Nonparametric Multiple Regression by Projection on Non-compactly Supported Bases

Florian Dussap

Annals of the Institute of Statistical Mathematics manuscript No.
(will be inserted by the editor)

Nonparametric Multiple Regression by Projection on Non-compactly Supported Bases

Florian Dussap

Received: date / Revised: date

Abstract We study the nonparametric regression estimation problem with a random design in $\mathbb{R}^p$ with $p \geq 2$. We do so by using a projection estimator obtained by least squares minimization. Our contribution is to consider non-compact estimation domains in $\mathbb{R}^p$ on which we recover the function, and to provide a theoretical study of the risk of the estimator relative to a norm weighted by the distribution of the design. We propose a model selection procedure in which the model collection is random and takes into account the discrepancy between the empirical norm and the norm associated with the distribution of design. We prove that the resulting estimator automatically optimizes the bias-variance trade-off in both norms, and we illustrate the numerical performance of our procedure on simulated data.

Keywords nonparametric estimation · nonparametric regression · hermite basis · model selection

1 Introduction

We consider the following random design regression model:

$$Y_i = b(\boldsymbol{X}_i) + \varepsilon_i, \quad i = 1, \ldots, n,$$

where the variables $\boldsymbol{X}_i \in \mathbb{R}^p$ are independent but not necessarily identically distributed, the noise variables $\varepsilon_i \in \mathbb{R}$ are i.i.d. centered with finite variance σ^2 and independent from the $\boldsymbol{X}_i$s, and $b\colon \mathbb{R}^p \to \mathbb{R}$ is a regression function. We seek to recover the function b on a domain $A \subset \mathbb{R}^p$ from the observations $(\boldsymbol{X}_i, Y_i)_{i=1,\ldots,n}$.

This work was supported by a grant from Région Île-de-France.

Florian Dussap
MAP5 Laboratory
Université Paris Cité
45 rue des Saints-Pères, 75006 Paris, France
E-mail: florian.dussap@gmail.com

More precisely, we consider the following framework. We assume that the variance of the noise σ^2 is known. We assume that the variables $\boldsymbol{X}_i$ are independent but not identically distributed, we call μ_i the distribution of $\boldsymbol{X}_i$, but we do not assume that μ_i is known. However, we fix ν a reference measure on A and we assume that $\mu := \frac{1}{n}\sum_{i=1}^n \mu_i$ admits a bounded density with respect to ν, so that we have $\mathrm{L}^2(A,\mu) \subset \mathrm{L}^2(A,\nu)$. In particular, this assumption implies that $\mathrm{supp}(\mu) \subset A$. Finally, we consider domains $A \subset \mathbb{R}^p$ of the form $A_1 \times \cdots \times A_p$ where $A_k \subset \mathbb{R}$ and we consider a measure ν on A that is of the form $\nu_1 \otimes \cdots \otimes \nu_p$ with ν_k supported on A_k. Our goal is to estimate the regression function b on the domain A and to control the expected error with respect to the norm $\|\cdot\|_\mu$ associated with the distribution of the $\boldsymbol{X}_i$s:

$$\forall t \in \mathrm{L}^2(A,\mu), \quad \|t\|_\mu^2 := \int_A t(\boldsymbol{x})^2 \,\mathrm{d}\mu(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^n \int_A t(\boldsymbol{x})^2 \,\mathrm{d}\mu_i(\boldsymbol{x}).$$

We can interpret the error with respect to this norm as a prediction risk: if $\boldsymbol{X}'_1, \ldots, \boldsymbol{X}'_n$ are independent copies of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, then we have:

$$\forall \hat{b} \text{ estimator}, \quad \|b - \hat{b}\|_\mu^2 = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\Big[\big(b(\boldsymbol{X}'_i) - \hat{b}(\boldsymbol{X}'_i)\big)^2 \,\Big|\, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\Big],$$

which is the mean quadratic error of a new observation drawn uniformly from one of the distributions μ_i.

Nonparametric regression problems have a long history, and a large number of methods have been proposed. In this introduction, we focus on two main families of methods: kernel estimators and projection estimators. For reference books on the subject, see Efromovich (1999) regarding the projection method and Györfi et al (2002) for the kernel method.

The classical estimator of Nadaraya (1964) and Watson (1964) consists of a quotient of estimators $\widehat{bf}/\hat{f}$, where $\widehat{bf}$ and $\hat{f}$ are kernel estimators of the functions bf and f (the function f being the common density of the $\boldsymbol{X}_i$s in the i.i.d case). This estimator can also be interpreted as locally fitting a constant by averaging the Y_is, the locality being determined by the kernel, see the book of Györfi et al (2002) or Tsybakov (2009). This method can then be generalized by replacing the local constant by a local polynomial, leading to the so-called *local polynomial estimator*.

The main drawback of the Nadaraya–Watson estimator is that it relies on an estimator of the density of the $\boldsymbol{X}_i$s. As such, the rate of convergence depends on the regularity of f, and two smoothing parameters have to be chosen. A popular solution is to choose the same bandwidth for both estimators using leave-one-out cross validation. This method works well in practice and has been proven consistent by Härdle and Marron (1985) (see also Chapter 8 in Györfi et al (2002)). Recently, Comte and Marie (2021) have proposed to use the Penalized Comparison to Overfitting method (PCO), a bandwidth selection method developed by Lacour et al (2017) for kernel density estimation, to select separately the bandwidths of the numerator and the denominator of the Nadaraya–Watson estimator. Their estimator matches the performances of the single bandwidth CV estimator when the noise is high, but the latter is better when the noise is small. Other bandwidth selection methods exist such as

plug-in or bootstrap; see Köhler et al (2014) for an extensive survey and comparison of the different bandwidth selection methods for the local linear estimator.

Another approach is to use a projection estimator. The idea is to minimize a least squares contrast over finite-dimensional spaces of functions $\{S_{\boldsymbol{m}} : \boldsymbol{m} \in \mathscr{M}_n\}$ called *models*:

$$\hat{b}_{\boldsymbol{m}} \coloneqq \operatorname*{arg\,min}_{t \in S_{\boldsymbol{m}}} \frac{1}{n} \sum_{i=1}^{n} \big(Y_i - t(\boldsymbol{X}_i)\big)^2,$$

the model collection $\mathscr{M}_n$ being allowed to depend on the number of observations. This method overcomes the problems of the Nadaraya–Watson estimator: it does not need to estimate the density of the $\boldsymbol{X}_i$s, and only one model selection procedure is required. Moreover, it can provide a sparse representation of the estimator. This approach was developed in a fixed design setting by Birgé and Massart (1998), Barron et al (1999) and Baraud (2000). In particular, the papers of Baraud (2000, 2002) provide a model selection procedure that optimizes the bias-variance compromise under weak assumptions on the moments of the noise distribution. They obtain an estimator that is adaptive both in the fixed and random design setting when the domain A is compact.

The non-compact case have been studied recently in the simple regression setting ($p = 1$) by Comte and Genon-Catalot (2020a,b). They use non-compactly supported bases, specifically the Hermite basis (supported on $\mathbb{R}$) and the Laguerre basis (supported on $\mathbb{R}_+$), to construct their estimator. Significant attention has been paid to these bases in the past years since they exhibit nice mathematical properties that are useful for solving inverse problems (Mabon 2017; Comte and Genon-Catalot 2018; Sacko 2020). Non-compactly supported bases also avoid issues concerning the choice of support. When A is compact, the theory assumes it is fixed *a priori*. In practice, however, the support is generally determined using the data, although this dependency between data and support is not taken into account in the theoretical development. Working with a non-compact domain, for example $\mathbb{R}$ or $\mathbb{R}_+$, allows us to bypass this issue.

Concerning the regression problem, difficulties arise when we go from the compact case to the non-compact case. When A is compact, it is usual to assume that the density of the $\boldsymbol{X}_i$s is bounded from below by some positive constant f_0. In the non-compact case, this assumption fails. Instead, the study of the minimum eigenvalue of some random matrix must be done. This question has been studied in the simple regression case ($p = 1$) by Cohen et al (2013) by using the matrix concentration inequalities of Tropp (2012). However, their results are obtained under the assumption that the regression function is bounded by a known quantity and they do not provide a model selection procedure.

We make the following contributions in our paper. We extend the results of Comte and Genon-Catalot (2020a) to the multiple regression case ($p \geq 2$) with more general assumptions on the design, and we improve their result on the oracle inequality under the empirical norm (see Theorem 2). Our work generalizes the results of Baraud (2002) to the non-compact case and improves their results in the compact case (see Theorem 3). We do so by combining the fixed design results of Baraud (2000) with a more refined study of the discrepancy between the empirical norm and the μ-norm.

This discrepancy is expressed in terms of the deviation of the minimum eigenvalue of a random matrix, of which we control the probability with the concentration inequalities of Tropp (2012) and Gittens and Tropp (2011). Finally, our estimator is constructed as a projection estimator on a tensorized basis whose coefficients are computed using hypermatrix calculus and can be implemented in practice. This feasibility is illustrated in Section 5 which also shows that the procedure works well.

Outline of the paper In Section 2 we define the projection estimator. In Section 3 we study the probability that the empirical norm and the μ-norm depart from each other and we derive an upper bound on the μ-risk of our estimator. In Section 4 we propose a model selection procedure and we prove that it satisfies an oracle inequality both in empirical norm and in μ-norm. Finally, in Section 5 we study numerically the performance of our estimator. All the proofs are gathered in Section 7.

Notations

- $\mathbb{E}_{\boldsymbol{X}} \coloneqq \mathbb{E}[\cdot\,|\,\boldsymbol{X}_1,\dots,\boldsymbol{X}_n]$, $\mathbb{P}_{\boldsymbol{X}} \coloneqq \mathbb{P}[\cdot\,|\,\boldsymbol{X}_1,\dots,\boldsymbol{X}_n]$, $\mathrm{Var}_{\boldsymbol{X}} \coloneqq \mathrm{Var}(\cdot\,|\,\boldsymbol{X}_1,\dots,\boldsymbol{X}_n)$, where $\boldsymbol{X} = (\boldsymbol{X}_1,\dots,\boldsymbol{X}_n)$.
- If π is a measure on A, we write $\|\cdot\|_\pi$ and $\langle\cdot,\cdot\rangle_\pi$ the norm and the inner product weighted by the measure π.
- We denote by $\langle\cdot,\cdot\rangle_n$ and $\|\cdot\|_n$ the empirical inner product and the empirical norm[1], defined as $\langle t,s\rangle_n \coloneqq \frac{1}{n}\sum_{i=1}^n t(\boldsymbol{X}_i)s(\boldsymbol{X}_i)$ and $\|t\|_n^2 \coloneqq \frac{1}{n}\sum_{i=1}^n t(\boldsymbol{X}_i)^2$. If $\mathbf{u}\in\mathbb{R}^n$ is a vector, we also write $\|\mathbf{u}\|_n^2 \coloneqq \frac{1}{n}\sum_{i=1}^n u_i^2$.

2 Projection estimator

In our setting, the domain is a Cartesian product $A = A_1\times\cdots\times A_p$ and $\nu = \nu_1\otimes\cdots\otimes\nu_p$ where ν_k is supported on A_k. For each $i\in\{1,\dots,p\}$, we consider $(\varphi_j^i)_{j\in\mathbb{N}}$ an orthonormal basis of $\mathrm{L}^2(A_i,\mathrm{d}\nu_i)$ and we form an orthonormal basis of $\mathrm{L}^2(A,\mathrm{d}\nu)$ by tensorization:

$$\forall \boldsymbol{j}\in\mathbb{N}^p,\quad \forall \boldsymbol{x}\in A,\quad \boldsymbol{\varphi_j}(\boldsymbol{x}) \coloneqq (\varphi_{j_1}^1\otimes\cdots\otimes\varphi_{j_p}^p)(\boldsymbol{x}) \coloneqq \varphi_{j_1}^1(x_1)\times\cdots\times\varphi_{j_p}^p(x_p).$$

For $\boldsymbol{m}\in\mathbb{N}_+^p$, we set $S_{\boldsymbol{m}} \coloneqq \mathrm{Span}(\boldsymbol{\varphi_j} : \boldsymbol{j}\le \boldsymbol{m}-\mathbf{1})$ and we write $D_{\boldsymbol{m}} \coloneqq m_1\cdots m_p$ its dimension. We estimate b by minimizing a least squares contrast on $S_{\boldsymbol{m}}$:

$$\hat{b}_{\boldsymbol{m}} \coloneqq \operatorname*{arg\,min}_{t\in S_{\boldsymbol{m}}} \frac{1}{n}\sum_{i=1}^n \big(Y_i - t(\boldsymbol{X}_i)\big)^2.$$

If we expand $\hat{b}_{\boldsymbol{m}}$ on the basis $(\boldsymbol{\varphi_j})_{\boldsymbol{j}\in\mathbb{N}^p}$, this problem can be written as:

$$\hat{b}_{\boldsymbol{m}} = \sum_{\boldsymbol{j}\le\boldsymbol{m}-\mathbf{1}} \hat{a}_{\boldsymbol{j}}^{(\boldsymbol{m})}\,\boldsymbol{\varphi_j},\qquad \hat{\mathbf{a}}^{(\boldsymbol{m})} \coloneqq \operatorname*{arg\,min}_{\boldsymbol{a}\in\mathbb{R}^{\mathbf{m}}} \|\mathbf{Y} - \widehat{\boldsymbol{\Phi}}_{\boldsymbol{m}}\times_p \mathbf{a}\|_{\mathbb{R}^n}^2, \tag{1}$$

[1] in general it is a semi-norm but we will only consider subspaces on which it is a norm.

where $\mathbf{Y} \coloneqq (Y_1, \ldots, Y_n) \in \mathbb{R}^n$ and $\widehat{\boldsymbol{\Phi}}_{\boldsymbol{m}} \in \mathbb{R}^{n \times \boldsymbol{m}}$ is defined as:

$$\forall i \in \{1, \ldots, n\}, \quad \forall \boldsymbol{j} \leq \boldsymbol{m} - \mathbf{1}, \quad \left[\widehat{\boldsymbol{\Phi}}_{\boldsymbol{m}}\right]_{i,\boldsymbol{j}} \coloneqq \varphi_{\boldsymbol{j}}(\boldsymbol{X}_i).$$

Using Lemma 8 in Appendix, the problem (1) has a unique solution if and only if $\widehat{\boldsymbol{\Phi}}_{\boldsymbol{m}}$ is injective and in that case:

$$\begin{aligned} \hat{\mathbf{a}}^{(\boldsymbol{m})} &= (\widehat{\boldsymbol{\Phi}}_{\boldsymbol{m}}^* \times_1 \widehat{\boldsymbol{\Phi}}_{\boldsymbol{m}})^{-1} \times_p \widehat{\boldsymbol{\Phi}}_{\boldsymbol{m}}^* \times_1 \mathbf{Y} \\ &= \frac{1}{n} \widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1} \times_p \widehat{\boldsymbol{\Phi}}_{\boldsymbol{m}}^* \times_1 \mathbf{Y}, \end{aligned}$$

where $[\widehat{\boldsymbol{\Phi}}_{\boldsymbol{m}}^*]_{\boldsymbol{j},i} = [\widehat{\boldsymbol{\Phi}}_{\boldsymbol{m}}]_{i,\boldsymbol{j}}$ and where $\widehat{\mathbf{G}}_{\boldsymbol{m}}$ is the Gram hypermatrix of $(\varphi_{\boldsymbol{j}})_{\boldsymbol{j} \leq \boldsymbol{m}-\mathbf{1}}$ relatively to the empirical inner product $\langle \cdot, \cdot \rangle_n$:

$$\forall \boldsymbol{j}, \boldsymbol{k} \leq \boldsymbol{m} - \mathbf{1}, \quad \left[\widehat{\mathbf{G}}_{\boldsymbol{m}}\right]_{\boldsymbol{j},\boldsymbol{k}} \coloneqq \langle \varphi_{\boldsymbol{j}}, \varphi_{\boldsymbol{k}} \rangle_n.$$

Notice that $\widehat{\boldsymbol{\Phi}}_{\boldsymbol{m}}$ is injective if and only if $\widehat{\mathbf{G}}_{\boldsymbol{m}}$ is invertible, that is if and only if $\|\cdot\|_n$ is a norm on $S_{\boldsymbol{m}}$.

3 Bound on the risk of the estimator

Let us start with the classical bias-variance decomposition of the empirical risk. In our context this result is given by the next Proposition.

Proposition 1 *If $\widehat{\mathbf{G}}_{\boldsymbol{m}}$ is invertible, then we have:*

$$\mathbb{E}_{\boldsymbol{X}} \|b - \hat{b}_{\boldsymbol{m}}\|_n^2 = \inf_{t \in S_{\boldsymbol{m}}} \|b - t\|_n^2 + \sigma^2 \frac{D_{\boldsymbol{m}}}{n}.$$

As a consequence, if $\widehat{\mathbf{G}}_{\boldsymbol{m}}$ is invertible a.s, then we have:

$$\mathbb{E} \|b - \hat{b}_{\boldsymbol{m}}\|_n^2 \leq \inf_{t \in S_{\boldsymbol{m}}} \|b - t\|_\mu^2 + \sigma^2 \frac{D_{\boldsymbol{m}}}{n}.$$

Hereafter, we always assume that $\hat{\mathbf{G}}_{\boldsymbol{m}}$ is invertible a.s.

If we want to obtain a similar result for the μ-norm, we need to understand how the empirical norm can deviate from the μ-norm. More generally, we need to understand the relations between the different norms we have on the subspace $S_{\boldsymbol{m}}$ ($\|\cdot\|_n$, $\|\cdot\|_\mu$, $\|\cdot\|_\nu$ and $\|\cdot\|_\infty$). It is well known that all norms are equivalent on finite dimensional spaces; our question concerns the constants in this equivalence. We introduce the following notation: if $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$ are two norms on a space S, we define:

$$K_\beta^\alpha(S) \coloneqq \sup_{t \in S \setminus \{0\}} \frac{\|t\|_\alpha^2}{\|t\|_\beta^2},$$

and when $S = S_{\boldsymbol{m}}$, we use the notation $K_\beta^\alpha(\boldsymbol{m}) \coloneqq K_\beta^\alpha(S_{\boldsymbol{m}})$. The next lemma gives the value of $K_\alpha^\beta(S)$ when the norms are Euclidean.

Lemma 1 *Let $(S, \langle\cdot,\cdot\rangle_\alpha)$ be a d-dimensional Euclidean vector space equipped with an orthonormal basis $(\phi_1,\dots,\phi_d)$. Let $\langle\cdot,\cdot\rangle_\beta$ be another inner product on E and let* $\mathbf{G}$ *be the Gram matrix of the basis $(\phi_1,\dots,\phi_d)$ relatively to $\langle\cdot,\cdot\rangle_\beta$, that is:*

$$\mathbf{G} \coloneqq \left[\langle \phi_j, \phi_k\rangle_\beta\right]_{1\le j,k\le d}.$$

We have:

$$K_\alpha^\beta(S) = \|\mathbf{G}\|_{\mathrm{op}} = \lambda_{\max}(\mathbf{G}), \qquad K_\beta^\alpha(S) = \|\mathbf{G}^{-1}\|_{\mathrm{op}} = \frac{1}{\lambda_{\min}(\mathbf{G})}.$$

The proof of Lemma 1 is identical to the proof of Lemma 3.1 in Baraud (2000), so we leave it out.

The next lemma provides a way to compute $K_\alpha^\infty(S)$ from an orthonormal basis when $\|\cdot\|_\alpha$ is Euclidean. It is essentially the same as Lemma 1 in Birgé and Massart (1998).

Lemma 2 *Let S be a space of bounded functions on A such that $d \coloneqq \dim(S)$ is finite. Let $\langle\cdot,\cdot\rangle_\alpha$ be an inner product on S. If $(\psi_1,\dots,\psi_d)$ is an orthonormal basis of S, then we have:*

$$K_\alpha^\infty(S) = \left\| \sum_{j=1}^d \psi_j^2 \right\|_\infty.$$

The question we are interested in is how close are the norms $\|\cdot\|_n$ and $\|\cdot\|_\mu$ on $S_{\boldsymbol{m}}$. Following a similar idea of Cohen et al (2013), let us define the event:

$$\forall \delta \in (0,1), \quad \Omega_{\boldsymbol{m}}(\delta) \coloneqq \left\{ \forall t \in S_{\boldsymbol{m}},\, \|t\|_\mu^2 \le \frac{1}{1-\delta}\|t\|_n^2 \right\} = \left\{ K_n^\mu(\boldsymbol{m}) \le \frac{1}{1-\delta} \right\}. \quad (2)$$

The key decomposition of the μ-risk of $\hat{b}_{\boldsymbol{m}}$ is given by the following Proposition.

Proposition 2 *For all $\delta \in (0,1)$, we have:*

$$\begin{aligned}\mathbb{E}\|b-\hat{b}_{\boldsymbol{m}}\|_\mu^2 \le \left(1+\frac{2}{1-\delta}\left[\frac{K_\mu^\infty(\boldsymbol{m})}{(1-\delta)n}\wedge 1\right]\right)\inf_{t\in S_{\boldsymbol{m}}}\|b-t\|_\mu^2 + \frac{2\sigma^2 D_{\boldsymbol{m}}}{(1-\delta)n} \\ + 2\|b\|_\mu^2\,\mathbb{P}[\Omega_{\boldsymbol{m}}(\delta)^c] + \mathbb{E}\left[K_n^\mu(\boldsymbol{m})\|\mathbf{Y}\|_n^2\mathbf{1}_{\Omega_{\boldsymbol{m}}(\delta)^c}\right],\end{aligned}$$

where $K_n^\mu(\boldsymbol{m})$ and $K_\mu^\infty(\boldsymbol{m})$ are given by Lemmas 1 and 2.

We see that we need an upper bound on the probability of the event $\Omega_{\boldsymbol{m}}(\delta)^c$. The following proposition is a consequence of the matrix Chernoff bound of Tropp (2012) (Theorem 5 in Appendix) .

Proposition 3 *For all $\delta \in (0,1)$, we have:*

$$\mathbb{P}[\Omega_{\boldsymbol{m}}(\delta)^c] \le D_{\boldsymbol{m}} \exp\left(-h(\delta)\frac{n}{K_\mu^\infty(\boldsymbol{m})}\right),$$

where $h(\delta) \coloneqq \delta + (1-\delta)\log(1-\delta)$ and $K_\mu^\infty(\boldsymbol{m})$ is given by Lemma 2.

Remark 1 The quantity $K_\mu^\infty(\boldsymbol{m})$ is unknown but we have the following upper bound using Lemmas 1 and 2:

$$K_\mu^\infty(\boldsymbol{m}) \leq K_\nu^\infty(\boldsymbol{m}) K_\mu^\nu(\boldsymbol{m}) = \left(\sup_{\boldsymbol{x}\in A} \sum_{\boldsymbol{j}\leq \boldsymbol{m}-\mathbf{1}} \varphi_{\boldsymbol{j}}(\boldsymbol{x})^2 \right) \|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}.$$

The quantity $\|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}$ is still unkown but can be estimated by plugging in $\widehat{\mathbf{G}}_{\boldsymbol{m}}$.

Comte and Genon-Catalot (2020a) show in their Proposition 8 that, when one uses the Hermite or the Laguerre basis, the inverse of the Gram matrix is unbounded (it satisfies $\|\mathbf{G}_m^{-1}\|_{\mathrm{op}} \gtrsim \sqrt{m}$), while it is bounded in the compact case:

$$\|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}} = \sup_{t\in S_{\boldsymbol{m}}\setminus\{0\}} \frac{\|t\|_\nu^2}{\|t\|_\mu^2} \leq \frac{1}{f_0}, \tag{3}$$

where f_0 is a positive lower bound of the covariates density. Hence, the least squares minimization problem will become highly unstable as the dimension of the projection space grows. That is why a form of regularization is needed if we want to control the μ-risk of the estimator. For α a positive constant, let us consider the following model collection:

$$\mathscr{M}_{n,\alpha}^{(1)} := \left\{ \boldsymbol{m} \in \mathbb{N}_+^p \;\middle|\; K_\nu^\infty(\boldsymbol{m}) \left(\|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}} \vee 1 \right) \leq \alpha \frac{n}{\log n} \right\}. \tag{4}$$

Gathering Propositions 2 and 3, we obtain the following bound on the μ-risk of $\hat{b}_{\boldsymbol{m}}$ when $\boldsymbol{m}$ belongs to $\mathscr{M}_{n,\alpha}^{(1)}$.

Theorem 1 *Let us assume that $b \in \mathrm{L}^{2r}(\mu)$ for some $r \in (1,+\infty]$ and let $r' \in [1,+\infty)$ be the conjugated index of r, that is: $\frac{1}{r} + \frac{1}{r'} = 1$. For all $\alpha \in (0, \frac{1}{2r'+1})$ and for all $\boldsymbol{m} \in \mathscr{M}_{n,\alpha}^{(1)}$ we have:*

$$\mathbb{E}\|b - \hat{b}_{\boldsymbol{m}}\|_\mu^2 \leq C_n(\alpha,r') \inf_{t\in S_{\boldsymbol{m}}} \|b-t\|_\mu^2 + C'(\alpha,r')\,\sigma^2 \frac{D_{\boldsymbol{m}}}{n} + \frac{C''\left(b,\sigma^2,\alpha,r\right)}{n\log n},$$

where the constants $C_n(\alpha,r')$ and $C'(\alpha,r')$ are given by:

$$C_n(\alpha,r') := 1 + \frac{2}{1-\delta(\alpha,r')} \left(\frac{\alpha}{\left(1-\delta(\alpha,r')\right)\log n} \wedge 1 \right), \quad C'(\alpha,r') := \frac{2}{1-\delta(\alpha,r')},$$

where $\delta(\alpha,r') \in (0,1)$ tends to 1 as α tends to $\frac{1}{2r'+1}$, and where $C''\left(b,\sigma^2,\alpha,r\right)$ is defined by (18).

Remark 2 Let us make some statements concerning the behavior of $C_n(\alpha,r')$ and $C'(\alpha,r')$:

- $C_n(\alpha,r')$ is bounded relatively to n;
- $C_n(\alpha,r') \geq 1$ and $C'(\alpha,r') \geq 2$;
- as $\alpha \to \frac{1}{2r'+1}$ with n fixed, $C_n(\alpha,r')$ and $C'(\alpha,r')$ tend to $+\infty$;
- as $n \to +\infty$ with α and r' fixed, $C_n(\alpha,r')$ tends to 1.

4 Adaptive estimator

We consider the empirical version of the model collection $\mathscr{M}_{n,\alpha}$ defined by (4):

$$\widehat{\mathscr{M}}_{n,\beta}^{(1)} \coloneqq \left\{ \boldsymbol{m} \in \mathbb{N}_+^p \;\middle|\; K_V^\infty(\boldsymbol{m}) \big(\|\widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}} \vee 1 \big) \leq \beta \frac{n}{\log n} \right\},$$

with β a positive constant. We choose $\hat{\boldsymbol{m}}_1 \in \widehat{\mathscr{M}}_{n,\beta}^{(1)}$ by minimizing the following penalized least squares criterion:

$$\hat{\boldsymbol{m}}_1 \coloneqq \operatorname*{arg\,min}_{\boldsymbol{m} \in \widehat{\mathscr{M}}_{n,\beta}^{(1)}} \left(-\|\hat{b}_{\boldsymbol{m}}\|_n^2 + (1+\theta)\sigma^2 \frac{D_{\boldsymbol{m}}}{n} \right), \qquad \theta > 0. \tag{5}$$

Based on a result of Baraud (2000) for fixed design regression, we prove that $\hat{b}_{\hat{\boldsymbol{m}}_1}$ automatically optimizes the bias-variance compromise in empirical norm on $\mathscr{M}_{n,\alpha}$, up to a constant and a remainder term.

Theorem 2 *If $b \in \mathrm{L}^{2r}(\mu)$ for some $r \in (1,+\infty]$ and if $\mathbb{E}|\varepsilon_1|^q$ is finite for some $q > 6$, then there exists a constant $\alpha_{\beta,r'} > 0$ depending on β and r' (the conjugated index of r) such that for all $\alpha \in (0, \alpha_{\beta,r'})$, the following upper bound on the risk of the estimator $\hat{b}_{\hat{\boldsymbol{m}}_1}$ with $\hat{\boldsymbol{m}}_1$ defined by* (5) *holds:*

$$\mathbb{E}\|b - \hat{b}_{\hat{\boldsymbol{m}}_1}\|_n^2 \leq C(\theta) \inf_{\boldsymbol{m} \in \mathscr{M}_{n,\alpha}^{(1)}} \left(\inf_{t \in S_{\boldsymbol{m}}} \|b-t\|_\mu^2 + \sigma^2 \frac{D_{\boldsymbol{m}}}{n} \right) + \sigma^2 \frac{\Sigma(\theta,q)}{n} + R_n,$$

where $C(\theta) \coloneqq (2 + 8\theta^{-1})(1+\theta)$, and where:

$$\Sigma(\theta,q) \coloneqq C''(\theta,q) \frac{\mathbb{E}|\varepsilon_1|^q}{\sigma^q} \sum_{\boldsymbol{m} \in \mathbb{N}_+^p} D_{\boldsymbol{m}}^{-(\frac{q}{2}-2)}, \quad R_n \coloneqq C'(\|b\|_{\mathrm{L}^{2r}(\mu)}, \sigma^2) \frac{(\log n)^{(p-1)/r'}}{n^{\kappa(\alpha,\beta)/r'}},$$

with $\kappa(\alpha,\beta)$ a positive constant satisfying $\frac{\kappa(\alpha,\beta)}{r'} > 1$ and $\frac{\kappa(\alpha,\beta)}{r'} \to 1$ as $\alpha \to \alpha_{\beta,r'}$.

Remark 3 The term $\Sigma(\theta,q)$ is finite if $q > 6$. Indeed, let $2\varepsilon \coloneqq (\frac{q}{2}-2) - 1 > 0$, we have:

$$\sum_{\boldsymbol{m} \in \mathbb{N}_+^p} D_{\boldsymbol{m}}^{-(\frac{q}{2}-2)} = \sum_{d=1}^{+\infty} \mathrm{Card}\big\{ \boldsymbol{m} \in \mathbb{N}_+^p \,\big|\, D_{\boldsymbol{m}} = d \big\} \times d^{-(\frac{q}{2}-2)} \leq \sum_{d=1}^{+\infty} \frac{\mathrm{o}(d^\varepsilon)}{d^{1+2\varepsilon}} < +\infty,$$

where we use Theorem 7 in Appendix.

Remark 4 The constant $\alpha_{\beta,r'}$ is increasing with β and goes from 0 to $\frac{1}{2r'+1}$. It is also decreasing with r' (so increasing with r) and tends to 0 as $r' \to +\infty$ (as $r \to 1$).

To transfer the previous adaptive result from the empirical norm into the μ-norm, we use once again concentration inequalities on the matrix $\widehat{\mathbf{G}}_{\boldsymbol{m}}$. However, we need to make a distinction between the compact case and the non-compact case. Indeed, when A is compact, we can make the usual assumption that the density $\frac{\mathrm{d}\mu}{\mathrm{d}\nu}$ is bounded from below and apply the matrix Chernoff bound of Gittens and Tropp (2011), see

Lemma 6. This lemma relies critically on the "bounded from below" assumption so it cannot work in the non-compact case.

To handle the non-compact case, we make use of the matrix Bernstein bound of Tropp (2012) instead (Theorem 6 in appendix), see Lemma 7. This inequality is different from the matrix Chernoff bounds we have used so far, so we have to consider smaller model collections to make it work. In the following, we consider two cases:

1. *Compact case*. We assume that there exists $f_0 > 0$ such that for all $x \in A$, $\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(x) > f_0$. In that case, $\mathbf{G}_{\boldsymbol{m}}$ is always invertible and we have $\|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}} \leq \frac{1}{f_0}$, see (3).
2. *General case*. We consider smaller model collections:

$$\mathscr{M}_{n,\alpha}^{(2)} \coloneqq \left\{ \boldsymbol{m} \in \mathbb{N}_+^p \;\middle|\; K_\nu^\infty(\boldsymbol{m}) \left(\|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}^2 \vee 1 \right) \leq \alpha \frac{n}{\log n} \right\},$$

$$\widehat{\mathscr{M}}_{n,\beta}^{(2)} \coloneqq \left\{ \boldsymbol{m} \in \mathbb{N}_+^p \;\middle|\; K_\nu^\infty(\boldsymbol{m}) \left(\|\widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}^2 \vee 1 \right) \leq \beta \frac{n}{\log n} \right\},$$

where α and β are positive constants and we choose $\hat{\boldsymbol{m}}_2 \in \widehat{\mathscr{M}}_{n,\beta}^{(2)}$ as:

$$\hat{\boldsymbol{m}}_2 \coloneqq \operatorname*{arg\,min}_{\boldsymbol{m} \in \widehat{\mathscr{M}}_{n,\beta}^{(2)}} \left(-\|\hat{b}_{\boldsymbol{m}}\|_n^2 + (1+\theta)\sigma^2 \frac{D_{\boldsymbol{m}}}{n} \right), \quad \theta > 0. \tag{6}$$

Theorem 3 *Let $r \in (1, +\infty]$, let $r' \in [1, +\infty)$ be its conjugated index and let us assume that b belongs to $\mathrm{L}^{2r}(\mu)$ and that $\mathbb{E}|\varepsilon_1|^q$ is finite for some $q > 6$.*

• ***Compact case***. *Let $f_0 > 0$ such that $\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(x) \geq f_0$ for all $x \in A$, there exists $\beta_{f_0,r'} > 0$ such that for all $\beta \in (0, \beta_{f_0,r'})$, there exists $\alpha_{\beta,r'} > 0$ such that for all $\alpha \in (0, \alpha_{\beta,r'})$, the following upper bound on the risk of the estimator $\hat{b}_{\hat{\boldsymbol{m}}_1}$ with $\hat{\boldsymbol{m}}_1$ defined by* (5) *holds:*

$$\mathbb{E}\|b - \hat{b}_{\hat{\boldsymbol{m}}_1}\|_\mu^2 \leq C(\theta,\beta,r) \inf_{\boldsymbol{m} \in \mathscr{M}_{n,\alpha}^{(1)}} \left(\inf_{t \in S_{\boldsymbol{m}}} \|b - t\|_\mu^2 + \sigma^2 \frac{D_{\boldsymbol{m}}}{n} \right) + C'(\beta,r)\sigma^2 \frac{\Sigma(\theta,q)}{n} + R_n,$$

where the remainder term is given by:

$$R_n = C''\left(\|b\|_{\mathrm{L}^{2r}(\mu)}, \sigma^2, \beta, r\right) \left(n^{-\frac{\kappa(\alpha,\beta)}{r'}} (\log n)^{\frac{p-1}{r'}} + n^{-\lambda(\beta,r,f_0)} (\log n)^{\frac{p-1}{r'}-1} \right),$$

with $\lambda(\beta,r,f_0) > 1$ and $\frac{\kappa(\alpha,\beta)}{r'} > 1$.

• ***General case***. *Let $B \coloneqq (\|\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\|_\infty + \frac{2}{3})^{-1}$, there exists $\beta_{B,r'} > 0$ such that for all $\beta \in (0, \beta_{B,r'})$, there exists $\tilde{\alpha}_{\beta,r'} > 0$ such that for all $\alpha \in (0, \tilde{\alpha}_{\beta,r'})$, the following upper bound on the risk of the estimator $\hat{b}_{\hat{\boldsymbol{m}}_2}$ with $\hat{\boldsymbol{m}}_2$ defined by* (6) *holds:*

$$\mathbb{E}\|b - \hat{b}_{\hat{\boldsymbol{m}}_2}\|_\mu^2 \leq C(\theta,\beta,r) \inf_{\boldsymbol{m} \in \mathscr{M}_{n,\alpha}^{(2)}} \left(\inf_{t \in S_{\boldsymbol{m}}} \|b - t\|_\mu^2 + \sigma^2 \frac{D_{\boldsymbol{m}}}{n} \right) + C'(\beta,r)\sigma^2 \frac{\Sigma(\theta,q)}{n} + R_n,$$

where the remainder term is given by:

$$R_n = C''\left(\|b\|_{\mathrm{L}^{2r}(\mu)}, \sigma^2, \beta, r\right)\left(n^{-\frac{\tilde{\kappa}(\alpha,\beta)}{r'}}(\log n)^{\frac{p-1}{r'}} + n^{-\lambda(\beta,r,B)}(\log n)^{\frac{p-1}{r'}-1}\right),$$

with $\lambda(\beta,r,B) > 1$ *and* $\frac{\tilde{\kappa}(\alpha,\beta)}{r'} > 1$.

This result shows that there is a range of values for the constant β that depends on the integrability of b and on f_0 (compact case) or $\|\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\|_\infty$ (general case), such that for the μ-norm, the estimator $\hat{b}_{\hat{\boldsymbol{m}}}$ automatically optimizes the bias-variance trade-off (up to a constant and a rest) on $\mathscr{M}_{n,\alpha}$ for all α in a range that depends on β.

Remark 5 Theorem 3 improves previous results in the literature:

1. In the compact case, we improve the result of Baraud (2002). Indeed in this article, the model collections considered are built by picking an "envelope model", that is a linear space $\mathscr{S}_n$ with finite dimension N_n, whose all models are a subspace. Their assumptions concern the space $\mathscr{S}_n$: they assume that $K_\nu^\infty(\mathscr{S}_n) \leq C^2 N_n$ for some constant $C > 0$ and they require that $N_n \leq C^{-1}\sqrt{n/(\log n)^3}$. In comparison, our procedure avoids the choice *a priori* of an envelope model, and uses a looser constraint on the dimension of the models.
2. In the non-compact case, we extend the results of Comte and Genon-Catalot (2020a) to the case $p \geq 2$ without losing much on the assumptions: their result requires a moment of order 6 on the noise whereas our result is obtained with a moment of order q, with $q > 6$. We also generalize their result by considering a non i.i.d. design and by using a more general moment assumption on the regression function.

Remark 6 (Unknown variance) During all of our work, we assume that σ^2 is known. To handle the case of an unknown variance, we can use the same method proposed by Baraud (2000) in the fixed design setting. Using a residual least-squares estimator of σ^2 in the penalized criterion for choosing the model, they prove (Theorem 6.1) that the resulting estimator of the regression function satisfies an oracle inequality. Starting from Baraud's result, and using the same arguments we used in this paper, we think one can obtain an oracle inequality for a projection estimator, in the random design framework with unknown variance. We omit such development for the sake of conciseness.

5 Numerical illustrations

In this section, we compare our estimator with the Nadaraya–Watson estimator on simulated data in the case $p = 1$ and $p = 2$.

Regression function We consider the following regression functions:

1. $b_1(x) = \exp((x-1)^2) + \exp((x+1)^2)$,
2. $b_2(x) := \frac{1}{1+x^2}$,

3. $b_3(x) := x\cos(x)$,
4. $b_4(x) := |x|$,
5. $b_5(x_1,x_2) := \exp(-\frac{1}{2}[(x_1-1)^2+(x_2-1)^2]) + \exp(-\frac{1}{2}[(x_1+1)^2+(x_2+1)^2])$,
6. $b_6(x_1,x_2) := 1/(1+x_1^2+x_2^2)$,
7. $b_7(x_1,x_2) := \cos(x_1)\sin(x_2)$,
8. $b_8(x_1,x_2) := |x_1x_2|$.

The functions b_2 and b_6 are smooth bounded functions and have a unique maximum at 0, so they should be an easy case. The functions b_1 and b_5 are smooth and bounded with two maximums. The functions b_3 and b_7 are smooth oscillating functions. Finally the functions b_4 and b_8 are not smooth nor bounded, and should be a harder case.

Distribution of $\boldsymbol{X}$ For the sake of simplicity, we consider the case where $\boldsymbol{X}_1,\dots,\boldsymbol{X}_n$ are i.i.d. and have a density with respect to Lebesgue measure (i.e. $\nu = \text{Leb}$). For the case $p=1$, we consider the following distributions: $X \sim \mathcal{N}(0,1)$, and $X \sim$ Laplace. Both distributions are symmetric and centered at 0, but the normal distribution is more concentrated around its mean than the Laplace distribution. For the case $p=2$, we use independent marginals for the distribution of the covariates: $\boldsymbol{X} \sim \mathcal{N}(0,1)\otimes\mathcal{N}(0,1)$, and $\boldsymbol{X} \sim \text{Laplace}\otimes\text{Laplace}$.

Noise distribution We consider the normal distribution: $\varepsilon \sim \mathcal{N}(0,\sigma^2)$. The variance σ^2 is chosen such that the signal-to-noise ratio is the same for each choice of regression function and distribution of $\boldsymbol{X}$, where we define the signal-to-noise ratio as:

$$\text{SNR} := \frac{\|b\|_\mu^2}{\sigma^2}.$$

We consider the following values: $\text{SNR}=2$ (*High noise*), and $\text{SNR}=20$ (*Low noise*).

Parameters of the projection estimator Since the distributions of $\boldsymbol{X}$ are supported on $\mathbb{R}$ or $\mathbb{R}^2$, we choose the Hermite basis. The Hermite functions are defined as:

$$\varphi_j(x) := c_j H_j(x)\,\mathrm{e}^{-\frac{x^2}{2}}, \quad H_j(x) := (-1)^j \mathrm{e}^{x^2}\frac{\mathrm{d}^j}{\mathrm{d}x^j}\left[\mathrm{e}^{-x^2}\right], \quad c_j := \left(2^j j!\sqrt{\pi}\right)^{-1/2}.$$

and form a basis of $\mathrm{L}^2(\mathbb{R})$. We form a basis of $\mathrm{L}^2(\mathbb{R}^2)$ by tensorizing the Hermite basis as explained in Section 2. We choose the parameter $\hat{\boldsymbol{m}}$ with the model selection procedure (6). This procedure requires two additional parameters: the constant θ in the penalty and the constant β in the model collection $\widehat{\mathscr{M}}_{n,\beta}^{(2)}$.

We choose β such that the model collection $\widehat{\mathscr{M}}_{n,\beta}^{(2)}$ is not too small, especially for small sample sizes. Indeed, we find that the operator norm $\|\widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1}\|_{\text{op}}$ can grow very fast with $\boldsymbol{m}$, which can result in model collections with very few models. In our case, we choose $\beta = 10^4$.

The constant $\kappa := 1+\theta$ in front of the penalty is chosen following the "minimum penalty heuristic" (Arlot and Massart 2009). On several preliminary simulations, we compute the selected dimension $D_{\hat{\boldsymbol{m}}}$ as a function of κ and we find $\kappa_{\min}$ such that for $\kappa < \kappa_{\min}$ the dimension is too high and for $\kappa > \kappa_{\min}$ it is acceptable. Then, we choose $\kappa_\star = 2\kappa_{\min}$. In our case, we find $\kappa_\star = 2$ when $p=1$ and $p=2$.

Nadaraya–Watson estimator Let us define the Nadaraya–Watson estimator in the case $p = 1$. For all $h \in (0,1)$, let K_h be the pdf of the $\mathcal{N}(0,h)$ distribution. The Nadaraya–Watson estimator is defined as:

$$\forall x \in \mathbb{R}, \qquad \hat{b}_h^{\mathrm{NW}}(x) := \frac{\sum_{i=1}^n Y_i K_h(x - X_i)}{\sum_{i=1}^n K_h(x - X_i)}.$$

The bandwidth h is selected by leave-one-out cross validation, that is:

$$\hat{h} := \arg\min_h \sum_{i=1}^n \left(Y_i - \hat{b}_{h,-i}^{\mathrm{NW}}(X_i)\right)^2,$$

where $\hat{b}_{h,-i}^{\mathrm{NW}}$ is the Nadaraya–Watson estimator computed from the data set:

$$\left\{(X_j, Y_j) : j \in \{1, \dots n\} \setminus \{i\}\right\}.$$

In the case $p = 2$, the definition of the estimator is the same but with a couple of bandwidths $\boldsymbol{h} = (h_1, h_2) \in (0,1)^2$, and with $K_{\boldsymbol{h}}$ the pdf of the $\mathcal{N}_2(\mathbf{0}, \mathbf{H})$ distribution, where $\mathbf{H} := \mathrm{diag}(h_1, h_2)$.

Computation of the risk We consider samples of size $n = 250$ and $n = 1000$ in the case $p = 1$, and samples of size $n = 500$ and $n = 2000$ in the case $p = 2$. For each choice of regression function, distribution of $\boldsymbol{X}$ and SNR, we generate $N = 100$ samples of size n. For each sample, we compute the Hermite projection estimator and the Nadaraya–Watson estimator, then we compute the relative μ-error of the estimators, that is:

$$\text{relative error} := \frac{\|\hat{b} - b\|_\mu^2}{\|b\|_\mu^2} = \frac{\int_{\mathbb{R}^p} |\hat{b}(\boldsymbol{x}) - b(\boldsymbol{x})|^2 f(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}}{\int_{\mathbb{R}^p} b(\boldsymbol{x})^2 f(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}},$$

where f is the density of the distribution μ. We compute an approximation of these integrals: we consider a compact domain $I \times I$ with I an interval such that $\mathbb{P}[X \in I] = 95\%$ in the case $p = 1$ and $\mathbb{P}[\boldsymbol{X} \in I \times I] = 95\%$ in the case $p = 2$. Then, we consider a discretization with 200 points of I. In the case $p = 1$, we use Simpson's rule with this discretization of I to approximate the integrals. In the case $p = 2$, we approximate the integrals by a sum over the grid of $I \times I$:

$$\iint_{\mathbb{R}^2} \left|\hat{b}(\boldsymbol{x}) - b(\boldsymbol{x})\right|^2 f(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} \approx \sum_{i=1}^{200} \sum_{j=1}^{200} \left|\hat{b}(x_{1,i}, x_{2,j}) - b(x_{1,i}, x_{2,j})\right|^2 f(x_{1,i}, x_{2,j}) \Delta^2,$$

where Δ is the discretization step.

Results In the case $p = 1$, we show our results on Table 1. First of all, we see that the results are superior when X has a Normal distribution compared to a Laplace distribution. This can be explained by the fact that the Laplace distribution is less concentrated around 0 than the normal distribution, so the X_is are more scattered and the mu-risk covers a larger range. In addition, in the normal setting, we see that the Hermite estimator is better than the Nadaraya–Watson estimator for estimating b_1, b_2 and b_3, and both estimators are equivalent for estimating b_4. In the Laplace

X distrib.	Reg. fun.	Estim.	SNR $= 2$		SNR $= 20$	
			$n = 250$	$n = 1000$	$n = 250$	$n = 1000$
Norm.	b_1	Hermite	1.23 [1.22, 1.24] 4	0.288 [0.284, 0.292] 5	0.138 [0.136, 0.140] 6	0.034 [0.034, 0.035] 6
		NW	1.50 [1.49, 1.51] 0.307	0.468 [0.463, 0.472] 0.212	0.255 [0.253, 0.258] 0.724	0.076 [0.075, 0.076] 0.763
	b_2	Hermite	1.00 [0.99, 1.01] 3	0.362 [0.358, 0.366] 5	0.159 [0.157, 0.161] 6	0.047 [0.047, 0.047] 8
		NW	1.38 [1.37, 1.40] 0.281	0.475 [0.470, 0.480] 0.214	0.236 [0.234, 0.238] 0.161	0.075 [0.074, 0.076] 0.126
	b_3	Hermite	1.77 [1.76, 1.79] 10	0.477 [0.472, 0.482] 12	0.206 [0.204, 0.208] 11	0.050 [0.049, 0.050] 13
		NW	2.80 [2.78, 2.82] 0.138	0.823 [0.817, 0.829] 0.107	0.808 [0.799, 0.818] 0.088	0.160 [0.160, 0.161] 0.066
	b_4	Hermite	1.94 [1.92, 1.97] 9	0.532 [0.528, 0.536] 12	0.288 [0.286, 0.290] 11	0.116 [0.115, 0.116] 13
		NW	1.86 [1.84, 1.88] 0.216	0.585 [0.581, 0.590] 0.162	0.344 [0.341, 0.347] 0.120	0.108 [0.107, 0.108] 0.096
Lap.	b_1	Hermite	1.81 [1.78, 1.83] 5	0.400 [0.394, 0.405] 6	0.162 [0.159, 0.164] 6	0.047 [0.046, 0.047] 7
		NW	2.20 [2.18, 2.23] 0.347	0.686 [0.681, 0.691] 0.260	0.335 [0.332, 0.338] 0.182	0.104 [0.103, 0.105] 0.147
	b_2	Hermite	1.45 [1.43, 1.47] 3	0.426 [0.421, 0.430] 5	0.202 [0.199, 0.204] 7	0.064 [0.063, 0.064] 9
		NW	1.94 [1.92, 1.95] 0.315	0.725 [0.720, 0.731] 0.249	0.0337 [0.334, 0.339] 0.180	0.113 [0.112, 0.114] 0.145
	b_3	Hermite	4.56 [4.49, 4.63] 19	0.985 [0.979, 0.991] 27	1.39 [1.32, 1.47] 20	0.121 [0.120, 0.123] 29
		NW	3.57 [3.52, 3.61] 0.225	0.974 [0.968, 0.980] 0.184	1.09 [1.06, 1.11] 0.155	0.258 [0.254, 0.261] 0.137
	b_4	Hermite	8.61 [8.23, 8.98] 19	1.04 [1.04, 1.05] 28	1.59 [1.53, 1.65] 20	0.177 [0.175, 0.180] 29
		NW	2.30 [2.28, 2.33] 0.294	0.729 [0.724, 0.733] 0.224	0.454 [0.451, 0.457] 0.171	0.133 [0.133, 0.134] 0.127

Table 1 Risk comparison, $p = 1$**.** Table showing the relative μ-risks of the Hermite projection estimator and the Nadaraya–Watson estimator. For each distribution of X, regression function, SNR and n, we display the estimated relative μ-risk over $N = 100$ samples with a 95% confidence interval, multiplied by 100. For the projection estimator, we display the mean selected model, and for the Nadaraya–Watson estimator, we display the mean selected bandwidth.

$\boldsymbol{X}$ distrib.	Reg. fun.	Estim.	SNR $= 2$		SNR $= 20$	
			$n = 500$	$n = 2000$	$n = 500$	$n = 2000$
Norm.	b_5	Hermite	1.69 [1.68, 1.71] 12	0.587 [0.583, 0.591] 16	0.294 [0.191, 0.196] 21	0.067 [0.066, 0.067] 25
		NW	2.31 [2.29, 2.32] (0.382, 0.388)	0.845 [0.841, 0.848] (0.295, 0.297)	0.566 [0.564, 0.568] (0.231, 0.238)	0.217 [0.216, 0.218] (0.190, 0.188)
	b_6	Hermite	1.41 [1.40, 1.43] 5	0.732 [0.728, 0.735] 14	0.333 [0.331, 0.336] 26	0.094 [0.094, 0.095] 29
		NW	2.80 [2.78, 2.81] (0.327, 0.356)	1.10 [1.09, 1.10] (0.273, 0.272)	0.630 [0.628, 0.633] (0.213, 0.210)	0.249 [0.248, 0.250] (0.172, 0.172)
	b_7	Hermite	3.32 [3.29, 3.35] 26	0.916 [0.912, 0.919] 35	0.650 [0.645, 0.654] 43	0.123 [0.123, 0.124] 59
		NW	3.72 [3.70, 3.74] (0.280, 0.285)	1.45 [1.45, 1.46] (0.229, 0.225)	1.29 [1.28, 1.29] (0.181, 0.192)	0.420 [0.419, 0.421] (0.151, 0.147)
	b_8	Hermite	9.00 [8.89, 9.12] 50	2.01 [2.00, 2.02] 67	4.80 [3.66, 4.93] 51	0.847 [0.841, 0.853] 70
		NW	5.47 [5.44, 5.49] (0.255, 0.250)	2.08 [2.07, 2.08] (0.197, 0.197)	2.56 [2.55, 2.57] (0.179, 0.174)	0.769 [0.767, 0.771] (0.138, 0.137)
Lap.	b_5	Hermite	1.91 [1.90, 1.93] 12	0.703 [0.698, 0.708] 17	0.366 [0.359, 0.373] 21	0.076 [0.076, 0.077] 27
		NW	3.79 [3.77, 3.80] (0.451, 0.441)	1.66 [1.66, 1.67] (0.354, 0.357)	1.01 [1.01, 1.02] (0.252, 0.254)	0.404 [0.403, 0.405] (0.212, 0.208)
	b_6	Hermite	2.09 [2.07, 2.11] 7	0.962 [0.956, 0.968] 18	0.416 [0.412, 0.420] 27	0.172 [0.171, 0.173] 39
		NW	4.21 [4.19, 4.22] (0.422, 0.403)	1.80 [1.79, 1.80] (0.324, 0.339)	0.944 [0.941, 0.947] (0.231, 0.236)	0.401 [0.400, 0.402] (0.203, 0.199)
	b_7	Hermite	10.3 [10.1, 10.5] 30	5.56 [5.50, 5.62] 115	14.3 [13.9, 14.6] 76	1.49 [1.46, 1.52] 128
		NW	7.43 [7.40, 7.46] (0.350, 0.391)	2.80 [2.80, 2.81] (0.292, 0.235)	3.02 [3.01, 3.03] (0.230, 0.201)	0.931 [0.929, 0.933] (0.187, 0.167)
	b_8	Hermite	415 [406, 424] 77	74.1 [72.1, 76.0] 136	330 [322, 338] 79	71.2 [69.5, 72.9] 135
		NW	9.59 [9.55, 9.64] (0.351, 0.356)	3.34 [3.33, 3.35] (0.284, 0.275)	6.20 [6.17, 6.23] (0.257, 0.264)	1.75 [1.74, 1.76] (0.211, 0.209)

Table 2 Risk comparison, $p = 2$**.** Table showing the relative μ-risks of the Hermite projection estimator and the Nadaraya–Watson estimator. For each distribution of $\boldsymbol{X}$, regression function, SNR and n, we display the estimated relative μ-risk over $N = 100$ samples with a 95% confidence interval, multiplied by 100. For the projection estimator, we display the mean selected dimension, and for the Nadaraya–Watson estimator, we display the mean selected bandwidths.

setting, the Hermite estimator is still better for b_1 and b_2, but for b_3 it has similar performances as the Nadaraya–Watson estimator. For estimating b_4, the latter is better, although the difference becomes small as n increases.

In the case $p = 2$, we show our results on Table 2. In the normal setting, the Hermite projection estimator is better for estimating b_5, b_6 and b_7. For b_8, its performances are worse than the kernel estimator on small samples but they are equivalent on large samples. In the Laplace setting, our estimator is better for estimating b_5 and b_6, but it is worse for estimating b_7. Moreover, the Hermite estimator has very poor performances for estimating b_8. We think that the functions b_7 and b_8 are hard to approximate with the Hermite basis, so that the Hermite projection estimator performs poorly. This can be seen by looking at the mean selected dimension, which grows quickly as n grows, showing that the estimator needs a large number of coefficients to reconstruct the regression function. This is especially true for b_8, as it is a non differentiable and unbounded function.

In addition, we observe that the Hermite estimator is faster to compute than the Nadaraya–Watson estimator with leave-one-out cross validation. The difference is small when n is small, but for example, when $n = 2000$ and $p = 2$, the Hermite estimator is about 3 time faster. In conclusion, the Hermite projection estimator is a good alternative to the Nadaraya–Watson estimator.

6 Concluding remark

In this paper, we have considered the nonparametric regression problem with a random design. The covariates are assumed to be independent but not identically distributed, and the variance of the noise is assumed to be known. We estimate the regression function on a non-compact domain of $\mathbb{R}^p$ with a projection estimator, using tensorised orthonormal bases. The projection space is chosen by a penalized criterion, as in Birgé and Massart (1998) and Baraud (2000). Our model collection depends on the design, and is thus random. Indeed, we consider subspaces $S_{\boldsymbol{m}}$ on which the operator norm of the Gram hypermatrix associated to the least squared minimization problem is constrained. This constraint on the operator norm comes from a refined study of the discrepancy between the norms $\|\cdot\|_n$ and $\|\cdot\|_\mu$ on $S_{\boldsymbol{m}}$. This study relies on Matrix concentration inequalities of Tropp (2012) and Gittens and Tropp (2011), as it has been suggested by the work of Cohen et al (2013). Doing so, we obtain oracle bounds for the selected estimator, in both norms. Our work extends and improves the results of Baraud (2002) and Comte and Genon-Catalot (2020a), as explained by Remark 5.

Different extension of our work can be pursued. A natural extension would be to consider the heteroskedastic regression model, in which the observations $(\boldsymbol{X}_i, Y_i)$ satisfy:

$$Y_i = b(\boldsymbol{X}_i) + \sigma(\boldsymbol{X}_i)\varepsilon_i,$$

were ε_is have unit variance. Using the same projection estimator, Comte and Genon-Catalot (2020b) have obtained similar results for this model in the one-dimensional case. The extension to the multivariate case could be done in two ways. The first way would be to generalize the fixed design results of Baraud (2000) to the case of noise

variables with different variance, and then to apply the same arguments we used in this paper to deduce the results for the random design setting. The second way would be be to follow the approach of Comte and Genon-Catalot (2020b), that is based on Talagrand's inequality, and to see if it can be extended to the multivariate case.
Another extension of our work would be to investigate the use of more general approximation spaces S_m, as does Baraud (2002). We want to known if the same method we used could handle approximation spaces that are not constructed from an orthonormal basis. A typical example we have in mind is splines approximation. We suspect that our results on the comparison between the norms $\|\cdot\|_n$ and $\|\cdot\|_\mu$ still hold in this context, so that adaptive strategies could be derived from it.

7 Proofs

7.1 Proofs of Section 2

Proof (Proposition 1) Let $\Pi_{\boldsymbol{m}}^{(n)}$ be the projector on $S_{\boldsymbol{m}}$ for the empirical inner product. We have the decomposition:

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{X}}\|b-\hat{b}_{\boldsymbol{m}}\|_n^2 &= \|b-\Pi_{\boldsymbol{m}}^{(n)}b\|_n^2 + \mathbb{E}_{\boldsymbol{X}}\|\hat{b}_{\boldsymbol{m}}-\Pi_{\boldsymbol{m}}^{(n)}b\|_n^2 \\
&= \inf_{t\in S_{\boldsymbol{m}}}\|b-t\|_n^2 + \mathbb{E}_{\boldsymbol{X}}\|\Pi_{\boldsymbol{m}}^{(n)}\boldsymbol{\varepsilon}\|_n^2 \\
&= \inf_{t\in S_{\boldsymbol{m}}}\|b-t\|_n^2 + \sigma^2\frac{\mathrm{Tr}\left(\Pi_{\boldsymbol{m}}^{(n)}\right)}{n} \\
&= \inf_{t\in S_{\boldsymbol{m}}}\|b-t\|_n^2 + \sigma^2\frac{D_{\boldsymbol{m}}}{n}.
\end{aligned}$$

Taking the expected value in this equality, we obtain:

$$\begin{aligned}
\mathbb{E}\|b-\hat{b}_{\boldsymbol{m}}\|_n^2 = \mathbb{E}\left[\inf_{t\in S_{\boldsymbol{m}}}\|b-t\|_n^2\right] + \sigma^2\frac{D_{\boldsymbol{m}}}{n} &\leq \inf_{t\in S_{\boldsymbol{m}}}\mathbb{E}\|b-t\|_n^2 + \sigma^2\frac{D_{\boldsymbol{m}}}{n} \\
&= \inf_{t\in S_{\boldsymbol{m}}}\mathbb{E}\|b-t\|_\mu^2 + \sigma^2\frac{D_{\boldsymbol{m}}}{n}.
\end{aligned}$$

□

7.2 Proofs of Section 3

Proof (Lemma 2) Let $x\in A$ and let $t=\sum_{j=1}^d a_j\,\psi_j\in S$. The family of functions $(\psi_1,\cdots,\psi_d)$ is orthonormal with respect to $\langle\cdot,\cdot\rangle_\alpha$, so by the Cauchy–Schwarz inequality we have:

$$t^2(x) = \left(\sum_{j=1}^d a_j\,\psi_j(x)\right)^2 \leq \left(\sum_{j=1}^d a_j^2\right)\left(\sum_{j=1}^d \psi_j^2(x)\right) = \|t\|_\alpha^2\sum_{j=1}^d\psi_j^2(x),$$

with equality if $(\alpha_1,\dots,\alpha_d)$ is proportional to $(\psi_1(x),\dots,\psi_d(x))$. Hence we have:

$$\sum_{j=1}^{d}\psi_j^2(x)=\sup_{t\in S\setminus\{0\}}\frac{t^2(x)}{\|t\|_\alpha^2}.$$

Taking the supremum for $x\in A$, we obtain:

$$\sup_{x\in A}\sum_{j=1}^{d}\psi_j^2(x)=\sup_{x\in A}\sup_{t\in S\setminus\{0\}}\frac{t^2(x)}{\|t\|_\alpha^2}=\sup_{t\in S\setminus\{0\}}\frac{\sup_{x\in A}t^2(x)}{\|t\|_\alpha^2},$$

that is:

$$\left\|\sum_{j=1}^{d}\psi_j^2\right\|_\infty=\sup_{t\in S\setminus\{0\}}\frac{\|t\|_\infty^2}{\|t\|_\alpha^2}=:K_\alpha^\infty(S).$$

□

To prove Proposition 3 and Theorem 2, we need the following lemma.

Lemma 3 *Let $(\psi_1,\dots,\psi_{D_{\boldsymbol{m}}})$ be an orthnormal basis of $S_{\boldsymbol{m}}$ relatively to an inner product $\langle\cdot,\cdot\rangle_\alpha$. Let $\widehat{\mathbf{H}}_{\boldsymbol{m}}$ be the Gram matrix of this basis relatively to the empirical inner product and let $\mathbf{H}_{\boldsymbol{m}}:=\mathbb{E}[\widehat{\mathbf{H}}_{\boldsymbol{m}}]$, that is:*

$$\forall j,k\in\{1,\dots,D_{\boldsymbol{m}}\},\quad \left[\widehat{\mathbf{H}}_{\boldsymbol{m}}\right]_{j,k}:=\langle\psi_j,\psi_k\rangle_n \text{ and } \left[\mathbf{H}_{\boldsymbol{m}}\right]_{j,k}:=\langle\psi_j,\psi_k\rangle_\mu.$$

For all $\delta\in(0,1)$ we have:

$$\mathbb{P}\left[\lambda_{\min}(\widehat{\mathbf{H}}_{\boldsymbol{m}})\le(1-\delta)\lambda_{\min}(\mathbf{H}_{\boldsymbol{m}})\right]\le D_{\boldsymbol{m}}\exp\left(-h(\delta)\frac{n\lambda_{\min}(\mathbf{H}_{\boldsymbol{m}})}{K_\alpha^\infty(\boldsymbol{m})}\right),$$

with $h(\delta):=\delta+(1-\delta)\log(1-\delta)$ and where $K_\alpha^\infty(\boldsymbol{m})$ is given by Lemma 2.

Proof We use Theorem 5 in Appendix. Indeed, $\widehat{\mathbf{H}}_{\boldsymbol{m}}$ can be written as a sum $\mathbf{Z}_1+\ldots+\mathbf{Z}_n$ where:

$$\forall j,k\in\{1,\dots,D_{\boldsymbol{m}}\},\quad \left[\mathbf{Z}_i\right]_{j,k}:=\frac{1}{n}\psi_j(\boldsymbol{X}_i)\psi_k(\boldsymbol{X}_i),$$

so we have using Lemma 2:

$$\lambda_{\max}(\mathbf{Z}_i)=\|\mathbf{Z}_i\|_{\text{op}}=\frac{1}{n}\sum_{k=1}^{D_{\boldsymbol{m}}}\psi_k(\boldsymbol{X}_i)^2\le\frac{1}{n}\left\|\sum_{k=1}^{D_{\boldsymbol{m}}}\psi_k^2\right\|_\infty=\frac{1}{n}K_\alpha^\infty(\boldsymbol{m}).$$

Therefore, applying inequality (29) of Theorem 5 with $\mu_{\min}=\lambda_{\min}(\mathbf{H}_{\boldsymbol{m}})$ and $R=\frac{1}{n}K_\alpha^\infty(\boldsymbol{m})$ yields:

$$\mathbb{P}\left[\lambda_{\min}(\widehat{\mathbf{H}}_{\boldsymbol{m}})\le(1-\delta)\lambda_{\min}(\mathbf{H}_{\boldsymbol{m}})\right]\le D_{\boldsymbol{m}}\exp\left(-h(\delta)\frac{n\lambda_{\min}(\mathbf{H}_{\boldsymbol{m}})}{K_\alpha^\infty(\boldsymbol{m})}\right).$$

□

Proof (Proposition 3) Let $\psi_1, \dots, \psi_{D_{\boldsymbol{m}}}$ be an orthonormal basis of $S_{\boldsymbol{m}}$ relatively to the inner product $\langle \cdot, \cdot \rangle_\mu$. Let $\widehat{\mathbf{H}}_{\boldsymbol{m}}$ be their Gram matrix relatively to the empirical inner product. According to Lemma 1, we have $K_n^\mu(\boldsymbol{m}) = \|\widehat{\mathbf{H}}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}} = \lambda_{\min}(\widehat{\mathbf{H}}_{\boldsymbol{m}})^{-1}$ and we have $\mathbb{E}[\widehat{\mathbf{H}}_{\boldsymbol{m}}] = \mathbf{I}_{\boldsymbol{m}}$ because $(\psi_1, \dots, \psi_{D_{\boldsymbol{m}}})$ is orthonormal for the inner product associated with μ, so the event $\Omega_{\boldsymbol{m}}(\delta)^c$ can be written as:

$$\Omega_{\boldsymbol{m}}(\delta)^c = \Big\{ \lambda_{\min}(\widehat{\mathbf{H}}_{\boldsymbol{m}}) \leq 1 - \delta \Big\} = \Big\{ \lambda_{\min}(\widehat{\mathbf{H}}_{\boldsymbol{m}}) \leq (1-\delta)\lambda_{\min}(\mathbb{E}[\widehat{\mathbf{H}}_{\boldsymbol{m}}]) \Big\}.$$

Applying Lemma 3 yields the result. □

Proof (Proposition 2) We start with the decomposition:

$$\mathbb{E}\|b - \hat{b}_{\boldsymbol{m}}\|_\mu^2 = \mathbb{E}\big[\|b - \hat{b}_{\boldsymbol{m}}\|_\mu^2 \mathbf{1}_{\Omega_{\boldsymbol{m}}(\delta)}\big] + \mathbb{E}\big[\|b - \hat{b}_{\boldsymbol{m}}\|_\mu^2 \mathbf{1}_{\Omega_{\boldsymbol{m}}(\delta)^c}\big]. \tag{7}$$

We consider these two terms separately. The expectation of the first term is controlled as in Theorem 3 in Cohen et al (2013). On the event $\Omega_{\boldsymbol{m}}(\delta)$ we have $(1-\delta)\|t\|_\mu^2 \leq \|t\|_n^2$ for all $t \in S_{\boldsymbol{m}}$, so if $b_{\boldsymbol{m}}^{(\mu)}$ is the projection of b on $S_{\boldsymbol{m}}$ for the norm $\|\cdot\|_\mu$, we have:

$$\begin{aligned}
\|b - \hat{b}_{\boldsymbol{m}}\|_\mu^2 \mathbf{1}_{\Omega_{\boldsymbol{m}}(\delta)} &\leq \|b - b_{\boldsymbol{m}}^{(\mu)}\|_\mu^2 + \|\hat{b}_{\boldsymbol{m}} - b_{\boldsymbol{m}}^{(\mu)}\|_\mu^2 \mathbf{1}_{\Omega_{\boldsymbol{m}}(\delta)} \\
&\leq \|b - b_{\boldsymbol{m}}^{(\mu)}\|_\mu^2 + 2\|\hat{b}_{\boldsymbol{m}} - b_{\boldsymbol{m}}^{(n)}\|_\mu^2 \mathbf{1}_{\Omega_{\boldsymbol{m}}(\delta)} + 2\|b_{\boldsymbol{m}}^{(n)} - b_{\boldsymbol{m}}^{(\mu)}\|_\mu^2 \mathbf{1}_{\Omega_{\boldsymbol{m}}(\delta)} \\
&\leq \|b - b_{\boldsymbol{m}}^{(\mu)}\|_\mu^2 + \frac{2}{1-\delta}\|\hat{b}_{\boldsymbol{m}} - b_{\boldsymbol{m}}^{(n)}\|_n^2 + 2\|b_{\boldsymbol{m}}^{(n)} - b_{\boldsymbol{m}}^{(\mu)}\|_\mu^2 \mathbf{1}_{\Omega_{\boldsymbol{m}}(\delta)}
\end{aligned}$$

Taking the expectation, we obtain:

$$\mathbb{E}\big[\|b - \hat{b}_{\boldsymbol{m}}\|_\mu^2 \mathbf{1}_{\Omega_{\boldsymbol{m}}(\delta)}\big] \leq \|b - b_{\boldsymbol{m}}^{(\mu)}\|_\mu^2 + \frac{2}{1-\delta}\sigma^2\frac{D_{\boldsymbol{m}}}{n} + 2\,\mathbb{E}\Big[\|b_{\boldsymbol{m}}^{(n)} - b_{\boldsymbol{m}}^{(\mu)}\|_\mu^2 \mathbf{1}_{\Omega_{\boldsymbol{m}}(\delta)}\Big]. \tag{8}$$

We give an upper bound on the last term in two ways. Firstly, we have:

$$\begin{aligned}
\mathbb{E}\Big[\|b_{\boldsymbol{m}}^{(n)} - b_{\boldsymbol{m}}^{(\mu)}\|_\mu^2 \mathbf{1}_{\Omega_{\boldsymbol{m}}(\delta)}\Big] &\leq \mathbb{E}\Big[K_n^\mu(\boldsymbol{m})\|b_{\boldsymbol{m}}^{(n)} - b_{\boldsymbol{m}}^{(\mu)}\|_n^2 \mathbf{1}_{\Omega_{\boldsymbol{m}}(\delta)}\Big] \\
&\leq \frac{1}{1-\delta}\mathbb{E}\|b_{\boldsymbol{m}}^{(n)} - b_{\boldsymbol{m}}^{(\mu)}\|_n^2
\end{aligned}$$

since $K_n^\mu(\boldsymbol{m}) \leq \frac{1}{1-\delta}$ on the event $\Omega_{\boldsymbol{m}}(\delta)$, see (2). Let $\Pi_{\boldsymbol{m}}^{(n)}$ be the empirical projector on $S_{\boldsymbol{m}}$, we have:

$$\|b_{\boldsymbol{m}}^{(n)} - b_{\boldsymbol{m}}^{(\mu)}\|_n^2 = \Big\|\Pi_{\boldsymbol{m}}^{(n)}\big(b - b_{\boldsymbol{m}}^{(\mu)}\big)\Big\|_n^2 \leq \|b - b_{\boldsymbol{m}}^{(\mu)}\|_n^2.$$

Thus, we have shown:

$$\mathbb{E}\Big[\|b_{\boldsymbol{m}}^{(n)} - b_{\boldsymbol{m}}^{(\mu)}\|_\mu^2 \mathbf{1}_{\Omega_{\boldsymbol{m}}(\delta)}\Big] \leq \frac{1}{1-\delta}\mathbb{E}\|b - b_{\boldsymbol{m}}^{(\mu)}\|_n^2 = \frac{1}{1-\delta}\|b - b_{\boldsymbol{m}}^{(\mu)}\|_\mu^2. \tag{9}$$

Secondly, let $g \coloneqq b - b_{\boldsymbol{m}}^{(\mu)}$ and let $\Pi_{\boldsymbol{m}}^{(n)}$ be the empirical projector on $S_{\boldsymbol{m}}$ we have:

$$\mathbb{E}\Big[\|b_{\boldsymbol{m}}^{(n)} - b_{\boldsymbol{m}}^{(\mu)}\|_\mu^2 \mathbf{1}_{\Omega_{\boldsymbol{m}}(\delta)}\Big] = \mathbb{E}\Big[\|\Pi_{\boldsymbol{m}}^{(n)} g\|_\mu^2 \mathbf{1}_{\Omega_{\boldsymbol{m}}(\delta)}\Big].$$

Let $(\psi_1,\dots,\psi_{D_{\boldsymbol{m}}})$ be an orthonormal basis of $S_{\boldsymbol{m}}$ for the inner product $\langle\cdot,\cdot\rangle_\mu$, we have:

$$\Pi_{\boldsymbol{m}}^{(n)} g = \operatorname*{arg\,min}_{t\in S_{\boldsymbol{m}}} \|g-t\|_n^2 = \sum_{j=1}^{D_{\boldsymbol{m}}} c_j^\star\,\psi_j, \quad \mathbf{c}^\star := \operatorname*{arg\,min}_{\mathbf{c}\in\mathbb{R}^{D_{\boldsymbol{m}}}} \|\mathbf{g}-\Psi_{\boldsymbol{m}}\mathbf{c}\|_{\mathbb{R}^n}^2,$$

where $\Psi_{\boldsymbol{m}}\in\mathbb{R}^{n\times D_{\boldsymbol{m}}}$ is the matrix defined by $[\Psi_{\boldsymbol{m}}]_{i,j} := \psi_j(\boldsymbol{X}_i)$, and where $\mathbf{g}$ is the vector $\big(g(\boldsymbol{X}_1),\dots,g(\boldsymbol{X}_n)\big)\in\mathbb{R}^n$. By Lemma 8, $\mathbf{c}^\star$ is given by:

$$\mathbf{c}^\star = (\Psi_{\boldsymbol{m}}^*\Psi_{\boldsymbol{m}})^{-1}\Psi_{\boldsymbol{m}}^*\mathbf{g} = \frac{1}{n}\mathbf{H}_{\boldsymbol{m}}^{-1}\Psi_{\boldsymbol{m}}^*\mathbf{g},$$

where $\mathbf{H}_{\boldsymbol{m}}$ is the Gram matrix of $(\psi_1,\dots,\psi_{D_{\boldsymbol{m}}})$ relatively to the empirical inner product. Using Lemma 1, we get:

$$\big\|\Pi_{\boldsymbol{m}}^{(n)} g\big\|_\mu^2 = \big\|\mathbf{c}^\star\big\|_{\mathbb{R}^{D_{\boldsymbol{m}}}}^2 \le \big\|\mathbf{H}_{\boldsymbol{m}}^{-1}\big\|_{\mathrm{op}}^2 \Big\|\frac{1}{n}\Psi_{\boldsymbol{m}}^*\mathbf{g}\Big\|_{\mathbb{R}^{D_{\boldsymbol{m}}}}^2 = K_n^\mu(\boldsymbol{m})^2\sum_{j=1}^{D_{\boldsymbol{m}}}\langle g,\psi_j\rangle_n^2.$$

Hence, on the event $\Omega_{\boldsymbol{m}}(\delta)$ we obtain:

$$\big\|\Pi_{\boldsymbol{m}}^{(n)} g\big\|_\mu^2 \mathbf{1}_{\Omega_{\boldsymbol{m}}(\delta)} \le \frac{1}{(1-\delta)^2}\sum_{j=1}^{D_{\boldsymbol{m}}}\langle g,\psi_j\rangle_n^2.$$

Since $g = b - b_{\boldsymbol{m}}^{(\mu)}$ is orthogonal to $\psi_1,\dots,\psi_{D_{\boldsymbol{m}}}$ relatively to the inner product $\langle\cdot,\cdot\rangle_\mu$, we have $\mathbb{E}[\langle g,\psi_j\rangle_n] = \langle g,\psi_j\rangle_\mu = 0$, so we get:

$$\begin{aligned}
\mathbb{E}\left[\sum_{k=1}^{D_{\boldsymbol{m}}}\langle g,\psi_k\rangle_n^2\right] = \sum_{k=1}^{D_{\boldsymbol{m}}}\mathrm{Var}\big(\langle g,\psi_k\rangle_n\big) &= \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^{D_{\boldsymbol{m}}}\mathrm{Var}\big(g(\boldsymbol{X}_i)\psi_j(\boldsymbol{X}_i)\big)\\
&= \frac{1}{n^2}\sum_{i=1}^n\mathbb{E}\left[g(\boldsymbol{X}_i)^2\sum_{j=1}^{D_{\boldsymbol{m}}}\psi_j(\boldsymbol{X}_i)^2\right]\\
&\le \frac{1}{n^2}\sum_{i=1}^n\mathbb{E}\big[g(\boldsymbol{X}_i)^2\big]\sup_{x\in A}\sum_{j=1}^{D_{\boldsymbol{m}}}\psi_j(x)^2\\
&= \frac{1}{n}\|g\|_\mu^2 K_\mu^\infty(\boldsymbol{m}) = \frac{K_\mu^\infty(\boldsymbol{m})}{n}\|b-b_{\boldsymbol{m}}^{(\mu)}\|_\mu^2,
\end{aligned}$$

where the last equality comes from Lemma 2. Hence we have shown:

$$\mathbb{E}\Big[\|b_{\boldsymbol{m}}^{(n)}-b_{\boldsymbol{m}}^{(\mu)}\|_\mu^2\,\mathbf{1}_{\Omega_{\boldsymbol{m}}(\delta)}\Big] \le \frac{1}{(1-\delta)^2}\frac{K_\mu^\infty(\boldsymbol{m})}{n}\|b-b_{\boldsymbol{m}}^{(\mu)}\|_\mu^2. \tag{10}$$

Combining (9) and (10) yields:

$$\mathbb{E}\Big[\|b_{\boldsymbol{m}}^{(n)}-b_{\boldsymbol{m}}^{(\mu)}\|_\mu^2\,\mathbf{1}_{\Omega_{\boldsymbol{m}}(\delta)}\Big] \le \frac{1}{1-\delta}\|b-b_{\boldsymbol{m}}^{(\mu)}\|_\mu^2\left(1\wedge\frac{K_\mu^\infty(\boldsymbol{m})}{(1-\delta)n}\right). \tag{11}$$

For the second term in (7), we have:

$$\mathbb{E}\big[\|b-\hat{b}_{\boldsymbol{m}}\|_\mu^2\mathbf{1}_{\Omega_{\boldsymbol{m}}(\delta)^c}\big] \le 2\|b\|_\mu^2\,\mathbb{P}[\Omega_{\boldsymbol{m}}(\delta)^c] + 2\,\mathbb{E}\big[\|\hat{b}_{\boldsymbol{m}}\|_\mu^2\mathbf{1}_{\Omega_{\boldsymbol{m}}(\delta)^c}\big].$$

We have the following upper bound on $\|\hat{b}_{\boldsymbol{m}}\|_\mu^2$:

$$\|\hat{b}_{\boldsymbol{m}}\|_\mu^2 \leq K_n^\mu(\boldsymbol{m})\,\|\hat{b}_{\boldsymbol{m}}\|_n^2 \leq K_n^\mu(\boldsymbol{m})\,\|\mathbf{Y}\|_n^2, \tag{12}$$

where the last inequality comes from the fact that $\hat{b}_{\boldsymbol{m}}$ is the empirical projection of $\mathbf{Y}$. Hence, we get:

$$\mathbb{E}\big[\|b-\hat{b}_{\boldsymbol{m}}\|_\mu^2 \mathbf{1}_{\Omega_{\boldsymbol{m}}(\delta)^c}\big] \leq 2\|b\|_\mu^2\,\mathbb{P}[\Omega_{\boldsymbol{m}}(\delta)^c] + 2\,\mathbb{E}\big[K_n^\mu(\boldsymbol{m})\,\|\mathbf{Y}\|_n^2 \mathbf{1}_{\Omega_{\boldsymbol{m}}(\delta)^c}\big]. \tag{13}$$

The inequality of Proposition 2 is obtained using (8), (11) and (13) in (7). □

Proof (Theorem 1) Let $\boldsymbol{m} \in \mathscr{M}_{n,\alpha}^{(1)}$ and let $\delta \in (0,1)$ (we choose it later in the proof). By Remark 1, we have by definition of $\mathscr{M}_{n,\alpha}^{(1)}$:

$$K_\mu^\infty(\boldsymbol{m}) \leq K_\nu^\infty(\boldsymbol{m})\|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}} \leq \alpha\frac{n}{\log n}, \tag{14}$$

so Proposition 2 yields:

$$\mathbb{E}\|b-\hat{b}_{\boldsymbol{m}}\|_\mu^2 \leq C_n(\delta,\alpha)\inf_{t\in S_{\boldsymbol{m}}}\|b-t\|_\mu^2 + C'(\delta)\sigma^2\frac{D_{\boldsymbol{m}}}{n} + R_n,$$

with $C_n(\alpha,\delta) \coloneqq \left(1+\frac{2}{1-\delta}\left[\frac{\alpha}{(1-\delta)\log n}\wedge 1\right]\right)$, $C'(\delta) \coloneqq \frac{2}{1-\delta}$ and:

$$R_n \coloneqq 2\|b\|_\mu^2\,\mathbb{P}[\Omega_{\boldsymbol{m}}(\delta)^c] + \mathbb{E}\big[K_n^\mu(\boldsymbol{m})\|\mathbf{Y}\|_n^2\,\mathbf{1}_{\Omega_{\boldsymbol{m}}(\delta)^c}\big].$$

For the first term in R_n, we apply Proposition 3 with (14):

$$\mathbb{P}[\Omega_{\boldsymbol{m}}(\delta)^c] \leq D_{\boldsymbol{m}}\,n^{-\frac{h(\delta)}{\alpha}} \leq n^{-\frac{h(\delta)}{\alpha}+1}. \tag{15}$$

For the second term in R_n, since $\|\cdot\|_\mu \leq \|\cdot\|_\infty$ and $\boldsymbol{m} \in \mathscr{M}_{n,\alpha}^{(1)}$ we have:

$$K_n^\mu(\boldsymbol{m}) \leq K_\nu^\mu(\boldsymbol{m})\,K_n^\nu(\boldsymbol{m}) \leq K_\nu^\infty(\boldsymbol{m})\,\|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}} \leq \alpha\frac{n}{\log n}, \tag{16}$$

and we have using the independence of $(\boldsymbol{X}_i)_{1\leq i\leq n}$ and $(\varepsilon_i)_{1\leq i\leq n}$:

$$\begin{aligned}\mathbb{E}\big[\|\mathbf{Y}\|_n^2\mathbf{1}_{\Omega_{\boldsymbol{m}}(\delta)^c}\big] &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}\Big[\big(b(\boldsymbol{X}_i)+\varepsilon_i\big)^2\mathbf{1}_{\Omega_{\boldsymbol{m}}(\delta)^c}\Big] \\ &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n b(\boldsymbol{X}_i)^2\mathbf{1}_{\Omega_{\boldsymbol{m}}(\delta)^c}\right] + \sigma^2\,\mathbb{P}[\Omega_{\boldsymbol{m}}(\delta)^c].\end{aligned}$$

We apply Hölder's inequality with $r,r' \in (1,+\infty)$ such that $\frac{1}{r}+\frac{1}{r'}=1$:

$$\begin{aligned}\mathbb{E}\big[\|\mathbf{Y}\|_n^2\mathbf{1}_{\Omega_{\boldsymbol{m}}(\delta)^c}\big] &\leq \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n b(\boldsymbol{X}_i)^2\right)^r\right]^{\frac{1}{r}}\mathbb{P}\big[\Omega_{\boldsymbol{m}}(\delta)^c\big]^{\frac{1}{r'}} + \sigma^2\,\mathbb{P}\big[\Omega_{\boldsymbol{m}}(\delta)^c\big] \\ &\leq \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n b(\boldsymbol{X}_i)^{2r}\right]^{\frac{1}{r}}\mathbb{P}\big[\Omega_{\boldsymbol{m}}(\delta)^c\big]^{\frac{1}{r'}} + \sigma^2\,\mathbb{P}\big[\Omega_{\boldsymbol{m}}(\delta)^c\big] \\ &\leq \|b\|_{\mathrm{L}^{2r}(\mu)}^2\,n^{-\frac{h(\delta)}{\alpha r'}+\frac{1}{r'}} + \sigma^2\,n^{-\frac{h(\delta)}{\alpha}+1},\end{aligned}$$

and if $b \in \mathrm{L}^\infty(\mu)$, the last inequality also holds for $r = \infty$ and $r' = 1$ (just take the limit as $r \to +\infty$). Hence, we obtain:

$$\mathbb{E}\left[K_n^\mu(\boldsymbol{m})\|\mathbf{Y}\|_n^2 \mathbf{1}_{\Omega_{\boldsymbol{m}}(\delta)^c}\right] \leq \frac{\alpha}{\log n}\left(\|b\|^2_{\mathrm{L}^{2r}(\mu)} n^{-\frac{h(\delta)}{\alpha r'}+\frac{1}{r'}+1} + \sigma^2 n^{-\frac{h(\delta)}{\alpha}+2}\right). \quad (17)$$

If we choose δ such that $h(\delta) \geq (2r'+1)\alpha$, then all the exponents of n in (15) and (17) are less than -1. The function h is an increasing function from $[0,1]$ to itself so it is invertible on $[0,1]$. Since $\alpha \in (0, \frac{1}{2r'+1})$, we can choose $\delta = \delta(\alpha, r') := h^{-1}((2r'+1)\alpha)$. For this choice, we obtain:

$$\mathbb{E}\|b - \hat{b}_{\boldsymbol{m}}\|_\mu^2 \leq C_n\big(\delta(\alpha,r'),\alpha\big) \inf_{t \in S_{\boldsymbol{m}}} \|b-t\|_\mu^2 + C'\big(\delta(\alpha,r')\big)\sigma^2 \frac{D_{\boldsymbol{m}}}{n} + \frac{C''\big(b,\sigma^2,\alpha,r\big)}{n\log n},$$

where $C_n(\delta,\alpha)$ and $C'(\delta)$ were defined at the beginning of the proof, and where:

$$C''\big(b,\sigma^2,\alpha,r\big) \leq 2\|b\|^2_{\mathrm{L}^2(\mu)} + \alpha\left(\|b\|^2_{\mathrm{L}^{2r}(\mu)} + \sigma^2\right). \quad (18)$$

□

7.3 Proof of Theorem 2

The proof of Theorem 2 is based on a result for fixed design regression of Baraud (2000). Let $\widehat{\mathscr{M}_n}$ be a finite collection of models, that may depend on $(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n)$, such that for all $\boldsymbol{m} \in \widehat{\mathscr{M}_n}$, $\widehat{\mathbf{G}}_{\boldsymbol{m}}$ is invertible. Let $\hat{\boldsymbol{m}} \in \widehat{\mathscr{M}_n}$ be the minimizer of the following penalized least squares criterion:

$$\hat{\boldsymbol{m}} := \operatorname*{arg\,min}_{\boldsymbol{m} \in \widehat{\mathscr{M}_n}} \left(-\|\hat{b}_{\boldsymbol{m}}\|_n^2 + \mathrm{pen}(\boldsymbol{m})\right), \quad \mathrm{pen}(\boldsymbol{m}) := (1+\theta)\sigma^2\frac{D_{\boldsymbol{m}}}{n}, \quad \theta > 0. \quad (19)$$

Theorem 4 (Corollary 3.1 in Baraud (2000)) *If $\mathbb{E}|\varepsilon_1|^q$ is finite for some $q > 4$, then the following upper bound on the risk of the estimator $\hat{b}_{\hat{\boldsymbol{m}}}$ with $\hat{\boldsymbol{m}}$ defined by* (19) *holds:*

$$\mathbb{E}_{\boldsymbol{X}}\|b - \hat{b}_{\hat{\boldsymbol{m}}}\|_n^2 \leq C(\theta) \inf_{\boldsymbol{m} \in \widehat{\mathscr{M}_n}} \left(\inf_{t \in S_{\boldsymbol{m}}} \|b-t\|_n^2 + \sigma^2\frac{D_{\boldsymbol{m}}}{n}\right) + \sigma^2\frac{\Sigma_n(\theta,q)}{n},$$

with:

$$\Sigma_n(\theta,q) := C'(\theta,q)\frac{\mathbb{E}|\varepsilon_1|^q}{\sigma^q} \sum_{\boldsymbol{m} \in \widehat{\mathscr{M}_n}} D_{\boldsymbol{m}}^{-(\frac{q}{2}-2)},$$

where $C(\theta) := (2+8\theta^{-1})(1+\theta)$ and $C'(\theta,q)$ is a positive constant.

Proof (Theorem 2) Let $\Delta_{n,\alpha,\beta} \coloneqq \{\mathscr{M}_{n,\alpha}^{(1)} \subset \widehat{\mathscr{M}}_{n,\beta}^{(1)}\}$, we have:

$$\mathbb{E}\|b-\hat{b}_{\hat{\boldsymbol{m}}_1}\|_n^2 = \mathbb{E}\Big[\mathbb{E}_{\boldsymbol{X}}\|b-\hat{b}_{\hat{\boldsymbol{m}}_1}\|_n^2 \mathbf{1}_{\Delta_{n,\alpha,\beta}}\Big] + \mathbb{E}\Big[\|b-\hat{b}_{\hat{\boldsymbol{m}}_1}\|_n^2 \mathbf{1}_{\Delta_{n,\alpha,\beta}^c}\Big].$$

For the first term, on $\Delta_{n,\alpha,\beta}$ we have $\inf_{\boldsymbol{m}\in\widehat{\mathscr{M}}_{n,\beta}^{(1)}}(\ldots) \leq \inf_{\boldsymbol{m}\in\mathscr{M}_{n,\alpha}^{(1)}}(\ldots)$ so by applying Theorem 4 we obtain:

$$\begin{aligned}\mathbb{E}\Big[\mathbb{E}_{\boldsymbol{X}}\|b-\hat{b}_{\hat{\boldsymbol{m}}_1}\|_n^2 \mathbf{1}_{\Delta_{n,\alpha,\beta}}\Big] &\leq \mathbb{E}\left[C(\theta)\inf_{m\in\mathscr{M}_{n,\alpha}^{(1)}}\left(\inf_{t\in S_{\boldsymbol{m}}}\|b-t\|_n^2 + \sigma^2\frac{D_{\boldsymbol{m}}}{n}\right) + \sigma^2\frac{\Sigma(\theta,q)}{n}\right] \\ &\leq C(\theta)\inf_{m\in\mathscr{M}_{n,\alpha}^{(1)}}\left(\inf_{t\in S_{\boldsymbol{m}}}\|b-t\|_\mu^2 + \sigma^2\frac{D_{\boldsymbol{m}}}{n}\right) + \sigma^2\frac{\Sigma(\theta,q)}{n}.\end{aligned}$$

For the second term, we have:

$$\|b-\hat{b}_{\hat{\boldsymbol{m}}_1}\|_n^2 \mathbf{1}_{\Delta_{n,\alpha,\beta}^c} \leq 2\|b\|_n^2 \mathbf{1}_{\Delta_{n,\alpha,\beta}^c} + 2\|\hat{b}_{\hat{\boldsymbol{m}}_1}\|_n^2 \mathbf{1}_{\Delta_{n,\alpha,\beta}^c}.$$

Using Hölder's inequality with $r, r' \in (1,\infty)$ such that $\frac{1}{r}+\frac{1}{r'}=1$, we obtain:

$$\mathbb{E}\Big[\|b\|_n^2 \mathbf{1}_{\Delta_{n,\alpha,\beta}^c}\Big] \leq \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n b(\boldsymbol{X}_i)^2\right)^r\right]^{1/r}\mathbb{P}\Big[\Delta_{n,\alpha,\beta}^c\Big]^{1/r'} \leq \|b\|_{\mathrm{L}^{2r}(\mu)}^2\,\mathbb{P}\Big[\Delta_{n,\alpha,\beta}^c\Big]^{1/r'},$$

and if $b \in \mathrm{L}^\infty(\mu)$, the inequality also holds for $r=\infty$ and $r'=1$. Since $\hat{b}_{\hat{\boldsymbol{m}}_1}$ is the empirical projection of $\mathbf{Y}$ on $S_{\hat{\boldsymbol{m}}_1}$, we have $\|\hat{b}_{\hat{\boldsymbol{m}}_1}\|_n^2 \leq \|\mathbf{Y}\|_n^2$. Hence, we get:

$$\begin{aligned}\mathbb{E}\Big[\|\hat{b}_{\hat{\boldsymbol{m}}_1}\|_n^2 \mathbf{1}_{\Delta_{n,\alpha,\beta}^c}\Big] \leq \mathbb{E}\Big[\|\mathbf{Y}\|_n^2 \mathbf{1}_{\Delta_{n,\alpha,\beta}^c}\Big] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n b(\boldsymbol{X}_i)^2 \mathbf{1}_{\Delta_{n,\alpha,\beta}^c}\right] + \sigma^2\,\mathbb{P}\Big[\Delta_{n,\alpha,\beta}^c\Big] \\ &\leq \|b\|_{\mathrm{L}^{2r}(\mu)}^2\,\mathbb{P}\Big[\Delta_{n,\alpha,\beta}^c\Big]^{\frac{1}{r'}} + \sigma^2\,\mathbb{P}\Big[\Delta_{n,\alpha,\beta}^c\Big]. \quad (20)\end{aligned}$$

To conclude, we give an upper bound on $\mathbb{P}\Big[\Delta_{n,\alpha,\beta}^c\Big]$:

$$\begin{aligned}\mathbb{P}\Big[\Delta_{n,\alpha,\beta}^c\Big] &= \mathbb{P}\Big[\exists \boldsymbol{m}\in\mathbb{N}_+^p,\ \boldsymbol{m}\in\mathscr{M}_{n,\alpha}^{(1)} \text{ and } \boldsymbol{m}\notin\widehat{\mathscr{M}}_{n,\beta}^{(1)}\Big] \\ &\leq \sum_{\boldsymbol{m}\in\mathscr{M}_{n,\alpha}^{(1)}}\mathbb{P}\Big[\boldsymbol{m}\in\mathscr{M}_{n,\alpha}^{(1)} \text{ and } \boldsymbol{m}\notin\widehat{\mathscr{M}}_{n,\beta}^{(1)}\Big].\end{aligned}$$

Using the following inclusion of events:

$$\begin{aligned}&\Big\{\boldsymbol{m}\in\mathscr{M}_{n,\alpha}^{(1)} \text{ and } \boldsymbol{m}\notin\widehat{\mathscr{M}}_{n,\beta}^{(1)}\Big\} \\ &\subset \left\{K_\nu^\infty(\boldsymbol{m})\left(\|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}\vee 1\right) \leq \alpha\frac{n}{\log n}\right\} \cap \left\{K_\nu^\infty(\boldsymbol{m})\left(\|\widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}\vee 1\right) \geq \beta\frac{n}{\log n}\right\} \\ &\subset \left\{\frac{\|\widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}}{\|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}} \geq \frac{\beta}{\alpha}\right\} = \left\{\lambda_{\min}(\widehat{\mathbf{G}}_{\boldsymbol{m}}) \leq \frac{\alpha}{\beta}\lambda_{\min}(\mathbf{G}_{\boldsymbol{m}})\right\},\end{aligned}$$

we get:

$$\mathbb{P}\Big[\Delta_{n,\alpha,\beta}^{c}\Big] \leq \sum_{\boldsymbol{m}\in\mathscr{M}_{n,\alpha}^{(1)}} \mathbb{P}\Big[\lambda_{\min}(\widehat{\mathbf{G}}_{\boldsymbol{m}}) \leq \frac{\alpha}{\beta}\lambda_{\min}(\mathbf{G}_{\boldsymbol{m}})\Big]. \tag{21}$$

Using Lemma 3 with the inequality $K_V^{\infty}(\boldsymbol{m})\|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}} \leq \alpha\frac{n}{\log n}$ for $\boldsymbol{m}\in\mathscr{M}_{n,\alpha}^{(1)}$, we obtain:

$$\begin{aligned}\forall \boldsymbol{m}\in\mathscr{M}_{n,\alpha}^{(1)},\ \mathbb{P}\Big[\lambda_{\min}(\widehat{\mathbf{G}}_{\boldsymbol{m}}) \leq \frac{\alpha}{\beta}\lambda_{\min}(\mathbf{G}_{\boldsymbol{m}})\Big] &\leq D_{\boldsymbol{m}}\exp\left(h(1-\tfrac{\alpha}{\beta})\frac{n}{K_V^{\infty}(\boldsymbol{m})\|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}}\right)\\ &\leq D_{\boldsymbol{m}}\, n^{-h(1-\frac{\alpha}{\beta})/\alpha}.\end{aligned}$$

Hence, we get:

$$\mathbb{P}\Big[\Delta_{n,\alpha,\beta}^{c}\Big] \leq \sum_{\boldsymbol{m}\in\mathscr{M}_{n,\alpha}^{(1)}} D_{\boldsymbol{m}}\, n^{-h(1-\frac{\alpha}{\beta})/\alpha} \leq \mathrm{Card}(\mathscr{M}_{n,\alpha}^{(1)})\, n^{1-h(1-\frac{\alpha}{\beta})/\alpha}.$$

Using Proposition 4 in appendix, we obtain:

$$\mathbb{P}\Big[\Delta_{n,\alpha,\beta}^{c}\Big] \leq n^{2-h(1-\frac{\alpha}{\beta})/\alpha} H_n^{p-1} = n^{-\kappa(\alpha,\beta)} H_n^{p-1},$$

with $H_n := \sum_{k=1}^{n}\frac{1}{k}$ and $\kappa(\alpha,\beta) := \frac{h(1-\frac{\alpha}{\beta})}{\alpha} - 2$. We know that $H_n \sim \log n$, so we want a condition on α such that the $\kappa(\alpha,\beta)$ is strictly greater than r'. Let $x := \frac{\beta}{\alpha} \geq 1$, we have:

$$\begin{aligned}\kappa(\alpha,\beta) > r' &\iff h\left(1-\frac{\alpha}{\beta}\right) > (2+r')\alpha\\ &\iff 1-\frac{\alpha}{\beta}+\frac{\alpha}{\beta}\log\left(\frac{\alpha}{\beta}\right) > (2+r')\alpha\\ &\iff 1-\frac{1+\log(x)}{x} > \frac{(2+r')\beta}{x}\\ &\iff \frac{1+(2+r')\beta+\log(x)}{x} < 1. \end{aligned}\tag{22}$$

The function:

$$f_{\beta,r'}(x) := \frac{1+(2+r')\beta+\log(x)}{x},$$

is decreasing on $[1,+\infty)$, we have $f_{\beta,r'}(1) > 1$ and $f_{\beta,r'}(x) \to 0$ when $x \to +\infty$, so there exists a unique $x_{\beta,r'} \in (1,+\infty)$ such that $f_{\beta,r'}(x_{\beta,r'}) = 1$. Thus, we have:

$$(22) \iff x \in (x_{\beta,r'},+\infty) \iff \alpha \in (0,\alpha_{\beta,r'}),$$

where $\alpha_{\beta,r'} := \frac{\beta}{x_{\beta,r'}}$. Hence, if $\alpha \in (0,\alpha_{\beta,r'})$ then we have:

$$\mathbb{P}\Big[\Delta_{n,\alpha,\beta}^{c}\Big]^{1/r'} \leq n^{-\frac{\kappa(\alpha,\beta)}{r'}} H_n^{\frac{p-1}{r'}},$$

with $\frac{\kappa(\alpha,\beta)}{r'} > 1$ and $\frac{\kappa(\alpha,\beta)}{r'} \to 1$ as $\alpha \to \alpha_{\beta,r'}$. □

Remark 7 If we use the collections $\mathscr{M}_{n,\alpha}^{(2)}$ and $\widehat{\mathscr{M}}_{n,\beta}^{(2)}$ instead, we obtain the inequality (21) with α and β replaced by $\alpha' := \sqrt{\alpha}$ and $\beta' := \sqrt{\beta}$. The rest of the proof is unchanged.

Proof (Remark 4) We have $\alpha_{\beta,r'} := \frac{\beta}{x_{\beta,r'}}$ where $x_{\beta,r'}$ is the unique solution in $(1,+\infty)$ of the equation $f_{\beta,r'}(x) = 1$ with:

$$f_{\beta,r'}(x) := \frac{1+(2+r')\beta+\log x}{x}.$$

Hence, x_β satisfies the relation:

$$x_{\beta,r'} - \log x_{\beta,r'} = 1+(2+r')\beta. \tag{23}$$

Since the functions $f_{\beta,r'}$ are decreasing on $(1,+\infty)$ and since $\forall x$, $f_{\beta,r'}(x)$ is increasing with β and r', we see that $x_{\beta,r'}$ is increasing with β and r'. Thus, the limits of $x_{\beta,r'}$ when $\beta \to 0$ and $\beta \to +\infty$ exist. Using the relation (23), we obtain:

$$\lim_{\beta\to 0} x_{\beta,r'} = 1, \qquad \lim_{\beta\to+\infty} x_{\beta,r'} = +\infty, \qquad \lim_{r'\to\infty} x_{\beta,r'} = +\infty,$$

and we have $x_{\beta,r'} \sim (2+r')\beta$ when $\beta \to +\infty$. Thus, the limits of $\alpha_{\beta,r'}$ are:

$$\lim_{\beta\to 0} \alpha_{\beta,r'} = 0, \qquad \lim_{\beta\to+\infty} \alpha_{\beta,r'} = \frac{1}{2+r'}, \qquad \lim_{r'\to+\infty} \alpha_{\beta,r'} = 0.$$

Since $x_{\beta,r'}$ is increasing with r', we see that $\alpha_{\beta,r'}$ is deacreasing with r'. Finally, using the relation (23) again, we have:

$$\alpha_{\beta,r'} = \frac{\beta}{x_{\beta,r'}} = \frac{1}{2+r'}\left(1 - \frac{1}{x_{\beta,r'}} - \frac{\log x_{\beta,r'}}{x_{\beta,r'}}\right).$$

It is easy to see that the function $x \mapsto 1 - \frac{1}{x} - \frac{\log x}{x}$ is increasing on $[1,+\infty)$ so $\alpha_{\beta,r'}$ is also increasing with β. □

7.4 Proof of Theorem 3

Before proving Theorem 3, we need some preliminary results.

Lemma 4 *For all $x > 0$ and all $\boldsymbol{m} \in \mathbb{N}_+^p$ we have:*

$$\begin{aligned}\mathbb{P}\Big[\|\widehat{\mathbf{G}}_{\boldsymbol{m}} - \mathbf{G}_{\boldsymbol{m}}\|_{\mathrm{op}} \geq x\Big] &\leq D_{\boldsymbol{m}} \exp\left(\frac{-nx^2/2}{K_\nu^\infty(\boldsymbol{m})\big(\|\mathbf{G}_{\boldsymbol{m}}\|_{\mathrm{op}} + \frac{2}{3}x\big)}\right) \\ &\leq D_{\boldsymbol{m}} \exp\left(\frac{-nx^2/2}{K_\nu^\infty(\boldsymbol{m})\big(\|\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\|_\infty + \frac{2}{3}x\big)}\right).\end{aligned}$$

Proof The set $\{\varphi_{\boldsymbol{j}} : \boldsymbol{j} \leq \boldsymbol{m}-\mathbf{1}\}$ has cardinality $D_{\boldsymbol{m}}$ so let $\{\phi_1,\dots,\phi_{D_{\boldsymbol{m}}}\}$ be its elements. We define the matrix $\widehat{\mathbf{H}}_{\boldsymbol{m}}$ as:

$$\forall j,k \in \{1,\dots,D_{\boldsymbol{m}}\}, \quad \left[\widehat{\mathbf{H}}_{\boldsymbol{m}}\right]_{j,k} \coloneqq \langle \phi_j, \phi_k\rangle_n,$$

and we denote its expectation $\mathbf{H}_{\boldsymbol{m}}$, of which the components are $\langle \phi_j, \phi_k\rangle_\mu$. In other words, we have reshaped the hypermatrices $\widehat{\mathbf{G}}_{\boldsymbol{m}}$ and $\mathbf{G}_{\boldsymbol{m}}$ into $D_{\boldsymbol{m}} \times D_{\boldsymbol{m}}$ matrices. Moreover, this operation preserves the operator norm:

$$\|\mathbf{G}_{\boldsymbol{m}}\|_{\mathrm{op}} = \|\mathbf{H}_{\boldsymbol{m}}\|_{\mathrm{op}}.$$

Indeed, let $d \coloneqq D_{\boldsymbol{m}}$, we have:

$$\|\mathbf{G}_{\boldsymbol{m}}\|_{\mathrm{op}} = \sup_{\substack{\mathbf{a}\in\mathbb{R}^{\boldsymbol{m}} \\ \|\mathbf{a}\|_{\mathbb{R}^{\boldsymbol{m}}}=1}} \|\mathbf{G}_{\boldsymbol{m}} \times_p \mathbf{a}\|^2_{\mathbb{R}^{\boldsymbol{m}}} = \sup_{\substack{\mathbf{a}\in\mathbb{R}^{\boldsymbol{m}} \\ \|\mathbf{a}\|_{\mathbb{R}^{\boldsymbol{m}}}=1}} \sum_{\boldsymbol{\ell}\leq \boldsymbol{m}-\mathbf{1}} \left(\sum_{\boldsymbol{k}\leq \boldsymbol{m}-\mathbf{1}} \langle \varphi_{\boldsymbol{\ell}}, \varphi_{\boldsymbol{k}}\rangle a_{\boldsymbol{k}} \right)^2,$$

$$\|\mathbf{H}_{\boldsymbol{m}}\|_{\mathrm{op}} = \sup_{\substack{\mathbf{a}\in\mathbb{R}^{d} \\ \|\mathbf{a}\|_{\mathbb{R}^{d}}=1}} \|\mathbf{H}_{\boldsymbol{m}}\mathbf{a}\|^2_{\mathbb{R}^{d}} = \sup_{\substack{\mathbf{a}\in\mathbb{R}^{d} \\ \|\mathbf{a}\|_{\mathbb{R}^{d}}=1}} \sum_{j=1}^{d} \left(\sum_{i=1}^{d} \langle \psi_j, \psi_i\rangle a_i \right)^2.$$

Since the sets $\{\varphi_{\boldsymbol{j}} : \boldsymbol{j} \leq \boldsymbol{m}-\mathbf{1}\}$ and $\{\phi_1,\dots,\phi_d\}$ are equal, these two quantities are also equal. Hence we have:

$$\|\widehat{\mathbf{G}}_{\boldsymbol{m}} - \mathbf{G}_{\boldsymbol{m}}\|_{\mathrm{op}} = \|\widehat{\mathbf{H}}_{\boldsymbol{m}} - \mathbf{H}_{\boldsymbol{m}}\|_{\mathrm{op}},$$

so we work on $\widehat{\mathbf{H}}_{\boldsymbol{m}}$ and $\mathbf{H}_{\boldsymbol{m}}$ from now on. We write:

$$\widehat{\mathbf{H}}_{\boldsymbol{m}} - \mathbf{H}_{\boldsymbol{m}} = \sum_{i=1}^{n} \mathbf{Z}_i, \quad \mathbf{Z}_i \coloneqq \frac{1}{n}\left(\mathbf{V}_i\mathbf{V}_i^\top - \mathbb{E}\left[\mathbf{V}_i\mathbf{V}_i^\top\right]\right), \quad \mathbf{V}_i \coloneqq \begin{bmatrix} \phi_1(\boldsymbol{X}_i) \\ \vdots \\ \phi_{D_{\boldsymbol{m}}}(\boldsymbol{X}_i) \end{bmatrix},$$

and we use the Matrix Bernstein bound (Theorem 6 in appendix).

1. Bound on $\|\mathbf{Z}_i\|_{\mathrm{op}}$:

$$\frac{1}{n}\|\mathbf{V}_i\mathbf{V}_i^\top\|_{\mathrm{op}} = \frac{1}{n}\|\mathbf{V}_i\|^2 = \frac{1}{n}\sum_{j=1}^{D_{\boldsymbol{m}}} \phi_j(\boldsymbol{X}_i)^2 \leq \frac{K_V^\infty(\boldsymbol{m})}{n},$$

   where the last inequality comes from Lemma 2. Hence, $\|\mathbf{Z}_i\|_{\mathrm{op}} \leq R$, with $R \coloneqq \frac{K_V^\infty(\boldsymbol{m})}{n}$.
2. Bound on $\|\sum_{i=1}^n \mathbb{E}\left[\mathbf{Z}_i^2\right]\|_{\mathrm{op}}$:

$$\left\|\sum_{i=1}^{n} \mathbb{E}\left[\mathbf{Z}_i^2\right]\right\|_{\mathrm{op}} = \sup_{\|\mathbf{a}\|=1} \sum_{i=1}^{n} \mathbb{E}\left[\|\mathbf{Z}_i\mathbf{a}\|^2\right] = \sup_{\|\mathbf{a}\|=1} \sum_{i=1}^{n}\sum_{j=1}^{D_{\boldsymbol{m}}} \mathbb{E}\left[(\mathbf{Z}_i\mathbf{a})_j^2\right]$$
$$= \sup_{\|\mathbf{a}\|=1} \sum_{i=1}^{n}\sum_{j=1}^{D_{\boldsymbol{m}}} \mathrm{Var}\left[(\mathbf{Z}_i\mathbf{a})_j\right],$$

since $\mathbb{E}\mathbf{Z}_i = \mathbf{0}$. We compute the variance:

$$\begin{aligned}\operatorname{Var}\big[(\mathbf{Z}_i\,\mathbf{a})_j\big] = \operatorname{Var}\left[\frac{1}{n}\phi_j(\boldsymbol{X}_i)\sum_{k=1}^{D_{\boldsymbol{m}}}\phi_k(\boldsymbol{X}_i)\,a_k\right] &\le \frac{1}{n^2}\mathbb{E}\left[\left(\phi_j(\boldsymbol{X}_i)\sum_{k=1}^{D_{\boldsymbol{m}}}\phi_k(\boldsymbol{X}_i)\,a_k\right)^2\right]\\ &= \frac{1}{n}\mathbb{E}\big[\phi_j(\boldsymbol{X}_i)^2\,t_{\mathbf{a}}(\boldsymbol{X}_i)^2\big],\end{aligned}$$

where $t_{\mathbf{a}} \coloneqq \sum_{k=1}^{D_{\boldsymbol{m}}} a_k\,\phi_k$. Using Lemmas 1 and 2 yields:

$$\begin{aligned}\sum_{i=1}^{n}\sum_{j=1}^{D_{\boldsymbol{m}}}\operatorname{Var}\big[(\mathbf{Z}_i\,\mathbf{a})_j\big] \le \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left[\sum_{j=1}^{D_{\boldsymbol{m}}}\phi_j(\boldsymbol{X}_i)^2\,t_{\mathbf{a}}(\boldsymbol{X}_i)^2\right] &\le \frac{1}{n}K_\nu^\infty(\boldsymbol{m})\,\|t_{\mathbf{a}}\|_\mu^2\\ &\le \frac{1}{n}K_\nu^\infty(\boldsymbol{m})\,K_\nu^\mu(\boldsymbol{m})\,\|t_{\mathbf{a}}\|_\nu^2\\ &= \frac{1}{n}K_\nu^\infty(\boldsymbol{m})\,\|\mathbf{G}_{\boldsymbol{m}}\|_{\mathrm{op}}\,\|\mathbf{a}\|^2.\end{aligned}$$

Hence, $\|\sum_{i=1}^n \mathbb{E}\big[\mathbf{Z}_i^2\big]\|_{\mathrm{op}} \le \frac{1}{n}K_\nu^\infty(\boldsymbol{m})\|\mathbf{G}_{\boldsymbol{m}}\|_{\mathrm{op}} =: v$.

Applying Theorem 6 yields:

$$\mathbb{P}\Big[\|\widehat{\mathbf{H}}_{\boldsymbol{m}} - \mathbf{H}_{\boldsymbol{m}}\|_{\mathrm{op}} \ge x\Big] \le D_{\boldsymbol{m}}\exp\left(-\frac{nx^2/2}{K_\nu^\infty(\boldsymbol{m})\big(\|\mathbf{G}_{\boldsymbol{m}}\|_{\mathrm{op}} + \frac{2}{3}x\big)}\right),$$

which is the first inequality of Lemma 4. The second inequality follows from the following upper bound on $\|\mathbf{G}_{\boldsymbol{m}}\|_{\mathrm{op}}$:

$$\|\mathbf{G}_{\boldsymbol{m}}\|_{\mathrm{op}} = \sup_{t\in S_{\boldsymbol{m}}\setminus\{0\}}\frac{\|t\|_\mu^2}{\|t\|_\nu^2} \le \left\|\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right\|_\infty.$$

□

In order to prove Theorem 3, let us consider the events:

$$\Lambda_n^{(\iota)}(\beta,\gamma) \coloneqq \Big\{\widehat{\mathscr{M}}_{n,\beta}^{(\iota)} \subset \mathscr{M}_{n,\gamma}^{(\iota)}\Big\}, \qquad \widetilde{\Omega}_n^{(\iota)}(\delta,\gamma) \coloneqq \bigcap_{\boldsymbol{m}\in\mathscr{M}_{n,\gamma}^{(\iota)}}\Omega_{\boldsymbol{m}}(\delta), \qquad \iota\in\{1,2\}, \tag{24}$$

where $\Omega_{\boldsymbol{m}}(\delta)$ is defined by (2).

Lemma 5 *For $\iota\in\{1,2\}$, we have for all $\delta\in(0,1)$ and all $\gamma>0$:*

$$\mathbb{P}\Big[\widetilde{\Omega}_n^{(\iota)}(\delta,\gamma)^c\Big] \le n^{-\frac{h(\delta)}{\gamma}+2}H_n^{p-1},$$

where $H_n \coloneqq \sum_{k=1}^n \frac{1}{k}$ is the n-th harmonic number.

Proof We use Proposition 3 with Remark 1:

$$\begin{aligned}\mathbb{P}\left[\widetilde{\Omega}_n^{(\iota)}(\delta,\gamma)^c\right] &\le \sum_{\boldsymbol{m}\in\mathscr{M}_{n,\gamma}^{(\iota)}} \mathbb{P}[\Omega_{\boldsymbol{m}}(\delta)^c] \le \sum_{\boldsymbol{m}\in\mathscr{M}_{n,\gamma}^{(\iota)}} D_{\boldsymbol{m}} \exp\left(-h(\delta)\frac{n}{K_\mu^\infty(\boldsymbol{m})}\right) \\ &\le \sum_{\boldsymbol{m}\in\mathscr{M}_{n,\gamma}^{(\iota)}} D_{\boldsymbol{m}} \exp\left(-h(\delta)\frac{n}{K_\nu^\infty(\boldsymbol{m})\|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}}\right) \\ &\le \sum_{\boldsymbol{m}\in\mathscr{M}_{n,\gamma}^{(\iota)}} D_{\boldsymbol{m}} n^{-\frac{h(\delta)}{\gamma}} \le n^{-\frac{h(\delta)}{\gamma}+2} H_n^{p-1},\end{aligned}$$

where the last inequality comes from Proposition 4. □

Lemma 6 (Compact case) *We have for all $\gamma > \beta > 0$:*

$$\mathbb{P}\left[\Lambda_n^{(1)}(\beta,\gamma)^c\right] \le n^{-h(1-\frac{\gamma}{\beta})\frac{f_0}{\beta}+1} H_n^{p-1},$$

where $h(\delta) = \delta + (1-\delta)\log(1-\delta)$, $f_0 > 0$ is such that $\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(x) \ge f_0$ for all $x \in A$ and $H_n := \sum_{k=1}^n \frac{1}{k}$.

Proof We start with a union bound:

$$\begin{aligned}\mathbb{P}\left[\Lambda_n^{(1)}(\beta,\gamma)^c\right] &= \mathbb{P}\left[\exists \boldsymbol{m}\in\mathbb{N}_+^p,\ \boldsymbol{m}\in\widehat{\mathscr{M}}_{n,\beta}^{(1)} \text{ and } \boldsymbol{m}\notin\mathscr{M}_{n,\gamma}^{(1)}\right] \\ &\le \sum_{\substack{\boldsymbol{m}\in\mathbb{N}_+^p \\ K_\nu^\infty(\boldsymbol{m})\le\beta\frac{n}{\log n}}} \mathbb{P}\left[\boldsymbol{m}\in\widehat{\mathscr{M}}_{n,\beta}^{(1)} \text{ and } \boldsymbol{m}\notin\mathscr{M}_{n,\gamma}^{(1)}\right].\end{aligned}$$

We have the following inclusion of events:

$$\begin{aligned}&\left\{\boldsymbol{m}\in\widehat{\mathscr{M}}_{n,\beta}^{(1)} \text{ and } \boldsymbol{m}\notin\mathscr{M}_{n,\gamma}^{(1)}\right\} \\ &\subset \left\{K_\nu^\infty(\boldsymbol{m})\left(\|\widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}\vee 1\right) \le \beta\frac{n}{\log n}\right\} \cap \left\{K_\nu^\infty(\boldsymbol{m})\left(\|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}\vee 1\right) \ge \gamma\frac{n}{\log n}\right\} \\ &\subset \left\{\frac{\|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}}{\|\widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}} \ge \frac{\gamma}{\beta}\right\} \subset \left\{\lambda_{\min}(\widehat{\mathbf{G}}_{\boldsymbol{m}}) \ge \frac{\gamma}{\beta}\lambda_{\min}(\mathbf{G}_{\boldsymbol{m}})\right\},\end{aligned}$$

hence we obtain:

$$\mathbb{P}\left[\Lambda_n^{(1)}(\beta,\gamma)^c\right] \le \sum_{\substack{\boldsymbol{m}\in\mathbb{N}_+^p \\ K_\nu^\infty(\boldsymbol{m})\le\beta\frac{n}{\log n}}} \mathbb{P}\left[\lambda_{\min}(\widehat{\mathbf{G}}_{\boldsymbol{m}}) \ge \frac{\gamma}{\beta}\lambda_{\min}(\mathbf{G}_{\boldsymbol{m}})\right].$$

We apply inequality (30) of Theorem 5 with $R = \frac{1}{n}K_\nu^\infty(\boldsymbol{m})$:

$$\mathbb{P}\left[\lambda_{\min}(\widehat{\mathbf{G}}_{\boldsymbol{m}}) \ge \frac{\gamma}{\beta}\lambda_{\min}(\mathbf{G}_{\boldsymbol{m}})\right] \le \exp\left(-h\left(1-\frac{\gamma}{\beta}\right)\frac{n}{K_\nu^\infty(\boldsymbol{m})\|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}}\right).$$

In the compact case, we have $\|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}} \leq \frac{1}{f_0}$, see (3). Using Proposition 4, we obtain:

$$\mathbb{P}\Big[\Lambda_n^{(1)}(\beta,\gamma)^c\Big] \leq \sum_{\substack{\boldsymbol{m}\in\mathbb{N}_+^p \\ K_V^\infty(\boldsymbol{m})\leq\beta\frac{n}{\log n}}} n^{-h(1-\frac{\gamma}{\beta})\frac{f_0}{\beta}} \leq n^{-h(1-\frac{\gamma}{\beta})\frac{f_0}{\beta}+1} H_n^{p-1}.$$

□

Lemma 7 (General case) *We have for all $\gamma > \beta > 0$:*

$$\mathbb{P}\Big[\Lambda_n^{(2)}(\beta,\gamma)^c\Big] \leq n^{-C(\beta,\gamma)\frac{B}{2\beta}+2} H_n^{p-1},$$

where $C(\beta,\gamma) \coloneqq \left(1-\sqrt{\beta/\gamma}\right)^2$, $B \coloneqq \left(\|\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\|_\infty + \frac{2}{3}\right)^{-1}$ and $H_n \coloneqq \sum_{k=1}^n \frac{1}{k}$.

Proof We start with a union bound:

$$\begin{aligned}\mathbb{P}\Big[\Lambda_n^{(2)}(\beta,\gamma)^c\Big] &= \mathbb{P}\Big[\exists \boldsymbol{m}\in\mathbb{N}_+^p, \boldsymbol{m}\in\widehat{\mathscr{M}}_{n,\beta}^{(2)} \text{ and } \boldsymbol{m}\notin\mathscr{M}_{n,\gamma}^{(2)}\Big] \\ &\leq \sum_{\substack{\boldsymbol{m}\in\mathbb{N}_+^p \\ K_V^\infty(\boldsymbol{m})\leq\beta\frac{n}{\log n}}} \mathbb{P}\Big[\boldsymbol{m}\in\widehat{\mathscr{M}}_{n,\beta}^{(2)} \text{ and } \boldsymbol{m}\notin\mathscr{M}_{n,\gamma}^{(2)}\Big].\end{aligned}$$

We have the following inclusion of events:

$$\begin{aligned}&\Big\{\boldsymbol{m}\in\widehat{\mathscr{M}}_{n,\beta}^{(2)} \text{ and } \boldsymbol{m}\notin\mathscr{M}_{n,\gamma}^{(2)}\Big\} \\ &\subset \left\{K_V^\infty(\boldsymbol{m})\left(\|\widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}^2\vee 1\right)\leq\beta\frac{n}{\log n}\right\}\cap\left\{K_V^\infty(\boldsymbol{m})\left(\|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}^2\vee 1\right)\geq\gamma\frac{n}{\log n}\right\} \\ &\subset \left\{K_V^\infty(\boldsymbol{m})\|\widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}^2\leq\beta\frac{n}{\log n}\right\}\cap\left\{K_V^\infty(\boldsymbol{m})\|\widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1}-\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}^2\geq\left(\sqrt{\gamma}-\sqrt{\beta}\right)^2\frac{n}{\log n}\right\} \\ &\subset \left\{\|\widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}^2\leq\frac{\beta}{K_V^\infty(\boldsymbol{m})}\frac{n}{\log n}\right\}\cap\left\{\|\widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1}-\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}\geq\left(\sqrt{\frac{\gamma}{\beta}}-1\right)\|\widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}\right\}.\end{aligned}$$

Let $\eta \coloneqq \sqrt{\frac{\gamma}{\beta}}-1$ and let $\varepsilon\in(0,1)$. We consider the following decomposition:

$$\left\{\|\widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1}-\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}\geq\eta\|\widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}\right\} = E_1\cup E_2,$$

with:

$$\begin{aligned}E_1 &\coloneqq \left\{\|\widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1}-\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}\geq\eta\|\widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}\right\}\cap\left\{\left\|\widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1}(\mathbf{G}_{\boldsymbol{m}}-\widehat{\mathbf{G}}_{\boldsymbol{m}})\right\|_{\mathrm{op}}<\varepsilon\right\}, \\ E_2 &\coloneqq \left\{\|\widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1}-\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}\geq\eta\|\widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}\right\}\cap\left\{\left\|\widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1}(\mathbf{G}_{\boldsymbol{m}}-\widehat{\mathbf{G}}_{\boldsymbol{m}})\right\|_{\mathrm{op}}\geq\varepsilon\right\}.\end{aligned}$$

– For E_1, we apply Lemma 9 with $\mathbf{A} \coloneqq \widehat{\mathbf{G}}_{\boldsymbol{m}}$ and $\mathbf{B} \coloneqq \mathbf{G}_{\boldsymbol{m}}-\widehat{\mathbf{G}}_{\boldsymbol{m}}$:

$$\begin{aligned}E_1 &\subset \left\{\frac{\|\widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}^2\|\widehat{\mathbf{G}}_{\boldsymbol{m}}-\mathbf{G}_{\boldsymbol{m}}\|_{\mathrm{op}}}{1-\left\|\widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1}(\mathbf{G}_{\boldsymbol{m}}-\widehat{\mathbf{G}}_{\boldsymbol{m}})\right\|_{\mathrm{op}}}\geq\eta\|\widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}\right\}\cap\left\{\left\|\widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1}(\mathbf{G}_{\boldsymbol{m}}-\widehat{\mathbf{G}}_{\boldsymbol{m}})\right\|_{\mathrm{op}}<\varepsilon\right\} \\ &\subset \left\{\|\widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}\|\widehat{\mathbf{G}}_{\boldsymbol{m}}-\mathbf{G}_{\boldsymbol{m}}\|_{\mathrm{op}}\geq(1-\varepsilon)\eta\right\}.\end{aligned}$$

– For E_2, we have directly:

$$E_2 \subset \left\{ \left\| \widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1} (\mathbf{G}_{\boldsymbol{m}} - \widehat{\mathbf{G}}_{\boldsymbol{m}}) \right\|_{\mathrm{op}} \geq \varepsilon \right\} \subset \left\{ \|\widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1}\| \|\mathbf{G}_{\boldsymbol{m}} - \widehat{\mathbf{G}}_{\boldsymbol{m}}\|_{\mathrm{op}} \geq \varepsilon \right\}.$$

Thus, we obtain:

$$\forall \varepsilon \in (0,1), \qquad E_1 \cup E_2 \subset \left\{ \|\widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}} \|\mathbf{G}_{\boldsymbol{m}} - \widehat{\mathbf{G}}_{\boldsymbol{m}}\|_{\mathrm{op}} \geq (1-\varepsilon)\eta \wedge \varepsilon \right\}.$$

We now choose ε maximizing $(1-\varepsilon)\eta \wedge \varepsilon$. This maximum is achieved when $\varepsilon = (1-\varepsilon)\eta$, that is:

$$\varepsilon = \frac{\eta}{1-\eta} = 1 - \sqrt{\beta/\gamma} =: c(\beta,\gamma) \in (0,1).$$

Thus, we obtain:

$$\begin{aligned}
&\mathbb{P}\left[\Lambda_n^{(2)}(\beta,\gamma)^c \right] \\
&\leq \sum_{\substack{\boldsymbol{m} \in \mathbb{N}_+^p \\ K_\nu^\infty(\boldsymbol{m}) \leq \beta \frac{n}{\log n}}} \mathbb{P}\left[\left\{ \|\widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}^2 \leq \frac{\beta}{K_\nu^\infty(\boldsymbol{m})} \frac{n}{\log n} \right\} \cap \left\{ \|\widehat{\mathbf{G}}_{\boldsymbol{m}} - \mathbf{G}_{\boldsymbol{m}}\|_{\mathrm{op}} \geq \frac{c(\beta,\gamma)}{\|\widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}} \right\} \right] \\
&\leq \sum_{\substack{\boldsymbol{m} \in \mathbb{N}_+^p \\ K_\nu^\infty(\boldsymbol{m}) \leq \beta \frac{n}{\log n}}} \mathbb{P}\left[\|\widehat{\mathbf{G}}_{\boldsymbol{m}} - \mathbf{G}_{\boldsymbol{m}}\|_{\mathrm{op}} \geq c(\beta,\gamma) \sqrt{\frac{K_\nu^\infty(\boldsymbol{m})}{\beta} \frac{\log n}{n}} \right].
\end{aligned}$$

Let $x := c(\beta,\gamma)\sqrt{\frac{K_\nu^\infty(\boldsymbol{m})}{\beta}\frac{\log n}{n}}$ and notice that $x \leq 1$ if $K_\nu^\infty(\boldsymbol{m}) \leq \beta \frac{n}{\log n}$. We apply Lemma 4 and Proposition 4:

$$\begin{aligned}
&\mathbb{P}\left[\Lambda_n^{(2)}(\beta,\gamma)^c \right] \\
&\leq \sum_{\substack{\boldsymbol{m} \in \mathbb{N}_+^p \\ K_\nu^\infty(\boldsymbol{m}) \leq \beta \frac{n}{\log n}}} D_{\boldsymbol{m}} \exp\left(-\frac{n}{2} c^2(\beta,\gamma) \frac{K_\nu^\infty(\boldsymbol{m})}{\beta} \frac{\log n}{n} \left[K_\nu^\infty(\boldsymbol{m}) \left(\left\| \frac{\mathrm{d}\mu}{\mathrm{d}\nu} \right\|_\infty + \frac{2}{3} x \right) \right]^{-1} \right) \\
&\leq \sum_{\substack{\boldsymbol{m} \in \mathbb{N}_+^p \\ K_\nu^\infty(\boldsymbol{m}) \leq \beta \frac{n}{\log n}}} D_{\boldsymbol{m}}\, n^{-c^2(\beta,\gamma)\frac{B}{2\beta}} \leq n^{-c^2(\beta,\gamma)\frac{B}{2\beta}+2} H_n^{p-1},
\end{aligned}$$

where $B := (\|\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\|_\infty + \frac{2}{3})^{-1}$. □

Now we can prove Theorem 3.

Proof (Theorem 3) Let $\delta \in (0,1)$ and $\gamma > \beta$ be constants to be chosen later. Let us introduce the event $\Xi_n^{(\iota)}(\beta,\gamma,\delta) := \Lambda_n^{(\iota)}(\beta,\gamma) \cap \widetilde{\Omega}_n^{(\iota)}(\delta,\gamma)$ where $\Lambda_n^{(\iota)}(\beta,\gamma)$ and

$\widetilde{\Omega}_n^{(\iota)}(\delta,\gamma)$ are defined by (24). On the event $\Xi_n^{(\iota)}(\beta,\gamma,\delta)$, for all $\boldsymbol{m}\in\mathscr{M}_{n,\alpha}^{(\iota)}$, for all $t\in S_{\boldsymbol{m}}$ we have:

$$\begin{aligned}\|b-\hat{b}_{\hat{\boldsymbol{m}}_\iota}\|_\mu^2 &\le 2\|b-t\|_\mu^2+2\|\hat{b}_{\hat{\boldsymbol{m}}_\iota}-t\|_\mu^2\\ &\le 2\|b-t\|_\mu^2+\frac{2}{1-\delta}\|\hat{b}_{\hat{\boldsymbol{m}}_\iota}-t\|_n^2\\ &\le 2\|b-t\|_\mu^2+\frac{4}{1-\delta}\|b-t\|_n^2+\frac{4}{1-\delta}\|b-\hat{b}_{\hat{\boldsymbol{m}}_\iota}\|_n^2.\end{aligned}$$

Taking the expectation yields for all $t\in S_{\boldsymbol{m}}$:

$$\mathbb{E}\Big[\|b-\hat{b}_{\hat{\boldsymbol{m}}_\iota}\|_\mu^2\mathbf{1}_{\Xi_n^{(\iota)}(\beta,\gamma,\delta)}\Big]\le\left(2+\frac{4}{1-\delta}\right)\|b-t\|_\mu^2+\frac{4}{1-\delta}\mathbb{E}\|b-\hat{b}_{\hat{\boldsymbol{m}}_\iota}\|_n^2. \tag{25}$$

On the event $\Xi_n^{(\iota)}(\beta,\gamma,\delta)^c$, we use inequalities (12) and (16):

$$\begin{aligned}\|b-\hat{b}_{\hat{\boldsymbol{m}}_\iota}\|_\mu^2\le 2\|b\|_\mu^2+2\|\hat{b}_{\hat{\boldsymbol{m}}_\iota}\|_\mu^2 &\le 2\|b\|_\mu^2+2K_n^\mu(\hat{\boldsymbol{m}}_\iota)\|\mathbf{Y}\|_n^2\\ &\le 2\|b\|_\mu^2+2K_\nu^\infty(\hat{\boldsymbol{m}}_\iota)\|\widehat{\mathbf{G}}_{\hat{\boldsymbol{m}}_\iota}^{-1}\|_{\mathrm{op}}\|\mathbf{Y}\|_n^2\\ &\le 2\|b\|_\mu^2+4\beta\frac{n}{\log n}\|\mathbf{Y}\|_n^2.\end{aligned}$$

Using Hölder's inequality as we did in (20), we obtain:

$$\begin{aligned}&\mathbb{E}\Big[\|b-\hat{b}_{\hat{\boldsymbol{m}}_\iota}\|_\mu^2\mathbf{1}_{\Xi_n^{(\iota)}(\beta,\gamma,\delta)^c}\Big]\le 2\|b\|_\mu^2\,\mathbb{P}\big[\Xi_n^{(\iota)}(\beta,\gamma,\delta)^c\big]\\ &+8\beta\frac{n}{\log n}\left(\|b\|_{\mathrm{L}^{2r}(\mu)}^2\,\mathbb{P}\big[\Xi_n^{(\iota)}(\beta,\gamma,\delta)^c\big]^{1/r'}+\sigma^2\,\mathbb{P}\big[\Xi_n^{(\iota)}(\beta,\gamma,\delta)^c\big]\right).\end{aligned} \tag{26}$$

We see we need to control $\mathbb{P}\big[\Xi_n^{(\iota)}(\beta,\gamma,\delta)^c\big]$ by a term of order $n^{-2r'}$.

We have decomposed the risk as the sum of (25) and (26). We give different upper bounds on these two terms depending on whether we are in the compact case or the general case.

• *Compact case*. In equation (25), we apply Theorem 2: for all $\alpha\in(0,\alpha_{\beta,r'})$ we have:

$$\begin{aligned}&\mathbb{E}\Big[\|b-\hat{b}_{\hat{\boldsymbol{m}}_1}\|_\mu^2\mathbf{1}_{\Xi_n^{(1)}(\beta,\gamma,\delta)}\Big]\\ &\le\left(2+\frac{4}{1-\delta}\big(1+C(\theta)\big)\right)\inf_{\boldsymbol{m}\in\mathscr{M}_{n,\alpha}}\left(\inf_{t\in S_{\boldsymbol{m}}}\|b-t\|_\mu^2+\sigma^2\frac{D_{\boldsymbol{m}}}{n}\right)\\ &+\frac{4\sigma^2}{1-\delta}\frac{\Sigma(\theta,q)}{n}+\frac{4}{1-\delta}C'\big(\|b\|_{\mathrm{L}^{2r}(\mu)}^2,\sigma^2\big)\frac{(\log n)^{(p-1)/r'}}{n^{\kappa(\alpha,\beta)/r'}},\end{aligned}$$

with $\frac{\kappa(\alpha,\beta)}{r'}>1$. To obtain an upper bound on (26), we apply Lemmas 5 and 6:

$$\begin{aligned}\mathbb{P}\big[\Xi_n^{(1)}(\beta,\gamma,\delta)^c\big]&\le\mathbb{P}\Big[\widetilde{\Omega}_n^{(1)}(\delta,\gamma)^c\Big]+\mathbb{P}\Big[\Lambda_n^{(1)}(\beta,\gamma)^c\Big]\\ &\le\left(n^{-\frac{h(\delta)}{\gamma}+2}+n^{-h(1-\frac{\gamma}{\beta})\frac{f_0}{\beta}+1}\right)H_n^{p-1},\end{aligned}$$

where $h(\delta) \coloneqq \delta + (1-\delta)\log(1-\delta)$ and $H_n \coloneqq \sum_{k=1}^{n} \frac{1}{k}$. In order to obtain a term of order $n^{-2r'}$, we need:

$$\begin{cases} \dfrac{h(\delta)}{\gamma} - 2 > 2r', \\ h\left(1-\dfrac{\gamma}{\beta}\right)\dfrac{f_0}{\beta} - 1 > 2r', \end{cases} \iff \begin{cases} h(\delta) > 2(1+r')\gamma, \\ h\left(1-\dfrac{\gamma}{\beta}\right) > (2r'+1)\dfrac{\beta}{f_0}, \end{cases}$$

$$\iff \begin{cases} \delta > h^{-1}\big(2(1+r')\gamma\big), \\ \gamma < \dfrac{1}{2(1+r')}, \\ h\left(1-\dfrac{\gamma}{\beta}\right) > (2r'+1)\dfrac{\beta}{f_0}. \end{cases}$$

Let us work on the last two conditions. Let $x \coloneqq \frac{\gamma}{\beta} > 1$, the conditions on (β,γ) become:

$$\begin{cases} x < \dfrac{1}{2(1+r')\beta}, \\ x\log x - x + 1 > (2r'+1)\dfrac{\beta}{f_0}. \end{cases}$$

The function $x \mapsto x\log x - x + 1$ is increasing on $(1,+\infty)$ and ranges from 0 to $+\infty$, so there exists $x_{f_0,\beta} > 1$ such that for all $x > x_{f_0,\beta}$ we have $x\log x - x + 1 > (2r'+1)\frac{\beta}{f_0}$. Hence we need to choose x such that:

$$x_{f_0,\beta} < x < \frac{1}{(2r'+2)\beta}. \tag{27}$$

This is possible only if $x_{f_0,\beta} < \frac{1}{(2r'+2)\beta}$, that is if:

$$(2r'+1)\frac{\beta}{f_0} < \frac{1}{(2r'+2)\beta}\log\left(\frac{1}{(2r'+2)\beta}\right) - \frac{1}{(2r'+2)\beta} + 1.$$

Let us introduce a new variable $y \coloneqq (2r'+2)\beta$ and let $R = \frac{2r'+1}{2r'+2}$, the last inequality becomes:

$$\frac{R}{f_0}y + \frac{1+\log y}{y} < 1. \tag{28}$$

The function $y \mapsto \frac{R}{f_0}y + \frac{1+\log y}{y}$ is increasing on $(0,1)$, it tends to $-\infty$ at 0 and for $y = 1$ it is greater that 1, so there exists $y_{f_0,r'} \in (0,1)$ such that the condition (28) is satisfied on $(0, y_{f_0,r'})$. To sum up, we have shown that there exists $\beta_{f_0,r'} \in (0, \frac{1}{2r'+2})$ such that for every $\beta < \beta_{f_0,r'}$, the condition (27) is not empty. We choose:

$$\gamma \coloneqq \beta x, \qquad x \text{ satisfying (27)}, \qquad \delta \coloneqq \frac{1 + h^{-1}\big(2(1+r')\gamma\big)}{2},$$

and we obtain that:

$$\mathbb{E}\Big[\|b - \hat{b}_{\hat{\boldsymbol{m}}_1}\|_\mu^2 \mathbf{1}_{\Xi_n^{(1)}(\beta,\gamma,\delta)^c}\Big] \leq C''(\|b\|_{\mathrm{L}^{2r}(\mu)}, \beta, \sigma^2)\, n^{-\lambda(\beta,r,f_0)}\, (\log n)^{\frac{p-1}{r'}-1},$$

where $\lambda(\beta,r,f_0)>1$.

• *General case*. In equation (25), if we follow the proof of Theorem 2 (see Remark 7), we see that if $\alpha\in(0,\alpha^2_{\beta^{1/2},r'})$ then we have:

$$\mathbb{E}\|b-\hat{b}_{\hat{\boldsymbol{m}}_2}\|_n^2 \le C(\theta)\|b-t\|_\mu^2+\sigma^2\frac{D_{\boldsymbol{m}}}{n}+\sigma^2\frac{\Sigma(\theta,q)}{n} + C'\big(\|b\|^2_{\mathrm{L}^{2r}(\mu)},\sigma^2\big)\frac{(\log n)^{(p-1)/r'}}{n^{\kappa(\alpha^{\frac12},\beta^{\frac12})/r'}},$$

with $\frac{\kappa(\alpha^{\frac12},\beta^{\frac12})}{r'}>1$. Thus, we obtain:

$$\begin{aligned}&\mathbb{E}\Big[\|b-\hat{b}_{\hat{\boldsymbol{m}}_2}\|_\mu^2\mathbf{1}_{\Xi_n^{(2)}(\beta,\gamma,\delta)}\Big]\\ &\le\left(2+\frac{4}{1-\delta}\big(1+C(\theta)\big)\right)\inf_{\boldsymbol{m}\in\mathscr{M}^{(2)}_{n,\alpha}}\left(\inf_{t\in S_{\boldsymbol{m}}}\|b-t\|_\mu^2+\sigma^2\frac{D_{\boldsymbol{m}}}{n}\right)\\ &+\frac{4\sigma^2}{1-\delta}\frac{\Sigma(\theta,q)}{n}+\frac{4}{1-\delta}C'\big(\|b\|^2_{\mathrm{L}^{2r}(\mu)},\sigma^2\big)\frac{(\log n)^{(p-1)/r'}}{n^{\kappa(\alpha^{\frac12},\beta^{\frac12})/r'}}.\end{aligned}$$

To obtain an upper bound on (26), we apply Lemmas 5 and 7:

$$\begin{aligned}\mathbb{P}\big[\Xi_n^{(2)}(\beta,\gamma,\delta)^c\big]&\le\mathbb{P}\Big[\widetilde{\Omega}_n^{(2)}(\delta,\gamma)^c\Big]+\mathbb{P}\Big[\Lambda_n^{(2)}(\beta,\gamma)^c\Big]\\ &\le\left(n^{-\frac{h(\delta)}{\gamma}+2}+n^{-C(\beta,\gamma)\frac{B}{2\beta}+2}\right)H_n^{p-1},\end{aligned}$$

where $C(\beta,\gamma):=\left(1-\sqrt{\beta/\gamma}\right)^2$, $B:=(\|\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\|_\infty+\frac23)^{-1}$ and $H_n:=\sum_{k=1}^n\frac1k$. To obtain a term of order $n^{-2r'}$, we need:

$$\begin{cases}\dfrac{h(\delta)}{\gamma}-2>2r',\\ C(\beta,\gamma)\dfrac{B}{2\beta}-2>2r',\end{cases}\iff\begin{cases}h(\delta)>2(1+r')\gamma,\\ C(\beta,\gamma)\dfrac{B}{2}>2(1+r')\beta,\end{cases}$$

$$\iff\begin{cases}\delta>h^{-1}\big(2(1+r')\gamma\big),\\ \gamma<\dfrac{1}{2(1+r')},\\ \dfrac{C(\beta,\gamma)B}{4(1+r')}>\beta.\end{cases}$$

Let $x:=\sqrt{\beta/\gamma}\in(0,1)$, the conditions on (β,γ) can be rewritten as:

$$\begin{cases}\dfrac{\beta}{x^2}<\dfrac{1}{2(1+r')},\\ \beta<(1-x)^2\dfrac{B}{4(1+r')},\end{cases}\iff\beta<\frac{1}{2(1+r')}\left(x^2\wedge(1-x)^2\frac{B}{2}\right).$$

We choose x maximizing this bound. This maximum is achieved when $x^2 = (1-x)^2 \frac{B}{2}$, that is $x = \frac{\sqrt{B/2}}{1+\sqrt{B/2}}$. Finally we choose:

$$x := \frac{\sqrt{B/2}}{1+\sqrt{B/2}}, \qquad \gamma := \frac{\beta}{x^2}, \qquad \delta := \frac{1+h^{-1}\big(2(1+r')\gamma\big)}{2},$$

and we obtain that for all $\beta \in (0, \beta_{B,r'})$ with:

$$\beta_{B,r'} := \frac{1}{2(1+r')}\left(\frac{\sqrt{B/2}}{1+\sqrt{B/2}}\right)^2,$$

we have:

$$\mathbb{E}\Big[\|b-\hat{b}_{\hat{\boldsymbol{m}}_2}\|_\mu^2 \mathbf{1}_{\Xi_n^{(2)}(\beta,\gamma,\delta)^c}\Big] \leq C''(\|b\|_{\mathrm{L}^{2r}(\mu)},\beta,\sigma^2)\, n^{-\lambda(\beta,r,B)} (\log n)^{\frac{p-1}{r'}-1},$$

where $\lambda(\beta,r,B) > 1$. □

Acknowledgements I want to thank Fabienne Comte and Céline Duval for their helpful advice and their support of my work. I also want to thank Florence Merlevède for her help with the second inequality of the Matrix Chernoff bound. Finally, I want to thank Herb Susmann for proofreading this article.

A Linear Algebra

Lemma 8 *Let E be a Euclidean vector space and let $\ell\colon E \to \mathbb{R}^n$ be an injective linear map. For $y \in \mathbb{R}^n$, the solution of the problem:*

$$\hat{a} := \operatorname*{arg\,min}_{a\in E} \|y-\ell(a)\|_{\mathbb{R}^n}^2$$

is given by:

$$\hat{a} = \big[(\ell^*\circ\ell)^{-1}\circ\ell^*\big](y),$$

where $\ell^*\colon \mathbb{R}^n \to E$ is characterized by the relation $\langle y, \ell(a)\rangle_{\mathbb{R}^n} = \langle \ell^*(y), a\rangle_E$.

Lemma 9 *Let $\mathbf{A}$, $\mathbf{B}$ be square matrices. If $\mathbf{A}$ is invertible and $\|\mathbf{A}^{-1}\mathbf{B}\|_{\mathrm{op}} < 1$, then $\mathbf{A}+\mathbf{B}$ is invertible and it holds:*

$$\big\|(\mathbf{A}+\mathbf{B})^{-1} - \mathbf{A}^{-1}\big\|_{\mathrm{op}} \leq \frac{\|\mathbf{A}^{-1}\|_{\mathrm{op}}^2 \|\mathbf{B}\|_{\mathrm{op}}}{1-\|\mathbf{A}^{-1}\mathbf{B}\|_{\mathrm{op}}}.$$

B Concentration inequalities

You can find the proofs of the following bounds in Tropp (2012) and Gittens and Tropp (2011).

Theorem 5 (Matrix Chernoff bound) *Let $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ be independent random self-adjoint positive semi-definite matrices with dimension d, such that $\sup_k \lambda_{\max}(\mathbf{Z}_k) \leq R$ a.s. If we define:*

$$\mu_{\min} := \lambda_{\min}\left(\sum_{k=1}^n \mathbb{E}[\mathbf{Z}_k]\right),$$

then we have:

$$\forall \delta \in (0,1), \quad \mathbb{P}\left[\lambda_{\min}\left(\sum_{k=1}^{n} \mathbf{Z}_k\right) \leq (1-\delta)\mu_{\min}\right] \leq d \times \left(\frac{\mathrm{e}^{-\delta}}{(1-\delta)^{(1-\delta)}}\right)^{\mu_{\min}/R}, \tag{29}$$

$$\forall \delta > 0, \quad \mathbb{P}\left[\lambda_{\min}\left(\sum_{k=1}^{n} \mathbf{Z}_k\right) \geq (1+\delta)\mu_{\min}\right] \leq \left(\frac{\mathrm{e}^{\delta}}{(1+\delta)^{(1+\delta)}}\right)^{\mu_{\min}/R}. \tag{30}$$

Theorem 6 (Matrix Bernstein bound) *Let $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ be independent random self-adjoint positive semi-definite matrices with dimension d, such that $\mathbb{E}[\mathbf{Z}_k] = \mathbf{0}$ and that $\sup_k \lambda_{\max}(\mathbf{Z}_k) \leq R$ a.s. If $v > 0$ is such that:*

$$\left\| \sum_{k=1}^{n} \mathbb{E}\left[\mathbf{Z}_k^2\right] \right\|_{\mathrm{op}} \leq v,$$

then for all $x > 0$ we have:

$$\mathbb{P}\left[\lambda_{\max}\left(\sum_{k=1}^{n} \mathbf{Z}_i\right) \geq x\right] \leq d \times \exp\left(\frac{-x^2/2}{v + \frac{R}{3}x}\right).$$

C Combinatorics

Proposition 4 *For $n \geq 1$ and $p \geq 2$ we have:*

$$\mathrm{Card}\left\{\boldsymbol{m} \in \mathbb{N}_+^p \,\middle|\, m_1 \cdots m_p \leq n\right\} \leq n H_n^{p-1},$$

where $H_n \coloneqq \sum_{k=1}^{n} \frac{1}{k}$ is the n-th harmonic number.

Proof We compute:

$$\begin{aligned}
\mathrm{Card}\left\{\boldsymbol{m} \in \mathbb{N}_+^p \,\middle|\, D_{\boldsymbol{m}} \leq n\right\} &= \sum_{m_1=1}^{n} \cdots \sum_{m_p=1}^{n} \mathbf{1}_{m_1 \cdots m_p \leq n} \\
&= \sum_{m_1=1}^{n} \cdots \sum_{m_p=1}^{n} \mathbf{1}_{m_p \leq \frac{n}{m_1 \cdots m_{p-1}}} \\
&= \sum_{m_1=1}^{n} \cdots \sum_{m_{p-1}=1}^{n} \left\lfloor \frac{n}{m_1 \cdots m_{p-1}} \right\rfloor \\
&\leq \sum_{m_1=1}^{n} \cdots \sum_{m_{p-1}=1}^{n} \frac{n}{m_1 \cdots m_{p-1}} = n H_n^{p-1}.
\end{aligned}$$

Theorem 7 (Divisor bound) *Let $N \in \mathbb{N}_+$ and let $\mathrm{div}(N)$ be the set of divisors of N. We have for all $\varepsilon > 0$:*

$$\mathrm{Card}\big(\mathrm{div}(N)\big) = \mathrm{o}(N^{\varepsilon}).$$

As a consequence, we have for all $\varepsilon > 0$:

$$\mathrm{Card}\left\{m \in \mathbb{N}_+^p \,\middle|\, m_1 \cdots m_p = N\right\} \leq \mathrm{Card}\big(\mathrm{div}(N)\big)^p = \mathrm{o}(N^{\varepsilon}).$$

A proof of this result can be found in Tao (2008).

References

Arlot S, Massart P (2009) Data-driven Calibration of Penalties for Least-Squares Regression. *Journal of Machine Learning Research* 10(10):245–279

Baraud Y (2000) Model selection for regression on a fixed design. *Probability Theory and Related Fields* 117(4):467–493

Baraud Y (2002) Model selection for regression on a random design. *ESAIM: Probability and Statistics* 6:127–146

Barron A, Birgé L, Massart P (1999) Risk bounds for model selection via penalization. *Probability Theory and Related Fields* 113(3):301–413

Birgé L, Massart P (1998) Minimum Contrast Estimators on Sieves: Exponential Bounds and Rates of Convergence. *Bernoulli* 4(3):329–375

Cohen A, Davenport MA, Leviatan D (2013) On the Stability and Accuracy of Least Squares Approximations. *Foundations of Computational Mathematics* 13(5):819–834

Comte F, Genon-Catalot V (2018) Laguerre and Hermite bases for inverse problems. *Journal of the Korean Statistical Society* 47(3):273–296

Comte F, Genon-Catalot V (2020a) Regression function estimation as a partly inverse problem. *Annals of the Institute of Statistical Mathematics* 72(4):1023–1054

Comte F, Genon-Catalot V (2020b) Regression function estimation on non compact support in an heteroscesdastic model. *Metrika* 83(1):93–128

Comte F, Marie N (2021) On a Nadaraya-Watson estimator with two bandwidths. *Electronic Journal of Statistics* 15(1):2566–2607

Efromovich S (1999) *Nonparametric curve estimation: methods, theory and applications*. Springer series in statistics, Springer, New York

Gittens A, Tropp JA (2011) Tail bounds for all eigenvalues of a sum of random matrices. *ArXiv*:1104.4513 [math]

Györfi L, Kohler M, Krzyżak A, Walk H (2002) *A Distribution-Free Theory of Nonparametric Regression*. Springer Series in Statistics, Springer New York, New York, NY

Härdle W, Marron JS (1985) Optimal Bandwidth Selection in Nonparametric Regression Function Estimation. *The Annals of Statistics* 13(4):1465–1481

Köhler M, Schindler A, Sperlich S (2014) A Review and Comparison of Bandwidth Selection Methods for Kernel Regression: Review of Bandwidth Selection for Regression. *International Statistical Review* 82(2):243–274

Lacour C, Massart P, Rivoirard V (2017) Estimator Selection: a New Method with Applications to Kernel Density Estimation. *Sankhya A* 79(2):298–335

Mabon G (2017) Adaptive Deconvolution on the Non-negative Real Line: Adaptive deconvolution on R+. *Scandinavian Journal of Statistics* 44(3):707–740

Nadaraya EA (1964) On Estimating Regression. *Theory of Probability & Its Applications* 9(1):141–142

Sacko O (2020) Hermite density deconvolution. *Latin American Journal of Probability and Mathematical Statistics* 17(1):419–443

Tao T (2008) The divisor bound. URL `https://terrytao.wordpress.com/2008/09/23/the-divisor-bound`

Tropp JA (2012) User-Friendly Tail Bounds for Sums of Random Matrices. *Foundations of Computational Mathematics* 12(4):389–434

Tsybakov AB (2009) *Introduction to nonparametric estimation*. Springer series in statistics, Springer, New York ; London

Watson GS (1964) Smooth Regression Analysis. *Sankhyā: The Indian Journal of Statistics, Series A* 26(4):359–372