Aggregation using input-output trade-off

Aurélie Fischer & Mathilde Mougeot

Laboratoire de Probabilités et Modèles Aléatoires
Université Paris Diderot
75013 Paris, France

March 28, 2017

Abstract

In this paper, we introduce a new learning strategy based on a seminal idea of Mojirsheibani (1999, 2000, 2002a, 2002b), who proposed a smart method for combining several classifiers, relying on a consensus notion. In many aggregation methods, the prediction for a new observation x is computed by building a linear or convex combination over a collection of basic estimators $r_1(x), \ldots, r_m(x)$ previously calibrated using a training data set. Mojirsheibani proposes to compute the prediction associated to a new observation by combining selected outputs of the training examples. The output of a training example is selected if some kind of consensus is observed: the predictions computed for the training example with the different machines have to be "similar" to the prediction for the new observation. This approach has been recently extended to the context of regression in Biau et al. (2016).

In the original scheme, the agreement condition is actually required to hold for all individual estimators, which appears inadequate if there is one bad initial estimator. In practice, a few disagreements are allowed ; for establishing the theoretical results, the proportion of estimators satisfying the condition is required to tend to 1.

In this paper, we propose an alternative procedure, mixing the previous consensus ideas on the predictions with the Euclidean distance computed between entries. This may be seen as an alternative approach allowing to reduce the effect of a possibly bad estimator in the initial list, using a constraint on the inputs.

We prove the consistency of our strategy in classification and in regression. We also provide some numerical experiments on simulated and real data to illustrate the benefits of this new aggregation method. On the whole, our practical study shows that our method may perform much better than the original combination technique, and, in particular, exhibit far less variance. We also show on simulated examples that this procedure mixing inputs and outputs is still robust to high dimensional inputs.

Keywords : Classification, regression estimation, aggregation, nonlinearity, consistency.

AMS classification : 62G05, 62G08, 62H30

1   Introduction

Given the growing number of available statistical estimation strategies, trying to combine several procedures in a smart way is a very natural idea, and so, an abundant literature on aggregation of estimators for different types of statistical models has sprung up in recent years. A widespread aggregation method consists in building a linear or a convex combination of a bunch of initial estimators (see, for instance, 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). Observe that the model selection approach, which aims at selecting the best estimator in the list, is also linked to the same goal (see, for example, the monograph by Massart (2007)).

In this paper, we focus on the question of combining estimators for two kinds of statistical procedures, pertaining to the supervised learning framework: classification, which consists in assigning to an observation a label, out of a finite number of candidates, and regression, whose aim is to associate real outputs to entries.

Beside the usual linear aggregation and model selection methods, a quite different point of view has been introduced by Mojirsheibani (1999) for classification. The idea consists in an original combination method, which is non linear in the initial estimators and is based on a consensus concept. More specifically, in this combining scheme, an observation is considered to be reliable for contributing to the classification of a new query point if all initial classifiers predict the same label for both points. Then, the label for the query point is estimated thanks to a majority vote among the labels of the observations which have been retained this way. Note that more regular versions, based on smoothing kernels, have also been developed (Mojirsheibani (2000)). A numerical comparison study of several combining schemes is available in Mojirsheibani (2002b), and recently, a variant of the method has been proposed in Balakrishnan and Mojirsheibani (2015).

This strategy has just been adapted in the regression framework by Biau et al. (2016). In this context, an observation is used in the combination step if all the initial estimators predict a similar value for the observation and the new point: the difference between both predictions is required to be less than some prespecified threshold. Then, the new prediction by the combined estimator is the average of the outputs corresponding to the selected entries. Note that the functional data framework has also been considered, by Cholaquidis et al. (2016).

In the classification case as well as in the regression one, when the initial list contains a consistent estimator, it can be shown that the combined estimator inherits this consistency property. Let us mention that the techniques of proof, and consequently, the assumptions made in both situations, are quite different. For instance, the number of initial estimators is expected to tend to infinity with the sample size in Mojirsheibani (2000) whereas it is fixed in Biau et al. (2016).

It is worth pointing out that, in both contexts described above, the condition for an observation to be reliable is in principle required to be satisfied for all estimators. In a further paper, Mojirsheibani (2002a) notes that this rule may seem too restrictive and proposes to allow a few disagreements (typically, a single one). The resulting classifier is still consistent provided that the number of initial classifiers keeps tending to infinity after removing those with disagreement. Similarly, in Biau et al. (2016), this unanimity constraint is relaxed in practice by demanding that the distance condition for keeping an observation is true at least for a certain proportion α of the estimators (for example, $\alpha = 80\%$). The theoretical results remain true provided that α tends to 1.

Here, our purpose is to investigate a new approach, based on distances between observations, which also aims at reducing the effect of a possibly bad initial estimator. Roughly, choosing a kernel point of view, we will propose a combined estimator with weights constructed by mixing distances between entries with distances between predictions coming from the individual estimators. Our motivation for introducing such a strategy is the intuition that taking advantage of the efficiency of the consensus idea of Mojirsheibani (1999) and Biau et al. (2016) without for all that forgetting the information related to the proximity between entries shall help improving the prediction, especially in the presence of an initial estimator that does not perform very well.

Our modified rule will be shown to be consistent under general assumptions. In particular, the combined estimator may perfectly be consistent even if the list of initial estimators does not contain any consistent estimator. We also conduct numerical experiments, both on simulated and real data, which demonstrate the benefits of our strategy, with respect to the original combining method and the individual estimators. The paper is organized as follows. The new combined estimator is defined in Section 2. Then, the main theoretical results are stated in Section 3. Section 4 is devoted to numerical experiments with simulated and real examples. A few perspectives are presented in a brief conclusive paragraph, Section 5. For the sake of clarity, proofs are postponed to Section 6.

2 Notation and definition of the estimator

Let (X, Y) denote a random pair taking its values in $\mathbb{R}^d \times \mathcal{Y}$. The variable X has distribution μ. We are interested in two different situations: $\mathcal{Y} = \{0, 1\}$, which corresponds to the binary classification problem, and $\mathcal{Y} = [0, 1]$, that is bounded regression. Let η stand for the regression function $\eta(x) = \mathbb{E}[Y|X = x]$.

Note that $\eta(x) = \mathbb{P}(Y = 1 | X = x)$ in the classification context.

Let $\psi^\star$ denote the Bayes classifier, given by

$$\psi^\star(x) = \begin{cases} 1 & \text{if } \eta(x) > 1/2 \\ 0 & \text{otherwise.} \end{cases}$$

It is well-known that $\psi^\star$ minimizes over all possible classifiers ψ the missclassification error $L(\psi) = \mathbb{P}(\psi(X) \neq Y)$.

Throughout the document, we assume that we are given a sample $\mathcal{D}_n = \{(X_1, Y_1) \ldots, (X_n, Y_n)\}$ of the random pair (X, Y).

Our goal in regression is to estimate the function η using $\mathcal{D}_n$. In classification, we aim at building a classifier based on $\mathcal{D}_n$ whose error mimics the Bayes classifier error.

For every $k \geq 1$, if $u \in \mathbb{R}^k$, $\|u\|_k$ will denote the Euclidean norm of the vector u, and $B_k(x, r) = \{y \in \mathbb{R}^k, \|x - y\|_k \leq r\}$ the closed ball centered at $x \in \mathbb{R}^k$, with radius $r > 0$. To simplify notation, as there is no ambiguity, we will most of the time drop the index k and simply write $\|u\|$ and $B(x, r)$.

Let $K : \mathbb{R}^{d+p} \mapsto \mathbb{R}_+$ be a kernel, that is a nonnegative and monotone decreasing function along rays starting from the origin. The next assumption will be made on the kernel K (see Devroye et al. (1996)).

Assumption 2.1. *We suppose that the kernel K is regular, that is, there exist $c > 0$ and $\rho > 0$ such that*

- *For all z, $K(z) \geq c\mathbf{1}_{B_{d+p}(0,\rho)}(z)$.*

- *$\int \sup_{t \in B_{d+p}(z,\rho)} K(t) dz < \infty$.*

We propose to combine the predictions of p initial estimators, denoted by $c_1, \ldots, c_p$ in the classification context or $r_1, \ldots, r_p$ in the regression context. Sometimes, the notation $f_1, \ldots, f_p$, embracing both cases, will be employed. For $x \in \mathbb{R}^d$, let $\mathbf{c}(x) = (c_1(x), \ldots, c_p(x))$, $\mathbf{r}(x) = (r_1(x), \ldots, r_p(x))$ and $\mathbf{f}(x) = (f_1(x), \ldots, f_p(x))$. For ease of exposition, we will assume throughout that $c_1, \ldots, c_p$ and $r_1, \ldots, r_p$ do not depend on the sample $\mathcal{D}_n$.

Remark 2.1. *Using a simple sample-splitting device, our results extend to the case where the individual estimators depend on the sample. More specifically, we may split $\mathcal{D}_n$ into two independent parts $\mathcal{D}_{n_1}$ and $\mathcal{D}_{n_2}$, assuming that $f_1, \ldots, f_p$ are based on $\mathcal{D}_{n_1}$ alone, and using $\mathcal{D}_{n_2}$ for the combination step.*

The definition of our combined estimator is first introduced in the regression framework.

Let the function $g : \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}_+$ be such that $g(v_1, v_2) = K(v)$, where $v = (v_1, v_2) \in \mathbb{R}^{d+p}$.

Definition 2.1. *Suppose that we are given a set of initial regression estimators $r_1, \ldots, r_p$. The regression combined estimator $\mathcal{T}_n$ is defined by*

$$\mathcal{T}_n(x) = \frac{\sum_{i=1}^n Y_i K(\frac{Z_i^r - z^r}{\alpha})}{\sum_{i=1}^n K(\frac{Z_i^r - z^r}{\alpha})}$$

$$= \frac{\sum_{i=1}^n Y_i g\left(\frac{X_i - x}{\alpha}, \frac{\mathbf{r}(X_i) - \mathbf{r}(x)}{\beta}\right)}{\sum_{i=1}^n g\left(\frac{X_i - x}{\alpha}, \frac{\mathbf{r}(X_i) - \mathbf{r}(x)}{\beta}\right)},$$

where $Z_i^r = (X_{i1}, \ldots, X_{id}, \frac{\alpha}{\beta} r_1(X_i), \ldots, \frac{\alpha}{\beta} r_p(X_i))$, $i = 1, \ldots, n$, $z^r = (x_1, \ldots, x_d, \frac{\alpha}{\beta} r_1(x) \ldots, \frac{\alpha}{\beta} r_p(x))$.

To lighten the equations, we will sometimes use the notation $x \mapsto K_\alpha(x)$ to mean the function $x \mapsto K(\frac{x}{\alpha})$ and $(v_1, v_2) \mapsto g_{\alpha,\beta}(v_1, v_2)$ for $(v_1, v_2) \mapsto g(\frac{v_1}{\alpha}, \frac{v_2}{\beta})$.

We introduce the notation $\mathcal{T}_n^\star$, for the quantity defined by

$$\frac{\sum_{i=1}^n Y_i g\left(\frac{X_i - x}{\alpha}, \frac{\mathbf{f}(X_i) - \mathbf{f}(x)}{\beta}\right)}{n\mathbb{E}\left[g\left(\frac{X - x}{\alpha}, \frac{\mathbf{f}(X) - \mathbf{f}(x)}{\beta}\right)\right]}.$$

We now state the definition of the estimator in the context of classification.

3

Definition 2.2. *Suppose that we are given a set of initial classifiers* $c_1, \ldots, c_p$. *The combined classifier* $\mathcal{C}_n$ *be defined by*

$$\mathcal{C}_n(x) = \begin{cases} 0 & \text{if } \frac{\sum_{i=1}^n Y_i K_\alpha(Z_i^c - z^c)}{n \mathbb{E}[K_\alpha(Z^c - z^c)]} \leq \frac{\sum_{i=1}^n (1 - Y_i) K_\alpha(Z_i^c - z^c)}{n \mathbb{E}[K_\alpha(Z^c - z^c)]} \\ 1 & \text{otherwise} \end{cases}$$

$$= \begin{cases} 0 & \text{if } \frac{\sum_{i=1}^n Y_i g_{\alpha,\beta}(X_i - x, \mathbf{c}(X_i) - \mathbf{c}(x))}{n \mathbb{E}[g_{\alpha,\beta}(X - x, \mathbf{c}(X) - \mathbf{c}(x))]} \leq \frac{\sum_{i=1}^n (1 - Y_i) g_{\alpha,\beta}(X_i - x, \mathbf{c}(X_i) - \mathbf{c}(x))}{n \mathbb{E}[g_{\alpha,\beta}(X - x, \mathbf{c}(X) - \mathbf{c}(x))]} \\ 1 & \text{otherwise,} \end{cases}$$

where $Z_i^c = (X_{i1}, \ldots, X_{id}, \frac{\alpha}{\beta} c_1(X_i), \ldots, \frac{\alpha}{\beta} c_p(X_i))$, $i = 1, \ldots, n$, *and* $z^c = (x_1, \ldots, x_d, \frac{\alpha}{\beta} c_1(x) \ldots, \frac{\alpha}{\beta} c_p(x))$.

Remark 2.2. *The random vectors* $Z_1^r, \ldots, Z_n^r$ *form an i.i.d. sequence, and so is the sequence* $Z_1^c, \ldots, Z_n^c$.

Here, the exponent r or c will be dropped whenever there is no need to specify the classification or regression context.

Remark 2.3. *In the particular case where* K *can be expressed in function of the squared Euclidean norm, i.e. we have, for* $x \in \mathbb{R}^{d+p}$, $K(x) = H(\|x\|^2)$, *then the quantity* $K(\frac{Z-z}{\alpha}) = g\left(\frac{X-x}{\alpha}, \frac{\mathbf{r}(X) - \mathbf{r}(x)}{\beta}\right)$ *takes the somewhat more explicit form* $H\left(\left\|\frac{X-x}{\alpha}\right\|^2 + \left\|\frac{\mathbf{r}(X) - \mathbf{r}(x)}{\beta}\right\|^2\right)$. *This is the case for a Gaussian kernel for instance.*

3 Main results

We are now in a position to state the main results of the paper.

Theorem 3.1 (Regression case). *If* $\alpha \to 0$ *and* $n\alpha^d \beta^p \to \infty$ *as* $n \to \infty$, *then, for every* $\varepsilon > 0$, *there exists* n_0 *such that for* $n \geq n_0$, *the following exponential bound holds:*

$$\mathbb{P}\left(\int |\eta(x) - \mathcal{T}_n(x)| \mu(dx) > \varepsilon\right) \leq 2 \exp\left(-\frac{n\varepsilon^2}{32R^2}\right),$$

where $R \geq 0$ *is a constant depending on* K *and* $d + p$.

Let $L^\star$ denote the Bayes error and L_n the missclassification error of $\mathcal{C}_n$.

Theorem 3.2 (Classification case). *If* $\alpha \to 0$ *and* $n\alpha^d \beta^p \to \infty$ *as* $n \to \infty$, *then, for every* $\varepsilon > 0$, *there exists* n_0 *such that for* $n \geq n_0$, *the following exponential bound holds:*

$$\mathbb{P}(L_n - L^\star > \varepsilon) \leq 2 \exp\left(-\frac{n\varepsilon^2}{32R^2}\right),$$

where $R \geq 0$ *is a constant depending on* K *and* $d + p$.

These results may be seen as "combining" versions of the strong consistency results for kernel regression and the kernel classification rule, described in Devroye and Krzyżak (1989) (see also the books of Devroye et al. (1996) and Györfi et al. (2002)).

For Theorem 3.1 (regression context), let us write

$$\int |\eta(x) - \mathcal{T}_n(x)| \mu(dx) \leq \int |\eta(x) - \mathcal{T}_n^\star(x)| \mu(dx) + \int |\mathcal{T}_n^\star(x) - \mathcal{T}_n(x)| \mu(dx). \tag{3.1}$$

Thus, the result will be obtained by replacing, on the one hand, $\mathcal{T}_n$ by $\mathcal{T}_n^\star$, and by controlling, on the other hand, the error due to the difference between the two terms. This is done respectively in Lemma 3.1 and 3.2 below.

Lemma 3.1. *If* $\alpha \to 0$ *and* $n\alpha^d \beta^p \to \infty$ *as* $n \to \infty$, *then there exists* $R \geq 0$, *depending on* K *and* $d + p$, *such that for every* $\varepsilon > 0$, *if* n *is large enough,*

$$\mathbb{P}\left(\int |\eta(x) - \mathcal{T}_n^\star(x)| \mu(dx) > \varepsilon/2\right) \leq \exp\left(-\frac{n\varepsilon^2}{32R^2}\right).$$

The next lemma is devoted to the control of the difference between $\mathcal{T}_n$ and $\mathcal{T}_n^\star$.

Lemma 3.2. *If $\alpha \to 0$ and $n\alpha^d\beta^p \to \infty$ as $n \to \infty$, then there exists $R \geq 0$, depending on K and $d+p$, such that for every $\varepsilon > 0$, if n is large enough,*

$$\mathbb{P}\left(\int |\mathcal{T}_n^\star(x) - \mathcal{T}_n(x)|\mu(dx) > \varepsilon/2\right) \leq \exp\left(-\frac{n\varepsilon^2}{8R^2}\right).$$

For Theorem 3.2 (classification context), let us note that Theorem 2.3 in Devroye et al. (1996) may be adapted to obtain the next lemma. For the sake of completeness, it is proved in section 6.

Lemma 3.3. *The following upper bound holds:*

$$L_n - L^\star \leq \int \left|1 - \eta(x) - \frac{\sum_{i=1}^n K_\alpha(Z_i - z)}{n\mathbb{E}[K_\alpha(Z - z)]} + \mathcal{T}_n^\star(x)\right| \mu(dx) + \int |\eta(x) - \mathcal{T}_n^\star(x)|\mu(dx).$$

We deduce from this upper bound that

$$\mathbb{P}(L_n - L^\star > \varepsilon) \leq \mathbb{P}\left(\int \left|1 - \eta(x) - \frac{\sum_{i=1}^n K_\alpha(Z_i - z)}{n\mathbb{E}[K_\alpha(Z - z)]} + \mathcal{T}_n^\star(x)\right| \mu(dx) + \int |\eta(x) - \mathcal{T}_n^\star(x)|\mu(dx) > \varepsilon\right)$$

$$\leq \mathbb{P}\left(\int \left|1 - \eta(x) - \frac{\sum_{i=1}^n K_\alpha(Z_i - z)}{n\mathbb{E}[K_\alpha(Z - z)]} + \mathcal{T}_n^\star(x)\right| \mu(dx) > \varepsilon/2\right)$$

$$+ \mathbb{P}\left(\int |\eta(x) - \mathcal{T}_n^\star(x)|\mu(dx) > \varepsilon/2\right). \quad (3.2)$$

Consequently, Theorem 3.2 (classification context) follows from Lemma 3.1, applied to Y and to $1 - Y$.

To control the quantity $\int |\eta(x) - \mathcal{T}_n^\star(x)|\mu(dx)$ in order to prove Lemma 3.1, we may use the following decomposition, for $x \in \mathbb{R}^d$:

$$|\eta(x) - \mathcal{T}_n^\star(x)| = \mathbb{E}[|\eta(x) - \mathcal{T}_n^\star(x)|] + (|\eta(x) - \mathcal{T}_n^\star(x)| - \mathbb{E}[|\eta(x) - \mathcal{T}_n^\star(x)|]). \quad (3.3)$$

The term $\int \mathbb{E}[|\eta(x) - \mathcal{T}_n^\star(x)|]\mu(dx)$ will be studied first and then McDiarmid's inequality will be employed to handle the deviation $\int |\eta(x) - \mathcal{T}_n^\star(x)|\mu(dx) - \int \mathbb{E}[|\eta(x) - \mathcal{T}_n^\star(x)|]\mu(dx)$.
Let $\varepsilon' > 0$ and $f : \mathbb{R}^d \to [0, 1]$ (recall that Y equals 0 or 1 in the classification case and we consider bounded regression with $Y \in [0, 1]$) be a continuous function with compact support such that

$$\int |\eta(x) - f(x)|\mu(dx) < \varepsilon'. \quad (3.4)$$

The following inequality holds:

$$\mathbb{E}[|\eta(x) - \mathcal{T}_n^\star(x)|] \leq |\eta(x) - f(x)| + \left|f(x) - \frac{\mathbb{E}\left[f(X)g\left(\frac{X-x}{\alpha}, \frac{\mathbf{f}(X)-\mathbf{f}(x)}{\beta}\right)\right]}{\mathbb{E}\left[g\left(\frac{X-x}{\alpha}, \frac{\mathbf{f}(X)-\mathbf{f}(x)}{\beta}\right)\right]}\right|$$

$$+ \left|\frac{\mathbb{E}\left[f(X)g\left(\frac{X-x}{\alpha}, \frac{\mathbf{f}(X)-\mathbf{f}(x)}{\beta}\right)\right]}{\mathbb{E}\left[g\left(\frac{X-x}{\alpha}, \frac{\mathbf{f}(X)-\mathbf{f}(x)}{\beta}\right)\right]} - \mathbb{E}[\mathcal{T}_n^\star(x)]\right| + \mathbb{E}|\mathbb{E}[\mathcal{T}_n^\star(x)] - \mathcal{T}_n^\star(x)|. \quad (3.5)$$

We will derive an upper bound for the integral against μ of each term. The next result is proved in Section 6.

Proposition 3.1. *For a function f as defined in (3.4), the following statements hold.*

1. If $\alpha \to 0$ as $n \to \infty$, then

$$\limsup_{n \to \infty} \int \left|f(x) - \frac{\mathbb{E}\left[f(X)g\left(\frac{X-x}{\alpha}, \frac{\mathbf{f}(X)-\mathbf{f}(x)}{\beta}\right)\right]}{\mathbb{E}\left[g\left(\frac{X-x}{\alpha}, \frac{\mathbf{f}(X)-\mathbf{f}(x)}{\beta}\right)\right]}\right| \mu(dx) \leq \varepsilon'.$$

5

2. We have $\displaystyle\int \left| \frac{\mathbb{E}\left[f(X)g\left(\frac{X-x}{\alpha}, \frac{\mathbf{f}(X)-\mathbf{f}(x)}{\beta}\right)\right]}{\mathbb{E}\left[g\left(\frac{X-x}{\alpha}, \frac{\mathbf{f}(X)-\mathbf{f}(x)}{\beta}\right)\right]} - \mathbb{E}[\mathcal{T}_n^\star(x)] \right| \mu(dx) \leq R\varepsilon'$.

3. Let us write

$$\int \mathbb{E}[|\mathbb{E}[\mathcal{T}_n^\star(x)] - \mathcal{T}_n^\star(x)|]\mu(dx) = \int_{[-\ell,\ell]^d} \mathbb{E}[|\mathbb{E}[\mathcal{T}_n^\star(x)] - \mathcal{T}_n^\star(x)|]\mu(dx) + \int_{([-\ell,\ell]^d)^c} \mathbb{E}[|\mathbb{E}[\mathcal{T}_n^\star(x)] - \mathcal{T}_n^\star(x)|]\mu(dx).$$

Then,

$$\int_{[-\ell,\ell]^d} \mathbb{E}[|\mathbb{E}[\mathcal{T}_n^\star(x)] - \mathcal{T}_n^\star(x)|]\mu(dx) \leq C(n\alpha^d\beta^p)^{-1/2},$$

for some constant $C \geq 0$ depending on ℓ, the kernel K and the dimensions d and p. Moreover, if $\alpha \to 0$ as $n \to \infty$, then, for ℓ and n large enough,

$$\int_{([-\ell,\ell]^d)^c} \mathbb{E}[|\mathbb{E}[\mathcal{T}_n^\star(x)] - \mathcal{T}_n^\star(x)|]\mu(dx) \leq \varepsilon'.$$

The proof of this proposition rests upon the following important result, which extends the covering lemma of Devroye and Krzyżak (1989) to our context.

Lemma 3.4 (Covering lemma). *1. There exists $R \geq 0$, depending on K and $d + p$, such that*

$$\sup_u \int \frac{g\left(\frac{u-x}{\alpha}, \frac{\mathbf{f}(u)-\mathbf{f}(x)}{\beta}\right)}{\mathbb{E}\left[g\left(\frac{X-x}{\alpha}, \frac{\mathbf{f}(X)-\mathbf{f}(x)}{\beta}\right)\right]} \mu(dx) \leq R < +\infty.$$

2. $\forall \delta, \varepsilon$, there exists α_0 such that

$$\sup_{u, \alpha \leq \alpha_0} \int \frac{g\left(\frac{u-x}{\alpha}, \frac{\mathbf{f}(u)-\mathbf{f}(x)}{\beta}\right) \mathbf{1}_{\{\|x-u\| \geq \delta\}}}{\mathbb{E}\left[g\left(\frac{X-x}{\alpha}, \frac{\mathbf{f}(X)-\mathbf{f}(x)}{\beta}\right)\right]} \mu(dx) \leq \varepsilon.$$

4 Numerical Experiments

The section presents numerical experiments to illustrate the benefits of using the new combining approach. The classification case is illustrated with numerical simulations and the regression case with real operational data recorded from two applications: modeling of the electrical power consumption of an air compressor and modeling of the electricity production of different wind turbines in a wind farm.

Two aggregation strategies are run, the strategy by Mojirsheibani (1999) or Biau et al. (2016), which only combines output predictions, and our strategy, which combines the input distance and output predictions. In the sequel, the original approach is called *Cobra*, from the name of the R package implementing the method (Guedj, 2013), and our version is called *MixCobra*.

4.1 Classification

Figure 1 presents the 6 examples simulated to illustrate the classification performances. For each example, a data set of N observations is simulated, split into two classes (with half of the observations in each class). All the figures are depicted with $N = 1000$ observations, to get graphics reflecting more accurately the distribution; however, to be more realistic with respect to available data in most applications, only $N = 200$ observations (100 per class) are used to evaluate the performances. The first four graphs present mixtures of uniform or Gaussian distributions. The two last graphs represent respectively two concentric circles and two interlocked spirals (see annex for details on simulated data).

Parametric and non parametric machines are here introduced for classification. For parametric methods, linear discriminant analysis (lda) and logistic regression (logit) are used. For nonparametric methods, k-nearest neighbors (knn) (with $k = 5$), support vector machines (svm), classification and regression trees (cart) and random forest (rf) are introduced. In this classification framework, the *Cobra* aggregation
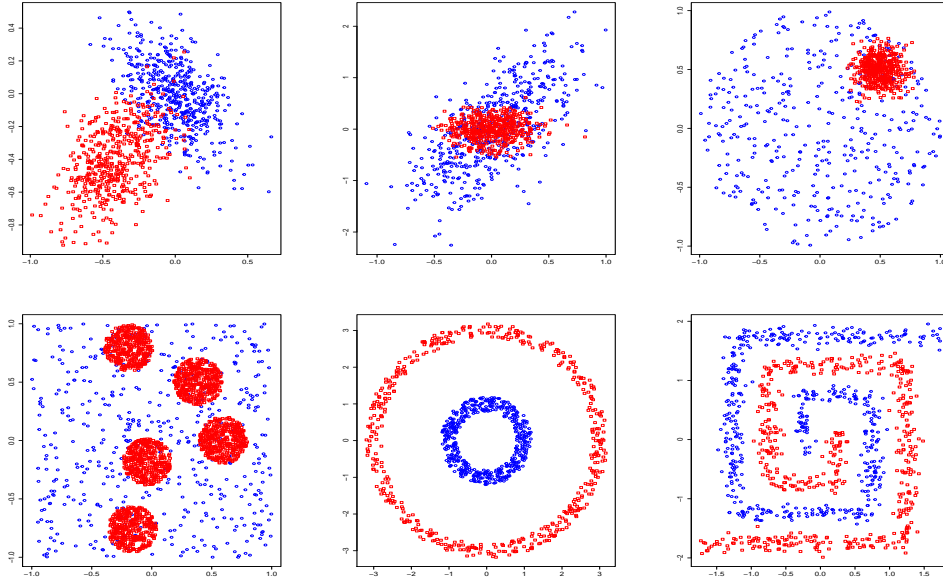
Figure 1: Simulated data for classification. From left to right, from up to bottom: "gauss", "comete", "nuclear", "spot", "circles" and "spirals" examples.

procedure is performed using all the available machines, meaning that the consensus is required for all machines. For each simulated example, the performances of each machine are computed using $K = 100$ repetitions. For each repetition, $N = 200$ observations are generated: 75% randomly chosen observations are used to calibrate the model and the remaining 25% to compute the test performances.

One first interesting question is to analyze whenever a specific machine always shows the best performance over all the repetitions. During an off-line study, if a practitioner observes that a given machine performs always the best, then the choice of the model to use is easy to state and an aggregation procedure relying on all machines may not be useful in this particularly case. Table 1 shows, for each machine, the number of runs where this machine provides the best model (i.e. has the best performance). For each simulated example (each column), the sum of the rows corresponding to the different methods equals the total number of repetitions ($K = 100$). We observe that, except for the "circles" example for which the knn machine outperforms all the others, almost every machine wins at least once over all the runs. For all simulated examples except the "circles" example, an aggregation procedure may be interesting.

	gauss	comete	nuclear	spot	circles	spirals
lda	75	9	35	4	0	2
logit	13	0	13	2	0	0
knn	5	46	27	44	100	76
svm	5	20	13	16	0	9
cart	0	18	8	19	0	2
bag	1	5	4	12	0	11
rf	1	2	0	3	0	0

Table 1: Number of runs where every machine provides the best model for the simulated classification examples (smallest test error).

Table 2 presents the average performances for all the classification examples. As expected for the "gauss" example corresponding to the well separated mixture of two Gaussian distributions, both parametric and non parametric machines perform well; otherwise, the nonparametric methods perform in general better. Concerning the aggregation procedures, *Cobra* and *MixCobra* succeed in yielding very satisfactory performances. When local information brings indirectly strong knowledge on the class label, as for the "circles"

7

	gauss	comete	nuclear	spot	circles	spirals
lda	7.0 (8.7)	51.0 (15.6)	12.9 (10.6)	17.6 (7.2)	52.5 (16.5)	34.3 (15.8)
logit	7.5 (8.2)	51.1 (15.6)	9.6 (9.0)	17.8 (7.3)	52.4 (16.3)	34.4 (16.0)
knn	7.0 (8.1)	25.0 (14.5)	6.0 (7.1)	9.2 (6.6)	0.0 (0.0)	4.1 (6.4)
svm	8.7 (9.0)	25.9 (14.6)	5.9 (8.0)	10.3 (6.7)	1.8 (4.1)	12.2 (10.8)
cart	7.2 (8.3)	23.0 (14.8)	4.0 (6.0)	7.2 (5.8)	0.0 (0.0)	17.8 (13.7)
bag	8.7 (8.8)	28.3 (15.8)	5.1 (7.5)	7.5 (5.5)	0.5 (2.2)	3.0 (5.4)
rf	8.9 (9.1)	27.5 (15.3)	4.7 (6.8)	7.4 (5.5)	0.3 (1.7)	2.2 (5.0)
Cobra	8.6 (8.6)	25.8 (15.2)	8.2 (8.9)	8.5 (5.9)	1.1 (3.5)	6.2 (8.3)
Mixcobra	7.7 (9.2)	28.2 (15.4)	5.4 (6.7)	7.2 (5.8)	0.3 (1.7)	2.8 (6.2)

Table 2: Average classification error (and standard deviation into brackets) for the different machines and both aggregation methods *Cobra* and *MixCobra*, computed for the simulated classification examples (1 unit = 10^{-2}).

and "spirals" examples, *MixCobra* outperforms compared to *Cobra*. This may be explained by the fact that the input part of *MixCobra* weights behaves like a nonparametric kernel-like method, performing well in such cases, as shown by the low error of the knn classifier.

4.2 Regression

For the regression framework, we use the eponymous R package COBRA available on CRAN (Guedj, 2013) to compute the performance results for *Cobra* aggregation. It should be noted that this package implements the procedure only for the regression framework and not for classification (so, it was not possible to use the package in the classification framework). In the R package COBRA, two parameters, called hereafter δ and γ, drive the numerical results of the aggregation strategy (see Guedj, 2013). The smoothing parameter δ governs the consensus, insofar that it sets a threshold for the distance between the prediction $r_m(x)$ for a new observation x and the prediction $r_m(x_i)$ for an observation x_i in the training set: for every machine m,

$$|r_m(x_i) - r_m(x)| \leq \delta.$$

The collective estimator T is a weighted average of the outputs of the training data set: for $x \in \mathbb{R}^d$,

$$T(x) = \sum_{i=1}^{n} W_i(x) Y_i.$$

The value of the weight $W_i(x)$ attributed to each output observation Y_i, $1 \leq i \leq n$, of the training data set depends on the consensus, possibly computed only for a certain proportion γ of machines. This approach will be referred to as "Cobra Adaptive method" by opposition to the "Cobra Fixed method" when the number of machines is fixed.

$$W_i(x) = \frac{\mathbf{1}_{\{\sum_{m=1}^{p} \mathbf{1}_{\{|r_m(x) - r_m(X_i)| \leq \delta\}} \geq p\gamma\}}}{\sum_{j=1}^{\ell} \mathbf{1}_{\{\sum_{m=1}^{p} \mathbf{1}_{\{|r_m(x) - r_m(X_j)| \leq \delta\}} \geq p\gamma\}}}.$$

A too small δ value leads to many zero weights and discards a large number of data points, whereas a large value provides just an average of all the predictions. The parameter γ drives the proportion of machines to be actually considered among all initial machines. With no a priori information on the performances of the different machines, both parameters δ and γ are quite hard to choose. Practically, in the R COBRA package, these parameters are computed by cross-validation, using a subset of the observations in the training data set.

Concerning *MixCobra*, two parameters need also to be chosen, as in the initial *Cobra* method. The α parameter drives the influence of the inputs (through the proximity between a new input observation and the input observations of the training data set) and the β parameter drives the consensus of the predictions. It may be interesting to note that, in *MixCobra*, the two parameters play more symmetric roles than the parameters of *Cobra*. In practice, as in the COBRA package, both parameters are chosen by cross-validation.

The following paragraphs illustrate, in two industrial applications, the benefit of combining several regressors relying on a consensus notion using the *Cobra* and the *MixCobra* method.

Modeling Air Compressors   Air compressors are crucial elements of production and distribution of air gases. Compressors are instrumented with sensors that provide valuable information on the compressor status. Power consumption is measured on the electric motor driving the compressor and other information as flow, input pressure, output pressure, cooling water temperature and air temperature are also recorded. Monitoring compressor power consumption is essential to an Air Separation Unit safe and efficient operation (Cadet et al., 2005). In this study, $p = 8$ regression machines are used to model the air compressor consumption: linear regression model (lm), regression tree (cart), bagging regression tree (bag), random forest (rf), Support Vector Machines (svm) and k-nearest neighbors (knn) with different values of k ($k = 2, 5, 10$).

The data set contains $N = 1000$ hourly observations of a working air compressor. Each variable is first transformed to have zero mean and unit variance. The performances are here computed by cross-validation using $K = 100$ replications. For each replication, 2/3 of the observations are randomly chosen to calibrate the machines and 1/3 of the remaining observations are used to estimate the performances of the different machines. The performances on the test sets are presented in Table 3: for each case, the mean squared error is computed. We observed that the cart bagging and the svm machines provide in average the smallest error. Moreover, we observe that the best machine (i.e. the machine providing the smallest test performance error) changes from run to run. For $K = 100$ replications, the best machine is alternatively the cart bagging machine (49 runs), the svm machine (49 runs) or the lm machine (2 runs). In this case, aggregation methods seem particularly well appropriate. Figure 2 shows the boxplots computed with the different machines and the 3 aggregation algorithms: *Cobra* with all machines (CobraF), *Cobra* with an adaptive number of machines (CobraA) and *MixCobra*. As expected, the Cobra algorithm with all machines yields in average the worst performance among the three aggregation techniques. Choosing adaptively the number of machines (CobraA) allows to discard a possibly bad machine and, so, improves the performance of Cobra. We observe that *MixCobra* provides, in average, the best performance, associated with a low standard deviation compared to the CobraF and CobraA aggregation methods.

	mean	std
lm	11.64	0.00
cart	26.43	0.02
bag	10.75	0.01
rf	20.05	0.02
svm	10.77	0.01
knn 2	18.35	0.01
knn 5	18.79	0.01
knn 10	20.49	0.01
Cobra Fixed	21.63	0.06
Cobra Adaptive	14.81	0.05
MixCobra	10.85	0.01

Table 3: Average test performances of the machines and aggregation methods for the air compressor equipment (1 unit $= 10^{-2}$).

Modeling Wind Turbines   The second application aims at modeling 6 different wind turbines on a wind plant in France. Each turbine is described by 5 variables, representing half-hourly information. The production of electricity is here the target variable. The explanatory variables are the wind power, the wind direction (sine, cosine), and the temperature. For each wind turbine, each variable is first transformed to have zero mean and unit variance.

Table 4 presents the performances obtained in modeling the 6 wind turbines (W_k, $1 \leq k \leq 6$), for $N = 1000$ operational data points, using the same $p = 8$ set of learning machines as in the previous paragraph and the corresponding *Cobra* and *MixCobra* aggregations. For all wind turbines, the cart
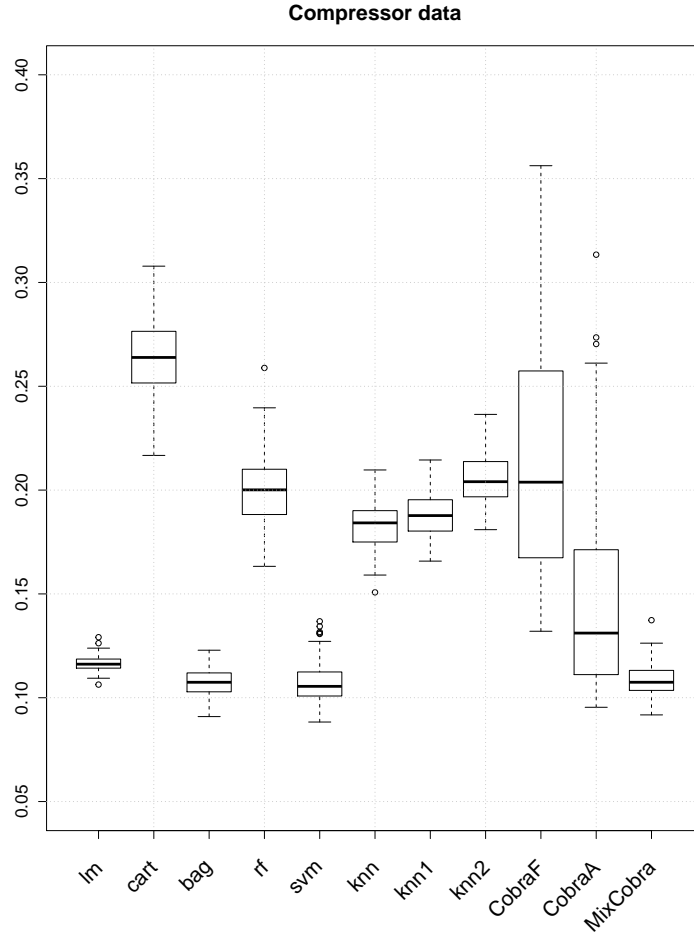
Figure 2: Performances of *MixCobra* and *Cobra* for a fixed or adaptive number of machines. The machines are lm, cart, bag, rf, svm, knn (3 machines $k = 2, 5, 10$).

bagging machine provides in average the best performances. However, we study in detail, for each wind turbine, the performances obtained for each run over the $K = 100$ repetitions. Table 5 presents the number of runs where every given machine provides the smallest error. For the $K = 100$ repetitions, the model which provides the smallest error is never the same, and is alternatively either cart bagging or svm. We observe that the proportion differs for each wind turbine. In this case, aggregation methods, and particularly *MixCobra*, are especially interesting, providing a unique global approach to model all wind turbines, as repetitions show that the best model is never the same.

	W1	W2	W3	W4	W5	W6
lm	14.39 (0.9)	18.22 (1.2)	16.65 (1.2)	15.72 (1.7)	15.16 (1.2)	16.97 (1.8)
cart	18.39 (1.7)	19.53 (1.8)	19.46 (1.9)	19.10 (2.0)	19.02 (1.7)	19.22 (1.8)
bag	9.13 (0.7)	9.95 (1.2)	9.61 (1.0)	9.35 (0.7)	9.75 (0.8)	10.17 (1.1)
rf	11.43 (0.9)	16.32 (2.7)	15.41 (2.2)	13.89 (1.8)	15.12 (2.0)	14.87 (1.7)
svm	9.28 (1.0)	13.55 (3.5)	12.72 (3.5)	11.12 (2.3)	12.01 (2.1)	12.00 (2.6)
knn 2	13.00 (0.6)	15.59 (1.5)	14.65 (1.1)	14.62 (1.0)	15.75 (1.2)	15.70 (1.1)
knn 5	12.12 (0.7)	15.26 (1.8)	13.90 (1.1)	13.60 (1.0)	14.95 (1.3)	14.75 (1.3)
knn 10	12.69 (0.8)	16.78 (1.8)	15.00 (1.3)	14.41 (1.2)	15.93 (1.4)	15.66 (1.5)
Cobra Fixed	18.83 (4.5)	29.55 (10.6)	31.76 (10.9)	24.01 (8.5)	24.73 (9.2)	23.92 (8.4)
Cobra Adaptive	14.15 (5.5)	17.76 (8.1)	17.38 (9.8)	16.43 (7.4)	17.71 (9.0)	15.29 (7.2)
MixCobra	9.02 (0.6)	10.47 (1.4)	10.04 (1.0)	9.45 (0.8)	10.37 (0.9)	10.57 (1.1)

Table 4: Average test performances (and standard deviation into brackets) for the wind turbines (1 unit $= 10^{-2}$).

	W1	W2	W3	W4	W5	W6
lm	0	0	0	0	0	0
cart	0	0	0	0	0	0
bag	52	89	88	76	93	77
rf	0	0	0	0	0	0
svm	48	11	12	24	7	23
knn 2	0	0	0	0	0	0
knn 5	0	0	0	0	0	0
knn 10	0	0	0	0	0	0

Table 5: Number of runs where every machine provides the best model for the wind turbines (smallest test error).

Increasing the dimension or the number of estimators   Since our combining method relies on a kernel rule, it is expected to be affected in some sense by the curse of dimensionality. However, an interesting feature of our approach is that the term based on the distance between the entries, $(X-x)/\alpha$, and the term involving the preliminary estimations, $(\mathbf{r}(X) - \mathbf{r}(x))/\beta$, are not affected the same way, since the "dimension" in not the same in both cases. Indeed, in the first case, the dimension is the actual dimension d of the space $\mathbb{R}^d$ containing the entries, whereas in the second case, the role of the dimension is played by the number p of individual estimators. In fact, the final combined estimator shows an interesting flexibility through the calibration of α and β. When the dimension d increases, the method will give more weight to the combining part, which is not affected in itself by increasing d and may only be affected through the particular behavior of the initial estimators considered. Conversely, for reasonable values of d, the effect of an increase of the number p of estimators may be balanced by the distance term between the entries.

	E1 $d = 6$	E2, $d = 11$	E3, $d = 16$	E4, $d = 21$	E5, $d = 26$
lm	11.6 (0.4)	11.7 (0.3)	11.8 (0.4)	11.9 (0.4)	11.8 (0.3)
cart	26.4 (1.9)	26.0 (2.0)	26.6 (1.9)	26.8 (2.2)	26.2 (1.8)
bag	10.7 (0.6)	11.4 (0.7)	12.2 (0.7)	12.3 (0.6)	12.6 (0.6)
rf	20.1 (1.6)	17.0 (1.3)	17.0 (1.5)	18.8 (1.4)	18.2 (1.7)
svm	10.8 (1.0)	14.8 (1.1)	16.4 (0.7)	16.6 (0.7)	16.2 (0.7)
knn 2	18.4 (1.2)	41.3 (2.1)	52.9 (2.3)	60.3 (2.8)	65.1 (3.4)
knn 5	18.8 (1.0)	36.5 (1.9)	45.5 (1.8)	51.9 (2.5)	56.6 (3.1)
knn10	20.5 (1.1)	36.3 (2.1)	44.5 (1.9)	50.3 (2.1)	54.4 (3.6)
MixCobra	10.8 (0.7)	13.2 (0.9)	13.9 (0.8)	14.5 (0.9)	14.1 (0.4)

Table 6: Performances with the initial air compressor data embedded into high dimensional spaces, adding respectively 5, 10, 15, 20, 25 random variables.

To study the effect of the inputs dimensionality, the original compressor data are embedded in successive high dimensional spaces of size $p = 11, 16, 21, 26$, by artificially adding to the 6 initial variables 5, 10, 15, 20 independent random variables uniformly distributed.

Table 6 shows the performances computed by cross validation using $K = 100$ replications. As previously, for each replication, 2/3 of the observations are randomly chosen to calibrate the machines and 1/3 of the remaining observations are used to estimate the performances for the different machines. We observe, as expected, that the performances of the knn machines strongly deteriorate as the dimension increases. The performances of the random forest and the svm machines are also impacted with an increase of the dimension but to a lower extent. The performances of the lm machines and the cart machines are stable. Table 6 shows that *MixCobra* performs well, even if the performances of some machines decrease, and adapts to the increase of dimensionality.

5 Perspectives

It is worth pointing out that further definitions of combining estimate linking an aggregation part and a distance between entries part could be investigated. A point of view that seem very appealing is to employ general multivariate kernels and allow the bandwidth to be different depending on the direction, that is the machine and also the direction in $\mathbb{R}^d$. The advantage of this method would be a great flexibility through his data-driven adaptive nature, whereas the increase of the number of parameters necessary for the estimation of a bandwidth matrix would represent a challenge both for estimation and computational aspects.

In our future research work, we plan to study more specifically the high dimensional case. In this framework, it could be interesting to compute more appropriate distances between entries. In particular, a promising strategy could be to try to bring into play the correlations between the output and the different explanatory variables.

6 Proofs

6.1 Proof of Lemma 3.3

For any classifier ψ, we may write

$$
\begin{aligned}
\mathbb{P}(\psi(X) \neq Y | X = x) &= 1 - \mathbb{P}(\psi(X) = Y | X = x) \\
&= 1 - \mathbb{P}(\psi(X) = 1, Y = 1 | X = x) - \mathbb{P}(\psi(X) = 0, Y = 0 | X = x) \\
&= 1 - \mathbf{1}_{\{\psi(X)=1\}} \mathbb{P}(Y = 1 | X = x) - \mathbf{1}_{\{\psi(X)=0\}} \mathbb{P}(Y = 0 | X = x) \\
&= 1 - \mathbf{1}_{\{\psi(X)=1\}} \eta(x) - \mathbf{1}_{\{\psi(X)=0\}} (1 - \eta(x)).
\end{aligned}
$$

Then, applying this to $\mathcal{C}_n$ and $\psi^\star$, we obtain

$$
\begin{aligned}
L_n - L^\star &= \mathbb{P}(\mathcal{C}_n(X) \neq Y | X = x) - \mathbb{P}(\psi^\star(X) \neq Y | X = x) \\
&= 1 - \mathbf{1}_{\{\mathcal{C}_n(X)=1\}}\eta(x) - \mathbf{1}_{\{\mathcal{C}_n(X)=0\}}(1 - \eta(x)) - 1 + \mathbf{1}_{\{\psi^\star(X)=1\}}\eta(x) + \mathbf{1}_{\{\psi^\star(X)=0\}}(1 - \eta(x)) \\
&= \eta(x)(-\mathbf{1}_{\{\mathcal{C}_n(X)=1\}} + \mathbf{1}_{\{\mathcal{C}_n(X)=0\}} + \mathbf{1}_{\{\psi^\star(X)=1\}} - \mathbf{1}_{\{\psi^\star(X)=0\}}) - \mathbf{1}_{\{\mathcal{C}_n(X)=0\}} + \mathbf{1}_{\{\psi^\star(X)=0\}} \\
&= 2\eta(x)(\mathbf{1}_{\{\mathcal{C}_n(X)=0\}} - \mathbf{1}_{\{\psi^\star(X)=0\}}) - \mathbf{1}_{\{\mathcal{C}_n(X)=0\}} + \mathbf{1}_{\{\psi^\star(X)=0\}} \\
&= (2\eta(x) - 1)(\mathbf{1}_{\{\mathcal{C}_n(X)=0\}} - \mathbf{1}_{\{\psi^\star(X)=0\}}).
\end{aligned}
$$

This implies

$$
L_n - L^\star = \int |2\eta(x) - 1| \mathbf{1}_{\{\psi^\star(x) \neq \mathcal{C}_n(x)\}} \mu(dx).
$$

The event $\{\psi^\star(x) \neq \mathcal{C}_n(x)\}$ corresponds to either one of the two following situations:

- $\psi^\star(x) = 0$, $\mathcal{C}_n(x) = 1$, which can be rewritten $\eta(x) \leq 1/2$ and $2\mathcal{T}_n^\star(x) - \frac{\sum_{i=1}^n K_\alpha(Z_i - z)}{n\mathbb{E}[K_\alpha(Z_i - z)]} > 0$. In this case, we have

$$
\begin{aligned}
|2\eta(x) - 1| &= 1 - 2\eta(x) \\
&= 1 - \frac{\sum_{i=1}^n K_\alpha(Z_i - z)}{n\mathbb{E}[K_\alpha(Z - z)]} + \frac{\sum_{i=1}^n K_\alpha(Z_i - z)}{n\mathbb{E}[K_\alpha(Z - z)]} - 2\eta(x) \\
&< 1 - \frac{\sum_{i=1}^n K_\alpha(Z_i - z)}{n\mathbb{E}[K_\alpha(Z - z)]} + 2\mathcal{T}_n^\star(x) - 2\eta(x) \\
&\leq \left| 1 - \eta(x) - \frac{\sum_{i=1}^n K_\alpha(Z_i - z)}{n\mathbb{E}[K_\alpha(Z - z)]} + \mathcal{T}_n^\star(x) \right| + |\eta(x) - \mathcal{T}_n^\star(x)|.
\end{aligned}
$$

- $\psi^\star(x) = 1$, $\mathcal{C}_n(x) = 0$, which means that $\eta(x) > 1/2$ and $2\mathcal{T}_n^\star(x) - \frac{\sum_{i=1}^n K_\alpha(Z_i - z)}{n\mathbb{E}[K_\alpha(Z - z)]} \leq 0$. We have

$$
\begin{aligned}
|2\eta(x) - 1| &= 2\eta(x) - 1 \\
&= 2\eta(x) - \frac{\sum_{i=1}^n K_\alpha(Z_i - z)}{n\mathbb{E}[K_\alpha(Z - z)]} + \frac{\sum_{i=1}^n K_\alpha(Z_i - z)}{n\mathbb{E}[K_\alpha(Z - z)]} - 1 \\
&< 2\eta(x) + \frac{\sum_{i=1}^n K_\alpha(Z_i - z)}{n\mathbb{E}[K_\alpha(Z - z)]} - 2\mathcal{T}_n^\star(x) - 1 \\
&\leq \left| 1 - \eta(x) - \frac{\sum_{i=1}^n K_\alpha(Z_i - z)}{n\mathbb{E}[K_\alpha(Z - z)]} + \mathcal{T}_n^\star(x) \right| + |\eta(x) - \mathcal{T}_n^\star(x)|.
\end{aligned}
$$

Hence,

$$
L_n - L^\star \leq \int \left| 1 - \eta(x) - \frac{\sum_{i=1}^n K_\alpha(Z_i - z)}{n\mathbb{E}[K_\alpha(Z - z)]} + \mathcal{T}_n^\star(x) \right| \mu(dx) + \int |\eta(x) - \mathcal{T}_n^\star(x)| \mu(dx).
$$

6.2   Proof of the covering lemma 3.4

1. We recover $\mathbb{R}^{d+p}$ with balls with radius $\rho/2$ and centers $v_1, \ldots, v_j, \ldots$. Each element of $\mathbb{R}^{d+p}$ is in a finite number of such balls. This number depends on the dimension $d + p$.

   For the first bound, we search for an upper bound for $g\left(\frac{u-x}{\alpha}, \frac{\mathbf{f}(u)-\mathbf{f}(x)}{\beta}\right)$, and a lower bound for $\mathbb{E}\left[g\left(\frac{X-x}{\alpha}, \frac{\mathbf{f}(X)-\mathbf{f}(x)}{\beta}\right)\right]$ for every x such that $\left\|\frac{z-t}{\alpha} - v_j\right\| \leq \frac{\rho}{2}$. Here, we use the notation $z = \left(x_1, \ldots, x_d, \frac{\alpha}{\beta}f_1(x) \ldots, \frac{\alpha}{\beta}f_p(x)\right)$ and $t = \left(u_1, \ldots, u_d, \frac{\alpha}{\beta}f_1(u) \ldots, \frac{\alpha}{\beta}f_p(u)\right)$.

   For the upper bound, we note that the element of $\mathbb{R}^{d+p}$

$$
\frac{z - t}{\alpha} = \left(\frac{x - u}{\alpha}, \frac{\mathbf{f}(x) - \mathbf{f}(u)}{\beta}\right) = \frac{1}{\alpha}\left(x_1 - u_1, \ldots, x_d - u_d, \frac{\alpha}{\beta}f_1(x) - \frac{\alpha}{\beta}f_1(u), \ldots, \frac{\alpha}{\beta}f_p(x) - \frac{\alpha}{\beta}f_p(u)\right)
$$

belongs to a finite number of balls of radius $\rho/2$ and center one of the v_j's. We are going to upper bound the quantity of interest by the sum of the suprema over these different balls. The fact that $\frac{z-t}{\alpha}$ belongs to such a ball (with a center v_j) means that $\left\|\frac{z-t}{\alpha} - v_j\right\| \le \frac{\rho}{2}$. We may write

$$g\left(\frac{u-x}{\alpha}, \frac{\mathbf{f}(u)-\mathbf{f}(x)}{\beta}\right)$$

$$\le \sum_{j=1}^{+\infty} \sup_{x', \left\|\frac{z'-t}{\alpha}-v_j\right\|\le\frac{\rho}{2}} g\left(\frac{u-x'}{\alpha}, \frac{\mathbf{f}(u)-\mathbf{f}(x')}{\beta}\right) \mathbf{1}_{\left\{\left\|\frac{z-t}{\alpha}-v_j\right\|\le\frac{\rho}{2}\right\}}(x).$$

For the lower bound, we use the property $K(z) \ge c\mathbf{1}_{B(0,\rho)}(z)$. We have

$$\mathbb{E}\left[g\left(\frac{X-x}{\alpha}, \frac{\mathbf{f}(X)-\mathbf{f}(x)}{\beta}\right)\right] = \mathbb{E}\left[K\left(\frac{z-Z}{\alpha}\right)\right]$$

$$\ge c\mathbb{E}\left[\mathbf{1}_{\left\{\left\|\frac{z-Z}{\alpha}\right\|\le\rho\right\}}\right]$$

$$\ge c\mathbb{P}\left(\left\|\frac{z-Z}{\alpha}\right\| \le \rho\right).$$

Now, if $\left\|\frac{z-t}{\alpha} - v_j\right\| \le \frac{\rho}{2}$, let us check that

$$\left\|\frac{Z-t}{\alpha} - v_j\right\| \le \frac{\rho}{2}$$

implies that

$$\left\|\frac{Z-z}{\alpha}\right\| \le \rho.$$

We have

$$\left\|\frac{z-Z}{\alpha}\right\| = \left\|\frac{z-t+t-Z}{\alpha} - v_j + v_j\right\|$$

$$\le \left\|\frac{z-t}{\alpha} - v_j\right\| + \left\|\frac{t-Z}{\alpha} + v_j\right\|$$

$$\le \rho.$$

Thus, when $\left\|\frac{z-t}{\alpha} - v_j\right\| \le \frac{\rho}{2}$, we obtain

$$\mathbb{E}\left[g\left(\frac{X-x}{\alpha}, \frac{\mathbf{f}(X)-\mathbf{f}(x)}{\beta}\right)\right] \ge c\mathbb{P}\left(\left\|\frac{Z-t}{\alpha} - v_j\right\| \le \frac{\rho}{2}\right).$$

Consequently, we have

$$\int \frac{g\left(\frac{u-x}{\alpha}, \frac{\mathbf{f}(u)-\mathbf{f}(x)}{\beta}\right)}{\mathbb{E}\left[g\left(\frac{X-x}{\alpha}, \frac{\mathbf{f}(X)-\mathbf{f}(x)}{\beta}\right)\right]} \mu(dx)$$

$$\leq \frac{1}{c} \sum_{j=1}^{+\infty} \frac{\int \sup_{x', \left\|\frac{z'-t}{\alpha}-v_j\right\| \leq \frac{\rho}{2}} g\left(\frac{u-x'}{\alpha}, \frac{\mathbf{f}(u)-\mathbf{f}(x')}{\beta}\right) \mathbf{1}_{\left\{\left\|\frac{z-t}{\alpha}-v_j\right\| \leq \frac{\rho}{2}\right\}}(x) \mu(dx)}{\mathbb{P}\left(\left\|\frac{Z-t}{\alpha}-v_j\right\| \leq \frac{\rho}{2}\right)}$$

$$\leq \frac{1}{c} \sum_{j=1}^{+\infty} \frac{\sup_{x', \left\|\frac{z'-t}{\alpha}-v_j\right\| \leq \frac{\rho}{2}} g\left(\frac{u-x'}{\alpha}, \frac{\mathbf{f}(u)-\mathbf{f}(x')}{\beta}\right) \mathbb{P}\left(\left\|\frac{Z-t}{\alpha}-v_j\right\| \leq \frac{\rho}{2}\right)}{\mathbb{P}\left(\left\|\frac{Z-t}{\alpha}-v_j\right\| \leq \frac{\rho}{2}\right)}$$

$$\leq \frac{1}{c} \sum_{j=1}^{+\infty} \sup_{x', \left\|\frac{z'-t}{\alpha}-v_j\right\| \leq \frac{\rho}{2}} g\left(\frac{u-x'}{\alpha}, \frac{\mathbf{f}(u)-\mathbf{f}(x')}{\beta}\right)$$

$$= \frac{1}{c} \sum_{j=1}^{+\infty} \sup_{x', \left\|\frac{z'-t}{\alpha}-v_j\right\| \leq \frac{\rho}{2}} K\left(\frac{t-z'}{\alpha}\right)$$

$$\leq \frac{1}{c} \sum_{j=1}^{+\infty} \sup_{s \in B(v_j, \rho/2)} K(s)$$

$$\leq \frac{1}{c} \frac{1}{\lambda_{d+p}\left(B(0, \rho/2)\right)} \sum_{j=1}^{+\infty} \int_{B(v_j, \rho/2)} \sup_{s \in B(v_j, \rho/2)} K(s) dt$$

$$\leq \frac{1}{c} \frac{1}{\lambda_{d+p}\left(B(0, \rho/2)\right)} \sum_{j=1}^{+\infty} \int_{B(v_j, \rho/2)} \sup_{s \in B(t, \rho)} K(s) dt$$

$$\leq \frac{1}{c} \frac{k_{d+p}}{\lambda_{d+p}\left(B(0, \rho/2)\right)} \int \sup_{s \in B(t, \rho)} K(s) dt = R.$$

Here, λ_{d+p} denotes the Lebesgue measure of $\mathbb{R}^{d+p}$ and k_{d+p} stands for the number of covering balls containing a certain element of $\mathbb{R}^{d+p}$: integrating over all such balls consists in integrating over $\mathbb{R}^{d+p}$ and multiplying by the number of balls k_{d+p}.

Finally, we use that the kernel is regular.

2. For the second result, we use the same upper and lower bounds.

   We write

   $$g\left(\frac{u-x}{\alpha}, \frac{\mathbf{f}(u)-\mathbf{f}(x)}{\beta}\right) \mathbf{1}_{\{\|x-u\| \geq \delta\}}$$

   $$\leq \sum_{j=1}^{+\infty} \sup_{x', \left\|\frac{z'-t}{\alpha}-v_j\right\| \leq \frac{\rho}{2}} g\left(\frac{u-x'}{\alpha}, \frac{\mathbf{f}(u)-\mathbf{f}(x')}{\beta}\right) \mathbf{1}_{\{\|x'-u\| \geq \delta\}} \mathbf{1}_{\left\{\left\|\frac{z-t}{\alpha}-v_j\right\| \leq \frac{\rho}{2}\right\}}(x).$$

As above, when $\left\|\frac{z-t}{\alpha}-v_j\right\| \leq \frac{\rho}{2}$,

$$\mathbb{E}\left[g\left(\frac{X-x}{\alpha}, \frac{\mathbf{f}(X)-\mathbf{f}(x)}{\beta}\right)\right] \geq c\mathbb{P}\left(\left\|\frac{Z-t}{\alpha}-v_j\right\| \leq \frac{\rho}{2}\right).$$

Then,

$$\int \frac{g\left(\frac{u-x}{\alpha}, \frac{\mathbf{f}(u)-\mathbf{f}(x)}{\beta}\right) \mathbf{1}_{\{\|x-u\| \geq \delta\}}}{\mathbb{E}\left[g\left(\frac{X-x}{\alpha}, \frac{\mathbf{f}(X)-\mathbf{f}(x)}{\beta}\right)\right]} \mu(dx)$$

$$\leq \frac{1}{c} \sum_{j=1}^{+\infty} \frac{\int \sup_{x', \left\|\frac{z'-t}{\alpha}-v_j\right\| \leq \frac{\rho}{2}} g\left(\frac{u-x'}{\alpha}, \frac{\mathbf{f}(u)-\mathbf{f}(x')}{\beta}\right) \mathbf{1}_{\{\|x'-u\| \geq \delta\}} \mathbf{1}_{\left\{\left\|\frac{z-t}{\alpha}-v_j\right\| \leq \frac{\rho}{2}\right\}}(x) \mu(dx)}{\mathbb{P}\left(\left\|\frac{Z-t}{\alpha}-v_j\right\| \leq \frac{\rho}{2}\right)}$$

$$\leq \frac{1}{c} \sum_{j=1}^{+\infty} \frac{\sup_{x', \left\|\frac{z'-t}{\alpha}-v_j\right\| \leq \frac{\rho}{2}} g\left(\frac{u-x'}{\alpha}, \frac{\mathbf{f}(u)-\mathbf{f}(x')}{\beta}\right) \mathbf{1}_{\{\|x'-u\| \geq \delta\}} \mathbb{P}\left(\left\|\frac{Z-t}{\alpha}-v_j\right\| \leq \frac{\rho}{2}\right)}{\mathbb{P}\left(\left\|\frac{Z-t}{\alpha}-v_j\right\| \leq \frac{\rho}{2}\right)}$$

$$\leq \frac{1}{c} \sum_{j=1}^{+\infty} \sup_{x', \left\|\frac{z'-t}{\alpha}-v_j\right\| \leq \frac{\rho}{2}} g\left(\frac{u-x'}{\alpha}, \frac{\mathbf{f}(u)-\mathbf{f}(x')}{\beta}\right) \mathbf{1}_{\{\|x'-u\| \geq \delta\}}$$

$$\leq \frac{1}{c} \sum_{j=1}^{+\infty} \sup_{s \in B(v_j, \rho/2)} K(s) \mathbf{1}_{\{\|s_1\| \geq \delta/\alpha\}},$$

where $s_1 \in \mathbb{R}^d$ denotes the vector of the d first coordinates of $s \in \mathbb{R}^{d+p}$. Let us check that this quantity converges to 0 as α tends to 0. Indeed, it is a series with nonnegative terms, of the type

$$\sum_{j=1}^{+\infty} \varphi_j(\alpha),$$

which is finite for every α thanks to the first part of the covering lemma. Since $\lim_{\alpha \to 0} \varphi_j(\alpha) = 0$ for every $j \geq 1$, the series converges to 0.

6.3 Proof of Proposition 3.1

1. Since f is supposed to be continuous with compact support, it is bounded and uniformly continuous on its support: there exists $\delta > 0$ such that $\|x - u\| < \delta$ implies $|f(x) - f(u)| < \varepsilon'$. We have

$$\int \left| f(x) - \frac{\mathbb{E}\left[f(X) g\left(\frac{X-x}{\alpha}, \frac{\mathbf{f}(X) - \mathbf{f}(x)}{\beta} \right) \right]}{\mathbb{E}\left[g\left(\frac{X-x}{\alpha}, \frac{\mathbf{f}(X) - \mathbf{f}(x)}{\beta} \right) \right]} \right| \mu(dx)$$

$$= \int \left| f(x) - \frac{\int f(u) g\left(\frac{u-x}{\alpha}, \frac{\mathbf{f}(u) - \mathbf{f}(x)}{\beta} \right) \mu(du)}{\mathbb{E}\left[g\left(\frac{X-x}{\alpha}, \frac{\mathbf{f}(X) - \mathbf{f}(x)}{\beta} \right) \right]} \right| \mu(dx)$$

$$= \int \left| \frac{f(x) \mathbb{E}\left[g\left(\frac{X-x}{\alpha}, \frac{\mathbf{f}(X) - \mathbf{f}(x)}{\beta} \right) \right] - \int f(u) g\left(\frac{u-x}{\alpha}, \frac{\mathbf{f}(u) - \mathbf{f}(x)}{\beta} \right) \mu(du)}{\mathbb{E}\left[g\left(\frac{X-x}{\alpha}, \frac{\mathbf{f}(X) - \mathbf{f}(x)}{\beta} \right) \right]} \right| \mu(dx)$$

$$= \int \left| \frac{f(x) \int g\left(\frac{u-x}{\alpha}, \frac{\mathbf{f}(u) - \mathbf{f}(x)}{\beta} \right) \mu(du) - \int f(u) g\left(\frac{u-x}{\alpha}, \frac{\mathbf{f}(u) - \mathbf{f}(x)}{\beta} \right) \mu(du)}{\mathbb{E}\left[g\left(\frac{X-x}{\alpha}, \frac{\mathbf{f}(X) - \mathbf{f}(x)}{\beta} \right) \right]} \right| \mu(dx)$$

$$= \int \left| \frac{\int g\left(\frac{u-x}{\alpha}, \frac{\mathbf{f}(u) - \mathbf{f}(x)}{\beta} \right) (f(x) - f(u)) \mu(du)}{\mathbb{E}\left[g\left(\frac{X-x}{\alpha}, \frac{\mathbf{f}(X) - \mathbf{f}(x)}{\beta} \right) \right]} \right| \mu(dx)$$

$$\leq \int \int \frac{g\left(\frac{u-x}{\alpha}, \frac{\mathbf{f}(u) - \mathbf{f}(x)}{\beta} \right)}{\mathbb{E}\left[g\left(\frac{X-x}{\alpha}, \frac{\mathbf{f}(X) - \mathbf{f}(x)}{\beta} \right) \right]} |f(x) - f(u)| \mu(du) \mu(dx)$$

$$= \int \int_{B(x,\delta)} \frac{g\left(\frac{u-x}{\alpha}, \frac{\mathbf{f}(u) - \mathbf{f}(x)}{\beta} \right)}{\mathbb{E}\left[g\left(\frac{X-x}{\alpha}, \frac{\mathbf{f}(X) - \mathbf{f}(x)}{\beta} \right) \right]} |f(x) - f(u)| \mu(du) \mu(dx)$$

$$+ \int \int \frac{g\left(\frac{u-x}{\alpha}, \frac{\mathbf{f}(u) - \mathbf{f}(x)}{\beta} \right)}{\mathbb{E}\left[g\left(\frac{X-x}{\alpha}, \frac{\mathbf{f}(X) - \mathbf{f}(x)}{\beta} \right) \right]} \mathbf{1}_{\{\|x-u\| \geq \delta\}}(u) |f(x) - f(u)| \mu(du) \mu(dx)$$

The first term may be upper bounded by ε' by uniform continuity of f, since, for every x,

$$\int_{B(x,\delta)} \frac{g\left(\frac{u-x}{\alpha}, \frac{\mathbf{f}(u) - \mathbf{f}(x)}{\beta} \right)}{\mathbb{E}\left[g\left(\frac{X-x}{\alpha}, \frac{\mathbf{f}(X) - \mathbf{f}(x)}{\beta} \right) \right]} \mu(dy) \leq 1.$$

For the second term, we first use $|f(x) - f(u)| \leq 1$, then, the second point in the covering lemma 3.4 implies that the integral against $\mu(du)$ tends to 0, and the convergence to 0 of the integral against $\mu(dx)$ follows from Lebesgue's dominated convergence theorem, since the integrand is bounded by R according to the first point in the covering lemma 3.4. Thus, we get

$$\limsup_{n \to \infty} \int \left| f(x) - \frac{\mathbb{E}\left[f(X) g\left(\frac{X-x}{\alpha}, \frac{\mathbf{f}(X) - \mathbf{f}(x)}{\beta} \right) \right]}{\mathbb{E}\left[g\left(\frac{X-x}{\alpha}, \frac{\mathbf{f}(X) - \mathbf{f}(x)}{\beta} \right) \right]} \right| \mu(dx) \leq \varepsilon'.$$

2. We have

$$\int \left| \frac{\mathbb{E}\left[f(X) g\left(\frac{X-x}{\alpha}, \frac{\mathbf{f}(X)-\mathbf{f}(x)}{\beta} \right) \right]}{\mathbb{E}\left[g\left(\frac{X-x}{\alpha}, \frac{\mathbf{f}(X)-\mathbf{f}(x)}{\beta} \right) \right]} - \mathbb{E}[\mathcal{T}_n^\star(x)] \right| \mu(dx)$$

$$= \int \left| \frac{\int f(u) g\left(\frac{u-x}{\alpha}, \frac{\mathbf{f}(u)-\mathbf{f}(x)}{\beta} \right) \mu(du)}{\mathbb{E}\left[g\left(\frac{X-x}{\alpha}, \frac{\mathbf{f}(X)-\mathbf{f}(x)}{\beta} \right) \right]} - \mathbb{E}[\mathcal{T}_n^\star(x)] \right| \mu(dx)$$

$$= \int \left| \frac{\int f(u) g\left(\frac{u-x}{\alpha}, \frac{\mathbf{f}(u)-\mathbf{f}(x)}{\beta} \right) \mu(du)}{\mathbb{E}\left[g\left(\frac{X-x}{\alpha}, \frac{\mathbf{f}(X)-\mathbf{f}(x)}{\beta} \right) \right]} - \frac{\mathbb{E}\left[\sum_{i=1}^n Y_i g\left(\frac{X_i-x}{\alpha}, \frac{\mathbf{f}(X_i)-\mathbf{f}(x)}{\beta} \right) \right]}{n \mathbb{E}\left[g\left(\frac{X-x}{\alpha}, \frac{\mathbf{f}(X)-\mathbf{f}(x)}{\beta} \right) \right]} \right| \mu(dx)$$

$$= \int \left| \frac{\int f(u) g\left(\frac{u-x}{\alpha}, \frac{\mathbf{f}(u)-\mathbf{f}(x)}{\beta} \right) \mu(du) - \mathbb{E}\left[\eta(X) g\left(\frac{X-x}{\alpha}, \frac{\mathbf{f}(X)-\mathbf{f}(x)}{\beta} \right) \right]}{\mathbb{E}\left[g\left(\frac{X-x}{\alpha}, \frac{\mathbf{f}(X)-\mathbf{f}(x)}{\beta} \right) \right]} \right| \mu(dx)$$

$$= \int \left| \frac{\int (f(u) - \eta(u)) g\left(\frac{u-x}{\alpha}, \frac{\mathbf{f}(u)-\mathbf{f}(x)}{\beta} \right) \mu(du)}{\mathbb{E}\left[g\left(\frac{X-x}{\alpha}, \frac{\mathbf{f}(X)-\mathbf{f}(x)}{\beta} \right) \right]} \right| \mu(dx)$$

$$\leq \int \int |f(u) - \eta(u)| \frac{g\left(\frac{u-x}{\alpha}, \frac{\mathbf{f}(u)-\mathbf{f}(x)}{\beta} \right)}{\mathbb{E}\left[g\left(\frac{X-x}{\alpha}, \frac{\mathbf{f}(X)-\mathbf{f}(x)}{\beta} \right) \right]} \mu(du) \mu(dx)$$

$$= \int \int \frac{g\left(\frac{u-x}{\alpha}, \frac{\mathbf{f}(u)-\mathbf{f}(x)}{\beta} \right)}{\mathbb{E}\left[g\left(\frac{X-x}{\alpha}, \frac{\mathbf{f}(X)-\mathbf{f}(x)}{\beta} \right) \right]} \mu(dx) |f(u) - \eta(u)| \mu(du)$$

$$\leq \int R |f(u) - \eta(u)| \mu(du) \leq R\varepsilon'.$$

Here, we used the Fubini Theorem, then the covering lemma, and finally the definition of f.

3. For the last term $\int \mathbb{E}\left[|\mathbb{E}[\mathcal{T}_n^\star(x)] - \mathcal{T}_n^\star(x)| \right] \mu(dx)$, let us write, for every x,

$$\left(\mathbb{E}\left[|\mathbb{E}[\mathcal{T}_n^\star(x)] - \mathcal{T}_n^\star(x)| \right] \right)^2 \leq \mathbb{E}\left[|\mathbb{E}[\mathcal{T}_n^\star(x)] - \mathcal{T}_n^\star(x)|^2 \right]$$

$$= \mathbb{E}\left[\left| \frac{\frac{1}{n} \sum_{i=1}^n \mathbb{E}\left[Y_i g\left(\frac{X_i-x}{\alpha}, \frac{\mathbf{f}(X_i)-\mathbf{f}(x)}{\beta} \right) \right]}{\mathbb{E}\left[g\left(\frac{X-x}{\alpha}, \frac{\mathbf{f}(X)-\mathbf{f}(x)}{\beta} \right) \right]} - \frac{\frac{1}{n} \sum_{i=1}^n Y_i g\left(\frac{X_i-x}{\alpha}, \frac{\mathbf{f}(X_i)-\mathbf{f}(x)}{\beta} \right)}{\mathbb{E}\left[g\left(\frac{X-x}{\alpha}, \frac{\mathbf{f}(X)-\mathbf{f}(x)}{\beta} \right) \right]} \right|^2 \right]$$

$$= \frac{\mathbb{E}\left[\left| \frac{1}{n} \sum_{i=1}^n \left(\mathbb{E}\left[Y_i g\left(\frac{X_i-x}{\alpha}, \frac{\mathbf{f}(X_i)-\mathbf{f}(x)}{\beta} \right) \right] - Y_i g\left(\frac{X_i-x}{\alpha}, \frac{\mathbf{f}(X_i)-\mathbf{f}(x)}{\beta} \right) \right) \right|^2 \right]}{\left(\mathbb{E}\left[g\left(\frac{X-x}{\alpha}, \frac{\mathbf{f}(X)-\mathbf{f}(x)}{\beta} \right) \right] \right)^2}$$

$$= \frac{\mathbb{E}\left[\left| \sum_{i=1}^n \left(\mathbb{E}\left[Y g\left(\frac{X-x}{\alpha}, \frac{\mathbf{f}(X)-\mathbf{f}(x)}{\beta} \right) \right] - Y_i g\left(\frac{X_i-x}{\alpha}, \frac{\mathbf{f}(X_i)-\mathbf{f}(x)}{\beta} \right) \right) \right|^2 \right]}{n^2 \left(\mathbb{E}\left[g\left(\frac{X-x}{\alpha}, \frac{\mathbf{f}(X)-\mathbf{f}(x)}{\beta} \right) \right] \right)^2}$$

Using that the variance of a sum of n i.i.d variables is n times the variance of one such variable, and then upper bounding this variance by the expectation of the squared variable, we obtain, for

18

every x,

$$\left(\mathbb{E}[|\mathbb{E}[\mathcal{T}_n^\star(x)] - \mathcal{T}_n^\star(x)|]\right)^2$$

$$\leq \frac{n\mathbb{E}\left[\left(\mathbb{E}\left[Yg\left(\frac{X-x}{\alpha}, \frac{\mathbf{f}(X)-\mathbf{f}(x)}{\beta}\right)\right] - Yg\left(\frac{X-x}{\alpha}, \frac{\mathbf{f}(X)-\mathbf{f}(x)}{\beta}\right)\right)^2\right]}{n^2\left(\mathbb{E}\left[g\left(\frac{X-x}{\alpha}, \frac{\mathbf{f}(X)-\mathbf{f}(x)}{\beta}\right)\right]\right)^2}$$

$$= \frac{\mathbb{E}\left[\left(\mathbb{E}\left[Yg\left(\frac{X-x}{\alpha}, \frac{\mathbf{f}(X)-\mathbf{f}(x)}{\beta}\right)\right] - Yg\left(\frac{X-x}{\alpha}, \frac{\mathbf{f}(X)-\mathbf{f}(x)}{\beta}\right)\right)^2\right]}{n\left(\mathbb{E}\left[g\left(\frac{X-x}{\alpha}, \frac{\mathbf{f}(X)-\mathbf{f}(x)}{\beta}\right)\right]\right)^2}$$

$$\leq \frac{\mathbb{E}\left[\left(Yg\left(\frac{X-x}{\alpha}, \frac{\mathbf{f}(X)-\mathbf{f}(x)}{\beta}\right)\right)^2\right]}{n\left(\mathbb{E}\left[g\left(\frac{X-x}{\alpha}, \frac{\mathbf{f}(X)-\mathbf{f}(x)}{\beta}\right)\right]\right)^2}$$

$$\leq \frac{\mathbb{E}\left[g\left(\frac{X-x}{\alpha}, \frac{\mathbf{f}(X)-\mathbf{f}(x)}{\beta}\right)^2\right]}{n\left(\mathbb{E}\left[g\left(\frac{X-x}{\alpha}, \frac{\mathbf{f}(X)-\mathbf{f}(x)}{\beta}\right)\right]\right)^2},$$

since Y is bounded by 1. Let G denote an upper bound on g, which exists because the kernel K is regular. Thus, for every x, $g\left(\frac{X-x}{\alpha}, \frac{\mathbf{f}(X)-\mathbf{f}(x)}{\beta}\right) \leq G$, so that

$$\frac{1}{G}g\left(\frac{X-x}{\alpha}, \frac{\mathbf{f}(X)-\mathbf{f}(x)}{\beta}\right) \leq 1.$$

Consequently, we get

$$\frac{g\left(\frac{X-x}{\alpha}, \frac{\mathbf{f}(X)-\mathbf{f}(x)}{\beta}\right)^2}{G^2} \leq \frac{g\left(\frac{X-x}{\alpha}, \frac{\mathbf{f}(X)-\mathbf{f}(x)}{\beta}\right)}{G}.$$

Then

$$\frac{\mathbb{E}\left[g\left(\frac{X-x}{\alpha}, \frac{\mathbf{f}(X)-\mathbf{f}(x)}{\beta}\right)^2\right]}{G^2} \leq \frac{\mathbb{E}\left[g\left(\frac{X-x}{\alpha}, \frac{\mathbf{f}(X)-\mathbf{f}(x)}{\beta}\right)\right]}{G},$$

and thus,

$$\frac{\mathbb{E}\left[g\left(\frac{X-x}{\alpha}, \frac{\mathbf{f}(X)-\mathbf{f}(x)}{\beta}\right)^2\right]}{\mathbb{E}\left[g\left(\frac{X-x}{\alpha}, \frac{\mathbf{f}(X)-\mathbf{f}(x)}{\beta}\right)\right]} \leq G.$$

Applying this result to the inequality above yields

$$\left(\mathbb{E}[|\mathbb{E}[\mathcal{T}_n^\star(x)] - \mathcal{T}_n^\star(x)|]\right)^2 \leq \frac{G}{n\mathbb{E}\left[g\left(\frac{X-x}{\alpha}, \frac{\mathbf{f}(X)-\mathbf{f}(x)}{\beta}\right)\right]}$$

$$\leq \frac{G}{nc\mathbb{P}(\|\frac{z-Z}{\alpha}\| \leq \rho)},$$

with $Z = (X_1, \ldots, X_d, \frac{\alpha}{\beta}f_1(X), \ldots, \frac{\alpha}{\beta}f_p(X))$ and $z = (x_1, \ldots, x_d, \frac{\alpha}{\beta}f_1(x) \ldots, \frac{\alpha}{\beta}f_p(x))$. We are interested in the integral against $\mu(dx)$ of this quantity.

Let us divide this integral between an integral over a cube of side 2ℓ centered at the origin $\mathcal{Q} = [-\ell, \ell]^d$ and an integral over the complement of this cube. We may write

$$\int_{\mathcal{Q}^c} \mathbb{E}[|\mathbb{E}[\mathcal{T}_n^\star(x)] - \mathcal{T}_n^\star(x)|]\mu(dx) \leq 2\int_{\mathcal{Q}^c} \mathbb{E}[\mathcal{T}_n^\star(x)]\mu(dx).$$

This quantity converges to

$$2 \int_{\mathcal{Q}^c} \eta(x) \mu(dx).$$

Indeed, the upper bounds obtained for the integral against $\mu(dx)$ of the first 3 terms in (3.5) remain true when integrating over $\mathcal{Q}^c \subset \mathbb{R}^d$, and they ensure the convergence of $\int_{\mathcal{Q}^c} \mathbb{E}[\mathcal{T}_n^\star(x)] \mu(dx)$ to $\int_{\mathcal{Q}^c} \eta(x) \mu(dx)$. Moreover, the quantity $2 \int_{\mathcal{Q}^c} \eta(x) \mu(dx)$ can be made arbitrarily small by choosing an appropriate value of ℓ, because $\left| 2 \int_{\mathcal{Q}^c} \eta(x) \mu(dx) \right| \leq 2\mu(\mathcal{Q}^c)$ since $|Y| \leq 1$. So, for ℓ and n large enough, we have

$$\int_{\mathcal{Q}^c} \mathbb{E}[|\mathbb{E}[\mathcal{T}_n^\star(x)] - \mathcal{T}_n^\star(x)|] \mu(dx) \leq \varepsilon'.$$

For the integral over $\mathcal{Q} = [-\ell, \ell]^d$, we will recover by balls $B(v_j, \frac{r\alpha}{2})$ a parallelepiped included in $\mathbb{R}^{d+p}$, containing all the elements of the form $(x_1, \ldots, x_d, \frac{\alpha}{\beta} f_1(x), \ldots, \frac{\alpha}{\beta} f_p(x))$, with $x \in [-\ell, \ell]^d$:

$$\left\{ s \in \mathbb{R}^{d+p}, (s_1, \ldots, s_d) \in [-\ell, \ell]^d, (s_{d+1}, \ldots, s_{d+p}) \in \left[-\frac{\alpha}{\beta}, \frac{\alpha}{\beta} \right]^p \right\}.$$

This set is the parallelepiped centered at the origin with d sides of length 2ℓ and p sides of length $2\frac{\alpha}{\beta}$. Let $\lceil x \rceil$ denote the ceil of x. For a covering, the following number of balls is needed:

$$N = \left\lceil \frac{4\ell}{r\alpha} \right\rceil^d \left\lceil \frac{4}{r\beta} \right\rceil^p \leq \left(\frac{8\ell}{r\alpha} \right)^d \left(\frac{8}{r\beta} \right)^p = \frac{8^{d+p} \ell^d}{r^{d+p} \alpha^d \beta^p}.$$

We have

$$\int \frac{1}{\sqrt{\mathbb{P}(\| \frac{z-Z}{\alpha} \| \leq \rho)}} \mu(dx)$$

$$\leq \sum_{j=1}^N \int \frac{\mathbf{1}\{\| \frac{z-v_j}{\alpha} \| \leq \frac{\rho}{2}\}(x)}{\sqrt{\mathbb{P}(\| \frac{z-Z}{\alpha} \| \leq \rho)}} \mu(dx)$$

Let us check that, if $\left\| \frac{z-v_j}{\alpha} \right\| \leq \frac{\rho}{2}$, then $\left\| \frac{Z-v_j}{\alpha} \right\| \leq \frac{\rho}{2}$ implies $\left\| \frac{z-Z}{\alpha} \right\| \leq \rho$. Indeed,

$$\left\| \frac{z-Z}{\alpha} \right\| \leq \left\| \frac{z-v_j}{\alpha} \right\| + \left\| \frac{v_j-Z}{\alpha} \right\|$$

$$\leq \frac{\rho}{2} + \frac{\rho}{2} = \rho.$$

Thus,

$$\int \frac{1}{\sqrt{\mathbb{P}\left(\left\| \frac{z-Z}{\alpha} \right\| \leq \rho\right)}} \mu(dx)$$

$$\leq \sum_{j=1}^N \int \frac{\mathbf{1}\{\left\| \frac{z-v_j}{\alpha} \right\| \leq \frac{\rho}{2}\}(x)}{\sqrt{\mathbb{P}\left(\left\| \frac{Z-v_j}{\alpha} \right\| \leq \frac{\rho}{2}\right)}} \mu(dx)$$

$$\leq \sum_{j=1}^N \sqrt{\mathbb{P}\left(\left\| \frac{Z-v_j}{\alpha} \right\| \leq \frac{\rho}{2}\right)}$$

$$\leq \sqrt{N \sum_{j=1}^N \mathbb{P}\left(\left\| \frac{Z-v_j}{\alpha} \right\| \leq \frac{\rho}{2}\right)} \quad \text{(Cauchy-Schwarz)}$$

$$\leq \sqrt{\frac{8^{d+p} \ell^d}{r^{d+p} \alpha^d \beta^p} \times k_{p+d}},$$

20

where k_{p+d} is the number of balls containing an element of $\mathbb{R}^{d+p}$. Indeed, the sum is over all balls, so that an upper bound is obtained by multiplying the maximal measure, equal to 1, by the number of balls containing one element.

Then, we obtain

$$\int_{\mathcal{Q}} \mathbb{E}[|\mathbb{E}[\mathcal{T}_n^\star(x)] - \mathcal{T}_n^\star(x)|] \mu(dx) \leq \sqrt{\frac{G}{nc}} \sqrt{\frac{8^{d+p} \ell^d}{r^{d+p} \alpha^d \beta^p} \times k_{p+d}}$$
$$\leq C(n\alpha^d\beta^p)^{-1/2},$$

where $C \geq 0$ is a constant depending on ℓ, the kernel K and the dimensions d and p. Finally, we have proved that for n sufficiently large,

$$\int \mathbb{E}[|\mathbb{E}[\mathcal{T}_n^\star(x)] - \mathcal{T}_n^\star(x)|] \mu(dx) \leq \varepsilon' + C(n\alpha^d\beta^p)^{-1/2}.$$

6.4 Proof of Lemma 3.1

According to equation (3.3), we have, for every $\varepsilon > 0$,

$$\mathbb{P}\left(\int |\eta(x) - \mathcal{T}_n^\star(x)| \mu(dx) \geq \varepsilon/2\right)$$
$$\leq \mathbb{P}\left(\int \mathbb{E}[|\eta(x) - \mathcal{T}_n^\star(x)|] \mu(dx) \geq \varepsilon/4\right)$$
$$+ \mathbb{P}\left(\int |\eta(x) - \mathcal{T}_n^\star(x)| \mu(dx) - \int \mathbb{E}[|\eta(x) - \mathcal{T}_n^\star(x)|] \mu(dx) \geq \varepsilon/4\right). \quad (6.1)$$

Thanks to Equation (3.5) and Proposition 3.1, provided that α tends to 0 and $n\alpha^d\beta^p$ tends to ∞, the quantity $\int \mathbb{E}[|\eta(x) - \mathcal{T}_n^\star(x)|] \mu(dx)$ tends to 0 as n tends to ∞. So, for n large enough,

$$\mathbb{P}\left(\int |\eta(x) - \mathcal{T}_n^\star(x)| \mu(dx) \geq \varepsilon/2\right) \leq \mathbb{P}\left(\int |\eta(x) - \mathcal{T}_n^\star(x)| \mu(dx) - \int \mathbb{E}[|\eta(x) - \mathcal{T}_n^\star(x)|] \mu(dx) \geq \varepsilon/4\right)$$
$$\leq \mathbb{P}\left(\int |\eta(x) - \mathcal{T}_n^\star(x)| \mu(dx) - \mathbb{E}\left[\int |\eta(x) - \mathcal{T}_n^\star(x)| \mu(dx)\right] \geq \varepsilon/4\right).$$

For controlling this quantity, McDiarmid's inequality will be used. Let $t_n^\star(x)$ denote $\mathcal{T}_n^\star(x)$ for values $(x_1, \ldots, x_n)$ and $t_n^{\star\prime}(x)$ be the version of $t_n^\star(x)$ where (x_i, y_i) has been replaced by (x_i', y_i'). We compute

$$\int |\eta(x) - t_n^\star(x)| \mu(dx) - \int |\eta(x) - t_n^{\star\prime}(x)| \mu(dx) \leq \int |t_n^\star(x) - t_n^{\star\prime}(x)| \mu(dx)$$
$$\leq \frac{\int \left|y_i g\left(\frac{x_i - x}{\alpha}, \frac{\mathbf{f}(x_i) - \mathbf{f}(x)}{\beta}\right) - y_i' g\left(\frac{x_i' - x}{\alpha}, \frac{\mathbf{f}(x_i') - \mathbf{f}(x)}{\beta}\right)\right| \mu(dx)}{n\mathbb{E}\left[g\left(\frac{X - x}{\alpha}, \frac{\mathbf{f}(X) - \mathbf{f}(x)}{\beta}\right)\right]}$$
$$\leq 2\frac{\sup_u \int g\left(\frac{u - x}{\alpha}, \frac{\mathbf{f}(u) - \mathbf{f}(x)}{\beta}\right) \mu(dx)}{n\mathbb{E}\left[g\left(\frac{X - x}{\alpha}, \frac{\mathbf{f}(X) - \mathbf{f}(x)}{\beta}\right)\right]}$$
$$\leq \frac{2R}{n}.$$

Then, McDiarmid's inequality implies

$$\mathbb{P}\left(\int |\eta(x) - \mathcal{T}_n^\star(x)| \mu(dx) - \mathbb{E}\left[\int |\eta(x) - \mathcal{T}_n^\star(x)| \mu(dx)\right] \geq \varepsilon/4\right) \leq \exp\left(-\frac{2\varepsilon^2}{16\sum_{i=1}^n (\frac{2R}{n})^2}\right) = \exp\left(-\frac{n\varepsilon^2}{32R^2}\right).$$

6.5 Proof of Lemma 3.2

For $x \in \mathbb{R}^d$,

$$
\begin{aligned}
|\mathcal{T}_n^\star(x) - \mathcal{T}_n(x)| &= \left| \frac{\sum_{i=1}^n Y_i K_\alpha(Z_i - z)}{n \mathbb{E}[K_\alpha(Z - z)]} - \frac{\sum_{i=1}^n Y_i K_\alpha(Z_i - z)}{\sum_{i=1}^n K_\alpha(Z_i - z)} \right| \\
&= \left| \sum_{i=1}^n Y_i K_\alpha(Z_i - z) \right| \left| \frac{1}{n\mathbb{E}[K_\alpha(Z - z)]} - \frac{1}{\sum_{i=1}^n K_\alpha(Z_i - z)} \right| \\
&\leq \left| \sum_{i=1}^n K_\alpha(Z_i - z) \right| \left| \frac{1}{n\mathbb{E}[K_\alpha(Z - z)]} - \frac{1}{\sum_{i=1}^n K_\alpha(Z_i - z)} \right| \\
&= \left| \frac{\sum_{i=1}^n K_\alpha(Z_i - z)}{n\mathbb{E}[K_\alpha(Z - z)]} - 1 \right|.
\end{aligned}
$$

So, by Lemma 3.1, applied in the particular case where Y is constant, equal to 1, we obtain for every $\varepsilon > 0$,

$$
\mathbb{P}\left(\int |\mathcal{T}_n^\star(x) - \mathcal{T}_n(x)| \mu(dx) > \varepsilon/2 \right) \leq \exp\left(-\frac{n\varepsilon^2}{8R^2} \right).
$$

6.6 Proof of the main theorems

Finally, to prove Theorem 3.1, it suffices to use Equation (3.1) and to combine Lemma 3.1 and Lemma 3.2. For every $\varepsilon > 0$, we obtain, as soon as n is large enough,

$$
\begin{aligned}
\mathbb{P}\left(\int |\eta(x) - \mathcal{T}_n(x)| \mu(dx) > \varepsilon \right) &\leq \mathbb{P}\left(\int |\eta(x) - \mathcal{T}_n^\star(x)| \mu(dx) > \varepsilon/2 \right) + \mathbb{P}\left(\int |\mathcal{T}_n^\star(x) - \mathcal{T}_n(x)| \mu(dx) > \varepsilon/2 \right) \\
&\leq \exp\left(-\frac{n\varepsilon^2}{32R^2} \right) + \exp\left(-\frac{n\varepsilon^2}{8R^2} \right) \\
&\leq 2\exp\left(-\frac{n\varepsilon^2}{32R^2} \right).
\end{aligned}
$$

For Theorem 3.2, Equation (3.2) and Lemma 3.1 show that, for every $\varepsilon > 0$, if n is large enough,

$$
\begin{aligned}
&\mathbb{P}(L_n - L^\star > \varepsilon) \\
&\leq \mathbb{P}\left(\int \left| 1 - \eta(x) - \frac{\sum_{i=1}^n K_\alpha(Z_i - z)}{n\mathbb{E}[K_\alpha(Z - z)]} + \mathcal{T}_n^\star(x) \right| \mu(dx) > \varepsilon/2 \right) + \mathbb{P}\left(\int |\eta(x) - \mathcal{T}_n^\star(x)| \mu(dx) > \varepsilon/2 \right) \\
&\hspace{10cm} \leq 2\exp\left(-\frac{n\varepsilon^2}{32R^2} \right).
\end{aligned}
$$

7 Annex

- The "gauss" example is a mixture of two Gaussian distributions $\mathcal{N}\left(\left(\begin{smallmatrix} 0 \\ 2 \end{smallmatrix} \right), \left(\begin{smallmatrix} 1 & -0.5 \\ -0.5 & 1 \end{smallmatrix} \right) \right)$ and $\mathcal{N}\left(\left(\begin{smallmatrix} -1 \\ 2 \end{smallmatrix} \right), \left(\begin{smallmatrix} 1 & 0.5 \\ 0.5 & 1 \end{smallmatrix} \right) \right)$.

- The "comete" example is a mixture of two Gaussian distributions $\mathcal{N}\left(\left(\begin{smallmatrix} 0 \\ 2 \end{smallmatrix} \right), \left(\begin{smallmatrix} 3 & 9/4 \\ 9/4 & 15 \end{smallmatrix} \right) \right)$ and $\mathcal{N}\left(\left(\begin{smallmatrix} 0 \\ 2 \end{smallmatrix} \right), \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix} \right) \right)$.

- The "nuclear" example is a mixture of a uniform distribution $\mathcal{U}(\mathcal{D}^1)$ on the unit disk and a Gaussian distribution $\mathcal{N}\left(\left(\begin{smallmatrix} 0.5 \\ 0.5 \end{smallmatrix} \right), \left(\begin{smallmatrix} 0.1 & 0 \\ 0 & 0.1 \end{smallmatrix} \right) \right)$.

- The "spot" example is a mixture of a uniform $\mathcal{U}([-1,1]^2)$ and 5 Gaussian distributions arbitrary chosen on the unit square.

8 Acknowledgement

References

Jean-Yves Audibert. Aggregated estimators and empirical complexity for least square regression. *Annales de l'Institut Henri Poincaré (B) Probabilités et Statistique*, 40:685–736, 2004.

N. Balakrishnan and M. Mojirsheibani. A simple method for combining estimates to improve the overall error rates in classification. *Computational Statistics*, pages 1–17, 2015.

G. Biau, A. Fischer, B. Guedj, and J.D. Malley. Cobra:a combined regression strategy. *Journal of Multivariate Analysis*, 146(C):18–28, 2016.

Florentina Bunea, Alexandre B. Tsybakov, and Marten H. Wegkamp. Aggregation and sparsity via ℓ_1-penalized least squares. In Gábor Lugosi and H. U. Simon, editors, *Proceedings of 19th Annual Conference on Learning Theory (COLT 2006), Lecture Notes in Artificial Intelligence*, volume 35, pages 379–391. Springer-Verlag, Berlin-Heidelberg, 2006.

Florentina Bunea, Alexandre B. Tsybakov, and Marten H. Wegkamp. Aggregation for gaussian regression. *The Annals of Statistics*, 35:1674–1697, 2007a.

Florentina Bunea, Alexandre B. Tsybakov, and Marten H. Wegkamp. Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics*, 35:169–194, 2007b.

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

Olivier Catoni. *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, 2004.

A Cholaquidis, R. Fraiman, J. Kalemkerian, and P. Llop. A nonlinear aggregation type classifier. *Journal of Multivariate Analysis*, 146:269–281, 2016.

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

L Devroye and A. Krzyżak. An equivalence theorem for l1 convergence of the kernel regression estimate. *Journal of Statistical Planning and Inference*, 23:71–82, 1989.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Applications of Mathematics. Springer, New York, 1996.

Benjamin Guedj. *COBRA: Nonlinear Aggregation of Predictors*, 2013. URL `http://cran.r-project.org/web/packages/COBRA/index.html`. R package version 0.99.4.

L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A distribution-free theory of nonparametric regression*. Springer Series in Statistics. Springer, New York, 2002.

Anatoli Juditsky and Arkadi Nemirovski. Functional aggregation for nonparametric estimation. *The Annals of Statistics*, 28:681–712, 2000.

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

Majid Mojirsheibani. Combining classifiers via discretization. *Journal of the American Statistical Association*, 94:600–609, 1999.

Majid Mojirsheibani. A kernel-based combined classification rule. *Statistics & Probability Letters*, 48: 411–419, 2000.

Majid Mojirsheibani. An almost surely optimal combined classification rule. *Journal of Multivariate Analysis*, 81:28–46, 2002a.

Majid Mojirsheibani. A comparison study of some combined classifiers. *Communications in Statistics - Simulation and Computation*, 31:245–260, 2002b.

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

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

Yuhong Yang. Combining different procedures for adaptive regression. *Journal of Multivariate Analysis*, 74:135–161, 2000.

Yuhong Yang. Adaptive regression by mixing. *Journal of the American Statistical Association*, 96: 574–588, 2001.

Yuhong Yang. Aggregating regression procedures to improve performance. *Bernoulli*, 10:25–47, 2004.