HAL open science

A Kernel-based Consensual Aggregation for Regression

Sothea Has

To cite this version:

Sothea Has. A Kernel-based Consensual Aggregation for Regression. 2021. hal-02884333v5

HAL Id: hal-02884333

https://hal.science/hal-02884333v5

Preprint submitted on 27 Apr 2021 (v5), last revised 13 Jun 2023 (v6)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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 $r_1, r_2, ..., r_M$ using a weighted average where the weights are defined based on some kernel function to build a target prediction. 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. Moreover, an optimization method based on gradient descent algorithm is proposed to efficiently and rapidly estimate the key parameter of the strategy. The numerical experiments carried out on several simulated and real datasets are also provided to illustrate the speed-up of gradient descent algorithm in estimating the key parameter and the improvement of overall performance of the method with the introduction of smoother kernel functions.

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 bunch 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).

Another approach of model selection, which aims at selecting the best estimator among the candidate estimators, has also been proposed (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 may be able to capture different features of the input data and 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)). Moreover, it has also been applied in filtering to improve the image denoising (see Guedj and Rengot (2020)). In a complementary manner to the earlier works, we present another kernel-based consensual regression aggregation method in this paper, 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, an evidence of numerical simulation carried out on a similar set of simulated models, and some real datasets shows that the present method outperforms

the classical one.

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 term of the method for a subclass of regular kernel functions. A method based on gradient descent algorithm to estimate the bandwidth parameter is described in Section 3. Section 4 illustrates the performances of the proposed method through several numerical examples of simulated and real datasets. Lastly, Section 6 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 $\ell + k = n$, which 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 or machines $r_{k,1}, r_{k,2}, ..., r_{k,M}$ are constructed using only the data points in $\mathcal{D}_k$. These basic machines 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 machines 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 $\mathbf{r}_k(X)$.

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 machines $r_{k,1}, r_{k,2}, ..., r_{k,M}$, 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 (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 is based only on $\mathcal{D}_\ell$ but the whole construction of the method depends on the whole training data $\mathcal{D}_n$ as the basic machines 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 $\mathbf{r}_k(x)$ separately.

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 $g_n(\boldsymbol{r}_k(x))$ be the combined estimator defined in* (1) *with the weights given in* (5) *computed at point $x \in \mathbb{R}^d$. Then, for all distributions of (X, Y) with $\mathbb{E}[|Y|^2] < +\infty$,*

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

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,

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

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 machines, the first term cannot be controlled as it depends on the performance of the best constructed machine, and it will be there as the asymptotic control of the 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 machines. 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 machines 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. Our goal is not to theoretically do better than the classical method but to investigate such a similar result in a broader class of kernel functions. For those kernels which the tails decrease fast enough, the convergence rate of the variance-type term can be attained as described in the following main theorem of this paper.

Theorem 1 *Assume that the response variable Y and all the basic machines $r_{k,m}, m = 1, 2, ..., M$, 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

$$\exists R_K, C_k > 0 : K(z)\|z\|^2 \leq \frac{C_K}{1 + \|z\|^M}, \forall z \in \mathbb{R}^M \text{ such that } \|z\| \geq R_K. \quad (7)$$

Then, with the choice of $h \propto \ell^{-\frac{M+2}{M^2+2M+4}}$, one has

$$\mathbb{E}[|g_n(\boldsymbol{r}_k(X)) - g^*(X)|^2] \leq \min_{1 \leq m \leq M} \mathbb{E}[|r_{k,m}(X) - g^*(X)|^2] + C\ell^{-\frac{4}{M^2+2M+4}} \quad (8)$$

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

Moreover, if there exists a consistent estimator named r_{k,m_0} among $\{r_{k,m}\}_{m=1}^M$ i.e.,

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

then the combing 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

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

Remark 1 *The assumption on the upper bound of the kernel K in the theorem above is very weak, chosen so that the result holds for a large subclass of regular kernels. However, the convergence rate is indeed slow for this subclass*

of kernel functions. If we strengthen this condition, we can obtain a much nicer result. For instance, if we assume that the tails decrease at least of exponential speed i.e.,

$$\exists R_K, C_K > 0 \text{ and } \alpha \in (0,1) : K(z) \leq C_K e^{-\|z\|^\alpha}, \forall z \in \mathbb{R}^M, \|z\| \geq R_K,$$

by following the same procedure as in the proof of the above theorem (Section 6), one can easily check that the convergence rate of the variance-type term is of order $O(\ell^{-2\alpha/(M+2\alpha)})$*. This rate approaches the state of the art of the classical method by Biau et al. (2016) when* α *approaches* 1.

3 Bandwidth parameter estimation thanks to 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 parts $\mathcal{D}_k$ where all the candidate machines $\{\mathbf{r}_{k,m}\}_{m=1}^M$ are built, and two other parts $\mathcal{D}_{\ell_1}$ and $\mathcal{D}_{\ell_2}$. $\mathcal{D}_{\ell_1}$ is used for the combination defined in equation (1), and $\mathcal{D}_{\ell_2}$ is the validation set used to learn the bandwidth parameter h of equation (2) and the proportion α of equation (3) by minimizing the average quadratic error evaluated on $\mathcal{D}_{\ell_2}$ defined as follows,

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

where $|\mathcal{D}_{\ell_2}|$ denotes the cardinality of $\mathcal{D}_{\ell_2}$, $g_n(\mathbf{r}_k(X_j)) = \sum_{(X_i,Y_i)\in\mathcal{D}_{\ell_1}} W_{n,i}(X_j)Y_i$ defined in equation (1), and the weight $W_{n,i}(X_j)$ is given in equation (2) and (6) for Biau et al. (2016) and Guedj and Srinivasa Desikan (2020) respectively. 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 and X_j of $\mathcal{D}_{\ell_1}$ and $\mathcal{D}_{\ell_2}$ respectively. In this case, constructing a combining estimation g_n is equivalent to searching for an optimal parameter h^* over a given grid $\mathcal{H} = \{h_{\min}, ..., h_{\max}\}$ i.e.,

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

The parameter α of equation (3) can be tuned easily by considering $\varphi_{\alpha M}(h)$ where $\alpha \in \{1/2, 1/3, ..., 1\}$ 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.,

$$(\alpha^*, h^*) = \underset{(\alpha,h)\in\{1/2,1/3,\ldots,1\}\times\mathcal{H}}{\operatorname{argmin}} \varphi_{\alpha M}(h).$$

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 machines using $\mathcal{D}_k$, and 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. We often observe the convex-like curves of the cross-validation quadratic error on many simulations; and from this observation, we propose to use 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:

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

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

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

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

In our numerical experiment, the numerical gradient of (10) can be computed efficiently and rapidly thanks to `grad` function contained in `pracma` library of `R` program (see Borchers (2019)). We observe that the algorithm works much faster, and more importantly it does not require the information of the interval containing the optimal parameter as the grid search does. Most of the time, the parameter h vanishing the numerical gradient of the objective function can be attained, leading to a good construction of the corresponding combining estimation 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 machines of the combination. In this experiment, we compare the performances of the proposed methods with the classical one and all the basic machines. 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.

Kernel	**Formula**
Naive[1]	$K(x) = \prod_{i=1}^{d} \mathbb{1}_{\{\lvert x_i \rvert \leq 1\}}$
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.

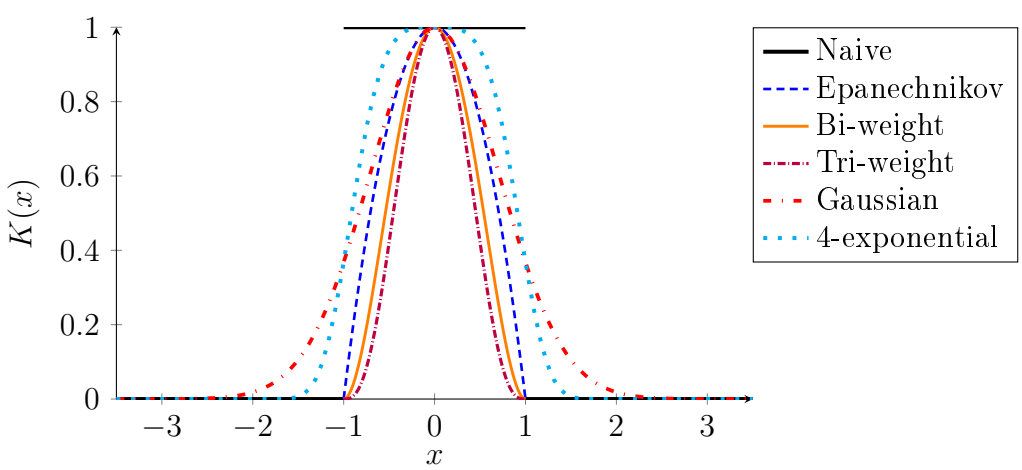

Figure 1: The shapes of some kernels.

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

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

Model 6 : $n = 500, d = 30, Y = \sum_{k=1}^{10} \mathbb{1}_{\{X_k<0\}} - \mathbb{1}_{\{\mathcal{N}(0,1)>1.25\}}$.

Model 7 : $n = 600, d = 300, Y = X_1^2 + X_2^2X_3\exp(-|X_4|) + X_6 - X_8 + \mathcal{N}(0,0.5)$.

Model 8 : $n = 600, d = 50, Y = \mathbb{1}_{\{X_1+X_4^3+X_9+\sin(X_{12}X_{18})+\mathcal{N}(0,0.01)>0.38\}}$.

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 the last one is not from Biau et al. (2016) but a made-up one.

Model 9 : $n = 500, d = 1000, Y = X_1 + 3X_3^2 - 2\exp(-X_5) + X_6$.

Model 10 : $n = 500, d = 1500, Y = \exp(X_1) + \exp(-X_1) + \sum_{j=2}^{d}[\cos(X_j^j)) - 2\sin(X_j^j) - \exp(-|X_j|)]$.

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% will be treated as the corresponding testing data. We measure the performance of any regression method f using *mean square error* (MSE) evaluated on the 20%-testing data defined by

$$\mathrm{MSE}(f) = \frac{1}{n_{\mathrm{test}}} \sum_{i=1}^{n_{\mathrm{test}}} (y_i^{\mathrm{test}} - f(x_i^{\mathrm{test}}))^2. \tag{11}$$

Table 2 and 3 below contain the average MSEs 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 machines $\mathbf{r}_k = (r_{k,m})_{m=1}^5$:

- **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 seven columns corresponding to the kernel functions used in the combining method where **COBRA**[2], **Epan**, **Bi-wgt**, **Tri-wgt**, **C-Gaus**, **Gauss** and **Exp4** respectively stand for classical COBRA, Epanechnikov, Bi-weight, Tri-weight, Compact-support Gaussian, Gaussian and 4-exponential kernels as listed in Table 1. In this block, the smallest MSE of each case is again written in **boldface**. For all the compactly supported kernels, we consider 500 values of 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, we set $\rho_1 = 3$ and $\sigma = 1$ therefore its support is $[-3, 3]$. Lastly, 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. Note that the results in the first block are not necessarily exactly the same as the ones reported in Biau et al. (2016) due to the choices of the parameters of the basic machines.

We can easily compare the performances of the combining estimation methods with all the basic machines and among themselves as the results reported in the second block are the straight combinations of those in the first block. In each table,

[2]We use the relaxed version of Biau et al. (2016) with the weights given in equation (3). `COBRA` library of `R` programming is used (see, Guedj (2013)).

we are interested in comparing the smallest average MSE in the first block to all the columns in the second block. First of all, we can see that all columns of the second block always outperform the best machine of the first block, which illustrates the theoretical result of the combining estimation methods. Secondly, the kernel-based methods beat the first column (classical COBRA) of the second block for almost all kernels. Lastly, the combing estimation method with Gaussian kernel is the absolute winner as the corresponding column is bold in both tables. Note that with the proposed gradient descent algorithm, we can obtain the value of bandwidth parameter with null gradient of cross-validation error defined in equation (10), which is often better and much faster than the one obtained by the grid search algorithm (2 or 3 times faster). Figure 2 below contains boxplots of runtimes of 100 runs of Model 1 and 9 of both correlated and uncorrelated cases computed on a 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.

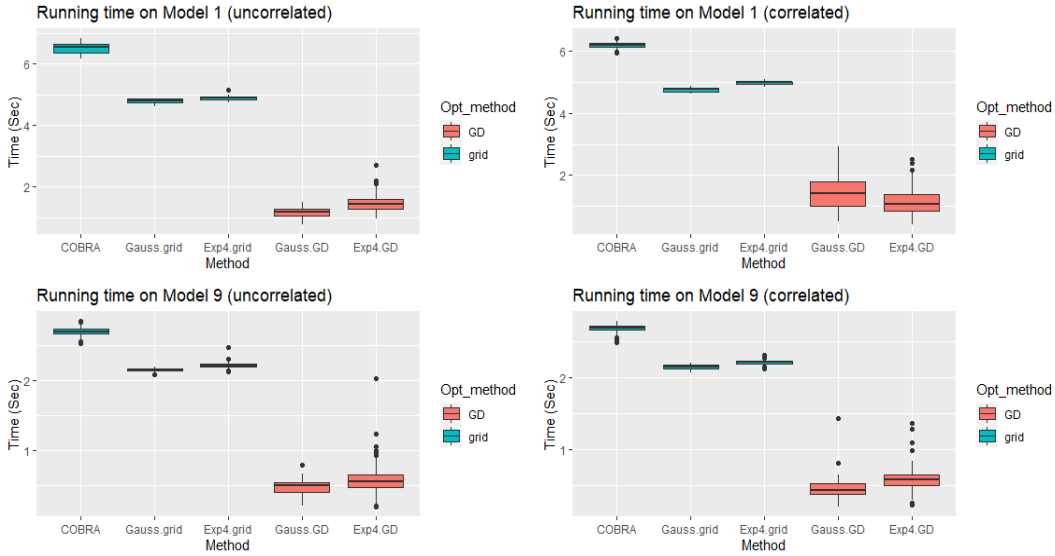

Figure 2: Boxplots of runtimes of GD and grid search algorithm implemented on some models.

Table 2: Average MSEs in the uncorrelated case.

Model	Las	Rid	kNN	Tr	RF	COBRA	Epan	Bi-wgt	Tri-wgt	C-Gaus	Gauss	Exp4
1	0.156 (0.016)	0.134 (0.013)	0.144 (0.014)	**0.027** (0.004)	0.033 (0.004)	0.022 (0.004)	0.020 (0.003)	0.019 (0.003)	0.019 (0.003)	0.019 (0.003)	**0.018** (0.002)	0.019 (0.003)
2	1.301 (0.216)	0.784 (0.110)	0.873 (0.123)	1.124 (0.165)	**0.707** (0.097)	0.722 (0.065)	0.718 (0.079)	0.712 (0.080)	0.715 (0.079)	0.712 (0.079)	**0.709** (0.078)	0.710 (0.079)
3	0.664 (0.107)	0.669 (0.255)	1.477 (0.192)	0.797 (0.135)	**0.629** (0.091)	0.554 (0.069)	0.482 (0.062)	0.478 (0.060)	0.476 (0.060)	0.479 (0.063)	**0.475** (0.060)	0.483 (0.060)
4	7.783 (1.121)	6.550 (1.115)	10.238 (1.398)	3.796 (0.840)	**3.774** (0.523)	3.608 (0.526)	3.231 (0.383)	3.185 (0.382)	3.153 (0.384)	3.189 (0.371)	**2.996** (0.384)	3.186 (0.464)
5	0.508 (0.051)	0.518 (0.073)	0.699 (0.084)	0.575 (0.081)	**0.436** (0.051)	0.429 (0.035)	0.389 (0.031)	0.387 (0.030)	0.386 (0.030)	0.387 (0.030)	**0.383** (0.030)	0.387 (0.028)
6	2.693 (0.537)	1.958 (0.292)	2.675 (0.349)	3.065 (0.475)	**1.826** (0.262)	1.574 (0.270)	1.274 (0.129)	1.259 (0.130)	**1.254** (0.130)	1.270 (0.125)	1.273 (0.130)	1.286 (0.130)
7	1.971 (0.410)	0.796 (0.132)	1.074 (0.152)	0.737 (0.109)	**0.515** (0.073)	0.506 (0.063)	0.472 (0.049)	0.468 (0.048)	0.467 (0.049)	0.469 (0.049)	**0.451** (0.049)	0.477 (0.067)
8	0.134 (0.016)	0.131 (0.020)	0.200 (0.020)	0.174 (0.034)	**0.127** (0.013)	0.104 (0.013)	0.092 (0.013)	**0.091** (0.013)	**0.091** (0.013)	**0.091** (0.013)	**0.091** (0.011)	0.094 (0.016)
9	1.592 (0.219)	2.948 (0.436)	3.489 (0.516)	1.830 (0.373)	**1.488** (0.267)	1.130 (0.151)	0.929 (0.128)	0.918 (0.127)	0.914 (0.130)	0.918 (0.124)	**0.895** (0.126)	0.993 (0.186)
10	2012.660 (284.391)	**1485.065** (210.816)	1778.955 (261.396)	3058.381 (486.504)	1618.977 (231.555)	1511.283 (129.796)	1462.509 (143.976)	1458.306 (142.988)	1459.558 (142.602)	1452.523 (141.168)	**1400.365** (143.330)	1414.316 (144.929)

Table 3: Average MSEs in the correlated case.

Model	Las	Rid	kNN	Tr	RF	COBRA	Epan	Bi-wgt	Tri-wgt	C-Gaus	Gauss	Exp4
1	2.294 (0.544	1.947 (0.507)	1.941 (0.487)	**0.320** (0.145)	0.542 (0.231)	0.307 (0.129)	0.304 (0.105)	0.301 (0.111)	0.288 (0.103)	0.297 (0.104)	**0.269** (0.092)	0.291 (0.098)
2	14.273 (2.593)	8.442 (1.912)	8.572 (1.751)	6.796 (1.548)	**5.135** (1.372)	5.345 (1.194)	4.582 (0.941)	4.529 (0.934)	4.491 (0.922)	4.541 (0.896)	**4.377** (0.905)	4.910 (1.181)
3	7.996 (3.393)	6.266 (3.296)	8.704 (3.523)	4.110 (2.894)	**3.722** (2.956)	3.327 (1.006)	2.598 (0.912)	2.536 (0.944)	2.444 (0.840)	2.554 (0.907)	**2.168** (0.680)	2.357 (0.756)
4	61.474 (13.986)	42.351 (11.622)	46.934 (12.543)	**8.855** (3.480)	13.381 (5.549)	9.599 (4.125)	10.511 (2.961)	9.963 (3.101)	9.682 (2.860)	10.085 (2.904)	**9.056** (2.407)	9.713 (2.695)
5	6.805 (3.685)	7.479 (5.336)	10.342 (5.425)	**4.000** (3.144)	4.880 (3.787)	3.225 (2.088)	2.640 (1.455)	2.401 (1.387)	2.235 (1.250)	2.412 (1.355)	**1.792** (0.913)	2.194 (1.242)
6	4.221 (0.848)	2.087 (0.485)	4.461 (0.599)	3.408 (0.636)	**1.701** (0.288)	1.493 (0.326)	1.271 (0.149)	1.238 (0.146)	1.217 (0.143)	1.248 (0.148)	**1.097** (0.145)	1.270 (0.386)
7	17.875 (5.632)	4.695 (1.318)	5.591 (1.418)	4.132 (1.360)	**3.081** (1.091)	3.304 (0.799	2.819 (0.636)	2.779 (0.614)	2.736 (0.605)	2.788 (0.623)	**2.640** (0.590)	2.979 (0.764)
8	0.139 (0.016)	0.133 (0.020)	0.201 (0.019)	0.159 (0.035)	**0.121** (0.013)	0.102 (0.021)	0.100 (0.020)	0.100 (0.021)	0.100 (0.020)	0.100 (0.020)	**0.092** (0.021)	**0.092** (0.018)
9	43.445 (12.210)	37.827 (12.201)	43.991 (12.920)	**15.258** (8.119)	16.957 (8.774)	13.505 (4.822)	11.303 (3.891)	11.007 (3.815)	11.067 (3.949)	11.206 (3.960)	**10.303** (3.634)	12.346 (5.014)
10	7235.062 (1100.579)	**5244.843** (996.181)	7636.811 (1159.445)	13014.596 (2020.133)	7092.741 (1030.249)	5147.950 (835.384)	4717.225 (703.049)	4669.516 (696.027)	4663.430 (687.474)	4697.019 (681.370)	**4660.043** (764.363)	5073.591 (1022.894)

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.

Notice that Model 6 and 8 of the previous subsection are about predicting integer labels of the response variable. Analogously, 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.

The five primary machines 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 machines are combined using the classical method by Biau et al. (2016) and the kernel-based method with Gaussian kernel as it is the most outstanding one among all the kernel functions. In this case, the search for parameter h for the classical COBRA method is performed using a grid of size 300. In addition, due to the scaling issue, we measure the performance of any method f in this case using average *root mean square error* (RMSE) defined by,

$$\mathrm{RMSE}(f) = \sqrt{\mathrm{MSE}(f)} = \sqrt{\frac{1}{n_{\mathrm{test}}} \sum_{i=1}^{n_{\mathrm{test}}} (y_i^{\mathrm{test}} - f(x_i^{\mathrm{test}}))^2}. \tag{12}$$

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 machines in the first block, and the proposed method either outperforms other columns (**Wine** and **Abalone**) or biases towards the best basic machine (**House**).

Moreover, the performances of kernel-based method always exceed the ones of the classical method by Biau et al. (2016).

4.3 Real private datasets

The results presented in this subsection are obtained from two 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. The results obtained from 100 independent runs of the methods are presented in the last two rows (**Air** and **Turbine**) of Table 4 below. We observe on one hand that the proposed method (**Gauss**) outperforms both the best basic machines (**RF**) and the classical method by Biau et al. (2016) in the case of **Turbine** dataset. On the other hand, the performance of our method approaches the performance of the best basic machine (**Las**) and outperforms the classical COBRA in the case of **Air** dataset. Moreover, boxplots of runtimes (100 runs) measured on **Wine** and **Turbine** datasets (computed using the same machine as described in the subsection of simulated data) are also given in Figure 3 below.

Table 4: Average RMSEs of real datasets.

Model	**Las**	**Rid**	**kNN**	**Tr**	**RF**	**COBRA**	**Gauss**
House	241083.959 (8883.107)	241072.974 (8906.332)	245153.608 (23548.367)	254099.652 (9350.885)	**205943.768** (7496.766)	223596.317 (13299.934)	**209955.276** (7815.623)
Wine	0.0.660 (0.029)	0.685 (0.053)	0.767 (0.031)	0.711 (0.030)	**0.623** (0.028)	0.650 (0.026)	**0.617** (0.020)
Abalone	2.204 (0.071)	2.215 (0.075)	2.175 (0.062)	2.397 (0.072)	**2.153** (0.060)	2.171 (0.081)	**2.128** (0.057)
Air	**163.099** (3.694)	164.230 (3.746)	241.657 (5.867)	351.317 (31.876)	174.836 (6.554)	172.858 (7.644)	**163.253** (3.333)
Turbine	70.051 (4.986)	68.987 (3.413)	44.516 (1.671)	81.714 (4.976)	**38.894** (1.506)	38.927 (1.561)	**37.135** (1.555)

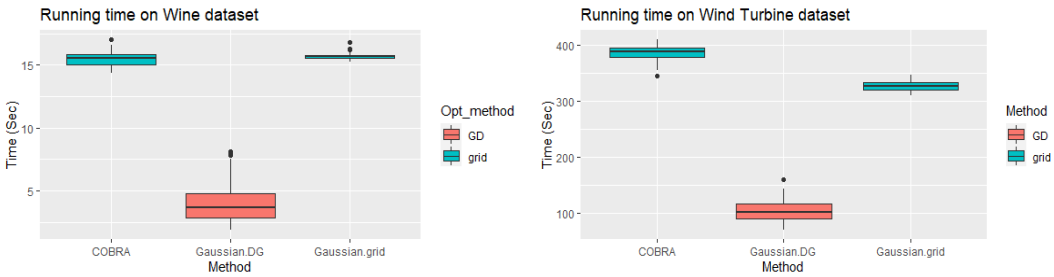

Figure 3: Boxplots of runtimes of GD and grid search algorithm implemented on **Wine** and **Turbine** datasets.

5 Conclusion

In this study, we investigate and extend the context of a naive kernel-based consensual regression aggregation method by Biau et al. (2016) to a more general regular kernel-based framework. Moreover, an optimization algorithm based on gradient descent is proposed to efficiently and rapidly estimate the key parameter of the method. It is also shown through several numerical simulations that the performance of the method is improved significantly with smoother kernel functions.
In practice, the performance of the consensual aggregation depends both on the performance of the individual regression machines, and on the final combination here involving kernel functions. Since the calibration of hyperparameters may be critical in both steps, it could be very interesting to investigate in future work how automated machine learning models can improve the performances of the global model.

6 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 $g^*(\boldsymbol{r}_k(X)) = \mathbb{E}[g^*(X)|\boldsymbol{r}_k(X)]$ thanks 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,

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

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

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

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

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

■

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

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

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,*

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

$$= \ell \mathbb{E}_{\mathcal{D}_k}\Big[\int |g^*(\boldsymbol{r}_k(u)) - \tilde{g}(\boldsymbol{r}_k(u))|^2 \times$$

$$\mathbb{E}_{\{X_j\}_{j=2}^{\ell}}\Big[\int \frac{K_h(\boldsymbol{r}_k(v) - \boldsymbol{r}_k(u))\mu(dv)}{K_h(\boldsymbol{r}_k(v) - \boldsymbol{r}_k(u)) + \sum_{j=2}^{\ell} K_h(\boldsymbol{r}_k(v) - \boldsymbol{r}_k(X_j))}\Big|\mathcal{D}_k\Big]\mu(du)\Big]$$

$$= \ell \mathbb{E}_{\mathcal{D}_k}\Big[\int |g^*(\boldsymbol{r}_k(u)) - \tilde{g}(\boldsymbol{r}_k(u))|^2 \times I(u,\ell)\mu(du)\Big].$$

Fubini's theorem 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 $\mathcal{B} \stackrel{\text{def}}{=} \{B_M(x_i, \rho/2) : i = 1, 2, ...\}$ 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

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

Thus, one has

$$I(u,\ell) \stackrel{\text{def}}{=} \mathbb{E}_{\{X_j\}_{j=2}^{\ell}}\Big[\int \frac{K_h(\boldsymbol{r}_k(v) - \boldsymbol{r}_k(u))\mu(dv)}{K_h(\boldsymbol{r}_k(v) - \boldsymbol{r}_k(u)) + \sum_{j=2}^{\ell} K_h(\boldsymbol{r}_k(v) - \boldsymbol{r}_k(X_j))}\Big|\mathcal{D}_k\Big]$$

$$\leq \mathbb{E}_{\{X_j\}_{j=2}^{\ell}}\Big[\sum_{i=1}^{+\infty}\int_{v:\|\boldsymbol{r}_k(v) - \boldsymbol{r}_k(u) - hx_i\| < h\rho/2}$$

$$\frac{K_h(\boldsymbol{r}_k(v) - \boldsymbol{r}_k(u))\mu(dv)}{K_h(\boldsymbol{r}_k(v) - \boldsymbol{r}_k(u)) + \sum_{j=2}^{\ell} K_h(\boldsymbol{r}_k(v) - \boldsymbol{r}_k(X_j))}\Big|\mathcal{D}_k\Big]$$

$$\leq \mathbb{E}_{\{X_j\}_{j=2}^{\ell}}\Big[\sum_{i=1}^{+\infty}\int_{A_{i,h}(u)}$$

$$\frac{\sup_{z:\|z - hx_i\| < h\rho/2} K_h(z)\mu(dv)}{\sup_{z:\|z - hx_i\| < h\rho/2} K_h(z) + \sum_{j=2}^{\ell} K_h(\boldsymbol{r}_k(v) - \boldsymbol{r}_k(X_j))}\Big|\mathcal{D}_k\Big]$$

$$\leq \mathbb{E}_{\{X_j\}_{j=2}^{\ell}}\Big[\sum_{i=1}^{+\infty}\int_{A_{i,h}(u)}$$

$$\frac{\sup_{z:\|z - hx_i\| < h\rho/2} K_h(z)\mu(dv)}{\sup_{z:\|z - hx_i\| < h\rho/2} K_h(z) + b\sum_{j=2}^{\ell} \mathbb{1}_{\{\|\boldsymbol{r}_k(v) - \boldsymbol{r}_k(X_j)\| < h\rho\}}}\Big|\mathcal{D}_k\Big]$$

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

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

$$\begin{aligned}
\sum_{i=1}^{\ell} W_{n,i}^2(X) &= \sum_{i=1}^{\ell} \Big(\frac{K_h(\boldsymbol{r}_k(X) - \boldsymbol{r}_k(X_i))}{\sum_{j=1}^{\ell} K_h(\boldsymbol{r}_k(X) - \boldsymbol{r}_k(X_j))}\Big)^2 \\
&\leq \frac{\sum_{i=1}^{\ell} K_h(\boldsymbol{r}_k(X) - \boldsymbol{r}_k(X_i))}{\Big(\sum_{j=1}^{\ell} K_h(\boldsymbol{r}_k(X) - \boldsymbol{r}_k(X_j))\Big)^2} \\
&\leq \min\Big\{\delta, \frac{\mathbb{1}_{\{\sum_{j=1}^{\ell} K_h(\boldsymbol{r}_k(X) - \boldsymbol{r}_k(X_j))>0\}}}{\sum_{j=1}^{\ell} K_h(\boldsymbol{r}_k(X) - \boldsymbol{r}_k(X_j))}\Big\} \\
&\leq \min\Big\{\delta, \frac{\mathbb{1}_{\{\sum_{j=1}^{\ell} \mathbb{1}_{\{\|\boldsymbol{r}_k(X) - \boldsymbol{r}_k(X_j)\| < h\rho\}}>0\}}}{b\sum_{j=1}^{\ell} \mathbb{1}_{\{\|\boldsymbol{r}_k(X) - \boldsymbol{r}_k(X_j)\| < h\rho\}}} \\
&\leq \delta + \frac{\mathbb{1}_{\{\sum_{j=1}^{\ell} \mathbb{1}_{\{\|\boldsymbol{r}_k(X) - \boldsymbol{r}_k(X_j)\| < h\rho\}}>0\}}}{b\sum_{j=1}^{\ell} \mathbb{1}_{\{\|\boldsymbol{r}_k(X) - \boldsymbol{r}_k(X_j)\| < h\rho\}}}. \qquad (13)
\end{aligned}$$

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 $B \subset \cup_{j \in I_{h,M}} B_M(x_j, h\rho/2)$.

$$
\begin{aligned}
&\mathbb{E}\Big[\frac{\mathbb{1}_{\{\boldsymbol{r}_k(X) \in B\}}}{(\ell+1)\mu(\{v \in \mathbb{R}^d : \|\boldsymbol{r}_k(v) - \boldsymbol{r}_k(X)\| < h\rho\})}\Big] \\
&\leq \sum_{j \in I_{h,M}} \int_{u:\|\boldsymbol{r}_k(u) - x_j\| < h\rho/2} \frac{\mu(du)}{(\ell+1)\mu(\{v \in \mathbb{R}^d : \|\boldsymbol{r}_k(v) - \boldsymbol{r}_k(u)\| < h\rho\})} \\
&\quad + \mu(\{v \in \mathbb{R}^d : \boldsymbol{r}_k(v) \in B^c\}) \\
&\leq \sum_{j \in I_{h,M}} \int_{u:\|\boldsymbol{r}_k(u) - x_j\| < h\rho/2} \frac{\mu(du)}{(\ell+1)\mu(\{v \in \mathbb{R}^d : \|\boldsymbol{r}_k(v) - x_j\| < h\rho/2\})} \\
&\quad + \mu(\{v \in \mathbb{R}^d : \boldsymbol{r}_k(v) \in B^c\}) \\
&= \sum_{j \in I_{h,M}} \frac{\mu(\{u \in \mathbb{R}^d : \|\boldsymbol{r}_k(u) - x_j\| < h\rho/2\})}{(\ell+1)\mu(\{v \in \mathbb{R}^d : \|\boldsymbol{r}_k(v) - x_j\| < h\rho/2\})} + \mu(\{v \in \mathbb{R}^d : \boldsymbol{r}_k(v) \in B^c\}) \\
&= \frac{|I_{h,M}|}{\ell+1} + \mu(\{v \in \mathbb{R}^d : \boldsymbol{r}_k(v) \in B^c\}) \\
&\leq \frac{C_0}{h^M(\ell+1)} + \mu(\{v \in \mathbb{R}^d : \boldsymbol{r}_k(v) \in B^c\}) \qquad (14) \\
&\xrightarrow[h^M \ell \to +\infty]{\ell \to +\infty, h \to 0} \mu(\{v \in \mathbb{R}^d : \boldsymbol{r}_k(v) \in B^c\}).
\end{aligned}
$$

It is easy to check the following fact,

$$
|I_{h,M}| \leq \frac{C_0}{h^M} \text{ for some } C_0 > 0. \qquad (15)
$$

To prove this, 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 $\mu(\{v \in \mathbb{R}^d : \boldsymbol{r}_k(v) \in B^c\}) = 0$ thanks to the boundedness of the basic machines.

Remark 2 *The assumption on the boundedness of the constructed machines 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,

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

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

where

$$
A_h(x) \stackrel{\text{def}}{=} \{v\in\mathbb{R}^d : \|\boldsymbol{r}_k(x)-\boldsymbol{r}_k(v)\| < h\rho\}. \tag{16}
$$

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 $\delta>0$:

$$
\begin{aligned}
E_1 &\stackrel{\text{def}}{=} \mathbb{E}\Big[\Big|g_n(\boldsymbol{r}_k(x)) - \mathbb{E}[g_n(\boldsymbol{r}_k(x))|\{\boldsymbol{r}_k(X_p)\}_{p=1}^{\ell}, \mathcal{D}_k]\Big|^2\Big|\{\boldsymbol{r}_k(X_p)\}_{p=1}^{\ell}, \mathcal{D}_k\Big] \\
&= \mathbb{E}\Big[\Big|\sum_{i=1}^{\ell} W_{n,i}(x)(Y_i - \mathbb{E}[Y_i|\boldsymbol{r}_k(X_i)])\Big|^2\Big|\{\boldsymbol{r}_k(X_p)\}_{p=1}^{\ell}, \mathcal{D}_k\Big] \\
&= \mathbb{E}\Big[\sum_{i=1}^{\ell} W_{n,i}^2(x)(Y_i - \mathbb{E}[Y_i|\boldsymbol{r}_k(X_i)])^2\Big|\{\boldsymbol{r}_k(X_p)\}_{p=1}^{\ell}, \mathcal{D}_k\Big] \\
&= \sum_{i=1}^{\ell} W_{n,i}^2(x)\mathbb{E}_{Y_i}[(Y_i - \mathbb{E}[Y_i|\boldsymbol{r}_k(X_i)])^2|\boldsymbol{r}_k(X_i)] \\
&= \mathbb{V}[Y_1|\boldsymbol{r}_k(X_1)]\sum_{i=1}^{\ell} W_{n,i}^2(x) \\
&\stackrel{(13)}{\leq} \frac{4R^2}{b}\Big(\delta + \frac{\mathbb{1}_{\{\sum_{j=1}^{\ell}\mathbb{1}_{\{\|\boldsymbol{r}_k(x)-\boldsymbol{r}_k(X_j)\|<h\rho\}}>0\}}}{\sum_{j=1}^{\ell}\mathbb{1}_{\{\|\boldsymbol{r}_k(x)-\boldsymbol{r}_k(X_j)\|<h\rho\}}}\Big)
\end{aligned}
$$

where the notation $\mathbb{V}(Z)$ stands for the variance of a random variable Z. Therefore, using the result of inequality (14), one has

$$
\mathbb{E}(E_1) \leq \frac{4R^2}{b}\Big(\delta + \frac{C_0}{h^M(\ell+1)}\Big) \tag{17}
$$

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

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

The second term E_2 is hard to control as it depends on the behavior of $g^*(\boldsymbol{r}_k(.))$. That is why a weak smoothness assumption of the theorem is required to connect this behavior to the behavior of the input machines. Using this assumption, one has

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

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

for any $\alpha \in (0,1)$ 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*

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

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

$$
\mathbb{E}(E_2^1) \leq L^2 h^{2\alpha} R_K^2 \tag{18}
$$

- *Secondly, we bound the second term E_2^2 using the tail assumption of the kernel*

K given equation (7), thus for any $h > 0$:

$$
\begin{aligned}
E_2^2 &\stackrel{\text{def}}{=} h^2 L^2 \sum_{i=1}^{\ell} \frac{K_h(\boldsymbol{r}_k(x) - \boldsymbol{r}_k(X_i)) \|(\boldsymbol{r}_k(x) - \boldsymbol{r}_k(X_i))/h\|^2}{\sum_{j=1}^{\ell} K_h(\boldsymbol{r}_k(x) - \boldsymbol{r}_k(X_j))} \mathbb{1}_{\{D_h^\ell(x)>0\}} \times \\
&\quad \mathbb{1}_{\{\|\boldsymbol{r}_k(x) - \boldsymbol{r}_k(X_i)\| \geq R_K h^\alpha\}} \\
&\leq h^2 L^2 \sum_{i=1}^{\ell} \frac{C_K \mathbb{1}_{\{\|(\boldsymbol{r}_k(x) - \boldsymbol{r}_k(X_i))/h\| \geq R_K/h^{1-\alpha}\}} \mathbb{1}_{\{D_h^\ell(x)>0\}}}{(1 + \|(\boldsymbol{r}_k(x) - \boldsymbol{r}_k(X_i))/h\|^M) \sum_{j=1}^{\ell} K_h(\boldsymbol{r}_k(x) - \boldsymbol{r}_k(X_j))} \\
&\leq h^{M+2} L^2 \sum_{i=1}^{\ell} \frac{C_K \mathbb{1}_{\{\|\boldsymbol{r}_k(x) - \boldsymbol{r}_k(X_i)\| \geq R_K h^\alpha\}} \mathbb{1}_{\{D_h^\ell(x)>0\}}}{(h^M + (R_K h^\alpha)^M) \sum_{j=1}^{\ell} K_h(\boldsymbol{r}_k(x) - \boldsymbol{r}_k(X_j))} \\
&\leq h^{M+2} L^2 C_K \sum_{i=1}^{\ell} \frac{\mathbb{1}_{\{\|\boldsymbol{r}_k(x) - \boldsymbol{r}_k(X_i)\| \geq R_K h^\alpha\}}}{(h^M + R_K^M h^{\alpha M}) \sum_{j=1}^{\ell} K_h(\boldsymbol{r}_k(x) - \boldsymbol{r}_k(X_j))} \mathbb{1}_{\{D_h^\ell(x)>0\}} \\
&\leq \frac{h^{M+2-\alpha M} L^2 C_K}{h^{M(1-\alpha)} + R_K^M} \times \frac{\sum_{i=1}^{\ell} \mathbb{1}_{\{\|\boldsymbol{r}_k(x) - \boldsymbol{r}_k(X_i)\| \geq R_K h^\alpha\}}}{\sum_{j=1}^{\ell} K_h(\boldsymbol{r}_k(x) - \boldsymbol{r}_k(X_j))} \mathbb{1}_{\{D_h^\ell(x)>0\}} \\
&\leq \frac{h^{(1-\alpha)M+2} L^2 C_K}{b R_K^M} \times \frac{\sum_{i=1}^{\ell} \mathbb{1}_{\{\|\boldsymbol{r}_k(x) - \boldsymbol{r}_k(X_i)\| \geq R_K h^\alpha\}}}{\sum_{j=1}^{\ell} \mathbb{1}_{\{\|\boldsymbol{r}_k(x) - \boldsymbol{r}_k(X_j)\| < h\rho\}}} \mathbb{1}_{\{C_h^\ell(x)>0\}}.
\end{aligned}
$$

Therefore,

$$
E_2^2 \leq \frac{h^{(1-\alpha)M+2} L^2 C_K \ell}{b R_K^M} \times \frac{\mathbb{1}_{\{\sum_{j=1}^{\ell} \mathbb{1}_{\{\|\boldsymbol{r}_k(x) - \boldsymbol{r}_k(X_j)\| < h\rho\}} > 0\}}}{\sum_{j=1}^{\ell} \mathbb{1}_{\{\|\boldsymbol{r}_k(x) - \boldsymbol{r}_k(X_j)\| < h\rho\}}}.
$$

Again, applying the result of inequality (14), one has

$$
\mathbb{E}(E_2^2) \leq \frac{h^{(1-\alpha)M+2} L^2 C_K \ell}{b R_K^M} \times \frac{C_0}{h^M(\ell+1)} \leq \frac{C_1 \ell}{(\ell+1)} h^{2-\alpha M} \tag{19}
$$

for some $C_1 > 0$ and $\alpha < 2/M$.

- *Lastly with $A_h(x)$ defined in (16), we bound the expectation of E_2^3 by,*

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

$$
\begin{aligned}
&\leq \sup_{u\in\mathbb{R}^d} (g^*(\boldsymbol{r}_k(u)))^2 \frac{\ell\mu(A_h(x))e^{-\ell\mu(A_h(x))}}{\ell\mu(A_h(x))} \\
&\leq \sup_{u\in\mathbb{R}^d} (g^*(\boldsymbol{r}_k(u)))^2 \frac{\max_{u\in\mathbb{R}^d} ue^{-u}}{\ell\mu(A_h(x))} \\
&\leq \sup_{u\in\mathbb{R}^d} (g^*(\boldsymbol{r}_k(u)))^2 \frac{e^{-1}}{\ell\mu(A_h(x))} \\
&\leq \frac{C_2}{\ell\mu(A_h(x)))} \qquad (20)
\end{aligned}
$$

for some $C_2 > 0$.

From (17), (18), (19) *and* (20), *one has*

$$
\begin{aligned}
\mathbb{E}[|g_n(\boldsymbol{r}_k(X)) - g^*(\boldsymbol{r}_k(X))|^2] &\leq \int_{\mathbb{R}^d} \mathbb{E}[|g_n(\boldsymbol{r}_k(x)) - g^*(\boldsymbol{r}_k(x))|^2]\mu(dx) \\
&\leq \int_{\mathbb{R}^d} \mathbb{E}(E_1 + E_2^1 + E_2^2 + E_2^3)\mu(dx) \\
&\leq \int_{\mathbb{R}^d} \Big[\frac{4R^2}{b}\Big(\delta + \frac{C_0}{h^M(\ell+1)}\Big) + L^2h^{2\alpha}R_K^2 \\
&\quad + \frac{C_1\ell}{(\ell+1)}h^{2-\alpha M} + \frac{C_2}{\ell\mu(A_h(x)))}\Big]\mu(dx).
\end{aligned}
$$

Therefore following the same procedure of proving inequality (14), one has

$$
\begin{aligned}
&\mathbb{E}[|g_n(\boldsymbol{r}_k(X)) - g^*(\boldsymbol{r}_k(X))|^2] \\
&\leq \frac{4R^2}{b}\Big(\delta + \frac{C_0}{h^M(\ell+1)}\Big) + L^2h^{2\alpha}R_K^2 + \frac{C_1\ell}{(\ell+1)}h^{2-\alpha M} + \int_{\mathbb{R}^d} \frac{C_2\mu(dx)}{\ell\mu(A_h(x)))} \\
&\leq \frac{4R^2}{b}\Big(\delta + \frac{C_0}{h^M(\ell+1)}\Big) + L^2h^{2\alpha}R_K^2 + \frac{C_1\ell}{(\ell+1)}h^{2-\alpha M} \\
&\quad + \sum_{j\in J_{h,M}} \int_{\|\boldsymbol{r}_k(x)-x_j\|<h\rho} \frac{C_2\mu(dx)}{\ell\mu(\{v\in\mathbb{R}^d : \|\boldsymbol{r}_k(v) - \boldsymbol{r}_k(x)\| < h\rho\})} \\
&\leq \frac{4R^2}{b}\Big(\delta + \frac{C_0}{h^M(\ell+1)}\Big) + L^2h^{2\alpha}R_K^2 + \frac{C_1\ell}{(\ell+1)}h^{2-\alpha M} \\
&\quad + \sum_{j\in J_{h,M}} \int_{\|\boldsymbol{r}_k(x)-x_j\|<h\rho} \frac{C_2\mu(dx)}{\ell\mu(\{v\in\mathbb{R}^d : \|\boldsymbol{r}_k(v) - x_j\| < h\rho\})}
\end{aligned}
$$

$$
\begin{aligned}
&\leq \frac{4R^2}{b}\Big(\delta + \frac{C_0}{h^M(\ell+1)}\Big) + L^2 h^{2\alpha} R_K^2 + \frac{C_1 \ell}{(\ell+1)} h^{2-\alpha M} \\
&\quad + \frac{C_2}{\ell} \sum_{j \in J_{h,M}} \frac{\mu(\{v \in \mathbb{R}^d : \|\boldsymbol{r}_k(v) - x_j\| < h\rho\})}{\mu(\{v \in \mathbb{R}^d : \|\boldsymbol{r}_k(v) - x_j\| < h\rho\})} \\
&\leq \frac{4R^2}{b}\Big(\delta + \frac{C_0}{h^M(\ell+1)}\Big) + L^2 h^{2\alpha} R_K^2 + \frac{C_1 \ell}{(\ell+1)} h^{2-\alpha M} + \frac{C_2 |J_{h,M}|}{\ell} \\
&\leq \frac{4R^2}{b}\Big(\delta + \frac{C_0}{h^M(\ell+1)}\Big) + L^2 R_K^2 h^{2\alpha} + \frac{C_1 \ell}{(\ell+1)} h^{2-\alpha M} + \frac{C_2'}{h^M \ell}
\end{aligned}
$$

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 $\delta > 0$ can be arbitrarily small, and with the choice of $\alpha = 2/(M+2)$, we can deduce that

$$
\mathbb{E}[|g_n(\boldsymbol{r}_k(X)) - g^*(\boldsymbol{r}_k(X))|^2] \leq \frac{\tilde{C}_1}{h^M \ell} + \tilde{C}_2 h^{4/(M+2)}. \tag{21}
$$

From this bound, for $h \propto \ell^{-(M+2)/(M^2+2M+4)}$ we obtain the desire result with the upper bound of order $O(\ell^{-\frac{4}{M^2+2M+4}})$ i.e.,

$$
\mathbb{E}[|g_n(\boldsymbol{r}_k(X)) - g^*(\boldsymbol{r}_k(X))|^2] \leq C \ell^{-\frac{4}{M^2+2M+4}}
$$

for some constant $C > 0$ independent of ℓ.

■

***Proof of Remark* 1** *To prove the result in this case, which means, under the following assumption:*

$$
\exists R_K, C_K > 0 \text{ and } \alpha \in (0,1) : K(z) \leq C_K e^{-\|z\|^\alpha}, \forall z \in \mathbb{R}^M, \|z\| \geq R_K,
$$

we only need to check the new bound of E_2^2 defined in the previous case. One has

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

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

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

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

Applying the result of inequality (14), one has

$$\begin{aligned}
\mathbb{E}(E_2^2) &\leq \frac{\ell h^{2\alpha}L^2C_KR_K^2e^{-R_K^\alpha h^{-\alpha(1-\alpha)}}}{b}\times\frac{C_0}{h^M(\ell+1)}\\
&\leq C_1h^{2\alpha-M}e^{-R_K^\alpha h^{-\alpha(1-\alpha)}} \qquad (22)
\end{aligned}$$

for some $C_1>0$ and $\alpha\in(0,1)$.
Therefore from (17), (18), (20) *and* (22), *one has*

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

Following the same procedure as in the previous proof of theorem 1, one has

$$\begin{aligned}
&\mathbb{E}[|g_n(\boldsymbol{r}_k(X))-g^*(\boldsymbol{r}_k(X))|^2]\\
&\leq \frac{4R^2}{b}\Big(\delta+\frac{C_0}{h^M(\ell+1)}\Big)+L^2h^{2\alpha}R_K^2+C_1h^{2\alpha-M}e^{-R_K^\alpha h^{-\alpha(1-\alpha)}}+\frac{C_2'}{h^M\ell}.
\end{aligned}$$

Since $\delta > 0$ is chosen arbitrarily and the third term of the last inequality decreases exponentially fast as $h \to 0$ for any $\alpha \in (0,1)$ fixed, hence it is negligible comparing to other terms. Finally, with the choice of $h \propto \ell^{-1/(M+2\alpha)}$, we obtain the desire result:

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

for some $C > 0$ independent of ℓ.

■

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.

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

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

Breiman, L., 1995. Stacked regression. Machine Learning 24, 49–64.

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.

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.

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.

Dalalyan, A., Tsybakov, A.B., 2008. Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity. Machine Learning 72, 39–61.

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.

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.

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software 33, 1–22.

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.

Juditsky, A., Nemirovski, A., 2000. Functional aggregation for nonparametric estimation. The Annals of Statistics 28, 681–712.

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

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.

Mojirsheibani, M., 2000. A kernel-based combined classification rule. Journal of Statistics and Probability Letters 48, 411–419.

Mojirsheibani, M., Kong, J., 2016. An asymptotically optimal kernel combined classifier. Journal of Statistics and Probability Letters 119, 91–100.

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.

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

Wegkamp, M.H., 2003. Model selection in nonparametric regression. The Annals of Statistics 31, 252–273.

Yang, Y., 2000. Combining different procedures for adaptive regression. Journal of multivariate analysis 74, 135–161.

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.