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

Let (X,Y) be a random variable taking its values in $\mathscr{X} \times \mathbb{R}$, where $\mathscr{X} \subset \mathbb{R}^p$. We study the estimation of the regression function $r(x) := \mathbb{E}(Y|X=x)$ from a minimax point of view, assuming that r belongs to a class of functions $\mathscr{F}$ of the form

$$\mathscr{G} \circ \mathscr{H} := \Big\{ g \circ h : g \in \mathscr{G},\ h \in \mathscr{H} \Big\}.$$

Here, $\mathscr{G}$ is a class of β-Hölder functions $g : \mathbb{R}^p \to \mathbb{R}$ and $\mathscr{H}$ a class of functions $h : \mathscr{X} \to \mathbb{R}^p$ which need not be linear nor regular. We define the *reduced dimension* as the smallest ℓ for which $r = g \circ h$ for some $g \in \mathscr{G}$ and some $h \in \mathscr{H}$ such that $h(\mathscr{X})$ spans a subspace of dimension ℓ. In this context, we construct an adaptive estimate which converges at a rate depending only on the *reduced dimension* under entropy conditions on $\mathscr{H}$.

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

AMS 2000 Classification – 62H12, 62G08.

1 Introduction

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 $\mathscr{X}$-valued predictor variable X where $\mathscr{X} \subset \mathbb{R}^p$. 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 (X, Y).

When addressing this task, the statistician considers a set of assumptions on the underlying distribution of the observations based on some physical understanding of the phenomenon under study. Roughly speaking, this modelization step consists in specifying a class $\mathscr{D}$ of distributions of the random variable (X, Y) for which the regression function r belongs to some class $\mathscr{F}$ of functions $f : \mathscr{X} \to \mathbb{R}$. Given such a regression model, one usually refers to optimal rates of convergence in the minimax sense as benchmarks to assess the performance of a particular estimate. In this context, optimal rates of convergence may be defined as follows. The sequence (υ_n) is said to be an optimal rate of convergence in the minimax sense if it is a lower minimax rate, *i.e.* if

$$\liminf_{n\to+\infty} \upsilon_n^{-2} \inf_{\hat{r}} \sup_{P\in\mathscr{D}} \mathbb{E}_{(X,Y)\sim P} \left(\hat{r}(X) - \mathbb{E}(Y|X)\right)^2 > 0, \tag{1.2}$$

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

$$\limsup_{n\to+\infty} \upsilon_n^{-2} \sup_{P\in\mathscr{D}} \mathbb{E}_{(X,Y)\sim P} \left(\hat{r}(X) - \mathbb{E}(Y|X)\right)^2 < +\infty. \tag{1.3}$$

In (1.2) and (1.3), we have used the subscript $(X, Y) \sim P$ to stress that (X, Y) is drawn from P. When no confusion may arise, we will omit it in the rest of the paper.

A major problem, known as the *curse of dimensionality*, is that for most models used across the literature, the optimal rates of convergence depend on the dimension p of the predictor X. For instance, suppose $\mathscr{F}$ is the class of all $[0, 1]$-valued, β-Hölder functions defined on the open unit Euclidian ball in $\mathbb{R}^p$. Then

it can be shown that, under some technical conditions, the optimal rate of convergence in this context is $n^{-\beta/(2\beta+p)}$ (see *e.g.* Theorem 3.2 in Györfi *et al.* (2002) and Theorem 1 in Kolher *et al.* (2009)). Hence, when the dimension p is large, the optimal rate of convergence is slow. To face this drawback when dealing with high-dimensional data, one usually considers a model which encodes so called *structural assumptions* in addition to reasonable regularity conditions, thus reducing the overall complexity of the model. Many such models have been studied and proved effective in practice among which we mention additive regression models, projection pursuit, single index models (see *e.g.* Chapter 22 in the book by Györfi *et al.* (2002) and the references cited therein) and the body of theory and methods known as sufficient dimension reduction (see *e.g.* the foundational article of Li (1991) which introduces the SIR method or the article of Cook and Li (2002) which defines the important notion of central mean subspace).

In the present article we study a regression model with a structural assumption inspired by sufficient dimension reduction methodology. Precisely, we assume that the regression function belongs to a class $\mathscr{F}$ of the form

$$\mathscr{G} \circ \mathscr{H} := \big\{ g \circ h : g \in \mathscr{G},\ h \in \mathscr{H} \big\}, \tag{1.4}$$

where $\mathscr{G}$ denotes a class of functions $g : \mathbb{R}^p \to \mathbb{R}$ and where $\mathscr{H}$ denotes a class of functions $h : \mathscr{X} \to \mathbb{R}^p$. In this context, having $r \in \mathscr{F}$ means that the two conditions

$$(i)\ \ \mathbb{E}(Y|h(X)) = \mathbb{E}(Y|X), \quad \text{and} \quad (ii)\ \ \mathbb{E}(Y|h(X) = .) \in \mathscr{G},$$

are satisfied for at least one function $h \in \mathscr{H}$. From a heuristic point of view, the main interest here is that condition (i) should allow to replace the high dimensional predictor X by a possibly low dimensional predictor $h(X)$. In this approach, it is expected that the rates of convergence depend on the dimension of $h(X)$ rather than on the dimension of X. Note that condition (i) generalizes the usual assumption made in sufficient dimension reduction where one assumes that h is a matrix. In the present paper the functions in $\mathscr{H}$ need not be linear nor regular. The class $\mathscr{G}$ is a class of regular functions the construction of which is detailed in the next section. In the classical sufficient dimension reduction context, a similar approach has been successfully applied by Cadre and Dong (2010) to obtain fast rates of convergence.

From a technical point of view, the impact of dimensionality on rates of convergence can be explained as follows. A first important fact is that optimal rates

of convergence usually depend on a measure of the complexity of $\mathscr{F}$. Suppose that the marginal distribution μ of X is fixed and that the complexity of $\mathscr{F}$ is measured in terms of the ε-metric entropy of $\mathscr{F}$ in $\mathbb{L}^2(\mu)$ denoted $H(\varepsilon,\mathscr{F},\mu)$ (i.e. the logarithm of the minimal number of $\mathbb{L}^2(\mu)$-balls of radius ε that are needed to cover $\mathscr{F}$). Then, if there exists $c>0$ such that $H(\varepsilon,\mathscr{F},\mu)\geq c\varepsilon^{-s}$, the optimal rates of convergence are lower bounded by $n^{-1/(2+s)}$ under some technical assumptions (see *e.g.* Theorem 6 in Yang and Barron (1999)). Suppose this time the complexity of $\mathscr{F}$ is measured in terms of $H(\varepsilon,\mathscr{F}):=\sup H(\varepsilon,\mathscr{F},Q)$, where the supremum is taken over all probability measures with finite support in $\mathscr{X}$. Then if there exists $C>0$ such that $H(\varepsilon,\mathscr{F})\leq C\varepsilon^{-s}$, the optimal rates of convergence are upper bounded by $n^{-1/(2+s)}$ under some technical assumptions (see Appendix A). The second important fact is that the value of the exponent s usually depends on the dimension p of the predictor X. For example, when $\mathscr{F}$ is the class of all $[0,1]$-valued, β-Hölder functions defined on the open unit Euclidian ball in $\mathbb{R}^p$, the typical value for the exponent s is p/β which explains that the optimal rate in this context is $n^{-1/(2+s)}=n^{-\beta/(2\beta+p)}$. In the case of a model as defined in (1.4), we will use the composite structure to get bounds on the entropy of $\mathscr{F}$ which depend only the regularity of the functions in $\mathscr{G}$ and the entropy of $\mathscr{H}$.

The paper is organized as follows. In Section 2, we describe precisely our model and the assumptions made on the classes $\mathscr{G}$ and $\mathscr{H}$. In Section 3, we fix an integer $\ell\in\{1,\dots,p\}$ and derive bounds for the optimal rates in the context of our composite model when all functions $h\in\mathscr{H}$ satisfy $\dim S(h)\leq\ell$ where $S(h)$ denotes the subspace of $\mathbb{R}^p$ spanned by $h(\mathscr{X})$. In Section 4 we present our main result, namely Theorem 4.3, which studies the general case where no restriction is made on the dimension of $S(h)$ for $h\in\mathscr{H}$. In this case, using the results of Section 3, we construct an estimate $\hat{r}$ of the regression function which adapts to the *reduced dimension* d defined as the smallest ℓ for which one has $r=g\circ h$ for some $g\in\mathscr{G}$ and some $h\in\mathscr{H}$ with $\dim S(h)\leq\ell$. More precisely, we prove that uniformly over a certain class of distributions, our estimate converges to r at a rate depending only on the reduced dimension d. Furthermore, this rate corresponds to the rate one would obtain if the true value of the reduced dimension where known as exposed in Section 3. We have reported in the appendix a useful result concerning rates of convergence for least-squares estimates in the case of bounded regression. The proof follows from the lines devised in Koltchinskii (2006) and is available in the supplementary material.

2 Definition of the model

Let (X,Y) be an $\mathscr{X}\times\mathbb{R}$-valued random variable of distribution P where $\mathscr{X}\subset\mathbb{R}^p$. We denote μ the distribution of X and for $x\in\mathscr{X}$ we set

$$r(x) := \mathbb{E}(Y|X=x). \tag{2.1}$$

Let $\mathscr{F}$ be a class of functions $f:\mathscr{X}\to\mathbb{R}$ of the form

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

where $\mathscr{G}$ and $\mathscr{H}$ are taken as follows. Let $R>0$. For all $h\in\mathscr{H}$ we assume that

$$\|h\|_{\mathscr{X}} := \sup_{x\in\mathscr{X}} \|h(x)\| < R, \tag{2.2}$$

where $\|.\|$ stands for the Euclidean norm in $\mathbb{R}^p$. Let $\mathbf{B}$ be the open Euclidean ball in $\mathbb{R}^p$ with center the origin and radius R. Let $\beta>0$. We denote $\mathscr{C}^{\lfloor\beta\rfloor}(\mathbf{B})$ the space of $\lfloor\beta\rfloor$-times continuously differentiable functions from $\mathbf{B}$ to $\mathbb{R}$, where $\lfloor\beta\rfloor$ stands for the greatest integer strictly smaller than β. For all g in $\mathscr{C}^{\lfloor\beta\rfloor}(\mathbf{B})$ we set

$$\|g\|_\beta := \max_{|s|\le\lfloor\beta\rfloor} \|\partial^s g\|_\infty + \max_{|s|=\lfloor\beta\rfloor} \sup_{u\neq u'} \frac{|\partial^s g(u)-\partial^s g(u')|}{\|u-u'\|^{\beta-\lfloor\beta\rfloor}}, \tag{2.3}$$

where for every multi-index $s=(s_1,\dots,s_p)\in\mathbb{N}^p$, we have denoted $|s| := \sum_i s_i$ and $\partial^s := \partial_1^{s_1}\cdots\partial_p^{s_p}$. For $L>0$, the class $\mathscr{G}$ is defined by

$$\mathscr{G} := \Big\{ g\in\mathscr{C}^{\lfloor\beta\rfloor}(\mathbf{B}) : \|g\|_\beta \le L \Big\}. \tag{2.4}$$

In the sequel, we suppose that the distribution P of (X,Y) belongs to the following class of distributions.

Class $\mathscr{D}$ of distributions. Let $\tau>0$ and $B>0$ be fixed. Let the marginal distribution μ be fixed. We denote $\mathscr{D}$ the class of distributions P of the random variable (X,Y) such that X is of distribution μ, such that the regression function r belongs to $\mathscr{F}$ and such that Y satisfies the exponential moment condition

$$\mathbb{E}\, e^{\tau|Y|} \le B. \tag{2.5}$$

3 Non adaptive case

Fix $\ell \in \{1,\dots,p\}$. In this section, we derive bounds on the optimal rates of convergence associated to the class of distributions $\mathscr{D}$ under the assumption that all functions $h \in \mathscr{H}$ satisfy

$$\dim S(h) \leq \ell, \tag{3.1}$$

where $S(h)$ denotes the subspace of $\mathbb{R}^p$ spanned by $h(\mathscr{X})$. Let $(X_1,Y_1),\dots,(X_n,Y_n)$ be a sample of n independent and identically distributed random variable with same distribution P as (X,Y).

First we present an upper bound. For any Borel measure Q on $\mathscr{X}$, we denote by $\mathbb{L}^2(Q)$ the set of measurable functions $f : \mathscr{X} \to \mathbb{R}$ such that

$$\|f\|_Q^2 := \int_{\mathscr{X}} |f|^2 \mathrm{d}Q < +\infty. \tag{3.2}$$

For a class $\mathscr{C} \subset \mathbb{L}^2(Q)$, we recall that the ε-covering number of $\mathscr{C}$ with respect to $\|.\|_Q$, denoted $N(\varepsilon,\mathscr{C},Q)$, is the minimal number of $\|.\|_Q$-balls of radius ε that are needed to cover $\mathscr{C}$. The ε-metric entropy of $\mathscr{C}$ with respect to $\|.\|_Q$ is defined by $H(\varepsilon,\mathscr{C},Q) := \ln N(\varepsilon,\mathscr{C},Q)$. In this paper, we set

$$H(\varepsilon,\mathscr{C}) := \sup_Q H(\varepsilon,\mathscr{C},Q), \tag{3.3}$$

where the supremum is taken over all probability measures Q with finite support in $\mathscr{X}$. These notations will also be used for classes of vector valued functions. In this case, it will be understood that the absolute value $|.|$ in (3.2) is replaced by the appropriate Euclidean norm.

We fix a constant $\alpha > 2$ and define the least-squares estimate $\hat{f}$ as any element in $\mathscr{F}$ satisfying

$$\hat{f} \in \arg\min_{f\in\mathscr{F}} \frac{1}{n}\sum_{i=1}^{n} \Big(Y_i \mathbf{1}\{|Y_i| \leq T_n\} - f(X_i)\Big)^2, \tag{3.4}$$

where $T_n := (\ln n)^{\alpha/2}$. Implicitly, it will be assumed that such an element exists, though it might not be unique. We denote

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

Theorem 3.1. *Suppose that $\beta \geq 1$, that $\beta > \ell/2$ and suppose that there exists $C > 0$ and $0 < s \leq \ell/\beta$ such that for all $\varepsilon > 0$ we have*

$$H(\varepsilon, \mathscr{H}) \leq C\varepsilon^{-s}. \tag{3.5}$$

Then we have

$$\limsup_{n\to+\infty} \upsilon_{n,\ell}^{-2} \sup_{P\in\mathscr{D}} \mathbb{E}\left(\hat{f}(X) - \mathbb{E}(Y|X)\right)^2 < +\infty. \tag{3.6}$$

An immediate consequence of Theorem 3.1 is that, up to a logarithmic factor, the optimal rate of convergence associated to $\mathscr{D}$ is upper bounded by $n^{-\beta/(2\beta+\ell)}$ which depends only on ℓ and not anymore on the dimension p of the predictor X. Note that, under some technical conditions, $n^{-\beta/(2\beta+\ell)}$ corresponds to the optimal rate of convergence for the estimation of a β-Hölder function in the case where X is ℓ-dimensional (see *e.g.* Theorem 3.2 in Györfi *et al.* (2002) and Theorem 1 in Kolher *et al.* (2009)). If the exponential moment condition (2.5) is replaced by a boundedness assumption on Y, then a slight modification of the proof reveals that Theorem 3.1 holds with $\upsilon_{n,\ell}$ replaced by $n^{-\beta/(2\beta+\ell)}$.

Example 3.2. *An example where condition* (3.5) *is fulfilled is the following. Consider the case where $\mathscr{H}$ is a parametric class of the form*

$$\mathscr{H} = \left\{h_\theta : \theta \in \Theta \subset \mathbb{R}^k\right\}, \tag{3.7}$$

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

$$\|h_\theta - h_{\theta'}\|_{\mathscr{X}} \leq C|\theta - \theta'|, \tag{3.8}$$

where $|.|$ stands for the Euclidian norm in $\mathbb{R}^k$. Then, for all $\varepsilon > 0$ we have $H(\varepsilon, \mathscr{H}) \leq H(\varepsilon/C, \Theta)$ 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 $\varepsilon > 0$ we have

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

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

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

and condition (3.5) *is satisfied.*

Now we give a lower bound for the optimal rates of convergence. The proof of this result uses arguments from Yang and Barron (1999).

Theorem 3.3. *Suppose that $\beta > 0$. Suppose that there exists $h \in \mathscr{H}$ with $\dim S(h) = \ell$ and a constant $c > 0$ such that*

$$\mu \circ h^{-1} \geq c\, \lambda_h (. \cap \mathbf{B}), \tag{3.9}$$

where λ_h denotes the Lebesgue measure in $S(h)$. Then, we have

$$\liminf_{n \to +\infty} n^{2\beta/(2\beta+\ell)} \inf_{\hat{f}} \sup_{P \in \mathscr{D}} \mathbb{E}\left(\hat{f}(X) - \mathbb{E}(Y|X)\right)^2 > 0,$$

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

The condition in (3.9) is not restrictive. As an example, it is satisfied if $\mathscr{X} = \mathbf{B}$, 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 $\mathscr{H}$ and if μ has a density with respect to the Lebesgue measure which is lower bounded by a positive constant on $\mathbf{B}$. Theorem 3.3 shows that the optimal rate associated to $\mathscr{D}$ is lower bounded by $n^{-\beta/(2\beta+\ell)}$ which, up to a logarithmic factor, corresponds to the upper bound on the optimal rate given in Theorem 3.1.

4 Dimension adaptivity

In Section 3 we have studied the case where all functions $h \in \mathscr{H}$ satisfy $\dim S(h) \leq \ell$ for some $\ell \in \{1, \ldots, p\}$. In this section, we impose no restriction on the dimension of $S(h)$ for $h \in \mathscr{H}$. In Theorem 4.1 we study an estimate of the *reduced dimension* defined hereafter. Based on this estimate, we construct an estimate of the regression function and show in our main result, Theorem 4.3, that it adapts to the reduced dimension.

First we need some notations. For all $\ell \in \{1, \ldots, p\}$, let $\mathscr{H}_\ell$ be the class of all functions $h \in \mathscr{H}$ such that $S(h)$ is at most ℓ-dimensional, *i.e.*

$$\mathscr{H}_\ell := \Big\{ h \in \mathscr{H} : \dim\ S(h) \leq \ell \Big\}. \tag{4.1}$$

We define the subset $\mathscr{F}_\ell$ of $\mathscr{F}$ by

$$\mathscr{F}_\ell := \Big\{ g \circ h : g \in \mathscr{G},\ h \in \mathscr{H}_\ell \Big\}. \tag{4.2}$$

The $\mathscr{F}_\ell$'s form a nested family of models, that is $\mathscr{F}_1 \subset \mathscr{F}_2 \subset \cdots \subset \mathscr{F}_p = \mathscr{F}$. The *reduced dimension* associated to the underlying distribution $P \in \mathscr{D}$ of our observations is defined by

$$d_P \;:=\; \min\Big\{\ell : r \in \mathscr{F}_\ell\Big\}. \tag{4.3}$$

When no confusion may arise, we will drop the dependency on P in the notation. Our first task will be to derive a tractable representation of the reduced dimension, suitable for estimation purposes. We shall use the following assumption.

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

Let R_ℓ be the risk defined by

$$R_\ell := \inf_{f\in\mathscr{F}_\ell} \mathbb{E}(Y - f(X))^2. \tag{4.4}$$

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

$$d = \min\Big\{\ell : R_\ell = R_p\Big\}. \tag{4.5}$$

Consequently, for all $0 < \delta < \Delta$, we have

$$d = \min\Big\{\ell = 1,\ldots,p : R_\ell \le R_p + \delta\Big\}, \tag{4.6}$$

where Δ is defined as

$$\Delta := \min\Big\{\; R_\ell - R_p :\; R_\ell > R_p \;\Big\}, \tag{4.7}$$

with the convention $\min\emptyset = +\infty$. Observe that $\Delta > 0$ and that, when $d \ge 2$, Δ corresponds to the distance from r to $\mathscr{F}_{d-1}$ in $\mathbb{L}^2(\mu)$, that is

$$\Delta = \inf_{f\in\mathscr{F}_{d-1}} \|f - r\|_\mu^2. \tag{4.8}$$

Equations (4.5) and (4.8) are proved in Appendix B.

Based on (4.6), one may construct a natural estimate of d. First we introduce the empirical counterpart of the risk R_ℓ defined in (4.4) by setting

$$\hat{R}_\ell := \inf_{f\in\mathscr{F}_\ell} \frac{1}{n}\sum_{i=1}^n \Big(Y_i \mathbf{1}\{|Y_i| \le a_n\} - f(X_i)\Big)^2,$$

where (a_n) denotes an increasing sequence of positive numbers. Then, given a decreasing sequence of positive numbers (b_n), we define the estimate

$$\hat{d} := \min\Big\{\ell = 1,\dots,p : \hat{R}_\ell \leq \hat{R}_p + b_n\Big\}. \tag{4.9}$$

The asymptotic behavior of $\hat{d}$ is given by the following result.

Theorem 4.1. *Suppose that:*

(1) $\beta \geq 1$,

(2) *There exists* $C > 0$ *and* $0 < s \leq p/\beta$ *such that* $H(\varepsilon, \mathscr{H}) \leq C\varepsilon^{-s}$ *for all* $\varepsilon > 0$,

(3) *Assumption* $(\mathbf{A})$ *is satisfied.*

Suppose that $a_n = n^u$ *and* $b_n = n^{-u'}$ *where* $u > 0$, *where* $u' \geq 0$ *and where*

$$0 < 2u + u' < \frac{2\beta}{2\beta + p}.$$

Then the following two statements hold:

(i) *For all* $\vartheta > 0$ *we have*

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

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

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

where we have denoted $\mathscr{D}(\underline{\delta}) := \big\{P \in \mathscr{D} : \Delta \geq \underline{\delta}\big\}$.

Theorem 4.1 reveals that for all $\underline{\delta} > 0$, the estimate $\hat{d}$ converges exponentially fast to d, uniformly for all $P \in \mathscr{D}(\underline{\delta})$. This is not the case for $\underline{\delta} = 0$, as shown by the following theorem.

Theorem 4.2. *Let* $\mathscr{D}_0 \subset \mathscr{D}$ *be an arbitrary subset of* $\mathscr{D}$ *such that*

$$\inf\big\{\Delta : P \in \mathscr{D}_0\big\} = 0.$$

Under the conditions of Theorem 4.1 and for all $\vartheta > 0$ *we have*

$$\sup_{P\in\mathscr{D}_0} \mathbb{P}\Big(\hat{d} < d\Big) \geq 1 - \frac{1}{n^{\vartheta}},$$

provided n is large enough.

Now we apply these results to the estimation of the regression function. First, for all $\ell \in \{1,\dots,p\}$, let $\hat{f}_\ell$ be defined as any element in $\mathscr{F}_\ell$ satisfying

$$\hat{f}_\ell \in \operatorname*{arg\,min}_{f\in\mathscr{F}_\ell} \frac{1}{n}\sum_{i=1}^{n}\Big(Y_i \mathbf{1}\{|Y_i| \le T_n\} - f(X_i)\Big)^2, \tag{4.10}$$

where T_n is defined as in (3.4). Our next result proves that the estimate $\hat{f}$ defined by

$$\hat{f} := \hat{f}_{\hat{d}}, \tag{4.11}$$

adapts to the reduced dimension d in the sense that for all $\underline{\delta} > 0$ and uniformly for all $P \in \mathscr{D}(\underline{\delta})$ it converges to r at a rate depending only on d.

Theorem 4.3. *Suppose that:*

$(1')$ $\beta \ge 1$ *and* $\beta > p/2$,

$(2')$ *For all* $\ell \in \{1,\dots,p\}$, *there exists* $C > 0$ *and* $0 < s \le \ell/\beta$ *such that for all* $\varepsilon > 0$ *we have* $H(\varepsilon, \mathscr{H}_\ell) \le C\varepsilon^{-s}$,

(3) *Assumption* $(\mathbf{A})$ *is satisfied.*

Suppose that $a_n = n^u$ *and* $b_n = n^{-u'}$ *where* $u > 0$, *where* $u' \ge 0$ *and where*

$$0 < 2u + u' < \frac{2\beta}{2\beta + p}.$$

Then for all $\underline{\delta} > 0$ *we have*

$$\limsup_{n\to+\infty} \sup_{P\in\mathscr{D}(\underline{\delta})} \upsilon_{n,d}^{-2}\, \mathbb{E}\left(\hat{f}(X) - \mathbb{E}(Y|X)\right)^2 < +\infty,$$

where it is understood that $d = d_P$.

Theorem 4.3 shows that the estimate $\hat{f}$ defined in (4.11) converges to r as fast as the estimate $\hat{f}_d$ one would choose if the true value of the reduced dimension d where known as exposed in Section 3. Therefore, for all $\underline{\delta} > 0$, the estimate $\hat{f}$ adapts to the reduced dimension d, uniformly over all $P \in \mathscr{D}(\underline{\delta})$. Note that for condition $(2')$ to be satisfied for all $\ell \in \{1,\dots,p\}$, it is sufficient to impose that $H(\varepsilon, \mathscr{H}) \le C\varepsilon^{-s}$ for some $C > 0$ and some $0 < s \le 1/\beta$. For example, this condition is satisfied when $\mathscr{H}$ is chosen as in Example (3.2).

The condition $\beta > p/2$ in Theorem 4.3 allows adaptation for all values of $d \in \{1,\dots,p\}$. We do not know whether this result holds for $\beta \le p/2$. That being

said, if adaptation is required only for small dimensions we have the following result. For all $\beta \geq 1$ and all $\underline{\delta} > 0$, by considering the smaller class

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

where

$$\ell_\beta := [2\beta] - 1,$$

and where $[x]$ stands for the greatest integer smaller or equal to x, we obtain readily that

$$\limsup_{n \to +\infty} \sup_{P \in \mathscr{D}(\underline{\delta}, \beta)} \upsilon_{n,d}^{-2} \, \mathbb{E}\left(\hat{f}(X) - \mathbb{E}(Y|X) \right)^2 < +\infty.$$

Finaly, note that if the exponential moment condition (2.5) is replaced by a boundedness assumption on Y, then Theorem 4.3 holds with $\upsilon_{n,d}$ replaced by $n^{-\beta/(2\beta+d)}$.

5 Proofs

5.1 Proof of Theorem 3.1

Lemma 5.1. *Assume $\beta \geq 1$. Then for all $\varepsilon > 0$ we have*

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

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

$$N := N\left(\tfrac{\varepsilon}{2L}, \mathscr{H}, Q \right),$$

and choose an $\frac{\varepsilon}{2L}$-covering of $\mathscr{H}$ with respect to $\|.\|_Q$ of minimum cardinality $\{h_1, \ldots, h_N\}$. For all $i \in \{1, \ldots, N\}$, let

$$N_i := N\left(\tfrac{\varepsilon}{2}, \mathscr{G} \circ h_i, Q \right),$$

and let $\{g_1^i \circ h_i, \ldots, g_{N_i}^i \circ h_i\}$ be an $\frac{\varepsilon}{2}$-covering of $\mathscr{G} \circ h_i$ with respect to $\|.\|_Q$. Then, let $g \in \mathscr{G}$ and $h \in \mathscr{H}$ be chosen arbitrarily. By definition, there exists $i \in \{1, \ldots, N\}$ and $j \in \{1, \ldots, N_i\}$ 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, we have

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

According to the mean value Theorem, all $g \in \mathscr{G}$ are L-Lipschitz. Therefore, we have

$$\begin{aligned}
\sqrt{\int \left(g\circ h - g_j^i \circ h_i\right)^2 \mathrm{d}Q} &\leq \sqrt{\int (g\circ h - g\circ h_i)^2 \mathrm{d}Q} + \frac{\varepsilon}{2} \\
&\leq L\sqrt{\int \|h - h_i\|^2 \mathrm{d}Q} + \frac{\varepsilon}{2} \\
&\leq \varepsilon. \qquad (5.1)
\end{aligned}$$

As a result, we have obtained that

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

which implies that

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

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

Lemma 5.2. *Suppose $\beta \geq 1$. Suppose there exists a constant $C > 0$ and $0 < s \leq \ell/\beta$ such that for all $\varepsilon > 0$ we have $H(\varepsilon, \mathscr{H}) \leq C\varepsilon^{-s}$. Then, there exists a constant $A > 0$ depending only on ℓ, β, L and R such that for all $\varepsilon > 0$ we have*

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

Proof –According to Lemma 5.1, we only need to prove that there exists a constant $A > 0$ depending only on ℓ, β, L and R such that for all $\varepsilon > 0$ we have

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

To that aim, fix a probability measure Q with support in $\mathscr{X}$ and $h \in \mathscr{H}$. For all $g \in \mathscr{G}$, let g_h be the restriction of g to $S(h) \cap \mathbf{B}$ and let $\mathscr{G}_h := \{g_h : g \in \mathscr{G}\}$. Then, from the transfer theorem, for all $\varepsilon > 0$ we have

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

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 radius R and center the origin. Also, $Q \circ h^{-1}$ may be seen as of support in $\mathbf{B}_{\ell'}$ and $\mathscr{G}_h$ as a subset of

$$\mathscr{G}_{\ell'} := \left\{ g \in \mathscr{C}^{\lfloor \beta \rfloor}(\mathbf{B}_{\ell'}) : \ \|g\|_\beta \leq L \right\},$$

where here $\|.\|_\beta$ is defined as in (2.3) but with the Euclidean norm in $\mathbb{R}^{\ell'}$. According to Theorem 9.19 in Kosorok (2008), we know that there exists a constant K' depending only on ℓ', β, R, and L such that for all $\varepsilon > 0$ we have

$$\sup_D H(\varepsilon, \mathscr{G}_{\ell'}, D) \leq K' \varepsilon^{-\ell'/\beta},$$

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

$$\sup_D H(\varepsilon, \mathscr{G}_{\ell'}, D) \leq K \varepsilon^{-\ell/\beta},$$

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

$$H(\varepsilon, \mathscr{G} \circ h, Q) \leq K\varepsilon^{-\ell/\beta}.$$

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

$$\sup_{h \in \mathscr{H}} H(\varepsilon, \mathscr{G} \circ h) \leq K\varepsilon^{-\ell/\beta},$$

which completes the proof. □

Proof of Theorem 3.1

Let $P \in \mathscr{D}$. 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

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

We have

$$\mathbb{E}\left(\hat{f}(X) - \mathbb{E}(Y|X)\right)^2 \leq 2\mathbb{E}\left(\hat{f}(X) - r_n(X)\right)^2 + 2\mathbb{E}\left(r_n(X) - \mathbb{E}(Y|X)\right)^2. \quad (5.2)$$

Using Theorem A.1 of the Appendix with $\varepsilon := 1$, $Z := Y\mathbf{1}\left\{|Y| \leq T_n\right\}$ and Lemma 5.2, we deduce that there exists a constant $C > 0$ such that

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

where in the second inequality we have used the fact that $r \in \mathscr{F}$. Since T_n goes to $+\infty$ and T_n^2/n goes to 0 as n goes to $+\infty$, we deduce that there exists a constant $C > 0$ and an integer n_0 depending only on L, β and ℓ such that $n \geq n_0$ implies that

$$\mathbb{E}\left(\hat{f}(X) - r_n(X)\right)^2 \leq 2\mathbb{E}\left(r_n(X) - \mathbb{E}(Y|X)\right)^2 + C\left(\frac{T_n^2}{n}\right)^{2\beta/(2\beta+\ell)}. \quad (5.3)$$

We deduce from (5.2) and (5.3) that there exists a constant $C > 0$ such that for all $n \geq n_0$ we have

$$\mathbb{E}\left(\hat{f}(X) - \mathbb{E}(Y|X)\right)^2 \leq 6\mathbb{E}\left(r_n(X) - \mathbb{E}(Y|X)\right)^2 + C\left(\frac{T_n^2}{n}\right)^{2\beta/(2\beta+\ell)}. \quad (5.4)$$

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

$$\begin{aligned}
\mathbb{E}\left(r_n(X) - \mathbb{E}(Y|X)\right)^2 &= \mathbb{E}\left[\mathbb{E}\left(Y\mathbf{1}\{|Y| \leq T_n\}|X\right) - \mathbb{E}\left(Y|X\right)\right]^2 \\
&= \mathbb{E}\left[\mathbb{E}\left(Y\mathbf{1}\{|Y| > T_n\}|X\right)\right]^2 \\
&\leq \mathbb{E}\left(Y^2\mathbf{1}\{|Y| > T_n\}\right) \\
&\leq \sqrt{\mathbb{E}Y^4}\sqrt{\mathbb{P}\left(|Y| > T_n\right)}. \quad (5.5)
\end{aligned}$$

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

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

we deduce from the exponential moment condition (2.5) that

$$\mathbb{E}Y^4 \leq \frac{4!}{\tau^4}\mathbb{E}e^{\tau|Y|} \leq \frac{24B}{\tau^4}. \tag{5.6}$$

Using Markov's inequality and the exponential moment condition (2.5), we obtain that

$$\mathbb{P}\left(|Y| > T_n\right) = \mathbb{P}\Big(e^{\tau|Y|} > e^{\tau T_n}\Big) \leq Be^{-\tau T_n}. \tag{5.7}$$

Combining (5.6) and (5.7) we deduce from (5.5) that

$$\mathbb{E}\left(r_n(X) - \mathbb{E}(Y|X)\right)^2 \leq \frac{\sqrt{24}B}{\tau^2}e^{-\tau T_n/2}. \tag{5.8}$$

Equations (5.4) and (5.8) imply that there exists a constant $C > 0$ such that for all $n \geq n_0$ we have

$$\mathbb{E}\left(\hat{f}(X) - \mathbb{E}(Y|X)\right)^2 \leq \frac{6\sqrt{24}B}{\tau^2}e^{-\tau T_n/2} + C\left(\frac{T_n^2}{n}\right)^{2\beta/(2\beta+\ell)}. \tag{5.9}$$

Since the integer n_0 and the constants involved on the right hand side of (5.9) do not depend on $P \in \mathscr{D}$, we deduce that there exists a constant $C > 0$ such that for all $n \geq n_0$ we have

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

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

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

It follows that

$$\limsup_{n\to+\infty} \upsilon_{n,\ell}^{-2} \sup_{P\in\mathscr{D}} \mathbb{E}\left(\hat{f}(X) - \mathbb{E}(Y|X)\right)^2 < +\infty.$$

The proof is complete. □

5.2 Proof of Theorem 3.3

Let $\mathscr{D}^\circ$ be the class of distributions P of (X,Y) where X is of distribution μ, where

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

for $f \in \mathscr{F}$ and where ξ is independent from X and with distribution $\mathscr{N}(0,\sigma^2)$. It may be easily verified that the exponential moment condition (2.5) holds in this context so that we have $\mathscr{D}^\circ \subset \mathscr{D}$. Therefore, it is clear that

$$\inf_{\hat{f}} \sup_{P\in\mathscr{D}^\circ} \mathbb{E}\left(\hat{f}(X) - \mathbb{E}(Y|X)\right)^2 \leq \inf_{\hat{f}} \sup_{P\in\mathscr{D}} \mathbb{E}\left(\hat{f}(X) - \mathbb{E}(Y|X)\right)^2.$$

As a result, in order to prove Theorem 3.3 we need only to prove that

$$\liminf_{n\to+\infty} n^{2\beta/(2\beta+\ell)} \inf_{\hat{f}} \sup_{P\in\mathscr{D}^\circ} \mathbb{E}\left(\hat{f}(X) - \mathbb{E}(Y|X)\right)^2 > 0. \tag{5.10}$$

According to Theorem 6 in Yang and Barron (1999), inequality (5.10) is satisfied provided there exists a lower bound $N(\varepsilon)$ for the covering number $N(\varepsilon,\mathscr{F},\mu)$ such that any ε_n satisfying

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

is of order $n^{-\beta/(2\beta+\ell)}$. To obtain such a lower bound, let $h \in \mathscr{H}$ satisfying the conditions of Theorem 3.3. Then, we have $\left\{g\circ h : g \in \mathscr{G}\right\} \subset \mathscr{F}$ which implies that for all $\varepsilon > 0$ we have

$$N(\varepsilon,\mathscr{F},\mu) \geq N\left(\varepsilon, \left\{g\circ h : g\in\mathscr{G}\right\}, \mu\right). \tag{5.11}$$

The right hand side of inequality (5.11) may be lower bounded as follows. For all $g \in \mathscr{G}$, let g_h be the restriction of g to $S(h)\cap\mathbf{B}$. Then, for all $g, g' \in \mathscr{G}$, we have

$$\begin{aligned}
\|g\circ h - g'\circ h\|_\mu^2 &= \int (g\circ h(x) - g'\circ h(x))^2 \mu(\mathrm{d}x) \\
&= \int (g_h(u) - g'_h(u))^2 \mu\circ h^{-1}(\mathrm{d}u) \\
&\geq c\int (g_h(u) - g'_h(u))^2 \lambda_h(\mathrm{d}u),
\end{aligned} \tag{5.12}$$

where in (5.12) we have used (3.9). Hence, for all $\varepsilon > 0$ we have

$$N\left(\varepsilon, \left\{g\circ h : g\in\mathscr{G}\right\}, \mu\right) \geq N\left(\tfrac{\varepsilon}{c}, \mathscr{G}_h, \lambda_h\right), \tag{5.13}$$

where we have denoted $\mathscr{G}_h := \{g_h : g \in \mathscr{G}\}$. 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 the book by Lorentz et al (1996), there exists a constant $c' > 0$ such that for all $\varepsilon > 0$, we have

$$\ln N\left(\tfrac{\varepsilon}{c}, \mathscr{G}_h, \lambda_h\right) \geq c' \varepsilon^{-\ell/\beta}. \tag{5.14}$$

Combining (5.11), (5.13) and (5.14) we deduce that

$$\ln N(\varepsilon, \mathscr{F}, \mu) \geq c' \varepsilon^{-\ell/\beta} =: \ln N(\varepsilon).$$

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. □

5.3 Proof of Theorem 4.1

Lemma 5.3. *Suppose $\beta \geq 1$ and suppose there exists a constant $C > 0$ and $0 < s \leq p/\beta$ such that for all $\varepsilon > 0$ we have $H(\varepsilon, \mathscr{H}) \leq C\varepsilon^{-s}$. Suppose $a_n = n^u$ and $b_n = n^{-u'}$ with $u > 0$, $u' \geq 0$ and $0 < 2u + u' < 2\beta/(2\beta + p)$. Then for all $\ell \in \{1, \ldots, p\}$ and for all $\vartheta > 0$ we have*

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

Proof – Let

$$\tilde{R}_\ell := \inf_{f \in \mathscr{F}_\ell} \mathbb{E}\left(Y^{a_n} - f(X)\right)^2,$$

where $Y^{a_n} := Y\mathbf{1}\{|Y| \leq a_n\}$. We have

$$\begin{aligned} |\hat{R}_\ell - R_\ell| &\leq |\hat{R}_\ell - \tilde{R}_\ell| + |\tilde{R}_\ell - R_\ell| \\ &\leq \sup_{f \in \mathscr{F}} \left| \frac{1}{n} \sum_{i=1}^n \left(Y_i^{a_n} - f(X_i)\right)^2 - \mathbb{E}\left(Y^{a_n} - f(X)\right)^2 \right| \\ &\quad + \sup_{f \in \mathscr{F}} \left| \mathbb{E}\left(Y^{a_n} - f(X)\right)^2 - \mathbb{E}\left(Y - f(X)\right)^2 \right|. \end{aligned} \tag{5.15}$$

Using Cauchy-Schwarz's inequality we obtain that for all $f \in \mathscr{F}$, we have

$$\begin{aligned} &\left| \mathbb{E}\left(Y^{a_n} - f(X)\right)^2 - \mathbb{E}\left(Y - f(X)\right)^2 \right| \\ &= \left| \mathbb{E}\left(Y^{a_n} + Y - 2f(X)\right)\left(Y^{a_n} - Y\right) \right| \\ &\leq \mathbb{E}\left[\left|Y^{a_n} + Y - 2f(X)\right| \left|Y^{a_n} - Y\right| \right] \\ &\leq \sqrt{\mathbb{E}\left(Y^{a_n} + Y - 2f(X)\right)^2} \sqrt{\mathbb{E}\left(Y^{a_n} - Y\right)^2}. \end{aligned} \tag{5.16}$$

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

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

we deduce from the exponential moment condition (2.5) that for every positive integer k we have

$$\mathbb{E}Y^k \leq \frac{k!}{\tau^k}\mathbb{E}e^{\tau|Y|} \leq \frac{k!B}{\tau^k}. \tag{5.17}$$

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

$$\begin{aligned}
\sqrt{\mathbb{E}\left(Y^{a_n} + Y - 2f(X)\right)^2} &\leq 2\sqrt{\mathbb{E}Y^2} + 2\sqrt{\mathbb{E}f(X)^2} \\
&\leq 2\sqrt{\frac{2!}{\tau^2}\mathbb{E}e^{\tau|Y|}} + 2L \\
&\leq 2\sqrt{\frac{2B}{\tau^2}} + 2L. \qquad (5.18)
\end{aligned}$$

Using Markov's inequality and the exponential moment condition (2.5) we obtain that

$$\mathbb{P}\left(|Y| > a_n\right) = \mathbb{P}\Big(e^{\tau|Y|} > e^{\tau a_n}\Big) \leq Be^{-\tau a_n}. \tag{5.19}$$

Then, using Cauchy-Schwarz's inequality and equations (5.17) and (5.19) we deduce that

$$\begin{aligned}
\mathbb{E}\left(Y^{a_n} - Y\right)^2 &= \mathbb{E}\left(Y^2 \mathbf{1}\left\{|Y| > a_n\right\}\right) \\
&\leq \sqrt{\mathbb{E}Y^4}\sqrt{\mathbb{P}\left(|Y| > a_n\right)} \\
&\leq \sqrt{\frac{4!}{\tau^4}\mathbb{E}e^{\tau|Y|}}\sqrt{\mathbb{P}\Big(e^{\tau|Y|} > e^{\tau a_n}\Big)} \\
&\leq B\sqrt{\frac{24}{\tau^4}}e^{-\tau a_n/2}. \qquad (5.20)
\end{aligned}$$

Hence, combining (5.16), (5.18) and (5.20) we deduce that

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

Therefore, denoting by $\kappa_n := b_n - Ue^{-\tau a_n/4}$ and

$$\bar{\mathscr{F}}^{a_n} := \Big\{ (x,y) \mapsto (y\mathbf{1}\{|y| \le a_n\} - f(x))^2 : f \in \mathscr{F} \Big\},$$

we deduce from (5.15) and Theorem 9.1 in the book by Györfi *et al.* (2002) that there exists a universal constant $C > 0$ such that

$$\begin{aligned}
\mathbb{P}\big(|\hat{R}_\ell - R_\ell| \ge b_n\big) &\le \mathbb{P}\left(\sup_{f\in\mathscr{F}} \left| \frac{1}{n}\sum_{i=1}^n \big(Y_i^{a_n} - f(X_i)\big)^2 - \mathbb{E}\,(Y^{a_n} - f(X))^2 \right| \ge \kappa_n \right) \\
&\le C\mathbb{E}\left[N_1\left(\tfrac{\kappa_n}{C}, \bar{\mathscr{F}}^{a_n}, P_n\right)\right] e^{-\frac{n\kappa_n^2}{C(a_n+L)^4}}, \qquad (5.21)
\end{aligned}$$

where $P_n := n^{-1}\sum_{i=1}^n \delta_{(X_i,Y_i)}$ denotes the empirical distribution associated with the sample $(X_1,Y_1),\ldots,(X_n,Y_n)$ and where $N_1(\varepsilon,\mathscr{C},Q)$ denotes the minimal number of metric balls of radius ε in $\mathbb{L}^1(Q)$ that are needed to cover $\mathscr{C}$. For all $f, f' \in \mathscr{F}$ we have

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

Therefore, we obtain that

$$N_1\left(\tfrac{\kappa_n}{C}, \bar{\mathscr{F}}^{a_n}, P_n\right) \le N\left(\tfrac{\kappa_n}{2C(a_n+L)}, \mathscr{F}, \mu_n\right),$$

where $\mu_n := n^{-1}\sum_{i=1}^n \delta_{X_i}$. Hence, we deduce from (5.21) and Lemma 5.2 that there exists a universal constant $C > 0$ such that

$$\mathbb{P}\Big(|\hat{R}_\ell - R_\ell| \ge b_n\Big) \le C\exp\left(\left(\frac{C(a_n+L)}{\kappa_n}\right)^{p/\beta} - \frac{n\kappa_n^2}{C(a_n+L)^4} \right). \qquad (5.22)$$

Recalling that $a_n = n^u$, that $b_n = n^{-u'}$ and that $\kappa_n = b_n - Ue^{-\tau a_n/4}$, it follows that there exists a universal constant $C > 0$ such that

$$\sup_{P\in\mathscr{D}} \mathbb{P}\Big(|\hat{R}_\ell - R_\ell| \ge b_n\Big) \le C\exp\left(n^{(u+u')p/\beta} - n^{1-2u'-4u} \right). \qquad (5.23)$$

Consequently, the term on the left hand side of (5.23) converges exponentially fast to 0 as n goes to $+\infty$ provided

$$(1-2u'-4u) > (u+u')\frac{p}{\beta},$$

which is implied by

$$0 < 2u+u' < \frac{2\beta}{2\beta+p}.$$

The proof is complete. □

Proof of Theorem 4.1

Let $P \in \mathscr{D}$. Since the function $\ell \mapsto \hat{R}_\ell$ is non increasing, for all integer $q \in \{1,\dots,p\}$ and all $n \geq 1$ we have

$$\min\Big\{\ell = 1,\dots,p : \hat{R}_\ell - \hat{R}_p \leq b_n\Big\} \leq q \;\Leftrightarrow\; \hat{R}_q - \hat{R}_p \leq b_n. \tag{5.24}$$

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

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

Using Lemma 5.3, we deduce that for all $\vartheta > 0$ we have

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

and

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

which gives

$$\lim_{n\to+\infty} n^\vartheta \sup_{P\in\mathscr{D}} \mathbb{P}\Big(\hat{d} > d\Big) = 0. \tag{5.25}$$

Now let $P \in \mathscr{D}(\underline{\delta})$ 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 $\underline{\delta} - b_n \geq b_n$, we obtain that

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

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

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

which concludes the proof. □

5.4 Proof of Theorem 4.2

Since we have

$$\inf\big\{\Delta : P \in \mathscr{D}_0\big\} = 0,$$

for all $\varepsilon > 0$ there exists $P(\varepsilon) \in \mathscr{D}_0$ such that

$$\Delta_{P(\varepsilon)} \leq \varepsilon.$$

For all $n \geq 1$ let

$$Q_n := P(b_n/2).$$

Now assume that the sample $(X_1,Y_1),\ldots,(X_n,Y_n)$ is drawn from Q_n and let $\hat{d}$ be defined by

$$\hat{d} := \min\Big\{\ell : \hat{R}_\ell \leq \hat{R}_p + b_n\Big\}.$$

Then, we obtain as in the proof of Theorem 4.1 that

$$\begin{aligned}
\mathbb{P}\Big(\hat{d} < d\Big) &= \mathbb{P}\Big(\min\Big\{\ell = 1,\dots,p : \hat{R}_\ell - \hat{R}_p \le b_n\Big\} \le d-1\Big) \\
&= \mathbb{P}\Big(\hat{R}_{d-1} - \hat{R}_p \le b_n\Big) \\
&= \mathbb{P}\Big(\big(\hat{R}_{d-1} - R_{d-1}\big) + \Delta_{Q_n} + \big(R_p - \hat{R}_p\big) \le b_n\Big) \\
&\ge \mathbb{P}\Big(\big(\hat{R}_{d-1} - R_{d-1}\big) + \tfrac{b_n}{2} + \big(R_p - \hat{R}_p\big) \le b_n\Big) \\
&= \mathbb{P}\Big(\big(\hat{R}_{d-1} - R_{d-1}\big) + \big(R_p - \hat{R}_p\big) \le \tfrac{b_n}{2}\Big) \\
&\ge 1 - \mathbb{P}\Big(\big(\hat{R}_{d-1} - R_{d-1}\big) + \big(R_p - \hat{R}_p\big) \ge \tfrac{b_n}{2}\Big) \\
&\ge 1 - \mathbb{P}\Big(|\hat{R}_{d-1} - R_{d-1}| \ge \tfrac{b_n}{4}\Big) - \mathbb{P}\Big(|\hat{R}_p - R_p| \ge \tfrac{b_n}{4}\Big).
\end{aligned}$$

According to Lemma 5.3, we know that for all $\vartheta > 0$ we have

$$\lim_{n\to+\infty} n^\vartheta \mathbb{P}\Big(|\hat{R}_{d-1} - R_{d-1}| \ge \tfrac{b_n}{4}\Big) = 0 \quad \text{and} \quad \lim_{n\to+\infty} n^\vartheta \mathbb{P}\Big(|\hat{R}_p - R_p| \ge \tfrac{b_n}{4}\Big) = 0.$$

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

$$\mathbb{P}\Big(|\hat{R}_{d-1} - R_{d-1}| \ge \tfrac{b_n}{4}\Big) \le \frac{1}{2n^\vartheta} \quad \text{and} \quad \mathbb{P}\Big(|\hat{R}_p - R_p| \ge \tfrac{b_n}{4}\Big) \le \frac{1}{2n^\vartheta}.$$

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

$$\sup_{P\in\mathscr{D}_0} \mathbb{P}\Big(\hat{d} < d\Big) \ge \mathbb{P}_{(X,Y)\sim Q_n}\Big(\hat{d} < d\Big) \ge 1 - \frac{1}{n^\vartheta},$$

which concludes the proof. □

5.5 Proof of Theorem 4.3

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

$$\begin{aligned}
\mathbb{E}\left(\hat{f}(X) - \mathbb{E}(X|Y)\right)^2 &= \mathbb{E}\Big[\mathbf{1}\left\{\hat{d} \ne d\right\}\left(\hat{f}(X) - \mathbb{E}(X|Y)\right)^2\Big] \\
&\quad + \mathbb{E}\Big[\mathbf{1}\left\{\hat{d} = d\right\}\left(\hat{f}(X) - \mathbb{E}(X|Y)\right)^2\Big]. \quad (5.26)
\end{aligned}$$

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

$$\mathbb{E}\Big[\mathbf{1}\left\{\hat{d} \ne d\right\}\left(\hat{f}(X) - \mathbb{E}(X|Y)\right)^2\Big] \le 4L^2\mathbb{P}\Big(\hat{d} \ne d\Big). \quad (5.27)$$

Then, we observe that

$$\begin{aligned}\mathbb{E}\Big[\mathbf{1}\left\{\hat{d}=d\right\}\left(\hat{f}(X)-\mathbb{E}(X|Y)\right)^2\Big] &= \mathbb{E}\Big[\mathbf{1}\left\{\hat{d}=d\right\}\left(\hat{f}_d(X)-\mathbb{E}(Y|X)\right)^2\Big] \\ &\leq \mathbb{E}\left(\hat{f}_d(X)-\mathbb{E}(Y|X)\right)^2. \qquad (5.28)\end{aligned}$$

Hence, we deduce from (5.26), (5.27) and (5.28) that

$$\begin{aligned}\upsilon_{n,d}^{-2}\,\mathbb{E}\left(\hat{f}(X)-\mathbb{E}(X|Y)\right)^2 &\leq 4L^2\upsilon_{n,d}^{-2}\,\mathbb{P}\Big(\hat{d}\neq d\Big) \\ &\quad + \upsilon_{n,d}^{-2}\,\mathbb{E}\left(\hat{f}_d(X)-\mathbb{E}(Y|X)\right)^2. \qquad (5.29)\end{aligned}$$

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

$$\upsilon_{n,d}^{-2}\,\mathbb{E}\left(\hat{f}_d(X)-\mathbb{E}(Y|X)\right)^2 \leq C. \qquad (5.30)$$

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

$$\sup_{P\in\mathscr{D}(\underline{\delta})}\upsilon_{n,d}^{-2}\,\mathbb{E}\left(\hat{f}_d(X)-\mathbb{E}(Y|X)\right)^2 \leq C. \qquad (5.31)$$

Then, we deduce from Theorem 4.1 that

$$\sup_{P\in\mathscr{D}(\underline{\delta})}\upsilon_{n,d}^{-2}\,\mathbb{P}\Big(\hat{d}\neq d\Big) \leq \upsilon_{n,1}^{-2}\sup_{P\in\mathscr{D}(\underline{\delta})}\mathbb{P}\Big(\hat{d}\neq d\Big) \underset{n\to+\infty}{\to} 0. \qquad (5.32)$$

Finally, it follows from (5.29), (5.31) and (5.32) that there exists a constant depending only on τ, B, β, R and L such that

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

which concludes the proof. □

A Performance of least-squares estimates

Let (X,Z) be an $\mathscr{X} \times \mathbb{R}$-valued random variable of distribution P where $\mathscr{X}$ denotes a metric space. We fix $T > 0$ and assume that $|Z| \leq T$. Let μ be the distribution of X. For $x \in \mathscr{X}$ we set

$$f_*(x) := \mathbb{E}(Z|X = x). \tag{A.1}$$

Let $L > 0$ be fixed and $\mathscr{F}$ be a class of functions $f : \mathscr{X} \to \mathbb{R}$ such that

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

Let $(X_1,Z_1),\ldots,(X_n,Z_n)$ be a sample of n independent and identically distributed random variables with distribution P. The least-squares estimate f_n of f_* is defined as any element in $\mathscr{F}$ satisfying

$$f_n \in \operatorname*{arg\,min}_{f \in \mathscr{F}} \frac{1}{n} \sum_{i=1}^{n} (Z_i - f(X_i))^2 .$$

Implicitly, it is assumed that such an element exists. Following the lines devised in Koltchinskii (2006), we establish the following result. The proof may be found in the supplementary material Paris (2012).

Theorem A.1. *Let $0 < s < 2$ and assume there exists $A > 0$ such that for all $u > 0$ we have $H(u,\mathscr{F}) \leq Au^{-s}$. Then, there exists a constant C depending only on s, A and L such that for all $\varepsilon \in (0,1]$ we have*

$$\mathbb{E}\|f_n - f_*\|_\mu^2 \leq (1+\varepsilon) \inf_{f \in \mathscr{F}} \|f - f_*\|_\mu^2 + C\varepsilon^{-\frac{2-s}{2+s}} \left(\frac{b}{n}\right)^{\frac{2}{2+s}} + \frac{C}{\varepsilon}\left(\frac{b}{n}\right),$$

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

B Reduced dimension d and parameter Δ

In this appendix, we prove equations (4.5) and (4.8). First, observe that since $\mathscr{F}_\ell$ is compact in $\mathbb{L}^2(\mu)$ and since $r \in \mathscr{F}$, we have

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

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

$$\begin{aligned}
d &:= \min\Big\{\ell : r \in \mathscr{F}_\ell\Big\} \\
&= \min\Big\{\ell : R_\ell = R_p\Big\},
\end{aligned}$$

which proves equation (4.5). Using (4.5) and the fact that $r \in \mathscr{F}$ we obtain that

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

which proves (4.8).

Acknowledgments – The author is indebted to Benoit Cadre and Bruno Pelletier for their time and help. Nicolas Klutchnikoff is also to be thanked for several fruitful discussions and advices concerning the redaction of the manuscript.

References

Cadre, B. and Dong, Q. (2010). Dimension reduction in regression estimation with nearest neighbor. *Electronic Journal of Statistics*. Vol. 4, pp. 436-460.

Cook, R.D. and Li, B. (2002). Dimension reduction for conditional mean in regression. *The Annals of Statistics*. Vol. 30, pp. 455-474.

Cucker, F. and Smale, S. (2001). On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*. Vol. 39, pp. 1-49.

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

Ibragimov, I.A. and Khasminskii, R.Z. (1981). *Statistical Estimation: Asymptotic Theory*. Springer-Verlag, New-York NY.

Kohler, M., Krzyzak, A. and Walk, H. (2009). Optimal global rates of convergence in nonparametric regression with unbounded data. *Journal of Statistical Planning and Inference*. Vol. 123, pp. 1286-1296.

Koltchinskii, V. (2006). Local Rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*. Vol. 34, pp. 2593-2656.

Koltchinskii, V. (2011). *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems.* Ecole d'été de Probabilités de Saint-Flour 2008. Lectures Notes in Mathematics 2033.

Kosorok, M.R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. Springer-Verlag, New-York NY.

Li, K.C. (1991). Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association*. Vol. 86, pp. 316-342.

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

Paris, Q. (2012). Supplement to "Minimax adaptative dimension reduction for regression".

Yang, Y. and Barron, A. (1999). Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics*. Vol. 27, pp. 1564-1599.