Gradient COBRA: A kernel-based consensual aggregation for regression

Sothea Has

Gradient COBRA: A Kernel-based Consensual Aggregation for Regression

Sothea Has

LPSM, Sorbonne Université Pierre et Marie Curie (Paris 6)
75005 Paris, France
sothea.has@lpsm.paris

Abstract

In this article, we introduce a kernel-based consensual aggregation method for regression problems. We aim to flexibly combine individual regression estimators $r_1, ..., r_M$ using a weighted average where the weights are defined based on predicted features given by all the basic estimators and some kernel function. This work extends the context of Biau et al. (2016) to a more general kernel-based framework. We show that this more general configuration also inherits the consistency of the basic consistent estimators, and the same convergence rate as in the classical method is achieved. Moreover, an optimization method based on gradient descent algorithm is proposed to efficiently and rapidly estimate the key parameter of the strategy. Various numerical experiments carried out on several simulated and real datasets are also provided to illustrate the efficiency and accuracy of the proposed method. Moreover, a domain adaptation-like property of the aggregation strategy is also illustrated on a physics data provided by Commissariat à l'Énergie Atomique (CEA).

Keywords: Consensual aggregation, kernel, regression.

2010 Mathematics Subject Classification: 62G08, 62J99, 62P30

1 Introduction

Aggregation methods, given the high diversity of available estimation strategies, are now of great interest in constructing predictive models. To this goal, several aggregation methods consisting of building a linear or convex combination of a collection of initial estimators have been introduced, for instance, in Catoni (2004), Juditsky and Nemirovski (2000), Nemirovski (2000), Yang

(2000, 2001, 2004), Györfi et al. (2002), Wegkamp (2003), Audibert (2004), Bunea et al. (2006, 2007a,b), and Dalalyan and Tsybakov (2008). Other than aggregating, another possible approach is selecting the best estimator among the candidate estimators which is known as model selection technique (see, for example, Massart (2007)).

Apart from the usual linear combination and model selection methods, a different technique has been introduced in classification problems by Mojirsheibani (1999). In his paper, the combination is the *majority vote* among all the points for which their predicted classes, given by all the basic classifiers, *coincide* with the predicted classes of the query point. Roughly speaking, instead of predicting a new point based on the structure of the original input, we look at the topology defined by the predictions of the candidate estimators. Each estimator was constructed differently so it may be able to capture different features of the input data and be useful in defining "closeness". Consequently, two points having similar predictions or classes seem reasonably having similar actual response values or belonging to the same actual class.

Later, Mojirsheibani (2000) and Mojirsheibani and Kong (2016) introduced exponential and general kernel-based versions of the primal idea to improve the smoothness in selecting and weighting individual data points in the combination. In this context, the kernel function transforms the level of *disagreements* between the predicted classes of a training point x_i and the query point x into a contributed weight given to the corresponding point in the vote. Besides, Biau et al. (2016) configured the original idea of Mojirsheibani (1999) as a regression framework where a training point x_i is "close" to the query point x if each of their predictions given by all the basic regression estimators is "close". Each of the close neighbors of x will be given a uniformly 0-1 weight contributing to the combination. It was shown theoretically in these former papers that the combinations inherit the consistency property of consistent basic estimators.

Recently from a practical point of view, a kernel-based version of Biau et al. (2016) called `KernelCobra` has been implemented in `pycobra` python library (see Guedj and Srinivasa Desikan (2018)). This method has also been applied in filtering to improve the image denoising (see Guedj and Rengot (2020)). Moreover, consensual aggregation methods such as Biau et al. (2016), Fischer and Mougeot (2019) and the present method are also incorporated in a three-step methodology called `KFC procedure`, which combines unsupervised clustering and supervised prediction for (energy) data modeling (see Has et al. (2021)). Such an idea of consensual aggregation was also used

in unsupervised classification known as `Clustering Aggregation` (see, for example, Gionis et al. (2005) and Wu et al. (2012)). On top of that, the aggregation method can also be used to handle the parameter tuning problem when different types of estimators are considered. It has been shown in Has (2022) that the method also maintains its good performance on highly correlated high-dimensional features of predictions that are plainly constructed without model selection or cross-validation.

In a complementary manner to the earlier works, we present in this paper a kernel-based consensual regression aggregation method, as well as its theoretical and numerical performances. More precisely, we show that the consistency inheritance property shown in Biau et al. (2016) also holds for this kernel-based configuration for a broad class of regular kernels. Moreover, evidence of numerical simulation carried out on several simulated models, and some real datasets, shows that the present method outperforms the classical one in both accuracy and efficiency.

This paper is organized as follows. Section 2 introduces some notation, the definition of the proposed method, and presents the theoretical results, namely consistency and convergence rate of the variance-type term of the aggregation strategy. An optimization method based on gradient descent algorithm for estimating the bandwidth parameter is described in Section 3. Section 4 illustrates the performances of the proposed method through several numerical experiments computed on different simulated and real datasets. Next, the conclusion and perspective, followed by the reproducibility of this study are given in Section 5 and Section 6 respectively. Lastly, Section 7 collects all the proofs of the theoretical results given in Section 2.

2 The kernel-based combining regression

2.1 Notation

We consider a training sample $\mathcal{D}_n = \{(X_i, Y_i)_{i=1}^n\}$ where $(X_i, Y_i), i = 1, 2, ..., n$, are *iid* copies of the generic couple (X, Y). We assume that (X, Y) is an $\mathbb{R}^d \times \mathbb{R}$-valued random variable with a suitable integrability which will be specified later.

We randomly split the training data $\mathcal{D}_n$ into two parts of size ℓ and k such that $\ell + k = n$. These are denoted by $\mathcal{D}_\ell = \{(X_i^{(\ell)}, Y_i^{(\ell)})_{i=1}^\ell\}$ and $\mathcal{D}_k = \{(X_i^{(k)}, Y_i^{(k)})_{i=1}^k\}$ respectively (a common choice is $k = \lceil n/2 \rceil = n - \ell$). The

M basic regression estimators $r_{k,1}, r_{k,2}, ..., r_{k,M}$ are constructed using only the data points in $\mathcal{D}_k$. These basic estimators can be any regression estimators such as linear regression, kNN, kernel smoother, SVR, lasso, ridge, neural networks, naive Bayes, bagging, gradient boosting, random forests, etc. They could be parametric, nonparametric or semi-parametric with their possible tuning parameters. For the combination, we only need the predictions given by all these basic estimators of the remaining part $\mathcal{D}_\ell$ and the query point x.

In the sequel, for any $x \in \mathbb{R}^d$, the following notation is used:

- $\mathbf{r}_k(x) = (r_{k,1}(x), r_{k,2}(x), ..., r_{k,M}(x))$: the vector of predictions of x.
- $\|x\| = \|x\|_2 = \sqrt{\sum_{i=1}^d x_i^2}$: Euclidean norm on $\mathbb{R}^d$.
- $\|x\|_1 = \sum_{i=1}^d |x_i|$: ℓ_1 norm on $\mathbb{R}^d$.
- $g^*(x) = \mathbb{E}[Y|X = x]$: the regression function.
- $g^*(\mathbf{r}_k(x)) = \mathbb{E}[Y|\mathbf{r}_k(x)]$: the conditional expectation of the response variable given all the predictions. This can be proven to be the optimal estimator in regression over the set of predictions $\mathbf{r}_k(X)$.
- $\mathbb{1}_{\{p\}} = \begin{cases} 1, & \text{if } p \text{ is true} \\ 0, & \text{otherwise} \end{cases}$: the indicator function.

The consensual regression aggregation is the weighted average defined by

$$g_n(\mathbf{r}_k(x)) = \sum_{i=1}^{\ell} W_{n,i}(x) Y_i^{(\ell)}. \tag{1}$$

Recall that given all the basic estimators $r_{k,1}, r_{k,2}, ..., r_{k,M}$, the aggregation method proposed by Biau et al. (2016) corresponds to the following naive weights:

$$W_{n,i}(x) = \frac{\prod_{m=1}^M \mathbb{1}_{\{|r_{k,m}(X_i) - r_{k,m}(x)| < h\}}}{\sum_{j=1}^{\ell} \prod_{m=1}^M \mathbb{1}_{\{|r_{k,m}(X_j) - r_{k,m}(x)| < h\}}}, i = 1, 2, ..., \ell. \tag{2}$$

Moreover, the condition of "closeness for all" predictions, can be relaxed to "some" predictions, which corresponds to the following weights:

$$W_{n,i}(x) = \frac{\mathbb{1}_{\{\sum_{m=1}^M \mathbb{1}_{\{|r_{k,m}(X_i) - r_{k,m}(x)| < h\}} \geq \alpha M\}}}{\sum_{j=1}^{\ell} \mathbb{1}_{\{\sum_{m=1}^M \mathbb{1}_{\{|r_{k,m}(X_j) - r_{k,m}(x)| < h\}} \geq \alpha M\}}}, i = 1, 2, ..., \ell \tag{3}$$

where $\alpha \in \{1/M, 2/M, ..., 1\}$ is the proportion of consensual predictions required and $h > 0$ is the bandwidth or window parameter to be determined. Constructing the proposed method is equivalent to searching for the best possible value of these parameters over a given grid, minimizing some quadratic error which will be described in Section 3.

In the present paper, $K : \mathbb{R}^M \to \mathbb{R}_+$ denotes a regular kernel which is a decreasing function satisfying:

$$\exists b, \kappa_0, \rho > 0 \text{ such that } \begin{cases} b\mathbb{1}_{B_M(0,\rho)}(z) \leq K(z) \leq 1, \forall z \in \mathbb{R}^M \\ \int_{\mathbb{R}^M} \sup_{u \in B_M(z,\rho)} K(u)dz = \kappa_0 < +\infty \end{cases} \tag{4}$$

where $B_M(c, r) = \{z \in \mathbb{R}^M : \|c - z\| < r\}$ denotes the open ball of center $c \in \mathbb{R}^M$ and radius $r > 0$ of $\mathbb{R}^M$. We propose in equation (1) a method associated to the weights defined at any query point $x \in \mathbb{R}^d$ by

$$W_{n,i}(x) = \frac{K_h(\mathbf{r}_k(X_i^{(\ell)}) - \mathbf{r}_k(x))}{\sum_{j=1}^{\ell} K_h(\mathbf{r}_k(X_j^{(\ell)}) - \mathbf{r}_k(x))}, i = 1, 2, ..., \ell \tag{5}$$

where $K_h(z) = K(z/h)$ for some bandwidth parameter $h > 0$ with the convention of $0/0 = 0$. Observe that the combination in equation (1) is computed based only on $\mathcal{D}_\ell$ but the construction of the method depends on the whole training data $\mathcal{D}_n$ as the basic estimators are all constructed using $\mathcal{D}_k$. In our setting, we treat the vector of predictions $\mathbf{r}_k(x)$ as an M-dimensional feature, and the kernel function is applied on the whole vector at once. Note that the implementation of `KernelCobra` in Guedj and Srinivasa Desikan (2020) corresponds to the following weights:

$$W_{n,i}(x) = \frac{\sum_{m=1}^{M} K_h(r_{k,m}(X_i^{(\ell)}) - r_{k,m}(x))}{\sum_{j=1}^{\ell} \sum_{m=1}^{M} K_h(r_{k,m}(X_j^{(\ell)}) - r_{k,m}(x))}, i = 1, 2, ..., \ell \tag{6}$$

where the univariate kernel function K is applied on each component of the predicted vector $\mathbf{r}_k(.)$ separately. In this case, the weight $W_{n,i}(x)$ defined in equation (6) above is more costly in computing than the one in the proposed method since the univariate kernel function has to be applied on all the entries of vectors $\mathbf{r}_k(X_i^\ell) - \mathbf{r}_k(x) = (r_{k,1}(X_i^\ell) - r_{k,1}(x), ..., r_{k,M}(X_i^\ell) - r_{k,M}(x))$ for all $i = 1, ..., \ell$. This entry-wise operation prevents us from trading off memory storage for computational complexity. On the other hand, the weights in equation (5) of the proposed method depend on pair-wise distances between

the predicted vectors of the training points $X_i^{(\ell)}$'s and the query point x, $d'(\mathbf{r}_k(X_i), \mathbf{r}_k(x))$, for some distance d' (associated to the kernel function). This dependency allows us to trade the memory storage off for computational complexity, yielding more efficient computation and the implementation of an optimization procedure based on gradient descent algorithm (section 3).

2.2 Theoretical performance

The performance of the combining estimation g_n is measured using the quadratic risk defined by

$$\mathbb{E}\Big[|g_n(\mathbf{r}_k(X)) - g^*(X)|^2\Big]$$

where the expectation is taken with respect to both X and the training sample $\mathcal{D}_n$. Firstly, we begin with a simple decomposition of the distortion between the proposed method and the optimal regression estimator $g^*(X)$ by introducing the optimal regression estimator over the set of predictions $g^*(\mathbf{r}_k(X))$. The following proposition shows that the nonasymptotic-type control of the distortion, presented in Proposition.2.1 of Biau et al. (2016), also holds for this case of regular kernels.

Proposition 1 *Let $\boldsymbol{r}_k = (r_{k,1}, r_{k,2}, ..., r_{k,M})$ be the collection of all basic estimators, and let $g_n(\boldsymbol{r}_k(x))$ be the combined estimator defined in equation (1) with the weights given in equation (5) computed at point $x \in \mathbb{R}^d$. Then, for all distributions of (X, Y) with $\mathbb{E}[|Y|^2] < +\infty$,*

$$\begin{aligned}\mathbb{E}\Big[|g_n(\boldsymbol{r}_k(X)) - g^*(X)|^2\Big] \leq \inf_{f \in \mathcal{G}} \mathbb{E}\Big[|f(\boldsymbol{r}_k(X)) - g^*(X)|^2\Big] \\ + \mathbb{E}\Big[|g_n(\boldsymbol{r}_k(X)) - g^*(\boldsymbol{r}_k(X))|^2\Big]\end{aligned}$$

where $\mathcal{G}$ is the class of any function $f : \mathbb{R}^M \to \mathbb{R}$ satisfying $\mathbb{E}[f(\boldsymbol{r}_k(X))|^2] < +\infty$. In particular,

$$\begin{aligned}\mathbb{E}\Big[|g_n(\boldsymbol{r}_k(X)) - g^*(X)|^2\Big] \leq \min_{1 \leq m \leq M} \mathbb{E}\Big[|r_{k,m}(X) - g^*(X)|^2\Big] \\ + \mathbb{E}\Big[|g_n(\boldsymbol{r}_k(X)) - g^*(\boldsymbol{r}_k(X))|^2\Big].\end{aligned}$$

The two terms of the last bound can be viewed as a bias-variance decomposition where the first term $\min_{1 \leq m \leq M} \mathbb{E}[|r_{k,m}(X) - g^*(X)|^2]$ can be seen

as the bias and $\mathbb{E}[|g_n(\mathbf{r}_k(X)) - g^*(\mathbf{r}_k(X))|^2]$ is the variance-type term (Biau et al. (2016)). Given all the estimators, the first term cannot be controlled as it depends on the performance of the best constructed estimator, and it will be the asymptotic performance of the proposed method. Our main task is to deal with the second term, which can be proven to be asymptotically negligible in the following key proposition.

Proposition 2 *Assume that $r_{k,m}$ is bounded for all $m = 1, 2, .., M$. Let $h \to 0$ and $\ell \to +\infty$ such that $h^M \ell \to +\infty$. Then*

$$\mathbb{E}\Big[|g_n(\boldsymbol{r}_k(X)) - g^*(\boldsymbol{r}_k(X))|^2\Big] \to 0 \text{ as } \ell \to +\infty$$

for all distribution of (X, Y) with $\mathbb{E}[|Y|^2] < +\infty$. Thus,

$$\limsup_{\ell \to +\infty} \mathbb{E}\Big[|g_n(\boldsymbol{r}_k(X)) - g^*(X)|^2\Big] \leq \inf_{f \in \mathcal{G}} \mathbb{E}\Big[|f(\boldsymbol{r}_k(X)) - g^*(X)|^2\Big].$$

And in particular,

$$\limsup_{\ell \to +\infty} \mathbb{E}\Big[|g_n(\boldsymbol{r}_k(X)) - g^*(X)|^2\Big] \leq \min_{1 \leq m \leq M} \mathbb{E}\Big[|r_{k,m}(X) - g^*(X)|^2\Big].$$

Proposition 2 above is an analogous setup of Proposition 2.2 in Biau et al. (2016). To prove this result, we follow the procedure of Stone's theorem (see, for example, Stone (1977) and Chapter 4 of Györfi et al. (2002)) of weak universal consistency of non-parametric regression. However, showing this result for the class of regular kernels is not straightforward. Most of the previous studies provided such a result of L_2-consistency only for the class of compactly supported kernels (see, for example, Chapter 5 of Györfi et al. (2002)). In this study, we can derive the result for this broader class thanks to the boundedness of all basic estimators. However, the price to pay for the universality for this class of regular kernels is the lack of convergence rate. To this goal, a weak smoothness assumption of g^* with respect to the basic estimators is required. For example, the convergence rate of the variance-type term in Biau et al. (2016) is of order $O(\ell^{-2/(M+2)})$ under the same smoothness assumption, and this result also holds for all the compactly support kernels. In this study, we can derive the same convergence rate for the class of kernel functions with the tails increase at least of exponential speed. This main theoretical result is given in the following theorem.

Theorem 1 *Assume that the response variable Y and all the basic estimators $r_{k,m}, m = 1, 2, ..., M$, are bounded by some constant R. Suppose that there exists a constant $L \geq 0$ such that, for every $k \geq 1$,*

$$|g^*(\boldsymbol{r}_k(x)) - g^*(\boldsymbol{r}_k(y))| \leq L\|\boldsymbol{r}_k(x) - \boldsymbol{r}_k(y)\|, \forall x, y \in \mathbb{R}^d.$$

We assume moreover that there exists some positive constants α, R_K and C_K such that

$$K(z) \leq C_K \exp(-\|z\|^\alpha), \forall z \in \mathbb{R}^M, \|z\| \geq R_K. \tag{7}$$

Then, one has

$$\mathbb{E}[|g_n(\boldsymbol{r}_k(X)) - g^*(X)|^2] \leq \min_{1 \leq m \leq M} \mathbb{E}[|r_{k,m}(X) - g^*(X)|^2] + C\ell^{-\frac{2}{M+2}} \tag{8}$$

for some positive constant $C = C(b, L, R, R_K, C_K)$ independent of ℓ.

From this result, if there exists a consistent estimator named r_{k,m_0} in the list $\{r_{k,m}\}_{m=1}^M$ i.e.,

$$\mathbb{E}[|r_{k,m_0}(X) - g^*(X)|^2] \to 0 \text{ as } k \to +\infty,$$

then the combining estimator g_n is also consistent for all distribution of (X, Y) in some class $\mathcal{M}$. Consequently, under the assumption of Theorem 1, one has

$$\lim_{k,\ell \to +\infty} \mathbb{E}[|g_n(\mathbf{r}_k(X)) - g^*(X)|^2] = 0.$$

3 Bandwidth estimation using gradient descent

In earlier works by Biau et al. (2016) and Guedj and Srinivasa Desikan (2020), the training data $\mathcal{D}_n$ is practically broken down into three balanced parts: $\mathcal{D}_k$ for constructing all candidate estimators $\{\mathbf{r}_{k,m}\}_{m=1}^M$, $\mathcal{D}_{\ell_1}$ for building aggregation defined in equation (1), and $\mathcal{D}_{\ell_2}$ for tuning the key parameters of the methods. Within these frameworks, the bandwidth parameter h is estimated by minimizing the following loss,

$$\varphi_M(h) = \frac{1}{|\mathcal{D}_{\ell_2}|} \sum_{(X_j, Y_j) \in \mathcal{D}_{\ell_2}} [g_n(\mathbf{r}_k(X_j)) - Y_j]^2, \tag{9}$$

where $|\mathcal{D}_{\ell_2}|$ denotes the cardinality of $\mathcal{D}_{\ell_2}$, and $g_n(\mathbf{r}_k(X_j)) = \sum_{(X_i, Y_i) \in \mathcal{D}_{\ell_1}} W_{n,i}(X_j) Y_i$ is given in equation (1). Note that the subscript M of $\varphi_M(h)$ indicates the full

consensus between the M components of the predictions $\mathbf{r}_k(X_i)$ and $\mathbf{r}_k(X_j)$ for any X_i of $\mathcal{D}_{\ell_1}$ and X_j of $\mathcal{D}_{\ell_2}$. In this case, constructing an aggregation method g_n is equivalent to searching for an optimal parameter h^* over a given grid $\mathcal{H} = \{h_{\min}, ..., h_{\max}\}$ i.e.,

$$h^* = \underset{h \in \mathcal{H}}{\operatorname{argmin}}\, \varphi_M(h).$$

The parameter α of equation (3) can be tuned easily by considering $\varphi_{\alpha M}(h)$ where $\alpha \in \{1/M, 2/M, ..., 1\}$ referring to the proportion of consensuses required among the M components of the predictions. In this case, the optimal parameters α^* and h^* are chosen to be the minimizer of $\varphi_{\alpha M}(h)$ i.e.,

$$(\alpha^*, h^*) = \underset{(\alpha,h) \in \{1/M, 2/M, ..., 1\} \times \mathcal{H}}{\operatorname{argmin}} \varphi_{\alpha M}(h).$$

Note that in both papers, the grid search algorithm is used in searching for the optimal bandwidth parameter.

In this paper, the training data is broken down into only two parts, $\mathcal{D}_k$ and $\mathcal{D}_\ell$. Again, we construct the basic estimators using $\mathcal{D}_k$, and for any κ folds $F_1, ..., F_\kappa$ ($\kappa \geq 2$) of $\mathcal{D}_\ell$, we propose the following κ-fold cross-validation error which is a function of the bandwidth parameter $h > 0$ defined by

$$\varphi^\kappa(h) = \frac{1}{\kappa} \sum_{p=1}^{\kappa} \sum_{(X_j, Y_j) \in F_p} [g_n(\mathbf{r}_k(X_j)) - Y_j]^2 \tag{10}$$

where in this case, $g_n(\mathbf{r}_k(X_j)) = \sum_{(X_i,Y_i) \in \mathcal{D}_\ell \setminus F_p} W_{n,i}(X_j) Y_i$, is computed using the remaining $\kappa - 1$ folds of $\mathcal{D}_\ell$ leaving $F_p \subset \mathcal{D}_\ell$ as the corresponding validation fold[1]. We often observe the convex-like curves of the cross-validation quadratic error on many simulations; and from this observation, we propose using a gradient descent algorithm to estimate the optimal bandwidth parameter. The associated gradient descent algorithm used to estimate the optimal parameter h^* is implemented as follows:

[1]In this part, we simply write $(X_i, Y_i) \in \mathcal{D}_\ell$ without the superscript (ℓ).

Algorithm 1 *: Gradient descent for estimating h^*:*

1. *Initialization: h_0, a learning rate $\lambda > 0$, threshold $\delta > 0$ and the maximum number of iteration N.*
2. *For $k = 1, 2, ..., N$, **while** $\left|\frac{d}{dh}\varphi^{\kappa}(h_{k-1})\right| > \delta$ do:*

$$h_k \leftarrow h_{k-1} - \lambda \frac{d}{dh}\varphi^{\kappa}(h_{k-1})$$

3. *return h_k violating the **while** condition or h_N to be the estimation of h^*.*

From equation (10), for any $(X_j, Y_j) \in F_p$, one has

$$\frac{d}{dh}\varphi^{\kappa}(h) = \frac{1}{\kappa}\sum_{p=1}^{\kappa}\sum_{(X_j,Y_j)\in F_p} 2\frac{\partial}{\partial h} g_n(\mathbf{r}_k(X_j))(g_n(\mathbf{r}_k(X_j)) - Y_j)$$

where

$$g_n(\mathbf{r}_k(X_j)) = \frac{\sum_{(X_i,Y_i)\mathcal{D}_\ell\in\setminus F_p} Y_i K_h(\mathbf{r}_k(X_j) - \mathbf{r}_k(X_i))}{\sum_{(X_q,Y_q)\in\mathcal{D}_\ell\setminus F_p} K_h(\mathbf{r}_k(X_j) - \mathbf{r}_k(X_q))}.$$

This implies that

$$\frac{\partial}{\partial h} g_n(\mathbf{r}_k(X_j)) = \sum_{(X_i,Y_i),(X_q,Y_q)\in\mathcal{D}_\ell\setminus F_p} (Y_i - Y_q)\frac{\frac{\partial}{\partial h}K_h(\mathbf{r}_k(X_j) - \mathbf{r}_k(X_i))K_h(\mathbf{r}_k(X_j) - \mathbf{r}_k(X_q))}{\Big[\sum_{(X_i,Y_i)\mathcal{D}_\ell\in\setminus F_p} K_h(\mathbf{r}_k(X_j) - \mathbf{r}_k(X_i))\Big]^2}.$$

The differentiability of g_n depends entirely on the kernel function K. Therefore, for suitable kernels, the implementation of the algorithm is straightforward. For example, in the case of Gaussian kernel $K_h(x) = \exp(-h\|x\|^2/(2\sigma^2))$ for some $\sigma > 0$, one has

$$\frac{\partial}{\partial h} g_n(\mathbf{r}_k(X_j)) = \sum_{(X_i,Y_i),(X_q,Y_q)\in\mathcal{D}_\ell\setminus F_p} (Y_q - Y_i)\|\mathbf{r}_k(X_j) - \mathbf{r}_k(X_i)\|^2 \times \frac{\exp\Big(-h(\|\mathbf{r}_k(X_j) - \mathbf{r}_k(X_i)\|^2 + \|\mathbf{r}_k(X_j) - \mathbf{r}_k(X_q)\|^2)/(2\sigma^2)\Big)}{2\sigma^2\Big(\sum_{(X_q,Y_q)\notin F_p}\exp(-h\|\mathbf{r}_k(X_j) - \mathbf{r}_k(X_q)\|^2/(2\sigma^2))\Big)^2}.$$

This suggests that we only need to store the distance matrices $D_p = (d'_{qj})$ where $d'_{qj} = \|\mathbf{r}_k(X_q) - \mathbf{r}_k(X_j)\|^2$ is the squared Euclidean distance between predictions of the input data from the $\kappa - 1$ folds $\mathcal{D}_\ell \setminus F_p$ and the corresponding validation fold F_p for $p = 1, ..., \kappa$. Then, the gradient can be computed straight away for any smoothing parameter $h > 0$.

To prevent the algorithm from reaching negative values of the bandwidth parameter during operation, a few adjustments have been implemented. Firstly, the predicted features are normalized for example, to be in the range $[0, 1]^M$. Then, the error is computed at a few randomly selected bandwidth parameters, and the algorithm begins at the parameter with the lowest error. Additionally, the learning rate λ is decreased when the parameter takes negative values, which may occur due to a large learning rate. To handle cases where the error curve is very flat around the optimal bandwidth, an option has been included to adjust the speed of the learning rate. This approach has resulted in faster algorithm performance, without requiring knowledge of the interval containing the optimal parameter, as with grid search. Moreover, it is possible to estimate the parameter that causes the gradient of the objective function to vanish. This leads to a well-constructed aggregation method, as reported in the next section.

4 Numerical examples

This section is devoted to numerical experiments to illustrate the performance of our proposed method. It is shown in Biau et al. (2016) that the classical method mostly outperforms the basic estimators of the combination. In this experiment, we compare the performances of the proposed methods with the classical one and all the basic regressors. Several options of kernel functions are considered. Most kernels are compactly supported on $[-1, 1]$, taking nonzero values only on $[-1, 1]$, except for the case of compactly supported Gaussian which is supported on $[-\rho_1, \rho_1]$, for some $\rho_1 > 0$. Moreover to implement the gradient descent algorithm in estimating the bandwidth parameter, we also present the results of non-compactly supported cases such as classical Gaussian and 4-exponential kernels. All kernels considered in this paper are listed in Table 1, and some of them are displayed (univariate case) in Figure 1 below.

4.1 Simulated datasets

In this subsection, we study the performances of our proposed method on the same set of simulated datasets of size n as provided in Biau et al. (2016). The input data

[2]The naive kernel corresponds to the method by Biau et al. (2016).

Kernel	Formula
Naive[2]	$K(x) = \prod_{i=1}^{d} \mathbb{1}_{\{\lvert x_i \rvert \leq 1\}}$
Epanechnikov	$K(x) = (1 - \lVert x \rVert^2)\mathbb{1}_{\{\lVert x \rVert \leq 1\}}$
Bi-weight	$K(x) = (1 - \lVert x \rVert^2)^2\mathbb{1}_{\{\lVert x \rVert \leq 1\}}$
Tri-weight	$K(x) = (1 - \lVert x \rVert^2)^3\mathbb{1}_{\{\lVert x \rVert \leq 1\}}$
Compact-support Gaussian	$K(x) = \exp\{-\lVert x \rVert^2/(2\sigma^2)\}\mathbb{1}_{\{\lVert x \rVert \leq \rho_1\}}, \sigma, \rho_1 > 0$
Gaussian	$K(x) = \exp\{-\lVert x \rVert^2/(2\sigma^2)\}, \sigma > 0$
4-exponential	$K(x) = \exp\{-\lVert x \rVert^4/(2\sigma^4)\}, \sigma > 0$

Table 1: Kernel functions used.

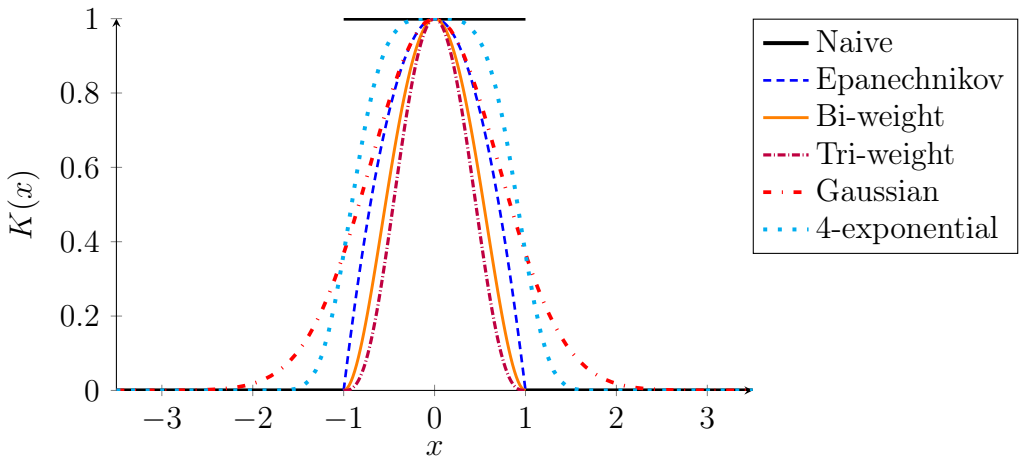

Figure 1: The shapes of some kernels.

is either independent and uniformly distributed over $(-1, 1)^d$ (*uncorrelated* case) or distributed from a Gaussian distribution $\mathcal{N}(0, \Sigma)$ where the covariance matrix Σ is defined by $\Sigma_{ij} = 2^{-|i-j|}$ for $1 \leq i, j \leq d$ (*correlated* case). We consider the following models:

Model 1 : $n = 800, d = 50, Y = X_1^2 + \exp(-X_2^2)$.

Model 2 : $n = 600, d = 100, Y = X_1X_2 + X_3^2 - X_4X_7 + X_8X_{10} - X_6^2 + \mathcal{N}(0, 0.5)$.

Model 3 : $n = 600, d = 100, Y = -\sin(2X_1) + X_2^2 + X_3 - \exp(-X_4) + \mathcal{N}(0, 0.5)$.

Model 4 : $n = 600, d = 100, Y = X_1 + (2X_2 - 1)^2 + \sin(2\pi X_3)/(2 - \sin(2\pi X_3)) + \sin(2\pi X_4) + 2\cos(2\pi X_4) + 3\sin^2(2\pi X_4) + 4\cos^2(2\pi X_4) + \mathcal{N}(0, 0.5)$.

Model 5 : $n = 700, d = 20, Y = \mathbb{1}_{\{X_1>0\}} + X_2^3 + \mathbb{1}_{\{X_4+X_6-X_8-X_9>1+X_{14}\}} + \exp(-X_2^2) + \mathcal{N}(0, 0.05)$.

These first five models are taken from Biau et al. (2016) which allows us to compare the performance of the methods. Note that by the design, there are not many active predictors contributing to the target, and most of them act as the noise. To see how the proposed method behaves on different type of datasets where more active independent variables are presented, we introduce the following models:

Model 6 : $n = 500, d = 20, Y = (\sum_{j=1}^{5}\sum_{k=0}^{3} X_{j+5k})\cos((\prod_{k=1}^{5} X_{4k})\pi/2) + \mathcal{N}(0, 0.25)$

Model 7 : $n = 600, d = 30, Y = \sum_{j=1}^{15} e^{0.25-X_j^2}\sin(\pi X_{j+15}) + \mathcal{N}(0, 0.25)$

Model 8 : $n = 700, d = 50, Y = (\sum j = 1^2 5 X_{2j}\sin(\pi/X_{2j-1}))e^{\sum k=1^5 X_{10k}^2/10} + \mathcal{N}(0, 0.75)$

Moreover, it is interesting to consider some high-dimensional cases as many real problems such as image and signal processing involve these kinds of datasets. Therefore, we also consider the following two high-dimensional models where all the independent variables contribute to the target via the coefficient β_j's.

Model 9 : $n = 600, d = 1500, Y = \pi + \sum_{j=1}^{d} \beta_j \frac{X_j \log|5+X_j|}{1+e^{X_j}} + \mathcal{N}(0, 1)$, *where* $\beta_j = 2^{-(d+1-j)/50} + 3^{-j/50}, j = 1, ..., d$.

Model 10 : $n = 700, d = 1500, Y = e + \sum_{j=1}^{d} \beta_j \frac{X_j e^{-X_j}}{1-\log|10-X_j|} + \mathcal{N}(0, 1.25)$, *where* $\beta_j = \frac{e^{-j/30}}{1-e^{-(d+1-j)/30}}, j = 1, ..., d$.

For each model, the proposed method is implemented over 100 replications. We randomly split 80% of each simulated dataset into two equal parts, $\mathcal{D}_\ell$ and $\mathcal{D}_k$ where $\ell = \lceil 0.8 \times n/2 \rceil - k$, and the remaining 20% is treated as the corresponding testing data. We measure the performance of any regression method f using *root mean square error* (RMSE) evaluated on the 20%-testing data defined by

$$\text{RMSE}(f) = \left(\frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} (y_i^{\text{test}} - f(x_i^{\text{test}}))^2 \right)^{1/2}. \tag{11}$$

Table 2 and 3 below contain the average RMSEs and the corresponding standard errors (into brackets) over 100 runs of *uncorrelated* and *correlated* cases respectively. In each table, the first block contains five columns corresponding to the following five basic regressors $\mathbf{r}_k = (r_{k,m})_{m=1}^5$:

- **Rid**: Ridge regression (R package `glmnet`, see Friedman et al. (2010)).
- **Las**: Lasso regression (R package `glmnet`).
- **kNN**: k-nearest neighbors regression (R package `FNN`, see Li (2019)).
- **Tr**: Regression tree (R package `tree`, see Ripley (2019)).
- **RF**: Random Forest regression (R package `randomForest`, see Liaw and Wiener (2002)).

We choose $k = 5$ for k-NN and $ntree = 300$ for random forest algorithm, and other methods are implemented using the default parameters. The best performance of each method in this block is given in **boldface**. The second block contains the last eight columns corresponding to kernel functions and different types of aggregation methods. The abbreviations of all the methods in this block are given below:

- **COBRA**: the *classical COBRA* by Biau et al. (2016).
- **Epan**: the aggregation method using *Epanechnikov* kernel.
- **Bi-wgt**: the aggregation method using *Bi-weight* kernel.
- **Tri-wgt**: the aggregation method using *Tri-weight* kernel.
- **C-Gaus**: the aggregation method using *Compact Gaussian* kernel.
- **Gauss**: the aggregation method using *Gaussian* kernel.
- **Exp4**: the aggregation method using 4-*Exponential* kernel.
- **KCOBRA**: the *KernelCobra* by Guedj and Srinivasa Desikan (2018).

The optimal RMSEs of each model in this block is also written in **boldface**. For all the compactly supported kernels, we consider 500 values of bandwidth parameter h in a uniform grid $\{10^{-100}, ..., h_{\max}\}$ where $h_{\max} = 10$, which is chosen to be large enough, likely to contain the optimal parameter to be searched. For the compactly supported Gaussian kernel, we set $\rho_1 = 3$ and $\sigma = 1$ therefore its support is $[-3, 3]$. For the two non-compactly supported kernels, Gaussian and 4-exponential, the optimal parameters are estimated using gradient descent algorithm described in the previous section. Lastly, Gaussian kernel is used for `KernelCobra`, and the optimal bandwidth is estimated using `optimal_kernelbandwidth` method of `pycobra` library.

In each table, we are interested in comparing the smallest average RMSEs in the first block to all the columns in the second block. First of all, we can

Table 2: Average MSEs in the uncorrelated case.

Model	Las	Rid	kNN	Tr	RF	COBRA	Epan	Bi-wgt	Tri-wgt	C-Gaus	Gauss	Exp4	KCOBRA
1	0.156	0.133	0.143	**0.027**	0.032	0.020	0.018	0.017	0.017	0.017	**0.015**	0.016	0.061
	(0.016)	(0.013)	(0.014)	(0.004)	(0.004)	(0.004)	(0.003)	(0.003)	(0.003)	(0.003)	(0.002)	(0.003)	(0.027)
2	1.301	0.784	0.873	1.124	**0.707**	0.722	0.718	0.712	0.715	0.712	**0.709**	0.710	0.788
	(0.216)	(0.110)	(0.123)	(0.165)	(0.097)	(0.065)	(0.079)	(0.080)	(0.079)	(0.079)	(0.078)	(0.079)	(0.085)
3	0.664	0.669	1.477	0.797	**0.629**	0.554	0.482	0.478	0.476	0.479	**0.475**	0.483	0.558
	(0.107)	(0.255)	(0.192)	(0.135)	(0.091)	(0.069)	(0.062)	(0.060)	(0.060)	(0.063)	(0.060)	(0.060)	(0.056)
4	7.783	6.550	10.238	3.796	**3.774**	3.608	3.231	3.185	3.153	3.189	2.996	3.186	**2.883**
	(1.121)	(1.115)	(1.398)	(0.840)	(0.523)	(0.526)	(0.383)	(0.382)	(0.384)	(0.371)	(0.384)	(0.464)	(0.212)
5	0.508	0.518	0.699	0.575	**0.436**	0.429	0.389	0.387	0.386	0.387	**0.383**	0.387	0.486
	(0.051)	(0.073)	(0.084)	(0.081)	(0.051)	(0.035)	(0.031)	(0.030)	(0.030)	(0.030)	(0.030)	(0.028)	(0.077)
6	**1.015**	1.020	1.405	1.774	1.290	1.004	0.934	0.943	0.941	0.947	**0.914**	0.936	0.957
	(0.054)	(0.053)	(0.098)	(0.145)	(0.083)	(0.085)	(0.050)	(0.062)	(0.060)	(0.053)	(0.049)	(0.049)	(0.076)
7	**1.887**	1.893	2.408	2.870	2.152	1.939	1.858	1.854	1.851	1.867	**1.828**	1.852	1.998
	(0.105)	(0.105)	(0.125)	(0.201)	(0.116)	(0.109)	(0.097)	(0.097)	(0.097)	(0.098)	(0.094)	(0.096)	(0.160)
8	1.475	**1.461**	1.578	1.919	1.464	1.426	1.416	**1.416**	**1.415**	**1.419**	**1.415**	1.416	1.456
	(0.079)	(0.078)	(0.089)	(0.121)	(0.074)	(0.085)	(0.080)	(0.080)	(0.080)	(0.081)	(0.079)	(0.080)	(0.099)
9	**3.343**	3.581	3.885	4.656	3.436	3.332	3.279	3.279	3.273	3.293	**3.240**	3.277	3.592
	(0.187)	(0.499)	(0.199)	(0.292)	(0.186)	(0.172)	(0.164)	(0.170)	(0.169)	(0.175)	(0.167)	(0.168)	(0.176)
10	**2.328**	2.489	2.797	3.381	2.541	2.308	2.214	2.216	2.212	2.232	**2.171**	2.210	2.342
	(0.135)	(0.158)	(0.154)	(0.243)	(0.141)	(0.163)	(0.143)	(0.154)	(0.153)	(0.163)	(0.153)	(0.153)	(0.158)

Table 3: Average MSEs in the correlated case.

Model	Las	Rid	kNN	Tr	RF	COBRA	Epan	Bi-wgt	Tri-wgt	C-Gaus	Gauss	Exp4	KCOBRA
1	2.294	1.947	1.941	**0.320**	0.542	0.307	0.304	0.301	0.288	0.297	**0.269**	0.291	0.449
	(0.544	(0.507)	(0.487)	(0.145)	(0.231)	(0.129)	(0.105)	(0.111)	(0.103)	(0.104)	(0.092)	(0.098)	(2.50)
2	14.273	8.442	8.572	6.796	**5.135**	5.345	4.582	4.529	4.491	4.541	**4.377**	4.910	4.946
	(2.593)	(1.912)	(1.751)	(1.548)	(1.372)	(1.194)	(0.941)	(0.934)	(0.922)	(0.896)	(0.905)	(1.181)	(1.271)
3	7.996	6.266	8.704	4.110	**3.722**	3.327	2.598	2.536	2.444	2.554	2.168	2.357	**1.853**
	(3.393)	(3.296)	(3.523)	(2.894)	(2.956)	(1.006)	(0.912)	(0.944)	(0.840)	(0.907)	(0.680)	(0.756)	(0.443)
4	61.474	42.351	46.934	**8.855**	13.381	9.599	10.511	9.963	9.682	10.085	9.056	9.713	**8.957**
	(13.986)	(11.622)	(12.543)	(3.480)	(5.549)	(4.125)	(2.961)	(3.101)	(2.860)	(2.904)	(2.407)	(2.695)	(0.954)
5	6.805	7.479	10.342	**4.000**	4.880	3.225	2.640	2.401	2.235	2.412	**1.792**	2.194	2.873
	(3.685)	(5.336)	(5.425)	(3.144)	(3.787)	(2.088)	(1.455)	(1.387)	(1.250)	(1.355)	(0.913)	(1.242)	(0.750)
6	24.078	23.883	22.216	24.612	**20.202**	19.573	18.475	18.901	16.718	17.186	**14.982**	16.597	18.541
	(5.547)	(5.527)	(5.255)	(5.351)	(5.291)	(5.919)	(4.886)	(5.703)	(5.569)	(6.232)	(5.556)	(5.479)	(6.863)
7	2.358	2.357	2.602	2.890	**2.260**	2.312	2.221	2.223	2.220	2.236	**2.216**	2.223	2.294
	(0.122)	(0.122)	(0.125)	(0.165)	(0.112)	(0.124)	(0.106)	(0.112)	(0.111)	(0.121)	(0.111)	(0.110)	(0.168)
8	4.013	**3.929**	4.276	5.151	3.986	4.046	3.948	3.953	3.949	3.973	**3.937**	**3.948**	4.213
	(0.258)	(0.253)	(0.259)	(0.394)	(0.241)	(0.253)	(0.225)	(0.244)	(0.246)	(0.249)	(0.244)	(0.245)	(0.341)
9	**6.072**	9.764	8.308	10.647	7.862	6.450	6.017	5.954	5.906	5.923	**5.732**	5.841	6.407
	(0.672)	(0.610)	(0.593)	(0.645)	(0.560)	(0.629)	(0.527)	(0.572)	(0.566)	(0.516)	(0.498)	(0.515)	(0.950)
10	**15.402**	17.611	19.287	20.819	16.039	15.754	16.629	14.970	15.017	14.806	**14.346**	14.568	16.666
	(1.644)	(2.225)	(1.885)	(1.844)	(1.748)	(1.825)	(2.348)	(1.913)	(1.967)	(1.705)	(1.506)	(1.517)	(1.534)

see that all columns of the second block often outperform the best estimator of the first block, which illustrates the theoretical result of the combining estimation methods. Secondly, the proposed methods (second to seventh column of the second block) always outperform the classical COBRA (first column) and KernelCOBRA (last column) for almost all kernels. Lastly, the combining estimation method with Gaussian kernel is the best one in both tables. In addition, Figure 2 below contains boxplots of RMSEs obtained from 100 independent runs of Model 1 and 10 (correlated and uncorrelated cases), computed on a computational machine with the following characteristics:

- Processor: 2x AMD Opteron 6174, 12C, 2.2GHz, 12x512K L2/12M L3 Cache, 80W ACP, DDR3-1333MHz.
- Memory: 64GB Memory for 2 CPUs, DDR3, 1333MHz.

These boxplots clearly show that the proposed method is around 3 to 10 times faster than the classical method by Biau et al. (2016), and is up to hundred times faster than KernelCOBRA by Guedj and Srinivasa Desikan (2018) with 500 values of bandwidth parameters.

4.2 Real public datasets

In this part, we consider three public datasets which are available and easily accessible on the internet. The first dataset (**Abalone**, available at Dua and Graff (2017a)) contains 4177 rows and 9 columns of measurements of abalones observed in Tasmania, Australia. We are interested in predicting the age of each abalone through the number of rings using its physical characteristics such as gender, size, weight, etc. The second dataset (**House**, available at Kaggle (2016)) comprises house sale prices for King County including Seattle. It contains homes sold between May 2014 and May 2015. The dataset consists of 21613 rows of houses and 21 columns of characteristics of each house including ID, Year of sale, Size, Location, etc. In this case, we want to predict the price of each house using all of its quantitative characteristics.

Finally, the last dataset (**Wine**, see Dua and Graff (2017b); Cortez et al. (2009)), which was also considered in Biau et al. (2016), containing 1599 rows of different types of wines, and 12 columns corresponding to different substances of red wines including the amount of different types of acids, sugar, chlorides, PH, etc. The variable of interest is *quality* which scales from 3 to 8 where 8 represents the best quality. We aim at predicting the quality of each wine, which is treated as a continuous variable, using all of its substances.

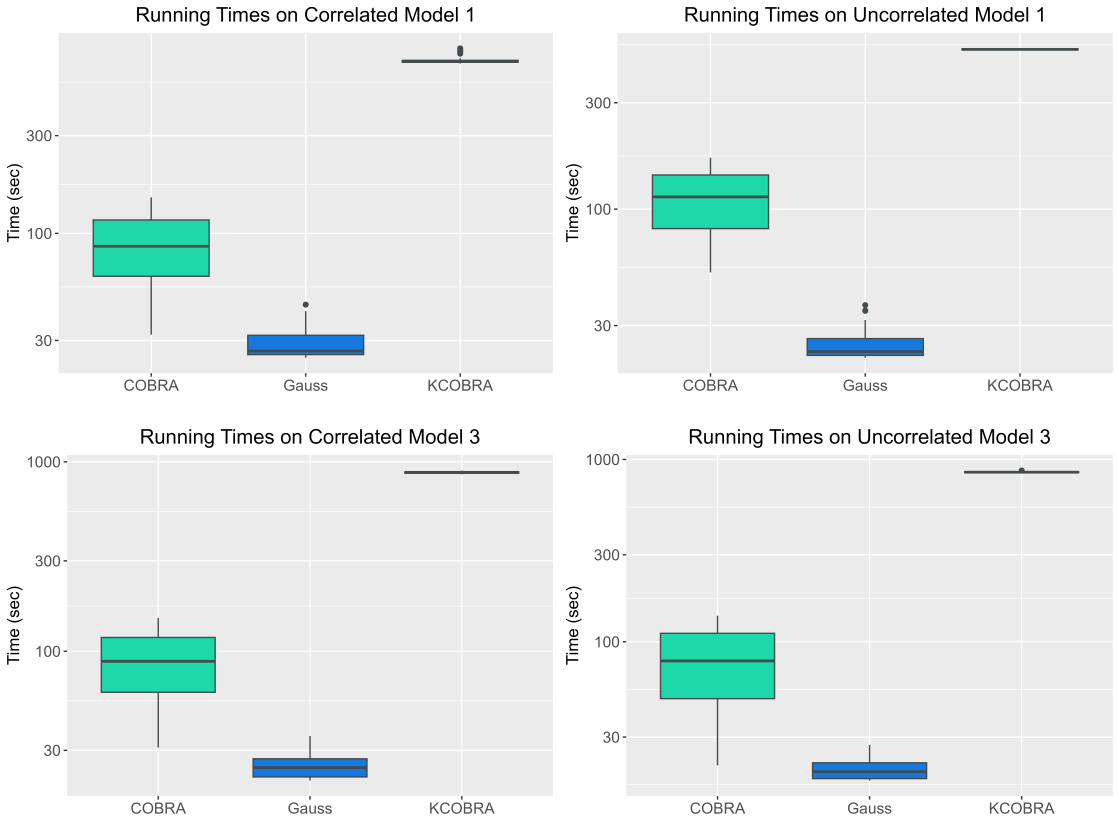

Figure 2: Boxplots of computational times of the three aggregation strategies implemented on model 1 and 3, with 500 of bandwidth parameters. Note that "Gauss" corresponds to the proposed method with Gaussian kernel, and the "Time" axis is in logarithmic scale.

The five primary regressors are Ridge, LASSO, kNN, Tree and Random Forest regression. In this case, the parameter $ntree = 500$ for random forest, and kNN is implemented using $k = 20, 12$ and 5 for Abalone, House and Wine dataset respectively. The five regressors are combined using the classical method by Biau et al. (2016), the proposed method using Gaussian kernel, and the KernelCOBRA by Guedj and Srinivasa Desikan (2018). In this case, 300 values of parameter h are considered for the classical COBRA and KernelCOBRA.

The average RMSEs obtained from 100 independent runs, evaluated on 20%-testing data of the three public datasets, are provided in Table 4 below (the first three rows). We observe that random forest is the best estimator among all the

basic estimators in the first block, and the proposed method (**Gauss**) either outperforms other columns (**Wine** and **Abalone**) or biases towards the best basic estimator (**House**). Moreover, the performances of the proposed method always exceed the ones of the classical COBRA and the KernelCOBRA.

4.3 Real private datasets

In this section, we provide the performances of the aggregation methods on other two (real) private datasets. The first dataset contains six columns corresponding to the six variables including *Air temperature, Input Pressure, Output Pressure, Flow, Water Temperature* and *Power Consumption* along with 2026 rows of hourly observations of these measurements of an air compressor machine provided by Cadet et al. (2005). The goal is to predict the power consumption of this machine using the five remaining explanatory variables. The second dataset is provided by the wind energy company Maïa Eolis. It contains 8721 observations of seven variables representing 10-minute measurements of *Electrical power, Wind speed, Wind direction, Temperature, Variance of wind speed* and *Variance of wind direction* measured from a wind turbine of the company (see, Fischer et al. (2017)). In this case, we aim at predicting the electrical power produced by the turbine using the remaining six measurements as explanatory variables. We use the same set of parameters as in the previous subsection except for kNN where in this case $k = 10$ and $k = 7$ are used for air compressor and wind turbine dataset respectively.

Table 4: Average RMSEs of real datasets.

Model	Las	Rid	kNN	Tr	RF	COBRA	Gauss	KCOBRA
Abalone	2.20 (0.07)	2.22 (0.08)	2.18 (0.06)	2.40 (0.07)	**2.15** (0.06)	2.17 (0.08)	**2.13** (0.06)	2.67 (0.12)
House	241083.96 (8883.11)	241072.97 (8906.33)	245153.61 (23548.37)	254099.65 (9350.89)	**205943.77** (7496.77)	223596.32 (13299.93)	**209955.28** (7815.62)	650943.60 (29565.23)
Wine	0.66 (0.03)	0.69 (0.05)	0.77 (0.03)	0.71 (0.03)	**0.62** (0.03)	0.65 (0.03)	**0.62** (0.02)	0.67 (0.04)
Air	**163.10** (3.69)	164.23 (3.75)	241.66 (5.87)	351.32 (31.88)	174.84 (6.55)	172.86 (7.64)	**163.25** (3.33)	1468.30 (78.47)
Turbine	70.05 (4.99)	68.99 (3.41)	44.52 (1.67)	81.71 (4.98)	**38.89** (1.51)	38.93 (1.56)	**37.14** (1.56)	515.41 (58.14)

The results obtained from 100 independent runs of the methods are presented in the last two rows (**Air** and **Turbine**) of Table 4 above. We observe on one hand that the proposed method (**Gauss**) outperforms the best basic estimators (**RF**) and the two other competitors (**COBRA** and **KCOBRA**) in the case of **Turbine** dataset. On the other hand, the performance of our method approaches the best basic estimator (**Las**) and outperforms the other aggregation methods in the case of **Air** dataset.

Moreover, boxplots of 100 runs measured on **Wine** and **Turbine** datasets (computed using the same computational machine as described in section 4.1) are also provided in Figure 3 below.

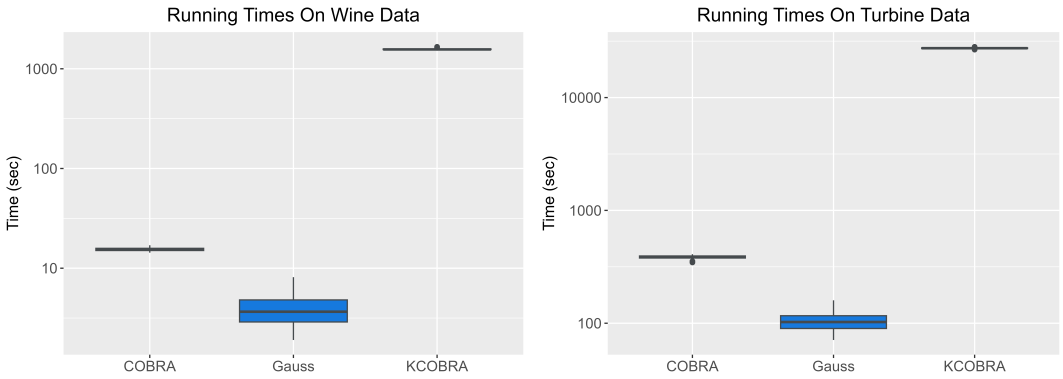

Figure 3: Boxplots of computational times of the proposed method and the two competitors implemented on **Wine** and **Turbine** datasets.

4.4 Application on a data of Magnetosphere- Ionosphere System provided by CEA

This section presents an application of the proposed method on a data provided by researchers of Commissariat à l'Énergie Atomique (CEA). In a collaboration with researchers of CEA on a research topic in Magnetosphere-Ionosphere System (see Kluth et al. (2022)), we are interested in constructing a global machine learning model of event-driven for estimating a physical quantity called *Pitch Angle Diffusion Coefficient* ($D_{\alpha\alpha}$) using three input data: electron at L-shell L, energy E, and equatorial pitch angle α. Pitch angle diffusion coefficient is one of the major mechanisms that drives the structure of the Van Allen radiation belts and causes the well-known two belt structure. Whistler mode waves which are known to play a crucial role in thermodynamics, electron acceleration, and electron precipitation in the atmosphere are also caused by the physical process of pitch angle diffusion. This quantity can be computed from statistical models derived from years of satellite observations of the hiss waves properties of different missions, or using a method called event-driven approach (Thorne et al. (2013)). We use in this study a database of event-driven diffusion coefficients that was generated for the studies of Ripoll et al. (2019). A very large fully observed dataset containing around two hundred million observations is available. However, one wants to construct

predictive models using reasonably small training data, therefore, a 3-dimensional grid made up of 4 values of $L \in \{2, 3, 4, 5\}$, 60 values of E and 256 values of α is considered. This filtering process creates a training dataset of size 61 440, simply called $\mathcal{D}_0$. Then, two training datasets are extracted: high-resolution ($\mathcal{D}_{\text{HR}}$) and low-resolution datasets ($\mathcal{D}_{\text{LR}}$). High-resolution dataset is composed of 84 pitch angles (α) and 60 energies bins (E), thus contains 20 160 data points. The low-resolution dataset is composed of only 14 pitch angles and 13 energies bins, thus contains only 728 data points. The table 5 below provides the structure of the described training datasets.

Data	L	E	α	**Size**
$\mathcal{D}_0$	4	60	256	61 440
$\mathcal{D}_{\text{HR}}$	4	60	84	20 160
$\mathcal{D}_{\text{LR}}$	4	13	14	728

Table 5: The high and low resolution training datasets.

It should be pointed out that the training datasets are noiseless (see Figure 4), and the relationship of $D_{\alpha\alpha}$ and α at some fixed couples (L, E) are illustrated in Figure 4 below.

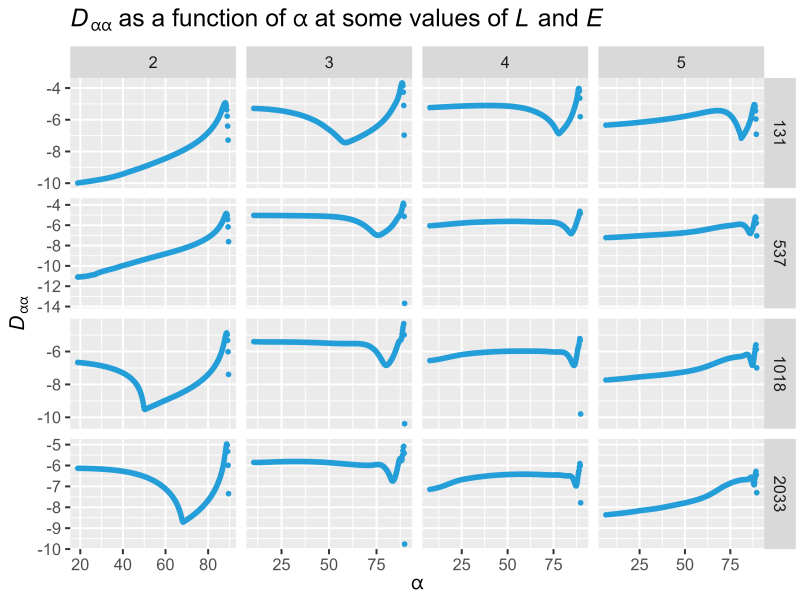

Figure 4: The relation between $D_{\alpha\alpha}$ and α at some cuts of L and E values.

In this part, we considered several regression models, including k-nearest neighbors (kNN), kernel regression (KerReg), regression tree (Tree), bagging (Bag), random forest (RF), radial basis (Radial), splines (Spline), and deep neural networks (DNN). These models were trained separately on the high-resolution ($\mathcal{D}_{\text{HR}}$) and low-resolution ($\mathcal{D}_{\text{LR}}$) training datasets.

To evaluate the prediction capability of these models, we extracted three different testing datasets from the fully observed data, which contains two hundred million observations. By using these testing datasets, we were able to compare the performance of the different regression models and identify the most effective one for the task at hand. Table 6 below describes the three testing datasets.

Data	**Description**
$\mathcal{D}_{\text{testHR}}$	For testing the models built on D_{HR}.
$\mathcal{D}_{\text{testLR}}$	For testing the models built on D_{LR}.
$\mathcal{D}_{\text{testL}}$	Contains more values of L other than $\{2, 3, 4, 5\}$. For testing the models built on both training data.

Table 6: The three testing datasets.

In both cases, the regression estimators were constructed using the entire training data ($\mathcal{D}_{\text{HR}}$ or $\mathcal{D}_{\text{LR}}$), which left no training data for aggregation. To avoid violating the independence assumption between the data used to train the individual estimators and the data used for aggregation, we randomly divided each testing dataset into two parts. The first part is used to optimize the bandwidth parameter h for the aggregation, while the remaining part is used as the actual testing dataset. The numerical results obtained from 50 independent runs of this procedure are presented in Figure 5 below.

The kernel-based consensual aggregation method is implemented using Gaussian kernel and is denoted by Gaussian. We observe that the tree-based models behave similarly and are the weak ones, and DNN is the best individual estimator as it provides the lowest average testing RMSE. On the other hand, the aggregation outperforms other basic estimators in the last three cases, and biases towards the best basic estimator on D_{testHR}.

Remark 1 *As the training data in our study are extracted selectively from the full observed data, the distributions of the training and testing data are not the same. For instance, L only takes values in* $\{2, 3, 4, 5\}$ *in the training data, while the testing data may have more decimal values. To overcome this limitation, we split the testing data into two parts, allowing us to fine-tune the smoothing parameter h and adjust the weights for predicting new observations coming from a different distribution.*

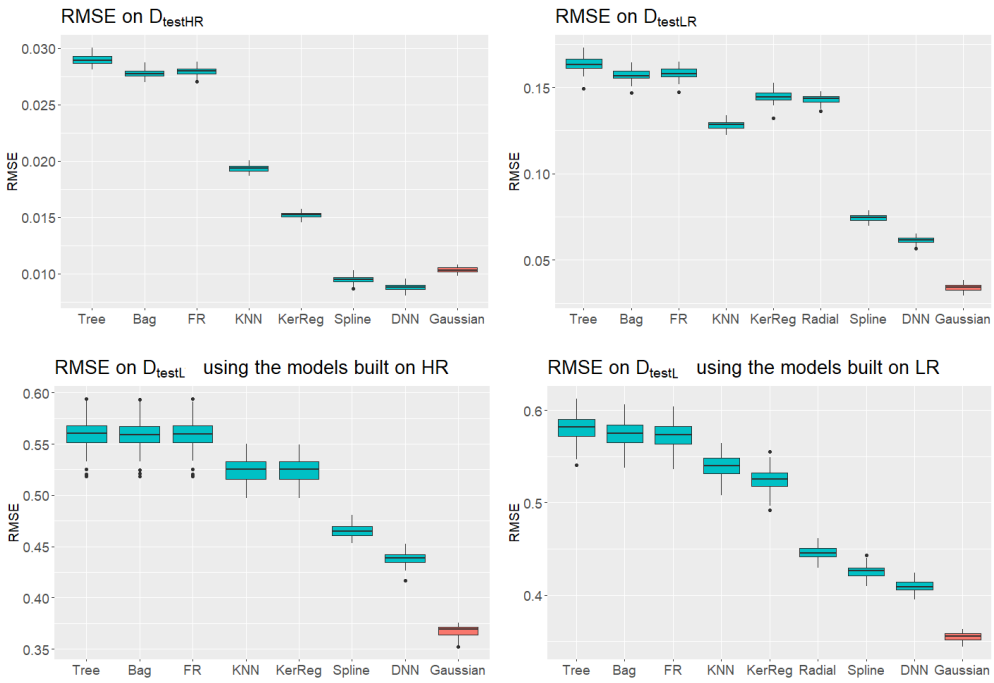

Figure 5: Boxplots of RMSEs over 50 runs of the algorithm. Note that Radial is built only on the training data $\mathcal{D}_{\text{LR}}$, therefore it is not presented in the two boxplots on the left-hand side (where the model are built using $\mathcal{D}_{\text{HR}}$). The last boxplot is the performance of the proposed aggregation method using Gaussian kernel.

This approach is practically useful because the basic models can be built on one source of an underlying distribution $\mathcal{L}_0$ and then used to predict observations from another source of distribution $\mathcal{L}_1$, which may be different from $\mathcal{L}_0$. In such cases, access to a part of the new source is required to adjust the weights in the aggregation, akin to a domain adaptation-like property. This adaptability of the aggregation method is a remarkable advantage and can lead to improved performance in diverse settings.

5 Conclusion

In conclusion, this study extends the context of a naive kernel-based consensual regression aggregation method to a more general regular kernel-based framework, and it demonstrates the consistency inheritance property of the method with the

same convergence rate. Additionally, we propose an optimization algorithm based on gradient descent to efficiently estimate the key parameter of the method with the computational speed up to several hundred times faster than the classical grid search. Our numerical simulations show that the performance of the method is significantly improved with smoother kernel functions. Furthermore, we demonstrate, in a real-world project with physics data, that the method exhibits a domain adaptation-like property, which opens up interesting directions for further study.

In practice, the performance of the consensual aggregation method depends on both the individual regression estimators and the final combination, which involves kernel functions. Therefore, calibration of hyperparameters in both steps is critical, and automated machine learning models may be useful for improving the performance of the global model.

6 Reproducibility of the experiments

For readers interested in reproducing our experiments, we have made some public datasets used in this article and the official source codes written in `python` and `R` of the algorithm available on our Github repository: https://github.com/hassothea/AggregationMethods.

7 Proofs

The following lemma, which is a variant of lemma 4.1 in Györfi et al. (2002) related to the property of binomial random variables, is needed.

Lemma 1 *Let $B(n,p)$ be the binomial random variable with parameters n and p. Then*

1. *For any $c > 0$,*

$$\mathbb{E}\Big[\frac{1}{c + B(n,p)}\Big] \leq \frac{2}{p(n+1)}.$$

2.

$$\mathbb{E}\Big[\frac{1}{B(n,p)}\mathbb{1}_{B(n,p)>0}\Big] \leq \frac{2}{p(n+1)}.$$

***Proof of Lemma* 1** *1. For any $c > 0$, one has*

$$\begin{aligned}
\mathbb{E}\Big[\frac{1}{c+B(n,p)}\Big] &= \sum_{k=0}^{n} \frac{1}{c+k} \times \frac{n!}{(n-k)!k!} p^k (1-p)^{n-k} \\
&= \sum_{k=0}^{n} \frac{1}{k+1} \times \frac{k+1}{k+c} \times \frac{n!}{(n-k)!k!} p^k (1-p)^{n-k} \\
&\leq \frac{2}{p(n+1)} \sum_{k=0}^{n} \frac{(n+1)! p^{k+1} (1-p)^{n+1-(k+1)}}{[n+1-(k+1)]!(k+1)!} \\
&\leq \frac{2}{p(n+1)} \sum_{k=0}^{n+1} \frac{(n+1)! p^{k} (1-p)^{n+1-k}}{[n+1-k]!k!} \\
&= \frac{2}{p(n+1)} (p+1-p)^{n+1} \\
&= \frac{2}{p(n+1)}
\end{aligned}$$

2.

$$\begin{aligned}
\mathbb{E}\Big[\frac{1}{B(n,p)} \mathbb{1}_{B(n,p)>0}\Big] &\leq \mathbb{E}\Big[\frac{2}{1+B(n,p)}\Big] \\
&= \sum_{k=0}^{n} \frac{2}{k+1} \times \frac{n!}{(n-k)!k!} p^k (1-p)^{n-k} \\
&= \frac{2}{p(n+1)} \sum_{k=0}^{n} \frac{(n+1)! p^{k+1} (1-p)^{n+1-(k+1)}}{[n+1-(k+1)]!(k+1)!} \\
&\leq \frac{2}{p(n+1)} \sum_{k=0}^{n+1} \frac{(n+1)! p^{k} (1-p)^{n+1-k}}{[n+1-k]!k!} \\
&= \frac{2}{p(n+1)} (p+1-p)^{n+1} \\
&= \frac{2}{p(n+1)}
\end{aligned}$$

■

***Proof of Proposition* 1** *For any square integrable function with respect to $\boldsymbol{r}_k(X)$, one has*

$$\begin{aligned}\mathbb{E}\Big[|g_n(\boldsymbol{r}_k(X)) - g^*(X)|^2\Big] &= \mathbb{E}\Big[|g_n(\boldsymbol{r}_k(X)) - g^*(\boldsymbol{r}_k(X)) + g^*(\boldsymbol{r}_k(X)) - g^*(X)|^2\Big] \\ &= \mathbb{E}\Big[|g_n(\boldsymbol{r}_k(X)) - g^*(\boldsymbol{r}_k(X))|^2\Big] \\ &\quad + 2\mathbb{E}\Big[(g_n(\boldsymbol{r}_k(X)) - g^*(\boldsymbol{r}_k(X)))(g^*(\boldsymbol{r}_k(X)) - g^*(X))\Big] \\ &\quad + \mathbb{E}\Big[|g^*(\boldsymbol{r}_k(X)) - g^*(X)|^2\Big].\end{aligned}$$

We consider the second term of the right hand side of the last equality,

$$\begin{aligned}&\mathbb{E}\Big[(g_n(\boldsymbol{r}_k(X)) - g^*(\boldsymbol{r}_k(X)))(g^*(\boldsymbol{r}_k(X)) - g^*(X))\Big] \\ &= \mathbb{E}_{\boldsymbol{r}_k(X)}\Big[\mathbb{E}_X\Big[(g_n(\boldsymbol{r}_k(X)) - g^*(\boldsymbol{r}_k(X)))(g^*(\boldsymbol{r}_k(X)) - g^*(X))\Big|\boldsymbol{r}_k(X)\Big]\Big] \\ &= \mathbb{E}_{\boldsymbol{r}_k(X)}\Big[(g_n(\boldsymbol{r}_k(X)) - g^*(\boldsymbol{r}_k(X)))(g^*(\boldsymbol{r}_k(X)) - \mathbb{E}[g^*(X)|\boldsymbol{r}_k(X)])\Big] \\ &= 0\end{aligned}$$

where $g^*(\boldsymbol{r}_k(X)) = \mathbb{E}[g^*(X)|\boldsymbol{r}_k(X)]$ *due to the definition of* $g^*(\boldsymbol{r}_k(X))$ *and the tower property of conditional expectation. It remains to check that*

$$\mathbb{E}\Big[|g^*(\boldsymbol{r}_k(X)) - g^*(X)|^2\Big] \leq \inf_{f\in\mathcal{G}} \mathbb{E}\Big[|f(\boldsymbol{r}_k(X)) - g^*(X)|^2\Big].$$

For any function f *s.t* $\mathbb{E}\Big[|f(\boldsymbol{r}_k(X))|^2\Big] < +\infty$*, one has*

$$\begin{aligned}\mathbb{E}\Big[|f(\boldsymbol{r}_k(X)) - g^*(X)|^2\Big] &= \mathbb{E}\Big[|f(\boldsymbol{r}_k(X)) - g^*(\boldsymbol{r}_k(X)) + g^*(\boldsymbol{r}_k(X)) - g^*(X)|^2\Big] \\ &= \mathbb{E}\Big[|f(\boldsymbol{r}_k(X)) - g^*(\boldsymbol{r}_k(X))|^2\Big] \\ &\quad + 2\mathbb{E}\Big[(f(\boldsymbol{r}_k(X)) - g^*(\boldsymbol{r}_k(X)))(g^*(\boldsymbol{r}_k(X)) - g^*(X))\Big] \\ &\quad + \mathbb{E}\Big[|g^*(\boldsymbol{r}_k(X)) - g^*(X)|^2\Big].\end{aligned}$$

Similarly,

$$\mathbb{E}\Big[(f(\boldsymbol{r}_k(X)) - g^*(\boldsymbol{r}_k(X)))(g^*(\boldsymbol{r}_k(X)) - g^*(X))\Big] = 0.$$

Therefore,

$$\mathbb{E}\Big[|f(\boldsymbol{r}_k(X)) - g^*(X)|^2\Big] = \mathbb{E}\Big[|f(\boldsymbol{r}_k(X)) - g^*(\boldsymbol{r}_k(X))|^2\Big] + \mathbb{E}\Big[|g^*(\boldsymbol{r}_k(X)) - g^*(X)|^2\Big].$$

As the first term of the right-hand side is nonnegative thus,

$$\mathbb{E}\Big[|g^*(\boldsymbol{r}_k(X)) - g^*(X)|^2\Big] \leq \inf_{f\in\mathcal{G}} \mathbb{E}\Big[|f(\boldsymbol{r}_k(X)) - g^*(X)|^2\Big].$$

Finally, we can conclude that

$$\mathbb{E}\Big[|g_n(\boldsymbol{r}_k(X)) - g^*(X)|^2\Big] \leq \mathbb{E}\Big[|g_n(\boldsymbol{r}_k(X)) - g^*(\boldsymbol{r}_k(X))|^2\Big] + \inf_{f\in\mathcal{G}} \mathbb{E}\Big[|f(\boldsymbol{r}_k(X)) - g^*(X)|^2\Big].$$

We obtain the particular case by restricting $\mathcal{G}$ to be the coordinates of $\mathbf{r}_k$, one has

$$\mathbb{E}\Big[|g_n(\mathbf{r}_k(X)) - g^*(X)|^2\Big] \leq \mathbb{E}\Big[|g_n(\mathbf{r}_k(X)) - g^*(\mathbf{r}_k(X))|^2\Big] + \min_{1\leq m\leq M} \mathbb{E}\Big[|r_{k,m}(X) - g^*(X)|^2\Big].$$

■

Proof of Proposition 2 *The procedure of proving this result is indeed the procedure of checking the conditions of Stone's theorem (see, for example, Stone (1977) and Chapter 4 of Györfi et al. (2002)) which is also used in the classical method by Biau et al. (2016). First of all, using the inequality:* $(a+b+c)^2 \leq 3(a^2+b^2+c^2)$, *one has*

$$\begin{aligned}
\mathbb{E}\Big[|g_n(\boldsymbol{r}_k(X)) - g^*(\boldsymbol{r}_k(X))|^2\Big] &= \mathbb{E}\Big[\Big|\sum_{i=1}^{\ell} W_{n,i}(X)Y_i - g^*(\boldsymbol{r}_k(X))\Big|^2\Big] \\
&= \mathbb{E}\Big[\Big|\sum_{i=1}^{\ell} W_{n,i}(X)[Y_i - g^*(\boldsymbol{r}_k(X_i))] \\
&\quad + \sum_{i=1}^{\ell} W_{n,i}(X)[g^*(\boldsymbol{r}_k(X_i)) - g^*(\boldsymbol{r}_k(X))] \\
&\quad + \sum_{i=1}^{\ell} W_{n,i}(X)g^*(\boldsymbol{r}_k(X)) - g^*(\boldsymbol{r}_k(X))\Big|^2\Big] \\
&\leq 3\mathbb{E}\Big[\Big|\sum_{i=1}^{\ell} W_{n,i}(X)[g^*(\boldsymbol{r}_k(X_i)) - g^*(\boldsymbol{r}_k(X))]\Big|^2\Big] \\
&\quad + 3\mathbb{E}\Big[\Big|\sum_{i=1}^{\ell} W_{n,i}(X)[Y_i - g^*(\boldsymbol{r}_k(X_i))]\Big|^2\Big] \\
&\quad + 3\mathbb{E}\Big[\Big|g^*(\boldsymbol{r}_k(X))\sum_{i=1}^{\ell}(W_{n,i}(X) - 1)\Big|^2\Big].
\end{aligned}$$

The three terms of the right-hand side are denoted by $A.1, A.2$ and $A.3$ respectively, thus one has

$$\mathbb{E}\Big[|g_n(\boldsymbol{r}_k(X)) - g^*(\boldsymbol{r}_k(X))|^2\Big] \leq 3(A.1 + A.2 + A.3).$$

*To prove the result, it is enough to prove that the three terms $A.1, A.2$ and $A.3$ vanish under the assumptions of **Proposition 2**. We deal with the first term $A.1$ in the following proposition.*

Proposition A.1 *Under the assumptions of **Proposition 2**,*

$$\lim_{\ell \to +\infty} \mathbb{E}\Big[\Big|\sum_{i=1}^{\ell} W_{n,i}(X)[g^*(\boldsymbol{r}_k(X_i)) - g^*(\boldsymbol{r}_k(X))]\Big|^2\Big] = 0.$$

***Proof of Proposition* A.1** *Using Cauchy-Schwarz's inequality, one has*

$$\begin{aligned}
A.1 &= \mathbb{E}\Big[\Big|\sum_{i=1}^{\ell} W_{n,i}(X)[g^*(\boldsymbol{r}_k(X_i)) - g^*(\boldsymbol{r}_k(X))]\Big|^2\Big] \\
&= \mathbb{E}\Big[\Big|\sum_{i=1}^{\ell} \sqrt{W_{n,i}(X)}\sqrt{W_{n,i}(X)}[g^*(\boldsymbol{r}_k(X_i)) - g^*(\boldsymbol{r}_k(X))]\Big|^2\Big] \\
&\leq \mathbb{E}\Big[\Big(\sum_{i=1}^{\ell} W_{n,i}(X)\Big)\sum_{i=1}^{\ell} W_{n,i}(X)[g^*(\boldsymbol{r}_k(X_i)) - g^*(\boldsymbol{r}_k(X))]^2\Big] \\
&= \mathbb{E}\Big[\sum_{i=1}^{\ell} W_{n,i}(X)[g^*(\boldsymbol{r}_k(X_i)) - g^*(\boldsymbol{r}_k(X))]^2\Big] \\
&\stackrel{\text{def}}{=} A_n.
\end{aligned}$$

Note that the regression function g^* satisfies $\mathbb{E}[|g^*(\boldsymbol{r}_k(X))|^2] < +\infty$, thus it can be approximated in L_2 sense by a continuous function with compact support named $\tilde{g}$ (see, for example, Theorem A.1 in Devroye et al. (1997)). This means that for any $\varepsilon > 0$, there exists a continuous function with compact support $\tilde{g}$ such that,

$$\mathbb{E}[|g^*(\boldsymbol{r}_k(X)) - \tilde{g}(\boldsymbol{r}_k(X))|^2] < \varepsilon.$$

Thus, one has

$$A_n = \mathbb{E}\Big[\sum_{i=1}^{\ell} W_{n,i}(X)[g^*(\boldsymbol{r}_k(X_i)) - g^*(\boldsymbol{r}_k(X))]^2\Big]$$

$$
\begin{aligned}
&\leq 3\mathbb{E}\Big[\sum_{i=1}^{\ell} W_{n,i}(X)[g^*(\boldsymbol{r}_k(X_i)) - \tilde{g}(\boldsymbol{r}_k(X_i))]^2\Big] \\
&\quad + 3\mathbb{E}\Big[\sum_{i=1}^{\ell} W_{n,i}(X)[\tilde{g}(\boldsymbol{r}_k(X_i)) - \tilde{g}(\boldsymbol{r}_k(X))]^2\Big] \\
&\quad + 3\mathbb{E}\Big[\sum_{i=1}^{\ell} W_{n,i}(X)[\tilde{g}(\boldsymbol{r}_k(X)) - g^*(\boldsymbol{r}_k(X))]^2\Big] \\
&\overset{\text{def}}{=} 3(A_{n1} + A_{n2} + A_{n3}).
\end{aligned}
$$

We deal with each term of the last upper bound as follows.

- *Computation of A_{n3}: applying the definition of $\tilde{g}$,*

$$
\begin{aligned}
A_{n3} &= \mathbb{E}\Big[\sum_{i=1}^{\ell} W_{n,i}(X)[\tilde{g}(\boldsymbol{r}_k(X)) - g^*(\boldsymbol{r}_k(X))]^2\Big] \\
&\leq \mathbb{E}\Big[|\tilde{g}(\boldsymbol{r}_k(X)) - g^*(\boldsymbol{r}_k(X))|^2\Big] < \varepsilon.
\end{aligned}
$$

- *Computation of A_{n1}: denoted by μ the distribution of X. Thus,*

$$
\begin{aligned}
A_{n1} &= \mathbb{E}\Big[\sum_{i=1}^{\ell} W_{n,i}(X)|g^*(\boldsymbol{r}_k(X_i)) - \tilde{g}(\boldsymbol{r}_k(X_i))|^2\Big] \\
&= \ell\mathbb{E}\Big[W_{n,1}(X)|g^*(\boldsymbol{r}_k(X_1)) - \tilde{g}(\boldsymbol{r}_k(X_1))|^2\Big] \\
&= \ell\mathbb{E}\Big[\frac{K_h(\boldsymbol{r}_k(X) - \boldsymbol{r}_k(X_1))}{\sum_{j=1}^{\ell} K_h(\boldsymbol{r}_k(X) - \boldsymbol{r}_k(X_j))}|g^*(\boldsymbol{r}_k(X_1)) - \tilde{g}(\boldsymbol{r}_k(X_1))|^2\Big] \\
&= \ell\mathbb{E}_{\mathcal{D}_k}\Big[\mathbb{E}_{\{X_j\}_{j=1}^{\ell}}\Big[\int \frac{K_h(\boldsymbol{r}_k(v) - \boldsymbol{r}_k(X_1))}{\sum_{j=1}^{\ell} K_h(\boldsymbol{r}_k(v) - \boldsymbol{r}_k(X_j))}|g^*(\boldsymbol{r}_k(X_1)) - \tilde{g}(\boldsymbol{r}_k(X_1))|^2\mu(dv)\Big|\mathcal{D}_k\Big]\Big] \\
&= \ell\mathbb{E}_{\mathcal{D}_k}\Big[\mathbb{E}_{\{X_j\}_{j=2}^{\ell}}\Big[\int\int |g^*(\boldsymbol{r}_k(u)) - \tilde{g}(\boldsymbol{r}_k(u))|^2 \times \\
&\quad \frac{K_h(\boldsymbol{r}_k(v) - \boldsymbol{r}_k(u))}{K_h(\boldsymbol{r}_k(v) - \boldsymbol{r}_k(u)) + \sum_{j=2}^{\ell} K_h(\boldsymbol{r}_k(v) - \boldsymbol{r}_k(X_j))}\mu(du)\mu(dv)\Big|\mathcal{D}_k\Big]\Big] \\
&= \ell\mathbb{E}_{\mathcal{D}_k}\Big[\int |g^*(\boldsymbol{r}_k(u)) - \tilde{g}(\boldsymbol{r}_k(u))|^2 \times \\
&\quad \mathbb{E}_{\{X_j\}_{j=2}^{\ell}}\Big[\int \frac{K_h(\boldsymbol{r}_k(v) - \boldsymbol{r}_k(u))\mu(dv)}{K_h(\boldsymbol{r}_k(v) - \boldsymbol{r}_k(u)) + \sum_{j=2}^{\ell} K_h(\boldsymbol{r}_k(v) - \boldsymbol{r}_k(X_j))}\Big|\mathcal{D}_k\Big]\mu(du)\Big] \\
&= \ell\mathbb{E}_{\mathcal{D}_k}\Big[\int |g^*(\boldsymbol{r}_k(u)) - \tilde{g}(\boldsymbol{r}_k(u))|^2 \times I(u,\ell)\mu(du)\Big].
\end{aligned}
$$

Fubini's theorem (Folland (1999)) is employed to obtain the result of the last bound where the inner conditional expectation is denoted by $I(u,\ell)$. We bound $I(u,\ell)$ using the argument of covering $\mathbb{R}^M$ with a countable family of balls $\mathcal{B}\overset{\text{def}}{=}\{B_M(x_i,\rho/2):i=1,2,....\}$ and the facts that

1. $\boldsymbol{r}_k(v)\in B_M(\boldsymbol{r}_k(u)+hx_i,h\rho/2)\Rightarrow B_M(\boldsymbol{r}_k(u)+hx_i,h\rho/2)\subset B_M(\boldsymbol{r}_k(v),h\rho)$.
2. $b\mathbb{1}_{\{B_M(0,\rho)\}}(z)<K(z)\leq 1,\forall z\in\mathbb{R}^M$.

Now, let

- $A_{i,h}(u)\overset{\text{def}}{=}\{v\in\mathbb{R}^d:\|\boldsymbol{r}_k(v)-\boldsymbol{r}_k(u)-hx_i\|<h\rho/2\}$.
- $B^{\ell}_{i,h}(u)\overset{\text{def}}{=}\sum_{j=2}^{\ell}\mathbb{1}_{\{\|\boldsymbol{r}_k(X_j)-\boldsymbol{r}_k(u)-hx_i\|<h\rho/2\}}$.

Thus, one has

$$
\begin{aligned}
I(u,\ell)&\overset{\text{def}}{=}\mathbb{E}_{\{X_j\}_{j=2}^{\ell}}\Big[\int\frac{K_h(\boldsymbol{r}_k(v)-\boldsymbol{r}_k(u))\mu(dv)}{K_h(\boldsymbol{r}_k(v)-\boldsymbol{r}_k(u))+\sum_{j=2}^{\ell}K_h(\boldsymbol{r}_k(v)-\boldsymbol{r}_k(X_j))}\Big|\mathcal{D}_k\Big]\\
&\leq\mathbb{E}_{\{X_j\}_{j=2}^{\ell}}\Big[\sum_{i=1}^{+\infty}\int_{v:\|\boldsymbol{r}_k(v)-\boldsymbol{r}_k(u)-hx_i\|<h\rho/2}\\
&\quad\frac{K_h(\boldsymbol{r}_k(v)-\boldsymbol{r}_k(u))\mu(dv)}{K_h(\boldsymbol{r}_k(v)-\boldsymbol{r}_k(u))+\sum_{j=2}^{\ell}K_h(\boldsymbol{r}_k(v)-\boldsymbol{r}_k(X_j))}\Big|\mathcal{D}_k\Big]\\
&\leq\mathbb{E}_{\{X_j\}_{j=2}^{\ell}}\Big[\sum_{i=1}^{+\infty}\int_{A_{i,h}(u)}\\
&\quad\frac{\sup_{z:\|z-hx_i\|<h\rho/2}K_h(z)\mu(dv)}{\sup_{z:\|z-hx_i\|<h\rho/2}K_h(z)+\sum_{j=2}^{\ell}K_h(\boldsymbol{r}_k(v)-\boldsymbol{r}_k(X_j))}\Big|\mathcal{D}_k\Big]\\
&\leq\mathbb{E}_{\{X_j\}_{j=2}^{\ell}}\Big[\sum_{i=1}^{+\infty}\int_{A_{i,h}(u)}\\
&\quad\frac{\sup_{z:\|z-hx_i\|<h\rho/2}K_h(z)\mu(dv)}{\sup_{z:\|z-hx_i\|<h\rho/2}K_h(z)+b\sum_{j=2}^{\ell}\mathbb{1}_{\{\|\boldsymbol{r}_k(v)-\boldsymbol{r}_k(X_j)\|<h\rho\}}}\Big|\mathcal{D}_k\Big]\\
&\leq\frac{1}{b}\mathbb{E}_{\{X_j\}_{j=2}^{\ell}}\Big[\sum_{i=1}^{+\infty}\int_{A_{i,h}(u)}\\
&\quad\frac{\sup_{z:\|z-hx_i\|<h\rho/2}K_h(z)\mu(dv)}{\sup_{z:\|z-hx_i\|<h\rho/2}K_h(z)+\sum_{j=2}^{\ell}\mathbb{1}_{\{\|\boldsymbol{r}_k(X_j)-\boldsymbol{r}_k(u)-hx_i\|<h\rho/2\}}}\Big|\mathcal{D}_k\Big]
\end{aligned}
$$

$$\leq \frac{1}{b}\sum_{i=1}^{+\infty}\mathbb{E}_{\{X_j\}_{j=2}^{\ell}}\Big[\frac{\sup_{z:\|z-hx_i\|<h\rho/2}K_h(z)\mu(A_{i,h}(u))}{\sup_{z:\|z-hx_i\|<h\rho/2}K_h(z)+B_{i,h}^{\ell}(u)}\Big|\mathcal{D}_k\Big].$$

Note that $B_{i,h}^{\ell}(u)$ is a binomial random variable $B(\ell-1,\mu(A_{i,h}(u)))$ under the law of $\{X_j\}_{j=2}^{\ell}$. Applying part 1 of lemma 1, one has

$$\begin{aligned}
I(u,\ell) &\leq \frac{1}{b}\sum_{i=1}^{+\infty}\frac{2\sup_{z:\|z-hx_i\|<h\rho/2}K_h(z)\mu(A_{i,h}(u))}{\ell\mu(A_{i,h}(u))}\\
&\leq \frac{2}{b\ell}\sum_{i=1}^{+\infty}\sup_{w:\|w-x_i\|<\rho/2}K(w)\\
&= \frac{2}{b\ell}\sum_{i=1}^{+\infty}\sup_{w\in B_M(x_i,\rho/2)}K(w)\\
&\leq \frac{2}{b\ell}\sum_{i=1}^{+\infty}\sup_{w\in B_M(x_i,\rho/2)}K(w)\\
&\leq \frac{2}{b\ell\lambda_M(B_M(0,\rho/2))}\sum_{i=1}^{+\infty}\int_{B_M(x_i,\rho/2)}\sup_{w\in B_M(x_i,\rho/2)}K(w)dy\\
&\leq \frac{2}{b\ell\lambda_M(B_M(0,\rho/2))}\sum_{i=1}^{+\infty}\int_{B_M(x_i,\rho/2)}\sup_{w\in B_M(y,\rho)}K(w)dy\\
&\leq \frac{2\kappa_M}{b\ell\lambda_M(B_M(0,\rho/2))}\underbrace{\int\sup_{w\in B_M(y,\rho)}K(w)dy}_{=\ \kappa_0\ \text{by (4)}}\\
&\leq \frac{2\kappa_M\kappa_0}{b\ell\lambda_M(B_M(0,\rho))}\stackrel{\text{def}}{=}\frac{C(b,\rho,\kappa_0,M)}{\ell}<+\infty
\end{aligned}$$

where λ_M denotes the Lebesque measure on of $\mathbb{R}^M$, κ_M denotes the number of balls covering a certain element of $\mathbb{R}^M$, and the constant part is denoted by $C(b,\rho,\kappa_0,M)$ depending on the parameters indicated in the bracket. The last inequality is attained from the fact that the overlapping integrals $\sum_{i=1}^{+\infty}\int_{B_M(x_i,\rho/2)}\sup_{z\in B_M(y,\rho/2)}K(z)dy$ is bounded above by the integral over the entire space $\int\sup_{z\in B_M(y,\rho/2)}K(z)dy$ multiplying by the number of covering balls k_M. Therefore,

$$
\begin{aligned}
A_{n1} &\leq \ell \frac{C(b,\rho,\kappa_0,M)}{\ell} \mathbb{E}_{\mathcal{D}_k}\Big[\int |g^*(\boldsymbol{r}_k(u)) - \tilde{g}(\boldsymbol{r}_k(u))|^2 \mu(du)\Big] \\
&= C(b,\rho,\kappa_0,M)\mathbb{E}\Big[|\tilde{g}(\boldsymbol{r}_k(X)) - g^*(\boldsymbol{r}_k(X))|^2\Big] \\
&< C(b,\rho,\kappa_0,M)\varepsilon.
\end{aligned}
$$

- *Computation of A_{n2}: for any $\delta > 0$ one has*

$$
\begin{aligned}
A_{n2} &= \mathbb{E}\Big[\sum_{i=1}^{\ell} W_{n,i}(X)|\tilde{g}(\boldsymbol{r}_k(X_i)) - \tilde{g}(\boldsymbol{r}_k(X))|^2\Big] \\
&= \mathbb{E}\Big[\sum_{i=1}^{\ell} W_{n,i}(X)|\tilde{g}(\boldsymbol{r}_k(X_i)) - \tilde{g}(\boldsymbol{r}_k(X))|^2 \mathbb{1}_{\{\|\boldsymbol{r}_k(X_i)-\boldsymbol{r}_k(X)\|\geq\delta\}}\Big] \\
&\quad + \mathbb{E}\Big[\sum_{i=1}^{\ell} W_{n,i}(X)|\tilde{g}(\boldsymbol{r}_k(X_i)) - \tilde{g}(\boldsymbol{r}_k(X))|^2 \mathbb{1}_{\{\|\boldsymbol{r}_k(X_i)-\boldsymbol{r}_k(X)\|<\delta\}}\Big] \\
&\leq 4 \sup_{u\in\mathbb{R}^d} |\tilde{g}(\boldsymbol{r}_k(u))|^2 \mathbb{E}\Big[\sum_{i=1}^{\ell} W_{n,i}(X)\mathbb{1}_{\{\|\boldsymbol{r}_k(X_i)-\boldsymbol{r}_k(X)\|\geq\delta\}}\Big] \\
&\quad + \sup_{u,v\in\mathbb{R}^d:\|\boldsymbol{r}_k(u)-\boldsymbol{r}_k(v)\|<\delta} |\tilde{g}(\boldsymbol{r}_k(u)) - \tilde{g}(\boldsymbol{r}_k(v))|^2
\end{aligned}
$$

Using the uniform continuity of $\tilde{g}$, the second term of the upper bound of A_{n2} tends to 0 when δ tends 0. Thus, we only need to prove that the first term of this upper bound also tends to 0. We follow a similar procedure as in the previous part:

$$
\begin{aligned}
&\mathbb{E}\Big[\sum_{i=1}^{\ell} W_{n,i}(X)\mathbb{1}_{\{\|\boldsymbol{r}_k(X_i)-\boldsymbol{r}_k(X)\|\geq\delta\}}\Big] \\
&= \mathbb{E}_{\mathcal{D}_k}\Big[\sum_{i=1}^{\ell} \mathbb{E}_{X,\{X_j\}_{j=1}^{\ell}}\Big[W_{n,i}(X)\mathbb{1}_{\{\|\boldsymbol{r}_k(X)-\boldsymbol{r}_k(X_i)\|\geq\delta\}}\Big|\mathcal{D}_k\Big]\Big] \\
&= \mathbb{E}_{\mathcal{D}_k}\Big[\sum_{i=1}^{\ell} \mathbb{E}_{\{X_j\}_{j=1}^{\ell}}\Big[\int \frac{K_h(\boldsymbol{r}_k(v)-\boldsymbol{r}_k(X_i))\mathbb{1}_{\{\|\boldsymbol{r}_k(v)-\boldsymbol{r}_k(X_i)\|\geq\delta\}}}{\sum_{j=1}^{\ell} K_h(\boldsymbol{r}_k(v)-\boldsymbol{r}_k(X_j))}\mu(dv)\Big|\mathcal{D}_k\Big]\Big] \\
&= \ell\mathbb{E}_{\mathcal{D}_k}\Big[\mathbb{E}_{\{X_j\}_{j=2}^{\ell}}\Big[\int\int \frac{K_h(\boldsymbol{r}_k(v)-\boldsymbol{r}_k(u))\mathbb{1}_{\{\|\boldsymbol{r}_k(v)-\boldsymbol{r}_k(u)\|\geq\delta\}}\mu(du)\mu(dv)}{K_h(\boldsymbol{r}_k(v)-\boldsymbol{r}_k(u)) + \sum_{j=2}^{\ell} K_h(\boldsymbol{r}_k(v)-\boldsymbol{r}_k(X_j))}\Big|\mathcal{D}_k\Big]\Big] \\
&= \ell\mathbb{E}_{\mathcal{D}_k}\Big[\int J(u,\ell)\mu(du)\Big].
\end{aligned}
$$

Fubini's theorem is applied to obtain the last equation where for any $u \in \mathbb{R}^d$,

$$
\begin{aligned}
J(u,\ell) &\stackrel{\text{def}}{=} \mathbb{E}_{\{X_j\}_{j=2}^{\ell}}\Big[\int \frac{K_h(\boldsymbol{r}_k(v)-\boldsymbol{r}_k(u))\mathbb{1}_{\{\|\boldsymbol{r}_k(v)-\boldsymbol{r}_k(u)\|\geq\delta\}}\mu(dv)}{K_h(\boldsymbol{r}_k(v)-\boldsymbol{r}_k(u))+\sum_{j=2}^{\ell}K_h(\boldsymbol{r}_k(v)-\boldsymbol{r}_k(X_j))}\Big|\mathcal{D}_k\Big] \\
&\leq \mathbb{E}_{\{X_j\}_{j=2}^{\ell}}\Big[\sum_{i=1}^{+\infty}\int_{v:\|\boldsymbol{r}_k(v)-\boldsymbol{r}_k(u)-hx_i\|<h\rho/2} \\
&\quad \frac{K_h(\boldsymbol{r}_k(v)-\boldsymbol{r}_k(u))\mathbb{1}_{\{\|\boldsymbol{r}_k(v)-\boldsymbol{r}_k(u)\|\geq\delta\}}}{K_h(\boldsymbol{r}_k(v)-\boldsymbol{r}_k(u))+\sum_{j=2}^{\ell}K_h(\boldsymbol{r}_k(v)-\boldsymbol{r}_k(X_j))}\mu(dv)\Big|\mathcal{D}_k\Big] \\
&\leq \mathbb{E}_{\{X_j\}_{j=2}^{\ell}}\Big[\sum_{i=1}^{+\infty}\int_{A_{i,h}(u)} \\
&\quad \frac{\sup_{z:\|z-hx_i\|<h\rho/2}K_h(z)\mathbb{1}_{\{\|z\|\geq\delta\}}}{\sup_{z:\|z-hx_i\|<h\rho/2}K_h(z)+\sum_{j=2}^{\ell}K_h(\boldsymbol{r}_k(v)-\boldsymbol{r}_k(X_j))}\mu(dv)\Big|\mathcal{D}_k\Big] \\
&\leq \sum_{i=1}^{+\infty}\sup_{z:\|z-hx_i\|<h\rho/2}K_h(z)\mathbb{1}_{\{\|z\|\geq\delta\}}\times\mathbb{E}_{\{X_j\}_{j=2}^{\ell}}\Big[\int_{A_{i,h}(u)} \\
&\quad \frac{\mu(dv)}{\sup_{z:\|z-hx_i\|<h\rho/2}K_h(z)+b\sum_{j=2}^{\ell}\mathbb{1}_{\{\|\boldsymbol{r}_k(X_j)-\boldsymbol{r}_k(v)\|<h\rho\}}}\Big|\mathcal{D}_k\Big] \\
&\leq \sum_{i=1}^{+\infty}\sup_{z:\|z-hx_i\|<h\rho/2}K_h(z)\mathbb{1}_{\{\|z\|\geq\delta\}}\times\mathbb{E}_{\{X_j\}_{j=2}^{\ell}}\Big[\int_{A_{i,h}(u)} \\
&\quad \mathbb{E}_{\{X_j\}_{j=2}^{\ell}}\Big[\frac{\mu(dv)}{\sup_{z:\|z-hx_i\|<h\rho/2}K_h(z)+b\sum_{j=2}^{\ell}\mathbb{1}_{\{\|\boldsymbol{r}_k(X_j)-\boldsymbol{r}_k(u)-hx_i\|<h\rho/2\}}}\Big|\mathcal{D}_k\Big] \\
&\leq \sum_{i=1}^{+\infty}\sup_{z:\|z-hx_i\|<h\rho/2}K_h(z)\mathbb{1}_{\{\|z\|\geq\delta\}}\mu(A_{i,h}(u))\times \\
&\quad \frac{1}{b}\mathbb{E}_{\{X_j\}_{j=2}^{\ell}}\Big[\frac{1}{\sup_{z:\|z-hx_i\|<h\rho/2}K_h(z)+B_{i,h}^{\ell}(u)}\Big|\mathcal{D}_k\Big] \\
&\leq \frac{1}{b}\sum_{i=1}^{+\infty}\frac{2\sup_{z:\|z-hx_i\|<h\rho/2}K_h(z)\mu(A_{i,h}(u))\mathbb{1}_{\{\|z\|\geq\delta\}}}{\ell\mu(A_{i,h}(u))} \\
&\leq \frac{2}{b\ell}\sum_{i=1}^{+\infty}\sup_{w:\|w-x_i\|<\rho/2}K(w)\mathbb{1}_{\{\|w\|\geq\delta/h\}}.
\end{aligned}
$$

Thus, one has

$$
\mathbb{E}\Big[\sum_{i=1}^{\ell}W_{n,i}(X)\mathbb{1}_{\{\|\boldsymbol{r}_k(X_i)-\boldsymbol{r}_k(X)\|\geq\delta\}}\Big] \leq \ell\frac{2}{b\ell}\sum_{i=1}^{+\infty}\sup_{w\in B_M(x_i,\rho/2)}K(w)\mathbb{1}_{\{\|w\|\geq\delta/h\}}
$$

When both $h \to 0$ and $\delta \to 0$ satisfying $\delta/h \to +\infty$, the upper bound series converges to zero. Indeed, it is a non-negative convergent series thanks to the proof of $I(u,l)$ in the previous part. Moreover, the general term of the series, $s_k = \sup_{w \in B_M(x_k, \rho/2)} K(w) \mathbb{1}_{\{\|w\| \geq \delta/h\}}$, satisfying $\lim_{\delta/h \to +\infty} s_k = 0$ for all $k \geq 1$. Therefore, this series converges to zero when $h \to 0, \delta \to 0$ such that $\delta/h \to +\infty$.

In conclusion, when $\ell \to +\infty$ and $\varepsilon, h, \delta \to 0$ such that $\delta/h \to +\infty$, all the three terms of the upper bound of A_n tend to 0, so does A_n.

■

Proposition A.2 *Under the assumptions of **Proposition 2**,*

$$\lim_{\ell \to +\infty} \mathbb{E}\Big[\Big| \sum_{i=1}^{\ell} W_{n,i}(X)[Y_i - g_n(\boldsymbol{r}_k(X_i))]\Big|^2\Big] = 0.$$

***Proof of Proposition* A.2** *Using the independence between (X_i, Y_i) and (X_j, Y_j) for all $i \neq j$, one has*

$$\begin{aligned} A.2 &= \mathbb{E}\Big[\Big| \sum_{i=1}^{\ell} W_{n,i}(X)[Y_i - g_n(\boldsymbol{r}_k(X_i))]\Big|^2\Big] \\ &= \sum_{1 \leq i,j \leq \ell} \mathbb{E}\Big[W_{n,i}(X) W_{n,j}(X)[Y_i - g_n(\boldsymbol{r}_k(X_i))][Y_j - g_n(\boldsymbol{r}_k(X_j))]\Big] \\ &= \mathbb{E}\Big[\sum_{i=1}^{\ell} W_{n,i}^2(X)|Y_i - g_n(\boldsymbol{r}_k(X_i))|^2\Big] = \mathbb{E}\Big[\sum_{i=1}^{\ell} W_{n,i}^2(X)\sigma^2(\boldsymbol{r}_k(X_i))\Big] \end{aligned}$$

where

$$\sigma^2(\boldsymbol{r}_k(x)) \stackrel{\text{def}}{=} \mathbb{E}[(Y_i - g_n(\boldsymbol{r}_k(X_i)))^2 | \boldsymbol{r}_k(x)].$$

Thus, based on the assumption of X and Y we have $\sigma^2 \in L_1(\mu)$. Therefore, σ^2 can be approximated in L_1 sense i.e., for any $\varepsilon > 0, \exists \tilde{\sigma}^2$ a continuous function with compact support such that

$$\mathbb{E}[|\sigma^2(\boldsymbol{r}_k(X)) - \tilde{\sigma}^2(\boldsymbol{r}_k(X))|] < \varepsilon.$$

Thus, one has

$$\begin{aligned} A.2 &\leq \mathbb{E}\Big[\sum_{i=1}^{\ell} W_{n,i}^2(X)\tilde{\sigma}^2(\boldsymbol{r}_k(X_i))\Big] + \mathbb{E}\Big[\sum_{i=1}^{\ell} W_{n,i}^2(X)|\sigma^2(\boldsymbol{r}_k(X_i)) - \tilde{\sigma}^2(\boldsymbol{r}_k(X_i))|\Big] \\ &\leq \sup_{u \in \mathbb{R}^d} |\tilde{\sigma}^2(\boldsymbol{r}_k(u))| \mathbb{E}\Big[\sum_{i=1}^{\ell} W_{n,i}^2(X)\Big] + \mathbb{E}\Big[\sum_{i=1}^{\ell} W_{n,i}^2(X)|\sigma^2(\boldsymbol{r}_k(X_i)) - \tilde{\sigma}^2(\boldsymbol{r}_k(X_i))|\Big]. \end{aligned}$$

Using similar argument as in the case of A_{n1} and the fact that $W_{n,i}(x) \leq 1, \forall i = 1, 2, ..., \ell$, thus for any $\varepsilon > 0$, one has

$$\mathbb{E}\Big[\sum_{i=1}^{\ell} W_{n,i}^2(X)|\sigma^2(\boldsymbol{r}_k(X_i)) - \tilde{\sigma}^2(\boldsymbol{r}_k(X_i))|\Big] \leq \mathbb{E}\Big[\sum_{i=1}^{\ell} W_{n,i}(X)|\sigma^2(\boldsymbol{r}_k(X_i)) - \tilde{\sigma}^2(\boldsymbol{r}_k(X_i))|\Big] < C(b, \rho, \kappa_0, M)\varepsilon.$$

Therefore, it remains to prove that $\mathbb{E}[\sum_{i=1}^{\ell} W_{n,i}^2(X)] \to 0$ as $\ell \to +\infty$. As $b\mathbb{1}_{\{B_M(0,\rho)\}}(z) < K(z) \leq 1, \forall z \in \mathbb{R}^M$ with the convention of $0/0 = 0$, for a fixed $\delta > 0$, one has

$$\begin{aligned}
\sum_{i=1}^{\ell} W_{n,i}^2(X) &= \sum_{i=1}^{\ell} \Big(\frac{K_h(\boldsymbol{r}_k(X) - \boldsymbol{r}_k(X_i))}{\sum_{j=1}^{\ell} K_h(\boldsymbol{r}_k(X) - \boldsymbol{r}_k(X_j))}\Big)^2 \\
&\leq \frac{\sum_{i=1}^{\ell} K_h(\boldsymbol{r}_k(X) - \boldsymbol{r}_k(X_i))}{\Big(\sum_{j=1}^{\ell} K_h(\boldsymbol{r}_k(X) - \boldsymbol{r}_k(X_j))\Big)^2} \\
&\leq \min\Big\{\delta, \frac{\mathbb{1}_{\{\sum_{j=1}^{\ell} K_h(\boldsymbol{r}_k(X) - \boldsymbol{r}_k(X_j)) > 0\}}}{\sum_{j=1}^{\ell} K_h(\boldsymbol{r}_k(X) - \boldsymbol{r}_k(X_j))}\Big\} \\
&\leq \min\Big\{\delta, \frac{\mathbb{1}_{\{\sum_{j=1}^{\ell} \mathbb{1}_{\{\|\boldsymbol{r}_k(X) - \boldsymbol{r}_k(X_j)\| < h\rho\}} > 0\}}}{b\sum_{j=1}^{\ell} \mathbb{1}_{\{\|\boldsymbol{r}_k(X) - \boldsymbol{r}_k(X_j)\| < h\rho\}}}\Big\} \\
&\leq \delta + \frac{\mathbb{1}_{\{\sum_{j=1}^{\ell} \mathbb{1}_{\{\|\boldsymbol{r}_k(X) - \boldsymbol{r}_k(X_j)\| < h\rho\}} > 0\}}}{b\sum_{j=1}^{\ell} \mathbb{1}_{\{\|\boldsymbol{r}_k(X) - \boldsymbol{r}_k(X_j)\| < h\rho\}}}. \qquad (12)
\end{aligned}$$

Therefore, it is enough to show that

$$\mathbb{E}\Big[\frac{\mathbb{1}_{\{\sum_{j=1}^{\ell} \mathbb{1}_{\{\|\boldsymbol{r}_k(X) - \boldsymbol{r}_k(X_j)\| < h\rho\}} > 0\}}}{\sum_{j=1}^{\ell} \mathbb{1}_{\{\|\boldsymbol{r}_k(X) - \boldsymbol{r}_k(X_j)\| < h\rho\}}}\Big] \xrightarrow{\ell \to +\infty} 0.$$

One has

$$\begin{aligned}
&\mathbb{E}\Big[\frac{\mathbb{1}_{\{\sum_{j=1}^{\ell} \mathbb{1}_{\{\|\boldsymbol{r}_k(X) - \boldsymbol{r}_k(X_j)\| < h\rho\}} > 0\}}}{\sum_{j=1}^{\ell} \mathbb{1}_{\{\|\boldsymbol{r}_k(X) - \boldsymbol{r}_k(X_j)\| < h\rho\}}}\Big] \\
&\leq \mathbb{E}\Big[\frac{\mathbb{1}_{\{\sum_{j=1}^{\ell} \mathbb{1}_{\{\|\boldsymbol{r}_k(X) - \boldsymbol{r}_k(X_j)\| < h\rho\}} > 0\}}}{\sum_{j=1}^{\ell} \mathbb{1}_{\{\|\boldsymbol{r}_k(X) - \boldsymbol{r}_k(X_j)\| < h\rho\}}}\mathbb{1}_{\{\boldsymbol{r}_k(X) \in B\}}\Big] + \mu(\{v \in \mathbb{R}^d : \boldsymbol{r}_k(v) \in B^c\}) \\
&= \mathbb{E}\Big[\mathbb{1}_{\{\boldsymbol{r}_k(X) \in B\}}\mathbb{E}\Big[\frac{\mathbb{1}_{\{\sum_{j=1}^{\ell} \mathbb{1}_{\{\|\boldsymbol{r}_k(X) - \boldsymbol{r}_k(X_j)\| < h\rho\}} > 0\}}}{\sum_{j=1}^{\ell} \mathbb{1}_{\{\|\boldsymbol{r}_k(X) - \boldsymbol{r}_k(X_j)\| < h\rho\}}}\Big|X\Big]\Big] + \mu(\{v \in \mathbb{R}^d : \boldsymbol{r}_k(v) \in B^c\}) \\
&\leq 2\mathbb{E}\Big[\frac{\mathbb{1}_{\{\boldsymbol{r}_k(X) \in B\}}}{(\ell+1)\mu(\{v \in \mathbb{R}^d : \|\boldsymbol{r}_k(v) - \boldsymbol{r}_k(X)\| < h\rho\})}\Big] + \mu(\{v \in \mathbb{R}^d : \boldsymbol{r}_k(v) \in B^c\})
\end{aligned}$$

where B is a M-dimensional ball centered at the origin chosen so that the second term $\mu(\{v \in \mathbb{R}^d : \boldsymbol{r}_k(v) \in B^c\})$ is small. The last inequality is attained by applying part 2 of lemma 1. Moreover, as $\boldsymbol{r}_k = (\boldsymbol{r}_{k,m})_{m=1}^M$ is bounded then there exists a finite number of balls in $\mathcal{B} = \{B_M(x_j, h\rho/2) : j = 1, 2, ...\}$ such that B is contained in the union of these balls i.e., $\exists I_{h,M}$ finite, such that $B \subset \cup_{j \in I_{h,M}} B_M(x_j, h\rho/2)$.

$$
\begin{aligned}
&\mathbb{E}\Big[\frac{\mathbb{1}_{\{\boldsymbol{r}_k(X) \in B\}}}{(\ell+1)\mu(\{v \in \mathbb{R}^d : \|\boldsymbol{r}_k(v) - \boldsymbol{r}_k(X)\| < h\rho\})}\Big] \\
&\leq \sum_{j \in I_{h,M}} \int_{u:\|\boldsymbol{r}_k(u) - x_j\| < h\rho/2} \frac{\mu(du)}{(\ell+1)\mu(\{v \in \mathbb{R}^d : \|\boldsymbol{r}_k(v) - \boldsymbol{r}_k(u)\| < h\rho\})} \\
&\quad + \mu(\{v \in \mathbb{R}^d : \boldsymbol{r}_k(v) \in B^c\}) \\
&\leq \sum_{j \in I_{h,M}} \int_{u:\|\boldsymbol{r}_k(u) - x_j\| < h\rho/2} \frac{\mu(du)}{(\ell+1)\mu(\{v \in \mathbb{R}^d : \|\boldsymbol{r}_k(v) - x_j\| < h\rho/2\})} \\
&\quad + \mu(\{v \in \mathbb{R}^d : \boldsymbol{r}_k(v) \in B^c\}) \\
&= \sum_{j \in I_{h,M}} \frac{\mu(\{u \in \mathbb{R}^d : \|\boldsymbol{r}_k(u) - x_j\| < h\rho/2\})}{(\ell+1)\mu(\{v \in \mathbb{R}^d : \|\boldsymbol{r}_k(v) - x_j\| < h\rho/2\})} + \mu(\{v \in \mathbb{R}^d : \boldsymbol{r}_k(v) \in B^c\}) \\
&= \frac{|I_{h,M}|}{\ell+1} + \mu(\{v \in \mathbb{R}^d : \boldsymbol{r}_k(v) \in B^c\}) \\
&\leq \frac{C_0}{h^M(\ell+1)} + \mu(\{v \in \mathbb{R}^d : \boldsymbol{r}_k(v) \in B^c\}) \qquad (13) \\
&\xrightarrow[h^M\ell \to +\infty]{\ell \to +\infty, h \to 0} \mu(\{v \in \mathbb{R}^d : \boldsymbol{r}_k(v) \in B^c\}).
\end{aligned}
$$

It is easy to check the following fact,

$$|I_{h,M}| \leq \frac{C_0}{h^M} \text{ for some } C_0 > 0. \qquad (14)$$

To prove inequality (14), we consider again the cover $\mathcal{B} = \{B_M(x_j, h\rho/2) : j = 1, 2, ...\}$ of $\mathbb{R}^M$. For any $\rho > 0$ fixed and $h > 0$, note that the covering number $|I_{h,M}|$ is proportional to the ratio between the volume of B and the volume of the ball $B_M(0, h\rho/2)$ i.e.,

$$
\begin{aligned}
|I_{h,M}| &\propto \frac{Vol(B)}{Vol(B_M(0, h\rho/2))} \\
&\propto \frac{Vol(B)}{(h\rho/2)^M} \\
&\leq \frac{C_0}{h^M}
\end{aligned}
$$

for some positive constant C_0 proportional to the volume of B. Finally, we can conclude the proof of the proposition as we can choose B such that $\mu(\{v \in \mathbb{R}^d : \boldsymbol{r}_k(v) \in B^c\}) = 0$ using the boundedness of the basic regressors.

Remark 2 *The assumption on the boundedness of the constructed estimators is crucial. This assumption allows us to choose a ball B which can be covered using a finite number $|I_{h,M}|$ of balls $B_M(x_j, h\rho/2)$, therefore makes it possible to prove the result of this proposition for this class of regular kernels. Note that for the class of compactly supported kernels, it is easy to obtain such a result directly from the begging of the evaluation of each integral (see, for example, Chapter 5 of Györfi et al. (2002)).*

■

Proposition A.3 *Under the assumptions of **Proposition 2**,*

$$\lim_{\ell \to +\infty} \mathbb{E}\Big[\Big|g^*(\boldsymbol{r}_k(X))\Big(\sum_{i=1}^{\ell} W_{n,i}(X) - 1\Big)\Big|^2\Big] = 0.$$

***Proof of Proposition* A.3** *Note that $|\sum_{i=1}^{\ell} W_{n,i}(X) - 1| \leq 1$ thus one has*

$$\Big|g^*(\boldsymbol{r}_k(X))\Big(\sum_{i=1}^{\ell} W_{n,i}(X) - 1\Big)\Big|^2 \leq |g^*(\boldsymbol{r}_k(X))|^2.$$

Consequently, by Lebesque's dominated convergence theorem, to prove this proposition, it is enough to show that $\sum_{i=1}^{\ell} W_{n,i}(X) \to 1$ almost surely. Note that $1 - \sum_{i=1}^{\ell} W_{n,i}(X) = \mathbb{1}_{\{\sum_{i=1}^{\ell} K_h(\boldsymbol{r}_k(X) - \boldsymbol{r}_k(X_i)) = 0\}}$ therefore,

$$\begin{aligned}
\mathbb{P}\Big[\sum_{i=1}^{\ell} W_{n,i}(X) \neq 1\Big] &= \mathbb{P}\Big[\sum_{i=1}^{\ell} K_h(\boldsymbol{r}_k(X) - \boldsymbol{r}_k(X_i)) = 0\Big] \\
&\leq \mathbb{P}\Big(\sum_{j=1}^{\ell} \mathbb{1}_{\{\|\boldsymbol{r}_k(X) - \boldsymbol{r}_k(X_j)\| < h\rho\}} = 0\Big) \\
&= \int \mathbb{P}\Big(\sum_{j=1}^{\ell} \mathbb{1}_{\{\|\boldsymbol{r}_k(x) - \boldsymbol{r}_k(X_j)\| < h\rho\}} = 0\Big)\mu(dx) \\
&= \int \mathbb{P}\Big(\cap_{j=1}^{\ell} \{\|\boldsymbol{r}_k(x) - \boldsymbol{r}_k(X_j)\| \geq h\rho\}\Big)\mu(dx) \\
&= \int \Big[1 - \mathbb{P}\Big(\{\|\boldsymbol{r}_k(x) - \boldsymbol{r}_k(X_1)\| < h\rho\}\Big)\Big]^{\ell}\mu(dx)
\end{aligned}$$

$$
\begin{aligned}
&= \int \Big[1 - \mu\Big(\{v \in \mathbb{R}^d : \|\boldsymbol{r}_k(x) - \boldsymbol{r}_k(v)\| < h\rho\}\Big)\Big]^{\ell} \mu(dx) \\
&\leq \int e^{-\ell\mu(A_h(x))}\mu(dx) \\
&= \int e^{-\ell\mu(A_h(x))}\mathbb{1}_{\{\boldsymbol{r}_k(x)\in B\}}\mu(dx) + \mu(\{v \in \mathbb{R}^d : \boldsymbol{r}_k(v) \in B^c\}) \\
&\leq \frac{\max_u\{ue^{-u}\}}{\ell} \int \frac{\mathbb{1}_{\{\boldsymbol{r}_k(x)\in B\}}}{\mu(A_h(x))}\mu(dx) + \mu(\{v \in \mathbb{R}^d : \boldsymbol{r}_k(v) \in B^c\})
\end{aligned}
$$

where

$$
A_h(x) \stackrel{\text{def}}{=} \{v \in \mathbb{R}^d : \|\boldsymbol{r}_k(x) - \boldsymbol{r}_k(v)\| < h\rho\}. \tag{15}
$$

Therefore,

$$
\begin{aligned}
\mathbb{P}\Big[\sum_{i=1}^{\ell} W_{n,i}(X) \neq 1\Big] \leq \frac{e^{-1}}{\ell}\mathbb{E}\Big[\frac{\mathbb{1}_{\{\boldsymbol{r}_k(X)\in B\}}}{\mu(\{v \in \mathbb{R}^d : \|\boldsymbol{r}_k(v) - \boldsymbol{r}_k(X)\| < h\rho\})}\Big] \\
+ \mu(\{v \in \mathbb{R}^d : \boldsymbol{r}_k(v) \in B^c\}).
\end{aligned}
$$

Following the same procedure as in the proof of A.2 we obtain the desire result.

■

Proof of Theorem 1 *Choose a new observation $x \in \mathbb{R}^d$, given the training data $\mathcal{D}_k$ and the predictions $\{\boldsymbol{r}_k(X_p)\}_{p=1}^{\ell}$ on $\mathcal{D}_\ell$, taking expectation with respect to the response variables $\{Y_p^{(\ell)}\}_{p=1}^{\ell}$, it is easy to check that*

$$
\begin{aligned}
&\mathbb{E}[|g_n(\boldsymbol{r}_k(x)) - g^*(\boldsymbol{r}_k(x))|^2|\{\boldsymbol{r}_k(X_p)\}_{p=1}^{\ell}, \mathcal{D}_k] \\
&= \mathbb{E}\Big[\Big|g_n(\boldsymbol{r}_k(x)) - \mathbb{E}[g_n(\boldsymbol{r}_k(x))|\{\boldsymbol{r}_k(X_p)\}_{p=1}^{\ell}, \mathcal{D}_k] \\
&\quad + \mathbb{E}[g_n(\boldsymbol{r}_k(x))|\{\boldsymbol{r}_k(X_p)\}_{p=1}^{\ell}, \mathcal{D}_k] - g^*(\boldsymbol{r}_k(x))\Big|^2\Big|\{\boldsymbol{r}_k(X_p)\}_{p=1}^{\ell}, \mathcal{D}_k\Big] \\
&= \mathbb{E}[|g_n(\boldsymbol{r}_k(x)) - \mathbb{E}[g_n(\boldsymbol{r}_k(x))|\{\boldsymbol{r}_k(X_p)\}_{p=1}^{\ell}, \mathcal{D}_k]|^2|\{\boldsymbol{r}_k(X_p)\}_{p=1}^{\ell}, \mathcal{D}_k] \\
&\quad + |g^*(\boldsymbol{r}_k(x)) - \mathbb{E}[g_n(\boldsymbol{r}_k(x))|\{\boldsymbol{r}_k(X_p)\}_{p=1}^{\ell}, \mathcal{D}_k]|^2 \\
&\stackrel{\text{def}}{=} E_1 + E_2.
\end{aligned}
$$

On one hand by using the independence between Y_i and (Y_j, X_j) for all $i \neq j$, we develop the square and obtain for any $\delta > 0$:

$$
\begin{aligned}
E_1 &\overset{\text{def}}{=} \mathbb{E}\Big[\Big|g_n(\boldsymbol{r}_k(x)) - \mathbb{E}[g_n(\boldsymbol{r}_k(x))|\{\boldsymbol{r}_k(X_p)\}_{p=1}^{\ell}, \mathcal{D}_k]\Big|^2\Big|\{\boldsymbol{r}_k(X_p)\}_{p=1}^{\ell}, \mathcal{D}_k\Big] \\
&= \mathbb{E}\Big[\Big|\sum_{i=1}^{\ell} W_{n,i}(x)(Y_i - \mathbb{E}[Y_i|\boldsymbol{r}_k(X_i)])\Big|^2\Big|\{\boldsymbol{r}_k(X_p)\}_{p=1}^{\ell}, \mathcal{D}_k\Big] \\
&= \mathbb{E}\Big[\sum_{i=1}^{\ell} W_{n,i}^2(x)(Y_i - \mathbb{E}[Y_i|\boldsymbol{r}_k(X_i)])^2\Big|\{\boldsymbol{r}_k(X_p)\}_{p=1}^{\ell}, \mathcal{D}_k\Big] \\
&= \sum_{i=1}^{\ell} W_{n,i}^2(x)\mathbb{E}_{Y_i}[(Y_i - \mathbb{E}[Y_i|\boldsymbol{r}_k(X_i)])^2|\boldsymbol{r}_k(X_i)] \\
&= \mathbb{V}[Y_1|\boldsymbol{r}_k(X_1)]\sum_{i=1}^{\ell} W_{n,i}^2(x) \\
&\overset{(12)}{\leq} \frac{4R^2}{b}\Big(\delta + \frac{\mathbb{1}_{\{\sum_{j=1}^{\ell}\mathbb{1}_{\{\|\boldsymbol{r}_k(x)-\boldsymbol{r}_k(X_j)\|<h\rho\}}>0\}}}{\sum_{j=1}^{\ell}\mathbb{1}_{\{\|\boldsymbol{r}_k(x)-\boldsymbol{r}_k(X_j)\|<h\rho\}}}\Big)
\end{aligned}
$$

where the notation $\mathbb{V}(Z)$ stands for the variance of a random variable Z. Therefore, using the result of inequality (13), one has

$$
\mathbb{E}(E_1) \leq \frac{4R^2}{b}\Big(\delta + \frac{C_0}{h^M(\ell+1)}\Big) \tag{16}
$$

for some $C_0 > 0$. On the other hand, set

- $C_h^{\ell}(x) \overset{\text{def}}{=} \sum_{j=1}^{\ell}\mathbb{1}_{\{\|\boldsymbol{r}_k(X_j)-\boldsymbol{r}_k(x)\|<h\rho\}}$.
- $D_h^{\ell}(x) \overset{\text{def}}{=} \sum_{j=1}^{\ell} K_h(r_k(X_j) - \boldsymbol{r}_k(x))$.

The second term E_2 is much harder to control as it depends on $g^*(\boldsymbol{r}_k(.))$, that is why a weak smoothness assumption of the theorem is made. Using this assumption and Jensen's inequality (Jensen (1906)), one has

$$
\begin{aligned}
E_2 &\overset{\text{def}}{=} \Big|g^*(\boldsymbol{r}_k(x)) - \mathbb{E}[g_n(\boldsymbol{r}_k(x))|\{\boldsymbol{r}_k(X_p)\}_{p=1}^{\ell}, \mathcal{D}_k]\Big|^2 \\
&= \Big(\sum_{i=1}^{\ell} W_{n,i}(X)(g^*(\boldsymbol{r}_k(x)) - \mathbb{E}[Y_i|\boldsymbol{r}_k(X_i)])\Big)^2\mathbb{1}_{\{D_h^{\ell}(x)>0\}} + (g^*(\boldsymbol{r}_k(x)))^2\mathbb{1}_{\{D_h^{\ell}(x)=0\}} \\
&\overset{(Jensen)}{\leq} \sum_{i=1}^{\ell} W_{n,i}(x)(g^*(\boldsymbol{r}_k(x)) - \mathbb{E}[Y_i|\boldsymbol{r}_k(X_i)])^2\mathbb{1}_{\{D_h^{\ell}(x)>0\}} + (g^*(\boldsymbol{r}_k(x)))^2\mathbb{1}_{\{D_h^{\ell}(x)=0\}}
\end{aligned}
$$

$$
\begin{aligned}
&\leq \sum_{i=1}^{\ell} \frac{K_h(\boldsymbol{r}_k(x)-\boldsymbol{r}_k(X_i))(g^*(\boldsymbol{r}_k(x))-g^*(\boldsymbol{r}_k(X_i)))^2}{\sum_{j=1}^{\ell} K_h(\boldsymbol{r}_k(x)-\boldsymbol{r}_k(X_j))}\mathbb{1}_{\{D_h^\ell(x)>0\}} + (g^*(\boldsymbol{r}_k(x)))^2\mathbb{1}_{\{D_h^\ell(x)=0\}} \\
&\leq L^2\sum_{i=1}^{\ell} \frac{K_h(\boldsymbol{r}_k(x)-\boldsymbol{r}_k(X_i))\|\boldsymbol{r}_k(x)-\boldsymbol{r}_k(X_i)\|^2}{\sum_{j=1}^{\ell} K_h(\boldsymbol{r}_k(x)-\boldsymbol{r}_k(X_j))}\mathbb{1}_{\{D_h^\ell(x)>0\}} + (g^*(\boldsymbol{r}_k(x)))^2\mathbb{1}_{\{D_h^\ell(x)=0\}} \\
&\leq L^2\Big[\sum_{i=1}^{\ell} \frac{K_h(\boldsymbol{r}_k(x)-\boldsymbol{r}_k(X_i))\|\boldsymbol{r}_k(x)-\boldsymbol{r}_k(X_i)\|^2\mathbb{1}_{\{\|\boldsymbol{r}_k(x)-\boldsymbol{r}_k(X_i)\|<R_K h^\beta\}}}{\sum_{j=1}^{\ell} K_h(\boldsymbol{r}_k(x)-\boldsymbol{r}_k(X_j))} \\
&\quad + \sum_{i=1}^{\ell} \frac{K_h(\boldsymbol{r}_k(x)-\boldsymbol{r}_k(X_i))\|\boldsymbol{r}_k(x)-\boldsymbol{r}_k(X_i)\|^2\mathbb{1}_{\{\|\boldsymbol{r}_k(x)-\boldsymbol{r}_k(X_i)\|\geq R_K h^\beta\}}}{\sum_{j=1}^{\ell} K_h(\boldsymbol{r}_k(x)-\boldsymbol{r}_k(X_j))}\Big]\mathbb{1}_{\{D_h^\ell(x)>0\}} \\
&\quad + (g^*(\boldsymbol{r}_k(x)))^2\mathbb{1}_{\{C_h^\ell(x)=0\}} \\
&\overset{\text{def}}{=} E_2^1 + E_2^2 + E_2^3.
\end{aligned}
$$

for any $\beta > 0$ chosen arbitrarily at this point. Now, we bound the expectation of the three terms of the last inequality.

- *Firstly, E_2^1 can be easily bounded from above by*

$$
\begin{aligned}
E_2^1 &= L^2\sum_{i=1}^{\ell} \frac{K_h(\boldsymbol{r}_k(x)-\boldsymbol{r}_k(X_i))\|\boldsymbol{r}_k(x)-\boldsymbol{r}_k(X_i)\|^2}{\sum_{j=1}^{\ell} K_h(\boldsymbol{r}_k(x)-\boldsymbol{r}_k(X_j))}\mathbb{1}_{\{D_h^\ell(x)>0\}}\mathbb{1}_{\{\|\boldsymbol{r}_k(x)-\boldsymbol{r}_k(X_i)\|<R_K h^\beta\}} \\
&\leq L^2 h^{2\beta} R_K^2 \sum_{i=1}^{\ell} \frac{K_h(\boldsymbol{r}_k(x)-\boldsymbol{r}_k(X_i))}{\sum_{j=1}^{\ell} K_h(\boldsymbol{r}_k(x)-\boldsymbol{r}_k(X_j))}\mathbb{1}_{\{D_h^\ell(x)>0\}} \\
&= L^2 h^{2\beta} R_K^2.
\end{aligned}
$$

  Therefore, its expectation is simply bounded by the same upper bound i.e.,

$$
\mathbb{E}(E_2^1) \leq L^2 h^{2\beta} R_K^2 \tag{17}
$$

- *Secondly, we bound the second term E_2^2 using the tail assumption of the kernel K given equation (7), thus for any $h > 0$:*

$$
\begin{aligned}
E_2^2 &= L^2\sum_{i=1}^{\ell} \frac{K_h(\boldsymbol{r}_k(x)-\boldsymbol{r}_k(X_i))\|\boldsymbol{r}_k(x)-\boldsymbol{r}_k(X_i)\|^2\mathbb{1}_{\{D_h^\ell(x)>0\}}}{\sum_{j=1}^{\ell} K_h(\boldsymbol{r}_k(x)-\boldsymbol{r}_k(X_j))}\mathbb{1}_{\{\|\boldsymbol{r}_k(x)-\boldsymbol{r}_k(X_i)\|\geq h^\beta R_K\}} \\
&\leq L^2 h^2 \sum_{i=1}^{\ell} \frac{K_h(\boldsymbol{r}_k(x)-\boldsymbol{r}_k(X_i))\|(\boldsymbol{r}_k(x)-\boldsymbol{r}_k(X_i))/h\|^2\mathbb{1}_{\{D_h^\ell(x)>0\}}}{\sum_{j=1}^{\ell} K_h(\boldsymbol{r}_k(x)-\boldsymbol{r}_k(X_j))}\times \\
&\quad \mathbb{1}_{\{(\|\boldsymbol{r}_k(x)-\boldsymbol{r}_k(X_i))/h\|\geq R_K/h^{1-\beta}\}}
\end{aligned}
$$

$$\leq \frac{h^2L^2}{b}\sum_{i=1}^{\ell}\frac{C_K e^{-\|(\boldsymbol{r}_k(x)-\boldsymbol{r}_k(X_i))/h\|^\alpha}\|(\boldsymbol{r}_k(x)-\boldsymbol{r}_k(X_i))/h\|^2}{\sum_{j=1}^{\ell}\mathbb{1}_{\{\|\boldsymbol{r}_k(x)-\boldsymbol{r}_k(X_j)\|<h\rho\}}}\times$$
$$\mathbb{1}_{\{\|(\boldsymbol{r}_k(x)-\boldsymbol{r}_k(X_i))/h\|\geq R_K/h^{1-\beta}\}}\mathbb{1}_{\{C_h^\ell(x)>0\}}.$$

As for any $\alpha > 0$, $t \mapsto \lambda(t) = t^2e^{-t^\alpha}$ is strictly decreasing for all $t \geq (2/\alpha)^{1/\alpha}$. Thus, for $h > 0$ small enough such that $R_K/h^{1-\beta} \geq (2/\alpha)^{1/\alpha}$, one has

$$E_2^2 \leq \frac{h^2L^2C_K}{b}\sum_{i=1}^{\ell}\frac{(R_K/h^{1-\beta})^2e^{-(R_K/h^{1-\beta})^\alpha}\mathbb{1}_{\{\|(\boldsymbol{r}_k(x)-\boldsymbol{r}_k(X_i))/h\|\geq R_K/h^{1-\beta}\}}}{\sum_{j=1}^{\ell}\mathbb{1}_{\{\|\boldsymbol{r}_k(x)-\boldsymbol{r}_k(X_j)\|<h\rho\}}}\mathbb{1}_{\{C_h^\ell(x)>0\}}$$
$$\leq \frac{h^{2\beta}L^2C_KR_K^2e^{-R_K^\alpha h^{-\alpha(1-\beta)}}}{b}\sum_{i=1}^{\ell}\frac{\mathbb{1}_{\{\sum_{j=1}^{\ell}\mathbb{1}_{\{\|\boldsymbol{r}_k(x)-\boldsymbol{r}_k(X_j)\|<h\rho\}}>0\}}}{\sum_{j=1}^{\ell}\mathbb{1}_{\{\|\boldsymbol{r}_k(x)-\boldsymbol{r}_k(X_j)\|<h\rho\}}}$$
$$\leq \frac{\ell h^{2\beta}L^2C_KR_K^2e^{-R_K^\alpha h^{-\alpha(1-\beta)}}}{b}\times\frac{\mathbb{1}_{\{\sum_{j=1}^{\ell}\mathbb{1}_{\{\|\boldsymbol{r}_k(x)-\boldsymbol{r}_k(X_j)\|<h\rho\}}>0\}}}{\sum_{j=1}^{\ell}\mathbb{1}_{\{\|\boldsymbol{r}_k(x)-\boldsymbol{r}_k(X_j)\|<h\rho\}}}.$$

Applying the result of inequality (13), one has

$$\mathbb{E}(E_2^2) \leq \frac{\ell h^{2\beta}L^2C_KR_K^2e^{-R_K^\alpha h^{-\alpha(1-\beta)}}}{b}\times\frac{C_0}{h^M(\ell+1)}$$
$$\leq C_1h^{2\beta-M}e^{-R_K^\alpha h^{-\alpha(1-\beta)}} \tag{18}$$

for some $C_1 > 0$.

- *Lastly with $A_h(x)$ defined in (15), we bound the expectation of E_2^3 by,*

$$\mathbb{E}(E_2^3) \leq \mathbb{E}\Big[(g^*(\boldsymbol{r}_k(x)))^2\mathbb{1}_{\{C_h^\ell(x)=0\}}\Big]$$
$$\leq \sup_{u\in\mathbb{R}^d}(g^*(\boldsymbol{r}_k(u)))^2\mathbb{E}\Big[\mathbb{1}_{\{C_h^\ell(x)=0\}}\Big]$$
$$= \sup_{u\in\mathbb{R}^d}(g^*(\boldsymbol{r}_k(u)))^2(1-\mu(A_h(x)))^\ell$$
$$\leq \sup_{u\in\mathbb{R}^d}(g^*(\boldsymbol{r}_k(u)))^2e^{-\ell\mu(A_h(x))}$$
$$\leq \sup_{u\in\mathbb{R}^d}(g^*(\boldsymbol{r}_k(u)))^2\frac{\ell\mu(A_h(x))e^{-\ell\mu(A_h(x))}}{\ell\mu(A_h(x))}$$
$$\leq \sup_{u\in\mathbb{R}^d}(g^*(\boldsymbol{r}_k(u)))^2\frac{\max_{u\in\mathbb{R}^d}ue^{-u}}{\ell\mu(A_h(x))}$$

$$\leq \sup_{u\in\mathbb{R}^d}(g^*(\boldsymbol{r}_k(u)))^2\frac{e^{-1}}{\ell\mu(A_h(x))}$$
$$\leq \frac{C_2}{\ell\mu(A_h(x)))} \tag{19}$$

for some $C_2 > 0$.

From (16), (17), (18) *and* (19), *one has*

$$\begin{aligned}\mathbb{E}[|g_n(\boldsymbol{r}_k(X)) - g^*(\boldsymbol{r}_k(X))|^2] &\leq \int_{\mathbb{R}^d}\mathbb{E}[|g_n(\boldsymbol{r}_k(x)) - g^*(\boldsymbol{r}_k(x))|^2]\mu(dx)\\ &\leq \int_{\mathbb{R}^d}\mathbb{E}(E_1 + E_2^1 + E_2^2 + E_2^3)\mu(dx)\\ &\leq \int_{\mathbb{R}^d}\Big[\frac{4R^2}{b}\Big(\delta + \frac{C_0}{h^M(\ell+1)}\Big) + L^2h^{2\beta}R_K^2\\ &\quad + C_1h^{2\beta-M}e^{-R_K^\alpha h^{-\alpha(1-\beta)}} + \frac{C_2}{\ell\mu(A_h(x)))}\Big]\mu(dx).\end{aligned}$$

Therefore, by following the same procedure of proving inequality (13), one has

$$\begin{aligned}&\mathbb{E}[|g_n(\boldsymbol{r}_k(X)) - g^*(\boldsymbol{r}_k(X))|^2]\\ &\leq \frac{4R^2}{b}\Big(\delta + \frac{C_0}{h^M(\ell+1)}\Big) + L^2h^{2\beta}R_K^2 + C_1h^{2\beta-M}e^{-R_K^\alpha h^{-\alpha(1-\beta)}} + \int_{\mathbb{R}^d}\frac{C_2\mu(dx)}{\ell\mu(A_h(x)))}\\ &\leq \frac{4R^2}{b}\Big(\delta + \frac{C_0}{h^M(\ell+1)}\Big) + L^2h^{2\beta}R_K^2 + C_1h^{2\beta-M}e^{-R_K^\alpha h^{-\alpha(1-\beta)}}\\ &\quad + \sum_{j\in J_{h,M}}\int_{\|\boldsymbol{r}_k(x)-x_j\|<h\rho}\frac{C_2\mu(dx)}{\ell\mu(\{v\in\mathbb{R}^d : \|\boldsymbol{r}_k(v) - \boldsymbol{r}_k(x)\| < h\rho\})}\\ &\leq \frac{4R^2}{b}\Big(\delta + \frac{C_0}{h^M(\ell+1)}\Big) + L^2h^{2\beta}R_K^2 + C_1h^{2\beta-M}e^{-R_K^\alpha h^{-\alpha(1-\beta)}}\\ &\quad + \sum_{j\in J_{h,M}}\int_{\|\boldsymbol{r}_k(x)-x_j\|<h\rho}\frac{C_2\mu(dx)}{\ell\mu(\{v\in\mathbb{R}^d : \|\boldsymbol{r}_k(v) - x_j\| < h\rho\})}\\ &\leq \frac{4R^2}{b}\Big(\delta + \frac{C_0}{h^M(\ell+1)}\Big) + L^2h^{2\beta}R_K^2 + C_1h^{2\beta-M}e^{-R_K^\alpha h^{-\alpha(1-\beta)}}\\ &\quad + \frac{C_2}{\ell}\sum_{j\in J_{h,M}}\frac{\mu(\{v\in\mathbb{R}^d : \|\boldsymbol{r}_k(v) - x_j\| < h\rho\})}{\mu(\{v\in\mathbb{R}^d : \|\boldsymbol{r}_k(v) - x_j\| < h\rho\})}\\ &\leq \frac{4R^2}{b}\Big(\delta + \frac{C_0}{h^M(\ell+1)}\Big) + L^2h^{2\beta}R_K^2 + C_1h^{2\beta-M}e^{-R_K^\alpha h^{-\alpha(1-\beta)}} + \frac{C_2|J_{h,M}|}{\ell}\\ &\leq \frac{4R^2}{b}\Big(\delta + \frac{C_0}{h^M(\ell+1)}\Big) + L^2R_K^2h^{2\beta} + C_1h^{2\beta-M}e^{-R_K^\alpha h^{-\alpha(1-\beta)}} + \frac{C_2'}{h^M\ell}\end{aligned}$$

where $|J_{h,M}|$ denotes the number of balls covering the ball B (introduced in the proof of A.2) by the cover $\{B_M(x_j, h\rho) : j = 1, 2, ...\}$. Similarly, one has $|J_{h,M}| \leq \frac{C_0}{h^M}$ for some constant $C_0 > 0$ proportional to the volume of B. Since $\delta > 0$ is chosen arbitrarily and the third term of the last inequality decreases exponentially fast when $h \to 0$ for any $\beta \in (0, 1)$, hence, it is negligible comparing to other terms. Finally, with the choice of $h \propto \ell^{-1/(M+2\beta)}$, one has

$$\mathbb{E}[|g_n(\boldsymbol{r}_k(X)) - g^*(\boldsymbol{r}_k(X))|^2] \leq \frac{\tilde{C}_1}{h^M \ell} + \tilde{C}_2 h^{2\beta} \leq C\ell^{-2\beta/(M+2\beta)}.$$

for some $C > 0$ independent of ℓ and for any positive $\beta < 1$ chosen arbitrarily. Thus, by letting $\beta \to 1$, we obtain the desire result:

$$\mathbb{E}[|g_n(\boldsymbol{r}_k(X)) - g^*(\boldsymbol{r}_k(X))|^2] \leq C\ell^{-2/(M+2)}.$$

■

Acknowledgments

The author gratefully acknowledges the support of Prof. Aurélie Fischer and Prof. Mathilde Mougeot for valuable feedback and suggestions during the process of writing this article.

References

Audibert, J.Y., 2004. Aggregated estimators and empirical complexity for least square regression. Annales de l'Institut Henri Poincaré (B) Probabilités et Statistique 40, 685–736. doi:10.1016/j.anihpb.2003.11.006.

Biau, G., Fischer, A., Guedj, B., Malley, J.D., 2016. COBRA: a combined regression strategy. Journal of Multivariate Analysis 146, 18–28. doi:10.1016/j.jmva.2015.04.007.

Borchers, H.W., 2019. pracma: Practical numerical math functions.

Breiman, L., 1995. Stacked regression. Machine Learning 24, 49–64. doi:10.1007/BF00117832.

Brian, R., Bill, V., Douglas, M.B., Kurt, H., Albrecht, G., David, F., 2021. Mass: Support functions and datasets for venables and ripley's mass. URL: https://CRAN.R-project.org/package=MASS.

Bunea, F., Tsybakov, A.B., Wegkamp, M.H., 2006. Aggregation and sparsity via ℓ_1-penalized least squares, in: Lugosi, G., Simon, H.U. (Eds.), Proceedings of 19th Annual Conference on Learning Theory (COLT 2006), Lecture Notes in Artificial Intelligence, Springer-Verlag, Berlin-Heidelberg. pp. 379–391.

Bunea, F., Tsybakov, A.B., Wegkamp, M.H., 2007a. Aggregation for gaussian regression. The Annals of Statistics 35, 1674–1697.

Bunea, F., Tsybakov, A.B., Wegkamp, M.H., 2007b. Sparsity oracle inequalities for the lasso. Electronic Journal of Statistics 35, 169–194. doi:10.1214/07-EJS008.

Cadet, O., Harper, C., Mougeot, M., 2005. Monitoring energy performance of compressors with an innovative auto-adaptive approach., in: Instrumentation System and Automation -ISA- Chicago.

Catoni, O., 2004. Statistical Learning Theory and Stochastic Optimization. Lectures on Probability Theory and Statistics, Ecole d'Eté de Probabilités de Saint-Flour XXXI - 2001, Lecture Notes in Mathematics, Springer.

Chen, T., Guestrin, C., 2016. Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, New York, NY, USA. p. 785–794. URL: https://doi.org/10.1145/2939672.2939785, doi:10.1145/2939672.2939785.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., Li, Y., contributors, X., 2021. xgboost: Extreme gradient boosting. URL: https://CRAN.R-project.org/package=xgboost.

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., Reis, J., 2009. Modeling wine preferences by data mining from physicochemical properties. Decision Support Systems, Elsevier 47, 547–553. doi:10.1016/j.dss.2009.05.016.

Dalalyan, A., Tsybakov, A.B., 2008. Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity. Machine Learning 72, 39–61. doi:10.1007/s10994-008-5051-0.

Devroye, L., Györfi, L., Lugosi, G., 1997. A Probabilistic Theory of Pattern Recognition. Springer.

Devroye, L., Krzyżak, A., 1989. An equivalence theorem for l1 convergence of the kernel regression estimate. Journal of Statistical Planning and Inference 23, 71–82. doi:10.1016/0378-3758(89)90040-2.

Dua, D., Graff, C., 2017a. UCI machine learning repository: Abalone data set.

Dua, D., Graff, C., 2017b. UCI machine learning repository: Wine quality data set.

Fischer, A., Montuelle, L., Mougeot, M., Picard, D., 2017. Statistical learning for wind power: A modeling and stability study towards forecasting. Wiley Online Library 20, 2037–2047. doi:10.1002/we.2139.

Fischer, A., Mougeot, M., 2019. Aggregation using input-output trade-off. Journal of Statistical Planning and Inference 200, 1–19. doi:10.1016/j.jspi.2018.08.001.

Folland, G.B., 1999. Real Analysis: Modern Techniques and Their Applications. John Willey & Sonsy, Inc., New York.

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software 33, 1–22. doi:10.18637/jss.v033.i01.

Gionis, A., Mannila, H., Tsaparas, P., 2005. Clustering aggregation, in: 21st International Conference on Data Engineering (ICDE'05), pp. 341–352. doi:10.1109/ICDE.2005.34.

Guedj, B., 2013. COBRA: Nonlinear Aggregation of Predictors. R package version 0.99.4.

Guedj, B., Rengot, J., 2020. Non-linear aggregation of filters to improve image denoising, in: Arai, K., Kapoor, S., Bhatia, R. (Eds.), Intelligent Computing, Springer International Publishing, Cham. pp. 314–327.

Guedj, B., Srinivasa Desikan, B., 2018. Pycobra: A python toolbox for ensemble learning and visualisation. Journal of Machine Learning Research 18, 1–5.

Guedj, B., Srinivasa Desikan, B., 2020. Kernel-based ensemble learning in python. Information 11, 63. doi:10.3390/info11020063.

Györfi, L., Kohler, M., Krzyżak, A., Walk, H., 2002. A Distribution-Free Theory of Nonparametric Regression. Springer.

Has, S., 2022. Consensual aggregation on random projected high-dimensional features for regression. URL: https://hal.archives-ouvertes.fr/hal-03631715. preprint.

Has, S., Fischer, A., Mougeot, M., 2021. Kfc: A clusterwise supervised learning procedure based on the aggregation of distances. Journal of Statistical Computation and Simulation 91, 2307–2327. doi:10.1080/00949655.2021.1891539.

Jensen, J.L.W.V., 1906. Sur les fonctions convexes et les inégalités entre les valeurs moyennes. Acta mathematica 30, 175–193. doi:10.1007/BF02418571.

Juditsky, A., Nemirovski, A., 2000. Functional aggregation for nonparametric estimation. The Annals of Statistics 28, 681–712. doi:10.1214/aos/1015951994.

Kaggle, 2016. House sales in king county, usa.

Kluth, G., Ripoll, J.F., Has, S., Fischer, A., Mougeot, M., Camporeale, E., 2022. Machine learning methods applied to the global modeling of event-driven pitch angle diffusion coefficients during high speed streams. Frontiers in Physics 10. URL: https://www.frontiersin.org/article/10.3389/fphy.2022.786639, doi:10.3389/fphy.2022.786639.

Li, S., 2019. Fnn: Fast nearest neighbor search algorithms and applications.

Liaw, A., Wiener, M., 2002. Classification and regression by randomforest. R News 2, 18–22.

Massart, P., 2007. Concentration Inequalities and Model Selection. École d'Été de Probabilités de Saint-Flour XXXIII – 2003, Lecture Notes in Mathematics, Springer, Berlin, Heidelberg.

Mojirsheibani, M., 1999. Combined classifiers via disretization. Journal of the American Statistical Association 94, 600–609. doi:10.1080/01621459.1999.10474154.

Mojirsheibani, M., 2000. A kernel-based combined classification rule. Journal of Statistics and Probability Letters 48, 411–419. doi:10.1016/S0167-7152(00)00024-9.

Mojirsheibani, M., Kong, J., 2016. An asymptotically optimal kernel combined classifier. Journal of Statistics and Probability Letters 119, 91–100. doi:10.1016/j.spl.2016.07.017.

Nemirovski, A., 2000. Topics in Non-Parametric Statistics. École d'Été de Probabilités de Saint-Flour XXVIII – 1998, Springer.

Ripley, B., 2019. tree: Classification and regression trees.

Ripoll, J.F., Loridan, V., Denton, M.H., Cunningham, G., Reeves, G., Santolík, O., Fennell, J., Turner, D.L., Drozdov, A.Y., Cervantes Villa, J.S., Shprits, Y.Y., Thaller, S.A., Kurth, W.S., Kletzing, C.A., Henderson, M.G., Ukhorskiy, A.Y., 2019. Observations and fokker-planck simulations of the l-shell, energy, and pitch angle structure of earth's electron radiation belts during quiet times. Journal of Geophysical Research: Space Physics 124, 1125–1142. doi:https://doi.org/10.1029/2018JA026111.

Stone, C.J., 1977. Consistent nonparametric regression. Ann. Statist. 5, 595–620. doi:10.1214/aos/1176343886.

Thorne, R.M., Li, W., Ni, B., Ma, Q., Bortnik, J., Chen, L., Baker, D.N., Spence, H.E., Reeves, G.D., Henderson, M.G., Kletzing, C.A., Kurth, W.S., Hospodarsky, G.B., Blake, J.B., Fennell, J.F., Claudepierre, S.G., Kanekal, S.G., 2013. Rapid local acceleration of relativistic radiation-belt electrons by magnetospheric chorus. Nature 504, 411–414. URL: https://doi.org/10.1038/nature12889, doi:10.1038/nature12889.

Wegkamp, M.H., 2003. Model selection in nonparametric regression. The Annals of Statistics 31, 252–273. doi:10.1214/aos/1046294464.

Wu, O., Hu, W., Maybank, S.J., Zhu, M., Li, B., 2012. Efficient clustering aggregation based on data fragments. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 42, 913–926. doi:10.1109/TSMCB.2012.2183591.

Yang, Y., 2000. Combining different procedures for adaptive regression. Journal of multivariate analysis 74, 135–161. doi:10.1006/jmva.1999.1884.

Yang, Y., 2001. Adaptive regression by mixing. Journal of the American Statistical Association 96, 574–588. doi:10.1198/016214501753168262.

Yang, Y., 2004. Aggregating regression procedures to improve performance. Bernoulli 10, 25–47. doi:10.3150/bj/1077544602.