectation of the increments of the main empirical process on the slices, it remains to combine the above lemmas in order to establish the probability estimates controlling the empirical excess risk on the slices.

Lemma 18 *Let $\beta, A_-, A_+, A_l, C > 0$. Assume that **(H1)**, **(H2)**, **(H3)** and (17) hold. A positive constant A_4 exists, only depending on $A, A_{3,M}, \sigma_{\min}, \beta$, such that, if*

$$A_l \frac{D}{n} \leq C \leq \frac{1}{4} (1 + A_4 \nu_n)^2 \frac{D}{n} \mathcal{K}_{1,M}^2 \quad \text{and} \quad A_+ \frac{n}{(\ln n)^2} \geq D \geq A_- (\ln n)^2$$

where $\nu_n = \max \left\{ \sqrt{\frac{\ln n}{D}}, \sqrt{\frac{D \ln n}{n}}, R_{n,D,\alpha} \right\}$ is defined in (100), then for all $n \geq n_0(A_\infty, A_{\text{cons}}, A_+, A_l)$,

$$\mathbb{P} \left[\sup_{s \in \mathcal{F}_C} P_n(Ks_M - Ks) \geq (1 + L_{A_\infty, A, A_{3,M}, \sigma_{\min}, A_-, A_l, \beta} \times \nu_n) \sqrt{\frac{CD}{n}} \mathcal{K}_{1,M} - C \right] \leq 2n^{-\beta} .$$

Proof. Start with

$$\begin{aligned}
\sup_{s \in \mathcal{F}_C} P_n(Ks_M - Ks) &= \sup_{s \in \mathcal{F}_C} \{P_n(\psi_{1,M} \cdot (s_M - s) - \psi_2 \circ (s - s_M))\} \\
&= \sup_{s \in \mathcal{F}_C} \{(P_n - P)(\psi_{1,M} \cdot (s_M - s)) - (P_n - P)(\psi_2 \circ (s - s_M)) - P(Ks - Ks_M)\} \\
&\leq \sup_{s \in \mathcal{F}_C} \{(P_n - P)(\psi_{1,M} \cdot (s_M - s)) - P(Ks - Ks_M)\} \\
&\quad + \sup_{s \in \mathcal{F}_C} |(P_n - P)(\psi_2 \circ (s - s_M))|.
\end{aligned} \tag{192}$$

Next, recall that by definition,

$$D_L = \left\{ s \in B_{(M, L_\infty)}(s_M, \tilde{R}_{n,D,\alpha}), P(Ks - Ks_M) = L \right\},$$

so we have

$$\begin{aligned}
&\sup_{s \in \mathcal{F}_C} \{(P_n - P)(\psi_{1,M} \cdot (s_M - s)) - P(Ks - Ks_M)\} \\
&= \sup_{0 \leq L \leq C} \sup_{s \in D_L} \{(P_n - P)(\psi_{1,M} \cdot (s_M - s)) - L\} \\
&\leq \sup_{0 \leq L \leq C} \left\{ \sqrt{L} \sqrt{\sum_{k=1}^D (P_n - P)^2(\psi_{1,M} \cdot \varphi_k)} - L \right\}
\end{aligned}$$

where the last bound follows from Cauchy-Schwarz inequality. Hence, we deduce from Lemma 13 that

$$\mathbb{P} \left[\sup_{s \in \mathcal{F}_C} \{(P_n - P)(\psi_{1,M} \cdot (s_M - s)) - P(Ks - Ks_M)\} \geq \sup_{0 \leq L \leq C} \left\{ \sqrt{L} (1 + \tau_n) \sqrt{\frac{D}{n}} \mathcal{K}_{1,M} - L \right\} \right] \leq n^{-\beta}, \tag{193}$$

where

$$\begin{aligned}
\tau_n &= L_{A, A_3, M, \sigma_{\min}, \beta} \left(\sqrt{\frac{\ln n}{D}} \vee \frac{\sqrt{\ln n}}{n^{1/4}} \right) \\
&\leq L_{A, A_3, M, \sigma_{\min}, \beta} \left(\sqrt{\frac{\ln n}{D}} \vee \sqrt{\frac{D \ln n}{n}} \right) \\
&\leq L_{A, A_3, M, \sigma_{\min}, \beta} \times \nu_n.
\end{aligned} \tag{194}$$

So, injecting (194) in (193) we have

$$\mathbb{P} \left[\sup_{s \in \mathcal{F}_C} \{(P_n - P)(\psi_{1,M} \cdot (s_M - s)) - P(Ks - Ks_M)\} \geq \sup_{0 \leq L \leq C} \left\{ \sqrt{L} (1 + L_{A, A_3, M, \sigma_{\min}, \beta} \times \nu_n) \sqrt{\frac{D}{n}} \mathcal{K}_{1,M} - L \right\} \right] \leq n^{-\beta}$$

and since we assume $C \leq \frac{1}{4} (1 + L_{A, A_3, M, \sigma_{\min}, \beta} \times \nu_n)^2 \frac{D}{n} \mathcal{K}_{1,M}^2$ we see that

$$\sup_{0 \leq L \leq C} \left\{ \sqrt{L} (1 + L_{A, A_3, M, \sigma_{\min}, \beta} \nu_n) \sqrt{\frac{D}{n}} \mathcal{K}_{1,M} - L \right\} = \sqrt{C} (1 + L_{A, A_3, M, \sigma_{\min}, \beta} \times \nu_n) \sqrt{\frac{D}{n}} \mathcal{K}_{1,M} - C$$

and therefore

$$\mathbb{P} \left[\sup_{s \in \mathcal{F}_C} \{(P_n - P)(\psi_{1,M} \cdot (s_M - s)) - P(Ks - Ks_M)\} \geq (1 + L_{A, A_3, M, \sigma_{\min}, \beta} \nu_n) \sqrt{\frac{CD}{n}} \mathcal{K}_{1,M} - C \right] \leq n^{-\beta}. \tag{195}$$

Moreover, as $C \geq A_l \frac{D}{n}$, we derive from Lemma 17 that it holds, for all $n \geq n_0(A_\infty, A_{cons}, A_+, A_l)$,

$$\mathbb{P} \left[\sup_{s \in \mathcal{F}_C} |(P_n - P)(\psi_2 \circ (s - s_M))| \geq L_{A_-, A_l, \beta} \sqrt{\frac{CD}{n}} \tilde{R}_{n,D,\alpha} \right] \leq n^{-\beta}. \quad (196)$$

Finally, noticing that

$$\begin{aligned} \tilde{R}_{n,D,\alpha} &= \max \left\{ R_{n,D,\alpha}, A_\infty \sqrt{\frac{D \ln n}{n}} \right\} \\ &\leq L_{A_\infty, \sigma_{\min}} \max \left\{ R_{n,D,\alpha}, \sqrt{\frac{D \ln n}{n}} \right\} \times \mathcal{K}_{1,M} \quad \text{by (104)} \\ &\leq L_{A_\infty, \sigma_{\min}} \times \nu_n \times \mathcal{K}_{1,M}, \end{aligned}$$

we deduce from (196) that, for all $n \geq n_0(A_\infty, A_{cons}, A_+, A_l)$,

$$\mathbb{P} \left[\sup_{s \in \mathcal{F}_C} |(P_n - P)(\psi_2 \circ (s - s_M))| \geq L_{A_\infty, \sigma_{\min}, A_-, A_l, \beta} \times \nu_n \sqrt{\frac{CD}{n}} \mathcal{K}_{1,M} \right] \leq n^{-\beta} \quad (197)$$

and the conclusion follows by making use of (195) and (197) in inequality (192). ■

The second deviation bound for the empirical excess risk we need to establish on the upper slice is proved in a similar way.

Lemma 19 *Let $\beta, A_-, A_+, C \geq 0$. Assume that (H1), (H2), (H3) and (17) hold. A positive constant A_5 , depending on $A, A_{3,M}, A_\infty, \sigma_{\min}, A_-$ and β , exists such that, if it holds*

$$C \geq \frac{1}{4} (1 + A_5 \nu_n)^2 \frac{D}{n} \mathcal{K}_{1,M}^2 \quad \text{and} \quad A_+ \frac{n}{(\ln n)^2} \geq D \geq A_- (\ln n)^2$$

where $\nu_n = \max \left\{ \sqrt{\frac{\ln n}{D}}, \sqrt{\frac{D \ln n}{n}}, R_{n,D,\alpha} \right\}$ is defined in (100), then for all $n \geq n_0(A_\infty, A_{cons}, A_+)$,

$$\mathbb{P} \left[\sup_{s \in \mathcal{F}_{>C}} P_n(Ks_M - Ks) \geq (1 + A_5 \nu_n) \sqrt{\frac{CD}{n}} \mathcal{K}_{1,M} - C \right] \leq 2n^{-\beta}.$$

Moreover, when we only assume $C \geq 0$, we have for all $n \geq n_0(A_\infty, A_{cons}, A_+)$,

$$\mathbb{P} \left[\sup_{s \in \mathcal{F}_{>C}} P_n(Ks_M - Ks) \geq \frac{1}{4} (1 + A_5 \nu_n)^2 \frac{D}{n} \mathcal{K}_{1,M}^2 \right] \leq 2n^{-\beta}. \quad (198)$$

Proof. First observe that

$$\begin{aligned} \sup_{s \in \mathcal{F}_{>C}} P_n(Ks_M - Ks) &= \sup_{s \in \mathcal{F}_{>C}} \{P_n(\psi_{1,M} \cdot (s_M - s) - \psi_2 \circ (s - s_M))\} \\ &= \sup_{s \in \mathcal{F}_{>C}} \{(P_n - P)(\psi_{1,M} \cdot (s_M - s)) - (P_n - P)(\psi_2 \circ (s - s_M)) - P(Ks - Ks_M)\} \\ &= \sup_{s \in \mathcal{F}_{>C}} \{(P_n - P)(\psi_{1,M} \cdot (s_M - s)) - P(Ks - Ks_M) - (P_n - P)(\psi_2 \circ (s - s_M))\} \\ &= \sup_{L > C} \sup_{s \in D_L} \{(P_n - P)(\psi_{1,M} \cdot (s_M - s)) - L - (P_n - P)(\psi_2 \circ (s - s_M))\} \\ &\leq \sup_{L > C} \left\{ \sqrt{L} \sqrt{\sum_{k=1}^D (P_n - P)^2(\psi_{1,M} \cdot \varphi_k) - L} + \sup_{s \in \mathcal{F}_L} |(P_n - P)(\psi_2 \circ (s - s_M))| \right\} \end{aligned} \quad (199)$$

where the last bound follows from Cauchy-Schwarz inequality. Now, the end of the proof is similar to that of Lemma 18 and follows from the same kind of computations. Indeed, from Lemma 13 we deduce that

$$\mathbb{P} \left[\sqrt{\sum_{k=1}^D (P_n - P)^2 (\psi_{1,M} \cdot \varphi_k)} \geq (1 + L_{A,A_3,M,\sigma_{\min},\beta} \times \nu_n) \sqrt{\frac{D}{n}} \mathcal{K}_{1,M} \right] \leq n^{-\beta} \quad (200)$$

and, since

$$C \geq \frac{1}{4} \frac{D}{n} \mathcal{K}_{1,M}^2 \geq \sigma_{\min}^2 \frac{D}{n},$$

we apply Lemma 17 with $A_l = \sigma_{\min}^2$, and deduce that, for all $n \geq n_0(A_\infty, A_{\text{cons}}, A_+)$,

$$\mathbb{P} \left[\forall L > C, \sup_{s \in \mathcal{F}_L} |(P_n - P)(\psi_{2,M}^s \cdot (s - s_M))| \geq L_{A_\infty, \sigma_{\min}, A_-, \beta} \times \nu_n \sqrt{\frac{LD}{n}} \mathcal{K}_{1,M} \right] \leq n^{-\beta}. \quad (201)$$

Now using (200) and (201) in (199) we obtain, for all $n \geq n_0(A_\infty, A_{\text{cons}}, A_+)$,

$$\mathbb{P} \left[\sup_{s \in \mathcal{F}_{>C}} P_n(Ks_M - Ks) \geq \sup_{L > C} \left\{ (1 + L_{A,A_3,M,A_\infty, \sigma_{\min}, A_-, \beta} \times \nu_n) \sqrt{\frac{LD}{n}} \mathcal{K}_{1,M} - L \right\} \right] \leq 2n^{-\beta} \quad (202)$$

and we set $A_5 = L_{A,A_3,M,A_\infty, \sigma_{\min}, A_-, \beta}$ where $L_{A,A_3,M,A_\infty, \sigma_{\min}, A_-, \beta}$ is the constant in (202). For $C \geq \frac{1}{4} (1 + A_5 \nu_n)^2 \frac{D}{n} \mathcal{K}_{1,M}^2$ we get

$$\sup_{L > C} \left\{ \sqrt{L} (1 + A_5 \nu_n) \sqrt{\frac{D}{n}} \mathcal{K}_{1,M} - L \right\} = (1 + A_5 \nu_n) \sqrt{\frac{CD}{n}} \mathcal{K}_{1,M} - C$$

and by consequence,

$$\mathbb{P} \left[\sup_{s \in \mathcal{F}_{>C}} P_n(Ks_M - Ks) \geq (1 + A_5 \nu_n) \sqrt{\frac{CD}{n}} \mathcal{K}_{1,M} - C \right] \leq 2n^{-\beta},$$

which gives the first part of the lemma. The second part comes from (202) and the fact that, for any value of $C \geq 0$,

$$\sup_{L > C} \left\{ \sqrt{L} (1 + A_5 \nu_n) \sqrt{\frac{D}{n}} \mathcal{K}_{1,M} - L \right\} \leq (1 + A_5 \nu_n)^2 \frac{D}{4n} \mathcal{K}_{1,M}^2.$$

■

Lemma 20 Let $r > 1$ and $C, \beta > 0$. Assume that **(H1)**, **(H2)**, **(H4)** and (17) hold and let $\varphi = (\varphi_k)_{k=1}^D$ be an orthonormal basis of $(M, \|\cdot\|_2)$ satisfying **(H4)**. If positive constants A_-, A_+, A_l, A_u exist such that

$$A_+ \frac{n}{(\ln n)^2} \geq D \geq A_- (\ln n)^2 \quad \text{and} \quad A_l \frac{D}{n} \leq rC \leq A_u \frac{D}{n},$$

and if the constant A_∞ defined in (99) satisfies

$$A_\infty \geq 64B_2 A \sqrt{2A_u} \sigma_{\min}^{-1} r_M(\varphi),$$

then a positive constant $L_{A_-, A_l, A_u, A, A_\infty, \sigma_{\min}, r_M(\varphi), \beta}$ exists such that, for all $n \geq n_0(A_-, A_+, A_u, A_l, A, A_\infty, A_{\text{cons}}, B_2, r_M(\varphi), \sigma_{\min})$,

$$\mathbb{P} \left(\sup_{s \in \mathcal{F}_{(C, rC)}} P_n(Ks_M - Ks) \leq (1 - L_{A_-, A_l, A_u, A, A_\infty, \sigma_{\min}, r_M(\varphi), \beta} \times \nu_n) \sqrt{\frac{rCD}{n}} \mathcal{K}_{1,M} - rC \right) \leq 2n^{-\beta},$$

where $\nu_n = \max \left\{ \sqrt{\frac{\ln n}{D}}, \sqrt{\frac{D \ln n}{n}}, R_{n,D,\alpha} \right\}$ is defined in (100).

Proof. Start with

$$\begin{aligned}
& \sup_{s \in \mathcal{F}_{(C, rC]}} P_n (Ks_M - Ks) \\
&= \sup_{s \in \mathcal{F}_{(C, rC]}} \{ (P_n - P) (Ks_M - Ks) + P (Ks_M - Ks) \} \\
&\geq \sup_{s \in \mathcal{F}_{(C, rC]}} (P_n - P) (\psi_{1,M} \cdot (s_M - s)) - \sup_{s \in \mathcal{F}_{(C, rC]}} (P_n - P) (\psi_2 \circ (s - s_M)) - \sup_{s \in \mathcal{F}_{(C, rC]}} P (Ks - Ks_M) \\
&\geq \sup_{s \in \mathcal{F}_{(C, rC]}} (P_n - P) (\psi_{1,M} \cdot (s_M - s)) - \sup_{s \in \mathcal{F}_{rC}} (P_n - P) (\psi_2 \circ (s - s_M)) - rC
\end{aligned} \tag{203}$$

and set

$$\begin{aligned}
S_{1,r,C} &= \sup_{s \in \mathcal{F}_{(C, rC]}} (P_n - P) (\psi_{1,M} \cdot (s_M - s)) \\
M_{1,r,C} &= \mathbb{E} \left[\sup_{s \in \mathcal{F}_{(C, rC]}} (P_n - P) (\psi_{1,M} \cdot (s_M - s)) \right] \\
b_{1,r,C} &= \sup_{s \in \mathcal{F}_{(C, rC]}} \left\| \psi_{1,M} \cdot (s_M - s) - P (\psi_{1,M} \cdot (s_M - s)) \right\|_{\infty} \\
\sigma_{1,r,C}^2 &= \sup_{s \in \mathcal{F}_{(C, rC]}} \text{Var} (\psi_{1,M} \cdot (s_M - s)).
\end{aligned}$$

By Klein-Rio's Inequality (217), we get, for all $\delta, x > 0$,

$$\mathbb{P} \left(S_{1,r,C} \leq (1 - \delta) M_{1,r,C} - \sqrt{\frac{2\sigma_{1,r,C}^2 x}{n}} - \left(1 + \frac{1}{\delta} \right) \frac{b_{1,r,C} x}{n} \right) \leq \exp(-x). \tag{204}$$

Then, notice that all conditions of Lemma 14 are satisfied, and that it gives by (168), for all $n \geq n_0(A_-, A_+, A_u, A_l, A, B_2, r_M(\varphi))$

$$M_{1,r,C} \geq \left(1 - \frac{L_{A,A_l,A_u,\sigma_{\min}}}{\sqrt{D}} \right) \sqrt{\frac{rCD}{n}} \mathcal{K}_{1,M}. \tag{205}$$

In addition, observe that

$$\sigma_{1,r,C}^2 \leq \sup_{s \in \mathcal{F}_{(C, rC]}} P (\psi_{1,M}^2 \cdot (s_M - s)^2) \leq 16A^2 rC \quad \text{by (102)} \tag{206}$$

and

$$b_{1,r,C} = \sup_{s \in \mathcal{F}_{(C, rC]}} \left\| \psi_{1,M} \cdot (s_M - s) \right\|_{\infty} \leq 4Ar_M(\varphi) \sqrt{rCD} \quad \text{by (102) and (H4)} \tag{207}$$

Hence, using (205), (206) and (207) in inequality (204), we get for all $x > 0$ and all $n \geq n_0(A_-, A_+, A_u, A_l, A, B_2, r_M(\varphi), \sigma_{\min})$,

$$\begin{aligned}
& \mathbb{P} \left(S_{1,r,C} \leq (1 - \delta) \left(1 - \frac{L_{A,A_l,A_u,\sigma_{\min}}}{\sqrt{D}} \right) \sqrt{\frac{rCD}{n}} \mathcal{K}_{1,M} - \sqrt{\frac{32A^2 rCx}{n}} - \left(1 + \frac{1}{\delta} \right) \frac{4Ar_M(\varphi) \sqrt{rCD} x}{n} \right) \\
& \leq \exp(-x).
\end{aligned}$$

Now, taking $x = \beta \ln n$, $\delta = \frac{\sqrt{\ln n}}{n^{1/4}}$ and using (104), we deduce by simple computations that for all $n \geq n_0(A_-, A_+, A_u, A_l, A, B_2, r_M(\varphi), \sigma_{\min})$,

$$\mathbb{P} \left(S_{1,r,C} \leq \left(1 - L_{A,A_l,A_u,\sigma_{\min},r_M(\varphi),\beta} \times \left(\sqrt{\frac{\ln n}{D}} \vee \frac{\sqrt{\ln n}}{n^{1/4}} \right) \right) \sqrt{\frac{rCD}{n}} \mathcal{K}_{1,M} \right) \leq n^{-\beta} \tag{208}$$

and as

$$\sqrt{\frac{\ln n}{D}} \vee \frac{\sqrt{\ln n}}{n^{1/4}} \leq \sqrt{\frac{\ln n}{D}} \vee \sqrt{\frac{D \ln n}{n}} \leq \nu_n$$

(208) gives, for all $n \geq n_0(A_-, A_+, A_u, A_l, A, B_2, r_M(\varphi), \sigma_{\min})$,

$$\mathbb{P} \left(S_{1,r,C} \leq (1 - L_{A,A_l,A_u,\sigma_{\min},r_M(\varphi),\beta} \times \nu_n) \sqrt{\frac{rCD}{n}} \mathcal{K}_{1,M} \right) \leq n^{-\beta}. \quad (209)$$

Moreover, from Lemma 17 we deduce that, for all $n \geq n_0(A_\infty, A_{\text{cons}}, A_+, A_l)$,

$$\mathbb{P} \left[\sup_{s \in \mathcal{F}_{rC}} |(P_n - P)(\psi_2 \circ (s - s_M))| \geq L_{A_-, A_l, \beta} \sqrt{\frac{rCD}{n}} \tilde{R}_{n,D,\alpha} \right] \leq n^{-\beta} \quad (210)$$

and noticing that

$$\begin{aligned} \tilde{R}_{n,D,\alpha} &= \max \left\{ R_{n,D,\alpha} ; A_\infty \sqrt{\frac{D \ln n}{n}} \right\} \\ &\leq L_{A_\infty, \sigma_{\min}} \max \left\{ R_{n,D,\alpha} ; \sqrt{\frac{D \ln n}{n}} \right\} \times \mathcal{K}_{1,M} \quad \text{by (104)} \\ &\leq L_{A_\infty, \sigma_{\min}} \times \nu_n \times \mathcal{K}_{1,M}, \end{aligned}$$

we deduce from (210) that for all $n \geq n_0(A_\infty, A_{\text{cons}}, A_+, A_l)$,

$$\mathbb{P} \left[\sup_{s \in \mathcal{F}_{rC}} |(P_n - P)(\psi_2 \circ (s - s_M))| \geq L_{A_-, A_l, A_\infty, \sigma_{\min}, \beta} \times \nu_n \times \sqrt{\frac{rCD}{n}} \mathcal{K}_{1,M} \right] \leq n^{-\beta}. \quad (211)$$

Finally, using (209) and (211) in (203) we get that,

for all $n \geq n_0(A_-, A_+, A_u, A_l, A, A_\infty, A_{\text{cons}}, B_2, r_M(\varphi), \sigma_{\min})$,

$$\mathbb{P} \left(\sup_{s \in \mathcal{F}_{(C, rC)}} P_n(Ks_M - Ks) \leq (1 - L_{A_-, A_l, A_u, A, A_\infty, \sigma_{\min}, r_M(\varphi), \beta} \times \nu_n) \sqrt{\frac{rCD}{n}} \mathcal{K}_{1,M} - rC \right) \leq 2n^{-\beta},$$

which concludes the proof. ■

6.5 Probabilistic Tools

We recall here the main probabilistic results that are instrumental in our proofs.

Let us begin with the L_p -version of Hoffmann-Jørgensen's inequality, that can be found for example in [21], Proposition 6.10, p.157.

Theorem 21 *For any independent mean zero random variables Y_j , $j = 1, \dots, n$ taking values in a Banach space $(\mathcal{B}, \|\cdot\|)$ and satisfying $\mathbb{E}[\|Y_j\|^p] < +\infty$ for some $p \geq 1$, we have*

$$\mathbb{E}^{1/p} \left\| \sum_{j=1}^n Y_j \right\|^p \leq B_p \left(\mathbb{E} \left\| \sum_{j=1}^n Y_j \right\| + \mathbb{E}^{1/p} \left(\max_{1 \leq j \leq n} \|Y_j\| \right)^p \right)$$

where B_p is a universal constant depending only on p .

We will use this theorem for $p = 2$ in order to control suprema of empirical processes. In order to be more specific, let \mathcal{F} be a class of measurable functions from a measurable space \mathcal{Z} to \mathbb{R} and (X_1, \dots, X_n) be independent variables of common law P taking values in \mathcal{Z} . We then denote by $\mathcal{B} = l^\infty(\mathcal{F})$ the space of uniformly bounded functions on \mathcal{F} and, for any $b \in \mathcal{B}$, we set $\|b\| = \sup_{f \in \mathcal{F}} |b(f)|$. Thus $(\mathcal{B}, \|\cdot\|)$ is a Banach

space. Indeed we shall apply Theorem 21 to the independent random variables, with mean zero and taking values in \mathcal{B} , defined by

$$Y_j = \{f(X_j) - Pf, f \in \mathcal{F}\}.$$

More precisely, we will use the following result, which is a straightforward application of Theorem 21. Denote by

$$P_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$$

the empirical measure associated to the sample (X_1, \dots, X_n) and by

$$\|P_n - P\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} |(P_n - P)(f)|$$

the supremum of the empirical process over \mathcal{F} .

Corollary 22 *If \mathcal{F} is a class of measurable functions from a measurable space \mathcal{Z} to \mathbb{R} satisfying*

$$\sup_{z \in \mathcal{Z}} \sup_{f \in \mathcal{F}} |f(z) - Pf| = \sup_{f \in \mathcal{F}} \|f - Pf\|_{\infty} < +\infty$$

and (X_1, \dots, X_n) are n i.i.d. random variables taking values in \mathcal{Z} , then an absolute constant B_2 exists such that,

$$\mathbb{E}^{1/2} \left[\|P_n - P\|_{\mathcal{F}}^2 \right] \leq B_2 \left(\mathbb{E} [\|P_n - P\|_{\mathcal{F}}] + \frac{\sup_{f \in \mathcal{F}} \|f - Pf\|_{\infty}}{n} \right). \quad (212)$$

Another tool we need is a comparison theorem for Rademacher processes, see Theorem 4.12 of [21]. A function $\varphi : \mathbb{R} \rightarrow \mathbb{R}$ is called a contraction if $|\varphi(u) - \varphi(v)| \leq |u - v|$ for all $u, v \in \mathbb{R}$. Moreover, for a subset $T \subset \mathbb{R}^n$ we set

$$\|h(t)\|_T = \|h\|_T = \sup_{t \in T} |h(t)|.$$

Theorem 23 *Let $(\varepsilon_1, \dots, \varepsilon_n)$ be n i.i.d. Rademacher variables and $F : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ be a convex and increasing function. Furthermore, let $\varphi_i : \mathbb{R} \rightarrow \mathbb{R}$, $i \leq n$, be contractions such that $\varphi_i(0) = 0$. Then, for any bounded subset $T \subset \mathbb{R}^n$,*

$$\mathbb{E} F \left(\left\| \sum_i \varepsilon_i \varphi_i(t_i) \right\|_T \right) \leq 2 \mathbb{E} F \left(\left\| \sum_i \varepsilon_i t_i \right\|_T \right).$$

The next tool is the well known Bernstein's inequality, that can be found for example in [23], Proposition 2.9.

Theorem 24 (Bernstein's inequality) *Let (X_1, \dots, X_n) be independent real valued random variables and define*

$$S = \frac{1}{n} \sum_{i=1}^n (X_i - \mathbb{E}[X_i]).$$

Assuming that

$$v = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[X_i^2] < \infty$$

and

$$X_i \leq b \text{ a.s.}$$

we have, for every $x > 0$,

$$\mathbb{P} \left[|S| \geq \sqrt{2v \frac{x}{n}} + \frac{bx}{3n} \right] \leq 2 \exp(-x). \quad (213)$$

We now turn to concentration inequalities for the empirical process around its mean. Bousquet's inequality [8] provides optimal constants for the deviations above the mean. Klein-Rio's inequality [13] gives sharp constants for the deviations below the mean, that slightly improves Klein's inequality [14].

Theorem 25 Let (ξ_1, \dots, ξ_n) be n i.i.d. random variables having common law P and taking values in a measurable space \mathcal{Z} . If \mathcal{F} is a class of measurable functions from \mathcal{Z} to \mathbb{R} satisfying

$$|f(\xi_i) - Pf| \leq b \quad \text{a.s., for all } f \in \mathcal{F}, i \leq n,$$

then, by setting

$$\sigma_{\mathcal{F}}^2 = \sup_{f \in \mathcal{F}} \left\{ P(f^2) - (Pf)^2 \right\},$$

we have, for all $x \geq 0$,

Bousquet's inequality :

$$\mathbb{P} \left[\|P_n - P\|_{\mathcal{F}} - \mathbb{E}[\|P_n - P\|_{\mathcal{F}}] \geq \sqrt{2(\sigma_{\mathcal{F}}^2 + 2b\mathbb{E}[\|P_n - P\|_{\mathcal{F}}]) \frac{x}{n}} + \frac{bx}{3n} \right] \leq \exp(-x) \quad (214)$$

and we can deduce that, for all $\varepsilon, x > 0$, it holds

$$\mathbb{P} \left[\|P_n - P\|_{\mathcal{F}} - \mathbb{E}[\|P_n - P\|_{\mathcal{F}}] \geq \sqrt{2\sigma_{\mathcal{F}}^2 \frac{x}{n}} + \varepsilon \mathbb{E}[\|P_n - P\|_{\mathcal{F}}] + \left(\frac{1}{\varepsilon} + \frac{1}{3} \right) \frac{bx}{n} \right] \leq \exp(-x). \quad (215)$$

Klein-Rio's inequality :

$$\mathbb{P} \left[\mathbb{E}[\|P_n - P\|_{\mathcal{F}}] - \|P_n - P\|_{\mathcal{F}} \geq \sqrt{2(\sigma_{\mathcal{F}}^2 + 2b\mathbb{E}[\|P_n - P\|_{\mathcal{F}}]) \frac{x}{n}} + \frac{bx}{n} \right] \leq \exp(-x) \quad (216)$$

and again, we can deduce that, for all $\varepsilon, x > 0$, it holds

$$\mathbb{P} \left[\mathbb{E}[\|P_n - P\|_{\mathcal{F}}] - \|P_n - P\|_{\mathcal{F}} \geq \sqrt{2\sigma_{\mathcal{F}}^2 \frac{x}{n}} + \varepsilon \mathbb{E}[\|P_n - P\|_{\mathcal{F}}] + \left(\frac{1}{\varepsilon} + 1 \right) \frac{bx}{n} \right] \leq \exp(-x). \quad (217)$$

The following result is due to Ledoux [20]. We will use it along the proofs through Corollary 27 which is stated below. From now on, we set for short $Z = \|P_n - P\|_{\mathcal{F}}$.

Theorem 26 Let (ξ_1, \dots, ξ_n) be independent random with values in some measurable space $(\mathcal{Z}, \mathcal{T})$ and \mathcal{F} be some countable class of real-valued measurable functions from \mathcal{Z} . Let (ξ'_1, \dots, ξ'_n) be independent from (ξ_1, \dots, ξ_n) and with the same distribution. Setting

$$v = \mathbb{E} \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n (f(\xi_i) - f(\xi'_i))^2 \right]$$

then

$$\mathbb{E}[Z^2] - \mathbb{E}[Z]^2 \leq \frac{v}{n}.$$

Corollary 27 Under notations of Theorem 25, if some $\varkappa_n \in (0, 1)$ exists such that

$$\varkappa_n^2 \mathbb{E}[Z^2] \geq \frac{\sigma^2}{n}$$

and

$$\varkappa_n^2 \sqrt{\mathbb{E}[Z^2]} \geq \frac{b}{n}$$

then we have, for a numerical constant $A_{1,-}$,

$$(1 - \varkappa_n A_{1,-}) \sqrt{\mathbb{E}[Z^2]} \leq \mathbb{E}[Z].$$

Proof of Corollary 27. Just use Theorem 26, noticing the fact that

$$\sqrt{\mathbb{E}[Z^2]} - \mathbb{E}[Z] \leq \sqrt{\mathbb{V}(Z)}$$

and that, with notations of Theorem 26,

$$v \leq 2\sigma^2 + 32b\mathbb{E}[Z] .$$

The result then follows from straightforward calculations. ■

References

- [1] Sylvain Arlot and Pascal Massart. Data-driven calibration of penalties for least-squares regression. *J. Mach. Learn. Res.*, 10:245–279 (electronic), 2009.
- [2] A.R. Barron. Approximation and estimation bounds for artificial neural networks. *Mach. Learning*, 14:115–133, 1994.
- [3] Peter L. Bartlett and Shahar Mendelson. Empirical Minimization. *Probab. Theory Related Fields*, 135(3):311–334, 2006.
- [4] Lucien Birgé and Pascal Massart. Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields*, 138(1-2):33–73, 2007.
- [5] L. Birgé and P. Massart. Rates of convergence for minimum contrast estimators. *Probab. Theory Related Fields*, 97:113–150, 1993.
- [6] L. Birgé and P. Massart. Minimum contrast estimators on sieves: Exponential bounds and rates of convergence. *Bernoulli*, 4(3):329–375, 1998.
- [7] S. Boucheron and P. Massart. A high dimensional Wilks phenomenon. *Probability Theory and Related Fields*, 2010. To appear.
- [8] Olivier Bousquet. A Bennett concentration inequality and its application to suprema of empirical processes. *C. R. Math. Acad. Sci. Paris*, 334(6):495–500, 2002.
- [9] N. N. Čencov. *Statistical decision rules and optimal inference*, volume 53 of *Translations of Mathematical Monographs*. American Mathematical Society, Providence, R.I., 1982. Translation from the Russian edited by Lev J. Leifman.
- [10] Michel Crouzeix and Alain L. Mignot. *Analyse numérique des équations différentielles*. Collection Mathématiques Appliquées pour la Maîtrise. [Collection of Applied Mathematics for the Master’s Degree]. Masson, Paris, 1984.
- [11] E. Giné and V. Koltchinskii. Concentration inequalities and asymptotic results for ratio type empirical processes. *Ann. Probab.*, 33:1143–1216, 2006.
- [12] Ulf Grenander. *Abstract inference*. New York: Wiley, 1981.
- [13] R. Klein and E. Rio. Concentration around the mean for maxima of empirical processes. *Annals of Probability*, 1:63–87 (electronic), 2005.
- [14] T. Klein. Une inégalité de concentration à gauche pour les processus empiriques. *C.R. Acad. Sci. Paris, Ser I*, 334:500–505, 2002.
- [15] Vladimir Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimisation. *Ann. Statist.*, 34(6):2593–2656, 2006.

- [16] Charles Kooperberg and Charles J. Stone. A study of logspline density estimation. *Comput. Statist. Data Anal.*, 12(3):327–347, 1991.
- [17] L.M. Le Cam. Convergence of estimates under dimensionality restrictions. *Ann. Statist.*, 1:38–53, 1973.
- [18] L.M. Le Cam. *Asymptotic Methods in Statistical Decision Theory*. Springer-Verlag, New York, 1986.
- [19] L.M. Le Cam and G.L. Yang. *Asymptotics in Statistics : Some Basic Concepts*. Springer-Verlag, New York, 1990.
- [20] M. Ledoux. On Talagrand’s deviation inequalities for product measures. *ESAIM: Probability and Statistics*, 1:63–87, 1996.
- [21] M. Ledoux and M. Talagrand. *Probability in Banach spaces*. Springer, Berlin, 1991.
- [22] E. Mammen and A. Tsybakov. Smooth discrimination analysis. *Ann.Stat.*, 27:1808–1829, 1999.
- [23] P. Massart. *Concentration Inequalities and Model Selection*. Springer-Verlag, 2007.
- [24] P. Massart and E. Nédélec. Risks bounds for statistical learning. *Ann.Stat.*, 34(5):2326–2366, 2006.
- [25] Xiatong Shen and Wing Hung Wong. Convergence rate of sieve estimates. *Ann.Stat.*, 22(2):580–615, 1994.
- [26] Charles J. Stone. Uniform error bounds involving logspline models. In *Probability, statistics, and mathematics*, pages 335–355. Academic Press, Boston, MA, 1989.
- [27] Charles J. Stone. Large-sample inference for log-spline models. *Ann. Statist.*, 18(2):717–741, 1990.
- [28] Charles J. Stone. The use of polynomial splines and their tensor products in multivariate function estimation. *Ann. Statist.*, 22(1):118–171, 1994.
- [29] Charles J. Stone. Nonparametric M -regression with free knot splines. *J. Statist. Plann. Inference*, 130(1-2):182–206, 2005.
- [30] A. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Ann.Stat.*, 32:135–166, 2004.
- [31] Sara van de Geer. Estimating a regression function. *Ann. Statist.*, 18:907–924, 1990.
- [32] Sara van de Geer. The method of sieves and minimum contrast estimators. *Mathematical Methods of Statistics*, 4:20–28, 1995.
- [33] Sara van de Geer. M -estimation using penalties or sieves. *Journal of Statistical Planning and Inference*, 108(1-2):55 – 69, 2002.
- [34] V. Vapnik. *Statistical Learning Theory*. John Wiley and Sons, New York, 1998.
- [35] V. Vapnik and A. Chervonenkis. Theory of pattern recognition. *Nauka, Moscow*, 1974.
- [36] Wing Hung Wong and Xiatong Shen. Probability inequalities for likelihood ratios and convergence rates of sieve mles. *Ann.Statist.*, 23:339–362, 1995.