Minimax adaptive dimension reduction for regression

Quentin Paris

Minimax adaptive dimension reduction for regression

Quentin Paris

IRMAR, ENS Cachan Bretagne, CNRS, UEB
Campus de Ker Lann
Avenue Robert Schuman, 35170 Bruz, France
quentin.paris@bretagne.ens-cachan.fr

Abstract

In this paper, we address the problem of regression estimation in the context of a p-dimensional predictor when p is large. We propose a general model in which the regression function is a composite function. Our model consists in a nonlinear extension of the usual sufficient dimension reduction setting. The strategy followed for estimating the regression function is based on the estimation of a new parameter, called the reduced dimension. We adopt a minimax point of view and provide both lower and upper bounds for the optimal rates of convergence for the estimation of the regression function in the context of our model. We prove that our estimate adapts, in the minimax sense, to the unknown value d of the reduced dimension and achieves therefore fast rates of convergence when $d \ll p$.

Index Terms – Regression estimation, dimension reduction, minimax rates of convergence, empirical risk minimization, metric entropy.

AMS 2000 Classification – 62H12, 62G08.

Contents

1 Introduction **2**
1.1 The curse of dimensionality in regression 2
1.2 A general model for dimension reduction in regression 3
1.3 Organization of the paper 4

2 Model and statistical methodology **5**
2.1 The model 5
2.2 The reduced dimension 5
2.3 Estimation of the regression function 7

3 Results **8**
3.1 Performance of the estimates $\hat{r}_\ell$ 8
3.2 Performance of $\hat{d}$ 10
3.3 Dimension adaptivity of $\hat{r}$ 11

4 Proofs **13**
4.1 Proof of Theorem 3.1 13
4.2 Proof of Theorem 3.2 17
4.3 Proof of Theorem 3.3 18
4.4 Proof of Theorem 3.4 22
4.5 Proof of Theorem 3.5 23

A Appendix **24**
A.1 Reduced dimension d and parameter Δ 24
A.2 Performance of least-squares estimates 24

1 Introduction

1.1 The curse of dimensionality in regression

From a general point of view, the goal of regression is to infer about the conditional distribution of a real-valued response variable Y given an $\mathcal{X}$-valued predictor variable X where $\mathcal{X} \subset \mathbb{R}^p$. In the statistical framework, one usually focuses on the estimation of the regression function

$$r(x) = \mathbb{E}(Y|X = x), \tag{1.1}$$

based on a sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ of n independent and identically distributed random variables with same distribution P as the generic random couple (X, Y).

A major issue in regression, known as the curse of dimensionality, is basically that the rates of convergence of estimates of the regression function are slow when the dimension p of the predictor variable X is high. For instance, if r is assumed to be β-Hölder and if $\hat{r}$ refers to any classical estimate (say a kernel, a nearest-neighbors or a least-squares estimate), the mean squared error $\mathbb{E}\left(\hat{r}(X) - r(X)\right)^2$ of $\hat{r}$ converges to 0 at the rate $n^{-2\beta/(2\beta+p)}$, which gets slower as p increases. To get a deeper understanding of the problem, one may refer to the minimax point of view. First, we recall the definition of optimal rates of convergence in the minimax sense. Given a set $\mathcal{D}$ of distributions P of the random couple (X, Y), v_n is said to be an optimal rate of convergence in the minimax sense for $\mathcal{D}$ if it is a lower minimax rate, i.e.,

$$\liminf_{n\to+\infty} v_n^{-2} \inf_{\hat{r}} \sup_{P\in\mathcal{D}} \mathbb{E}\left(\hat{r}(X) - r(X)\right)^2 > 0,$$

where the infimum is taken over all estimates based on our sample, and if there exists an estimate $\hat{r}$ such that

$$\limsup_{n\to+\infty} \upsilon_n^{-2} \sup_{P\in\mathcal{D}} \mathbb{E}\left(\hat{r}(X) - r(X)\right)^2 < +\infty.$$

Then, in a word, when $\mathcal{D}$ is taken as the set of all distributions P of the random couple (X, Y) for which r is β-Hölder, the optimal rate of convergence for $\mathcal{D}$ is $\upsilon_n = n^{-\beta/(2\beta+p)}$ (for more details on optimal rates of convergence, we refer the reader to Stone, 1982; Györfi et al., 2002; Kohler et al., 2009; Tsybakov, 2009). Accordingly, there is no hope of constructing an estimate which converges at a faster rate under the only general assumption that r is regular. Hence, the only alternative to obtain faster rates is to exploit additional information on the regression function.

1.2 A general model for dimension reduction in regression

In practice, when such additional information is available, it is often encoded in regression models as so called structural assumptions on the regression function. Statistical procedures based on such models are usually referred to as dimension reduction techniques. In the recent years, much attention has been paid to dimension reduction techniques due to the increasing complexity of the data considered in applications. Among popular models for dimension reduction in regression, one can mention for example the single index model (see, e.g., Alquier and Biau, 2013, and the references therein), the additive regression model or the projection pursuit model (see, e.g., Chapter 22 in Györfi et al., 2002). Another important dimension reduction framework is called sufficient dimension reduction. In this framework, one assumes that

$$\mathbb{E}\left(Y|X\right) = \mathbb{E}\left(Y|\Lambda X\right) \quad \text{and} \quad \mathbb{E}\left(Y|\Lambda X = .\right) \in \mathcal{G}, \tag{1.2}$$

are satisfied for a matrix $\Lambda \in \mathcal{M}_p(\mathbb{R})$ of rank smaller than p, and a class $\mathcal{G}$ of regular functions (see, e.g., Härdle and Stoker, 1989; Li, 1991; Cook, 1998, and the references therein). The motivation for studying such a model is that, provided the matrix Λ may be estimated, the predictor variable X may be replaced by ΛX which takes its values in a lower dimensional space. Many methods have been introduced in the litterature to estimate Λ among which we mention average derivative estimation (ADE) (Härdle and Stoker, 1989), sliced inverse regression (SIR) (Li, 1991), principal Hessian directions (PHD) (Li, 1992), sliced average variance estimation (SAVE) (Cook and Weisberg, 1991), kernel dimension reduction (KSIR) (Fukumizu et al., 2009) and, more recently, the optimal transformation procedure (Delyon and Portier, 2013). Discussions, improvements and other relevant papers on that topic can be found in Cook and Li (2002); Fung et al. (2002); Xia et al. (2002); Cook and Ni (2005); Yin et al. (2008), and in the references therein. In the last years, little attention

has been paid to measuring the impact of the estimation of Λ in terms of the estimation of r. Recently, Cadre and Dong (2010) have used these methods to show that, in the context of model (1.2), one could indeed construct an estimate $\hat{r}$ of the regression function such that

$$\mathbb{E}\left(\hat{r}(X) - r(X)\right)^2 = O\left(n^{-2/(2+\mathrm{rank}(\Lambda))}\right),$$

when $\mathcal{G}$ is taken as a class of Lipschitz functions.

In the present article, we tackle the problem of dimension reduction for regression by studying a model which consists in a nonlinear extension of (1.2) and which is described as follows.

Our model – For a given class $\mathcal{H}$ of functions $h : \mathcal{X} \to \mathbb{R}^p$ and a given class $\mathcal{G}$ of regular functions $g : \mathbb{R}^p \to \mathbb{R}$, we assume that the two conditions

$$(i)\ \mathbb{E}\left(Y|X\right) = \mathbb{E}\left(Y|h(X)\right) \quad \text{and} \quad (ii)\ \mathbb{E}(Y|h(X) = .) \in \mathcal{G}, \tag{1.3}$$

are satisfied for at least one function $h \in \mathcal{H}$. In other words, denoting

$$\mathcal{F} = \{g \circ h : g \in \mathcal{G}, h \in \mathcal{H}\}, \tag{1.4}$$

we assume that $r \in \mathcal{F}$.

This model generalizes (1.2) in the sense that the functions $h \in \mathcal{H}$ need not be linear nor regular. The motivation for such a generalization comes from the fact that one may find a much lower dimensional representation $h(X)$ of X satisfying (i) by relaxing the linear requirement made in the usual sufficient dimension reduction setting. In the existing literature, some nonlinear extensions of the classical sufficient dimension reduction framework have been introduced (see, e.g., Cook, 2007) and the estimation of a nonlinear h satisfying (i) has been studied for instance in Wu (2008); Wang and Yin (2008) and Yeh et al. (2009). In this paper, and in the context of our model, we introduce a statistical methodology for estimating the regression function which does not require to estimate such a function h. Our dimension reduction approach is done in the spirit of model selection and is based on the estimation of a new parameter of our model, called the reduced dimension. We adopt a minimax point of view and provide both upper and lower bounds for optimal rates of convergence for the estimation of the regression function in the context of our model. Our constructed estimate of r is shown to adapt to the unknown value d of the reduced dimension in the minimax sense and achieves fast rates when $d << p$, thus reaching the goal of dimension reduction.

1.3 Organization of the paper

In Section 2, we describe our model in further details. We define the reduced dimension d and describe our strategy for defining an estimate of the regression

function which adapts to the unknown value of d. Our results are exposed in Section 3 and proofs are postponed to Section 4.

2 Model and statistical methodology

2.1 The model

As mentioned in the introduction, we assume that the regression function r belongs to the class $\mathcal{F}$ defined by

$$\mathcal{F} = \{g \circ h : \ g \in \mathcal{G}, \ h \in \mathcal{H}\},$$

where $\mathcal{H}$ is a class of functions $h : \mathcal{X} \to \mathbb{R}^p$ and $\mathcal{G}$ a class of functions $g : \mathbb{R}^p \to \mathbb{R}$ that are taken as follows. First, for $R > 0$ fixed, we assume that every function $h \in \mathcal{H}$ satisfies

$$\|h\|_{\mathcal{X}} = \sup_{x \in \mathcal{X}} \|h(x)\| < R, \tag{2.1}$$

where $\|.\|$ stands for the Euclidean norm in $\mathbb{R}^p$. Then, for $\beta > 0$ and $L > 0$ fixed, the class $\mathcal{G}$ is taken as the class of β-Hölder functions with constant L. In other words, $\mathcal{G}$ is the set of all functions $g : \mathbb{R}^p \to \mathbb{R}$ such that

$$\|g\|_{\beta} = \max_{|s| \leq \lfloor \beta \rfloor} \sup_{u} |\partial^s g(u)| + \max_{|s| = \lfloor \beta \rfloor} \sup_{u \neq u'} \frac{|\partial^s g(u) - \partial^s g(u')|}{\|u - u'\|^{\beta - \lfloor \beta \rfloor}} \leq L, \tag{2.2}$$

where $\lfloor \beta \rfloor$ stands for the greatest integer stricly smaller than β and where, for every multi-index $s = (s_1, \ldots, s_p) \in \mathbb{N}^p$, we have denoted $|s| = \sum_i s_i$ and $\partial^s = \partial_1^{s_1} \cdots \partial_p^{s_p}$. A particular aspect of this model is that only functions in $\mathcal{G}$ are assumed regular. Functions in $\mathcal{H}$ may well be nonlinear and nonregular.

For our study, we need to define the following set of distributions of the random couple (X, Y). Let μ be a fixed probability measure on $\mathcal{X}$. Let $\tau > 0$ and $B > 0$ be fixed. Then, we denote $\mathcal{D}$ the set of distributions P of (X, Y) such that the three following conditions are satisfied:

(a) X is of distribution μ;
(b) Y satisfies the exponential moment condition

$$\mathbb{E} \exp(\tau |Y|) \leq B;$$

(c) The regression function $r(.) = \mathbb{E}(Y|X = .)$ belongs to $\mathcal{F}$.

2.2 The reduced dimension

Roughly speaking, the reduced dimension associated to our model is the dimension of the lowest dimensional representation $h(X)$ of X satisfying equations (i)

and (ii). To be more specific, we need some notations. For all $\ell \in \{1, \ldots, p\}$, let

$$\mathcal{H}_\ell = \{h \in \mathcal{H} : \dim S(h) \leq \ell\},$$

where $S(h)$ denotes the subspace of $\mathbb{R}^p$ spanned by $h(\mathcal{X})$. Hence, if $h \in \mathcal{H}_\ell$, the variable $h(X)$ takes its values in an ℓ-dimensional space. Then, set

$$\mathcal{F}_\ell = \{g \circ h : g \in \mathcal{G}, h \in \mathcal{H}_\ell\}.$$

We therefore obtain a nested family of subsets $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots \subset \mathcal{F}_p = \mathcal{F}$ and the reduced dimension is defined as

$$d = \min\{\ell = 1, \ldots, p : r \in \mathcal{F}_\ell\}.$$

This parameter plays a fundamental role in our study and an important part of our work is devoted to its estimation. Our first task will be to derive a tractable representation of the reduced dimension, suitable for estimation purposes. We shall use the following assumption. (Recall that μ denotes the distribution of X.)

Assumption (**A**). *For all $\ell \in \{1, \ldots, p\}$, the set $\mathcal{F}_\ell$ is compact in $\mathbb{L}^2(\mu)$.*

Let R_ℓ be the risk defined by

$$R_\ell = \inf_{f \in \mathcal{F}_\ell} \mathbb{E}(Y - f(X))^2.$$

Since the $\mathcal{F}_\ell$'s are nested, the function $\ell \in \{1, \ldots, p\} \mapsto R_\ell$ is nonincreasing. Then, using Assumption (**A**), we deduce that

$$d = \min\{\ell = 1, \ldots, p : R_\ell = R_p\}. \tag{2.3}$$

(The proof of equation (2.3) may be found in Appendix A.1.) Consequently, denoting

$$\Delta = \min\{R_\ell - R_p : \ R_\ell > R_p\}, \tag{2.4}$$

with the convention $\min \emptyset = +\infty$, we observe that, for all $0 \leq \delta < \Delta$,

$$d = \min\{\ell = 1, \ldots, p : R_\ell \leq R_p + \delta\}. \tag{2.5}$$

Note that $\Delta > 0$ and that, when $d \geq 2$, Δ corresponds to the distance from r to $\mathcal{F}_{d-1}$ in $\mathbb{L}^2(\mu)$. In other words, when $d \geq 2$,

$$\Delta = \inf_{f \in \mathcal{F}_{d-1}} \int_{\mathcal{X}} (f - r)^2 \, \mathrm{d}\mu. \tag{2.6}$$

(The proof of equation (2.6) has been reported to Appendix A.1.)

Based on the representation (2.5) of the reduced dimension, a natural estimate of d may be obtained as follows. For $\ell \in \{1, \dots, p\}$, we introduce an empirical counterpart of the risk R_ℓ defined by

$$\hat{R}_\ell = \inf_{f \in \mathcal{F}_\ell} \frac{1}{n} \sum_{i=1}^{n} \left[Y_i \mathbf{1}\{|Y_i| \le a_n\} - f(X_i) \right]^2,$$

where $a_n > 0$ is a tuning parameter to be fixed later on. Then, for some $\delta_n > 0$, we define the estimate

$$\hat{d} = \min \left\{ \ell = 1, \dots, p : \hat{R}_\ell \le \hat{R}_p + \delta_n \right\}. \tag{2.7}$$

2.3 Estimation of the regression function

Our next task is to construct an estimate of the regression function which achieves fast rates when $d \ll p$. In this paper, we reach this goal by using the following strategy. First, for all $\ell \in \{1, \dots, p\}$, we consider the least-squares type estimate $\hat{r}_\ell$ defined as any random element in $\mathcal{F}_\ell$ satisfying

$$\hat{r}_\ell \in \arg\min_{f \in \mathcal{F}_\ell} \frac{1}{n} \sum_{i=1}^{n} \left[Y_i \mathbf{1}\{|Y_i| \le T_n\} - f(X_i) \right]^2, \tag{2.8}$$

where $T_n > 0$ is a truncation parameter to be tuned later on. Then, we define our final estimate $\hat{r}$ by

$$\hat{r} = \hat{r}_{\hat{d}}, \tag{2.9}$$

where $\hat{d}$ has been defined in (2.7).

This strategy can be motivated as follows. Suppose that, for some known $\ell < p$, the information that r belongs to $\mathcal{F}_\ell$ where available. Then, since the class $\mathcal{F}_\ell$ is smaller than the whole class $\mathcal{F}$, one would naturally be brought to consider the estimate $\hat{r}_\ell$ instead of the estimate $\hat{r}_p$ as it involves a minimization over a smaller class and therefore is expected to converge more rapidly as the sample size n grows. In this respect, the idealized estimate $\hat{r}_d$ appears as the best choice one could possibly make as, by definition of the reduced dimension, d is the smallest ℓ for which r belongs to $\mathcal{F}_\ell$. Now, since the value of d is unknown, we simply replace it by our estimate $\hat{d}$. In fact, by doing so, we obtain an estimate $\hat{r}$ which performance is not too far from that of the idealized estimate $\hat{r}_d$ as shown by the inequality

$$\mathbb{E}\left(\hat{r}(X) - r(X)\right)^2 \le \mathbb{E}\left(\hat{r}_d(X) - r(X)\right)^2 + 4L^2 \mathbb{P}\left(\hat{d} \ne d\right). \tag{2.10}$$

(This inequality is derived in the proof of Theorem 3.5.) In other words, the performance of $\hat{r}$ corresponds to that of $\hat{r}_d$ up to the error term $\mathbb{P}(\hat{d} \ne d)$. Therefore, we will study independently the performance of the estimates $\hat{r}_\ell$ and the performance of the estimate $\hat{d}$ of the reduced dimension to obtain finally the performance of $\hat{r}$.

3 Results

3.1 Performance of the estimates $\hat{r}_\ell$

In this subsection, we study the performance of the estimates $\hat{r}_\ell$ from a minimax point of view. To that aim, we introduce the set $\mathcal{D}_\ell$ of distributions P of the random couple (X, Y) for which the following conditions are satisfied:

(a) X is of distribution μ;
(b) Y satisfies the exponential moment condition

$$\mathbb{E}\exp\left(\tau|Y|\right) \leq B;$$

(c, ℓ) The regression function $r(.) = \mathbb{E}(Y|X = .)$ belongs to $\mathcal{F}_\ell$.

Note that $\mathcal{D}_\ell$ is a subset of the set $\mathcal{D}$ and that $\mathcal{D}_p = \mathcal{D}$. For any measure Q on $\mathcal{X}$, and any class $\mathcal{C} \subset \mathbb{L}^2(Q)$ of real (or vector-valued) functions, we recall that the ε-covering number of $\mathcal{C}$ in $\mathbb{L}^2(Q)$, denoted $N(\varepsilon, \mathcal{C}, \mathbb{L}^2(Q))$, is the minimal number of metric balls of radius ε in $\mathbb{L}^2(Q)$ that are needed to cover $\mathcal{C}$. The ε-metric entropy of $\mathcal{C}$ in $\mathbb{L}^2(Q)$ is defined by $H(\varepsilon, \mathcal{C}, \mathbb{L}^2(Q)) = \ln\ N\left(\varepsilon, \mathcal{C}, \mathbb{L}^2(Q)\right)$. In this paper, we set

$$H(\varepsilon, \mathcal{C}) = \sup_Q\ H\left(\varepsilon, \mathcal{C}, \mathbb{L}^2(Q)\right), \tag{3.1}$$

where the supremum is taken over all probability measures Q with finite support in $\mathcal{X}$. We are now in position to state our first result.

Theorem 3.1. *Let $\ell \in \{1, \ldots, p\}$ be fixed. Let $\alpha > 2$ and set $T_n = (\ln n)^{\alpha/2}$. Suppose that $\beta \geq 1$ and that $\beta > \ell/2$. Suppose, in addition, that there exist $C > 0$ and $0 < s \leq \ell/\beta$ such that, for all $\varepsilon > 0$,*

$$H\left(\varepsilon, \mathcal{H}_\ell\right) \leq C\varepsilon^{-s}. \tag{3.2}$$

Then,

$$\limsup_{n\to+\infty} \left(\frac{n}{(\ln n)^\alpha}\right)^{2\beta/(2\beta+\ell)} \sup_{P\in\mathcal{D}_\ell} \mathbb{E}\left(\hat{r}_\ell(X) - r(X)\right)^2 < +\infty. \tag{3.3}$$

Remarks – An immediate consequence of Theorem 3.1 is that the optimal rate of convergence associated to $\mathcal{D}_\ell$ is upper bounded by

$$\left(\frac{(\ln n)^\alpha}{n}\right)^{\beta/(2\beta+\ell)},$$

for all $\alpha > 2$, which does not depend anymore on the dimension p of the predictor X. Here it is noticeable that, up to a logarithmic factor, we recover the optimal rate $n^{-\beta/(2\beta+\ell)}$ corresponding to the case where X is ℓ-dimensional

and r is only assumed to be β-Hölder (see, e.g., Theorem 1 in Kohler et al., 2009). Since $\mathcal{H}_\ell \subset \mathcal{H}$, condition 3.2 may be satisfied (resp. satisfied for all ℓ) if there exist $C > 0$ and $0 < s \leq \ell/\beta$ (resp. $0 < s \leq 1/\beta$) such that, for all $\varepsilon > 0$, $H(\varepsilon, \mathcal{H}) \leq C\varepsilon^{-s}$. This entropy condition allows for a large variety of classes $\mathcal{H}$ in our model as shown in the next examples. Finally, we mention that if the exponential moment condition (b) is replaced by a boundedness assumption on Y, then a slight modification of the proof reveals that Theorem 3.1 holds without the logarithmic factor.

Examples – 1. *Parametric class.* An first example where condition (3.2) is fulfilled is the following. Consider the case where $\mathcal{H}$ is a parametric class of the form

$$\mathcal{H} = \left\{ h_\theta : \theta \in \Theta \subset \mathbb{R}^k \right\}, \tag{3.4}$$

where Θ is bounded. Suppose that there exists a constant $C > 0$ such that for all $\theta, \theta' \in \Theta$ we have

$$\|h_\theta - h_{\theta'}\|_{\mathcal{X}} \leq C|\theta - \theta'|, \tag{3.5}$$

where $|.|$ stands for the Euclidian norm in $\mathbb{R}^k$. Then, for all $\varepsilon > 0$, $H(\varepsilon, \mathcal{H}) \leq H(\varepsilon/C, \Theta)$ where $H(\varepsilon, \Theta)$ stands for the logarithm of the minimal number of Euclidean balls of radius ε that are needed to cover Θ. Since Θ is bounded, it is included in a Euclidean ball of radius ρ for some $\rho > 0$. Therefore, it follows from Proposition 5 in Cucker and Smale (2001) that, for all $\varepsilon > 0$,

$$H(\varepsilon, \Theta) \leq k \ln(4\rho/\varepsilon).$$

As a result, there exists a constant $C' > 0$ such that, for all $\varepsilon > 0$,

$$H(\varepsilon, \mathcal{H}) \leq C' \ln(1/\varepsilon).$$

Hence, as $\mathcal{H}_\ell \subset \mathcal{H}$, condition (3.2) is satisfied for all $\ell \in \{1, \ldots, p\}$.

2. *Class of regular functions.* In this second example, we show that condition (3.2) may be satisfied when $\mathcal{H}$ is a general (possibly nonparametric) class of regular functions. Suppose $\mathcal{X}$ is bounded, convex and with nonempty interior. Suppose, in addition, that for some constants $\gamma > 0$ and $M > 0$, $\mathcal{H}$ is the class of functions $h : x \in \mathcal{X} \mapsto (h_1(x), \ldots, h_p(x)) \in \mathbb{R}^p$ such that, for each i, $\|h_i\|_\gamma \leq M$. (Norm $\|.\|_\gamma$ has been defined in (2.2).) Then, an easy application of Theorem 9.19 in Kosorok (2008), shows that there exists a constant $K > 0$, depending only on γ, on the diameter of $\mathcal{X}$ and on p such that, for all $\ell \in \{1, \ldots, p\}$,

$$H(\varepsilon, \mathcal{H}_\ell) \leq K\varepsilon^{-\ell/\gamma}.$$

As a result, condition (3.2) is satisfied for all $\ell \in \{1, \ldots, p\}$ in this case, provided $\gamma \geq \beta$.

In our next result, we provide a lower bound for the optimal rate of convergence associated to $\mathcal{D}_\ell$ in order to assess the tightness of the upper bound obtained in Theorem 3.1.

Theorem 3.2. *Let $\ell \in \{1, \ldots, p\}$ be fixed. Suppose that $\beta > 0$. Suppose, in addition, that there exist $h \in \mathcal{H}_\ell$ with $\dim S(h) = \ell$ and a constant $c > 0$ such that*

$$\mu \circ h^{-1} \geq c\, \lambda_h \left(. \cap \mathbf{B}\right). \tag{3.6}$$

Here, λ_h denotes the Lebesgue measure in $S(h)$ and $\mathbf{B}$ stands for the open Euclidean ball in $\mathbb{R}^p$ with center the origin and radius R. Then,

$$\liminf_{n \to +\infty} n^{2\beta/(2\beta+\ell)} \inf_{\hat{r}} \sup_{P \in \mathcal{D}_\ell} \mathbb{E}\left(\hat{r}(X) - r(X)\right)^2 > 0, \tag{3.7}$$

where the infinimum is taken over all estimates $\hat{r}$.

Remarks – Theorem 3.2 indicates that the optimal rate of convergence associated to $\mathcal{D}_\ell$ is lower bounded by $n^{-\beta/(2\beta+\ell)}$ which, up to a logarithmic factor, corresponds to the upper bound found in Theorem 3.1. It is important to mention that condition (3.6) is not restrictive. As an example, it is satisfied if $\mathcal{X} = \mathbf{B}$, if the function $h : (x_1, \ldots, x_p) \in \mathbb{R}^p \mapsto (x_1, \ldots, x_\ell, 0, \ldots, 0) \in \mathbb{R}^p$ belongs to $\mathcal{H}$ and if μ has a density with respect to the Lebesgue measure which is lower bounded by a positive constant on $\mathbf{B}$. For the proof of Theorem 3.2, we have used results from Yang and Barron (1999).

3.2 Performance of $\hat{d}$

Our second task is to study the behavior of the estimate $\hat{d}$ of the reduced dimension d introduced in (2.7). To that aim, we need to introduce a notation. For all $\underline{\delta} \geq 0$, we set

$$\mathcal{D}(\underline{\delta}) = \{P \in \mathcal{D} : \Delta \geq \underline{\delta}\}.$$

When $d \geq 2$, and according to the interpretation of Δ given in (2.6), the set $\mathcal{D}(\underline{\delta})$ corresponds to the subset of all distributions P of (X, Y) that are in $\mathcal{D}$ and for which r satisfies

$$\inf_{f \in \mathcal{F}_{d-1}} \int_{\mathcal{X}} (f - r)^2 \, \mathrm{d}\mu \geq \underline{\delta}.$$

As shown by the next result, for all $\underline{\delta} > 0$, the estimate $\hat{d}$ performs uniformly well over $\mathcal{D}(\underline{\delta})$.

Theorem 3.3. *Suppose that Assumption (**A**) is satisfied. Suppose that $\beta \geq 1$. Suppose, in addition, that there exist $C > 0$ and $0 < s \leq p/\beta$ such that, for all $\varepsilon > 0$,*

$$H(\varepsilon, \mathcal{H}) \leq C\varepsilon^{-s}.$$

Then, if we take $a_n = n^u$ and $\delta_n = n^{-u'}$ with

$$u > 0,\ u' > 0 \quad \text{and} \quad (u + u')\left(2 + \frac{p}{\beta}\right) + 2u < 1,$$

the two following statements hold:

(*i*) *For all* $\vartheta > 0$,

$$\lim_{n\to+\infty} n^{\vartheta} \sup_{P\in\mathcal{D}} \mathbb{P}\left(\hat{d} > d\right) = 0.$$

(*ii*) *For all* $\underline{\delta} > 0$ *and for all* $\vartheta > 0$,

$$\lim_{n\to+\infty} n^{\vartheta} \sup_{P\in\mathcal{D}(\underline{\delta})} \mathbb{P}\left(\hat{d} < d\right) = 0.$$

Remarks – A straightforward consequence of Theorem 3.3 is that, provided $P \in \mathcal{D}$, the probability $\mathbb{P}(\hat{d} \neq d)$ converges to 0 faster than any power of $1/n$. Furthermore, for all $\underline{\delta} > 0$, the estimate $\hat{d}$ behaves uniformly well over the set $\mathcal{D}(\underline{\delta})$ in the sense that, for all $\vartheta > 0$, there exists a constant $C_{\underline{\delta},\vartheta} > 0$ such that

$$\sup_{P\in\mathcal{D}(\underline{\delta})} \mathbb{P}\left(\hat{d} \neq d\right) \leq \frac{C_{\underline{\delta},\vartheta}}{n^{\vartheta}}.$$

The next result shows that, however, the performance of $\hat{d}$ is not uniform over the whole set $\mathcal{D}$ of distributions.

Theorem 3.4. *Let* $\mathcal{D}_0 \subset \mathcal{D}$ *be any subset of* $\mathcal{D}$ *such that* $\inf\{\Delta : P \in \mathcal{D}_0\} = 0$. *Then, under the conditions of Theorem 3.3,*

$$\lim_{n\to+\infty} \sup_{P\in\mathcal{D}_0} \mathbb{P}\left(\hat{d} < d\right) = 1.$$

3.3 Dimension adaptivity of $\hat{r}$

Now we apply the results of the two previous subsections to show that the estimate $\hat{r}$ defined by

$$\hat{r} = \hat{r}_{\hat{d}},$$

adapts to the unknown value of the reduced dimension d.

Theorem 3.5. *Let* $\alpha > 2$ *and set* $T_n = (\ln n)^{\alpha/2}$. *Suppose that Assumption* (**A**) *is satisfied. Suppose that* $\beta \geq 1$ *and that* $\beta > p/2$. *Suppose, in addition, that for all* $\ell \in \{1, \ldots, p\}$, *there exist* $C > 0$ *and* $0 < s \leq \ell/\beta$ *such that, for all* $\varepsilon > 0$,

$$H(\varepsilon, \mathcal{H}_\ell) \leq C\varepsilon^{-s}.$$

Then, if we take $a_n = n^u$ *and* $\delta_n = n^{-u'}$ *with*

$$u > 0,\ u' > 0 \quad \textit{and} \quad (u + u')\left(2 + \frac{p}{\beta}\right) + 2u < 1,$$

we obtain, for all $\underline{\delta} > 0$,

$$\limsup_{n\to+\infty} \sup_{P\in\mathcal{D}(\underline{\delta})} \left(\frac{n}{(\ln n)^{\alpha}}\right)^{2\beta/(2\beta+d)} \mathbb{E}\left(\hat{r}(X) - r(X)\right)^2 < +\infty.$$

Remarks – (1) A first remark concerns the impact of this result in terms of individual rates of convergence. Details of the proof of Theorem 3.5 reveal that, provided $P \in \mathcal{D}$, $\beta > d/2$, $H(\varepsilon, \mathcal{H}_d) \leq C\varepsilon^{-s}$ is satisfied for some $0 < s \leq d/\beta$ and $H(\varepsilon, \mathcal{H}) \leq C\varepsilon^{-s'}$ is satisfied for some $0 < s' \leq p/\beta$, our estimate $\hat{r}$ satisfies

$$\mathbb{E}\left(\hat{r}(X) - r(X)\right)^2 = O\left(\frac{(\ln n)^\alpha}{n}\right)^{2\beta/(2\beta+d)}.$$

In other words, from the point of view of individual rates of convergence, one may impose less strict conditions. In particular, we observe that the regularity parameter β of functions in $\mathcal{G}$ needs only to satisfy condition $\beta > d/2$, which becomes less restrictive as the reduced dimension d gets smaller.

(2). This result is an improvement on three levels. First, in the context of a high-dimensional predictor X, we have introduced a regression model with a structural assumption that has been successfully exploited to construct an estimate $\hat{r}$ which achieves faster rates when $d \ll p$. Second, since the value of d is unknown, our estimate $\hat{r}$ is adaptive. Third, for all $\underline{\delta} > 0$, the adaptivity of $\hat{r}$ is uniform over the set $\mathcal{D}(\underline{\delta}) \subset \mathcal{D}$ of distributions P in the sense that there exists $C_{\underline{\delta}} > 0$ such that, for all $P \in \mathcal{D}(\underline{\delta})$,

$$\mathbb{E}\left(\hat{r}(X) - r(X)\right)^2 \leq C_{\underline{\delta}}\left(\frac{(\ln n)^\alpha}{n}\right)^{2\beta/(2\beta+d)}.$$

(3). Another important remark is the following. In the ideal situation where the value of the reduced dimension d where known, we have seen that one may not construct an estimate of r which converges at a rate faster than $n^{-\beta/(2\beta+d)}$ since, according to Theorem 3.2,

$$\liminf_{n\to+\infty} n^{2\beta/(2\beta+d)} \inf_{\hat{r}} \sup_{P\in\mathcal{D}_d} \mathbb{E}\left(\hat{r}(X) - r(X)\right)^2 > 0.$$

In other words, without knowing the value of the reduced dimension, we have constructed an estimate of r which, up to a logarithmic factor, converges at the best possible rate that one could obtain knowing d.

(4). Details of the proof of Theorem 3.5 show that if the exponential moment condition (b) is replaced by a boundedness assumption on Y, then a slight adaptation allows to obtain the same result without the logarithmic factor.

(5). We conclude with a technical remark. The condition $\beta > p/2$ in Theorem 3.5 allows adaptation for all values of $d \in \{1, \ldots, p\}$. We do not know whether this result holds for $\beta \leq p/2$. That being said, if adaptation is required only for small dimensions, we have the following result. For all $\beta \geq 1$ and all $\underline{\delta} > 0$, let

$$\mathcal{D}(\underline{\delta}, \beta) := \mathcal{D}(\underline{\delta}) \cap \{P \in \mathcal{D} : d \leq \ell_\beta\},$$

where $\ell_\beta := [2\beta] - 1$, and where $[x]$ stands for the greatest integer smaller or equal to x. Then, we readily obtain, for all $\alpha > 2$,

$$\limsup_{n\to+\infty} \sup_{P\in\mathcal{D}(\underline{\delta},\beta)} \left(\frac{n}{(\ln n)^\alpha}\right)^{2\beta/(2\beta+d)} \mathbb{E}\left(\hat{r}(X) - r(X)\right)^2 < +\infty.$$

4 Proofs

4.1 Proof of Theorem 3.1

Lemma 4.1. *Let $\ell \in \{1, \dots, p\}$ be fixed. Suppose that $\beta \geq 1$. Then, for all $\varepsilon > 0$,*

$$H\left(\varepsilon, \mathcal{F}_\ell\right) \leq \sup_{h \in \mathcal{H}_\ell} H\left(\tfrac{\varepsilon}{2}, \mathcal{G} \circ h\right) + H\left(\tfrac{\varepsilon}{2L}, \mathcal{H}_\ell\right).$$

Proof – Let $\varepsilon > 0$ be fixed and let Q be any probability measure with support in $\mathcal{X}$. We denote

$$N = N\left(\tfrac{\varepsilon}{2L}, \mathcal{H}_\ell, \mathbb{L}^2(Q)\right),$$

and choose an $\frac{\varepsilon}{2L}$-covering $\{h_1, \dots, h_N\}$ of $\mathcal{H}_\ell$ in $\mathbb{L}^2(Q)$ of minimum cardinality. For all $i \in \{1, \dots, N\}$, let

$$N_i = N\left(\tfrac{\varepsilon}{2}, \mathcal{G} \circ h_i, \mathbb{L}^2(Q)\right),$$

and let $\{g_1^i \circ h_i, \dots, g_{N_i}^i \circ h_i\}$ be an $\frac{\varepsilon}{2}$-covering of $\mathcal{G} \circ h_i = \{g \circ h_i : g \in \mathcal{G}\}$ in $\mathbb{L}^2(Q)$. Then, let $g \in \mathcal{G}$ and $h \in \mathcal{H}_\ell$ be chosen arbitrarily. By definition, there exists $i \in \{1, \dots, N\}$ and $j \in \{1, \dots, N_i\}$ such that

$$\sqrt{\int \|h - h_i\|^2 \mathrm{d}Q} \leq \tfrac{\varepsilon}{2L} \quad \text{and} \quad \sqrt{\int \left(g \circ h_i - g_j^i \circ h_i\right)^2 \mathrm{d}Q} \leq \tfrac{\varepsilon}{2}.$$

Therefore,

$$\begin{aligned}
&\sqrt{\int \left(g \circ h - g_j^i \circ h_i\right)^2 \mathrm{d}Q} \\
&\quad \leq \quad \sqrt{\int \left(g \circ h - g \circ h_i\right)^2 \mathrm{d}Q} + \sqrt{\int \left(g \circ h_i - g_j^i \circ h_i\right)^2 \mathrm{d}Q} \\
&\quad \leq \quad \sqrt{\int \left(g \circ h - g \circ h_i\right)^2 \mathrm{d}Q} + \frac{\varepsilon}{2}.
\end{aligned}$$

According to the mean value Theorem, each $g \in \mathcal{G}$ is L-Lipschitz. As a consequence,

$$\begin{aligned}
\sqrt{\int \left(g \circ h - g_j^i \circ h_i\right)^2 \mathrm{d}Q} \quad &\leq \quad \sqrt{\int \left(g \circ h - g \circ h_i\right)^2 \mathrm{d}Q} + \frac{\varepsilon}{2} \\
&\leq \quad L\sqrt{\int \|h - h_i\|^2 \mathrm{d}Q} + \frac{\varepsilon}{2} \\
&\leq \quad \varepsilon. \qquad (4.1)
\end{aligned}$$

Hence, by denoting $|\mathcal{A}|$ the cardinality of a set $\mathcal{A}$, we have obtained

$$\begin{aligned} N\left(\varepsilon, \mathcal{F}_\ell, \mathbb{L}^2(Q)\right) &\leq \left|\left\{g_j^i \circ h_i : i \in \{1,\dots,N\},\ j \in \{1,\dots,N_i\}\right\}\right| \\ &= \sum_{i=1}^{N} N_i \\ &\leq \sup_{h\in\mathcal{H}_\ell} N\left(\tfrac{\varepsilon}{2}, \mathcal{G}\circ h, \mathbb{L}^2(Q)\right) N\left(\tfrac{\varepsilon}{2L}, \mathcal{H}_\ell, \mathbb{L}^2(Q)\right). \end{aligned}$$

Therefore,

$$\begin{aligned} H\left(\varepsilon, \mathcal{F}_\ell, \mathbb{L}^2(Q)\right) &\leq \ln\left(\sup_{h\in\mathcal{H}_\ell} N\left(\tfrac{\varepsilon}{2}, \mathcal{G}\circ h, \mathbb{L}^2(Q)\right)\right) + H\left(\tfrac{\varepsilon}{2L}, \mathcal{H}_\ell, \mathbb{L}^2(Q)\right) \\ &= \sup_{h\in\mathcal{H}_\ell} H\left(\tfrac{\varepsilon}{2}, \mathcal{G}\circ h, \mathbb{L}^2(Q)\right) + H\left(\tfrac{\varepsilon}{2L}, \mathcal{H}_\ell, \mathbb{L}^2(Q)\right), \end{aligned}$$

by continuity. Taking the supremum over all probability measures Q with finite support in $\mathcal{X}$, we obtain the expected result. □

Lemma 4.2. *Let $\ell \in \{1,\dots,p\}$ be fixed. Suppose that $\beta \geq 1$. Suppose, in addition, that there exist $C > 0$ and $0 < s \leq \ell/\beta$ such that, for all $\varepsilon > 0$, $H(\varepsilon, \mathcal{H}_\ell) \leq C\varepsilon^{-s}$. Then, there exist a constant $A > 0$ depending only on ℓ, β, L and R such that, for all $\varepsilon > 0$,*

$$H(\varepsilon, \mathcal{F}_\ell) \leq A\varepsilon^{-\ell/\beta}.$$

Proof – Let $\ell \in \{1,\dots,p\}$ be fixed. According to Lemma 4.1, we only need to prove that there exist a constant $A > 0$, depending only on ℓ, β, L and R, such that, for all $\varepsilon > 0$,

$$\sup_{h\in\mathcal{H}_\ell} H(\varepsilon, \mathcal{G}\circ h) \leq A\varepsilon^{-\ell/\beta}.$$

To that aim, fix a probability measure Q with support in $\mathcal{X}$ and fix $h \in \mathcal{H}_\ell$. For all $g \in \mathcal{G}$, let g_h be the restriction of g to $S(h) \cap \mathbf{B}$, where $\mathbf{B}$ denotes the open Euclidean ball in $\mathbb{R}^p$ with center the origin and radius R. Then, set $\mathcal{G}_h = \{g_h : g \in \mathcal{G}\}$. From the transfer theorem, for all $\varepsilon > 0$,

$$H\left(\varepsilon, \mathcal{G}\circ h, \mathbb{L}^2(Q)\right) = H\left(\varepsilon, \mathcal{G}_h, \mathbb{L}^2(Q\circ h^{-1})\right),$$

where, for any Borel set $A \subset S$, $Q \circ h^{-1}(A) := Q(h^{-1}(A))$. Since $S(h)$ is a vector space of dimension $\ell' \leq \ell$, $S(h) \cap \mathbf{B}$ may be identified to the open Euclidean ball $\mathbf{B}_{\ell'}$ in $\mathbb{R}^{\ell'}$ with center the origin and radius R. Also, $Q \circ h^{-1}$ may be seen as of support in $\mathbf{B}_{\ell'}$ and $\mathcal{G}_h$ as a subset of

$$\mathcal{G}_{\ell'} = \{g : \mathbf{B}_{\ell'} \to \mathbb{R} :\ \|g\|_\beta \leq L\},$$

where $\|.\|_\beta$ is defined as in (2.2) with the appropriate Euclidean norm. According to Theorem 9.19 in Kosorok (2008), there exist a constant $K > 0$ depending only on ℓ', β, R, and L such that, for all $\varepsilon > 0$,

$$\sup_D H\left(\varepsilon, \mathcal{G}_{\ell'}, \mathbb{L}^2(D)\right) \leq K\varepsilon^{-\ell'/\beta},$$

where the supremum is taken over all probability distributions D in $\mathbb{R}^{\ell'}$. Since $\ell' \leq \ell$, and since $\sup_D H\left(\varepsilon, \mathcal{G}_{\ell'}, \mathbb{L}^2(D)\right)$ is equal to 0 for ε sufficiently large, we deduce finally that there exist a constant $K' > 0$ depending only on ℓ, β, R, and L such that, for all $\varepsilon > 0$,

$$\sup_D H\left(\varepsilon, \mathcal{G}_{\ell'}, \mathbb{L}^2(D)\right) \leq K'\varepsilon^{-\ell/\beta},$$

where the supremum is taken over all probability distributions D in $\mathbb{R}^{\ell'}$. Hence, we deduce that, for all $\varepsilon > 0$,

$$H\left(\varepsilon, \mathcal{G} \circ h, \mathbb{L}^2(Q)\right) \leq K'\varepsilon^{-\ell/\beta}.$$

Since the result holds uniformly for all $h \in \mathcal{H}_\ell$ and for all Q with support in $\mathcal{X}$, we deduce finally that, for all $\varepsilon > 0$,

$$\sup_{h \in \mathcal{H}_\ell} H\left(\varepsilon, \mathcal{G} \circ h\right) \leq K'\varepsilon^{-\ell/\beta},$$

which completes the proof. □

Proof of Theorem 3.1 – Let $\ell \in \{1, \ldots, p\}$ be fixed and let $P \in \mathcal{D}_\ell$. In the proof, $C > 0$ will denote a constant depending only on ℓ, β, R and L, and which value may change from line to line. We denote

$$r_n(x) = \mathbb{E}\left(Y\mathbf{1}\left\{|Y| \leq T_n\right\}|X = x\right).$$

Then,

$$\mathbb{E}\left(\hat{r}_\ell(X) - r(X)\right)^2 \leq 2\mathbb{E}\left(\hat{r}_\ell(X) - r_n(X)\right)^2 + 2\mathbb{E}\left(r_n(X) - r(X)\right)^2. \tag{4.2}$$

Using Theorem A.1 of the Appendix with $\varepsilon = 1$, $Z = Y\mathbf{1}\left\{|Y| \leq T_n\right\}$ and Lemma 4.2, we deduce that there exist a constant $C > 0$ such that

$$\begin{aligned}
&\mathbb{E}\left(\hat{r}_\ell(X) - r_n(X)\right)^2 \\
&\quad\leq\quad 2\inf_{f \in \mathcal{F}_\ell} \mathbb{E}\left(f(X) - r_n(X)\right)^2 + C\left(\frac{(T_n + L)^2}{n}\right)^{2\beta/(2\beta+\ell)} + C\left(\frac{(T_n + L)^2}{n}\right) \\
&\quad\leq\quad 2\mathbb{E}\left(r_n(X) - r(X)\right)^2 + C\left(\frac{(T_n + L)^2}{n}\right)^{2\beta/(2\beta+\ell)} + C\left(\frac{(T_n + L)^2}{n}\right),
\end{aligned}$$

where, in the second inequality, we have used the fact that $r \in \mathcal{F}_\ell$. Since $T_n \to +\infty$ and $T_n^2/n \to 0$ as n goes to $+\infty$, we deduce that there exist a constant $C > 0$ such that

$$\mathbb{E}\left(\hat{r}_\ell(X) - r_n(X)\right)^2 \leq 2\mathbb{E}\left(r_n(X) - r(X)\right)^2 + C\left(\frac{T_n^2}{n}\right)^{2\beta/(2\beta+\ell)}. \tag{4.3}$$

Hence, invoking (4.2) and (4.3),

$$\mathbb{E}\left(\hat{r}_\ell(X) - r(X)\right)^2 \leq 6\mathbb{E}\left(r_n(X) - r(X)\right)^2 + C\left(\frac{T_n^2}{n}\right)^{2\beta/(2\beta+\ell)}. \tag{4.4}$$

Using Jensen's inequality and Cauchy-Schwarz's inequality, we obtain

$$\begin{aligned}
\mathbb{E}\left(r_n(X) - r(X)\right)^2 &= \mathbb{E}\Big[\mathbb{E}\left(Y\mathbf{1}\{|Y| \leq T_n\}|X\right) - \mathbb{E}\left(Y|X\right)\Big]^2 \\
&= \mathbb{E}\Big[\mathbb{E}\left(Y\mathbf{1}\{|Y| > T_n\}|X\right)\Big]^2 \\
&\leq \mathbb{E}\left(Y^2\mathbf{1}\{|Y| > T_n\}\right) \\
&\leq \sqrt{\mathbb{E}Y^4}\sqrt{\mathbb{P}\left(|Y| > T_n\right)}.
\end{aligned} \tag{4.5}$$

Then, using the fact that, for all $u \in \mathbb{R}$,

$$\frac{(\tau u)^4}{4!} \leq e^{\tau|u|},$$

we deduce from the exponential moment condition (b) that

$$\mathbb{E}Y^4 \leq \frac{4!}{\tau^4}\mathbb{E}e^{\tau|Y|} \leq \frac{24B}{\tau^4}. \tag{4.6}$$

Using Markov's inequality and the exponential moment condition (b), we obtain

$$\mathbb{P}\left(|Y| > T_n\right) = \mathbb{P}\Big(e^{\tau|Y|} > e^{\tau T_n}\Big) \leq Be^{-\tau T_n}. \tag{4.7}$$

Combining (4.6) and (4.7) we deduce from (4.5) that

$$\mathbb{E}\left(r_n(X) - r(X)\right)^2 \leq \frac{\sqrt{24}B}{\tau^2}e^{-\tau T_n/2}. \tag{4.8}$$

Equations (4.4) and (4.8) imply

$$\mathbb{E}\left(\hat{r}_\ell(X) - r(X)\right)^2 \leq \frac{6\sqrt{24}B}{\tau^2}e^{-\tau T_n/2} + C\left(\frac{T_n^2}{n}\right)^{2\beta/(2\beta+\ell)}. \tag{4.9}$$

Since the constants involved on the right hand side of (4.9) do not depend on $P \in \mathcal{D}_\ell$, we deduce that there exist a constant $C > 0$ such that

$$\sup_{P\in\mathcal{D}_\ell}\mathbb{E}\left(\hat{r}_\ell(X) - r(X)\right)^2 \leq \frac{6\sqrt{24}B}{\tau^2}e^{-\tau T_n/2} + C\left(\frac{T_n^2}{n}\right)^{2\beta/(2\beta+\ell)}.$$

Since $\alpha > 2$, the choice of $T_n = (\ln n)^{\alpha/2}$ leads to

$$e^{-\tau T_n/2} \underset{n\to+\infty}{\ll} \left(\frac{T_n^2}{n}\right)^{2\beta/(2\beta+\ell)} = \left(\frac{(\ln n)^\alpha}{n}\right)^{2\beta/(2\beta+\ell)}.$$

Then, it follows that

$$\limsup_{n\to+\infty} \left(\frac{n}{(\ln n)^\alpha}\right)^{2\beta/(2\beta+\ell)} \sup_{P\in\mathcal{D}_\ell} \mathbb{E}\left(\hat{r}_\ell(X) - r(X)\right)^2 < +\infty,$$

which completes the proof. □

4.2 Proof of Theorem 3.2

Let $\ell \in \{1, \ldots, p\}$ be fixed. Let $\mathcal{D}_\ell^\circ$ be the class of distributions P of (X, Y) such that X is of distribution μ and such that

$$Y = f(X) + \xi,$$

where $f \in \mathcal{F}_\ell$ and where ξ is independent from X and with distribution $\mathcal{N}(0, \sigma^2)$. It may be easily verified that the exponential moment condition (b) holds in this context so that $\mathcal{D}_\ell^\circ \subset \mathcal{D}_\ell$. Therefore,

$$\inf_{\hat{r}} \sup_{P\in\mathcal{D}_\ell^\circ} \mathbb{E}\left(\hat{r}(X) - r(X)\right)^2 \leq \inf_{\hat{r}} \sup_{P\in\mathcal{D}_\ell} \mathbb{E}\left(\hat{r}(X) - r(X)\right)^2.$$

As a result, in order to prove Theorem 3.2, we need only to prove that

$$\liminf_{n\to+\infty} n^{2\beta/(2\beta+\ell)} \inf_{\hat{r}} \sup_{P\in\mathcal{D}_\ell^\circ} \mathbb{E}\left(\hat{r}(X) - r(X)\right)^2 > 0. \tag{4.10}$$

According to Theorem 6 in Yang and Barron (1999), inequality (4.10) is satisfied provided there exist a lower bound $N(\varepsilon)$ for the covering number $N\left(\varepsilon, \mathcal{F}_\ell, \mathbb{L}^2(\mu)\right)$ such that any solution ε_n of

$$\ln N(\varepsilon) = n\varepsilon^2,$$

is of order $n^{-\beta/(2\beta+\ell)}$. To obtain such a lower bound, let $h \in \mathcal{H}_\ell$ be satisfying the conditions of Theorem 3.2. Then, $\mathcal{G}_h = \{g \circ h : g \in \mathcal{G}\} \subset \mathcal{F}_\ell$, which implies that, for all $\varepsilon > 0$,

$$N\left(\varepsilon, \mathcal{F}_\ell, \mathbb{L}^2(\mu)\right) \geq N\left(\varepsilon, \mathcal{G}_h, \mathbb{L}^2(\mu)\right). \tag{4.11}$$

The right hand side of inequality (4.11) may be lower bounded as follows. For all $g \in \mathcal{G}$, let g_h be the restriction of g to $S(h) \cap \mathbf{B}$, where $\mathbf{B}$ stands for the open Euclidean ball in $\mathbb{R}^p$ with center the origin and radius R. (Note that, with this notation, $\mathcal{G}_h = \{g_h : g \in \mathcal{G}\}$.) Then, for all $g, g' \in \mathcal{G}$, we have

$$\begin{aligned}\int (g \circ h(x) - g' \circ h(x))^2 \mu(\mathrm{d}x) &= \int (g_h(u) - g'_h(u))^2 \mu \circ h^{-1}(\mathrm{d}u) \\ &\geq c \int (g_h(u) - g'_h(u))^2 \lambda_h(\mathrm{d}u), \end{aligned} \tag{4.12}$$

where in (4.12) we have used (3.6). Hence, for all $\varepsilon > 0$,

$$N\left(\varepsilon, \left\{g \circ h : g \in \mathcal{G}\right\}, \mathbb{L}^2(\mu)\right) \geq N\left(\tfrac{\varepsilon}{c}, \mathcal{G}_h, \mathbb{L}^2(\lambda_h)\right). \tag{4.13}$$

Now, let us identify $S(h) \cap \mathbf{B}$ to the open Euclidean ball in $\mathbb{R}^\ell$ with center the origin and radius $\mathbb{R}^\ell$, and λ_h to the Lebesgue measure in $\mathbb{R}^\ell$. Then, according to Corollary 2.4 of chapter 15 in Lorentz et al. (1996), there exist a constant $c' > 0$ such that, for all $\varepsilon > 0$,

$$\ln N\left(\tfrac{\varepsilon}{c}, \mathcal{G}_h, \mathbb{L}^2(\lambda_h)\right) \geq c' \varepsilon^{-\ell/\beta}. \tag{4.14}$$

Combining (4.11), (4.13) and (4.14), we obtain

$$\ln N\left(\varepsilon, \mathcal{F}_\ell, \mathbb{L}^2(\mu)\right) \geq c' \varepsilon^{-\ell/\beta} =: \ln N(\varepsilon).$$

It may be easily verified that the solution ε_n of $c' \varepsilon_n^{-\ell/\beta} = n\varepsilon_n^2$ is given by a constant times $n^{-\beta/(2\beta+\ell)}$, and this concludes the proof. □

4.3 Proof of Theorem 3.3

Lemma 4.3. *Suppose that $\beta \geq 1$. Suppose, in addition, that there exist $C > 0$ and $0 < s \leq p/\beta$ such that, for all $\varepsilon > 0$, $H(\varepsilon, \mathcal{H}) \leq C\varepsilon^{-s}$. Take $a_n = n^u$ and $\delta_n = n^{-u'}$ with*

$$u > 0, \ u' > 0 \quad \text{and} \quad (u + u')\left(2 + \frac{p}{\beta}\right) + 2u < 1.$$

Then, for all $\ell \in \{1, \ldots, p\}$ and for all $\vartheta > 0$,

$$\lim_{n \to +\infty} n^\vartheta \sup_{P \in \mathcal{D}} \mathbb{P}\left(\left|\hat{R}_\ell - R_\ell\right| \geq \delta_n\right) = 0.$$

Proof – Let

$$\bar{R}_\ell = \inf_{f \in \mathcal{F}_\ell} \mathbb{E}\left(Y\mathbf{1}\{|Y| \leq a_n\} - f(X)\right)^2.$$

Then,

$$\begin{aligned}
|\hat{R}_\ell - R_\ell| &\leq |\hat{R}_\ell - \bar{R}_\ell| + |\bar{R}_\ell - R_\ell| \\
&\leq \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \left(Y_i \mathbf{1}\{|Y_i| \leq a_n\} - f(X_i)\right)^2 - \mathbb{E}\left(Y\mathbf{1}\{|Y| \leq a_n\} - f(X)\right)^2 \right| \\
&\quad + \sup_{f \in \mathcal{F}} \left| \mathbb{E}\left(Y\mathbf{1}\{|Y| \leq a_n\} - f(X)\right)^2 - \mathbb{E}\left(Y - f(X)\right)^2 \right|. \qquad (4.15)
\end{aligned}$$

Using Cauchy-Schwarz's inequality, for all $f \in \mathcal{F}$,

$$\begin{aligned}
&\left|\mathbb{E}\left(Y\mathbf{1}\{|Y| \leq a_n\} - f(X)\right)^2 - \mathbb{E}\left(Y - f(X)\right)^2\right| \\
&\quad = \left|\mathbb{E}\left[\left(Y\mathbf{1}\{|Y| \leq a_n\} + Y - 2f(X)\right)\left(Y\mathbf{1}\{|Y| \leq a_n\} - Y\right)\right]\right| \\
&\quad \leq \mathbb{E}\left[\left|Y\mathbf{1}\{|Y| \leq a_n\} + Y - 2f(X)\right| \left|Y\mathbf{1}\{|Y| > a_n\}\right|\right] \\
&\quad \leq \sqrt{\mathbb{E}\left(Y\mathbf{1}\{|Y| \leq a_n\} + Y - 2f(X)\right)^2} \sqrt{\mathbb{E}Y^2\mathbf{1}\{|Y| > a_n\}}. \qquad (4.16)
\end{aligned}$$

Using the fact that, for all $u \in \mathbb{R}$ and for every positive integer k,

$$\frac{(\tau u)^k}{k!} \leq e^{\tau|u|},$$

we deduce from the exponential moment condition (b) that, for every positive integer k,

$$\mathbb{E}Y^k \leq \frac{k!}{\tau^k}\mathbb{E}e^{\tau|Y|} \leq \frac{k!B}{\tau^k}. \tag{4.17}$$

Then, using Minkowski's inequality and the fact that all functions $f \in \mathcal{F}$ are bounded by L, we deduce that

$$\begin{aligned}\sqrt{\mathbb{E}\left(Y\mathbf{1}\{|Y| \leq a_n\} + Y - 2f(X)\right)^2} &\leq 2\sqrt{\mathbb{E}Y^2} + 2\sqrt{\mathbb{E}f(X)^2} \\ &\leq 2\sqrt{\frac{2}{\tau^2}\mathbb{E}e^{\tau|Y|}} + 2L \\ &\leq 2\sqrt{\frac{2B}{\tau^2}} + 2L. \end{aligned} \tag{4.18}$$

Using Markov's inequality, and the exponential moment condition (b), we obtain

$$\mathbb{P}\left(|Y| > a_n\right) = \mathbb{P}\left(e^{\tau|Y|} > e^{\tau a_n}\right) \leq Be^{-\tau a_n}. \tag{4.19}$$

Then, using Cauchy-Schwarz's inequality and equations (4.17) and (4.19), we deduce that

$$\begin{aligned}\mathbb{E}Y^2\mathbf{1}\left\{|Y| > a_n\right\} &\leq \sqrt{\mathbb{E}Y^4}\sqrt{\mathbb{P}\left(|Y| > a_n\right)} \\ &\leq \sqrt{\frac{4!}{\tau^4}\mathbb{E}e^{\tau|Y|}}\sqrt{\mathbb{P}\left(e^{\tau|Y|} > e^{\tau a_n}\right)} \\ &\leq B\sqrt{\frac{24}{\tau^4}}e^{-\tau a_n/2}. \end{aligned} \tag{4.20}$$

Hence, combining (4.16), (4.18) and (4.20) yields

$$\begin{aligned}&\sup_{f\in\mathcal{F}}\left|\mathbb{E}\left(Y\mathbf{1}\{|Y| \leq a_n\} - f(X)\right)^2 - \mathbb{E}\left(Y - f(X)\right)^2\right| \\ &\quad\leq \sqrt{\left(\tfrac{2\sqrt{2B}}{\tau} + 2L\right)}\sqrt{\left(\tfrac{B\sqrt{24}}{\tau^2}\right)}e^{-\tau a_n/4} \\ &\quad=: Ue^{-\tau a_n/4}.\end{aligned}$$

Therefore, denoting letting $\kappa_n = \delta_n - Ue^{-\tau a_n/4}$ and

$$\bar{\mathcal{F}}^{a_n} = \left\{(x,y) \mapsto \left(y\mathbf{1}\{|y| \leq a_n\} - f(x)\right)^2 : f \in \mathcal{F}\right\},$$

we deduce from (4.15) and Theorem 9.1 in Györfi et al. (2002) that there exist a universal constant $C > 0$ such that

$$\begin{aligned}
&\mathbb{P}\left(|\hat{R}_\ell - R_\ell| \geq \delta_n\right) \\
&\leq \mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\left(Y_i\mathbf{1}\{|Y_i|\leq a_n\} - f(X_i)\right)^2 - \mathbb{E}\left(Y\mathbf{1}\{|Y|\leq a_n\} - f(X)\right)^2\right| \geq \kappa_n\right) \\
&\leq C\mathbb{E}\left[N\left(\tfrac{\kappa_n}{C}, \bar{\mathcal{F}}^{a_n}, \mathbb{L}^1(P_n)\right)\right] e^{-\frac{n\kappa_n^2}{C(a_n+L)^4}}. \qquad (4.21)
\end{aligned}$$

Here, $P_n = n^{-1}\sum_{i=1}^{n}\delta_{(X_i,Y_i)}$ denotes the empirical distribution associated with the sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ and $N(\varepsilon, \mathcal{C}, \mathbb{L}^1(Q))$ denotes the minimal number of metric balls of radius ε in $\mathbb{L}^1(Q)$ that are needed to cover $\mathcal{C}$. For all $f, f' \in \mathcal{F}$ we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left|\left(Y_i\mathbf{1}\{|Y_i|\leq a_n\} - f(X_i)\right)^2 - \left(Y_i\mathbf{1}\{|Y_i|\leq a_n\} - f'(X_i)\right)^2\right| \\
&\quad = \frac{1}{n}\sum_{i=1}^{n}\left|2Y_i\mathbf{1}\{|Y_i|\leq a_n\} - f(X_i) - f'(X_i)\right|\left|f(X_i) - f'(X_i)\right| \\
&\quad \leq \frac{2(a_n+L)}{n}\sum_{i=1}^{n}\left|f(X_i) - f'(X_i)\right| \\
&\quad \leq 2(a_n+L)\left\{\frac{1}{n}\sum_{i=1}^{n}\left(f(X_i) - f'(X_i)\right)^2\right\}^{1/2}.
\end{aligned}$$

Therefore, we obtain

$$N\left(\tfrac{\kappa_n}{C}, \bar{\mathcal{F}}^{a_n}, \mathbb{L}^1(P_n)\right) \leq N\left(\tfrac{\kappa_n}{2C(a_n+L)}, \mathcal{F}, \mathbb{L}^2(\mu_n)\right),$$

where $\mu_n = n^{-1}\sum_{i=1}^{n}\delta_{X_i}$. Hence, we deduce from (4.21), from the entropy condition

$$H(\varepsilon, \mathcal{H}) \leq C\varepsilon^{-p/\beta},$$

and from Lemma 4.2 that there exist a universal constant $C > 0$ such that

$$\mathbb{P}\left(|\hat{R}_\ell - R_\ell| \geq \delta_n\right) \leq C\exp\left[\left(\frac{C(a_n+L)}{\kappa_n}\right)^{p/\beta} - \frac{n\kappa_n^2}{C(a_n+L)^4}\right]. \qquad (4.22)$$

Hence, since $C > 0$ is universal,

$$\sup_{P\in\mathcal{D}}\mathbb{P}\left(|\hat{R}_\ell - R_\ell| \geq \delta_n\right) \leq C\exp\left[\left(\frac{C(a_n+L)}{\kappa_n}\right)^{p/\beta} - \frac{n\kappa_n^2}{C(a_n+L)^4}\right].$$

Now recall that $a_n = n^u$, that $\delta_n = n^{-u'}$ and that $\kappa_n = \delta_n - Ue^{-\tau a_n/4}$. Then, it may be easily observed that, provided

$$(u+u')\left(2 + \frac{p}{\beta}\right) + 2u < 1,$$

we obtain, for all $\vartheta \geq 0$,

$$\lim_{n\to+\infty} n^{\vartheta} \sup_{P\in\mathcal{D}} \mathbb{P}\left(|\hat{R}_\ell - R_\ell| \geq \delta_n\right) = 0.$$

This completes the proof. □

Proof of Theorem 3.3 – Let $P \in \mathcal{D}$. Since the function $\ell \mapsto \hat{R}_\ell$ is nonincreasing, for all integer $q \in \{1, \ldots, p\}$ and all $n \geq 1$ we have

$$\min\left\{\ell = 1, \ldots, p : \hat{R}_\ell - \hat{R}_p \leq \delta_n\right\} \leq q \;\Leftrightarrow\; \hat{R}_q - \hat{R}_p \leq \delta_n. \tag{4.23}$$

Therefore, using the fact that $R_d = R_p$, we obtain

$$\begin{aligned}
\mathbb{P}\left(\hat{d} > d\right) &= \mathbb{P}\left(\min\left\{\ell = 1, \ldots, p : \hat{R}_\ell - \hat{R}_p \leq \delta_n\right\} > d\right) \\
&= \mathbb{P}\left(\hat{R}_d - \hat{R}_p > \delta_n\right) \\
&= \mathbb{P}\left(\left(\hat{R}_d - R_d\right) + \left(R_p - \hat{R}_p\right) > \delta_n\right) \\
&\leq \mathbb{P}\left(|\hat{R}_d - R_d| \geq \tfrac{\delta_n}{2}\right) + \mathbb{P}\left(|\hat{R}_p - R_p| \geq \tfrac{\delta_n}{2}\right).
\end{aligned}$$

Using Lemma 4.3, we deduce that for all $\vartheta > 0$ we have

$$\lim_{n\to+\infty} n^{\vartheta} \sup_{P\in\mathcal{D}} \mathbb{P}\left(|\hat{R}_d - R_d| \geq \tfrac{\delta_n}{2}\right) = 0,$$

and

$$\lim_{n\to+\infty} n^{\vartheta} \sup_{P\in\mathcal{D}} \mathbb{P}\left(|\hat{R}_p - R_p| \geq \tfrac{\delta_n}{2}\right) = 0,$$

which gives

$$\lim_{n\to+\infty} n^{\vartheta} \sup_{P\in\mathcal{D}} \mathbb{P}\left(\hat{d} > d\right) = 0. \tag{4.24}$$

Now let $\underline{\delta} > 0$ be fixed, let $P \in \mathcal{D}(\underline{\delta})$ and assume $d \geq 2$. Using the fact that $R_{d-1} - R_p = \Delta \geq \underline{\delta}$ and provided n is large enough to have $\underline{\delta} - \delta_n \geq \delta_n$, we obtain that

$$\begin{aligned}
\mathbb{P}\left(\hat{d} < d\right) &= \mathbb{P}\left(\min\left\{\ell = 1, \ldots, p : \hat{R}_\ell - \hat{R}_p \leq \delta_n\right\} \leq d - 1\right) \\
&= \mathbb{P}\left(\hat{R}_{d-1} - \hat{R}_p \leq \delta_n\right) \\
&= \mathbb{P}\left(\left(\hat{R}_{d-1} - R_{d-1}\right) + \Delta + \left(R_p - \hat{R}_p\right) \leq \delta_n\right) \\
&\leq \mathbb{P}\left(\left(R_{d-1} - \hat{R}_{d-1}\right) + \left(\hat{R}_p - R_p\right) \geq \underline{\delta} - \delta_n\right) \\
&\leq \mathbb{P}\left(\left(R_{d-1} - \hat{R}_{d-1}\right) + \left(\hat{R}_p - R_p\right) \geq \delta_n\right) \\
&\leq \mathbb{P}\left(|\hat{R}_{d-1} - R_{d-1}| \geq \tfrac{\delta_n}{2}\right) + \mathbb{P}\left(|\hat{R}_p - R_p| \geq \tfrac{\delta_n}{2}\right).
\end{aligned}$$

From the same argument as in the beginning of the proof, we have

$$\lim_{n\to+\infty} n^{\vartheta} \sup_{P\in\mathcal{D}(\underline{\delta})} \mathbb{P}\left(\hat{d} < d\right) = 0,$$

which concludes the proof. □

4.4 Proof of Theorem 3.4

Since we have

$$\inf\left\{\Delta : P \in \mathcal{D}_0\right\} = 0,$$

for all $\varepsilon > 0$ there exists $P(\varepsilon) \in \mathcal{D}_0$ such that $\Delta_{P(\varepsilon)} \leq \varepsilon$. For all $n \geq 1$ let $Q_n = P(\delta_n/2)$. Now assume that the sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ is drawn from Q_n. Then, as in the proof of Theorem 3.3, we obtain

$$\begin{aligned}
\mathbb{P}\left(\hat{d} < d\right) &= \mathbb{P}\left(\min\left\{\ell = 1, \ldots, p : \hat{R}_\ell - \hat{R}_p \leq \delta_n\right\} \leq d-1\right) \\
&= \mathbb{P}\left(\hat{R}_{d-1} - \hat{R}_p \leq \delta_n\right) \\
&= \mathbb{P}\left(\left(\hat{R}_{d-1} - R_{d-1}\right) + \Delta_{Q_n} + \left(R_p - \hat{R}_p\right) \leq \delta_n\right) \\
&\geq \mathbb{P}\left(\left(\hat{R}_{d-1} - R_{d-1}\right) + \tfrac{\delta_n}{2} + \left(R_p - \hat{R}_p\right) \leq \delta_n\right) \\
&= \mathbb{P}\left(\left(\hat{R}_{d-1} - R_{d-1}\right) + \left(R_p - \hat{R}_p\right) \leq \tfrac{\delta_n}{2}\right) \\
&\geq 1 - \mathbb{P}\left(\left(\hat{R}_{d-1} - R_{d-1}\right) + \left(R_p - \hat{R}_p\right) \geq \tfrac{\delta_n}{2}\right) \\
&\geq 1 - \mathbb{P}\left(|\hat{R}_{d-1} - R_{d-1}| \geq \tfrac{\delta_n}{4}\right) - \mathbb{P}\left(|\hat{R}_p - R_p| \geq \tfrac{\delta_n}{4}\right).
\end{aligned}$$

According to Lemma 4.3, we know that

$$\lim_{n\to+\infty} n\,\mathbb{P}\left(|\hat{R}_{d-1} - R_{d-1}| \geq \tfrac{\delta_n}{4}\right) = 0 \quad \text{and} \quad \lim_{n\to+\infty} n\,\mathbb{P}\left(|\hat{R}_p - R_p| \geq \tfrac{\delta_n}{4}\right) = 0.$$

As a result, there exists an integer n_0 such that for all $n \geq n_0$ we have

$$\mathbb{P}\left(|\hat{R}_{d-1} - R_{d-1}| \geq \tfrac{\delta_n}{4}\right) \leq \frac{1}{2n} \quad \text{and} \quad \mathbb{P}\left(|\hat{R}_p - R_p| \geq \tfrac{\delta_n}{4}\right) \leq \frac{1}{2n}.$$

Therefore, for all $n \geq n_0$, we have

$$\sup_{P\in\mathcal{D}_0} \mathbb{P}\left(\hat{d} < d\right) \geq \mathbb{P}_{(X,Y)\sim Q_n}\left(\hat{d} < d\right) \geq 1 - \frac{1}{n},$$

which concludes the proof. □

4.5 Proof of Theorem 3.5

Let $\underline{\delta} > 0$ and $P \in \mathcal{D}(\underline{\delta})$. We have

$$\begin{aligned}\mathbb{E}\left(\hat{r}(X) - r(X)\right)^2 &= \mathbb{E}\left[\mathbf{1}\left\{\hat{d} \neq d\right\}\left(\hat{r}(X) - r(X)\right)^2\right] \\ &\quad + \mathbb{E}\left[\mathbf{1}\left\{\hat{d} = d\right\}\left(\hat{r}(X) - r(X)\right)^2\right]. \quad (4.25)\end{aligned}$$

Since both $\hat{r}$ and r belong to $\mathcal{F}$, they are bounded by L and therefore we have

$$\mathbb{E}\left[\mathbf{1}\left\{\hat{d} \neq d\right\}\left(\hat{r}(X) - r(X)\right)^2\right] \leq 4L^2\mathbb{P}\left(\hat{d} \neq d\right). \quad (4.26)$$

Then, we observe that

$$\begin{aligned}\mathbb{E}\left[\mathbf{1}\left\{\hat{d} = d\right\}\left(\hat{r}(X) - r(X)\right)^2\right] &= \mathbb{E}\left[\mathbf{1}\left\{\hat{d} = d\right\}\left(\hat{r}_d(X) - r(X)\right)^2\right] \\ &\leq \mathbb{E}\left(\hat{r}_d(X) - r(X)\right)^2. \quad (4.27)\end{aligned}$$

Hence, denoting for simplicity

$$\upsilon_{n,\ell} = \left(\frac{(\ln n)^\alpha}{n}\right)^{\beta/(2\beta+\ell)},$$

for all $\ell \in \{1, \ldots, p\}$, we deduce from (4.25), (4.26) and (4.27) that

$$\begin{aligned}\upsilon_{n,d}^{-2}\,\mathbb{E}\left(\hat{r}(X) - r(X)\right)^2 &\leq 4L^2\,\upsilon_{n,d}^{-2}\,\mathbb{P}\left(\hat{d} \neq d\right) \\ &\quad + \upsilon_{n,d}^{-2}\,\mathbb{E}\left(\hat{r}_d(X) - r(X)\right)^2. \quad (4.28)\end{aligned}$$

According to the proof of Theorem 3.1, there exists a constant $C > 0$ depending only on τ, B, β, R and L such that

$$\upsilon_{n,d}^{-2}\,\mathbb{E}\left(\hat{r}_d(X) - r(X)\right)^2 \leq C. \quad (4.29)$$

In particular, since this constant C does not depend on $P \in \mathcal{D}$, we obtain

$$\sup_{P \in \mathcal{D}(\underline{\delta})} \upsilon_{n,d}^{-2}\,\mathbb{E}\left(\hat{r}_d(X) - r(X)\right)^2 \leq C. \quad (4.30)$$

Then, we deduce from Theorem 3.3 that

$$\sup_{P \in \mathcal{D}(\underline{\delta})} \upsilon_{n,d}^{-2}\,\mathbb{P}\left(\hat{d} \neq d\right) \leq \upsilon_{n,1}^{-2} \sup_{P \in \mathcal{D}(\underline{\delta})} \mathbb{P}\left(\hat{d} \neq d\right) \underset{n \to +\infty}{\to} 0. \quad (4.31)$$

Finally, it follows from (4.28), (4.30) and (4.31) that there exists a constant depending only on τ, B, β, R and L such that

$$\limsup_{n \to +\infty} \sup_{P \in \mathcal{D}(\underline{\delta})} \upsilon_{n,d}^{-2}\,\mathbb{E}\left(\hat{r}(X) - r(X)\right)^2 \leq C,$$

which concludes the proof. □

A Appendix

A.1 Reduced dimension d and parameter Δ

In this appendix, we prove equations (2.3) and (2.6). First, observe that since $\mathcal{F}_\ell$ is compact in $\mathbb{L}^2(\mu)$ and since $r \in \mathcal{F}$,

$$\begin{aligned} r \in \mathcal{F}_\ell &\Leftrightarrow \inf_{f\in\mathcal{F}_\ell} \mathbb{E}\left(f(X)-r(X)\right)^2 = 0 \\ &\Leftrightarrow \inf_{f\in\mathcal{F}_\ell} \mathbb{E}\left(Y-f(X)\right)^2 - \mathbb{E}\left(Y-r(X)\right)^2 = 0 \\ &\Leftrightarrow \inf_{f\in\mathcal{F}_\ell} \mathbb{E}\left(Y-f(X)\right)^2 - \inf_{f\in\mathcal{F}} \mathbb{E}\left(Y-f(X)\right)^2 = 0 \\ &\Leftrightarrow R_\ell = R_p. \end{aligned}$$

Therefore, since the function $\ell \in \{1,\dots,p\} \mapsto R_\ell$ is nonincreasing, we deduce that

$$\begin{aligned} d &:= \min\left\{\ell = 1,\dots,p : r \in \mathcal{F}_\ell\right\} \\ &= \min\left\{\ell = 1,\dots,p : R_\ell = R_p\right\}, \end{aligned}$$

which proves equation (2.3). Using (2.3), and the fact that $r \in \mathcal{F}$, we obtain

$$\begin{aligned} \Delta &= \min\left\{R_\ell - R_p :\ R_\ell > R_p\right\} \\ &= R_{d-1} - R_p \\ &= \inf_{f\in\mathcal{F}_{d-1}} \mathbb{E}\left(Y-f(X)\right)^2 - \inf_{f\in\mathcal{F}} \mathbb{E}\left(Y-f(X)\right)^2 \\ &= \inf_{f\in\mathcal{F}_{d-1}} \mathbb{E}\left(Y-f(X)\right)^2 - \mathbb{E}\left(Y-r(X)\right)^2 \\ &= \inf_{f\in\mathcal{F}_{d-1}} \mathbb{E}\left(f(X)-r(X)\right)^2 \\ &= \inf_{f\in\mathcal{F}_{d-1}} \int_{\mathcal{X}} (f-r)^2 \,\mathrm{d}\mu, \end{aligned}$$

which proves (2.6).

A.2 Performance of least-squares estimates

Let $\mathcal{X}$ be a metric space, let P be a probability measure on $\mathcal{X}\times\mathbb{R}$ and let (X,Z) be an $\mathcal{X}\times\mathbb{R}$-valued random variable. The regression function $f^\star$ of Z given X is defined for $x \in \mathcal{X}$ by

$$f^\star(x) := \mathbb{E}\left(Z|X=x\right). \tag{A.1}$$

In this appendix, we study the performance of the least-squares estimation of $f^\star$ based on a given class $\mathcal{F}$ of real functions defined on $\mathcal{X}$. For some $L > 0$, it will be assumed that each $f \in \mathcal{F}$ satisfies

$$\sup_{x\in\mathcal{X}} |f(x)| \le L. \tag{A.2}$$

Let $(X_1, Z_1), \ldots, (X_n, Z_n)$ be a sample of n i.i.d. random variables with same distribution P as (X, Z). The least-squares estimate f_n of $f^\star$ based on $\mathcal{F}$ is defined as any random element in $\mathcal{F}$ satisfying

$$f_n \in \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \left(Z_i - f(X_i)\right)^2 . \tag{A.3}$$

Implicitly, it will be assumed that such an element exists. The performance of f_n will be measured in terms of the mean squared error

$$\mathbb{E}\|f_n - f^\star\|^2 := \mathbb{E} \int_{\mathcal{X}} (f_n - f^\star)^2 \mathrm{d}\mu, \tag{A.4}$$

where μ stands for the distribution of X, and shown to be related to the metric entropy of $\mathcal{F}$. We are now in position to state the main result of this appendix.

Theorem A.1 (Koltchinskii, 2006)**.** *Suppose that $|Z| \leq T$ and that there exist two constants $A > 0$ and $0 < s < 2$ such that, for all $u > 0$, $H(u, \mathcal{F}) \leq Au^{-s}$. Then, there exists a constant C depending only on s and A such that, for all $\varepsilon \in (0, 1]$,*

$$\mathbb{E}\|f_n - f^\star\|^2 \leq (1+\varepsilon) \inf_{f \in \mathcal{F}} \|f - f^\star\|^2 + C\varepsilon^{-\frac{2-s}{2+s}} \left(\frac{b}{n}\right)^{\frac{2}{2+s}} + \frac{Cb}{\varepsilon n},$$

where $b := (T + L)^2$.

A detailed proof of Theorem A.1 may be found in the supplementary material.

Acknowledgments – The author is indebted to Benoit Cadre, Bruno Pelletier and Nicolas Klutchnikoff for many fruitful discussions and advices concerning the redaction of the paper. The author would also like to thank the Associate Editor as well as two anonymous referees for valuable comments that helped improve the original version of the manuscript.

Supplementary material – Supplement to: Minimax adaptive dimension reduction for regression.

References

P. Alquier and G. Biau. Sparse single-index model. *Journal of Machine Learning Research*, 14:243–280, 2013.

B. Cadre and Q. Dong. Dimension reduction in regression estimation with nearest neighbor. *Electronic Journal of Statistics*, 4:436–460, 2010.

R.D. Cook. *Regression Graphics: Ideas for Studying Regressions Through Graph- ics*. Wiley, New York, 1998.

R.D. Cook. Fisher lecture: Dimension reduction in regression. *Statistical Science*, 22:1–26, 2007.

R.D. Cook and B. Li. Dimension reduction for conditional mean in regression. *The Annals of Statistics*, 30:455–474, 2002.

R.D. Cook and L. Ni. Sufficient dimension reduction via inverse regression: a minimum discrepancy approach. *Journal of the American Statistical Association*, 100:410–428, 2005.

R.D. Cook and S. Weisberg. Discussion of sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86:316–342, 1991.

F. Cucker and S. Smale. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39:1–49, 2001.

B. Delyon and F. Portier. Optimal transformation: a new approach for covering the central subspace (to appear). *Journal of Multivariate Analysis*, 2013.

K. Fukumizu, F. Bach, and M. Jordan. Kernel dimension reduction in regression. *The Annals of Statistics*, 37:1871–1905, 2009.

W.K. Fung, X. He, L. Liu, and P. Shi. Dimension reduction based on canonical correlation. *Statistica Sinica*, 12:1093–1113, 2002.

L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer-Verlag, New York, 2002.

W. Härdle and T.M. Stoker. Investigating smooth multiple regression by the method of average derivative. *Journal of the American Statistical Association*, 84:986–995, 1989.

M. Kohler, A. Krzyżak, and H. Walk. Optimal global rates of convergence in nonparametric regression with unbounded data. *Journal of Statistical Planning and Inference*, 123:1286–1296, 2009.

V. Koltchinskii. Local rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34:2593–2656, 2006.

M.R. Kosorok. *Introduction to Empirical Processes and Semiparametric Inference*. Springer, New York, 2008.

K.C. Li. Sliced inverse regression for dimension reduction (with discussions). *Journal of the American Statistical Association*, 86:316–342, 1991.

K.C. Li. On principal hessian directions for data visualization and dimension reduction: another application of steins lemma. *Journal of the American Statistical Association*, 87:1025–1039, 1992.

G.G. Lorentz, M.v. Golitschek, and Y. Makovoz. *Constructive Approximation: Advanced Problems*. Springer-Verlag, New York, 1996.

C.J. Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10:1040–1053, 1982.

A.B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, New York, 2009.

Q. Wang and X. Yin. A nonlinear multi-dimensional variable selection method for high dimensional data: Sparse mave. *Computational Statistics and Data Analysis*, 52:4512–4520, 2008.

H.M. Wu. Kernel sliced inverse regression with applications on classification. *Journal of Computational and Graphical Statistics*, 17:590–610, 2008.

Y. Xia, H. Tong, W.K. Li, and L.-X. Zhu. An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society (B)*, 64:1–28, 2002.

Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics*, 27:1564–1599, 1999.

Y.-R. Yeh, S.-Y. Huang, and Y.-Y. Lee. Nonlinear dimension reduction with kernel sliced inverse regression. *IEEE Transactions on Knowledge and Data Engineering*, 21:1590–1603, 2009.

X. Yin, B. Li, and R.D. Cook. Successive direction extraction for estimating the central subspace in a multiple-index regression. *Journal of Multivariate Analysis*, 99:1733–1757, 2008.