Iterative Isotonic Regression

Arnaud GUYADER[1]
Université Rennes 2, INRIA and IRMAR
Campus de Villejean, Rennes, France
arnaud.guyader@uhb.fr

Nick HENGARTNER
Los Alamos National Laboratory
NM 87545, USA
nickh@lanl.gov

Nicolas JÉGOU
Université Rennes 2
Campus de Villejean, Rennes, France
nicolas.jegou@uhb.fr

Eric MATZNER-LØBER
Université Rennes 2
Campus de Villejean, Rennes, France
eml@uhb.fr

Abstract

This article explores some theoretical aspects of a recent nonparametric method for estimating a univariate regression function of bounded variation. The method exploits the Jordan decomposition which states that a function of bounded variation can be decomposed as the sum of a non-decreasing function and a non-increasing function. This suggests combining the backfitting algorithm for estimating additive functions with isotonic regression for estimating monotone functions. The resulting iterative algorithm is called Iterative Isotonic Regression (I.I.R.). The main result in this paper states that the estimator is consistent if the number of iterations k_n grows appropriately with the sample size n. The proof requires two auxiliary results that are of interest in and by themselves: firstly, we generalize the well-known consistency property of isotonic regression to the framework of a non-monotone regression function, and secondly, we relate the backfitting algorithm to von Neumann's algorithm in convex analysis. We also analyse how the algorithm can be stopped in practice using a data-splitting procedure.

Index Terms — Nonparametric statistics, isotonic regression, additive models, metric projection onto convex cones.

2010 Mathematics Subject Classification: 52A05, 62G08, 62G20.

[1]Corresponding author.

1 Introduction

Consider the regression model

$$Y = r(X) + \varepsilon, \tag{1}$$

where X and ε are independent real-valued random variables, with X distributed according to a non-atomic law μ on $[0,1]$, $\mathbb{E}[\varepsilon] = 0$ and $\mathbb{E}[\varepsilon^2] = \sigma^2$. We want to estimate the regression function r, assuming it is of bounded variation. Since μ is non-atomic, we will further assume, without loss of generality, that r is right-continuous. The Jordan decomposition states that r can be written as the sum of a non-decreasing function u and a non-increasing function b

$$r(x) = u(x) + b(x). \tag{2}$$

This decomposition is not unique in general. However, if one requires that both terms on the right-hand side have singular associated Stieltjes measures and that

$$\int_{[0,1]} r(x)\mu(dx) = \int_{[0,1]} u(x)\mu(dx), \tag{3}$$

then the decomposition is unique and the model is identifiable. Let us emphasize that, from a statistical point of view, this assumption on r is mild. The classical counterexample of a function that is not of bounded variation is $r(x) = \sin(1/x)$ for $x \in (0,1]$, with $r(0) = 0$.

Our idea for estimating a regression function of bounded variation consists in viewing the Jordan decomposition (2) as an additive model involving the increasing and the decreasing parts of r. It leads to an "Iterative Isotonic Regression" estimator (abbreviated to I.I.R.) that combines the isotonic regression and backfitting algorithms, two well-established algorithms for estimating monotone functions and additive models, respectively.

Estimating a monotone regression function is the archetypical shape restriction estimation problem. Specifically, assume that the regression function r in (1) is non-decreasing, and suppose we are given a sample $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ of i.i.d. $\mathbb{R}\times\mathbb{R}$ valued random variables distributed as a generic pair (X, Y). Then denote $x_1 = X_{(1)} < \ldots < x_n = X_{(n)}$, the ordered sample and $y_1, \ldots, y_n$ the corresponding observations. In this framework, the Pool-Adjacent-Violators Algorithm (PAVA) determines a collection of non-decreasing level sets solution to the least square minimization problem

$$\min_{u_1 \leq \ldots \leq u_n} \frac{1}{n} \sum_{i=1}^{n} (y_i - u_i)^2 . \tag{4}$$

Early works on the maximum likelihood estimators of distribution parameters subject to order restriction date back to the 50's, starting with Ayer *et al.* [2] and Brunk [7]. Comprehensive treatises on isotonic regression include Barlow *et al.* [3] and Robertson *et al.* [29]. For improvements and extensions of the PAVA approach to more general order

restrictions, see Best and Chakravarti [6], Dykstra [13], and Lee [23], among others.

The additive model was originally suggested by Friedman and Stuetzle [14] and popularized by Hastie and Tibshirani [18] as a way to accommodate the so-called curse of dimensionality in a multivariate setting. It assumes that the regression function is the sum of one-dimensional univariate functions. Buja *et al.* [9] proposed the backfitting algorithm as a practical method for estimating additive models. It consists in iteratively fitting, in each direction, the partial residuals from earlier steps until convergence is achieved.

In the present context, the backfitting algorithm is applied to estimate the univariate regression function r in (2) by alternating isotonic and antitonic regressions on the partial residuals in order to estimate the additive components u and b of the Jordan decomposition (2). Guyader *et al.* [15] introduced the I.I.R. algorithm and investigated its finite sample size behavior. Among other results, the authors state that the sequence $\hat{r}_n^{(k)}$ of estimators obtained in this way converges, when increasing the number k of iterations, to an interpolant of the raw data (see Section 2 below for details). As any interpolant overfits the data, iterating the procedure until convergence is not desirable. In the present paper, we go one step further and prove the consistency of this estimator. Before going into more details, let us specify why it is impossible to apply known results from isotonic regression theory as well as from additive models literature.

First, the consistency of the PAVA estimator was established by Brunk [7] and Hanson *et al.* [17]. Brunk [8] proved its cube-root convergence at a fixed point and obtained the pointwise asymptotic distribution, and Durot [12] provided a central limit theorem for the L_p-error. We wish to emphasize that all these asymptotic results assume monotonicity of the regression function r. In our context, at each stage of the iterative process, we apply an isotonic regression to an arbitrary function (of bounded variation). Consequently, their consistency theorems do not apply. Our first result, namely Theorem 1, is to prove the consistency of an isotonic regression estimator for the L_2 projection of the regression function onto the cone of monotone increasing functions.

Second, backfitting procedures and statistical properties of the resulting estimators have been studied in a linear framework by Härdle and Hall [19], Opsomer and Ruppert [27], Mammen, Linton and Nielsen [24], Horowitz, Klemelä and Mammen [20]. Alternative estimation procedures for additive models have been considered by Kim, Linton and Hengartner [22], and by Hengartner and Sperlich [21]. However, as the solution of (4) can be seen as the metric projection of the raw data onto the cone consisting in vectors with increasing components, the isotonic regression is *not* a linear smoother. It follows that the results in the previous references do *not* apply in the study of the I.I.R procedure. Interestingly, backfitted estimators in a non-linear case have also been studied by Mammen and Yu [25]. Specifically, in a multivariate setting, they assume that the regression function is the sum of isotonic one-dimensional functions, and estimate each component by iterating the PAVA in a backfitting fashion. However, as I.I.R. consists in applying

monotone regressions to a non-monotone regression function, we can not share their results either.

As mentioned before, the main result addressed in this paper, i.e. Theorem 2, states the consistency of the I.I.R. estimator. Denoting $\hat{r}_n^{(k)}$ the I.I.R. estimator resulting from k iterations of the algorithm, we prove the existence of a sequence of iterations (k_n), increasing with the sample size n, such that

$$\mathbb{E}\left[\|\hat{r}_n^{(k_n)} - r\|^2\right] \underset{n\to\infty}{\longrightarrow} 0$$

where $\|.\|$ is the quadratic norm with respect to the law μ of X. Our analysis identifies two error terms: an estimation error that comes from the isotonic regression, and an approximation error that is governed by the number of iterations k.

The approximation term can be controlled by increasing the number of iterations. This is made possible thanks to the interpretation of I.I.R. as a von Neumann's algorithm, and by applying related results in convex analysis (see Proposition 3). Let us remark that, as far as we know, rates of convergence of von Neumann's algorithm have not yet been studied in the context of bounded variation functions. Hence, at this time, it seems difficult to establish rates of convergence for the I.I.R. estimator without further restrictions on the shape of the underlying regression function. Thus, the results we present here may be considered as a starting point in the study of novel methods which would consist in applying isotonic regression with no particular shape assumption on the regression function.

The remainder of the paper is organised as follows. We first give further details and notations about the construction of I.I.R. in Section 2. The general consistency result for isotonic regression is given in Section 3. The main result of this article, the consistency of I.I.R., is established in Section 4, and we show how the algorithm can be stopped in practice using a data splitting procedure. Most of the proofs are postponed to Section 5, while related technical results are gathered in Section 6.

2 The I.I.R. procedure

For completeness, we recall the notations and some of the results presented in Guyader *et al.* [15]. Denote by $y = (y_1, \ldots, y_n)$ the vector of observations corresponding to the ordered sample $x_1 = X_{(1)} < \ldots < X_{(n)} = x_n$. We implicitly assume in this writing that the law μ of X has no atoms. Let us introduce the isotone cone $\mathcal{C}_n^+$:

$$\mathcal{C}_n^+ = \{u = (u_1, \ldots, u_n) \in \mathbb{R}^n : u_1 \leq \ldots \leq u_n\}.$$

We denote by $\mathrm{iso}(y)$ (resp. $\mathrm{anti}(y)$) the metric projection of y with respect to the Euclidean norm onto the isotone cone $\mathcal{C}_n^+$ (resp. $\mathcal{C}_n^- = -\mathcal{C}_n^+$):

$$\mathrm{iso}(y) = \underset{u \in \mathcal{C}_n^+}{\mathrm{argmin}} \frac{1}{n} \sum_{i=1}^{n} (y_i - u_i)^2 = \underset{u \in \mathcal{C}_n^+}{\mathrm{argmin}} \|y - u\|_n^2$$
$$\mathrm{anti}(y) = \underset{b \in \mathcal{C}_n^-}{\mathrm{argmin}} \frac{1}{n} \sum_{i=1}^{n} (y_i - b_i)^2 = \underset{b \in \mathcal{C}_n^-}{\mathrm{argmin}} \|y - b\|_n^2.$$

The backfitting algorithm consists in updating each component by smoothing the partial residuals, i.e., the residuals resulting from the current estimate in the other direction. Thus, the Iterative Isotonic Regression algorithm goes like this:

Algorithm 1 Iterative Isotonic Regression (I.I.R.)

(1) Initialization: $\hat{b}_n^{(0)} = \left(\hat{b}_1^{(0)}[1], \ldots, \hat{b}_n^{(0)}[n]\right) = 0$

(2) Cycle: for $k \geq 1$

$$\begin{aligned} \hat{u}_n^{(k)} &= \mathrm{iso}\left(y - \hat{b}_n^{(k-1)}\right) \\ \hat{b}_n^{(k)} &= \mathrm{anti}\left(y - \hat{u}_n^{(k)}\right) \\ \hat{r}_n^{(k)} &= \hat{u}_n^{(k)} + \hat{b}_n^{(k)}. \end{aligned}$$

(3) Iterate (2) until a stopping condition to be specified is achieved.

Guyader *et al.* [15] prove that the terms of the decomposition $\hat{r}_n^{(k)} = \hat{u}_n^{(k)} + \hat{b}_n^{(k)}$ have singular Stieltjes measures. Furthermore, by starting with isotonic regression, the terms $\hat{u}_n^{(k)}$ have all the same empirical mean as the original data y, while all the $\hat{b}_n^{(k)}$ are centered. Hence, for each k, the decomposition $\hat{r}_n^{(k)} = \hat{u}_n^{(k)} + \hat{b}_n^{(k)}$ satisfies the discrete translation of condition (3) so that it is unique (identifiable).

Algorithm 1 returns vectors of fitted values that we extend into piecewise constant functions defined on the interval $[0, 1]$. Specifically, the vector $\hat{u}_n^{(k)} = (\hat{u}_n^{(k)}[1], \ldots, \hat{u}_n^{(k)}[n])$ is associated to the real-valued function $\hat{u}_n^{(k)}$ defined on $[0, 1]$ by

$$\hat{u}_n^{(k)}(x) = \hat{u}_n^{(k)}[1] \mathbb{1}_{[0, X_{(2)})}(x) + \sum_{i=2}^{n-1} \hat{u}_n^{(k)}[i] \mathbb{1}_{[X_{(i)}, X_{(i+1)})}(x) + \hat{u}_n^{(k)}[n] \mathbb{1}_{[X_{(n)}, 1]}(x). \tag{5}$$

Observe that our definition of $\hat{u}_n^{(k)}(x)$ makes it right-continuous. Obviously, equivalent formulations hold for $\hat{b}_n^{(k)}$ and $\hat{r}_n^{(k)}$ as well.

Figure 1 illustrates the application of I.I.R. on an example. The top left-hand side displays the regression function r, and $n = 100$ points (x_i, y_i), with $y_i = r(x_i) + \varepsilon_i$, where the ε_i's are Gaussian centered random variables. The three other figures show the estimations

$\hat{r}_n^{(k)}$ obtained on this sample for $k = 1, 10$, and $1,000$ iterations. According to (5), our method fits a piecewise constant function. Moreover, increasing the number of iterations tends to increase the number of jumps.

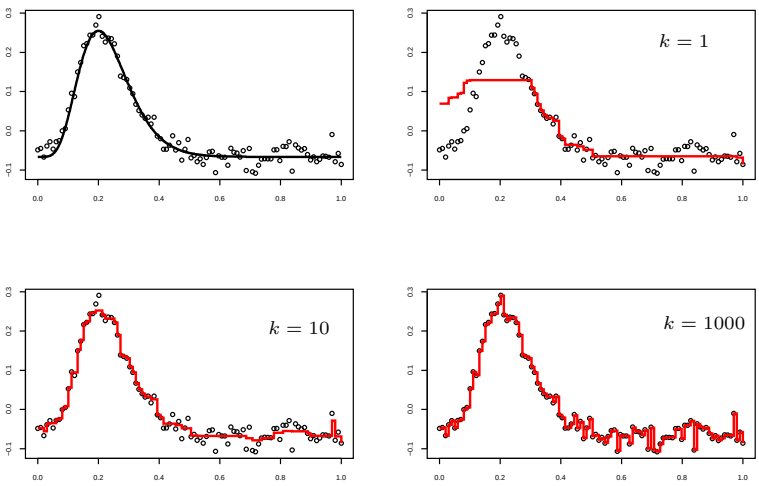

Figure 1: Application of the I.I.R. algorithm for $k = 1, 10$, and $1,000$ iterations.

The bottom right figure illustrates that, as established in Guyader *et al.* [15], for fixed sample size n, the function $\hat{r}_n^{(k)}(x)$ converges to an interpolant of the data when the number of iterations k tends to infinity, i.e., for all $i = 1, \ldots, n$,

$$\lim_{k \to \infty} \hat{r}_n^{(k)}(x_i) = y_i.$$

One interpretation of the above result is that increasing the number of iterations leads to overfitting. Thus, iterating the procedure until convergence is not desirable. On the other hand, as illustrated on Figure 1, iterations beyond the first step typically improve the fit. This suggests that the bias-variance trade-off is governed by the number of iterations and we need to couple the I.I.R. algorithm with a stopping rule. This will be discussed at the end of Section 4.

3 Isotonic regression: a general result of consistency

In this section, we focus on the first half step of the algorithm, which consists in applying isotonic regression to the original data. To simplify the notations, we omit in this section the exponent related to the number of iterations k, and simply denote $\hat{u}_n$ the isotonic regression on the data, that is,

$$\hat{u}_n = \underset{u \in \mathcal{C}_n^+}{\operatorname{argmin}} \|y - u\|_n = \underset{u \in \mathcal{C}_n^+}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (y_i - u_i)^2 .$$

Let u_+ denote the closest non-decreasing function to the regression function r with respect to the $L_2(\mu)$ norm. Thus, u_+ is defined as

$$u_+ = \underset{u \in \mathcal{C}^+}{\operatorname{argmin}} \|r - u\| = \underset{u \in \mathcal{C}^+}{\operatorname{argmin}} \int_{[0,1]} (r(x) - u(x))^2 \mu(dx),$$

where $\mathcal{C}^+$ denotes the cone of non-decreasing functions in $L_2(\mu)$. Since $\mathcal{C}^+$ is closed and convex, the metric projection u_+ exists and is unique in $L_2(\mu)$.

For mathematical purpose, we also introduce u_n, the result from applying isotonic regression to the sample $(x_i, r(x_i))$, $i = 1, \ldots, n$, that is

$$u_n = \underset{u \in \mathcal{C}_n^+}{\operatorname{argmin}} \|r - u\|_n = \underset{u \in \mathcal{C}_n^+}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (r(x_i) - u_i)^2 . \tag{6}$$

Finally, we note that, since r is bounded (say, by a constant denoted C in all what follows) so are u_+ and u_n, independently of the sample size n (see for example Lemma 2 in Anevski and Soulier [1]). Figure 2 displays the three terms involved.

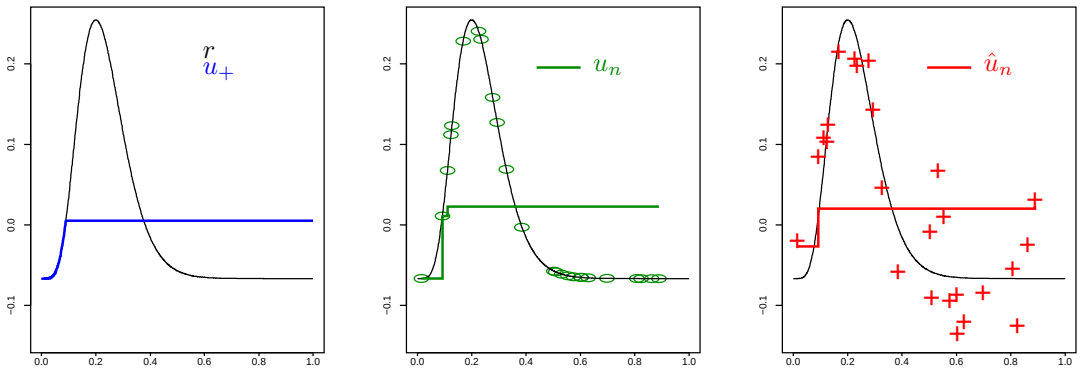

Figure 2: Isotonic regression on a non-monotone regression function.

The main result of this section states that, under a mild assumption on the noise ε,

$$\mathbb{E}\left[\|\hat{u}_n - u_+\|^2\right] \underset{n \to \infty}{\longrightarrow} 0,$$

where the expectation is taken with respect to the sample $\mathcal{D}_n$. Our analysis decomposes $\|\hat{u}_n - u_+\|$ into two distinct terms, namely:

$$\|\hat{u}_n - u_+\| \leq \|\hat{u}_n - u_n\| + \|u_n - u_+\|.$$

As $\|u_n - u_+\|$ does not depend on the response variable Y_i, one could interpret it as a bias term, whereas $\|\hat{u}_n - u_n\|$ plays the role of a variance term.

Throughout this section, our results are stated for both the empirical norm $\|.\|_n$ and the $L_2(\mu)$ norm $\|.\|$, as both are informative. The following proposition states the convergence of the bias term (its proof is postponed to Section 5.1).

Proposition 1 *With the previous notations, we have*

$$\lim_{n\to\infty} \|u_n - u_+\|_n = 0 \qquad a.s.,$$

and

$$\lim_{n\to\infty} \|u_n - u_+\| = 0 \qquad a.s.$$

Applying Lebesgue's dominated convergence Theorem ensures that both

$$\lim_{n\to\infty} \mathbb{E}\left[\|u_n - u_+\|_n^2\right] = 0 \qquad \text{and} \qquad \lim_{n\to\infty} \mathbb{E}\left[\|u_n - u_+\|^2\right] = 0.$$

Analysis of the variance term requires assumptions on the noise ε. More precisely, we will make two types of hypothesis. More details about sub-Gaussian random variables may be found for example in Chapter 1 of Buldygin and Kozachenko [10].

Assumption [A] The random variable ε satisfies $\mathbb{E}[\varepsilon] = 0$ and $\mathbb{E}[\varepsilon^2] = \sigma^2$.
Assumption [B] The random variable ε is centered and sub-Gaussian, i.e., there exists a number $a \in [0, \infty)$ such that the inequality

$$\mathbb{E}[\exp(t\varepsilon)] \leq \exp\left\{\frac{a^2t^2}{2}\right\}$$

holds for all $t \in \mathbb{R}$.

The proof of the following result is given in Section 5.2.

Proposition 2 *Under Assumption [A], we have*

$$\lim_{n\to\infty} \mathbb{E}\left[\|\hat{u}_n - u_n\|_n^2\right] = 0.$$

Under Assumption [B], we have

$$\lim_{n\to\infty} \mathbb{E}\left[\|\hat{u}_n - u_n\|^2\right] = 0.$$

Combining Proposition 1 and Proposition 2 yields the following theorem. This result generalizes the consistency of isotonic regression when applied in a more general context than the one of monotone functions.

Theorem 1 *Consider the model $Y = r(X) + \varepsilon$, with $r : [0,1] \to \mathbb{R}$ a bounded function belonging to $L_2(\mu)$, where μ is a non-atomic distribution on $[0,1]$. Denote u_+ and $\hat{u}_n$ the functions resulting from the isotonic regression applied on r and on the sample $\mathcal{D}_n$, respectively. Then, under Assumption [A], we have*

$$\mathbb{E}\left[\|\hat{u}_n - u_+\|_n^2\right] \to 0,$$

when the sample size n tends to infinity. Under Assumption [B], we have

$$\mathbb{E}\left[\|\hat{u}_n - u_+\|^2\right] \to 0,$$

when the sample size n tends to infinity.

This result will be of constant use when iterating our algorithm. This is the topic of the upcoming section.

4 Consistency of iterative isotonic regression

We now proceed with our main result, which states that there exists a sequence of numbers of iterations (k_n), increasing with the sample size n, such that

$$\mathbb{E}\left[\|\hat{r}_n^{(k_n)} - r\|^2\right] \underset{n\to\infty}{\longrightarrow} 0.$$

In order to control the expectation of the L_2 distance between the estimator $\hat{r}_n^{(k)}$ and the regression function r, we shall split $\|\hat{r}_n^{(k)} - r\|$ as follows: let $r^{(k)}$ be the result from applying the algorithm on the regression function r itself k times, that is $r^{(k)} = u^{(k)} + b^{(k)}$, where

$$u^{(k)} = \underset{u\in\mathcal{C}^+}{\operatorname{argmin}} \|r - b^{(k-1)} - u\| \qquad \text{and} \qquad b^{(k)} = \underset{b\in\mathcal{C}^-}{\operatorname{argmin}} \|r - u^{(k)} - b\|.$$

We then upper-bound

$$\|\hat{r}_n^{(k)} - r\| \leq \|r^{(k)} - r\| + \|\hat{r}_n^{(k)} - r^{(k)}\|. \tag{7}$$

In this decomposition, the first term is an approximation error, while the second one corresponds to an estimation error.

Figure 3 displays the function $r^{(k)}$ for two particular values of k. One can see that, after k steps of the algorithm, there generally remains an approximation error $\|r^{(k)} - r\|$. Nonetheless, one also observes that this error decreases when iterating the algorithm.

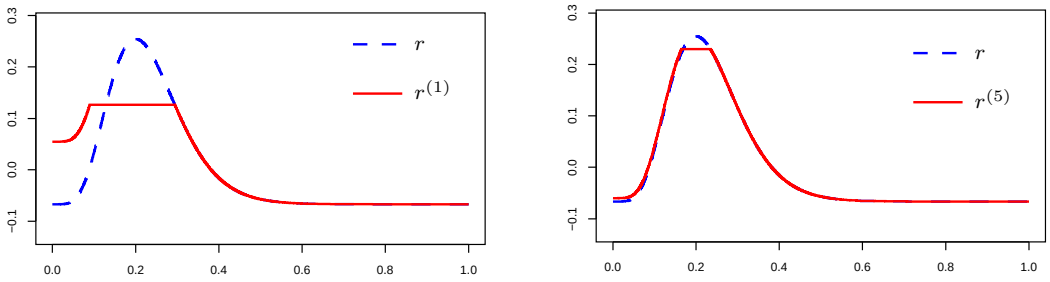

Figure 3: Decreasing of the approximation error $\|r^{(k)} - r\|$ with k.

The following proposition states that the approximation error can indeed be controlled by increasing the number of iterations k. Its proof relies on the interpretation of I.I.R. as a von Neumann's algorithm.

Proposition 3 *Assume that r is a right-continuous function of bounded variation and μ a non-atomic law on $[0,1]$. Then the approximation term $\|r^{(k)} - r\|$ tends to 0 when the number of iterations grows:*

$$\lim_{k\to\infty} \|r^{(k)} - r\| = 0,$$

where $\|.\|$ denotes the quadratic norm in $L_2(\mu)$.

Von Neumann's algorithm originally solved the problem of finding the projection of a given point onto the intersection of two closed subspaces. Since then, many related methods have extended the primary idea to the case of closed convex sets in Hilbert spaces (see Deutsch [11], Bauschke and Borwein [4] and references therein for further details).

Figure 4 provides a very simple interpretation of the I.I.R algorithm in terms of von Neumann sequences: namely, it illustrates that the sequences of functions $u^{(k)}$ and $r - b^{(k)}$ might be seen as alternate projections onto the cone $\mathcal{C}^+$ and the translated cone $r + \mathcal{C}^+ = \{r + u, u \in \mathcal{C}^+\}$ respectively. This geometric interpretation and its application to establish Proposition 3 are justified in Section 5.3.

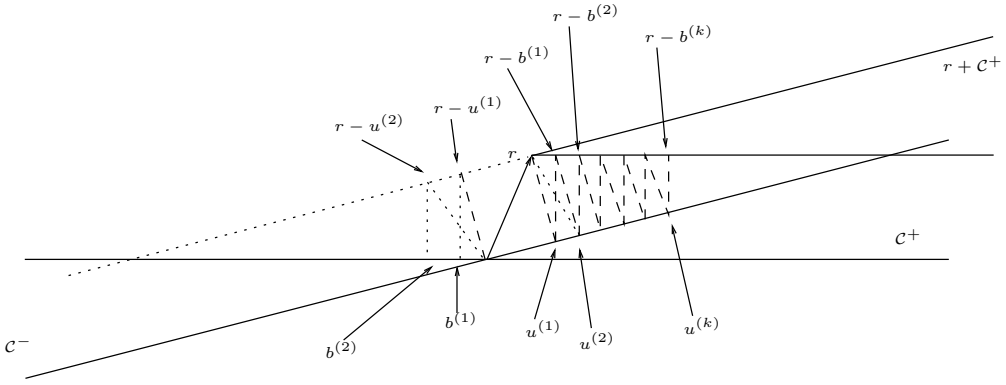

Figure 4: Interpretation of I.I.R. as a von Neumann's algorithm.

Coming back to (7), we further decompose the estimation error into a bias and a variance term to obtain

$$\underbrace{\|\hat{r}_n^{(k)} - r\|}_{\text{Estimation}} \leq \overbrace{\|\hat{r}_n^{(k)} - r_n^{(k)}\|}^{\downarrow \text{ Variance}} + \overbrace{\|r_n^{(k)} - r^{(k)}\|}^{\downarrow \text{ Bias}} + \underbrace{\|r^{(k)} - r\|}_{\text{Approximation}}.$$

The function $r_n^{(k)}$ results from k iterations of the algorithm on the sample $(x_i, r(x_i))$, $i = 1, \ldots, n$, and can be seen as the equivalent of the function u_n defined in (6). This decomposition allows us to make use of the consistency results of the previous section, and to control the estimation error when the sample size n goes to infinity. We now state the main theorem of this paper, whose proof is detailed in Section 5.4.

Theorem 2 *Consider the model $Y = r(X) + \varepsilon$, where $r : [0,1] \to \mathbb{R}$ is a right-continuous function of bounded variation, μ a non-atomic distribution on $[0,1]$, and ε a centered and*

sub-Gaussian random variable. Then there exists a sequence of numbers of iterations (k_n) *such that, for any* n, $k_n \leq n$ *and*

$$\mathbb{E}\left[\|\hat{r}_n^{(k_n)} - r\|^2\right] \underset{n\to\infty}{\longrightarrow} 0,$$

where $\|.\|$ *denotes the quadratic norm in* $L_2(\mu)$.

In the remainder of this section, we present a data-dependent way for choosing the number of iterations k_n and show that, for bounded Y, this procedure is consistent. To this end, we split the sample $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ in two parts, denoted by $\mathcal{D}_n^\ell$ (learning set) and $\mathcal{D}_n^t$ (testing set), of size $\lfloor n/2 \rfloor$ and $n - \lfloor n/2 \rfloor$, respectively. The first half is used to construct the I.I.R. estimate

$$\hat{r}_{\lfloor n/2 \rfloor}^{(k)}(x, \mathcal{D}_n^\ell)$$

for each $k \in \mathcal{K} = \{1, \ldots, \lfloor n/2 \rfloor\}$. The second half is used to choose k by picking $\hat{k}_n \in \mathcal{K}$ to minimize the empirical risk on the testing set, that is

$$\hat{k}_n = \underset{k\in\mathcal{K}}{\operatorname{argmin}} \frac{1}{n - \lfloor n/2 \rfloor} \sum_{i=\lfloor n/2 \rfloor + 1}^{n} \left(Y_i - \hat{r}_{\lfloor n/2 \rfloor}^{(k)}(X_i, \mathcal{D}_n^\ell)\right)^2.$$

Define the estimate

$$\hat{r}_n^{(\hat{k}_n)}(x) = \hat{r}_{\lfloor n/2 \rfloor}^{(\hat{k}_n)}(x, \mathcal{D}_n^\ell),$$

and note that $\hat{r}_n^{(\hat{k}_n)}$ depends on the entire data $\mathcal{D}_n$. The following theorem ensures the consistency of this method, provided that Y is assumed bounded. The proof is detailed in Section 5.5.

Theorem 3 *Suppose that* $|Y| \leq L$ *almost surely, and let* $\hat{r}_n^{(\hat{k}_n)}$ *be the I.I.R. estimate with* $\hat{k}_n \in \mathcal{K} = \{1, \ldots, \lfloor n/2 \rfloor\}$ *chosen by data-splitting. Then*

$$\mathbb{E}[\|\hat{r}_n^{(\hat{k}_n)} - r\|^2] \to 0,$$

when n *goes to infinity.*

To conclude this section, let us notice that the choice of a stopping criterion as a model selection also suggests stopping rules based, for example, on Akaike Information Criterion, Bayesian Information Criterion or Generalized Cross Validation. These criteria can be written in the generic form

$$\underset{p}{\operatorname{argmin}} \left\{ \log \frac{1}{n} \operatorname{RSS}(p) + \phi(p) \right\}, \tag{8}$$

where RSS denotes the residual sum of squares and ϕ is an increasing function. The parameter p stands for the number (or equivalent number) of parameters. For isotonic regression, we refer to Meyer and Woodroofe [26] to consider that the number of jumps provides the effective dimension of the model. Therefore, a natural extension for I.I.R. is to replace p by the number of jumps of $\hat{r}_n^{(k)}$ in (8). The comparisons of these criteria and the practical behavior of the I.I.R. procedure will be addressed elsewhere by the authors.

5 Proofs

5.1 Proof of Proposition 1

For g and h two functions from $[0,1]$ to $[-C,C]$, we denote $\Delta_n(g-h)$ the random variable

$$\Delta_n(g-h) = \|g-h\|_n^2 - \|g-h\|^2 = \frac{1}{n}\sum_{i=1}^{n}\left\{(g(X_i)-h(X_i))^2 - \mathbb{E}\left[(g(X)-h(X))^2\right]\right\}.$$

We first show that

$$\|r-u_n\|_n \to \|r-u_+\| \qquad a.s. \tag{9}$$

To this end, we proceed in two steps, proving in a first time that

$$\limsup \|r-u_n\|_n \leq \|r-u_+\| \qquad a.s., \tag{10}$$

and in a second time that

$$\liminf \|r-u_n\|_n \geq \|r-u_+\| \qquad a.s. \tag{11}$$

For the first inequality, let us denote

$$A_n = \left\{|\Delta_n(r-u_+)| > n^{-1/3}\right\} = \left\{|\|r-u_+\|_n^2 - \|r-u_+\|^2| > n^{-1/3}\right\}.$$

By the definition of u_n, note that for all n,

$$\|r-u_n\|_n \leq \|r-u_+\|_n,$$

so that on $\overline{A_n}$,

$$\|r-u_n\|_n^2 \leq \|r-u_+\|_n^2 \leq \|r-u_+\|^2 + n^{-1/3}.$$

Consequently

$$B_n = \left\{\|r-u_n\|_n^2 \leq \|r-u_+\|^2 + n^{-1/3}\right\} \supset \overline{A_n}.$$

Therefore

$$\mathbb{P}\left(\liminf B_n\right) \geq \mathbb{P}\left(\liminf \overline{A_n}\right) = 1 - \mathbb{P}\left(\limsup A_n\right).$$

Since $|r(X_i) - u_+(X_i)| \leq 2C$, Hoeffding's inequality gives for all $t > 0$

$$\mathbb{P}\left(|\Delta_n(r-u_+)| > t\right) \leq 2\exp\left(-\frac{t^2 n}{8C^2}\right).$$

Taking $t = n^{-1/3}$, we deduce that

$$\mathbb{P}(A_n) = \mathbb{P}\left(|\Delta_n(r-u_+)| > n^{-1/3}\right) \leq 2\exp\left(-\frac{n^{1/3}}{8C^2}\right).$$

By Borel-Cantelli Lemma, we conclude that $\mathbb{P}\left(\limsup A_n\right) = 0$, and hence $\mathbb{P}\left(\liminf B_n\right) = 1$. On the set $\liminf B_n$, we have

$$\limsup \|r-u_n\|_n^2 \leq \|r-u_+\|^2,$$

which proves Equation (10).

Conversely, we now establish Equation (11). By definition of u_+, observe that for all n,

$$\|r - u_+\| \leq \|r - u_n\|.$$

Consider the sets

$$C_n = \left\{ \sup_{h \in \mathcal{C}^+_{[0,1]}} |\Delta_n(r-h)| > n^{-1/4} \right\} \text{ and } D_n = \left\{ \|r-u_n\|_n^2 \geq \|r-u_+\|^2 - n^{-1/4} \right\},$$

so that $\overline{C_n} \subset D_n$, and by applying Lemma 1,

$$\mathbb{P}\left(\liminf D_n\right) \geq 1 - \mathbb{P}\left(\limsup C_n\right) = 1.$$

On the set $\liminf D_n$, one has

$$\liminf \|r-u_n\|_n^2 \geq \|r-u_+\|^2,$$

which proves (11). Combining Equations (10) and (11) leads to (9).

Next, using Lemma 1 again, we get

$$\lim_{n\to\infty} \|r-u_n\|_n - \|r-u_n\| = 0 \qquad a.s.$$

Combined with (9), this leads to

$$\|r-u_n\| \to \|r-u_+\| \qquad a.s. \tag{12}$$

It remains to prove the almost sure convergence of u_n to u_+. For this, it suffices to use the parallelogram law. Indeed, noting $m_n = (u_n + u_+)/2$, we have

$$\|u_n - u_+\|^2 = 2\left(\|r-u_+\|^2 + \|u_n - r\|^2\right) - 4\|m_n - r\|^2.$$

Since both u_+ and u_n belong to the convex set $\mathcal{C}^+$, so does m_n. Hence $\|r-u_+\|^2 \leq \|r-m_n\|^2$, and

$$\|u_n - u_+\|^2 \leq 2\left(\|u_n - r\|^2 - \|r-u_+\|^2\right).$$

Combining this with (12), we conclude that

$$\lim_{n\to\infty} \|u_n - u_+\| = 0 \qquad a.s.$$

Finally, Lemma 1 guarantees the same result for the empirical norm, that is

$$\lim_{n\to\infty} \|u_n - u_+\|_n = 0 \qquad a.s.,$$

and the proof is complete.

5.2 Proof of Proposition 2

Let us denote $\langle \cdot, \cdot \rangle_n$ the inner product associated to the empirical norm $\|.\|_n$. Since isotonic regression corresponds to the metric projection onto the closed convex cone $\mathcal{C}_n^+$ with respect to this empirical norm, the vectors $\hat{u}_n$ and u_n are characterized by the following inequalities: for any vector $u \in \mathcal{C}_n^+$,

$$\langle y - \hat{u}_n, u - \hat{u}_n \rangle_n \leq 0 \tag{13}$$
$$\langle r - u_n, u - u_n \rangle_n \leq 0 \tag{14}$$

Setting $u = u_n$ in (13) and $u = \hat{u}_n$ in (14), we get

$$\langle y - \hat{u}_n, u_n - \hat{u}_n \rangle_n \leq 0 \qquad \text{and} \qquad \langle r - u_n, \hat{u}_n - u_n \rangle_n \leq 0.$$

Since $\varepsilon = y - r$, this leads to

$$\|\hat{u}_n - u_n\|_n^2 \leq \langle \varepsilon, \hat{u}_n - u_n \rangle_n. \tag{15}$$

Next, we have to use an approximation result, namely Lemma 3 in Section 6.2. The underlying idea is to exploit the fact that any non-decreasing bounded sequence can be approached by the element of a subspace H_n^+ at distance less than δ_n. Specifically, if C_n is an upper-bound for the absolute value of the considered non-decreasing bounded sequences, we can construct such a subspace H_n^+ with dimension N_n where $N_n = (8C_n^2)/\delta_n^2$. From now on, we will take $N_n \leq n$.

Let us introduce the vectors $\hat{h}_n$ and h_n defined by

$$\hat{h}_n = \inf_{h \in H_n^+} \|\hat{u}_n - h\|_n \qquad \text{and} \qquad h_n = \inf_{h \in H_n^+} \|u_n - h\|_n,$$

so that

$$\|\hat{u}_n - \hat{h}_n\|_n \leq \delta_n \qquad \text{and} \qquad \|u_n - h_n\|_n \leq \delta_n.$$

From this, we get

$$\begin{aligned}
\langle \varepsilon, \hat{u}_n - u_n \rangle_n =& \langle \varepsilon, \hat{u}_n - \hat{h}_n \rangle_n + \langle \varepsilon, \hat{h}_n - h_n \rangle_n + \langle \varepsilon, h_n - u_n \rangle_n \\
\leq& \|\hat{h}_n - h_n\|_n \left\langle \varepsilon, \frac{\hat{h}_n - h_n}{\|\hat{h}_n - h_n\|_n} \right\rangle_n + 2\delta_n \|\varepsilon\|_n \\
\leq& \left\{ \|\hat{h}_n - \hat{u}_n\|_n + \|\hat{u}_n - u_n\|_n + \|u_n - h_n\|_n \right\} \sup_{v \in H_n^+, \|v\|_n = 1} \langle \varepsilon, v \rangle_n + 2\delta_n \|\varepsilon\|_n \\
\leq& \{ \|\hat{u}_n - u_n\|_n + 2\delta_n \} \sup_{v \in H_n^+, \|v\|_n = 1} \langle \varepsilon, v \rangle_n + 2\delta_n \|\varepsilon\|_n.
\end{aligned}$$

According to (15), we deduce

$$\|\hat{u}_n - u_n\|_n^2 \leq \{ \|\hat{u}_n - u_n\|_n + 2\delta_n \} \sup_{v \in H_n^+, \|v\|_n = 1} \langle \varepsilon, v \rangle_n + 2\delta_n \|\varepsilon\|_n$$

so that

$$\|\hat{u}_n - u_n\|_n^2 \leq \{\|\hat{u}_n - u_n\|_n + 2\delta_n\} \|\pi_{H_n^+}(\varepsilon)\|_n + 2\delta_n \|\varepsilon\|_n,$$

where $\pi_{H_n^+}(\varepsilon)$ stands for the metric projection of ε onto H_n^+. Put differently, we have

$$\|\hat{u}_n - u_n\|_n^2 \leq \|\hat{u}_n - u_n\|_n \times \|\pi_{H_n^+}(\varepsilon)\|_n + 2\delta_n \left\{\|\pi_{H_n^+}(\varepsilon)\|_n + \|\varepsilon\|_n\right\},$$

and taking the expectation on both sides leads to

$$\mathbb{E}\left[\|\hat{u}_n - u_n\|_n^2\right] \leq \mathbb{E}\left[\|\hat{u}_n - u_n\|_n \times \|\pi_{H_n^+}(\varepsilon)\|_n\right] + 2\delta_n \left\{\mathbb{E}\left[\|\pi_{H_n^+}(\varepsilon)\|_n\right] + \mathbb{E}\left[\|\varepsilon\|_n\right]\right\}.$$

If we denote

$$\left\{\begin{array}{rl} x & = \sqrt{\mathbb{E}\left[\|\hat{u}_n - u_n\|_n^2\right]} \\ \alpha_n & = \sqrt{\mathbb{E}\left[\|\pi_{H_n^+}(\varepsilon)\|_n^2\right]} \\ \beta_n & = 2\delta_n \left\{\mathbb{E}\left[\|\pi_{H_n^+}(\varepsilon)\|_n\right] + \mathbb{E}\left[\|\varepsilon\|_n\right]\right\} \end{array}\right.$$

an application of Cauchy-Schwarz inequality gives

$$x^2 - \alpha_n x - \beta_n \leq 0 \;\Rightarrow\; x \leq \frac{\alpha_n + \sqrt{\alpha_n^2 + 4\beta_n}}{2},$$

which means that

$$\mathbb{E}\left[\|\hat{u}_n - u_n\|_n^2\right] \leq \left(\frac{\alpha_n + \sqrt{\alpha_n^2 + 4\beta_n}}{2}\right)^2.$$

Under Assumption [A], a straightforward computation shows that

$$\mathbb{E}\left[\|\pi_{H_n^+}(\varepsilon)\|_n^2\right] = \frac{1}{n}\mathbb{E}\left[(\pi_{H_n^+}\varepsilon)'(\pi_{H_n^+}\varepsilon)\right] = \frac{1}{n}\mathbb{E}\left[\mathrm{tr}\left((\pi_{H_n^+}\varepsilon)'(\pi_{H_n^+}\varepsilon)\right)\right] = \frac{1}{n}\,\mathrm{tr}\left(\mathbb{E}\left[\varepsilon\varepsilon'\right]\pi_{H_n^+}\right),$$

and since H_n^+ has dimension $N_n = (8C_n^2)/\delta_n^2$, this gives

$$\mathbb{E}\left[\|\pi_{H_n^+}(\varepsilon)\|_n^2\right] = \sigma^2\frac{N_n}{n} \;\Rightarrow\; \alpha_n = \sigma\sqrt{\frac{N_n}{n}} = \sigma \times \frac{2\sqrt{2}C_n}{\delta_n\sqrt{n}}.$$

Set $\delta_n = n^{-\alpha}$ and $C_n = n^{\gamma}$ with α and γ strictly positive, it then follows that α_n goes to zero when n goes to infinity, provided that $\alpha + \gamma < 1/2$. Moreover, Jensen's inequality implies

$$\beta_n \leq 2\delta_n(\alpha_n + \sigma).$$

As both δ_n and α_n tend to zero when n goes to infinity, we have proved the first part of Proposition 2, that is

$$\lim_{n\to\infty} \mathbb{E}\left[\|\hat{u}_n - u_n\|_n^2\right] = 0.$$

For the second part of Proposition 2, introducing the random variable

$$\Delta_n = \|\hat{u}_n - u_n\|_n^2 - \|\hat{u}_n - u_n\|^2,$$

our goal is to show that $\mathbb{E}[\Delta_n]$ goes to zero when n goes to infinity. For this, let us denote

$$M_n = \sup_{1\leq i\leq n} |Y_i|,$$

and consider the decomposition

$$\mathbb{E}[\Delta_n] = \mathbb{E}[\Delta_n \mathbb{1}_{M_n<C_n}] + \mathbb{E}[\Delta_n \mathbb{1}_{M_n\geq C_n}],$$

where, as previously, $C_n = n^\gamma$ with $0 < \gamma < 1/2$. Notice that, since $\|r\|_\infty \leq C < \infty$, one has $\|r\|_\infty \leq C_n$ for n large enough. Then, on the set $\{M_n \geq C_n\}$, both $\hat{u}_n$ and u_n are bounded by M_n. From the definition of Δ_n, we deduce that, for n large enough,

$$\mathbb{E}[\Delta_n \mathbb{1}_{M_n\geq C_n}] \leq 8\mathbb{E}[M_n^2 \mathbb{1}_{M_n\geq C_n}].$$

Recall that for any non negative random variable X,

$$\mathbb{E}[X] = \int_0^{+\infty} \mathbb{P}\left(X \geq t\right) dt,$$

whose application in our case gives

$$\mathbb{E}[\Delta_n \mathbb{1}_{M_n\geq C_n}] \leq 8C_n^2\, \mathbb{P}(M_n \geq C_n) + 8\int_{C_n^2}^{+\infty} \mathbb{P}\left(M_n \geq \sqrt{t}\right) dt.$$

Then, observing that

$$\mathbb{P}(|Y_i| \geq \sqrt{t}) \leq \mathbb{P}(|\varepsilon_i| \geq \sqrt{t} - C),$$

we get, from the fact that the ε_i's are i.i.d.,

$$\mathbb{P}\left(M_n \geq \sqrt{t}\right) \leq 1 - \Big(1 - \mathbb{P}(|\varepsilon| \geq \sqrt{t} - C)\Big)^n.$$

Since ε is sub-Gaussian (Assumption [B]), there exists $\tau > 0$ such that

$$\mathbb{P}(|\varepsilon| \geq \sqrt{t} - C) \leq 2\exp\left(-\frac{(\sqrt{t} - C)^2}{2\tau^2}\right),$$

for all $t \geq C^2$ (see Lemma 1.3 in [10]). Then the inequality $(1-x)^n \geq 1 - nx$, valid for any x small enough, leads to

$$\mathbb{P}\left(M_n \geq \sqrt{t}\right) \leq 2n\exp\left(-\frac{(\sqrt{t} - C)^2}{2\tau^2}\right).$$

Thus, for n large enough and taking into account that $C_n = n^\gamma$, we get

$$\int_{C_n^2}^{+\infty} \mathbb{P}\left(M_n \geq \sqrt{t}\right) dt \leq 2n\int_{n^{2\gamma}}^{+\infty} \exp\left(-\frac{(\sqrt{t} - C)^2}{2\tau^2}\right) dt,$$

which goes to zero when n goes to infinity. In the same way, one has

$$C_n^2 \, \mathbb{P}(M_n \geq C_n) \leq 2nC_n^2 \exp\left(-\frac{(C_n - C)^2}{2\tau^2}\right),$$

or, equivalently,

$$C_n^2 \, \mathbb{P}(M_n \geq C_n) \leq 2n^{1+2\gamma} \exp\left(-\frac{(n^\gamma - C)^2}{2\tau^2}\right),$$

which goes to zero when n goes to infinity. Therefore, we get

$$\mathbb{E}[\Delta_n \mathbb{1}_{M_n \geq C_n}] \to 0.$$

Next, we have

$$\mathbb{E}[\Delta_n \mathbb{1}_{M_n < C_n}] = \int_0^{+\infty} \mathbb{P}(\Delta_n \mathbb{1}_{M_n < C_n} \geq t) dt = \int_0^{+\infty} \mathbb{P}(\Delta_n \geq t, M_n < C_n) dt.$$

Again, if $M_n < C_n$, with n large enough so that $C_n \geq C$, then both $\hat{u}_n$ and u_n are bounded by C_n. As a consequence, from Lemma 2, we know that for any $t > 0$,

$$\mathbb{P}(\Delta_n \geq t, M_n < C_n) \leq \exp\left(2 \left\lceil \frac{64C_n^2}{t} \right\rceil \log n - \frac{t^2 n}{32C_n^2}\right).$$

Thus, setting

$$f_n(t) = \min\left(1, \exp\left(2 \left\lceil \frac{64C_n^2}{t} \right\rceil \log n - \frac{t^2 n}{32C_n^2}\right)\right),$$

we have

$$\mathbb{E}[\Delta_n \mathbb{1}_{M_n < C_n}] \leq \int_0^{+\infty} f_n(t) \, dt.$$

Then, it remains to see that for n large enough and for all $t \geq 0$, one has $f_n(t) \leq f_2(t)$. Since for all $t > 0$ fixed, $f_n(t)$ goes to 0 when n tends to infinity, Lebesgue's dominated convergence Theorem ensures that

$$\mathbb{E}[\Delta_n \mathbb{1}_{M_n < C_n}] \to 0.$$

This terminates the proof of Proposition 2.

5.3 Proof of Proposition 3

Consider the translated cone

$$r + \mathcal{C}^+ = \{r + u, u \in \mathcal{C}^+\}.$$

As mentioned above, Figure 4 provides a very simple interpretation of the algorithm: namely, it illustrates that the sequences of functions $u^{(k)}$ and $r - b^{(k)}$ might be seen as

alternate projections onto the cones $\mathcal{C}^+$ and $r + \mathcal{C}^+$. In what follows, we justify this illuminating geometric interpretation in a rigorous way, and we explain its key role in the proof of the convergence as k goes to infinity.

By definition, we have $u^{(1)} = P_{\mathcal{C}^+}(r)$ where $P_{\mathcal{C}^+}$ denotes the metric projection onto $\mathcal{C}^+$. Classical properties of projections ensure that

$$P_{r+\mathcal{C}^+}(u^{(1)}) = r + P_{\mathcal{C}^+}(u^{(1)} - r) = r - P_{\mathcal{C}^-}(r - u^{(1)}).$$

Coming back to the definition of $b^{(1)} = P_{\mathcal{C}^-}(r - u^{(1)})$, we are led to

$$r - b^{(1)} = P_{r+\mathcal{C}^+}(u^{(1)}).$$

In the same manner, since $u^{(2)} = P_{\mathcal{C}^+}(r - b^{(1)})$, we get

$$r - b^{(2)} = r - P_{\mathcal{C}^-}(r - u^{(2)}) = r + P_{\mathcal{C}^+}(r - u^{(2)}) = P_{r+\mathcal{C}^+}(u^{(2)}).$$

More generally, denoting $b^{(0)} = 0$, this yields for all $k \geq 1$ (see also Figure 4)

$$u^{(k)} = P_{\mathcal{C}^+}(r - b^{(k-1)}) \qquad \text{and} \qquad r - b^{(k)} = P_{r+\mathcal{C}^+}(u^{(k)}).$$

It remains to invoke Theorem 4.8 in Bauschke and Borwein [5] to conclude that

$$(r - b^{(k)}) - u^{(k)} = r - r^{(k)} \xrightarrow[k\to\infty]{} 0,$$

which ends the proof of Proposition 3.

5.4 Proof of Theorem 2

Coming back to the original notation, Theorem 1 states that

$$\lim_{n\to\infty} \mathbb{E}\left[\|\hat{u}_n^{(1)} - u^{(1)}\|_n^2\right] = 0 \qquad \text{and} \qquad \lim_{n\to\infty} \mathbb{E}\left[\|\hat{u}_n^{(1)} - u^{(1)}\|^2\right] = 0. \tag{16}$$

In the following, we show that this result still holds when applying the backfitting algorithm.

We first describe the end of the first step by showing that $\mathbb{E}\left[\|\hat{b}_n^{(1)} - b^{(1)}\|^2\right] \to 0$.
Recall the definitions

$$b^{(1)} = \operatorname*{argmin}_{b\in\mathcal{C}^-} \|r - u^{(1)} - b\| \qquad \text{and} \qquad \hat{b}_n^{(1)} = \operatorname*{argmin}_{b\in\mathcal{C}_n^-} \|y - \hat{u}_n^{(1)} - b\|_n.$$

In order to mimic the previous step, let us consider the vectors

$$\tilde{y} = y - u^{(1)} \qquad \text{and} \qquad \tilde{b}_n^{(1)} = \operatorname*{argmin}_{b\in\mathcal{C}_n^-} \|\tilde{y} - b\|_n,$$

so that

$$\tilde{y} = \left(r - u^{(1)}\right) + \varepsilon$$

and

$$\tilde{b}_n^{(1)} = \underset{b \in \mathcal{C}_n^-}{\operatorname{argmin}} \|(r - u^{(1)}) + \varepsilon - b\|_n.$$

To study the term $\|\tilde{b}_n^{(1)} - b^{(1)}\|$, one can apply *mutatis mutandis* the result of Theorem 1, replacing $\hat{u}_n^{(1)}$ by $\tilde{b}_n^{(1)}$, r by $r - u^{(1)}$, and isotonic regression by antitonic regression. Hence,

$$\lim_{n\to\infty} \mathbb{E}\left[\|\tilde{b}_n^{(1)} - b^{(1)}\|_n^2\right] = 0 \qquad \text{and} \qquad \lim_{n\to\infty} \mathbb{E}\left[\|\tilde{b}_n^{(1)} - b^{(1)}\|^2\right] = 0. \tag{17}$$

As projection reduces distances, we also have

$$\|\hat{b}_n^{(1)} - \tilde{b}_n^{(1)}\|_n \leq \|y - \hat{u}_n^{(1)} - \tilde{y}\|_n = \|\hat{u}_n^{(1)} - u^{(1)}\|_n.$$

Thanks to equations (16) and (17), we deduce

$$\mathbb{E}\left[\|\hat{b}_n^{(1)} - b^{(1)}\|_n^2\right] \leq 2 \times \left\{\mathbb{E}\left[\|\hat{b}_n^{(1)} - \tilde{b}_n^{(1)}\|_n^2\right] + \mathbb{E}\left[\|\tilde{b}_n^{(1)} - b^{(1)}\|_n^2\right]\right\} \to 0.$$

Invoking the same arguments as those at the end of the proof of Proposition 2, we also have

$$\lim_{n\to\infty} \mathbb{E}\left[\|\hat{b}_n^{(1)} - b^{(1)}\|^2\right] = 0.$$

Finally, at the end of the first iteration, we have

$$\mathbb{E}\left[\|\hat{r}_n^{(1)} - r^{(1)}\|^2\right] \leq 2 \times \left\{\mathbb{E}\left[\|\hat{u}_n^{(1)} - u^{(1)}\|^2\right] + \mathbb{E}\left[\|\hat{b}_n^{(1)} - b^{(1)}\|^2\right]\right\} \to 0.$$

For the beginning of the second iteration, consider this time

$$\hat{u}_n^{(2)} = \underset{u \in \mathcal{C}_n^+}{\operatorname{argmin}} \|y - \hat{b}_n^{(1)} - u\|_n \qquad \text{and} \qquad u^{(2)} = \underset{u \in \mathcal{C}^+}{\operatorname{argmin}} \|r - b^{(1)} - u\|.$$

Let us introduce

$$\tilde{y} = y - b^{(1)} = (r - b^{(1)}) + \varepsilon \qquad \text{and} \qquad \tilde{u}_n^{(2)} = \underset{u \in \mathcal{C}_n^+}{\operatorname{argmin}} \|\tilde{y} - u\|_n = \underset{u \in \mathcal{C}_n^+}{\operatorname{argmin}} \|(r - b^{(1)}) + \varepsilon - u\|_n.$$

We apply Theorem 1 again, replacing r by $r - b^{(1)}$, and $\hat{u}_n^{(1)}$ by $\tilde{u}_n^{(2)}$. This leads to

$$\lim_{n\to 0} \mathbb{E}\left[\|\tilde{u}_n^{(2)} - u^{(2)}\|_n^2\right] = 0.$$

Thanks to the reduction property of isotonic regression and using the conclusion of the first iteration, we get

$$\mathbb{E}\left[\|\hat{u}_n^{(2)} - \tilde{u}_n^{(2)}\|_n^2\right] \leq \mathbb{E}\left[\|y - \hat{b}_n^{(1)} - ((r - b^{(1)}) + \varepsilon)\|_n^2\right] = \mathbb{E}\left[\|\hat{b}_n^{(1)} - b^{(1)}\|_n^2\right] \to 0.$$

Therefore

$$\mathbb{E}\left[\|\hat{u}_n^{(2)} - u^{(2)}\|_n^2\right] \leq 2 \times \left\{\mathbb{E}\left[\|\hat{u}_n^{(2)} - \tilde{u}_n^{(2)}\|_n^2\right] + \mathbb{E}\left[\|\tilde{u}_n^{(2)} - u^{(2)}\|_n^2\right]\right\} \to 0,$$

and, as before, we also have

$$\lim_{n\to\infty} \mathbb{E}\left[\|\hat{u}_n^{(2)} - u^{(2)}\|^2\right] = 0.$$

The same scheme leads to $\lim_{n\to\infty} \mathbb{E}\left[\|\hat{b}_n^{(2)} - b^{(2)}\|^2\right] = 0$, so that

$$\mathbb{E}\left[\|\hat{r}_n^{(2)} - r^{(2)}\|^2\right] \leq 2 \times \left\{\mathbb{E}\left[\|\hat{u}_n^{(2)} - u^{(2)}\|^2\right] + \mathbb{E}\left[\|\hat{b}_n^{(2)} - b^{(2)}\|^2\right]\right\} \to 0.$$

By iterating this process, it is readily seen that, for all $k \geq 1$,

$$\lim_{n\to\infty} \mathbb{E}\left[\|\hat{r}_n^{(k)} - r^{(k)}\|^2\right] = 0,$$

which means that, at each iteration, the estimation error goes to 0 when the sample size tends to infinity.

We deduce that we can construct an increasing sequence (n_k) such that for each $k \geq 1$ and for all $n \geq n_k$

$$\mathbb{E}\left[\|\hat{r}_n^{(k)} - r^{(k)}\|\right] \leq \|r^{(k)} - r\| + \frac{1}{k}.$$

Notice that the term $\|r^{(k)} - r\|$ might be equal to zero (*e.g.*, $r^{(1)} = r$ if r is monotone), hence the additive term $1/k$ in the previous inequality. Consequently,

$$\mathbb{E}\left[\|\hat{r}_n^{(k)} - r\|\right] \leq 2\|r^{(k)} - r\| + \frac{1}{k}.$$

Then let us consider the sequence (k_n) defined as: $k_n = 0$ if $n < n_1$, $k_n = 1$ if $n_1 \leq n < n_2$, and so on. Obviously, one has $k_n \leq n$ for any n, (k_n) tends to infinity, and

$$\mathbb{E}\left[\|\hat{r}_n^{(k_n)} - r\|\right] \leq 2\|r^{(k_n)} - r\| + \frac{1}{k_n} \xrightarrow[n\to\infty]{} 0.$$

This ends the proof of Theorem 2.

5.5 Proof of Theorem 3

If $|Y| \leq L < \infty$ almost surely, the adaptation of Theorem 7.1 in Györfi *et al.* [16] to our setting reveals that, for any $\delta > 0$,

$$\mathbb{E}[\hat{r}_n^{(\hat{k}_n)}(X) - r(X)]^2 \leq (1+\delta) \inf_{k\in\mathcal{K}} \mathbb{E}[\hat{r}_{\lfloor n/2\rfloor}^{(k)}(X, \mathcal{D}_n^\ell) - r(X)]^2 + c\,\frac{1 + \ln\lfloor n/2\rfloor}{n - \lfloor n/2\rfloor}, \qquad (18)$$

for some positive constant c depending only on L and δ. Next, Theorem 2 says that there exists a sequence of numbers of iterations $(k_{\lfloor n/2 \rfloor})$ such that $k_{\lfloor n/2 \rfloor} \leq \lfloor n/2 \rfloor$ for any n, and

$$\mathbb{E}\left[\hat{r}_{\lfloor n/2 \rfloor}^{(k_{\lfloor n/2 \rfloor})}(X, \mathcal{D}_n^\ell) - r(X)\right]^2 \underset{n\to\infty}{\longrightarrow} 0.$$

Coming back to (18), this ensures that

$$\inf_{k\in\mathcal{K}} \mathbb{E}[\hat{r}_{\lfloor n/2 \rfloor}^{(k)}(X, \mathcal{D}_n^\ell) - r(X)]^2 \leq \mathbb{E}\left[\hat{r}_{\lfloor n/2 \rfloor}^{(k_{\lfloor n/2 \rfloor})}(X, \mathcal{D}_n^\ell) - r(X)\right]^2 \underset{n\to\infty}{\longrightarrow} 0.$$

For the second term of (18), one has clearly

$$\frac{1 + \ln\lfloor n/2 \rfloor}{n - \lfloor n/2 \rfloor} \underset{n\to\infty}{\longrightarrow} 0,$$

and the proof is complete.

6 Technical results

6.1 Concentration inequalities

Throughout the previous proofs, we repeatedly needed to pass from the empirical norm $\|.\|_n$ to the $L_2(\mu)$ norm $\|.\|$. This was made possible thanks to several exponential inequalities that we justify in this section.

Specifically, let g and h denote two mappings from $I = [0, 1]$ to $[-C, C]$, and consider the random variable

$$\Delta_n(g - h) = \frac{1}{n}\sum_{i=1}^n \left\{(g(X_i) - h(X_i))^2 - \mathbb{E}\left[(g(X) - h(X))^2\right]\right\} = \|g - h\|_n^2 - \|g - h\|^2.$$

In what follows, we focus on the concentration of $\Delta_n(g-h)$ around zero. First note that, since $|g(X_i) - h(X_i)| \leq 2C$, Hoeffding's inequality gives for all $t > 0$

$$\mathbb{P}\left(|\Delta_n(g - h)| > t\right) \leq 2\exp\left(-\frac{t^2 n}{8C^2}\right). \tag{19}$$

The following lemma goes one step further, by considering, for fixed g, the tail distribution of

$$\sup_{h\in\mathcal{C}^+_{[0,1]}} |\Delta_n(g - h)|.$$

For obvious reasons, this type of result is sometimes called a maximal inequality. The proof shares elements with the one of Theorem 3.1 of van de Geer and Wegkamp [30].

Lemma 1 *Let g be a function from $[0,1]$ to $[-C,C]$ and let $\mathcal{C}^+_{[0,1]}$ denote the set of non-decreasing functions from $[0,1]$ to $[-C,C]$. For any $\alpha \in (0,1/3)$, there exist positive real numbers c_1 and c_2 depending only on α and C and such that*

$$\mathbb{P}\left(\sup_{h\in\mathcal{C}^+_{[0,1]}} |\Delta_n(g-h)| > n^{-\alpha}\right) \leq c_1 \exp\left(-c_2 n^{1-2\alpha}\right).$$

Proof. The first step consists in showing that the mapping $h \mapsto \Delta_n(g-h)$ is Lipschitz. For any pair of functions h and $\tilde{h}$, we have

$$\begin{aligned}\Delta_n(g-h) - \Delta_n(g-\tilde{h}) =& \frac{1}{n}\sum_{i=1}^n \left\{2g(X_i) - h(X_i) - \tilde{h}(X_i)\right\}\left(\tilde{h}(X_i) - h(X_i)\right) \\ & - \mathbb{E}\left[\left\{2g(X) - h(X) - \tilde{h}(X)\right\}\left(\tilde{h}(X) - h(X)\right)\right].\end{aligned}$$

Since h and $\tilde{h}$ take values in $[-C,C]$, we get

$$|\Delta_n(g-h) - \Delta_n(g-\tilde{h})| \leq 4C \times \left\{\frac{1}{n}\sum_{i=1}^n |h(X_i) - \tilde{h}(X_i)| + \mathbb{E}\left[|h(X) - \tilde{h}(X)|\right]\right\},$$

and according to Jensen's inequality,

$$|\Delta_n(g-h) - \Delta_n(g-\tilde{h})| \leq 4C \times \left\{\|h-\tilde{h}\|_n + \|h-\tilde{h}\|\right\}.$$

Now, since $\|h-\tilde{h}\| = \mathbb{E}\left[\|h-\tilde{h}\|_n\right]$, if the inequality $\|h-\tilde{h}\|_n \leq \delta$ is satisfied, we also have $\|h-\tilde{h}\| \leq \delta$. Thus,

$$\forall \delta > 0, \qquad \|h-\tilde{h}\|_n \leq \delta \Rightarrow |\Delta_n(g-h) - \Delta_n(g-\tilde{h})| \leq 8C\delta$$

and the mapping $h \mapsto \Delta_n(g-h)$ is Lipschitz for the empirical norm $\|\cdot\|_n$.

Next, let us consider a δ-covering $\mathcal{E}^* = \{e_j^*, j = 1, \cdots, M\}$ of $\mathcal{C}^+_{[0,1]}$ for the empirical norm $\|.\|_n$. We stress that this set $\mathcal{E}^*$ is random since it depends on the points X_i, but its cardinality M may be chosen deterministic and upper-bounded as follows (see for example the proof of Lemma 2.2 in van de Geer [31]): denoting $N = \left\lceil \frac{2C}{\delta} \right\rceil$, where $\lceil\rceil$ stands for the ceiling function, we have

$$M = \binom{n+N}{N} \leq n^N, \tag{20}$$

where the last inequality is satisfied for any integer $n \geq 2$ as soon as $N \geq 3$.

Then, for any h in $\mathcal{C}^+_{[0,1]}$, there exists e^* in $\mathcal{E}^*$ such that $\|h-e^*\|_n \leq \delta$. From the previous Lipschitz property, we know that

$$|\Delta_n(g-h) - \Delta_n(g-e^*)| \leq 8C\delta.$$

Letting $t > 0$ and $\delta = t/(16C)$, our objective is to upper bound

$$\mathbb{P}\left(\sup_{h \in \mathcal{C}^+_{[0,1]}} |\Delta_n(g-h)| > t \right).$$

In this aim, for any h in $\mathcal{C}^+_{[0,1]}$ and any e^* in $\mathcal{E}^*$, we start with the decomposition

$$|\Delta_n(g-h)| \leq |\Delta_n(g-h) - \Delta_n(g-e^*)| + |\Delta_n(g-e^*)|.$$

For any h such that $|\Delta_n(g-h)| > t$, since there exists e^* in $\mathcal{E}^*$ such that

$$|\Delta_n(g-h) - \Delta_n(g-e^*)| \leq t/2,$$

we necessarily have $|\Delta_n(g-e^*)| > t/2$, and consequently

$$\mathbb{P}\left(|\Delta_n(g-h)| > t\right) \leq \mathbb{P}\left(\max_{j=1\cdots M} |\Delta_n(g-e_j^*)| > t/2 \right).$$

In other words,

$$\begin{aligned}
\mathbb{P}\left(\sup_{h \in \mathcal{C}^+_{[0,1]}} |\Delta_n(g-h)| > t \right) &\leq \mathbb{P}\left(\max_{j=1\cdots M} |\Delta_n(g-e_j^*)| > t/2 \right) \\
&\leq \mathbb{P}\left(\bigcup_{j=1}^{M} |\Delta_n(g-e_j^*)| > t/2 \right) \\
&\leq \sum_{j=1}^{M} \mathbb{P}\left(|\Delta_n(g-e_j^*)| > t/2 \right).
\end{aligned}$$

According to (19) and to the fact that

$$M \leq n^N = n^{\lceil \frac{2C}{\delta} \rceil},$$

fixing $\delta = t/(16C)$ leads to

$$\mathbb{P}\left(\sup_{h \in \mathcal{C}^+_{[0,1]}} |\Delta_n(g-h)| > t \right) \leq 2M \exp\left(-\frac{t^2 n}{8C^2} \right) \leq 2 \exp\left(\left\lceil \frac{32C^2}{t} \right\rceil \log n - \frac{t^2 n}{32C^2} \right).$$

Finally, for any $\alpha \in (0, 1/3)$, there exists $c_2 = c_2(\alpha)$ such that for any integer n,

$$\left\lceil \frac{32C^2}{n^{-\alpha}} \right\rceil \log n - \frac{n^{-2\alpha} n}{32C^2} \leq -c_2 n^{1-2\alpha},$$

hence the desired result. □

The last concentration inequality is a generalization of the previous one: this time, neither g nor h are assumed fixed.

Lemma 2 *Denoting $\mathcal{C}^+_{[0,1]}$ the set of non decreasing mappings from $[0,1]$ to $[-C,C]$, we have for all $t>0$*

$$\mathbb{P}\left(\sup_{h_1\in\mathcal{C}^+_{[0,1]},h_2\in\mathcal{C}^+_{[0,1]}}|\Delta_n(h_1-h_2)|>t\right)\leq\exp\left(2\left\lceil\frac{64C^2}{t}\right\rceil\log n-\frac{t^2n}{32C^2}\right).$$

Proof. With the same notations as before, just note that for any mapping $h_1\in\mathcal{C}^+_{[0,1]}$ (respectively h_2), there exists h_1^* (respectively h_2^*) in the δ-covering $\mathcal{E}^*$ of $\mathcal{C}^+_{[0,1]}$, such that

$$\|h_1-h_1^*\|_n\leq\delta\qquad\text{and}\qquad\|h_2-h_2^*\|_n\leq\delta.$$

Following the same line as in the proof of Lemma 1, we have that, for any mapping g with values in $[-C,C]$,

$$|\Delta_n(g-h_1)-\Delta_n(g-h_1^*)|\leq 8C\delta\qquad\text{and}\qquad|\Delta_n(g-h_2)-\Delta_n(g-h_2^*)|\leq 8C\delta.$$

In particular

$$|\Delta_n(h_2-h_1)-\Delta_n(h_2-h_1^*)|\leq 8C\delta\qquad\text{and}\qquad|\Delta_n(h_1^*-h_2)-\Delta_n(h_1^*-h_2^*)|\leq 8C\delta.$$

Moreover,

$$|\Delta_n(h_1-h_2)|\leq|\Delta_n(h_2-h_1)-\Delta_n(h_2-h_1^*)|+|\Delta_n(h_2-h_1^*)|.$$

Set $\delta=t/(32C)$, then

$$|\Delta_n(h_1-h_2)|>t\Rightarrow|\Delta_n(h_2-h_1^*)|>3t/4.$$

In the same manner,

$$|\Delta_n(h_2-h_1^*)|\leq|\Delta_n(h_1^*-h_2)-\Delta_n(h_1^*-h_2^*)|+|\Delta_n(h_1^*-h_2^*)|,$$

and

$$|\Delta_n(h_2-h_1^*)|>3t/4\Rightarrow|\Delta_n(h_1^*-h_2^*)|>t/2.$$

Hence, for any h_1 and h_2 in $\mathcal{C}^+_{[0,1]}$,

$$\mathbb{P}\left(|\Delta_n(h_1-h_2)|>t\right)\leq\mathbb{P}\left(\max_{h_1^*,h_2^*\in\mathcal{E}^*}|\Delta_n(h_1^*-h_2^*)|>t/2\right).$$

As a consequence, the choice $\delta=t/(32C)$ gives

$$\begin{aligned}\mathbb{P}\left(\sup_{h_1\in\mathcal{C}^+_{[0,1]},h_2\in\mathcal{C}^+_{[0,1]}}|\Delta_n(h_1-h_2)|>t\right)&\leq\mathbb{P}\left(\max_{h_1^*,h_2^*\in\mathcal{E}^*}|\Delta_n(h_j^*-h_{j'}^*)|>t/2\right)\\&\leq\sum_{1\leq j_1\neq j_2\leq M}\mathbb{P}\left(|\Delta_n(e_j^*-e_{j'}^*)|>t/2\right)\\&\leq M^2\exp\left(-\frac{t^2n}{32C^2}\right).\end{aligned}$$

According to (20), we are led to

$$\mathbb{P}\left(\sup_{h_1\in\mathcal{C}^+_{[0,1]},h_2\in\mathcal{C}^+_{[0,1]}}|\Delta_n(h_1-h_2)|>t\right)\leq\exp\left(2\left\lceil\frac{64C^2}{t}\right\rceil\log n-\frac{t^2n}{32C^2}\right).$$

□

Note that for any $\alpha\in(0,1/3)$, there exists $c_2=c_2(\alpha)>0$ such that for any integer n

$$2\left\lceil\frac{64C^2}{n^{-\alpha}}\right\rceil\log n-\frac{n^{-2\alpha}n}{32C^2}\leq-c_2n^{1-2\alpha}.$$

Thus, for any $\alpha\in(0,1/3)$, the following concentration inequality holds

$$\mathbb{P}\left(\sup_{h_1\in\mathcal{C}^+_{[0,1]},h_2\in\mathcal{C}^+_{[0,1]}}|\Delta_n(h_1-h_2)|>n^{-\alpha}\right)\leq\exp\left(-c_2n^{1-2\alpha}\right).$$

6.2 An approximation result

Consider the subset $\mathcal{C}^+_{n,C}$ of $\mathcal{C}^+_n$ consisting in all vectors whose absolute values of the components are bounded by a real number C. Consider $N\in\mathbb{N}$ such that $N\leq n$. For each $j=0,\ldots,N-1$, let us introduce the vector $h^+_j=\left(h^+_j[1],\cdots,h^+_j[n]\right)'$ of $\mathbb{R}^n$ as follows

$$h^+_j[i]=\begin{cases}0 & \text{if } i\leq\lfloor\frac{jn}{N}\rfloor\\ 1 & \text{otherwise}\end{cases}$$

and define

$$H_+=\mathrm{Vect}(h^+_0,\cdots,h^+_{N-1}).$$

Finally, set $\delta=2\sqrt{2}C/\sqrt{N}\geq2\sqrt{2}C/\sqrt{n}$.

Lemma 3 *With the previous notations, we have for all f in $\mathcal{C}^+_{n,C}$,*

$$\inf_{h\in H_+}\|f-h\|_n\leq\delta.$$

Proof. We denote $f=(f[1],\ldots,f[n])'$, with

$$-C\leq f[1]\leq\cdots\leq f[n]\leq C.$$

Set $\alpha_N=f[n]$ and, for $j=0,\ldots,N-1$,

$$\alpha_j=\min_{i:h^+_j[i]=1}f[i].$$

We define also the vectors f_- and f_+ of H_+ as follows

$$f_-=\alpha_0h^+_0+\sum_{j=1}^{N-1}(\alpha_j-\alpha_{j-1})h^+_j,$$

and

$$f_+ = \alpha_1 h_0^+ + \sum_{j=1}^{N-1} (\alpha_{j+1} - \alpha_j) h_j^+.$$

Observe that $f_- \leq f \leq f_+$, therefore

$$\|f - f_-\|_n^2 \leq \|f_+ - f_-\|_n^2$$

with

$$f_+ - f_- = \sum_{j=1}^{N-1} (\alpha_j - \alpha_{j-1})(h_{j-1}^+ - h_j^+) + (\alpha_N - \alpha_{N-1}) h_{N-1}^+. \tag{21}$$

Remark that, for all $j = 1, \ldots, N-1$,

$$\|h_{j-1}^+ - h_j^+\|_n^2 \leq \frac{1}{n} \left(\lfloor \frac{jn}{N} \rfloor - \lfloor \frac{(j-1)n}{N} \rfloor \right) \leq \frac{1}{n} \left(\frac{n}{N} + 1 \right) \leq \frac{2}{N},$$

and $\|h_{N-1}^+\|_n^2 \leq 2/N$ as well. Thus, taking into account that the decomposition (21) is orthogonal, we get

$$\|f_+ - f_-\|_n^2 \leq \frac{2}{N} \sum_{j=1}^{N} (\alpha_j - \alpha_{j-1})^2 = \frac{8C^2}{N} \sum_{j=1}^{N} \left(\frac{\alpha_j - \alpha_{j-1}}{2C} \right)^2.$$

Since $0 \leq (\alpha_j - \alpha_{j-1})/(2C) \leq 1$ and $0 \leq (\alpha_N - \alpha_1)/2C \leq 1$, we are led to

$$\|f_+ - f_-\|_n^2 \leq \frac{8C^2}{N} \sum_{j=1}^{N} \frac{\alpha_j - \alpha_{j-1}}{2C} \leq \frac{8C^2}{N}.$$

Considering that $\delta^2 = 8C^2/N$, we finally get the desired result, that is

$$\inf_{h \in H_+} \|f - h\|_n^2 \leq \delta^2.$$

□

For the subset $\mathcal{C}_{n,C}^-$ of $\mathcal{C}_n^-$, we proceed in the same way. We conclude that there exists a vector space H_- with dimension $N = 8C^2/\delta^2$ such that, for all f in $\mathcal{C}_{n,C}^-$,

$$\inf_{h \in H^-} \|f - h\|_n \leq \delta.$$

Acknowledgments. We wish to thank Dragi Anevski and Enno Mammen to have made us aware of reference [1]. Arnaud Guyader is greatly indebted to Bernard Delyon for fruitful discussions on von Neumann's algorithm.

References

[1] D. Anevski and P. Soulier (2011). Monotone spectral density estimation. *The Annals of Statistics*, **39**(1), 418-438.

[2] M. Ayer, H.D. Brunk, G.M. Ewing, W.T. Reid, and E. Silverman (1955). An empirical distribution function for sampling with incomplete information. *The Annals of Mathematical Statistics*, 641-647.

[3] R.E. Barlow, D.J. Bartholomew, J.M. Bremner, and H.D. Brunk (1972). *Statistical inference under order restrictions: Theory and application of isotonic regression.* John Wiley & Sons.

[4] H. H. Bauschke and J. M. Borwein (1993). On the Convergence of von Neumann's Alternating Projection Algorithm for Two Sets. *Set-Valued Analysis*, **1**(2), 185-212.

[5] H.H. Bauschke and J.M. Borwein (1994). Dykstra's alternating projection algorithm for two sets. *Journal of Approximation Theory*, **79**(3), 418-443.

[6] M.J. Best and N. Chakravarti (1990). Active set algorithms for isotonic regression; An unifying framework. *Mathematical Programming*, **47**(1), 425-439.

[7] H.D. Brunk (1955). Maximum likelihood estimates of monotone parameters. *The Annals of Mathematical Statistics*, 607-616.

[8] H.D. Brunk (1970). Estimation of isotonic regression. *Cambridge University Press*, 177-195.

[9] A. Buja, T.J. Hastie, and R.J. Tibshirani (1989). Linear smoothers and additive models. *The Annals of Statistics*, **17**(2), 453-510.

[10] V.V. Buldygin and Yu.V. Kozachenko (1972). *Metric Characterization of Random Variables and Random Processes.* American Mathematical Society.

[11] F. Deutsch (1991). The method of alternating orthogonal projections. *Approximation Theory , Spline Functions and Applications* (S.P. Singh, Ed), 105-121.

[12] C. Durot (2007). On the Lp-error of monotonicity constrained estimators. *The Annals of Statistics*, **35**(3), 1080-1104.

[13] R.L. Dykstra (1981). An isotonic regression algorithm. *Journal of Statistical Planning and Inference*, **5**(4), 355-363.

[14] J.H. Friedman and W. Stuetzle (1981). Projection pursuit regression. *Journal of the American Statistical Association*, 817-823.

[15] A. Guyader, N. Jégou, A.B. Németh, and S.N. Németh (2014). A Geometrical Approach to Iterative Isotone Regression. *Applied Mathematics and Computation*, **227**, 359-369.

[16] L. Györfi, M. Kohler, A. Kryzak, and H. Walk (1990). *A distribution-free theory of nonparametric regression*. Springer-Verlag, New York

[17] D.L. Hanson, G. Pledger, and F.T. Wright (1973). On consistency in monotonic regression. *The Annals of Statistics*, **1**(3), 401-421.

[18] T.J. Hastie and R.J. Tibshirani (1990). *Generalized additive models*. Chapman & Hall/CRC.

[19] W. Härdle and P. Hall (1993). On the backfitting algorithm for additive regression models. *Statistica Neerlandica*, **47**(1), 43-57.

[20] J. Horowitz, J. Klemelä, and E. Mammen (2006). Optimal estimation in additive regression models. *Bernoulli*, **12**(2), 271-298.

[21] N.W. Hengartner and S. Sperlich (1999). Rate optimal estimation with the integration method in the presence of many covariates. *Journal of Multivariate Analysis*, **95**(2), 246-272.

[22] W. Kim, O.B. Linton, and N.W. Hengartner (1999). A computationally efficient oracle estimator for additive nonparametric regression with bootstrap confidence intervals. *Journal of Computational and Graphical Statistics*, **8**(2), 278-297.

[23] C.I.C. Lee (1983). The min-max algorithm and isotonic regression. *The Annals of Statistics*, **11**(2), 467-477.

[24] E. Mammen, O. Linton, and J. Nielsen (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *The Annals of Statistics*, **27**(5), 1443-1490.

[25] E. Mammen and K. Yu (2007). Additive isotone regression. *Asymptotics: Particles, Processes and Inverse Problems, IMS Lecture Notes-Monograph Series*, **55**, 179-195.

[26] M. Meyer and M. Woodroofe (2000). On the Degrees of Freedom in Shape-Restricted Regression. *The Annals of Statistics*, **28**(4), 1083-1104.

[27] J.D. Opsomer and D. Ruppert (1997). Fitting a bivariate additive model by local polynomial regression. *The Annals of Statistics*, **25**(1), 186-211.

[28] J.D. Opsomer (2000). Asymptotic properties of backfitting estimators. *Journal of Multivariate Analysis*, **73**(2), 166-179.

[29] T. Robertson, F.T. Wright, and R.L. Dykstra (1988). *Order Restricted Statistical Inference*. Wiley, New York.

[30] S. van de Geer and M. Wegkamp (1996). Consistency for the least squares estimator in nonparametric regression. *The Annals of Statistics*, **24**(6), 2513-2523.

[31] S. van de Geer (2000). *Empirical Process in M-Estimation*. Cambridge University Press.