Dimension reduction for regression over a general class of functions

Quentin Paris

Dimension reduction for regression over a general class of functions

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

Given an $\mathbb{R}^p \times \mathbb{R}$ valued random variable (X,Y), we investigate a new and nonparametric dimension reduction approach for estimating the regression function $r(x) = \mathbb{E}(Y|X = x)$ when p is large. We assume given a class $\mathscr{F}$ of functions $\varphi : \mathbb{R}^p \to \mathbb{R}^p$ such that there exists $\varphi \in \mathscr{F}$ with

$$\mathbb{E}(Y|\varphi(X)) = \mathbb{E}(Y|X).$$

In classical sufficient dimension reduction, one considers linear transformations of the predictor variable X that preserve the conditional expectation. Extending this approach, the class $\mathscr{F}$ is considered here to be a general and possibly nonparametric class of functions. In this context, we introduce the *reduced dimension* $d_{\mathscr{F}}$ associated with $\mathscr{F}$, defined as the dimension of the lowest dimensional subspace of $\mathbb{R}^p$ spanned by the range of a function $\varphi \in \mathscr{F}$ satisfying the former equality. Then, we define an estimate $\hat{r}$ of r and we prove that $\hat{r}$ achieves the optimal rate of convergence as if the predictor X where $d_{\mathscr{F}}$-dimensional.

Index Terms — Dimension reduction, regression, empirical risk minimization, nearest neighbor.

AMS 2000 Classification – 62H12, 62G08.

1 Introduction

In a general setting, regression analysis deals with the problem of retrieving information about the conditional distribution of a real-valued response variable Y

given an $\mathbb{R}^p$-valued predictor X and is often understood as a study of the regression function

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

Estimation of the regression function faces the *curse of dimensionality* which means that the expected rate of convergence of a given estimate slows down as the dimension of the predictor X increases. This statement is usually understood in terms of *optimal rates of convergence*. Given a class $\mathscr{M}$ of functions $\mathbb{R}^p \to \mathbb{R}$, the optimal rate of convergence associated to $\mathscr{M}$ is basically the best rate of convergence one can expect for an estimate $\hat{r}$ of r assuming only that r belongs to $\mathscr{M}$. For instance, if $L > 0$ is fixed, the optimal rate of convergence associated with the class of L-Lipschitz functions $\mathbb{R}^p \to \mathbb{R}$ is $n^{-2/(2+p)}$, under some technical assumptions (see e.g. Theorem 3.2 in the book by Györfi et al, 2002).

Therefore, the only way to circumvent the curse of dimensionality, in terms of optimal rates of convergence, is to consider a model $\mathscr{M}$ that encodes structural assumptions on the regression function in addition to the regularity assumptions. For more details on rates of convergence, we refer the reader to the book by Györfi et al (2002) or Ibragimov and Khasminskii (1981) and the references therein.

Methods to overcome the curse of dimensionality are usually referred to as *dimension reduction methods*. These methods usually consist in two fundamental steps. Based on structural assumptions on the regression function, the first step aims at finding an appropriate lower dimensional predictor variable and the second step focuses on using this predictor variable to build an estimate $\hat{r}$ of r. In this approach, it is naturally expected that the rate of convergence then only depends on the reduced dimension, namely the dimension of the new predictor variable.

An efficient approach to reach this first step is sufficient dimension reduction. Sufficient dimension reduction is a body of theory and methods that aim at replacing the predictor variable X by its projection $P_V X$ onto a lower dimensional subspace V without loss of information or in other words such that the conditional distribution of Y given $P_V X$ is equal to the conditional distribution of Y given X (see Li, 1991, 1992 and Cook and Weisberg, 1991). Under some mild conditions on the conditional distribution of Y given X, there exists a subspace $\mathscr{S}_{Y|X}$ of minimum dimension such that the former equality holds. When it exists, it is refered to as the *central subspace* and becomes in that context an important issue for dimension reduction (see Cook and Li, 2002). Many methods have been

introduced to estimate this subspace, among which we mention average derivative estimation (ADE; Härdle and Stoker, 1989), sliced inverse regression (SIR; Li, 1991), sliced average variance estimation (SAVE; Cook and Weisberg, 1991), kernel dimension reduction (Fukumizu et al, 2009) and more recently the optimal transformation procedure (Delyon and Portier, 2012). When focus is made on the regression function r instead of the conditional law of Y given X, it is sufficient to consider the *central mean subspace* $\mathscr{S}_{\mathbb{E}(Y|X)} \subset \mathscr{S}_{Y|X}$ that is, the subspace of minimum dimensionality among those V satisfying $\mathbb{E}(Y|P_V X) = \mathbb{E}(Y|X)$. Estimation of $\mathscr{S}_{\mathbb{E}(Y|X)}$ has been extensively studied; we mention for instance principal Hessian direction (pHd; Li, 1992) or minimum average variance estimation (MAVE; Xia et al, 2002). Discussions, improvements and other relevant papers can be found in Ye and Weiss (2003), Cook and Ni (2005), Zhu and Zeng (2006) or Delyon and Portier (2012).

Using sufficient dimension reduction methods, Cadre and Dong (2010) have constructed an estimate of the regression function with improved rate of convergence. They proved that given a matrix Λ that spans the central subspace $\mathscr{S}_{Y|X}$ and given a proper estimate $\hat{\Lambda}$ of Λ one may construct an estimate $\hat{r}$ of r such that

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

where d stands for the dimension of $\mathscr{S}_{Y|X}$, thus recovering the optimal rate of convergence should the predictor X be d-dimensional.

The motivation for our paper is that better performances may be expected when not only projections, or linear transformations, of the predictor X are considered but general transformations belonging to some prescribed and possibly nonparametric class $\mathscr{F}$. In the spirit of classical sufficient dimension reduction methods, we assume that the class $\mathscr{F}$ contains a function $\varphi : \mathbb{R}^p \to \mathbb{R}^p$ such that the following equality holds:

$$\mathbb{E}(Y|\varphi(X)) = \mathbb{E}(Y|X). \tag{1.1}$$

In this context, we introduce the *reduced dimension associated with* $\mathscr{F}$, denoted $d_{\mathscr{F}}$, defined as the dimension of lowest dimensional subspace of $\mathbb{R}^p$ spanned by the range of a function $\varphi \in \mathscr{F}$, satisfying equation (1.1). Provided such a function φ may be estimated, it is shown in this paper that an estimate $\hat{r}$ of r may be defined, and which converges at the optimal rate of convergence should the predictor be of

dimension $d_{\mathscr{F}}$, *i.e.* such that

$$\mathbb{E}\left(r(X)-\hat{r}(X)\right)^2 = O\left(n^{-2/(2+d_{\mathscr{F}})}\right).$$

Our main result, Theorem 4.1, holds for all $n \geq 1$ and involves explicit constants. Other important contributions in the field of nonlinear dimension reduction or nonlinear feature extraction can be found in Wu (2008), Yeh et al (2009) and Li et al (2011).

The paper is organized as follows. The model and tools for dimension reduction on a class $\mathscr{F}$ are defined in section 2. Section 3 is devoted to the study of an estimate of the reduced dimension $d_{\mathscr{F}}$ based on empirical risk minimization. In Section 4 we study an estimate of the regression function based on the Nearest Neighbors procedure. Section 5 is devoted to the study of some examples of classes of functions $\mathscr{F}$, including the linear setting, *i.e.* when only projections of the predictor are considered. In Section 6 we present a simulation study. Proofs of the main results are presented in Section 7 and technical results are collected in Section 8.

2 Model and tools for dimension reduction

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

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

Let $\mathscr{F}$ be a class of functions $\varphi : \mathscr{X} \to \mathbb{R}^p$ satisfying the following basic assumption.

Basic assumption – *There exists a function $\varphi\in\mathscr{F}$ such that*

$$\mathbb{E}(Y|\varphi(X)) = \mathbb{E}(Y|X). \tag{2.1}$$

This assumption generalizes the assumption made in sufficient dimension reduction where it is assumed that there exists a matrix $\Lambda\in M_p(\mathbb{R})$ of rank less than p such that $\mathbb{E}(Y|\Lambda X) = \mathbb{E}(Y|X)$. Here, we allow the class $\mathscr{F}$ to be a general, possibly nonparametric, class of functions, allowing therefore equation (2.1) to be satisfied for functions φ that may not be linear.

To assess the dimension reduction potential of $\mathscr{F}$, we introduce the *reduced dimension* associated with $\mathscr{F}$ given by

$$d_{\mathscr{F}} := \min\Big\{ \dim S(\varphi) : \varphi \in \mathscr{F},\ \mathbb{E}(Y|\varphi(X)) = \mathbb{E}(Y|X) \Big\},$$

where $S(\varphi)$ stands for the linear subspace of $\mathbb{R}^p$ spanned by the set $\varphi(\mathscr{X}) \subset \mathbb{R}^p$. A first part of our work is the estimation of $d_{\mathscr{F}}$ and Section 3 is devoted to this task. We will often write d instead of $d_{\mathscr{F}}$ keeping in mind the dependency on $\mathscr{F}$.

According to equation (2.1), there exists $\varphi \in \mathscr{F}$ and a measurable function $f : \mathbb{R}^p \to \mathbb{R}$ such that $r = f \circ \varphi$. At this stage, no specific information on function f is made available. To circumvent this drawback, we make a slightly stronger assumption in the spirit of our basic assumption. Let $\mathscr{L}$ be a given class of functions $\mathbb{R}^p \to \mathbb{R}$.

Assumption $(\mathbf{A1})$ – *The regression function r belongs to model $\mathscr{M}$ defined by*

$$\mathscr{M} := \Big\{ f \circ \varphi : f \in \mathscr{L},\ \varphi \in \mathscr{F} \Big\}. \tag{2.2}$$

Equation (2.2) allows the estimation of the regression function throughout the estimation of both a function f in $\mathscr{L}$ and a function φ in $\mathscr{F}$ in a way that will be described in Section 4. Now we proceed to giving a more tractable representation of the reduced dimension. For all $\ell \in \{1, \ldots, p\}$, let $\mathscr{F}_\ell$ be the class of all functions $\varphi \in \mathscr{F}$ such that $S(\varphi)$ is at most ℓ-dimensional, i.e.

$$\mathscr{F}_\ell := \Big\{ \varphi \in \mathscr{F} : \dim S(\varphi) \leq \ell \Big\},$$

and $\mathscr{M}_\ell \subset \mathscr{M}$ the submodel given by

$$\mathscr{M}_\ell := \Big\{ f \circ \varphi : f \in \mathscr{L},\ \varphi \in \mathscr{F}_\ell \Big\}.$$

The $\mathscr{M}_\ell$'s form a nested family of models, that is

$$\mathscr{M}_1 \subset \mathscr{M}_2 \subset \cdots \subset \mathscr{M}_p = \mathscr{M},$$

and under assumption $(\mathbf{A1})$, the reduced dimension may be written as

$$d = \min\Big\{ \ell = 1, \ldots, p : r \in \mathscr{M}_\ell \Big\}. \tag{2.3}$$

In other words, the reduced dimension d is the smallest ℓ for which there exists a function $\varphi \in \mathscr{F}_\ell$ and a function $f \in \mathscr{L}$ such that $r = f \circ \varphi$. Going one step further, we give a representation of d in terms of risk which will be considered in the sequel. If for all $\ell \in \{1, \ldots, p\}$, we denote by

$$R_\ell := \inf_{m \in \mathscr{M}_\ell} \mathbb{E}(Y - m(X))^2, \tag{2.4}$$

then (2.3) reveals that

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

We refer the reader to Figure 1 for an illustration of equation (2.5).

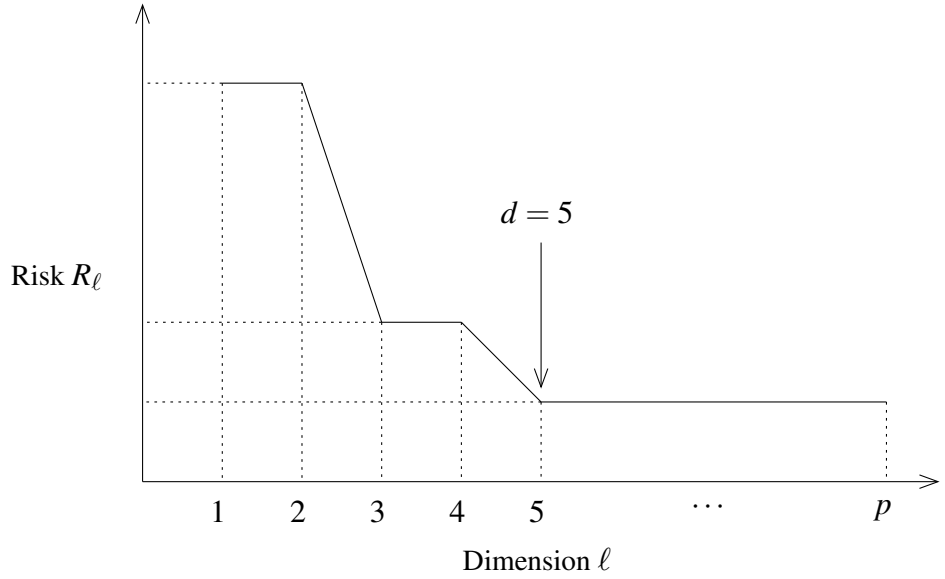

Figure 1: Illustration of the dependence of the risk R_ℓ on the dimension ℓ. In this example, the reduced dimension d is equal to 5.

3 Estimation of the reduced dimension

In this section, we focus on the estimation of the reduced dimension d. Consider a sample of n i.i.d. random variables $(X_1, Y_1), \ldots, (X_n, Y_n)$ with distribution P,

independent from (X,Y). For all $\ell \in \{1,\dots,p\}$, we denote by

$$\hat{R}_\ell := \inf_{m \in \mathscr{M}_\ell} \frac{1}{n} \sum_{i=1}^{n} (Y_i - m(X_i))^2,$$

the empirical version of the risk R_ℓ given in equation (2.4). Our estimation procedure is inspired by the representation given by equation (2.5). Define the *last drop of the risk* by

$$\gamma(\mathscr{M}) := \min \Big\{ R_\ell - R_p \ : \ R_\ell > R_p,\ \ell = 1,\dots,p \Big\},$$

with the convention $\min \emptyset = +\infty$. Note that the case where $\gamma(\mathscr{M}) = +\infty$ corresponds to $d = 1$. Then, choosing $0 \le \gamma < \gamma(\mathscr{M})$ leads to

$$d = \min \Big\{ \ell = 1,\dots,p : R_\ell \le R_p + \gamma \Big\}. \tag{3.1}$$

Accordingly, we define for all $\gamma \ge 0$ the estimate $\hat{d}(\gamma)$ of the reduced dimension d by

$$\hat{d}(\gamma) := \min \Big\{ \ell = 1,\dots,p : \hat{R}_\ell \le \hat{R}_p + \gamma \Big\}. \tag{3.2}$$

An illustration of equation (3.1) is provided in Figure 2.

Let us introduce some notations. For all $\ell \in \{1,\dots,p\}$, let $\|.\|_\ell$ be the Euclidean norm in $\mathbb{R}^\ell$. For $z \in \mathbb{R}^\ell$, we denote by $B_\ell(z,\varepsilon)$ the open Euclidean ball in $\mathbb{R}^\ell$ with center z and radius ε. When no confusion may arise, we denote $\|.\| = \|.\|_p$ and $B(u,\varepsilon) = B_p(u,\varepsilon)$ for $u \in \mathbb{R}^p$. For all $\ell \in \{1,\dots,p\}$ and any function $h : \mathscr{X} \to \mathbb{R}^\ell$ we denote by $\|h\|_\infty$ its supremum norm defined by

$$\|h\|_\infty = \sup_{x \in \mathscr{X}} \|h(x)\|_\ell.$$

Next, we describe the specific choice made in this paper for the class $\mathscr{L}$. First, we require $\mathscr{F}$ to be totally bounded with respect to the supremum norm $\|.\|_\infty$.

Assumption $(\mathbf{A2})$ – *$\mathscr{F}$ is totally bounded with respect to the supremum norm $\|.\|_\infty$ and we let $R > 0$ be such that $\|\varphi\|_\infty < R$ for all $\varphi \in \mathscr{F}$.*

This assumption on $\mathscr{F}$ implies that the elements of $\mathscr{F}$ can be approximated by a finite number of functions in a way that is described in Section 4.

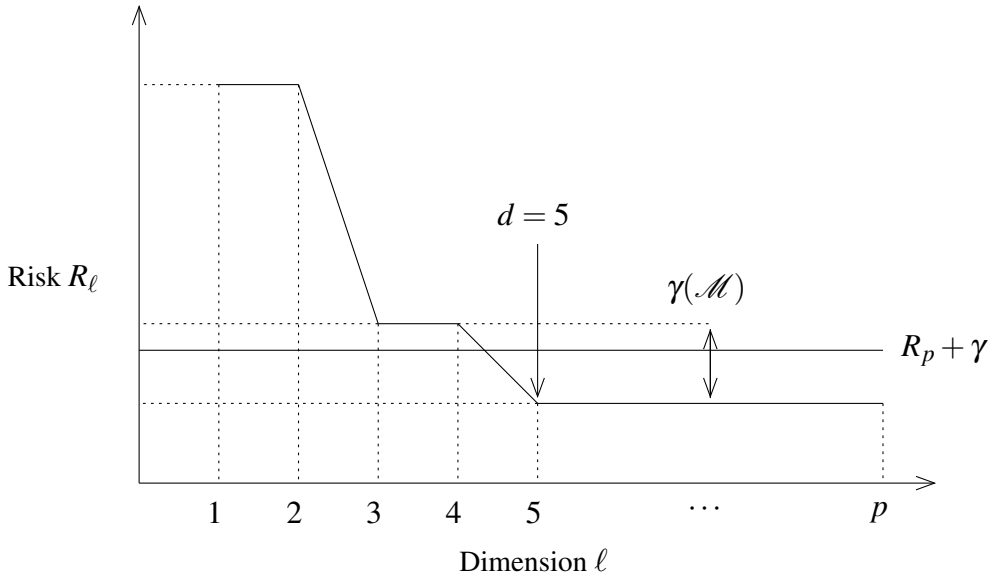

Figure 2: Illustration of the last drop of the risk $\gamma(\mathscr{M})$ and of the impact of the choice of the parameter γ. In this example, $0 < \gamma < \gamma(\mathscr{M})$ and the smallest ℓ for which $R_\ell \leq R_p + \gamma$ is equal to the reduced dimension d as in equation (3.1). Notice that if $\gamma > \gamma(\mathscr{M})$, then the smallest ℓ for which $R_\ell \leq R_p + \gamma$ is at most $d - 1$.

We denote $\mathbf{B} := B_p(0,R)$ in the sequel. We assume that $\mathscr{L}$ is composed of uniformly bounded Lipschitz functions. Precisely, for a fixed $L > 0$, we assume that

$$\mathscr{L} \subset \Big\{ f \in \mathscr{C}(\mathbf{B}, \mathbb{R}) : \|f\|_{Lip} \leq L \Big\}, \tag{3.3}$$

where $\mathscr{C}(\mathbf{B}, \mathbb{R})$ stands for the space of continuous functions $f : \mathbf{B} \to \mathbb{R}$ and where

$$\|f\|_{Lip} := \sup_u |f(u)| + \sup_{u \neq u'} \frac{|f(u) - f(u')|}{\|u - u'\|}. \tag{3.4}$$

From now on, it will be understood that $\mathscr{L}$ is chosen as in equation (3.3).

Remark 3.1. *For computational reasons one may have to choose $\mathscr{L}$ as a finite or countable set satisfying* (3.3)*. This choice has no significative impact in the following results and we decide to consider a general class $\mathscr{L}$ for reasons of ease.*

We shall need the following assumption in order to control the concentration of the empirical risk $\hat{R}_\ell$ around R_ℓ.

Assumption (**A3**) – *There exists $B > 0$ such that for all $m \in \mathscr{M}$: $|Y - m(X)| \leq B$ almost surely.*

Given $\ell \in \{1, \ldots, p\}$ and a set $\mathscr{G}$ of functions $h : \mathscr{X} \to \mathbb{R}^\ell$, recall that the ε-covering number of $\mathscr{G}$ with respect to $\|.\|_\infty$, denoted $N(\mathscr{G}, \|.\|_\infty, \varepsilon)$, is defined as the minimal number of $\|.\|_\infty$-balls of radius ε needed to cover $\mathscr{G}$.

Theorem 3.2. *Suppose assumptions* (**A1**)*,* (**A2**) *and* (**A3**) *are satisfied. Then, the following statements hold.*

(i) *If $0 < \gamma < \gamma(\mathscr{M})$, we have*

$$\mathbb{P}\Big(\hat{d}(\gamma) < d\Big) \leq 4\, N\left(\mathscr{M}, \|.\|_\infty, \tfrac{\gamma(\mathscr{M})-\gamma}{12(B+2L)}\right) \exp\left(-\tfrac{n(\gamma(\mathscr{M})-\gamma)^2}{18B^4}\right),$$

$$\mathbb{P}\Big(\hat{d}(\gamma) > d\Big) \leq 4\, N\left(\mathscr{M}, \|.\|_\infty, \tfrac{\gamma}{12(B+2L)}\right) \exp\left(-\tfrac{n\gamma^2}{18B^4}\right).$$

(ii) *If $\gamma > \gamma(\mathscr{M})$, we have*

$$\mathbb{P}\Big(\hat{d}(\gamma) < d\Big) \geq 1 - 4\, N\left(\mathscr{M}, \|.\|_\infty, \tfrac{\gamma-\gamma(\mathscr{M})}{12(B+2L)}\right) \exp\left(-\tfrac{n(\gamma-\gamma(\mathscr{M}))^2}{18B^4}\right).$$

We can easily deduce from Theorem 3.2 that provided $0 < \gamma < \gamma(\mathscr{M})$, we have

$$\hat{d}(\gamma) \underset{n\to+\infty}{\longrightarrow} d, \quad \text{a.s.}$$

In the next result, we prove that with the choice of class $\mathscr{L}$ given by (3.3), the covering number of model $\mathscr{M}$ is controlled in terms of the covering number of the class $\mathscr{F}$.

Proposition 3.3. *If class $\mathscr{F}$ satisfies assumption* (**A2**) *and if $\mathscr{L}$ is chosen as in* (3.3)*, there exists a constant C depending only on p, R and L such that for all $\varepsilon > 0$ we have*

$$N\Big(\mathscr{M}, \|.\|_\infty, \varepsilon\Big) \leq N\left(\mathscr{F}, \|.\|_\infty, \tfrac{\varepsilon}{2L}\right) \exp\left(\tfrac{C}{\varepsilon^p}\right).$$

4 Fast-rate estimation of the regression function

In this section we present an estimate of the regression function and give our main result. For each function $\varphi \in \mathscr{F}$, we denote r_φ the regression function defined for

all $u \in \mathbb{R}^p$ by

$$r_\varphi(u) := \mathbb{E}(Y|\varphi(X) = u). \tag{4.1}$$

Under assumption (**A1**) and by definition of the reduced dimension d, there exists $\varphi^* \in \mathscr{F}_d$ such that

$$r = r_{\varphi^*} \circ \varphi^* \quad \text{and} \quad r_{\varphi^*} \in \mathscr{L}. \tag{4.2}$$

The function φ^* is assumed fixed in the sequel. The estimation procedure presented here is inspired by this representation of r and consists in two steps. First, for all $\varphi \in \mathscr{F}$, we estimate r_φ using the k-nearest neighbors (k-NN) method. In a second step, we estimate φ^* through the minimization of an empirical criterion. Consider a second data set of i.i.d. copies $(X_{n+1}, Y_{n+1}), \ldots, (X_{2n}, Y_{2n})$ of (X, Y) independent from the first data set $(X_1, Y_1), \ldots, (X_n, Y_n)$ introduced in Section 3. Denote

$$\mathscr{D}_1 := \{(X_1, Y_1), \ldots, (X_n, Y_n)\} \quad \text{and} \quad \mathscr{D}_2 := \{(X_{n+1}, Y_{n+1}), \ldots, (X_{2n}, Y_{2n})\},$$

and fix a real number $\gamma > 0$.

We start with the estimation of r_φ and recall the k-NN procedure in this context. For all $\varphi \in \mathscr{F}$ and all $i \in \{n+1, \ldots, 2n\}$, we let

$$X_i^\varphi := \varphi(X_i).$$

If $u \in \mathbb{R}^p$, we reorder the transformed data $\left(X_{n+1}^\varphi, Y_{n+1}\right), \ldots, \left(X_{2n}^\varphi, Y_{2n}\right)$ according to increasing values of $\left\{\|X_i^\varphi - u\|, i = n+1, \ldots, 2n\right\}$. The reordered data sequence is denoted

$$\left(X_{(1)}^\varphi(u), Y_{(1)}^\varphi(u)\right), \left(X_{(2)}^\varphi(u), Y_{(2)}^\varphi(u)\right), \ldots, \left(X_{(n)}^\varphi(u), Y_{(n)}^\varphi(u)\right),$$

which means that

$$\|X_{(1)}^\varphi(u) - u\| \leq \|X_{(2)}^\varphi(u) - u\| \leq \cdots \leq \|X_{(n)}^\varphi(u) - u\|.$$

In this approach, $X_{(i)}^\varphi(u)$ is called the i-th NN of u. Note that if X_i^φ and X_j^φ are equidistant from u, i.e. $\|X_i^\varphi - u\| = \|X_j^\varphi - u\|$, then we have a tie. As usual, we then declare X_i^φ closer to u than X_j^φ if $i < j$. For any $i \in \{n+1, \ldots, 2n\}$ and $k \in \{1, \ldots, n\}$, we define

$$W_i[k](\varphi, u) = \begin{cases} 1/k & \text{if } X_i^\varphi \text{ is among the } k\text{-NN of } u \text{ in } \left\{X_{n+1}^\varphi, \ldots, X_{2n}^\varphi\right\}; \\ 0 & \text{elsewhere.} \end{cases}$$

Observe that we have $\sum_{i=n+1}^{2n} W_i[k](\varphi,u) = 1$. For all $k \in \{1,\dots,n\}$, we define the estimate $\hat{r}_\varphi[k]$ of r_φ for all $u \in \mathbb{R}^p$ by

$$\hat{r}_\varphi[k](u) = \sum_{i=n+1}^{2n} W_i[k](\varphi,u) Y_i = \frac{1}{k} \sum_{i=1}^{k} Y_{(i)}^{\varphi}(u).$$

For more information on the k-NN method, we refer the reader to Chapter 6 of the monography by Györfi et al (2002).

Now we focus on the estimation of φ^* through minimization of an empirical criterion over a finite covering of $\mathscr{F}$. To this aim, for all $\rho > 0$ and all $\ell \in \{1,\dots,p\}$, let $\mathbf{F}_\ell(\rho)$ be a ρ-covering of $\mathscr{F}_\ell$ of minimum cardinality. We set $\mathbf{F}(\rho) := \cup_\ell \mathbf{F}_\ell(\rho)$ and we denote

$$\mathbf{N}(\rho) := \text{Card } \mathbf{F}(\rho),$$

the cardinality of $\mathbf{F}(\rho)$. Now, for all $\ell \in \{1,\dots,p\}$, all $k \in \{1,\dots,n\}$ and all $\rho > 0$, we define

$$\hat{\varphi}_\ell[k,\rho] := \underset{\varphi \in \mathbf{F}_\ell(\rho)}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \hat{r}_\varphi[k](\varphi(X_i))\right)^2,$$

and we put

$$\hat{\varphi}[k,\rho] := \hat{\varphi}_{\hat{d}(\gamma)}[k,\rho],$$

where $\hat{d}(\gamma)$ is defined by equation (3.2).

For all $k \in \{1,\dots,n\}$ and all $\rho > 0$, we define the estimate $\hat{r}[k,\rho]$ of the regression function r by

$$\hat{r}[k,\rho] := \hat{r}_{\hat{\varphi}[k,\rho]}[k] \circ \hat{\varphi}[k,\rho].$$

Next, we describe a data driven choice of the number k of neighbors and of the radius ρ of the covering $\mathbf{F}(\rho)$. For all $\ell \in \{1,\dots,p\}$, denote

$$\upsilon_n(\ell) = n^{-2/(2+\ell)}. \tag{4.3}$$

Note that $\upsilon_n(\ell)$ is the optimal rate of convergence corresponding to the class of L-Lipschitz functions $\mathbb{R}^\ell \to \mathbb{R}$. Then, let

$$\hat{k} := \left\lfloor \upsilon_n^{-1}\left(\hat{d}(\gamma)\right) \right\rfloor \quad \text{and} \quad \hat{\rho} := \upsilon_n^{1/2}\left(\hat{d}(\gamma)\right), \tag{4.4}$$

where $\lfloor x \rfloor$ stands for the greatest integer smaller than x. Finally, for all $x \in \mathscr{X}$, let

$$\hat{r}(x) := \hat{r}\left[\hat{k}, \hat{\rho}\right](x). \tag{4.5}$$

We will assume that $\mathscr{F}$ satisfies the following property which roughly means that r_φ is close to r_{φ^*} provided φ is close to φ^*.

Assumption (**A4**) – *For all $\varphi \in \mathscr{F}$ and all $u \in \mathbb{R}^p$ we have*

$$|r_\varphi(u) - r_{\varphi^*}(u)| \leq L\|\varphi - \varphi^*\|_\infty,$$

where r_φ and φ^* have been defined in (4.1) and (4.2).

Theorem 4.1. *Suppose that assumptions* (**A1**) *to* (**A4**) *hold and that $d \geq 3$. For any $0 < \gamma < \gamma(\mathscr{M})$ and all $n \geq 1$ we have*

$$\mathbb{E}\Big[\left(r(X) - \hat{r}(X)\right)^2\Big] \leq 5C_1 n^{-2/(2+d)} + C_1\left\{2\mathbf{A}_n(\gamma) + \mathbf{B}_n + \frac{1}{\sqrt{n}}\right\},$$

where

$$\begin{aligned}
\mathbf{A}_n(\gamma) &:= N\left(\mathscr{M}, \|.\|_\infty, \tfrac{\gamma(\mathscr{M})-\gamma}{12(B+2L)}\right)\exp\left(-\tfrac{n(\gamma(\mathscr{M})-\gamma)^2}{18B^4}\right) \\
&\quad + \quad N\left(\mathscr{M}, \|.\|_\infty, \tfrac{\gamma}{12(B+2L)}\right)\exp\left(-\tfrac{n\gamma^2}{18B^4}\right), \\
\mathbf{B}_n &:= \frac{p}{n^{d/(2+d)}} N\left(\mathscr{F}, \|.\|_\infty, n^{-1/(2+d)}\right)\exp\left(-\tfrac{n^{(d-2)/(2+d)}}{C_1}\right),
\end{aligned}$$

and where

$$C_1 := \max\left\{1;\ 8(B+L)^2;\ 8(B+L)^4;\ 20L^2;\ 96L^2R^2 4^{1/d}\left(\tfrac{d-2}{2}\right)^{4/d}\right\}.$$

Remark 4.2. *When $d \leq 2$, under the additional conditions of Problem* 6.7 *in the book by Györfi et al (2002), a slight adaptation of the proof of Theorem 4.1 enables us to derive the same convergence rate.*

In Theorem 4.1, the quadratic risk of the estimator $\hat{r}$ is bounded by a sum of two terms. The first term

$$5C_1 n^{-2/(2+d)},$$

goes to 0 at the optimal rate of convergence associated with the class of L-Lipschitz functions $\mathbb{R}^d \to \mathbb{R}$. The second term

$$C_1\left\{2\mathbf{A}_n(\gamma) + \mathbf{B}_n + \frac{1}{\sqrt{n}}\right\},$$

is induced by our dimension reduction procedure. An important observation here is that since $0 < \gamma < \gamma(\mathscr{M})$ and $d \geq 3$, we have

$$\mathbf{A}_n(\gamma) = O\left(n^{-2/(2+d)}\right) \quad \text{and} \quad \frac{1}{\sqrt{n}} = O\left(n^{-2/(2+d)}\right).$$

Therefore, the rate of convergence of $\hat{r}$ depends on that of term $\mathbf{B}_n$. The next assumption consists in a restriction on the complexity of $\mathscr{F}$ and is known to be satisfied by many examples as described in Section 5.

Assumption $(\mathbf{A5})$ – *There exists* $A > 0$ *and* $0 < \beta < d-2$ *such that for all* $\varepsilon > 0$

$$\log N(\mathscr{F}, \|.\|_\infty, \varepsilon) \leq A\varepsilon^{-\beta}.$$

Under this assumption, we have

$$\mathbf{B}_n \leq \frac{p}{n^{d/(2+d)}} \exp\left(An^{\beta/(2+d)} - \frac{n^{(d-2)/(2+d)}}{C_1}\right) = O\left(n^{-2/(2+d)}\right),$$

which leads to the following result.

Corollary 4.3. *Under the assumptions of Theorem 4.1 and under assumption* $(\mathbf{A5})$*, we have*

$$\mathbb{E}\left[\left(r(X) - \hat{r}(X)\right)^2\right] = O\left(n^{-2/(2+d)}\right).$$

In other words, our estimator reaches the optimal rate of convergence, should X be taking its values in $\mathbb{R}^d$.

5 Examples

In this section, we study three examples and illustrate our main assumptions in different settings. The linear case is studied first. In a second example, we give a simple regression model for which linear dimension reduction is ineffective and for which our method applies. The third example provides an example where the class $\mathscr{F}$ is a non-parametric class composed with smooth functions.

5.1 Linear dimension reduction

Linear dimension reduction techniques (i.e. techniques that involve linear transformations of the predictor variable X) have proven effective in a large class of

examples and practical situations (see e.g. Cadre and Dong, 2010 or Chapter 22 of the book by Györfi et al, 2002). In classical sufficient dimension reduction for instance, one considers orthogonal projections of X.

Assume $\mathscr{X} \subset \mathbf{B}$ where $\mathbf{B}$ has been defined in Section 3. For any subspace V of $\mathbb{R}^p$ (abreviated $V \subset \mathbb{R}^p$), denote by $\pi_V : \mathbb{R}^p \to \mathbb{R}^p$ the orthogonal projector onto V and $P_V : \mathscr{X} \to \mathbb{R}^p$ the restriction of π_V to $\mathscr{X}$. Define

$$\mathscr{F} := \left\{P_V : V \subset \mathbb{R}^p\right\}.$$

The class $\mathscr{F}$ satisfies assumption $(\mathbf{A2})$ since for all $V \subset \mathbb{R}^p$

$$\|P_V\|_\infty = \sup_{x\in\mathscr{X}} \|P_V(x)\| = \sup_{x\in\mathscr{X}} \|\pi_V(x)\| \leq \sup_{x\in\mathscr{X}} \|\pi_V\|.\|x\| = \sup_{x\in\mathscr{X}} \|x\| < R.$$

Therefore, $\mathscr{F}$ may be seen as a subset of the Euclidean ball with center 0 and radius R in $\mathbb{R}^{2p}$ and it follows that

$$\log N(\mathscr{F}, \|.\|_\infty, \varepsilon) \leq C \log\left(\tfrac{1}{\varepsilon}\right),$$

for a constant C depending only on p and R (see e.g. Proposition 5 in Cucker and Smale, 2001). Hence, assumption $(\mathbf{A5})$ is also satisfied in this case. The fact that assumptions $(\mathbf{A1})$, $(\mathbf{A3})$ and $(\mathbf{A4})$ are satisfied depends on the joint distribution P of (X,Y).

5.2 Radial functions

Assume that X is uniformly distributed over $B_p(0,1)$. Let $K : \mathbb{R}_+ \to \mathbb{R}$ be non-constant, and assume the regression function r is given for all $x \in \mathscr{X}$ by

$$r(x) = K\left(\|x\|^2\right).$$

Then, it may be easily verified that every matrix $\Lambda \in M_p(\mathbb{R}) \setminus \{0\}$ satisfying $\mathbb{E}(Y|X) = \mathbb{E}(Y|\Lambda X)$ is of rank p. Therefore, in the case where the regression function is modeled by a radial function, dimension reduction using a class of linear transformations is ineffective. However, our general approach applies efficiently.

Let $\mathscr{F}$ be the class of functions with polynomial coordinate functions whose coefficients are bounded by 1 and degree at most 2. The function φ defined by

$\varphi(x) := \left(\sum_i x_i^2, 0, \ldots, 0\right)$ belongs to $\mathscr{F}$ and the regression function r satisfies $r = f \circ \varphi$ for $f(u) := K(u_1)$ which is Lipschitz if so is K. Hence assumption (**A1**) is satisfied in this context. Furthermore, one may verify easily that assumptions (**A2**) and (**A5**) are also verified. The general approach studied in this paper proves that one can find an estimate $\hat{r}$ of r such that

$$\mathbb{E}(\hat{r}(X) - r(X))^2 = O(n^{-2/3}),$$

since the reduced dimension is here $d = 1$, and under the additional assumptions (**A3**) and (**A4**).

5.3 Smooth functions

In this example we show that one may consider classes $\mathscr{F}$ much wider than the class of projectors or the parametric class introduced in the previous example. The class $\mathscr{F}$ introduced here consists in a nonparametric class of smooth functions. Fix two constants $R, \alpha > 0$. Denote by $\lfloor \alpha \rfloor$ the greatest integer strictly smaller than α and $\mathscr{C}^{\lfloor \alpha \rfloor}(\mathscr{X}, \mathbb{R})$ the space of $\lfloor \alpha \rfloor$-times continuously differentiable functions $h : \mathscr{X} \to \mathbb{R}$. For all $h \in \mathscr{C}^{\lfloor \alpha \rfloor}(\mathscr{X}, \mathbb{R})$ we define

$$\|h\|_\alpha := \max_{|s| \leq \lfloor \alpha \rfloor} \sup_x \|\partial^s h(x)\| + \max_{|s| = \lfloor \alpha \rfloor} \sup_{x \neq y} \frac{\|\partial^s h(x) - \partial^s h(y)\|}{\|x - y\|^{\alpha - \lfloor \alpha \rfloor}},$$

where, for all multi-index $s = (s_1, \ldots, s_p) \in \mathbb{N}^p$, we have denoted $|s| := \sum_i s_i$ and $\partial^s := \partial_1^{s_1} \cdots \partial_p^{s_p}$. Now we may define

$$\mathscr{F} := \left\{ h \in \mathscr{C}^{\lfloor \alpha \rfloor}(\mathscr{X}, \mathbb{R}) : \|h\|_\alpha \leq R \right\}^{\otimes p},$$

where for any set $\mathscr{G}$ of functions $\mathscr{X} \to \mathbb{R}$, $\mathscr{G}^{\otimes p}$ stands for the set of functions $\mathscr{X} \to \mathbb{R}^p$ such that each coordinate function belongs to $\mathscr{G}$. By definition of $\|.\|_\alpha$, assumption (**A2**) is satisfied. Now provided $\mathscr{X}$ is compact, there exists a constant C depending only on α, p and the diameter of $\mathscr{X}$ such that for all $\varepsilon > 0$

$$\log N(\mathscr{F}, \|.\|_\infty, \varepsilon) \leq C \varepsilon^{-p/\alpha}$$

(see e.g. Theorem 2.7.1 in van der Vaart and Wellner, 1996). Hence, provided $\alpha > p/(d-2)$, assumption (**A5**) is also satisfied.

6 A small simulation study

Here, we illustrate the improvement that our dimension reduction step induces compared to the classical NN approach, for a simple model. We let $p = 4$ and generate our data as follows. We let $X = (X^{(1)}, \dots, X^{(4)})$ be a 4-dimensional vector uniformly distributed over the unit open Euclidean ball B in $\mathbb{R}^4$ and let

$$Y = X^{(1)}X^{(4)} + X^{(2)}X^{(3)} + \sigma\varepsilon.$$

Here, $\sigma > 0$ and ε is a real random variable independent from X with standard normal distribution. We assume we have enough information about the phenomenon under study to know that the regression function is of the form $f \circ \varphi$ for $f \in \mathscr{L}$ and $\varphi \in \mathscr{F}$ where $\mathscr{L}$ and $\mathscr{F}$ are known and taken as follows. The class $\mathscr{L}$ is the class of functions f of the form $f(u_1, \dots, u_4) = a_1u_1 + \cdots + a_4u_4$ where $a_i \in \{-10, -9.9, \dots, 9.9, 10\}$. The class $\mathscr{F}$ is the class of functions $\varphi = (\varphi_1, \dots, \varphi_4)$ where $\varphi_i(x_1, \dots, x_4) = x_{i,a}x_{i,b}$ and where $x_{i,a}$, $x_{i,b} \in \{x_1, \dots, x_4\}$. In this context, the reduced dimension is $d = 2$.

First, we study the performance of the estimate $\hat{d}(\gamma)$ of d defined in (3.2) and describe an empirical procedure to select γ. We take a data set $\mathscr{D}_1 := \{(x_i, y_i) : i = 1, \dots, 1600\}$ generated by our model. We divide $\mathscr{D}_1$ into 40 sets of 40 data points and for $j = 1, \dots, 40$ we denote

$$\mathscr{S}_j := \{(x_i, y_i) : i = (j-1)n + 1, \dots, (j-1)n + 40\}.$$

For all $j = 1, \dots, 40$ and all $\gamma \in \{\kappa/2000 : \kappa = 0, 1, \dots, 1000\}$ we compute the estimate $\hat{d}_j(\gamma)$ of d based on the subsample $\mathscr{S}_j$ as in (3.2). Then, for $\ell = 1, 2, 3, 4$ we plot in Figure 3 the proportion of $\hat{d}_j(\gamma)$ that are equal to ℓ, *i.e.*

$$p_\ell(\gamma) := \frac{1}{40} \sum_{j=1}^{40} \mathbf{1}\left\{\hat{d}_j(\gamma) = \ell\right\},$$

when γ ranges over $\{\kappa/2000 : \kappa = 0, 1, \dots, 1000\}$. We select γ using the following heuristic approach. For small values of $\gamma > 0$ and when the number of data points is large enough, the probability that $\hat{d}(\gamma) = \ell$ should be close to 1 for $\ell = d$ and close to 0 for $\ell \neq d$ according to Theorem 3.2. Therefore, the value of ℓ for which $p_\ell(\gamma)$ is close to 1 for the small values of γ should correspond to the reduced dimension. Then, for this fixed value of ℓ, we select γ as the smallest maximizer of $p_\ell(\gamma)$. In our example, this heuristic applies successfully as we can

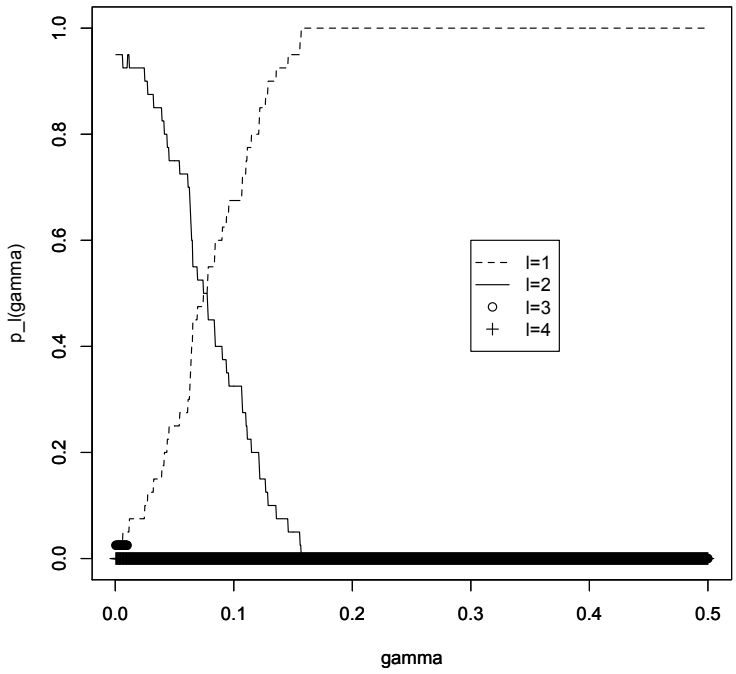

Figure 3: Plot of $p_\ell(\gamma)$ as a function of $\gamma \in (0, 0.5]$ for $\ell = 1, 2, 3, 4$.

see in Figure 3 that $p_2(\gamma) > 0.9$ for $0 < \gamma < 0.04$. Here we select $\gamma = 0.01$.

For the estimation of the regression function we consider two additional independent data sets $\mathscr{D}_2 := \{(x_i, y_i) : i = 1601, ..., 3200\}$ and $\mathscr{D}_{MC} := \{(x_i, y_i) : i = 3201, ..., 4800\}$ generated by our model for $\sigma = 0.1$, 0.5 and 0.9. For each $i = 3201, ..., 4800$ we compute our estimates $\hat{r}(x_i)$ based on the subsample $\mathscr{D} := \mathscr{D}_1 \cup \mathscr{D}_2$ with our methods as in (4.5) with $\gamma = 0.01$ and with the classical NN-method. Finally, we estimate $\mathbb{E}(\hat{r}(X) - \mathbb{E}(Y|X))^2$ from the Monte-Carlo approximation

$$\frac{1}{1600} \sum_{i=3201}^{4800} (\hat{r}(x_i) - \mathbb{E}(Y|X = x_i))^2 = \frac{1}{1600} \sum_{i=3201}^{4800} \left(\hat{r}(x_i) - x_i^{(1)} x_i^{(4)} - x_i^{(2)} x_i^{(3)} \right)^2,$$

where $x_i^{(j)}$ denotes the j-th coordinate of x_i. The obtained results are given in Table 1 and the variance of each experiment is given in parenthesis. As expected from our main theorem, our method performs better than the classical 3-dimensional NN method.

	Our meth.	3-dim NN-meth.
$\sigma = 0.1$	$0.027(8.4e-05)$	$0.039(1.2e-04)$
$\sigma = 0.5$	$0.216(7.3e-03)$	$0.336(8.4e-03)$
$\sigma = 0.9$	$0.720(4.7e-02)$	$0.898(4.9e-02)$

Table 1: Estimated mean squared error. The variance of the MC approximation is given in parenthesis.

7 Proofs

7.1 Proof of Theorem 3.2

Proof of (i) – Since the function $\ell \mapsto \hat{R}_\ell$ is non increasing, one has for all integer $q \in \{1, \dots, p\}$ and every $\gamma \geq 0$

$$\min \left\{ \ell = 1, \dots, p : \hat{R}_\ell - \hat{R}_p \leq \gamma \right\} \leq q \quad \Leftrightarrow \quad \hat{R}_q - \hat{R}_p \leq \gamma.$$

Now assume $0 < \gamma < \gamma(\mathscr{M})$. Using $R_{d-1} - R_p = \gamma(\mathscr{M})$ and the equivalence above, we have

$$\begin{aligned}
\mathbb{P}\Big(\hat{d}(\gamma) < d\Big) &= \mathbb{P}\Big(\min\Big\{\ell = 1,\ldots,p : \hat{R}_\ell - \hat{R}_p \le \gamma\Big\} \le d-1\Big) \\
&= \mathbb{P}\Big(\hat{R}_{d-1} - \hat{R}_p \le \gamma\Big) \\
&= \mathbb{P}\Big(\big(\hat{R}_{d-1} - R_{d-1}\big) + \gamma(\mathscr{M}) + \big(R_p - \hat{R}_p\big) \le \gamma\Big) \\
&\le \mathbb{P}\Big(|\hat{R}_{d-1} - R_{d-1}| \ge \tfrac{\gamma(\mathscr{M})-\gamma}{2}\Big) + \mathbb{P}\Big(|\hat{R}_p - R_p| \ge \tfrac{\gamma(\mathscr{M})-\gamma}{2}\Big) \\
&\le 4N\Big(\mathscr{M}, \|.\|_\infty, \tfrac{\gamma(\mathscr{M})-\gamma}{12(B+2L)}\Big) \exp\Big(-\tfrac{n(\gamma(\mathscr{M})-\gamma)^2}{18B^4}\Big),
\end{aligned}$$

where the last inequality follows from Lemma 8.1. Next, using $R_d = R_p$, observe that

$$\begin{aligned}
\mathbb{P}\Big(\hat{d}(\gamma) > d\Big) &= \mathbb{P}\Big(\min\Big\{\ell = 1,\ldots,p : \hat{R}_\ell - \hat{R}_p \le \gamma\Big\} > d\Big) \\
&= \mathbb{P}\Big(\hat{R}_d - \hat{R}_p > \gamma\Big) \\
&= \mathbb{P}\Big(\big(\hat{R}_d - R_d\big) + \big(R_p - \hat{R}_p\big) > \gamma\Big) \\
&\le \mathbb{P}\Big(|\hat{R}_p - R_p| \ge \tfrac{\gamma}{2}\Big) + \mathbb{P}\Big(|\hat{R}_d - R_d| \ge \tfrac{\gamma}{2}\Big) \\
&\le 4N\Big(\mathscr{M}, \|.\|_\infty, \tfrac{\gamma}{12(B+2L)}\Big) \exp\Big(-\tfrac{n\gamma^2}{18B^4}\Big),
\end{aligned}$$

where again the last inequality follows from Lemma 8.1.

Proof of (ii) – Now assume $\gamma > \gamma(\mathscr{M})$. We have

$$\begin{aligned}
\mathbb{P}\Big(\hat{d}(\gamma) \ne d\Big) &\ge \mathbb{P}\Big(\hat{d}(\gamma) < d\Big) \\
&= \mathbb{P}\Big(\min\Big\{\ell = 1,\ldots,p : \hat{R}_\ell - \hat{R}_p \le \gamma\Big\} \le d-1\Big) \\
&= \mathbb{P}\Big(\hat{R}_{d-1} - \hat{R}_p \le \gamma\Big) \\
&= \mathbb{P}\Big(\big(\hat{R}_{d-1} - R_{d-1}\big) + \gamma(\mathscr{M}) + \big(R_p - \hat{R}_p\big) \le \gamma\Big) \\
&= 1 - \mathbb{P}\Big(\big(\hat{R}_{d-1} - R_{d-1}\big) + \big(R_p - \hat{R}_p\big) > \gamma - \gamma(\mathscr{M})\Big) \\
&\ge 1 - \mathbb{P}\Big(|\hat{R}_{d-1} - R_{d-1}| \ge \tfrac{\gamma-\gamma(\mathscr{M})}{2}\Big) - \mathbb{P}\Big(|\hat{R}_p - R_p| \ge \tfrac{\gamma-\gamma(\mathscr{M})}{2}\Big) \\
&\ge 1 - 4N\Big(\mathscr{M}, \|.\|_\infty, \tfrac{\gamma-\gamma(\mathscr{M})}{12(B+2L)}\Big) \exp\Big(-\tfrac{n(\gamma-\gamma(\mathscr{M}))^2}{18B^4}\Big),
\end{aligned}$$

by Lemma 8.1 once again.□

7.2 Proof of Proposition 3.3

Fix $\varepsilon > 0$ and denote $\mathrm{N} := N\left(\mathscr{F}, \|.\|_\infty, \frac{\varepsilon}{2L}\right)$. By definition, there exists functions $\varphi_1, \ldots, \varphi_\mathrm{N} \in \mathscr{F}$ such that the $\|.\|_\infty$-balls with radius $\frac{\varepsilon}{2L}$ and centers the φ_i's cover $\mathscr{F}$. For all $i \in \{1, \ldots, \mathrm{N}\}$, define

$$N_i := N\Big(\mathscr{L} \circ \varphi_i, \|.\|_\infty, \tfrac{\varepsilon}{2}\Big).$$

Then, for all $i \in \{1, \ldots, \mathrm{N}\}$, there exists functions $f_{i,1}, \ldots, f_{i,N_i}$ such that for every $f \in \mathscr{L}$ there is at least one $j \in \{1, \ldots, N_i\}$ with

$$\|f \circ \varphi_i - f_{i,j} \circ \varphi_i\|_\infty \leq \frac{\varepsilon}{2}.$$

Now, for any two functions $f \in \mathscr{L}$ and $\varphi \in \mathscr{F}$, there exists two integers $k \in \{1, \ldots, \mathrm{N}\}$ and $j \in \{1, \ldots, N_k\}$ such that

$$\|\varphi - \varphi_k\|_\infty \leq \frac{\varepsilon}{2L} \quad \text{and} \quad \|f \circ \varphi_k - f_{k,j} \circ \varphi_k\|_\infty \leq \frac{\varepsilon}{2}.$$

Therefore, we have

$$\begin{aligned} \|f \circ \varphi - f_{k,j} \circ \varphi_k\|_\infty &\leq \|f \circ \varphi - f \circ \varphi_k\|_\infty + \|f \circ \varphi_k - f_{k,j} \circ \varphi_k\|_\infty \\ &\leq L\|\varphi - \varphi_k\|_\infty + \frac{\varepsilon}{2} \\ &\leq L\frac{\varepsilon}{2L} + \frac{\varepsilon}{2} = \varepsilon, \end{aligned}$$

since, according to (3.4), every function in $\mathscr{L}$ is L-Lipschitz. We have proved that

$$N\Big(\mathscr{M}, \|.\|_\infty, \varepsilon\Big) \leq \sum_{i=1}^{\mathrm{N}} N_i.$$

To complete the proof, we need only to show that there exists a constant C depending only on p, R and L such that

$$\log N_i \leq \frac{C}{\varepsilon^p},$$

for all $i \in \{1,\ldots,\mathrm{N}\}$. For that purpose, observe that for any function $\varphi \in \mathscr{F}$ and any $\varepsilon > 0$ we have

$$N\Big(\mathscr{L} \circ \varphi, \|.\|_\infty, \varepsilon\Big) \leq N\Big(\mathscr{L}, \|.\|_\infty, \varepsilon\Big),$$

since for any two functions $f, f' \in \mathscr{L}$

$$\|f \circ \varphi - f' \circ \varphi\|_\infty \leq \|f - f'\|_\infty.$$

Now according to Theorem 2.7.1 in van der Vaart and Wellner (1996), there exists a constant C depending only on p, R and L such that

$$\log N\Big(\mathscr{L}, \|.\|_\infty, \varepsilon\Big) \leq \frac{C}{\varepsilon^p}.$$

Hence, for all $i \in \{1,\ldots,N\}$ and all $\varepsilon > 0$

$$\log N_i = \log N\Big(\mathscr{L} \circ \varphi_i, \|.\|_\infty, \varepsilon\Big) \leq \log N\Big(\mathscr{L}, \|.\|_\infty, \varepsilon\Big) \leq \frac{C}{\varepsilon^p},$$

which yields the desired result.□

7.3 Proof of Theorem 4.1

First, we give an intermediate result.

Proposition 7.1. *Suppose assumptions* (**A1**) *to* (**A4**) *hold. Suppose* $d \geq 3$ *and* $0 < \gamma < \gamma(\mathscr{M})$. *Then for all* $k \in \{1,\ldots,n\}$ *and all* $\rho > 0$ *we have*

$$\mathbb{E}\Big[\,(r(X) - \hat{r}[k,\rho](X))^2\,\Big] \leq C_1 \left\{\frac{1}{k} + \left(\frac{k}{n}\right)^{2/d}\right\} + C_1 \left\{\mathbf{D}_n(\gamma) + \mathbf{T}_n(\rho) + \frac{1}{\sqrt{n}}\right\},$$

where

$$\begin{aligned}
\mathbf{D}_n(\gamma) &:= N\left(\mathscr{M}, \|.\|_\infty, \tfrac{\gamma(\mathscr{M})-\gamma}{12(B+2L)}\right) \exp\left(-\tfrac{n(\gamma(\mathscr{M})-\gamma)^2}{18B^4}\right), \\
\mathbf{T}_n(\rho) &:= \rho^2 + \upsilon_n(d) + \tfrac{p}{n\upsilon_n(d)} N\left(\mathscr{F}, \|.\|_\infty, \rho\right) \exp\left(-\tfrac{n\upsilon_n^2(d)}{C_1}\right),
\end{aligned}$$

and where

$$C_1 := \max\left\{1;\ 8(B+L)^2;\ 8(B+L)^4;\ 20L^2;\ 96L^2R^2 4^{1/d}\left(\tfrac{d-2}{2}\right)^{4/d}\right\}.$$

Proof of Proposition 7.1 – Fix $k \in \{1,\dots,n\}$, $\rho > 0$ and a function $\varphi \in \mathbf{F}_d(\rho)$ such that $\|\varphi - \varphi^*\|_\infty \leq \rho$. We have

$$\begin{aligned}\mathbb{E}\Big[\,(r(X)-\hat{r}[k,\rho](X))^2\,\Big] &= \mathbb{E}\Big[\,(r(X)-\hat{r}[k,\rho](X))^2\,\mathbf{1}\left\{\hat{d}(\gamma)<d\right\}\Big] \\ &\quad + \mathbb{E}\Big[\,(r(X)-\hat{r}[k,\rho](X))^2\,\mathbf{1}\left\{\hat{d}(\gamma)\geq d\right\}\Big] \\ &=: T_1+T_2.\end{aligned}$$

Under assumptions (**A1**) and (**A3**), we have $|Y-r(X)| \leq B$. Since the functions in $\mathscr{L}$ are uniformly bounded by L by definition, r is uniformly bounded by L and $|Y| \leq |Y-r(X)|+|r(X)| \leq B+L$. Hence, from the construction of $\hat{r}[k,\rho]$, we have $|\hat{r}[k,\rho](X)| \leq B+L$, and so

$$(r(X)-\hat{r}[k,\rho](X))^2 \leq (B+2L)^2.$$

Therefore

$$T_1 \leq (B+2L)^2\mathbb{P}\Big(\hat{d}(\gamma)<d\Big).$$

Using Theorem 3.2, we deduce that

$$T_1 \;\leq\; 4(B+2L)^2 N\left(\mathscr{M},\|.\|_\infty,\tfrac{\gamma(\mathscr{M})-\gamma}{12(B+2L)}\right)\exp\left(-\tfrac{n(\gamma(\mathscr{M})-\gamma)^2}{18B^4}\right).$$

Next, we have

$$\begin{aligned}T_2 &= \mathbb{E}\Big[\,(r(X)-\hat{r}[k,\rho](X))^2\,\mathbf{1}\left\{\hat{d}(\gamma)\geq d\right\}\Big] \\ &= \mathbb{E}\Big[\mathbb{E}\Big[\,(r(X)-\hat{r}[k,\rho](X))^2\,\big|\mathscr{D}\Big]\mathbf{1}\left\{\hat{d}(\gamma)\geq d\right\}\Big] \\ &= \mathbb{E}\Big[\Big\{\mathbb{E}\Big[\,(Y-\hat{r}[k,\rho](X))^2\,\big|\mathscr{D}\Big]-\mathbb{E}\Big[\,(Y-r(X))^2\,\Big]\Big\}\mathbf{1}\left\{\hat{d}(\gamma)\geq d\right\}\Big] \\ &=: \mathbb{E}\Big[\,(I_1+I_2+I_3)\,\mathbf{1}\left\{\hat{d}(\gamma)\geq d\right\}\Big],\end{aligned}$$

where I_1, I_2 and I_3 are defined by

$$\begin{aligned}I_1 &= \left\{\mathbb{E}\left[(Y-\hat{r}[k,\rho](X))^2\,|\mathscr{D}\right]-\frac{1}{n}\sum_{i=1}^n(Y_i-\hat{r}[k,\rho](X_i))^2\right\}, \\ I_2 &= \left\{\frac{1}{n}\sum_{i=1}^n(Y_i-\hat{r}[k,\rho](X_i))^2-\mathbb{E}\left[\left(Y-\hat{r}_\varphi[k](\varphi(X))\right)^2|\mathscr{D}\right]\right\}, \\ I_3 &= \left\{\mathbb{E}\left[\left(Y-\hat{r}_\varphi[k](\varphi(X))\right)^2|\mathscr{D}\right]-\mathbb{E}\,(Y-r(X))^2\right\}.\end{aligned}$$

By taking $a = \upsilon_n(d)$ in Lemma 8.2, we obtain

$$\mathbb{E}\Big[I_1 \mathbf{1}\left\{\hat{d}(\gamma) \geq d\right\}\Big] \;\leq\; \mathbb{E}\Big[\big|I_1\big|\Big] \leq\; C'\left\{\upsilon_n(d) + \frac{\mathbf{N}(\rho)}{n\upsilon_n(d)} \exp\left(-\tfrac{n\upsilon_n^2(d)}{C'}\right)\right\},$$

where we recall that $\upsilon_n(d)$ has been defined in equation (4.3) and where $C' = \max\{1;\ 8(B+L)^2;\ 8(B+L)^4\}$. From the construction of $\mathbf{F}(\rho)$, we deduce that

$$\mathbb{E}\Big[I_1 \mathbf{1}\left\{\hat{d}(\gamma) \geq d\right\}\Big] \leq C'\left\{\upsilon_n(d) + \frac{pN(\mathscr{F}, \|.\|_\infty, \rho)}{n\upsilon_n(d)} \exp\left(-\tfrac{n\upsilon_n^2(d)}{C'}\right)\right\}. \quad (7.1)$$

Now on the event $\left\{\hat{d}(\gamma) \geq d\right\}$, we have the inclusion $\mathbf{F}_d(\rho) \subset \mathbf{F}_{\hat{d}(\gamma)}(\rho)$, and therefore

$$\frac{1}{n}\sum_{i=1}^{n} (Y_i - \hat{r}[k,\rho](X_i))^2 \leq \frac{1}{n}\sum_{i=1}^{n} \big(Y_i - \hat{r}_\varphi[k](\varphi(X_i))\big)^2.$$

We conclude that on the event $\left\{\hat{d}(\gamma) \geq d\right\}$ we have

$$\begin{aligned} I_2 \;\leq\; J_2 \;\;&:=\;\; \left\{\frac{1}{n}\sum_{i=1}^{n} \big(Y_i - \hat{r}_\varphi[k](\varphi(X_i))\big)^2 - \mathbb{E}\left[\big(Y - \hat{r}_\varphi[k](\varphi(X))\big)^2 | \mathscr{D}\right]\right\} \\ &=\;\; \left\{\frac{1}{n}\sum_{i=1}^{n} \big(Y_i - \hat{r}_\varphi[k](\varphi(X_i))\big)^2 - \mathbb{E}\left[\big(Y - \hat{r}_\varphi[k](\varphi(X))\big)^2 | \mathscr{D}_2\right]\right\}. \end{aligned}$$

Recall that, for all $\varphi \in \mathscr{F}$, $\hat{r}_\varphi[k]$ is constructed on $\mathscr{D}_2 = \{(X_i, Y_i); i = n+1, \ldots, 2n\}$. Therefore, conditionally to $\mathscr{D}_2$, the variables

$$\big(Y - \hat{r}_\varphi[k](\varphi(X))\big)^2 \quad \text{and} \quad \big(Y_i - \hat{r}_\varphi[k](\varphi(X_i))\big)^2,\ i = 1, \ldots, n,$$

are i.i.d. and bounded by $4(B+L)^2$. Hence, for all $\varepsilon > 0$, Hoeffding's inequality yields

$$\mathbb{P}\Big(|J_2| > \varepsilon \big| \mathscr{D}_2\Big) \leq 2\exp\left(-\tfrac{n\varepsilon^2}{8(B+L)^4}\right).$$

Therefore

$$\begin{aligned}
\mathbb{E}\Big[I_2\mathbf{1}\left\{\hat{d}(\gamma)\geq d\right\}\Big] &\leq \mathbb{E}\Big[|J_2|\mathbf{1}\left\{\hat{d}(\gamma)\geq d\right\}\Big] \\
&\leq \mathbb{E}\Big[\big|J_2\big|\Big] \\
&= \int_0^{+\infty}\mathbb{P}\Big(|J_2|>\varepsilon\Big)\mathrm{d}\varepsilon \\
&= \int_0^{+\infty}\mathbb{E}\left[\mathbb{P}\Big(|J_2|>\varepsilon|\mathscr{D}_2\Big)\right]\mathrm{d}\varepsilon \\
&\leq 2\int_0^{+\infty}\exp\left(-\tfrac{n\varepsilon^2}{8(B+L)^4}\right)\mathrm{d}\varepsilon \\
&= \frac{4(B+L)^2}{\sqrt{n}}\int_0^{+\infty}\exp\left(-\tfrac{\tau^2}{2}\right)\mathrm{d}\tau \\
&= \frac{2\sqrt{2\pi}(B+L)^2}{\sqrt{n}} \\
&\leq \frac{C''}{\sqrt{n}}, \qquad (7.2)
\end{aligned}$$

where $C'' := 8(B+L)^2$. Finally, we have

$$\begin{aligned}
\mathbb{E}\Big[I_3\mathbf{1}\left\{\hat{d}(\gamma)\geq d\right\}\Big] &= \mathbb{E}\Big[\mathbb{E}\left[\big(r(X)-\hat{r}_\varphi[k](\varphi(X))\big)^2|\mathscr{D}\right]\mathbf{1}\left\{\hat{d}(\gamma)\geq d\right\}\Big] \\
&\leq \mathbb{E}\Big[\big(r(X)-\hat{r}_\varphi[k](\varphi(X))\big)^2\Big] \\
&\leq C'''\left\{\frac{1}{k}+\left(\frac{k}{n}\right)^{2/d}+\rho^2\right\}, \qquad (7.3)
\end{aligned}$$

where the last inequality follows from Lemma 8.3 an where constant C''' can be taken equal to $\max\Big\{2B^2;\ 20L^2\ ;\ 96L^2R^2 4^{1/d}\left(\frac{d-2}{2}\right)^{4/d}\Big\}$. Combining (7.1), (7.2) and (7.3) and denoting

$$C_1 := \max\Big\{1;\ 8(B+L)^2;\ 8(B+L)^4;\ 20L^2;\ 96L^2R^2 4^{1/d}\left(\tfrac{d-2}{2}\right)^{4/d}\Big\},$$

we conclude that

$$T_2 \leq C_1\left\{\frac{1}{k}+\left(\frac{k}{n}\right)^{2/d}\right\}+C_1\left\{\mathbf{T}_n(\rho)+\frac{1}{\sqrt{n}}\right\},$$

where

$$\mathbf{T}_n(\rho) := \rho^2 + \upsilon_n(d) + \frac{pN(\mathscr{F}, \|.\|_\infty, \rho)}{n\upsilon_n(d)} \exp\left(-\frac{n\upsilon_n^2(d)}{C_1}\right).$$

As a result, we obtain

$$\begin{aligned}
\mathbb{E}\Big[(r(X) - \hat{r}[k,\rho](X))^2 \Big] &\leq T_1 + T_2 \\
&\leq C_1 \left\{ \frac{1}{k} + \left(\frac{k}{n}\right)^{2/d} \right\} + C_1 \left\{ \mathbf{D}_n(\gamma) + \mathbf{T}_n(\rho) + \frac{1}{\sqrt{n}} \right\},
\end{aligned}$$

where

$$\mathbf{D}_n(\gamma) := N\left(\mathscr{M}, \|.\|_\infty, \tfrac{\gamma(\mathscr{M})-\gamma}{12(B+2L)}\right) \exp\left(-\tfrac{n(\gamma(\mathscr{M})-\gamma)^2}{18B^4}\right),$$

as desired. □

Proof of Theorem 4.1 – Under assumptions (**A1**) and (**A3**), we have $|Y - r(X)| \leq B$. Since the functions in $\mathscr{L}$ are uniformly bounded by L by definition, r is uniformly bounded by L and so $|Y| \leq |Y - r(X)| + |r(X)| \leq B + L$. Hence, from the construction of $\hat{r}$, we have $|\hat{r}(X)| \leq B + L$, and so

$$(r(X) - \hat{r}(X))^2 \leq (B + 2L)^2.$$

Therefore

$$\begin{aligned}
\mathbb{E}\Big[(r(X) - \hat{r}(X))^2 \Big] &= \mathbb{E}\Big[(r(X) - \hat{r}(X))^2 \mathbf{1}\left\{\hat{d}(\gamma) \neq d\right\} \Big] \\
&\quad + \mathbb{E}\Big[(r(X) - \hat{r}(X))^2 \mathbf{1}\left\{\hat{d}(\gamma) = d\right\} \Big] \\
&= \mathbb{E}\Big[(r(X) - \hat{r}(X))^2 \mathbf{1}\left\{\hat{d}(\gamma) \neq d\right\} \Big] \\
&\quad + \mathbb{E}\Big[\Big(r(X) - \hat{r}\left[\lfloor \upsilon_n^{-1}(d) \rfloor, \upsilon_n^{1/2}(d)\right](X) \Big)^2 \mathbf{1}\left\{\hat{d}(\gamma) = d\right\} \Big] \\
&\leq (B+2L)^2 \mathbb{P}\Big(\hat{d}(\gamma) \neq d\Big) \\
&\quad + \mathbb{E}\Big[\Big(r(X) - \hat{r}\left[\lfloor \upsilon_n^{-1}(d) \rfloor, \upsilon_n^{1/2}(d)\right](X) \Big)^2 \Big], \\
&=: U_1 + U_2. \qquad (7.4)
\end{aligned}$$

According to Theorem 3.2 we have

$$\begin{aligned}
\mathbb{P}\Big(\hat{d}(\gamma) \neq d\Big) &\leq N\left(\mathscr{M}, \|.\|_\infty, \tfrac{\gamma(\mathscr{M})-\gamma}{12(B+2L)}\right) \exp\left(-\tfrac{n(\gamma(\mathscr{M})-\gamma)^2}{18B^4}\right) \\
&\quad + N\left(\mathscr{M}, \|.\|_\infty, \tfrac{\gamma}{12(B+2L)}\right) \exp\left(-\tfrac{n\gamma^2}{18B^4}\right) \\
&=: \mathbf{A}_n(\gamma).
\end{aligned}$$

Hence

$$U_1 \leq (B+2L)^2 \mathbf{A}_n(\gamma). \tag{7.5}$$

Now, by Proposition 7.1 applied with $k = \lfloor \upsilon_n^{-1}(d) \rfloor$ and $\rho = \upsilon_n^{1/2}(d)$, we have

$$\begin{aligned} U_2 &\leq C_1 \left\{ \frac{1}{\lfloor \upsilon_n^{-1}(d) \rfloor} + \left(\frac{\lfloor \upsilon_n^{-1}(d) \rfloor}{n} \right)^{2/d} \right\} \\ &+ C_1 \left\{ \mathbf{D}_n(\gamma) + \mathbf{T}_n(\upsilon_n^{1/2}(d)) + \frac{1}{\sqrt{n}} \right\}, \end{aligned}$$

where C_1, $\mathbf{D}_n(\gamma)$ and $\mathbf{T}_n(\rho)$ have been defined in Proposition 7.1. Since $x-1 < \lfloor x \rfloor \leq x$, for all positive number x, we deduce that

$$\begin{aligned} &\frac{1}{\lfloor \upsilon_n^{-1}(d) \rfloor} + \left(\frac{\lfloor \upsilon_n^{-1}(d) \rfloor}{n} \right)^{2/d} \\ &\leq \frac{1}{\upsilon_n^{-1}(d)-1} + \left(\frac{\upsilon_n^{-1}(d)}{n} \right)^{2/d} \\ &= \left(\frac{1}{\upsilon_n^{-1}(d)-1} - \frac{1}{\upsilon_n^{-1}(d)} \right) + \frac{1}{\upsilon_n^{-1}(d)} + \left(\frac{\upsilon_n^{-1}(d)}{n} \right)^{2/d} \\ &= \frac{1}{(\upsilon_n^{-1}(d)-1)\upsilon_n^{-1}(d)} + 2\upsilon_n(d) \\ &\leq 3\upsilon_n(d). \end{aligned} \tag{7.6}$$

From the definitions, it is clear that

$$\mathbf{D}_n(\gamma) \leq \mathbf{A}_n(\gamma). \tag{7.7}$$

Next, we have

$$\begin{aligned} \mathbf{T}_n(\upsilon_n^{1/2}(d)) &= 2\upsilon_n(d) + \frac{pN\left(\mathscr{F}, \|.\|_\infty, \upsilon_n^{1/2}(d)\right)}{n\upsilon_n(d)} \exp\left(-\tfrac{n\upsilon_n^2(d)}{C_1}\right) \\ &=: 2\upsilon_n(d) + \mathbf{B}_n. \end{aligned} \tag{7.8}$$

From inequalities (7.6), (7.7) and (7.8), we deduce that

$$U_2 \leq 5C_1\upsilon_n(d) + C_1 \left\{ \mathbf{A}_n(\gamma) + \mathbf{B}_n + \frac{1}{\sqrt{n}} \right\}. \tag{7.9}$$

As a consequence, we conclude from (7.4), (7.5) and (7.9) that

$$\begin{aligned}\mathbb{E}\Big[\,(r(X)-\hat{r}(X))^2\,\Big] &\leq U_1+U_2\\ &\leq (B+2L)^2\mathbf{A}_n(\gamma)+5C_1\upsilon_n(d)+C_1\left\{\mathbf{A}_n(\gamma)+\mathbf{B}_n+\frac{1}{\sqrt{n}}\right\}\\ &\leq 5C_1\upsilon_n(d)+C_1\left\{2\mathbf{A}_n(\gamma)+\mathbf{B}_n+\frac{1}{\sqrt{n}}\right\},\end{aligned}$$

since $(B+2L)^2 \leq C_1$. □

8 Technical results

Lemma 8.1. *Suppose assumptions* (**A1**) *to* (**A3**) *hold. Then, for all* $\ell \in \{1,\ldots,p\}$ *and all* $\varepsilon > 0$*, we have:*

$$\mathbb{P}\Big(|\hat{R}_\ell - R_\ell| \geq \varepsilon\Big) \leq 2N\left(\mathscr{M}_\ell, \|.\|_\infty, \tfrac{\varepsilon}{6(B+2L)}\right)\exp\left(-\tfrac{2n\varepsilon^2}{9B^4}\right).$$

Proof – First, we have

$$\begin{aligned}\mathbb{P}\Big(|\hat{R}_\ell - R_\ell| \geq \varepsilon\Big) &= \mathbb{P}\left(\Big|\inf_{m\in\mathscr{M}_\ell}\frac{1}{n}\sum_{i=1}^n (Y_i-m(X_i))^2 - \inf_{m\in\mathscr{M}_\ell}\mathbb{E}(Y-m(X))^2\Big| \geq \varepsilon\right)\\ &\leq \mathbb{P}\left(\sup_{m\in\mathscr{M}_\ell}\Big|\frac{1}{n}\sum_{i=1}^n (Y_i-m(X_i))^2 - \mathbb{E}(Y-m(X))^2\Big| \geq \varepsilon\right).\end{aligned}$$

Then, notice that $|Y| \leq |Y-r(X)|+|r(X)| \leq B+L$. For all $m \in \mathscr{M}_\ell$, denote $g_m : (x,y) \in \mathscr{X}\times[-B-L,B+L] \mapsto (y-m(x))^2$ and define

$$\widetilde{\mathscr{M}_\ell} := \Big\{g_m : m \in \mathscr{M}_\ell\Big\}.$$

For all $m \in \mathscr{M}_\ell$, we have $|g_m(X,Y)| \leq B^2$. Therefore, according to Lemma 9.1 in Györfi et al (2002) we have

$$\mathbb{P}\Big(|\hat{R}_\ell - R_\ell| \geq \varepsilon\Big) \leq 2N\left(\widetilde{\mathscr{M}_\ell}, \|.\|_\infty, \tfrac{\varepsilon}{3}\right)\exp\left(-\tfrac{2n\varepsilon^2}{9B^4}\right). \tag{8.1}$$

Then, for any two functions $m, m' \in \mathscr{M}_\ell$, we have

$$\begin{aligned}|g_m(x,y)-g_{m'}(x,y)| &= |(y-m(x))^2-(y-m'(x))^2|\\ &= |(2y-m(x)-m'(x))(m(x)-m'(x))|\\ &\leq 2(B+2L)|m(x)-m'(x)|,\end{aligned}$$

for all $(x,y) \in \mathscr{X} \times [-B-L, B+L]$. Therefore, for all $\varepsilon > 0$, we have

$$N\Big(\widetilde{\mathscr{M}_\ell}, \|.\|_\infty, \varepsilon\Big) \leq N\Big(\mathscr{M}_\ell, \|.\|_\infty, \tfrac{\varepsilon}{2(B+2L)}\Big). \tag{8.2}$$

Combining (8.1) and (8.2) yields the expected result. □

Lemma 8.2. *Suppose that assumptions* (**A1**) *to* (**A3**) *hold. Then, for all* $k \in \{1,\dots,n\}$*, all* $\rho > 0$ *and all* $a > 0$*, we have*

$$\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{r}[k,\rho](X_i))^2 - \mathbb{E}\Big[(Y - \hat{r}[k,\rho](X))^2 \,\Big|\, \mathscr{D}\Big]\right|\right] \leq C\left\{a + \frac{\mathbf{N}(\rho)}{na}\exp\left(-\tfrac{na^2}{C}\right)\right\},$$

where we can take $C := \max\{1;\ 8(B+L)^2;\ 8(B+L)^4\}$.

Proof – Fix $k \in \{1,\dots,n\}$, $\rho > 0$ and $a > 0$. Denote $\mathbf{F}(\rho) = \big\{\psi_j : j = 1,\dots,\mathbf{N}(\rho)\big\}$. Since (X,Y) is independent from $\mathscr{D}$, since for all $\varphi \in \mathscr{F}$, $\hat{r}_\varphi[k] \circ \varphi$ depends only on $\mathscr{D}_2 = \{(X_i,Y_i); i = n+1,\dots,2n\}$, and since $\hat{\varphi}[k,\rho]$ takes its values in $\mathbf{F}(\rho)$, we have

$$\begin{aligned}
\mathbb{E}\Big[(Y - \hat{r}[k,\rho](X))^2 \,\Big|\, \mathscr{D}\Big] &= \mathbb{E}\left[\left(Y - \sum_{j=1}^{\mathbf{N}(\rho)} \hat{r}_{\psi_j}[k](\psi_j(X))\mathbf{1}\big\{\hat{\varphi}[k,\rho] = \psi_j\big\}\right)^2 \,\middle|\, \mathscr{D}\right] \\
&= \mathbb{E}\left[\sum_{j=1}^{\mathbf{N}(\rho)} \big(Y - \hat{r}_{\psi_j}[k](\psi_j(X))\big)^2 \mathbf{1}\big\{\hat{\varphi}[k,\rho] = \psi_j\big\} \,\middle|\, \mathscr{D}\right] \\
&= \sum_{j=1}^{\mathbf{N}(\rho)} \mathbb{E}\Big[\big(Y - \hat{r}_{\psi_j}[k](\psi_j(X))\big)^2 \,|\, \mathscr{D}\Big] \mathbf{1}\big\{\hat{\varphi}[k,\rho] = \psi_j\big\} \\
&= \sum_{j=1}^{\mathbf{N}(\rho)} \mathbb{E}\Big[\big(Y - \hat{r}_{\psi_j}[k](\psi_j(X))\big)^2 \,|\, \mathscr{D}_2\Big] \mathbf{1}\big\{\hat{\varphi}[k,\rho] = \psi_j\big\}.
\end{aligned}$$

Therefore, denoting $E_j := \big\{\hat{\varphi}[k,\rho] = \psi_j\big\}$, we obtain

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{r}[k,\rho](X_i))^2 - \mathbb{E}\Big[(Y - \hat{r}[k,\rho](X))^2 \,|\, \mathscr{D}\Big] \\
&\quad = \sum_{j=1}^{\mathbf{N}(\rho)} \left\{\frac{1}{n}\sum_{i=1}^{n}\big(Y_i - \hat{r}_{\psi_j}[k](\psi_j(X_i))\big)^2 - \mathbb{E}\Big[\big(Y - \hat{r}_{\psi_j}[k](\psi_j(X))\big)^2 \,|\, \mathscr{D}_2\Big]\right\} \mathbf{1}\big\{E_j\big\}.
\end{aligned}$$

Now for all $i \in \{1,\dots,n\}$ and all $j \in \{1,\dots,\mathbf{N}(\rho)\}$, let

$$Z_{i,j} := \big(Y_i - \hat{r}_{\psi_j}[k](\psi_j(X_i))\big)^2 - \mathbb{E}\Big[\big(Y - \hat{r}_{\psi_j}[k](\psi_j(X))\big)^2 \,|\mathscr{D}_2\Big].$$

Using the fact that the events E_j are pairwise disjoint, we deduce that for all $\varepsilon > 0$

$$\begin{aligned}
&\mathbb{P}\left(\Big|\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{r}[k,\rho](X_i))^2 - \mathbb{E}\Big[(Y - \hat{r}[k,\rho](X))^2\,|\mathscr{D}\Big]\Big| \geq \varepsilon\right)\\
&= \quad \mathbb{P}\left(\Big|\sum_{j=1}^{\mathbf{N}(\rho)}\left\{\frac{1}{n}\sum_{i=1}^{n} Z_{i,j}\right\}\mathbf{1}\{E_j\}\Big| \geq \varepsilon\right)\\
&= \quad \mathbb{P}\left(\max_{j=1,\dots,\mathbf{N}(\rho)}\Big|\frac{1}{n}\sum_{i=1}^{n} Z_{i,j}\Big| \geq \varepsilon\right)\\
&\leq \quad \mathbf{N}(\rho)\max_{j=1,\dots,\mathbf{N}(\rho)}\mathbb{P}\left(\Big|\frac{1}{n}\sum_{i=1}^{n} Z_{i,j}\Big| \geq \varepsilon\right),
\end{aligned}$$

where the last inequality follows from the union bound. Now, conditionally to $\mathscr{D}_2$, and for all $j \in \{1,\dots,\mathbf{N}(\rho)\}$, the variables

$$\big(Y - \hat{r}_{\psi_j}[k](\psi_j(X))\big)^2 \quad \text{and} \quad \big(Y_i - \hat{r}_{\psi_j}[k](\psi_j(X_i))\big)^2,\ i \in \{1,\dots,n\}$$

are i.i.d. an bounded by $4(B+L)^2$. By Hoeffding's inequality, it follows that for all $j \in \{1,\dots,\mathbf{N}(\rho)\}$

$$\mathbb{P}\left(\Big|\frac{1}{n}\sum_{i=1}^{n} Z_{i,j}\Big| \geq \varepsilon \Big| \mathscr{D}_2\right) \leq 2\exp\left(-\tfrac{n\varepsilon^2}{8(B+L)^4}\right).$$

Therefore

$$\begin{aligned}
&\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n}(Y_i-\hat{r}[k,\rho](X_i))^2-\mathbb{E}\Big[(Y-\hat{r}[k,\rho](X))^2\,\Big|\,\mathscr{D}\Big]\right|\right]\\
&= \int_0^{+\infty}\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}(Y_i-\hat{r}[k,\rho](X_i))^2-\mathbb{E}\Big[(Y-\hat{r}[k,\rho](X))^2\,\Big|\,\mathscr{D}\Big]\right|\geq\varepsilon\right)\mathrm{d}\varepsilon\\
&\leq a+\int_a^{+\infty}\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}(Y_i-\hat{r}[k,\rho](X_i))^2-\mathbb{E}\Big[(Y-\hat{r}[k,\rho](X))^2\,\Big|\,\mathscr{D}\Big]\right|\geq\varepsilon\right)\mathrm{d}\varepsilon\\
&\leq a+\mathbf{N}(\rho)\int_a^{+\infty}\max_{j=1,\ldots,\mathbf{N}(\rho)}\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}Z_{i,j}\right|\geq\varepsilon\right)\mathrm{d}\varepsilon\\
&= a+\mathbf{N}(\rho)\int_a^{+\infty}\max_{j=1,\ldots,\mathbf{N}(\rho)}\mathbb{E}\left[\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}Z_{i,j}\right|\geq\varepsilon\Big|\mathscr{D}_2\right)\right]\mathrm{d}\varepsilon\\
&\leq a+2\mathbf{N}(\rho)\int_a^{+\infty}\exp\left(-\tfrac{n\varepsilon^2}{8(B+L)^4}\right)\mathrm{d}\varepsilon.
\end{aligned}$$

Now, using the fact that for all $x>0$

$$\int_x^{+\infty}\exp\left(-\tfrac{\tau^2}{2}\right)\mathrm{d}\tau\leq\frac{1}{x}\exp\left(-\tfrac{x^2}{2}\right),$$

we obtain

$$\begin{aligned}
\int_a^{+\infty}\exp\left(-\tfrac{n\varepsilon^2}{8(B+L)^4}\right)\mathrm{d}\varepsilon &= \frac{2(B+L)^2}{\sqrt{n}}\int_{\frac{a\sqrt{n}}{2(B+L)^2}}^{+\infty}\exp\left(-\tfrac{\tau^2}{2}\right)\mathrm{d}\tau\\
&\leq \frac{4(B+L)^4}{an}\exp\left(-\tfrac{na^2}{8(B+L)^2}\right).
\end{aligned}$$

This leads to

$$\begin{aligned}
&\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n}(Y_i-\hat{r}[k,\rho](X_i))^2-\mathbb{E}\Big[(Y-\hat{r}[k,\rho](X))^2\,\Big|\,\mathscr{D}\Big]\right|\right]\\
&\leq a+\frac{8(B+L)^4\mathbf{N}(\rho)}{an}\exp\left(-\tfrac{na^2}{8(B+L)^2}\right)\\
&\leq C\left\{a+\frac{\mathbf{N}(\rho)}{an}\exp\left(-\tfrac{na^2}{C}\right)\right\},
\end{aligned}$$

with $C:=\max\{1;\,8(B+L)^2;\,8(B+L)^4\}$, as desired. □

Lemma 8.3. *Suppose assumptions* (**A1**) *to* (**A4**) *hold. Then, for all* $k \in \{1,\dots,n\}$, *all* $\rho > 0$ *and all* $\varphi \in \mathbf{F}_d(\rho)$ *satisfying* $\|\varphi - \varphi^*\|_\infty \leq \rho$, *we have*

$$\mathbb{E}\Big[\big(r(X) - \hat{r}_\varphi[k](\varphi(X))\big)^2\Big] \leq C\left\{\frac{1}{k} + \left(\frac{k}{n}\right)^{2/d}\right\} + C\rho^2,$$

where $C := \max\left\{2B^2;\ 20L^2\ ;\ 96L^2R^2 4^{1/d}\left(\frac{d-2}{2}\right)^{4/d}\right\}$.

The proof of Lemma 8.3 is very similar to that of Theorem 2.2 in Cadre and Dong (2010) and has therefore been reported in the supplementary material Paris (2012).

Acknowledgments – The author is indebted to Benoit Cadre and Bruno Pelletier for their time and help.

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.

Cook, R.D. and Ni, L. (2005). Sufficient dimension reduction via inverse regression: A minimum discrepancy approach. *Journal of the American Statistical Association*. Vol. 100, pp. 410-428.

Cook, R.D. and Weisberg, S. (1991). Discussion of Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*. Vol. 86, pp. 316-342.

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

Delyon, B. and Portier. F. (2012). Optimal transformation: A new approach for covering the central subspace. (to appear in *Journal of Multivariate Analysis*)

Fukumizu, K., Bach, F. and Jordan, M. (2009). Kernel dimension reduction in regression. *The Annals of Statistics*. Vol. 37, pp. 1871-1905.

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

Härdle, W. and Stoker, T.M. (1989). Investigating smooth multiple regression by the method of average derivative. *Journal of the American Statistical Association*. Vol. 84, pp. 986-995.

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

Li, B., Artemiou, A. and Li, L. (2011). Principal support vector machine for linear and nonlinear dimension reduction. *The Annals of Statistics*. Vol. 39, pp. 3182-3210.

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

Li, K.C. (1992). On principal Hessian directions for data visualization and dimension reduction: Another application of Steins lemma. *Journal of the American Statistical Association*. Vol. 87, pp. 1025-1039.

Paris, Q. (2012). Supplement to "Dimension reduction for regression over a general class of function".

van der Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes.* Springer-Verlag, New-York NY.

Wu, H.M. (2008). Kernel sliced inverse regression with applications on classification. *Journal of Computational and Graphical Statistics*. Vol. 17, pp. 590-610.

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

Ye, Z. and Weiss, R.E. (2003). Using the bootstrap to select one of a new class of dimension reduction methods. *Journal of the American Statistical Association*. Vol. 98, pp. 968979.

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

Zhu, Y. and Zeng P. (2006). Fourier methods for estimating the central subspace and the central mean subspace in regression. *Journal of the American Statistical Association*. Vol. 101, pp. 1638-1651.