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

Adrien Saumard

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

February 24, 2012

Abstract

We consider the estimation of a bounded regression function with nonparametric heteroscedastic noise and random design. We study the true and empirical excess risks of the least-squares estimator on finite-dimensional vector spaces. We give upper and lower bounds on these quantities that are nonasymptotic and optimal to first order, allowing the dimension to depend on sample size. 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 with the projection of the regression function onto the considered model. Consistency in the 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, Sup-norm, Localized basis, Empirical process.

1 Introduction

A few years ago, Birgé and Massart [6] introduced a data-driven calibration method for penalized criteria in model selection, called the *Slope Heuristics*. Their algorithm is based on the concept of the minimal penalty, under which a model selection procedure fails. Given the shape of the ideal penalty, which in their Gaussian setting is a known function of the dimension of the considered models, the algorithm first provides a data-driven estimate of the minimal penalty. This is done by taking advantage of a sudden change in the behavior of the model selection procedure around this level of penalty. Then, the algorithm selects a model by using a penalty that is twice the estimated minimal penalty. Birgé and Massart prove in [6] that an asymptotically optimal penalty is twice the minimal one, in the sense that the associated selected model achieves a nonasymptotic oracle inequality with leading constant converging to one when the sample size tends to infinity.

The slope heuristics algorithm has been recently extended by Arlot and Massart [2] to the selection of M-estimators, whenever the number of models is not more than polynomial in the sample size. Arlot and Massart highlight that, in this context, the mean of the empirical excess risk on each model should be a good, rather general candidate for the - unknown - minimal penalty. In addition, they note that an optimal penalty is roughly given by the sum of the true and the empirical excess risks on each model. A key fact underlying the asymptotic optimality of the slope heuristics algorithm is the equivalence - in the sense that the ratio tends to one when the sample size tends to infinity - between the true and empirical excess risk, for each model which is likely to be selected. Generally, these models are of moderate dimension, typically between $(\log(n))^c$ and $n/(\log(n))^c$, where c is a positive constant and n is the sample size. This equivalence leads, quite straightforwardly, to the factor two between the minimal penalty and the optimal one.

Arlot and Massart prove in [2], by considering the selection of finite-dimensional models of histograms in heteroscedastic regression with a random design, that the slope heuristics algorithm is asymptotically optimal.

*Research partly supported by the french Agence Nationale de la Recherche (ANR 2011 BS01 010 01 projet Calibration).

The authors conjecture in [2], Section 1, that the restriction to histograms is "mainly technical", and that the slope heuristics "remains valid at least in the general least squares regression framework".

The first motivation of the present paper is thus to tackle the challenging mathematical problem raised by Arlot and Massart in [2], concerning the validity slope heuristics. More precisely, we isolate the question of the equivalence, for a fixed model, between the true and empirical excess risks. As emphasized in [2], this constitutes the principal part of the conjecture, since other arguments leading to model selection results are now well understood. We thus postpone model selection issues to a forthcoming paper, and focus on the fixed model case.

We consider least squares regression with heteroscedastic noise and random design, using a finite-dimensional linear model. Our analysis is nonasymptotic in the sense that our results are available for a fixed value of the sample size. It is also worth noticing that the dimension of the considered model is allowed to depend on the sample size and consequently is not treated as a constant of the problem. In order to determine the possible equivalence between the true and empirical excess risks, we investigate upper and lower deviation bounds for each quantity. We obtain first order optimal bounds, thus exhibiting the first part of the asymptotic expansion of the excess risks. This requires to determine not only the right rates of convergence, but also the optimal constant on the leading order term. We give two examples of models that satisfy our conditions: models of histograms and models of piecewise polynomials, whenever the partition defining these models satisfy some regularity condition with respect to the unknown distribution of data. Our results concerning histograms roughly recover those derived for a fixed model by Arlot and Massart [2], but with different techniques. Moreover, the case of piecewise polynomials strictly extend these results, and thus tends to confirm Arlot and Massart conjecture on the validity of the slope heuristics.

We believe that our deviation bounds, especially those concerning the true excess risk, are interesting by themselves. Indeed, the optimization of the excess risk is, from a general perspective, at the core of many nonparametric approaches, especially those related to statistical learning theory. Hence, any sharp control of this quantity is likely to be useful in many contexts.

In the general bounded M-estimation framework, rates of convergence and upper bounds for the excess risk are now well understood, see [18], [17], [13], [4], [10]. However, the values of the constants in these deviation bounds are suboptimal - or even unknown -, due in particular to the use of chaining techniques. Concerning lower deviation bounds, there is no convincing contribution to our knowledge, except the work of Bartlett and Mendelson [4], where an additional assumption on the behavior of underlying empirical process is used to derive such a result. However, this assumption is in general hard to check.

More specific frameworks, such as least squares regression with a fixed design on linear models (see for instance [6], [3] and [1]), least squares estimation of density on linear models (see [**?**] and references therein), or least squares regression on histograms as in [2], allow for sharp, explicit computations that lead to optimal upper and lower bounds for the excess risks. Hence, a natural question is: is there a framework, between the general one and the special cases, that would allow to derive deviation bounds that are optimal at the first order ? In other words, how far could optimal results concerning deviation bounds been extended ? The results presented in this article can be seen as a first attempt to answer these questions.

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. In Section 3, we detail the main steps of our approach at a heuristic level and give a summary of the results presented in the paper. We then derive some general results in Section 4. These theorems are then applied to the case of histograms and piecewise polynomials in Sections 5 and 6 respectively, where in particular, 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 $\xi_i = (X_i, Y_i) \in \mathcal{X} \times \mathbb{R}$, $i \in \{1, ..., n\}$, are n i.i.d. observations with distribution P. The marginal law of X_i is denoted by P^X. We assume that the

data satisfy the following relation

$$Y_i = s_* (X_i) + \sigma (X_i) \varepsilon_i \ , \tag{1}$$

where $s_* \in L_2\left(P^X\right)$, ε_i are i.i.d. random variables with mean 0 and variance 1 conditionally to X_i and $\sigma : \mathcal{X} \longrightarrow \mathbb{R}$ is a heteroscedastic noise level. A generic random variable of law P, independent of $(\xi_1, ..., \xi_n)$, is denoted by $\xi = (X, Y)$.
Hence, s_* is the regression function of Y with respect to X, to be estimated. Given a finite dimensional linear vector space M, that we will call a "model", we denote by s_M the linear projection of s_* onto M in $L_2\left(P^X\right)$ and by D the linear dimension of M.
We consider on the model M a least squares estimator s_n (possibly non unique), defined as follows

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

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 $K : L_2\left(P^X\right) \longrightarrow L_1 (P)$ the least squares contrast, defined by

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

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

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

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 $L_2\left(P^X\right)$ of the mean of the contrast over the unknown law P,

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

where

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

is called the risk of the function s. In particular we have $PKs_* = \mathbb{E}\left[\sigma^2 (X)\right]$. We first notice that for any $s \in L_2\left(P^X\right)$, 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,

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

The quantity $PKs - PKs_*$ is called the excess risk of s. Now, if we denote by s_M the linear projection of s_* onto M in $L_2\left(P^X\right)$, we have

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

and for all $s \in M$

$$P^X\left(s \cdot (s_M - s_*)\right) = 0 \ . \tag{5}$$

From (4), we deduce that

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

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

$$P\left(Ks_n - Ks_*\right) = P\left(Ks_n - Ks_M\right) + P\left(Ks_M - Ks_*\right)$$

we naturally focus on the quantity

$$P\left(Ks_n - Ks_M\right) \geq 0$$

that we want to bound from upper and from below, with high probability. We will often call this last quantity the excess risk of the estimator on M or the true excess risk of s_n, in opposition to the empirical excess risk for which the expectation is taken over the empirical measure,

$$P_n\left(Ks_M - Ks_n\right) \geq 0 \ .$$

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

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

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

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

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

Proof. Start with

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

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

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

We have

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

and, by (5),

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

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

3 Outline of the approach

Having introduced the framework and notations in Section 2 above, we are now able to explain more precisely the major steps of our approach to the problem of deriving optimal upper and lower bounds for the excess risks. As mentioned in the introduction, one of our main motivations is to determine whether the true excess risk is equivalent to the empirical one or not:

$$P\left(Ks_n - Ks_M\right) \sim P_n\left(Ks_M - Ks_n\right) \ ? \tag{11}$$

Indeed, such an equivalence is a keystone to justify the slope heuristics, a data-driven calibration method first proposed by Birgé and Massart [6] in a Gaussian setting and then extended by Arlot and Massart [2] to the selection of M-estimators.
The goal of this section is twofold. Firstly, it helps the reader to understand the role of the assumptions made in the forthcoming sections. Secondly, it provides an outline of the proof of our main result, Theorem 2 below. We suggest the reader interested in our proofs to read this section before entering the proofs.
We start by rewriting the lower and upper bound problems, for the true and empirical excess risks. Let C and α be two positive numbers. The question of bounding the true excess risk from upper and with high probability can be stated as follows: find, at a fixed $\alpha > 0$, the smallest $C > 0$ such that

$$\mathbb{P}\left[P\left(Ks_n - Ks_M\right) > C\right] \leq n^{-\alpha} \ .$$

We then write, by definition of the M-estimator s_n as a minimizer of the empirical excess risk over the model M,

$$\begin{aligned} & \mathbb{P}\left[P\left(Ks_n - Ks_M\right) > C\right] \\ \leq \ & \mathbb{P}\left[\inf_{s \in M_C} P_n\left(Ks - Ks_M\right) \geq \inf_{s \in M_{>C}} P_n\left(Ks - Ks_M\right)\right] \\ = \ & \mathbb{P}\left[\sup_{s \in M_C} P_n\left(Ks_M - Ks\right) \leq \sup_{s \in M_{>C}} P_n\left(Ks_M - Ks\right)\right] \ , \end{aligned} \tag{12}$$

where

$$M_C := \left\{s \in M \ ; \ P\left(Ks - Ks_M\right) \leq C\right\}$$

and

$$M_{>C} := M \backslash M_C = \left\{s \in M \ ; \ P\left(Ks - Ks_M\right) > C\right\}$$

are subsets of the model M, localized in terms of excess risk. As a matter of fact, M_C is the closed ball of radius C in $(M, \|\cdot\|_2)$. In the same manner, the question of bounding the true excess risk from below and with high probability is formalized as follows: find the larger $C > 0$ such that

$$\mathbb{P}\left[P\left(Ks_n - Ks_M\right) \leq C\right] \leq n^{-\alpha} \ .$$

We then have, by definition of the M-estimator s_n,

$$\begin{aligned} & \mathbb{P}\left[P\left(Ks_n - Ks_M\right) \leq C\right] \\ \leq \ & \mathbb{P}\left[\inf_{s \in M_C} P_n\left(Ks - Ks_M\right) \leq \inf_{s \in M_{>C}} P_n\left(Ks - Ks_M\right)\right] \\ = \ & \mathbb{P}\left[\sup_{s \in M_C} P_n\left(Ks_M - Ks\right) \geq \sup_{s \in M_{>C}} P_n\left(Ks_M - Ks\right)\right] \ . \end{aligned} \tag{13}$$

Expressions obtained in (12) and (13) allow to reduce both upper and lower bounds problems for the excess risk to the comparison of two quantities of interest,

$$\sup_{s \in M_C} P_n\left(Ks_M - Ks\right) \quad \text{and} \quad \sup_{s \in M_{>C}} P_n\left(Ks_M - Ks\right) \ .$$

Moreover, by setting $\mathcal{D}_L = \{s \in M \ ; \ P(Ks_n - Ks_M) = L\}$, we get

$$\begin{aligned}
\sup_{s \in M_C} P_n(Ks_M - Ks) &= \sup_{0 \le L \le C} \left\{ \sup_{s \in \mathcal{D}_L} P_n(Ks_M - Ks) \right\} \\
&= \sup_{0 \le L \le C} \left\{ \sup_{s \in \mathcal{D}_L} \{(P_n - P)(Ks_M - Ks) + P(Ks_M - Ks)\} \right\} \\
&= \sup_{0 \le L \le C} \left\{ \sup_{s \in \mathcal{D}_L} \{(P_n - P)(Ks_M - Ks)\} - L \right\} \qquad (14)
\end{aligned}$$

and also

$$\sup_{s \in M_{>C}} P_n(Ks_M - Ks) = \sup_{L > C} \left\{ \sup_{s \in \mathcal{D}_L} \{(P_n - P)(Ks_M - Ks)\} - L \right\} . \qquad (15)$$

The study of the excess risk thus reduces to the control of the following suprema, on the spheres $\mathcal{D}_L$ of radius L in $(M, \|\cdot\|_2)$, of the empirical process indexed by contrasted increments of functions in M around the projection s_M of the target,

$$\sup_{s \in \mathcal{D}_L} \{(P - P_n)(Ks - Ks_M)\} \ , \ L \ge 0 . \qquad (16)$$

Similarly, the empirical excess risk can be written, by definition of the M-estimator s_n,

$$\begin{aligned}
P_n(Ks_M - Ks_n) &= \sup_{s \in M} P_n(Ks_M - Ks) \\
&= \sup_{L \ge 0} \left\{ \sup_{s \in \mathcal{D}_L} P_n(Ks_M - Ks) \right\} \\
&= \sup_{L \ge 0} \left\{ \sup_{s \in \mathcal{D}_L} \{(P_n - P)(Ks_M - Ks)\} - L \right\} . \qquad (17)
\end{aligned}$$

Hence, the study of the empirical excess risk reduces again to the control of the quantities given in (16). As these quantities are (local) suprema of an empirical process, we can handle, under the right hypotheses, the deviations from their mean *via* the use of concentration inequalities - deviations from the right being described with optimal constants by Bousquet inequality (Bousquet, [8], recalled in Section 7.5 at the end of the present paper) and deviations from left being controlled with sharp constants by Klein and Rio inequality (Klein and Rio [12], also recalled in Section 7.5). We can thus expect that, under standard assumptions, the deviations are negligible compared to the means with large enough probability, at least for radii L not too small,

$$\sup_{s \in \mathcal{D}_L} \{(P - P_n)(Ks - Ks_M)\} \sim \mathbb{E}\left[\sup_{s \in \mathcal{D}_L} \{(P - P_n)(Ks - Ks_M)\} \right] . \qquad (18)$$

Remark 1 *It is worth noting that the above computations, which allow to investigate both upper and lower bound problems, only rely on the definition of s_n as a minimizer of the empirical risk over the model M, and not on the particular structure of the least squares contrast. Thus, formula (12), (13), (14), (15) and (17) are general facts of M-estimation - whenever the projection s_M of the target onto the model M exists. Moreover, although presented in a quite different manner, our computations related to the control of the true excess risk are in essence very similar to those developed by Bartlett and Mendelson in [4], concerning what they call "a direct analysis of the empirical minimization algorithm". Indeed, the authors highlight in Section 3 of [4] that, under rather mild hypotheses, the true excess risk is essentially the maximizer of the function $\mathcal{V}_n(L) - L$, where we set*

$$\mathcal{V}_n(L) := \mathbb{E}\left[\sup_{s \in \mathcal{D}_L} \{(P - P_n)(Ks - Ks_M)\} \right] .$$

Now, combining (12), (13), (14) and (15), it is easily seen that in the case where s_n is unique and where

$$\forall C \ge 0, \ \sup_{s \in \mathcal{D}_C} P_n(Ks_M - Ks) \ \textit{is achieved} \ \left(= \max_{s \in \mathcal{D}_C} P_n(Ks_M - Ks) \right) ,$$

we have in fact the following exact formula,

$$\begin{aligned} P\left(Ks_n - Ks_M\right) &= \arg\max_{L\geq 0}\left\{\max_{s\in\mathcal{D}_L} P_n\left(Ks_M - Ks\right)\right\} \\ &= \arg\max_{L\geq 0}\left\{\max_{s\in\mathcal{D}_L}\left(P-P_n\right)\left(Ks-Ks_M\right) - L\right\} \ . \end{aligned} \tag{19}$$

So, if (18) is satisfied with high probability, we recover Bartlett and Mendelson's observation, which is

$$P\left(Ks_n - Ks_M\right) \sim \arg\max_{L\geq 0}\left\{\mathcal{V}_n\left(L\right) - L\right\} \ . \tag{20}$$

In Theorem 3.1 of [4], a precise sense is given to (20), in a rather general framework. In particular, a lower bound for the excess risk is given but only through an additional condition controlling the supremum of the empirical process of interest itself over a subset of functions of "small" excess risks. This additional condition remains the major restriction concerning the related result of Bartlett and Mendelson. In the following, we show in our more restricted framework how to take advantage of the linearity of the model, as well as the existence of an expansion of the least squares contrast around the projection s_M of the target, to derive lower bounds without additional assumptions on the behavior of the empirical process of interest. Moreover, our methodology allow to explicitly calculate the first order of the quantity given at the right side of (20), thus exhibiting a rather simple complexity term controlling the rate of convergence of the excess risk in the regression setting and relating some geometrical characteristics of the model M to the unknown law P of data.

Remark 2 *Formula (17) and (19) above show that the true and empirical excess risks are of different nature, in the sense that the first one is referred to the arguments of the function*

$$\Gamma_n : L\,(\geq 0) \mapsto \max_{s\in\mathcal{D}_L}\left(P-P_n\right)\left(Ks-Ks_M\right) - L \ ,$$

whereas the second one is measured from the values of the function Γ_n. Hence, the equivalence between the true and the empirical excess risks, when satisfied, is in general not straightforward. It is a consequence of the following "fixed point type" equation,

$$\arg\max_{\mathbb{R}_+}\left\{\Gamma_n\right\} \sim \max_{\mathbb{R}_+}\left\{\Gamma_n\right\} \ .$$

Considering that the approximation stated in (18) is suitably satisfied, it remains to get an asymptotic first order expansion of its right-hand term. Such a control is obtained through the use of the least squares contrast expansion given in (6). Indeed, using (6), we get

$$\begin{aligned} & \mathbb{E}\left[\sup_{s\in\mathcal{D}_L}\left\{\left(P-P_n\right)\left(Ks-Ks_M\right)\right\}\right] \\ = \ & \underbrace{\mathbb{E}\left[\sup_{s\in\mathcal{D}_L}\left\{\left(P-P_n\right)\left(\psi_{1,M}\cdot\left(s-s_M\right)\right)\right\}\right]}_{\text{principal part}} + \underbrace{\mathbb{E}\left[\sup_{s\in\mathcal{D}_L}\left\{\left(P-P_n\right)\left(\left(s-s_M\right)^2\right)\right\}\right]}_{\text{residual term}} . \end{aligned} \tag{21}$$

In order to show that the residual term is negligible compared with the principal part, it is natural to use a contraction principle (see Theorem 4.12 of [15], also recalled in Section 7.5). Indeed, arguments of the empirical process appearing in the residual term are related to the square of the arguments defining the empirical process in the principal part. Moreover, it appears by using the contraction principle, that the ratio of the residual term over the principal part is roughly given by the supremum norm of the indexes: $\sup_{s\in\mathcal{D}_L}\left|\left(s-s_M\right)(x)\right|$ (see Lemma 14 in Section 7.4 for more details). Now, using assumption (**H3**) of Section 4.1, concerning the unit envelope of the linear model M, we get that the last quantity is of order $\sqrt{DL}$. Since the values L of interest are typically of order D/n, the quantity controlling the ratio is not sharp enough as it does not converge to zero as soon as the dimension D is of order at least $\sqrt{n}$.

We thus have to refine our analysis in order to be able to neglect the residual term. The assumption of sup-norm consistency, of the least squares estimator s_n toward the projection s_M of the target onto the model M, appears here to be essential. Indeed, if assumption (**H5**) of Section 4.1 is satisfied, then all the above computations can be restricted with high probability to the subset where belongs the estimator s_n, this subset being more precisely

$$B_{L_\infty}(s_M, R_{n,D,\alpha}) = \{s \in M \ ; \ \|s - s_M\|_\infty \leq R_{n,D,\alpha}\} \subset M \ , \tag{22}$$

$R_{n,D,\alpha} \ll 1$ being the rate of convergence in sup-norm of s_n toward s_M, defined in (**H5**). In particular, the spheres of interest $\mathcal{D}_L$ are now replaced in the calculations by their intersection $\tilde{\mathcal{D}}_L$ with the ball of radius $R_{n,D,\alpha}$ in $(M, \|\cdot\|_\infty)$,

$$\tilde{\mathcal{D}}_L = \mathcal{D}_L \cap B_{L_\infty}(s_M, R_{M,n,\alpha}) \ .$$

The ratio between the consequently modified residual term and principal part of (21) is then roughly controlled by $R_{n,D,\alpha}$ (see again Lemma 14 in Section 7.4), a quantity indeed converging to zero as desired. Hence, under the assumption (**H5**), we get

$$\mathbb{E}\left[\sup_{s \in \tilde{\mathcal{D}}_L} \{(P - P_n)(Ks - Ks_M)\}\right] \sim \mathbb{E}\left[\sup_{s \in \tilde{\mathcal{D}}_L} \left\{(P - P_n)\left(\psi_{1,M} \cdot (s - s_M)\right)\right\}\right] \ . \tag{23}$$

A legitimate and important question is: how restrictive is assumption (**H5**) of consistency in sup-norm of the least squares estimator ? We prove in Lemma 5 of Section 5 that this assumption is satisfied for models of histograms defined on a partition satisfying some regularity condition, at a rate of convergence of order $\sqrt{D \ln(n)/n}$. Moreover, in Lemma 8, Section 6, we extend this result for models of piecewise polynomials uniformly bounded in their degrees, again under some lower-regularity assumption on the partition defining the model; the rate of convergence being also preserved. A systematical study of consistency in sup-norm of least squares estimators, on more general finite-dimensional linear models, is also postponed to a forthcoming paper.

The control of the right-hand side of (23), which is needed to be sharp, is particularly technical, and is essentially contained in Lemmas 12 and 13 of Section 7.4. Let us shortly describe the mathematical figures underlying this control. First, by bounding the variance of the considered supremum of the empirical process - by using a result due to Ledoux [14], see Theorem 24 and also Corollary 25 in Section 7.5 -, we roughly get, for values of L of interest,

$$\mathbb{E}\left[\sup_{s \in \tilde{\mathcal{D}}_L} \left\{(P - P_n)\left(\psi_{1,M} \cdot (s - s_M)\right)\right\}\right] \sim \mathbb{E}^{1/2}\left[\left(\sup_{s \in \tilde{\mathcal{D}}_L} \left\{(P - P_n)\left(\psi_{1,M} \cdot (s - s_M)\right)\right\}\right)^2\right] \ . \tag{24}$$

Then, by assuming that the model M is fulfilled with a localized orthonormal basis, as stated in assumption (**H4**) of Section 4.1, it can be shown that the localization on the ball $B_{L_\infty}(s_M, R_{M,n,\alpha})$ can be removed from the right-hand side of (24), in the sense that

$$\mathbb{E}^{1/2}\left[\left(\sup_{s \in \tilde{\mathcal{D}}_L} \left\{(P - P_n)\left(\psi_{1,M} \cdot (s - s_M)\right)\right\}\right)^2\right] \sim \mathbb{E}^{1/2}\left[\left(\sup_{s \in \mathcal{D}_L} \left\{(P - P_n)\left(\psi_{1,M} \cdot (s - s_M)\right)\right\}\right)^2\right] \ . \tag{25}$$

The property of localized basis is standard in model selection theory (see for instance Chapter 7 of [16]) and was first introduced by Birgé and Massart in [5], also for deriving sharp exponential bounds in a M-estimation context. We show in Lemmas 4 and 7 that this assumption is satisfied for models of histograms and piecewise polynomials respectively, when they satisfy a certain regularity assumption concerning the underlying partition.

Finally, as $\mathcal{D}_L$ is a sphere in $(M, \|\cdot\|_2)$, we simply get, by the use of Cauchy-Schwarz inequality, that the right-hand side of (25) is equal to $\sqrt{(L/n) \cdot \sum_{k=1}^{D} \mathrm{Var}\left(\psi_{1,M} \cdot \varphi_k\right)}$, where $(\varphi_k)_{k=1}^{D}$ is an orthonormal basis of

M. Gathering our arguments, we then obtain

$$\begin{aligned} P\left(Ks_n - Ks_M\right) &\sim \arg\max_{L\geq 0}\left\{\sup_{s\in\mathcal{D}_L}\mathbb{E}\left[\left(P_n - P\right)\left(Ks_M - Ks\right)\right] - L\right\} \\ &\sim \arg\max_{L\geq 0}\left\{\sqrt{\frac{L\cdot\sum_{k=1}^{D_M}\mathrm{Var}\left(\psi_{1,M}\cdot\varphi_k\right)}{n}} - L\right\} = \frac{1}{4}\frac{D_M}{n}\mathcal{K}_{1,M}^2\ , \end{aligned} \tag{26}$$

where $\mathcal{K}_{1,M}^2 := \frac{1}{D_M}\sum_{k=1}^{D_M}\mathrm{Var}\left(\psi_{1,M}\cdot\varphi_k\right)$. As shown in Section 4.3 below, the (normalized) complexity term $\mathcal{K}_{1,M}$ is independent of the choice of the basis $(\varphi_k)_{k=1}^{D}$ and is, under our assumptions, of the order of a constant. Concerning the empirical excess risk, we have

$$\begin{aligned} P_n\left(Ks_M - Ks_n\right) &= \max_{L\geq 0}\left\{\sup_{s\in\mathcal{D}_L}\mathbb{E}\left[\left(P_n - P\right)\left(Ks_M - Ks\right)\right] - L\right\} \\ &\sim \max_{L\geq 0}\left\{\sqrt{\frac{L\cdot\sum_{k=1}^{D_M}\mathrm{Var}\left(\psi_{1,M}\cdot\varphi_k\right)}{n}} - L\right\} = \frac{1}{4}\frac{D_M}{n}\mathcal{K}_{1,M}^2\ . \end{aligned} \tag{27}$$

In particular, the equivalence

$$P\left(Ks_n - Ks_M\right) \sim P_n\left(Ks_M - Ks_n\right)\left(\sim \frac{1}{4}\frac{D_M}{n}\mathcal{K}_{1,M}^2\right)$$

is justified.

In Theorem 2 below, a precise, non-asymptotic sense, is given to equivalences described in (26) and (27). This is done under the structural constraints stated in conditions (**H4**) and (**H5**), for models of reasonable dimension. Moreover, we give in Theorem 3 upper bounds for the true and empirical excess risks, that are less precise than the bounds of Theorem 2, but that are also valid for models of small dimension. Corollaries of these theorems are given in the case of histograms and piecewise polynomials, in Corollaries 6 and 9 respectively. Indeed, we show that in these particular cases, our general conditions (**H4**) and (**H5**) essentially reduce to a simple lower-regularity assumption on the underlying partition.

4 True and empirical excess risk bounds

In this section, we derive under general constraints on the linear model M, upper and lower bounds for the true and empirical excess risk, that are optimal - and equal - at the first order. In particular, we show that the true excess risk is equivalent to the empirical one when the model is of reasonable dimension. For smaller dimensions, we only achieve some upper bounds.

4.1 Main assumptions

We turn now to the statement of some assumptions that will be needed to derive our results in Section 4.2. These assumptions will be further discussed in Section 4.3.

Boundedness assumptions:

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

$$|Y_i| \leq A \ a.s. \tag{28}$$

and

$$\|s_M\|_\infty \leq A\ . \tag{29}$$

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

$$\|s_*\|_\infty = \|\mathbb{E}[Y|X=\cdot]\|_\infty \leq A \tag{30}$$

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

$$\sigma^2(X_i) \leq \sigma_{\max}^2 \leq A^2 \quad a.s. \tag{31}$$

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

$$\left|\psi_{1,M}(X_i, Y_i)\right| \leq 4A \quad a.s. \tag{32}$$

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

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

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

$$\Psi_M(x) = \frac{1}{\sqrt{D}} \sup_{s \in M, \|s\|_2 \leq 1} |s(x)| \ . \tag{33}$$

As M is a finite dimensional real vector space, the supremum in (33) 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 $|\beta|_\infty = \max\{|\beta_k| ; k \in \{1, ..., D\}\}$ is the sup-norm of the D-dimensional vector β.

Remark 3 *(**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

$$R_{n,D,\alpha} \leq \frac{A_{cons}}{\sqrt{\ln n}} \tag{34}$$

  and by setting

$$\Omega_{\infty,\alpha} = \{\|s_n - s_M\|_\infty \leq R_{n,D,\alpha}\} \ , \tag{35}$$

  it holds for all $n \geq n_1$,

$$\mathbb{P}[\Omega_{\infty,\alpha}] \geq 1 - n^{-\alpha} \ . \tag{36}$$

4.2 Theorems

We state here the general results of this article, that will be applied in Section 5 and 6 in the case of piecewise constant functions and piecewise polynomials respectively.

Theorem 2 *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*

$$A_- (\ln n)^2 \leq D \leq A_+ \frac{n}{(\ln n)^2} \ , \tag{37}$$

*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 assumptions **(H1)**, **(H2)** and **(H4)** respectively, such that by setting*

$$\varepsilon_n = A_0 \max\left\{ \left(\frac{\ln n}{D}\right)^{1/4}, \ \left(\frac{D \ln n}{n}\right)^{1/4}, \ \sqrt{R_{n,D,\alpha}} \right\} \ , \tag{38}$$

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

$$\mathbb{P}\left[P(Ks_n - Ks_M) \geq (1 - \varepsilon_n) \frac{1}{4} \frac{D}{n} \mathcal{K}_{1,M}^2 \right] \geq 1 - 5n^{-\alpha} \ , \tag{39}$$

$$\mathbb{P}\left[P(Ks_n - Ks_M) \leq (1 + \varepsilon_n) \frac{1}{4} \frac{D}{n} \mathcal{K}_{1,M}^2 \right] \geq 1 - 5n^{-\alpha} \ , \tag{40}$$

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

$$\mathbb{P}\left[P_n(Ks_M - Ks_n) \leq \left(1 + \varepsilon_n^2\right) \frac{1}{4} \frac{D}{n} \mathcal{K}_{1,M}^2 \right] \geq 1 - 3n^{-\alpha} \ , \tag{42}$$

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

$$\mathbb{P}\left(P_n(Ks_M - Ks_n) \geq \left(1 - A_0 \max\left\{ \sqrt{\frac{\ln n}{D}}, \sqrt{\frac{D \ln n}{n}} \right\}\right) \frac{D}{4n} \mathcal{K}_{1,M}^2 \right) \geq 1 - 2n^{-\alpha} \ . \tag{43}$$

In Theorem 2 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 2 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 (43), 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 3 *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 $A, A_{cons}, A_{3,M}$ and α, such that for all $n \geq n_0(A_{cons}, n_1)$,*

$$\mathbb{P}\left[P(Ks_n - Ks_M) \geq A_u \frac{D \vee \ln n}{n} \right] \leq 3n^{-\alpha} \tag{44}$$

and

$$\mathbb{P}\left[P_n\left(Ks_M - Ks_n\right) \geq A_u \frac{D \vee \ln n}{n}\right] \leq 3n^{-\alpha} \ . \tag{45}$$

Notice that on contrary to the situation of Theorem 2, 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 2 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 3, 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 (44) and (45) are possibly larger than the corresponding ones of Theorem 2, but our results still hold true for small models. Moreover, in the case of models with reasonable dimensions, that is dimensions satisfying assumption (37) of Theorem 2, the rate of decay is preserved compared to Theorem 2 and is proportional to D/n.
The proofs of the above theorems can be found in Section 7.3.

4.3 Some additional comments

Let us first comment on the assumptions given in Section 4.1. Assumptions (28) and (**H2**) are rather mild and can also be found in the work of Arlot and Massart [2] related to the case of histograms, where they are respectively denoted by (**Ab**) and (**An**). These assumptions state respectively that the response variable Y is uniformly bounded and that the noise level is uniformly bounded away from zero. In [2], Arlot and Massart also notice that their results can be extended to the unbounded case, where assumption (**Ab**) is replaced by some condition on the moments of the noise, and where (**An**) is weakened into mild regularity conditions for the noise level. We believe that moments conditions on the noise, in the spirit of assumptions stated by Arlot and Massart, could also been taken into account in our study in order to weaken (28), but at the prize of many technical efforts that are beyond the scope of the present paper. However, we explain at the end of this section how condition (**H2**) can be relaxed - see hypothesis (**H2bis**) below.
In assumption (**H4**) we require that the model M is provided with an orthonormal localized basis in $L_2\left(P^X\right)$. 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 [16], where it is shown that models of histograms, piecewise polynomials and compactly supported wavelets are typical examples of models with localized basis for the $L_2\left(\text{Leb}\right)$ structure, considering that $\mathcal{X} \subset \mathbb{R}^k$. In Sections 5 and 6, we show that models of piecewise constant and piecewise polynomials respectively can also have a localized basis for the $L_2\left(P^X\right)$ structure, under rather mild assumptions on P^X. Assumption (**H4**) is needed in Theorem 2, whereas in Theorem 3 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 2 to achieve sharp lower bounds for the quantities of interest, whereas in Theorem 3 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 $\left(\ln n\right)^{-1/2}$. This can be considered as a rather strong assumption, but it is essential for our methodology. Moreover, we show in Sections 5 and 6 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 2 for models of reasonable dimensions. As we can see in Theorem 2, 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

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

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 $P\left(\psi_{1,M}\cdot\varphi_k\right) = 0$, 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 (33), as pointed out by Birgé and Massart [5], we get

$$\Psi_M^2 = \frac{1}{D}\sum_{k=1}^{D}\varphi_k^2 \tag{46}$$

and so

$$\begin{aligned}
\mathcal{K}_{1,M}^2 &= P\left(\psi_{1,M}^2\Psi_M^2\right) \\
&= 4\mathbb{E}\left[\mathbb{E}\left[\left(Y - s_M(X)\right)^2 | X\right]\Psi_M^2(X)\right] \\
&= 4\left(\mathbb{E}\left[\sigma^2(X)\,\Psi_M^2(X)\right] + \mathbb{E}\left[\left(s_M - s_*\right)^2(X)\,\Psi_M^2(X)\right]\right) \ .
\end{aligned} \tag{47}$$

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

$$\mathcal{K}_{1,M} \le 2\sigma_{\max} + 4A \le 6A \ . \tag{48}$$

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

$$\mathcal{K}_{1,M} \ge 2\sigma_{\min} > 0 \ . \tag{49}$$

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 we have

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

Now, under notations of Lemma 4 below,

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

thus we deduce, by (46) and the previous equality, that

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

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

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

By (47), we explicitly get

$$\mathcal{K}_{1,M}^2 = 4\left(\sigma^2 + \frac{1}{|\mathcal{P}|}\sum_{I\in\mathcal{P}} \mathbb{V}[\mathbb{E}[Y|X]|X \in I]\right) \ . \tag{50}$$

A careful look at the proof of Theorem 2 given in Section 7.3 show that condition (**H2**) is only used through the lower bound (49), 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 49 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 $x \in \mathcal{X}$,

$$\Psi_M(x) \geq A_\Psi^- > 0 \ , \tag{51}$$

and this allows to consider vanishing noise level, as we then have by (47),

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

As we will see in Sections 5 and 6, Inequality (51) can be satisfied for histogram and piecewise polynomial models on a partition achieving some upper regularity assumption with respect to the law P^X .

5 The histogram case

In this section, we particularize the results stated in Section 4 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 $L_2\left(P^X\right)$ and (**H5**) of consistency in sup-norm of the M-estimator towards the linear projection s_M are satisfied.

5.1 Existence of a localized basis

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

Lemma 4 *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 $c_{M,P}$,*

$$\sqrt{|\mathcal{P}| \inf_{I\in\mathcal{P}} P^X(I)} \geq c_{M,P} > 0 \ . \tag{52}$$

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

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

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

$$\text{for all } \beta = (\beta_I)_{I\in\mathcal{P}} \in \mathbb{R}^D, \quad \left\|\sum_{I\in\mathcal{P}} \beta_I\varphi_I\right\|_\infty \leq c_{M,P}^{-1}\sqrt{D}\,|\beta|_\infty \ . \tag{53}$$

Condition (52) can also be found in Arlot and Massart [2] 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. (52) is equivalent to (53). The proof of Lemma 4 is straightforward and can be found in Section 7.1.

5.2 Rates of convergence in sup-norm

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

Lemma 5 *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 (28) holds, that is, a positive constant A exists such that $|Y| \leq A$ a.s. Moreover, assume that for some positive finite constant $c_{M,P}$,*

$$\sqrt{|\mathcal{P}| \inf_{I \in \mathcal{P}} P^X(I)} \geq c_{M,P} > 0 \tag{54}$$

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,

$$\|s_n - s_M\|_\infty \leq L_{A_+,A,c_{M,P},\alpha} \sqrt{\frac{D \ln n}{n}} \ . \tag{55}$$

In Lemma 5 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 5, if we set

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

then Assumption **(H5)** is satisfied. To derive Inequality (55), 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 [2] and we can thus view the general set of assumptions exposed in Section 4.1 as a natural generalization for linear models of the framework developed in [2] in the case of histograms. The proof of Lemma 5 can be found in Section 7.1.

5.3 Bounds for the excess risks

The next results is a straightforward application of Lemmas 4, 5 and Theorems 2, 3.

Corollary 6 *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*

$$\sqrt{|\mathcal{P}| \inf_{I \in \mathcal{P}} P^X(I)} \geq c_{M,P} > 0 \ . \tag{56}$$

*If (**H1**) and (**H2**) of Section 4.1 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 $n \geq n_0(A_-, A_+, A, , \sigma_{\min}, c_{M,P}, \alpha)$,

$$\mathbb{P}\left[(1+\varepsilon_n) \frac{1}{4} \frac{D}{n} \mathcal{K}_{1,M}^2 \geq P(Ks_n - Ks_M) \geq (1-\varepsilon_n) \frac{1}{4} \frac{D}{n} \mathcal{K}_{1,M}^2 \right] \geq 1 - 10 n^{-\alpha} \tag{57}$$

and

$$\mathbb{P}\left[\left(1+\varepsilon_n^2\right) \frac{1}{4} \frac{D}{n} \mathcal{K}_{1,M}^2 \geq P_n(Ks_M - Ks_n) \geq \left(1-\varepsilon_n^2\right) \frac{1}{4} \frac{D}{n} \mathcal{K}_{1,M}^2 \right] \geq 1 - 5 n^{-\alpha} \ . \tag{58}$$

*If (56) 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 $n \geq n_0 (A, c_{M,P}, A_+, \alpha)$,

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

and

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

We recover in Corollary 6 the general results of Section 4.2 for the case of histograms on a lower-regular partition. Moreover, in the case of histograms, assumption (29) which is part of (**H1**) is a straightforward consequence of (28). 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

$$s_M = \sum_{I \in \mathcal{P}} \mathbb{E}\left[Y \,|X \in I\right] \mathbf{1}_I \ . \tag{59}$$

Under (28), we have $|\mathbb{E}\left[Y \,|X \in I\right]| \leq \|Y\|_\infty \leq A$ for every $I \in \mathcal{P}$ and we deduce by (59) that $\|s_M\|_\infty \leq A$.

5.4 Comments

Our bounds in Corollary 6 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 7.3 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 [2], in the case of a fixed model. Such results can be found in Propositions 10, 11 and 12 of [2]. The strategy adopted by the authors in this case is as follows. They first notice that the mean of the empirical excess risk on histograms is given by

$$\mathbb{E}\left[P_n\left(Ks_M - Ks_n\right)\right] = \frac{D}{4n} \mathcal{K}_{1,M}^2 \ .$$

Then they derive concentration inequalities for the true excess risk and its empirical counterpart around their mean. Finally, the authors compare the mean of the true excess risk to the mean of the empirical excess risk. More precisely, using our notations, inequality (34) of Proposition 10 in [2] states that for every $x \geq 0$ there exists an event of probability at least $1 - e^{1-x}$ on which,

$$\begin{aligned} &\left|P_n\left(Ks_M - Ks_n\right) - \mathbb{E}\left[P_n\left(Ks_M - Ks_n\right)\right]\right| \\ &\leq \frac{L}{\sqrt{D_M}} \left[P(Ks_M - Ks_*) + \frac{A^2 \mathbb{E}\left[P_n\left(Ks_M - Ks_n\right)\right]}{\sigma_{\min}^2} \left(\sqrt{x} + x\right) \right] \ , \end{aligned} \tag{60}$$

for some absolute constant L. One can notice that inequality (60), 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 [2], 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 for 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 as ours, inequality (60) becomes significantly better than inequality (58) for large models.

Concentration inequalities for the true excess risk given in Proposition 11 of [2] 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 Corollary 6. 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 [2] 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 tends to indicate that, up to a logarithmic factor, the term proportional to $\left(\frac{D \ln n}{n}\right)^{1/4}$ appearing in ε_n is not improvable in general, and that the empirical excess risk concentrates better around its mean than the true excess risk.
We conclude that the bounds given in Proposition 10, 11 and 12 of [2] are essentially more accurate than ours, apart for 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 sharper than ours for large models.
Arlot and Massart [2] 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 [2] in this case. Indeed, we have noticed in Section 4.3 that assumption (**H2**) can be weakened into (**H2bis**), where we assume that

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

for some positive constant $A_{\min}$. So, it suffices to bound from below the normalized complexity. We have from identity (47),

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

Moreover, from identity (46), 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, if we assume the upper regularity of the partition $\mathcal{P}$ with respect to P^X, that is

$$|\mathcal{P}| \sup_{I \in \mathcal{P}} P^X(I) \leq c_{M,P}^+ < +\infty \tag{61}$$

for some positive constant $c_{M,P}^+$, we then 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**).

6 The case of piecewise polynomials

In this Section, we generalize the results given in Section 5 for models of piecewise constant functions to models of piecewise polynomials uniformly bounded in their degrees.

6.1 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 7 *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* $\left\{\varphi_{I,j},\ I \in \mathcal{P},\ j \in \{0,...,r\}\right\}$ *of* $(M, \|\cdot\|_2)$ *such that,*

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

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

$$\max_{j \in \{0,...,r\}} \left\|\varphi_{I,j}\right\|_\infty \leq L_{r,c_{\min}} \frac{1}{\sqrt{\text{Leb}(I)}} \ . \tag{62}$$

As a consequence, if it holds

$$\sqrt{|\mathcal{P}| \inf_{I \in \mathcal{P}} \text{Leb}(I)} \geq c_{M,\text{Leb}} \tag{63}$$

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* $\beta = \left(\beta_{I,j}\right)_{I \in \mathcal{P}, j \in \{0,...,r\}} \in \mathbb{R}^D$*,*

$$\left\| \sum_{I,j} \beta_{I,j} \varphi_{I,j} \right\|_\infty \leq L_{r,c_{\min},c_{M,\text{Leb}}} \sqrt{D}\, |\beta|_\infty \tag{64}$$

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

Lemma 7 states that if $\mathcal{X} = [0,1]$ is the unit interval and if P^X has a density with respect to the Lebesgue measure Leb on $\mathcal{X}$, which is uniformly bounded away form zero, then there exists an orthonormal basis in $(M, \|\cdot\|_2)$ satisfying good enough properties in terms of the sup-norm of its elements. Moreover, if we assume the lower regularity of the partition with respect to Leb, then the orthonormal basis is localized and the constant of localization given in (64) 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 7 can be found in Section 7.2.

6.2 Rates of convergence in sup-norm

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

Lemma 8 *Assume that Inequality (28) 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* $c_{\max}$*,*

$$0 < c_{\min} \leq f(x) \leq c_{\max} < +\infty, \ x \in [0,1] \ . \tag{65}$$

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* $c_{M,\text{Leb}}$

$$\sqrt{|\mathcal{P}| \inf_{I \in \mathcal{P}} \text{Leb}(I)} \geq c_{M,\text{Leb}} > 0 \ . \tag{66}$$

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* $n \geq n_0(r, A_+, c_{\min}, c_{M,\text{Leb}}, \alpha)$*,*

$$\|s_n - s_M\|_\infty \leq L_{A,r,A_+,c_{\min},c_{\max},c_{M,\text{Leb}},\alpha} \sqrt{\frac{D \ln n}{n}} \ . \tag{67}$$

In Lemma 5, we thus obtain the convergence in sup-norm of the M-estimator s_n toward the linear projection s_M at the rate $\sqrt{D\ln(n)/n}$. It is worth noting that, for a model of piecewise polynomials satisfying the assumptions of Lemma 5, if we set

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

then Assumption **(H5)** is satisfied. The proof of Lemma 8 can be found in Section 7.2.

6.3 Bounds for the excess risks

The forthcoming result is a straightforward application of Lemmas 7, 8 and Theorems 2, 3.

Corollary 9 *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* $c_{\max}$*,*

$$0 < c_{\min} \leq f(x) \leq c_{\max} < +\infty, \ x \in [0,1] \ . \tag{68}$$

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* $c_{M,\mathrm{Leb}}$*,*

$$\sqrt{|\mathcal{P}| \inf_{I\in\mathcal{P}} \mathrm{Leb}(I)} \geq c_{M,\mathrm{Leb}} > 0 \ . \tag{69}$$

*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* $\alpha, A, \sigma_{\min}, A_-, A_+, r, c_{M,\mathrm{Leb}}, c_{\min}$ *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 $n \geq n_0(A_-, A_+, A, r, \sigma_{\min}, c_{M,\mathrm{Leb}}, c_{\min}, c_{\max}, \alpha)$*,*

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

and

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

*Moreover, if (68) and (69) 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* $A_+, A, r, c_{M,\mathrm{Leb}}, c_{\min}$ *and* α*, such that for all* $n \geq n_0(A_+, A, r, c_{\min}, c_{\max}, c_{M,\mathrm{Leb}}, \alpha)$*,*

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

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 Corollary 9 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 essentially the bounds of Corollary 6, since by Lemma 8, we still have $R_{n,D,\alpha} \propto \sqrt{D \ln(n)/n}$.
Moreover, as in the case of histograms, assumption (29) which is part of (**H1**), is a straightforward consequence of (28). 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

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

where $\varphi_{I,j}$ is the orthonormal basis given in Lemma 7. It is then easy to show, using (62) and (28), 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 4.3, 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 Corollary 9 still hold. Now, from identity (47) we have

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

Moreover, from identity (46), it holds in the case of piecewise polynomials, for all $x \in \mathcal{X}$,

$$\Psi_M^2(x) = \frac{1}{(r+1)\,|\mathcal{P}|}\sum_{(I,j)\in\mathcal{P}\times\{0,\dots,r\}} \varphi_{I,j}^2 \geq \frac{1}{(r+1)\,|\mathcal{P}|}\sum_{I\in\mathcal{P}}\frac{\mathbf{1}_{x\in I}}{P^X(I)} \ . \tag{70}$$

Furthermore, if we ask that

$$|\mathcal{P}|\sup_{I\in\mathcal{P}}\mathrm{Leb}(I) \leq c_{M,P}^+ < +\infty \tag{71}$$

for some positive constant $c_{M,P}^+$, then by using (68), (70) and (71), we obtain for all $x \in \mathcal{X}$,

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

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

7 Proofs

We begin with the simpler proofs of Sections 5 and 6, in Sections 7.1 and 7.2 respectively. The proofs of Theorems 2 and 3 of Section 4.2 can be found in Section 7.3.

7.1 Proofs of Section 5

Proof of Lemma 4. It suffices to observe that

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

■

We now intend to prove (55) under the assumptions of Lemma 5.

Proof of Lemma 5. Along the proof, we denote by abusing the 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

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

we get by Bernstein's inequality (230), for any $x > 0$ and $I \in \mathcal{P}$,

$$\mathbb{P}\left[|(P_n - P)(I)| \geq \sqrt{\frac{2P(I)x}{n}} + \frac{x}{3n}\right] \leq 2\exp(-x) . \tag{72}$$

Further note that by (54), $D \geq c_{M,P}^2 P(I)^{-1} > 0$ for any $I \in \mathcal{P}$, and thus by taking $x = \beta \ln n$, we easily deduce from inequality (72) that there exists a positive constant $L^{(1)}_{\beta, c_{M,P}, A_+}$ only depending on $c_{M,P}$ and β such that, for any $I \in \mathcal{P}$,

$$\mathbb{P}\left[\frac{|(P_n - P)(I)|}{P(I)} \geq L^{(1)}_{\beta, c_{M,P}, A_+}\sqrt{\frac{D \ln n}{n}}\right] \leq 2n^{-\beta} . \tag{73}$$

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

$$L^{(1)}_{\beta, c_{M,P}, A_+}\sqrt{\frac{D \ln n}{n}} \leq \frac{1}{2}, \text{ for all } n \geq n_0(\beta, c_{M,P}, A_+) . \tag{74}$$

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

$$\begin{aligned}
&\mathbb{P}[\forall I \in \mathcal{P},\ P_n(I) > 0] \\
&\geq \mathbb{P}\left[\forall I \in \mathcal{P},\ \frac{P(I)}{2} > |(P_n - P)(I)|\right] \\
&\geq \mathbb{P}\left[\forall I \in \mathcal{P},\ \frac{|(P_n - P)(I)|}{P(I)} < L^{(1)}_{\beta, c_{M,P}, A_+}\sqrt{\frac{D \ln n}{n}}\right] \quad \text{by (74)} \\
&\geq 1 - 2Dn^{-\beta} .
\end{aligned}$$

Introduce the event

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

We have shown that

$$\mathbb{P}[\Omega_+] \geq 1 - 2Dn^{-\beta} . \tag{75}$$

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

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

Hence it holds on Ω_+,

$$\begin{aligned}
\left\| s_n - s_M \right\|_\infty &= \sup_{I \in \mathcal{P}} \left| \frac{P_n\left(y \mathbf{1}_{x \in I}\right)}{P_n\left(I\right)} - \frac{P\left(y \mathbf{1}_{x \in I}\right)}{P\left(I\right)} \right| \\
&= \sup_{I \in \mathcal{P}} \left| \frac{P_n\left(y \mathbf{1}_{x \in I}\right)}{P\left(I\right)\left(1 + \frac{(P_n - P)(I)}{P(I)}\right)} - \frac{P\left(y \mathbf{1}_{x \in I}\right)}{P\left(I\right)} \right| \\
&\leq \sup_{I \in \mathcal{P}} \left| \frac{\left(P_n - P\right)\left(y \mathbf{1}_{x \in I}\right)}{P\left(I\right)\left(1 + \frac{(P_n - P)(I)}{P(I)}\right)} \right| \\
&\quad + \sup_{I \in \mathcal{P}} \left| \frac{P\left(y \mathbf{1}_{x \in I}\right)}{P\left(I\right)} \right| \times \sup_{I \in \mathcal{P}} \left| 1 - \frac{1}{1 + \frac{(P_n - P)(I)}{P(I)}} \right| \ . \qquad (76)
\end{aligned}$$

Moreover, by Bernstein's inequality (230), as

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

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

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

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

$$\mathbb{P}\left[\frac{\left| \left(P_n - P\right)\left(y \mathbf{1}_{x \in I}\right) \right|}{P\left(I\right)} \geq L^{(2)}_{A, c_{M,P}, \beta, A_+} \sqrt{\frac{D \ln n}{n}} \right] \leq 2 n^{-\beta} \ . \qquad (77)$$

Now define

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

Clearly, since $D \leq n$ we have, by (73) and (77),

$$\mathbb{P}\left[\Omega_{1,2}^c \right] \leq 4 n^{-\beta + 1} \ . \qquad (78)$$

Moreover, for all $n \geq n_0\left(\beta, c_{M,P}, A_+\right)$, we get by (74) that

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

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

$$\begin{aligned}
&\sup_{I \in \mathcal{P}} \left| \frac{\left(P_n - P\right)\left(y \mathbf{1}_{x \in I}\right)}{P\left(I\right)\left(1 + \frac{(P_n - P)(I)}{P(I)}\right)} \right| + \sup_{I \in \mathcal{P}} \left| \frac{P\left(y \mathbf{1}_{x \in I}\right)}{P\left(I\right)} \right| \times \sup_{I \in \mathcal{P}} \left| 1 - \frac{1}{1 + \frac{(P_n - P)(I)}{P(I)}} \right| \\
&\leq 2 \sup_{I \in \mathcal{P}} \left| \frac{\left(P_n - P\right)\left(y \mathbf{1}_{x \in I}\right)}{P\left(I\right)} \right| + 2 \sup_{I \in \mathcal{P}} \left| \frac{P\left(y \mathbf{1}_{x \in I}\right)}{P\left(I\right)} \right| \times \sup_{I \in \mathcal{P}} \left| \frac{\left(P_n - P\right)\left(I\right)}{P\left(I\right)} \right| \\
&\leq 2 L^{(2)}_{A, c_{M,P}, \beta, A_+} \sqrt{\frac{D \ln n}{n}} + 2 L^{(1)}_{\beta, c_{M,P}, A_+} \sqrt{\frac{D \ln n}{n}} \times \sup_{I \in \mathcal{P}} \left| \frac{P\left(y \mathbf{1}_{x \in I}\right)}{P\left(I\right)} \right| \ . \qquad (79)
\end{aligned}$$

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

$$|P(y\mathbf{1}_{x\in I})| \leq P(|y|\,\mathbf{1}_{x\in I}) \leq AP(I), \tag{80}$$

so by (76), (79) and (80) we finally get, on the event $\Omega_{1,2}$ and for all $n \geq n_0(\beta, c_{M,P}, A_+)$,

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

Taking $\beta = \alpha + 3$, we get by (78) for all $n \geq 2$, $\mathbb{P}\left[\Omega^c_{1,2}\right] \leq n^{-\alpha}$ which implies (55).

■

7.2 Proofs of Section 6

Under the assumptions of Lemma 7, we intend to establish (64).

Proof of Lemma 7. 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

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

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 $\|.\|_{I,w}$ 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 $g, h \in L_2\left(I, f_{|I}\right)$,

$$\begin{aligned}&\int_{x\in I} g(x)\,h(x)\,f_{|I}(x)\,\frac{dx}{\text{Leb}(I)}\\ &= \int_{y\in[0,1]} g((b-a)\,y+a)\,h((b-a)\,y+a)\,f_{|I}((b-a)\,y+a)\,dy\ .\end{aligned} \tag{81}$$

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

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

we deduce from equality (81) that $\left(\tilde{\varphi}_{I,j}\right)_{j=0}^r$ is an orthonormal family of polynomials in $L_2\left(I, f_{|I}\right)$ such that $\deg\left(\tilde{\varphi}_{I,j}\right) = \deg(p_{I,j})$.
Now, it is a classical fact of orthogonal polynomials theory (see for example Theorems 1.11 and 1.12 of [9]) that there exists a unique family $(q_{I,0}, q_{I,1}, ...q_{I,r})$ 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

$$q_{I,j}(x) = \prod_{k=1}^{j}\left(x - \alpha^k_{I,j}\right),\quad \alpha^k_{I,j} \in\ ]0,1[\ \text{and}\ \alpha^k_{I,j} \neq \alpha^l_{I,j}\ \text{for}\ k \neq l\ . \tag{82}$$

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

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

Now we set $B(\alpha, r) =]\alpha - r, \alpha + r[$ for $\alpha \in \mathbb{R}$, so that by (82) we get

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

and

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

Therefore,

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

Finally, introduce $p_{I,j} = \|q_{I,j}\|^{-1}_{[0,1],f^I} q_{I,j}$ and denote by $\varphi_{I,j}$ its associated orthonormal family of $L_2\left(I, f_{|I}\right)$. 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

$$\left\{\varphi_{I,j},\ I \in \mathcal{P},\ j \in \{0,...,r\}\right\}$$

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

$$\begin{aligned}\left\|\varphi_{I,j}\right\|_\infty &= \left\|\tilde{\varphi}_{I,j}\right\|_\infty \\ &= \|q_{I,j}\|^{-1}_{[0,1],f^I} \|q_{I,j}\|_\infty \text{Leb}(I)^{-1/2} \\ &\leq \sqrt{2} c_{\min}^{-1/2} (4r)^r \text{Leb}(I)^{-1/2} \qquad (83) \\ &\leq \sqrt{2} c_{M,\text{Leb}}^{-1} c_{\min}^{-1/2} (4r)^r (r+1)^{-1/2} \sqrt{D} \qquad (84)\end{aligned}$$

where in the last inequality we used the fact that

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

For all $j \in \{0,...,r\}$, $\varphi_{I,j}$ is supported by the element I of $\mathcal{P}$, hence we deduce from (83) that the orthonormal basis $\left\{\varphi_{I,j},\ I \in \mathcal{P},\ j \in \{0,...,r\}\right\}$ of $(M, \|\cdot\|_2)$ satisfies (62) with

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

To conclude, observe that

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

and thus, by plugging (84) into the right-hand side of the last inequality, we finally obtain that the value

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

gives the desired bound (64). ■

We now turn to the proof of (67) under the assumptions of Lemma 8. The proof is based on concentration inequalities recalled in Section 7.5 and on inequality (62) of Lemma 7, that allows us to control the sup-norm of elements of an orthonormal basis for a model of piecewise polynomials.

Proof of Lemma 8. Let $\alpha > 0$ be fixed and $\gamma > 0$ to be chosen later. The partition $\mathcal{P}$ associated to M will be denoted by

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

so that $|\mathcal{P}| = m$ and $D = (r+1)\, m$ where D is the dimension of the model M. By (62) of Lemma 7 there exist an orthonormal basis $\left\{\varphi_{I_k,j};\ k \in \{0, ..., m-1\},\ j \in \{0, ..., r\}\right\}$ of $\left(M, L_2\left(P^X\right)\right)$ such that,

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

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

$$\max_{j \in \{0,...,r\}} \left\|\varphi_{I_k,j}\right\|_\infty \leq L_{r,c_{\min}} \frac{1}{\sqrt{\mathrm{Leb}\left(I_k\right)}}, \ \text{ for all } k \in \{0, ..., m-1\}. \tag{85}$$

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

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

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

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

and consider the total ordering $\preceq$ defined to be

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

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

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

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

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

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

$$\begin{aligned} &\frac{1}{2}\frac{\partial}{\partial\beta_{I_k,j}} P_n\left[\left(y - \left(\sum_{(I_k,j)\in\mathcal{I}} \beta_{I_k,j}\varphi_{I_k,j}\left(x\right)\right)\right)^2\right] \\ &\quad = -P_n\left(y\varphi_{I_k,j}\left(x\right)\right) + \sum_{(I_l,i)\in\mathcal{I}} \beta_{I_l,i} P_n\left(\varphi_{I_k,j}\varphi_{I_l,i}\right) \ . \end{aligned} \tag{86}$$

We see that if $\beta^{(n)} = \left(\beta^{(n)}_{I_k,j}\right)_{(I_k,j)\in\mathcal{I}} \in \mathbb{R}^D$ 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

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

and by combining (86) with the fact that

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

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

$$\left(I_D + L_{n,D}\right)\beta^{(n)} = X_{y,n} \tag{87}$$

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

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

Now, by inequality (99) in Lemma 10 below, one can find a positive integer $n_0\left(r, A_+, c_{\min}, c_{M,\mathrm{Leb}}, \gamma\right)$ such that for all $n \geq n_0$, we have on an event Ω_n of probability at least $1 - 3Dn^{-\gamma}$,

$$\left\|L_{n,D}\right\| \leq \frac{1}{2} , \tag{88}$$

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 (88) that $(I_D + L_{n,D})$ is a non-singular $D\times D$ matrix and, as a consequence, that the linear system (87) admits a unique solution $\beta^{(n)}$ on Ω_n for any $n \geq n_0\left(r, A_+, c_{\min}, c_{M,\mathrm{Leb}}, \gamma\right)$. 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 $\left(\beta_{I_k,j}\right)_{(I_k,j)\in\mathcal{I}} \in \mathbb{R}^D$ 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,

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

is the unique least squares estimator on M, and by (87) it holds,

$$\beta^{(n)}_{I_k,j}\left(1 + \sum_{(I_l,i)\in\mathcal{I}}\left(P_n - P\right)\left(\varphi_{I_k,j}\varphi_{I_l,i}\right)\right) = P_n\left(y\varphi_{I_k,j}(x)\right) \text{ , for all } (I_k,j)\in\mathcal{I}. \tag{89}$$

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 (89) reduces to

$$\beta^{(n)}_{I_k,j}\times\left(1 + \sum_{i=0}^{r}\left(P_n - P\right)\left(\varphi_{I_k,j}\varphi_{I_k,i}\right)\right) = P_n\left(y\varphi_{I_k,j}(x)\right) \text{ , for all } (I_k,j)\in\mathcal{I} . \tag{90}$$

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

$$\begin{aligned}
\|s_n - s_M\|_\infty &= \left\| \sum_{(I_k,j)\in\mathcal{I}} \left(\beta^{(n)}_{I_k,j} - P\left(y\varphi_{I_k,j}(x)\right)\right)\varphi_{I_k,j} \right\|_\infty \\
&\leq \max_{k\in\{0,\dots,m-1\}} \left\| \sum_{j=0}^{r} \left(\beta^{(n)}_{I_k,j} - P\left(y\varphi_{I_k,j}(x)\right)\right)\varphi_{I_k,j} \right\|_\infty \\
&\leq (r+1) \max_{k\in\{0,\dots,m-1\}} \left\{ \left(\max_{j\in\{0,\dots,r\}} \left|\beta^{(n)}_{I_k,j} - P\left(y\varphi_{I_k,j}(x)\right)\right| \right) \right. \\
&\qquad \left. \times \max_{j\in\{0,\dots,r\}} \left\|\varphi_{I_k,j}\right\|_\infty \right\} \qquad (91)
\end{aligned}$$

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 (91). Let the index (I_k, j) be fixed. By subtracting the quantity $\left(1 + \sum_{i=0}^{r} (P_n - P)\left(\varphi_{I_k,j}\varphi_{I_k,i}\right)\right) \times P\left(y\varphi_{I_k,j}(x)\right)$ in each side of equation (90), we get

$$\begin{aligned}
&\left(\beta^{(n)}_{I_k,j} - P\left(y\varphi_{I_k,j}(x)\right)\right) \times \left(1 + \sum_{i=0}^{r} (P_n - P)\left(\varphi_{I_k,j}\varphi_{I_k,i}\right)\right) \\
&= (P_n - P)\left(y\varphi_{I_k,j}(x)\right) - \left(\sum_{i=0}^{r} (P_n - P)\left(\varphi_{I_k,j}\varphi_{I_k,i}\right)\right) \times P\left(y\varphi_{I_k,j}(x)\right) \ . \qquad (92)
\end{aligned}$$

Moreover, by Inequality (100) of Lemma 10, we have for all $n \geq n_0\left(r, A_+, c_{\min}, c_{M,\mathrm{Leb}}, \gamma\right)$,

$$\sum_{i=0}^{r} \left|(P_n - P)\left(\varphi_{I_k,j}\varphi_{I_k,i}\right)\right| \leq L_{r,A_+,c_{\min},c_{M,\mathrm{Leb}},\gamma} \sqrt{\frac{\ln n}{n \operatorname{Leb}(I_k)}} \leq \frac{1}{2} \qquad (93)$$

on the event Ω_n. We thus deduce that

$$\left| \left(\beta^{(n)}_{I_k,j} - P\left(y\varphi_{I_k,j}(x)\right)\right) \times \left(1 + \sum_{i=0}^{r} (P_n - P)\left(\varphi_{I_k,j}\varphi_{I_k,i}\right)\right) \right| \geq \frac{1}{2} \left|\beta^{(n)}_{I_k,j} - P\left(y\varphi_{I_k,j}(x)\right)\right| \qquad (94)$$

and

$$\left| \left(\sum_{i=0}^{r} (P_n - P)\left(\varphi_{I_k,j}\varphi_{I_k,i}\right)\right) \times P\left(y\varphi_{I_k,j}(x)\right) \right| \leq L_{r,A_+,c_{\min},c_{M,\mathrm{Leb}},\gamma} \sqrt{\frac{\ln n}{n \operatorname{Leb}(I_k)}} \times \left|P\left(y\varphi_{I_k,j}(x)\right)\right| \ . \qquad (95)$$

Moreover, by (28), (65) and (85) we have

$$\begin{aligned}
\left|P\left(y\varphi_{I_k,j}(x)\right)\right| &\leq A \left\|\varphi_{I_k,j}\right\|_\infty P(I_k) \\
&\leq A c_{\max} \left\|\varphi_{I_k,j}\right\|_\infty \operatorname{Leb}(I_k) \\
&\leq A c_{\max} L_{r,c_{\min}} \sqrt{\operatorname{Leb}(I_k)} \\
&\leq L_{A,r,c_{\min},c_{\max}} \sqrt{\operatorname{Leb}(I_k)} \ . \qquad (96)
\end{aligned}$$

Putting inequality (96) in (95) we obtain

$$\left| \left(\sum_{i=0}^{r} (P_n - P)\left(\varphi_{I_k,j}\varphi_{I_k,i}\right)\right) \times P\left(y\varphi_{I_k,j}(x)\right) \right| \leq L_{r,A_+,c_{\min},c_{\max},c_{M,\mathrm{Leb}},\gamma} \sqrt{\frac{\ln n}{n}} \ . \qquad (97)$$

Hence, using inequalities (94), (97) and inequality (101) of Lemma 10 in equation (92), we obtain that

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

on Ω_n. Since the constant $L_{A,r,A_+,c_{\min},c_{\max},c_{M,\mathrm{Leb}},\gamma}$ does not depend on the index (I_k,j) we deduce by (85) that

$$\begin{aligned}
&\left(\max_{j\in\{0,\dots,r\}}\left|\beta_{I_k,j}^{(n)}-P\left(y\varphi_{I_k,j}(x)\right)\right|\right)\times\max_{j\in\{0,\dots,r\}}\left\|\varphi_{I_k,j}\right\|_\infty\\
&\leq L_{A,r,A_+,c_{\min},c_{\max},c_{M,\mathrm{Leb}},\gamma}\sqrt{\frac{\ln n}{n}}\times\max_{j\in\{0,\dots,r\}}\left\|\varphi_{I_k,j}\right\|_\infty\\
&\leq L_{A,r,A_+,c_{\min},c_{\max},c_{M,\mathrm{Leb}},\gamma}\sqrt{\frac{\ln n}{n\,\mathrm{Leb}\left(I_k\right)}}\ . \qquad (98)
\end{aligned}$$

Finally, by using (66) and (98) in (91), we get for all $n\geq n_0\left(r,A_+,c_{\min},c_{M,\mathrm{Leb}},\gamma\right)$, on the event Ω_n of probability at least $1-3Dn^{-\gamma}$,

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

To conclude, simply take $\gamma=\frac{\ln 3}{\ln 2}+\alpha+1$, so that it holds for $n\geq 2$, $\mathbb{P}\left[\Omega_n^c\right]\leq n^{-\alpha}$ which implies (67).
It remains to prove the following lemma that has been used all along the proof.

Lemma 10 *Recall that* $L_{n,D}=\left(\left(L_{n,D}\right)_{(I_k,j),(I_l,i)}\right)_{(I_k,j),(I_l,i)\in\mathcal{I}\times\mathcal{I}}$ *is a* $D\times D$ *matrix such that for all* $(k,l)\in\left\{0,...,m-1\right\}^2$, $(j,i)\in\left\{0,...,r\right\}^2$,

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

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 8, a positive integer $n_0\left(r,A_+,c_{\min},c_{M,\mathrm{Leb}},\gamma\right)$ *exists such that, for all* $n\geq n_0\left(r,A_+,c_{\min},c_{M,\mathrm{Leb}},\gamma\right)$*, the following inequalities hold on an event* Ω_n *of probability at least* $1-3Dn^{-\gamma}$*,*

$$\left\|L_{n,D}\right\|\leq L_{r,A_+,c_{\min},c_{M,\mathrm{Leb}},\gamma}\sqrt{\frac{D\ln n}{n}}\leq\frac{1}{2} \qquad (99)$$

and for all $k\in\left\{0,...,m-1\right\}$,

$$\max_{j\in\{0,\dots,r\}}\left\{\sum_{i=0}^{r}\left|\left(P_n-P\right)\left(\varphi_{I_k,j}\varphi_{I_k,i}\right)\right|\right\}\leq L_{r,A_+,c_{\min},c_{M,\mathrm{Leb}},\gamma}\sqrt{\frac{\ln n}{n\,\mathrm{Leb}\left(I_k\right)}}\leq\frac{1}{2}\ , \qquad (100)$$

$$\max_{j\in\{0,\dots,r\}}\left|\left(P_n-P\right)\left(y\varphi_{I_k,j}(x)\right)\right|\leq L_{A,A_+,r,c_{\min},c_{M,\mathrm{Leb}},\gamma}\sqrt{\frac{\ln n}{n}}\ . \qquad (101)$$

Proof of Lemma 10. Let us begin with the proof of inequality (101). Let the index $(I_k,j)\in\mathcal{I}$ be fixed. By using Bernstein's inequality (230) and observing that, by (28),

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

and, by (28), (85) and (66),

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

we get

$$\mathbb{P}\left[\left|(P_n-P)\left(y\varphi_{I_k,j}(x)\right)\right|\ge\sqrt{2A^2\frac{x}{n}}+\frac{L_{A,r,c_{\min},c_{M,\mathrm{Leb}}}\sqrt{D}}{3n}x\right]\le 2\exp(-x)\ . \tag{102}$$

By taking $x=\gamma\ln n$ in inequality (102), we obtain that

$$\mathbb{P}\left[\left|(P_n-P)\left(y\varphi_{I_k,j}(x)\right)\right|\ge\sqrt{2A^2\gamma\frac{\ln n}{n}}+\frac{L_{A,r,c_{\min},c_{M,\mathrm{Leb}}}\sqrt{D}\gamma\ln n}{3n}\right]\le 2n^{-\gamma}\ . \tag{103}$$

Now, as $D\le A_+ n(\ln n)^{-2}$, we deduce from (103) that for some well chosen positive constant $L_{A,A_+,r,c_{\min},c_{M,\mathrm{Leb}},\gamma}$, we have

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

and by setting

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

we deduce that

$$\mathbb{P}\left(\Omega_n^{(1)}\right)\ge 1-2Dn^{-\gamma}\ . \tag{104}$$

Hence the expected bound (101) holds on $\Omega_n^{(1)}$, for all $n\ge 1$.
We turn now to the proof of inequality (100). Let the index $(I_k,j)\in\mathcal{I}$ be fixed. By Cauchy-Schwarz inequality, we have

$$\sum_{i=0}^{r}\left|(P_n-P)\left(\varphi_{I_k,j}\varphi_{I_k,i}\right)\right|\le\sqrt{r+1}\sqrt{\sum_{i=0}^{r}\left((P_n-P)\left(\varphi_{I_k,j}\varphi_{I_k,i}\right)\right)^2}\ . \tag{105}$$

Let write

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

By Cauchy-Schwarz inequality again, it holds

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

Then, Bousquet's inequality (231), applied with $\varepsilon=1$ and $\mathcal{F}=B_{I_k}$, implies that

$$\mathbb{P}\left[\chi_{I_k,j}-\mathbb{E}\left[\chi_{I_k,j}\right]\ge\sqrt{2\sigma_{I_k,j}^2\frac{x}{n}}+\mathbb{E}\left[\chi_{I_k,j}\right]+\frac{4}{3}\frac{b_{I_k,j}x}{n}\right]\le\exp(-x) \tag{106}$$

where, by (85),

$$\sigma_{I_k,j}^2=\sup_{s\in B_{I_k}}\mathrm{Var}\left(\varphi_{I_k,j}s\right)\le\left\|\varphi_{I_k,j}\right\|_\infty^2\le\frac{L_{r,c_{\min}}}{\mathrm{Leb}(I_k)} \tag{107}$$

and

$$b_{I_k,j} \leq 2 \sup_{s \in B_{I_k}} \left\| \varphi_{I_k,j} s \right\|_\infty \leq 2 \left\| \varphi_{I_k,j} \right\|_\infty \sup_{s \in B_{I_k}} \|s\|_\infty \ . \tag{108}$$

Moreover, for $s = \sum_{i=0}^r \beta_{I_k,i} \varphi_{I_k,i} \in B_{I_k}$, we have $\max_i \left|\beta_{I_k,i}\right| \leq \sqrt{\sum_{i=0}^r \beta_{I_k,i}^2} \leq 1$, so by (85),

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

and injecting the last bound in (108) we get

$$b_{I_k,j} \leq \left\| \varphi_{I_k,j} \right\|_\infty \frac{L_{r,c_{\min}}}{\sqrt{\mathrm{Leb}\left(I_k\right)}} \leq \frac{L_{r,c_{\min}}}{\mathrm{Leb}\left(I_k\right)} \ . \tag{109}$$

In addition, we have

$$\begin{aligned}
\mathbb{E}\left[\chi_{I_k,j}\right] &\leq \sqrt{\mathbb{E}\left[\chi_{I_k,j}^2\right]} = \sqrt{\frac{\sum_{i=0}^r \mathrm{Var}\left(\varphi_{I_k,j}\varphi_{I_k,i}\right)}{n}} \\
&\leq \left\| \varphi_{I_k,j} \right\|_\infty \sqrt{\frac{\sum_{i=0}^r P\left(\varphi_{I_k,i}^2\right)}{n}} \\
&= \left\| \varphi_{I_k,j} \right\|_\infty \sqrt{\frac{r+1}{n}} \\
&\leq L_{r,c_{\min}} \sqrt{\frac{1}{n\,\mathrm{Leb}\left(I_k\right)}} \ .
\end{aligned} \tag{110}$$

Therefore, combining (107), (109), (110) and (106) while taking $x = \gamma \ln n$, we get

$$\mathbb{P}\left[\chi_{I_k,j} \geq L_{r,c_{\min},\gamma} \left(\sqrt{\frac{1}{n\,\mathrm{Leb}\left(I_k\right)}} + \sqrt{\frac{\ln n}{n\,\mathrm{Leb}\left(I_k\right)}} + \frac{\ln n}{n\,\mathrm{Leb}\left(I_k\right)} \right) \right] \leq n^{-\gamma} \ . \tag{111}$$

Now, since by (66) and the fact that $D \leq A_+ n \left(\ln n\right)^{-2}$ we have

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

we obtain from (111) that a positive constant $L_{r,A_+,c_{\min},c_{M,\mathrm{Leb}},\gamma}$ exists, depending only on $\gamma, r, A_+, c_{\min}$ and $c_{M,\mathrm{Leb}}$ such that

$$\mathbb{P}\left[\chi_{I_k,j} \geq L_{r,A_+,c_{\min},c_{M,\mathrm{Leb}},\gamma} \sqrt{\frac{\ln n}{n\,\mathrm{Leb}\left(I_k\right)}} \right] \leq n^{-\gamma} \ . \tag{112}$$

Finally, define

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

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

$$\begin{aligned}
&\sqrt{r+1} \times L_{r,A_+,c_{\min},c_{M,\mathrm{Leb}},\gamma} \sqrt{\frac{\ln n}{n\,\mathrm{Leb}\left(I_k\right)}} \\
&\leq L_{r,A_+,c_{\min},c_{M,\mathrm{Leb}},\gamma} \sqrt{\frac{D \ln n}{n}} \\
&\leq L_{r,A_+,c_{\min},c_{M,\mathrm{Leb}},\gamma} \frac{1}{\sqrt{\ln n}} \leq \frac{1}{2} \ .
\end{aligned} \tag{113}$$

Moreover by (112) it holds

$$\mathbb{P}\left(\Omega_n^{(2)}\right) \geq 1 - Dn^{-\gamma} \tag{114}$$

and, by (105), the expected bound (100) holds on $\Omega_n^{(2)}$, for all $n \geq n_0\left(r, A_+, c_{\min}, c_{M,\mathrm{Leb}}, \gamma\right)$.
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,

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

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

$$\begin{aligned}\|L_{n,D}\| &= \max_{(I_k,j)\in\mathcal{I}} \sum_{(I_l,i)\in\mathcal{I}} \left|\left(P_n - P\right)\left(\varphi_{I_k,j}\varphi_{I_l,i}\right)\right| \\ &= \max_{k\in\{0,\ldots,m-1\}} \max_{j\in\{0,\ldots,r\}} \left\{ \sum_{(I_l,i)\in\mathcal{I}} \left|\left(P_n - P\right)\left(\varphi_{I_k,j}\varphi_{I_l,i}\right)\right| \right\} \ . \end{aligned} \tag{115}$$

Thus, using formula (115), inequalities (100), (66) and (113) give that for all $n \geq n_0\left(r, A_+, c_{\min}, c_{M,\mathrm{Leb}}, \gamma\right)$, we have on $\Omega_n^{(2)}$,

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

Finally, by setting $\Omega_n = \Omega_n^{(1)} \bigcap \Omega_n^{(2)}$, we have $\mathbb{P}\left(\Omega_n\right) \geq 1 - 3Dn^{-\gamma}$, and inequalities (100), (99) and (101) are satisfied on Ω_n for all $n \geq n_0\left(r, A_+, c_{\min}, c_{M,\mathrm{Leb}}, \gamma\right)$, which completes the proof of Lemma 10. ■

7.3 Proofs of Section 4

In order to express the quantities of interest in the proofs of Theorems 2 and 3, we need preliminary definitions. Let $\alpha > 0$ be fixed and for $R_{n,D,\alpha}$ defined in (**H5**), see Section 4.1, we set

$$\tilde{R}_{n,D,\alpha} = \max\left\{ R_{n,D,\alpha} \ ; \ A_\infty\sqrt{\frac{D\ln n}{n}} \right\} \tag{116}$$

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

$$\nu_n = \max\left\{ \sqrt{\frac{\ln n}{D}} \ ; \ \sqrt{\frac{D\ln n}{n}} \ ; \ R_{n,D,\alpha} \right\} \ . \tag{117}$$

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

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

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

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

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

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

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

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

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 $s \in M$,

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

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

$$P(\psi_{1,M} \cdot s) = 0 \tag{118}$$

and by (**H1**) inequality (32) holds true, that is

$$\|\psi_{1,M}\|_\infty \leq 4A . \tag{119}$$

Also, for $\mathcal{K}_{1,M}$ defined in Section 4.3, we have

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

for any orthonormal basis $(\varphi_k)_{k=1}^D$ of $(M, \|\cdot\|_2)$. Moreover, inequality (48) holds under (**H1**) and we have

$$\mathcal{K}_{1,M} \leq 2\sigma_{\max} + 4A \leq 6A . \tag{120}$$

Assuming (**H2**), we have from (49)

$$0 < 2\sigma_{\min} \leq \mathcal{K}_{1,M} . \tag{121}$$

Finally, when (**H3**) holds (it is the case when (**H4**) holds), we have by (33),

$$\sup_{s \in M,\ \|s\|_2 \leq 1} \|s\|_\infty \leq A_{3,M}\sqrt{D} \tag{122}$$

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

$$\|\varphi_k\|_\infty \leq A_{3,M}\sqrt{D} . \tag{123}$$

7.3.1 Proofs of the theorems

The proof of Theorem 2 relies on Lemmas 16, 17 and 18 stated in Section 7.4, and that give sharp estimates of suprema of the empirical process on the contrasted functions over slices of interest.

Proof of Theorem 2. 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 2 into four parts, corresponding to the four Inequalities (39), (40), (41) and (42). The values of A_0 and A_∞, respectively defined in (38) and (116), will then be chosen at the end of the proof.

Proof of Inequality (39). Let $r \in (1,2]$ to be chosen later and $C > 0$ such that

$$rC = \frac{D}{4n}\mathcal{K}_{1,M}^2 \ . \tag{124}$$

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

$$\mathbb{P}\left(P\left(Ks_n - Ks_M\right) \leq C\right) \leq \mathbb{P}\left(\left\{P\left(Ks_n - Ks_M\right) \leq C\right\} \bigcap \Omega_{\infty,\alpha}\right) + n^{-\alpha} \tag{125}$$

and also

$$\begin{aligned}
&\mathbb{P}\left(\left\{P\left(Ks_n - Ks_M\right) \leq C\right\} \bigcap \Omega_{\infty,\alpha}\right) \\
&\leq \mathbb{P}\left(\inf_{s\in\mathcal{F}_C} P_n\left(Ks - Ks_M\right) \leq \inf_{s\in\mathcal{F}_{>C}} P_n\left(Ks - Ks_M\right)\right) \\
&\leq \mathbb{P}\left(\inf_{s\in\mathcal{F}_C} P_n\left(Ks - Ks_M\right) \leq \inf_{s\in\mathcal{F}_{(C,rC]}} P_n\left(Ks - Ks_M\right)\right) \\
&= \mathbb{P}\left(\sup_{s\in\mathcal{F}_C} P_n\left(Ks_M - Ks\right) \geq \sup_{s\in\mathcal{F}_{(C,rC]}} P_n\left(Ks_M - Ks\right)\right).
\end{aligned} \tag{126}$$

Now, by (124) and (121) we have

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

where A_4 is defined in Lemma 16. Hence we can apply Lemma 16 with $\alpha = \beta$, $A_l = \sigma_{\min}^2/2$ and $A_{3,M} = r_M(\varphi)$, by Remark 3. Therefore it holds, for all $n \geq n_0\left(A_\infty, A_{cons}, A_+, \sigma_{\min}, \alpha\right)$,

$$\mathbb{P}\left[\sup_{s\in\mathcal{F}_C} P_n\left(Ks_M - Ks\right) \geq \left(1 + L_{A_\infty,A,r_M(\varphi),\sigma_{\min},A_-,\alpha} \times \nu_n\right)\sqrt{\frac{CD}{n}}\mathcal{K}_{1,M} - C\right] \leq 2n^{-\alpha} \ . \tag{127}$$

Moreover, by using (121) and (120) in (124) we get

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

We then apply Lemma 18 with

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

and

$$A_\infty \geq 64\sqrt{2}B_2 A\left(\sigma_{\max} + 2A\right)\sigma_{\min}^{-1} r_M(\varphi) \ , \tag{128}$$

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

$$\mathbb{P}\left(\sup_{s\in\mathcal{F}_{(C,rC]}} P_n\left(Ks_M - Ks\right) \leq \left(1 - L_{A_-,A,A_\infty,\sigma_{\max},\sigma_{\min},r_M(\varphi),\alpha} \times \nu_n\right)\sqrt{\frac{rCD}{n}}\mathcal{K}_{1,M} - rC\right) \leq 2n^{-\alpha} \ . \tag{129}$$

Now, from (127) and (129) 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 $n \geq n_0\left(A_-, A_+, A, A_\infty, A_{cons}, B_2, r_M(\varphi), \sigma_{\max}, \sigma_{\min}, \alpha\right)$, there exists an event of probability at least $1 - 4n^{-\alpha}$ on which

$$\sup_{s\in\mathcal{F}_C} P_n\left(Ks_M - Ks\right) \leq \left(1 + \tilde{A}_0\nu_n\right)\sqrt{\frac{CD}{n}}\mathcal{K}_{1,M} - C \tag{130}$$

and

$$\sup_{s\in\mathcal{F}_{(C,rC]}} P_n\left(Ks_M - Ks\right) \geq \left(1-\tilde{A}_0\nu_n\right)\sqrt{\frac{rCD}{n}}\mathcal{K}_{1,M} - rC\ . \tag{131}$$

Hence, from (130) and (131) we deduce, using (125) and (126), that if we choose $r \in (1,2]$ such that

$$\left(1+\tilde{A}_0\nu_n\right)\sqrt{\frac{CD}{n}}\mathcal{K}_{1,M} - C < \left(1-\tilde{A}_0\nu_n\right)\sqrt{\frac{rCD}{n}}\mathcal{K}_{1,M} - rC \tag{132}$$

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

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

with probability at least $1-5n^{-\alpha}$. Now, by (124) 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 (132) is equivalent to

$$\left(1-2\tilde{A}_0\nu_n\right)r - 2\left(1+\tilde{A}_0\nu_n\right)\sqrt{r} + 1 > 0\ . \tag{133}$$

Moreover, we have by (117) and (**H5**), for all $n \geq n_0\left(A_+, A_-, A_{cons}, \tilde{A}_0, \alpha\right)$,

$$\tilde{A}_0\nu_n \leq \frac{1}{4} \tag{134}$$

and so, for all $n \geq n_0\left(A_+, A_-, A_{cons}, \tilde{A}_0, \alpha\right)$, simple computations involving (134) show that by taking

$$r = 1 + 48\sqrt{\tilde{A}_0\nu_n} \tag{135}$$

inequality (133) is satisfied. Notice that, for all $n \geq n_0\left(A_+, A_-, A_{cons}, \tilde{A}_0, \alpha\right)$ we have $0 < 48\sqrt{\tilde{A}_0\nu_n} < 1$, so that $r \in (1,2)$. Finally, we compute C by (124) and (135), in such a way that for all $n \geq n_0\left(A_+, A_-, A_{cons}, \tilde{A}_0, \alpha\right)$,

$$C = \frac{rC}{r} = \frac{1}{1+48\sqrt{\tilde{A}_0\nu_n}}\frac{1}{4}\frac{D}{n}\mathcal{K}_{1,M}^2 \geq \left(1-48\sqrt{\tilde{A}_0\nu_n}\right)\frac{1}{4}\frac{D}{n}\mathcal{K}_{1,M}^2 > 0 \tag{136}$$

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

Proof of Inequality (40). Let $C > 0$ and $\delta \in \left(0, \frac{1}{2}\right)$ to be chosen later in such a way that

$$\left(1-\delta\right)C = \frac{D}{4n}\mathcal{K}_{1,M}^2 \tag{137}$$

and

$$C \geq \frac{1}{4}\left(1+A_5\nu_n\right)^2\frac{D}{n}\mathcal{K}_{1,M}^2\ , \tag{138}$$

where A_5 is defined in Lemma 17. We have by (**H5**), for all $n \geq n_1$,

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

and also

$$\begin{aligned}
&\mathbb{P}\left(\left\{P\left(Ks_n - Ks_M\right) > C\right\} \bigcap \Omega_{\infty,\alpha}\right) \\
&\leq \mathbb{P}\left(\inf_{s\in\mathcal{F}_C} P_n\left(Ks - Ks_M\right) \geq \inf_{s\in\mathcal{F}_{>C}} P_n\left(Ks - Ks_M\right)\right) \\
&= \mathbb{P}\left(\sup_{s\in\mathcal{F}_C} P_n\left(Ks_M - Ks\right) \leq \sup_{s\in\mathcal{F}_{>C}} P_n\left(Ks_M - Ks\right)\right) \\
&\leq \mathbb{P}\left(\sup_{s\in\mathcal{F}_{\left[\frac{C}{2},(1-\delta)C\right]}} P_n\left(Ks_M - Ks\right) \leq \sup_{s\in\mathcal{F}_{>C}} P_n\left(Ks_M - Ks\right)\right) . \qquad (140)
\end{aligned}$$

Now by (138) we can apply Lemma 17 with $\alpha = \beta$ and we obtain, for all $n \geq n_0\left(A_\infty, A_{cons}, A_+, \alpha\right)$,

$$\mathbb{P}\left[\sup_{s\in\mathcal{F}_{>C}} P_n\left(Ks_M - Ks\right) \geq \left(1 + A_5\nu_n\right)\sqrt{\frac{CD}{n}}\mathcal{K}_{1,M} - C\right] \leq 2n^{-\alpha} \qquad (141)$$

where A_5 only depends on $A, A_{3,M}, A_\infty, \sigma_{\min}, A_-$ and α. Moreover, we can take $A_{3,M} = r_M\left(\varphi\right)$ by Remark 3. Also, by (137), (121) and (120) we can apply Lemma 18 with the quantity C in Lemma 18 replaced by $C/2$, $\alpha = \beta$, $r = 2\left(1-\delta\right)$, $A_u = \left(\sigma_{\max} + 2A\right)^2$, $A_l = \sigma_{\min}^2$ and the constant A_∞ satisfying

$$A_\infty \geq 64\sqrt{2}B_2A\left(\sigma_{\max} + 2A\right)\sigma_{\min}^{-1}r_M\left(\varphi\right) , \qquad (142)$$

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

$$\mathbb{P}\left(\begin{array}{c} \sup_{s\in\mathcal{F}_{\left[\frac{C}{2},(1-\delta)C\right]}} P_n\left(Ks_M - Ks\right) \\ \leq \left(1 - L_{A_-,A,A_\infty,\sigma_{\max},\sigma_{\min},r_M(\varphi),\alpha} \times \nu_n\right)\sqrt{\frac{(1-\delta)CD}{n}}\mathcal{K}_{1,M} - \left(1-\delta\right)C \end{array}\right) \leq 2n^{-\alpha} . \qquad (143)$$

Hence from (141) and (143), we deduce that a positive constant $\check{A}_0$ exists, only depending on $A_-, A, A_\infty, \sigma_{\max}, \sigma_{\min}, r_M\left(\varphi\right)$ and α, such that
for all $n \geq n_0\left(A_-, A_+, A, A_\infty, A_{cons}, B_2, r_M\left(\varphi\right), \sigma_{\max}, \sigma_{\min}, \alpha\right)$ it holds on an event of probability at least $1 - 4n^{-\alpha}$,

$$\sup_{s\in\mathcal{F}_{\left[\frac{C}{2},(1-\delta)C\right]}} P_n\left(Ks_M - Ks\right) \geq \left(1 - \check{A}_0\nu_n\right)\sqrt{\frac{\left(1-\delta\right)CD}{n}}\mathcal{K}_{1,M} - \left(1-\delta\right)C \qquad (144)$$

and

$$\sup_{s\in\mathcal{F}_{>C}} P_n\left(Ks_M - Ks\right) \leq \left(1 + \check{A}_0\nu_n\right)\sqrt{\frac{CD}{n}}\mathcal{K}_{1,M} - C . \qquad (145)$$

Now, from (144) and (145) we deduce, using (139) and (140), that if we choose $\delta \in \left(0, \frac{1}{2}\right)$ such that (138) and

$$\left(1 + \check{A}_0\nu_n\right)\sqrt{\frac{CD}{n}}\mathcal{K}_{1,M} - C < \left(1 - \check{A}_0\nu_n\right)\sqrt{\frac{\left(1-\delta\right)CD}{n}}\mathcal{K}_{1,M} - \left(1-\delta\right)C \qquad (146)$$

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

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

with probability at least $1 - 5n^{-\alpha}$. By (137) it holds

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

and by consequence, inequality (146) is equivalent to

$$\left(1 - 2\check{A}_0\nu_n\right)\left(1-\delta\right) - 2\left(1 + \check{A}_0\nu_n\right)\sqrt{1-\delta} + 1 > 0 . \qquad (147)$$

Moreover, we have by (117) and (**H5**), for all $n \geq n_0\left(A_+, A_-, A_{cons}, \check{A}_0, A_5, \alpha\right)$,

$$\left(\check{A}_0 \vee A_5\right)\nu_n < \frac{1}{72} \tag{148}$$

and so, for all $n \geq n_0\left(A_+, A_-, A_{cons}, \check{A}_0, \alpha\right)$, simple computations involving (148) show that by taking

$$\delta = 6\left(\sqrt{\check{A}_0} \vee \sqrt{A_5}\right)\sqrt{\nu_n}\ , \tag{149}$$

inequalities (147) and (138) are satisfied and $\delta \in \left(0, \frac{1}{2}\right)$. Finally, we can compute C by (137) and (149), in such a way that for all $n \geq n_0\left(A_+, A_-, A_{cons}, \check{A}_0, \alpha\right)$

$$0 < C = \frac{(1-\delta)\,C}{(1-\delta)} = \frac{1}{(1-\delta)}\frac{1}{4}\frac{D}{n}\mathcal{K}_{1,M}^2 \leq \left(1 + 12\left(\sqrt{\check{A}_0} \vee \sqrt{A_5}\right)\sqrt{\nu_n}\right)\frac{1}{4}\frac{D}{n}\mathcal{K}_{1,M}^2\ , \tag{150}$$

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

Proof of Inequality (41). Let $C = \frac{D}{8n}\mathcal{K}_{1,M}^2 > 0$ and let $r = 2$. By (120) and (121) we have

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

so we can apply Lemma 18 with $\alpha = \beta$, $A_l = \sigma_{\min}^2$ and $A_u = \left(\sigma_{\max} + 2A\right)^2$. So if

$$A_\infty \geq 64\sqrt{2}B_2 A\left(\sigma_{\max} + 2A\right)\sigma_{\min}^{-1} r_M\left(\varphi\right)\ , \tag{151}$$

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

$$\mathbb{P}\left(\sup_{s \in \mathcal{F}_{(C,rC]}} P_n\left(Ks_M - Ks\right) \leq \left(1 - L_{A_-,A,A_\infty,\sigma_{\max},\sigma_{\min},r_M(\varphi),\alpha} \times \nu_n\right)\sqrt{\frac{rCD}{n}}\mathcal{K}_{1,M} - rC\right) \leq 2n^{-\alpha}\ . \tag{152}$$

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 (152), we get

$$\mathbb{P}\left(\sup_{s \in \mathcal{F}_{(C,rC]}} P_n\left(Ks_M - Ks\right) \leq \left(1 - \hat{A}_0\nu_n\right)\frac{D}{4n}\mathcal{K}_{1,M}^2\right) \leq 2n^{-\alpha}\ . \tag{153}$$

Notice that

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

so from (153) we deduce that

$$\mathbb{P}\left(P_n\left(Ks_M - Ks_n\right) \geq \left(1 - \hat{A}_0\nu_n\right)\frac{D}{4n}\mathcal{K}_{1,M}^2\right) \geq 1 - 2n^{-\alpha}\ . \tag{154}$$

Remark 4 *Notice that in the proof of inequality (41), we do not need to assume the consistency of the least squares estimator s_n towards the projection s_M. Straightforward adaptations of Lemma 18 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 (117). This readily gives the expected bound (43) of Theorem 2.

Proof of Inequality (42). Let

$$C = \frac{1}{4}\left(1 + A_5\nu_n\right)^2 \frac{D}{n}\mathcal{K}_{1,M}^2 > 0 \tag{155}$$

where A_5 is defined in Lemma 17 applied with $\beta = \alpha$. By (**H5**) we have

$$\mathbb{P}\left(P_n\left(Ks_M - Ks_n\right) > C\right) \leq \mathbb{P}\left(\left\{P_n\left(Ks_M - Ks_n\right) > C\right\} \bigcap \Omega_{\infty,\alpha}\right) + n^{-\alpha} \ . \tag{156}$$

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

$$\begin{aligned} P_n\left(Ks_M - Ks_n\right) &= \sup_{s \in B_{(M,L_\infty)}\left(s_M, \tilde{R}_{n,D,\alpha}\right)} P_n\left(Ks_M - Ks\right) \\ &= \sup_{s \in \mathcal{F}_{>0}} P_n\left(Ks_M - Ks\right) \end{aligned} \tag{157}$$

and by (215) of Lemma 17 applied with $\alpha = \beta$ it holds, for all $n \geq n_0\left(A_\infty, A_{cons}, A_+, \alpha\right)$,

$$\mathbb{P}\left(\sup_{s \in \mathcal{F}_{>0}} P_n\left(Ks_M - Ks\right) > C\right) \leq 2n^{-\alpha} \ . \tag{158}$$

Finally, using (157) and (158) in (156) we get, for all $n \geq n_0\left(A_\infty, A_{cons}, n_1, A_+, \alpha\right)$,

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

Conclusion. To complete the proof of Theorem 2, just notice that by (128), (142) and (151) we can take

$$A_\infty = 64\sqrt{2}B_2 A\left(\sigma_{\max} + 2A\right)\sigma_{\min}^{-1} r_M\left(\varphi\right)$$

and by (136), (150), (154) and (155),

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

is convenient. ■

Proof of Theorem 3. We localize our analysis in the subset

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

Unlike in the proof of Theorem 2, see (116), 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 2 is only needed in Lemma 12, 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 3, we just need upper bounds, and Lemma 12 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 $C \geq 0$,

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

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

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

I. Proof of Inequality (44). Let $C_1 > 0$ to be fixed later, satisfying

$$C_1 \geq \frac{D}{n} =: C_- > 0 \ . \tag{159}$$

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

$$\mathbb{P}(P(Ks_n - Ks_M) > C_1) \leq \mathbb{P}\left(\{P(Ks_n - Ks_M) > C_1\} \bigcap \Omega_{\infty,\alpha}\right) + n^{-\alpha} \tag{160}$$

and also

$$\begin{aligned}
&\mathbb{P}\left(\{P(Ks_n - Ks_M) > C_1\} \bigcap \Omega_{\infty,\alpha}\right) \\
&\leq \mathbb{P}\left(\inf_{s \in \mathcal{G}_{C_1}} P_n(Ks - Ks_M) \geq \inf_{s \in \mathcal{G}_{>C_1}} P_n(Ks - Ks_M)\right) \\
&= \mathbb{P}\left(\sup_{s \in \mathcal{G}_{C_1}} P_n(Ks_M - Ks) \leq \sup_{s \in \mathcal{G}_{>C_1}} P_n(Ks_M - Ks)\right) \\
&\leq \mathbb{P}\left(0 \leq \sup_{s \in \mathcal{G}_{>C_1}} P_n(Ks_M - Ks)\right) \ .
\end{aligned} \tag{161}$$

Moreover, it holds

$$\begin{aligned}
&\sup_{s \in \mathcal{G}_{>C_1}} P_n(Ks_M - Ks) \\
&= \sup_{s \in \mathcal{G}_{>C_1}} \left\{P_n\left(\psi_{1,M} \cdot (s_M - s) - \psi_2 \circ (s - s_M)\right)\right\} \\
&= \sup_{s \in \mathcal{G}_{>C_1}} \left\{(P_n - P)\left(\psi_{1,M} \cdot (s_M - s)\right) - (P_n - P)(\psi_2 \circ (s - s_M)) - P(Ks - Ks_M)\right\} \\
&= \sup_{s \in \mathcal{G}_{>C_1}} \left\{(P_n - P)\left(\psi_{1,M} \cdot (s_M - s)\right) - P(Ks - Ks_M) - (P_n - P)(\psi_2 \circ (s - s_M))\right\} \\
&= \sup_{U > C_1} \sup_{s \in \mathcal{D}_U} \left\{(P_n - P)\left(\psi_{1,M} \cdot (s_M - s)\right) - U - (P_n - P)(\psi_2 \circ (s - s_M))\right\} \\
&\leq \sup_{U > C_1} \left\{\sqrt{U}\sqrt{\sum_{k=1}^{D} (P_n - P)^2\left(\psi_{1,M} \cdot \varphi_k\right)} - U + \sup_{s \in \mathcal{G}_U} |(P_n - P)(\psi_2 \circ (s - s_M))|\right\} \ .
\end{aligned} \tag{162}$$

Now, from inequality (181) of Lemma 11 applied with $\beta = \alpha$, we get

$$\mathbb{P}\left[\sqrt{\sum_{k=1}^{D} (P_n - P)^2\left(\psi_{1,M} \cdot \varphi_k\right)} \geq L_{A,A_{3,M},\alpha}\sqrt{\frac{D \vee \ln n}{n}}\right] \leq n^{-\alpha} \ . \tag{163}$$

In addition, we handle the empirical process indexed by the second order terms by straightforward modifications of Lemmas 14 and 15 as well as their proofs. It thus holds, by the same type of arguments as those given in Lemma 14,

$$\mathbb{E}\left[\sup_{s \in \mathcal{G}_{C_1}} \left|(P_n - P)\left(\psi^s_{2,M} \cdot (s - s_M)\right)\right|\right] \leq 8\sqrt{\frac{CD}{n}} R_{n,D,\alpha} \ . \tag{164}$$

Moreover, using (164), the same type of arguments as those leading to inequality (208) of Lemma 15, allow to show that for any $q \geq 1$ and $j \in \mathbb{N}^*$, for all $x > 0$,

$$\begin{aligned}
&\mathbb{P}\left[\sup_{s \in \mathcal{G}_{q^j C_-}} |(P_n - P)(\psi_2 \circ (s - s_M))| \geq 16\sqrt{\frac{q^j C_- D}{n}} R_{n,D,\alpha} + \sqrt{\frac{2R^2_{n,D,\alpha} q^j C_- x}{n}} + \frac{8}{3}\frac{R^2_{n,D,\alpha} x}{n}\right] \\
&\leq \exp(-x) \ .
\end{aligned} \tag{165}$$

Hence, taking $x = \gamma \ln n$ in (165) and using the fact that $C_- = Dn^{-1} \geq n^{-1}$, we get

$$\mathbb{P}\left[\sup_{s\in\mathcal{G}_{q^j C_-}} |(P_n - P)(\psi_2 \circ (s - s_M))| \geq L_{A_{cons},\gamma} R_{n,D,\alpha} \sqrt{\frac{q^j C_- (D \vee \ln n)}{n}}\right] \leq n^{-\gamma} . \tag{166}$$

Now, by straightforward modifications of the proof of Lemma 15, we get that for all $n \geq n_0(A_{cons})$,

$$\mathbb{P}\left[\forall U > C_-, \quad \sup_{s\in\mathcal{G}_U} |(P_n - P)(\psi_2 \circ (s - s_M))| \leq L_{A_{cons},\alpha} R_{n,D,\alpha} \sqrt{\frac{U (D \vee \ln n)}{n}}\right] \geq 1 - n^{-\alpha} . \tag{167}$$

Combining (162), (163) and (167), we have on an event of probability at least $1 - 2n^{-\alpha}$, for all $n \geq n_0(A_{cons})$,

$$\begin{aligned}
\sup_{s\in\mathcal{G}_{>C_1}} P_n(Ks_M - Ks) &\leq \sup_{U>C_1} \left\{ L_{A,A_{3,M},\alpha} \sqrt{\frac{U (D \vee \ln n)}{n}} - U + L_{A_{cons},\alpha} R_{n,D,\alpha} \sqrt{\frac{U (D \vee \ln n)}{n}} \right\} \\
&\leq \sup_{U>C_1} \left\{ L_{A,A_{cons},A_{3,M},\alpha} (1 + R_{n,D,\alpha}) \sqrt{\frac{U (D \vee \ln n)}{n}} - U \right\} .
\end{aligned} \tag{168}$$

Now, as $R_{n,D,\alpha} \leq A_{cons} (\ln n)^{-1/2}$, we deduce from (168) that for

$$C_1 = L_{A,A_{cons},A_{3,M},\alpha} \frac{D \vee \ln(n)}{n} > C_- \tag{169}$$

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 (160) and (161), this yields inequality (44).

II. Proof of Inequality (45). Let $C_2 > 0$ to be fixed later, satisfying

$$C_2 \geq \frac{D}{n} = C_- > 0 . \tag{170}$$

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

$$\mathbb{P}(P_n(Ks_M - Ks_n) > C_2) \leq \mathbb{P}\left(\{P_n(Ks_M - Ks_n) > C_2\} \bigcap \Omega_{\infty,\alpha}\right) + n^{-\alpha} . \tag{171}$$

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

$$\begin{aligned}
P_n(Ks_M - Ks_n) &= \sup_{s\in B_{(M,L_\infty)}(s_M, R_{n,D,\alpha})} P_n(Ks_M - Ks) \\
&= \max\left\{ \sup_{s\in\mathcal{G}_{C_1}} P_n(Ks_M - Ks) \; ; \; \sup_{s\in\mathcal{G}_{>C_1}} P_n(Ks_M - Ks) \right\} ,
\end{aligned} \tag{172}$$

where C_1 is defined in the first part of the proof dedicated to the establishment of inequality (44). 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 $n \geq n_0(A_{cons})$,

$$\sqrt{\sum_{k=1}^{D} (P_n - P)^2 (\psi_{1,M} \cdot \varphi_k)} \leq L_{A,A_{3,M},\alpha} \sqrt{\frac{D \vee \ln n}{n}} , \tag{173}$$

$$\forall U > C_-, \quad \sup_{s\in\mathcal{G}_U} |(P_n - P)(\psi_2 \circ (s - s_M))| \leq L_{A_{cons},\alpha} R_{n,D,\alpha} \sqrt{\frac{U (D \vee \ln n)}{n}} , \tag{174}$$

and

$$\sup_{s\in\mathcal{G}_{>C_1}} P_n\left(Ks_M - Ks\right) < 0 \ . \tag{175}$$

By (172) and (175), we thus have on $\Omega_{\infty,\alpha}\bigcap\Omega_1$, for all $n \geq n_0\left(A_{cons}\right)$,

$$0 \leq P_n\left(Ks_M - Ks_n\right) = \sup_{s\in\mathcal{G}_{C_1}} P_n\left(Ks_M - Ks\right) \ . \tag{176}$$

In addition, it holds

$$\begin{aligned}
&\sup_{s\in\mathcal{G}_{C_1}} P_n\left(Ks_M - Ks\right) \\
&= \sup_{s\in\mathcal{G}_{C_1}} \left\{P_n\left(\psi_{1,M}\cdot\left(s_M - s\right) - \psi_2\circ\left(s - s_M\right)\right)\right\} \\
&= \sup_{s\in\mathcal{G}_{C_1}} \left\{\left(P_n - P\right)\left(\psi_{1,M}\cdot\left(s_M - s\right)\right) - \left(P_n - P\right)\left(\psi_2\circ\left(s - s_M\right)\right) - P\left(Ks - Ks_M\right)\right\} \\
&\leq \sup_{s\in\mathcal{G}_{C_1}} \left\{\left(P_n - P\right)\left(\psi_{1,M}\cdot\left(s_M - s\right)\right)\right\} + \sup_{s\in\mathcal{G}_{C_1}} \left|\left(P_n - P\right)\left(\psi_2\circ\left(s - s_M\right)\right)\right| \ .
\end{aligned} \tag{177}$$

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

$$\begin{aligned}
\sup_{s\in\mathcal{G}_{C_1}} \left\{\left(P_n - P\right)\left(\psi_{1,M}\cdot\left(s_M - s\right)\right)\right\} &\leq \sqrt{C_1}\sqrt{\sum_{k=1}^{D}\left(P_n - P\right)^2\left(\psi_{1,M}\cdot\varphi_k\right)} \\
&\leq L_{A,A_{3,M},\alpha}\sqrt{\frac{C_1\left(D\vee\ln n\right)}{n}} \qquad \text{by (173)} \\
&= L_{A,A_{cons},A_{3,M},\alpha}\frac{D\vee\ln\left(n\right)}{n} \qquad \text{by (169)} \ ,
\end{aligned} \tag{178}$$

and also, by (174) and (169),

$$\begin{aligned}
\sup_{s\in\mathcal{G}_{C_1}} \left|\left(P_n - P\right)\left(\psi_2\circ\left(s - s_M\right)\right)\right| &\leq L_{A_{cons},\alpha}R_{n,D,\alpha}\sqrt{\frac{C_1\left(D\vee\ln n\right)}{n}} \\
&\leq L_{A,A_{cons},A_{3,M},\alpha}R_{n,D,\alpha}\frac{D\vee\ln\left(n\right)}{n} \ .
\end{aligned} \tag{179}$$

Finally, as $R_{n,D,\alpha} \leq A_{cons}\left(\ln n\right)^{-1/2}$, we deduce from (176), (177), (178) and (179), that it holds on $\Omega_{\infty,\alpha}\bigcap\Omega_1$, for all $n \geq n_0\left(A_{cons}\right)$,

$$P_n\left(Ks_M - Ks_n\right) \leq L_{A,A_{cons},A_{3,M},\alpha}\frac{D\vee\ln\left(n\right)}{n} \ ,$$

and so, this yields to inequality (45) by using (171) and this concludes the proof of Theorem **3**. ■

7.4 Technical Lemmas

We state here some lemmas needed in the proofs of Section 7.3. First, in Lemmas 11, 12 and 13, 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 14 and 15 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 16, 17 and 18 to derive upper and lower bounds for the empirical process indexed by the contrasted functions over slices of interest.

Lemma 11 *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* $(M, \|\cdot\|_2)$,

$$\mathbb{P}\left[\sqrt{\sum_{k=1}^{D}\left(P_n - P\right)^2\left(\psi_{1,M}\cdot\varphi_k\right)} \geq (1+\tau_n)\sqrt{\frac{D}{n}}\mathcal{K}_{1,M}\right] \leq n^{-\beta} \ . \tag{180}$$

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

$$\mathbb{P}\left[\sqrt{\sum_{k=1}^{D}\left(P_n - P\right)^2\left(\psi_{1,M}\cdot\varphi_k\right)} \geq L_{A,A_{3,M},\beta}\sqrt{\frac{D\vee\ln n}{n}}\right] \leq n^{-\beta} \ . \tag{181}$$

Proof. By Cauchy-Schwarz inequality we have

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

Hence, we get by Bousquet's inequality (232) applied with $\mathcal{F} = \left\{\psi_{1,M}\cdot s\ ;\ s\in M,\ \|s\|_2 \leq 1\right\}$, for all $x > 0$, $\delta > 0$,

$$\mathbb{P}\left[\chi_M \geq \sqrt{2\sigma^2\frac{x}{n}} + (1+\delta)\,\mathbb{E}\left[\chi_M\right] + \left(\frac{1}{3}+\frac{1}{\delta}\right)\frac{bx}{n}\right] \leq \exp\left(-x\right) \tag{182}$$

where

$$\sigma^2 \leq \sup_{s\in M,\ \|s\|_2\leq 1} P\left[\left(\psi_{1,M}\cdot s\right)^2\right] \leq \left\|\psi_{1,M}\right\|_\infty^2 \leq 16A^2 \qquad \text{by (119)}$$

and

$$b \leq \sup_{s\in M,\ \|s\|_2\leq 1}\left\|\psi_{1,M}\cdot s - P\left(\psi_{1,M}\cdot s\right)\right\|_\infty \leq 4A\sqrt{D}A_{3,M} \qquad \text{by (118), (119) and (122).}$$

Moreover,

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

So, from (182) it follows that, for all $x > 0$, $\delta > 0$,

$$\mathbb{P}\left[\chi_M \geq \sqrt{32A^2\frac{x}{n}} + (1+\delta)\sqrt{\frac{D}{n}}\mathcal{K}_{1,M} + \left(\frac{1}{3}+\frac{1}{\delta}\right)\frac{4A\sqrt{D}A_{3,M}x}{n}\right] \leq \exp\left(-x\right) \ . \tag{183}$$

Hence, taking $x = \beta\ln n$, $\delta = \frac{\sqrt{\ln n}}{n^{1/4}}$ in (183), we derive by (121) 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 (180). By (120) 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 (183), simple computations give

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

and by consequence, (181) 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 contrasted functions of M belonging to a slice of excess risk. This is done for a model of reasonable dimension.

Lemma 12 *Let $r > 1$ and $C > 0$. Assume that (**H1**), (**H2**), (**H4**) and (34) 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 (116) satisfies

$$A_\infty \geq 64 B_2 A \sqrt{2A_u} \sigma_{\min}^{-1} r_M(\varphi) \ , \tag{184}$$

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

$$\mathbb{E}\left[\sup_{s \in \mathcal{F}_{(C,rC]}} (P_n - P)\left(\psi_{1,M} \cdot (s_M - s)\right)\right] \geq \left(1 - \frac{L_{A,A_l,A_u,\sigma_{\min}}}{\sqrt{D}}\right) \sqrt{\frac{rCD}{n}} \mathcal{K}_{1,M} \ . \tag{185}$$

Our argument leading to Lemma 12 shows that we have to assume that the constant A_∞ introduced in (116) is large enough. In order to prove Lemma 12 the following result is needed.

Lemma 13 *Let $r > 1$, $\beta > 0$ and $C \geq 0$. Assume that (**H1**), (**H2**), (**H4**) and (34) 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

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

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

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

Proof of Lemma 13. By Cauchy-Schwarz inequality, we get

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

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 (234) to χ_M by taking $\mathcal{F} = S_M$ and use the fact that

$$\sup_{s \in S_M} \left\|\psi_{1,M} \cdot s - P\left(\psi_{1,M} \cdot s\right)\right\|_\infty \leq 4A\sqrt{D} r_M(\varphi) \quad \text{by (118), (119) and (**H4**).} \tag{186}$$

$$\sup_{s \in S_M} \mathrm{Var}\left(\psi_{1,M} \cdot s\right) = \sup_{s \in S_M} P\left(\psi_{1,M} \cdot s\right)^2 \leq 16A^2 \quad \text{by (118), (119)}$$

and also, by using (186) in Inequality (229) 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 (234), for all $\varepsilon, x > 0$,

$$\mathbb{P}\left(\chi_M \leq (1-\varepsilon) B_2^{-1} \sqrt{\frac{D}{n}} \mathcal{K}_{1,M} - \sqrt{32A^2 \frac{x}{n}} - \left(1 - \varepsilon + \left(1 + \frac{1}{\varepsilon}\right) x\right) \frac{4A\sqrt{D} r_M(\varphi)}{n}\right) \leq \exp(-x) \ . \quad (187)$$

So, by taking $\varepsilon = \frac{1}{2}$ and $x = \beta \ln n$ in (187), and by observing that $D \geq A_- (\ln n)^2$ and $\mathcal{K}_{1,M} \geq 2\sigma_{\min}$, we conclude that, for all $n \geq n_0(A_-, A, B_2, r_M(\varphi), \sigma_{\min}, \beta)$,

$$\mathbb{P}\left[\chi_M \leq \frac{B_2^{-1}}{8} \sqrt{\frac{D}{n}} \mathcal{K}_{1,M}\right] \leq n^{-\beta} \ . \quad (188)$$

Furthermore, combining Bernstein's inequality (230), with the observation that we have, for every $k \in \{1, ..., D\}$,

$$\begin{aligned} \left\|\psi_{1,M} \cdot \varphi_k\right\|_\infty &\leq 4A\sqrt{D} r_M(\varphi) && \text{by (119) and (\textbf{H4})} \\ P\left(\psi_{1,M} \cdot \varphi_k\right)^2 &\leq \left\|\psi_{1,M}\right\|_\infty^2 \leq 16A^2 && \text{by (119)} \end{aligned}$$

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

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

and so

$$\mathbb{P}\left[\max_{k \in \{1,...,D\}} \left|(P_n - P)\left(\psi_{1,M} \cdot \varphi_k\right)\right| \geq \sqrt{32A^2 \frac{x}{n}} + \frac{4A\sqrt{D} r_M(\varphi)}{3} \frac{x}{n}\right] \leq 2D\exp(-x) \ . \quad (189)$$

Hence, taking $x = \beta \ln n$ in (189), it comes

$$\mathbb{P}\left[\max_{k \in \{1,...,D\}} \left|(P_n - P)\left(\psi_{1,M} \cdot \varphi_k\right)\right| \geq \sqrt{\frac{32A^2 \beta \ln n}{n}} + \frac{4A\sqrt{D} r_M(\varphi) \beta \ln n}{3n}\right] \leq \frac{2D}{n^\beta} \ , \quad (190)$$

then, by using (188) and (190), we get for all $n \geq n_0(A_-, A, B_2, r_M(\varphi), \sigma_{\min}, \beta)$,

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

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

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

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

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

which readily gives the result. ■

We are now ready to prove the lower bound (185) for the expected value of the largest increment of the empirical process over $\mathcal{F}_{(C,rC]}$.

Proof of Lemma 12. Let us begin with the lower bound of

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

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

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

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

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

By Lemma 13 we have that for all $\beta>0$, if $A_\infty\geq 32B_2\sqrt{2A_uA^2\beta}\sigma_{\min}^{-1}r_M\left(\varphi\right)$ then, for all $n\geq n_0\left(A_-,A_+,A,B_2,r_M\left(\varphi\right),\sigma_{\min},\beta\right)$,

$$\mathbb{P}\left(\tilde{\Omega}\right)\geq 1-\frac{2D+1}{n^\beta} . \tag{191}$$

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 $\tilde{\Omega}$,

$$\left(s_M+\sum_{k=1}^{D}\beta_{k,n}\varphi_k\right)\in\mathcal{F}_{(C,rC]} . \tag{192}$$

As a consequence, by (192) it holds

$$\begin{aligned}
&\mathbb{E}^{\frac{1}{2}}\left(\sup_{s\in\mathcal{F}_{(C,rC]}}\left(P_n-P\right)\left(\psi_{1,M}\cdot\left(s_M-s\right)\right)\right)^2\\
&\geq\mathbb{E}^{\frac{1}{2}}\left[\left(\left(P_n-P\right)\left(\psi_{1,M}\cdot\left(\sum_{k=1}^{D}\beta_{k,n}\varphi_k\right)\right)\right)^2\mathbf{1}_{\tilde{\Omega}}\right]\\
&=\sqrt{rC}\sqrt{\mathbb{E}\left[\left(\sum_{k=1}^{D}\left(P_n-P\right)^2\left(\psi_{1,M}\cdot\varphi_k\right)\right)\mathbf{1}_{\tilde{\Omega}}\right]} .
\end{aligned} \tag{193}$$

Furthermore, since by (118) $P\left(\psi_{1,M}\cdot\varphi_k\right)=0$ and by (**H4**) $\left\|\varphi_k\right\|_\infty\leq\sqrt{D}r_M\left(\varphi\right)$ for all $k\in\{1,...,D\}$, we have

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

and it ensures

$$\mathbb{E}\left[\left(\sum_{k=1}^{D}\left(P_n-P\right)^2\left(\psi_{1,M}\cdot\varphi_k\right)\right)1_{\tilde{\Omega}}\right]\geq\mathbb{E}\left[\left(\sum_{k=1}^{D}\left(P_n-P\right)^2\left(\psi_{1,M}\cdot\varphi_k\right)\right)\right]-16A^2D^2r_M^2\left(\varphi\right)\mathbb{P}\left[\left(\tilde{\Omega}\right)^c\right]\ . \quad (194)$$

Comparing inequality (194) with (193) and using (191), we obtain the following lower bound for all $n\geq n_0\left(A_-,A_+,A,B_2,r_M\left(\varphi\right),\sigma_{\min},\beta\right)$,

$$\begin{aligned}\mathbb{E}^{\frac{1}{2}}\left(\sup_{s\in\mathcal{F}_{(C,rC]}}\left(P_n-P\right)\left(\psi_{1,M}\cdot\left(s_M-s\right)\right)\right)^2&\geq\sqrt{rC}\sqrt{\mathbb{E}\left[\left(\sum_{k=1}^{D}\left(P_n-P\right)^2\left(\psi_{1,M}\cdot\varphi_k\right)\right)\right]}\\&\quad-4Ar_M\left(\varphi\right)D\sqrt{rC}\sqrt{\mathbb{P}\left[\left(\tilde{\Omega}\right)^c\right]}\\&\geq\sqrt{\frac{rCD}{n}}\mathcal{K}_{1,M}-4Ar_M\left(\varphi\right)D\sqrt{rC}\sqrt{\frac{2D+1}{n^\beta}}\ . \quad (195)\end{aligned}$$

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

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

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

$$4Ar_M\left(\varphi\right)D\sqrt{rC}\sqrt{\frac{2D+1}{n^\beta}}\leq\frac{1}{\sqrt{D}}\times\sqrt{\frac{rCD}{n}}\mathcal{K}_{1,M} \quad (196)$$

and so, by combining (195) and (196), for all $n\geq n_0\left(A_-,A_+,A,B_2,r_M\left(\varphi\right),\sigma_{\min}\right)$, it holds

$$\mathbb{E}^{\frac{1}{2}}\left(\sup_{s\in\mathcal{F}_{(C,rC]}}\left(P_n-P\right)\left(\psi_{1,M}\cdot\left(s_M-s\right)\right)\right)^2\geq\left(1-\frac{1}{\sqrt{D}}\right)\sqrt{\frac{rCD}{n}}\mathcal{K}_{1,M}\ . \quad (197)$$

Now, as $D\geq A_-\left(\ln n\right)^2$ we have for all $n\geq n_0\left(A_-\right)$, $D^{-1/2}\leq 1/2$. Moreover, we have $\mathcal{K}_{1,M}\geq 2\sigma_{\min}$ by (**H2**) and $rC\geq A_lDn^{-1}$, so we finally deduce from (197) that, for all $n\geq n_0\left(A_-,A_+,A,B_2,A_l,r_M\left(\varphi\right),\sigma_{\min}\right)$,

$$\mathbb{E}^{\frac{1}{2}}\left(\sup_{s\in\mathcal{F}_{(C,rC]}}\left(P_n-P\right)\left(\psi_{1,M}\cdot\left(s_M-s\right)\right)\right)^2\geq\sigma_{\min}\sqrt{A_l}\frac{D}{n}\ . \quad (198)$$

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

$$\mathbb{E}\left[\sup_{s\in\mathcal{F}_{(C,rC]}}\left(P_n-P\right)\left(\psi_{1,M}\cdot\left(s_M-s\right)\right)\right]=\mathbb{E}\left[\sup_{s\in\mathcal{F}_{(C,rC]}}\left|\left(P_n-P\right)\left(\psi_{1,M}\cdot\left(s_M-s\right)\right)\right|\right]. \quad (199)$$

In the next step, we apply Corollary 25. More precisely, using notations of Corollary 25, we set

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

and

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

Now, since for all $n\geq n_0\left(A_+,A_-,A_\infty,A_{cons}\right)$ we have $\tilde{R}_{n,D,\alpha}\leq 1$, we get by (118) and (119), for all $n\geq n_0\left(A_+,A_-,A_\infty,A_{cons}\right)$,

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

we set $b = 4A$. Since we assume that $rC \leq A_u \frac{D}{n}$, it moreover holds by (119),

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

and so we set $\sigma^2 = 16A^2 A_u \frac{D}{n}$. Now, by (198) we have, for all $n \geq n_0\left(A_-, A_+, A, B_2, A_l, r_M\left(\varphi\right), \sigma_{\min}\right)$,

$$\sqrt{\mathbb{E}\left[Z^2\right]} \geq \sigma_{\min} \sqrt{A_l} \frac{D}{n} \ . \tag{200}$$

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

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

we get, using (200), that, for all $n \geq n_0\left(A_-, A_+, A_l, A_u, A, B_2, r_M\left(\varphi\right), A_{cons}, \sigma_{\min}\right)$,

$$\varkappa_n^2 \mathbb{E}\left[Z^2\right] \geq \frac{\sigma^2}{n} \ ,$$

$$\varkappa_n^2 \sqrt{\mathbb{E}\left[Z^2\right]} \geq \frac{b}{n} \ .$$

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

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

So, using (199) and Corollary 25, it holds for all $n \geq n_0\left(A_-, A_+, A_l, A_u, A, B_2, r_M\left(\varphi\right), \sigma_{\min}\right)$,

$$\begin{aligned} &\mathbb{E}\left[\sup_{s \in \mathcal{F}_{(C,rC]}} \left(P_n - P\right)\left(\psi_{1,M} \cdot \left(s_M - s\right)\right)\right] \\ &\geq \left(1 - \frac{L_{A,A_l,A_u,\sigma_{\min}}}{\sqrt{D}}\right) \mathbb{E}^{\frac{1}{2}}\left(\sup_{s \in \mathcal{F}_{(C,rC]}} \left(P_n - P\right)\left(\psi_{1,M} \cdot \left(s_M - s\right)\right)\right)^2 \ . \end{aligned} \tag{201}$$

Finally, by comparing (197) and (201), we deduce that for all $n \geq n_0\left(A_-, A_+, A_l, A_u, A, B_2, r_M\left(\varphi\right), \sigma_{\min}\right)$,

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

and so (185) 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

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

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

$$\Omega_C\left(x\right) = 2 \sup_{s \in \mathcal{F}_C} \left\{\left|s\left(x\right) - s_M\left(x\right)\right|\right\} \tag{202}$$

$$\leq 2\left(\tilde{R}_{n,D,\alpha} \wedge \sqrt{CD} A_{3,M}\right) \ . \tag{203}$$

Lemma 14 *Let $C \geq 0$. Under (**H3**), it holds*

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

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

$$\begin{aligned}
&\mathbb{E}_\varepsilon\left[\sup_{s\in\mathcal{F}_C} |\mathcal{R}_n(\psi_2 \circ (s - s_M))|\right] \\
&= \mathbb{E}_\varepsilon\left[\sup_{s\in\mathcal{F}_C} \left|\mathcal{R}_n\left((s - s_M)^2\right)\right|\right] \\
&\leq \left(\max_{1\leq i\leq n} \Omega_C(X_i)\right) \mathbb{E}_\varepsilon\left[\sup_{s\in\mathcal{F}_C} \left|\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \varphi_i((s - s_M)(X_i))\right|\right] \qquad (204)
\end{aligned}$$

where the functions $\varphi_i : \mathbb{R} \longrightarrow \mathbb{R}$ 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 (202) we deduce that φ_i is a contraction mapping with $\varphi_i(0) = 0$. We thus apply Theorem 21 to get

$$\begin{aligned}
\mathbb{E}_\varepsilon\left[\sup_{s\in\mathcal{F}_C} \left|\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \varphi_i((s - s_M)(X_i))\right|\right] &\leq 2\mathbb{E}_\varepsilon\left[\sup_{s\in\mathcal{F}_C} \left|\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i (s - s_M)(X_i)\right|\right] \\
&= 2\mathbb{E}_\varepsilon\left[\sup_{s\in\mathcal{F}_C} |\mathcal{R}_n(s - s_M)|\right] \qquad (205)
\end{aligned}$$

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

$$\begin{aligned}
&\mathbb{E}\left[\sup_{s\in\mathcal{F}_C} |\mathcal{R}_n(\psi_2 \circ (s - s_M))|\right] = \mathbb{E}\left[\mathbb{E}_\varepsilon\left[\sup_{s\in\mathcal{F}_C} |\mathcal{R}_n(\psi_2 \circ (s - s_M))|\right]\right] \\
&\leq \mathbb{E}\left[\left(\max_{1\leq i\leq n} \Omega_C(X_i)\right) \mathbb{E}_\varepsilon\left[\sup_{s\in\mathcal{F}_C} \left|\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \varphi_i((s - s_M)(X_i))\right|\right]\right] \quad \text{by (204)} \\
&\leq 2\mathbb{E}\left[\left(\max_{1\leq i\leq n} \Omega_C(X_i)\right) \mathbb{E}_\varepsilon\left[\sup_{s\in\mathcal{F}_C} |\mathcal{R}_n(s - s_M)|\right]\right] \quad \text{by (205)} \\
&= 2\mathbb{E}\left[\left(\max_{1\leq i\leq n} \Omega_C(X_i)\right) \sup_{s\in\mathcal{F}_C} |\mathcal{R}_n(s - s_M)|\right] \\
&\leq 2\sqrt{\mathbb{E}\left[\max_{1\leq i\leq n} \Omega_C^2(X_i)\right]} \sqrt{\mathbb{E}\left[\left(\sup_{s\in\mathcal{F}_C} |\mathcal{R}_n(s - s_M)|\right)^2\right]}
\end{aligned}$$

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

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

to complete the proof, it remains to observe that, by (203),

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

■

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 15 *Let $A_+, A_-, A_l, \beta, C_- > 0$, and assume (**H3**) and (34). If $C_- \geq A_l\frac{D}{n}$ and $A_+ n\left(\ln n\right)^{-2} \geq D \geq A_-\left(\ln n\right)^2$, then a positive constant $L_{A_-,A_l,\beta}$ exists such that, for all $n \geq n_0\left(A_\infty, A_{cons}, A_+, A_l\right)$,*

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

Proof. First notice that, as $A_+ n\left(\ln n\right)^{-2} \geq D$, we have by (34),

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

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

$$\tilde{R}_{n,D,\alpha} \leq 1\ . \tag{206}$$

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

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

we have by (206), for all $s \in \cup_{C>C_-}\mathcal{F}_C$ and for all $n \geq n_0\left(A_\infty, A_{cons}, A_+\right)$,

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

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

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

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 $C \longmapsto \sup_{s \in \mathcal{F}_C} |(P_n - P)(\psi_2 \circ (s - s_M))|$, we have uniformly in $C \in \left(q^{j-1} C_-, q^j C_-\right]$,

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

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 $L > 0$,

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

Now, for any $L > 0$,

$$\begin{aligned}
&\mathbb{P}\left[\forall j \in \{1, ..., J\}, \quad \sup_{s \in \mathcal{F}_{q^j C_-}} |(P_n - P)(\psi_2 \circ (s - s_M))| \leq L\sqrt{\frac{q^j C_- D}{n}} \tilde{R}_{n,D,\alpha}\right] \\
&= 1 - \mathbb{P}\left[\exists j \in \{1, ..., J\}, \quad \sup_{s \in \mathcal{F}_{q^j C_-}} |(P_n - P)(\psi_2 \circ (s - s_M))| > L\sqrt{\frac{q^j C_- D}{n}} \tilde{R}_{n,D,\alpha}\right] \\
&\geq 1 - \sum_{j=1}^{J} \mathbb{P}\left[\sup_{s \in \mathcal{F}_{q^j C_-}} |(P_n - P)(\psi_2 \circ (s - s_M))| > L\sqrt{\frac{q^j C_- D}{n}} \tilde{R}_{n,D,\alpha}\right] . \qquad (207)
\end{aligned}$$

Given $j \in \{1, ..., J\}$, Lemma 14 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 (232) to handle the deviations around the mean. We have

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

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

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

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

$$\mathbb{P}\left[\sup_{s\in\mathcal{F}_{q^jC_-}} |(P_n - P)(\psi_2 \circ (s - s_M))| \geq 16\sqrt{\frac{q^jC_-D}{n}}\tilde{R}_{n,D,\alpha} + \sqrt{\frac{2\tilde{R}^2_{n,D,\alpha}q^jC_-x}{n}} + \frac{8}{3}\frac{\tilde{R}^2_{n,D,\alpha}x}{n}\right] \leq \exp(-x) \ . \tag{208}$$

By consequence, as $D \geq A_-(\ln n)^2$ and as $\tilde{R}_{n,D,\alpha} \leq 1$ for all $n \geq n_0(A_\infty, A_{cons}, A_+)$, taking $x = \gamma \ln n$ in (208) for some $\gamma > 0$, easy computations show that a positive constant $L_{A_-,A_l,\gamma}$ independent of j exists such that for all $n \geq n_0(A_\infty, A_{cons}, A_+)$,

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

Hence, using (207), we get for all $n \geq n_0(A_\infty, A_{cons}, A_+)$,

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

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

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 16 *Let $\beta, A_-, A_+, A_l, C > 0$. Assume that **(H1)**, **(H2)**, **(H3)** and (34) hold. A positive constant A_4 exists, only depending on $A, A_{3,M}, \sigma_{\min}, \beta$, such that, if*

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

where $\nu_n = \max\left\{\sqrt{\frac{\ln n}{D}}, \ \sqrt{\frac{D\ln n}{n}}, \ R_{n,D,\alpha}\right\}$ is defined in (117), then for all $n \geq n_0(A_\infty, A_{cons}, A_+, A_l)$,

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

Proof. Start with

$$\begin{aligned}\sup_{s\in\mathcal{F}_C} P_n(Ks_M - Ks) &= \sup_{s\in\mathcal{F}_C}\left\{P_n\left(\psi_{1,M}\cdot(s_M - s) - \psi_2\circ(s - s_M)\right)\right\} \\ &= \sup_{s\in\mathcal{F}_C}\left\{(P_n - P)\left(\psi_{1,M}\cdot(s_M - s)\right) - (P_n - P)(\psi_2\circ(s - s_M)) - P(Ks - Ks_M)\right\} \\ &\leq \sup_{s\in\mathcal{F}_C}\left\{(P_n - P)\left(\psi_{1,M}\cdot(s_M - s)\right) - P(Ks - Ks_M)\right\} \\ &\quad + \sup_{s\in\mathcal{F}_C}|(P_n - P)(\psi_2\circ(s - s_M))| \ . \end{aligned} \tag{209}$$

Next, recall that by definition,

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

so we have

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

where the last bound follows from Cauchy-Schwarz inequality. Hence, we deduce from Lemma 11 that

$$\mathbb{P}\left[\sup_{s\in\mathcal{F}_C}\left\{(P_n-P)\left(\psi_{1,M}\cdot(s_M-s)\right)-P\left(Ks-Ks_M\right)\right\}\ge\sup_{0\le L\le C}\left\{\sqrt{L}\left(1+\tau_n\right)\sqrt{\frac{D}{n}}\mathcal{K}_{1,M}-L\right\}\right]\le n^{-\beta}\ , \tag{210}$$

where

$$\begin{aligned}
\tau_n&=L_{A,A_{3,M},\sigma_{\min},\beta}\left(\sqrt{\frac{\ln n}{D}}\vee\frac{\sqrt{\ln n}}{n^{1/4}}\right)\\
&\le L_{A,A_{3,M},\sigma_{\min},\beta}\left(\sqrt{\frac{\ln n}{D}}\vee\sqrt{\frac{D\ln n}{n}}\right)\\
&\le L_{A,A_{3,M},\sigma_{\min},\beta}\times\nu_n\ .
\end{aligned}\tag{211}$$

So, injecting (211) in (210) we have

$$\mathbb{P}\left[\begin{array}{c}\sup_{s\in\mathcal{F}_C}\left\{(P_n-P)\left(\psi_{1,M}\cdot(s_M-s)\right)-P\left(Ks-Ks_M\right)\right\}\\ \ge\sup_{0\le L\le C}\left\{\sqrt{L}\left(1+L_{A,A_{3,M},\sigma_{\min},\beta}\times\nu_n\right)\sqrt{\frac{D}{n}}\mathcal{K}_{1,M}-L\right\}\end{array}\right]\le n^{-\beta}$$

and since we assume $C\le\frac{1}{4}\left(1+L_{A,A_{3,M},\sigma_{\min},\beta}\times\nu_n\right)^2\frac{D}{n}\mathcal{K}_{1,M}^2$ we see that

$$\sup_{0\le L\le C}\left\{\sqrt{L}\left(1+L_{A,A_{3,M},\sigma_{\min},\beta}\nu_n\right)\sqrt{\frac{D}{n}}\mathcal{K}_{1,M}-L\right\}=\sqrt{C}\left(1+L_{A,A_{3,M},\sigma_{\min},\beta}\times\nu_n\right)\sqrt{\frac{D}{n}}\mathcal{K}_{1,M}-C$$

and therefore

$$\mathbb{P}\left[\sup_{s\in\mathcal{F}_C}\left\{(P_n-P)\left(\psi_{1,M}\cdot(s_M-s)\right)-P\left(Ks-Ks_M\right)\right\}\ge\left(1+L_{A,A_{3,M},\sigma_{\min},\beta}\nu_n\right)\sqrt{\frac{CD}{n}}\mathcal{K}_{1,M}-C\right]\le n^{-\beta}\ . \tag{212}$$

Moreover, as $C\ge A_l\frac{D}{n}$, we derive from Lemma 15 that it holds, for all $n\ge n_0\left(A_\infty,A_{cons},A_+,A_l\right)$,

$$\mathbb{P}\left[\sup_{s\in\mathcal{F}_C}\left|(P_n-P)\left(\psi_2\circ(s-s_M)\right)\right|\ge L_{A_-,A_l,\beta}\sqrt{\frac{CD}{n}}\tilde{R}_{n,D,\alpha}\right]\le n^{-\beta}\ . \tag{213}$$

Finally, noticing that

$$\begin{aligned}
\tilde{R}_{n,D,\alpha}&=\max\left\{R_{n,D,\alpha},A_\infty\sqrt{\frac{D\ln n}{n}}\right\}\\
&\le L_{A_\infty,\sigma_{\min}}\max\left\{R_{n,D,\alpha},\sqrt{\frac{D\ln n}{n}}\right\}\times\mathcal{K}_{1,M}\qquad\text{by (121)}\\
&\le L_{A_\infty,\sigma_{\min}}\times\nu_n\times\mathcal{K}_{1,M}\ ,
\end{aligned}$$

we deduce from (213) that, for all $n \geq n_0\left(A_\infty, A_{cons}, A_+, A_l\right)$,

$$\mathbb{P}\left[\sup_{s\in\mathcal{F}_C}\left|\left(P_n-P\right)\left(\psi_2\circ\left(s-s_M\right)\right)\right| \geq L_{A_\infty,\sigma_{\min},A_-,A_l,\beta}\times\nu_n\sqrt{\frac{CD}{n}}\mathcal{K}_{1,M}\right] \leq n^{-\beta} \tag{214}$$

and the conclusion follows by making use of (212) and (214) in inequality (209). ■
The second deviation bound for the empirical excess risk we need to establish on the upper slice is proved in a similar way.

Lemma 17 *Let $\beta, A_-, A_+, C \geq 0$. Assume that **(H1)**, **(H2)**, **(H3)** and (34) 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*

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

where $\nu_n = \max\left\{\sqrt{\frac{\ln n}{D}},\ \sqrt{\frac{D\ln n}{n}},\ R_{n,D,\alpha}\right\}$ is defined in (117), then for all $n \geq n_0\left(A_\infty, A_{cons}, A_+\right)$,

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

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

$$\mathbb{P}\left[\sup_{s\in\mathcal{F}_{>C}} P_n\left(Ks_M - Ks\right) \geq \frac{1}{4}\left(1+A_5\nu_n\right)^2\frac{D}{n}\mathcal{K}_{1,M}^2\right] \leq 2n^{-\beta}\ . \tag{215}$$

Proof. First observe that

$$\begin{aligned}
\sup_{s\in\mathcal{F}_{>C}} P_n\left(Ks_M - Ks\right) &= \sup_{s\in\mathcal{F}_{>C}}\left\{P_n\left(\psi_{1,M}\cdot\left(s_M-s\right)-\psi_2\circ\left(s-s_M\right)\right)\right\}\\
&= \sup_{s\in\mathcal{F}_{>C}}\left\{\left(P_n-P\right)\left(\psi_{1,M}\cdot\left(s_M-s\right)\right)-\left(P_n-P\right)\left(\psi_2\circ\left(s-s_M\right)\right)-P\left(Ks-Ks_M\right)\right\}\\
&= \sup_{s\in\mathcal{F}_{>C}}\left\{\left(P_n-P\right)\left(\psi_{1,M}\cdot\left(s_M-s\right)\right)-P\left(Ks-Ks_M\right)-\left(P_n-P\right)\left(\psi_2\circ\left(s-s_M\right)\right)\right\}\\
&= \sup_{L>C}\sup_{s\in D_L}\left\{\left(P_n-P\right)\left(\psi_{1,M}\cdot\left(s_M-s\right)\right)-L-\left(P_n-P\right)\left(\psi_2\circ\left(s-s_M\right)\right)\right\}\\
&\leq \sup_{L>C}\left\{\sqrt{L}\sqrt{\sum_{k=1}^D\left(P_n-P\right)^2\left(\psi_{1,M}\cdot\varphi_k\right)}-L+\sup_{s\in\mathcal{F}_L}\left|\left(P_n-P\right)\left(\psi_2\circ\left(s-s_M\right)\right)\right|\right\}
\end{aligned} \tag{216}$$

where the last bound follows from Cauchy-Schwarz inequality. Now, the end of the proof is similar to that of Lemma 16 and follows from the same kind of computations. Indeed, from Lemma 11 we deduce that

$$\mathbb{P}\left[\sqrt{\sum_{k=1}^D\left(P_n-P\right)^2\left(\psi_{1,M}\cdot\varphi_k\right)} \geq \left(1+L_{A,A_{3,M},\sigma_{\min},\beta}\times\nu_n\right)\sqrt{\frac{D}{n}}\mathcal{K}_{1,M}\right] \leq n^{-\beta} \tag{217}$$

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 15 with $A_l = \sigma_{\min}^2$, and deduce that, for all $n \geq n_0\left(A_\infty, A_{cons}, A_+\right)$,

$$\mathbb{P}\left[\forall L > C,\quad \sup_{s\in\mathcal{F}_L}\left|\left(P_n-P\right)\left(\psi_{2,M}^s\cdot\left(s-s_M\right)\right)\right| \geq L_{A_\infty,\sigma_{\min},A_-,\beta}\times\nu_n\sqrt{\frac{LD}{n}}\mathcal{K}_{1,M}\right] \leq n^{-\beta}\ . \tag{218}$$

Now using (217) and (218) in (216) we obtain, for all $n \geq n_0 (A_\infty, A_{cons}, A_+)$,

$$\mathbb{P}\left[\sup_{s\in\mathcal{F}_{>C}} P_n (Ks_M - Ks) \geq \sup_{L>C}\left\{\left(1 + L_{A,A_{3,M},A_\infty,\sigma_{\min},A_-,\beta} \times \nu_n\right)\sqrt{\frac{LD}{n}}\mathcal{K}_{1,M} - L\right\}\right] \leq 2n^{-\beta} \qquad (219)$$

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 (219). 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 (219) 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 18 *Let $r > 1$ and $C, \beta > 0$. Assume that (**H1**), (**H2**), (**H4**) and (34) 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 (116) satisfies

$$A_\infty \geq 64 B_2 A\sqrt{2A_u}\sigma_{\min}^{-1} r_M(\varphi),$$

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

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

where $\nu_n = \max\left\{\sqrt{\frac{\ln n}{D}},\ \sqrt{\frac{D\ln n}{n}},\ R_{n,D,\alpha}\right\}$ is defined in (117).

Proof. Start with

$$\begin{aligned}
&\sup_{s\in\mathcal{F}_{(C,rC]}} P_n (Ks_M - Ks) \\
&= \sup_{s\in\mathcal{F}_{(C,rC]}} \{(P_n - P)(Ks_M - Ks) + P(Ks_M - Ks)\} \\
&\geq \sup_{s\in\mathcal{F}_{(C,rC]}} (P_n - P)\left(\psi_{1,M}\cdot(s_M - s)\right) - \sup_{s\in\mathcal{F}_{(C,rC]}} (P_n - P)(\psi_2 \circ (s - s_M)) - \sup_{s\in\mathcal{F}_{(C,rC]}} P(Ks - Ks_M) \\
&\geq \sup_{s\in\mathcal{F}_{(C,rC]}} (P_n - P)\left(\psi_{1,M}\cdot(s_M - s)\right) - \sup_{s\in\mathcal{F}_{rC}} (P_n - P)(\psi_2 \circ (s - s_M)) - rC \qquad (220)
\end{aligned}$$

and set

$$
\begin{aligned}
S_{1,r,C} &= \sup_{s\in\mathcal{F}_{(C,rC]}} (P_n - P)\left(\psi_{1,M}\cdot(s_M - s)\right)\\
M_{1,r,C} &= \mathbb{E}\left[\sup_{s\in\mathcal{F}_{(C,rC]}} (P_n - P)\left(\psi_{1,M}\cdot(s_M - s)\right)\right]\\
b_{1,r,C} &= \sup_{s\in\mathcal{F}_{(C,rC]}} \left\|\psi_{1,M}\cdot(s_M - s) - P\left(\psi_{1,M}\cdot(s_M - s)\right)\right\|_\infty\\
\sigma^2_{1,r,C} &= \sup_{s\in\mathcal{F}_{(C,rC]}} \mathrm{Var}\left(\psi_{1,M}\cdot(s_M - s)\right).
\end{aligned}
$$

By Klein-Rio's Inequality (234), we get, for all $\delta, x > 0$,

$$
\mathbb{P}\left(S_{1,r,C} \le (1-\delta)M_{1,r,C} - \sqrt{\frac{2\sigma^2_{1,r,C}x}{n}} - \left(1+\frac{1}{\delta}\right)\frac{b_{1,r,C}x}{n}\right) \le \exp(-x)\ . \tag{221}
$$

Then, notice that all conditions of Lemma 12 are satisfied, and that it gives by (185), for all $n \ge n_0(A_-, A_+, A_u, A_l, A, B_2, r_M(\varphi), \sigma_{\min})$,

$$
M_{1,r,C} \ge \left(1 - \frac{L_{A,A_l,A_u,\sigma_{\min}}}{\sqrt{D}}\right)\sqrt{\frac{rCD}{n}}\mathcal{K}_{1,M}\ . \tag{222}
$$

In addition, observe that

$$
\sigma^2_{1,r,C} \le \sup_{s\in\mathcal{F}_{(C,rC]}} P\left(\psi^2_{1,M}\cdot(s_M - s)^2\right) \le 16A^2rC \quad \text{by (119)} \tag{223}
$$

and

$$
b_{1,r,C} = \sup_{s\in\mathcal{F}_{(C,rC]}} \left\|\psi_{1,M}\cdot(s_M - s)\right\|_\infty \le 4Ar_M(\varphi)\sqrt{rCD} \quad \text{by (119) and (\textbf{H4})} \tag{224}
$$

Hence, using (222), (223) and (224) in inequality (221), we get for all $x > 0$ and for all $n \ge n_0(A_-, A_+, A_u, A_l, A, B_2, r_M(\varphi), \sigma_{\min})$,

$$
\begin{aligned}
&\mathbb{P}\left(S_{1,r,C} \le (1-\delta)\left(1 - \frac{L_{A,A_l,A_u,\sigma_{\min}}}{\sqrt{D}}\right)\sqrt{\frac{rCD}{n}}\mathcal{K}_{1,M} - \sqrt{\frac{32A^2rCx}{n}} - \left(1+\frac{1}{\delta}\right)\frac{4Ar_M(\varphi)\sqrt{rCD}x}{n}\right)\\
&\le \exp(-x)\ .
\end{aligned}
$$

Now, taking $x = \beta\ln n$, $\delta = \frac{\sqrt{\ln n}}{n^{1/4}}$ and using (121), we deduce by simple computations that for all $n \ge n_0(A_-, A_+, A_u, A_l, A, B_2, r_M(\varphi), \sigma_{\min})$,

$$
\mathbb{P}\left(S_{1,r,C} \le \left(1 - L_{A,A_l,A_u,\sigma_{\min},r_M(\varphi),\beta} \times \left(\sqrt{\frac{\ln n}{D}} \vee \frac{\sqrt{\ln n}}{n^{1/4}}\right)\right)\sqrt{\frac{rCD}{n}}\mathcal{K}_{1,M}\right) \le n^{-\beta} \tag{225}
$$

and as

$$
\sqrt{\frac{\ln n}{D}} \vee \frac{\sqrt{\ln n}}{n^{1/4}} \le \sqrt{\frac{\ln n}{D}} \vee \sqrt{\frac{D\ln n}{n}} \le \nu_n
$$

(225) gives, for all $n \ge n_0(A_-, A_+, A_u, A_l, A, B_2, r_M(\varphi), \sigma_{\min})$,

$$
\mathbb{P}\left(S_{1,r,C} \le \left(1 - L_{A,A_l,A_u,\sigma_{\min},r_M(\varphi),\beta} \times \nu_n\right)\sqrt{\frac{rCD}{n}}\mathcal{K}_{1,M}\right) \le n^{-\beta}\ . \tag{226}
$$

Moreover, from Lemma 15 we deduce that, for all $n \ge n_0(A_\infty, A_{cons}, A_+, A_l)$,

$$
\mathbb{P}\left[\sup_{s\in\mathcal{F}_{rC}} |(P_n - P)(\psi_2 \circ (s - s_M))| \ge L_{A_-,A_l,\beta}\sqrt{\frac{rCD}{n}}\tilde{R}_{n,D,\alpha}\right] \le n^{-\beta} \tag{227}
$$

and noticing that

$$\begin{aligned}\tilde{R}_{n,D,\alpha} &= \max\left\{R_{n,D,\alpha}\ ;\ A_\infty\sqrt{\frac{D\ln n}{n}}\right\}\\ &\leq L_{A_\infty,\sigma_{\min}}\max\left\{R_{n,D,\alpha}\ ;\ \sqrt{\frac{D\ln n}{n}}\right\}\times\mathcal{K}_{1,M}\qquad\text{by (121)}\\ &\leq L_{A_\infty,\sigma_{\min}}\times\nu_n\times\mathcal{K}_{1,M}\ ,\end{aligned}$$

we deduce from (227) that for all $n\geq n_0\left(A_\infty,A_{cons},A_+,A_l\right)$,

$$\mathbb{P}\left[\sup_{s\in\mathcal{F}_{rC}}\left|\left(P_n-P\right)\left(\psi_2\circ\left(s-s_M\right)\right)\right|\geq L_{A_-,A_l,A_\infty,\sigma_{\min},\beta}\times\nu_n\times\sqrt{\frac{rCD}{n}}\mathcal{K}_{1,M}\right]\leq n^{-\beta}\ . \tag{228}$$

Finally, using (226) and (228) in (220) we get that,
for all $n\geq n_0\left(A_-,A_+,A_u,A_l,A,A_\infty,A_{cons},B_2,r_M\left(\varphi\right),\sigma_{\min}\right)$,

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

which concludes the proof. ■

7.5 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 [15], Proposition 6.10, p.157.

Theorem 19 *For any independent mean zero random variables Y_j, $j=1,...,n$ taking values in a Banach space $(\mathcal{B},\|.\|)$ and satisfying $\mathbb{E}\left[\|Y_j\|^p\right]<+\infty$ 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 $(X_1,...,X_n)$ be independent variables of common law P taking values in $\mathcal{Z}$. We then denote by $\mathcal{B}=l^\infty\left(\mathcal{F}\right)$ the space of uniformly bounded functions on $\mathcal{F}$ and, for any $b\in\mathcal{B}$, we set $\|b\|=\sup_{f\in\mathcal{F}}|b\left(f\right)|$. Thus $(\mathcal{B},\|.\|)$ is a Banach space. Indeed we shall apply Theorem 19 to the independent random variables, with mean zero and taking values in $\mathcal{B}$, defined by

$$Y_j=\left\{f\left(X_j\right)-Pf,\ f\in\mathcal{F}\right\}\ .$$

More precisely, we will use the following result, which is a straightforward application of Theorem 19. Denote by

$$P_n=\frac{1}{n}\sum_{i=1}^n\delta_{X_i}$$

the empirical measure associated to the sample $(X_1,...,X_n)$ and by

$$\|P_n-P\|_{\mathcal{F}}=\sup_{f\in\mathcal{F}}\left|\left(P_n-P\right)\left(f\right)\right|$$

the supremum of the empirical process over $\mathcal{F}$.

Corollary 20 *If $\mathcal{F}$ is a class of measurable functions from a measurable space $\mathcal{Z}$ to $\mathbb{R}$ satisfying*

$$\sup_{z\in\mathcal{Z}}\sup_{f\in\mathcal{F}}\left|f\left(z\right)-Pf\right|=\sup_{f\in\mathcal{F}}\left\|f-Pf\right\|_{\infty}<+\infty$$

and $(X_1,...,X_n)$ are n i.i.d. random variables taking values in $\mathcal{Z}$, then an absolute constant B_2 exists such that,

$$\mathbb{E}^{1/2}\left[\left\|P_n-P\right\|_{\mathcal{F}}^2\right]\leq B_2\left(\mathbb{E}\left[\left\|P_n-P\right\|_{\mathcal{F}}\right]+\frac{\sup_{f\in\mathcal{F}}\left\|f-Pf\right\|_{\infty}}{n}\right)\ . \tag{229}$$

Another tool we need is a comparison theorem for Rademacher processes, see Theorem 4.12 of [15]. A function $\varphi:\mathbb{R}\to\mathbb{R}$ is called a contraction if $\left|\varphi\left(u\right)-\varphi\left(v\right)\right|\leq\left|u-v\right|$ for all $u,v\in\mathbb{R}$. Moreover, for a subset $T\subset\mathbb{R}^n$ we set

$$\left\|h\left(t\right)\right\|_T=\left\|h\right\|_T=\sup_{t\in T}\left|h\left(t\right)\right|\ .$$

Theorem 21 *Let $(\varepsilon_1,...,\varepsilon_n)$ be n i.i.d. Rademacher variables and $F:\mathbb{R}_+\longrightarrow\mathbb{R}_+$ be a convex and increasing function. Furthermore, let $\varphi_i:\mathbb{R}\longrightarrow\mathbb{R}$, $i\leq n$, be contractions such that $\varphi_i\left(0\right)=0$. Then, for any bounded subset $T\subset\mathbb{R}^n$,*

$$\mathbb{E}F\left(\left\|\sum_i\varepsilon_i\varphi_i\left(t_i\right)\right\|_T\right)\leq2\mathbb{E}F\left(\left\|\sum_i\varepsilon_it_i\right\|_T\right).$$

The next tool is the well known Bernstein's inequality, that can be found for example in [16], Proposition 2.9.

Theorem 22 *(Bernstein's inequality) Let $(X_1,...,X_n)$ be independent real valued random variables and define*

$$S=\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\mathbb{E}\left[X_i\right]\right).$$

Assuming that

$$v=\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[X_i^2\right]<\infty$$

and

$$\left|X_i\right|\leq b\quad a.s.$$

we have, for every $x>0$,

$$\mathbb{P}\left[\left|S\right|\geq\sqrt{2v\frac{x}{n}}+\frac{bx}{3n}\right]\leq2\exp\left(-x\right). \tag{230}$$

We turn now to concentration inequalities for the empirical process around its mean. Bousquet's inequality [8] provides optimal constants for the deviations at the right. Klein-Rio's inequality [12] gives sharp constants for the deviations at the left, that slightly improves Klein's inequality [11].

Theorem 23 *Let $(\xi_1,...,\xi_n)$ 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*

$$\left|f\left(\xi_i\right)-Pf\right|\leq b\quad a.s.,\ \text{for all } f\in\mathcal{F},\ i\leq n,$$

then, by setting

$$\sigma_{\mathcal{F}}^2=\sup_{f\in\mathcal{F}}\left\{P\left(f^2\right)-\left(Pf\right)^2\right\},$$

we have, for all $x\geq0$,

***Bousquet's inequality** :*

$$\mathbb{P}\left[\left\|P_n-P\right\|_{\mathcal{F}}-\mathbb{E}\left[\left\|P_n-P\right\|_{\mathcal{F}}\right]\geq\sqrt{2\left(\sigma_{\mathcal{F}}^2+2b\mathbb{E}\left[\left\|P_n-P\right\|_{\mathcal{F}}\right]\right)\frac{x}{n}}+\frac{bx}{3n}\right]\leq\exp\left(-x\right) \tag{231}$$

and we can deduce that, for all $\varepsilon, x > 0$, it holds

$$\mathbb{P}\left[\|P_n - P\|_{\mathcal{F}} - \mathbb{E}\left[\|P_n - P\|_{\mathcal{F}}\right] \geq \sqrt{2\sigma_{\mathcal{F}}^2 \frac{x}{n}} + \varepsilon \mathbb{E}\left[\|P_n - P\|_{\mathcal{F}}\right] + \left(\frac{1}{\varepsilon} + \frac{1}{3}\right)\frac{bx}{n}\right] \leq \exp\left(-x\right). \tag{232}$$

Klein-Rio's inequality :

$$\mathbb{P}\left[\mathbb{E}\left[\|P_n - P\|_{\mathcal{F}}\right] - \|P_n - P\|_{\mathcal{F}} \geq \sqrt{2\left(\sigma_{\mathcal{F}}^2 + 2b\mathbb{E}\left[\|P_n - P\|_{\mathcal{F}}\right]\right)\frac{x}{n}} + \frac{bx}{n}\right] \leq \exp\left(-x\right) \tag{233}$$

and again, we can deduce that, for all $\varepsilon, x > 0$, it holds

$$\mathbb{P}\left[\mathbb{E}\left[\|P_n - P\|_{\mathcal{F}}\right] - \|P_n - P\|_{\mathcal{F}} \geq \sqrt{2\sigma_{\mathcal{F}}^2 \frac{x}{n}} + \varepsilon \mathbb{E}\left[\|P_n - P\|_{\mathcal{F}}\right] + \left(\frac{1}{\varepsilon} + 1\right)\frac{bx}{n}\right] \leq \exp\left(-x\right). \tag{234}$$

The following result is due to Ledoux [14]. We will use it along the proofs through Corollary 25 which is stated below. From now on, we set for short $Z = \|P_n - P\|_{\mathcal{F}}$.

Theorem 24 *Let $(\xi_1, ..., \xi_n)$ 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 $\left(\xi_1', ..., \xi_n'\right)$ be independent from $(\xi_1, ..., \xi_n)$ and with the same distribution. Setting*

$$v = \mathbb{E}\left[\sup_{f \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\left(f\left(\xi_i\right) - f\left(\xi_i'\right)\right)^2\right]$$

then

$$\mathbb{E}\left[Z^2\right] - \mathbb{E}\left[Z\right]^2 \leq \frac{v}{n}\ .$$

Corollary 25 *Under notations of Theorem 23, if some $\varkappa_n \in (0,1)$ exists such that*

$$\varkappa_n^2 \mathbb{E}\left[Z^2\right] \geq \frac{\sigma^2}{n}$$

and

$$\varkappa_n^2 \sqrt{\mathbb{E}\left[Z^2\right]} \geq \frac{b}{n}$$

then we have, for a numerical constant $A_{1,-}$,

$$\left(1 - \varkappa_n A_{1,-}\right)\sqrt{\mathbb{E}\left[Z^2\right]} \leq \mathbb{E}\left[Z\right]\ .$$

Proof of Corollary 25. Just use Theorem 24, noticing the fact that

$$\sqrt{\mathbb{E}\left[Z^2\right]} - \mathbb{E}\left[Z\right] \leq \sqrt{\mathbb{V}\left(Z\right)}$$

and that, with notations of Theorem 24,

$$v \leq 2\sigma^2 + 32b\mathbb{E}\left[Z\right]\ .$$

The result then follows from straightforward calculations. ■

Acknowledgements

The author thanks gratefully the editor David Ruppert, the associate editor and the two anonymous referees for their comments and suggestions, that greatly improved the paper. I am also grateful to Gilles Celeux for having helped me to edit the English in the text.

References

[1] S. Arlot and F. Bach. Data-driven calibration of linear estimators with minimal penalties. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 46–54, 2009.

[2] S. Arlot and P. Massart. Data-driven calibration of penalties for least-squares regression. *J. Mach. Learn. Res.*, 10:245–279 (electronic), 2009.

[3] Y. Baraud, C. Giraud, and S. Huet. Gaussian model selection with an unknown variance. *Ann. Statist.*, 37(2):630–672, 2009.

[4] P. L. Bartlett and S. Mendelson. Empirical Minimization. *Probab. Theory Related Fields*, 135(3):311–334, 2006.

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

[6] L. Birgé and P. Massart. Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields*, 138(1-2):33–73, 2007.

[7] S. Boucheron and P. Massart. A high-dimensional Wilks phenomenon. *Probab. Theory Related Fields*, 150(3-4):405–433, 2011.

[8] O. 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] M. Crouzeix and A.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.

[10] E. Giné and V. Koltchinskii. Concentration inequalities and asymptotic results for ratio type empirical processes. *Ann.Probab.*, 33:1143–1216, 2006.

[11] T. Klein. Une inégalité de concentration à gauche pour les processus empiriques. *C. R. Math. Acad. Sci. Paris*, 334(6):501–504, 2002.

[12] T. Klein and E. Rio. Concentration around the mean for maxima of empirical processes. *Ann. Probab.*, 33(3):1060–1077, 2005.

[13] V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimisation. *Ann. Statist.*, 34(6):2593–2656, 2006.

[14] M. Ledoux. On Talagrand's deviation inequalities for product measures. *ESAIM Probab. Statist.*, 1:63–87 (electronic), 1995/97.

[15] M. Ledoux and M. Talagrand. *Probability in Banach spaces*. Springer, Berlin, 1991.

[16] P. Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

[17] P. Massart and E. Nédélec. Risks bounds for statistical learning. *Ann.Stat.*, 34(5):2326–2366, 2006.

[18] A. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Ann.Stat.*, 32:135–166, 2004.